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

Visualization
- investigating
- structuring
- presenting

Big Data
- processing
- parallelizing
- optimizing
50 Best Jobs in America

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

1. Data Scientist

- 4.8/5 Job Score
- 4.2/5 Job Satisfaction
- $110,000 Median Base Salary
- 4,524 Job Openings

View Jobs

# 50 Best Jobs in America for 2019

<table>
<thead>
<tr>
<th>Rank</th>
<th>Job Title</th>
<th>Median Base Salary</th>
<th>Job Satisfaction</th>
<th>Job Openings</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Data Scientist</td>
<td>$108,000</td>
<td>4.3/s</td>
<td>6,510</td>
</tr>
<tr>
<td>#2</td>
<td>Nursing Manager</td>
<td>$83,000</td>
<td>4/s</td>
<td>13,931</td>
</tr>
<tr>
<td>#3</td>
<td>Marketing Manager</td>
<td>$82,000</td>
<td>4.2/s</td>
<td>7,395</td>
</tr>
</tbody>
</table>
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 stood in the corner sipping your drink...and you probably have email.”

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_
Case studies

01 Sports

02 Medicine

03 Politics
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
2-5 years
Basic Science Research

1:10 success rate
1-2 years
Preclinical Testing

5-7 years
Clinical Trials

1/2-2 years
Government Approval

$350M to $5.5B cost
Approved Drug

Data Science in Drug Discovery

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

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
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.
Data Science in Politics

Forecasting the race for the Senate

1 in 6
Chance Democrats win control (18.0%)

Forecasting the race for the House

5 in 6
Chance Republicans keep control (82.0%)

7 in 8
Chance Democrats win control (86.6%)

1 in 8
Chance Republicans keep control (13.4%)

https://fivethirtyeight.com/tag/2018-election/
Who will win the presidency?

Chance of winning

Hillary Clinton 71.4%
Donald Trump 28.6%
Data Science in Commerce

Recommendations for you in Electronics & Photo

Pick of the day  See all →

Amazon

Ebay

£27.95

£24.00

£179.99

£24.99

£14.59

£42.99
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.

<table>
<thead>
<tr>
<th>movie</th>
<th>user</th>
<th>date</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56</td>
<td>2004-02-14</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>25363</td>
<td>2004-03-01</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>855321</td>
<td>2004-07-29</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>44562</td>
<td>2004-07-30</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>42357</td>
<td>2004-12-10</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1345</td>
<td>2005-01-08</td>
<td>2</td>
</tr>
</tbody>
</table>
Data Science in Climate Control

How Data Science can help solve Climate Change
Data-driven solutions will lead the Transition to Clean Energy

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
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).

Data Science in Climate Control

Machine learning can increase the value of wind energy

Illustrative results from 2018 Google/DeepMind field study

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 (%)
<table>
<thead>
<tr>
<th>Country Name</th>
<th>GDP per Capita (PPP USD)</th>
<th>Population Density (persons per sq km)</th>
<th>Population Growth Rate (%)</th>
<th>Urban Population (%)</th>
<th>Life Expectancy at Birth (avg years)</th>
<th>Fertility Rate (births per woman)</th>
<th>Infant Mortality (deaths per 1000 births)</th>
<th>Enrollment Rate, Tertiary (%)</th>
<th>Unemployment, Total (%)</th>
<th>Estimated Government Effectiveness (scale -2.5 to 2.5)</th>
<th>Internet Users (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>1560.67, 44.62, 2.44, 23.86, 60.07, 5.39, 71.3, 33.85, 1.41, 1.45, 4.5</td>
<td>403.43, 115.11, 0.26, 54.45, 77.16, 1.75, 15.54, 85.14, 0.72, 0.28, 54.66</td>
<td>8515.35, 15.86, 1.89, 73.71, 76.75, 2.83, 25.6, 31.46, 0.10, -0.54, -0.55, 15.23</td>
<td>19640.35, 200.35, 1.63, 29.87, 75.5, 2.12, 9.2, 14.37, 8.4, 1.29, 0.48, 83.79</td>
<td>12016.2, 14.88, 0.88, 92.64, 75.84, 2.2, 12.7, 74.83, 7.2, -0.49, -0.25, 55.8</td>
<td>8416.82, 104.08, 0.17, 64.16, 74.33, 1.74, 14.7, 48.94, 18.4, -0.62, -0.04, 39.16</td>
<td>44597.83, 2.91, 1.69, 83.34, 81.85, 1.87, 4.1, 83.24, 5.2, 2, 1.61, 82.35</td>
<td>43661.15, 102.22, 0.46, 67.88, 81.03, 1.42, 3.3, 71.43, 1.35, 1.66, 81</td>
<td>10125.23, 110.98, 1.35, 53.39, 70.55, 1.92, 38.5, 19.65, 5.2, -1.13, -0.79, 54.2</td>
<td>24590.49, 1701.01, 1.92, 88.76, 76.4, 2, 12, 8.2, 33.46, 1.1, 0.39, 0.65, 88</td>
<td>1883.05, 1174.33, 1.19, 28.89, 69.89, 2.24, 33.1, 13.15, 5, -0.87, -0.83, 6.3</td>
</tr>
</tbody>
</table>
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 runtime 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/
Python Refresher

```python
In [1]: import random

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

camel 0.5333896529549417
elephant 0.8289440919886492
crocodile 0.5635699354595317
```

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

```python
import pandas as pd

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

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

**CSV - comma-separated values**

<table>
<thead>
<tr>
<th>Country</th>
<th>1.92</th>
<th>88.76</th>
<th>76.4</th>
<th>2.12</th>
<th>8.2</th>
<th>33.46</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bahrain</td>
<td>24590.49</td>
<td>1701.01</td>
<td>1.92</td>
<td>88.76</td>
<td>76.4</td>
<td>2.12</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>1883.05</td>
<td>1174.33</td>
<td>1.19</td>
<td>28.89</td>
<td>69.89</td>
<td>2.24</td>
</tr>
<tr>
<td>Barbados</td>
<td>26487.77</td>
<td>655.36</td>
<td>0.5</td>
<td>44.91</td>
<td>74.97</td>
<td>1.84</td>
</tr>
<tr>
<td>Belgium</td>
<td>39751.48</td>
<td>364.85</td>
<td>0.85</td>
<td>97.51</td>
<td>80.49</td>
<td>1.84</td>
</tr>
</tbody>
</table>

**TSV - tab-separated values**

<table>
<thead>
<tr>
<th>Country</th>
<th>1.92</th>
<th>88.76</th>
<th>76.4</th>
<th>2.12</th>
<th>8.2</th>
<th>33.46</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bahrain</td>
<td>24590.49</td>
<td>1701.01</td>
<td>1.92</td>
<td>88.76</td>
<td>76.4</td>
<td>2.12</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>1883.05</td>
<td>1174.33</td>
<td>1.19</td>
<td>28.89</td>
<td>69.89</td>
<td>2.24</td>
</tr>
<tr>
<td>Barbados</td>
<td>26487.77</td>
<td>655.36</td>
<td>0.5</td>
<td>44.91</td>
<td>74.97</td>
<td>1.84</td>
</tr>
<tr>
<td>Belgium</td>
<td>39751.48</td>
<td>364.85</td>
<td>0.85</td>
<td>97.51</td>
<td>80.49</td>
<td>1.84</td>
</tr>
</tbody>
</table>
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

Average GDP by unemployment

- Low unemployment
- High unemployment
Calculating Statistics over the Data

Average GDP by unemployment

GDP

Low unempl.  High unempl.
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

![Diagram showing the relationship between GDP per capita (PPP USD) and unemployment rate (total %)]
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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>Friday, 8 November</td>
</tr>
<tr>
<td>2. Linear Regression</td>
<td>Monday, 11 November</td>
</tr>
<tr>
<td>3. <strong>Practical1</strong>: Linear Regression</td>
<td>Tuesday, 12 November</td>
</tr>
<tr>
<td>4. Classification</td>
<td>Wednesday, 13 November</td>
</tr>
<tr>
<td>5. <strong>Practical2</strong>: Classification</td>
<td>Thursday, 14 November</td>
</tr>
<tr>
<td>6. Ensemble-based models</td>
<td>Monday, 18 November</td>
</tr>
<tr>
<td>7. <strong>Practical3</strong>: Ensemble models</td>
<td>Tuesday, 19 November</td>
</tr>
<tr>
<td>8. Visualization, part I</td>
<td>Wednesday, 20 November</td>
</tr>
</tbody>
</table>
# Course Syllabus

<table>
<thead>
<tr>
<th>Course Title</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Visualization, part II</td>
<td>Friday, 22 November</td>
</tr>
<tr>
<td>10. Deep Learning basics</td>
<td>Monday, 25 November</td>
</tr>
<tr>
<td><strong>11. Practical4: Visualization</strong></td>
<td>Tuesday, 26 November</td>
</tr>
<tr>
<td>12. Deep Learning with TensorFlow</td>
<td>Wednesday, 27 November</td>
</tr>
<tr>
<td><strong>13. Practical5: Deep Learning I</strong></td>
<td>Thursday, 28 November</td>
</tr>
<tr>
<td>14. Deep Learning architectures</td>
<td>Friday, 29 November</td>
</tr>
<tr>
<td>15. Challenges in Data Science</td>
<td>Monday, 2 December</td>
</tr>
<tr>
<td><strong>16. Practical6: Deep Learning II</strong></td>
<td>Tuesday, 3 December</td>
</tr>
</tbody>
</table>
Lecturers

Ekaterina Kochmar
ek358

Guy Emerson
gete2

Damon Wischik
dwj1005
Course Pages

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


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