Deanonymization: Re-identification Attacks

Dr. Balázs Pejó

www.crysys.hu

Agenda

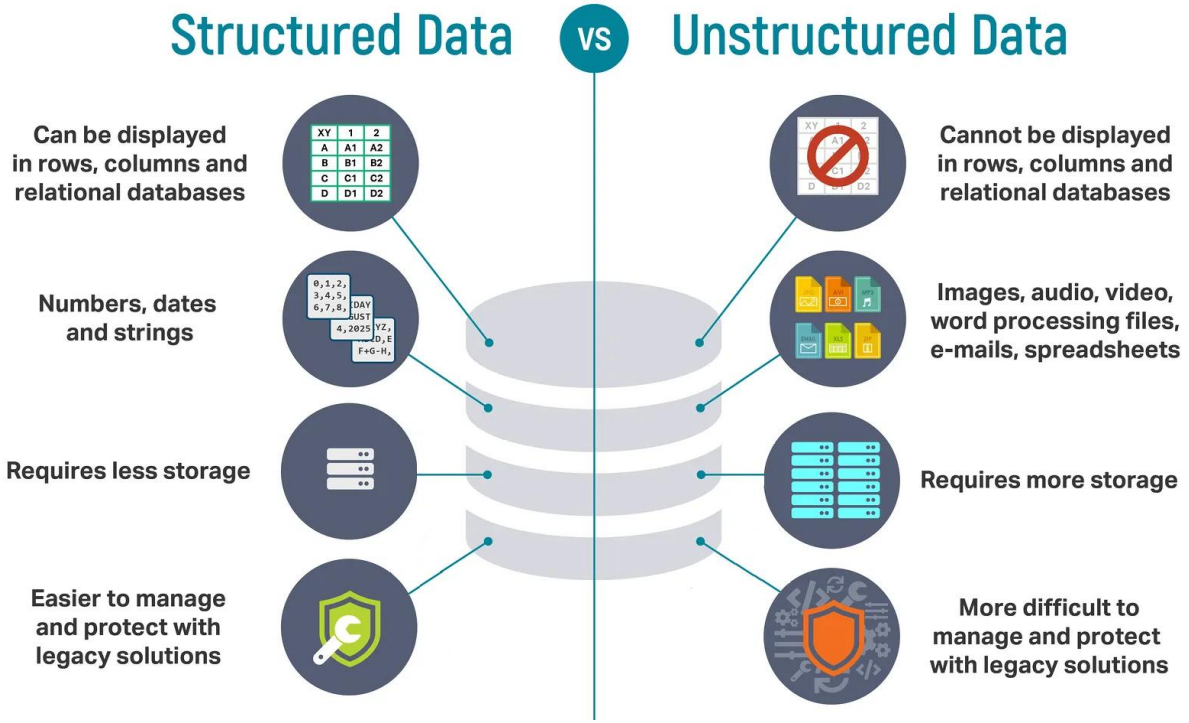
- Dark Patterns
- Tracking
- GDPR
- Deidentification
- Machine Learning
- Anonymization
- Cryptography

- Data Types
- GDPR & The Opinion
- Matching Attacks
- Re-identification Attacks on Micro Data
- Re-identification Attacks on Unstructured Data
  - Stylometry
  - TOR

Data Types

Structured vs Unstructured

- Unstructured data
  - Text (e.g., medical prescriptions, invoices, emails)
  - Images (e.g., X-ray, MRI, photos)
- Structured data
  - Micro-data: each record represents a single individual's data.
  - Statistical (aggregate) data: each record represents multiple individual's data (e.g., histogram).

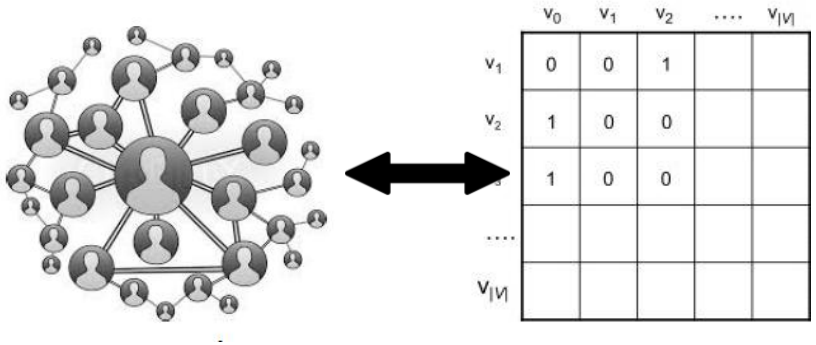


Micro-data

Pseudo-id	Gender	Age	ZIP	Degree	Income	...
00013701	Male	21	77005	Bachelor	13,000	...
08936402	Male	37	77065	Master's	90,000	...
42330327	Female	60	89123	High School	46,000	...
...	...	...	...	...	...	...

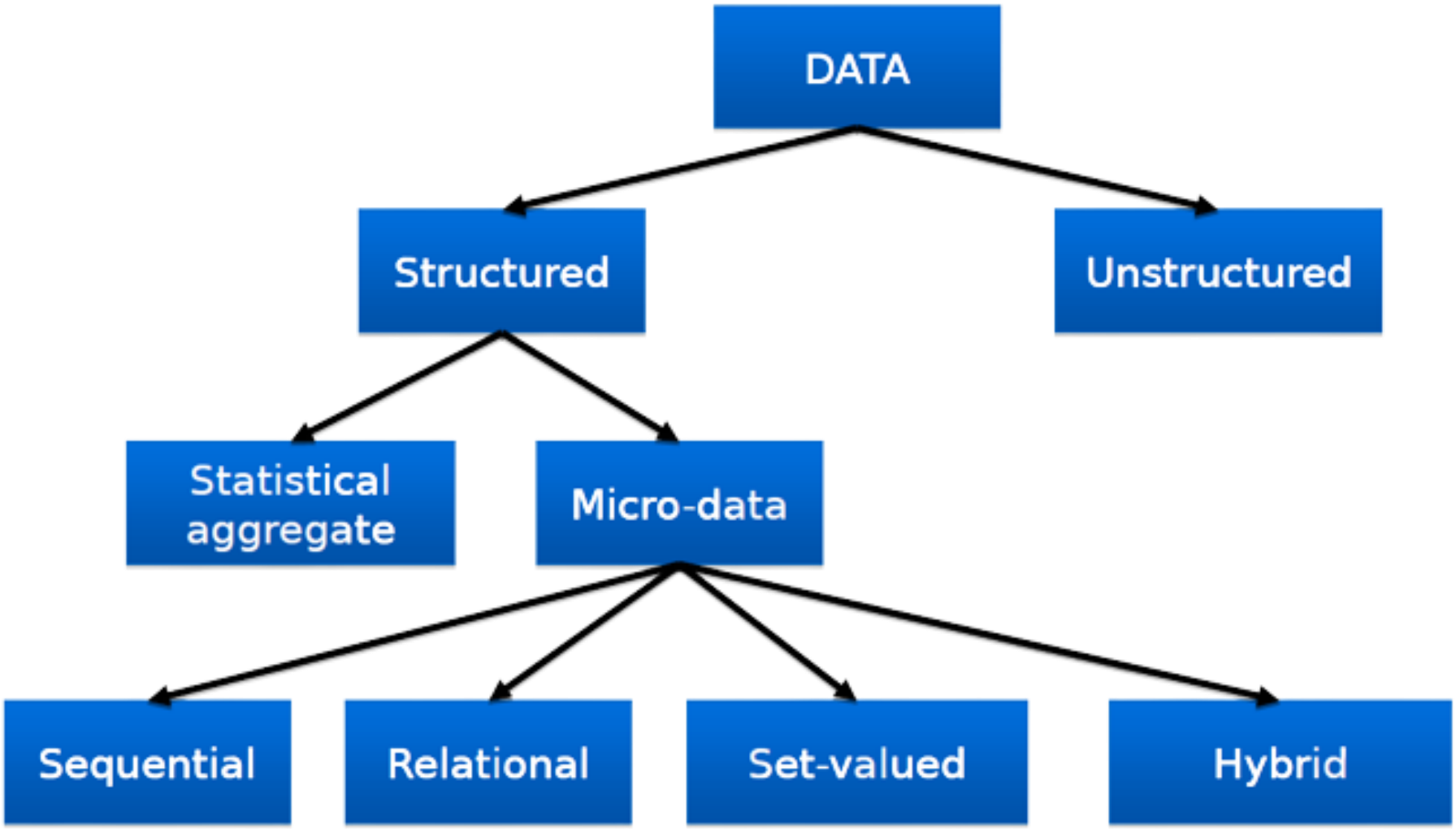
- Relational (tabular) data
- Sequential data: time-series (time × value), e.g., location trajectories, power consumption.
- Set-valued (transactional) data: represented as a relational table with binary attributes, e.g., purchased item, watched movies, social graphs.
- Hybrid data: contains set valued data and other attributes.

Rec #	Sex	Zip	Date	Purchased goods
1	Male	3400	2/7/2015 14:32	{Bread, Wine, Diaper}
2	Female	3600	2/8/2015 10:11	{Bread, Bier, Cold meat, Shampoo, Viagra, Sushi, Milk, Cheese}



Sequence Database	Sequence	Element (Transaction)	Event (Item)
Customer	Purchase history of a given customer	A set of items bought by a customer at time t	Books, diary products, CDs, etc
Web Data	Browsing activity of a particular Web visitor	A collection of files viewed by a Web visitor after a single mouse click	Home page, index page, contact info, etc
Event data	History of events generated by a given sensor	Events triggered by a sensor at time t	Types of alarms generated by sensors
Genome sequences	DNA sequence of a particular species	An element of the DNA sequence	Bases A,T,G,C

Summary


DATA
Structured
Unstructured
Statistical aggregate
Micro-data
Sequential
Relational
Set-valued
Hybrid

Dimensionality

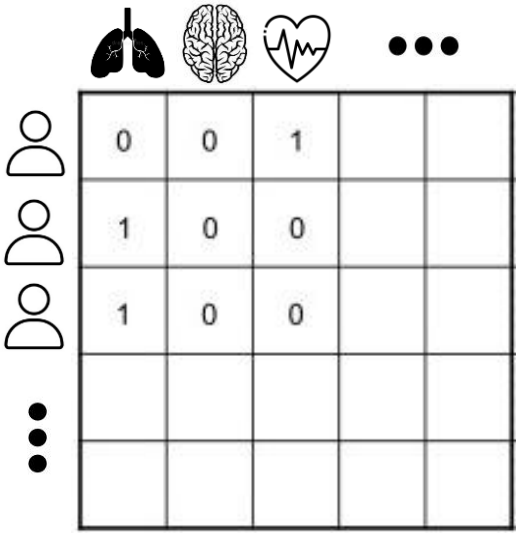
- Refers to how many attributes a dataset has.
  - In an ideal world, data could be represented in a spreadsheet, with one column representing each dimension.
  - The actual dimensions can be unknown, as some columns could be correlated (e.g., duplicates), some are useless, etc.

- High (>1000) dimension is bad for privacy!
  - Healthcare has vast amounts of variables (e.g., blood pressure, weight, cholesterol level).

- Low (<100) dimensions is bad for utility!

- Inevitable Trade-Off between Privacy and Accuracy!

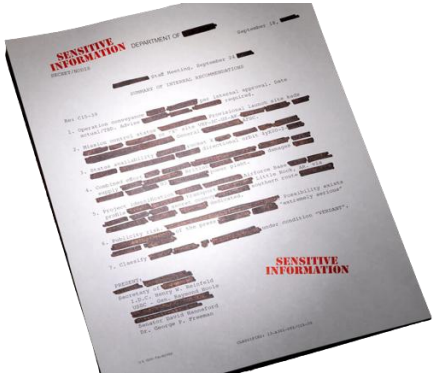
Column 1	Column 2
Peter	Human
John	Human
Kate	Human



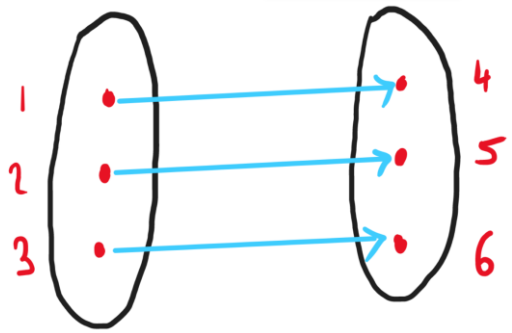
De-anonymization

Anonymized Data

- A simply anonymized dataset does not contain *personally identifiable information* (PII) such as name, address, phone number, etc.



- If individual patterns are unique enough, outside information can be used to link the data back to an individual via Matching Attacks.
  - Adversary has some prior (background) knowledge about its target.
  - The attacker's task is to match pieces of information from the first source to pieces of information from the second source that correspond to the same underlying user.



Adversary's Goal

- Identity disclosure: to single out the target in the dataset (also called re-identification).

- Attribute disclosure: to learn anything new about the target.

- Attribute disclosure is stronger:
  - Adversary cannot single out the target (who is 70 years old) but it can learn that the target has cancer.
  - Assuming the target is in the dataset.

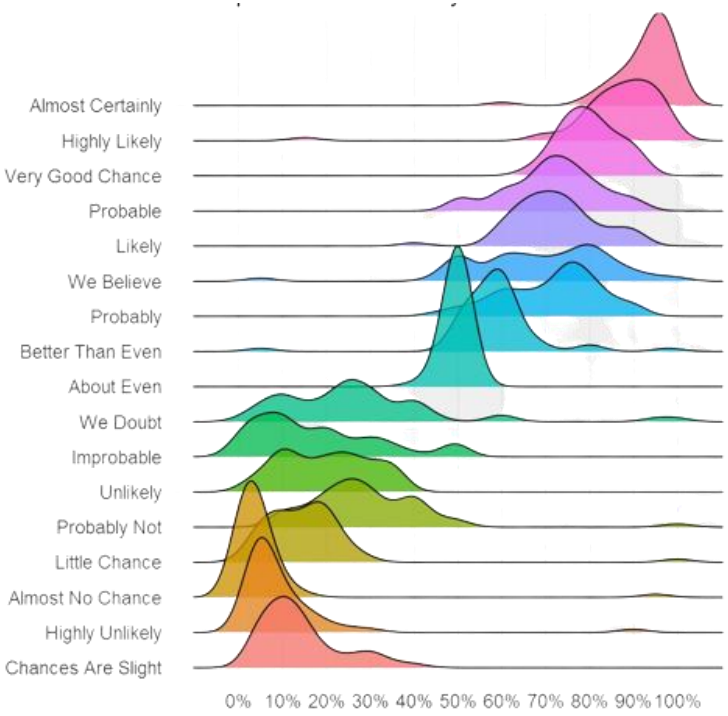
Age	*Sex*	*Diagnosis*
[60,80[	*	Colon cancer
[60,80[	*	Breast cancer
[20,50[	*	Hodgkin disease
[20,50[	*	Breast cancer
[20,50[	*	Colon cancer

GDPR vs Anonymity

- The principles of data protection does not apply to anonymous information.
  - It is critical to understand what the definition of anonymization, as it bears huge legal consequences.
- The logic behind these provisions is that once data are anonymized, they are no longer linkable to a specific subject.
  - Thus, they are no longer considered as personal data.
- However, even an anonymized dataset can be de-anonymized.
  - Re-identification turns non-personal data into personal data, which changes the applicable legal framework.
- The GDPR does not specify what techniques are compliant with the anonymization requirement.

GDPR vs Reidentification

- Data is personal ← if the data subject is identifiable ← if there is a *plausible* attack reidentifying the person and it has a *reasonable* chance of succeeding.
  - Plausible: enough motivation to attack.
- There are no explicit pre-defined thresholds of plausibility and reasonable chance in GDPR (it is context-dependent).
- A successful identity disclosure almost always implies that the data is personal.
- A successful attribute disclosure without reidentification does not necessarily imply that the data is personal.

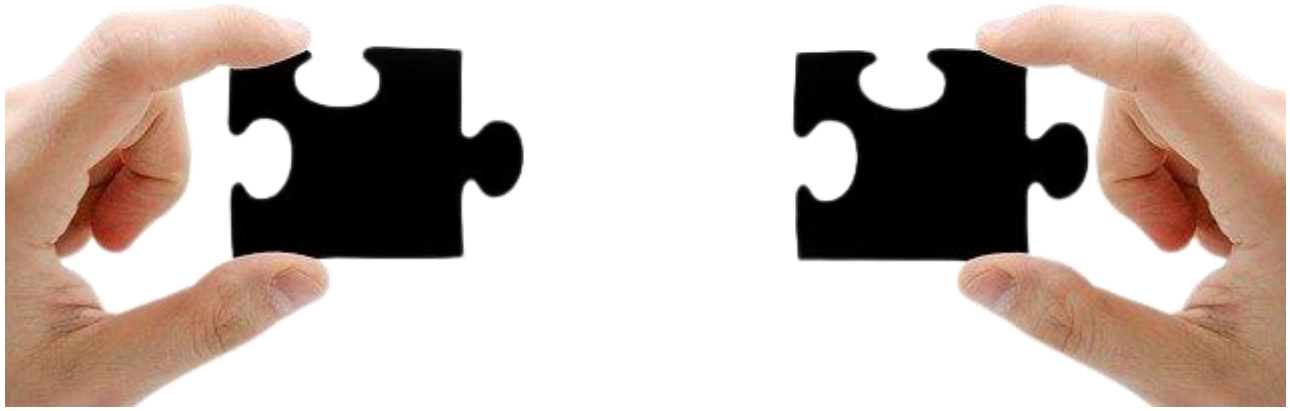


The Opinion

- The Article 29 Working Party issued an opinion in 2014 on anonymity via three essential risks.

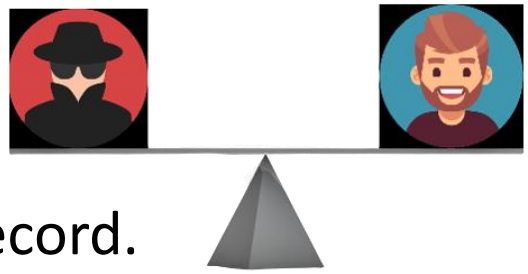
  - Singling out: which corresponds to the possibility to isolate some or all records which identify an individual in the dataset.
  - Linkability: which is the ability to link, at least, two records concerning the same data subject or a group of data subjects.
  - Inference: which is the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.

- The Opinion presents an absolute definition of acceptable risk in the form of *zero risk*, which is practically not achievable.
- The Opinion uses the Likability criterion to evaluate anonymization techniques, which makes longitudinal datasets (e.g., multiple treatment in the hospital, etc.) useless.



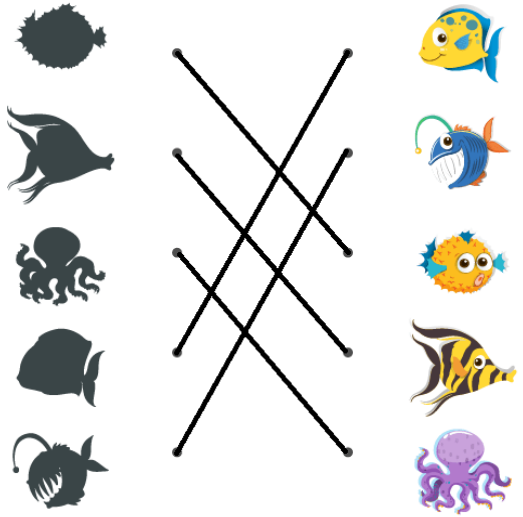
Probabilistic Matching

Matching



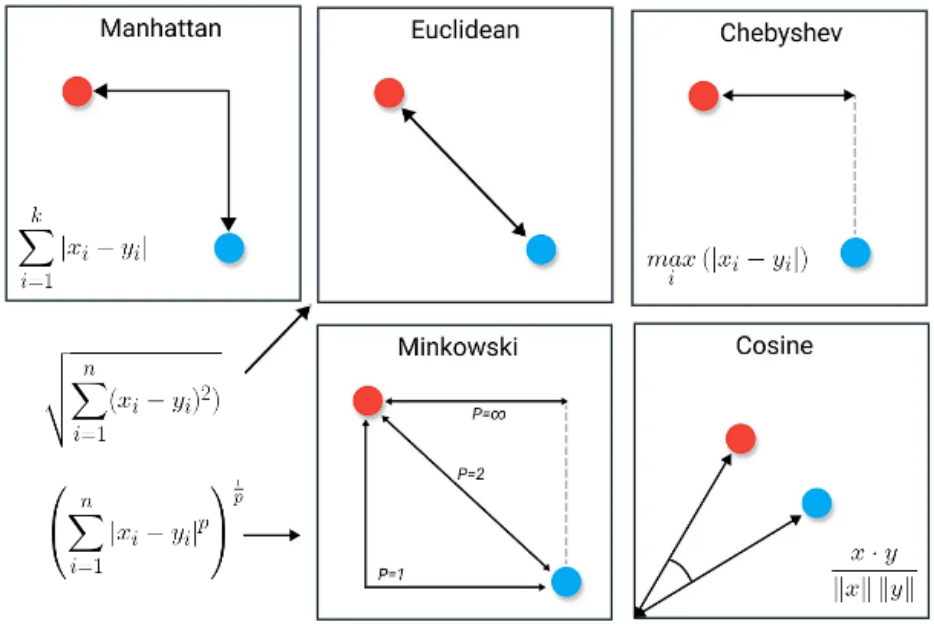
- Adversary compares the background knowledge to every anonymized record and selects the best matching anonymized record.
- The task is to match pieces of information from the first source to pieces of information from the second source that correspond to the same underlying user.

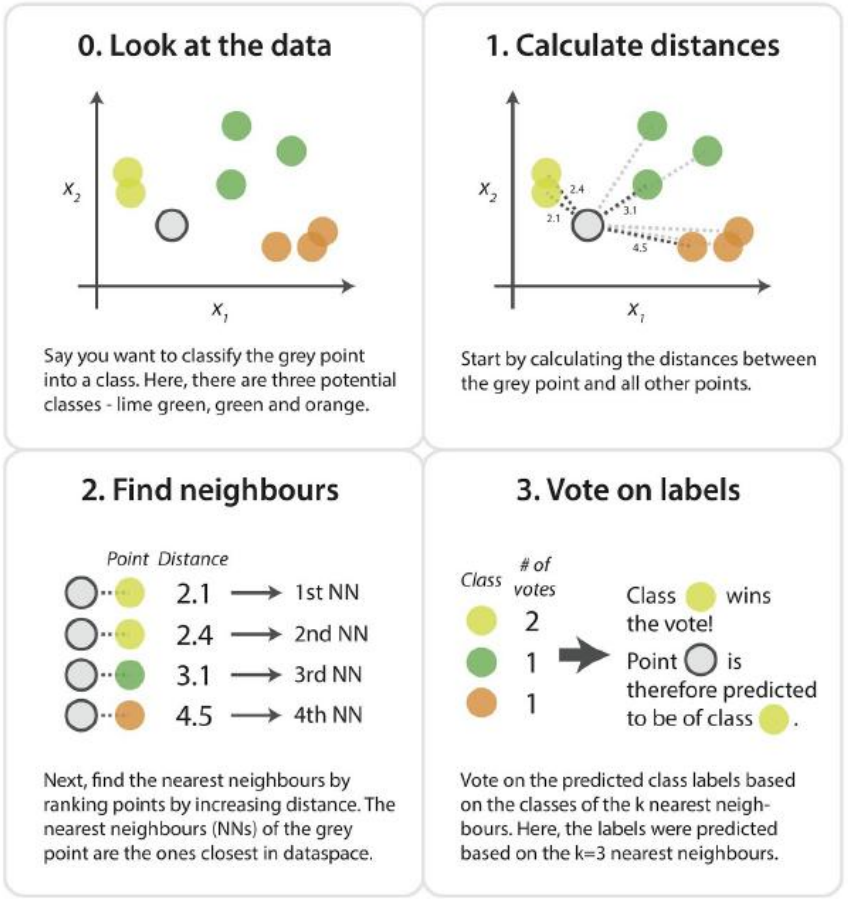
- If an adversary can link a simply anonymized dataset with auxiliary information about the users from publicly available databases, then there is a privacy breach.



k-Nearest Neighbor Classifier

- K-NN is a non-parametric, supervised learning classifier, which uses proximity to make predictions about the grouping of an individual data point.
  - Simple and fast.
  - No need for training.
  - Not so accurate in high dimension.
  - How to choose the distance?





Selecting Distance

- Homogenic (location)
  - Euclidian, Earth Mover Distance

- Heterogenic (gender, job title)
  - Hamming Distance, Mahalanobis Distance
  - You need to assign weights to the attributes.

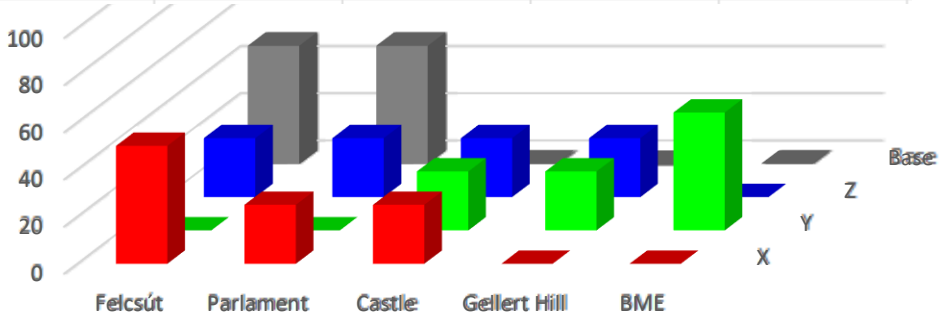
- PCA / Entropy: calculate the variance / entropy corresponding to one attribute.
  - Higher variance / entropy → more important.

- Curse of dimensionality: a record can have many attributes, that constitutes very long vectors, which can cause the vectors to be approximately the same distance from each other.

Example Attack

- Attacker knows the home and work location of the target.
  - Work: Parliament
  - Home: Felcsút

Pseudo ID				
X	Felcsút	Parlament	Castle	Felcsút
Y	BME	Gellert Hill	BME	Castle
Z	Gellert Hill	Parlament	Felcsút	Castle

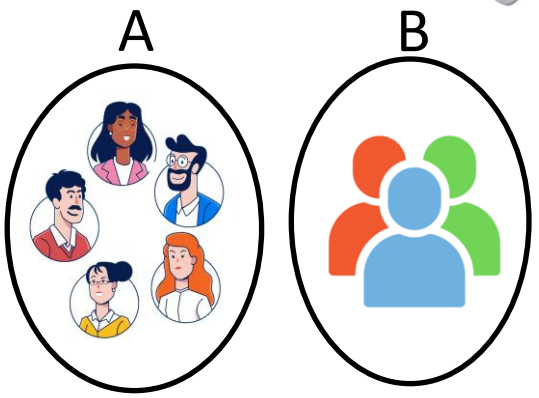


- Adversary represents its background knowledge as a histogram.
- Adversary builds a histogram from each anonymized record.
- Adversary compares its background knowledge with anonymized records using a similarity metric.
  - EMD(X, Base) = 25
  - EMD(Y, Base) = 100
  - EMD(Z, Base) = 50

Y Z

Mass Attack

- Adversary may target multiple users at once.
  - It crawls Facebook and has background knowledge about multiple users in the anonymized dataset.

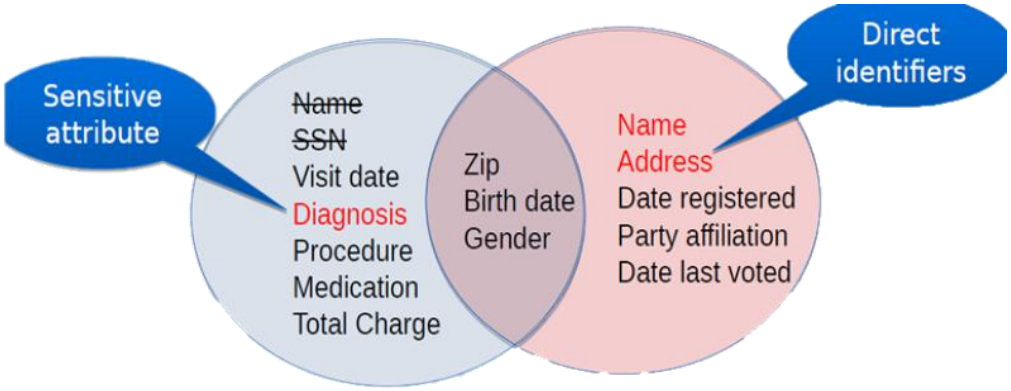


- It wants to know an optimal mapping of background knowledge to anonymized records.
  - If the background knowledge B contains each user from the anonymized dataset A (i.e., $A \subseteq B$), then the assignment problem can be solved in $O(|B| \cdot |E|^2)$ time with the Hungarian algorithm (E=number of edges).
  - If the background knowledge not contain every anonymized user, and the anonymized dataset may have users that are not in the background knowledge, but the adversary has a good estimation of $r = |B \cap A| < |A|$, then it can solve the minimum-weight matching problem with cardinality r (minimum-cost imperfect matching) with an adapted version of the Hungarian algorithm in $O(|A| \cdot |B| \cdot r)$.

Re-identification Attacks

Targeted Attack on Medical Data

- In 1997 the Governor of Massachusetts strongly advocated that anonymized datasets are safe.
- Latanya Sweeney de-identified him amongst the anonymized records by a Matching Attack.
  - She bought voter registration records for $10.
- She showed that data protection techniques used by American public administrations are at stake.
- In 2000 she concluded that 87% of the US population could potentially be identified by their ZIP code combined with gender and date of birth.



Untargeted Attack on Search Queries

- In 2006 AOL publicly released information about users' search history (including queries about political views and medical conditions).

- The web search query logs were pseudonymized, but, when cross-correlated with auxiliary information available from other sources, re-identification was possible.

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.

GO TO LESSON PLAN

Knowledge Tools

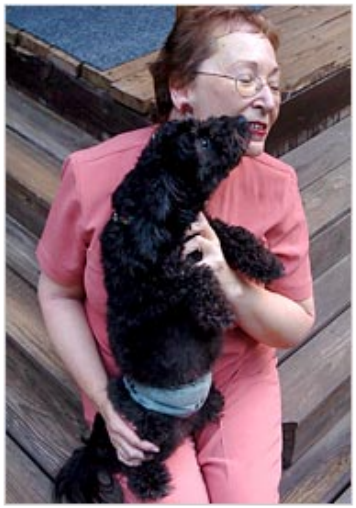
Turn Vocabulary On: Link words to the Merriam-Webster Collegiate® Dictionary.

Turn Geography On: Link countries and states to the Merriam-Webster Atlas®

Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."



Erik S. Lesser for The New York Times

Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, Dudley, who clearly has a problem.

Untargeted Attack on Movie Ratings

- In 2007 Netflix has offered $1,000,000 for a 10% improvement in its movie recommendation system.
  - A pseudonymised training set was released containing information about its users' movie ratings.
  - Without notice to or consent by their subscribers.

DeID 1

Fresh Example

- Vitaly Shmatikov successfully linked 99% of this dataset with IMDB, where some user profiles are public.
  - This led to the disclosure of sensitive information, such as sexual orientations.

- The second Prize competition was cancelled due to a lawsuit and the privacy concerns of the Federal Trade Commission.

Uniqueness of Top-K Data

- In 2011 researchers used 3 months of call data with 25 million users and 30 billion calls.

Anonymity set with top 1 location

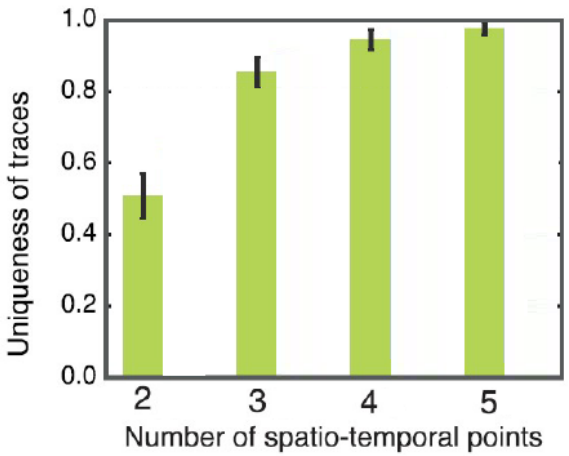
Location granularity	Size of anonymity set
	Median
Sector	372
Cell	967
Zip code	3125
City	7638
County	55649
State	7.2e+05

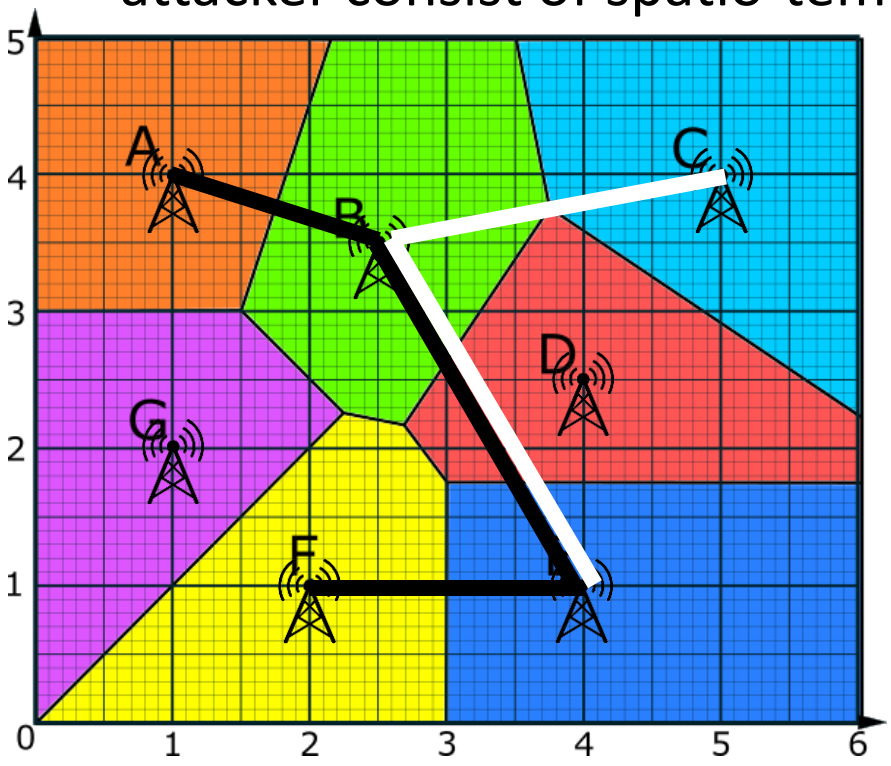
- Coarsing the data (in either the time or the space domain) has limited effect while decrease the utility significantly.

- Publishing anonymized location data will lead to inevitable privacy risks.

Uniqueness of Location Data

- GPS coordinates are shared exhaustively, e.g., app requests, geotagged posts, etc.
- In 2013, researchers used 15 months of mobility data from 1.5M people.
- The information available to the attacker consist of spatio-temporal point.



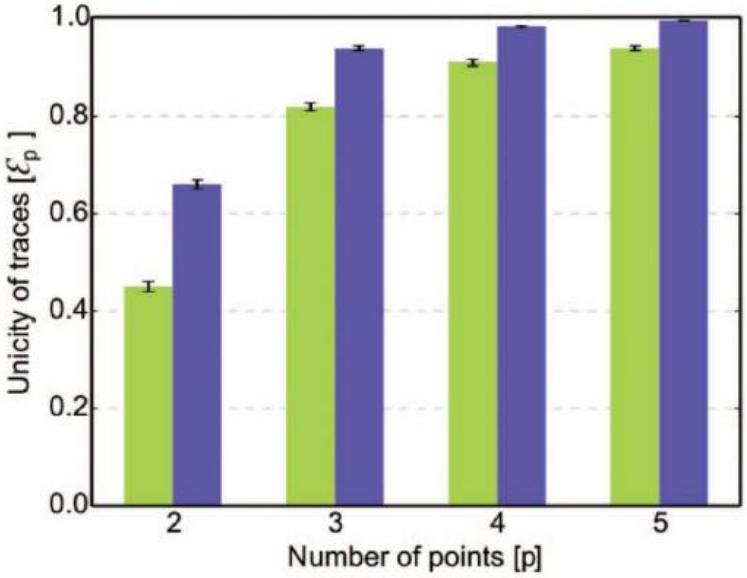


  - A - 8:00, B - 9:30, E - 14:45, F - 15:15
  - C - 8:40, B - 9:20, E - 12:00

- Four randomly chosen points are enough to uniquely characterize 95% of the users, whereas two randomly chosen points still uniquely characterize more than 50% of the users.

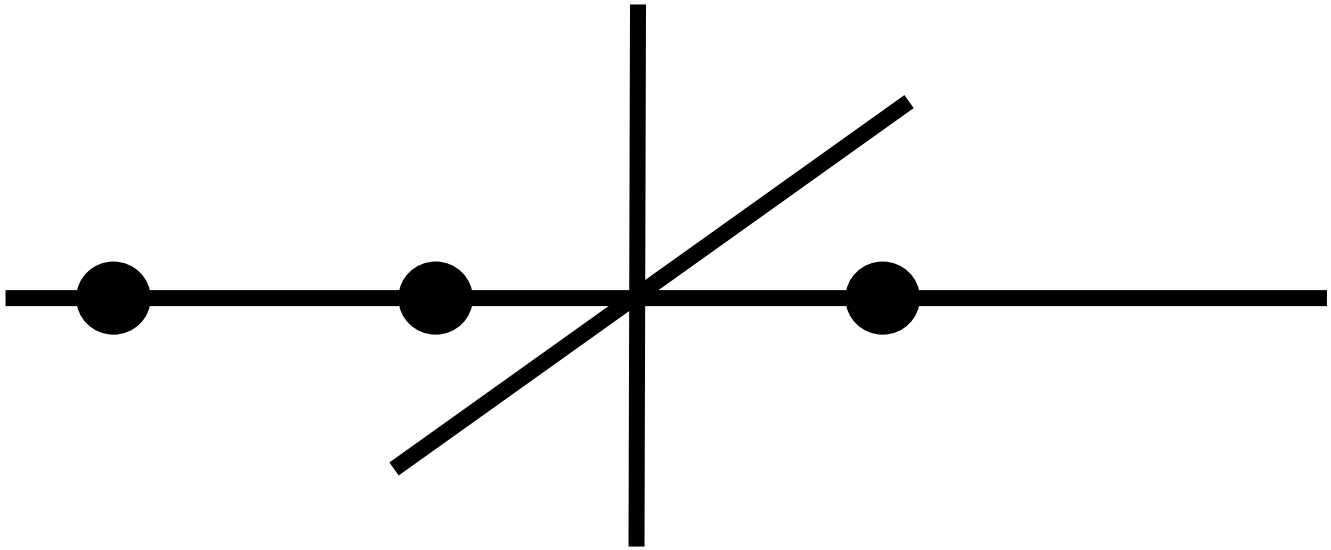
Uniqueness of Transaction Data

- In 2015 researchers used 3 months of credit card transactions for 1.1 million users in 10,000 shops.
- Simply anonymized, price tags are binned, time-stamps coarsed to a resolution of 1 day.
- The green bars represent unicity when random spatio-temporal tuples are known.
  - This shows that four spatio-temporal points taken at random are enough to uniquely characterize 90% of individuals.
- The blue bars represent unicity when price is also considered.



Curse of Dimensionality

- When the dimensionality increases, the volume of the space increases so fast that the available data become sparse.
- In high dimension, all records appear to be dissimilar in many ways, which assist re-identification.
  - If there are many attributes in a dataset, it is more likely that the adversary knows some of them that are enough to re-identify someone.



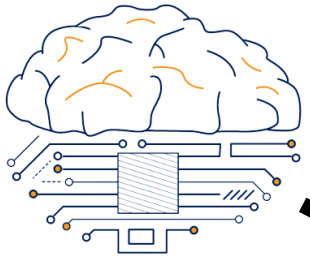
Reidentification via Machine Learning

- Modelling correlation between features is difficult, but machines can learn these patterns automatically.
  - If there are limited training data, use k-NN.
  - If multiple samples are available about a target person, use neural networks, etc.
- Train a classifier on the background knowledge to predict the user's identity (or one of its attribute) and test the model on the anonymized data.

Anonymized Data

Background Knowledge

Training



Machine Learning Model

Prediction

Attacking Anonymized Social Graphs

- In 2009 researchers used Twitter (as anonymized) data with 224 thousand users and 8.5 million connections.
- They used Flickr (as background knowledge) data with 3.3 million users and 53 million connections.
- The attacker solely relies on network topology.
  - 31% of the mappings were re-identified correctly.
  - 5% were mapped to nodes at distance 1 from the true mapping.
  - 6% were mapped to the same geographic location category.

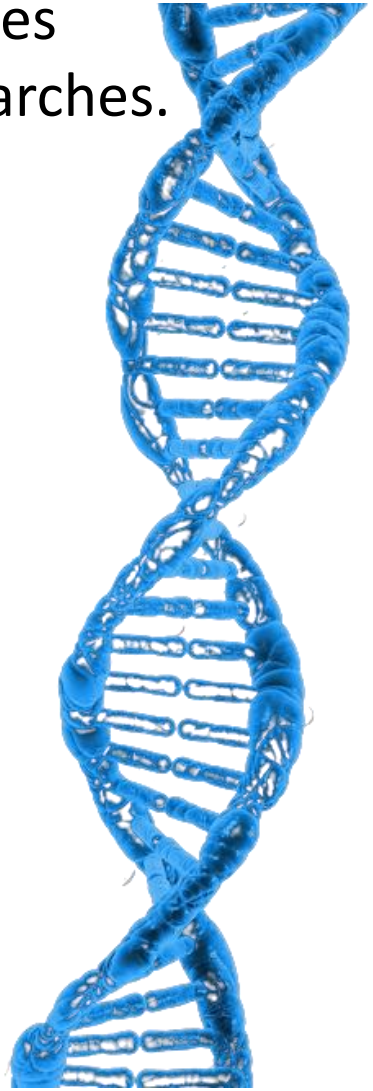
DeID 2

Fresh Example

- In reality, anonymized graphs are usually released with attributes in their nodes and edges, making deanonymization even easier.

Attacking Anonymized DNA

- Law enforcement agencies are using genetics databases to identify anonymous DNA via long-range familial searches.
- If combined with a genealogical database of online ancestry services, it is extremely powerful.
- Via genealogical triangulation, over 50% of anonymous DNA can be identified (matched to the correct individual or same-sex sibling) when the genetic database includes just 1% of the population.
- Snowball identification: a successful identification adds to the genetic genealogical database, increasing the identification accuracy for future instances.



- Researchers (in 2019) used artificial simulated datasets.

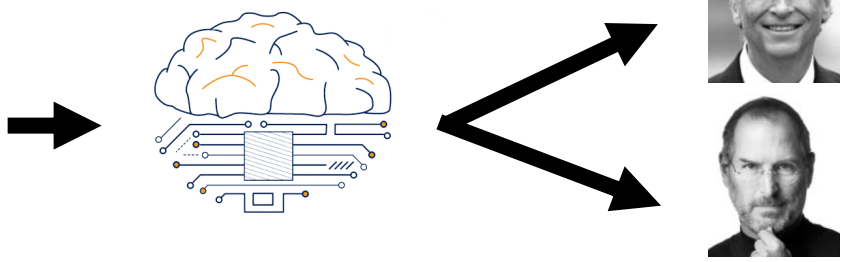
23andMe

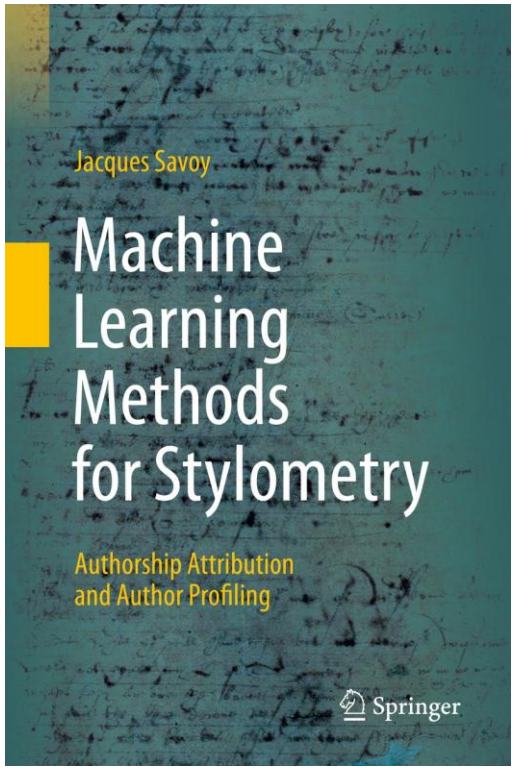
De-anonymization of Unstructured Data

Stylometry

- Stylometry is the analysis of style features that can be statistically quantified, such as sentence length, vocabulary diversity, and frequencies (of words, word forms, etc.).
- Stylometry can be used to differentiate texts, codes, etc.
- Matching anonymized data with background knowledge is possible as well.

```
for i in people.data.users:
    response = client.api.statuses.user_timeline.get(screen_name=i.scre
    print 'Got', len(response.data), 'tweets from', i.screen_name
    if len(response.data) != 0:
        ltdate = response.data[0]['created_at']
        ltdate2 = datetime.strptime(ltdate,'%a %b %d %H:%M:%S +0000 %Y'
        today = datetime.now()
        howlong = (today-ltdate2).days
        if howlong < daywindow:
            print i.screen_name, 'has tweeted in the past' , daywindow,
            totaltweets += len(response.data)
            for j in response.data:
                if j.entities.urls:
                    for k in j.entities.urls:
                        newurl = k['expanded_url']
                        urlset.add((newurl, j.user.screen_name))
    else:
        print i.screen_name, 'has not tweeted in the past', daywind
```

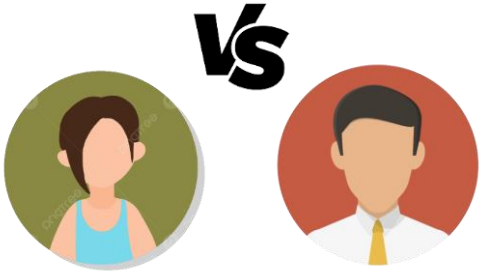


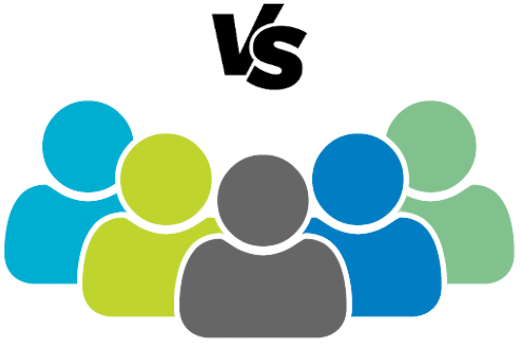


- Many programmers have unique coding style.
  - Preference for spaces over tabs, while loops over for loops, etc.
  - The model can distinguish programmers present in the training data.

Options

- One-vs-One
  - Train a model per user pairs to distinguish the two users.
  - There are two output classes of the model (one per user)
- One-vs-Rest
  - Train a model per user to distinguish the user from the rest.
  - There are two output classes of the model (one for the user and one for the rest)
- All-vs-All
  - Train a single model to distinguish all users at once.
  - There are multiple output classes of the model (one per user).

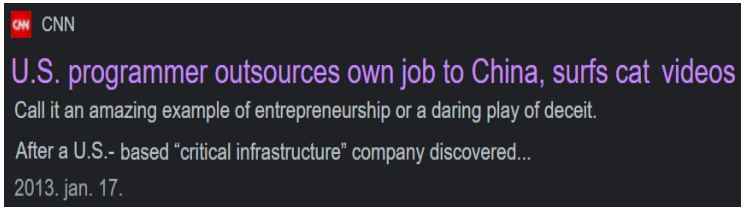




Use Cases

- Programmer de-anonymization: attribute an anonymous codes to known programmers, e.g., finding Satoshi Nakamoto …
- Copyright investigation: decide the validity of an ownership claim of developed software. ©
  - One vs One

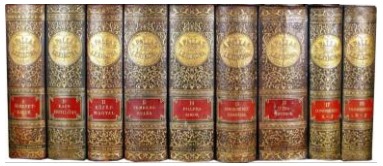
- Authorship verification: the code may not be written by the programmer.
  - One vs Rest



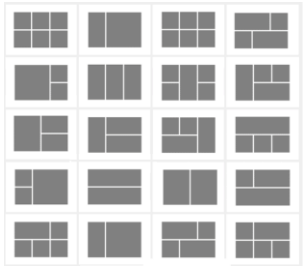
- Software forensics: attribution of malware code to a set of known candidate programmers.
  - All vs All with an extra class (as the code might not be written by the candidates).

- Ghostwriting: is your homework written by another student?

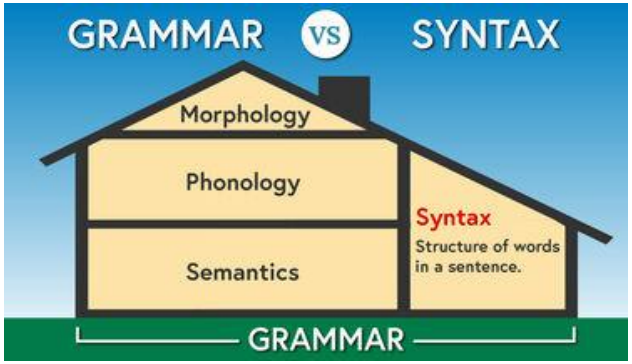
Feature Extraction



- Lexical features
  - Obtained directly from source code.
  - E.g., frequency of functions, operators, comments, variables, macros, average number of parameters of functions, nesting depth of loops, etc.



- Layout features
  - Obtained directly from source code.
  - E.g., ratio of white spaces/tabs to the file size, number of empty lines, newline between new code blocks, etc.
- Syntactic features
  - Properties of the language dependent abstract syntax tree (AST), and keywords.
  - Grammatical features, invariant to changes in source-code layout.
  - E.g., depth of the AST, frequency of AST bigrams, etc.



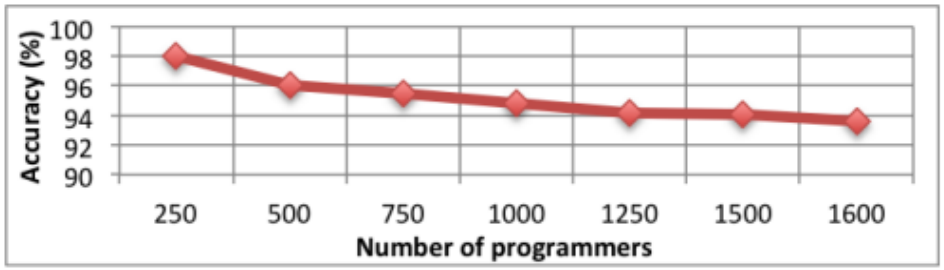
Abstract Syntax Tree

- A tree representation of the abstract syntactic structure of source code written in a formal language.
- Each node denotes a construct.

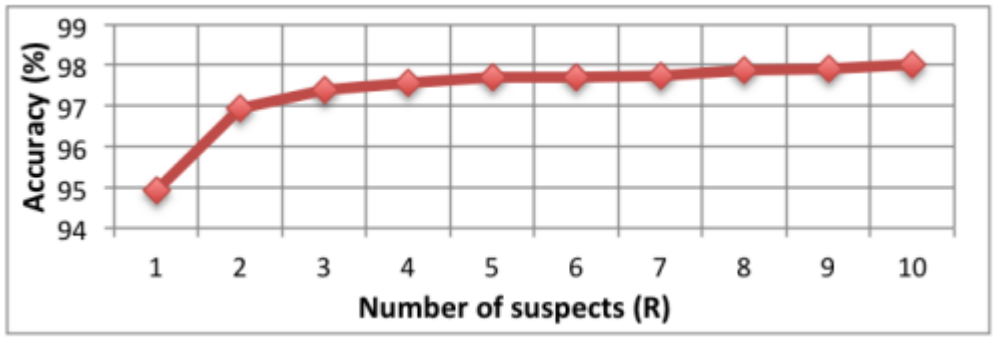
```
while b ≠ 0:
    if a > b:
        a := a - b
    else:
        b := b - a
return a
```

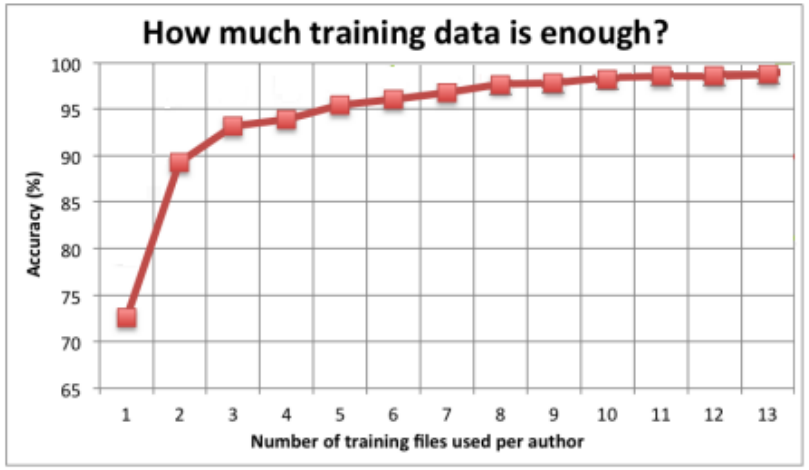
Attack on Code

- In 2015 researchers used C/C++ code from Google Code Jam from 250 programmers, each with 9 samples, consisting in average 70 lines of code.
  - Lexical: ca. 55,000
  - Layout: 6
  - Syntactic: ca. 60,000



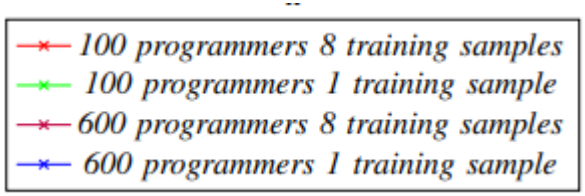
- The model (random forest) achieved 95% of accuracy for programmer de-anonymization.

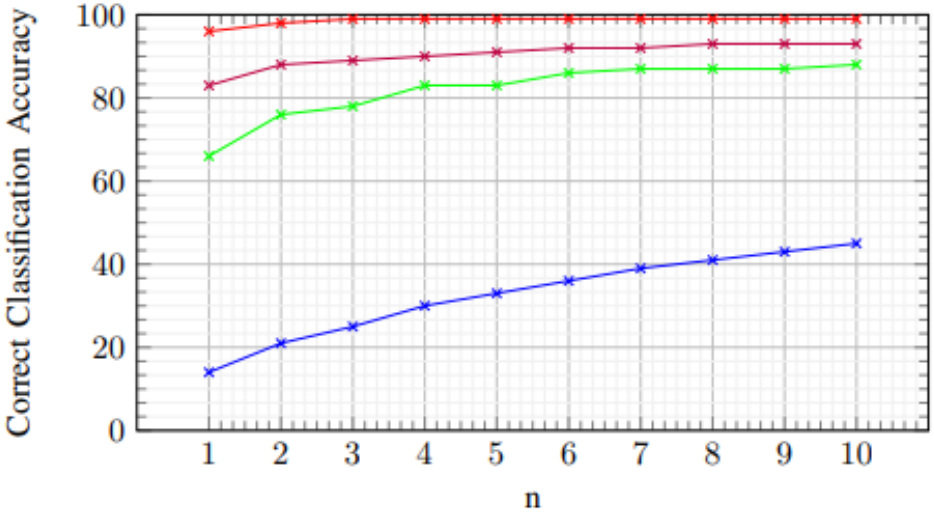


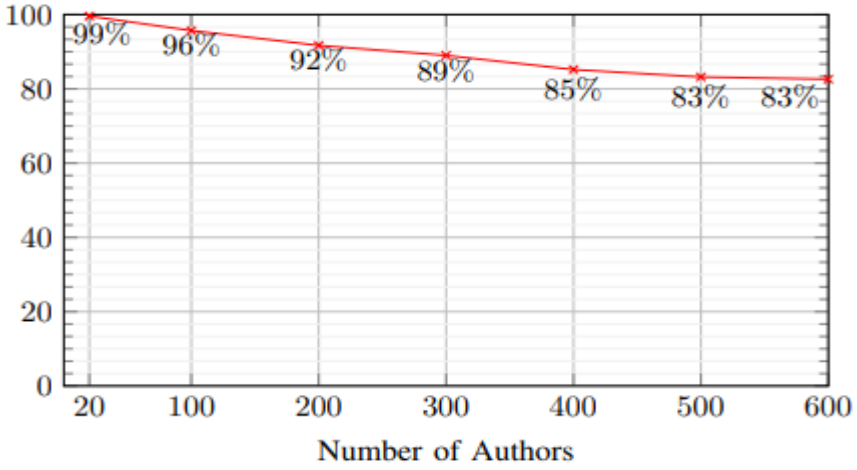


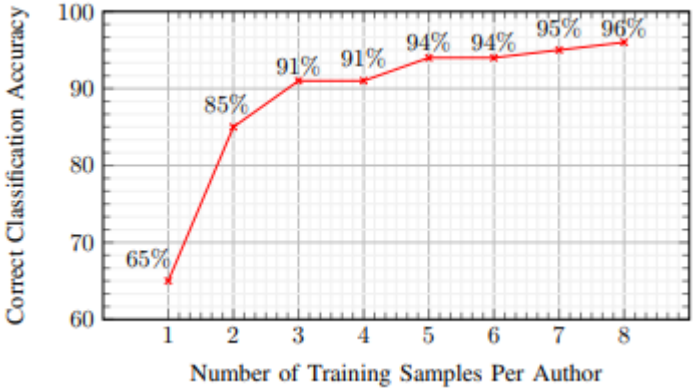
Attack on Binaries

- In 2015 researchers used C/C++ code from Google Code Jam from 100 programmers, each with 9 binary samples.









Number of Programmers	Number of Training Samples	Compiler Optimization Level	Accuracy
100	8	None	96%
100	8	1	93%
100	8	2	89%
100	8	3	89%

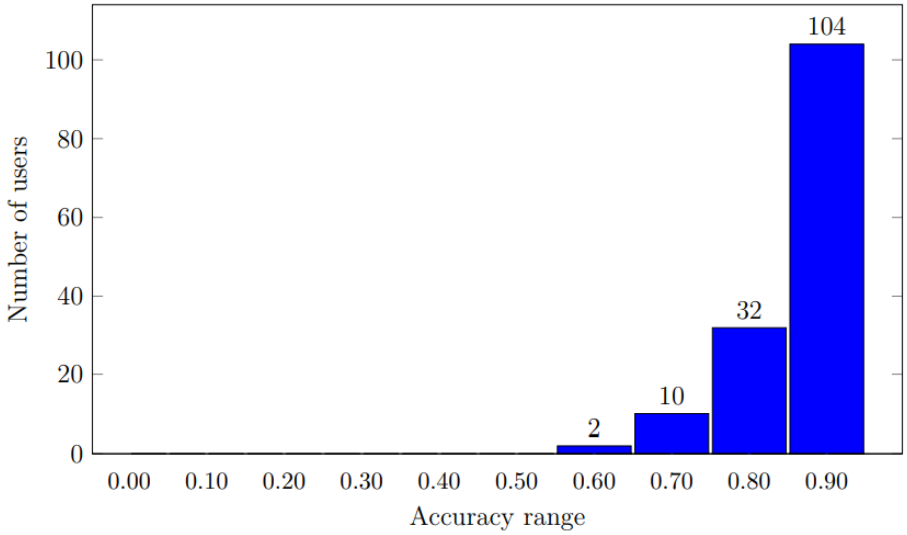
Attack on Typing Style

- In 2023, 148 programmer typing style was used.
  - Time of push down keys
  - Time between pushing down different keys
  - Time between letting go keys

- Only 8 times was the detection below 75%, 36 times it was between 75-90%, and for the rest 104 cases it was over >90%.

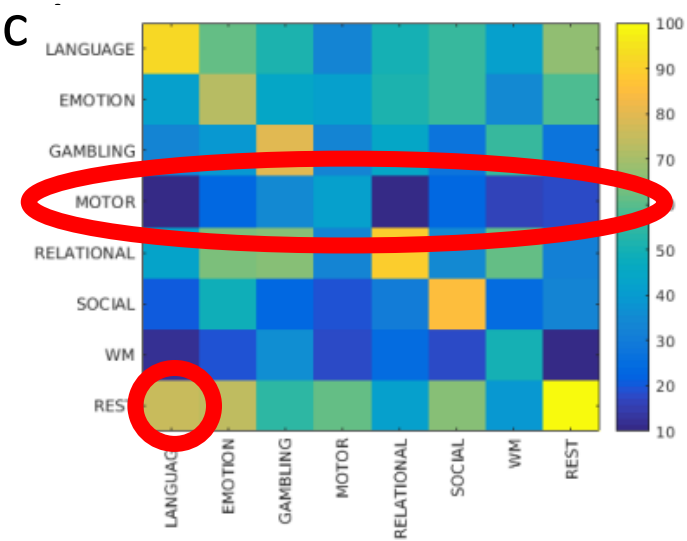
- Similar result is expected on touchscreen devices.

- Such uniqueness can also be used as authentication methods.



Deanonymizing MRI Images

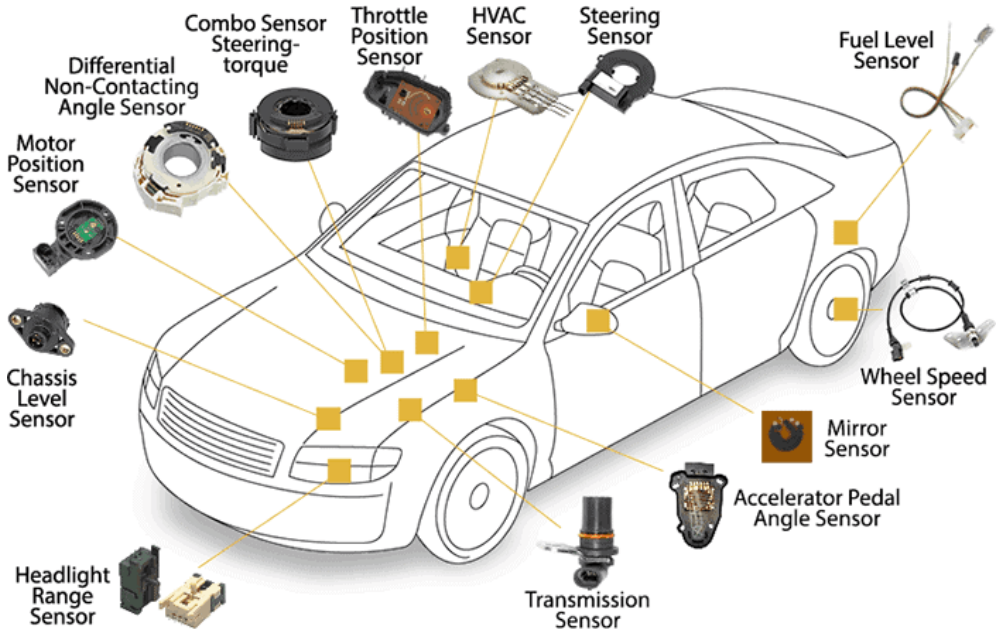
- In 2021 researchers used multiple MRI scans from 362 children diagnosed with Attention Deficit Hyper-active Disorder (ADHD), along with 585 controls while the subjects were doing various tasks.
  - REST is the background knowledge; LANGUAGE is the anonymized data: matching the corresponding subjects for these different tasks is possible with 94% accuracy.
  - MOTOR tasks are ineffective in predic even for the same task.
  - The matrix is asymmetric, implying that de-anonymization of different tasks has varying impact on the anonymity of other datasets.



Deanonymizing Driving Data

- A Controller Area Network (CAN bus) is a vehicle bus standard designed to allow microcontrollers and devices to communicate with each other.

- Driving the same car in a constrained parking lot setting and a longer fixed route with all sensor signals available from manufacturer's documentation.
- In 2016 researchers achieved 100% re-identification rate of 15 drivers.
- Car manufacturers do not reveal the exact signal location within CAN logs.



CrySyS Research (2019)

- Extract signals from CAN logs using ML.
  - Train and test data come from different manufacturer's car.
- Statistics of extracted features used for driver re-identification: achieved 70-87% accuracy with 33 drivers.
- Reidentify drivers directly from the logs: 85% re-identification based on 2 minutes of driving in urban areas (even if drivers follow different routes).
  - Drivers with substantial experience are the easiest to distinguish.

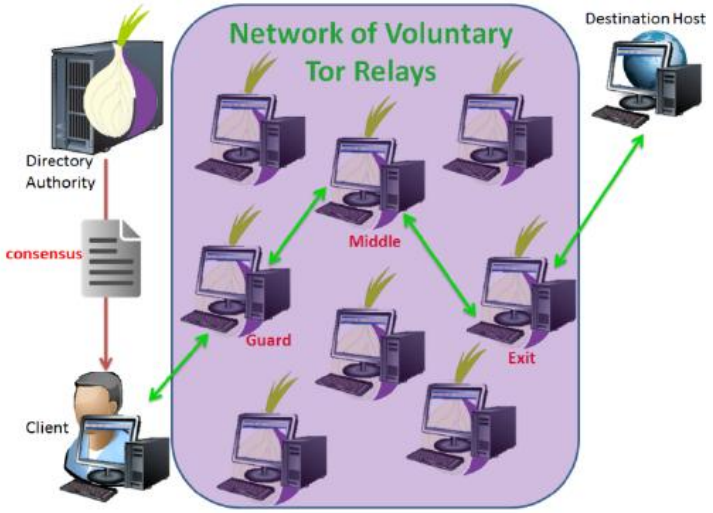
ID	DLC	DATA	Timestamp
05f0	2	00 00 0e 00 00 00 00 00	2.084334
04f0	8	00 00 00 00 00 00 00 00	2.116588
0690	8	00 00 00 00 00 00 00 00	2.124836
04f0	8	00 00 00 00 00 00 00 00	2.126036
05f0	2	00 00 00 00 00 00 00 00	2.134353
0690	8	00 00 00 00 00 00 00 00	2.135407
04f0	8	00 00 00 80 00 eb b6 13	2.144996
0130	8	00 00 40 ff 00 00 41 3d	2.156946
0131	8	00 00 40 00 00 00 41 9b	2.157975
0140	8	00 00 00 00 02 29 21 f0	2.158382
04f0	8	00 00 00 80 00 eb b6 13	2.165245
0130	8	00 00 40 ff 00 00 42 1a	2.176891
0131	8	00 00 40 00 00 00 42 bc	2.177941
0140	8	00 00 00 00 04 25 22 a5	2.178345
05f0	2	00 00 00 00 04 25 22 a5	2.184367
04f0	8	00 00 00 80 00 eb b6 13	2.185408
0130	8	00 00 40 ff 00 00 43 07	2.196924
0131	8	00 00 40 00 00 00 43 a1	2.197951
0140	8	00 00 00 00 06 26 23 6f	2.198356
04f0	8	00 00 00 80 00 eb b6 13	2.205049
0430	8	00 00 00 00 00 00 00 00	2.206446
04b1	8	00 00 00 00 00 00 00 00	2.207510
01f1	8	0f 00 00 00 00 00 00 00	2.207884
0153	8	00 00 00 ff 00 ff 00 00	2.208233
0002	8	00 00 00 00 00 08 00 50	2.208581
02b0	5	57 00 00 07 50 08 00 50	2.210872
0002	8	00 00 01 00 00 00 01 27	2.216572
0153	8	00 00 00 ff 00 ff 00 00	2.217625

DeID 3

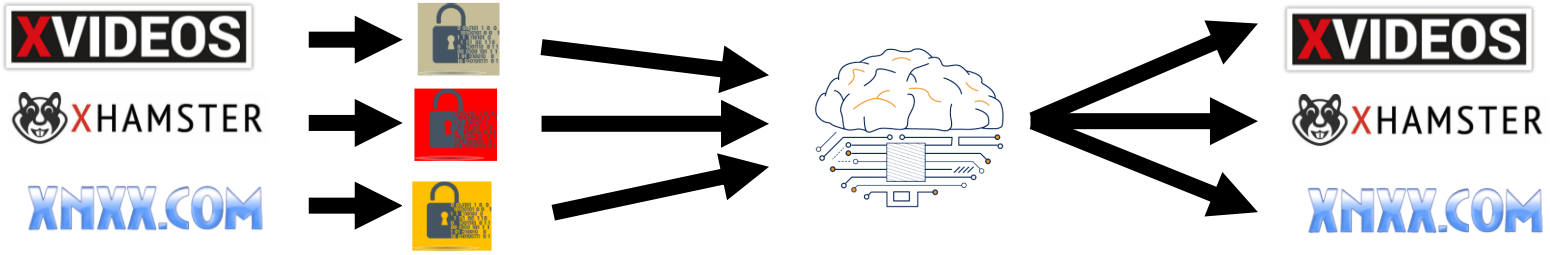
Another Example

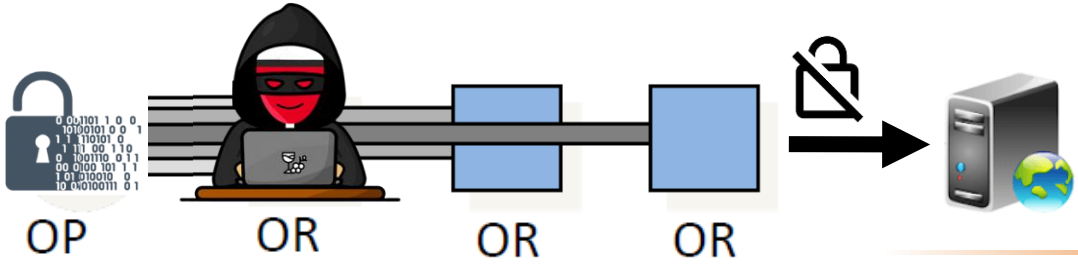
TOR

- Tor is an anonymous communication service which makes compromise between anonymity and speed.
  - Does not alter the size, direction, and timing of packets.



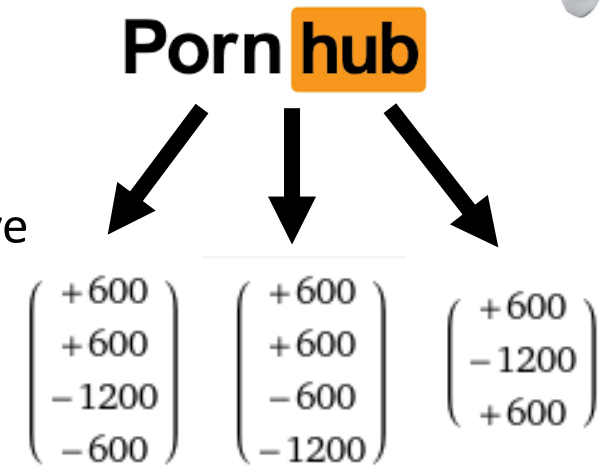
- Web pages could be fingerprinted.
  - Does not break the encryption!

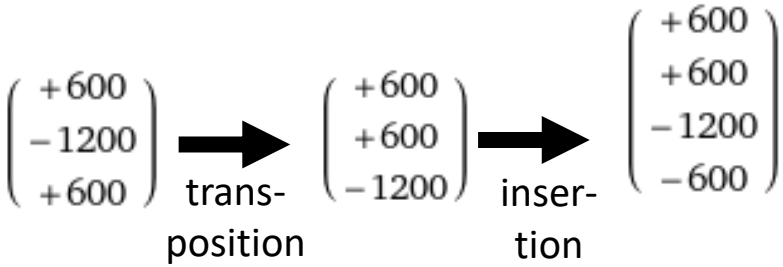




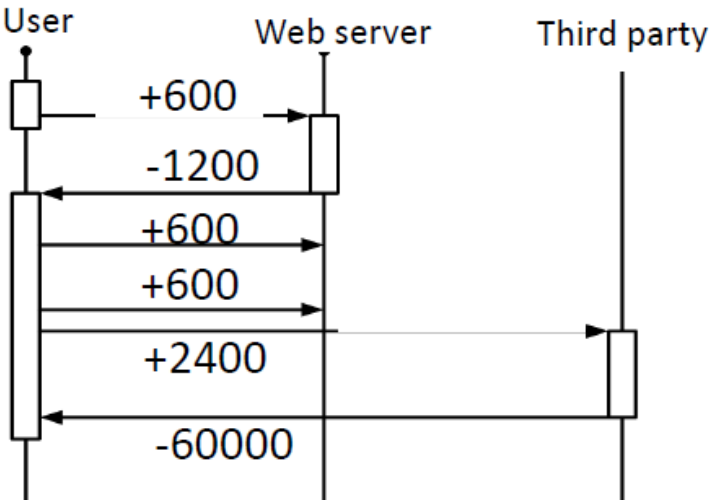
Attack on Page Visits

- Differentiable factor: inherent stability in the ordering of requests.
  - Browsers cannot request an object until they have received the portion of a page that references it.
- The features are the encoded packet sizes and directions.
  - Due to packet re-ordering / dropping the web-page can have features vectors with different sizes.
- Damerau–Levenshtein edit distance
  - DL(a ,b): the length of the shortest sequence of insertions, deletions, substitutions, and transpositions required to transform a to b.



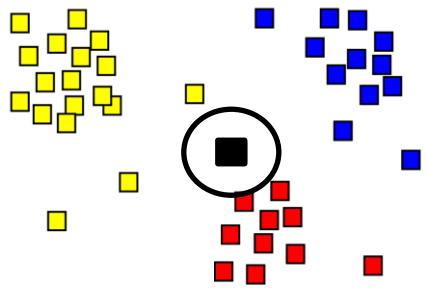


DL = 2



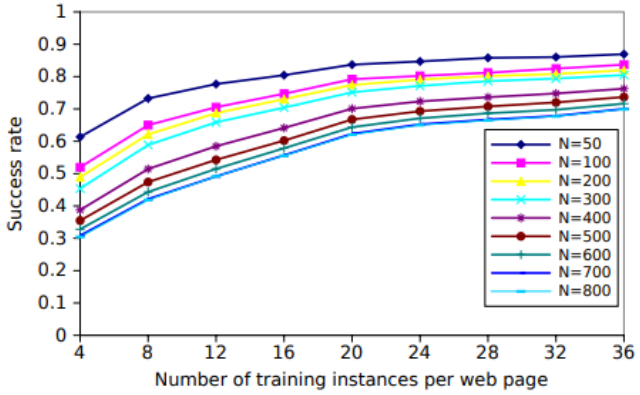
Deanonymizing Page Visits

- The adversary collects n vectors (loads) of every web-page, and classifies them using 1-NN.

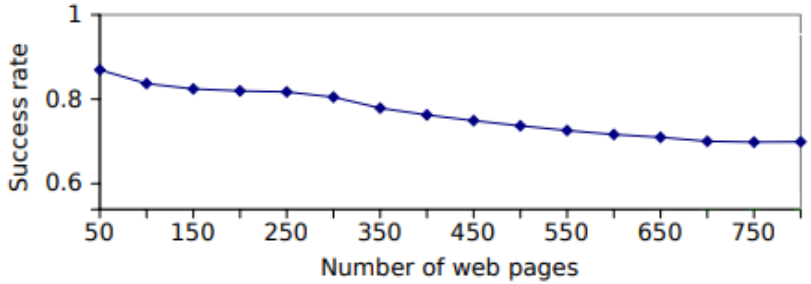


Defense	Rate
SSH tunnel	91.6%
Tor	83.7%

- Closed Word (2012)
  - Researchers used 100 pages.
  - Attack aims to identify the target web page.



- Follow-up study (on TLS) scaled the attack to thousands of webpages.
  - Consequently, TLS only protects private information (e.g., credit card numbers) but not the browsing habits.



Take Away

- All data type can be attacked.
  - Targeted / Untargeted / Mass attack
  - Structured Micro Data / Unstructured Data
- GDPR: if de-anonymization is reasonably possible, the data is personal (even it is encrypted).
  - By linking it to a physical person.
- Systematic (and automatic) de-anonymization is done by machine learning.
  - User is linked to anonymous data with a classifier.
  - One vs One / One vs Rest / All vs All

Control Questions

- Give an example of how the entry node can figure out the webpage the user is visiting in TOR!

- What are singling out, linkability, and inference?

- What kind of features can be extracted from program codes? Give examples for all three!

References

- k-anonymity: a model for protecting privacy
- Simple demographics often identify people uniquely
- How To Break Anonymity of the Netflix Prize Dataset
- Unique in the crowd: The privacy bounds of human mobility
- Unique in the shopping mall: On the reidentifiability of credit card metadata
- Anonymization of location data does not work: A large-scale measurement study
- De-anonymizing social networks
- Identification of anonymous DNA using genealogical triangulation
- De-anonymizing programmers via code stylometry
- When coding style survives compilation
- Keystroke Dynamics for User Identification
- De-anonymization attacks on neuroimaging datasets
- Extracting vehicle sensor signals from CAN logs for driver re-identification
- Automatic driver identification from in-vehicle network logs
- Touching from a distance: Website fingerprinting attacks and defenses
- Adaptive Webpage Fingerprinting from TLS Traces
- Deanonymizing Tor hidden service users through Bitcoin transactions analysis