# Big Data Terms, Tools and Algorithms

What i’ve learned in the past 12 months

Kenneth P. Sanford, Ph.D.

eKenomics@gmail.com

@eKenomics

# outline

What I've learned in the past year

Economists as "storytellers" and analytics architects in this space

The rise of ML and AI

- What is ML?
- Why ML is here to stay.

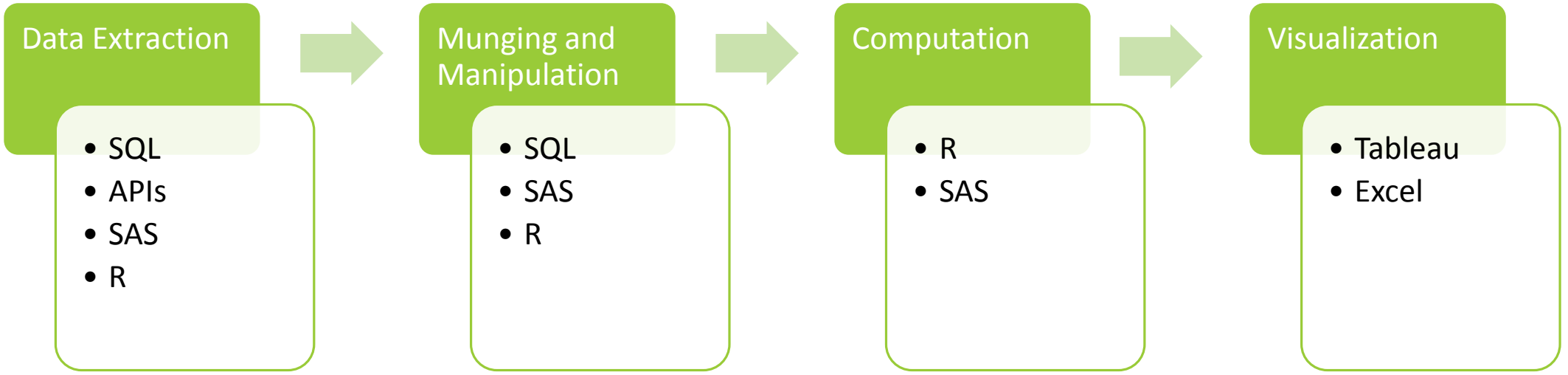
Technological changes (Spark, Streaming)

Language changes (SAS→R→Python)

Methodological changes (Deep Learning, Online Learning)

What Economists should do to learn

# Economists in Data Science (Year Ago)

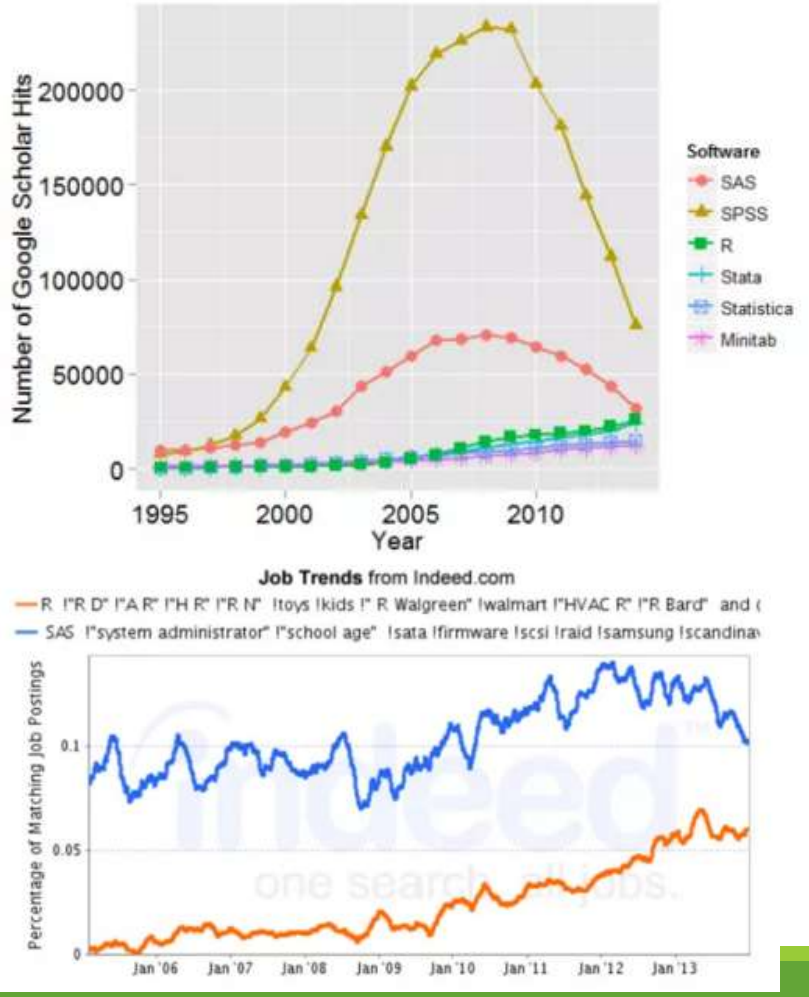


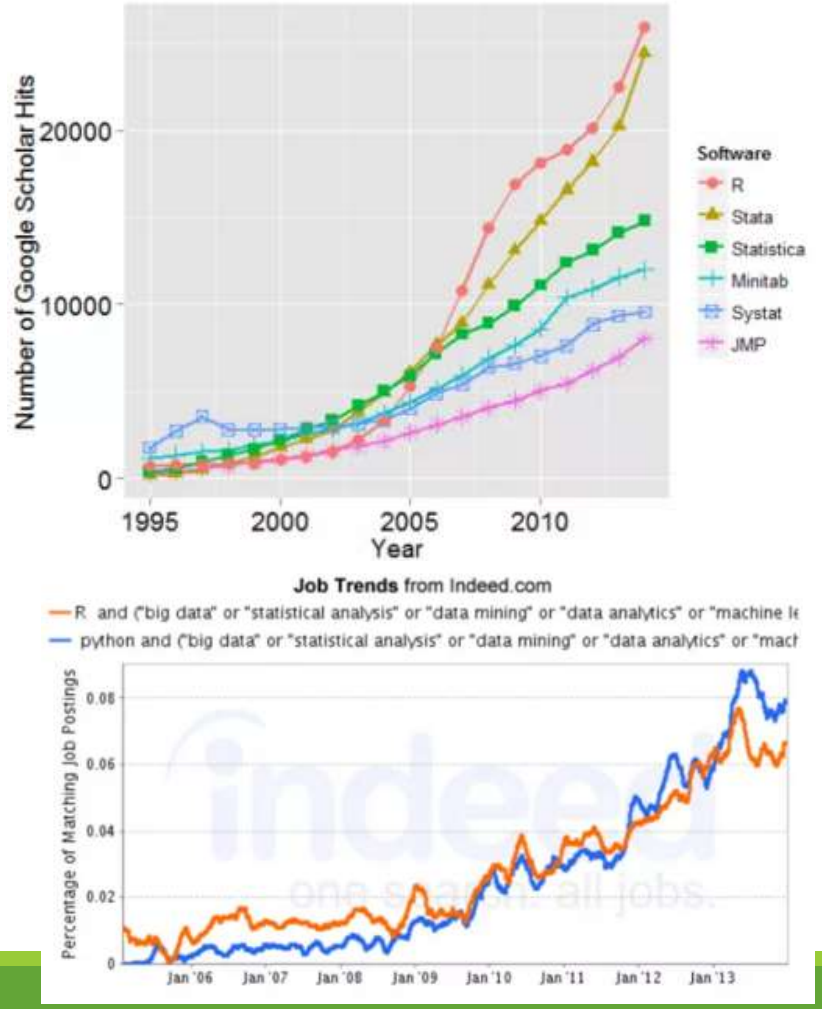
# Why Economics is a Data Science

- Understand objective functions
  - Academic answer vs. useful answer
- Great storytellers
- Solid visualization skills
- Observational data and causality
- Interdisciplinary training
- Solid knowledge of regression (background of predictive modeling)

# Over the past year…

# Proprietary to Free AND Open Source



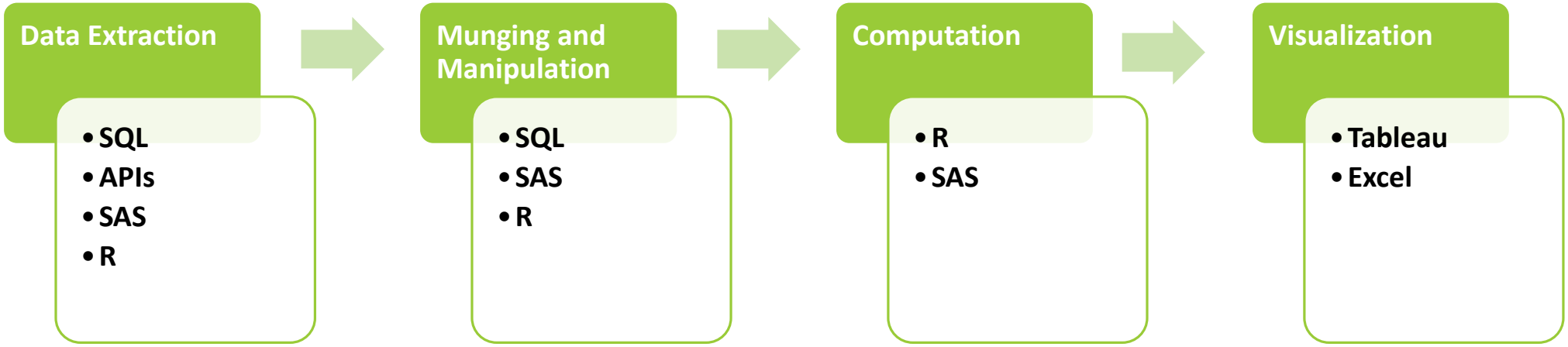

http://r4stats.com/articles/popularity/

# Analytics for Operations vs. Analytics as Product

"Hypothetical" Examples

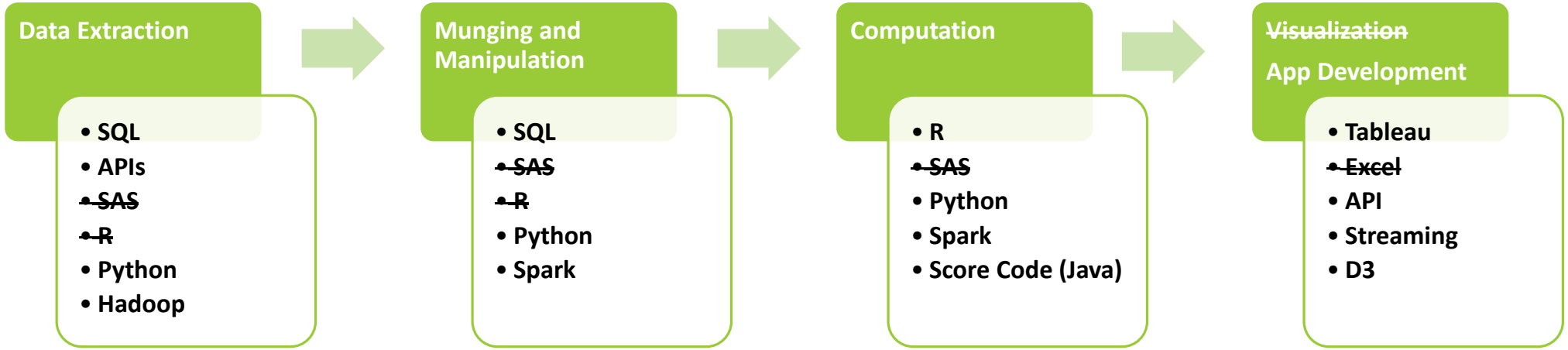
- Amazon:
  - OA: Retail needs estimates of cost of delivery, timing, etc.
  - AP: Create and sell customer cross-sell data (Customer 360)
- UBER:
  - OA: Where to suggest that drivers locate
  - AP: Targeted list of drivers for maintenance coupons
- LinkedIn
  - OA: Who you may know
  - AP: Who might buy machine learning software

# This cloud stuff is real……




**Static Analytics (On Premise)**

---

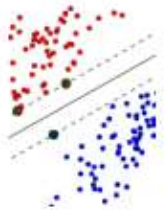



**Production Analytics (Cloud)**

# Machine Learning and Artificial Intelligence

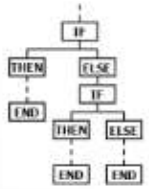
# Machine Learning and Artificial Intelligence

## What it is:



- "Field of study that gives computers the ability to learn without being explicitly programmed." (Samuel, 1959)
- "Machine learning and statistics are closely related fields. The ideas of machine learning, from methodological principles to theoretical tools, have had a long pre-history in statistics." (Jordan, 2014)
- M.I. Jordan also suggested the term ***data science*** as a placeholder to call the overall field.

## What it's not:




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 and AI Vocabulary

| Concept | Statistics\Econometrics | Machine Learning |
|---|---|---|
| “Computation” | Fit\Estimate | Train |
| “Left-hand side” | Dependent variable | Target |
| “Right-hand side” | Regressor\Predictor\Class | Feature\Factor\Enum |
| “Goal” | Estimation\Explanation | Prediction |

# Statistics vs. Machine Learning

Statistics: Good estimators are….

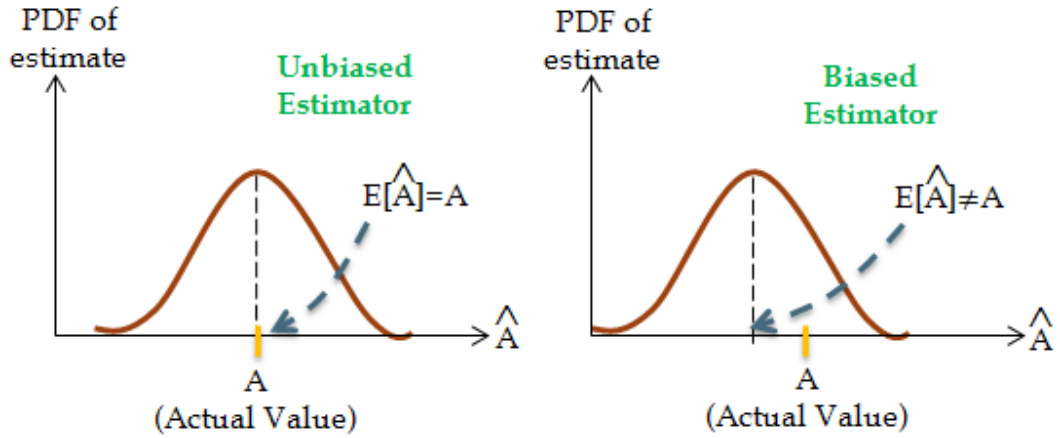
Unbiased in small samples

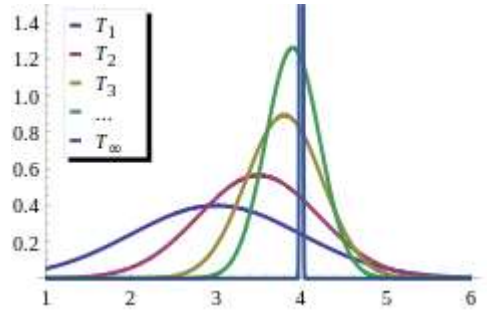
Consistent if not unbiased

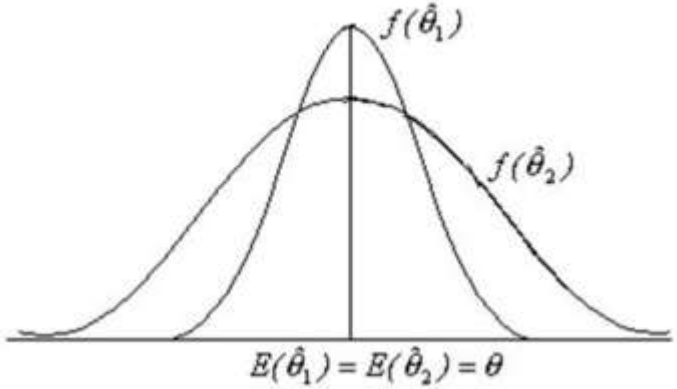
Efficient

Machine Learning: Good models….

Predict well.

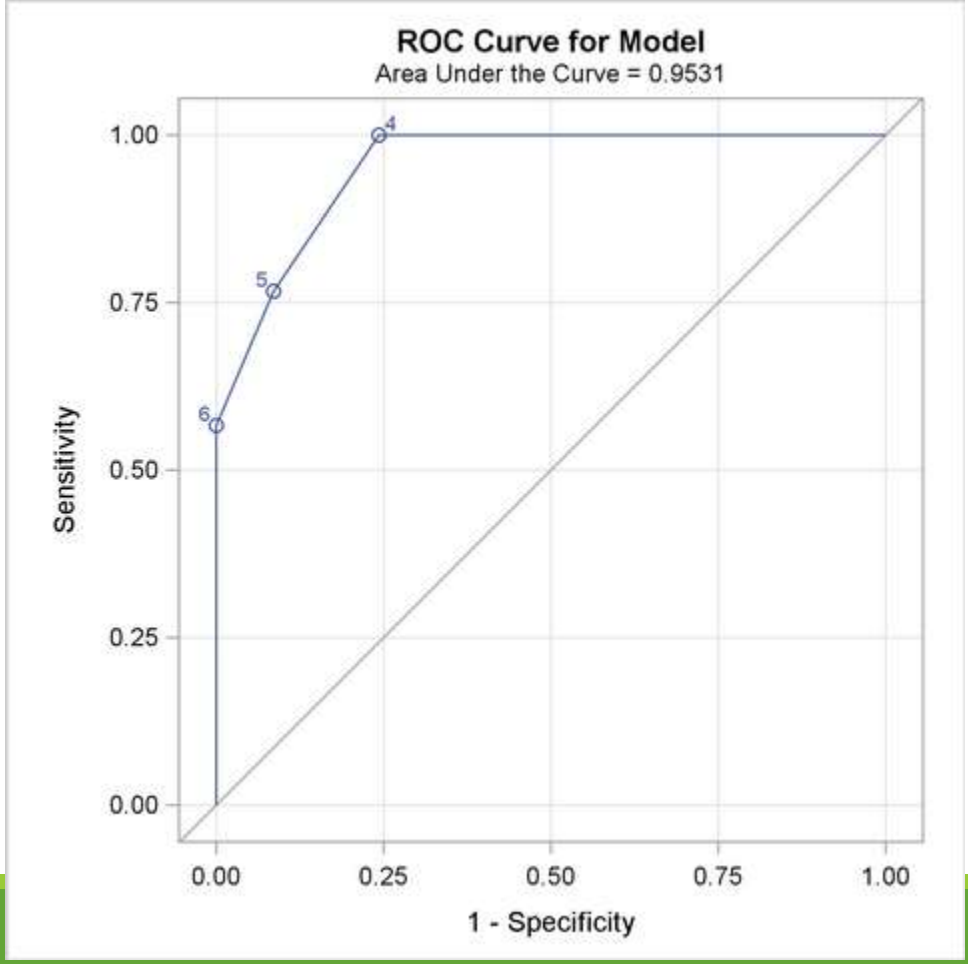


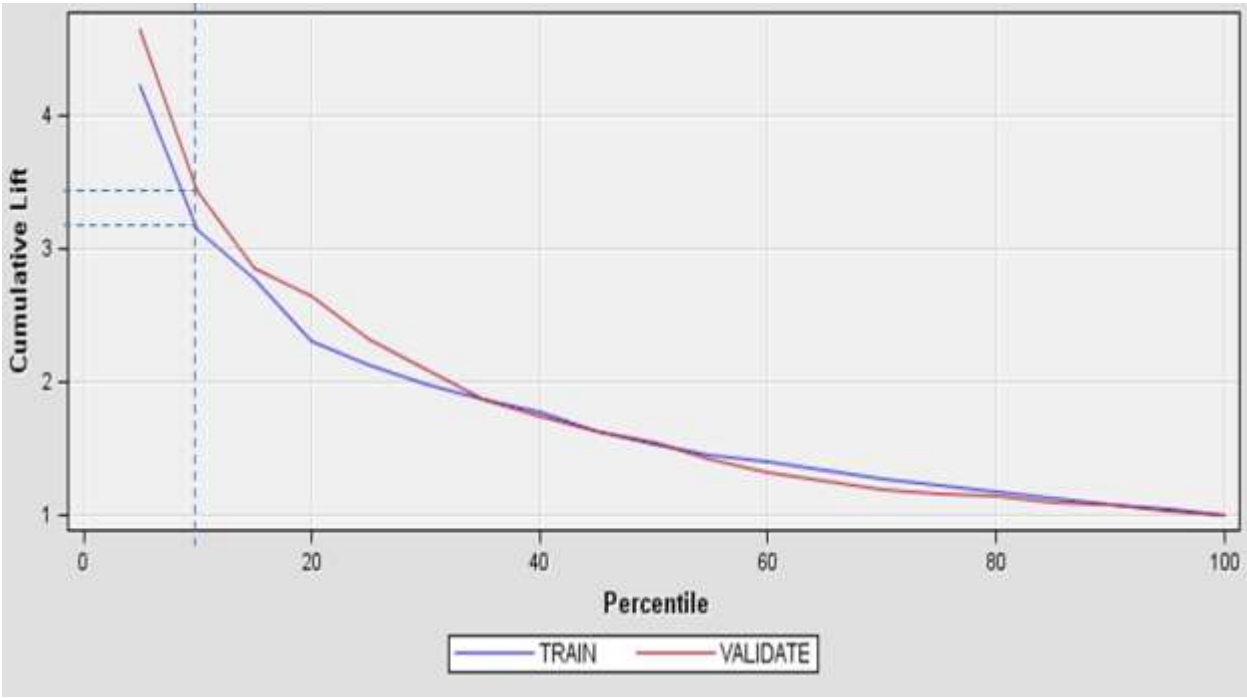



# Diagnostics: Evaluating Your Model

## Receiver Operating Characteristic (ROC) Curve




## Lift Curves



## Supervised Learning Methods

- Regression (GLM)
  - Lasso
  - Ridge
  - E
- De
- Ra
- Gradient Boosted Models
- Support Vector Machine
- Neural Network
- Deep Learning

Know Y

## Unsupervised Learning Methods

- Clustering
  - Kmeans
  - Hierarchical
- Princ
- Auto
- Non-
- Generalized Low Rank Models

Don’t know Y

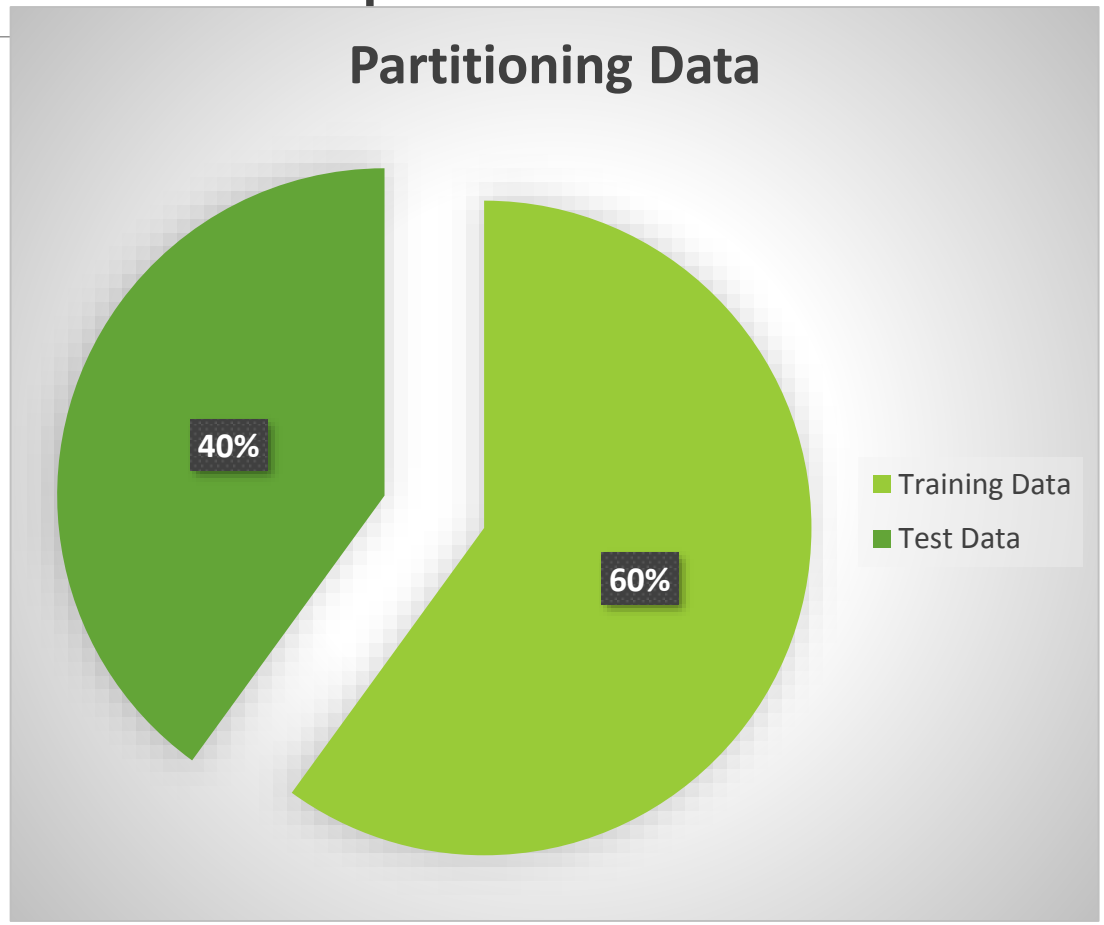
# Fitting: Training and Test Samples

Why?

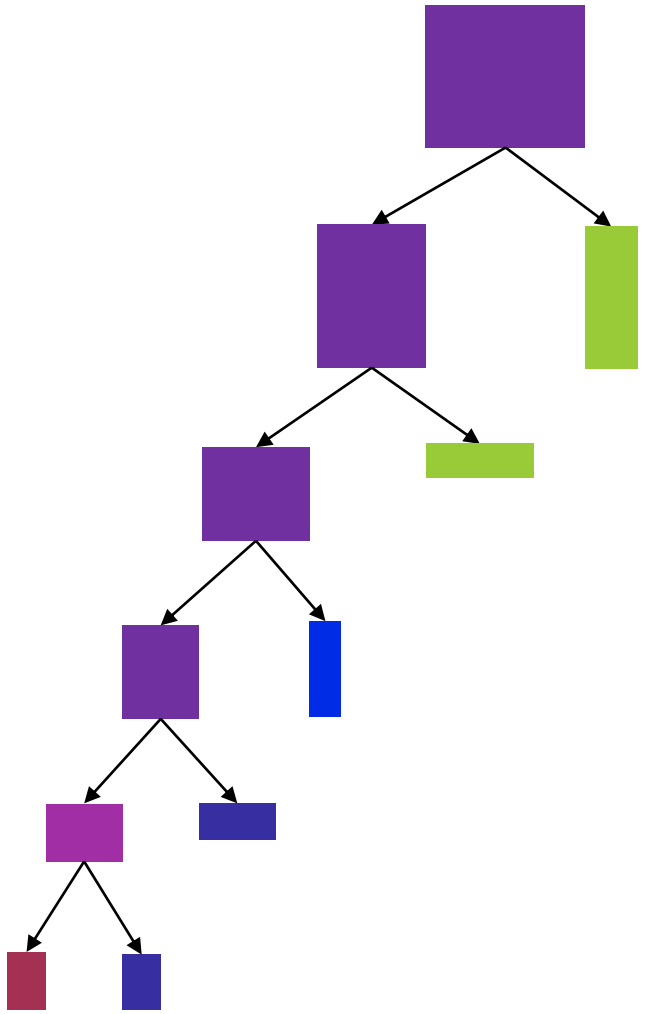
- As our objective is to predict, and with “big data” we might have lots of observations, lets reserve some data to objectively evaluate out-of-sample performance.

Train: Estimate one or more models

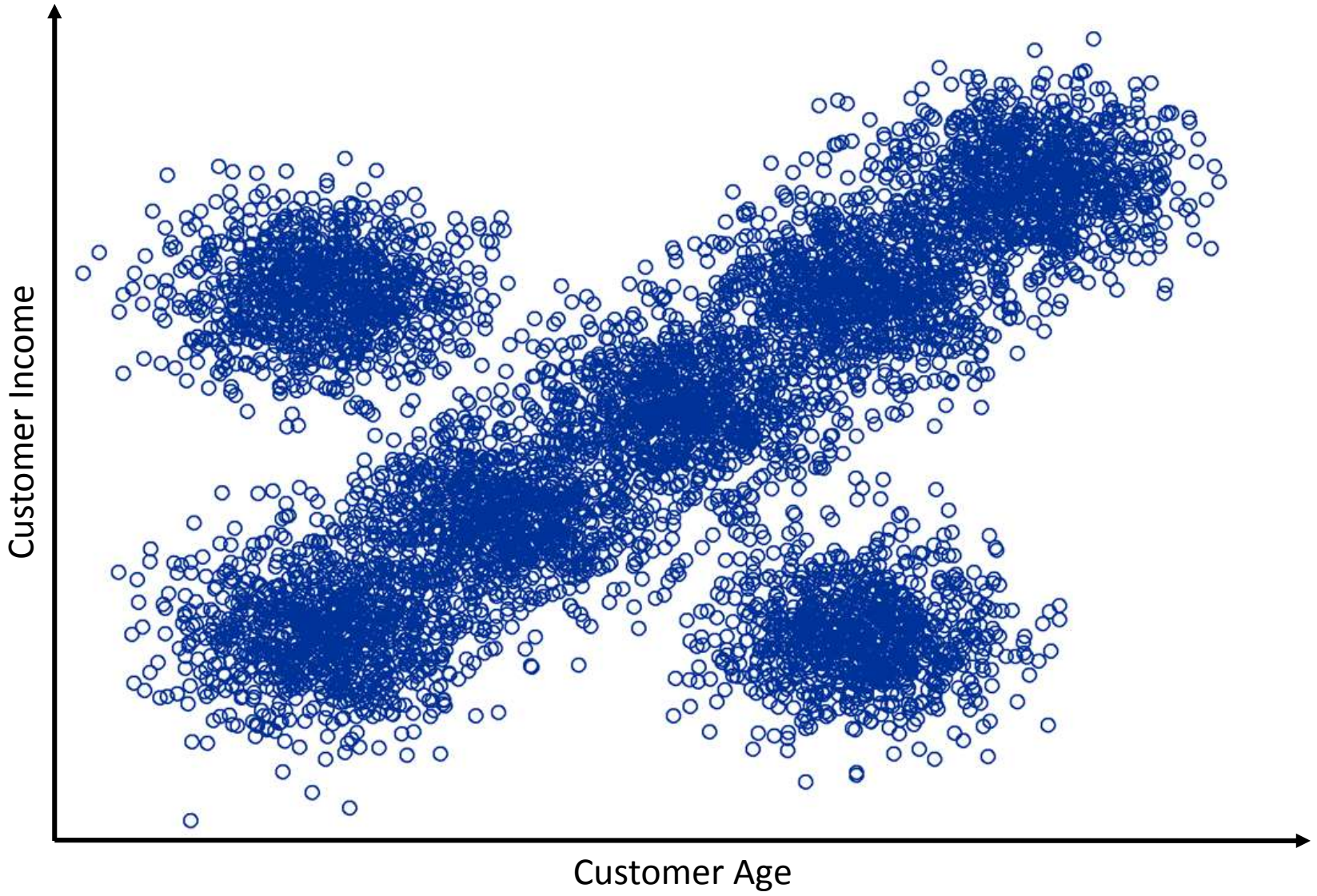
Test: Compare the predictive qualities of the model



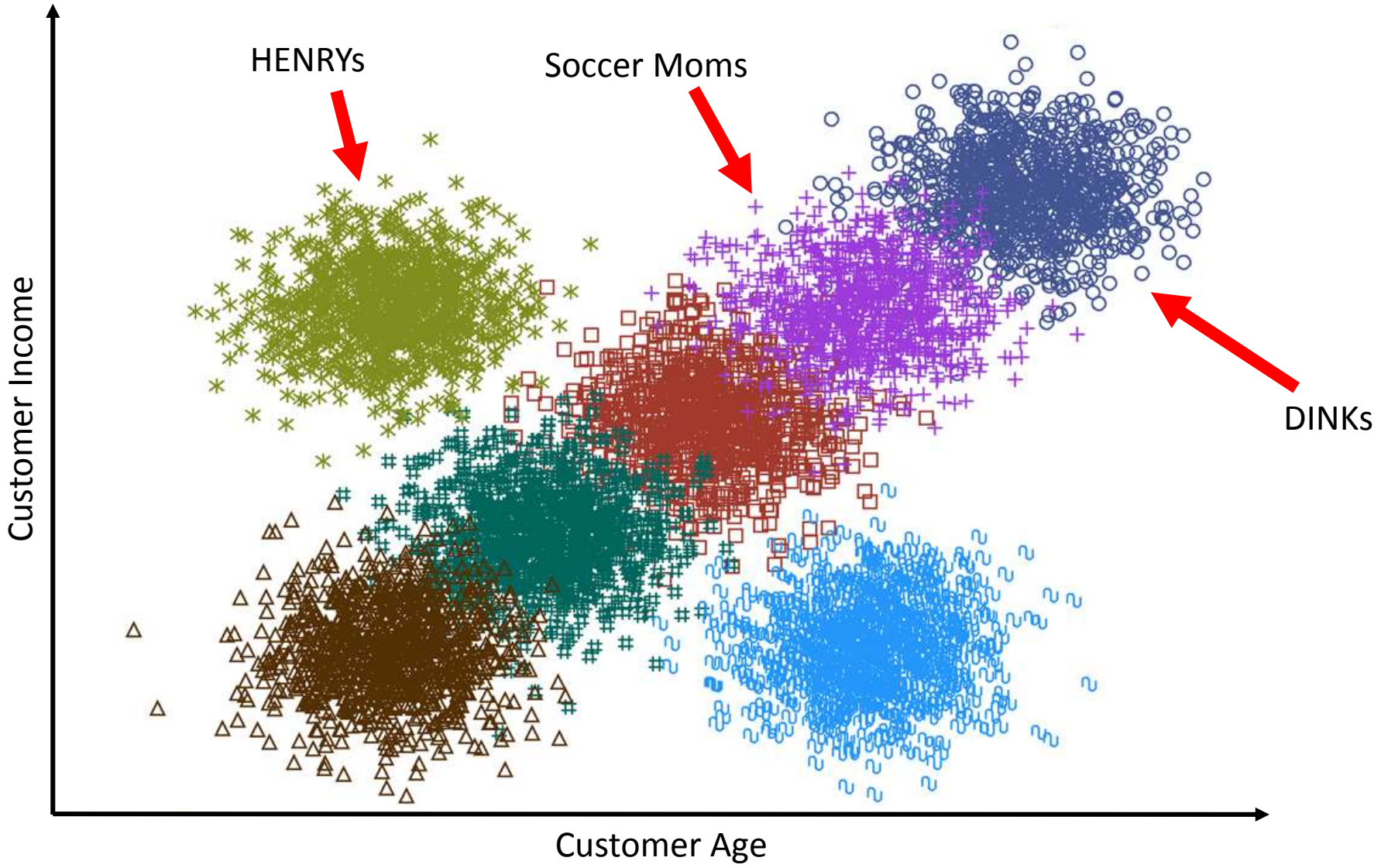
# Decision Tree

# Clustering



# Clustering



# Algorithm Improvements and Adoption

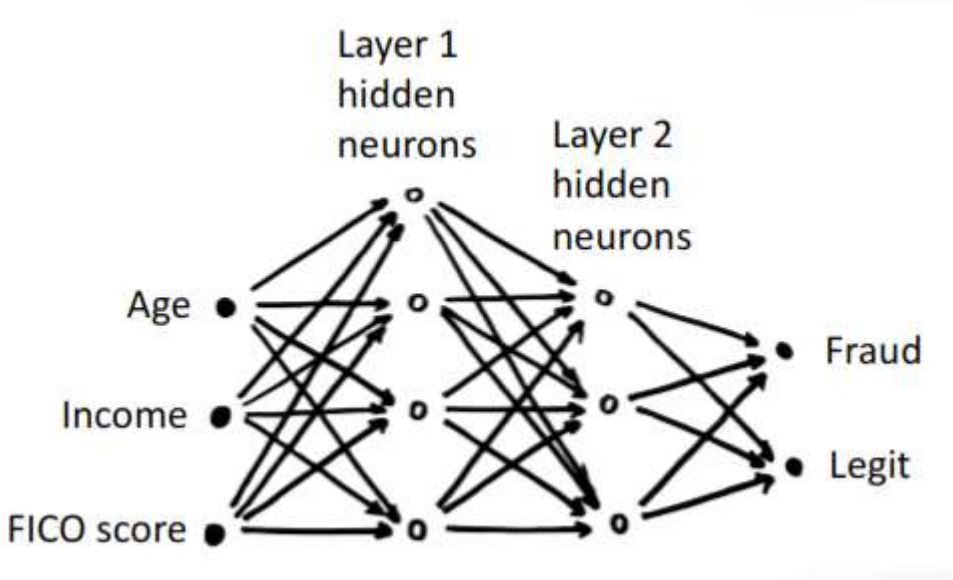
Deep Learning learns a **hierarchy of non-linear transformations**

Neurons transform their input in a non-linear way

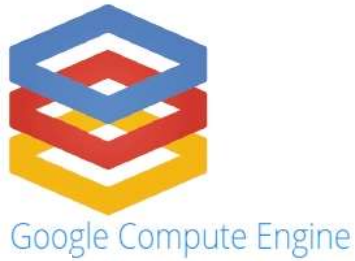
**Black-box, brute-force method, really good at pattern recognition**

Deep Learning got a boost in the last decade due to **faster hardware and algorithmic advances**

**Great for image recognition/text**



# Software Technologies of “Big” ML






# What Economists can do

# Where to learn more

- DataCamp
- Codeschool
- Kaggle Competitions

- Read Athey, Imbens, Bajari, etc.

- Conferences
  - ODSC
  - MLConf