Doing the right thing is just good business

Sameer

What is privacy?

- Privacy is a concept in disarray
- Privacy is “protean”,“vague”, “evanescent”, and suffering an embarrassment of meanings

> *“It is apparent that the word “privacy” has proven to be a powerful rhetorical battle cry in a plethora of unrelated contexts. . . . Like the emotive word “freedom,” “privacy” means so many different things to so many different people that it has lost any precise legal connotation that it might once have had.”*

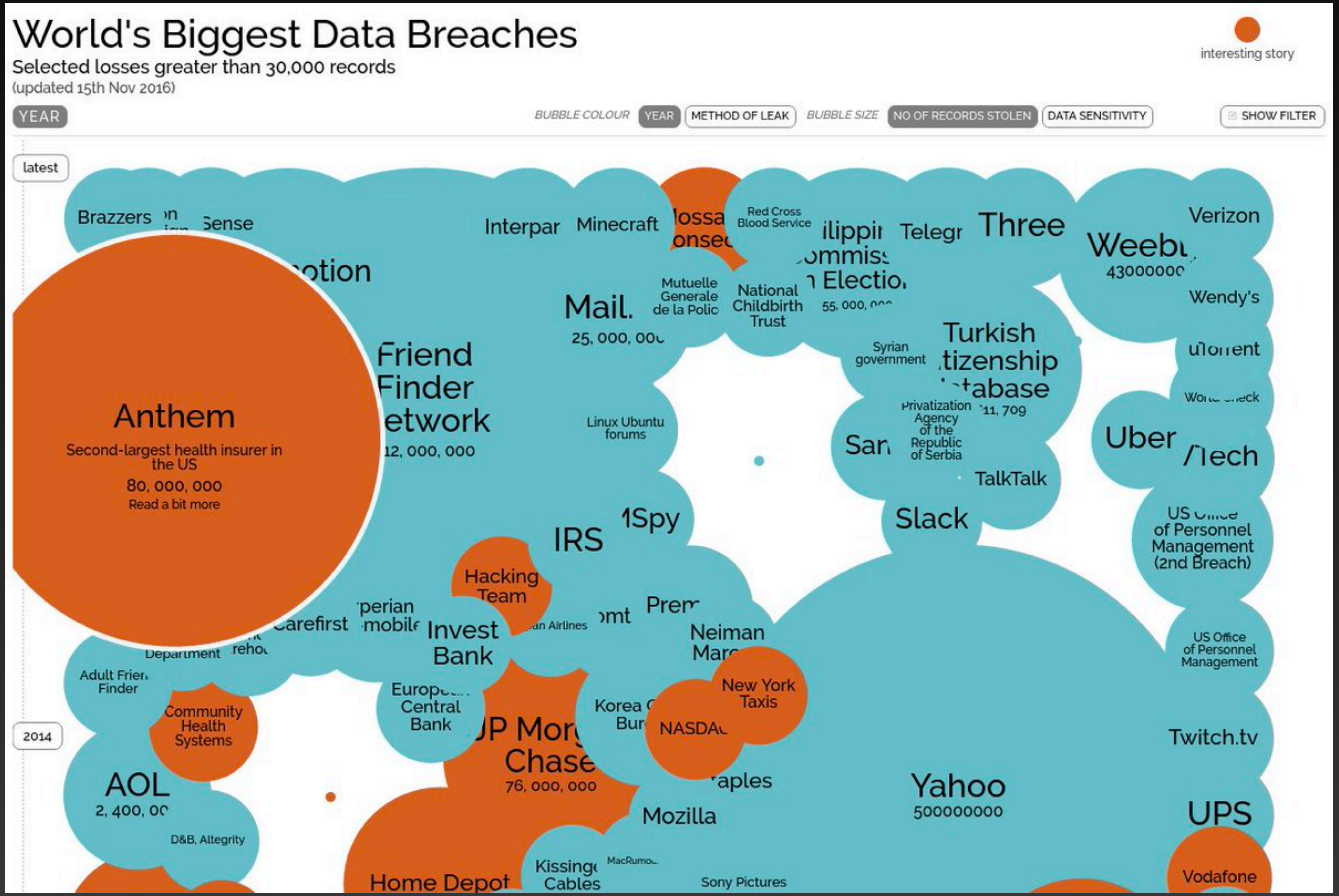
Privacy vs Security aren't they the same?

Data Privacy is for protecting the user freedom and rights

- Security is responsible for protecting data from malicious attacks, exploitation and stealing data for profit
- Privacy is responsible for data governance, collection and use

Data breaches

- An “inside job” breach at data broker Court Ventures was once one of the world’s largest single losses of records at 200 million.
- Malicious hacks at yahoo compromised over 1.5 Billion records
- Over 93% of healthcare companies have been in some kind of lawsuit

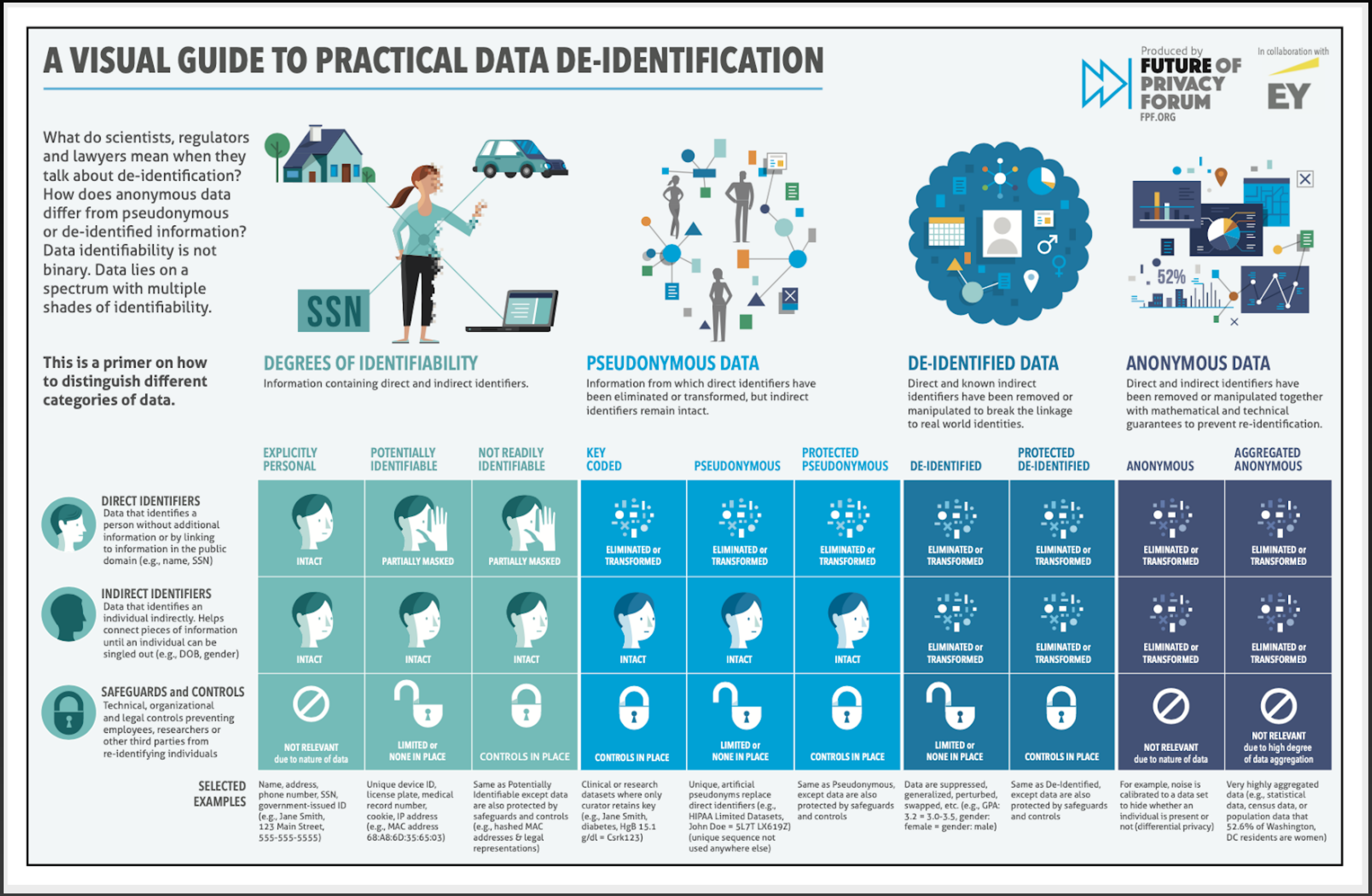


World's Biggest Data Breaches

Selected losses greater than 30,000 records

(updated 15th Nov 2016)

A visual of data privacy protection



A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

Produced by FUTURE OF PRIVACY FORUM FPF.ORG

In collaboration with EY

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.

DEGREES OF IDENTIFIABILITY

Information containing direct and indirect identifiers.

PSEUDONYMOUS DATA

Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

DE-IDENTIFIED DATA

Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

ANONYMOUS DATA

Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

DIRECT IDENTIFIERS Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)

INDIRECT IDENTIFIERS Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)

SAFEGUARDS and CONTROLS Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals

	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
DIRECT IDENTIFIERS	INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INDIRECT IDENTIFIERS	INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
SAFEGUARDS and CONTROLS	NOT RELEVANT due to nature of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to nature of data	NOT RELEVANT due to high degree of data aggregation
SELECTED EXAMPLES	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-Identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

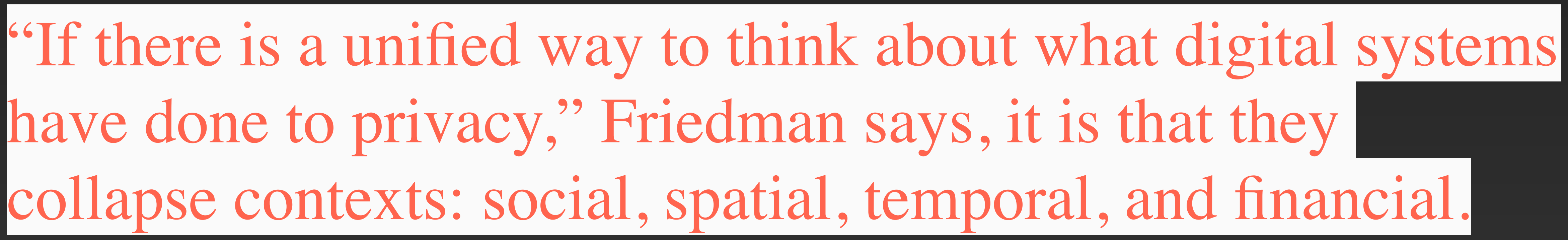
Its a smaller world

- Technology has forced people rethink privacy
- Private, public and really really public

“We think because we have a word for privacy that it is something we can put our arms around. But it’s not.” - Hal Abelson

End of anonymity

- Because personal information is everywhere it leaks
- Botnets-network of machines - are able to find vulnerable information using brute force



"If there is a unified way to think about what digital systems have done to privacy," Friedman says, it is that they collapse contexts: social, spatial, temporal, and financial.

Legal Patch work

- Data breach notifications have increased significantly
- More companies are getting fined
- Is this helpful in protecting privacy? Is legal privacy the right way to do it?

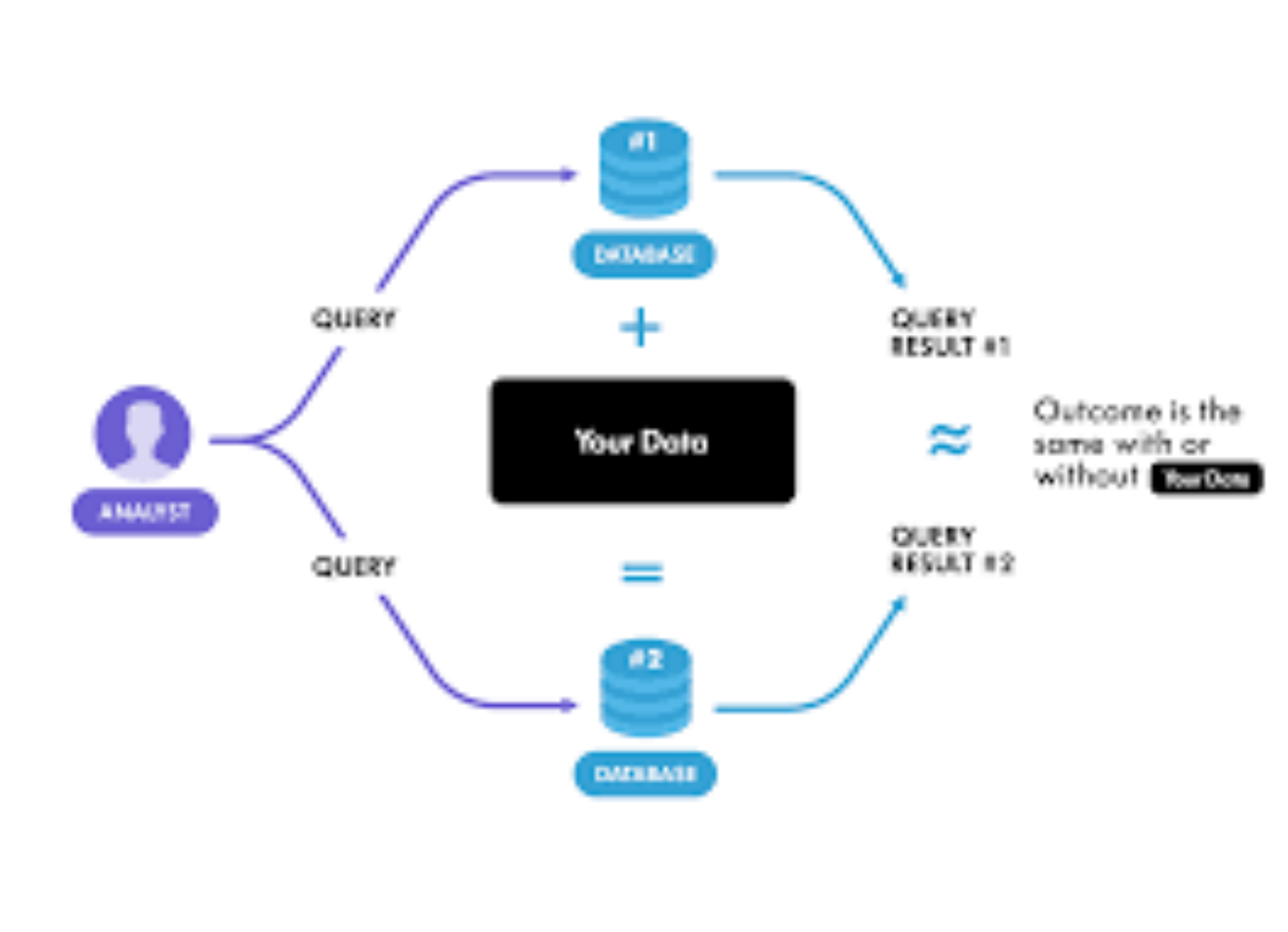
“We need Transparency Around PII and Data Collection”

“Legal policies are letting companies Ignore the General Sentiment”

“Keep the Door Ajar to Let in AI-as-a-Service in 2020”

Differential Privacy

- Strongest privacy guarantees
- The guarantee of a differentially private algorithm is that its behavior hardly changes when a single individual joins or leaves the dataset
- Gives a formal guarantee that individual information from data is not leaked
- Defines privacy as a quantifiable measure


#1
DATABASE
QUERY
+
QUERY RESULT #1
ANALYST
Your Data
≈
Outcome is the same with or without Your Data
QUERY
=
QUERY RESULT #2
#2
DATABASE