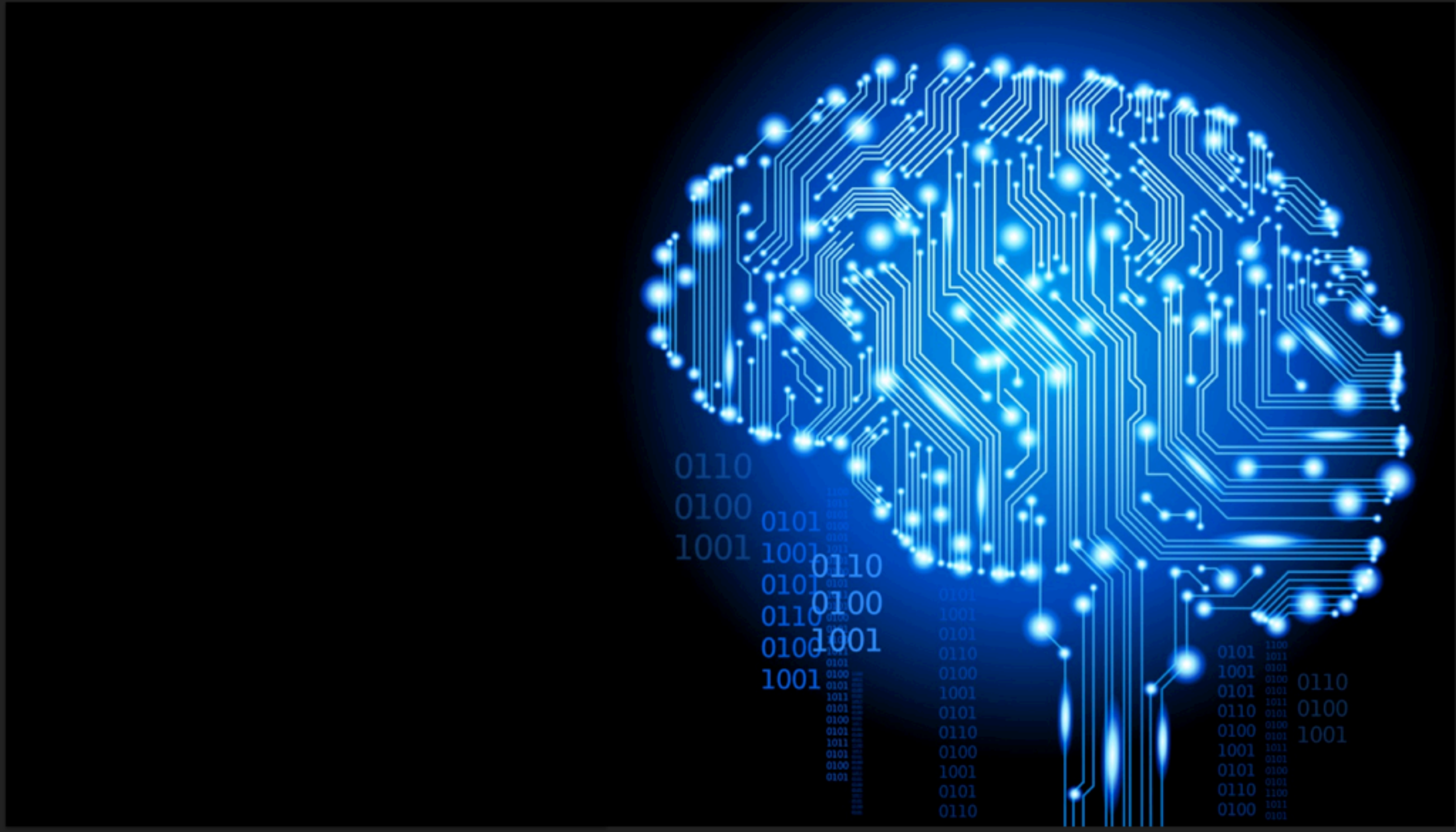


AN INTRODUCTION TO DATA SCIENCE

Chris Chen

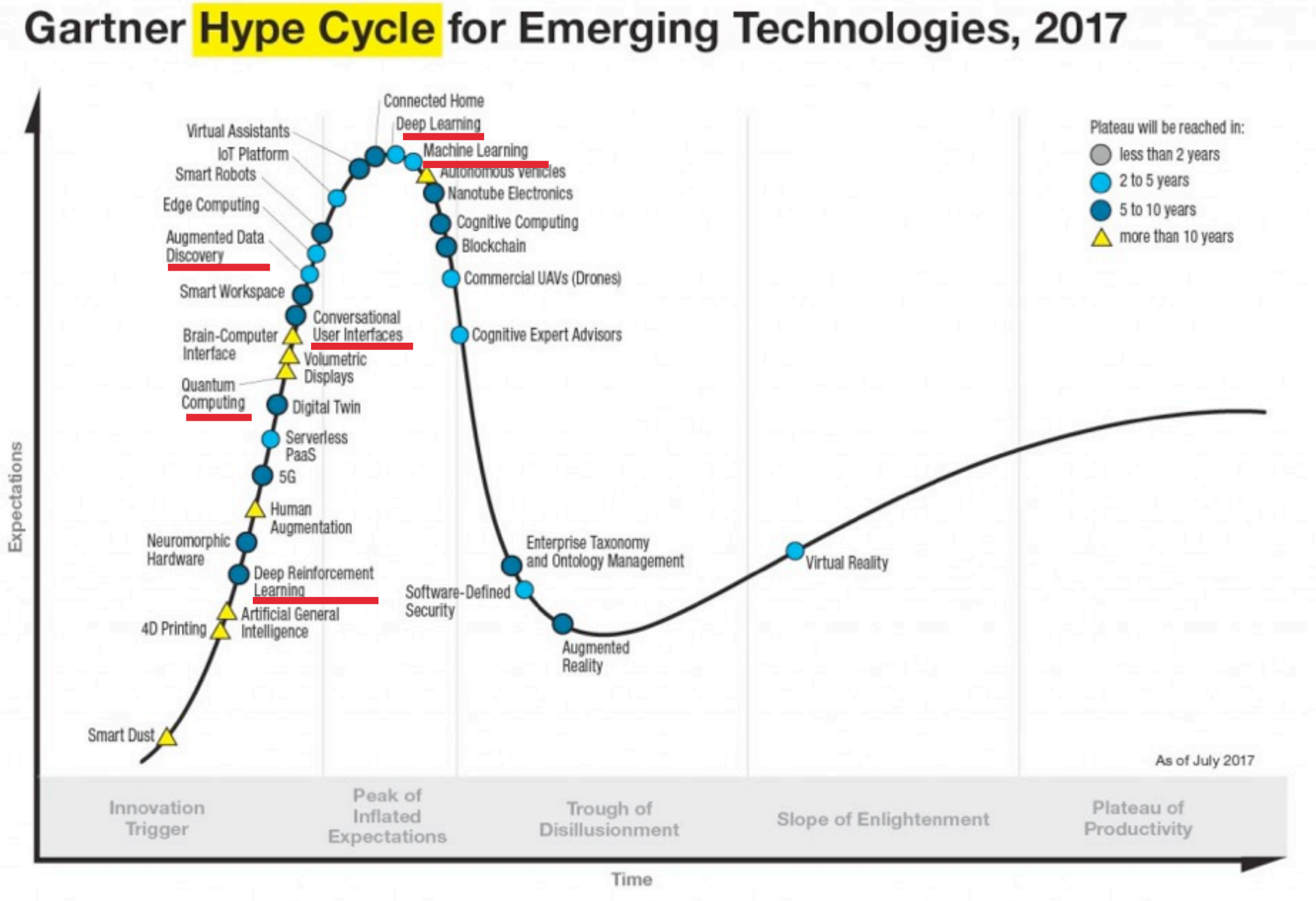
Data Scientist@AER

Kaggle Master

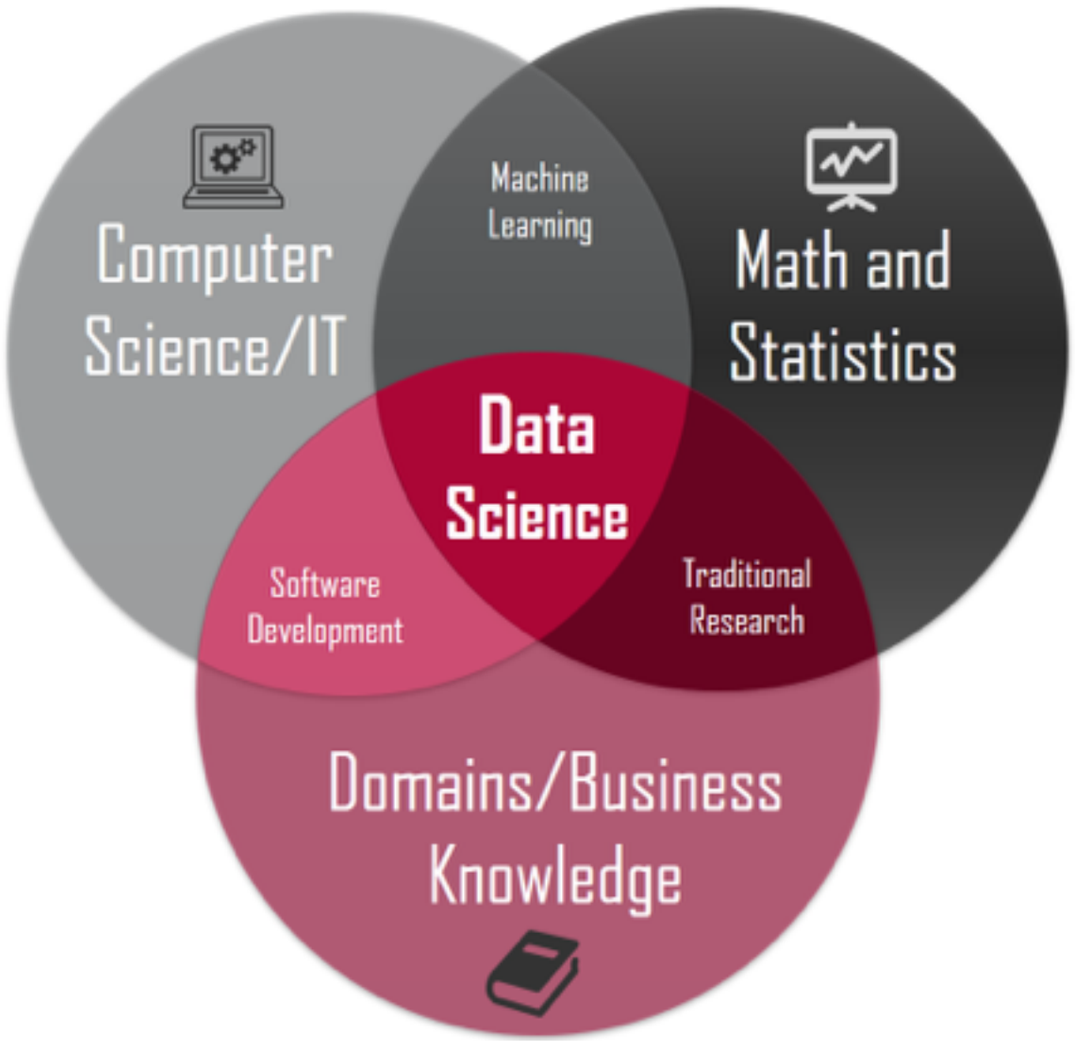
IN AN ERA OF


WEBDESIGN
SOCIAL MEDIA
SEARCHING
ADVERTISING
DATA
MARKETING
FORUM
BLOGGING
SECURITY
INTERFACE
NEWSFEED

AI IS EVERYWHERE



WHAT IS DATA SCIENCE



At its core, data science involves using automated methods to analyze massive amounts of data and to extract knowledge from them.

AN OVERSIMPLIFIED COMPARISON OF BUZZ WORDS

Computer Science

Programming

Java

C/C++
Python
....

Java Script

Data Structure
Computer System
Data Base
...

Data Science

Machine Learning

Neural Networks

an algorithm

Linear Regression
Decision Tree
SVM
...

Deep Learning

Statistical Inference
Pattern Recognition
Data Visualization
...

Data Visualization

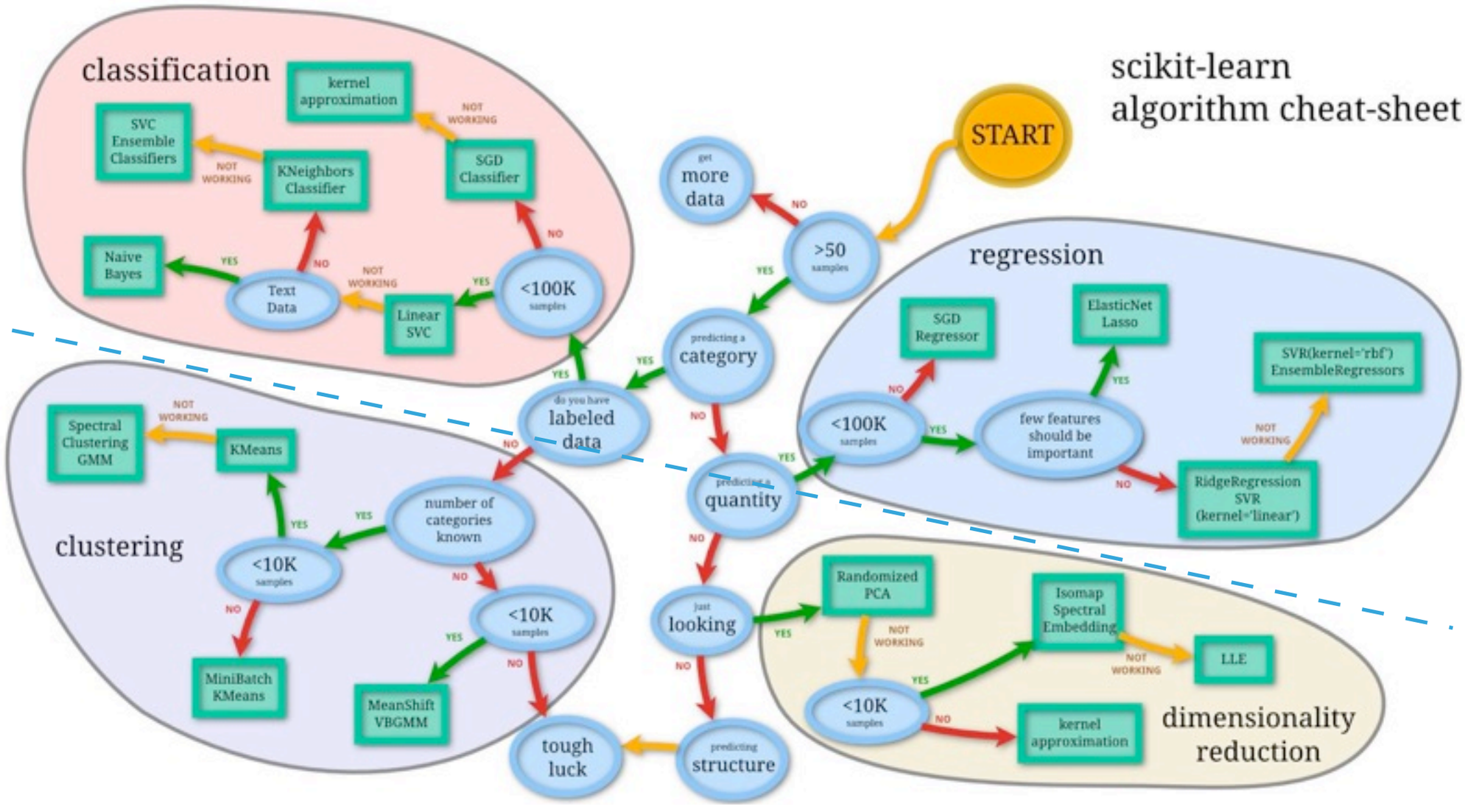
Artificial Intelligence

AlphaGo

MACHINE LEARNING ALGORITHMS

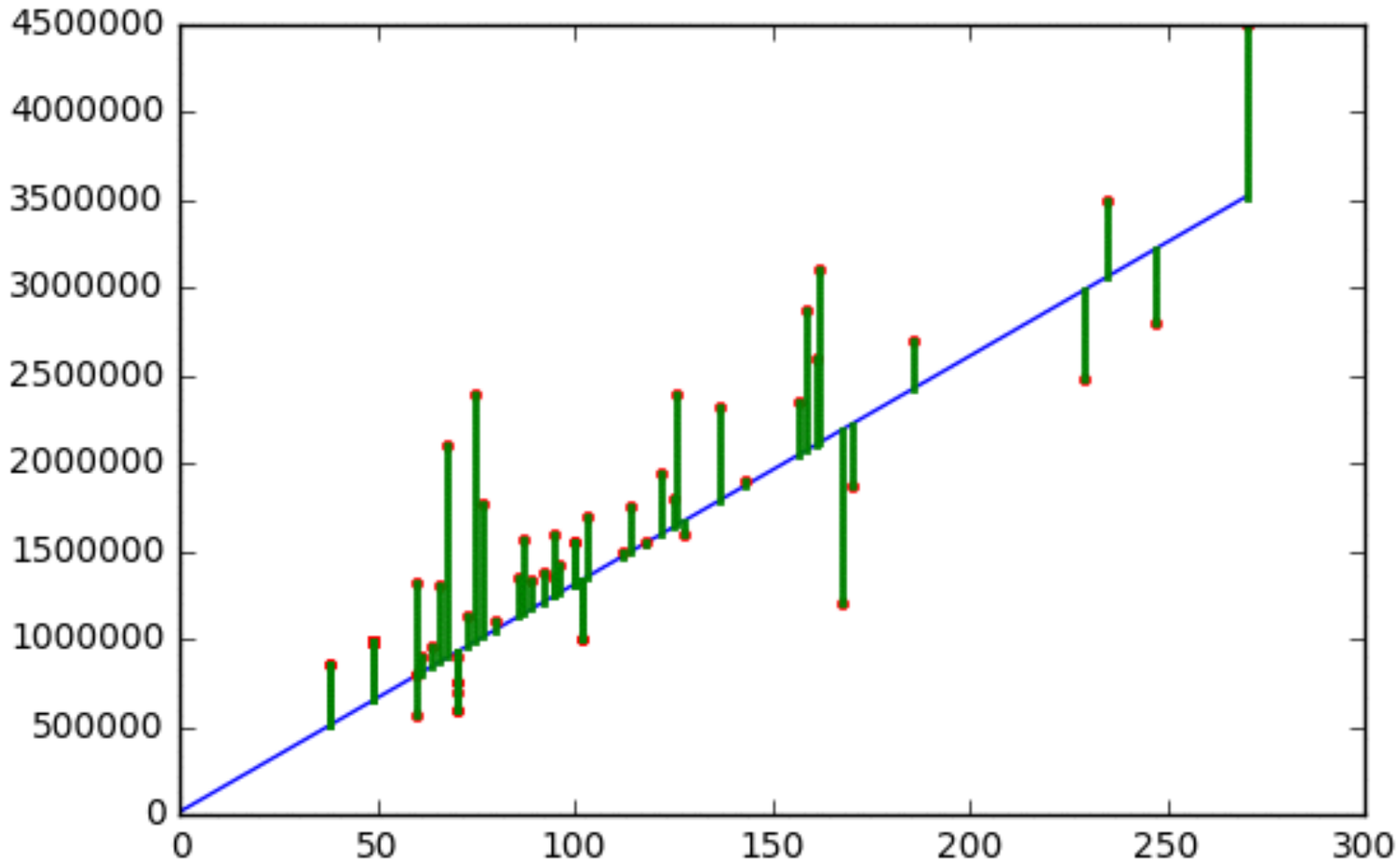
Supervised Learning

Unsupervised Learning



MACHINE LEARNING EXPLAINED – REGRESSION

	price	size
0	980000	49
1	980000	49
2	1550000	118
3	1350000	86
4	2400000	75
5	980000	49
6	1300000	66
7	1950000	122
8	980000	49



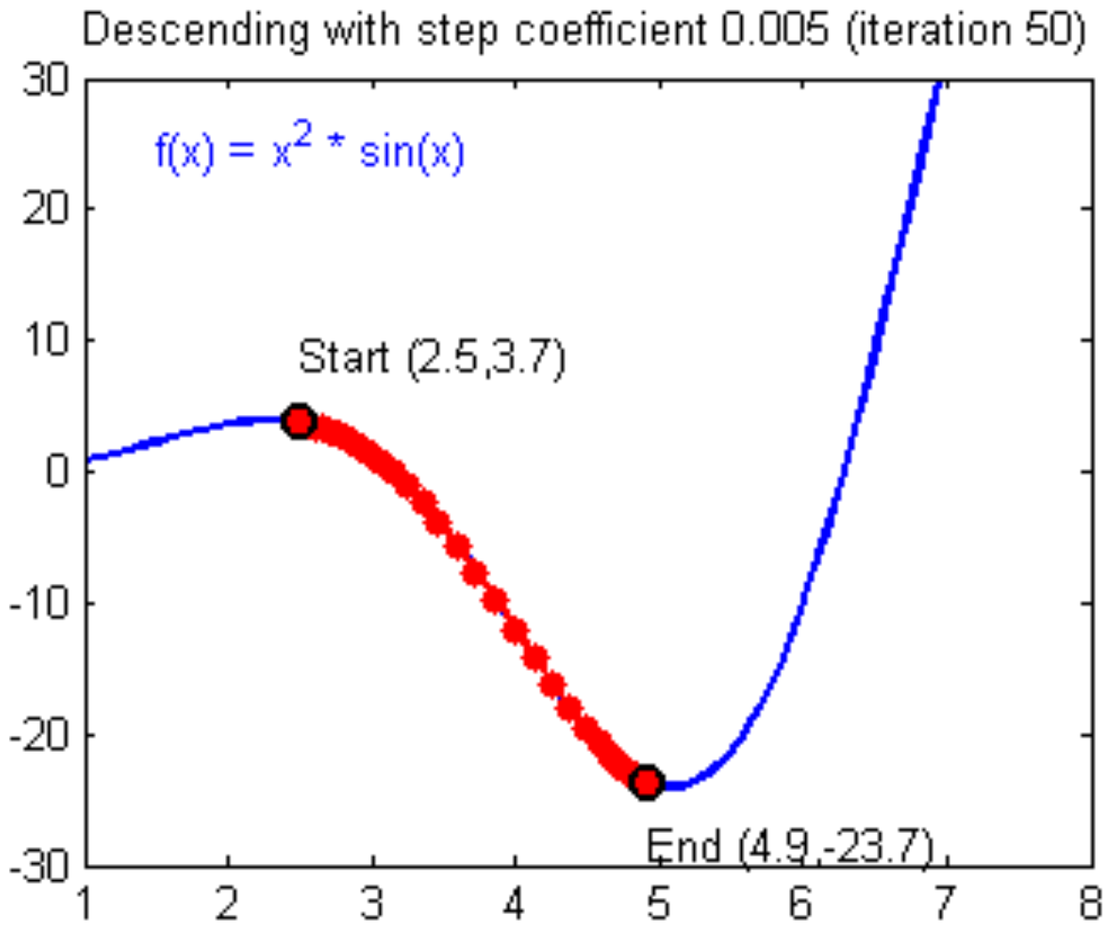
$$h(x) = \theta_0 + x\theta_1$$

The hypothesis

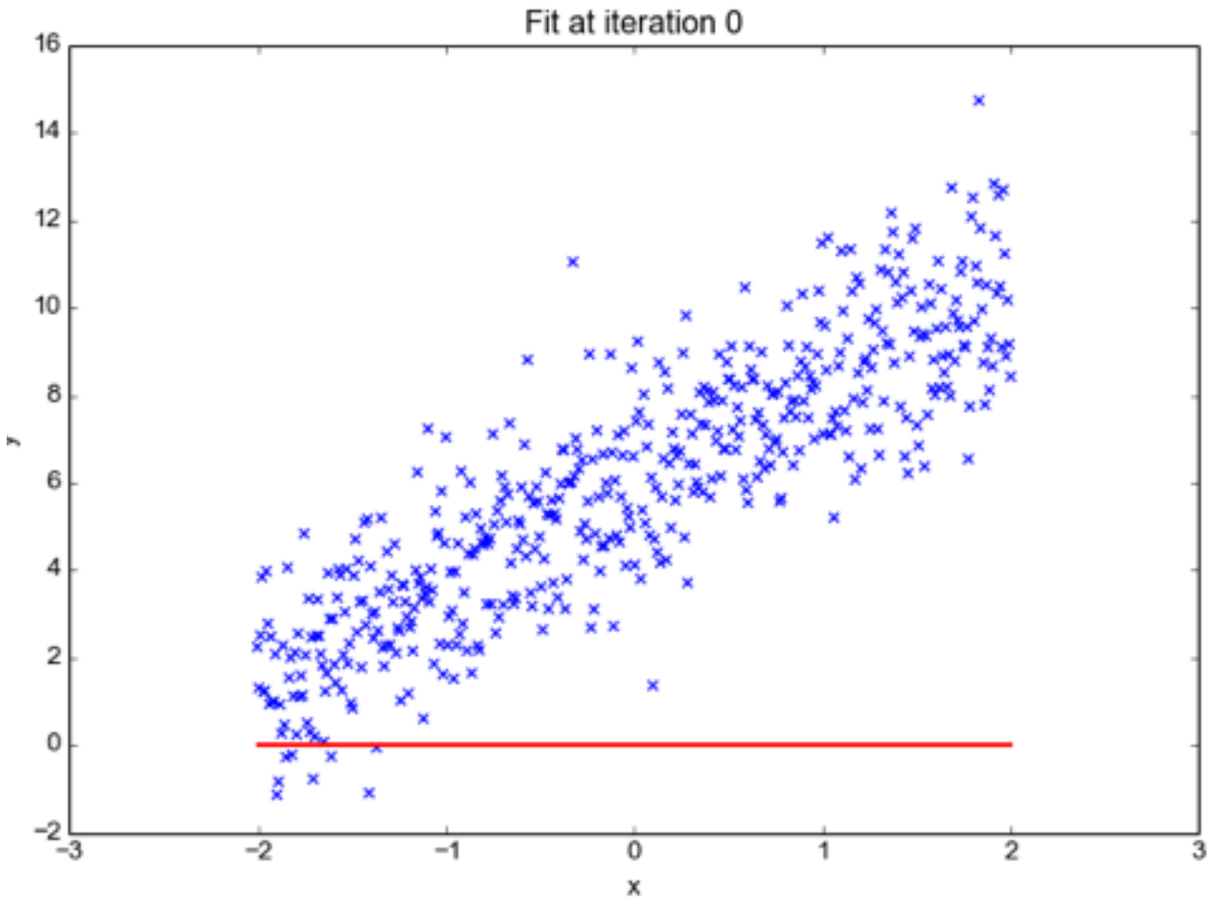
$$minJ(\theta_0, \theta_1) = \sum_{i=1}^{m} |h(X^i) - Y^i|$$

The object function

MACHINE LEARNING EXPLAINED – REGRESSION



Optimize cost - Gradient decent



Fit labeled data - Linear Regression

MACHINE LEARNING EXPLAINED - CLASSIFICATION

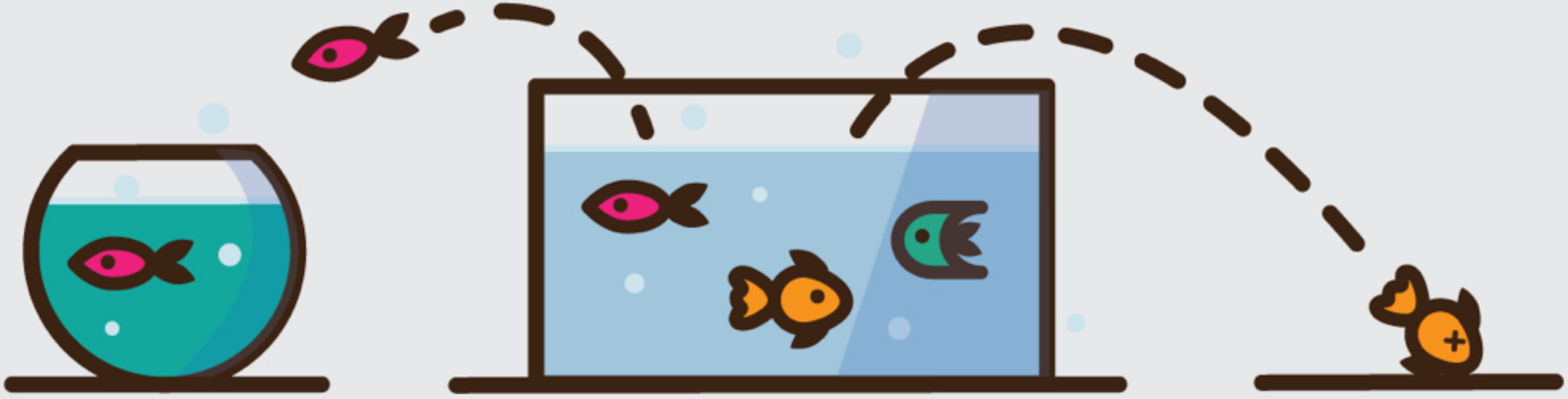
** R2D3 - A visual introduction to machine learning http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

USE CASE – CUSTOMER CHURN

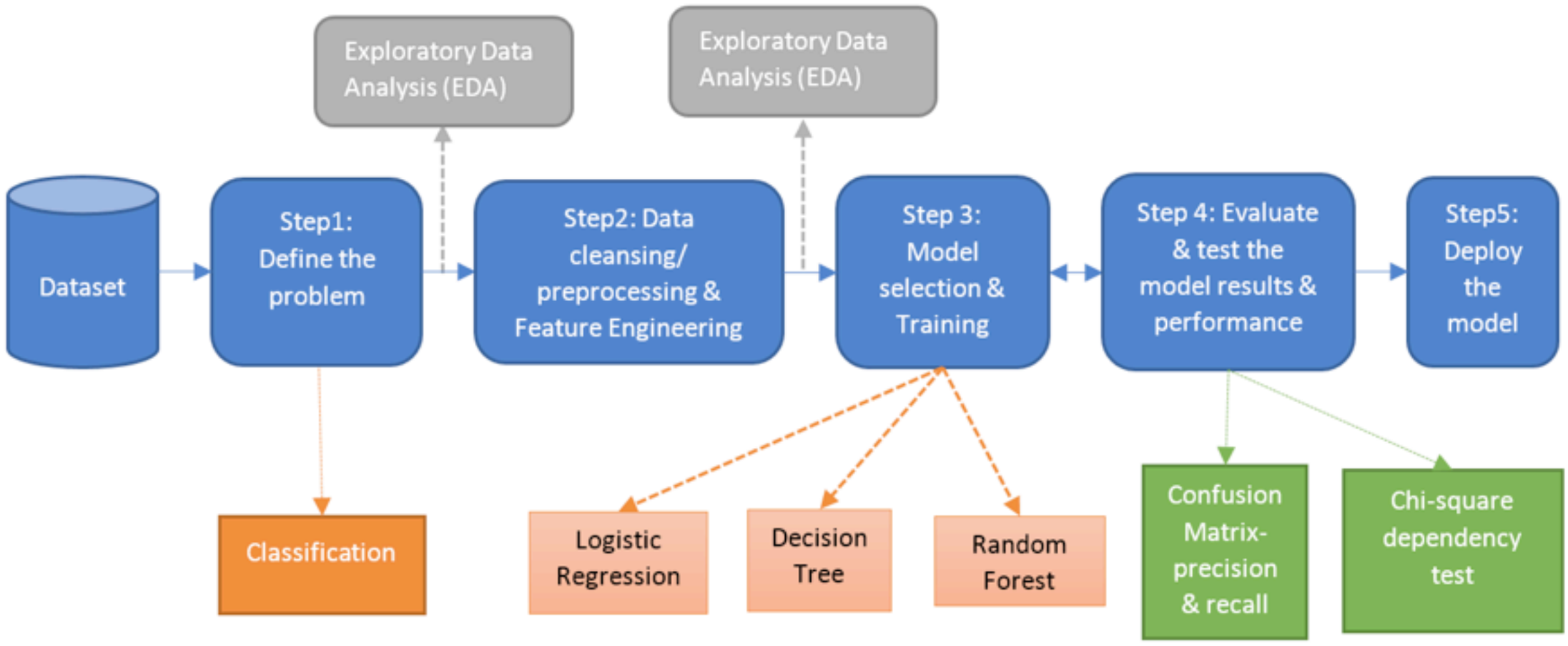
Which customers are under risky of churn?

How risky are they?

What can be done to retain them?

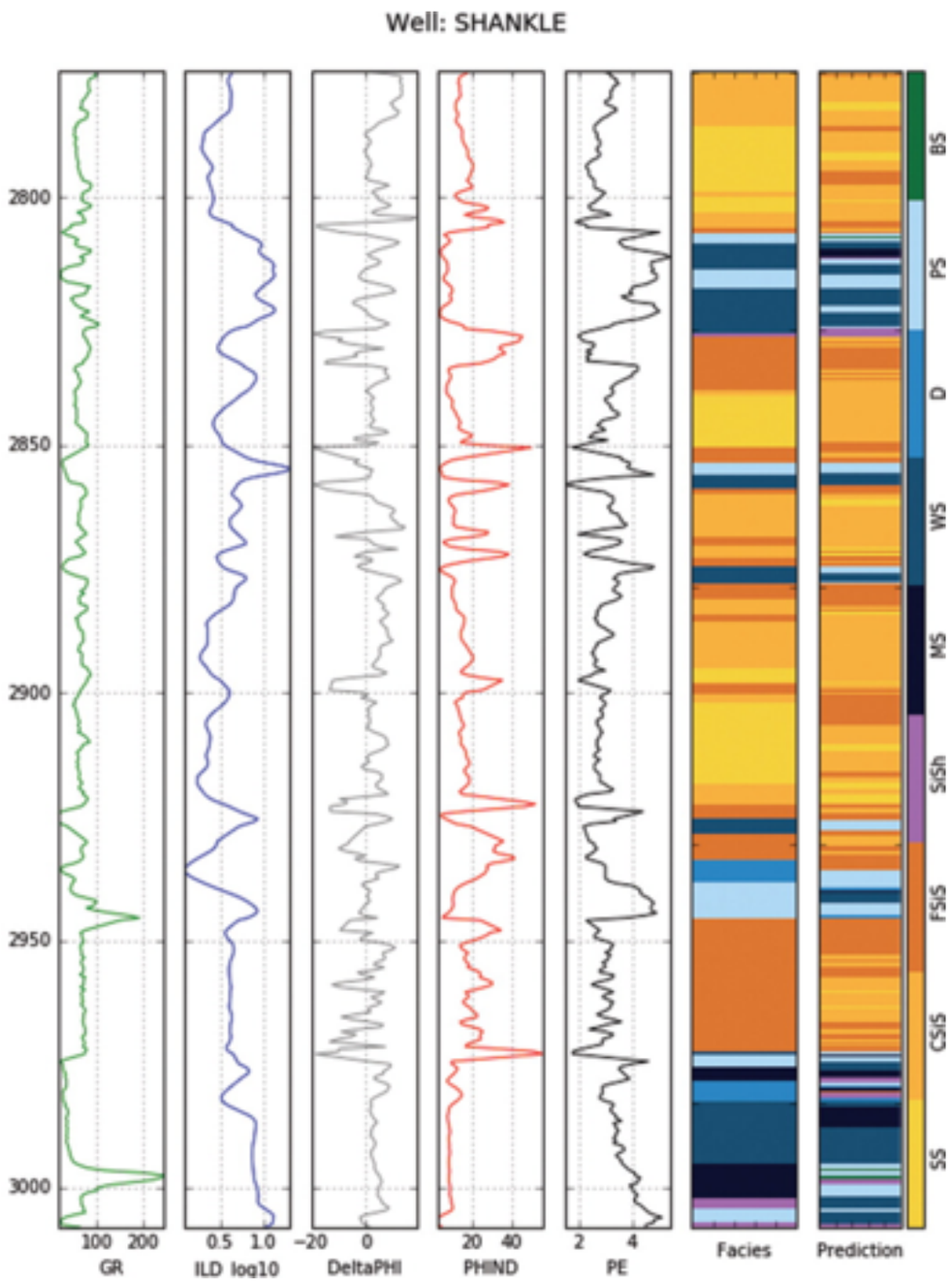


A TYPICAL PROCESS



USE CASE – OBJECT SEGMENTATION

USE CASE – FORMATION CLASSIFICATION



- Data
  - Core log
  - Well log
  - Seismic data
- Machine Learning
  - Neural Networks
  - Gradient Boosting Trees

OTHER USE CASES

Marketing

- Customer segmentation
- Target Marketing
- Lifetime Value (LTV) Prediction

Health Care

- Survival Analysis
- MRI Image Analysis
- Medical Resource Allocation

Financial

- Credit Risk
- Fraud Detection
- Anti-money Laundry

Energy

- Production Optimization
- Anomaly Detection
- Exploration and Discovery

ROADMAP TOWARDS A DATA SCIENTIST

TAKE ADVANTAGE OF MOOCS

Massive

Open

Online

Courses

Mathematics & Statistics

- Linear Algebra - MIT OpenCourseWare
- Intro to Statistics - Udacity
- Statistical Learning - Stanford Online

R / Python / SQL

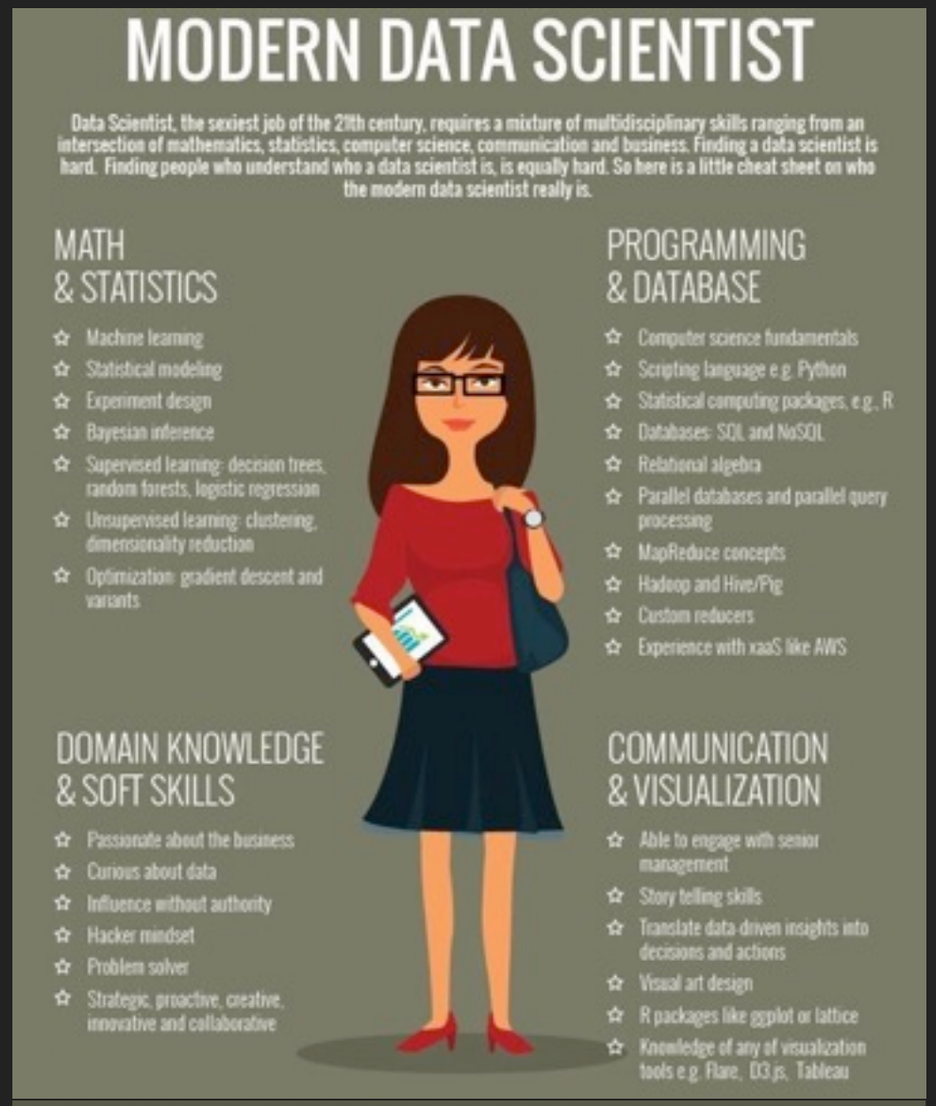
- Python for Everybody - U of Michigan / Coursera
- Complete Python Bootcamp: Go from zero to hero in Python - Udemy
- Statistics with R Specialization - Duke Univ. / Coursera
- Intro to Relational Databases - Udacity
- SQL for Data Science - UC Davis / Coursera

Machine Learning / Deep Learning

- Machine Learning Specialization - University of Washington/ Coursera
- Advanced Machine Learning Specialization - National Research University (Russia)
- Deep Learning - Udacity
- Deep Learning - deeplearning.ai / Coursera
- CS231n - Stanford Online
- Making neural nets uncool again - fast.ai
- 动手学深度学习 - DMLC (Strongly recommended)

CHOOSE THE RIGHT LEARNING PATH

- Pick a **TOPIC** you're passionate or curious about.
- Take a **LEARNING PATH** that works best for you.
  - Campus academic study: U of C Berkeley, Georgia Tech, UIUC
  - MOOCs: Cousera, Edx, Udacity
  - Data Science boot camps: Insight Data Science, Data Incubator, Data Application Lab, BitTiger.
- Get your hands dirty with a **PROBLEM** you are mostly interested in resolving.
  - Real-world challenges in your industry.
  - Data Science competitions: Kaggle, Tianchi, Data Science Bowl, KDD
  - Open datasets.



QUESTIONS?