PRIVACY & BIG DATA: CAN THEY GET ALONG?

Mike Shapiro
Chief Privacy Officer

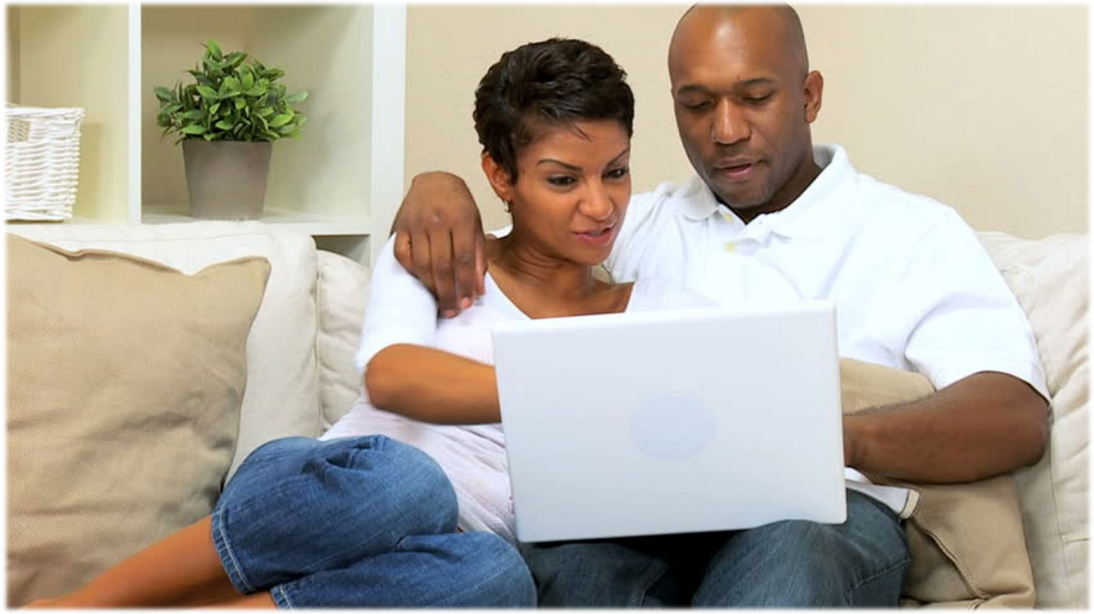
THE EVOLVING DATA LANDSCAPE BRINGS PRIVACY FRONT AND CENTER

"Between the dawn of civilization and 2003, we only created five exabytes of data; now we're creating that amount every two days. By 2020, that figure is predicted to sit at 53 zettabytes (53 trillion gigabytes, the equivalent of about half a billion HD movie downloads) – an increase of 50 times."

Hal Varian, Chief Economist, Google

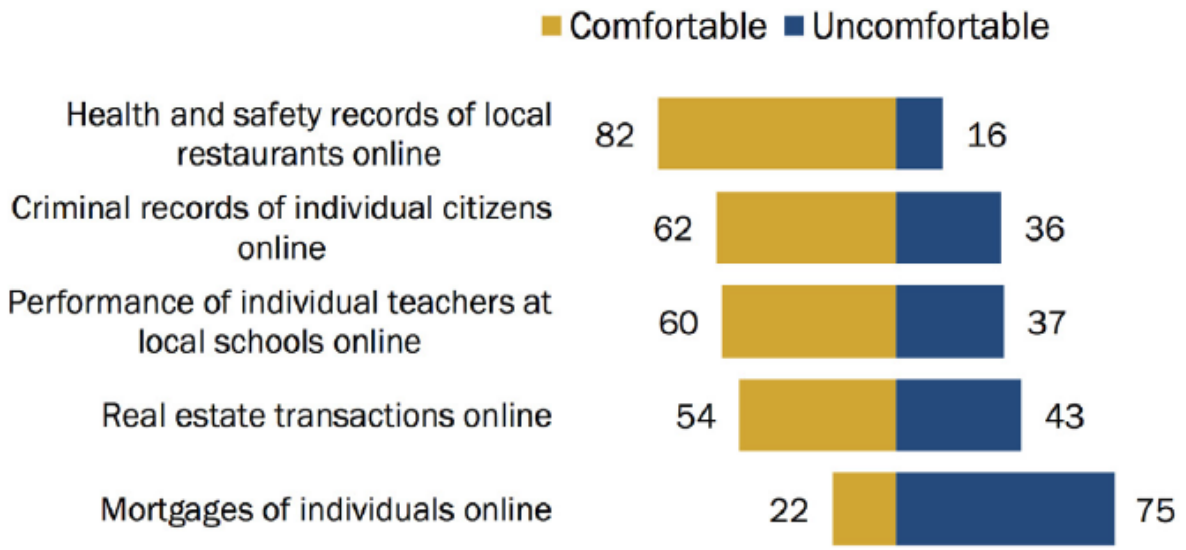
- **Data = Value** and with a treasure trove of data available, privacy breaches are becoming more prevalent and far-reaching:
  - **Facebook / Cambridge Analytica**: 87 million profiles transferred and analyzed for behavioral patterns and targeted for political persuasion
  - **Equifax**: 148 million records exposed containing personally identifiable information (PII), to include credit data, driver's license numbers, social security numbers (SSN), dates of birth, phone numbers, and email addresses
  - **Office of Personnel Management (OPM)**: 25 million compromised files on current and former federal government employees included personal information, background checks, fingerprints, and adjudications for security clearances. OPM's director and Chief Information Officer *resigned*.

PUBLIC SENTIMENT



People are Generally Comfortable with Local Government Data Sharing — Until it Hits Close to Home

% of adults who are comfortable/uncomfortable with local government data sharing about these issues



Source: Online survey of 3,212 adults in Pew Research's American Trends Panel, Nov. 17-Dec. 15, 2014.

PEW RESEARCH CENTER

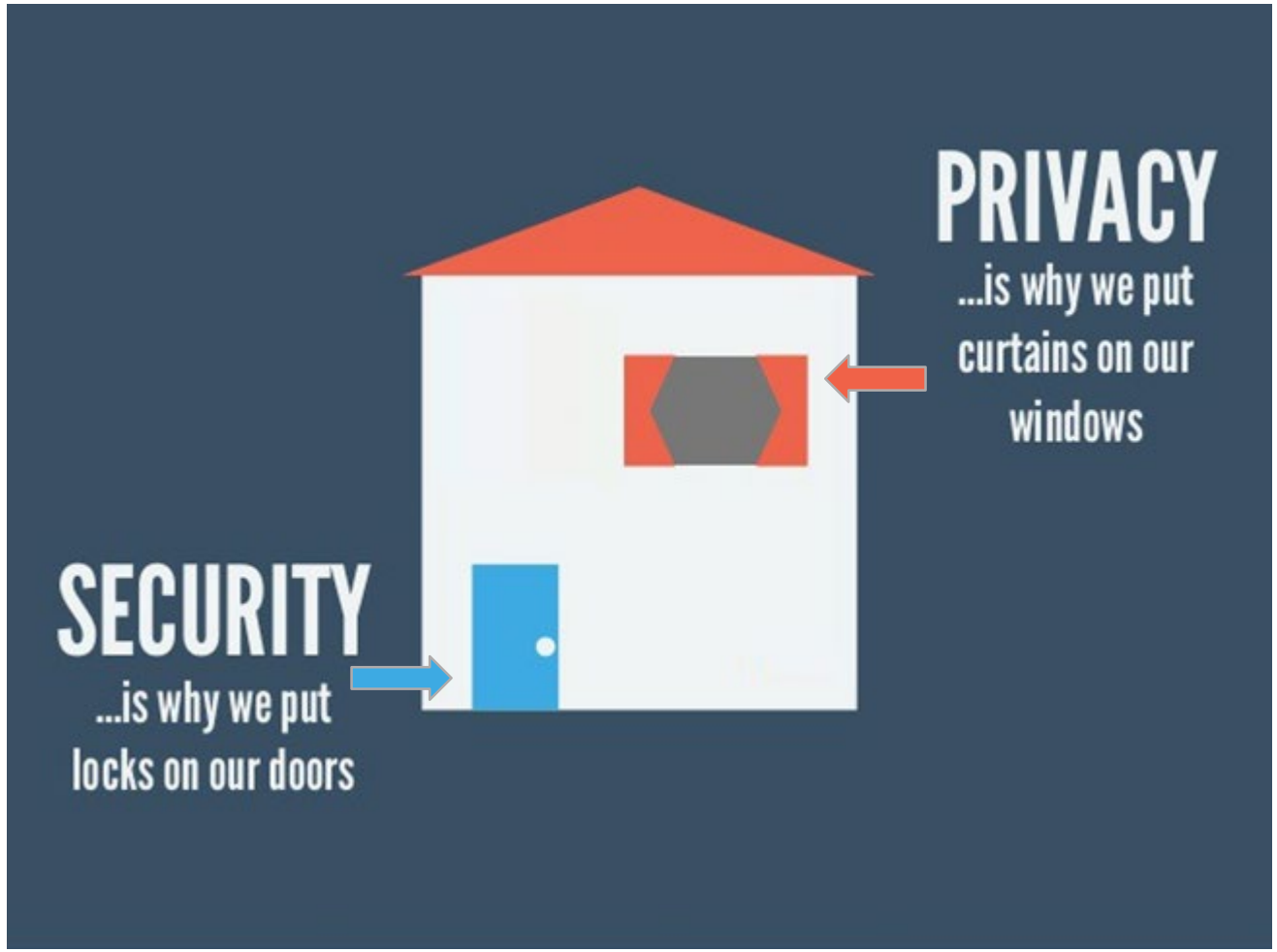
Pew Research Center. "Americans' Views on Open Government Data." (2015) http://www.pewinternet.org/2015/04/21/open-government-data/.

SECURITY AND PRIVACY: WHAT'S THE DIFFERENCE

While security and privacy do share common ground, you can think of it this way:

Security
protects systems

Privacy
protects information



EXPANDING DEFINITION OF PERSONAL INFORMATION

SB 1386: California Data Breach Notification Security Act

"Personal information" means an individual's first name or first initial and last name in combination with any one or more of the following data elements, when either the name or the data elements are not encrypted: (1) Social security number. (2) Driver's license number or California Identification Card number. (3) Account number, credit or debit card number, in combination with any required security code, access code, or password that would permit access to an individual's financial account.

California Consumer Privacy Act of 2018

- Identifiers (e.g., name address, email, IP, SSN, DL)
- Commercial data, purchases
- Biometrics
- Online activity
- Geolocation
- Inferences drawn from info (to develop profiles)
- Employment data
- Education
- Audio, visual, other sensory

COLLAGE OF PRIVACY AND DATA PROTECTION LAWS

FEDERAL | STATE | COUNTY

- Electronic Communications Privacy Act
- Census Confidentiality Statute
- Health Insurance Portability and Accountability Act (HIPAA)
- Children's Online Privacy Protection Act (COPPA)

- SB 24 Data Breach Notification Requirements
- Cal. Welfare & Institutions Code § 10850
- Cal. Consumer Privacy Act of 2018

- Information Practices and Individual Privacy (Div. A16)
- Surveillance–Technology and Community–Safety (Div. A40)

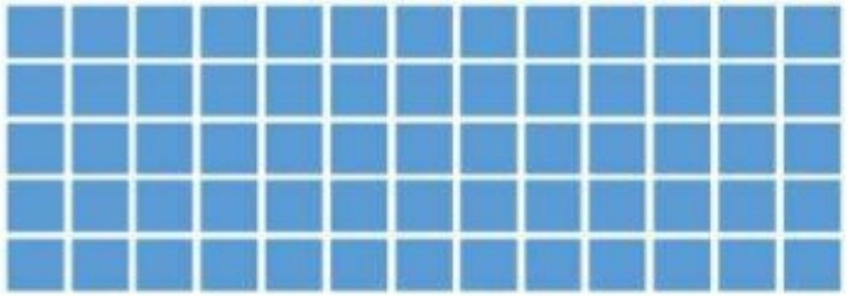
The benefits of data sharing can offer organizations the ability to identify gaps in service, recognize trends, and target resources to improve operations and help people

Big data services are now able to combine, overlay, and analyze data from multiple sources more efficiently and economically than ever before

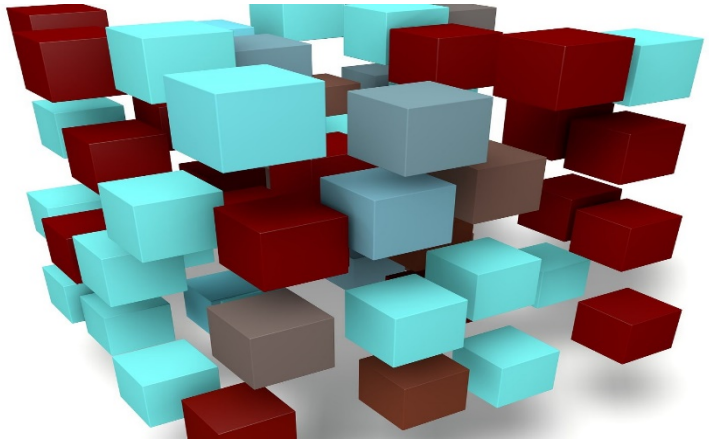
While many organizations see the upside in sharing information to address current and future needs, they may not be considering privacy, compliance, and ethical implications

Data that was once isolated in a single database or department may have had limited value and limited privacy risks, but with data commingling from other sources, privacy risks may be elevated

WHILE THE BENEFITS OF DATA SHARING ARE USUALLY WHERE ORGANIZATIONS FOCUS THEIR ATTENTION, PRIVACY AND COMPLIANCE ISSUES MUST ALSO BE CONSIDERED



Structured Data



Unstructured Data

WHAT IS THE DIFFERENCE BETWEEN STRUCTURED AND UNSTRUCTURED DATA?

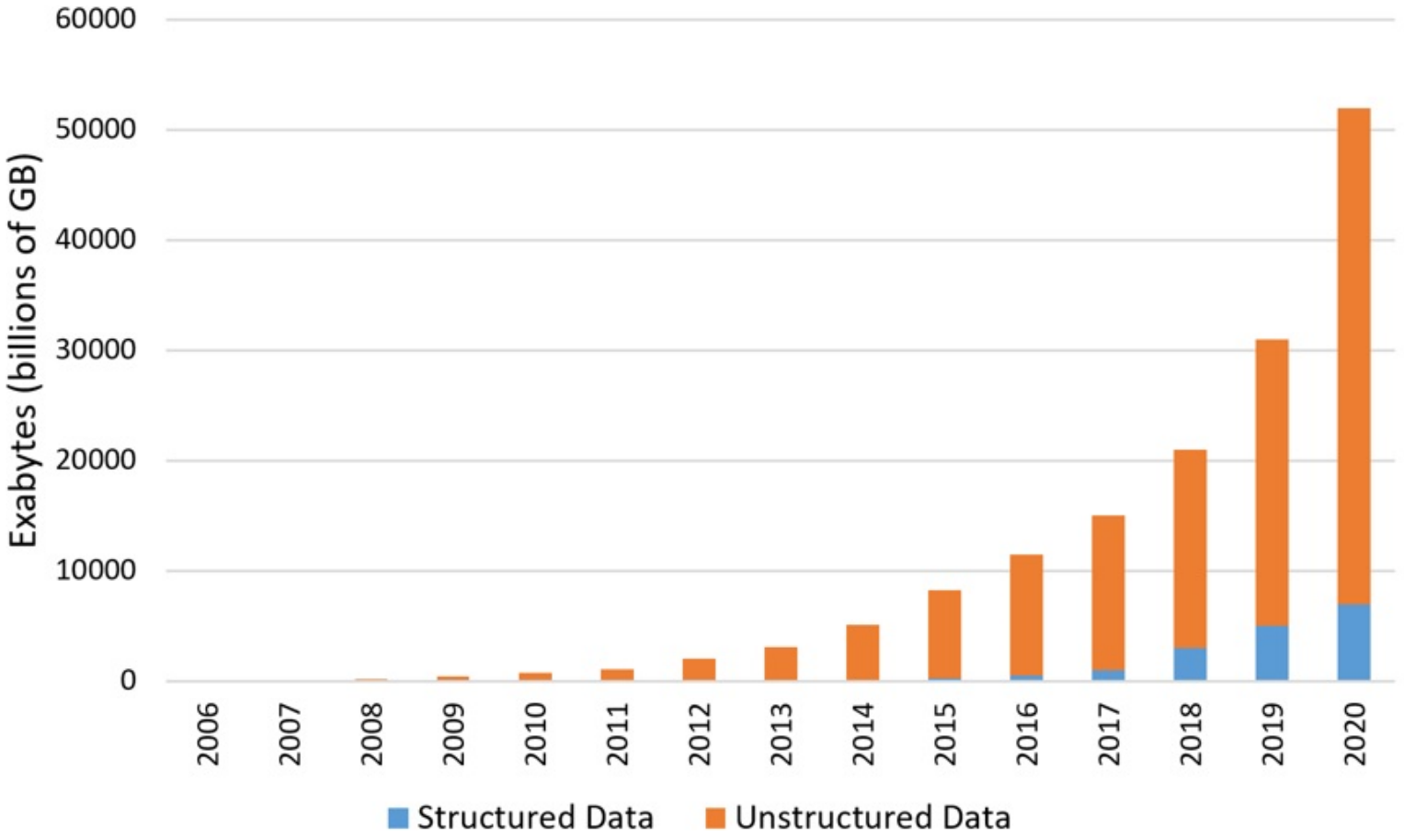
Structured data is highly organized and can be made up of tables with rows and columns that define their meaning.

- Examples are Excel spreadsheets and relational databases.

Unstructured data is everything else.

- Examples include the following:
  - Email messages, instant messages, text messages
  - Text files, including Word documents, PDFs, written documents, audio and video transcripts
  - PowerPoints and SlideShare presentations
  - Audio files of music, voicemails, customer recordings
  - Video files that include movies, personal videos, YouTube uploads
  - Images of pictures, illustrations, memes

UNSTRUCTURED DATA IS EXPLODING AND PRESENTS UNIQUE PRIVACY CHALLENGES



Electronic Engineering Times, https://www.eetimes.com/author.asp?section_id=36&doc_id=1330462

THE GROWTH OF STRUCTURED VERSUS UNSTRUCTURED DATA OVER THE PAST DECADE SHOWS THAT UNSTRUCTURED DATA ACCOUNTS FOR MORE THAN 90% OF ALL DATA, WHICH SIGNIFICANTLY ELEVATES PRIVACY CONCERNS

- Lack of data awareness (didn't know PII was included)
- Data Loss Prevention (DLP) tools can only do so much (lexicons have limitations)
- Unintended profile building of individuals (data commingling can potentially re-identify people or build out profiles)

UNINTENDED CONSEQUENCES MAY OCCUR WHEN PRIVACY CONSIDERATIONS ARE NOT INTEGRATED

- While we want to provide data in an open platform to offer value, it is essential to make sure that we understand the business reasons for sharing data and do not place our constituents, employees, or the County at undue risk
- As an example, while unstructured data (e.g., freeform text fields, open comments boxes, images) may offer valuable insight into people's mindsets, issues, or complaints; oftentimes individuals not only include general data, but may also provide highly sensitive or personal information
- Working with privacy, legal, and security teams (as needed) will help to recognize those risks in an open data platform, determine if adequate protections/restrictions can be put in place, and then decide whether or not any residual risks are acceptable when compared to the benefits of sharing data

In 2012, the Philadelphia's Department of Licenses & Inspections published gun permit appeals as part of its open data initiative. These permits included freeform text fields in which applicants explained why they needed the permit, and where some people wrote that they carry large sums of cash at night. As a consequence for publishing this information, the City was ultimately charged $1.4 million as part of a class-action lawsuit. One of the lawyers behind the suit stated that the information released "was a road map for criminals."

Objective	Potential Privacy and Compliance Issues
Bring together criminal justice information and social services information to identify resources to help released prisoners	• Criminal justice information and social services information may have legal requirements for confidentiality • Are employees in one department trained in the confidentiality requirements of another?
Gather addresses from multiple sources along with geo-location imagery to update the census master address file to improve congressional representation and funding	• Public concerns about reporting information to federal law enforcement (e.g., DHS/ICE) • Concerns about sharing information within the County (e.g., zoning, housing inspector, Assessor, Sheriff)
Provide schools with information from juvenile services to address children's needs in the classroom and at home	• Teachers may respond to children differently knowing their backgrounds • Are teachers and administrators trained on the confidentiality requirements of juvenile criminal information?

ORGANIZATIONS NATURALLY LOOK TO THE VALUE THAT DATA SHARING CAN PROVIDE, BUT WHAT ARE THE RISKS?

These examples indicate the good intentions and potential benefits of data sharing. However, what are the privacy and compliance issues that may arise? What are the possible unintended consequences?

PRIVACY AND COMPLIANCE CHECKPOINTS CAN HELP MAKE SURE THAT DATA IS SHARED RESPONSIBLY

- An example of responsible data sharing can be illustrated between the Public Justice Institute (PJI) and Pretrial Services (PTS)
- PJI requested PTS data, along with other data, to support research initiatives that may help identify trends, determine if marginalized communities are being treated equitably, and understand where resources and assistance may be best applied
- Working with County Counsel, a Memorandum of Understanding (MOU) was put in place explaining confidentiality requirements and expectations
- In addition, to reduce the likelihood that individuals could not be singled out, the data set was de-identified
- Names and identifying numbers, for instance, were removed and location data was modified to regional zones
- This method helps to support analysis activities while reducing re-identification potential on the public portal
- As an extra precaution, data will be reviewed by PTS and other County stakeholders (e.g., Counsel, Privacy), as necessary, prior to publishing

DATA IN CONTEXT: HOW PRIVACY RISK CAN CHANGE DEPENDING ON CONTEXT

Do the following scenarios increase privacy risk?

I would like to share a dataset containing employee names, employee ID number, and performance evals. Is that OK?

Who's it going to? Another co-worker assigned to the project with authorization. ✓

Anyone else? Wellll, we may ask for some help from one of our new contractors. ?

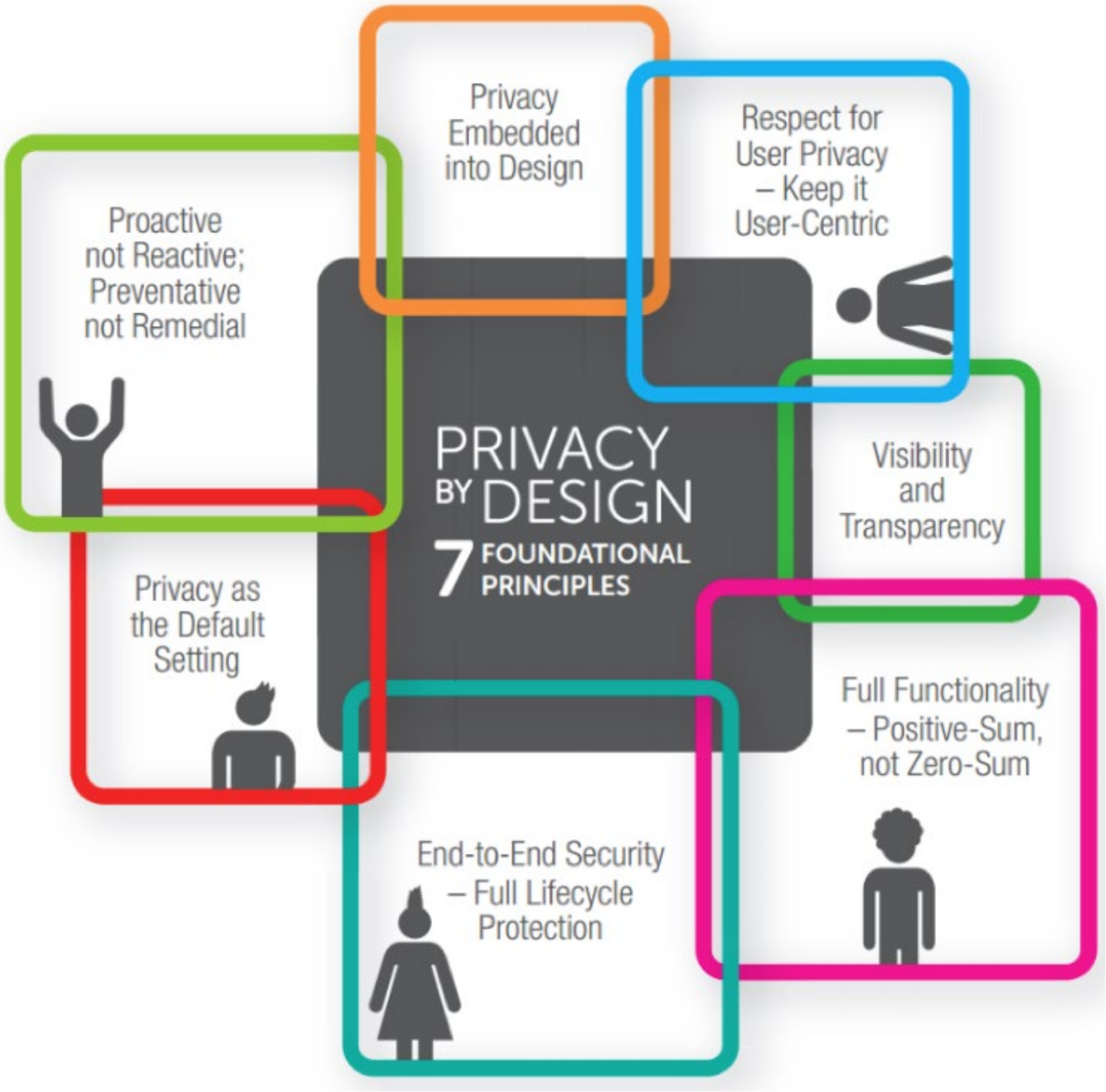
I would like to share a dataset containing client files with names, addresses, and phone numbers. Is that OK?

What's the purpose? Just to provide basic contact information for our partners to help people most in need. ✓

What's the context of the data? The title of the dataset is "HIV Positive Subjects." ?

Has consent been received, etc… What do you mean? ✗

"Before I write my name on the board, I'll need to know how you're planning to use that data."

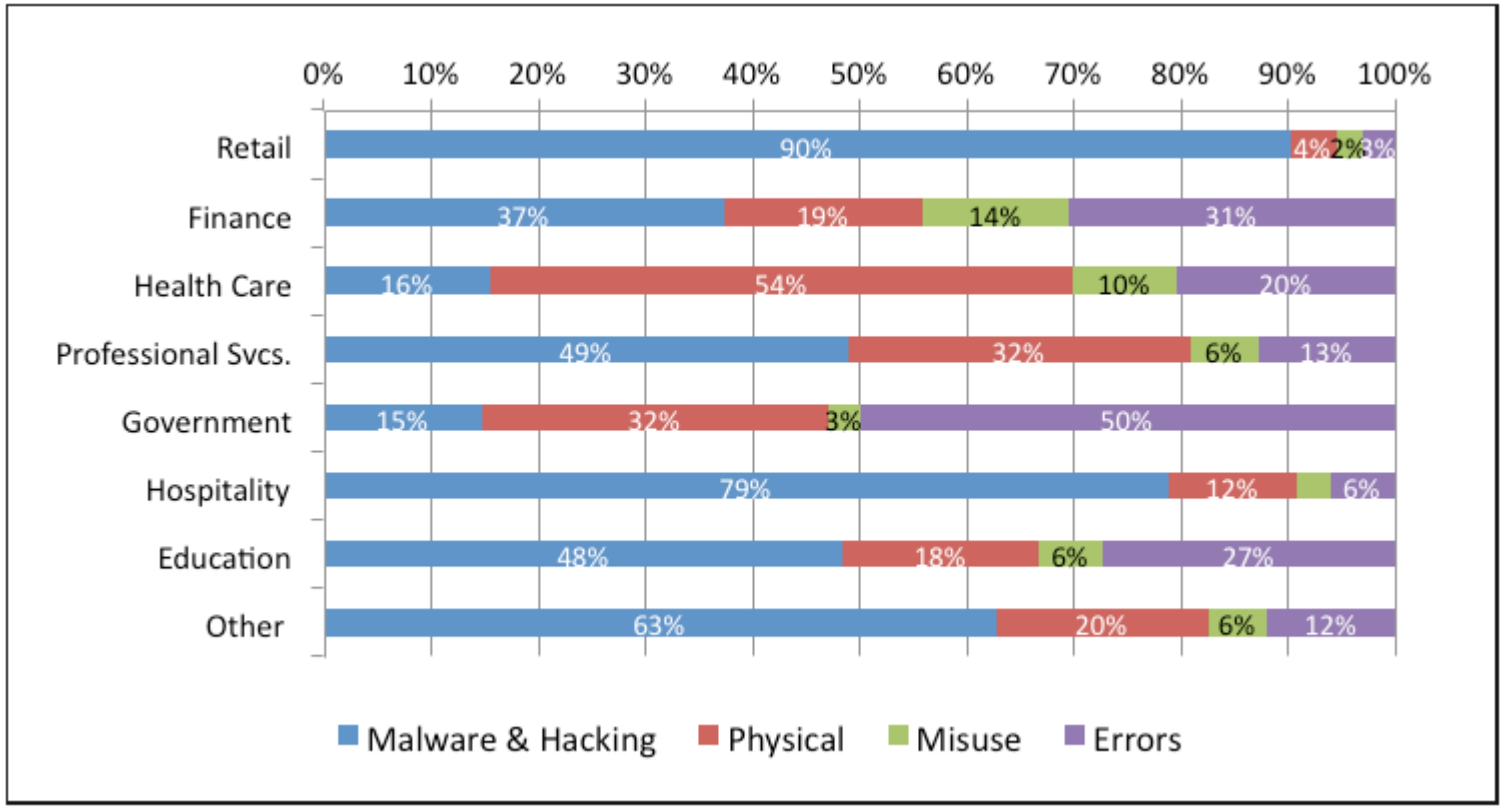


PRIVACY-BY-DESIGN

- Privacy-by-design integrates privacy considerations throughout the information life cycle
- It helps to make sure that people and organizations are asking the right questions so that data sharing is balanced with data protection
- For an open data platform, not only do we want to make sure the data offers value, but that we also do not inadvertently put individual privacy at risk
- What is the County doing?
  - Established an Enterprise Data Risk Governance Committee & Working Group
  - Beginning to develop charter and governance around data sharing
  - Projects involving data sharing and requests for data (among other projects) will be vetted prior to moving forward
  - The goal is to make sure that leadership knows about upcoming project requests, potential risks to privacy/security/legal obligations

WHAT ARE THE CAUSES OF DATA BREACHES

- California Attorney General released a report in 2016 on data breaches from 2012 to 2015
- California law requires reporting data breaches to the Attorney General
- Government sector data losses were primarily due to employee error



What are error breaches?

Errors by insiders (i.e., leadership, employees, contractors) that resulted in breaches included sending information by email to unintended people and unintentionally making information available to unauthorized persons (e.g., screen views or online posts including text or images).

WHAT ARE THE CHANCES OF A DATA BREACH AND HOW MUCH DOES IT COST?

$$$

- The Ponemon Institute conducts annual study to determine cost of data breaches
- For 2018, the likelihood of a data breach is 1 in 4, or 27% over a two year period
- For public sector agencies, the average cost is $75 per record
- Total average cost for a data breach for a public sector agency is $2.8 million

WHAT METHODS CAN BE USED TO PROTECT IDENTITY?

METHOD	BASIC DESCRIPTION
Removing fields	Deleting fields that contain sensitive information.
Removing records	Deleting records that are particularly sensitive, either because of the type of event represented or because of rare (and hence more easily identifiable) features.
Aggregating data	Summarizing data across the population and releasing a report of those statistics.
Generalizing data	Reducing the precision of fields in order to make each entry less unique.
k-anonymity	Generalizing fields such that at least k individuals exhibit each feature within those fields. Different traits will require a different level of generalization, depending on how many other entries exhibit that trait.
Adding noise (random perturbation)	Adjusting data with randomness to offset its original information.
Creating anonymous identifiers	Replacing attributes with randomly generated codes that have no underlying connection to the attribute they replace. This is done through a correspondence table, in which each unique attribute is paired with a random identifier that will replace that attribute wherever it appears.
Differential privacy	Differential privacy is a formal mathematical definition of privacy that provides a provable guarantee of privacy against a wide range of potential attacks. It is not a single tool, but rather a standard of privacy that many tools have been devised to satisfy.

Open Data Privacy, A risk-benefit, process-oriented approach to sharing an protecting municipal data, Berman Klein Center, Harvard University.

WORKING WITH PRIVACY IN MIND

- Only share data with those who are authorized to receive it
- Use County-approved filesharing & collaboration tools (not Dropbox or Google Docs)
- Securely store paper records and don't leave files in plain sight
- Report any known or even suspected data breaches
- Lock your screens before you step away (Ctrl-Alt-Del)
- Manage privacy settings

"Do you mind? It's private."

PRIVACY OFFICE

COUNTY OF SANTA CLARA

Email: PrivacyOffice@ceo.sccgov.org

Internal website: https://sccconnect.sharepoint.com/sites/cpo

External website for constituents: https://www.sccgov.org/sites/cpo