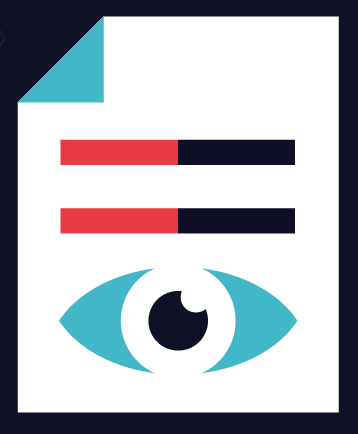
CAP5768/IDC4140:

Introduction to Data Science

Guang Wang Department of Computer Science

Florida State University

# What is Privacy



* Wikipedia  *“ability of an individual or group to seclude themselves or information about themselves, and thereby express themselves selectively”*
* Merriam-Webster  *“freedom from unauthorized intrusion” , “the quality or state of being apart from company or observation”*

# Defining Privacy

* Private information sometimes equated to **personal data** / **personally identifiable information (PII)**
  + *“(1) any* ***information that can be used to distinguish or trace an individual's identity****, such as name, social security number, date and place of birth, mother's maiden name, or biometric records;*
  + *(2) any other* ***information that is linked or linkable to an individual****, such as medical, educational, financial, and employment information." [1]*

***What should be considered private?***

1. *NIST: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf*

# Why Data Privacy it Important?

* + Data privacy is a guideline for how data should be collected or handled, based on its sensitivity and importance. Data privacy is typically applied to personal health information (PHI) and personally identifiable information (PII). This includes financial information, medical records, social security or ID numbers, names, birthdates, and contact information.
  + Data privacy concerns apply to all sensitive information that organizations handle, including that of customers, shareholders, and employees. Often, this information plays a vital role in business operations, development, and finances.
  + Data privacy helps ensure that sensitive data is only accessible to approved parties. It prevents criminals from being able to maliciously use data and helps ensure that organizations meet regulatory requirements.

# Privacy Principles

**Minimization**: what you collect, who has access to it, and how long you keep it

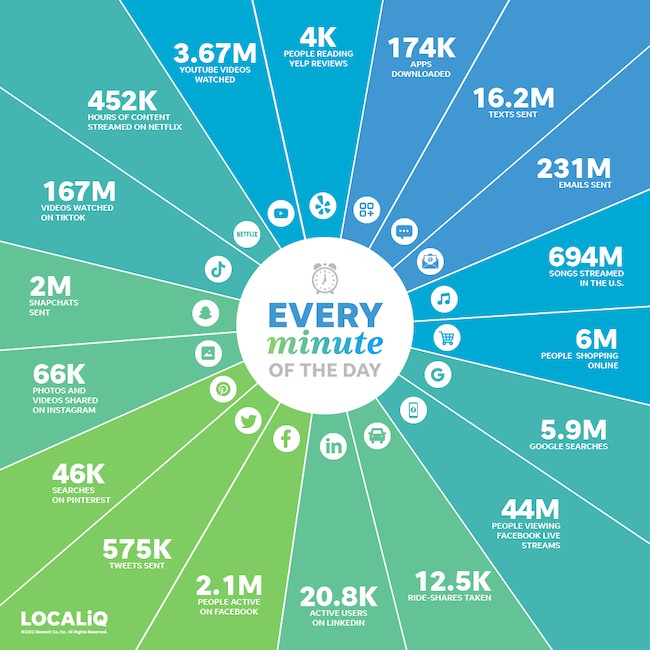
**Choice**: user’s option to share information 6543tgrepcheck2&&\*\*(11

**Access**: user’s right to review, correct and possibly delete the data you hold

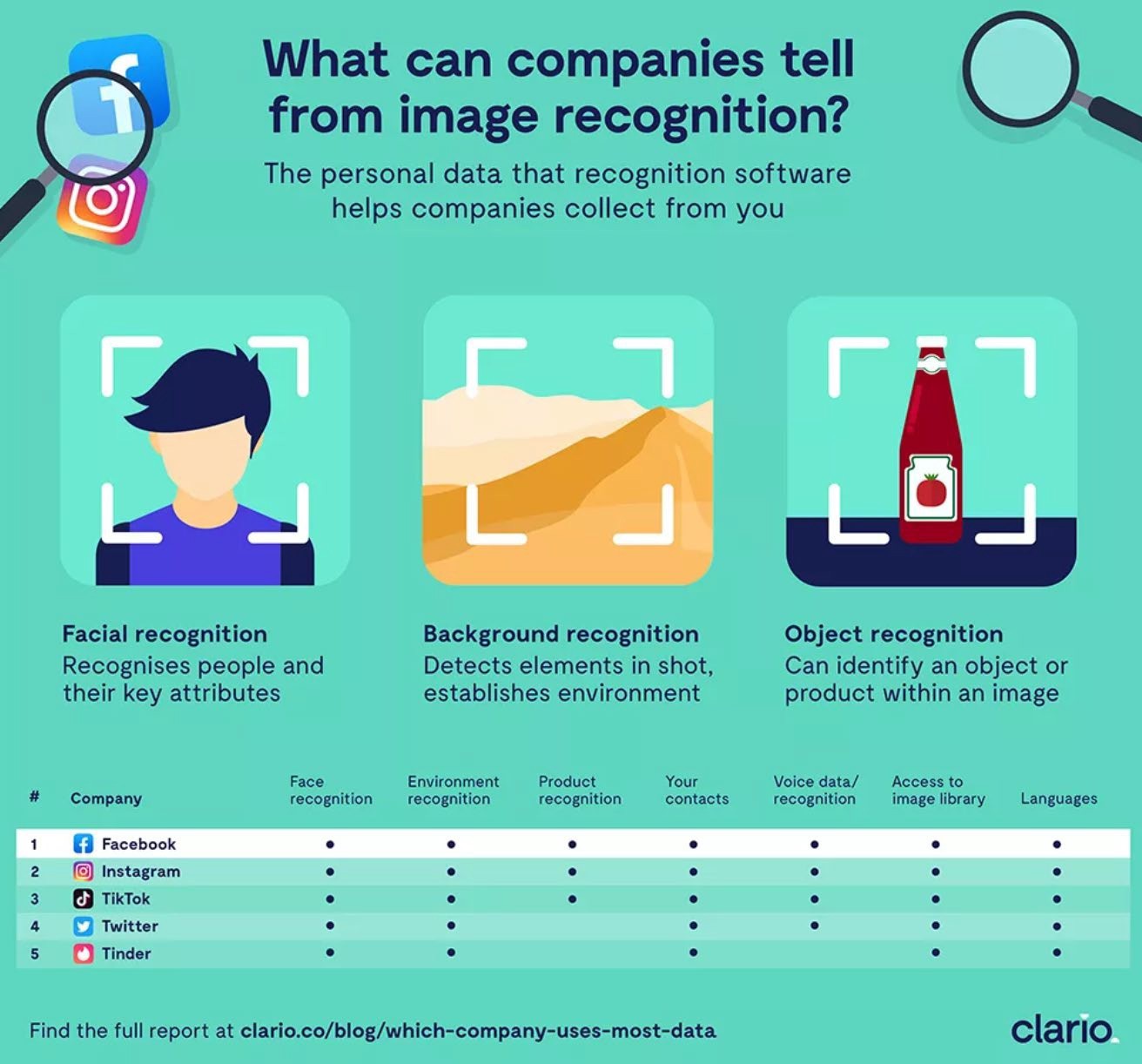
**Transparency**: disclosure of what, how and with whom data will be shared

# Do your care about Data Privacy?

Today’s Privacy Challenge



*Source: https://localiq.com/blog/what-happens-in-an-internet-minute/*



# Where is all this data coming from?



Where is all this data coming from?

* Census surveys
* IRS Records
* Medical records
* Insurance records
* Vehicle Registration
* Search logs
* Browse logs
* Shopping histories
* Photos
* Videos
* Smart phone Sensors
* Mobility trajectories
* Smartcard payment
* Credit card payment
* …

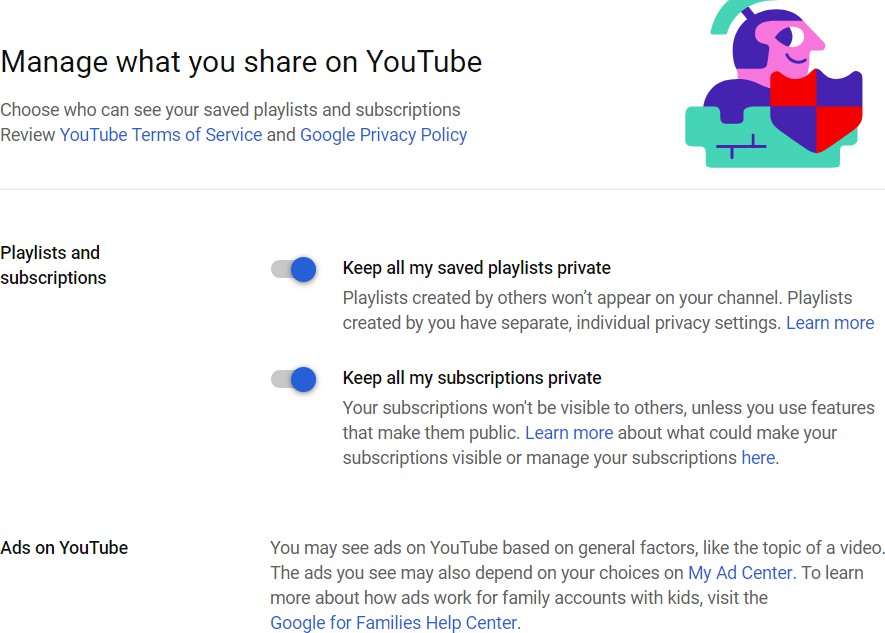
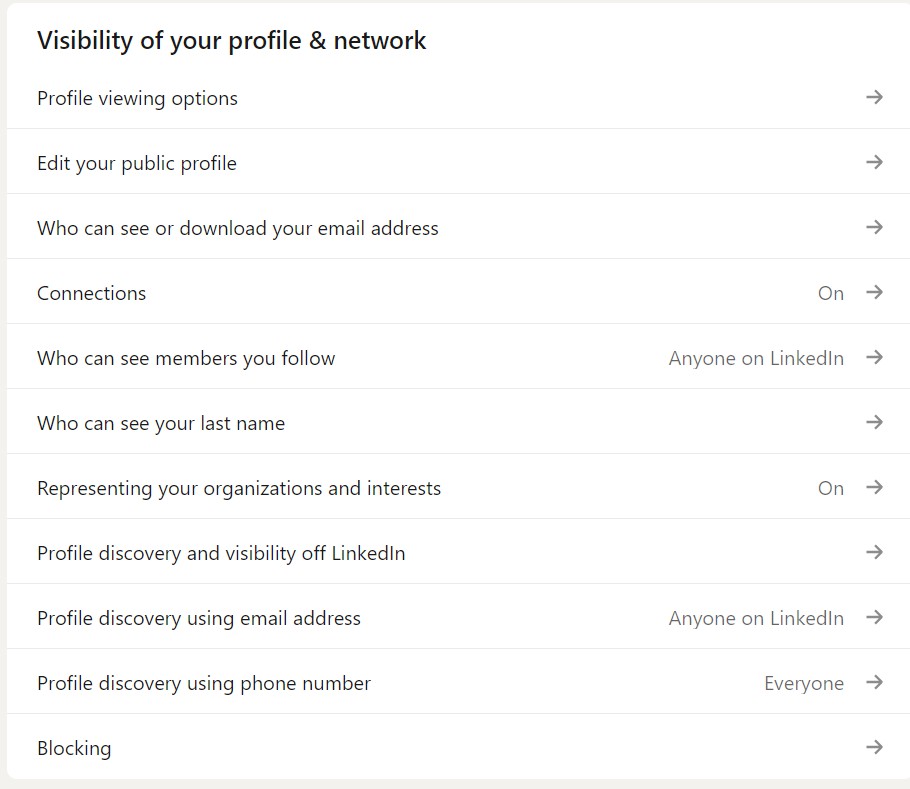
# Data Is A Double-Edged Sword



Source: https://jacobsmedia.com/why-data-is-a-double-edged-sword-for-radio/

# Data Privacy

Sometimes users can know and control who sees their information



# Data Privacy

But not always!!! 6543tgrepcheck2&&\*\*(11



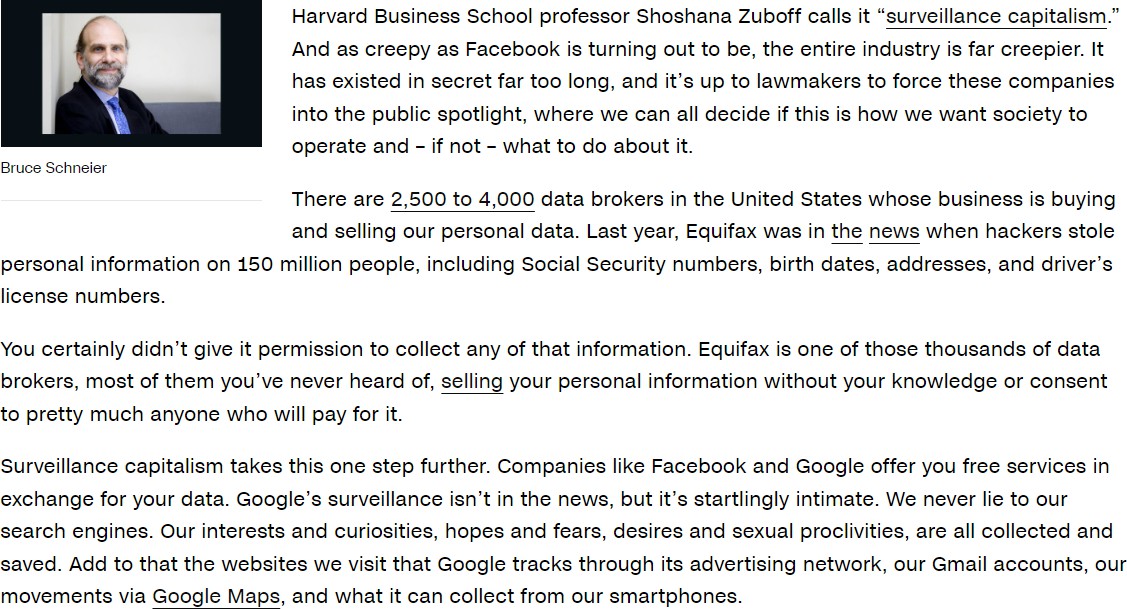
Protecting Your Digital Life Covers your webcam when not in use and prevents web hackers from spying on you.

# Data Privacy: Tracking

https://[www.digitaltrends.com/web/top-100-websites-how-are-they-tracking-you/](http://www.digitaltrends.com/web/top-100-websites-how-are-they-tracking-you/)

Top 100 websites: How they track your every move online

# Data Privacy: Sharing of Data



https://edition.cnn.com/2018/03/26/opinions/data-company-spying-opinion-schneier/index.html



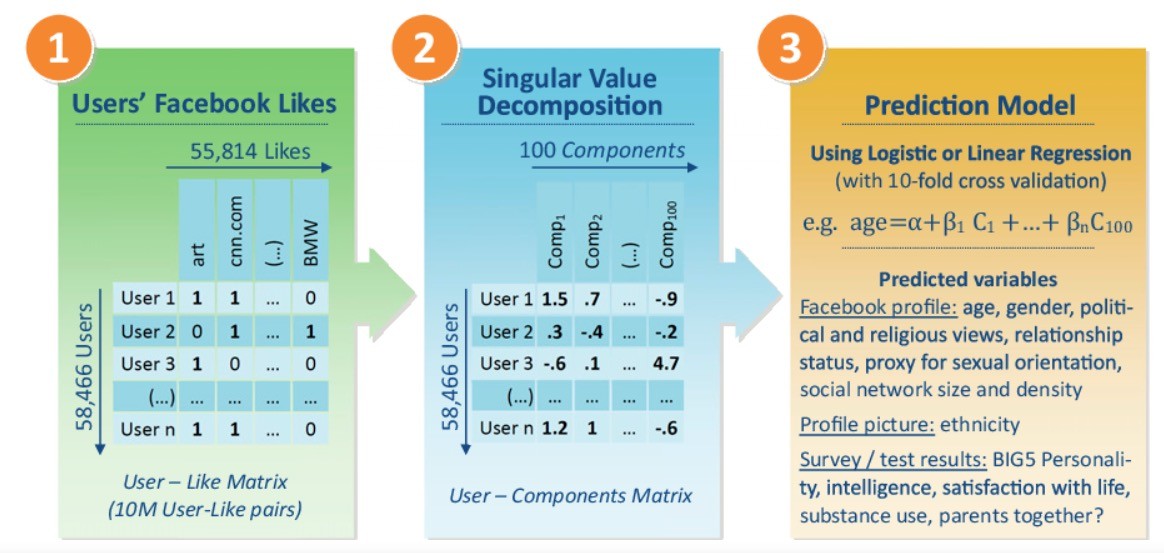
https://[www.pcmag.com/how-to/social-media-and-food-delivery-apps-sell-the-most-personal-data](http://www.pcmag.com/how-to/social-media-and-food-delivery-apps-sell-the-most-personal-data)

# Data Reveals More Than is Evident



Poster by David Stillwell, Michal Kosinski, R.J. Tunney

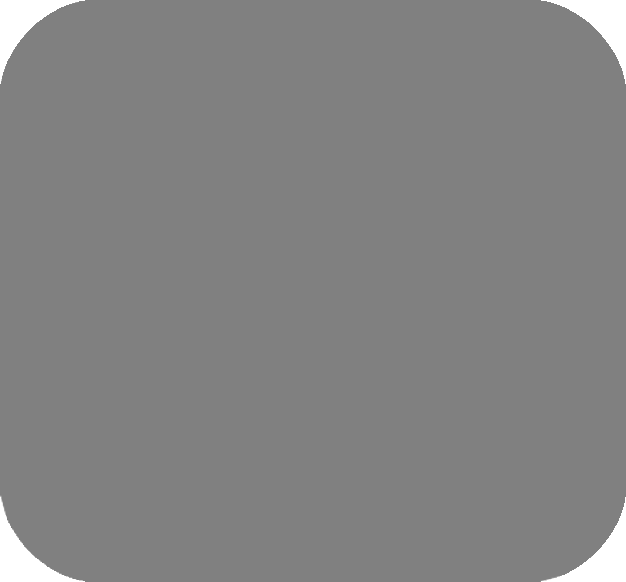
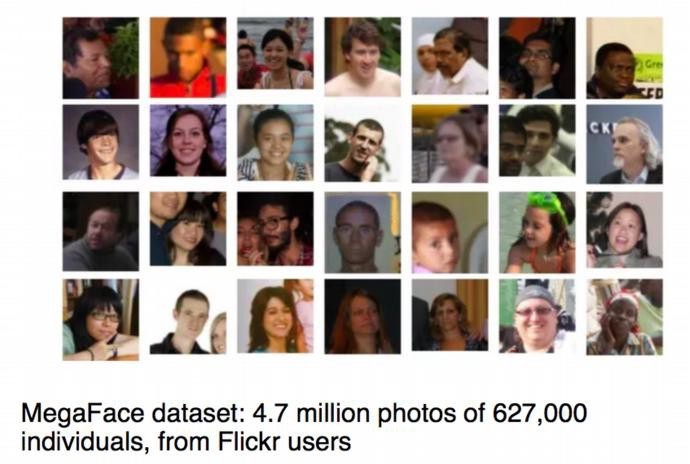
Data Reveals More Than is Evident



Michal Kosinski, David Stillwell and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. PNAS 2013

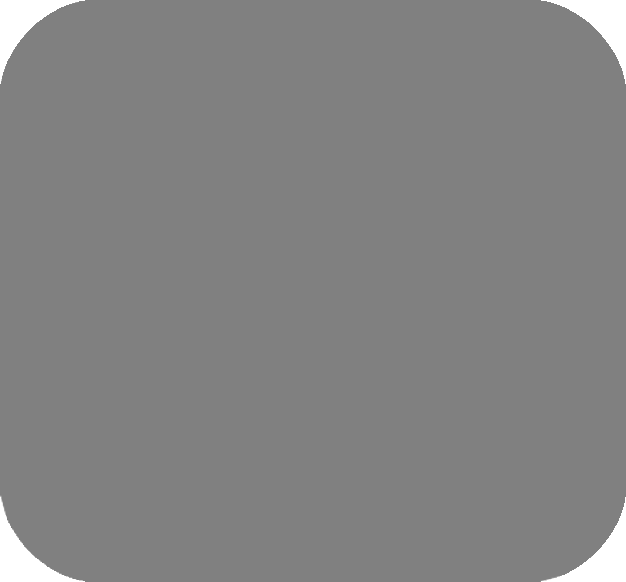
# Data Privacy: Sharing of Data

### Shared data often ends up being passed on and used in unforeseen ways



Source: Michael Paul, U Colorado Boulder

# Data Privacy: Sharing of Data

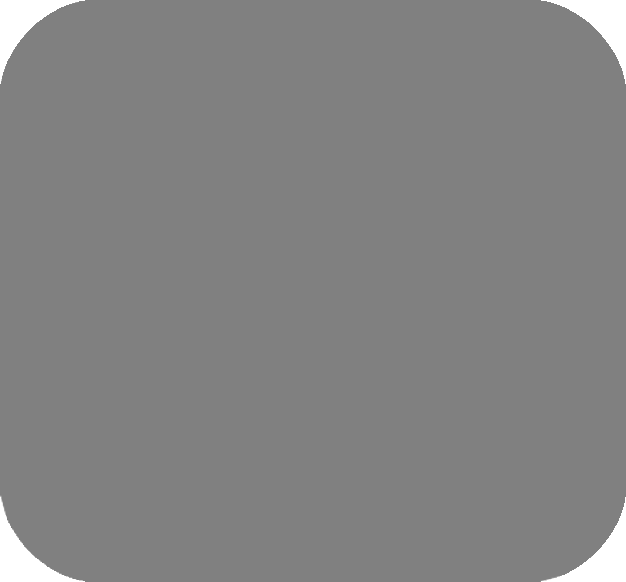
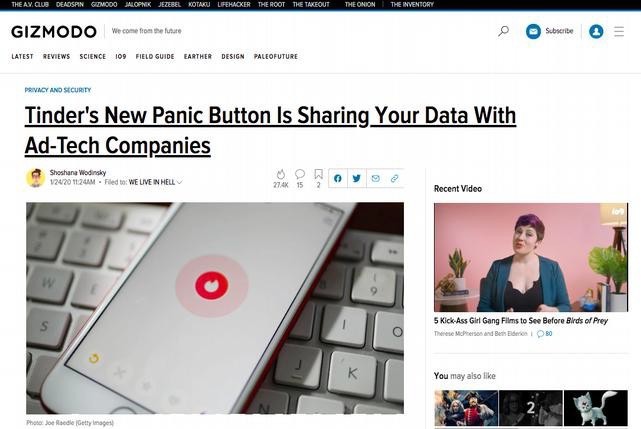


**Shared data often ends up being passed on and used in unforeseen ways**

[www.forbes.com/sites/thomasbrewster/2018/04/16/huge-facebook-facial-recognition-database-built-by-ex-israeli-spies/](http://www.forbes.com/sites/thomasbrewster/2018/04/16/huge-facebook-facial-recognition-database-built-by-ex-israeli-spies/)

# Data Privacy: Sharing of Data

### Shared data often ends up being passed on and used in unforeseen ways



https://gizmodo.com/tinders-new-panic-button-is-sharing-your-data-with-ad-t-1841184919

# Data Means Power



https://[www.wired.com/2015/12/psychology-of-clickbait/](http://www.wired.com/2015/12/psychology-of-clickbait/)

# Data Means Power



**How many of you**

**compulsively visit some site/app/service multiple times per day?**

Data Means Power



Source: https://privacyinternational.org/learn/data-and-elections

# Data Means Power



Source: https://[www.brookings.edu/blog/techtank/2022/06/21/data-misuse-and-disinformation-technology-and-the-2022-elections/](http://www.brookings.edu/blog/techtank/2022/06/21/data-misuse-and-disinformation-technology-and-the-2022-elections/)

Does it matter … 6543tgrepcheck2&&\*\*(11 I am anonymous, right?



Source: Duke CompSci 516

# Data Privacy: Sharing of Data

YOUR IDENTITY IS SAFE WITH US

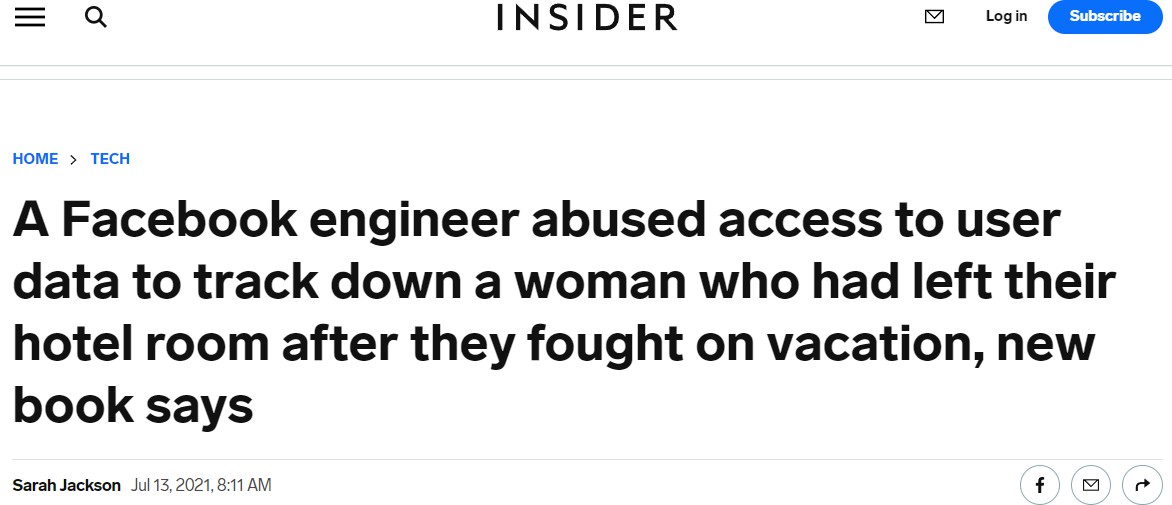


Source: https:[//w](http://www.pinterest.com/pin/470344754831578164/)ww[.pinterest.com/pin/470344754831578164/](http://www.pinterest.com/pin/470344754831578164/)

## Data Privacy: Sharing of Data

*Always?*





## …but, only algorithms access the data…

* Always?

## …and they don’t share data…

* + *Never?*



## …but, only algorithms access the data…

* + - Always?

## …and they 6543tgrepcheck2&&\*\*(11 don’t share data…

* + - * Never?

## …but they anonymize data before sharing!

*Yes (maybe if it’s not a breach)…*

***Is anonymization enough?***

|  |  |  |
| --- | --- | --- |
| **~~Name~~** | **Age** | **Condition** |
| ~~John Doe~~ | 59 | Cancer |
| ~~Mary Smith~~ | 78 | Covid-19 |

# Famous Case: The Netflix Prize



*dannypeled.com*

* 2007 -- Netflix competition to create the best collaborative filtering algorithm to predict what rating a user would give to a movie based on their previous ratings
* Netflix Prize had a huge positive impact in the research area on recommendations

# Famous Case: The Netflix Prize

* Dataset: Ratings of almost 500K customers (more than 100M ratings in total)
* Netflix **anonymized** the dataset
  + From their FAQ *“[...] all customer identifying information has been removed; all that remains are ratings and dates.”*

***What could go wrong?***

# Famous Case: The Netflix Prize

* Researchers from UT at Austin presented a **statistical de-anonymization attack** using IMDB as 6543TGREPCHECK2&&\*\*(11 background knowledge [1]
* Able to **identify individual users** and determine potentially sensitive information (political views, religious views, or sexual orientation).
* Four Netflix users filled a class action 6543tgrepcheck2&&\*\*(11 lawsuit.
  + Netflix ended up reaching a settlement with the plaintiffs in 2010.
* Cancellation of projected sequel to the Netflix Prize

*[1] Narayanan, A., & Shmatikov, V. (2006). How to break anonymity of the netflix prize dataset. arXiv preprint cs/0610105.*

# Another Case: AOL Search Data [1]

Dataset: 20M search queries for 650K users from 2006 Why released: allow researchers to understand search patterns How anonymized: user identifiers removed

* + All searches from same user linked by an arbitrary identifier

Attacks: many successful attacks identified individual users

* + Ego-surfers: people typed in their own names
  + Zip codes and town names identify an area
  + NY Times identified 4417749 as 62yr old GA widow [2]

Consequences: CTO resigned, two researchers fired

* + - Well-intentioned effort failed due to inadequate anonymization

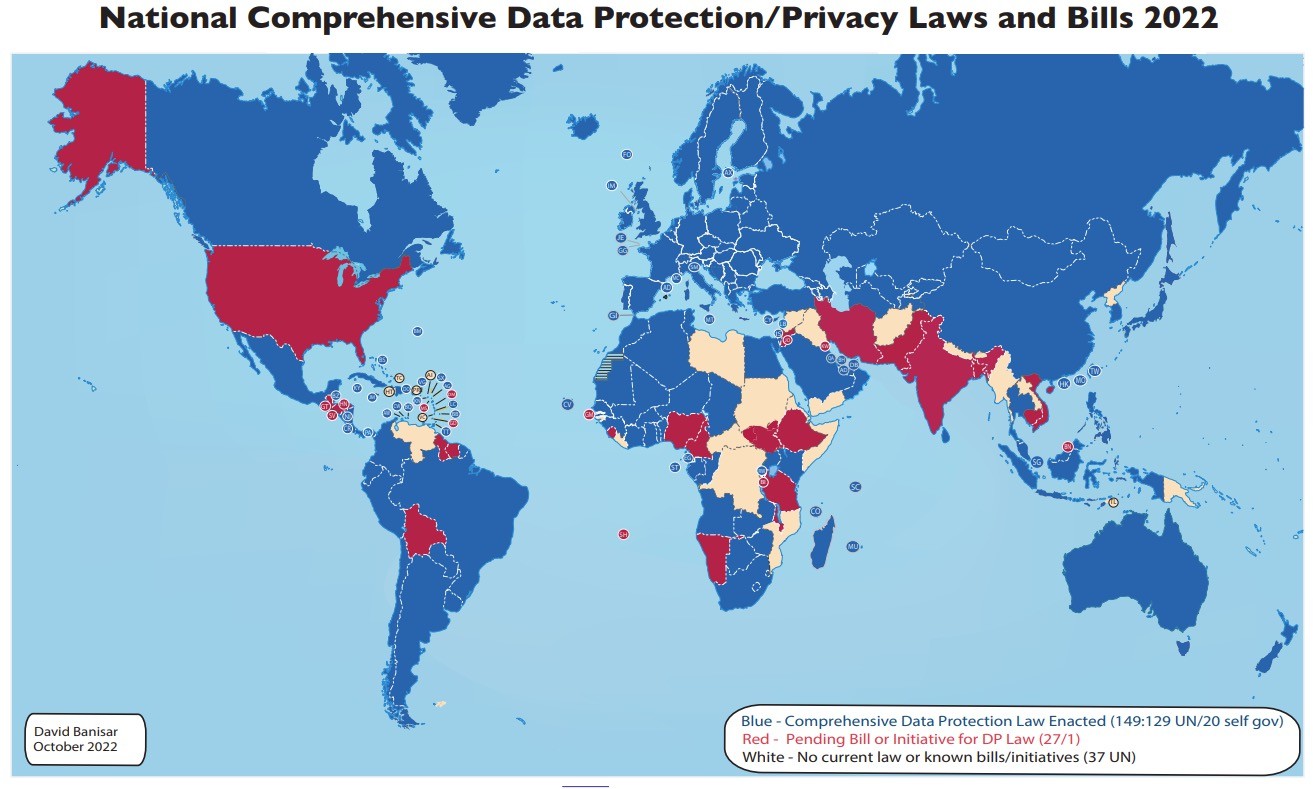
1. *Slide extracted from Cormode, G., & Srivastava, D. (2009, June). Anonymized data: generation, models, usage. SIGMOD (pp. 1015-1018).*
2. *https://*[*www.nytimes.com/2006/08/09/technology/09aol.html*](http://www.nytimes.com/2006/08/09/technology/09aol.html)

**Data Privacy Regulations**

# Data Protection Regulations

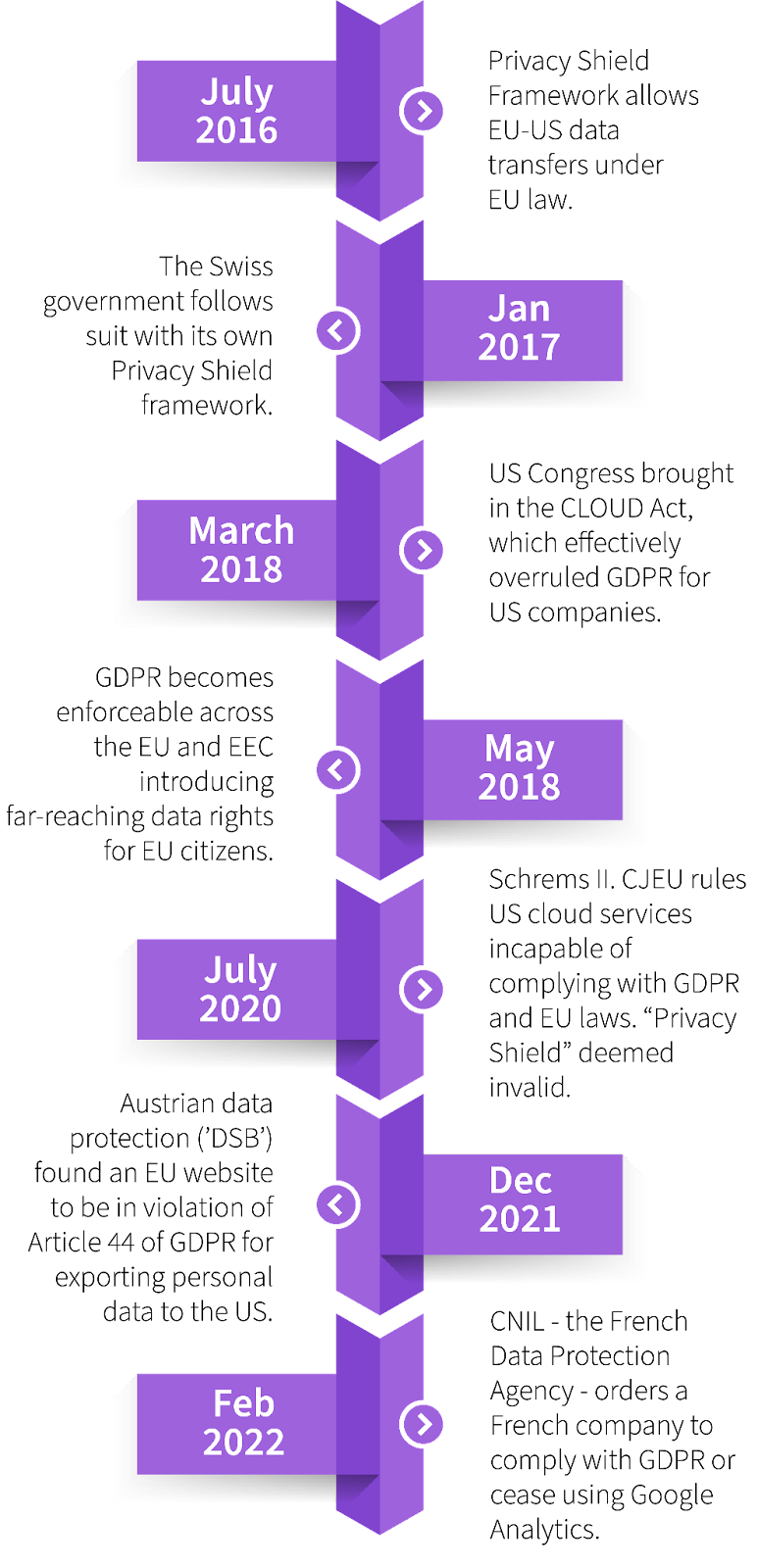
* Nearly 150 countries and self-governing jurisdictions and territories around the world have now adopted comprehensive data protection/privacy laws to protect personal data held by private and public bodies.[1]

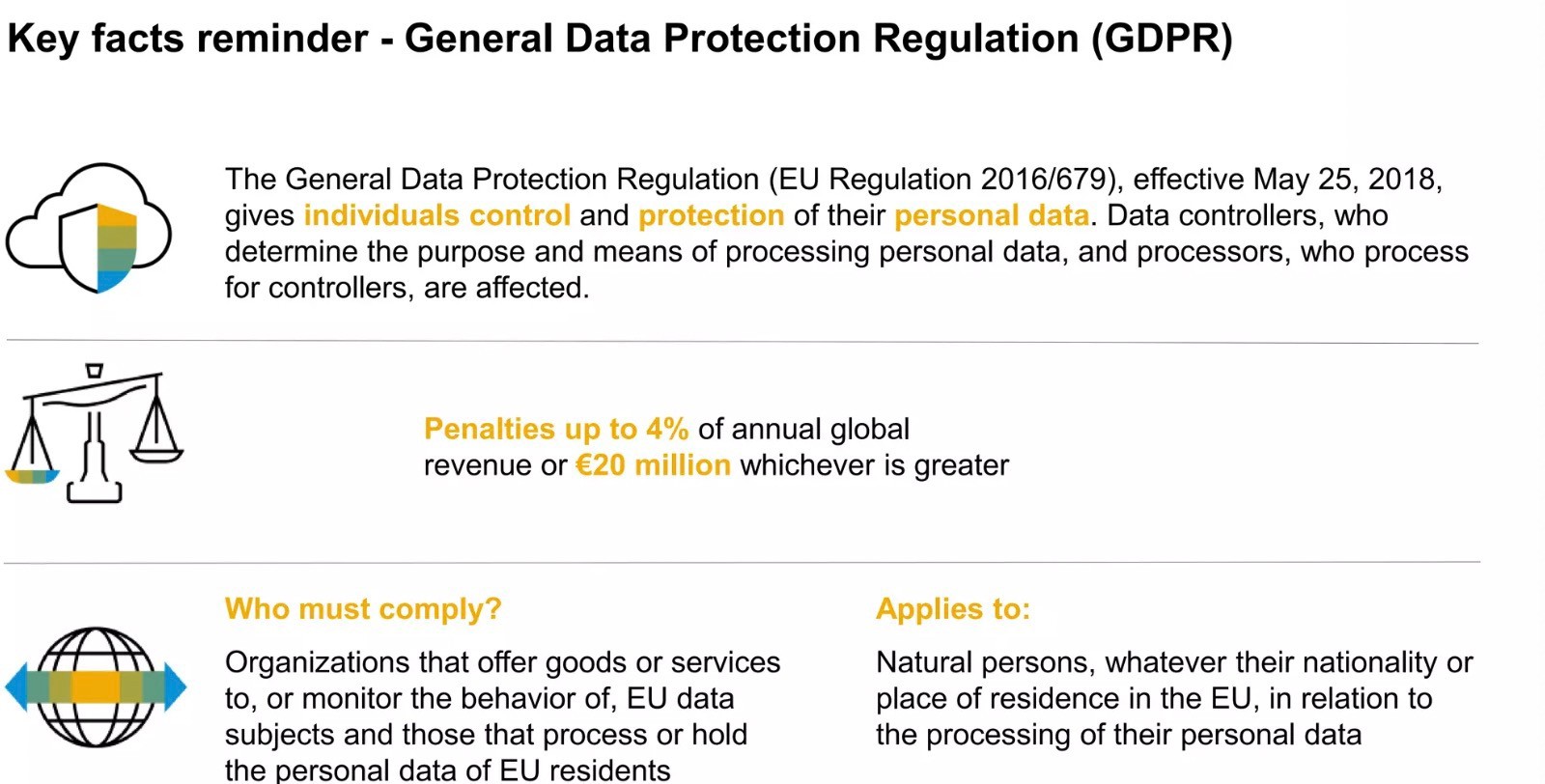
CCPA in California



*[1] National Comprehensive Data Protection/Privacy Laws and Bills 2021, D Banisar - Privacy Laws and Bills, 2021*

European General Data Protection Regulation (GDPR)

GDPR is a law



https://snowplow.io/blog/gdpr-multiple-data-pipelines/ GDPR, Data Privacy and Cybersecurity - MIT Symposium 2018

European General Data Protection Regulation (GDPR)

* + In effect starting May 25th, 2018
  + Unified data protection rules in all 27 member states of the European Union
  + Replaced previous Data Protection Directive
    - Unlike Data Protection Directive, GDPR is a law (i.e., other, national laws are not required)
  + Is applicable to all services offered within the EU (regardless of where the company is located)

# New and Expanded Rights

* Right to be informed
* Right of access
* Right to rectification
* Right to erasure
* Right to restriction
* Right to data portability
* Right to object
* Right to prevent automated processing, including profiling

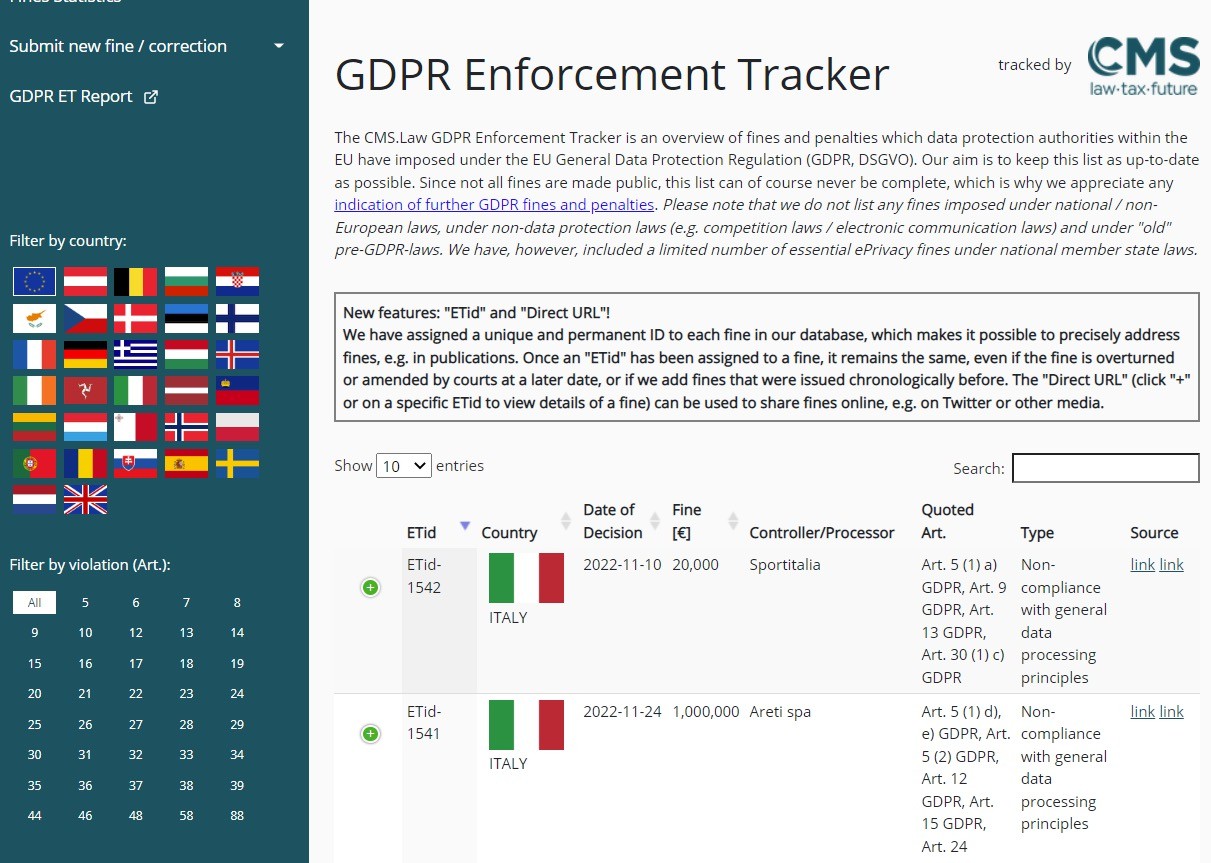
# Regulatory Measures

* **High fines**: up to 20,000,000 EUR or up to 4% of the annual worldwide turnover
* **Data breaches** have 6543TGREPCHECK2&&\*\*(11 to be reported within 72 hours, data subjects have to be notified
* Need for **data protection impact assessment**

**The biggest GDPR fines**

1. Am azon —€746 m illion ($877 m illion) ...
2. Whats App —€225 m illion ($255 m illion) ...
3. Google Irela nd —€90 m illion ($102 m illion) ...
4. Facebook —€60 m illion ($68 m illion) ...
5. Google LLC —€60 m illion ($68 m illion) ...
6. Google – €50 m illion ($56.6 m illion)

[*www.tessian.com/blog/biggest-gdpr-fines-2020/*](http://www.tessian.com/blog/biggest-gdpr-fines-2020/)



*https://*[*www.enforcementtracker.com/*](http://www.enforcementtracker.com/)

# Private data analysis methods

* + **Bare Minimum protection:**
    - K-anonymity [Sweeney IJUFKS 2002]
    - L-diversity [Machanavajjhala et al ICDE 2006]
    - T-closeness [Li et al ICDE 2007]
  + **Ideal: Differential Privacy**

# Differential Privacy

[Dwork et al TCC 2006]

## Consider two datasets

* + - * *D1*: with Bob as one of the participants
      * *D2* : without Bob
    - Answers are roughly the same whether or not Bob is in the data

# Example

* Consider an individual who is deciding whether to allow their data to be included in a database. For example, it may be a patient deciding whether their medical records can be used in a study, or someone deciding whether to answer a survey. A useful notion of privacy would be an assurance that allowing their data to be included should have negligible impact on them in the future. As we’ve already seen, absolute privacy is inherently impossible but what is being guaranteed here is that that the chance of a privacy violation is small. This is precisely what differential privacy (DP) provides.

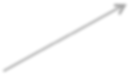
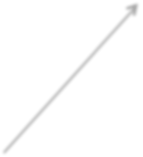
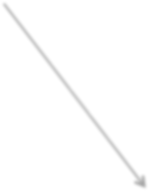
# Differential Privacy

## Algorithm *A* satisfies ε-differential privacy if:

* + for **every pair** of *neighboring tables D1, D2*
  + for **every output** O

Pr[A(D1) = O] ≤ eε Pr[A(D2) = O]

# Meaning …



**A(D1) = O1**

***D1***

**.**

**.**

**.**

***D2***

**Set of all outputs**

**P [ A(D1) = O1 ]**



**Bob in the data**



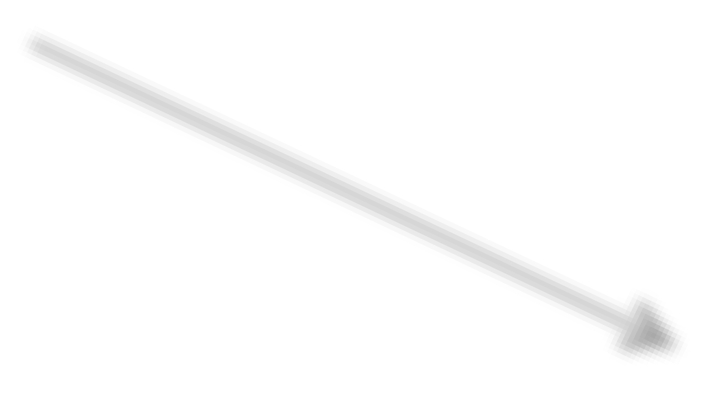
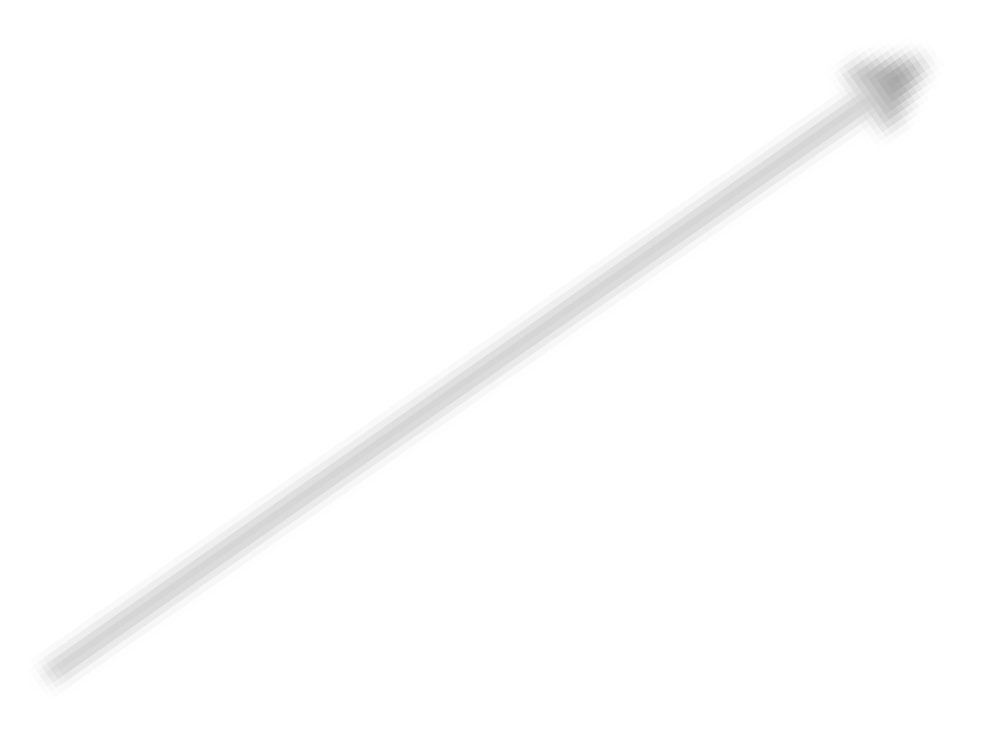
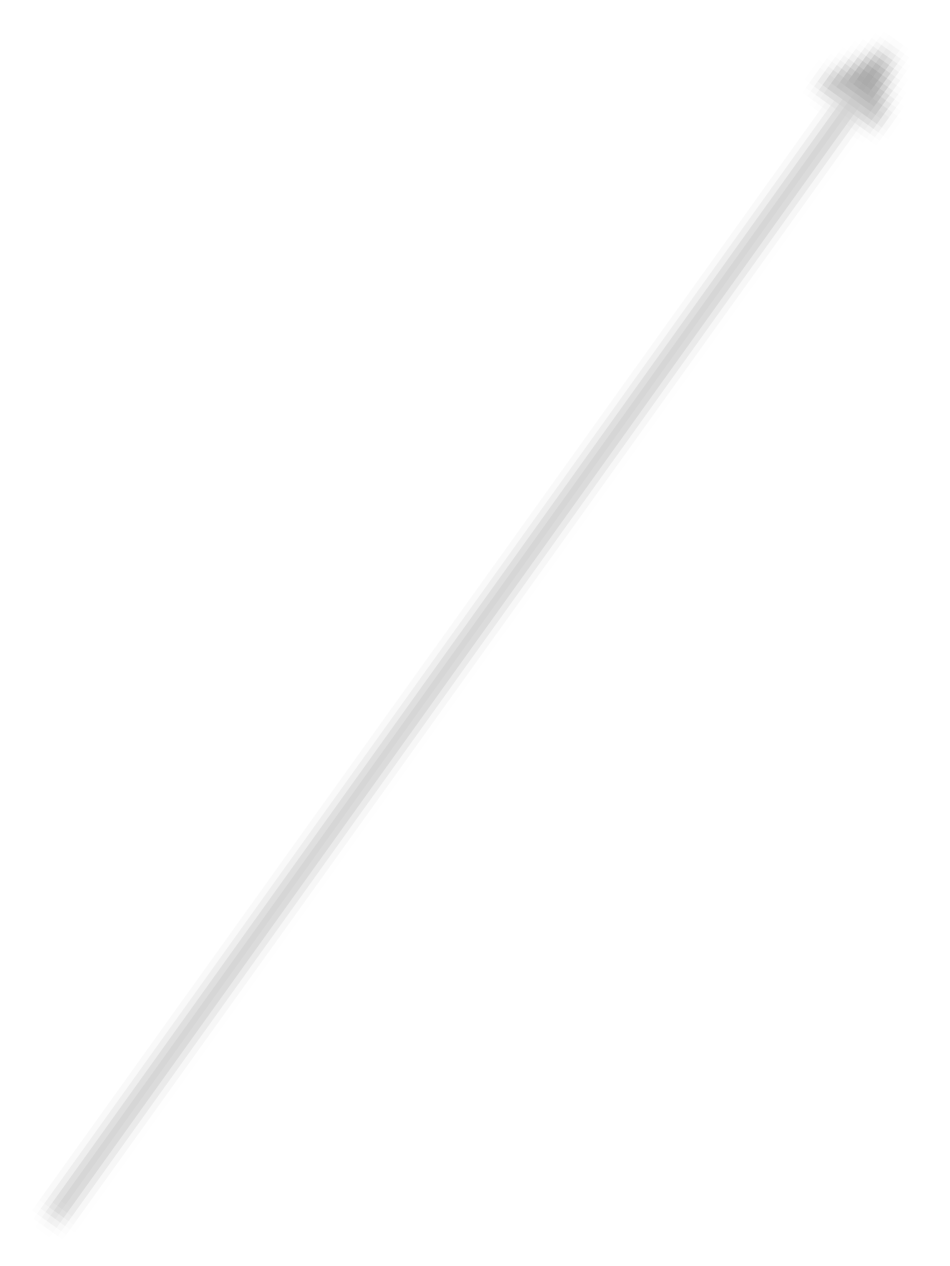
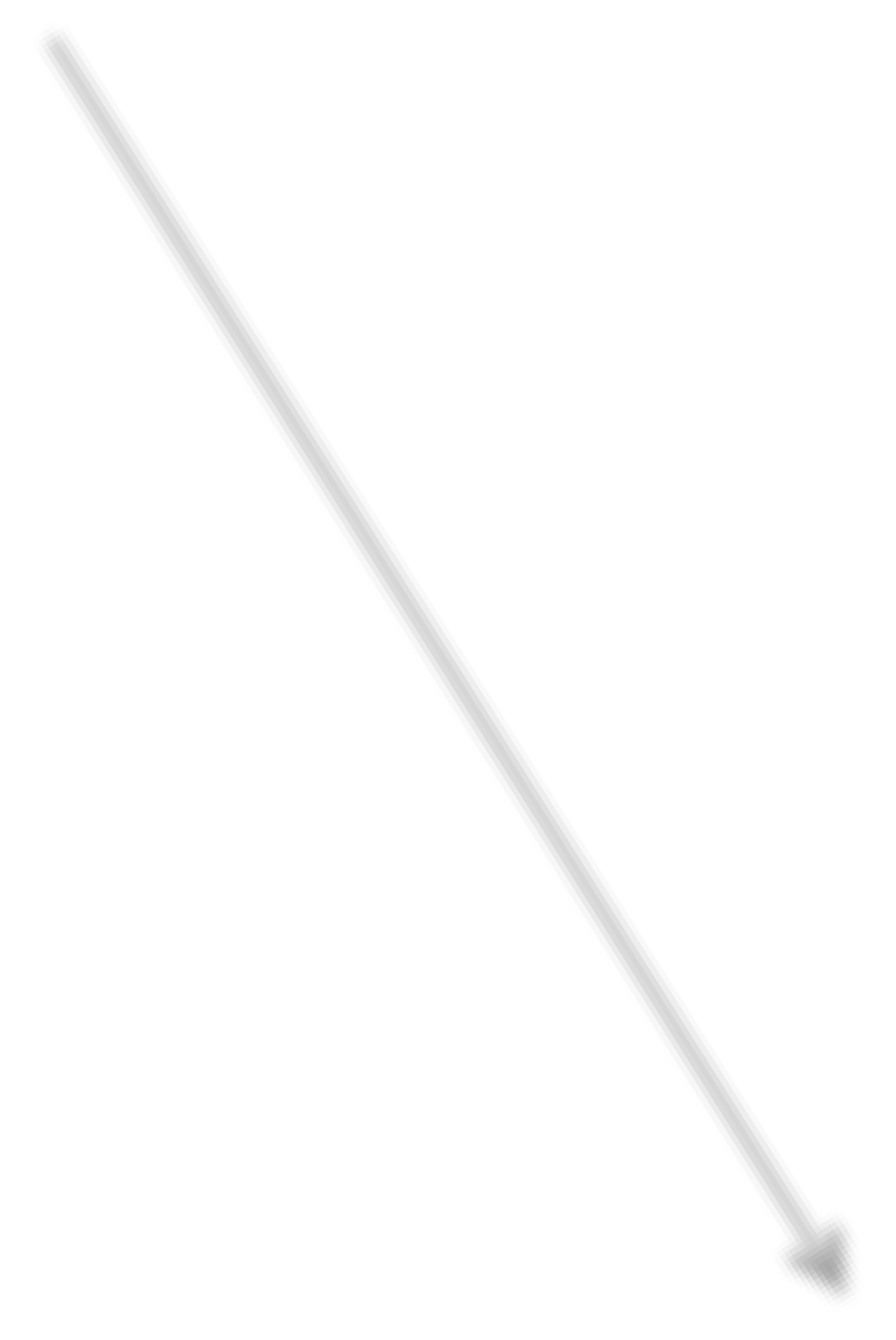
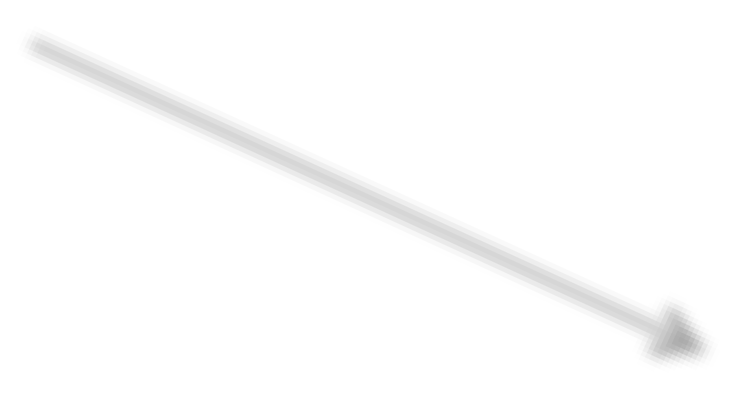
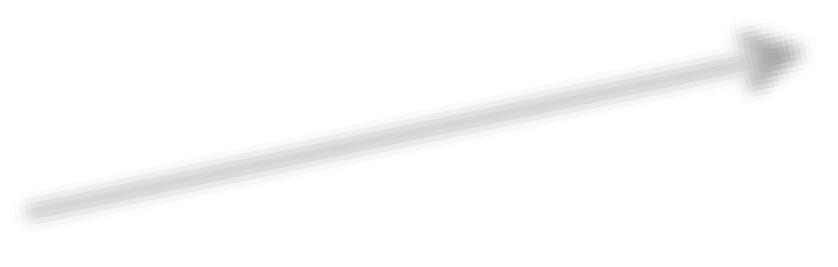
**Bob not in the data**

**P [ A(D2) = Ok ]**

# Meaning …

***O1***

**Worst discrepancy in probabilities**



***D1***

**.**

**.**

**.**

***D2***

# Privacy loss parameter ε

## Algorithm *A* satisfies ε-differential privacy if:

* for **every pair** of *neighboring tables D1, D2*
* for **every output** O

## Pr[A(D1) = O] ≤ eε Pr[A(D2) = O]

* Smaller the ε more the privacy (and better the utility)

# Privacy loss parameter ε

## Algorithm *A* satisfies ε-differential privacy if:

* for **every pair** of *neighboring tables D1, D2*
* for **every output** O

## Pr[A(D1) = O] ≤ eε Pr[A(D2) = O]

***what the adversary learns about an individual is the same even if the individual is not in the data (or lied about his/her value)***

**What are DP applications?**

* U.S. Census Bureau started to use differential privacy with the 2020 Census data. The dataset contains detailed demographic information about U.S. citizens. Without a privacy measure, this information can be traced back to individuals. The Bureau states that traditional anonymization techniques became obsolete. This is because of re- identification methods that make it possible to reveal information about a specific individual from an anonymized dataset.
* In 2014, Google introduced a differential privacy tool called Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR) to Chrome browsers. It helps Google to analyze and draw insights from browser usage while preventing sensitive information from being traced. Google also made its differential privacy libraries open source in 2019.
* Apple uses differential privacy in iOS and macOS devices for personal data such as emojis, search queries and health information.
* Microsoft uses differential privacy for collecting telemetry data from Windows devices.
* Differential privacy is also used in applications of other privacy-preserving methods in artificial intelligence such as federated learning or synthetic data generation.

Discussion

* Why Data 6543TGREPCHECK2&&\*\*(11 Privacy is important to you?
* How do you protect data privacy?

# Further reading

* A Survey on Privacy in Social Media: Identification, Mitigation, and Applications
* A Survey on Privacy in Human 6543tgrepcheck2&&\*\*(11 Mobility
* When Machine Learning Meets Privacy: A Survey and Outlook