Data Science

Saurabh Jain

About Me

Who	Saurabh Jain
Where	Persistent Systems
What	Data Scientist, Enterprise Applications
How	MS in Data Science, Indiana University
	BTech in CS, Visvesvaraya National Institute of Technology
When	Since 12+ years

Aim

Entice you into Data Science as a career

Flow

Why to discuss Data Science

What is Data Science

The Fourth Paradigm

Four V's of Data

Data Science Umbrella, Technologies, Lifecycle

How machines learn, AI, ML, DL

Machine Learning – Intro

Applications of Data Science, ML

Learnings from Industry

Learn Data Science

What Big Guns are talking about Data Science

Data Scientist: The Sexiest Job of the 21st Century – HBR 2012

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

Second Fastest growing Profession – LinkedIn 2017

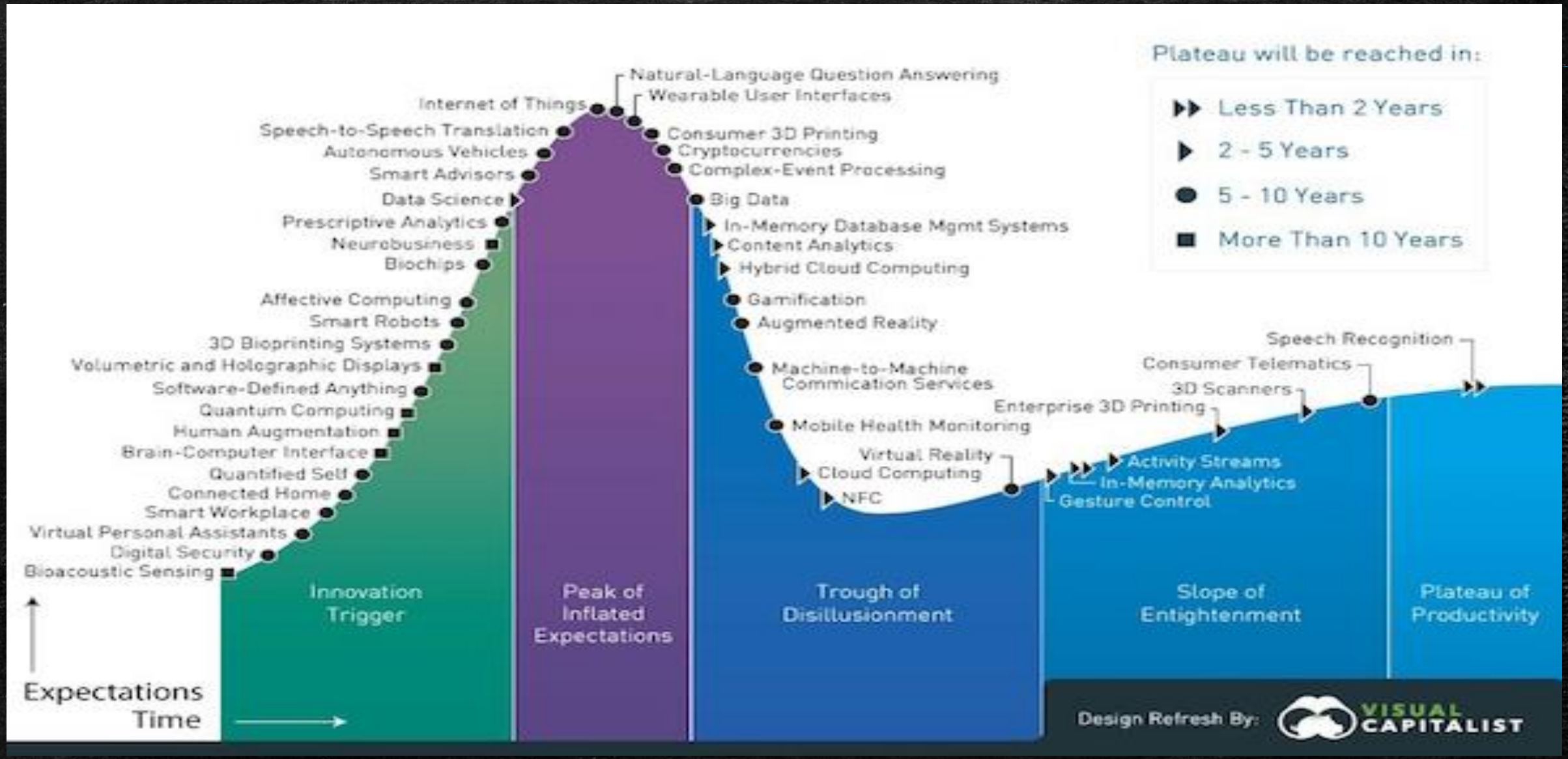
https://blog.linkedin.com/2017/december/7/the-fastest-growing-jobs-in-the-u-s-based-on-linkedin-data

Which is at Top ☺ ?

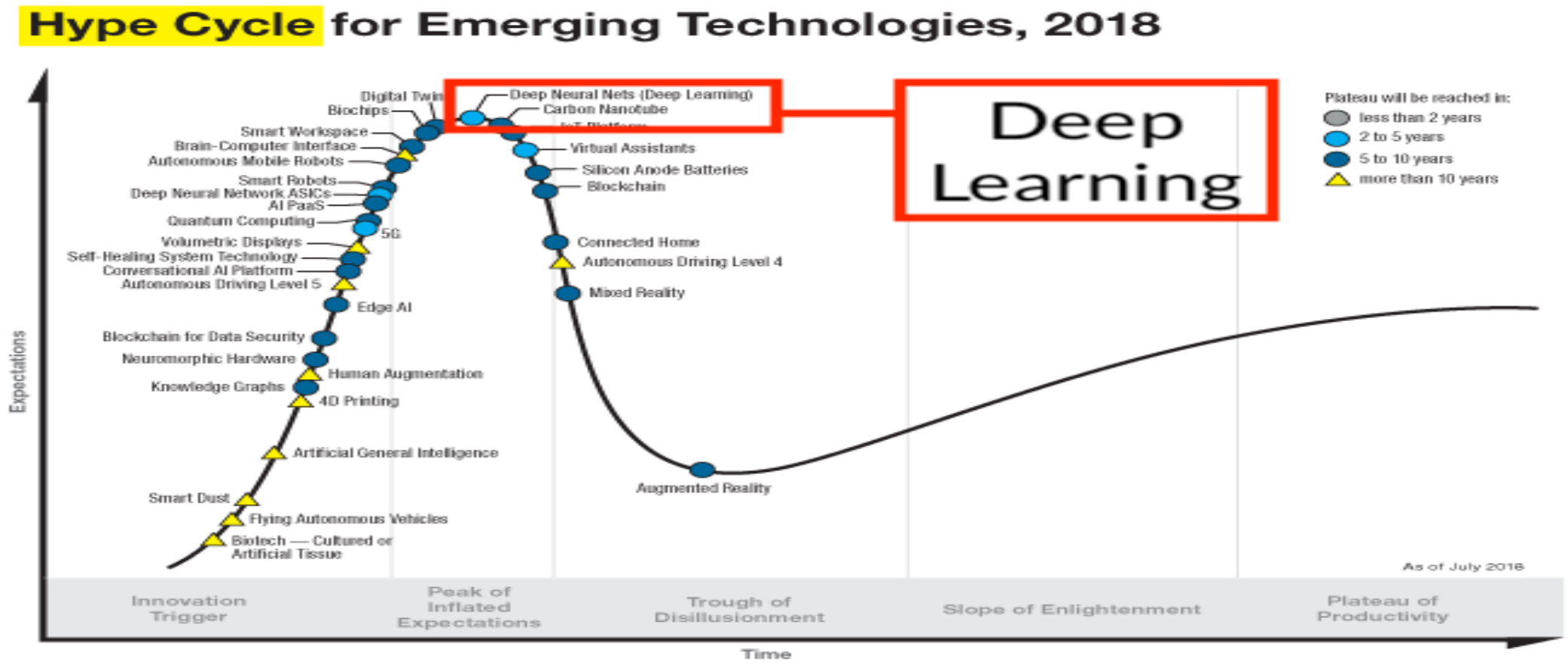
50 percent gap in the supply – Mckinsey, Amazon

https://blog.alexa.com/know-data-science-important/

Why Data Science? – Gartner Hype Cycle



Why Data Science?



What is Data Science

Poll

https://pollev.com/saurabhjain220

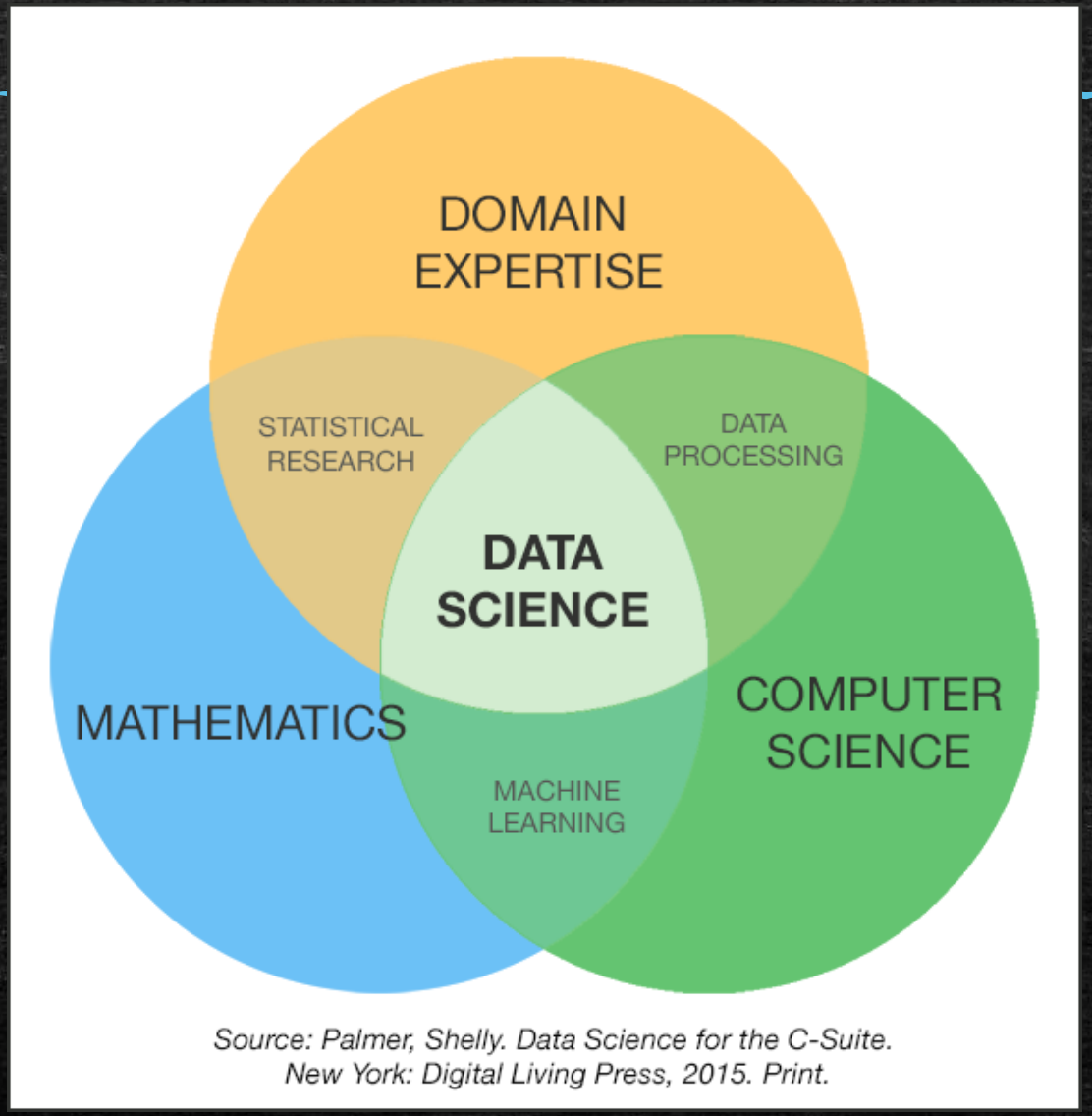
Data Science - Definition

- Applying advanced statistical tools to existing data to generate new insights
- Data science is a multidisciplinary blend of data inference, algorithm development, and technology in order to solve analytically complex problems
- Wikipedia - Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured,[1][2] similar to data mining

DIKW

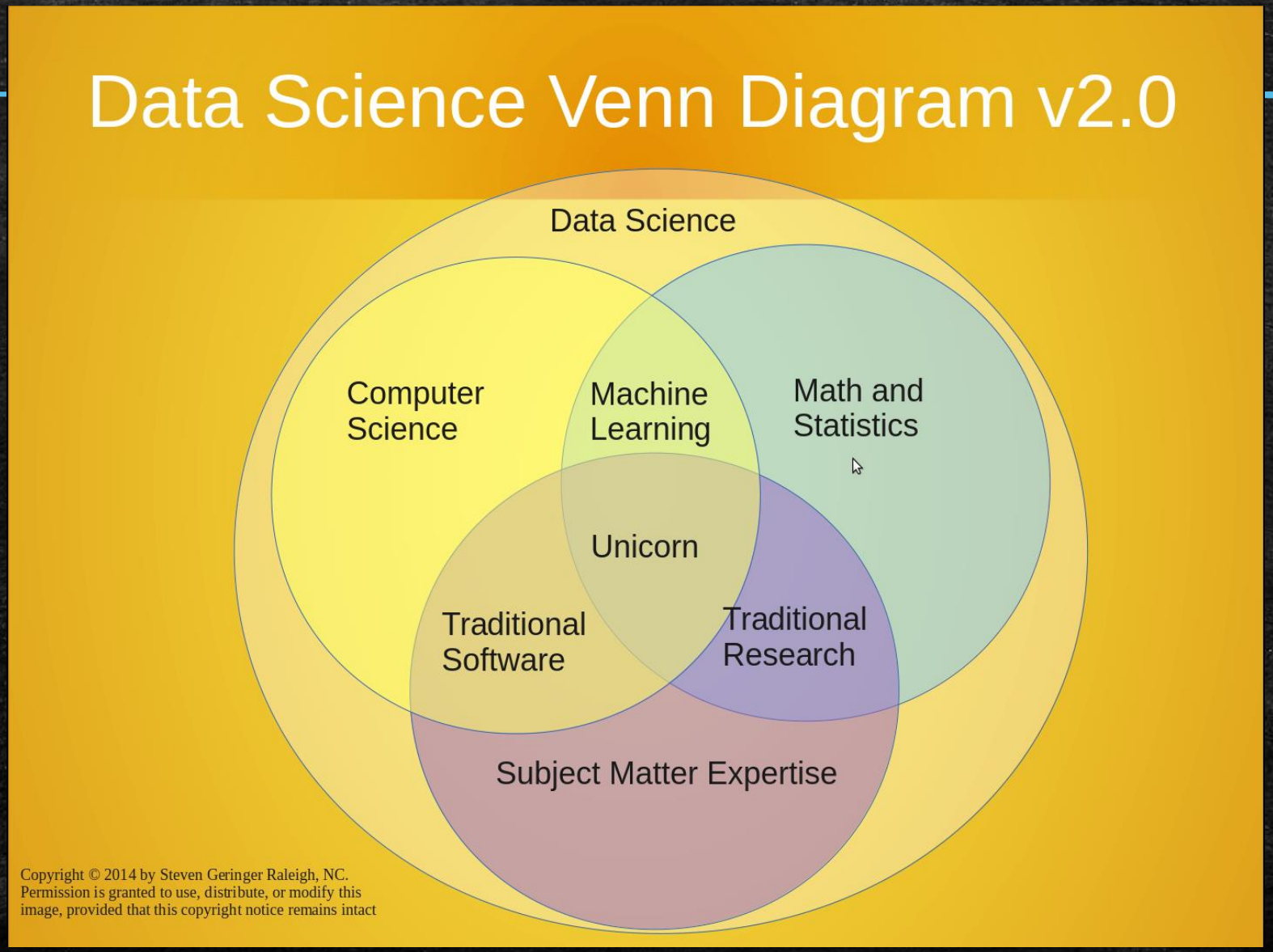
- Datum (singular of data) (an observation): 12
- Information (data in context): 12 degrees Fahrenheit
- Knowledge (information in context): 12 degrees Fahrenheit, today, at 7:30 AM, in Summerville, Oregon
- Wisdom (application of knowledge in context): I need to put on a coat when I go out.

Data Science?



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

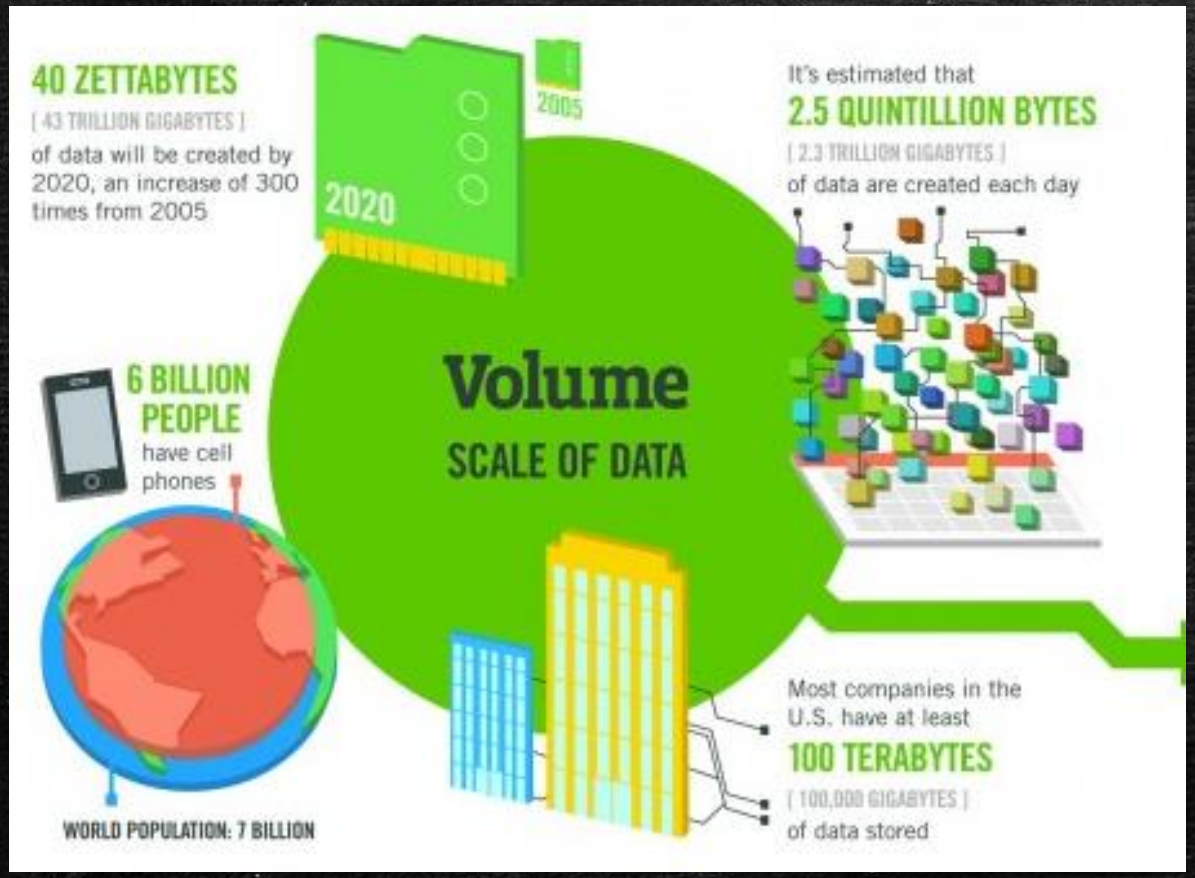
Data Science?

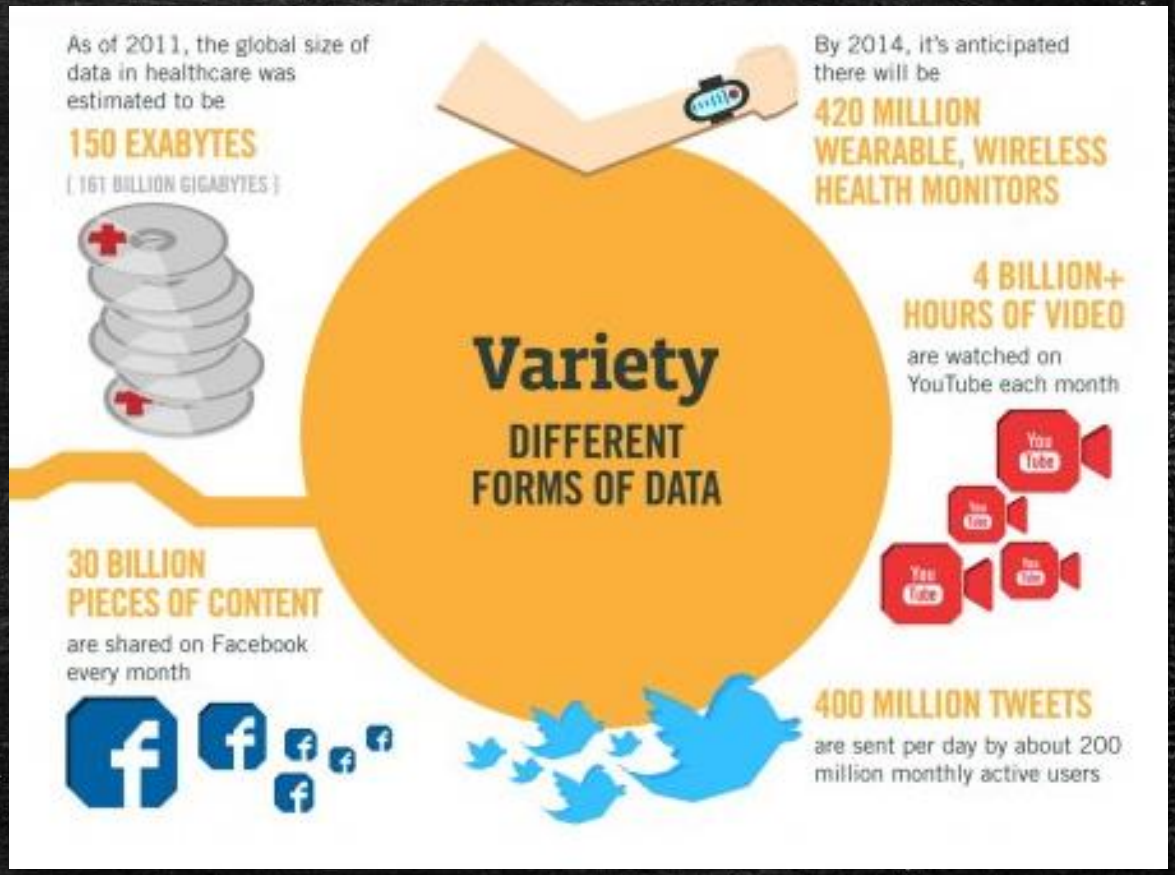


The Fourth Paradigm

- Thousand Years Ago – Empirical, Experimental
  - Describing natural phenomenon, Fire
- Last Few hundred years – Theoretical
  - Newtons Laws, Generalizations
- Last few decades – Computational Science
  - Simulation of complex Phenomena, Predict Global warning?
- Today – Data Science
  - Science based, data intensive computing
  - Scientists overwhelmed with vast datasets
    - From instruments, IOT
    - From simulations
    - From networks
    - From Users i.e. Google, Facebook

Explosion of data - Data is New Oil





Source: https://www.ibmbigdatahub.com/infographic/four-vs-big-data

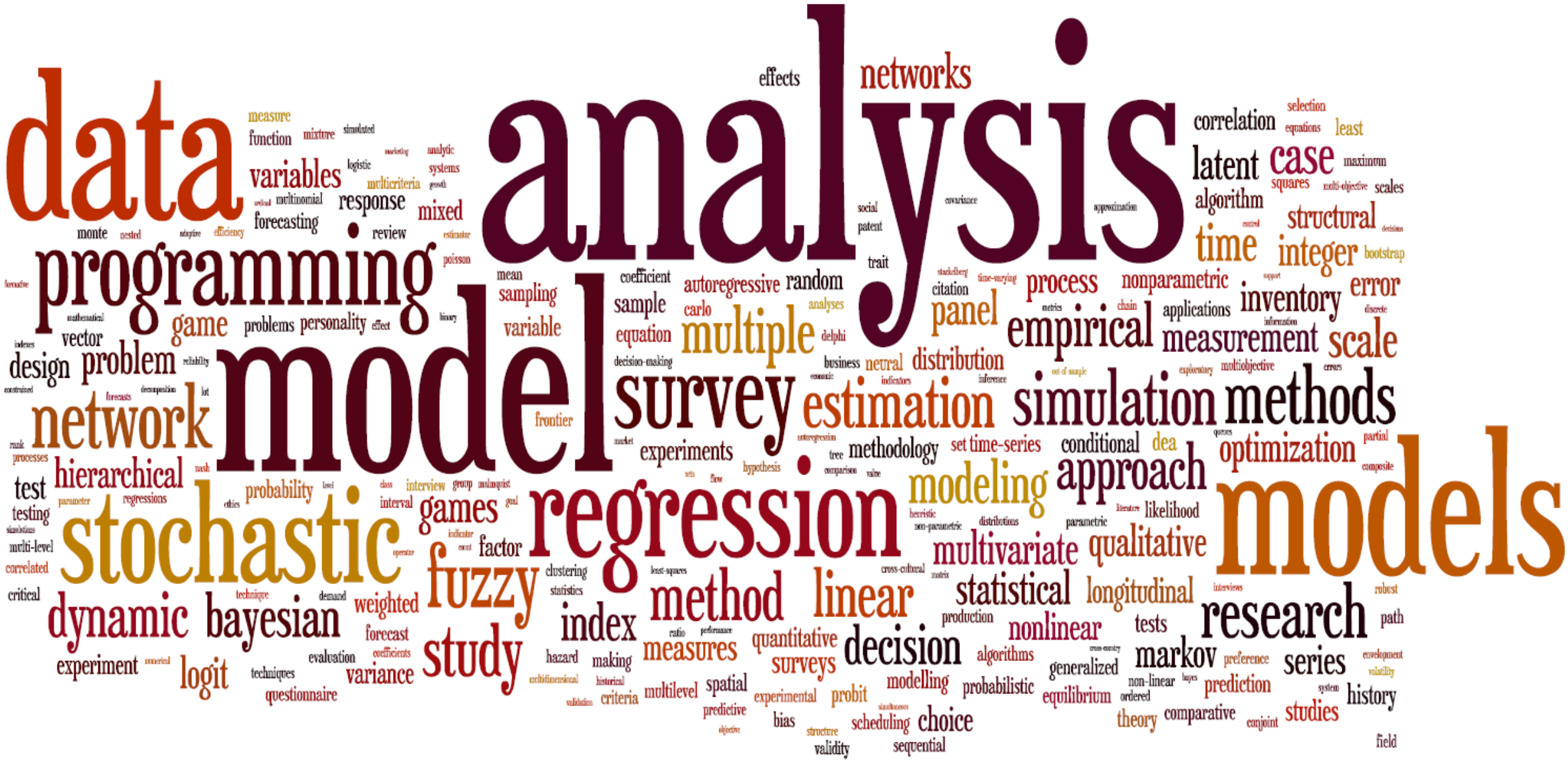
Four V's of Data

- Volume
  - Scale of data
- Variety
  - Different forms of data
- Velocity
  - Streaming data, Data Flow
- Veracity
  - Uncertainty, Quality of data

What is the need for Machines?

- Massive amount of data to process
- Computers don't fight
- Computers don't have prejudice
- Computers don't have bias
- No meetings

Word Cloud for Data Science



Data Science Umbrella

Data Analytics
Big Data
Machine Learning
Artificial Intelligence
Statistics
Network Theory
Visualization
Cloud
Data Engineering
Predictive Analytics
Stream Analytics
Deep Learning

Information Engineering
Data Warehouse
Data Mining
Data Wrangling
Mathematics
IOT
Business Analytics
No SQL Database
Data Virtualization
Distributed Storage
Knowledge Discovery Tools
Computer Science

Toolkit

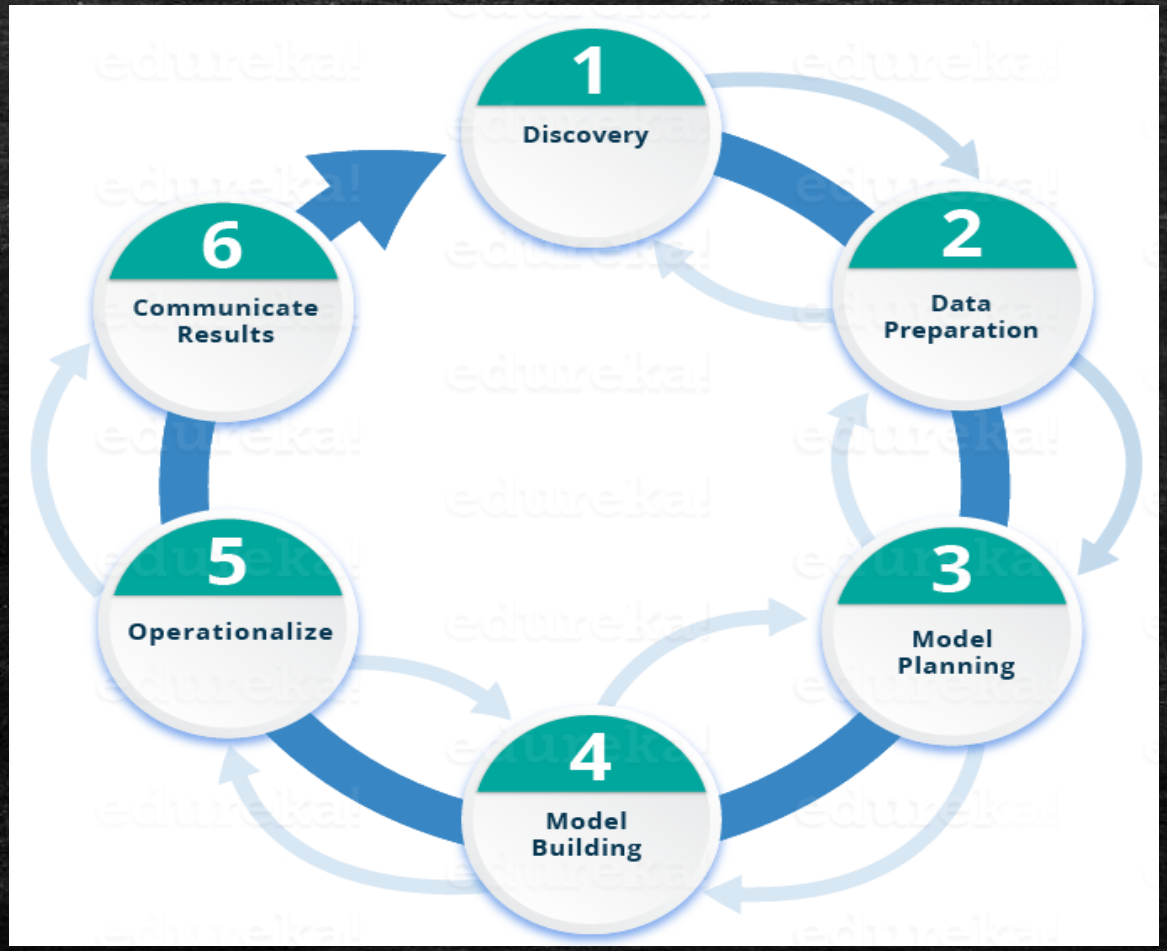
Languages
Python
R
SQL

Libraries
Tensorflow
Keras
Pandas
Scikit-learn
OpenNLP
Spacy
Stanford NLP

+many others

Visualization
D3.js
Gephi
R
PowerBI
ggplot2
Shiny
Tableau

Lifecycle



- Source: https://www.edureka.co/blog/what-is-data-science/

How do Humans learn?

- Consider a child

- Scenario
  - Frequent high volume tasks –
    - Grading an essay
    - Sorting mails
    - Spam vs Non Spam
    - Predicting weather

How do Machines learn?

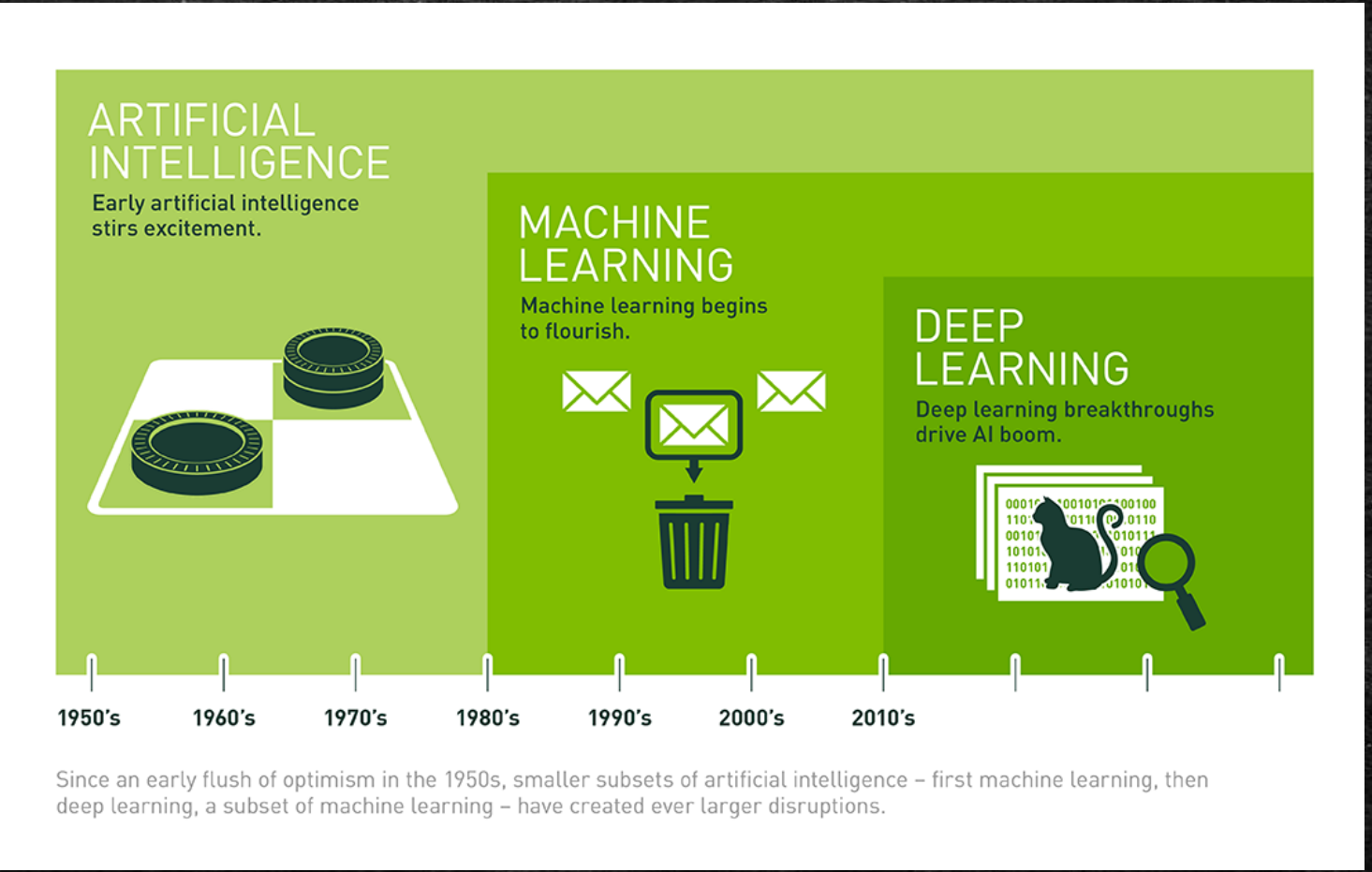
- Artificial Intelligence
- Machine Learning
- Deep Learning

- Requirement - **Large volume of past data**

AI vs ML vs DL

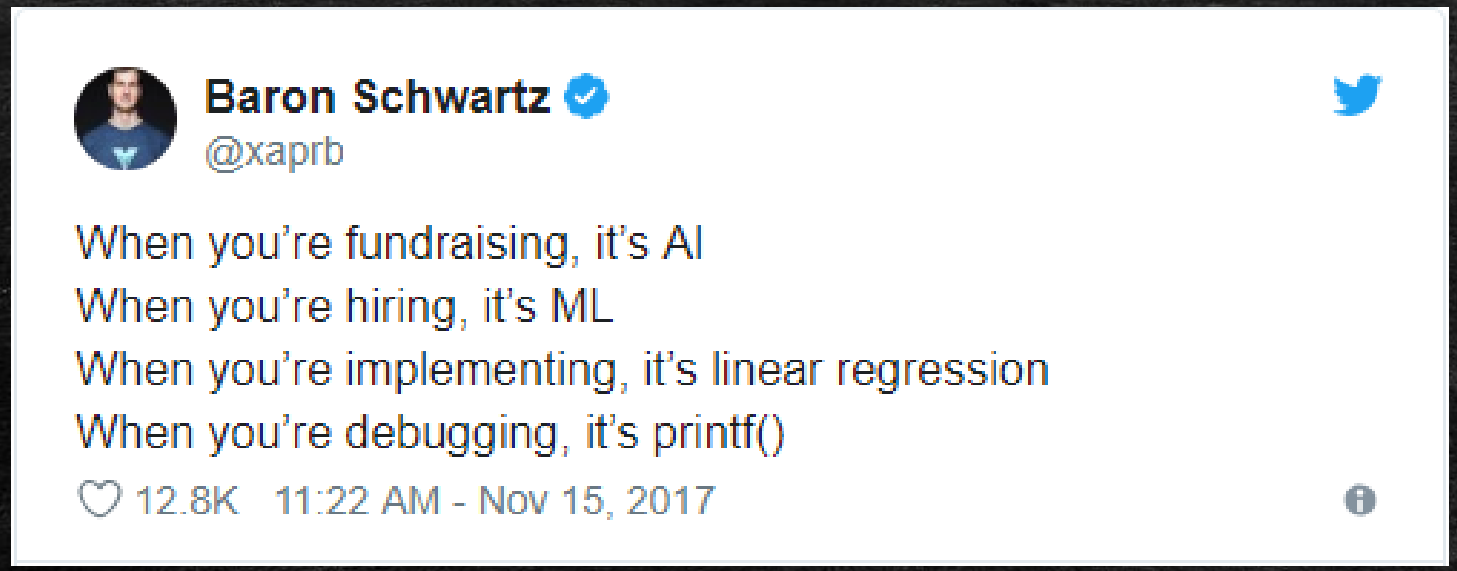
- Artificial Intelligence –
  - Human Intelligence Exhibited by Machines
  - A technique which enables machines to mimic human behaviour
- Machine Learning
  - An Approach to Achieve Artificial Intelligence
  - Programs that alter themselves
  - Subset of AI technique which use statistical emthods to enable machines to improve with experience
- Deep Learning - A Technique for Implementing Machine Learning
  - Subset of ML which make the computation of multi layer neural network feasible

AI vs ML vs DL



- Source: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

AI vs ML vs DL



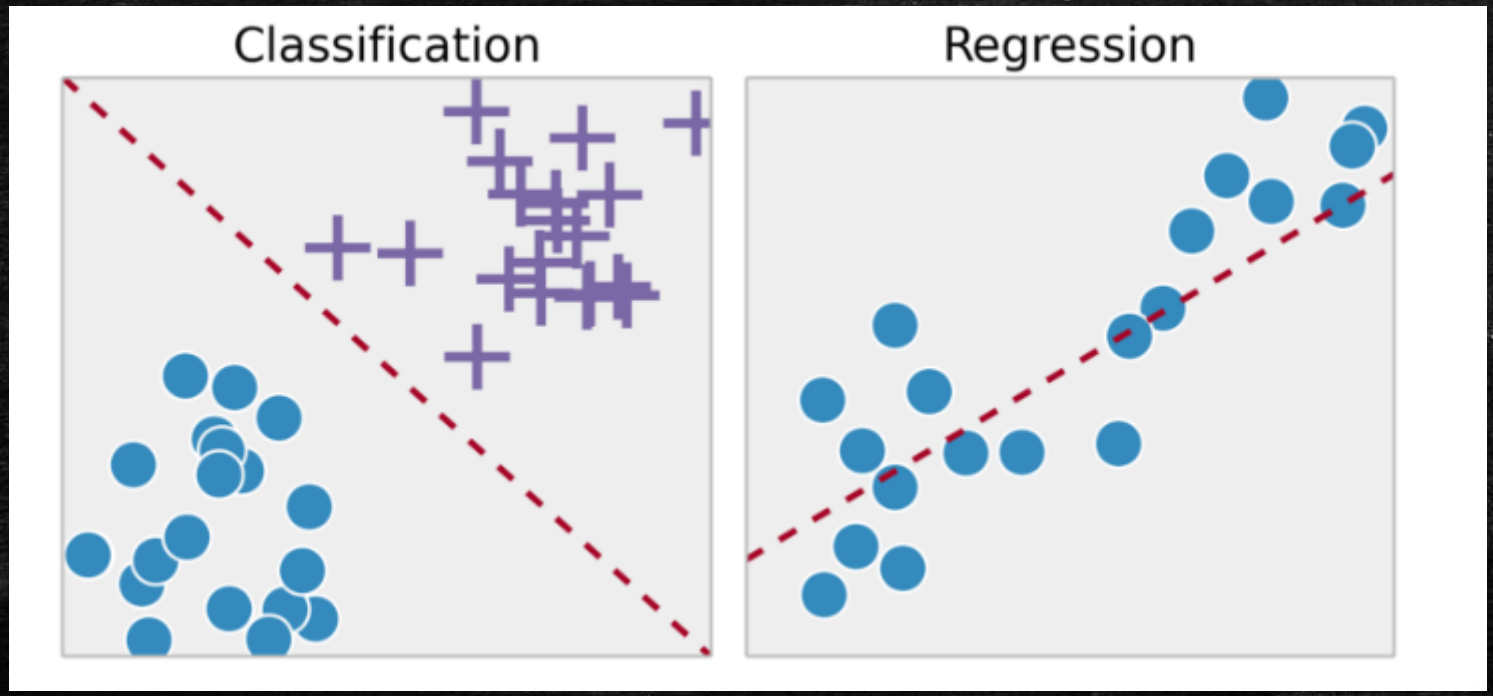
- Source: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Machine Learning types

- **Supervised Learning**
  - *The outcome or output for the given input is known before itself*
- **Unsupervised Learning**
  - *The outcome or output for the given inputs is unknown*
- **Reinforcement Learning**
  - The machine is exposed to an ***environment where it gets trained by trial and error method***

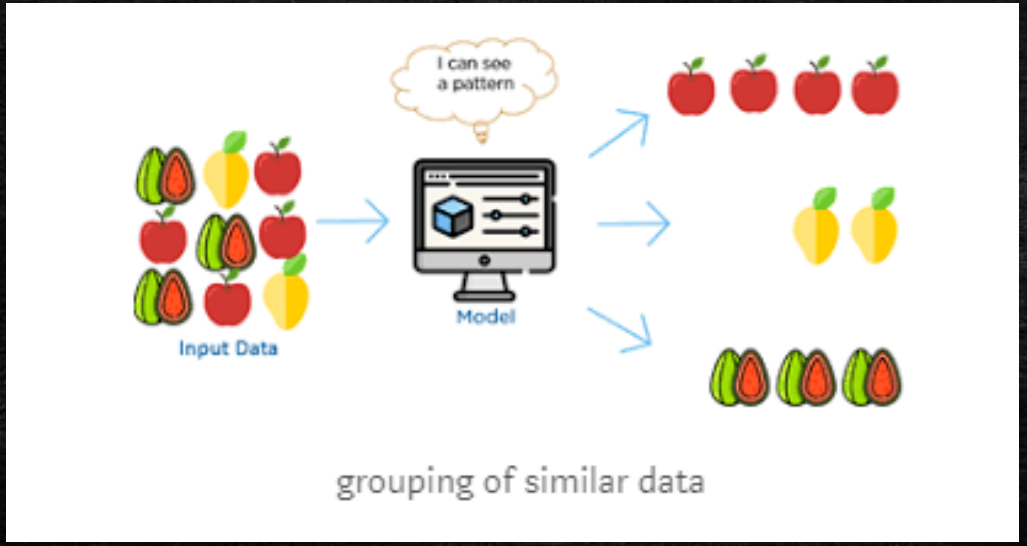
Machine Learning types

- **Supervised Learning**
  - ***Regression and classification*** problems are mainly solved here.
  - ***Labelled data*** is used for training here.
  - ***Popular Algorithms:*** Linear Regression, Support Vector Machines (SVM), Neural Networks, Decision Trees, Naive Bayes, Nearest Neighbor.



Machine Learning types

- **UnSupervised Learning**
  - ***Clustering problems(***grouping***), Anomaly Detection (in banks for unusual transactions)*** where there is a need for finding relationships among the data given.
  - ***Unlabeled data*** is used in unsupervised learning.
  - ***Popular Algorithms:*** *k-means clustering,* Association rule.

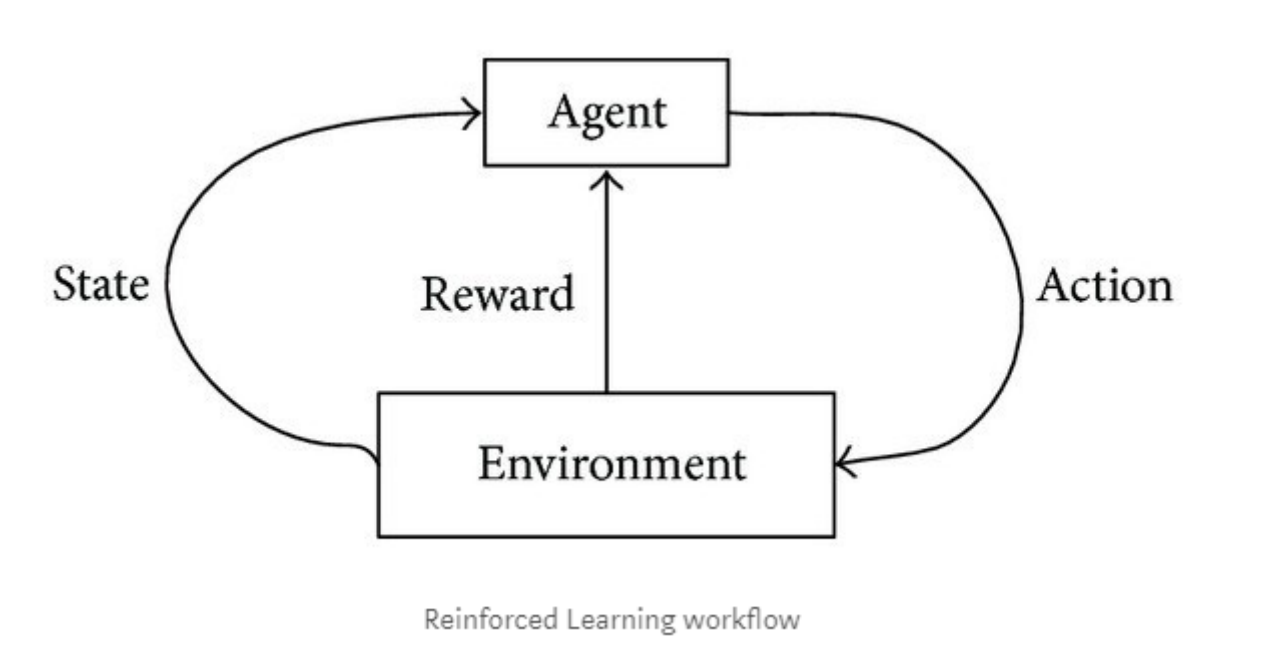


Machine Learning types

- **Semi-supervised Learning**
  - in-between that of ***Supervised and Unsupervised Learning***
  - ***labelled and unlabeled data***

Machine Learning types

- **Reinforced Learning**
  - Machine learns from past experience and tries to capture the best possible knowledge to make ***accurate decisions*** based on the feedback received



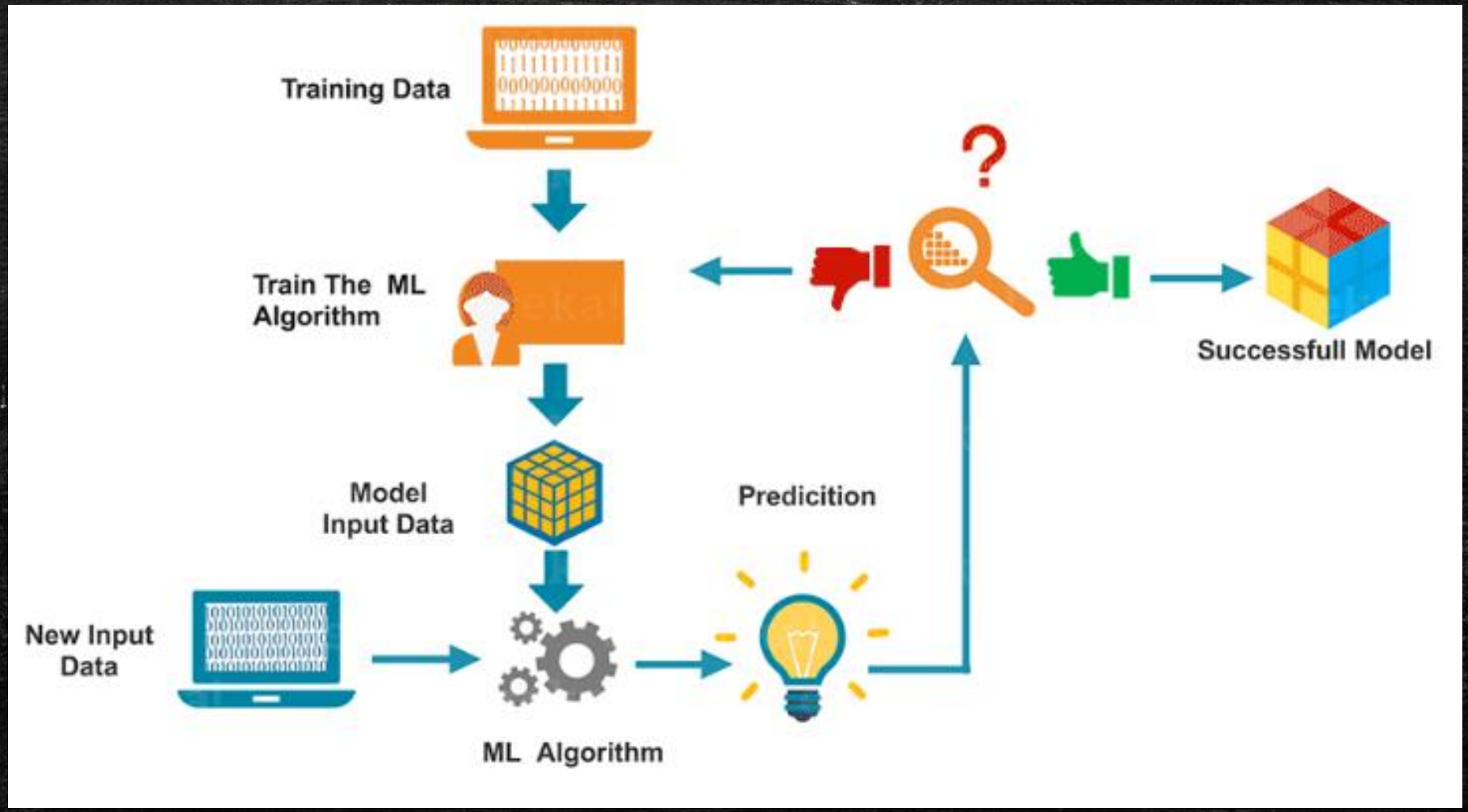
Reinforced Learning workflow

- **Source:** https://towardsdatascience.com/machine-learning-types-and-algorithms-d8b79545a6ec

Machine Learning common algorithms

- Linear Regression
- Logistic Regression
- Decision Tree
- SVM
- Naive Bayes
- kNN
- K-Means
- Random Forest
- Dimensionality Reduction Algorithms
- Gradient Boosting algorithms

Machine Learning Flow



Source: https://www.edureka.co/blog/what-is-machine-learning/

Applications of AI, ML, DL

- **Virtual Personal Assistants**

- Virtual Assistants are integrated to a variety of platforms. For example:
  - Smart Speakers: Amazon Echo and Google Home
  - Smartphones: Samsung Bixby on Samsung S8
  - Mobile Apps: Google Allo

Source: https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Applications of AI, ML, DL

- **Predictions while Commuting**

- *Traffic Predictions*
  - *Google Map*
- *Airline delay predictions*
  - *MS project*
- *Railway ticket confirmation prediction*
  - Various android apps
- *Online Transportation Networks*
  - Ubser

Source: https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Applications of AI, ML, DL

- **Videos Surveillance**

- *Single person monitoring?*
- *AI backed monitoring*
  - detect crime before they happen
  - track unusual behaviour of people like standing motionless for a long time, stumbling
- *Traffic signals*

Source: https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Applications of AI, ML, DL

- **Social Media Services**

- *People You May Know – LinkedIn, Facebook*
- *Face Recognition – Mobile unclock, Google photos, Facebook*
- *Similar Pins - Pinterest*

Source: https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Applications of AI, ML, DL

- **Email Spam and Malware Filtering**

- Spam vs HAM
- Multi Layer Perceptron, C 4.5 Decision Tree Induction
- Over 325, 000 malwares are detected everyday and each piece of code is 90–98% similar to its previous versions

Source: https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Applications of AI, ML, DL

- **Chatbots**

- Online Customer Support
  - State Bank of India's SIA chatbot
  - ICICI Bank's iPal
- Facebook Messenger
- Insomnia
- Disney: Solving Crimes with Fictional Characters, Zootopia

Source: https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Applications of AI, ML, DL

- **Search Engine Result Refining**

- Google
  - Need more examples??
  - Used to be Page Rank alone, now uses 200+ parameters and AI, ML
  - Stay on the web page for long
  - Search results but do not open any of the results

Source: https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Applications of AI, ML, DL

- **Product Recommendations**

- Netflix
  - Main differentiator
  - $1 Million coding contest
  - BellKor's Pragmatic Chaos team which bested Netflix's own algorithm for predicting ratings by 10.06%
- Amazon
  - Moment you start browsing
  - Buy this along with this
- Alibaba
  - E-commerce Brain - bookmarking, commenting, browsing history

Source: https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Applications of AI, ML, DL

- **Online Fraud Detection - Finance**

- Paypal is using ML for protection against money laundering
  - Outlier, anomaly detection
- You are watching "Game of Thrones" when you get a call from your bank asking if you have swiped your card for "$X" at a store in your city to buy a gadget

Source: https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

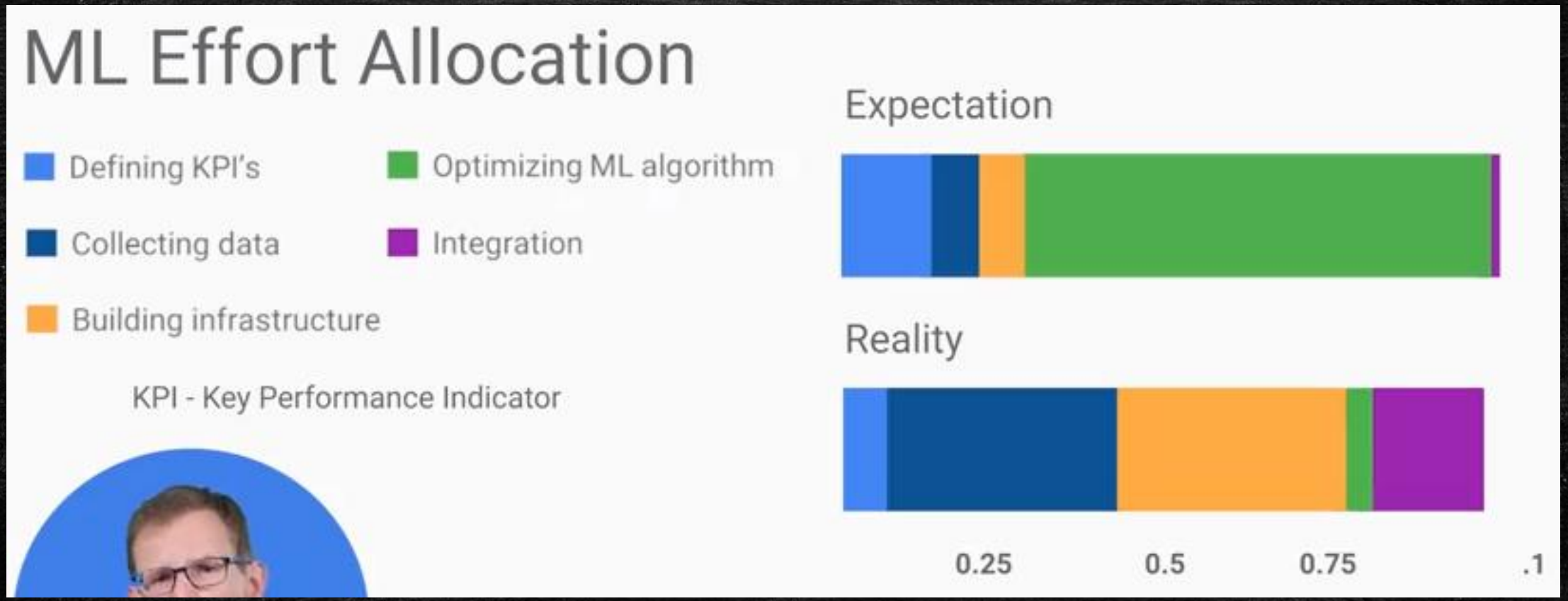
Applications of AI, ML, DL

- **Drug Discovery/Manufacturing**

- Pfizer is using IBM Watson on its immuno-oncology (a technique that uses body's immune system to help fight cancer) research
- **Personalized Treatment/Medication**
  - Genentech, a member of the Roche Group

Source: https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0

Learnings from Industry



Source: Tensorflow, Coursera

Learnings from Industry

- One approach not sufficient
- Solve using multiple ways and ensemble

Surviving ML onslaught

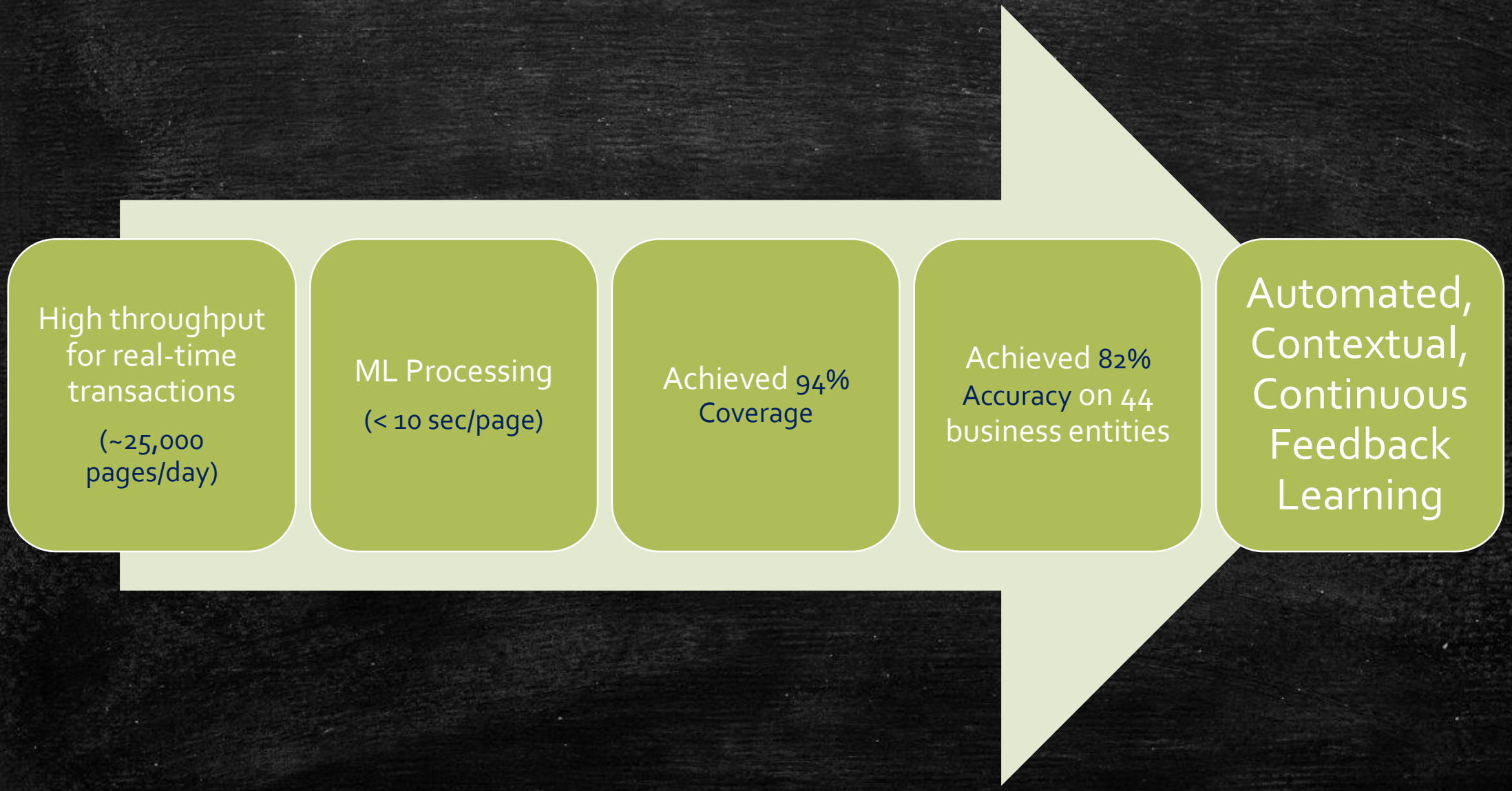
Everyday what you do, make sure it generates new challenges frequently

Learnings from Industry - Usecase

NER for Trade Finance for world leading bank

- **Spacy**
- **Tensorflow**
- **Checkmark recognition**
  - https://towardsdatascience.com/check-mark-state-recognition-will-take-nlp-projects-to-the-next-level-668a1013408f
- **GATE**
- **KEM**
- **Ensembler**
- **Abby, Nuance**
- **Document Classification**
- **Goods Classification**

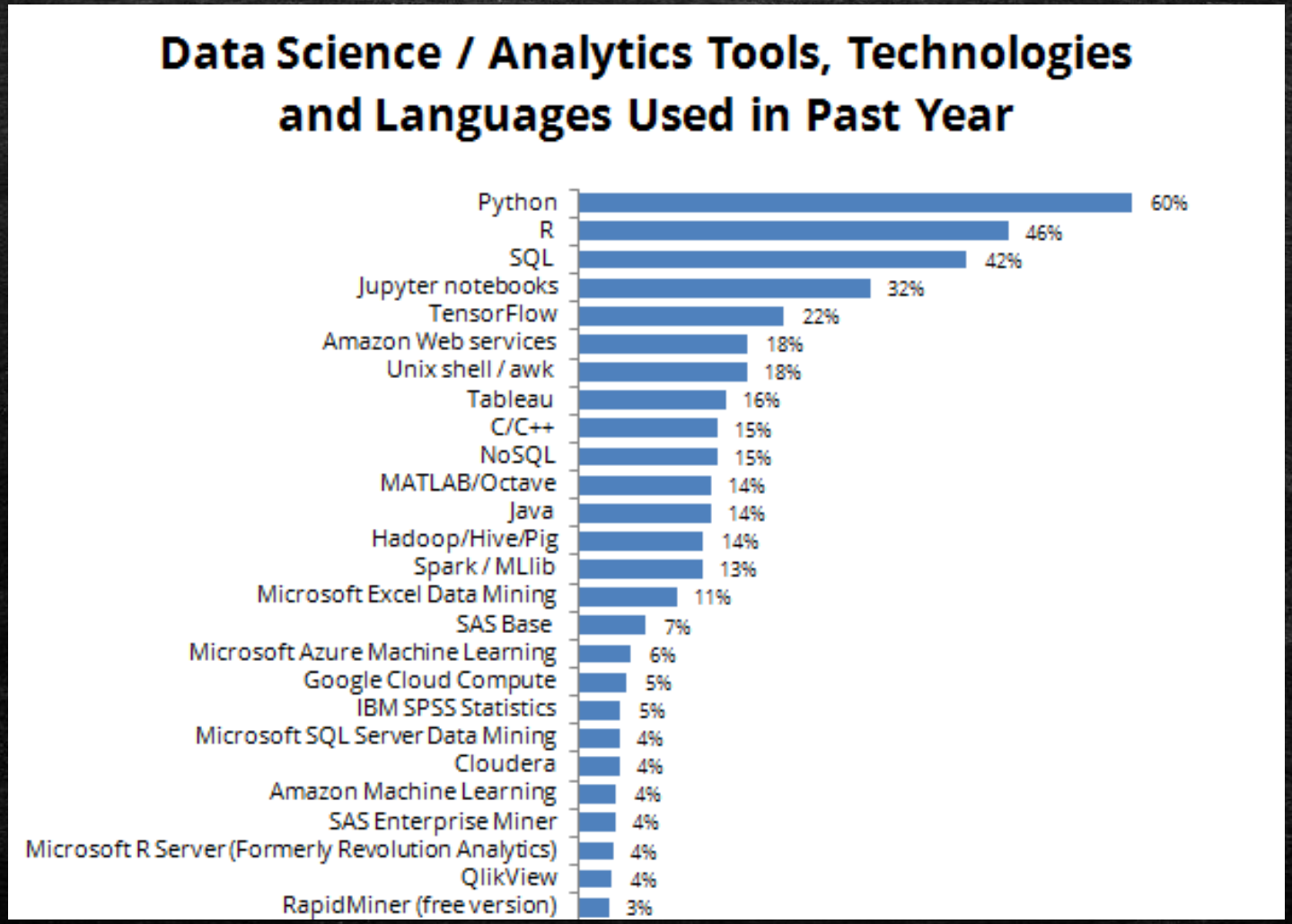
Metrics


High throughput for real-time transactions
(~25,000 pages/day)
ML Processing
(< 10 sec/page)
Achieved 94% Coverage
Achieved 82% Accuracy on 44 business entities
Automated, Contextual, Continuous Feedback Learning

Current buzz - Top 5 Data Science GitHub Repositories

- Flair (State-of-the-Art NLP Library)
  - Based on PyTorch
- face.evoLVe – High Performance Face Recognition Library
  - High performance deep face recognition
- YOLOv3
  - Object detection tasks
- FaceBoxes: A CPU Real-Time Face Detector with High Accuracy
  - face detecting approach, No GPU
- Transformer-XL from Google AI
  - Google AI team, NLP

Current buzz – Technologies used



Learn Data Science

- **Coursera**
  - Data Science Specialization – John Hopkins
    https://www.coursera.org/specializations/jhu-data-science
  - Deep Learning Specialization – Andrew Ng
    https://www.coursera.org/specializations/deep-learning
  - Machine Learning Primer – Stanford
    https://www.coursera.org/learn/machine-learning
- **Analytics Vidhya**
  - https://www.analyticsvidhya.com
- **Medium.com**
  - Data Science specific stream
    https://towardsdatascience.com/

Recap

Why to discuss Data Science

What is Data Science

The Fourth Paradigm

Four V's of Data

Data Science Umbrella, Technologies, Lifecycle

How machines learn, AI, ML, DL

Machine Learning – Intro

Applications of Data Science, ML

Learnings from Industry

Learn Data Science

Thank You !