Privacy-Preserving Medical Data Sharing

Aris Gkoulalas-Divanis*
arisdiva@ie.ibm.com
IBM Research - Ireland

Grigorios Loukides*
g.loukides@cs.cf.ac.uk
Cardiff University

SIAM Data Mining, Anaheim, CA, USA, April 2012

* Part of the reported research was conducted while the authors were employed by Vanderbilt University

Topic of the tutorial

demographics, billing info, DNA, clinical notes

improve healthcare provisioning, medical research

Allow medical data to be shared in a way that preserves patients' privacy and data utility

privacy legislation, attacks, disclosures, privacy models

support medical research, decision making, personalized medicine

Content

- Part 1: Motivation: medical data sharing and use

- Part 2: Research challenges and state-of-the-art solutions

- Part 3: Open problems and research directions

Part 1 - Content

- Part 1: *Medical data sharing and the need for privacy*
  - Patient data: EMRs, sharing, and use in applications
  - Introduction to privacy-preserving data sharing
- Part 2: *Research challenges and solutions*
- Part 3: *Open problems and research directions*

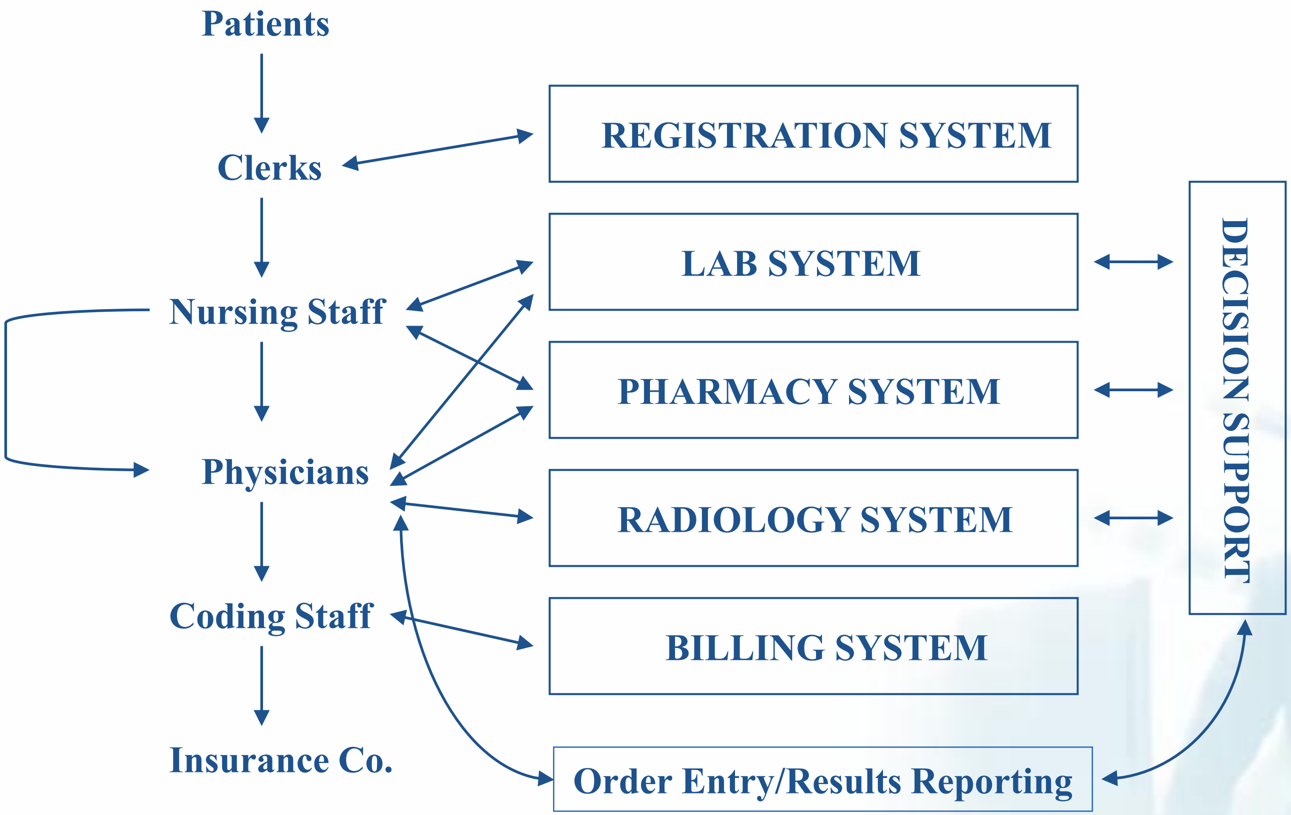
Patient data

- **Patient data**
  - Registration data (e.g., contact info, SSN)
  - Demographics (e.g., DOB, gender, race)
  - Billing information (e.g., diagnosis codes)
  - Genomic information (e.g., SNPs)
  - Medication and allergies
  - Immunization status
  - Laboratory test results
  - Radiology images
  - …

EMR System – *a system of systems*

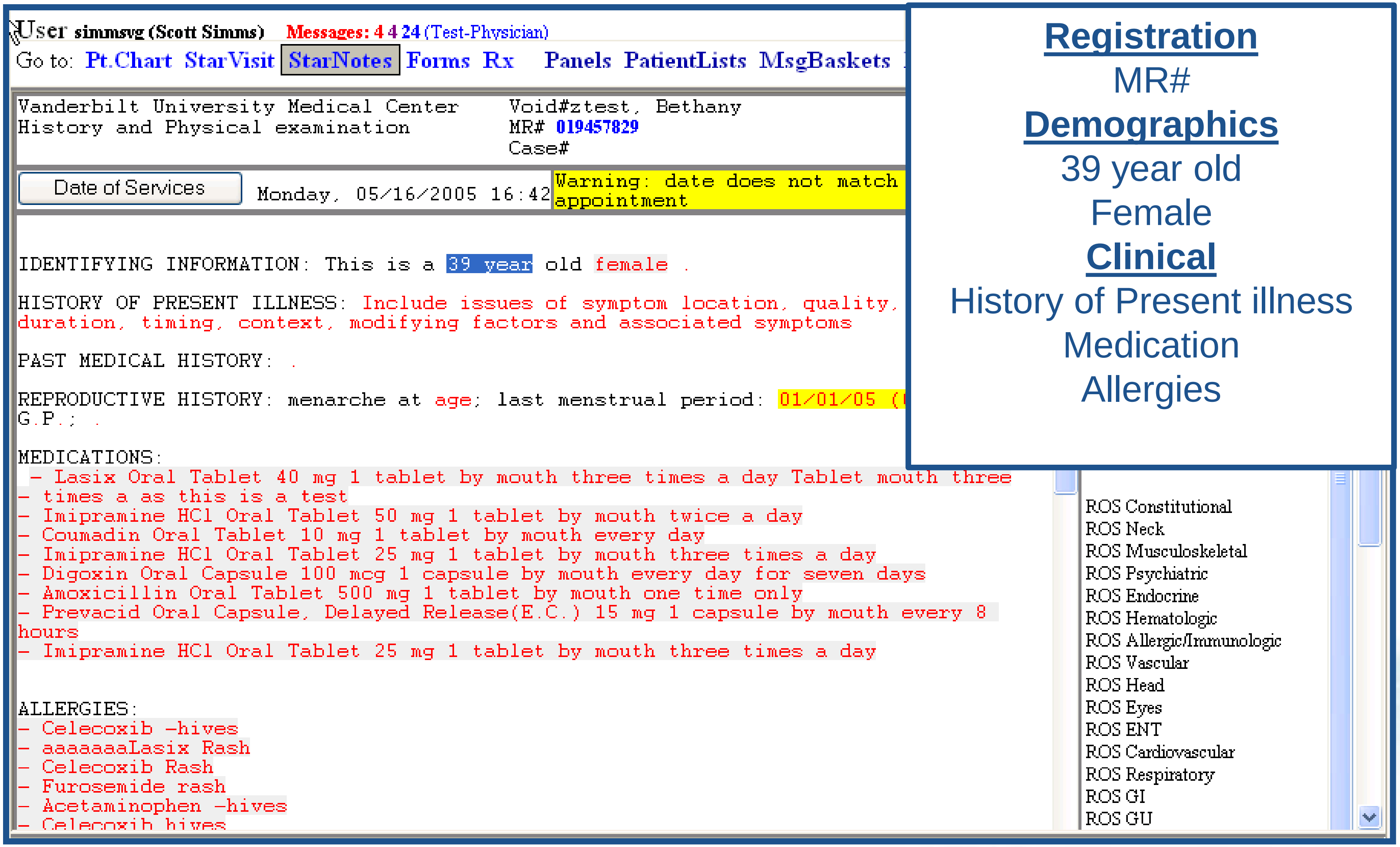
- Registration System (identifiers, date & time of visit)
- Billing System (diagnosis codes)
- Lab System (lab results)
- Radiology System (reports)
- Pharmacy System (medications)
- *Order Entry System (orders, prescriptions)*
- *Decision Support System (clinical knowledge, guidelines)*

Interaction with an EMR System



Slide adapted from W. Tierney

A view from VUMC’s EMR



User simmsvg (Scott Simms) Messages: 4 4 24 (Test-Physician)

Go to: Pt.Chart StarVisit StarNotes Forms Rx Panels PatientLists MsgBaskets

Vanderbilt University Medical Center
History and Physical examination

Void#ztest, Bethany
MR# 019457829
Case#

Date of Services Monday, 05/16/2005 16:42 Warning: date does not match appointment

IDENTIFYING INFORMATION: This is a 39 year old female .

HISTORY OF PRESENT ILLNESS: Include issues of symptom location, quality, duration, timing, context, modifying factors and associated symptoms

PAST MEDICAL HISTORY: .

REPRODUCTIVE HISTORY: menarche at age; last menstrual period: 01/01/05 (
G.P.; .

MEDICATIONS:
- Lasix Oral Tablet 40 mg 1 tablet by mouth three times a day Tablet mouth three
- times a as this is a test
- Imipramine HCl Oral Tablet 50 mg 1 tablet by mouth twice a day
- Coumadin Oral Tablet 10 mg 1 tablet by mouth every day
- Imipramine HCl Oral Tablet 25 mg 1 tablet by mouth three times a day
- Digoxin Oral Capsule 100 mcg 1 capsule by mouth every day for seven days
- Amoxicillin Oral Tablet 500 mg 1 tablet by mouth one time only
- Prevacid Oral Capsule, Delayed Release(E.C.) 15 mg 1 capsule by mouth every 8 hours
- Imipramine HCl Oral Tablet 25 mg 1 tablet by mouth three times a day

ALLERGIES:
- Celecoxib -hives
- aaaaaaaLasix Rash
- Celecoxib Rash
- Furosemide rash
- Acetaminophen -hives
- Celecoxib hives

Registration

MR#

Demographics

39 year old

Female

Clinical

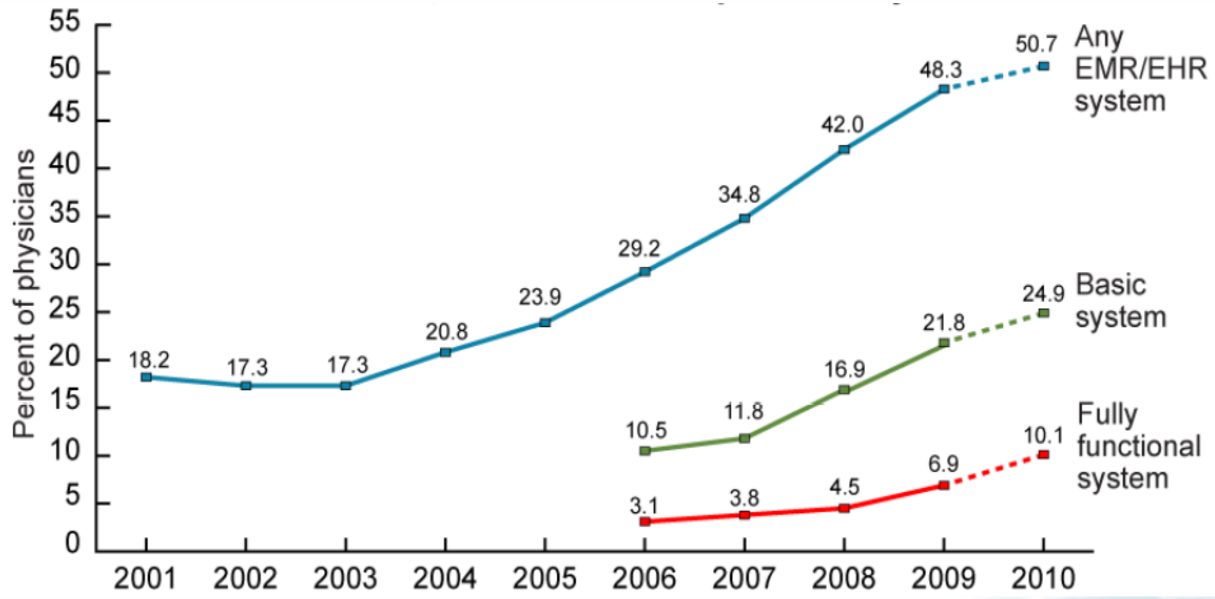
History of Present illness

Medication

Allergies

Adoption of EMRs

- **EMRs are increasingly adopted***



- Incentives by US stimulus bill ($50B) for adoption and meaningful use of EMR systems
- *Goal is to utilize an EMR for each person in the US by 2014*

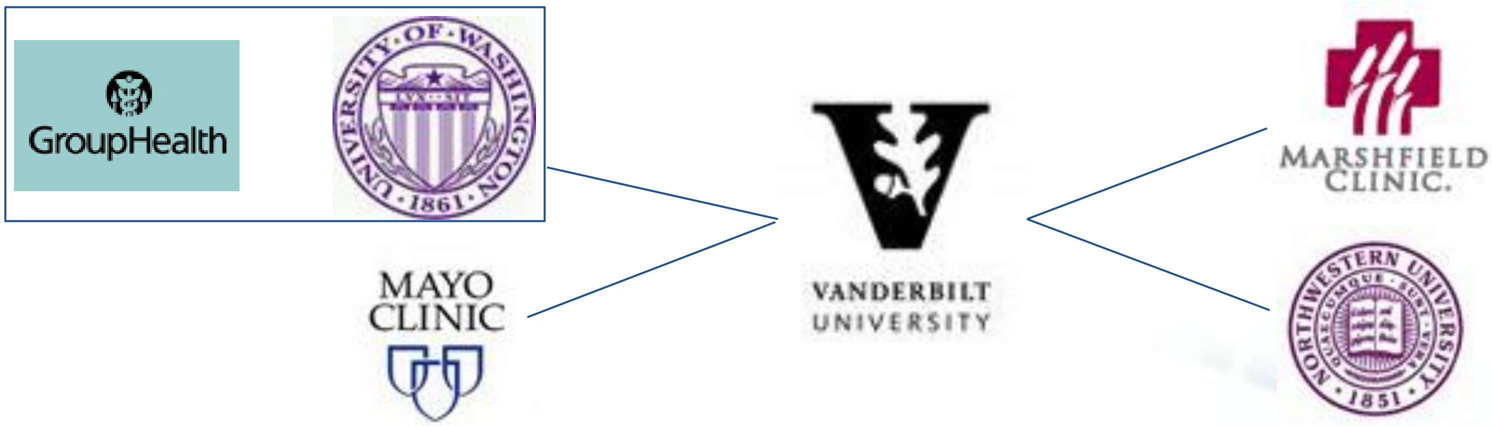
* National Ambulatory Medical Care Survey, National Center for Health Statistics, 2010

Usefulness of EMRs

- **EMRs help improve healthcare**
  - physicians to better diagnose and treat diseases
  - patients to be mobile and receive better services
- **… achieved by Health Information exchange**
  - improve accessibility of health information by physicians
  - create a standardized interoperable model that is
    - patient centric, trusted, longitudinal, scalable, sustainable, and reliable
  - e.g., Wisconsin Health Information Exchange,
    MidSouth E-health Alliance
  - HL7 – standard for information exchange between various
    healthcare systems

Usefulness of EMRs

- **EMRs help support “local” research**
  - **electronic Medical Records & Genomics (eMERGE) Consortium**



  - **Sharing diagnosis codes and DNA from EMRs to enable large-scale, low-cost GWAS for many disorders**
    - GWAS on asthma* - all patients with an ICD code of 493.xx, as well as all patients on asthma medications

* Pacheco et al. A Highly Specific Algorithm for Identifying Asthma Cases and Controls for Genome-Wide Association Studies. AMIA, 2009.

Patient data management and sharing

- **Support “broad” research**

- **Database of Genotypes and Phenotypes (dbGaP)**
  - archive and distribute data collected for GWAS
  - established in 2006 and funded by the National Center for Biotechnology Information (NCBI), NIH

  - **Tiered data access**
    - Aggregated data (e.g., questionnaires) – open to the public
    - Person-specific data (e.g., genotypes) – PIs need to apply for access

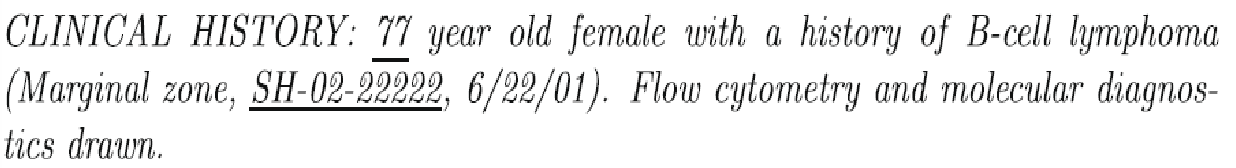
  - **Data protection**
    - Security (off-line servers, secure FTP, encryption)
    - Privacy (more on this later)

EMR data representation

- **Relational data**
  - Registration and demographic data
- **Transaction (set-valued) data**
  - Billing information
    - ICD codes are represented as numbers (up to 5 digits) and denote signs, findings, and causes of injury or disease*
- **Sequential data**
  - DNA
- **Text data**
  - Clinical notes

Electronic Medical Records

Name	YOB	ICD	DNA
Jim	1955	493.00, 185	C…T
Mary	1943	185, 157.3	A…G
Mary	1943	493.01	C…G
Carol	1965	493.02	C…G
Anne	1973	157.9, 493.03	G…C
Anne	1973	157.3	A…T



CLINICAL HISTORY: 77 year old female with a history of B-cell lymphoma (Marginal zone, SH-02-22222, 6/22/01). Flow cytometry and molecular diagnostics drawn.

* Centers for Medicare & Medicaid Services - https://www.cms.gov/icd9providerdiagnosticcodes/

EMR data in analytics

- **Statistical analysis**
  - **Correlation between YOB and ICD code 185** *(Malignant neoplasm of prostate)*
- **Querying**
- **Clustering**
  - Control epidemics*
- **Classification**
  - Predict domestic violence**
- **Association rule mining**
  - Formulate a S. Korea government policy on hypertension management***

IF age in [43,48] AND smoke = yes AND exercise=no AND drink=yes;

THEN hypertension=yes (sup=2.9%; conf=26%)。

Electronic Medical Records

Name	YOB	ICD	DNA
Jim	1955	493.00, 493.01	C…T
Mary	1943	185	A…G
Mary	1943	493.01, 493.02	C…G
Carol	1965	493.02, 157.9	C…G
Anne	1973	157.9, 157.3	G…C
Anne	1973	157.3	A…T

* Tildesley et al. Impact of spatial clustering on disease transmission and optimal control, PNAS, 2010.

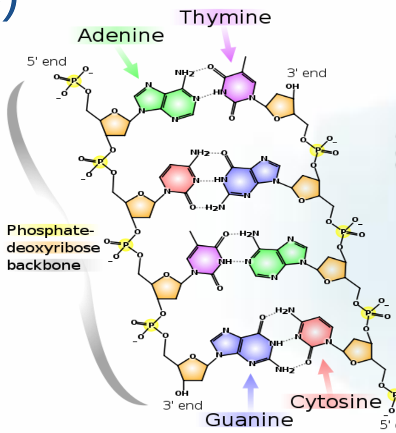
** Reis et al. Longitudinal Histories as Predictors of Future Diagnoses of Domestic Abuse: Modelling Study, BMJ: British Medical Journal, 2011

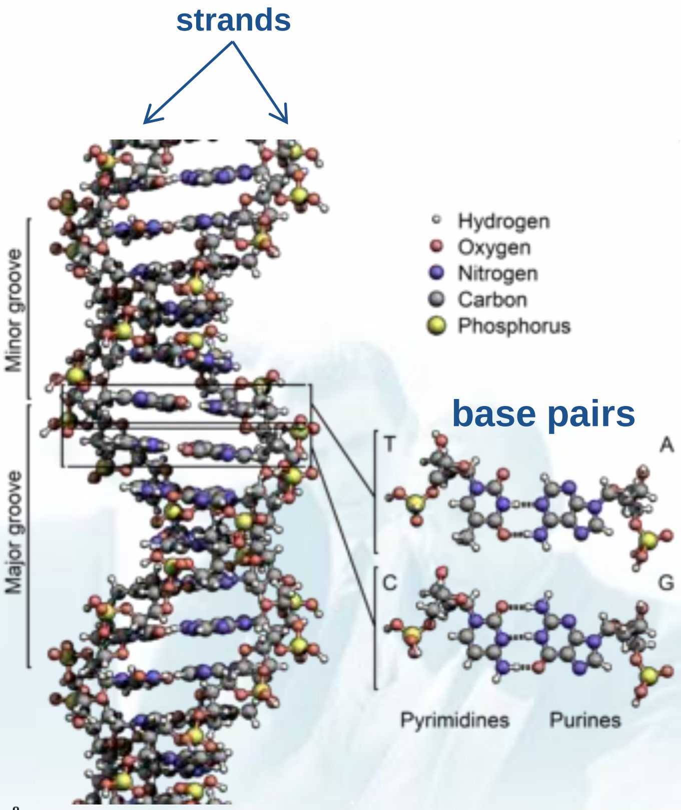
*** Chae et al. Data mining approach to policy analysis in a health insurance domain. Int. J. of Med. Inf., 2001

EMR data to support research on personalized medicine

- **Genome-Wide Association Studies (GWAS)**
  - aim to discover associations between diseases and genes
  - can help improve disease diagnosis and treatment
  - “the holy grail for personalized medicine”

- **DNA** (Deoxyribonucleic acid)
  - Genetic instructions for living organisms
  - Each strand consists of a sequence of nucleobases (A, T, G, C)
  - strands are correlated
  - DNA has 3B base pairs





structure of the DNA double helix

EMR data to support research on personalized medicine

- **Human genetic variation**

A	C	G	G	C	A	A	A	T	Bob
A	C	G	G	G	A	A	T	T	Alice
A	C	G	G	C	A	A	A	T	Tom

Single Nucleotide Polymorphism (SNP)

- **out of the 3B base pairs, less than 1% differ between any two persons worldwide!**

- **Scientists have identified about 11M SNPs**
  - They have specific (known) positions in the DNA
  - Are indicators of disease susceptibility, drug metabolism, ethnic heritage
  - Each SNP can have each of two possible bases ("values")

EMR data to support research on personalized medicine

- **Why SNPs are interesting?**
  - SNPs might be associated with diseases

	SNP	
	C	G
Disease		
Healthy		

- **What is a Genome-Wide Association Study ?**
  - Each GWAS studies a disease or trait and considers about 1M SNPs
  - People are split into two groups: *case (diseased)* vs. *control (non-diseased)*
  - Statistical tests (e.g., chi-square) are used to identify genetic markers (SNPs) that are associated to the disease/trait susceptibility
  - If the variation of some SNPs is found to be higher in the case group than in the control group, these SNPs are reported as a potential marker of the disease/trait (biomarker)
- **Why are GWAS important for personalized medicine ?**
  - Combinations of SNPs can reflect biomarkers of diseases (e.g., cancer)
  - People who have DNA compatible with a biomarker have predisposition for developing the corresponding disease
  - Medicine can be supplied at an early stage to these people to prevent the development of the disease

EMR data to support research on personalized medicine

- **Genome-Wide Association Studies (GWAS)**
    - 1,200 human GWASs have examined over 200 diseases and traits and found almost 4,000 SNP associations*

GWAS-related diseases**	
Asthma	*Lung cancer*
ADHD	*Pancreatic cancer*
Bipolar I disorder	*Platelet phenotypes*
Bladder cancer	*Pre-term birth*
Breast cancer	*Prostate cancer*
Coronary disease	*Psoriasis*
Dental caries	*Renal cancer*
Diabetes mellitus type 1	*Schizophrenia*
Diabetes mellitus type 2	*Sickle-cell disease*

* Johnson et al. An open access database of genome-wide association results. BMC medical genetics, 2009.

** Manolio et al. A HapMap harvest of insights into the genetics of common disease. J Clinic. Inv., 2008.

Part 1 - Content

- **Part 1: *Medical data sharing and the need for privacy***
  - **Patient data: EMRs, sharing, and use in applications**
  - **Introduction to privacy-preserving data sharing**
- **Part 2: *Research challenges and solutions***
- **Part 3: *Open problems and research directions***

Introduction to privacy-preserving data sharing

- Need for privacy
- Privacy scenarios
- Threats in data sharing
- Privacy policies

Need for privacy

- **Why we need privacy in medical data sharing?**
- **If privacy is breached, there are consequences to patients**

Consequences to patients

- Emotional and economical embarrassment
  - 62% of individuals worry their EMRs will not remain confidential*
  - 35% expressed privacy concerns regarding the publishing of their data to dbGaP**
- Opt-out or provide fake data → difficulty to conduct statistically powered studies

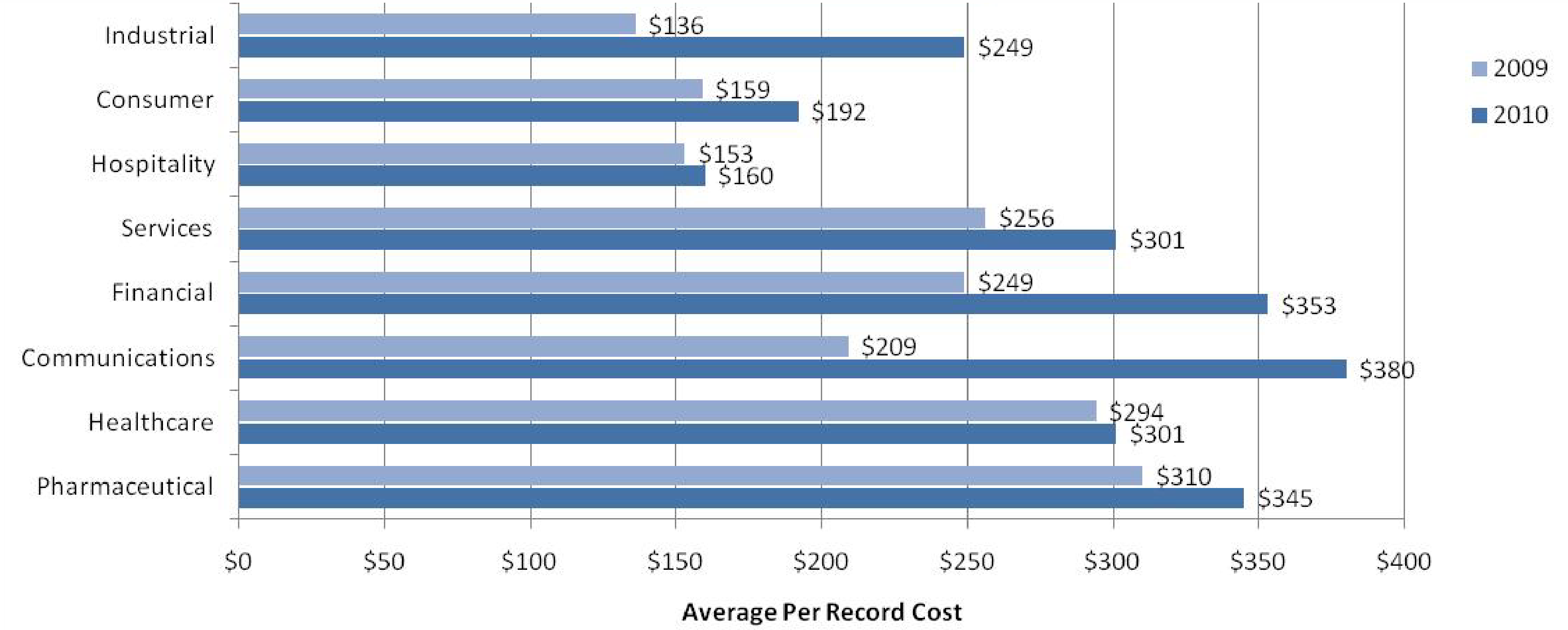
* Health Confidence Survey 2008, Employee Benefit Research Institute

** Ludman et al. Glad You Asked: Participants' Opinions of Re-Consent for dbGap Data Submission. Journal of Empirical Research on Human Research Ethics, 2010.

Need for privacy

- **If privacy is breached, there are consequences to organizations**
  - Legal → HIPAA, EU legislation (95/46/EC, 2002/58/EC, 2009/136/EC etc.)
  - Financial → It can cost an organization $7.2M on average*

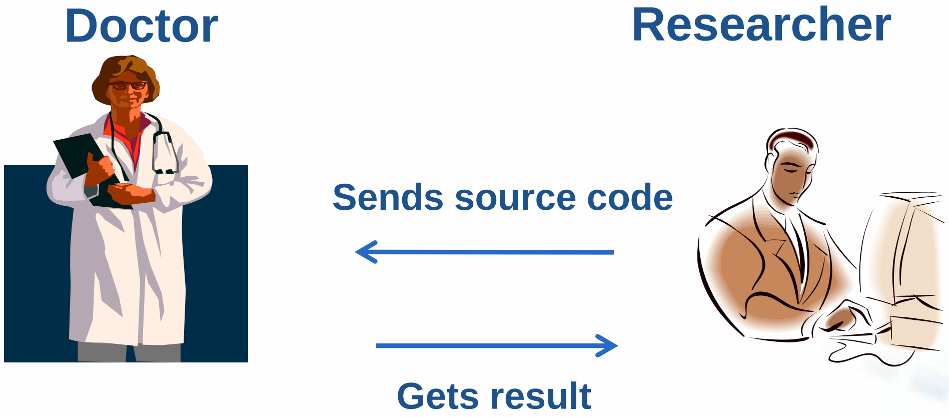
    and up to $35.3M



* Ponema Institute/Symantec corporation, 2010 Annual Study: US cost of a data breach.

Privacy-aware data sharing scenarios

- **“Send me your source code” scenario**



Pros:

- Attacker sees no data
- No infrastructure costs

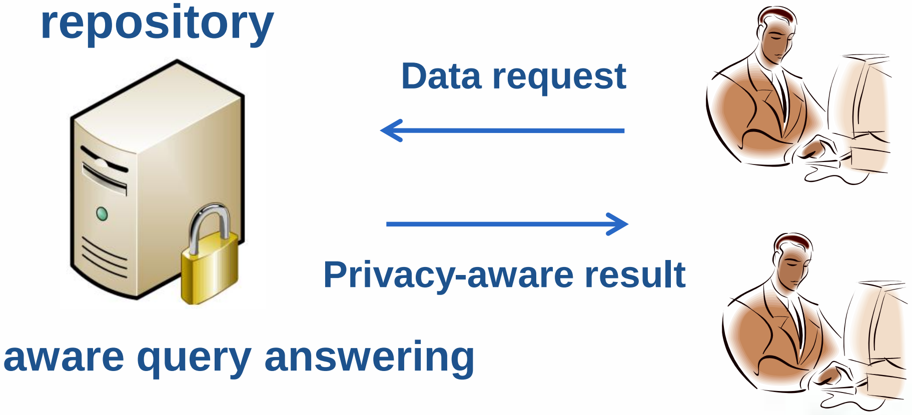
Cons:

- Only for hypothesis testing
- Result may breach privacy
- Code may be malicious
- Technical issues

Collaboration between researchers in CS & Medical Schools

Privacy-aware data sharing scenarios

- **Interactive scenario (akin to statistical databases)**



Protected data repository

Researchers

Data request

Privacy-aware result

Privacy aware query answering

Pros:

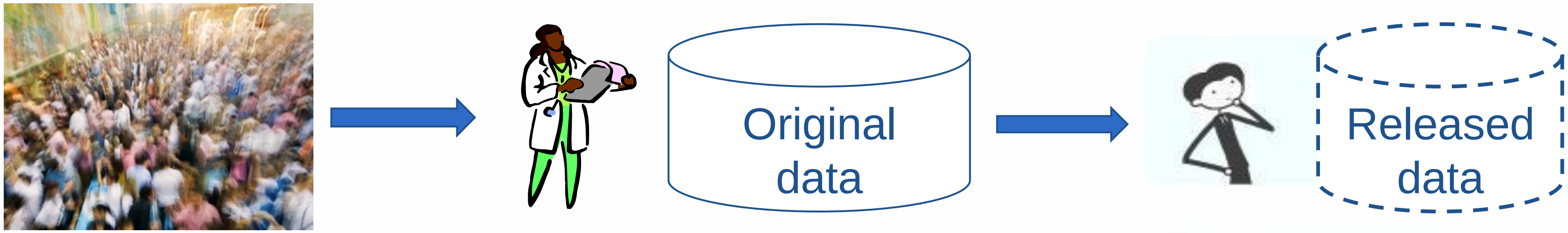
- Data kept in-house
- No need to specify utility requirements
- Strong privacy
- Attack identification and recovery from privacy breaches based on auditing

Cons:

- Difficulty to answer complex queries
- Data availability reduces with time
- Infrastructure costs
- Bad for hypothesis generation

Privacy-aware data sharing scenarios

- **Non-interactive scenario (a.k.a. *data publishing*)**



data owners **data publisher** *(trusted)* **data recipient** *(untrusted)*

Original data

Released data

Pros:

- Constant data availability
- No infrastructure costs
- Good for hypothesis generation and testing
- Seems to model most releases

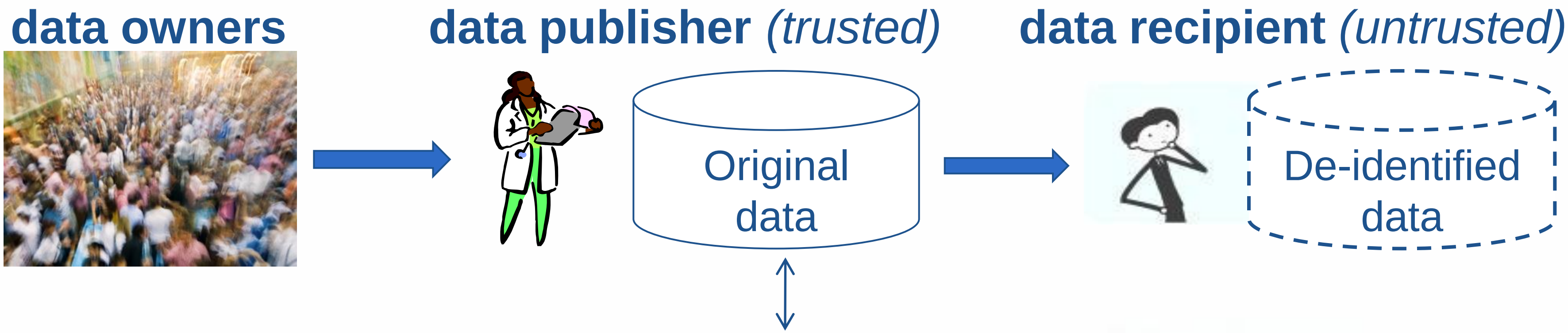
Cons:

- Privacy and utility requirements need to be specified
- Publisher has no control of the data
- No auditing

Hospitals release discharge summaries

Data publishing needs to preserve privacy

- **De-identification**



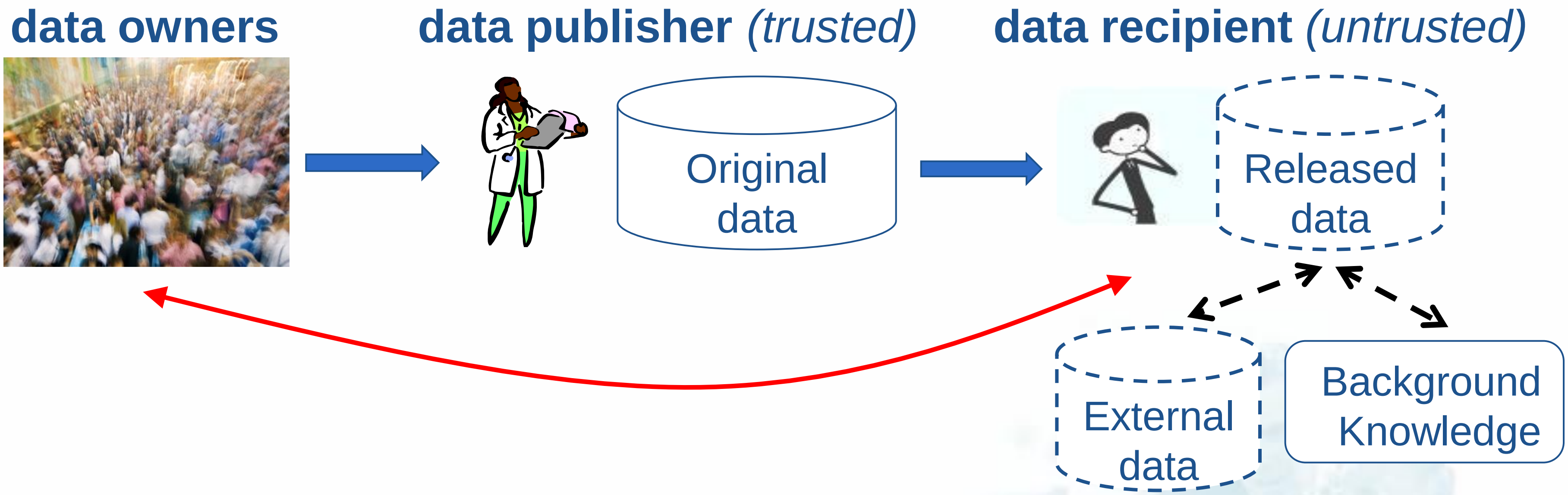
data owners
data publisher *(trusted)*
data recipient *(untrusted)*
Original data
De-identified data

  - Find out *identifiers* (attributes that uniquely identify an individual)
    - SSN, Patient ID, Phone number etc.
  - Remove them from the data prior to data publishing

Name	Search Query Terms
John Doe	Harry potter, King’s speech
Thelma Arnold	Hand tremors, bipolar,dry mouth, effect of nicotine on the body

Data publishing needs to preserve privacy

- **De-identification is not enough!**



data owners

data publisher *(trusted)*

Original data

data recipient *(untrusted)*

Released data

External data

Background Knowledge

- **Main types of threats to data privacy**
  - Identity disclosure
  - Sensitive information disclosure
  - Inferential disclosure

Privacy Threats: Identity Disclosure

- **Identity disclosure**
  - Individuals are linked to their published records based on quasi-identifiers (attributes that in combination can identify an individual)

Age	*Postcode*	*Sex*
20	NW10	M
45	NW15	M
22	NW30	M
50	NW25	F

De-identified data

⋈

Name	*Age*	*Postcode*	*Sex*
Greg	20	NW10	M
Jim	45	NW15	M
Jack	22	NW30	M
Anne	50	NW25	F

External data

Real-world identity disclosure cases – medical data

- Group Insurance Commission data → Voter list of Cambridge, MA

William Weld, Former Governor of MA

- Chicago Homicide database → Social security death index

  35% of murder victims

- Adverse Drug Reaction Database → Public obituaries

  26-year old girl who died from drug

Identity Disclosure in EMR data sharing

*De-identifying EMRs is not enough!

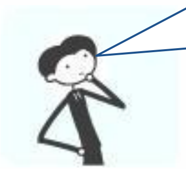
Identified EMR data

ID	ICD
Jim	333.4
Mary	401.0 401.1
Anne	401.0 401.2 401.3

↔

Released EMR Data

ICD	DNA
333.4	*CT…A*
401.0 401.1	*AC…T*
401.0 401.2 401.3	*GC…C*



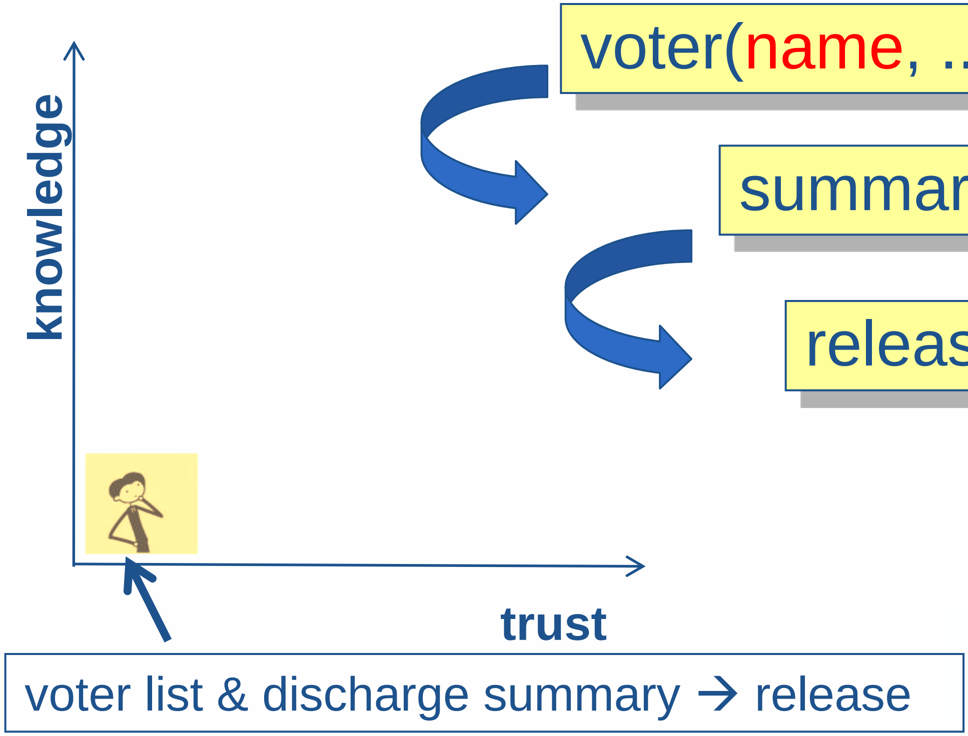
Mary is diagnosed with benign essential hypertension
(ICD code 401.1)
… the second record belongs to her → all her diagnosis codes

- **Disclosure based on diagnosis codes***
  - → general problem for other medical terminologies (e.g., ICD-10 used in EU)
  - → sharing data susceptible to the attack against legislation

* Loukides et al. The Disclosure of Diagnosis Codes Can Breach Research Participants' Privacy. JAMIA, 2010.

Identity Disclosure in EMR data sharing

- **Two-step attack using publicly available voter lists and hospital discharge summaries**



voter(name, ..., **zip, dob, sex**)

summary(**zip, dob, sex, diagnoses**)

release(**diagnoses**, DNA)

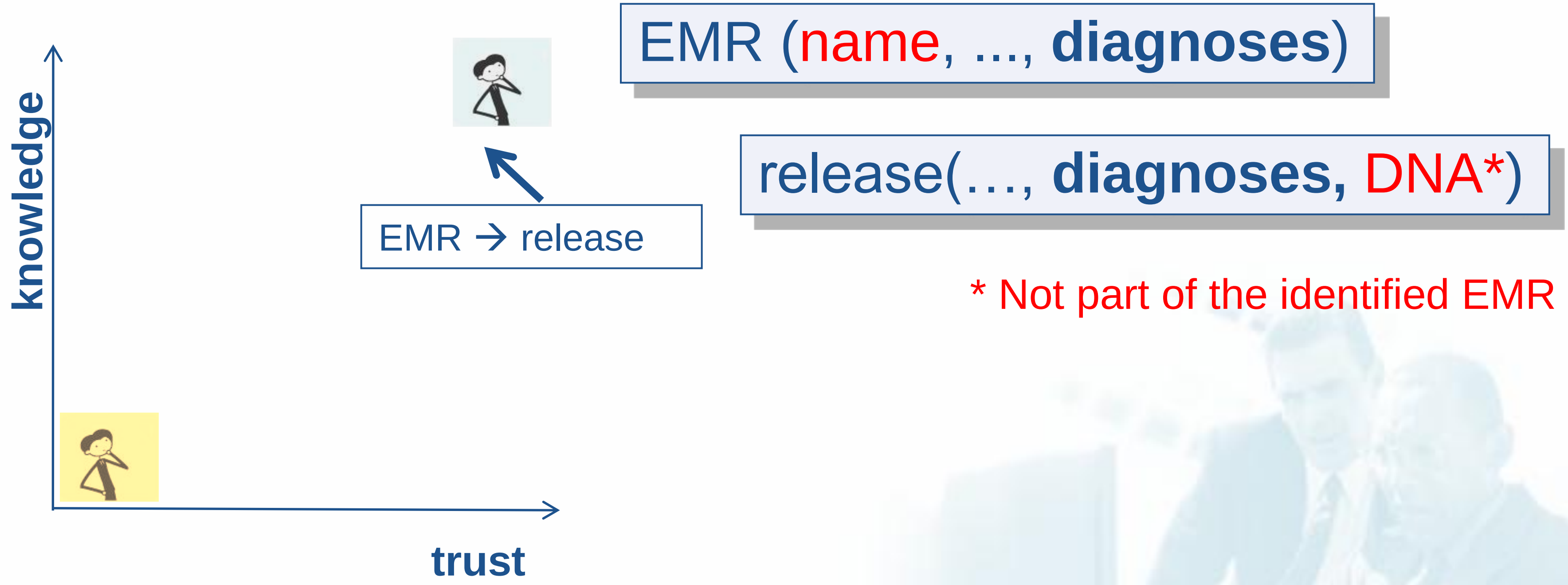
voter list & discharge summary → release

87% of US citizens can be identified by {dob, sex, ZIP-code}

* Sweeney, k-anonymity: a model for protecting privacy. IJUFKS, 2002.

Identity Disclosure in EMR data sharing

- **One-step attack using EMRs**



EMR → release

EMR (name, ..., **diagnoses**)

release(…, **diagnoses**, DNA*)

* Not part of the identified EMR

* Loukides et al. The Disclosure of Diagnosis Codes Can Breach Research Participants' Privacy. JAMIA, 2010.

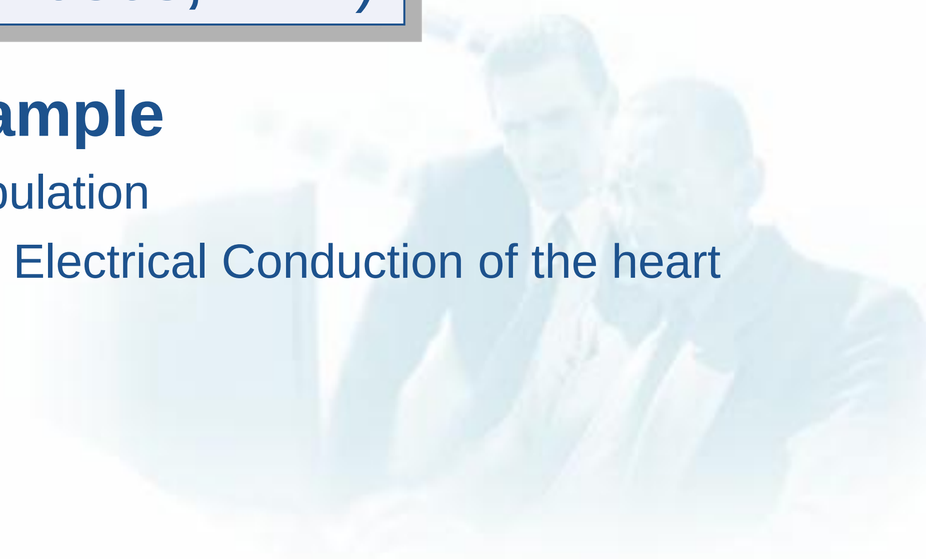
Identity Disclosure in EMR data sharing – Case Study

- **De-identified EMR population**
  - 1.2M records from Vanderbilt
  - a unique random number for ID

de-identified EMR (ID, ..., **diagnoses**)

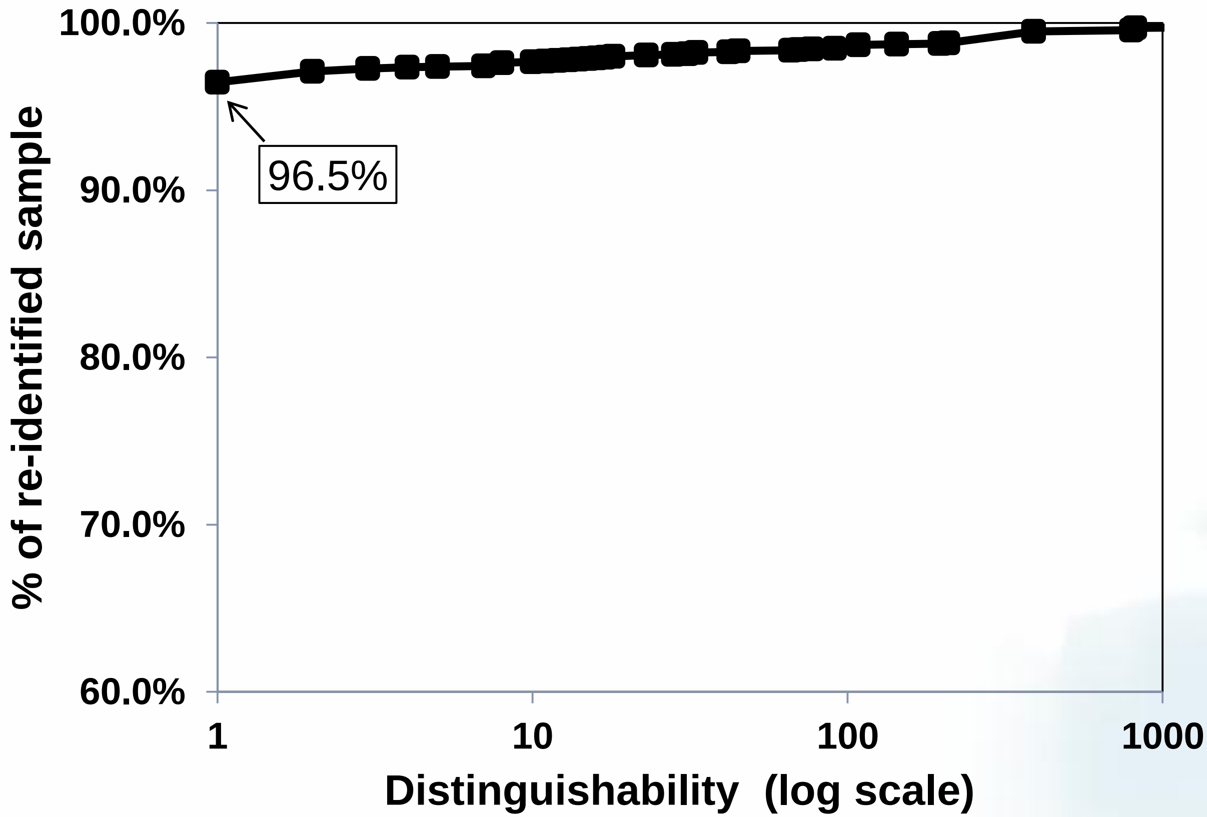
VNEC(…, **diagnoses**, DNA)

- **VNEC de-identified EMR sample**
  - 2762 records derived from the population
  - involved in a GWAS for the Native Electrical Conduction of the heart
  - will be deposited into dbGaP
  - useful for other GWAS



Identity Disclosure in EMR data sharing – Case Study

- **Vanderbilt's EMR - VNEC dataset linkage on ICD codes**



Number of times a set of ICD codes appears in the population
***Support* in the data mining literature**

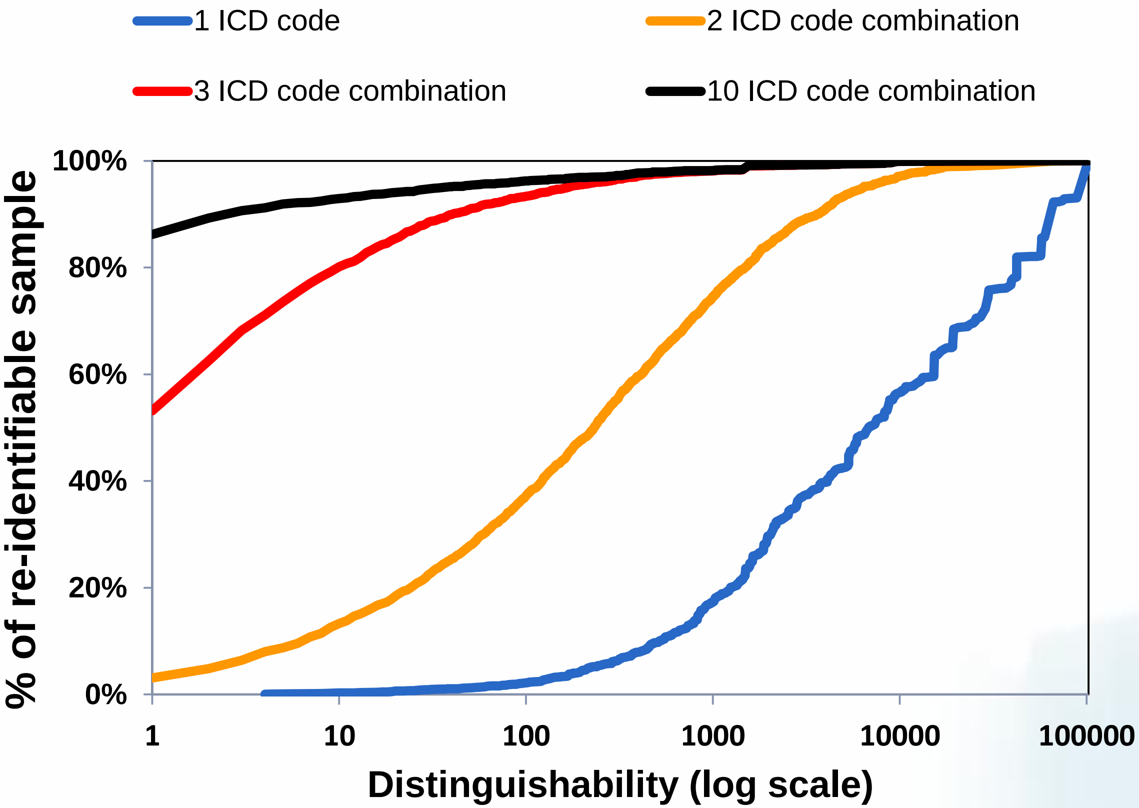
- We assume that all ICD codes are used to issue an attack

  (an "insider"'s attack)

- 96.5% of patients susceptible to identity disclosure

Identity Disclosure in EMR data sharing – Case Study

- **Vanderbilt's EMR - VNEC dataset linkage on ICD codes**

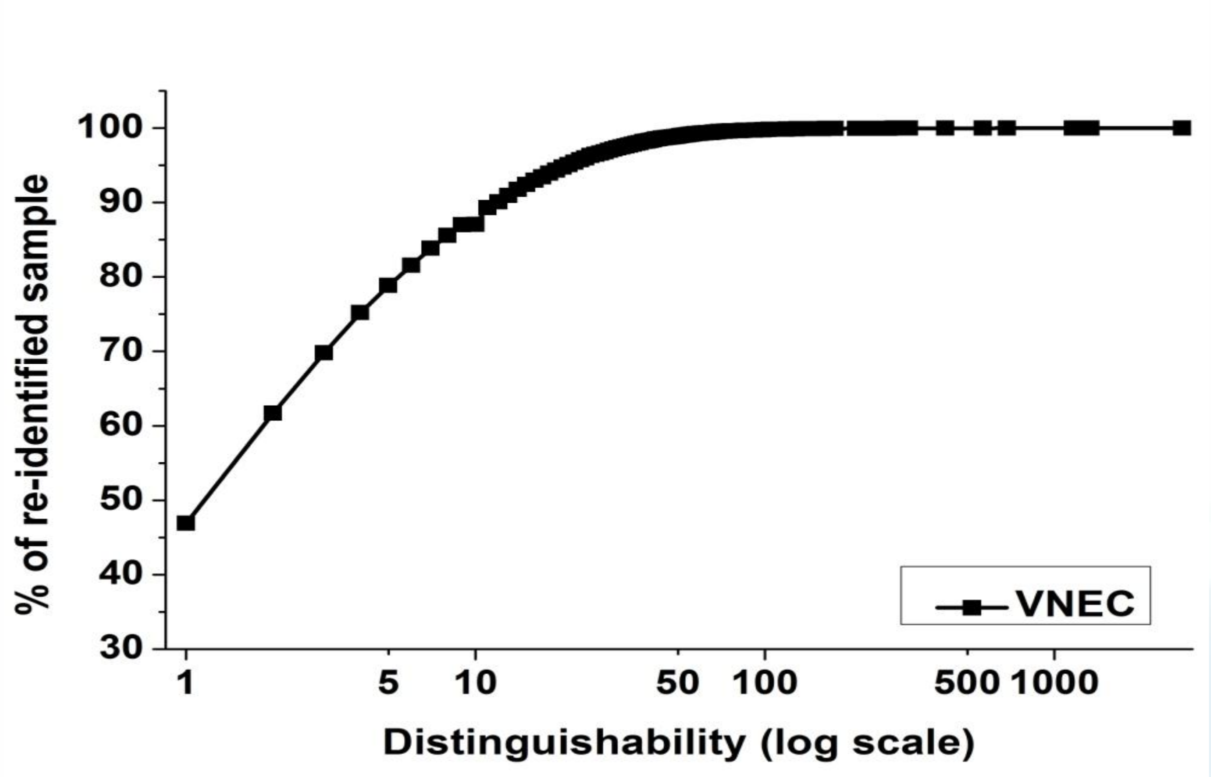


- A random subset of ICD codes that can be used in attack
- Knowing a random combination of 2 ICD codes can lead to unique re-identification

Number of times a set of ICD codes appears in the population
***Support* in data mining literature**

Identity Disclosure in EMR data sharing – Case Study

- **VNEC dataset linkage on ICD codes – Hospital discharge records**



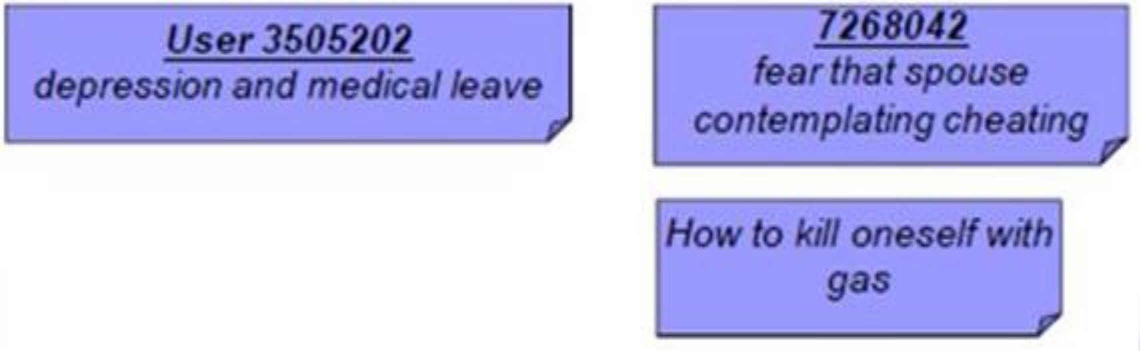
Number of times a set of ICD codes appears in the VNEC
Support in data mining literature

- All ICD codes for a single visit
- Difficult to know ICD codes that span visits when public discharge summaries are used
- 46% uniquely re-identifiable patients in VNEC

Privacy Threats: Sensitive information disclosure

- **Sensitive information disclosure**
  - Individuals are associated with *sensitive* information

Sensitive terms in AOL search logs



User 3505202
depression and medical leave

7268042
fear that spouse contemplating cheating

How to kill oneself with gas

* Narayanan et al. Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy '08.

Sensitive information disclosure – demographic data sharing

Sensitive Attribute (SA)

Age	*Postcode*	*Sex*	*Disease*
20	NW10	M	HIV
45	NW15	M	Cold
22	NW30	M	Cancer
50	NW25	F	Cancer

De-identified data

Name	*Age*	*Postcode*	*Sex*
Greg	20	NW10	M

External data

⋈

Age	*Postcode*	*Sex*	*Disease*
20	NW10	M	HIV
20	NW10	M	HIV
20	NW10	M	HIV
20	NW10	M	HIV

De-identified data

- Can occur without identity disclosure

Sensitive information disclosure – movie rate sharing

NETFLIX

- 100M dated ratings from 480K users to 18K movies
- data mining contest ($1M prize) to improve movie recommendation based on personal preferences
- movies reveal political, religious, and sexual beliefs and need protection according to Video Protection Act
- **“Anonymized”**
  - De-identification

A lawsuit was filed, Netflix settled the lawsuit

“We will find new ways to collaborate with researchers”

- **Researchers inferred movie rates of subscribers***
  - Data are linked with IMDB w.r.t. ratings and/or dates

* Narayanan et al. Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy ‘08.

Sensitive information disclosure – EMR data sharing

Identified EMR data

ID	ICD
Jim	401.0 401.1 295
Mary	401.0 401.1 303 295

Released EMR Data

ID	ICD	DNA
~~*Jim*~~	401.1 401.1 295	*C…A*
~~*Mary*~~	401.0 401.1 303 295	*A…T*

Schizophrenia

Mary is diagnosed with 401.0 and 401.1… she has Schizophrenia

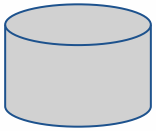
* Loukides et al. The Disclosure of Diagnosis Codes Can Breach Research Participants' Privacy. JAMIA , 2010.

Privacy-threats: Inferential disclosure

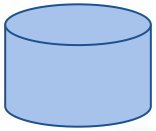
- **Sensitive knowledge patterns are inferred by data mining**[*,**]

75% of patients visit the same physician >4 times

Unsolicited advertisement



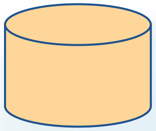
Stream data collected by health monitoring systems



Electronic medical records

60% of the white males >50 suffer from diabetes

Patient discrimination



Drug orders & costs

- Competitors can harm data publishers <u>and</u> insurance, pharmaceutical and marketing companies can harm data owners[*]

[*] Das et al. Privacy risks in health databases from aggregate disclosure. PETRA, 2009.

[**] Gkoulalas-Divanis et al. Revisiting sequential pattern hiding to enhance utility. KDD, 2011.

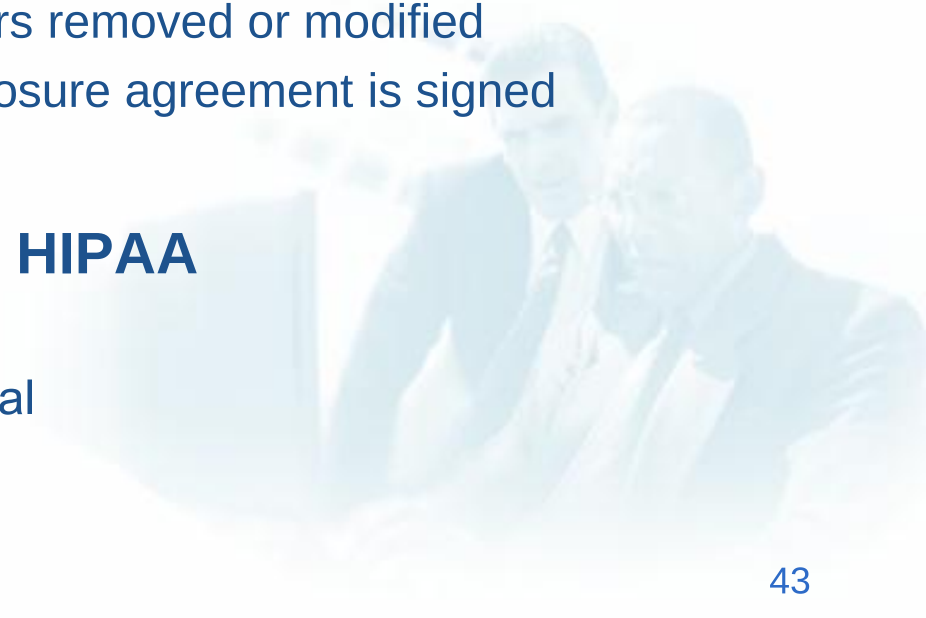
Privacy policies

- **Policies related to Protected Health Information (i.e., health information that may identify individuals) in the US**
  - Health Insurance Portability and Accountability Act (HIPAA), 1996
  - Health Information Technology for Economic and Clinical Health Act (HITECH), 2009
  - NIH GWAS policy, 2007
- **Similar policies world-wide**
  - EU Data Protection Directive 95/46/EC, UK Data Protection Act, etc.

HIPAA & HITECH

- **HIPAA specifies three routes for sharing data**
  - Expert determination – data are statistically verified to be de-identified by a person with appropriate knowledge
  - Safe Harbor – 17 identifiers (names, SSN etc.) are removed or modified
    – no knowledge that the remaining information can lead to identity disclosure
  - Limited Dataset – data are shared for research activities,
    – 16 identifiers removed or modified
    – a non disclosure agreement is signed

- **HITECH introduces changes to HIPAA**
  - Notification in case of privacy breach
  - Selling PHI requires patient's approval



NIH GWAS policy

- Applies to GWAS-related grants, contracts, intramural research projects submitted to the NIH on or after Jan. 25, 2008

- **NIH-funded investigators are expected to share de-identified GWAS data to dbGaP***
  - descriptive data (questionnaires, genotype – phenotype analysis)
  - patient-specific data (coded phenotypes, exposures, genotypes)

- **Not sharing is an exception**
  - should be justified
  - will be considered for funding on a case-by-case basis

* National Institutes of Health, Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies. 2007.

Content

- Part 1: *Medical data sharing and the need for privacy*

- **Part 2:** ***Research challenges and solutions***

- **Part 3:** ***Open problems and research directions***

Content

- Part 1: *Medical data sharing and the need for privacy*

- Part 2: *Research challenges and solutions*
  - **Identifying and modeling adversarial knowledge**
  - **Transforming data to guarantee privacy**
  - **Quantifying data utility**
  - **Privacy-preserving data publishing:**

    models, methods, case studies

- Part 3: *Open problems and research directions*

Identifying and modeling adversarial knowledge

- **Data adversary’s knowledge and data sources are unknown**
  - Assumptions based on general properties of data, availability of external datasets, or policies

{YOB, Gender, 3-digit Zip code} unique for 0.04% of US citizens

vs

{DOB, Gender, 5-digit Zip code} unique for 87% of US citizens*

* Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. IJUFKS. 2002.

Identifying and modeling adversarial knowledge

- **Data adversary's knowledge and data sources are unknown**
  - What if data publishers cannot make such assumptions?

    Automatic specification - based on the dataset to be published

    Mine the original data to find negative association rules*

    males do not have "ovarian cancer"

    female Japanese have low chance of heart attack

    Privacy is protected when these rules cannot be used to perform sensitive information disclosure

    No assumptions on adversarial background knowledge

    The line of work of differential privacy*,** we will examine later.

[1] Li et al. Injector: Mining Background Knowledge for Data Anonymization. ICDE, 2008.

[2] Li et al. Modeling and Integrating Background Knowledge in Data Anonymization. ICDE, 2009.

[3] Dwork, Differential Privacy, ICALP, 2006.

[4] Dwork, The Promise of Differential Privacy. A Tutorial on Algorithmic Techniques, FOCS, 2011

Identifying and modeling adversarial knowledge

- **We must preserve privacy and achieve data utility**

  … but utility and privacy can only be traded-off

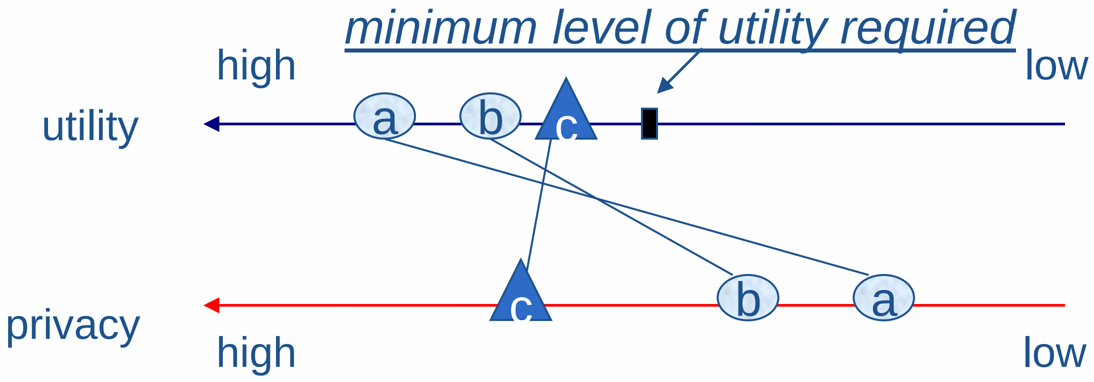
  - Max utility → Min privacy
  - Max privacy → Min utility

  - Models to capture privacy
  - Measures to capture utility

  - We will now focus on interesting solutions to trade-off privacy and utility

Identifying and modeling adversarial knowledge

- **Utility-bound approach**

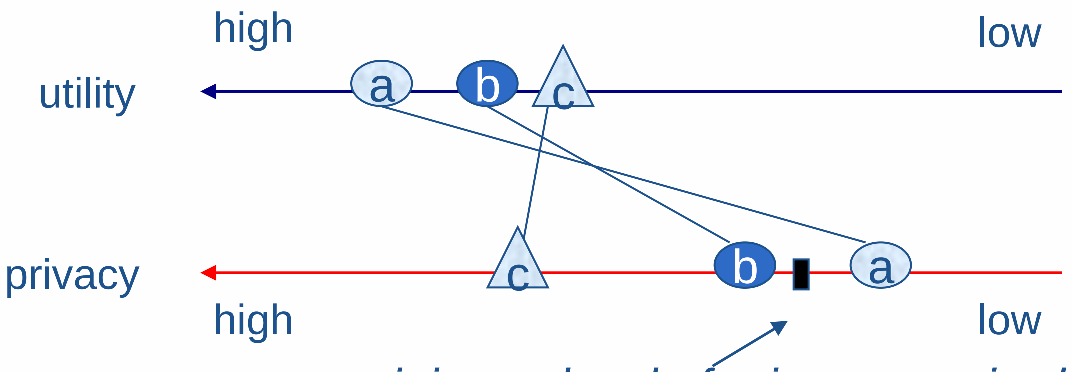


Best privacy the lowest tolerable level of utility

- Works well for some applications
  - classification accuracy in biomedical studies, LBS
- However, the minimum level of utility required may be difficult to be specified

Identifying and modeling adversarial knowledge

- **Privacy-bound approach**



minimum level of privacy required

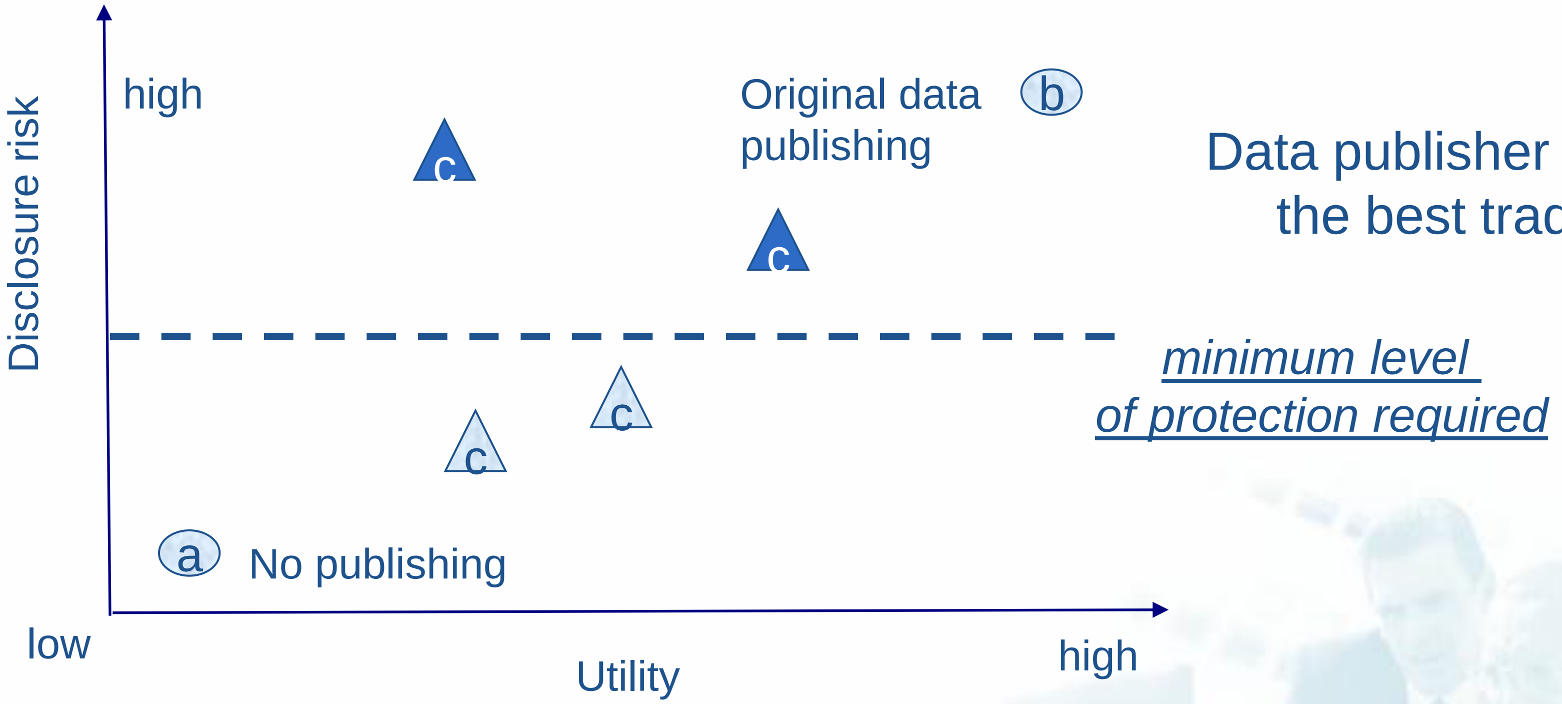
Best utility for a lower bound of privacy

- Adopted by the majority of works (e.g., k-anonymity, l-diversity)
- Utility quantification
  - with an optimization measure (e.g., level of information loss)
  - based on how well anonymized data supports a task compared to original data (e.g., workload of COUNT queries)*
- However, data publishers may still want to consider different solutions

* LeFevre et al. Workload-aware anonymization. KDD, 2006.

Identifying and modeling adversarial knowledge

- **R-U Confidentiality map to track the trade-off***

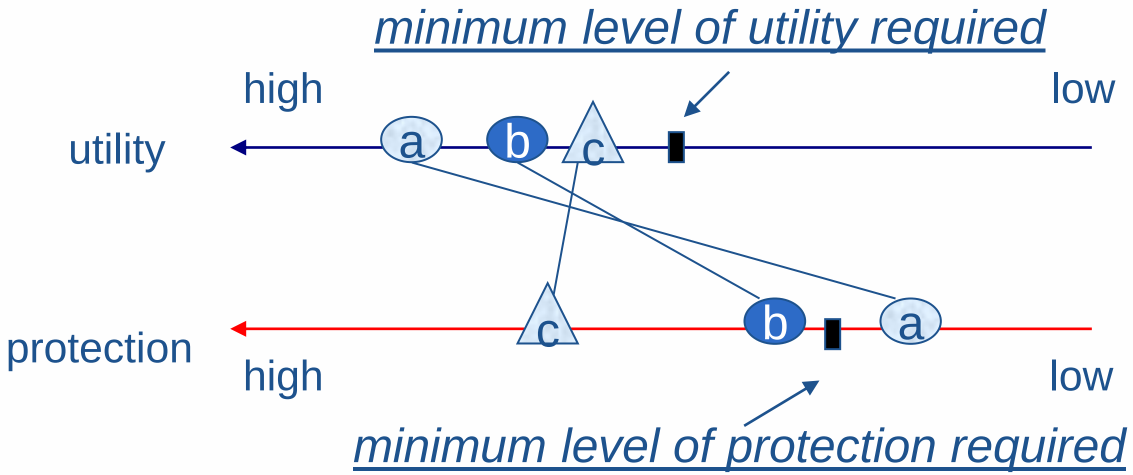


  - Allows comparing different anonymization techniques
  - Intuitive
  - Not easy to use it for comparing methods based on different privacy principles or more complex utility models

* Duncan et al. Disclosure Risk vs. Data Utility: The R-U Confidentiality map. Tech. Rep LA-UR-01-6428, Los Alamos National Library, 2001

Identifying and modeling adversarial knowledge

- **Utility-and-privacy constrained approach**



minimum level of utility required
high
low
utility
a
b
c
c
b
a
protection
high
low
minimum level of protection required

Lower bound for privacy
Minimum tolerable utility

- Constraints for utility and privacy
  - bound on information loss and privacy risk (on specific attributes or values)
- Guarantees privacy and utility
- Not always feasible (e.g., max privacy and max utility)
- Requires domain knowledge - reasonable in certain applications

Data transformation strategies

- **Synthetic data generation -** build a statistical model using a noise infused version of the data, and then synthetic data are generated by randomly sampling from this model

- **Masking methods**
  - **Perturbative –** aim to preserve privacy and aggregate statistics (e.g., means and correlation coefficients),
    - randomization, data swapping, microaggregation, rounding
    - falsify the data
  - **Non-perturbative** – aim to change the granularity of the reported data
    – do not falsify data

Non-perturbative methods – record suppression

- **Suppression of demographics**
  - **Record suppression –** all values in a record are deleted prior to data publishing

    - **Intuition:** An individual cannot be associated with a suppressed record or any of its values

Age	*Postcode*	*Sex*
20	**NW10**	**M**
20	**NW10**	**M**
45	**NW15**	**M**

De-identified data

Non-perturbative methods – record suppression

- **Suppression of demographics**
  - **Record suppression –** all values in a record are deleted prior to data publishing
    - **Intuition:** An individual cannot be associated with a suppressed record or any of its values

Name	*Age*	*Postcode*	*Sex*
Greg	20	NW10	M
Jim	45	NW15	M

External data

Age	*Postcode*	*Sex*
20	NW10	M
20	NW10	M

???

Suppressed data

- **Protects from both identity and sensitive information disclosure, but results in excessive information loss**

Non-perturbative methods – value suppression

- **Suppression of demographics**
  - **Value suppression –** certain values in quasi-identifiers are deleted (replaced by *) prior to data publishing

Intuition: An individual cannot be associated with a record based on a suppressed value

Name	*Age*	*Postcode*	*Sex*
Greg	20	NW10	M
Jim	45	NW15	M

External data

→

Age	*Postcode*	*Sex*	*Disease*
20	NW10	M	HIV
46	NW10	M	Flu

De-identified data

Non-perturbative methods – value suppression

- **Suppression of demographics**
  - **Value suppression –** certain values in quasi-identifiers are deleted (replaced by *) prior to data publishing

Intuition: An individual cannot be associated with a record based on a suppressed value

Name	*Age*	*Postcode*	*Sex*
Greg	20	NW10	M
Jim	45	NW10	M

External data

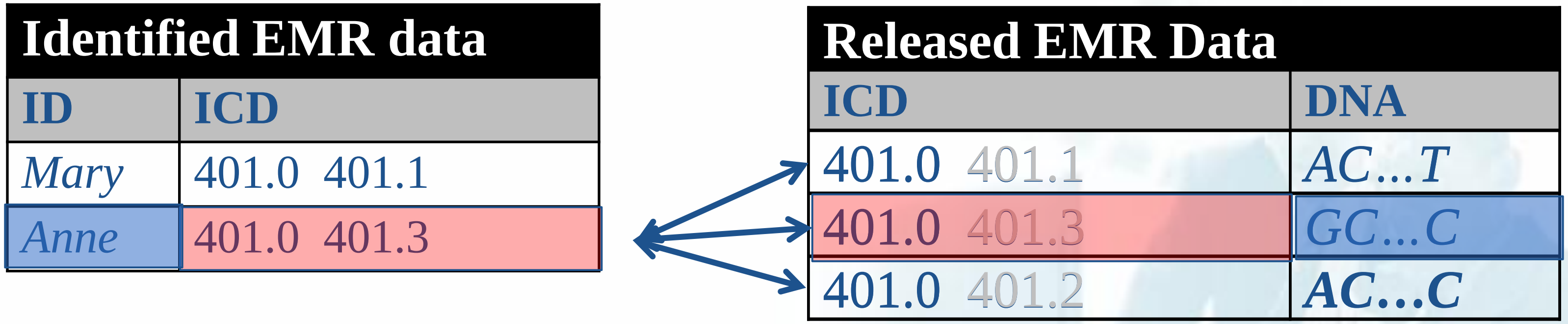
Age	*Postcode*	*Sex*	*Disease*
*	NW10	M	HIV
*	NW10	M	Flu

Suppressed data

- **Incurs less information loss than record suppression**
- ***… but identifying which values to suppress can be challenging***

Non-perturbative methods – code suppression

- **Suppression of ICD codes**
  - **Global** – removes an ICD code from all records
    - preserves the count of non-suppressed codes, which is beneficial in data mining applications

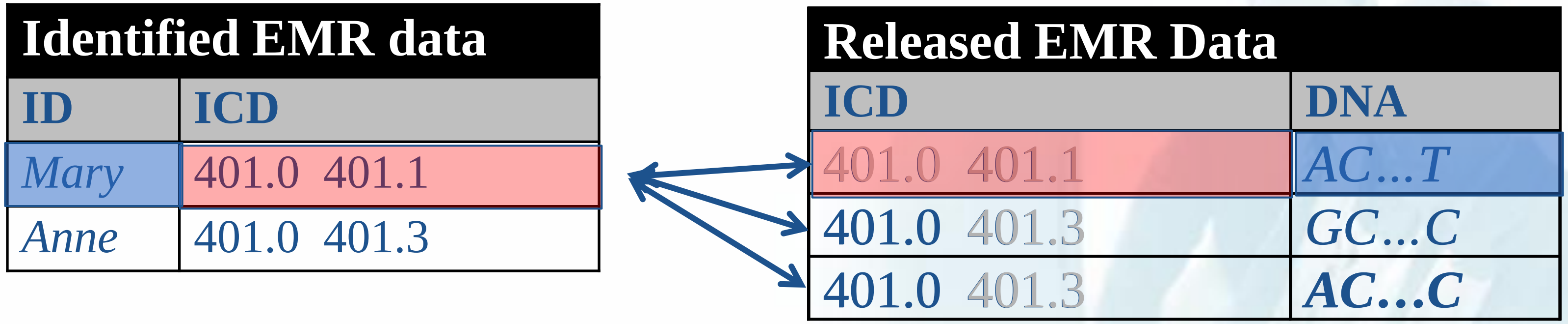


Identified EMR data	
ID	**ICD**
Mary	401.0 401.1
Anne	401.0 401.3

Released EMR Data	
ICD	**DNA**
401.0 401.1	*AC…T*
401.0 401.3	*GC…C*
401.0 401.2	***AC…C***

Non-perturbative methods – code suppression

- **Suppression of ICD codes**
  - **Local** – removes an ICD code from a number of records
    - preserves data utility better than global suppression

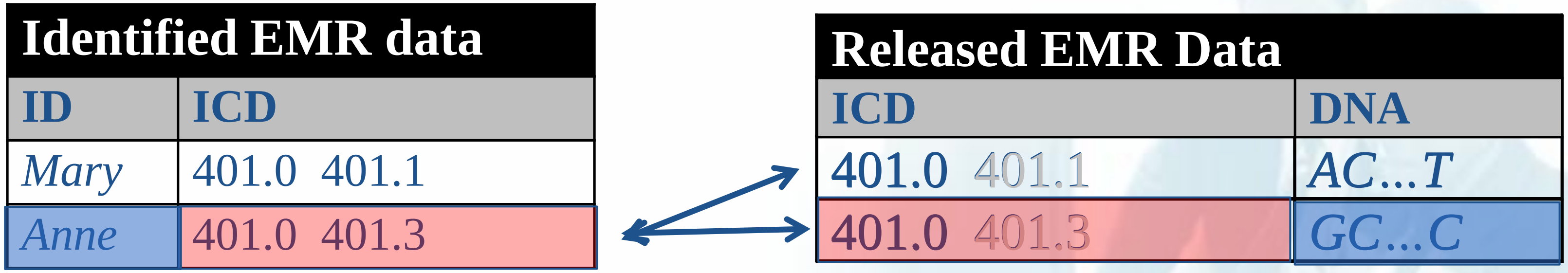


Identified EMR data	
ID	**ICD**
Mary	401.0 401.1
Anne	401.0 401.3

Released EMR Data	
ICD	**DNA**
401.0 401.1	*AC…T*
401.0 401.3	*GC…C*
401.0 401.3	***AC…C***

Non-perturbative methods – code suppression

- **We applied Vinterbo's method of suppression for ICD codes***
  - **Global** – removes an ICD code from all records
    - X% of least frequent ICD codes*
    - **Intuition:** they distinguish transactions from one another



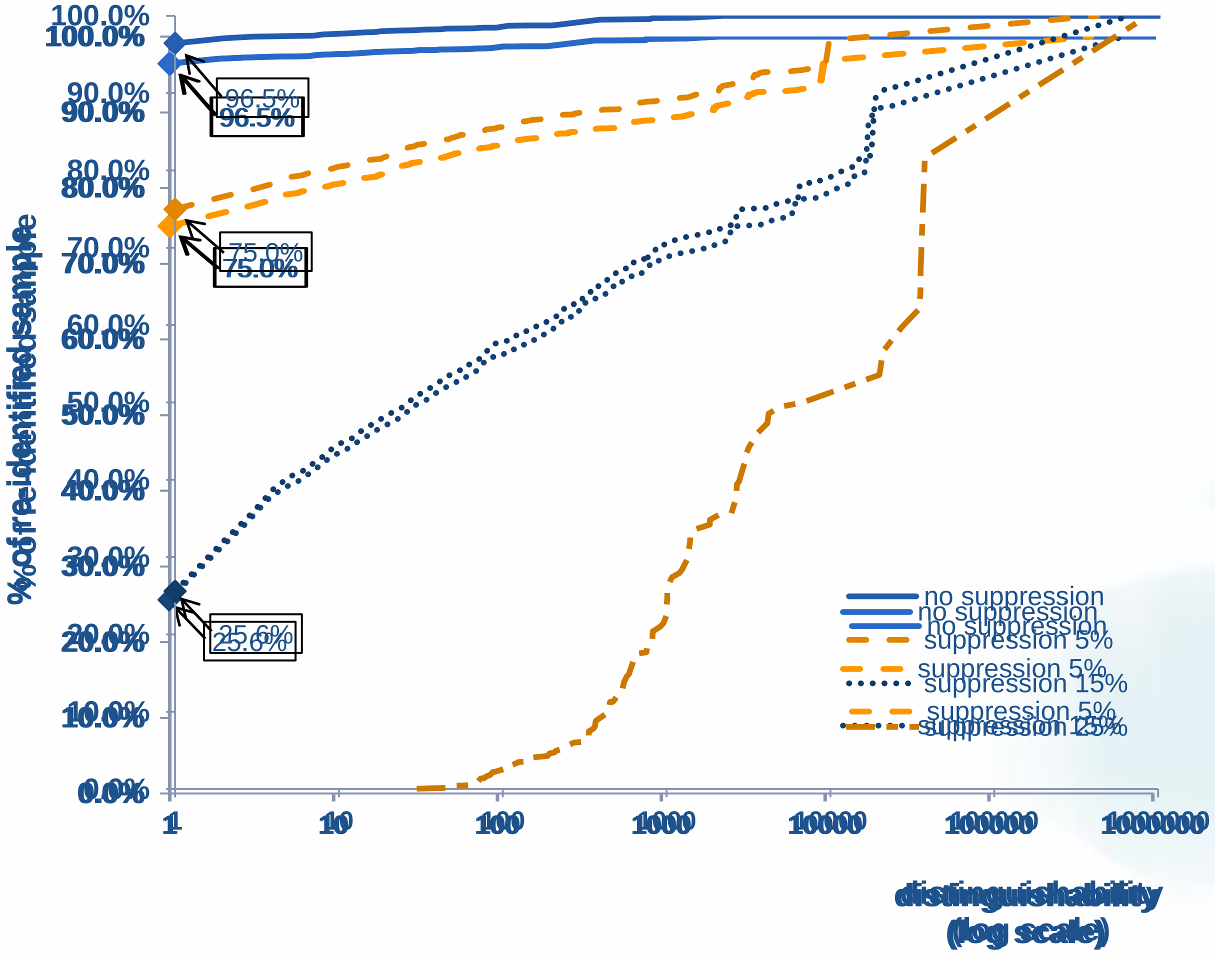
Identified EMR data	
ID	**ICD**
Mary	401.0 401.1
Anne	401.0 401.3

Released EMR Data	
ICD	**DNA**
401.0 401.1	*AC…T*
401.0 401.3	*GC…C*

* Vinterbo et al. Hiding information by cell suppression. AMIA Annual Symposium '01

Code suppression – a case study using Vanderbilt's EMR data

- **Vinterbo's method on VNEC –** suppress X% of least frequent codes



- Suppression of codes that appear in ≤ 25% of records to prevent re-identification

Code suppression – a case study using Vanderbilt’s EMR data

- **What can be safely released when privacy is achieved? – 5 out of ~6K ICD codes are released**

5-Digit ICD-9 Codes	3-Digit ICD-9 Codes	ICD-9 Sections
401.1- Benign essential hypertension →	*401-Essential hypertension* →	*Hypertensive disease*
780.79 - Other malaise and fatigue →	*780- Other soft tissue* →	*Rheumatism excluding the back*
729.5 - Pain in limb →	*729 - Other disorders of soft tissues* →	*Rheumatism excluding the back*
789.0 - Abdominal pain →	*789 – Other abdomen/pelvis symptoms* →	*Symptoms*
786.5 - Chest pain →	*786 -Respiratory system* →	*Symptoms*

Non-perturbative methods – data generalization

- **Generalization of demographics**
  - Values in quasi-identifiers are replaced by more general ones
    - **Intuition:** Fewer distinct values → data linkage becomes more difficult

Name	Age	Postcode	Sex
Greg	20	NW10	M
Jim	45	NW15	M

External data

Age	Postcode	Sex
20	NW10	M
45	NW15	M

De-identified data

Non-perturbative methods – data generalization

- **Generalization of demographics**
  - Values in quasi-identifiers are replaced by more general ones
    - **Intuition:** Fewer distinct values → data linkage becomes more difficult

Name	Age	Postcode	Sex
Greg	20	NW10	M
Jim	45	NW15	M

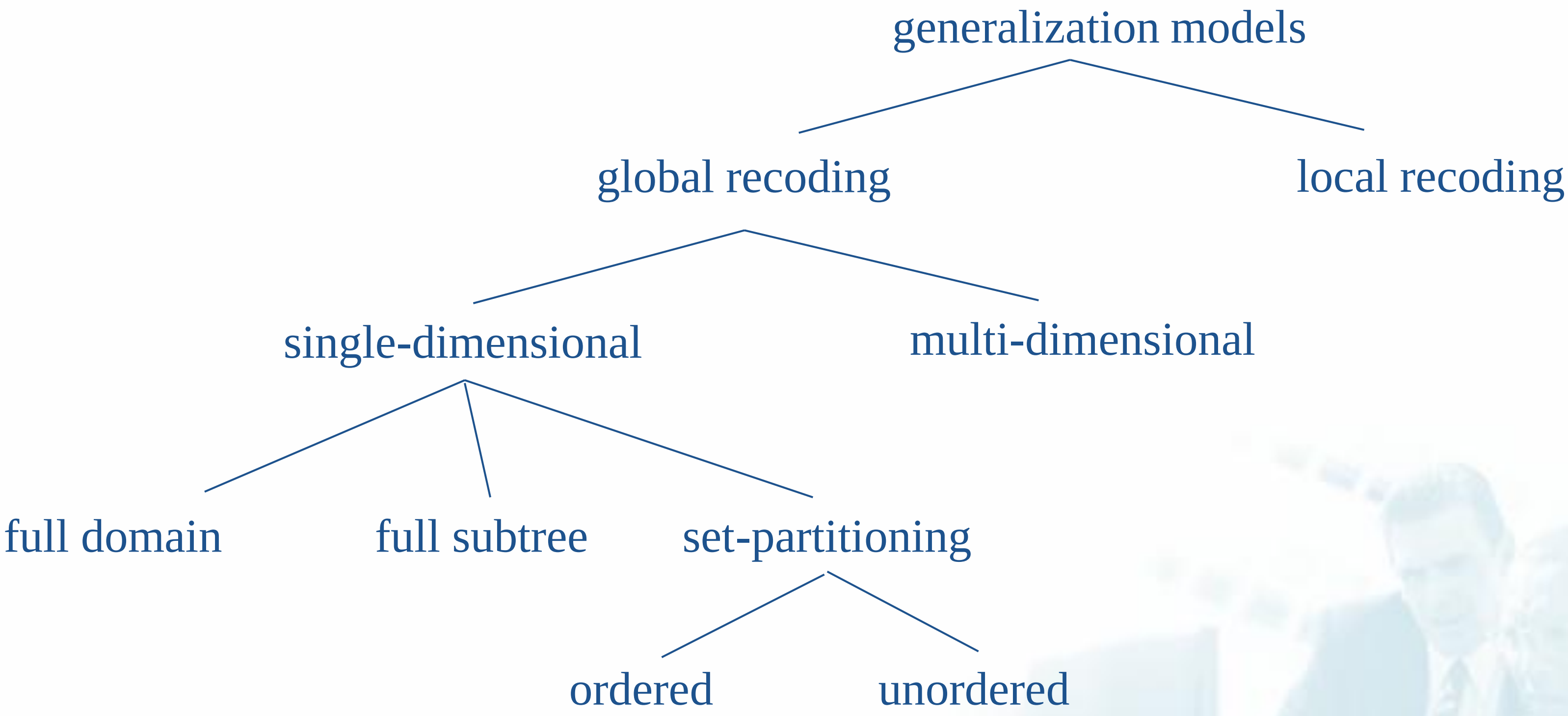
External data

Age	Postcode	Sex
[20-45]	NW1*	M
[20-45]	NW1*	M

Generalized data

- **Typically, it incurs less information loss than suppression**
- **However, identifying which values to generalize and how can be challenging**

Data generalization models for demographics



generalization models

global recoding

local recoding

single-dimensional

multi-dimensional

full domain

full subtree

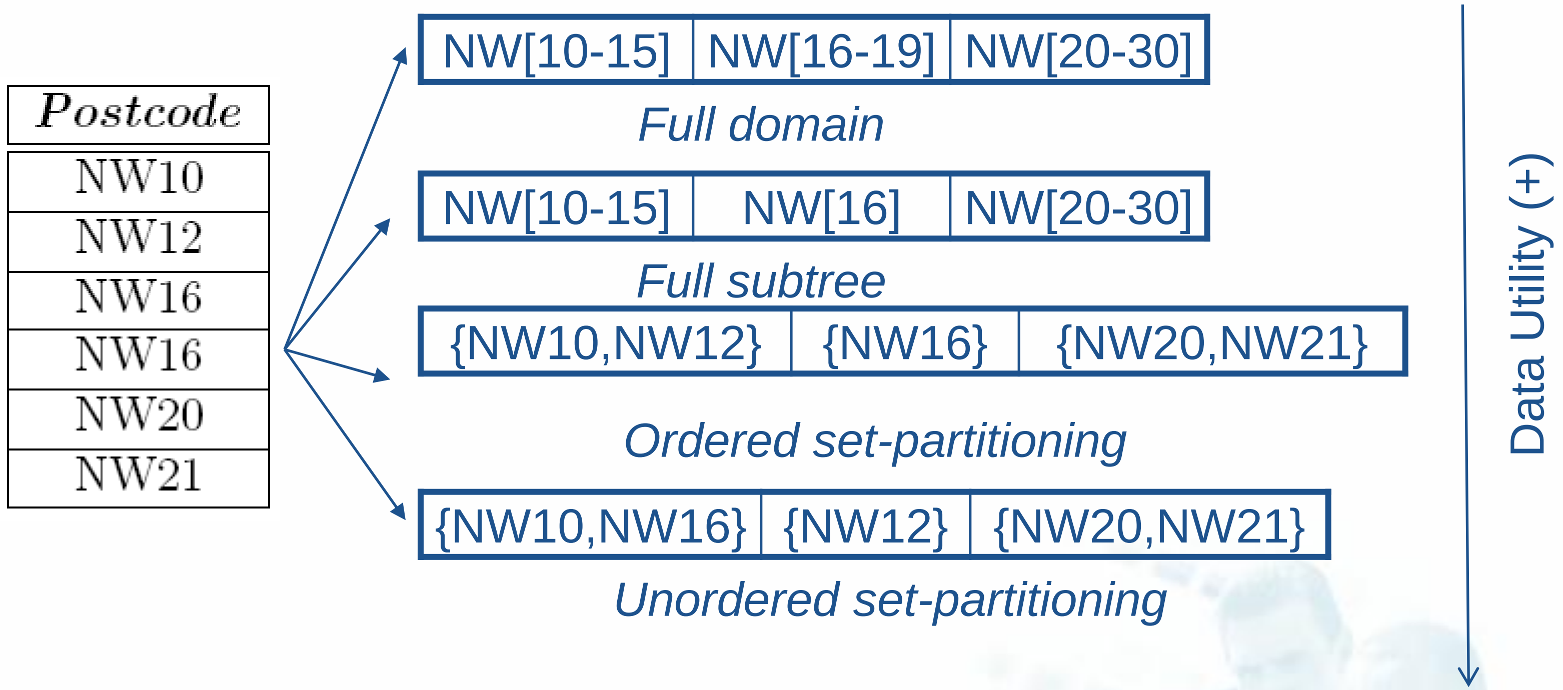
set-partitioning

ordered

unordered

- **Global** – a value is replaced by the same generalized value in all records

Global data generalization example



Postcode
NW10
NW12
NW16
NW16
NW20
NW21

NW[10-15]	NW[16-19]	NW[20-30]

Full domain

NW[10-15]	NW[16]	NW[20-30]

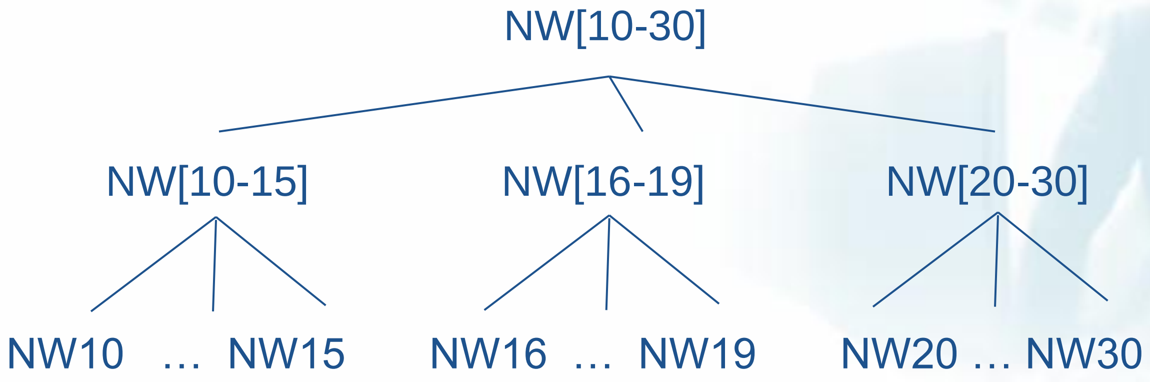
Full subtree

{NW10,NW12}	{NW16}	{NW20,NW21}

Ordered set-partitioning

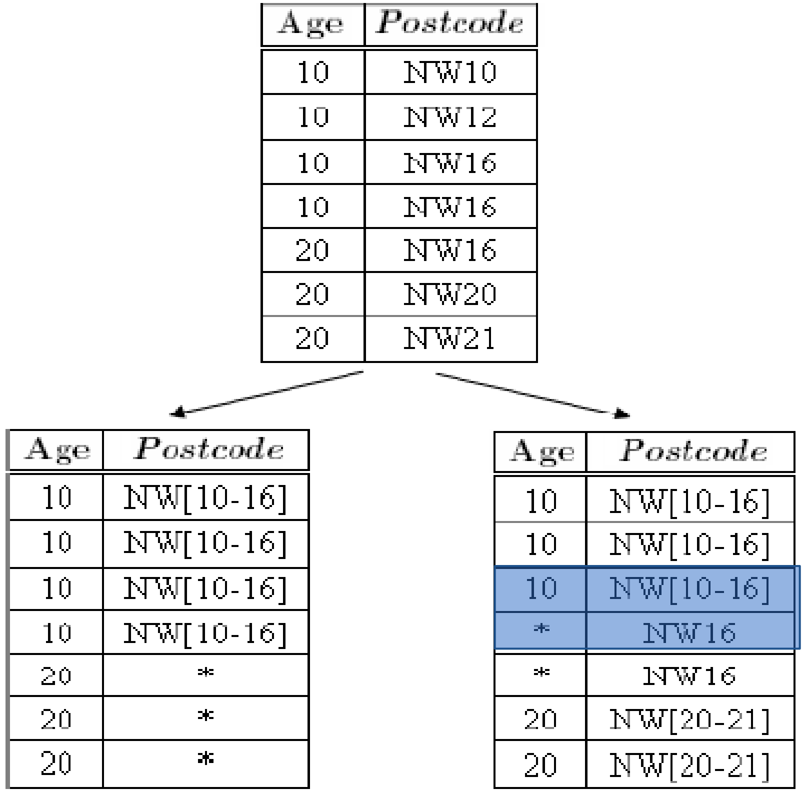
{NW10,NW16}	{NW12}	{NW20,NW21}

Unordered set-partitioning



Data generalization models – local generalization

- **Generalization of demographics**
  - **Local recoding** – a value can be replaced by multiple generalized values



Age	Postcode
10	NW10
10	NW12
10	NW16
10	NW16
20	NW16
20	NW20
20	NW21

Age	Postcode
10	NW[10-16]
10	NW[10-16]
10	NW[10-16]
10	NW[10-16]
20	*
20	*
20	*

Multi-dimensional global recoding

Age	Postcode
10	NW[10-16]
10	NW[10-16]
10	NW[10-16]
*	NW16
*	NW16
20	NW[20-21]
20	NW[20-21]

Local recoding

Pros: Allows exploring a larger number of generalizations than global recoding → less information loss

Cons: Anonymized data are difficult to be interpreted and/or mined (e.g., difficult to be used to train a classifier)

Data generalization models for ICD codes

- **Generalization of *ICD codes***
  - **Global** – an ICD code is replaced by a generalized code in all the records

401.1 - benign essential hypertension → 401- essential hypertension

Identified EMR data	
ID	**ICD**
Mary	401.0 401.1
Anne	401.0 401.3

Released EMR Data	
ICD	**DNA**
401.0 401.1	***AC…T***
401.0 401.3	*GC…C*

Data generalization models for ICD codes

- **Generalization of *ICD codes***
  - **Local** – an ICD code can be replaced by more than one *generalized codes in different records*

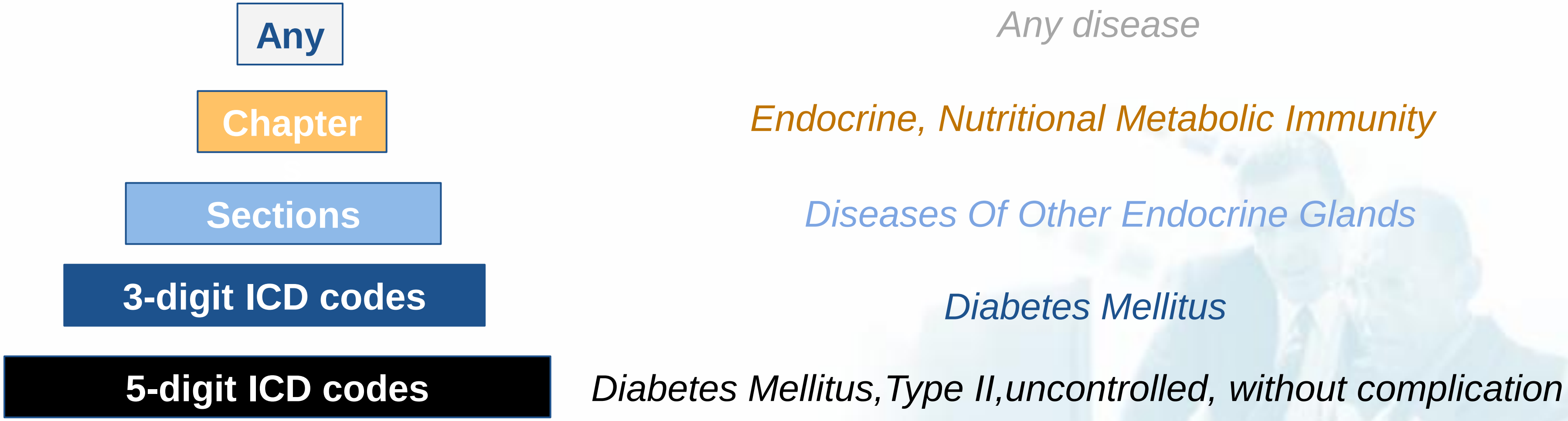
401.1 - benign essential hypertension → 401- essential hypertension → Any

Identified EMR data	
ID	**ICD**
Mary	401.0 401.1
Anne	401.0 401.3

Released EMR Data	
ICD	**DNA**
401.0 401.1 Any	*AC…T*
401.0 401.3 401	*GC…C*

Data generalization models for ICD codes – hierarchy-based model

- **Generalization of *ICD codes****
  - Hierarchy-based global generalization model



Any	*Any disease*
Chapter	*Endocrine, Nutritional Metabolic Immunity*
Sections	*Diseases Of Other Endocrine Glands*
3-digit ICD codes	*Diabetes Mellitus*
5-digit ICD codes	*Diabetes Mellitus,Type II,uncontrolled, without complication*

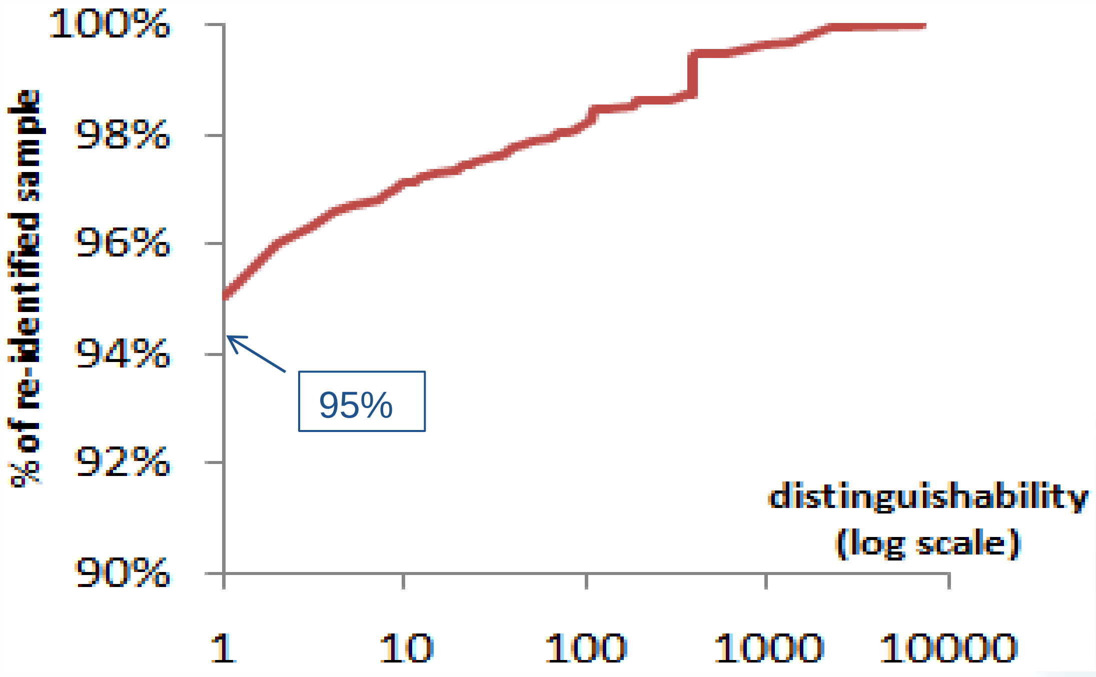
* Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. IJUFKS. 2002.

Code generalization – a case study using Vanderbilt’s EMR data

Generalizing ICD codes from VNEC*

5-digit ICD codes → **3-digit ICD codes**

coarsest allowable generalization for GWAS

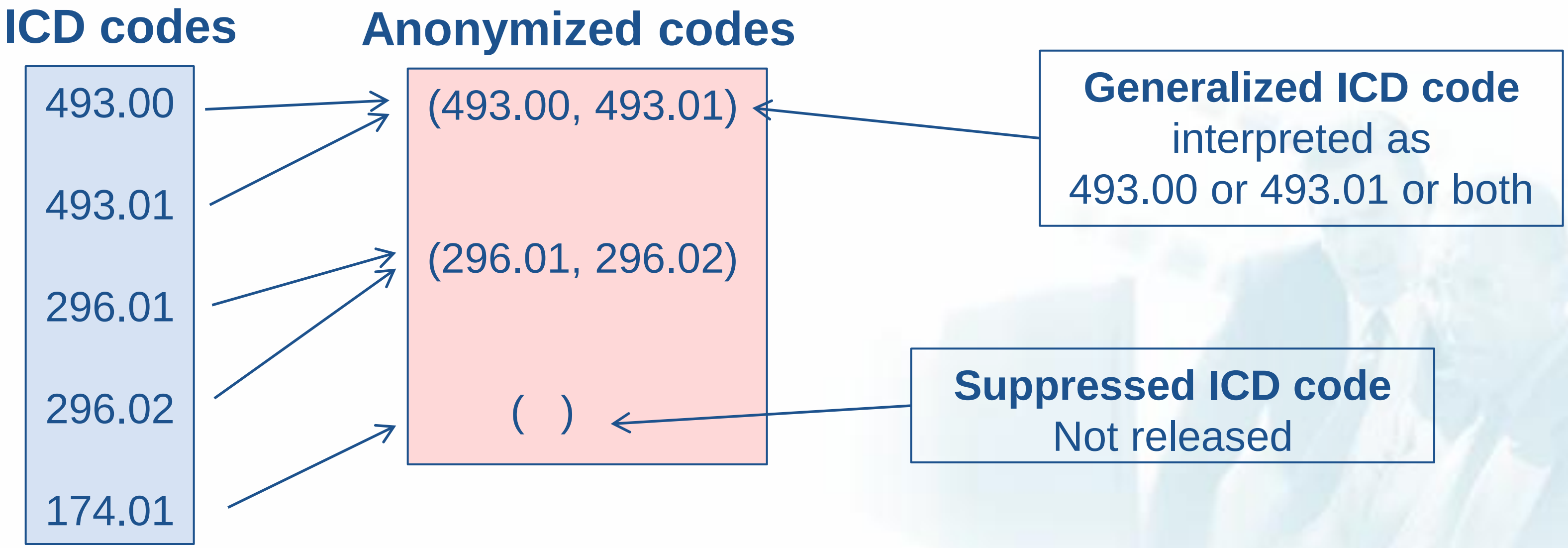


- 95% of the patients remain re-identifiable
- Combining generalization and suppression does not help privacy

* Loukides et al. The Disclosure of Diagnosis Codes Can Breach Research Participants’ Privacy. JAMIA, 2010.

Set-based anonymization model for ICD codes

- **Set-based anonymization***
  - Global model
  - Models both generalization and suppression
  - Each original ICD code is replaced by a unique set of ICD codes – *no need for generalization hierarchies*



ICD codes

493.00

493.01

296.01

296.02

174.01

Anonymized codes

(493.00, 493.01)

(296.01, 296.02)

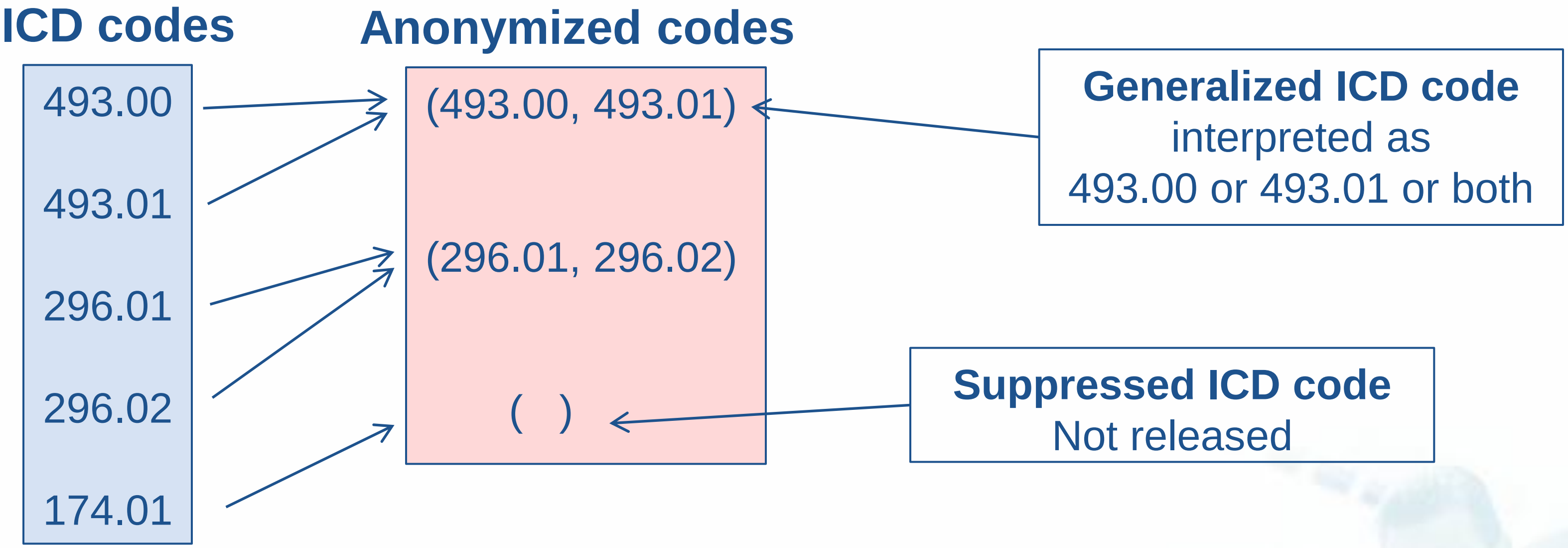
( )

Generalized ICD code
interpreted as
493.00 or 493.01 or both

Suppressed ICD code
Not released

*Loukides et al. Anonymization of Electronic Medical Records for Validating Genome- Wide Association Studies. PNAS, 2010.

Set-based anonymization model for ICD codes



ICD codes

493.00

493.01

296.01

296.02

174.01

Anonymized codes

(493.00, 493.01)

(296.01, 296.02)

( )

Generalized ICD code
interpreted as
493.00 or 493.01 or both

Suppressed ICD code
Not released

EMR Data

ICD	DNA
493.00 296.01 296.02	*CT…A*
493.00 493.01	*AC…T*
296.01	*GC…C*

Anonymized EMR Data

ICD	DNA
(493.00, 493.01) (296.01, 296.02)	*CT…A*
(493.00, 493.01)	*AC…T*
(296.01, 296.02)	*GC…C*

Quantifying data utility

- **Suppression and generalization reduce data utility**
- **Capture data utility by measuring information loss**
  - Assumes that we do not know the applications data will be used for
  - **Generalized group** – all records with the same values in all QIDs

Age	*Postcode*	*Disease*
[20-30]	CF[0-10]	HIV
[20-30]	CF[0-10]	Cold
[30-40]	CF[26-75]	Cancer
[30-40]	CF[26-75]	Cold

Generalized group g_1

Generalized group g_2

- **Capture data utility by measuring the accuracy of performing a specific task using anonymized data**
  - Reasonable for data shared between researchers

Quantifying data utility for demographics based on information loss

- **Group size-based measures**

  → **large groups more Information Loss**

  - Discernability Measure (**DM**)

$$DM = \sum_{j=1}^{r}(|g_j|^2) + \sum_{j=r+1}^{h}(|T| \times |g_j|)$$

Penalty for a generalized group g_j

Penalty for a suppressed group g_j

(removed records)

Age	*Postcode*	*Disease*
[20-30]	CF[0-10]	HIV
[20-30]	CF[0-10]	Cold
[30-40]	CF[26-75]	Cancer
[30-40]	CF[26-75]	Cold
[50-60]	CF[0-45]	HIV
[50-60]	CF[0-45]	Cancer
[60-90]	CF[50-95]	Cold
[60-90]	CF[50-95]	Cough
[60-90]	CF[50-95]	HIV

  - Normalized Average Equivalence Class Size Metric (**C$_{AVG}$**)

$$C_{AVG} = \frac{|T|}{h \times k}$$

Size of anonymized dataset

groups

# records in smallest generalized group

Quantifying data utility for demographics based on information loss

- **Range-based measures**

  → **large ranges more Information loss**

  Same DM scores

  - Normalized Certainty Penalty (**NCP**)

$$NCP = \sum_{j=1}^{h} \left(|g_j| \sum_{i=1}^{m} \frac{r(\pi_{a_i}(g_j))}{r(\pi_{a_i}(T))} \right)$$

# records in generalized group g_j

domain size of the QID a_i

range of the projection of g_j over the QID a_i

Age	*Postcode*	*Disease*
[20-30]	CF[0-10]	HIV
[20-30]	CF[0-10]	Cold
[30-40]	CF[26-75]	Cancer
[30-40]	CF[26-75]	Cold
[50-60]	CF[0-45]	HIV
[50-60]	CF[0-45]	Cancer
[60-90]	CF[50-95]	Cold
[60-90]	CF[50-95]	Cough
[60-90]	CF[50-95]	HIV

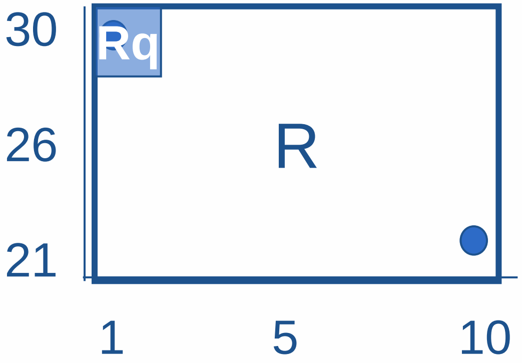
  - Loss Metric **(LM)**
  - Utility Measure (**UM**)

Quantifying data utility for demographics based on analytic tasks

- Average Relative Error (**AvgRE**)

COUNT(*) from T where Age=30 and Postcode is CF1

Age	*Postcode*	*Disease*
[21-30]	CF[1-10]	HIV
[21-30]	CF[1-10]	Cold
[30-40]	CF[26-75]	Cancer
[30-40]	CF[26-75]	Cold
[50-60]	CF[0-45]	HIV
[50-60]	CF[0-45]	HIV
[60-90]	CF[50-95]	Cold
[60-90]	CF[50-95]	Cough
[60-90]	CF[50-95]	HIV



$$Rq = \frac{1}{10} \times \frac{1}{10} = 0.01$$

$$est(q) = |g| \times \frac{R \cap Rq}{R} = 0.02$$

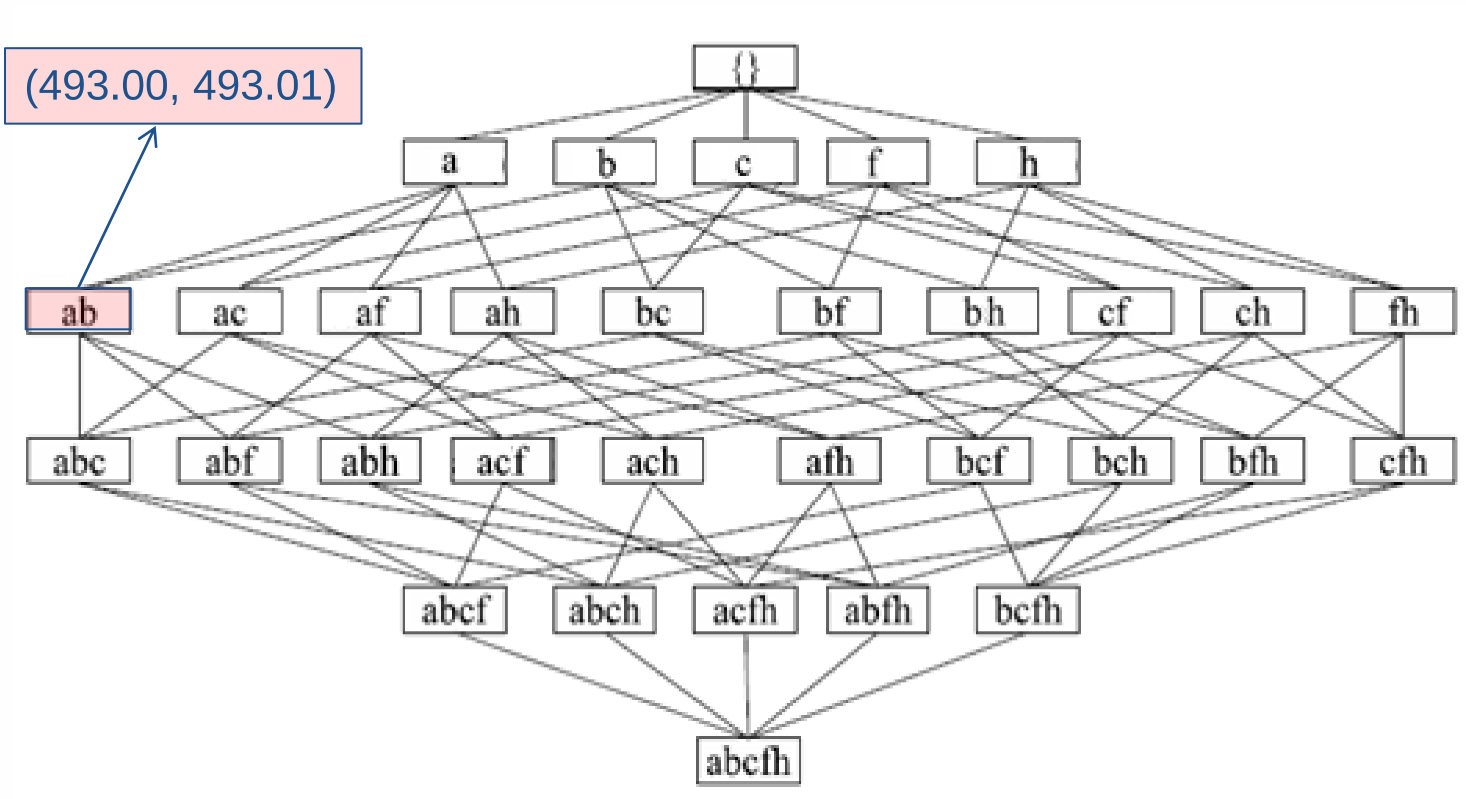
$$RE = \frac{|act(q) - est(q)|}{act(q)} = \frac{1 - 0.02}{1} = 0.98$$

- Classification Metric **(CM)**
  - Penalizes groups with different classification labels

Quantifying data utility for diagnosis codes based on information loss

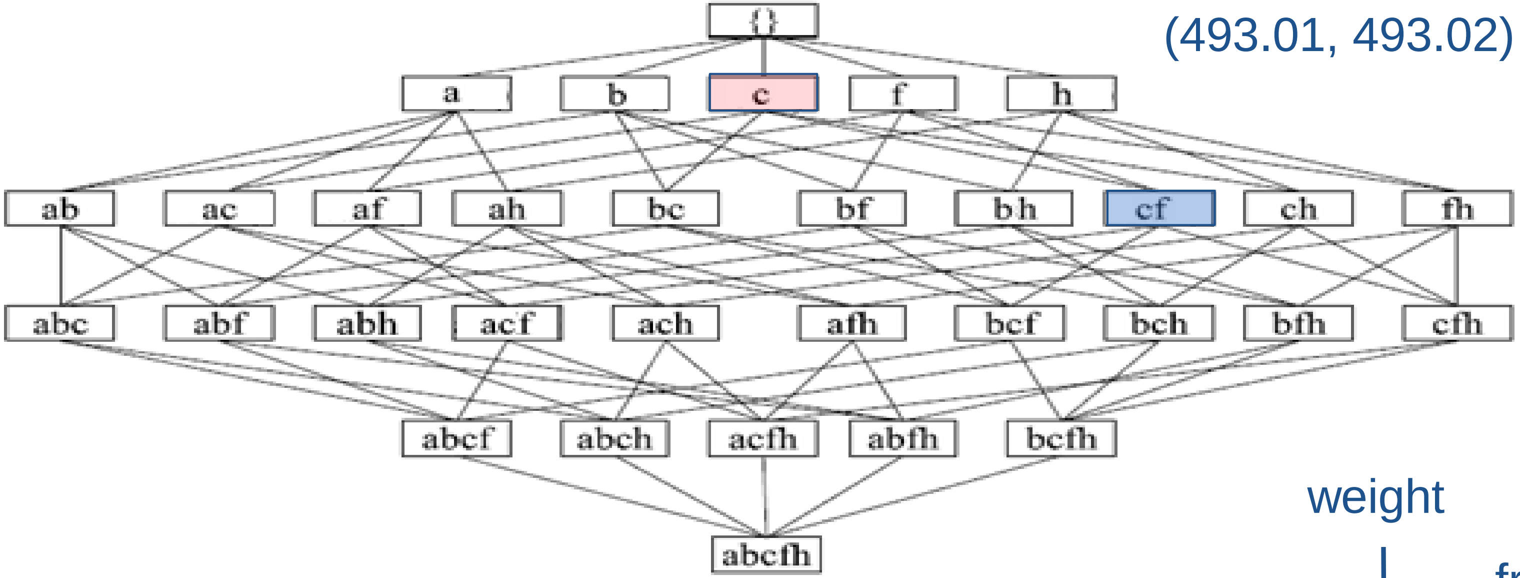
- 1.01×10^{1755} **possible set-based anonymizations for VNEC**

a:	493.00
b:	493.01
c:	296.01
f:	296.02
h:	174.01



Quantifying data utility for diagnosis codes based on information loss

- **Utility Loss (UL):** A measure to quantify the level of information loss incurred by anonymization
  - Favors (493.01) over (493.01, 493.02)



- captures the introduced uncertainty of interpreting an anonymized item
- customizable

weight

# of items mapped to generalized item

fraction of affected transactions

$$UL(\tilde{i_m}) = \frac{2^{|\tilde{i_m}|} - 1}{2^M - 1} \times w(\tilde{i_m}) \times \frac{sup(\tilde{i_m}, \tilde{D})}{N}$$

Quantifying data utility for diagnosis codes based on analytic tasks

- Average Relative Error (**AvgRE**)

ICD	DNA
401.0 401.1	*AC…T*
401.2 401.3	*GC…C*
401.0 401.1	*CC…A*
401.4 401.3	*CA…T*

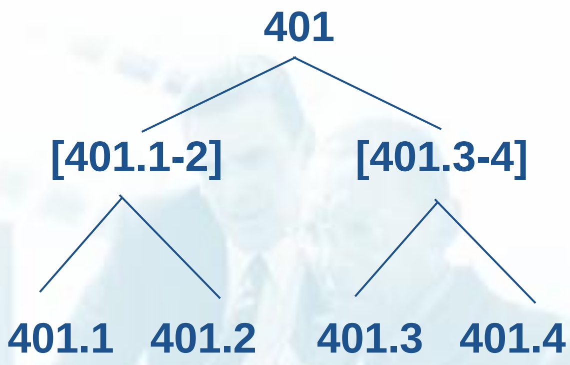
→

ICD	DNA
[401.1-2]	*AC…T*
[401.1-2] 401.3	*GC…C*
[401.1-2]	*CC…A*
401 401.3	*CA…T*

COUNT(*) from T
where Diagnosis is "401.2"

$$est(q) = | g | \times p = 3 \times \frac{2}{3}$$

$$RE = \frac{| act(q) - est(q) |}{act(q)} = \frac{|1-2|}{1} = 1$$



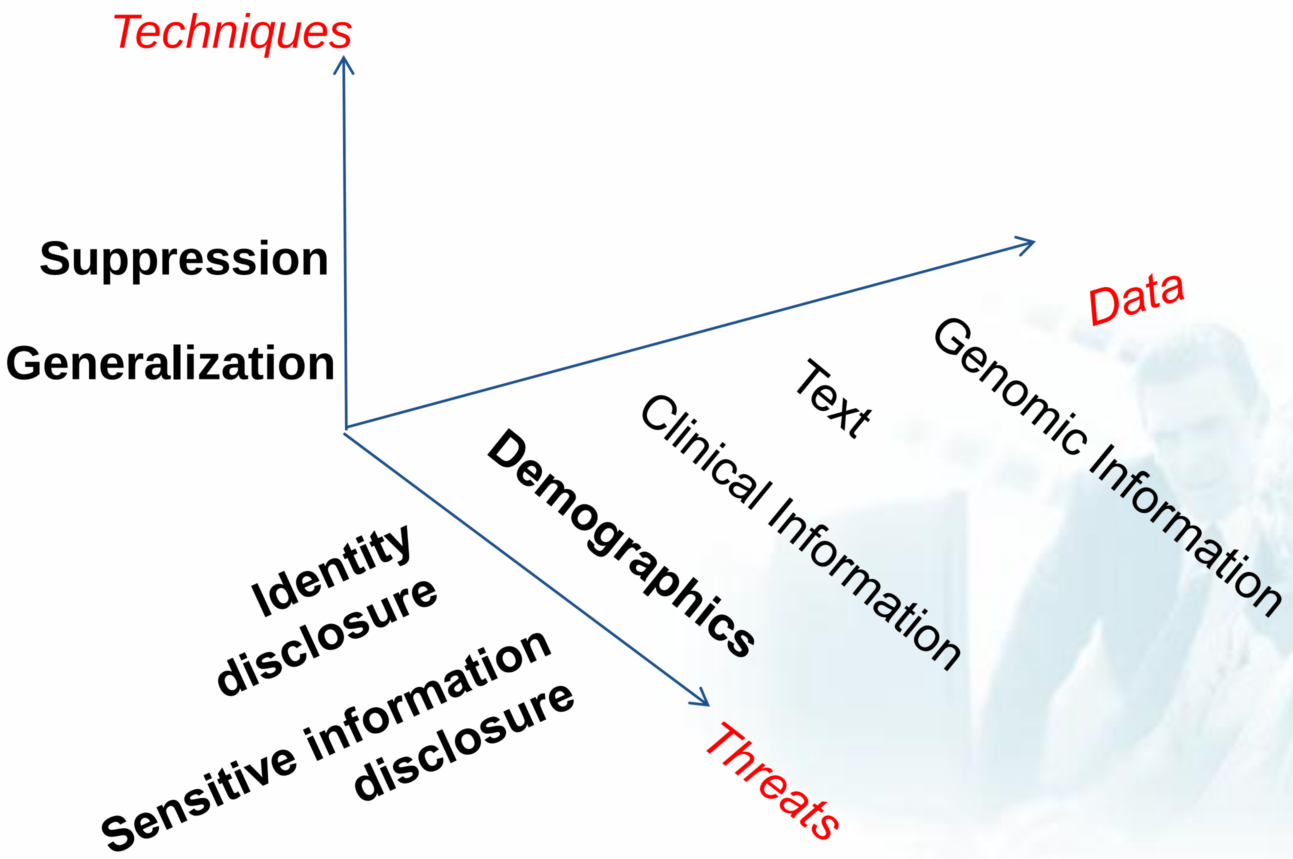
Content

- Part 1: *Medical data sharing and the need for privacy*
- **Part 2: *Research challenges and solutions***
  - Identifying and modeling adversarial knowledge
  - Transforming data to guarantee privacy
  - Quantifying data utility
  - **Privacy-preserving data publishing:**

    principles, methods, case studies
- **Part 3: *Open problems and research directions***

Research challenges and solutions

- **Privacy-preserving data publishing**



Privacy-preserving data sharing: Demographics

- **Principles**
  - k-anonymity
  - k-map
  - l-diversity
  - ρ_1-to-ρ_2 privacy
  - differential privacy

- **Anonymization algorithms**
  - Partition-based
  - Clustering-based

- **Case Study: US Census data**

Anonymization principles for demographics

- **k-anonymity***
  - Each record in a relational table T needs to have the same value over quasi-identifiers with at least k-1 other records in T
  - These records collectively form a *k*-anonymous group

  - Protects from identity disclosure
    - Makes linking to external data more difficult
    - Probability an identified individual is associated with their record is at most 1/k

Name	*Age*	*Postcode*	*Sex*
Greg	**40**	**NW10**	**M**
Jim	**45**	**NW15**	**M**
Jack	**22**	**NW30**	**M**
Anne	**50**	**NW25**	**F**

External data

⋈

Age	*Postcode*	*Sex*
4*	**NW1***	**M**
4*	**NW1***	**M**
*****	**NW***	*****
*****	**NW***	*****

2-anonymous data

* Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. IJUFKS. 2002.

Anonymization principles for demographics

- **k-anonymity**

Pros

- A baseline model
- Intuitive
- Has been implemented in real-world systems

Cons

- Known attacks
- Requires specifying QIDs and k

Name	***Age***	***Postcode***	***Sex***
Greg	**40**	**NW10**	**M**
Jim	**45**	**NW15**	**M**
Jack	**22**	**NW30**	**M**
Anne	**50**	**NW25**	**F**

External data

⋈

Age	***Postcode***	***Sex***
4*	**NW1***	**M**
4*	**NW1***	**M**
*****	**NW***	*****
*****	**NW***	*****

2-anonymous data

Anonymization principles for demographics

- **k-map***
  - Each record in a relational table T needs to have the same value over quasi-identifiers with at least k -1 records in a relational table P from which T is derived
  - Probability **an identified individual in P** is associated with their record is at most 1/k

Name	*Age*	*Postcode*	*Sex*
Greg	40	NW10	M
Jack	40	NW10	M
Jim	45	NW15	M
John	45	NW15	M

Population table

⋈

Age	*Postcode*	*Sex*
40	NW10	M
45	NW15	M

2-mapped data

* Sweeney, Computational Disclosure Control: Theory and Practice. . Massachusetts Institute of Technology, Laboratory for Computer Science, Tech Report, PhD Thesis. 2001.

Anonymization principles for demographics

- **k-map**

Pros

- May allow more useful data than k-anonymity

Cons

- Weaker than k-anonymity
  - attacker does not know whether a record in P is in T or not
- Assumes knowledge of P

- Variations explore different mappings for better utility
  - (k,k)-anonymization*

Name	***Age***	***Postcode***	***Sex***
Greg	**40**	**NW10**	**M**
Jack	**40**	**NW10**	**M**
Jim	**45**	**NW15**	**M**
John	**45**	**NW15**	**M**

Population table

⋈

Age	***Postcode***	***Sex***
40	**NW10**	**M**
45	**NW15**	**M**

2-mapped data

[*] Gionis et al. k-Anonymization revisited. ICDE, 2008.

Attack on k-anonymous data

- **Homogeneity attack***
  - All sensitive values in a k-anonymous group are the same → sensitive information disclosure

Name	*Age*	*Postcode*
Greg	40	NW10

External data

Age	*Postcode*	*Disease*
4*	NW1*	HIV
4*	NW1*	HIV
5*	NW*	Ovarian Cancer
5*	NW*	Flu

2-anonymous data

* Machanavajjhala et al, l-Diversity: Privacy Beyond k-Anonymity. ICDE 2006.

How to prevent homogeneity attack

- **Observation**
  - Given a k-anonymous group G, the probability of a sensitive value u being disclosed is $\frac{f(u)}{|G|}$

Can we limit this probability to prevent sensitive information disclosure?

Age	***Postcode***	***Disease***
4*	**NW1***	**HIV**
4*	**NW1***	**HIV**
5*	**NW***	**Ovarian Cancer**
5*	**NW***	**Flu**

The probability of “flu” being disclosed is 0.5

l-diversity principle for demographics

- ***l*** **-diversity***
  - A relational table is ***l*-diverse** if all groups of records with the same values over quasi-identifiers (QID groups) contain no less than ***l*** "well-represented" values for the SA

  - **Distinct *l*-diversity**

    l "well-represented" → ***l*** distinct

Age	*Postcode*	*Disease*
4*	NW1*	HIV
4*	NW1*	HIV
4*	NW1*	HIV
4*	NW1*	HIV
4*	NW1*	Flu
4*	NW1*	Cancer

Three distinct values, but the probability of *"HIV"* being disclosed is 0.67

l-diversity principles for demographics

- *l* -diversity*
  - <u>**Entropy *l* –diversity**</u>
    - each QID group needs to have ***l*** distinct values that are distributed equally enough: $Entropy(G) \geq \log(l)$
    - can be too restrictive if there are some frequent values in the table (e.g., hypertension in a patient dataset)
  - <u>**Recursive (c,*l*)-diversity**</u>
    - each QID group is (c, *l*)-diverse if and only if

      $$r_1 < c \times (r_l + r_{l+1} + ... + r_n)$$

      where r_i is the i-th most frequent SA value in the group

    - **Intuition:** the most frequent value should not appear “too” frequently in the QID group

Addressing the limitations of l-diversity

- Sensitive values may not need the same level of protection
  - **(a,k)-anonymity**[1]
- *l*-diversity is difficult to achieve when the SA values are skewed
  - **t-closeness**[2]
- Does not consider semantic similarity of SA values
  - **(e,m)-anonymity**[3] **, range diversity**[4]
- Can patients decide the level of protection for their SA values?
  - **Personalized privacy**[5]

[1] Wong et al., (alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing, KDD 2006.

[2] Li et al., t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, ICDE 2007.

[3] Li et al. Preservation of proximity privacy in publishing numerical sensitive data. SIGMOD 2008.

[4] Loukides et al. Preventing range disclosure in k-anonymised data. Expert Syst. Appl. 2011.

[5] Xiao et al. Personalized privacy preservation. SIGMOD, 2006.

ρ_1-to-ρ_2 privacy

- **Probabilistic disclosure -** prior knowledge of adversaries over SA values

- **ρ_1-to-ρ_2 privacy**[*,**] - bounds an adversary's posterior belief in a predicate of a sensitive value by $\boldsymbol{\rho_2}$, given a bound $\boldsymbol{\rho_1}$ on an adversary's prior belief

Definition

Given constants $\rho_1, \rho_2 \in [0,1]$ s.t. $\rho_1 < \rho_2$, X a sensitive value and Y its perturbed version, $Pr[Q(X)], Pr[Q(X)|Y = y]$ the adversary's belief in a predicate $Q(X)$ of X prior and after observing $Y = y$, respectively, the ρ_1-*to*-ρ_2 privacy states that

$$Pr[Q(X)] \leq \rho_1 \text{ implies that } Pr[Q(X)|Y = y] \leq \rho_2$$

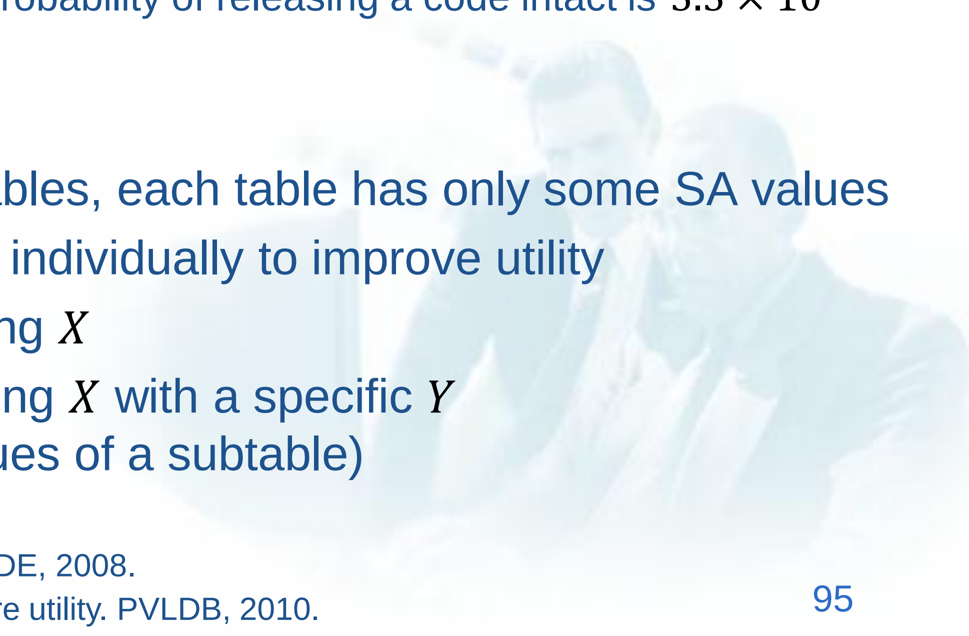
[*] Efvimievski et al. Limiting Privacy Breaches in Privacy Preserving Data Mining, PODS, 2003.

[**] We consider upward ρ1-to-ρ2 privacy breaches.

Limitations of ρ_1-to-ρ_2 privacy

- Does not limit the difference between adversary's prior and posterior belief
  - 0.1-to-0.5 privacy guards against an adversary with $\Pr[Q(x)] \leq 0.1$ by limiting $\Pr[Q(x)|Y = y]$ to 0.5, but not against adversaries with $\Pr[\mathrm{Q}(\mathrm{x})] > 0.1$.
  - Δ-growth* - satisfied when $\Pr[Q(x)] - \Pr[Q(x)|Y = y] \leq \Delta$, for $\Delta \in (0,1]$

- Large amount of noise needs to be added when SA has large domain – sensitive values are rarely released intact
  - There are ~15K distinct ICD-9 codes, the probability of releasing a code intact is 3.3×10^4
  - **Small-domain randomization***
    - Partition table into disjoint subtables, each table has only some SA values
    - Perturb values in each subtable individually to improve utility
      - Higher probability of retaining X
      - Higher probability of replacing X with a specific Y (chosen among the SA values of a subtable)



* Tao et al. On anti-corruption privacy preserving publication. ICDE, 2008.

** Chaytor et al. Small domain randomization: same privacy, more utility. PVLDB, 2010.

ϵ-Differential privacy

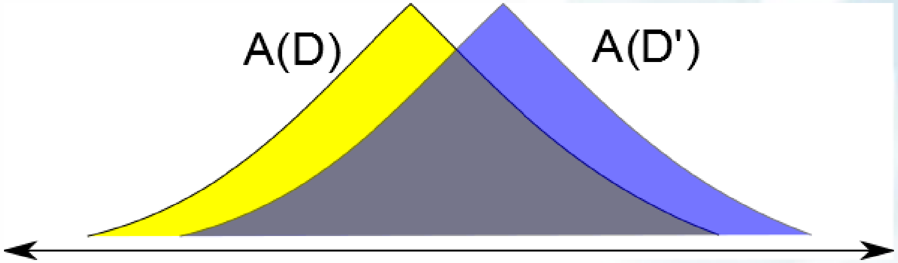
- **Objective –** Prevent an adversary from inferring any additional information about an individual, regardless of whether the published dataset contains the individual's record or not.

- **ϵ-Differential privacy** – satisfied by a randomized algorithm A if

$$\Pr\left[A(D) = \widetilde{D}\right] \leq e^{\epsilon} \times \Pr[A(D') = \widetilde{D}]$$

for all datasets D, D' that differ in one record, and for any possible anonymized dataset $\widetilde{D}$, where ϵ is a constant and the probabilities are over the randomness of A**

- Probability of any event increases by at most $e^{\epsilon} \approx 1 + \epsilon$



* Dwork. Differential privacy. ICALP, 2006.

** Definition from Mohammed et al. Differentially private data release for data mining. KDD, 2011.

ϵ-Differential privacy

- Add random noise to $f(D)$ (true output of a function f) to achieve ϵ-differential privacy

- **Laplace mechanism*** - Add noise from Laplace distribution $\Pr[x|\lambda] = \frac{1}{2\lambda} \times e^{-x/\lambda}$

Theorem*

For any function $f: D \rightarrow R^d$, the algorithm A that adds independently generated noise with distribution $\mathrm{Lap}(\Delta_f/\epsilon)$ to each of its d outputs satisfies ϵ-differential privacy, where $\Delta_f = \ max_{D,D'}|f(D) - f(D')|$ for all datasets D, D' that differ in one record.

Age	*Sex*
20	M
23	F
25	M
40	F

f- returns the number of patients with $Age \ < \ 40$

$f(D) = 3$

$\Delta_f \ = 1$

Add noise with distribution $Lap\left(\frac{1}{\epsilon}\right)$ to $f\ (D)$

$$f\big(\widetilde{D}\big) = 3 + Lap(\frac{1}{\epsilon})$$

[*] Dwork et al. Calibrating noise to sensitivity in private data analysis. TCC, 2006.

ϵ-Differential privacy

- **Exponential mechanism***
  - adding noise makes no sense in some tasks, when the output of a function is not a number (e.g., partition a dataset D along an attribute)
  - there is a function $u: (D \times t) \rightarrow R$ that measures the utility of an output $t \in T$ and induces a probability over the output domain
  - the exponential mechanism samples t from this distribution, favoring outputs with large utility

Theorem*

For any function u, an algorithm A that output t chosen from T with probability proportional to $exp(e \times \frac{u(D,t)}{2\Delta u})$ satisfies ϵ-differential privacy, where

$\Delta_u = max_{\forall t,D,D'}|u(D,t) - u(D',t)|$

Age	*Sex*
[20-41)	{M,F}
[20-41)	{M,F}
[25-41)	{M,F}
[25-41)	{M,F}

u- scores attribute to specialize according to utility loss

exponential mechanism to select Age or Sex

* McSherry et al. Mechanism design via differential privacy. FOCS, 2007.

ϵ -Differential privacy

(+)

- **semantic definition** – no assumptions on adversarial knowledge
- **composability**[1] – privacy holds even when multiple differentially-private datasets are obtained by an adversary
- **many mechanisms** for the interactive[2] and the non-interactive scenario [3,4]

(-)

- **data cannot be analyzed at a record-level (important in the medical domain)**
- **returned answers are noisy and, typically, of low utility**
  – several variations[5], improved mechanisms[6]
- **misconceptions**[7] and **susceptibility to attacks**[8]

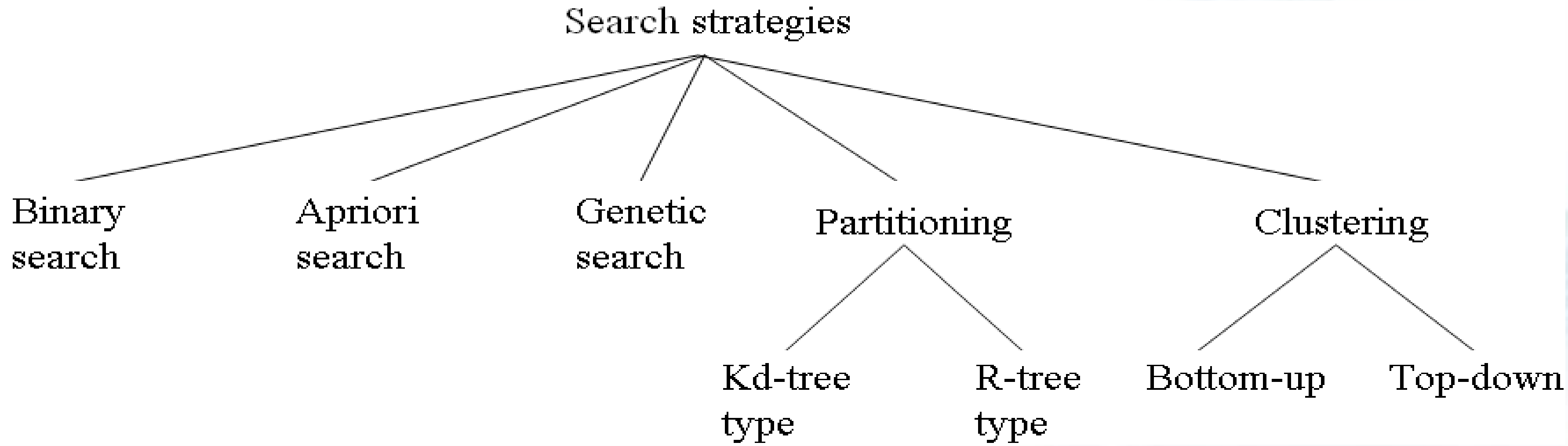
[1] Ganta et al. Composition attacks and auxiliary information in data privacy. KDD, 2008.
[2] Dwork. Differential privacy: a survey of results. TAMC, 2008.
[3] Mohammed. Differentially private release for data mining. KDD, 2011.
[4] Xiao et al. Differential privacy via wavelet transforms. ICDE, 2010.
[5] Machanavajjhala et al. Data Publishing against Realistic Adversaries. PVLDB, 2009.
[6] Ding et al. Differentially private data cubes: optimizing noise sources and consistency. SIGMOD, 2011.
[7] Kifer et al. No free lunch in data privacy. SIGMOD, 2011.
[8] Cormode. Personal privacy vs population privacy: learning to attack anonymization. KDD, 2011.

Anonymization algorithms for demographics

- **Goal -** Transform data in a way that satisfies privacy with minimal utility loss

- **Problem -** many different anonymizations and finding the one with best utility is NP-hard

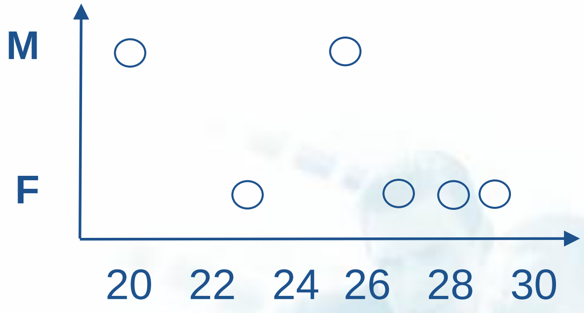
- **Optimal and heuristic algorithms**



Partition-based algorithms for k-anonymity

- **Main idea of partition-based algorithms**
  - A record projected over QIDs is treated as a multidimensional point
  - A subspace (hyper-rectangle) that contains at least ***k*** points can form a ***k***-anonymous group → multidimensional global recoding

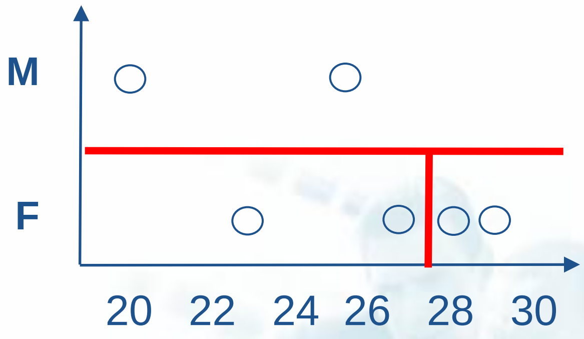
Age	***Sex***	***Disease***
20	**M**	**HIV**
23	**F**	**HIV**
25	**M**	**Obesity**
27	**F**	**HIV**
28	**F**	**Cancer**
29	**F**	**Obesity**



Partition-based algorithms for k-anonymity

- **Main idea of partition-based algorithms**
  - A record projected over QIDs is treated as a multidimensional point
  - A subspace (hyper-rectangle) that contains at least ***k*** points can form a ***k***-anonymous group → multidimensional global recoding

Age	*Sex*	*Disease*
20	M	HIV
23	F	HIV
25	M	Obesity
27	F	HIV
28	F	Cancer
29	F	Obesity



  - **How to partition the space?**
    - One attribute at a time – which to use?
    - How to split the selected attribute?

Mondrian algorithm

Mondrian(D,k)*

- Find the QID attribute ***Q*** with the largest domain

Attribute selection

- Find the median ***μ*** of ***Q***
- Create subspace ***S*** with all records of ***D*** whose value in ***Q*** is less than ***μ***
- Create subspace ***S'*** with all records of ***D*** whose value in ***Q*** is at least ***μ***

Attribute split

- **If |S|≥k** or **|S'| ≥ k**

  Return *Mondrian(**S**,k)* U Mondrian(***S'***,k)
- **Else Return *T***

Recursive execution

* LeFevre et al. Mondrian multidimensional k-anonymity, ICDE, 2006.

Mondrian algorithm

Mondrian(D,k)*

- Find the QID attribute ***Q*** with the largest domain

Attribute selection

- Find the median ***μ*** of ***Q***
- Create subspace ***S*** with all records of ***T*** whose value in ***Q*** is <u>less than</u> ***μ***
- Create subspace ***S'*** with all records of ***T*** whose value in ***Q*** is at least ***μ***

Attribute split

- **If |S|≥k** or **|S'| ≥ k**

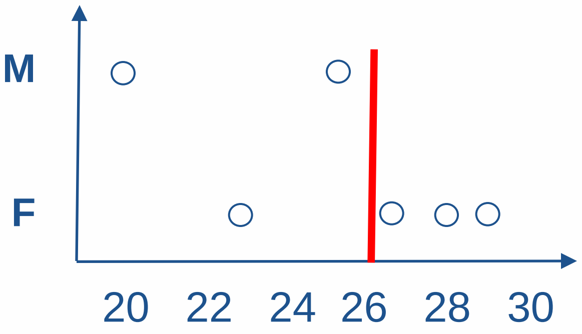
  Return *Mondrian(**S**,k)* U Mondrian(***S'***,k)
- **Else Return *T***

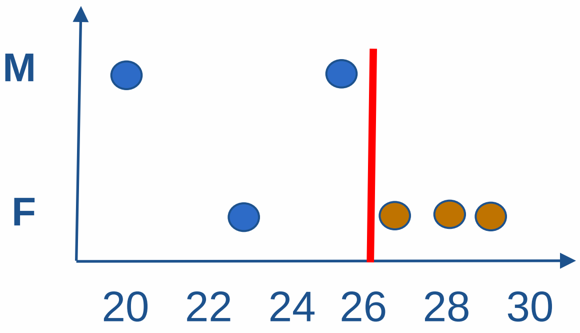
Recursive execution

Optimizes group size

Cost: $O(|T|\log(|T|))$, where T the size of original dataset

Example of applying Mondrian (k=2)





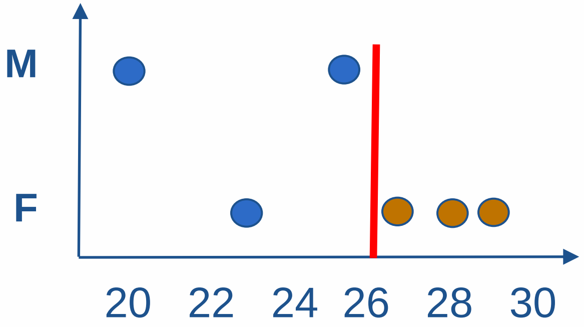
Age	*Sex*	*Disease*
20	M	HIV
23	F	HIV
25	M	Obesity
27	F	HIV
28	F	Cancer
29	F	Obesity

→

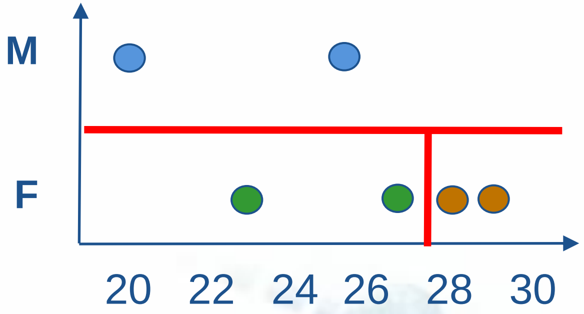
Age	*Sex*	*Disease*
[20-26]	{M,F}	HIV
[20-26]	{M,F}	HIV
[20-26]	{M,F}	Obesity
[27-29]	{M,F}	HIV
[27-29]	{M,F}	Cancer
[27-29]	{M,F}	Obesity

Example of applying Mondrian (k=2)

Example of Mondrian algorithm (k=2)



- Heuristic attribute selection for efficiency
  → there may be better splits



Age	*Sex*	*Disease*
[20-26]	{M,F}	HIV
[20-26]	{M,F}	HIV
[20-26]	{M,F}	Obesity
[27-29]	{M,F}	HIV
[27-29]	{M,F}	Cancer
[27-29]	{M,F}	Obesity

Age	*Sex*	*Disease*
[20-25]	M	HIV
[20-25]	M	Obesity
[23-27]	F	HIV
[23-27]	F	HIV
[28-29]	F	Cancer
[28-29]	F	Obesity

Other works on partition-based algorithms

- **R-tree based algorithm [1]**
- **Optimized partitioning for intended tasks [2]**
  - Classification
  - Regression
  - Query answering
- **Algorithms for disk-resident data [3]**
- **Algorithms to prevent sensitive information disclosure [4]**

[1] Iwuchukwu et al. K-anonymization as spatial indexing: toward scalable and incremental anonymization, VLDB, 2007.
[2] LeFevre et al. Workload-aware anonymization. KDD, 2006.
[3] LeFevre et al. Workload-aware anonymization techniques for large-scale datasets. TODS, 2008.
[4] Loukides et al. Preventing range disclosure in k-anonymised data. Expert Syst. Appl. 2011.

Clustering-based anonymization algorithms

- **Main idea of clustering-based anonymization**

1. Create clusters containing at least **k** records with "similar" values over QIDs

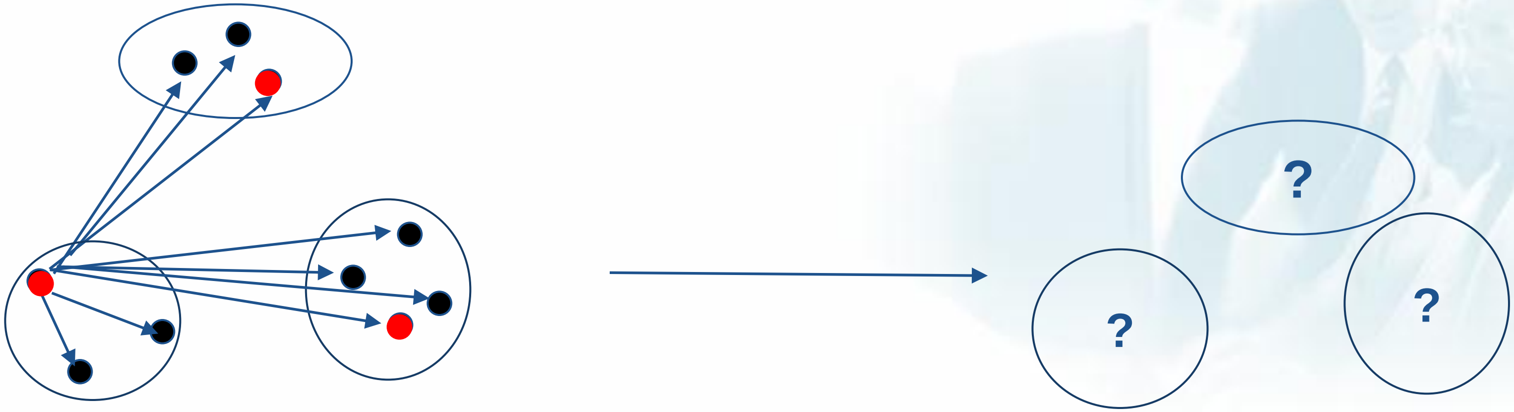
2. Anonymize records in each cluster separately

Seed selection

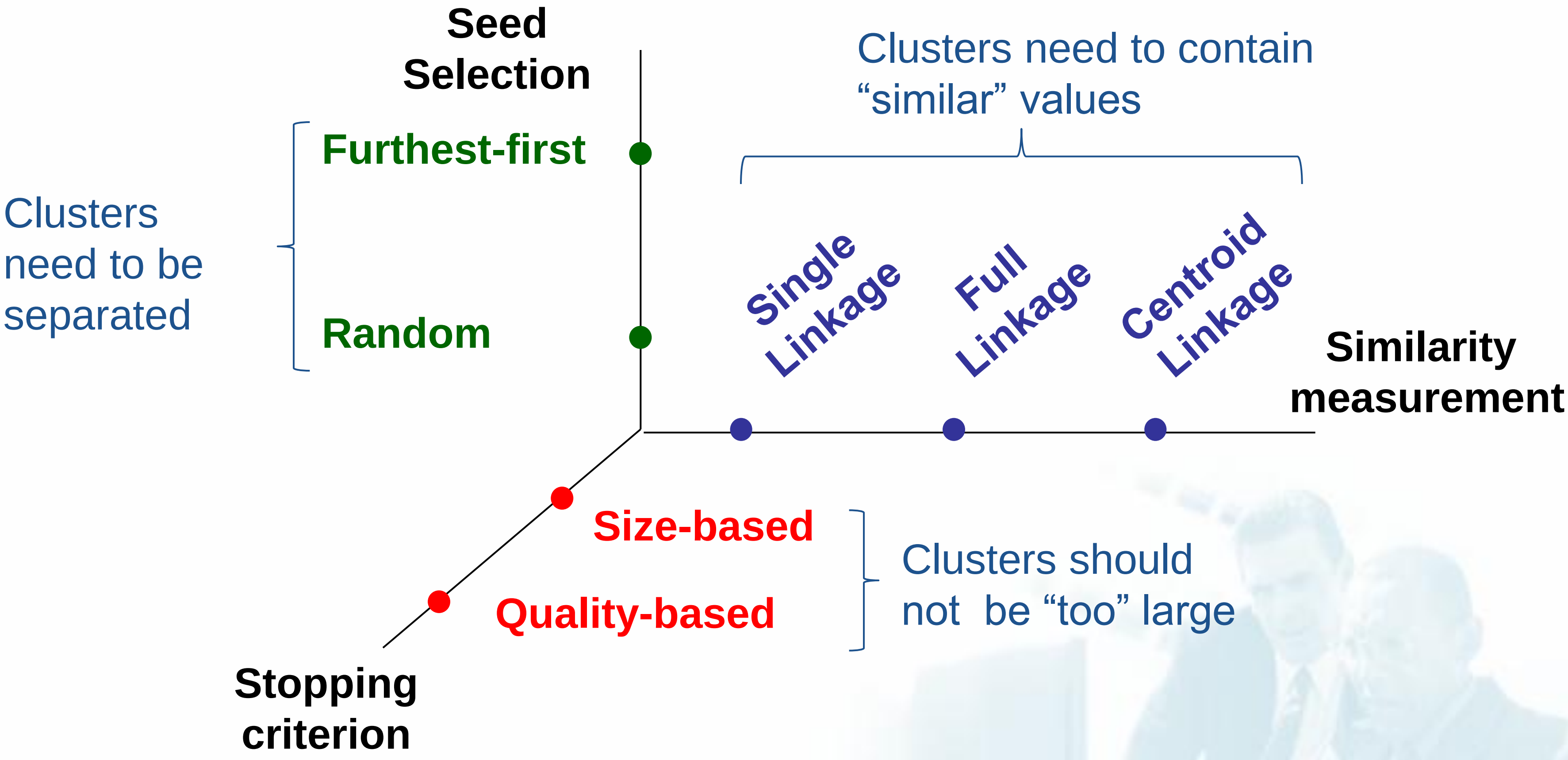
Similarity measurement

Stopping criterion

Local recoding and/or Suppression



Clustering-based anonymization algorithms



Seed Selection

Furthest-first

Random

Clusters need to be separated

Similarity measurement

Single Linkage

Full Linkage

Centroid Linkage

Clusters need to contain “similar” values

Stopping criterion

Size-based

Quality-based

Clusters should not be “too” large

- **All these heuristics attempt to improve data utility**

Bottom-up clustering algorithm

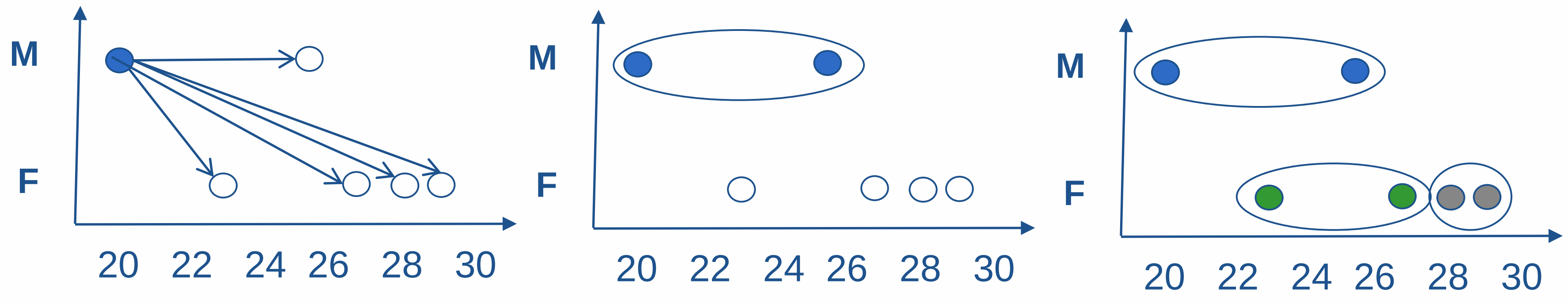
Bottom-up clustering algorithm*

- Each record is selected as a *seed* to start a cluster
- While there exists group G s.t. $|G|<k$
  - For each group G s.t. $|G|<k$
    - Find group G' s.t. $NCP(G \cup G')$ is min. and merge G and G'
  - For each group G s.t. $|G|>2\times k$
    - Split G into $\left\lfloor \frac{|G|}{k} \right\rfloor$ groups s.t. each group has at least k records
- Generalize the QID values in each group
- Return all groups

Cost: $O(|T|^2 \times \log(k))$

* Xu et al. Utility-Based Anonymization Using Local Recoding, KDD, 2006.

Example of Bottom-up clustering algorithm (k=2)



Age	*Sex*	*Disease*
20	M	HIV
23	F	HIV
25	M	Obesity
27	F	HIV
28	F	Cancer
29	F	Obesity

Age	*Sex*	*Disease*
[20-25]	M	HIV
[20-25]	M	Obesity
[23-27]	F	HIV
[23-27]	F	HIV
[28-29]	F	Cancer
[28-29]	F	Obesity

Top-down clustering algorithm

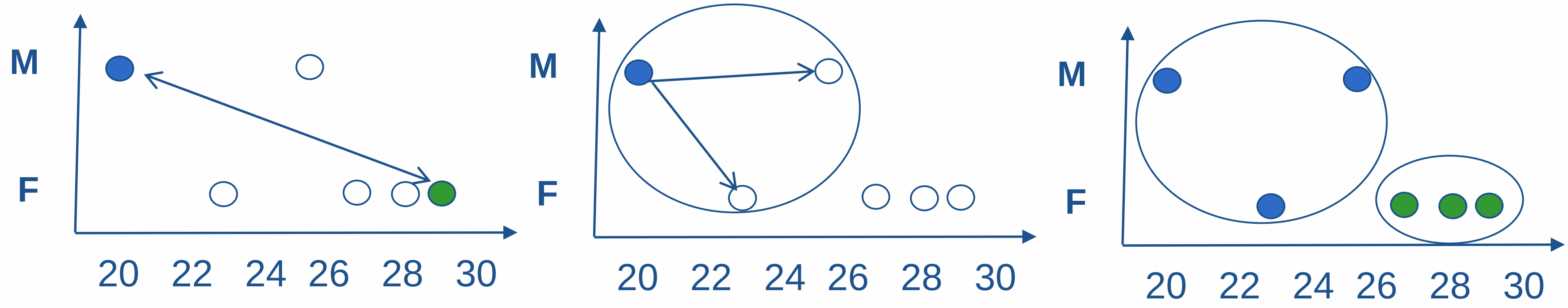
Top-down clustering algorithm*

- If $|T| \leq k$ then *Return* T
- Else
  - Chose two *seeds* s and s' from T *s.t.* $NCP(s \cup s')$ is maximum
  - Form a group G that contains s
  - Form a group G' that contains s'
  - For each record r in $T - \{G \cup G'\}$
    - If $NCP(G \cup r) < NCP(G' \cup r)$ then $G \leftarrow G \cup r$
    - Else $G' \leftarrow G' \cup r$
  - If $|G| \geq k$ then recursively partition G
  - If $|G'| \geq k$ then recursively partition G'
- Anonymize each of the final clusters separately

Cost: $O(|T|^2)$ - slightly lower than that of Bottom-up clustering

* Xu et al. Utility-Based Anonymization Using Local Recoding, KDD, 2006.

Example of top-down clustering algorithm (k=2)

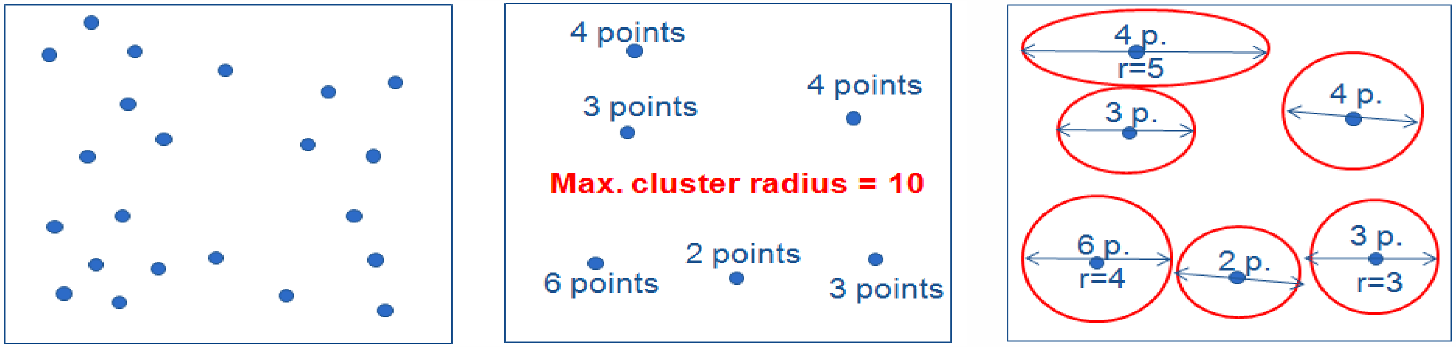


Age	*Sex*	*Disease*
20	M	HIV
23	F	HIV
25	M	Obesity
27	F	HIV
28	F	Cancer
29	F	Obesity

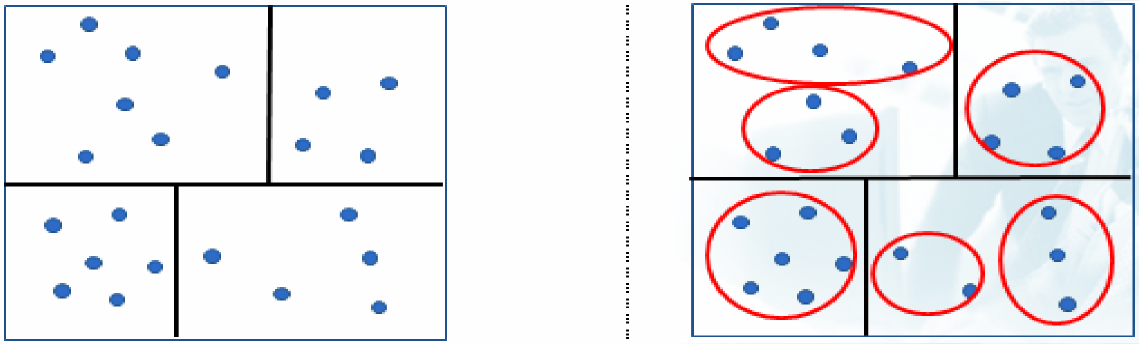
Age	*Sex*	*Disease*
[20-25]	{M,F}	HIV
[20-25]	{M,F}	HIV
[20-25]	{M,F}	Obesity
[27-29]	F	HIV
[27-29]	F	Cancer
[27-29]	F	Obesity

Other works on clustering-based anonymization

- Constant factor approximation algorithms*
  - Publish only the cluster centers along with radius information



- Combine partitioning with clustering for efficiency**



* Aggarwal et al. Achieving anonymity via clustering. ACM Trans. on Algorithms, 2010.

** Loukides et al. Preventing range disclosure in k-anonymised data. Expert Syst. Appl. 2011.

Case study: US Census Data

- **US Census data****
  - Adults dataset – 30162 records

Attribute	Domain Size
Age	74
Gender	2
Race	5
Salary	2
Country	41
Work-Class	7
Marital Status	7
Occupation	14
Education	16

- **Clustering – Bottom-up, Top-down***
- **Partitioning – Mondrian**

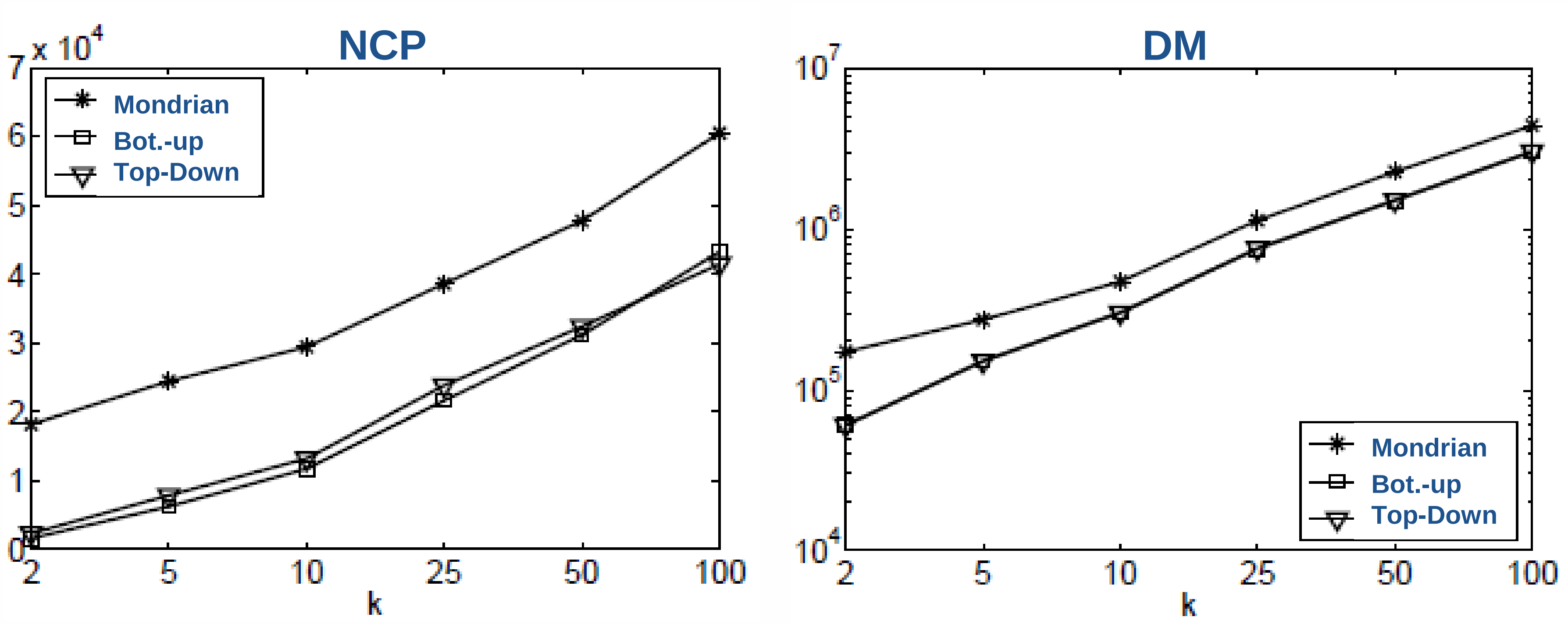
- **How much utility is lost by anonymization? ***
  - DM
  - NCP
  - RE

* Blake et al. UCI repository of machine learning databases, 1998.

** Some results are based on Xu et al. Utility-based anonymization using local recoding, KDD, 2006.

Case study: US Census Data

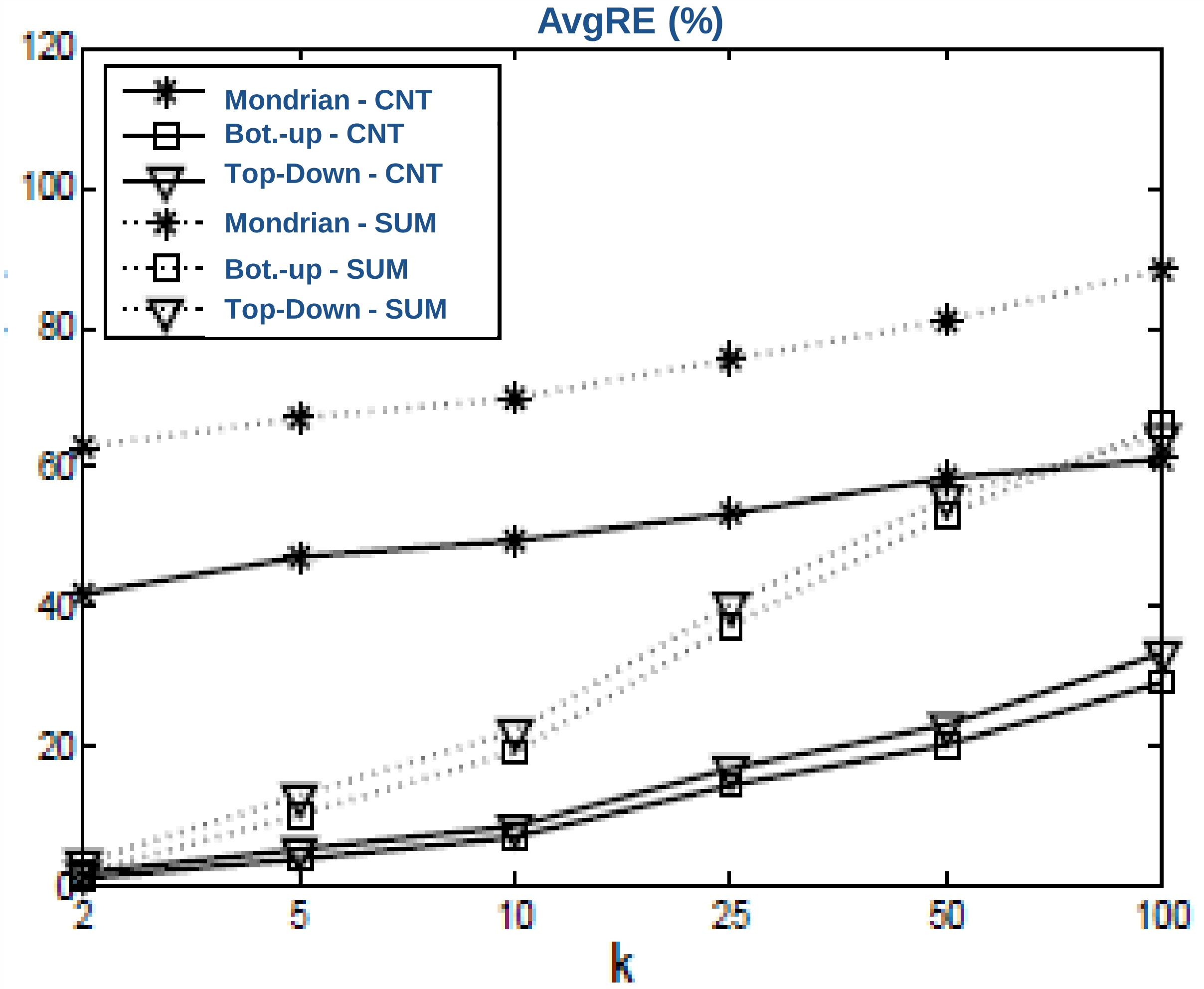
- Utility vs. Privacy (varying *k*) – Information Loss Metrics



  - Small *k* values better for utility
  - Clustering outperforms Mondrian
  - Bottom-up slightly better

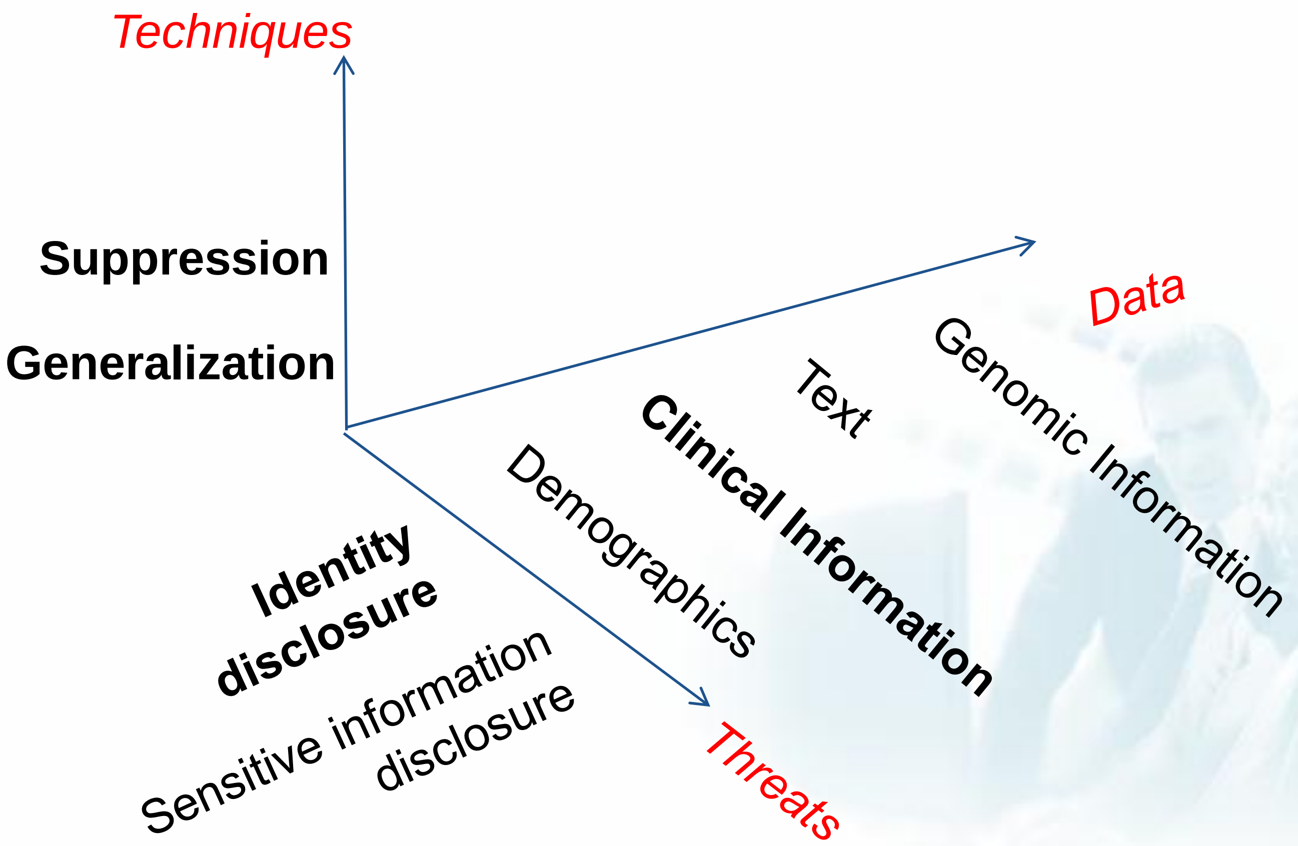
Case study: US Census Data

- **Utility vs. Privacy (varying *k*) – Query Answering**



Research challenges and solutions

- **Privacy-preserving data publishing**



Privacy-preserving data sharing: Clinical information

- **Focus on diagnosis codes**
  - High replication (each visit generates a number of diagnosis codes)
  - High availability (contained in publicly available discharge summaries)
  - High distinguishability (discussed already)

  compared to lab results and other clinical information

- **The problem**
  - prevent the association between a patient and their record based on diagnosis codes (identity disclosure)
    - Needed to satisfy policies (HIPAA, NIH GWAS policy,…)
    - Records can be associated with DNA sequences that are highly sensitive and can be misused or abused

Complete k-anonymity

- ***Complete k-anonymity:*** Knowing that an individual is associated with any itemset, an attacker should not be able to associate this individual to less than ***k*** transactions

ICD	DNA
401.0 401.1	*AC…T*
401.2 401.3	*GC…C*
401.0 401.1	*CC…A*
401.4 401.3	*CA…T*

Original data

ICD	DNA
401.0 401.1	*AC…T*
401 401.3	*GC…C*
401.0 401.1	*CC…A*
401 401.3	*CA…T*

2-complete anonymous data

- **Prevents identity disclosure**
  - Probability of linking an individual to their record is at most 1/k
- **Guards against attackers who know any part of the record**
  - e.g., physicians with access to identified EMRs

* He et al. Anonymization of Set-Valued Data via Top-Down, Local Generalization. PVLDB, 2009.

Complete k-anonymity

- ***Complete k-anonymity:*** Knowing that an individual is associated with any itemset, an attacker should not be able to associate this individual to less than ***k*** transactions

ICD	DNA
401.0 401.1	*AC…T*
401.2 401.3	*GC…C*
401.0 401.1	*CC…A*
401.4 401.3	*CA…T*

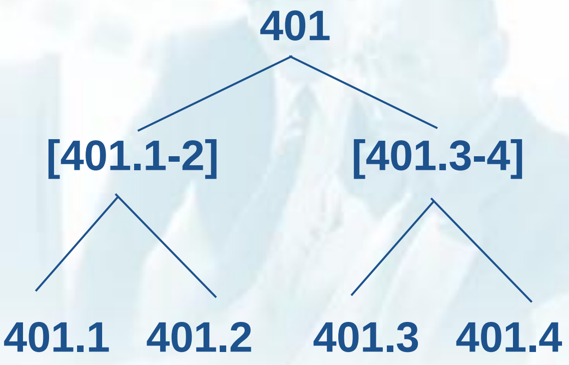
Original data

ICD	DNA
401.0 401.1	*AC…T*
401 401.3	*GC…C*
401.0 401.1	*CC…A*
401 401.3	*CA…T*

2-complete anonymous data

- **Hierarchy-based, local recoding generalization**

- **Information loss can be high!**



401

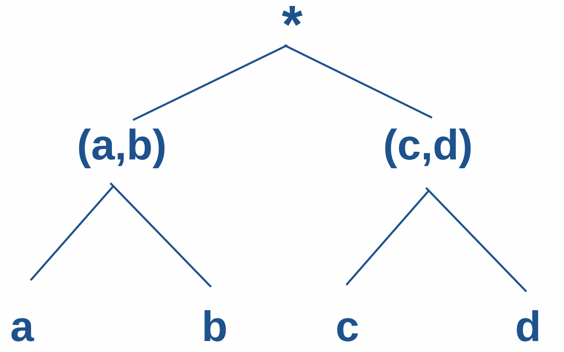
[401.1-2] **[401.3-4]**

401.1 **401.2** **401.3** **401.4**

Algorithm to enforce complete k-anonymity

Greedy partitioning (Sketch)

- Start with most general data P (all values are generalized to *)
- **If** *complete k-anonymity* is not satisfied
  - **Return** partition
- **Else**
  - **Find** the node u in the hierarchy that incurs minimum information loss if replaced by its ascendants
  - **Replace** u with its ascendants
  - Generate all possible subpartitions of P
  - **For each** transaction T in P
    - distribute T into a subpartition based on its generalized items
  - **Balance** subpartitions so that they have at least ***k*** transactions
  - **For each** subpartition
    Recursively execute ***Greedy partitioning***
- Construct anonymous dataset based on returned partitions



$P_{(a,b)}$ $P_{(c,d)}$ $P_{(a,b)(c,d)}$

ICD	DNA
a b	*AC…T*
c	*GC…C*
c d	*CC…A*
a b c d	*CA…T*

$\rightarrow P_{(a,b)}$

$\rightarrow P_{(c,d)}$

$\rightarrow P_{(c,d)}$

$\rightarrow P_{(a,b)(c,d)}$

k^m-anonymity

- ***k^m*-anonymity:** Knowing that an individual is associated with any ***m***-itemset, an attacker should not be able to associate this individual to less than ***k*** transactions

ICD	DNA
401.0 401.1	*AC…T*
401.2 401.3	*GC…C*
401.0 401.1	*CC…A*
401.4 401.3	*CA…T*

Original data

ICD	DNA
401	*AC…T*
401	*GC…C*
401	*CC…A*
401	*CA…T*

4^2- anonymous data

- **Prevents from identity disclosure**
- **Can be used to model different attacks**
  - e.g., discharge summaries contain < 10 diagnoses codes → no need for complete k-anonymity to prevent the "two-step" attack

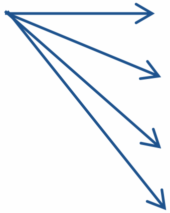
* Terrovitis et al. Privacy-preserving anonymization of set-valued data. PVLDB, 2008.

k^m-anonymity

- ***k^m*-anonymity:** Knowing that an individual is associated with any ***m***-itemset, an attacker should not be able to associate this individual to less than ***k*** transactions

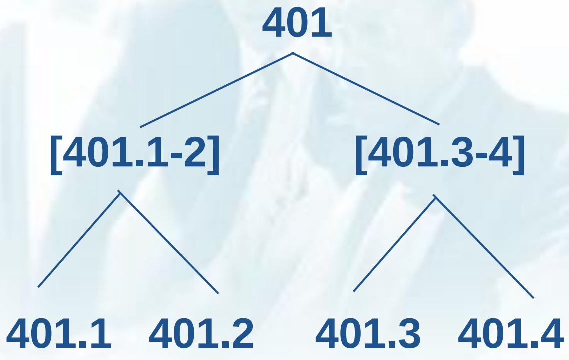
ICD	DNA
401.0 401.1	*AC…T*
401.2 401.3	*GC…C*
401.0 401.1	*CC…A*
401.4 401.3	*CA…T*

Original data



ICD	DNA
401	*AC…T*
401	*GC…C*
401	*CC…A*
401	*CA…T*

4^2- anonymous data



401

[401.1-2] **[401.3-4]**

401.1 **401.2** **401.3** **401.4**

- **Global, full-subtree recoding**
  - **more information loss than local recoding**

* Terrovitis et al. Privacy-preserving anonymization of set-valued data. PVLDB, 2008.

Algorithm to enforce k^m-anonymity

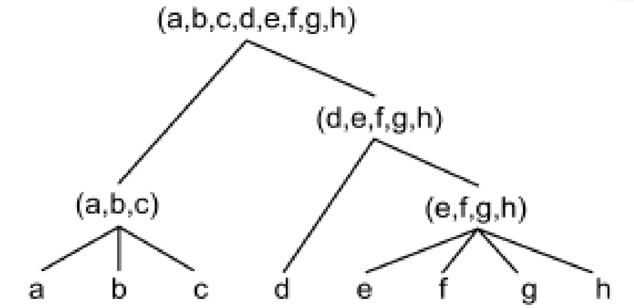
Apriori Anonymization (Sketch)

- Start with original data
- **For** ***j=1*** to ***m***
    - **For each** transaction ***T***
        - Consider all the ***j***-itemsets of ***T*** (generalized or not)
        - Find all those itemsets with support less than ***k***
        - For each of these itemsets
            - Generate all possible generalizations
        - Find the generalization that satisfies ***k^m-anonymity*** and has minimum information loss

Diagnosis Codes
a, b, c, d, e, f, g, h
a, c, e, f, g
c, d, e, f, h
a, c, e, f
e, f, g, h
d, e, f, g
a, b, d, e
a, c, f
a, c
b, h

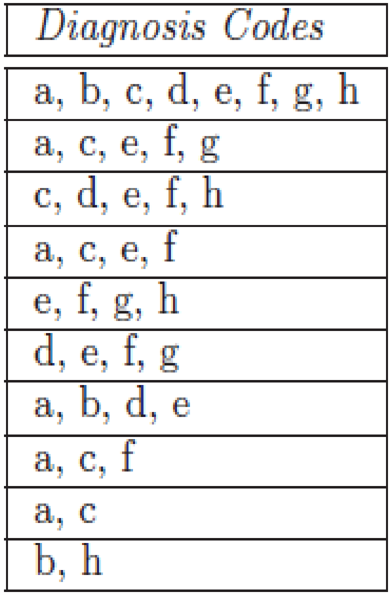
5^3-anonymity ↓

Diagnosis Codes
(a,b,c), (d,e,f,g,h)
(a,b,c), (d,e,f,g,h)
(a,b,c), (d,e,f,g,h)
(a,b,c), (d,e,f,g,h)
(d,e,f,g,h)
(d,e,f,g,h)
(a,b,c), (d,e,f,g,h)
(a,b,c), (d,e,f,g,h)
(a,b,c)
(a,b,c), (d,e,f,g,h)



Applicability of complete k-anonymity and k^m-anonymity to medical data

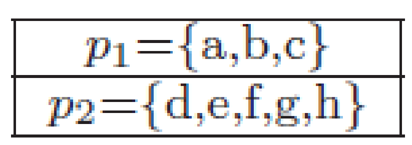
- **Limited in the specification of privacy requirements**
  - Assume powerful attackers
    - all *m*-itemsets (combinations of *m* diagnosis codes) need protection
  - but… medical data publishers have <u>detailed</u> privacy requirements



Diagnosis Codes
a, b, c, d, e, f, g, h
a, c, e, f, g
c, d, e, f, h
a, c, e, f
e, f, g, h
d, e, f, g
a, b, d, e
a, c, f
a, c
b, h

Attackers know who is diagnosed with **abc or defgh**

They protect all 5-itemsets instead of the 2 itemsets



p_1={a,b,c}
p_2={d,e,f,g,h}

privacy constraints

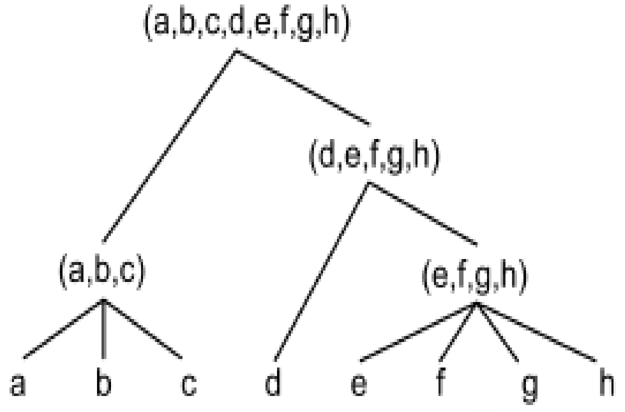
Applicability of complete k-anonymity and k^m-anonymity to medical data

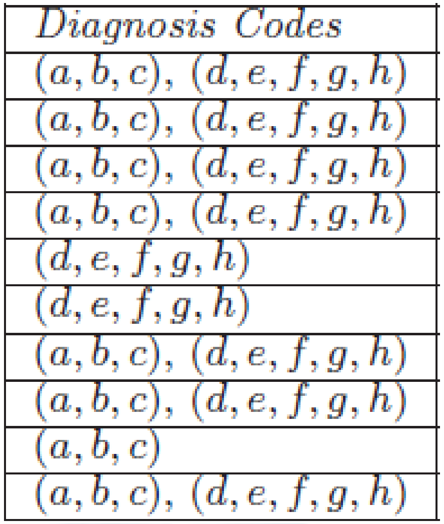
- **Explore a small number of possible generalizations**

Full sub-tree generalization

a,b cannot be replaced by **(a,b)**

c,e cannot be replaced by **(c,e)**





Diagnosis Codes
$(a, b, c), (d, e, f, g, h)$
$(a, b, c), (d, e, f, g, h)$
$(a, b, c), (d, e, f, g, h)$
$(a, b, c), (d, e, f, g, h)$
(d, e, f, g, h)
(d, e, f, g, h)
$(a, b, c), (d, e, f, g, h)$
$(a, b, c), (d, e, f, g, h)$
(a, b, c)
$(a, b, c), (d, e, f, g, h)$

- **Do not take into account utility requirements**
  - Can we perform GWAS as accurately as if we had original data?

Policy-based Anonymization: privacy model

- **Data publishers specify diagnosis codes that need protection**

- **Privacy Model:** Knowing that an individual is associated with <u>one or more specific itemsets</u> ***(privacy constraints)***, an attacker should not be able to associate this individual to less than ***k*** transactions

ICD	DNA
401.0 401.1	*AC…T*
401.2 401.3	*GC…C*
401.0 401.1	*CC…A*
401.4 401.3	*CA…T*

Original data

→

ICD	DNA
401.0 401.1	*AC…T*
(401.2, 401.4) 401.3	*GC…C*
401.0 401.1	*CC…A*
(401.2, 401.4) 401.3	*CA…T*

Anonymized data

- **Privacy Policy:** The set of all specified privacy constraints

Policy-based Anonymization: privacy model

- **Privacy achieved** when all privacy constraints are supported by at least k transactions in the published data or are not supported

ICD	DNA
401.0 401.1	*AC…T*
401.2 401.3	*GC…C*
401.0 401.1	*CC…A*
401.4 401.3	*CA…T*

Original data

ICD	DNA
401.0 401.1	*AC…T*
(401.2, 401.4) 401.3	*GC…C*
401.0 401.1	*CC…A*
(401.2, 401.4) 401.3	*CA…T*

Anonymized data

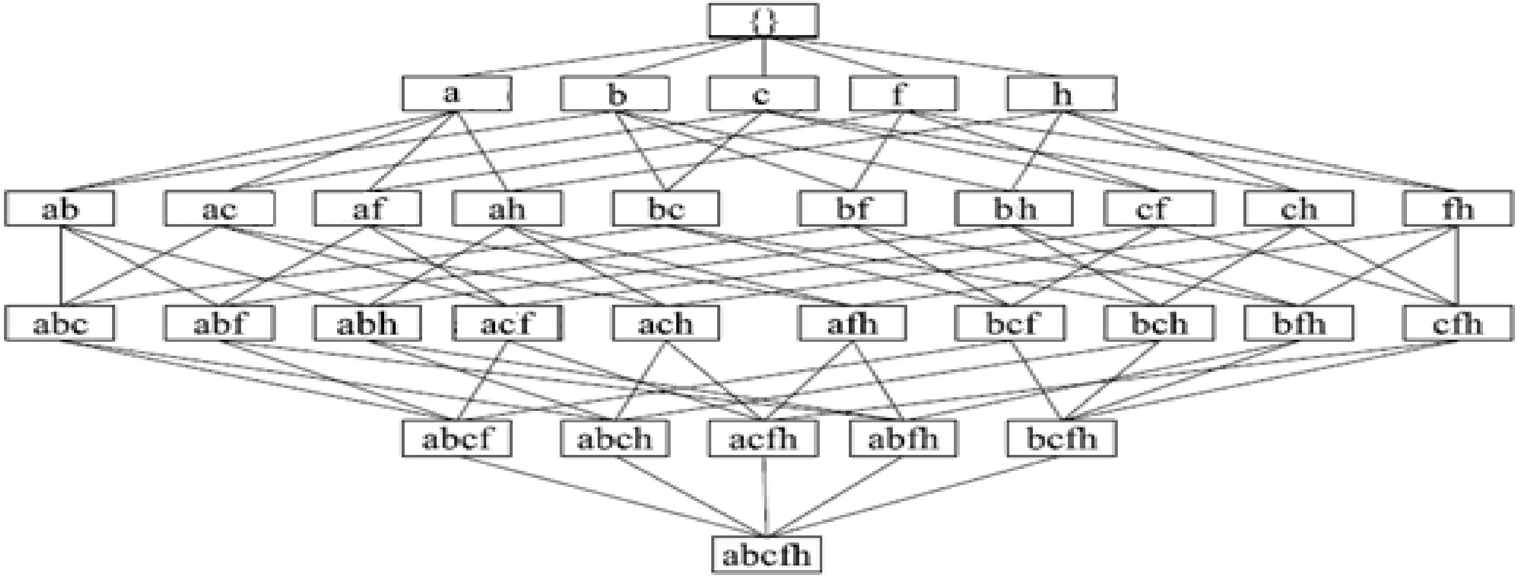
- **Protection against identity disclosure**
  - Probability of re-identification given the data and the specified sets of ICD codes ≤ ***1/k***

- Automatic construction of privacy policies from hospital discharge summaries – PPE algorithm

Policy-based Anonymization: data utility considerations

- **Published data must remain as useful as the original data for conducting a GWAS on a disease**

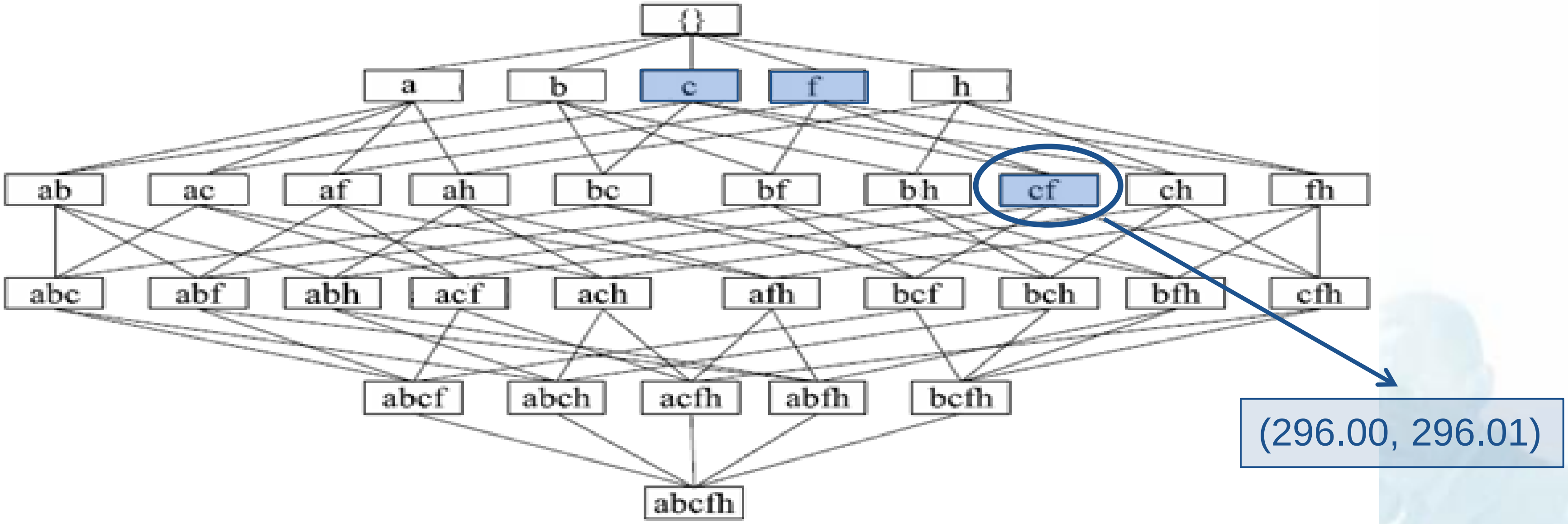
- Set-based anonymization to search a large part of the solution space



- Minimize the Utility Loss (UL) measure

Policy-based Anonymization: data utility considerations

- **Utility Constraints** to specify the maximum level of anonymization
- Enforcing utility constraints guarantees data utility for GWAS
  - the number of cases and controls are preserved



(296.00, 296.01)

- Utility constraints can be specified manually or extracted from electronic medical records (UPE algorithm)

UGACLIP algorithm

UGACLIP (sketch)

- **While** the Privacy Policy is not satisfied
  - Select the privacy constraint ***p*** that corresponds to most patients
  - **While *p*** is not satisfied
    - Select the ICD code ***i*** in ***p*** that corresponds to fewest patients
    - **If *i*** can be anonymized according to the Utility Policy
      - **generalize i** to **(i,i')**
    - **Else**

      suppress each unprotected ICD code in ***p***

Considers one privacy constraint at a time

Protects a privacy constraint by set-based anonymization

- Generalization when Utility Policy is satisfied
- otherwise suppression

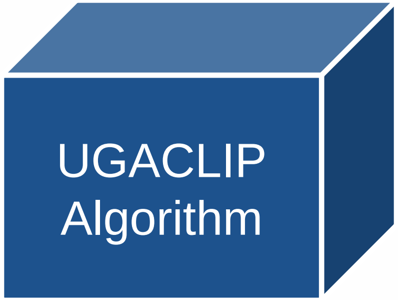
Example of UGACLIP algorithm (k=2)

Privacy Policy
296.00 296.01 296.02

Utility Policy
296.00 296.01

EMR data

ICD	DNA
296.00 296.01 296.02	*CT…A*
295.00 295.01 295.02	*AC…T*
296.00 296.02	GC…C



UGACLIP Algorithm

Anonymized EMR data

ICD	DNA
(296.00, 296.01) 296.02	*CT…A*
295.00 295.01 295.02	*AC…T*
(296.00, 296.01) 296.02	GC…C

Data is protected; {296.00, 296.01, 296.02} appears 2 times

Data remains useful for GWAS on Bipolar disorder; associations between cases and CT…A and controls and CT…A are preserved

Clustering-Based Anonymization (CBA) algorithm

CBA (Sketch)

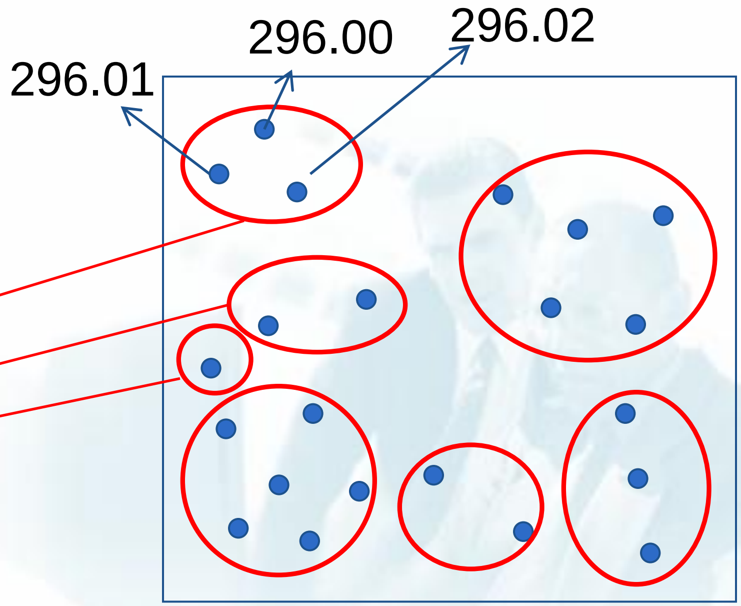
- **Retrieve the ICD codes that need less protection from the Privacy Policy**
  - Gradually build a cluster of codes that can be anonymized according to the utility policy and with minimal UL
- **If the ICD codes are not protected**
  - Suppress no more ICD codes than required to protect privacy

Privacy Policy
296.00 296.01 296.02

Utility Policy
296.00 296.01 296.02

Anonymized EMR data

ICD	DNA
(296.00, 296.01, 296.02)	*CT…A*
(295.00 295.01)	*AC…T*
295.02	GC…C



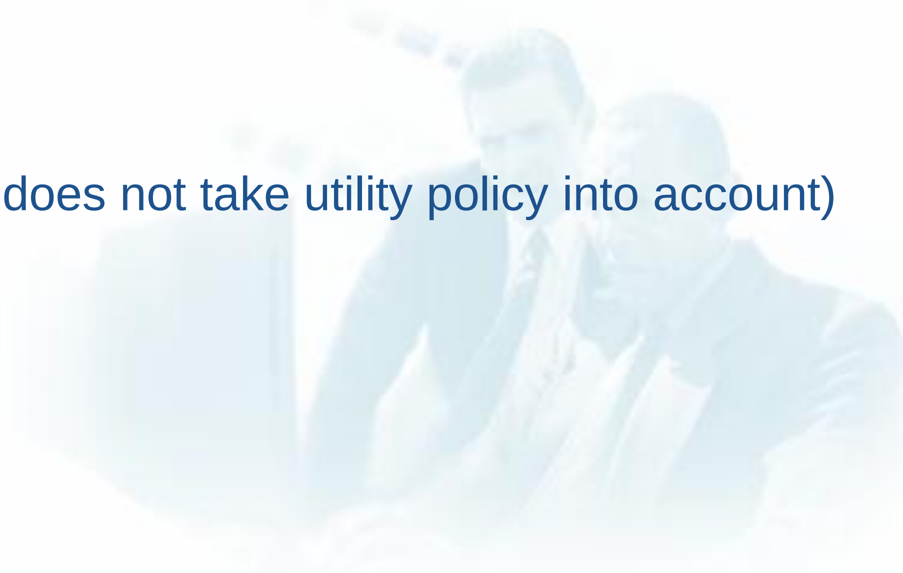
*Loukides et al. Privacy-Preserving publication of diagnosis codes for effective biomedical analysis. IEEE ITAB, 2010.

Case Study: EMRs from Vanderbilt University Medical Center

- **Datasets**
  - VNEC - 2762 de-identified EMRs from Vanderbilt – involved in a GWAS
  - VNECkc - subset of VNEC, we know which diseases are controls for others

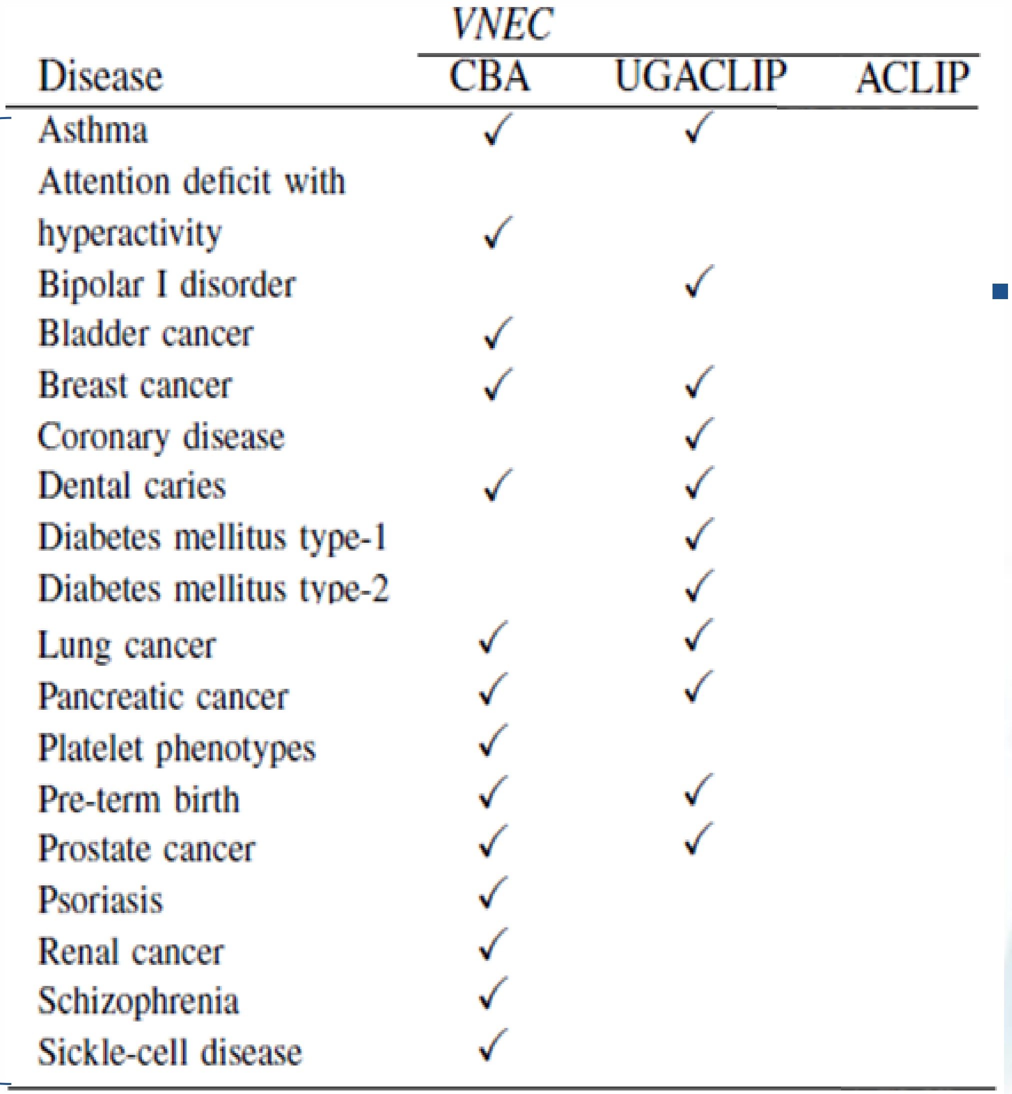
- We have seen that sharing VNEC and VNECkc intact risks identity disclosure and that simple solutions are insufficient

- **Methods**
  - UGACLIP and CBA
  - ACLIP (state-of-the-art method – it does not take utility policy into account)



Case Study: EMRs from Vanderbilt University Medical Center

Diseases related to all GWAS ever conducted*



	VNEC		
Disease	CBA	UGACLIP	ACLIP
Asthma	✓	✓	
Attention deficit with hyperactivity	✓		
Bipolar I disorder		✓	
Bladder cancer	✓		
Breast cancer	✓	✓	
Coronary disease		✓	
Dental caries	✓	✓	
Diabetes mellitus type-1		✓	
Diabetes mellitus type-2		✓	
Lung cancer	✓	✓	
Pancreatic cancer	✓	✓	
Platelet phenotypes	✓		
Pre-term birth	✓	✓	
Prostate cancer	✓	✓	
Psoriasis	✓		
Renal cancer	✓		
Schizophrenia	✓		
Sickle-cell disease	✓		

- Result of ACLIP is useless for validating GWAS

  UGACLIP preserves 11 out of 18 GWAS

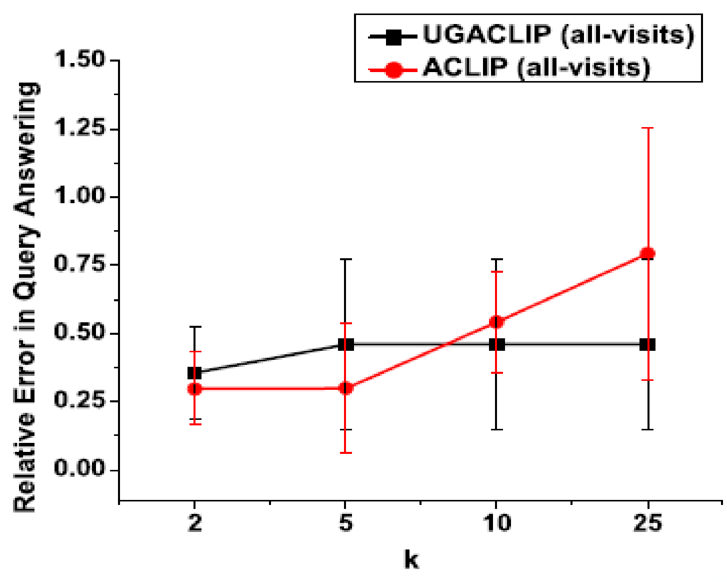
  CBA 14 out of 18 GWAS simultaneously

* Manolio et al. A HapMap harvest of insights into the genetics of common disease. J Clinic. Inv. '08.

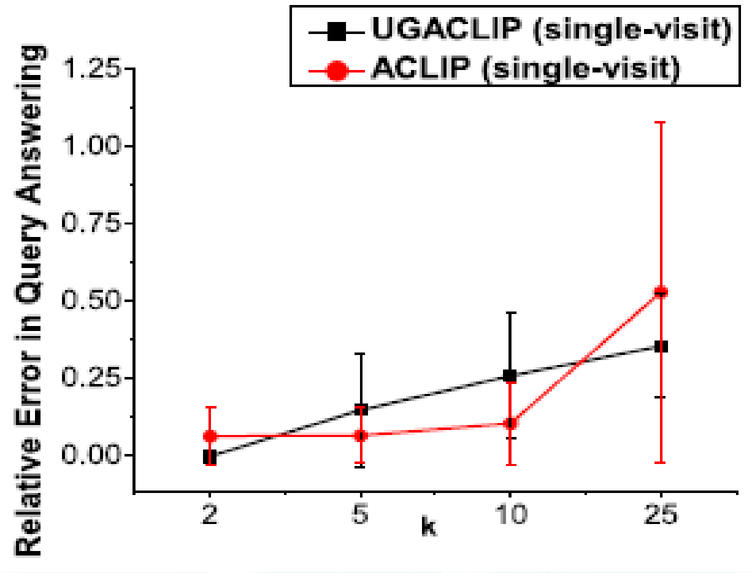
Case Study: EMRs from Vanderbilt University Medical Center

- **Supporting clinical case counts in addition to GWAS**
  - learn number of patients with sets of codes in ≥10% of the records
  - useful for epidemiology and data mining applications

$$\frac{|act. - estim.|}{act.}$$



VNECkc



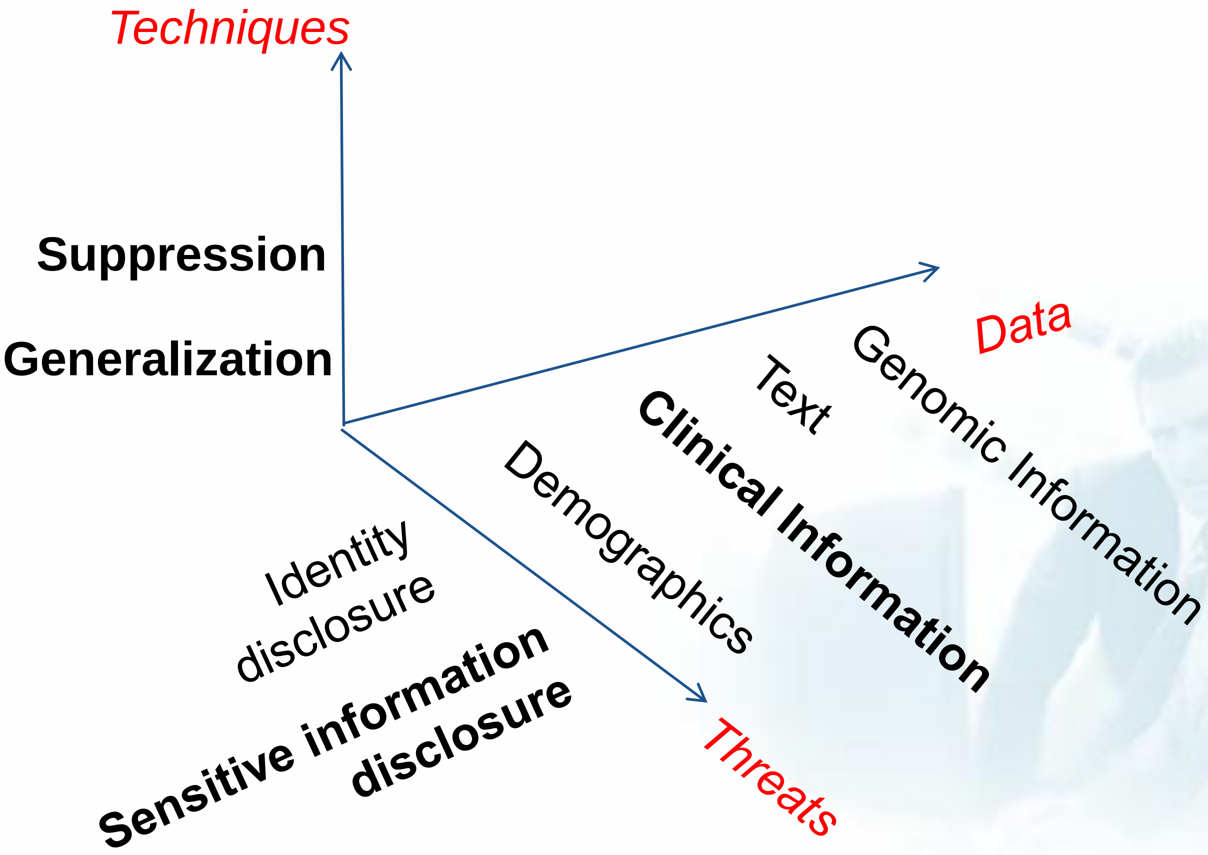
VNECkc

Queries can be estimated accurately (ARE <1.25), comparable to ACLIP

Anonymized data can support both GWAS and studies on clinical case counts

Research challenges and solutions

- **Privacy-preserving data publishing**



Techniques

Suppression

Generalization

Data

Genomic Information

Text

Clinical Information

Demographics

Threats

Identity disclosure

Sensitive information disclosure

Sensitive information disclosure for clinical information sharing

- Certain diagnosis codes are sensitive
  - HIV, Alcohol abuse, etc.

- Preventing identity disclosure may not be sufficient → homogeneity attacks on diagnosis codes

ICD		DNA
401.1 401.1 295		*C…A*
401.0 401.1 295		*A…T*

Schizophrenia

(h,k,p)-coherence

- **(h,k,p)-coherence:** Knowing that an individual is associated with any potentially identifying **p**-itemset, an attacker should not be able to:
  - associate this individual to < ***k*** and >0 transactions, and
  - infer a sensitive item with a probability larger than **1/*h***

ICD	DNA
401.0 401.1	*AC…T*
401.2 401.3 **295**	*GC…C*
401.0 401.1	*CC…A*
401.4	*CA…T*

Original data

ICD	DNA
401.0 401.1	*AC…T*
(401.2, 401.4) 295	*GC…C*
401.0 401.1	*CC…A*
(401.2, 401.4)	*CA…T*

(2,2,2)-coherent data

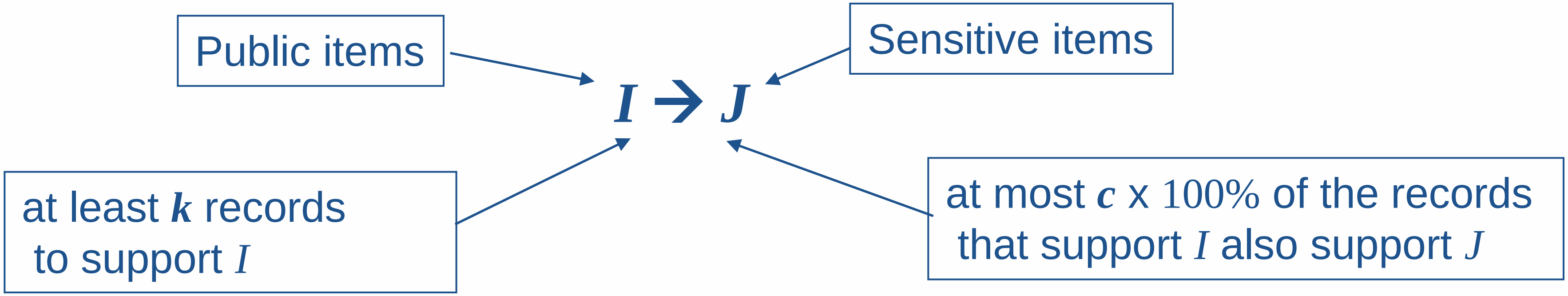
- **Protection from both identity and sensitive information disclosure**
  - ***p* plays the role of m in k^m-anonymity**
- **Enforced through a global suppression algorithm**

* Xu et al. Anonymizing transaction databases for publication. KDD, 2008.

Rule-based Anonymization

- **PS-rules model –** more general than (h,k,p)-coherence

  supports *detailed* privacy requirements



Public items

Sensitive items

I* → *J

at least ***k*** records to support *I*

at most ***c*** x 100% of the records that support *I* also support *J*

(preventing identity disclosure)

(preventing sensitive information disclosure)

ICD	**DNA**
401.0	*AC…T*
(401.2, 401.4) 295	*GC…C*
401.3	*CC…A*
(401.2, 401.4)	*CA…T*

401.2* → *295 is protected for k=2, c=0.5 because (*401.2*,*401.4*) is supported by 2 records and only one of them supports *295*

* Loukides et al. Anonymizing transaction data to eliminate sensitive inferences. DEXA, 2010.

Rule-Based Anonymization of Transactions (RBAT) algorithm

cd → hi, k=5, c=0.2

RBAT (Sketch)

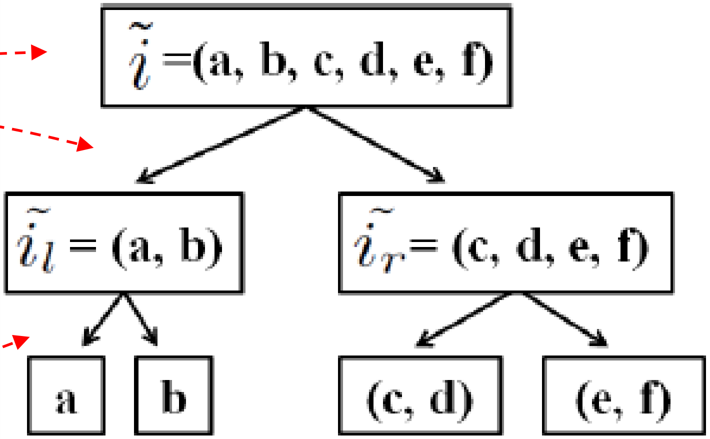
Start with all items generalized into one

Split it into two to enhance data utility
(more specific generalized items)

Check if rules are protected
by computing their support and
confidence in the anonymized dataset

Continue splitting to enhance utility

Return the anonymized dataset



$\tilde{i}$ =(a, b, c, d, e, f)

$\tilde{i}_l$ = (a, b)

$\tilde{i}_r$= (c, d, e, f)

a

b

(c, d)

(e, f)

ICD
a b **(c,d)** **g**
a **(c,d)** (e,f) **h i**
b **(c,d)** **g j**
(e,f) **g h**
a b **(c,d)** (e,f) **j**
(c,d) (e,f) **i**

Other works on anonymizing clinical information

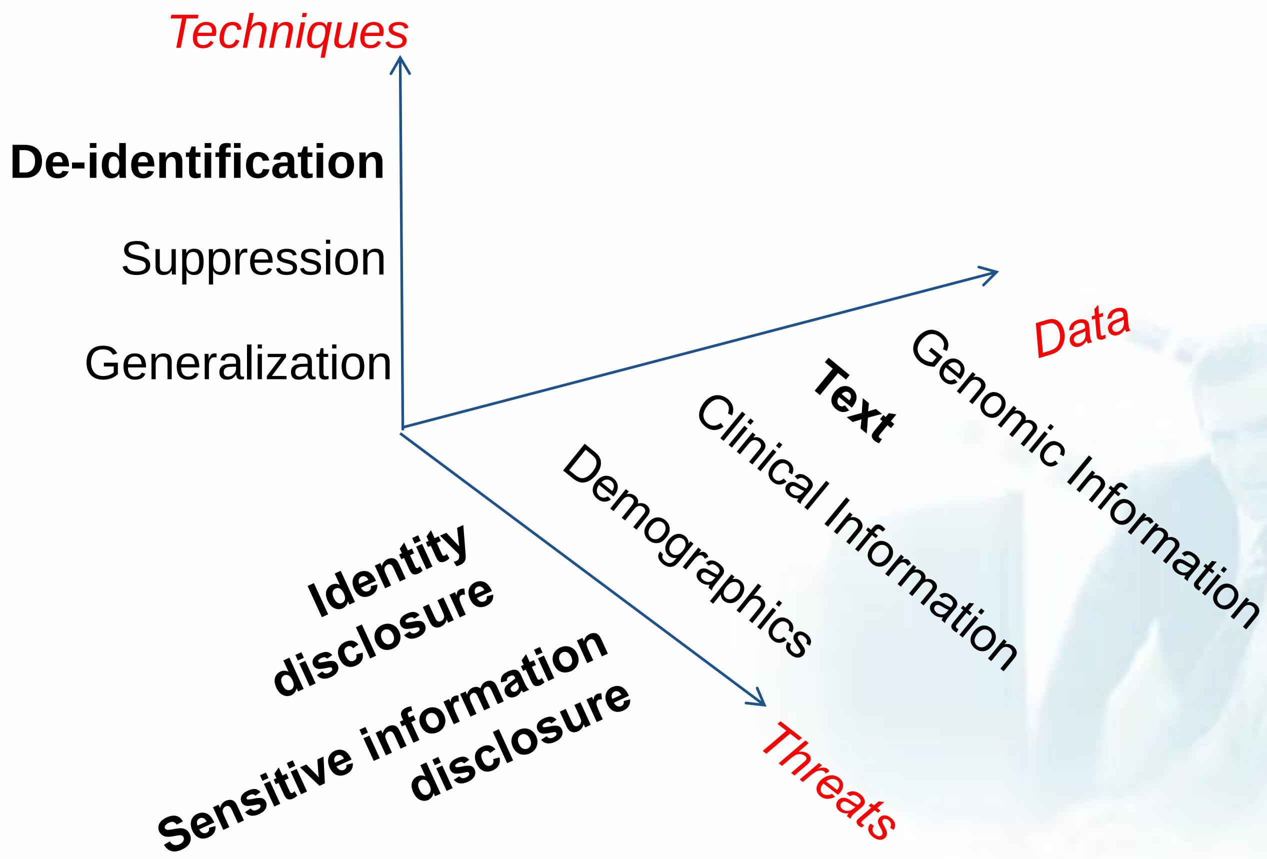
- **ρ-uncertainty[1]**
  - Attackers may use both public and sensitive items to infer sensitive information
  - Limit the probability of inferring any sensitive code
  - Enforced through non-sensitive code generalization and/or sensitive code suppression
  - Does not prevent identity disclosure

- **Other k^m-anonymity algorithms**
  - Local recoding[2]
  - Disassociation[3]

[1] Cao et al. ρ-uncertainty: Inference-Proof Transaction Anonymization. PVLDB, 2010.
[2] Terrovitis et al. Local and Global Recoding Methods for Anonymizing Set-valued Data. VLDBJ, 2010.
[3] Terrovitis et al. Privacy Preservation by Disassociation. TR-IMIS-2010-1, 2010.

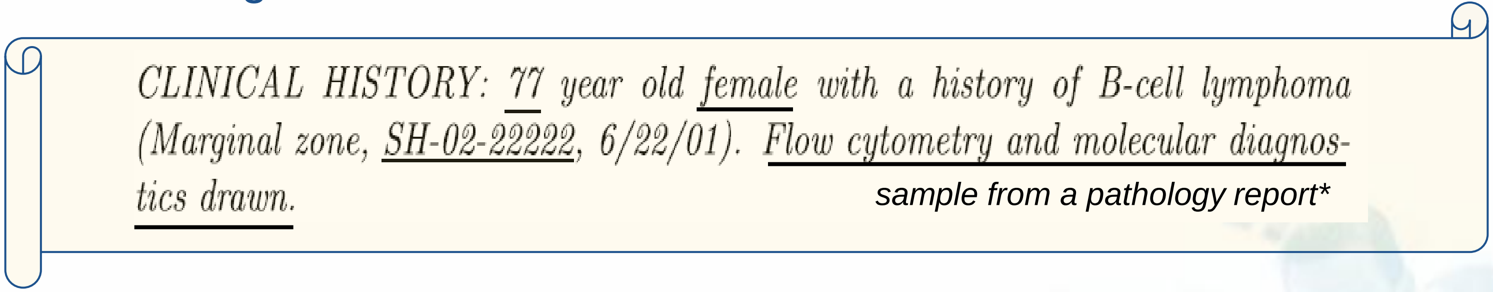
Research challenges and solutions

- **Privacy-preserving data publishing**



Clinical text de-identification

- **EMRs contain a considerable amount of unstructured data**
  - Clinical notes
  - SOAP (Subjective, Objective, Assessment, Patient care plan) notes
  - Radiology and pathology reports
  - Discharge summaries



CLINICAL HISTORY: 77 year old female with a history of B-cell lymphoma (Marginal zone, SH-02-22222, 6/22/01). Flow cytometry and molecular diagnostics drawn.

*sample from a pathology report**

- **Clinical text de-identification is a 2-step process**
  - Detect personal identifiers (e.g., name, record#, SSN)
  - Replace or remove the discovered personal identifiers
- **Goal: integrity of medical information remains intact while personal identity is effectively concealed**

* Xiong et al. Privacy-Preserving Information Discovery on EHRs. *Information Discovery on Electronic Health Records*, 2008.

Detecting personal identifiers

- **Named Entity Recognition (NER)**
  - Locate atomic elements in text (HIPAA-compliant personal identifiers)
  - Classify elements into pre-defined categories (e.g., name, address, phone)

- **Grammar-based or Rule-based approaches**
  - Hand-coded rules and dictionaries (e.g., common names)
  - Regular expressions for identifiers that follow a syntactic pattern (e.g., phones, zip codes)

- **Statistical learning approaches**
  - Rely on manually annotated training data with pre-labeled identifiers
  - Build a classifier to classify the terms of previously unseen (test) data as *identifier* or *non-identifier*
  - Feature sets: terms, local/global context, dictionary-related features
  - Techniques: Maximum Entropy model, HMMs, SVMs, etc.

Scrub system

- **Rule-based and dictionary-based system**
- **Detection strategy**
  - Several detection algorithms
  - Aim to recognize specific entities by using rules and lists
  - Operate in parallel to label entities as names, addresses, dates, etc.
  - Share results and compete based on the certainty of their findings
  - The algorithm with highest certainty prevails
- **Replacement strategy**
  - Associated with each detection algorithm is a replacement algorithm
  - Consistent replacement for names, cities, etc.; lumping for dates
- **Evaluation**
  - pediatric medical records: 275 patients; 3198 letters to referring physicians
  - 99-100% of personally identifying information was reported to be detected

* L. Sweeney. Replacing personally-identifying information in medical records, the Scrub system, JAMIA, 1996.

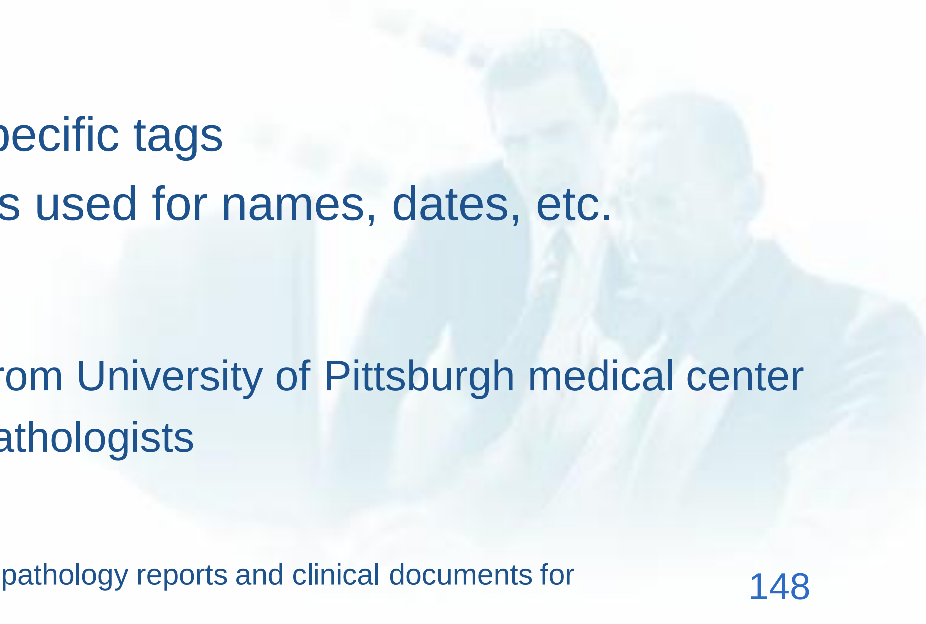
DE-ID system

- **Rule-based and dictionary-based software (DE-ID Data Corp 2004)**
- **Works with archives of several types of clinical documents**
- **Supports the 17 HIPAA-specified ids (excl. photo) + more**

- **Detection strategy**
  - Uses rules and dictionaries to identify patient and provider names
  - Uses the UMLS database to identify medical phrases
  - Uses pattern matching to detect phone numbers and zip codes

- **Replacement strategy**
  - Identifying terms are replaced by specific tags
  - A consistent replacement strategy is used for names, dates, etc.

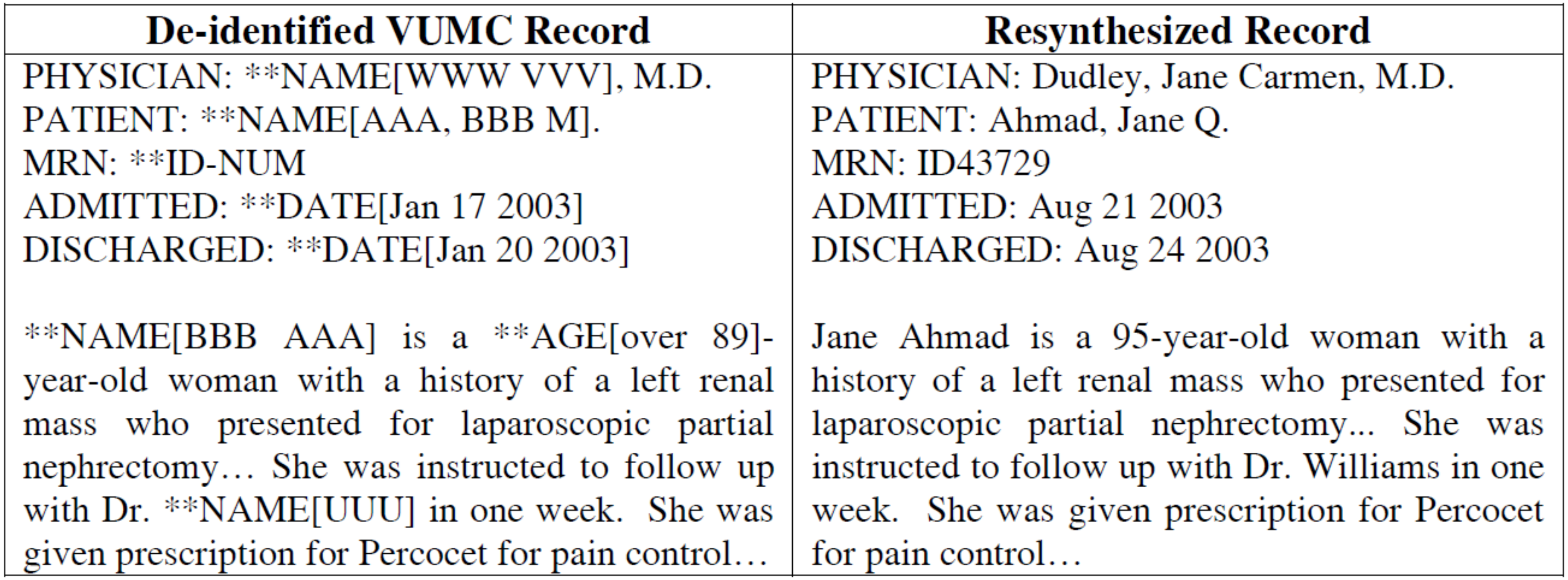
- **Evaluation**
  - Datasets of surgical pathology reports from University of Pittsburgh medical center
  - DE-ID reports were evaluated by four pathologists
  - No precision or recall were reported



* D. Gupta, et al., Evaluation of a de-identification software engine to share pathology reports and clinical documents for research, *American Journal of Clinical Pathology*, 2004.

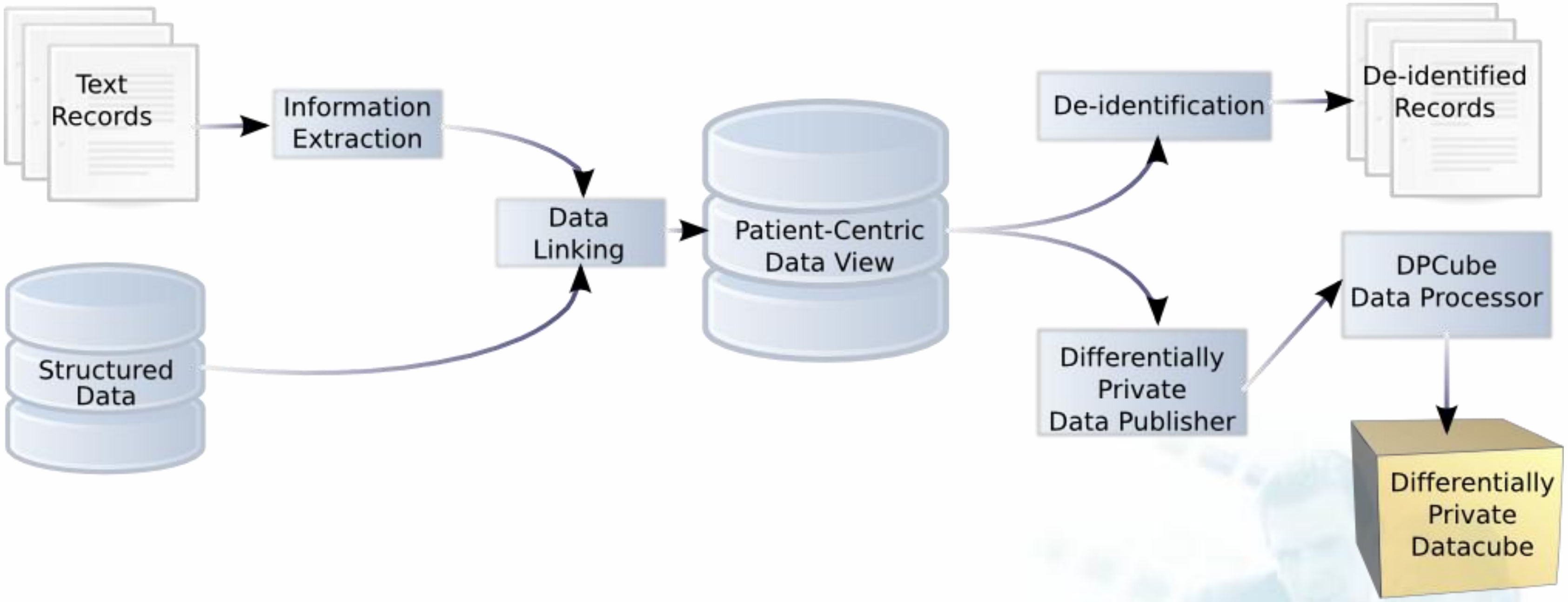
DE-ID system

Example of a clinical report that was de-identified using DE-ID



De-identified VUMC Record	Resynthesized Record
PHYSICIAN: **NAME[WWW VVV], M.D. PATIENT: **NAME[AAA, BBB M]. MRN: **ID-NUM ADMITTED: **DATE[Jan 17 2003] DISCHARGED: **DATE[Jan 20 2003] **NAME[BBB AAA] is a **AGE[over 89]-year-old woman with a history of a left renal mass who presented for laparoscopic partial nephrectomy… She was instructed to follow up with Dr. **NAME[UUU] in one week. She was given prescription for Percocet for pain control…	PHYSICIAN: Dudley, Jane Carmen, M.D. PATIENT: Ahmad, Jane Q. MRN: ID43729 ADMITTED: Aug 21 2003 DISCHARGED: Aug 24 2003 Jane Ahmad is a 95-year-old woman with a history of a left renal mass who presented for laparoscopic partial nephrectomy... She was instructed to follow up with Dr. Williams in one week. She was given prescription for Percocet for pain control…

Health Information DEidentification (HIDE)



*“a configurable, integrated framework for publishing and sharing health data while preserving data privacy”***

* L. Xiong et al. Privacy-Preserving Information Discovery on EHRs. *Information Discovery on Electronic Health Records*, 2008.
** http://www.mathcs.emory.edu/hide/ (open-source software, Emory University)

HIDE: Text de-identification

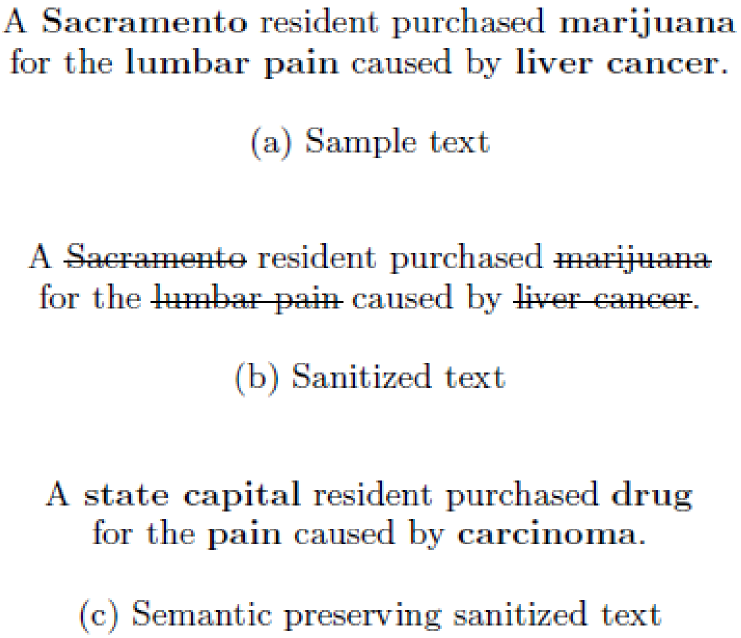
- **Open source system using statistical learning for text de-id**
- **Detection strategy:** iterative process for classifying + retagging
  - A tagging interface allows users to annotate medical data with identifying attributes to build the training set
  - A feature generation component extracts the features from text to build a Conditional Random Field (CRF) classifier
  - The CRF classifier is employed to classify terms into multiple classes
  - Data post-processing strategies are used to feed the classified data back to the tagging software for retagging and corrections
- **Replacement strategy**
  - Suppression or term generalization
- **Evaluation**
  - Dataset of pathology reports: 100 reports
  - Precision and recall are reported to be ~ 97%

* J. Gardner et al. An integrated framework for anonymizing unstructured medical data. *DKE*, 2009.

t-plausibility

- Generalizes sensitive terms to semantically related terms (e.g., “tuberculosis” → “infectious disease”)

- t-plausibility*: Given word ontologies and a threshold t, the sanitized text can be associated with at least t texts; any of them could be the original text



A **Sacramento** resident purchased **marijuana** for the **lumbar pain** caused by **liver cancer**.

(a) Sample text

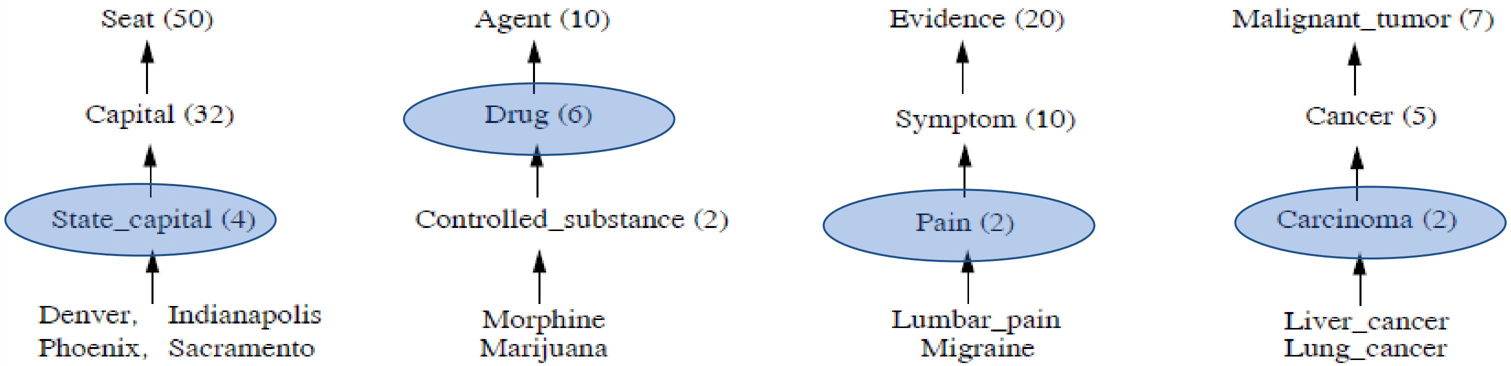
A ~~Sacramento~~ resident purchased ~~marijuana~~ for the ~~lumbar pain~~ caused by ~~liver cancer~~.

(b) Sanitized text

A **state capital** resident purchased **drug** for the **pain** caused by **carcinoma**.

(c) Semantic preserving sanitized text

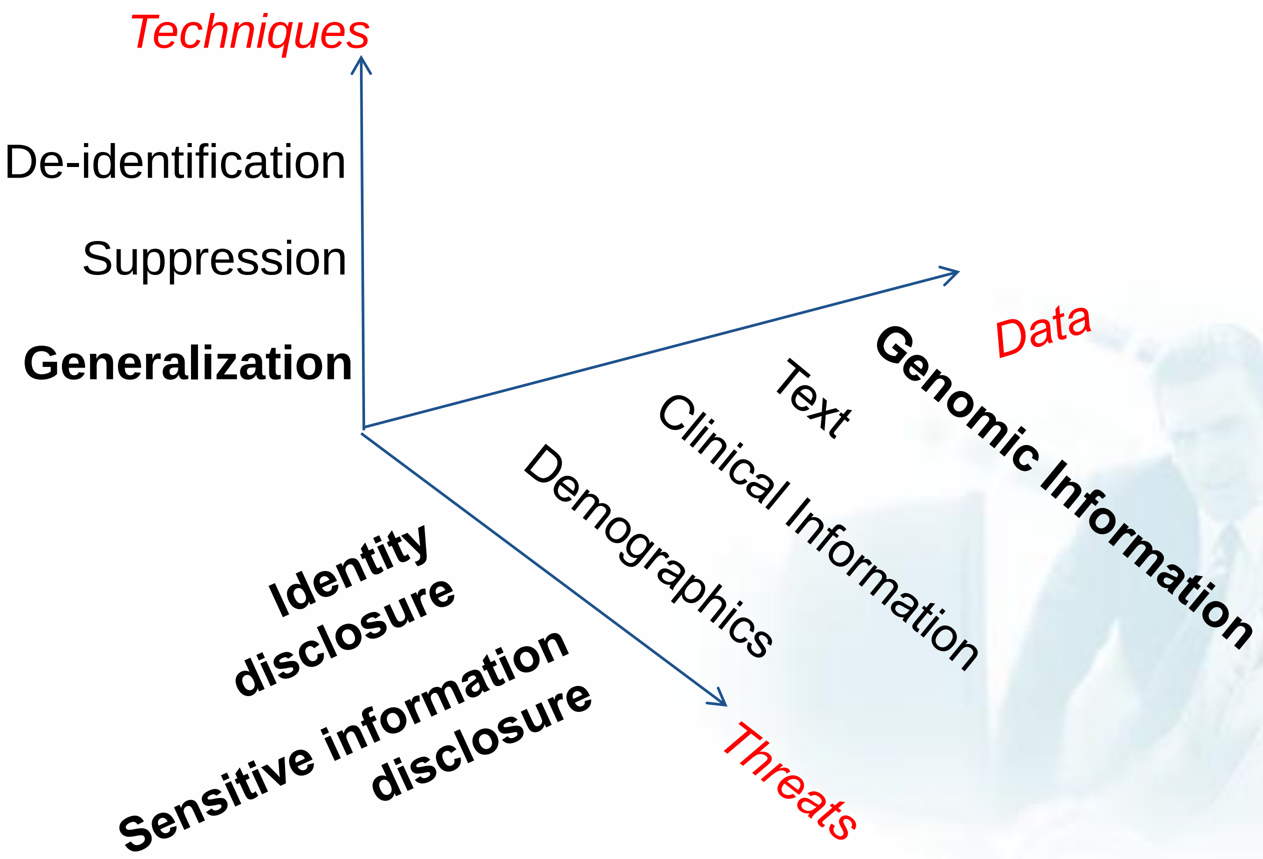
D can be associated with 96 texts



* Jiang et al. t-Plausibility: Semantic Preserving Text Sanitization. CSE, 2009.

Research challenges and solutions

- **Privacy-preserving data publishing**



Techniques

De-identification

Suppression

Generalization

Data

Genomic Information

Text

Clinical Information

Demographics

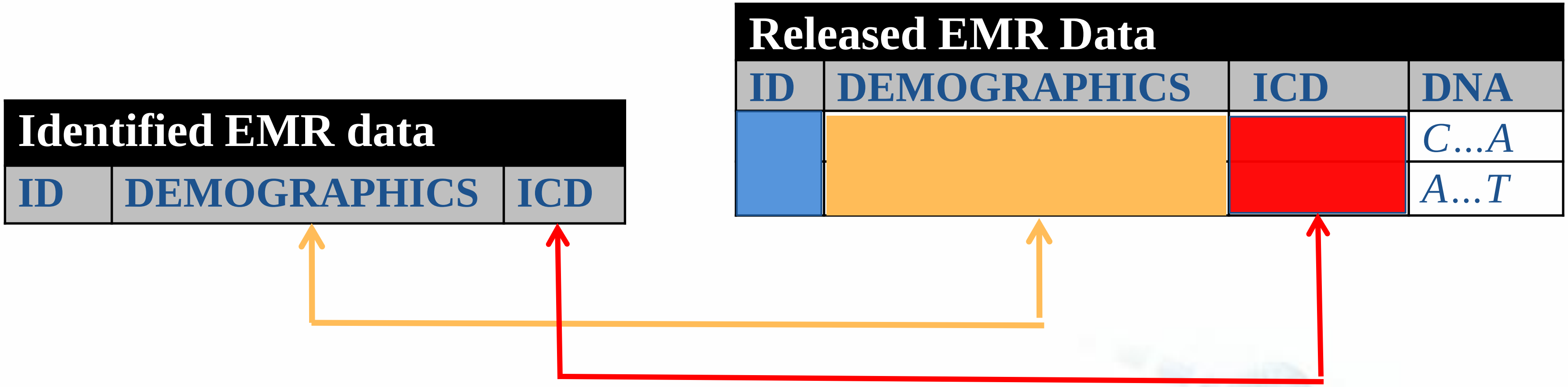
Threats

Identity disclosure

Sensitive information disclosure

DNA sequence privacy

- So far, we showed how to prevent two linkages

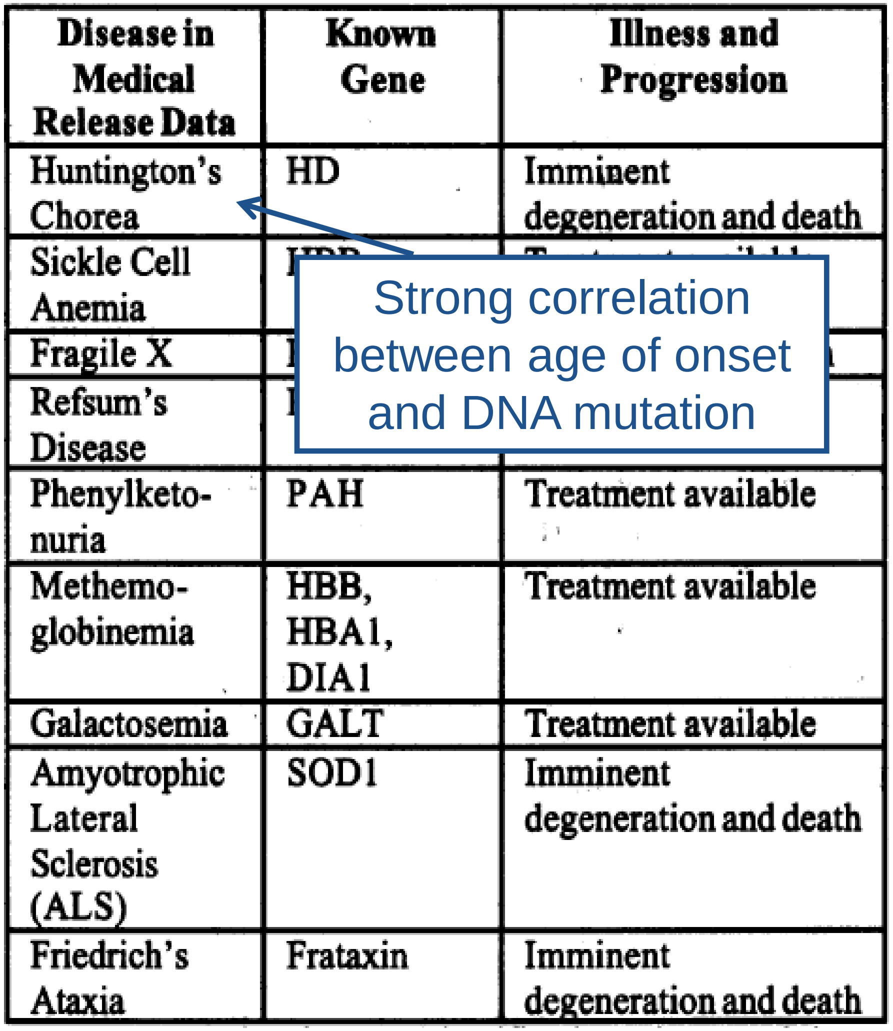


Identified EMR data		
ID	DEMOGRAPHICS	ICD

Released EMR Data			
ID	DEMOGRAPHICS	ICD	DNA
🟦	🟧	🟥	*C…A*
			A…T

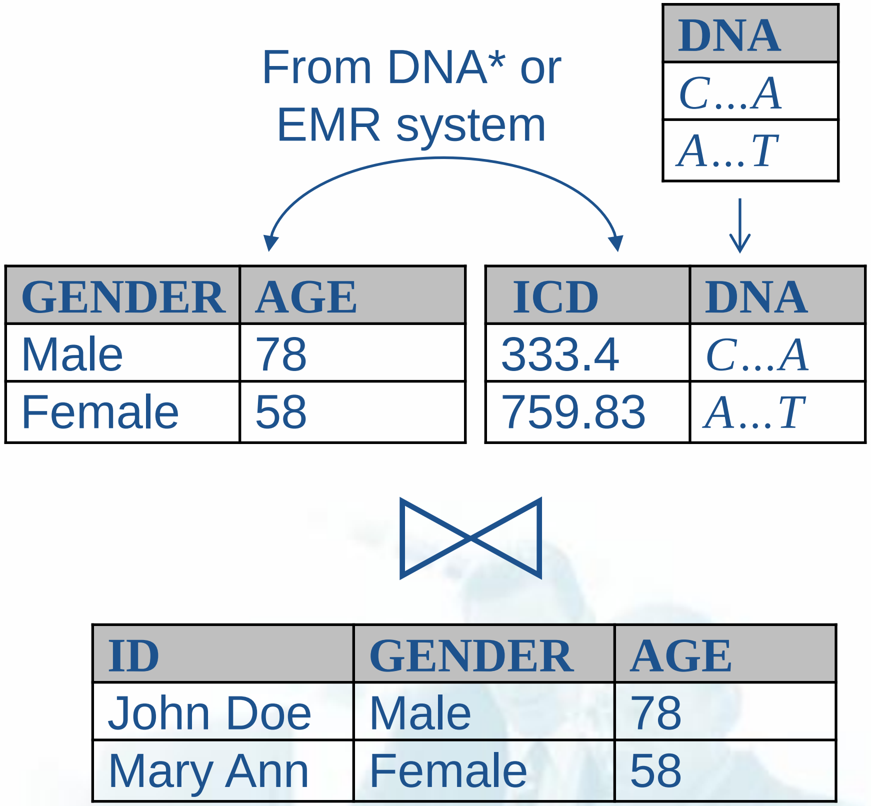
What if DNA sequences themselves reveal sensitive information?

Genotype-Phenotype attack



Disease in Medical Release Data	Known Gene	Illness and Progression
Huntington's Chorea	HD	Imminent degeneration and death
Sickle Cell Anemia	HBB	Treatment available
Fragile X	[illegible]	[illegible]
Refsum's Disease	[illegible]	[illegible]
Phenylketo-nuria	PAH	Treatment available
Methemo-globinemia	HBB, HBA1, DIA1	Treatment available
Galactosemia	GALT	Treatment available
Amyotrophic Lateral Sclerosis (ALS)	SOD1	Imminent degeneration and death
Friedrich's Ataxia	Frataxin	Imminent degeneration and death

Strong correlation between age of onset and DNA mutation



From DNA* or EMR system

DNA
C…A
A…T

GENDER	AGE
Male	78
Female	58

ICD	DNA
333.4	C…A
759.83	A…T

$\bowtie$

ID	GENDER	AGE
John Doe	Male	78
Mary Ann	Female	58

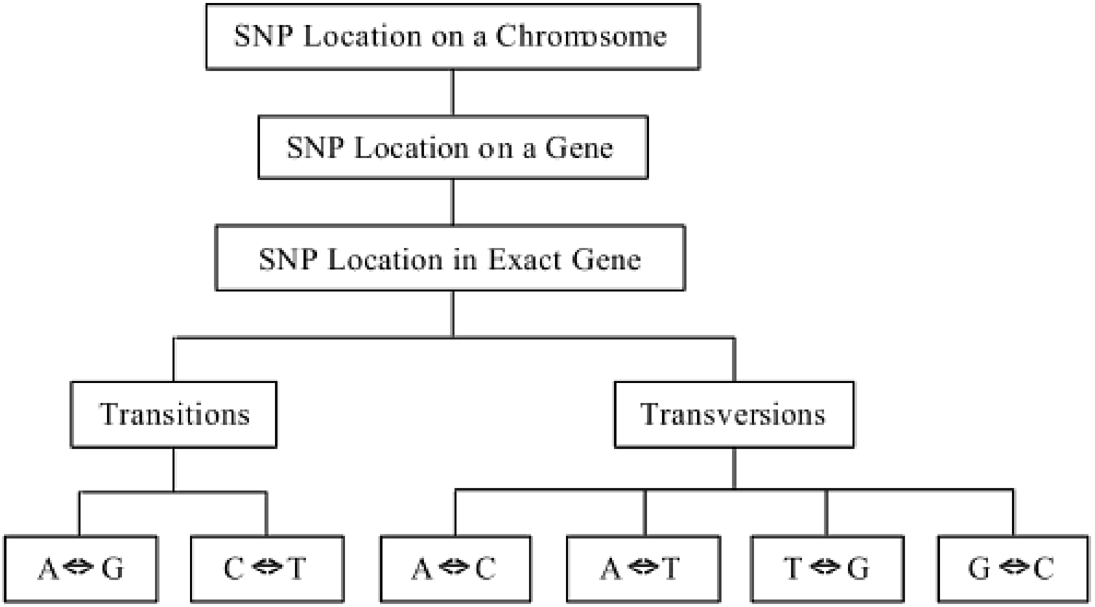
From Voter lists or EMR system

* Malin et al. Determining the Identifiabiity of DNA Database Entries. AMIA, 2000.

Binning

- **Main idea*:** Apply a two-step generalization on SNPs using a hierarchy-based model so that
  - at least ***B*** SNPs in a genomic sequence have the same value
  - at least ***B'*** genomic sequences have the same value for a specific set of SNPs.

- **Generalization hierarchy**



SNP Location on a Chromosome

SNP Location on a Gene

SNP Location in Exact Gene

Transitions

Transversions

A⇔G

C⇔T

A⇔C

A⇔T

T⇔G

G⇔C

* Lin et al. Using binning to maintain confidentiality of medical data. AMIA, 2002.

Binning strategies

- **To generalize SNPs in a genomic sequence**
  - Bottom-up search using the generalization hierarchy
    - nodes are generalized to their closest ancestors one by one

    until at least ***B*** SNPs have the same value

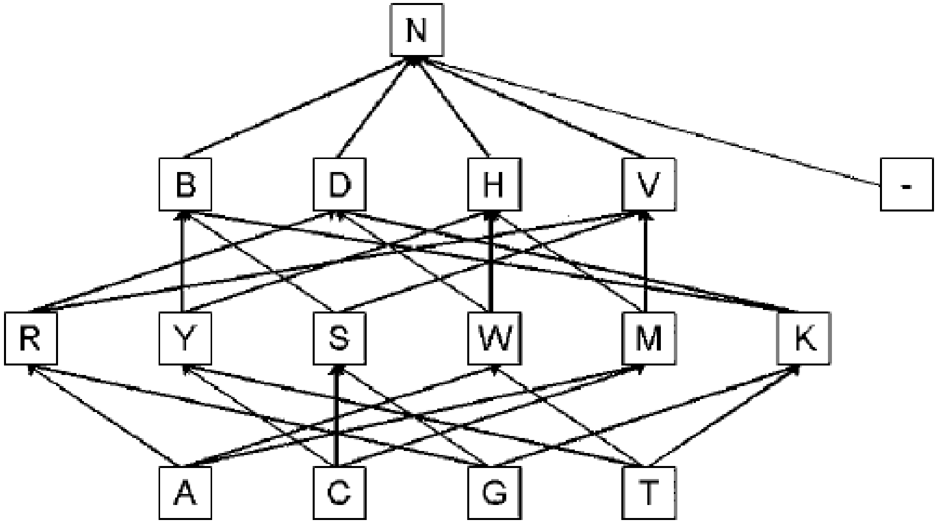
- **To generalize different SNPs of different genomic sequences**
  - Consider all combinations of SNPs one by one
    - starting with the one that is the least represented in the data

    until at least ***B’*** sequences are indistinguishable w.r.t. the SNPs

- ***B* and *B’* are bin size parameters to control the utility/privacy trade-off**
  - similar to ***k*** in *k*-anonymity

DNA lattice generalization

- **The DNA Lattice generalization method* attempts to reduce information loss by**
  - Using a lattice (the union of all possible trees for single nucleotide hierarchies) instead of a generalization hierarchy to represent a larger number of generalizations



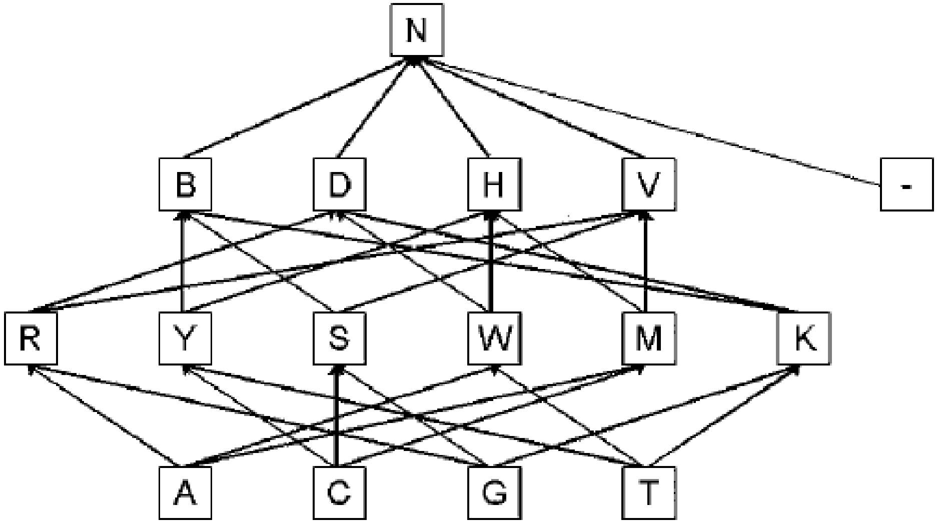
A: Adenine	**C**: Cytosine
G : Guanine	**T**: Thymine
R : Purine	**Y**: Pyrimadine
S : Strong hydrogen	**W**: Weak hydrogen
M : Amino group	**K**: Keto group
B : not A	**D**: not C
H : not G	**V**: not T
- : gap	**N**: Indeterminate

* Malin. Protecting DNA Sequence Anonymity with Generalization Lattices. Methods of Information in Medicine, 2005.

DNA lattice generalization

- **The DNA Lattice generalization method attempts to reduce information loss by**
  - Employing a distance measure based on the level of hierarchy to measure distance between two bases *x* and *y* generalized to *z*

$$d(x,y) = 2 \times level(z) - level(x) - level(y)$$

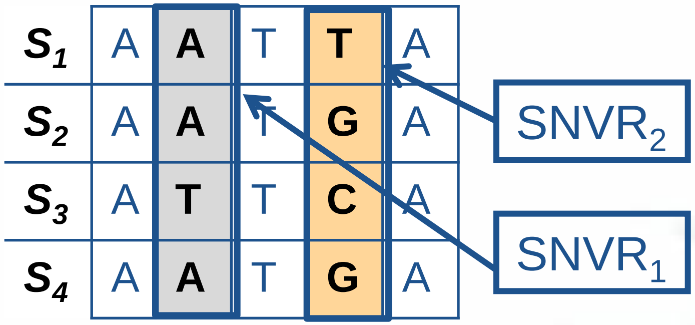


A: Adenine	**C**: Cytosine
G : Guanine	**T**: Thymine
R : Purine	**Y**: Pyrimadine
S : Strong hydrogen	**W**: Weak hydrogen
M : Amino group	**K**: Keto group
B : not A	**D**: not C
H : not G	**V**: not T
- : gap	**N**: Indeterminate

DNA lattice generalization algorithm

DNALA (Sketch)

- Identify Single Nucleotide Variable Regions (positions in which at least one sequence has a different value than another sequence) based on a sequence alignment algorithm



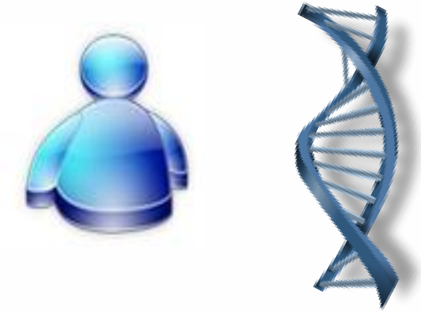
S_1	A	**A**	T	**T**	A
S_2	A	**A**	T	**G**	A
S_3	A	**T**	T	**C**	A
S_4	A	**A**	T	**G**	A

- Pair each sequence with its “closest” according to the sum of generalization distances between the set of SNVRs
- For each pair of sequences
  - Remove the gaps inserted during sequence alignment
  - Generalize according to the lattice

Homer’s attack

- **Homer’s attack*:** Infer whether an individual is in a complex genomic DNA mixture

Individual’s identity and DNA



Mixture DNA // (Similar) Population DNA

- Measure the difference between the distance of the individual from the mixture and the distance of the individual from the Population
  - Is individual most likely to be Case for a GWAS-related disease?
  - Is individual most likely to be Control …?
  - Is individual equally likely to be Case or Control … ?

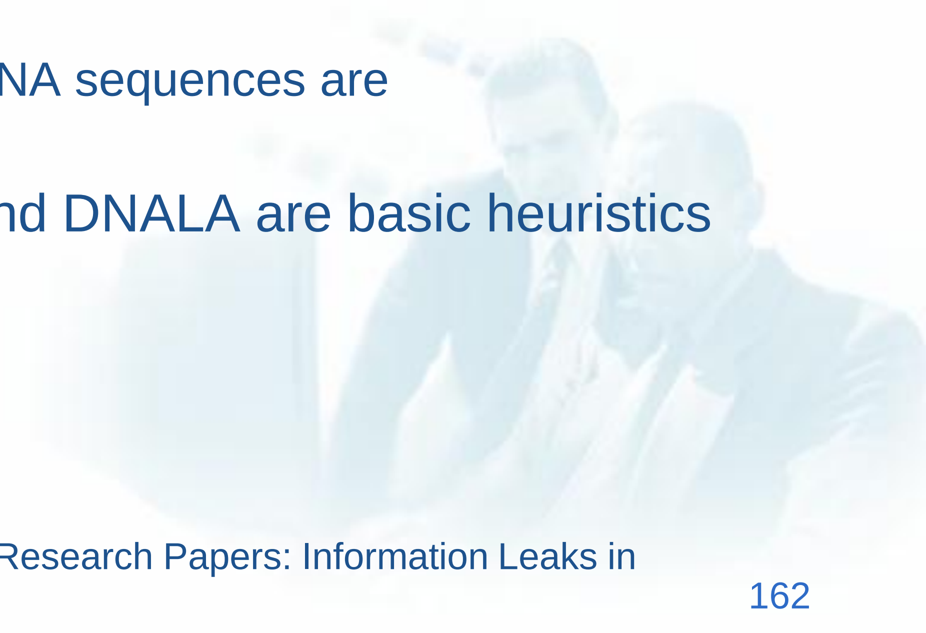
* Homer et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLOS Genetics, 2008.

DNA privacy issues

- **Privacy issues – are these threats real?**
  - **Availability of DNA is currently limited**
    - GWAS data in dbGaP is accessible only to Pis
  - **Attacks**
    - **complex– not just joins**
    - **more predictive than Homer's attack***

- **Utility issues**
  - DNA has complex semantics
  - Unclear how useful generalized DNA sequences are

- **Algorithmic issues -** binning and DNALA are basic heuristics
    - no utility guarantees
    - ad-hoc objective measures
    - inefficient



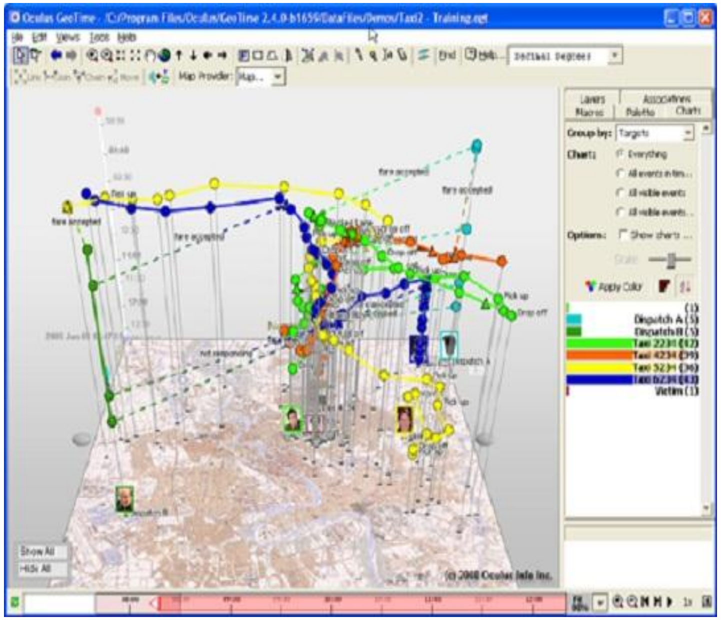
[*] Wang et al. Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study, CCS, 2009.

Content

- Part 1: Medical data sharing and the need for data privacy
- Part 2: Challenges and state-of-the-art solutions
- **Part 3: Open problems and research directions**

Complex data sharing

- **Medical data are inherently complex**



- **different types of data**
  - demographics, clinical notes, lab values, images, spatiotemporal information, etc.
- **lack of universal medical classification schemes**
  - ICD-9 vs. ICD-10 etc.
- **various forms of attacks that must be prevented while maintaining utility**
  - inferential and membership disclosures, etc.

... but most work focuses on simple data types and prevents a simple attack without offering utility guarantees

Large-scale, distributed data sharing

- **Medical data are provided by and shared with many parties**
  - **Health information exchange**
    - UK NHS reconsidered plans to build a centralized electronic medical record system because of privacy* and data management concerns**
  - **Collaborative research efforts**
    - Biobanks, medical data repositories

- **Lots of data, stored or processed, also remotely**

… but most work focuses on

a static dataset that can be processed in main memory

* Anderson. Undermining data privacy in health information, BMJ, 2001

** Zhang et al. A role-based delegation framework for healthcare information systems, SACMAT, 2002.

Summary

- Medical data sharing and the need for data privacy
- Research challenges and solutions for different types of data
- Open problems and research directions

Acknowledgements

- **Joshua Denny –** VANDERBILT
- **Hariklia Eleftherohorinou –** Imperial College London
- **Efi Kokiopoulou –** Google Research
- **Jianhua Shao –** CARDIFF UNIVERSITY
- **Michail Vlachos –** IBM

Thank you!
Questions?

References

1. National Ambulatory Medical Care Survey, National Center for Health Statistics, 2010.
2. J. A. Pacheco et al. A Highly Specific Algorithm for Identifying Asthma Cases and Controls for Genome-Wide Association Studies. AMIA Annual Symposium ’09.
3. Centers for Medicare & Medicaid Services - https://www.cms.gov/icd9providerdiagnosticcodes/
4. M. J. Tildesley et al. Impact of spatial clustering on disease transmission and optimal control, Proceedings of the National Academy of Sciences, 2010.
5. B. Reis, I. S. Kohance, and K. D. Mandl. Longitudinal Histories as Predictors of Future Diagnoses of Domestic Abuse: Modelling Study, BMJ: British Medical Journal, 2011.
6. Y.M. Chae et al. Data mining approach to policy analysis in a health insurance domain. International Journal of Medical Infprmatics, 2001.
7. A. D. Johnson and C. J. O’Donnell. An open access database of genome-wide association results". BMC Medical Genetics, 2009.
8. T. A. Manolio and F. S. Collins. A HapMap harvest of insights into the genetics of common disease. Journal of Clinical Investigation, 2008.
9. Health Confidence Survey 2008, Employee Benefit Research Institute
10. E. J. Ludman et al. Glad You Asked: Participants’ Opinions of Re-Consent for dbGap Data Submission. Journal of Empirical Research on Human Research Ethics, 2010.
11. Ponema Institute/Symantec corporation, 2010 Annual Study: US cost of a data breach.
12. M. Barbaro and T. Zeller. A face exposed for AOL searcher no. 4417749. NY Times. Aug 9, 2006.
13. G. Loukides, J. C. Denny and B. Malin. The Disclosure of Diagnosis Codes Can Breach Research Participants’ Privacy. Journal of the American Medical Informatics Association, 2010.
14. L. Sweeney, k-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 2002.
15. A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy, 2008.

References

16. A. Gkoulalas-Divanis and G. Loukides. Revisiting sequential pattern hiding to enhance utility. ACM SIGKDD International Conference on Knowledge Discovery and Data Engineering, 2011.
17. G. Das and N. Zhang, Privacy risks in health databases from aggregate disclosure. International Conference on Pervasive Technologies Related to Assistive Environments, 2009.
18. M. Grean and M. J. Shaw. Supply chain partnership between P&G and Wal-Mart. Chapter 3, Integrated Series in Information Systems. 2002.
19. National Institutes of Health, Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies. 2007.
20. K. Benitez and B. Malin. Evaluating re-identification risks with respect to the HIPAA privacy rule, Journal of the American Medical Informatics Association, 2010.
21. T. Li and N. Li. Injector: Mining Background Knowledge for Data Anonymization. International Conference on Data Engineering, 2008.
22. G. T. Duncan, S. A. Keller-McNulty, and S. L. Stokes, Disclosure Risk vs. Data Utility: The R-U Confidentiality map. Technical Report LA-UR-01-6428. Los Alamos National Library, 2001.
23. R. Agrawal and R. Srikant. Privacy-preserving data mining. SIGMOD, 2000.
24. A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2002.
25. H. Polat and W. Du. SVD-based collaborative filtering with privacy, ACM SAC, 2005.
26. H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. On the privacy preserving properties of random data perturbation techniques, IEEE International Conference on Data Mining, 2003.
27. C. C. Aggarwal. On Randomization, Public Information and the Curse of Dimensionality. IEEE International Conference on Data Engineering, 2007.
28. S. A. Vinterbo, L. Ohno-Machado, and S. Dreiseitl. Hiding information by cell suppression. AMIA Annual Symposium, 2001.
29. G. Loukides, A. Gkoulalas-Divanis, and B. Malin. Anonymization of Electronic Medical Records for Validating Genome-Wide Association Studies. Proceedings of the National Academy of Sciences, 2010.

References

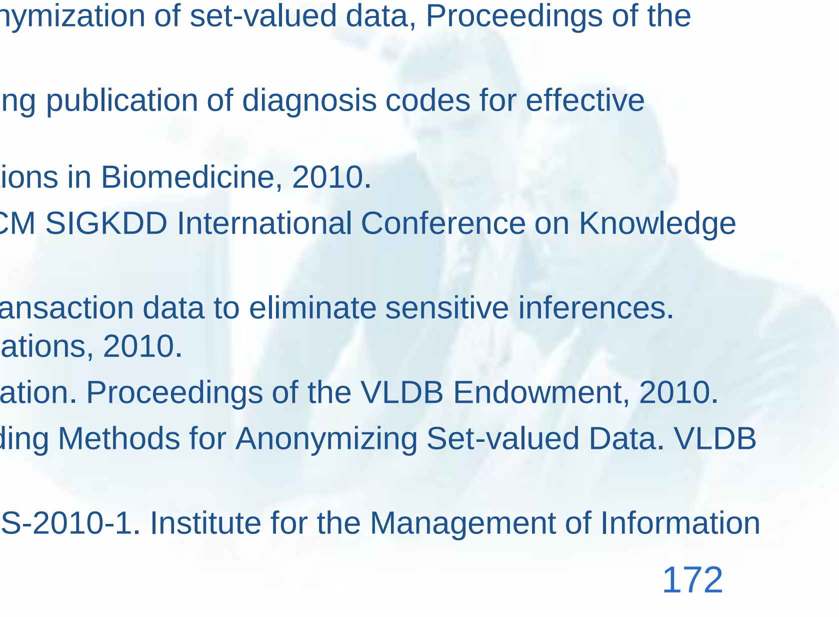
30. L. Sweeney, Computational Disclosure Control: Theory and Practice. . Massachusetts Institute of Technology, Laboratory for Computer Science, Tech Report, PhD Thesis. 2001.
31. A. Gionis, A. Mazza, and T. Tassa. k-Anonymization Revisited. International Conference on Data Engineering, 2008.
32. A. Machanavajjhala et al. l-diversity: Privacy beyond k-anonymity. International Conference on Data Engineering, 2006.
33. R. C. Wong et al., (alpha, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing, ACM SIGKDD International Conference on Knowledge Discovery and Data mining 2006.
34. N. Li , T. Li, and V. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity, International Conference on Data Engineering, 2007.
35. J. Li. Y. Tao, and X. Xiao. Preservation of proximity privacy in publishing numerical sensitive data. ACM SIGMOD International Conference on Management of Data, 2008.
36. G. Loukides and J. Shao. Preventing range disclosure in k-anonymised data. Expert Systems with Applications: An International Journal, 2011.
37. X. Xiao and Y. Tao, Personalized privacy preservation. ACM SIGMOD International Conference on Management of Data, 2006.
38. A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2003.
39. Y. Tao, X. Xiao, J. Li, and D. Zhang. On anti-corruption privacy preserving publication. IEEE International Conference on Data Engineering, 2008.
40. R. Chaytor and K. Wang. Small domain randomization: same privacy, more utility. Proceedings of the VLDB Endowment, 2010.
41. C. Dwork. Differential privacy. International Colloquium on Automata, Languages, and Programming. 2006.
42. N. Mohammed, R. Chen, B.C.M. Fung, and P.S. Yu. Differentially private data release for data mining. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011.
43. C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference, 2006.

References

44. F. McSherry, K. Talwar. Mechanism design via differential privacy. IEEE Symposium on Foundations of Computer Science, 2007.
45. S. R. Ganta and S. P. Kasiviswanathan, and A. Smith. Composition attacks and auxiliary information in data privacy. ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2008.
46. C. Dwork. Differential privacy: a survey of results. International Conference on Theory and Applications of Models of Computation. 2008.
47. X. Xiao, G. Wang, and J. Gehrke. Differential privacy via wavelet transforms. IEEE International Conference on Data Engineering, 2010.
48. A. Machanavajjhala, J. Gehrke, and M. Gotz. Data privacy against realistic adversaries. Proceedings of the VLDB Endowment, 2009.
49. B. Ding, M. Winslett, J. Han, Z. Li. Differentially private data cubes: optimizing noise sources and consistency. ACM SIGMOD International Conference on Management of Data, 2011.
50. D. Kifer, A. Machanavajjhala. No free lunch in data privacy. ACM SIGMOD International Conference on Management of Data, 2011.
51. G. Cormode. Personal privacy vs population privacy: learning to attack anonymization. ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2011.
52. J. Li. Y. Tao, and X. Xiao. Preservation of proximity privacy in publishing numerical sensitive data. ACM SIGMOD International Conference on Management of Data, 2008.
53. G. Loukides and J. Shao. Preventing range disclosure in k-anonymised data. Expert Systems with Applications: An International Journal, 2011.
54. X. Xiao and Y. Tao, Personalized privacy preservation. ACM SIGMOD International Conference on Management of Data, 2006.
55. K. LeFevre. D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity, International Conference on Data Engineering, 2006.
56. T. Iwuchukwu and J. F. Naughton. K-anonymization as spatial indexing: toward scalable and incremental anonymization, International Conference on Very Large Databases, 2007.

References

57. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware anonymization. ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2006.
58. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware anonymization techniques for large-scale datasets. ACM Transactions on Database Systems, 2008.
59. J. Xu et al. Utility-Based Anonymization Using Local Recoding, ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2006.
60. G. Aggarwal et al. Achieving anonymity via clustering. ACM Transactions on Algorithms, 2010.
61. C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998.
62. C. C. Aggarwal. On k-anonymity and the curse of dimensionality. International Conference on Very Large Databases, 2005.
63. Y. He and J. F. Naughton, Anonymization of Set-Valued Data via Top-Down, Local Generalization. Proceedings of the VLDB Endowment, 2009.
64. M. Terrovitis, N. Mamoulis, and P. Kalnis. Privacy-preserving anonymization of set-valued data, Proceedings of the VLDB Endowment, 2008.
65. G. Loukides, A. Gkoulalas-Divanis, and B. Malin. Privacy-Preserving publication of diagnosis codes for effective biomedical analysis.
    International Conference on Information Technology and Applications in Biomedicine, 2010.
66. Y. Xu et al. Anonymizing transaction databases for publication. ACM SIGKDD International Conference on Knowledge Discovery and Data mining, 2008.
67. G. Loukides, Aris Gkoulalas-Divanis, and J. Shao, Anonymizing transaction data to eliminate sensitive inferences. International Conference on Database and Expert Systems Applications, 2010.
68. J. Cao et al. ρ-uncertainty: Inference-Proof Transaction Anonymization. Proceedings of the VLDB Endowment, 2010.
69. M. Terrovitis, N. Mamoulis, and P. Kalnis. Local and Global Recoding Methods for Anonymizing Set-valued Data. VLDB Journal, 2010.
70. M. Terrovitis et al. Privacy Preservation by Disassociation. TR-IMIS-2010-1. Institute for the Management of Information Systems, ``Athena'' RC, Greece, 2010.



References

71. L. Xiong et al. Privacy-Preserving Information Discovery on EHRs. *Information Discovery on Electronic Health Records*, 2008.
72. L. Sweeney. Replacing personally-identifying information in medical records, the Scrub system, Journal of the American Medical Informatics Association, 1996.
73. J. J. Berman. Concept-match medical data scrubbing: how pathology text can be used in research, Archives of Pathology and Laboratory Medicine, 2003.
74. D. Gupta, M. Saul, and J. Gilbertson, Evaluation of a de-identification software engine to share pathology reports and clinical documents for research, *American Journal of Clinical Pathology*, 2004.
75. J. Gardner and L. Xiong, An integrated framework for anonymizing unstructured medical data. Data and Knowledge Engineering, 2009.
76. V. T. Venkatesan, et al., Efficient Techniques for Document Sanitization, ACM Conference on Information and Knowledge Management, 2008.
77. W. Jiang et al., t-Plausibility: Semantic Preserving Text Sanitization. International Conference on Computational Science and Engineering, 2009.
78. B. Malin and L. Sweeney, Determining the Identifiabiity of DNA Database Entries. AMIA Annual Symposium, 2000.
79. Z. Lin, M. Hewett, and R.B. Altman. Using binning to maintain confidentiality of medical data. AMIA Annual Symposium, 2002.
80. B. Malin, Protecting DNA Sequence Anonymity with Generalization Lattices. Methods of Information in Medicine, 2005.
81. R. Wang et al. Learning Your Identity and Disease from Research Papers: Information Leaks in Genome Wide Association Study. ACM Conference on Computer and Communications Security, 2009.
82. The Guardian, May 2011. http://www.guardian.co.uk/uk/2011/may/11/police-software-maps-digital-movements