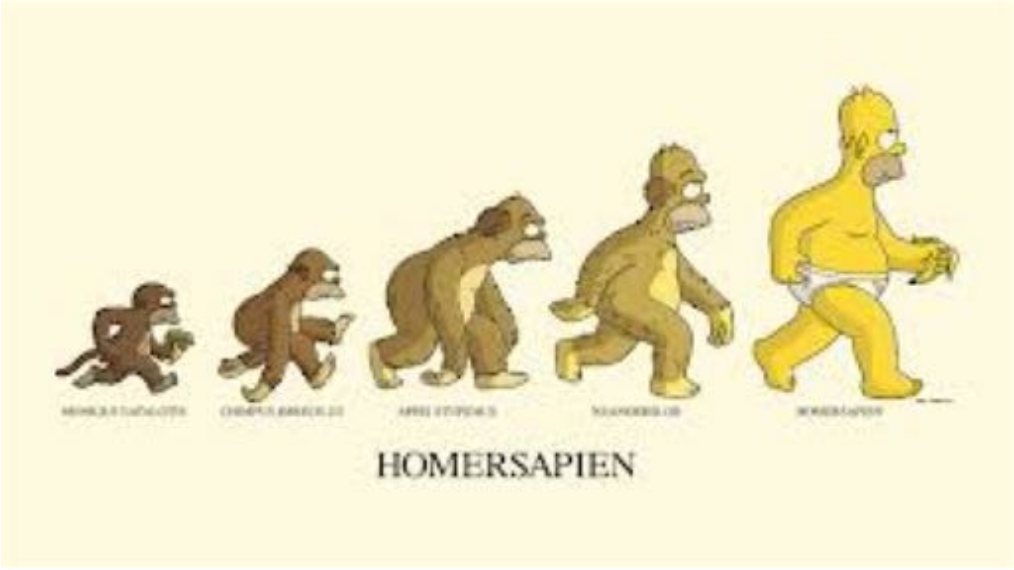
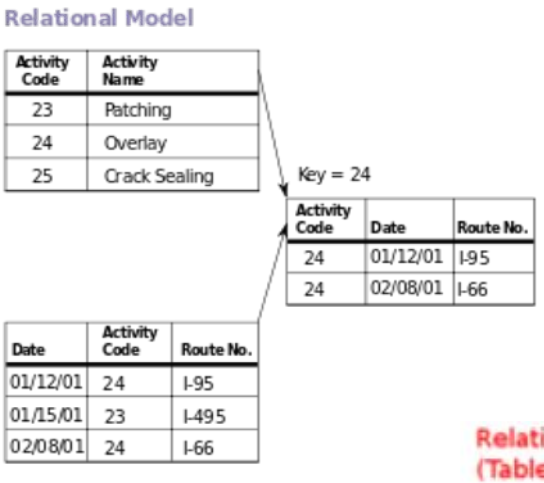
Lecture 25 Overview

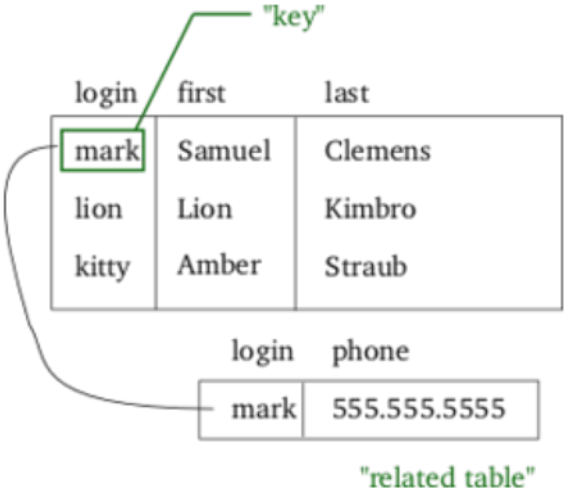
- Last Lecture
  - Query optimisation/query execution strategies
- This Lecture
  - Non-relational data models
  - Source: web pages, textbook chapters 20-22
- Next Lecture
  - Revision



Review - Relational Model Concepts

- First formulated and proposed in 1969 by Edgar F. Codd
- A database is a collection of *relations*
- Usually referred to as *tables*





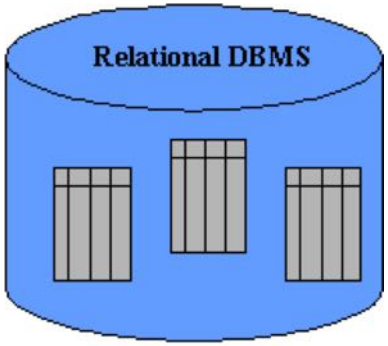
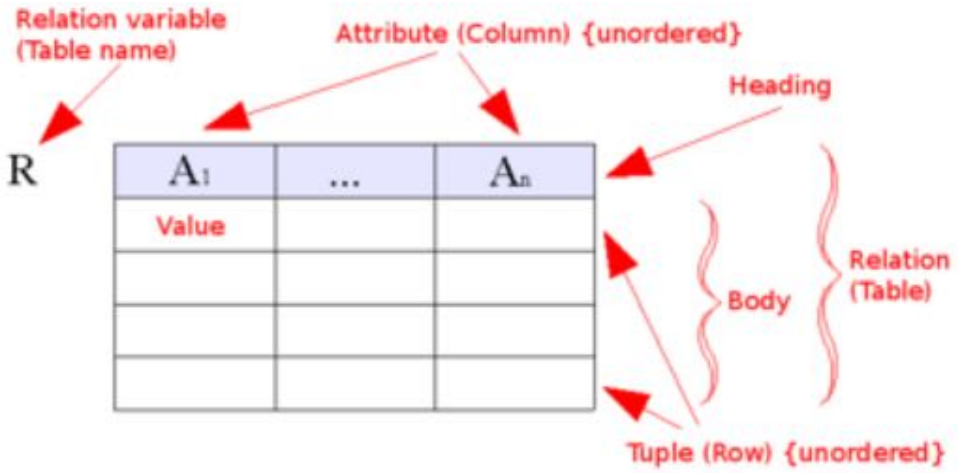
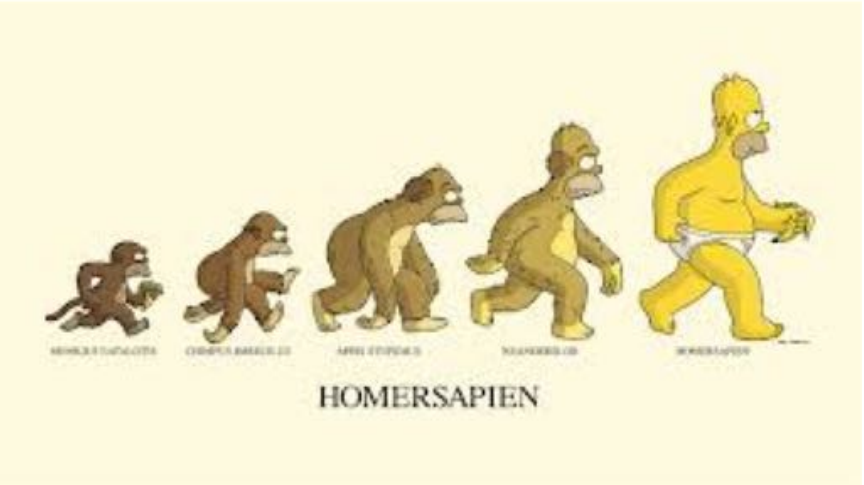


Figure 1: Relational Model Structure



Review - Relational Model Concepts

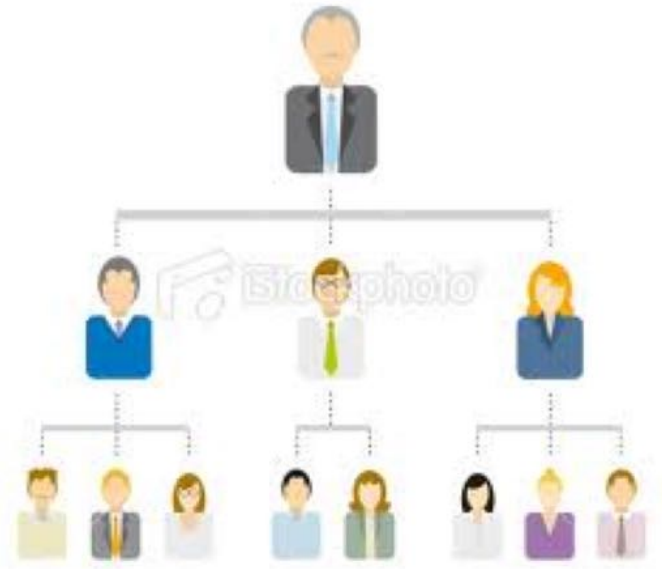
- Other models before and after Relational Model Concepts?



History of Database:
https://www.youtube.com/watch?annotation_id=annotation_3070369099&feature=iv&src_vid=qUV2j3XBRHc&v=FR4QIeZaPeM

The Hierarchical Model

- Began as IBM's database program IMS - Information Management System, introduced in the late 1960s
  - Organize data into a tree-like structure
  - records and parent-child relationships
  - 1:N relationship



Hierarchical schemas

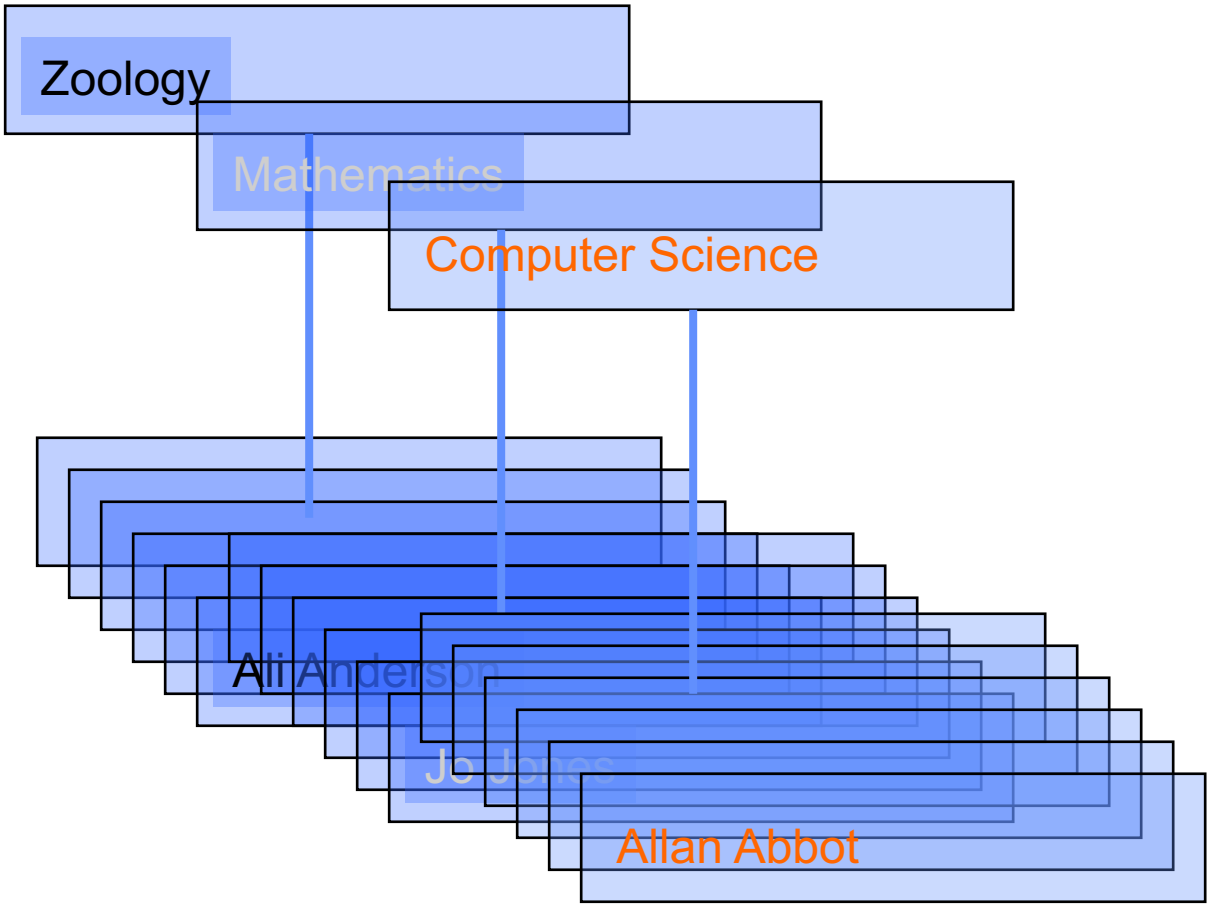
Department

Dname ....

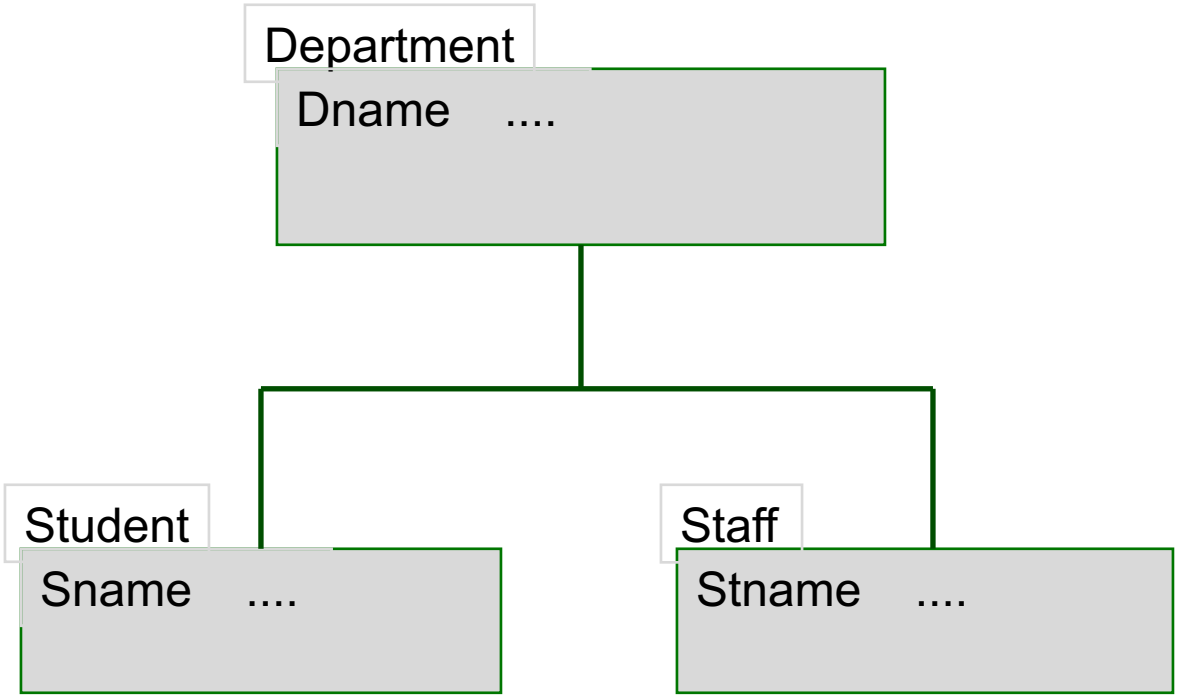
Student

Sname ....

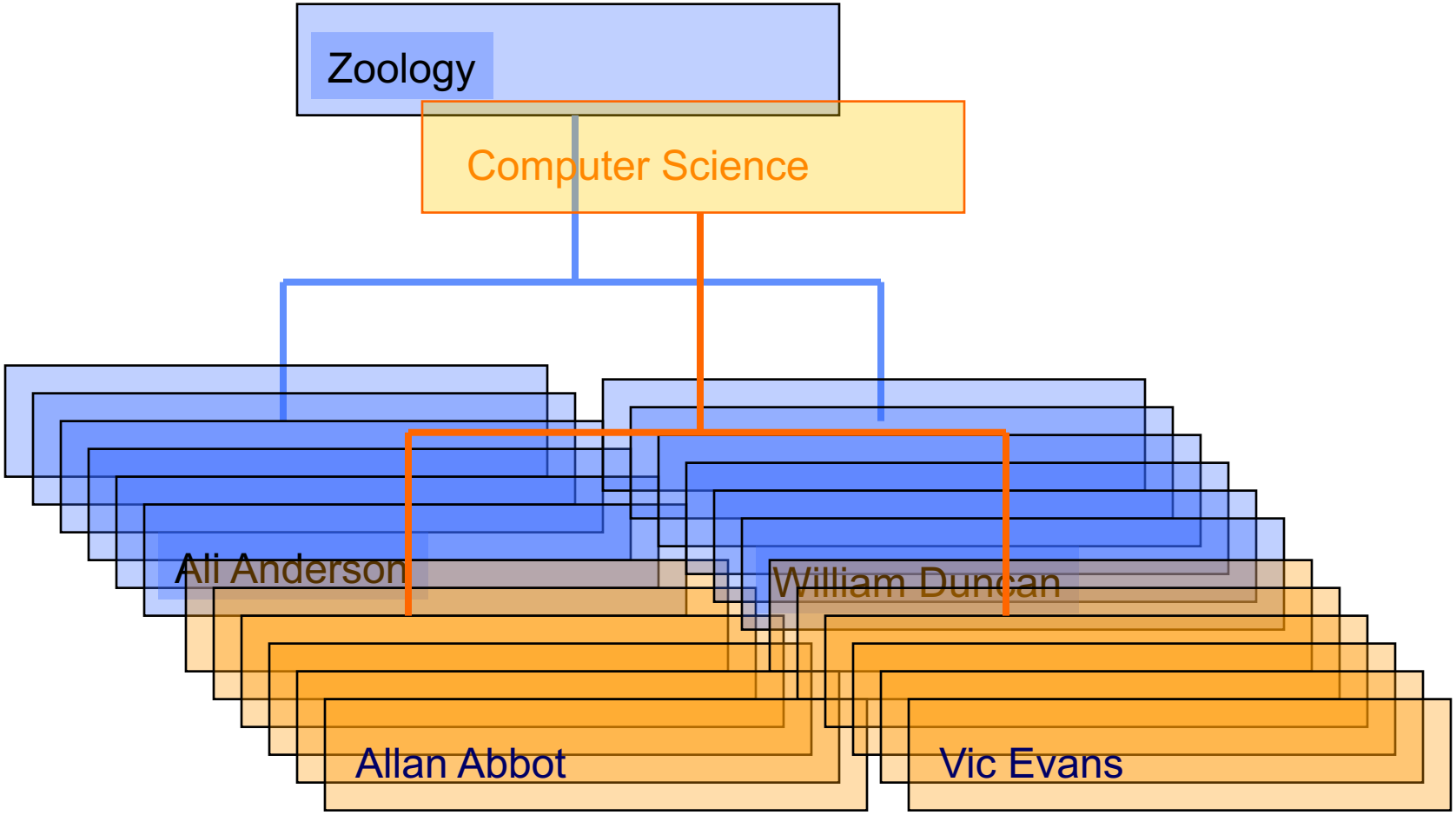
Sample data as seen by the model



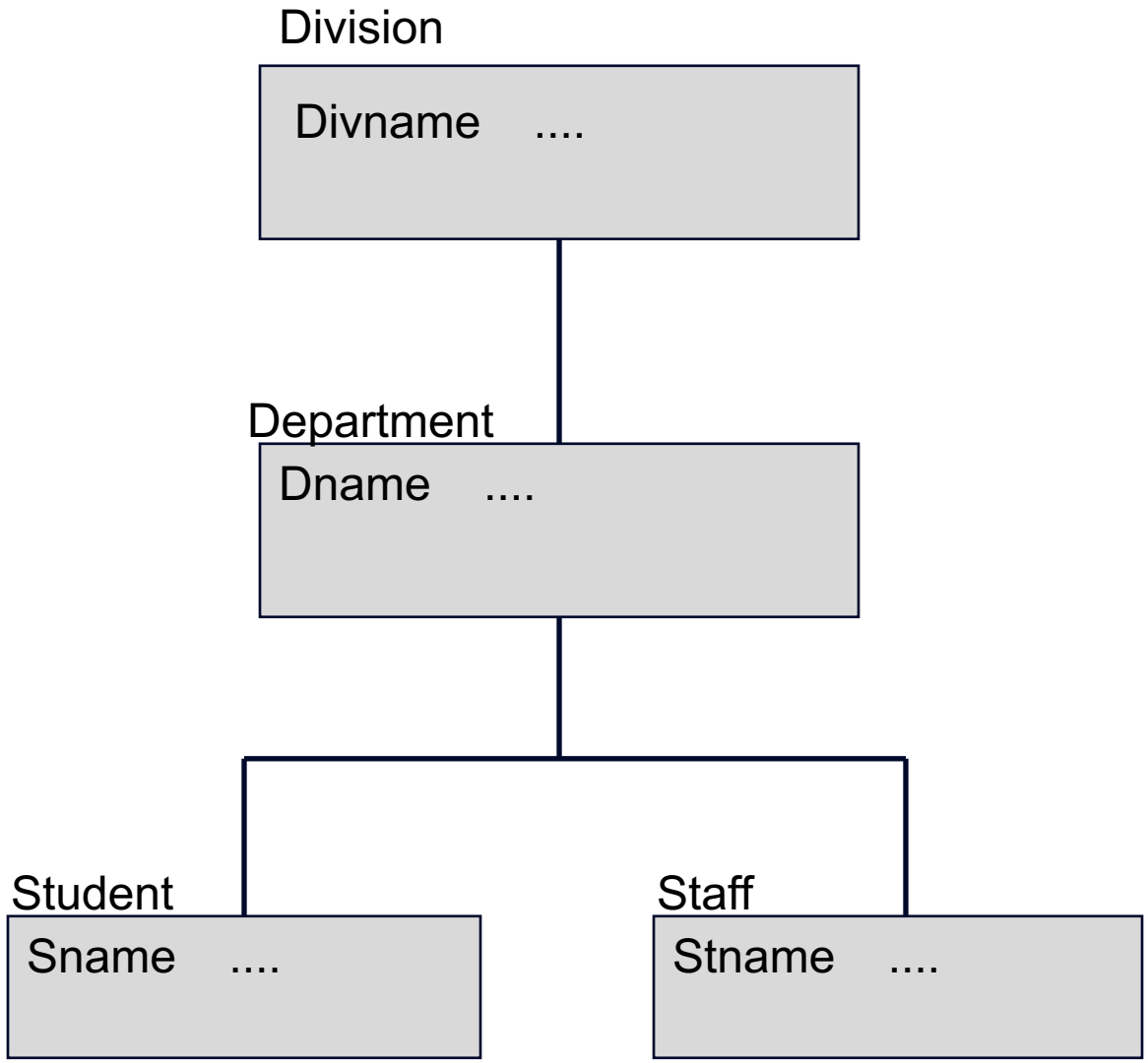
Hierarchical schemas



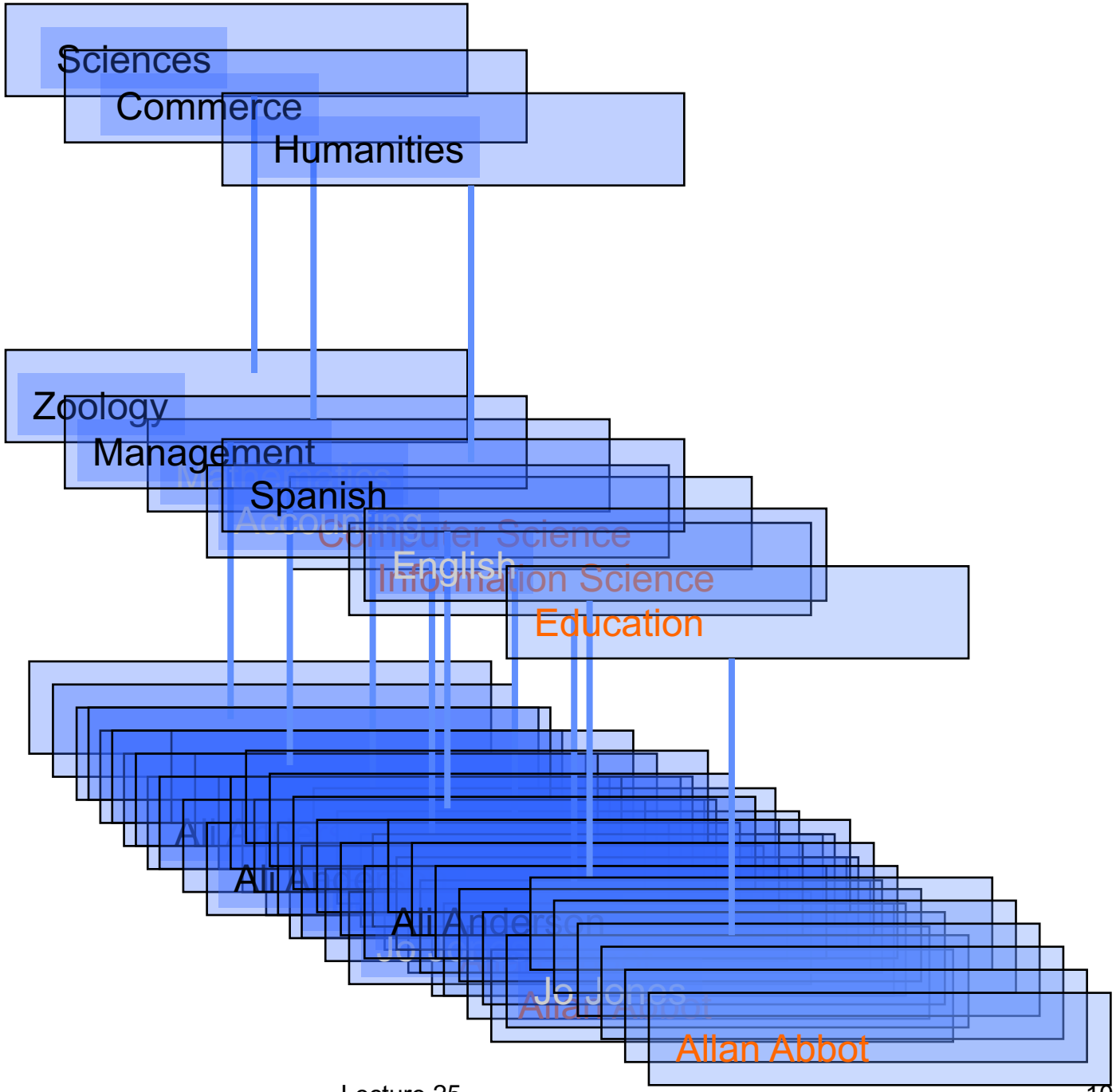
Sample data as seen by the model



Hierarchical schemas

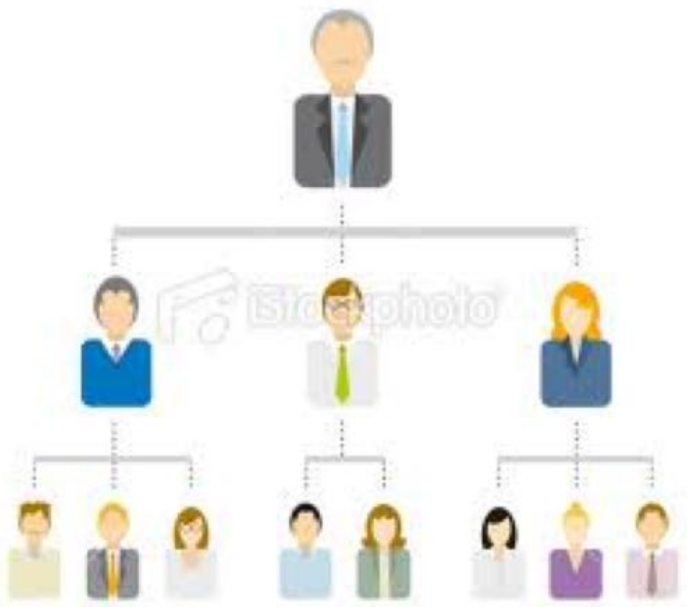


Sample data as seen by the model



Hierarchical comments

- It needs to be designed, defined and then built
- It is difficult to extend or alter
- It lacks the relational flexibility where tables can be added and dropped



The Network model

- The network database model was invented by Charles Bachman in 1969 as an enhancement of the hierarchical database model. (“CODASYL model”)
- An attempt to standardise
- the data model objects are records and sets
- record - usual meaning, a group of related data values
- set - a description of a 1:N relationship between two record types
- sets have names, owners and members and are represented using Bachman diagrams

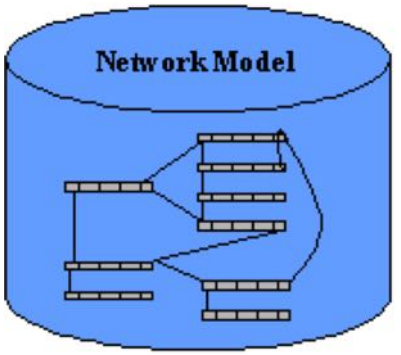
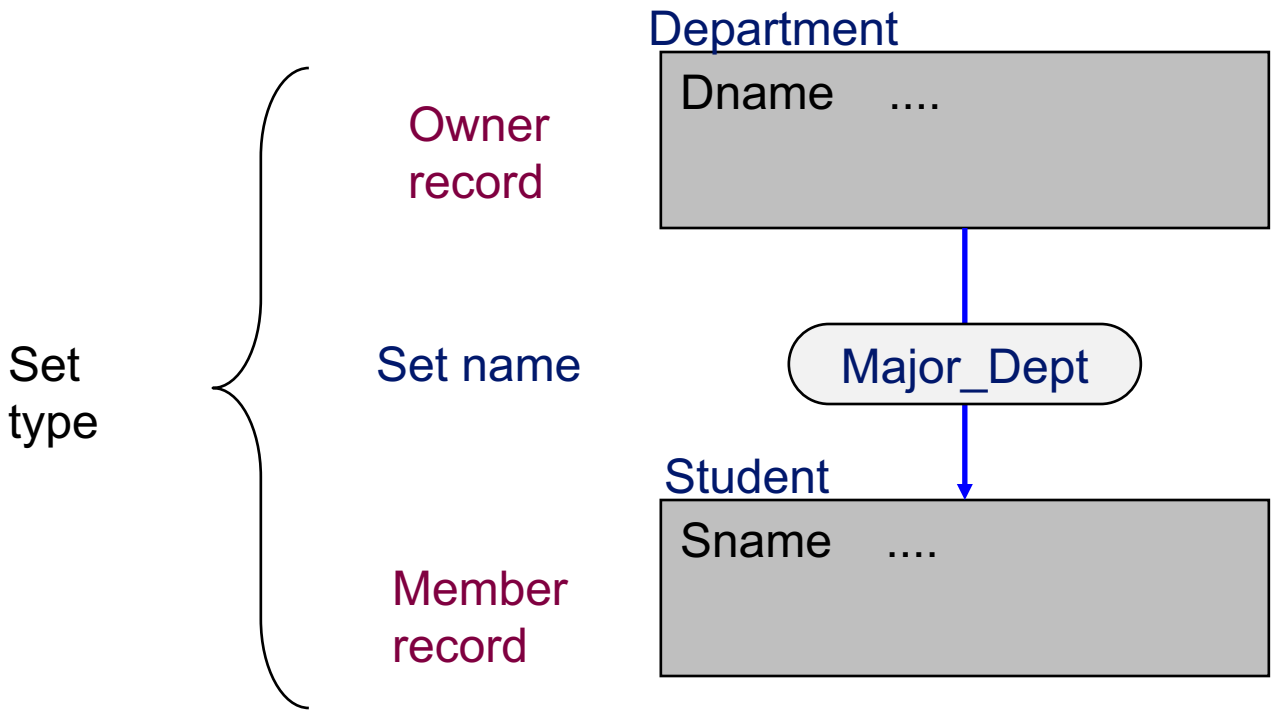


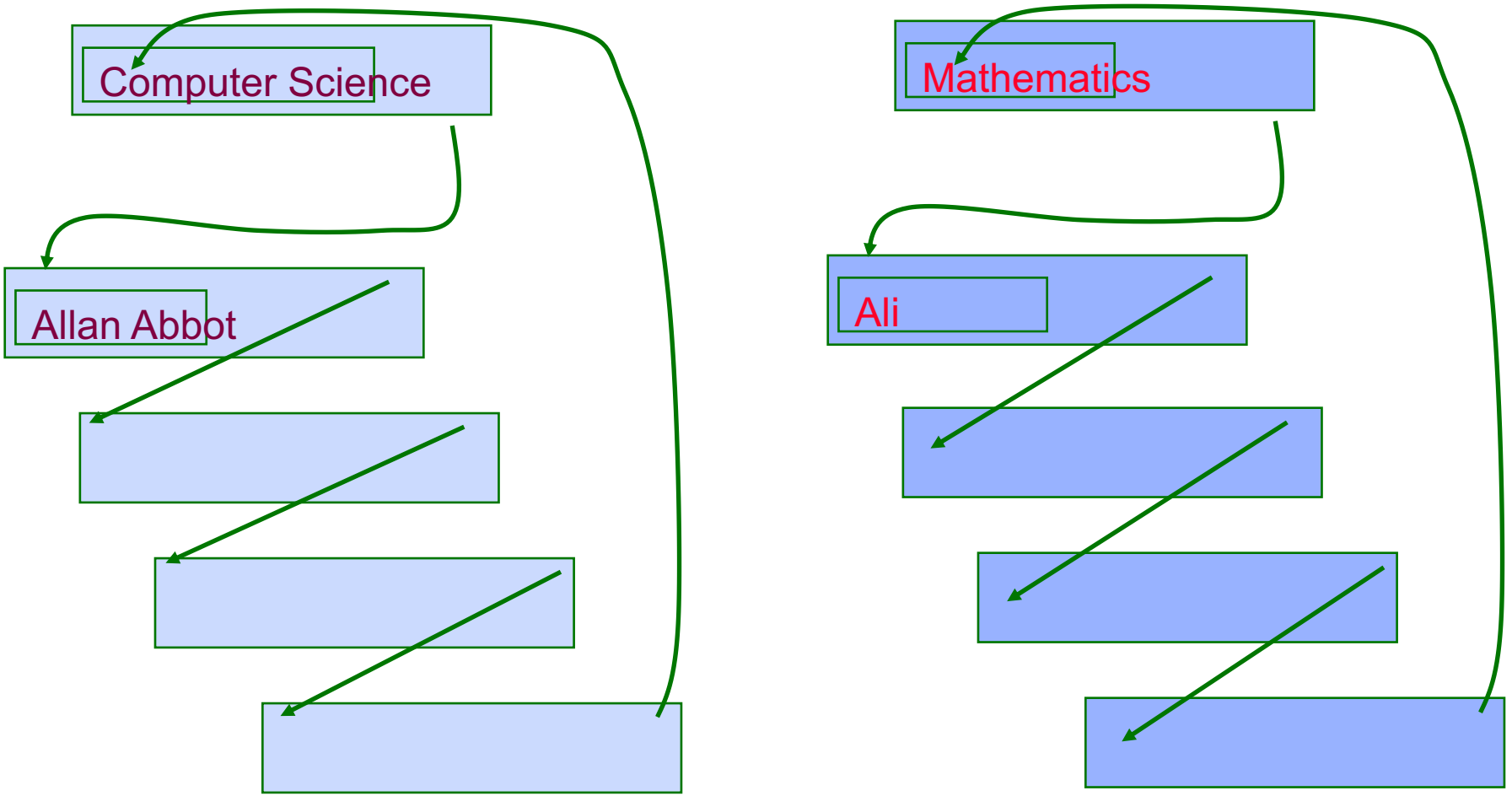
Figure 1: Network Model Structure

Bachman diagrams

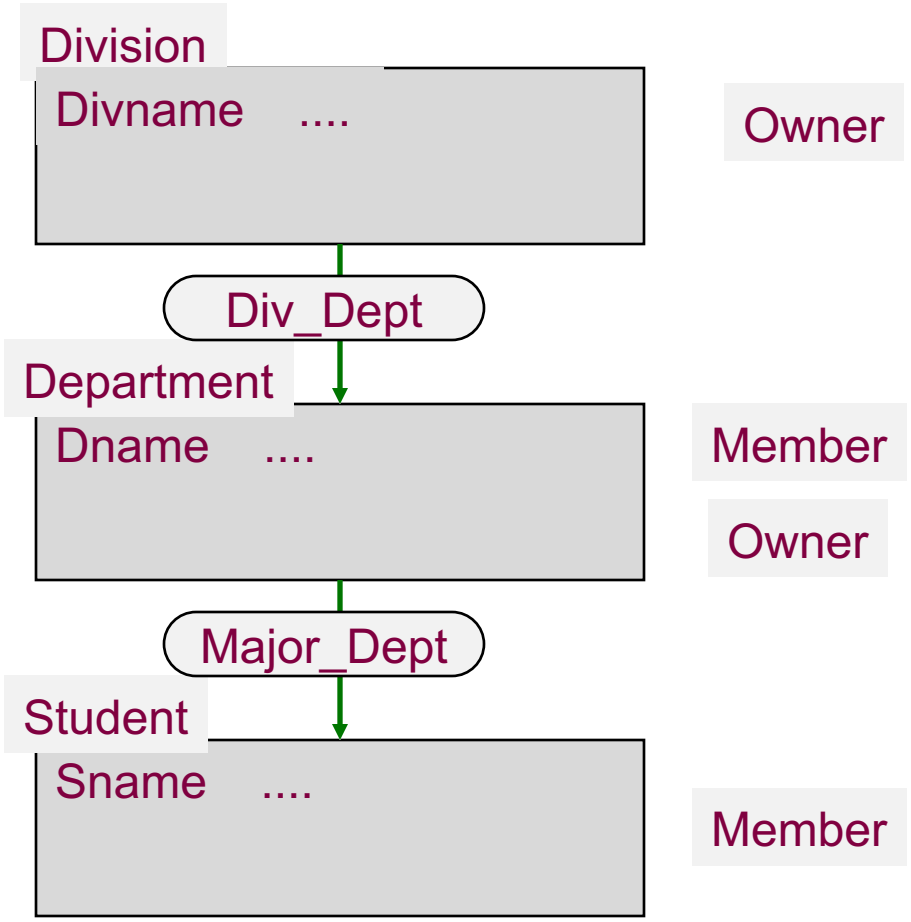


Sample data as seen by the model

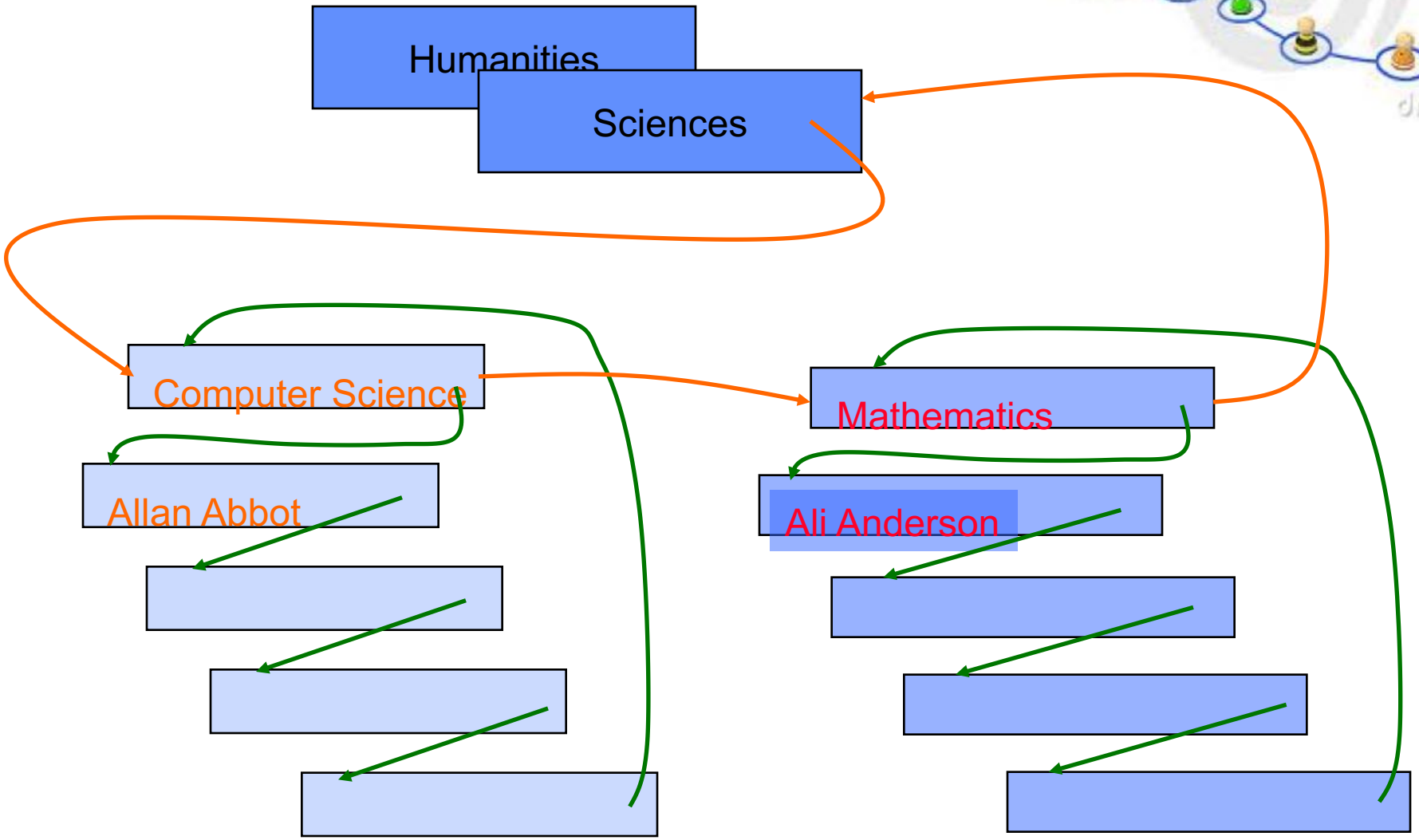
Ring (circular linked list) linking the owner record and all member records.



Bachman diagrams



Sample data as seen by the model



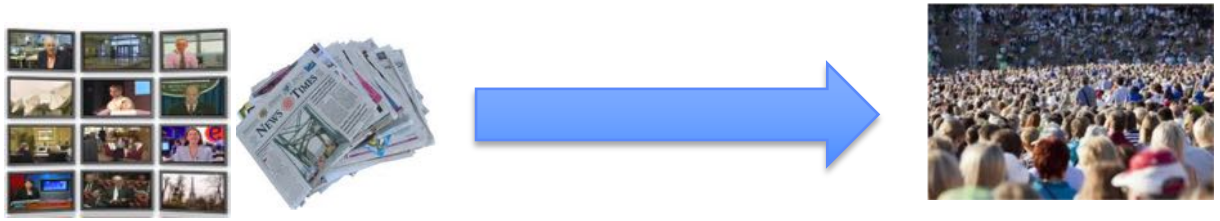
Introduction to Big Data

What is Big Data?
What makes data, "Big"?

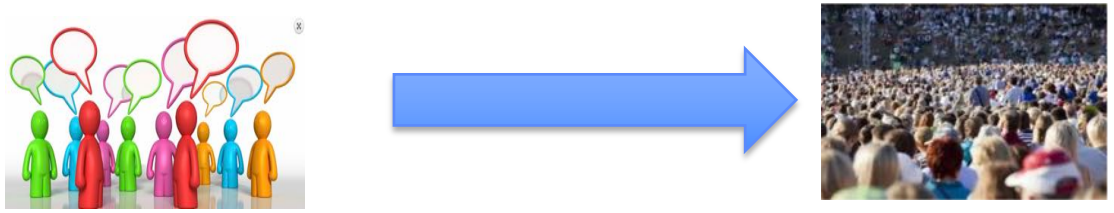
The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

Old Model: Few companies are generating data, all others are consuming data



New Model: all of us are generating data, and all of us are consuming data



Big Data Sources

Social Networking and Media
- Facebook
- Twitter
- LinkedIn
- Blogs
- Site Comments

Mobile Devices
- Call
- Text/IM
- Location
- In-App Activity

Big Data Sources

Internet Transactions
- Purchases (automated or manual)
- Banking Activity
- Investment Activity

Networked Devices/Sensors
- Internet Connected Hardware
- Sensors (Temperature, Movement, Pressure, Humidity)
- Beacon Interactions

ebay

trademe
Where Kiwis Buy and Sell

Who's Generating Big Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)

Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

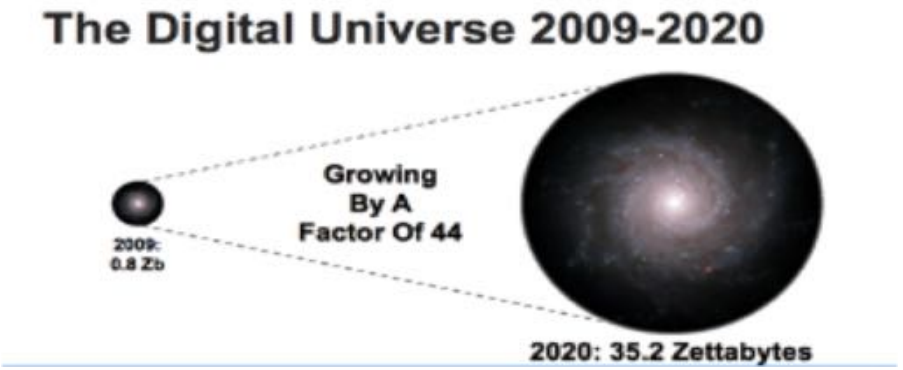
Big Data Definition

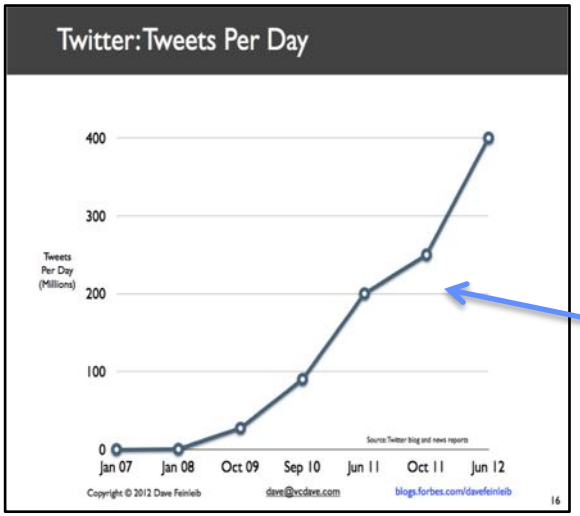
- No single standard definition…

"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

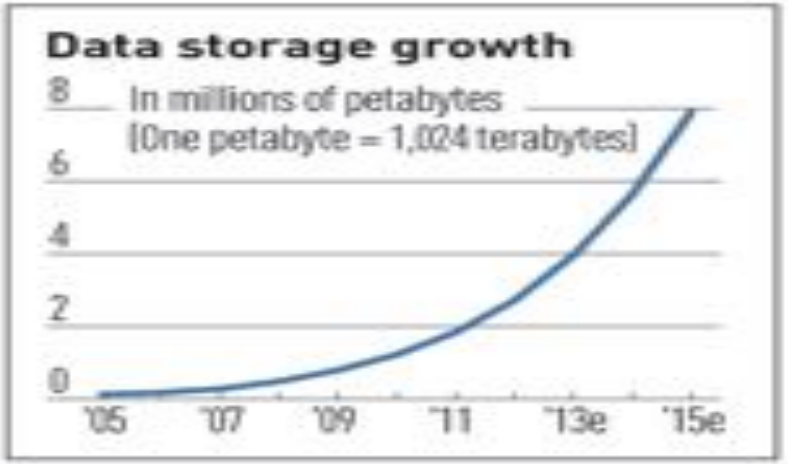
Characteristics of Big Data: 1-Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 2020
- Data volume is increasing exponentially



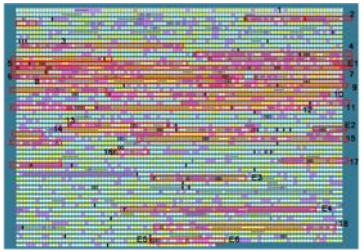


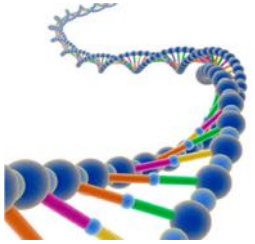
Exponential increase in collected/generated data

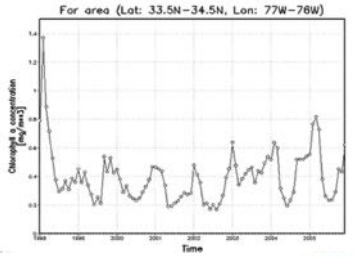


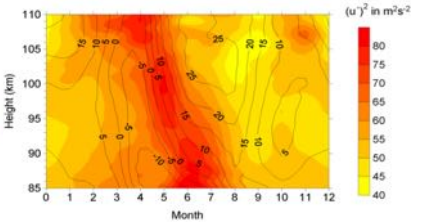
Characteristics of Big Data: 2-Complexity (Varity)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- A single application can be generating/collecting many types of data







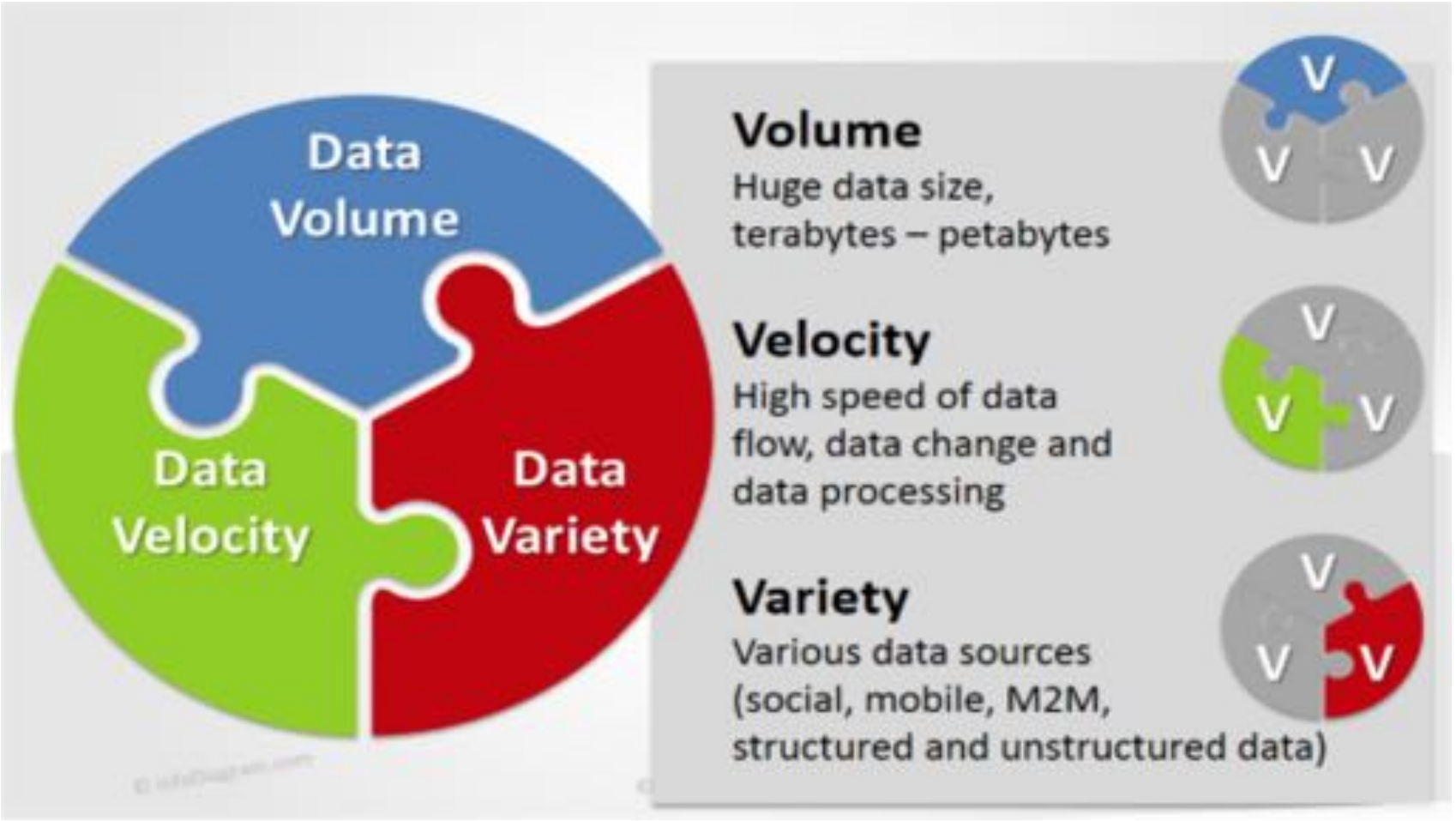


To extract knowledge➔ all these types of data need to linked together

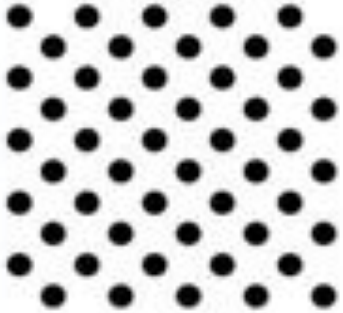
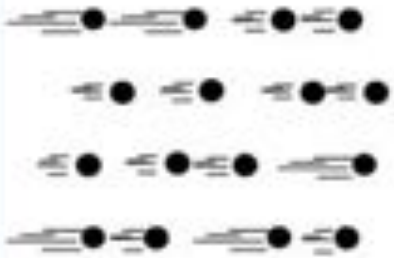
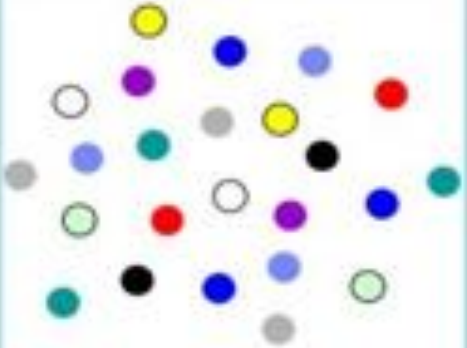
Characteristics of Big Data:
3-Speed (Velocity)

- Data is begin generated fast and need to be processed fast
- Late decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Big Data Characteristics



Some Make it 4V's

Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

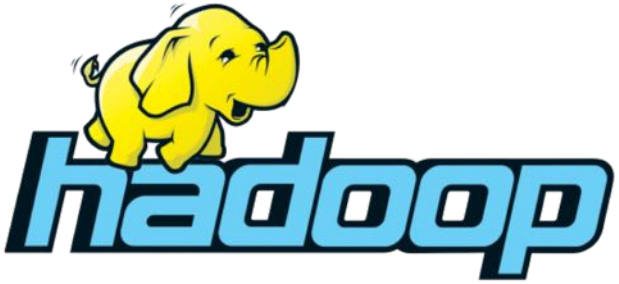
NoSQL

- Stands for **N**ot **O**nly **S**QL
- Looser schema definition
  - Usually do not require a fixed table schema
- Designed to handle distributed, large databases
- No joins

It's not a replacement for a RDBMS but compliments it.

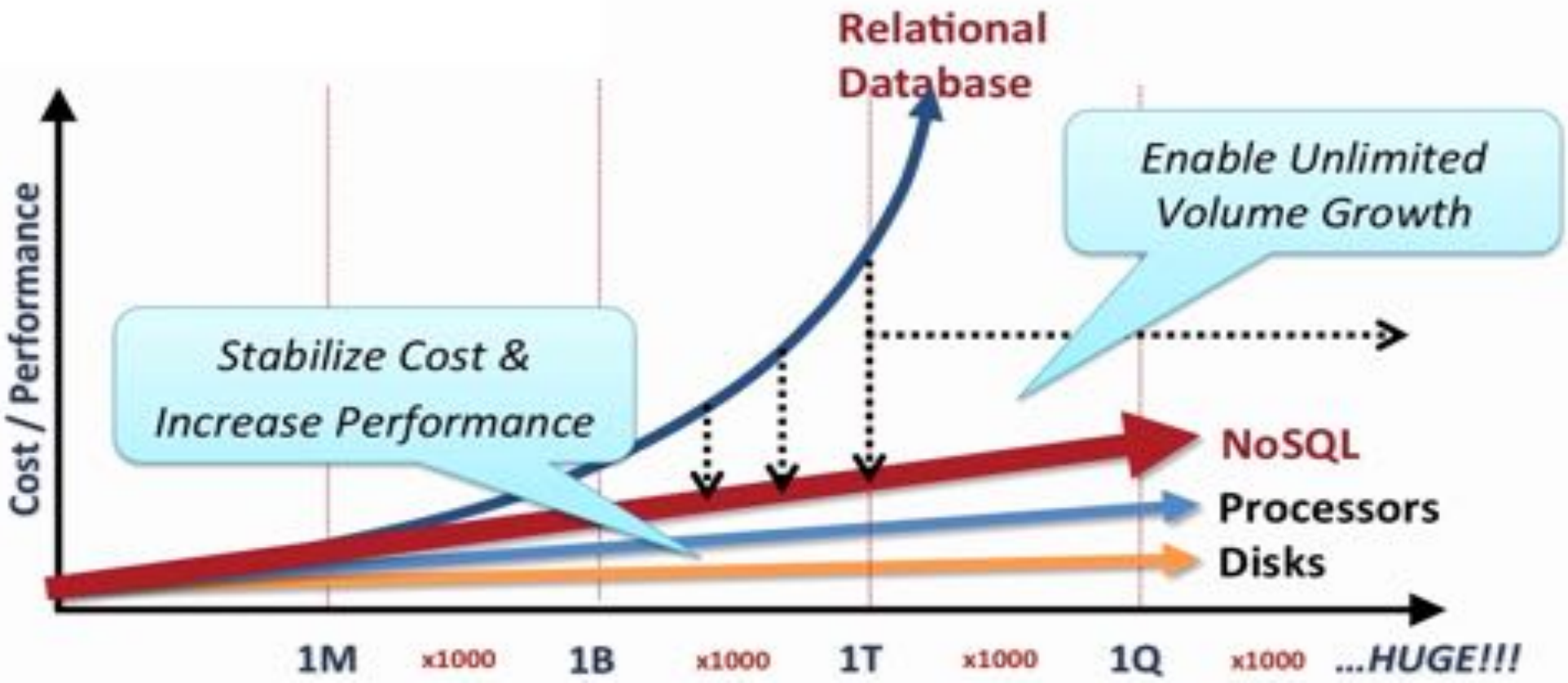
Where NoSQL is Used?

- Google (BigTable, Cloud Datastore)
- LinkedIn (Espresso)
- Facebook (Cassandra)
- Twitter (Hadoop/Hbase, FlockDB, Cassandra)
- Netflix(Simple DB, Hadoop/Hbase)



NoSQL Scaling

- Scaling horizontally is possible with NoSQL
- Scaling up / down is easy



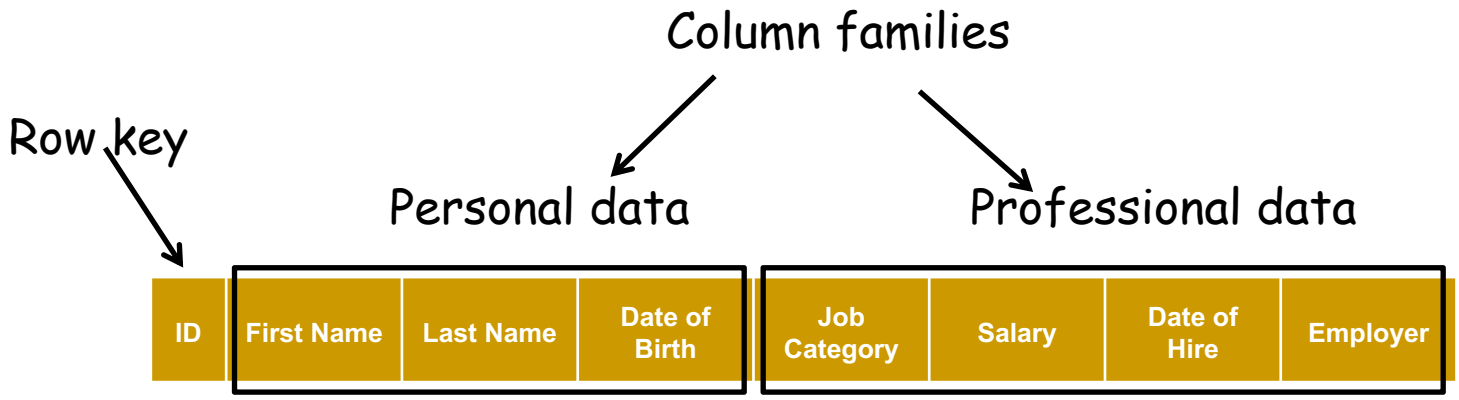
NoSQL Types

- Key-Value
  - Simple model, mapping between key and value
  - You cannot query without having the key
  - Easy to use, very scalable
  - Example: Redis (Widely used by Twitter and Flickr)
    https://www.youtube.com/watch?v=OG610oe_kxs

key	value
123	123 Main St.
126	(805) 477-3900

NoSQL Types

- Wide-Column
  - Similar to tables
  - Key -> (Column: Value)*
  - Example: Hbase, Cassandra (by Facebook)



NoSQL Types

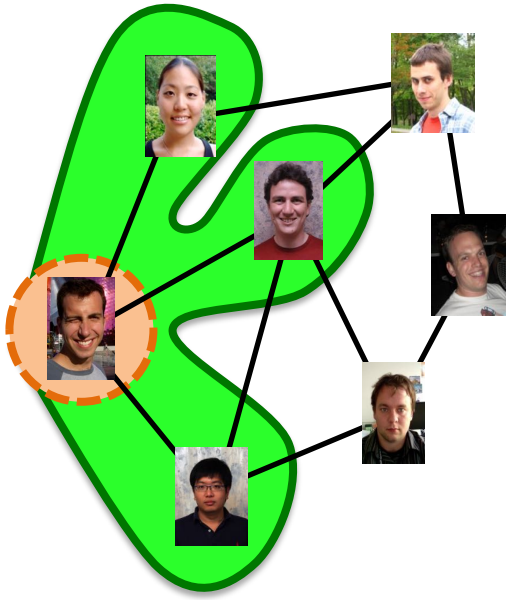
- Document
  - JSON (JavaScript Object Notation), XML
  - Use when data is semi-structured
  - Useful when the structure of data changes through the lifetime of application
  - Example: MongoDB

(https://www.youtube.com/watch?v=CvIr-2lMLsk)

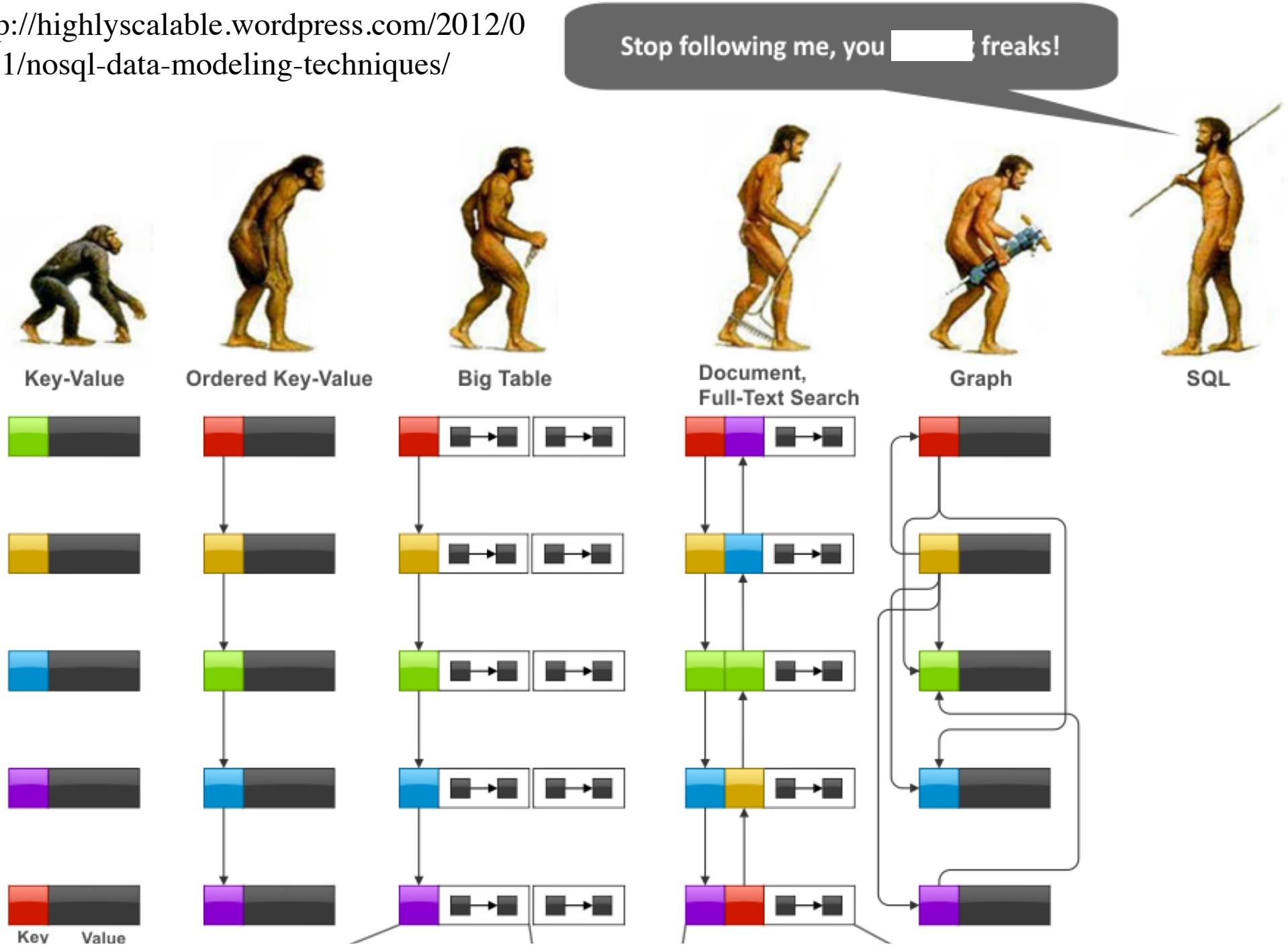
```
{ "user" :{
   "id": "124",
   "name": "Emmanuel",
   "addresses":[
     {"city": "Paris", "country": "France" },
     {"city": "Atlanta", "country": "USA" },
   ]
{
```

NoSQL Types

- Graph
  - Useful when objective is to quickly find connections, pattern and relationship between lots of data
  - Usage: Recommendations for product, social media
  - Example: FlockDB, Stardog



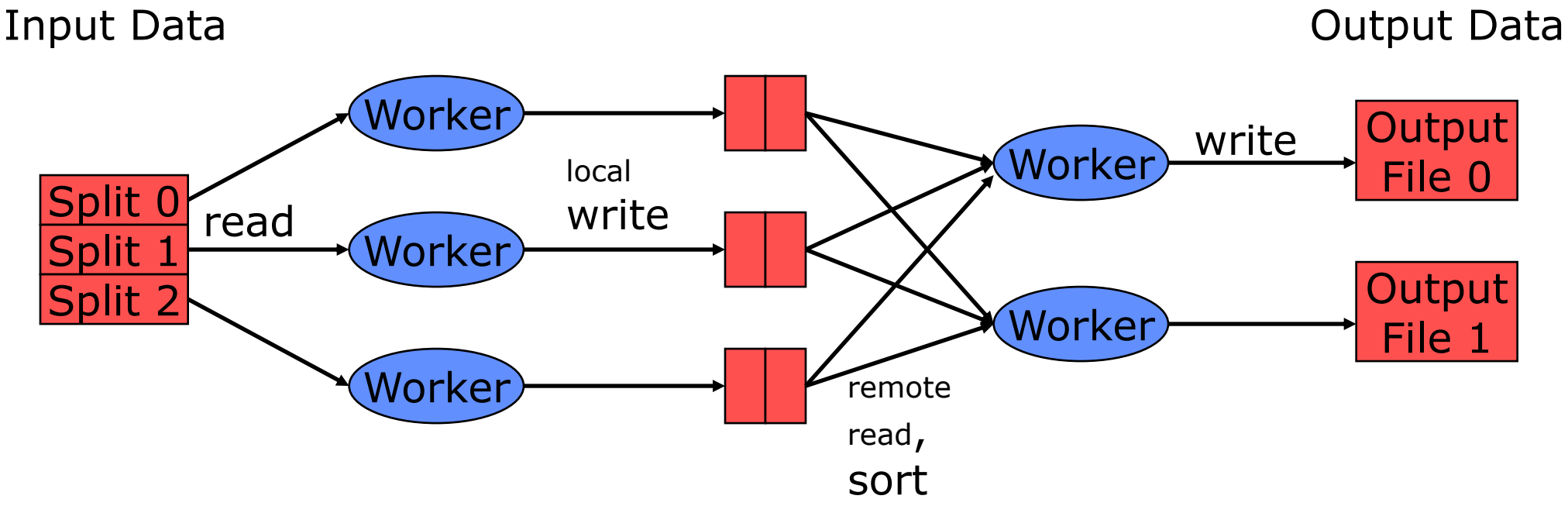
http://highlyscalable.wordpress.com/2012/03/01/nosql-data-modeling-techniques/



MapReduce Programming Model

- A single machine cannot serve all the data
- Need a distributed system to store and process **in parallel**
- How do you scale to more machines?
- **MapReduce** - programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster
- How does MapReduce work?
  - Read a lot of data
  - Map: extract something you care about from each record
  - Shuffle and Sort
  - Reduce: aggregate, summarize, filter, or transform
  - Write the results

MapReduce workflow



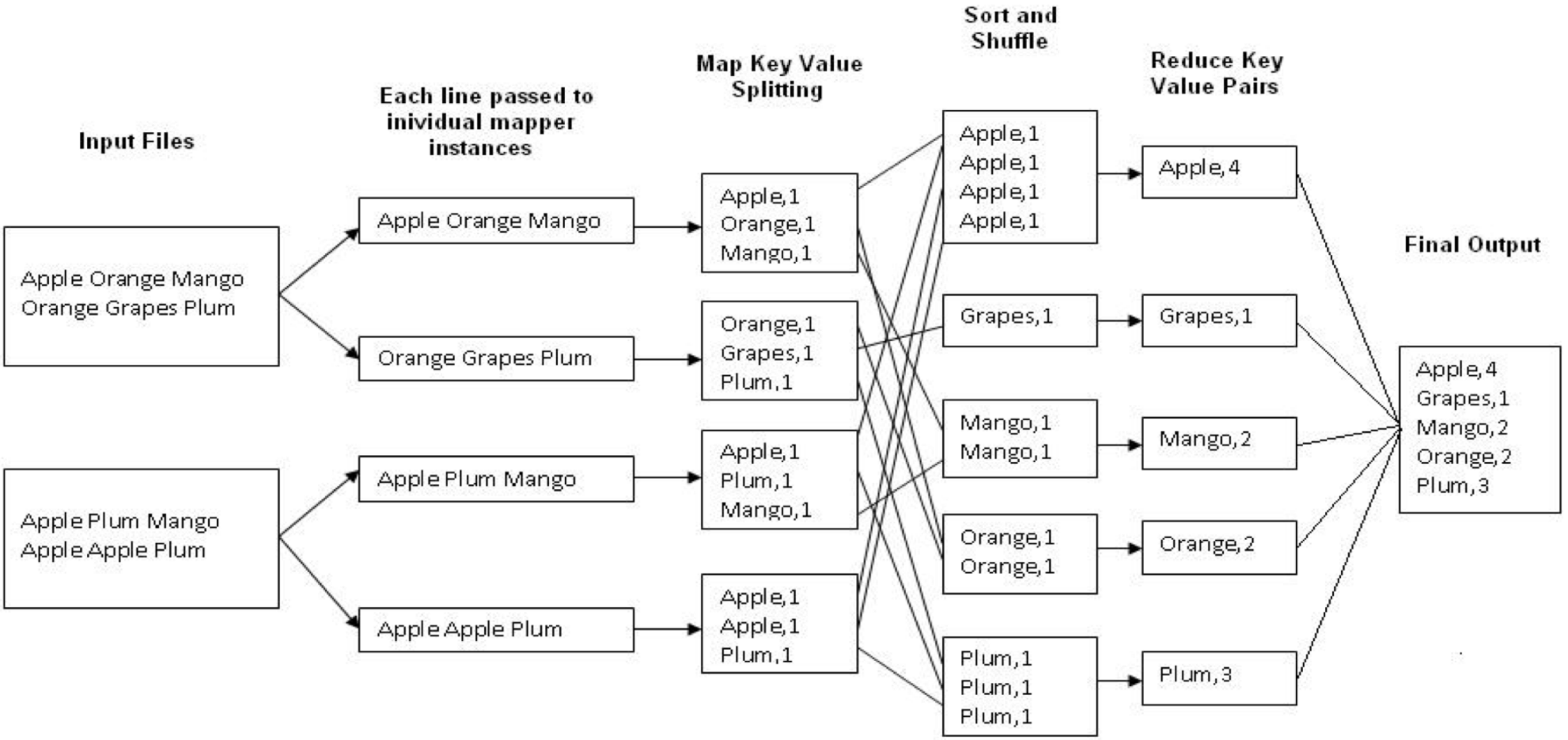
Map
extract something you care about from each record

Reduce aggregate, summarize, filter, or transform

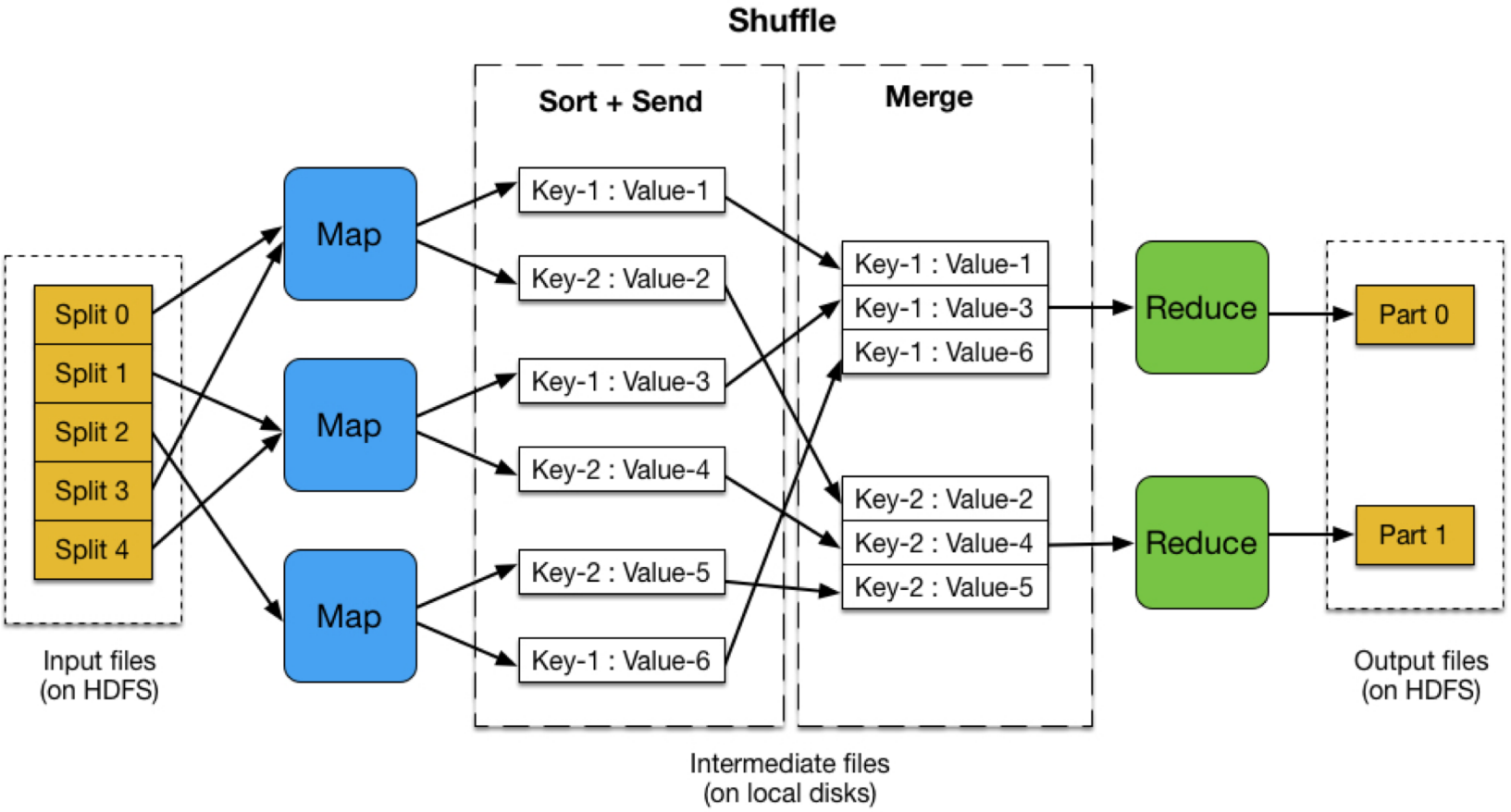
Example: Word Count

http://kickstarthadoop.blogspot.ca/2011/04/word-count-hadoop-map-reduce-example.html

Example: Word Count



MapReduce


Shuffle
Sort + Send
Merge
Split 0
Split 1
Split 2
Split 3
Split 4
Map
Map
Map
Key-1 : Value-1
Key-2 : Value-2
Key-1 : Value-3
Key-2 : Value-4
Key-2 : Value-5
Key-1 : Value-6
Key-1 : Value-1
Key-1 : Value-3
Key-1 : Value-6
Key-2 : Value-2
Key-2 : Value-4
Key-2 : Value-5
Reduce
Reduce
Part 0
Part 1
Input files
(on HDFS)
Intermediate files
(on local disks)
Output files
(on HDFS)

Exam

Review and exam guidance in next lecture

References

- NoSQL: Distributed and Scalable Non-Relational Database Systems,http://www.linux-mag.com/id/7579/
- What is NoSQL: https://www.youtube.com/watch?v=qUV2j3XBRHc
- NoSQL vs. SQL Summary, http://www.mongodb.com/nosql-explained
- NoSQL Wiki, http://en.wikipedia.org/wiki/NoSQL
- Cassandra, http://cassandra.apache.org/
- Mongodb, http://www.mongodb.org/
- https://www.youtube.com/watch?v=FR4QIeZaPeM
- https://www.mongodb.com/big-data-explained
- https://datajobs.com/what-is-hadoop-and-nosql
- http://bigdata-madesimple.com/a-deep-dive-into-nosql-a-complete-list-of-nosql-databases/
- Introduction to NoSQL and MongoDB - Northeastern University, www.ccs.neu.edu/home/kathleen/classes/cs3200/20-**NoSQL**MongoDB.**pdf**