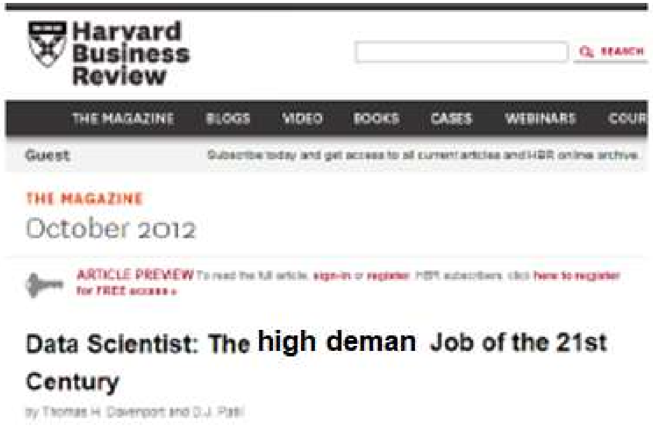
Big Data Analytics

Dr. Muhammad Affan Alim

Data Scientists are in high demand



Harvard Business Review

SEARCH

THE MAGAZINE BLOGS VIDEO BOOKS CASES WEBINARS COUR

Guest — Subscribe today and get access to all current articles and HBR online archive.

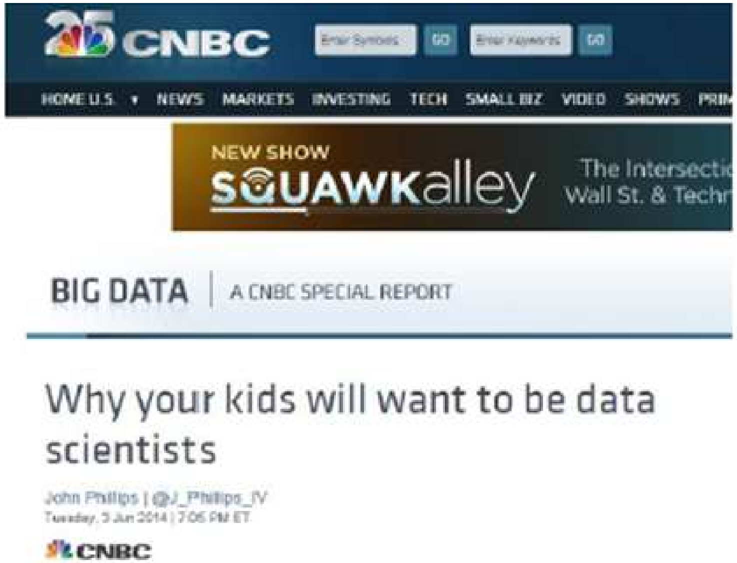
THE MAGAZINE

October 2012

ARTICLE PREVIEW To read the full article, sign-in or register. HBR subscribers, click here to register for FREE access »

Data Scientist: The high deman Job of the 21st Century

by Thomas H. Davenport and D.J. Patil



CNBC

Enter Symbols GO Enter Keywords GO

HOME U.S. ▾ NEWS MARKETS INVESTING TECH SMALL BIZ VIDEO SHOWS PRI

NEW SHOW SQUAWKalley The Intersecti Wall St. & Techn

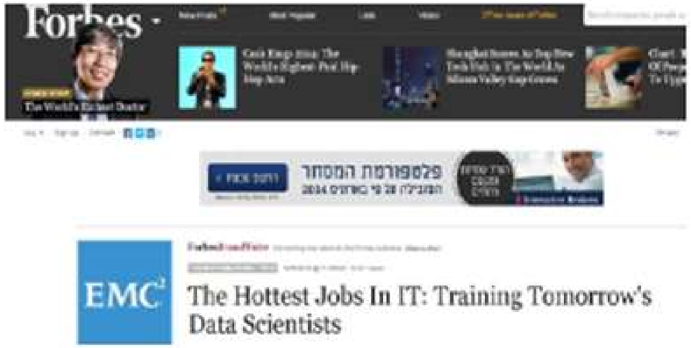
BIG DATA | A CNBC SPECIAL REPORT

Why your kids will want to be data scientists

John Phillips | @J_Phillips_IV

Tuesday, 3 Jun 2014 | 7:06 PM ET

CNBC



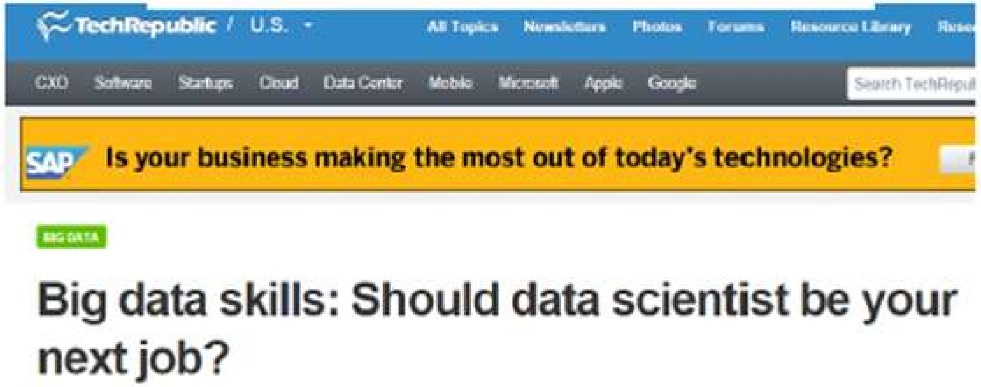
Forbes

The World's Richest Doctor

פלטפורמות המסחר

EMC

The Hottest Jobs In IT: Training Tomorrow's Data Scientists



TechRepublic / U.S. ▾ All Topics Newsletters Photos Forums Resource Library Rese

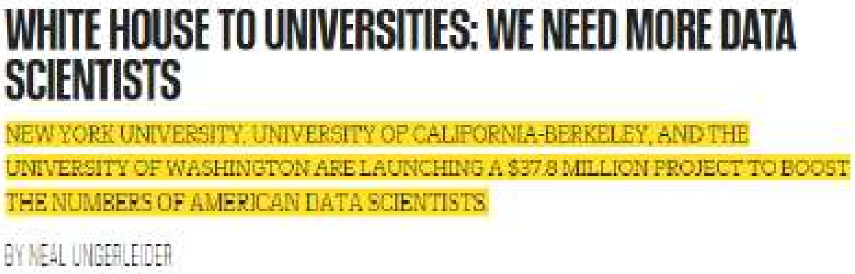
CXO Software Startups Cloud Data Center Mobile Microsoft Apple Google Search TechRepubl

SAP Is your business making the most out of today's technologies?

BIG DATA

Big data skills: Should data scientist be your next job?

Also in academia

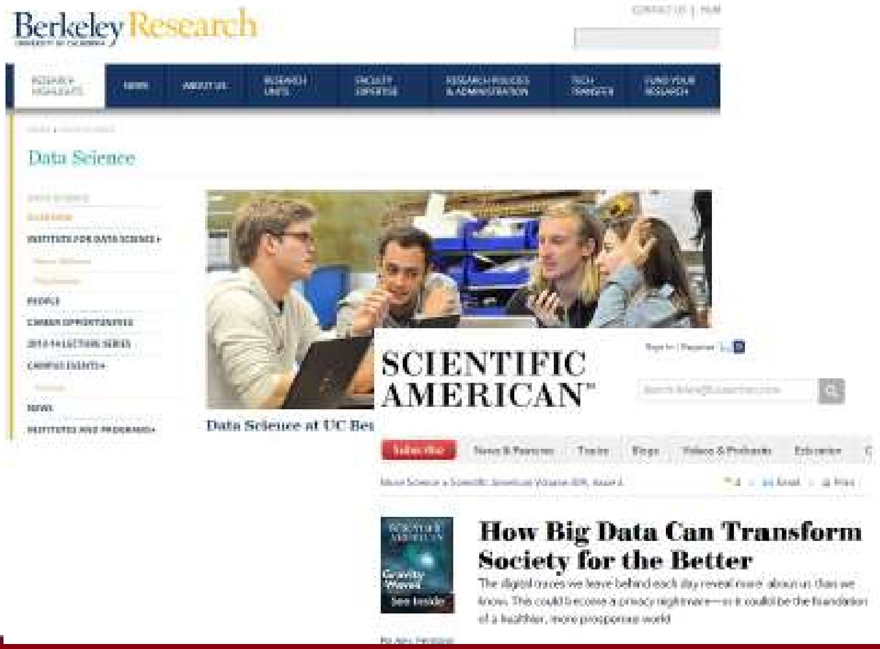


WHITE HOUSE TO UNIVERSITIES: WE NEED MORE DATA SCIENTISTS

NEW YORK UNIVERSITY, UNIVERSITY OF CALIFORNIA-BERKELEY, AND THE UNIVERSITY OF WASHINGTON ARE LAUNCHING A $37.8 MILLION PROJECT TO BOOST THE NUMBERS OF AMERICAN DATA SCIENTISTS

BY NEAL UNGERLEIDER

It's official: America needs more data scientists. This week, a $37.8 million project



Berkeley Research

CONTACT US | HUM

RESEARCH HIGHLIGHTS | NEWS | ABOUT US | RESEARCH UNITS | FACULTY EXPERTISE | RESEARCH POLICIES & ADMINISTRATION | TECH TRANSFER | FUND YOUR RESEARCH

Data Science

DATA SCIENCE
OVERVIEW
INSTITUTE FOR DATA SCIENCE ›
PEOPLE
CAREER OPPORTUNITIES
2013-14 LECTURE SERIES
CAMPUS EVENTS ›
NEWS
INSTITUTES AND PROGRAMS ›

Data Science at UC Ber

SCIENTIFIC AMERICAN™

Sign In | Register

Subscribe | News & Features | Topics | Blogs | Videos & Podcasts | Education

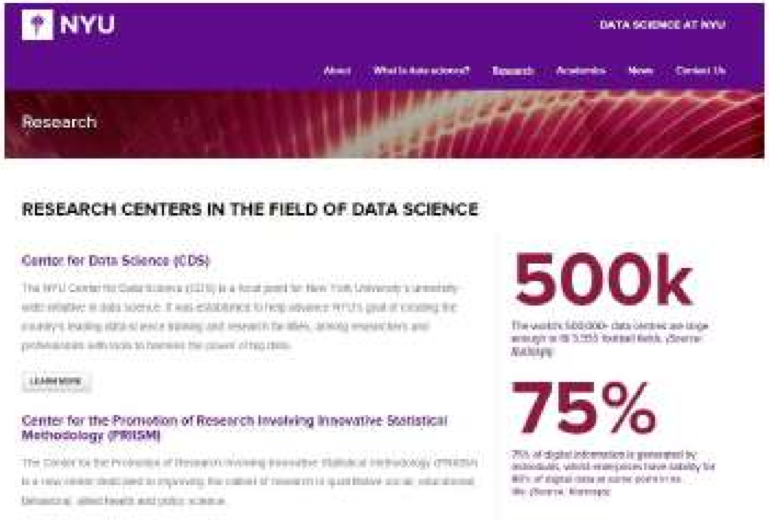
More Science » Scientific American Volume 309, Issue 4

Email | Print

How Big Data Can Transform Society for the Better

The digital traces we leave behind each day reveal more about us than we know. This could become a privacy nightmare—or it could be the foundation of a healthier, more prosperous world

By Alex Pentland



NYU

DATA SCIENCE AT NYU

About | What is data science? | Research | Academics | News | Contact Us

Research

RESEARCH CENTERS IN THE FIELD OF DATA SCIENCE

Center for Data Science (CDS)

The NYU Center for Data Science (CDS) is a focal point for New York University's university-wide initiative in data science. It was established to help advance NYU's goal of creating the country's leading data science training and research facilities, arming researchers and professionals with tools to harness the power of big data.

LEARN MORE

Center for the Promotion of Research Involving Innovative Statistical Methodology (PRISM)

The Center for the Promotion of Research Involving Innovative Statistical Methodology (PRISM) is a new center dedicated to improving the caliber of research in quantitative social, educational, behavioral, allied health and policy science.

500k

The world's 500,000+ data centres are large enough to fill 5,955 football fields. (Source: [illegible])

75%

75% of digital information is generated by individuals, whilst enterprises have liability for 80% of digital data at some point in its life. (Source: [illegible])

UNIVERSITY of WASHINGTON

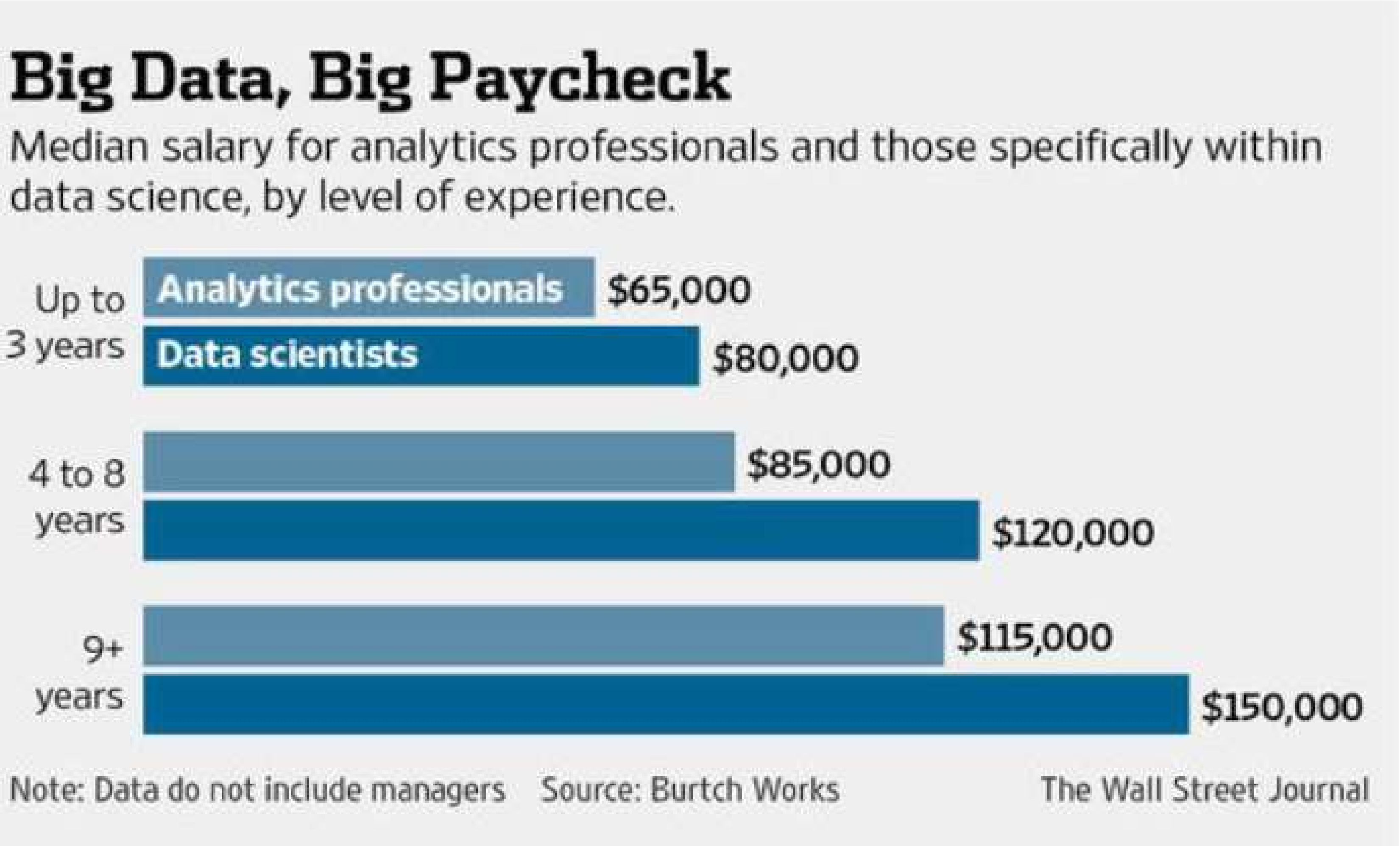
eScience Institute

Supporting Data-Driven Discovery in All Fields

WHO WE ARE

New Ph.D. Tracks in "Big Data"

Pays Well

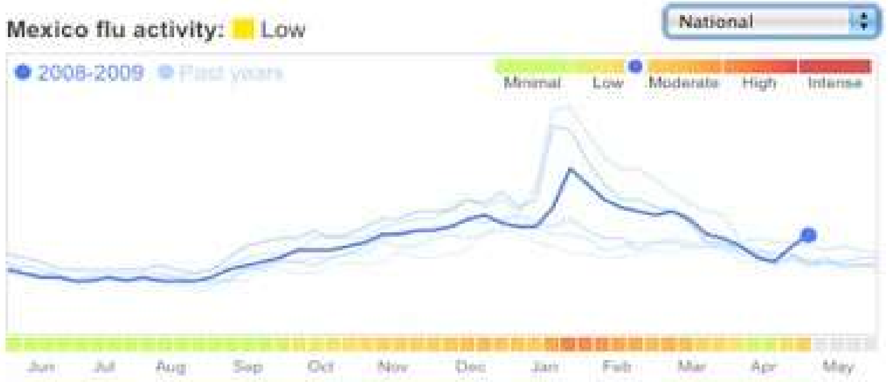
- 

Big Data, Big Paycheck

Median salary for analytics professionals and those specifically within data science, by level of experience.

Note: Data do not include managers Source: Burtch Works The Wall Street Journal

Data Science: Why all the Excitement?



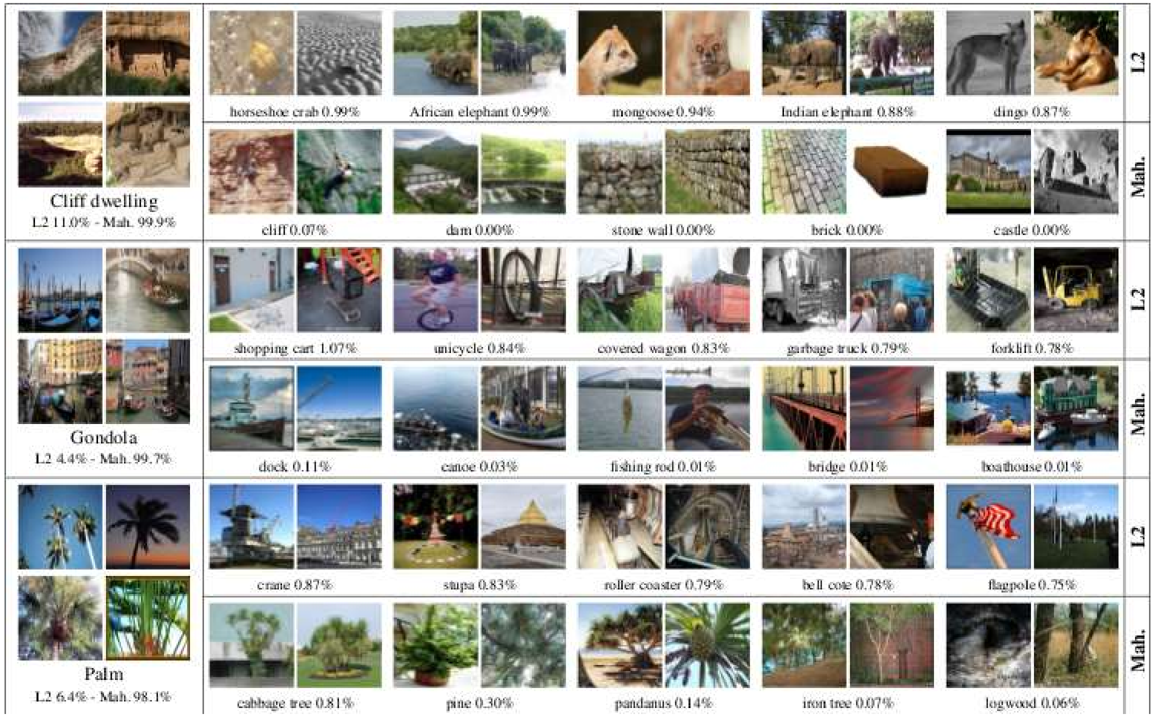
e.g.,
Google Flu Trends:

Detecting outbreaks
two weeks ahead
of CDC data

New models are estimating
which cities are most at risk
for spread of the Ebola virus.

The unreasonable effectiveness of Deep Learning (CNNs)

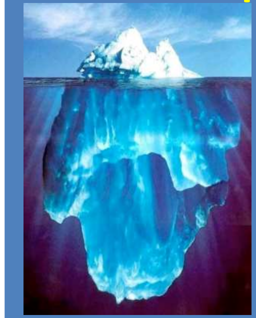
- 2012 Image-net challenge: Classify 1 million images into 1000 classes.



Where does data come from?

“Big Data” Sources

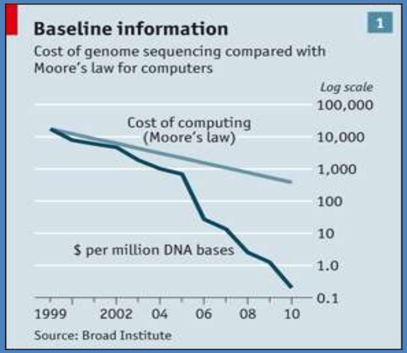
It’s All Happening On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause,…
Server request
Transaction
Network message
Fault
…

Internet of Things / M2M

Health/Scientific Computing



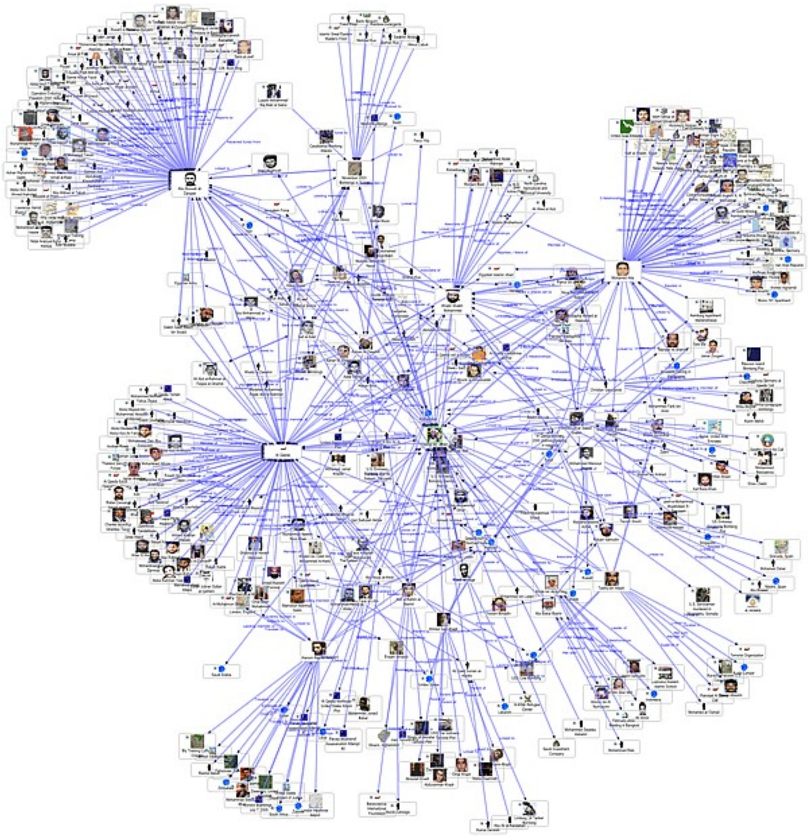
Source: Broad Institute

Graph Data

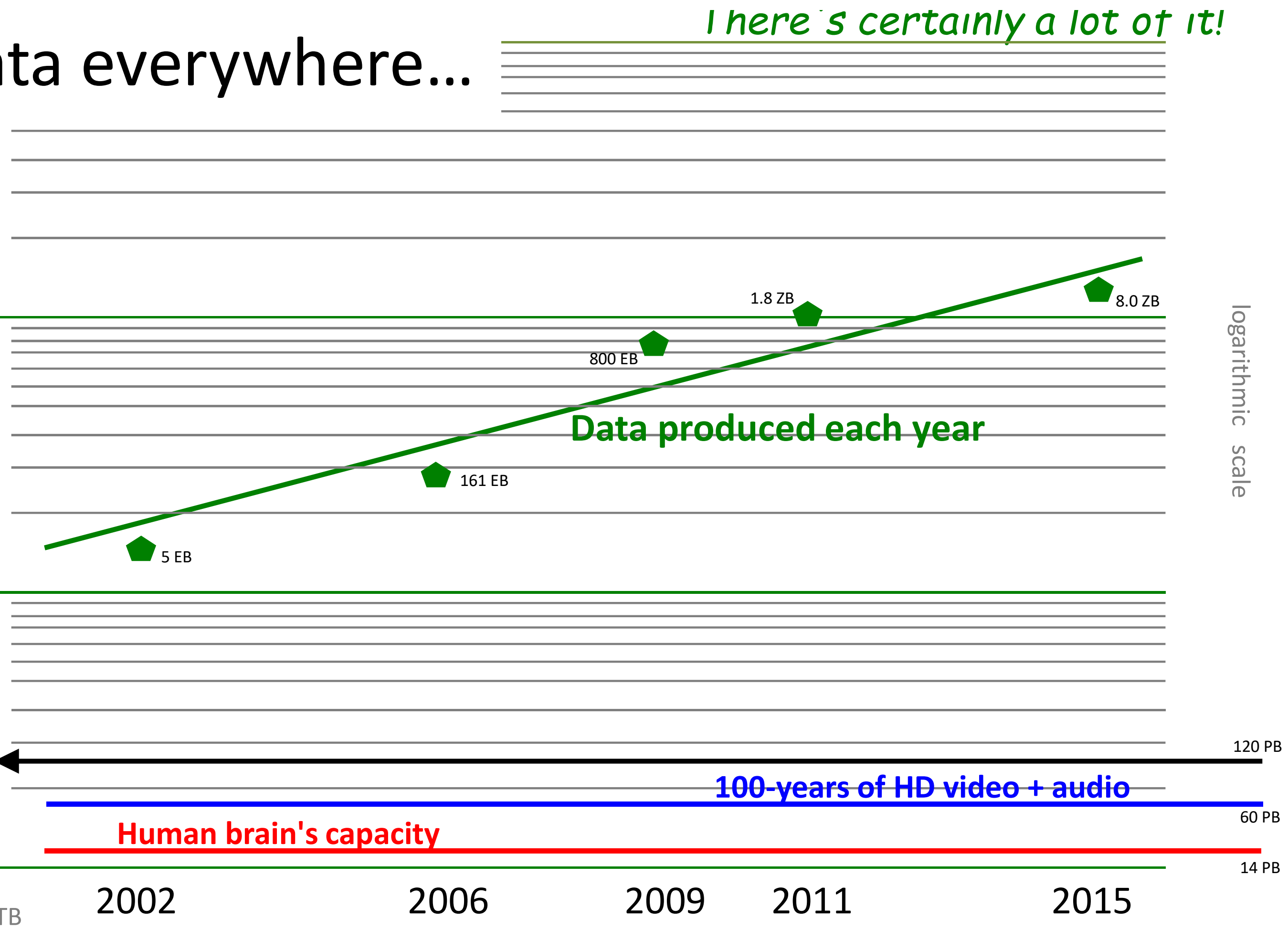
Lots of interesting data has a graph structure:

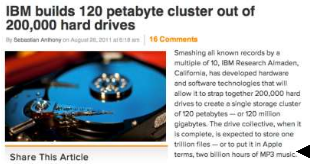
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

Some of these graphs can get quite large (e.g., Facebook* user graph)



Data, data everywhere…





IBM builds 120 petabyte cluster out of 200,000 hard drives

By Sebastian Anthony on August 26, 2011 at 8:18 am | 16 Comments

Smashing all known records by a multiple of 10, IBM Research Almaden, California, has developed hardware and software technologies that will allow it to strap together 200,000 hard drives to create a single storage cluster of 120 petabytes — or 120 million gigabytes. The drive collective, when it is complete, is expected to store one trillion files — or to put it in Apple terms, two billion hours of MP3 music.

Share This Article

1 Petabyte == 1000 TB
1 TB = 1000 GB

References

(2015) 8 ZB: http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

(2011) 1.8 ZB: http://www.emc.com/leadership/programs/digital-universe.htm

(2009) 800 EB: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf

(2006) 161 EB: http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf

(2002) 5 EB: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm

(life in video) 60 PB: in 4320p resolution, extrapolated from 16MB for 1:21 of 640x480 video (w/sound) – almost certainly a gross overestimate, as sleep can be compressed significantly!

(brain) 14 PB: http://www.quora.com/Neuroscience-1/How-much-data-can-the-human-brain-store

Data Science – A Definition

- **Data Science** is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

Goal of Data Science

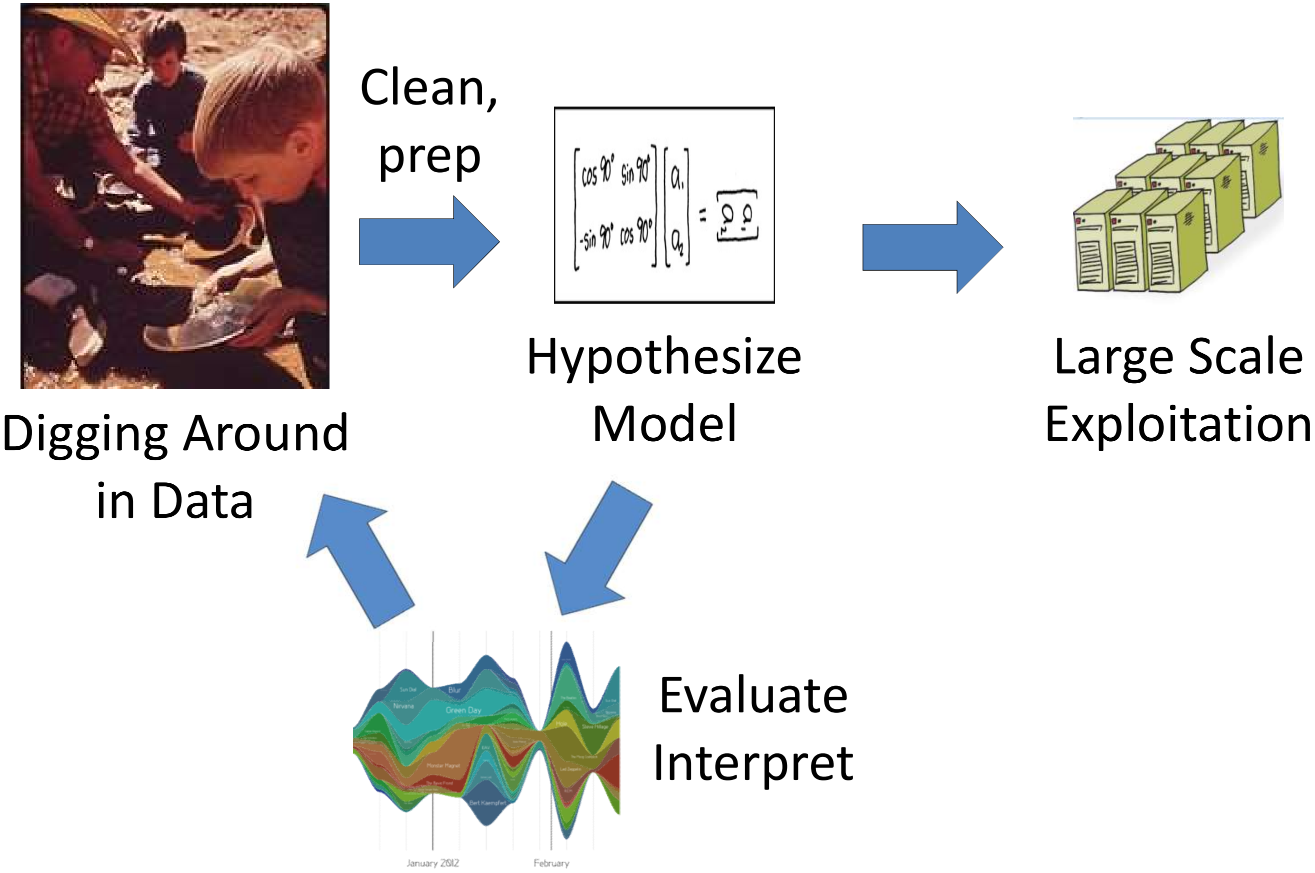
Turn data into data products

How to use data?

Data => exploratory analysis => knowledge models => product / decision marking

Data => predictive models => evaluate / interpret => product / decision making

Data Scientist’s Practice



Digging Around in Data

Clean, prep

Hypothesize Model

Large Scale Exploitation

Evaluate Interpret

Example data science applications

Marketing: predict the characteristics of high life time value (LTV) customers, which can be used to support customer segmentation, identify upsell opportunities, and support other marking initiatives

Logistics: forecast how many of which things you need and where will we need them, which enables learn inventory and prevents out of stock situations

Healthcare: analyze survival statistics for different patient attributes (age, blood type, gender, etc.) and treatments; predict risk of re-admittance based on patient attributes, medical history, etc.

Example data science applications cont…

Healthcare: analyze survival statistics for different patient attributes (age, blood type, gender, etc.) and treatments; predict risk of re-admittance based on patient attributes, medical history, etc.

Some more examples

Transaction Databases → Recommender systems (NetFlix), Fraud Detection (Security and Privacy)

Wireless Sensor Data → Smart Home, Real-time Monitoring, Internet of Things

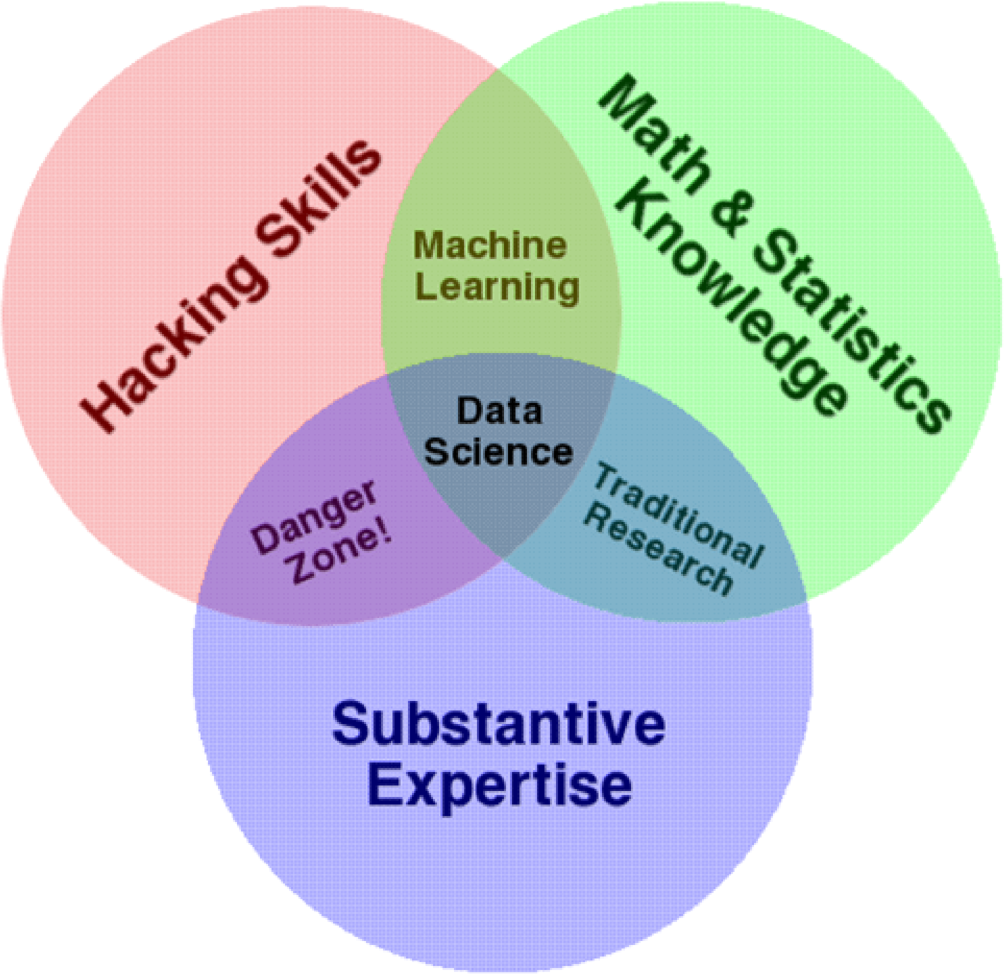
Text Data, Social Media Data → Product Review and Consumer Satisfaction (Facebook, Twitter, LinkedIn), E-discovery

Software Log Data → Automatic Trouble Shooting (Splunk)

Genotype and Phenotype Data → Epic, 23andme, Patient-Centered Care, Personalized Medicine

Data Science – One Definition

-



What’s Hard about Data Science

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)

About the course

- A mixture of theory and practice
- Introductory, broad overview of subjects
- Focus on practical aspects, but not on ever-changing technology and tools
- Seminar style - I am here to learn as well as to teach
- Language choice: python

  Relatively easy to learn (for computer scientist) compared to R (more popular among statisticians)

  Open source means easy access (as opposed to SAS or MATLAB)

  [Which one is more frequently used in data science]()?

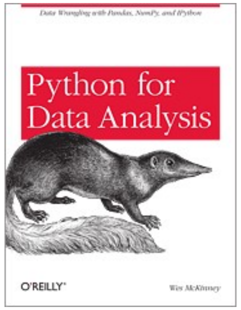
Textbook

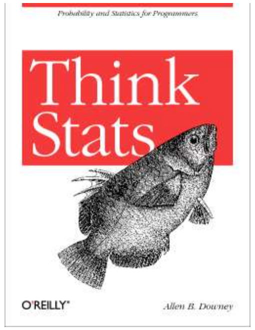
- **Required**: Python Data Science Handbook (**PDSH**)
- by Jake VanderPlas
- **Optional:**

Data Science from Scratch (**DSS**) by Joel Grus

Python for Data Analysis (**PDA**) by Wes McKinney

Free e-book: Think Stats (**TS**) by Allen B. Downey.





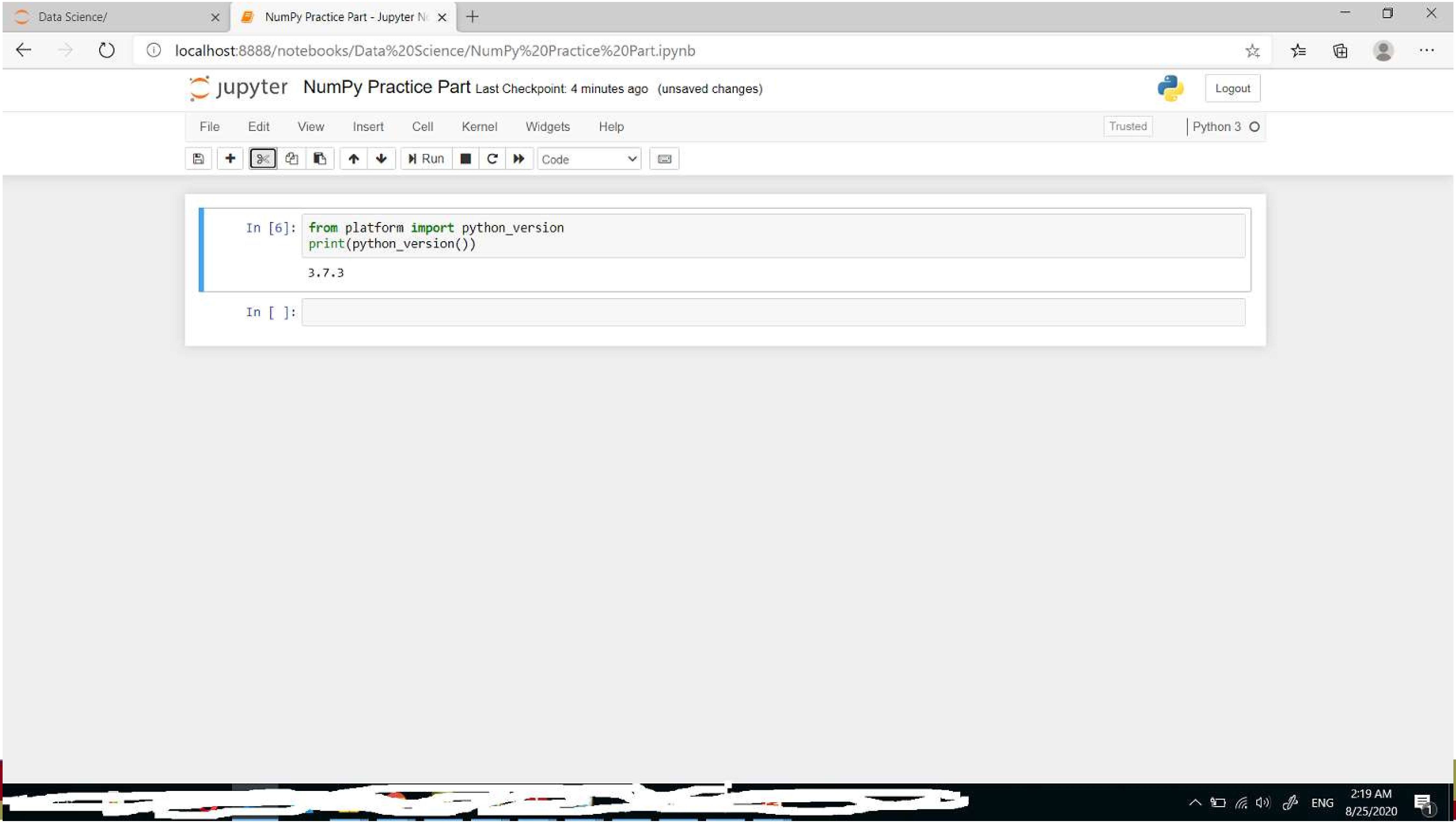
O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
Jake VanderPlas

Brief introduction of Python

- Invented in the Netherlands, early 90s by Guido van Rossum
- Open sourced from the beginning
- Considered a scripting language, but is much more
  - No compilation needed
  - Scripts are evaluated by the interpreter, line by line
  - Functions need to be defined before they are called

Installing the Anaconda

- Watch the video
- https://www.youtube.com/watch?v=G3Lt1JWBvL8


Data Science/
NumPy Practice Part - Jupyter N
localhost:8888/notebooks/Data%20Science/NumPy%20Practice%20Part.ipynb
jupyter
NumPy Practice Part Last Checkpoint: 4 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
In [6]:
In [ ]:
ENG
2:19 AM
8/25/2020

[redacted]

```
from platform import python_version
print(python_version())
```

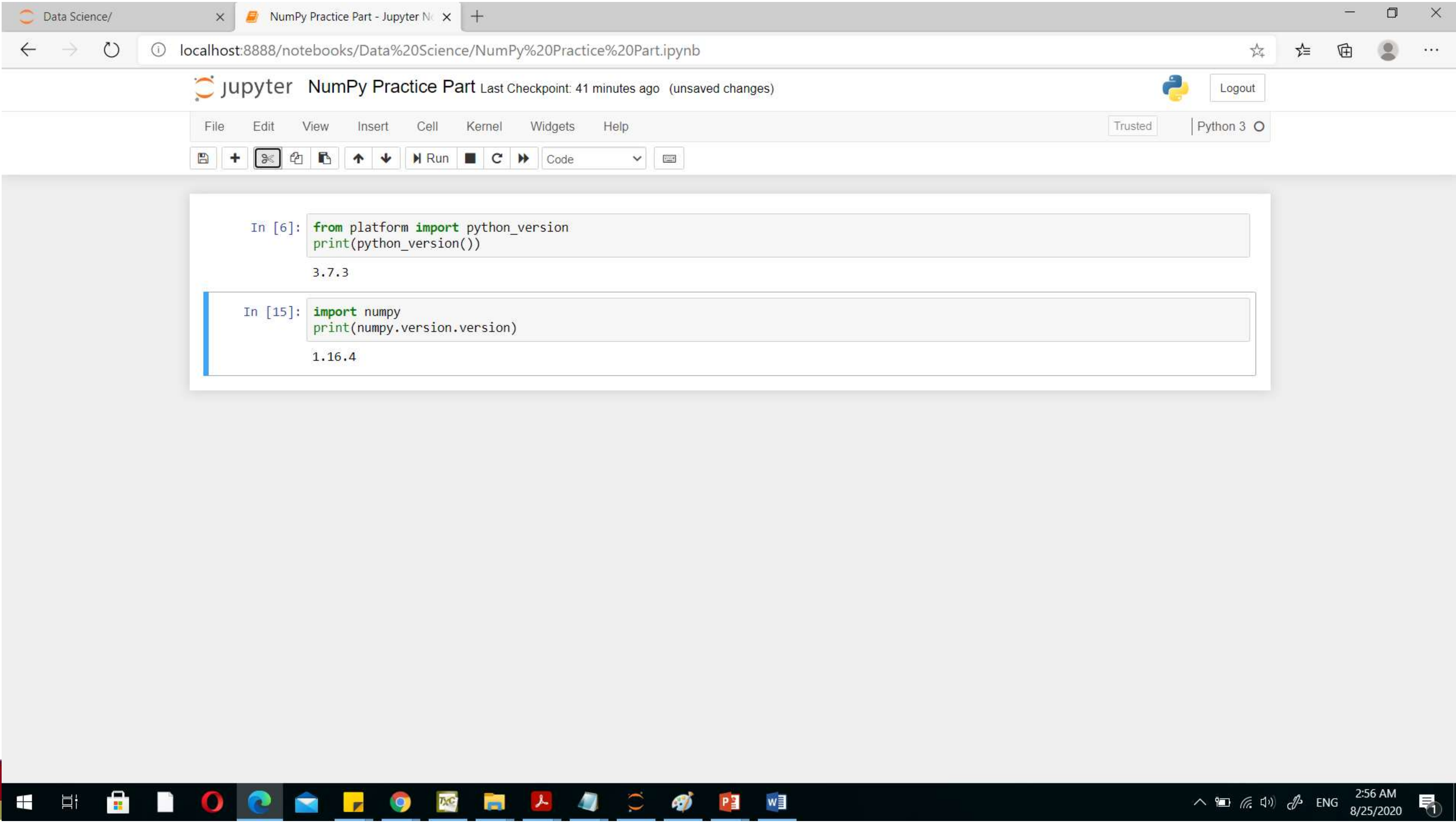
```
3.7.3
```

Introduction to NumPy

- NumPy (short for Numerical Python) provides an efficient interface to store and operate on dense data buffers.

- NumPy arrays from the core of nearly the entire ecosystem of data science tools in Python

- If you followed the installation the Anaconda stack, you already have NumPy


Data Science/
NumPy Practice Part - Jupyter N
localhost:8888/notebooks/Data%20Science/NumPy%20Practice%20Part.ipynb
jupyter
NumPy Practice Part Last Checkpoint: 41 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
2:56 AM
8/25/2020
ENG

```
In [6]: from platform import python_version
        print(python_version())

3.7.3
```

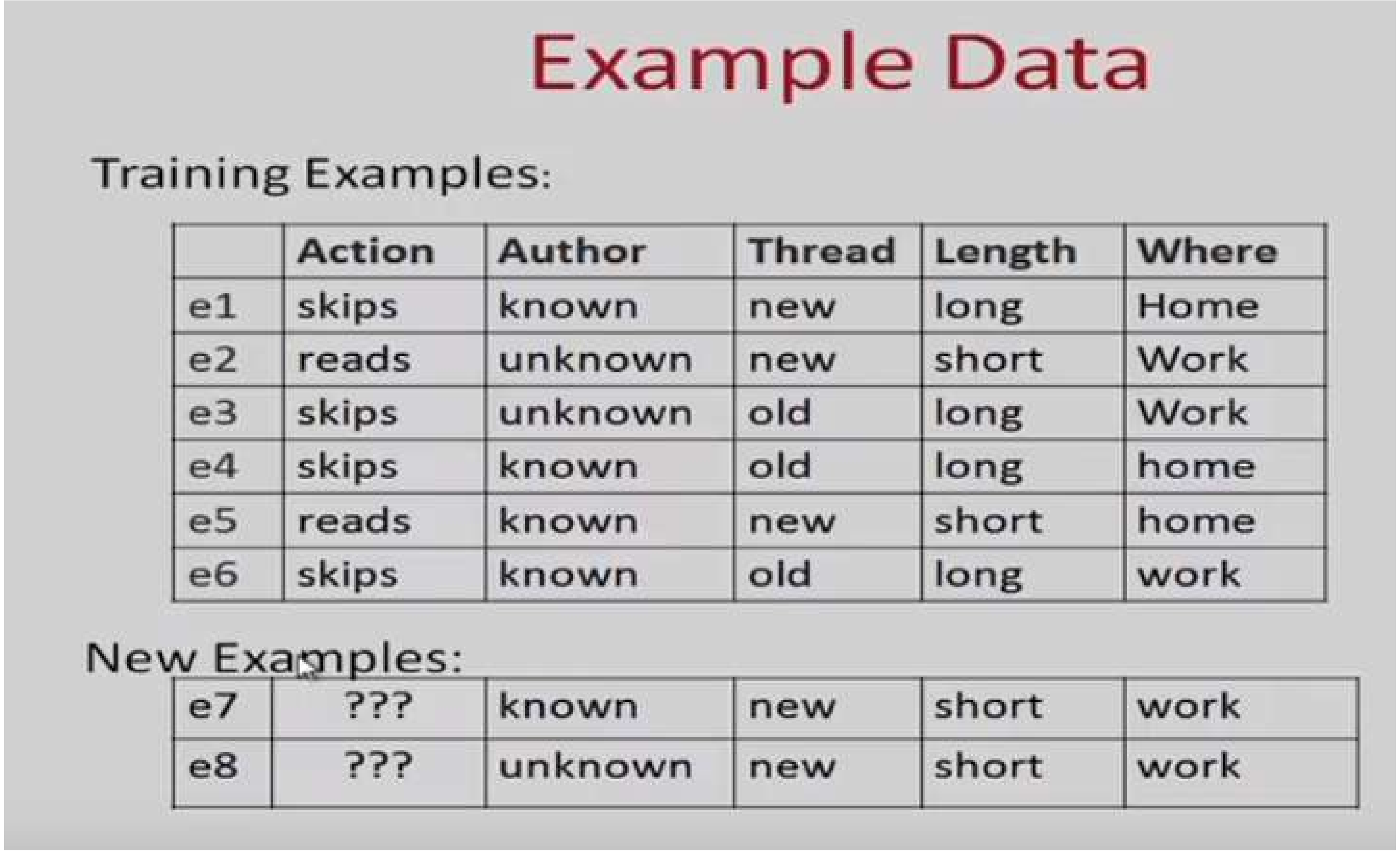
```
In [15]: import numpy
         print(numpy.version.version)

1.16.4
```

NumPy cont…

- By convention, you’ll find that most people in the SciPy/PyData world will import NumPy using np as an alias:

- In[2]: import numpy as np

Features

- 

Example Data

Training Examples:

	Action	Author	Thread	Length	Where
e1	skips	known	new	long	Home
e2	reads	unknown	new	short	Work
e3	skips	unknown	old	long	Work
e4	skips	known	old	long	home
e5	reads	known	new	short	home
e6	skips	known	old	long	work

New Examples:

e7	???	known	new	short	work
e8	???	unknown	new	short	work

Why is NumPy Faster Than Lists?

- NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently.

- This behavior is called locality of reference in computer science.

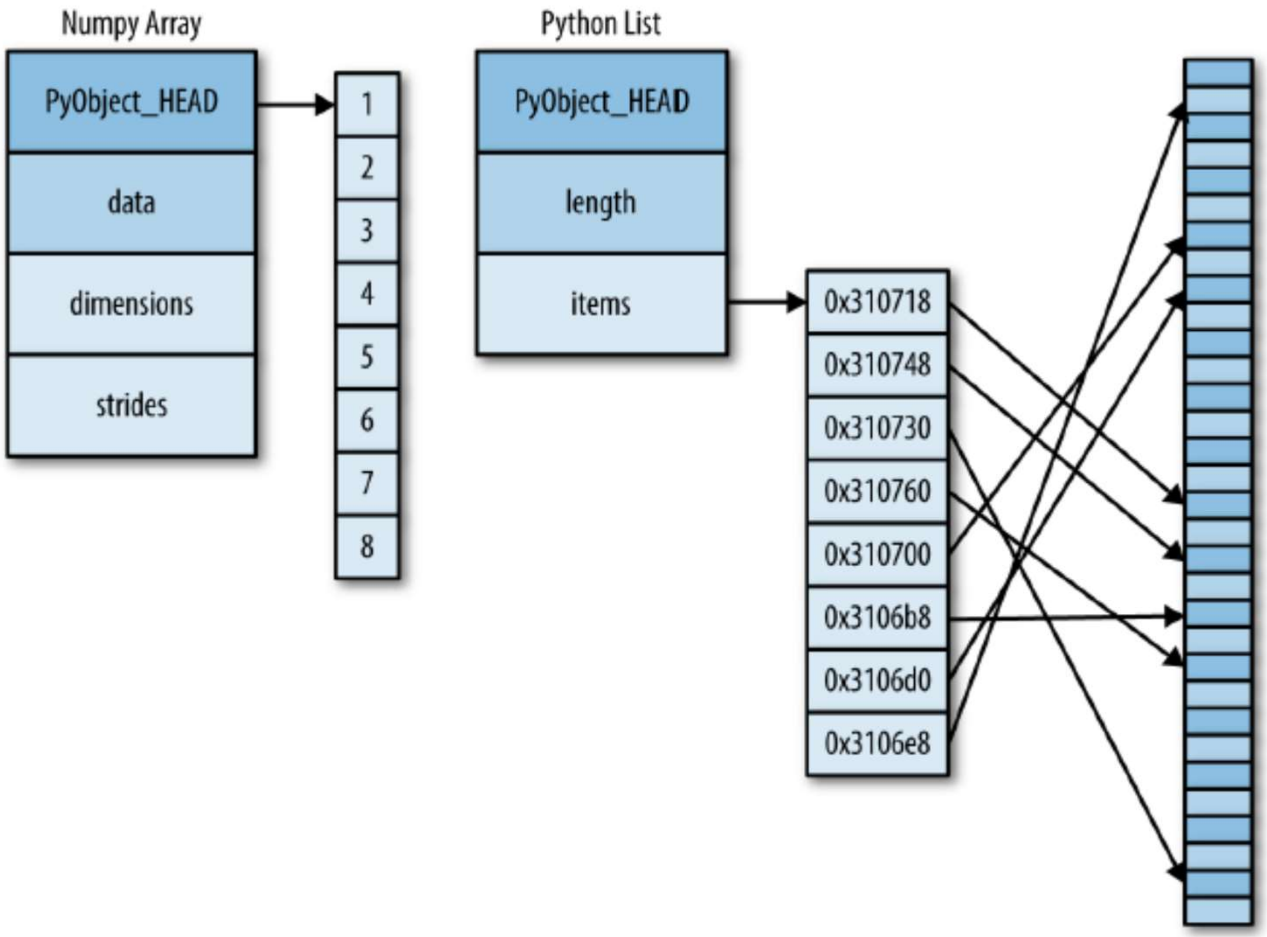
- This is the main reason why NumPy is faster than lists. Also it is optimized to work with latest CPU architectures.

c

- The Python list, on the other hand, contains a pointer to a block of pointers, each of which in turn points to a full Python object like the Python integer.
- Fixed-type NumPy-style arrays lack this flexibility, but are much more efficient for storing and manipulating data

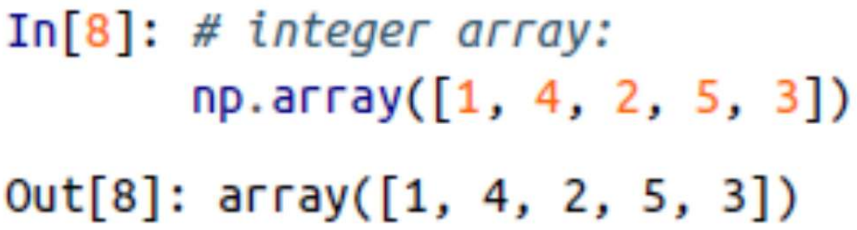
Difference between NumPy array and Python List

-



Creating arrays using NumPy

- First, we can use np.array to create arrays from Python lists:

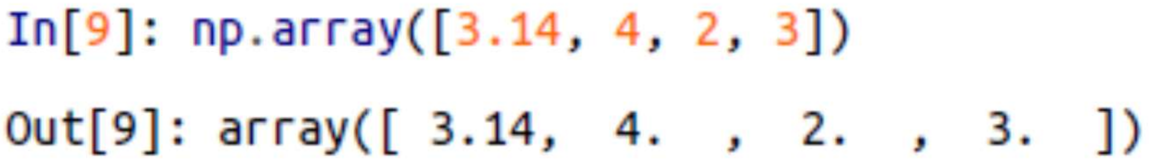


```
In[8]: # integer array:
       np.array([1, 4, 2, 5, 3])

Out[8]: array([1, 4, 2, 5, 3])
```

- Remember that unlike Python lists, NumPy is constrained to arrays that all contain the same type.

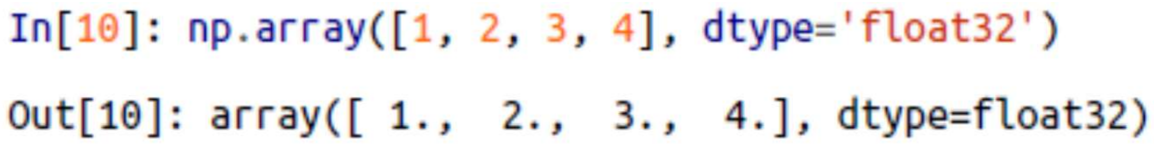
Creating arrays using NumPy cont…

- If types do not match, NumPy will upcast if possible (here, integers are upcast to floating point):



```
In[9]: np.array([3.14, 4, 2, 3])

Out[9]: array([ 3.14,  4.  ,  2.  ,  3.  ])
```

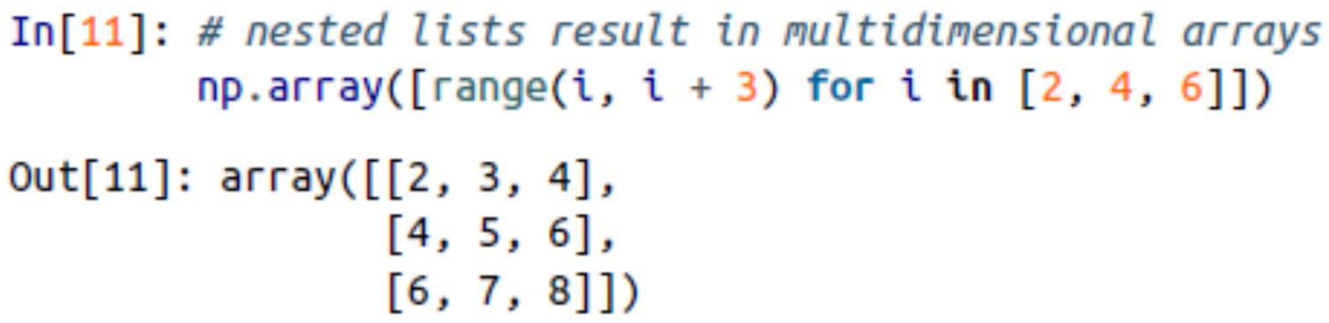
- If we want to explicitly set the data type of the resulting array, we can use the dtype keyword:



```
In[10]: np.array([1, 2, 3, 4], dtype='float32')

Out[10]: array([ 1.,  2.,  3.,  4.], dtype=float32)
```

Creating arrays using NumPy cont…

- NumPy arrays can explicitly be multidimensional; here’s one way of initializing a multidimensional array using a list of lists:



```
In[11]: # nested lists result in multidimensional arrays
        np.array([range(i, i + 3) for i in [2, 4, 6]])

Out[11]: array([[2, 3, 4],
                [4, 5, 6],
                [6, 7, 8]])
```

Data Science

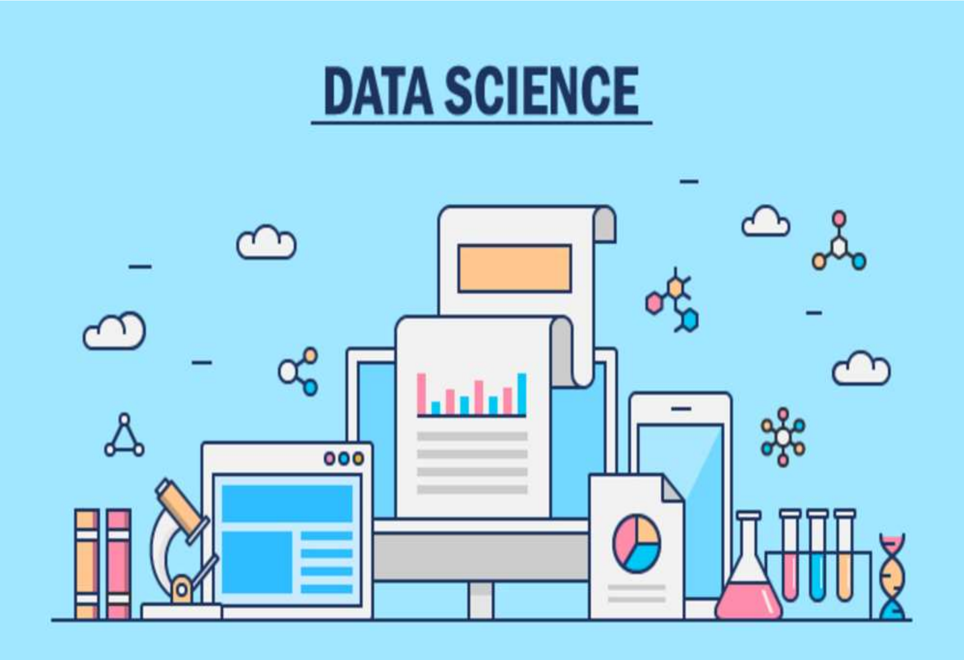
Revision

Data Science – A Definition

- **Data Science** is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

Data Science - Definition

- Data Science is an interdisciplinary field about processes to extract the knowledge or insights and predict from data in various forms, either structured or unstructured



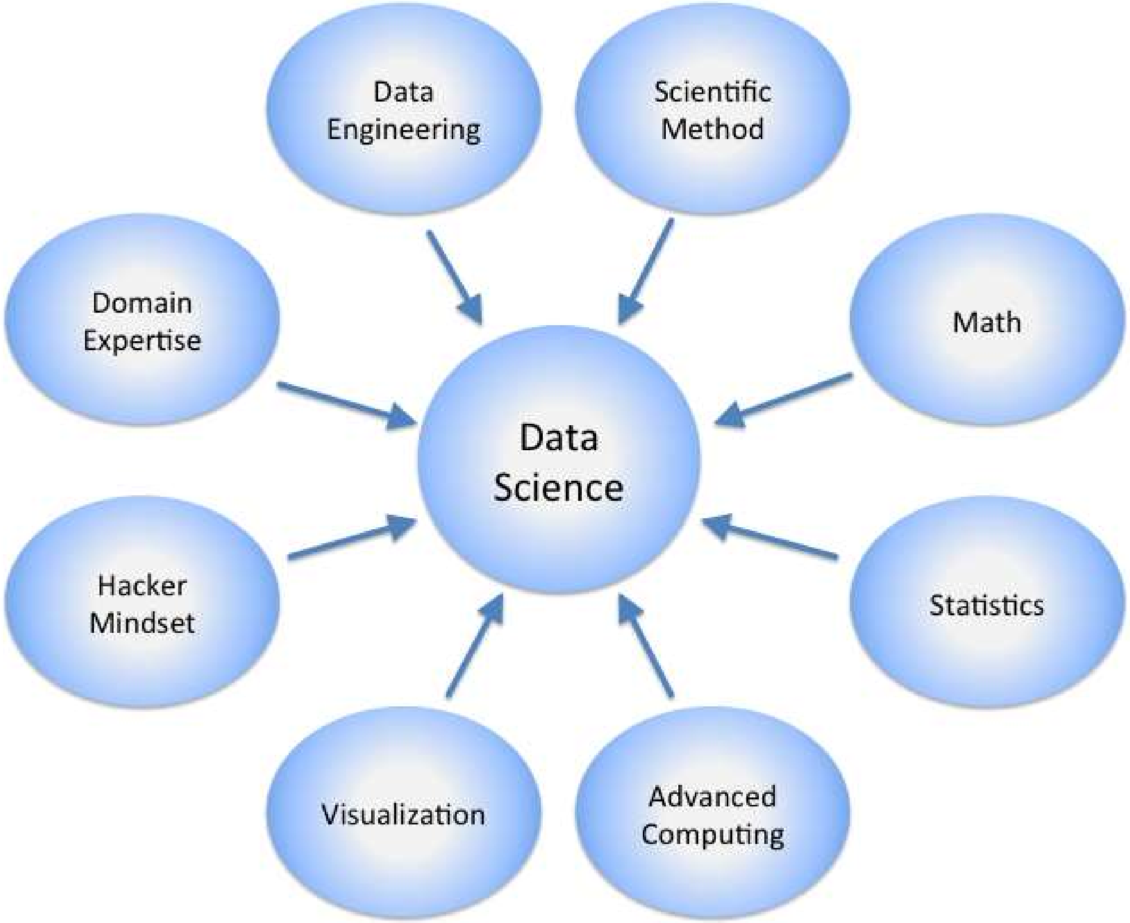
Data Analyst

- Data analytics is the science of examining raw data with the purpose of drawing conclusions about that information.

- Data analytics is used in many industries to allow companies and organization to make better decisions and in the sciences to verify or disprove existing models or theories

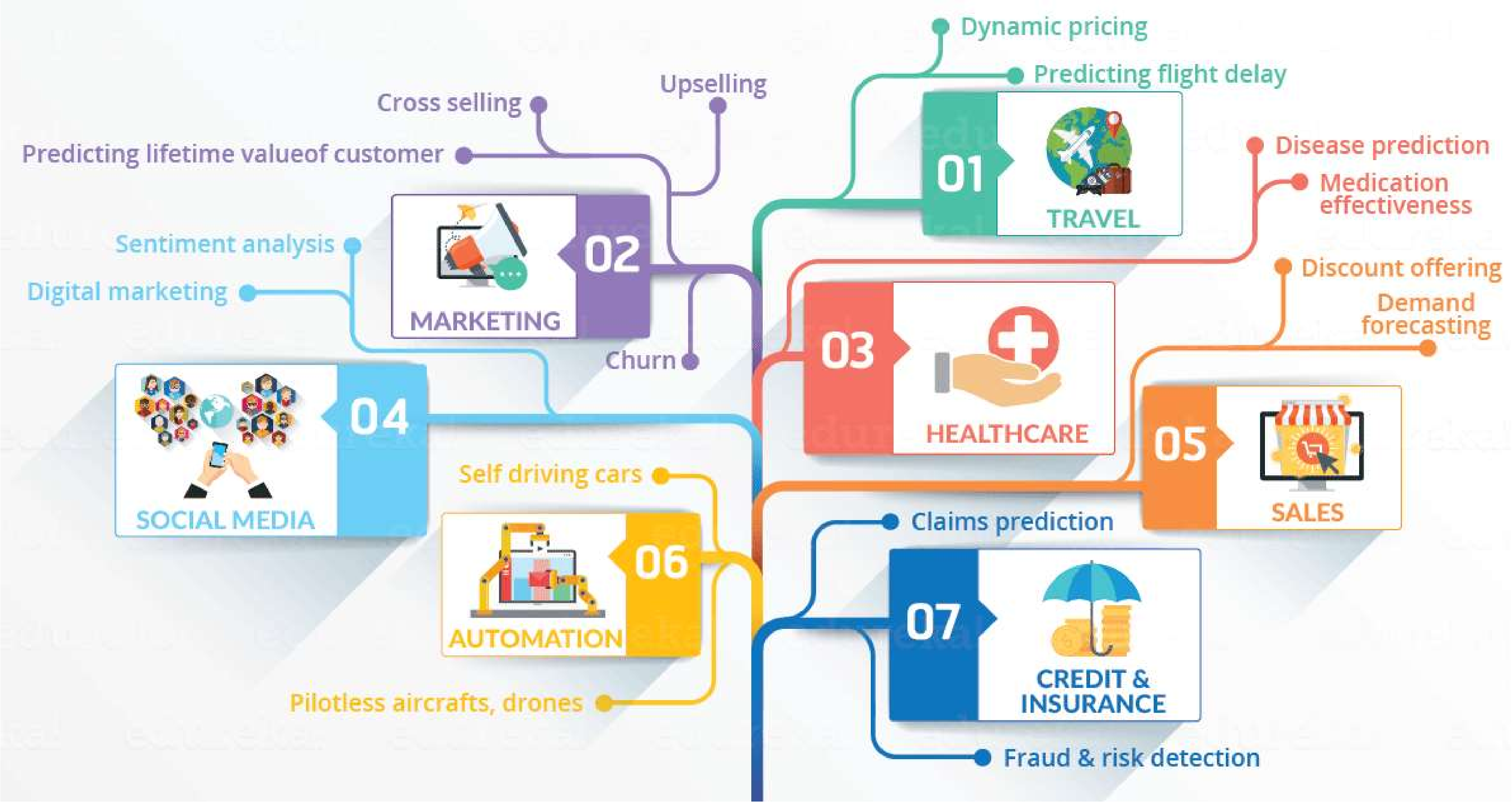
What to know of a data scientist

-



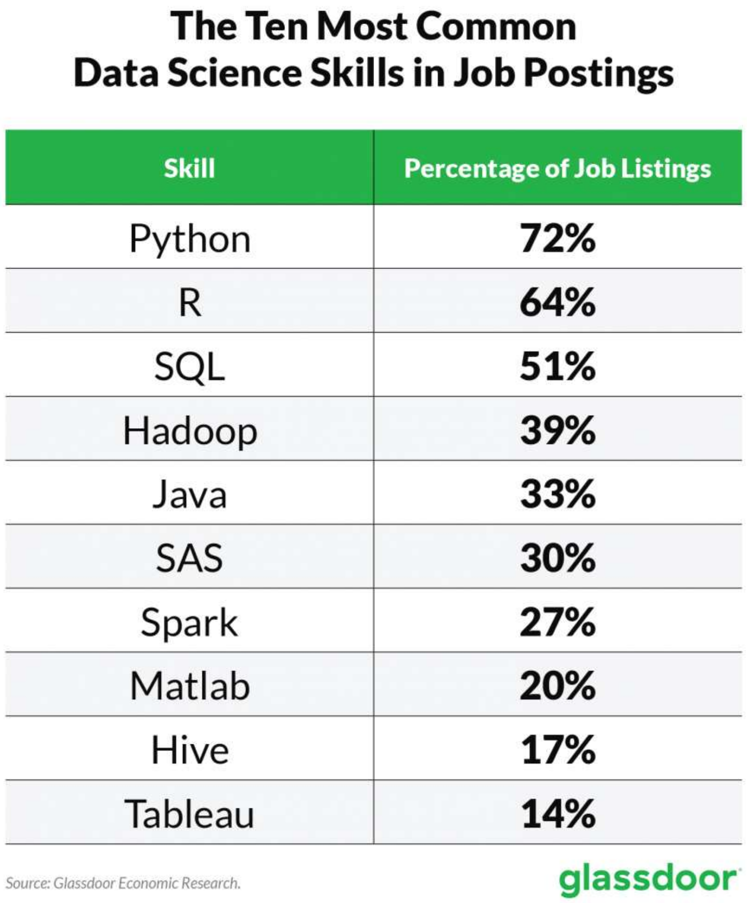
Targeted field of Data Science

-



Popular tools for Data Science

-



The Ten Most Common Data Science Skills in Job Postings

Skill	Percentage of Job Listings
Python	**72%**
R	**64%**
SQL	**51%**
Hadoop	**39%**
Java	**33%**
SAS	**30%**
Spark	**27%**
Matlab	**20%**
Hive	**17%**
Tableau	**14%**

Source: Glassdoor Economic Research.

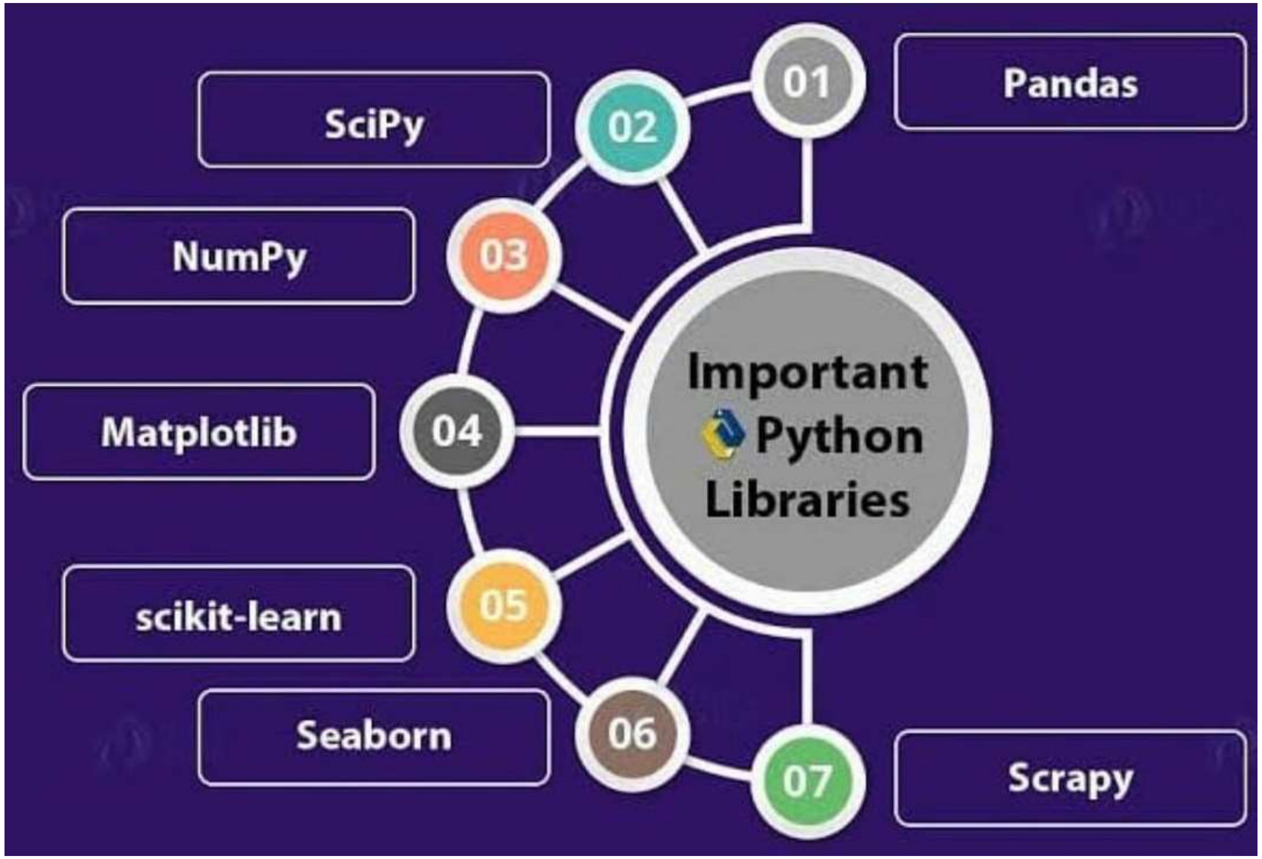
glassdoor

Discussion of our courses

- Python for Data Science
- Feature Engineering
- Machine Learning towards Data Science
- No SQL for BIG Data
- Real Time Application for Data Science

Python for Data Science

- As requirement of job market, python is the top most popular language
- Due to this reason we have selected python in our course



Feature Engineering

- Feature engineering is the process of using **domain knowledge** of the data to create features that make machine learning algorithms work

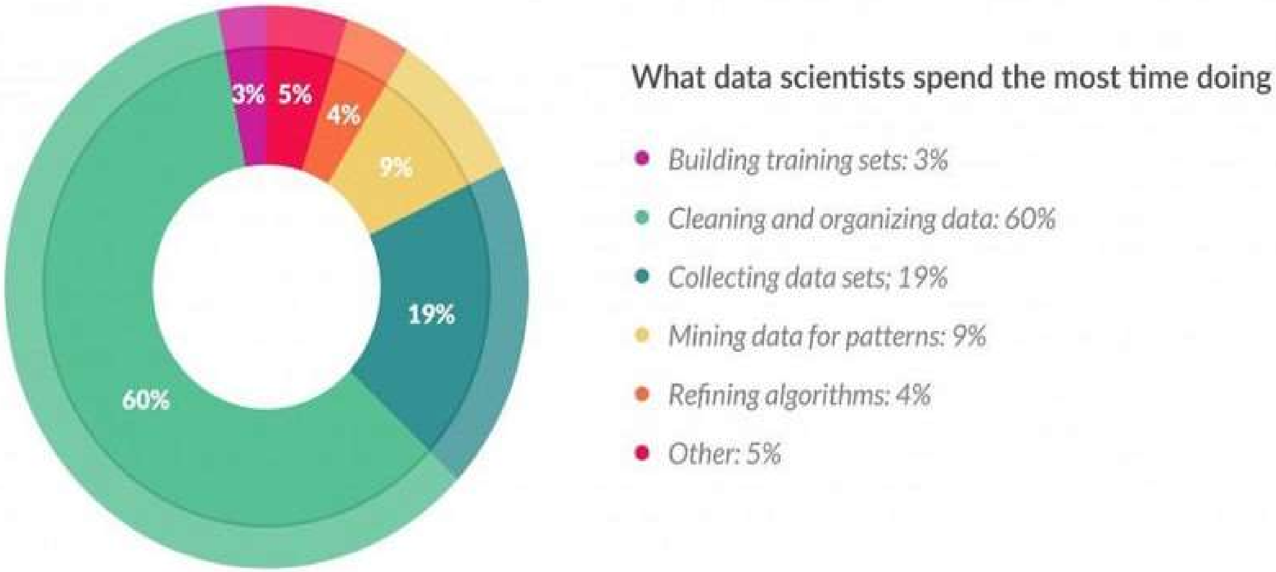
- Feature Engineering is a process of extracting useful feature from raw data using math, statistics and domain knowledge

-

Feature Engineering cont…

- Feature Engineering takes 80% energies of a modeling.

- Good feature engineering leads to good model insight and prediction



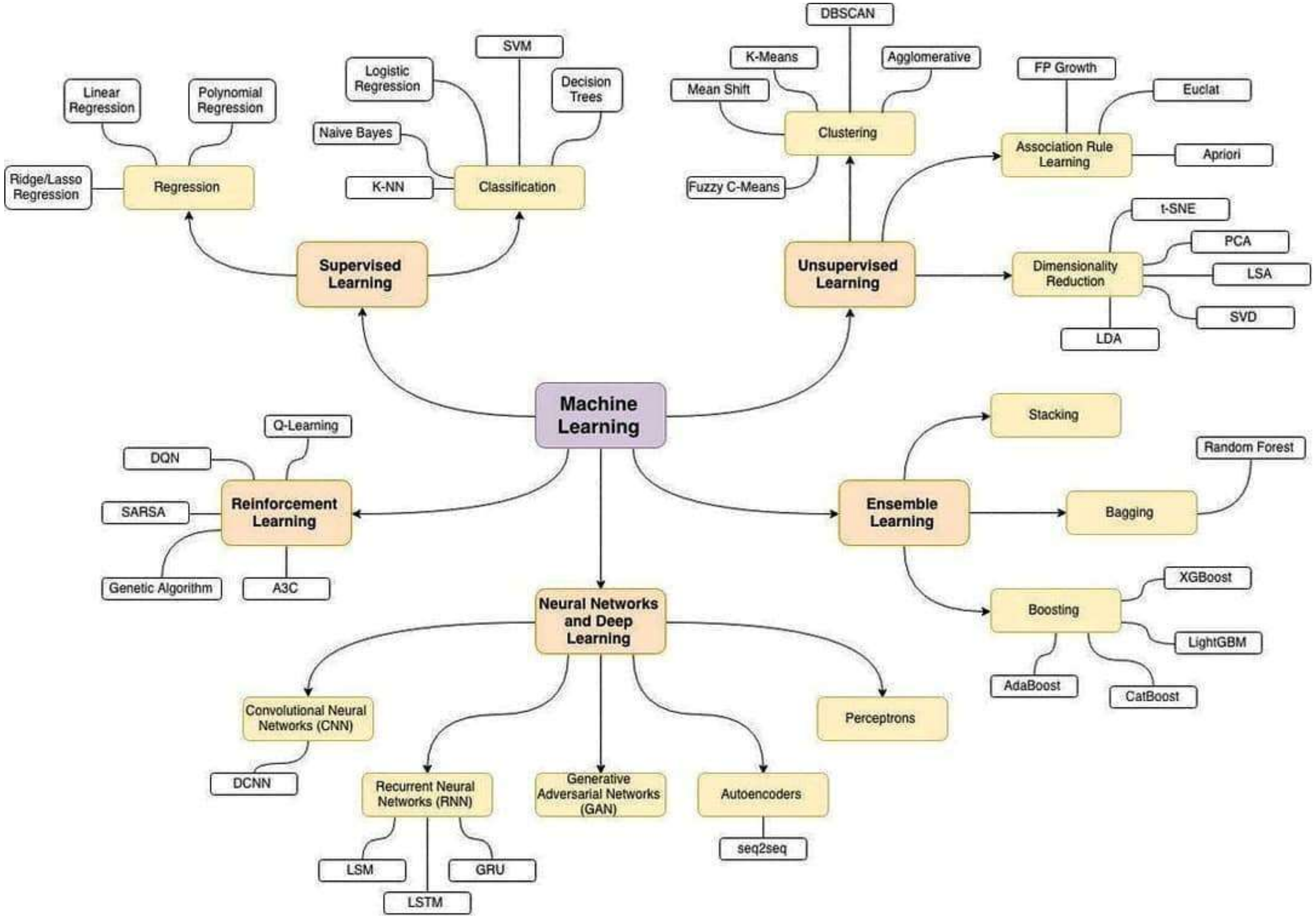
Feature Engineering cont…

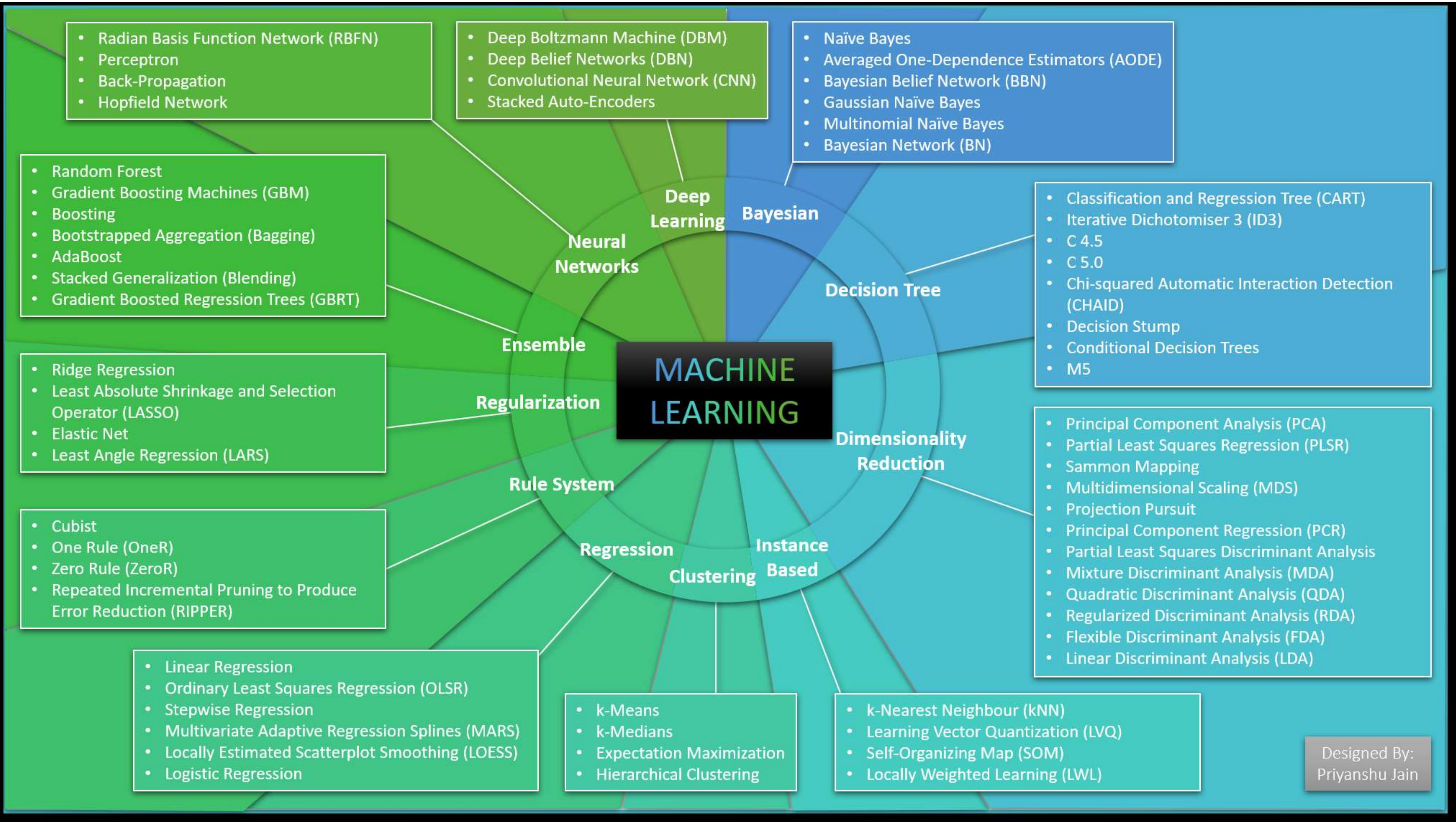
- **List of Techniques**
  1. Imputation
  2. Handling Outliers
  3. Binning
  4. Log Transform
  5. One-Hot Encoding
  6. Grouping Operations
  7. Feature Split
  8. Scaling
  9. Extracting Date

I think the best way to achieve expertise in feature engineering is practicing different techniques on various datasets and observing their effect on model performances

Machine Learning towards Data Science

- Machine learning is a backbone of Data Science.
- Even individual machine learning has good job market
- It helps modeling, feature extraction, feature reduction, etc.





MACHINE LEARNING

Neural Networks

- Radian Basis Function Network (RBFN)
- Perceptron
- Back-Propagation
- Hopfield Network

Deep Learning

- Deep Boltzmann Machine (DBM)
- Deep Belief Networks (DBN)
- Convolutional Neural Network (CNN)
- Stacked Auto-Encoders

Bayesian

- Naïve Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Gaussian Naïve Bayes
- Multinomial Naïve Bayes
- Bayesian Network (BN)

Decision Tree

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C 4.5
- C 5.0
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- Conditional Decision Trees
- M5

Dimensionality Reduction

- Principal Component Analysis (PCA)
- Partial Least Squares Regression (PLSR)
- Sammon Mapping
- Multidimensional Scaling (MDS)
- Projection Pursuit
- Principal Component Regression (PCR)
- Partial Least Squares Discriminant Analysis
- Mixture Discriminant Analysis (MDA)
- Quadratic Discriminant Analysis (QDA)
- Regularized Discriminant Analysis (RDA)
- Flexible Discriminant Analysis (FDA)
- Linear Discriminant Analysis (LDA)

Instance Based

- k-Nearest Neighbour (kNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)

Clustering

- k-Means
- k-Medians
- Expectation Maximization
- Hierarchical Clustering

Regression

- Linear Regression
- Ordinary Least Squares Regression (OLSR)
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)
- Logistic Regression

Rule System

- Cubist
- One Rule (OneR)
- Zero Rule (ZeroR)
- Repeated Incremental Pruning to Produce Error Reduction (RIPPER)

Regularization

- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least Angle Regression (LARS)

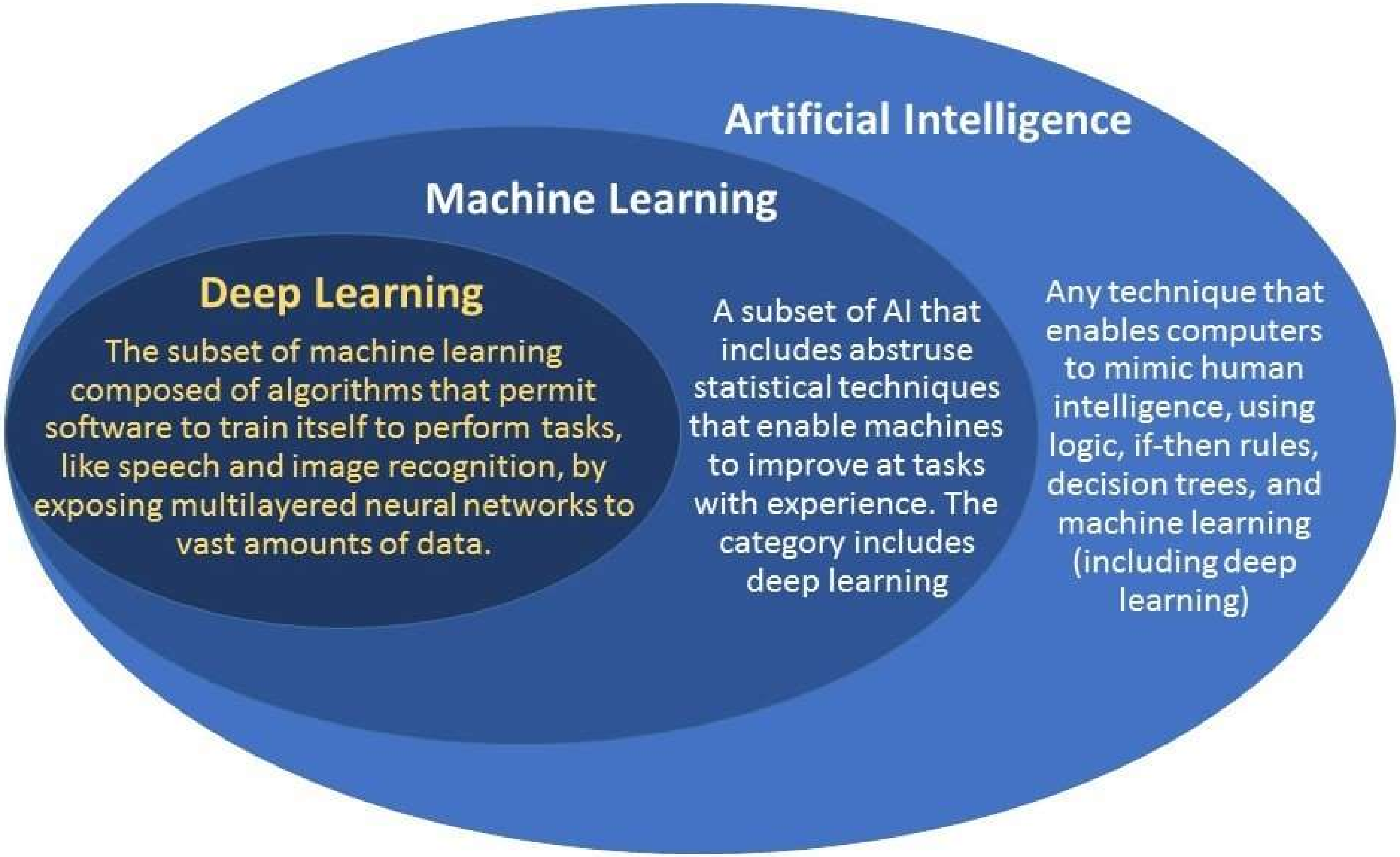
Ensemble

- Random Forest
- Gradient Boosting Machines (GBM)
- Boosting
- Bootstrapped Aggregation (Bagging)
- AdaBoost
- Stacked Generalization (Blending)
- Gradient Boosted Regression Trees (GBRT)

Designed By:
Priyanshu Jain

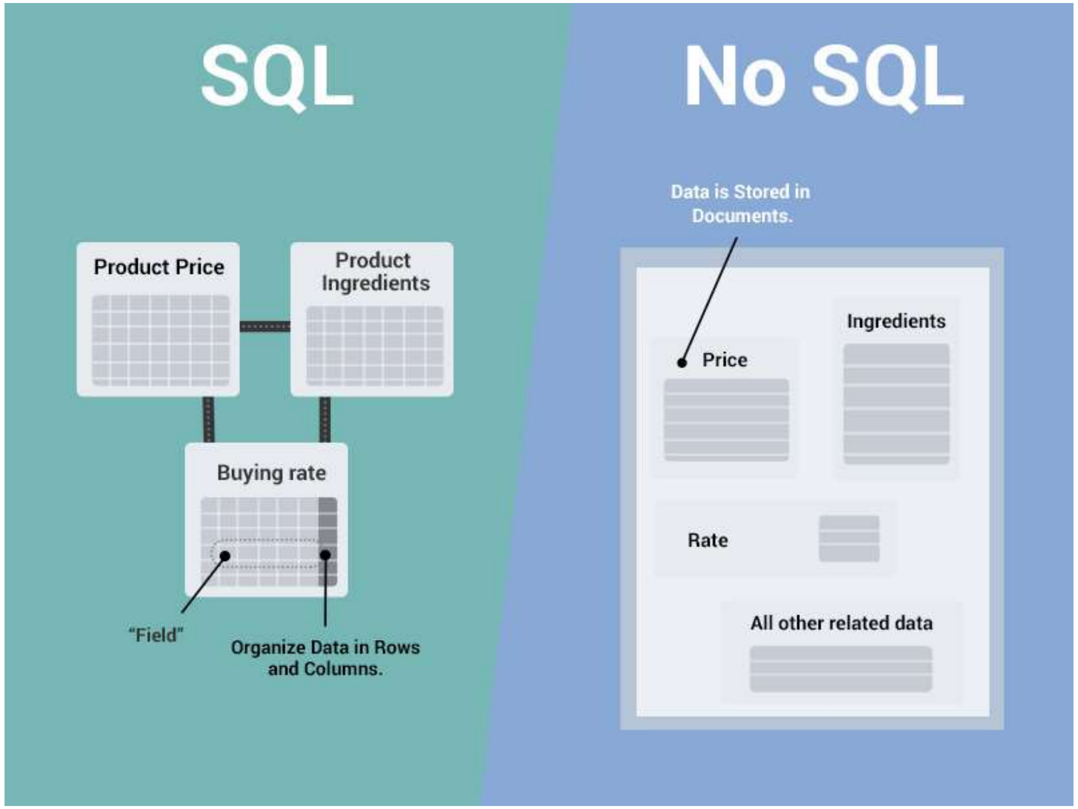
Differences between AI, ML, and DL,

-

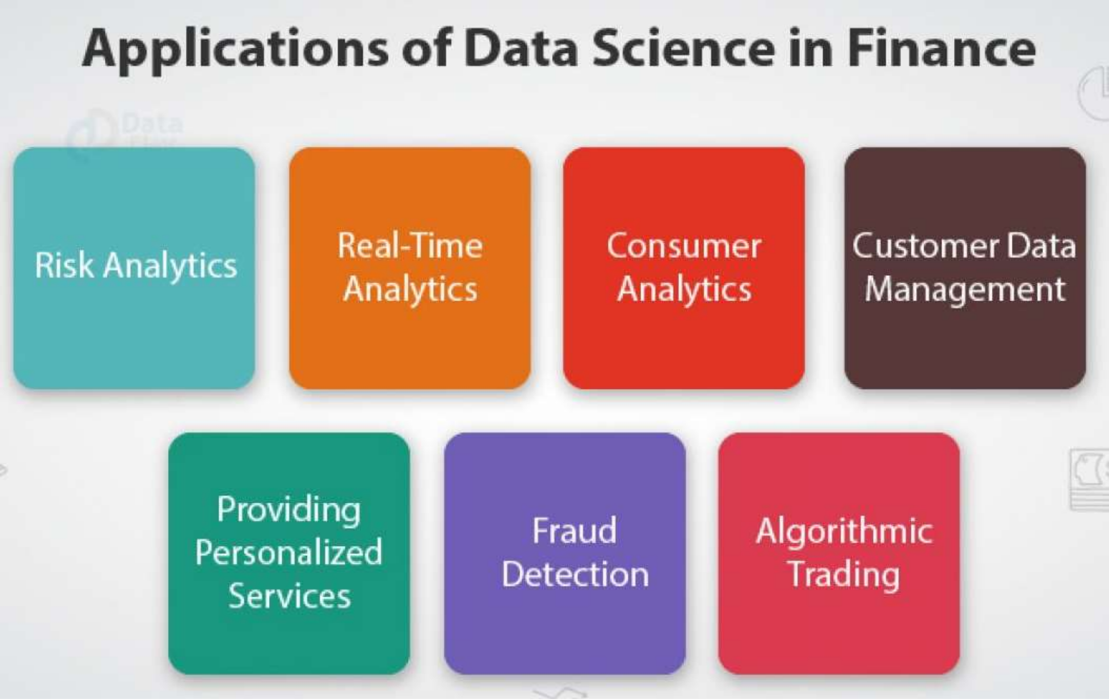

Artificial Intelligence
Machine Learning
Deep Learning
The subset of machine learning composed of algorithms that permit software to train itself to perform tasks, like speech and image recognition, by exposing multilayered neural networks to vast amounts of data.
A subset of AI that includes abstruse statistical techniques that enable machines to improve at tasks with experience. The category includes deep learning
Any technique that enables computers to mimic human intelligence, using logic, if-then rules, decision trees, and machine learning (including deep learning)

No SQL for big DATA

- Some time we need to access the partial data
- Whole available data will be stored in no SQL
- Mongo DB is very popular and easy to integrate with piton.



Real Time Application for Data Science



Applications of Data Science in Finance

Risk Analytics

Real-Time Analytics

Consumer Analytics

Customer Data Management

Providing Personalized Services

Fraud Detection

Algorithmic Trading

Applications of Data Science in Real time world

Data Wrangling and Munging

- Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making. It more appropriate and valuable for a variety of downstream purposes such as analytics.
- A data wrangler is a person who performs these transformation operations.

Data Wrangling and Munging cont…

- This may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses.
- Data munging as a process typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data using algorithms (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.