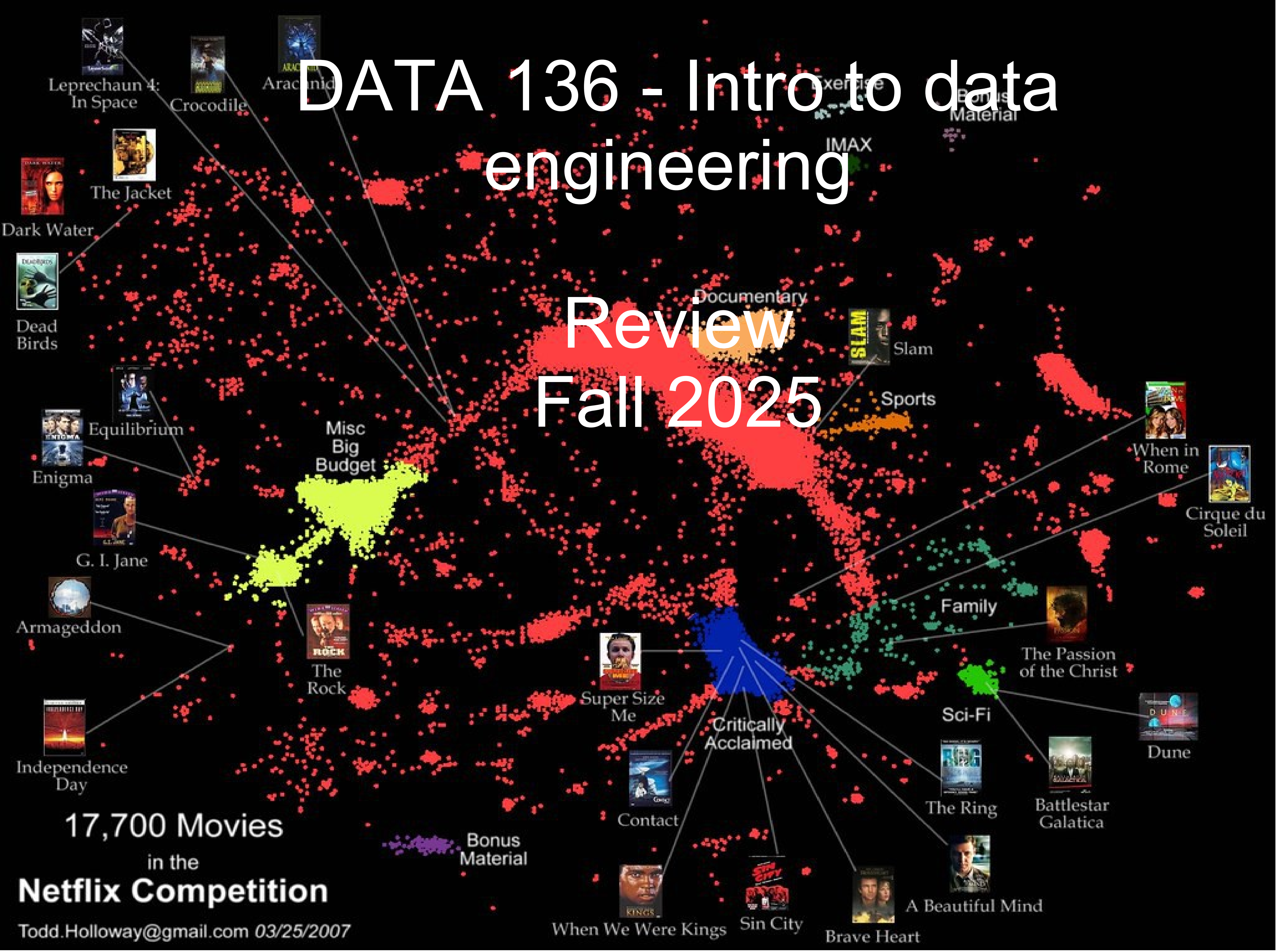
DATA 136 - Intro to data engineering

Review Fall 2025



17,700 Movies in the Netflix Competition

Todd.Holloway@gmail.com *03/25/2007*

Administrivia

- HW7 due Thursday Dec 11
- Early exam
  - Mon Dec 8 1pm-3:00 Rosenwald 011
- Final Exam
  - Friday Dec 12   7am-9am Kent 120

The exams will be designed to be equivalent, about 25 questions, 110 minutes.

If you take the exam twice, I'll be happy to assign you the lower of the two grades.

The three resources

$45/month
for
~30Mbits/s

$1500 /
3 years
$30/month

$0.02 / Gb
/ 5 years?

Normies start looking for engineers in hoodies when they run into resource limitations

Testing

- Automatically-run tests that run a part of the code and compare results against KNOWN CORRECT ANSWER.
- Tests are sometimes harder to write than implementation, but are **designed to be inexpensive to run.**
- Tests might be more specific than (or different from 4   the specification)
- Jargon: data used in tests called "test fixtures"
- Jargon: fraction of lines of code run during testing is "test coverage"
- **Q: Do we really trust our code?**

Testing best practices

- You want your tests to help you fix the problem.
- Include purpose of test + data specific to the failure, if possible in the fail message.
- Deterministic tests are preferred over randomized ones.
- Many little tests or one big test? A test with four asserts can only fail once.

A QA engineer walks into a bar and orders a beer.

She orders 2 beers.

She orders 0 beers.

She orders -1 beers.

She orders a lizard.

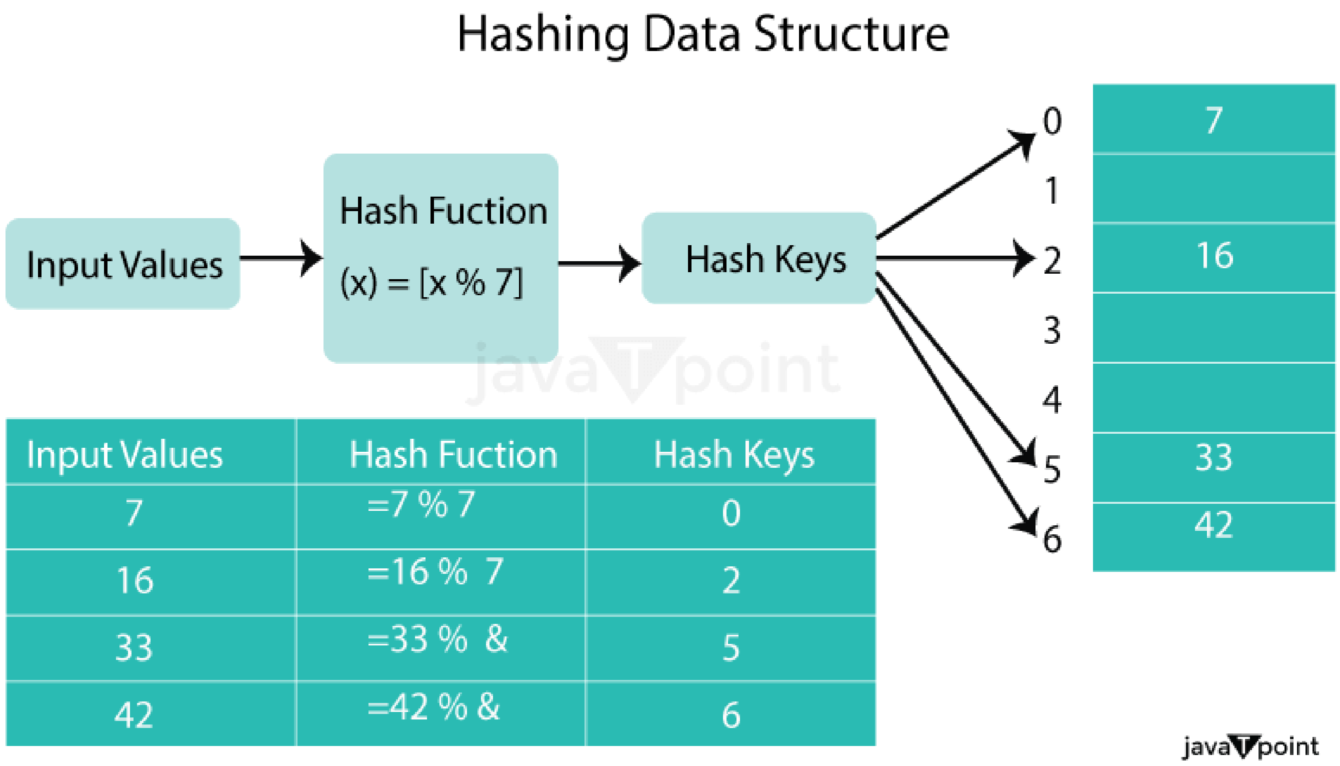
She tries to leave without paying.

Satisfied, she declares the bar ready for business. The first customer comes in an orders a beer. They finish their drink, and then ask where the bathroom is.

The bar explodes.

Some definitions

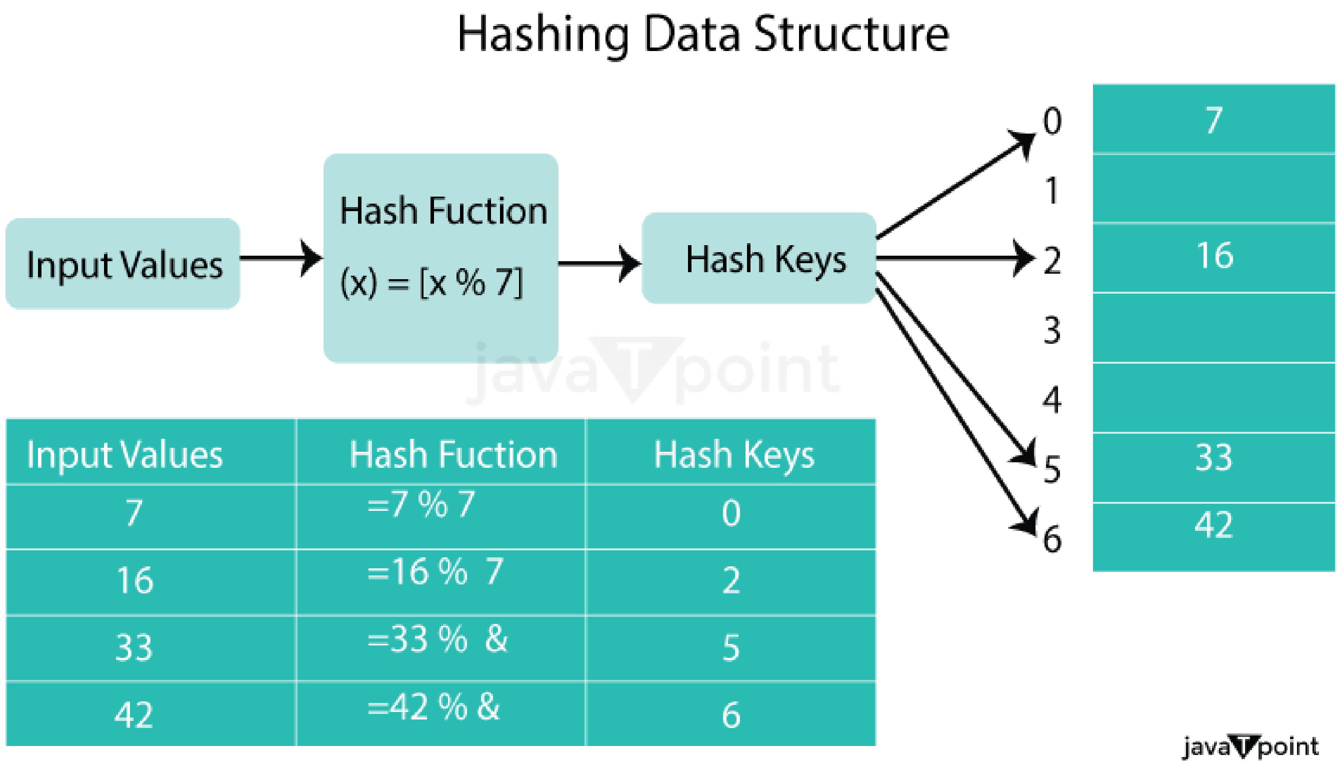
- A **data model** defines how different pieces of digital data are organized, defines how they relate to one another, and how these representations correspond to the properties of real-world entities.
- A **data dictionary** is essentially the same thing, but packaged as metadata.
- **Data Description Language** is the subset of SQL (or Django in models.py) that defines the database schema. CREATE, DROP, ALTER
- **Data Manipulation Language** adds, retrieves, changes data. INSERT, SELECT, DELETE...



Input Values	Hash Fuction	Hash Keys
7	=7 % 7	0
16	=16 % 7	2
33	=33 % &	5
42	=42 % &	6

0	7
1	
2	16
3	
4	
5	33
6	42

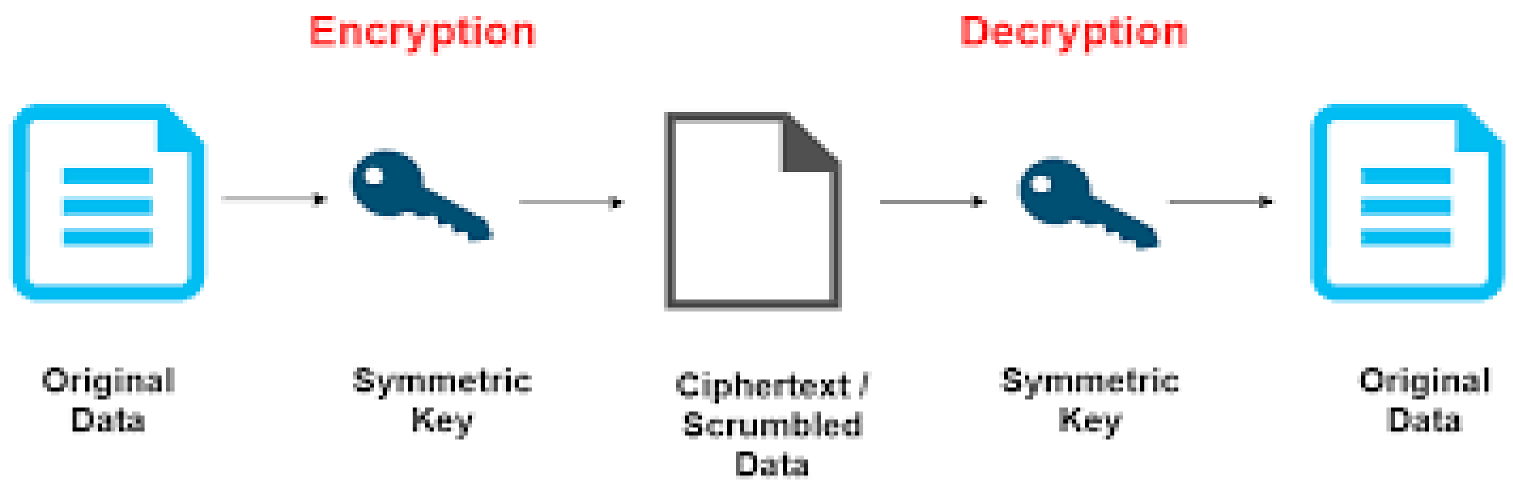
- Hashing for storage depends on hash functions, functions that take data and return a fixed-length number.
- Most commonly multiplication and modular division (using numbers vetted by cs-math people)
- Address depends deterministically on data


Hashing Data Structure

Input Values	Hash Fuction	Hash Keys
7	=7 % 7	0
16	=16 % 7	2
33	=33 % &	5
42	=42 % &	6

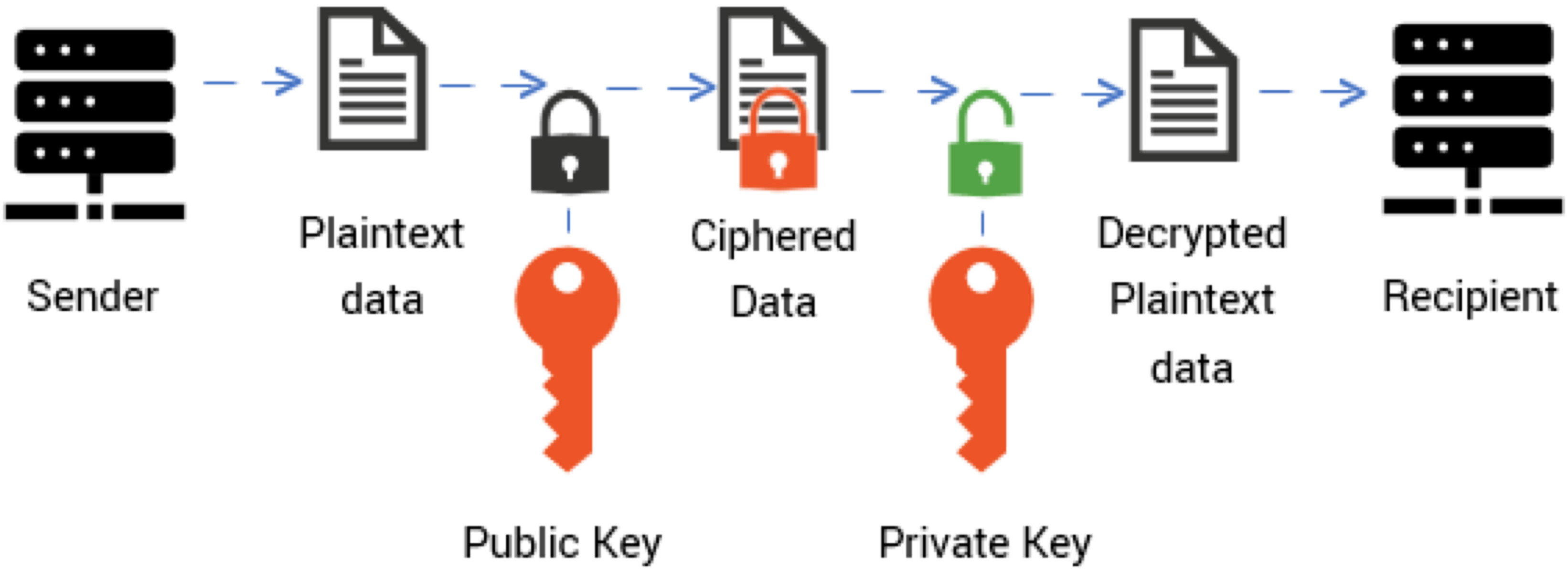
Because collisions are possible and expected in hash datastructures, the objects stored are allowed to contain multiple values (jargon: chains).

- Hash table needs lots of space to not become too crowded to work well.
- Constant-time lookup (!!!!)



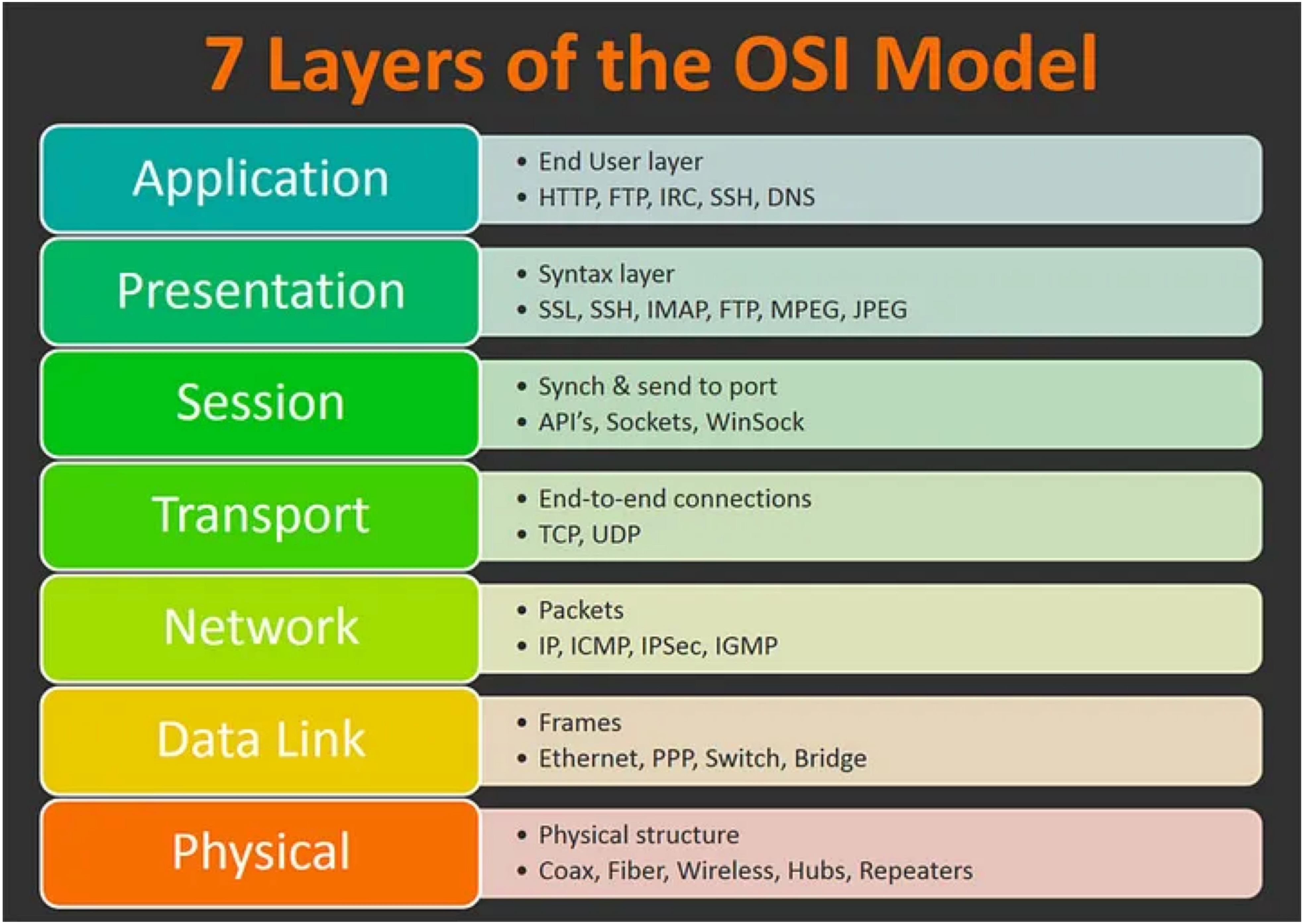
- Symmetric encryption: in use since the telegraph era
- Easy to compute, hard to transfer keys safely
- Built into SSH layer; "session keys" used for a brief time; Public key is used to distribute session keys
- Symmetric encryption for binary data uses XOR
- If the key is long enough, not stolen or re-used, can't be cracked by brute-force effort

Public Key Encryption (Asymmetric)



- PKE has essentially two applications:
  - secret communication (encrypt with public key, decrypt with private key)
  - integrity assurance / signing (encrypt a hash with private key, published public key ensures everyone that you have ownership of private key.)
- We trust PKE because certain math problems (factoring) are really hard. Can be cracked, but it is too expen$ive.

Open Systems Interconnection



7 Layers of the OSI Model

Application	• End User layer • HTTP, FTP, IRC, SSH, DNS
Presentation	• Syntax layer • SSL, SSH, IMAP, FTP, MPEG, JPEG
Session	• Synch & send to port • API's, Sockets, WinSock
Transport	• End-to-end connections • TCP, UDP
Network	• Packets • IP, ICMP, IPSec, IGMP
Data Link	• Frames • Ethernet, PPP, Switch, Bridge
Physical	• Physical structure • Coax, Fiber, Wireless, Hubs, Repeaters

Data models

- Logical data model: A model that describes the semantics of the data, as represented by the particular data analysis technology. We always worry about this if we don't want to embarrass ourselves.
- Physical data model : describes the physical means by which data are stored. This is concerned with partitions, CPUs, bits, data types, and so on. **We don't worry about this if it works / if the problem is easy.**

Metadata

- Data about data.
- Usually includes the names of the authors, among other things.
- **Does not fit in the schema for the data** itself. This is its defining feature. This makes it a universally-reviled aspect of data management; it doesn't fit in the box!
- Can include SEO terms to prevent the data from being forgotten.

Application Programming Interface

- Code that does stuff
- API codifies communication between the user or program using the API and the engine producing the API results.
- Documentation and examples show you what inputs are required, permitted, how can the behavior of the program be modified.
- We can talk about bash, pandas, pytorch, amazon S3, instagram, podcast publishers...
- If you aren't running the server running the API, usually not free -> authentication

Management tools: increase complexity in exchange for lower maintenance cost

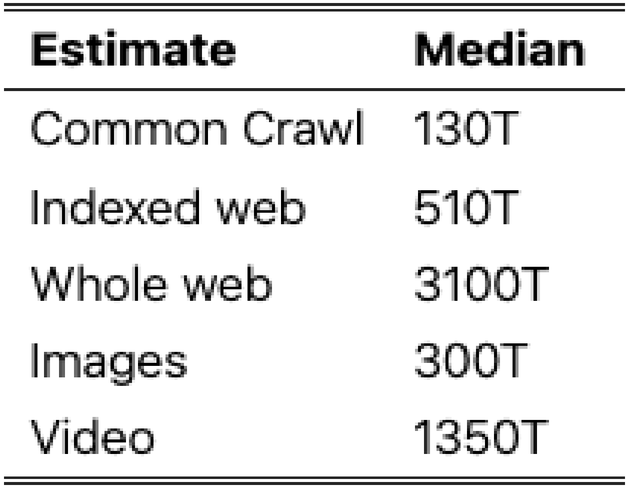
- Version control (!!)
- APIs / service-oriented architecture
  - lower performance, easier to maintain
- Containers
  - lower performance, high repeatability, portability
- Virtualization
  - lower performance, can be scaled and billed fractionally flexible (hence name "elastic compute")
- Test-driven development
  - writing tests is hard

Tradeoffs tradeoffs tradeoffs

Space, time, error rate, choose two.

• Indexing	+space –time
• Distributed computing	+space –time
• Hash tables	+space –time
• MINHASH/Bloom filters	+error –time
• Compression	+time –space
• Cacheing	+space –time
• Sampling	+error –time

Step 1: download the internet



Estimate	Median
Common Crawl	130T
Indexed web	510T
Whole web	3100T
Images	300T
Video	1350T

Estimates of the stock of data on the web in tokens.

Villalobos 2024
arXiv:2211.04325v2

Only 5 B words in wikipedia

40 M books in google books

10^{12} words in paperbacks orderable from Amazon

There's a lot of engineering on curating the training data

Feedback?