IST718

Big Data Analytics

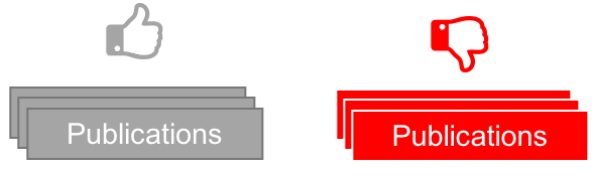
Unit 1: A Course Introduction

About me

- Daniel Acuna, Assistant Professor, iSchool
    - Ph.D. Computer Science, University of Minnesota, Twin Cities
    - Postdoctoral Researcher, Northwestern University & RIC
    - Member of Metaknowledge Research Network, University of Chicago
    - Affiliated to the Center for Computational and Data Science @ SU
- Research interests
    - “Science of science”, human-AI collaboration

Recommendation system

$$u = (1+\alpha)\frac{\sum_{i\in \text{Relevant}} d_i}{|\text{Relevant}|} - \beta\frac{\sum_{j\in \text{Not Relevant}} d_j}{|\text{Not Relevant}|}$$



Activity recognition: in the past

We have very good quality but invasive features:

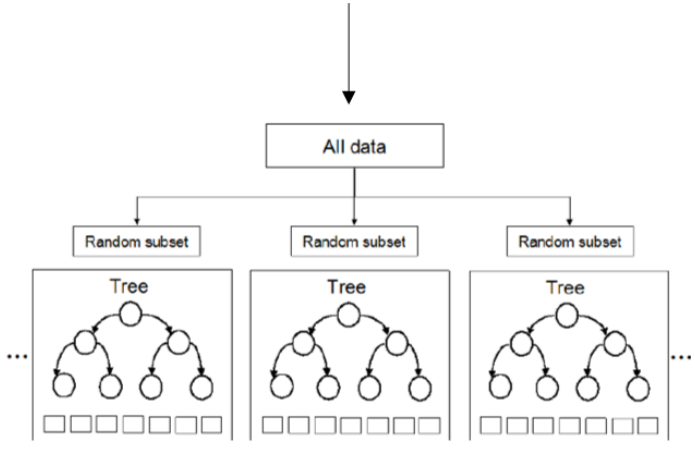
- GPS
- Audio
- WiFi location
- App activity

Predict location and activity

Activity recognition: big data

With big data we might not need such high quality features but just a ton of data

- App open and close activity
- And clock 2 billion seconds of phone activity

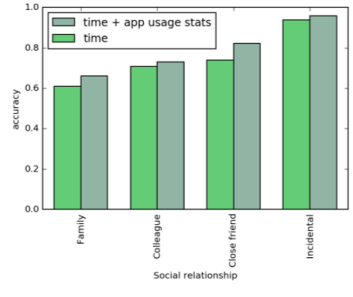


Random forest model

Show Me Your App Usage and I Will Tell Who Your Close Friends Are: Predicting User's Context from Simple Cellphone Activity

Alain Shema
School of Information Studies
Syracuse University, Syracuse, USA
sralain@syr.edu

Daniel E. Acuna
School of Information Studies
Syracuse University, Syracuse, USA
deacuna@syr.edu



About the course

- The main objective of the course is to develop predictive models that are interpretable using modern (very large) datasets
- As secondary goals:
    1. Understand opportunities and challenges of big data
    2. Understand why and how big data enables increases accuracy but lowers interpretability
    3. Understand the statistical bias, variance, and intrisic noise
    4. Understand model fitting, selection, and estimation of generalization error

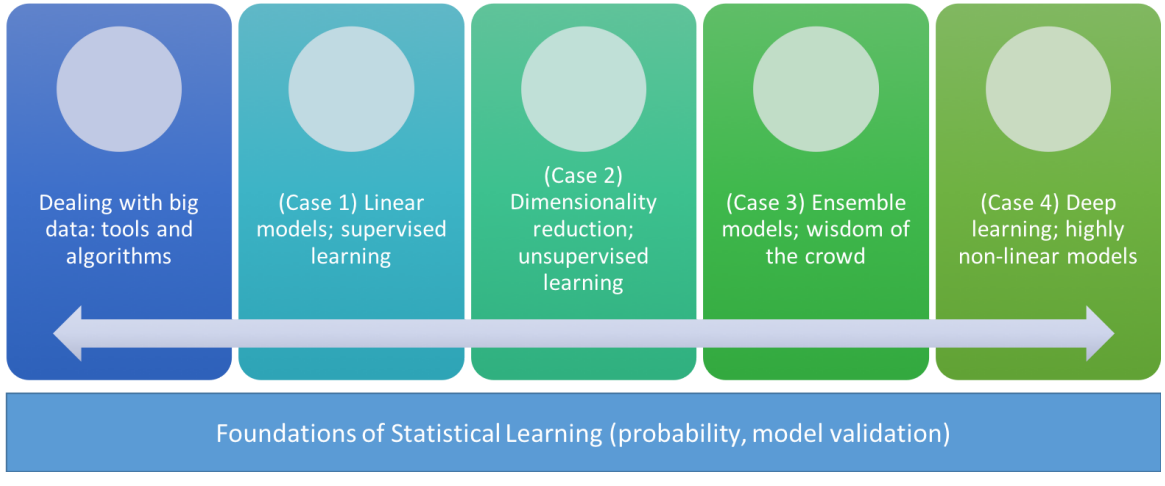
About the course (cont'd)

- Technology is seen as a **means** to an **end**
- We happen to use Python and Apache Spark as our tool but in the future it could be any other technology
- The fundamental principles of this class will apply in such future

About the course (cont'd)

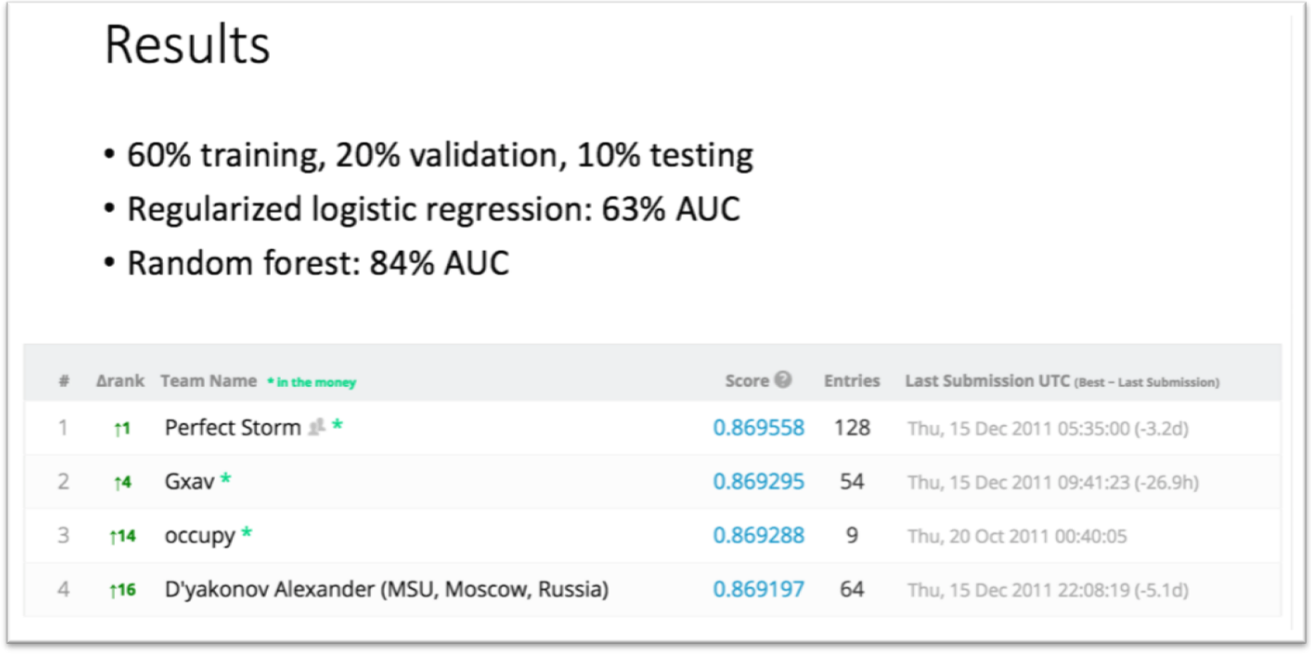
- 1/4 of the course covers prerequisite skills required for big data analytics
    - Python programming, linear algebra, calculus, statistics
- 1/4 of the course covers the Spark and Hadoop, providing skills required to perform big data analytics.
- 1/4 of the course consists of case studies, where you apply your skills and knowledge to real-world applications.
- 1/4 of the course consists of a project, where you work in groups in a real world application

Course roadmap



Course preview (1)

- Slides



Results

- 60% training, 20% validation, 10% testing
- Regularized logistic regression: 63% AUC
- Random forest: 84% AUC

#	Δrank	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	Perfect Storm *	0.869558	128	Thu, 15 Dec 2011 05:35:00 (-3.2d)
2	↑4	Gxav *	0.869295	54	Thu, 15 Dec 2011 09:41:23 (-26.9h)
3	↑14	occupy *	0.869288	9	Thu, 20 Oct 2011 00:40:05
4	↑16	D'yakonov Alexander (MSU, Moscow, Russia)	0.869197	64	Thu, 15 Dec 2011 22:08:19 (-5.1d)

Course preview (2)

- Notebooks and labs to be done on your own

School of Information Studies
Syracuse University

sentiment_analysis (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Markdown Validate Show running Spark jobs

Introduction to Spark ML: An application to Sentiment Analysis

Spark ML

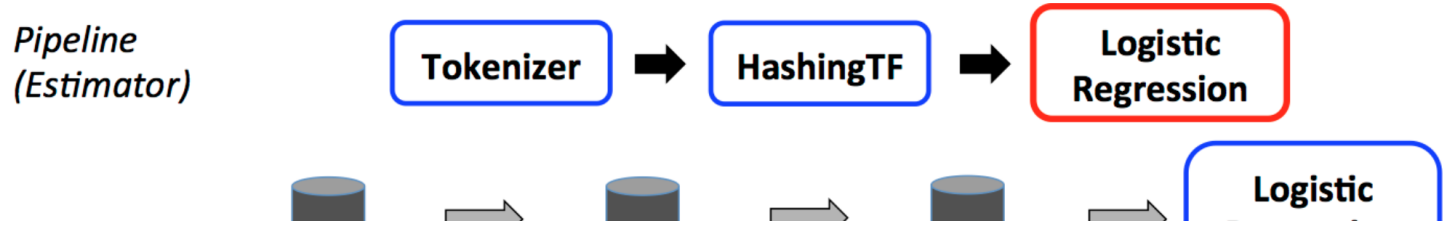
In previous versions of Spark, most Machine Learning funcionality was provided through RDD (Resilient Distributed Datasets). However, to improve performance and communicability of results, Spark developers ported the ML functionality to work almost exclusively with DataFrames. Future releases of Spark will not update the support of ML with RDDs.

In this modern Spark ML approach, there are *Estimators* and *Transformers*. Estimators have some parameters that need to be fit into the data. After fitting, Estimators return Transformers. Tranformers can be applied to dataframes, taking one (or several) columns as input and creating (or several) columns as output.

A *Pipeline* combines several *Tranformers* with a final *Estimator*. The *Pipeline*, therefore, can be fit to the data because the final step of the process (the *Estimator*) is fit to the data. The result of the fitting is a pipelined *Transformer* that takes an input dataframe through all the stages of the Pipeline.

There is a third type of functionality that allows to select features.

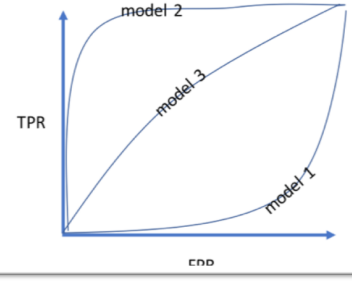
For example, for analyzing text, a typical pipelined estimator is as follows:



Course preview (3)

- Quizes

8 (2 pts) Multiple choice: In terms of area under the curve (AUC), from best to worse, rank the following models based on their ROC curve



a) Model 1, model 2, and model 3
b) Model 2, model 1, and model 3
c) Model 1, model 3, and model 2
d) Model 2, model 3, and model 1

Course preview (4)

- Project

Project Proposal, IST 718
Predicting the trends of cryptocurrencies based on historic data

Team members: Neha Humbal, Shikhar Agrawal, Smit Udani, Suchitra Deekshitula
Our project's aim is to look at different cryptocurrencies prevailing in the market. For example, Bitcoin is a type of digital currency in which encryption techniques are used to regulate the generation of units of currency and verify the transfer of funds, operating independently of a central bank. Since evolution, cryptocurrencies like Bitcoin have grown more than 1000%, trading at $14 in 2013 and currently at $4,500 per Bitcoin. Cryptocurrency's price varies on several factors like price of other cryptocurrencies, news, hype, buying and selling on exchange.
The broader idea is to understand the trends in each of these currencies to strategize investments. The biggest problem while understanding cryptocurrency is price fluctuation. Without understanding factors associated with the fluctuation, it is difficult to make it future predictions. Experts are considering Bitcoins for fundraising. Considering there is neither an optimal nor a perfect trading strategy, it misses a lot of profit opportunities. There are 21 million bitcoins to be mined in total and 15 million have been mined for now which shows inflation levels are controllable. Cryptocurrency has the potential of replacing and acting as an international credit card which could help centralize payments and transactions while reducing fraud. Our objective is to identify and analyze the factors associated with each of the several cryptocurrencies and understand the trends.

Goal: Predict the value of a particular cryptocurrency

Tasks

- **Data Acquisition**: Data has been obtained from Kaggle
- **Data Preparation**: Clean, format and identify relevant attributes from the data
- **Feature Extraction**: Check which attributes contribute better to the model and help in better predictions
- **Modelling and testing**: Create different models, evaluate models based on validation data, test and compare prediction accuracy on test data
- **Visualization**: Present the analysis and trends using appropriate visualization techniques

Expected result: Designing a model to predict future values and seasonal trend in price fluctuation based on historical prices/ market capitalizations of various currencies

Expected problems: Considering the data we have is of stock market past trends, the data could exhibit unexpected drops. We are taking several currencies into account for handling this inconsistency and this would be something we would have to be careful about

Dataset: This dataset has been obtained from Kaggle and has historical data of several cryptocurrencies. Data can be retrieved from:
https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory

Tools: Databricks, PyCharm, Spark, Scikit-learn, Seaborn, Pandas, Numpy, Matplotlib, etc

Models: Linear regression, Polynomial regression, Knn, Svm, Random forest, CART

Criteria: Final model will be selected by comparing the performance of different models using k-fold cross validation and checking accuracy.

Course preview (5)

- In class discussion

What is one of Google's most important assets?
1) Location
2) Cash
3) Data
4) Engineers

Sample of concepts from the course

- Maximum likelihood estimation; mean square error estimation; gradient descent
- Confusion matrix, bias–variance tradeoff, model selection: training, validating, and testing
- Supervised learning, logistic regression, regularized logistic regression, elastic net regularization, model interpretation
- Unsupervised learning, nearest neighbors, dimensionality reduction (Principal Component Analysis, PCA), clustering (k-means)
- "wisdom of the crowd", bagging, random forests, gradient boosting, feature importance
- Neural networks, multilayer perceptron, backpropagation for MLP
- Computation graph, stochastic and mini-batch gradient descent, loss function, convolutional and recurrent networks

Sample of technology

- Python, Numpy, Pandas, Matplotlib
- Apache Spark, Hadoop
- Distributed systems, columnar storage, redundancy

Question: What is data science?

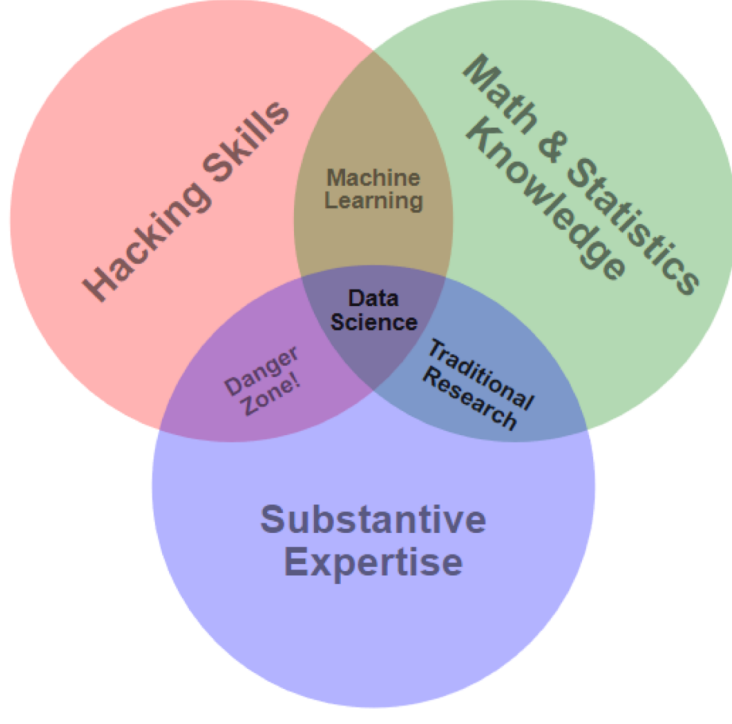
Data Science is

- A combination of disciplines:
  - Information / Computer science
  - Mathematics
  - Statistics
  - Research / Management Science
  - Domain Knowledge

- With the goal of:
  - Using data to make decisions and drive actions

If you do this, then you are a ***data scientist***, and the artifacts you produce are ***data products***.

Data Science Venn Diagram*



*Drew Conway:
https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html (https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html)

"Classic" data science

- Expert is in charge of creating model
- Expert is in charge of providing features that describe or predict new data
- Expert typically produces small models
- Expert typically produces very transparent and easy to understand models

Big Data

What is Big Data?

Data with the following characteristics:

- Data Volume too large to store on a single system.
- Data Velocity too fast for processing by a single computer.
- Data Variety too complex for traditional processing techniques.

These are known as the "three V’s" of big data.

”A new kind of data science”

Classic data science

- Expert is in charge of creating model
- Expert is in charge of providing features that describe or predict new data
- Expert typically produces small models
- Expert typically produces very transparent and easy to understand models

Big-data data science

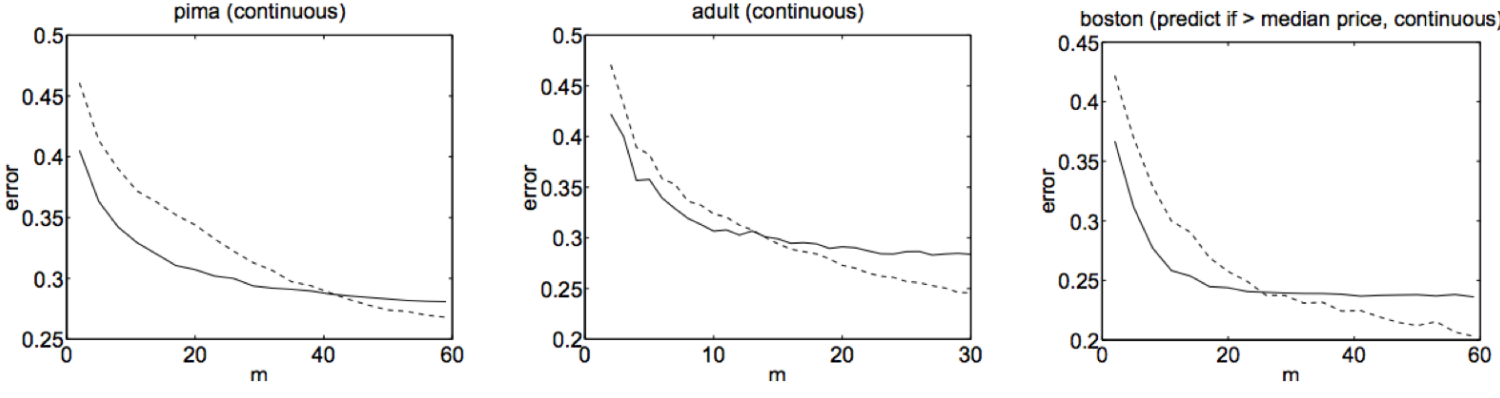
- Generally, there is no expert
- Model used is very general or there is no model at all!
- Features are very low level (e.g., raw transactions vs credit scores)
- Models are very large when fit (e.g., Baidu speech recognition is several terabytes)
- Models are black boxes and almost impossible to understand

On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes

(NIPS 2001)

Andrew Y. Ng
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720

Michael I. Jordan
C.S. Div. & Dept. of Stat.
University of California, Berkeley
Berkeley, CA 94720

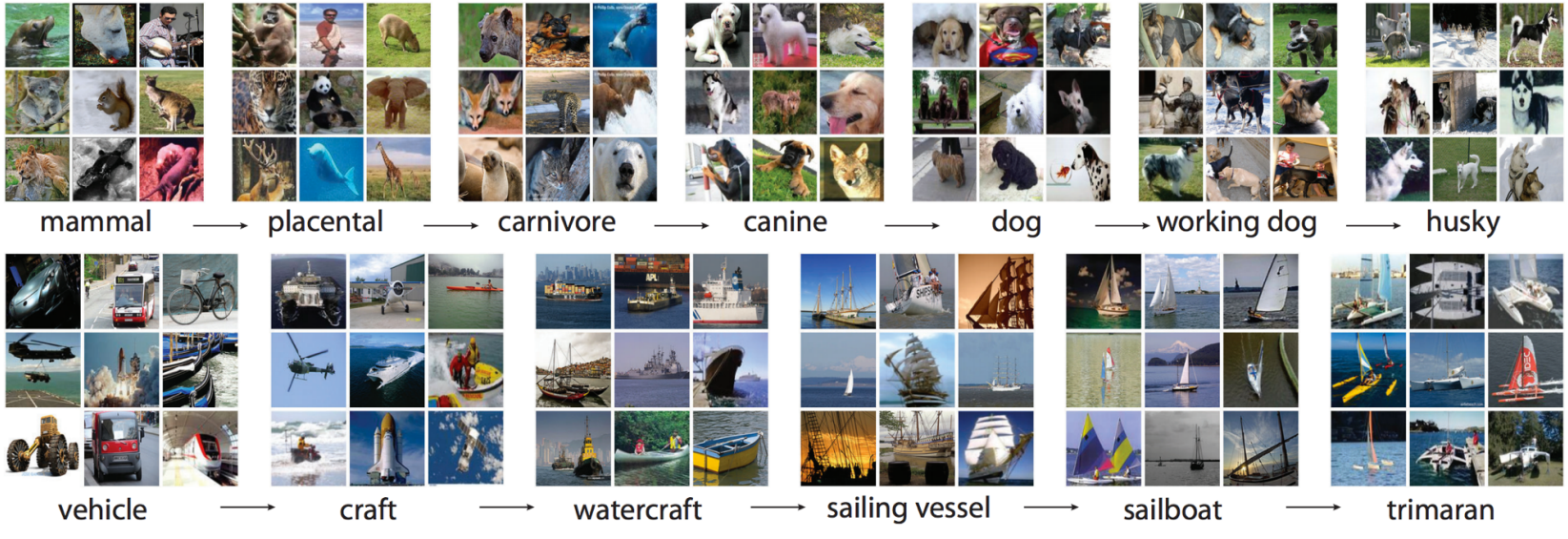


Proposition 1 *Let* h_{Gen} *and* h_{Dis} *be any generative-discriminative pair of classifiers, and* $h_{\mathrm{Gen},\infty}$ *and* $h_{\mathrm{Dis},\infty}$ *be their asymptotic/population versions. Then*[1] $\varepsilon(h_{\mathrm{Dis},\infty}) \leq \varepsilon(h_{\mathrm{Gen},\infty})$.

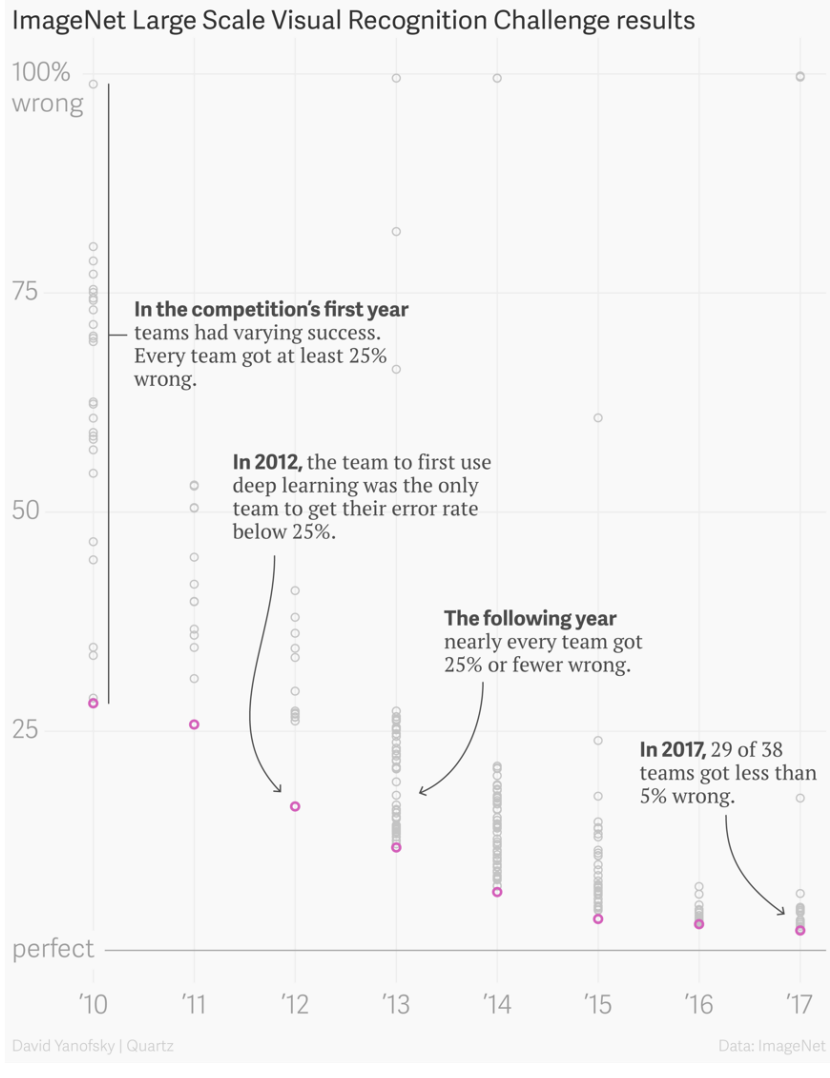
Image recognition

A task that was incredibly challenging and a benchmark for computer algorithms

Image recognition: ImageNet



First models were based on Nearest Neighbor and Naïve Bayes algorithms

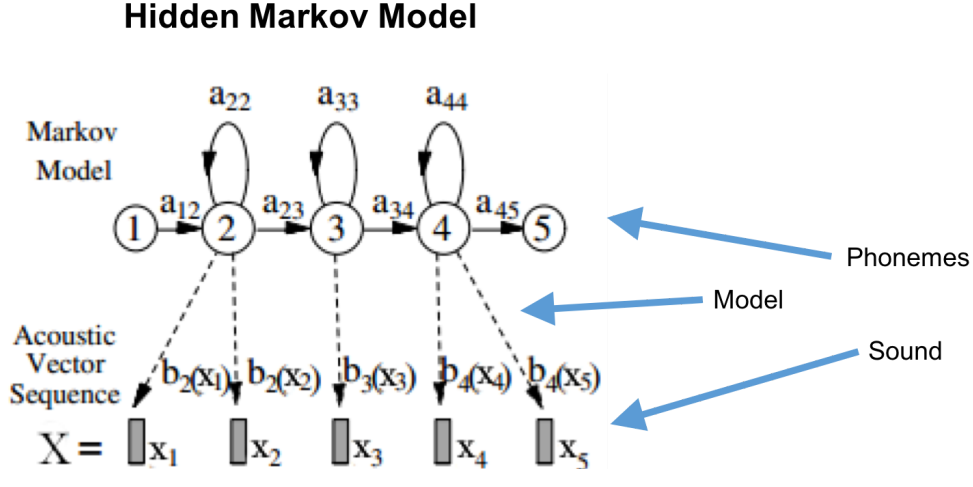

ImageNet Large Scale Visual Recognition Challenge results
100% wrong
75
50
25
perfect
'10
'11
'12
'13
'14
'15
'16
'17
In the competition's first year teams had varying success. Every team got at least 25% wrong.
In 2012, the team to first use deep learning was the only team to get their error rate below 25%.
The following year nearly every team got 25% or fewer wrong.
In 2017, 29 of 38 teams got less than 5% wrong.
David Yanofsky | Quartz
Data: ImageNet

State of the art ImageNet model

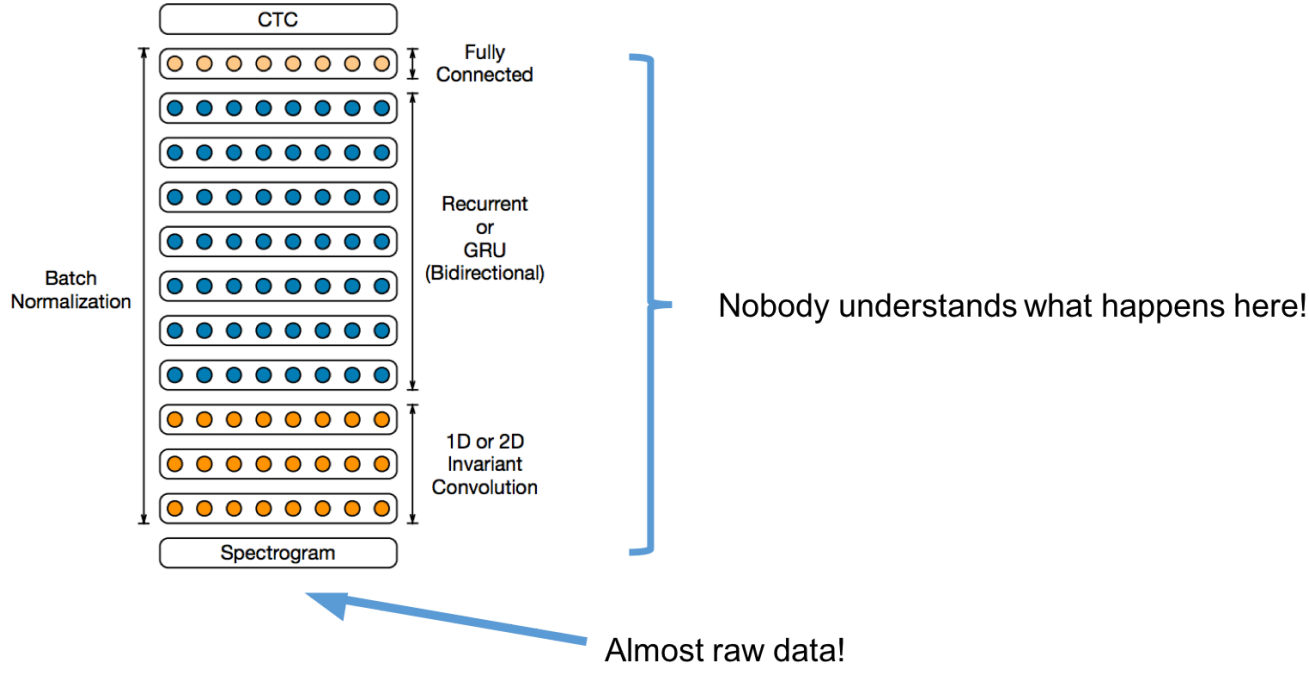
GoogleLetNet neural network model

Team	Year	Place	Error (top-5)	Uses external data
SuperVision	2012	1st	16.4%	no
SuperVision	2012	1st	15.3%	Imagenet 22k
Clarifai	2013	1st	11.7%	no
Clarifai	2013	1st	11.2%	Imagenet 22k
MSRA	2014	3rd	7.35%	no
VGG	2014	2nd	7.32%	no
GoogLeNet	2014	1st	6.67%	no

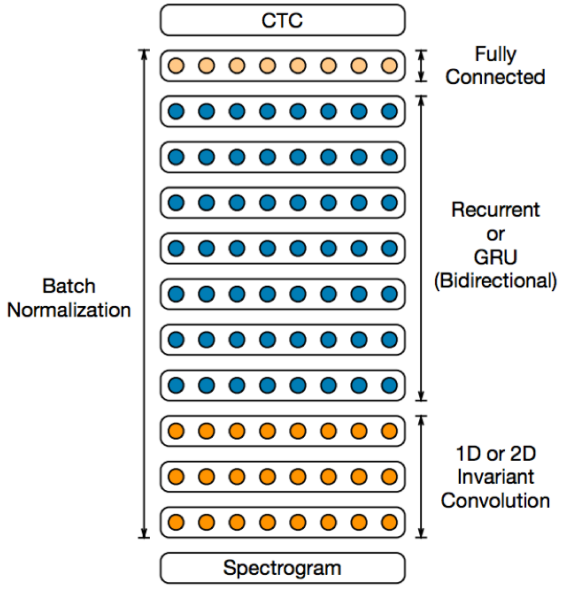
Speech recognition: Generative model



Speech recognition: Deep Speech 2 (Baidu) (1)

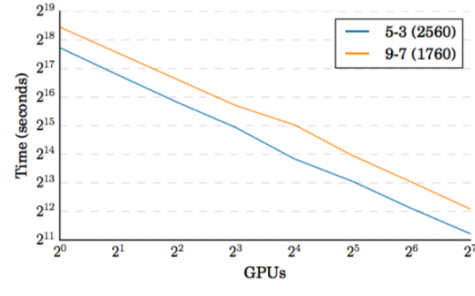


Speech recognition: Deep Speech 2 (Baidu) (2)

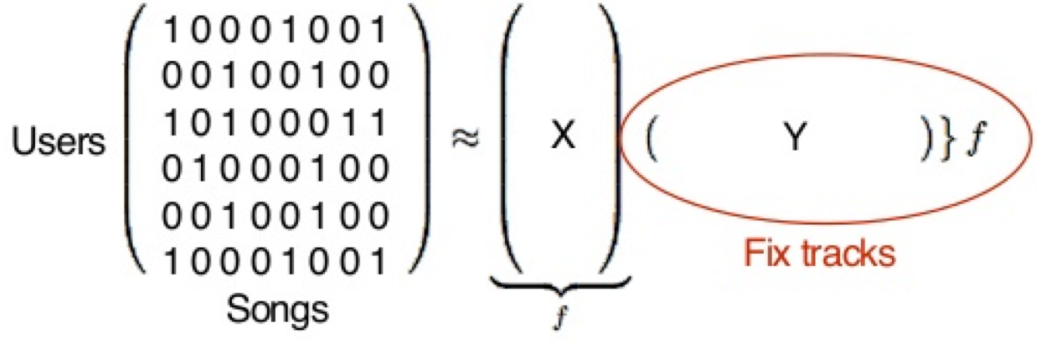


Read Speech			
Test set	DS1	DS2	Human
WSJ eval'92	4.94	3.60	5.03
WSJ eval'93	6.94	4.98	8.08
LibriSpeech test-clean	7.89	5.33	5.83
LibriSpeech test-other	21.74	13.25	12.69

Model size	Model type	Regular Dev	Noisy Dev
18×10^6	GRU	10.59	21.38
38×10^6	GRU	9.06	17.07
70×10^6	GRU	8.54	15.98
70×10^6	RNN	8.44	15.09
100×10^6	GRU	7.78	14.17
100×10^6	RNN	7.73	13.06



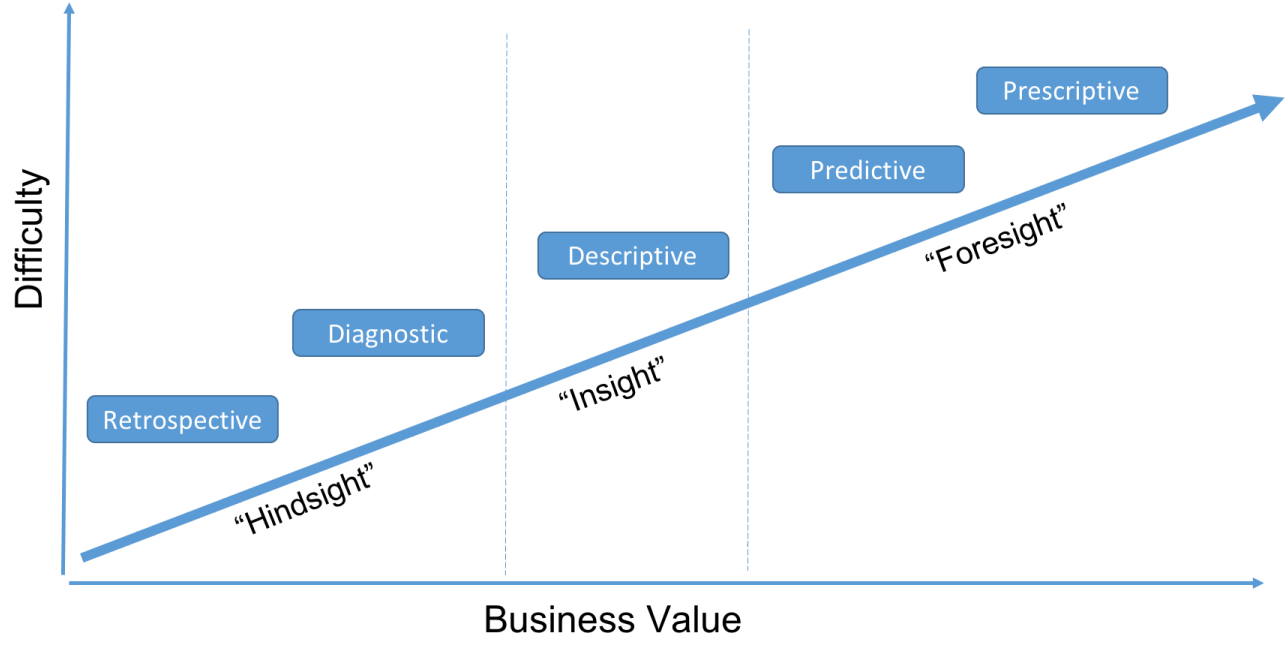
Music: Implicit-feedback ALS (Spotify)



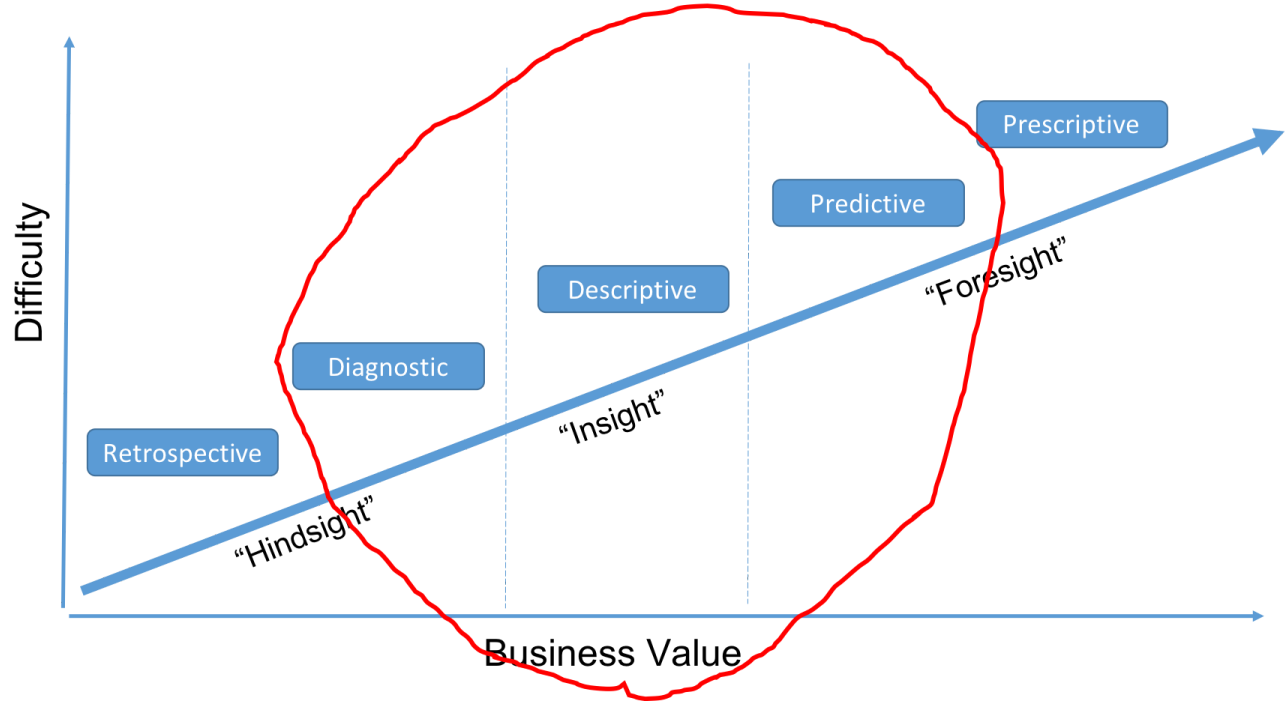
$$\min_{x,y} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i - \beta_u - \beta_i)^2 + \lambda(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2)$$

- p_{ui} : 1 if user u streamed track i else 0
- $c_{ui} = 1 + \alpha r_{ui}$
- x_u = user $u's$ latent factor vector
- x_i = item $i's$ latent factor vector
- β_u = bias for user u
- β_i = bias for item i
- λ = regularization parameter

What will this course cover? (1)



What will this course cover? (2)



Questions?

- Contact me at deacuna@syr.edu
- Visit office hours often