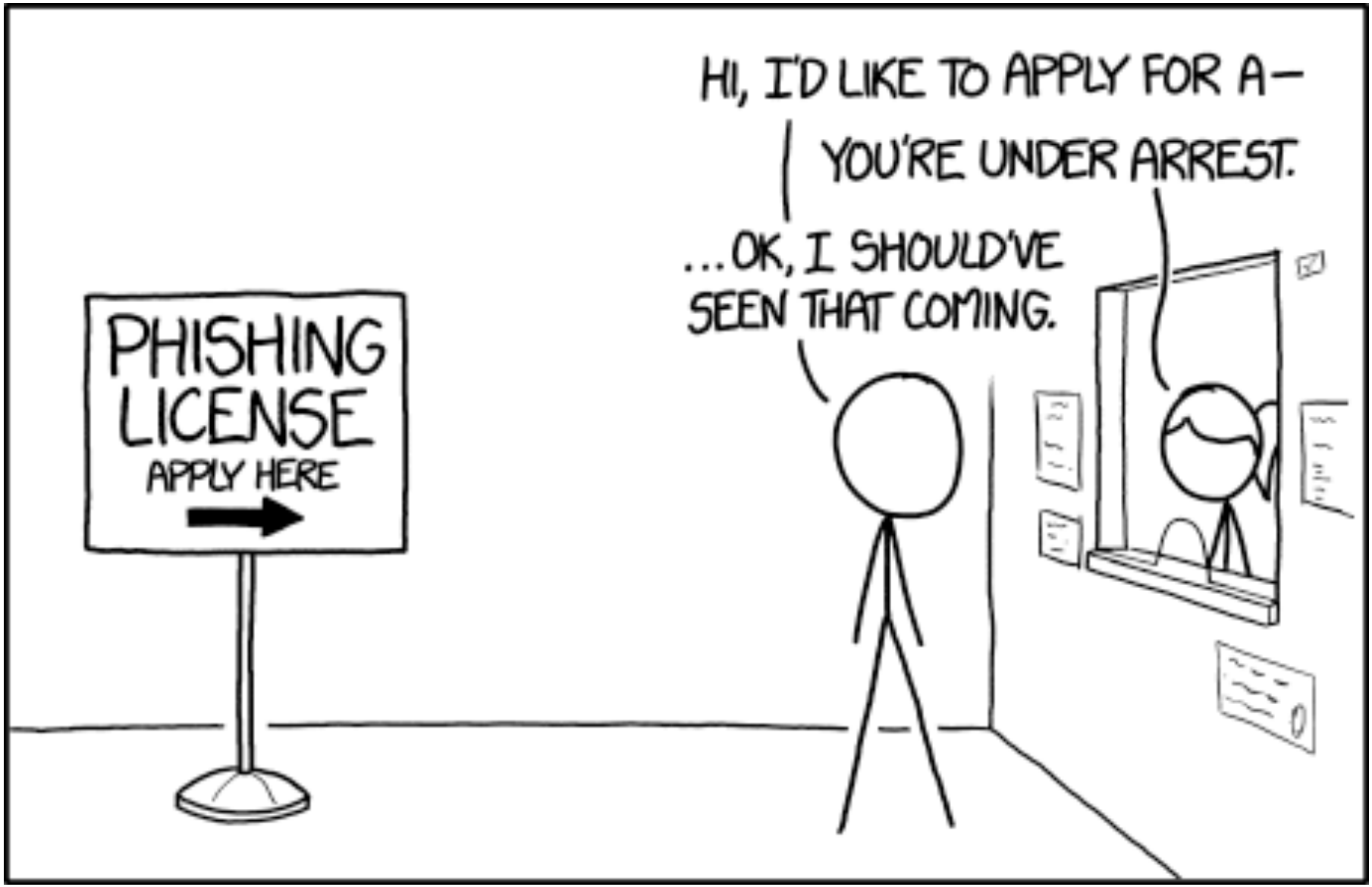
CS 55:
Security and Privacy

Legal issues, economics, ethics



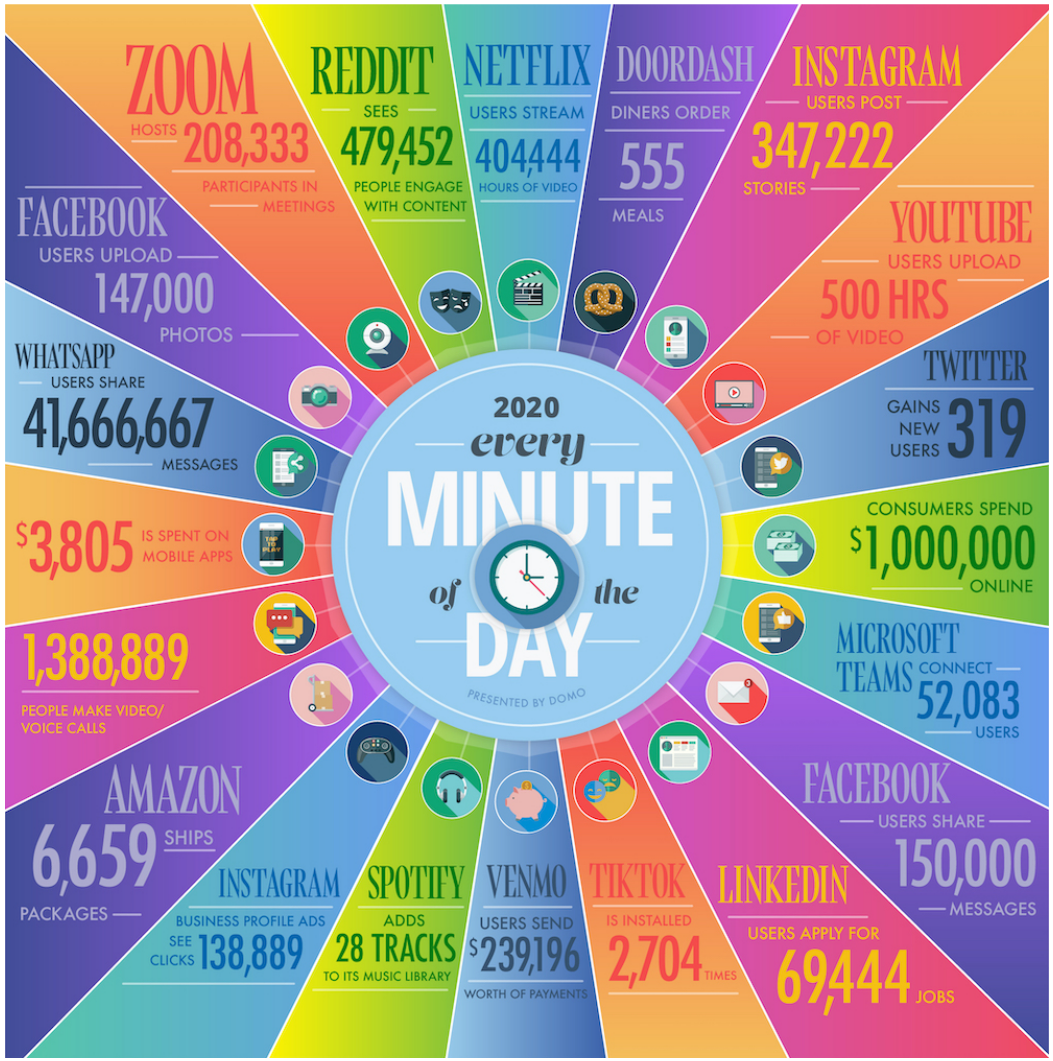
https://xkcd.com/1694/

Agenda

1. Legal aspects of privacy and data protection
2. Forensics and rules of evidence
3. Economics
4. Ethics

Today we collect lots of data…

Five V's of big data



https://www.domo.com/learn/data-never-sleeps-8

Big data characterized by five V's:

1. **Volume**: quantity of data to be stored, systems can be scaled
   - Vertically : "get a bigger box"
   - Horizontally: "get more boxes"
2. **Velocity**: speed at which data must be processed
   - Stream processing: analyze data as it comes
   - Feedback loop: data generates recommendations, recommendations lead to more data
3. **Variety**: store data in many forms
   - Structured data: fits into predefined data model
   - Unstructured data: does not fit data model
4. **Veracity**: can the data be trusted?
5. **Value**: can we exact value from the data, perhaps by correlating with other data?

Data privacy concerns abound!

As more data is collected and analysis techniques improve, privacy risks increase

More data:

- Increases the probability of re-identifying individuals in anonymous datasets
- Even in datasets which seem not to have personal linking information

Better analysis

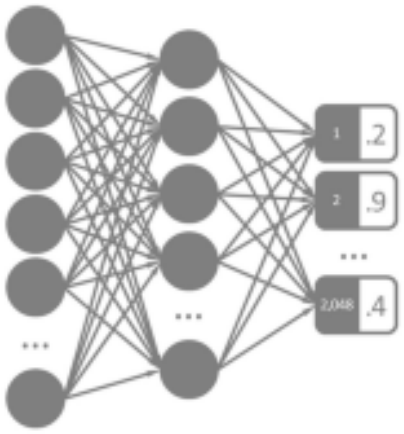
- Can infer sensitive information from “harmless” personal data
- Inferred information can much more privacy sensitive
- Can reveal information not intended to be revealed

Gruschka, Nils, et al. "Privacy issues and data protection in big data: a case study analysis under GDPR." *2018 IEEE International Conference on Big Data (Big Data)*. 2018.

Example: algorithm claimed to infer political orientation from facial features

Detect face (Face++)

Crop and resize (224 x 224 pixels)

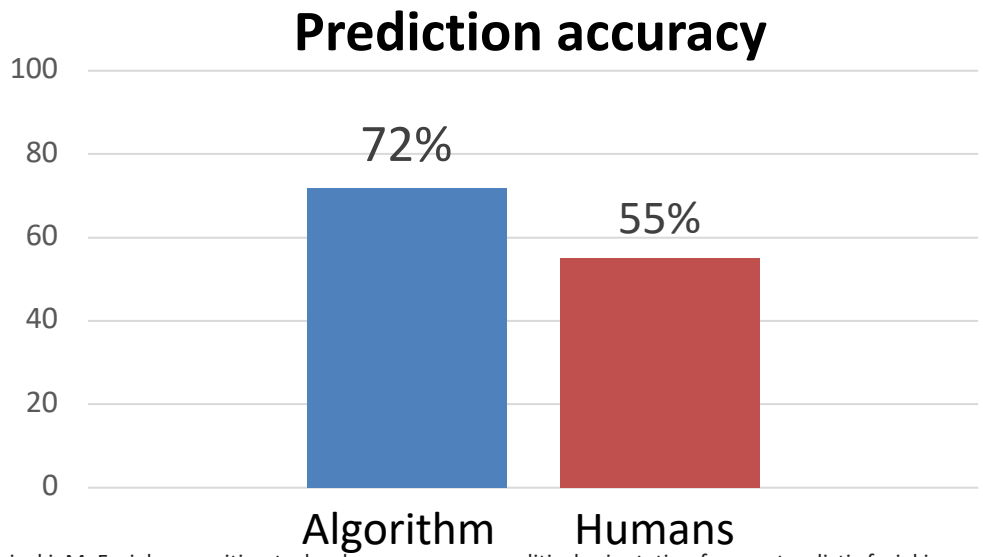


Extract 2,048 face descriptors (VGGFace2)

Cross-validated Logistic Regression *(or other similarity measure)*

$P_{liberal}$ = 38%

Compare with liberal and conservative faces



Other examples of algorithm accuracy

- Emotions: 94%
- Sexual orientation: 91%
- Personality: 64%

From just a photo of a face!

Kosinski, M. Facial recognition technology can expose political orientation from naturalistic facial images. *Sci Rep* **11,** 100 (2021).
Gruschka, Nils, et al. "Privacy issues and data protection in big data: a case study analysis under GDPR." *2018 IEEE International Conference on Big Data (Big Data)*. 2018.

Even anonymized data can be re-identified

Netflix challenge

- Netflix anonymized dataset of 500,000 subscribers
- No Personally Identifiable Information (PII), just subscriber ID and movie ratings

Re-identification

- Researchers used IMDB ratings with Netflix dataset
- Able to identify individual subscribers with high probability using ratings from both Netflix and IMDB and dates
- Also inferred political preferences, religious affiliations, sexual orientation, and other sensitive data

Another example: phone-based location tracking

- 95% of individuals positively identified by location traces

Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets." *2008 IEEE Symposium on Security and Privacy (Oakland)*. IEEE, 2008.
de Montjoye, YA., Hidalgo, C., Verleysen, M. *et al.* Unique in the Crowd: The privacy bounds of human mobility. *Sci Rep* **3,** 1376 (2013). https://doi.org/10.1038/srep01376

General Data Protection Regulation (GDPR) protects European citizen's data

General Data Protection Regulation (GDPR)

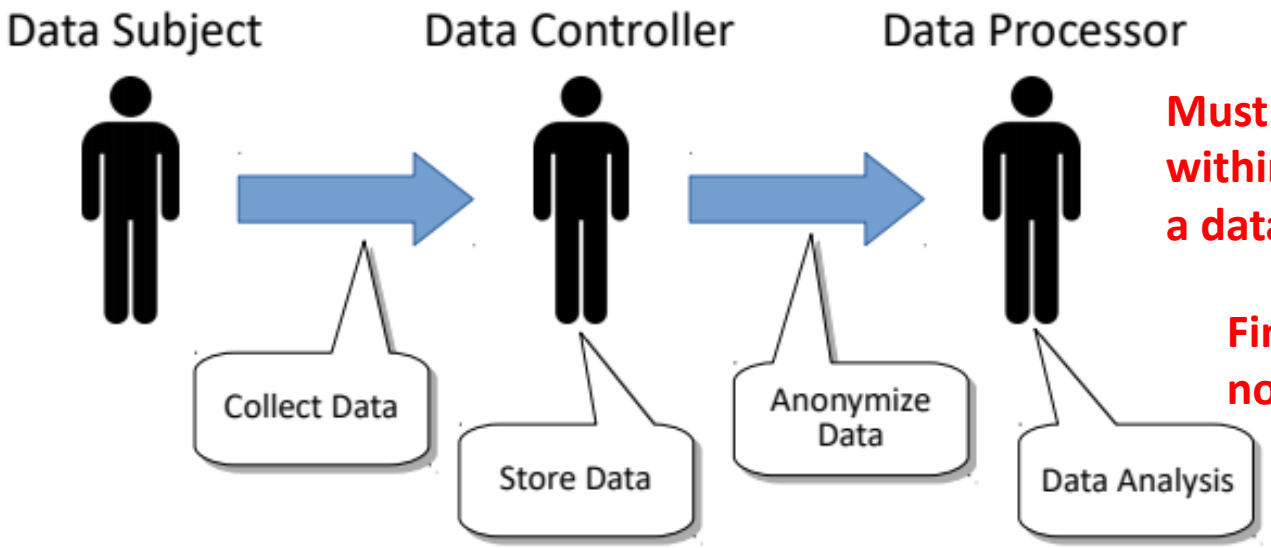
- Went into effect in May 2018
- Provisions are consistent across all 28 EU member states
- Importantly also extends to organizations from other countries that process data on European citizens

- Regulates the collection, storage, and processing of personal data
  - Personal identifiers (full name, national ID number)
  - Indirect identifiers (phone number, IP addresses, photos with identifiable people)
  - Data that do not include these identifiers are regarded as anonymous and out of scope of GDPR (e.g., statistical analysis without direct links to specific people)
- Includes the "right to be forgotten"

Gruschka, Nils, et al. "Privacy issues and data protection in big data: a case study analysis under GDPR." *2018 IEEE International Conference on Big Data (Big Data)*. 2018.

GDPR includes provisions to make personal identification difficult

General Data Protection Regulation (GDPR)



Collection

- In most cases processing of personal data is only allowed if the data subject has given consent
- Must be limited to specific purpose
- Cannot change purpose later
- Only collect what is necessary

Anti-identification

- Data should be pseudonymous
- Use random ID for each person
- Data controller can keep a look up table
- Appropriate technical means must be used to ensure security
- Anonymize using well-known techniques

Gruschka, Nils, et al. "Privacy issues and data protection in big data: a case study analysis under GDPR." *2018 IEEE International Conference on Big Data (Big Data)*. 2018.
https://www.cynet.com/cynet-for-compliance/gdpr-data-breach-notifications-everything-you-need-to-know/

Data can be distorted to make personal identification difficult

Data distortion

1. Suppression: remove data or replace with dummy value (e.g., *)
2. Generalization:
   - Date of birth -> age (in years)
   - Age (in years) -> range of years
   - Zip code -> first two digits
3. Permutation: partition data into groups, shuffle sensitive values
4. Perturbation: replace values, but keep statistical properties (e.g., add noise)

Gruschka, Nils, et al. "Privacy issues and data protection in big data: a case study analysis under GDPR." *2018 IEEE International Conference on Big Data (Big Data)*. 2018.

Secure Multi-Party Computation (MPC) is another option to protect identity

Secure Multi-Party Computation (MPC)

Compute average salary without revealing any person's salary

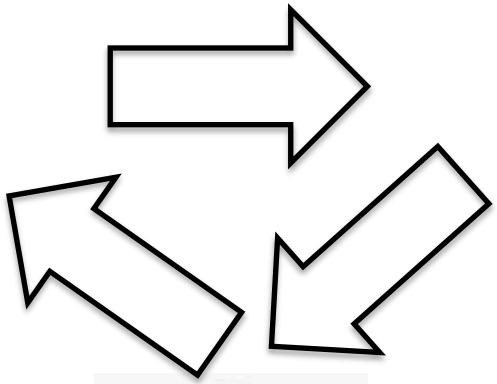
- Adds secret number to her salary
- Encrypts result with Bob's public key

Alice

- Decrypts with private key
- Subtracts secret number
- Divides by number of participants
- Shares result

Bob

- Decrypts with private key
- Adds salary to value from Alice
- Encrypts with Carol's public key



Carol

- Decrypts with private key
- Adds salary to value from Bob
- Encrypts with Alice's public key

Average is computed, but no one learns anyone else's salary

Adapted from https://cs.stanford.edu/people/eroberts/courses/soco/projects/2004-05/cryptography/smc.html

Agenda

1. Legal aspects of privacy and data protection
2. Forensics and rules of evidence
3. Economics
4. Ethics

Forensic investigations normally proceed through four stages

Forensic examination process

Identify, label, record, and acquire data from the possible sources of relevant data

- Video record entire collection process
- Wear gloves
- Use Faraday bags
- Bring power cables
- Create chain of custody

Extract data of interest while preserving data integrity

- May have to pull back up tapes
- May search data using keywords
- Often identifies numerous non-relevant documents

Analyze the results of examination stage

- Use legally justifiable methods to derive useful information
- Lawyers may review data to protect non-relevant or privileged communications (e.g., client-layer confidentiality)

Report results of analysis

- Explain tools and procedures used
- Identify actions that still need to be performed

Problems: Too much information!
Even small-time drug dealers have lots of devices
Some data stored in cloud (even if police have device)

NIST Publication 800-86

Oettinger, William. *Learn Computer Forensics: A beginner's guide to searching, analyzing, and securing digital evidence*. Packt Publishing Ltd, 2020.

Discussion

Why is computer crime hard to prosecute?

Agenda

1. Legal aspects of privacy and data protection
2. Forensics and rules of evidence
3. Economics
4. Ethics

Discussion

Every dollar that you spend on security comes out of my bonus!

Why are you spending so much on an offsite disaster recovery location?!?!?

What do you say?

If something happens, you'll find me in Montana, or better yet, you won't find me in Montana!

Annualize Loss Expectancy (ALE) is a common way to justify security spending

You shouldn't spend $5,000 on a safe to protect a $2 item

How much should you spend?

ALE = damage * probability

Pierson's comments from the trenches:

- ALE is easy to say but hard to use in practice
- Some things are hard to value
  - Customer confidence in firm if event happens
  - Staff morale
- Estimating the probability of an adverse event is also difficult
- Be wary of vendor's estimation of threats – they have an agenda!

Adapted from Cole, Eric. *Advanced persistent threat: understanding the danger and how to protect your organization*. Syngress, 2013.

Another approach is to benchmark spending by other firms in the industry

Industry overall spends >$120B on security hardware, software and services

Many firms survey industry participants and gather data about security spending

These firms sell reports summarizing their findings

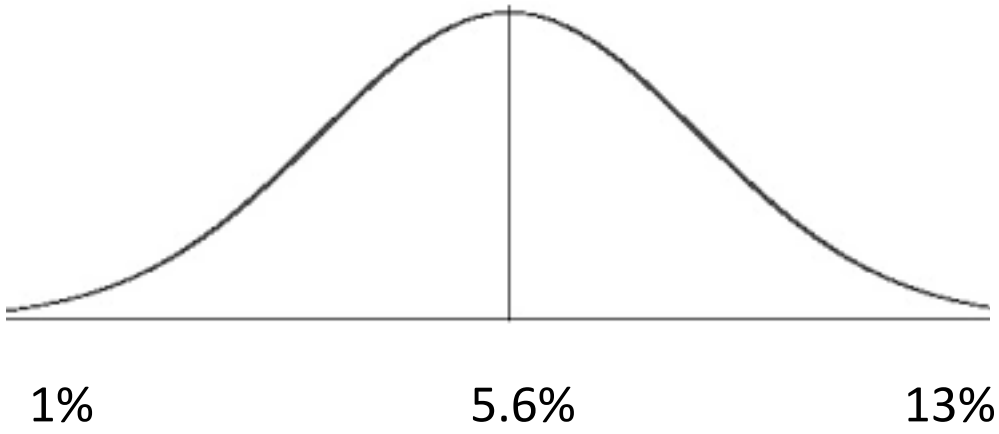
You can find how much your industry spends on security (and other items)

Normally presented as a percentage of sales

https://cybersecurity.att.com/blogs/security-essentials/how-to-justify-your-cybersecurity-budget

Companies tend to spend around 5% of their IT budget

Security spending as percentage of IT budget



Ranges from 1% to about 13% of IT budget

- Varies by industry
- Financial firms tend to spend more than firms in other industries

"We never had a cybersecurity budget until recently when I said we need a dedicated budget for it.

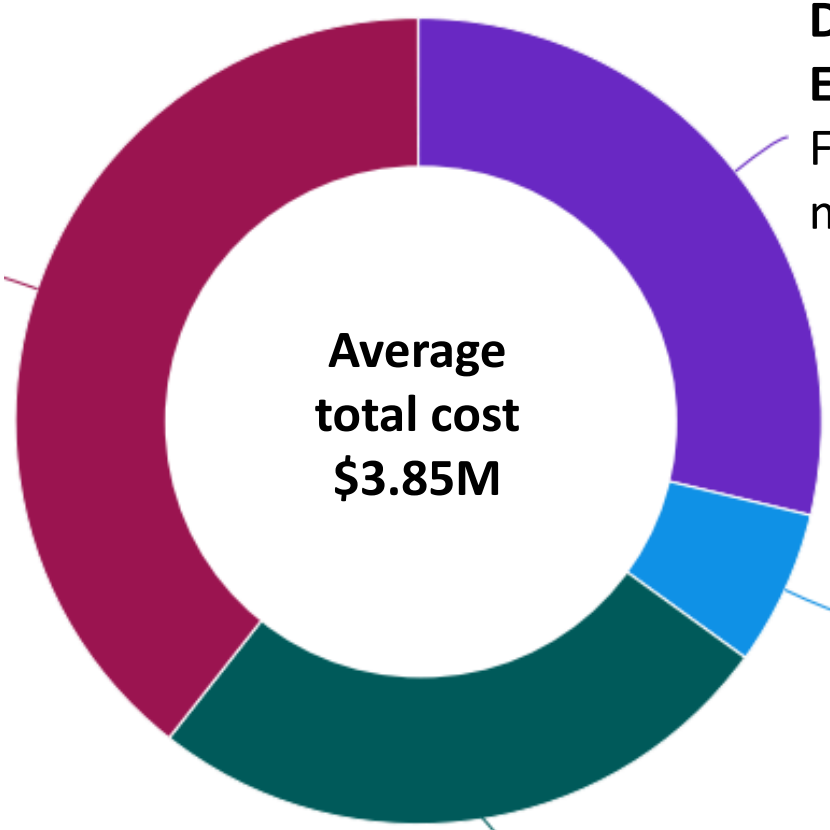
So, they took a chunk of the IT budget and told me to be grateful!

We have no more money overall, but some for cybersecurity!"
- Anonymous CISO

https://cybersecurity.att.com/blogs/security-essentials/how-to-justify-your-cybersecurity-budget

The global average cost of a data breach in 2020 was $3.86M

Global average cost ($M)



Detection and Escalation $1.11M
Forensics, crisis management

Lost Business $1.52M
Business disruption, revenue loss

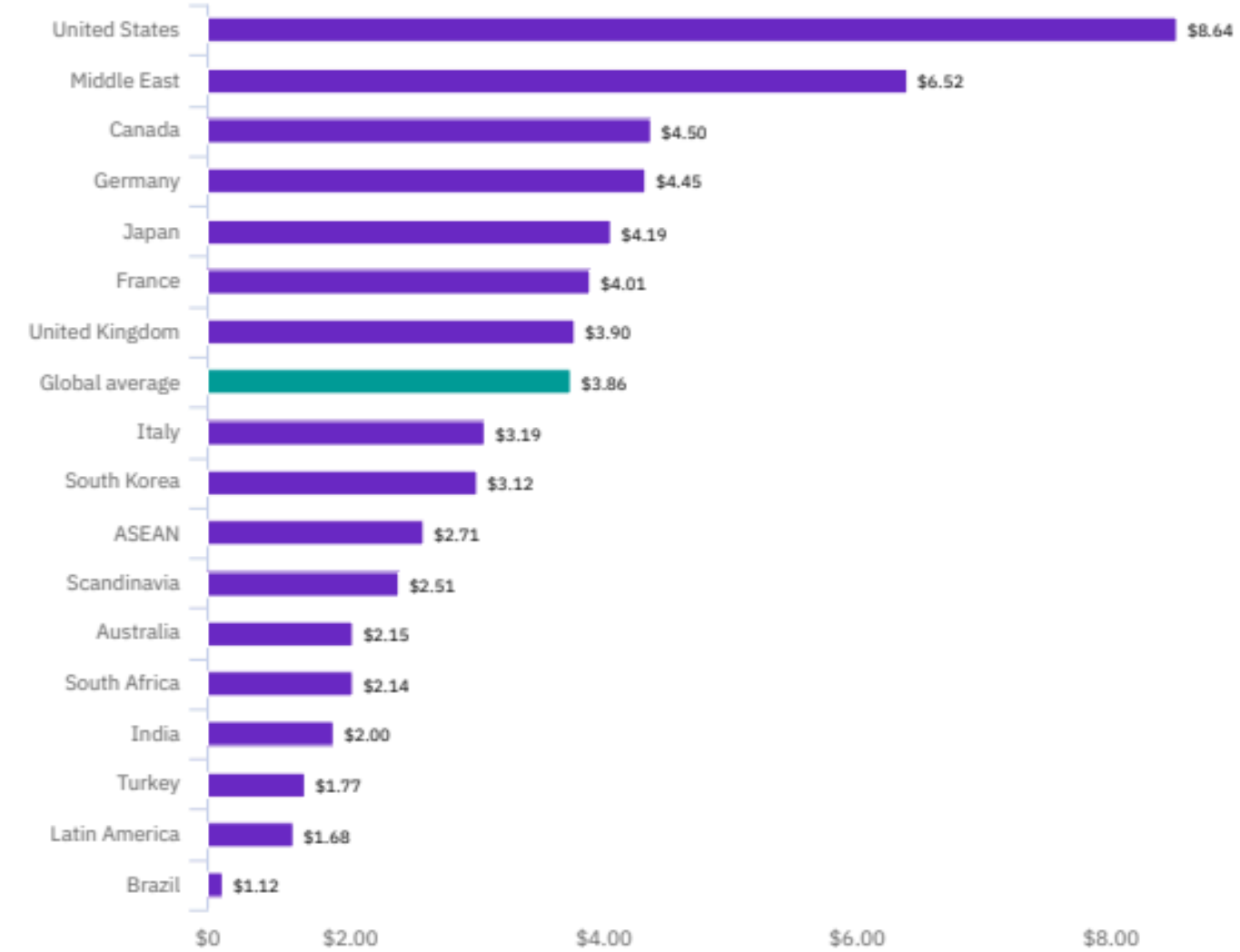
Notification $0.24M
Informing subjects of breach

Post incident $0.99M
Help desk inquiries, credit monitoring, legal expenses, fines

https://www.ibm.com/security/digital-assets/cost-data-breach-report/#/

The average cost of a data breach in the U.S. was more than double the average

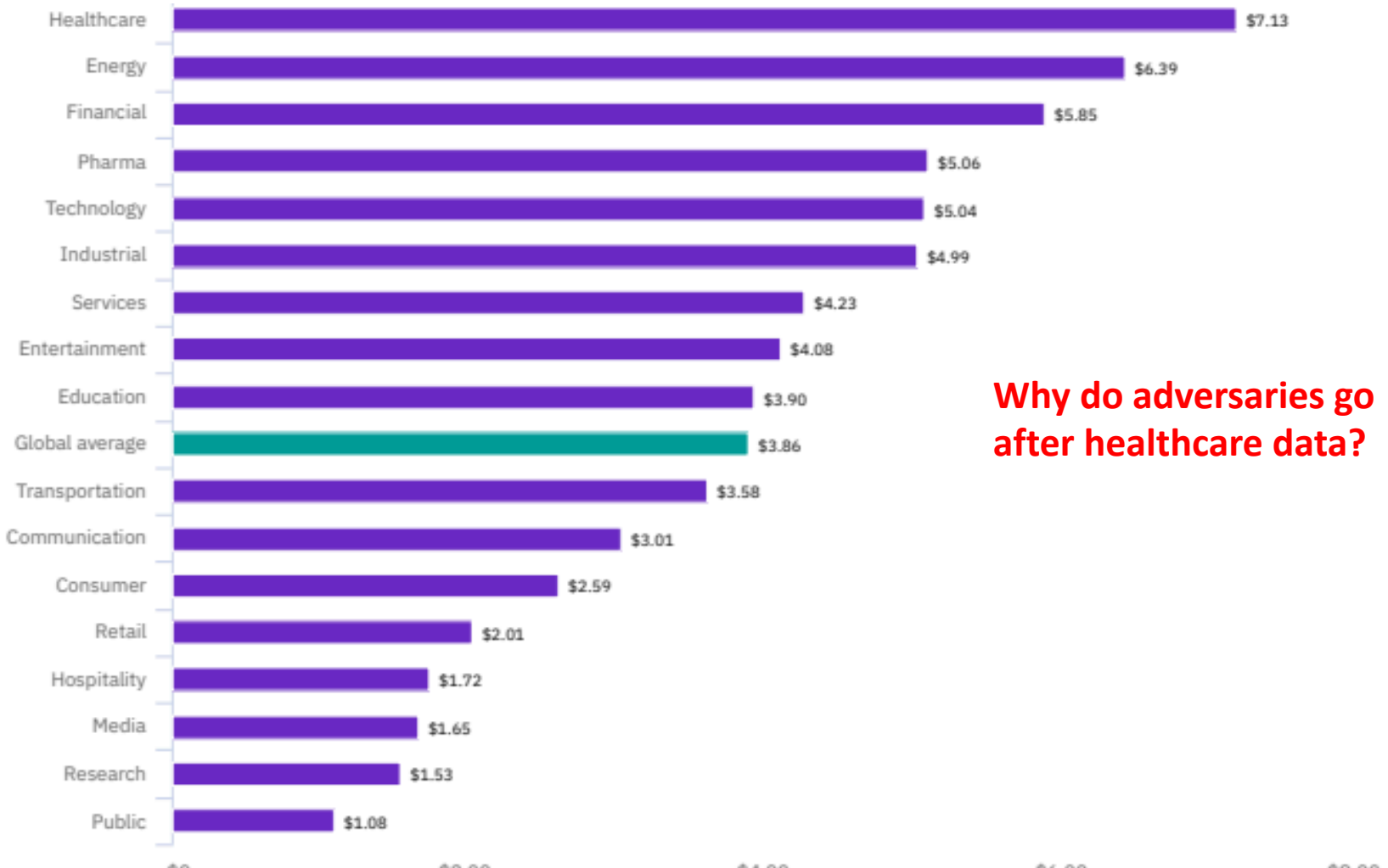
Average cost by country ($M)



https://www.ibm.com/security/digital-assets/cost-data-breach-report/#/

Breach costs were highest in healthcare, energy, and financial firms

Average cost by industry ($M)

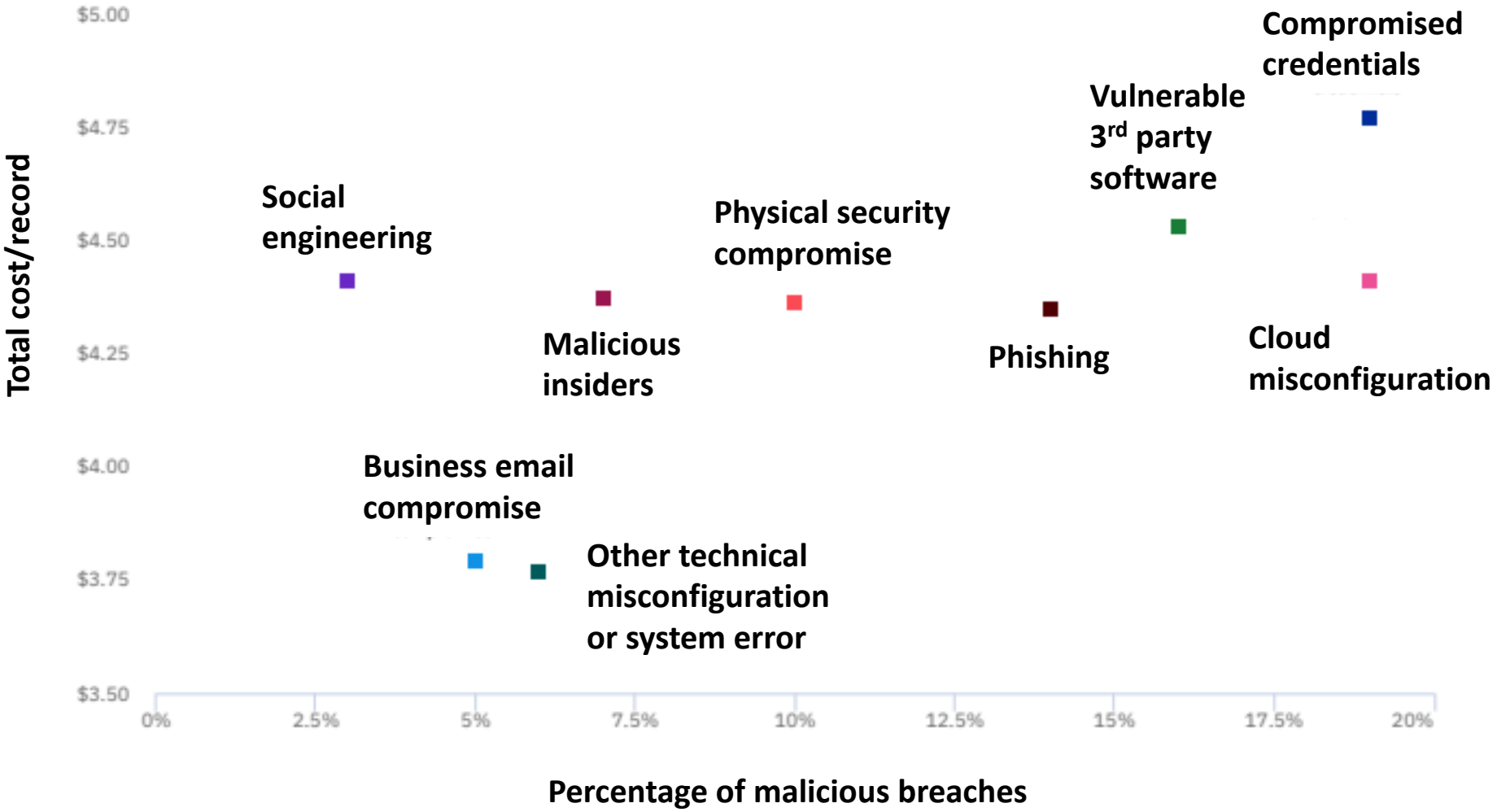


Why do adversaries go after healthcare data?

https://www.ibm.com/security/digital-assets/cost-data-breach-report/#/

Five root causes accounted for 78% of data breaches caused by malicious attacks

Root cause by threat vector



https://www.ibm.com/security/digital-assets/cost-data-breach-report/#/

An incident response plan is a set of instructions to respond to security events

Goal: respond efficiently to a security-related event

Steps to create an incident response plan

- Identify an incident recovery team
- Determine critical components of the network
- Identify single points of failure and address them
- Create a disaster recovery plan
- Create an incident response plan
  - List roles and responsibilities
  - Summarize tools, technologies, and physical resources that must in place
  - List critical network and data recovery processes (e.g., what order to bring up critical systems and who will accomplish them)
  - Plan for communications, both internal and external
- Train staff/practice plan

Firms saved $2M/incident on average if they had a robust incident response plan

Adapted from https://www.cisco.com/c/en/us/products/security/incident-response-plan.html#~how-to-create-a-plan
https://www.ibm.com/security/digital-assets/cost-data-breach-report/#/

Agenda

1. Legal aspects of privacy and data protection
2. Forensics and rules of evidence
3. Economics
4. Ethics

Discussion

Is video surveillance good or bad (or in between)?

Who should be accountable if a system is hacked and a negative outcome results?

Who is accountable for decisions made by autonomous systems such as a self-driving car?

Discussion

Alice works for a large software company

- She notices company servers are not busy overnight
- She uses company servers (without permission) over night to compute ML stock market analysis
- Analysis does not cause drain on other projects

Is it ethical for Alice to use these resources?

Discussion

Bob works as a programmer writing accounting software

- Bob's manager Carol asks him to write a routine to allow Carol to post entries directly to the company's ledger
- This action would violation double-entry GAAP accounting principles
- There would be no way to trace changes made to the ledger
- Bob raises concerns to Carol, but Carol says that fraudulent entries are her responsibility, that Bob should just write the code

Is Bob an accessory to fraud? What can Bob do?

Discussion

Donald works for a company's records department where he has access to property tax records

As part of a scientific study, Edith has been granted access to numerical data not associated with property owner names

Edith finds some information that should like to use, but she needs the names and addresses of certain properties

She asks Donald to provide that information

Should Donald release that information?

The average cost of a data breach in 2020 was $3.86M

Factors involved with breach cost:

- **Detection and escalation**
  Forensics, crisis management
- **Notification**
  Notify subjects of breach
- **Lost business**
  Business disruption, revenue loss
- **Post response**
  Help desk inquiries, credit monitoring, legal expenses, fines

Average cost/breach: $3.86M

Average cost/PII record: $150

https://www.ibm.com/security/digital-assets/cost-data-breach-report/#/