DATA7001
INTRODUCTION TO DATA SCIENCE

Module 1 Problem Solving with Data

Shazia Sadiq

Welcome to Master of Data Science

- First in Australia to offer an advanced level of knowledge applied in industry, government, social and scientific contexts.
- Emphasis on high level of graduate attributes through cross-disciplinary curriculum and innovative and disruptive thinking applied to complex problems.
- 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

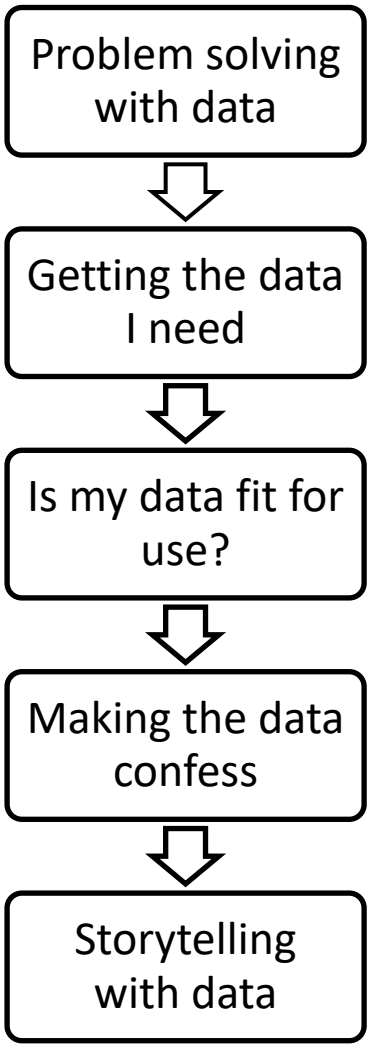
Piazza is a great tool for communication between students and between students and teaching team. Note that all resources and announcements will always be made via this site (blackboard), and piazza is used for discussions and Q&A.

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 ++

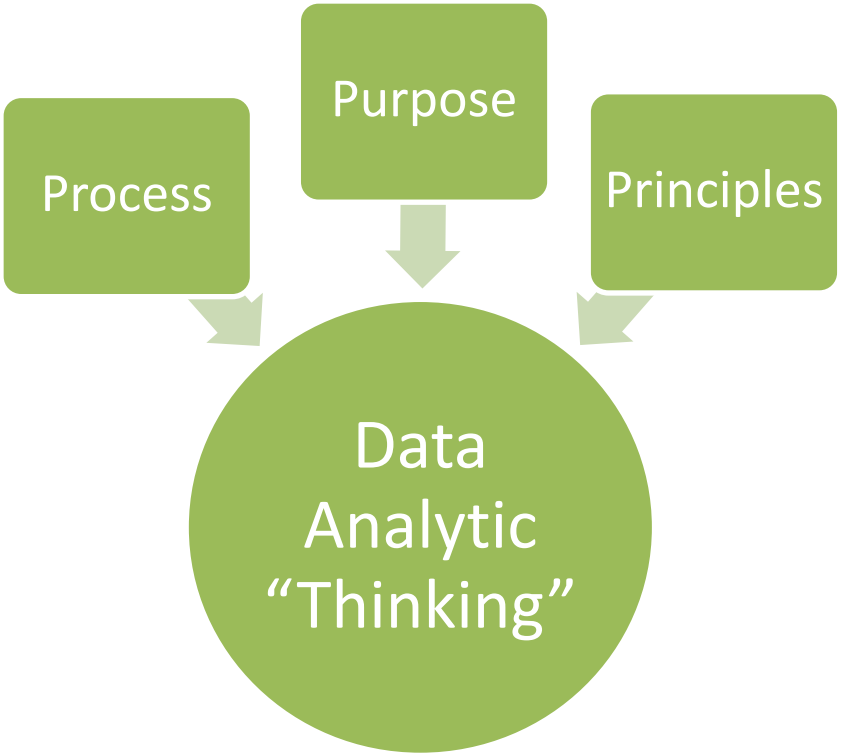
→ DATA7901 Capstone Propose
→ DATA7902/3 Capstone Build

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

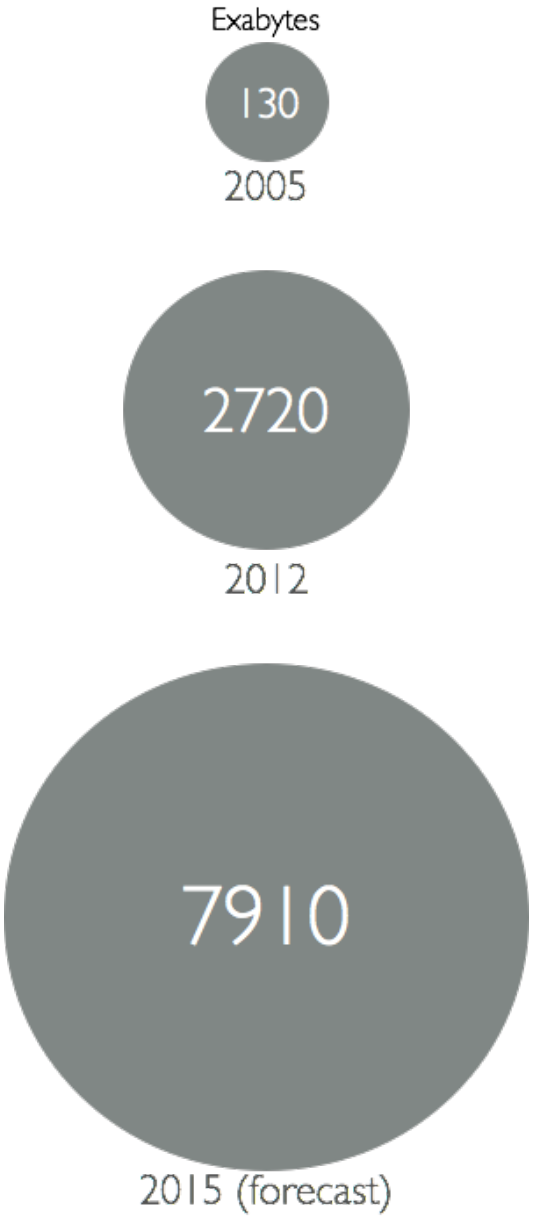
- Lecturers
  - Shazia Sadiq (course coordinator)
  - Thomas Taimre
- Tutors
  - Ajay Hemnath
  - Hrishikesh Patel
  - Reia Natu
  - Mubashir Imran
- Guest presenters

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}



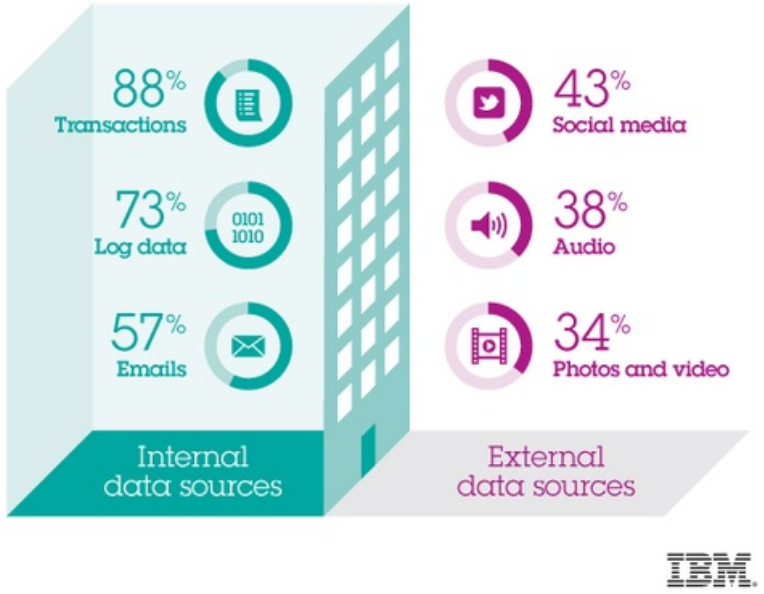
How big is Big Data?

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!

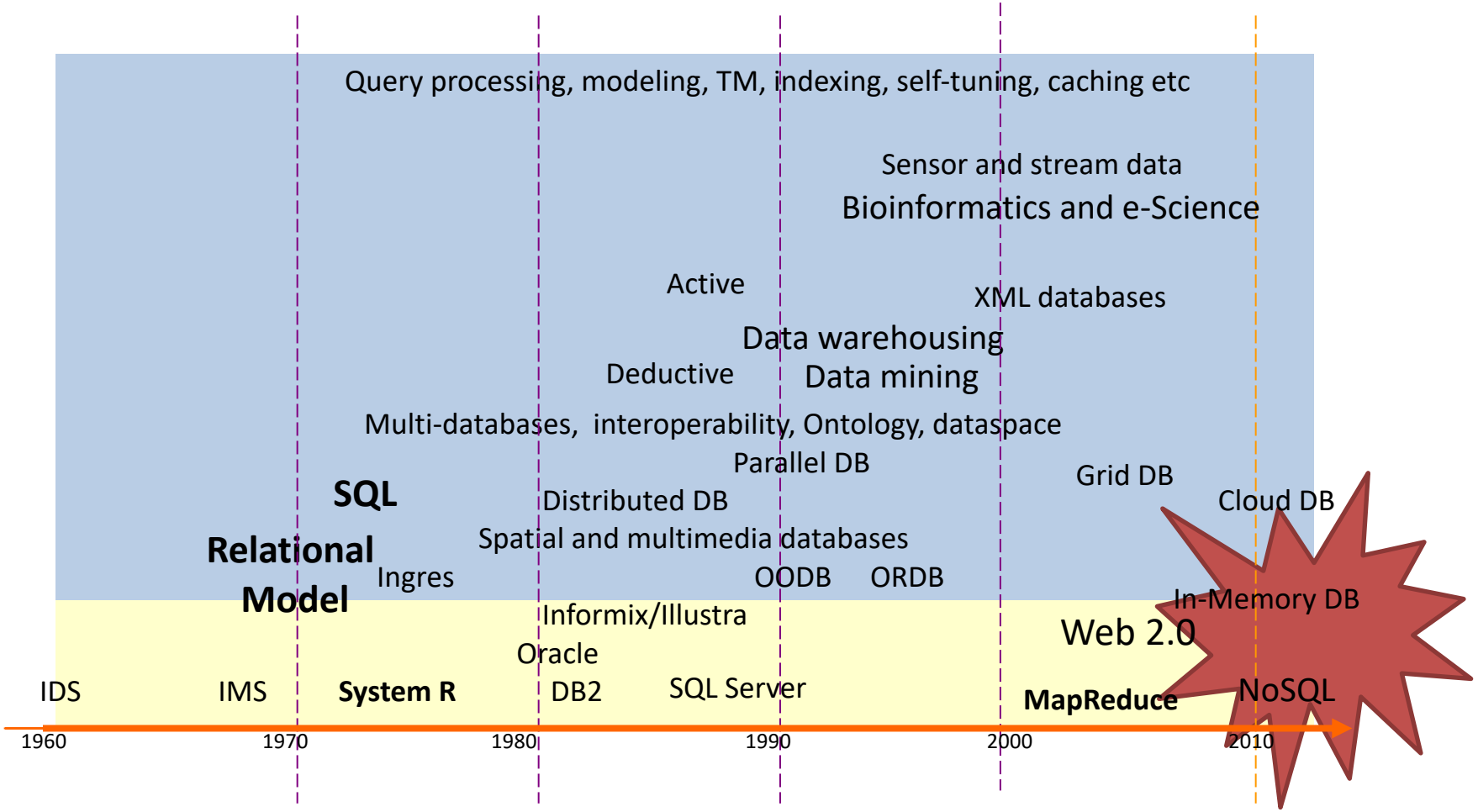
(Once) Advantages of DBMS

- 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

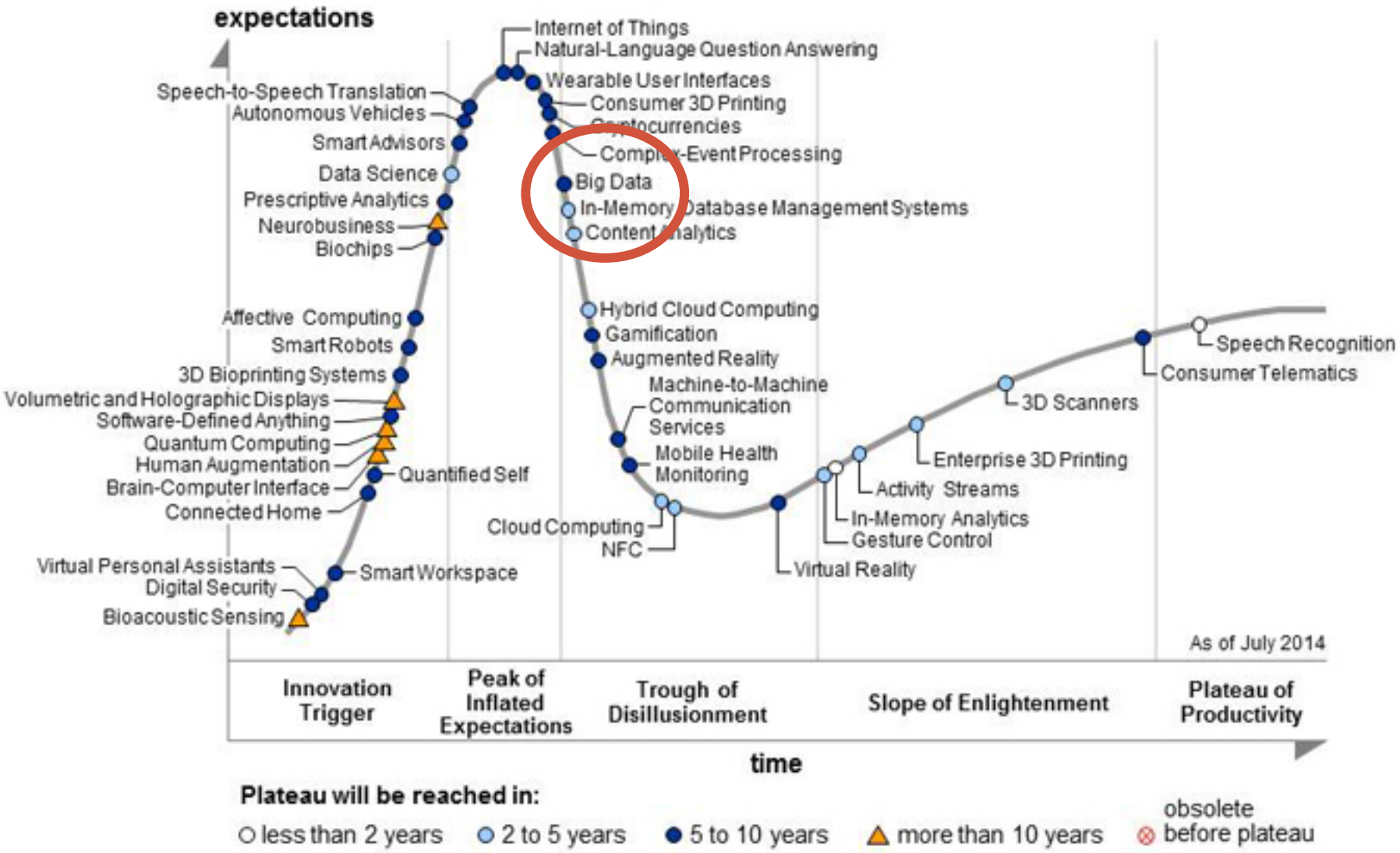
Sensory

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

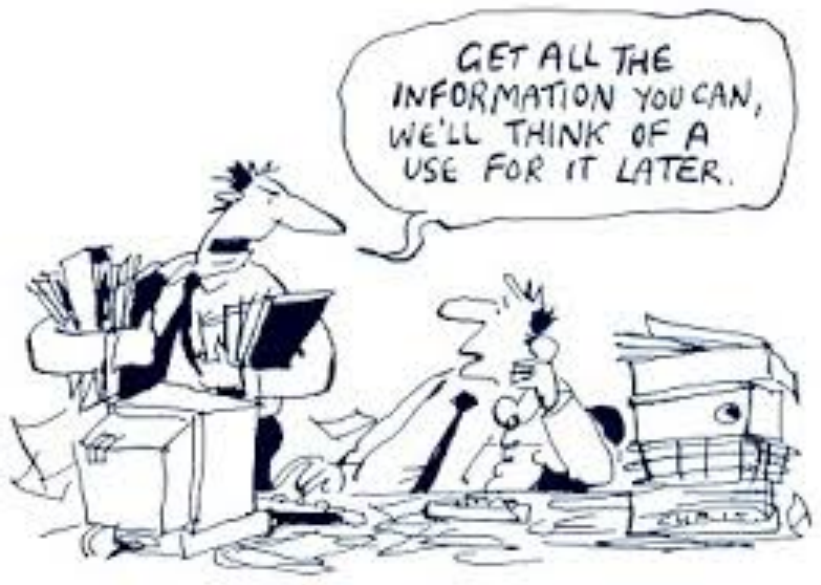
- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Not validating models
- Data pipeline integrity
- Using statistical tests correctly
- Making transition from prototype to production
- Data generation, process and use complexity (who do you ask?)
- …

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

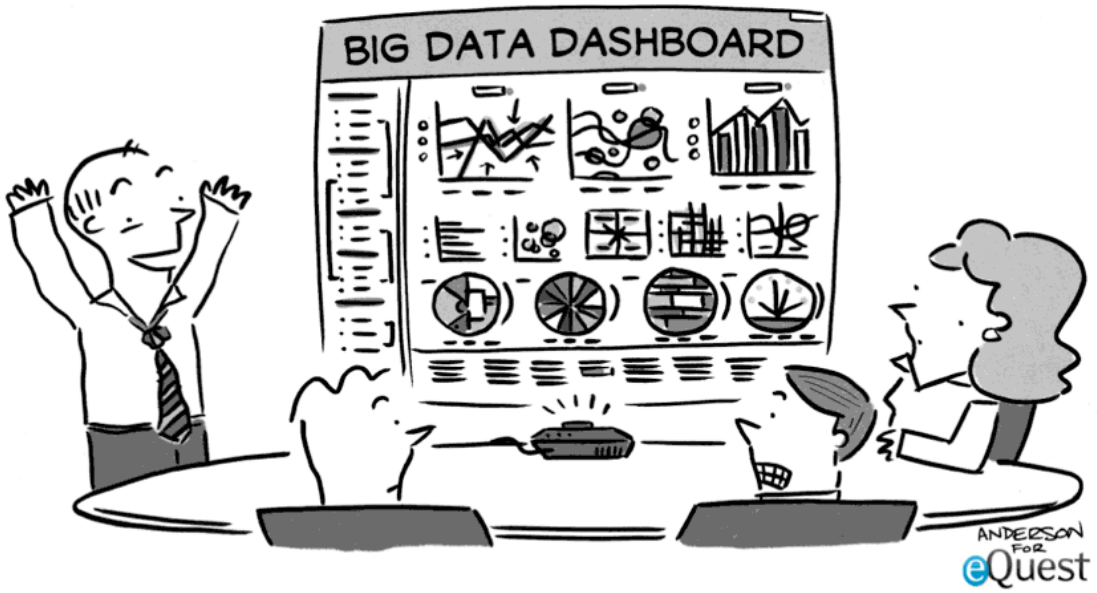


… if you don't know the right question to ask you discover nothing

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

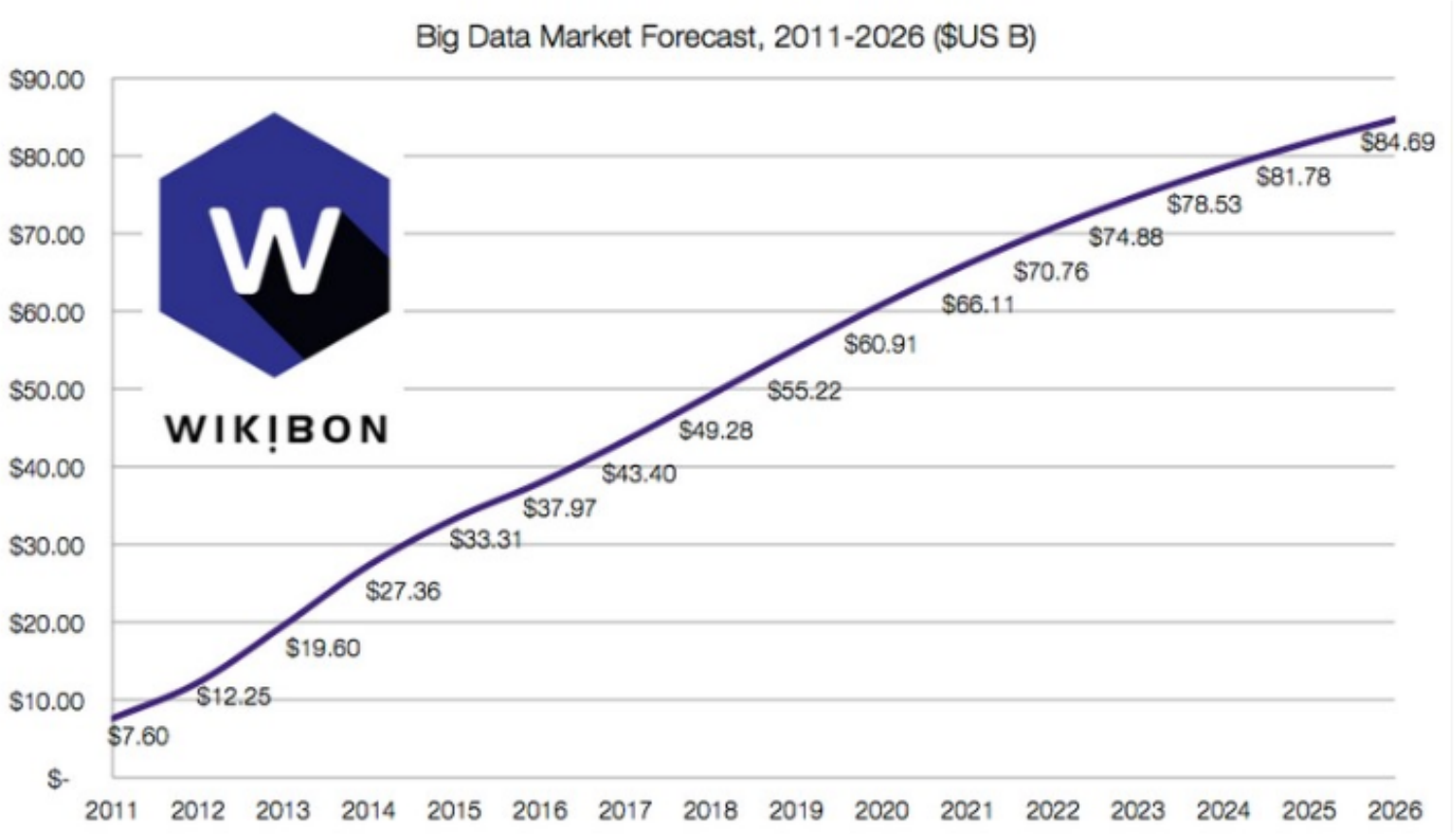
"Among Voters in New Jersey, GOP Sees Dead People," The New York Times, September 16, 2005 by David W. Chen." Comparing information from county voter registration lists, Social Security death records and other pubic information, Republican officials announced on Thursday (9/15/05) that 4,755 people who 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 …

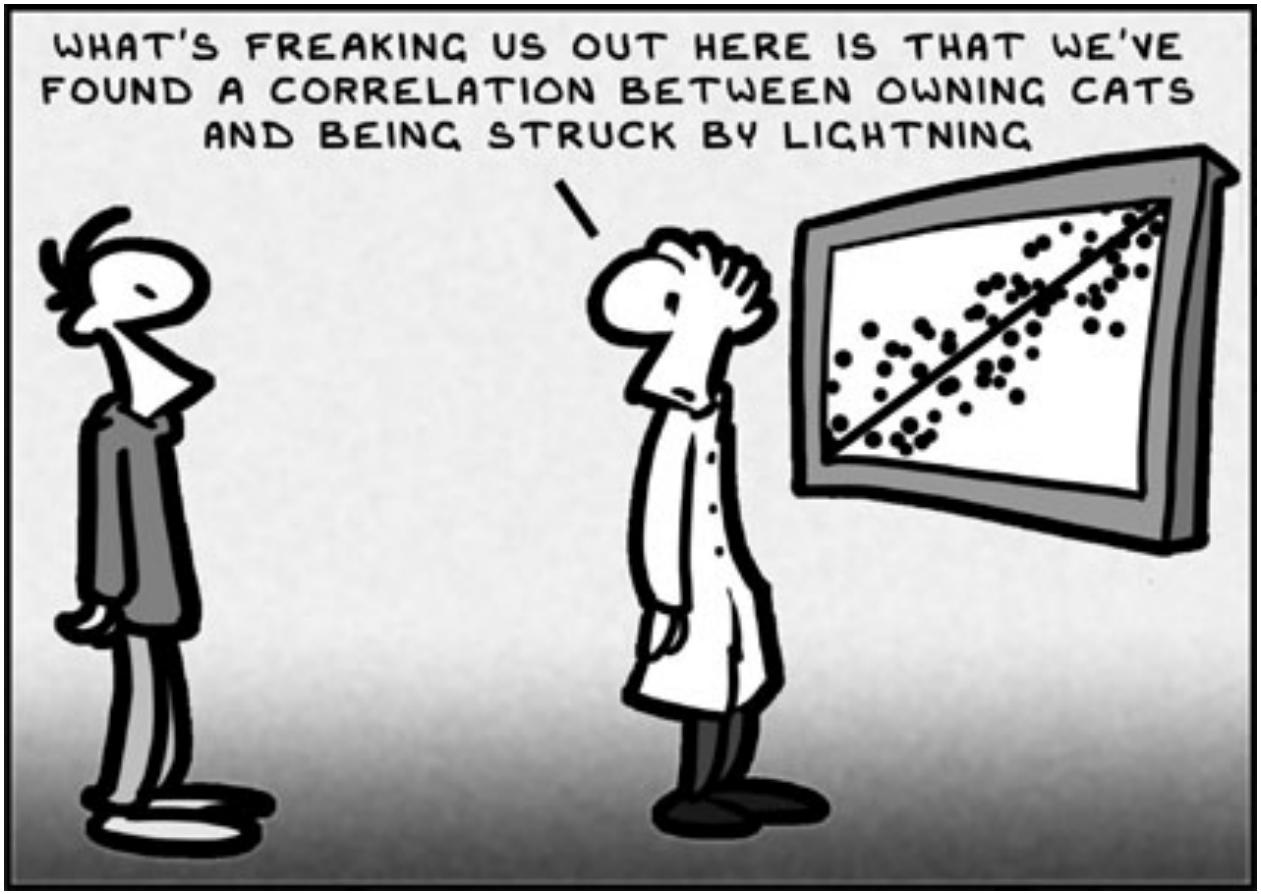
- Sloan Digital Sky Survey
- Next Generation Sequencing
- Large Hadron Collider
- Security and Surveillance
- Financial Planning and Risk Management
- Environmental Modelling
- Energy Saving
- Preventative Healthcare
- Intelligent Transportation
- Logistics
- Predictive Maintenance
- Brand/Product Protection
- 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