Data Science: Principles and Practice

Lecture 1: Introduction

Ekaterina Kochmar[1]

UNIVERSITY OF CAMBRIDGE

[1] Based on slides from Marek Rei

Data Science: Principles and Practice

01 Introduction and motivation

02 Practical basics

03 Course logistics

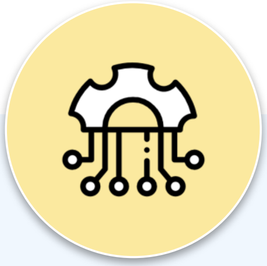
What is Data Science?

Data Processing

crawling
cleaning
connecting

Statistics

measuring
analyzing
exploring



Machine Learning

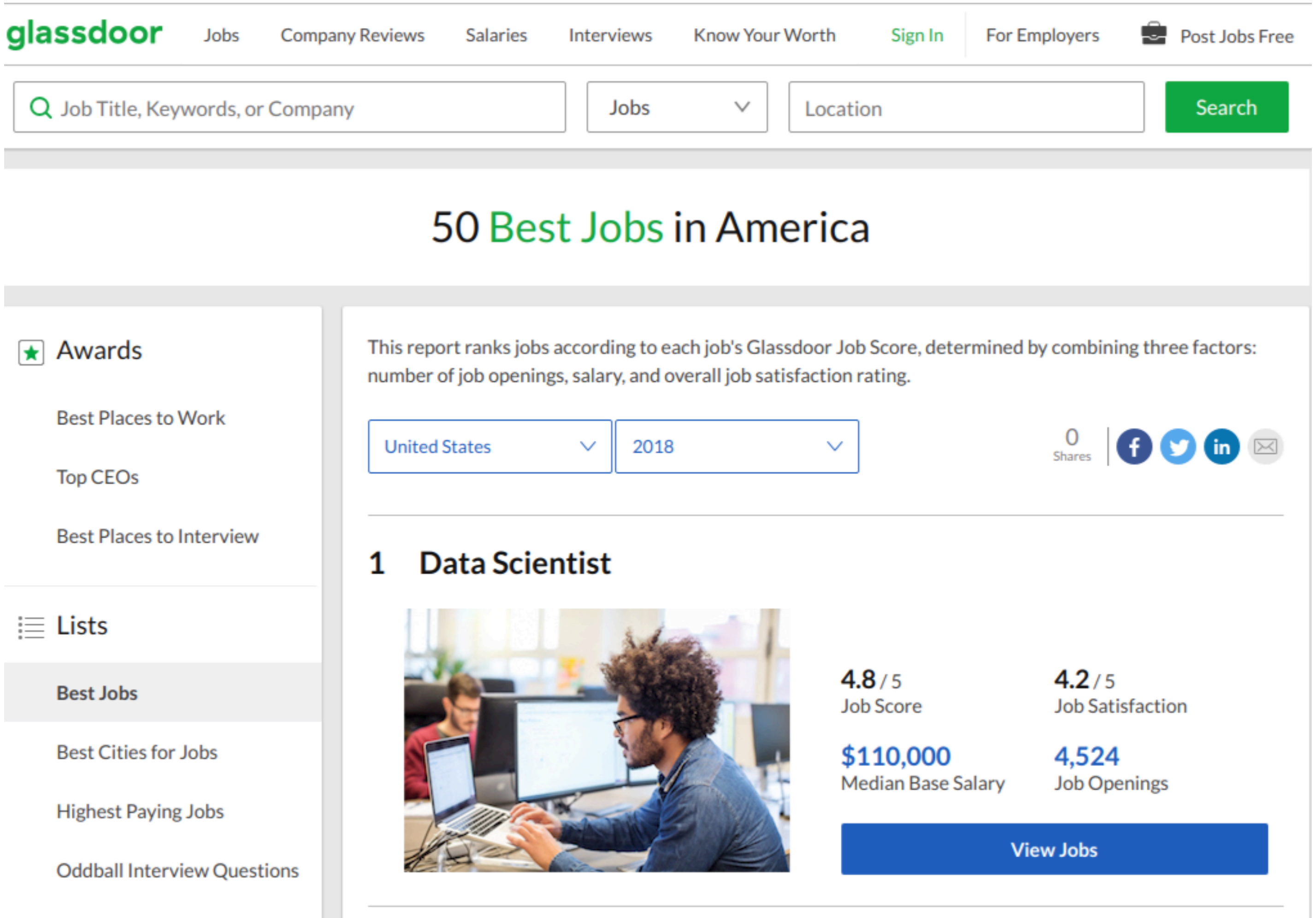
modeling
predicting
simulating

Visuali-zation

investigating
structuring
presenting

Big Data

processing
parallelizing
optimizing



https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm glassdoor Jobs Company Reviews Salaries Interviews Salary Calculator Write Review For Employers Post Jobs Free

Job Title, Keywords, or Company | Jobs | Location | Search

50 Best Jobs in America for 2019

Best Jobs | 2019 | United States — Share

	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Data Scientist	$108,000	4.3/5	6,510	View Jobs
#2	Nursing Manager	$83,000	4/5	13,931	View Jobs
#3	Marketing Manager	$82,000	4.2/5	7,395	View Jobs

https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

MENU Data Scientist: The Sexiest Job of the 21st Century Subscribe | Sign In | Register

SUMMARY SAVE SHARE 16 COMMENT TEXT SIZE PRINT $8.95 BUY COPIES

DATA

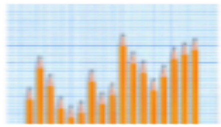
Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

WHAT TO READ NEXT



What Data Scientists Really Do, According to 35 Data Scientists

VIEW MORE FROM THE

October 2012 Issue

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

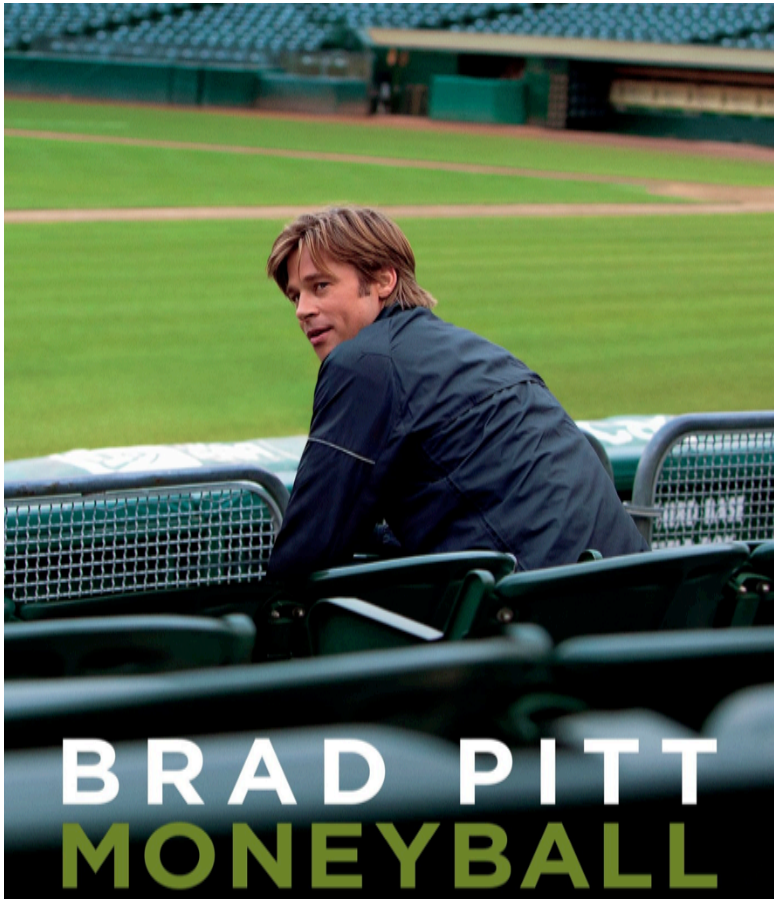
Case studies

01 Sports

02 Medicine

03 Politics

Data Science in Sports

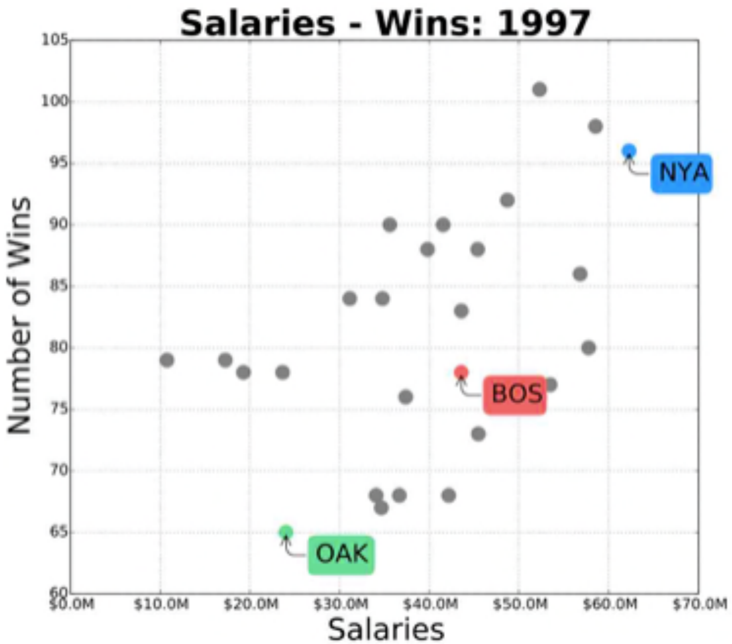


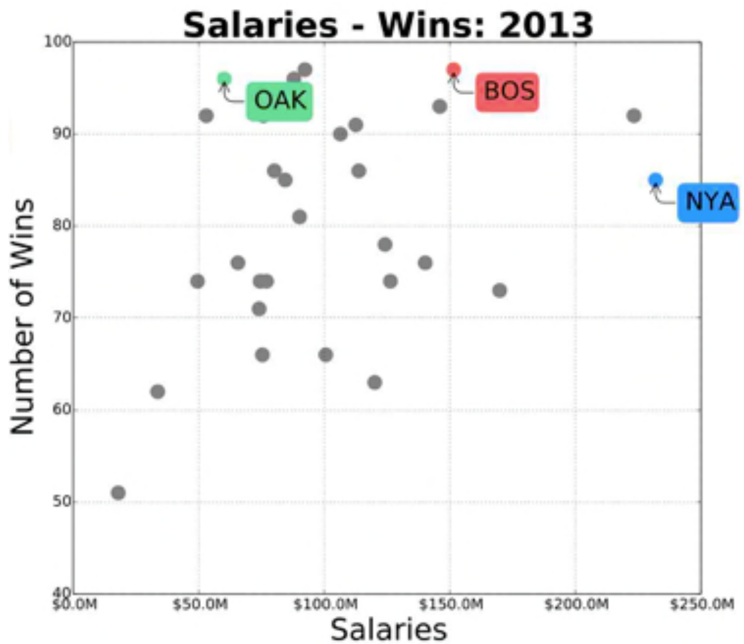
The market for baseball
players was so inefficient…
that superior management
could run circles around taller
piles of cash.
- Michael Lewis

Legendary 2002 season for Oakland Athletics.

Manager Billy Beane put together an unexpected team using data science.

Data Science in Sports





http://adilmoujahid.com/posts/2014/07/baseball-analytics/

Data Science in Drug Discovery

1:10,000 success rate

1:10 success rate

$350M to $5.5B cost

http://sitn.hms.harvard.edu/flash/2017/make-fda-great-trump-future-drug-approval-process/
https://en.wikipedia.org/wiki/Cost_of_drug_development

Data Science in Drug Discovery

nature

SPOTLIGHT · 30 MAY 2018

How artificial intelligence is changing drug discovery

Machine learning and other technologies are expected to make the hunt for new pharmaceuticals quicker, cheaper and more effective.

Nic Fleming

PDF version

RELATED ARTICLES

The drug-maker's guide to the galaxy

BenchSci Blog

106 Startups Using Artificial Intelligence in Drug Discovery

Simon Smith
Last Updated Oct 1, 2018

2.2k Shares | in 1.1k | 372 | 354 | 198

Some time ago, I wrote about how we're now in the long-tail of machine learning in drug discovery. I noted that we're moving past generalist applications of AI such as IBM Watson's to more specific, purpose-built tools. This got me thinking: What *are* all the startups applying artificial intelligence in drug discovery

https://www.nature.com/articles/d41586-018-05267-x
https://blog.benchsci.com/startups-using-artificial-intelligence-in-drug-discovery

FiveThirtyEight

ABC NEWS

Politics Sports Science & Health Economics Culture

NOV. 4, 2008, AT 6:16 PM

Today's Polls and Final Election Projection: Obama 349, McCain 189

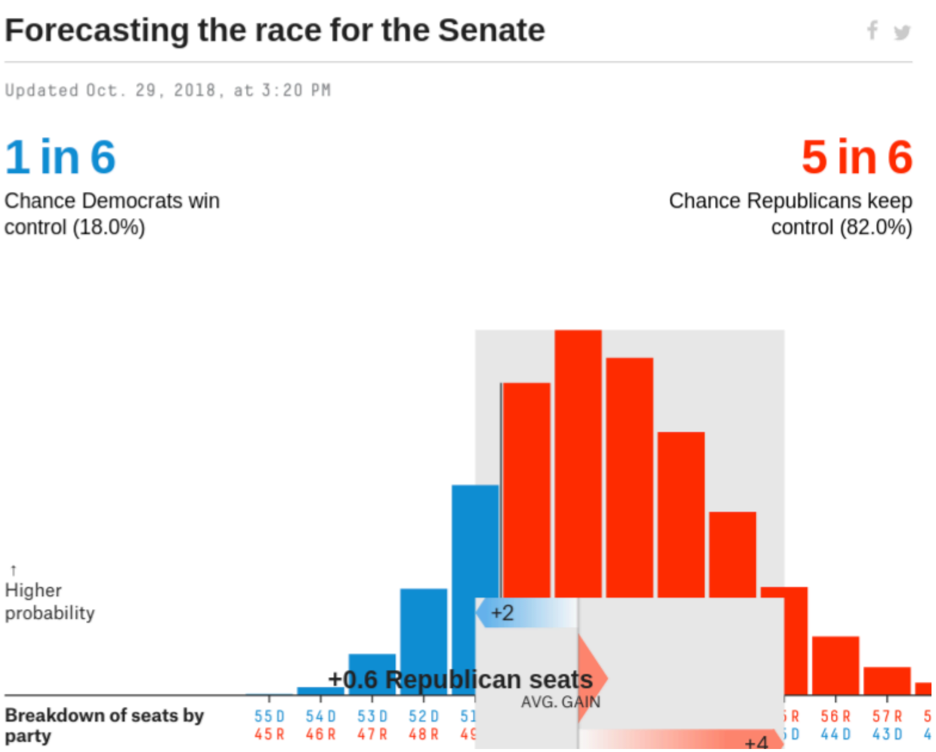
By Nate Silver

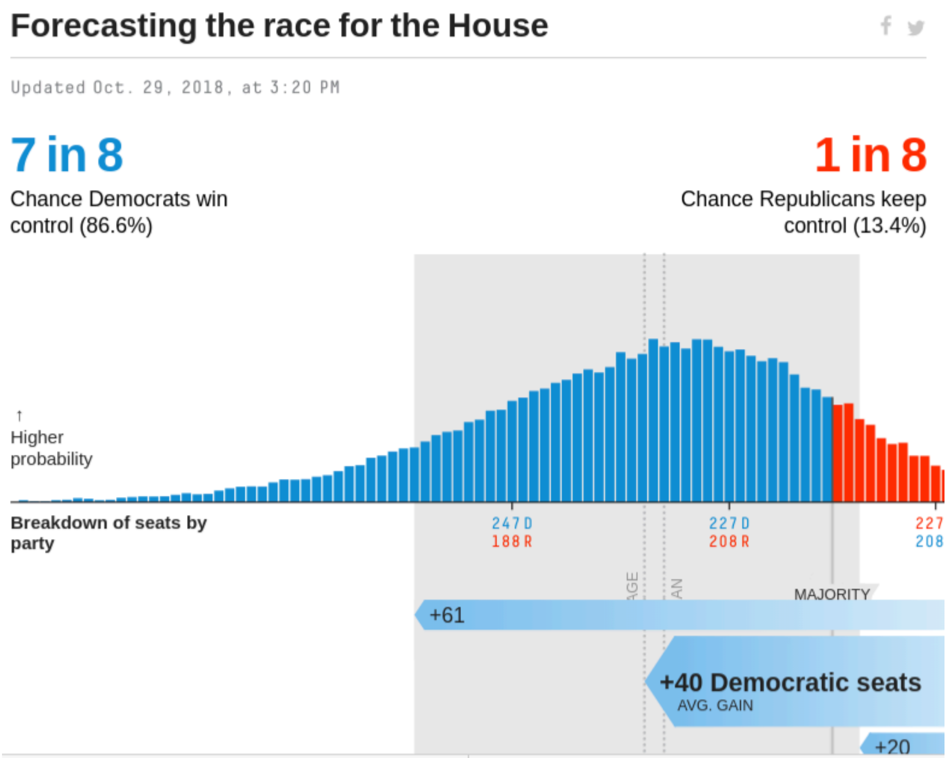
It's Tuesday, November 4th, 2008, Election Day in America. The last polls have straggled in, and show little sign of mercy for John McCain. Barack Obama appears poised for a decisive electoral victory.

Our model projects that Obama will win all states won by John Kerry in 2004, in addition to Iowa, New Mexico, Colorado, Ohio, Virginia, Nevada, Florida and North Carolina, while narrowly losing Missouri

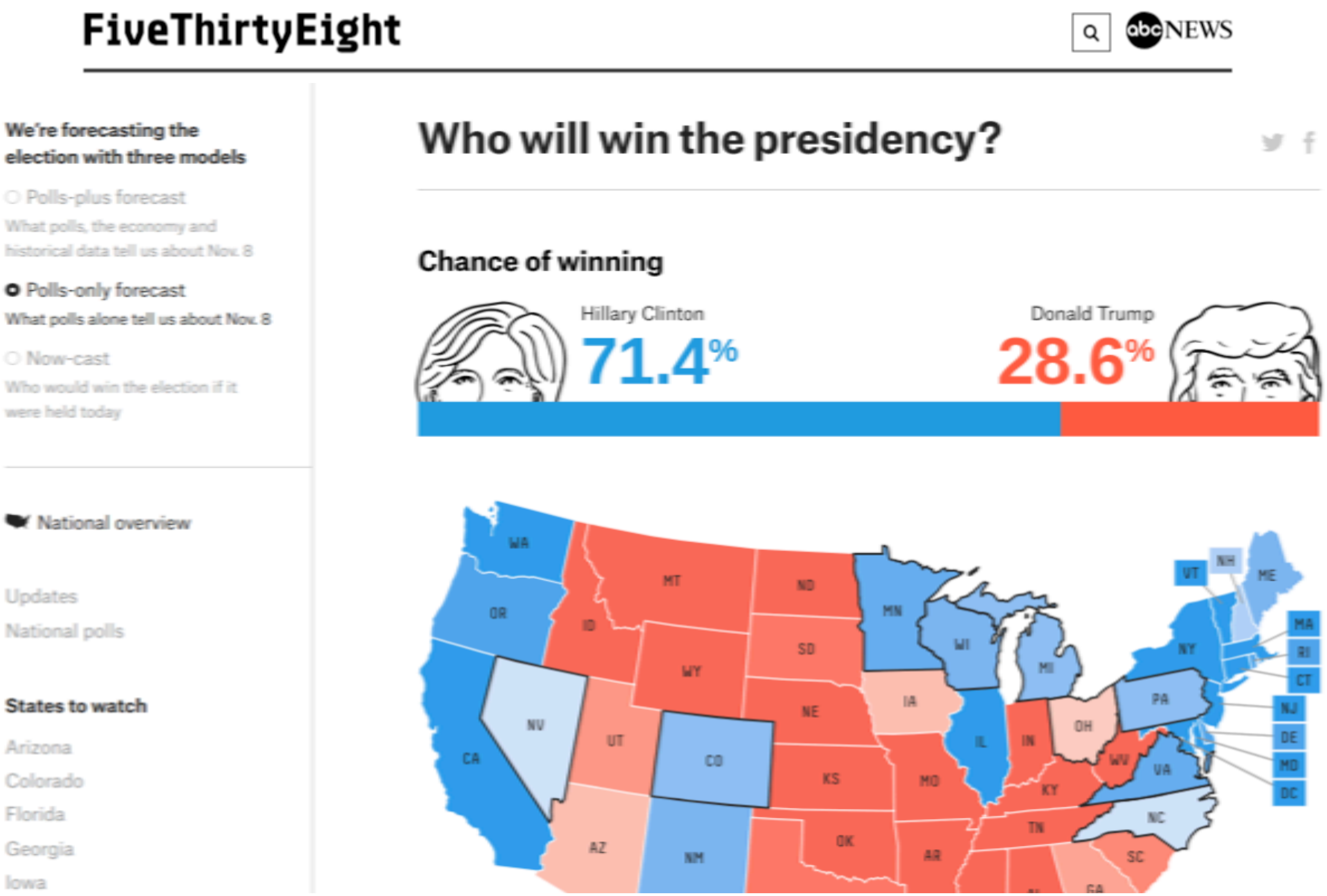
https://fivethirtyeight.com/features/todays-polls-and-final-election/

Data Science in Politics





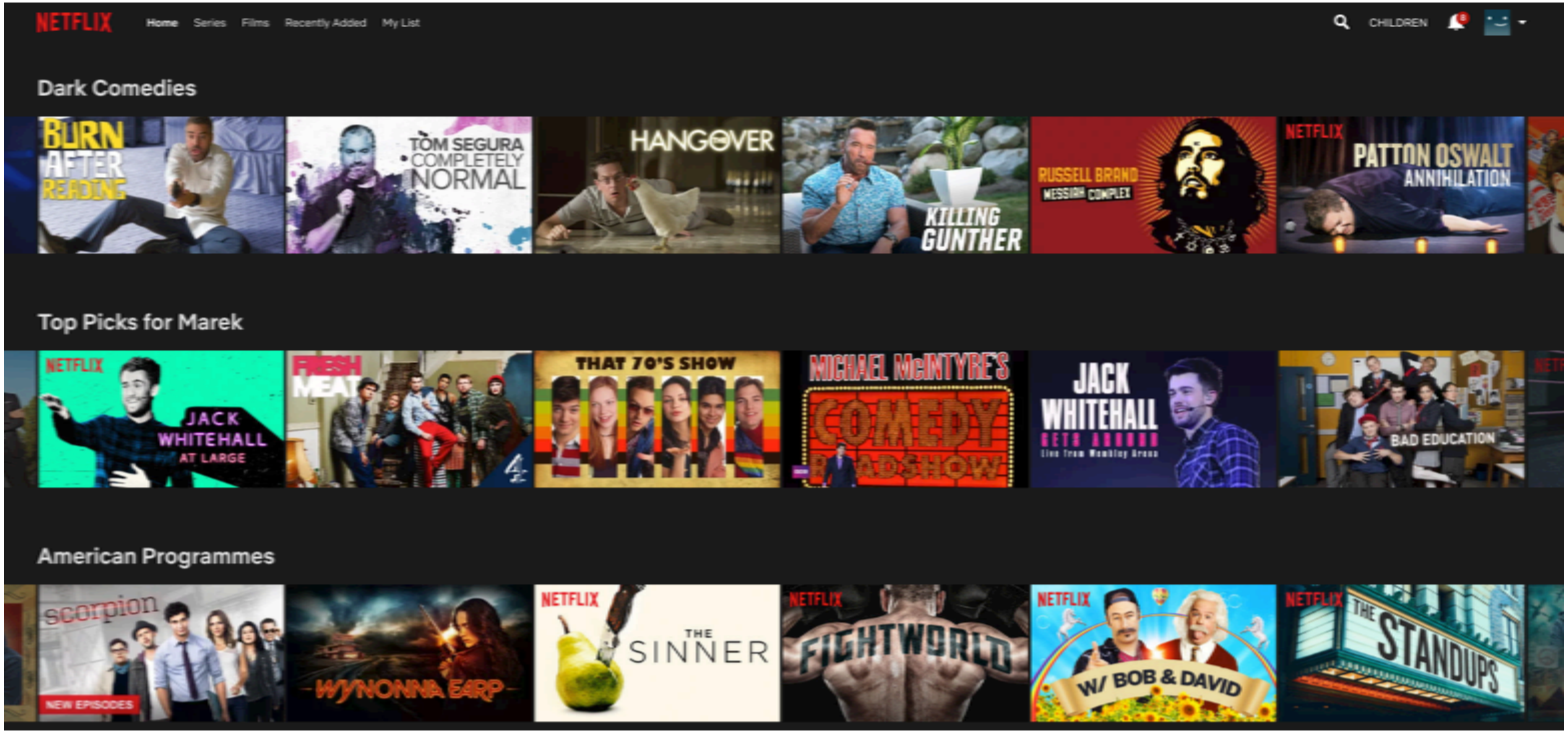
https://fivethirtyeight.com/tag/2018-election/



https://projects.fivethirtyeight.com/2016-election-forecast/

Data Science in Commerce

Data Science in Commerce



Netflix Challenge

In 2006, Netflix offered 1 million dollars for an improved movie recommendation algorithm.

Provided 100M movie ratings for training.

The goal: Improve over Netflix's own algorithm by 10% to get the prize.

Several teams joined up and claimed the prize on in 2009.

movie	user	date	score
1	56	2004-02-14	5
1	25363	2004-03-01	3
2	855321	2004-07-29	3
2	44562	2004-07-30	4
3	42357	2004-12-10	1
3	1345	2005-01-08	2

Data Science in Climate Control

How Data Science can help solve Climate Change

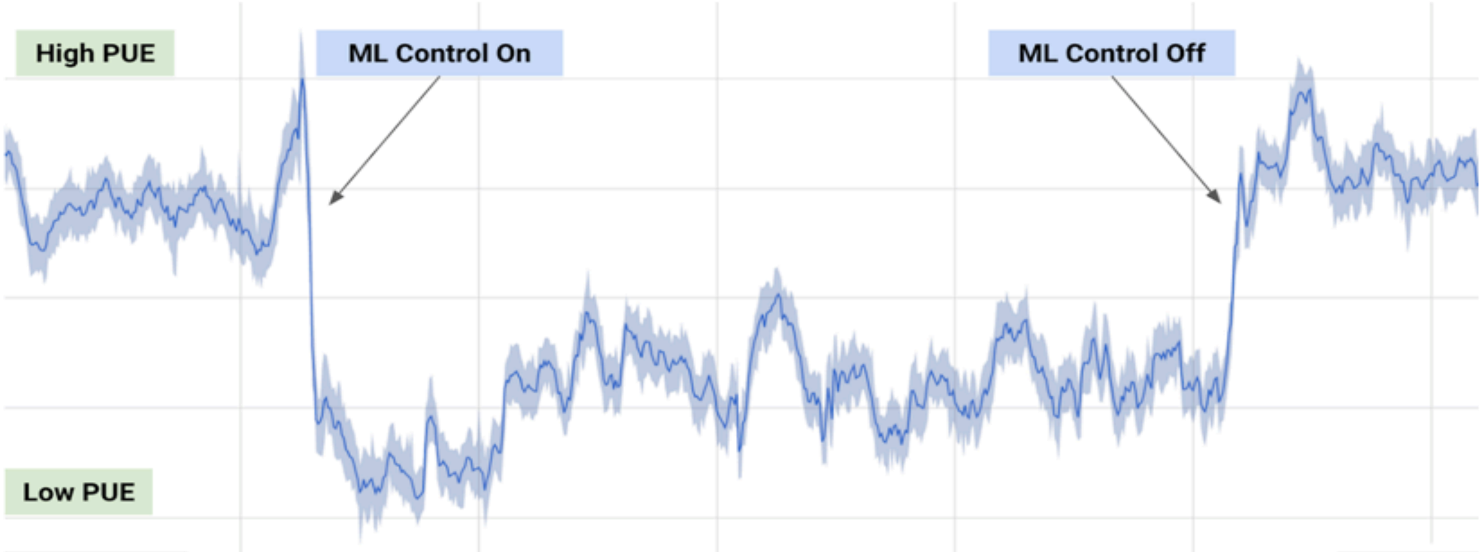
Data-driven solutions will lead the Transition to Clean Energy

Marco Pasini Follow
Aug 21 · 6 min read ★

Photo by Bogdan Pasca on Unsplash

https://towardsdatascience.com/how-data-science-can-help-solve-climate-change-12b28768e77b

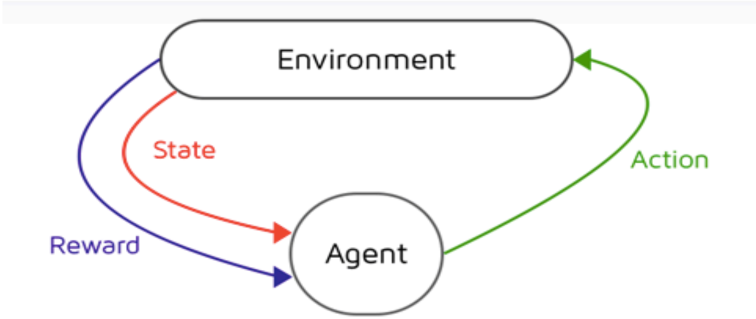
Data Science in Climate Control



Our machine learning system was able to consistently achieve a 40 percent reduction in the amount of energy used for cooling, which equates to a 15 percent reduction in overall PUE overhead after accounting for electrical losses and other non-cooling inefficiencies. It also produced the lowest PUE the site had ever seen.

https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40

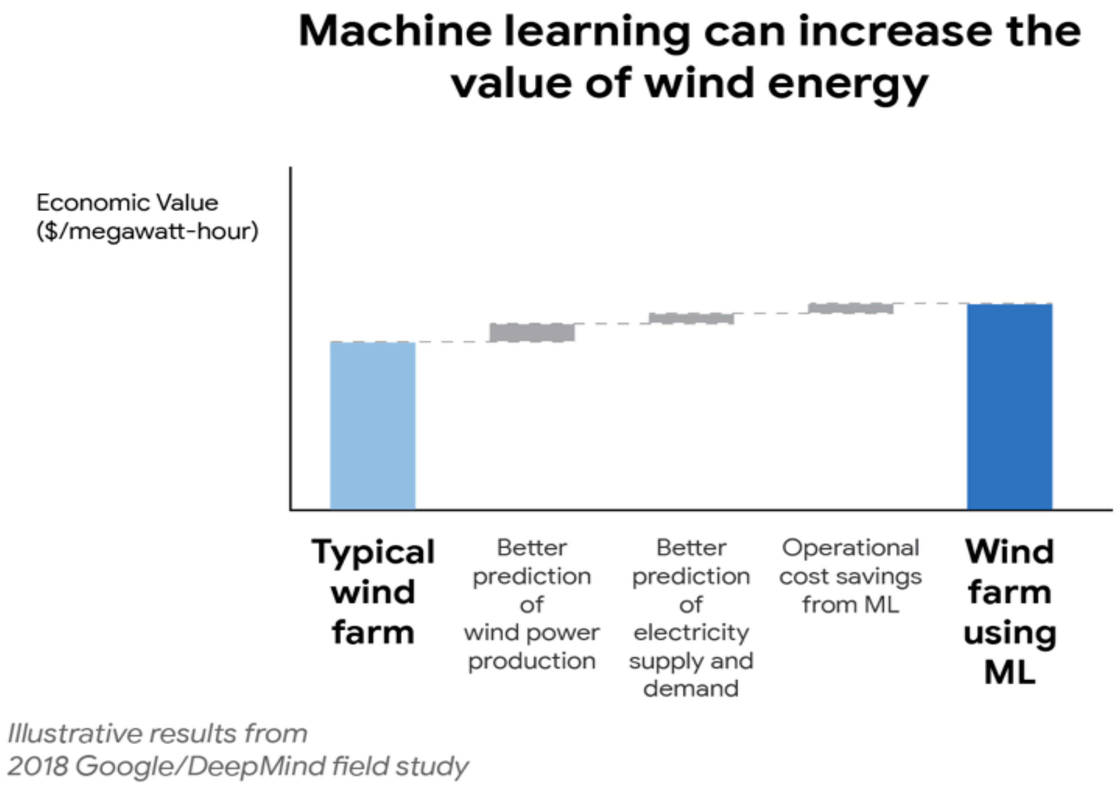
Data Science in Climate Control



A number of **recent studies** propose Reinforcement Learning (RL, a branch of machine learning in which an **agent** interacts with an **environment**, becoming progressively better at a specified **goal** defined by a reward function) as the solution: applying this kind of algorithm to increase efficiency of different buildings shows incredible and **promising results**, with **up to 70% (!!!) reduction** in HVAC energy usage (source).

https://ywang393.expressions.syr.edu/wp-content/uploads/2016/07/Deep-reinforcement-learning-for-HVAC-control-in-smart-buildings.pdf

Data Science in Climate Control



https://deepmind.com/blog/article/machine-learning-can-boost-value-wind-energy

Getting Practical

Dataset: Country Statistics

World Bank data about 161 countries

- Country Name
- GDP per Capita (PPP USD)
- Population Density (persons per sq km)
- Population Growth Rate (%)
- Urban Population (%)
- Life Expectancy at Birth (avg years)
- Fertility Rate (births per woman)
- Infant Mortality (deaths per 1000 births)
- Enrolment Rate, Tertiary (%)
- Unemployment, Total (%)
- Estimated Control of Corruption (scale -2.5 to 2.5)
- Estimated Government Effectiveness (scale -2.5 to 2.5)
- Internet Users (%)

Dataset: Country Statistics

```
Country Name,GDP per Capita (PPP USD),Population Density (persons per sq km),Population Growth Rate (%),Urban
Population (%),Life Expectancy at Birth (avg years),Fertility Rate (births per woman),Infant Mortality (deaths
per 1000 births),"Enrolment Rate, Tertiary (%)","Unemployment, Total (%)",Estimated Control of Corruption (scale
-2.5 to 2.5),Estimated Government Effectiveness (scale -2.5 to 2.5),Internet Users (%)
Afghanistan,1560.67,44.62,2.44,23.86,60.07,5.39,71,3.33,8.5,-1.41,-1.4,5.45
Albania,9403.43,115.11,0.26,54.45,77.16,1.75,15,54.85,14.2,-0.72,-0.28,54.66
Algeria,8515.35,15.86,1.89,73.71,70.75,2.83,25.6,31.46,10,-0.54,-0.55,15.23
Antigua and Barbuda,19640.35,200.35,1.03,29.87,75.5,2.12,9.2,14.37,8.4,1.29,0.48,83.79
Argentina,12016.2,14.88,0.88,92.64,75.84,2.2,12.7,74.83,7.2,-0.49,-0.25,55.8
Armenia,8416.82,104.08,0.17,64.16,74.33,1.74,14.7,48.94,18.4,-0.62,-0.04,39.16
Australia,44597.83,2.91,1.6,89.34,81.85,1.87,4.1,83.24,5.2,2,1.61,82.35
Austria,43661.15,102.22,0.46,67.88,81.03,1.42,3.3,71,4.3,1.35,1.66,81
Azerbaijan,10125.23,110.98,1.35,53.89,70.55,1.92,38.5,19.65,5.2,-1.13,-0.79,54.2
Bahrain,24590.49,1701.01,1.92,88.76,76.4,2.12,8.2,33.46,1.1,0.39,0.65,88
Bangladesh,1883.05,1174.33,1.19,28.89,69.89,2.24,33.1,13.15,5,-0.87,-0.83,6.3
Barbados,26487.77,655.36,0.5,44.91,74.97,1.84,16.9,60.84,11.6,1.66,1.45,73.33
Belgium,39751.48,364.85,0.85,97.51,80.49,1.84,3.4,69.26,7.5,1.55,1.59,82
Belize,7936.84,13.87,2.43,44.59,73.49,2.74,15.7,21.37,8.2,0.01,-0.18,25
Benin,1557.16,86.73,2.73,45.56,58.94,5.21,58.5,12.37,0.7,-0.92,-0.53,3.8
Bhutan,6590.69,19,1.68,36.34,67.28,2.32,35.7,8.74,2.1,0.82,0.48,25.43
Bolivia,5195.58,9.53,1.65,67.22,66.63,3.31,39.3,37.69,3.4,-0.7,-0.37,34.19
Bosnia and Herzegovina,9392.47,75.28,-0.14,48.81,75.96,1.25,6.7,37.74,28.1,-0.3,-0.47,65.36
Brazil,11715.7,23.28,0.87,84.87,73.35,1.81,12.9,25.63,6.7,-0.07,-0.12,49.85
Brunei,52482.33,77.14,1.4,76.32,78.07,2.03,6.7,24.34,4.7,0.64,0.83,60.27
Bulgaria,15932.63,67.69,-0.6,73.64,74.16,1.51,10.5,59.63,11.2,-0.24,0.14,55.15
Burkina Faso,1512.97,58.46,2.86,27.35,55.44,5.78,65.8,4.56,3.3,-0.52,-0.63,3.73
Burundi,551.27,371.51,3.19,11.21,53.14,6.21,66.9,3.17,0.5,-1.12,-1.33,1.22
Cambodia,2494.39,82.74,1.76,20.19,62.98,2.93,33.9,14.5,0.2,-1.04,-0.83,4.94
```

Using Python. Why Python?

Fast to write and modify

Great for working with datasets

Portable

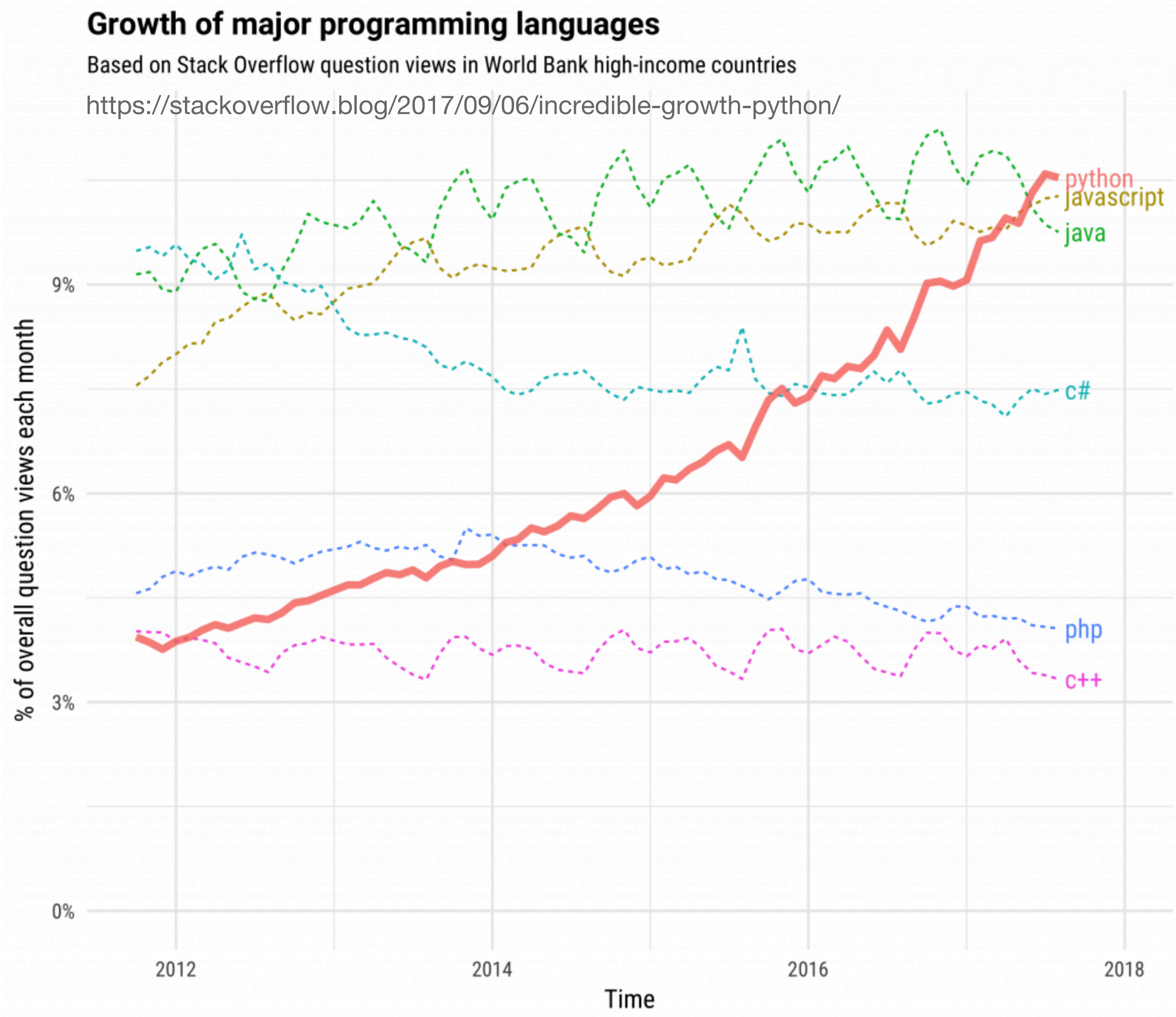
Most machine learning research happens in python

Actually useful for other things besides data science

Dynamically typed (can cause run-time errors)

Not as fast as lower-level languages (sometimes)

Not good for unusual platforms



Growth of major programming languages
Based on Stack Overflow question views in World Bank high-income countries
https://stackoverflow.blog/2017/09/06/incredible-growth-python/
% of overall question views each month
9%
6%
3%
0%
2012
2014
2016
2018
Time
python
javascript
java
c#
php
c++

Python Refresher

```
In [1]: import random

        my_list = ["camel", "elephant", "crocodile"]
        for word in my_list:
            print(word + " " +str(random.random()))
```

```
camel 0.5333896529549417
elephant 0.8289440919886492
crocodile 0.5635699354595317
```

Python tutorial: https://www.tutorialspoint.com/python/index.htm

Loading CSV files

In [2]:
```python
import pandas as pd

data = pd.read_csv('data/country-stats.csv')
data.head()
```

Out[2]:

	Country Name	GDP per Capita (PPP USD)	Population Density (persons per sq km)	Population Growth Rate (%)	Urban Population (%)	Life Expectancy at Birth (avg years)	Fertility Rate (births per woman)	Infant Mortality (deaths per 1000 births)
0	Afghanistan	1560.67	44.62	2.44	23.86	60.07	5.39	71.0
1	Albania	9403.43	115.11	0.26	54.45	77.16	1.75	15.0
2	Algeria	8515.35	15.86	1.89	73.71	70.75	2.83	25.6
3	Antigua and Barbuda	19640.35	200.35	1.03	29.87	75.50	2.12	9.2
4	Argentina	12016.20	14.88	0.88	92.64	75.84	2.20	12.7

Common File Formats

CSV - comma-separated values

```
Bahrain,24590.49,1701.01,1.92,88.76,76.4,2.12,8.2,33.46,1.1,0.39,0.65,88
Bangladesh,1883.05,1174.33,1.19,28.89,69.89,2.24,33.1,13.15,5,-0.87,-0.83,6.3
Barbados,26487.77,655.36,0.5,44.91,74.97,1.84,16.9,60.84,11.6,1.66,1.45,73.33
Belgium,39751.48,364.85,0.85,97.51,80.49,1.84,3.4,69.26,7.5,1.55,1.59,82
```

TSV - tab-separated values

```
Bahrain	24590.49	1701.01	1.92	88.76	76.4	2.12	8.2	33.46
Bangladesh	1883.05	1174.33	1.19	28.89	69.89	2.24	33.1	13.15
Barbados	26487.77	655.36	0.5	44.91	74.97	1.84	16.9	60.84
Belgium	39751.48	364.85	0.85	97.51	80.49	1.84	3.4	69.26
```

Common File Formats

JSON:
JavaScript Object Notation

```json
{
    "firstName": "John",
    "lastName": "Smith",
    "isAlive": true,
    "age": 27,
    "address": {
        "streetAddress": "21 2nd Street",
        "city": "New York",
        "state": "NY",
        "postalCode": "10021-3100"
    }
}
```

XML:
Extensible Markup Language

```xml
<?xml version="1.0" encoding="UTF-8"?>
<breakfast_menu>
   <food>
      <name>Belgian Waffles</name>
      <price>$5.95</price>
      <desc>Famous Belgian Waffles</desc>
      <calories>650</calories>
   </food>
</breakfast_menu>
```

Calculating Statistics over the Data

```
In [3]: data["GDP per Capita (PPP USD)"].mean()

Out[3]: 15616.289378881998
```

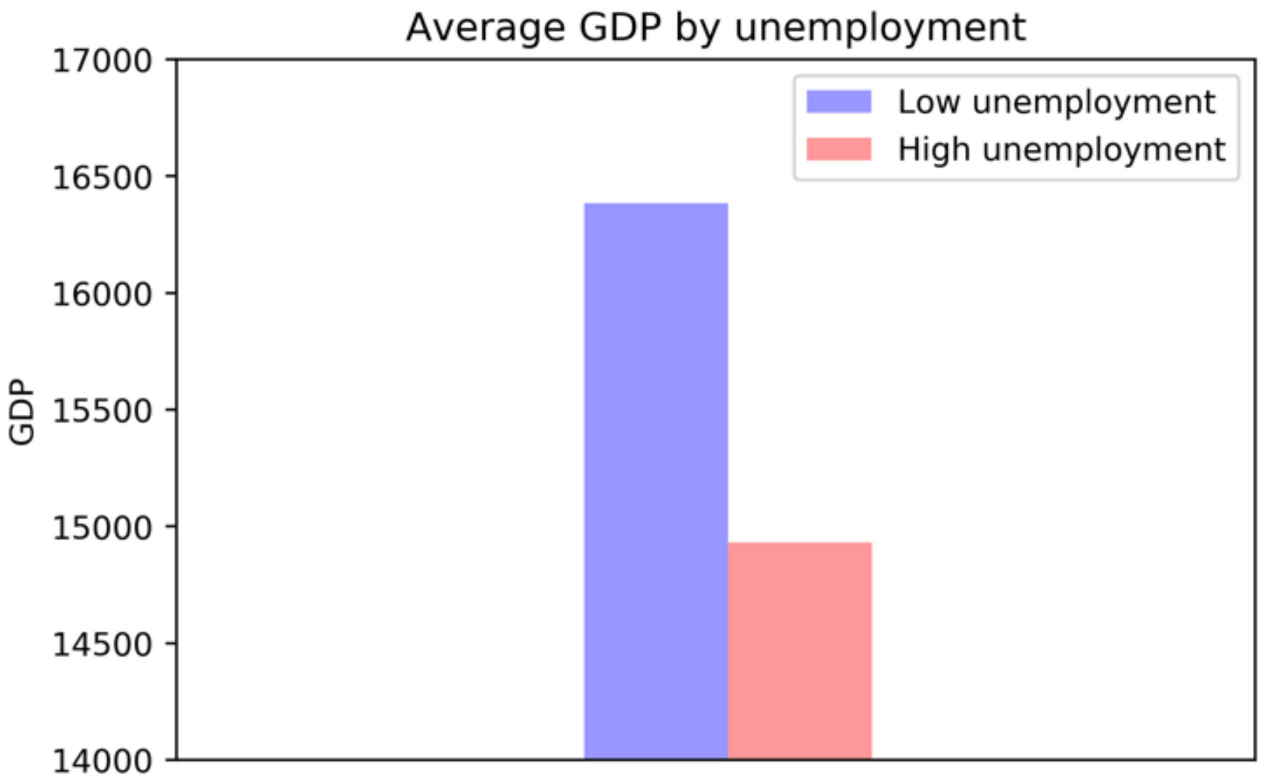
```
In [4]: low_unemployment_countries = data[data["Unemployment, Total (%)"] < 7]
        low_unemployment_countries["GDP per Capita (PPP USD)"].mean()

Out[4]: 16383.713421052627
```

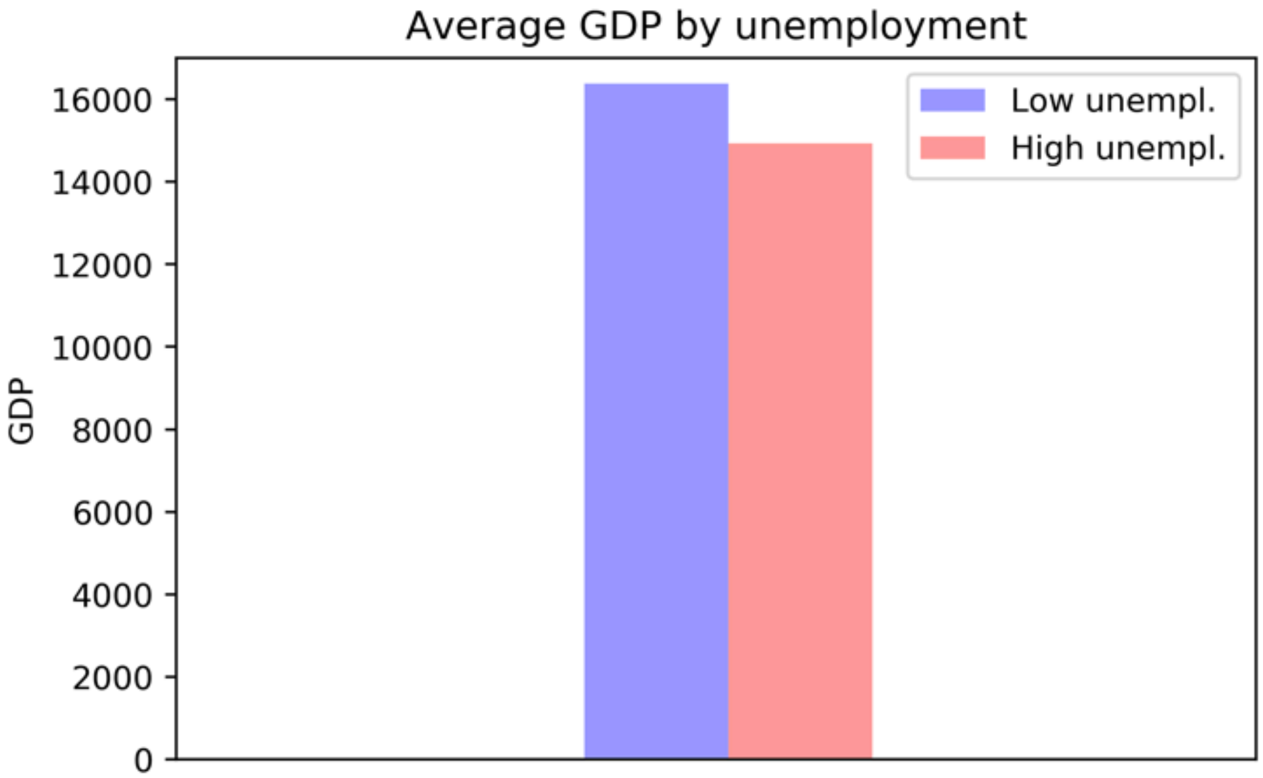
```
In [5]: high_unemployment_countries = data[data["Unemployment, Total (%)"] >= 7]
        high_unemployment_countries["GDP per Capita (PPP USD)"].mean()

Out[5]: 14930.121999999996
```

Calculating Statistics over the Data



Calculating Statistics over the Data



Calculating Statistics over the Data

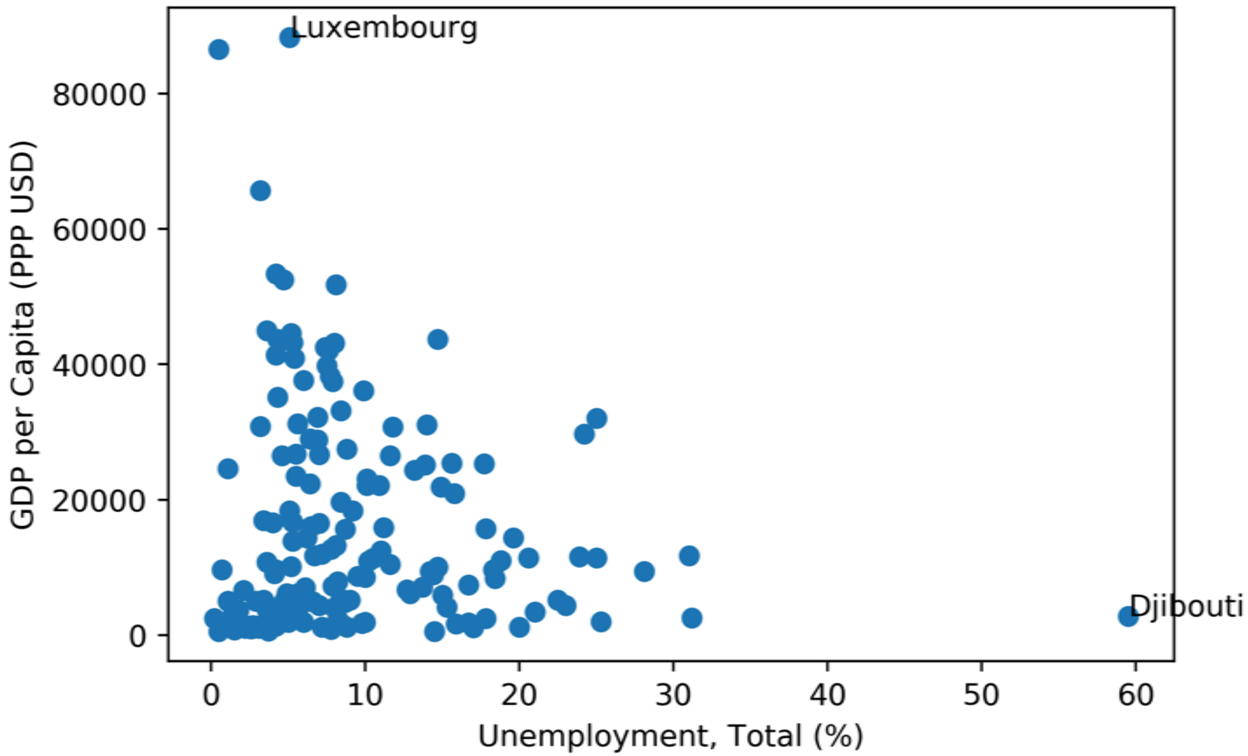
```
In [9]: low_unemployment_countries = data[data["Unemployment, Total (%)"] < 7]
        low_unemployment_countries["GDP per Capita (PPP USD)"].std()

Out[9]: 19752.912647780504
```

```
In [10]: high_unemployment_countries = data[data["Unemployment, Total (%)"] >= 7]
         high_unemployment_countries["GDP per Capita (PPP USD)"].std()

Out[10]: 12781.059320722152
```

Calculating Statistics over the Data



Course Logistics

Course Objectives

Focusing on the practical aspects of data science

After this course you should be able to

1. Understand the principles of data science
2. Use the necessary software tools for data processing, statistics and machine learning
3. Visualize data, both for exploration and presentation
4. Rigorously analyze your data using a variety of approaches

Course Format

10 lectures

6 practicals

Assessment

- 20% from practicals (pass/fail)
- 80% from take-home assignment

Final assignment

- Practical exercise
- Given out after the lecture on 25 November
- Submit a report
- The report will be marked by two assessors

Course Syllabus

1. Introduction	Friday, 8 November
2. Linear Regression	Monday, 11 November
3. **Practical1:** Linear Regression	Tuesday, 12 November
4. Classification	Wednesday, 13 November
5. **Practical2:** Classification	Thursday, 14 November
6. Ensemble-based models	Monday, 18 November
7. **Practical3:** Ensemble models	Tuesday, 19 November
8. Visualization, part I	Wednesday, 20 November

Course Syllabus

9. Visualization, part II	Friday, 22 November
10. Deep Learning basics	Monday, 25 November
11. **Practical4:** Visualization	Tuesday, 26 November
12. Deep Learning with TensorFlow	Wednesday, 27 November
13. **Practical5:** Deep Learning I	Thursday, 28 November
14. Deep Learning architectures	Friday, 29 November
15. Challenges in Data Science	Monday, 2 December
16. **Practical6:** Deep Learning II	Tuesday, 3 December

Lecturers

Ekaterina Kochmar
ek358

Guy Emerson
gete2

Damon Wischik
djw1005

Course Pages

Course homepage: https://www.cl.cam.ac.uk/teaching/1920/DataSciII/

Azure Notebooks: https://notebooks.azure.com/ek358/projects/data-science-pnp-1920

Getting started with Azure Notebooks: https://notebooks.azure.com/ek358/projects/data-science-pnp-1920/getting-started.ipynb

Github: https://github.com/ekochmar/cl-datasci-pnp