CSE 91, Lecture 1

Credit: Brad Voytek

Udacity (online course, Intro to Data Science.)

Edureka Data Science course video

Lecture plan

- Class Logistics
- Overview of the Data Science field
- Overview of the first project

Course structure

- Lectures on Tuesdays
- Office hours on Thursdays (instead of discussions).
  - Not mandatory
  - Depending on a number of students: either in office, or lab.
- Class Participation is counted towards your pass
  - Bring iclickers
  - First class does not count, can skip 1 more
- All class activities together worth 100 points
  - You need to score 60 or more in order to pass

Class activities (flexible)

- Participation
- Projects
  - About 2-3 projects. (can work in pairs)
- Developing test cases (maybe)
- Presentation of some sort (maybe)

Academic integrity

The same rules from DSC10 apply to CSE91. See the DSC10 Syllabus.

Source: http://www.wenlockjunior.co.uk/values/values-2016-17/integrity

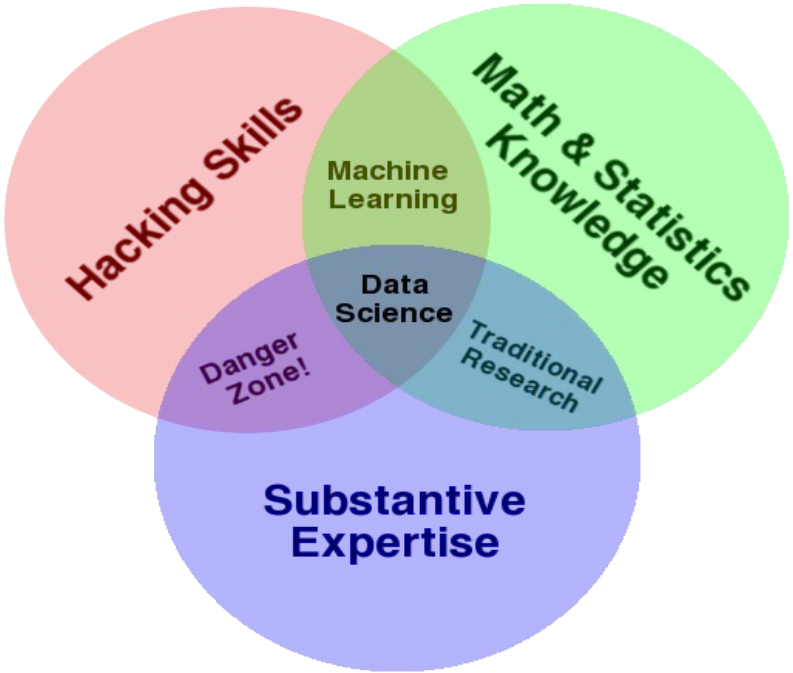
Data Science Overview

Credit: Udacity (online course, Intro to Data Science.)
Edureka Data Science course video

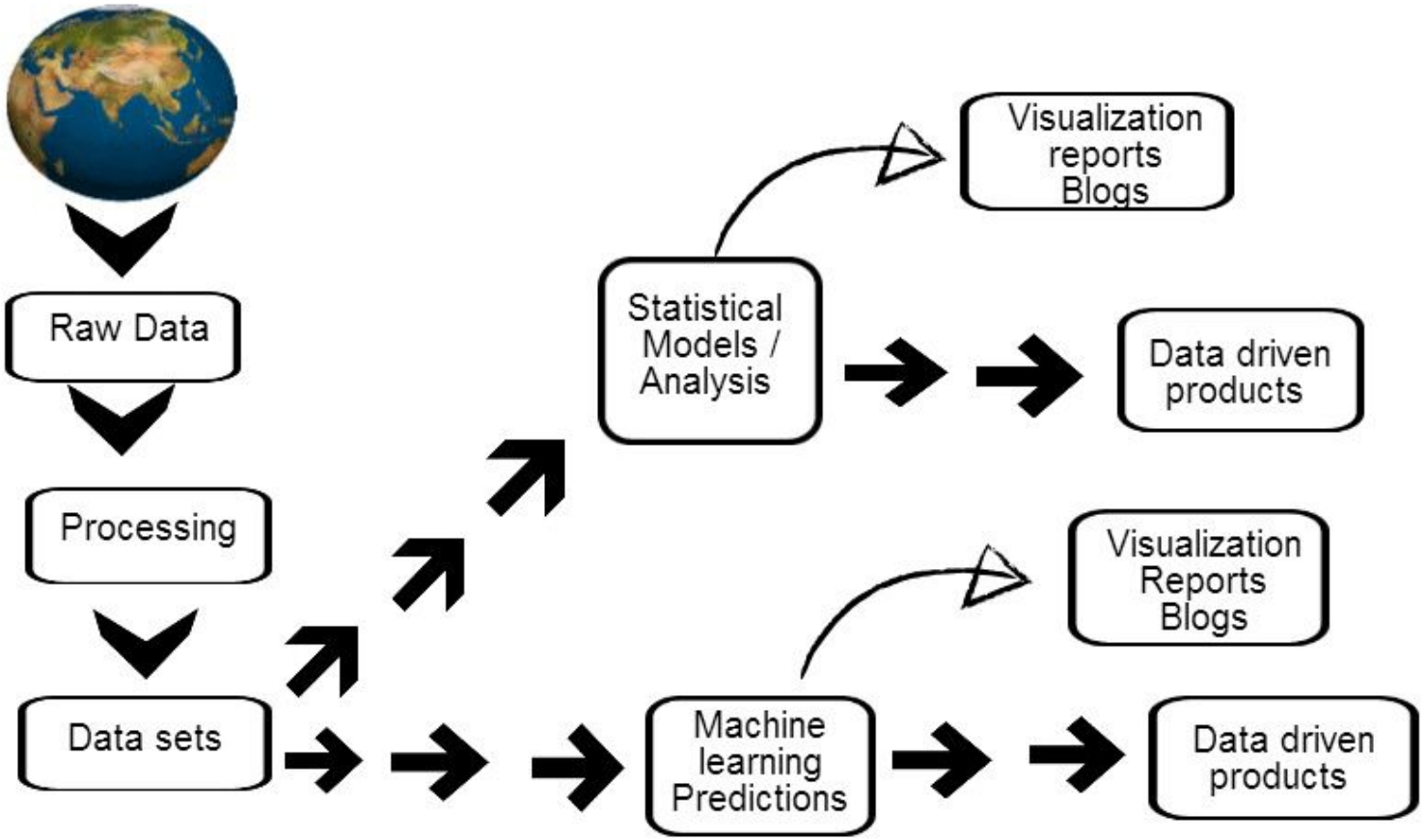
Discussion question

- What is a data scientist?
- What do you think data scientists do in day to day work?
- http://bigdata-madesimple.com/what-everybody-ought-to-know-about-data-scientist/

Data Scientist is



What does a data scientist do



Need Of Data Science

edureka!

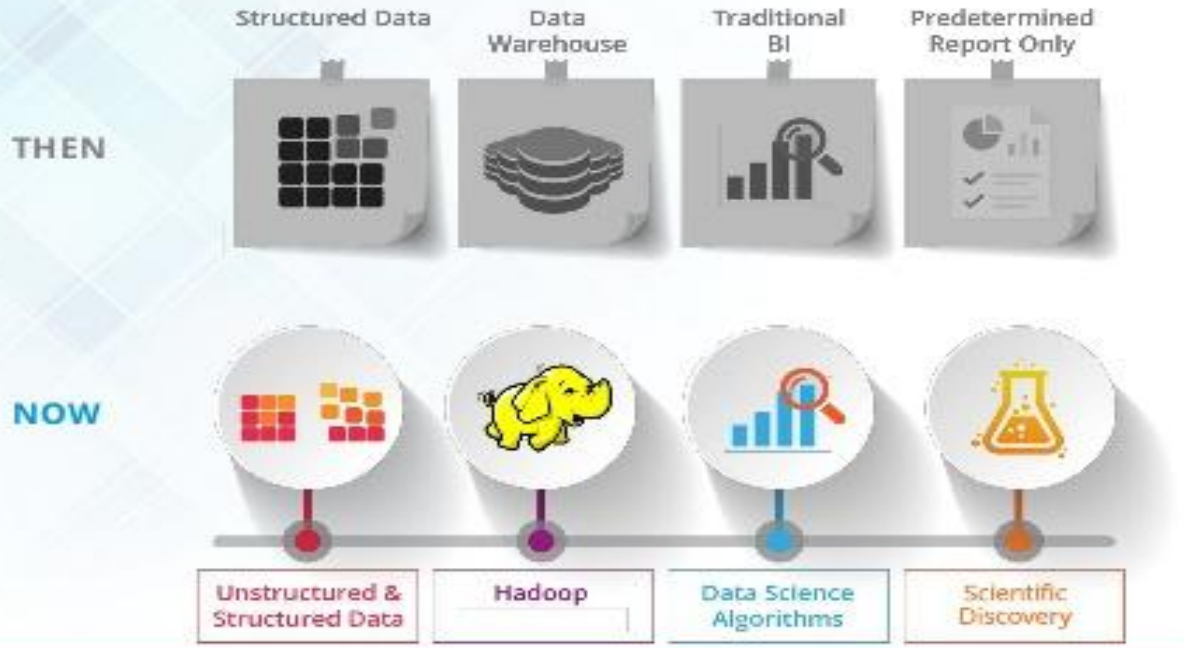
- Data Flow
- Unstructured Data
- Data Storage
- Lack of predictive analytics
- Lack of scientific insights

- Decision making
- Prediction
- Pattern discovery

Data Science

Need Of Data Science

edureka!



Guess which item goes next to diapers?

A: Formula

B: Beer

C: Newspaper

D: Wipes

E: Candy

Guess why :)

Need Of Data Science

edureka!

You can use Data Science to

- Recommend the right product to the right customer to enhance business.
- Predict the characteristics of high LTV customers and helps in customer segmentation.
- Build intelligence and ability in machines.
- Predict fraudulent transactions beforehand.
- Perform sentiment analysis to predict the outcome of elections.

What Is Data Science?

edureka!

- Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.
- Data Science is primarily used to make decisions and predictions.

What Is Data Science?

edureka!

- Google self driving car is a smart, driverless car.
- It collects data from environment through sensors.
- Takes decisions like when to speed up, when to speed down, when to overtake and when to turn.

Not perfect yet

https://www.theguardian.com/technology/2016/mar/09/google-self-driving-car-crash-video-accident-bus

BI Vs. Data Science

edureka!

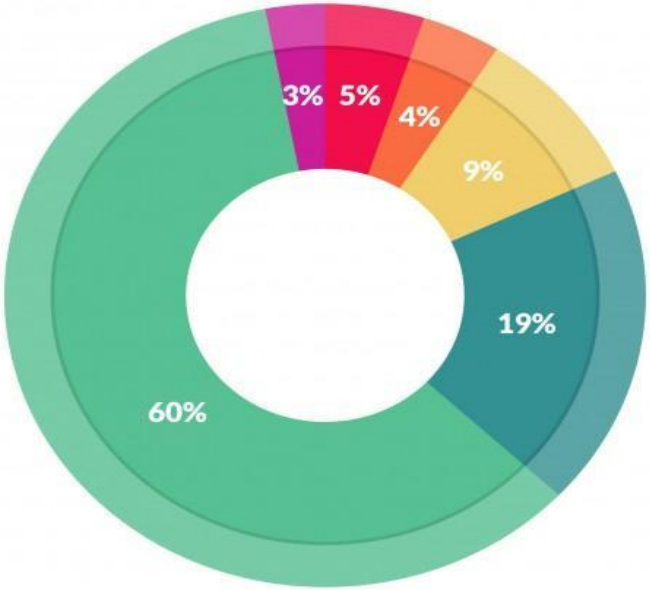
Characteristics	Business Intelligence	Data Science
Perspective	Looking Backward	Looking Forward
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured (logs, cloud data, SQL, NoSQL, text)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP)
Focus	Past and Present	Present and Future
Tools	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R

Lifecycle Of Data Science

edureka!

Data Collection and Data Preparation

How much time spent for data preparation?



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#7405fde6f637

First Project, part 1: Data Wrangling

Learning goals:

- Loading data from different sources
- Parsing the data
- Creating tables
- Manipulating tables
- Connecting together data from different sources

Done in Python 3

How many of you know how to program?

A: I’m confident

B: Not too bad, I know the basics

C: I’ve never programmed before

Data Wrangling. Data is everywhere

- Explore how to load and organize data into a usable formats
- A bit about file types, data formats, databases and APIs (Application Programming Interface)

Data Sources:

- Files
- Databases
- Web Scraping and API

File types: ‘friendly’ and ‘unfriendly’

1. txt
2. pdf
3. json
4. csv
5. docx
6. html
7. xml

What types are friendly?

A: 1, 2, 3, 4, 7
B: 1, 3, 5, 6
C: 1, 3, 4, 7
D: 2, 5, 6
E: All of them are fine, humans can understand all of them

File types: ‘friendly’ and ‘unfriendly’

Friendly

1. txt
2. tsv
3. csv
4. json
5. xml

Unfriendly

1. pdf
2. docx
3. html
4. Anything made to look nice for human

JSON format

- **J**ava**S**cript **O**bject **N**otation
- Used in Android, web technologies, APIs etc => very popular
- It is a *format* in which we pass data from client to server and back.

- Used to be HTML, CSS, JavaScript to create static pages
- Now we have dynamic pages
- Send a request to server, need to get data (only data) back in some format.

JSON format: JavaScript Object Notation

- We usually send a complex object back, not just a simple text
- Issue: parsing data (you will see that in project 1)
- We can send data in a JavaScript object
- Machine should be able to understand it, not you!

EMP

eid = 1
eName = Donald
Salary = 5000
Dep = CSE

JSON

```
{
        eid: “1”,
        eName: “Donald”,
        Salary: “5000”,
        Dep: “CSE”
}
```

JSON object can have JSON array
JSON array can have JSON object

So you can create pretty complex data objects.

(demo)

Project 1 Overview

Race

rider	horse_name	finish	date
Art	Ceasar	8	2/1/2001
Eric	Chunky	2	2/1/2001
Mike	Rambo	6	2/1/2001
Niklas	Flame	5	2/1/2001
Richard	Melody	1	2/1/2001
Tom	Mountain Rush	4	2/1/2001
Will	Dandy	3	2/1/2001
Greg	Ginger	7	2/1/2001
Steve	Heart of Stone	9	2/1/2001

Horses

horse_name	age	body_mark	color
Flame	3	Star	Chestnut
Gusty Wind	4	Star	Chestnut
Melody	4	Strip	Chestnut
Dandy	3	None	Black
Ginger	3	Star	Gray
Ceasar	3	None	Roan
Heart of Stone	2	Star	Black
Rocky	2	Star	Roan
Seven	2	None	
Batman	2	None	Dun
Rambo	1	None	Albino
Rodeo	1	Strip	
Mountain Rush	4	None	Albino
Beau Dancer	5	None	Chestnut
San Domingo	5	None	Chestnut
Chunky	5	Strip	Piebald

Join (merge) 2 tables by 'horse_name'

Merged Table

rider	horse_name	finish	date	age	body_mark	color
Art	Ceasar	8	2/1/01	3	None	Roan
Eric	Chunky	2	2/1/01	5	Strip	Piebald
Will	Dandy	3	2/1/01	3	None	Black
Niklas	Flame	5	2/1/01			Chestnut
Greg	Ginger	7	2/1/01			Gray
Steve	Heart of Stone	9	2/1/01	2	Star	Black
Richard	Melody	1	2/1/01	4	Strip	Chestnut
Tom	Mountain Rush	4	2/1/01	4	None	Albino
Mike	Rambo	6	2/1/01	1	None	Albino

Joining two tables by common column

How to look for help?

A: Ask Marina or Janine

B: Ask friends

C: Ask Google

D: Both A and B

E: A, B and C

Troubleshooting

- Tab complete - browse available methods
- Question mark - view documentation
- type() - see what type of object you are working with
- Google - look at examples of usage

Let's practice

Helpful link: http://data8.org/datascience/

1. Create a table with 4 columns and three rows.
    a. The content of the table is up to you

Now modify your table:

2. Change the value from second column and first row to something else
    a. Use a new notebook cell
    b. Count starts from 0
3. Relabel your leftmost column
4. Remove your last row