Data Wrangling For Kaggle Data Science Competitions

An Etude

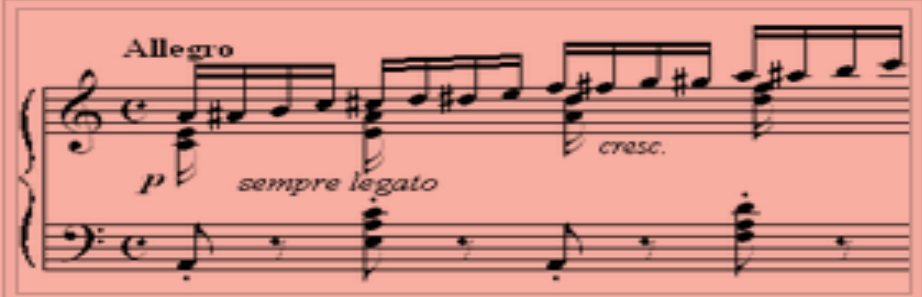
April 9, 2014

@ksankar // doubleclix.wordpress.com

etude

An **étude** (/ ˈeɪtjuːd /; French pronunciation: [eˈtyd], a French word meaning *study*) is an instrumental musical composition, usually short and of considerable difficulty, usually designed to provide practice material for perfecting a particular musical skill. The

We will focus on "short", "acquiring skill" & "having fun" !



Frédéric Chopin's Étude Op. 10, No. 2: a rapid chromatic scale in the right hand is used to develop the weaker fingers of the right hand. Most études are written to perfect a particular technical skill.

http://en.wikipedia.org/wiki/%C3%89tude, http://www.etudesdemarche.net/articles/etudes-sectorielles.htm, http://upload.wikimedia.org/wikipedia/commons/2/26/La_Cour_du_Palais_des_%C3%A9tudes_de_l%E2%80%99%C3%89cole_des_beaux-arts.jpg

Agenda [1 of 3]

- Intro, Goals, Logistics, Setup [10] [1:20-1:30)
- Anatomy of a Kaggle Competition [60] [1:30-2:30)
  - *Competition Mechanics*
  - *Register, download data, create sub directories*
  - *Trial Run : Submit Titanic*
- Algorithms for the Amateur Data Scientist [20] [2:30-2:50)
  - *Algorithms, Tools & frameworks in perspective*
  - *"Folk Wisdom"*
- Break (30 Min) [2:50-3:10)

Agenda [2 of 3]

- Model Evaluation & Interpretation [30] [3:10-3:40)
  - *Confusion Matrix, ROC Graph*
- Session 1 : The Art of a Competition – DS London + Scikit-learn[*30 min*] [*3:40-4:10*)
  - *Dataset Organization*
  - *Analytics Walkthrough*
    - *Algorithms – CART, RF, SVM*
    - *Feature Extraction*
  - *Hands-on Analytics programming of the challenge*
  - *Submit entry*
- Session 2 : The Art of a Competition – ASUS,PAKDD 2014 [*20 min*] [*4:10-4:30*)
  - *Dataset Organization*
  - *Analytics Walkthrough, Transformations*

Agenda [3 of 3]

- Questions, Discussions & Slack [*10 min*] [*4:30-4:40*]
- Schedule
  - *12:20 – 1:20 : Lunch*
  - *1:20 – 4:40 : Tutorial (1:20–2:50;2:50–3:10:Break;3:10–4:40)*

Overload Warning … There is enough material for a week's training … which is good & bad !
Read thru at your pace, refer, ponder & internalize

Goals & Assumptions

- Goals:
  - *Get familiar with the mechanics of Data Science Competitions*
  - *Explore the intersection of Algorithms, Data, Intelligence, Inference & Results*
  - *Discuss Data Science Horse Sense ;o)*
- At the end of the tutorial you should have :
  - *Submitted entries for 3 competitions*
  - *Applied Algorithms on Kaggle Data*
    - *CART, RF*
    - *Linear Regression*
    - *SVM*
  - *Explored Data, have a good understanding of the Analytics Pipeline viz. collect–store–transform–model–reason–deploy–visualize–recommend–infer–explore*
  - *Knowledge of Model Evaluation*
    - *Cross Validation, ROC Curves*

Close Encounters

- 1st
  - *This Tutorial*
- 2nd
  - *Do More Hands-on Walkthrough*
- 3nd
  - *Listen To Lectures*
  - *More competitions …*

Warm-up

Setup

- Install anaconda
- update (if needed)
  - conda update conda
  - conda update ipython
  - condo update python
- conda install pydot
- ipython notebook –-plab=inline
- Go to ipython tab in the browser
- Import packages & print version #

First setup check

In [2]:
```
import numpy
print 'numpy:', numpy.__version__

import scipy
print 'scipy:', scipy.__version__

import pandas
print 'pandas: ',pandas.__version__

import matplotlib
print 'matplotlib:', matplotlib.__version__

import sklearn
print 'scikit-learn:', sklearn.__version__
```

```
numpy: 1.8.0
scipy: 0.13.3
pandas:  0.13.1
matplotlib: 1.3.1
scikit-learn: 0.14.1
```

Tutorial Materials

- Github : https://github.com/xsankar/freezing-bear
- Clone or download zip
- Open terminal
- cd ~/freezing-bear/notebooks
- ipython notebook –pylab=inline
- Click on ipython dashboard
- Just look thru the ipython notebooks

Tutorial Data

- Setup an account in Kaggle (www.kaggle.com)
- We will be using the data from 3 Kaggle competitions

① Titanic: Machine Learning from Disaster

- Download data from http://www.kaggle.com/c/titanic-gettingStarted
- Directory ~/freezing-bear/notebooks/titanic

② Data Science London + Scikit-learn

- Download data from http://www.kaggle.com/c/data-science-london-scikit-learn
- Directory ~/freezing-bear/notebooks/dsl

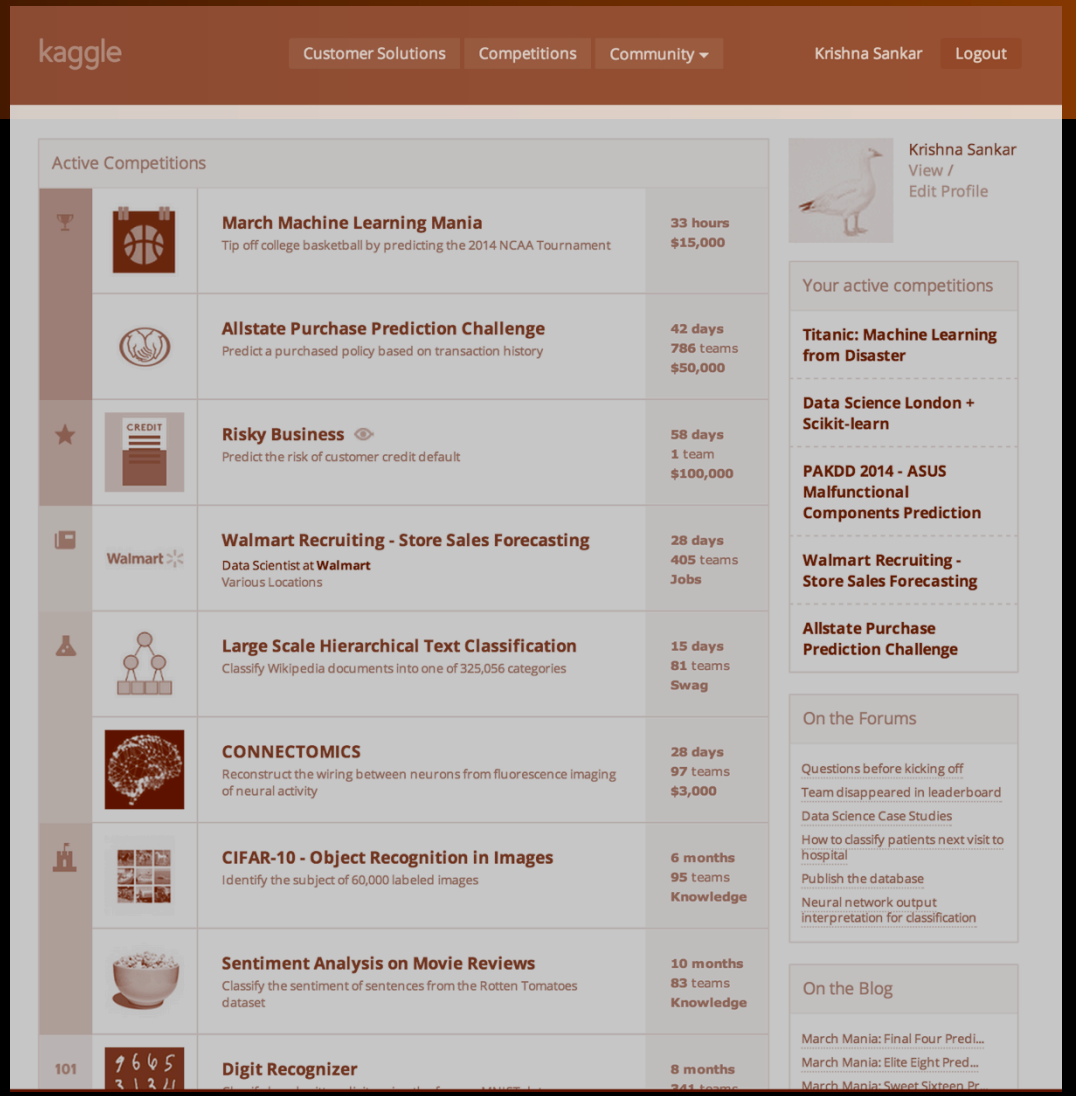
③ PAKDD 2014 - ASUS Malfunctional Components Prediction

- Download data from http://www.kaggle.com/c/pakdd-cup-2014
- Directory ~/freezing-bear/notebooks/asus

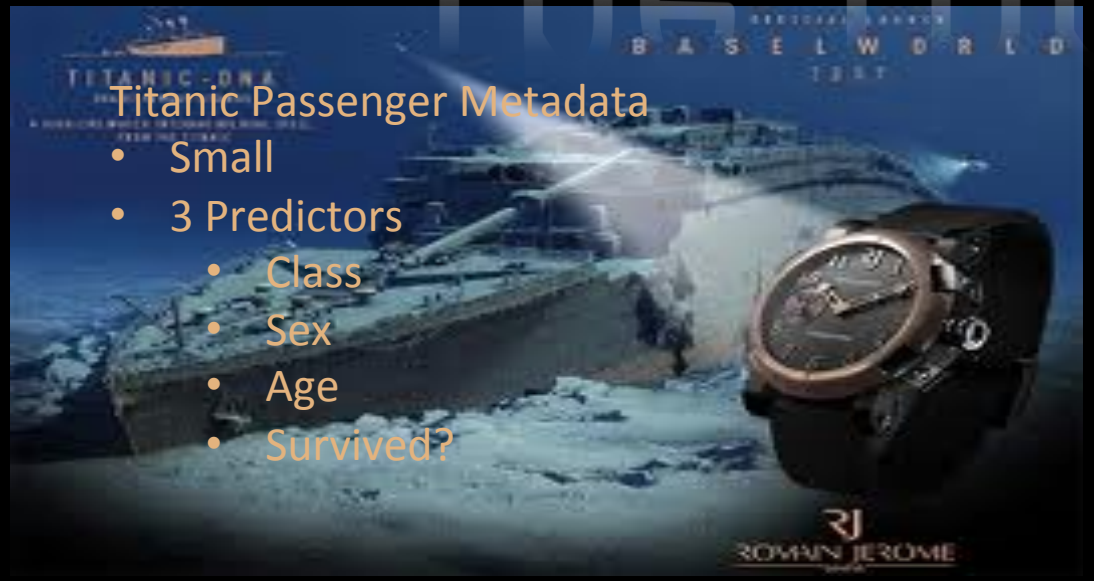
Anatomy Of a Kaggle Competition

Kaggle Data Science Competitions

- Hosts Data Science Competitions
- Competition Attributes:
  - *Dataset*
    - *Train*
    - *Test (Submission)*
    - *Final Evaluation Data Set (We don't see)*
  - *Rules*
  - *Time boxed*
  - *Leaderboard*
  - *Evaluation function*
  - *Discussion Forum*
  - *Private or Public*



The Three Datasets



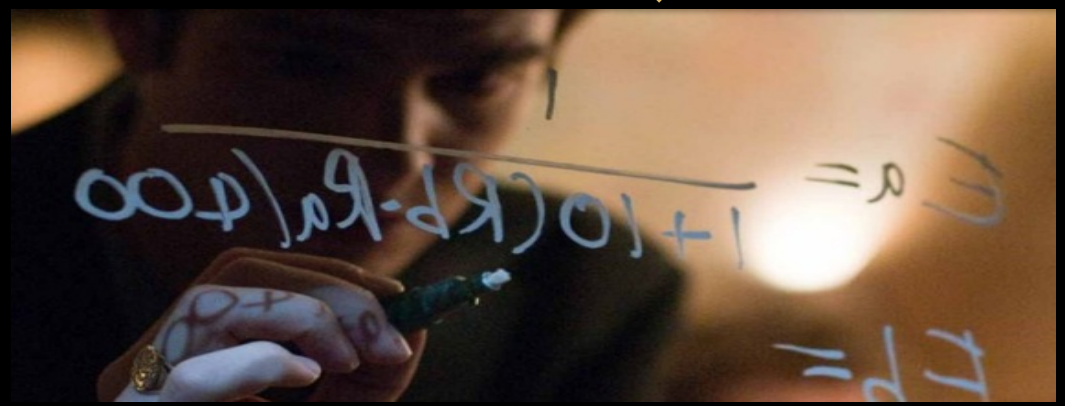
Titanic Passenger Metadata

- Small
- 3 Predictors
  - Class
  - Sex
  - Age
  - Survived?

Data Science London + Sci-Kit learn Competition

PAKDD 2014

Predict Component Failures … ASUS



http://www.ohgizmo.com/2007/03/21/romain-jerome-titanic

http://flyhigh-by-learnonline.blogspot.com/2009/12/at-movies-sherlock-holmes-2009.html

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, Mrs. John Bradley (Florence I	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101	7.925		S
5	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhe	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
886	885	0	3	Sutehall, Mr. Henry Jr	male	25	0	0	SOTON/OQ 392	7.05		S
887	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39	0	5	382652	29.125		Q
888	887	0	2	Montvila, Rev. Juozas	male	27	0	0	211536	13		S
889	888	1	1	Graham, Miss. Margaret Edith	female	19	0	0	112053	30	B42	S
890	889	0	3	Johnston, Miss. Catherine Helen "Carri	female		1	2	W./C. 6607	23.45		S
891	890	1	1	Behr, Mr. Karl Howell	male	26	0	0	111369	30	C148	C
892	891	0	3	Dooley, Mr. Patrick	male	32	0	0	370376	7.75		Q

Train.csv
Taken from Titanic Passenger Manifest

Titanic Passenger Metadata

- Small
- 3 Predictors
  - Class
  - Sex
  - Age
  - Survived?

Variable	Description
Survived	0-No, 1=yes
Pclass	Passenger Class (1^{st},2^{nd},3^{rd})
Sibsp	Number of Siblings/Spouses Aboard
Parch	Number of Parents/Children Aboard
Embarked	Port of Embarkation ○ C = Cherbourg ○ Q = Queenstown ○ S = Southampton

Submission

	A	B	C	D	E	F	G	H	I	J	K
1	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	892	3	Kelly, Mr. Jar	male	34.5	0	0	330911	7.8292		Q
3	893	3	Wilkes, Mrs.	female	47	1	0	363272	7		S
4	894	2	Myles, Mr. Tl	male	62	0	0	240276	9.6875		Q
5	895	3	Wirz, Mr. Alb	male	27	0	0	315154	8.6625		S
6	896	3	Hirvonen, Mı	female	22	1	1	3101298	12.2875		S
7	897	3	Svensson, M	male	14	0	0	7538	9.225		S
415	1305	3	Spector, Mr.	male		0	0	A.5. 3236	8.05		S
416	1306	1	Oliva y Ocan	female	39	0	0	PC 17758	108.9	C105	C
417	1307	3	Saether, Mr.	male	38.5	0	0	SOTON/O.Q.	7.25		S
418	1308	3	Ware, Mr. Fr	male		0	0	359309	8.05		S
419	1309	3	Peter, Maste	male		1	1	2668	22.3583		C

Test.csv

- 418 lines; 1st column should have 0 or 1 in each line
- Evaluation:
  - *% correctly predicted*

Approach

- This is a classification problem - 0 or 1
- Troll the forums !
- Opportunity for us to try different algorithms & compare them
  - *Simple Model*
  - *CART[Classification & Regression Tree]*
    - *Greedy, top-down binary, recursive partitioning that divides feature space into sets of disjoint rectangular regions*
  - *RF*
    - *Different parameters*
  - *SVM*
    - *Multiple kernels*
    - *Table the results*
- Use cross validation to predict our model performance & correlate with what Kaggle says

http://www.dabi.temple.edu/~hbling/8590.002/Montillo_RandomForests_4-2-2009.pdf

Simple Model – Our First Submission

Refer to iPython notebook <1-Intro-To-Kaggle>
at https://github.com/xsankar/freezing-bear

- #1 : Simple Model (M=survived)

1241	new	_T@y_	0.33493	1	Fri, 21 Mar 2014 11:08:41
1242	new	**Krishna Sankar**	0.23445	1	Sun, 23 Mar 2014 20:07:41

Your Best Entry
Congratulations on making your first submission!

- #2 : Simple Model (F=survived)

1059	new	Vasya Eriklintsev	0.76555	2	Sun, 23 Mar 2014 16:58:15
1060	new	RuudAlex	0.76555	6	Sun, 23 Mar 2014 19:26:42
1061	new	Ggglygy	0.76555	2	Sun, 23 Mar 2014 19:59:50
1062	new	**Krishna Sankar**	0.76555	2	Sun, 23 Mar 2014 20:19:23

Your Best Entry
You improved on your best score by 0.53110.
You just moved up 180 positions on the leaderboard.

https://www.kaggle.com/c/titanic-gettingStarted/details/getting-started-with-python

#3 : Simple CART Model

- CART (Classification & Regression Tree)

1062	new	RuudAlex	0.76555	6	Sun, 23 Mar 2014 19:26:42
1063	new	**Krishna Sankar**	0.76555	3	Sun, 23 Mar 2014 23:37:04 (-3.3h)

Your Best Entry
Your submission scored **0.75120**, which is not an improvement of your best score. Keep trying!

http://scikit-learn.org/stable/modules/tree.html

#4 : Random Forest Model

1063	new	Krishna Sankar	0.76555	4	Sun, 23 Mar 2014 23:57:58 (-3.6h)

Your Best Entry
Your submission scored **0.75598**, which is not an improvement of your best score. Keep trying!

- https://www.kaggle.com/wiki/GettingStartedWithPythonForDataScience
- *Chris Clark* *http://blog.kaggle.com/2012/07/02/up-and-running-with-python-my-first-kaggle-entry/*
  - https://www.kaggle.com/c/titanic-gettingStarted/details/getting-started-with-random-forests
  - https://github.com/RahulShivkumar/Titanic-Kaggle/blob/master/titanic.py

#5 : SVM

- Multiple Kernels
- kernel = ‘rbf’ #Radial Basis Function

1063	new	**Krishna Sankar**	0.76555	6	Mon, 24 Mar 2014 00:14:14 (-3.9h)

Your Best Entry
Your submission scored **0.74641**, which is not an improvement of your best score. Keep trying!

- Kernel = ‘sigmoid’

1063	new	**Krishna Sankar**	0.76555	8	Mon, 24 Mar 2014 00:23:26 (-4.1h)

Your Best Entry
Your submission scored **0.62679**, which is not an improvement of your best score. Keep trying!

- agconti's blog - Ultimate Titanic !
- http://fastly.kaggle.net/c/titanic-gettingStarted/forums/t/5105/ipython-notebook-tutorial-for-titanic-machine-learning-from-disaster/29713

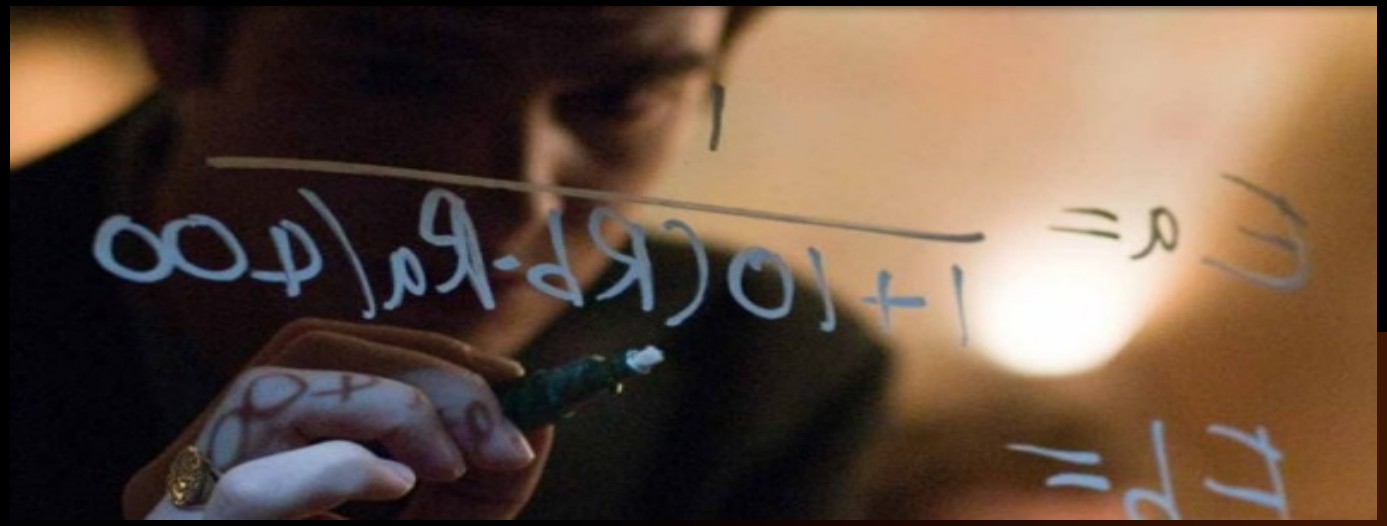
Feature Engineering - Homework

- Add attribute : Age
  - *In train 714/891 have age; in test 332/418 have age*
  - *Missing values can be just Mean Age of all passengers*
  - *We could be more precise and calculate Mean Age based on Title (Ms,Mrs,Master et al)*
  - *Box plot age*
- Add attribute : Mother, Family size et al
- Feature engineering ideas
  - *http://www.kaggle.com/c/titanic-gettingStarted/forums/t/6699/sharing-experiences-about-data-munging-and-classification-steps-with-python*
- More ideas at http://statsguys.wordpress.com/2014/01/11/data-analytics-for-beginners-pt-2/
- And https://github.com/wehrley/wehrley.github.io/blob/master/SOUPTONUTS.md
- Also Learning scikit-learn: Machine Learning in Python, By: Raúl Garreta; Guillermo Moncecchi, Publisher: Packt Publishing has more ideas on feature selection et al

What does it mean ? Let us ponder ….

Think different

- We have a training data set representing a domain
  - *We reason over the dataset & develop a model to predict outcomes*
- How good is our prediction when it comes to real life scenarios ?
- The assumption is that the dataset is taken at random
  - *Or Is it ? Is there a Sampling Bias ?*
  - *i.i.d ? Independent ? Identically Distributed ?*
  - *What about homoscedasticity ? Do they have the same finite variance ?*
- Can we assure that another dataset (from the same domain) will give us the same result ?
- Will our model & it's parameters remain the same if we get another data set ?
- How can we evaluate our model ?
- How can we select the right parameters for a selected model ?



2:30

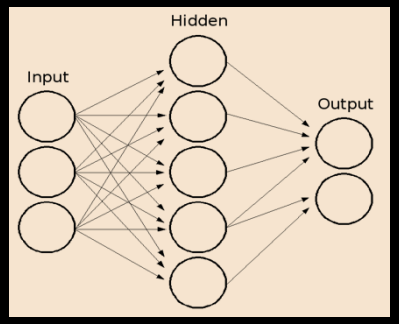
Algorithms ! The Most Massively useful thing an Amateur Data Scientist can have ···

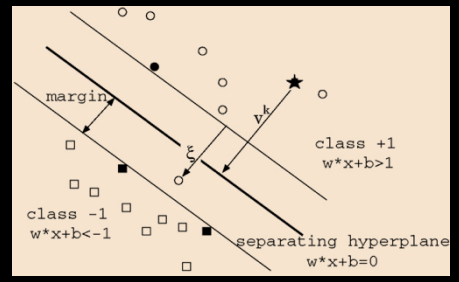
Algorithms for the Amateur Data Scientist

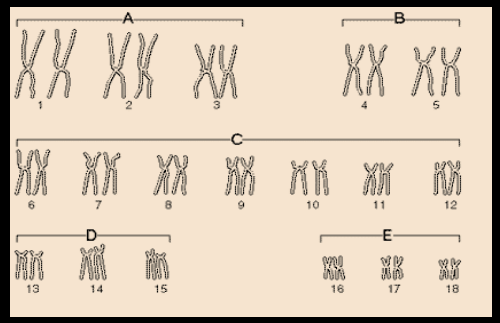
"A towel is about the most massively useful thing an interstellar hitchhiker can have ··· any man who can hitch the length and breadth of the Galaxy, rough it ··· win through, and still know where his towel is, is clearly a man to be reckoned with."

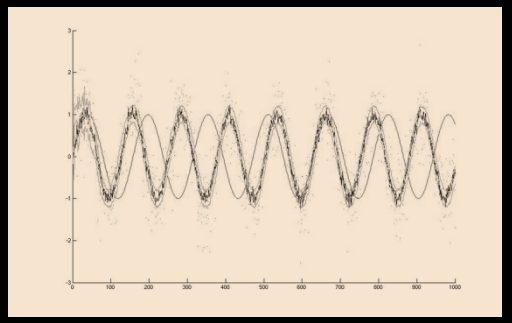
– From The Hitchhiker's Guide to the Galaxy, by Douglas Adams.

Data Scientists apply different techniques









- Support Vector Machine
- adaBoost
- Bayesian Networks
- Decision Trees
- Ensemble Methods
- Random Forest
- Logistic Regression

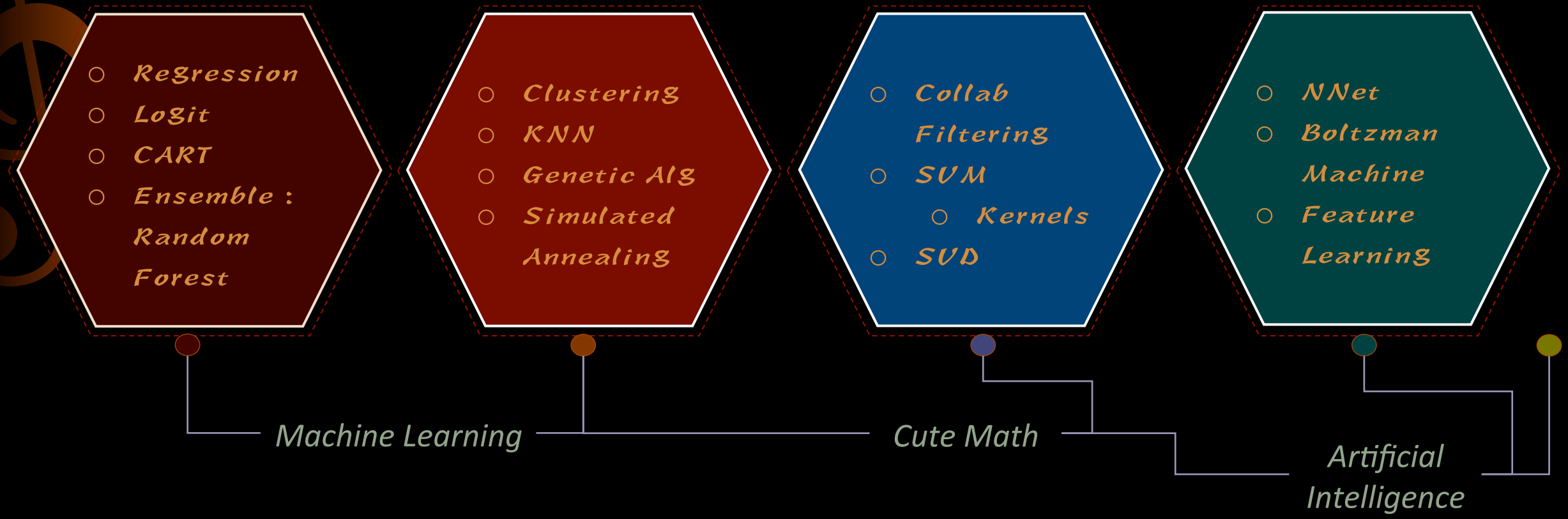
- Genetic Algorithms
- Monte Carlo Methods
- Principal Component Analysis
- Kalman Filter
- Evolutionary Fuzzy Modelling
- Neural Networks

Quora

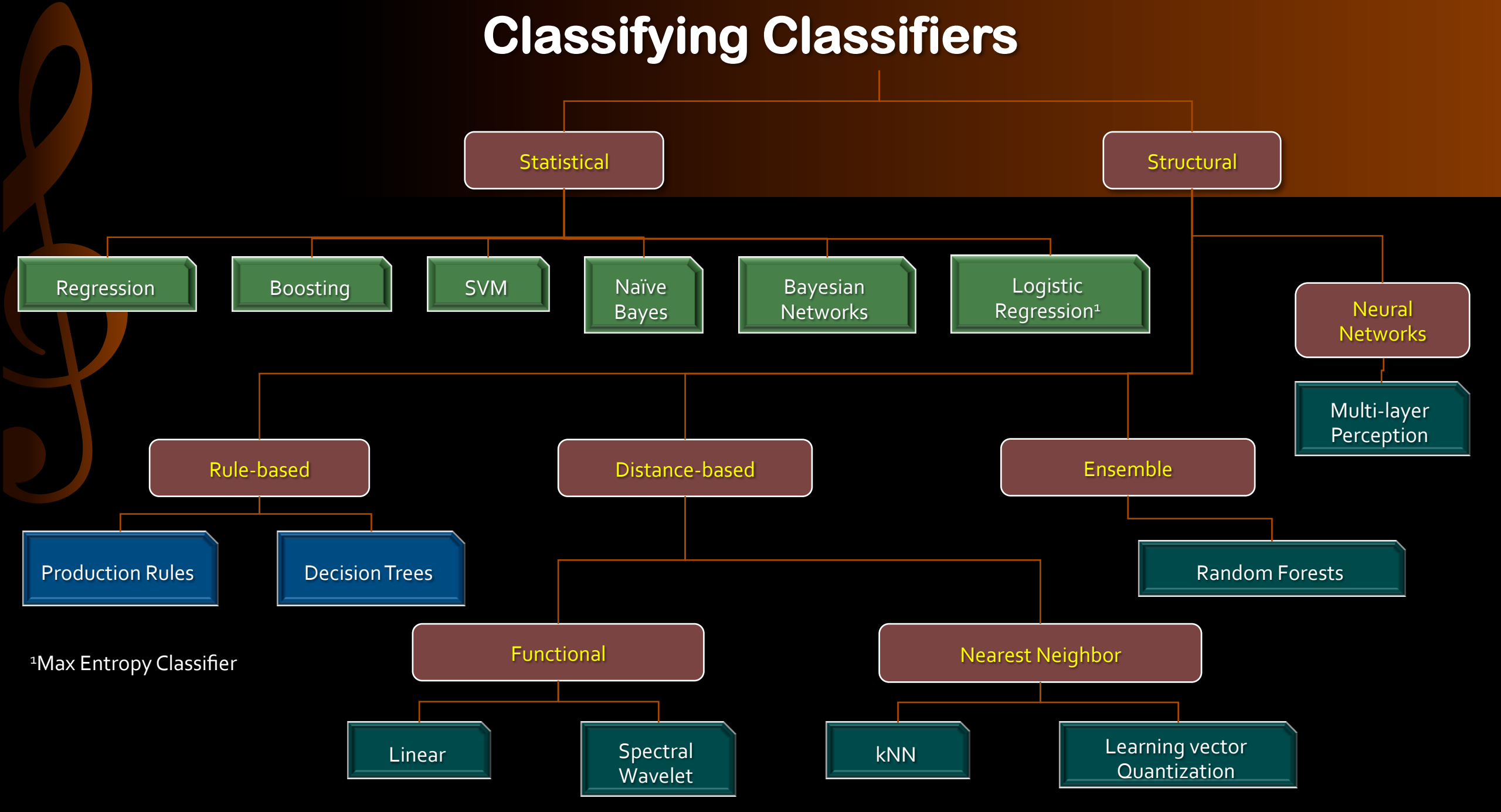
- http://www.quora.com/What-are-the-top-10-data-mining-or-machine-learning-algorithms

Ref: Anthony's Kaggle Presentation

Algorithm spectrum

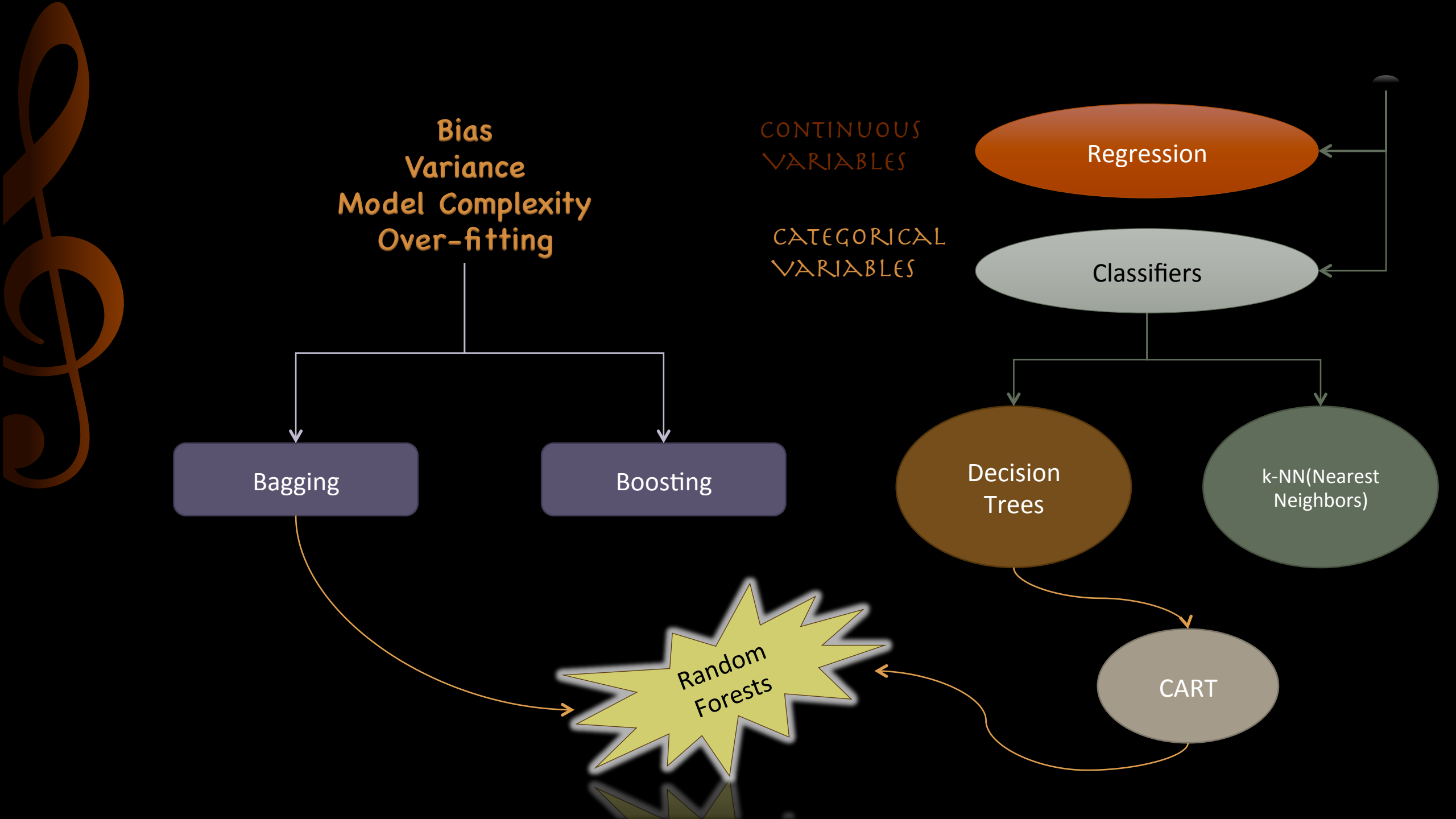


Classifying Classifiers



[1]Max Entropy Classifier

Ref: Algorithms of the Intelligent Web, Marmanis & Babenko


Bias
Variance
Model Complexity
Over-fitting
Bagging
Boosting
CONTINUOUS VARIABLES
CATEGORICAL VARIABLES
Regression
Classifiers
Decision Trees
k-NN(Nearest Neighbors)
CART
Random Forests

review articles

DOI:10.1145/2347736.2347755

Tapping into the "folk knowledge" needed to advance machine learning applications.

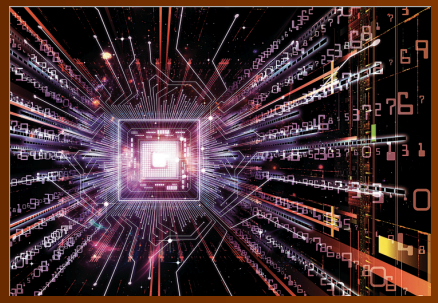
BY PEDRO DOMINGOS

A Few Useful Things to Know About Machine Learning

is needed to successfully develop machine learning applications is not readily available in them. As a result, many machine learning projects take much longer than necessary or wind

Data Science "folk knowledge"

Data Science “folk knowledge” (1 of A)



- "If you torture the data long enough, it will confess to anything." – Hal Varian, Computer Mediated Transactions
- Learning = Representation + Evaluation + Optimization
- It’s Generalization that counts
  - *The fundamental goal of machine learning is to generalize beyond the examples in the training set*
- Data alone is not enough
  - *Induction not deduction – Every learner should embody some knowledge or assumptions beyond the data it is given in order to generalize beyond it*
- Machine Learning is not magic – one cannot get something from nothing
  - *In order to infer, one needs the knobs & the dials*
  - *One also needs a rich expressive dataset*

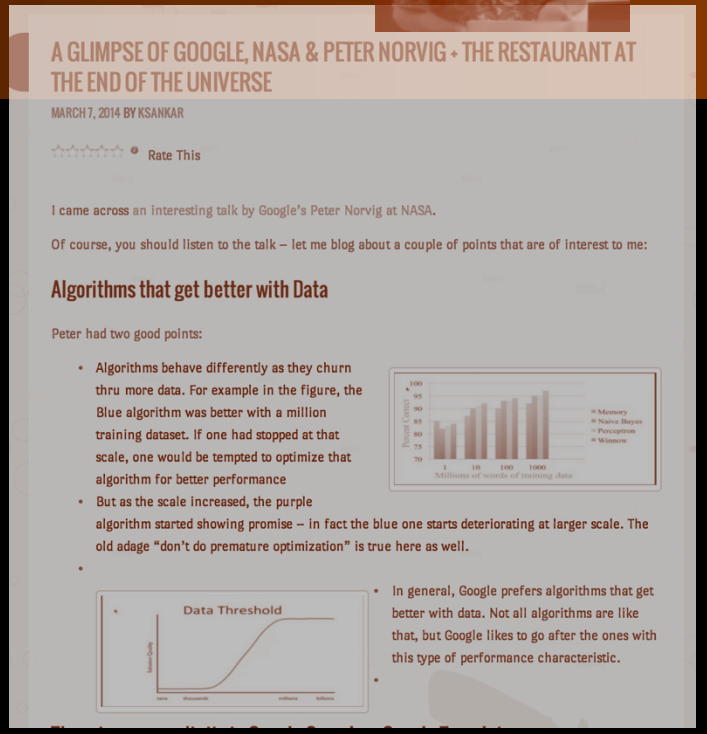
A few useful things to know about machine learning - by Pedro Domingos
http://dl.acm.org/citation.cfm?id=2347755

Data Science “folk knowledge” (2 of A)

- Over fitting has many faces
  - *Bias – Model not strong enough. So the learner has the tendency to learn the same wrong things*
  - *Variance – Learning too much from one dataset; model will fall apart (ie much less accurate) on a different dataset*
  - *Sampling Bias*
- Intuition Fails in high Dimensions –Bellman
  - *Blessing of non–conformity & lower effective dimension; many applications have examples not uniformly spread but concentrated near a lower dimensional manifold eg. Space of digits is much smaller then the space of images*
- Theoretical Guarantees are not What they seem
  - *One of the major developments o f recent decades has been the realization that we can have guarantees on the results of induction, particularly if we are willing to settle for probabilistic guarantees.*
- Feature engineering is the Key

A few useful things to know about machine learning - by Pedro Domingos
http://dl.acm.org/citation.cfm?id=2347755

Data Science "folk knowledge" (3 of A)

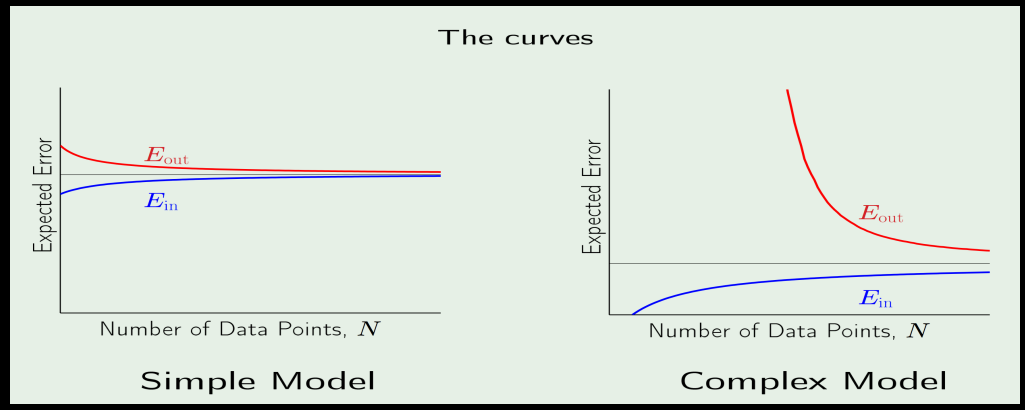


- More Data Beats a Cleverer Algorithm
  - *Or conversely select algorithms that improve with data*
  - *Don't optimize prematurely without getting more data*
- Learn many models, not Just One
  - *Ensembles ! – Change the hypothesis space*
  - *Netflix prize*
  - *E.g. Bagging, Boosting, Stacking*
- Simplicity Does not necessarily imply Accuracy
- Representable Does not imply Learnable
  - *Just because a function can be represented does not mean it can be learned*
- Correlation Does not imply Causation

- http://doubleclix.wordpress.com/2014/03/07/a-glimpse-of-google-nasa-peter-norvig/
- A few useful things to know about machine learning - by Pedro Domingos
  - http://dl.acm.org/citation.cfm?id=2347755

Data Science "folk knowledge" (4 of A)

- The simplest hypothesis that fits the data is also the most plausible
  - *Occam's Razor*
  - *Don't go for a 4 layer Neural Network unless you have that complex data*
  - *But that doesn't also mean that one should choose the simplest hypothesis*
  - *Match the impedance of the domain, data & the algorithms*
- Think of over fitting as memorizing as opposed to learning.
- Data leakage has many forms
- Sometimes the Absence of Something is Everything
- [Corollary] Absence of Evidence is not the Evidence of Absence

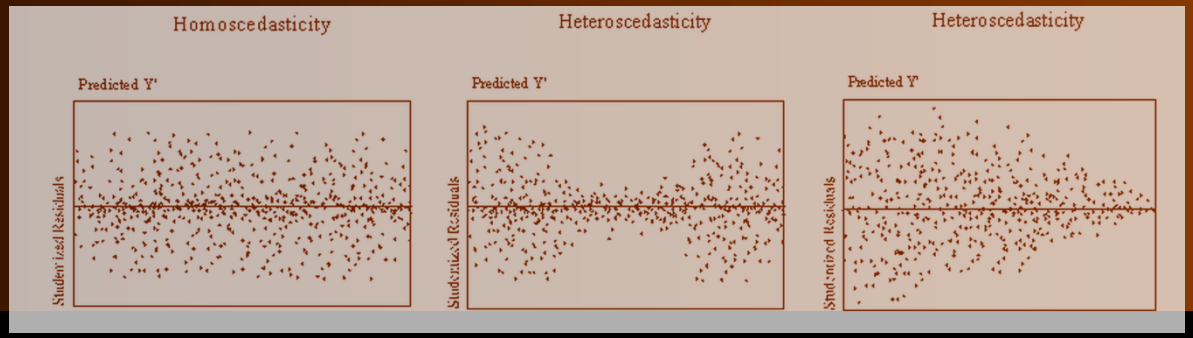


- Simple Model
  - High Error line that cannot be compensated with more data
  - Gets to a lower error rate with less data points
- Complex Model
  - Lower Error Line
  - But needs more data points to reach decent error

Ref: Andrew Ng/Stanford, Yaser S./CalTech

New to Machine Learning? Avoid these three mistakes, James Faghmous
https://medium.com/about-data/73258b3848a4

Check your assumptions



- The decisions a model makes, is directly related to the it's assumptions about the statistical distribution of the underlying data
- For example, for regression one should check that:

①Variables are normally distributed

- *Test for normality via visual inspection, skew & kurtosis, outlier inspections via plots, z-scores et al*

②There is a linear relationship between the dependent & independent variables

- *Inspect residual plots, try quadratic relationships, try log plots et al*

③Variables are measured without error

④Assumption of Homoscedasticity

- *Homoscedasticity assumes constant or near constant error variance*
- *Check the standard residual plots and look for heteroscedasticity*
  - *For example in the figure, left box has the errors scattered randomly around zero; while the right two diagrams have the errors unevenly distributed*

Jason W. Osborne and Elaine Waters, Four assumptions of multiple regression that researchers should always test, http://pareonline.net/getvn.asp?v=8&n=2

Data Science "folk knowledge" (5 of A)

Donald Rumsfeld is an armchair Data Scientist !

The World	You	
	UnKnown	Known
Knowns	○ Others know, you don't	○ What we do
Unknowns	○ Facts, outcomes or scenarios we have not encountered, nor considered ○ "Black swans", outliers, long tails of probability distributions ○ Lack of experience, imagination	○ Potential facts, outcomes we are aware, but not with certainty ○ Stochastic processes, Probabilities

- *Known Knowns*
  - *There are things we know that we know*
- *Known Unknowns*
  - *That is to say, there are things that we now know we don't know*
- *But there are also Unknown Unknowns*
  - *There are things we do not know we don't know*

http://smartorg.com/2013/07/valuepoint19/

Data Science "folk knowledge" (6 of A) - Pipeline

Collect

- Volume
- Velocity
- Streaming Data

Store

- Metadata
- Monitor counters & Metrics
- Structured vs. Multi-structured

Transform

- Canonical form
- Data catalog
- Data Fabric across the organization
- Access to multiple sources of data
- Think Hybrid – Big Data Apps, Appliances & Infrastructure

Reason

- Flexible & Selectable
  - Data Subsets
  - Attribute sets

Model

- Refine model with
  - Extended Data subsets
  - Engineered Attribute sets
- Validation run across a larger data set

Deploy

- Scalable Model Deployment
- Big Data automation & purpose built appliances (soft/ hard)
- Manage SLAs & response times

Data Management

Data Science

- Bytes to Business a.k.a. Build the full stack
- Find Relevant Data For Business
- Connect the Dots

Visualize

- Performance
- Scalability
- Refresh Latency
- In-memory Analytics

Recommend

- Advanced Visualization
- Interactive Dashboards
- Map Overlay
- Infographics

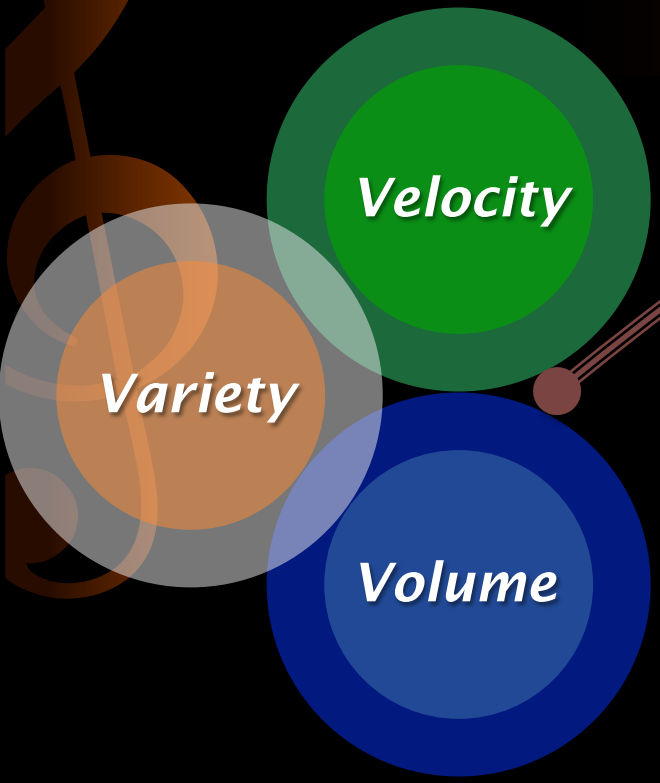
Predict

Explore

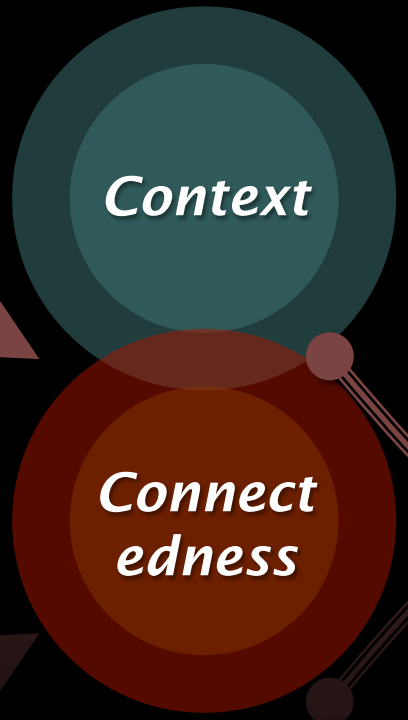
- Dynamic Data Sets
- 2 way key-value tagging of datasets
- Extended attribute sets
- Advanced Analytics

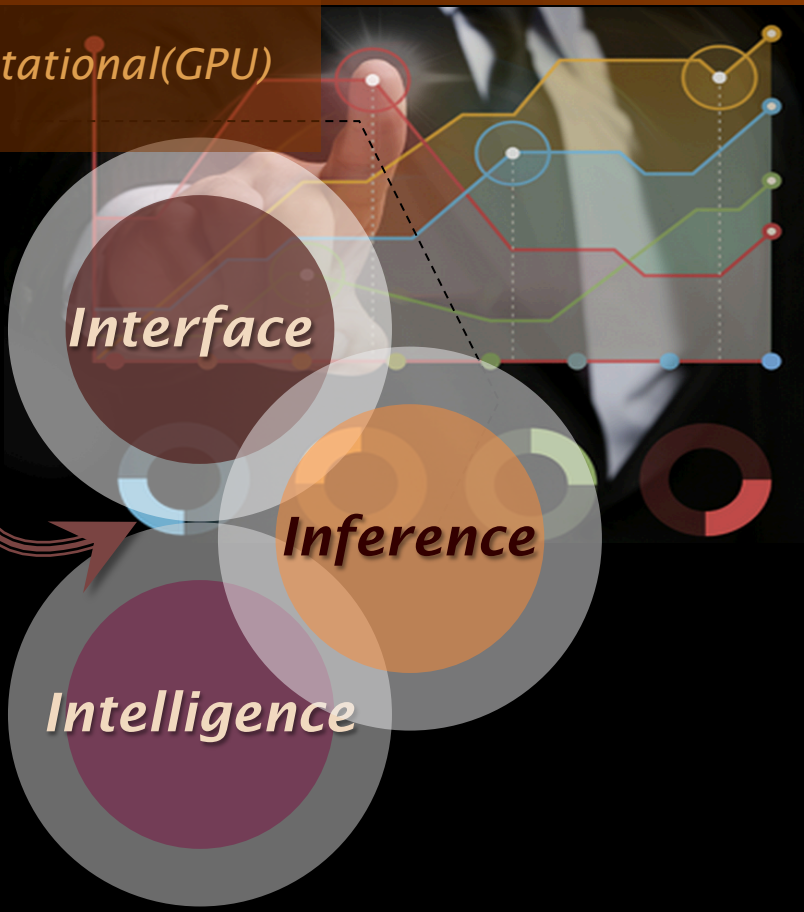
Data Science “folk knowledge” (7 of A)

- *Three Amigos*
  - *Interface = Cognition*
  - *Intelligence = Compute(CPU) & Computational(GPU)*
  - *Infer Significance & Causality*



“Data of unusual size”
that can’t be brute forced





Data Science "folk knowledge" (8 of A)
Jeremy's Axioms

- Iteratively explore data
- Tools
  - *Excel Format, Perl, Perl Book*
- Get your head around data
  - *Pivot Table*
- Don't over-complicate
- If people give you data, don't assume that you need to use all of it
- Look at pictures !
- History of your submissions – keep a tab
- Don't be afraid to submit simple solutions
  - *We will do this during this workshop*

Ref: http://blog.kaggle.com/2011/03/23/getting-in-shape-for-the-sport-of-data-sciencetalk-by-jeremy-howard/

Data Science “folk knowledge” (9 of A)

① Common Sense (some features make more sense then others)
② Carefully read these forums to get a peak at other peoples’ mindset
③ Visualizations
④ Train a classifier (e.g. logistic regression) and look at the feature weights
⑤ Train a decision tree and visualize it
⑥ Cluster the data and look at what clusters you get out
⑦ Just look at the raw data
⑧ Train a simple classifier, see what mistakes it makes
⑨ Write a classifier using handwritten rules
⑩ Pick a fancy method that you want to apply (Deep Learning/Nnet)

-- Maarten Bosma
-- http://www.kaggle.com/c/stumbleupon/forums/t/5761/methods-for-getting-a-first-overview-over-the-data

Data Science “folk knowledge” (A of A)
Lessons from Kaggle Winners

① Don’t over-fit

② All predictors are not needed

- *All data rows are not needed, either*

③ Tuning the algorithms will give different results

④ Reduce the dataset (Average, select transition data,…)

⑤ Test set & training set can differ

⑥ Iteratively explore & get your head around data

⑦ Don’t be afraid to submit simple solutions

⑧ Keep a tab & history your submissions

The curious case of the Data Scientist

- Data Scientist is multi-faceted & Contextual
- Data Scientist should be building Data Products
- Data Scientist should tell a story

Data Scientist (noun): Person who is better at statistics than any software engineer & better at software engineering than any statistician
– Josh Wills (Cloudera)

Large is hard; Infinite is much easier !
– Titus Brown

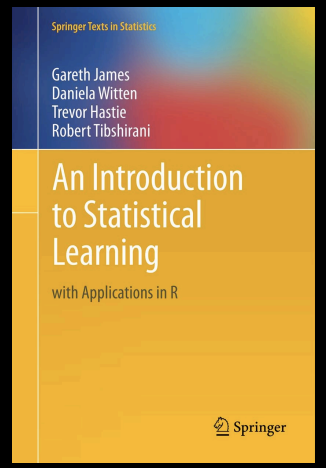
Data Scientist (noun): Person who is worse at statistics than any statistician & worse at software engineering than any software engineer
– Will Cukierski (Kaggle)

http://doubleclix.wordpress.com/2014/01/25/the-curious-case-of-the-data-scientist-profession/

Essential Reading List

- A few useful things to know about machine learning - by Pedro Domingos
  - http://dl.acm.org/citation.cfm?id=2347755
- The Lack of A Priori Distinctions Between Learning Algorithms by David H. Wolpert
  - http://mpdc.mae.cornell.edu/Courses/MAE714/Papers/lack_of_a_priori_distinctions_wolpert.pdf
- http://www.no-free-lunch.org/
- Controlling the false discovery rate: a practical and powerful approach to multiple testing Benjamini, Y. and Hochberg, Y. C
  - http://www.stat.purdue.edu/~doerge/BIOINFORM.D/FALL06/Benjamini%20and%20Y%20FDR.pdf
- A Glimpse of Googl, NASA,Peter Norvig + The Restaurant at the End of the Universe
  - http://doubleclix.wordpress.com/2014/03/07/a-glimpse-of-google-nasa-peter-norvig/
- Avoid these three mistakes, James Faghmo
  - https://medium.com/about-data/73258b3848a4
- Leakage in Data Mining: Formulation, Detection, and Avoidance
  - http://www.cs.umb.edu/~ding/history/470_670_fall_2011/papers/cs670_Tran_PreferredPaper_LeakingInDataMining.pdf

For your reading & viewing pleasure … An ordered List



① An Introduction to Statistical Learning
- http://www-bcf.usc.edu/~gareth/ISL/

② ISL Class Stanford/Hastie/Tibsharani at their best - Statistical Learning
- http://online.stanford.edu/course/statistical-learning-winter-2014

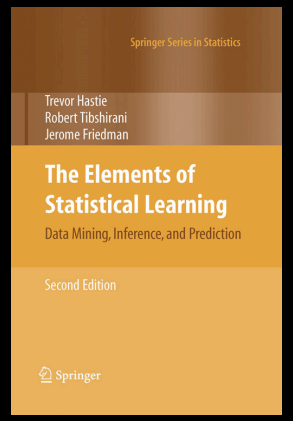
③ Prof. Pedro Domingo
- https://class.coursera.org/machlearning-001/lecture/preview

④ Prof. Andrew Ng
- https://class.coursera.org/ml-003/lecture/preview

⑤ Prof. Abu Mostafa, CaltechX: CS1156x: Learning From Data
- https://www.edx.org/course/caltechx/caltechx-cs1156x-learning-data-1120

⑥ Mathematicalmonk @ YouTube
- https://www.youtube.com/playlist?list=PLD0F06AA0D2E8FFBA

⑦ The Elements Of Statistical Learning
- http://statweb.stanford.edu/~tibs/ElemStatLearn/

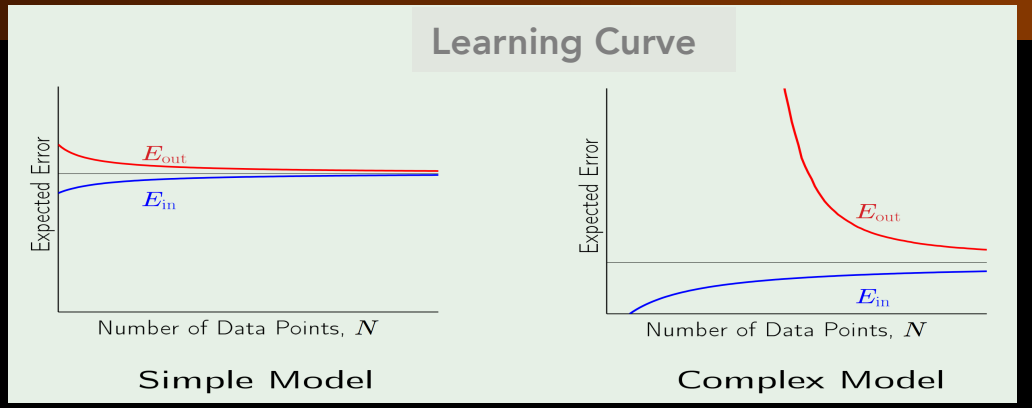


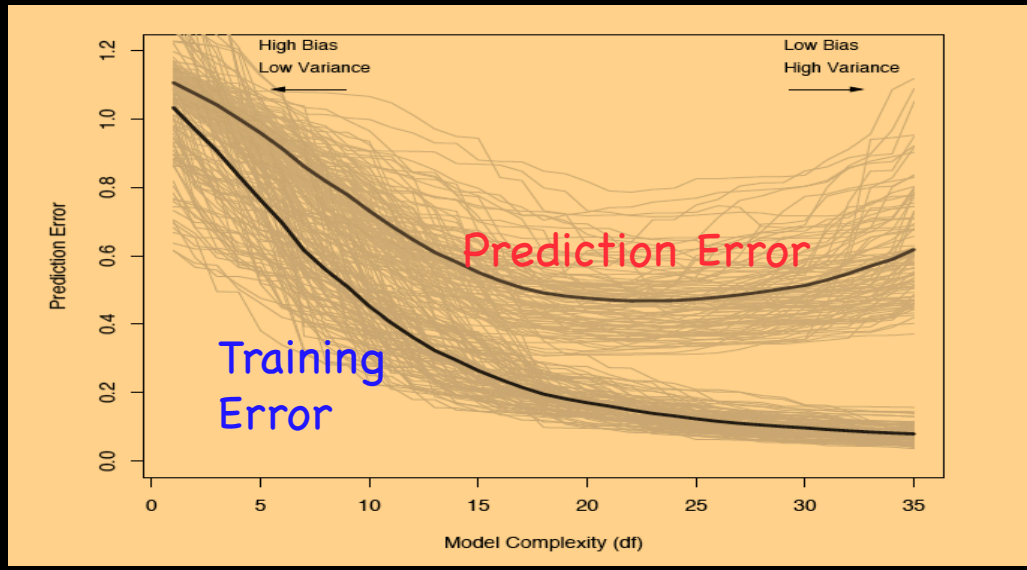
http://www.quora.com/Machine-Learning/Whats-the-easiest-way-to-learn-machine-learning/

Of Models, Performance, Evaluation & Interpretation

Bias/Variance (1 of 2)

- Model Complexity
  - *Complex Model increases the training data fit*
  - *But then it overfits & doesn't perform as well with real data*
- Bias vs. Variance
  - Classical diagram
  - *From ELSII, By Hastie, Tibshirani & Friedman*
- *Bias – Model learns wrong things; not complex enough; error gap small; more data by itself won't help*
- *Variance – Different dataset will give different error rate; over fitted model; larger error gap; more data could help*





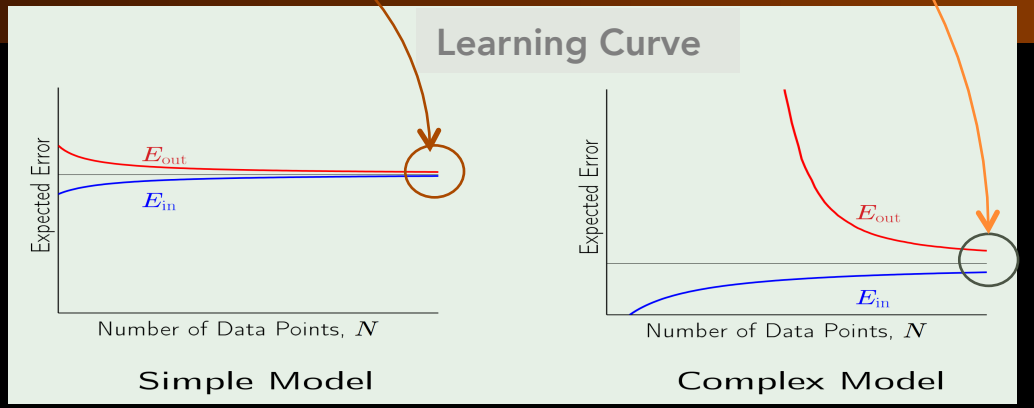
Ref: Andrew Ng/Stanford, Yaser S./CalTech

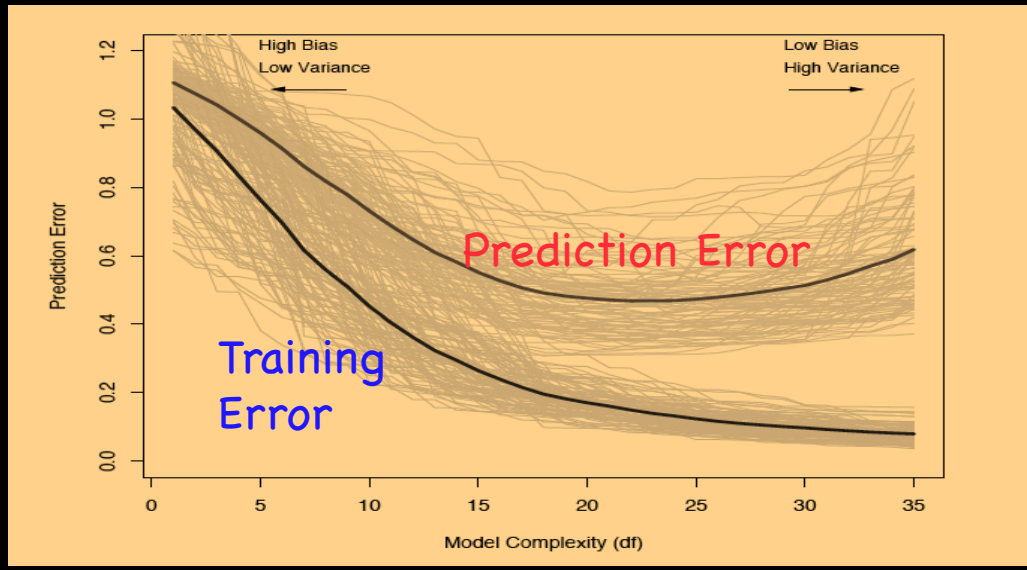
Bias/Variance (2 of 2)

- High Bias
  - *Due to Underfitting*
  - *Add more features*
  - *More sophisticated model*
    - *Quadratic Terms, complex equations,…*
  - *Decrease regularization*
- High Variance
  - *Due to Overfitting*
  - *Use fewer features*
  - *Use more training sample*
  - *Increase Regularization*

Need more features or more complex model to improve

Need more data to improve





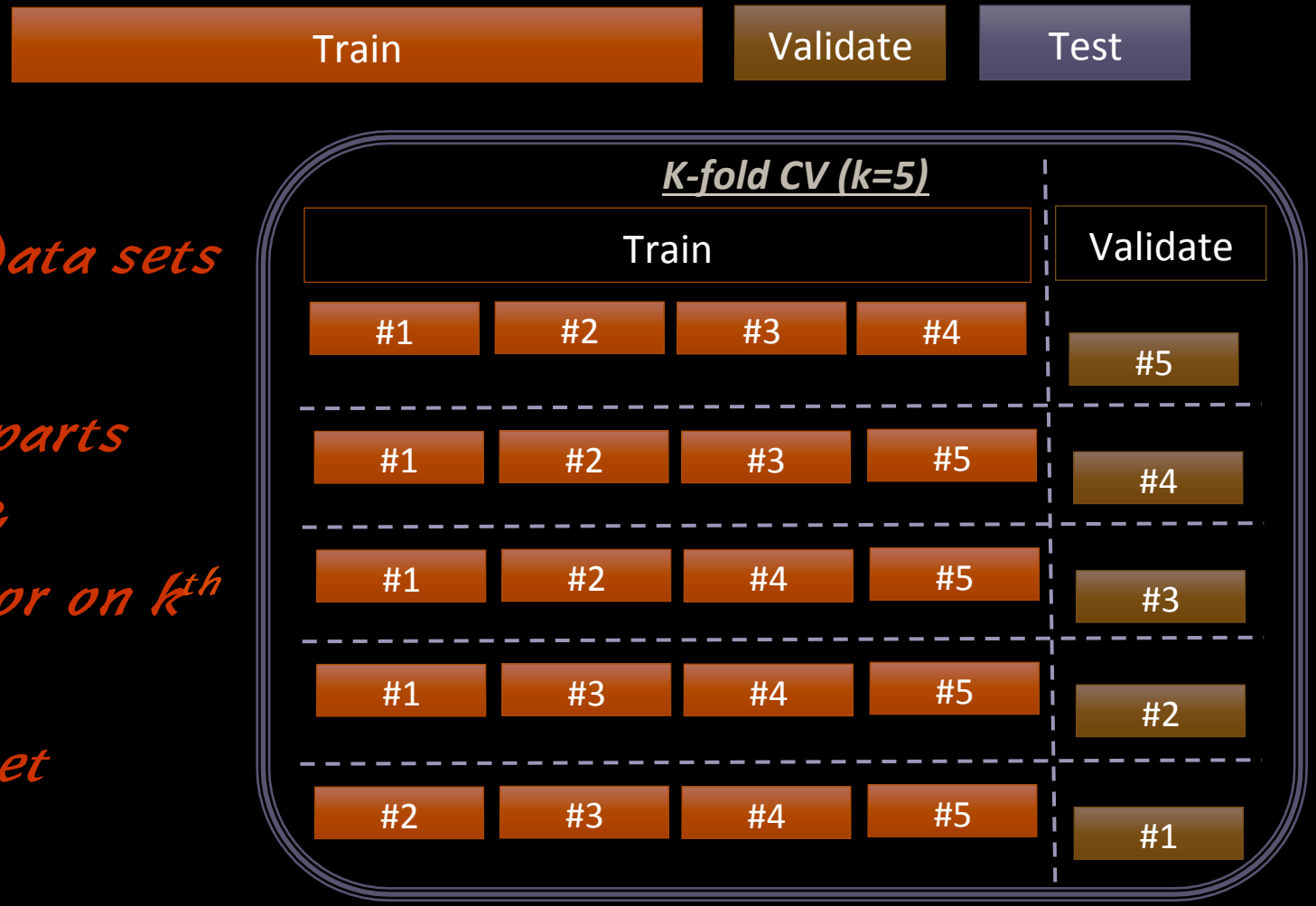
Ref: Strata 2013 Tutorial by Olivier Grisel

Data Partition & Cross-Validation

- **Goal**
  - Model Complexity (-)
  - Variance (-)
  - Prediction Accuracy (+)

Partition Data !

- Training (60%)
- Validation(20%) &
- "Vault" Test (20%) Data sets



k-fold Cross-Validation

- Split data into k equal parts
- Fit model to k–1 parts & calculate prediction error on k^{th} part
- Non-overlapping dataset

Bootstrap & Bagging

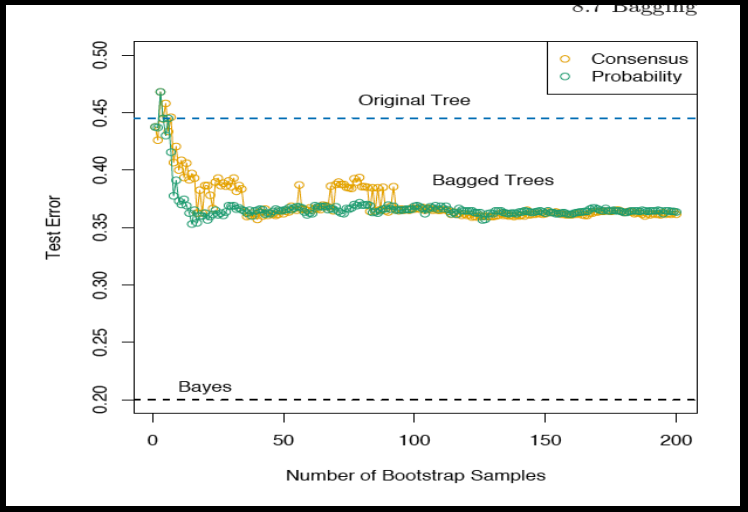
- *Goal*
  - *Model Complexity (-)*
  - *Variance (-)*
  - *Prediction Accuracy (+)*

Bootstrap

- *Draw datasets (with replacement) and fit model for each dataset*
  - *Remember : Data Partitioning (#1) & Cross Validation (#2) are without replacement*

Bagging (Bootstrap aggregation)

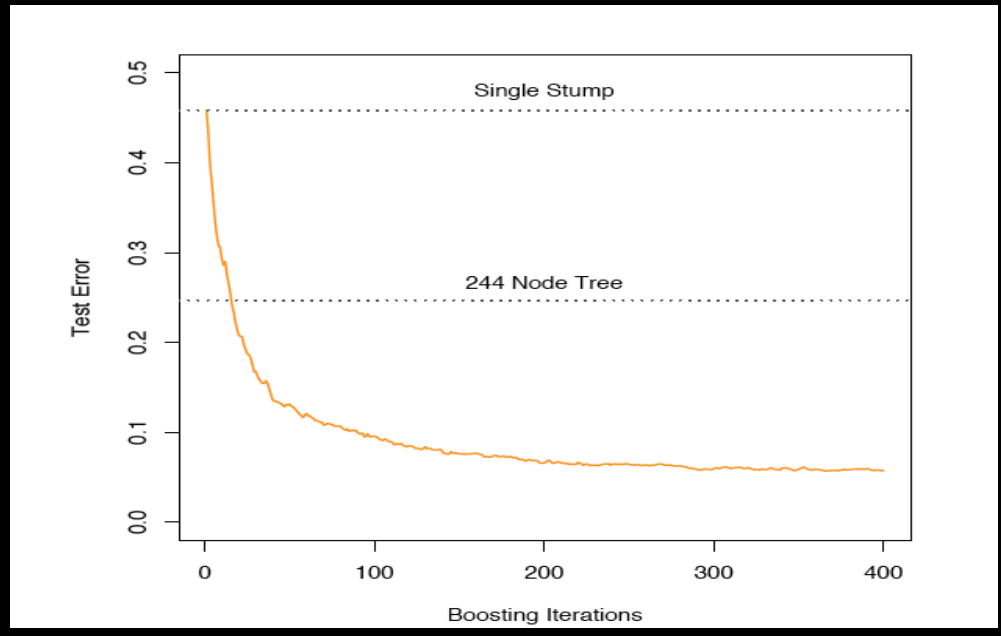
- Average prediction over a collection of bootstrap-ed samples, thus reducing variance



Boosting

- *Goal*
  - *Model Complexity (-)*
  - *Variance (-)*
  - *Prediction Accuracy (+)*

- "Output of weak classifiers into a powerful committee"
- Final Prediction = weighted majority vote
- Later classifiers get misclassified points
  - With higher weight,
  - So they are forced
  - To concentrate on them
- AdaBoost (*Ada*ptive*Boost*ing)
- Boosting vs Bagging
  - Bagging – independent trees
  - Boosting – successively weighted



Random Forests+

- *Goal*
  - *Model Complexity (-)*
  - *Variance (-)*
  - *Prediction Accuracy (+)*

- Builds large collection of de-correlated trees & averages them
- Improves Bagging by selecting i.i.d* random variables for splitting
- Simpler to train & tune
- "*Do remarkably well, with very little tuning required*" – ESLII
- Less susceptible to over fitting (than boosting)
- Many RF implementations
  - Original version - Fortran-77 ! By Breiman/Cutler
  - Python, R, Mahout, Weka, Milk (ML toolkit for py), matlab

* i.i.d – independent identically distributed
+ http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

Ensemble Methods

- *Goal*
  - *Model Complexity (-)*
  - *Variance (-)*
  - *Prediction Accuracy (+)*

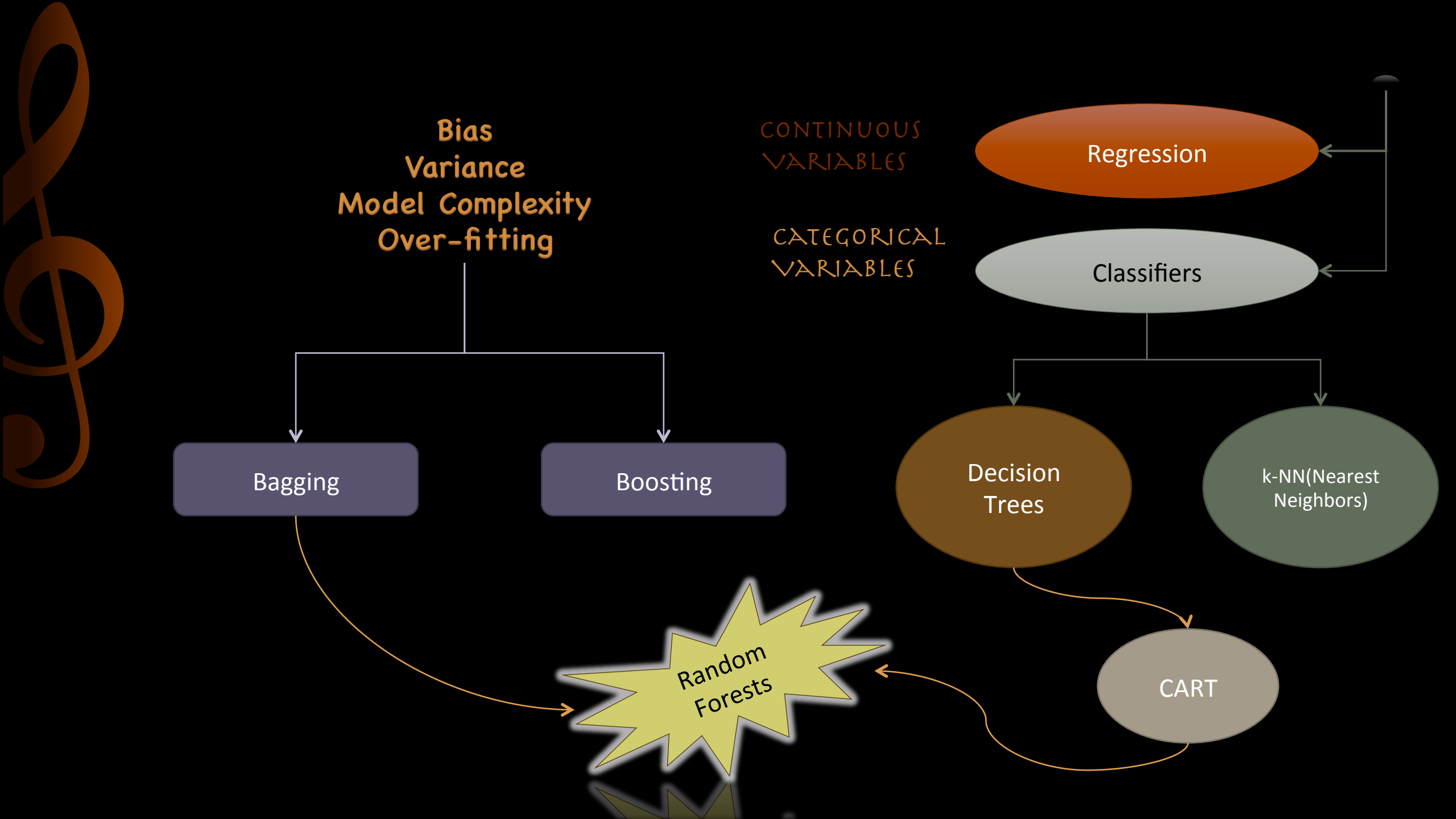
- Two Step
  - Develop a set of learners
  - Combine the results to develop a composite predictor
- Ensemble methods can take the form of:
  - Using different algorithms,
  - Using the same algorithm with different settings
  - Assigning different parts of the dataset to different classifiers
- Bagging & Random Forests are examples of ensemble method

Ref: Machine Learning In Action

Random Forests

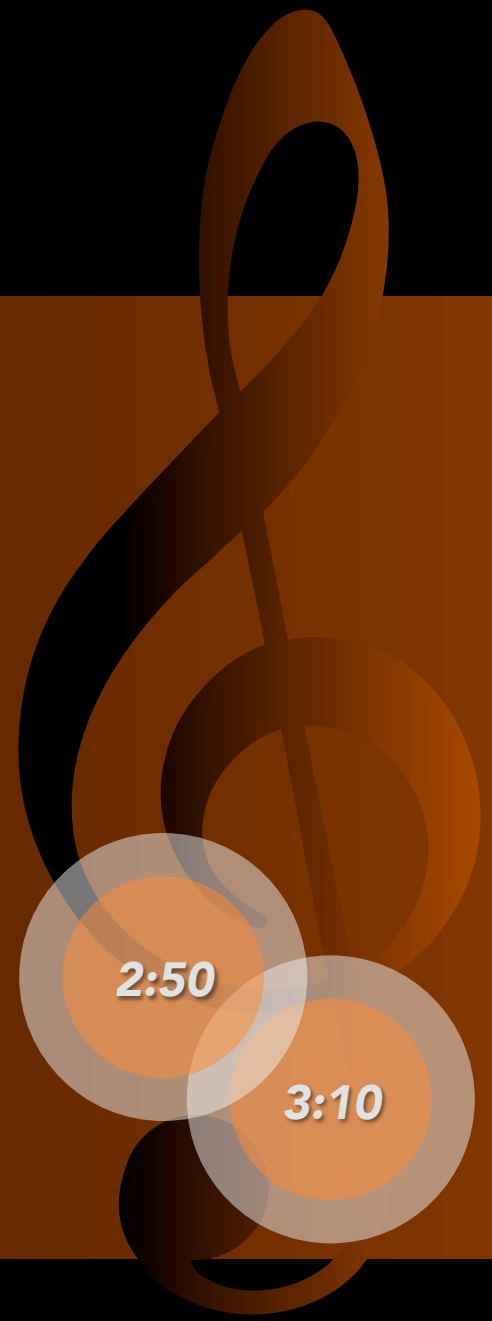
- While Boosting splits based on best among *all* variables, RF splits based on best among *randomly chosen* variables
- Simpler because it requires two variables – no. of Predictors (typically √k) & no. of trees (500 for large dataset, 150 for smaller)
- Error prediction
  - *For each iteration, predict for dataset that is not in the sample (OOB data)*
  - *Aggregate OOB predictions*
  - *Calculate Prediction Error for the aggregate, which is basically the OOB estimate of error rate*
    - *Can use this to search for optimal # of predictors*
  - *We will see how close this is to the actual error in the Heritage Health Prize*
- Assumes equal cost for mis-prediction. Can add a cost function
- Proximity matrix & applications like adding missing data, dropping outliers

Ref: R News Vol 2/3, Dec 2002
Statistical Learning from a Regression Perspective : Berk
A Brief Overview of RF by Dan Steinberg

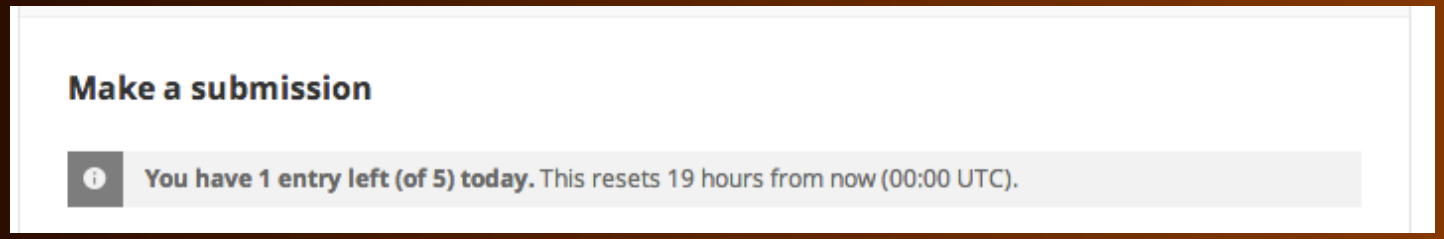

Bias
Variance
Model Complexity
Over-fitting
Bagging
Boosting
CONTINUOUS VARIABLES
CATEGORICAL VARIABLES
Regression
Classifiers
Decision Trees
k-NN(Nearest Neighbors)
CART
Random Forests

Model Evaluation & Interpretation

Relevant Digression



Cross Validation



Refer to iPython notebook <2-Model-Evaluation>
at https://github.com/xsankar/freezing-bear

- Reference:
  - https://www.kaggle.com/wiki/GettingStartedWithPythonForDataScience
  - Chris Clark 's blog : http://blog.kaggle.com/2012/07/02/up-and-running-with-python-my-first-kaggle-entry/
  - Predicive Modelling in py with scikit-learning, Olivier Grisel Strata 2013
    - titanic from pycon2014/parallelmaster/An introduction to Predictive Modeling in Python

Model Evaluation - Accuracy

	Predicted=1	Predicted=0
Actual =1	True+ (tp)	False- (fn)
Actual=0	False+ (fp)	True- (tn)

- Accuracy = $$\frac{tp + tn}{tp+fp+fn+tn}$$
- For cases where tn is large compared tp, a degenerate return(false) will be very accurate !
- Hence the F-measure is a better reflection of the model strength

Model Evaluation – Precision & Recall

	Predicted=1	Predicted=0
Actual =1	True+ (tp)	False- (fn)
Actual=0	False+ (fp)	True- (tn)

- Precision
- Accuracy
- Relevancy

$$\frac{tp}{tp+fp}$$

- Recall
- True +ve Rate
- Coverage
- Sensitivity
- Hit Rate

$$\frac{tp}{tp+fn}$$

- Type 1 Error Rate
- False +ve Rate
- False Alarm Rate

$$\frac{fp}{fp+tn}$$

- Specificity = 1 – fp rate

- Precision = How many items we identified are relevant
- Recall = How many relevant items did we identify
- Inverse relationship – Tradeoff depends on situations
  - *Legal – Coverage is important than correctness*
  - *Search – Accuracy is more important*
  - *Fraud*
    - *Support cost (high fp) vs. wrath of credit card co.(high fn)*

- Type 1 Error = fp
- Type 2 Error = fn

http://nltk.googlecode.com/svn/trunk/doc/book/ch06.html

Confusion Matrix

Actual		Predicted			
		C_1	C_2	C_3	C_4
	C_1	10	5	9	3
	C_2	4	20	3	7
	C_3	6	4	13	3
	C_4	2	1	4	15

Correct Ones (c_{ii})

$$\text{Precision} = \frac{c_{ii}}{\sum_{\text{Columns } i} c_{ij}}$$

$$\text{Recall} = \frac{c_{ii}}{\sum_{\text{Rows } j} c_{ij}}$$

Model Evaluation : F-Measure

	Predicted=1	Predicted=0
Actual =1	True+ (tp)	False- (fn)
Actual=0	False+ (fp)	True- (tn)

Precision = tp / (tp+fp) : Recall = tp / (tp+fn)

F-Measure

Balanced, Combined, Weighted Harmonic Mean, measures effectiveness

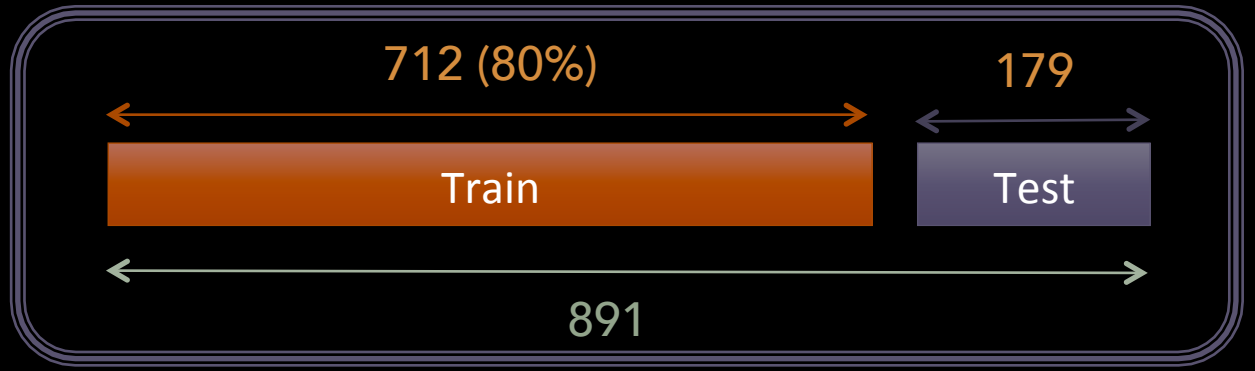
$$\frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Common Form (Balanced F1) : $\beta = 1$ ($\alpha = ½$) ; F1 = 2PR / P+R

Hands-on Walkthru - Model Evaluation

```
In [4]: from sklearn import metrics
        from sklearn.cross_validation import train_test_split
        #Split 80-20 train vs test data
        train_x, test_x, train_y, test_y = train_test_split(train_x_all,train_y_all, test_size=0.20, random_state=0)
        print (train_x_all.shape,train_y_all.shape)
        print (train_x.shape,train_y.shape)
        print (test_x.shape,test_y.shape)

        ((891, 5), (891,))
        ((712, 5), (712,))
        ((179, 5), (179,))
```



Refer to iPython notebook <2-Model-Evaluation>
at https://github.com/xsankar/freezing-bear

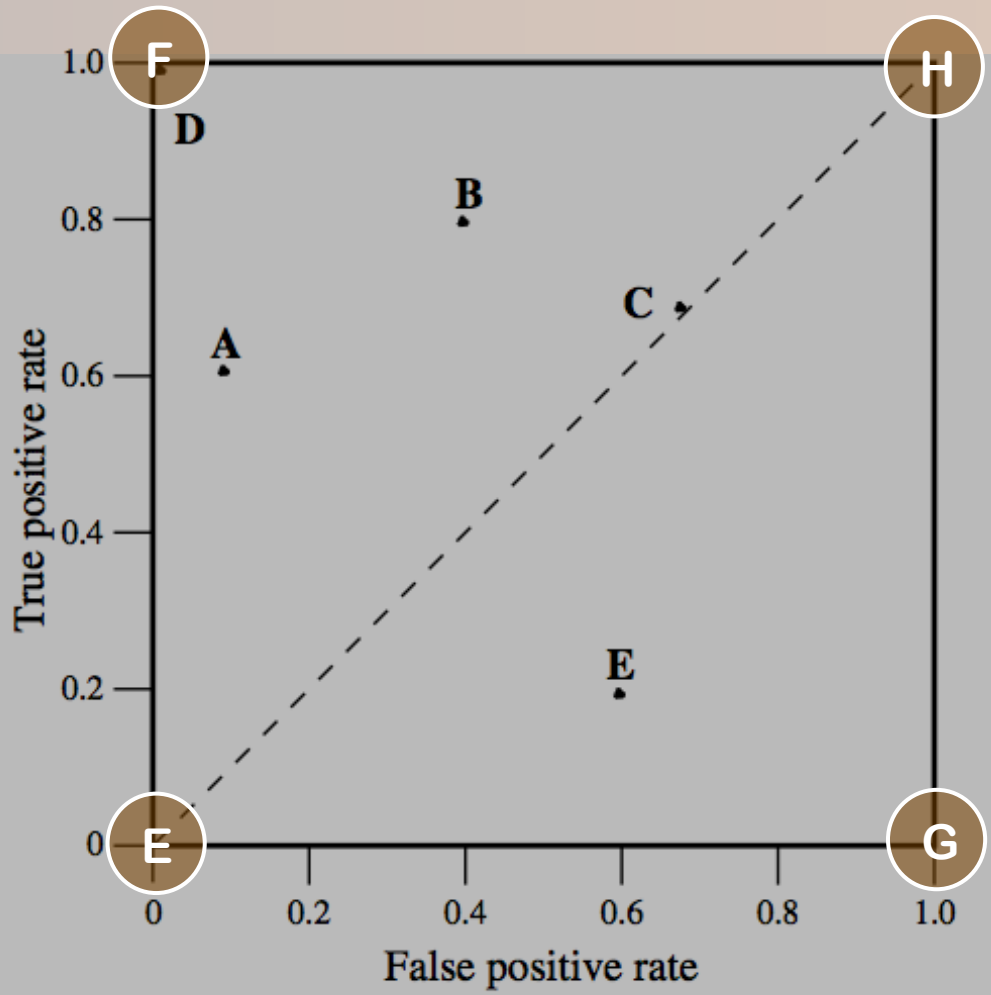
ROC Analysis

- "How good is my model?"
- Good Reference : http://people.inf.elte.hu/kiss/13dwhdm/roc.pdf
- *"A receiver operating characteristics (ROC) graph is a technique for visualizing, organizing and selecting classifiers based on their performance"*
- Much better than evaluating a model based on simple classification accuracy
- Plots tp rate vs. fp rate
- After understanding the ROC Graph, we will draw a few for our models in iPython notebook <2-Model-Evaluation> at https://github.com/xsankar/freezing-bear

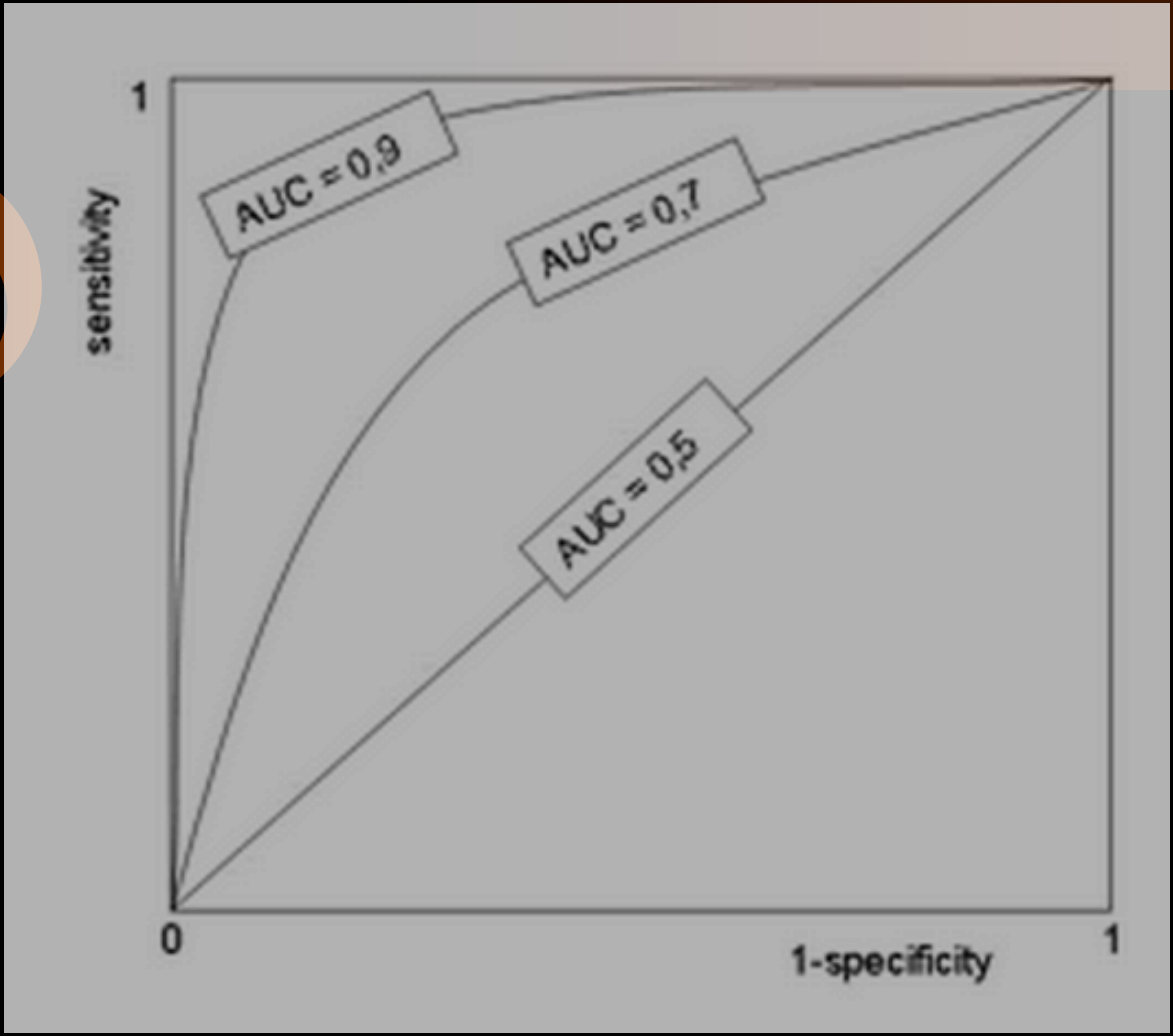
ROC Graph - Discussion

	Predicted=1	Predicted=0
Actual =1	True+ (tp)	False- (fn)
Actual=0	False+ (fp)	True- (tn)

- E = Conservative, Everything NO
- H = Liberal, Everything YES
- *Am not making any political statement !*
- *F = Ideal*
- *G = Worst*
- *The diagonal is the chance*
- *North West Corner is good*
- *South-East is bad*
  - *For example E*
  - *Believe it or Not – I have actually seen a graph with the curve in this region !*



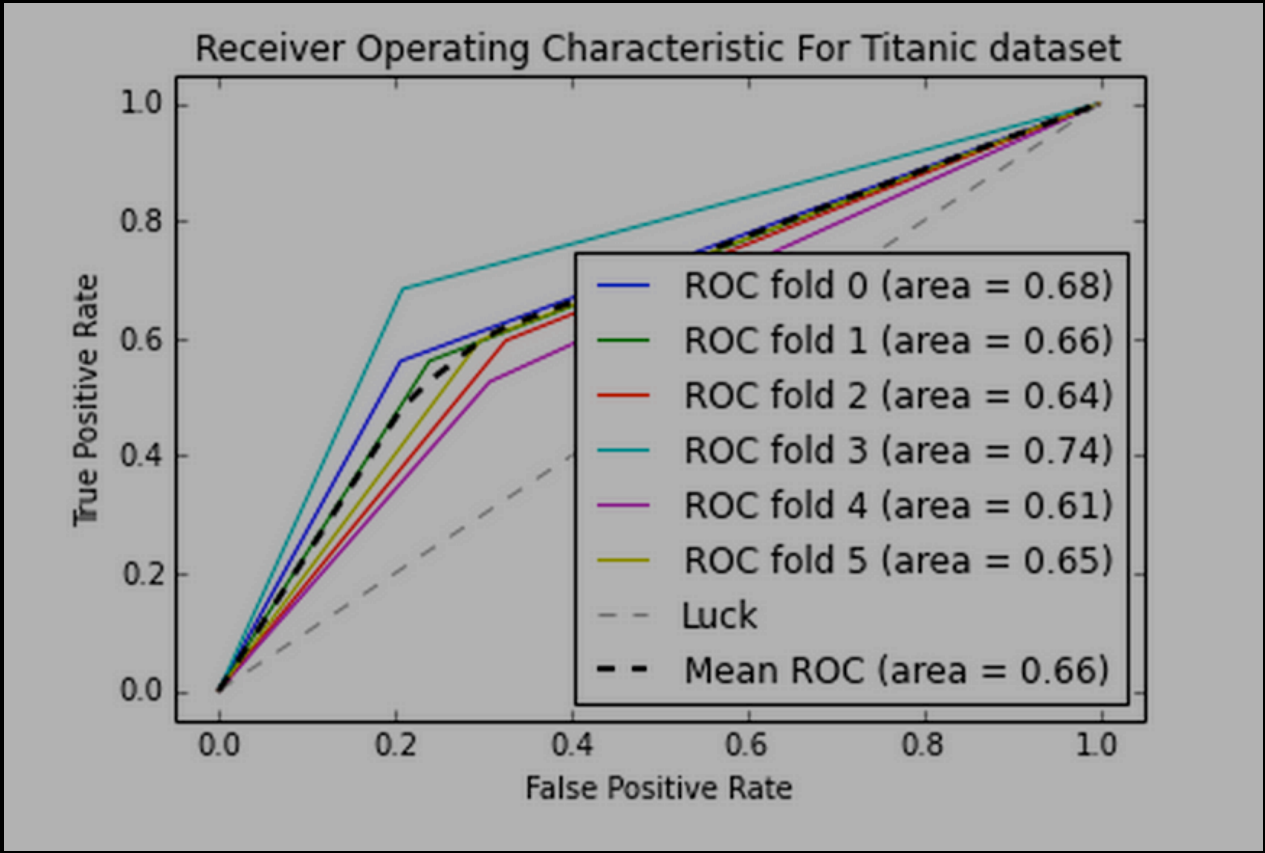
ROC Graph – Clinical Example

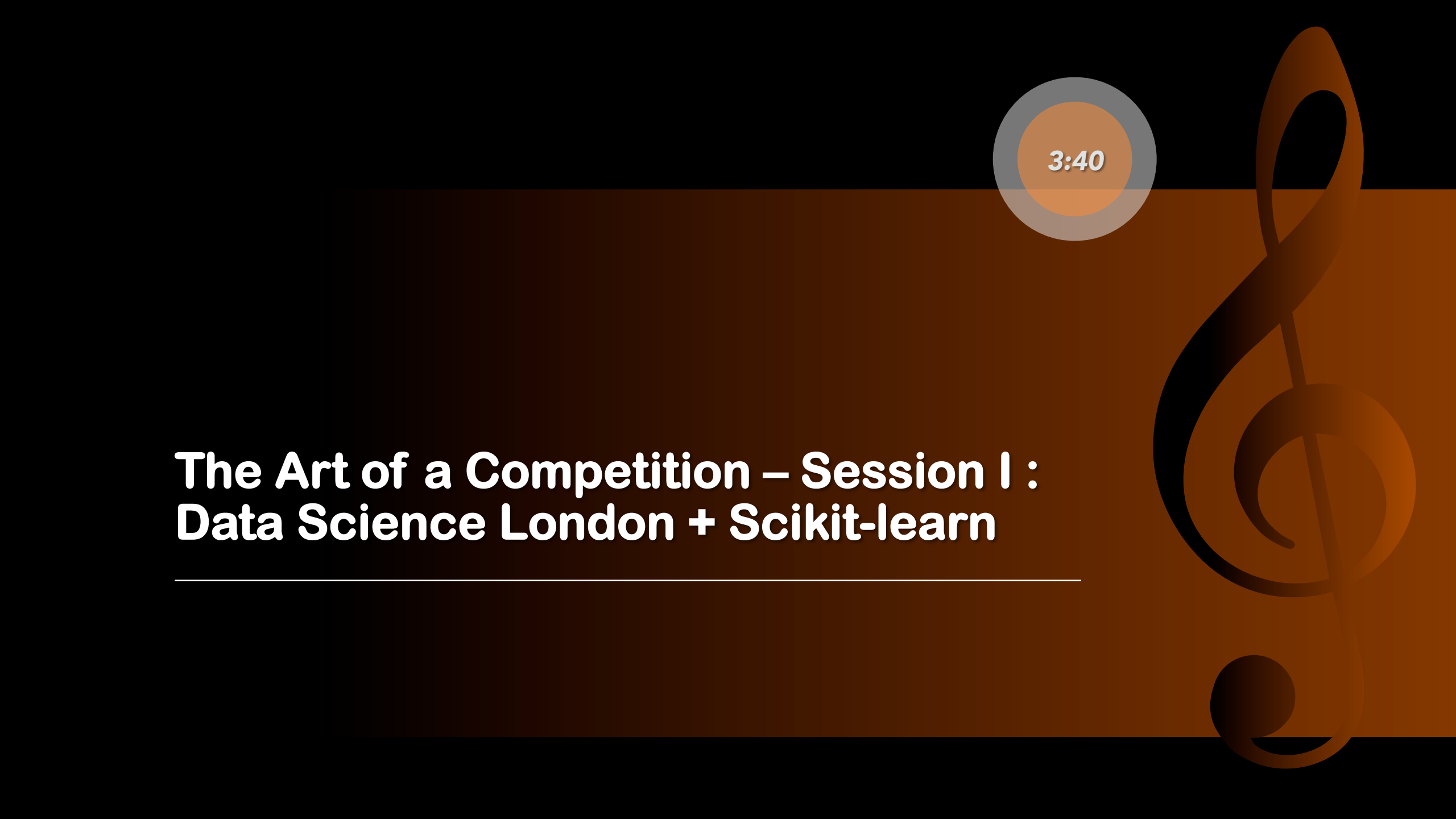


area	diagnostic accuracy
0.9 – 1.0	excellent
0.8 - 0.9	very good
0.7 - 0.8	good
0.6 - 0.7	sufficient
0.5 - 0.6	bad
< 0.5	test not useful

ROC Graph Walk thru

- iPython notebook <2-Model-Evaluation> at https://github.com/xsankar/freezing-bear





The Art of a Competition – Session I : Data Science London + Scikit-learn

Few interesting Links-troll the forums

- http://www.kaggle.com/c/data-science-london-scikit-learn/visualization/1113
  - *Will's Solution*
- Quick First prediction : http://www.kaggle.com/c/data-science-london-scikit-learn/visualization/1075
- CV http://www.kaggle.com/c/data-science-london-scikit-learn/visualization/1183
- In-depth Solution : http://www.kaggle.com/c/data-science-london-scikit-learn/forums/t/6528/solution-discussion
- Video http://datasciencelondon.org/machine-learning-python-scikit-learn-ipython-dsldn-data-science-london-kaggle/
- More in http://www.kaggle.com/c/data-science-london-scikit-learn/visualization

#1 : SVM-Will

Refer to iPython notebook <3-Session-I>
at https://github.com/xsankar/freezing-bear

108	new	Krishna Sankar	0.91282	1	Thu, 27 Mar 2014 02:29:04

Your Best Entry
Congratulations on making your first submission!

#3 : Unsupervised Feature Engineering

80	new	Krishna Sankar	0.91393	2	Sat, 29 Mar 2014 23:01:04

Your Best Entry
You improved on your best score by 0.00112.
You just moved up 30 positions on the leaderboard.

The Art of a Competition - Session II : ASUS, PAKDD2014

The 18th Pacific-Asia Conference on Knowledge Discovery & Data Mining

4:10

Data Organization & Approach

- Deceptively Simple Data
  - *Sales*
  - *Repairs*
  - *Predict Future Repairs*
- Explore data
  - *Module : M1–M9*
  - *Component : P1–P31*
- Repair before sales
- -ve sales
- Reason for exploring this competition is to get a feel for a complex dataset in terms of processing

Sales

	A	B	C	D
1	module_category	component_category	year/month	number_sale
2	M4	P27	2005/1	0
3	M4	P27	2005/5	1042
4	M4	P27	2005/9	1677
5	M4	P27	2005/10	918
6	M4	P27	2005/11	0
7	M0	P27	2006/8	210
8	M0	P27	2006/11	0
9	M0	P27	2007/4	0

Repairs

	A	B	C	D	E
1	module_category	component_category	year/month(sale)	year/month(repair)	number_repair
2	M6	P16	2007/9	2009/4	1
3	M2	P30	2007/9	2009/8	1
4	M1	P12	2006/10	2008/2	2
5	M1	P30	2006/5	2007/7	1
6	M3	P06	2007/8	2007/12	1
7	M7	P19	2006/7	2007/6	1
8	M7	P04	2006/3	2008/4	1
9	M3	P09	2006/11	2007/6	1

Future Repairs ?

	A	B	C	D
1	module_category	component_category	year	month
2	M1	P02	2010	1
3	M1	P02	2010	2
4	M1	P02	2010	3
5	M1	P02	2010	4
6	M1	P02	2010	5
7	M1	P02	2010	6
8	M1	P02	2010	7

Approach, Ideas & Results

Refer to iPython notebook <4-Session-II>
at https://github.com/xsankar/freezing-bear

- Discussions – troll the forums
  - http://www.kaggle.com/c/pakdd-cup-2014/forums/t/7573/what-did-you-do-to-get-to-the-top-of-the-board
    - Ran Locar, James King, https://github.com/aparij/kaggle__asus/blob/master/lin__comb__survival.py, Brandon Kam,
  - http://www.kaggle.com/c/pakdd-cup-2014/forums/t/6980/sample-submission-benchmark-in-leaderboard
    - Beat__benchmark by Chitrasen
    - Beat-benchmary.py
- Lots of interesting models & ideas
- Simple linear regression
- Quadratic Time ?
- Batch Mortality Rate
- Breaks with higher repair tendency ? Survival Analysis ?

#1 – All 0s, All 1s

Refer to iPython notebook <4-Session-II>
at https://github.com/xsankar/freezing-bear

#	Δ1w	Team Name * in the money	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑2	Veroljub Mihajlovic *	1.58694	79	Mon, 31 Mar 2014 01:38:49 (-0.3h)
2	↑11	Shize Su *	1.67648	31	Mon, 31 Mar 2014 01:01:03 (-15.3h)
3	↑4	eluk *	1.69831	64	Sun, 30 Mar 2014 02:40:32
4	↓3	gregl	1.71526	98	Sun, 30 Mar 2014 18:24:11 (-6d)
5	↑7	Manuel Lopez	1.81720	90	Sun, 30 Mar 2014 09:37:50 (-23.9h)
6	↓4	Herra Huu	1.88765	82	Mon, 31 Mar 2014 00:01:12 (-27.4h)
7	↑3	sunshine	1.92575	22	Sun, 30 Mar 2014 23:23:58

Submission	Files	Public Score
Mon, 31 Mar 2014 01:51:59 All ones-Test for pycon 2014 Tutorial Edit description	ones.csv	5.89991
Mon, 31 Mar 2014 01:47:00 All Zeros - Test for pycon 2014 Edit description	zeros.csv	5.65179

- All 0s
- All 1s

26 hours to go

Sunday, January 26, 2014 — $8,500 • 603 teams — Tuesday, April 1, 2014

This competition allows you to make 2 entries in a day and your team has already hit that limit. Please wait until midnight UTC (22 hours from now) to make your next submission.

#2 :

- Let us get serious & do some transformation
- Split
- Convert to int64

Refer to iPython notebook <4-Session-II>
at https://github.com/xsankar/freezing-bear

#3 : Hints To Try

- Decay
- Kaplan Meier Analysis
- Cox Model
- Parametric Survival Models

The Beginning As The End

4:10

The Beginning As the End

Goals & Assumptions

- Goals:
  - *Get familiar with the mechanics of Data Science Competitions*
  - *Explore the intersection of Algorithms, Data, Intelligence, Inference & Results*
  - *Discuss Data Science Horse Sense ;o)*
- At the end of the tutorial you should have :
  - *Submitted entries for 3 competitions*
  - *Applied Algorithms on Kaggle Data*
    - *CART, RF*
    - *Linear Regression*
    - *SVM*
  - *Explored Data, have a good understanding of the Analytics Pipeline viz. collect-store-transform-model-reason-deploy-visualize-recommend-infer-explore*
  - *Knowledge of Model Evaluation*
    - *Cross Validation, ROC Curves*

- We started with a set of goals
- Homework
  - For you
    - *Go through the slides*
    - *Do the walkthrough*
    - *Work on more competitions*
    - *Submit entries to Kaggle*

References:

- An Introduction to scikit-learn, pycon 2013, Jake Vanderplas
  - http://pyvideo.org/video/1655/an-introduction-to-scikit-learn-machine-learning
- Advanced Machine Learning with scikit-learn, pycon 2013, Strata 2014, Olivier Grisel
  - http://pyvideo.org/video/1719/advanced-machine-learning-with-scikit-learn
- Just The Basics, Strata 2013, William Cukierski & Ben Hamner
  - http://strataconf.com/strata2013/public/schedule/detail/27291
- The Problem of Multiple Testing
  - http://download.journals.elsevierhealth.com/pdfs/journals/1934-1482/PIIS1934148209014609.pdf

I enjoyed a lot preparing the materials …

Hope you enjoyed more attending …