Intro to Data Science

Erin LeDell Ph.D.
Statistician & Machine Learning Scientist
H2O.ai

H2O World 2015

Download our app, “H2O World 2015”

H2O World 2015

I have H2O Installed

I have R installed

I have Python installed

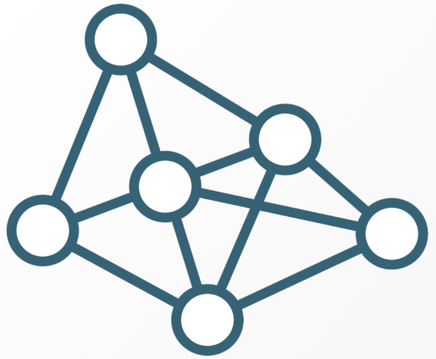
I have the H2O World data sets

Pick up stickers or get install help at the information booth

H2O WORLD

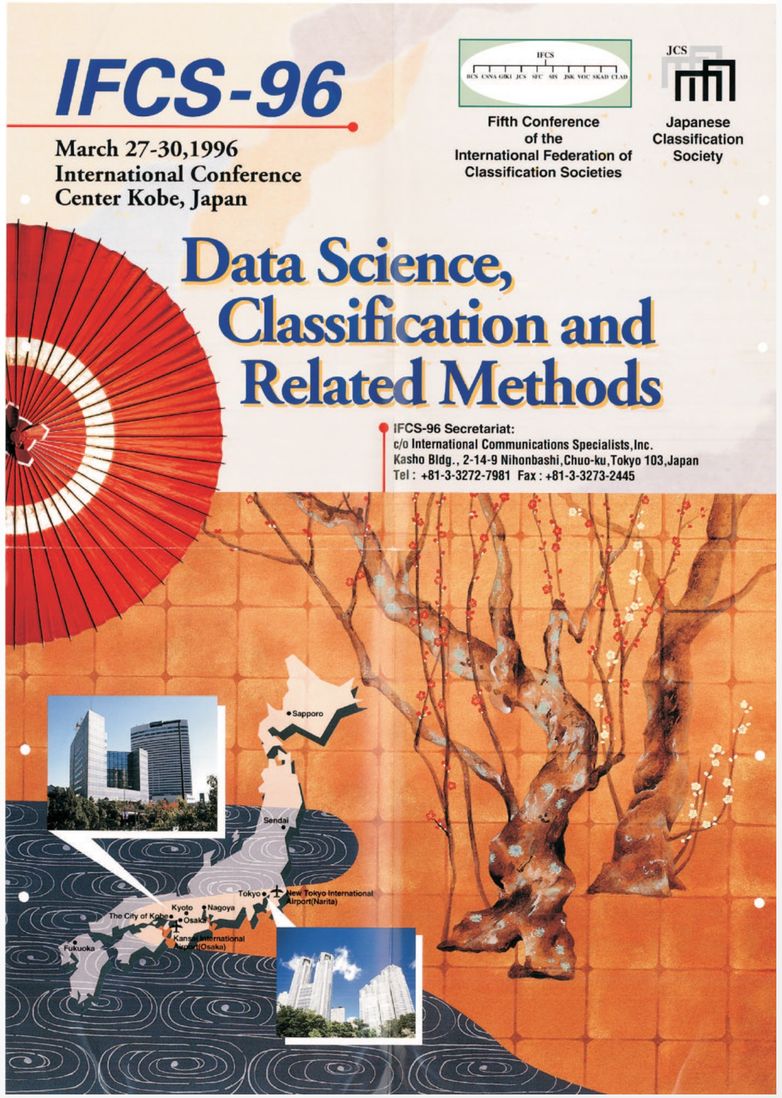
Intro to Data Science

- What is Data Science?
- The Data Scientist
- The Data Science Team
- Data Science Tools
- What is Machine Learning?
- What is Deep Learning?
- What is Ensemble Learning?
- Data Science Resources



What is Data Science?

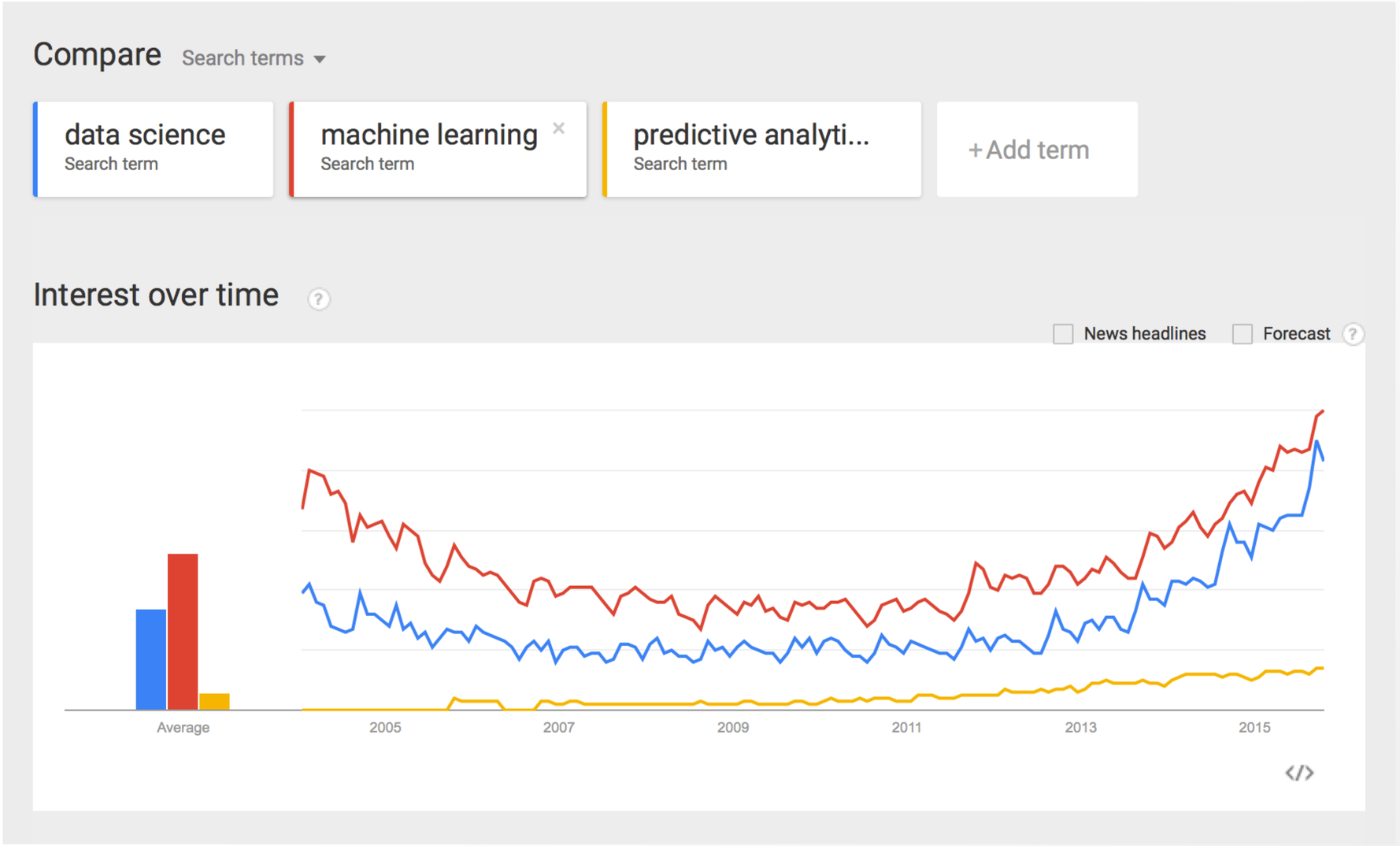
One of the earliest uses of the term "data science" occurred in the title of the 1996 International Federation of Classification Societies conference in Kobe, Japan.



What is Data Science?

- The term re-emerged and became popularized in 2001 by William Cleveland, then at Bell Labs, when he published, *"Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics"*.
- This publication describes a plan to enlarge the major areas of technical work of the field of statistics. Dr. Cleveland states, "Since plan is ambitious and implies substantial change, the altered field will be called Data Science."

What is Data Science?



What is Data Science?

Problem Formulation

- Identify a data task or prediction problem
- Collect relevant data

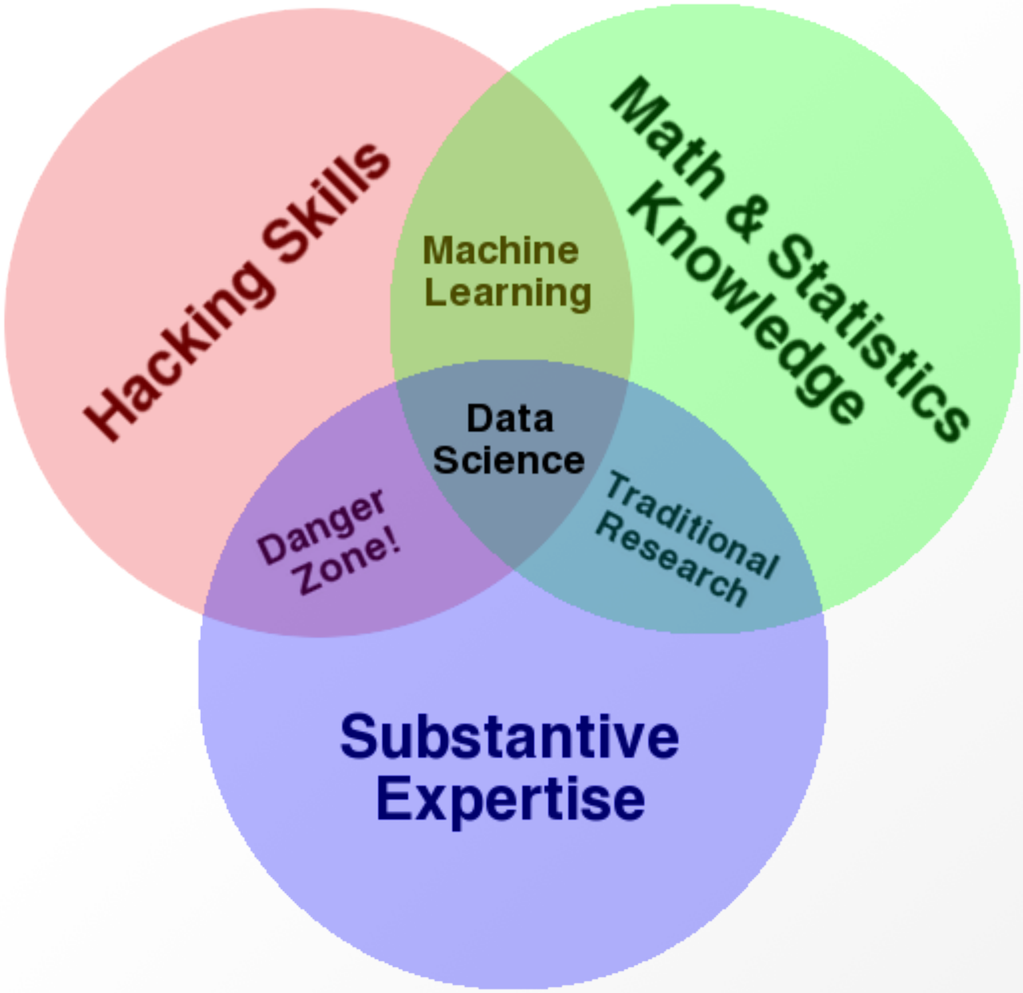
Data Processing

- Clean, transform, filter, aggregate, impute
- Convert into X and Y

Machine Learning

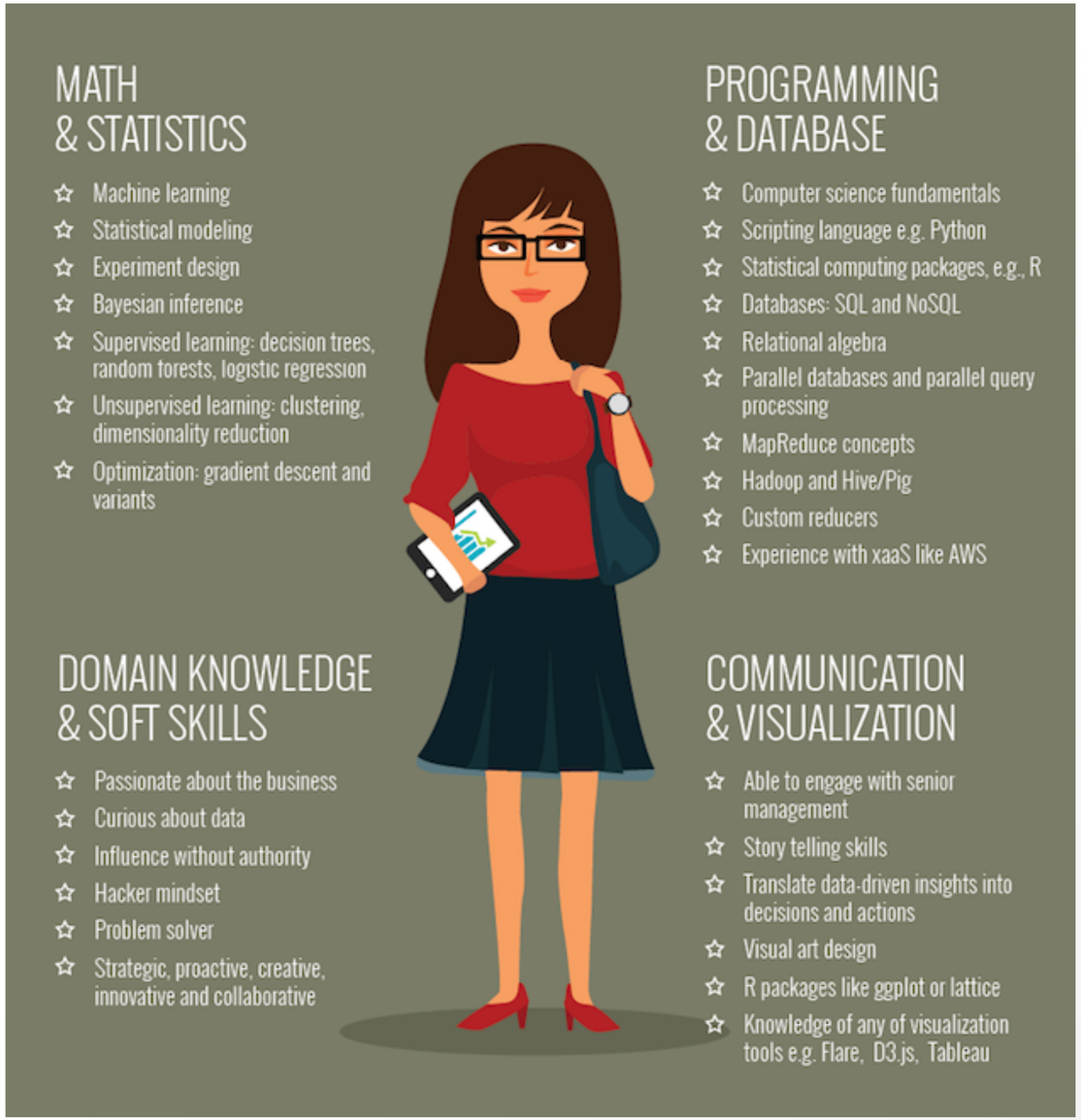
- Train models
- Evaluate models

The Data Science Venn Diagram

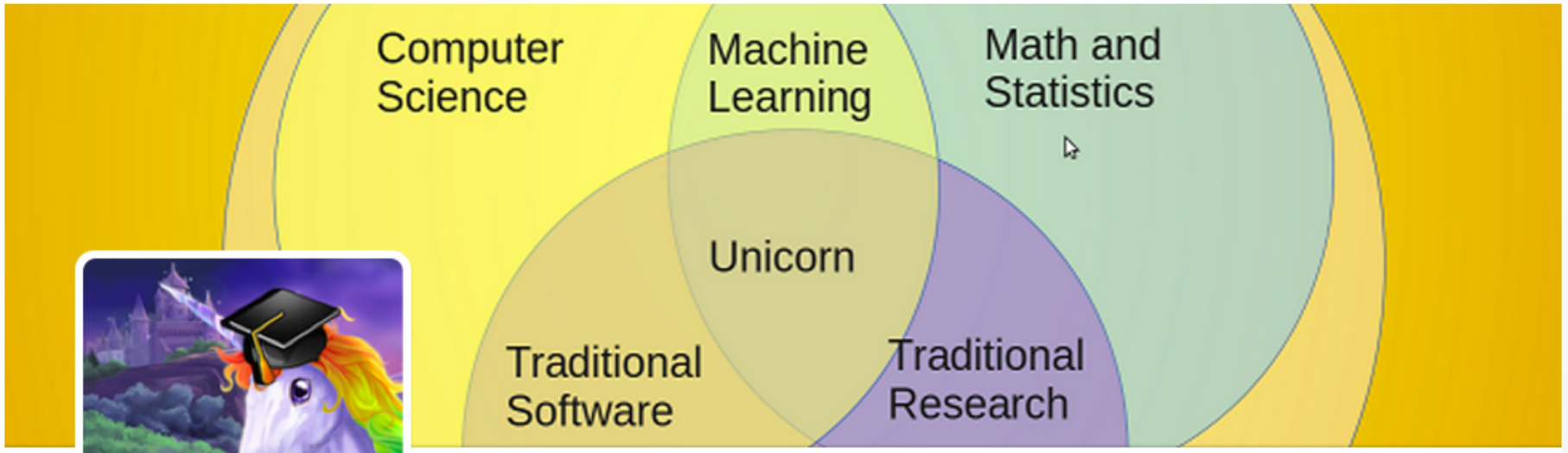


Drew Conway (2010)

The Data Scientist



The Data Scientist "Unicorn"



TWEETS 715 | FOLLOWING 357 | FOLLOWERS 435 | LIKES 2,411 | LISTS 9

Following

Data Science Unicorn
@DataScienceUni

All things #DataScience on the path to becoming the elusive unicorn #datascientist.

DataScienceUnicorn.com

Tweets | Tweets & replies

Data Science Unicorn @DataScienceUni · 5h
The Data Science Unicorn Daily is out! paper.li/e-1433455671?e…
#datascience Stories via @onwebanalytics @NYCDataSci @millie

View summary

Survey of Data Scientists on LinkedIn

The number of data scientists has doubled over the last 4 years.

The top five skills listed by data scientists:

1. Data Analysis
2. R
3. Python
4. Data Mining
5. Machine Learning

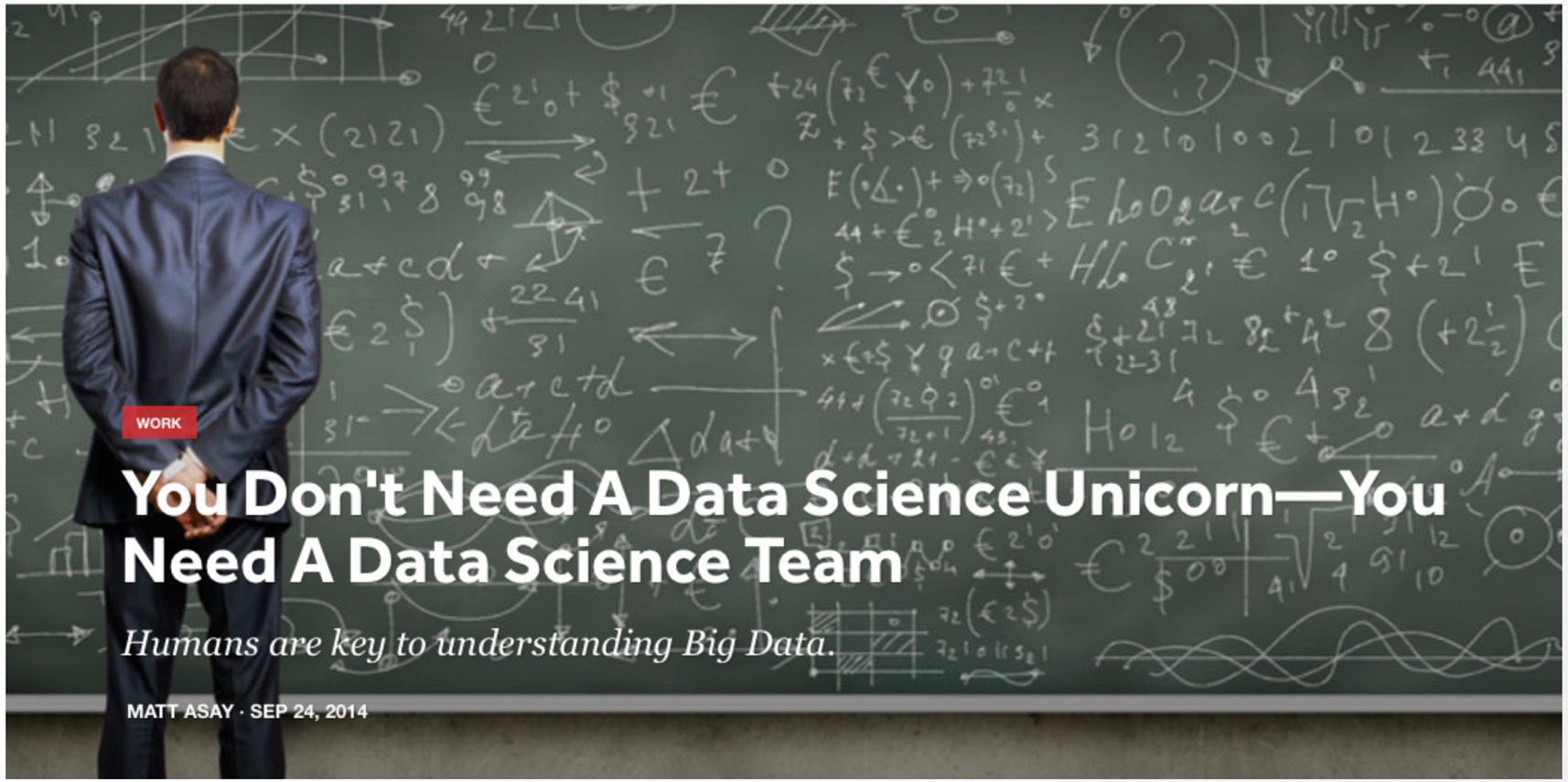
H2O WORLD

From Data Unicorns to Data Teams

readwrite

NEWSLETTER
Sign Up To Get ReadWrite In Your Inbox

CODE BUILD STRUCTURE CONNECT LEAD FUND OPERATE GROW



H2O WORLD

Data Science Teams

Data Analysts

- Strong data skills and the ability to use existing data analysis tools
- Able to communicate and tell a story using data

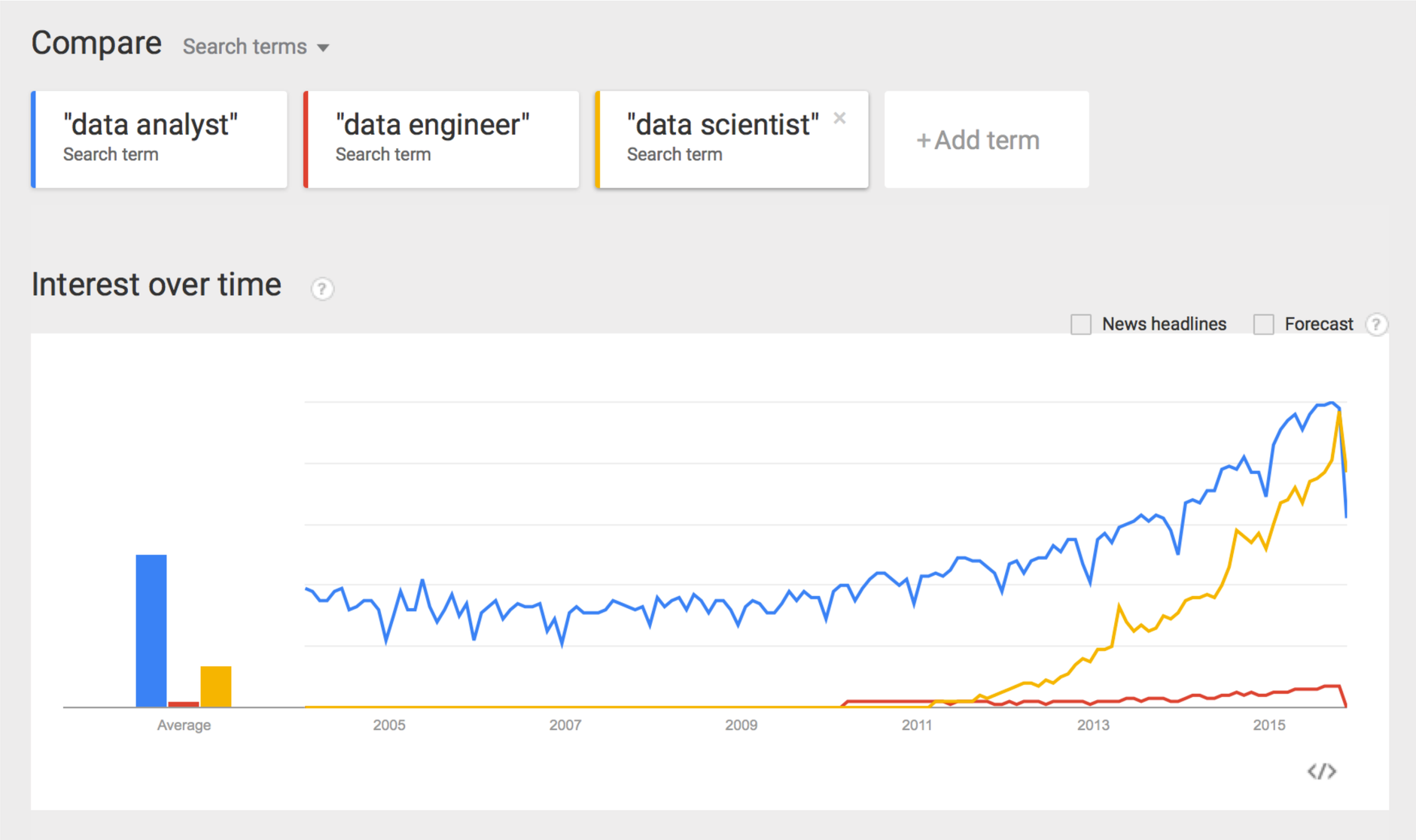
Data Engineers

- Usually a background in computer science or engineering
- Very good programming and DevOps skills

Data Scientists

- Strong math/stats background in addition to programming ability
- Understanding of machine learning algorithms

Data Science Teams



Data Science in the Enterprise

- Data Science teams develop “actionable insights” for business.
- They provide decision makers with information, guidance and confidence in the decision making process.

- Competitive advantage
- Cost minimization
- Data-driven products

Data Science in the Enterprise

Don't be a data dinosaur.

Embrace the data!

H2O WORLD

Data Science Tools

readwrite

NEWSLETTER Sign Up To Get ReadWrite In Your Inbox

CODE BUILD STRUCTURE CONNECT LEAD FUND OPERATE GROW

WORK

Python Displacing R As The Programming Language For Data Science

R remains popular with the PhDs of data science, but as data moves mainstream, Python is taking over.

MATT ASAY · NOV 25, 2013

12.3K SHARES

Data Science Wars: Python vs. R

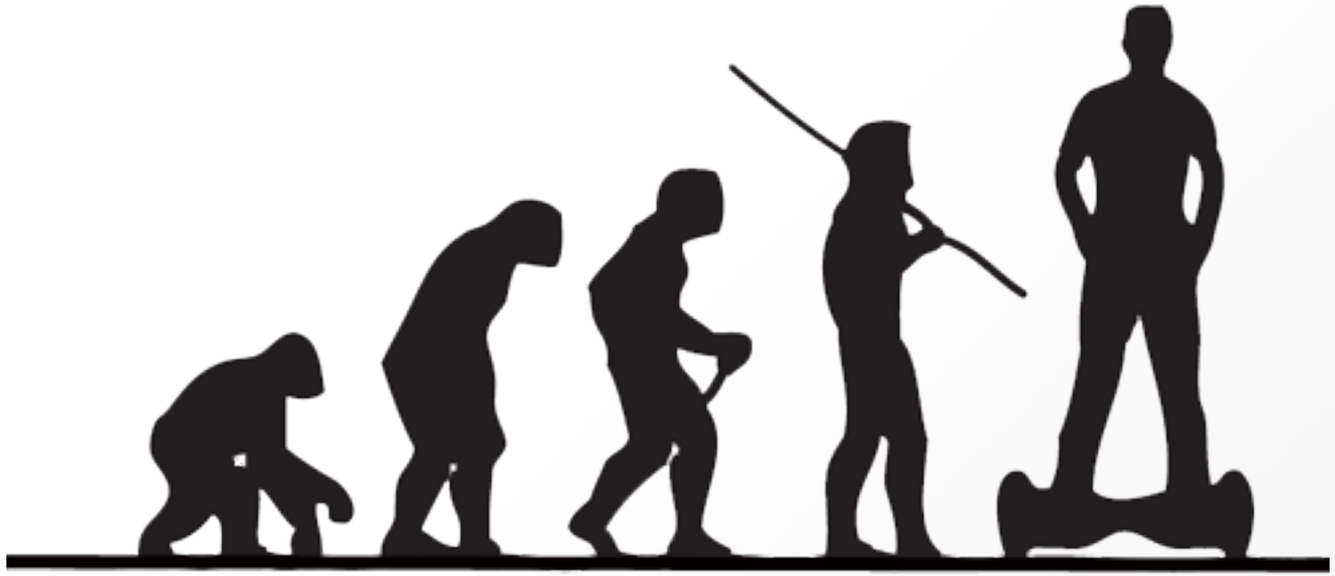
December 9, 2013 by Daniel Gutierrez 4 Comments

As I frequently travel in data science circles, I'm hearing more and more about a new kind of tech war: *Python vs. R*. I've lived through many tech wars in the past, e.g. Windows vs. Linux, iPhone vs. Android, etc., but this tech war seems to have a different flavor to it. What feels different in this case is that the application area is the same, namely performing work in data science where the solution often depends on the use of libraries that implement various machine learning algorithms. This being the case, the question is what language should you adopt as a data scientist?

2013 was the year of the
data science "language wars."

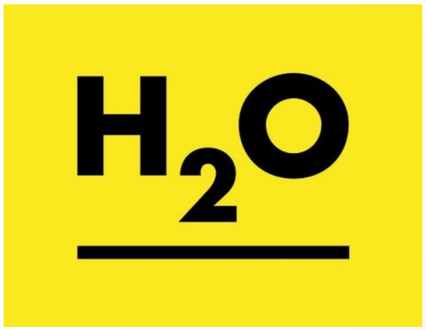
H2O WORLD

Data Science Tools



In 2015, we have evolved beyond this…
We are too busy doing actual data science!

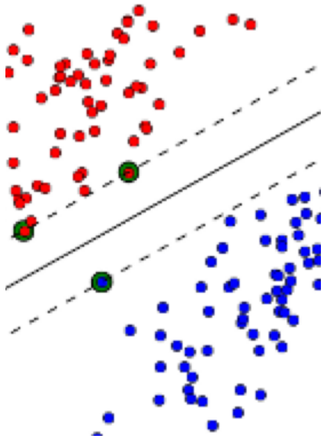
Data Science Tools



We are headed toward language agnostic data science, where friendly APIs connect to powerful data processing engines.

H2O WORLD

What is Machine Learning?



"Field of study that gives computers the ability to learn without being explicitly programmed."

— Arthur Samuel, 1959

Unlike rules-based systems which require a human expert to hard-code domain knowledge directly into the system, a machine learning algorithm learns how to make decisions from the data alone.

Machine Learning Tasks

Regression

- Predict a real-valued response (e.g. viral load, price)
- Gaussian, Gamma, Poisson, etc. distributed response
- Evaluate with MSE or R^2

Classification

- Multi-class or binary classification
- Ranking (e.g. Google Search results order)
- Evaluate with Classification Error or AUC

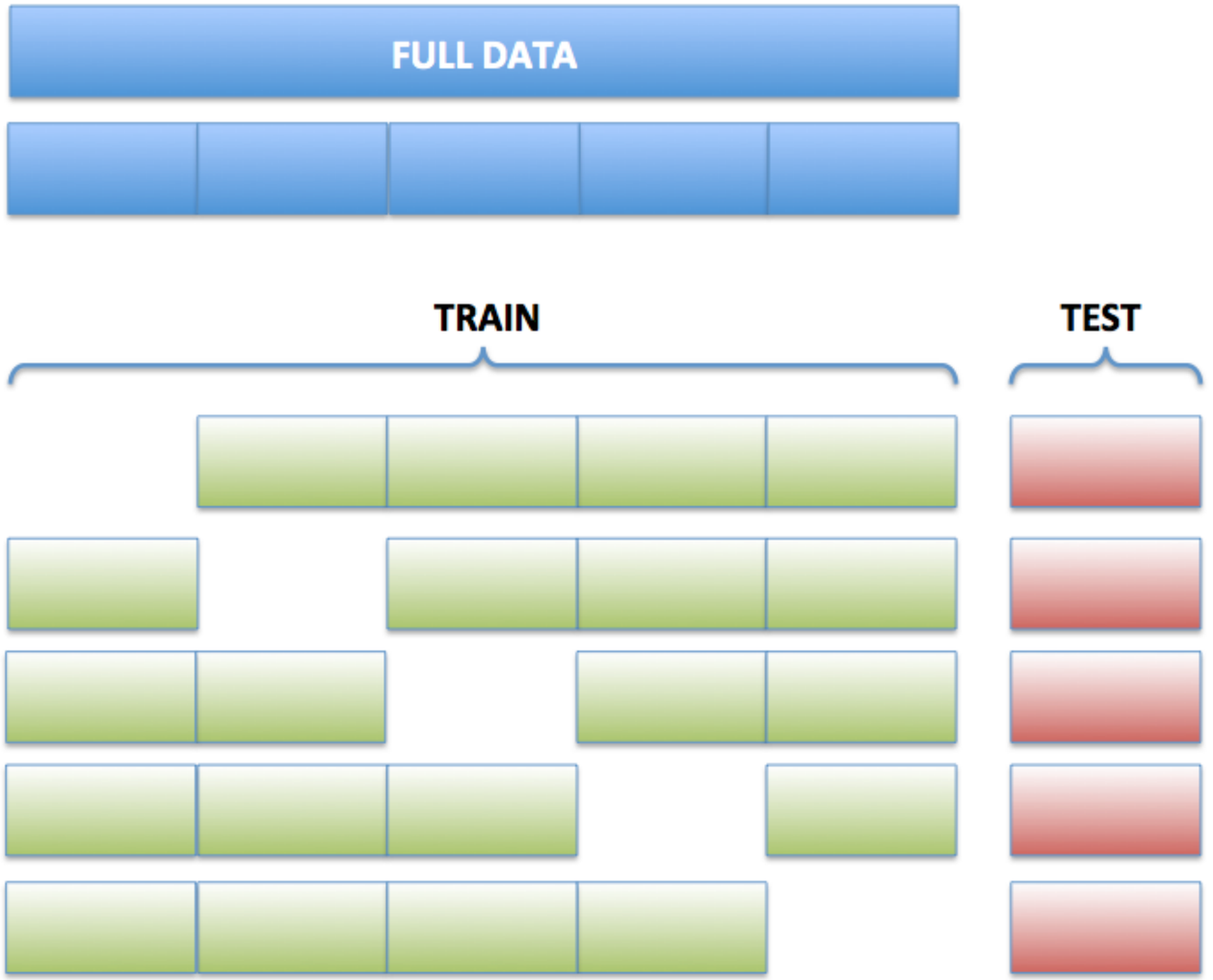
Clustering

- Unsupervised learning (no training labels)
- Partition the data; identify clusters or sub-populations
- Evaluate with AIC, BIC or Total Sum of Squares

Train, Validation and Test Set

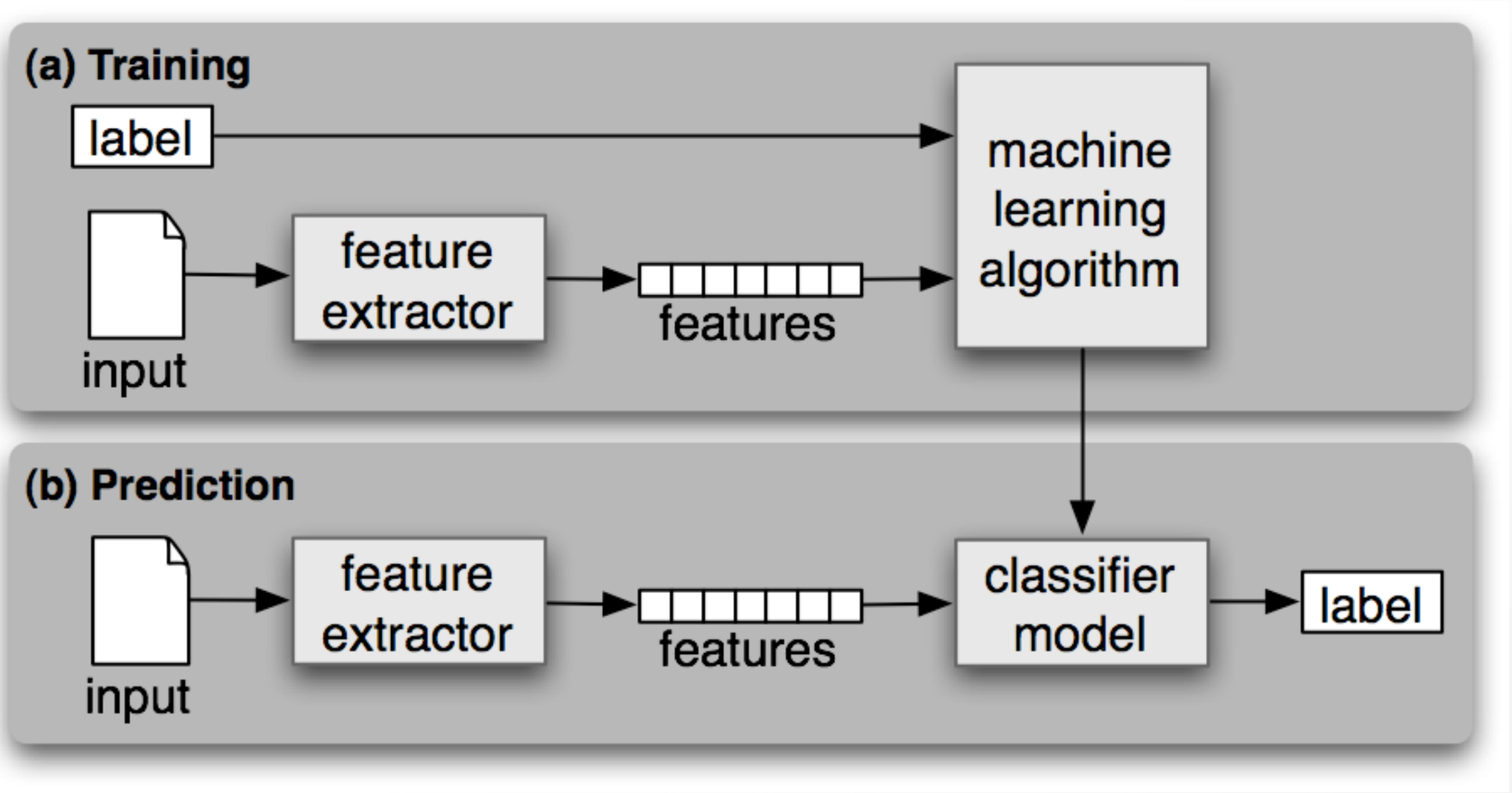
- If you plan on doing any model tuning, you should split your dataset into three parts: Train, Validation and Test
- There is no general rule for how you should partition the data and it will depend on how strong the signal in your data is, but an example could be:
  50% Train, 25% Validation and 25% Test
- The validation set is used strictly for model tuning (via validation of models with different parameters) and the test set is used to make a final estimate of the generalization

K-fold Cross-validation



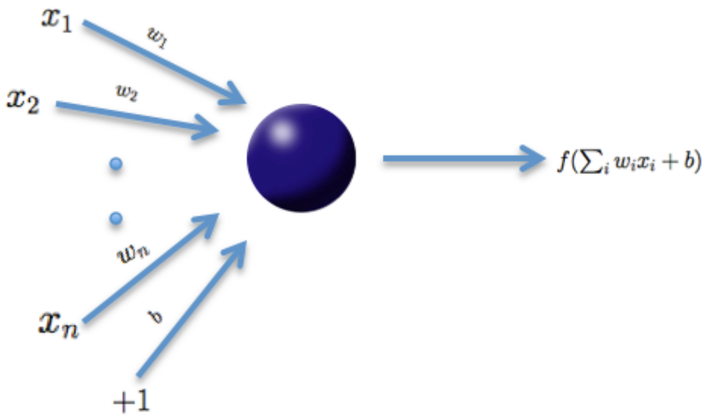
- K-fold Cross-validation (CV) is used to evaluate the performance of machine learning algorithms.
- CV will give you the most "mileage" on your training data.
- Performance metrics are averaged across k folds.

Machine Learning Workflow



Training and Prediction
in machine learning

What is Deep Learning?

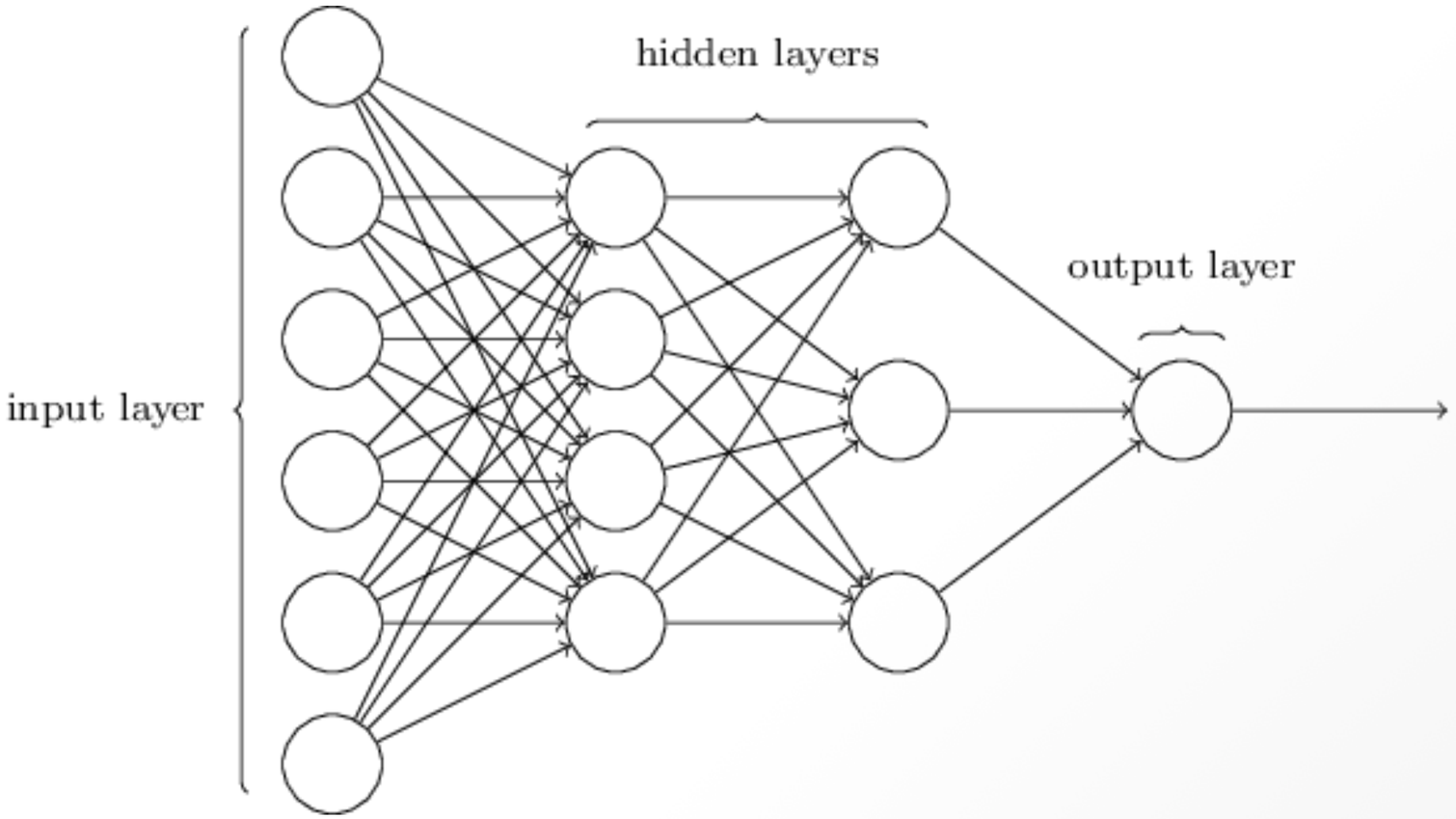


"A branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using model architectures, composed of multiple non-linear transformations."

— Wikipedia (2015)

- Deep neural networks have more than one hidden layer in their architecture. That's why they are called "deep" neural networks.
- Very useful for complex input data such as images, video, audio.

What is Deep Learning?



- Deep learning architectures, specifically artificial neural networks (ANNs) have been around since 1980.
- However, there were breakthroughs in training techniques that lead to their recent resurgence in the mid 2000's.
- Combined with modern computing power, they are quite effective.

What is Ensemble Learning?

"Ensemble methods use multiple learning algorithms to obtain better predictive performance that could be obtained from any of the constituent learning algorithms."

— Wikipedia (2015)

- Random Forests and Gradient Boosting Machines (GBM) are both ensembles of decision trees.
- Stacking, or Super Learning, is technique for combining various learners into a single, powerful learner using a second-level metalearning algorithm.

No Free Lunch

"Even after the observation of the frequent or constant conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience."

— David Hume (1711-1776)

- No general purpose algorithm to solve all problems.
- No right answer on optimal data preparation.
- Some algorithms may have such strong biases that they can only learn certain kinds of functions.

Where to Learn More?

- H2O Online Training (free): http://learn.h2o.ai
- H2O Slidedecks: http://www.slideshare.net/0xdata
- H2O Video Presentations: https://www.youtube.com/user/0xdata
- H2O Community Events & Meetups: http://h2o.ai/events
- Machine Learning & Data Science courses: http://coursebuffet.com