Invitation to Data Science

Donghui Yan

University of Massachusetts Dartmouth

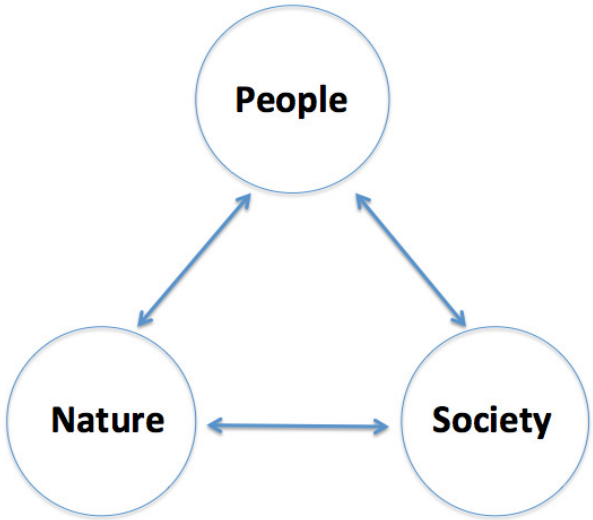
December 28, 2016

Outline

- Introduction
- Case studies
- Data science and the life cycle
- Deep learning by examples
- Data science at UMass Dartmouth

Introduction

- Data records activities about
  - *People, society, the nature, and their interactions*
  - Helps understand and gain insights
    - Just think how we infer about people and events in history
    - Written records, archaeological findings, folklore etc as *data*



- Of particular interest is high-level 'understanding' of data
  - Summary, visualization, and analysis
  - ♠ Where Data Science sets in.

Example Data Science applications

Examples

Examples

- E-commerce
- Weather
- Sports statistics
- Healthcare
- Credit scoring
- Retail stock selection
- Enterprise

♠ Data science is relevant in any domain as long as there is data.

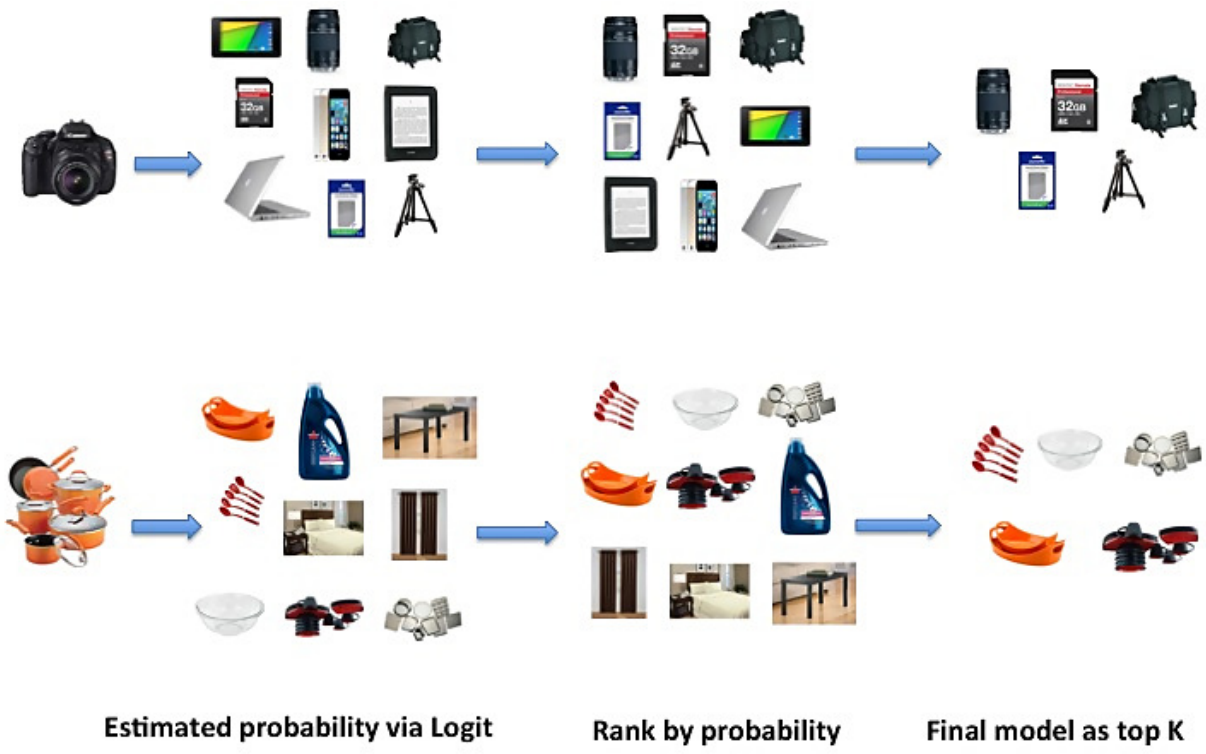
How to leverage the value of data in e-commerce

- Large number of user accesses to a typical e-commerce site
  - e.g., walmart.com has tens of millions of user access each day
  - Every mouse click captured by e-commerce server
    - e.g., view, cart, purchase of an item
  - ♠ Huge data of sales record, how to make use it?

- *Observation*: a user typically buys several items together
  - e.g., a user bought $B_1, ..., B_t$ in the same transaction
  - Any pair (or triple etc) of items from $B_1, ..., B_t$ called co-bought items
  - If a user buys an item, he tends to buy co-bought items as well.

Item recommendation in e-commerce

- Co-bought stats can be used to build a recommendation model
  - Users buying/viewing product A may also like B
  - An effective way to promote sales
    - e.g., *Amazon* boosts its sales by 20-30%, *Walmart* by 5-10%

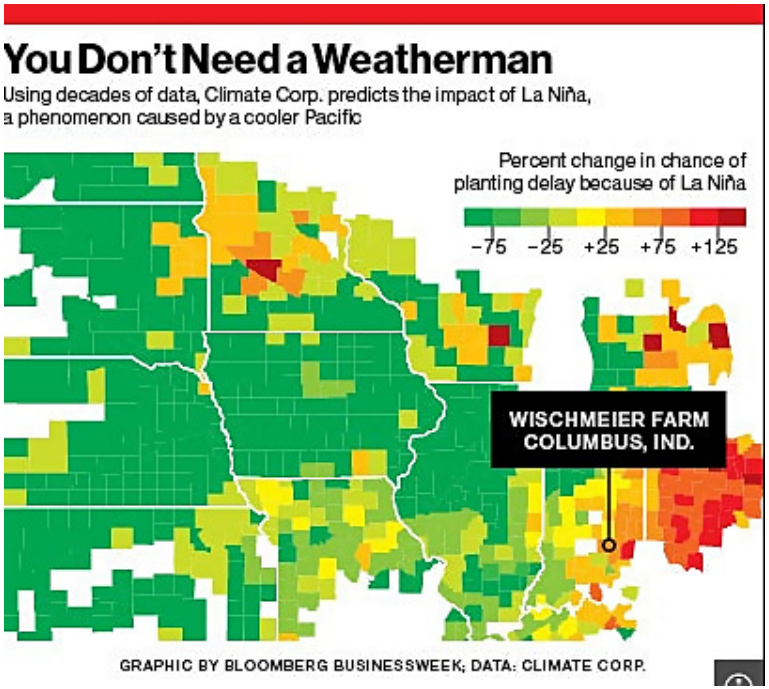


Big data @Walmart (continued)

- Market basket analysis
  - Affinity-based shelving for stores/warehouse
  - Affinity-aware allocation algorithm for logistics
    - Products often bought together are shipped together
    - To reduce inventory, package splits, and improve order fulfillment.

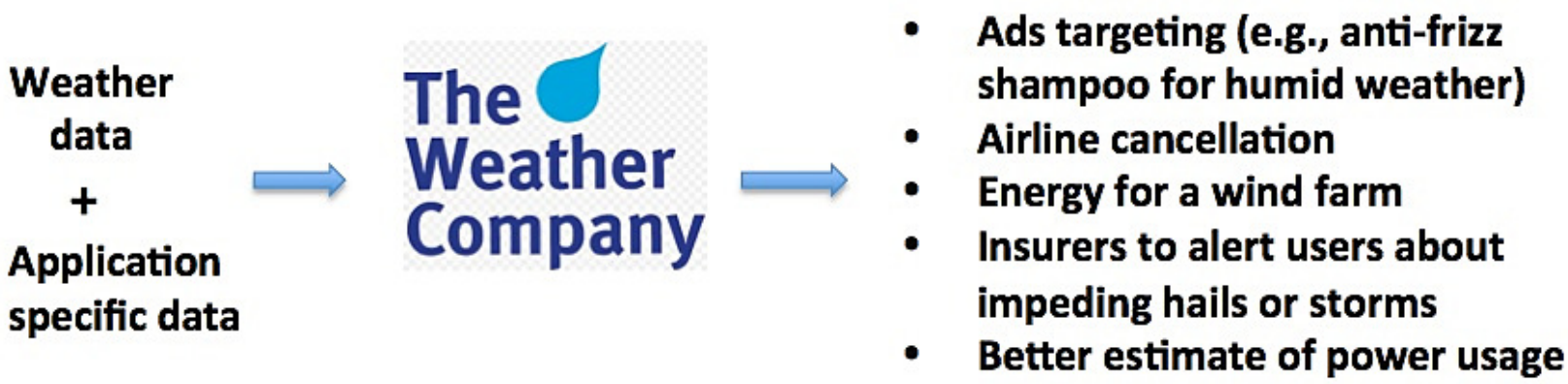
The Climate Corporation

- Provide insurance to farmers under extreme weather conditions
  - Global agriculture industry estimated at $3 trillion
  - Weather data at 2.5m locations + 150b soil measurements
  - 10 trillion weather simulation data points
- Acquired by Mosanto at $930m in 2013.



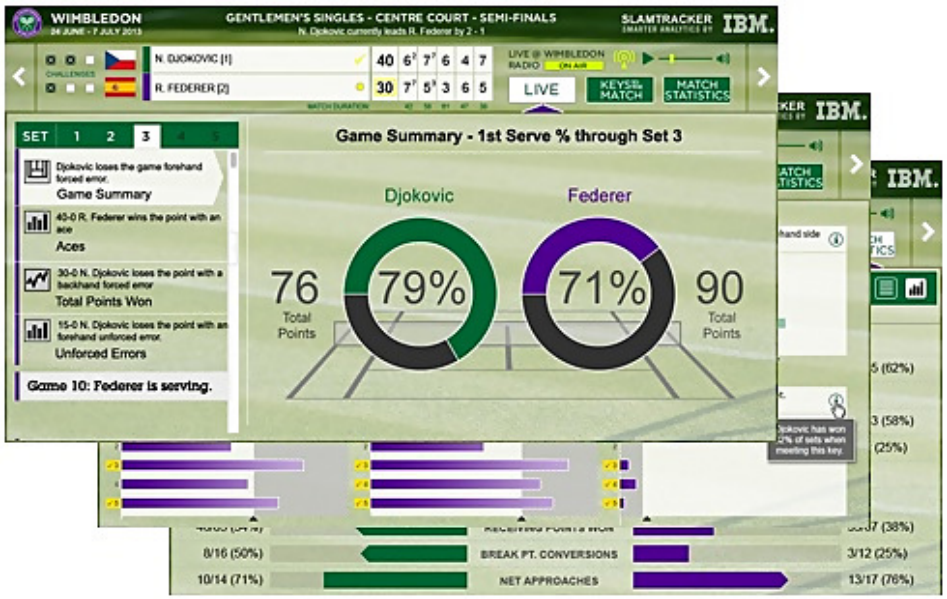
The Weather Company

- Weather is one largest external swing factor in business
  - Annual economic impact $\approx$ \$0.5 trillion in US
- To help understand behaviors of digital and mobile users
  - In 3 million locations worldwide
  - Along with climate data in each locale
- Half of its revenue from digital operations.



Sports statistics at IBM

- IBM's SlamTracker
- 8 years of *Grand Slam* tennis tournaments data
  - $\sim$ 41m data points
- Top 3 key actions to enhance chance of winning
  - Target serve percentages, rally counts, types of shots
- In 2012, USOpen.org logged 45.6m visits and 325m page.



Healthcare

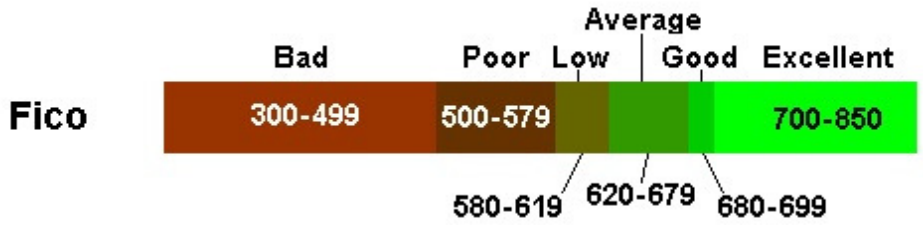
- Health data used in various ways
  - Prevention, intervention, diagnosis, and maybe re-admission
- Devices
  - Smart phones (e.g., for blood test from Berkeley, EPFL)
  - Wearable devices (FitBit, Jawbone, Samsung Gear Fit etc)

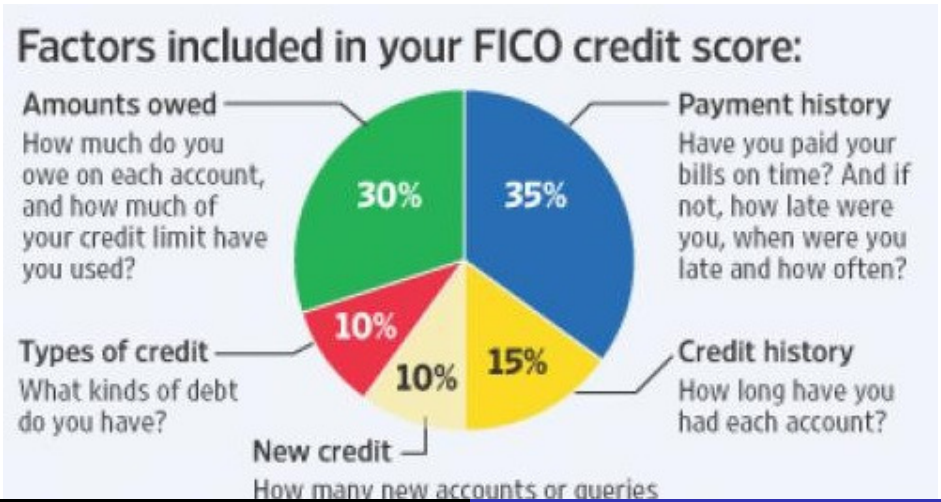
Healthcare (continued)

- e.g., *Pittsburgh Health Data Alliance* uses various data
  - Medical records, insurance records, wearable sensors
  - Genetic data
  - Usage of social media (connection to lifestyle)
- New way of treatment
  - From visiting *webmd.com* and self-diagnosis, to
  - Healthtap (one-to-one remote treatment, or telemedicine)
- Personalized medicine, privacy and security etc.

FICO analytics

- 10b+ FICO scores purchased in 2013
  - Excluding 30m personal purchases by US customers
- 65% credit cards managed worldwide
- 2.5b payment cards protected from fraud





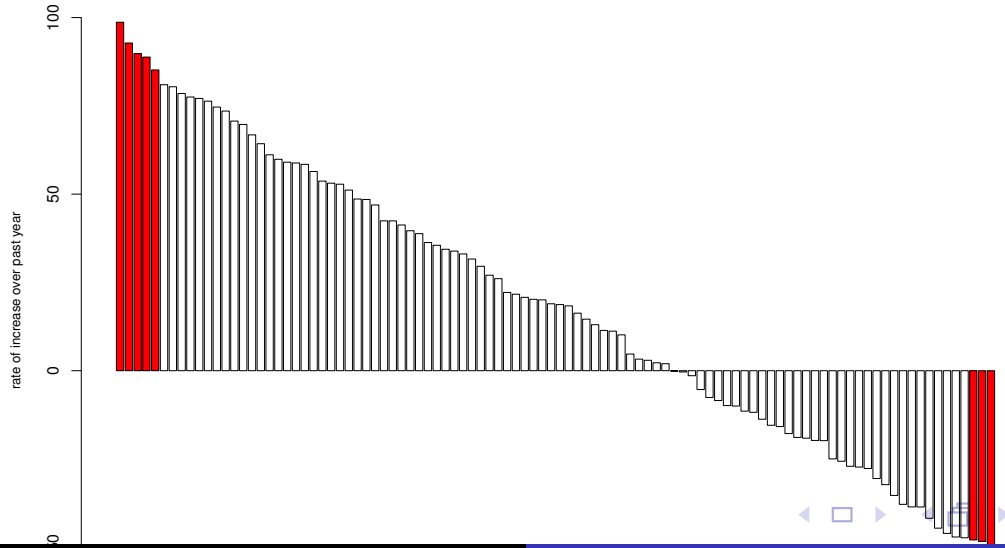
Variables used in FICO score calculation

- Individual
  - Credit history, credit usage, defaults
  - Demographic: age, gender, education, ethnicity
  - Buying behavior: products, channels, frequencies etc
  - Hobbies, affiliations, interests etc
- Aggregated (geographical)
  - Summarized credit info: avg credit limit, ave percent utilization
  - Census: Avg income, percent occupied houses etc
- Economic variables
- Property data
- Electronic medical records

$\Rightarrow$ Credit score (between 300-850).

A strategy for picking stocks (**Not** recommended)

- *Chase* sued two employees for violations in 2014
  - Query past year transactions by all Chase credit card holders
  - *A few days before the announcement of annual sales and aggregate by company*
  - Pick a few best or worst performing companies
  - Buy or short their stocks
- Easily earn a few million in the last few years
  - Earn $50k–200k from each transaction.



Everstring

- To help find or rank potential customers
  - Operating efficiency improves by 50%
  - Conversion rate up by 300%
- Profile of over 11m companies and vendors
  - Relevant customers information by web crawling
  - Third-party CRM (customer relationship management) data
- 3 rounds of investments about $78m
- 2 of 3 founders graduated from math@SJTU.

Other applications

- Search engine design
- Computational advertising
- Marketing, and hedge fund
- Digital journalism
- . . . . . .

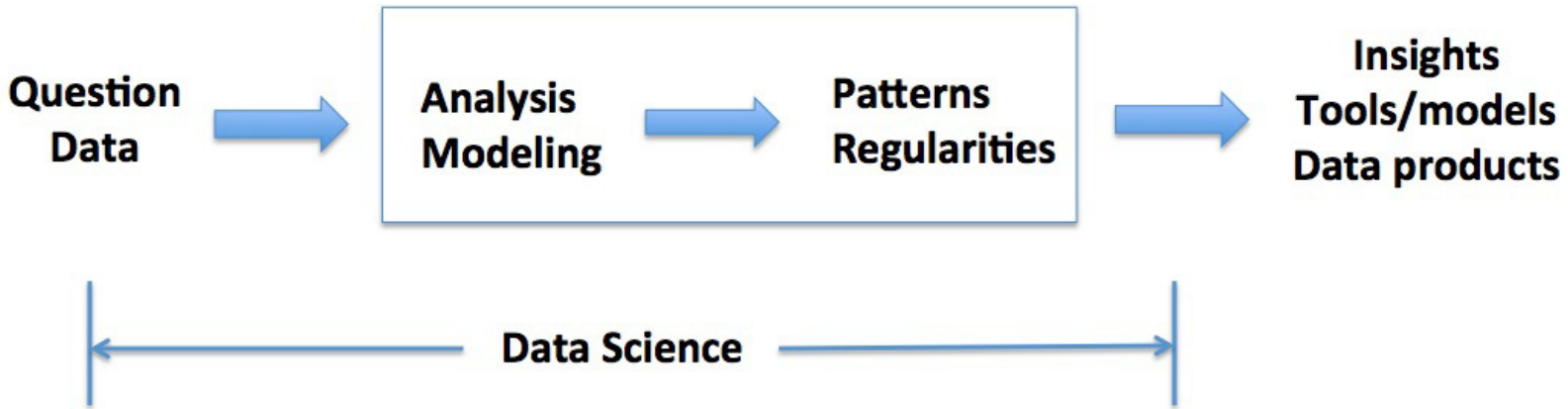
Data science and the life cycle

Data science and the life cycle

What is data science?

- Data Science is the science of data

  A discipline that provides principles, methodology, or guidelines for the analysis of data for tools, values, or insights.



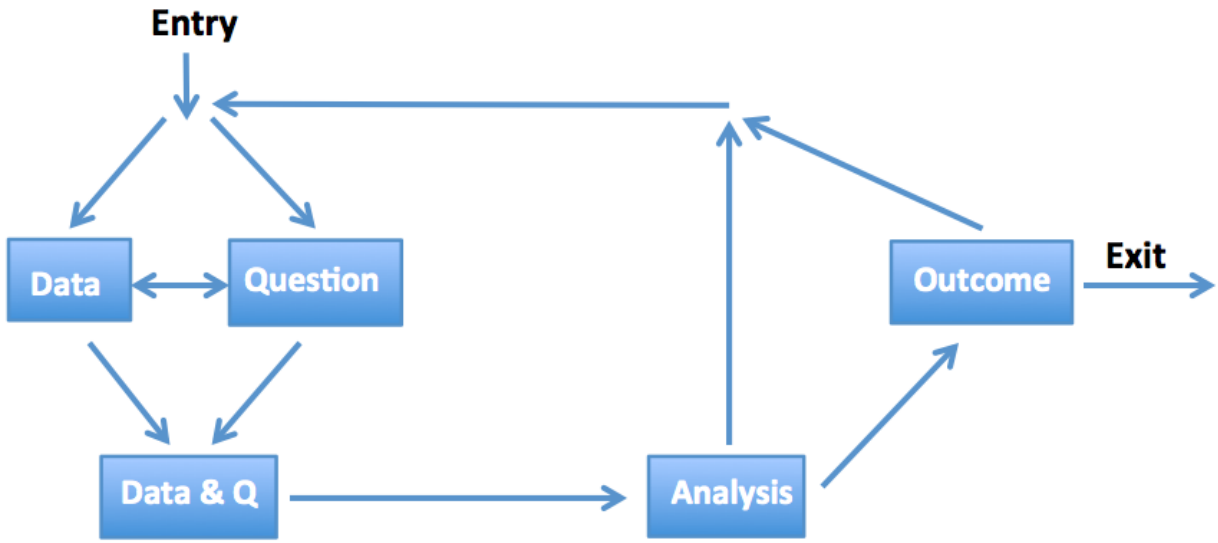
Tasks involved in data science

- Data collection (including preprocessing)
- Formulation of relevant Qs by knowledge or from data
- Exploratory data analysis (EDA)
  - Data summary/transformation, graphical visualization etc
- Confirmatory data analysis (CDA)
  - A sound framework for conclusions or assessment of results
  - Modeling
    - To find patterns or regularities from data
    - Data $\rightarrow$ patterns/regularity $\xrightarrow[\text{domain}]{}$ models

♠ Output as insights, models/tools, or data products.

Data science life cycle

- Data science practice as a process
  - A series of steps (states) with 'sequential' dependency
  - Called the *data science life cycle*
    - Analogy from *software life cycle*.



Data science Vs statistics

- Extend the "generative culture" in statistics
  - To embrace the "predictive" culture (Leo Breiman, 2001)
  - Machine learning seamlessly fits in
    - Convergence of statistics and machine learning
    - Two historically separate *science of learning from data*
- Why is this important?
  - Unified treatment of wide problems
    - Estimation, classification, clustering, ranking, topic modeling etc
    - Under the general term of "modeling" or "analysis"
  - Potentially better results than traditional statistics
    - e.g., SVM, boosting, RF often outperform logistic regression
  - May solve problems not feasible for traditional statistics
    - e.g., deep learning for tasks related to image/audio.

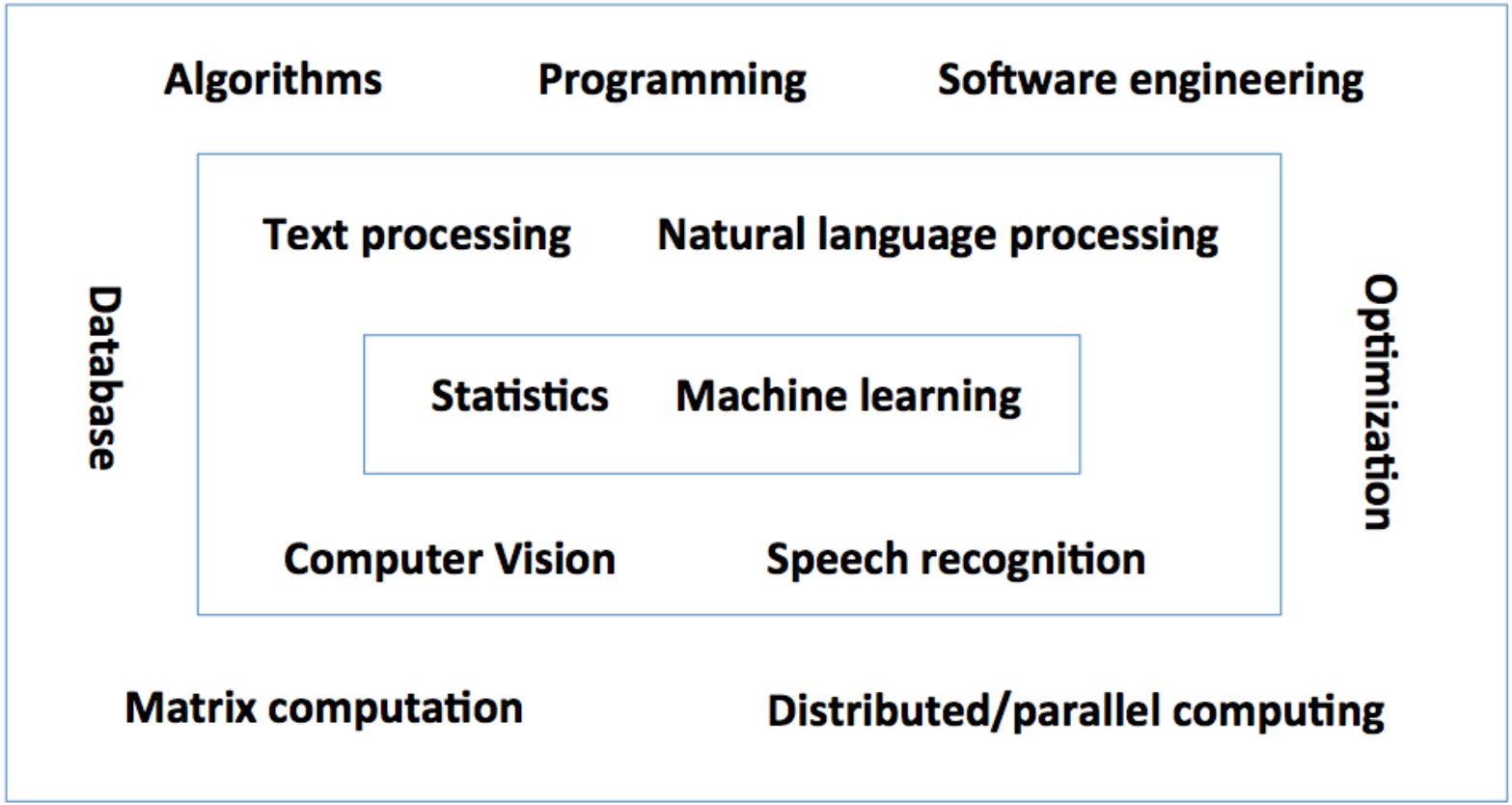
Data science Vs statistics

- May start by exploring data without a question in mind
  - Come up with interesting question via EDA
- Why is this important?
  - As a matter of fact, large data collected in various domains
    - General consensus is *data entails value*
    - Imagination is the limit
  - How to leverage the value of such data?
    - Pattern/regularity may inspire new lines of business
  - Driving force behind the explosive growth in data science
    - c.f. may examples discussed before.

Data science Vs statistics

- Data visualization becomes more important than ever
  - Only limited use in traditional statistics
  - Though a long history in statistics since 60-70's
    - e.g., Tukey's EDA, Huber's projection pursuit, Chernoff faces
- Why is this important?
  - No way to summarize big complex data with few simple statistics
  - Better sense about spatial, temporal, heterogeneous data
  - Reproducibility for data driven practice
    - e.g., data provenance.

Key enabling technology



Deep learning

Machine learning in real life

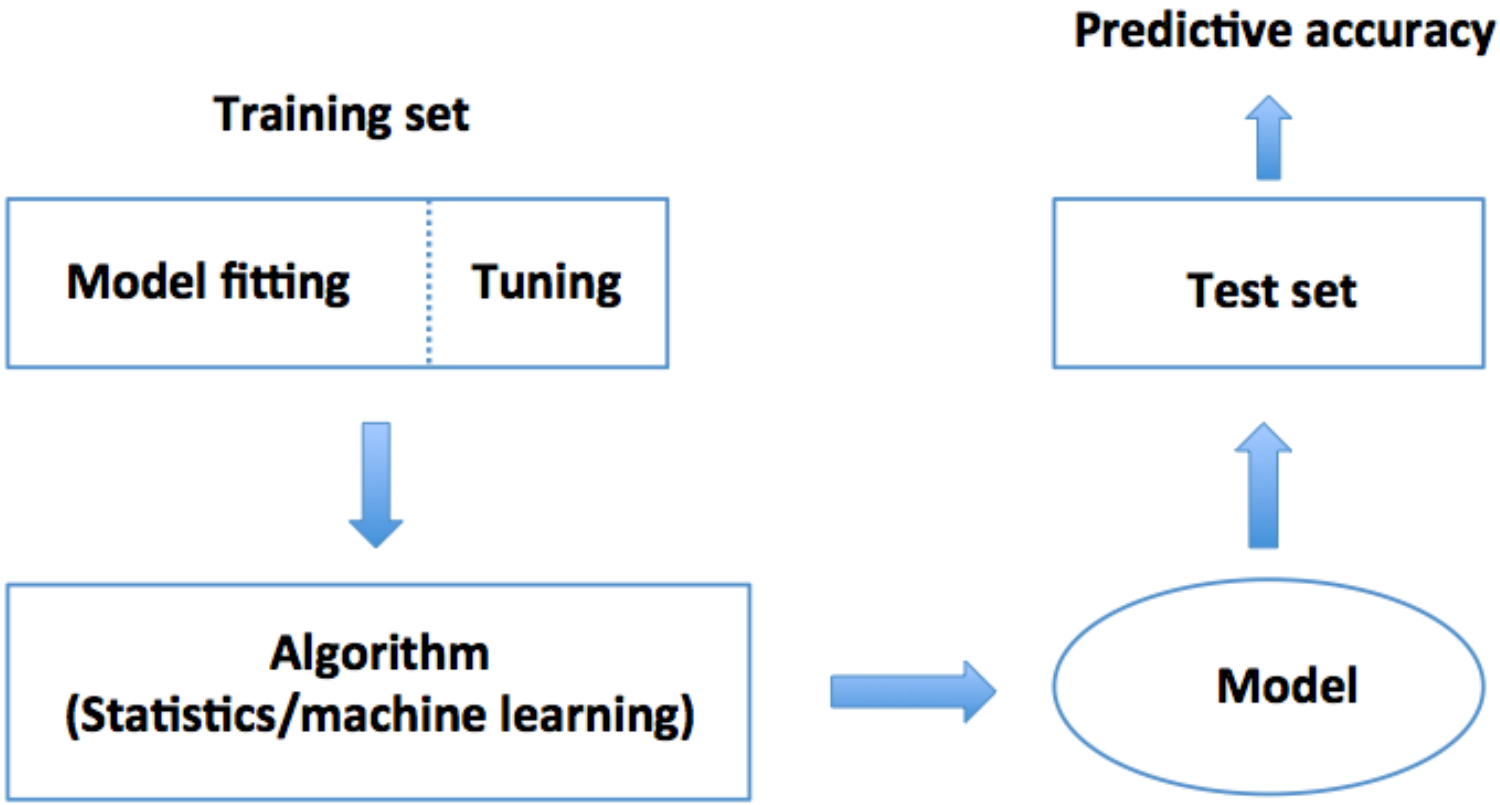
- Search engine design
  - To max chance one gets what he searches in top K entries
- Computational advertising
  - Placement of ads to maximize profit
- Design of e-commerce web site
  - Selection of selling items to max click thru rate (or profit)
- Selection of headline news
  - e.g., which news as headline in news portal at Yahoo, CNN etc
- Email spam filter.

What is machine learning (ML)?

Applications, methodology, and theory relevant to the automatic learning of patterns or regularities from data.

- Given data $(X_1, Y_1), ..., (X_n, Y_n)$, learn the relationship $f : X \mapsto Y$ s.t. the future is 'best' predicated.
  - X_i called features, $Y_i \in \{1, 2, ..., J\}$ called *labels or classes*
  - $(X_1, Y_1), ..., (X_n, Y_n)$ is called a *training sample*
  - $\hat{f}$ is called the *trained or fitted model*
- Two important *assumptions*
  - The future is related to the past
  - Knowledge about the problem under study
    - Features being the most elementary form.

Machine learning model fitting



Types of problems in machine learning

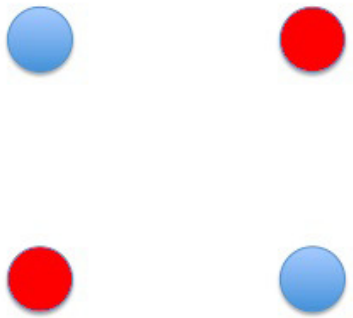
- Classification (supervised)
  - $Y \in \mathcal{C} = \{c_1, c_2, ..., c_k\}$, called labels
- Clustering (unsupervised)
  - Y not given (often called unsupervised learning)
- Ranking
- And a lot more new topics emerging in recent years
  - Topic model (e.g., what is the topic of a blogger article)
  - Manifold (topological) learning
  - Salient sentence extraction
  - Graph learning etc.

Machine learning topics

- Methods
  - Supervised: decision trees, logistic regression, neural network, deep learning, SVM, boosting, Random Forests, ...
  - Unsupervised: kNN, k-Means clustering, mixture analysis, spectral clustering, manifold learning, topic modeling, ...
  - Optimization: (stochastic) gradient/coordinate descent
  - Regularizations to learning algorithms
- Theory
  - Statistical learning theory (Vapnik and Chervonenkis, 1964-1974)
  - PAC learning (Valiant 1984)
  - Consistency results about various learning methods
- Learning paradigms
  - Active learning, co-training, group-learning, transfer learning, reinforcement learning.

A half-century of neural network

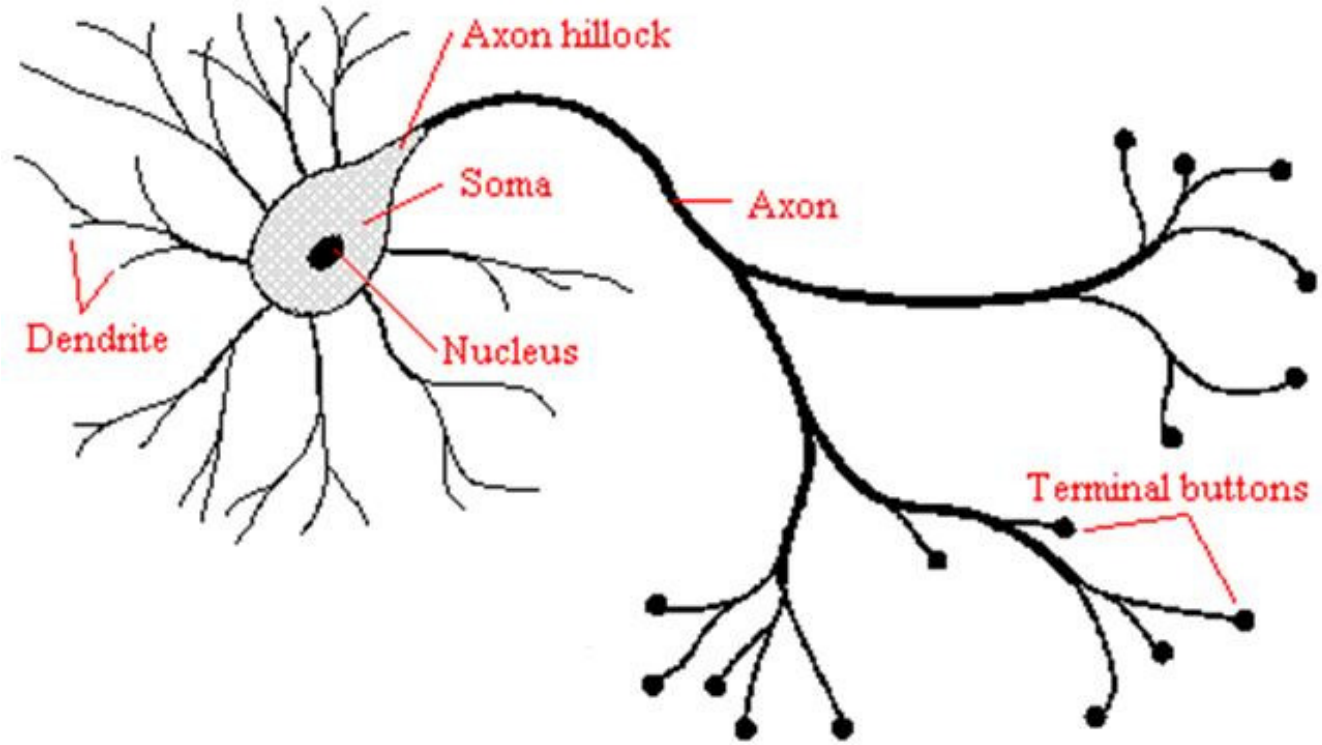
- Rooted in McCullouch and Pitts (1943)
  - A computational model for neural networks based on algorithms
  - The model is called *threshold logic*
- Hebbian learning (Later 1940's)
  - "Cells that fire together, wire together"
- Perceptron (Rosenblatt 1958)
  - The first neural network
- Pronunced death by Minsky and Papert (1969)
  - 2-layer network incapable of solving the XOR problem.



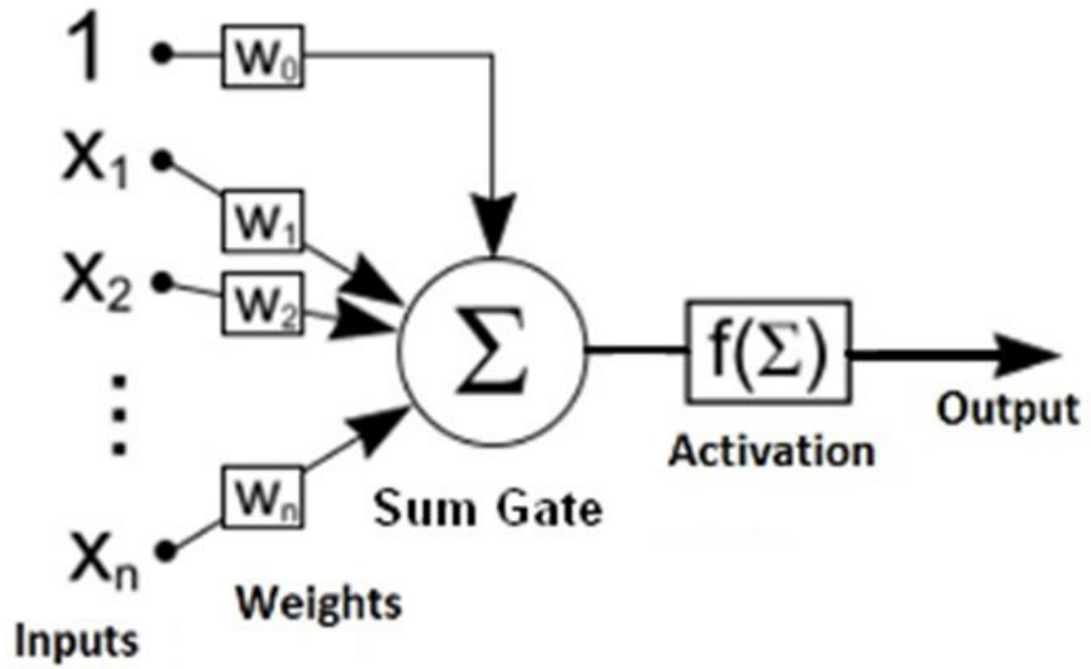
A half-century of neural network (continued)

- The backpropogation algorithm (Werbos, 1975)
- The Cognitron (Fukushima, 1975)
  - An early multilayered neural network with a training algorithm
- Hopfield's network (1982)
  - Ability for bi-directional flow of inputs between neurons/nodes
- Boltzman machine (Hinton and Sejnowski 1983)
  - A stochastic version of Hopfield's network, but hard to scale
- Rumelhart, Hinton and Williams (1986)
  - "Learning Internal Representations by Error Propagation"
- LeNet-1,2,3,4,5 (LeCan et al, 1989-1998)
  - State-of-the-art results on USPS pen digits recognition
- Neural networks regains popularity since 2008
  - Under guise of deep learning (Hinton and Salakhutdinovet, 2006).

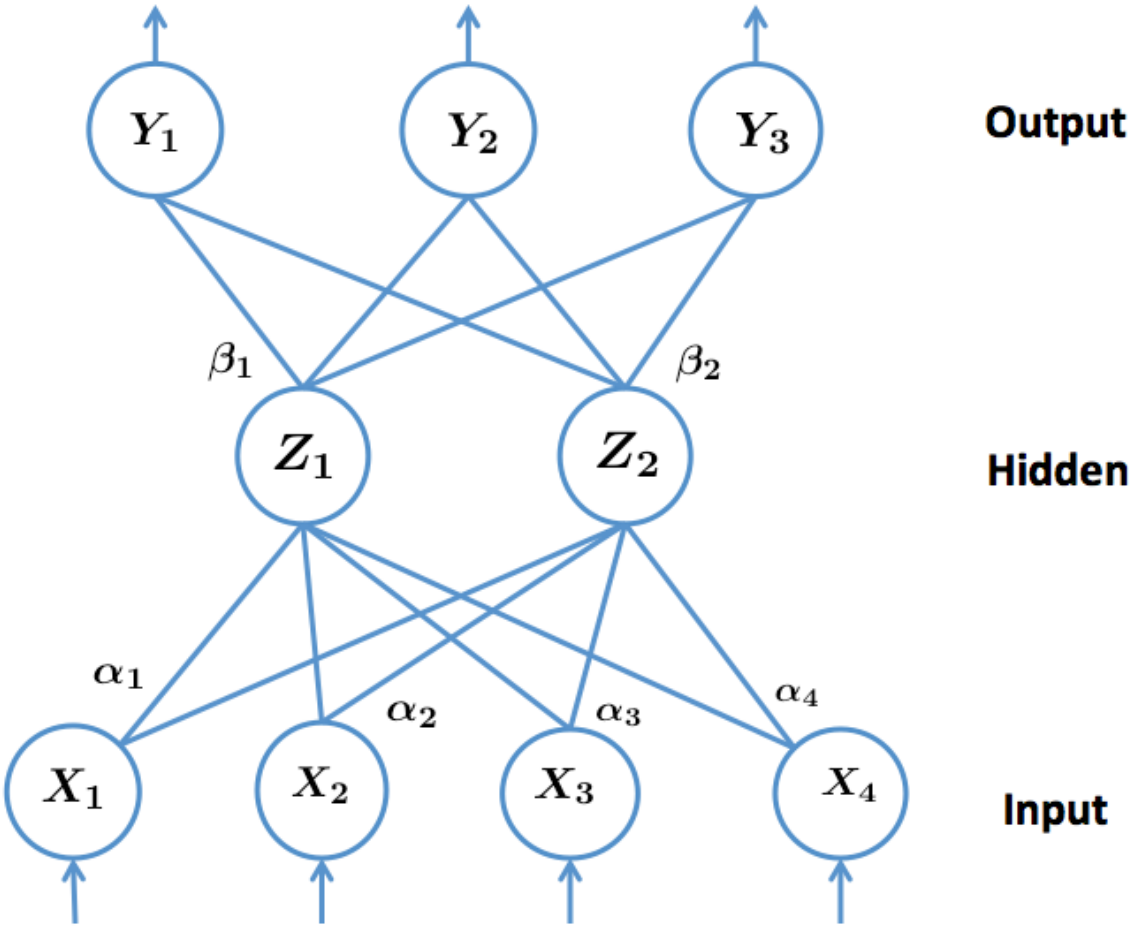
The biological neuron



The artificial neuron



The neural network architecture



The neural network architecture (continued)

- Input $\rightarrow$ Hidden layer

$$X \mapsto \sigma(\alpha_0 + \alpha^T X) \triangleq Z$$

  - A popular choice of σ is $\sigma(x) = 1/(1 + e^{-x})$
  - # parameters = # Input $\times$ # Hidden + # Hidden

- Hidden $\rightarrow$ Output

$$\begin{aligned} Z &\mapsto \beta_0 + \beta^T Z \triangleq T, \\ T &\mapsto g(T_1, ..., T_K) \triangleq f(X) \triangleq Y \end{aligned}$$

  - # parameters = # Hidden $\times$ # Output + # Output

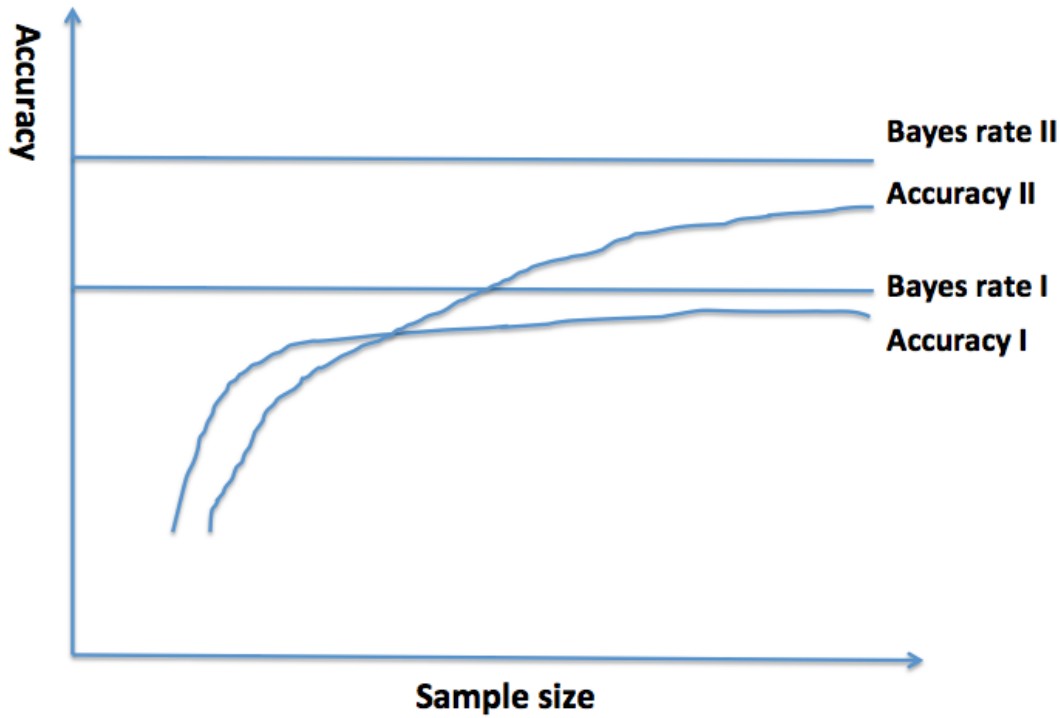
- # Total parameters = # Input $\times$ # Hidden + # Hidden $\times$ # Output + # Hidden + # Output.

The story of deep learning

- Hinton asked Salakhutdinov to work on a deep NN (2006)
  - About the computational issues
  - *Accidently*, some intermediate results turned out to be good representation of the data
  - Remarkable performance when using such representation
    - This marks the "birth" of deep learning
    - *However, deep learning should not be limited to NN*
- Deep Vs usual learning
  - Usual: features obtained by knowledge/feature engineering
  - Deep: features learned by algorithm
    - Typically *multiple levels of (hierarchical) representation* (some def of deep learning require only a representation)
    - e.g., edge, texture-like features found by NN/algorithm
    - Such features typically beyond usual feature engineering.

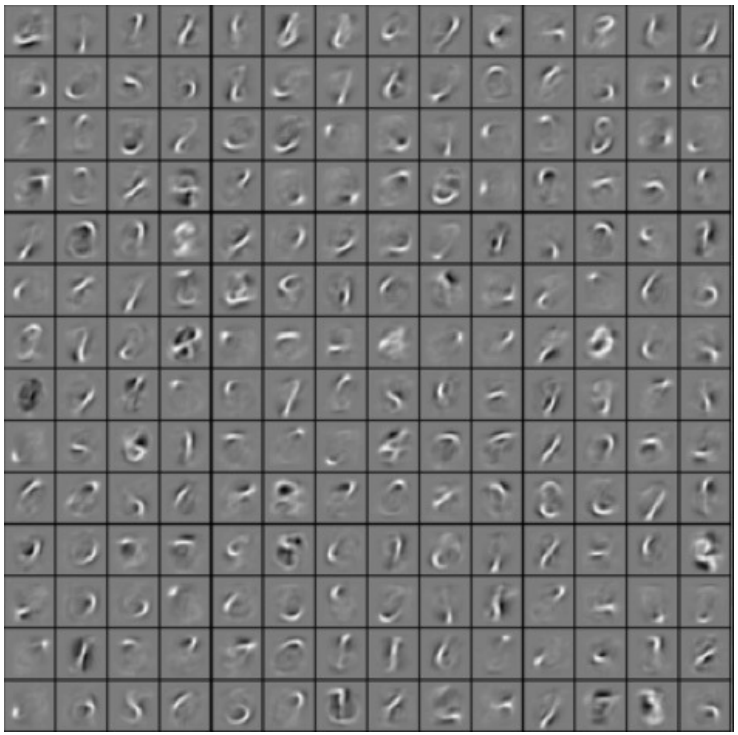
Why may it work? A crude view

- 'I' indicates original data, 'II' with deep features
  - 'II' leads to a larger model space
    - High Bayes rate if deep features *informative*
    - Better accuracy even if gap towards Bayes rate may be large.

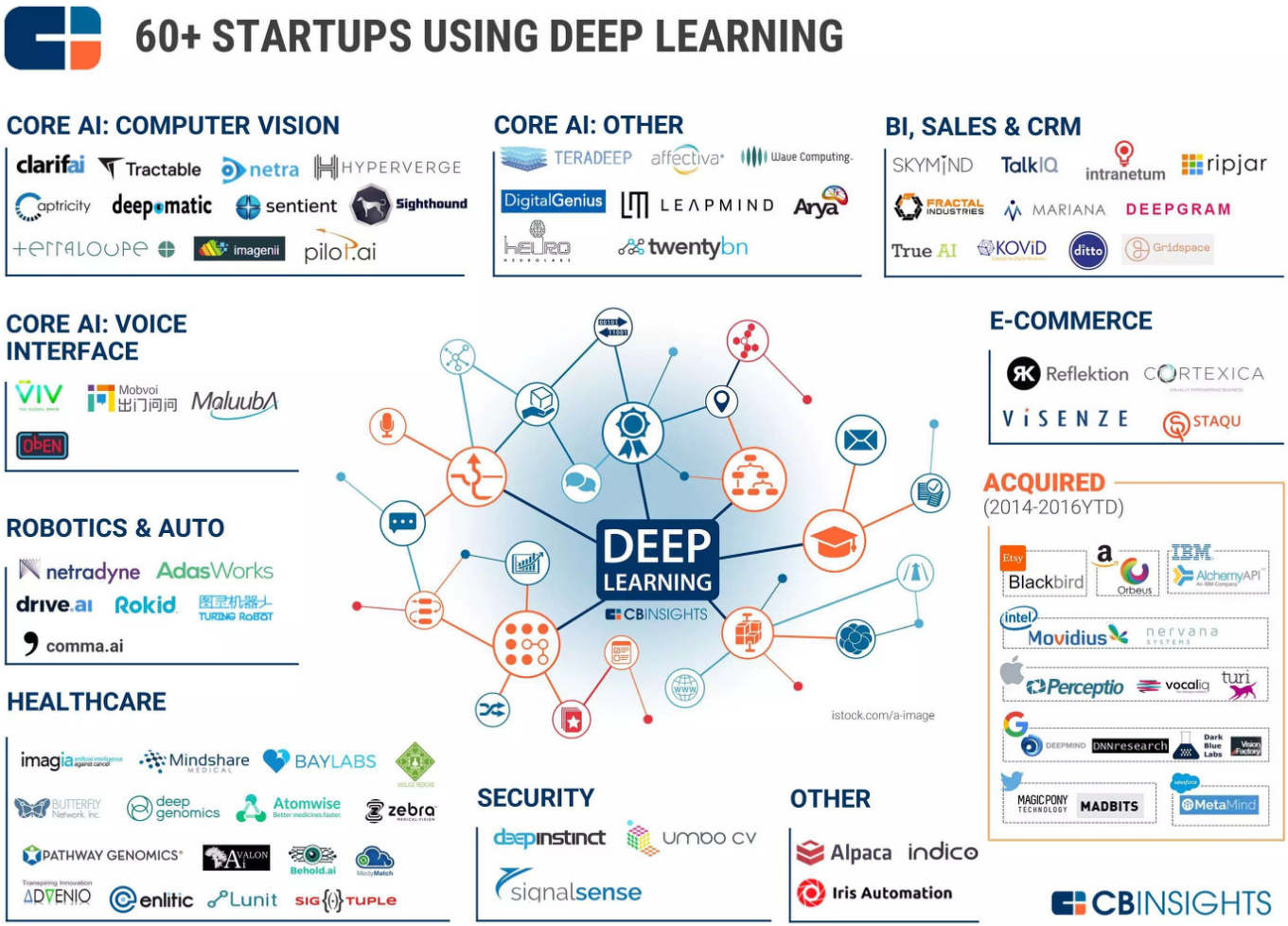


Deep features may be informative

- 1st layer of a deep NN gives edge-like feature
  - 2nd and higher layer texture or shape etc information.



Startups related to deep learning



MNIST digits recognition

- A database of handwritten digits
  - Image size 28 x 28
  - 60000 training, 10000 test images
  - State of the art error rate: 0.27
    - 6-layer CNN 784-50-100-500-1000-10-10

Output with 4 hidden layers

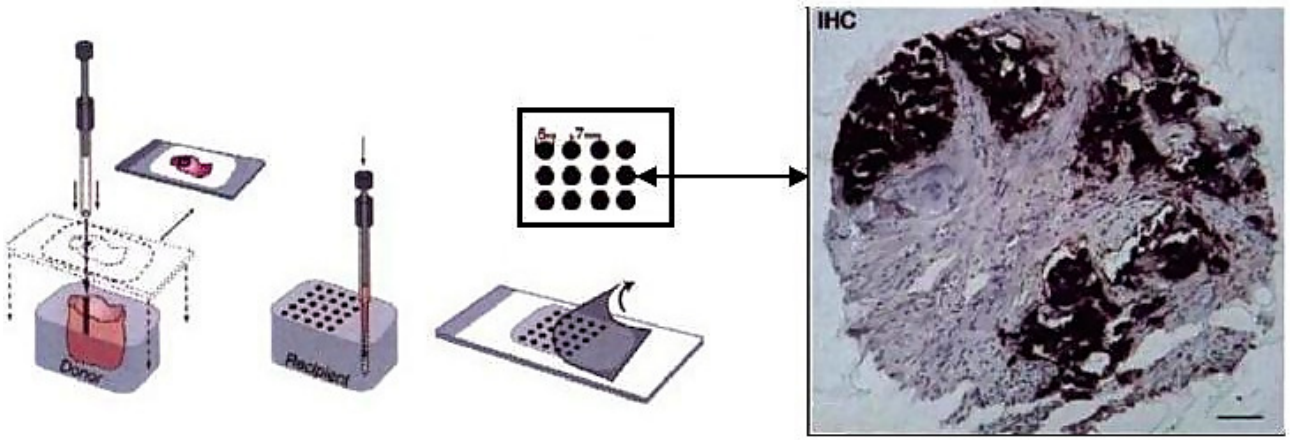
```
> library(deepnet);
> dnn <- dbn.dnn.train(train_x_n, mnist$train$yy,
          hidden = c(300, 80, 50, 20),
          hidden_dropout=0.6, numepochs = 5, cd = 3);
> err.dnn <- nn.test(dnn, test_x_n, mnist$test$yy);
> dnn_predict <- nn.predict(dnn, test_x_n);
> print(err.dnn)
[1] 0.05515
> print(dnn_predict[1,])
 [1] 3.876123e-03 3.326393e-04 2.584891e-03 1.029923e-02
 [5] 5.424675e-04 1.807185e-04 2.153286e-05 9.921616e-01
 [9] 1.166920e-04 3.344022e-02
> print(mnist$test$y[1])
[1] 7
```

Feasibility of deep neural networks

- Large training set
- Sufficient computational resources
- ♠ *Not always feasible*
  - Need alternative ways to find deep features.

Tissue microarray images

- Used in cancer/tumor diagnostics
- Obtain tissue cores from tumor site and store in archive
- Section slices of tissues and mount onto in form of array
- Apply biomarker (stain) and take images.

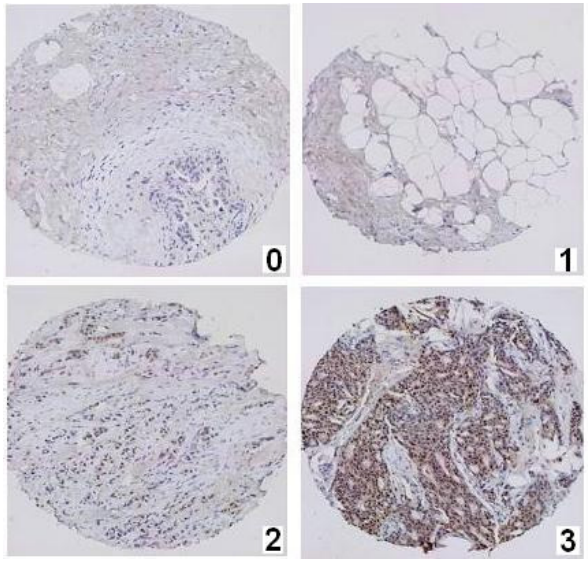


- Each cell in a TMA array $\Longleftrightarrow$ a tissue (image)
  - Each cell in a microarray image $\Longleftrightarrow$ a gene.

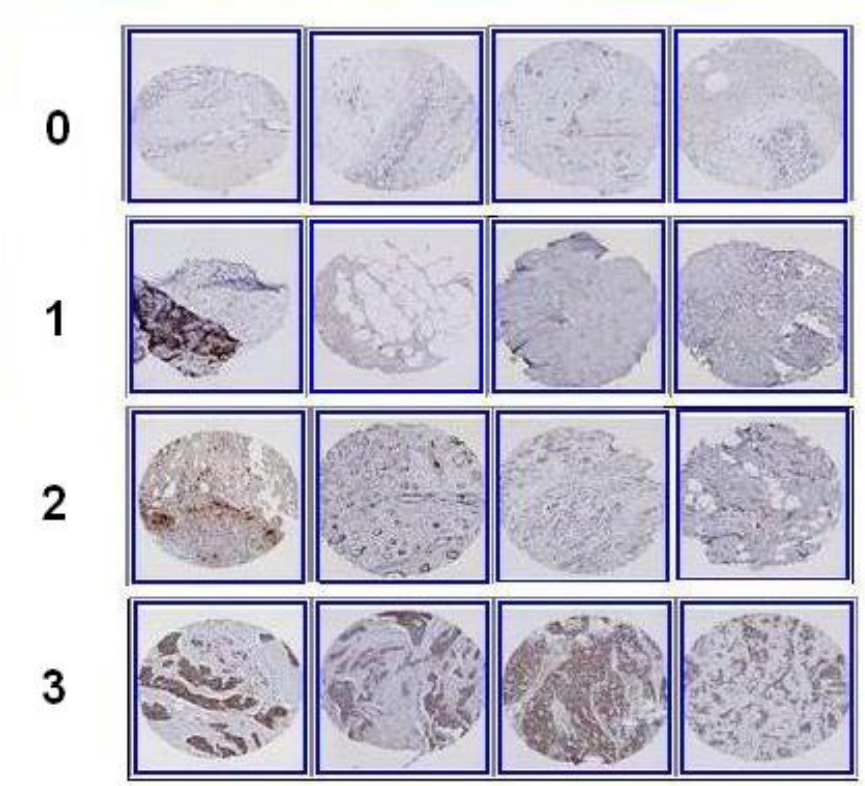
The scoring of TMA images

Measure tumor-specific protein expression level

- 0 - definite negative (no staining)
- 1 - ambiguous or weak staining in a minority of tumor cells
- 2 - weak positive (minor dark or major weak nucleus staining)
- 3 - definite positive (majority show dark nucleus staining)



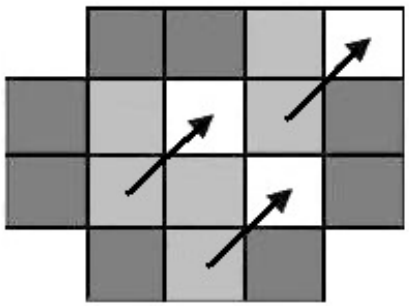
Challenges of TMA image scoring



- TMA images highly heterogenous
  - Staining patterns not localized in position, shape, or size

GLCM (Grey Level Co-occurrence Matrix)

- Each TMA image $\Longrightarrow$ a GLCM
  - GLCM as features for an image in classification
- GLCM entries count # transitions between pixel intensities of neighboring pixels with a spatial relationship



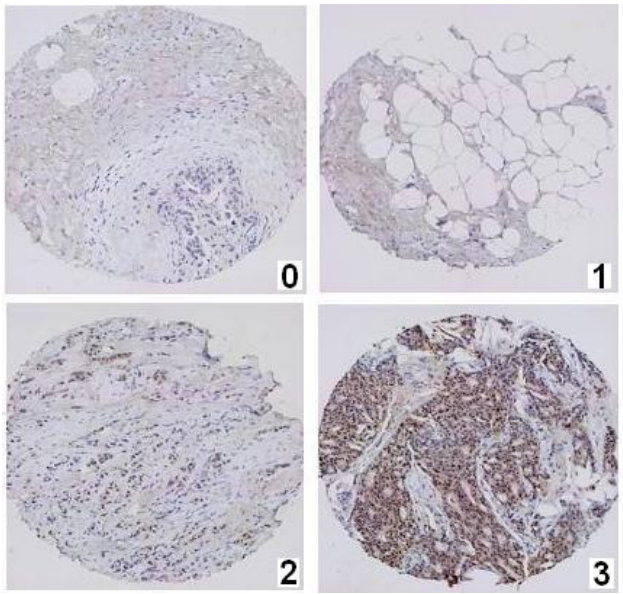
Input image
(Pixel gray values)

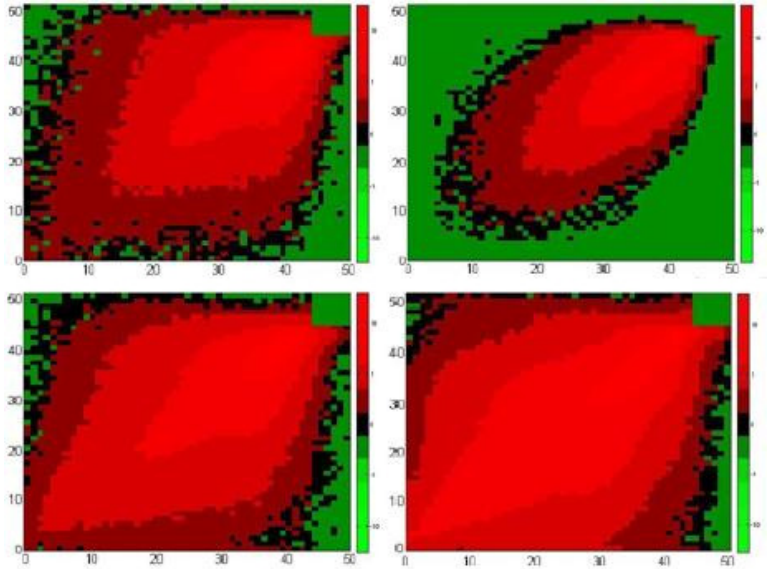
Spatial relationship= $(\nearrow, 1)$

	Dark	Grey	White
White	1	1	0
Grey	1	1	3
Dark	2	2	0

GLCM
(Co-occurrence counts)

GLCM of TMA images (heat map in log scale)





Experiments

- TMA images corresponding to ER (Estrogen Receptor)
  - Available at the Stanford TMA repository (*tma.stanford.edu*)
  - Totally 695 images with 50%-50% for training and test
  - Average test set accuracy over 100 runs
  - *Self repeatability* of pathologists: 75%-84%.

Deep features	**# clusters or leaf nodes**	**Error rate**
—	—	24.79%
K-means clustering	30-60	24.02%
hClustering (various)	[10,40]	**23.46%**
rpForests	30	**23.28%**

Data science@UMassD

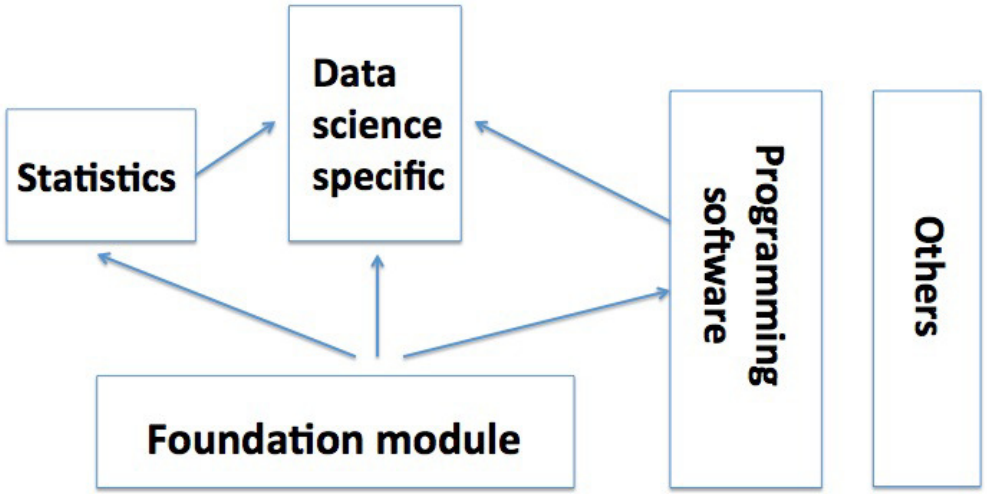
Data Science programs at UMass Dartmouth

- Officially launched in Fall 2015
- BS, MS, and accelerated 5-year BS/MS in Data Science
- Ph.D. in Engineering and Applied Science
  - Tailored to individuals (not discussed here)

♠ What is special about Data Science@UMassD?
  - Strong computation expertise of faculty
  - With a special interest in data visualization
  - Faculty have wide industry experience in data science.

BS in Data Science

- 120 credits to complete in 4 years
  - Foundation module (17)
    - Calculus, linear algebra and computation etc
  - Statistics module (9)
  - Programming and software module (21)
  - Data science specific module (26)
    - Required courses (17), technical electives (9)
  - Others (47).



Foundation module and statistics module

- Foundation module: 17 credits
  - Calculus I and II (8)
  - Discrete mathematics (3)
  - Linear algebra (3)
  - Computation (numerical and matrix, 3)
- Statistics module: 9 credits
  - Elementary statistics (3)
  - Mathematical statistics (3)
  - Probability (3).

Programming and software module

- 21 credits
  - Object-oriented programming I and II (8)
    - Jave programming
  - Algorithms and data structures (3)
  - Software specification and design (4)
  - Database design (3)
  - Social and ethical aspects of software (3).

Data science module

- Required courses: 17 credits
  - Introduction to data science (3)
  - Data visualization (3)
  - Matrix methods for data science (3)
  - Statistical modeling and machine learning (3)
  - Data science senior capstone I and II (5).
- Data science electives: 9 credits
  - 3 courses from a pre-approved list.

MS in Data Science

- Admission requirements
  - Basic programming skills
  - Elementary statistics
  - Quantitative skills (calculus, linear algebra etc)
  - Or makeup upon admission for special cases
- At least 30 credits, or 10 courses
- Expect to complete in 1.5–2 year(s)
  - Depends on preparation and learning pace.

Curriculum of MS in Data Science

- 6 required courses (mostly 3 credits)
- Data science project (3-4)
- 3 elective courses
  - Technical electives
    - e.g., matrix computation, algorithms, data mining etc
  - Or, a course sequence from an applied domain
    - e.g., biology, ecology, psychology, journalism, business etc.

List of required courses

- Introduction to data science (3)
- Data visualization (3)
- Mathematical statistics (3)
- Computational methods (3)
  - Numerical optimization or computational statistics
- Database design (3)
- Statistical modeling and machine learning (3).

Summary

- Examples of data science presented
- Data science and the life cycle discussed
- Deep learning discussed through examples
- Data science program at UMassD introduced.

The end

Thank you!