UNIVERSITAT DE BARCELONA

MSc in Fundamental Principles of Data Science

Ethical Data Science

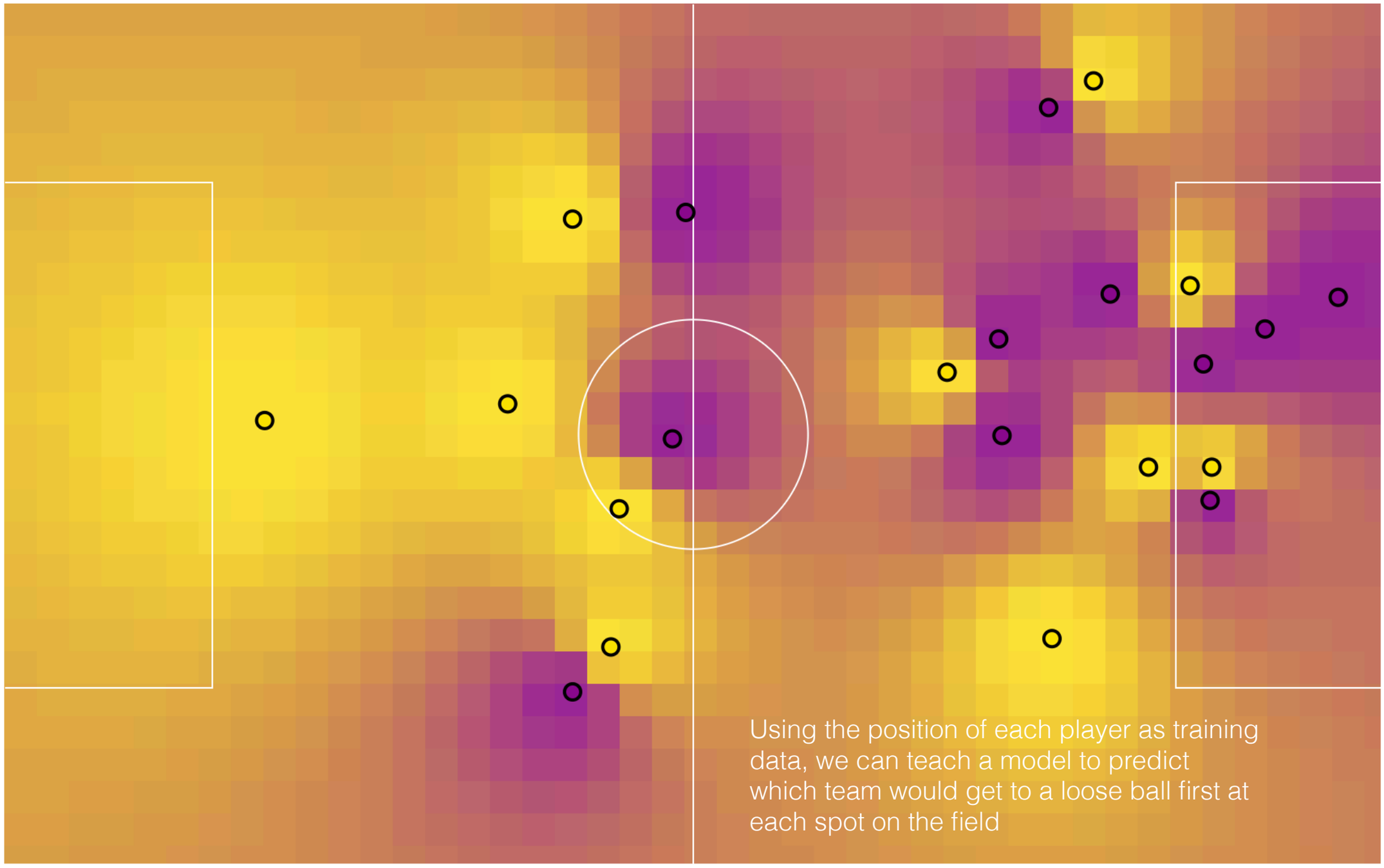
Privacy or the problem of data agency

Jordi Vitrià

2022-2023

Why some models leak data

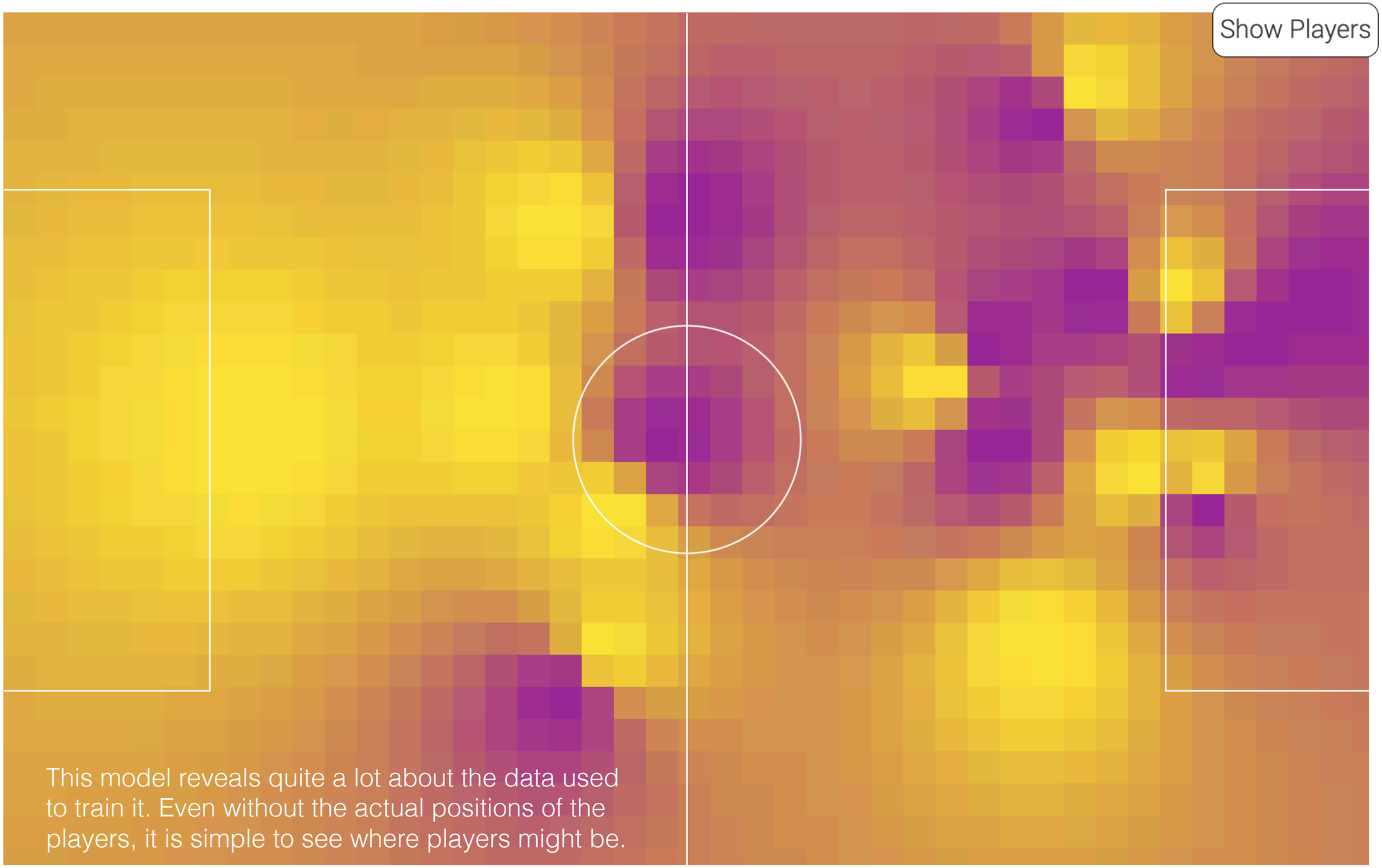
Game of soccer.



https://pair.withgoogle.com/explorables/data-leak/

Why some models leak data

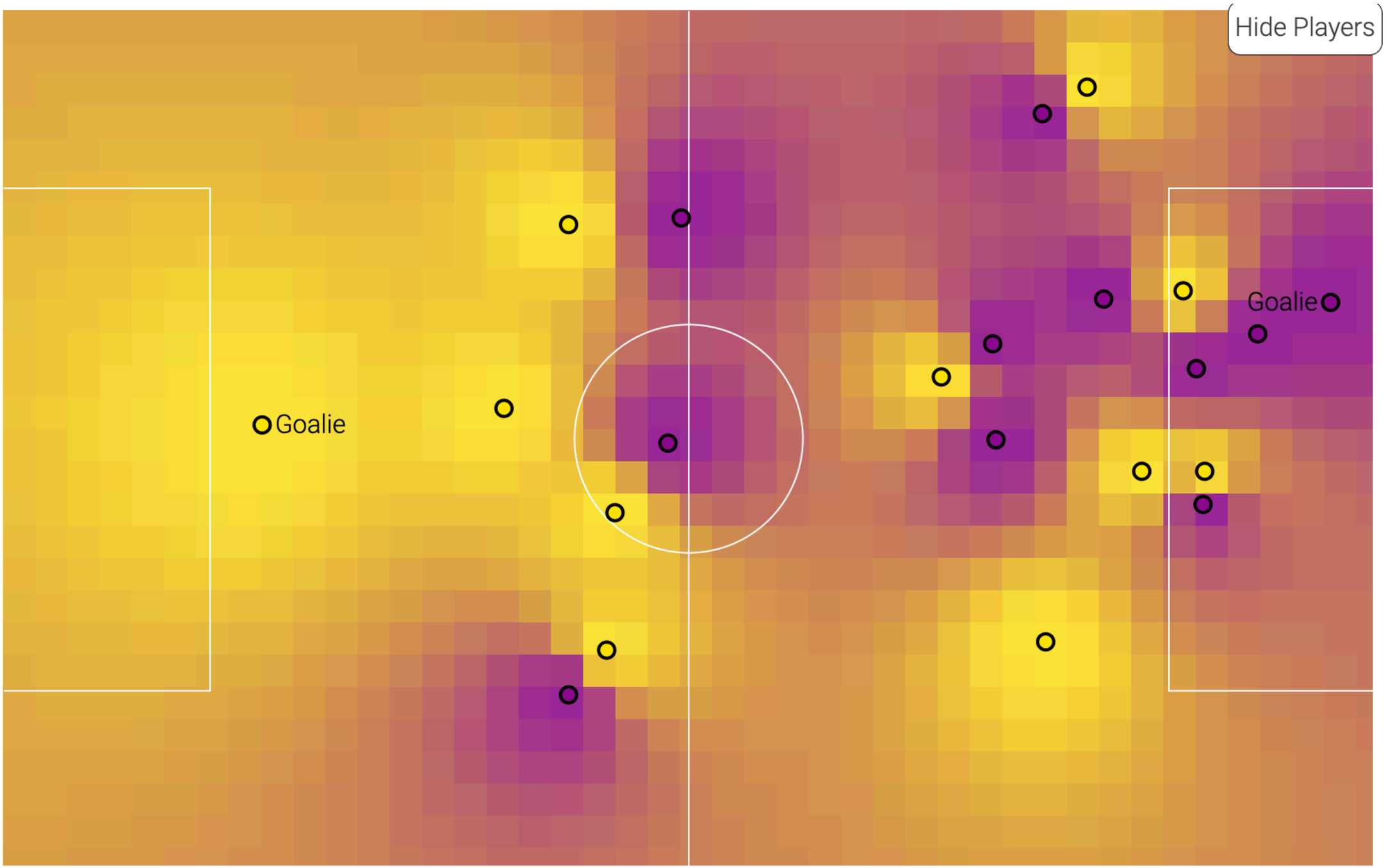
Game of soccer.



https://pair.withgoogle.com/explorables/data-leak/

Why some models leak data

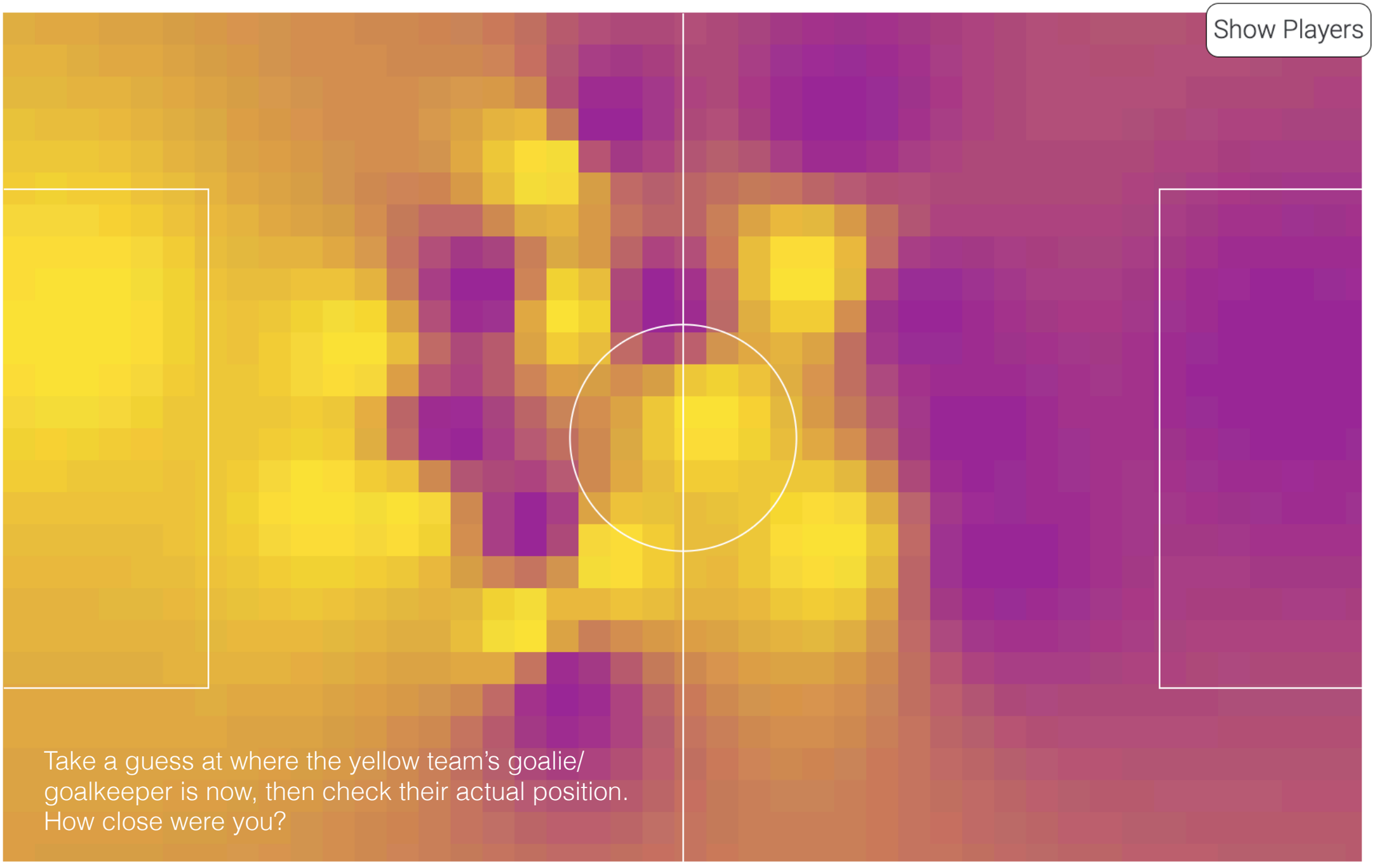
Game of soccer.



https://pair.withgoogle.com/explorables/data-leak/

Why some models leak data

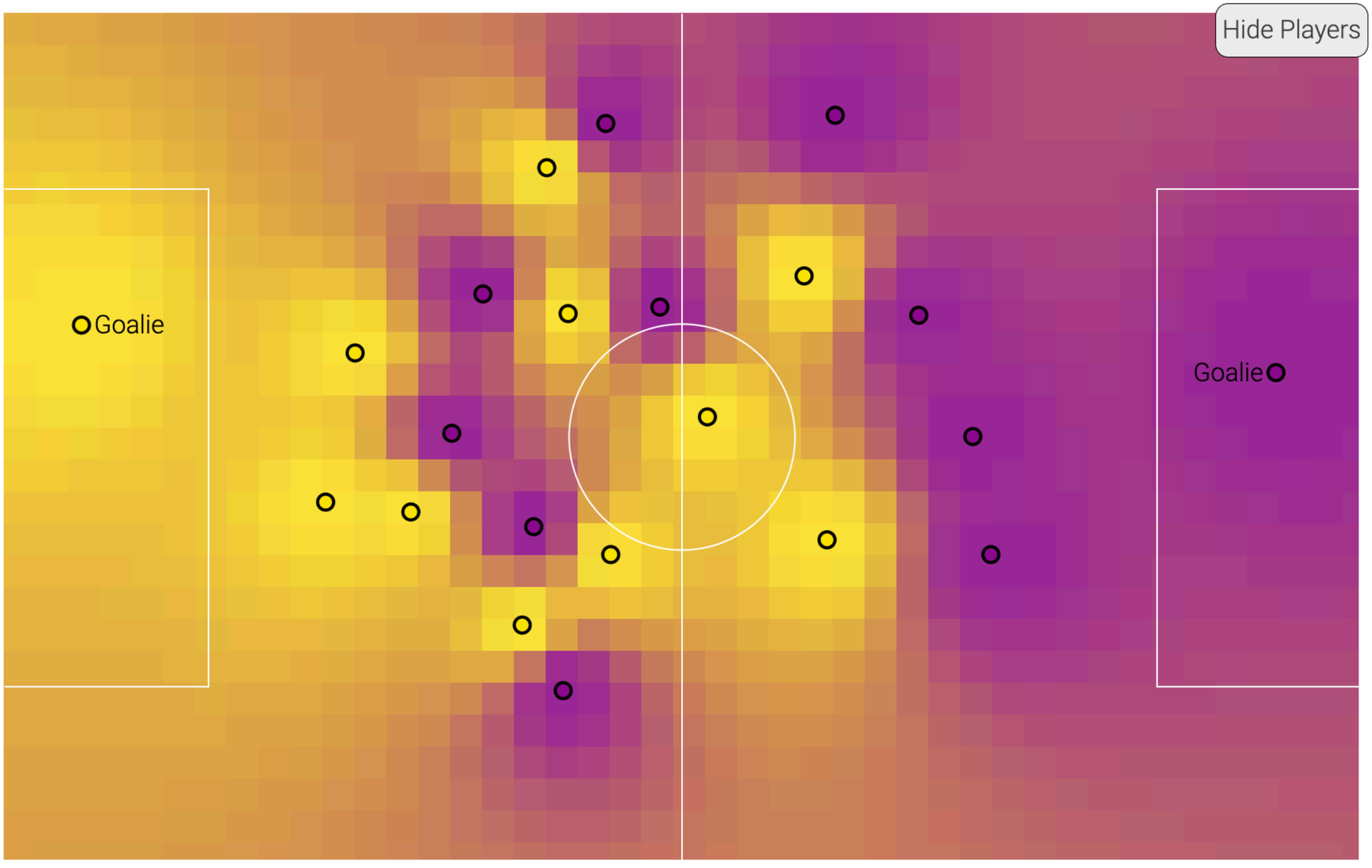
Game of soccer.



https://pair.withgoogle.com/explorables/data-leak/

Why some models leak data

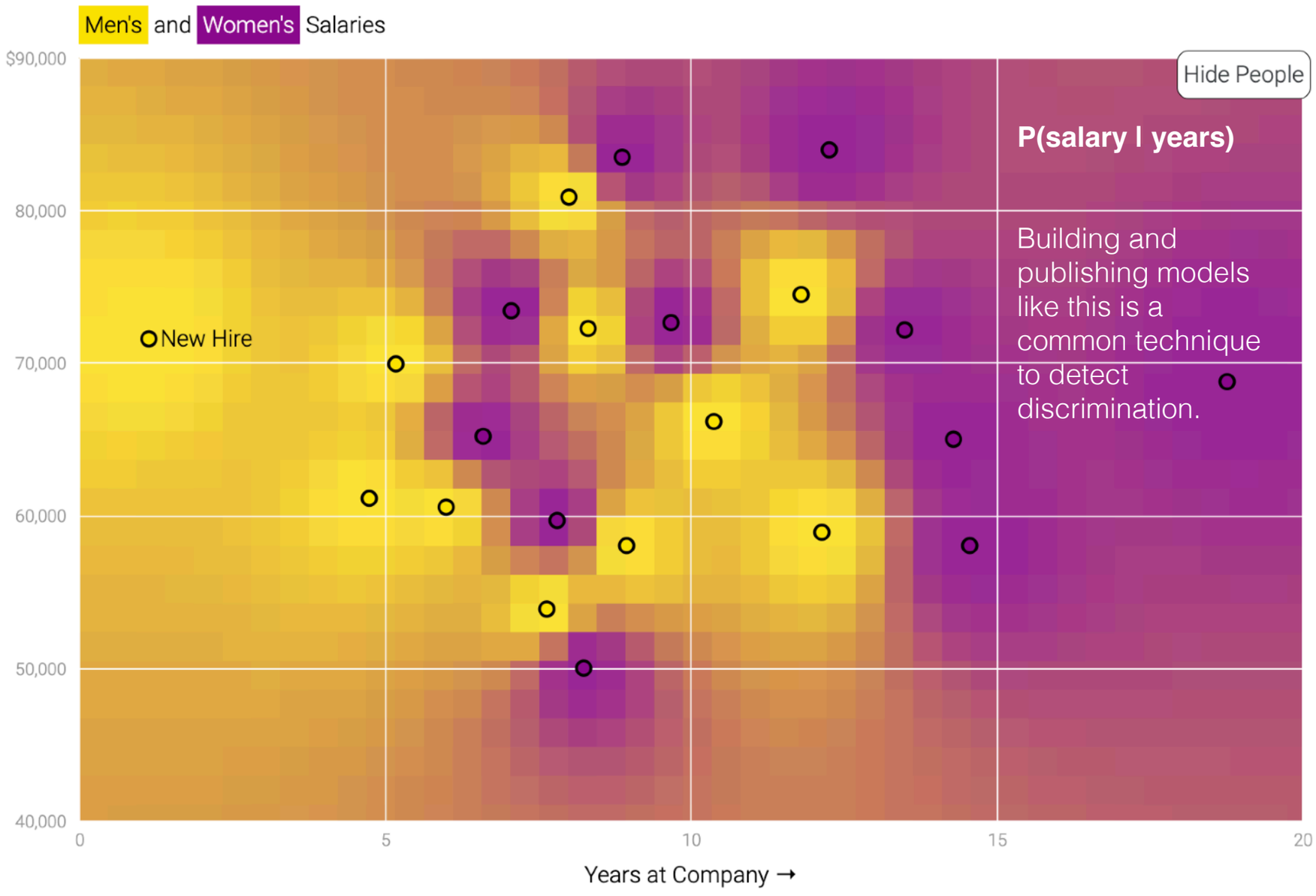
Game of soccer.



https://pair.withgoogle.com/explorables/data-leak/

Why some models leak data

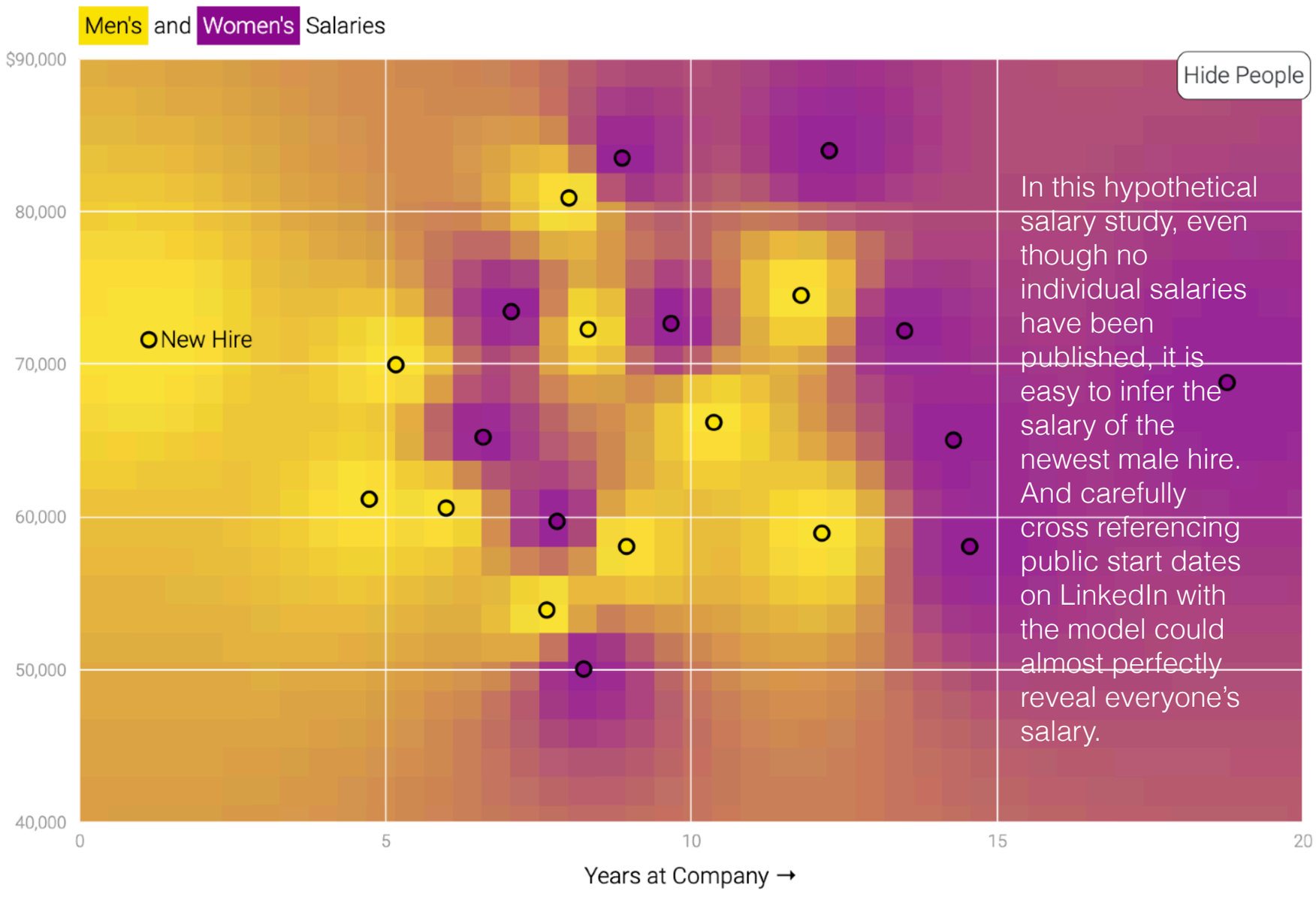
Sensitive Salary Data



https://pair.withgoogle.com/explorables/data-leak/

Why some models leak data

Sensitive Salary Data



https://pair.withgoogle.com/explorables/data-leak/

Why some models leak data

Real World Data

Models of real world data are often quite complex—this can improve accuracy, but makes them more susceptible to unexpectedly leaking information. Medical models have inadvertently revealed patients' genetic markers. Language models have memorized credit card numbers. Faces can even be reconstructed from image models:

M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in CCS, 2015.

Fredrikson et al were able to extract the image on the left by repeatedly querying a facial recognition API. It isn't an exact match with the individual's actual face (on the right), but this attack only required access to the model's predictions, not its internal state.

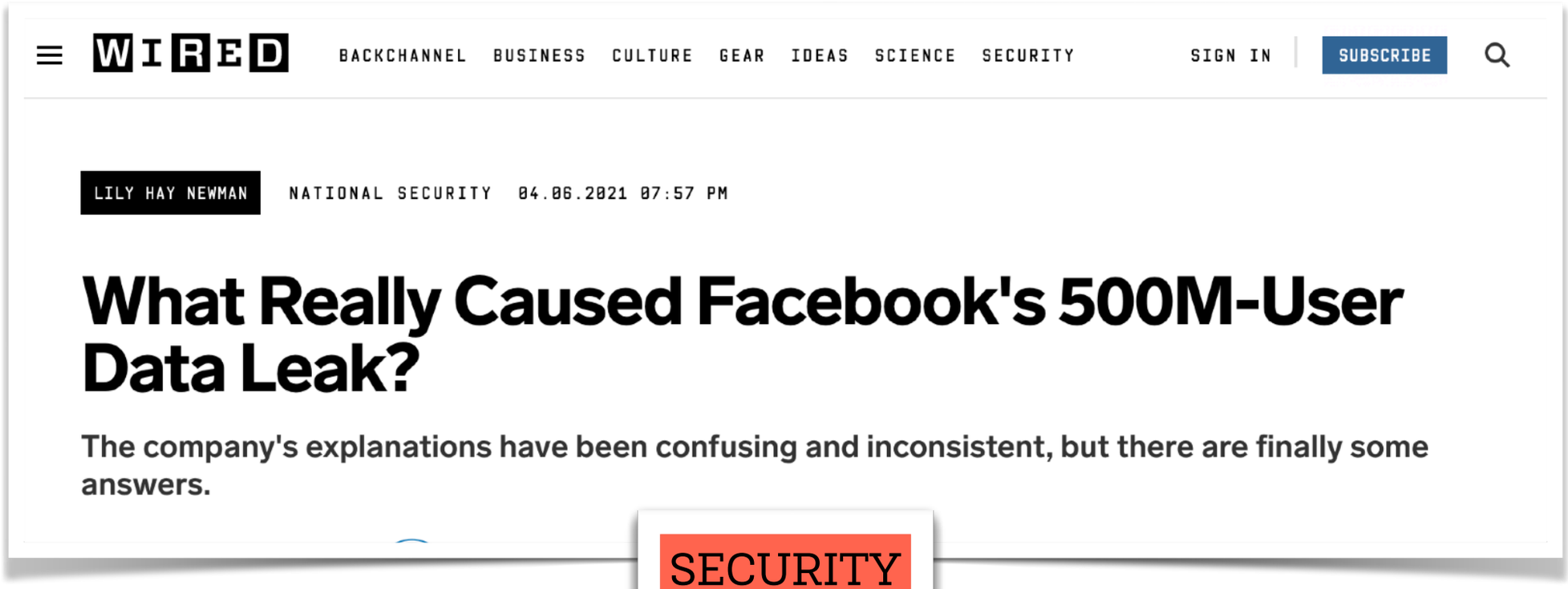
https://pair.withgoogle.com/explorables/data-leak/

Privacy vs Security

SECURITY: Control access to data by using locks, keys, crypto, etc.

PRIVACY: Disallow or **allow** use of data but **control** inferences and exfiltrations by using anonymization, differential privacy, federated learning, etc.

Traditional security issues



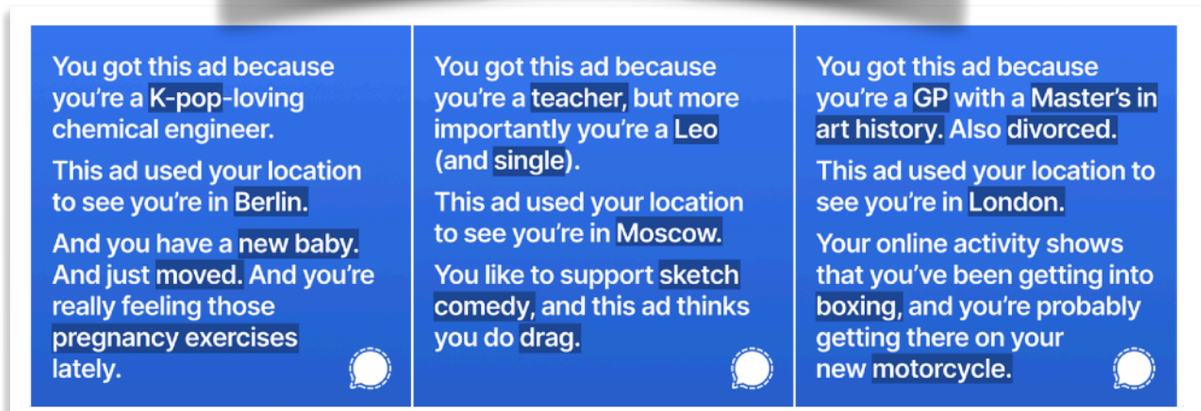
SECURITY

SINCE SATURDAY, A massive trove of Facebook data has circulated publicly, splashing information from roughly 533 million Facebook users across the internet. The data includes things like profile names, Facebook ID numbers, email addresses, and phone numbers. It's all the kind of information that may already have been leaked or scraped from some other source, but it's yet another resource that links all that data together—and ties it to each victim—presenting tidy profiles to scammers, phishers, and spammers on a silver platter.

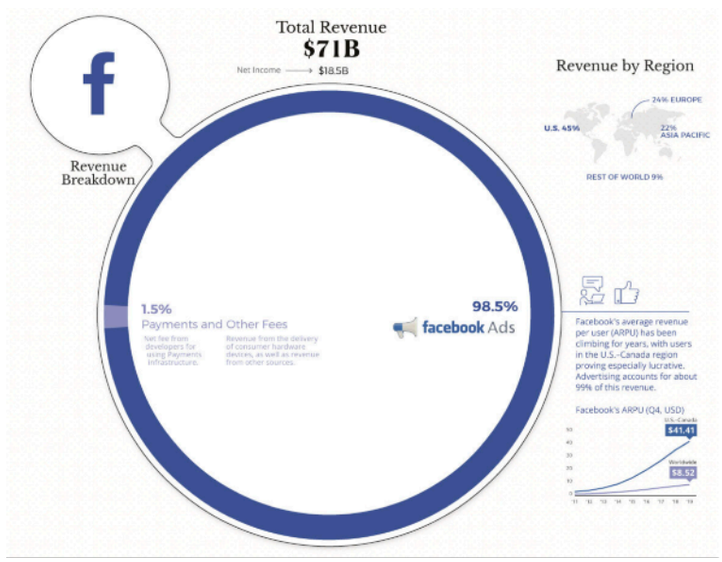
Sensible ML use cases

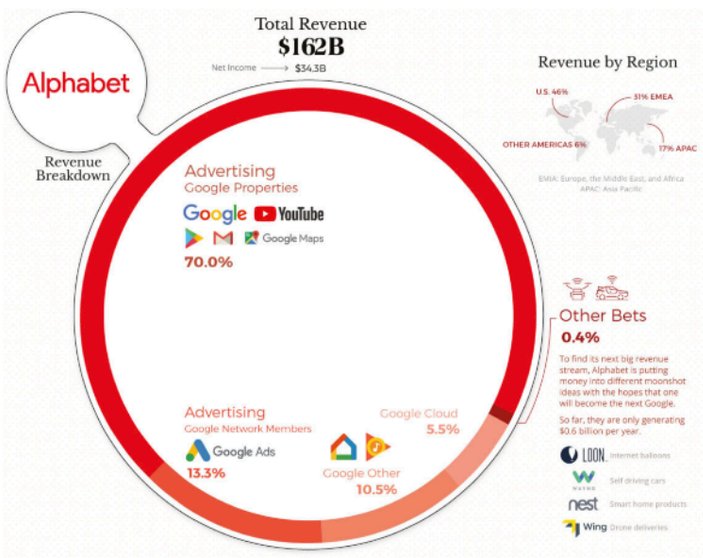
Sensible ML use cases

Unintentional data sharing



https://signal.org/blog/the-instagram-ads-you-will-never-see/





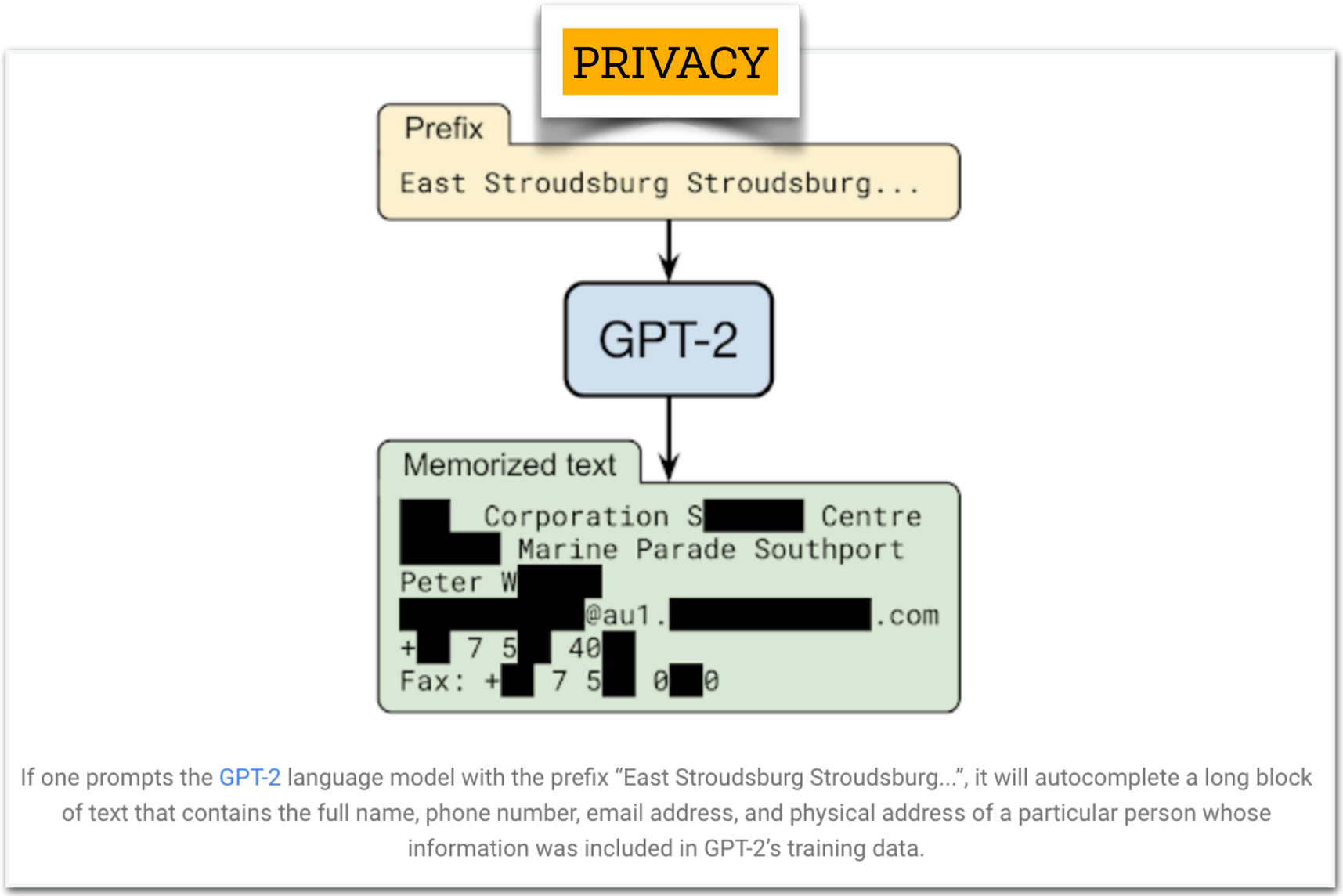
https://www.visualcapitalist.com/how-big-tech-makes-their-billions-2020/

Private information leaking

Privacy Considerations in Large Language Models

Tuesday, December 15, 2020

Posted by Nicholas Carlini, Research Scientist, Google Research

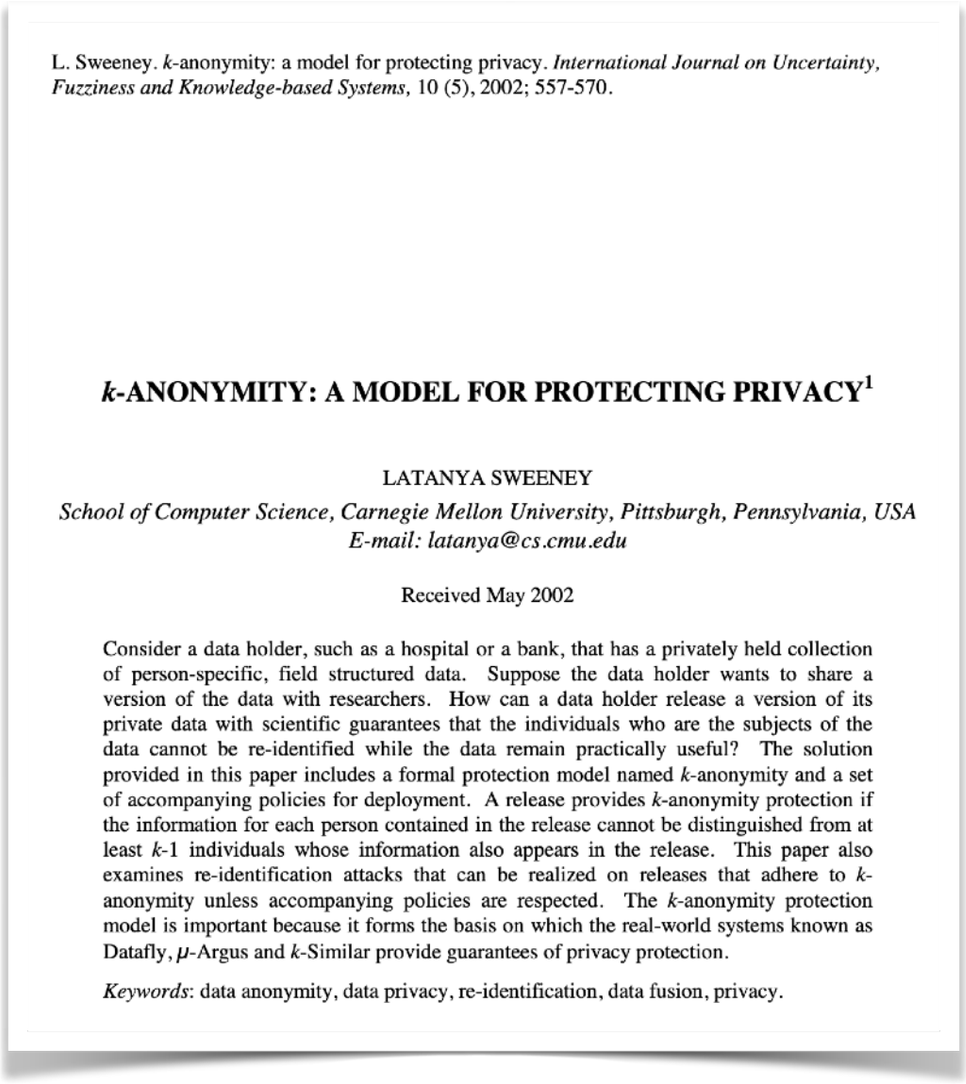


If one prompts the GPT-2 language model with the prefix "East Stroudsburg Stroudsburg...", it will autocomplete a long block of text that contains the full name, phone number, email address, and physical address of a particular person whose information was included in GPT-2's training data.

https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html

Why it is difficult?

Naive anonymization does not work.



L. Sweeney. *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 557-570.

***k*-ANONYMITY: A MODEL FOR PROTECTING PRIVACY[1]**

LATANYA SWEENEY

School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
E-mail: latanya@cs.cmu.edu

Received May 2002

Consider a data holder, such as a hospital or a bank, that has a privately held collection of person-specific, field structured data. Suppose the data holder wants to share a version of the data with researchers. How can a data holder release a version of its private data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful? The solution provided in this paper includes a formal protection model named *k*-anonymity and a set of accompanying policies for deployment. A release provides *k*-anonymity protection if the information for each person contained in the release cannot be distinguished from at least *k*-1 individuals whose information also appears in the release. This paper also examines re-identification attacks that can be realized on releases that adhere to *k*-anonymity unless accompanying policies are respected. The *k*-anonymity protection model is important because it forms the basis on which the real-world systems known as Datafly, μ-Argus and *k*-Similar provide guarantees of privacy protection.

Keywords: data anonymity, data privacy, re-identification, data fusion, privacy.

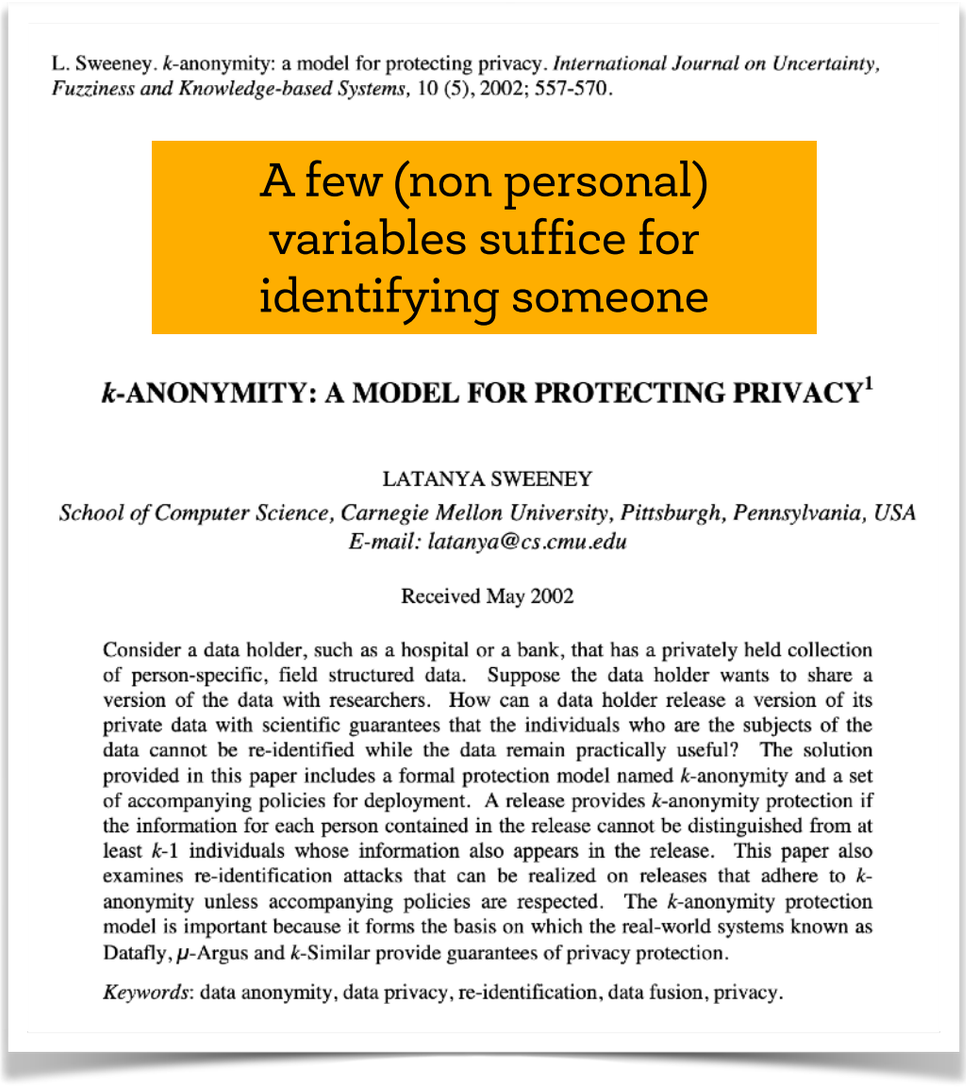
Sweeney demonstrated in an academic paper how she was able to **identify and retrieve sensitive medical data from individuals based on linking a public available data set of 'hospital visits' to the publicly available voting registrar in the United States**. Both datasets where assumed to be properly anonymized through the deletion of names and other direct identifiers.

Quasi-identifiers

Based on only the three parameters (1) Zip Code, (2) Gender and (3) Date of Birth, she showed that 87% of the entire US population could be re-identified by matching aforementioned attributes from both datasets.

Why it is difficult?

Naive anonymization does not work.



L. Sweeney. *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), 2002; 557-570.

A few (non personal) variables suffice for identifying someone

***k*-ANONYMITY: A MODEL FOR PROTECTING PRIVACY**[1]

LATANYA SWEENEY

School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
E-mail: latanya@cs.cmu.edu

Received May 2002

Consider a data holder, such as a hospital or a bank, that has a privately held collection of person-specific, field structured data. Suppose the data holder wants to share a version of the data with researchers. How can a data holder release a version of its private data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful? The solution provided in this paper includes a formal protection model named *k*-anonymity and a set of accompanying policies for deployment. A release provides *k*-anonymity protection if the information for each person contained in the release cannot be distinguished from at least *k*-1 individuals whose information also appears in the release. This paper also examines re-identification attacks that can be realized on releases that adhere to *k*-anonymity unless accompanying policies are respected. The *k*-anonymity protection model is important because it forms the basis on which the real-world systems known as Datafly, μ-Argus and *k*-Similar provide guarantees of privacy protection.

Keywords: data anonymity, data privacy, re-identification, data fusion, privacy.

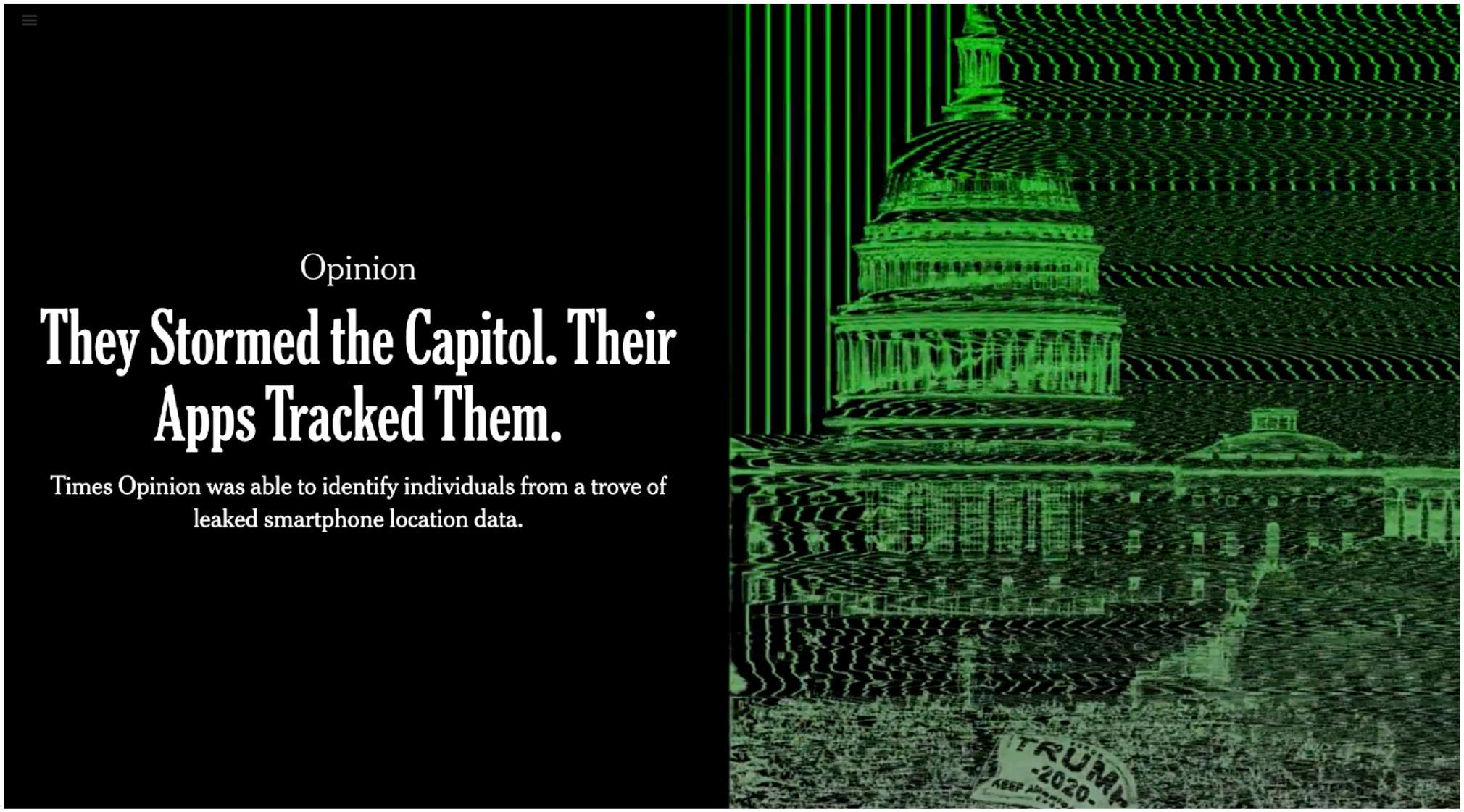
Sweeney demonstrated in an academic paper how she was able to **identify and retrieve sensitive medical data from individuals based on linking a public available data set of 'hospital visits' to the publicly available voting registrar in the United States**. Both datasets where assumed to be properly anonymized through the deletion of names and other direct identifiers.

Quasi-identifiers

Based on only the three parameters (1) Zip Code, (2) Gender and (3) Date of Birth, she showed that 87% of the entire US population could be re-identified by matching aforementioned attributes from both datasets.

Why it is difficult?

We are dealing with highly **identifiable** data



https://www.nytimes.com/2021/02/05/opinion/capitol-attack-cellphone-data.html

Why it is difficult?

We are dealing with highly **identifiable** data

https://www.nytimes.com/2021/02/05/opinion/capitol-attack-cellphone-data.html

Why it is difficult?

We are dealing with high **dimensional** data

SecurityFocus™

About Contact

PRINT COMMENT (page 1 of 2) next

Researchers reverse Netflix anonymization

Robert Lemos, SecurityFocus 2007-12-04

In a dramatic demonstration of the privacy dangers of databases that collect consumer habits, two researchers from the University of Texas at Austin have shown that a handful of movie ratings can identify a person as easily as a Social Security number.

The researchers -- graduate student Arvind Narayanan and professor Vitaly Shmatikov, both from the Department of Computer Sciences at the University of Texas at Austin -- claim to have identified two people out of the nearly half million anonymized users whose movie ratings were released by online rental company Netflix last year. The company published the large database as part of its $1 million Netflix Prize, a challenge to the world's researchers to improve the rental firm's movie-recommendation engine.

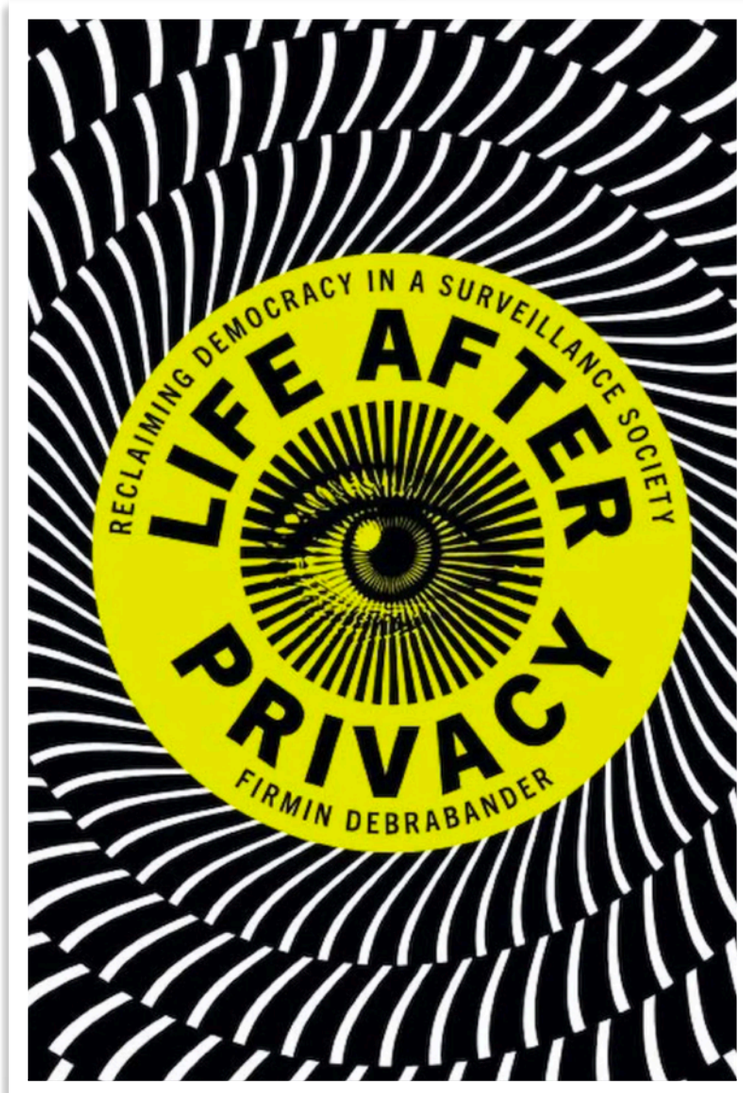
"Releasing the data and just removing the names does nothing for privacy," Shmatikov told SecurityFocus. "If you know their name and a few records, then you can identify that person in the other (private) database."

" Releasing the data and just removing the names does nothing for privacy. If you know their name and a few records, then you can identify that person in the other (private) database. "

Vitaly Shmatikov, Professor of Computer Science, University of Texas at Austin

Netflix:
Individual users matched with film ratings on the Internet Movie Database.

Is privacy a lost cause?



Some authors are pessimistic:

- We're living in a "**confessional culture**" that normalizes oversharing.

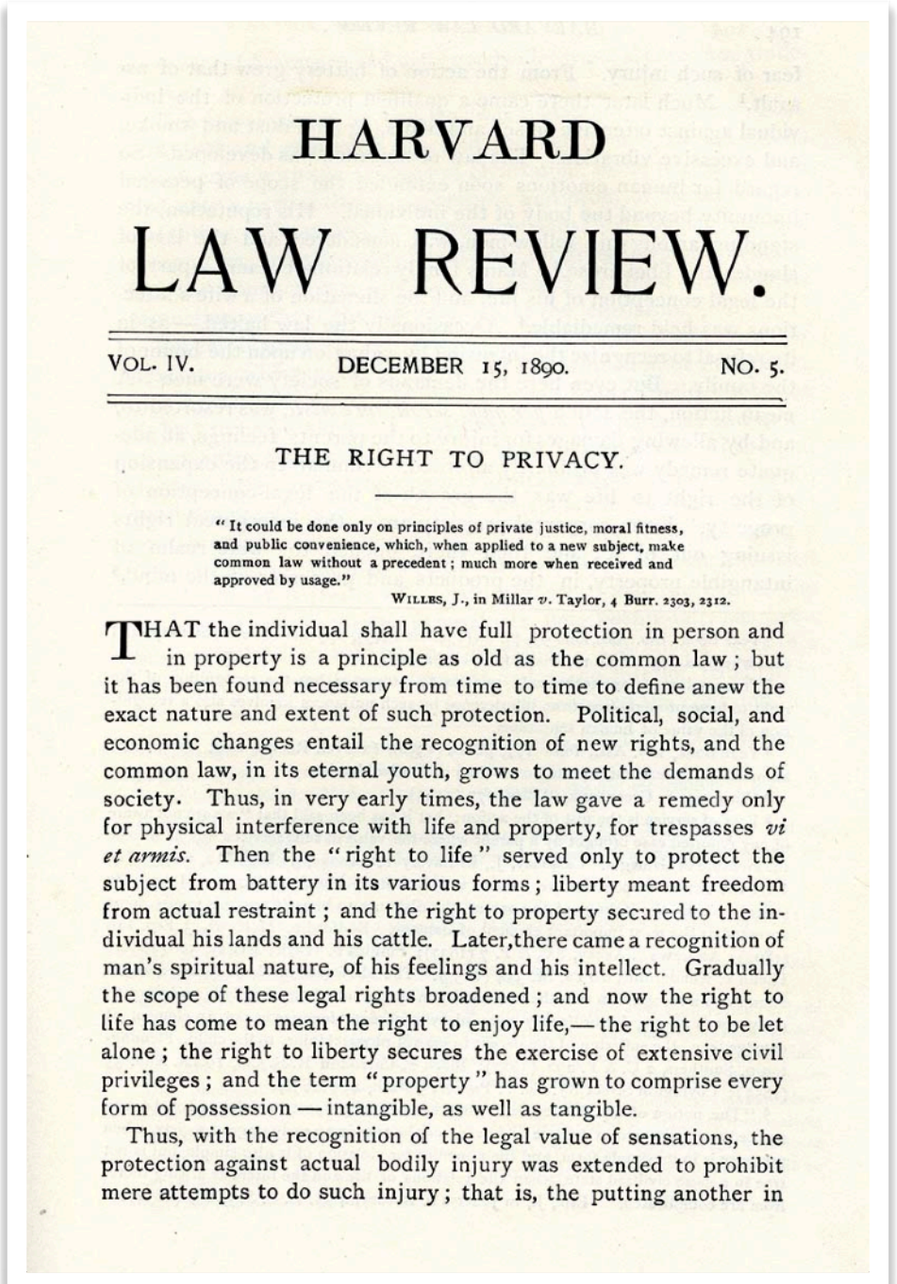
- Companies participating in the "**surveillance economy**" have an insatiable appetite for our personal information, and consumers don't fully comprehend just how much value these companies are able to extract from it through data analytics.

- Consumer privacy protection laws are doomed to be ineffective because they're fundamentally locked into a "**privacy self-management**" paradigm that presupposes an "**illusion of autonomy**". Consumers lack the bargaining power to negotiate with take-it-or-leave-it offers from tech companies.

- The modern conception of privacy is a recent historical invention, **not an enduring value** or **original virtue**.

The human factor

Not a new problem



HARVARD

LAW REVIEW.

VOL. IV. DECEMBER 15, 1890. NO. 5.

THE RIGHT TO PRIVACY.

"It could be done only on principles of private justice, moral fitness, and public convenience, which, when applied to a new subject, make common law without a precedent; much more when received and approved by usage."

WILLES, J., in Millar *v.* Taylor, 4 Burr. 2303, 2312.

THAT the individual shall have full protection in person and in property is a principle as old as the common law; but it has been found necessary from time to time to define anew the exact nature and extent of such protection. Political, social, and economic changes entail the recognition of new rights, and the common law, in its eternal youth, grows to meet the demands of society. Thus, in very early times, the law gave a remedy only for physical interference with life and property, for trespasses *vi et armis*. Then the "right to life" served only to protect the subject from battery in its various forms; liberty meant freedom from actual restraint; and the right to property secured to the individual his lands and his cattle. Later, there came a recognition of man's spiritual nature, of his feelings and his intellect. Gradually the scope of these legal rights broadened; and now the right to life has come to mean the right to enjoy life,—the right to be let alone; the right to liberty secures the exercise of extensive civil privileges; and the term "property" has grown to comprise every form of possession—intangible, as well as tangible.

Thus, with the recognition of the legal value of sensations, the protection against actual bodily injury was extended to prohibit mere attempts to do such injury; that is, the putting another in

Instantaneous photographs *and* ***newspaper*** *enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that "what is whispered in the closet shall be proclaimed from the house-tops." (...)*

Gossip is no longer the resource of the idle and of the vicious, but has become a trade, which is pursued with industry as well as effrontery (...)

To occupy the indolent, column upon column is filled with idle gossip, which can only be procured by intrusion upon the domestic circle.

S. D. Warren and L. D. Brandeis, 1890

The privacy paradox

Privacy paradox: **users' concerns about data privacy aren't reflected in their behaviors**.

A recent IBM study found that 81% of consumers say they have become more concerned about how their data is used online. But most users continue to hand over their data online and tick consent boxes impatiently.
(Source: HBR, Jan 30, 2020)

- Why don't users care enough to take actions that match their concerns?
- What are the possible solutions?
- Why is this so difficult?

The privacy paradox

Possible **causes**:

- Data is **intangible**. It is a byproduct of our online activity, it is easy to ignore or forget about.

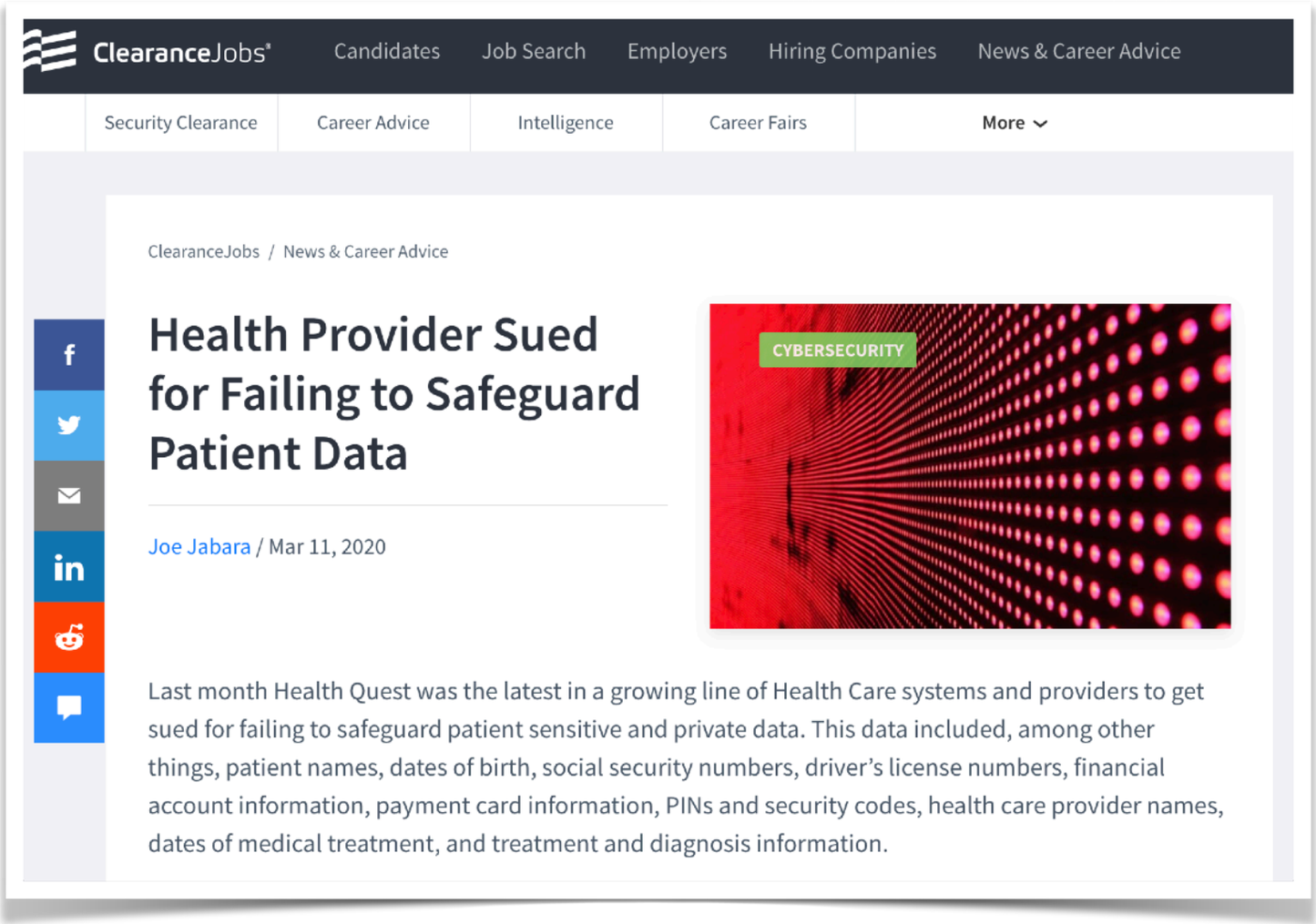
- Even if users wanted to negotiate more **data agency**, they have little leverage. It is not easy to change to a alternative provider.

Data Agency: The ability to **own the rights** to their personal data, manage access to this data and, potentially, be compensated fairly for such access.

Solution is not clear: Better regulation? Better management tools? Data trusts?

GDPR: Europe's General Data Protection Regulation (GDPR) gives citizens greater digital agency. However, there is a lack of practicality associated with the rules.

It is a real problem



Problem: Someone needs to access to (personal) data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure of (personal) data.

What is privacy?

Definition

Broadly speaking, **privacy** is the right to be let alone, or freedom from interference or intrusion.

Data privacy is the **right** to have some control over how your personal data is collected and used.

GDPR

The rules governing **personal data** have their basis in some **fundamental principles**. Article 5 of the GDPR lists the principles that apply to all personal data processing. These principles require that data is:

- processed in a lawful, fair and transparent manner (**principle of legality, fairness and transparency**)
- collected for specific, expressly stated and justified purposes and not treated in a new way that is incompatible with these purposes (**principle of purpose limitation**)
- adequate, relevant and limited to what is necessary for fulfilling the purposes for which it is being processed (**principle of data minimisation**)
- correct and, if necessary, updated (**accuracy principle**)
- not stored in identifiable form for longer periods than is necessary for the purposes (**principle relating to data retention periods**)
- processed in a way that ensures adequate personal data protection (**principle of integrity and confidentiality**)

GDPR & Data Science

DS meets the Fairness Principle

Algorithms and models are dependent on the personal data that is used for training. The model's result may be **incorrect** or **discriminatory** if the training data renders a biased picture reality.

DS meets the Principle of Purpose Limitation

Many of the models developed using artificial intelligence will be used in connection with good causes, such as cancer diagnosis. Are we permitted **to use personal data unrestrictedly as long as it is for a good cause**?

DS meets the Principle of Data Minimization

It is difficult to define **which data is necessary** for a purpose.
The developer must consider how to achieve the objective in a way that is least invasive for the data subjects.

GDPR & Data Science

DS meets the Principle of Transparent Processing

Transparency is achieved by providing data subjects with **process details**.
It can be challenging to explain how information is used.

DS meets the Principle of Integrity and Confidentiality

For scenarios in which users provide training data, **attackers** can repeatedly query a trained model to obtain users' private information. This is called a **membership inference attack**.

Data privacy techniques

SCENARIO Data Analyst

Someone needs to access to (personal) data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure of (personal) data.

Example 1

Let's suppose we provide to the data analyst this **database**:

```
DataBase = {(UNI:UB, LOC:Mollet del Vallès, Outcome:5),
            (UNI:UB, LOC:Mollet del Vallès, Outcome:4),
            (UNI:UAB, LOC:Barcelona, Outcome:9),
            (UNI:UAB, LOC:Barcelona, Outcome:3),
            (UNI:UPC, LOC:Sabadell, Outcome:4),
            (UNI:UB, LOC:Barcelona, Outcome:5),
            (UNI:UB, LOC:Barcelona, Outcome:8)}
```

Is it ok from a privacy perspective?

Example 1

```
DataBase = {(UNI:UB, LOC:Mollet del Vallès, Outcome:5),
            (UNI:UB, LOC:Mollet del Vallès, Outcome:4),
            (UNI:UAB, LOC:Barcelona, Outcome:9),
            (UNI:UAB, LOC:Barcelona, Outcome:3),
            (UNI:UPC, LOC:Sabadell, Outcome:4),
            (UNI:UB, LOC:Barcelona, Outcome:5),
            (UNI:UB, LOC:Barcelona, Outcome:8)}
```

Is it ok from a privacy perspective? **No!**

You can learn that a colleague got a 4!

Example 2

Let's suppose we publish the **result** of the analysis of a database (mean monthly income of student families in this course):

```
DataBase = {1000,2000,3000,2000,1000,6000,2000,10000,2000,4000,100000}
Mean = 12090
```

Is it ok from a privacy perspective?

Example 2

Let's suppose we publish the **result** of the analysis of a database (mean monthly income of student families in this course):

```
DataBase = {1000,2000,3000,2000,1000,6000,2000,10000,2000,4000,100000}
Mean = 12090
```

Is it ok from a privacy perspective? **No!**

We can infer that the son of Mr.Rich was considered in this database!

Privacy Models

DATA:

- **k-Anonymity**: A record is indistinguishable with $k-1$ other records.
- **l-diversity**: Records have diverse sensitive attribute values.

MODELS

- **Federated learning**: Training statistical models over remote devices or siloed data centers, while keeping data localized.
- **Differential privacy**: The output of a query to a database should not depend (much) on whether a record is in the database or not.

k-anonymity

k-anonymity

k-anonymity, the parent of all privacy definitions

2017-08-14 — updated 2017-10-01

In 1997, a PhD student named Latanya Sweeney heard about an interesting data release. A health insurance organization from Massachusetts had compiled a database of hospital visits by state employees, and had thought that giving it to researchers could encourage innovation and scientific discovery. Of course, there were privacy considerations: allowing researchers to look at other citizens health records seemed pretty creepy. So they decided to do the obvious thing, and remove all columns that indicated who a patient was: name, phone number, full address, social security number, etc.

Some demographic information was left in the database, so researchers could still compile useful stats: ZIP code, date of birth, and gender were all part of the data. Sweeney realized that the claims of the Massachusetts governor, who insisted that the privacy of state employees was respected (all identifiers were removed!), were perhaps a little bit over-optimistic. Since the governor himself was a state employee, Sweeney decided to do the obvious thing and reidentify which records of the "anonymized" database were the governor's.

Reidentification problem

With just $20, Sweeney bought the public voter records from Massachusetts, which had both full identifiers (names, addresses) and demographic data (ZIP code and date of birth), and contained the governor's information. Guess how many records matched the governor's gender, ZIP code, and date of birth inside the hospital database? Only one, and thus, Sweeney was able to know which prescriptions and visits in the data were the governor's. She posted all of it to his office, showing theatrically that their anonymization process wasn't as solid as it should have been.

https://desfontain.es/privacy/k-anonymity.html#how-to-prevent-this-attack

k-anonymity

Several factors made this attack possible:

- The hospital data contained demographic information that could be used to distinguish between different records.
- A secondary database was available to figure out the demographic information about the target.
- The target was in both datasets.
- And the demographic information of the target (ZIP code, date of birth, and gender) was unique.

Removing one of the factors should be enough to prevent attacks like these.

k-anonymity

Which ones can we afford to remove, while making sure that the data can be used for data analysis tasks?

Maybe suppressing all demographic values would render the data useless, but there might be a **middle ground** to make sure that the demographic values are no longer unique in the dataset.

This last suggestion is the basic idea of k-anonymity.

k-anonymity

A dataset is said to be **k-anonymous** if every combination of values for **demographic** columns in the dataset appears at least for k different records.

ZIP code	age
4217	34
4217	34
1742	77
1742	77
4217	34

k=2? Yes!

ZIP code	age
4217	34
1742	77
1743	77
4217	34

k=2? No!

The intuition is that when a dataset is k-anonymous for a sufficiently large k, the last requirement for a successful **reidentification attack is broken**.

k-anonymity

The two main building blocks used to transform a dataset into a k-anonymous table are generalization and suppression.

ZIP code	age
4217	34
4217	39
1742	75
1691	77

→

ZIP code	age
4217	30-39
4217	30-39
1***	75-79
1***	75-79

- **Generalization** is the process of making a quasi-identifier value less precise, so that records with different values are transformed (or generalized) into records that share the same values.

ZIP code	age
4217	34
4217	39
1742	75
1691	77
9755	13

→

ZIP code	age
4217	30-39
4217	30-39
1000-1999	75-79
1000-1999	75-79

- **Suppression.** In some cases, because of outliers, generalization can result in very large ranges of values, which would significantly reduce the utility of the resulting data. So a simple solution to deal with such outlier values is simply to remove them from the data.

Optimal *k*-anonymity is an NP-hard problem.
There are cheaper approximations.

k-anonymity

Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	30	Female	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	Kerala	Hindu	Viral infection
Salima	28	Female	Tamil Nadu	Muslim	TB
Sunny	27	Male	Karnataka	Parsi	No illness
Joan	24	Female	Kerala	Christian	Heart-related
Bahuksana	23	Male	Karnataka	Buddhist	TB
Rambha	19	Male	Kerala	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
Johnson	17	Male	Kerala	Christian	Heart-related
John	19	Male	Kerala	Christian	Viral infection

Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Viral infection

https://en.wikipedia.org/wiki/K-anonymity

k-anonymity

k-anonymity is simple to understand, and it seems intuitively obvious that **reidentification attacks** are well mitigated when a dataset is transformed to become k-anonymous.

However, it **only mitigates this particular kind of attack** ("this record corresponds to my target").

An attacker might figure out private information about someone, *without reidentifying their record*.

The leak of sensitive information associated to one given individual is the problem, not the reidentification itself!

l-diversity

Suppose you have the following database, which contains everyone in the country:

name	ZIP code	age	diagnostic
Alice	4217	34	Common flu
Bob	4212	39	Healthy
Camille	4732	39	Otitis
Dan	4743	23	Otitis

https://desfontain.es/privacy/l-diversity.html

Now, you want to release an anonymized version of this database, for research purposes.

l-diversity

Let's make this data k-anonymous. Here, k=2, because it's a small country...

ZIP code	age	diagnostic
421*	30-39	Common flu
421*	30-39	Healthy
47**	20-39	Otitis
47**	20-39	Otitis

https://desfontain.es/privacy/l-diversity.html

Suppose an attacker wants to find Camille's diagnostic. The attacker knows that Camille has ZIP code 4732 and age 23. They can easily figure out that Camille's record is the third or fourth one, but cannot know which.

And there's the obvious problem: **both records have the same diagnostic**. So the attacker can deduce that Camille's diagnostic is "Otitis". Even without knowing which record is Camille's!

k-anonymity wasn't enough to protect Camille's private information.

l-diversity

Columns, not necessarily sensitive themselves but which might be used in a reidentification attack.

Let's say that all users with the same **quasi-identifier** tuple are in the same *bucket*.

If all sensitive values are the same within a bucket, we might leak private information. The obvious solution? Imposing some **diversity in the sensitive values associated to the same (generalized) tuple**.

l-diversity states that each bucket must have at least l distinct sensitive values.

ZIP code	age	diagnostic
4***	20-39	Common flu
4***	39	Healthy
4***	39	Otitis
4***	20-39	Otitis

https://desfontain.es/privacy/l-diversity.html#wait-that-seems-too-easy

l-diversity

The key idea behind l-diversity is that if the attacker has uncertainty over the sensitive value, then we avoid leaking private info. But consider the following database, which satisfies 2-diversity:

ZIP code	age	diagnostic
42**	20-29	AIDS
42**	20-29	Hepatitis B
17**	30-39	Otitis
17**	30-39	Healthy

https://desfontain.es/privacy/l-diversity.html#wait-that-seems-too-easy

Suppose the attacker knows that their target has ZIP code 4235 and age 25. The target's record is one of the first two rows. The attacker can learn that their target either has AIDS, or hepatitis B. They can't be sure which one is the correct one... But they can infer that their target has a sexually transmitted infection. This is called **probabilistic information gain.**

t-closeness

To solve the probabilistic information gain problem, we need to also require that their *distribution* is roughly the same that the rest of the data.

If 40% of the records are "healthy" in the overall data, then each bucket must also have roughly 40% of "healthy" records.

This way, the attacker's knowledge can't change *too much* from the baseline.

This is the core idea behind another definition named **t-closeness**.

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian (2007). "*t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity" (PDF). *ICDE*. Purdue University. doi:10.1109/ICDE.2007.367856

Federated Learning

Federated Learning

Federated learning involves training statistical models over remote devices or siloed data centers, such as mobile phones or hospitals, while keeping data localized.

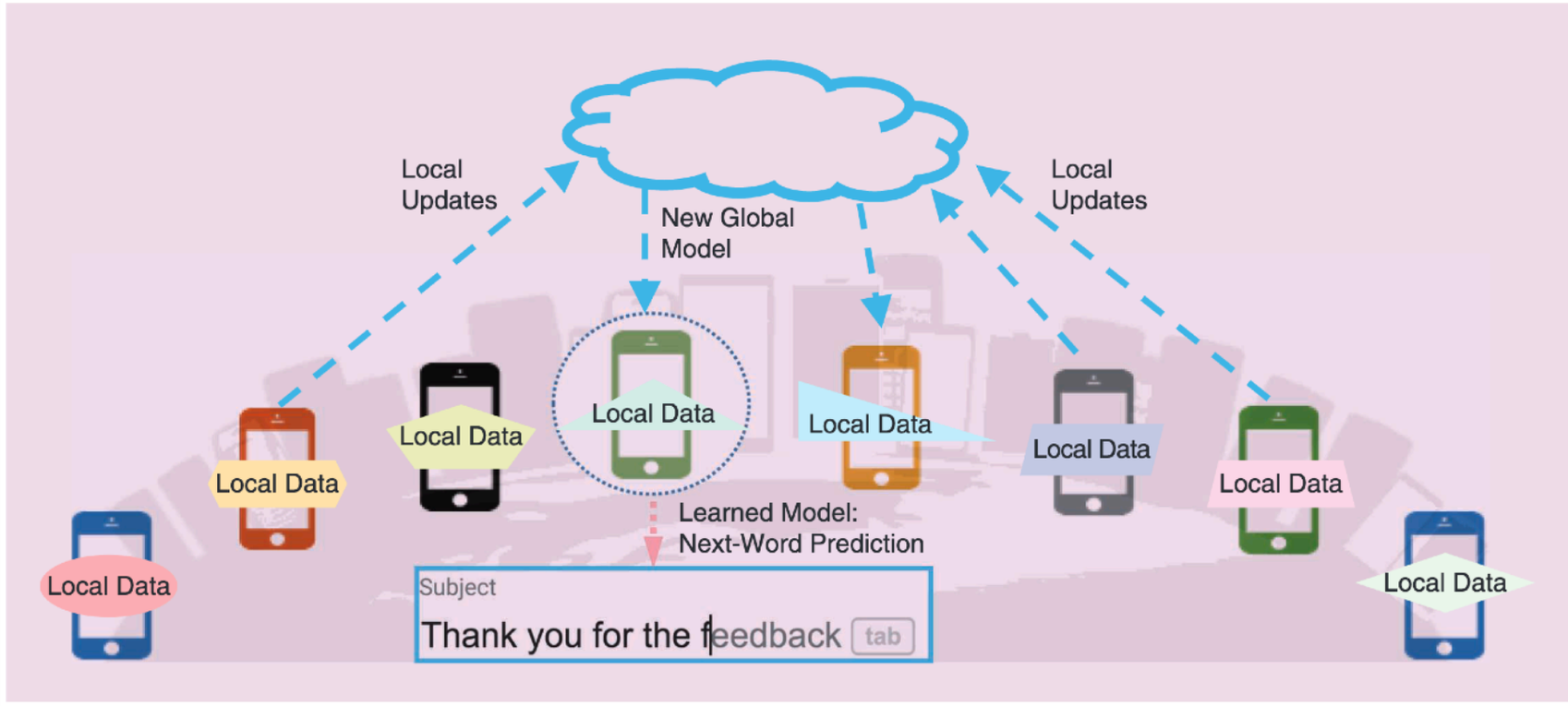


FIGURE 1. An example application of federated learning for the task of next-word prediction on mobile phones. To preserve the privacy of the text data and reduce strain on the network, we seek to train a predictor in a distributed fashion, rather than sending the raw data to a central server. In this setup, remote devices communicate with a central server periodically to learn a global model. At each communication round, a subset of selected phones performs local training on their nonidentically distributed user data, and sends these local updates to the server. After incorporating the updates, the server then sends back the new global model to another subset of devices. This iterative training process continues across the network until convergence is reached or some stopping criterion is met.

https://ieeexplore-ieee-org.sire.ub.edu/stamp/stamp.jsp?tp=&arnumber=9084352

Federated Learning

The standard federated learning problem involves learning a single global model from data stored on several devices.

We aim to learn this model under the constraint that device-data are stored and processed locally, with only intermediate updates being communicated periodically with a central server.

The goal is typically to minimize the following objective function:

$$\arg\min_{w} F(w) = \arg\min_{w} \sum_{k=1}^{m} p_k [\frac{1}{n_k} \sum_{j_k=1}^{n_k} f_k(w, x_{jk}, y_{jk})]$$

m is the number of devices, n_k is the number of samples that are available at device k, $p_k = (n_k/n)$, and n the total number of samples.

Federated Learning

Challenges:

- To fit a model to data generated by the devices in the federated network, it is important to develop communication-efficient methods that iteratively send **model updates** as part of the training process, as opposed to sending the entire data set over the network.

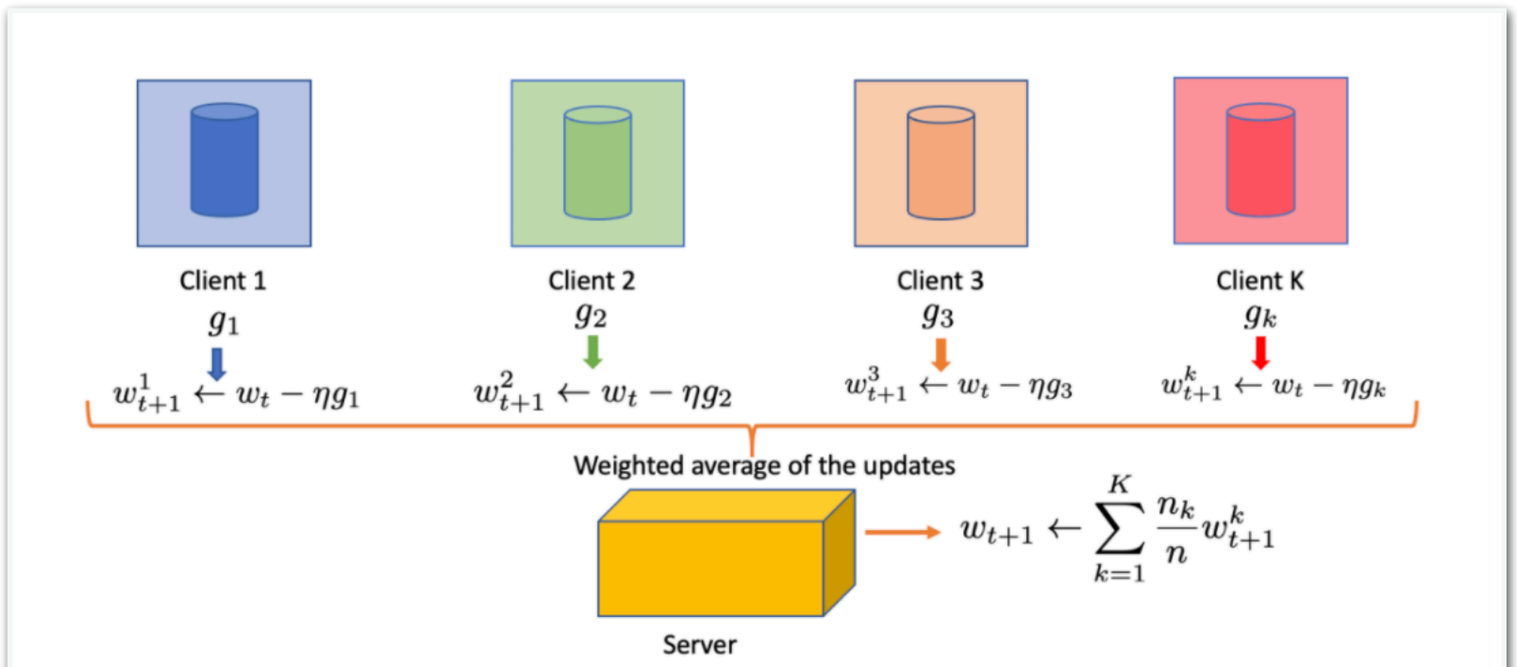
- Devices frequently generate and collect data in a highly **non-identically distributed** manner across the network, e.g., mobile phone users have varied use of language in the context of a next-word prediction task. This data-generation paradigm violates i.i.d. assumptions in optimization.

  There exist other alternatives such as simultaneously learning distinct local models via **multitask learning** frameworks.

- Federated learning makes a step toward protecting data generated on each device by sharing model updates, e.g., gradient information, instead of the raw data. However, **communicating model updates throughout the training process can nonetheless reveal sensitive information**, either to a third-party or the central server

Federated Learning

- The most commonly used method for federated learning is federated averaging, a method based on averaging local stochastic gradient descent (SGD) updates. FedAvg has been shown to work well empirically but comes without convergence guarantees and can diverge in practical settings when data are heterogeneous.



Federated Averaging(McMahan et. al, 2017) algorithm illustrated

Federated Learning

- MOCHA is an optimization framework designed for the federated setting. It can allow for personalization by learning **separate but related models for each device**, while leveraging a shared representation via multitask learning.

- This method has provable theoretical convergence guarantees for the considered objectives but is limited in its ability to scale to massive networks and is restricted to convex objectives.

MOCHA: V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," in Proc. Advances in Neural Information Processing Systems, 2017, pp. 4424–4434.

Beyond Federated Learning

The three main strategies in privacy-preserving machine learning are:

- **differential privacy** (DF) to communicate noisy data sketches,
- homomorphic encryption to operate on encrypted data,
- and multiparty computation,

beign the most widely used differential privacy.

Differential Privacy

Differential privacy

- Unlike k-Anonymity, differential privacy is a **property of algorithms**, and not a property of data.

- Differential privacy is most widely used due to its strong information **theoretic guarantees**, **algorithmic simplicity**, **and relatively small systems overhead**.

- Simply put, a randomized mechanism is **differentially private** if the change of one input element will not result in too much difference in the output distribution; this means that one cannot draw any conclusions about whether or not a specific sample is used in the learning process.

- For gradient-based learning methods, a popular approach is to apply differential privacy by randomly perturbing the intermediate output signal at each iteration.

Differential privacy

Two datasets are considered neighbors if they differ in the data of a single individual.

Definition

A function which satisfies differential privacy is often called a *mechanism*. We say that a *mechanism* F satisfies differential privacy if for all *neighboring datasets* x and x', and all possible outputs S,

$$\frac{\Pr[F(x) = S]}{\Pr[F(x') = S]} \leq e^{\epsilon} \tag{1}$$

Differential privacy

"How many individuals in the dataset are 40 years old or older?"

```
adult[adult['Age'] >= 40].shape[0]
```

```
14237
```

The easiest way to achieve differential privacy for this query is to add random noise to its answer. The key challenge is to add enough noise to satisfy the definition of differential privacy, but not so much that the answer becomes too noisy to be useful. To make this process easier, some basic *mechanisms* have been developed in the field of differential privacy, which describe exactly what kind of - and how much - noise to use. One of these is called the *Laplace mechanism* [4].

Definition

According to the Laplace mechanism, for a function $f(x)$ which returns a number, the following definition of $F(x)$ satisfies ϵ-differential privacy:

$$F(x) = f(x) + \text{Lap}\left(\frac{s}{\epsilon}\right) \tag{2}$$

where s is the *sensitivity* of f, and $\text{Lap}(S)$ denotes sampling from the Laplace distribution with center 0 and scale S.

Differential privacy

- Before applying the perturbation it is common to clip the gradients to bound the influence of each example on the overall update.

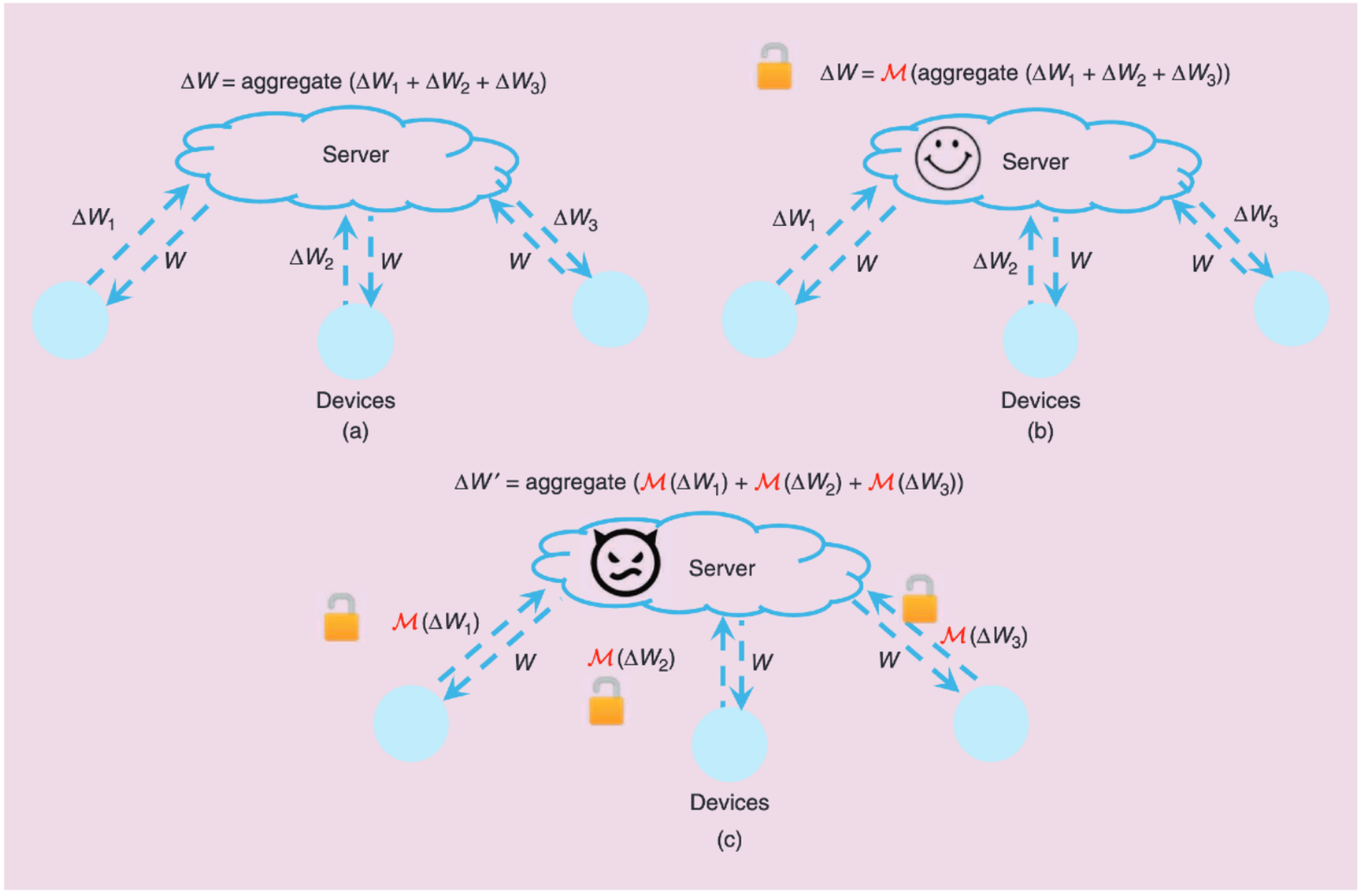
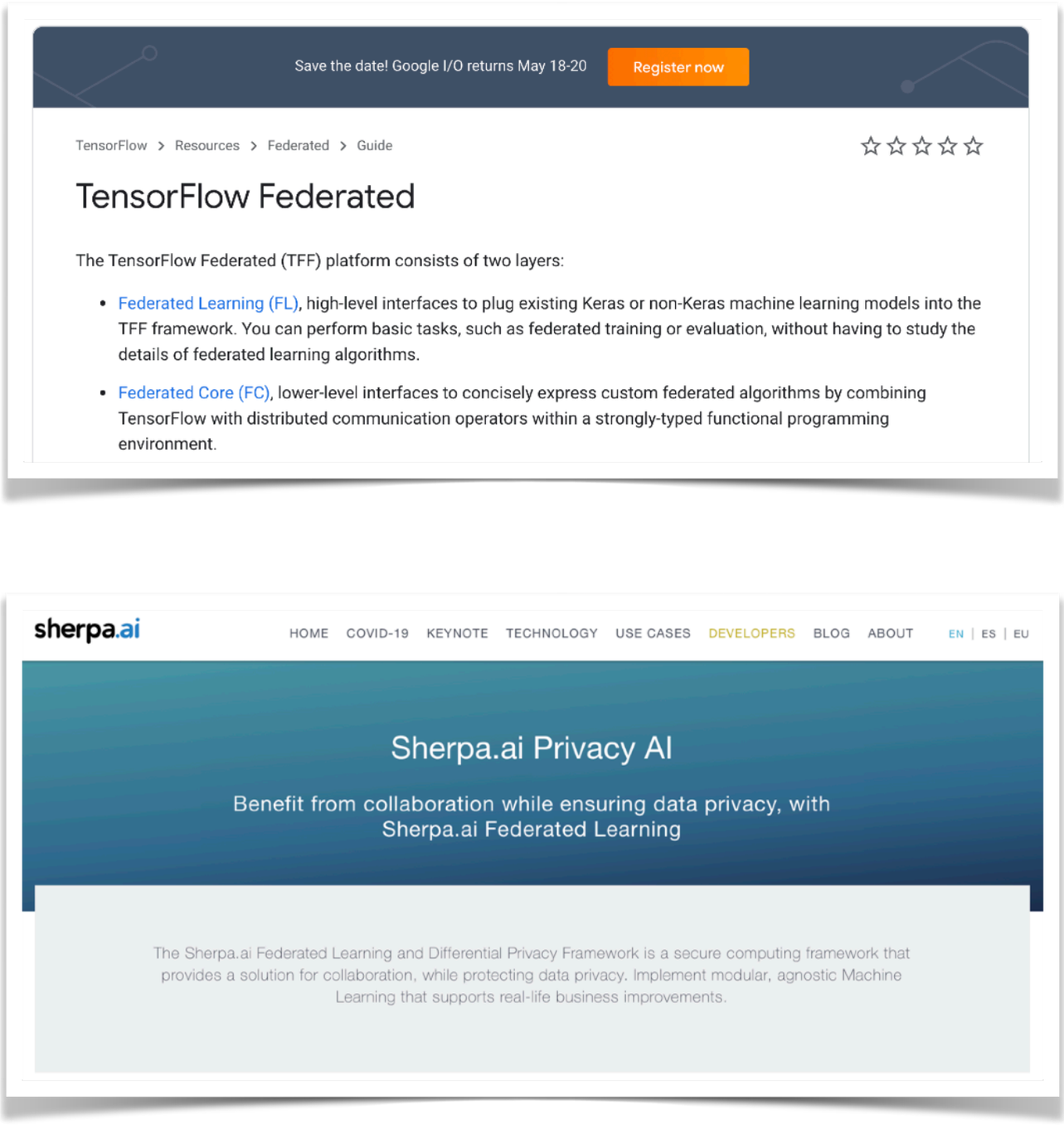


FIGURE 5. An illustration of different privacy-enhancing mechanisms in one round of federated learning. $\mathcal{M}$ denotes a randomized mechanism used to privatize the data. (a) Federated learning without additional privacy protection mechanisms, (b) global privacy, where a trusted server is assumed, and (c) local privacy, where the central server may be malicious.

https://ieeexplore-ieee-org.sire.ub.edu/stamp/stamp.jsp?tp=&arnumber=9084352

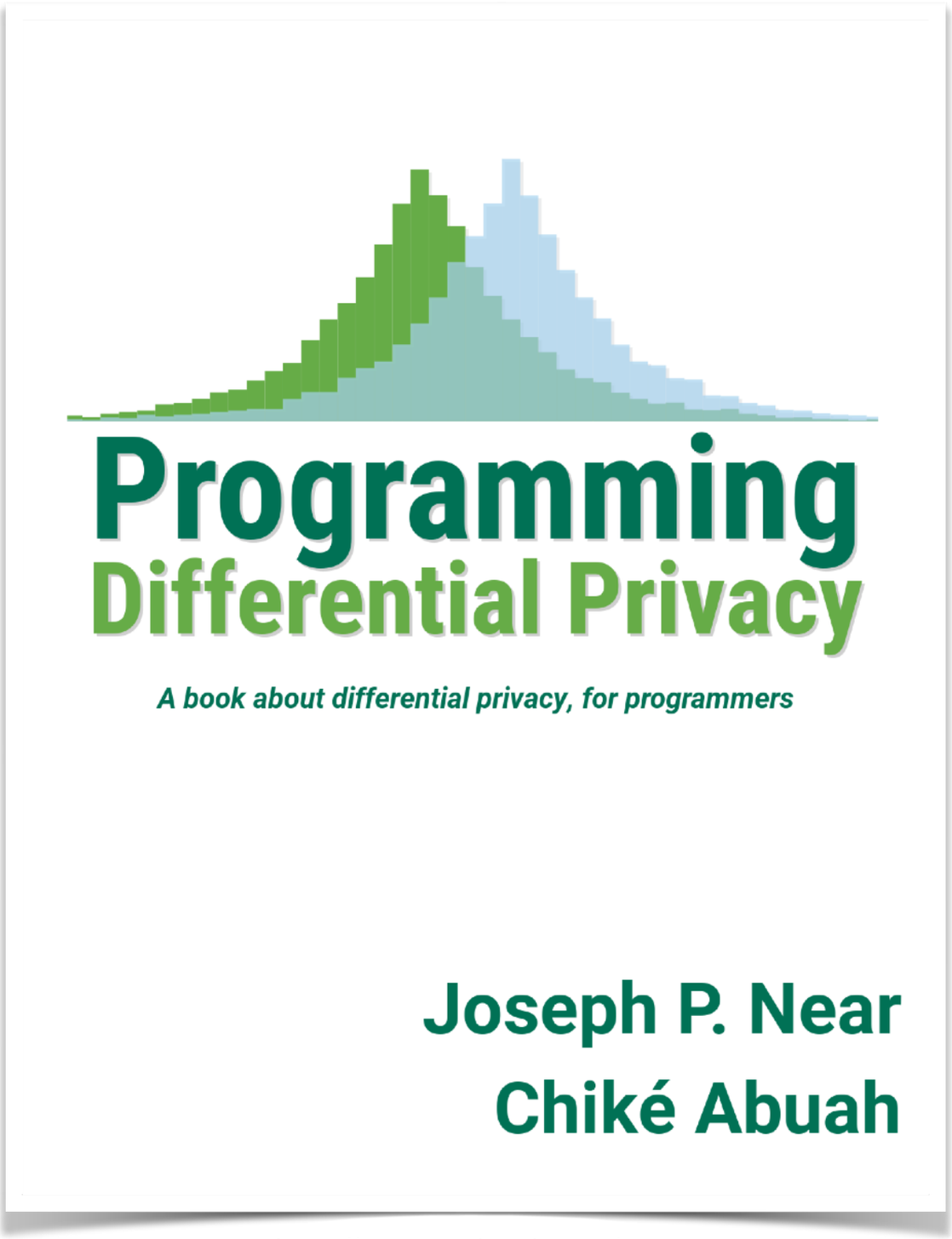
Resources

Software


Save the date! Google I/O returns May 18-20
Register now
TensorFlow > Resources > Federated > Guide
TensorFlow Federated
The TensorFlow Federated (TFF) platform consists of two layers:
Federated Learning (FL), high-level interfaces to plug existing Keras or non-Keras machine learning models into the TFF framework. You can perform basic tasks, such as federated training or evaluation, without having to study the details of federated learning algorithms.
Federated Core (FC), lower-level interfaces to concisely express custom federated algorithms by combining TensorFlow with distributed communication operators within a strongly-typed functional programming environment.

sherpa.ai
HOME COVID-19 KEYNOTE TECHNOLOGY USE CASES DEVELOPERS BLOG ABOUT EN | ES | EU
Sherpa.ai Privacy AI
Benefit from collaboration while ensuring data privacy, with Sherpa.ai Federated Learning
The Sherpa.ai Federated Learning and Differential Privacy Framework is a secure computing framework that provides a solution for collaboration, while protecting data privacy. Implement modular, agnostic Machine Learning that supports real-life business improvements.

Software



https://programming-dp.com/

Software

Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, et al. 2018. Differential Privacy: A Primer for a Non-Technical Audience. Vanderbilt Journal of Entertainment & Technology Law 21 (1): 209.

Differential Privacy: A Primer for a Non-Technical Audience

*Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke & Salil Vadhan**

ABSTRACT

Differential privacy is a formal mathematical framework for quantifying and managing privacy risks. It provides provable privacy protection against a wide range of potential attacks, including those

---

* Alexandra Wood is a Fellow at the Berkman Klein Center for Internet & Society at Harvard University. Micah Altman is Director of Research at MIT Libraries. Aaron Bembenek is a PhD student in computer science at Harvard University. Mark Bun is a Google Research Fellow at the Simons Institute for the Theory of Computing. Marco Gaboardi is an Assistant Professor in the Computer Science and Engineering department at the State University of New York at Buffalo. James Honaker is a Research Associate at the Center for Research on Computation and Society at the Harvard John A. Paulson School of Engineering and Applied Sciences. Kobbi Nissim is a McDevitt Chair in Computer Science at Georgetown University and an Affiliate Professor at Georgetown University Law Center; work towards this document was completed in part while the Author was visiting the Center for Research on Computation and Society at Harvard University. David R. O'Brien is a Senior Researcher at the Berkman Klein Center for Internet & Society at Harvard University. Thomas Steinke is a Research Staff Member at IBM Research – Almaden. Salil Vadhan is the Vicky Joseph Professor of Computer Science and Applied Mathematics at Harvard University.

This Article is the product of a working group of the *Privacy Tools for Sharing Research Data* project at Harvard University (http://privacytools.seas.harvard.edu). The working group discussions were led by Kobbi Nissim. Alexandra Wood and Kobbi Nissim are the lead Authors of this Article. Working group members Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke, Salil Vadhan, and Alexandra Wood contributed to the conception of the Article and to the writing. The Authors thank John Abowd, Scott Bradner, Cynthia Dwork, Simson Garfinkel, Caper Gooden, Deborah Hurley, Rachel Kalmar, Georgios Kellaris, Daniel Muise, Michel Reymond, and Michael Washington for their many valuable comments on earlier versions of this Article. A preliminary version of this work was presented at the 9th Annual Privacy Law Scholars Conference (PLSC 2017), and the Authors thank the participants for contributing thoughtful feedback. The original manuscript was based upon work supported by the National Science Foundation under Grant No. CNS-1237235, as well as by the Alfred P. Sloan Foundation. The Authors' subsequent revisions to the manuscript were supported, in part, by the US Census Bureau under cooperative agreement no. CB16ADR0160001. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the Authors and do not necessarily reflect the views of the National Science Foundation, the Alfred P. Sloan Foundation, or the US Census Bureau.