

# Lecture 4: The Lifecycles of Data

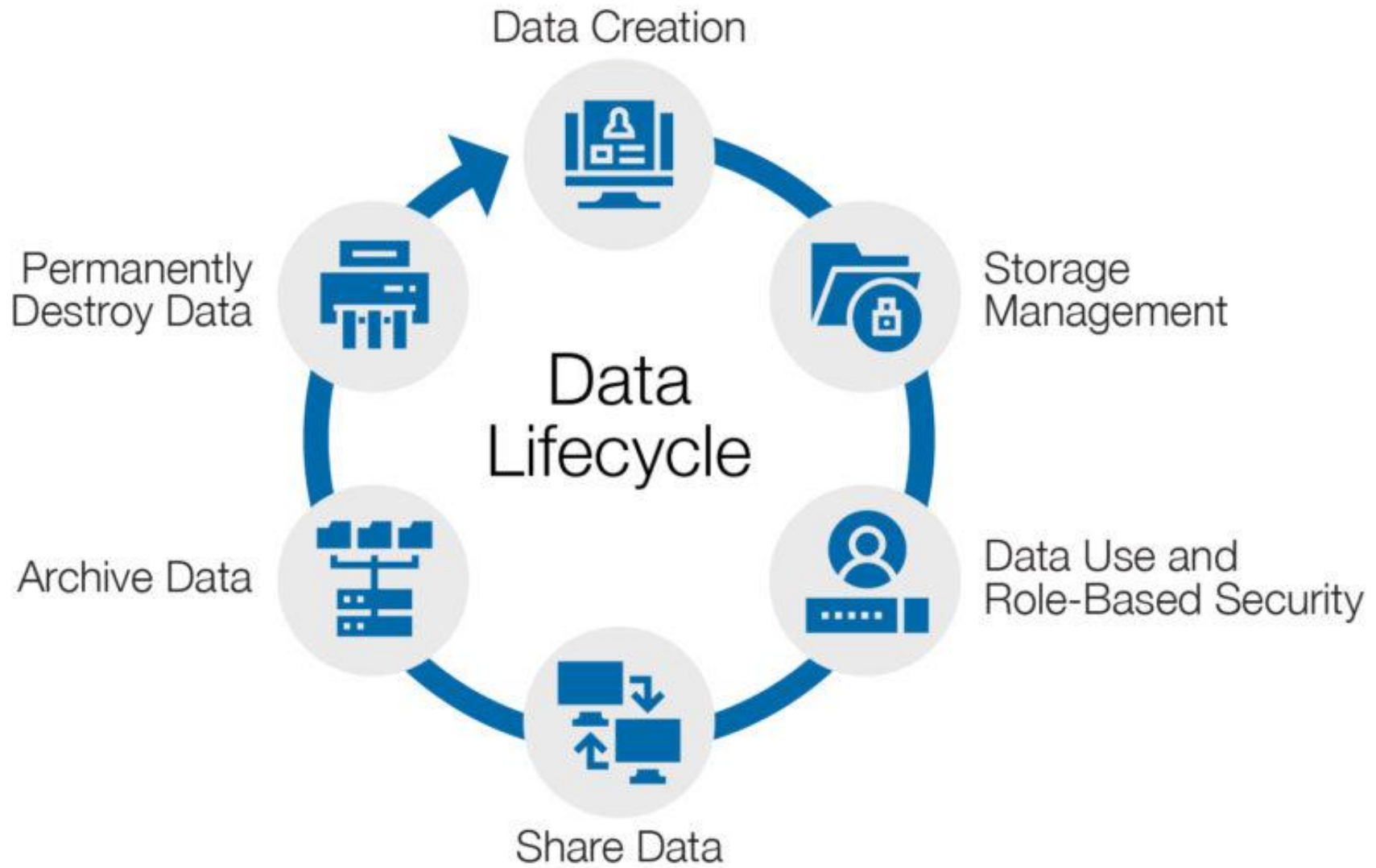
CMSC 25910

Spring 2024

The University of Chicago

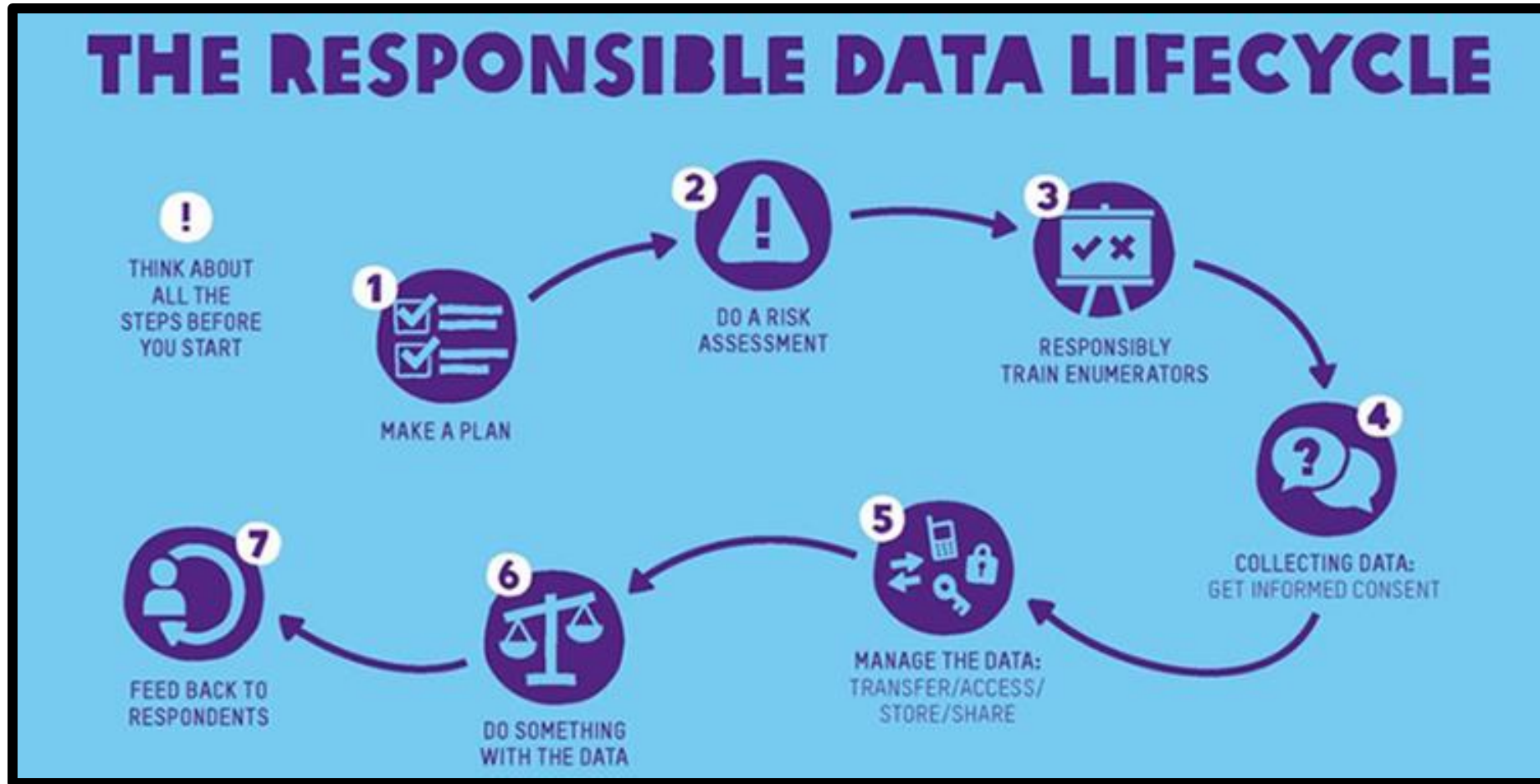


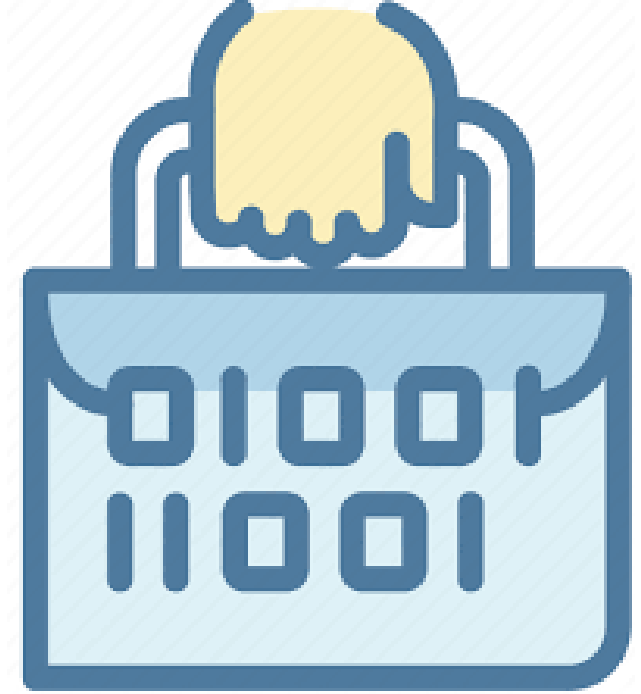
THE UNIVERSITY OF  
CHICAGO



Graphic from <https://www.spirion.com/data-lifecycle-management/>

“90% of the data in the world today has been created in the last two years. Possibly the greatest challenge of the information age is how to manage data properly. As data is increasingly used for needs assessments, feedback, accountability and monitoring; management of data is a particularly crucial challenge in humanitarian response.”





# Data Portability

Art. 20 GDPR

# Right to data portability

---

1. The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided, where:

# Data Access



# Designing a Data Subject Access Rights Tool

Chrome/SyncSettings.json

Upload zip file

Clear all data

Search for a term...

✕

go

↑

↓

0 out of 0

◆	remote_install: false
	disable_reasons: 0
	installed_by_custodian: false
	update_url: http://clients2.google.com/service/update2/crx
◆	name: Google Calendar
◆	id: ejjicmeblgpmajnghnppodondlgfn

# Overview

Legislation, DSARs, and data exports

Evaluating data exports

Consumer needs and wants

Designing future access tools



Legislation, DSARs, and data exports

Evaluating data exports

Consumer needs and wants

Designing future access tools

# DSARs

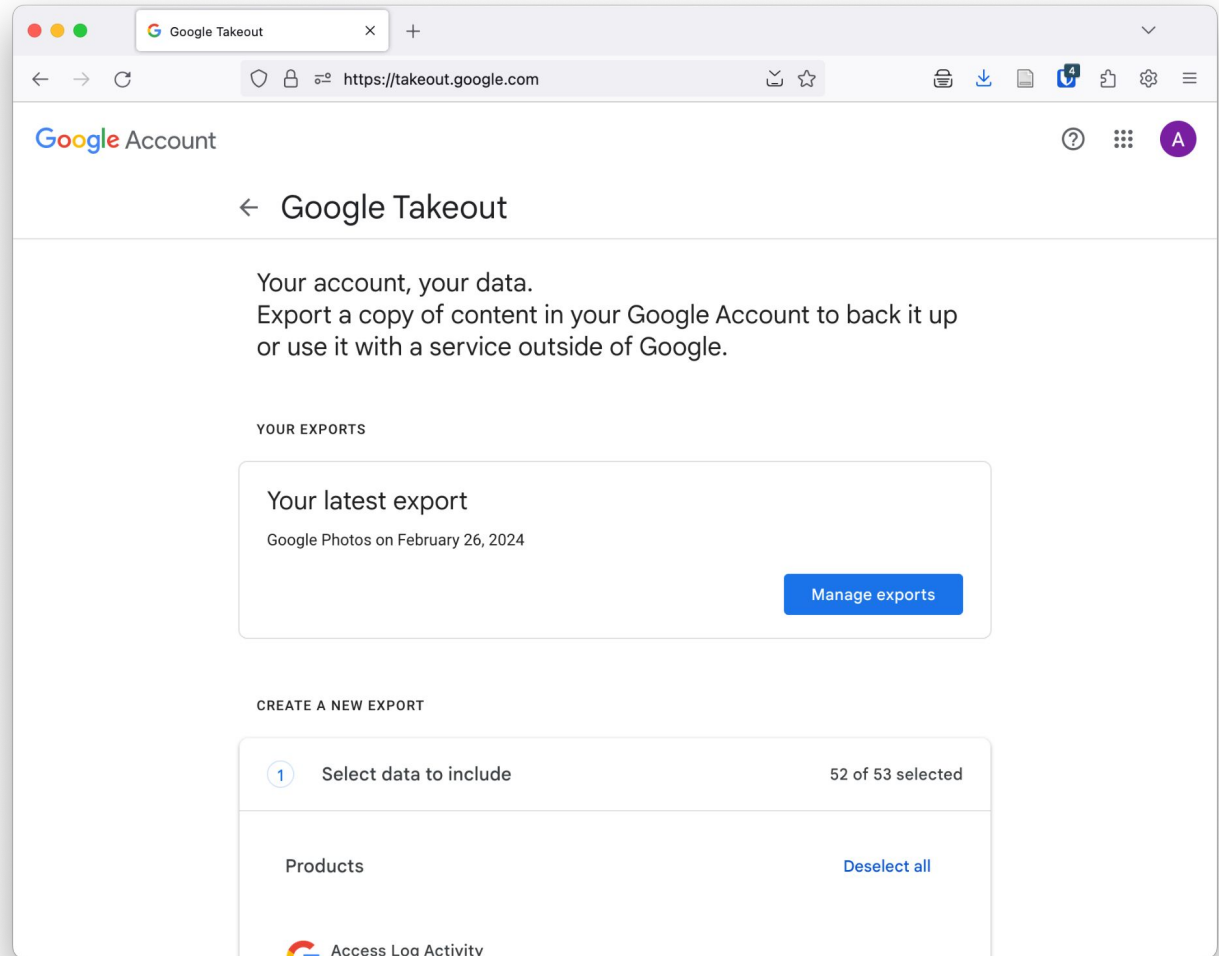
## **Data Subject Access Request**

Legal right for consumers to access their data.

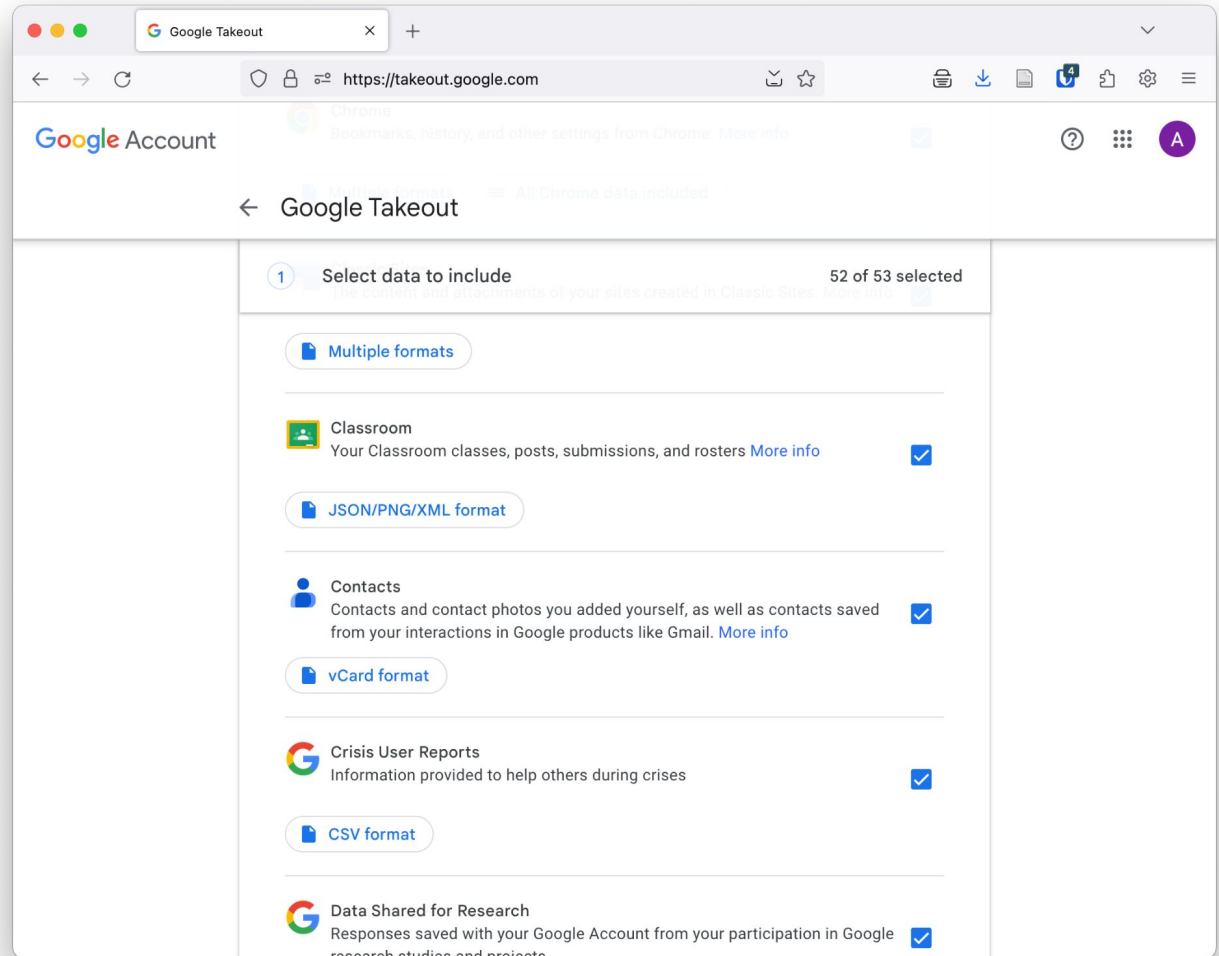
Access → knowledge → enacting preferences



# DSAR in Action



# DSAR in Action



# DSAR in Action

Google Takeout

Destination

Frequency

← Google Takeout

2 Choose file type, frequency & destination

Frequency

☒ Export once

1 export

☐ Export every 2 months for 1 year

6 exports

File type & size

File type:

.zip

Zip files can be opened on almost any computer.

1 GB

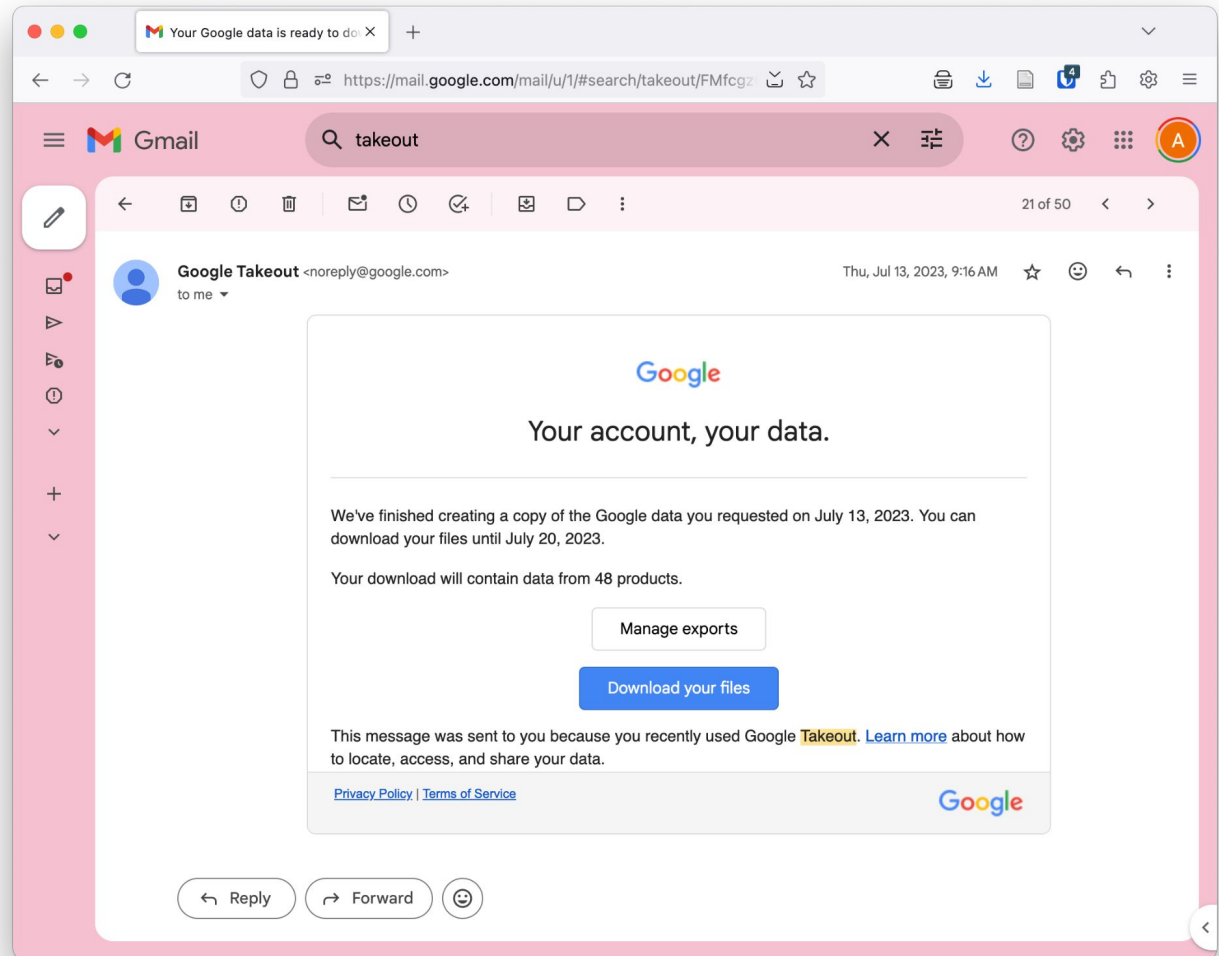
2 GB

4 GB

10 GB

Create export

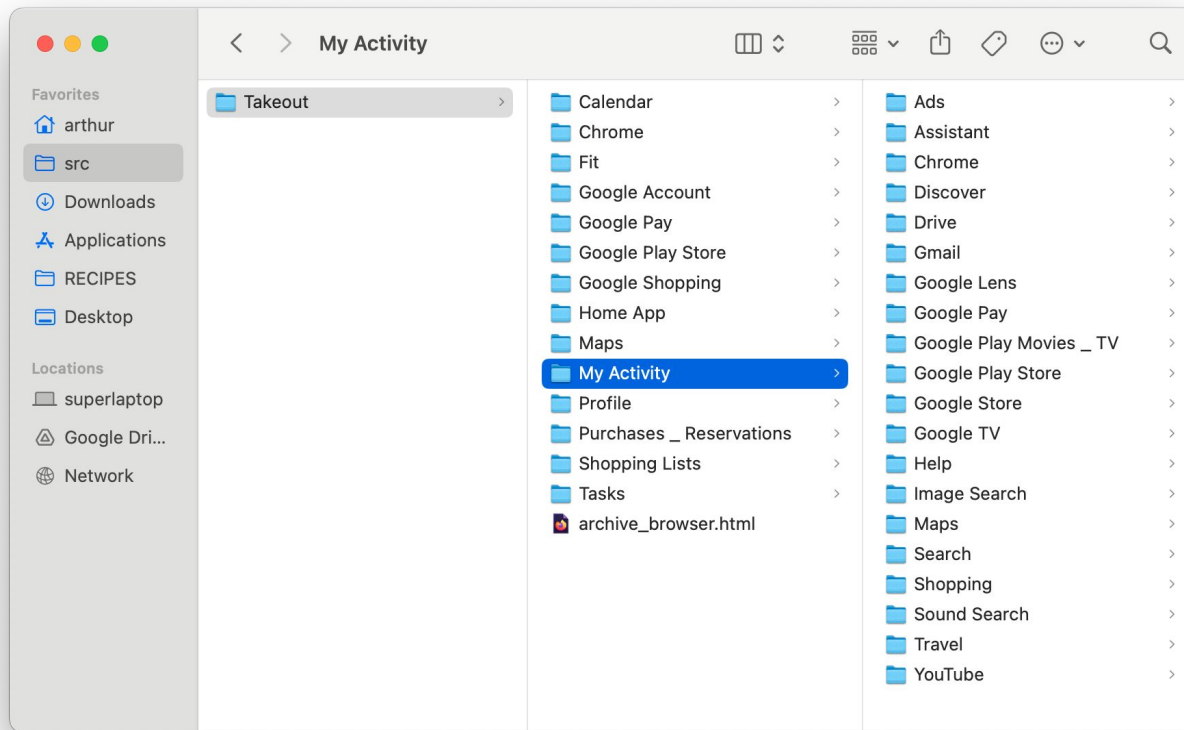
# DSAR in Action



# DSAR in Action

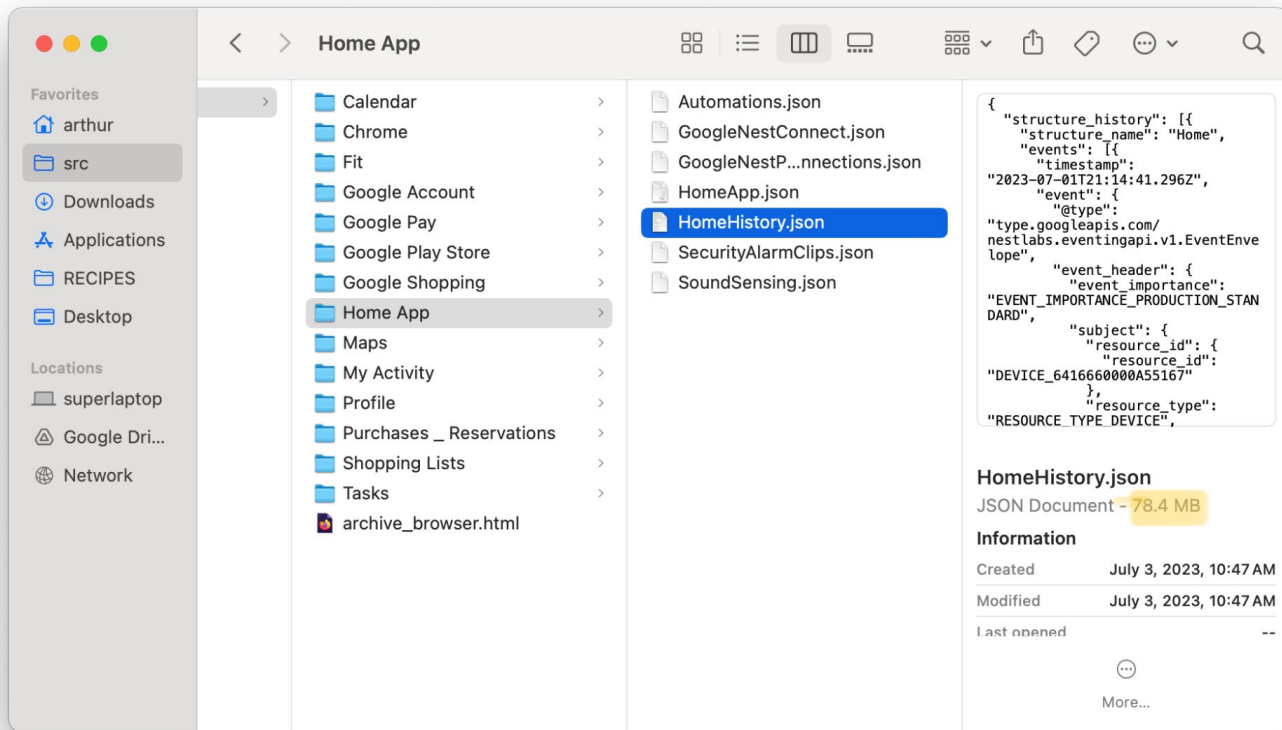
1 hour to 30 days later...

# DSAR in Action





# DSAR in Action



Legislation, DSARs, and data exports

**Evaluating data exports**

Consumer needs and wants

Designing future access tools

# Pursuing Usable and Useful Data Downloads Under GDPR/CCPA Access Rights via Co-Design, SOUPS 2021.

42 participants across 12 focus groups

- (1) Reactions to format and content.
- (2) What's important?
- (3) Could data exports be better?

*Sophie Veys,  
Daniel Serrano,  
Madison Stamos,  
Margot Herman,  
Nathan Reitingger,  
Michelle L. Mazurek,  
and Blase Ur.*

University of Chicago  
and University of  
Maryland

# Focus group protocol

Overview of GDPR/CCPA

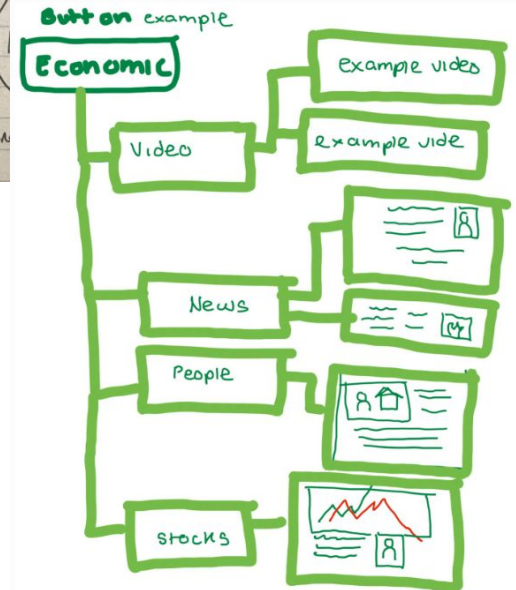
2 min data export exploration

Scavenger hunt for specific records (i.e., timestamp of purchase)

Highlight activity

Data visualization exploration

Sketching the ideal data export



# Participant reactions

- \* missing data that should be there
- \* creeped out by amount of data
- \* useful for finding lost information
- \* fear of misuse by government/police

*“It’s like they didn’t even try. They just kind of dumped it on you.”*

*“Most of the interesting data is stored in these files, that as a non-specialist, I can’t read... We’re effectively illiterate when it comes to reading this additional data they’ve been collecting.”*

# Participant desires and design suggestions

- \* sorting and filtering
- \* aggregation and inferences
- \* rich interactions
- \* enable action

*“What’s interesting to me is how my online behavior is affecting how this company and all the affiliates see me. And in what category, say, they put me or don’t put me... That has a way broader implication... Who is programming these algorithms?... Do they represent a broader part of society or are they all from a very similar group? ”*

Legislation, DSARs, and data exports

Evaluating data exports

**Consumer needs and wants**

Designing future access tools

# Data Subjects' Reactions to Exercising Their Right of Access, USENIX 2024

Interactive data exploration with 33 participants

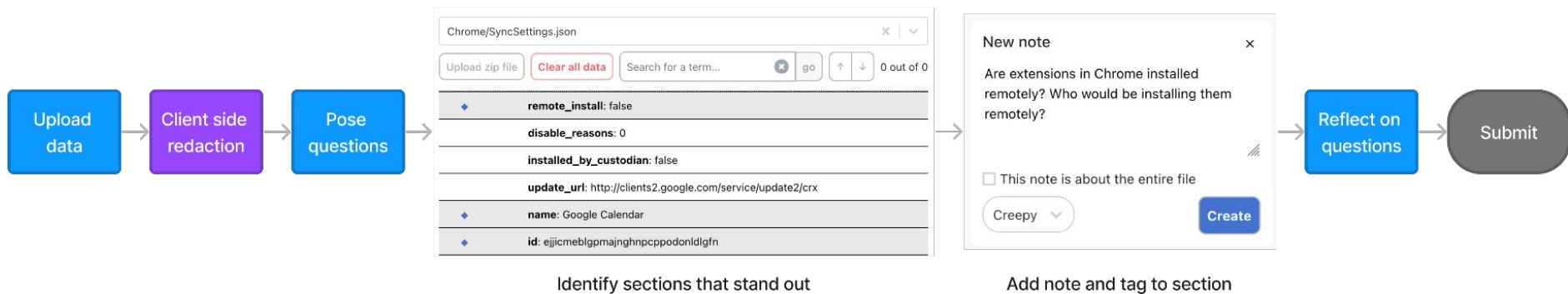
- (1) Data export complexity.
- (2) What do consumers want answered by data exports?
- (3) How effective are data exports?
- (4) How relevant are privacy laws to consumers' needs and wants?

*Arthur Borem,  
Elleen Pan,  
Olufunmilola Obielodan,  
Aurelie Roubinowitz,  
Luca Dovichi,  
Michelle L. Mazurek,  
Blase Ur.*

University of Chicago  
and University of  
Maryland



# Study protocol



# Data annotation tool

File selector displays files in a dropdown for participants to explore

Search bar finds searched text throughout the entire data export file

Notes can be created for specific lines or the whole file

Search text is highlighted throughout the file and in the file selector

Notes are tagged with a reaction label

Existing notes are highlighted and designated with an identifier

The screenshot shows the 'Data Illuminator' interface. At the top, a dark header contains the logo, 'Downloading My Data', and an 'About' link. Below the header is a file selector dropdown showing 'takeout-20230703T1150538Z-001/Takeout/Chrome/BrowserHistory.json'. Below the dropdown are buttons for 'Upload zip file', 'Clear all data', and a search bar containing 'astros' with a 'go' button and navigation arrows. The main content area displays a list of data entries. The first entry has a 'page\_transition' of 'RELOAD' and a 'title' of 'Astros 5-3 Rangers (Jul 2, 2023) Game Recap - ESPN'. The second entry has a 'page\_transition' of 'LINK' and the same title. The third entry has a 'client\_id' of 'i93+5JM3bmF5t/ZAPFVmVA=='. The search term 'astros' is highlighted in green in the search bar and in the file selector dropdown. The sidebar on the right is titled 'Your notes' and contains a 'New note' form with the text 'How did they know I watched this?'. Below the form is a checkbox for 'This note is about the entire file' and a 'Highlight' dropdown menu. A 'Create' button is at the bottom of the form. Below the form is a list of existing notes. The first note is labeled 'B.1' and has the text 'What does this mean?'. It is highlighted with a yellow background and has a 'Confusing' reaction label. The sidebar also has an 'All notes' button and a 'Submit' button.

# Data export structure

Metric	Amazon			Facebook			Google			Spotify			Uber		
# participants	2			9			17			3			7		
# unique keys	749			328			1000			56			99		
# exclusive unique keys	650			149			700			3			0		
Per participant:	min	med	max	min	med	max	min	med	max	min	med	max	min	med	max
# files	27	51.5	76	9	23	56	1	21	53	8	10	10	7	8	10
# unique keys	250	424	598	45	78	188	7	118	545	41	48	55	67	79	99
# directories	5	17	29	4	12	30	2	15	30	1	1	1	4	4	4
Directory depth	-	-	5	-	-	4	-	-	6	-	-	1	-	-	2
Export size (kB)	194	806	1,418	83	976	6,678	5	4,118	38,318	10	260	952	10	1,243	18,966

# Consumers want to know about...

## Platform information (93 questions)

*“Does the company share [its] users’ data with other companies?”*

59 answered  
34 unanswered

*“Why do you gather my data?”*

# Consumers want to know about...

## Specific records (59 questions)

*“Does Uber track my travel locations?”*

15 answered

44 unanswered

*“Are the devices where I’ve logged on stored and detailed?”*

# Consumers want to know about...

## **Their own information (30 questions)**

*“What locations have I been to based on my location history, and can I identify any significant travel patterns or places of interest?”*

7 answered  
23 unanswered

*“How many hours have I spent on Facebook in a certain period of time?”*

# Surprises and interesting finds

**“They have my address and I don’t like it”**  
**33 participants**

- \* unknown records (e.g., number of times each emoji is used)
- \* settings and privacy preferences
- \* lost data

# Barriers to comprehension

**“What the hell is a checkpoint?”**

**30 participants**

- \* key naming (i.e., `used_for_disbursements`)
- \* file naming
- \* missing data
- \* incorrect data



# Self-discovery, memories, and nostalgia

**“I’m surprised to know my Uber waiting times and [idle times] have increased over the years”**

**12 participants**

- \* Forgotten memories and connections
- \* Patterns and behaviors
- \* Positive associations

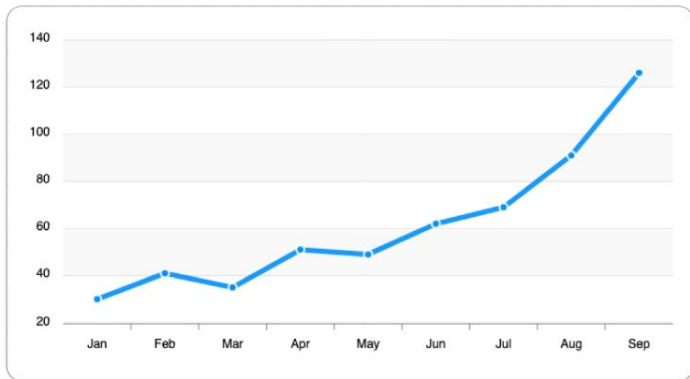
Legislation, DSARs, and data exports

Evaluating data exports

Consumer needs and wants

**A better access tool**

# Design features



```
search_queries: [  
  {  
    index: 0,  
    timestamp: 1679954336,  
    text: "UChicago CS"  
  },  
  {  
    index: 1,  
    timestamp: 1679954329,  
    text: "CS internships"  
  },  
  ...  
  {  
    index: 47201,  
    timestamp: 1679953456,  
    text: "italian coffee"  
  },  
  {  
    index: 47202,  
    timestamp: 1679959996,  
    text: "chicago pizza"  
  },  
]
```

make deletion  
request

You've made over 40  
thousand queries over  
the past 5 years!

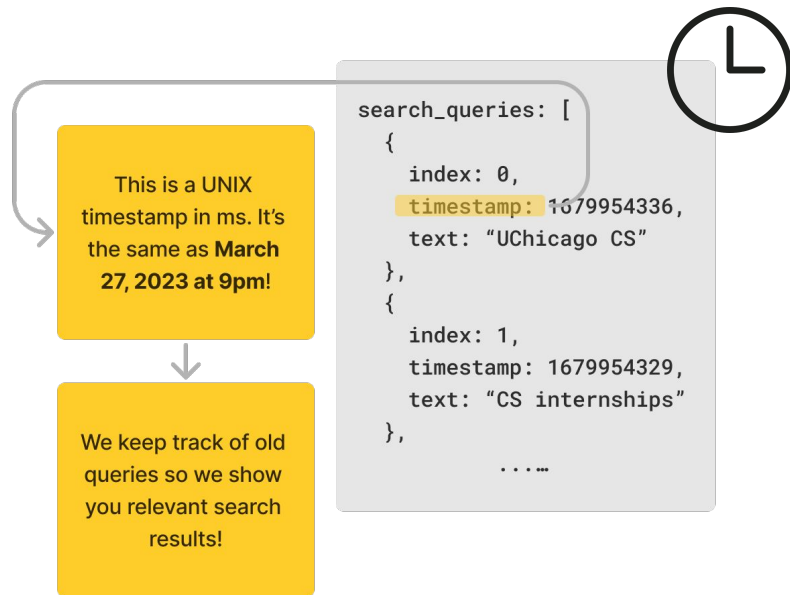
Search through them  
below:

# Policy recommendations

**Reduce delivery time**

**Data summaries and definitions**

**Justified data retention**



# Thanks!



## **Policy recommendations:**

- \* Reduce delivery time
- \* Data summaries and definitions
- \* Justified data retention



## **Developer recommendations:**

- \* Visualization without truncation
- \* Meaningful interactions
- \* Action in-band

# **Right to Erasure**

# GDPR Article 17

- **Right to erasure (formerly known as the ‘right to be forgotten’)**
- The data subject shall have the right to obtain from the controller the **erasure of personal data concerning him or her** without undue delay and the controller shall have the obligation to erase personal data without undue delay where one of the following grounds applies:
  - the **personal data are no longer necessary in relation to the purposes for which they were collected** or otherwise processed;
  - the data subject **withdraws consent** on which the processing is based according to point (a) of Article 6(1), or point (a) of Article 9(2), and where there is no other legal ground for the processing;
  - the **data subject objects to the processing** pursuant to Article 21(1) and there are no overriding legitimate grounds for the processing, or the data subject objects to the processing pursuant to Article 21(2);

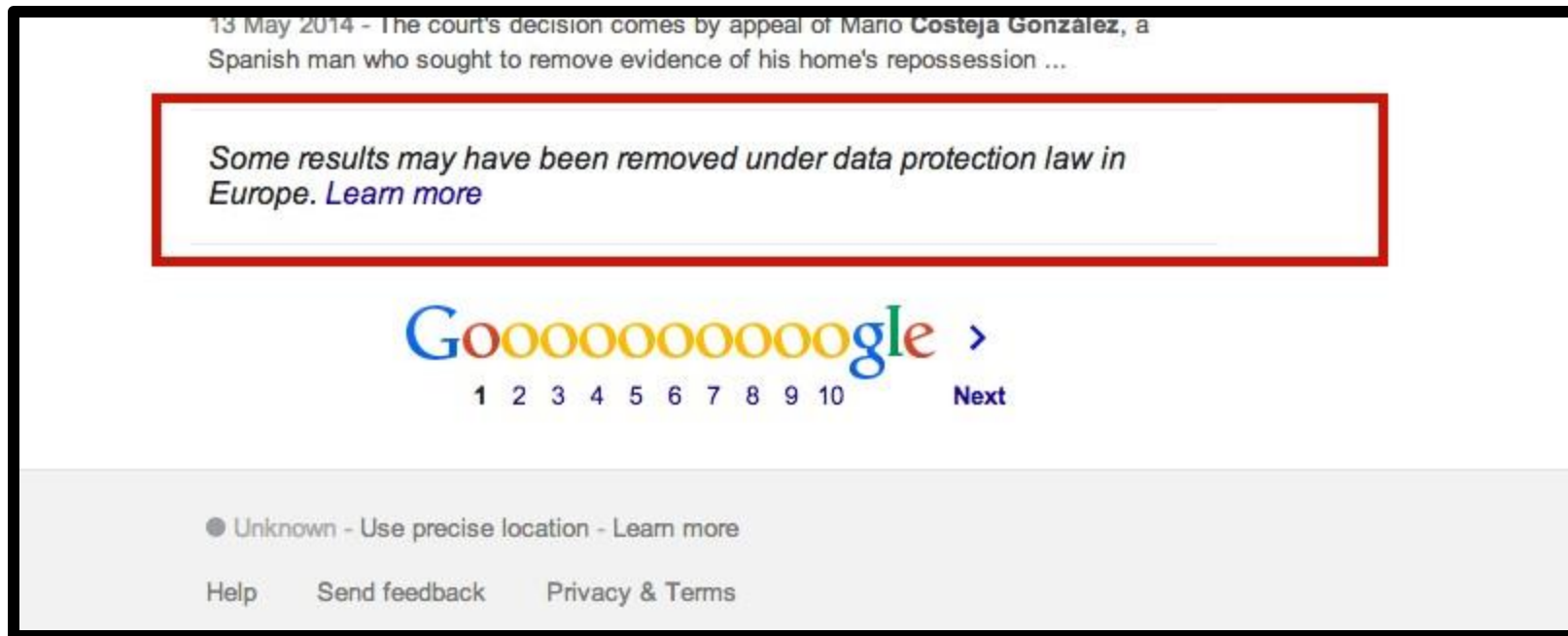
# CCPA (California Civic Code 1798.105)

- (a) A consumer shall have the **right to request that a business delete any personal information about the consumer which the business has collected from the consumer...**
- (c) A business that receives a verifiable consumer request from a consumer to delete the consumer's personal information pursuant to subdivision (a) of this section shall delete the consumer's personal information from its records and direct any service providers to delete the consumer's personal information from their records.
- (d) A business or a service provider shall not be required to comply with a consumer's request to delete the consumer's personal information if it is necessary for the business or service provider to maintain the consumer's personal information in order to:
  - (1) **Complete the transaction** for which the personal information was collected, fulfill the terms of a written warranty or product recall conducted in accordance with federal law, provide a good or service requested by the consumer, or reasonably anticipated within the context of a business' ongoing business relationship with the consumer, or otherwise perform a contract between the business and the consumer.
  - (2) **Detect security incidents, protect against malicious, deceptive, fraudulent, or illegal activity**; or prosecute those responsible for that activity.
  - (3) **Debug to identify and repair errors** that impair existing intended functionality...
  - (6) Engage in public or peer-reviewed scientific, historical, or statistical research in the public interest that adheres to all other applicable ethics and privacy laws, when the business' deletion of the information is likely to render impossible or seriously impair the achievement of such research, if the consumer has provided informed consent.
  - (7) To enable solely internal uses that are **reasonably aligned with the expectations of the consumer** based on the consumer's relationship with the business



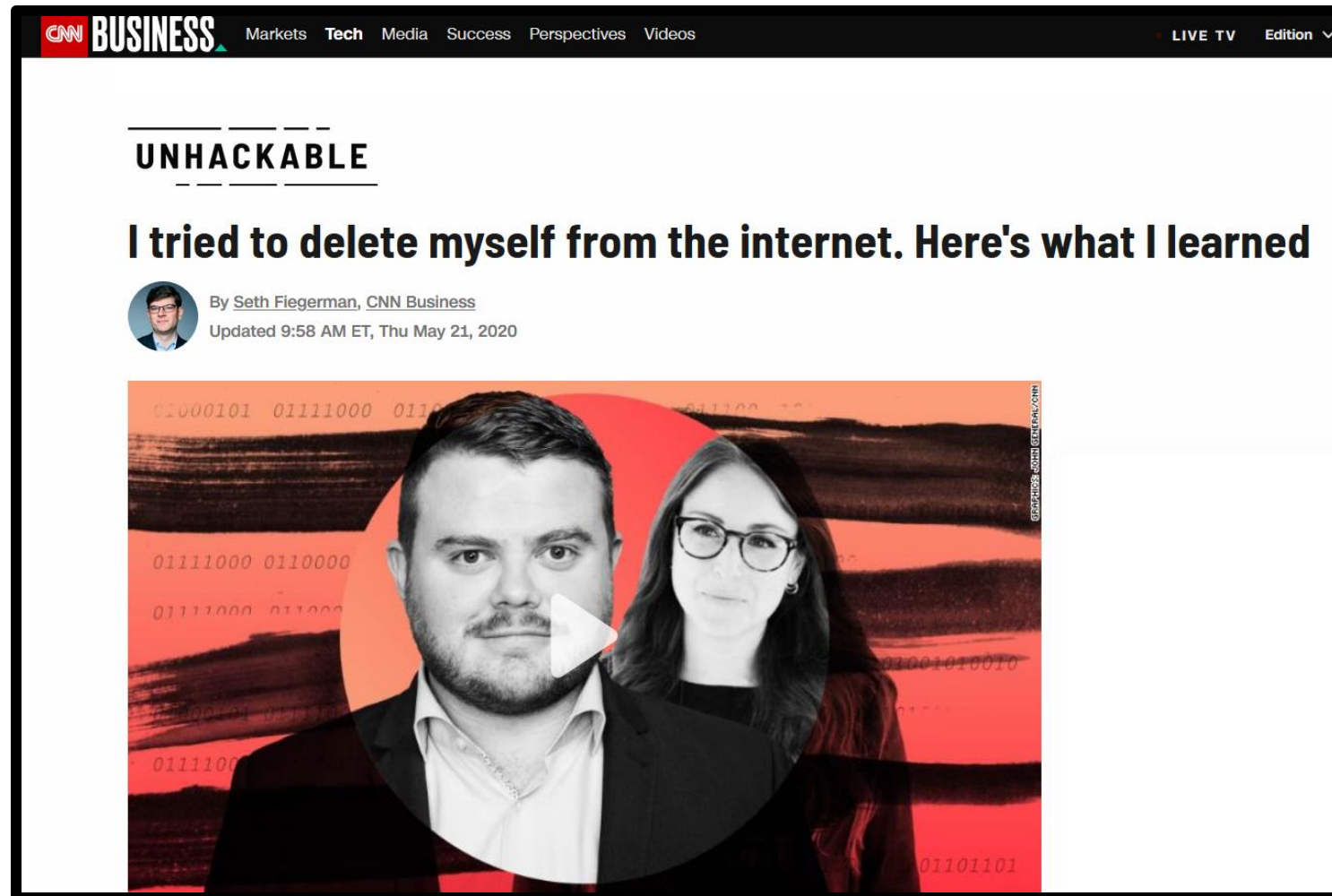
# Right to Erasure

- Bound to jurisdictions
- Not applicable to public figures / issues of public interest



# **The Difficulty of Data Deletion**

# Making Others Delete Your Data is Hard



<https://www.cnn.com/2020/05/21/tech/deleting-personal-data-online/index.html>

# Learning Who Has Your Data is Hard



<https://www.nytimes.com/2019/11/04/business/secret-consumer-score-access.html>

# Backup Copies Are Fun



# Shared Responsibility is Also Fun



<http://www.gdprtoons.com/2017/11/gdpr-addresses-joint-controllers-with.html>

# Challenges for data deletion

- Backup copies and data duplication
- Log files and other subtle records
- Removing the data that should be erased from aggregated data
  - How does this impact an ML model?
- Legal, contractual, or policy considerations for data *retention*
- Provenance tracking
  - How do you store metadata about the subject of particular data? How does this vary based on the data structure / data type?
  - What if the data is sold?
  - What if the entity holding the data is sold?

# Design exercise: data deletion

- You have a policy that all personal data must be deleted after 90 days
  - Data that has been pseudonymized can be kept forever
- You decide that you will also respond within 72 hours to erasure requests

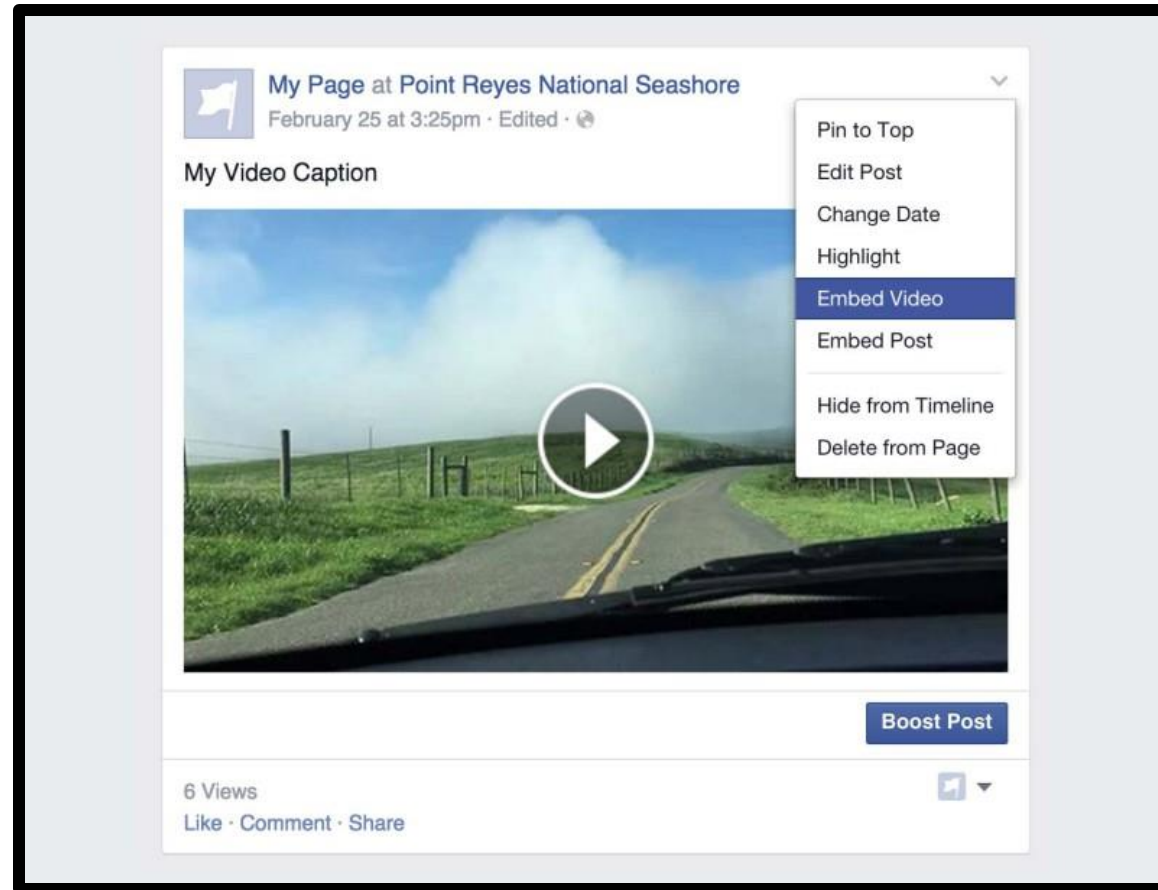


# Who is the Data Subject?



Image from <https://freesvg.org/group-of-people>

# The Nuances of Deletion



# The Nuances of Deletion

## Facebook Launches New Video Embeds & Comment Syncing From Site To Page

Facebook video now can be embedded independent of Facebook posts; conversation threads in the Facebook comment plugin will have the option to be mirrored on Facebook Pages.

Martin Beck on March 25, 2015 at 3:25 pm



# **Data Repurposing**

# Repurposing Google data

- Google Buzz was a social network that replaced Google Wave
- Users automatically opted into Buzz, with weak privacy settings, based on Gmail data
- Automatically add as friends people you talk to frequently on Google Chat (Gchat)



# Repurposing data

- Food delivery services have added restaurants to their site using their public menus (but without permission)

# Banner ads



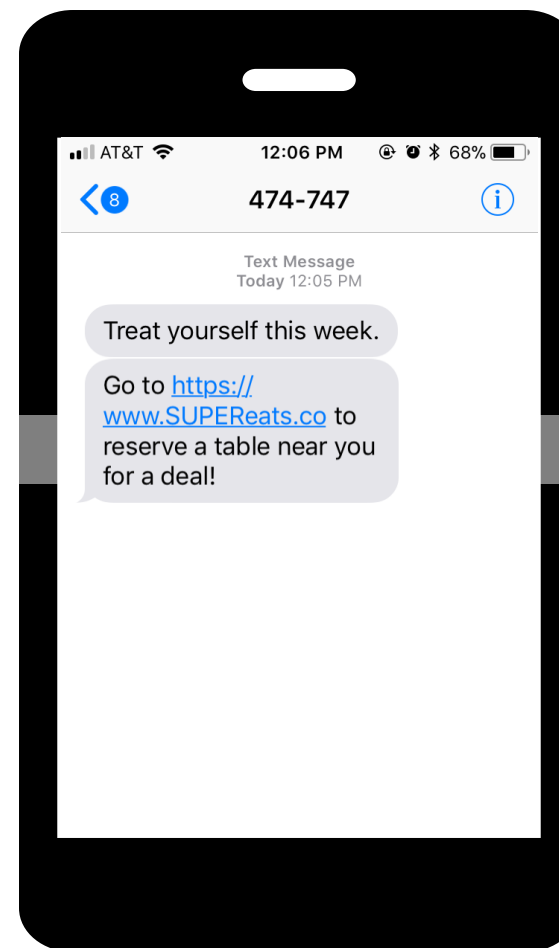
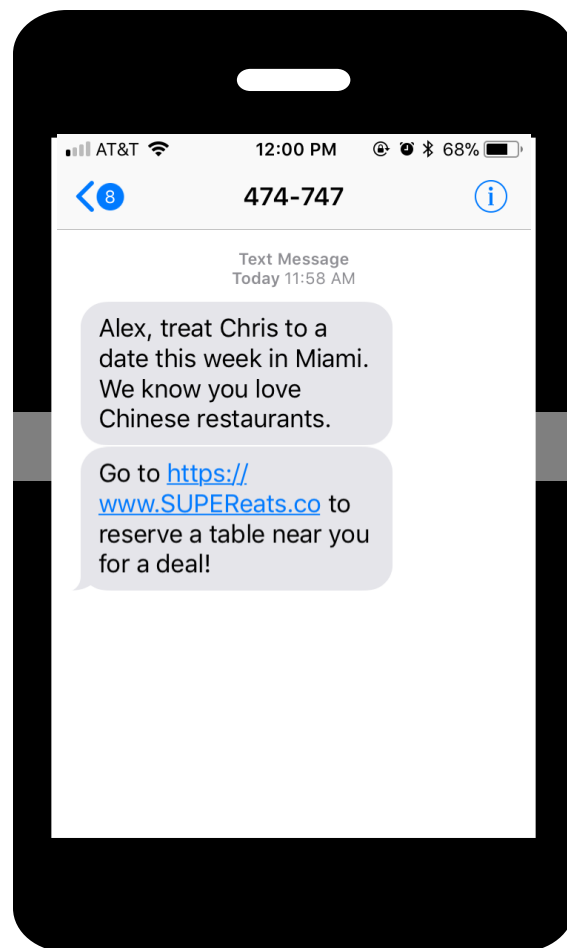
Personalized



Generic

# Robotext ads

Personalized

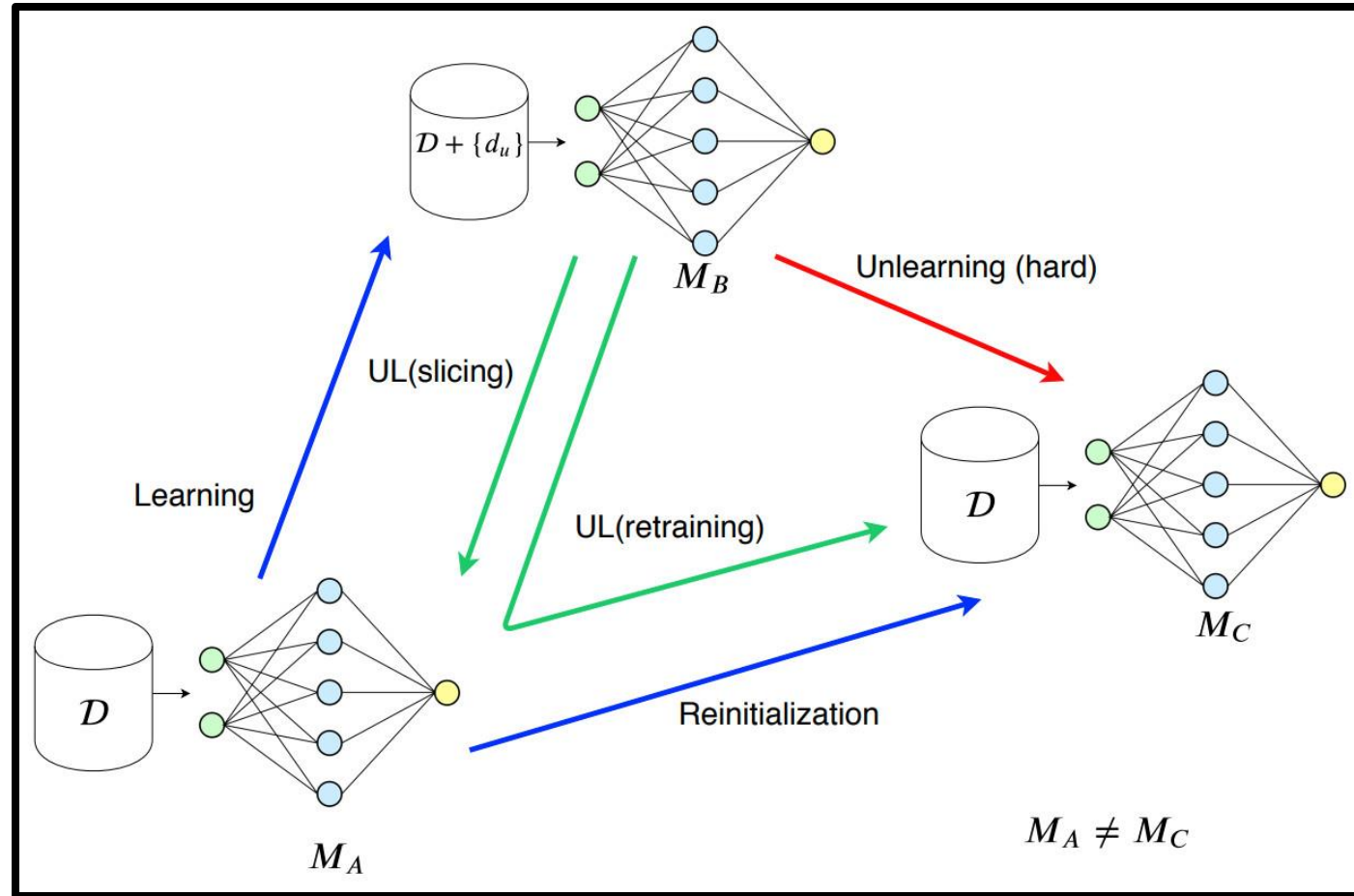


Generic



# Machine Unlearning

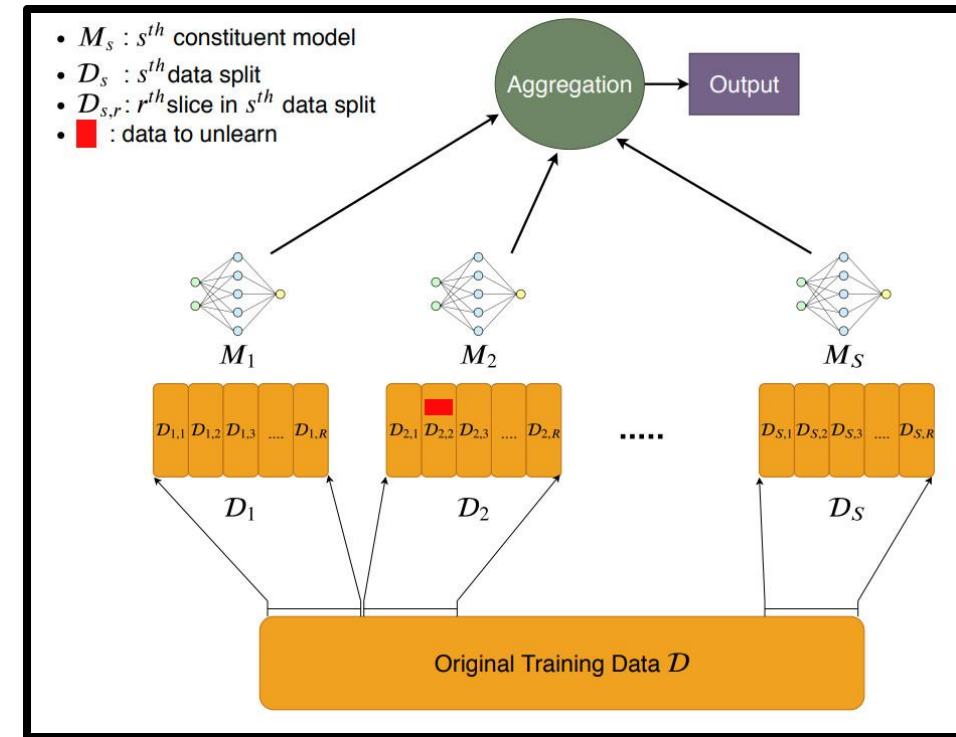
# Machine Unlearning



Images taken from <https://arxiv.org/abs/1912.03817>

The work is summarized on <http://www.cleverhans.io/2020/07/20/unlearning.html>

# Machine Unlearning



- Shard data: Each subject's data is contained in only one shard
  - You only need to retrain 1 of N shards
- Slice data within a shard: Train for Slice 1, Slices 1-2, Slices 1-3, and maintain state
  - Expectation that you can start retraining halfway through the slices

Images taken from <https://arxiv.org/abs/1912.03817>

The work is summarized on <http://www.cleverhans.io/2020/07/20/unlearning.html>

# Genetic Data

# The Ownership of Biological Data

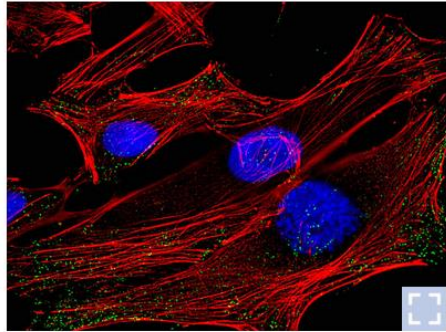
## 'Henrietta Lacks': A Donor's Immortal Legacy

February 2, 2010 · 12:00 PM ET  
Heard on **FRESH AIR**

**37-Minute Listen** [+ PLAYLIST](#) [Download](#) [Share](#) [Menu](#)

In 1951, an African-American woman named Henrietta Lacks was diagnosed with terminal cervical cancer. She was treated at Johns Hopkins University, where a doctor named George Gey snipped cells from her cervix without telling her. Gey discovered that Lacks' cells could not only be kept alive, but would also grow indefinitely.

For the past 60 years Lacks' cells have been cultured and used in experiments ranging from determining the long-term effects of radiation to

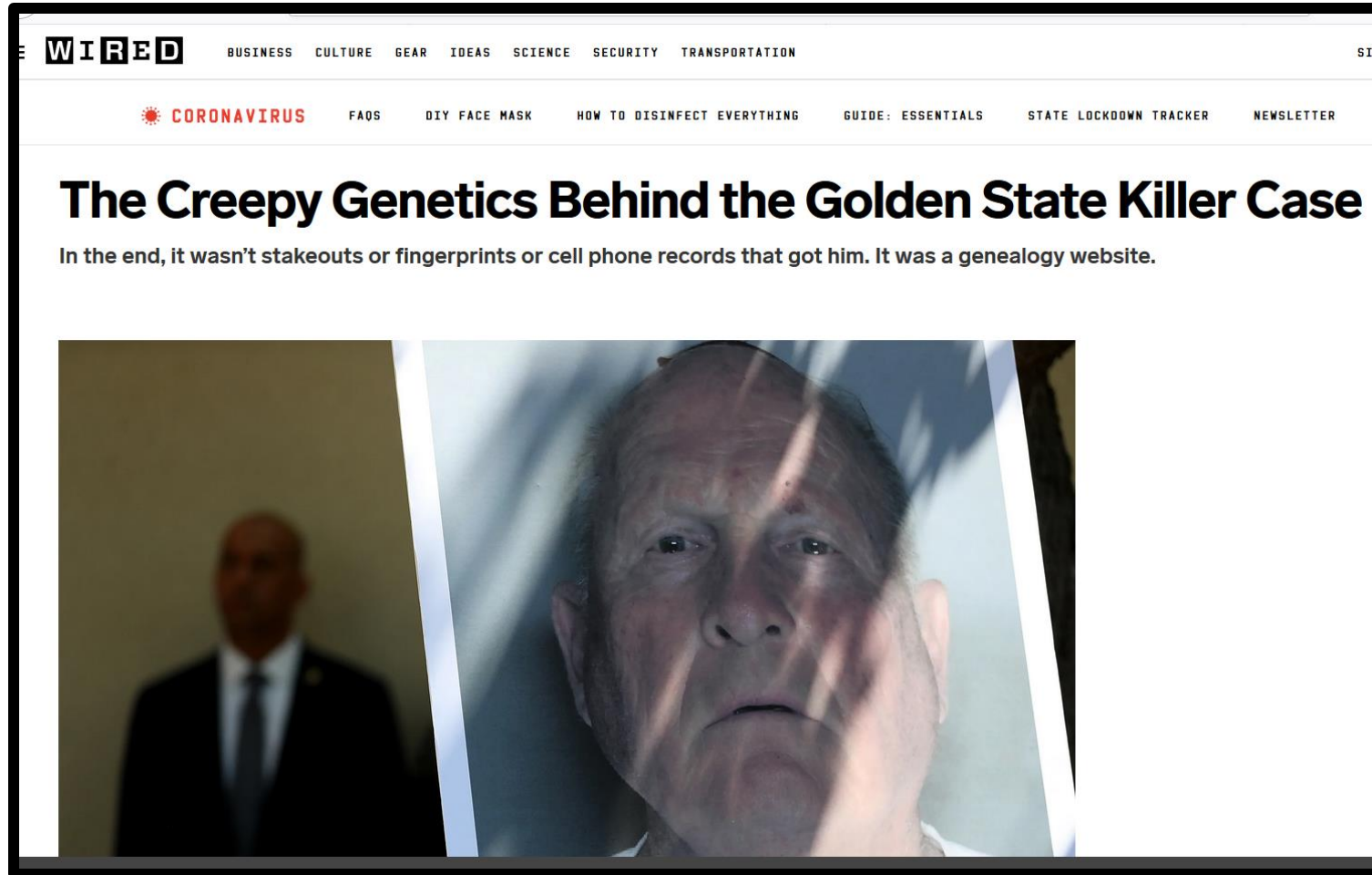


A fluorescence micrograph of HeLa cells, derived from cervical cancer cells taken from Henrietta Lacks and named in her honor

Tomasz Szul/Visuals Unlimited, Inc./Getty Images

<https://www.npr.org/2010/02/02/123232331/henrietta-lacks-a-donors-immortal-legacy>

# Wins (?) for DNA Data



<https://www.wired.com/story/detectives-cracked-the-golden-state-killer-case-using-genetics/>

# Losses (?) for DNA Data



<https://www.theguardian.com/lifeandstyle/2018/sep/18/your-fathers-not-your-father-when-dna-tests-reveal-more-than-you-bargained-for>