Python for Data Science

Data Science: Getting Value out of Data

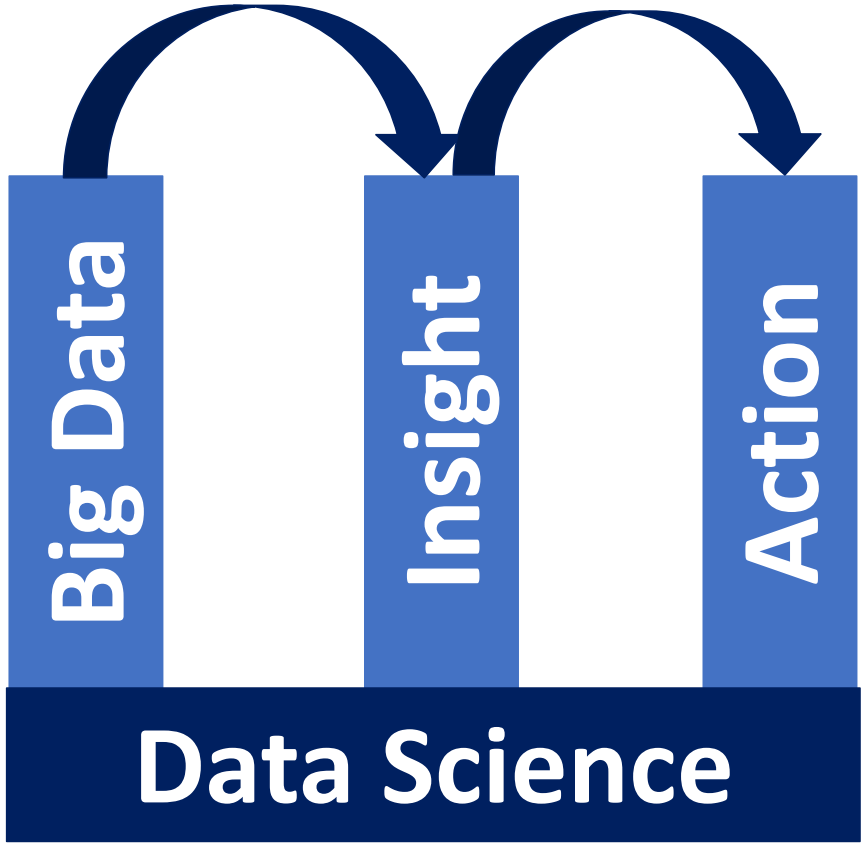
Dr. Ilkay Altintas and Dr. Leo Porter

Twitter: #UCSDpython4DS

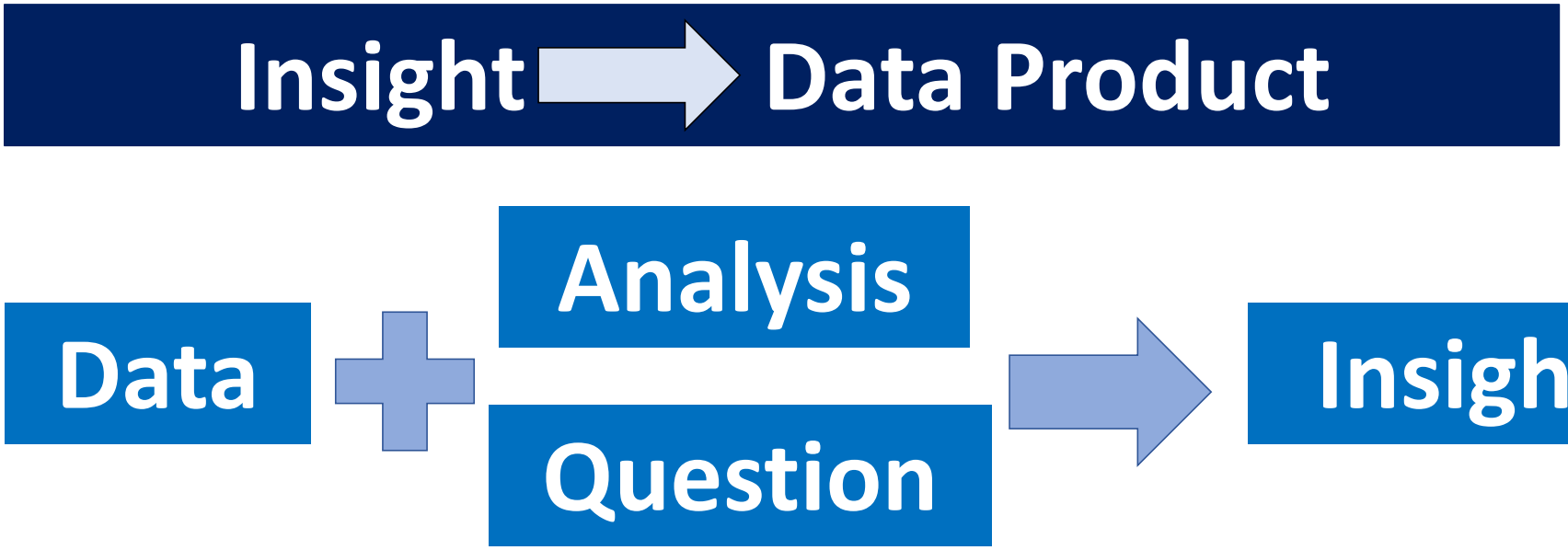
By the end of this video, you will be able:

- Describe what modern data science is
- Explain why data science is the key to getting value out of data and where the growing interest for it comes from
- List a recommended set of skills for a data scientist

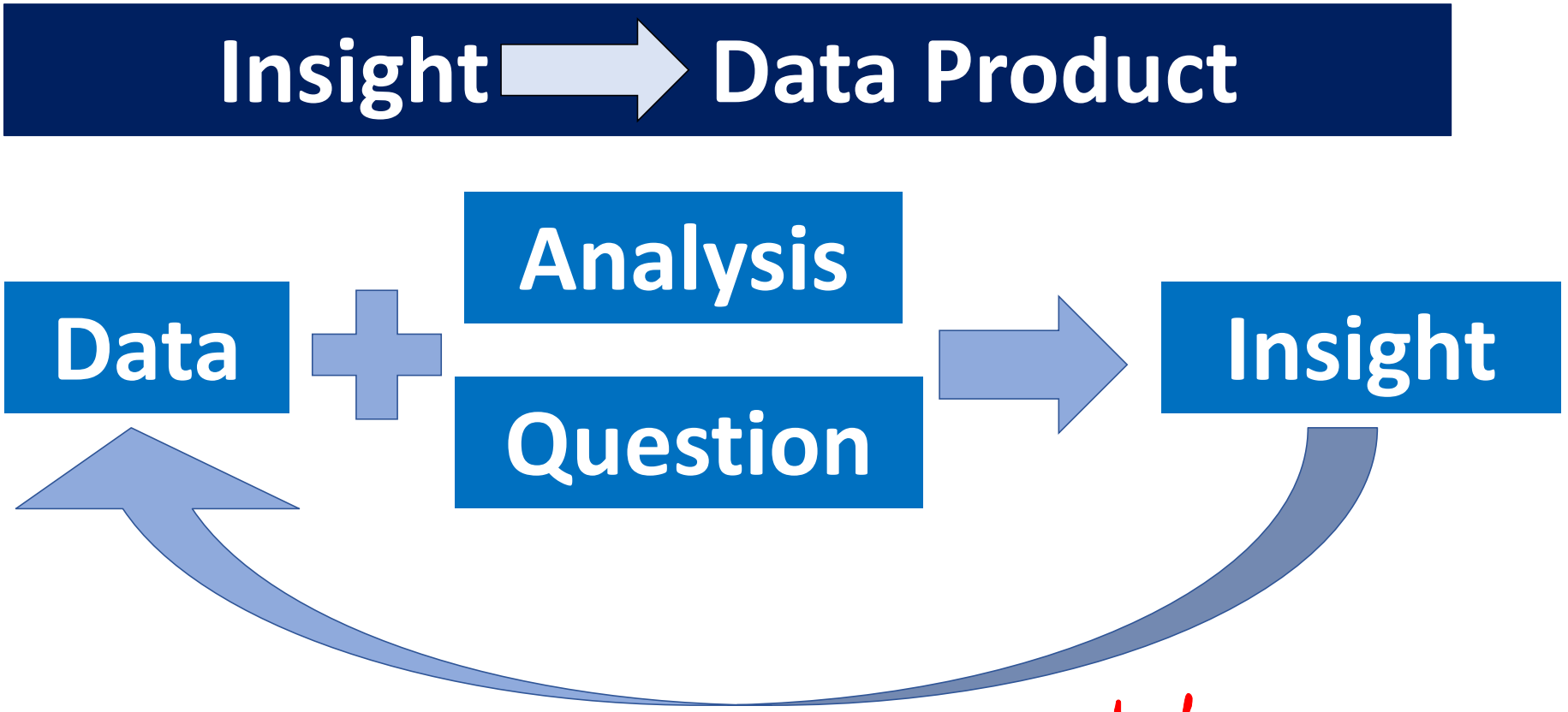
Data Science turns data into insights or even actions.



Data Science can be thought of as the basis of empirical research, for data is used to inform our hypotheses and provide observations.



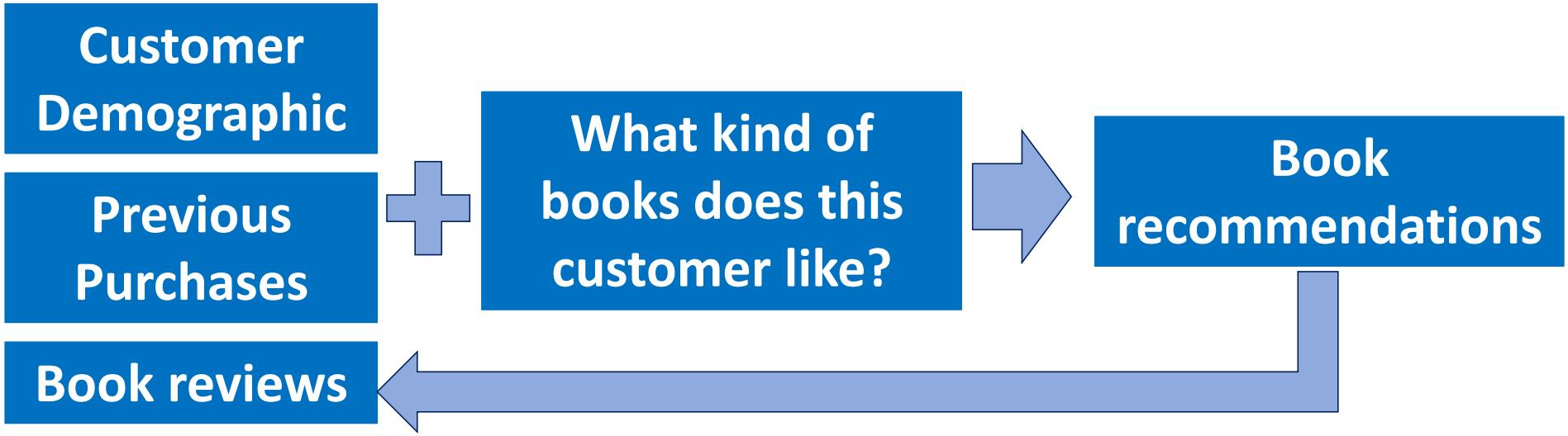
Insight is extracted from a diverse amount of data through a combination of exploratory data analysis and modeling.



not a one time product

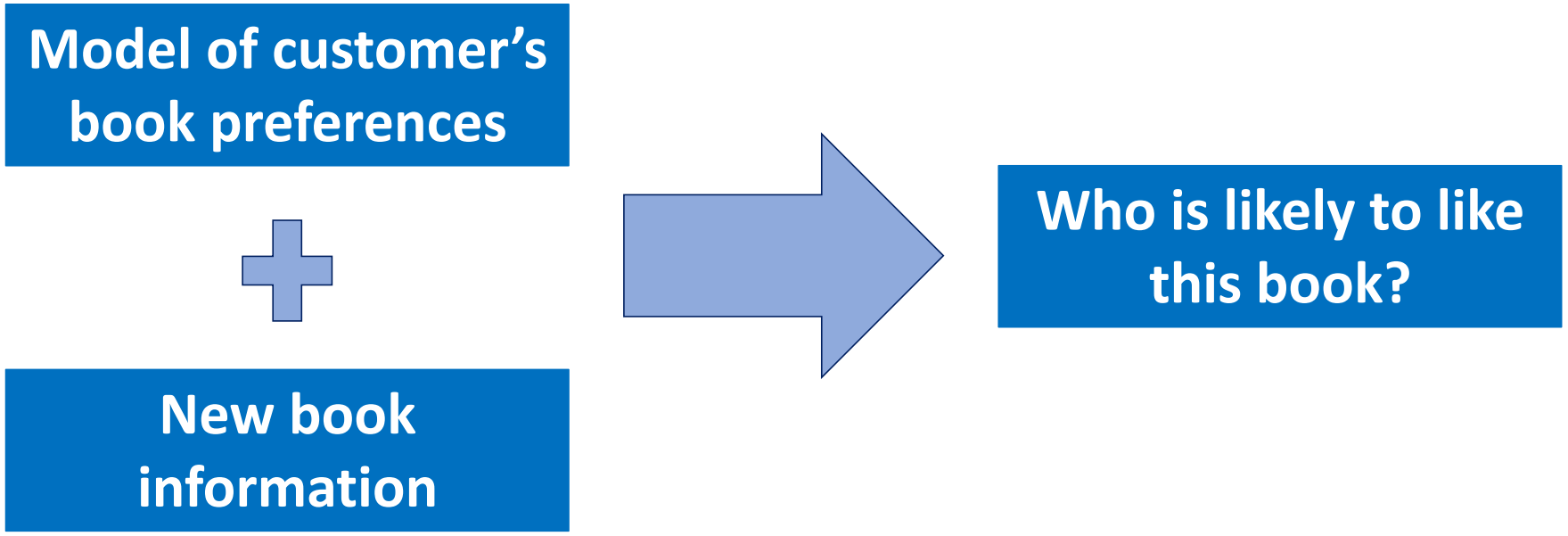
Improve insights by creating insights using insights.

Book Recommendations

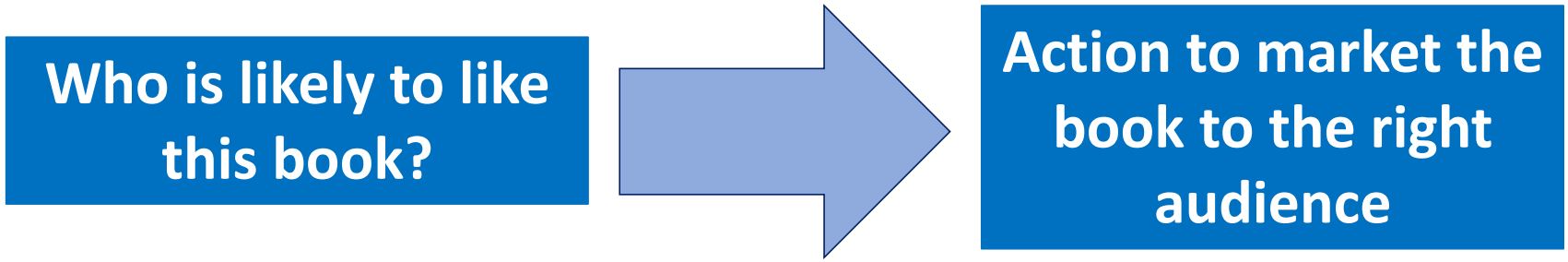


amazon.com

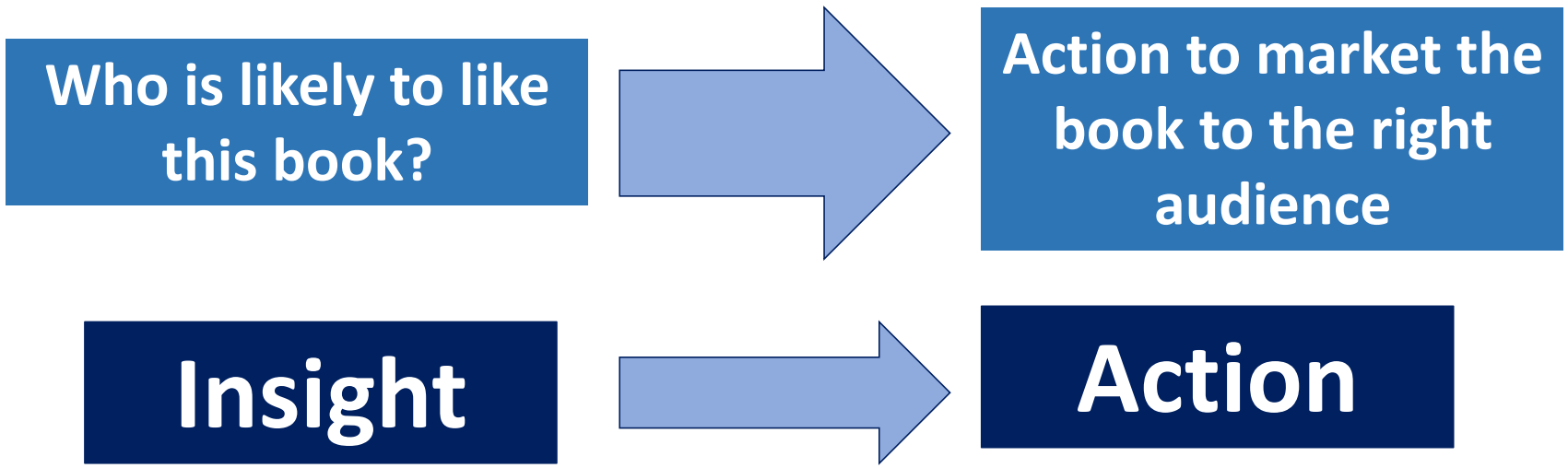
Find Potential Audience for a Book



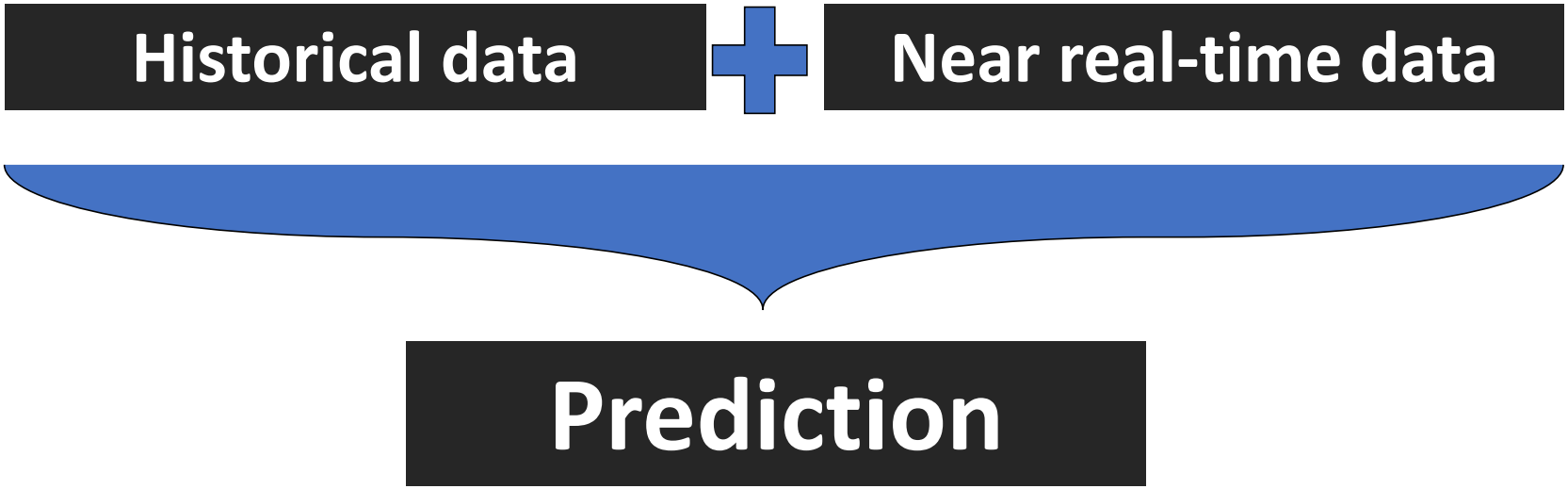
Market a New Book



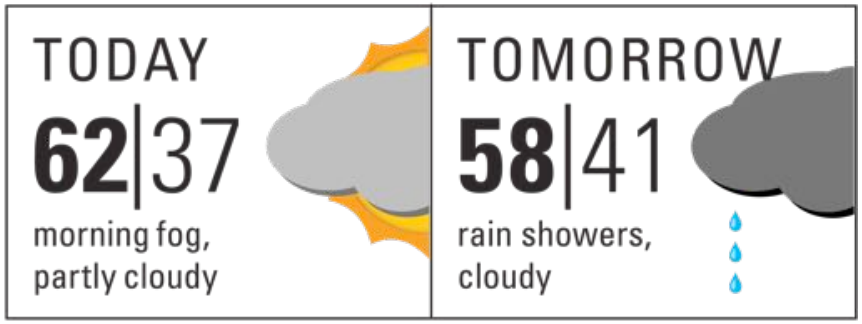
Market a New Book



Actionable Information



Prediction



Action

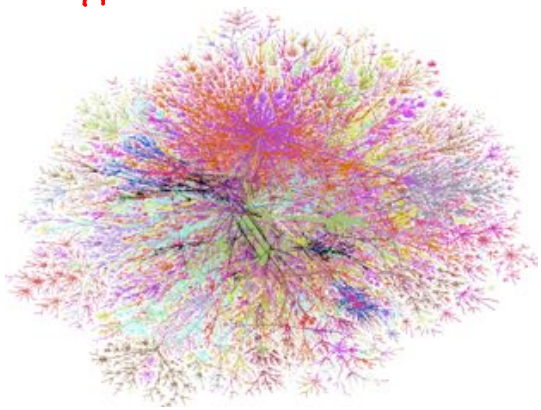
Why the Increased Interest in Data Science?

ability to collect data in real time

Storing

Networking

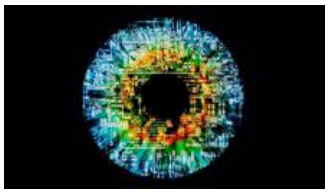
Python for Data Science



BIG DATA

COMPUTING AT SCALE

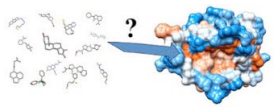
New era of data science!



Many dynamic data-driven applications

Smart Manufacturing

Computer-Aided Drug Discovery



Personalized Precision Medicine

Smart Cities

Smart Grid and Energy Management

Disaster Resilience and Response

How Much Data Is Big Data?

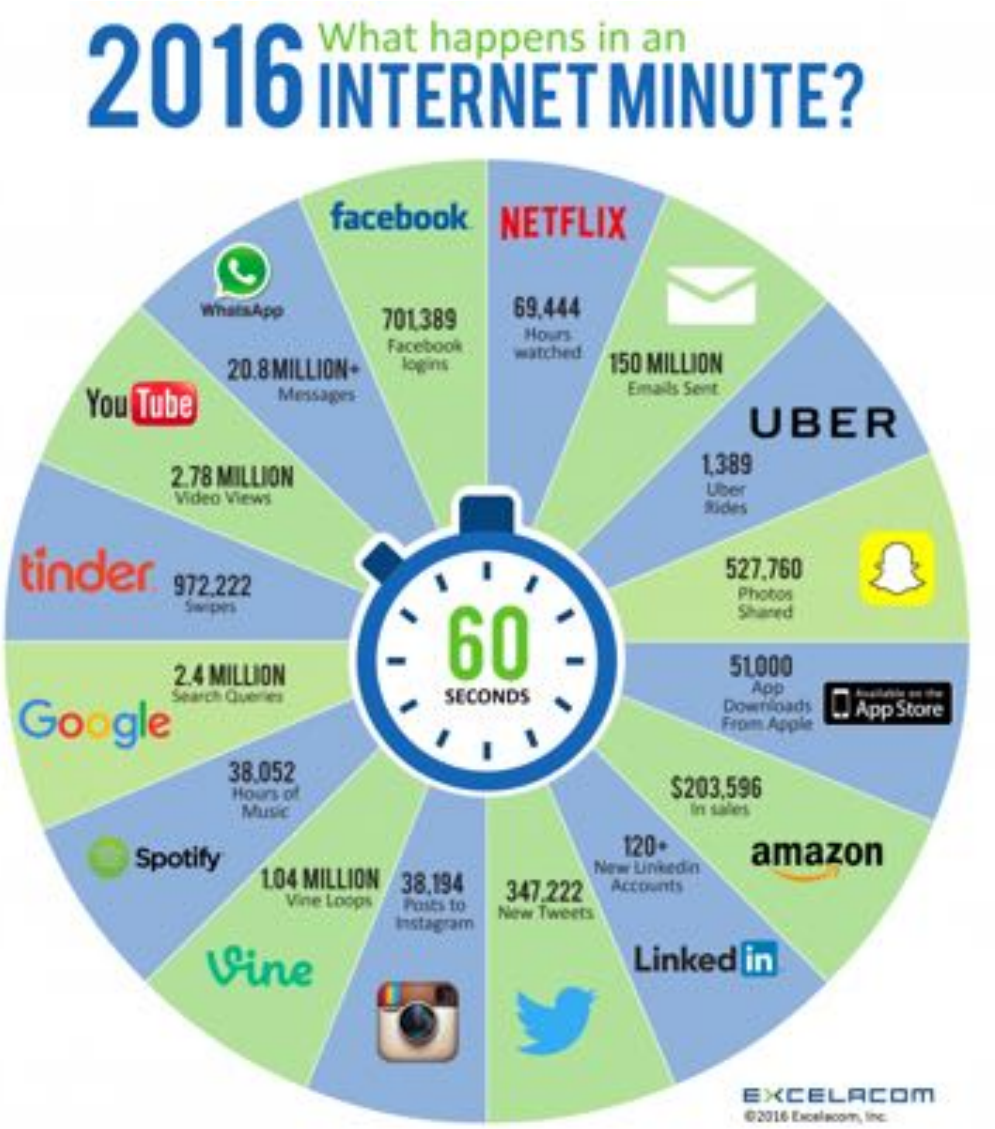


Image Source: http://www.marketwatch.com/story/one-chart-shows-everything-that-happens-on-the-internet-in-just-one-minute-2016-04-26

Every minute…

204 Million emails

200,000 photos

facebook

1.8 Million likes

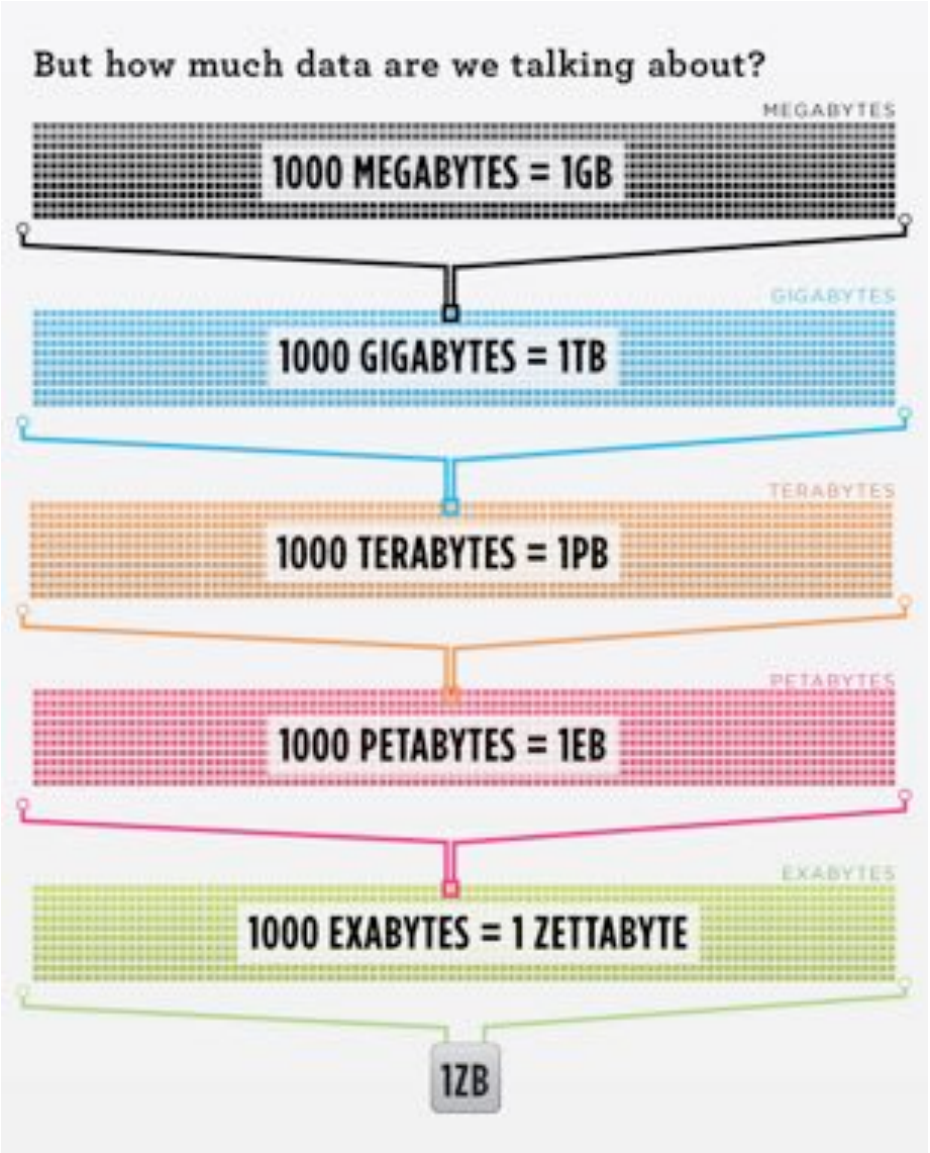
2.78 Million video views

YouTube

72 hours of video uploads

Scientific Data Management and Analysis

- HPWREN: hpwren.ucsd.edu
  - 30 TB of data annually
- MODIS: modis.gsfc.nasa.gov
  - 219 TB of data annually
- Precision Medicine
  - 4 EB (10^{18} bytes) of data in 2016 (www.fastcompany.com)
- LIGO, Deep Space Network, Protein Data Bank, …



100 MBs ~= couple of volumes of Encyclopedias

A DVD ~= 5 GBs

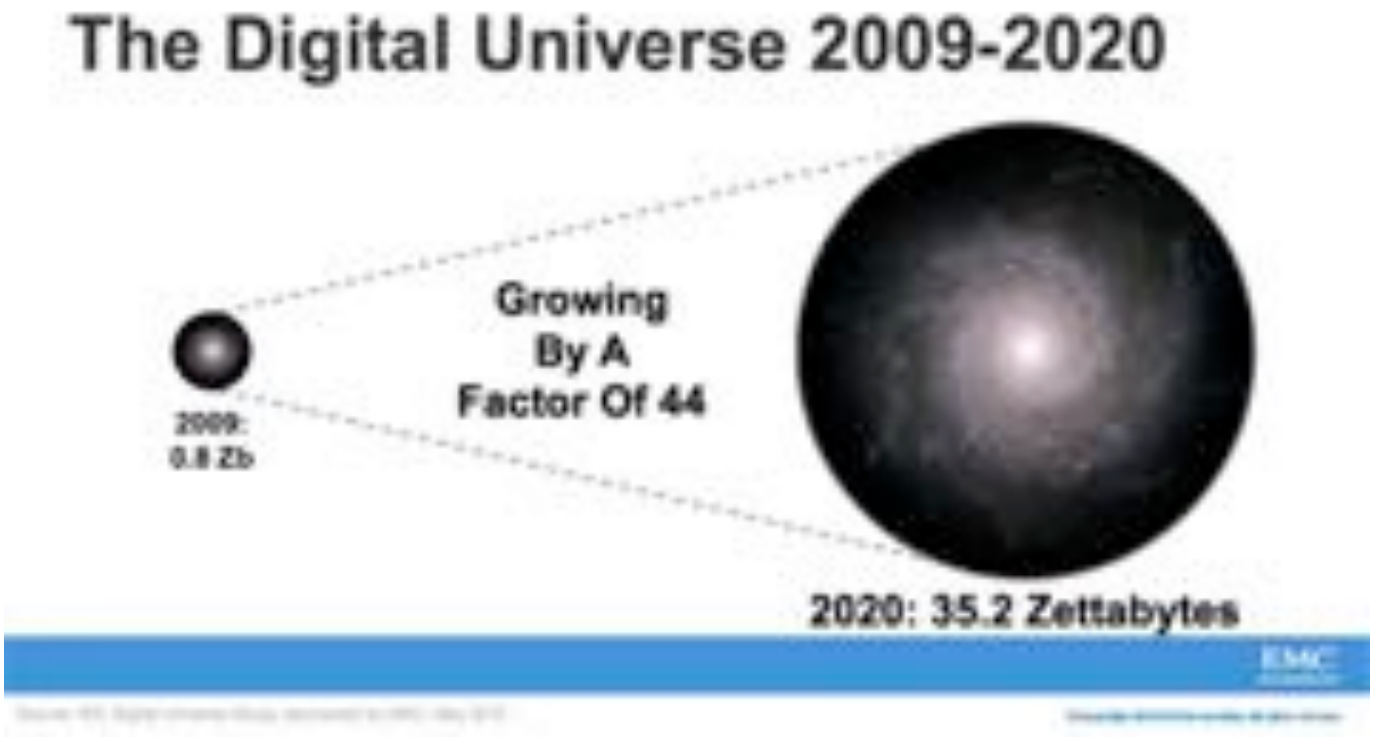
1 TB ~= 300 hours of good quality video

LHC ~= 15 PBs a year

10^{21} GB

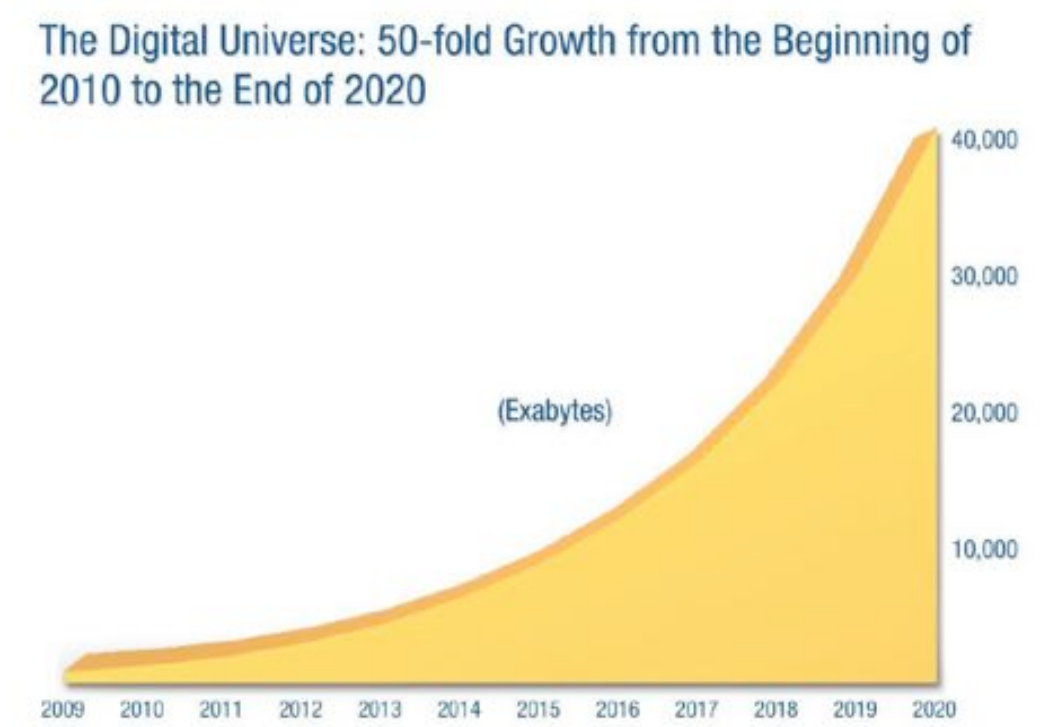
1 Tottabytes = 10^{24} GB

1 brontobytes = 10^{27} GB



data at astronomical scale.

Exponential data growth!



This IDC graph predicts exponential growth of data from around 3 zettabytes in 2013 to approximately 40 zettabytes by 2020. An exabyte equals 1,000,000,000,000,000,000 bytes and 1,000 exabytes equals one zettabyte. Source: IDC's Digital Universe Study, December 2012, http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

Data Deluge

"We are drowning in information and starving for knowledge"

– John Naisbitt

Source: Megatrends, 1982

Challenges

1) Information overload

Image Source: http://www.digitalzenway.com/2011/12/data-diet-a-resolution-you-can-stick-to/

How do we find the connections?

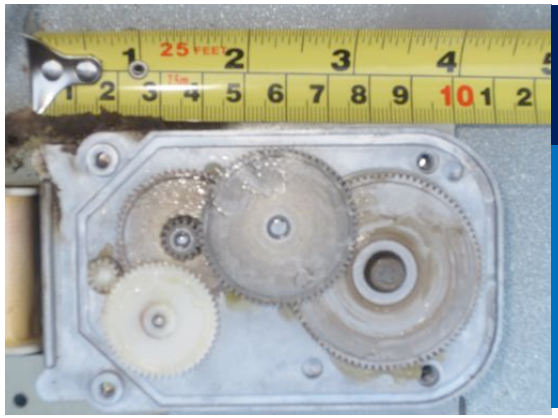
data management, data driven methods, scalable tools for dynamic coordinations, and scalable execution, and a skilled interdisciplinary workforce.

Modern Data Science Skills

- Programming in Python
- Statistics
- Machine Learning
- Scalable Big Data Analysis

Application

drug effectiveness analysis,
Crime pattern detection,
self-driving cars



Data Science

The sum is bigger than the parts!

Big Data → **Actionable Insight**

Modern Data Science Skills

- Python Programming
- Statistical Analysis
- Machine Learning
- Scalable Big Data Analysis

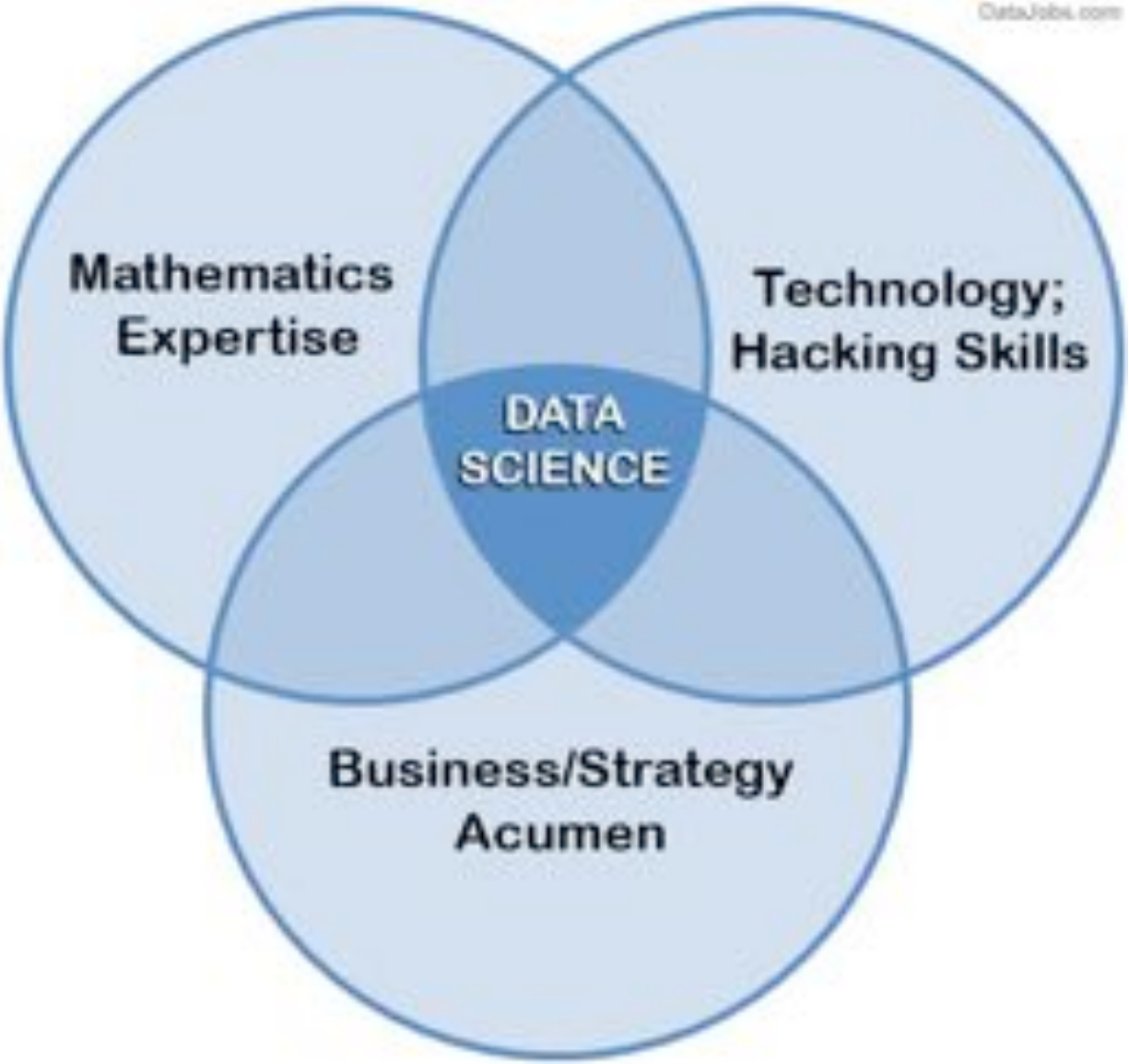
The Role of Python Programming in Data Science

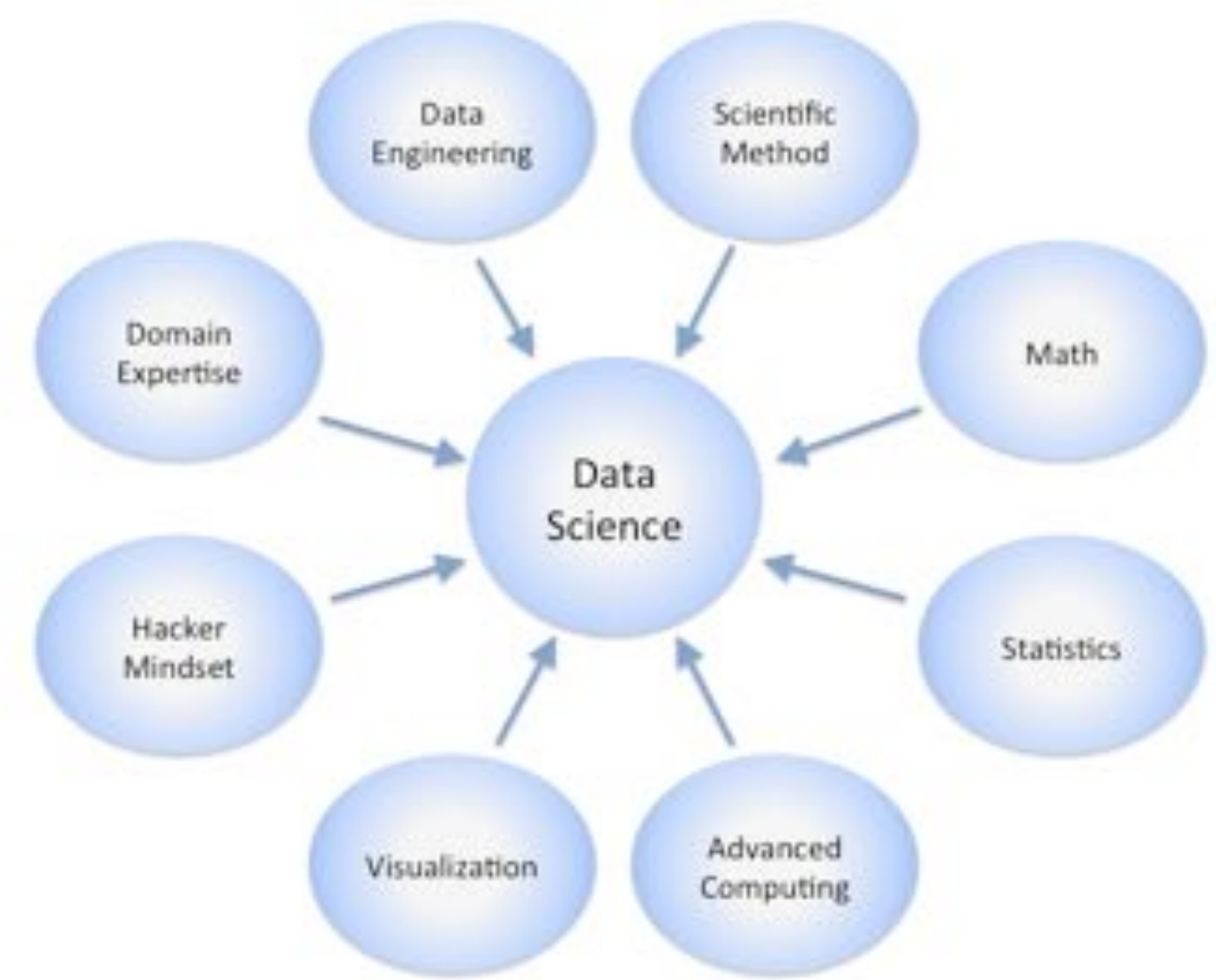
Dr. Ilkay Altintas and Dr. Leo Porter

Twitter: #UCSDpython4DS

By the end of this video, you will be able:

- List some of the traits of modern data scientists
- Explain why Python is a good programming language for data science
- Recite four major Python modules that are useful for data analysis


Mathematics Expertise
Technology; Hacking Skills
DATA SCIENCE
Business/Strategy Acumen


Data Engineering
Scientific Method
Domain Expertise
Math
Data Science
Hacker Mindset
Statistics
Visualization
Advanced Computing

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include marketing strategy and optimization: customer tracking and on site analytics: predictive analytics and econometrics: data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include marketing strategy and optimization: customer tracking and on site analytics: predictive analytics and econometrics: data warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

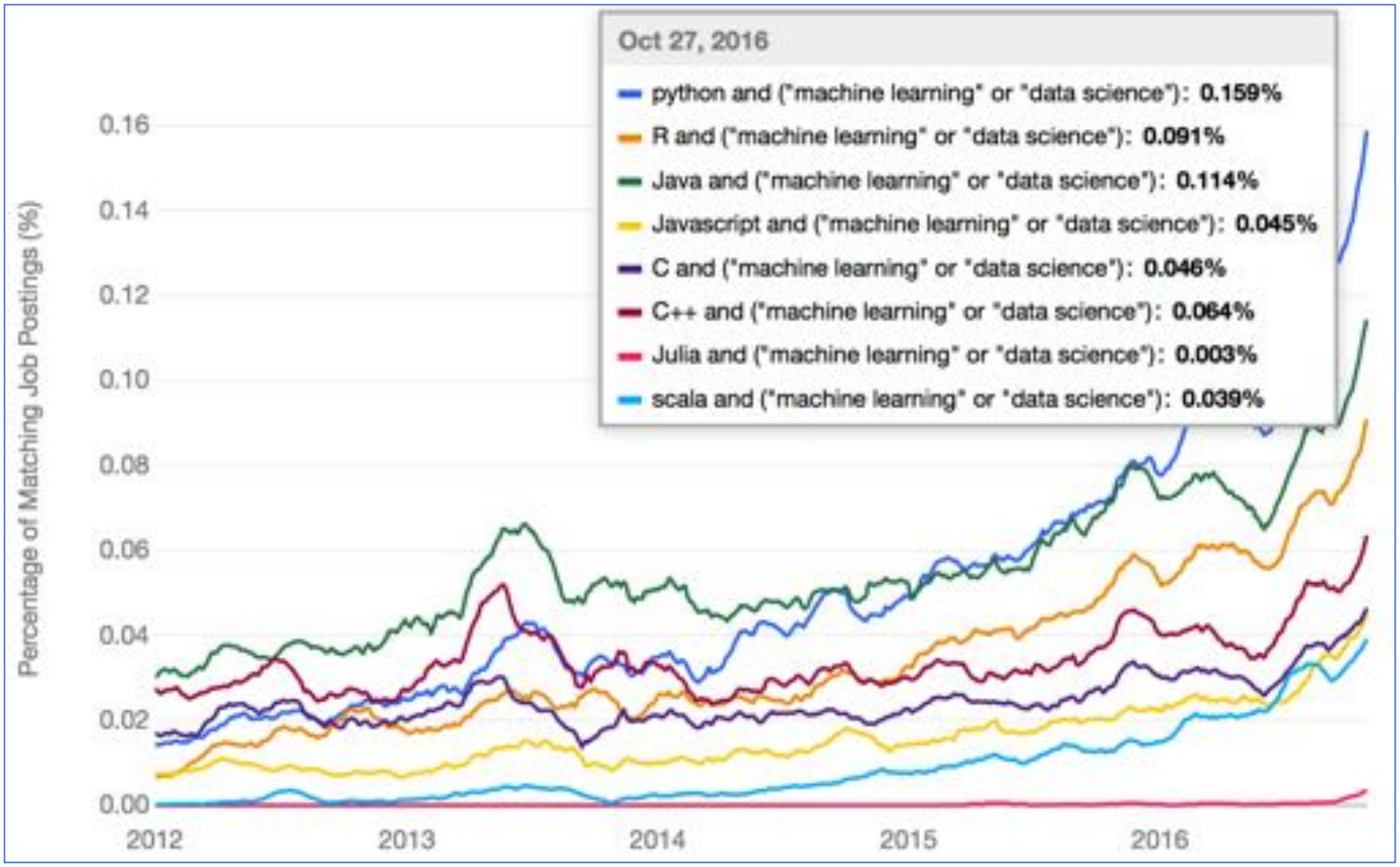
Are data scientists unicorns?

Data science is team sport!

Data scientists…

- Have passion for data
- Relate problems to analytics
- Care about engineering solutions
- Exhibit curiosity
- Communicate with teammates

Python for Data Science



http://www.kdnuggets.com/2017/01/most-popular-language-machine-learning-data-science.html

Why Python for Data Science?

- Easy-to-read and learn
- Vibrant community
- Growing and evolving set of libraries
  - Data management
  - Analytical processing
  - Visaualization
- Applicable to each step in the data science process
- Notebooks

What to look forward to!

- Jupyter notebooks
- NumPy
- Pandas
- Matplotlib
- Scikit-Learn
- BeautifulSoup

Python for Data Science

Case Study: Soccer Data Analysis

Dr. Ilkay Altintas and Dr. Leo Porter

Twitter: #UCSDpython4DS

By the end of this video, you will be able:

- Talk about the “Big Picture” of data science through a soccer case study
- Generate statistics about a soccer data set
- Summarize how data cleaning and correlations were applied to an existing dataset
- Recite the data visualization techniques employed in this study
- Explain how clustering similar groups and plotting these clusters helped the case study
- Recall what was used to drawing conclusion based on data analysis

Week 1 Case Study: Soccer Data Analysis

kaggle

Dataset location: https://www.kaggle.com/hugomathien/soccer

- Form meaningful player groups
- Discover other players that are similar to your favorite athlete
- Form strong teams by using analytics

Understanding the Benefits

Ask yourself:
"What insights do I expect to get!"

INSIGHTS

- Better understanding and insights on
  - player strengths
  - enhancing performance
  - critical attributes for a player's performance

ACTIONS

- Coach can design programs that improve these areas in teams

Basic Steps in a Data Science Project

ACQUIRE
- Import raw dataset into your analytics platform

PREPARE
- Explore & Visualize
- Perform Data Cleaning

ANALYZE
- Feature Selection
- Model Selection
- Analyze the results

REPORT
- Present your findings

ACT
- Use them

Data Collection from Diverse Sources

- Databases
  - Relational
  - Non-relational (NoSQL)
- Text files
  - CSV files
  - Text files
- Live feeds
  - Sensors
  - Online Platforms
    - Twitter
    - Live feeds of weather observations

Data Ingestion to Analytics Platform

Data Preparation: Explore using Statistics

```
df.describe().transpose()
```

	count	mean	std	min	25%	50%	75%	max
id	183978.0	91989.500000	53110.018250	1.0	45995.25	91989.5	137983.75	183978.0
player_fifa_api_id	183978.0	165671.524291	53851.094769	2.0	155798.00	183488.0	199848.00	234141.0
player_api_id	183978.0	135900.617324	136927.840510	2625.0	34763.00	77741.0	191080.00	750584.0
overall_rating	183142.0	68.600015	7.041139	33.0	64.00	69.0	73.00	94.0
potential	183142.0	73.460353	6.592271	39.0	69.00	74.0	78.00	97.0
crossing	183142.0	55.086883	17.242135	1.0	45.00	59.0	68.00	95.0
finishing	183142.0	49.921078	19.038705	1.0	34.00	53.0	65.00	97.0
heading_accuracy	183142.0	57.266023	16.488905	1.0	49.00	60.0	68.00	98.0
short_passing	183142.0	62.429672	14.194068	3.0	57.00	65.0	72.00	97.0
volleys	181265.0	49.468436	18.256618	1.0	35.00	52.0	64.00	93.0
dribbling	183142.0	59.175154	17.744688	1.0	52.00	64.0	72.00	97.0

Data Cleaning

- Why do we need to clean data?
  - Missing entries
  - Garbage values
  - NULLs
- How do we clean data?
  - Remove the entries
  - Impute these entries with a counterpart
    - Ex. Average values of the column
    - Ex. Assign 0, -1, etc

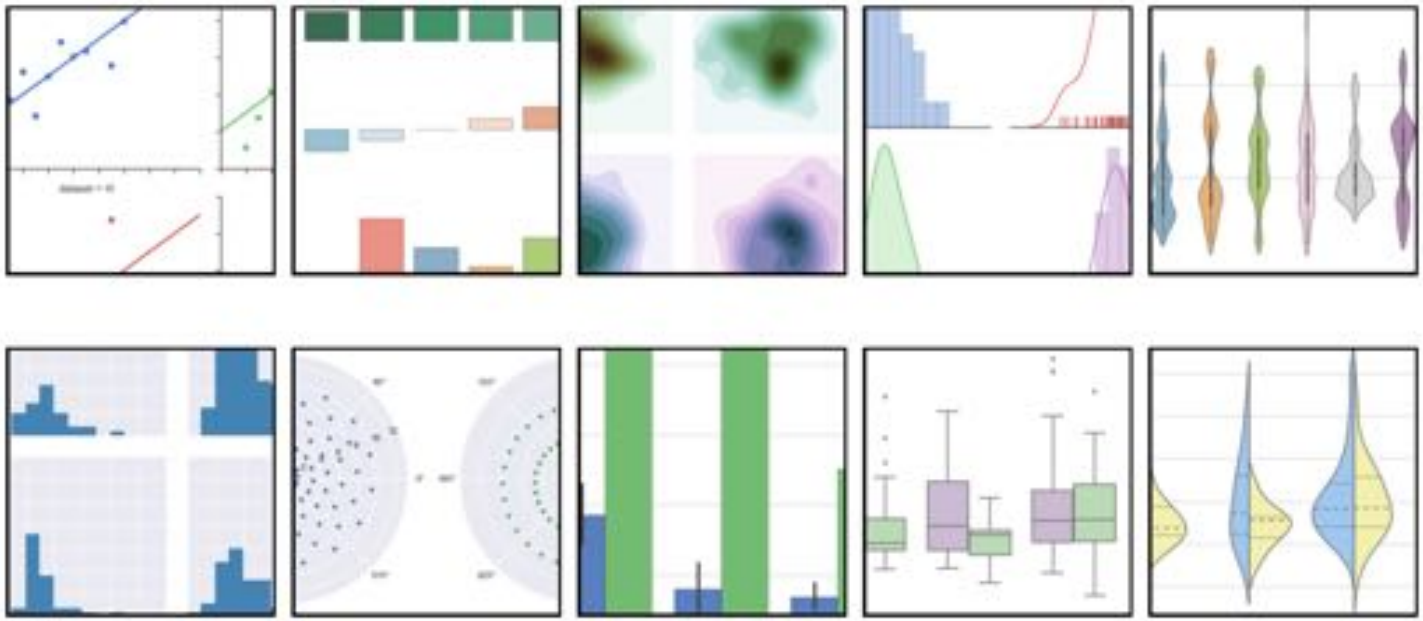
```
#is any row NULL ?

rows = df.shape[0]
df.isnull().any().any(), df.shape
```

```
# Fix it

df = df.dropna()
```

Data Visualization



Convey more in less space and time

Use Graphs when possible

http://seaborn.pydata.org/examples/

Python for Data Science

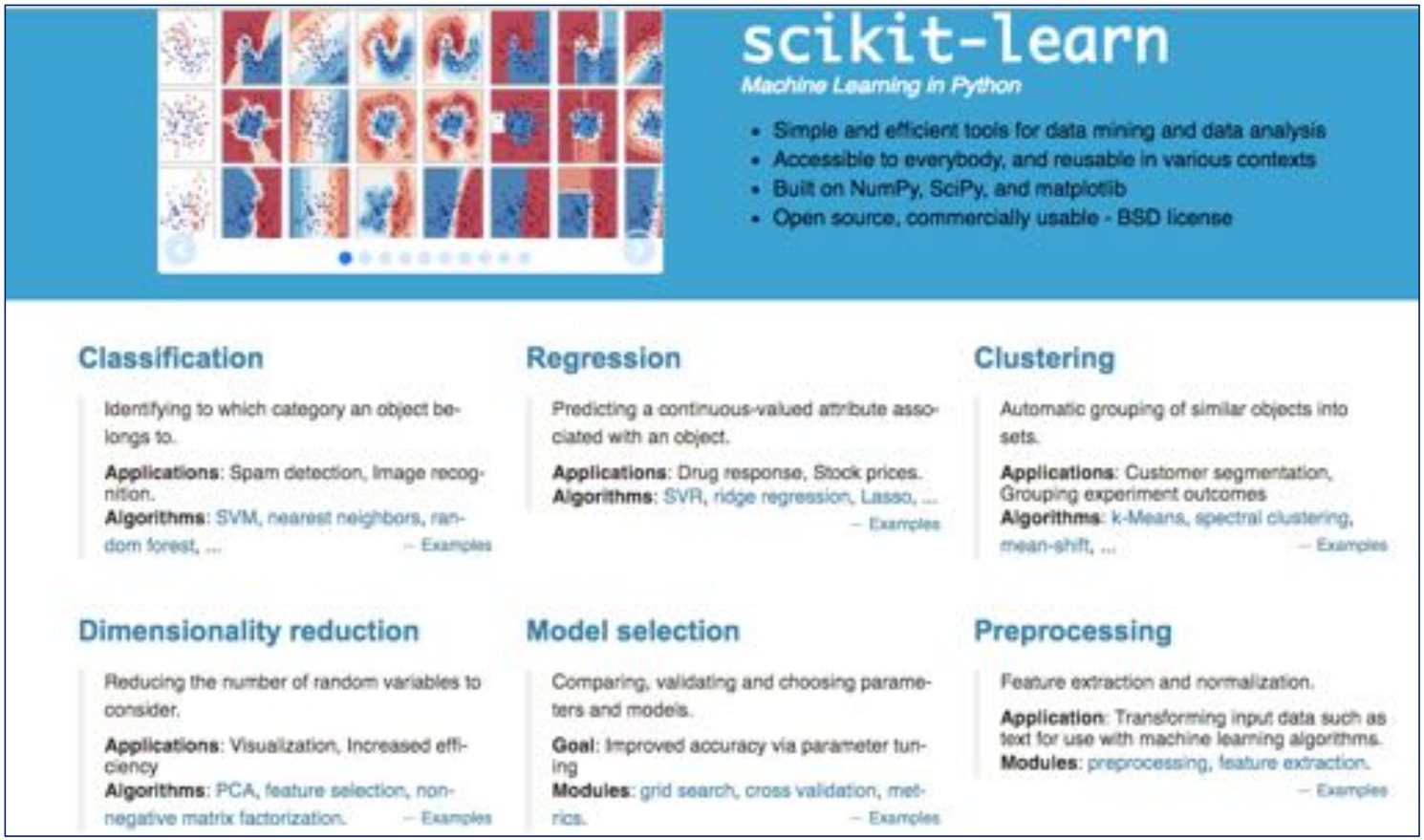
Analysis and Modeling

- Supervised Learning
- Unsupervised Learning
- Semi supervised Learning

Image Source: https://jixta.wordpress.com/2015/07/17/machine-learning-algorithms-mindmap/

Python for Data Science

scikit-learn for Machine Learning in Python



http://scikit-learn.org

Soccer Data Analysis: Feature Selection

- What are intrinsic attributes on which 'you' would group players ?

Agility
Reaction Time
Shot Power
Sprint Speed

Hair Style
Movies the player likes

- You can also build complex features

f (shot power, reaction time)

benefits

1. models that are easier to interpret
2. Models get trained much faster.
3. generalize newer scenarios.

Clustering in Python: **sklearn.cluster**

→ machine learning

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction.	Euclidean distance between points

http://scikit-learn.org/stable/modules/clustering.html

K-Means clustering in Python

```
from sklearn.cluster import Kmeans

…

Y = KMeans(n_clusters=3,random_state=random_state).fit_predict(X)

…
```

How to choose the right algorithm?

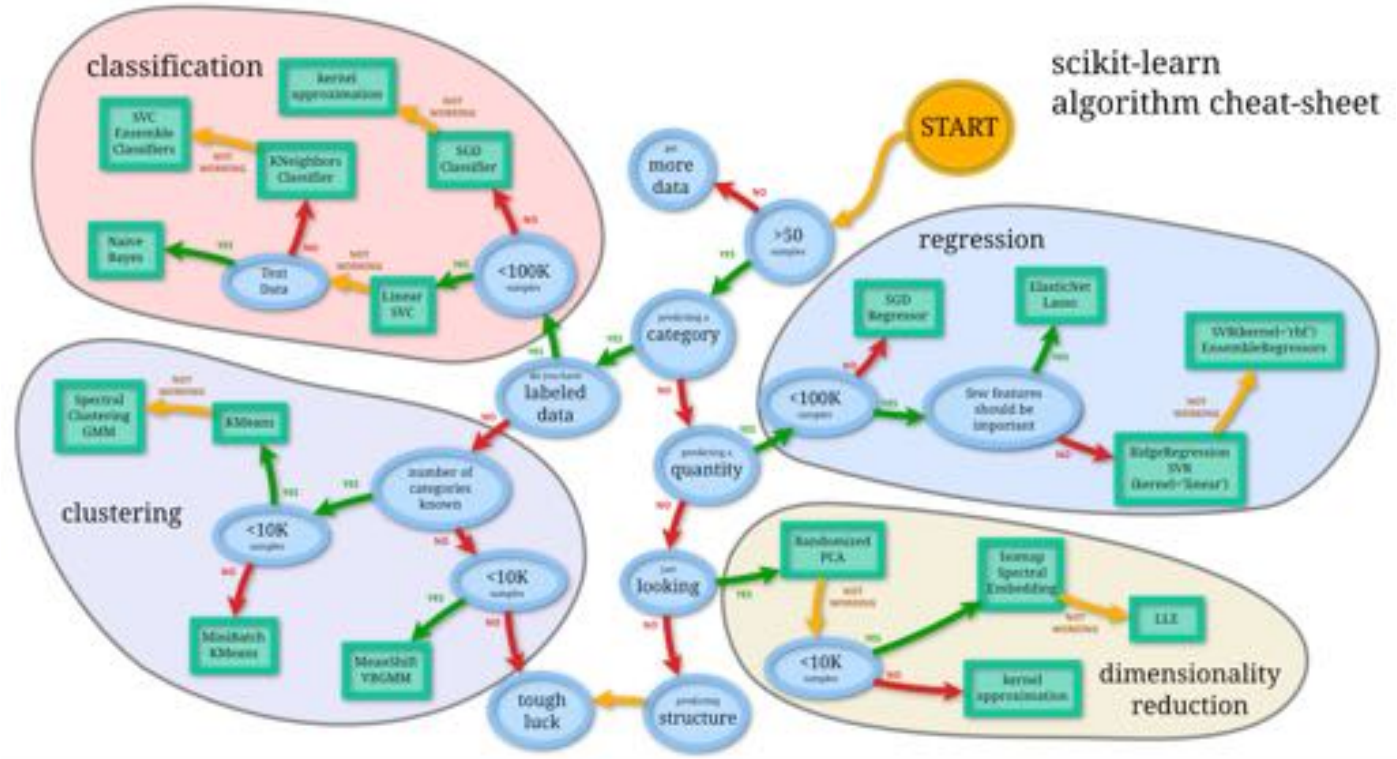


Image Source: http://4.bp.blogspot.com/-o0vLxYf6YZ4/UQVO9K2jxDI/AAAAAAAACt8/Z5w0bSgqkxw/s1600/machine_learning.png

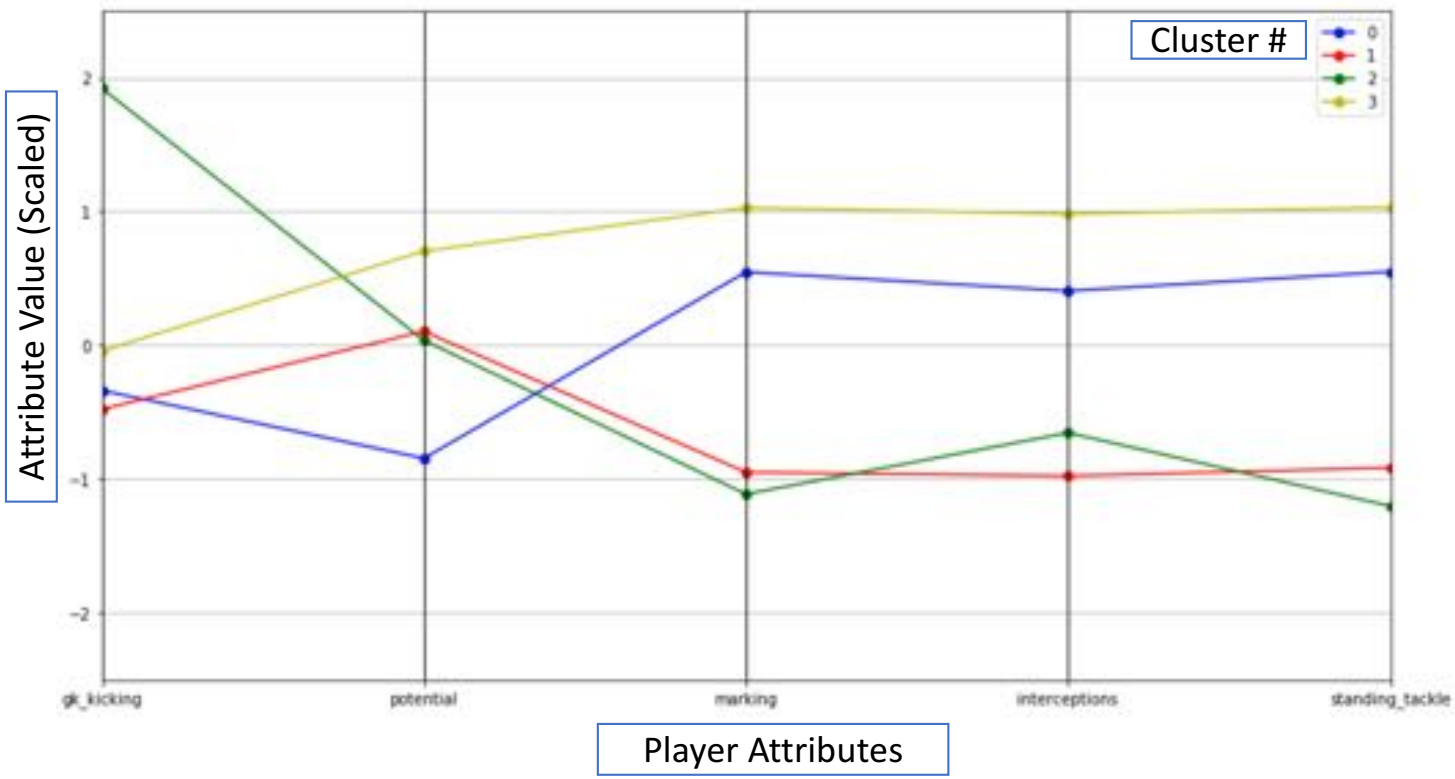
Interpreting Clustering Results

- How many players per cluster ?
  - Too many in few clusters ?
  - Too few ?

Count of players in each cluster	
0	50186
1	55889
2	23783
3	50496

- Look at distribution of features in each cluster
  - Investigate the values for each cluster
  - If few clusters → Plot for comparative analysis

Presenting Data Science Outcomes



Summary

ACQUIRE
PREPARE
ANALYZE
REPORT
ACT

INSIGHTS

- Better understanding and insights on
  - player strengths
  - enhancing performance
  - critical attributes for a player's performance

ACTIONS

- Coach can design programs that improve these areas in teams