



Tomas Pytlíček

@Pytlíček



🤔 This is really weird. In [#WhatsApp](#), I started to see messages that I know 100% that I deleted 2 days ago?! WTF is happening there? I think this is a really big violation of privacy! I see the messages from a month ago, with my disappearing messages setting turned on?! Gosh

10:05 AM · Oct 4, 2021 · Twitter Web App

32 Retweets **22** Quote Tweets **110** Likes



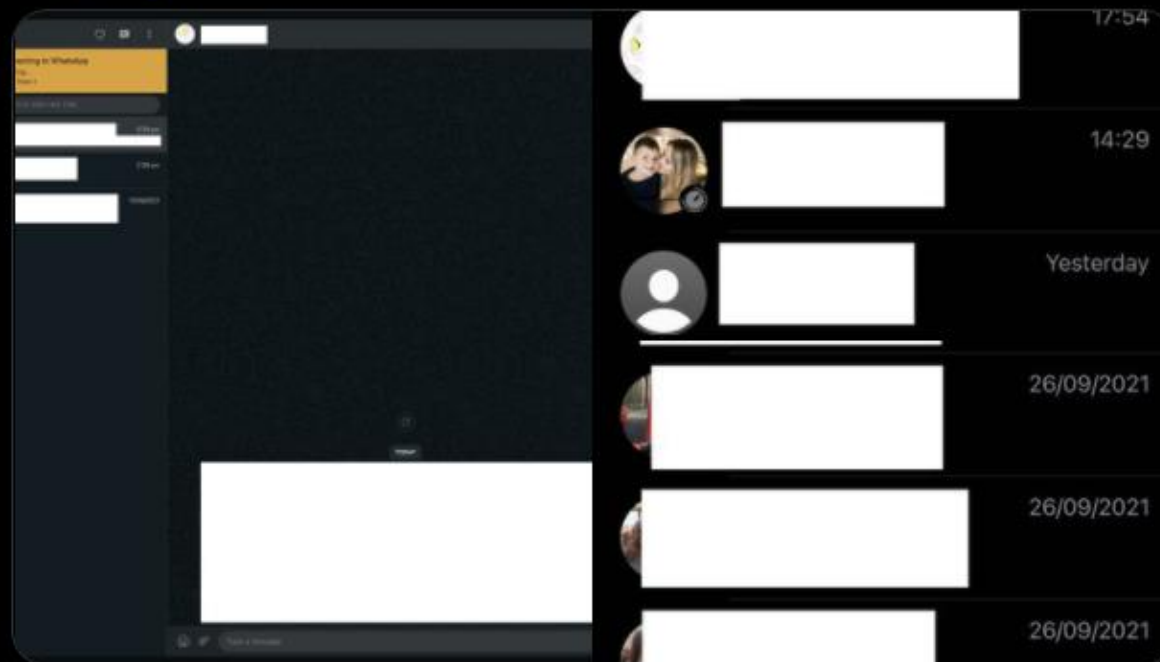


Tomas Pytlíček @Pytlícek · Oct 4

...

Replying to @Pytlíček

3 chats before the outage and now 15+ or more chats which I deleted before the week or two. This is really bad. #WhatsApp you lie about privacy. F*** you 🖐️



1



7



30





Tomas Pytlíček @Pytlíček · Oct 4

...

OK, 1 hour ago all chats appeared with messages, read status, etc... Now the text is gone. It looks like they are doing a replay of actions like deleting. But it doesn't change the fact that I deleted those messages a long time ago. It should no longer be on [#WhatsApp](#) servers 🇸🇰



8

4

37





DATA GOVERNANCE

OUR GOALS

Encoding categorical
PII Removal

Capital One
Legendary Pictures

Building skills for manipulating
and interpreting data

Systems in delivering data
forward products

Delivering business outcomes

Discoverability
Accountability
Security

WHEN TO BE CAUTIOUS ABOUT SHARING DATA

- <https://medium.com/97-things/accessible-data-empowers-organizations-but-it-can-also-cause-problems-5db1c12857d5>
- Part 1: Role play with your homework partner for 10 mins:
 - Partner A: Has the sales data for individual retail stores at the SKU level and DOES NOT want to share the data, aggregates are published and available as part of the monthly powerpoint sales update
 - Partner B: Works on the innovation team and trying to forecast seasonal / uncover new trends in data, NEED the underlying data
 - Aggregated metrics obscure the by product and smaller trends that you're trying to uncover
 - Make notes of the arguments each partner uses: what are the key arguments against sharing the data, what about for? Is there a middle ground? Are their arguments that are more convincing?
- Part 2: Volunteers to role play in front of the class

DATA SILOS

- Organizational / Cultural Posture
 - Is data default open or closed?
 - Sensitive data types: HR/employee salaries
- Trend: Democratization of data – increased focus on data based decision making
- Moving from closed to open paradigm very challenging
 - Encoded business logic and fiefdoms



SHADOW IT

- In big organizations, **shadow IT** refers to information technology (IT) systems deployed by departments other than the central IT department, to work around the shortcomings of the central information systems.
- Shadow IT systems are an important source of innovation, and shadow systems may become prototypes for future central IT solutions.
- Tension between the time to delivery, speed to market from the engineering team, and what insights the business might need.
- Often: security controls are not in place, lack of scale to go to product, or the code quality may not be exactly where you need it to be



There's always a way, we just need to work it out. And I think that my attitude being that, at the start I think it scared my team a little bit, because they're like, "Oh, saying no was our way. So now we're going to have to come up with a smarter answer." And I think we managed it, which is we created a team that is only small, it's less than 10 people, but that is dedicated to people, to help them self-serve, so do it in the right way, and they're only dedicated to providing the tools for them to do it in a supported way.

Stuart Hughes - Chief information and digital officer Rolls Royce

MODERN SOLUTIONS TO SILOED DATA

- From previous class: data exchanges
- Data clean rooms - A Data Clean Room is a secure, protected environment where PII (Personally Identifiable Information) data is anonymized, processed and stored to be made available for measurement, or data transformations in a privacy-compliant way. The raw PII, is made available to the brand and is only viewable by the brand.



A pair of blue and red binoculars is shown in a close-up, resting on a blue wooden bench. The binoculars have a textured blue body and red accents around the objective lenses. The background is a soft-focus landscape of a beach and ocean under a clear sky. The word "DISCOVERABILITY" is written in white, bold, sans-serif capital letters across the center of the image, with a thin white horizontal line underneath it.

DISCOVERABILITY

Many sources. One truth.

Metaphor goes beyond the traditional data catalog by providing a system-of-record for your data ecosystem



Data Catalog



Data Context



Data Discovery



Data Insights



DATA INSIGHTS

Optimize your data investments

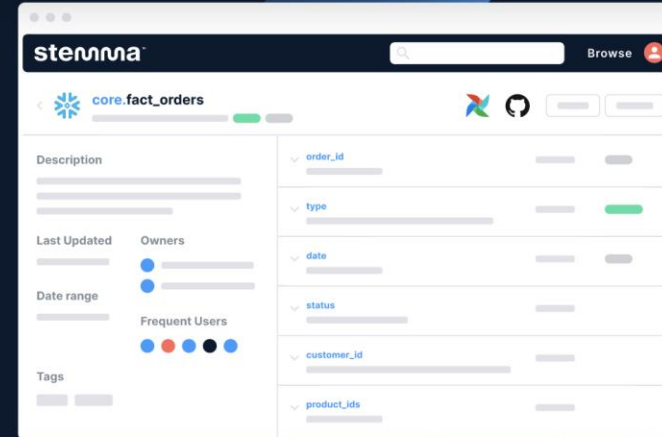
Optimize how your data team spends time and money with insight into how data is utilized. Shift investment away from underutilized datasets, dashboards and jobs and towards higher value data assets.

Gain Total Trust in Your Data

Everyone has access to data, but few know what exists, what's trustworthy and how to use it. Stemma makes finding trustworthy data easy and offers an always up-to-date view of your data's usage at any time.

Get Started

See Demo



Stemma for Data Scientists & Analysts

Rather than being the last to find out when data gets shut off or delayed, Stemma users are always in the loop and have oversight on what's changing with their data.

We keep users up-to-date through automated documentation based on common usage patterns.



Stemma for Data Engineers

Not knowing who or what a data change will affect means that data engineers simply have to spray and pray.

The volume of data that companies need to collect and process is growing exponentially. Stemma removes the complexity and stress of tracking and navigating the flow of new and ever-



Stemma for Data Governance

Organizations often rely on Slack conversations and continuous shoulder-tapping to gain trust in data. This is error-prone and time-consuming.

Stemma uniquely augments your data with automated documentation, so you don't have to document and curate every single data set.

All of your metadata in one place

Secoda is a single space for all your data knowledge. Build a full picture of your data with all the context you need. Let you and your team search, organize, and collaborate on data knowledge in one place.

Try for free

✓ No credit card required ✓ No coding needed ✓ On-prem available

The screenshot displays the Secoda Data catalog interface. At the top, there are four tabs: 'Data catalog' (selected), 'Data analysis', 'Data dictionary', and 'Data requests'. The main content area shows a table entry for 'rfam.database_link'. It includes a visualization of monthly business signups since 2019, created on 20/01/2020 and last updated on 04/03/2021. The table is owned by Adam Newman and was updated on 09/01/2021. It has 12 views, 21 likes, and 2 comments. The table has columns for 'Name', 'Description', 'Type', 'Mode', and 'Tags'. The 'Name' column lists '_id', '_cdc_batched_at', and '_cdc_batched_at'. The 'Description' column describes the core ID variable and timestamp without time zone. The 'Type' column lists 'character varying' and 'timestamp'. The 'Mode' column lists 'required' and 'required'. The 'Tags' column lists 'JOIN' and 'PI'. The 'Frequent Users' section shows three users and a query count of 231 times.

Secoda

Data catalog | Data analysis | Data dictionary | Data requests

rfam.database_link Visualization of monthly business signups since 2019 Created at: 20/01/2020 Last updated: 04/03/2021

Owner: Adam Newman Updated on 09/01/2021

12 Views 21 Likes 2 Comments

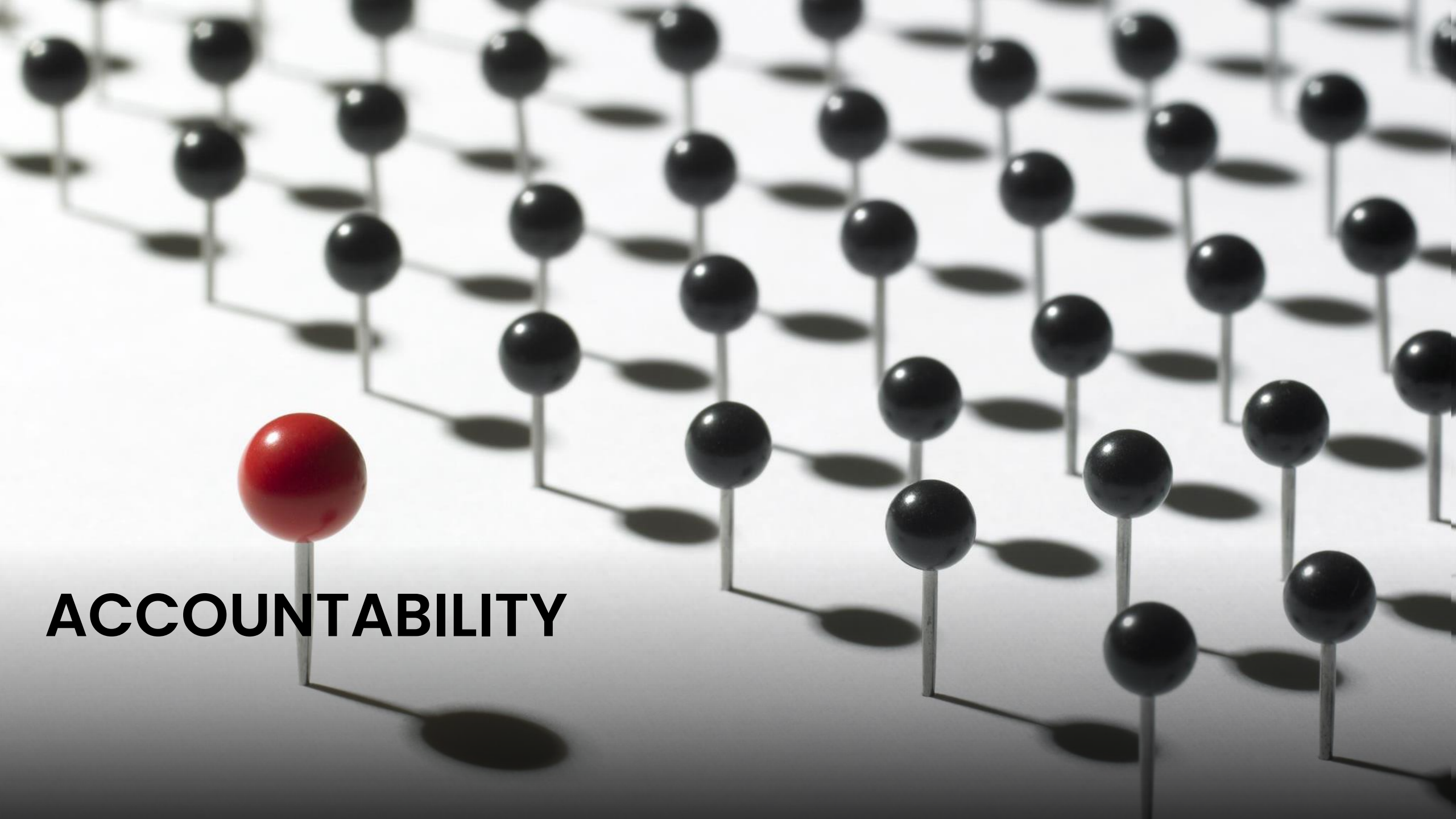
Frequent Users: Queried 231 times

Name	Description	Type	Mode	Tags
_id	The core ID variable for the table. Used to join to other tables	character varying	required	JOIN PI
_cdc_batched_at	timestamp without time zone	timestamp	required	
_cdc_batched_at				



DISCOVERABILITY PAIN POINTS

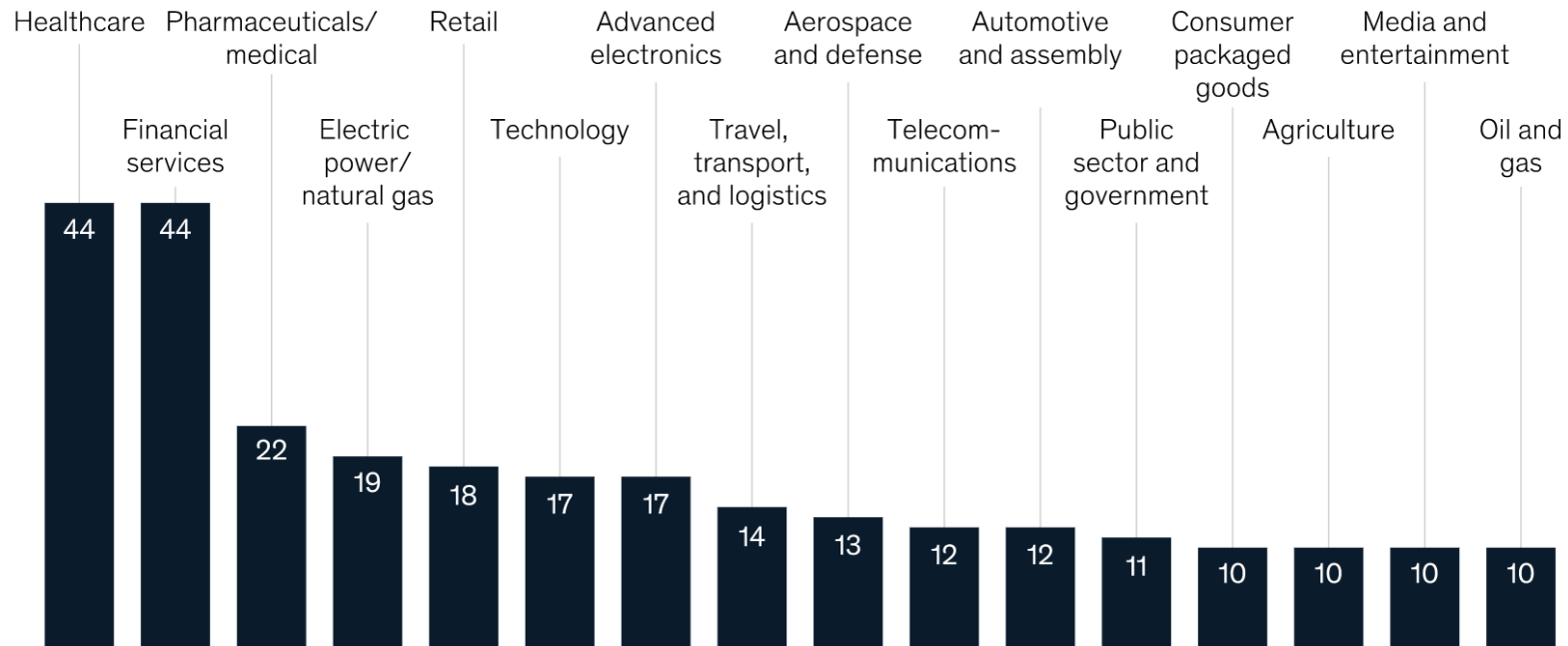
- Hard to know what data is available and how to get access to it
- Metadata Management – data about the data, what type of data is it.
 - PII, Confidential.
 - How is the data derived
- Lineage – where did the data come from – where is it being used
- Access control – who should be allowed to access the data



ACCOUNTABILITY

CONSUMERS MOST TRUST HEALTHCARE & FINANCIAL SERVICES WITH THEIR DATA

Respondents choosing a particular industry as most trusted in protecting of privacy and data,
% (n = 1,000)



Source: McKinsey Survey of North American Consumers on Data Privacy and Protection, 2019

MAJOR DATA BREACHES

The 10 Biggest Data Breaches of 2021

- *Nearly 215.4 million individuals were impacted by the 10 biggest data breaches of 2021, with three of the 10 largest breaches occurring at technology companies and four involving the exposure of sensitive records.*

<https://www.crn.com/slideshows/security/the-10-biggest-data-breaches-of-2021/2>



2. T-Mobile

Number Of Individuals Impacted: 47.8 Million

T-Mobile confirmed Aug. 17 that its systems had on March 18 been subject to a criminal cyberattack that **compromised data from millions of customers**, former customers and prospective customers. The compromised information included names, driver's licenses, government identification numbers, Social Security numbers, dates of birth, T-Mobile prepaid PINs, addresses and phone numbers, T-Mobile said.

The image is a screenshot of a data breach report table for OneMoreLead. The table has six columns: Name, Title, Company, Industry, Location, and Country. It contains 10 rows of data, showing various individuals and their associated information. The data is as follows:

Name	Title	Company	Industry	Location	Country
Kent Ba	Managing Branch Broker	Intel Co	Real Estate	Denver, Colorado, USA	USA
Alex Occhetti	Lead Designer	Macbook pro	Mechanical Engineering	Kanchipuram, Tamil Nadu, India	India
Gordon Ramsey	Head Pro Cook	Ramsey Kitchen	Education Management	Orlando, Florida, USA	USA
Kent Ba	Managing Branch Broker	Intel Co	Real Estate	Denver, Colorado, USA	USA
Alex Occhetti	Lead Designer	Macbook pro	Mechanical Engineering	Kanchipuram, Tamil Nadu, India	India
Gordon Ramsey	Head Pro Cook	Ramsey Kitchen	Education Management	Orlando, Florida, USA	USA
Kent Ba	Managing Branch Broker	Intel Co	Real Estate	Denver, Colorado, USA	USA
Gordon Ramsey	Head Pro Cook	Ramsey Kitchen	Education Management	Orlando, Florida, USA	USA
Kent Ba	Managing Branch Broker	Intel Co	Real Estate	Denver, Colorado, USA	USA
Alex Occhetti	Lead Designer	Macbook pro	Mechanical Engineering	Kanchipuram, Tamil Nadu, India	India

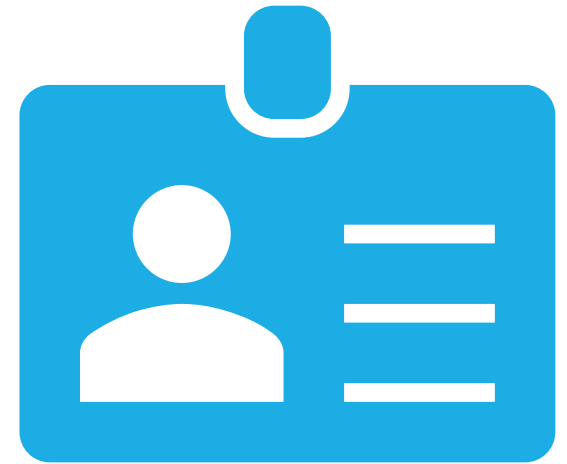
1. OneMoreLead

Number Of Individuals Impacted: 63 Million

vpnMentor's research team discovered in August that B2B marketing company OneMoreLead was storing the private data of at least 63 million Americans on an unsecured database, which the company had left completely open. As a result, names, email addresses and workplace information were **exposed to anyone with a web browser**, according to vpnMentor.

PII

- Personally Identifiable Information (PII) is a legal term pertaining to information security environments. While PII has several formal definitions, generally speaking, it is information that can be used by organizations on its own or with other information to identify, contact, or locate a single person, or to identify an individual in context.



OVERVIEW OF REGULATIONS GOVERNING PII

NOT LEGAL ADVICE

- GDPR – The [General Data Protection Regulation \(GDPR\)](#) is the toughest privacy and security law in the world. Though it was drafted and passed by the European Union (EU), it imposes obligations onto organizations anywhere, so long as they target or collect data related to people in the EU. The regulation was put into effect on May 25, 2018. The GDPR will levy harsh fines against those who violate its privacy and security standards, with penalties reaching into the tens of millions of euros.
 - Accountability, Security, People's Privacy Rights
- CCPA – The [California Consumer Privacy Act of 2018](#) (CCPA) gives consumers more control over the personal information that businesses collect about them and the [CCPA regulations](#) provide guidance on how to implement the law. This landmark law secures new privacy rights for California consumers, including:
 - Right to know, right to delete, right to opt out, right to non discriminate
- GDPR Fines: Amazon \$877M, WhatsApp \$255M, H&M, Google...
- Consumer health data in US: HIPAA, PCI Compliance with card networks

SARBANES OXLEY (SOX) COMPLIANCE

NOT LEGAL ADVICE

01

Keep data
secure and free
of tampering

02

Track attempted
security
breaches and
resolutions

03

Keep event logs
available for
independent
auditing

04

Prove
compliance for
past 90 days

Amazon Macie

Discover and protect your sensitive data at scale

Get started with Amazon Macie

Amazon Macie is a fully managed data security and data privacy service that uses machine learning and pattern matching to discover and protect your sensitive data in AWS.

As organizations manage growing volumes of data, identifying and protecting their sensitive data at scale can become increasingly complex, expensive, and time-consuming. Amazon Macie automates the discovery of sensitive data at scale and lowers the cost of protecting your data. Macie automatically provides an inventory of Amazon S3 buckets including a list of unencrypted buckets, publicly accessible buckets, and buckets shared with AWS accounts outside those you have defined in AWS Organizations. Then, Macie applies machine learning and pattern matching techniques to the buckets you select to identify and alert you to sensitive data, such as personally identifiable information (PII).

<https://www.youtube.com/watch?v=CenD1dq3xj8&feature=youtu.be>

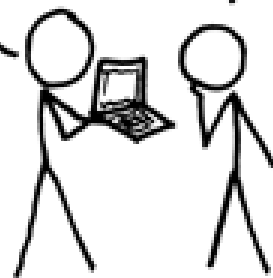
SECURITY

A CRYPTO NERD'S IMAGINATION:

HIS LAPTOP'S ENCRYPTED.
LET'S BUILD A MILLION-DOLLAR
CLUSTER TO CRACK IT.

NO GOOD! IT'S
4096-BIT RSA!

BLAST! OUR
EVIL PLAN
IS FOILED!



WHAT WOULD ACTUALLY HAPPEN:

HIS LAPTOP'S ENCRYPTED.
DRUG HIM AND HIT HIM WITH
THIS \$5 WRENCH UNTIL
HE TELLS US THE PASSWORD.



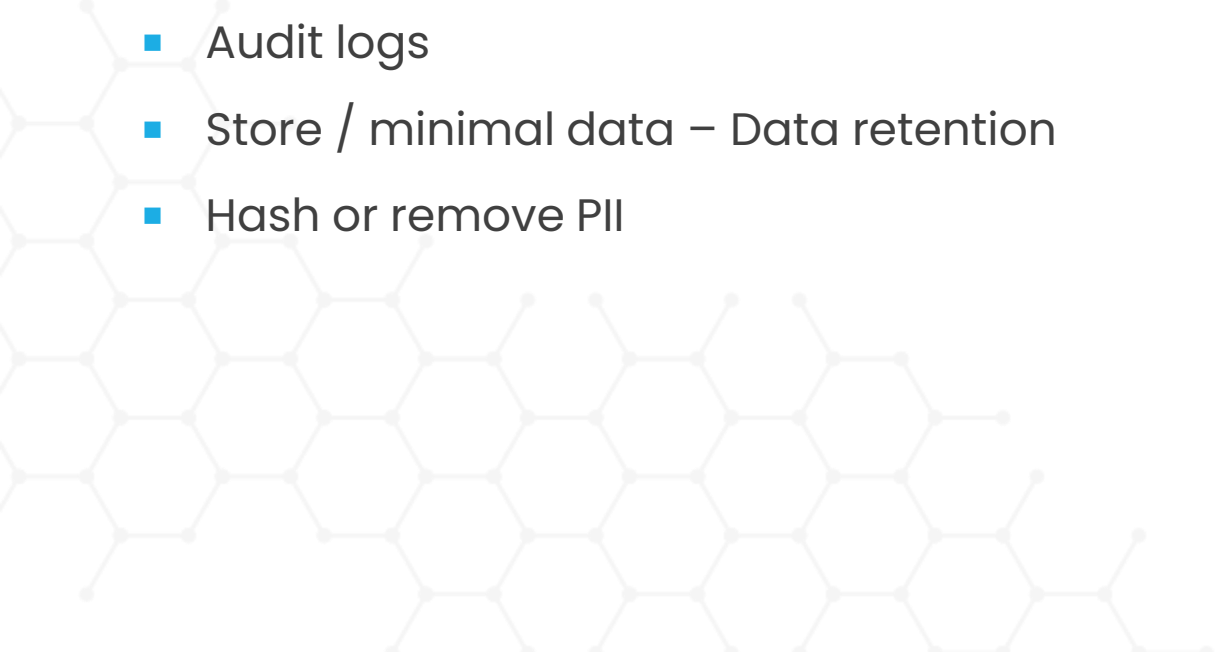
CAPITAL ONE 2019

- Affected 100 million individuals in the United States and approximately 6 million in Canada.
- Information on consumers and small businesses as they applied for credit card products from 2005 through early 2019.
 - names, addresses, zip codes/postal codes, phone numbers, email addresses, dates of birth, and self-reported income.
- The outside individual who took the data was captured by the FBI





LESSONS FROM CAP ONE

- Capital One settles a class-action lawsuit for \$190 million in a 2019 hacking.
 - Audit logs
 - Store / minimal data – Data retention
 - Hash or remove PII
- 

SECURITY



ACCESS CONTROL



AUDITING

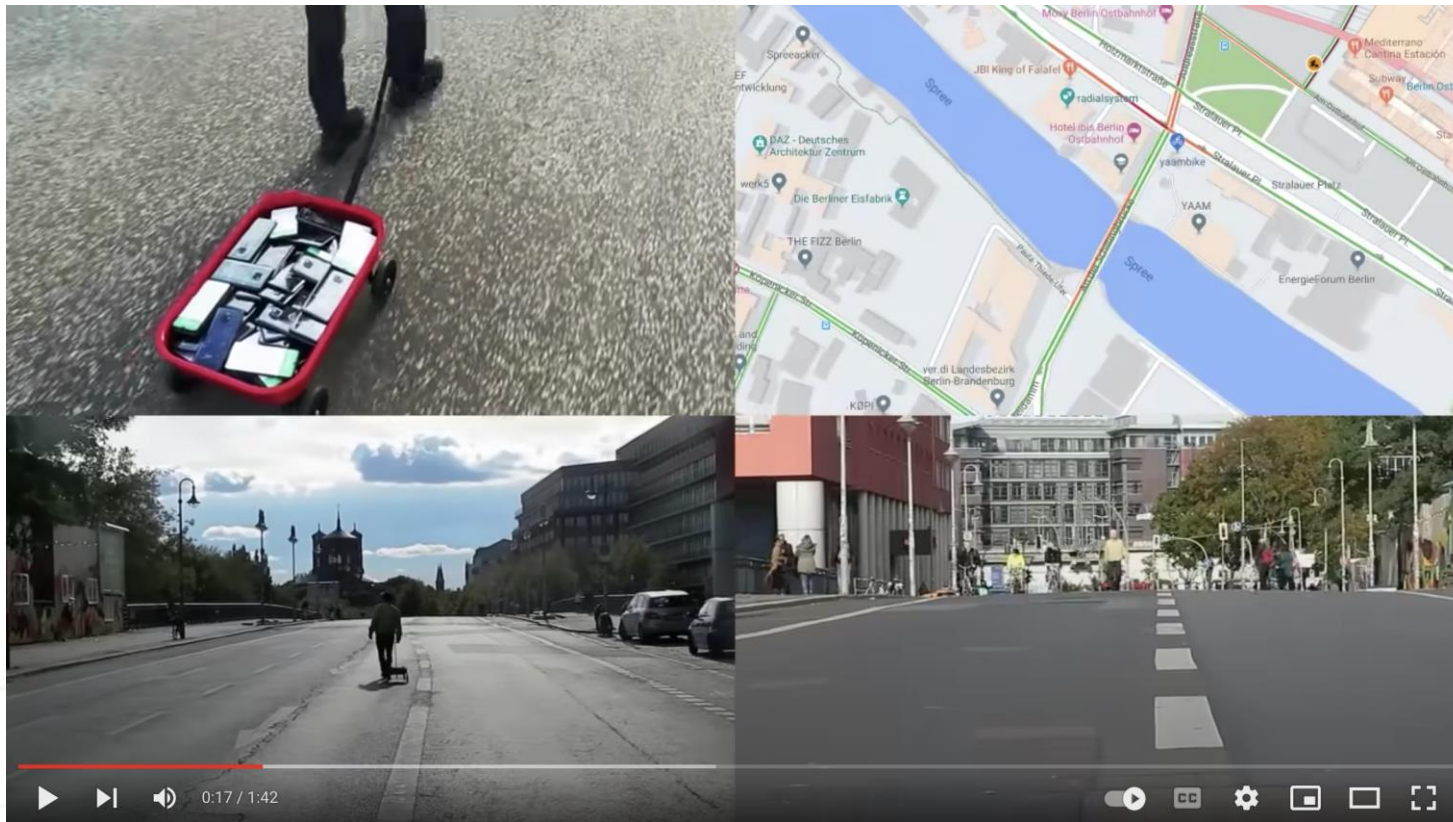


CONFIDENTIALITY /
PRIVACY



INTEGRITY

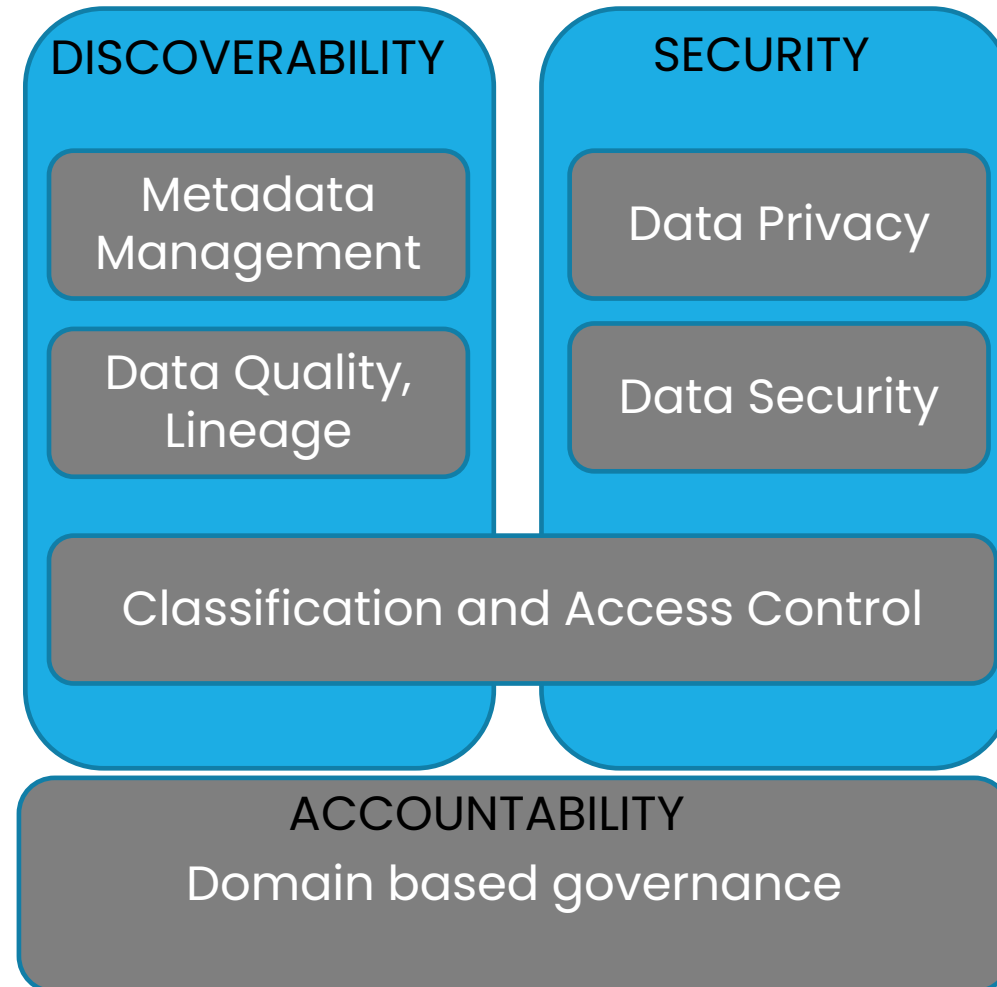
ADVERSARIAL DATA



https://www.youtube.com/watch?v=k5eL_al_m

DATA GOVERNANCE

Creating trust in data across stakeholders

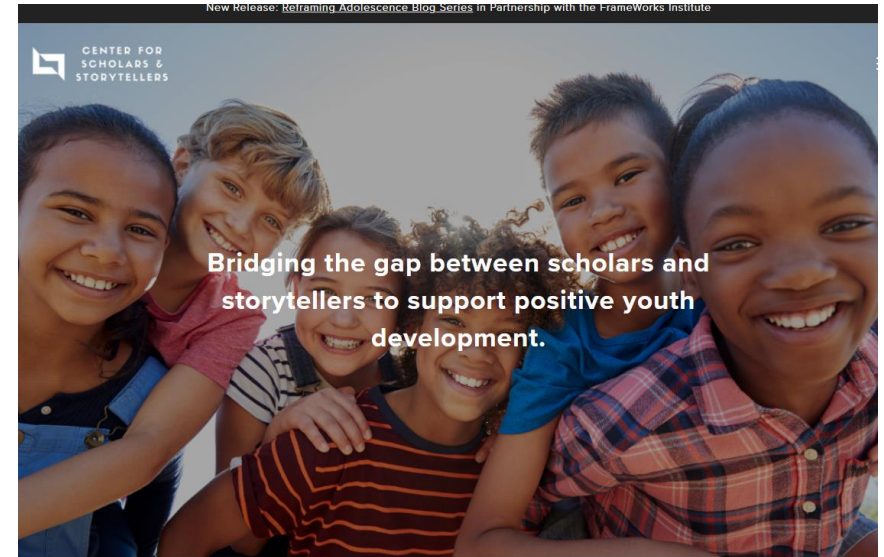




LEGENDARY

JUMPCUT

Connecting storytellers and audiences to unlock collective creativity



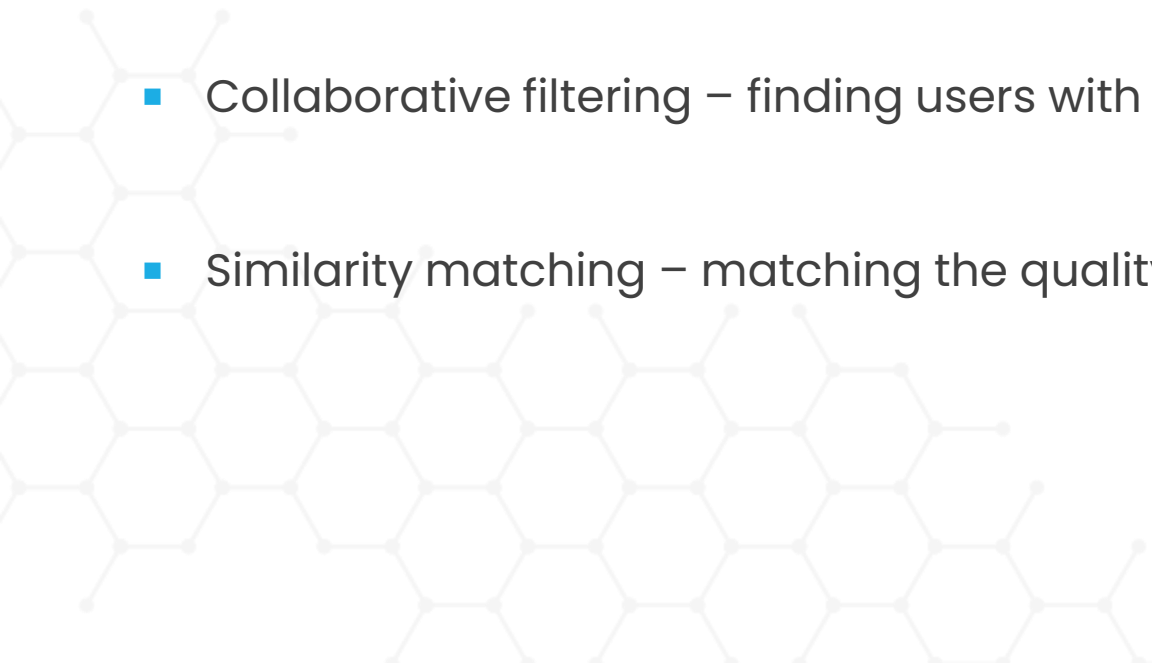
The Center for Scholars & Storytellers is an organization based at UCLA dedicated to bridging the gap between scholars and storytellers to promote positive youth development.

UCLA

OTHER ENTERTAINMENT USES



RECOMMENDER SYSTEMS

- Content based recommendation – finding songs in the same genre
 - Collaborative filtering – finding users with similar tastes and matching
 - Similarity matching – matching the quality or other latent aspects
- 

The Netflix logo is displayed in a large, bold, white, sans-serif font. Each letter has a thick black outline and a 3D shadow effect, giving it a sense of depth. The letters are set against a solid red rectangular background. Above the red background, there are three horizontal bars: a dark grey bar on the left, a light blue bar in the middle, and a light grey bar on the right.

NETFLIX

NETFLIX PRIZE

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of

and sparsity. Each record contains many attributes (*i.e.*, columns in a database schema), which can be viewed as dimensions. Sparsity means that for the average record, there are no “similar” records in the multi-dimensional space defined by the attributes. This sparsity is empirically well-established [7, 4, 19] and related to the “fat tail” phenomenon: individual transaction and preference records tend to include statistically rare attributes.

Our contributions. Our first contribution is a formal model for privacy breaches in anonymized micro-data (section 3). We present two definitions, one based on the probability of successful de-anonymization, the other on the amount of information recovered about the target. Unlike previous work [25], we do not assume *a priori* that the adversary's knowledge is limited to a fixed set of “quasi-identifier” attributes. Our model thus en-

These are the actual locations
for millions of Americans. At the New York
Stock Exchange ...

- <https://www.nytimes.com/interactive/2019/12/20/opinion/location-data-national-security.html>



APPENDIX





LEGENDARY PICTURES

- Marketing effectiveness, which movies to make (market sizing), which actors to cast, and when to release the finished product.
- Scaling marketing efforts: As a movie's release date nears, the analytics group looks for people who are talking about it enthusiastically on social networks like Twitter and Facebook — those people are likely movie-goers. It then starts making small purchases in online advertising targeted at those people, to see which taglines, trailers, and artwork are most likely to earn a click or a view. Marolda says it amounts to “tens of thousands” of mini ad campaigns to see which groups of consumers respond best to which ads — and who turns out just not to be interested. Then, the analytics group builds bigger campaigns, both online, on TV, and on other media, based on what messages worked in those tests, and what it has learned about the people who responded.