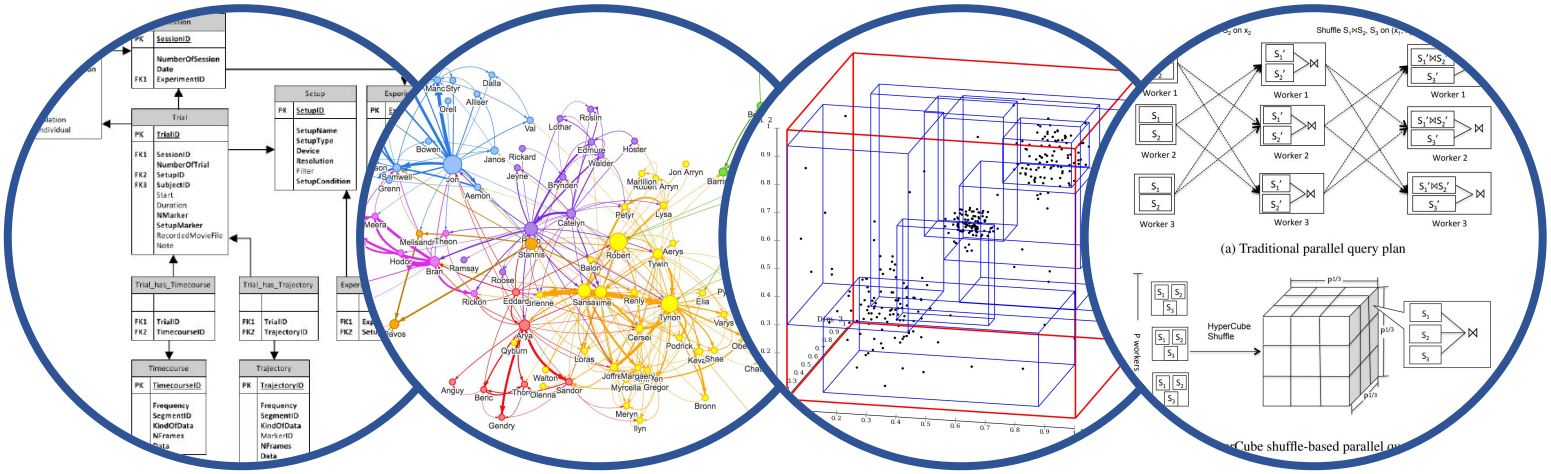


Introduction to Data Management

Practical Data Management

Alyssa Pittman
Based on slides by Jonathan Leang, Dan Suciu, et al

Paul G. Allen School of Computer Science and Engineering
University of Washington, Seattle

Goals for Today

- Talk about the parts of data management you might encounter in the real world without having theory to back you up

Outline

- Views
- Data Cleaning
  - ETL
  - Data wrangling on GCP Dataprep (Trifacta)
- Data Management Ethics and Best Practices

Views

- A view is a relation defined by a query:

```
Customer(cid, name, city)
Purchase(cid, pid, store)
Product(pid, name, price)
```

This is like a new relation StorePrice(store, price)

```
CREATE VIEW StorePrice AS
SELECT x.store, y.price
FROM Purchase x, Product y
WHERE x.pid = y.pid
```

Views

- Views can be queried just like tables:

```
SELECT DISTINCT z.name, u.store
FROM Customer z, Purchase u, StorePrice v
WHERE z.cid = u.cid AND u.store = v.store
      AND v.price > 1000
```

- The semantics are the same as using a subquery:

```
SELECT DISTINCT z.name, u.store
FROM Customer z, Purchase u,
     (SELECT x.store, y.price
      FROM Purchase x, Product y) v
WHERE x.pid = y.pidStorePrice v
WHERE z.cid = u.cid AND u.store = v.store
      AND v.price > 1000
```

Applications of views

- Logical data independence
- Security
- Increased physical data independence

Applications of views

- Logical data independence

Say I want to normalize my database, but have many old queries using the original schema.

…create views consistent with the old schema!

Applications of views

- Security

Give users access to only the data they need.

Name	Address	Balance
Mary	Houston	450.99
Sue	Seattle	-240
Joan	Seattle	60.23
Ann	Portland	-23.50

Advertising team shouldn't see the balances

But they can see this

```
CREATE VIEW PublicCustomers
  SELECT Name, Address
  FROM Customers
```

Applications of views

- Security

Give users access to only the data they need.

Name	Address	Balance
Mary	Houston	450.99
Sue	Seattle	-240
Joan	Seattle	60.23
Ann	Portland	-23.50

But they can see this

Collections team shouldn't see people with positive balances

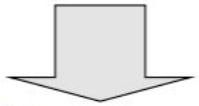
```
CREATE VIEW NegativeBalanceCustomers
SELECT * FROM Customers WHERE Balance < 0
```

Applications of views

- Increased physical data independence
  - Vertical partitioning

Resumes

SSN	Name	Address	Resume	Picture
234234	Mary	Huston	Clob1…	Blob1…
345345	Sue	Seattle	Clob2…	Blob2…
345343	Joan	Seattle	Clob3…	Blob3…
234234	Ann	Portland	Clob4…	Blob4…



T1

SSN	Name	Address
234234	Mary	Huston
345345	Sue	Seattle
. . .		

T2

SSN	Resume
234234	Clob1…
345345	Clob2…

T3

SSN	Picture
234234	Blob1…
345345	Blob2…

10

Applications of views

- Increased physical data independence
  - Vertical partitioning
    - Helpful for data warehousing
    - Can improve performance
      - Have large columns (e.g. photo blob)
      - Have lots of columns but each query only accesses a few

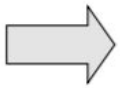
```
CREATE VIEW Resumes AS
SELECT T1.ssn, T1.name, T1.address,
T2.resume, T3.picture
FROM T1,T2,T3
WHERE T1.ssn=T2.ssn and T2.ssn=T3.ssn
```

Applications of views

- Increased physical data independence
  - Horizontal partitioning

Customers

SSN	Name	City	Country
234234	Mary	Houston	USA
345345	Sue	Seattle	USA
345343	Joan	Seattle	USA
234234	Ann	Portland	USA
--	Frank	Calgary	Canada
--	Jean	Montreal	Canada



CustomersInHouston

SSN	Name	City	Country
234234	Mary	Houston	USA

CustomersInSeattle

SSN	Name	City	Country
345345	Sue	Seattle	USA
345343	Joan	Seattle	USA

CustomersInCanada

SSN	Name	City	Country
--	Frank	Calgary	Canada
--	Jean	Montreal	Canada

15

Applications of views

- Increased physical data independence
  - Horizontal partitioning
    - Helpful for data warehousing

```
CREATE VIEW Customers AS
CustomersInHouston UNION ALL
CustomersInSeattle UNION ALL
. . .
```

Applications of views

- Increased physical data independence
  - Partitioning
    - Mostly helpful for

Views trivia

- Virtual views
  - Computed on the fly - potentially slow
  - Always up-to-date
- Materialized views
  - Pre-computed and stored - fast to access
  - May have stale data
- Can views be updated?
  - Some SQL variants let you update the data behind *simple* views

Where is my data coming from?

- You generate the data
  - Output data that is easy to use
- External sources or preexisting data
  - Sometimes doesn't fit your application needs
  - Need to translate the data into a usable form

Extract Transform Load (ETL)

"I know exactly what operations need to be done to get from data format A to data format B"

- Extract
  - Read relevant data
- Transform
  - Push data through mapping functions until done
    - Aggregations
    - Normalization
    - …
- Load
  - Write to destination

Extract Transform Load (ETL)

Data Wrangling

"I have no clue what's going on with my data"

- Essentially ETL but with **data exploration**
- Interactivity is important
  - Visualizations
  - Suggestions

Pivot

- Create a "summary table"
  - Generally used for reports to draw attention to interesting values
  - Able to make values into columns
- "Skinny and tall" □ "short and wide"

Name	Year	GDP
Angola	2015	100
Luxembourg	2015	50
Angola	2016	110
Angola	2018	115
Luxembourg	2017	55
Luxembourg	2018	65

Pivot

- Create a "summary table"
  - Generally used for reports to draw attention to interesting values
  - Able to make values into columns
- "Skinny and tall" □ "short and wide"

Name	2015	2016	2017	2018
Angola	100	110		115
Luxembourg	50		55	65

Unpivot

- Usually we want to store unpivoted data
  - Easier to manage
- "Short and wide" □ "skinny and tall"

Name	2015	2016	2017	2018
Angola	100	110		115
Luxembourg	50		55	65

Data Wrangling

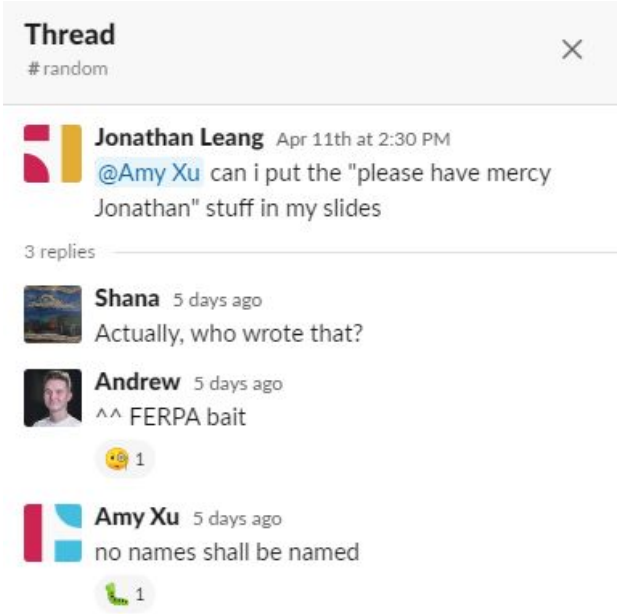
Quickstart - demo

Now what?

You can get data but what are you doing with it?

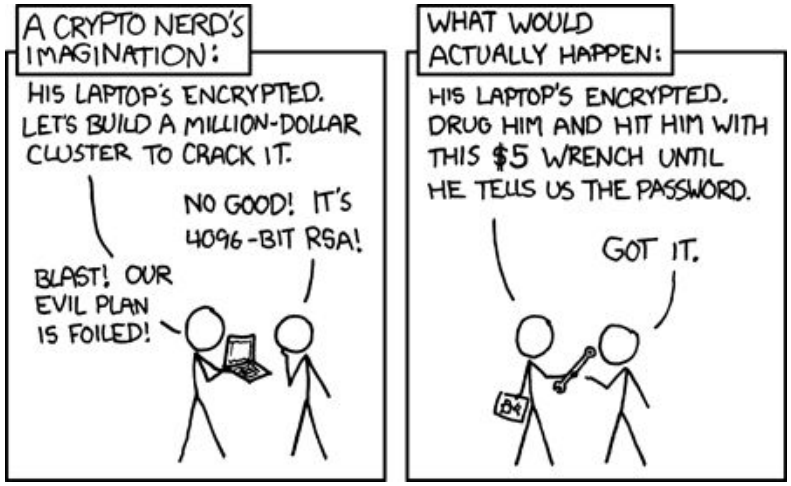
Existing Laws and Regulations

- FERPA (Family Education Rights and Privacy Act)
- Mandatory for education institutions
  - Requires written consent to disclose academic info
  - Allows the release of directory information



Existing Laws and Regulations

- HIPAA (Health Information Portability and Accountability Act)
- Mandatory for healthcare and health insurance institutions
  - Privacy Rule to protect Protected Health Information
  - Security Rule to ensure administrative, physical, and technical safeguards



Existing Laws and Regulations

- SOX (Sarbanes-Oxley Act)
- Requires auditability for companies' financial records
  - What does this have to do with data?
  - Can financial data be tampered with?
  - Can code touching financial data be tampered with?

Existing Laws and Regulations

- GDPR (European General Data Protection Regulation)
- Requires disclosure from companies about what user data they have and how they use it
  - …but can be exploited

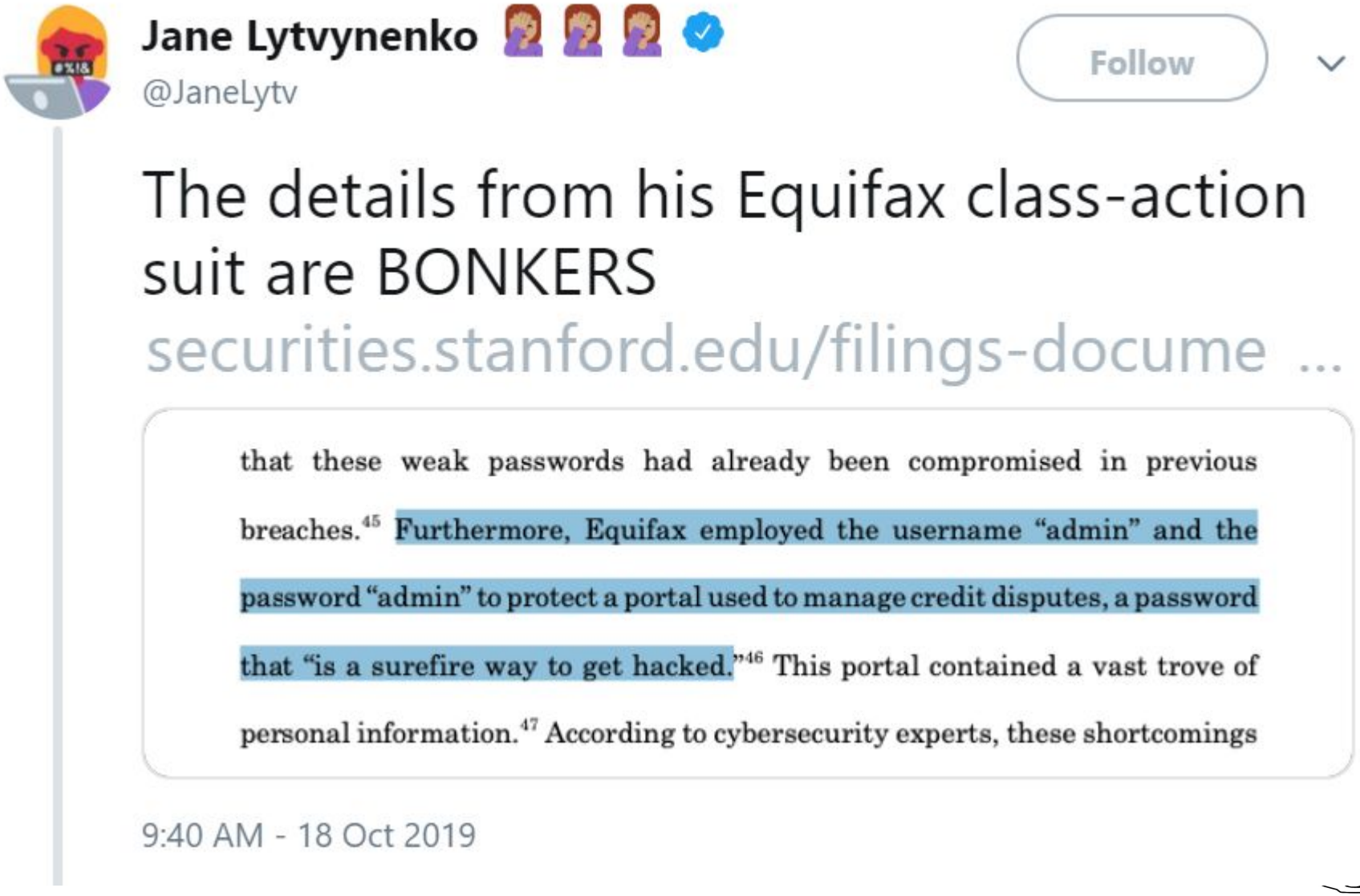
Laws and Regulations Today

- Social Media and Politics?
- Facebook-Cambridge Analytica Scandal
  - CA uses loophole in Facebook API through an online quiz to harvest personal information data

Whistleblower Christopher Wylie

Mark Zuckerberg's hearing

What's at Stake?



Sensitive Information

- PII = Personally identifying information
  - Names
  - Student ID
  - Social security number
  - License number
- Protected data (for legal and/or ethical reasons)
  - Academic records (FERPA)
  - Protected Health Information (HIPAA)
  - Customer records (GLBA)
- Passwords

Access Control

- Block people who shouldn't have access
  - Most large companies have a tiered-access hierarchy
- Databases usually have built-in access control:

```
GRANT <permissions>
  [ON <table>]
   TO <user/role>
```

```
GRANT SELECT, INSERT
   ON MySecureTable
   TO PUBLIC
```

Allow anyone who can connect to read and add data to MySecureTable

Permissions:
- Table-level operations (SELECT, DELETE, …)
- DB-level operations (CREATE TABLE, GRANT, …)

User/Role:
- Users like a user on your computer
- Roles (groups) can be predefined or created

Access Control

- SQL Injection application input acts as code
  - Union attack, tautology attack, illegal queries
  - Only possible if there is a place to inject code
  - Consistently one of the top web-based attacks
    - People simply don't realize its an issue or…
    - People know it's an issue and never get around to fixing it
- Considered a "solved" problem
  - **Parameterize queries with prepared statements**

Access Control

Other common techniques to limit access:

- Limit the number of rows that can be seen
  - Leaking a few tuples is better than leaking all of them
- Only allow aggregations
  - Grouping implicitly eliminates identification info
- Don't store data you don't need!

Anonymize Data

FERPA Deidentification

- ID to anonymous ID mapping should be secret
- Aggregate data (minimum n-size)
  - **Suppression** □ Don't provide data ☹
    - Necessary for very small groups
  - **Rounding** □ Bucket data or introduce noise 😊
    - More people means you can be more specific

Implicit Disclosure

- FERPA allows institutions to disclose “directory information” without consent (institution policies can be stronger)
  - Name
  - Email
  - Photographs
  - Phone Number
- If users can derive sensitive information like grades, it violates FERPA

Implicit Disclosure

- “Hey, can you give me the directory information for students with a GPA of 3.5?”

Implicit Disclosure

- "Hey, can you give me the directory information for students with a GPA of 3.5?"

Reveals sensitive information by context

```
SELECT D.*
  FROM Directory AS D, Grades AS G
 WHERE D.id = G.id AND
       G.gpa = 3.5
```

Implicit Disclosure

Re-identification of Mass. Governor William Weld

- Public voter data
  - Name
  - ZIP code
  - Sex
  - Birth date
  - …
- Anonymous insurance data
  - ZIP code
  - Sex
  - Birth date
  - Prescription
  - Diagnosis
  - …

Implicit Disclosure

Cambridge, MA Voter Data ($20)

Name	ZIP	Sex	Bday
…	…	…	…
W. Weld	12345	M	Feb 30
…	…	…	…

⋈

Anon. Insurance Data for Researchers

ZIP	Sex	Bday	MedInfo
…	…	…	…
12345	M	Feb 30	Affluenza
…	…	…	…

6 matches on ZIP
3 matches on Sex
1 match on Bday

Name	…	MedInfo
…	…	…
W. Weld	…	Affluenza
…	…	…

Implicit Disclosure

Cambridge, MA Voter Data ($20)

Name	ZIP	Sex	Bday
…	…	…	…
W. Weld	12345	M	Feb 30
…	…	…	…

⋈

Anon. Insurance Data for Researchers

ZIP	Sex	Bday	MedInfo
…	…	…	…
12345	M	Feb 30	Afluenza
…	…	…	…

Legal in 1997
Illegal since 2003

6 matches on ZIP
3 matches on Sex
1 match on Bday

Name	…	MedInfo
…	…	…
W. Weld	…	Afluenza
…	…	…

Storing Passwords

- Passwords are special
  - High potential for additional security compromises
  - Only operation that should be done is equality comparison

Storing Passwords

(bobtheninja246, password)

If you do this, Ted Codd will start rolling in his grave.

Username	Password
bobtheninja246	password
xXxDragonSlayerxXx	password
420_E-Sports_Masta	qwertyuiop

Storing Passwords

- Quick overview of hashing
  - Hash(input) □ hash value
  - Hashing is deterministic
  - Ideally hashing is noninvertible
  - Ideally hash values are uniformly spread out

Storing Passwords

Hash it!

(bobtheninja246, hash(password))

(bobtheninja246, FCgJFI9ryz)

Username	Hash
bobtheninja246	FCgJFI9ryz
xXxDragonSlayerxXx	FCgJFI9ryz
420_E-Sports_Masta	p8mel6usIF

Storing Passwords

Hash it!

(bobtheninja246, hash(password))

(bobtheninja246, FCgJFI9ryz)

Issues/pitfalls:

- Hashing functions have precomputed “rainbow tables”
- Some hashing functions are fast so brute forcing attacks can happen
- Patterns can occur for the same passwords

Username	Hash
bobtheninja246	FCgJFI9ryz
xXxDragonSlayerxXx	FCgJFI9ryz
420_E-Sports_Masta	p8mel6usIF

Storing Passwords

Salt it and hash it!

(bobtheninja246, slowhash(password * random salt), random salt)

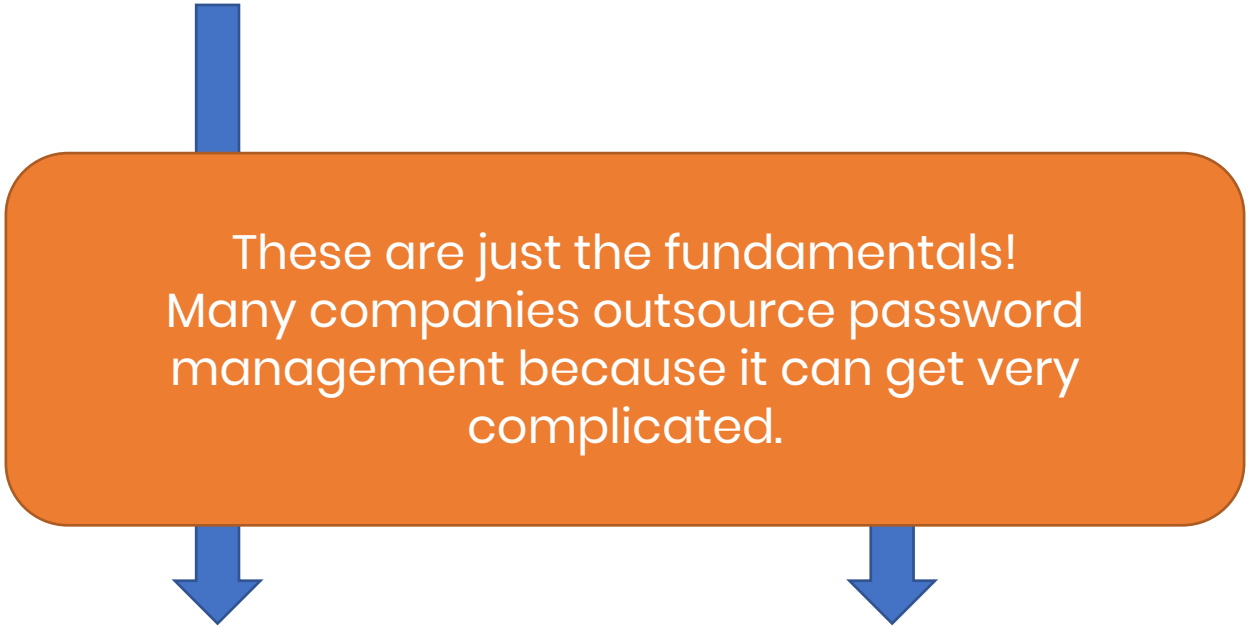
(bobtheninja246, slowhash(password * stored salt))

Username	Hash	Salt
bobtheninja246	HHxrd5o7Cn	WUKhhIFBLc
xXxDragonSlayerxXx	7rYFQIowpW	mq5rFL6JzF
420_E-Sports_Masta	cQF4DdSFfn	S8e0zpATNR

Storing Passwords

Salt it and hash it!

(bobtheninja246, slowhash(password * random salt), random salt)



stored salt))

Username	Hash	Salt
bobtheninja246	HHxrd5o7Cn	WUKhhIFBLc
xXxDragonSlayerxXx	7rYFQIowpW	mq5rFL6JzF
420_E-Sports_Masta	cQF4DdSFfn	S8e0zpATNR

Data Quality

- Quality is not only about cleanness
- Quality may also involve significance
  - Are certain groups large enough to draw meaningful aggregates?
  - If my data is a sample of a population, does it accurately depict that population?
  - Did I ask the right kinds of questions to get good data?

Even Affects Machine Learning

- Training data □ Prediction program
  - Prediction program believes that the training data is representative of a population and covers all cases
  - If there's bias in the training data, it will affect the model

Takeaways

- Be good stewards of the data you have
- There's more to data management than the technical bits