

COMP20008 Elements of Data Processing

Lecture 9 Ethics

Intellectual Property, Privacy, and Bias

Semester 2 2025

Lecturer: Abby Meng Yuan

Slides adapted from Prof. James Bailey



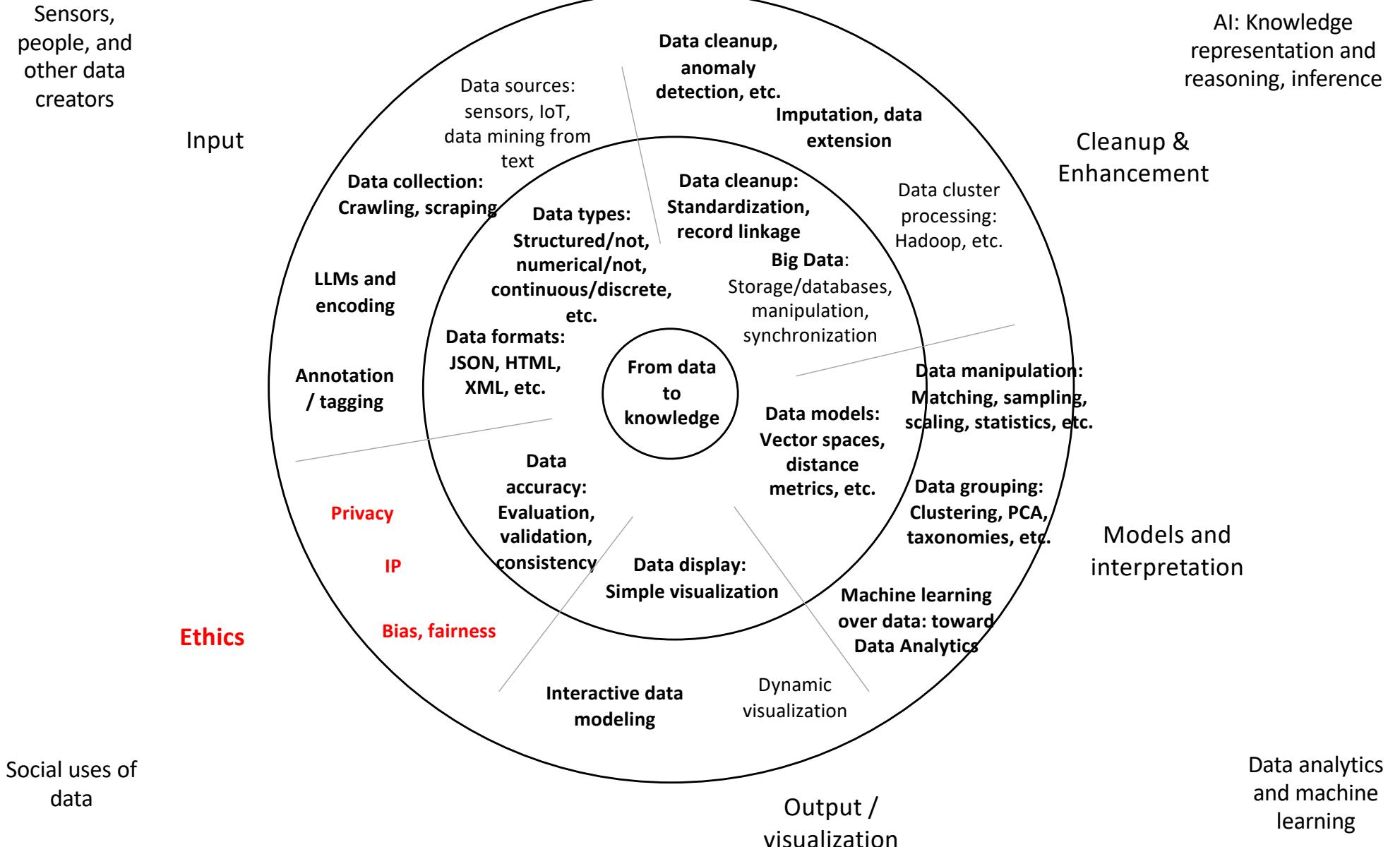
The problem

- There is value in information
- Perhaps someone can discover, steal and sell information from you or about you that might harm you or others
 - Computer scientists can identify individuals hidden in anonymized data with “astonishing ease” and others can use this information against the individuals

<https://fpf.org/wp-content/uploads/The-Re-Identification-of-Governor-Welds-Medical-Information-Daniel-Barth-Jones.pdf>

- Why is this bad?
- What should we (in data processing) do to protect ideas and people?

Data Processing



Lecture outline

1. Intellectual property (IP)
2. Privacy
3. Ethical issues

Intellectual Property

Sharing data electronically

- As soon as the internet / web came up in 1993 people started sharing (as opposed to mailing and posting)
 - Mostly text and music at first
 - Then images and other (structured) data
 - Then analyses and all kinds of data
 - The move from the web to social media accelerated this
- Large volume of shared information are collected for analytical purposes.
- ‘small’ bits of info may be valuable when aggregated and analyzed — who owns that?

IP control

- **Intellectual property (IP)** is a category of property that includes the intangible creations of the human intellect
- Main types:
 - **Patent:** The right granted by the government to an inventor, giving the owner the right to exclude others from making, using, selling, offering to sell, and importing an invention for a limited period of time, in exchange for the public disclosure of the invention
 - **Copyright:** Preventing an original work to be copied or used, except by the creator who has exclusive rights to it, usually for a limited time
 - **Trademark:** a recognizable sign or logo, a design or expression that distinguishes a particular trader's products or services from similar products or services of other traders
 - **Trade secret:** a formula, practice, process, design, instrument, pattern, or compilation of information which is not generally known or reasonably ascertainable, by which a business can obtain an economic advantage over competitors and customers
 - **License:** rules allowing someone to use the property



Your responsibility with data

- You are a data processing person
- When you build a database or use data you should know who owns it and how you and others may use it:
 1. You crawl / scrape other people's information
 2. You extract and process it to make it data
 3. You store it
 4. You process and analyse it, or learn new information
 5. You deliver the results or the data to others



- You should make sure that all legal, ethical, and professional obligations are met, esp. step 5
- There are different established codifications of the terms (licenses) for using data

The law: Copyright

- Copyright holds automatically as soon as you make an image or a text — you don't have to register it
- If you download and use someone's **image** you must give the attribution (unless they give written permission)
 - You may use a single photograph or illustration
 - You may not use more than 5 images by a single artist or photographer
 - You may use up to 10%, but no more than 15 images, from a single publication
 - You may use a single chart, graph, diagram, cartoon, or picture from a book or magazine
- If you repeat more than N words of someone's **text** you must quote and cite them (unless they give written permission)
 - You can use up to 10%, but no more than 1000 words, of essays, articles, or stories, of a single copyrighted work
 - You can use up to 250 words of an entire poem, or a portion of a poem
 - You may not use more than 3 poems or portions of poems by one poet, or by different poets in the same book
 - You may use up to 10%, but no more than 2500 fields or cells, from a database or data table

When is something truly new?

- “Good artists copy; great artists steal – as the saying goes. But what’s the difference between ‘tipping your hat’ to another musician for inspiration, and their lawyer accusing you of copyright infringement? A few million dollars, potentially.”
- Example (though of course Mona Lisa carries no copyright today):
 - How new is Duchamp’s moustachioed Mona Lisa?
 - How new is your own variation of it? And DALL-E’s?

BBC article, Eddie Mullan,
<https://www.bbc.com/culture/article/20190605-nine-most-notorious-copyright-cases-in-music-history>



Marcel Duchamp's moustachioed Mona Lisa sells for \$750,000

Duchamp reproduction of Leonardo Da Vinci's classic painting includes a beard and moustache



Marcel Duchamp's LHOOQ - Mona Lisa with added moustache and beard - has sold for €632,500. Photograph: Graham Turner/The Guardian



DALL-E



- OpenAI's DALL-E is a generative AI neural network trained on millions of images that can recombine fragments into new images on request
- Does DALL-E violate copyright?
- Leonardo cannot complain, but a living painter can — and they do!
- When you use someone's images in your slides you're supposed to give them credit (and pay, even)
- DALL-E (and its users) just combine pieces — that's ok?

ChatGPT and similar

- The same goes for the text equivalent in the Chatmodels of OpenAI (ChatGPT, GPT-4 etc.), Microsoft (Sydney), and Google (Bard)
- If you re-use someone's text you must quote it if it is longer than just a few words
- The Chatmodels and their users don't do this
- Their builders claim their text is an original and novel combination of words based on common usage patterns
- The world is hotly debating the IP here

The image shows a chat interface with two messages. The first message is from a user with the handle 'HO' and the text 'hello. Do you violate copyright?'. The second message is from a large language model with a green icon and the text: 'Hello! As an AI language model, I do not intentionally violate copyright laws. However, it is possible that I may generate responses that could potentially infringe on someone's copyright. It is the responsibility of the user to ensure that the content generated using my responses complies with copyright laws. If you have any specific concerns or questions regarding copyright, I would recommend seeking legal advice from a qualified attorney.'

```
graph TD; HO[HO] --- Q1[hello. Do you violate copyright?]; Model[Large Language Model] --- A1>Hello! As an AI language model, I do not intentionally violate  
copyright laws. However, it is possible that I may generate  
responses that could potentially infringe on someone's  
copyright. It is the responsibility of the user to ensure that the  
content generated using my responses complies with copyright  
laws. If you have any specific concerns or questions regarding  
copyright, I would recommend seeking legal advice from a  
qualified attorney.
```

Useful info

<https://help.figshare.com/article/what-is-the-most-appropriate-licence-for-my-data>

<https://data.research.cornell.edu/content/intellectual-property>

- [CC0 \(+BY\)](#). (Cohen, 2013): A call for using CC0 with data, tempered by an ethical obligation to attribute.
- [Copyright Guide for Scientific Software](#). (Albert et. al, 2019): A joint project of the Harvard Cyberlaw Clinic and the Center for Astrophysics, in association with the Software Preservation Network, that provides clear, easy-to-read answers to common questions about how scientific software and copyright interact.
- [Data Citation Developments](#) (Kratz, 2013): An update on efforts to standardize data attribution requirements.
- [How to License Research Data](#). (Ball, 2012): Written with British law in mind, but it has a good discussion of the pros and cons of the ODC licenses.
- [Licensing Open Data: A Practical Guide](#). (Korn & Oppenheim, 2011): Another guide written with UK law in mind, but with a helpful comparison of CC and ODC licensing options.
- "Open Data." (Wikipedia)
- [Open Licenses](#). (Project Open Data): The US Federal Government guide to open licenses and dedications.
- [RDA & CODATA Legal Interoperability of Research Data: Principles and Implementation Guidelines - Now Published](#) (Research Data Alliance): A set of practical guidelines for researchers dealing with legal aspects of sharing data.
- [Sharing Research Data and Intellectual Property Law: A Primer](#) (Carroll, 2015): An introduction to the various kinds of property rights that can be associated with research data.

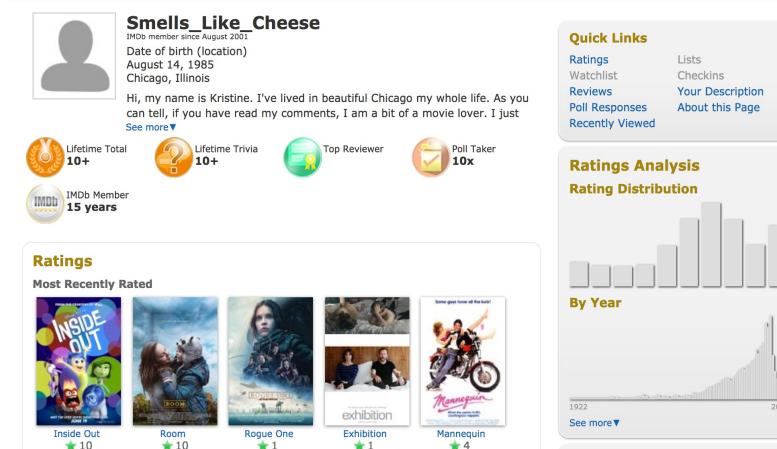
Privacy

Example 4: Netflix dataset

- 2006: Netflix publicly releases 6 years of data about its customers' viewing habits
 - Cinematch is a bit of software embedded in the Netflix Web site that analyzes each customer's movie-viewing habits and recommends other movies that the customer might enjoy
 - <https://www.nytimes.com/2008/11/21/technology/21iht-23netflixt.18049332.html>
 - An anonymous id is created for each user
 - Sampled 10% of their data
 - Slight data perturbation
- Netflix Prize: 2009 Competition to find better collaborative filtering algorithms (10% improvement over Cinematch)
 - \$1 million prize for a model, won by BellKore team

Next step: Linking Netflix data with public IMDb data

- Two researchers, Narayanan and Shmatikov:
 - <https://arxiv.org/pdf/cs/0610105v2.pdf>



- Given knowledge about a person’s “public” movie habits on IMDb, showed it was possible uncover their “private” movie habits in the Netflix dataset
 - 8 movie ratings (≤ 2 wrong ratings, dates ± 2 weeks):
 - 99% re-identified raters

The lesson

- One database might preserve privacy
- And so may another
- But **together** they may give enough information to violate privacy
- You NEVER know what some future database may provide
- So you can NEVER be sure your privacy measures will be enough

So, do we just give up?

What we would like:

1. Releasing *some* data, but with a guarantee of individual privacy no matter what
2. A flexible degree of overall privacy that can be controlled

How can we do this?

Privacy overview

Personal privacy:
each person
protects themself

Strategies:

- Don't disclose
- Lie

Local privacy: the
data owner makes
changes to fields
in a record

k-anonymity
l-diversity

Strategies:

- Add noise
- Remove features
- Generalise features

Global privacy: the
data owner makes
changes when
answering a query

budget k
global sensitivity G

Strategies:

- Change fields
- Hide info

Terminology

- **Explicit identifier:** Unique for an individual
 - name, national ID, Tax file number, account numbers
- **Quasi identifier:** A combination of non-sensitive attributes that can be linked with external data to identify an individual
 - e.g., {Gender, Age, Zip code} combination from earlier
- **Sensitive attribute**
 - Information that people don't wish to reveal (e.g., medical condition)

Personal choice privacy

Personal (self) privacy: Ask people what they want to share

- Idea: People know what they consider private about themselves
- So let them provide only the data they feel like sharing — the rest they can **hide or lie** about
 - Example: ask people to lie, and then make inferences based on the aggregate answers
 - “The future of privacy is lying”
(April 10 2013, Matt Buchanan, *New Yorker*)

Lying

- Each person is responsible for adding noise to their own data
- Classic survey example: each person has to answer question “Do you use drugs?”
 - They flip a coin in secret and answer “Yes” if it comes up heads but tell the truth otherwise
 - Plausible deniability about a “Yes” answer
- But lying doesn’t always work!

Privacy requires knowing who else is in the data/survey

- Imagine a survey is asking you:
 - Question: Are you a smoker?
 - Result: Number of smokers will be reported
- Would you take part in this dataset/survey?
- Only if you are not the only smoker!
- You don't know who else is being asked!

ID	Age	Sex	Smoker
sdhj5vbg	20	Male	False
wu234u4	25	Female	True
hi384yrh	17	Female	False
po92okwj	50	Male	False

← Only one smoker!

Instead: Centralised privacy management

- If I'm that smoker, then my presence or absence makes a big difference to the survey!
- This can only be managed by someone at the centre who can see all the data
- They can protect each individual
- They can protect the entire dataset

Local and global privacy

- We have a sensitive dataset, a trusted data owner Alice, and a researcher Bob. Alice does analysis on the raw data
- **Local:** Changes made to protect each individual
 - Alice wants to protect individuals
 - Alice decides to change some individual bits of data
- **Global:** Changes made in the response to a query
 - Bob asks Alice for some data
 - Alice adds noise or changes some fields in the answers, and reports the (noisy) answers to Bob
 - Alice can change what to hide depending on the query

Local privacy

Privacy overview

Self privacy: each person protects themselves

Strategies:

- Don't disclose
- Lie

Local privacy: the data owner makes changes to fields in a record

k-anonymity
l-diversity

Strategies:

- Add noise
- Remove features
- Generalise features

Global privacy: the data owner makes changes when answering a query

budget k
global sensitivity G

Strategies:

- Change fields
- Hide info

Measures of anonymity for individuals

How can the data manager ensure individual privacy?

Make sure that no-one can be identified uniquely

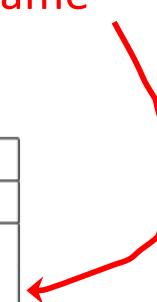
- But removing explicit identifiers from a dataset is not enough ... the intersection of datasets can break this
- 2 solutions:
 1. K-anonymity
 2. L-diversity

k-anonymity

- Produce a release of parts of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified, but make sure the data remains practically useful
- Definition: A table satisfies *k-anonymity* if every record in the table is indistinguishable from at least $k - 1$ other records with respect to every set of quasi-identifier attributes
- Such a table is called a *k-anonymous* table
 - For every combination of values of the quasi-identifiers in the k-anonymous table, there are at least k records that share those values

k-anonymity example

Doesn't matter which way you look at it — always at least 4 records with same attribute set



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

I-Diversity: Privacy Beyond k-Anonymity. Machanavajjhala, Gehrke, Kifer and Venkatasubramaniam, 2007

Exercise: What level of k-anonymity is satisfied here?

- Sensitive attribute: COMP20008 Grade
- Quasi identifier: {Gender, Age, Hair Colour}

Student Name	Gender	Age	Hair Colour	COMP20008 Grade
7930c	Male	20	Brown	78
1a985	Male	20	Brown	88
04ed9	Female	19	Red	75
82260	Female	19	Red	85
e461e	Female	19	Red	80
1e609	Female	21	Red Brown	90

If we can 'collapse' these two
 $k=1, 2, 3 \text{ or } 4?$ $k = 1$ values then we get 2-anonymity

What is maximal k for which it satisfies k -anonymity?

How to achieve k-anonymity, 1

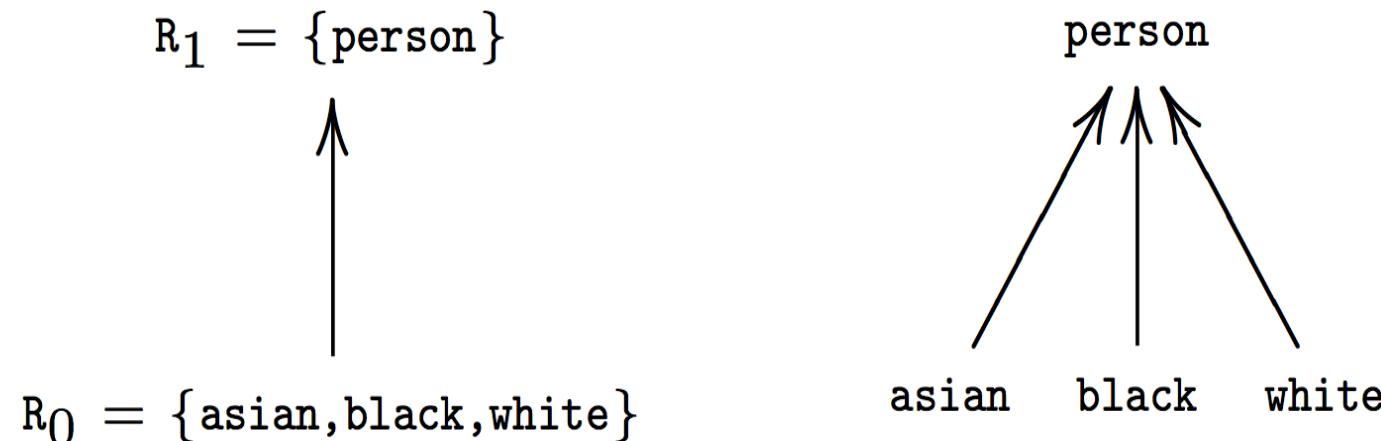
- Method 1: **Suppression** (“deletion”):
 - Just remove (suppress) the quasi identifiers completely
 - Leave the cells empty
 - This limits the number of outliers
 - Works for row, column, or cell level



Student Name	Gender	Age	Hair Colour	COMP2008 Grade
7930c	Male	20	Brown	78
1a985	Male	20	Brown	88
04ed9	Female	19	Red	75
82260	Female	19	Red	85
e461e	Female	19	Red	80
1e609	Female	21	Brown	90

How to achieve k-anonymity, 2

- Method 2: **Term generalization (“collapsing”):**
 - Make the quasi identifiers less specific
 - This works on the whole column level
 - Example: race



<http://www.springerlink.com/content/ht1571nl63563x16/fulltext.pdf>

How to achieve k-anonymity, 3

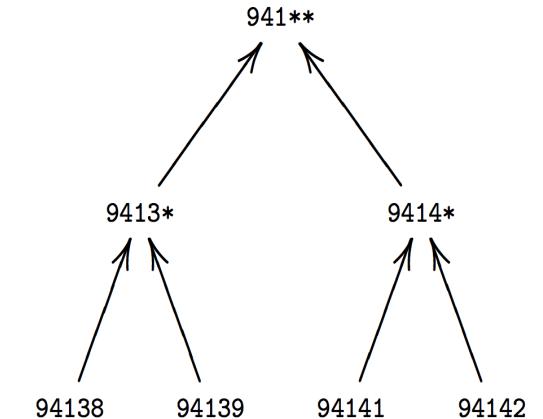
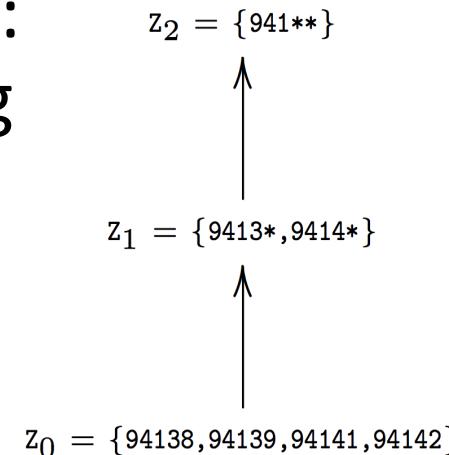
- **Method 3: Numerical generalization:**
 - Make the quasi identifiers less specific
 - This works on a section of the column
 - Example: Zip code numbers

Student Name	Gender	Age	Zip code	COMP20008 Grade
7930c	Male	20	30539	78
1a985	Male	20	30539	88
04ed9	Female	19	30541	75
82260	Female	19	30541	85
e461e	Female	19	30538	80
1e609	Female	21	30536	90

k-anonymity example 1

- Example: Zip code:
When generalizing
94138, which is a
better strategy?

- 9413*
- *4138



<http://www.springerlink.com/content/ht1571nl63563x16/fulltext.pdf>

- Why?
- Zones close by share weather and other details,
so rest of the info is pretty similar. Using very
different areas might reveal additional info

k-anonymity guarantee: What is the max amount of damage?

In the worst case, if data gets into the wrong hands, this can only narrow down a quasi identifier to a group of k individuals

Data publisher needs to

1. Choose acceptable parameter k
2. Determine which quasi identifier(s) to alter

But k-anonymity is not perfect!

Attack on k-anonymity, 1:

Sensitive data homogeneity problem

- k-anonymity is not perfect!
- It may create a group that leaks information due to similarity of the sensitive attribute
 - Alice knows that Bob is a 31-year-old American who lives in zip code 13053
 - So Alice knows that Bob's record number is 9,10,11, or 12
 - Alice can **still** conclude that Bob has cancer if she sees the data

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Attack on k-anonymity, 2: Background info problem

- k-anonymity also does not protect against attacks that use background knowledge
 - Alice knows that Umeko is a 21 year-old Japanese who currently lives in zip code 13068
 - So she knows that that Umeko's information is record number 1,2,3, or 4
 - She knows that Japanese usually have very low incidence of heart disease
 - So she concludes that Umeko has a viral infection!

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

I-Diversity: Privacy Beyond k-Anonymity.
 Machanavajjhala, Gehrke, Kifer and
 Venkatasubramaniam, 2007

Solution: L-diversity

- Make sure the sensitive attribute is also diverse *within each group*
- L-diversity: For each k anonymous group, make sure there are at least l different sensitive attribute values

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

At least one is different!

I-Diversity: Privacy Beyond k-Anonymity.
Machanavajjhala, Gehrke, Kifer and
Venkatasubramaniam, 2007

Figure 3. 3-Diverse Inpatient Microdata

L-diversity example

Everyone in 14850 has a Viral Infection!

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

But not everyone in 1485*

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	< 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Figure 3. 3-Diverse Inpatient Microdata

Doesn't matter which way you look at it — any group always has at least 4 records with different attributes

Limitations of l-diversity

1. It's difficult to do because your data has to be suitable: Imbalanced number of occurrences of a sensitive attribute within just one group may limit usefulness
2. May be unnecessary

K-anonymity and l-diversity

- **k-anonymity** comes in when you want to make sure there are at least k records in the dataset for every combination of non-sensitive parameters
- **l-diversity** comes in when you want to make sure there are at least l different values in the dataset for every set of non-sensitive parameters

Summary for simple privacy

- To reduce risk of re-identification of individuals in released datasets
 - Choose value of k
 - Manipulate data to make it k -anonymous, either
 - Replace categories by broader categories
 - Suppress attributes with a * (limited utility)
 - Further manipulate data to make it l -diverse
 - Ensure there are at least l different values of the sensitive attribute in *each* group
 - Privacy is difficult to maintain in high-dimensional datasets like trajectory datasets
 - Cloaking provides spatial k -anonymity
 - Obfuscation ensures location imprecision

Privacy overview

Self privacy: each person protects themselves

Strategies:

- Don't disclose
- Lie

Local privacy: the data owner makes changes to fields in a record

k-anonymity
l-diversity

Strategies:

- Add noise
- Remove features
- Generalise features

Global privacy: the data owner makes changes when answering a query

budget k
global sensitivity G

Strategies:

- Change fields
- Hide info

Global privacy

- Often we can achieve k-anonymity and l-diversity in different ways
- Different changes ‘damage’ the data to different degrees
- Sometimes we trust one querying person more than another with the data
- How careful/private do we want to be?
- How can we find the ‘best’ change for each querier?
- Idea: **Compare the global effect of each possible change**

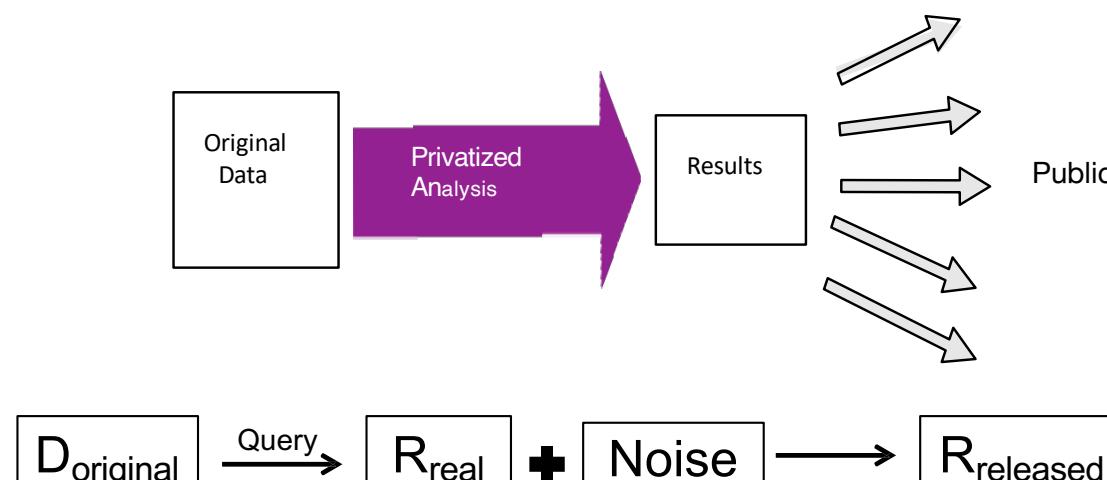
Which records to alter?

Try to keep the dataset as ‘true’ as possible
Try to alter as few records as possible

- Alter info that makes only a small difference to the overall statistics — the chance that the privatized result would be the truth (or some version **R**) is nearly the same whether or not the record is altered or dropped
- This can only be managed by someone at the centre who can see all the data

Global differential privacy: Treat each change as ‘noise’

- You can change the degree of privacy on a query-by-query basis
- Two steps:
 1. Query the data to find and consider every individual relative to the query (e.g., how many female students are in the survey?)
 2. Add random noise to hide the presence/absence of any individual for that query. Release noisy result to the user



Adapting noise for differential privacy

- Different queries ask for different sets of records
- Strategy: Answer with a variable amount of noise (= ‘lies’) included to protect privacy
 - Make the released results different each time
$$\textit{Released result} = \textit{True result} + \textit{noise}$$
- This will help us guarantee that the queryer cannot learn sensitive information about individuals from the released results
- Careful...too much noise makes the data valueless!

Example: Different amounts of noise

- Query: How many females in the dataset?
(true result = 32)
- Generate some random values, according to a distribution with mean value 0: {1,2,-2,-1,0,-3,1,0}, add to true result and release
 - 1. Released result=33 (32+1)
 - 2. Released result=34 (32+2)
 - 3. Released result=30 (32-2)
 - 4. Released result=31 (32-1)
 - 5. Released result=32 (32+0)
 - 6. Released result=29 (32-3)
 - 7. Released result=33 (32+1)
 - 8. Released result=32 (32+0)
- On average, the released result will still be 32
- But observing a single released result doesn't give the adversary exact knowledge

Accuracy vs privacy

- The less we change, the more ‘true’ / accurate our results are
- The more we change, the more we protect privacy of our dataset

The balance is a tradeoff: try to change as little as possible while guaranteeing as much privacy as possible

Picking the record(s) to change

- There may be thousands of records...I can alter any of them! Which record(s) to pick? How many?
- Pick the one(s) that affect the output the least:
 - Calculate the effect of changing a field's value
 - If the difference is 'too large' then don't change that value
- **Global sensitivity:** the maximum difference in answer accuracy that changing any field will cause (= maximum effect of an individual field)
- When handling multiple queries, global sensitivity is equal to the sum of these differences

Example: what to change?

- Example query: how many smokers are there?

- Change a **smoker**:
remove 1 of 2

- Change a **nonsmoker**:
remove 1 of 5

- Which is better?

Row	ID	Age	Sex	Smoker
1	sdhj5vbg	20	Male	False
2	wu234u4	25	Female	True
3	pl34kkj9	21	Female	False
4	n66rsy3	17	Male	True
5	hi384yrh	17	Male	False
6	po92okwj	50	Male	False
7	jj97830g	45	Female	False

- In the result we return, $\frac{4}{5}$ is closer to the truth than $\frac{1}{2}$
- So we should change one of the **nonsmoker** records

Changing records: Two questions

1. How many records should we alter?

=> Choose desired privacy; alter up to that limit

=> **Privacy loss budget k**

2. Which ones should we alter? Why them?

=> Alter ones that make the least global difference

=> **Global sensitivity G**

Definitions

- **Privacy loss budget k:** An integer expressing how much tolerance we have for altering our released data for the sake of privacy (how hard for the queryer to guess the true result). A small k means less change. This is determined by the data owner
- **Global sensitivity G:** A number expressing how much difference a change to an individual record field can make to the result. This is determined by calculating the difference in likelihood of an accurate result R when one or more records is altered

Global sensitivity example

Global sensitivity is the maximum difference in the answer accuracy from changing an individual fact in the dataset

An example:

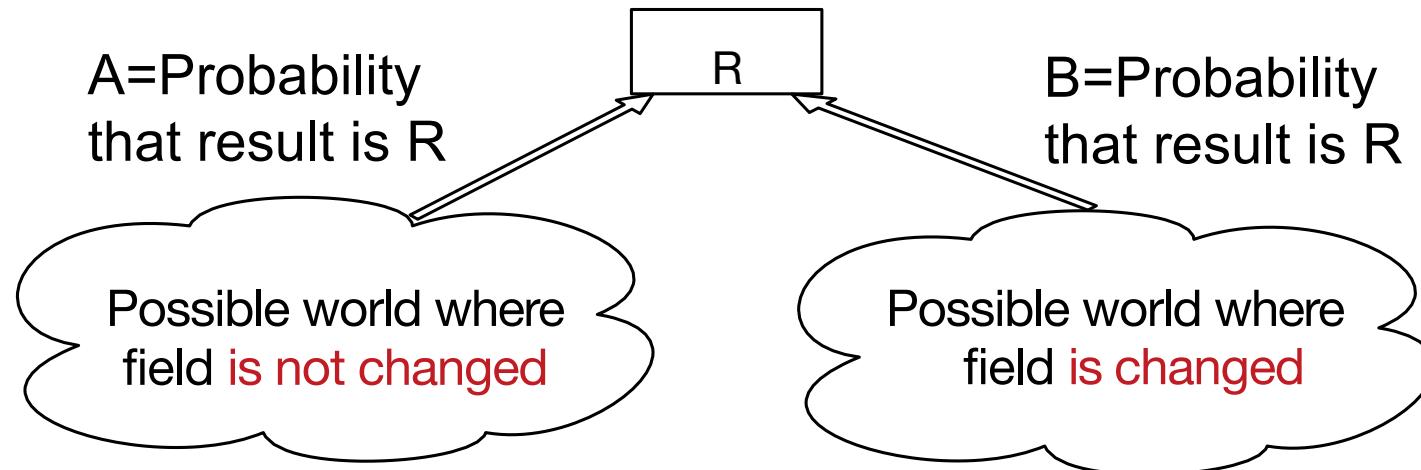
- Consider query *CountFemale* (counts how many females there are)
- To calculate the global sensitivity, consider the maximum difference in this query result when an individual is changed
 - For this query, the global sensitivity is 1 because adding/removing a female to * changes the count by 1
 - If there had been no females in the dataset then the sensitivity would be zero

Sex	Marital Status
M	Single
M	Married
F	Single
M	Single
F	Married

Note that when asking multiple queries, the global sensitivity is equal to the sum of the differences

The promise of differential privacy

- The chance that the noisy released result will be R (for example: the truth) must be nearly the same *whether or not* any individual fact is included in the dataset



- If we can guarantee $A \approx B$ (A is very close to B), then no one can guess which actual dataset resulted in R
- So we want to make a change with a small effect

Calculating E

Calculate the effect E
on the output when
you alter different fields
from their value to *

Row	ID	Age	Sex	Smoker
1	sdhj5vbg	20	Male	False
2	wu234u4	25	Female	True
3	pl34kkj9	21	Female	False
4	n66rsy3	17	Male	True
5	hi384yrh	17	Male	False
6	po92okwj	50	Male	False
7	jj97830g	45	Female	False

Row	Change in Sex	Counts	E
2	Male: no change Female: 1/3 (1 of 3)	0+1/3	0.33
1	M: 1/4 F: no change	1/4+0	0.25
Row	Change in Smoker	Counts	E
2	Smoker: 1/2 Nonsmoker: no change	1/2+0	0.5
1	S: no change NS: 1/5	0+1/5	0.2

Calculating Effect, continued

- Changing a Smoker to * makes the largest difference ($E=0.5$) because there are few smokers
- Changing a Nonsmoker to * makes the smallest difference ($E=0.2$) because there are many nonsmokers
- Changing a Male or Female makes less difference because their totals are not as large
- How about changing both fields *together*?
- What do you predict?

The privacy budget

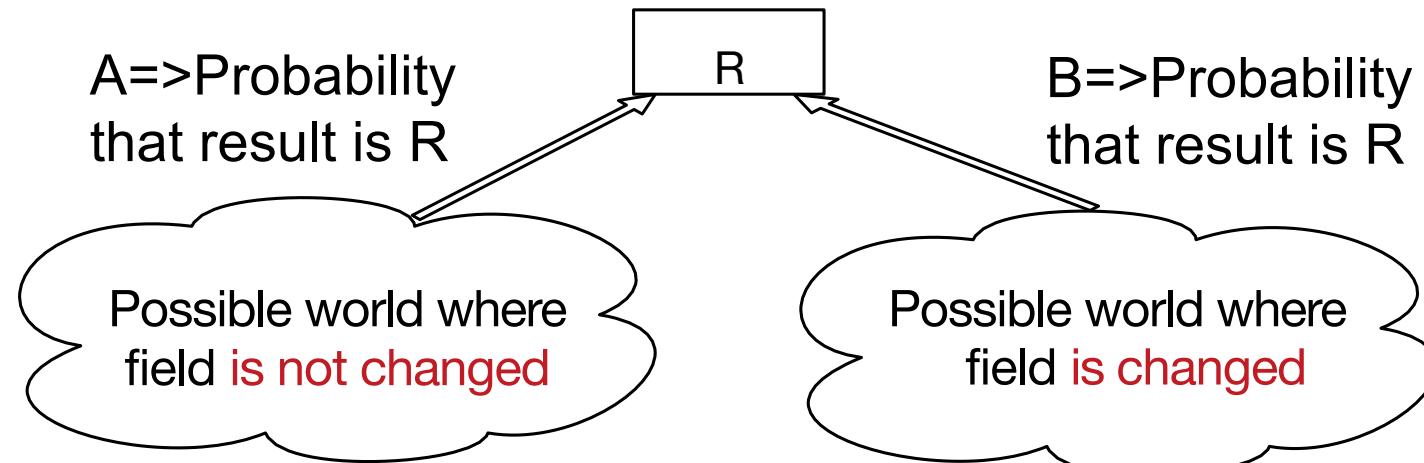
- The least change happens with just one field alone
- Is it enough? — How much change can we accept?

Change budget: We pick a number k that specifies how inaccurate we are allowed to be (= how willing we are to change data)

- We can pick a new k for each user and each query
(Note: less accurate = more data loss → less useful)

Privacy loss budget = k

- Privacy loss budget = k ($k \geq 0$)
- Calculate probabilities for data releases A and B
- Choose record(s) to change so that $\text{prob}(A) \leq 2^k \times \text{prob}(B)$



$k=0$: No privacy, no records changed $\rightarrow A=B \rightarrow R=1$

$k=\text{low}$: Small difference $\rightarrow A \approx B \rightarrow A$ and B close

$k=\text{high}$: Larger difference $\rightarrow A > B \rightarrow$ result diverges more

Calculating with the budget k

- The querier is trying to guess something about my data. He asks me for the data
 - Let's say the true data is A. But I don't want to give him the true data. So I can corrupt the data in various ways to make different possible answers B_1 , or B_2 , or B_3 ...
 - I can calculate the probability R that B_1 , or B_2 , or B_3 is 'true' (by comparing to A) ... the more corruption, the lower R becomes
 - Which data should I give him, A or B_1 , B_2 , B_3 ...?
 - Example probs: $\text{Prob}(A)=1$, $\text{prob}(B_1) = 0.5$, $\text{prob}(B_2) = 0.3$, $\text{prob}(B_3) = 0.1$... becoming less and less likely (= more and more corrupted)
- I don't want to lie *too* much
 - So I give myself a budget k = how much corruption I will tolerate
 - Let $\text{prob}(A) \leq 2^k \times \text{prob}(B)$, same as $\text{prob}(A)/\text{prob}(B) \leq 2^k$
 - So when $\text{prob}(B_i)$ is very small (= big corruption), I need a large k
 - Or: if I use a small k , then I must use B_i so that $\text{prob}(B_i) \approx \text{prob}(A)$

Remember:
 $\text{prob}(B) < 1!$

Example calculation

Change a record if the difference between including and changing it (\Rightarrow prob accuracy R) is less than the tolerance that k allows

A (included)	B (not included)	k	$2^k \times B$	$A \leq 2^k \times B?$
0.9	0.3	1	0.6	N
0.9	0.9	1	1.8	Y
0.9	0.3	3	2.4	Y
0.9	0.9	3	7.2	Y

$A \leftrightarrow B$ vs k

A big change if changed,
/ a small k — no change!

\ A small change if changed,
a small k — ok to change

A big change if changed,
/ a large k — ok to change

\ A small change if changed,
a large k — ok to change

Putting k and G together

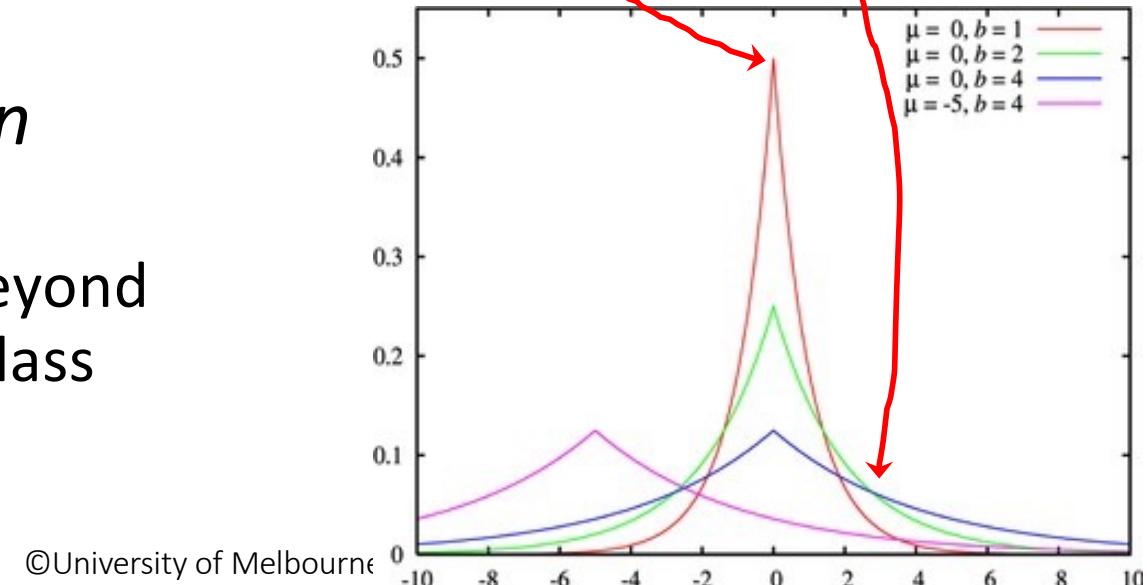
- We want to alter as little data as possible within our privacy loss budget k
- We want to alter the most ‘revealing’ data with little damage to the data accuracy — minimize G
- So...while keeping constant the average data, we vary the ‘spread’ of the noise:
 - alter more-revealing records but fewer of them,
 - or alter less-revealing records, but more of them
- Calculate the ratio G/k

Now we pick the noise value(s)

- *Released result = True result + noise*
- Noise is a number randomly sampled from a distribution with
 - average value = 0 (μ)
 - standard deviation (spread) = G/k (b)
- This is called the *Laplace distribution*
 - Details about the distribution are beyond the scope of the class

Stay accurate on average

Spread: Tend to be accurate or vary?

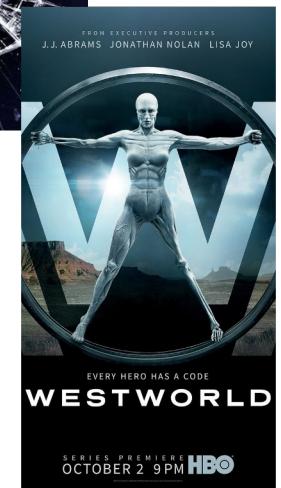
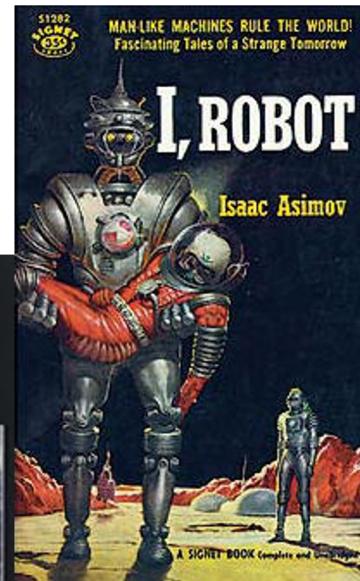
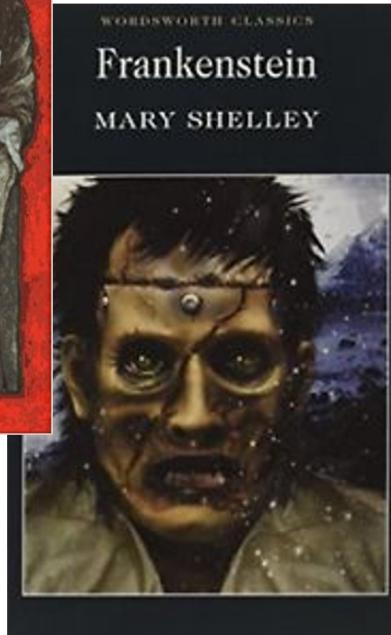
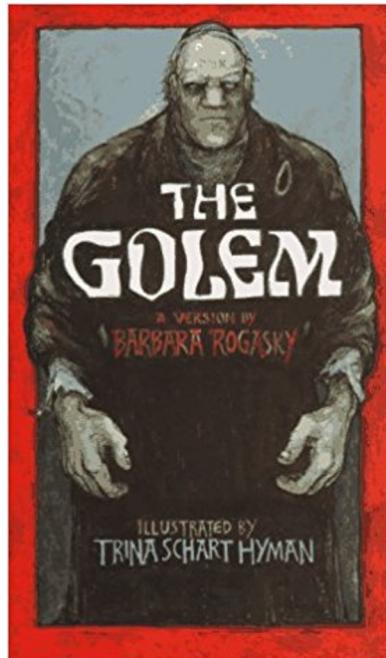


Summary

- Differential privacy guarantees that the presence or absence of a user cannot be revealed after releasing the query result
- It does not prevent queryers from drawing conclusions about individuals from the aggregate results over the population
- To determine the scale of the noise to be added we must decide the budget and find the global sensitivity

Ethical issues

The long history of ethics and CS



Outline

- Ethics of Data Processing and AI
- Ten rules for preventing problems

What is Ethical action?

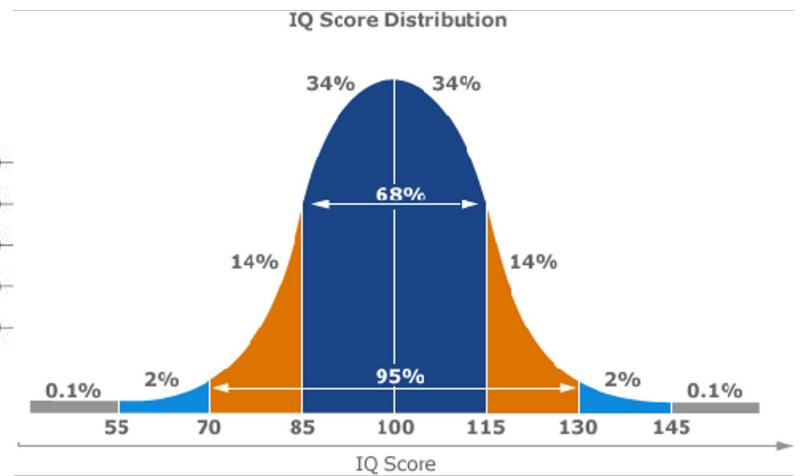
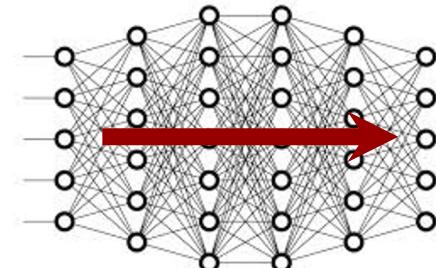
Decisions we make about our data, methods, and tools are tied up with their impact on people and societies

- It's the good things
- It's the right things

How easy is it to define
what's *good* and what's *right*?

An IQ classifier

- Let's instead train a classifier to predict people's IQ from their photos
 - Input: Face features + text features + intelligence estimator from language usage
 - Output: IQ score to express the apparent relative intelligence of a person



An IQ classifier

- If the classifier were 100% accurate
 - Who could benefit from such a classifier?
 - Who could be harmed from such a classifier? How can such a classifier be misused?
- If the classifier is only 90% accurate
 - Say accuracy is 95% for white blonde females
 - Say accuracy is only 60% for people with brown/black hair under age 25
- Would you object? Who is responsible?
 - Researcher/developer? Reviewer? University? Society?

Ten rules for preventing problems

Rule 1

Assume that ‘data are people’ so it can do harm

- All data are [about] people until proven otherwise
 - Social media
 - Heart rates from Youtube videos
 - Ocean rise measurements that change property risk profiles

Rule 2

Recognize that privacy is more than a binary value

- Privacy is contextual and situational
 - Single Instagram photo versus entire history of social media posts
 - Privacy preferences differ across individuals and societies

Rule 3

Guard against the re-identification of your data

- When supposedly anonymized datasets are combined with other variables, the combination may enable unexpected reidentification
 - Metadata associated with photos
 - Reverse image search — connect dating and professional profiles
 - Difficult to recognize the vulnerable points a priori!
 - Ex.: Battery usage on a phone can reveal a person's location
- Unintended consequence of 3rd party access to phone sensors

Rule 4

Practice ethical data sharing

- Seek consent from participants to share data

Rule 5

Consider the strengths and limitations of your data

- Big does not automatically mean better
 - “there’s no data like more data”
 - “...but *relevant* data is better than more data!”
- Document the provenance and evolution of your data. Do not overstate clarity; acknowledge messiness and multiple meanings
 - Is a Facebook post or an Instagram photo best interpreted as an approval/disapproval of a phenomenon, a simple observation, or an effort to improve status within a friend network?

Rule 6

Debate the tough, ethical choices/issues

- Importance of debating the issues within groups of peers — educate everyone
 - Examples mentioned earlier
 - Facebook emotional contagion
 - Facebook self-disclosure experiments, with or without thumbs-up/down feedback
 - Exposing teen girl's pregnancy by purchase patterns

Rule 7

Develop a code of conduct for your organization, research community, or industry

- Are we abiding by the terms of service or users' expectations?
- Does the general public consider our research "creepy"?

Rule 8

Design your data and systems for auditability

- Plan for and welcome audits of your big data practices
- Systems of auditability clarify how different datasets (and the subsequent analysis) differ from each other, aiding understanding and creating better research
 - “For example, many types of social media and other trace data are unstructured, and answers to even basic questions such as network links depend on the steps taken to collect and collate data.”

Rule 9

Engage with the broader consequences of data and analysis practices

- Recognize that doing big data research has societal-wide effects

Rule 10

Know when to break these rules

- It may be important to temporarily put aside questions of individual privacy in order to serve a larger public good
 - Natural disaster
 - Public health emergency
 - Hostile enemy
 - ...

Summary

- Although as a data processing person you now know HOW to work with data
(...classify, impute, make inferences, discover, display...)
- This doesn't mean you SHOULD always do what you can!



- There are legitimate and non-legitimate uses
(privacy, safety, wellbeing, etc.)
- Ask how your results could be used against someone
- Ask if YOU would be ok with it if it happened to you

What you must know

- For IP:
 - What is a license, copyright, trademark, and patent
 - What GDPR is and why it was created
 - How much text and images you can use without violating ©
- For Privacy:
 - What are explicit and quasi identifiers
 - What are k-anonymity and l-diversity and their limitations
 - What are budget k and global sensitivity G
 - How differential privacy works by withholding records to increase the amount of ‘noise’
- For Ethics:
 - The ethical and moral considerations of data processing and Big Data Analytics
 - The 10 rules of ethical data processing

Acknowledgements

- Intellectual Property: Thanks to James Bailey
- Privacy material adapted from
 - Massachusetts story
 - https://epic.org/privacy/reidentification/ohm_article.pdf
 - From a social science perspective
 - [http://papers.ssrn.com/sol3/papers.cfm? abstract_id=1450006](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006)
 - l-diversity
 - <https://www.cs.cornell.edu/~vmuthu/research/ldiversity.pdf>
- Ethics: Thanks to Yulia Tsvetkov and Pauline Lin

End

