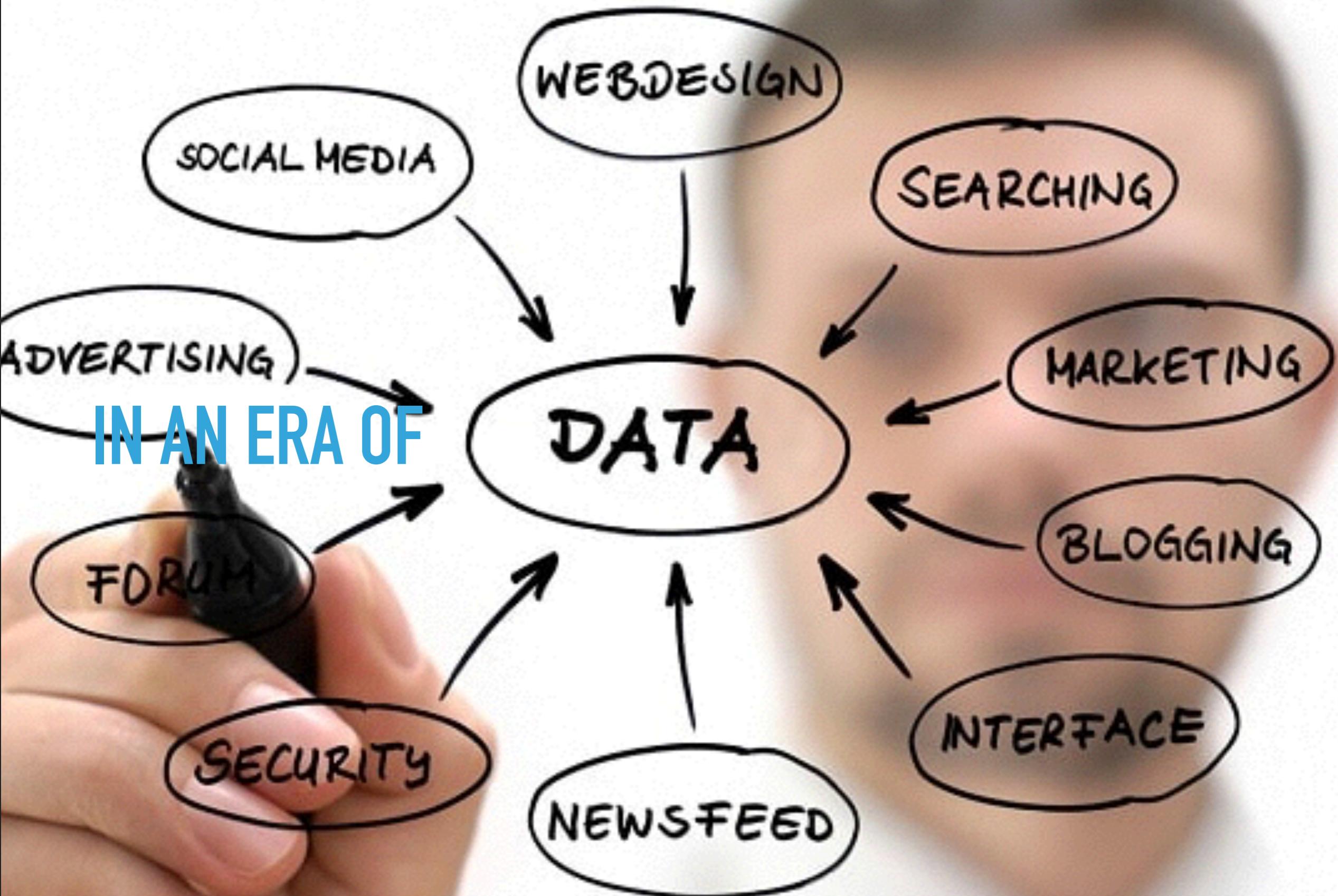


# AN INTRODUCTION TO DATA SCIENCE

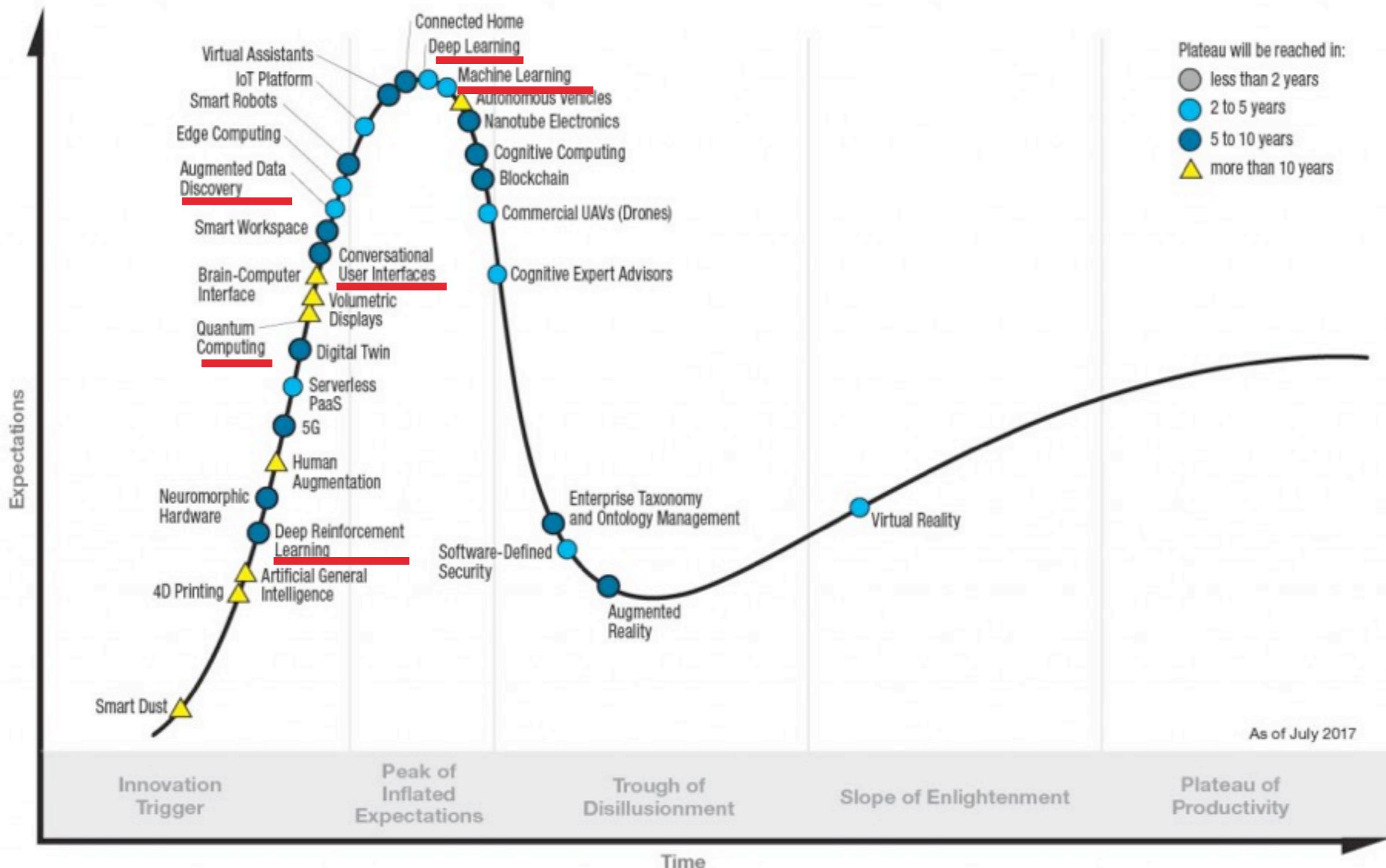


Chris Chen  
Data Scientist@AER  
Kaggle Master

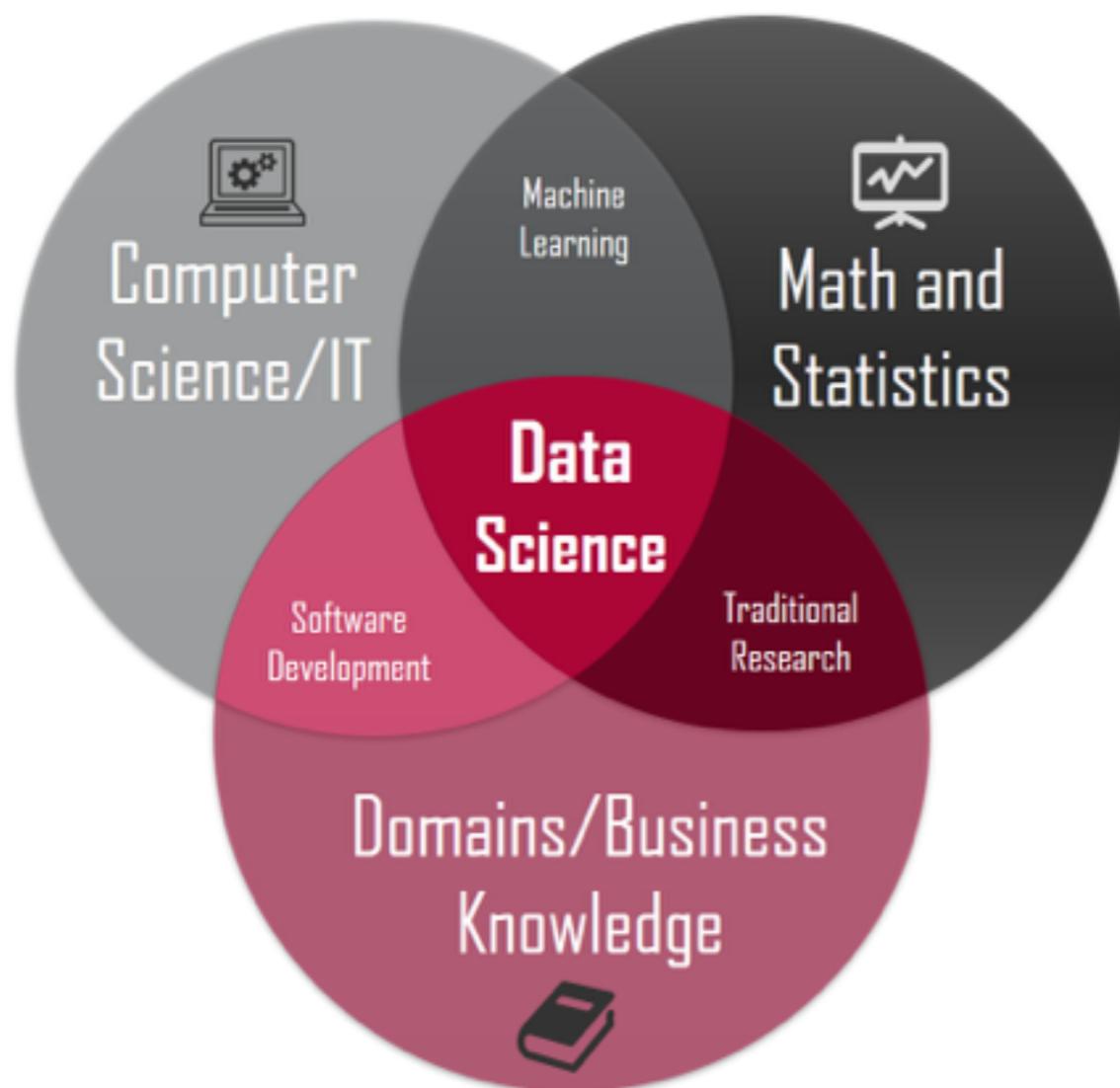


# AI IS EVERYWHERE

## Gartner Hype Cycle for Emerging Technologies, 2017



# 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

Java Script

Data Structure  
Computer System  
Data Base  
...

C/C++  
Python  
....

## Data Science

Machine Learning

Neural Networks  
an algorithm

Deep Learning

Statistical Inference  
Pattern Recognition  
Data Visualization  
...

Data Visualization

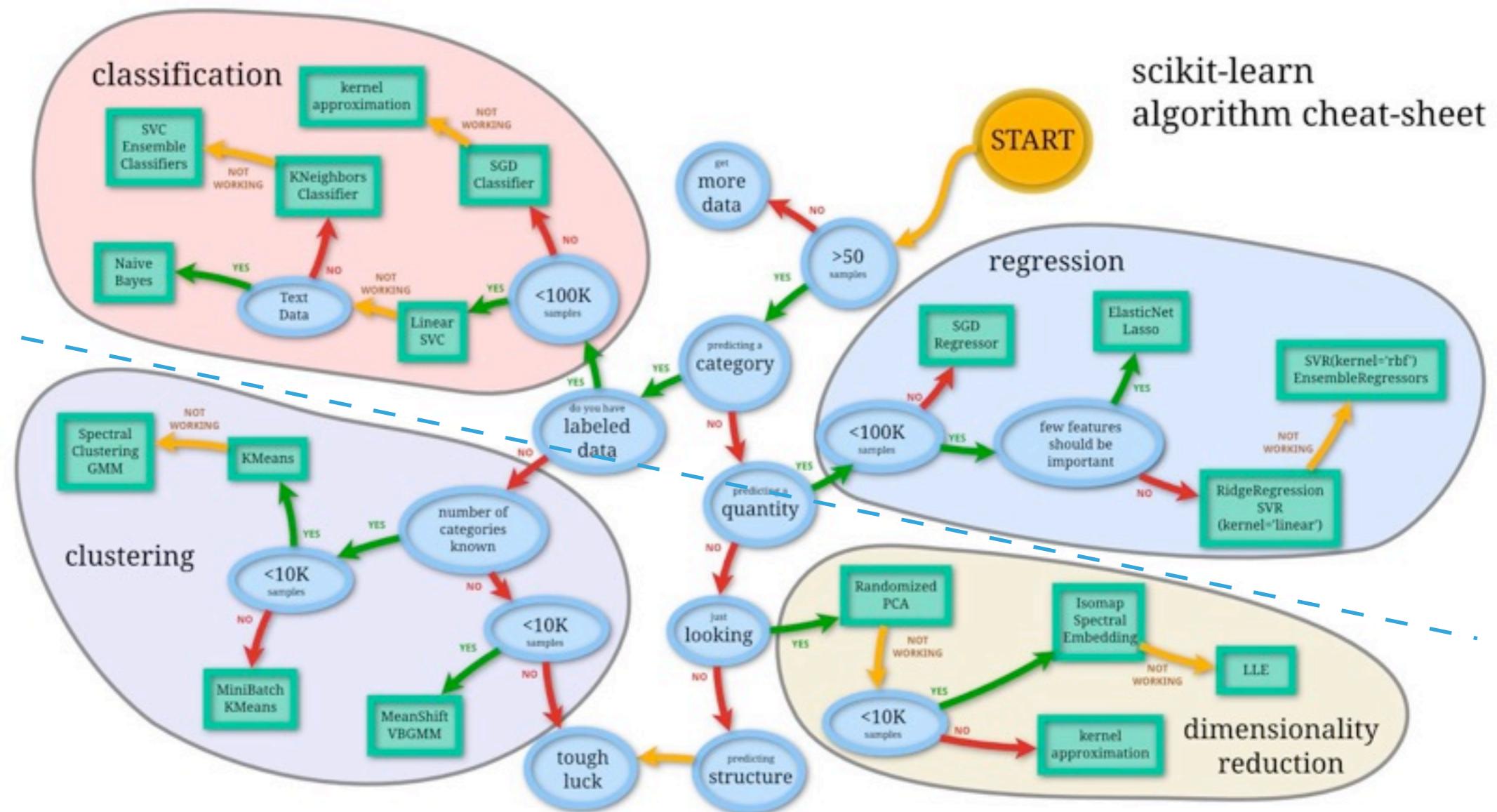
Artificial Intelligence

AlphaGo



# MACHINE LEARNING ALGORITHMS

Supervised Learning

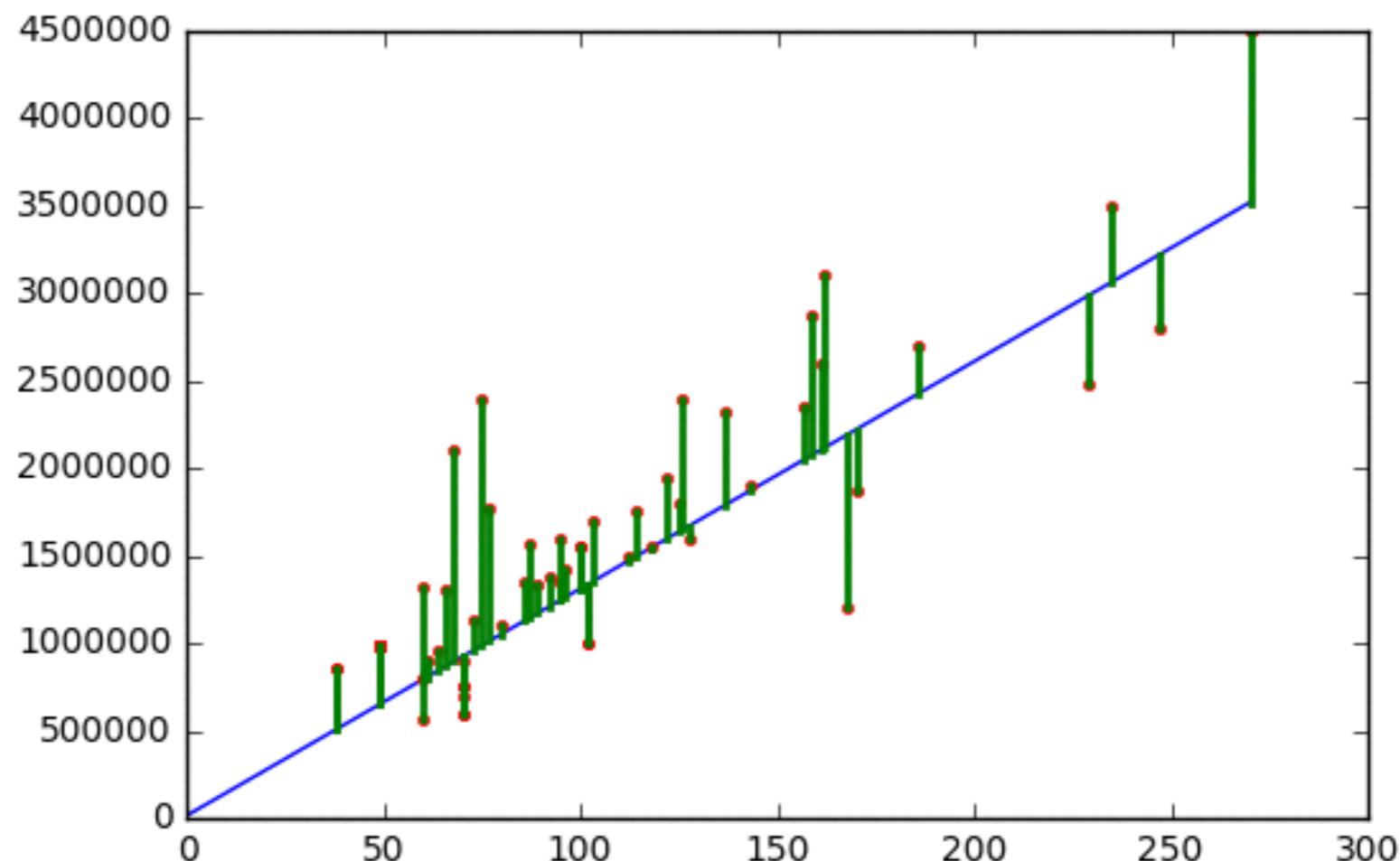


Unsupervised Learning

scikit-learn  
algorithm cheat-sheet

# 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



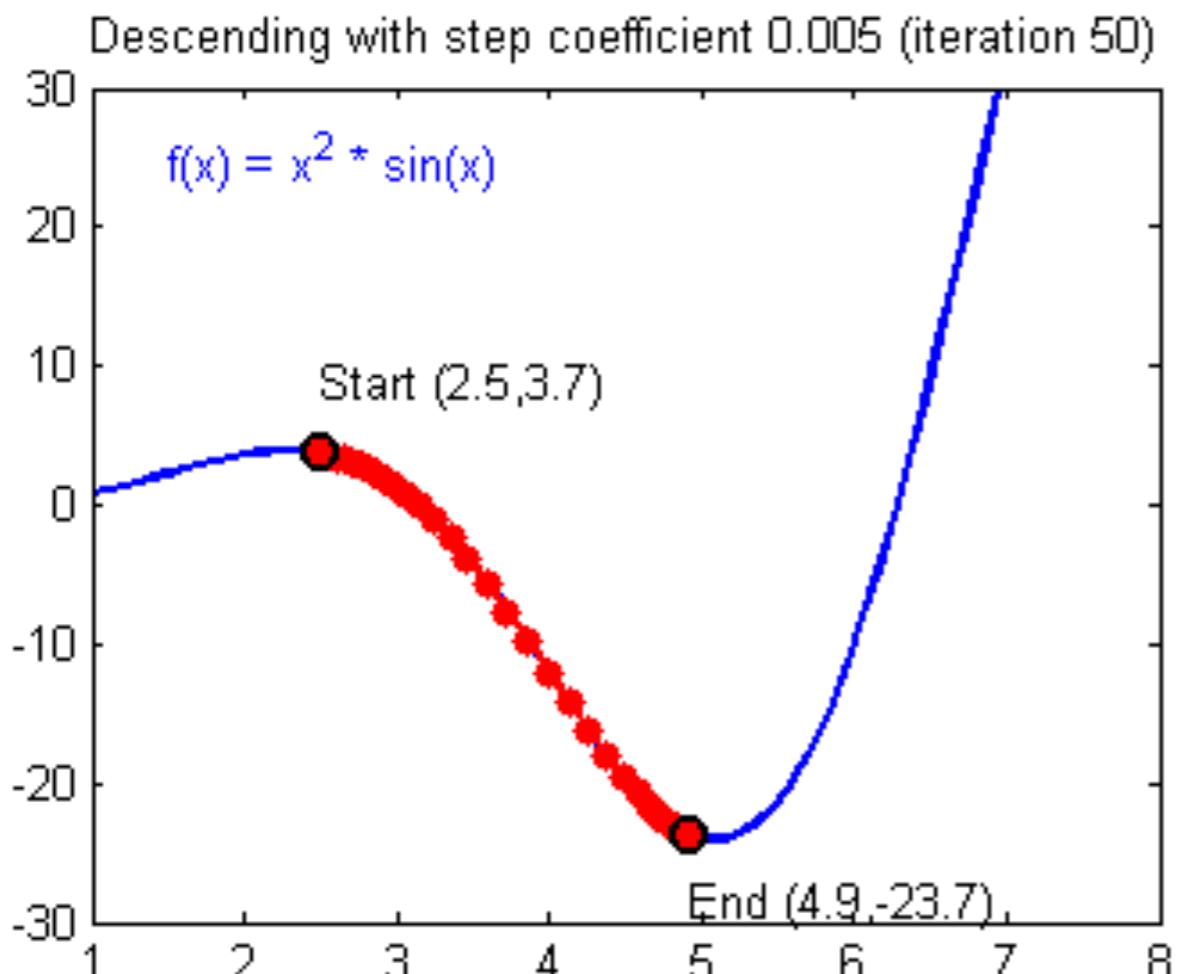
$$h(x) = \theta_0 + x\theta_1$$

The hypothesis

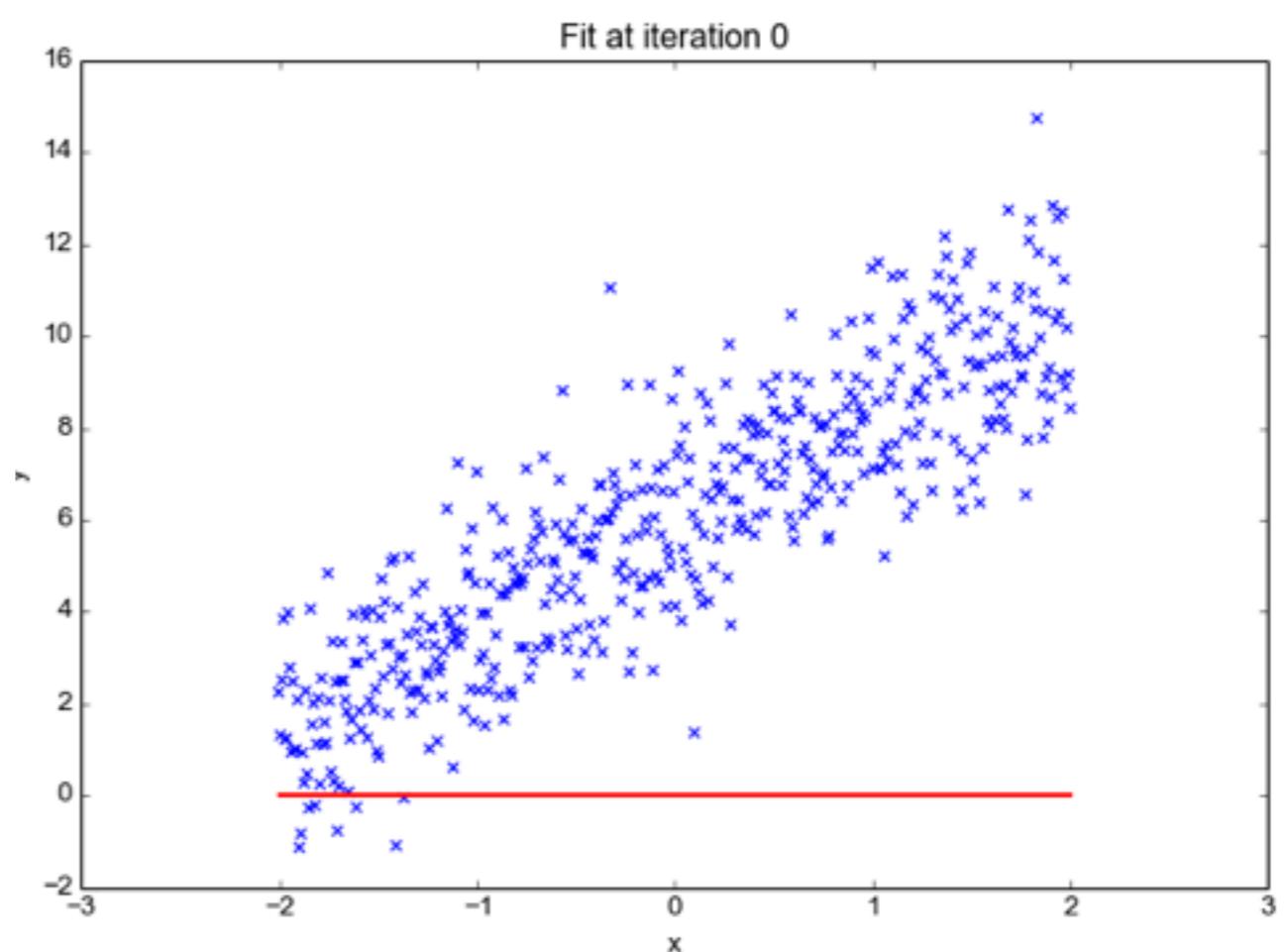
$$\min J(\theta_0, \theta_1) = \sum_{i=1}^m |h(X^i) - Y^i|$$

The object function

# MACHINE LEARNING EXPLAINED – REGRESSION



Optimize cost - Gradient decent



Fit labeled data - Linear Regression

# MACHINE LEARNING EXPLAINED – CLASSIFICATION

## Using the tree as a classifier

The newly-grown decision tree determines whether a home is in San Francisco or New York by running each data point through the branches.

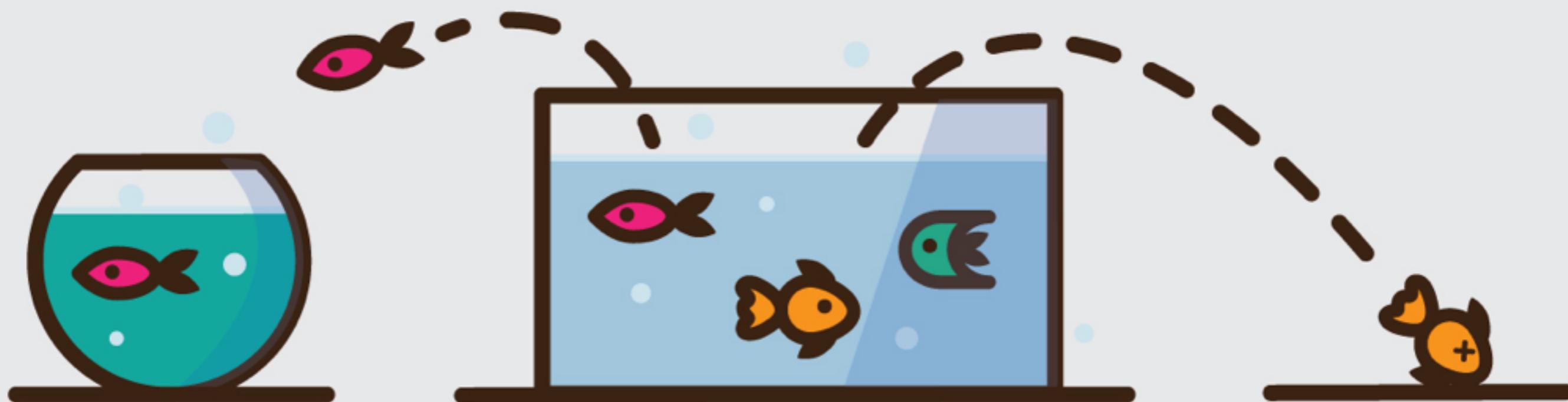


# USE CASE - CUSTOMER CHURN

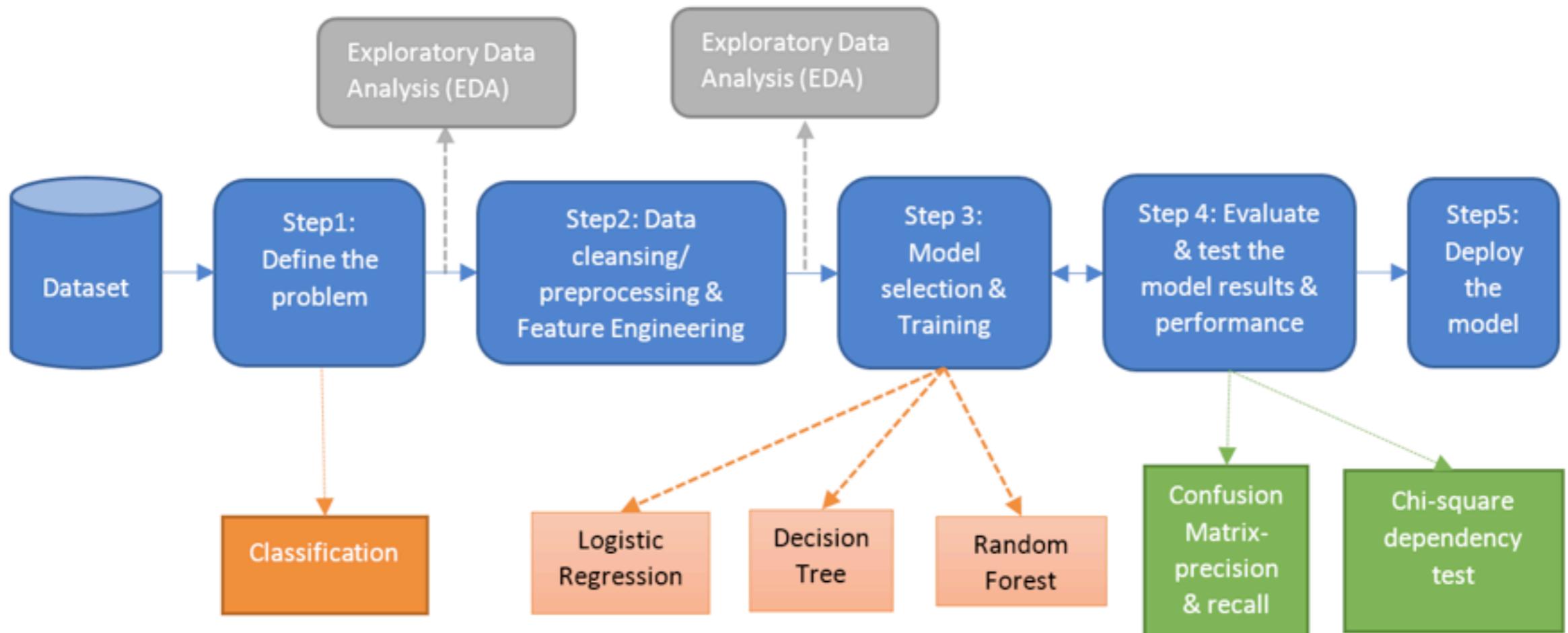
**Which** customers are under risk of churn?

**How** risky are they?

**What** can be done to retain them?



# A TYPICAL PROCESS

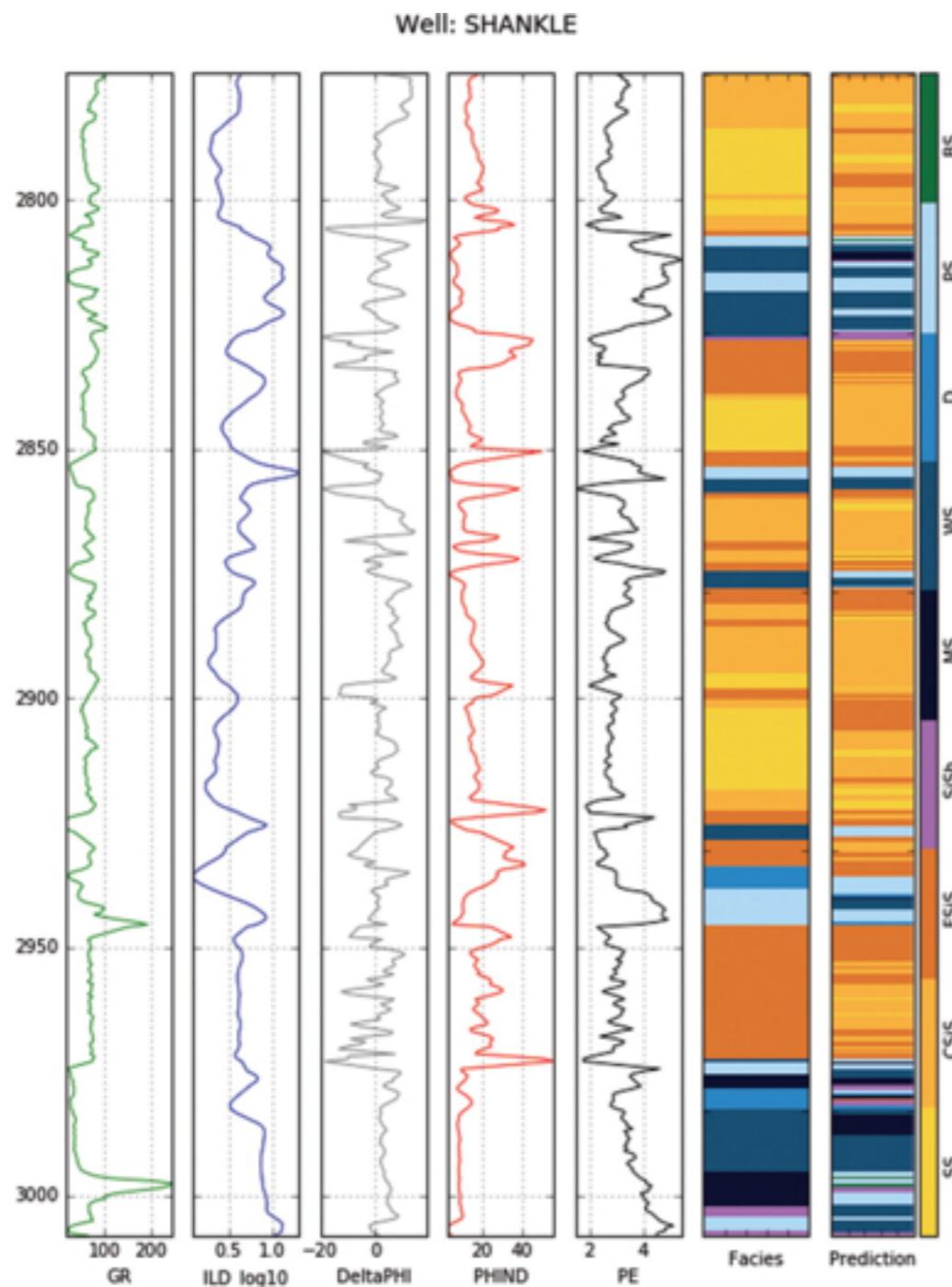


# USE CASE - OBJECT SEGMENTATION



- Buildings
- Misc.
- Roads
- Tracks
- Trees
- Crops
- Standing Water
- Large Vehicles
- Small Vehicles

# 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: Coursera, 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, Tiantchi, Data Science Bowl, KDD
  - Open datasets.

## 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	PROGRAMMING & DATABASE	DOMAIN KNOWLEDGE & SOFT SKILLS	COMMUNICATION & VISUALIZATION
<ul style="list-style-type: none"><li>★ Machine learning</li><li>★ Statistical modeling</li><li>★ Experiment design</li><li>★ Bayesian inference</li><li>★ Supervised learning: decision trees, random forests, logistic regression</li><li>★ Unsupervised learning: clustering, dimensionality reduction</li><li>★ Optimization: gradient descent and variants</li></ul>	<ul style="list-style-type: none"><li>★ Computer science fundamentals</li><li>★ Scripting language e.g. Python</li><li>★ Statistical computing packages, e.g. R</li><li>★ Databases: SQL and NoSQL</li><li>★ Relational algebra</li><li>★ Parallel databases and parallel query processing</li><li>★ MapReduce concepts</li><li>★ Hadoop and Hive/Pig</li><li>★ Custom reducers</li><li>★ Experience withaaS like AWS</li></ul>	<ul style="list-style-type: none"><li>★ Passionate about the business</li><li>★ Curious about data</li><li>★ Influence without authority</li><li>★ Hacker mindset</li><li>★ Problem solver</li><li>★ Strategic, proactive, creative, innovative and collaborative</li></ul>	<ul style="list-style-type: none"><li>★ Able to engage with senior management</li><li>★ Story telling skills</li><li>★ Translate data-driven insights into decisions and actions</li><li>★ Visual art design</li><li>★ R packages like ggplot or lattice</li><li>★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau</li></ul>

Skills required by a modern data scientist:

- ★ Solid understanding of statistics and probability
- ★ Proficiency in programming languages such as Python, R, and SQL
- ★ Knowledge of machine learning algorithms and their applications
- ★ Ability to work with large datasets and perform data analysis
- ★ Strong communication and presentation skills
- ★ Familiarity with data visualization tools like Tableau and ggplot2
- ★ Experience with big data technologies like Hadoop and Spark
- ★ Understanding of domain-specific knowledge in fields like finance, healthcare, or marketing
- ★ Ability to work effectively in a team and lead projects
- ★ Problem-solving skills and ability to think critically
- ★ Attention to detail and ability to handle complex data structures
- ★ Knowledge of data mining and data cleaning techniques
- ★ Familiarity with data storage and retrieval systems like MySQL and PostgreSQL
- ★ Experience with data visualization tools like Tableau and ggplot2
- ★ Understanding of data privacy and ethical considerations in data science
- ★ Ability to work effectively in a team and lead projects
- ★ Problem-solving skills and ability to think critically
- ★ Attention to detail and ability to handle complex data structures
- ★ Knowledge of data mining and data cleaning techniques
- ★ Familiarity with data storage and retrieval systems like MySQL and PostgreSQL

---

# QUESTIONS?

