P9120 - Statistical Learning and Data Mining

Lecture I - Overview of Statistical Learning

Min Qian, PhD

September 5, 2024

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

Outline

Machine Learning and Data Mining

- Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information, with an emphasis on large observational data bases.

- Machine learning concerns with design of algorithms that allow machines (computers) to learn from data or learn from examples.
  - The term "machine learning" comes from the artificial intelligence community, but is now a focus area in many branches of statistics, computer science and applied math

Traditional Statistics (A Hypothetical Example)

Research questions:

1. Hypothesis testing: Does treatment result in better outcome than control?
2. Effect estimation: What is the treatment effect on outcomes after controlling for other covariates?
3. Prediction: What is a good predictive model for the outcome of interest?

Traditional Statistics: Hypothesis Testing

1. Hypothesize:

   The new treatment may result in better outcome than control.

2. Collect data (e.g. randomized trials)

3. Analyze data (t-test, z-test, etc.)

4. Draw conclusion (based on effect size, p-value, confidence interval, etc.)

Traditional Statistics: Effect Estimation

What is the treatment effect on outcomes (after controlling for other covariates)?

- Often model-oriented

$$\text{Outcome} = \alpha + \beta * \text{treatment} + \text{adjusted variables},$$

  - β measures the size of the effect of treatment as compared to control.

- Estimate β using linear regression (e.g. $\beta = 1$ with p-value$= 0.01$).

- Interpretation: Holding the values of all adjusted covariates unchanged, on average the treatment will result in a significant increase of 1 unit in outcome as compared to control.

Traditional Statistics: Prediction

What is a good predictive model for the outcome of interest?

- Often model-oriented

$$\text{Outcome} = \beta_0 + \beta_1 * \text{treatment} + \beta_2 * age + \beta_3 * male,$$

- Estimate βs using linear regression (e.g. $\beta_0 = 5, \beta_1 = 1, \beta_2 = 0.05, \beta_3 = 0.5$).

- Interpretation: For a 40 years old male patient who would be treated with treatment, his predicted outcome would be $5 + 1 + 0.05 * 40 + 0.5 = 8.5$ (with 95% confidence interval xxx)

Traditional Statistics: Summary

- First hypothesize, then collect data, then analyze
- Often model-oriented

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots,$$

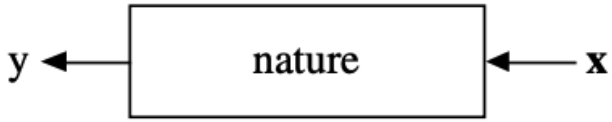
  where Y is the clinical outcome, and X includes treatment, demographic variables, commodities, etc.

- Emphasis on inference (p-value, confidence interval, etc.)
- Care about interpretation, causality, etc.

Machine Learning in Computer Science

- Emphasis on fully automatic methods. often algorithm-oriented
- Focus on computational tractability (i.e. how long and how many CPUs will it take to compute this?)
- Care less on sample size, p-values, causality, etc.
- Gold standard: prediction performance in the population

Statistical Modeling: The Two Cultures (Leo Breiman, 2001)

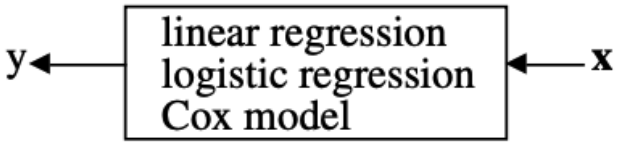


Two goals:

Prediction. To be able to predict what the responses are going to be to future input variables;

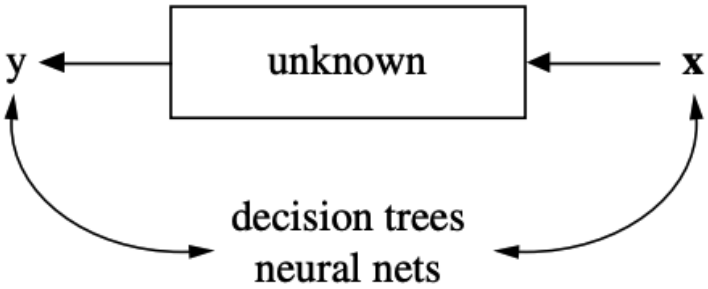
Information. To extract some information about how nature is associating the response variables to the input variables.

The Data Modeling Culture



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

Statistical Machine Learning

- Emphasize statistical analysis and methodology.
- Provides theoretical foundations for learning algorithms
- Gives useful tools to analyze an algorithm's statistical properties and performance guarantee
- Helps researchers deepen understanding of the approaches, design better algorithms, and select appropriate methods for a given problem.

Machine Learning Types

- Supervised Learning (Y observed) X Y
- Unsupervised Learning (Y unobserved) X
- Reinforcement Learning X<-A Y
- Semi-supervised Learning (Y partially observed)
- Active Learning
- ...

Broad Overview of Syllabus

- Overview of Machine Learning
- Machine learning methods for regression.
- Machine learning methods for classification.
- Deep learning for different types of tasks.
- Topics in unsupervised learning.
- Reinforcement learning for decision making tasks.

Course Learning Objectives

- Be able to identify and formulate Supervised learning, Unsupervised learning and Reinforcement learning problems.
- Understand a range of machine learning algorithms along with their strengths and weaknesses.
- Be able to choose an appropriate statistical learning method to solve open biomedical or public health research problems.
- Be able to implement the algorithms using standard statistical software to perform data analysis.
- Be able to implement deep learning (and reinforcement learning) in python.

Understand how/when/why those methods work!

Prerequisite

- Strong background in calculus and linear algebra.
- Knowledge of linear and logistic regression models.
- Knowledge of probability theory and statistical inference.
- Familiarity with R or Python programming.
- Some working knowledge of optimization techniques would be desirable.

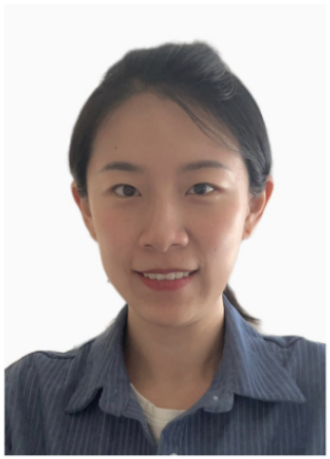
Class Details

- Time: 1pm-3:50pm Thursdays
- Classroom: Hammer 305
- Instructor info:

  Min Qian, PhD
  Office: 722 W 168th St., room 645
  Email: mq2158@cumc.columbia.edu
  Office hours by appointment

- TA info:

Yuqi Miao



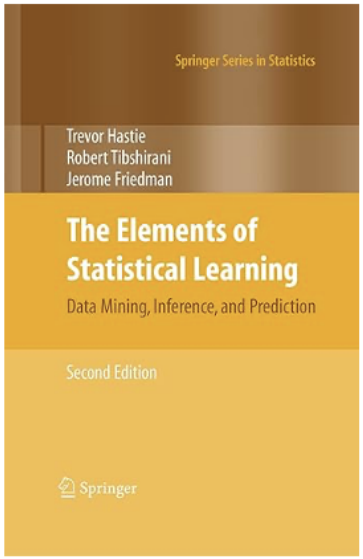
Bin Yang

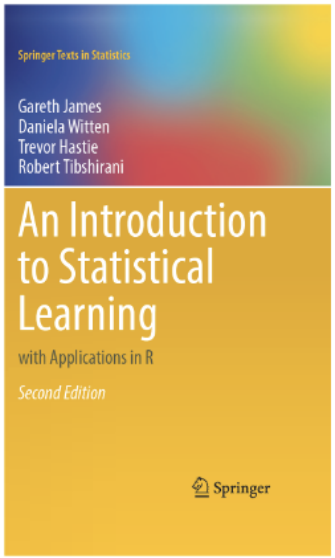
Yuqi Miao,
ym2771@cumc.columbia.edu
Bin Yang,
by2303@cumc.columbia.edu

Office Hour: TBA

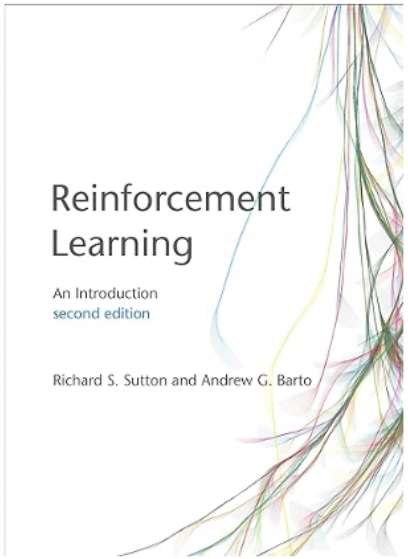
References

Machine Learning





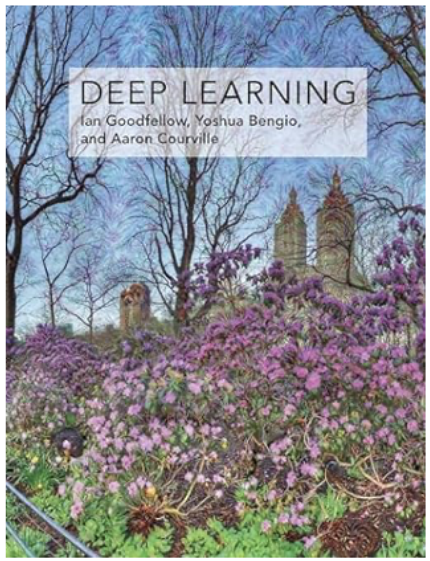
Reinforcement Learning



Deep Learning

DeepLearning.AI

Deep Learning Specialization — Course Slides



Software

- R: prior knowledge is required.
- Python: prior knowledge is not required. There will be 4 in-class lab sessions on Python.

Assessment and Grading Policy

- Homework: 40% (4 assignments, submit online)
- Class attendance and participation: 10%
- Weekly mini-quizzes: 15% (12 quizzes. The lowest 2 scores will be dropped.)
- Group paper presentation: 10% (instructor-assigned)
- Final project report: 25% (self-chosen, in consultation with instructor)

Homework

- May include analyses of real data sets, running simulation studies and derivations of theoretical properties
- Theoretical exercises can be hand-written, but data analysis/simulation part must be typed in using LaTeX/ MS Word
- Data analysis results should be incorporated in the main text, while the source code should be included in the Appendix.
- You are encouraged to work with your classmates on homework, but copying homework from someone is STRICTLY prohibited.
- Each homework assignment is due 2-3 weeks after it is given. Late homework will not be accepted.

Final Project

- In-depth exploration of one methodology.
- Do a thorough literature search, and cite any relevant work properly within your report.
- Include theoretical results and/or simulation studies and/or data analysis (comparison with other methods).
- Conclude with your results and discussion.
- Do not simply reproduce the results of a paper.
- Judged based on clarity, thoroughness, and originality.
- Reports must be $\leq$5 pages (not including tables, figures, and references), single spaced, 12 point font.
- A vert brief project proposal is due on November 21st.
- The Final Project Report is due on December 22th.

Supervised Learning

- Learn about the unknown relation between input X and output Y variables in a "supervised" way, i.e. through a set of examples in which both the inputs and outputs are given, during the "training" period.
  - Regression: Y continuous
  - Classification: $Y \in \{1, 2, \ldots, K\}$
- Training data: observed examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$.
  In this course, we consider the case that all observations are i.i.d.
- Goal: build a model $f(\mathbf{x})$, so that we can predict the output y when seeing a new input vector $\mathbf{x}$.

Supervised Learning Examples

Application	Input (X)	Output (Y)
Real Estate	home features	Price
Credit risk	user info	Likelihood to default
Medical diagnosis	patient characteristics	disease or not
Spam Filtering	Email	Spam or not
Image Classification	Image	Object
Speech recognition	Audio	text transcript
Machine Translation	English	Chinese

Terminologies

	Statistics	Machine Learning
$\mathbf{X}$	predictor, covariate	input, feature
Y	response, outcome	output
$\{(\mathbf{x}_i, y_i)\}_{i=1}^n$	data set, sample	training data, instances
θ, β	parameters	weights
main interest	bias, variance, sensitivity, specificity	prediction accuracy
Theory	consistency, inference, convergence rate	speed, risk bound, learning theory

Regression

- $(X, Y) \sim$ distribution $\mathcal{D}$, Y is continuous.

optim

- Loss $l(Y, f(X))$: measuring the discrepancy of Y and $f(X)$.
  - Squared error loss: $l(Y, f(X)) = [Y - f(X)]^2$ (commonly used).
  - Absolute error loss: $l(Y, f(X)) = |Y - f(X)|$ (less used).

sq - differentiable

- find beta

- Risk: prediction error (under squared error loss):

$$R(f) = El(Y, f(X)) = E[Y - f(X)]^2.$$

- Optimal prediction:

$$f^*(\mathbf{x}) = \arg\min_f R(f) = E(Y|X = \mathbf{x}).$$

Empirical Risk Minimization

- $\mathcal{D}$ is unknown, thus we can not compute $R(f)$.
- But the training data $(x_1, y_1), \ldots, (x_n, y_n)$ are i.i.d. from $\mathcal{D}$.
- Replace prediction error $R(f)$ by training error

$$R_{emp}(f) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2.$$

  $R_{emp}(f)$ is also called empirical risk.

- Least squares (Empirical Risk Minimization)

$$\min_f \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2$$

Linear Regression

- $f^*(\mathbf{x}) = E(Y|X = \mathbf{x})$: underlying truth. Unknown.
- There is no way to estimate $f^*(\mathbf{x})$ directly given a finite number of samples.
- We have to put some restrictions/structures on $f(\mathbf{x})$.
- Linear Model:

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

where β_j's are unknown parameters and β_0 is the intercept.

Ordinary Least Squares

- Estimation of $f(\mathbf{x})$ reduces to estimation of β_j's.

- Ordinary least squares estimate (OLS)

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \arg\min_{\beta_j' s} \frac{1}{n} \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \right]^2.$$

  Or in the matrix form

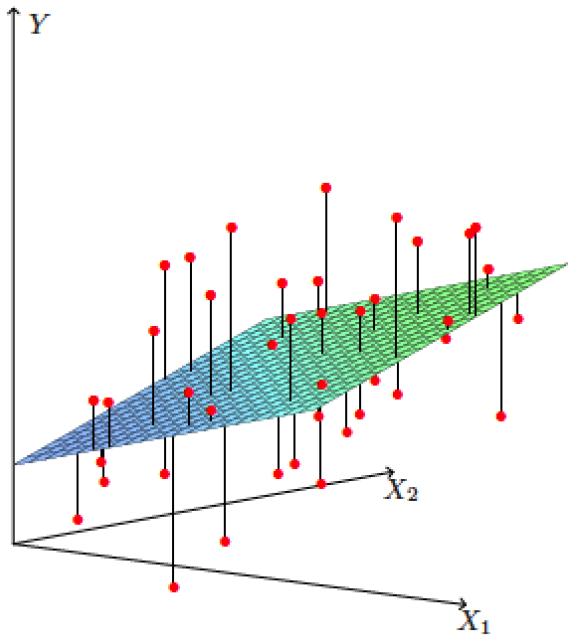
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

  where $\mathbf{y} \in \mathbb{R}^n$ is the vector of observed outcomes, $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ is the design matrix, and $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ is a vector of parameters.

- Fitted value $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. It is a very important quantity for diagnostics.

Why OLS estimate?

- OLS makes sense geometrically.



- We often assume

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i,$$

where the error terms $\epsilon_i, i = 1, \ldots, n$, are independent with $E\epsilon_i = 0$ and $Var(\epsilon_i) = \sigma^2$.

OLS estimate coincides with mle when the errors are iid

Regression Methods

- Linear model

$$f(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

  Shrinkage methods, PCR, etc.

- Go beyond linear models: splines, ensemble methods, neural networks

- Issues care about:
  - What's the prediction performance?
  - What is the relationship between Y and predictors X?
  - Is the linear model sufficient? Nonlinear effect? Interactions?
  - Which predictors are more relevant to the prediction?

Classification

- $Y \in \{1, 2, \ldots, K\}$.
- Classifier $G(X) : \mathcal{X} \to \{0, 1\}$.
- Loss: $L(Y, G(X)) = 1_{Y \neq G(X)}$. $L(k, l)$ denotes the price paid for misclassifying an observation belonging to class-k as class-l
- Risk: $R(G) = E[L(Y, G(X))] = P(Y \neq G(X))$.=1-P(Y=G(X))
  max
- Optimal classifier (Bayes classifier):
  P(Y=G(X)|X)=sumP(Y=G(X)=k|X)

$$G^*(X) = \arg\min_G R(G) = \arg\max_k (f_k^*(\mathbf{x})),$$

  =sumP(Y=k|x)1_G(X)=k

  where $f_k^*(\mathbf{x}) = P(Y = k|X = \mathbf{x})$, $k = 1, \ldots, K$.

Generative Methods

Generative methods learn the joint distribution $p(\mathbf{x}, y)$.

- Estimate $p(\mathbf{x}|Y = k)$ (and $Pr(Y = k)$), then use the Bayes rule

$$Pr(Y = k|\mathbf{x}) \propto p(\mathbf{x}|Y = k)Pr(Y = k).$$

- Linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), naive Bayes

Discriminative methods

Discriminative methods learn the conditional distribution $Pr(Y = k|\mathbf{x})$ directly.

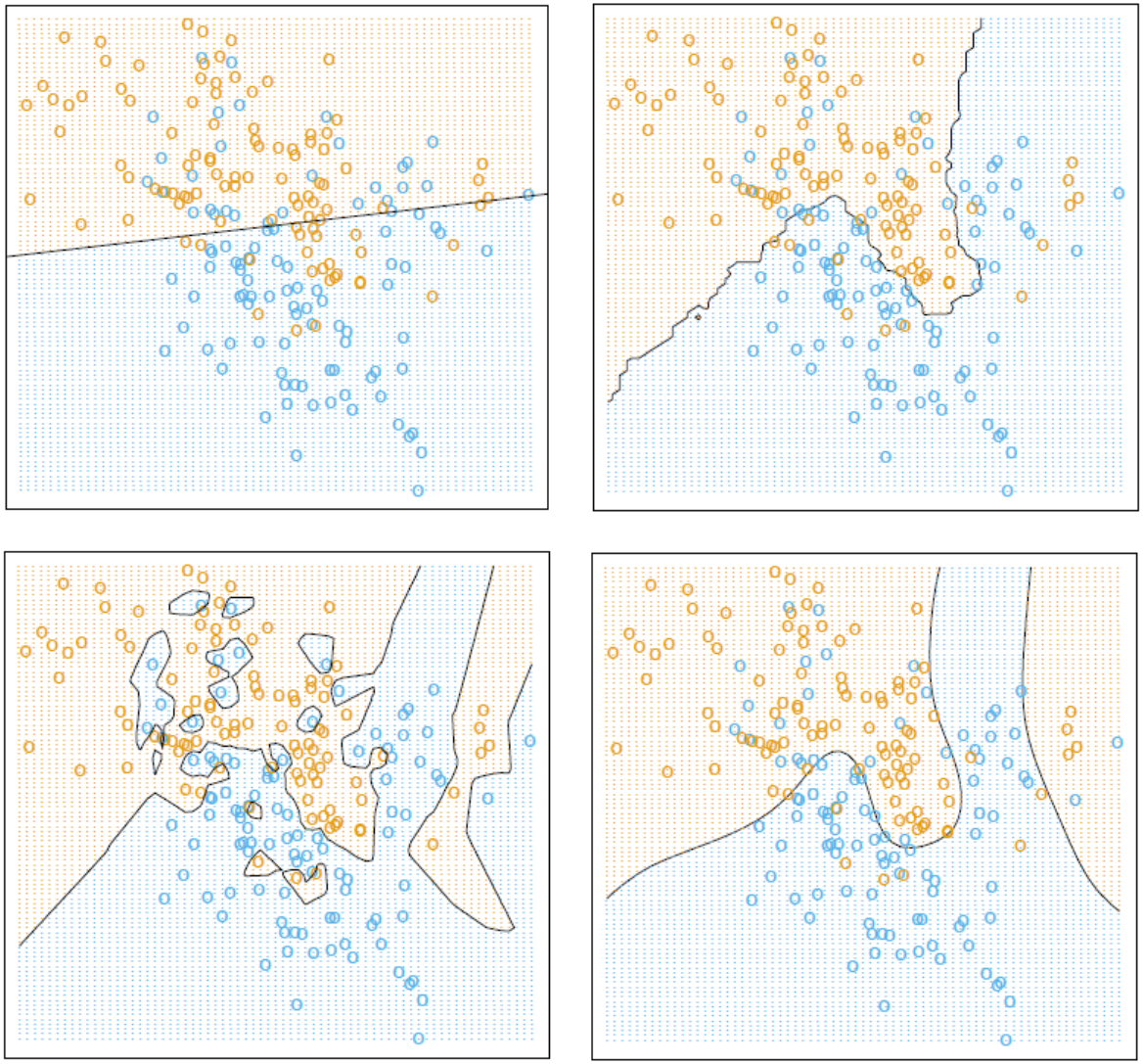
- Logistic regression
- K-nearest neighbor (KNN)
- Classification tree (CART)
- Support vector machines (SVM)
- Ensemble methods: Bagging, Boosting, Random Forest
- Neural networks

Generative vs. Discriminative Methods

For example, what's the big conceptual difference between LDA and Logistic Regression?

- In terms of Bias? Robustness?
- In terms of Variance? Efficiency?
- Why so?

Supervised Learning - Issues



- Overfitting, generalization, variable selection/model selection.

Variable Selection / Model Selection / Inference

Variable selection: Selecting a smaller subset of predictor variables that have the strongest effects.

Model selection: select the model that balance bias and variance.

- Best subset selection, forward-, backward-stepwise selection
- Information criterion: AIC, BIC, Mallow's C_p, ...
- Simultaneous methods: LASSO, SCAD, ...
- Re-sampling methods: cross-validation, bootstrap, ...

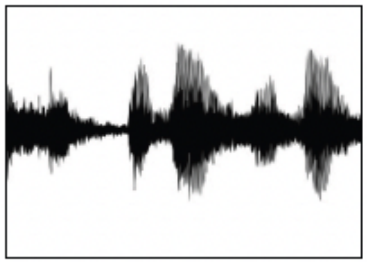
Structured vs Unstructured Data

Structured Data

Size	#bedrooms	...	Price (1000$s)
2104	3		400
1600	3		330
2400	3		369
...	...		...
3000	4		540

User Age	Ad Id	...	Click
41	93242		1
80	93287		0
18	87312		1
...	...		...
27	71244		1

Unstructured Data



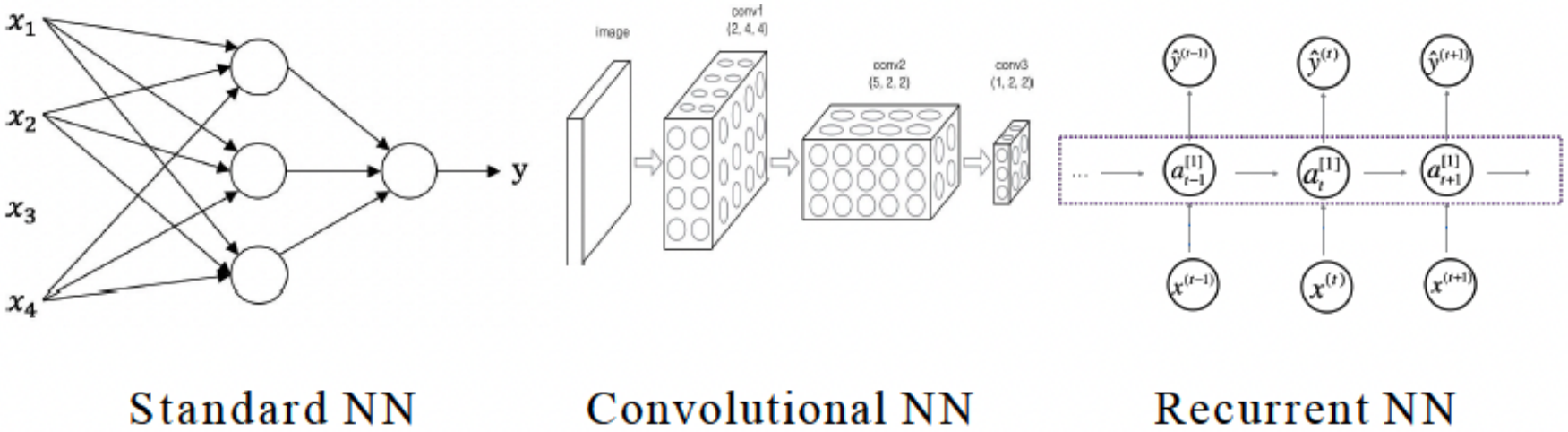
Audio

Image

Four scores and seven years ago...

Text

Deep Learning (Neural Networks)



Standard NN Convolutional NN Recurrent NN

Unsupervised Learning

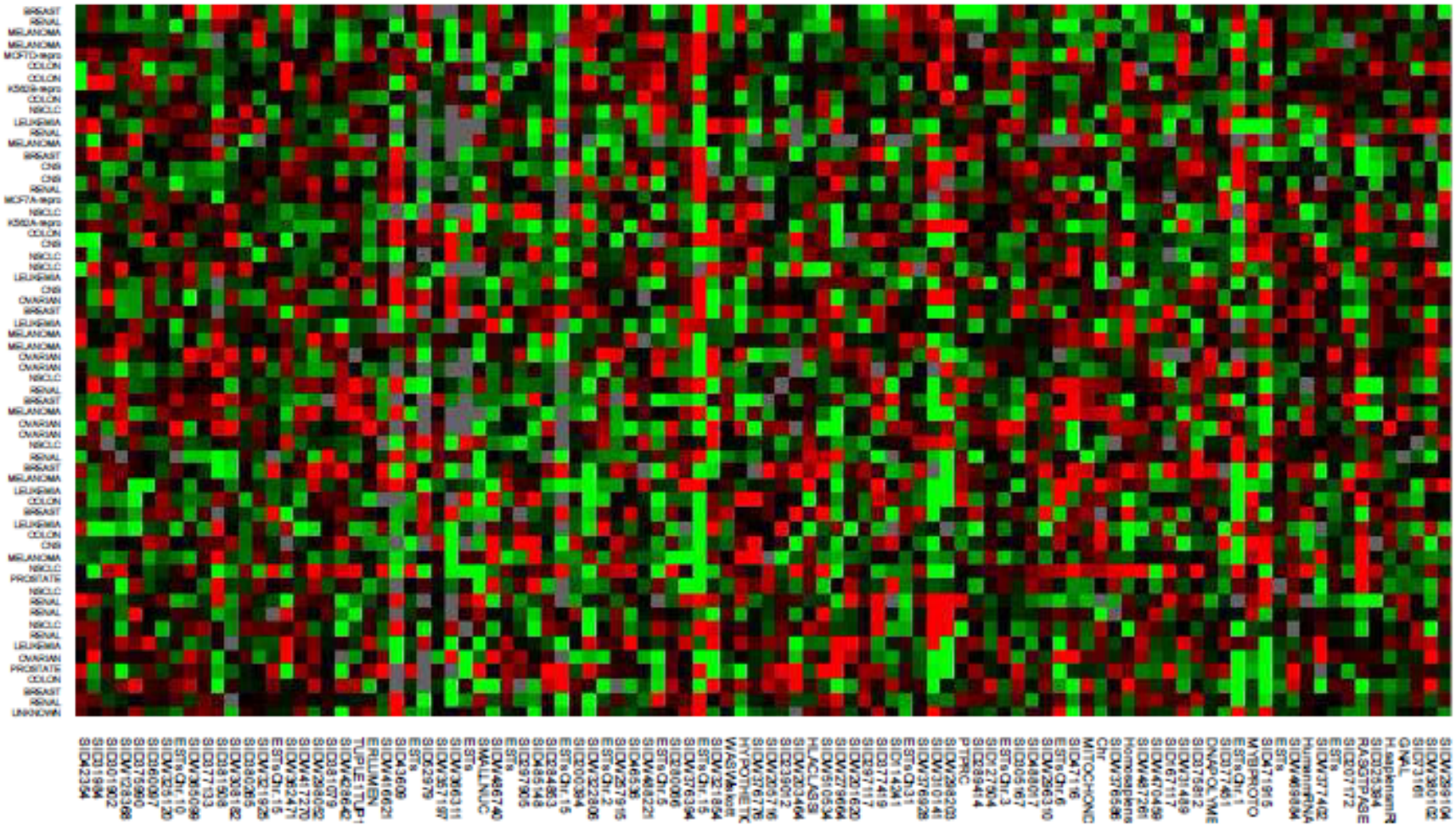
We observe $\mathbf{X} = (X_1, \ldots X_p)$ but no Y.

- p is often much larger than that in supervised learning.
- Goal: Infer the properties of $Pr(\mathbf{X})$. Properties of interest are often more complicated than simple location parameters.
- Difficult to measure quality of the results.

DNA Expression Microarrays

Training data: Gene expression data from 64 cancer tumors across 6830 genes.

- Input X: the level of expression for each gene



DNA Expression Microarrays

- Training data: Gene expression data from 64 cancer tumors across 6830 genes.
  - Input X: the level of expression for each gene

- Goal: understand how the genes and samples are organized.
  - Which samples are most similar to each other, in terms of their expression profiles across genes?
  - Which genes are most similar to each other, in terms of their expression profiles across samples?
  - Do certain genes show very high (or low) expression for certain cancer samples?

- This task can be formulated as
  - supervised learning: classification problem
  - unsupervised learning: Cluster Analysis

Cluster Analysis no clear rule; similarity

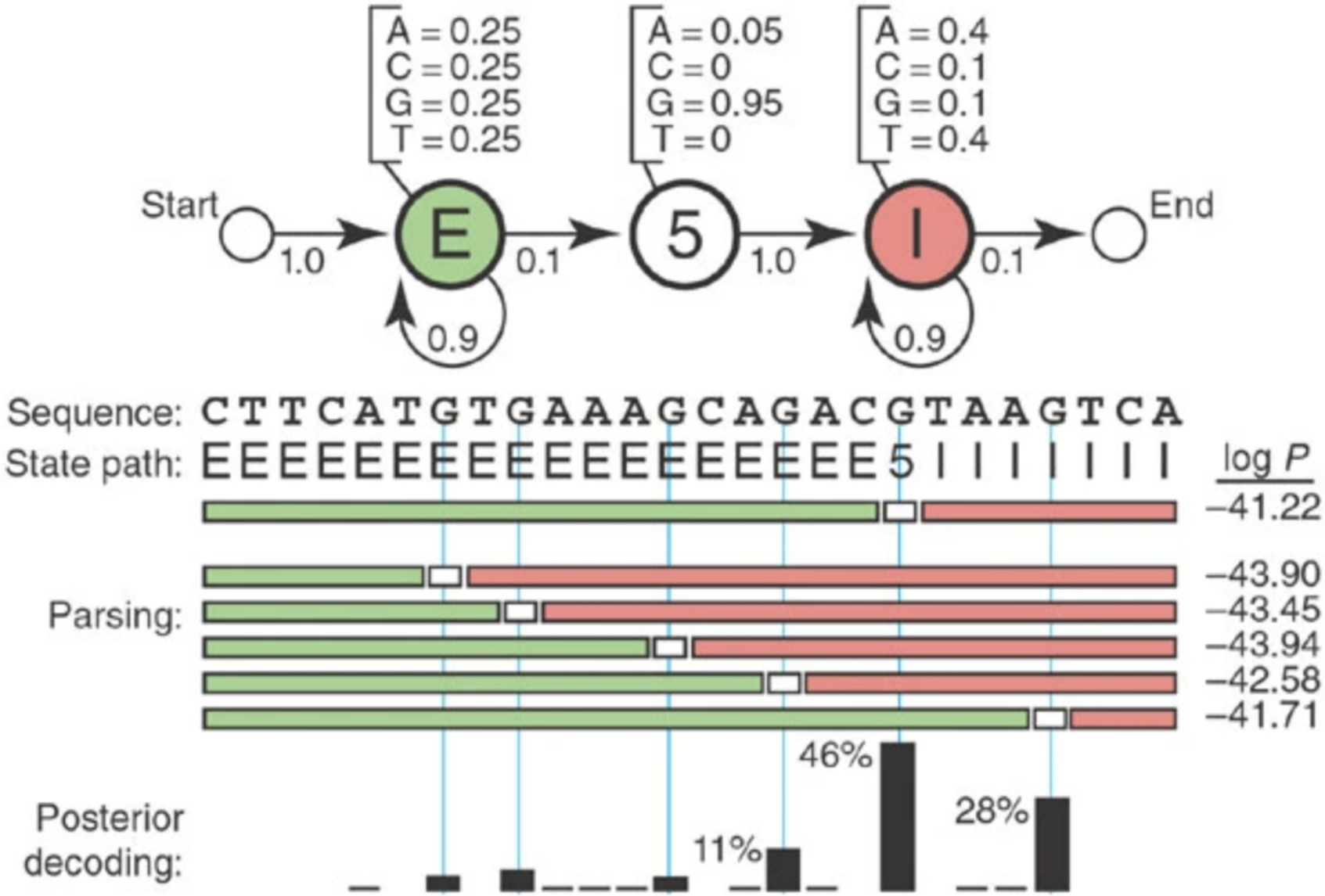
Goal: Grouping objects into clusters such that those within each cluster are more closely related to one another than those in different clusters.

Market Segmentation Example:

- Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix
- Collect different variables of customers based on their geographical and lifestyle related information
- Find clusters of similar customers

How many clusters? What kinds of clusters? What objective function?

Hidden Markov Models



Association Rules

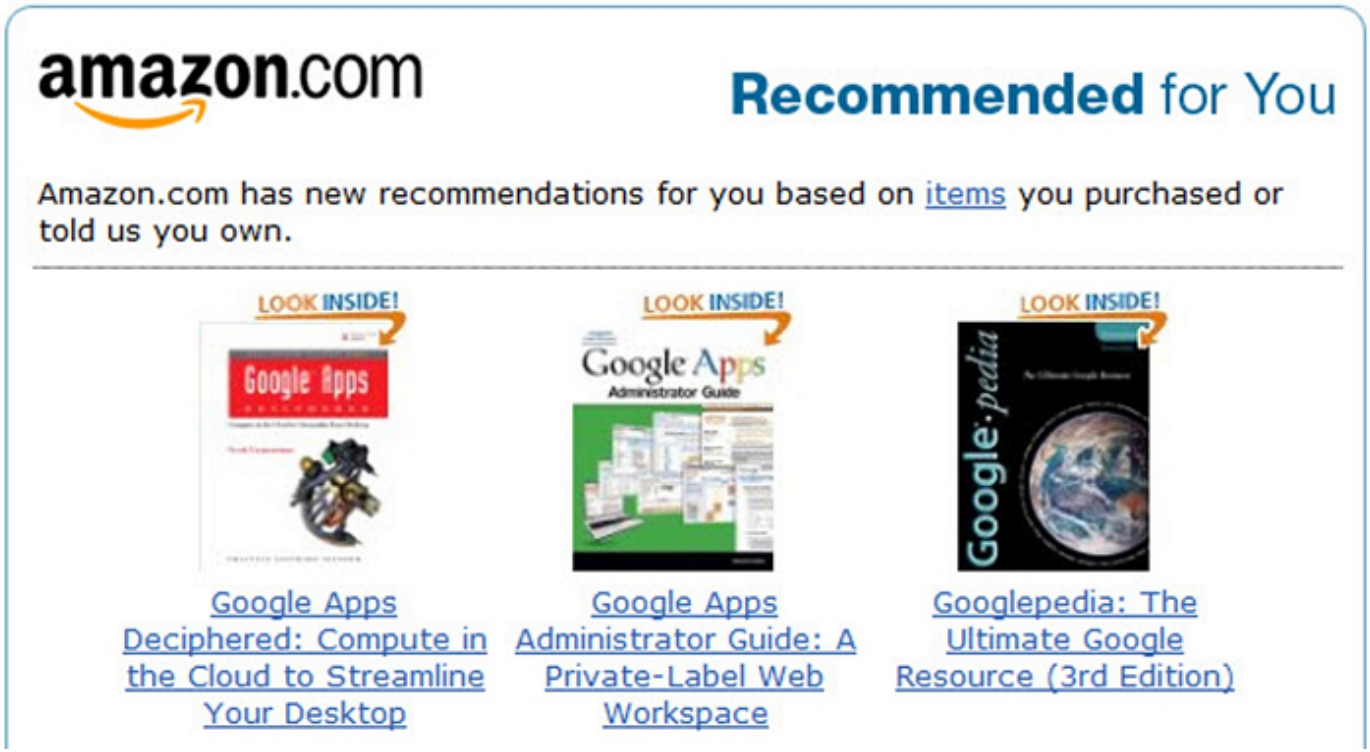
Training data: a set of records each of which contains some number of items from a given collection.

ID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Bread, Coke, Milk
4	Bread, Coke, Diaper, Milk
5	Coke, Diaper, Milk
⋮	⋮

- Goal: Find collections of items that occur together with high probability. (Produce dependency rules which will predict occurrence of an item based on occurrences of other items.)
  e.g., Rules discovered: {Coke} ⇒ {Milk}, {Beer} ⇒{Bread}

- focus on discovering interesting local patterns in the data rather than to characterize the data globally

Recommender Systems

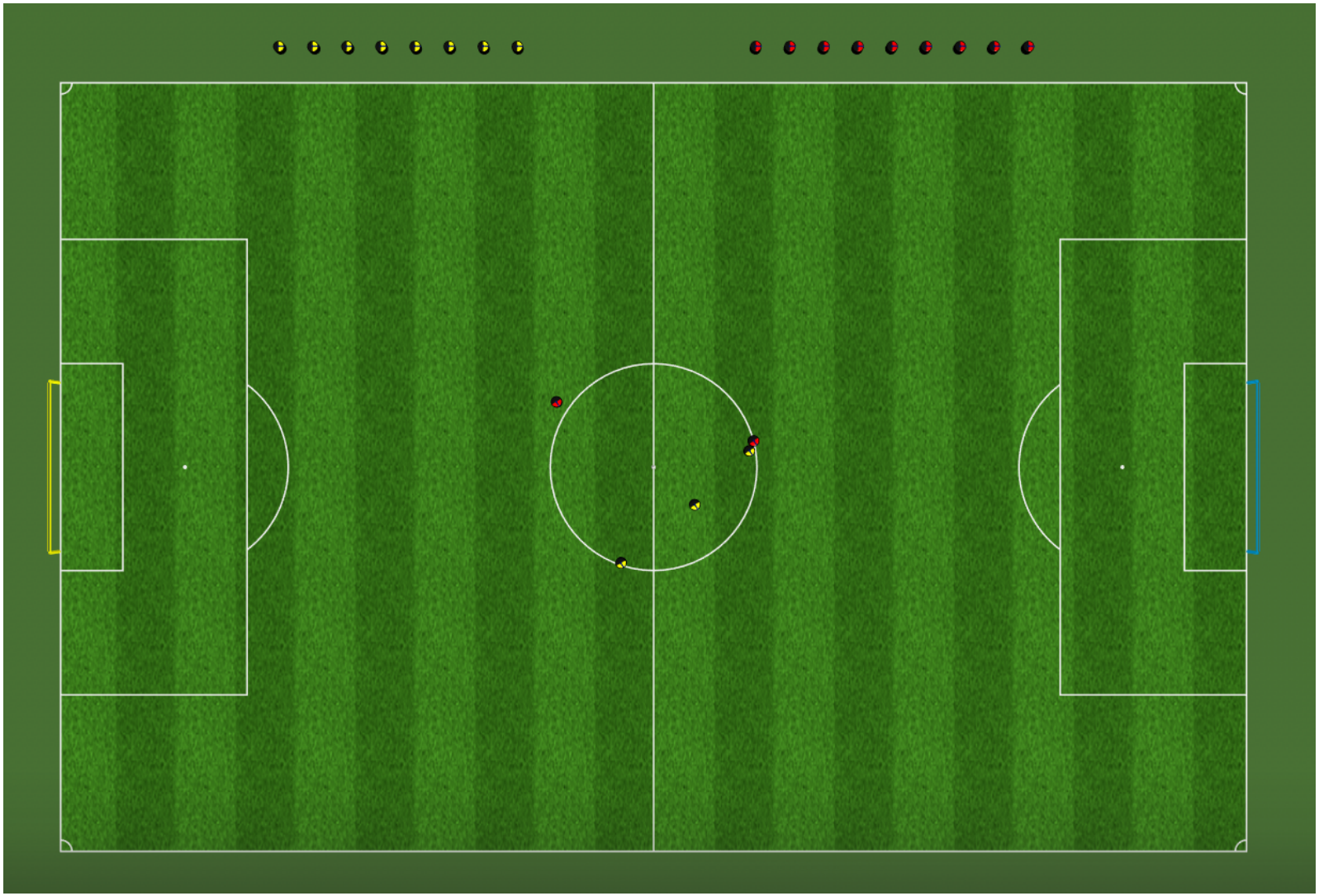


- Collaborative filtering
- Content-based filtering

Reinforcement Learning

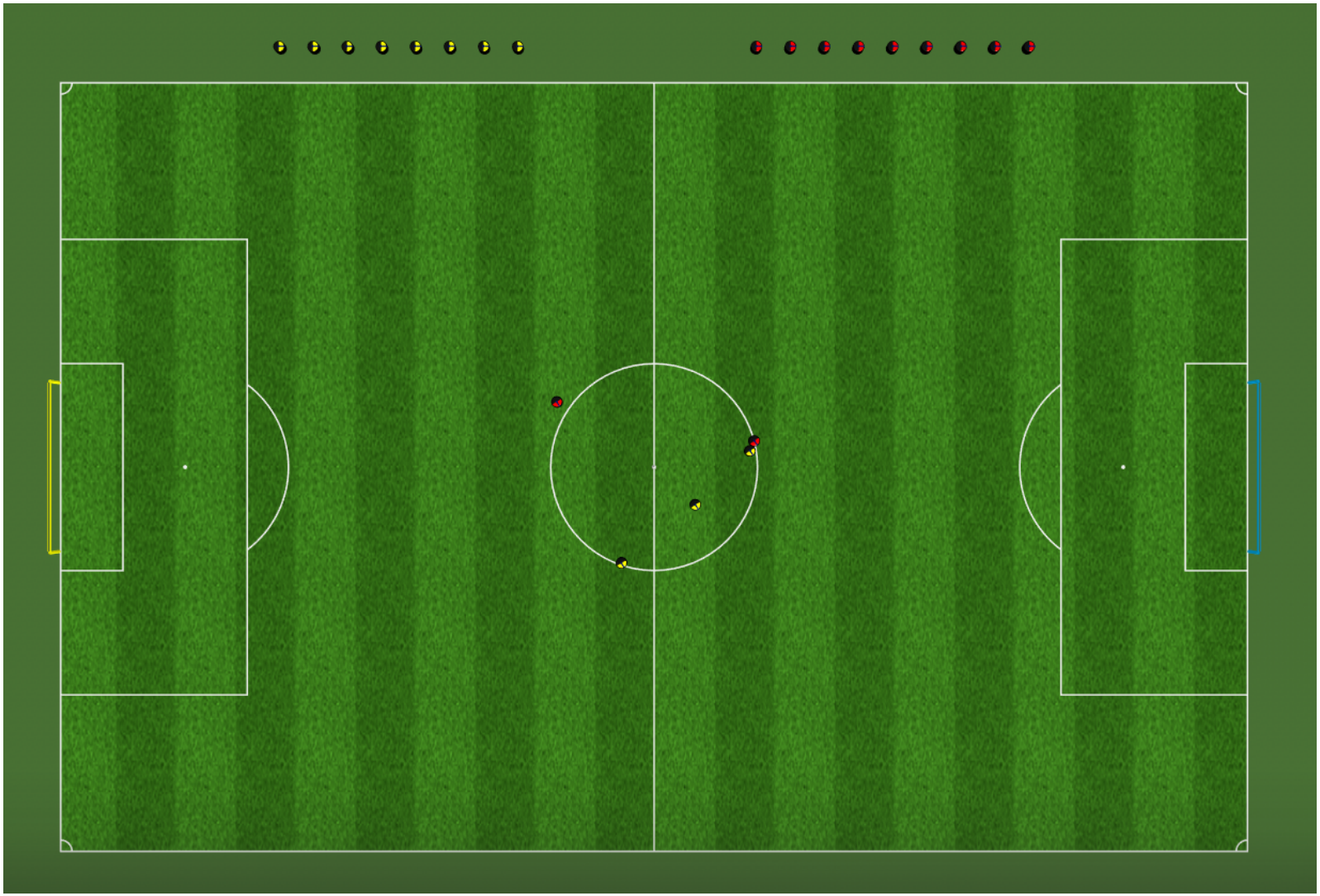
- Three elements: (state, action, reward) over time.
- Goal: Develop a policy: state space $\rightarrow$ action space so as to maximize the long-term return.
- Falls between the supervised and unsupervised learning: you have the input, but not output, instead you have "critic".

Keepaway Example: Random action



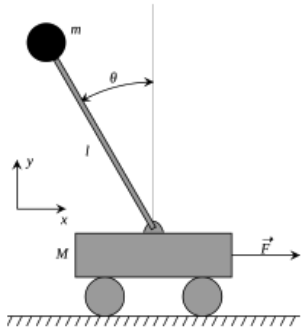
Keepers hold the ball for an average of 5.5 seconds.

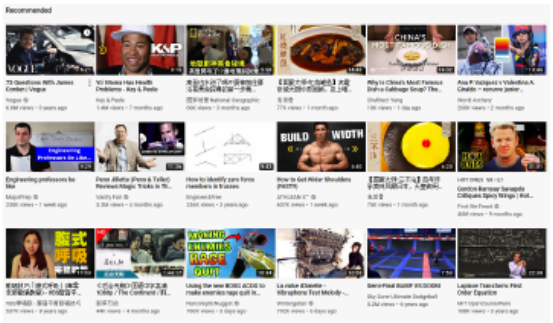
Keepaway Soccer Example: Learned action

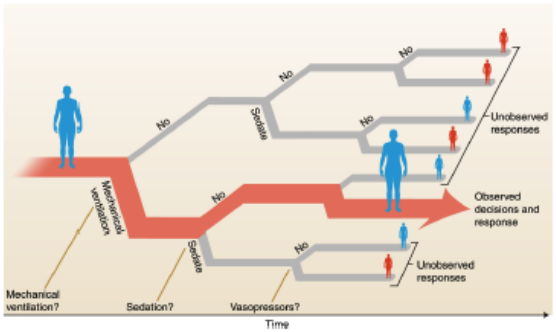


After learning, keepers could hold the ball for about 12 seconds on average.

RL Applications







A Proposal from Prof. Jeff Wu in 1997

Statistics = Data Science ?

C. F. Jeff Wu

University of Michigan, Ann Arbor

A proposal:

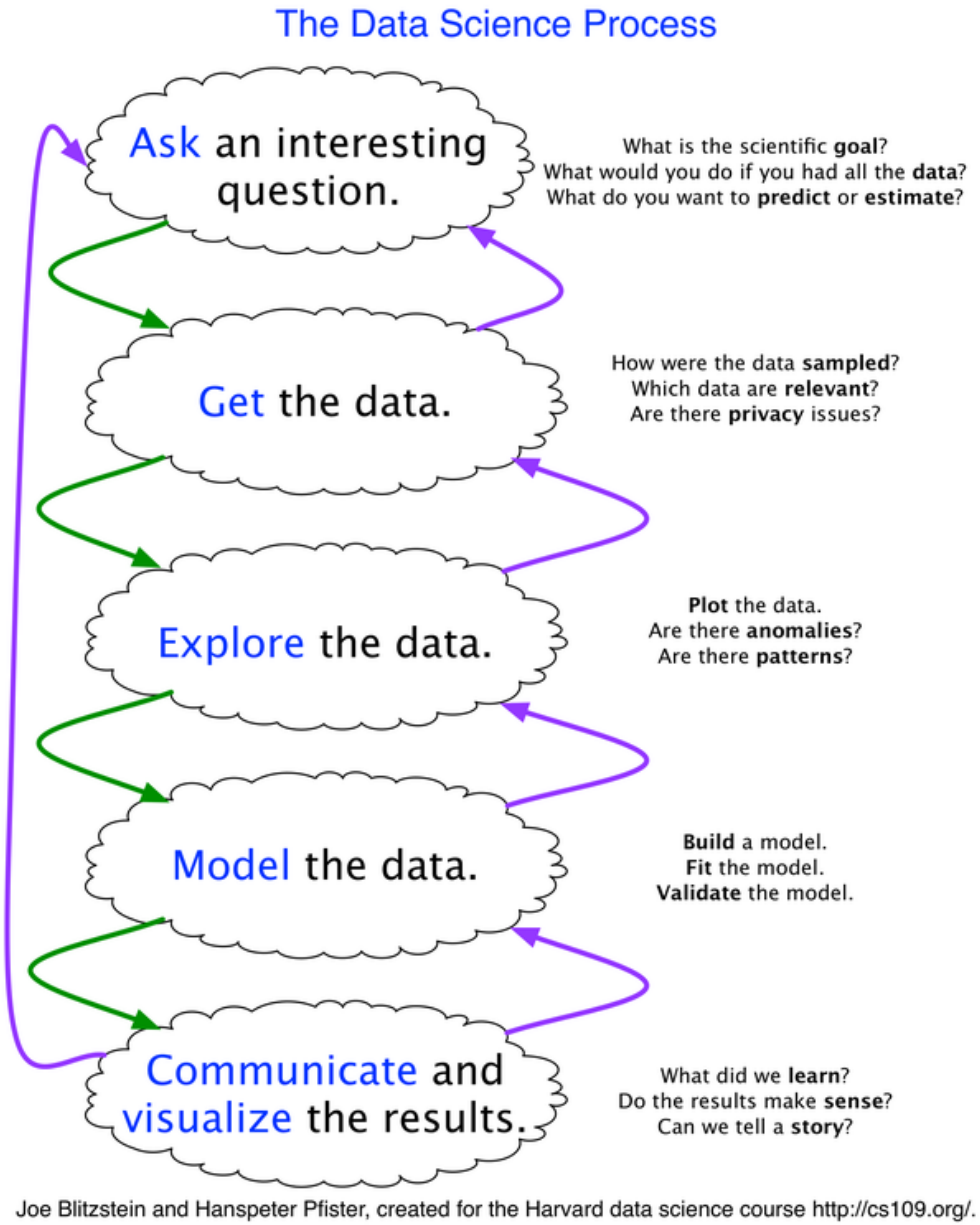
"Statistics" $\longrightarrow$ "Data Science"

"Statisticians" $\longrightarrow$ "Data Scientists"

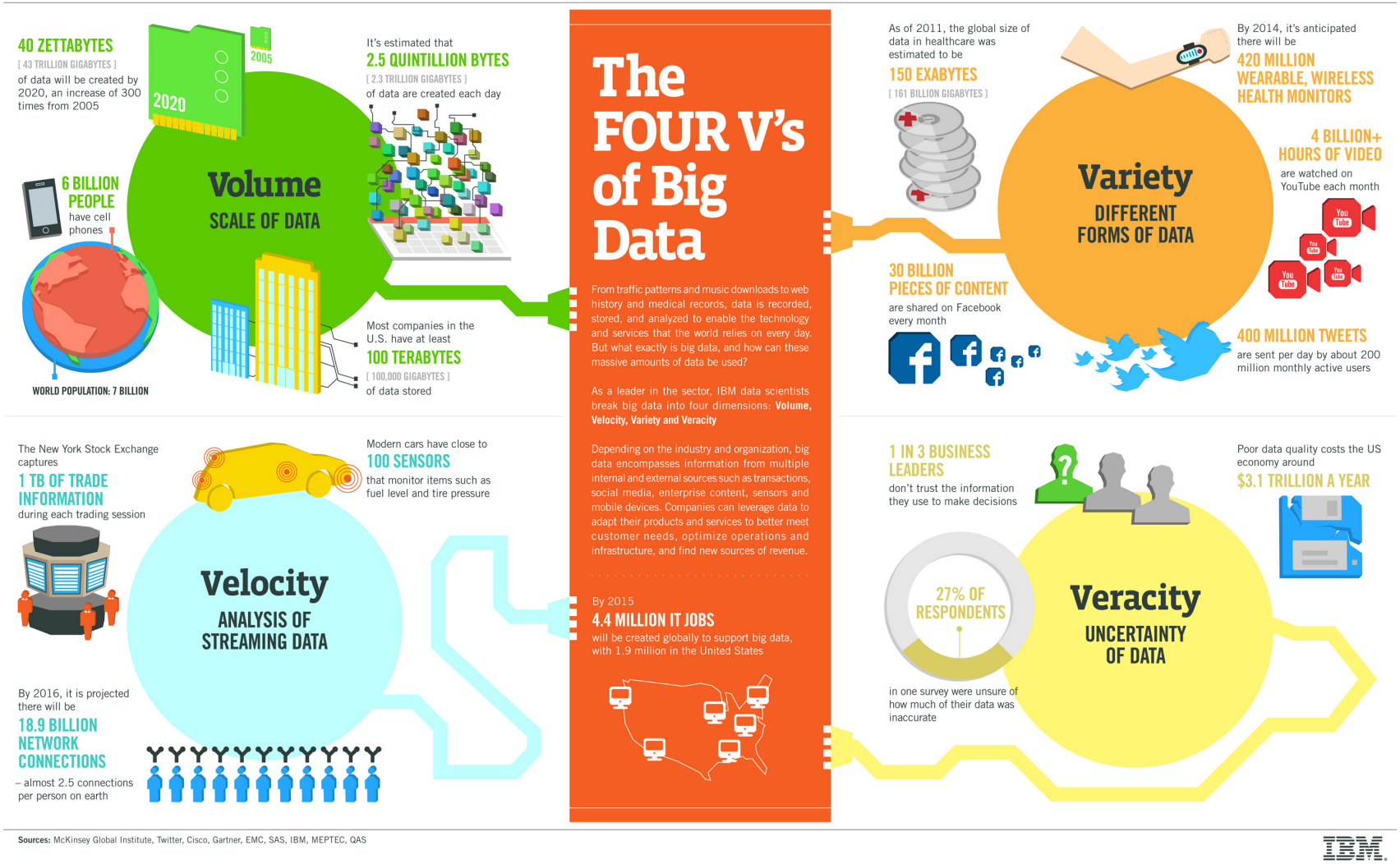
- What is "Statistics"?
- A Statistical Trilogy
- Frontier and Beyond
- A Bold Proposal

- Several good names have been taken up: computer science, information science, material science, cognitive science
- "Data Science" is likely the remaining good name reserved for us
- "Statistical Science" not as attractive, but much better than "Statistics"

Data Science



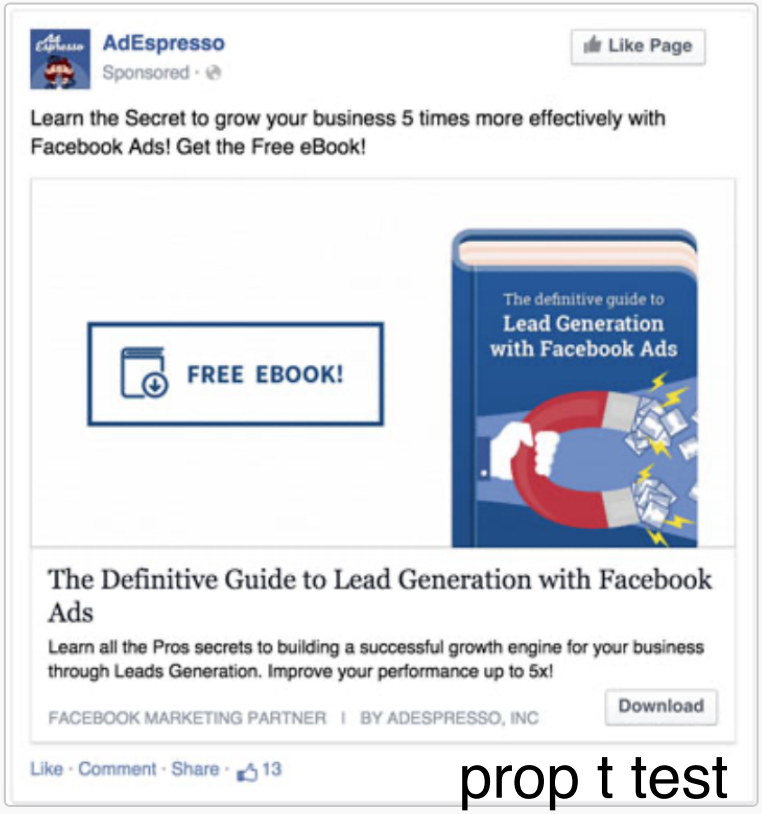
Four (+1) V's of Big Data (IBM version)



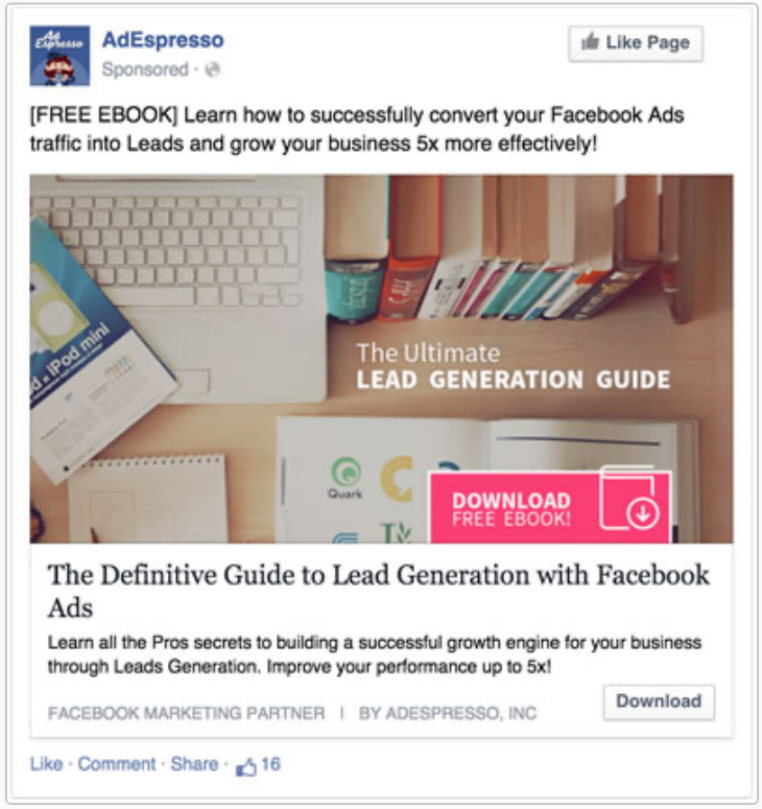
"Value": the ability of turning our data into value

How Scale is Enabling Deep Learning (Andrew Ng)

Goal: Solve the Problem! Sometimes simple is better ...



prop t test



10,000 Impression	10,000 Impression
237 Clicks *(CTR: 2.37%)*	187 Clicks *(CTR: 1.87%)*
28 Sales *(Conversion rate: 11.81%)*	16 Sales *(Conversion rate: 8.55%)*
Spent: $150	Spent: $150
Cost per Sale: $5.35	Cost per Sale: $9.37 (+75.14%)

Summary

- Standard (Supervised) Machine Learning (4 lectures).
  - Comprehensive understanding of the methods.
- Deep learning (3 lectures).
  - Understand the structures and optimization algorithms.
  - Implementation in Python.
- Reinforcement Learning (3 lectures).
  - Intuitive understanding.
  - Implementation of simple algorithms (in python or R).
- Topics in Unsupervised Learning (2 lectures).
  - Intuitive understanding.
  - Implementation of some methods.

The best way to understand the methods is to try them out!