ANALYZING BIG DATA – I

Week 1 – Introduction

Agenda

- Motivation
  - What is Big Data?
  - Is it important?
  - Is it something new?

- Topics covered
  - Syllabus

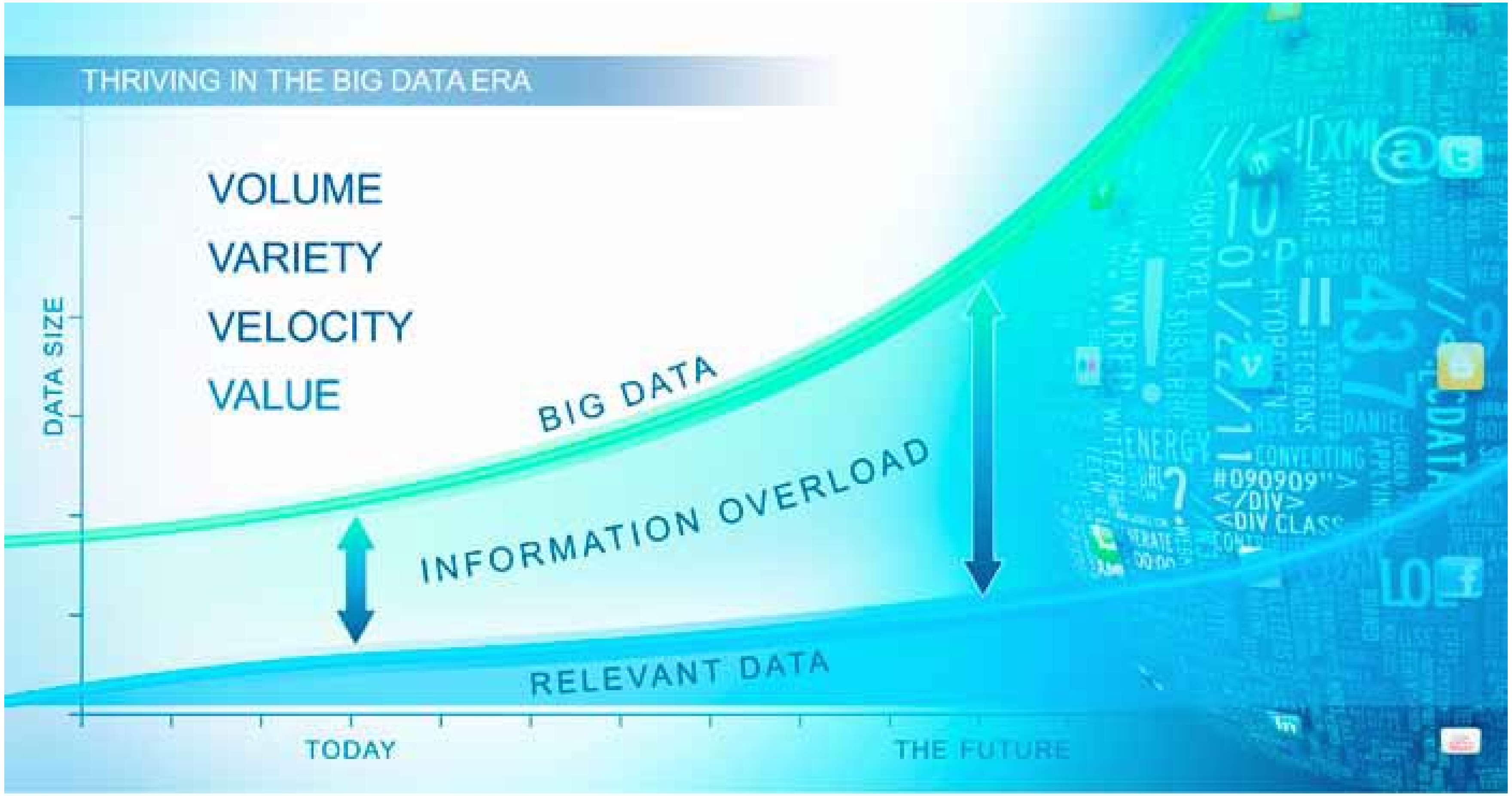
- Administrative Staff
  - Course Instructor
  - Course TA'S
  - Course Goals
  - Course Evaluation method

- Introducing Homework 1: Intro to SQL for Data Science

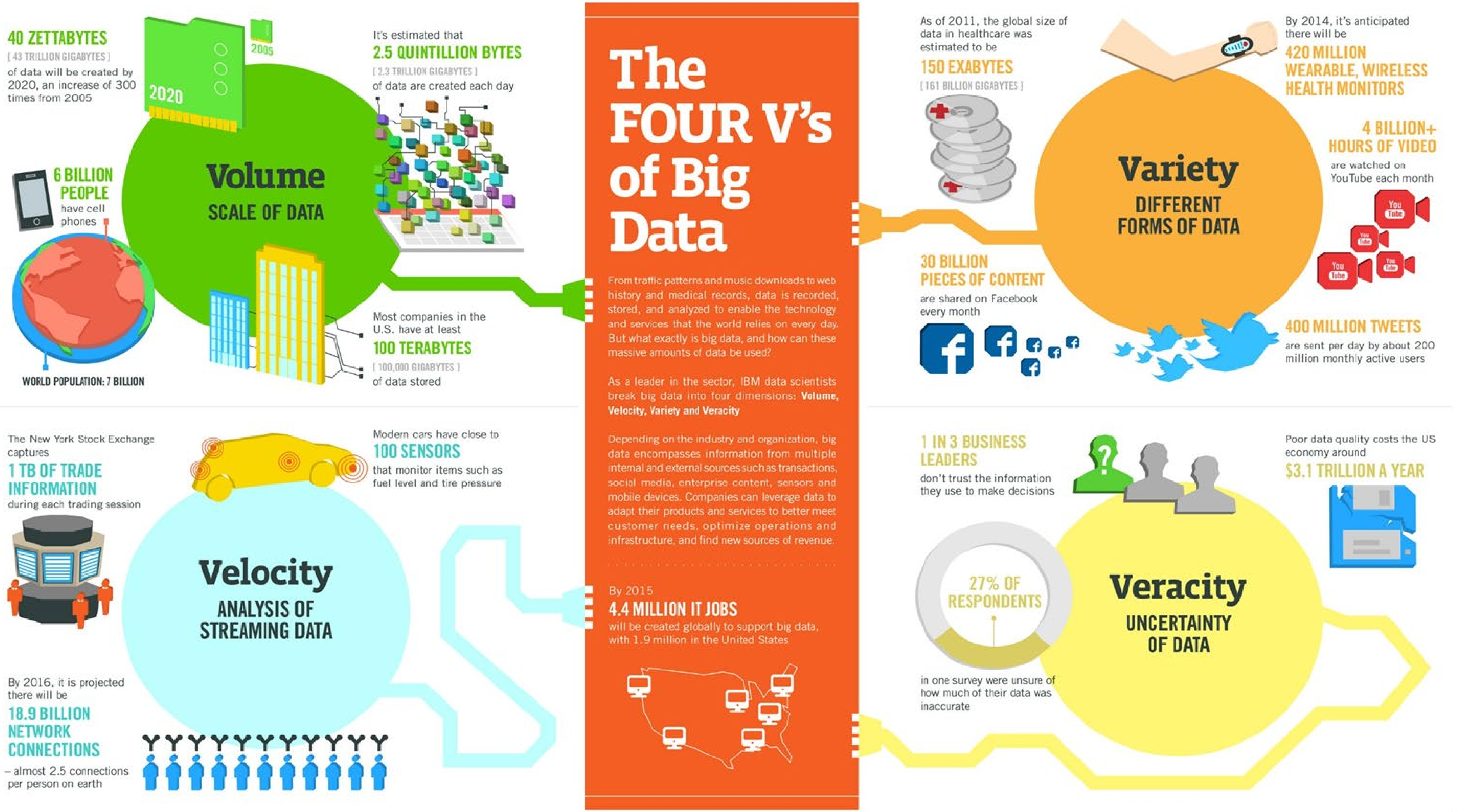
- First visit to MySQL workbench (if time permits)

MOTIVATION

What is Big Data?



Source: SAS Institute White Paper, Big Data Meets Big Data Analytics



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

– almost 2.5 connections per person on earth

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions

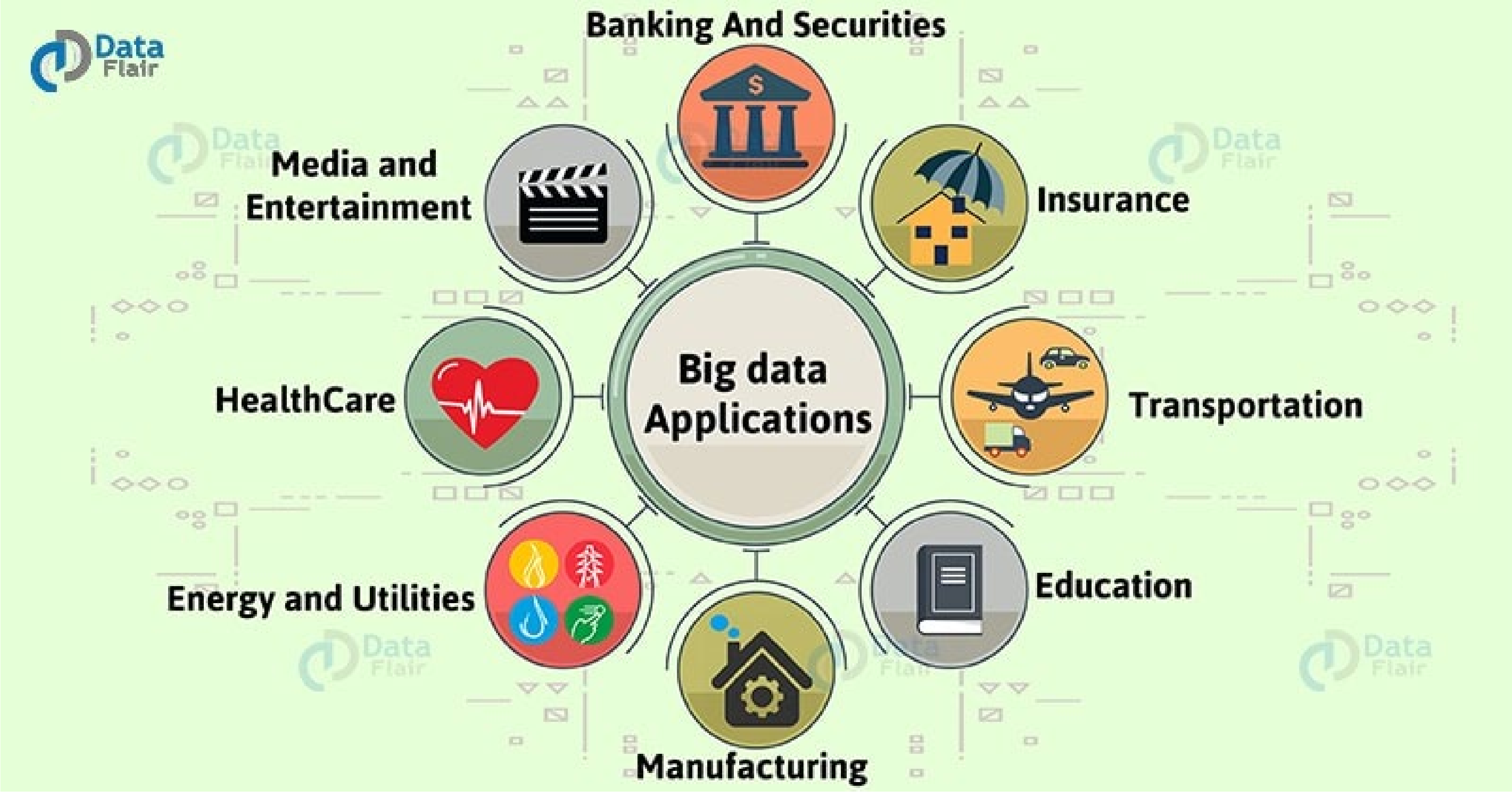
Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

27% OF RESPONDENTS

in one survey were unsure of how much of their data was inaccurate

Big Data Applications



Source: data-flair.com Blog - Top Real Time Big Data Applications in Various Domains

Data Science reshaping HealthCare

- $3.5 billion was invested in 188 digital health companies in 2017 Q1-Q2

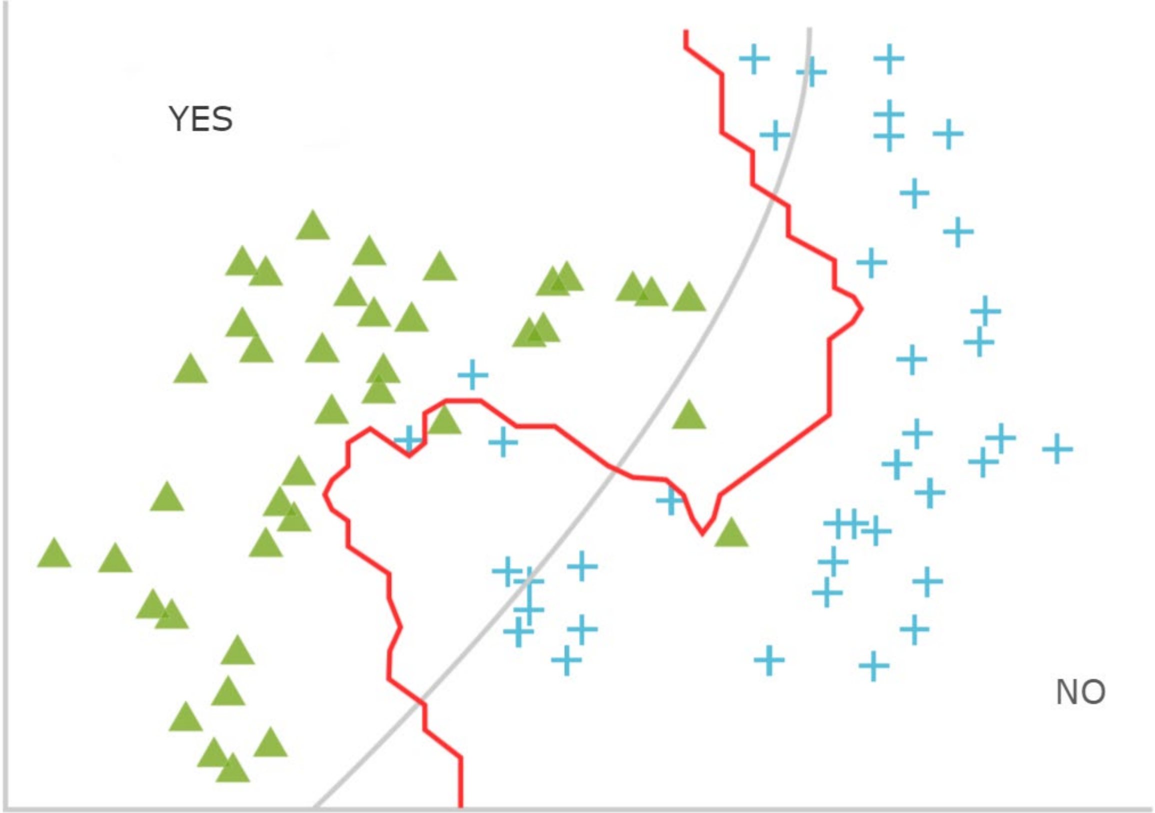
- Technology companies in the health space
  - IBM Health
  - Apple Research Kit

- Examples of data science applications in Health
  - Gathering health data
  - Optimizing clinic performance
  - Prescription errors, optimizing insurance payouts, hospital readmissions
  - Improving diagnostic accuracy:
    - Misdiagnosis
    - Genome sequencing
    - Pharmaceutical research

Data Science reshaping HealthCare

Is there a probability that a patient will experience heart failure?

Machine learning can answer this question!

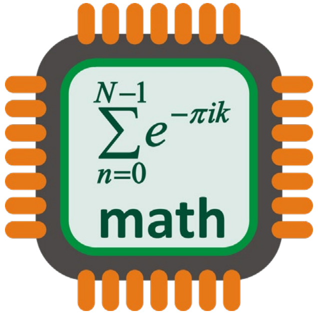


Ethical concerns

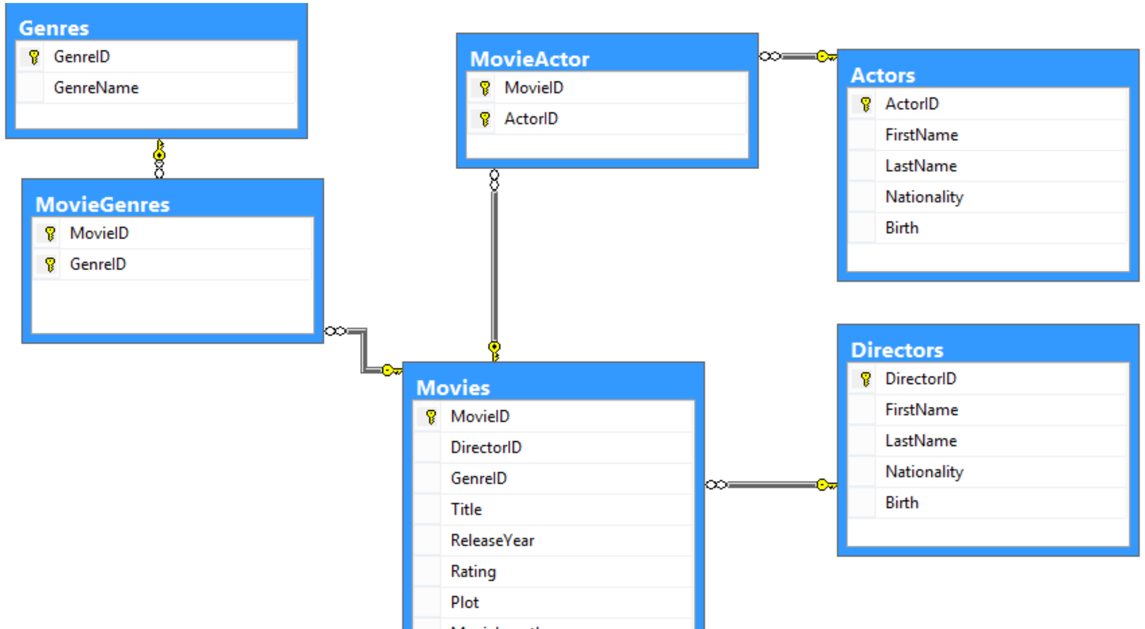
- Growing issue, maybe we are learning too much:
  - Privacy issues: Racial and Sexual Discrimination
  - Amazon Predicting
  - Debate about Telsa self-driving car
    - Maybe -> Difficult to understand

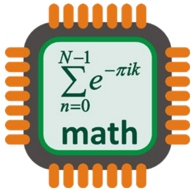
From Big Data Analytics with Data Bases

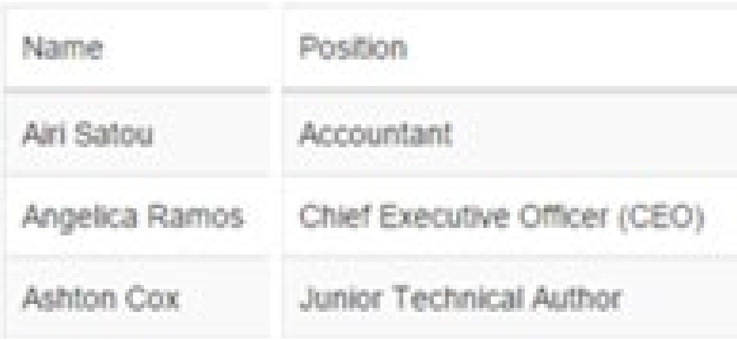
- Why not start talking about Programing languages for doing analytics?
  - Answer has to do the Computer Architecture
  - How do we sketch a Computer?



From Big Data Analytics with Data Bases

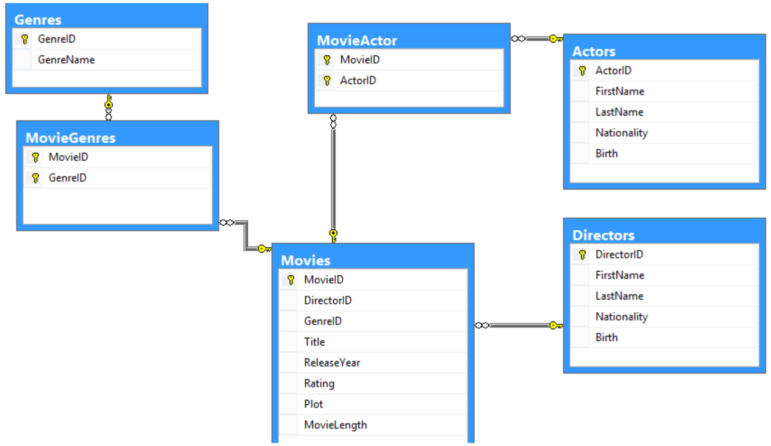


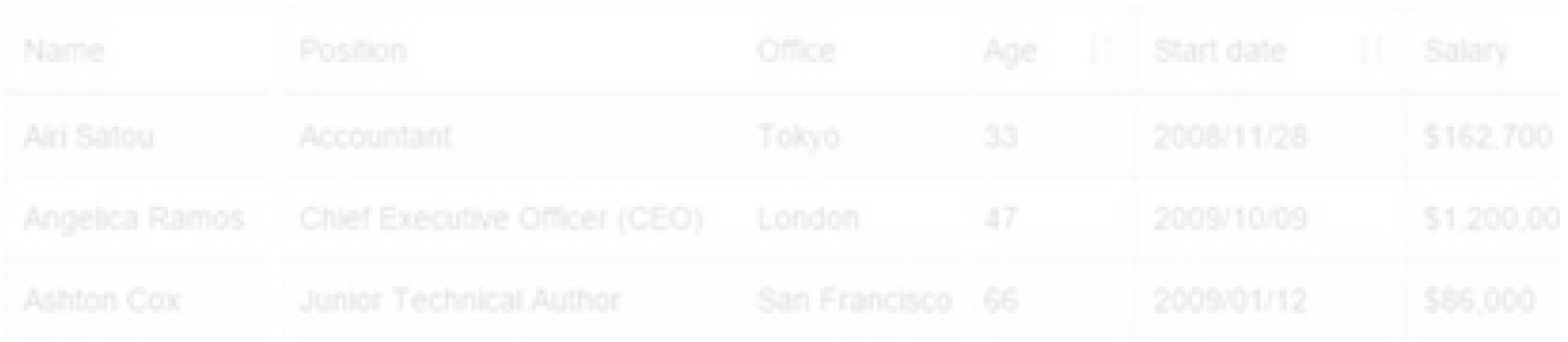




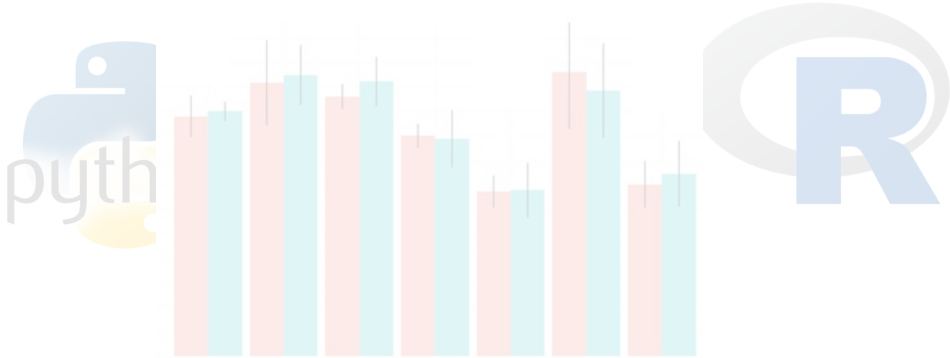
Name	Position
Airi Satou	Accountant
Angelica Ramos	Chief Executive Officer (CEO)
Ashton Cox	Junior Technical Author

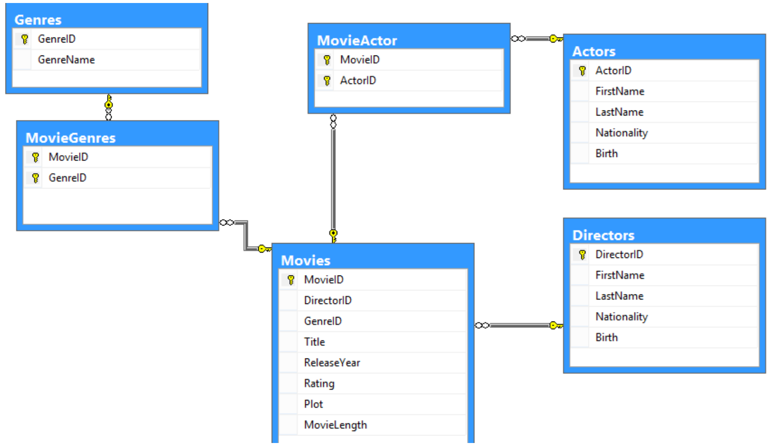
STRUCTURE OF THE COURSE

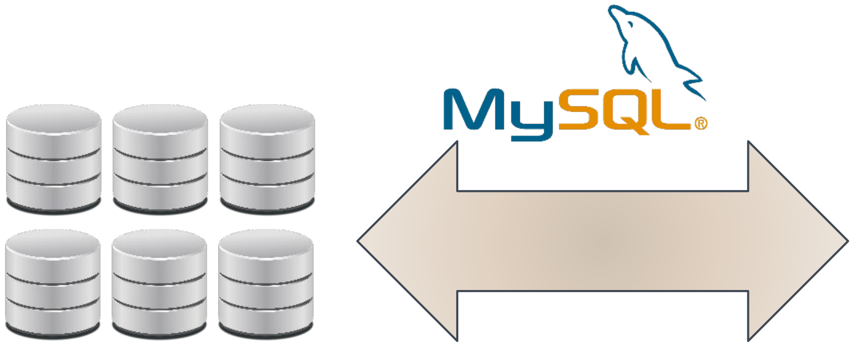


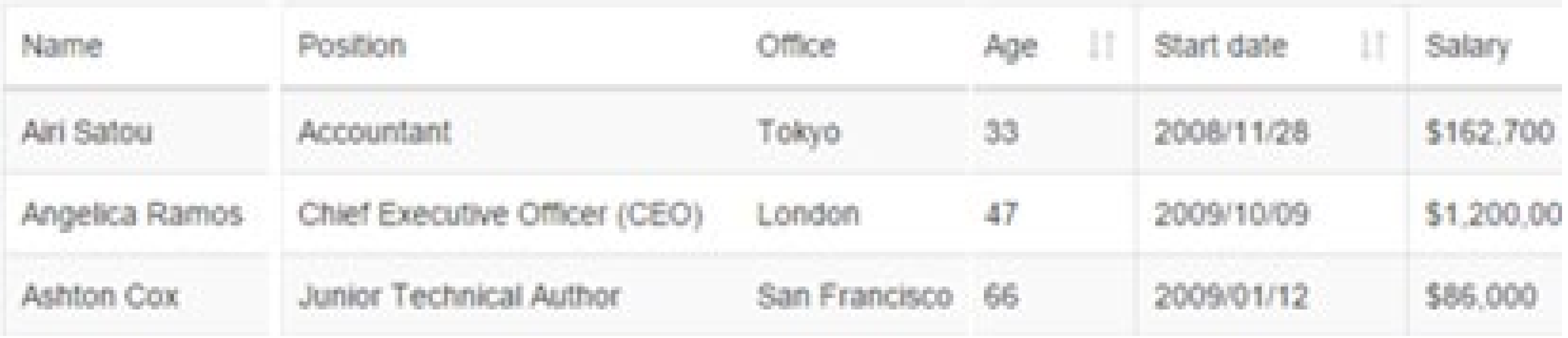


Name	Position	Office	Age	Start date	Salary
Airi Satou	Accountant	Tokyo	33	2008/11/28	$162,700
Angelica Ramos	Chief Executive Officer (CEO)	London	47	2009/10/09	$1,200,00
Ashton Cox	Junior Technical Author	San Francisco	66	2009/01/12	$86,000

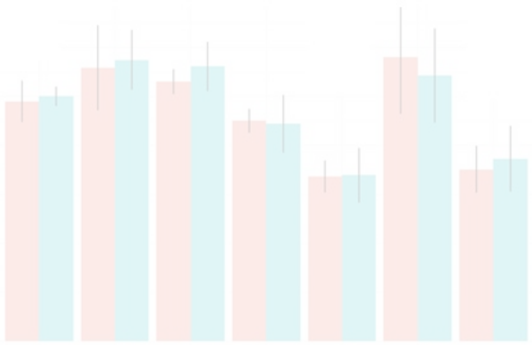


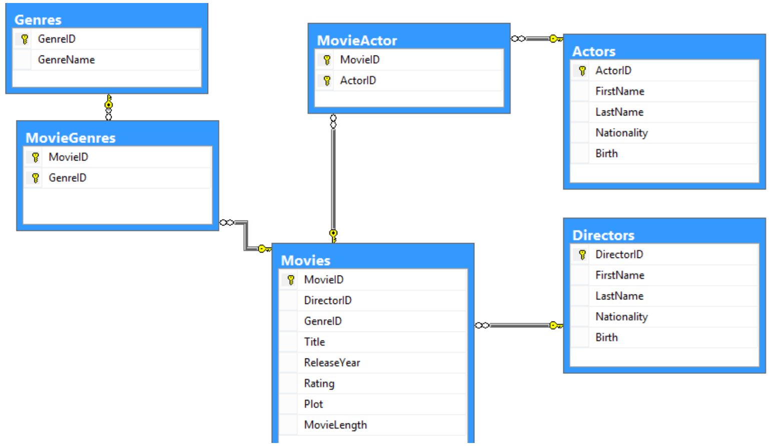






Name	Position	Office	Age	Start date	Salary
Airi Satou	Accountant	Tokyo	33	2008/11/28	$162,700
Angelica Ramos	Chief Executive Officer (CEO)	London	47	2009/10/09	$1,200,00
Ashton Cox	Junior Technical Author	San Francisco	66	2009/01/12	$86,000





Genres
- GenreID
- GenreName

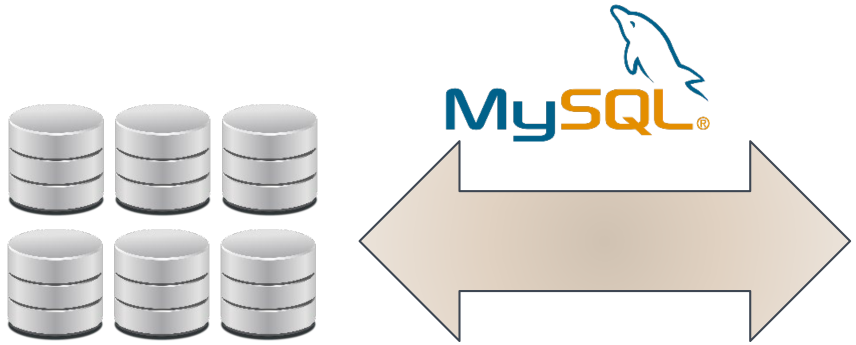
MovieActor
- MovieID
- ActorID

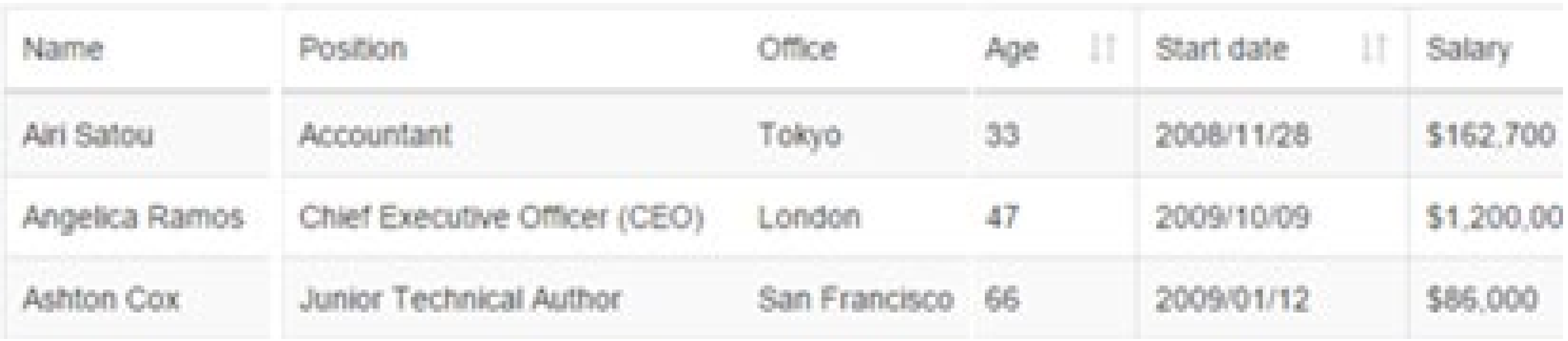
Actors
- ActorID
- FirstName
- LastName
- Nationality
- Birth

MovieGenres
- MovieID
- GenreID

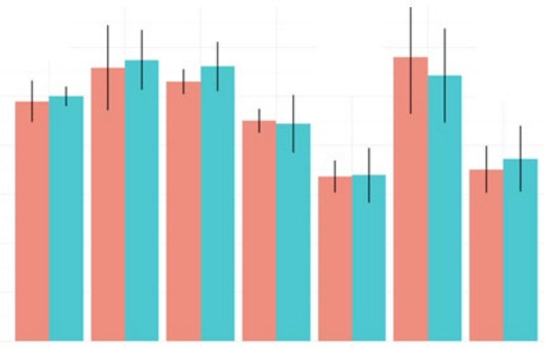
Movies
- MovieID
- DirectorID
- GenreID
- Title
- ReleaseYear
- Rating
- Plot
- MovieLength

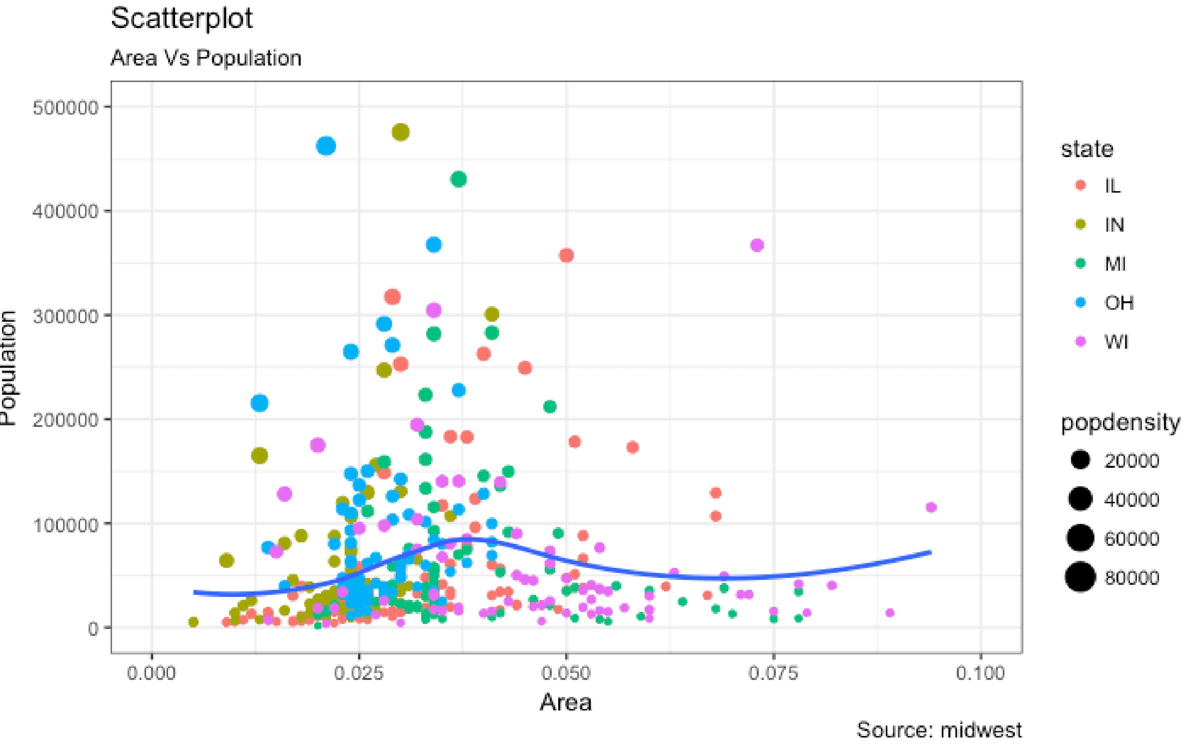
Directors
- DirectorID
- FirstName
- LastName
- Nationality
- Birth

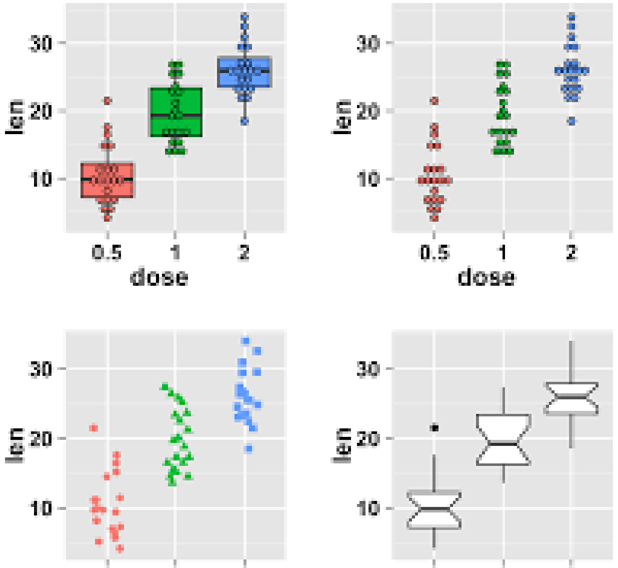


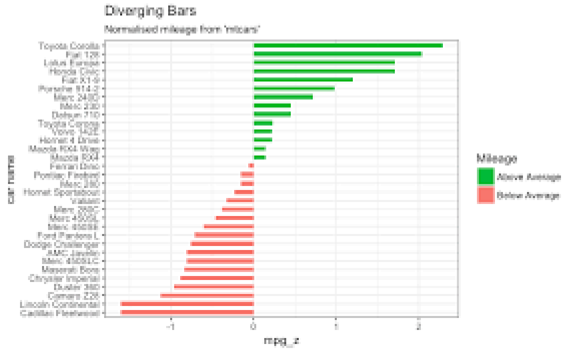


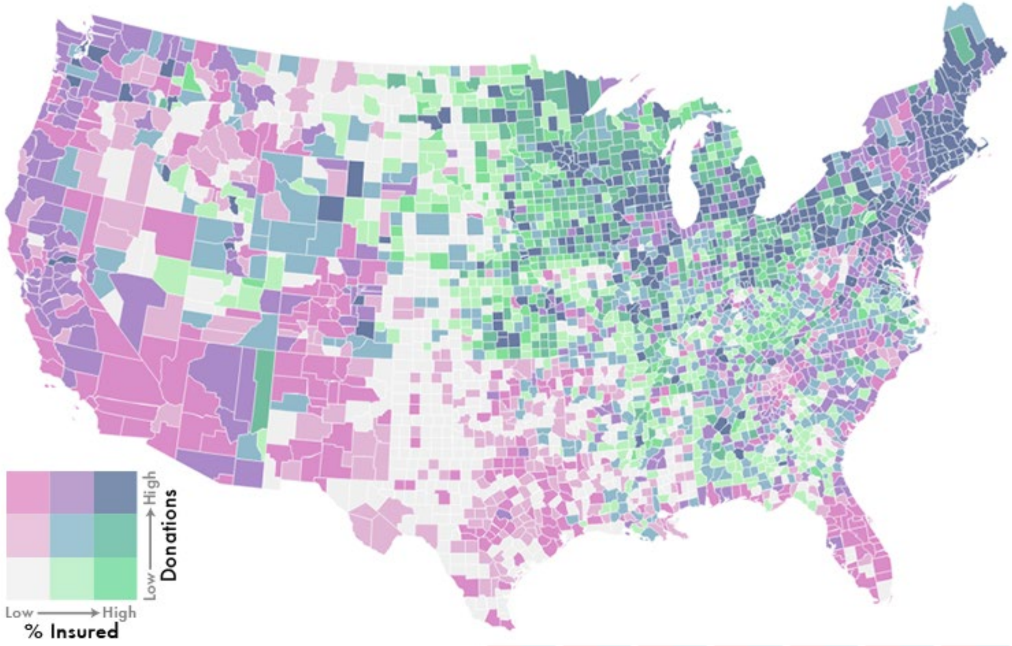
Name	Position	Office	Age	Start date	Salary
Airi Satou	Accountant	Tokyo	33	2008/11/28	$162,700
Angelica Ramos	Chief Executive Officer (CEO)	London	47	2009/10/09	$1,200,00
Ashton Cox	Junior Technical Author	San Francisco	66	2009/01/12	$86,000




Scatterplot
Area Vs Population
Population
500000
400000
300000
200000
100000
0
0.000
0.025
0.050
0.075
0.100
Area
state
IL
IN
MI
OH
WI
popdensity
20000
40000
60000
80000
Source: midwest


len
30
20
10
0.5
1
2
dose
len
30
20
10
0.5
1
2
dose
len
30
20
10
len
30
20
10


Diverging Bars
Normalised mileage from 'mtcars'
car name
Toyota Corolla
Fiat 128
Lotus Europa
Honda Civic
Fiat X1-9
Porsche 914-2
Merc 240D
Merc 230
Datsun 710
Toyota Corona
Volvo 142E
Hornet 4 Drive
Mazda RX4 Wag
Mazda RX4
Ferrari Dino
Pontiac Firebird
Merc 280
Hornet Sportabout
Valiant
Merc 280C
Merc 450SL
Merc 450SE
Ford Pantera L
Dodge Challenger
AMC Javelin
Merc 450SLC
Maserati Bora
Chrysler Imperial
Duster 360
Camaro Z28
Lincoln Continental
Cadillac Fleetwood
-1
0
1
2
mpg_z
Mileage
Above Average
Below Average


High
Donations
Low
Low
High
% Insured

TOPICS COVERED

Visiting the Syllabus

- Syllabus is continuously updated

ADMINISTRATIVE STAFF

Course: Instructor

- Assistant Professor of Marketing
- I’m an applied statistician. I work with large data sets to answer questions in business and policy.
- I draw on models from economics, and psychology, to inform the data.
- Specific research and consulting interests: brand valuation, consumer attention, pricing of consumer packaged goods
- Office: Sachar 214 - A

Course: Students

- Past experience with
  - Statistics
  - Programming
- Where you are from
- Major
- Previous University
- Anything else you want me to know

Course: TAs

- Office hours with Xavi: Friday 9:00 am – 10:00 am
  - In pairs

- Office Hours:
  - (S1) Zhiqi Chen
  - (S2) Ran Dou

- Office Hours and Location: TBD

Course: Goals

- Course is meant to be a “gateway” course for other courses in the Analytics concentration
  - Analyzing Big Data II
  - Marketing Analytics
  - Data Visualization
  - Analytics Field Projects

- We will spend ~ 4 weeks on SQL , and 3 weeks on learning R methods

Course: Recordings and support

- To help you learn class concepts and practice programming on your own, I will be posting the full, commented code to LATTE

- I am trying to make Classes Recorded
  - Videos links will be on LATTE

Course: Etiquette

- Come to class on time
- Cell phones should have their ringers off and should be out of sight
- Laptops:
  - Allowed for note taking and class activities
  - Should never disrupt class
- Ask questions and ask me to slow down if I am going too fast or the material is not clear.
- Help out the class by initiating and participating

Course: Honor Code

- Graded cases, assignments & team project:
  - Don’t consult solutions of other teams/individuals
  - Put your name on cases and assignments only if you contributed materially to solution
  - After cases presented, don’t share solutions/notes with others outside of class
  - Mind pairs/team instructions
  - Breaking these rules leads to an F!

Course: References

- Links for reference materials are available on LATTE
  - Russell, G. *Database eLearning*. online at https://db.grussell.org/index.html
  - Grolemund, G. and Wickham, H. *R for Data Science* (2017). online at http://r4ds.had.co.nz/ (Link on LATTE)
  - Chang, W. *Cookbook for R.* Online at http://www.cookbook-r.com/ (Link on LATTE)
  - Zhang, Y. *R and Data Mining.* Online pdf through LATTE

Course: Graded Deliverables

- Attendance & Class participation
- 1 Midterm
- Weekly Homeworks
- Final project

Course: Participation

- Attendance
  - Attendance is compulsory!
  - Lack of attendance will bring down participation grade

Course: Midterm

- Date is on the syllabus

Weekly Homeworks

- There will be homework every week

- DataCamp is a distance-learning website specifically for data science and programming techniques

- Class email invites from datacamp.com – Sign up!
  - If you DIDN'T receive an email from DataCamp to sign up, OR
  - If you are getting errors while trying to sign up ----- EMAIL Me ASAP!

- Homework 1 - Introduction to SQL - Due Next Thursday

Final Project

- Will involve analyzing a large dataset

- Focus is on whether you can create a SQL database and then use it to analyze a marketing problem

- Group formation – I will assign you to groups of 3-4 students, based on your Data Science Profiles

INTRODUCING HW 1

DataCamp Assignment

Due next Thursday: : Intro to SQL for Data Science

Thanks !!

NEXT WEEK…

- Presenting a datacamp course