Data Science

Introduction

Google Says…

2.5 quintillion bytes of **data** are produced by humans every **day**. If you've wondered how much **data** the average person uses **per** month, you can start by looking at how much **data** is **created** every **day** in **2020** by the average person. This currently stands at 2.5 quintillion bytes **per** person, **per day**. Sep 10, 2020

techjury.net › blog › how-much-data-is-created-every-day

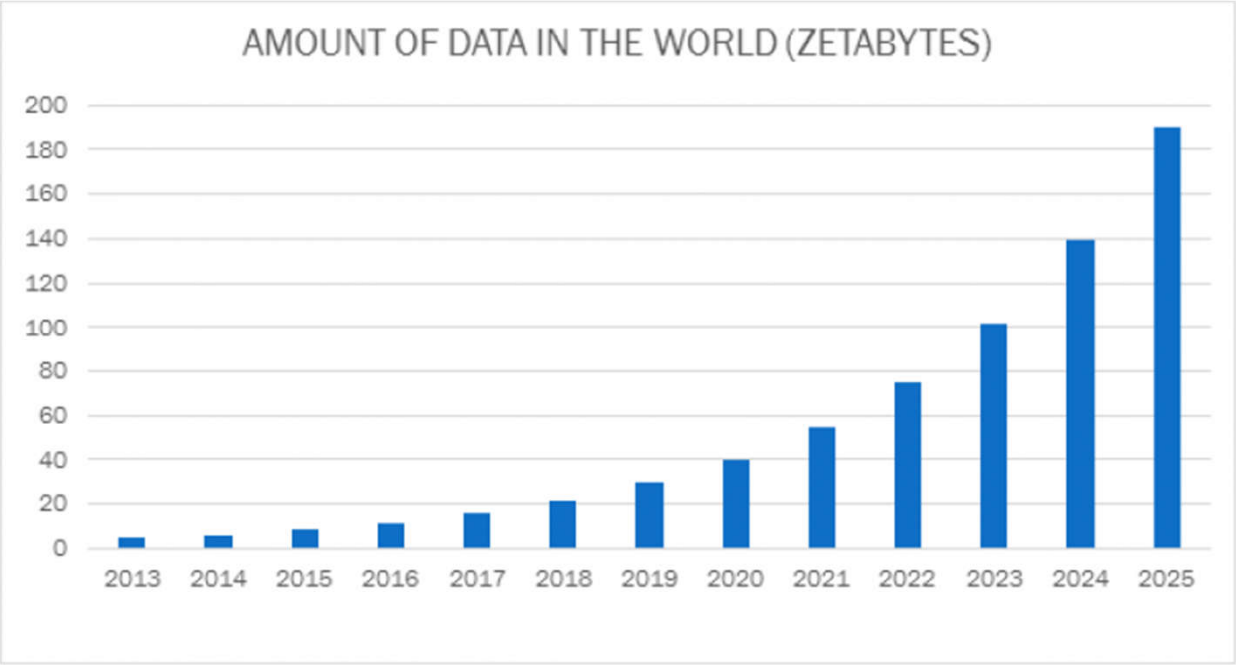
How Much Data Is Created Every Day in 2020? [You'll be ...

What does that mean?

Prefix		Base 10	Decimal	English
Name	Symbol			Short scale
yotta	Y	10^{24}	1 000 000 000 000 000 000 000 000	septillion
zetta	Z	10^{21}	1 000 000 000 000 000 000 000	sextillion
exa	E	10^{18}	1 000 000 000 000 000 000	quintillion
peta	P	10^{15}	1 000 000 000 000 000	quadrillion
tera	T	10^{12}	1 000 000 000 000	trillion
giga	G	10^{9}	1 000 000 000	billion

If data had mass, earth could be a blackhole!

Exponential Growth of Data



What is Data Science

- “Collecting, manipulating, and analyzing data in order to extracting value from it.”

What is Data Science

- “Collecting, manipulating, and analyzing data in order to extracting value from it.”
- Wikipedia
  - “Data Science is the extraction of knowledge from data, which is a continuation of the field of data mining and predictive analytics.”

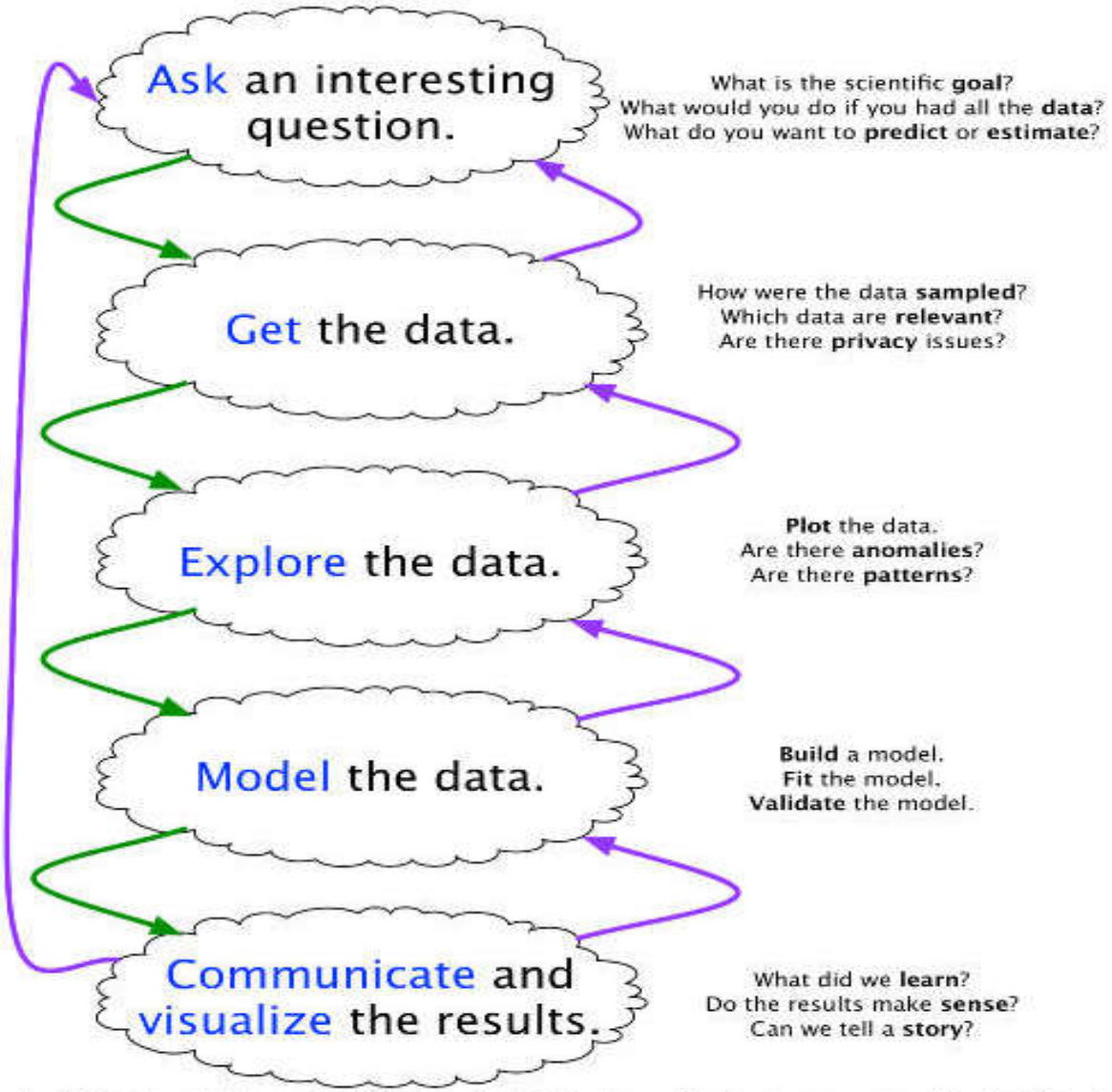
What is Data Science

- “Collecting, manipulating, and analyzing data in order to extracting value from it.”
- Wikipedia
  - “Data Science is the extraction of knowledge from data, which is a continuation of the field of data mining and predictive analytics.”
- NIST Big Data Working Group
  - “Data Science is the empirical synthesis of actionable knowledge from raw data through the complete data lifecycle process.”

Why Data Science Matters

- It empowers policy makers with quantifiable data driven evidences
- It helps organizations
  - to make better decisions and test those decisions
  - to and make plans for improvement
  - to track trends, analyze user behaviors and define goals
  - to identify and refine target audiences
  - to identify opportunities
  - to design effective business processes and adopt best practices

The Data Science Process



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

The Data Science Process

- The Data Science Process is similar to the scientific process - one of observation, model building, analysis and conclusion:
  - Ask questions
  - Data Collection
  - Data Exploration
  - Data Modeling
  - Data Analysis
  - Visualization and Presentation of Results
- Note: This process is by no means linear!

It’s all about story!
Can we tell a story?

Let’s look at an example

- Hubway is metro-Boston’s public bike share program, with more than 1600 bikes at 160+ stations across the Greater Boston area. Hubway is owned by four municipalities in the area.

Let’s look at an example

- Hubway is metro-Boston’s public bike share program, with more than 1600 bikes at 160+ stations across the Greater Boston area. Hubway is owned by four municipalities in the area.
- By 2016, Hubway operated 185 stations and 1750 bicycles, with 5 million ride since launching in 2011.
- In April 2017, Hubway held a Data Visualization Challenge at the Microsoft NERD Center in Cambridge, releasing 5 years of trip data.

Let's look at an example

- Hubway is metro-Boston's public bike share program, with more than 1600 bikes at 160+ stations across the Greater Boston area. Hubway is owned by four municipalities in the area.
- By 2016, Hubway operated 185 stations and 1750 bicycles, with 5 million ride since launching in 2011.
- In April 2017, Hubway held a Data Visualization Challenge at the Microsoft NERD Center in Cambridge, releasing 5 years of trip data.
- **The Question**
  - What does the data tell us about the ride share program?

The Data Exploration/Question Refinement Cycle

- Our original question
  - “What does the data tell us about the ride share program?”

The Data Exploration/Question Refinement Cycle

- Our original question
  - “What does the data tell us about the ride share program?”
- Perhaps a good slogan for a Hackathon, but not good for guiding scientific investigation!

The Data Exploration/Question Refinement Cycle

- Our original question
  - “What does the data tell us about the ride share program?”
- Perhaps a good slogan for a Hackathon, but not good for guiding scientific investigation!
- Before we can refine the question, we have to look at the data!

The Data Exploration/Question Refinement Cycle

- Our original question
  - "What does the data tell us about the ride share program?"
- Perhaps a good slogan for a Hackathon, but not good for guiding scientific investigation!
- Before we can refine the question, we have to look at the data!

	seq_id	hubway_id	status	duration	start_date	strt_statn	end_date	end_statn	bike_nr	subsc_type	zip_code	birth_date	gender
0	1	8	Closed	9	7/28/2011 10:12:00	23.0	7/28/2011 10:12:00	23.0	B00468	Registered	'97217	1976.0	Male
1	2	9	Closed	220	7/28/2011 10:21:00	23.0	7/28/2011 10:25:00	23.0	B00554	Registered	'02215	1966.0	Male
2	3	10	Closed	56	7/28/2011 10:33:00	23.0	7/28/2011 10:34:00	23.0	B00456	Registered	'02108	1943.0	Male
3	4	11	Closed	64	7/28/2011 10:35:00	23.0	7/28/2011 10:36:00	23.0	B00554	Registered	'02116	1981.0	Female
4	5	12	Closed	12	7/28/2011 10:37:00	23.0	7/28/2011 10:37:00	23.0	B00554	Registered	'97214	1983.0	Female

The Data Exploration/Question Refinement Cycle

- Our original question
  - "What does the data tell us about the ride share program?"
- Good slogan for a Hackathon, but not good for guiding scientific investigation!
- Before we can refine the question, we have to look at the data!

	seq_id	hubway_id	status	duration	start_date	strt_statn	end_date	end_statn	bike_nr	subsc_type	zip_code	birth_date	gender
0	1	8	Closed	9	7/28/2011 10:12:00	23.0	7/28/2011 10:12:00	23.0	B00468	Registered	'97217	1976.0	Male
1	2	9	Closed	220	7/28/2011 10:21:00	23.0	7/28/2011 10:25:00	23.0	B00554	Registered	'02215	1966.0	Male
2	3	10	Closed	56	7/28/2011 10:33:00	23.0	7/28/2011 10:34:00	23.0	B00456	Registered	'02108	1943.0	Male
3	4	11	Closed	64	7/28/2011 10:35:00	23.0	7/28/2011 10:36:00	23.0	B00554	Registered	'02116	1981.0	Female
4	5	12	Closed	12	7/28/2011 10:37:00	23.0	7/28/2011 10:37:00	23.0	B00554	Registered	'97214	1983.0	Female

- Based on the data, what kind of questions can we ask?

The Data Exploration/Question Refinement Cycle

- Who?
  - Who's using the bikes?

The Data Exploration/Question Refinement Cycle

- Who?
  - Who's using the bikes?
- Refine into specific hypotheses

The Data Exploration/Question Refinement Cycle

- Who?
  - Who's using the bikes?
- Refine into specific hypotheses
  - More men or more women?

The Data Exploration/Question Refinement Cycle

- Who?
  - Who's using the bikes?
- Refine into specific hypotheses
  - More men or more women?
  - Older or younger people?

The Data Exploration/Question Refinement Cycle

- Who?
  - Who's using the bikes?
- Refine into specific hypotheses
  - More men or more women?
  - Older or younger people?
  - Subscribers or one time users?

The Data Exploration/Question Refinement Cycle

- Where?
  - Where are bikes being checked out?

The Data Exploration/Question Refinement Cycle

- Where?
  - Where are bikes being checked out?
- Refine into specific hypotheses

The Data Exploration/Question Refinement Cycle

- Where?
  - Where are bikes being checked out?
- Refine into specific hypotheses
  - More in Boston than Cambridge?

The Data Exploration/Question Refinement Cycle

- Where?
  - Where are bikes being checked out?
- Refine into specific hypotheses
  - More in Boston than Cambridge?
  - More in commercial or residential?

The Data Exploration/Question Refinement Cycle

- Where?
  - Where are bikes being checked out?
- Refine into specific hypotheses
  - More in Boston than Cambridge?
  - More in commercial or residential?
  - More around tourist attractions?

The Data Exploration/Question Refinement Cycle

- Where?
  - Where are bikes being checked out?
- Refine into specific hypotheses
  - More in Boston than Cambridge?
  - More in commercial or residential?
  - More around tourist attractions?

- Sometimes the data need a lot of pre-processing.

The Data Exploration/Question Refinement Cycle

- When?
  - When are the bikes being checked out?

The Data Exploration/Question Refinement Cycle

- When?
  - When are the bikes being checked out?
- Refine into specific hypotheses

The Data Exploration/Question Refinement Cycle

- When?
  - When are the bikes being checked out?
- Refine into specific hypotheses
  - More during the weekend than on the weekdays?

The Data Exploration/Question Refinement Cycle

- When?
  - When are the bikes being checked out?
- Refine into specific hypotheses
  - More during the weekend than on the weekdays?
  - More during rush hour?

The Data Exploration/Question Refinement Cycle

- When?
  - When are the bikes being checked out?
- Refine into specific hypotheses
  - More during the weekend than on the weekdays?
  - More during rush hour?
  - More during the summer than the fall?

The Data Exploration/Question Refinement Cycle

- When?
  - When are the bikes being checked out?
- Refine into specific hypotheses
  - More during the weekend than on the weekdays?
  - More during rush hour?
  - More during the summer than the fall?

- Sometimes the feature we want to explore doesn't exist in the data, and must be engineered!

The Data Exploration/Question Refinement Cycle

- Why?
  - For what reasons/activities are people checking out bikes?
- Refine into specific hypotheses
  - More bikes are used for recreation than commute?
  - More bikes are used for touristic purposes?
  - Bikes are use to bypass traffic?

The Data Exploration/Question Refinement Cycle

- Why?
  - For what reasons/activities are people checking out bikes?
- Refine into specific hypotheses
  - More bikes are used for recreation than commute?
  - More bikes are used for touristic purposes?
  - Bikes are use to bypass traffic?

- Do we have the data to answer these questions with reasonable certainty?
- What data do we need to collect in order to answer these questions?

The Data Exploration/Question Refinement Cycle

- How?
  - Questions that combine variables.
  - How does user demographics impact the duration the bikes are being used? Or where they are being checked out?
  - How does weather or traffic conditions impact bike usage?
  - How do the characteristics of the station location affect the number of bikes being checked out?

- How questions are about modeling relationships between different variables.

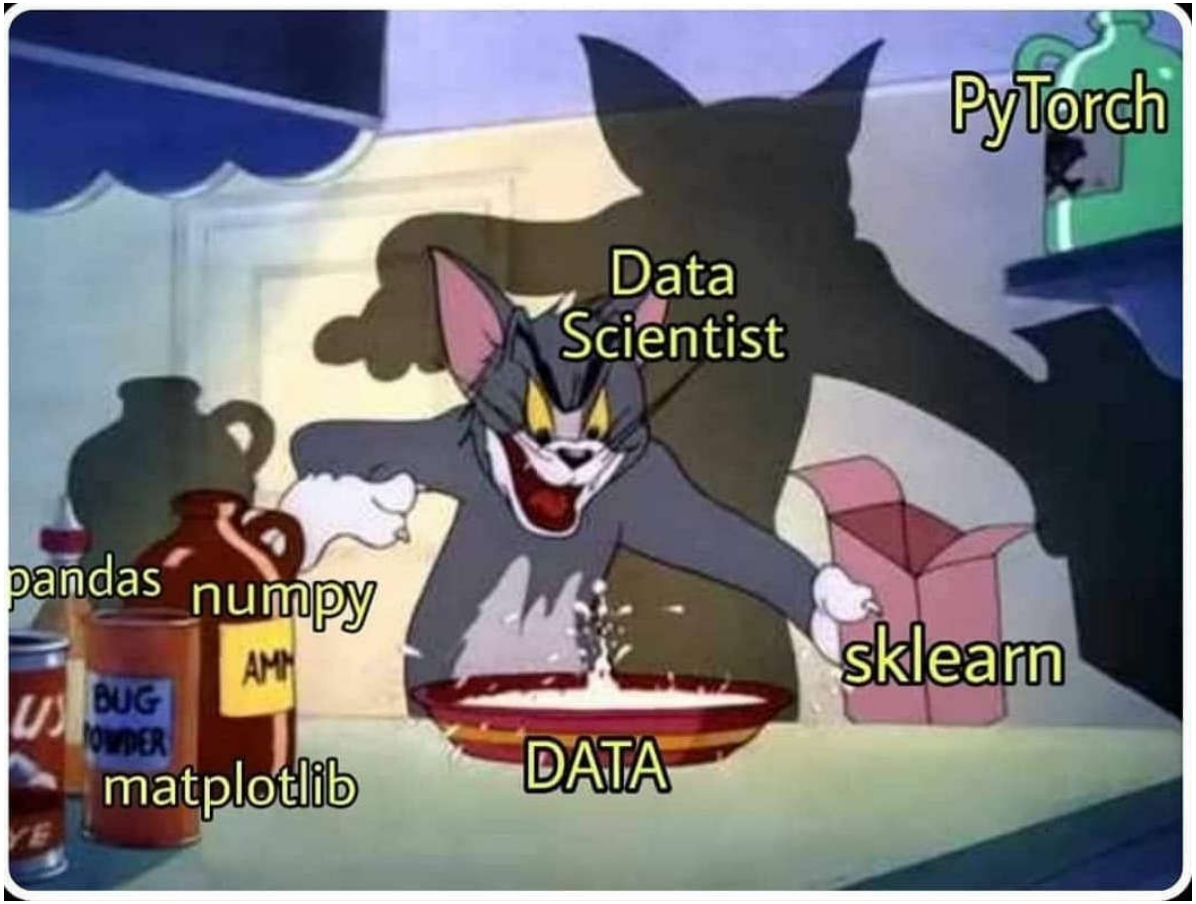
Communicate the results

What is your story?

Data Visualization

Contents of the course

- The scope of Data Science
- Descriptive Statistics
- Exploratory Data Analysis, Principles of Visualizing Data
- Data Scraping, Cleaning and Summarization
- Statistical Significance and P-values
- Building Models and Validating Models
- Linear Algebra Review
- Linear Regression and Logistic Regression
- Crowdsourcing and Ensemble Learning
- Large-scale Clustering
- Mining Massive Datasets
- Python for data analysis, data wrangling and modeling.


PyTorch
Data
Scientist
pandas
numpy
sklearn
matplotlib
DATA