DATA7001 INTRODUCTION TO DATA SCIENCE

[ɪntrəˈdʌkʃn]
n. 介绍；引进；

['saɪəns]
n. 科学；技术

Module 1 Problem Solving with Data

Shazia Sadiq

Welcome to Master of Data Science

['ɒfə(r)] vt. 提供

[ədˈvɑːnst] adj. 先进的；高级的

- First in Australia to offer an advanced level of knowledge applied in industry, government, social and scientific contexts.

[əˈplaɪd] adj. 应用的；实用的 ['ɪndəstri] n. 产业 ['gʌvənmənt] n. 政府；政体 ['səʊʃl] adj. 社会的 [ˌsaɪənˈtɪfɪk] adj. 科学的adj. 社会的 ['kɒntekst] n. 环境 ['emfəsɪs] n. 重点；强调；

- Emphasis on high level of graduate attributes through cross-disciplinary curriculum and innovative and disruptive thinking applied to complex problems.

['grædʒuət] n. 毕业生 [əˈtrɪbjuːt] n. 属性；

- Industry alignment is at the core of the program design, to ensure job-ready graduates capable of shaping the future of data science.

Program Design

- Compulsory
- Bridging
- Advanced
- Electives

Make a study plan today…

Seek academic advice!

Join the community

Master of Data Science cohort at piazza

piazza.com/uq.edu.au/other/allcohorts

UQ entrepreneurship program

https://ideahub.uq.edu.au

National and International Data Science Competitions

check out kaggle, govehack, …

Discussion board for the course

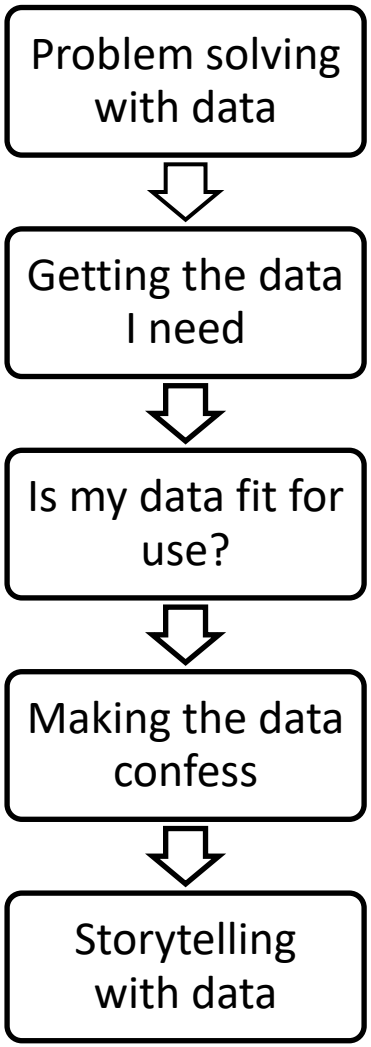
Piazza is a great tool for communication between students and between students and teaching team. Note that all resources and announcements will always be made via this site (blackboard), and piazza is used for discussions and Q&A.

Please sign up in case you are not already on it.

- Signup Link: piazza.com/uq.edu.au/semester22020/data7001
- Class Link: piazza.com/uq.edu.au/semester22020/data7001/home

About DATA7001

DATA7001 is a preamble for the rest of the program



→ DATA7901 Capstone Propose

→ DATA7002 Responsible Data Science
→ DATA7201 Data Analytics at Scale

→ DATA7201 Data Analytics at Scale ++
→ DATA7202 Advanced Statistics ++

→ DATA7201 Big Data Analytics ++
→ DATA7202 Advanced Statistics ++

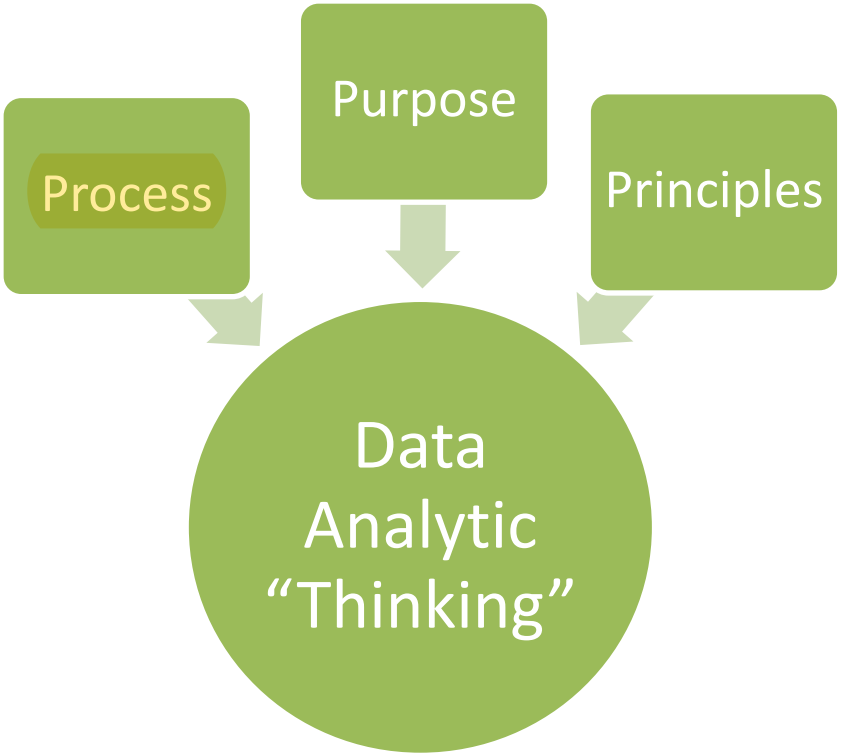
→ DATA7901 Capstone Propose
→ DATA7902/3 Capstone Build

Make the journey from learning data science to becoming a data scientist

Experimentalist ←→ Engineer

Learning Objectives

- Apply design thinking methodology to data science problems
- Design effective data science processes from problem formulation to persuasive story telling with data
- Reason with the fitness of basic computational analytical models in data science scenarios
- Develop data-centric approaches to complex business and scientific problems



Assessments

- Design thinking task – 10%
- Practicals – 15%
- Mid semester exam – 30%
- Project (multiple group and individual assessments) – 40%
- Adaptive learning – 5%

Course Structure

See Weekly Plan on Blackboard

DATA7001 Semester 2, 2020 (Tentative Weekly Plan)

	Monday Date	Lecture	Tutorial	Pracs	Assessment
1	3 Aug	Introduction to course What is data science Data Science Bootcamp (Guest Presentation by Nan Ye)	Data Science Bootcamp cont'd	Data Science Induction	
2	10 Aug	**1.Problem solving with data** Programming with R and Python	Collaborating on Code (Guest presentation by Richard Thomas)	P0 (Lab orientation)	
3	17 Aug	On Setting up Data Science Teams **2.Getting the data I need**	Problem solving with data	P1	
4	24 Aug	**3.Is my data fit for use**	Getting the data I need	P2	P1 (3%) Design Thinking Task (10%)
5	31 Aug	**4.Making the data confess**	Is my data fit for use	P3	P2 (3%)
6	7 Sep	4.Making the data confess cont'd	Making the data confess	P4	P3 (3%)
7	14 Sep	**5.Storytelling with data** Student Project Pitches (no marks)	Making the data confess cont'd	P4 cont'd	Project Pitches (no marks)
8	21 Sep	Recap of Data Science Process Mock Mid Sem exam	Storytelling with data	P5	P4 (3%) P5 (3%)
	28 Sep	Term Break			
9	5 Oct	Mid Semester Exam	No Tutorial	No scheduled practicals. Students will continue to have access to zones via uqcloud and Q&A via Zoom and Piazza	Mid Semester Exam (30%)
10	12 Oct	Group Project Consultations	Group Project Technical Q&A		
11	19 Oct	Group Project Consultations	Group Project Technical Q&A		
12	26 Oct	Project presentations			Project presentations (15%, in class, group)
	2 Nov	Exam Weeks			Project peer review (5%, individual) Project report (15%, group)
	9 Nov				Reflective essay (5%, individual) Adaptive Learning Task (5%)
	16 Nov				

Teaching team

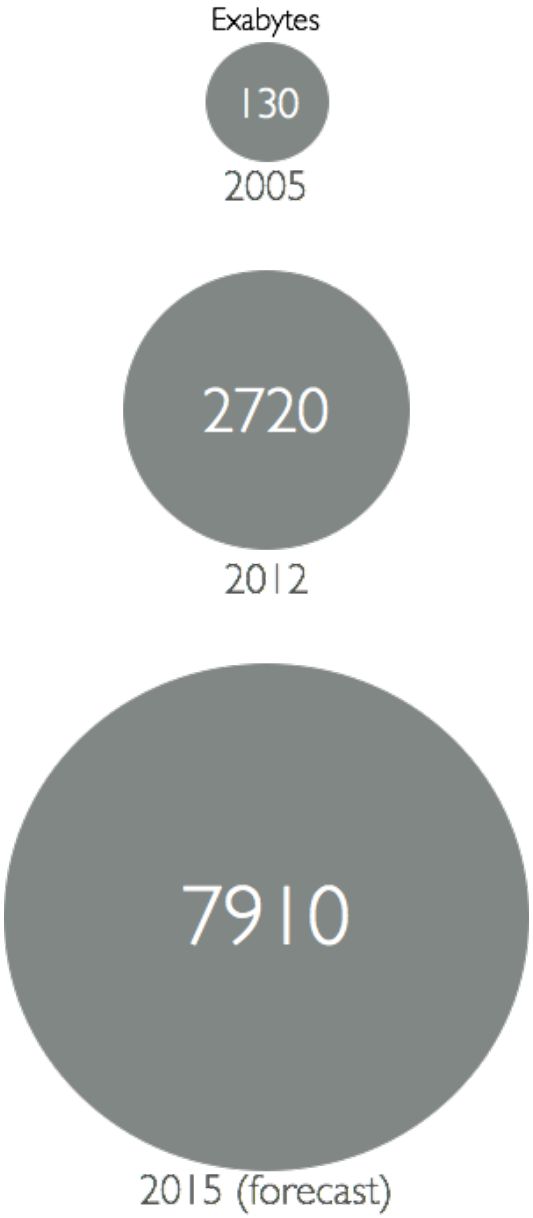
- Lecturers
  - Shazia Sadiq (course coordinator)
  - Thomas Taimre
- Tutors
  - Ajay Hemnath
  - Hrishikesh Patel
  - Reia Natu
  - Mubashir Imran
- Guest presenters

Module 1

- Emergence of Data Science as a discipline
  - Characteristics of (big) data
  - History of data management
  - Big data challenges
- Problem solving with data
  - Using design thinking to formulate authentic data science problems and develop well-targeted solutions

What Distinguishes Big Data?

Bytes (8 bits)
Kilobyte
1,024 bytes; 2^{10}; approx. 1,000 or 10^{3}
2 Kilobytes: Typewritten page
Megabyte
1,048,576 bytes; 2^{20};
approx 1,000,000 or 10^{6}
5 Megabytes: Complete works of Shakespeare
Gigabyte
1,073,741,824 bytes; 2^{30};
approx 1,000,000,000 or 10^{9}
20 Gigabytes: Audio collection of the works of Beethoven
Terabyte
1,099,511,627,776 or 2^{40};
approx. 1,000,000,000,000 or 10^{12}
10 Terabytes: Printed collection of the U. S. Library of Congress
with 130 million items on about 530 miles of bookshelves
Petabyte
1,125,899,906,842,624 bytes or 2^{50}
approx. 1,000,000,000,000,000 or 10^{15}
2 Petabytes: All U. S. academic research libraries
Exabyte
1,152,921,504,606,846,976 bytes or 2^{60}
approx. 1,000,000,000,000,000,000 or 10^{18}
5 Exabytes: All words ever spoken by human beings.
Zettabyte
1,180,591,620,717,411,303,424 bytes or 2^{70}
approx. 1,000,000,000,000,000,000,000 or 10^{21}
Yottabyte
1,208,925,819,614,629,174,706,176 bytes or 2^{80}
approx. 1,000,000,000,000,000,000,000,000 or 10^{24}

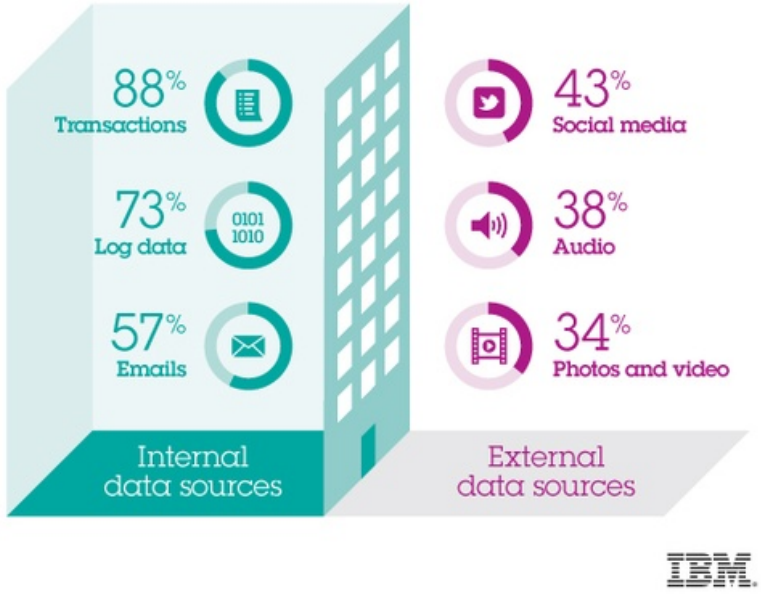


How big is Big Data?

EMC/IDC Digital Universe Study, 2011

Where is all this data coming from

According to IBM "Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone.



This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. "

…..more than 80% of all data is inactive, unmanaged, often unstructured, lacking meaningful metadata, and even unknown to the organisation. The proportion of this dark data is expected to reach 93% by 2020.

See also - Daniel Price. **Surprising Facts and Stats about the Big Data Industry.** Cloudtweaks.com, Mar 2015. http://cloudtweaks.com/2015/03/surprising-facts-and-stats-about-the-big-data-industry/

Example … Trajectory Data?

How Much Trajectory Data?

- A back-of-the-envelope calculation:
  - A simple point data (`x,y,t`): 24 bytes
  - **A car can generate 85KB a day** (10 hours a day, 10 seconds interval)
  - Beijing has 60,000 taxis, that is 5GB a day, or 1.72 TB a year
  - A car navigation service provider

	Current	Daily
Company X (in-car navigation provider)	17.6TB	15M trajectories
Company Y (map app provider)	14.5TB	5M trajectories
Company Z (social network)	0.68TB	18M trajectories

Every day, ~40M new trajectories, ~4 billion points

and …

- Volume
  - terabytes, petabytes, …
- Velocity
  - batch, real-time, streams, …
- Variety
  - structured, unstructured, multimedia,
- Veracity
  - reliability, availability, completeness, ….
- Value
  - insights, foresights, actions/decisions, ….

Famous Examples

http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=0
https://datafloq.com/read/big-data-obama-campaign/516
https://en.wikipedia.org/wiki/Google_Flu_Trends
http://www.datasciencecentral.com/profiles/blogs/the-amazing-ways-uber-is-using-big-data

A good example

10% of flights had 10 minute gap between ETA and ATA, 30% had 5 minute gap

- 2001 – PASSUR Aerospace builds RightETA
  - Public data on weather & flight schedules
  - Company proprietary data , including feed from passive radar stations
- 2012 – Collecting data every 4.6 seconds and maintaining historical data

Airline virtually eliminated gaps between ETA & ATA → multiple million $$$ at each airport!

Bad examples

- predict that someone searching for “used cars” might respond to an ad for used cars

- use social media to study unemployment rate (search for “jobs”)

http://www.theguardian.com/world/2015/jun/30/the-shrinking-of-the-big-data-promise
http://searchcio.techtarget.com/opinion/Big-data-bad-analytics

Task and Discussion

List two problems where you will need 2 or more datasets in order to develop a solution.

An airline company has up to 10 minutes gap between Estimated Time of Arrival (ETA) and Actual Time of Arrival (ATA). Reducing these gaps can save millions of dollars at each airport for the airline.

Airline can use (1) public data on weather; (2) flight schedules; (3) feed from radar stations; (4) historical data to accurately predict ATA.

DBMS: A Great Achievement!

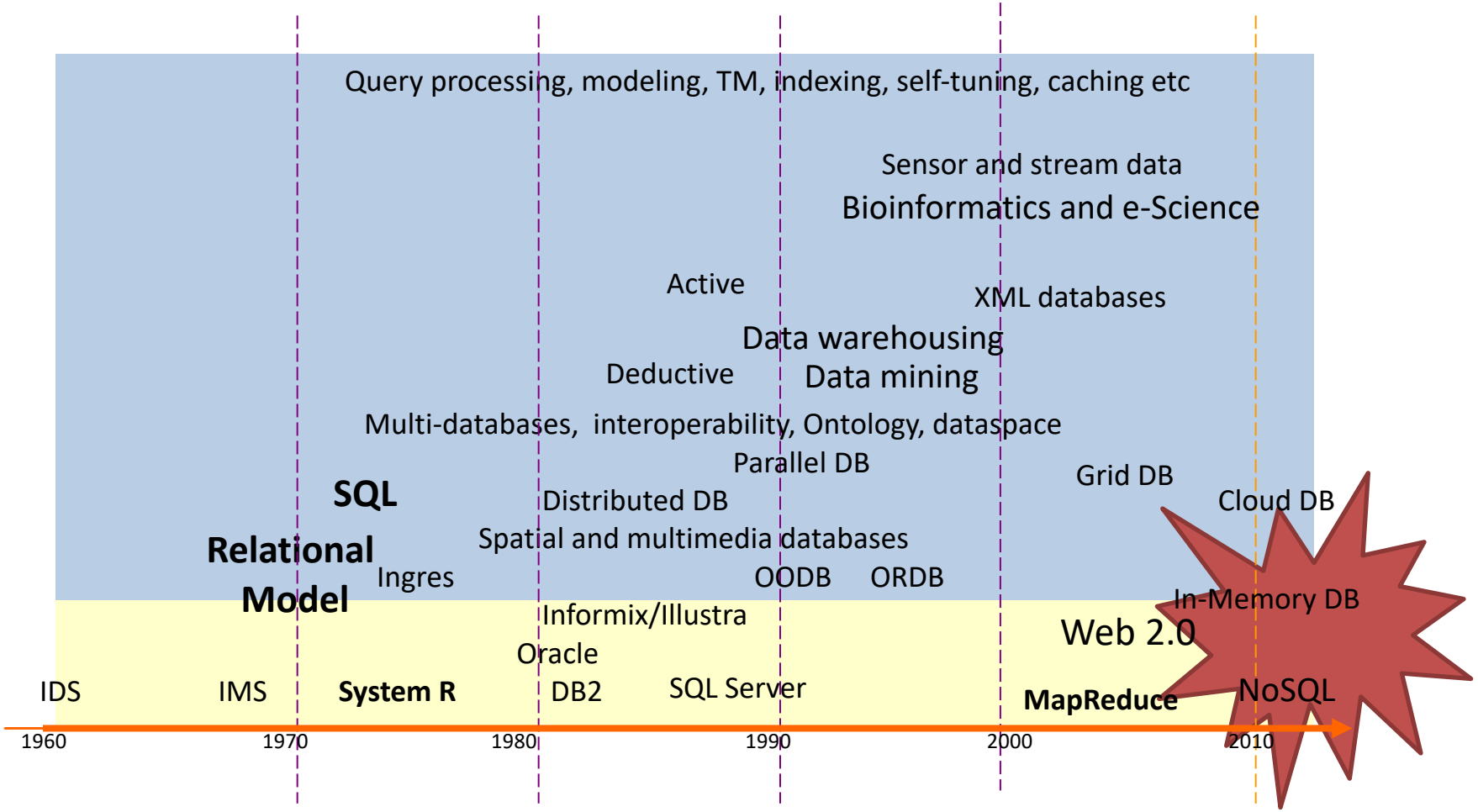
(Once) Advantages of DBMS

- Separation of data from applications
- Push-down common functions (general-purpose systems!)
- Separation of physical structures and logical structures
- Relational model and theory
- Non-procedural query language
- Concurrency control and recovery
- High performance query processing

DBMS in the Big Data Era

- The Closed-World assumption
- A piece of software, independent of hardware platforms (for too long!)
- A victim of its own success (extensions not well supported)
- Limited data types

A Brief History



See also - Gil Press. **A Very Short History of Data Science**. Forbes May 2013.
http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#3cc80d1f69fd

Data

Interactional

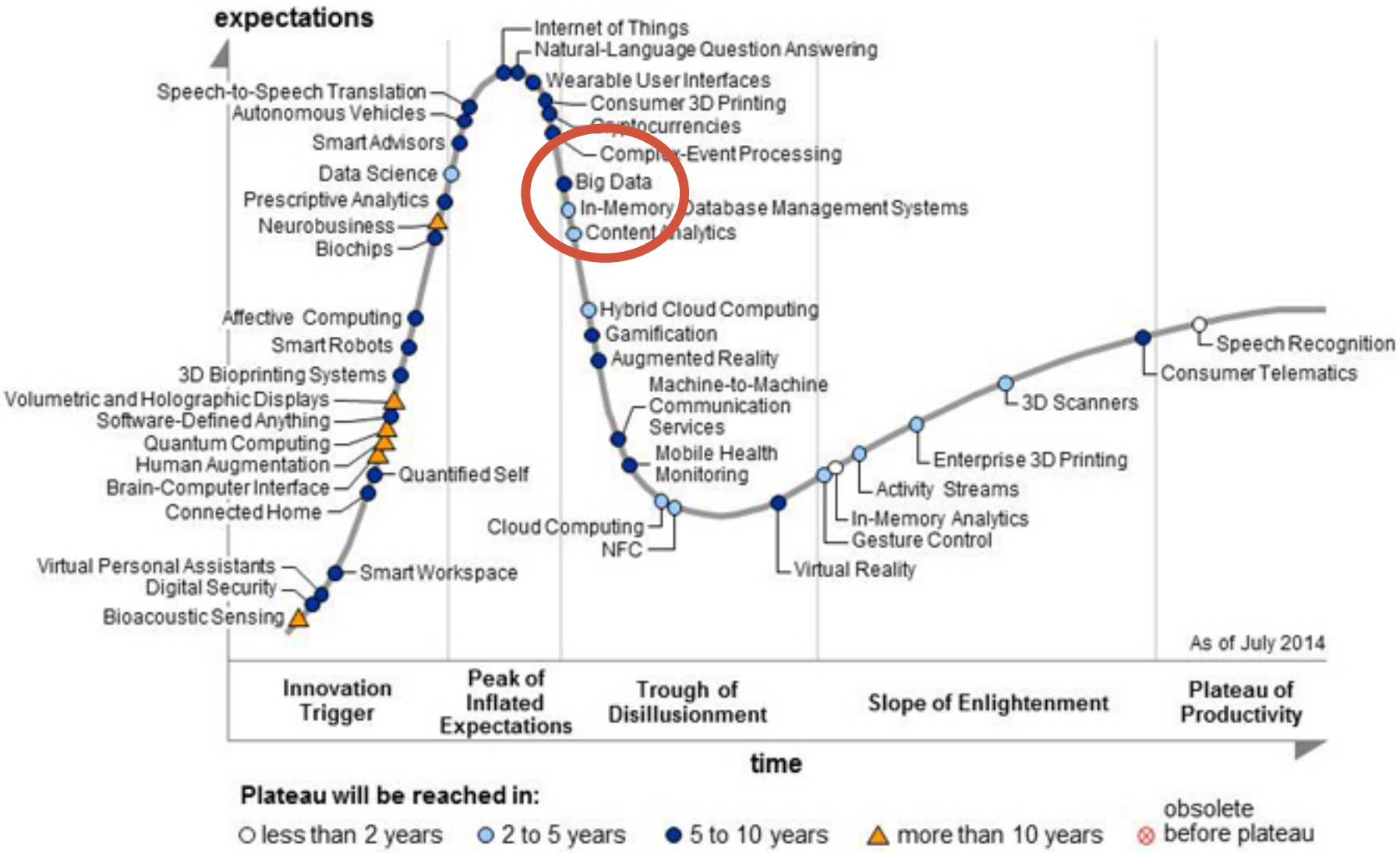
Sensory

Transactional

Time

Big Data vs DBMS

	DBMS	Big Data
Application-driven propriety solutions	1960s	2010s
Theory-based development	1970s	?
Commercialization	1980s	?
Universal adoption	1990s	?
Extensions	2000+	


expectations
Internet of Things
Natural-Language Question Answering
Speech-to-Speech Translation
Autonomous Vehicles
Wearable User Interfaces
Consumer 3D Printing
Smart Advisors
Data Science
Prescriptive Analytics
Neurobusiness
Biochips
Big Data
In-Memory Database Management Systems
Content Analytics
Affective Computing
Smart Robots
3D Bioprinting Systems
Volumetric and Holographic Displays
Software-Defined Anything
Quantum Computing
Human Augmentation
Brain-Computer Interface
Connected Home
Quantified Self
Hybrid Cloud Computing
Gamification
Augmented Reality
Machine-to-Machine Communication Services
Mobile Health Monitoring
Cloud Computing
NFC
Virtual Reality
Activity Streams
In-Memory Analytics
Gesture Control
Enterprise 3D Printing
3D Scanners
Consumer Telematics
Speech Recognition
Virtual Personal Assistants
Digital Security
Bioacoustic Sensing
Smart Workspace
As of July 2014
Innovation Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time
Plateau will be reached in:
less than 2 years
2 to 5 years
5 to 10 years
more than 10 years
obsolete before plateau

Big Data challenges

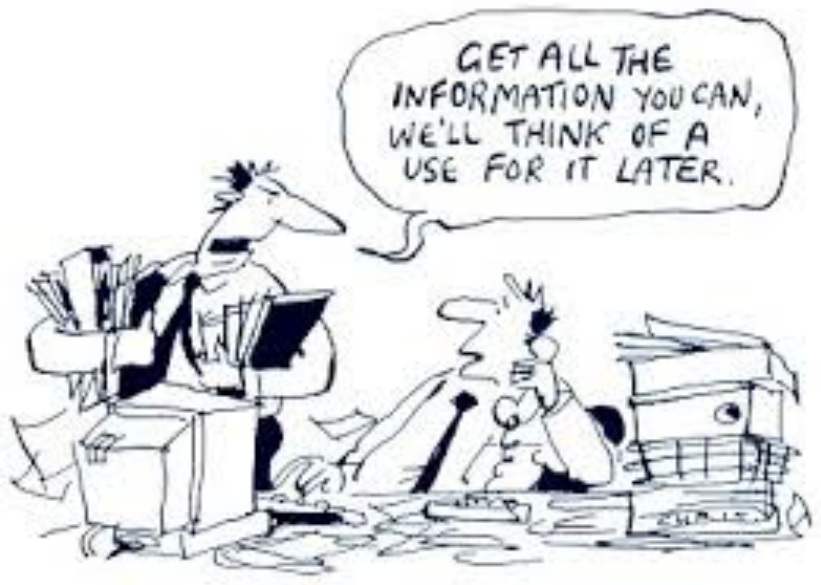
- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Not validating models
- Data pipeline integrity
- Using statistical tests correctly
- Making transition from prototype to production
- Data generation, process and use complexity (who do you ask?)
- …

Ack. John Canny and Mike Franklin UC Berkeley Lecture Notes

Bigger challenges (show stoppers)

- Lack of Purpose
  - Data before purpose
- Cultural divide
  - Business IT alignment
  - Privacy concerns
- Human Intelligence
  - Data literacy
  - Insight to Action
- Data Quality
  - Garbage in Garbage out!

Lack of Purpose

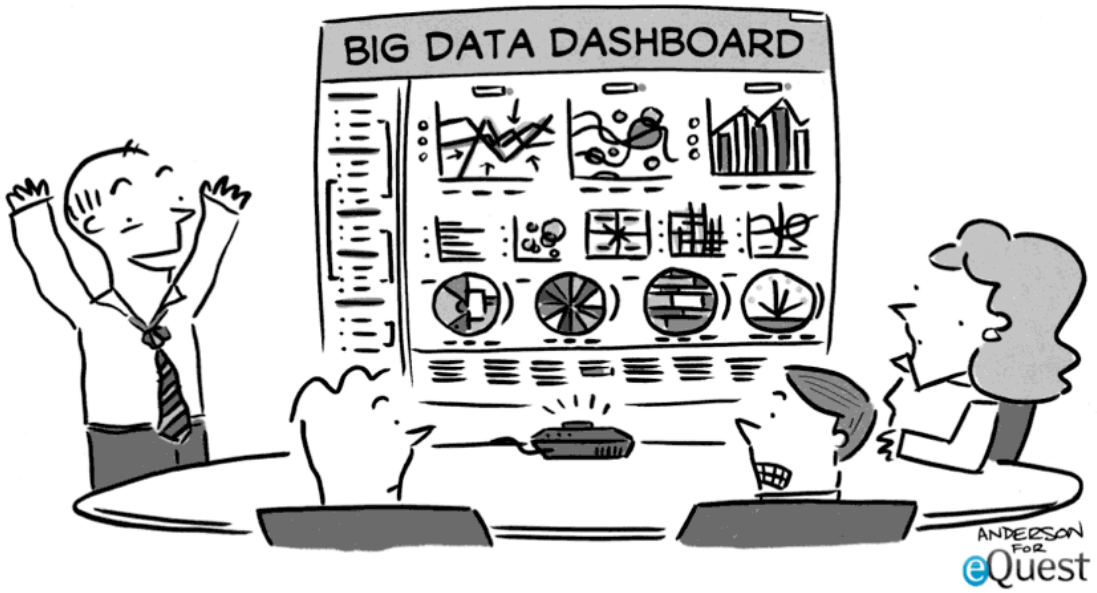


… if you don't know the right question to ask you discover nothing

Cultural Divide

… business IT divide!
Liability and Monetization ?

Human Intelligence



"After careful consideration of all 437 charts, graphs, and metrics, I've decided to throw up my hands, hit the liquor store, and get snockered. Who's with me?!"

… actionable insights?

Data Quality

- Poor quality data costs …
  - "$3 trillion to US government"
  - "$611 billion to US business for customer data alone"

You have to start with a very basic idea: **Data is super messy**, and data cleanup will always be literally 80% of the work. In other words, data is the problem.

"If you take something like LinkedIn in the early days, let's say, there were 4,000 variations of how people said they worked at IBM — IBM, IBM Research, Software Engineer, all the abbreviations, etc.," says Patil.

First US Chief Data Scientist at the White House

http://www.datablueprint.com/cost-of-bad-data-by-the-numbers/

Task and Discussion

Recall a data quality problem you may have encountered in your personal or professional life

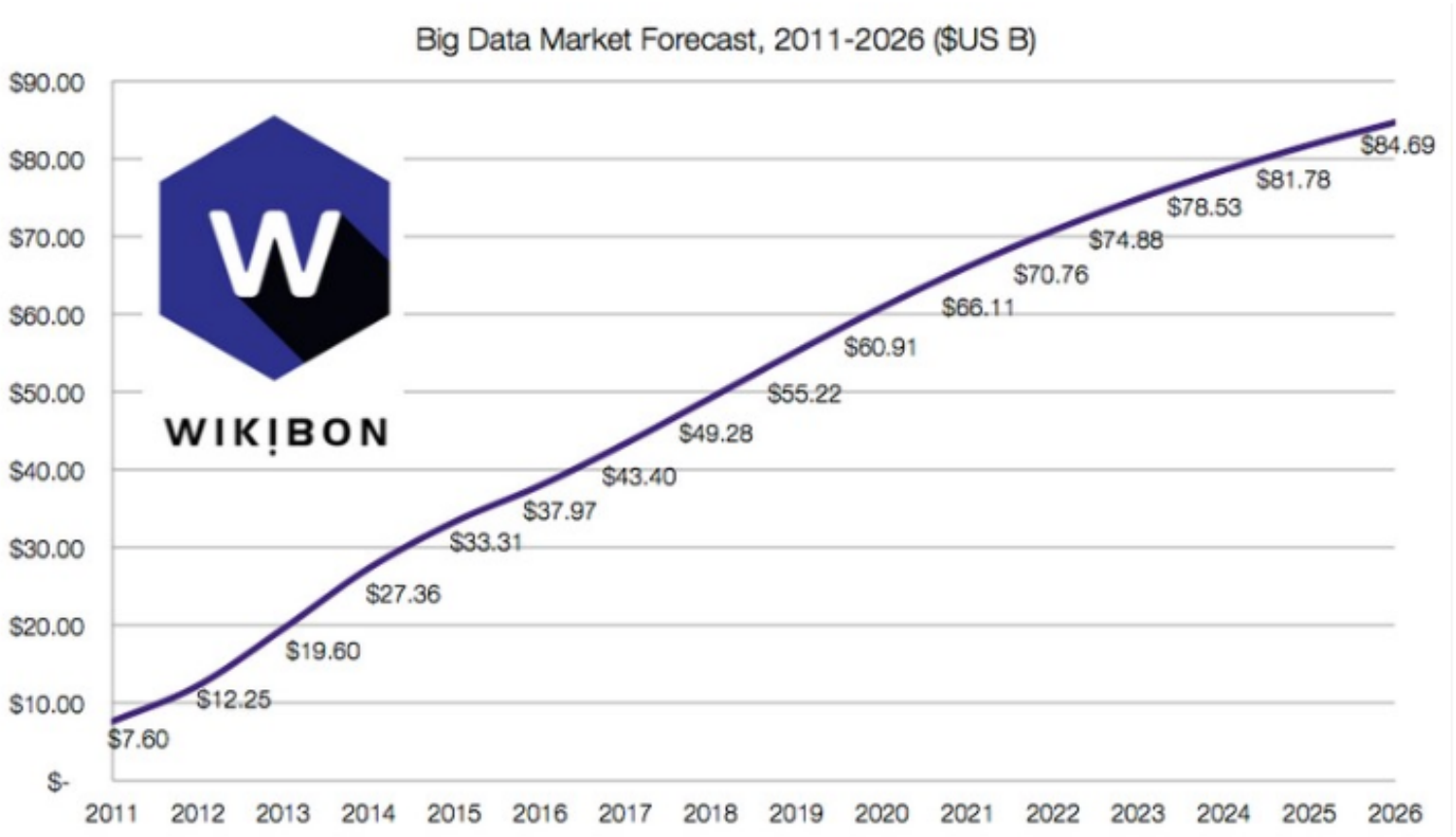
"Among Voters in New Jersey, GOP Sees Dead People," The New York Times, September 16, 2005 by David W. Chen." Comparing information from county voter registration lists, Social Security death records and other pubic information, Republican officials announced on Thursday (9/15/05) that 4,755 people who were listed as deceased appear to have voted in the 2004 general election. Another 4,397 people who were registered to vote in more than one county appeared to have voted twice, while 6,572 who were registered in New Jersey and in one of five other states selected for analysis voted in each state," according to Mr. Chen's article.

POLL QUESTIONS …

Data will only grow …

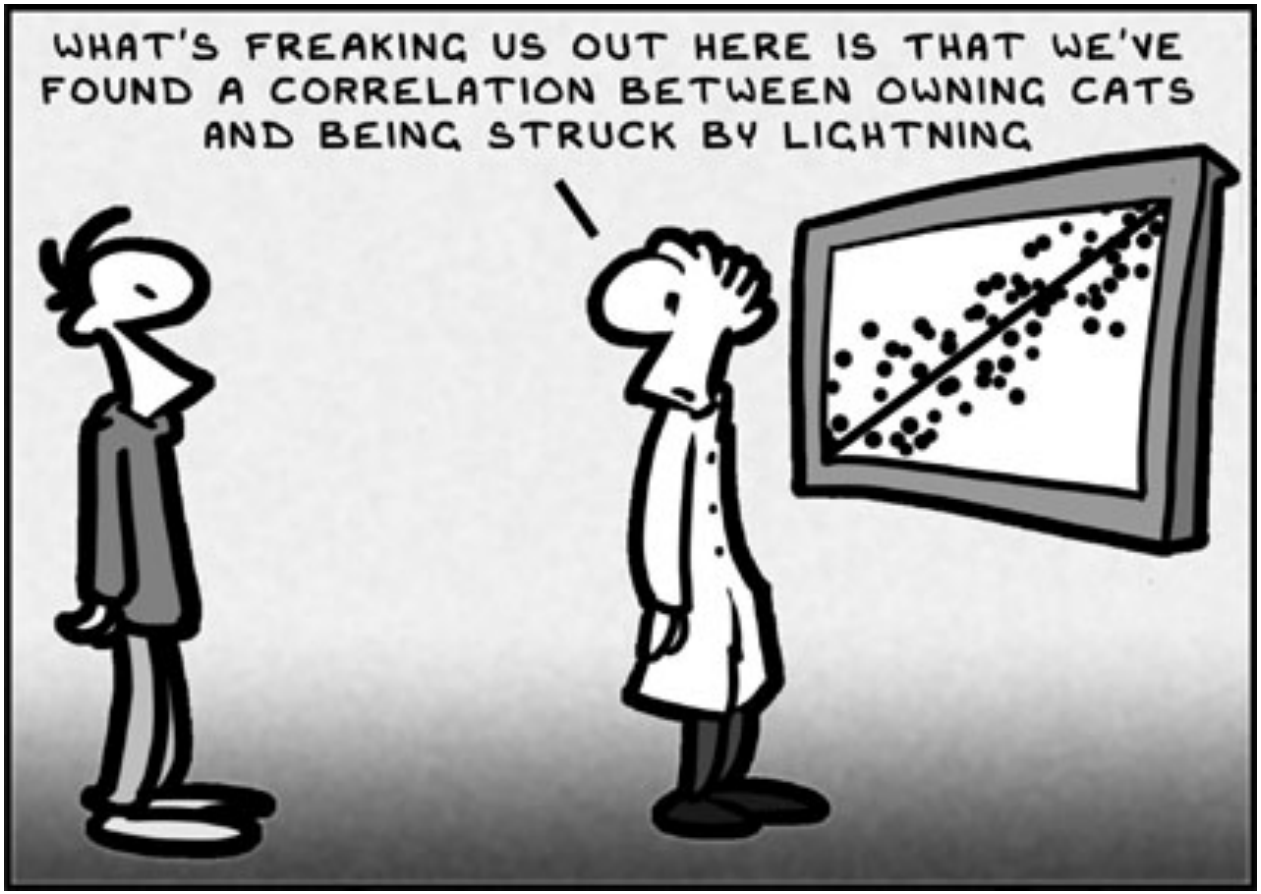
- Sloan Digital Sky Survey
- Next Generation Sequencing
- Large Hadron Collider
- Security and Surveillance
- Financial Planning and Risk Management
- Environmental Modelling
- Energy Saving
- Preventative Healthcare
- Intelligent Transportation
- Logistics
- Predictive Maintenance
- Brand/Product Protection
- Social Media and Marketing
- Credit Card Fraud Detection
- Computational Social Science

Big Data Growth and Adoption



http://www.forbes.com/sites/louiscolumbus/2015/05/25/roundup-of-analytics-big-data-business-intelligence-forecasts-and-market-estimates-2015/

Don’t be …


WHAT'S FREAKING US OUT HERE IS THAT WE'VE FOUND A CORRELATION BETWEEN OWNING CATS AND BEING STRUCK BY LIGHTNING

Module 1

- Emergence of Data Science as a discipline
  - Characteristics of (big) data
  - History of data management
  - Big data challenges
- Problem solving with data
  - Using design thinking to formulate authentic data science problems and develop well-targeted solutions

DATA7001
INTRODUCTION TO DATA SCIENCE

Module 2 Getting the Data I Need

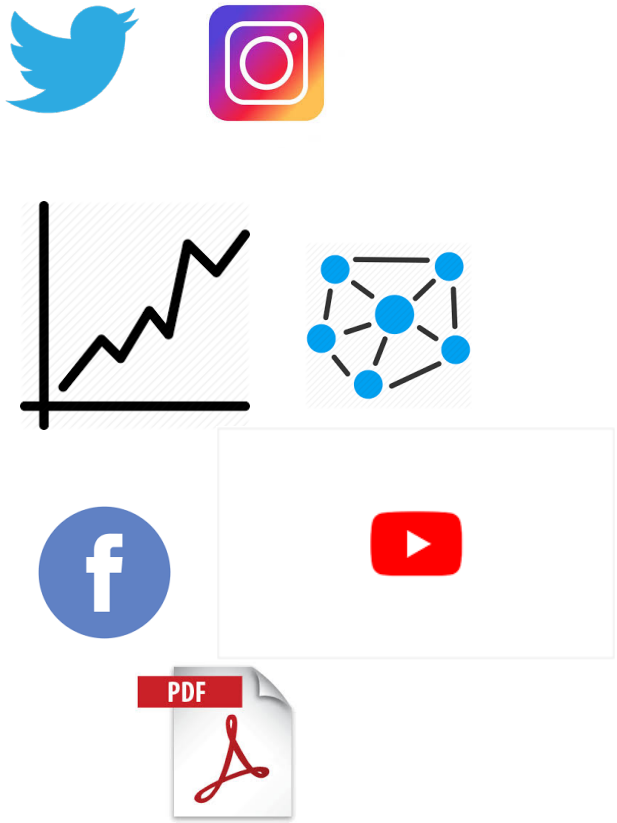
Module Topics

- Types of Data
- Data Ingestion
- Managing Data Privacy
- Sampling Big Data

Types of Data

- Structured
- Text
- Spatial
- Time Series
- Graph
- Multimedia

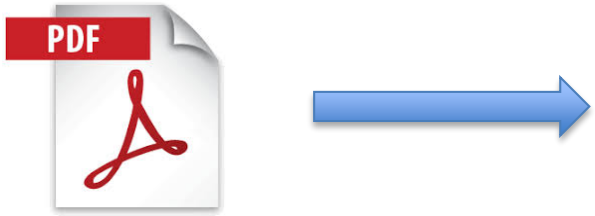
Name	Age	Course	GPA
xyz	25	D01	5.0
abc	22	D01	3.5
jlk	24	D01	6.5



- …and Meta-data

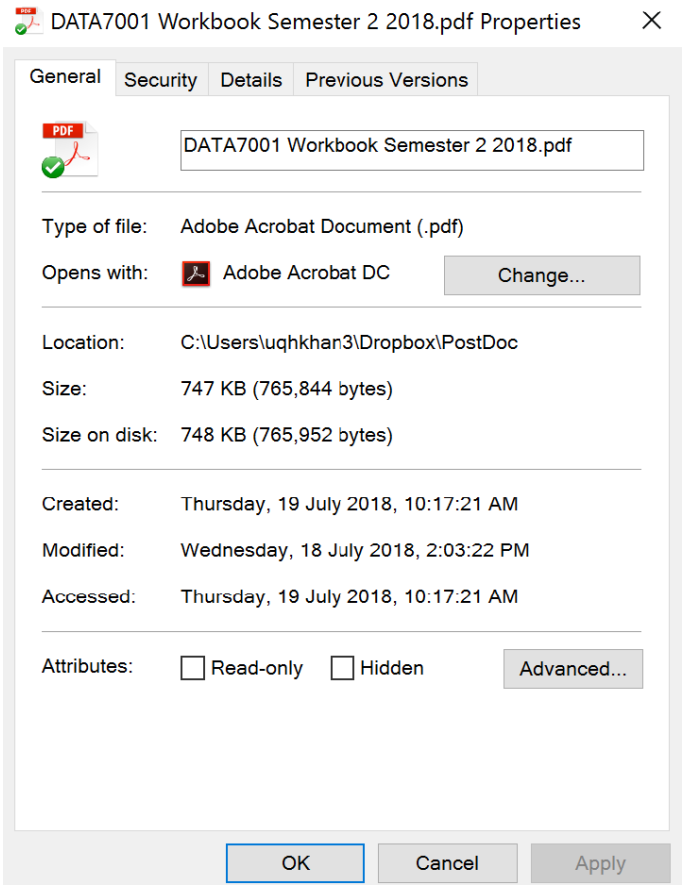
What is Meta Data?

Data



DATA7001 Workbook
Semester 2 2018

Meta Data

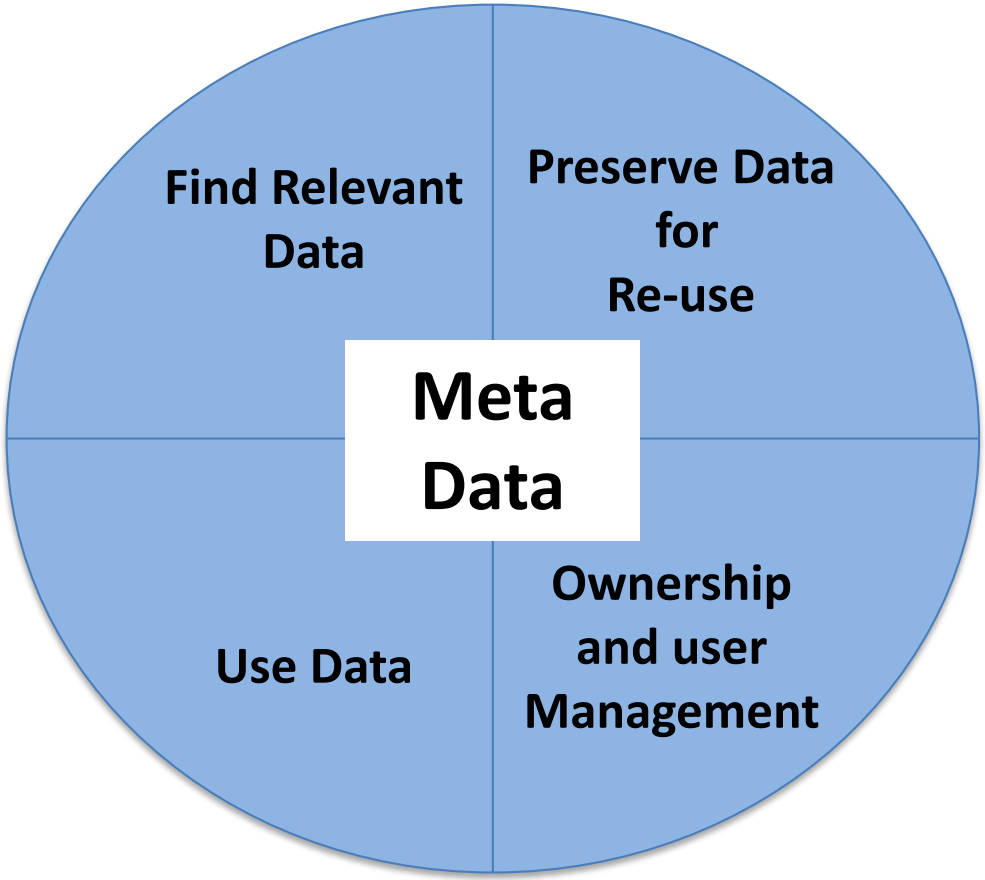


Meta-data is data about other data

Types of Meta-data

Type	Description	Use
Descriptive	• Describes the data, such as author, title, abstract, key terms • Difficult to create and manage	Enable searching and retrieval of data
Administrative	• Technical data on creation and quality control such as rights management, access control and use requirements • Created and managed by automation and by procedure.	Facilitates management of data for use and re-use
Structural	• Describes the information hierarchy of data, e.g., table of contents, frame index, versioning info. • Created when the data is created	Facilitates information navigation. Provides a higher view of how data is put together in a meaningful way.

Why Meta-data?



Where is meta-data ?

Embedded Storage

- Store metadata within the data file such as markups, file header or folders.
- Hard to disrupt the connection between data and metadata

Centralized Storage

- Store metadata in a centralized repository such as searchable indices and databases
- Easy to automate and standardize

Metadata format is routinely standardised, and for good reason. If the metadata cannot be understood, then the data cannot be understood….

Meta-data standards

Aspects of meta-data standards

Syntax - the layout, character encoding, punctuation of the metadata. XML is a good example.

Semantics - the terminology, interpretation and operationalisation of metadata. This includes dictionaries of tags, keywords, terms, classes in classification systems, and permitted ways for acting on each of these - such as how to render a tag, decrypt a codec, or route an address field.

Popular standards

There is a whole forest of metadata standards! The introduction of XML has helped standardise syntax in some areas, but the semantics of different disciplines are very different, and no two people interpret data in exactly the same way.

Semantic metadata standards :

- Library of Congress Classification,
- DDI for social science data
- TEI for archived text
- Dublin core for internet content
- MPEG for media
- RDF for relationships

More on Meta Data

Further reading on meta-data

The Wikipedia page on “metadata” is fine as a starting point. Metadata standards change frequently, so using this page as a jumping off point to read about ongoing standardisation projects is actually a good idea.

For more theory on library and information science, try:

Robert Colomb (2002). *Information Spaces, The Architecture of Cyberspace*. Springer London.

Gerard Salton (1989). *Automatic Text Processing*. Addison-Wesley Publishing.

Types of Data

- Structured
- Text
- Spatial
- Time Series
- Graph
- Multimedia

Structured (relational)

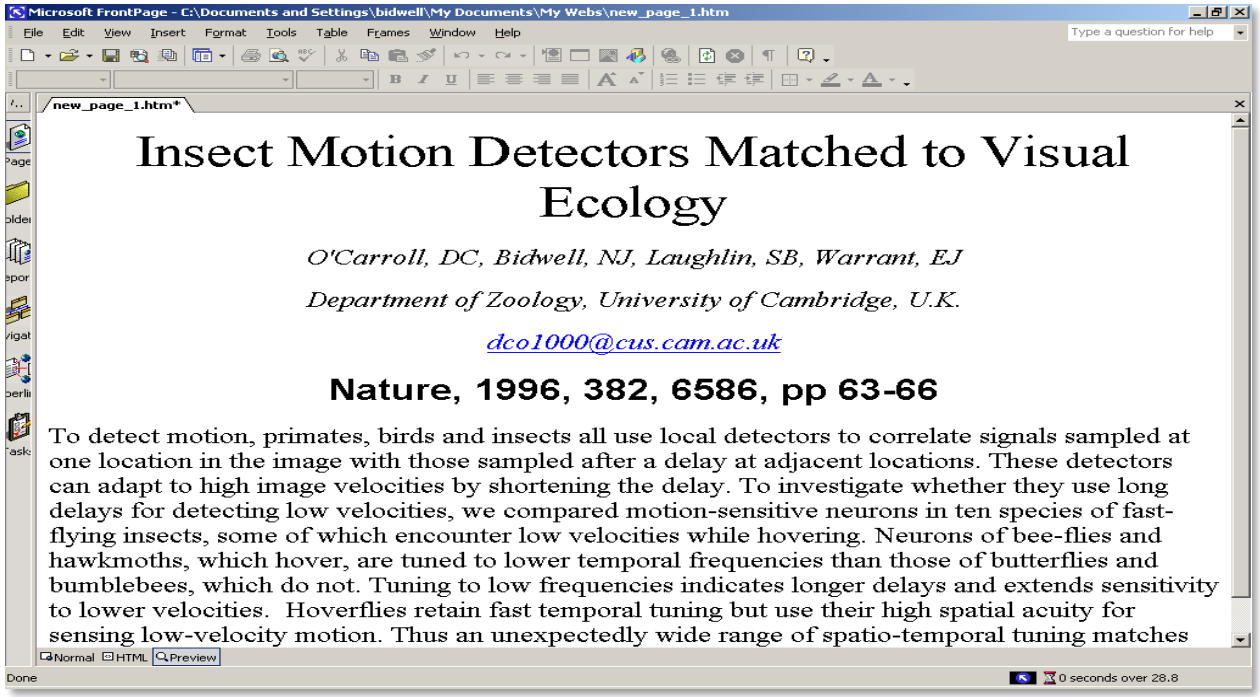
Title	Journal	Vol	Issue	Date	Pages
Insect Motion Detectors Matched to Visual Ecology	Nature	382	6586	1999	Pp 63-66
Information systems success: The quest for the dependent variable	Information Systems Research	3	1	NULL	60-95

Semi-structured (xml)

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema targetNamespace="http://www. http://www.itee.uq.edu.au/~nbidwell/schemas/paper”
        <xsd:complexType name=“publishType">
                <xsd:sequence>
                 <xsd:element ref=“journal"/>
                 <xsd:element name=“date" type="xsd:date“ minOccurs=“1"/>
                 <xsd:element name=“volume" type="xsd:string"/>
                 <xsd:element name=“issue" type="xsd:string"/>
                 <xsd:element name=“page" type="xsd:string"/>
                </xsd:sequence>
        </xsd:complexType>

</xsd:schema>
```

Unstructured (document)



Unstructured (short text)

Spatial

Any data with a location component

- Base-map data
  - Control points, topographic contours, building sites
- Natural area data
  - Soil types, landuse (industrial, agriculture, zoning etc), vegetation, water (rivers, ponds etc)
- Manmade area data
  - School districts, emergency service areas
- Land records data
  - Lot boundaries, zoning, easements
- Network data
  - Utilities (phones, sewers, water, electricity etc)
  - Roads (centerlines, curb lines, intersections, lights)

Spatial data representations

Vector Data

- A Canberra Suburb

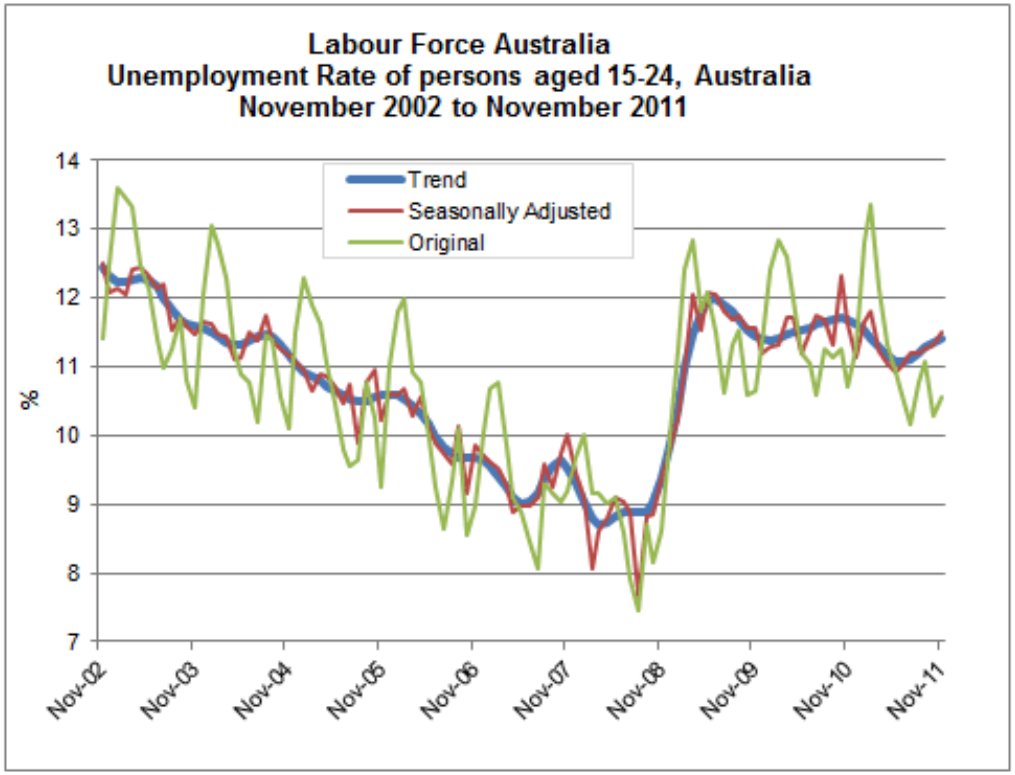
Raster Data

- Maroochy, Sunshine Coast

Time Series

A series of measurements taken at successive regularly spaced time intervals

"For example, measuring the level of unemployment each month of the year would comprise a time series. This is because employment and unemployment are well defined, and consistently measured at equally spaced intervals." [Australian Bureau of Statistics]



Types of Time Series Data

- A **stock series** is a measure of certain attributes at a point in time and can be thought of as “stock takes”. For example, the annual ABS *Prisoners in Australia* collection is a stock measure because it is a count of the number of persons in custody who were the legal responsibility of adult corrective services agencies on the night of 30 June each year.

- A **flow series** is a series which is a measure of activity over a given period. For example, the quarterly ABS *Corrective Services, Australia* collection is a flow measure as it provides prisoner counts taken on each day of the month which are summed and divided by the number of days in that month to determine the mean (average) daily prisoner number for that month

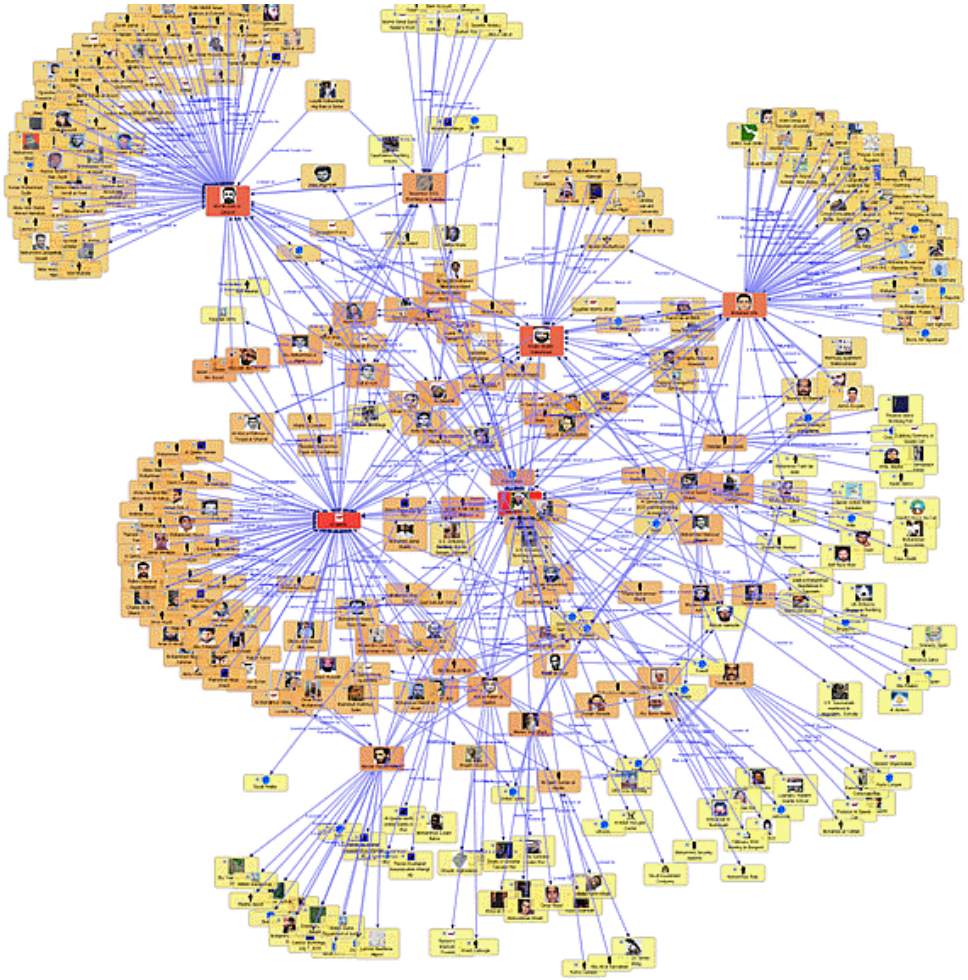
Graph Data

Widely used data structure

- Nodes: entities such as people, websites, objects
- Edges: relationships between nodes, e.g ‘friend of’
- Properties: attributes of nodes such as age group

Storage

- **Relational** (table) known to have flexibility and scalability limitations
- **Key-value** stores (~NoSQL databases e.g Neo4j) but also known to have issues of integrity and adoption



Multimedia Data

Multimedia data includes multiple data types including: text, images, graphics, audio and video.

Task and Discussion

Given two examples of *interesting* analytics that will need two or more different types of data

Example1. Sentiment analysis on product reviews social data (short text) displayed on a map (spatial data) with visualizations to support population demographics obtained from open census data (structured).

Example 2.

Example 3.

Module Topics

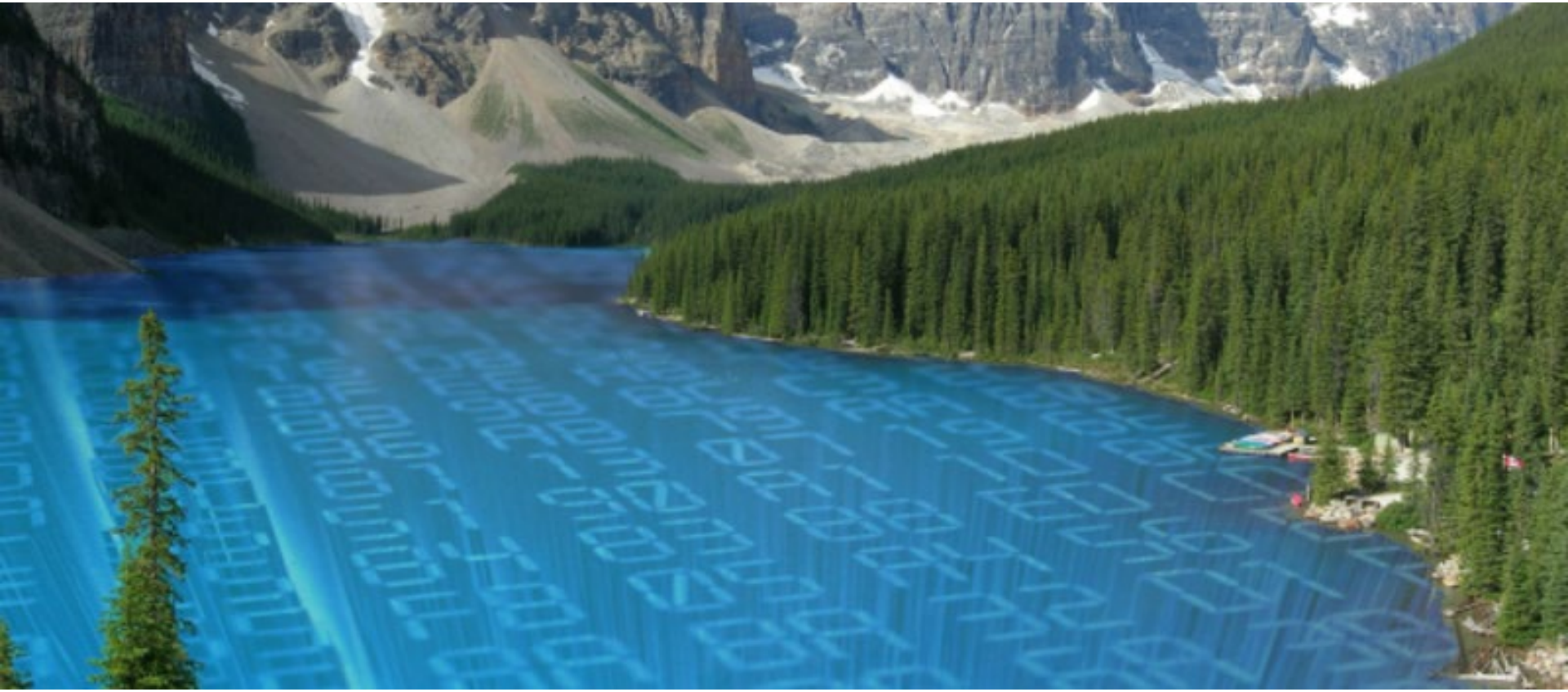
- Types of Data
- **Data Ingestion**
- Managing Data Privacy
- Sampling Big Data

- What data do you need?
  - **Why** do you need it?
  - Do you need **all of it**, or a **sample** will do?

- How do you get the data?
  - Do you own it, or will you beg, buy, or scrape it?
  - How fast is your data arriving (stream, batch, or snap)?
  - Are you authorized to acquire that data?

- Where do you keep the data?
  - In what form are you going to store (*ingest*) the data?
  - For how long do you need to keep the data?

Data Lake



A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed. While a hierarchical data warehouse stores data in files or folders, a data lake uses a flat architecture to store data.

Recall from Module 1 …

Dumping everything into e.g a Hadoop Distributed File System (HDFS), with a plan to do something interesting with it some day is going to turn your **DATA LAKE** into a **DATA SWAMP**

https://theconversation.com/use-it-or-lose-it-the-search-for-enlightenment-in-dark-data-68167

Automating Data Ingestion

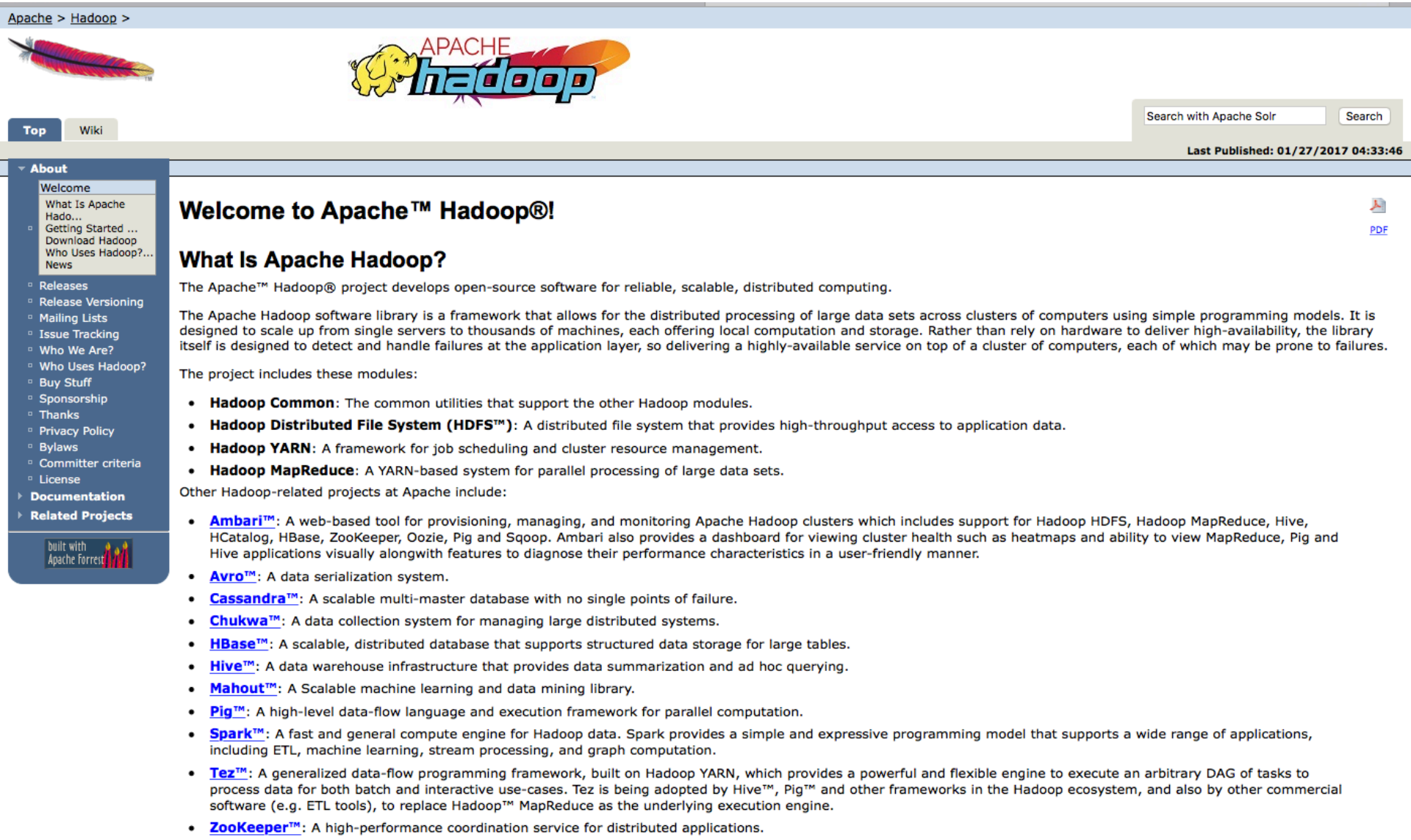
Challenges

- Need for customized scripts for each (multiple) source of data
- Huge demands on power, computing and bandwidth for high volume data (streams)
- Data loss due to outages from unstable connectivity
- Lack of security and difficulties in managing data access

- Data ingestion (typically) refers to HDFS (Hadoop Distributed File System)
- As the **Hadoop** ecosystem matures, many tools are available to simplify data ingestion e.g hortonworks and cloudera
- New frameworks have emerged to overcome some of the limitations of Hadoop e.g **Spark**

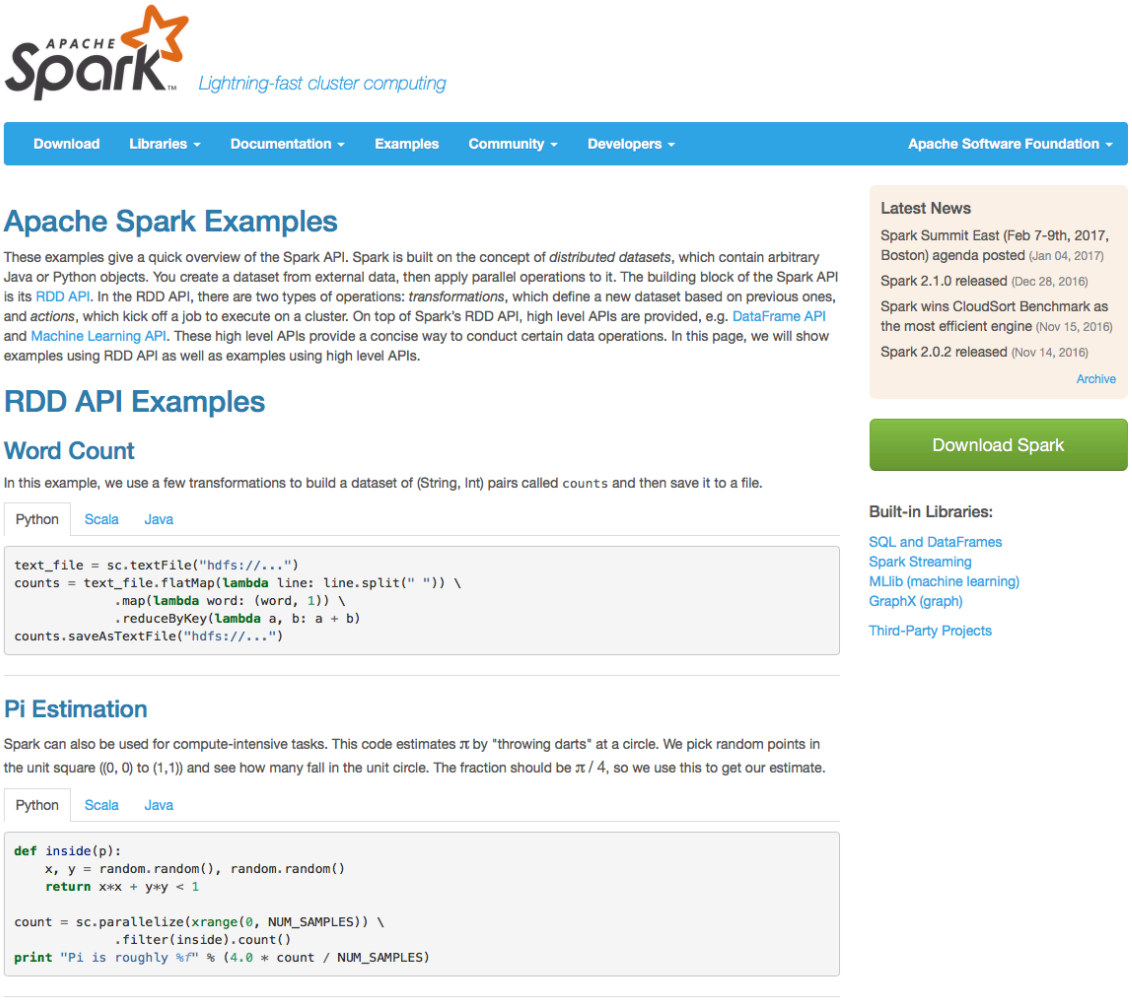
→ Accelerate the time to (start of) analytics

http://hadoop.apache.org



http://spark.apache.org

Apache Spark is an open-source cluster-computing framework. Originally developed at the University of California, Berkeley's AMPLab, the Spark codebase was later donated to the Apache Software Foundation, which has maintained it since. Spark provides an interface for programming entire clusters with implicit data parallelism and fault-tolerance. [wikipedia]



Examples of Data Ingestion

- CSV to MySQL (relational data)
- TXT to HDFS (hadoop distributed file system)

- Checkout tools for automating ingestion and other housekeeping tasks
  - Hortonworks
  - Cloudera

POLL QUESTIONS – DATA TYPES

Module Topics

- Types of Data
- Data Ingestion
- **Managing Data Privacy (next part)**
- Sampling Big Data

DATA7001
INTRODUCTION TO DATA SCIENCE

Module 2 Getting the Data I Need

Module Topics

- Types of Data
- Data Ingestion
- **Managing Data Privacy**
- Sampling Big Data

Patient data is private

Patient data has insights for scientific research on biomedicine, drug innovation, public health …

CCTV cameras (can) record private conversations from general public

Audio data can contain critical information on criminal, and terrorist activity

If you had access to data on people's movements, behaviors and social habits … what would you do with it?

Ethical use of big data

Some considerations..

Legal use of big data

Information Privacy Law

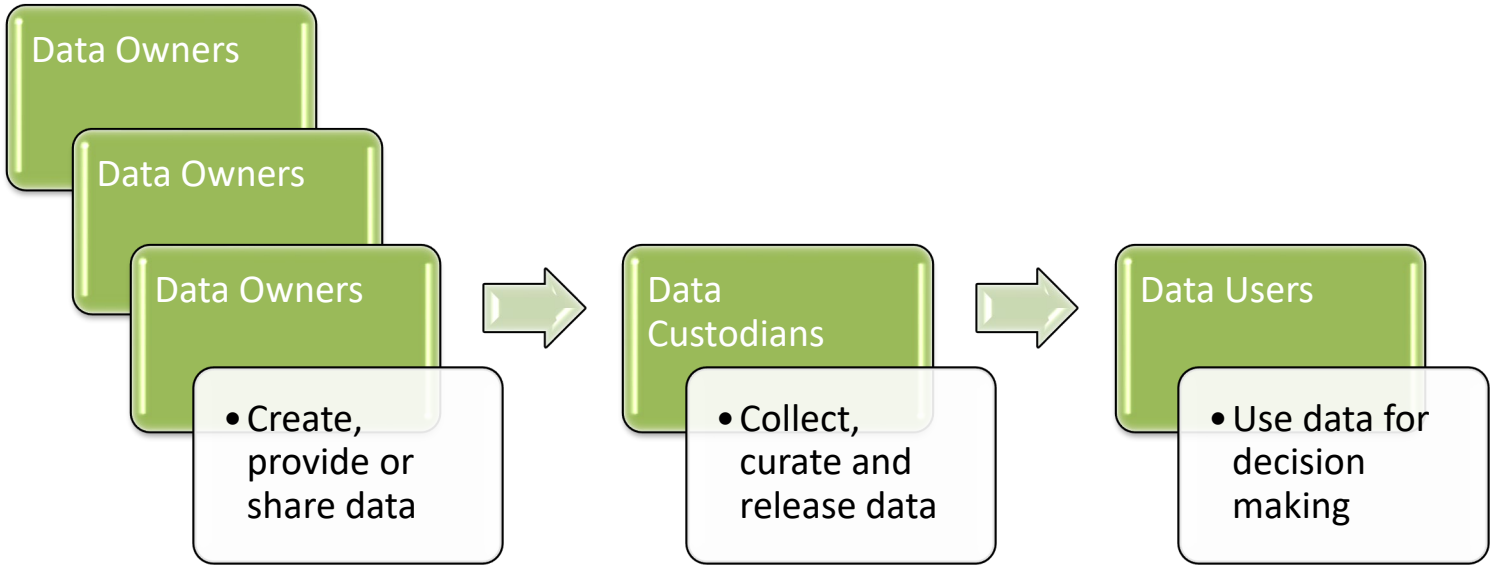
- Information privacy or data protection law is based on control concepts of privacy
  - Alan Westin, *Privacy and Freedom* (1967)
  - *Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others*
- Information privacy law is therefore about….the mechanics of personal information exchange
  1. Individuals have limited rights of control over collected personal information.
  2. Collecting organisations have legal obligations on how personal information is collected, stored and used.
- The law tries to balance individual rights and organisational requirements by providing
  - Fairness protections for individuals in the form of information privacy principles
  - Collecting organisations with flexibility to use and exchange personal information

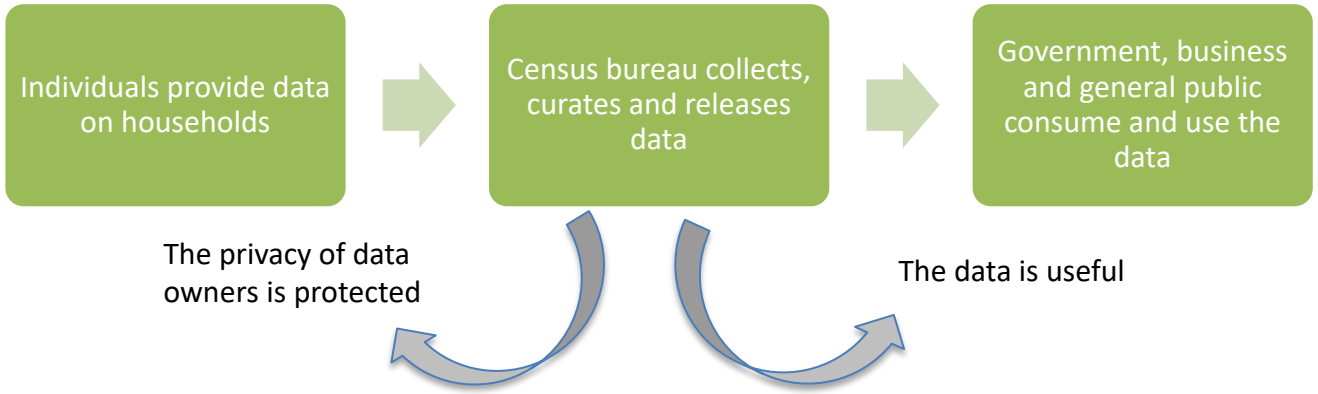
Questions for Data Scientists

- Does information privacy law apply to my de-identified data sets?
- Do I need to tell individuals about how I'm using their personal information?
- Do I have to keep my data sets up to date and secure?
- I can use my data sets for any purpose, right?

Acknowledgments: Mark Burdon (UQ Law School) and Andrew Crowden (UQ Philosophy)

Privacy preserving data release





Acknowledgements Xiaokui Xiao, Nanyan Technological University

Anonymisation Failure

- Massachusetts Group Insurance Commission
  - Medical records of state employees

Name	Birth Date	Gender	ZIP	Disease
Alice	1960/01/01	F	10000	flu
Bob	1965/02/02	M	20000	dyspepsia
Cathy	1970/03/03	F	30000	pneumonia
David	1975/04/04	M	40000	gastritis

Medical Records

At the time MGIC released the data, William Weld, then Governor of Massachusetts, assured the public that GIC had protected patient privacy by deleting identifiers. In response, then-graduate student **Latanya Sweeney** started hunting for the Governor's hospital records in the GIC data. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes.

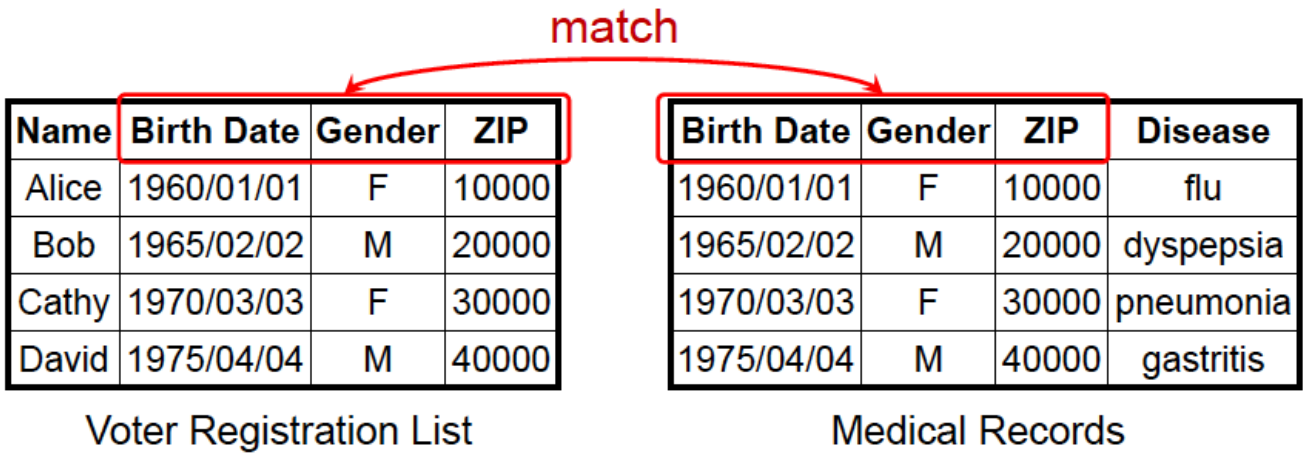
For twenty dollars, she purchased the complete voter rolls from the city of Cambridge, a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date, only three of them men, and of them, only he lived in his ZIP code. In a theatrical flourish, Dr. Sweeney sent the Governor's health records (which included diagnoses and prescriptions) to his office.

In 2000, Sweeney showed that 87 percent of all Americans could be uniquely identified using only three bits of information: ZIP code, birthdate, and gender.

https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/

Annonymisation Failure

- Massachusetts Group Insurance Commission
  - Medical records of state employees



Another Privacy Breach - AOL

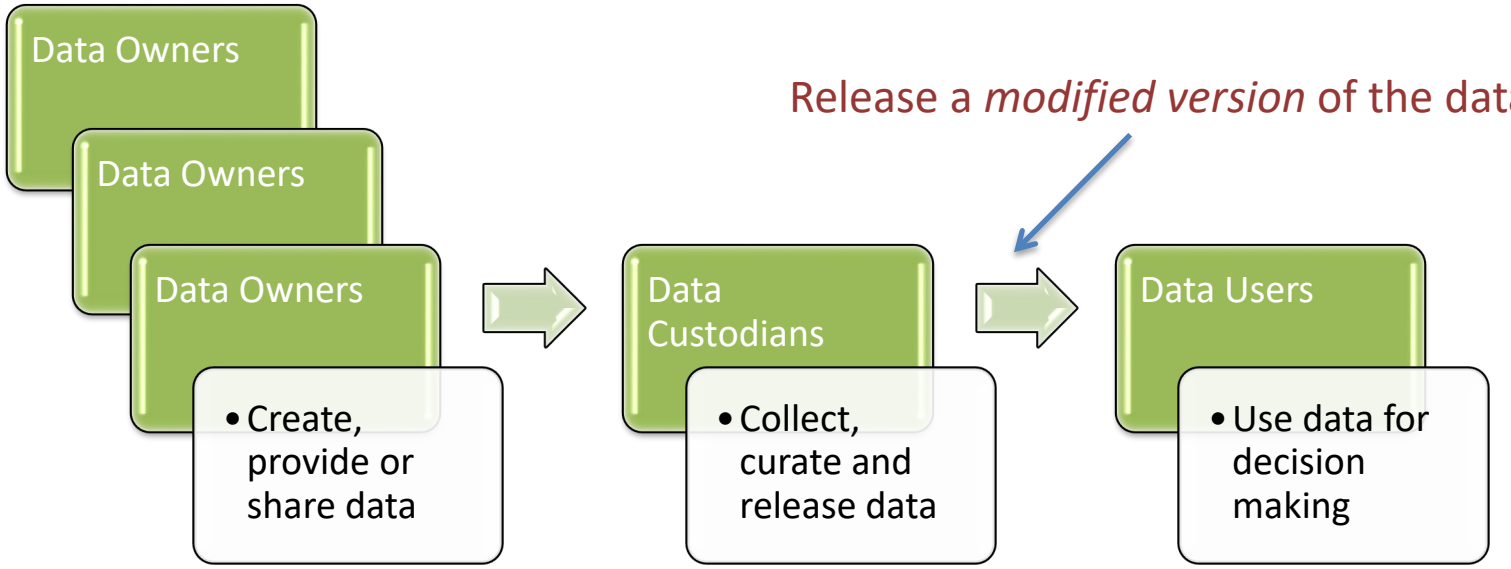
Log record: < User ID, Query, … >

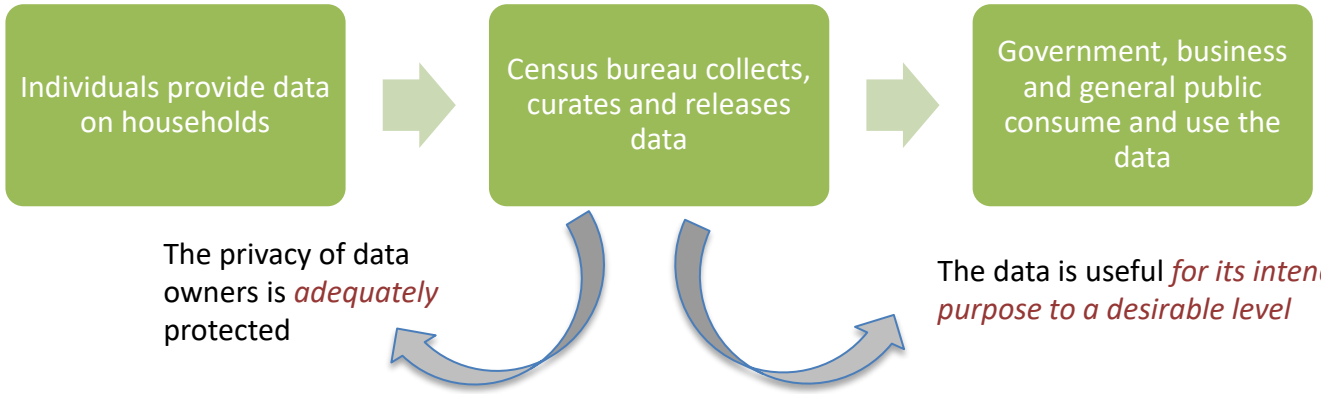
Example: < 4417749, “UQ”, … >

Method:

- Find all log entries for AOL user 4417749
- Many queries for businesses and services in Lilburn, GA (population 11K)
- A number of queries for different persons with the last name Arnold
- Lilburn has 14 people with the last name Arnold
- The **New York Times** contacted them and found that AOL User 4417749 is Thelma Arnold

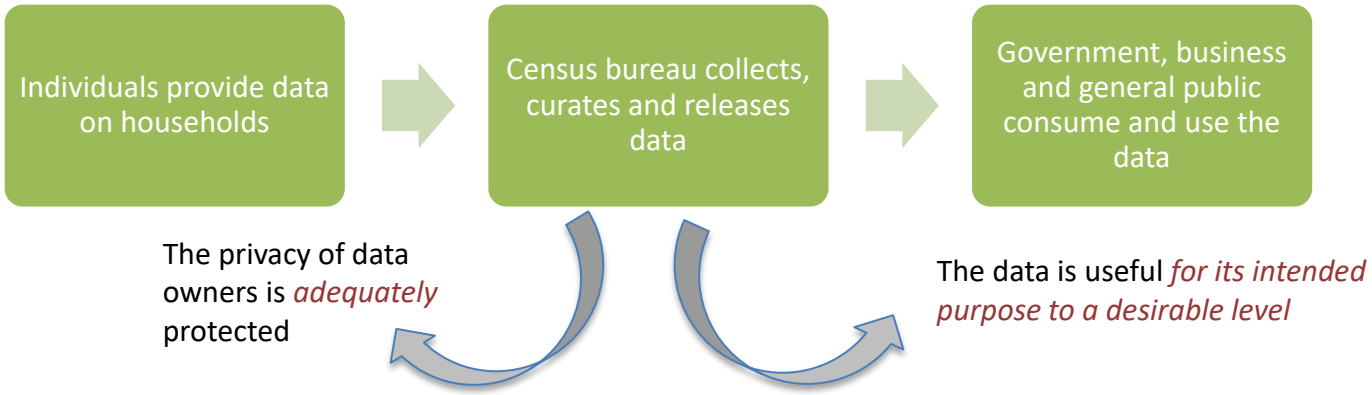
Privacy preserving data release





Acknowledgements Xiaokui Xiao, Nanyan Technological University

- privacy principle: what do we mean that privacy is by “adequately”protected?
- modification method: how should we modify the data to ensure adequate privacy while maximizing usefulness of the data for its intended purpose ?



Existing solutions (post 2000)

- ***K*-Anonymity**
- *l*-diversity
- Differential privacy

k-Anonymity

Example: We want to release medical records

Name	**Age**	**Zip**	**Disease**
John	20	1000	dyspepsia
Bob	30	2000	dyspepsia
Cathy	40	3000	pneumonia
Jane	50	4000	gastritis

Latanya Sweeney. 2002. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5 (October 2002), 557-570.

k-Anonymity

k-anonymity [Sweeney, 2002] requires that <Age, ZIP> combination can be matched to at least *k* patients.

This is done by making Age and ZIP less specific.

Name	Age	Zip
John	20	1000
Bob	30	2000
Cathy	40	3000
Jane	50	4000

adversary's knowledge

Age	Zip	Disease
20	1000	dyspepsia
30	2000	dyspepsia
40	3000	pneumonia
50	4000	gastritis

medical records

k-Anonymity

k-anonymity [Sweeney, 2002] requires that <Age, ZIP> combination can be matched to at least *k* patients. This is done by making Age and ZIP less specific.

generalization

Name	Age	Zip
John	20	1000
Bob	30	2000
Cathy	40	3000
Jane	50	4000

adversary's knowledge

Age	Zip	Disease
[20, 30]	[1000, 2000]	dyspepsia
[20, 30]	[1000, 2000]	dyspepsia
[40, 50]	[3000, 4000]	pneumonia
[40, 50]	[3000, 4000]	gastritis

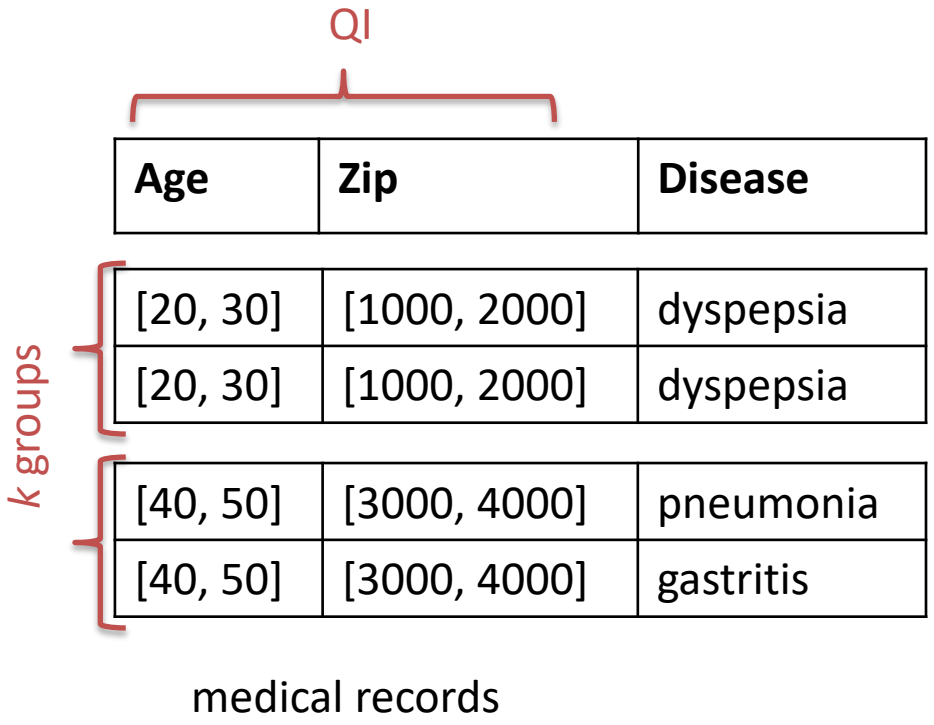
medical records

2-anonymous table

k-Anonymity

The general approach for *k*-anonymity required identification of attributes that an adversary may know e.g. Age and ZIP. These are called *Quasi-identifiers (QI)*.

You then divide the tuples into sizes of at least *k* and generalize the QI values of each group to make them identical.



medical records

k-Anonymity

k-anonymity requires that each combination of quasi-identifiers (QI) is hidden in a group of at least size *k*.

But what about the remaining attributes?

Name	Age	Zip
John	20	1000
Bob	30	2000
Cathy	40	3000
Jane	50	4000

adversary's knowledge

Age	Zip	Disease
[20, 30]	[1000, 2000]	dyspepsia
[20, 30]	[1000, 2000]	dyspepsia
[40, 50]	[3000, 4000]	pneumonia
[40, 50]	[3000, 4000]	gastritis

medical records

sensitive attribute!

k-Anonymity

What do you know about John?

Name	Age	Zip
John	20	1000
Bob	30	2000
Cathy	40	3000
Jane	50	4000

adversary's knowledge

Age	Zip	Disease
[20, 30]	[1000, 2000]	dyspepsia
[20, 30]	[1000, 2000]	dyspepsia
[40, 50]	[3000, 4000]	pneumonia
[40, 50]	[3000, 4000]	gastritis

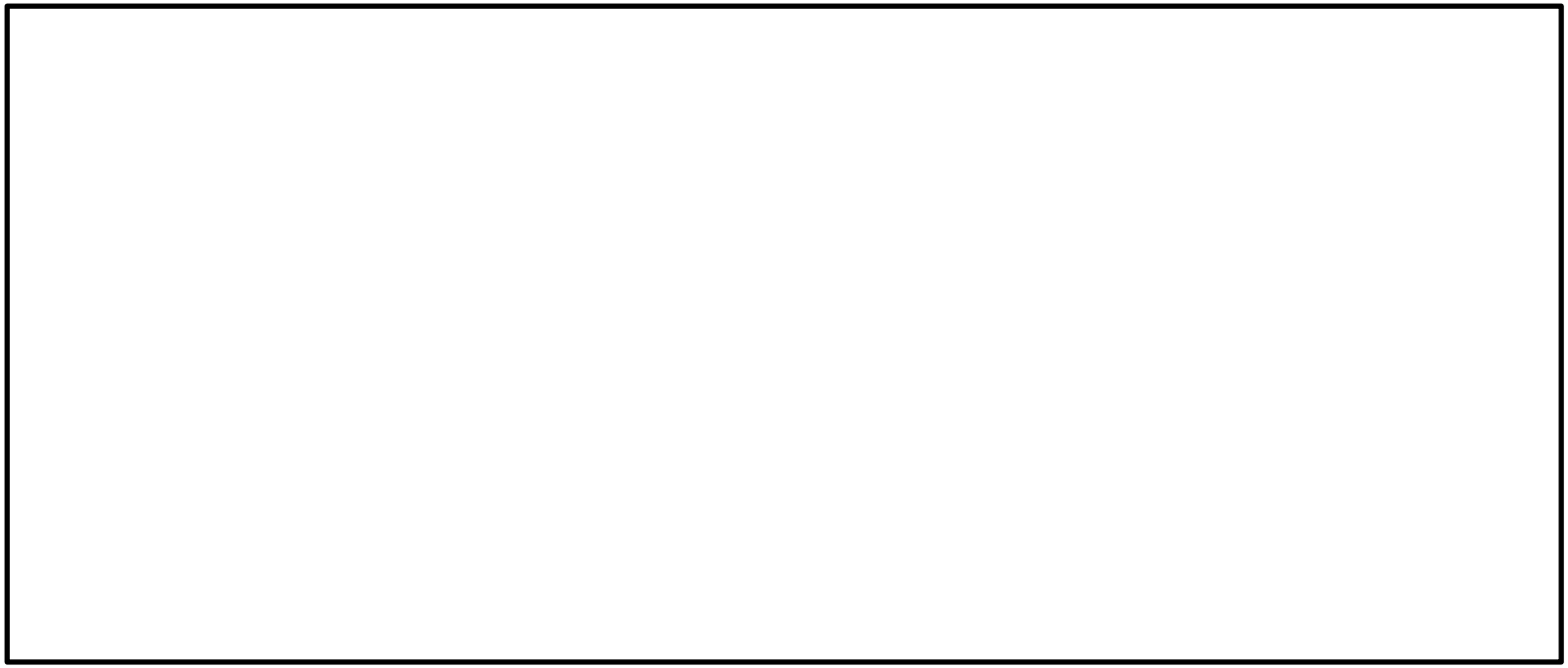
medical records

Vulnerability of Privacy Preserving Algorithms

- *k*-anonymity has been abandoned due to its vulnerability – disclosure of sensitive attributes is possible [Machanavajjhala et al. 2006].
- New algorithms have been proposed …
  - *l*-diversity
  - Differential privacy

Task and Discussion

What are the two principles of private data release?



POLL QUESTIONS - PRIVACY

DATA7001
INTRODUCTION TO DATA SCIENCE

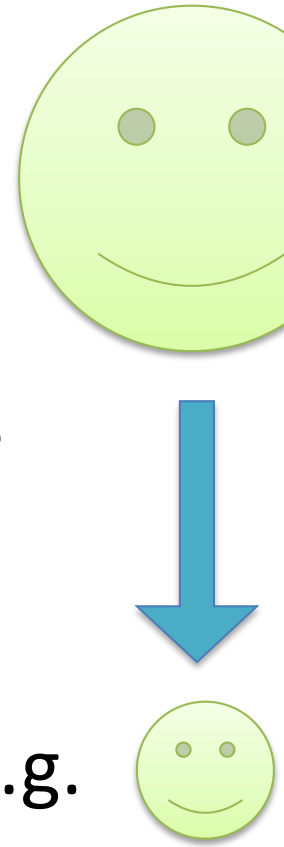
Module 2 Getting the Data I Need

Module Topics

- Types of Data
- Data Ingestion
- Managing Data Privacy
- **Sampling Big Data**

(Structured) Data Sampling – Why?

- Reduction of data
  - Volume of data – storage, accessibility
  - Convenience – laptop vs. cluster
  - Smaller dataset with same data structure
  - Generally applicable

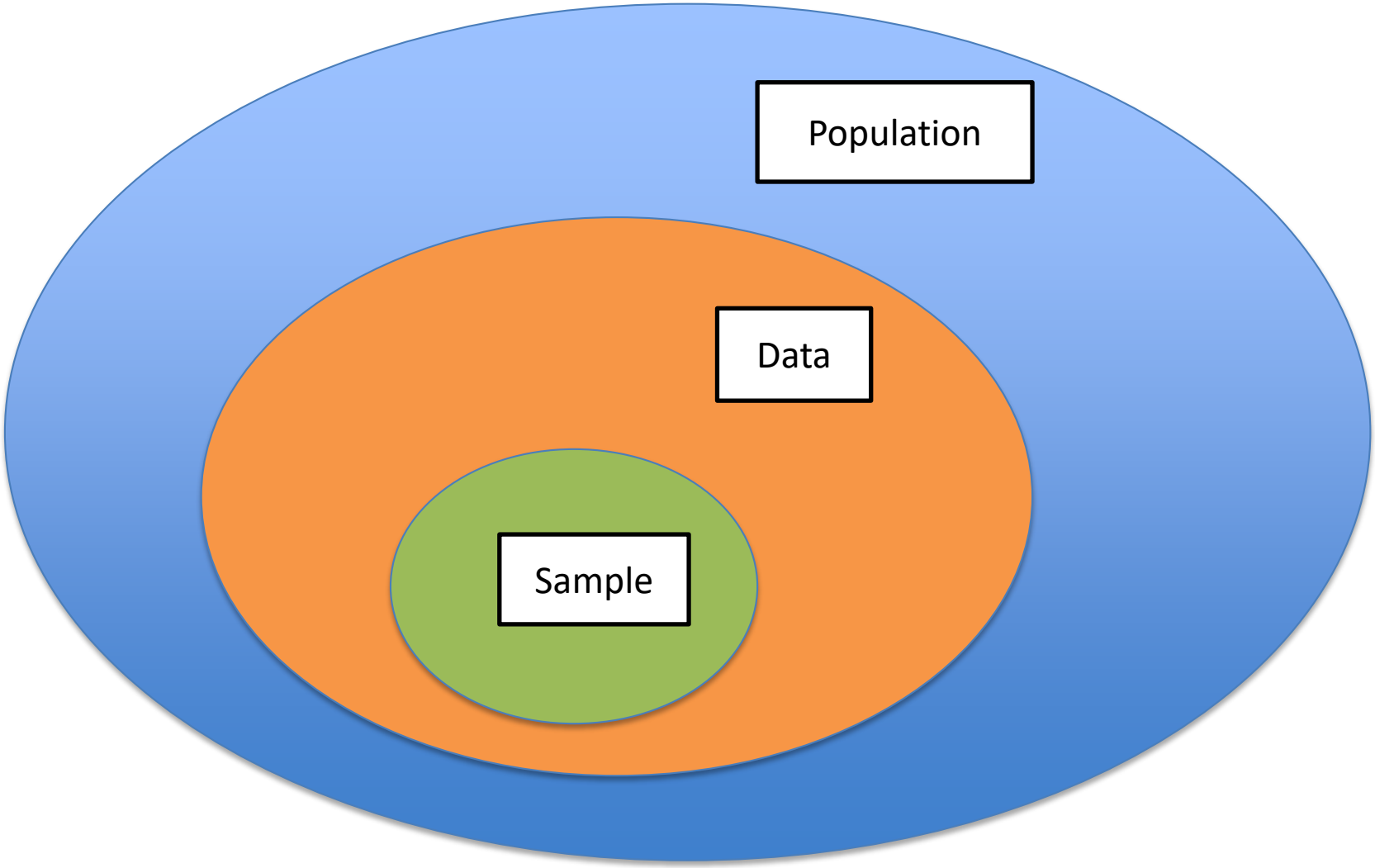


- Other data reduction methods exist (e.g. Summarization, PCA)

Data Sampling – What?

- Select data subset, usually according to probability rules
  - Simple Random Sampling
    - Each item has an equal chance of appearing in the sample
  - Weighted Random Sampling
    - Each item has a weight
    - Appears in sample proportional to weight
  - Stratified Sampling
    - Distinct groups (strata) present in data
    - Maintain representation of all groups in the sample
- Many other approaches (e.g. systematic sampling)

Data Sampling – What?



Data Sampling – What?

- Population
  - Set of items of interest (e.g. individuals, households)

- Data
  - Information pertaining to (usually part) of the population of interest
  - **NB: Often, we only have data on a sample of the population!**

- Sample
  - Subset of data, (random) representative of whole dataset

Data Sampling – How?

- Sampling Without Replacement (WOR)
  - Each time we add an item to the sample, it is excluded from being added again
  - No item is duplicated in the sample
  - Sampled items are DEPENDENT
- Sampling With Replacement (WR)
  - Each time we add an item to the sample, it is NOT excluded from being added again
  - Items could be duplicated in the sample
  - Sampled items are INDEPENENT
- **NB: We will ONLY consider WR!**

Data Sampling – How?

- Simple Random Sampling
  - Given n items in the dataset, want to select m items for the sample, WR (where $m << n$)
  - For each of the m items in the sample, choose item i in the dataset with probability $p_i = 1/n$

Data Sampling – How?

- Simple Random Sampling

	DATA ITEM	CATEGORY1	VALUE1
	1	F	27
	2	F	21
	3	F	18
	4	F	35
X 2	5	F	31
	6	F	22
	7	M	37
	8	F	21
	9	F	37
	10	M	55

SAMPLE ITEM	CATEGORY1	VALUE1
1	F	21
2	F	31
3	F	31
4	F	21

NB: *Sampling Error* with SRS can lead to loss of data features

Data Sampling – How?

- Weighted Random Sampling
  - Given n items in the dataset, each with a (positive) weight w_i, want to select m items for the sample, WR (where $m<<n$)
  - For each of the m items in the sample, choose item i in the dataset with probability p_i proportional to w_i
  - **NB: The weights should be designed to capture data features of particular interest**

Data Sampling – How?

- Weighted Random Sampling (e.g. PPS)

DATA ITEM	CATEGORY1	VALUE1
1	F	27
2	F	21
3	F	18
4	F	35
5	F	31
6	F	22
7	M	37
8	F	21
9	F	37
X 2 10	M	55

SAMPLE ITEM	CATEGORY1	VALUE1
1	M	55
2	F	35
3	F	37
4	M	55

PPS: Probability Proportional to Size

Data Sampling – How?

- Stratified Random Sampling
  - Given n items in the dataset, each belonging to one of s strata, want to select k items from each stratum giving $m=sk$ items for the sample, WR (where $m<<n$)
  - For each of the s strata, choose each of the k samples for that stratum uniformly at random (i.e. according to SRS within the stratum)
  - **NB: Strata can be created artificially by selecting ranges of a numerical variable (e.g. income bands)**

Data Sampling – How?

- Stratified Random Sampling

DATA ITEM	CATEGORY1	VALUE1
1	F	27
2	F	21
3	F	18
4	F	35
5	F	31
6	F	22
7	M	37
8	F	21
9	F	37
10	M	55

X 2

SAMPLE ITEM	CATEGORY1	VALUE1
1	F	21
2	F	21
3	M	37
4	M	37

NB: Two samples taken uniformly at random from each category `F' and `M'

Data Sampling – How?

- In general, stratified random sampling may not sample the same number of items from each stratum.
- Instead, the idea is to make sure that the “right” number of items are sampled from each stratum.
  - E.g. we may want to to preserve the proportion of the strata in some study
  - E.g. we may want to oversample a rare strata in order to perform meaningful statistical analysis on these rare strata.

Data Sampling – When?

- Sampling can occur during or after data collection
  - Here, we focus on the latter case
- Sampling methods (particularly SRS) are also used for analytic purposes (e.g. cross-validation of statistical models)
- Simple Random Sampling is easy; however, can lose data features (e.g. unusual items)
- Weighted Random Sampling or Stratified Sampling can be used to address this problem

Task and Discussion

For each of the three sampling methods, give an example of a dataset for which the method is appropriate.

POLL QUESTIONS - SAMPLING

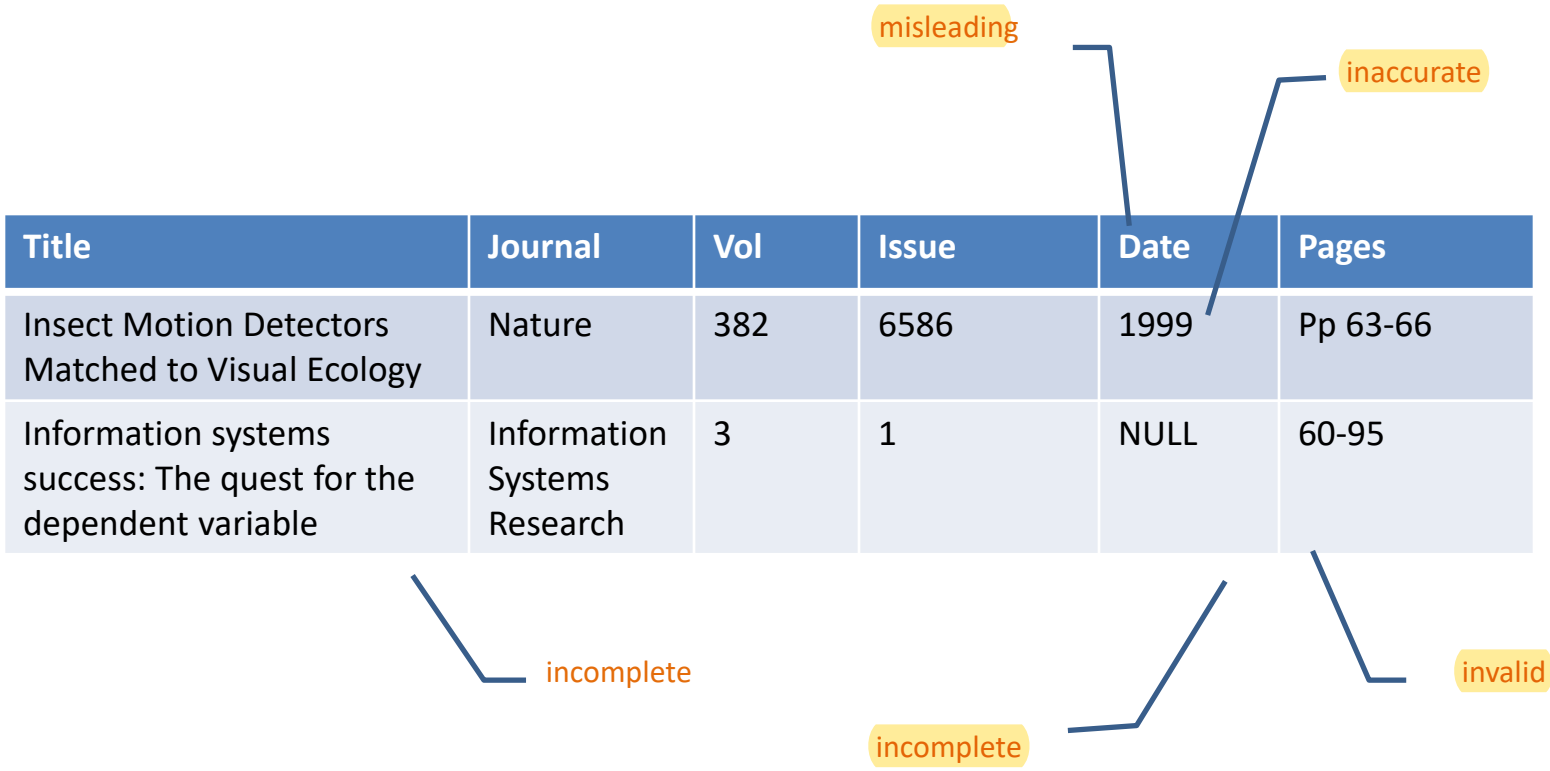
DATA7001
INTRODUCTION TO DATA SCIENCE

Module 3 Is my Data Fit for Use

Module Topics

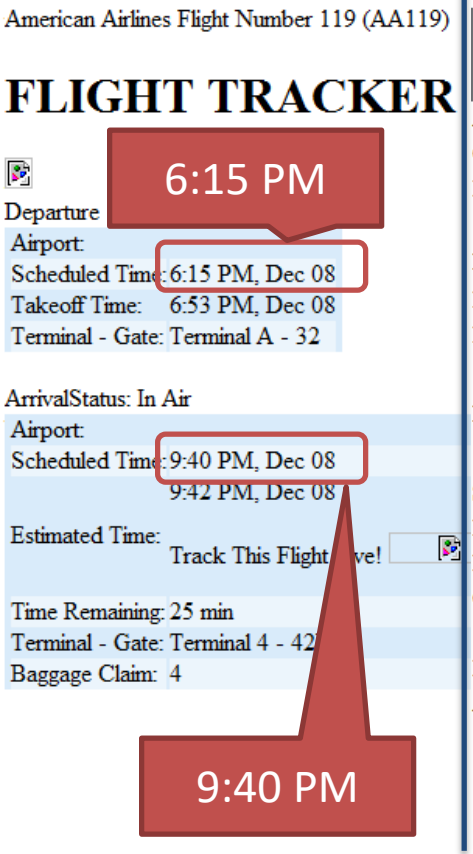
- **What is Data Quality**
- Data Exploration
  - Discovering and understanding the quality characteristics of the data through exploratory techniques
- Data Transformation
  - Transforming the data through cleaning, curating, repairing
- Data Enrichment
  - Enriching the data through data imputation and integration

Any Problems?

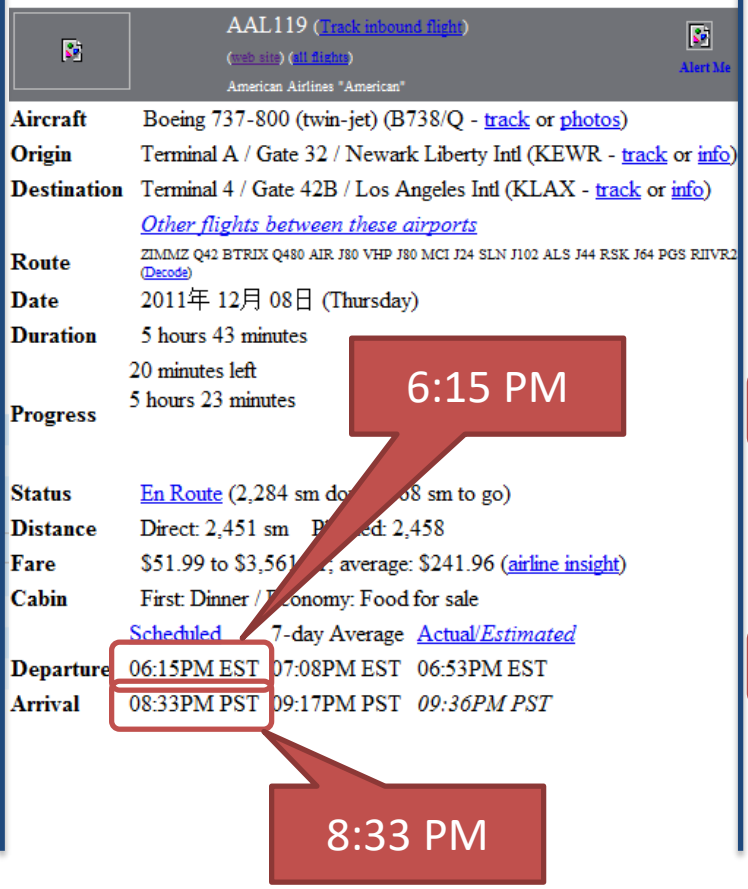


Any Problems?

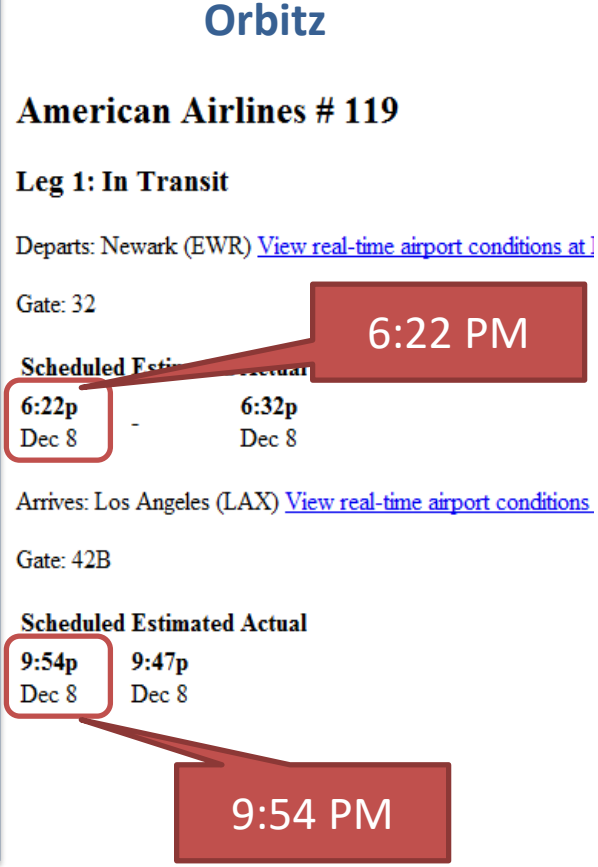
FlightView



FlightAware

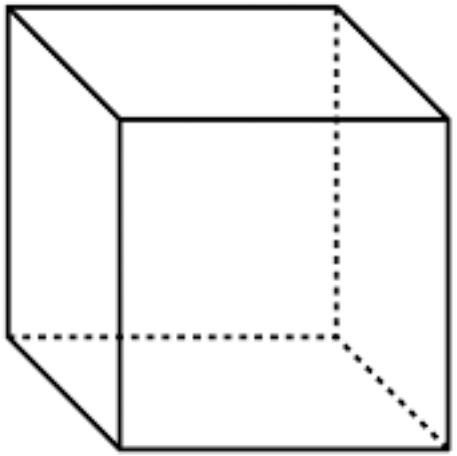


Orbitz



Quality Dimensions

- Software (it-CISQ.org)
  - Security
  - Reliability
  - Efficiency
  - Maintainability
  - …?
- Computer System
  - Throughput
  - Response time
  - Availability
  - …?



Dimension:

a central notion in Quality Domain

"a measureable extent of a particular kind"

Service quality dimensions [Russell and Taylor 2003]

Dimension	Definition
Time & Timeliness	Customer wait time, On-time completion
Completeness	Customers get all they ask for
Courtesy	Treatment by employees
Consistency	Same level of service for all customers
Accessibility and convenience	Ease of obtaining service
Accuracy	Performed correctly every time
Responsiveness	Reaction to special circumstances or requests

Product quality dimensions [Garvin 1987]

Dimension	Definition
Performance	The product's primary operating characteristic (such as acceleration, braking distance, steering, and handling of an automobile)
Features	The ``bells and whistles'' of a product (such as power option and a tape or CD deck of a car)
Reliability	The probability of a product's surviving over a specified period of time under stated conditions of use
Conformance	The degree to which physical and performance characteristics of a product match pre-established standards
Durability	The amount of use one gets from a product before it physically deteriorates or until replacement is preferable
Serviceability	The speed, courtesy, and competence of repair
Aesthetics	How a product looks, feels, sounds, tastes, or smells
Perceived quality	The subjective assessment of quality resulting from image, advertising, or brand names.

What is Data Quality?

- Degree to which data can be used for its intended purpose

- Degree to which data accurately represent the real-world

Dimensions of Data Quality

Completeness

Missing points on a trajectory

Accuracy

Postcode “4107” rather than “4017”

Freshness

Old telephone number

Consistency

ITEE vs. Information Technology and Electrical Engineering

Reliability

…

Wand &Wang (1996) Anchoring Data Quality Dimensions in Ontological Foundations. Comm. ACM 39/11. 1996

Shazia Sadiq, Naiem Khodabandehloo Yeganeh and Marta Indulska. 20 years of data quality research: Themes, trends and synergies. Proceedings of: The 22nd Australasian Database Conference (ADC 2011). Perth, WA, Australia, (1-10). 17-20 January 2011.

Dimensions of Data Quality

Data Completeness:

1) A measure of the availability and comprehensiveness of data compared to the total data universe or population of interest. [McGilvray, 2008]
2) A record exists for every Real-World Object or Event the Enterprise needs to know about. [English, 2009]
3) Quality of having all data that existed in the possession of the sender at time the data message was created. [ISO, 2012]
4) Completeness refers to the degree to which values are present in a data collection, as for as an individual datum is concerned, only two situations are possible: Either a value is assigned to the attribute in question or not. In the latter case, null, a special element of an attribute's domain can be assigned as the attribute's value. Depending on whether the attribute is mandatory, optional, or inapplicable, null can mean different things. [Redman, 1997]
5) Determined the extent to which data is not missing. For example, an order is not complete without a price and quantity.[Gatling et al, 2007]

+ 13 more …

A classification of data quality dimensions

User Independent

- Completeness of mandatory attributes
- Completeness of optional attributes
- Precision
- Business rules compliance
- Meta-data compliance
- Uniqueness
- Non-redundancy
- Semantic consistency
- Value consistency
- Format consistency
- Referential integrity

User Dependent

- Completeness of records
- Data volume
- Continuity of data access
- Data maintainability
- Data awareness
- Ease of data access
- Data punctuality
- Data access control
- Data timeliness
- Data freshness
- Accuracy to reference source
- Accuracy to reality
- Standards and regulatory compliance
- Statistical validity
- Source quality
- Objectivity
- Traceability
- Usefulness and relevance
- Understandability
- Appropriate presentation
- Interpretability
- Information value

good enough ≠ good data

Jayawardene, Vimukthi, Sadiq, Shazia and Indulska, Marta (2013). The curse of dimensionality in data quality. In Proceedings of the 24th Australasian Conference on Information Systems (ACIS 2013). Melbourne, VIC, Australia, (1-11). 4-6 December, 2013.

Whose problem is data quality?

- ❑ Management problem
- ❑ IT problem
- ❑ Computational/ Statistical problem
- ❑ All of the above

Ownership

- Who owns the data?
  - Possession, Responsibility, Power, or Control
- Who is liable if data is faulty
  - New legislation such as Data Transparency Act (DATA) - http://www.datacoalition.org
  - Open data initiatives - https://data.qld.gov.au
- Who profits from the value of data assets
  - How do you monetize data?

Consider a large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings.

Each of the subsidiaries may have its own partner suppliers along with item catalogs.

Consider the case that there is a large overlap of business with a particular supplier group, which may put LDC into a favorable bargaining position to negotiate significant discounts.

However, data differences do not reveal this position, and thus directly impact on the bottom line for LDC

1. Create a reference (synonym) table for suppliers
2. Load supplier data from all subsidiaries into the reference table
3. Use *matching* techniques to identify potential overlaps
4. Extract a master table for suppliers – represents a single version of truth
5. Retain original representations – represent multiple versions of truth
6. Allow access for subsidiaries to reference master data in all new (or update) transactions involving supplier data
7. Ensure data managers are accountable for continued master data checks
8. Introduce a periodic monitoring scheme

Consider a large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings.

Each of the subsidiaries may have its own partner suppliers along with item catalogs.

Consider the case that there is a large overlap of business with a particular supplier group, which may put LDC into a favorable bargaining position to negotiate significant discounts.

However, data differences do not reveal this position, and thus directly impact on the bottom line for LDC

Algorithms and Methods

1. Create a reference (synonym) table for suppliers
2. Load supplier data from all subsidiaries into the reference table
3. Use *matching* techniques to identify potential overlaps
4. Extra master table for suppliers – a single version of truth
5. inal representations – multiple versions of truth
6. Allow access for subsidiaries to reference master data in all new (or update) transactions involving supplier data
7. Ensure data managers are accountable for continued master data checks
8. Introduce a periodic monitoring scheme

Consider a large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings.

Each of the subsidiaries may have its own partner suppliers along with item catalogs.

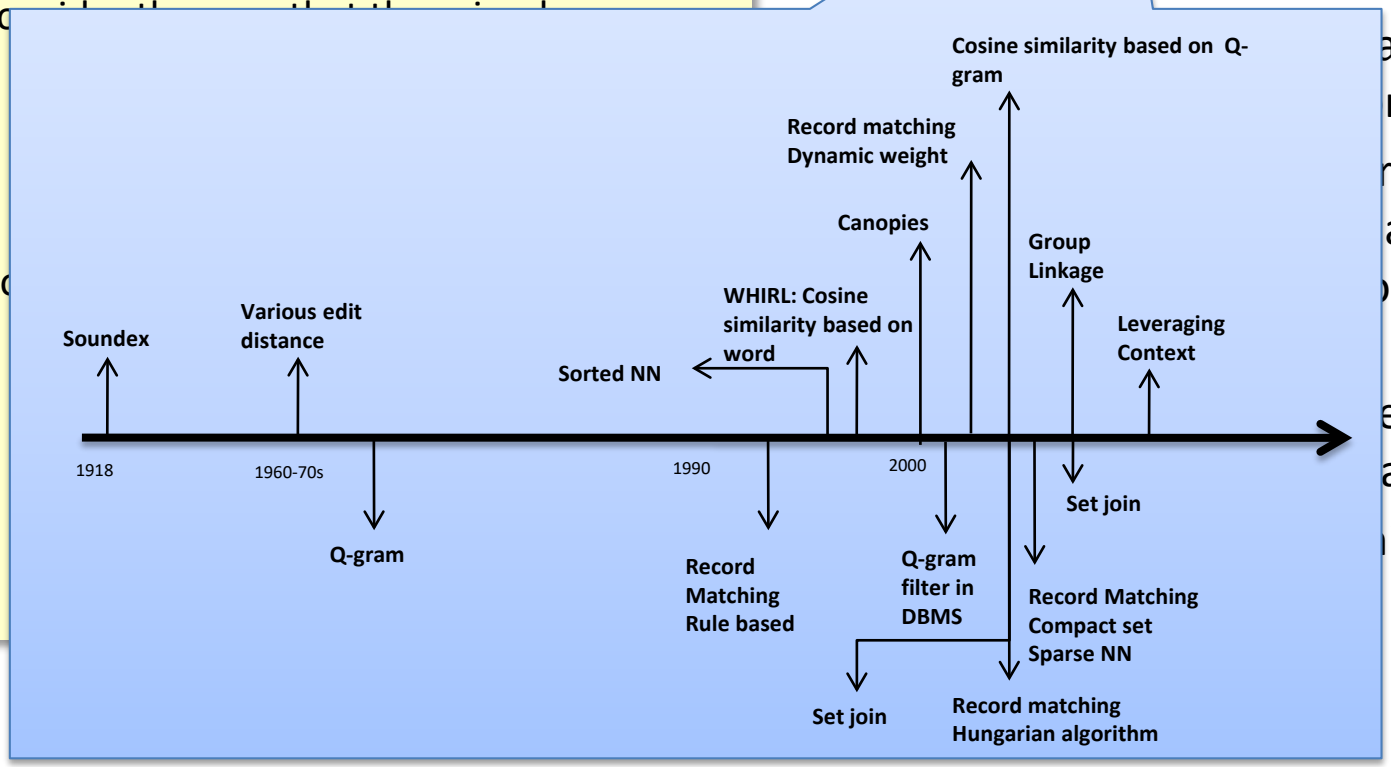
Co

Ho

1. Create a reference (synonym) table for suppliers
2. Load supplier data from all subsidiaries into the reference table
3. Use *matching* techniques to identify potenti erlaps
4. Ext ster table for suppliers – single version of truth

ations –
ns of truth
ries to
all new (or
olving supplier

e accountable
a checks
itoring scheme



Consider a large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings.

Each of the subsidiaries may have its own partner suppliers along with item catalogs.

Consider the case that there is a large overlap of business with a particular supplier group, which may put LDC into a favorable bargaining position to negotiate significant discounts.

However, data differences do not reveal this position, and thus directly impact on the bottom line for LDC

1. Create a reference (synonym) table for suppliers
2. Load supplier data from all subsidiaries into the reference table
3. Use *matching* techniques to identify potential overlaps
4. Extract a master table for suppliers – represents a single version of truth
5. Retain original representations – represent multiple versions of truth
6. Allow access for subsidiaries to reference master data in all new (or update) transactions involving supplier data
7. Ensure data managers for continued master d
8. Introduce a periodic monitoring scheme

Technology Infrastructure

Consider a large distribution company (LDC) that acquires two other distribution establishments, which will now form part of LDC operations as subsidiaries while maintaining their individual brandings.

Each of the subsidiaries may have its own partner suppliers along with item catalogs.

Consider the case that there is a large overlap of business with a particular supplier group, which may put LDC into a favorable bargaining position to negotiate significant discounts.

However, data differences do not reveal this position, and thus directly impact on the bottom line for LDC

1. Create a reference (synonym) table for suppliers
2. Load supplier data from all subsidiaries into the reference table
3. Use *matching* techniques to identify potential overlaps
4. Extract a master table for suppliers – represents a single version of truth
5. Retain original representations – represent multiple versions of truth
6. Allow access for su… reference master … update) transactio… …ier data
7. Ensure data managers are accountable for continued master data checks
8. Introduce a periodic monitoring scheme

Data Governance

Total Data Quality

Organizational	Development of data quality objectives for the organization and **strategies** to establish the people, processes, policies, and standards required to manage and ensure the data quality objectives are met
Architectural	The **technology** landscape required to deploy developed data quality management processes, standards and policies
Computational	Effective and efficient **methods** & techniques required to meet data quality objectives

Develop the capacity to understand how the quality of data affects the quality of the insight we derive from it

J. Stoyanovich; and F. Suchanek. *The elephant in the room: getting value from Big Data*. Available: http://wp.sigmod.org/?p=1519 (accessed on 24/2/2015)

Data Quality

- Poor quality data costs …
  - "$3 trillion to US government"
  - "$611 billion to US business for customer data alone"

You have to start with a very basic idea: **Data is super messy**, and data cleanup will always be literally 80% of the work. In other words, data is the problem.

"If you take something like LinkedIn in the early days, let's say, there were 4,000 variations of how people said they worked at IBM — IBM, IBM Research, Software Engineer, all the abbreviations, etc.," says Patil.

First US Chief Data Scientist at the White House

http://www.datablueprint.com/cost-of-bad-data-by-the-numbers/

How can you accelerate the time to value from big data in the presence of data quality problems?

Find out if your data is fit for use

- Data Exploration
  - Discovering and understanding the quality characteristics of the data through exploratory techniques
- Data Transformation
  - Transforming the data through cleaning, curating, repairing
- Data Enrichment
  - Enriching the data through data imputation and integration

POLL QUESTIONS – DATA QUALITY

DATA7001
INTRODUCTION TO DATA SCIENCE

Module 3 Is my Data Fit for Use

Module Topics

- What is Data Quality
- **Data Exploration**
  - **Discovering and understanding the quality characteristics of the data through exploratory techniques**
- Data Transformation
  - Transforming the data through cleaning, curating, repairing
- Data Enrichment
  - Enriching the data through data imputation and integration

Data Exploration

- EDA (exploratory data analysis) is an approach to summarize and discover key data characteristics, typically using visual methods.
- Example: plots of wage against a few other variables

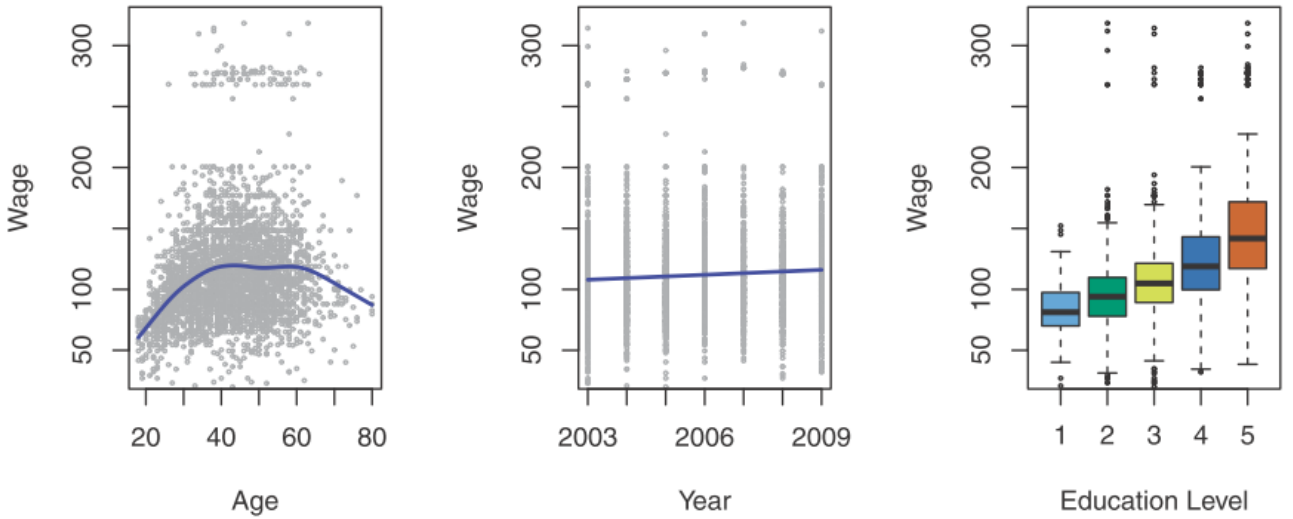


Figure reproduced from: G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning: with Applications in R. Corrected 6th printing, 2015. Springer, New York.

EDA is about Exploration

- Purposes of EDA
  - Suggest hypotheses for observed patterns
  - Assess assumptions for statistical learning
  - Support selection of appropriate statistical techniques
  - Guide further data collection

- EDA vs CDA (confirmatory data analysis)
  - CDA aims to confirm a hypothesis (e.g. via statistical hypothesis testing), rather than how to generate a hypothesis
  - EDA emphasize on coming up with a hypothesis
- EDA vs IDA (initial data analysis)
  - IDA: checking assumptions, transformation, imputation
  - IDA is part of EDA

We Need Both Exploratory and Confirmatory

JOHN W. TUKEY*

We often forget how science and engineering function. Ideas come from previous exploration more often than from lightning strokes. Important questions can demand the most careful planning for confirmatory analysis. Broad general inquiries are also important. Finding the question is often more important than finding the answer. Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught. Confirmatory data analysis, by contrast, is easier to teach and easier to computerize. We need to teach both; to think about science and engineering more broadly; to be prepared to randomize and avoid multiplicity.

2. How are designs guided? (Usually, by the best qualitative and semiquantitative information available, obtained by exploration of past data.)
3. How is data collection monitored? (By exploring the data, often as they come in, for unexpected behavior.)
4. How is analysis overseen; how do we avoid analysis that the data before us indicate should be avoided? (By exploring the data—before, during,

The Sin of Over-exploration

- Data Dredging: finding spurious patterns

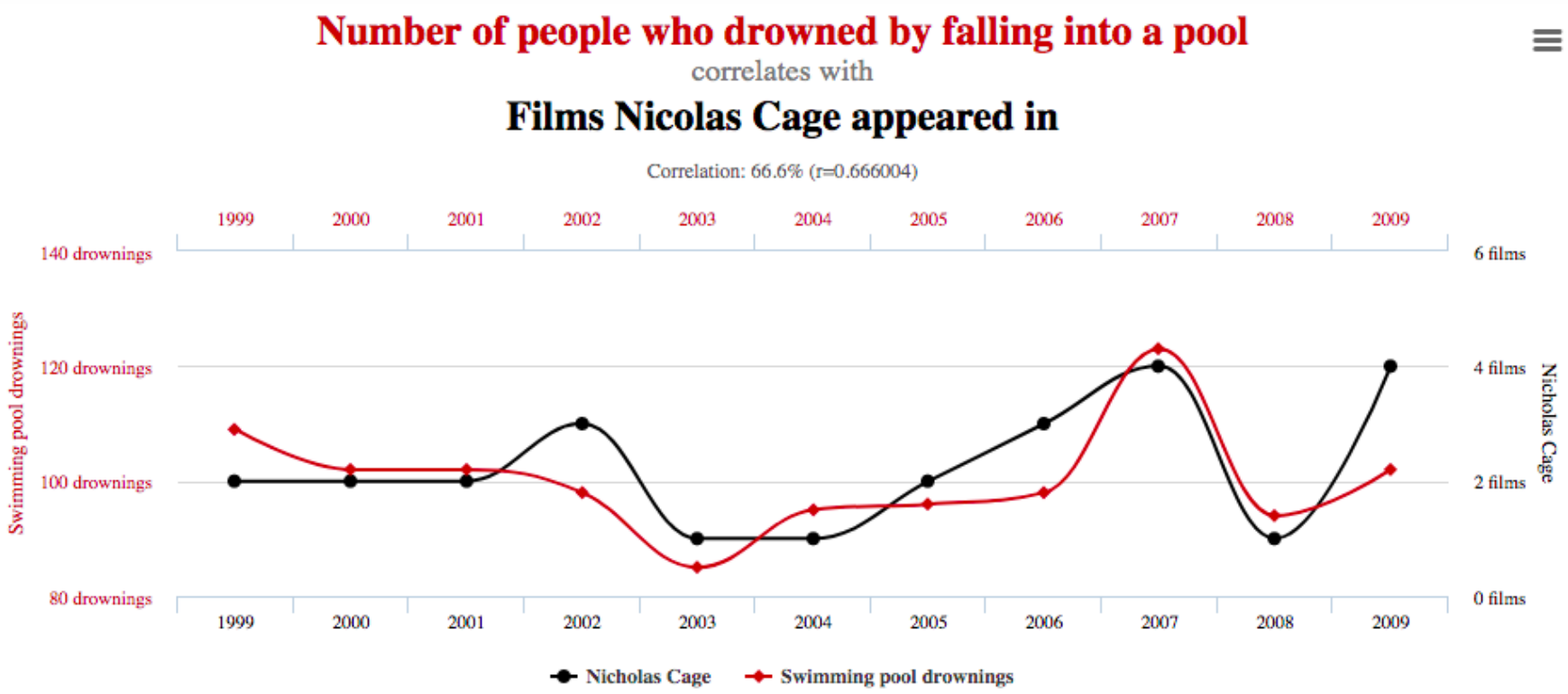


Figure from http://tylervigen.com

Data Exploration – Variables

- Variable Identification
  - Eg. Age, Wage, Year, Education Level

- Variable Type
  - Quantitative (numerical; discrete vs. continuous)
  - Qualitative (words, categories)

Data Exploration – Variables

	A	B	C	D	E	F	G	H	I	J	K
1	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales	salary	
2	0.38	0.53	2	157	3	0	1	0	sales	low	
3	0.8	0.86	5	262	6	0	1	0	sales	medium	
4	0.11	0.88	7	272	4	0	1	0	sales	medium	
5	0.72	0.87	5	223	5	0	1	0	sales	low	
6	0.37	0.52	2	159	3	0	1	0	sales	low	
7	0.41	0.5	2	153	3	0	1	0	sales	low	
8	0.1	0.77	6	247	4	0	1	0	sales	low	
9	0.92	0.85	5	259	5	0	1	0	sales	low	
10	0.89	1	5	224	5	0	1	0	sales	low	
11	0.42	0.53	2	142	3	0	1	0	sales	low	
12	0.45	0.54	2	135	3	0	1	0	sales	low	
13	0.11	0.81	6	305	4	0	1	0	sales	low	
14	0.84	0.92	4	234	5	0	1	0	sales	low	
15	0.41	0.55	2	148	3	0	1	0	sales	low	
16	0.36	0.56	2	137	3	0	1	0	sales	low	
17	0.38	0.54	2	143	3	0	1	0	sales	low	
18	0.45	0.47	2	160	3	0	1	0	sales	low	
19	0.78	0.99	4	255	6	0	1	0	sales	low	
20	0.45	0.51	2	160	3	1	1	1	sales	low	
21	0.76	0.89	5	262	5	0	1	0	sales	low	
22	0.11	0.83	6	282	4	0	1	0	sales	low	
23	0.38	0.55	2	147	3	0	1	0	sales	low	
24	0.09	0.95	6	304	4	0	1	0	sales	low	
25	0.46	0.57	2	139	3	0	1	0	sales	low	
26	0.4	0.53	2	158	3	0	1	0	sales	low	
27	0.89	0.92	5	242	5	0	1	0	sales	low	
28	0.82	0.87	4	239	5	0	1	0	sales	low	
29	0.4	0.49	2	135	3	0	1	0	sales	low	
30	0.41	0.46	2	128	3	0	1	0	accounting	low	
31	0.38	0.5	2	132	3	0	1	0	accounting	low	
32	0.09	0.62	6	294	4	0	1	0	accounting	low	
33	0.45	0.57	2	134	3	0	1	0	hr	low	
34	0.4	0.51	2	145	3	0	1	0	hr	low	
35	0.45	0.55	2	140	3	0	1	0	hr	low	
36	0.84	0.87	4	246	6	0	1	0	hr	low	
37	0.1	0.94	6	255	4	0	1	0	technical	low	
38	0.38	0.46	2	137	3	0	1	0	technical	low	
39	0.45	0.5	2	126	3	0	1	0	technical	low	

HR_comma_sep

Task and Discussion

Give an example variable for the following types:
(1) Categorical; (2) Discrete; (3) Continuous.

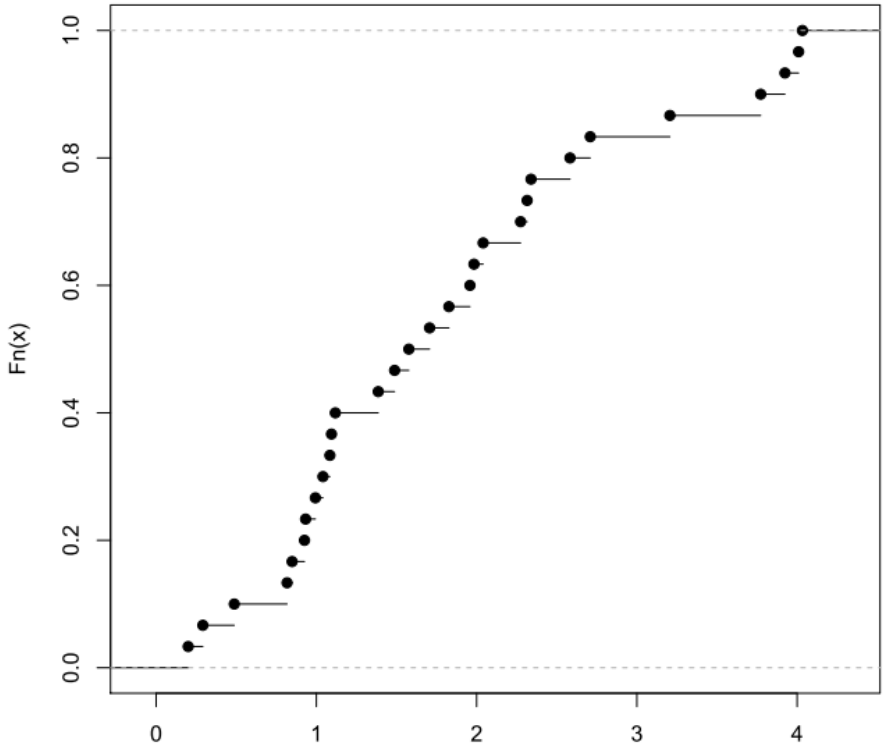
EDA Techniques

- EDA makes use of both visual and quantitative techniques.
- Basic visualisation techniques
  - Univariate: empirical cdf*, histogram, box plot
  - Bivariate: time series plot, scatter plot
- Quantitative techniques
  - Summary statistics: mean, quantiles, …
  - Dimension reduction techniques: e.g. PCA, …

*cdf: cumulative distribution function

Empirical cdf Plot

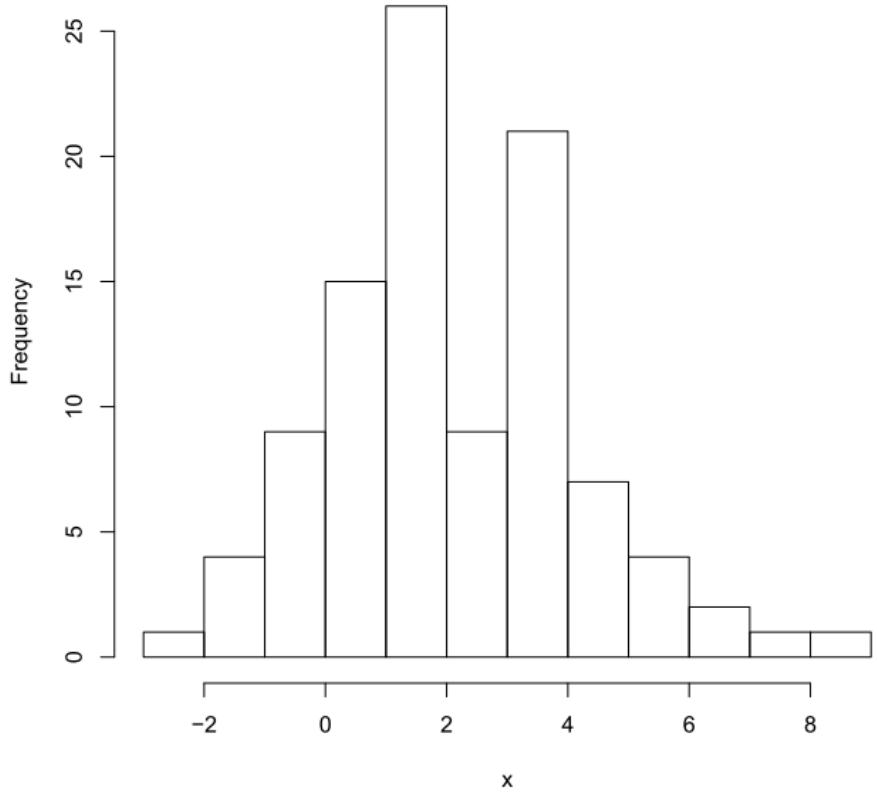
Displays the proportion of observations not more than x
- e.g. roughly 60% observations not more than 2 in the plot below



Histogram

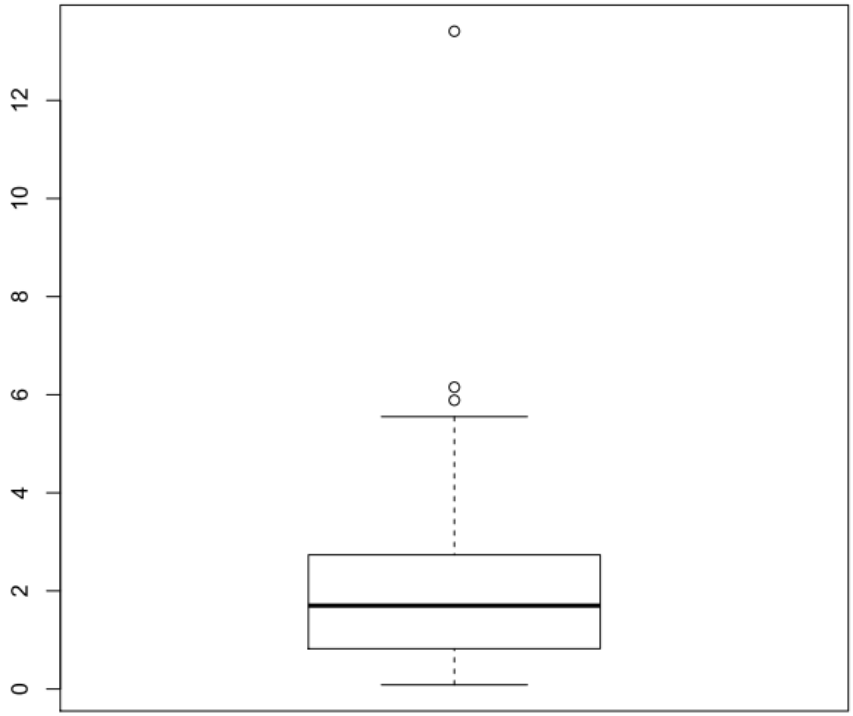
Plots counts of data falling into a set of *bins*

- e.g. 15 observations fall into the bin [0, 1] in the histogram below



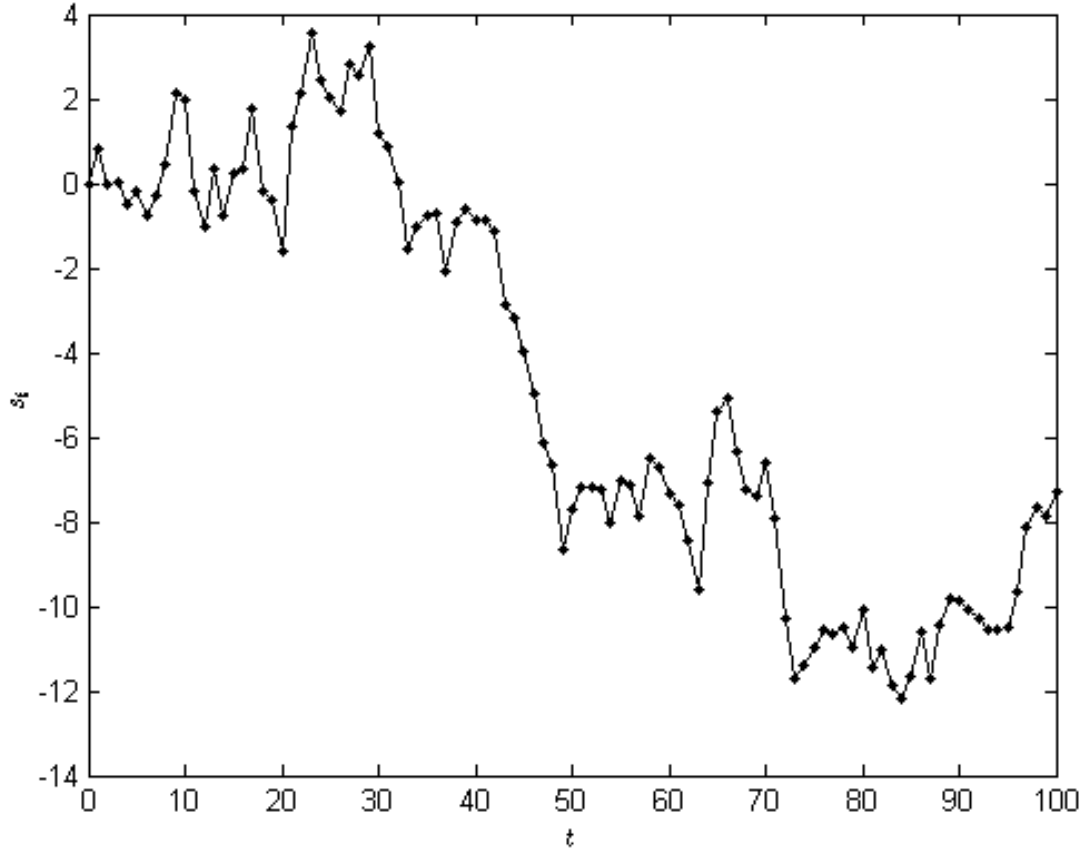
Box plot

Visualises
- a five-number summary: median, first- and third quartiles (the box); min and max of "typical observations" (the whiskers)
- and "unusual observations";



Time series plot

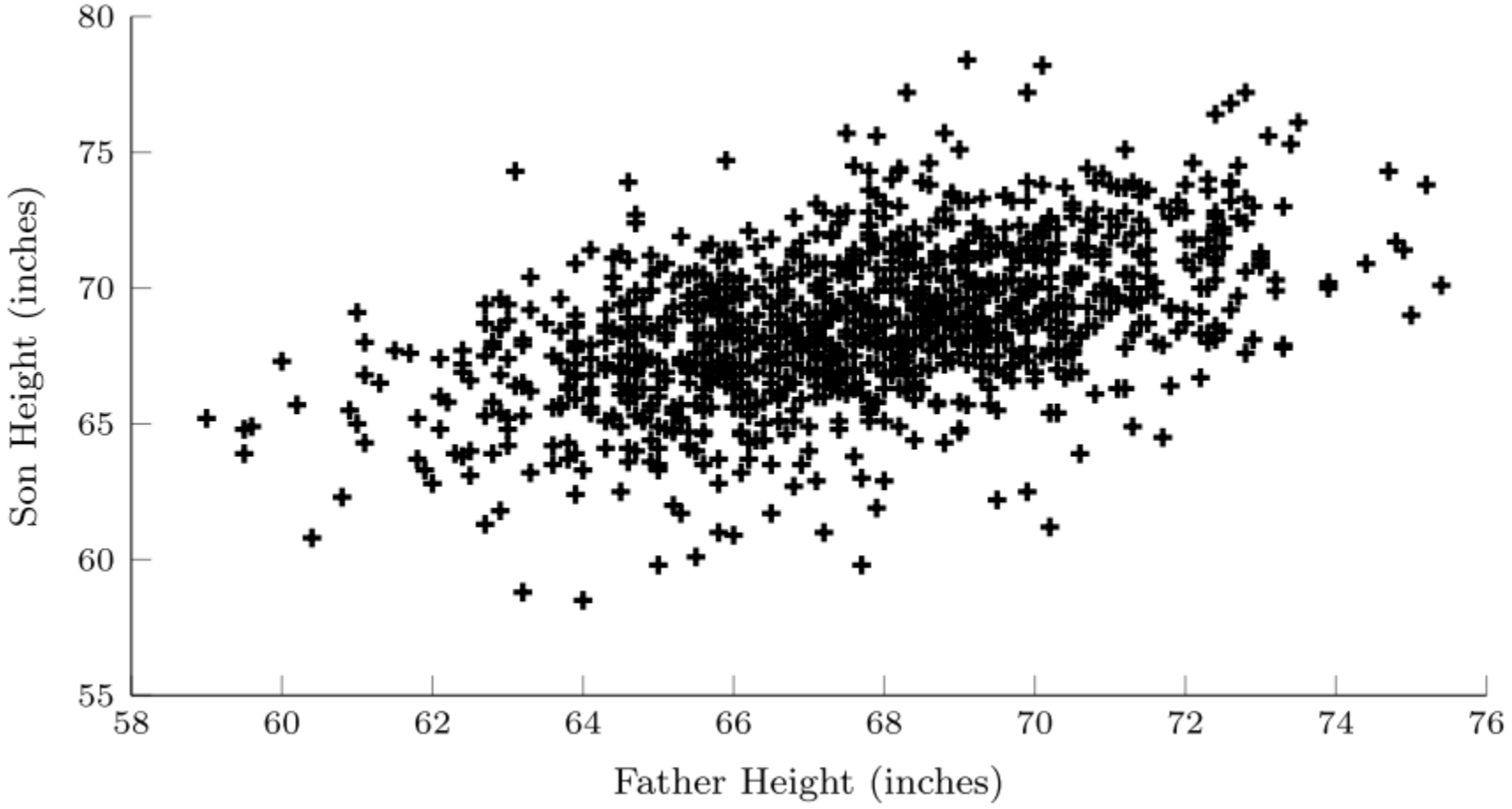
Plots a variable evolving over time (e.g. daily stock price)



Scatter plot

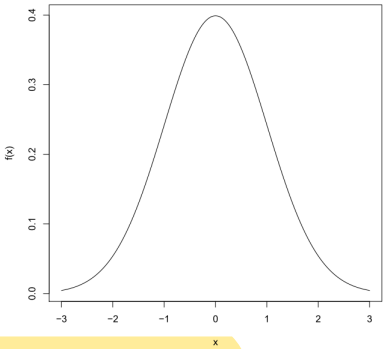
Plots pairs of observed variables

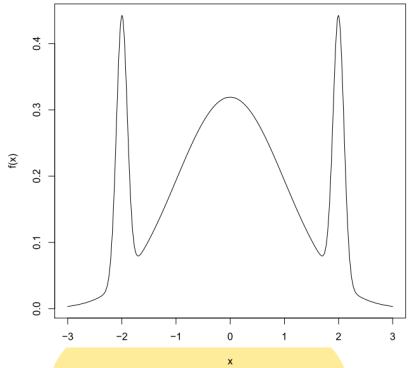
- Useful for determining whether two variables are correlated with each other
- e.g. the plot below shows son's heights are positively correlated with father's heights



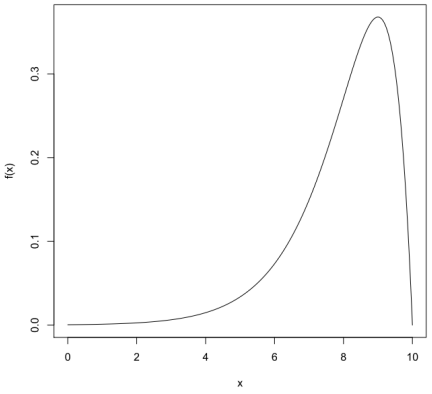
Shape Character of Numerical Data

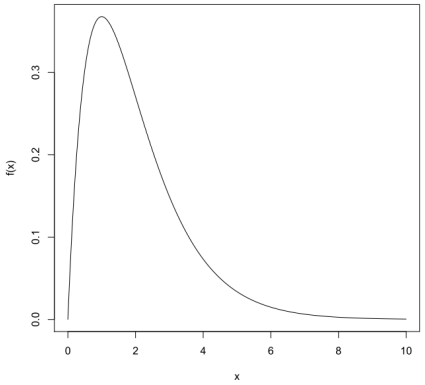
- Unimodal vs Multimodal





- Symmetric vs [left/right]-Skewed





Data Exploration – Missing Data

- Most real data sets contain missing values
  - E.g. nonresponse to survey questions for privacy concerns
  - E.g. a respondent missed a question

- To appropriately deal with missing data, it helps to understand *why* data is missing.

- Rubin's missing data mechanisms:
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Missing not at Random (MNAR)

Missing Completely at Random

- Missing mechanism does not depend on values of any (observed or unobserved) variables
  - E.g. (contrived) respondents answer a survey question if tossing a coin gives a head
  - Rarely happens in reality

- Complete cases
  - Can be considered a simple random sample from target population
  - Analysis based on them does not introduce any bias

Missing at Random

- Probability of missing for a particular variable depends only on observed variables

- Missing Categorical data: Add an extra category to indicate missingness!

- Imputation (Simple Random, Model-based,…)

Missing not at Random

- Probability of missing for a particular variable depends on other, unobserved Inputs.
  - Imputation, Augmentation

- Probability of missing for a particular variable *depends* on the value of the missing variable itself!
  - Model-based methods

Task and Discussion

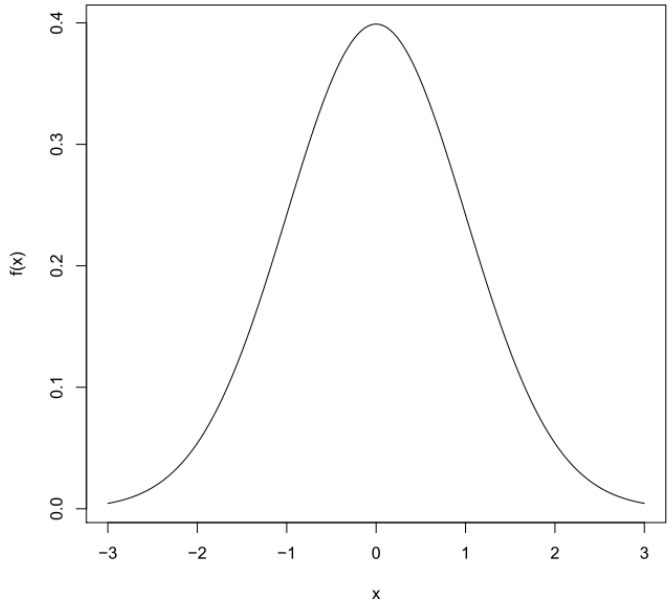
Give an example of (1) MCAR; (2) MAR; (3) MNAR.

Data Exploration – Unusual Data

- Often seek to identify unusual observations
- Outliers: Deviate significantly from statistical model assumptions
  - Erroneous Data (Sampling, Measurement, …)
    - Sometimes correctable (eg. Height in metres vs millimetres)
  - Natural Outliers
    - May indicate inappropriate model assumptions
- Influential observations: inclusion/exclusion significantly affects statistical analyses
  - NB: Influential observations may or may not be outliers

Basic Error Model Assumption

- A common assumption underpinning many statistical methods is that random variability is *normally distributed*.



Basic Outlier Detection

- For *normally distributed* data, slightly less than 1% of probability density lies outside the range: (Q1-1.5 × IQR, Q3+1.5 × IQR)**

- Consequently, a frequently applied rule is to flag any observations outside this range as outliers. [cf. box plot]

** Q1: First Quartile; Q3: Third Quartile; IQR: Inter-Quartile Range

What to do with Outliers?

- Examine them more closely; are they erroneous or natural?
- If erroneous, can treat as missing data
- If natural, should use *robust* statistical methods [trimmed mean, …]

POLL QUESTIONS – DATA EXPLORATION

DATA7001
INTRODUCTION TO DATA SCIENCE

Module 3 Is my Data Fit for Use

Module Topics

- What is Data Quality
- Data Exploration
  - Discovering and understanding the quality characteristics of the data through exploratory techniques
- **Data Transformation**
  - **Transforming the data through cleaning, curating, repairing**
- **Data Enrichment**
  - **Enriching the data through data** imputation **and integration**

Data Curation

Data curation is a broad term used to indicate (principled and controlled) processes and activities related to

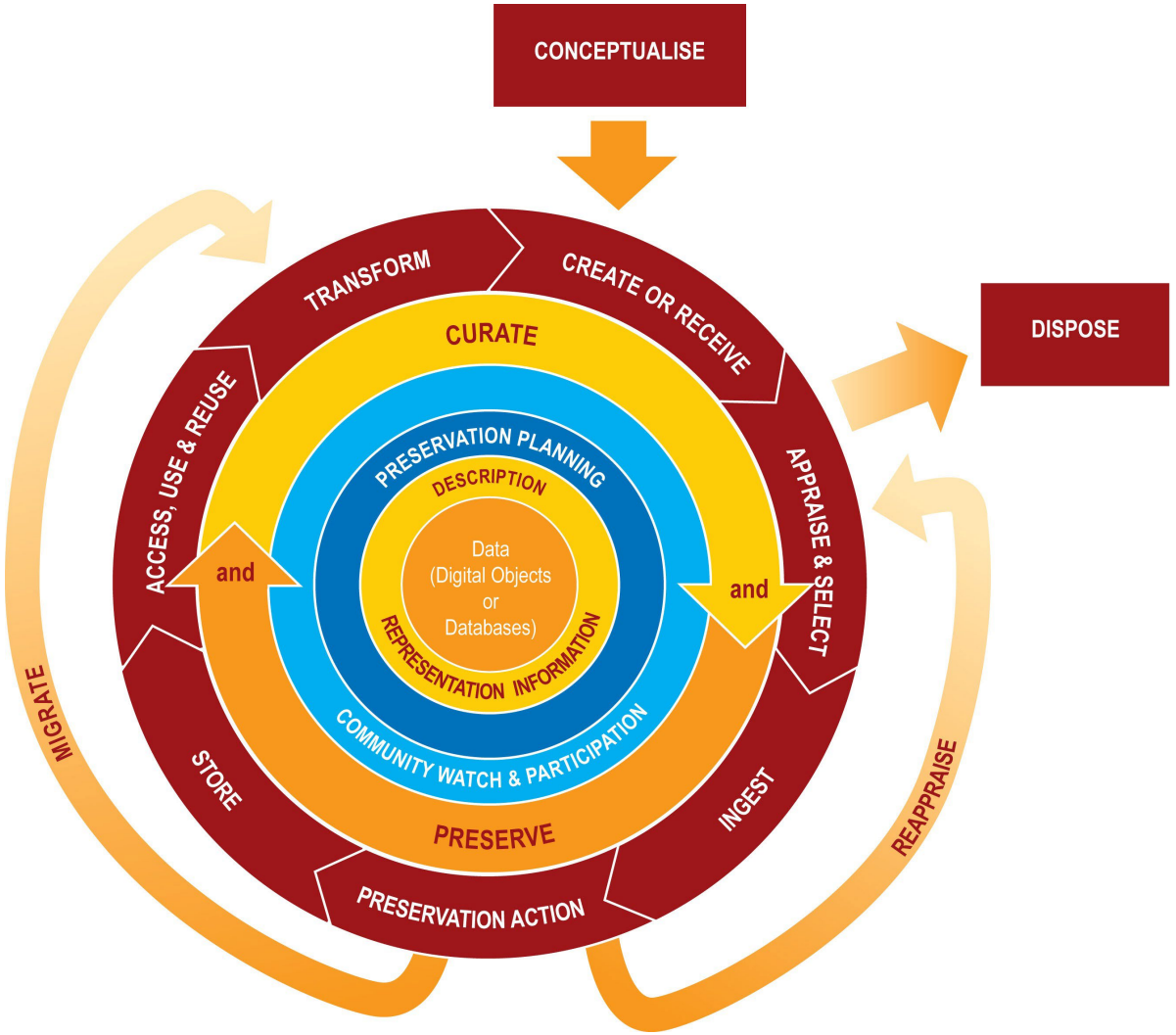
- the organization and integration of data collected from various sources,
- annotation of the data for documentation as well as lineage tracing purposes,
- And publication (storage) and presentation (release) of the data

such that the value of the data is maintained over time, and the data remains available for reuse and preservation.

For Big Data, curation processes have to scale and hence the focus is on software and tools to process high volume and complex data

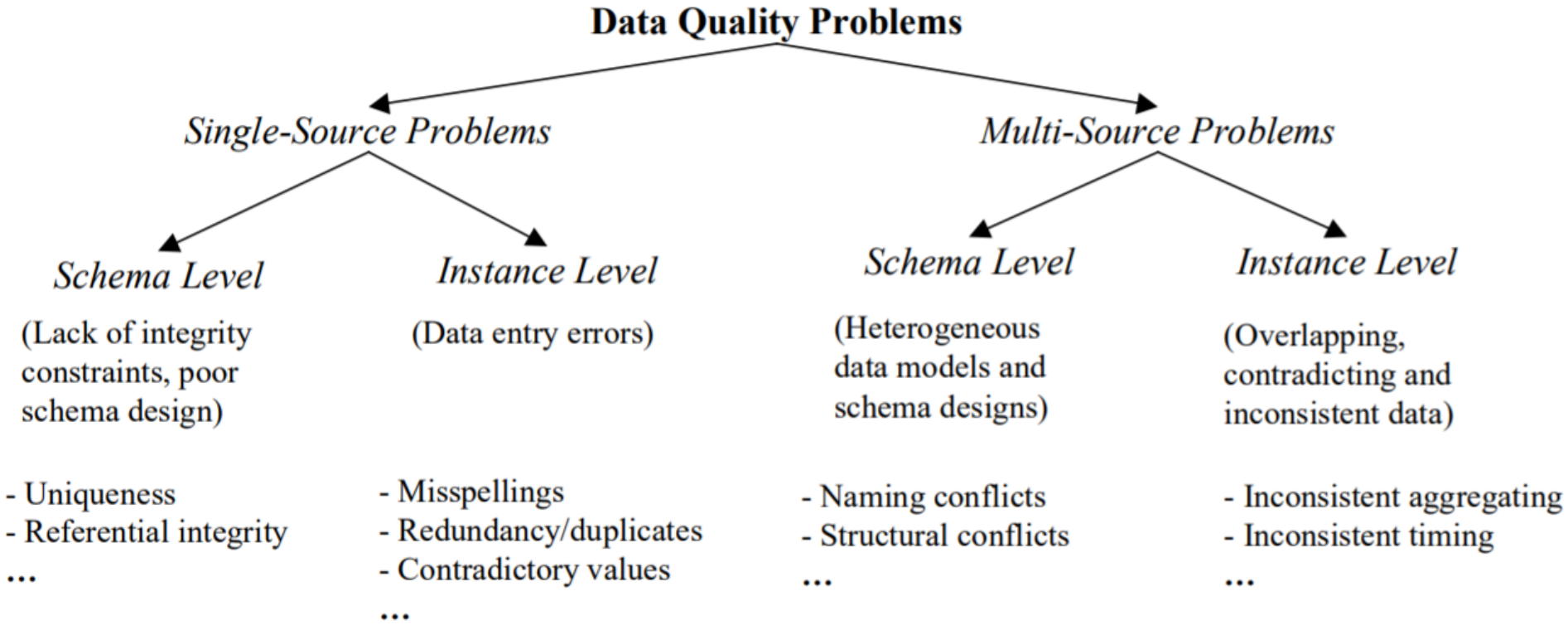
https://en.wikipedia.org/wiki/Data_curation

Data Curation is complex



Higgins, Sarah. "The DCC curation lifecycle model." (2008).

Data Cleaning



Rahm, Erhard, and Hong Hai Do. "Data cleaning: Problems and current approaches." *IEEE Data Eng. Bull.* 23.4 (2000): 3-13.

Data Cleaning (in a nutshell)

- Cleaning from rules
  - Remove personnel whose height is recorded as 0.
- Cleaning from filter
  - Don’t use old data, only use data from last year.
- Cleaning from source
  - Go and check why in the original data source height of a personnel is recorded as 0, then fix it.

Data Transformation Tools

Data Preparation tools for data exploration and transformation

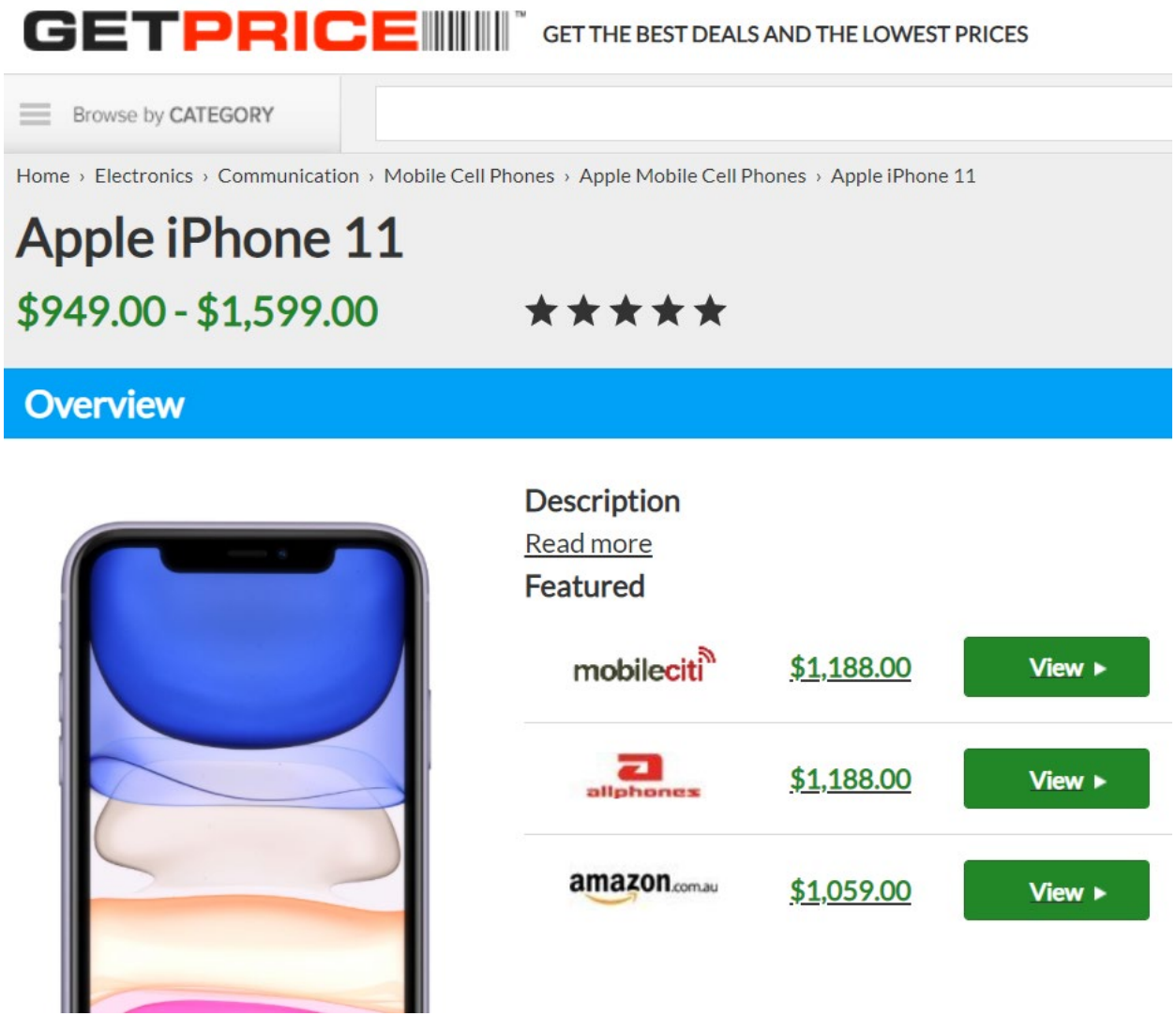
- MS PowerBI
- Tableau Desktop
- Qlik Sense
- Talend Data Preparation
- Trifacta
- Tamr
- SAP Analytics Cloud
- Datameer
- Informatica Enterprise Data Preparaiton
- IBM Cloud Park for Data
- …

Data Enrichment

- **Integration**
  - **Combine multiple data together**

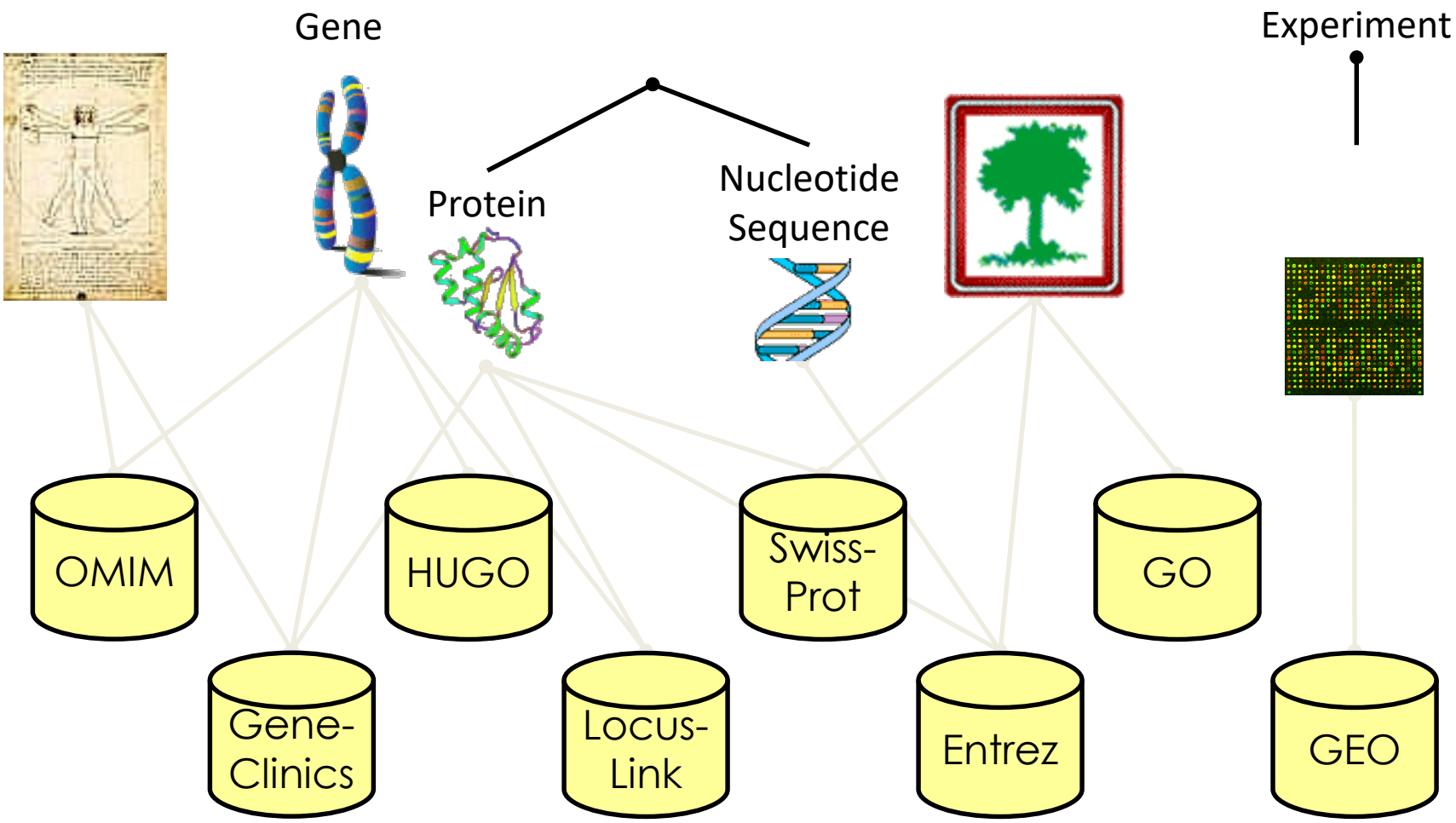
- Imputation
  - Impute missing values in the data

Why Data Integration? $$$



getprice.com.au

Why Data Integration? Science



Doan, AnHai, Alon Halevy, and Zachary Ives. *Principles of data integration*. Elsevier, 2012.

Data Integration

- Schema Level
  - Structural differences
  - Semantic differences
- Instance Level
  - Field linking (or matching)
  - Record linking
  - Entity linking

Example (schema level)

Consider two companies data models :

Company 1 records are stored in one table Emp;

Emp(Emp#, Fname, Lname, Bdate, Dept#, Rank, Salary)

Company 2 records are stored in many tables – one for each company department;

Dept XX(S-Id, Fname, Sname, Position, Phone#, e-mail, URL)

Build an integrated schema – number of changes must be made to at least one of component schemata and consequently to the local Application Programs.

Example (semantic level)

Consider you are rich and have bought both Woolworth and Coles and you want to integrate them:

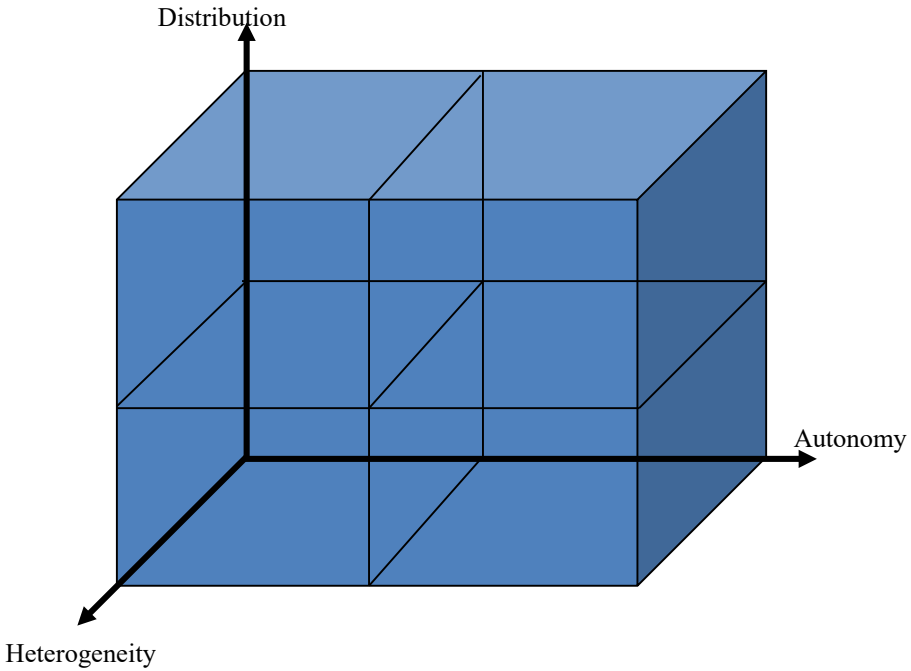
Coles stores prices in **cents** and **Woolworths** in **dollars**;
Product(id, name, price)
From Coles db: (1, delicious sourdough bread, 600)
From Woolworths db: (17, delicious sourdough bread, 6.1)

To integrate the two in one table/schema, you need to convert the semantics of data so that everything is in cents or dollars.

Data Integration (in databases)

More in INFS7907

- Autonomy (A)
  - A0: Tight integration
  - A1: Semi-autonomous
  - A2: Total isolation
- Distribution (D)
  - D0: Central
  - D1: Client Server
  - D2: Peer-to-peer
- Heterogeneity (H)
  - H0: Homogeneous
  - H1: Heterogeneousc



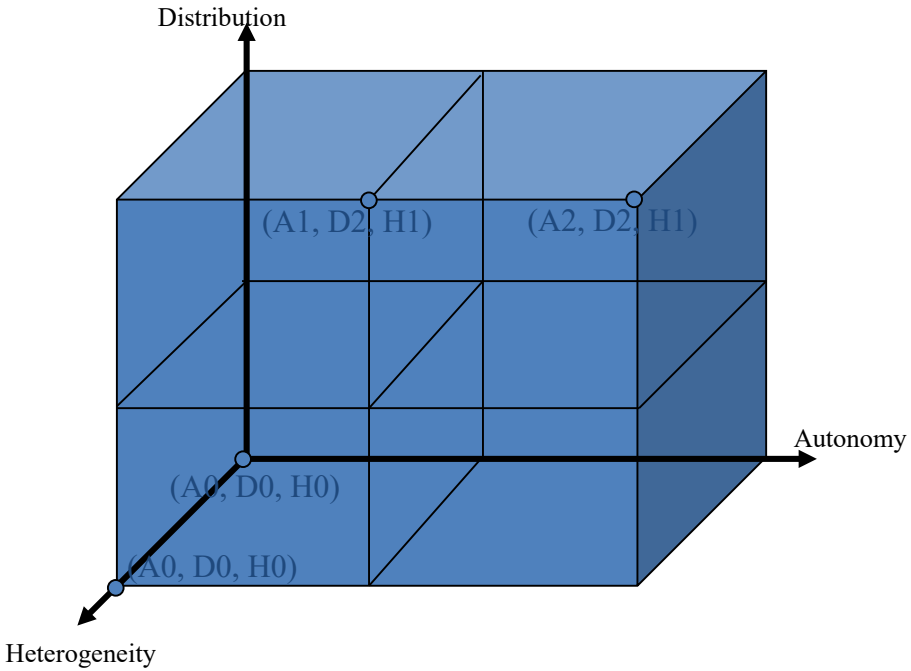
3*3*2 = 18 Alternatives

Some alternatives are meaningless or not practical!

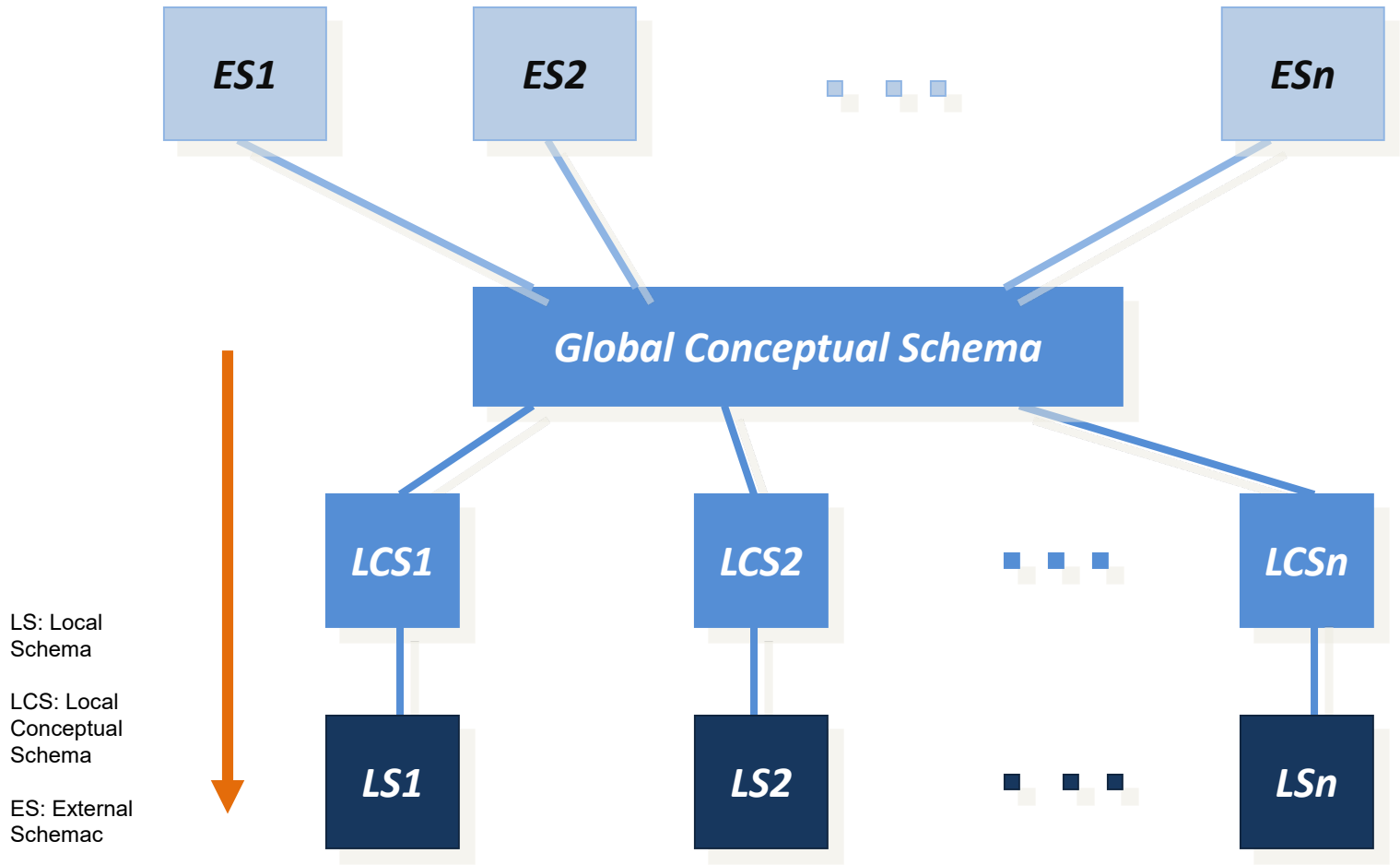
M. T. Ozsu, & P. Valduriez (1999) Principles of Distributed Distributed Database Systems. Prentice Hall

Major Alternatives

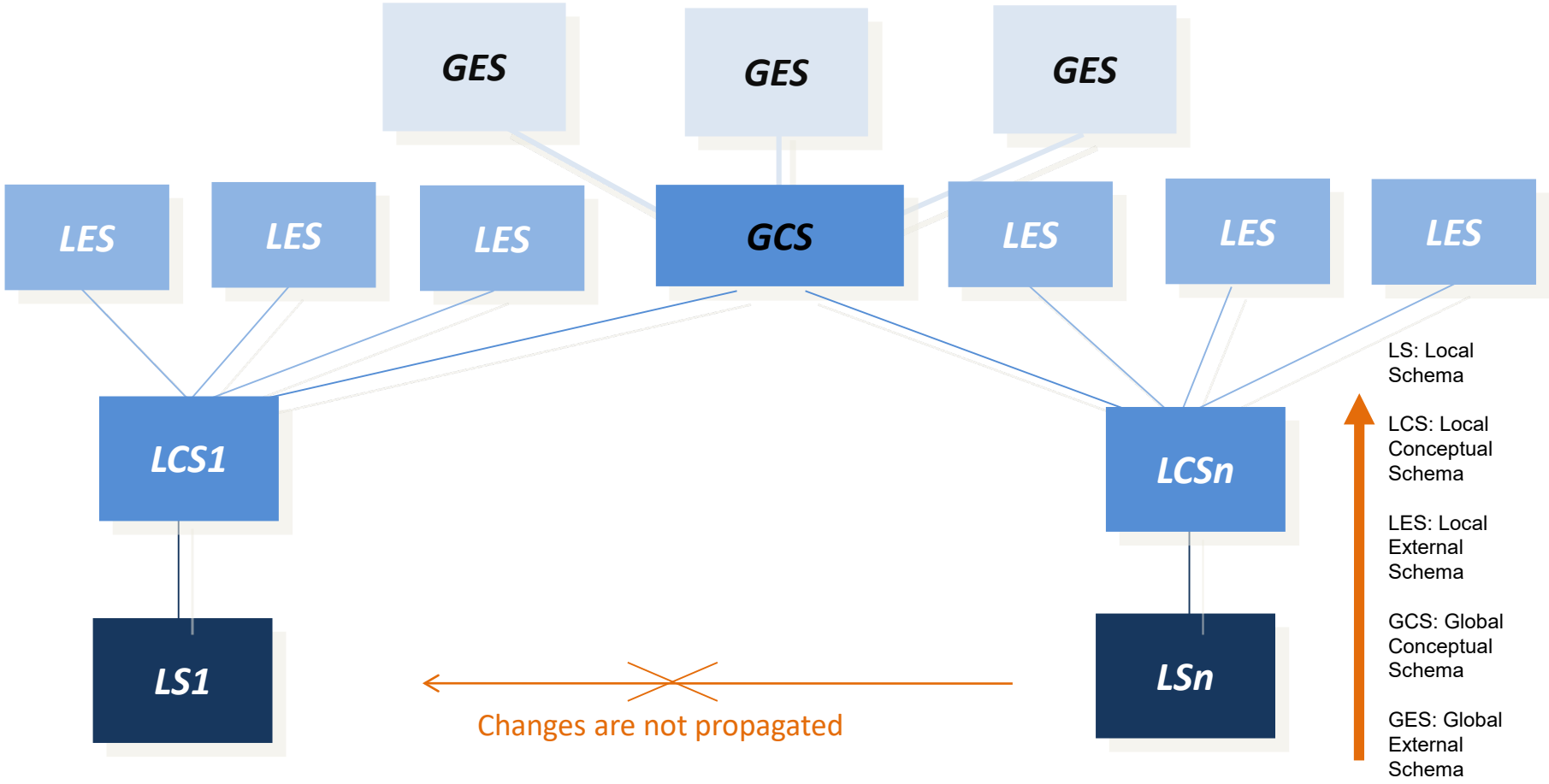
- (A0, D0, H0) Monolithic
- (A0, D0, H1) Data Warehouses
- (A1, D1/2, H1) Data Federations
- (A2, D2, H1) P2P Systems



Typical Architecture of Distributed Database



Typical Architecture of Global Schema Multi-Database



Example (instance level)

Value linking

John Smith| Jhn Smith | Mr J Smith | …

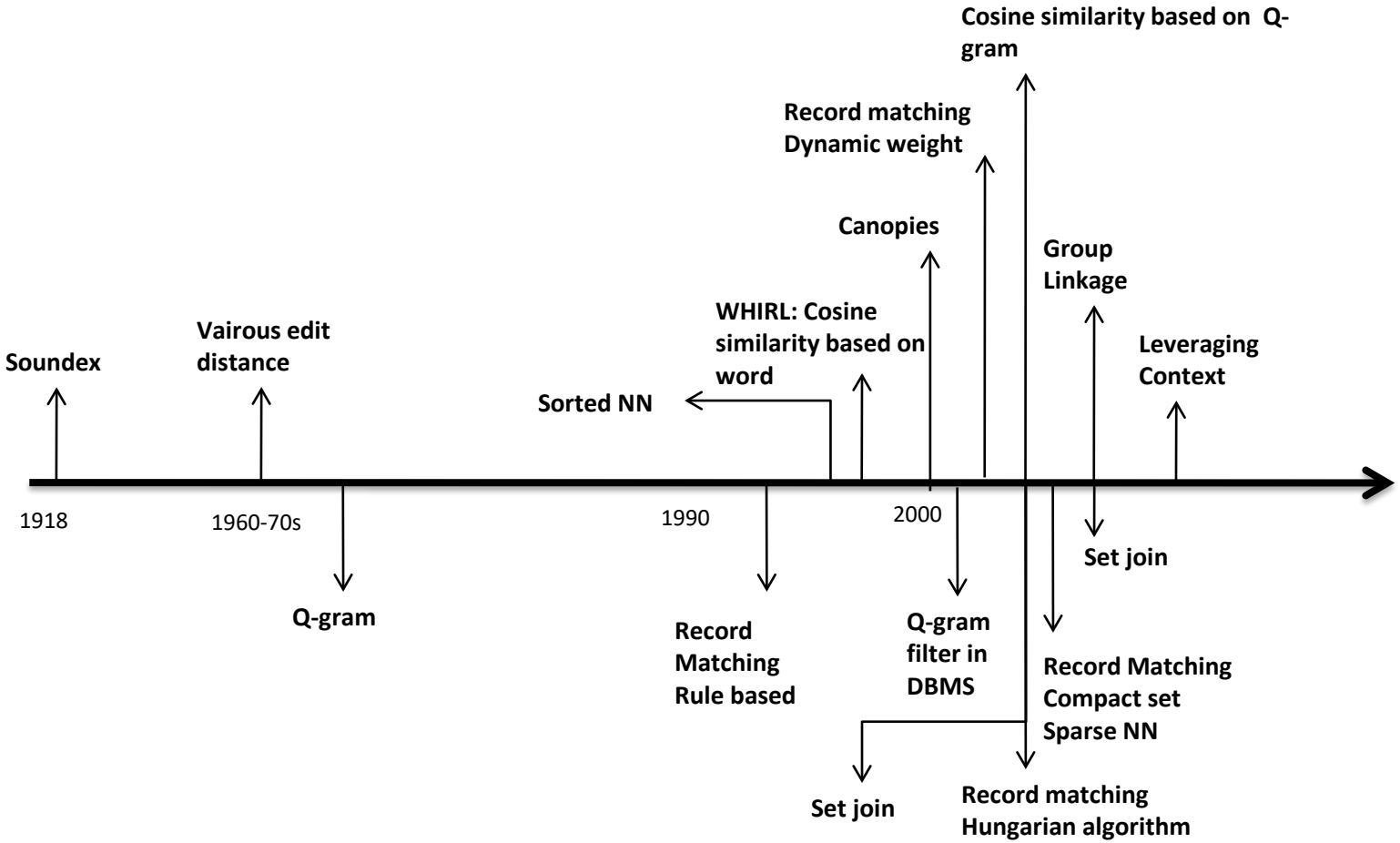
Record linking

Name	Address
Jenny Tylor	16, Finance St, South Bank, QLD
Jennifer Taylor	16 Finance Street, South Bank QLD 4077

Entity linking

Tiger Woods| Tiger Balm | Indian Tiger | …

Field Matching



Edit Distance

Classical distance function for string matching. Also known as Levenshtein Distance

The edit distance between two strings is the minimum number of operations to transform one string to another

- Operations: delete, insert or replace one character

What's the edit distance?

- 'John', 'Jon'
- 'John' , 'Jhn'
- 'John', 'Josh'

Task

What is the edit-distance between

- Shoe and Show
- Broad and Brand
- Kitten and Sitting

Edit Distance

X=dva y=dave

find Levenstein distance using dynamic programming.

		y0	y1	y2	y3	y4
			d	**a**	**v**	**e**
x0		0	1	2	3	4
x1	**d**	1	0	1		
x2	**v**	2				
x3	**a**	3				

		y0	y1	y2	y3	y4
			d	**a**	**v**	**e**
x0		0	1	2	3	4
x1	**d**	1	0	1	2	3
x2	**v**	2	1	1	1	2
x3	**a**	3	2	1	2	2

x = d – v a
| | | |
y = d a v e

substitute a with e
insert a (after d)

$$d(i,j) = \min \begin{cases} d(i-1,j-1) & \text{if } x_i = y_j \quad // \text{ copy} \\ d(i-1,j-1) + 1 & \text{if } x_i <> y_j \quad // \text{ substitute} \\ d(i-1,j) + 1 & // \text{ delete } x_i \\ d(i,j-1) + 1 & // \text{ insert } y_j \end{cases}$$

Doan, AnHai, Alon Halevy, and Zachary Ives. *Principles of data integration*. Elsevier, 2012.

Edit Distance

- Two strings are considered as the same (similar enough) if their ED is less than a pre-defined threshold
- May be costly operation for large strings
- Suitable for common typing mistakes
  - Comprehensive vs Comprenhensive
- Problematic for specific domains
  - AT&T Corporation vs AT&T Corp and
    IBM Corporation vs AT&T Corporation
  - What about numeric domains 500 vs {499, 800}
- Many variations and refinements exist (more in INFS7907)

Data Integration

- Schema Level
  - Structural differences
  - Semantic differences
- Instance Level
  - Field linking (or matching)
  - Record linking
  - Entity linking

In big data environment data is …

often external to the organization,

designed and created for a purpose other than it is being used for,

of a variety of data types,

and there is little or no knowledge of schema or data quality

How do you integrate is such an environment?

POLL QUESTIONS – DATA TRANSFORMATION AND INTEGRATION

DATA7001
INTRODUCTION TO DATA SCIENCE

Module 3 Is my Data Fit for Use

Data Enrichment

- Integration
  - Combine multiple data together

- **Imputation**
  - **Impute missing values in the data**

Data Imputation

- Imputation is not prediction
  - **The goal is not to re-create lost data**. Rather, to obtain statistically valid inferences from incomplete data.

- Recall Rubin's missing data mechanisms (MCAR, MAR, MNAR)
  - Impossible to prove data is MAR; assumption;

- We discuss only some of the simplest imputation methods.

Mean Imputation

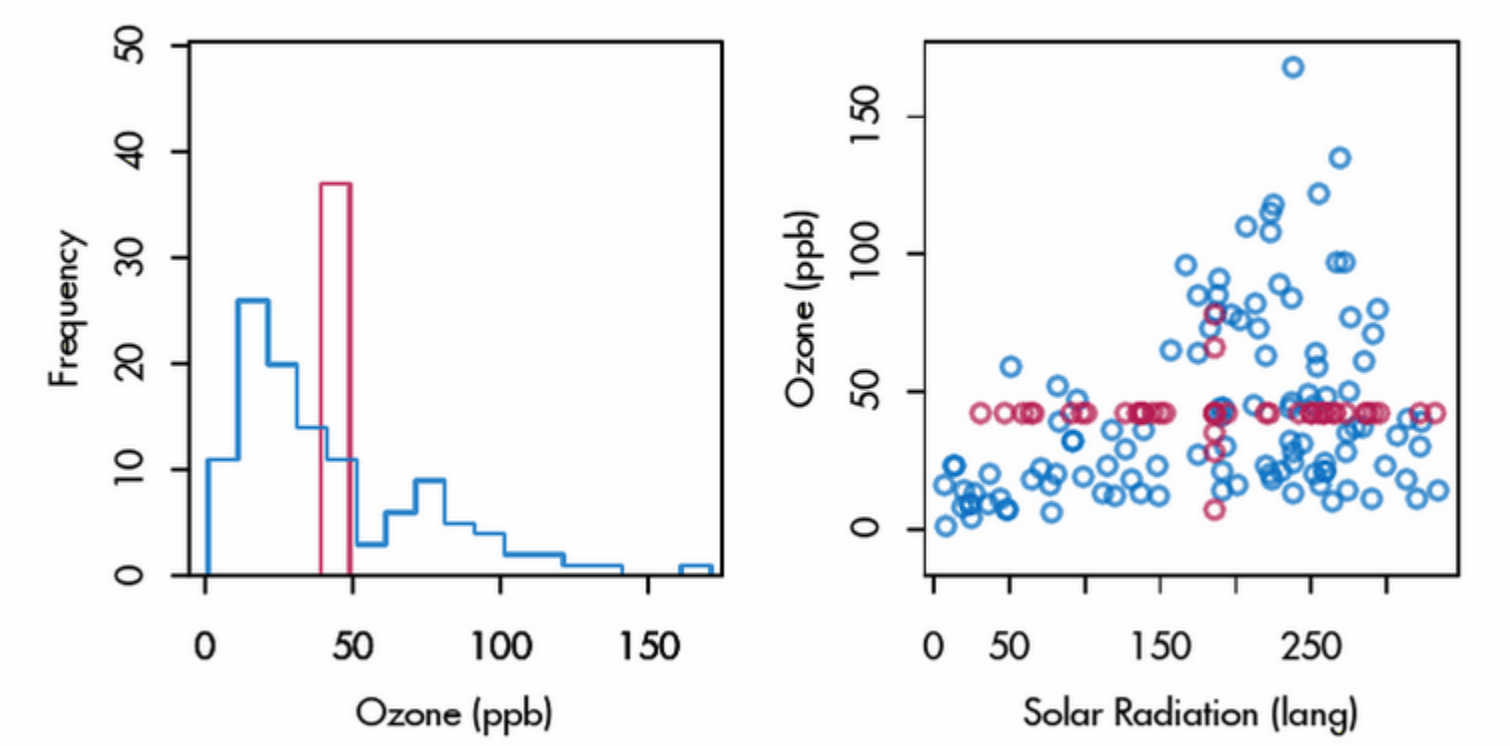
- One very simple solution is to replace each missing value for a particular variable by the mean of the non-missing values.

- Assumes MCAR for unbiassed univariate analysis

- Can severely distort the distribution of the variable;
  - Variability in data and estimates ``too low'';
  - Weakens relationships between variables
  - Biasses estimates other than the mean

- **Not ever recommended**

Mean Imputation



From Stef van Buuren. Flexible Imputation of Missing Data. CRC Press, Boca Raton FL, 2012.

Simple Random Imputation

- Fill in missing value for a particular variable with a simple random sample from of the non-missing values.

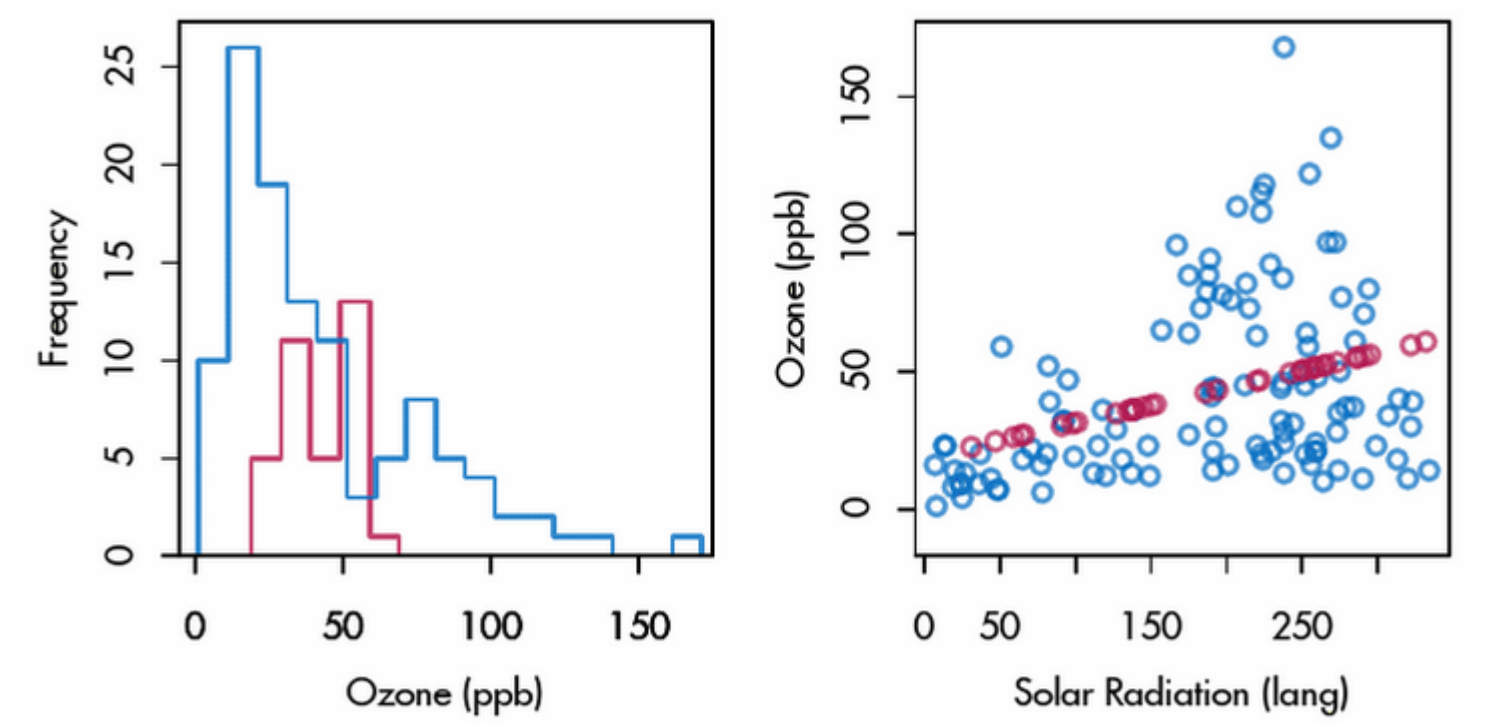
- Ignores information contained in other variables
  - Weakens relationships between variables
  - Variability in estimates ``too low''

- **Not recommended; use only as a first step**

(Deterministic) Regression Imputation

- Fit a regression model to non-missing values and replace missing value by its most likely value (prediction).
- Unbiassed estimates of regression coefficients under MAR
- Distorts distribution of data
  - Strengthens relationships between predictor and response
  - Variability in data and estimates ``too low''
- **Not recommended; use only as a first step**

(Deterministic) Regression Imputation

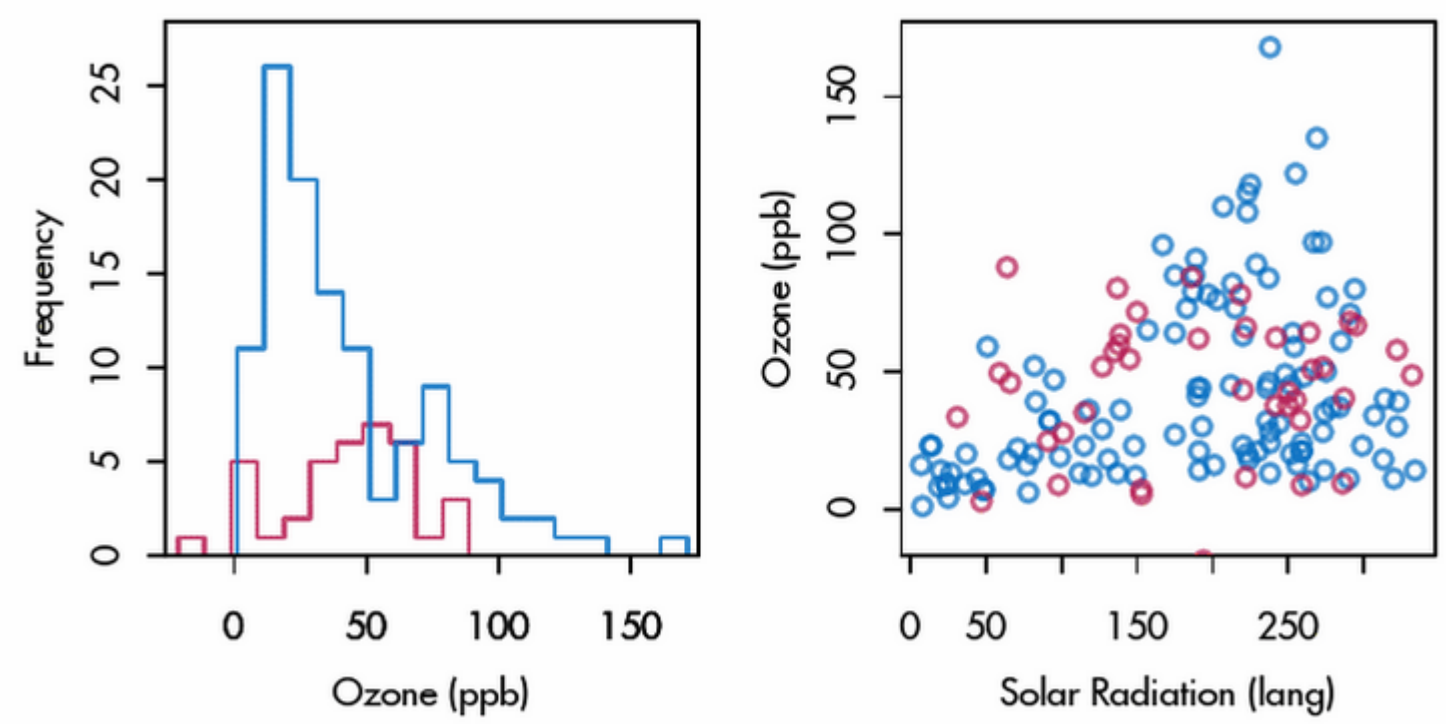


From Stef van Buuren. Flexible Imputation of Missing Data. CRC Press, Boca Raton FL, 2012.

(Stochastic) Regression Imputation

- Fit a regression model to non-missing values and replace missing value by its predicted value `plus noise'.
  - `Noise' distribution estimated from non-missing residuals
- Unbiassed estimates of regression coefficients and correlation under MAR
- Variability in estimates ``too low''
- **A step in the right direction**

(Stochastic) Regression Imputation

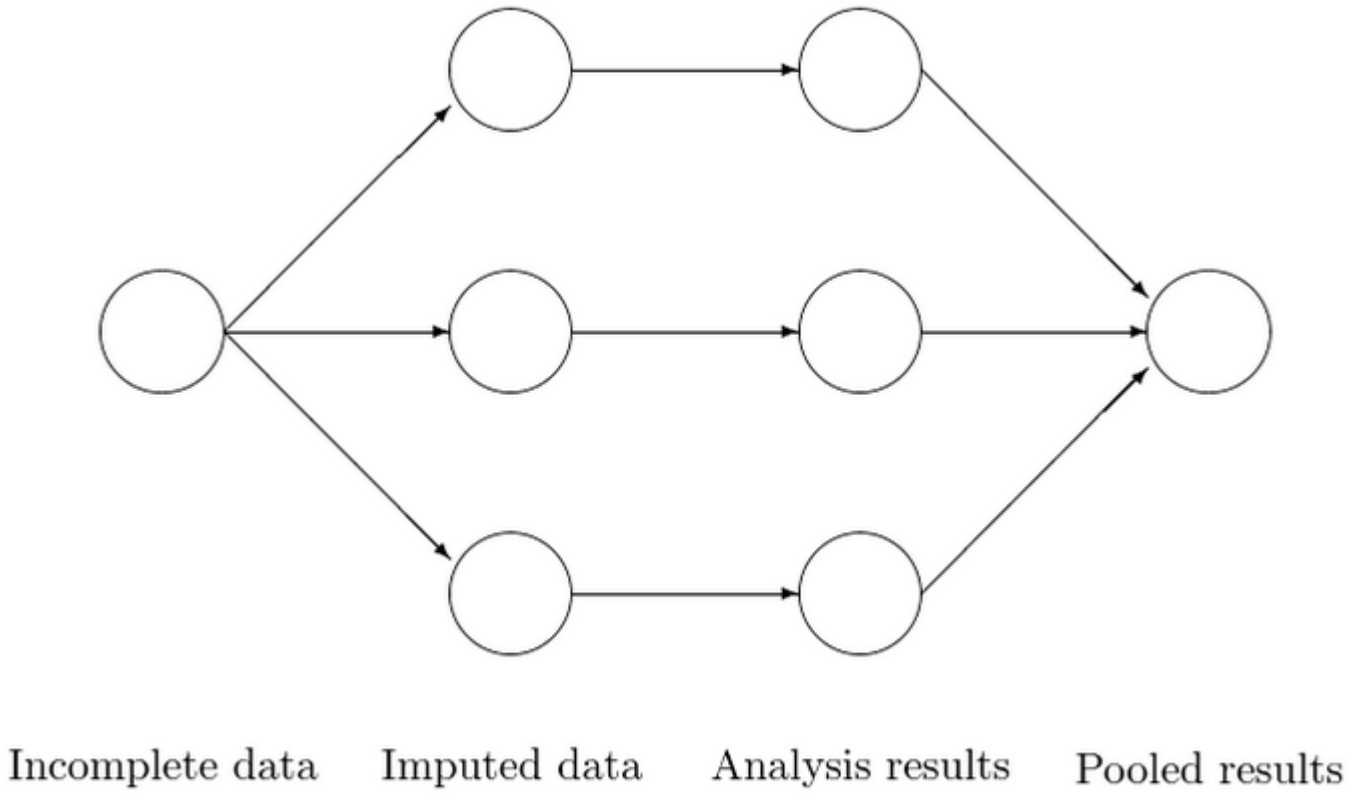


From Stef van Buuren. Flexible Imputation of Missing Data. CRC Press, Boca Raton FL, 2012.

Multiple Imputation

- Analysis framework to address the problem of variability in estimates ``too low''.
- Creates *m* complete datasets (missing data drawn from distribution).
- Each dataset analysed using standard techniques.
- The *m* results are pooled into a final point estimate, together with an estimate of its variance.

Multiple Imputation



From Stef van Buuren. Flexible Imputation of Missing Data. CRC Press, Boca Raton FL, 2012.

Other Methods

- Re-weighting complete cases
  - Cf. Surveys

- Likelihood approaches
  - Model for the observed data and missing data mechanism
  - Iterative maximum likelihood techniques such as the expectation maximisation (EM) algorithm

POLL QUESTIONS - IMPUTATION