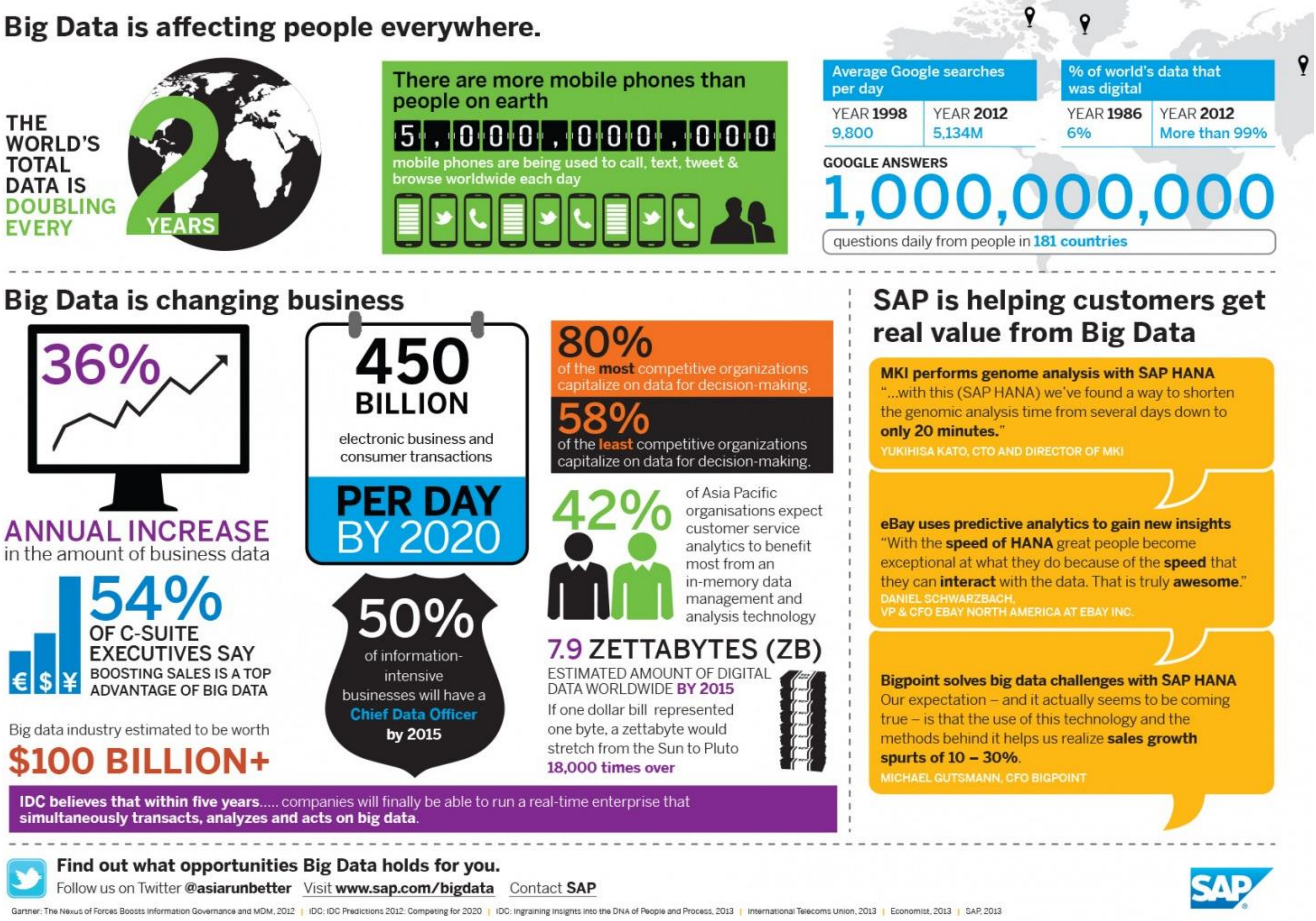
Data Science

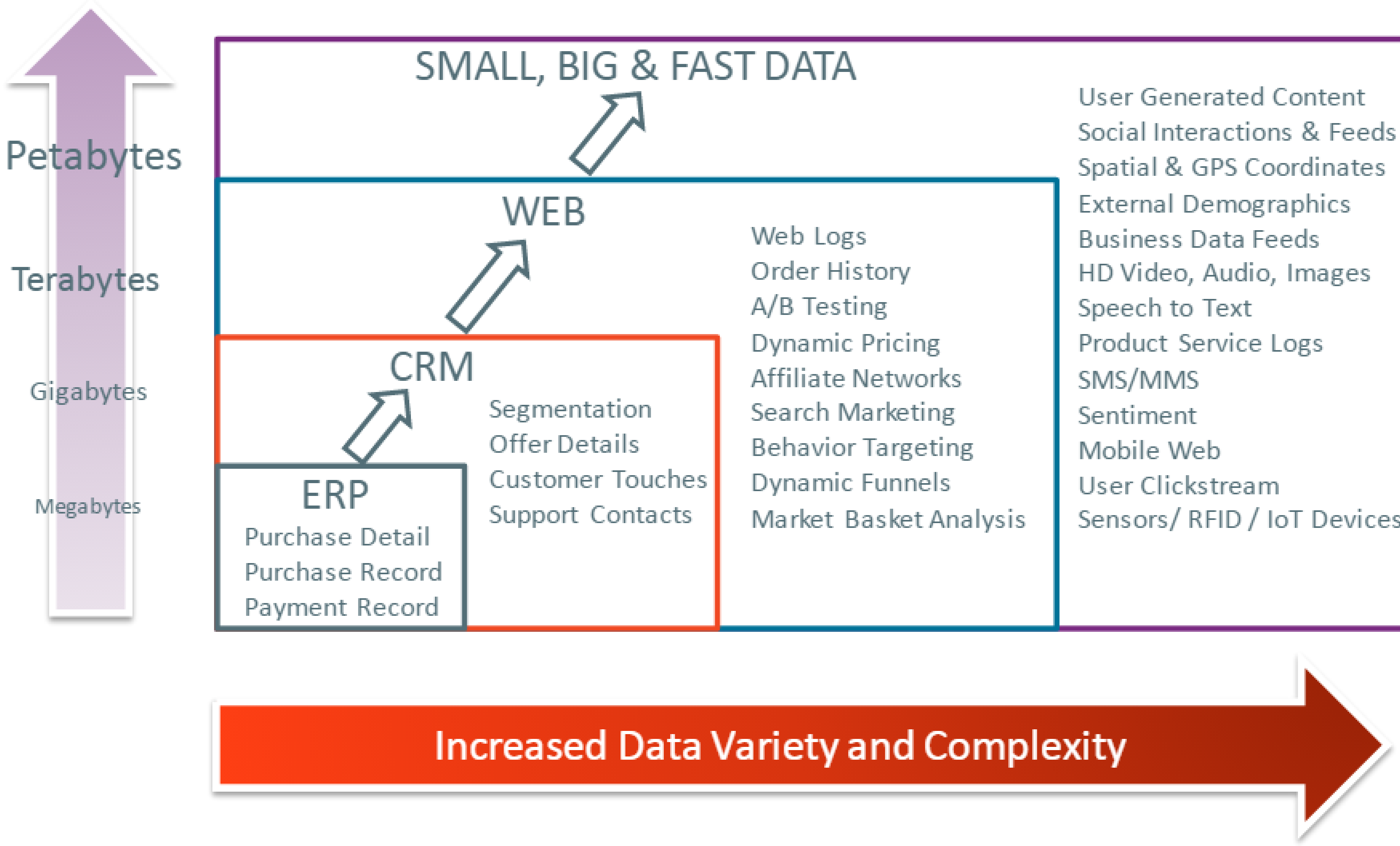
An Introduction

Setia Pramana

Data, data and data everywhere……

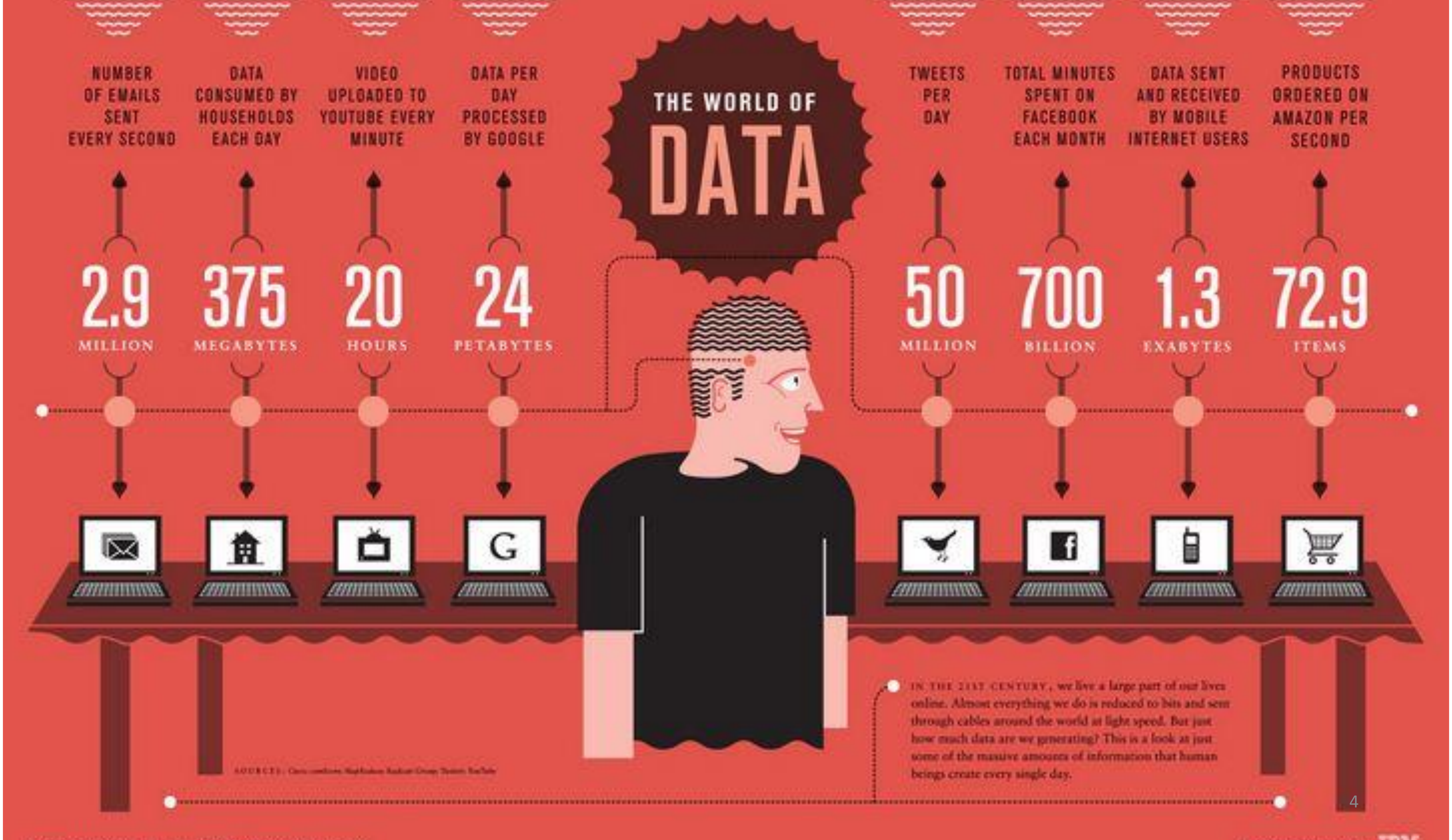


Data = Transactions + Interactions + Observations



Cervello

http://mycervello.com


THE WORLD OF
DATA
NUMBER OF EMAILS SENT EVERY SECOND
2.9 MILLION
DATA CONSUMED BY HOUSEHOLDS EACH DAY
375 MEGABYTES
VIDEO UPLOADED TO YOUTUBE EVERY MINUTE
20 HOURS
DATA PER DAY PROCESSED BY GOOGLE
24 PETABYTES
TWEETS PER DAY
50 MILLION
TOTAL MINUTES SPENT ON FACEBOOK EACH MONTH
700 BILLION
DATA SENT AND RECEIVED BY MOBILE INTERNET USERS
1.3 EXABYTES
PRODUCTS ORDERED ON AMAZON PER SECOND
72.9 ITEMS
G
f
IN THE 21ST CENTURY, we live a large part of our lives online. Almost everything we do is reduced to bits and sent through cables around the world at light speed. But just how much data are we generating? This is a look at just some of the massive amounts of information that human beings create every single day.

Type of Data

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- You can afford to scan the data once

What to do?

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

Analytics Approaches

- Descriptive: What happened or what is happening now?
- Diagnostic: Why did it happen or Why is it happening now?
- Predictive: What will happen next? What will happen under various conditions?
- Prescriptive: What are the options to create the most optimal/high value result/outcome?

Data Science

“Applying advanced statistical tools to existing data to solve problems, generate new insights, improve products/services”

“Everything that has something to do with data: Collecting, analyzing, modeling...... yet the most important part is its applications --- all sorts of application”

What is Data Science?

- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.

Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics

William S. Cleveland

Statistics Research, Bell Labs

DOI:10.1002/sam.11239
Published online in Wiley Online Library (wileyonlinelibrary.com).

Abstract: An action plan to expand the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations. Statistical Analysis and Data Mining 7: 414–417, 2014

1. SUMMARY OF THE PLAN

This article describes a plan to broaden the major areas of technical work of the field of statistics. Because the plan is ambitious and implies su

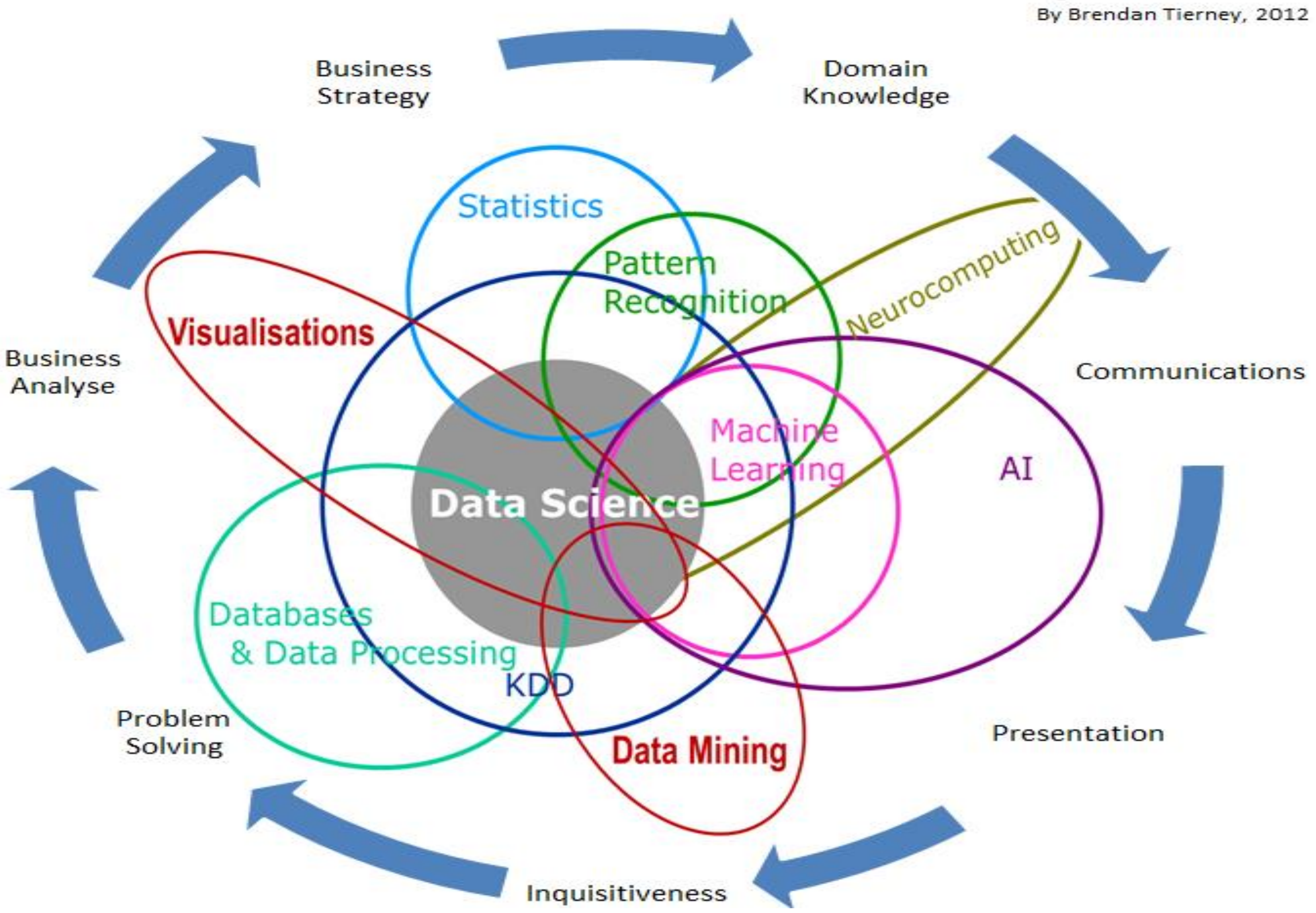
will be called 'data scien

The focus of the plan is

- *Models and Methods for Data (20%)*: statistical models; methods of model building; and methods of estimation and distribution based on probabilistic

hardware systems;

Data Science Is Multidisciplinary

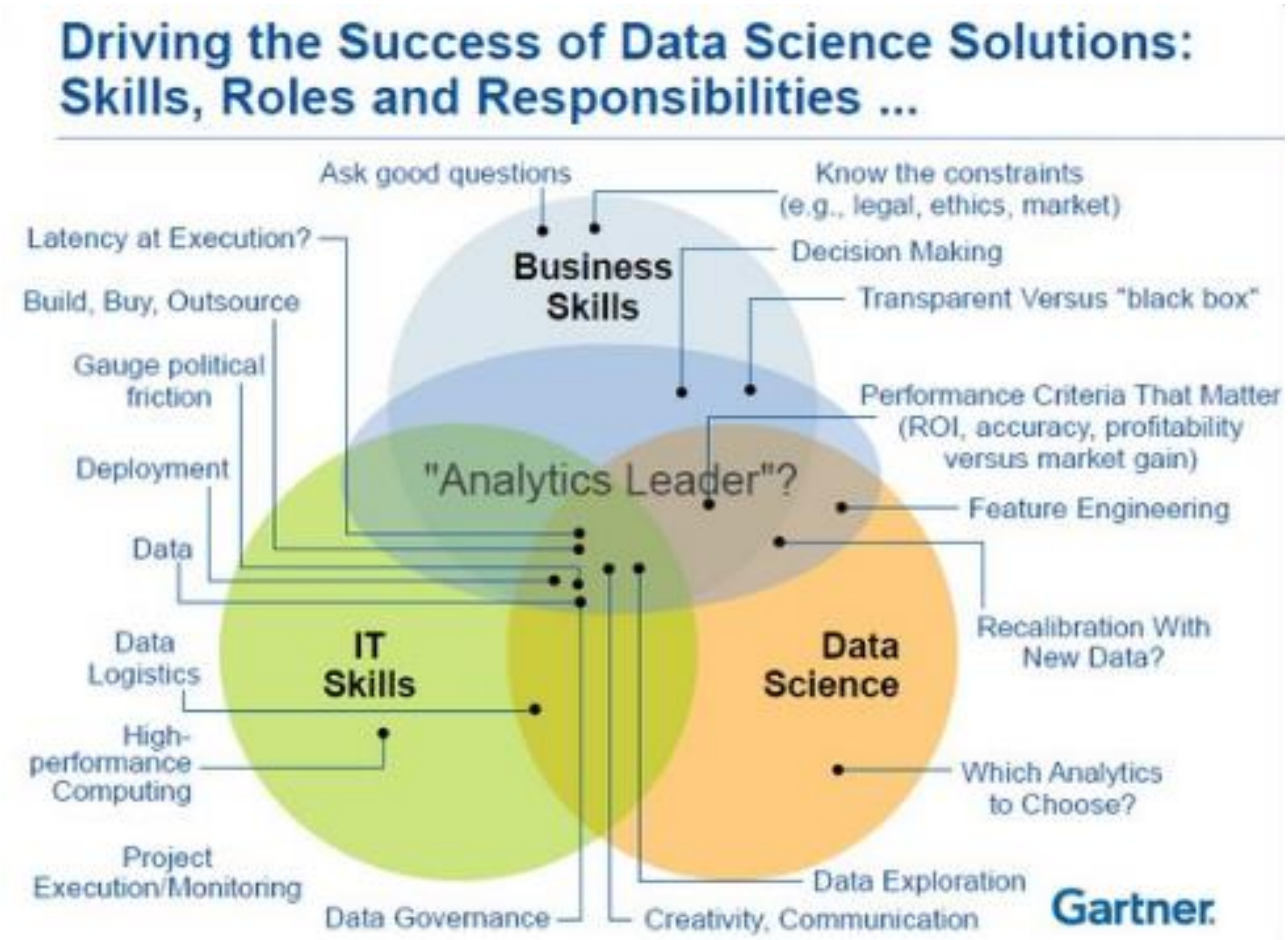


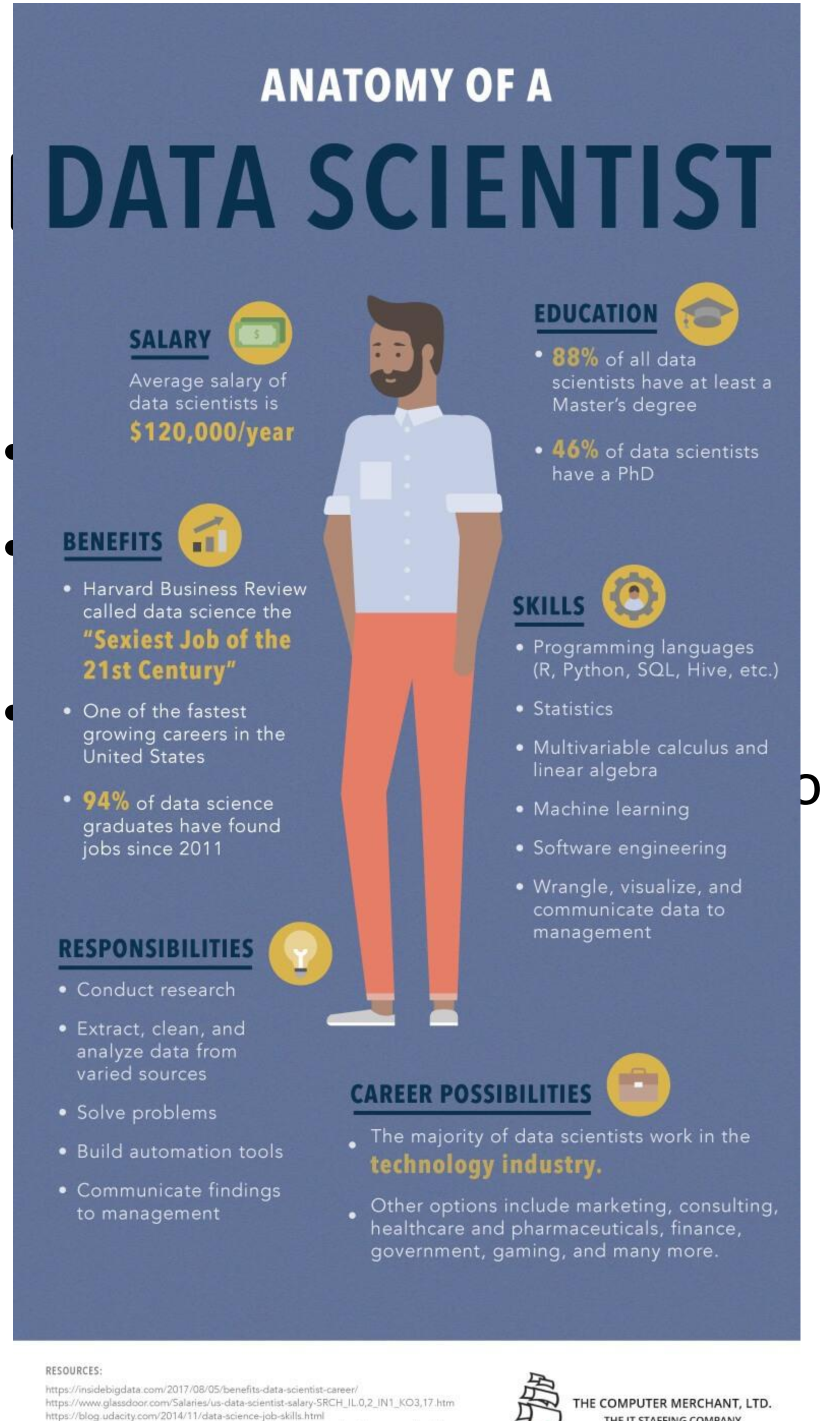
Data Science

- A Mashed Up Discipline
- A multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data

Math and Theory	• Statistics, Linear Algebra, Optimization, Time Series, etc.
Applied Algorithms	• Machine Learning, Data Structures, Parallel Algorithms, etc.
Engineering and Technologies	• Storage and computing platforms, statistical tools ,etc.
Domain Expertise	• Text, Finance, Images, Econometrics etc.
Art	• Visualization, Infographics
Best practices and hacks	• Handle missed values in data, transform and represent data, etc.

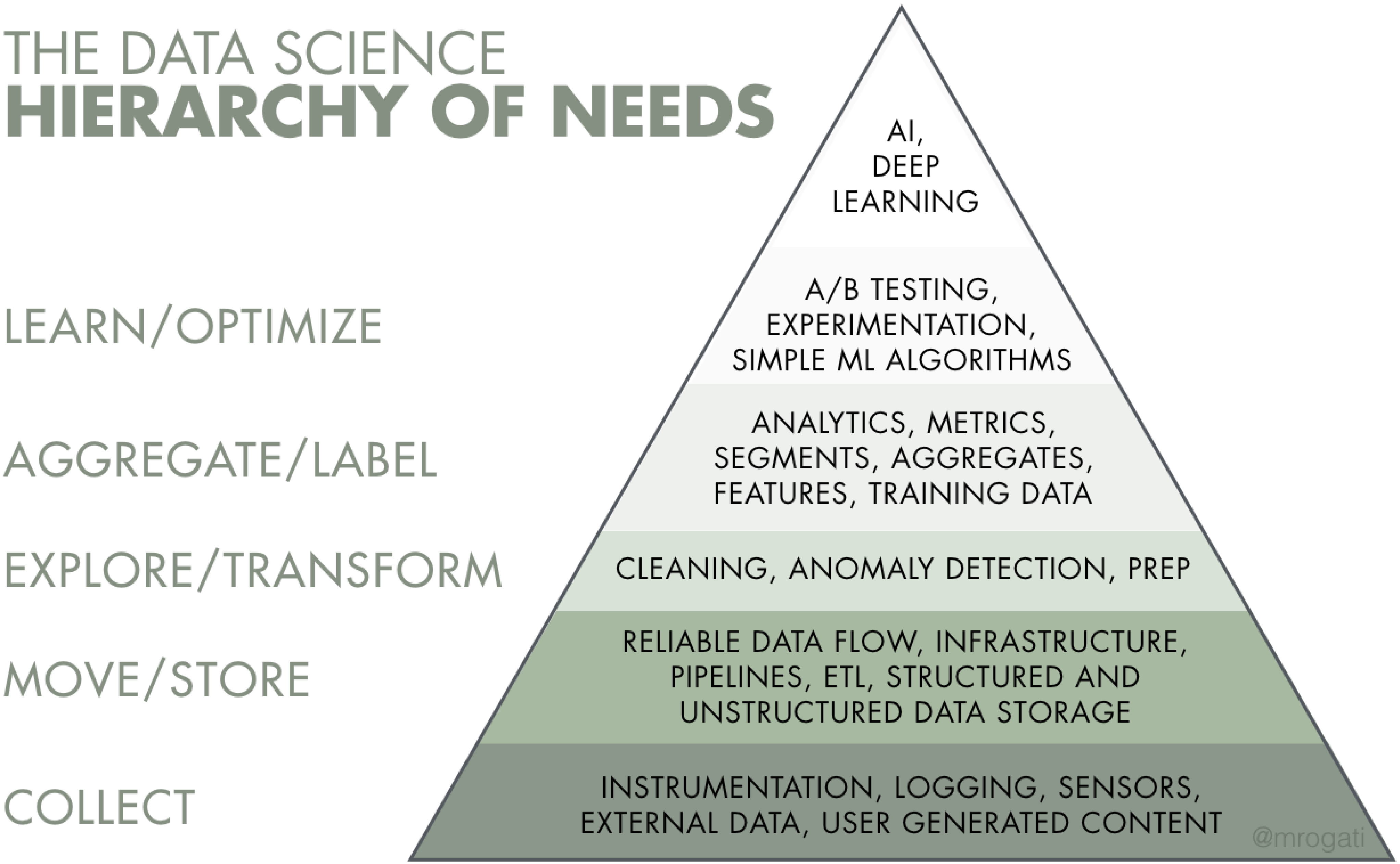
Data Science



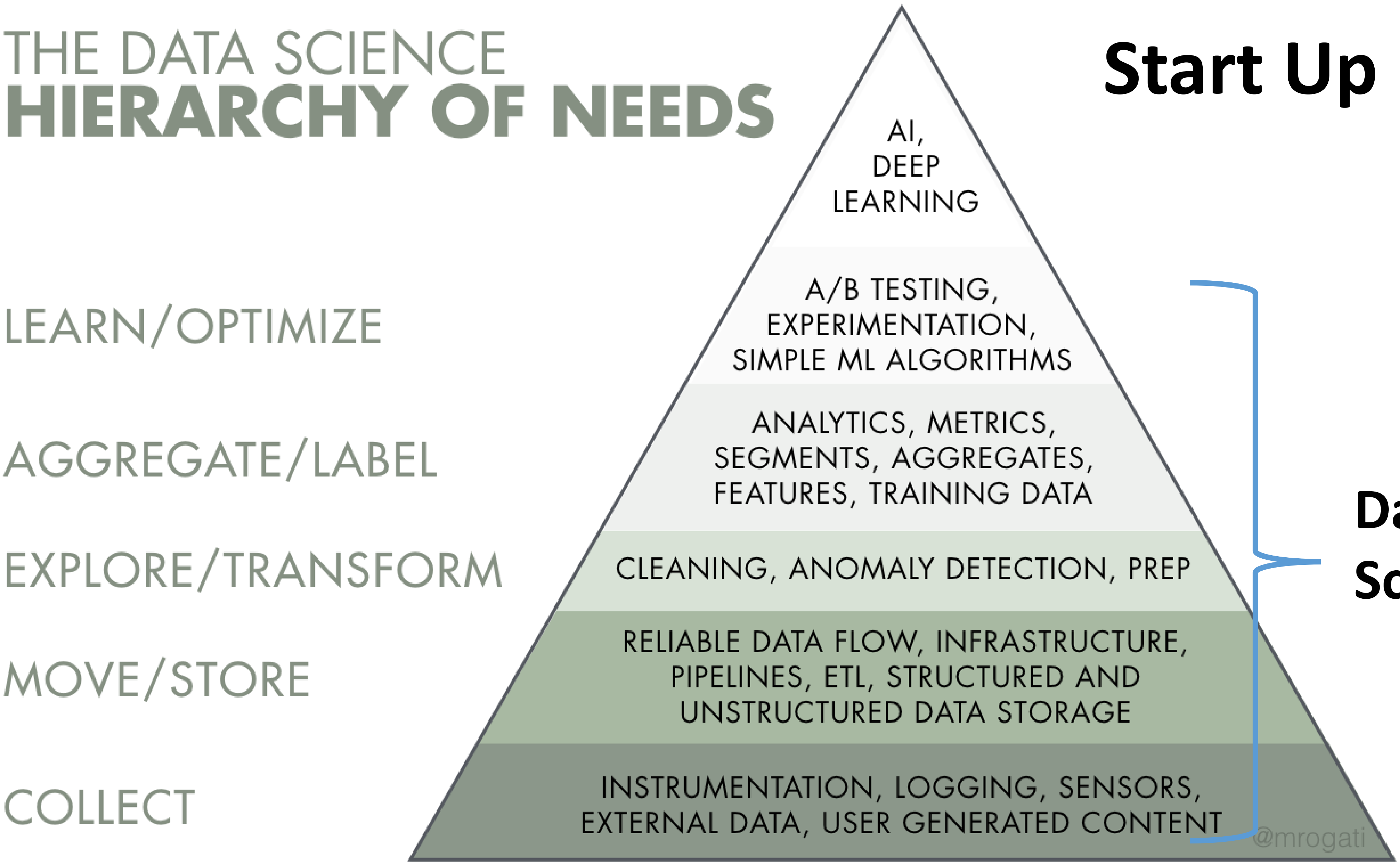

ANATOMY OF A
DATA SCIENTIST
SALARY
Average salary of data scientists is $120,000/year
EDUCATION
• 88% of all data scientists have at least a Master's degree
• 46% of data scientists have a PhD
BENEFITS
• Harvard Business Review called data science the "Sexiest Job of the 21st Century"
• One of the fastest growing careers in the United States
• 94% of data science graduates have found jobs since 2011
SKILLS
• Programming languages (R, Python, SQL, Hive, etc.)
• Statistics
• Multivariable calculus and linear algebra
• Machine learning
• Software engineering
• Wrangle, visualize, and communicate data to management
RESPONSIBILITIES
• Conduct research
• Extract, clean, and analyze data from varied sources
• Solve problems
• Build automation tools
• Communicate findings to management
CAREER POSSIBILITIES
• The majority of data scientists work in the technology industry.
• Other options include marketing, consulting, healthcare and pharmaceuticals, finance, government, gaming, and many more.
RESOURCES:
https://insidebigdata.com/2017/08/05/benefits-data-scientist-career/
https://www.glassdoor.com/Salaries/us-data-scientist-salary-SRCH_IL.0,2_IN1_KO3,17.htm
https://blog.udacity.com/2014/11/data-science-job-skills.html
THE COMPUTER MERCHANT, LTD.

orate


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



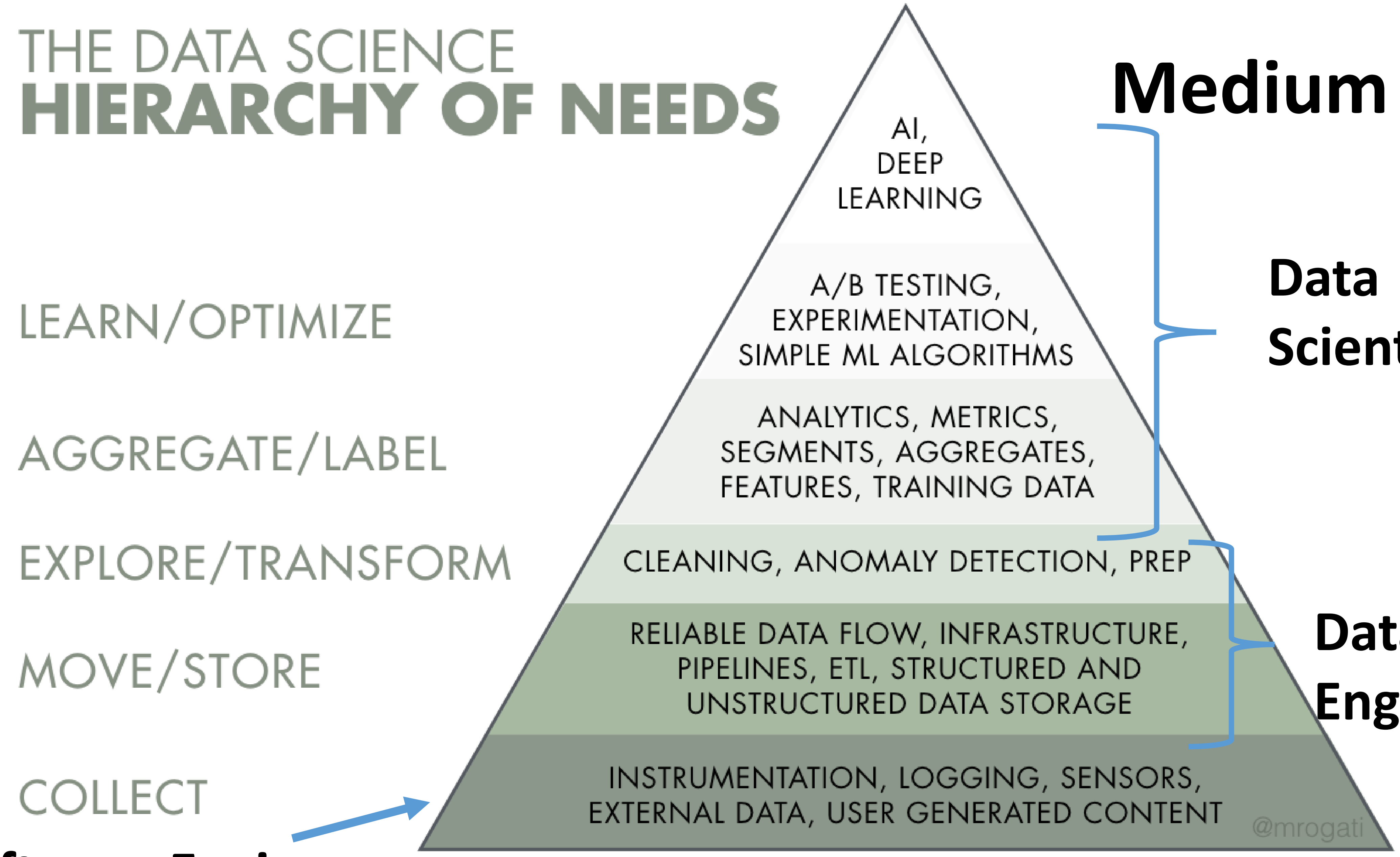
Monica Rogati https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007



Start Up

Data Scientist

[Monica Rogati](https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007) https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007



Medium

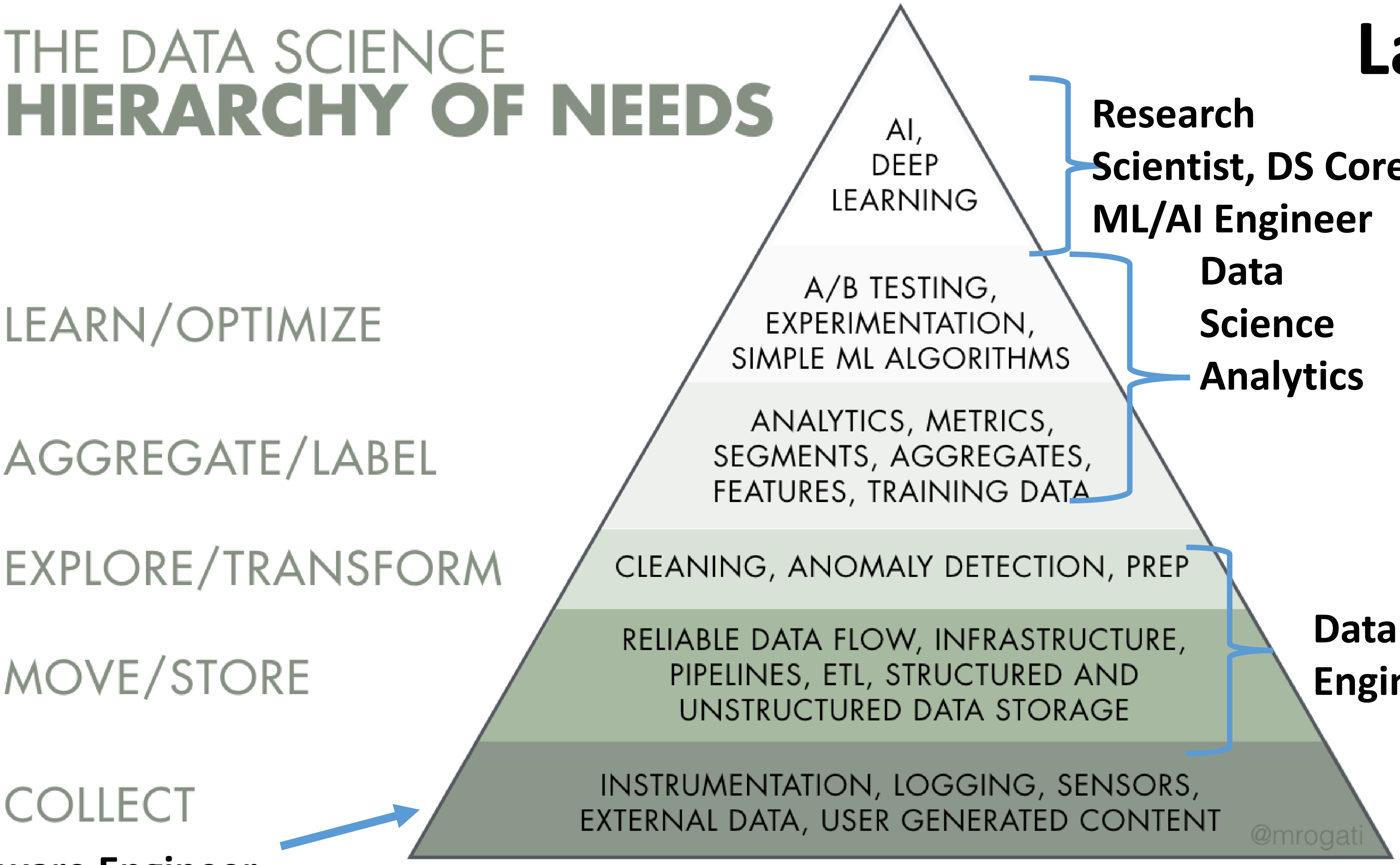
Data Scientist

Data Engineer

Software Engineer

[Monica Rogati](https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007) https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Large



Research Scientist, DS Core + ML/AI Engineer

Data Science Analytics

Data Engineer

Software Engineer

[Monica Rogati](https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007) https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007


DATA
Engineer
DataCamp
Learn Data Science By Doing
DATA
Scientist
Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.
Cleans, massages and organizes (big) data. Performs descriptive statistics and analysis to develop insights, build models and solve a business need.

Source: datacamp

Data Scientist

also known as Data Managers, statisticians.

A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication

Will use programmes such as:

SQL, Python, R

Data Engineers

also known as database administrators and data architects.

They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

Skills: Programming, Mathematics, Big data

Will use programmes such as:

Hadoop, NoSQL, and Python

Data Analysts

also known as business Analysts.

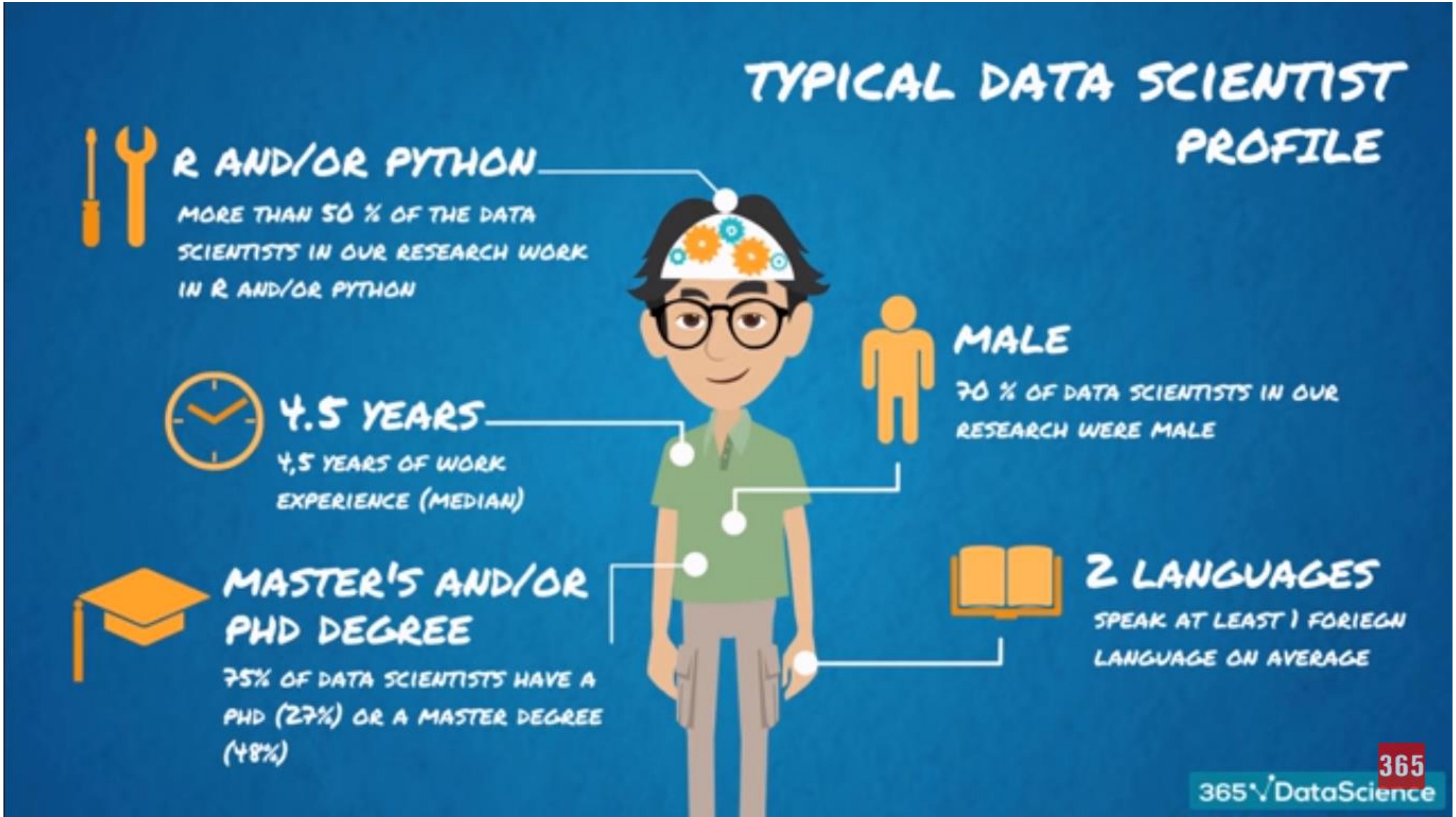
They typically help people from across the company understand specific queries with charts.

Skills: Statistics, Communication, Business knowledge

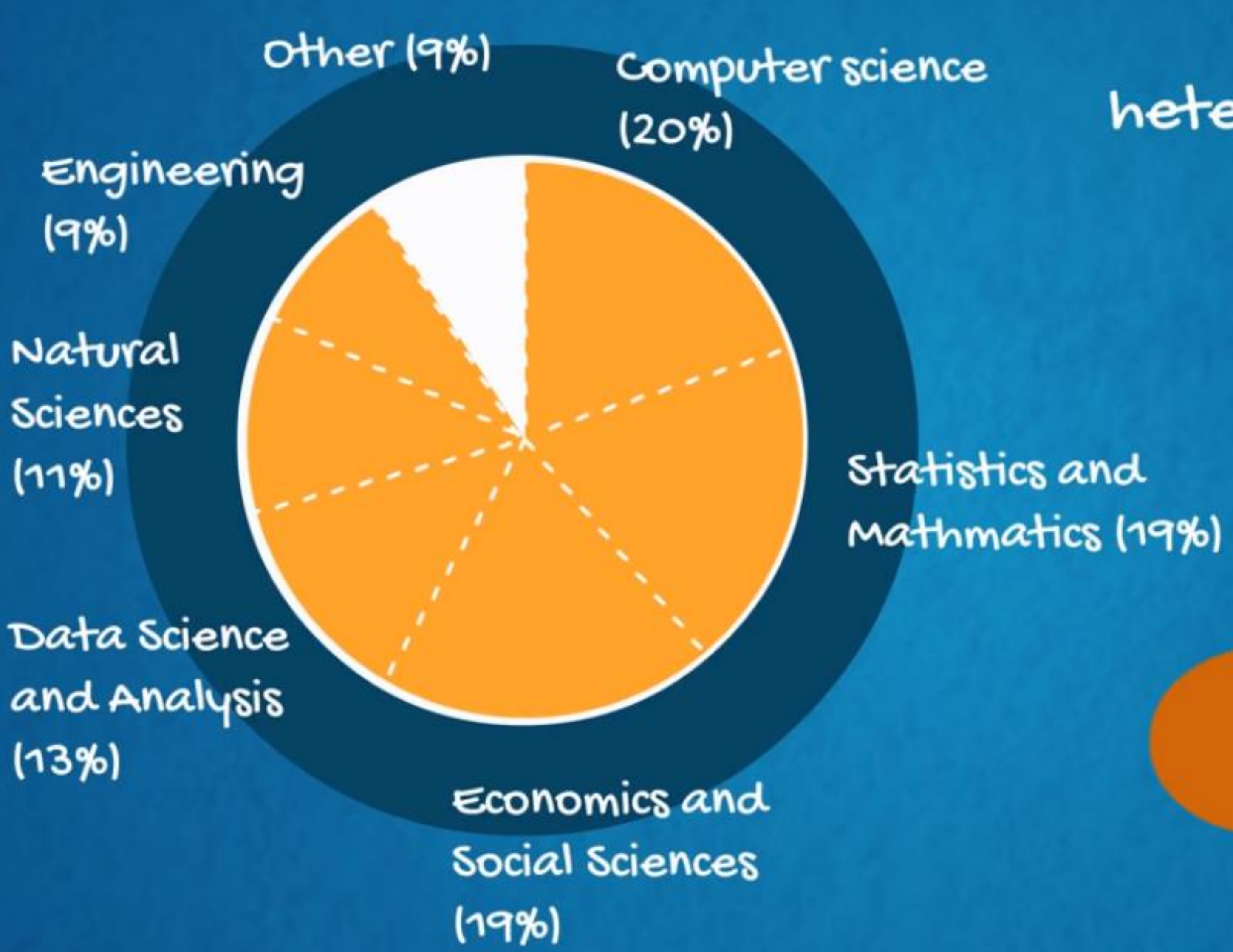
Will use programmes such as:

Excel, Tableau, SQL

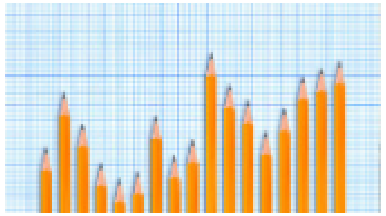
Current Data Scientists Profile



Data scientists have heterogeneous academic profiles



91%



What Data Scientists Really Do, According to 35 Data Scientists

Data Scientist: The Sexiest Job of the 21st Century

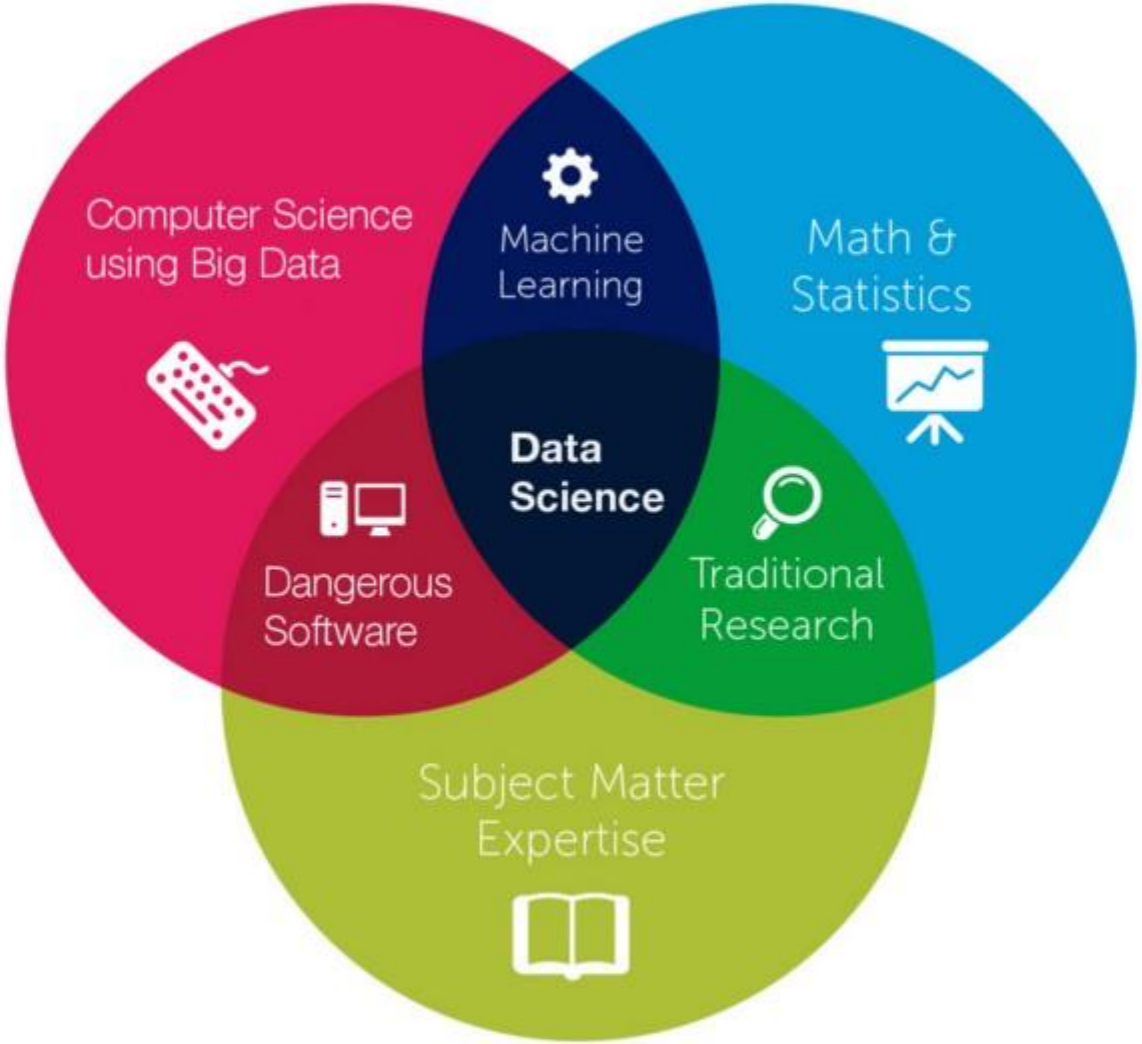
by Thomas H. Davenport and D.J. Patil

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

VIEW MORE FROM THE

October 2012 Issue



https://medium.com/@StepUpAnalytics/ai-vs-machine-learning-vs-deep-learning-vs-data-science-572b34452c3

Methods

Sentiment analysis

Time series analysis

Data mining

Multilevel modeling

Missing data imputations

Classification and clustering

Survival analysis

Pattern recognition

Principal component and factor analysis

AB testing

Machine learning

Forecasting

Propensity score matching

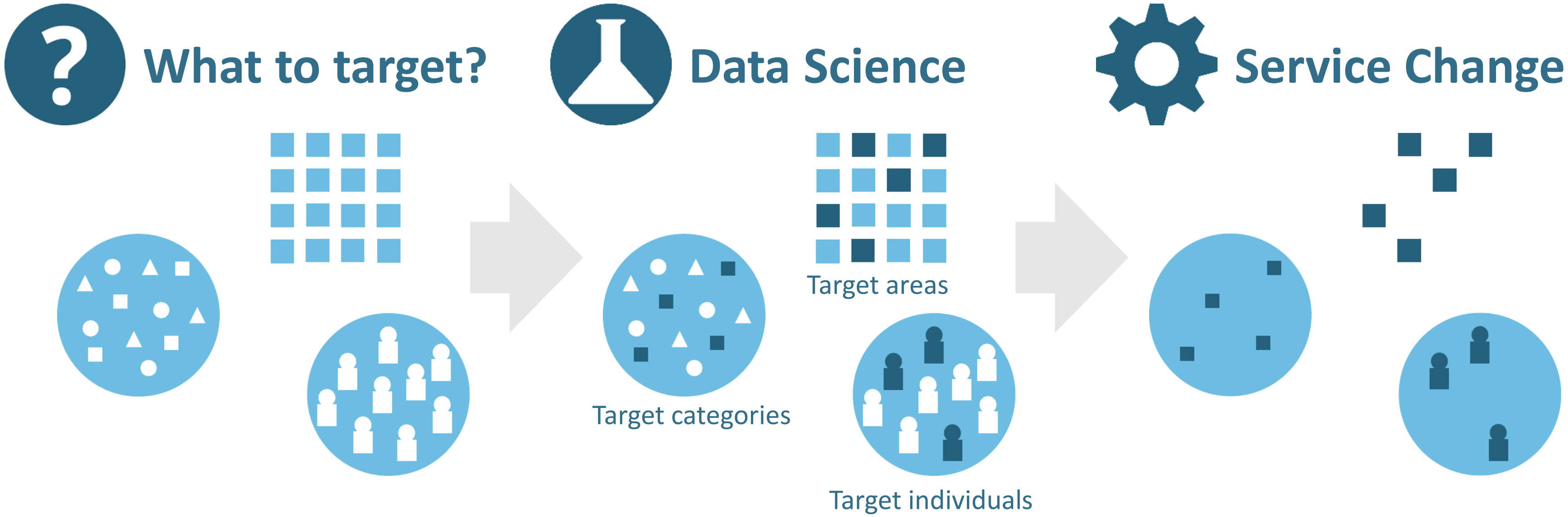
Logistic, multinomial and multiple linear regression techniques

Network analysis

Tools

Languages	Libraries	Data Engineering	Visualization
Python	SciPy	Profiling	D3.js
R	Pandas	ETL	Gephi
SQL	Scikit-learn	Job notices	R
Javascript	GPText	APIs	Leaflet
NodeJS	OpenNLP	Optimized data pipelines	PowerBI
	Mahout	Optimized data storage/access	ggplot2
	+many others		shiny

Targeting: Find the needle in the haystack



Service Issue:
Difficult to identify targets in a population

Data Science Process:
Use existing data and predictive modeling to identify targets

Service Change:
Engage with target subset of population

Result: Department resources are spent where most needed

Prioritizing

What to prioritize? | **Data Science** | **Service Change**

Service Issue:
Backlog is tackled via first in, first out (FIFO)

Data Science Process:
Create a model to categorize and group past and current cases

Service Change:
Prioritize cases based on categories in order of risk, need or opportunity

Result: Department addresses high priority cases first

Predictions

How to detect? | **Data Science** | **Service Change**

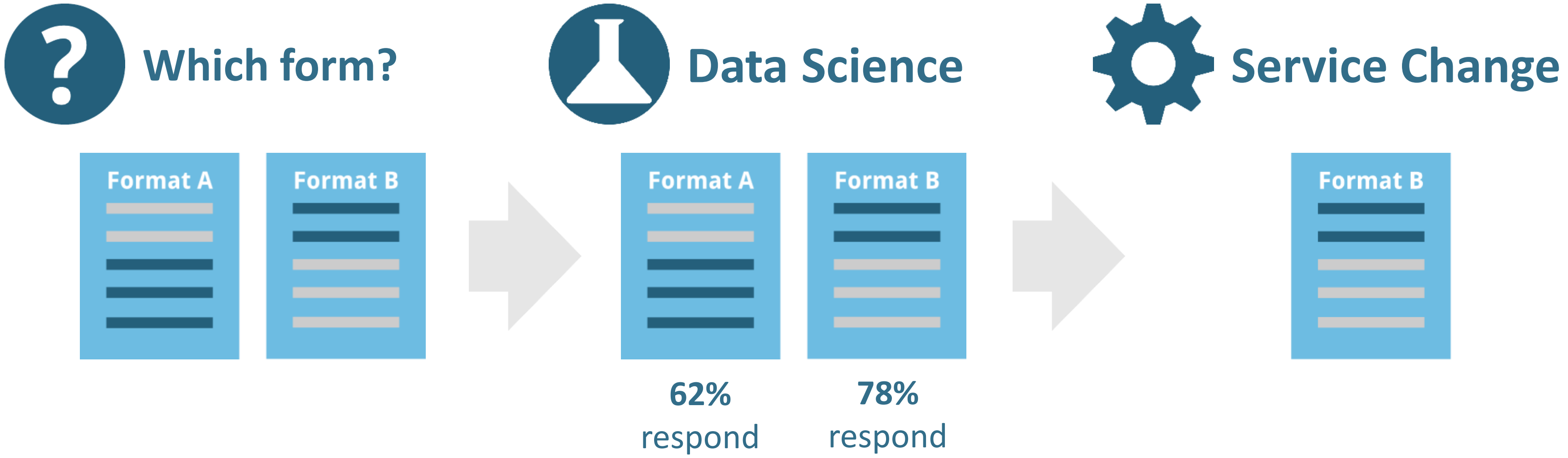
Service Issue:
Hard to predict future condition which leads to reactive services

Data Science Process:
Use historical and current data to create estimate ranges for potential outcomes

Service Change:
Use estimates to change and tailor intervention points

Result: Department provides pro-active early interventions

A/B test



Service Issue:
Costly outreach methods are not tested before implementation

Data Science Process:
Statistical testing on outreach methods to identify which, when, and to whom to send

Service Change:
Use statistically validated outreach method

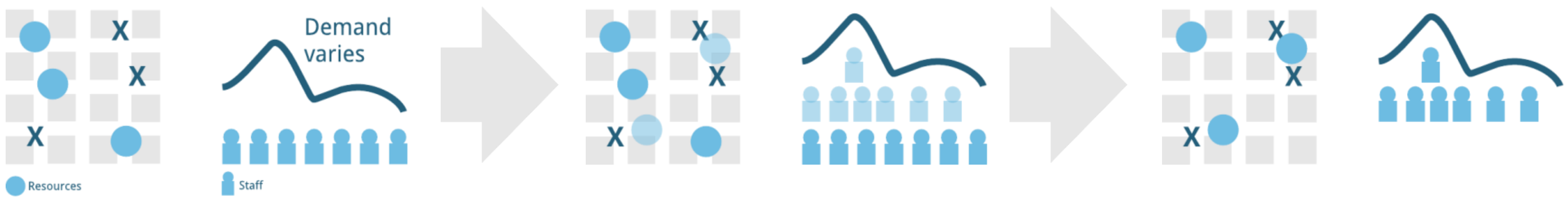
Result: Department increases response rates

Optimization

How to distribute?

Data Science

Service Change

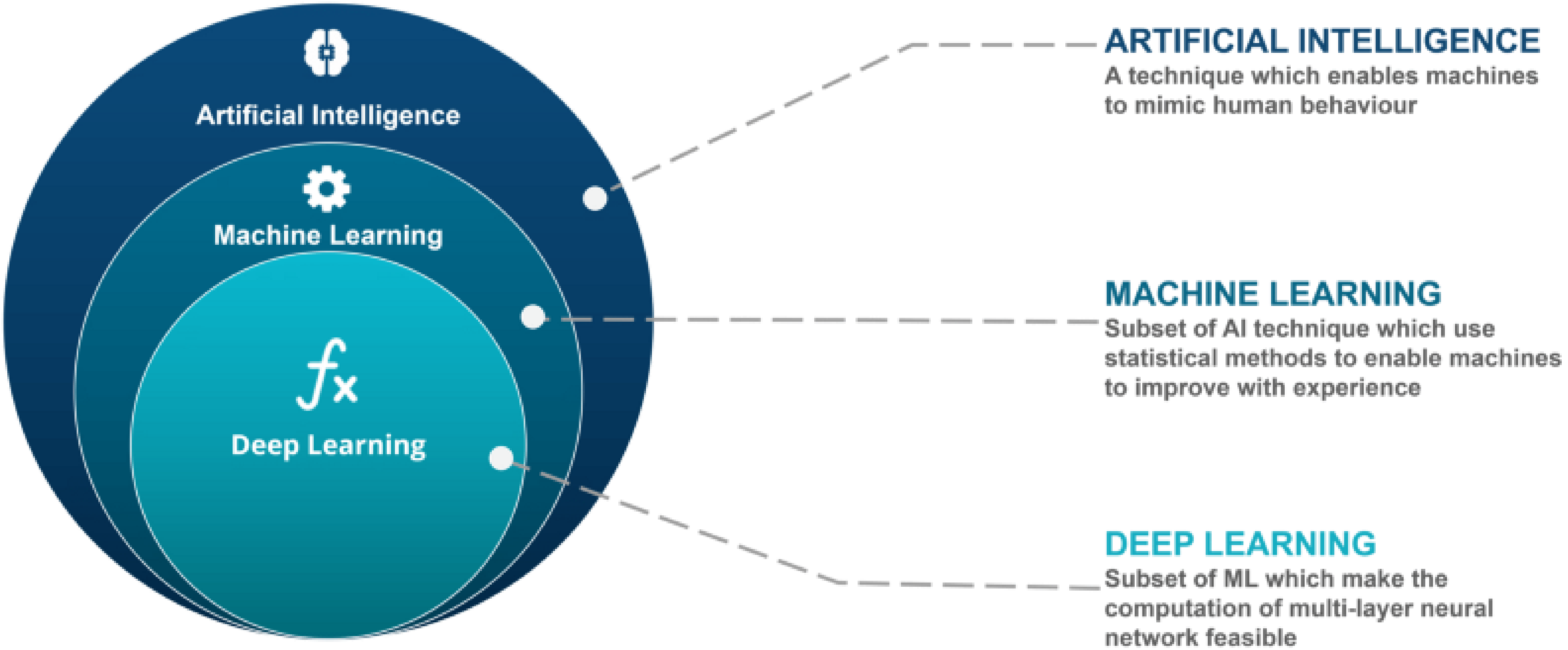


Service Issue:
Difficult to identify where to place or distribute resources to be most effective

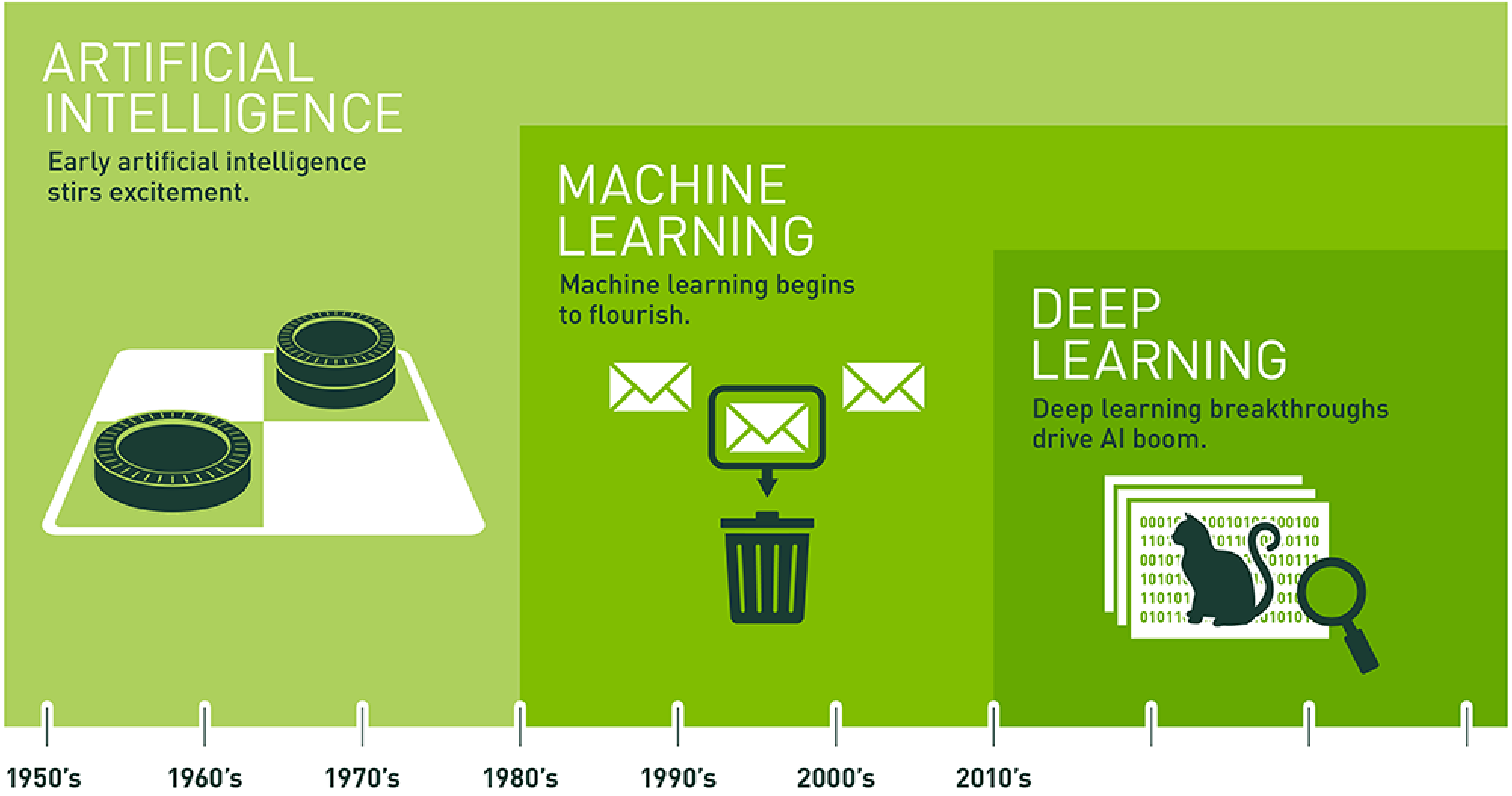
Data Science Process:
Use geospatial and/or other data to identify optimal distribution of resources

Service Change:
Re-allocates resources to optimal distribution

Result: Department decreases response times; increases volume



https://medium.com/@StepUpAnalytics/ai-vs-machine-learning-vs-deep-learning-vs-data-science-572b34452c3



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

https://www.datasciencecentral.com

Data Mining, AI and Machine Learning

- **Data Mining:** extract existing information to highlight patterns, and serves as foundation for AI and machine learning.
- **Artificial Intelligence:** creating machines that perform functions that require intelligence when performed by people.
- **Machine Learning:** Offers data necessary for a machine to learn & adapt. The machine must automatically learn the parameters of models from the data. It uses self-learning algorithms to improve its performance at a task with experience over time

AI in Sci-Fi Movies

- Terminator

http://starwars.com/

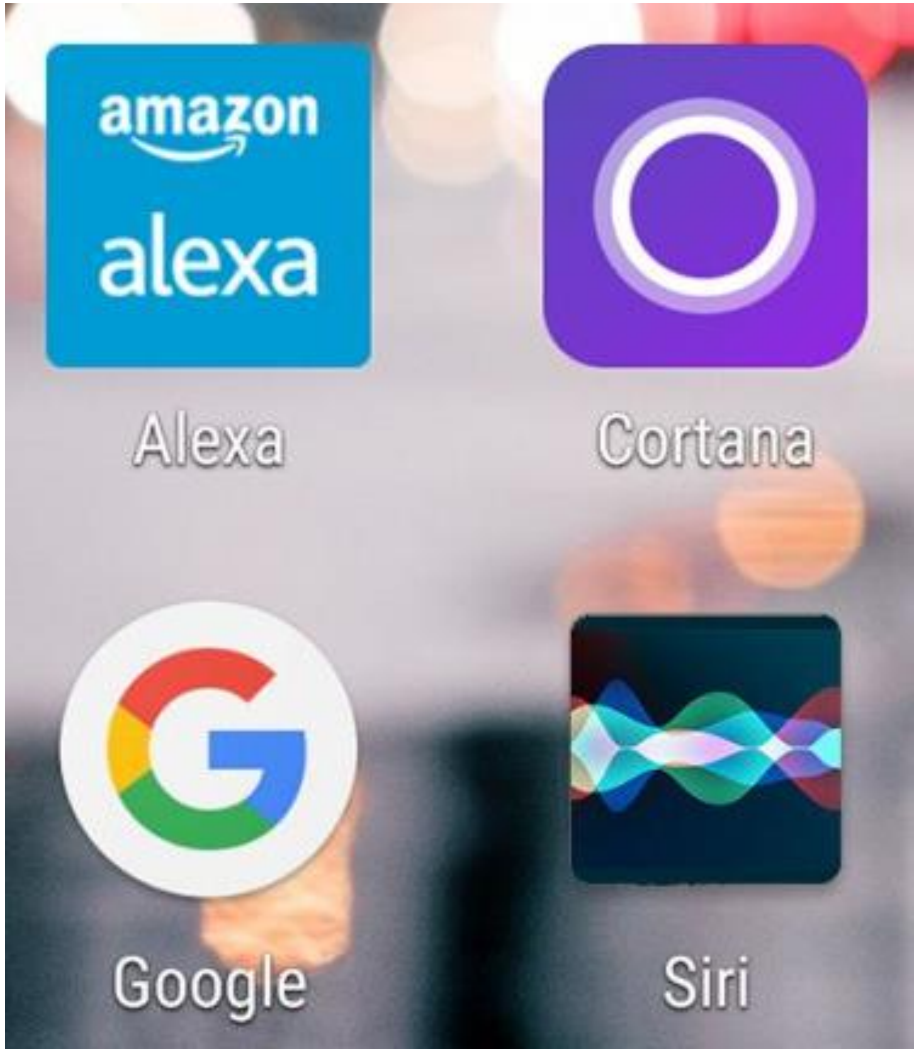
- Iron Man Marvel

Just A Rather Very Intelligent System

AI in Life

Vacuum Cleaning Robot

AI assistants



AI in Life

- Kiva warehouse robot

What is Artificial Intelligence ?

- The art of creating machines that perform functions that require intelligence when performed by people (Kurzweil, 1990)
- The study of how to make computers do things at which, at the moment, people are better (Rich and Knight, 1991)
- AI: acting humanly

Machine Learning

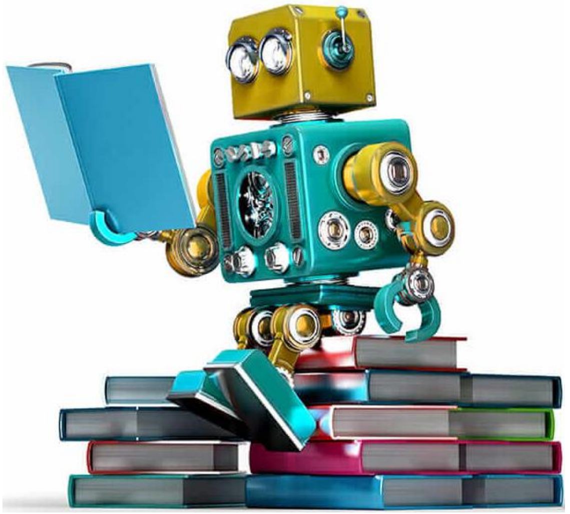
Learning from Experience

Learning from Data

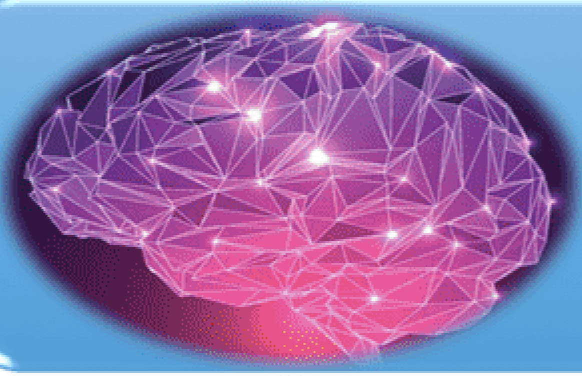
Follow Instructions

Machine Learning

- Machine learning is aimed to optimize a certain task using example data or past experience
- The extraction of knowledge from data
- Machine learning is preferred approach to
  - Business Intelligence
  - Speech recognition, Natural language processing
  - Computer vision
  - Robot control
  - Computational biology
  - Crime predictions
  - Etc..



Machine Learning & Some Use Cases



Machine Learning

Where business and experience meet emerging technology and decides to work together".

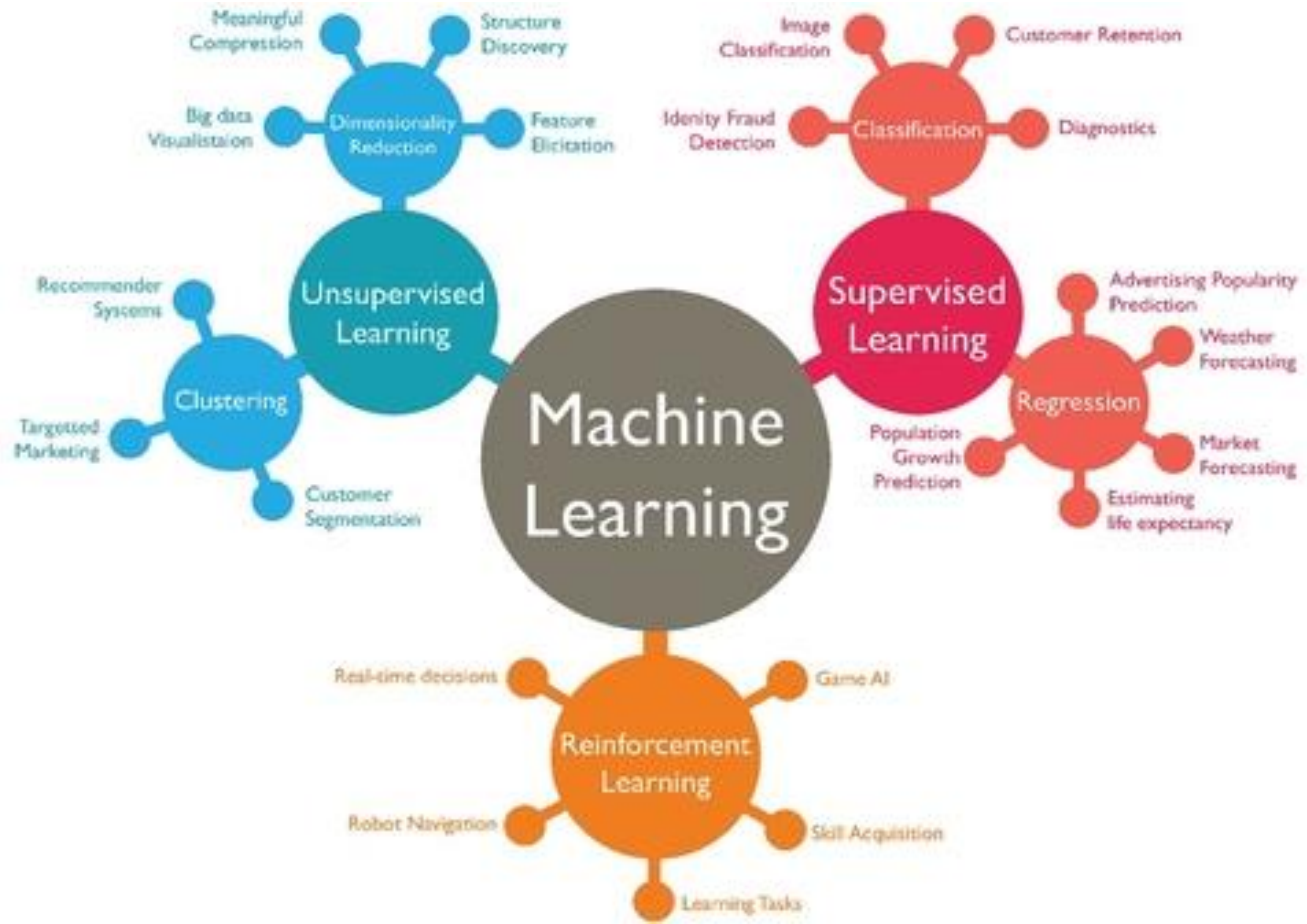
FB use it at extreme level for Spam control, discovering new content, recommendations and Ad sales

Google use it for many many reasons i.e. for maps, route calculations, data collection, translations, email spam and many more.

Credit Card Companies getting this now deeper and deeper to minimize the frauds and give safer transaction experience for customers

Types of Learning

- **Supervised (inductive) learning**
  - Training data includes desired outputs
- **Unsupervised learning**
  - Training data does not include desired outputs
- **Reinforcement learning**
  - Rewards from sequence of actions


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Methods

- **Supervised learning**
  - Decision tree induction
  - Rule induction
  - Naïve Bayes
  - Neural networks
  - Support vector machines
  - Model ensembles
  - Etc.

- **Unsupervised learning**
  - Clustering
  - Dimensionality reduction

- **Reinforcement learning**
  - Decision making (robot, chess machine)

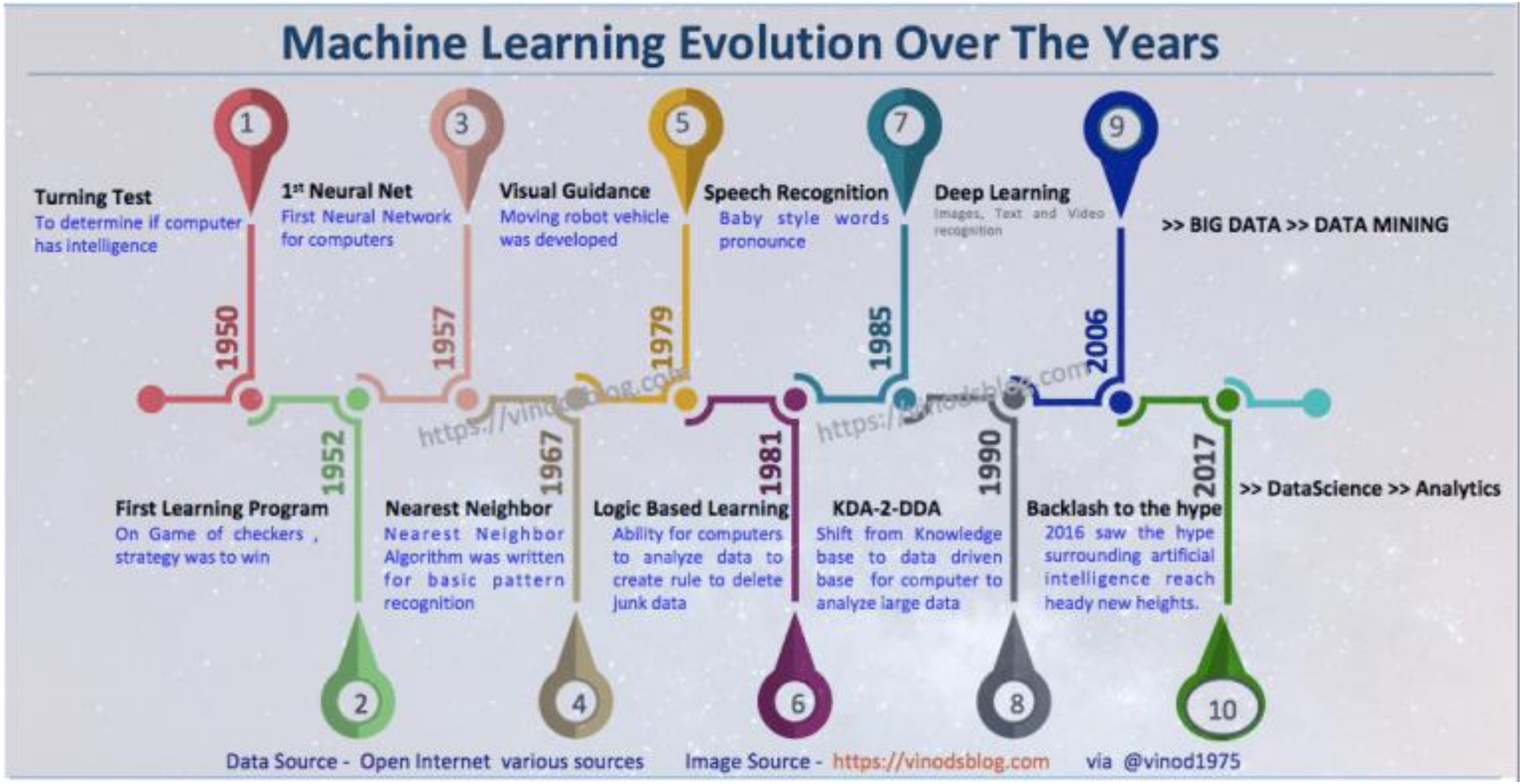
AI Magazine Volume 17 Number 3 (1996) (© AAAI)

From Data Mining to Knowledge Discovery in Databases

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth

■ Data mining and knowledge discovery in databases have been attracting a significant amount of research, industry, and media attention of late. What is all the excitement about? This article provides an overview of this emerging field, clarifying how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. The article mentions particular real-world applications, specific data-mining techniques, challenges involved in real-world applications of knowledge

This article begins by discussing the historical context of KDD and data mining and their intersection with other related fields. A brief summary of recent KDD real-world applications is provided. Definitions of KDD and data mining are provided, and the general multistep KDD process is outlined. This multistep process has the application of data-mining algorithms as one particular step in the process. The data-mining step is discussed in more detail in the context of specific data-mining al-

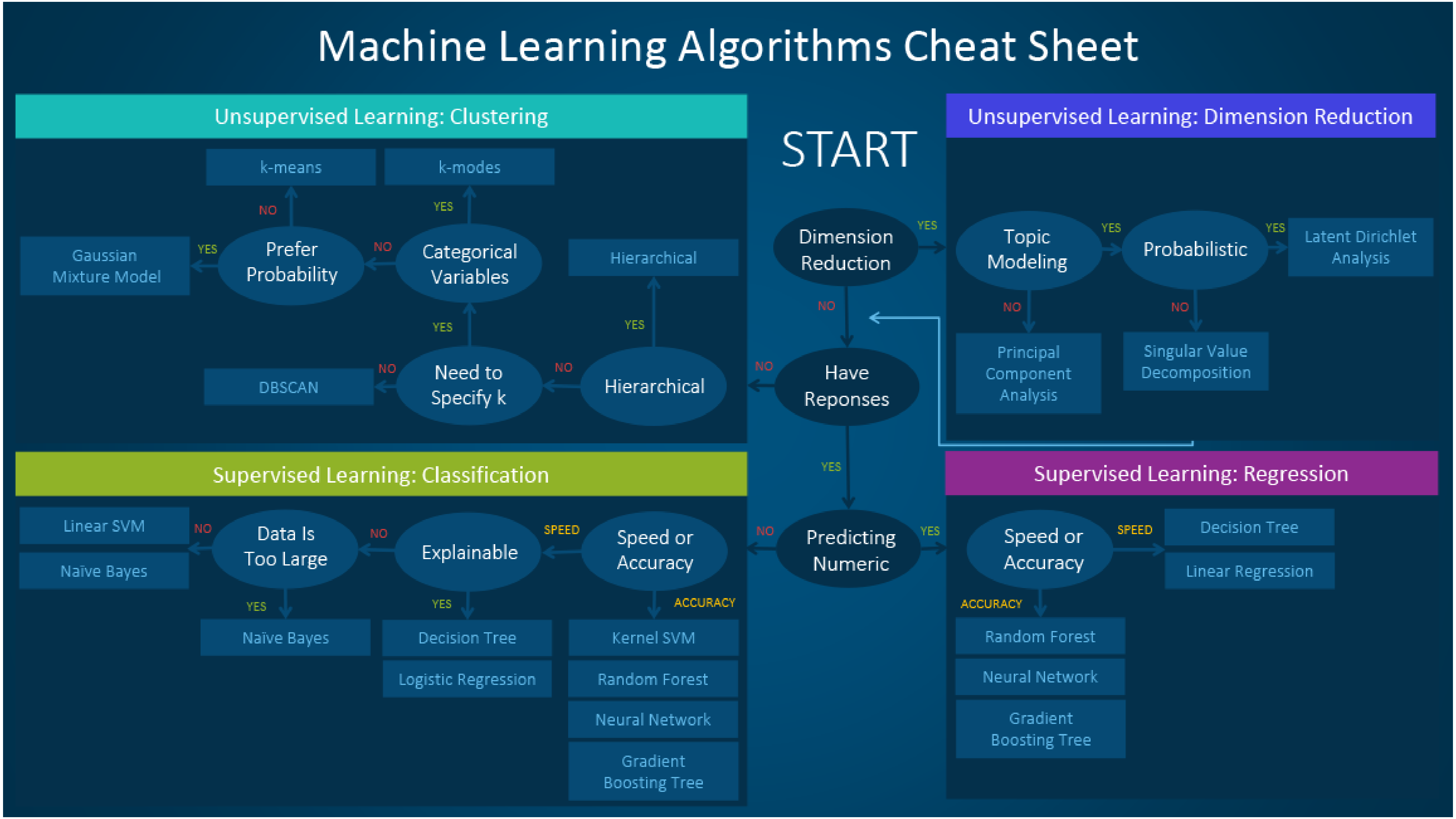

Machine Learning Evolution Over The Years
1
Turning Test
To determine if computer has intelligence
1950
2
1952
First Learning Program
On Game of checkers , strategy was to win
3
1st Neural Net
First Neural Network for computers
1957
4
1967
Nearest Neighbor
Nearest Neighbor Algorithm was written for basic pattern recognition
5
Visual Guidance
Moving robot vehicle was developed
1979
6
1981
Logic Based Learning
Ability for computers to analyze data to create rule to delete junk data
7
Speech Recognition
Baby style words pronounce
1985
8
1990
KDA-2-DDA
Shift from Knowledge base to data driven base for computer to analyze large data
9
Deep Learning
Images, Text and Video recognition
2006
>> BIG DATA >> DATA MINING
10
2017
Backlash to the hype
2016 saw the hype surrounding artificial intelligence reach heady new heights.
>> DataScience >> Analytics
https://vinodsblog.com
Data Source - Open Internet various sources
Image Source - https://vinodsblog.com
via @vinod1975

Business Intelligence in Banking

- Customer account data and demographics
- Core banking data
- Transactional data at every level of detail
- Wire and payment data
- Trade and position data
- General ledger data including accounts payable, accounts receivable, cash management, purchasing information
- Support data from banking reporting

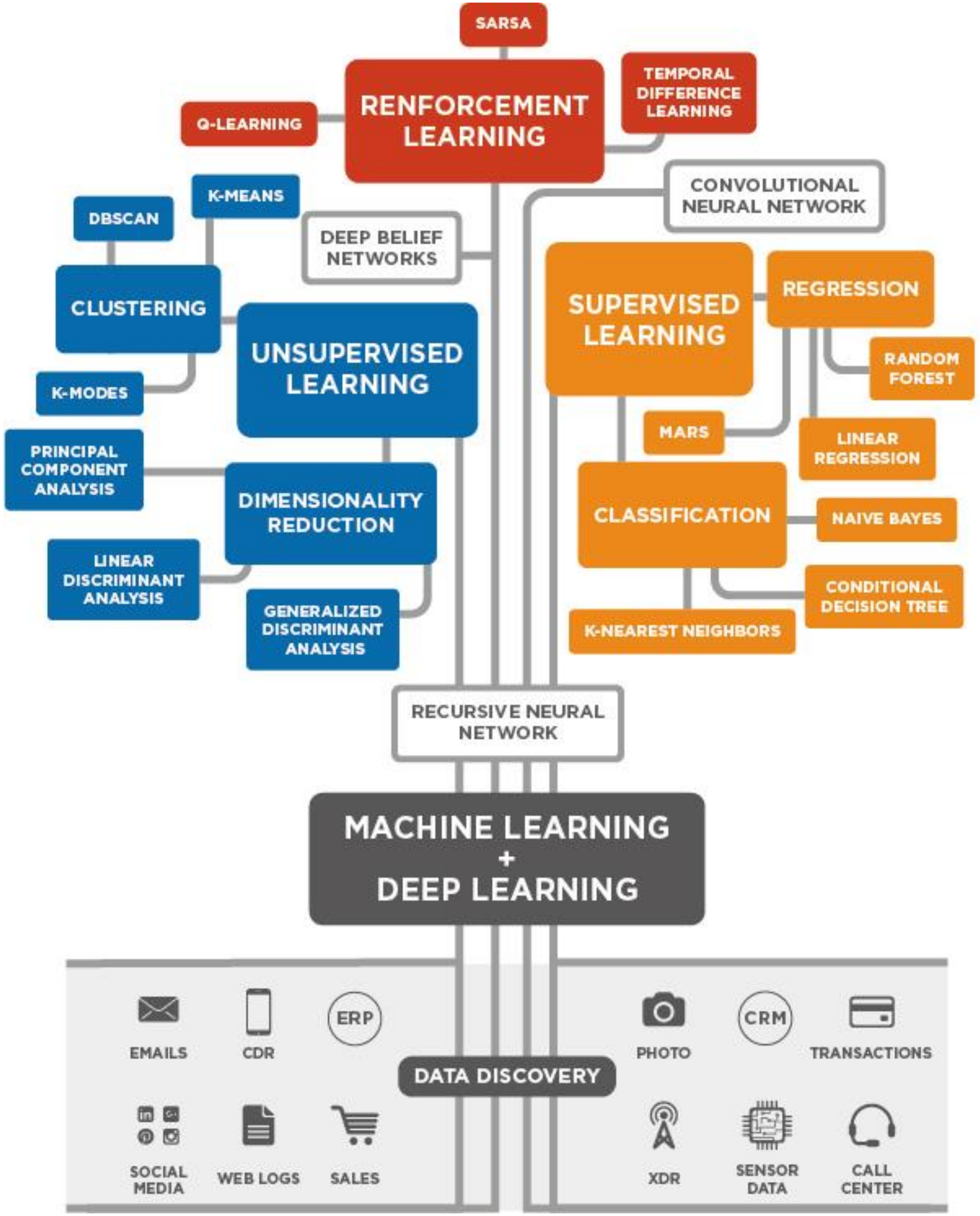
Machine Learning in Finance

- Fraud prevention
- Portfolio and Risk Management
- Investment predictions
- Customer service
- Digital assistants
- Marketing
- Network security
- Loan underwriting
- Algorithmic trading
- Customer Service (Chatbot)
- Process automation
- Document interpretation
- Content creation
- Trade settlements
- Money-laundering prevention
- Custom machine learning solutions
- Sales/Recommendations of Financial Products
- Sentiment/News Analysis



Machine Learning Algorithms Cheat Sheet
Unsupervised Learning: Clustering
k-means
k-modes
NO
YES
Gaussian Mixture Model
YES
Prefer Probability
NO
Categorical Variables
Hierarchical
YES
YES
DBSCAN
NO
Need to Specify k
NO
Hierarchical
START
Dimension Reduction
YES
NO
Have Reponses
NO
YES
Predicting Numeric
NO
YES
Unsupervised Learning: Dimension Reduction
Topic Modeling
YES
Probabilistic
YES
Latent Dirichlet Analysis
NO
NO
Principal Component Analysis
Singular Value Decomposition
Supervised Learning: Classification
Linear SVM
Naïve Bayes
NO
Data Is Too Large
NO
Explainable
SPEED
Speed or Accuracy
YES
YES
ACCURACY
Naïve Bayes
Decision Tree
Logistic Regression
Kernel SVM
Random Forest
Neural Network
Gradient Boosting Tree
Supervised Learning: Regression
Speed or Accuracy
SPEED
Decision Tree
Linear Regression
ACCURACY
Random Forest
Neural Network
Gradient Boosting Tree

Numerous (New) Algorithms

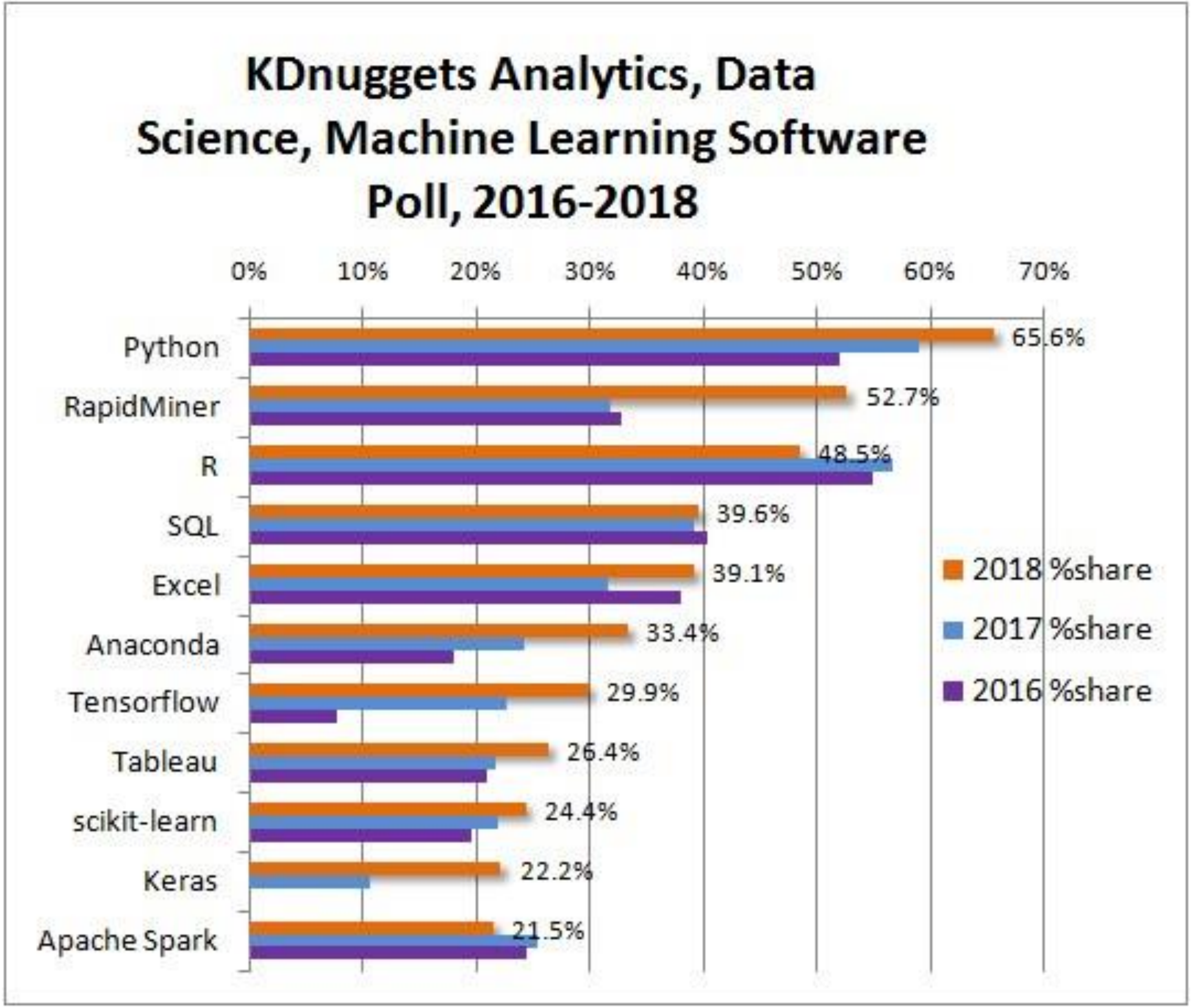


Numerous (New) Algorithms

Classification algorithms considered in the benchmarking study.

	BM selection	Classification algorithm	Acronym
Individual classifier	n.a.	Bayesian Network	B-Net
		CART	CART
		Extreme learning machine	ELM
		Kernalized ELM	ELM-K
		k-nearest neighbor	kNN
		J4.8	J4.8
		Linear discriminant analysis[a]	LDA
		Linear support vector machine	SVM-L
		Logistic regression[a]	LR
		Multilayer perceptron artificial neural network	ANN
		Naive Bayes	NB
		Quadratic discriminant analysis[a]	QDA
		Radial basis function neural network	RbfNN
		Regularized logistic regression	LR-R
		SVM with radial basis kernel function	SVM- Rbf
		Voted perceptron	VP
Classification models from individual classifiers			**16**
Homogenous ensembles	n.a.	Alternating decision tree	ADT
		Bagged decision trees	Bag
		Bagged MLP	BagNN
		Boosted decision trees	Boost
		Logistic model tree	LMT
		Random forest	RF
		Rotation forest	RotFor
		Stochastic gradient boosting	SGB

Heterogeneous ensembles	n.a.	Simple average ensemble	AvgS
		Weighted average ensemble	AvgW
		Stacking	Stack
	Static direct	Complementary measure	CompM
		Ensemble pruning via reinforcement learning	EPVRL
		GASEN	GASEN
		Hill-climbing ensemble selection	HCES
		HCES with bootstrap sampling	HCES-B
		Matching pursuit optimization ensemble	MPOE
		Top-*T* ensemble	Top-*T*
	Static indirect	Clustering using compound error	CuCE
		k-Means clustering	k-Mean
		Kappa pruning	KaPru
		Margin distance minimization	MDM
		Uncertainty weighted accuracy	UWA
	Dynamic	Probabilistic model for classifier competence	PMCC
		k-nearest oracle	kNORA
Classification models from heterogeneous ensembles			**17**


KDnuggets Analytics, Data Science, Machine Learning Software Poll, 2016-2018
0%
10%
20%
30%
40%
50%
60%
70%
Python
65.6%
RapidMiner
52.7%
R
48.5%
SQL
39.6%
Excel
39.1%
Anaconda
33.4%
Tensorflow
29.9%
Tableau
26.4%
scikit-learn
24.4%
Keras
22.2%
Apache Spark
21.5%
2018 %share
2017 %share
2016 %share

Next, we learn R… and Python….