Machine Learning per la Finanza

Docenti: Zelda Marino e Paolo Zanetti

Email: {zelda.marino;paolo.zanetti}@uniparthenope.it

Data: 16/04/2021

Calendario

Marzo							
L	M	M	G	V	S	D	Tot Ore
27	28	1	2	3	4	5	*7*
6	7	8	9	10	11	12	*14*
13	14	15	16	17	18	19	*21*
20	21	22	23	24	25	26	*28*
27	28	29	30	31			*35*

Aprile							
L	M	M	G	V	S	D	Tot Ore
					1	2	
3	4	5	6	7	8	9	
10	11	12	13	14	15	16	
17	18	19	20	21	22	23	*42*
24	25	26	27	28	29	30	*49*

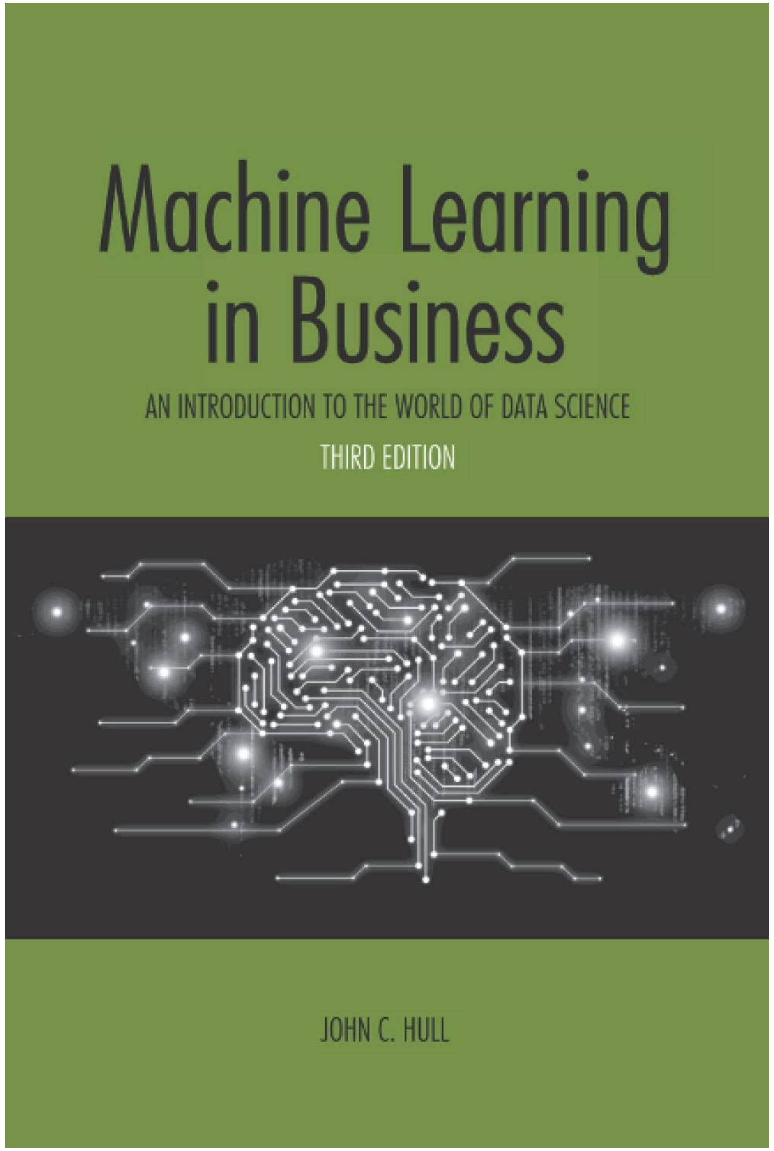
Maggio							
L	M	M	G	V	S	D	Tot Ore
1	2	3	4	5	6	7	*54*
8	9	10	11	12	13	14	*61*
15	16	17	18	19	20	21	*68*
22	23	24	25	26	27	28	*72*
29	30	31					

#	Lezione Marino	#	prove intercorso
#	Lezione Zanetti	#	recuperi
#	Festa accademica		

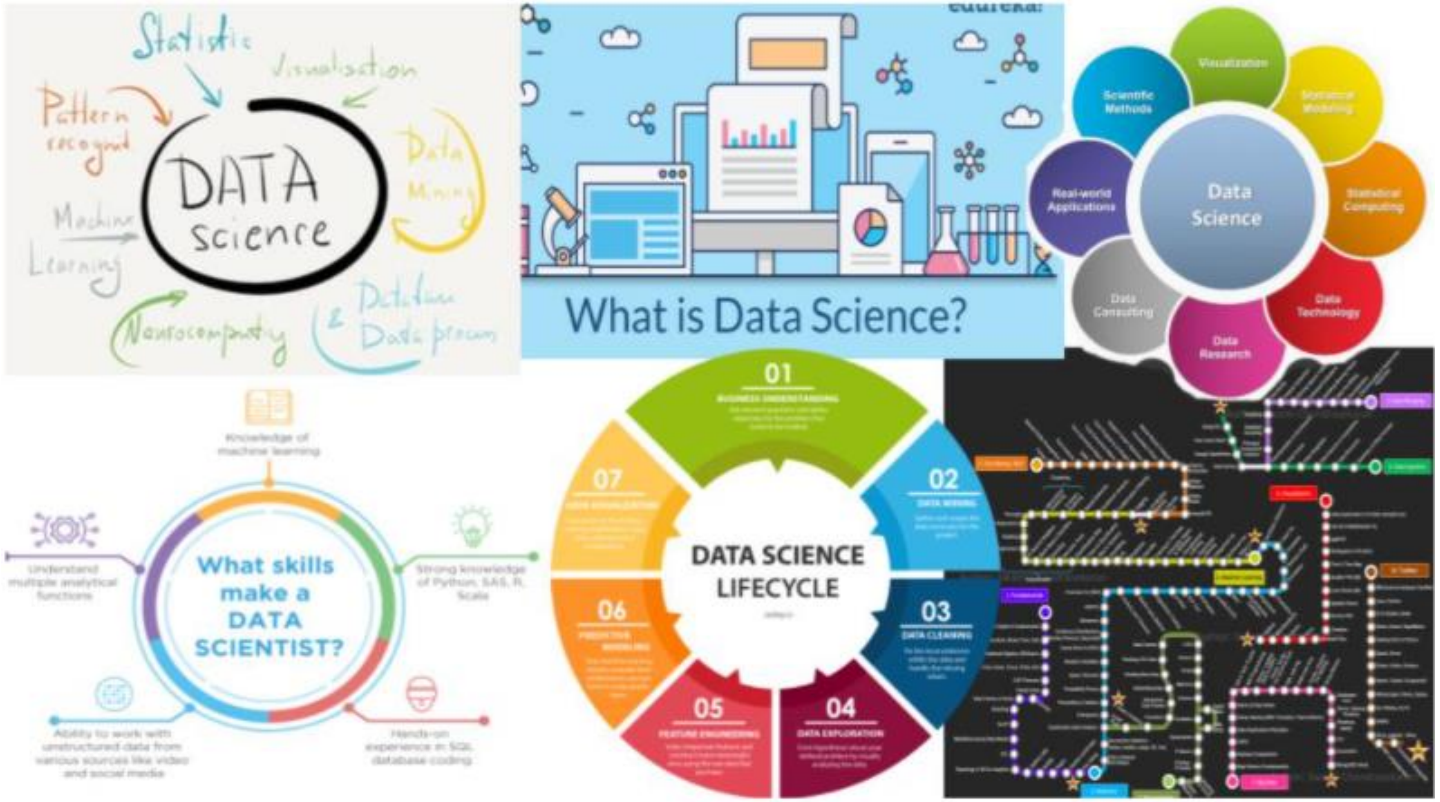
DATE ESAME	
5	giugno
19	giugno
10	luglio
11	settembre

LUN	15.00	17.00
GIOV	8.30	10.30
VEN	11.30	14.30

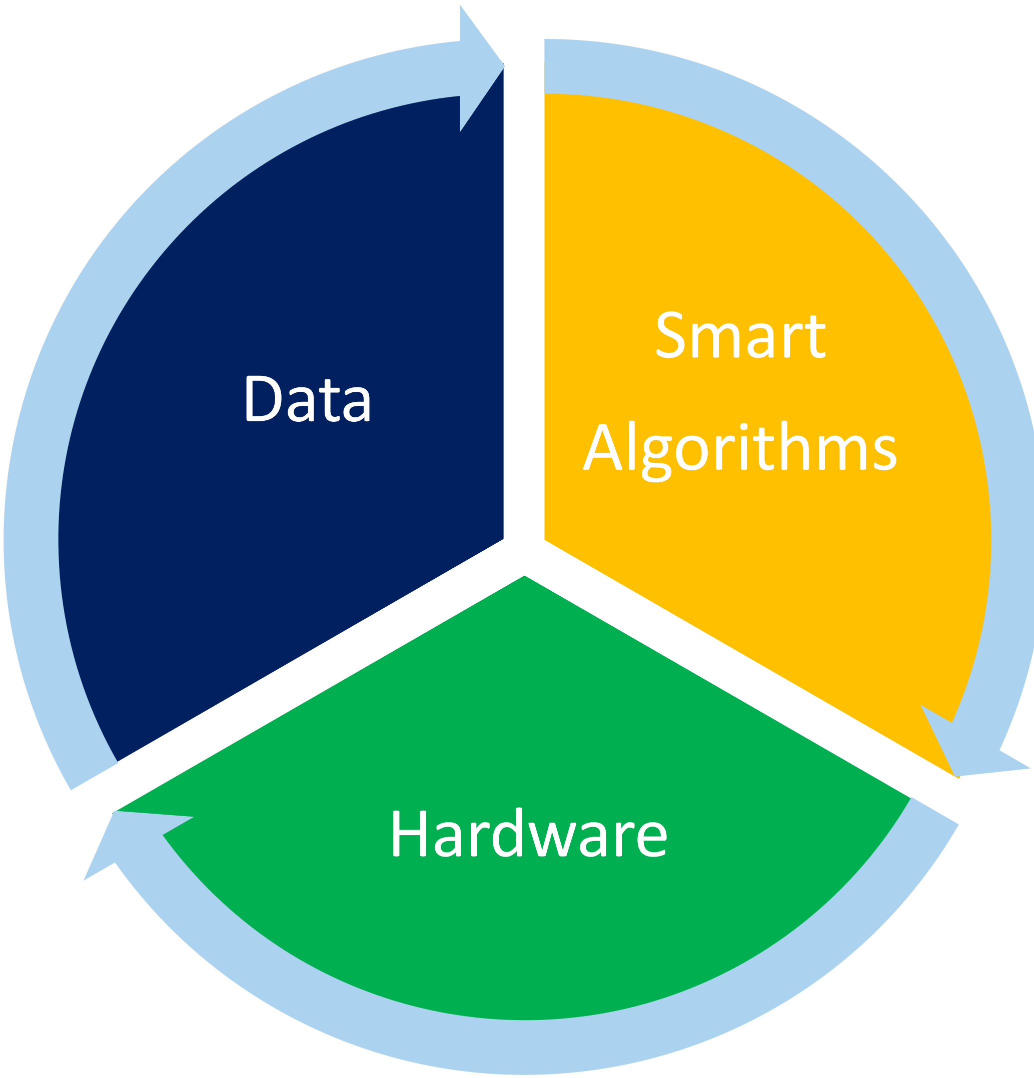
Libro di riferimento



What is Data Science? Ask google…



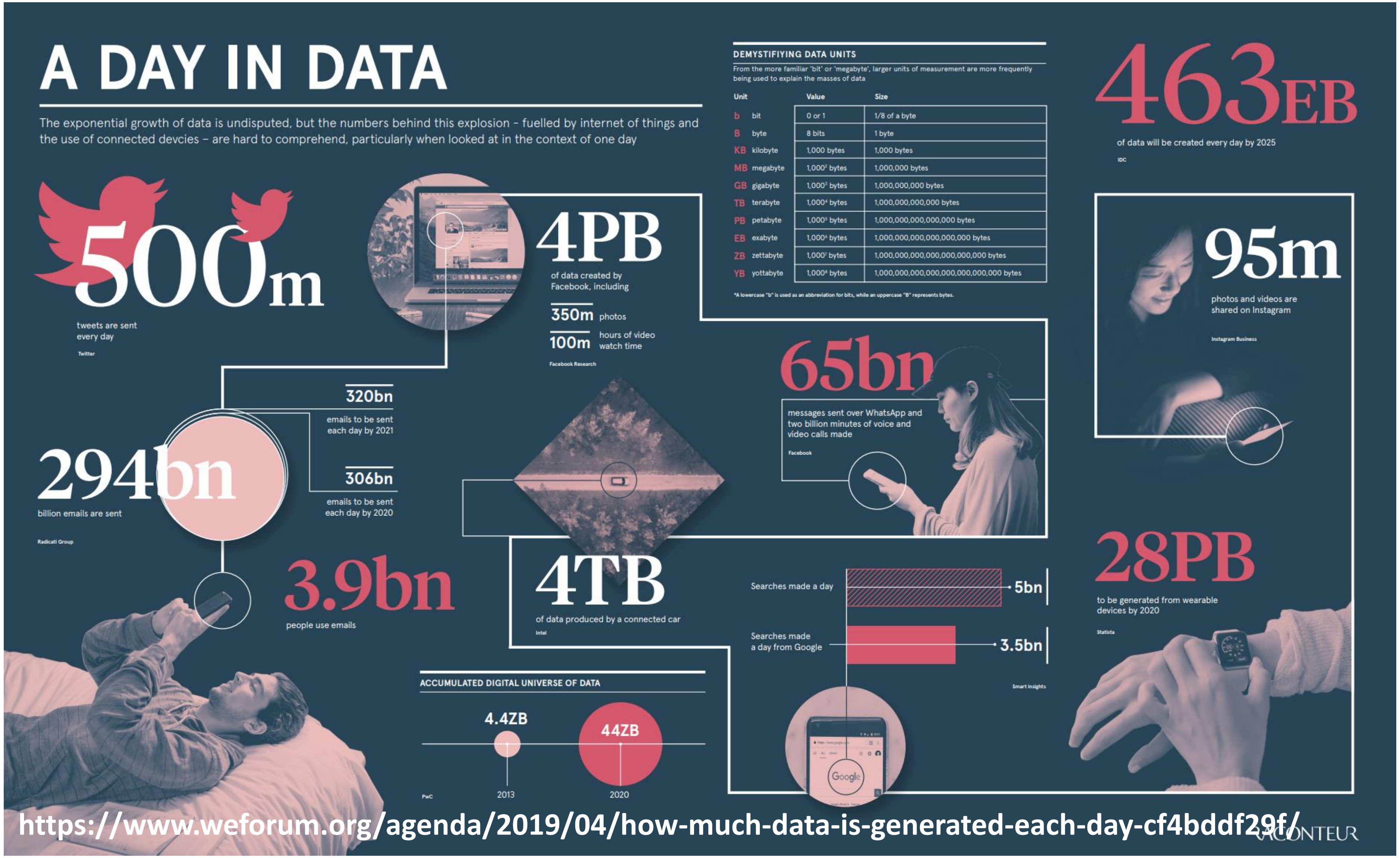
The Virtuous Circle of Machine Learning and AI



Making data for you

Use data to better describe the present or better predict the future

A day in data



https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

Hardware

Speed up calculations with 1000s of processors

NVIDIA TITAN V

NVIDIA's SUPERCOMPUTING GPU ARCHITECTURE, NOW FOR YOUR PC

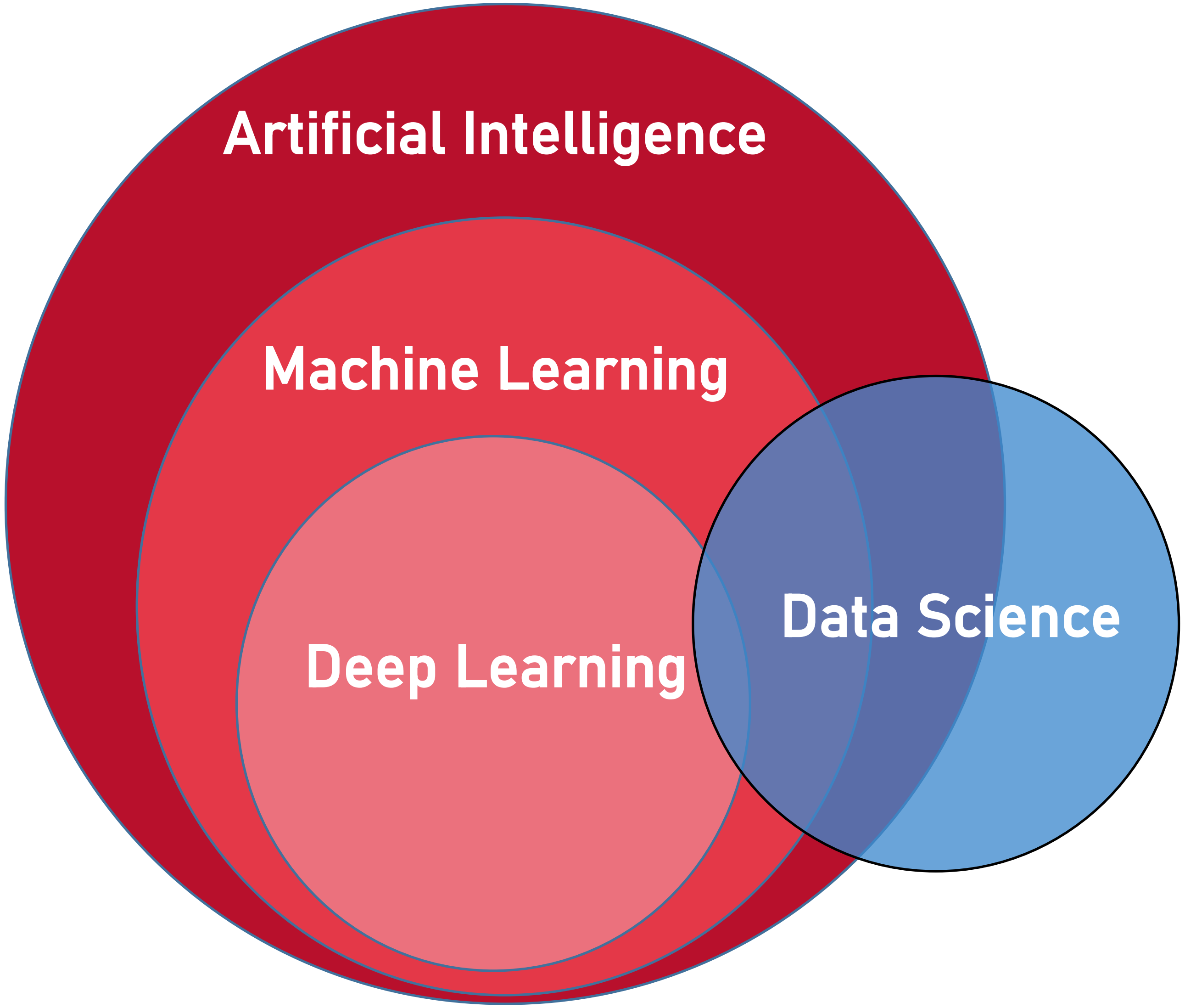
NVIDIA TITAN V is the most powerful Volta-based graphics card ever created for the PC. NVIDIA's supercomputing GPU architecture is now here for your PC, and fueling breakthroughs in every industry.

Scale computations with infinite compute power

What is Machine Learning

- Machine learning is a branch of AI
- The idea underlying machine learning is that we give a computer program access to lots of data and let it learn about relationships between variables and make predictions
- Some of the techniques of machine learning date back to the 1950s but improvements in computer speeds and data storage costs have now made machine learning a practical tool

Machine Learning, Deep Learning, Artificial Intelligence and Data Science



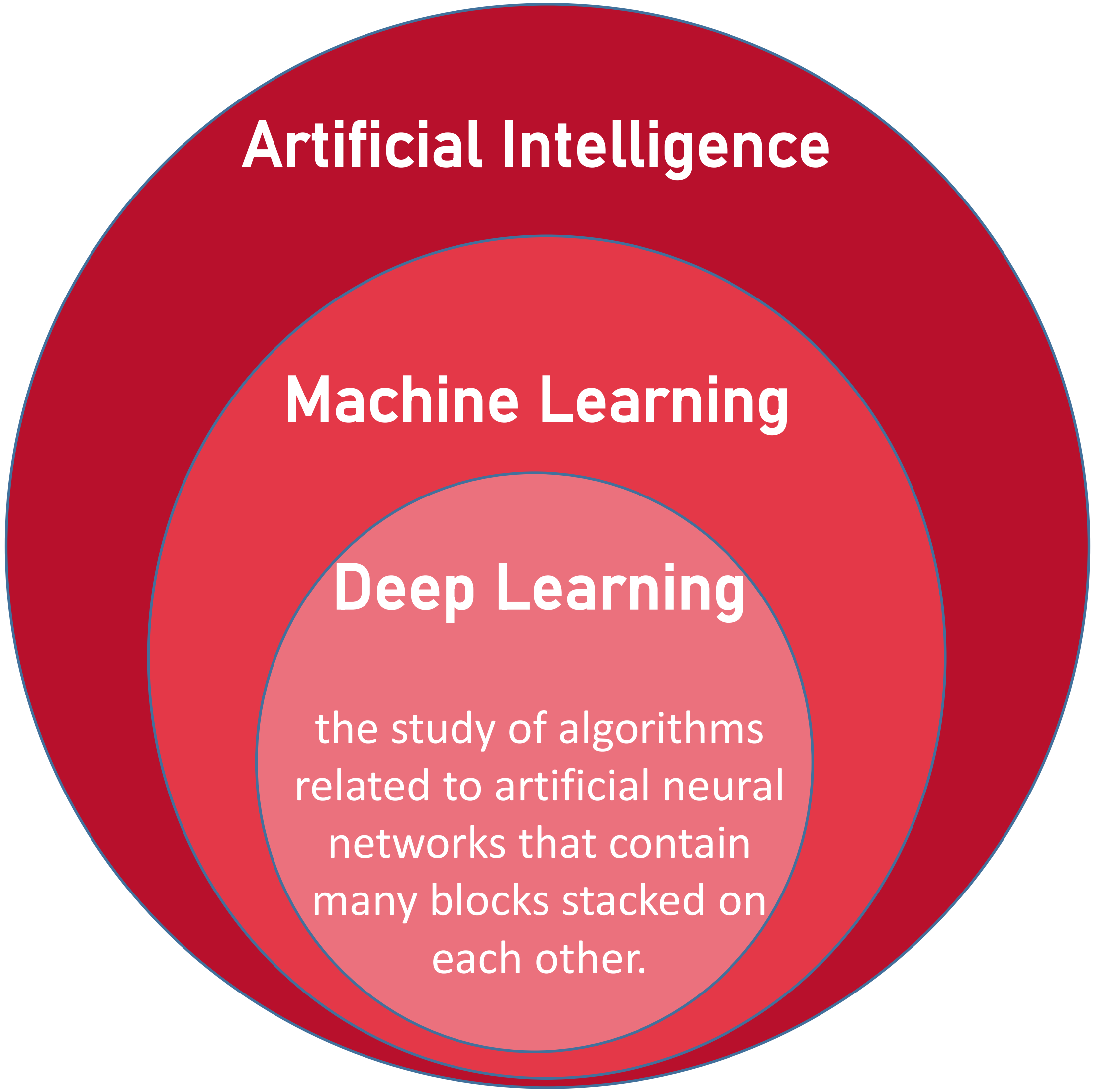
Artificial Intelligence

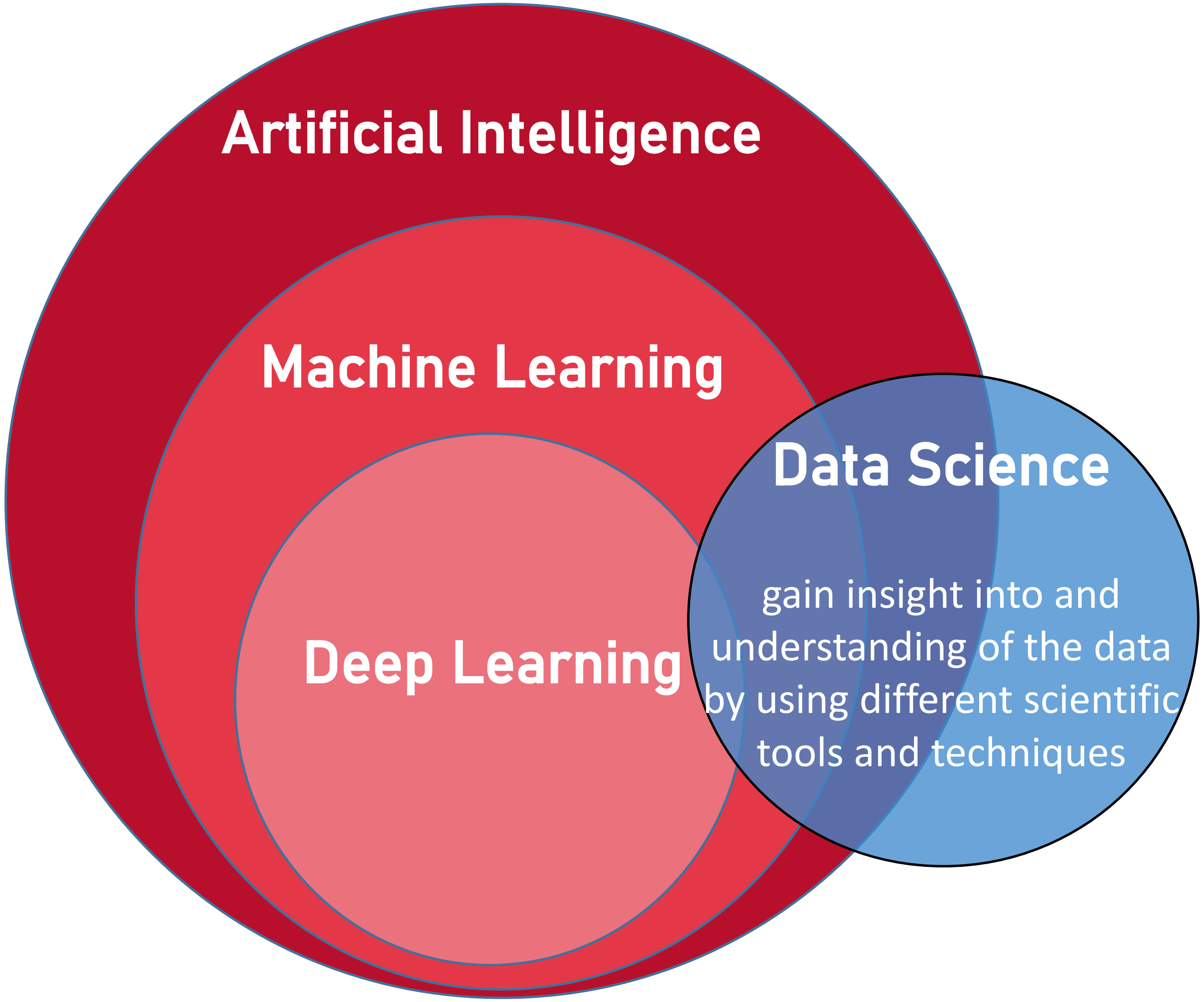
a branch of computer science dealing with the simulation of intelligent behavior in computers; (2) the capability of a machine to imitate intelligent human behavior.

Artificial Intelligence

Machine Learning

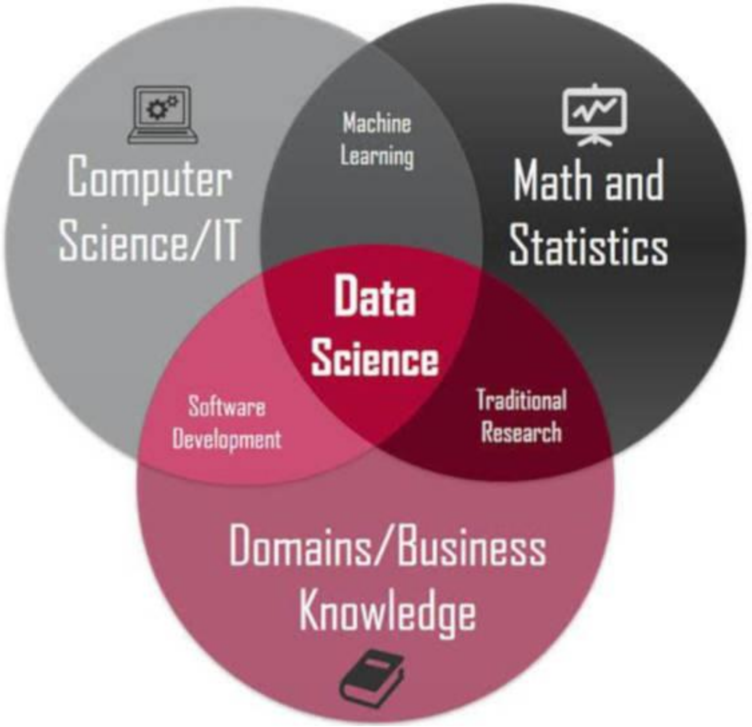
application of artificial intelligence that provides the AI system with the ability to automatically learn from the environment and apply those lessons to make better decisions.


Artificial Intelligence
Machine Learning
Deep Learning
the study of algorithms related to artificial neural networks that contain many blocks stacked on each other.


Artificial Intelligence
Machine Learning
Deep Learning
Data Science
gain insight into and understanding of the data by using different scientific tools and techniques

Data science skills

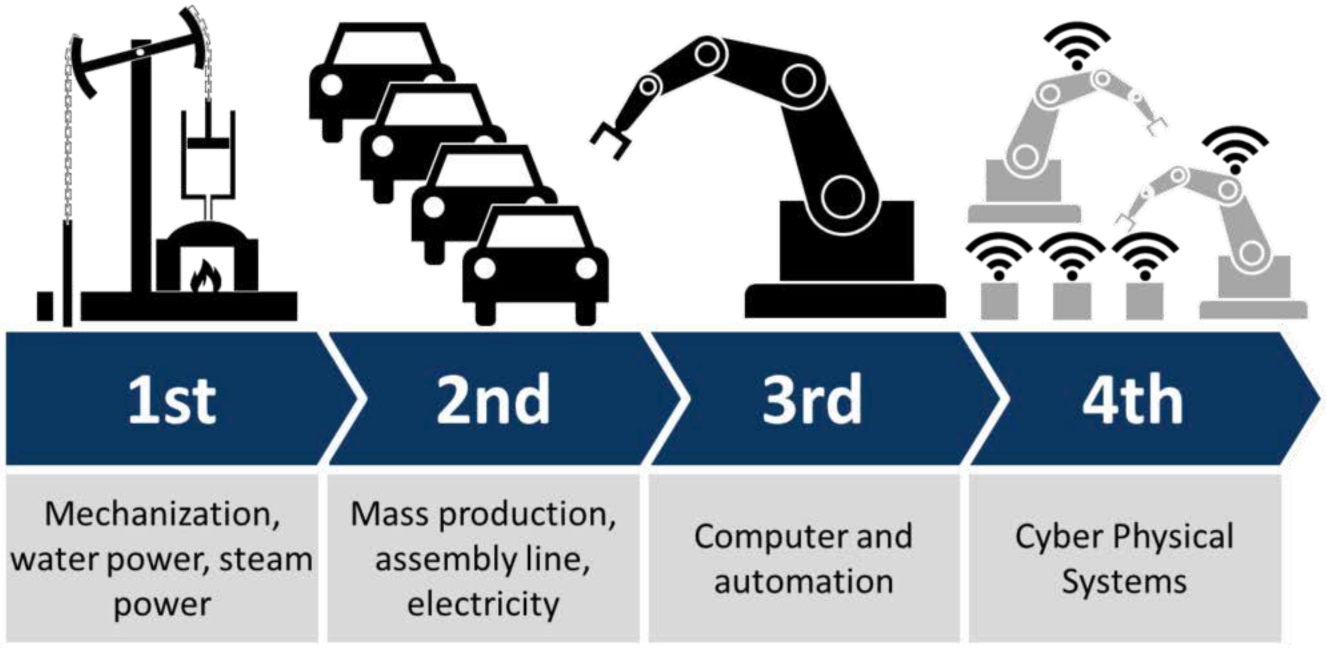
- **Technology**: manage data (structured and unstructured, big data)
- **Machine Learning**: Mathematics and Statistics
- **Programming**: open source Data Science programming suite (Python, R, ….)
- **Business knowledge:** transfer information got from data to Analysts
- **Communication**: data visualization tools



Machine Learning vs. Automation

- Computers have been used to automate many business decisions (payroll, sending out invoices, summarizing sales by region, etc)
- This is digitization: the third industrial revolution
- Machine learning is central to the fourth industrial revolution where computers are used to create intelligence

The 4th Industrial revolution is Here!



Source: Christoph Roser at AllAboutLean.com

As per Wikipedia*, "The 4th Industrial Revolution ….. marked by emerging technology breakthroughs in a number of fields, including robotics, **artificial intelligence**, nanotechnology, quantum computing, biotechnology, the Internet of Things, the Industrial Internet of Things (IIoT), decentralized consensus, fifth-generation wireless technologies (5G), additive manufacturing/3D printing and fully autonomous vehicles."

* https://en.wikipedia.org/wiki/Fourth_Industrial_Revolution

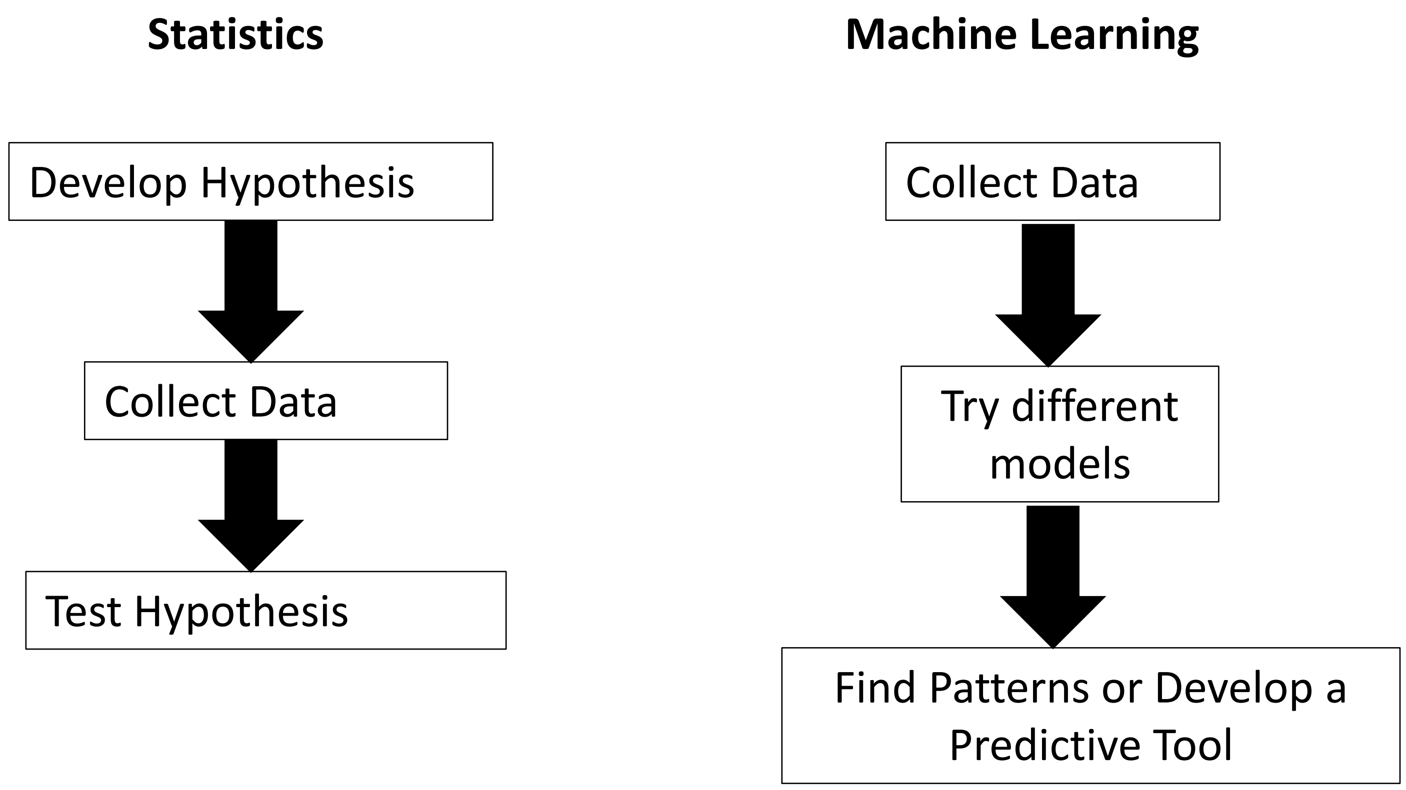
Example: Loan Applications (digitization vs. ML)

- If loan officers applied certain known rules we could digitize their activities
- If we did not know the rules used, we could use ML to determine them
- But we could go one step further and use ML to improve upon the rules for accepting or rejecting loans

The new world of statistics

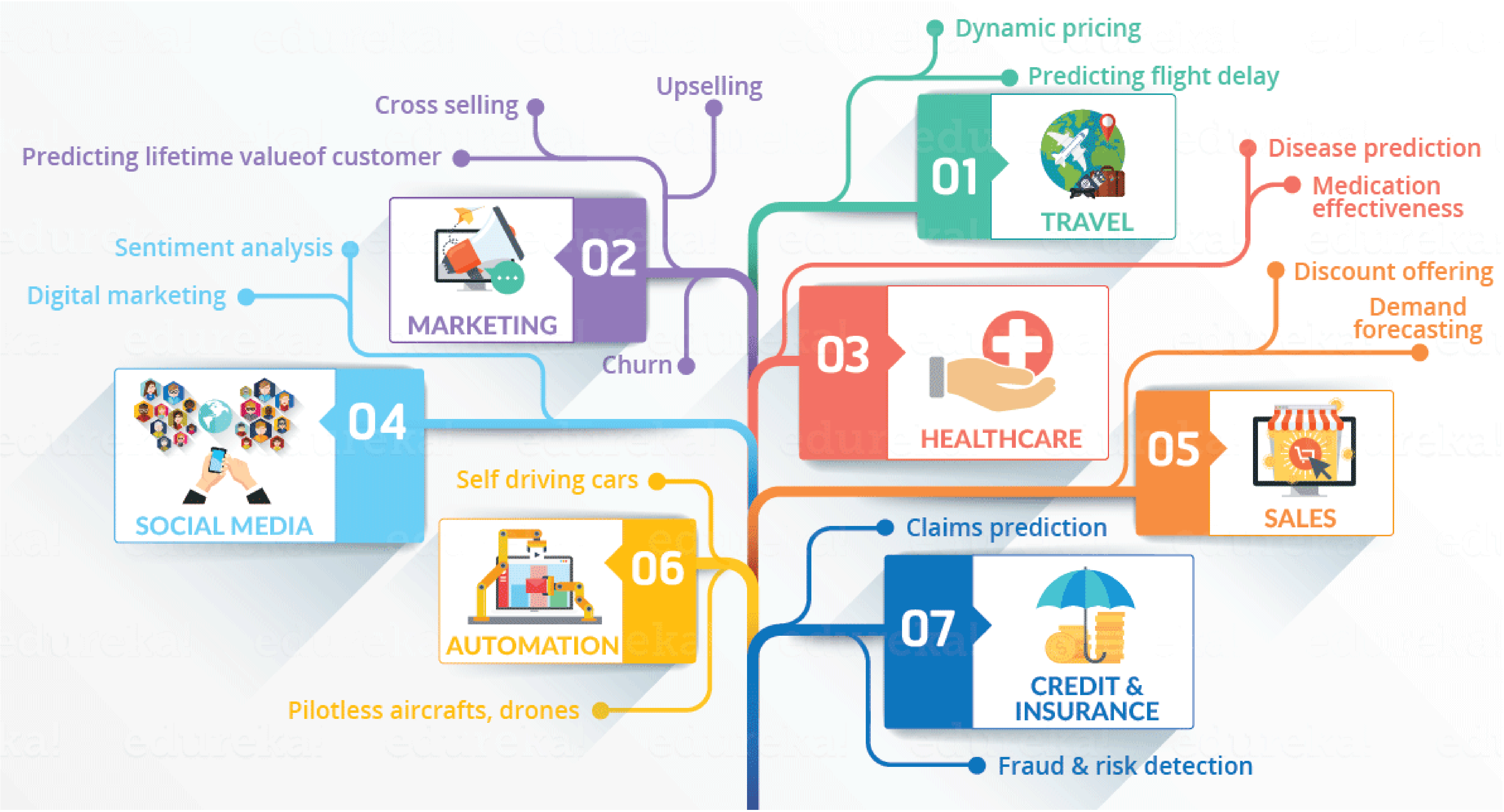
- Huge data sets
- Fantastic improvements in computer processing speeds and data storage costs
- Machine learning tools are now feasible
- Can now develop non-linear prediction models, find patterns in data in ways that were not possible before, and develop multi-stage decision strategies
- New terminology: features, labels, activation functions, target, bias, supervised/unsupervised learning……

Traditional Statistics vs Machine Learning (Figure 1.1)



What can data do?

- **Describe the current state of an organization or process**
- **Detect anomalous event**
- **Diagnose the causes of events and behaviors**
- **Predict future events**


Dynamic pricing
Predicting flight delay
01
TRAVEL
Cross selling
Upselling
Predicting lifetime valueof customer
02
MARKETING
Churn
Sentiment analysis
Digital marketing
04
SOCIAL MEDIA
Disease prediction
Medication effectiveness
03
HEALTHCARE
Discount offering
Demand forecasting
05
SALES
Self driving cars
06
AUTOMATION
Pilotless aircrafts, drones
Claims prediction
07
CREDIT & INSURANCE
Fraud & risk detection

Average Years of Experience

0 - 1	8%
2 - 4	80%
5 - 7	6%
8+	6%

Common Skill Sets

- Machine Learning
- Statistics
- Python
- Natural Language Processing
- Hadoop SPARK
- Algorithms
- SQL
- Programming Languages

Data Scientist Salaries

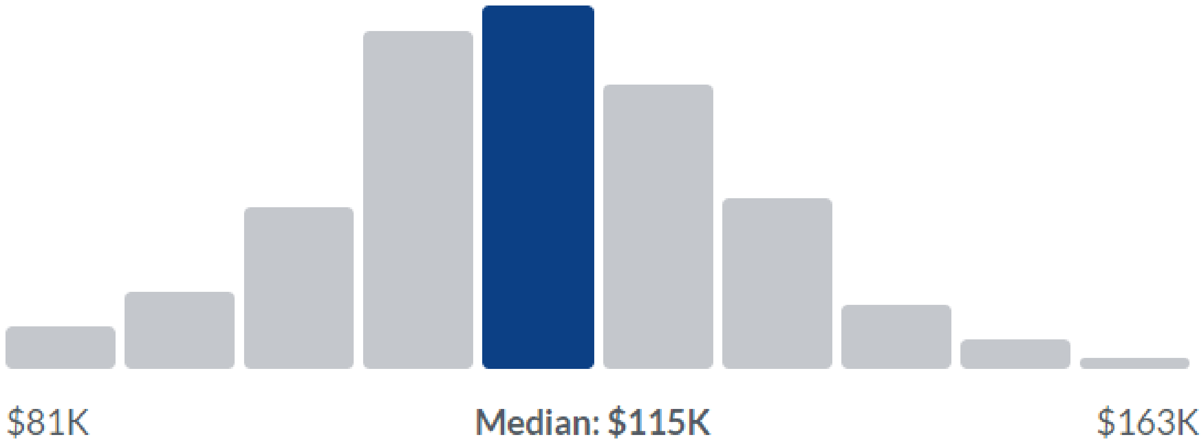
Average Base Pay

$114,673 /yr

Same as national average

Not including cash compensation

See More Insights



The Data scientist

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

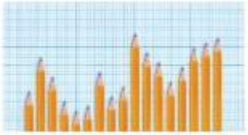
Data Scientist: The Sexiest Job of the 21st Century

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

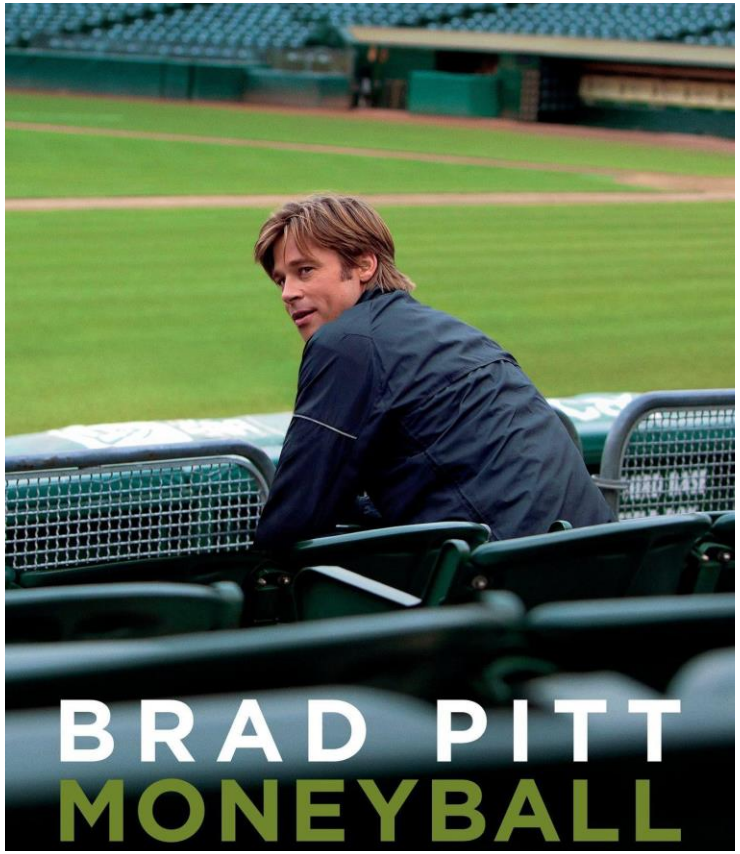
VIEW MORE FROM THE

October 2012 Issue

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

Data science in sports

Data Science in Sports

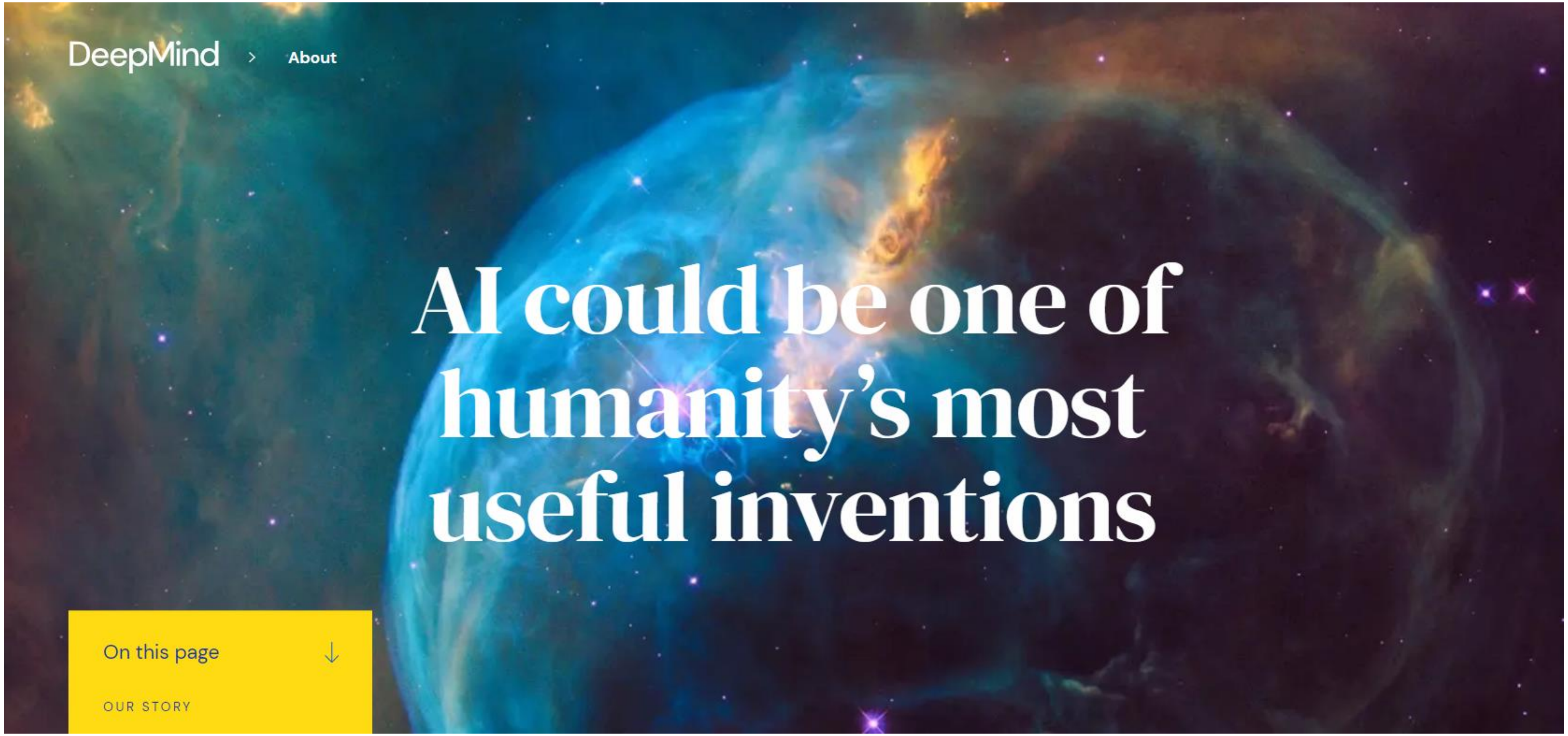


The market for baseball players was so inefficient
…
that superior management could run circles
around taller piles of cash
- Michael Lewis

Legendary 2002 season for Oakland Athletics.

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

Data Science in Drug Discovery



Data Science in Drug Discovery

nature
International journal of science

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

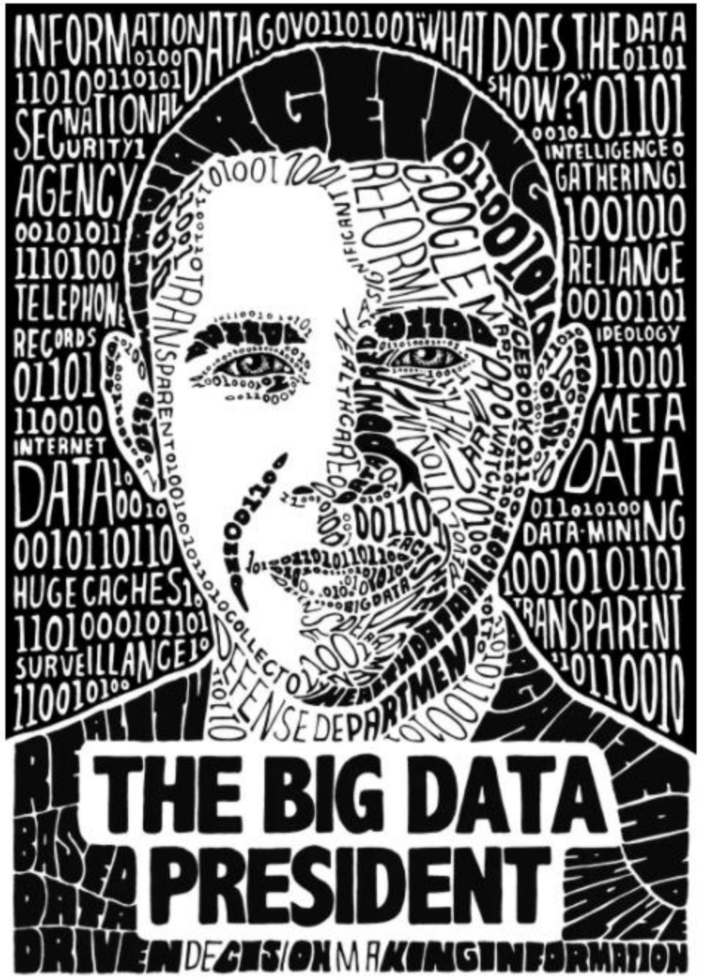
2.2k Shares | in 1.1k | 372 | 354 | 198

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

Data science for politics

The Washington Post
Democracy Dies in Darkness



Opinions Editorial Board The Opinions Essay Global Opinions Post Opinión First 100 Days Reimag

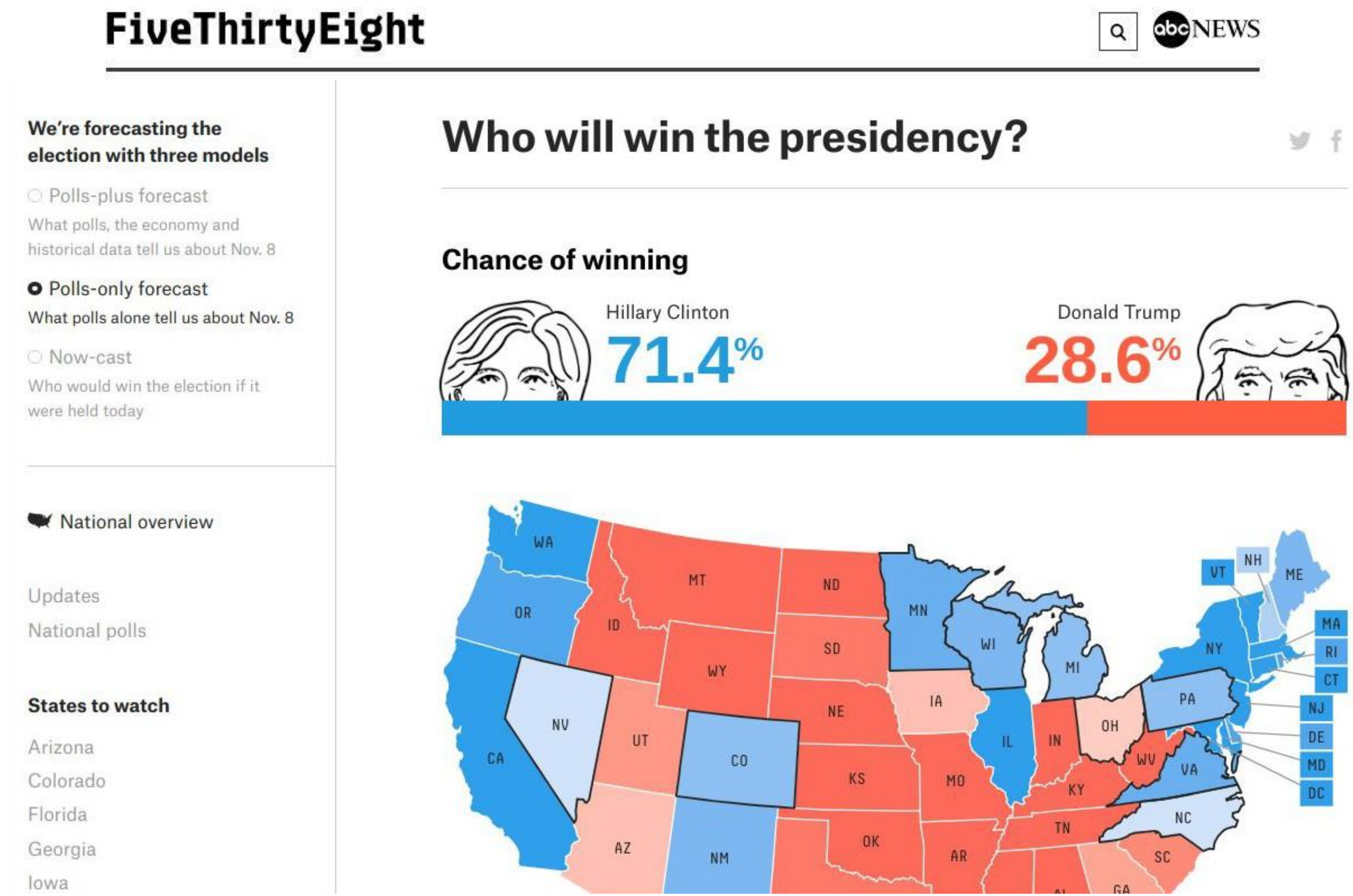
Opinions

Obama, the 'big data' president

By **Nancy Scola**
June 14, 2013

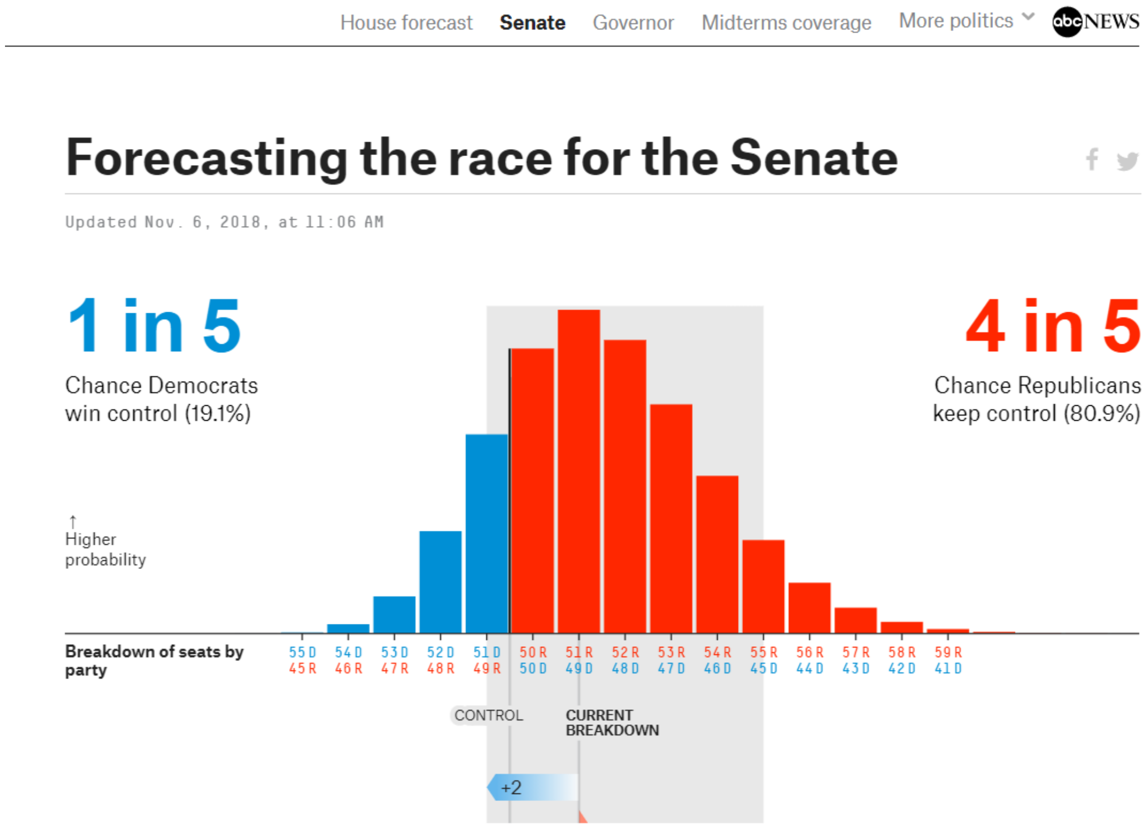
Nancy Scola is a journalist covering technology and politics. From 2001 to 2005, she served on the staff of the House government oversight committee.

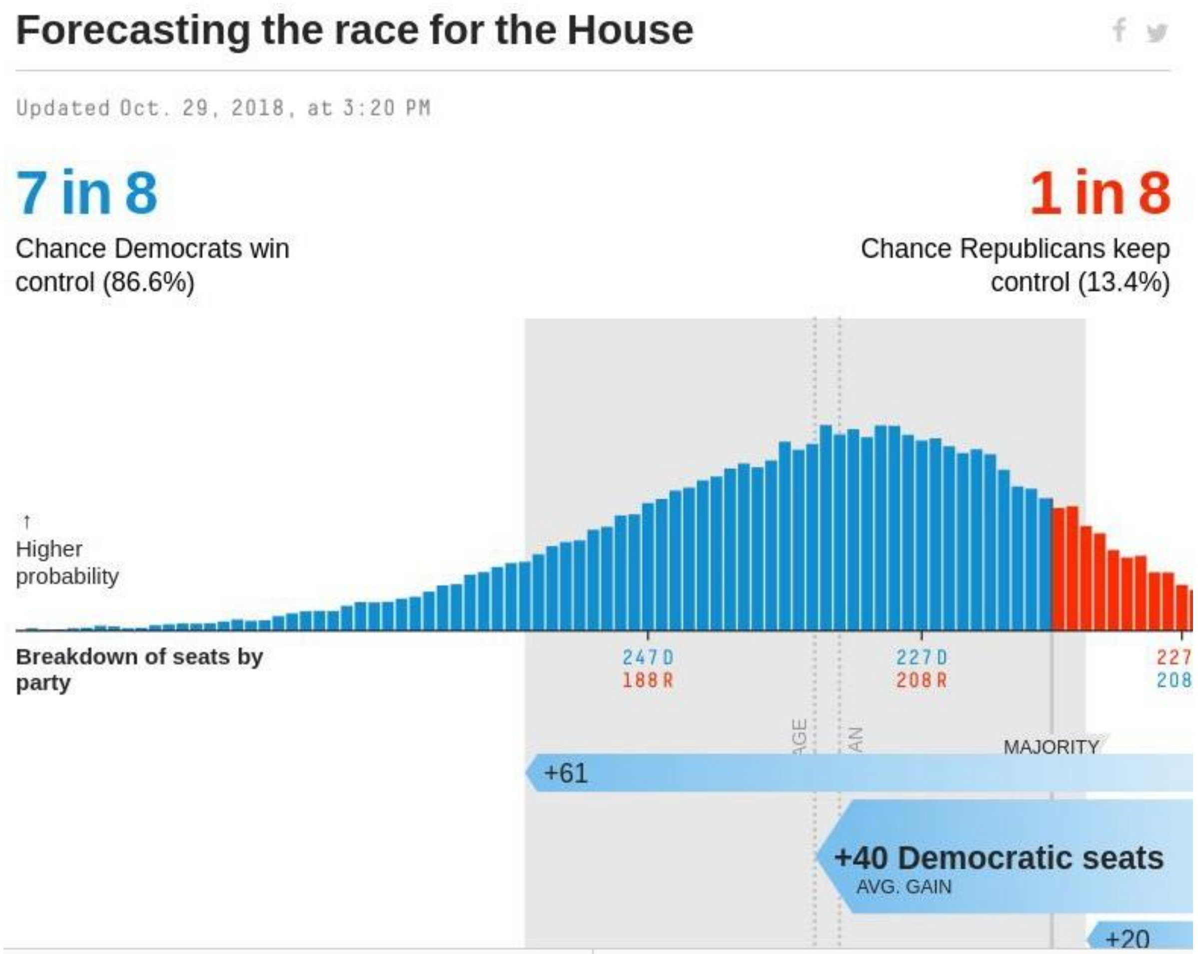
n the political world, the promise of data — whether it's Nate Silver's spot-on election predictions or President Obama's clearinghouse of government information, Data.gov — is that we no longer have to take so much on faith. "What do the data show?" is the new "What do you think?," the new "Is this a good idea?"



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

Data Science in Politics





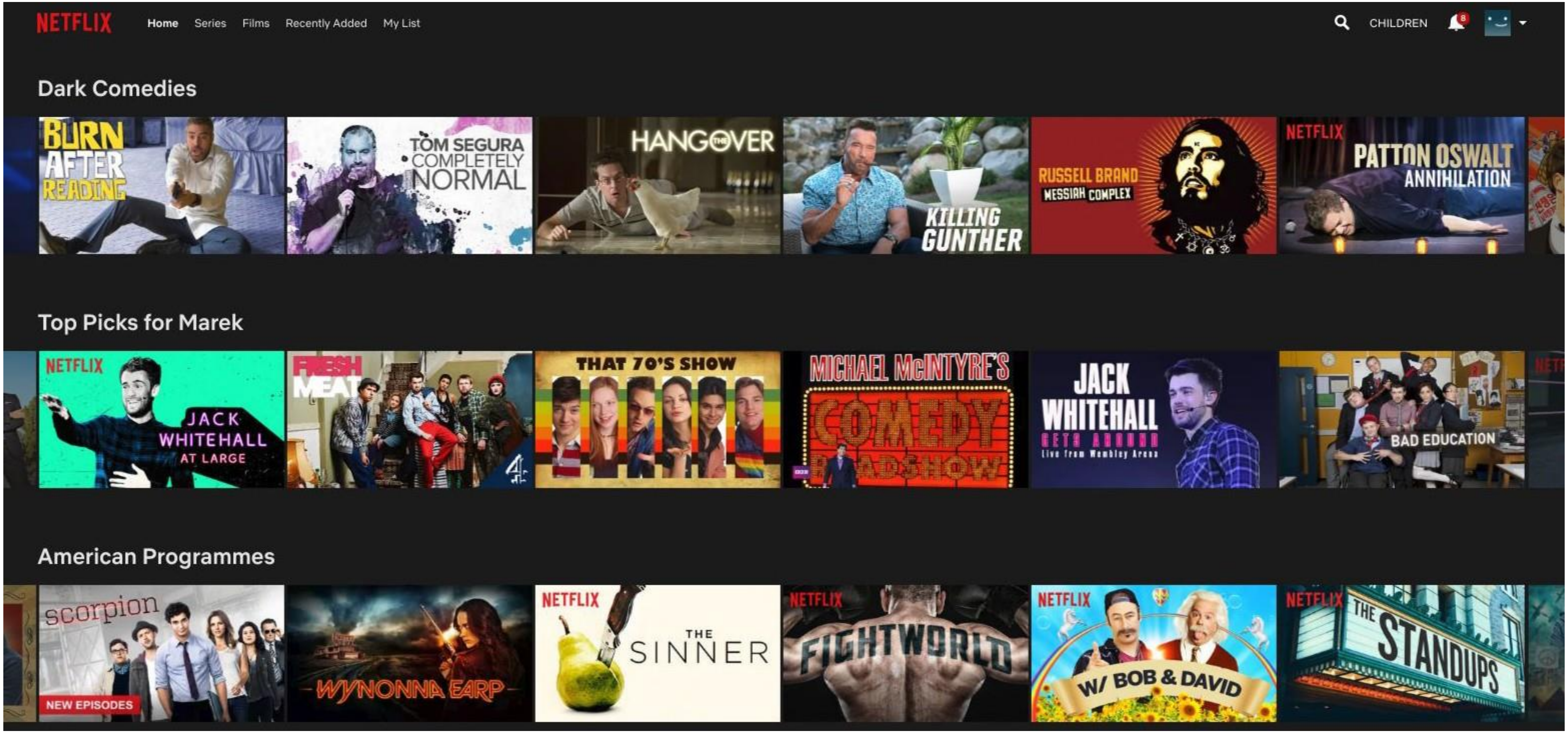
https://fivethirtyeight.com/tag/2018-election/

Data Science in Commerce

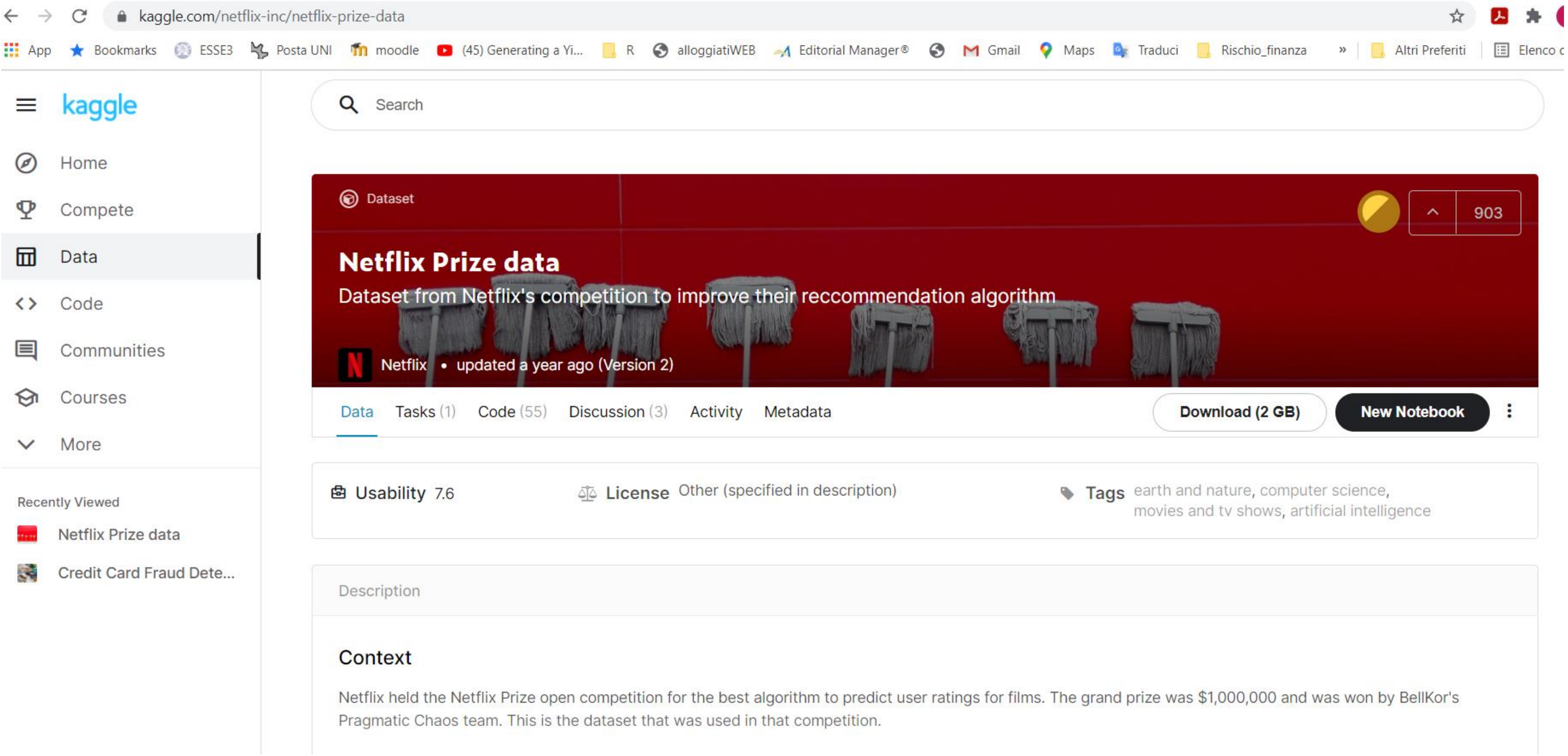
amazon

ebay

Data Science in Commerce



Netfix challenge



ML and AI is revolutionizing finance

Artificial intelligence and robotics + Add to myFT

BlackRock bulks up research into a[rtificial] intelligence

INVESTMENT BANKING

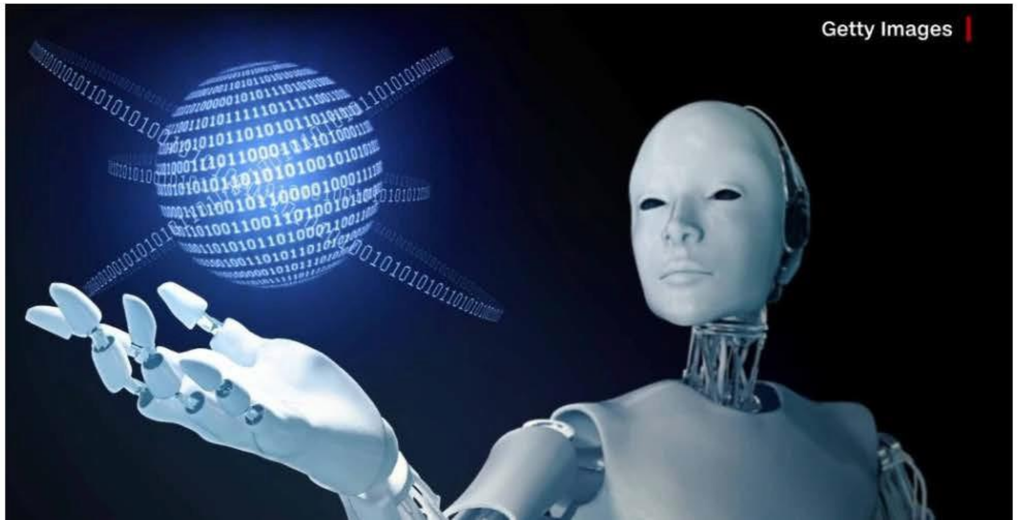
Goldman Sachs hunts AI experts for all-important quant team

US bank is building its vast strats department by hiring a new generation of machine learning and artificial intelligence specialists

JPMorgan's latest hire proves the bank is serious about artificial intelligence

by Julia Horowitz @juliakhorowitz

May 3, 2018: 7:46 PM ET



Social Surge - What's Trending

Ivanka Trump and Jared Kushner detail vast wealth: Real estate, fashion and investments

AT&T-Time Warner ruling: The media industry hangs in the balance

CNN anchors reflect on the life of Anthony Bourdain

Market impact at the speed of light!

Catherine Kang • 2nd
Event-Driven Feeds | Bloomberg for Enterprise
2d • Edited

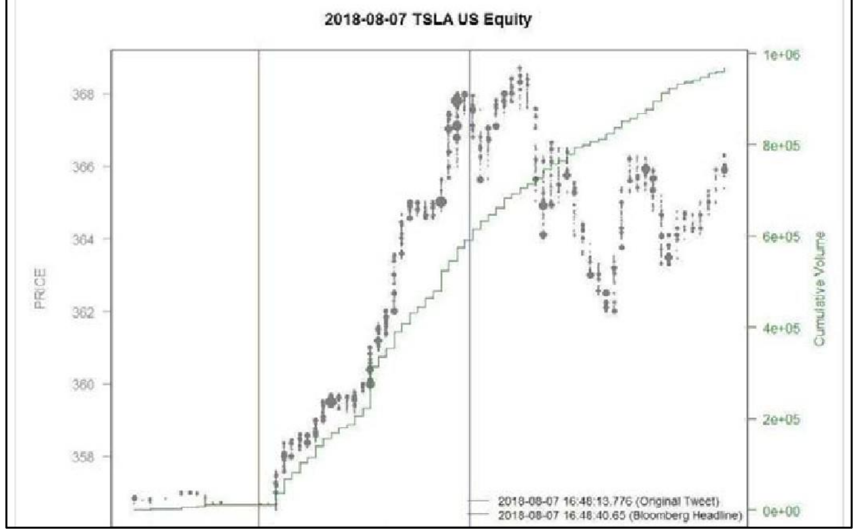
Elon Musk announced on Twitter today that he is considering taking Tesla private at $420/share.

Original tweet on Bloomberg Event-Driven Feeds:
Elon Musk: Am considering taking Tesla private at $420. Funding secured.
08/07/2018
12:48:13.776 ET

Bloomberg newsroom headline:
*MUSK: AM CONSIDERING TAKING TESLA PRIVATE AT $420
08/07/2018
12:48:40.650 ET

Please message me to learn more about how our Bloomberg newsroom verified and curated Twitter feed can benefit your trading.

Market impact:



Machine Learning & AI in finance: A paradigm shift

Quant	Data Scientist
Stochastic Models	Real-time analytics
Factor Models	Predictive analytics
Optimization	Machine Learning
Risk Factors	RPA
P/Q Quants	NLP
Derivative pricing	Deep Learning
Trading Strategies	Computer Vision
Simulations	Graph Analytics
Distribution fitting	Chatbots
	Sentiment Analysis
	Alternative Data

Technology drives finance!

«Financial technologies of «fintech» is used to describe a variety of

innovative business models

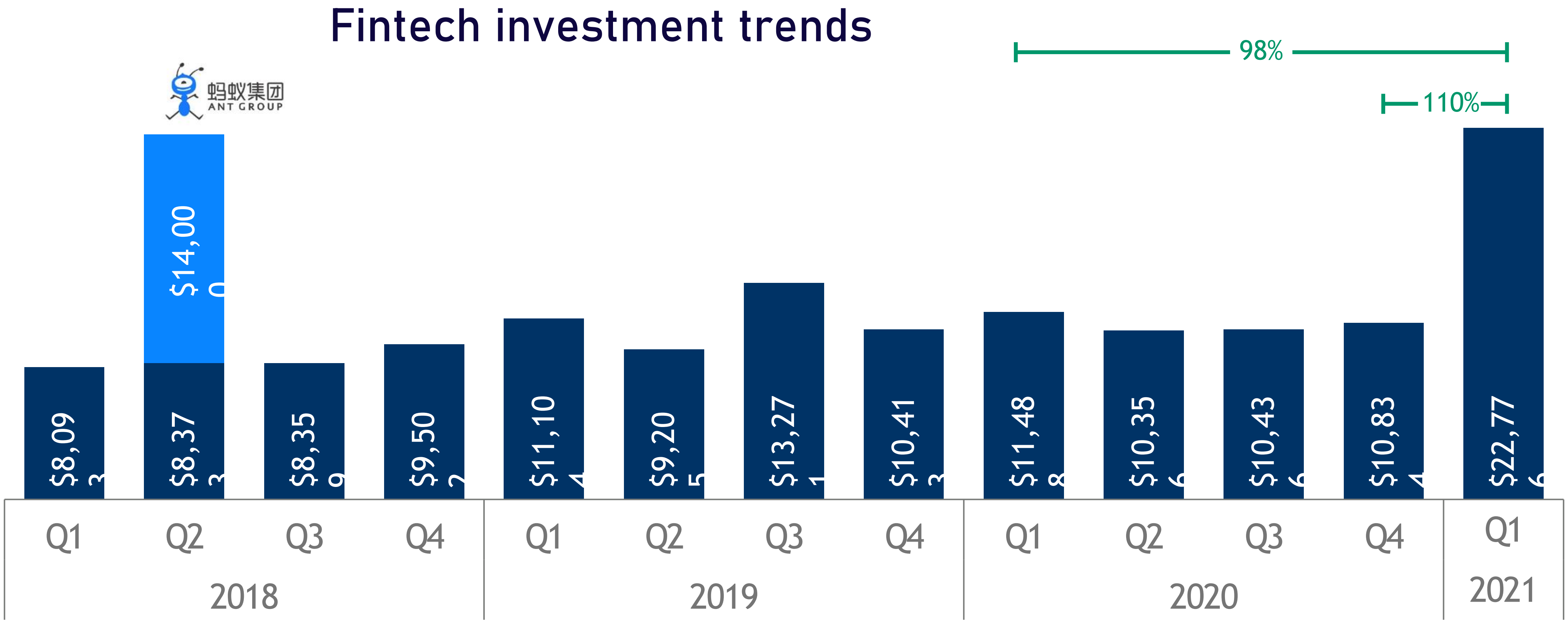
and

emerging technologies

that have the potential to transform the financial service industry»

https://www.iosco.org/library/pubdocs/pdf/IOSCOPD554.pdf

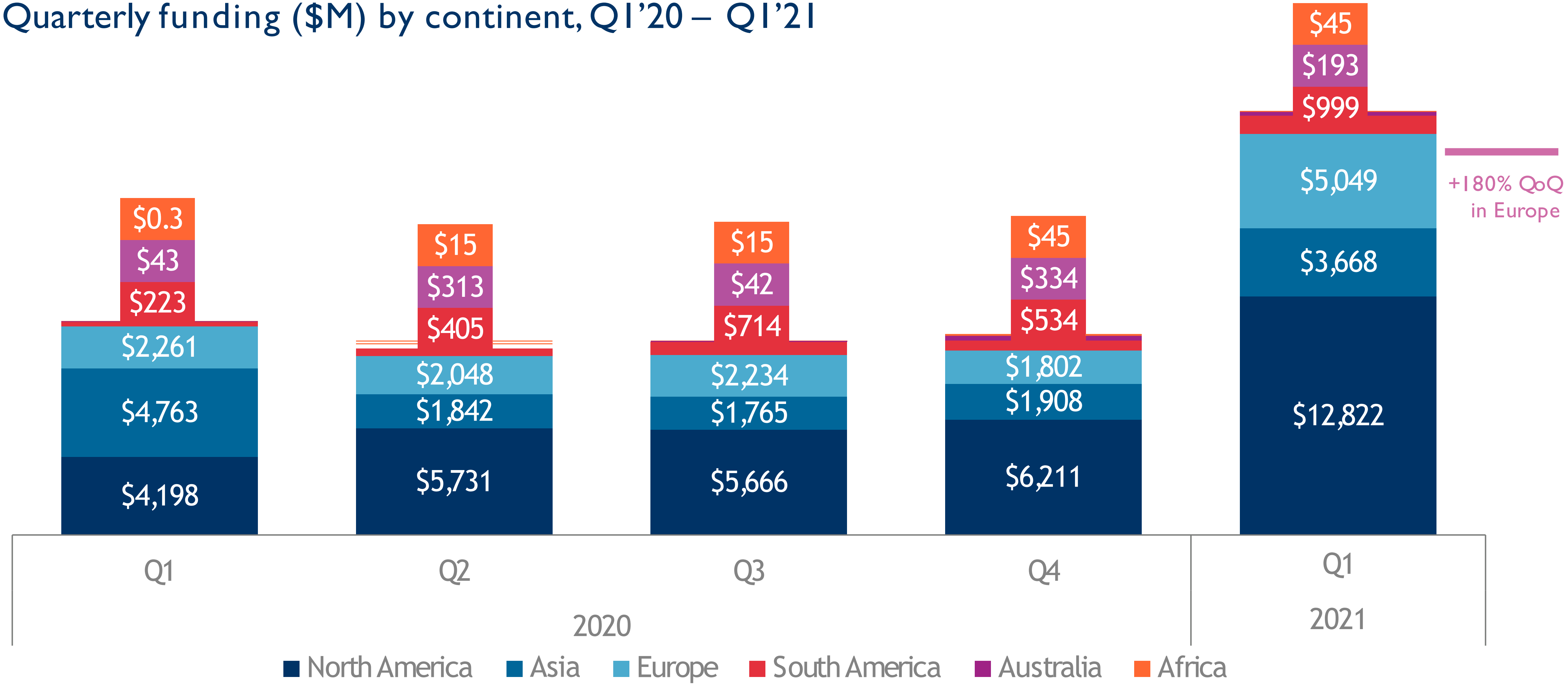
Fintech funding more than doubled QoQ
Global VC-backed fintech funding trends, Q1'18 – Q1'21



Europe saw the largest QoQ increase in funding

FINTECH INVESTMENT TRENDS

Quarterly funding ($M) by continent, Q1'20 – Q1'21



What the state of fintech covers

PAYMENTS

Payments processing, card developers, money transfer platforms, and tracking software

BANKING

Digital-first banks or companies digitizing banking services for credit and debit

DIGITAL LENDING

Companies creating new solutions for personal or commercial lending

WEALTH MANAGEMENT

Personal finance tools, investment and wealth management platforms, and analytics tools

INSURANCE

Companies selling or distributing insurance digitally or providing data analytics and software for (re)insurers

CAPITAL MARKETS

Sales and trading, analysis, and infrastructure tools for financial institutions

SMB

Companies focused on providing solutions to small- and medium-sized businesses

REAL ESTATE

Mortgage lending, transaction digitization, and financing platforms

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

Algorithmic trading (or simply algo trading) is the use of algorithms to conduct trades autonomously.

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

Robo-advisors, algorithms built to calibrate a financial portfolio to the goals and risk tolerance of the user. Additionally, they provide automated financial guidance and service to end investors and clients.

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

Fraud is a massive problem for financial institutions and one of the foremost reasons to leverage machine learning in finance.

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

Underwriting could be described as a perfect job for machine learning in finance, and indeed there is a great deal of worry in the industry that machines will replace a large swath of underwriting positions that exist today.

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

Automation is patently well suited to finance. It reduces the strain that repetitive, low-value tasks put on human employees. It tackles the routine, everyday processes, freeing up teams to finish their high-value work.

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

All aspects of understanding and controlling risk are being revolutionized through the growth of solutions driven by machine learning

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

Asset price prediction is considered the most frequently discussed and most sophisticated area in finance.

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

The classic derivative pricing models are built on several impractical assumptions to reproduce the empirical relationship between the underlying input data (strike price, time to maturity, option type) and the price of the derivatives observed in the market.

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

Sentiment analysis involves the perusal of enormous volumes of unstructured data, such as videos, transcriptions, photos, audio files, social media posts, articles, and business documents, to determine market sentiment.

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

Trade settlement is the process of transferring securities into the account of a buyer and cash into the seller's account following a transaction of a financial asset.

Current and Future Applications in Finance

- Algorithmic Trading
- Portfolio Management and Robo-Advisors
- Fraud Detection
- Loans/Credit Card/Insurance Underwriting
- Automation and Chatbots
- Risk Management
- Asset Price Prediction
- Derivative Pricing
- Sentiment Analysis
- Trade Settlement
- Money Laundering

A United Nations report estimates that the amount of money laundered worldwide per year is 2%–5% of global GDP

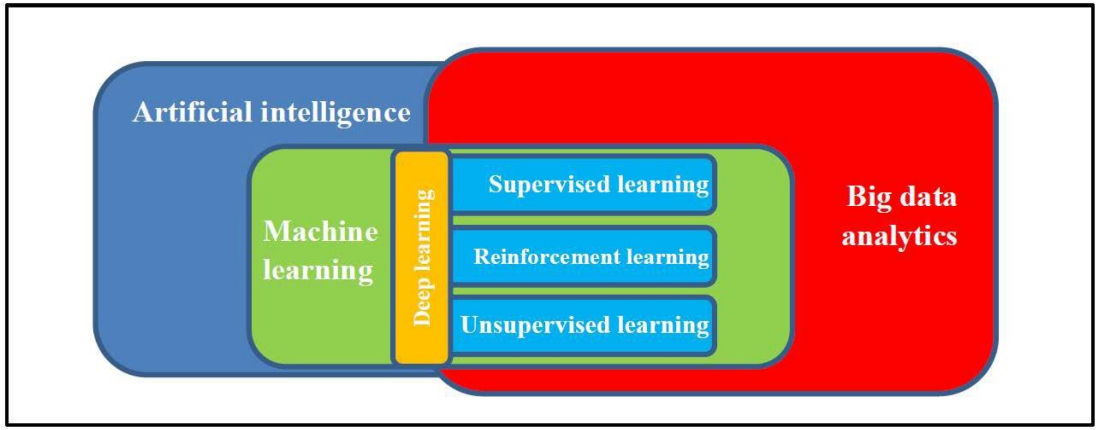
An intuitive introduction to AI and ML

Definitions: Machine Learning and AI

- Machine learning is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead[1]

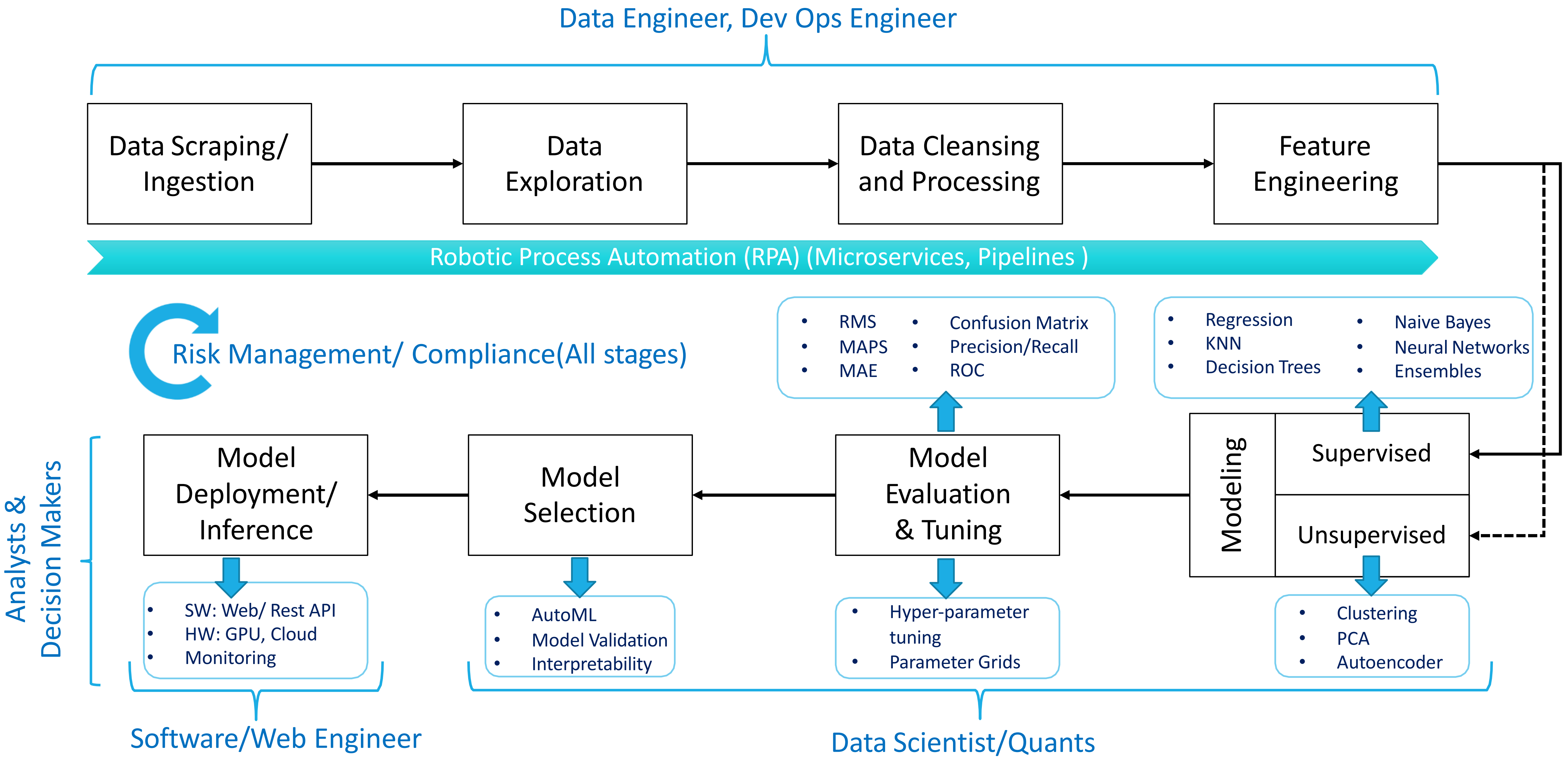
- Artificial intelligence is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and animals[1]

Figure 1: A schematic view of AI, machine learning and big data analytics



1. https://en.wikipedia.org/wiki/Machine_learning
2. Figure Source: http://www.fsb.org/wp-content/uploads/P011117.pdf

Machine Learning Workflow



Key steps involved

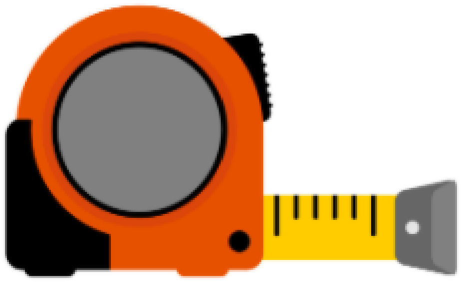
1. Data
2. Goals
3. Machine learning algorithms
4. Process
5. Performance evaluation

Data

Quantitative vs qualitative data

Quantitative data

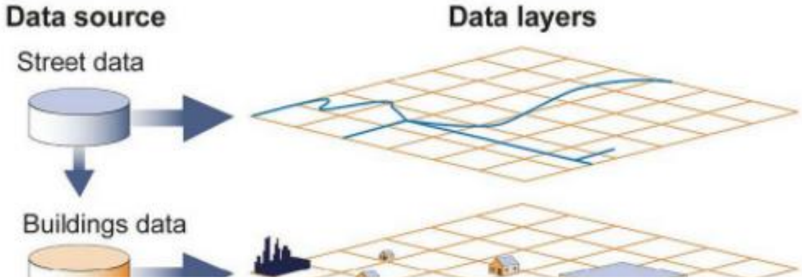
- Deals with numbers
- Data can be measured



Qualitative data

- Deals with descriptions
- Data can be observed but not measured

Quantitative vs qualitative data

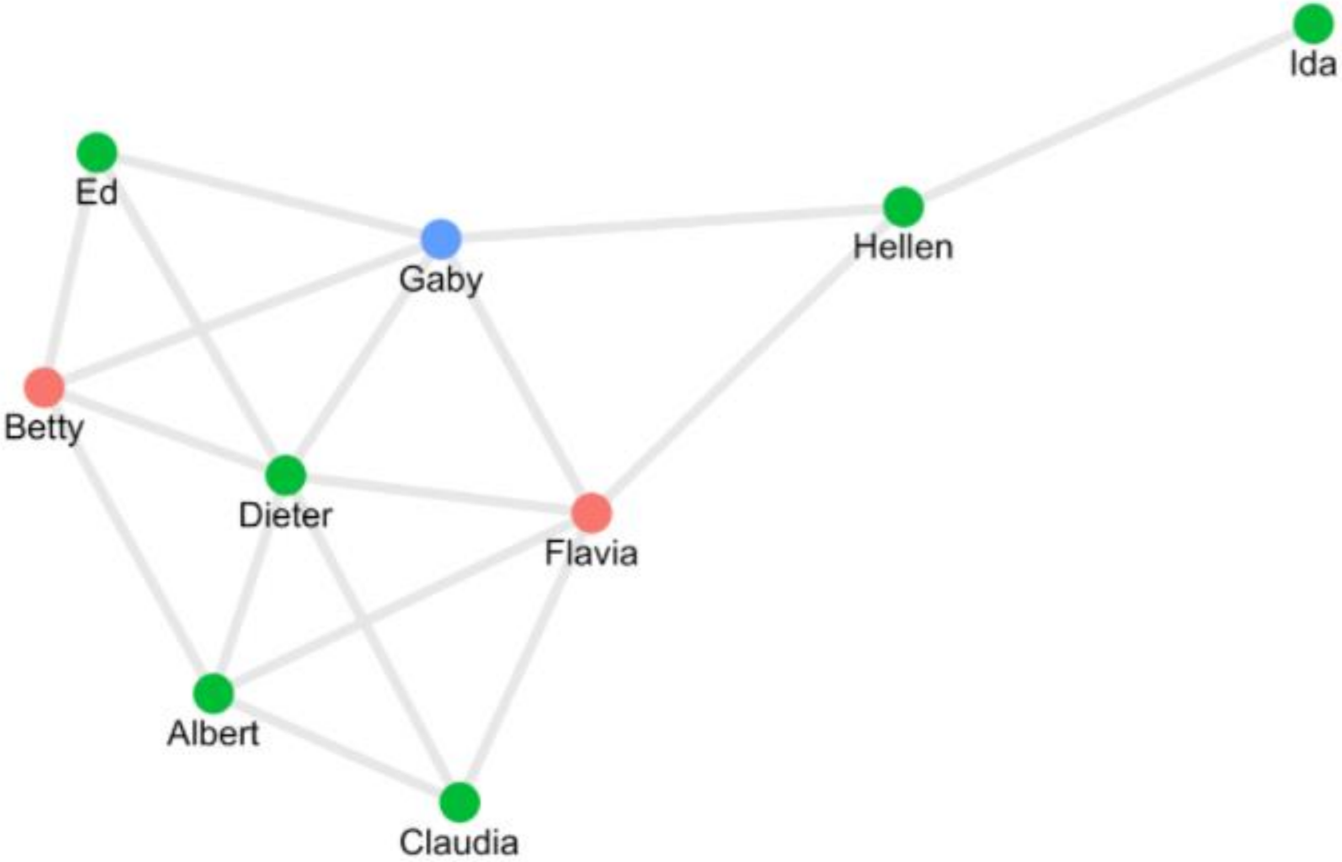


“Great evening, extremely good value”

Review of L'Ange 20 Restaurant

I went to this place with my boyfriend for a special occas
were greeted warmly by Christopher who guided us thro
delicious and I only wish that we could have had room f
excellent compared to other prices we had seen and we
hard to match during the rest of our stay.

I had the lamb which I can highly recommend. When we

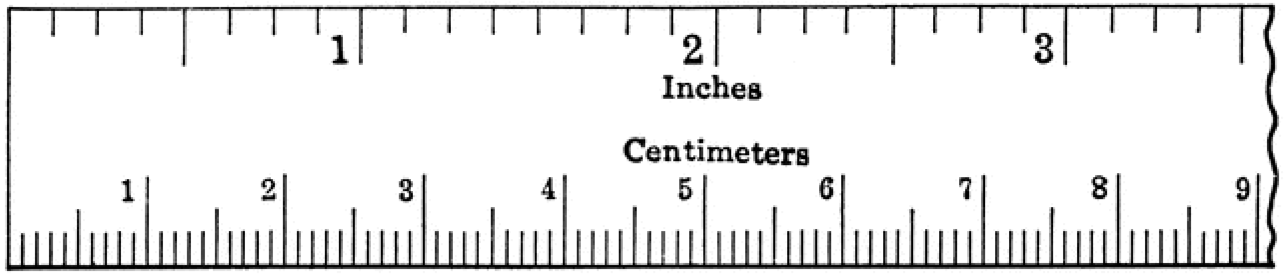


Variables

A variable could be:

- **Categorical**
  - Yes/No flags
  - AAA,BB ratings for bonds

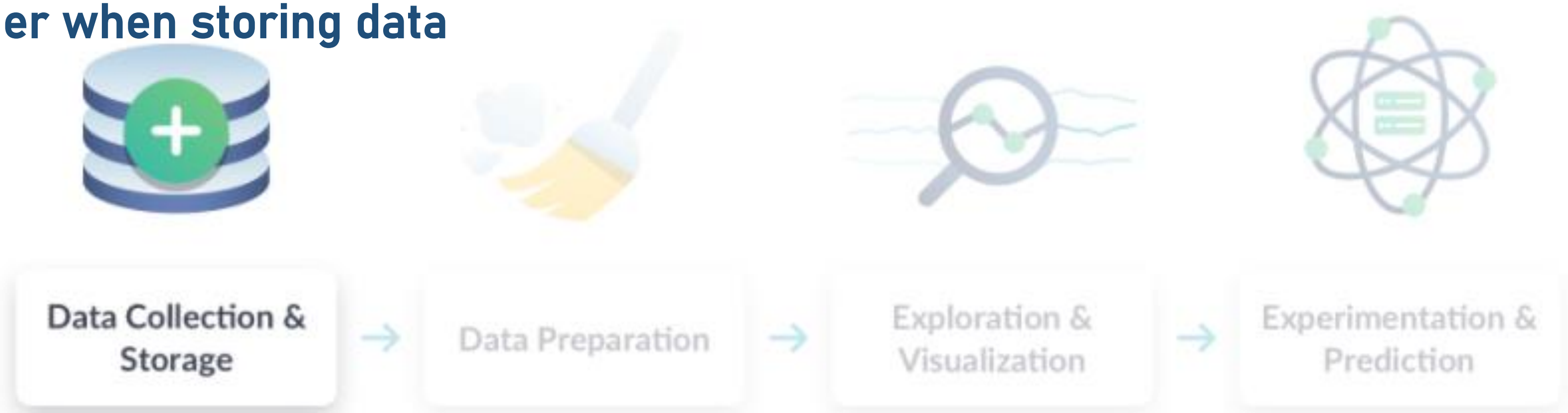
- **Numerical**
  - 35 mpg
  - $170K salary

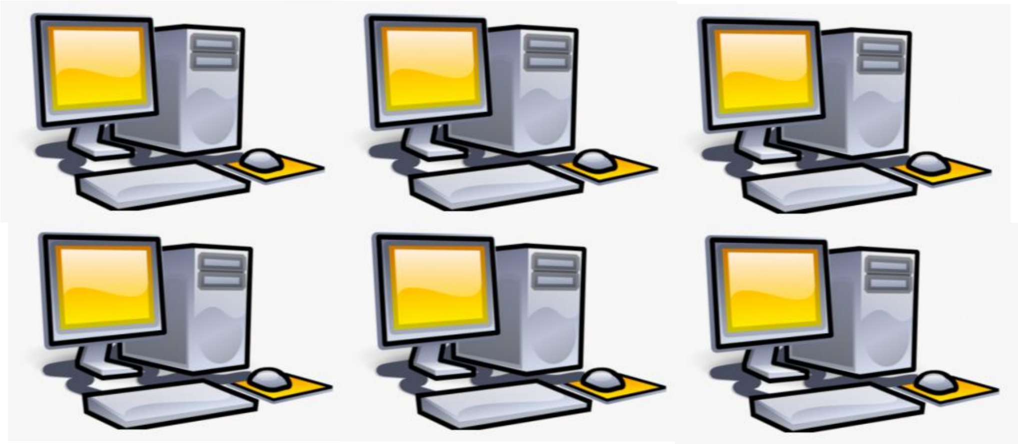


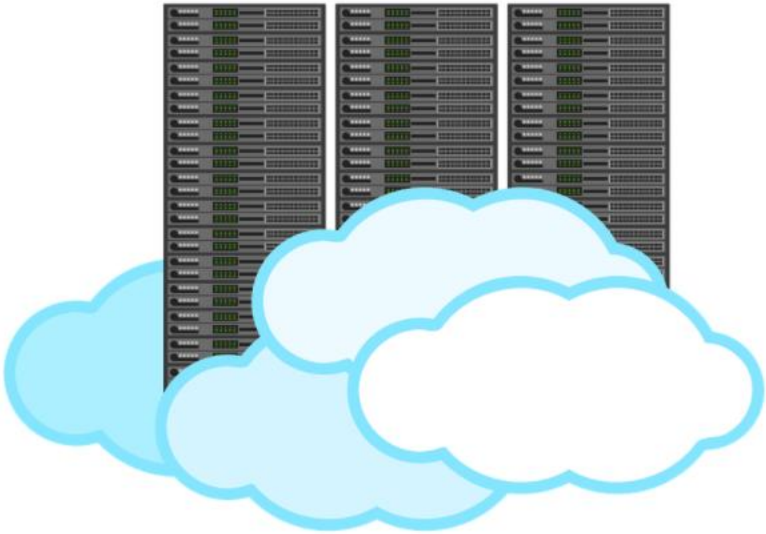
Data storage and retrieval

Things to consider when storing data

- Location
- Data type
- Retrieval







Types of data storage

Unstructured

- Email
- Text
- Video and audio files
- Web pages
- Social media

Document Database

Tabular

Customer Name	Customer Address	...
Jane Doe	123 Maple St.	...

Relational Database

Data Type	Query Language
Document Database	NoSQL
Relational Database	SQL

Datasets

Longitudinal

- Observations are dependent
- Temporal-continuity is required

	2009
January	339 778
February	343 438
March	339 228
April	338 344
May	339 873
June	342 912
July	342 489
August	350 800
September	343 687
October	347 641
November	354 467
December	354 085

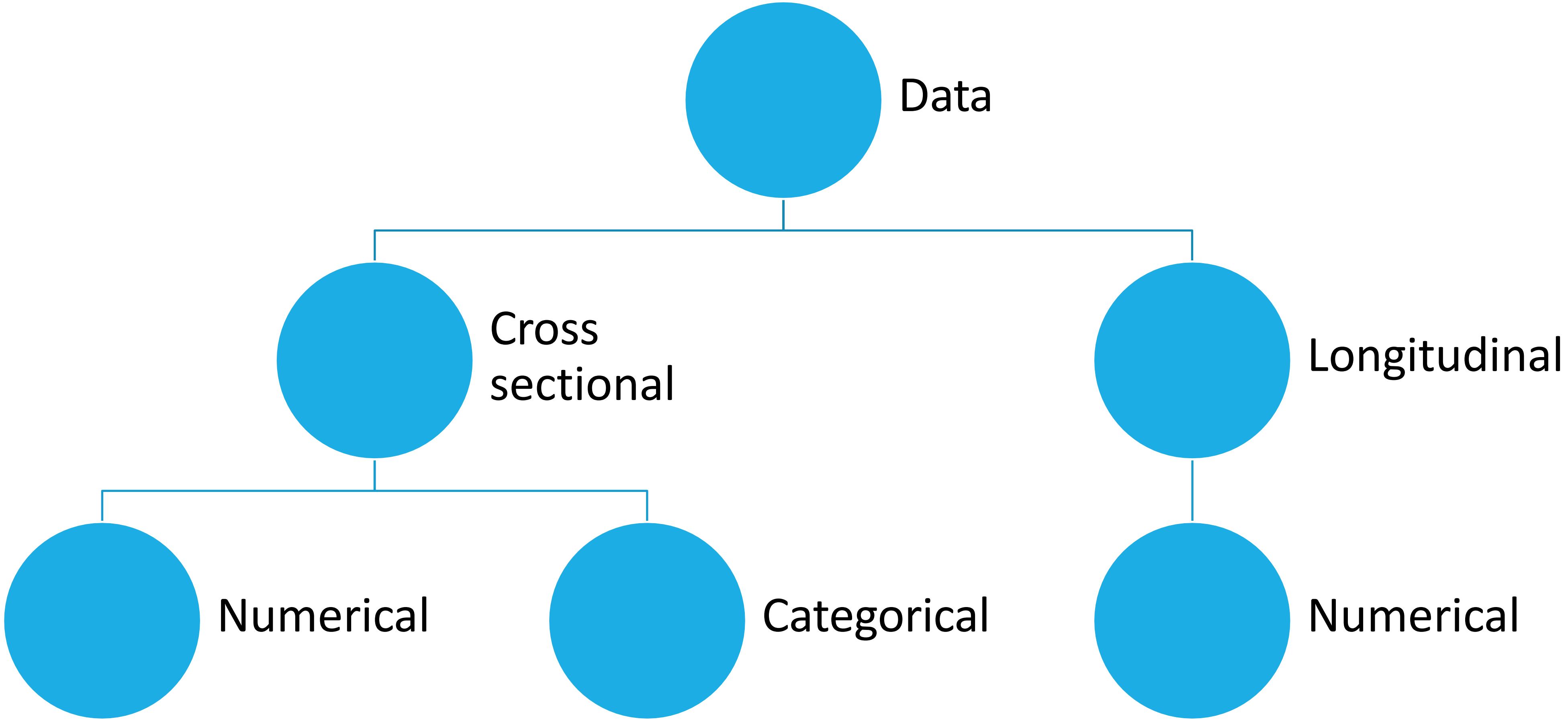
Cross-sectional

- Observations are independent

Ten Highest-Yielding Dow Stocks

December 31, 2007	2007 Dividends	Price	Yield
1 Citigroup	2.1600	29.44	7.34%
2 Pfizer	1.1600	22.73	5.10%
3 Altria Group	3.0500	75.58	4.04%
4 General Motors	1.0000	24.89	4.02%
5 Verizon	1.6450	43.69	3.77%
6 du Pont	1.5200	44.09	3.45%
7 AT&T	1.4200	41.56	3.42%
8 Home Depot	0.9000	26.94	3.34%
9 JP Morgan Chase	1.4400	43.65	3.30%
10 General Electric	1.1500	36.96	3.11%

Summary


Data
Cross sectional
Longitudinal
Numerical
Categorical
Numerical

Goals

Goal

- **Descriptive Statistics**
  - Goal is to describe the data at hand
  - Backward-looking
  - Statistical techniques employed here

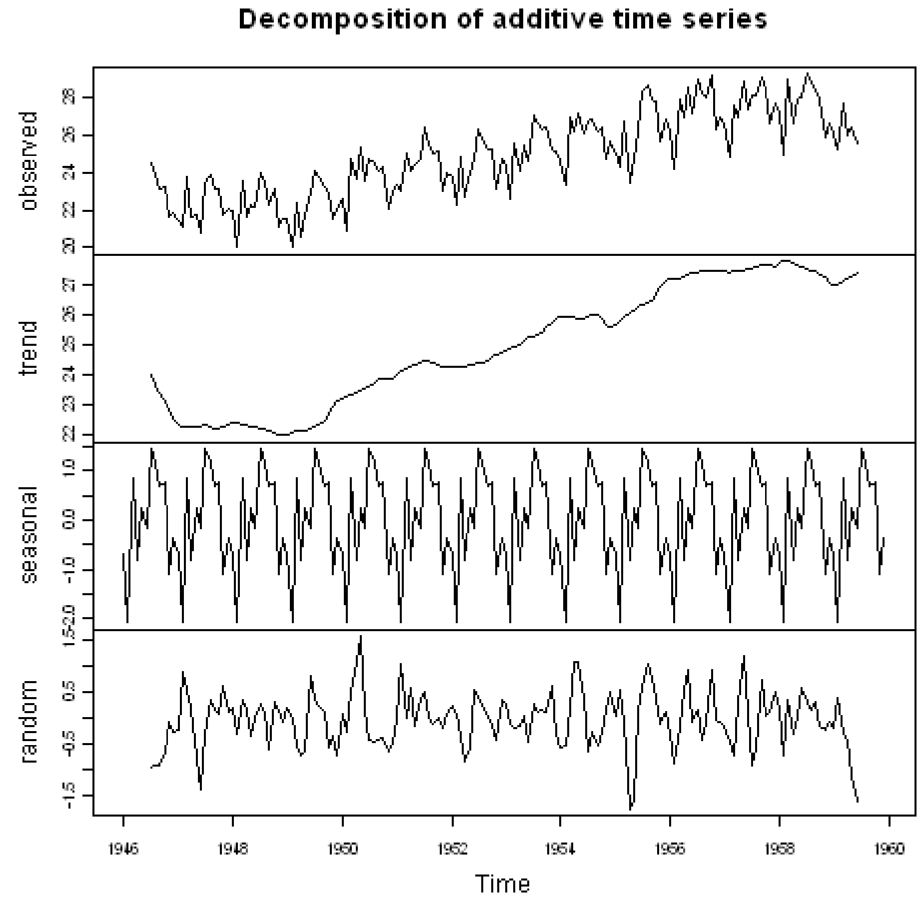
- **Predictive Analytics**
  - Goal is to use historical data to build a model for prediction
  - Forward-looking
  - Machine learning & AI techniques employed here

Descriptive Statistics – Cross sectional datasets

- How do you summarize numerical variables ?
- How do you summarize categorical variables ?
- How do you describe variability in numerical variables ?
- How do you summarize relationships between categorical and numerical variables ?
- How do you summarize relationships between 2 numerical variables?

Longitudinal datasets

Goal is to extract the various components



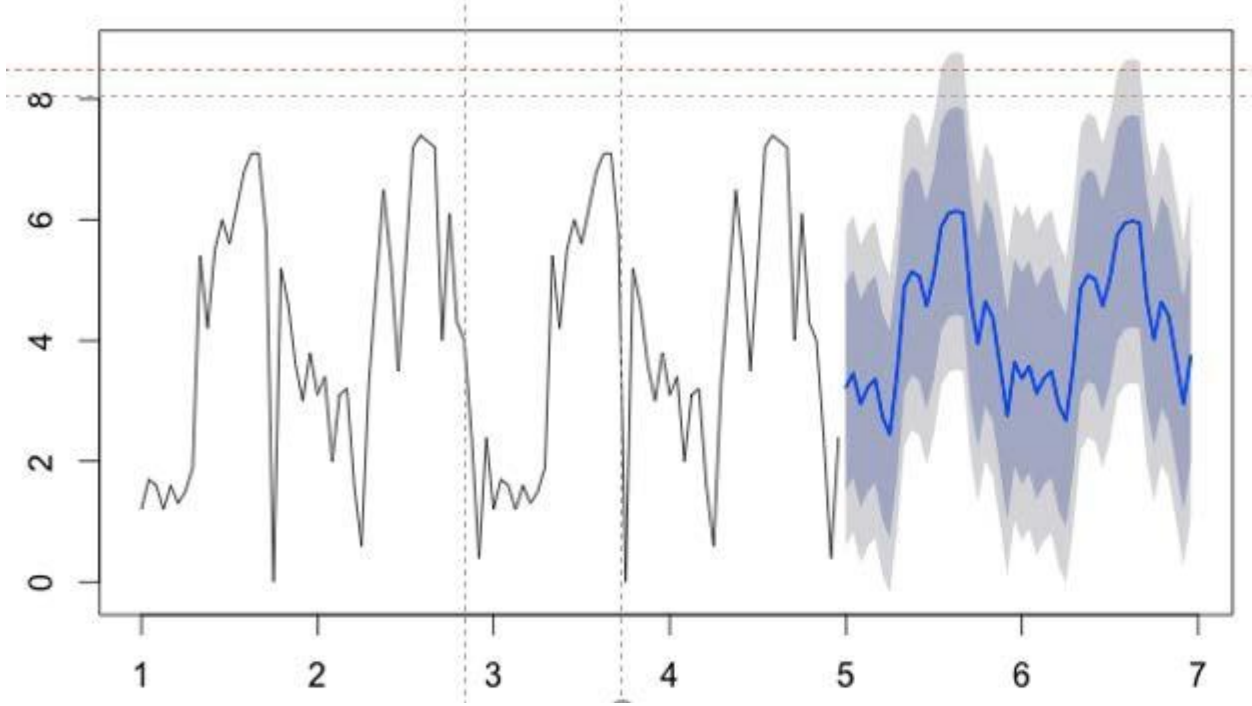
Predictive Analytics : Cross sectional datasets

- Given a dataset, build a model that captures the similarities in different observations and assigns them to different buckets.

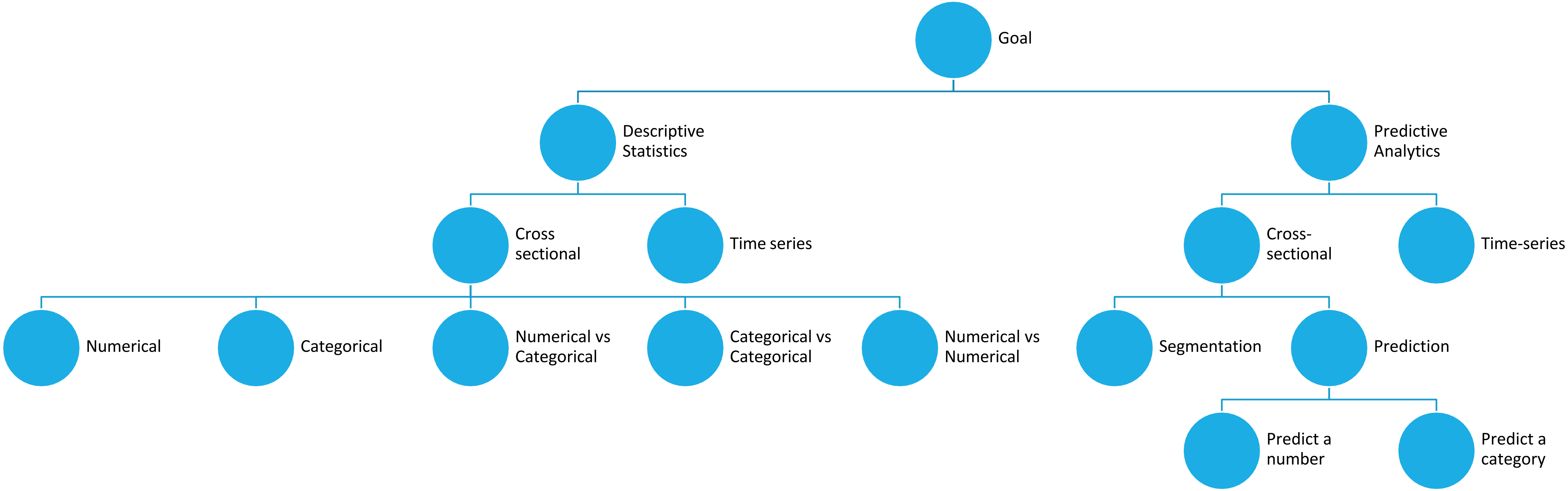
- Given a set of variables, predict the value of another variable in a given data set
  - Predict salaries given work experience, education etc.

  - Predict whether a loan would be approved given fico score, current loans, employment status etc.

- Given a time series dataset, build a model that can be used to forecast values in the future

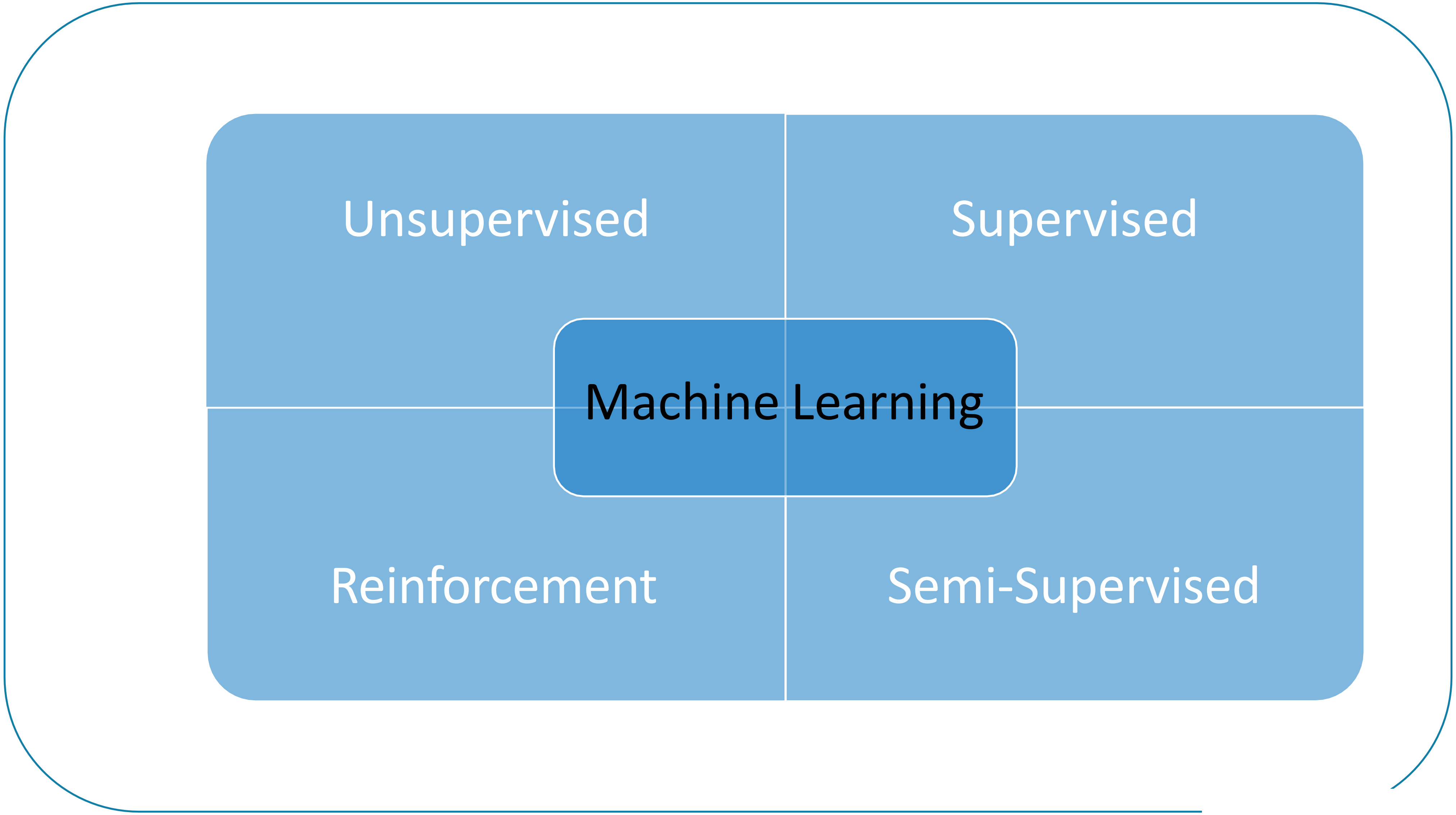


Summary

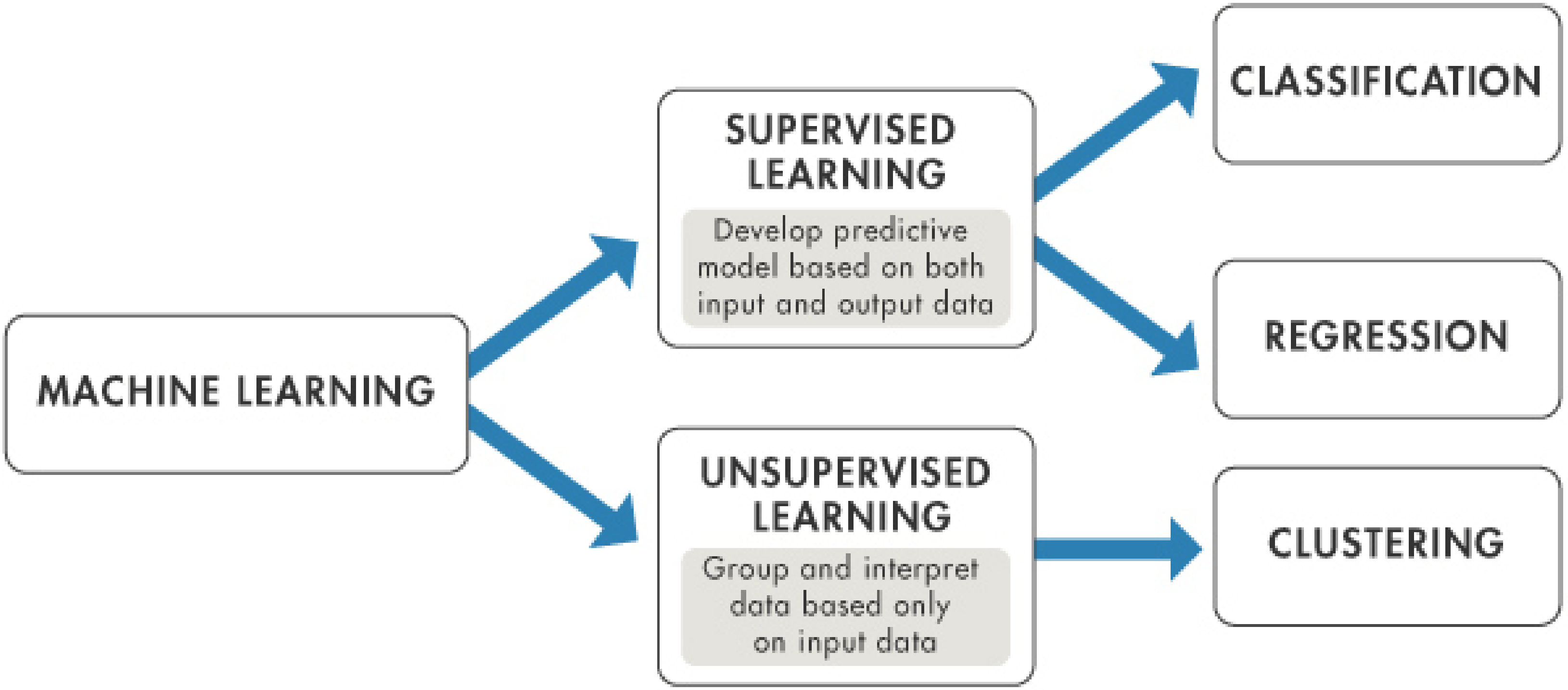

Goal
Descriptive Statistics
Predictive Analytics
Cross sectional
Time series
Cross-sectional
Time-series
Numerical
Categorical
Numerical vs Categorical
Categorical vs Categorical
Numerical vs Numerical
Segmentation
Prediction
Predict a number
Predict a category

Machine Learning algorithms

Machine Learning



Machine Learning Types



Supervised Algorithms

- Given a set of variables $\underline{x}$, predict the value of another variable y in a given data set such that

x1,x2,x3… → Model F(X) → y

- If y is numeric => **Prediction**
- If y is categorical => **Classification**
- **Example:** Given that a customer’s Debt-to-Income ratio increased 20%, what are the chances he/she would default in 3 months?

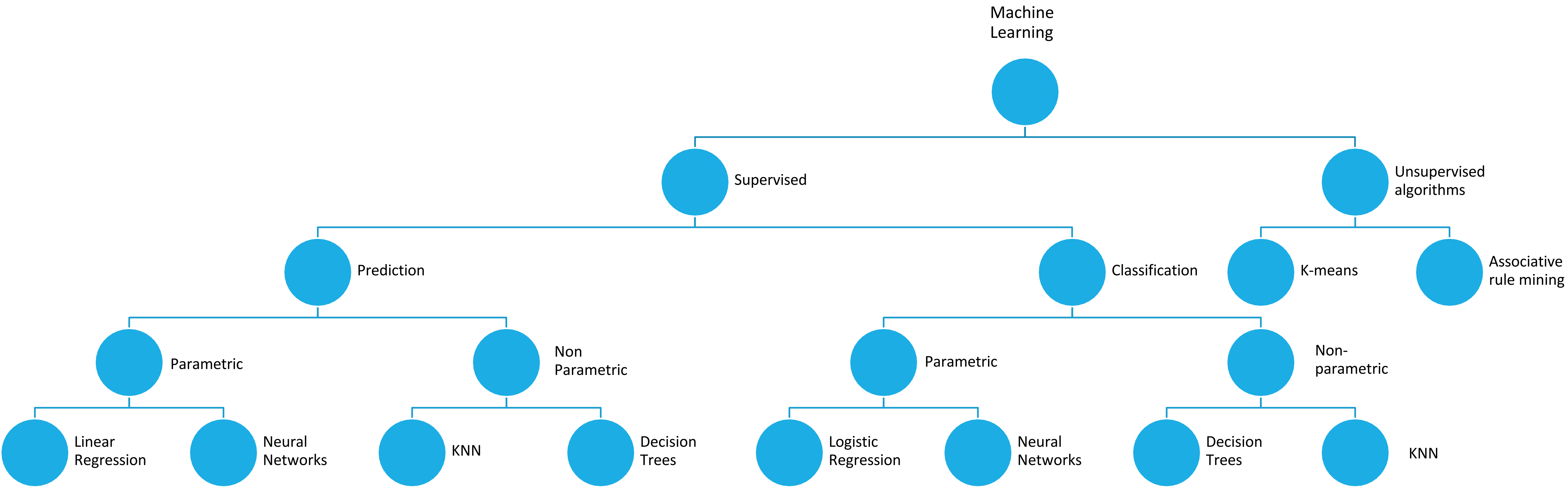
Unsupervised Algorithms

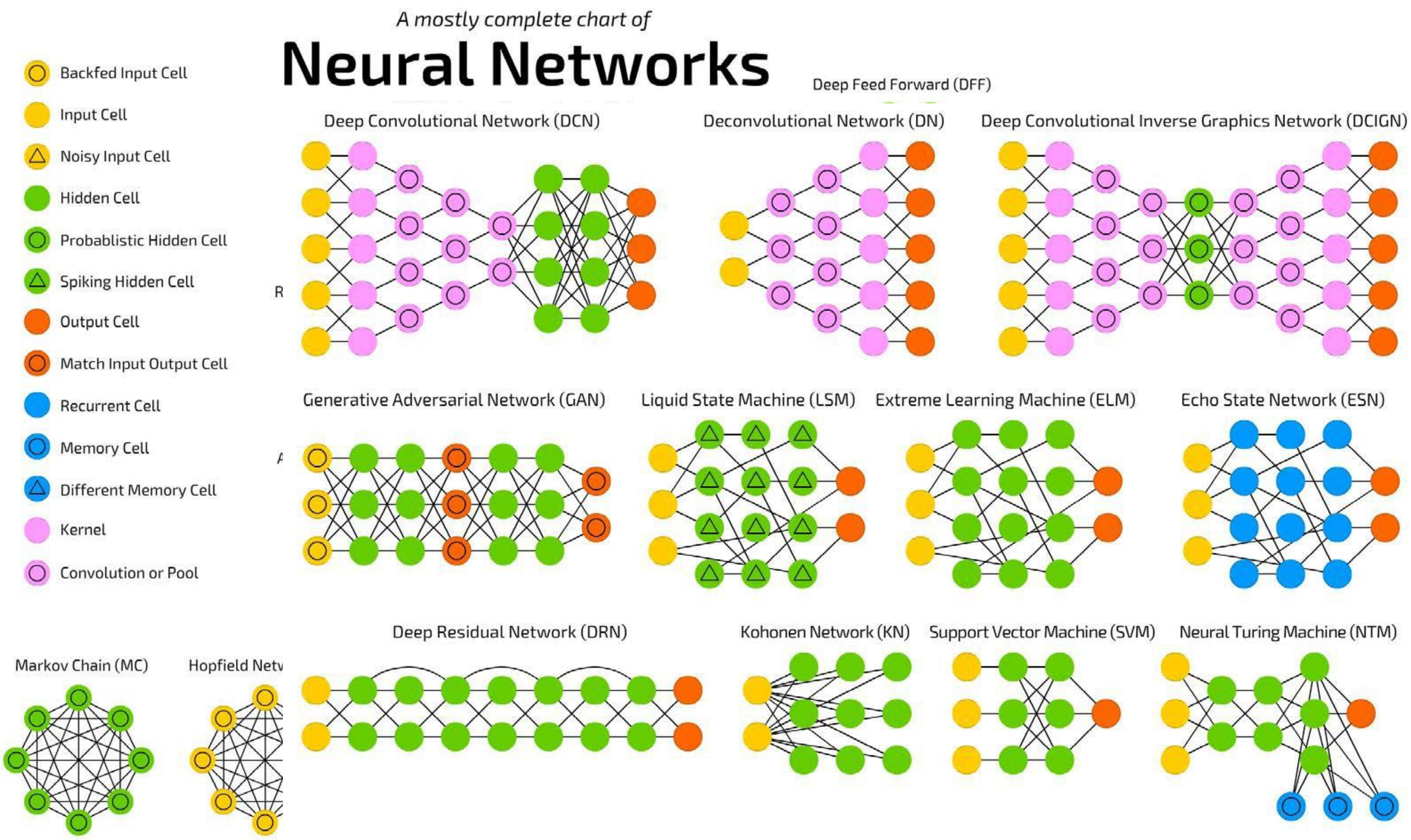
- Given a dataset with variables $\underline{x}$, build a model that captures the similarities in different observations and assigns them to different buckets => **Clustering**

Obs1, Obs2,Obs3 etc. → Model → Obs1- Class 1
Obs2- Class 2
Obs3- Class 1

- Example: Given a list of emerging market stocks, can we segment them into three buckets?

Machine Learning Algorithms





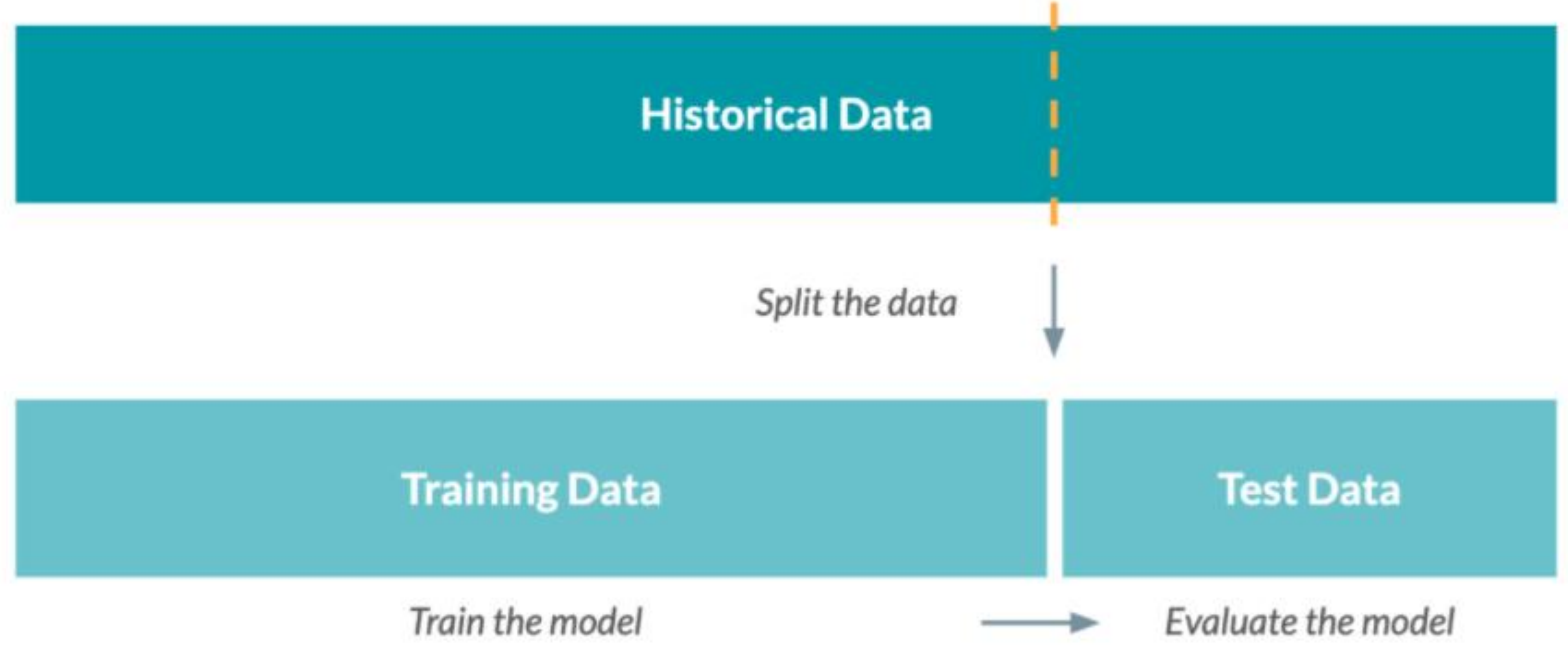
http://www.asimovinstitute.org/neural-network-zoo/

The Process

Data ingestion → Data cleaning → Feature engineering → Training and testing → Model building → Model selection

Model evaluation

Split historical data into training and testing sets



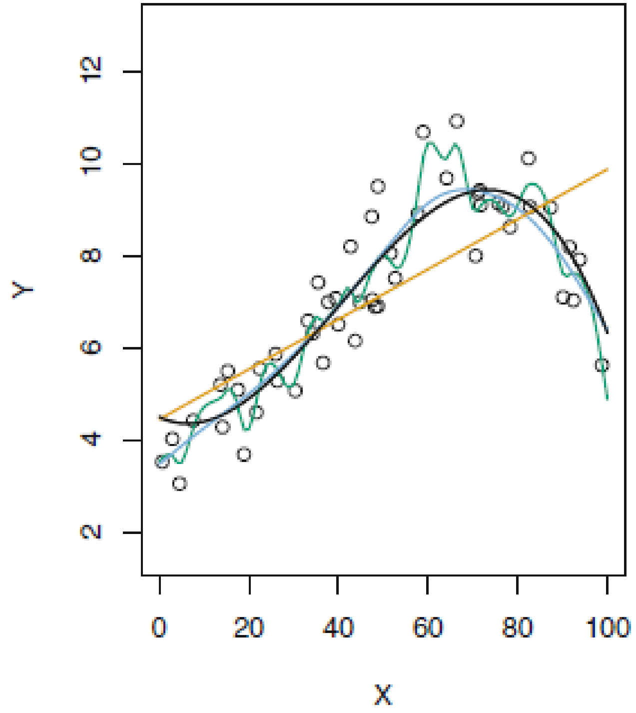
Performance evaluation

Model performance

- Over fitting
- Cross validation
- Evaluation metrics

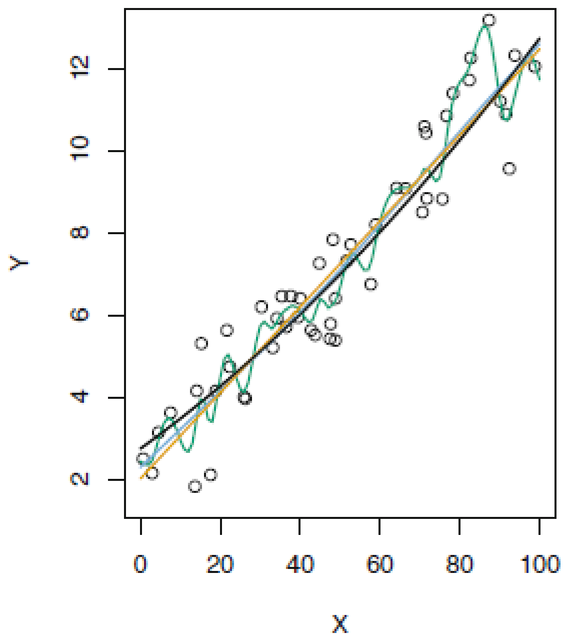
Overfitting

Assume that the "true" *f* is given by the black curve. The others are three possible estimates for *f*. The orange line is the linear regression fit. The blue and green curves were produced using flexible methods. The green curve is the most flexible and it is the best one in matching the data; however, we observe that it fits the true *f* poorly.



Overfitting

This situation is referred to as *overfitting*. This happens because the procedure is working too hard to find patterns in the training data, and may be picking up some patterns that are caused by random chance rather than by true properties of the unknown function. The picture shows overfitting when the true model is linear.



Cross Validation

Evaluation Metrics

Regression

- Mean absolute error (MAE)
- Mean squared error (MSE)
- R squared (R^2)
- Adjusted R squared (Adj-R^2)

Classification

- Accuracy
- Precision
- Recall
- Area under curve (AUC)
- Confusion matrix

Model accuracy

In order to evaluate the performance of a statistical learning method on a given data set, we have to measure how accurately its predictions match the observed data. This is usually done by estimating the error on the training set.

- Linear regression - *Mean Squared Error*:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

- Classification – Indicator function:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

Model accuracy

One is generally interested in the accuracy of the predictions obtained applying the method to previously unseen test data.

This theoretically corresponds to minimizing the average prediction error for large number of test observations (*average test MSE*), that could be not available.

Model accuracy

It can be shown that the *expected test MSE*, that is the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at a fixed point can be decomposed into the sum of three fundamental quantities:

1. variance, which refers to the amount by which the estimated function would change if we used a different training data set;
2. bias, which refers to the error that is introduced by approximating the "true" model by a simpler one;
3. irreducible error.

As a general rule, <u>as we use more flexible methods, the variance will increase and the bias will decrease</u>.

Evaluation Metrics

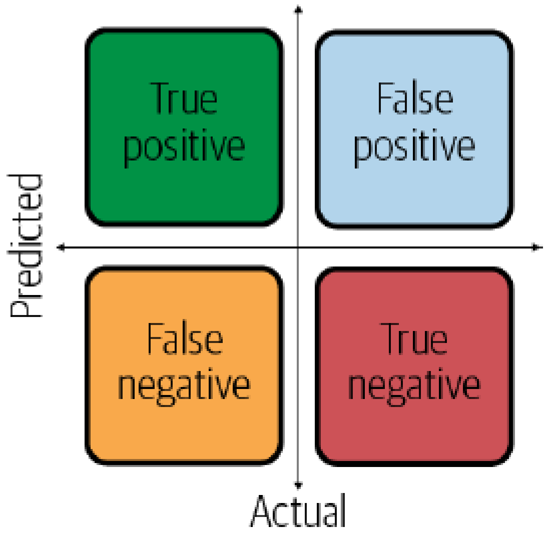
For simplicity, we will mostly discuss things in terms of a binary classification problem; some common terms are:

- True positives (TP) $\rightarrow$ Predicted positive and are actually positive.
- False positives (FP) $\rightarrow$ Predicted positive and are actually negative.
- True negatives (TN) $\rightarrow$ Predicted negative and are actually negative.
- False negatives (FN) $\rightarrow$ Predicted negative and are actually positive.

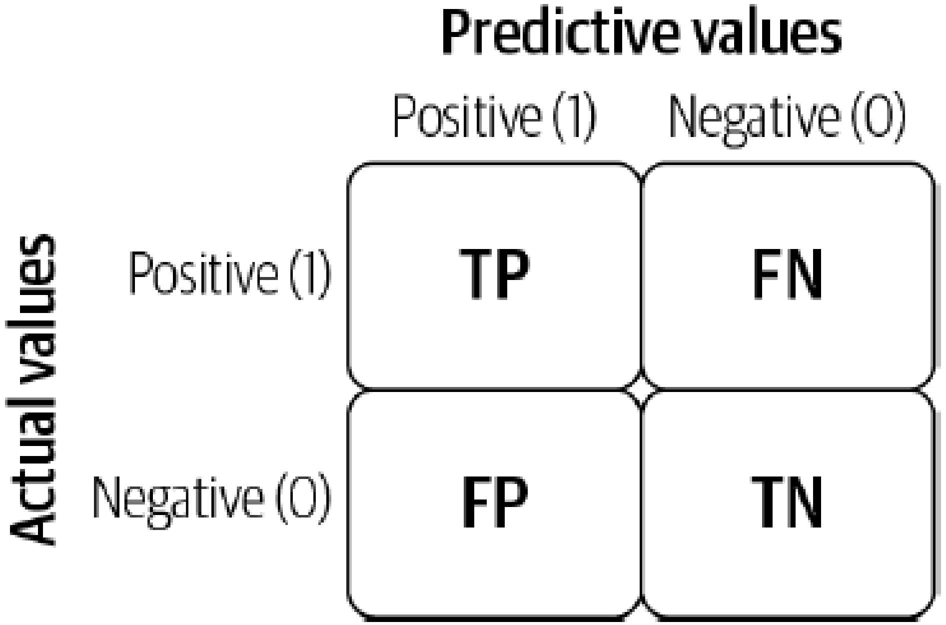
$$\textbf{Precision} = \frac{\text{True positive}}{\text{Actual results}} \text{ or } \frac{\text{True positive}}{\text{True positive + False positive}}$$

$$\textbf{Recall} = \frac{\text{True positive}}{\text{Predictive results}} \text{ or } \frac{\text{True positive}}{\text{True positive + False negative}}$$

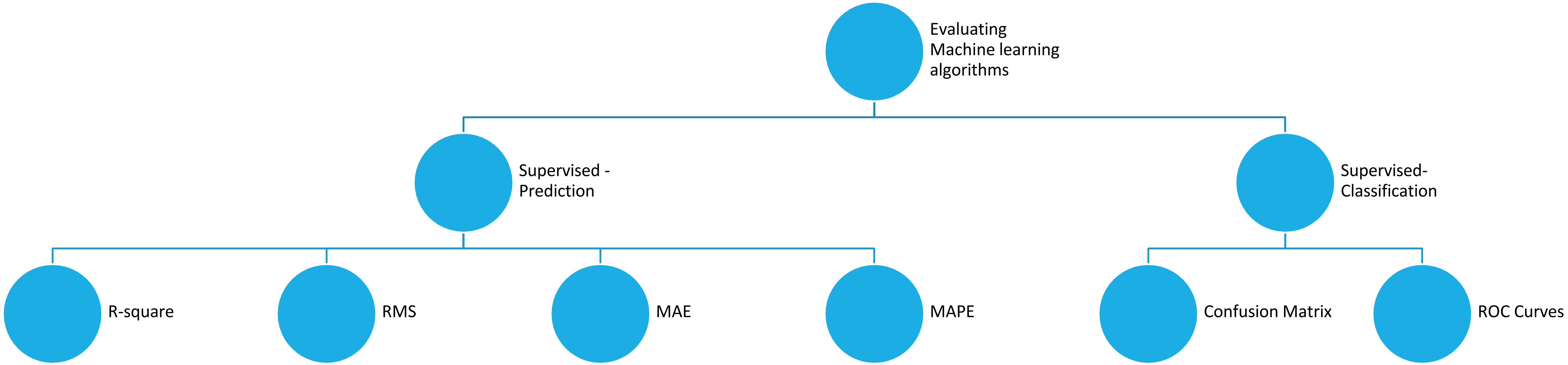
$$\textbf{Accuracy} = \frac{\text{True positive + True negative}}{\text{Total}}$$



Confusion matrix



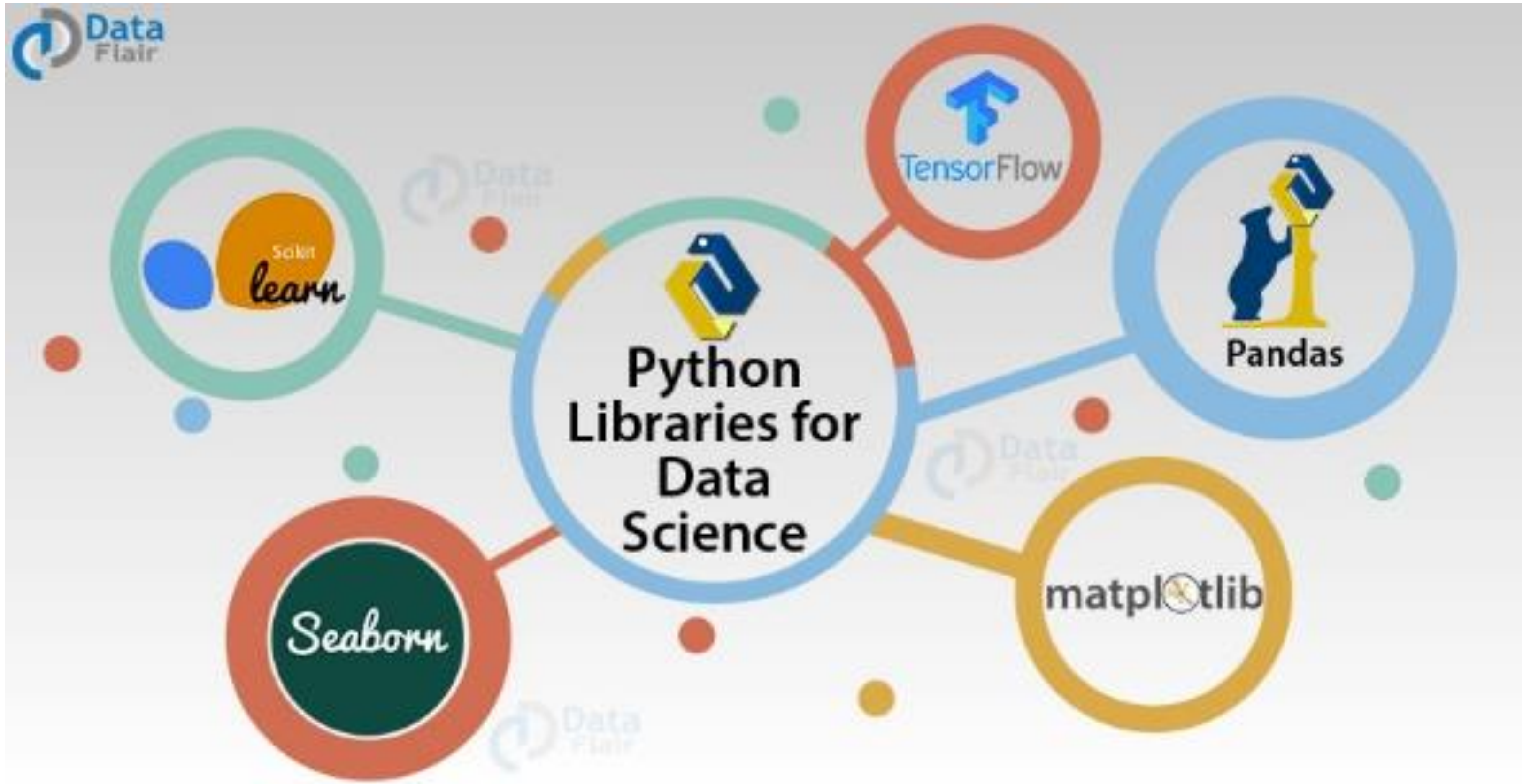
Evaluation framework



Model selection

	Linear regression	Logistic regression	SVM	CART	Gradient boosting	Random forest	Artificial neural network	KNN	LDA
Simplicity	✔	✔	✔	✔	✘	✘	✘	✔	✔
Training Time	✔	✔	✘	✔	✘	✘	✘	✔	✔
Handle non-linearity	✘	✘	✔	✔	✔	✔	✔	✔	✔
Robust to overfitting	✘	✘	✔	✘	✘	✔	✘	✔	✘
Large datasets	✘	✘	✘	✔	✔	✔	✔	✘	✔
Many features	✘	✘	✔	✔	✔	✔	✔	✘	✔
Model interpretation	✔	✔	✘	✔	✔	✔	✘	✔	✔
Feature scaling needed	✘	✘	✔	✘	✘	✘	✘	✘	✘

Python for Data Science

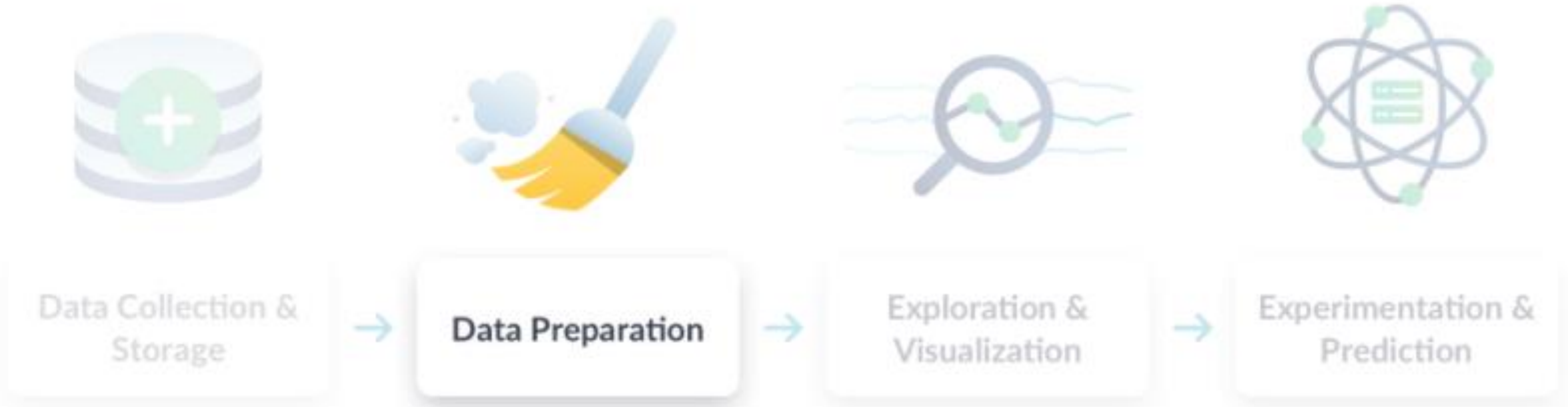


Main Python libraries for Data Science

- **NumPy**: storage and manipulation of dense data arrays.
- **Pandas**: DataFrame object for storage and manipulation of labeled/columnar data.
- **Matplotlib**: capabilities for a flexible range of data visualizations.
- **Scikit-learn**: machine learning.

Data cleaning

Data preparation



Why prepare data?

- Preparation is done to prevent:
- Errors
- Incorrect results
- Biasing algorithms

Tidy Data

	Sara	Lis	Hadrien	Lis
Age	"27"	"30"		"30"
Size	1.77	5.58	1.80	5.58
Country	"Belgium"	"USA"	"FR"	"USA"

Name	Age	Size	Country
Sara	"26"	1.78	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

Tidy Data

Name	Age	Size	Country
Sara	"26"	1.78	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"
Lis	"30"	5.58	"USA"

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

Tidy Data

Name	Age	Size	Country
Sara	"27"	1.77	"Belgium"
Lis	"30"	5.58	"USA"
Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

Homogeneity

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	5.58	"USA"
2	Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

Homogeneity

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"Belgium"
1	Lis	"30"	1.70	"USA"
2	Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

Data types

ID	Name	Age	Size	Country
0	Sara	"27"	1.77	"BE"
1	Lis	"30"	1.70	"US"
2	Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien		1.80	"FR"

Data types

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien		1.80	"FR"

ID	Name	Age	Size	Country
0	Sara	27	1.77	"BE"
1	Lis	30	1.70	"US"
2	Hadrien	28	1.80	"FR"

Missing values

Reasons:
- Data entry
- Error
- Valid missing value

Solutions:
- impute
- drop
- keep