

---

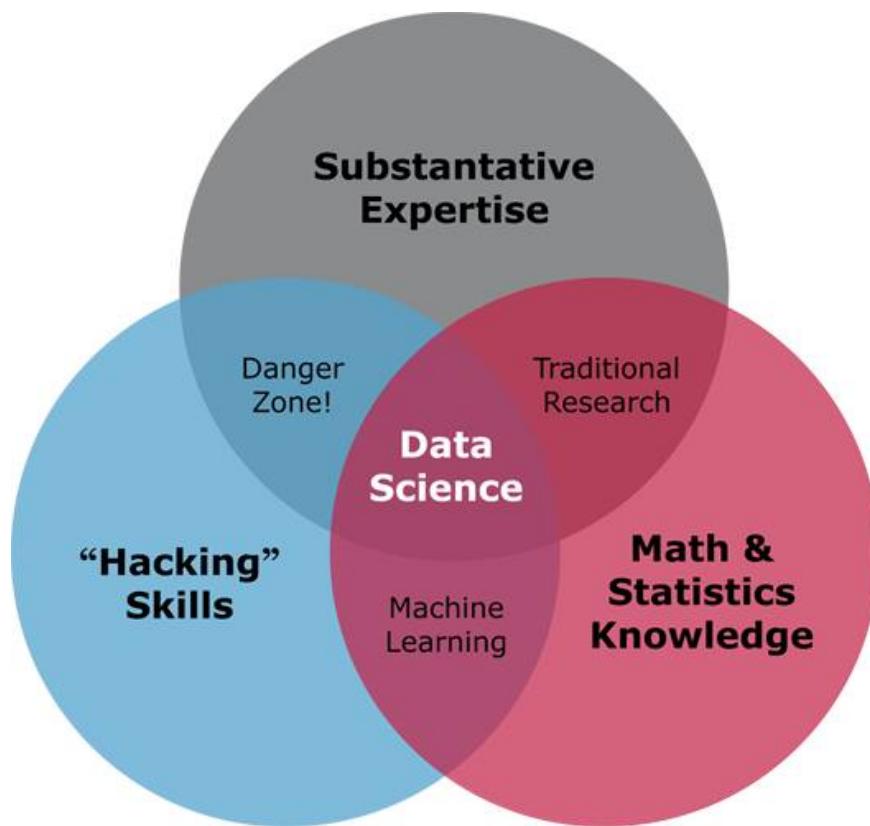
# MIS 382N: Advanced Predictive Modeling

Prof. Joydeep Ghosh  
ECE/UT

[www.ideal.ece.utexas.edu/~ghosh](http://www.ideal.ece.utexas.edu/~ghosh)  
[jghosh@utexas.edu](mailto:jghosh@utexas.edu)

# Data “Science”

---



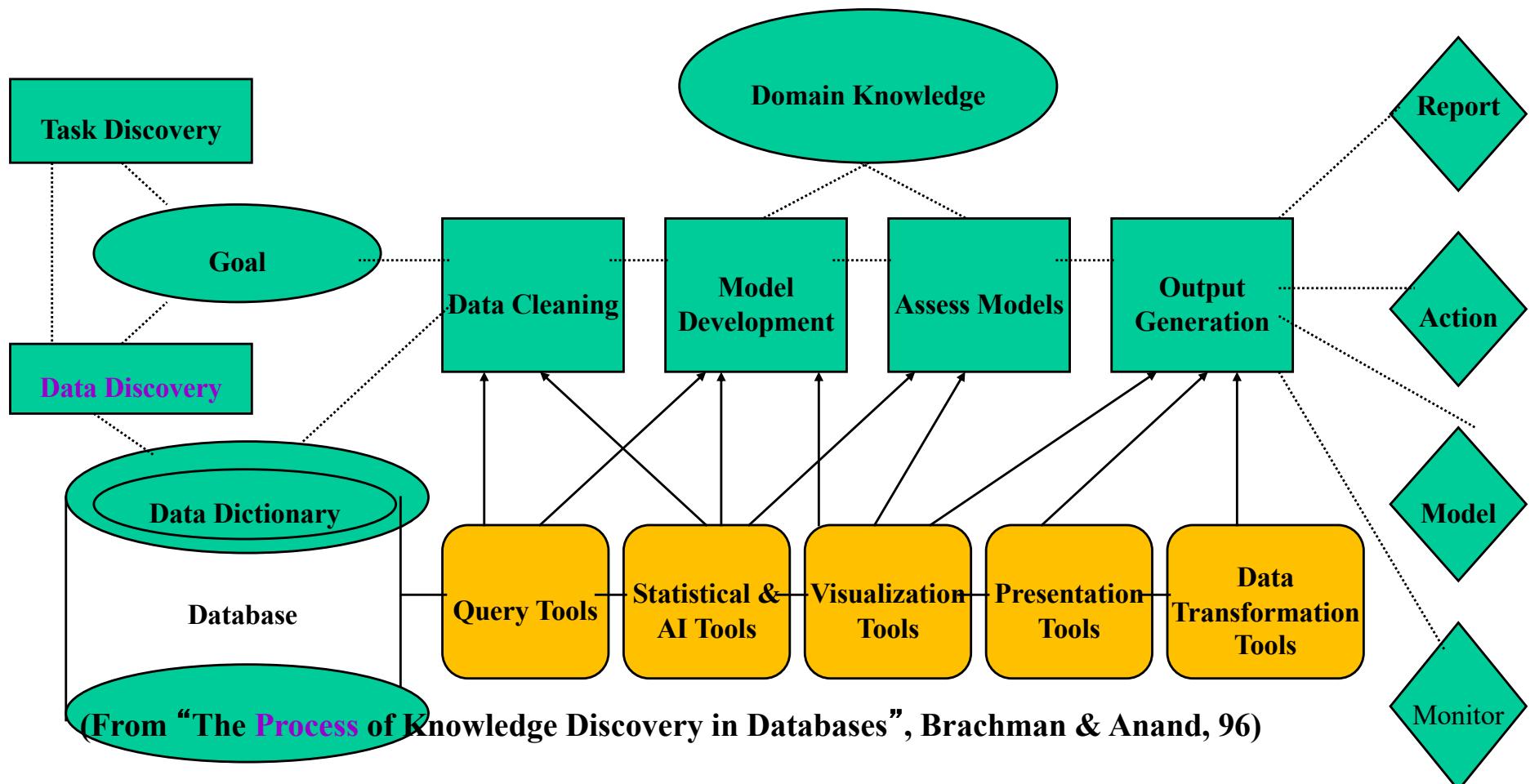
- *Business Problem → Data Science sub-problems*
- *Additional AI modalities*
- *Enterprise Delivery Platform*
- *U/I and U/X*

<https://cyborgus.com/2017/03/13/think-like-data-scientist/>



# (Inter-disciplinary) Process of Data Mining

*Collect      Prepare      Analyze      Consume*



**Iterative Process at multiple levels**

# Data Driven Modeling Approaches and Goals

---

- **Seek (aggregate) models**
  - (quick) large scale summary of data
    - E.g. characterize dominant customer types
- **Seek (local) patterns**
  - Characterise a small portion of data
    - e.g. “rare patterns”: fraud or intrusion detection
- Goals:
  - **Description:** Find human-interpretable patterns that describe the data.
  - **Prediction:** Use some variables to predict unknown or future values of other variables.
  - **Prescription:** (tough!!)

# Backup Texts

---

- **Main (only a few chapters; provided for you).**
  - **B:** Bishop, Pattern Recognition and Machine Learning (more mathematical, Bayesian)<http://www.rmki.kfki.hu/~banmi/elite/Bishop%20-%20Pattern%20Recognition%20and%20Machine%20Learning.pdf>
- **Basic**
  - **KJ:** Kjell and Johnson. Applied Predictive Modeling, Springer 2013.
  - <http://appliedpredictivemodeling.com/>
  - **JW: ISLR:** elementary stats with R
  - <http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf>
- **Advanced: HTF:** Hastie/Tibshirani/Friedman (stats)  
<http://www-stat.stanford.edu/~tibs/ElemStatLearn/>

# Languages and Software

---

- Stats oriented: R, Python (with packages)
  - Commercial: SAS, IBM SPSS,..
  - Open: GUI oriented: Knime, RapidMiner
- General purpose (Java for text analysis)
- Distributed/bigdata
  - Hadoop/Spark/MapReduce/PigLatin
  - HIVE (SQL like for Hadoop)
  - Various NoSQL
- **New (2018): AUTO-ML (DataRobot, H2O,...)**

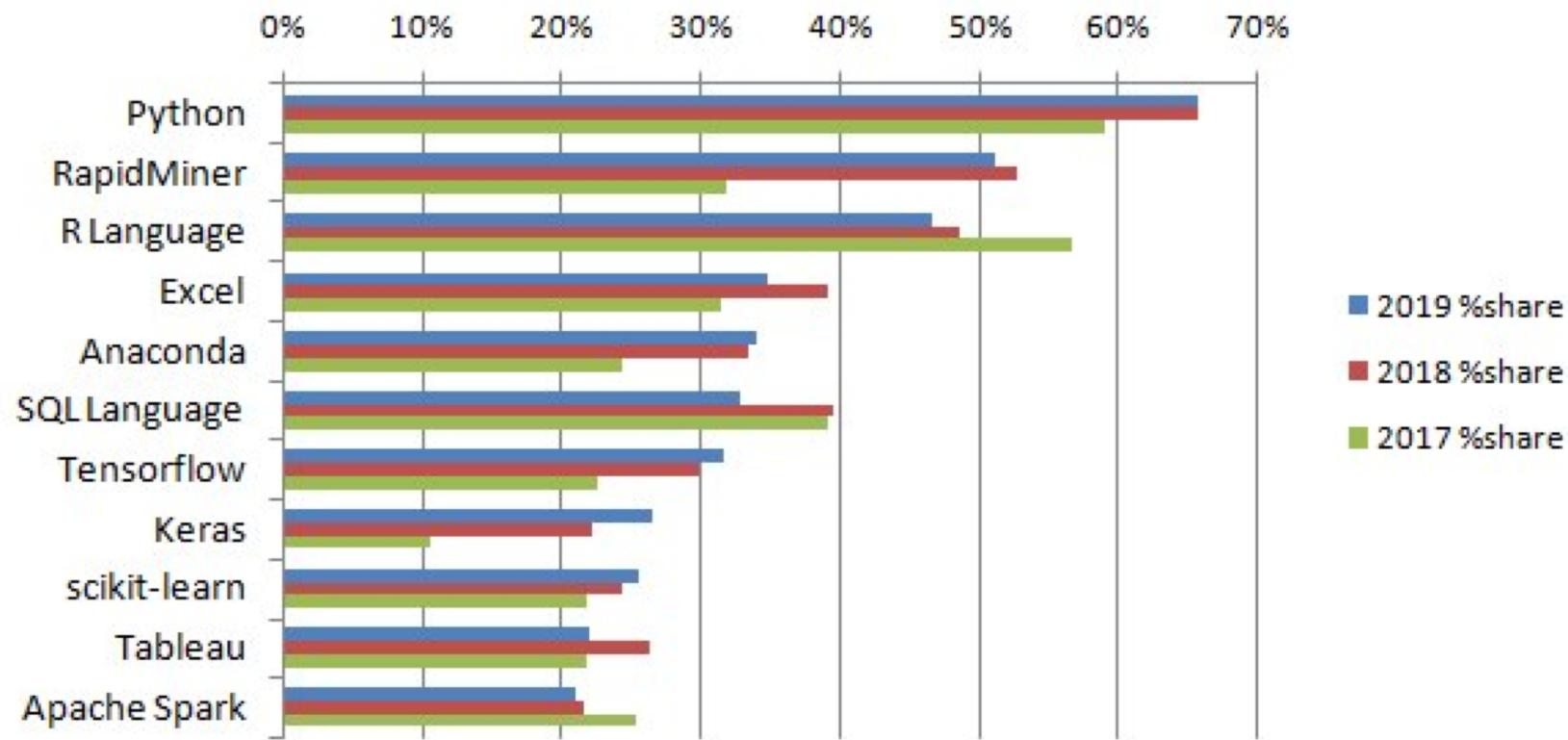
**See: How Did Python Become A Data Science Powerhouse?**

<https://www.youtube.com/watch?v=9by46AAqz70>

<https://www.datanami.com/2019/08/15/is-python-strangling-r-to-death/>

# KDD Nuggets Survey May 2019

## Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll

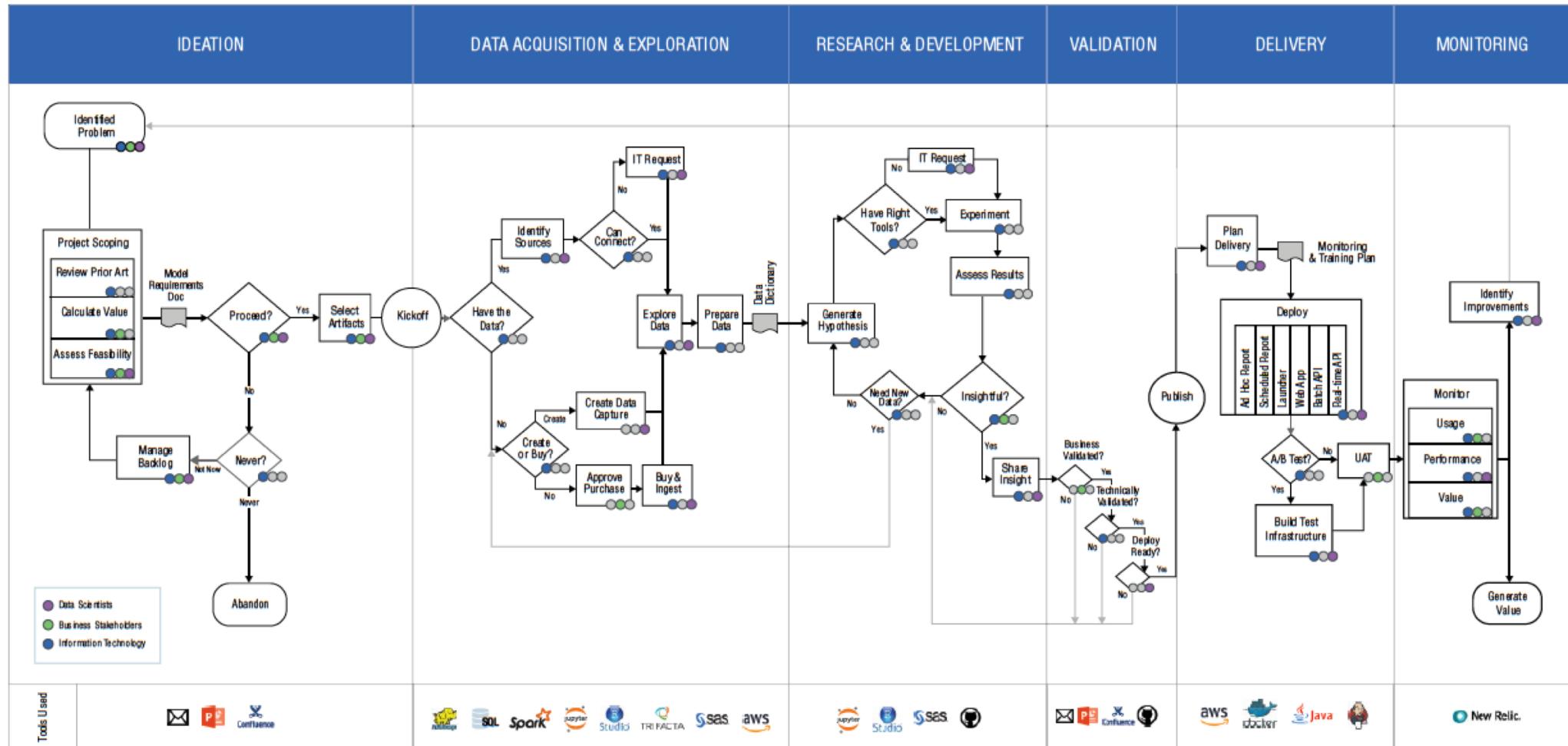


## *Trends*

---

- Integrating with Software Environment
  - Model lifecycle management
  - Kubernetes
  - Cloud Based Platforms (vs. On-prem vs. Edge)
- Integrating with Business KPIs, and with other Decision Making Systems
- Human in the loop
- FAT-ML
- AutoML
  - <https://www.topbots.com/automl-solutions-overview/>

# DATA SCIENCE LIFECYCLE

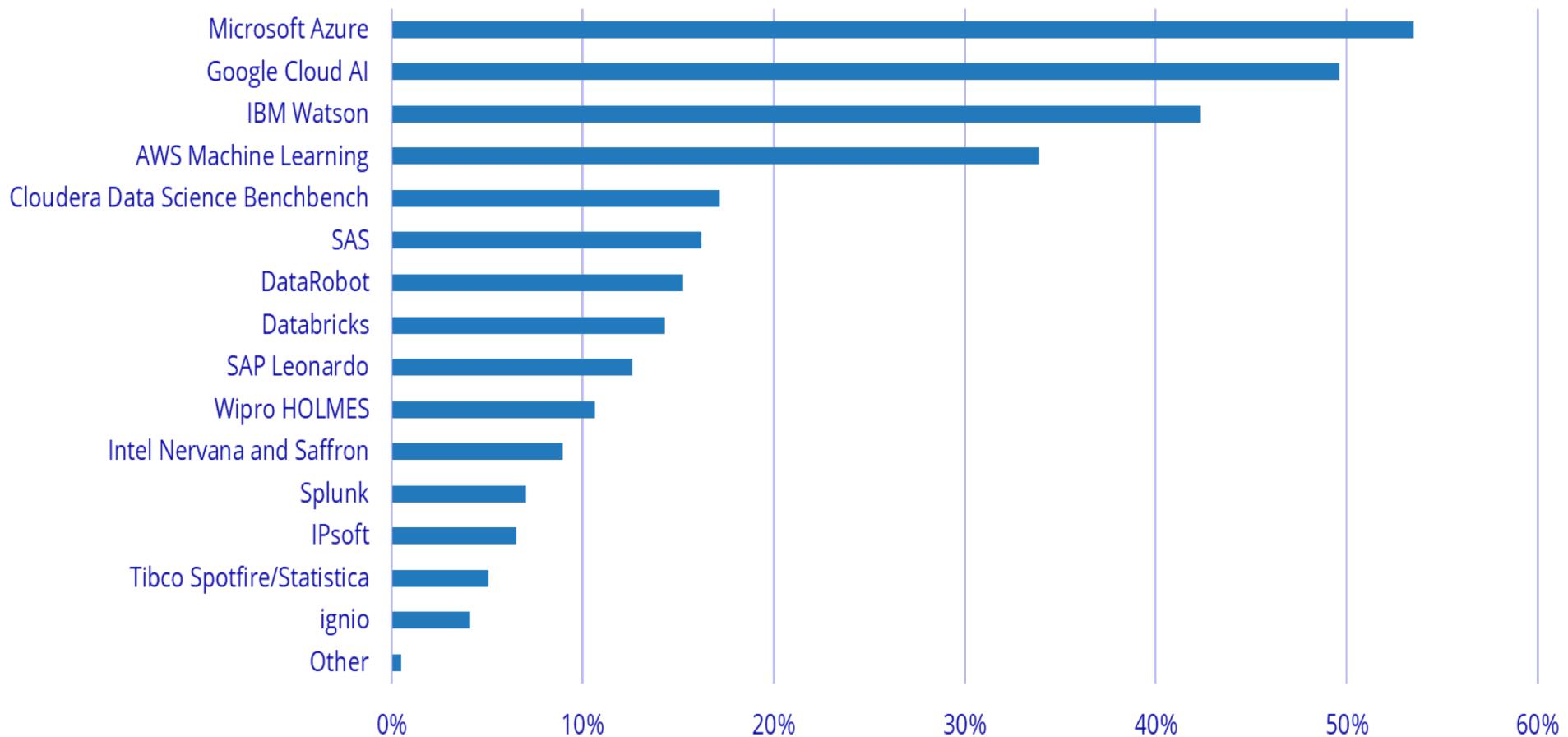


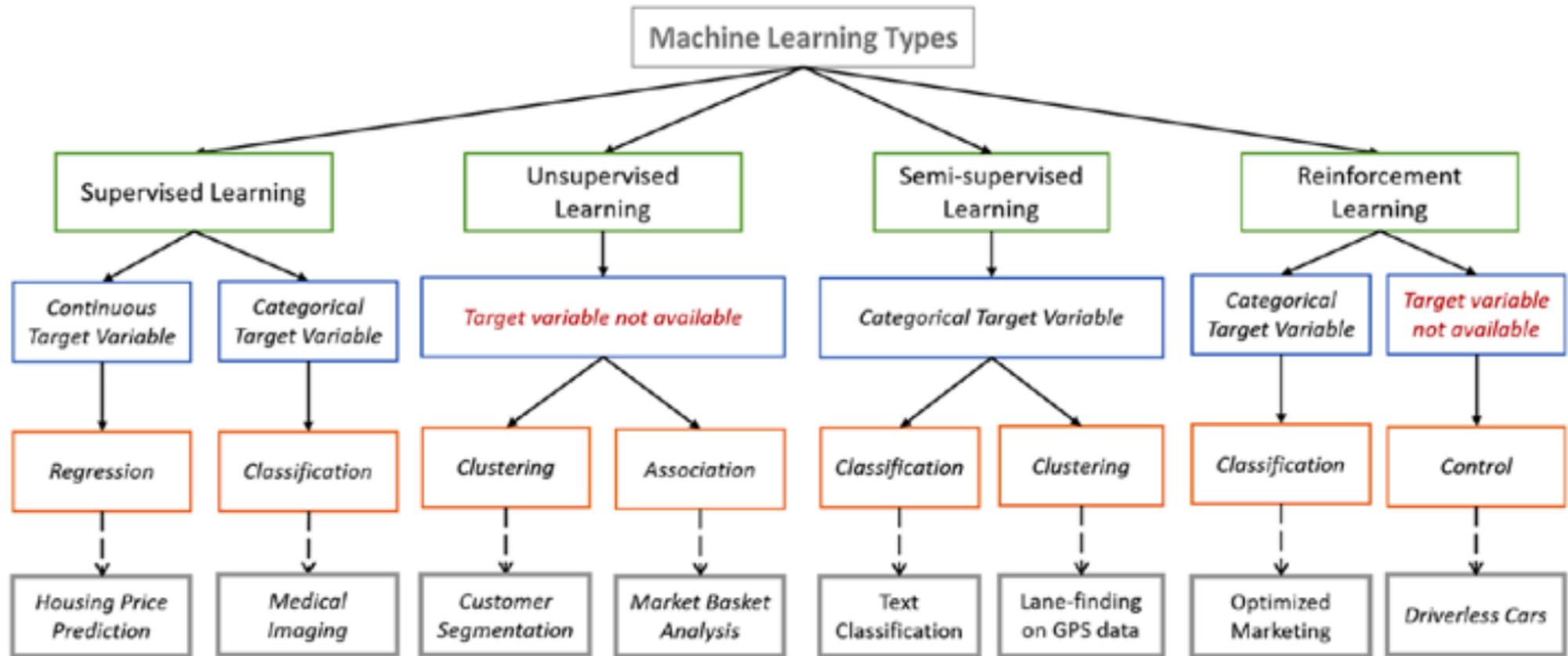
Read the “Domino” article before/while doing your project

*Which of the following general-purpose AI platforms are you actively working with?*

---

- (from a Feb 2019 Survey)





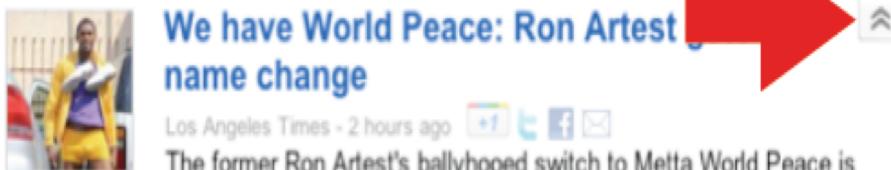
*Cold Start Problem: Mismatch → Online learning methods*

# 50 shades of supervision

- Ads



- Click feedback



- Emails



- Tags



- Ads on website
- Spam
- Product recommendations
  - Tinder
- Domain expert feedback on clusters produced

# No Free Lunch (NFL)

- No universally best model; so understand tradeoffs.
- Table from HTF

**TABLE 10.1.** Some characteristics of different learning methods. Key: ▲ = good, ◇ = fair, and ▼ = poor.

Characteristic	Neural Nets	SVM	Trees	MARS	k-NN, Kernels
Natural handling of data of “mixed” type	▼	▼	▲	▲	▼
Handling of missing values	▼	▼	▲	▲	▲
Robustness to outliers in input space	▼	▼	▲	▼	▲
Insensitive to monotone transformations of inputs	▼	▼	▲	▼	▼
Computational scalability (large $N$ )	▼	▼	▲	▲	▼
Ability to deal with irrelevant inputs	▼	▼	▲	▲	▼
Ability to extract linear combinations of features	▲	▲	▼	▼	◇
Interpretability	▼	▼	◇	▲	▼
Predictive power	▲	▲	▼	◇	▲

# It Depends

---

“all models are wrong, but some  
are useful”

- George Box, 1987

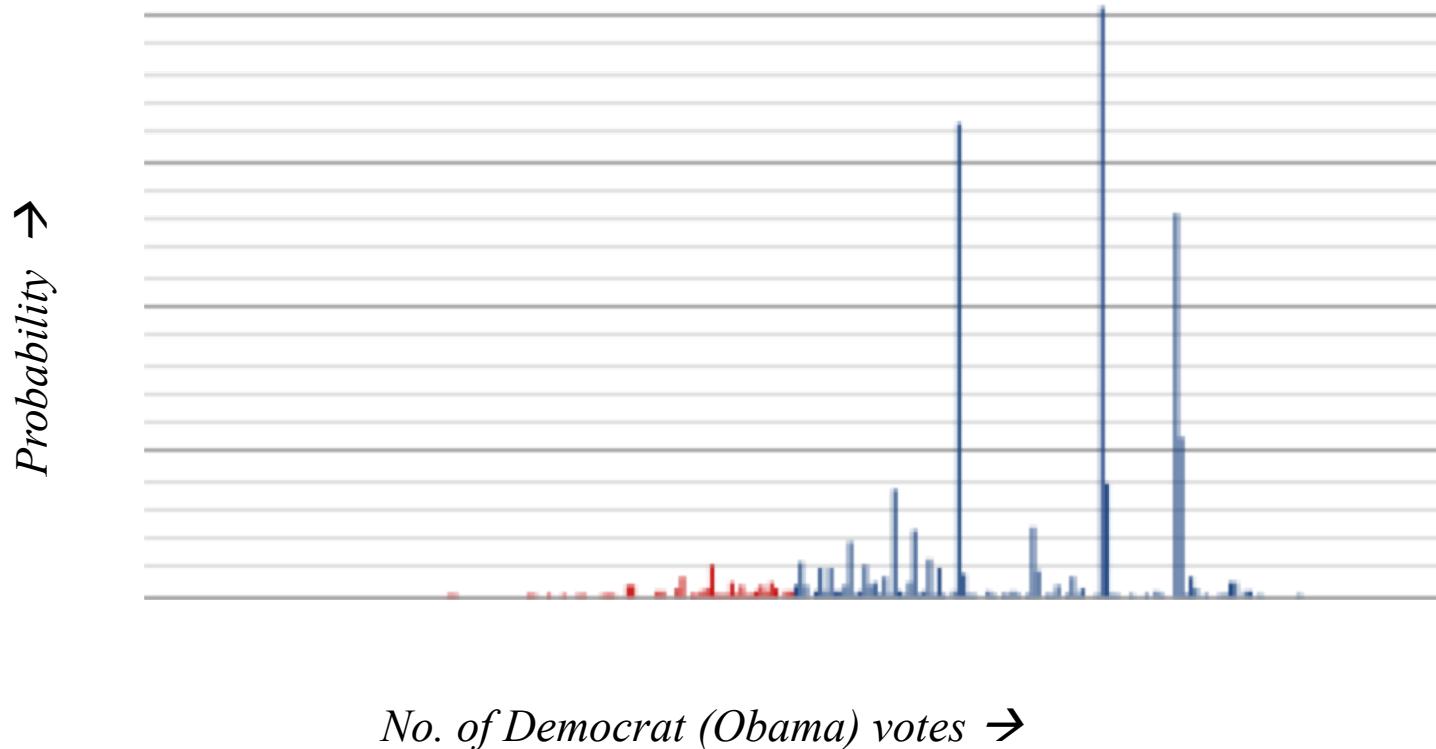
- All statistical models make assumptions
  - (Lets pretend...)
  - Given the situations, some assumptions are plausible, others are not

Visualize: <http://setosa.io/ev/ordinary-least-squares-regression/>

# Nate Silver's 2012 model

---

- Uncertainty is everywhere
  - even in the “perfect” model
  - (used weighted ensemble)



# Course Goals

---

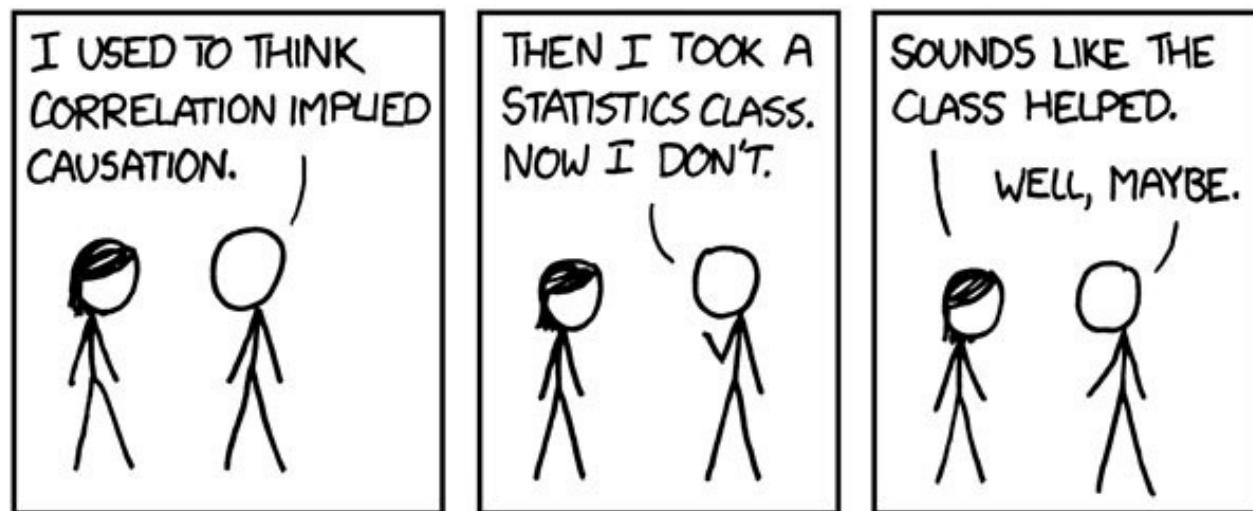
- study different predictive models for a given task
  - Properties, pros and cons
  - Evaluation metrics
    - Business relevance
  - Build predictive models in Python
- Process-oriented viewpoint
- Introduction to issues of scale and real data considerations

# Broader Goals

---

Reason about data analysis and the “results” obtained

- “Mr. Trump won 76% of Cracker Barrel counties and 22% of Whole Food counties”
  - Economist, Jan 28<sup>th</sup>, 2017, quoting the Cook Political Report
  - (compares 54% gap vs. 31 for Bush 2000 and 43 for Obama 2008)



XKCD

# Probability Recap

---

- Basic Concepts:

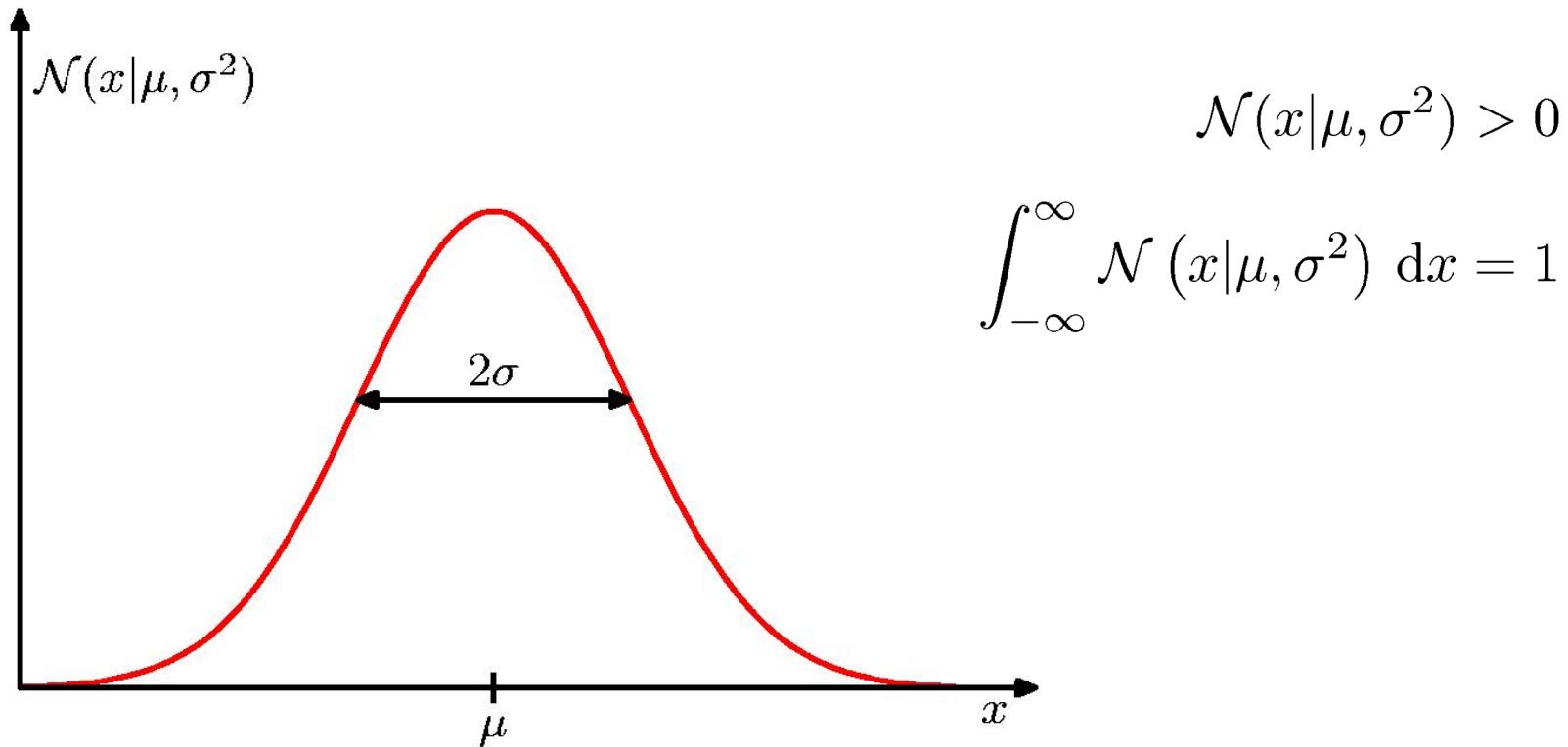
- Joint distribution
- Marginal distribution
- Conditional distribution
- Variance and covariance

Visualize: <http://setosa.io/ev/conditional-probability/>

# The Gaussian Distribution

---

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$



# Gaussian Mean and Variance

---

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x \, dx = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 \, dx = \mu^2 + \sigma^2$$

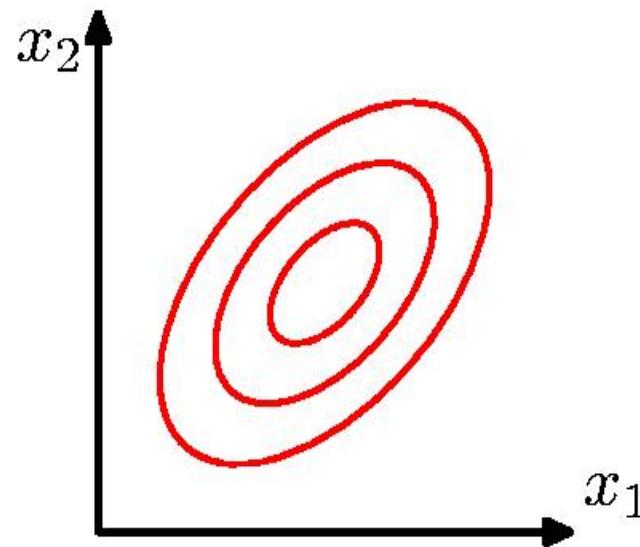
$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

*Denotes the “expectation” operator*

# The Multivariate Gaussian (in D dimensions)

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

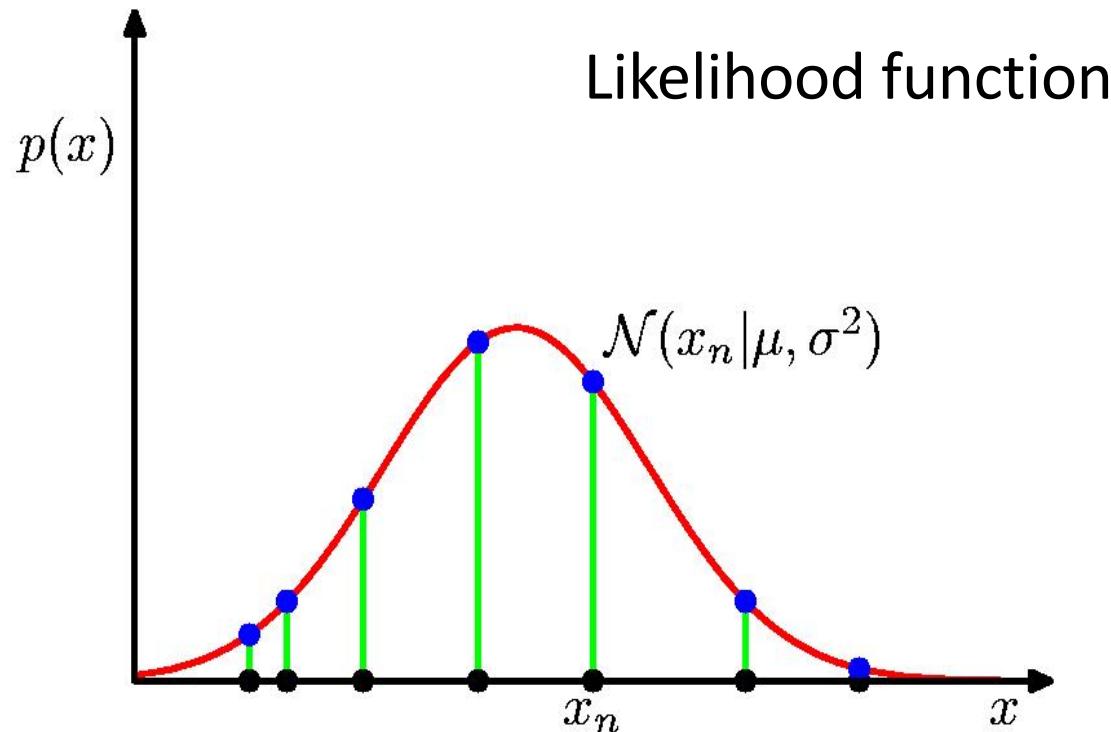
Vector Mean      D-by-D Covariance Matrix      Determinant of the covariance matrix



*Marginals and conditionals of multivariate Gaussians?*

# Gaussian Parameter Estimation

- What is the probability that a dataset  $X$  with  $N$  i.i.d. points was obtained from a specified Gaussian?



*Likelihood of observing  
the given set of points  $X$*

$$= \prod_{n=1}^N \mathcal{N}(x_n | \mu, \sigma^2)$$

# Maximum (Log) Likelihood

---

- Selects the Gaussian that most likely produced the given dataset  $\mathbf{x}$ .

$$\boxed{\text{Log Likelihood} = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)}$$

(Note: for fixed  $\sigma$ , “cost” is sum/mean squared error)

- Maximized when

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n$$

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$$

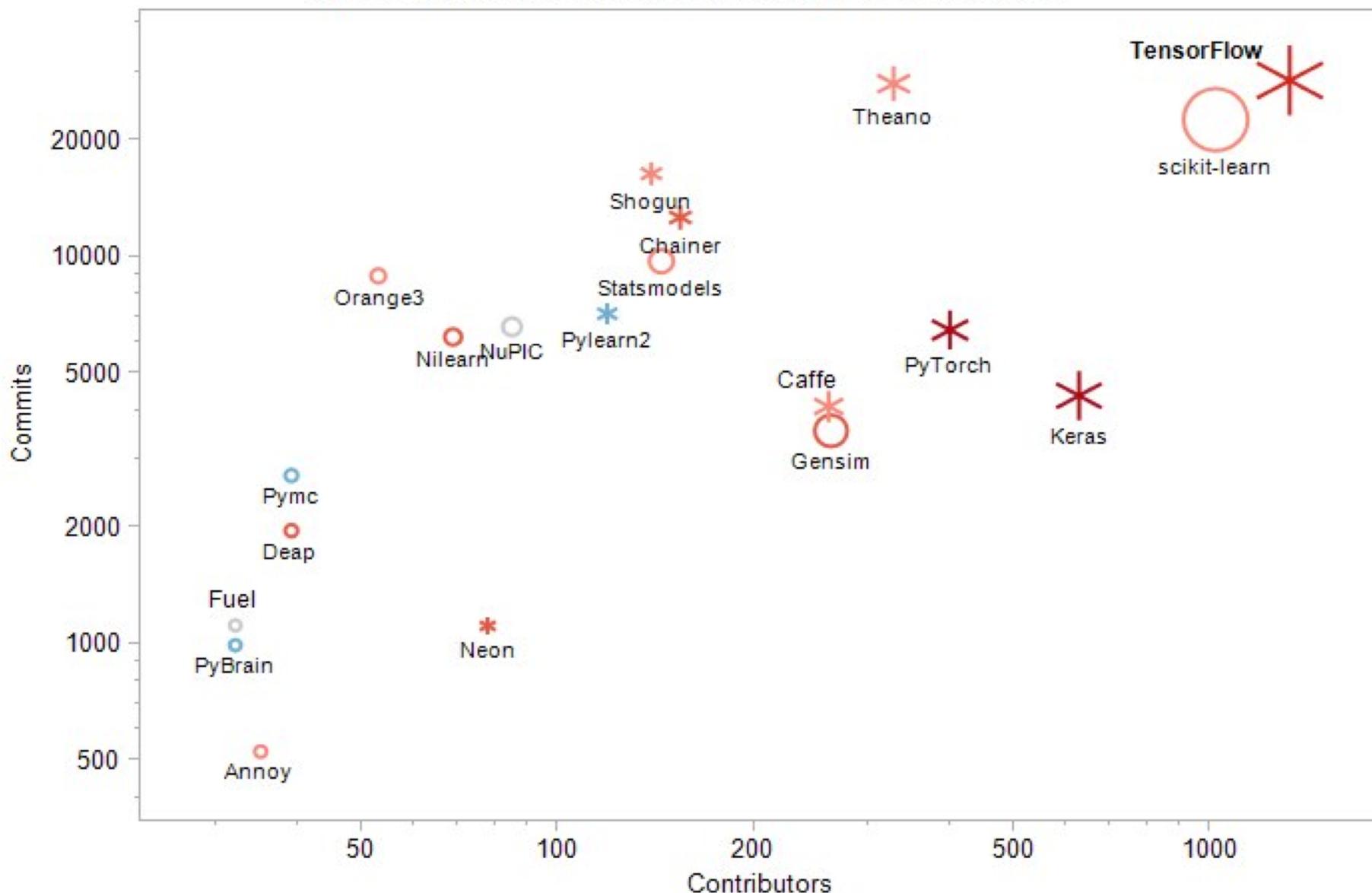
Are ML estimates *biased*?

# Extras

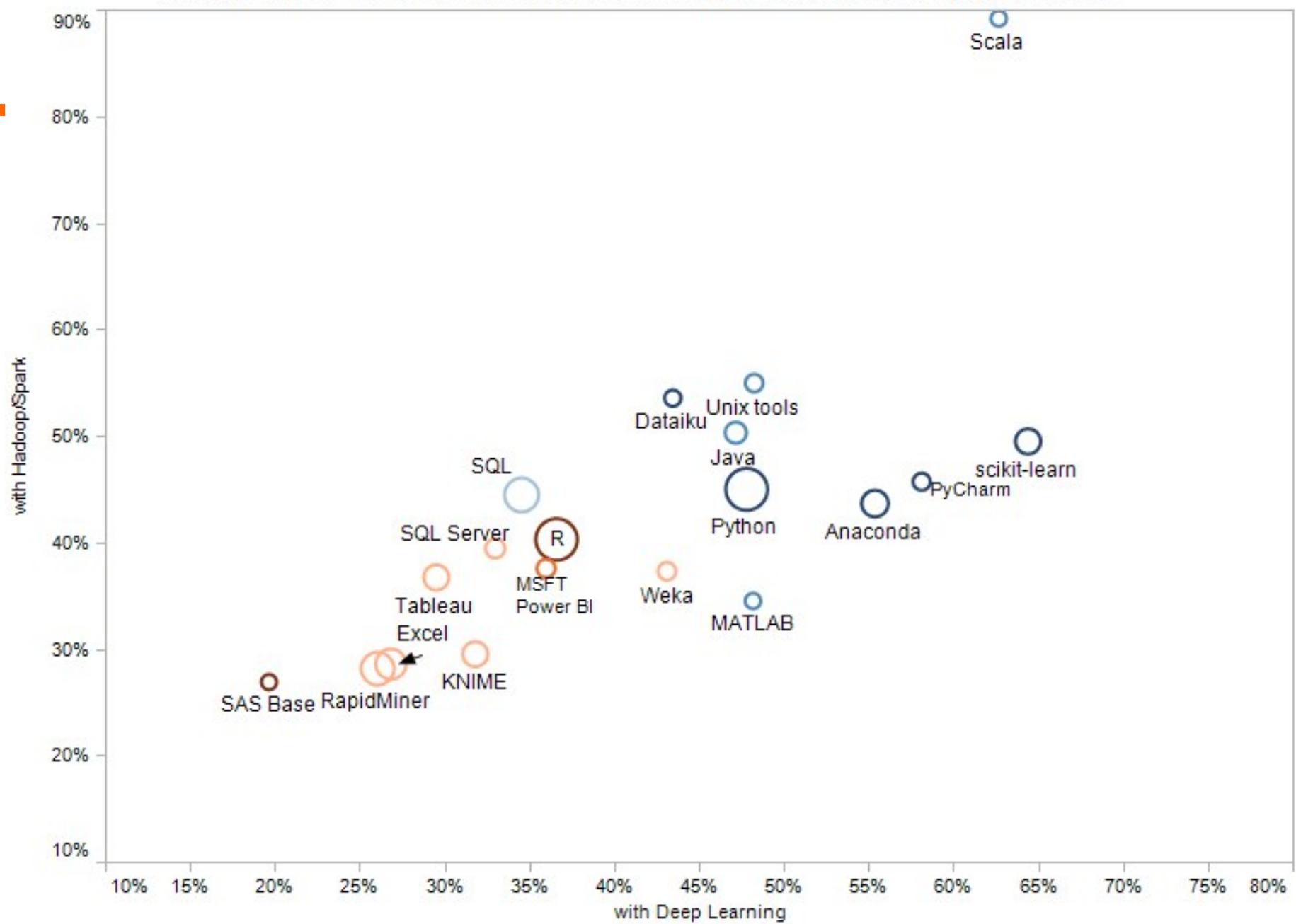
---

- **Top 20 Python ML Open Source Projects (on Github)**  
<https://www.kdnuggets.com/2018/02/top-20-python-ai-machine-learning-open-source-projects.html>

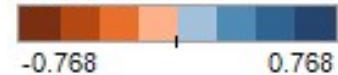
Top 20 Python AI and Machine Learning projects on Github



# KDnuggets 2017 Data Science Software Poll: Deep Learning vs Hadoop/Spark affinity



log(Py/R)



Joydeep Ghosh UT-ECE

27

# Types of Modeling

---

Consider a large collection of fruits

- How to characterize each fruit? (weight, volume, .....)
  - How many “attributes” to use?
- **Regression:** predicting shelf life based on other attributes....
- **Classification:** predicting what type of fruit is it? (class hierarchy!)
- **Ranking and Recommendations**
- **Clustering (grouping):** how many different types of fruits are there?
- **Anomaly detection**
- **Sequence analysis**
- **Causality:** what causes the fruit to rot?
- .....
- **Topics not new, but some data mining concerns/aspects are...**
- **Predictive Modeling:** Developing mathematical models that make accurate predictions
- **Prescriptive Analytics**

# A Brief History of Data Mining Society

---

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD started in 2007

# Conferences and Journals on Data Mining

---

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining ([KDD](#))
  - SIAM Data Mining Conf. ([SDM](#))
  - (IEEE) Int. Conf. on Data Mining ([ICDM](#))
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining ([PKDD](#))
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining ([PAKDD](#))
  - Predictive Analytic World ([PAW](#)) industry focussed.
- Other related conferences
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

# Where to Find References?

---

## DBLP, CiteSeer, Google Scholar

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD (new)
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: JMLR, Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

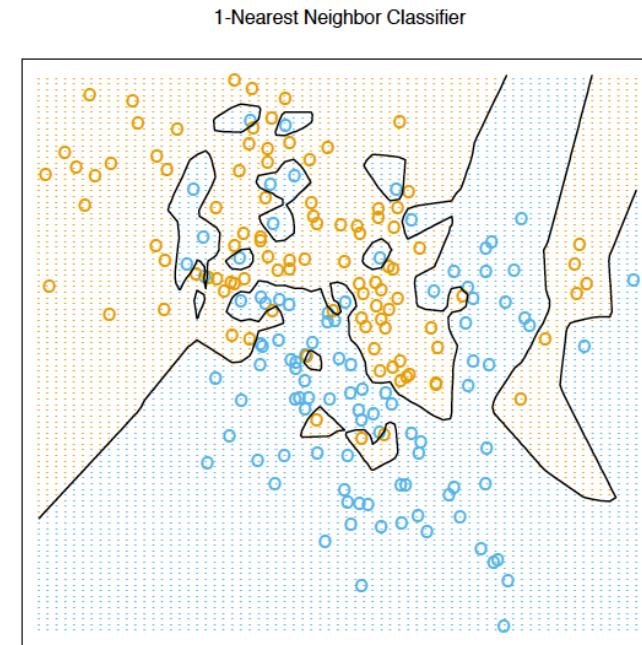
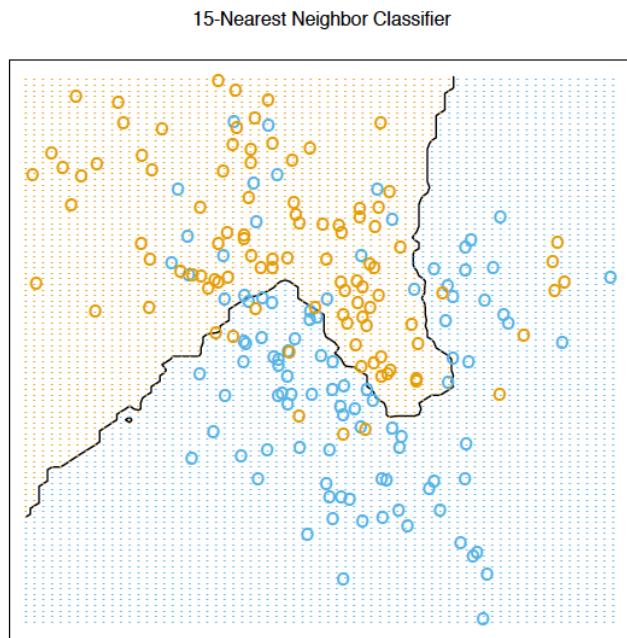
# Data Mining Group at the University of Texas

---

- Joydeep Ghosh (ECE):
  - web mining; adaptive systems; image/signals; system integration
- Raymond Mooney (CS):
  - Machine learning; natural language processing
- ML Theory Oriented: Sujay Sanghavi (ECE) sparsity); Caramanis (ECE) optimization; P. Ravikumar (CS) graphical models/stats;
- Applied Statisticians (McCombs)
- Inderjit Dhillon (CS) – linear algebra; Ed Marcotte (Biochem) – bioinformatics ; Mia Markey (Biomed); Kristen Grauman (CS) – computer vision; Maytal T. (marketing); Matt Lease (I-school) text; Jason Baldridge (linguistics)...
  - NEW: Division of Statistics and Scientific Computation  
<http://ssc.utexas.edu/>
  - **McCombs: Masters in Business Analytics**

# Local vs Global Models (from HTF Ch. 2)

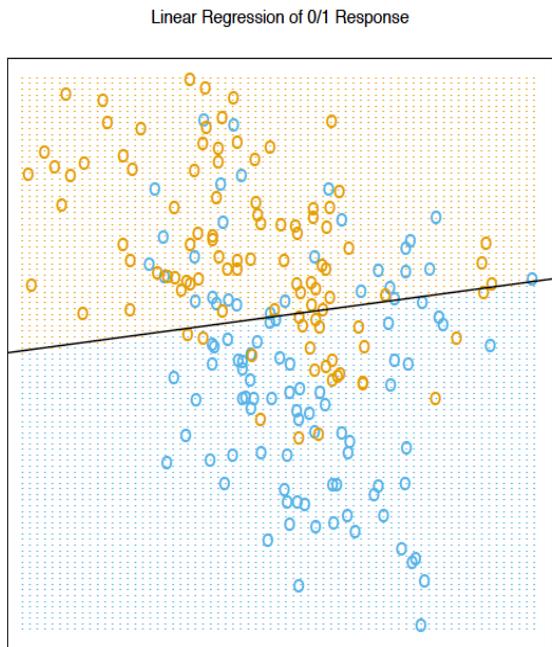
- K-nearest neighbor (KNN)



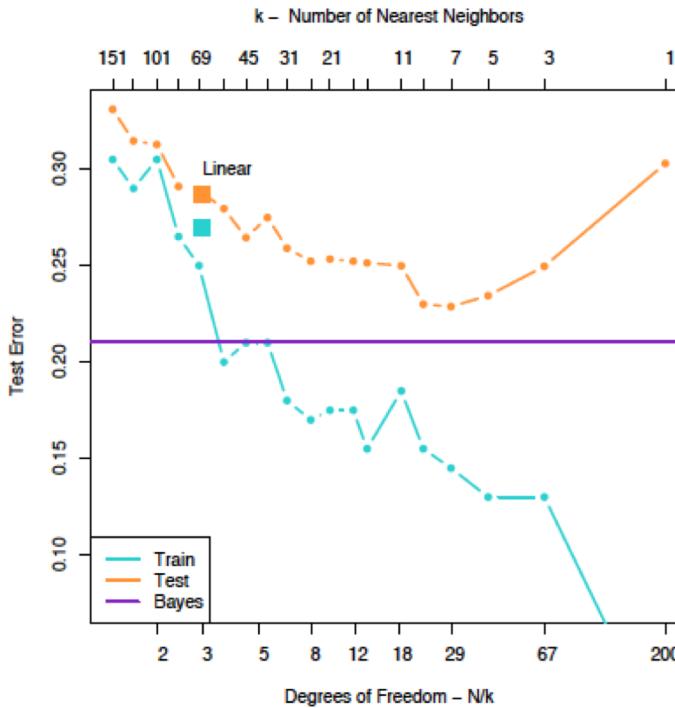
**FIGURE 2.2.** The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (**BLUE** = 0, **ORANGE** = 1) and then fit by 15-nearest-neighbor averaging as in (2.8). The predicted class is hence chosen by majority vote amongst the 15-nearest neighbors.

**FIGURE 2.3.** The same classification example in two dimensions as in Figure 2.1. The classes are coded as a binary variable (**BLUE** = 0, **ORANGE** = 1), and then predicted by 1-nearest-neighbor classification.

# KNN vs Linear Regression



**FIGURE 2.1.** A classification example in two dimensions. The classes are coded as a binary variable (**BLUE** = 0, **ORANGE** = 1), and then fit by linear regression. The line is the decision boundary defined by  $x^T \hat{\beta} = 0.5$ . The orange shaded region denotes that part of input space classified as **ORANGE**, while the blue region is classified as **BLUE**.



**FIGURE 2.4.** Misclassification curves for the simulation example used in Figures 2.1, 2.2 and 2.3. A single training sample of size 200 was used, and a test sample of size 10,000. The orange curves are test and the blue are training error for  $k$ -nearest-neighbor classification. The results for linear regression are the bigger orange and blue squares at three degrees of freedom. The purple line is the optimal Bayes error rate.

# Towards Good Predictive Models

---

- Use data driven models to complement domain expertise and intuition (see quotes in KJ 1.2)
  - Understand problem context
  - Get relevant data
  - Use versatile toolbox and select appropriately
    - Prediction vs. interpretation tradeoff
    - Tailor to data properties
      - » But do not overfit
  - Convey results effectively

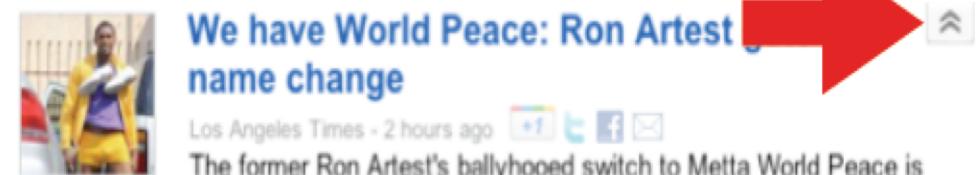
# Degrees of Supervision in Learning Algorithms

- 1. Unsupervised (clustering)
- 2. Semi-supervised
- 3. Noisy or implicit labels (examples to the right)
- 4. “Ground truth” – actual value of output known for some data
- Internet scale analytics relies a lot on 1-3:

- Ads



- Click feedback



- Emails



- Tags

