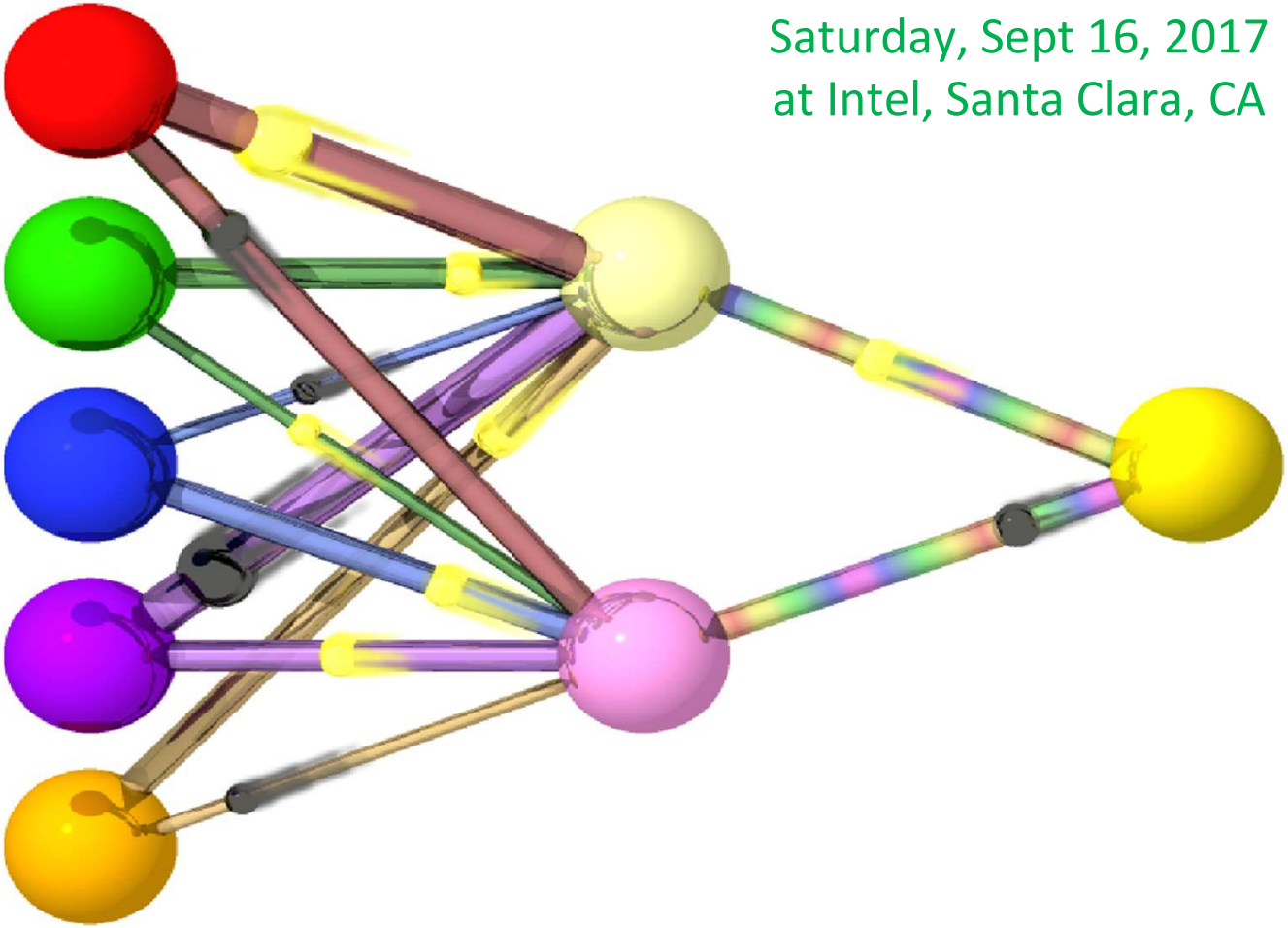
Predictive Data Science in R

SFbayACM Professional Development Seminar (PDS)
Saturday, Sept 16, 2017
at Intel, Santa Clara, CA



By Greg Makowski

http://www.LinkedIn.com/in/GregMakowski

Primary Class Take-Away's

Increase how you can be competitive in predictive projects by:

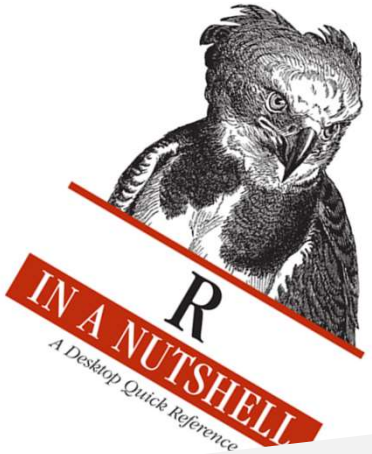
1. **DESIGN the Structure of the Solution**

   (data pull and business metric to optimize)

2. **DESIGN of Knowledge Representation and Preprocessing**
   1. Scalable in record volume
   2. Scalable in complexity, deal with 1000's vars

   Design before coding

3. **Practical R Mechanics training and going into production**



"Intro to Prediction with Small Data"

Focus on complementing existing training, not repeating

Value of EXPERIENCE

Target Student Background

- Has academic or professional programming experience
- Has tried some R programming

  The tutorial is not at a speed of "first exposure to R"

  MOVING FASTER
- Has had some exposure to analytic methods or statistics or matrices

  Will provide some review

- My goal is to start with people in the 20-50 percentile of experience
- point them in a direction the can continue to get to the 70-80 percentile of experience
- I hope all feel that they got at least 1-3 solid "lessons learned" to try in the future in your projects.

Survey of Student Background

- Show of hands on audience background
  - –Have some R experience?
  - –Have read about, or preliminary training in predictive algorithms?
  - –Have DS projects they could work on in next 2 months
  - –Training or experience in other quant analytics
    - •Clustering
    - •Recommender systems
    - •NLP
    - •Optimization
  - –Have deployed data mining

Outline

Part 1: Get started with R, play with data

R Resources, how to get started in R, overview of RStudio

Basics of R variables, lists, read csv → data.table, look at data

Review **HoMe EQuity** (**HMEQ**) loan example problem

Lab 1a: you play with R, looking at the data in data.table

Preliminary review of algorithms and parameters

Lab 1b: Train decision tree on HMEQ, TensorFlow Playground

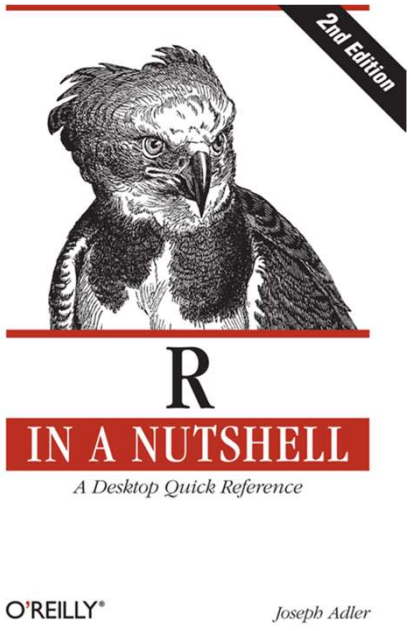
Part 2: Data Science Project Design

Part 3: Preprocessing Design

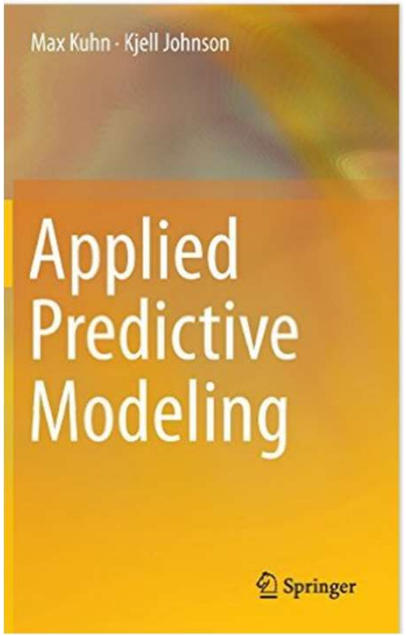
Part 4: Modeling Design

R Resources: Good Books on R for Data Science

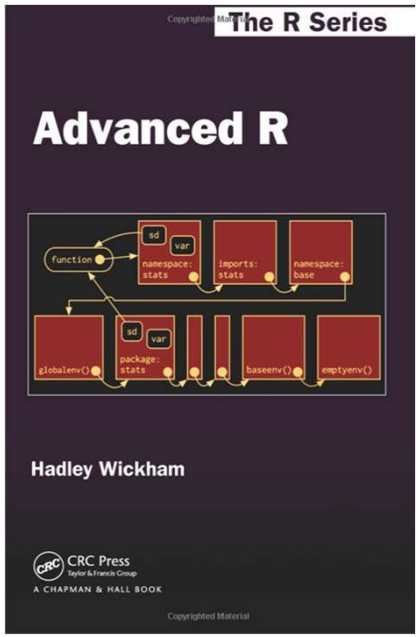
Broad R overview, Older.
Don't spend time on data.frames, use data.tables instead



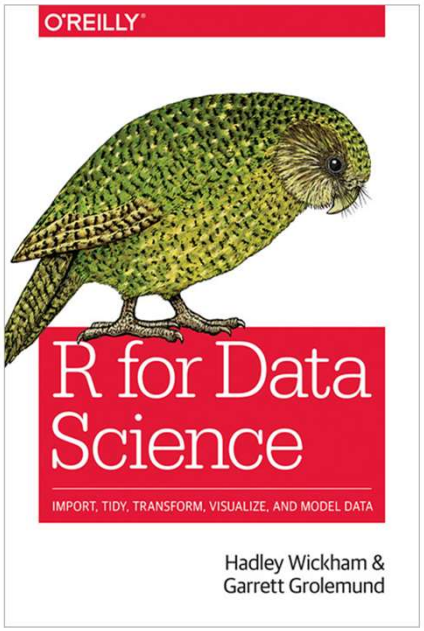
Caret library in detail from library author. A single interface for 233 predictive algorithms.



Hadley Wickham the Chief Scientist at Rstudio, and is Author of 14 libraries.
A good person to learn from.
http://hadley.nz/



http://r4ds.had.co.nz/
Free, all content in web



Other R Training

edX Course, “Programming with R for Data Science”

6 weeks, 4-8 hrs / week, Free (certificate for $99)

https://www.edx.org/course/programming-r-data-science-microsoft-dat209x-3

They have a similar course for Python

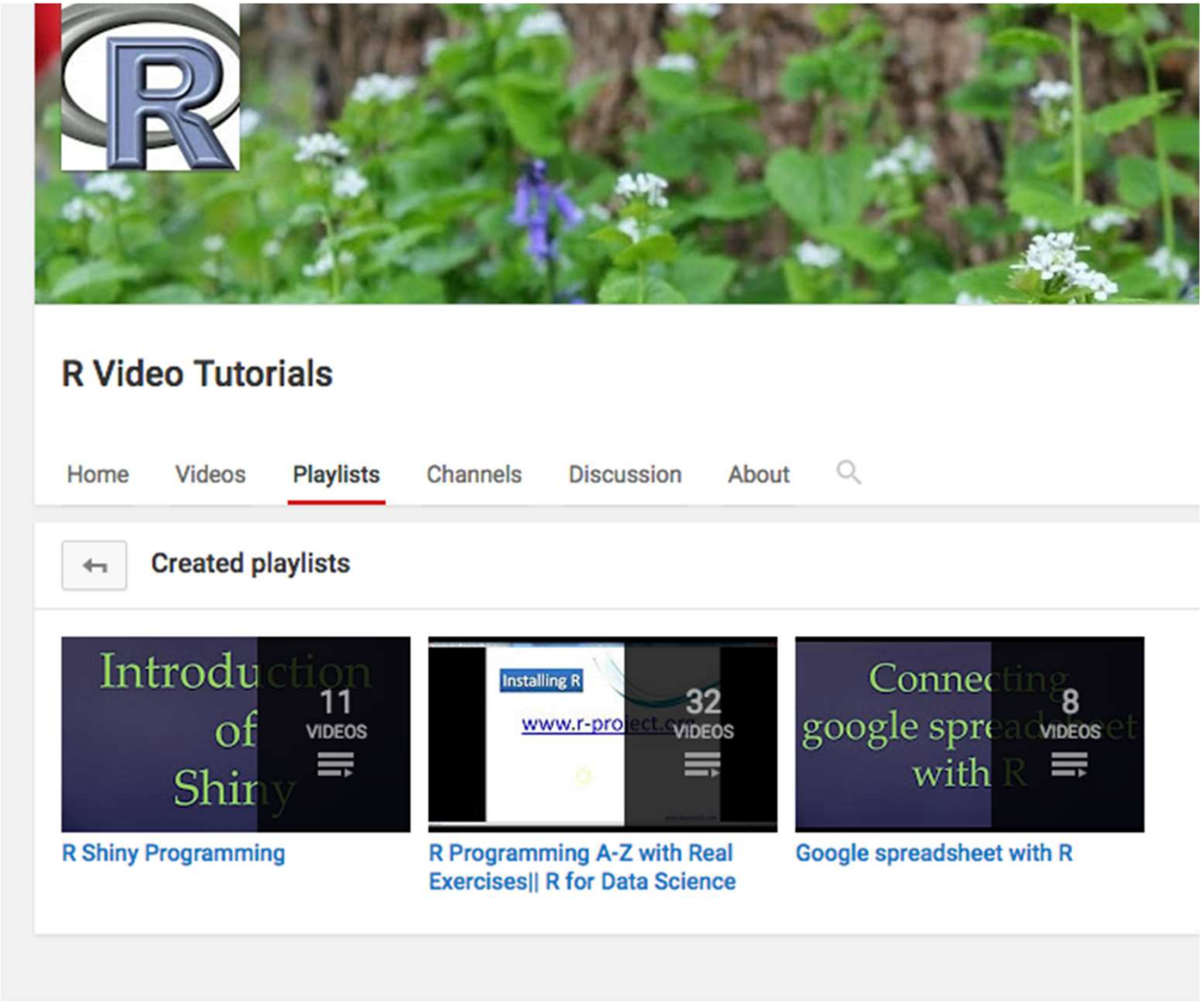
https://www.edx.org/course/programming-python-data-science-microsoft-dat210x-3

10 Microsoft Courses toward **“Professional Program Certificate in Data Science”**

https://academy.microsoft.com/en-us/professional-program/data-science?wt.mc_id=DX_877928&WT.srch=1

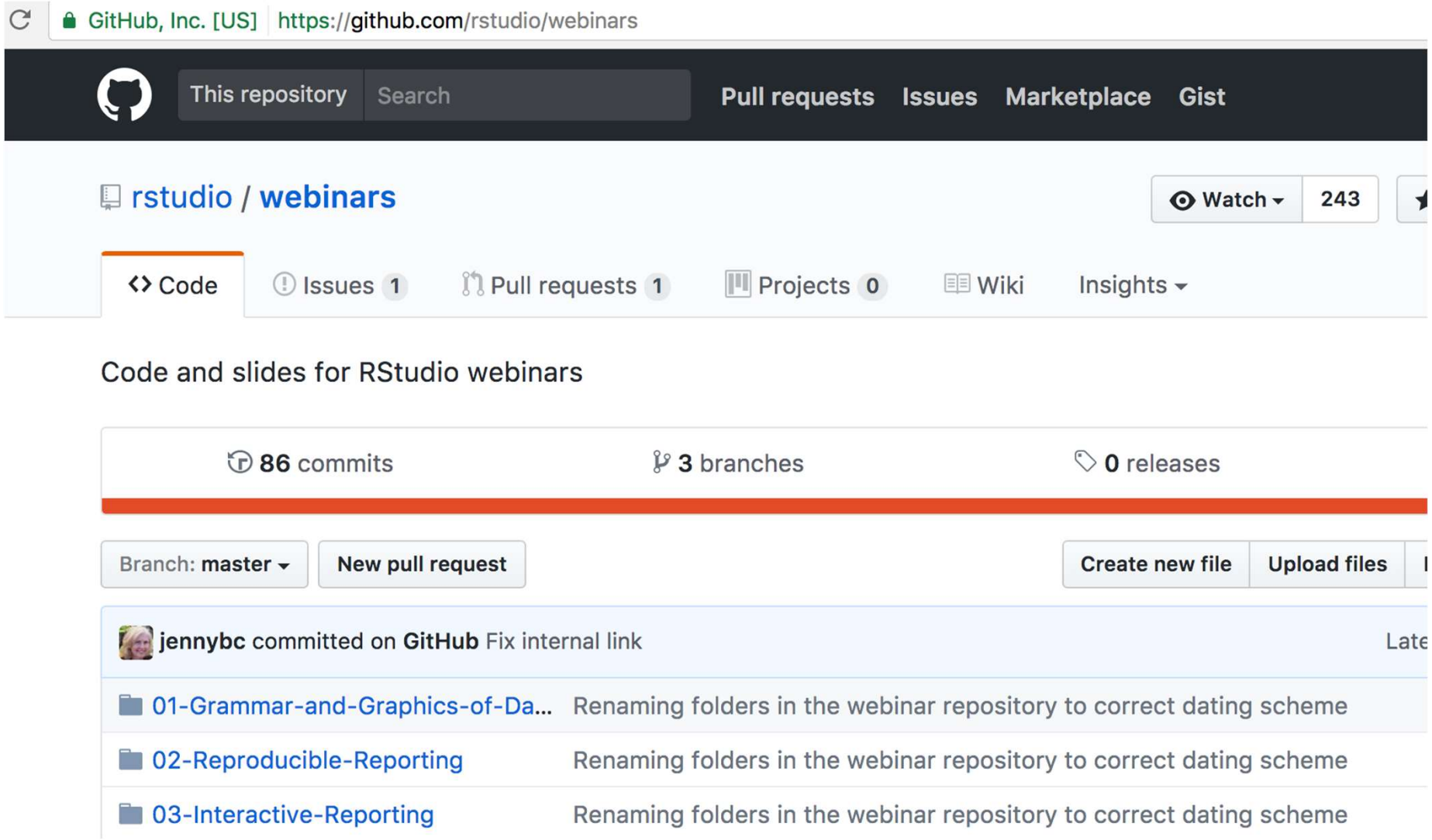
Other R Training

- On YouTube, search for the “R Video Tutorials” channel



Other R Training

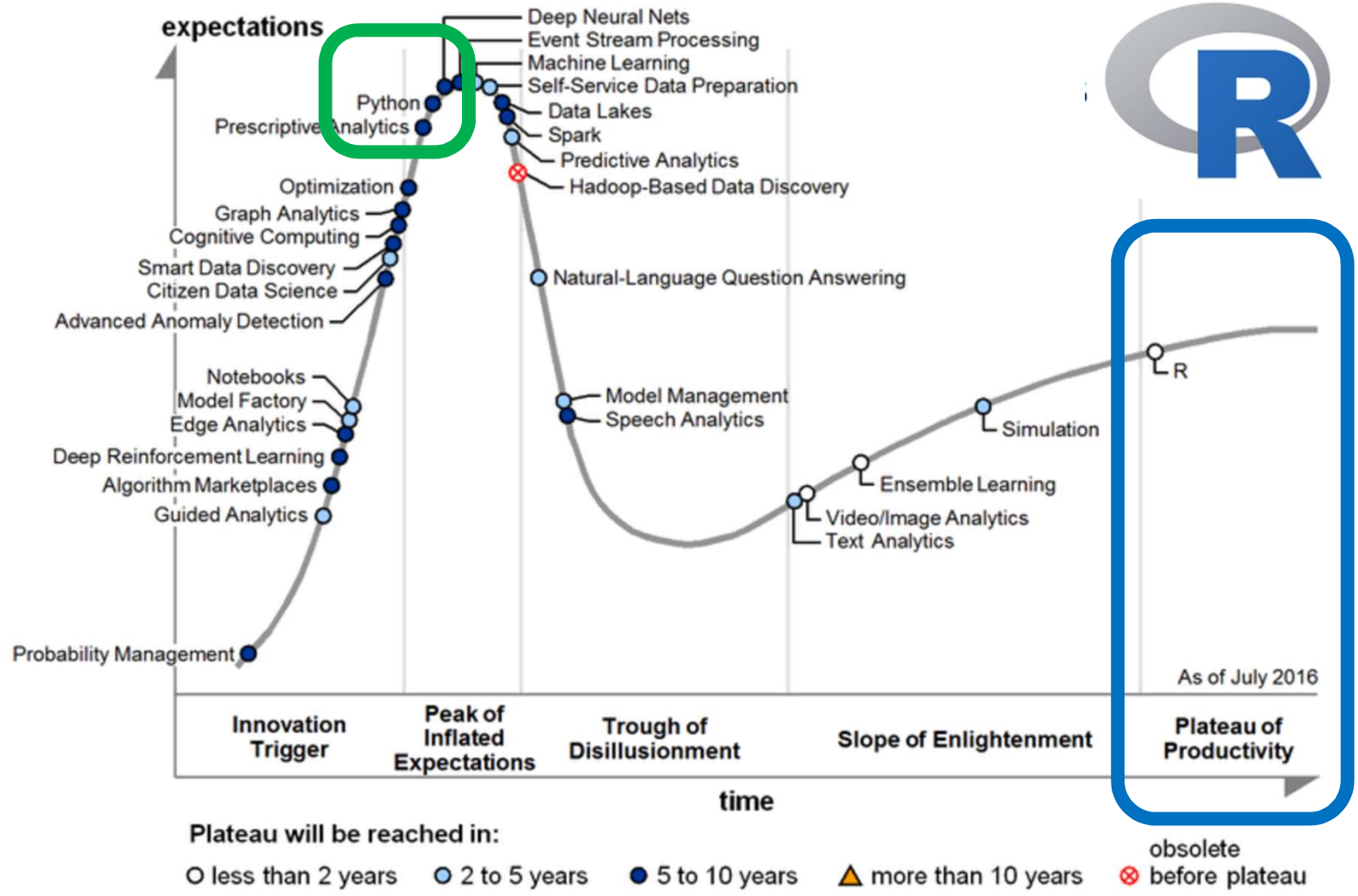
https://github.com/rstudio/webinars



Why R for Data Science?

Gartner Group's report on Hype Cycle in Data Science, in 2016

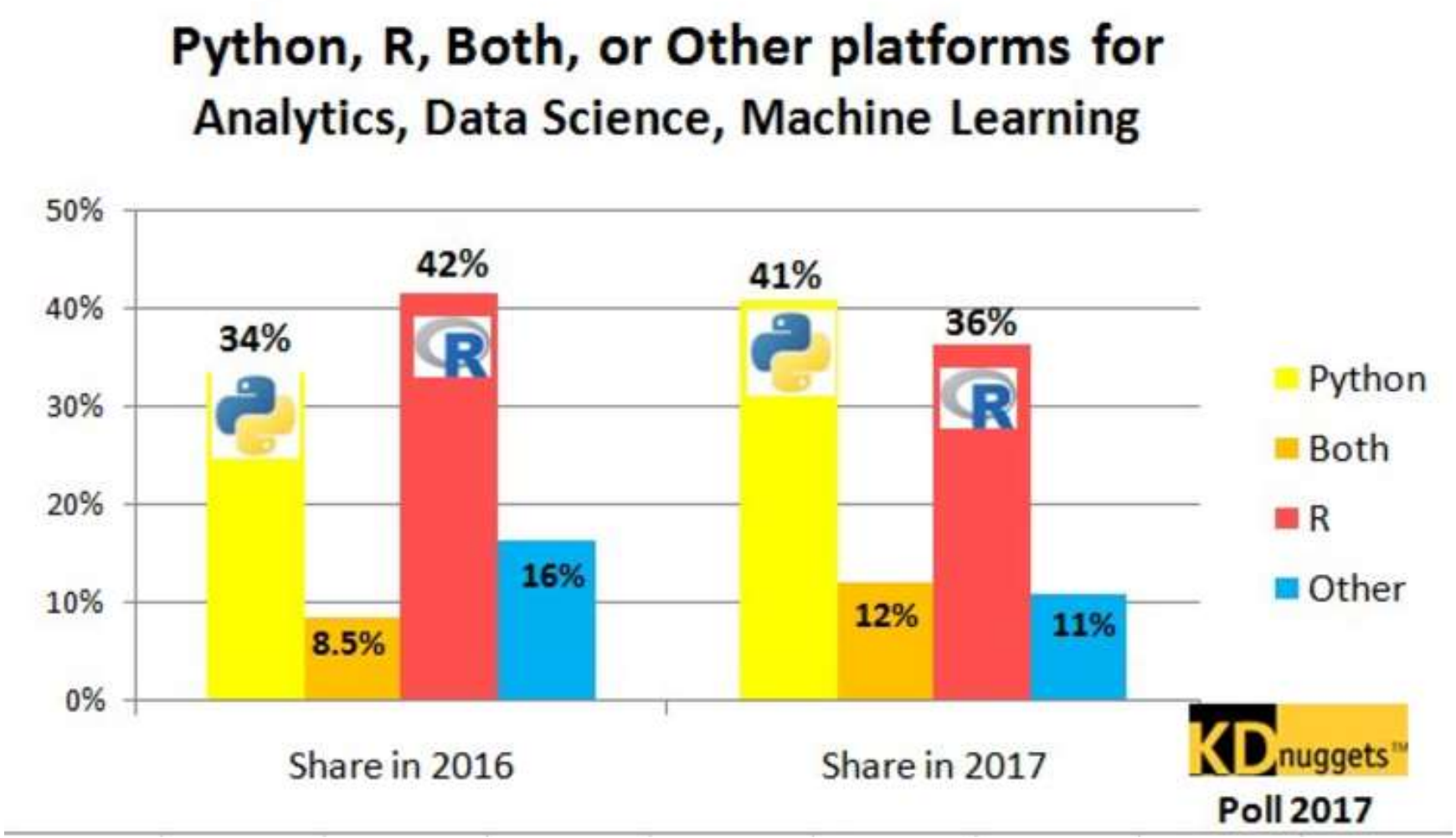
Figure 1. Hype Cycle for Data Science, 2016



Source: Gartner (July 2016)

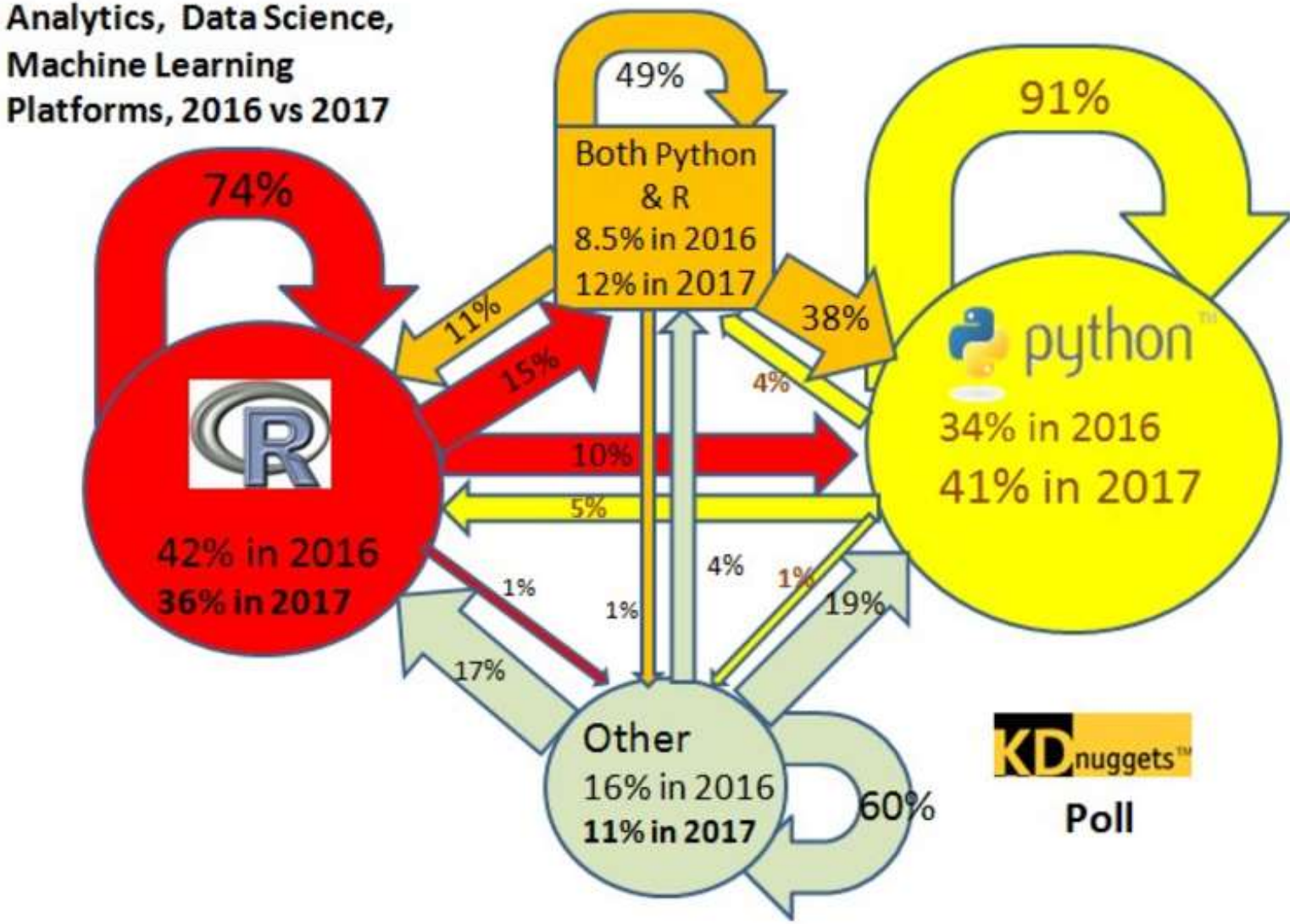
https://thomaswdinsmore.com/2017/02/14/spark-is-the-future-of-analytics/2016-hype-cycle-for-data-science/

Why R for Data Science?



http://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html

Why R for Data Science?



http://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html

Why R for Data Science?

- Why not both R and Python? (a decent idea)
- R is "by quants for quants", an interface to current research
  - Has **10,000 packages** (libraries) on CRAN alone (as of Jan 2017)

    http://blog.revolutionanalytics.com/2017/01/cran-10000.html (more on github, Microsoft,…)
  - Historically, was designed for ease for the analyst, all data was in memory only.

  - R was started in 1995 as an implementation of the language "S", by Ross Ihaka and Robert Gentleman
  - S was started in 1975-76 at Bell Laboratories, mostly by John Chambers, as well as Rick Becker and Allan Wilks
  - There is a commercial version of S, S-Plus, now sold by Tibco Software
  - Many recent improvements for efficiency and scalability

Primary Deep Learning libraries are in Python
- TensorFLow
- Theano

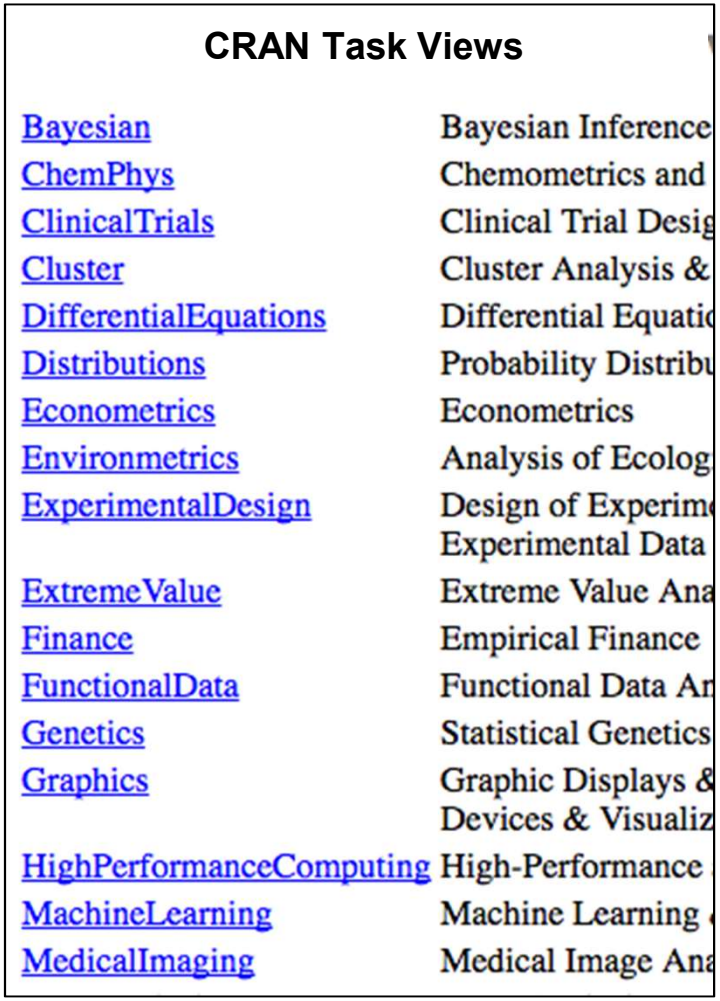
There is now TensorFlow for R - to connect to Python libraries

https://rstudio.github.io/tensorflow/

Getting Started in R

Task Views Give a Top Down Approach

- Install R (search with "R" in Google)
  https://cran.r-project.org/
- Install Rstudio, IDE
  https://www.rstudio.com/
- Install packages / libraries
  - See R "CRAN Task Views" – grouping of 100's of libraries by subject
    https://cran.r-project.org/web/views/
  - Machine Learning https://cran.r-project.org/web/views/MachineLearning.html
- Optional – Rattle, model building GUI
  http://rattle.togaware.com/rattle-install-mac.html
  (can be difficult install on Mac)
  http://marcoghislanzoni.com/blog/2014/08/29/solved-installing-rattle-r-3-1-mac-os-x-10-9/



CRAN Task Views

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and
ClinicalTrials	Clinical Trial Desig
Cluster	Cluster Analysis &
DifferentialEquations	Differential Equatic
Distributions	Probability Distribu
Econometrics	Econometrics
Environmetrics	Analysis of Ecolog
ExperimentalDesign	Design of Experim Experimental Data
ExtremeValue	Extreme Value Ana
Finance	Empirical Finance
FunctionalData	Functional Data An
Genetics	Statistical Genetics
Graphics	Graphic Displays & Devices & Visualiz
HighPerformanceComputing	High-Performance
MachineLearning	Machine Learning
MedicalImaging	Medical Image Ana

Getting Started in R Machine Learning **Task View**

Web search "r task view" now, read some, click around

CRAN Task View: Machine Learning & Statistical Learning

Maintainer: Torsten Hothorn
Contact: Torsten.Hothorn at R-project.org
Version: 2017-04-19
URL: https://CRAN.R-project.org/view=MachineLearning

Why I get Excited about R, 10,000 libraries !!!

Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field machine learning. The packages can be roughly structured into the following topics:

- *Neural Networks and Deep Learning* : Single-hidden-layer neural network are implemented in package nnet (shipped with base the Stuttgart Neural Network Simulator (SNNS). An interface to the FCNN library allows user-extensible artificial neural netwo recurrent neural networks. Packages implementing deep learning flavours of neural networks include darch (restricted Boltzman (feed-forward neural network, restricted Boltzmann machine, deep belief network, stacked autoencoders), RcppDL (denoising a restricted Boltzmann machine, deep belief network) and h2o (feed-forward neural network, deep autoencoders).
- *Recursive Partitioning* : Tree-structured models for regression, classification and survival analysis, following the ideas in the CA (shipped with base R) and tree. Package rpart is recommended for computing CART-like trees. A rich toolbox of partitioning alg RWeka provides an interface to this implementation, including the J4.8-variant of C4.5 and M5. The Cubist package fits rule-bas regression models in the terminal leaves, instance-based corrections and boosting. The C50 package can fit C5.0 classification tr versions of these.
  Two recursive partitioning algorithms with unbiased variable selection and statistical stopping criterion are implemented in pack non-parametrical conditional inference procedures for testing independence between response and each input variable whereas n models. Extensible tools for visualizing binary trees and node distributions of the response are available in package party as wel Tree-structured varying coefficient models are implemented in package vcrpart.
  For problems with binary input variables the package LogicReg implements logic regression. Graphical tools for the visualizatic maptree.
  Trees for modelling longitudinal data by means of random effects is offered by package REEMtree. Partitioning of mixture mod Computational infrastructure for representing trees and unified methods for predition and visualization is implemented in partyk evtree to implement evolutionary learning of globally optimal trees. Oblique trees are available in package oblique.tree.
- *Random Forests* : The reference implementation of the random forest algorithm for regression and classification is available in p bagging for regression, classification and survival analysis as well as bundling, a combination of multiple models via ensemble l

Help and R Examples (Vignettes)

Vignettes and Code Demonstrations: `browseVignettes()`, `vignette()` and `demo()`

Many packages include *vignettes*, which are discursive documents meant to illustrate and explain facilities in the package. You can discover vignettes by accessing the help page for a package, or via the `browseVignettes()` function: the command `browseVignettes()` opens a list of vignettes from *all* of your installed packages in your browser, while `browseVignettes(package=package-name)` (e.g., `browseVignettes(package="survival")`) shows the vignettes, if any, for a particular package. `vignette()` is employed similarly, but displays a list of vignettes in text form.

Getting Help with R

Helping Yourself

Before asking others for help, it's generally a good idea for you to try to help yourself. R includes extensive facilities for accessing documentation and searching for help. There are also specialized search engines for accessing information about R on the internet, and general internet search engines can also prove useful (see below).

R Help: `help()` and `?`

The `help()` function and `?` help operator in R provide access to the documentation pages for R functions, data sets, and other objects, both for packages in the standard R distribution and for contributed packages. To access documentation for the standard `lm` (linear model) function, for example, enter the command `help(lm)` or `help("lm")`, or `?lm` or `?"lm"` (i.e., the quotes are optional).

Vignettes found by "browseVignette

Vignettes in package amap

- Introduction to amap - PDF source R code

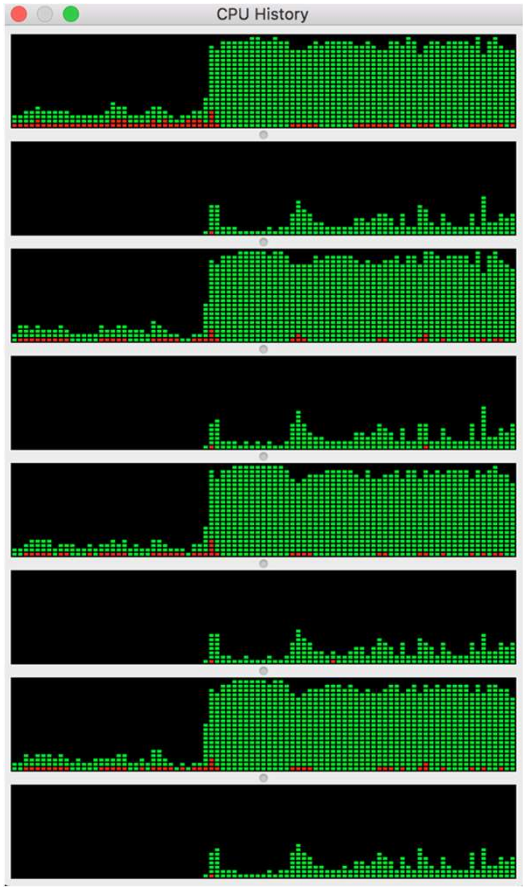
Vignettes in package arules

- Introduction to arules - PDF source R code

Vignettes in package arulesViz

- Visualizing Association Rules: Introduction to arulesViz - PD

R Support for Multicore
Parallel Processing on one Node
Keeps all 4 cores busy
caret training on dozens or 100's of models



To enable parallel processing, you first need to load an adaptor for the foreach package (doMC, doMPI, doParallel, doRedis, doRNG or doSNOW))

```
library(doMC)          # Unix/Mac: parallel for loop processing &
other DoMC(cores = 4)      # for MacPro.  Watch CPU load with
"activity monitor"

library(doParallel)    # for Windows
```

doFuture – a universal foreach adaptor ready to be used by 1000+ packages (multi-core, ad-hoc clusters, cloud clusters…)

Comparison of R to Python's Pandas

https://pandas.pydata.org/pandas-docs/stable/comparison_with_r.html

Querying, Filtering, Sampling

R (data.frames, not data.tables)	pandas
`dim(df)`	`df.shape`
`head(df)`	`df.head()`
`slice(df, 1:10)`	`df.iloc[:9]`
`filter(df, col1 == 1, col2 == 1)`	`df.query('col1 == 1 & col2 == 1')`
`df[df$col1 == 1 & df$col2 == 1,]`	`df[(df.col1 == 1) & (df.col2 == 1)]`
`select(df, col1, col2)`	`df[['col1', 'col2']]`
`select(df, col1:col3)`	`df.loc[:, 'col1':'col3']`
`select(df, -(col1:col3))`	`df.drop(cols_to_drop, axis=1)` but see [1]
`distinct(select(df, col1))`	`df[['col1']].drop_duplicates()`
`distinct(select(df, col1, col2))`	`df[['col1', 'col2']].drop_duplicates()`
`sample_n(df, 10)`	`df.sample(n=10)`
`sample_frac(df, 0.01)`	`df.sample(frac=0.01)`

[1] R's shorthand for a subrange of columns (`select(df, col1:col3)`) can be approached cleanly in pandas, if you have the list of columns, for example `df[cols[1:3]]` or `df.drop(cols[1:3])`, but doing this by column name is a bit messy.

Sorting

R	pandas
`arrange(df, col1, col2)`	`df.sort_values(['col1', 'col2'])`
`arrange(df, desc(col1))`	`df.sort_values('col1', ascending=False)`

Overview of RStudio

variables, lists, data.tables

- Details here…

HMEQ (Home Equity) Data

Line of credit loan application, using existing home as loan equity.
Use for home improvements, or combine bills together
5,960 records, 20% defaulted after 1 year

DATA ROLE	COLUMN	DESCRIPTION
Key	rec_ID	Record ID or key field, for each line of credit loan or person
Target	BAD	After 1 year, loan went in default, (=1, 20%) vs. still being paid (=0)
Applicant	CLAGE	Credit Line Age, in months (for another credit line)
Applicant	CLNO	Credit Line Number
Applicant	DEBTINC	Debt to Income ratio
Applicant	DELINQ	Number of delinquent credit lines
Applicant	DEROG	Number of major derogatory reports
Applicant	JOB	Job, 6 occupation categories
Loan applic	LOAN	Requested loan amount
Property	MORTDUE	Amount due on existing mortgage
Applicant	NINQ	Number of recent credit inquiries
Loan applic	REASON	"DebtCon" = debt consolidation, "HomeImp" = home improvement
Property	VALUE	Value of current property
Applicant	YOJ	Years on present job

HMEQ (Home Equity) Data

Lab 1a: Exploratory Data Analysis

- Switch to code in Rstudio

Compare Algorithm Families

by the number of inputs and outputs in practical use

A high level way to figure out what family of algorithms can be applied to a given problem

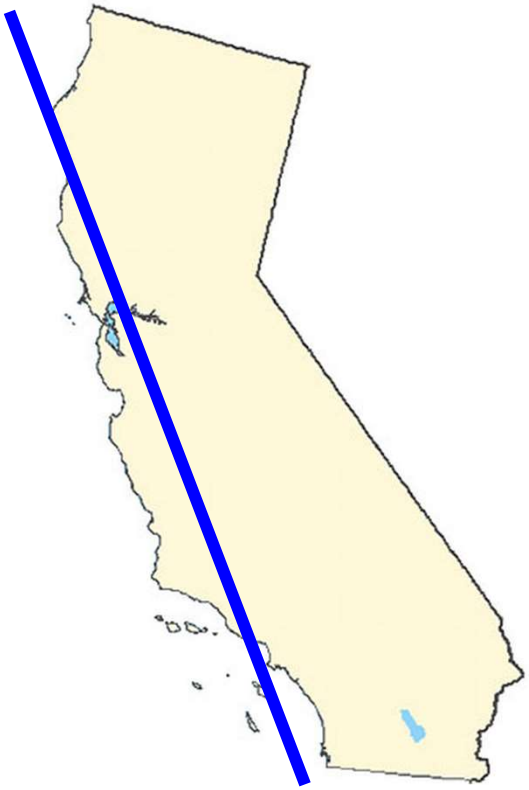
	Inputs	Outputs	Clusters Found, Internal States or Factors
Prediction, Forecasting	10 - 500	1	
Classificaiton	10 - 500	1, category	
Regression	10 - 500	1, number	
Classification: text mining, genomic	100k - 2mm	1	
Clustering, Outlier Detection	10 - 150	NONE	3 - K's
Association Rules, Sequence Mining	50k - 500k categories	NONE	10k - 100k rules
Recommendor Systems, Collaborative Filtering	100k - 10mm	100k - 10mm	~200 - 300
Deep Learning, Auto-Encoder	50 - 20k	50 - 20k	~200 - 300
Deep Learning, Recurrent Nets	20 - 200's	10 - 100's	hundreds
Deep Learning, Image/Speech	500 - 2k	100	5 - 150 layers
Bayesian Graphical Belief Networks	100's	NONE	

Think of Separating Land vs. Water

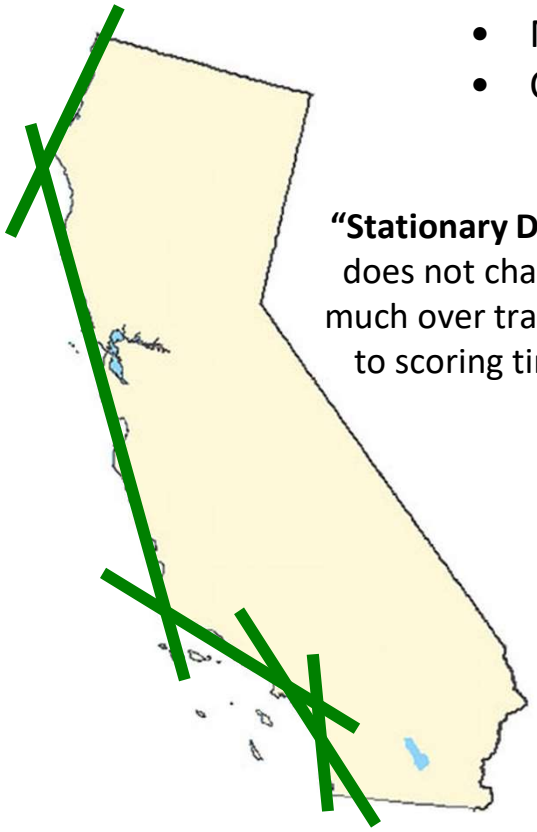
(using 2 input vars: latitude & longitude)

Accurate

1 line,
Regression
(more errors)



5 Hidden Nodes in
a Neural Network

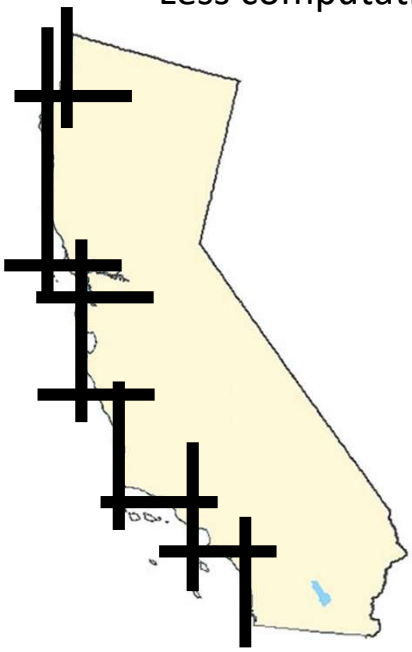


Different algorithms use different **Basis Functions**:

- One line
- Many horizontal & vertical lines
- Many diagonal lines
- Circles

"Stationary Data" does not change much over training to scoring time

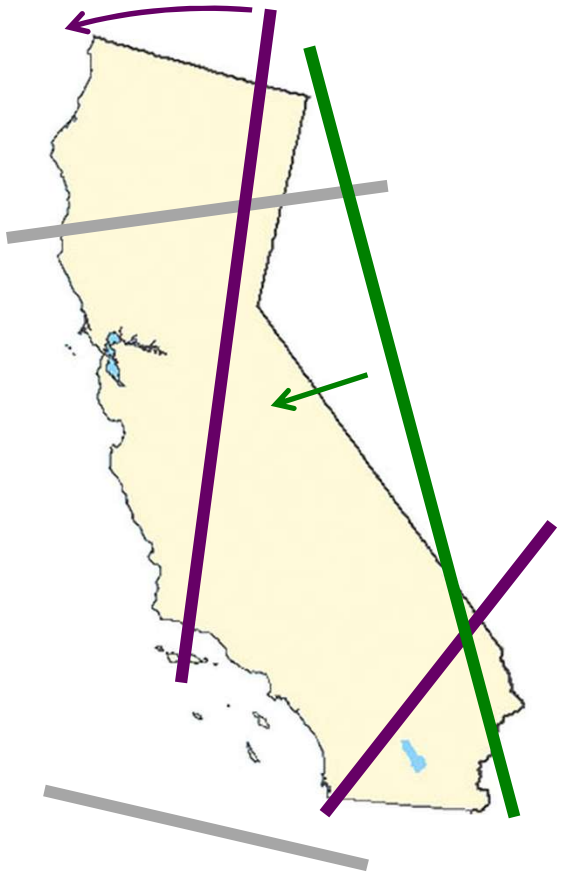
Decision Tree
12 splits
(more elements,
Less computation)



Q) What is too detailed? "Memorizing high tide boundary" …and applying it at all times
… only using training data with photos on sunny days, at similar times of day

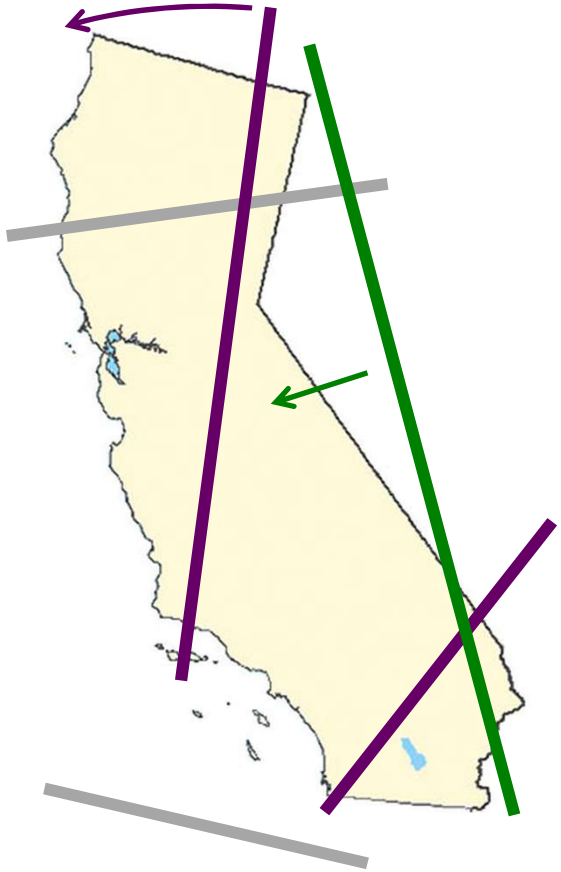
What happens during model training?

Literally, Starts
With Random Values
(just born)

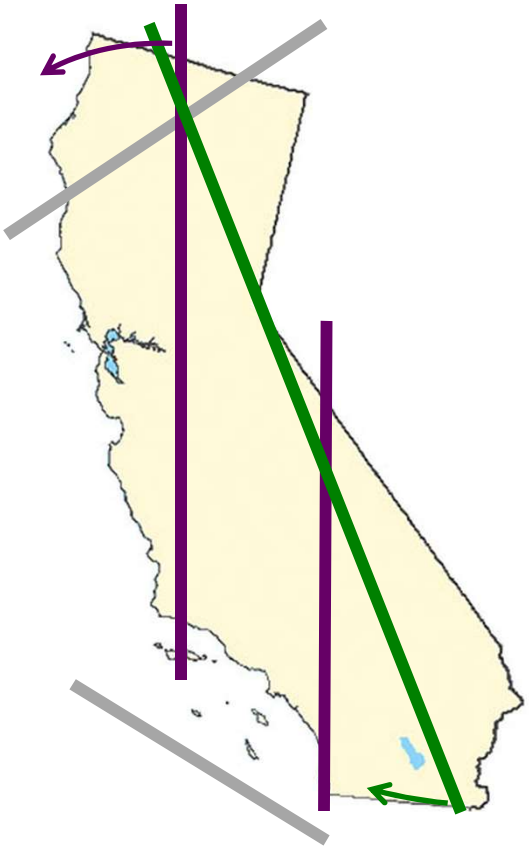


What happens during model training?

Literally, Starts
With Random Values
(just born)



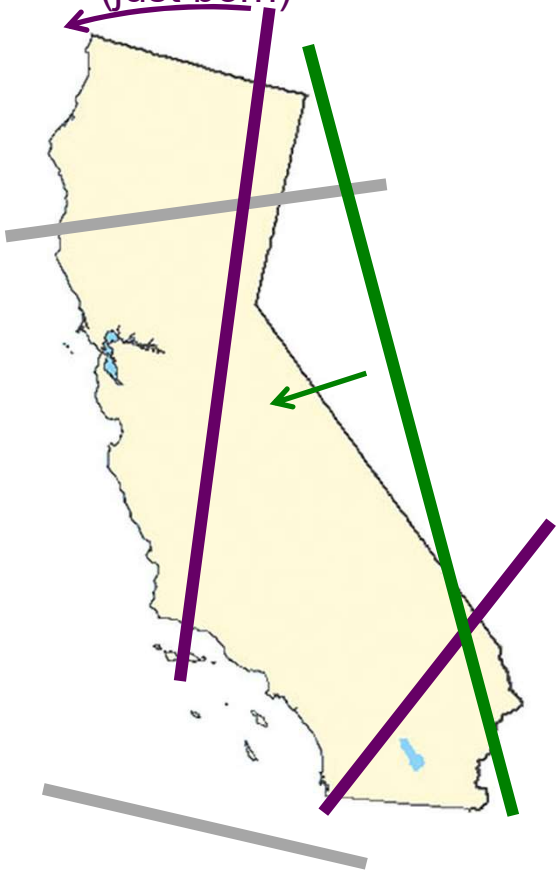
"Babbling" with a
Feedback loop, like a baby



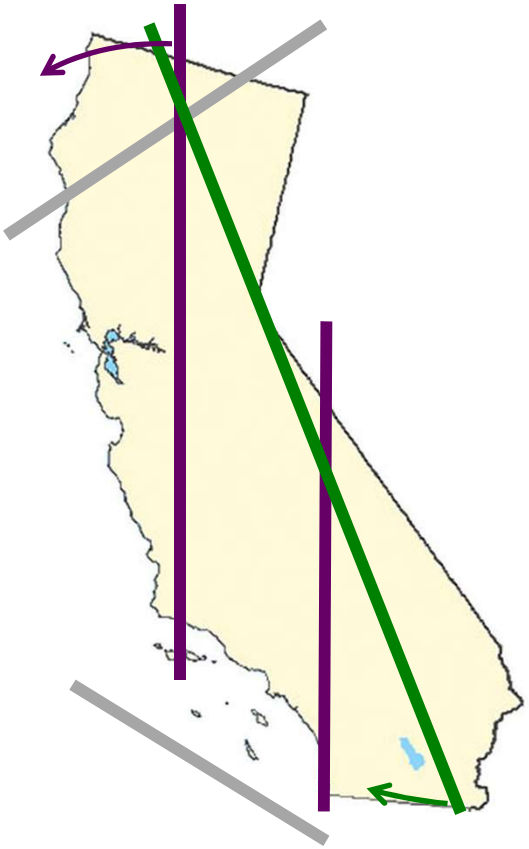
Changes get smaller as approach decent results

What happens during model training?

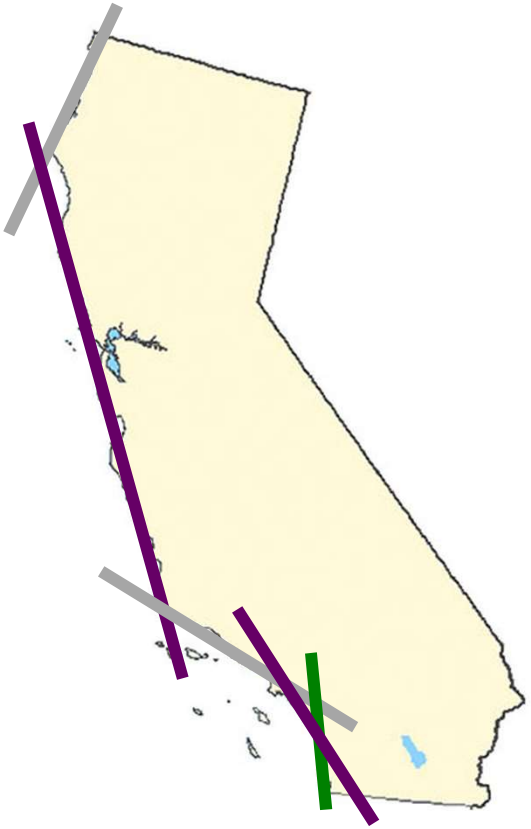
Literally, Starts
With Random Values
(just born)



"Babbling" with a
Feedback loop, like a baby

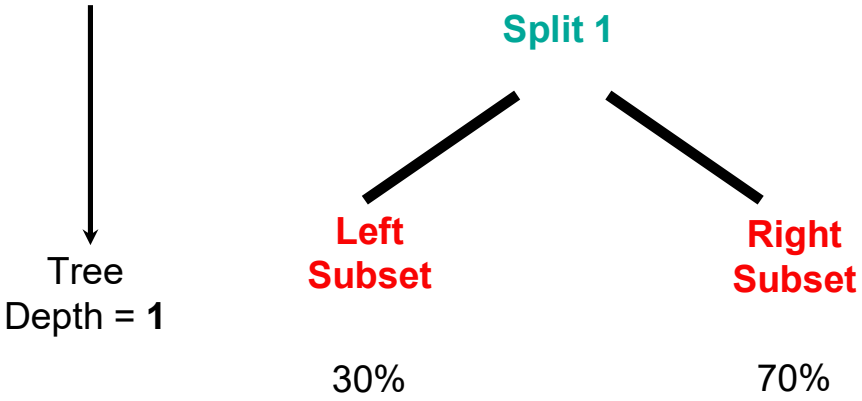


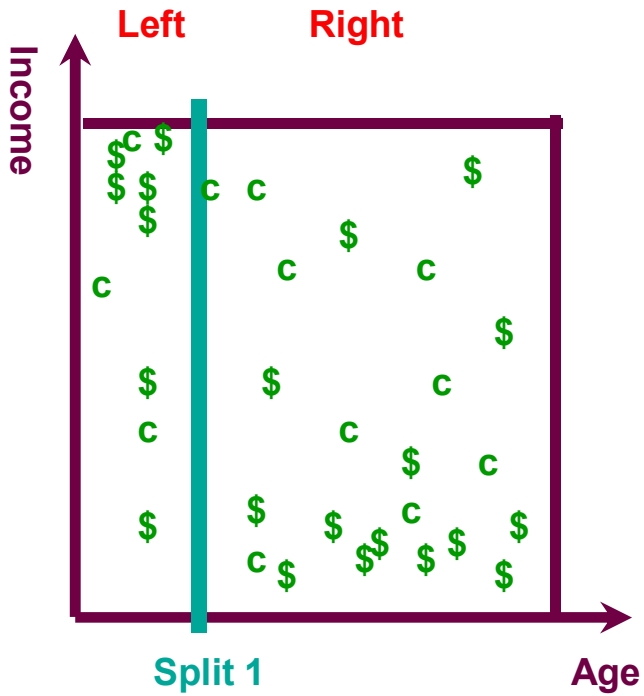
Getting to a functioning
state by
Minimizing error^2



Changes get smaller as approach decent results

What does **Training** a Decision Tree Look Like?





If (Age < Split1) then
:…

Left Subset

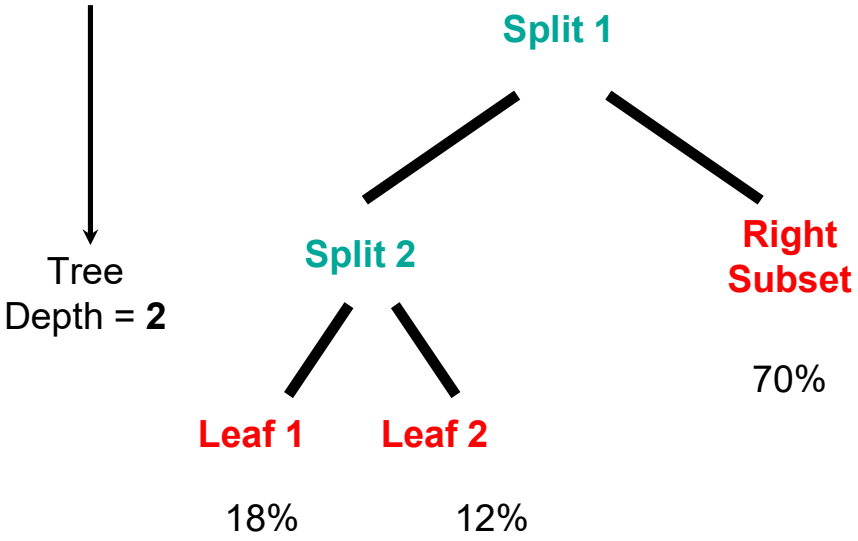
If (Age > Split1) then
:…

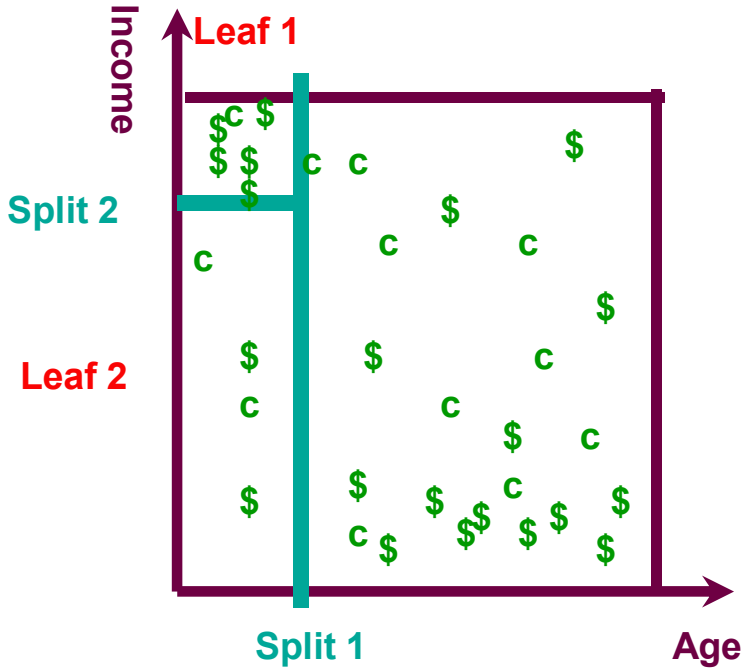
Right Subset

Decision Tree algorithms may find a split optimal to metrics like:

- Entropy
- Gini
- Chi-Square

What does **Training** a Decision Tree Look Like?

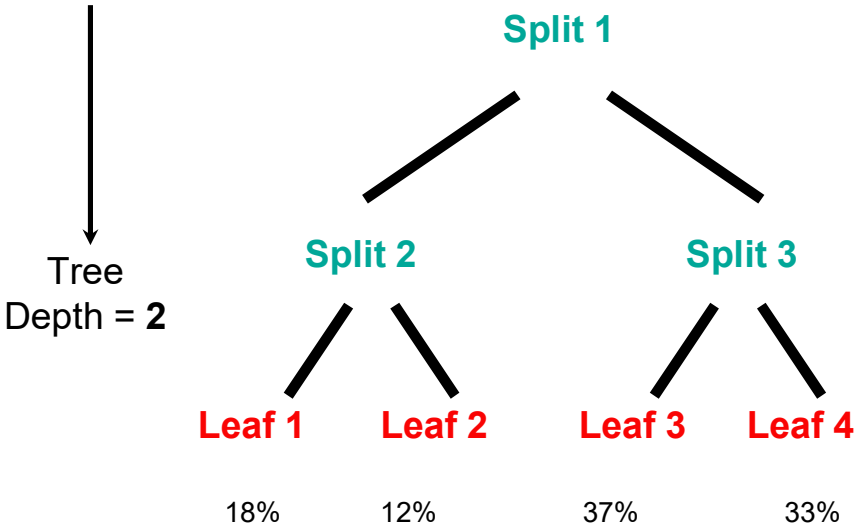


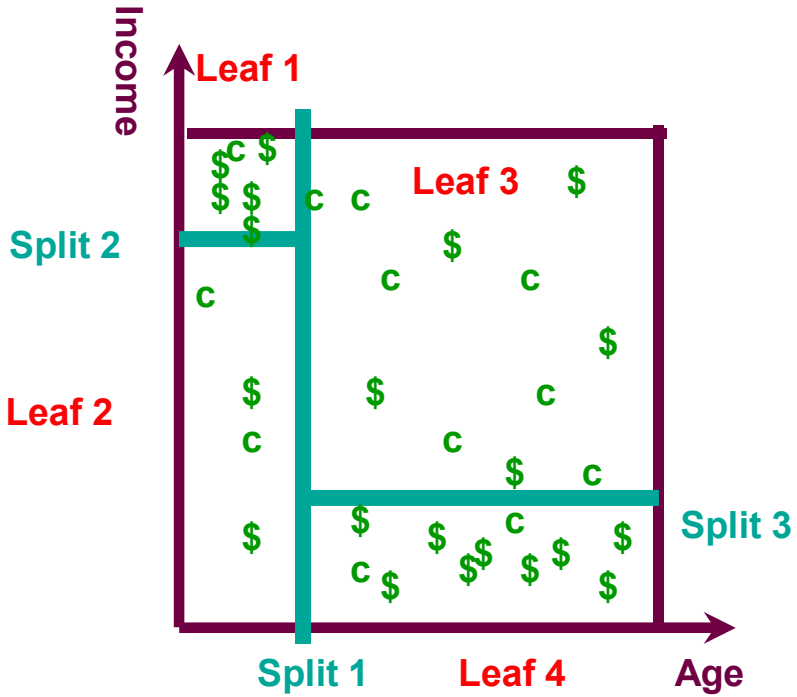


If (Age < Split1) then
:…If (Income > Split2) then Leaf1 with dollar_avg1
:…If (Income < Split2) then Leaf2 with dollar_avg2
If (Age > Split1) then
:…

Right Subset

What does **Training** a Decision Tree Look Like?



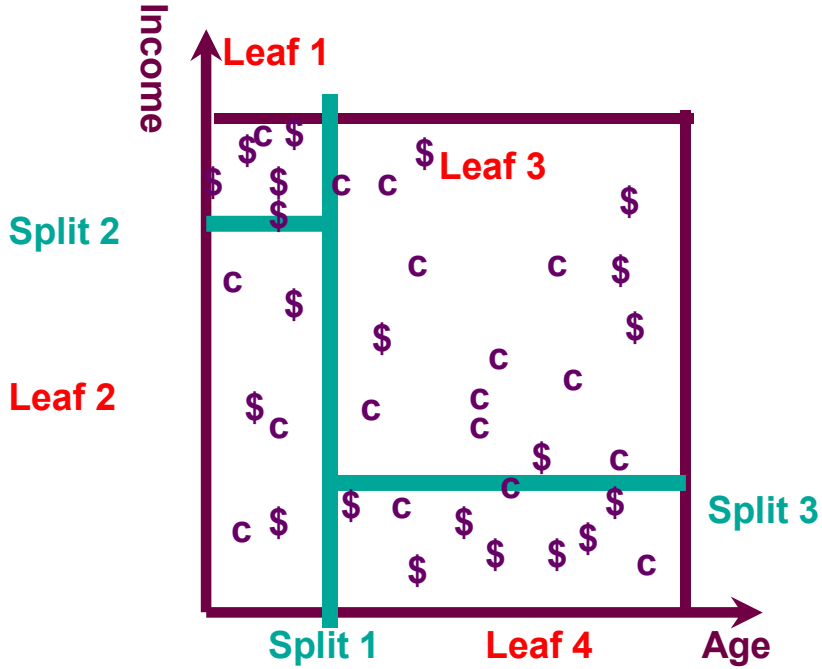


If (**Age** < **Split1**) then
:…If (**Income** > **Split2**) then **Leaf1** with **dollar_avg1**
:…If (**Income** < **Split2**) then **Leaf2** with **dollar_avg2**
If (**Age** > **Split1**) then
:…If (**Income** > **Split3**) then **Leaf3** with **dollar_avg3**
:…If (**Income** < **Split3**) then **Leaf4** with **dollar_avg4**

Scoring Model with a Test Set to Generalize Well

IMPORTANT
When a model is deployed,
It is expected to be general
enough to last awhile

Split all historic data available
70% training, 30% test
60% training, 20%, 20% test 1 & 2
"Science"



Can automatically build detectors for "model drift"

When the behavior represented in the training data drifts from the current production behavior

Lab 1b: TensorFlow Playground

http://playground.tensorflow.org

- If you have less exposure to hands on predictive modeling
- This is used to help give intuition
- The relationship between 2 input fields on data points (records) with 2 different colors (target values)

- The web site is very graphical
- Training time is quick and gives immediate feedback

- Start with the default data set (bulls-eye) and try some training

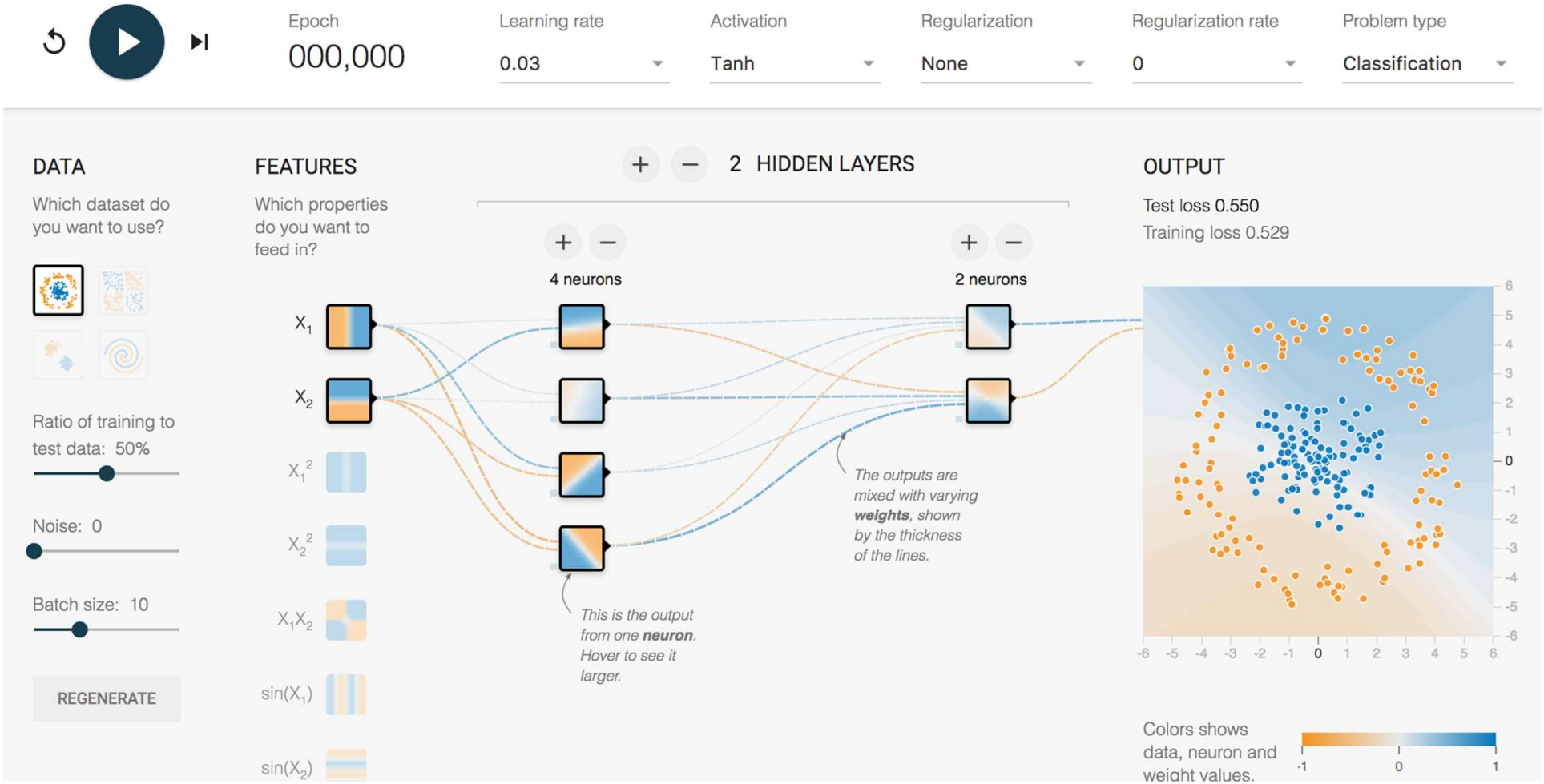
- Then move on to other data sets, (2 spirals)
- Try adding another layer

Lab 1b:

http://playground.tensorflow.org

Play with TensorFlow Neural Net Model Training

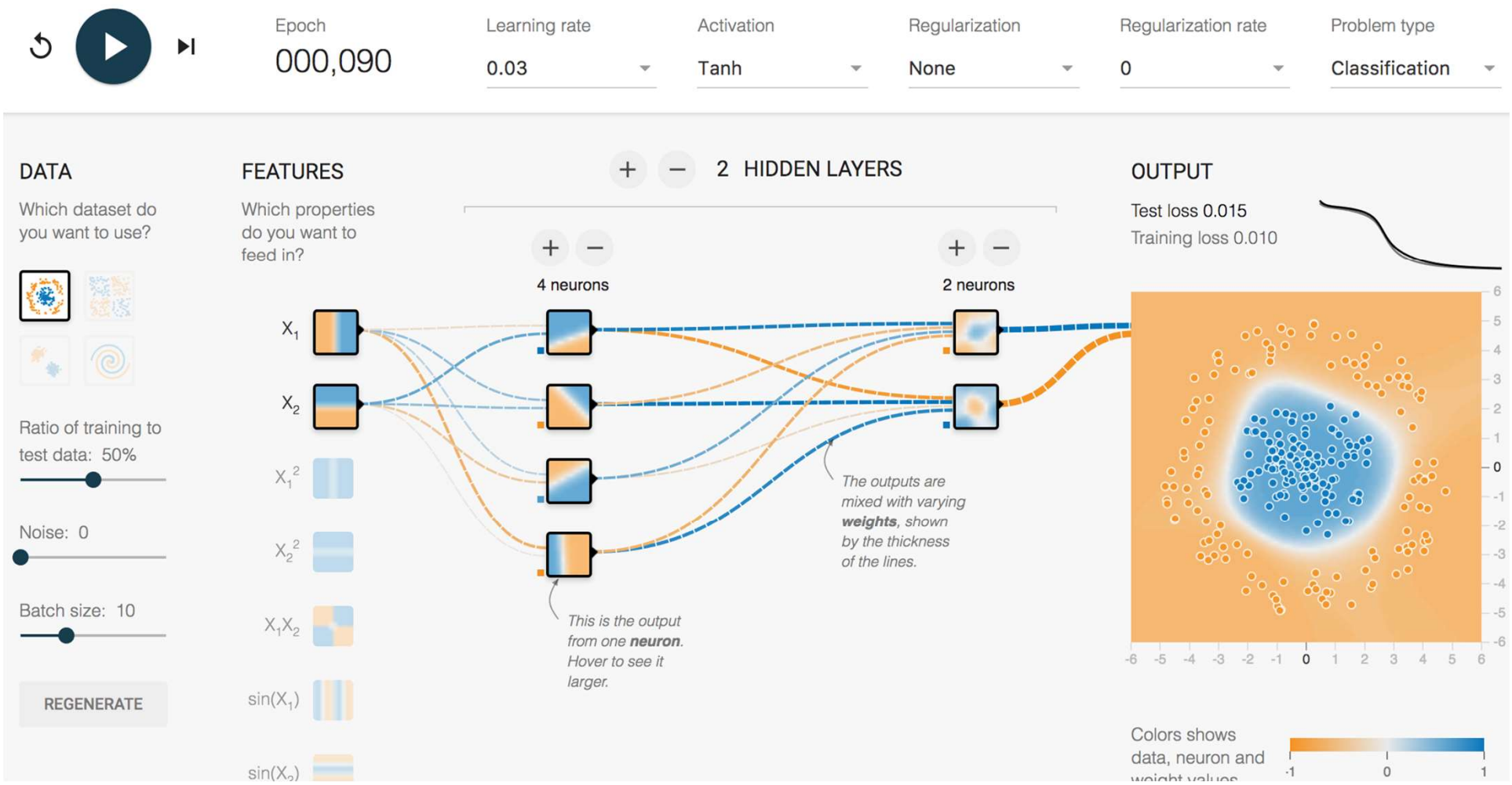
Simple Problem (gold donut) BEFORE training (random wts)



http://playground.tensorflow.org

Play with Model Training

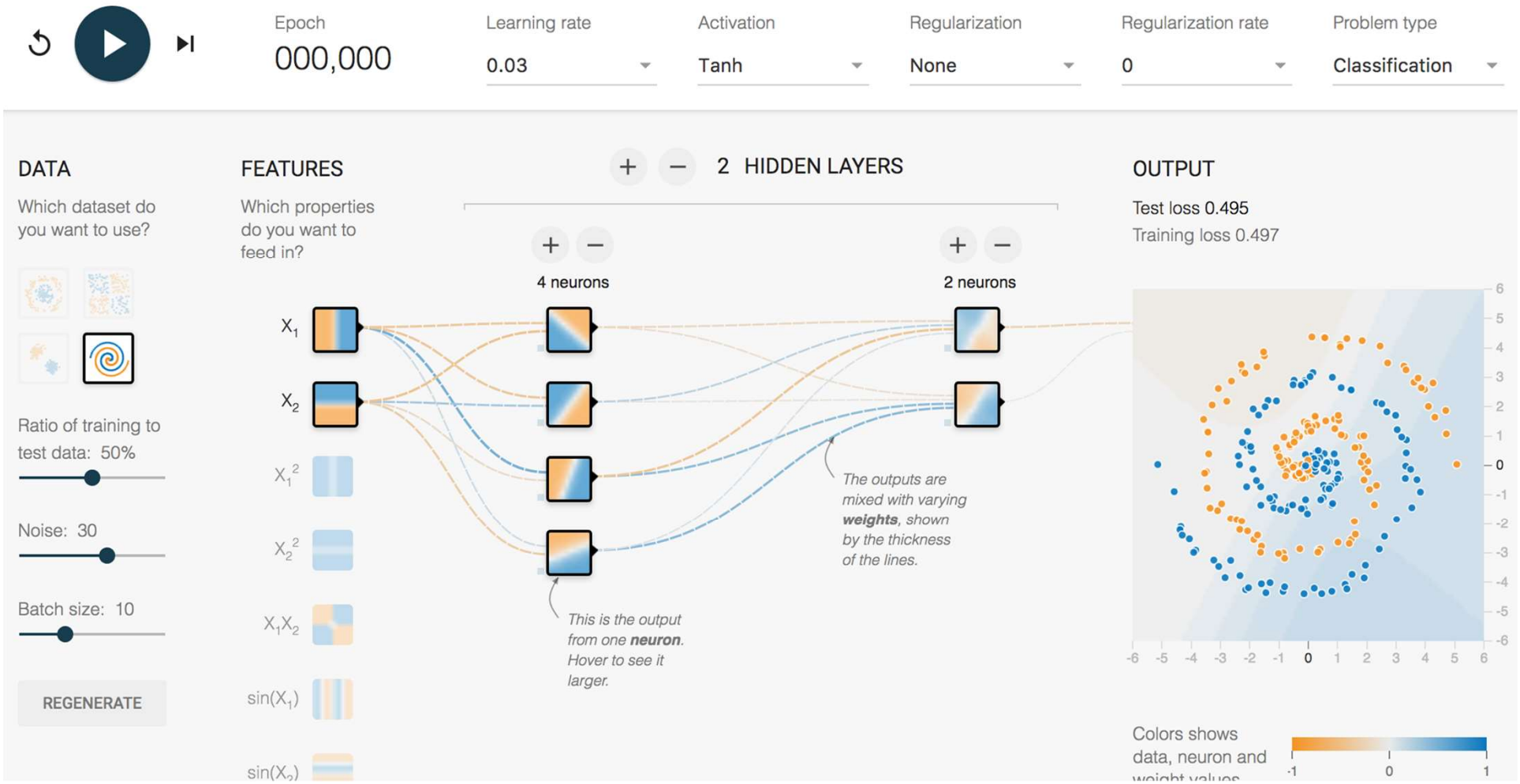
Simple Problem (gold donut) AFTER training



http://playground.tensorflow.org

Play with Model Training

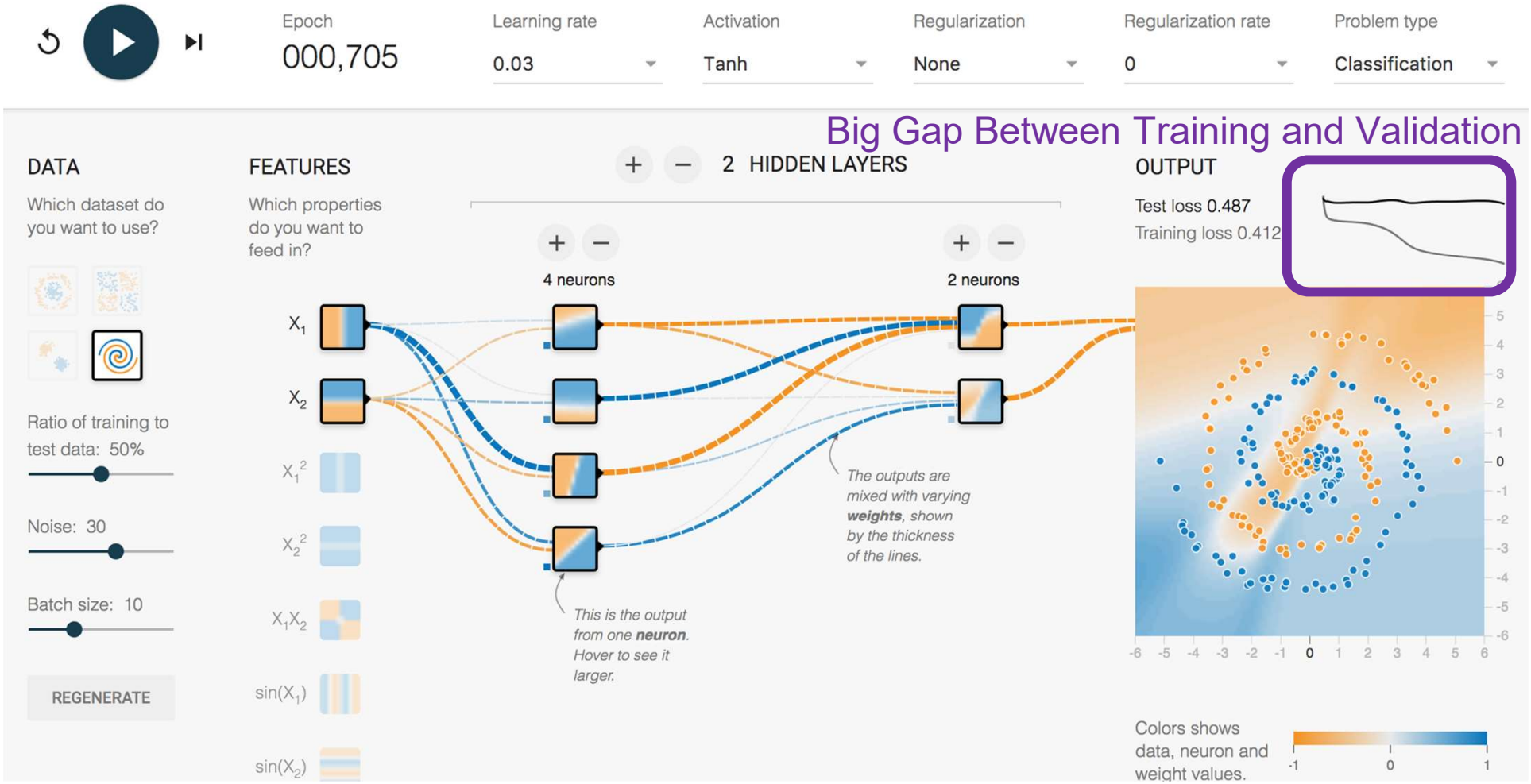
Difficult Problem (2 spirals) BEFORE training



http://playground.tensorflow.org

Play with Model Training

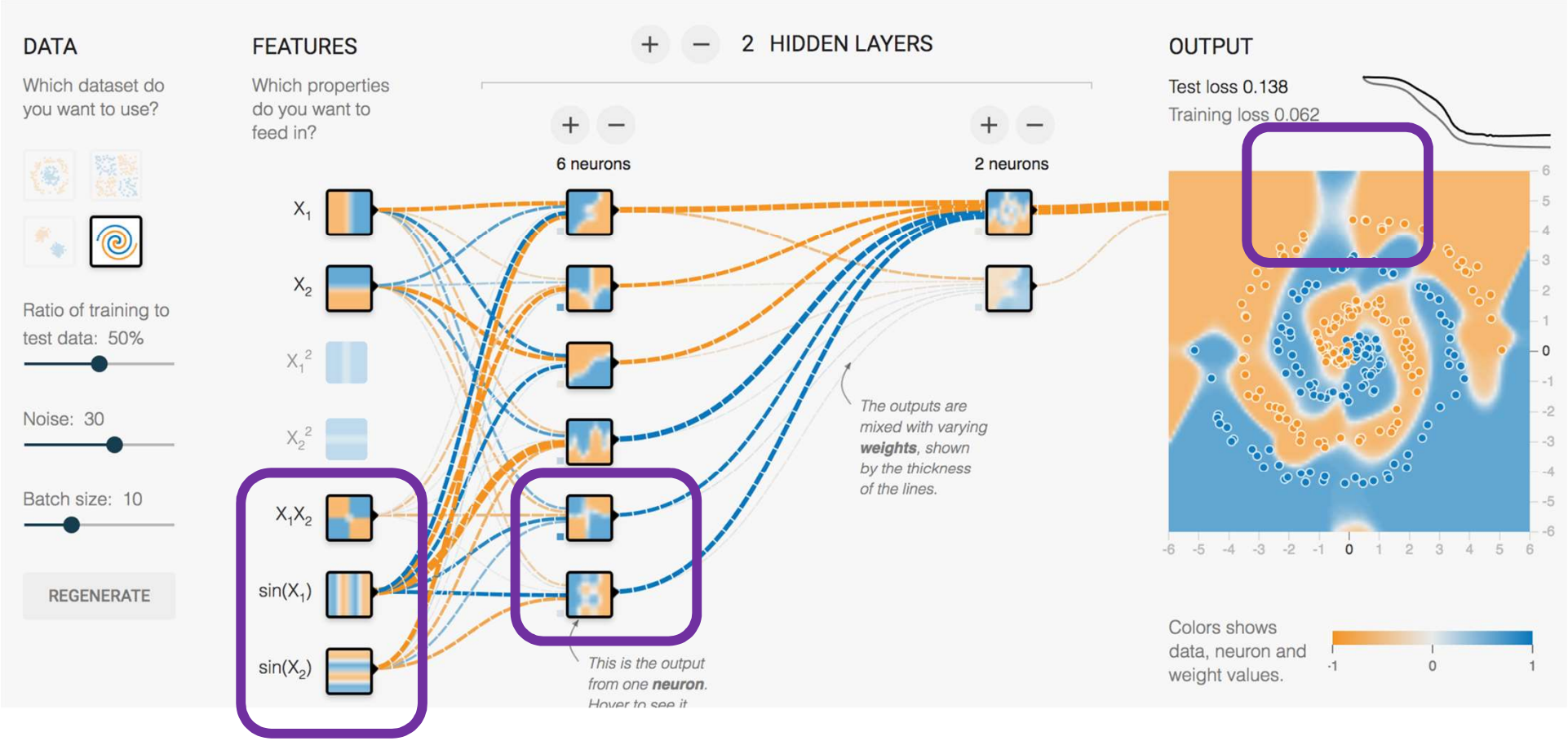
Difficult Problem (2 spirals) AFTER training (kinda stuck)



http://playground.tensorflow.org

Play with Model Training

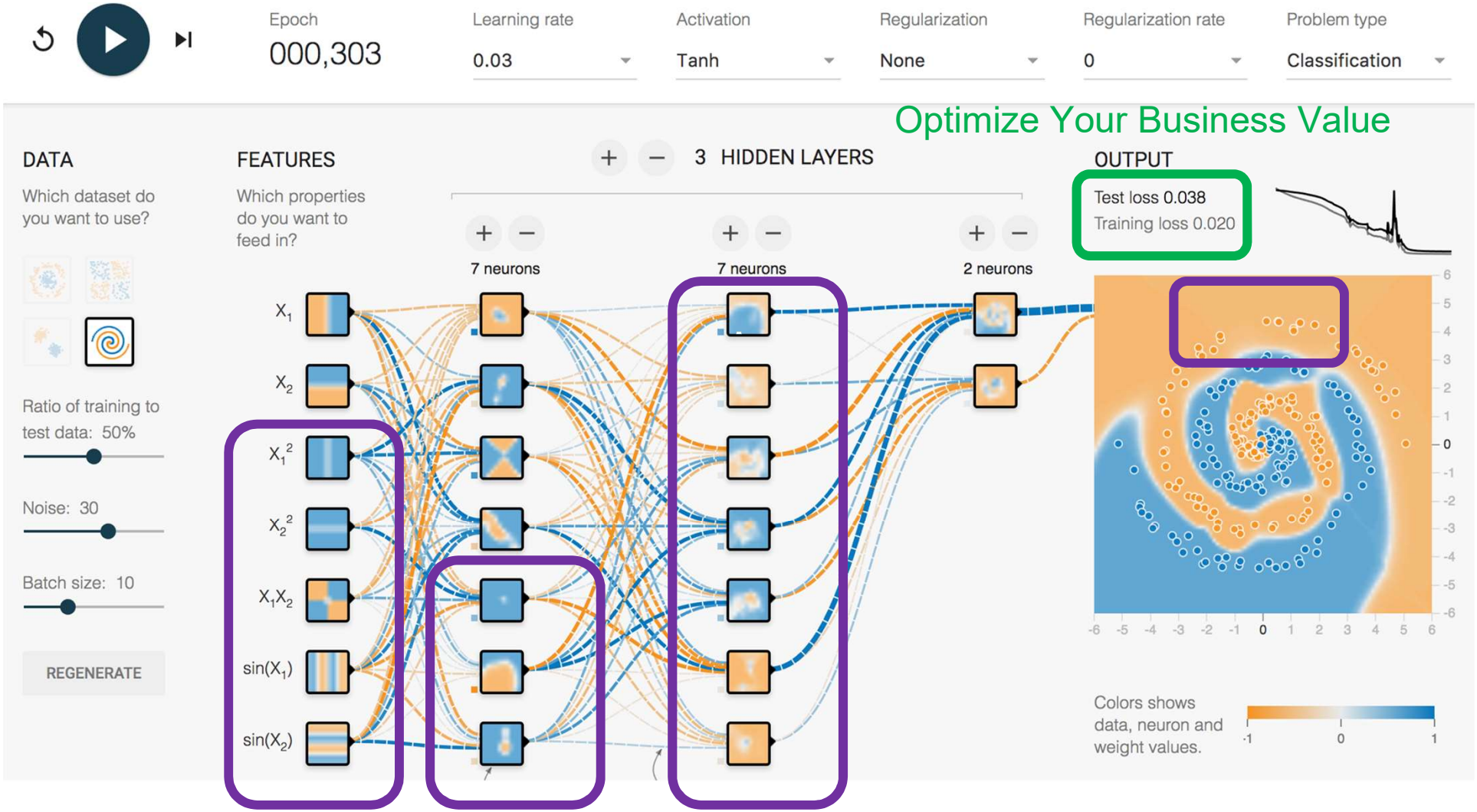
Difficult Problem (2 spirals) AFTER training (better prep)



http://playground.tensorflow.org

Play with Model Training

Difficult Problem (2 spirals) AFTER training (better model)



Outline

Part 1: Get started with R, play with data

Part 2: Data Science Project Design

- Model Evaluation Fundamentals, DS Model Loop Sprints
- Selling Data Mining to Executive Check Writers
- Find candidate projects with a Knowledge Discovery Workshop
- Design the ”Analysis Universe” of data
- Retraining Frequency (daily, monthly or over years)
- Reference Dates & staggering to increase training data variety
- Target & Weight variable variations
- Business Metrics to Optimize
- Big Data production, Lambda & Kappa Architecture
- **Lab 2** lecture / lab: data.tables, select rows, create vars, func

Pitching the Data Mining Project

Now, Repeat after me….

Pitching the Data Mining Project

Give me 2 months, and
I will give you $2 million

If you used 2 months of my salary or consulting fees
and invested in the stock market,
would you get a better return?

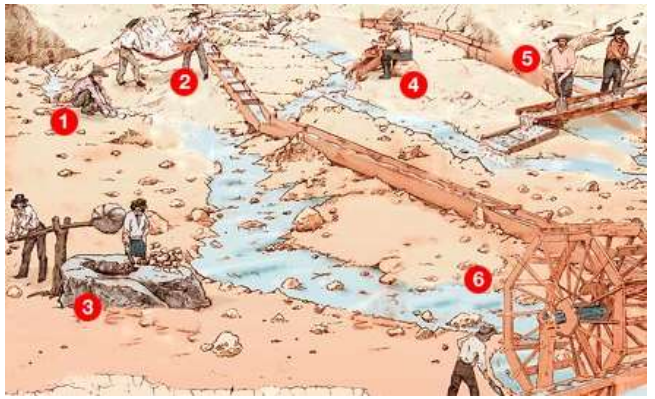
Easiest if offering a mining approach to a problem that did
not previously have data mining applied
(or new approach is non-linear, or much bigger data….)

Pitching the Data Mining Project

Advantage of Picking "New" Projects (Gold Rush Analogy)

http://www.calgoldrush.com/graphics/evolution.html

- In the CA Gold Rush there were "stages" of technology to remove value
  - –Start with panning (placer mining) → nuggets
  - –Panning dries up as a useful activity
  - –Crushing quarts → dust



  - –Rocker with water to process 200 bucket-fulls of earth / day
  - –Waterwheel (more work, less incremental return)
  - –….. Stock market options day trading on gold futures

- –Data Mining is easier, can more reliably provide incremental improvement on "New projects", or with some differentiation
  - –Linear → non-linear Batch → real time Data → 10* or 100* bigger data
  - –Add purchased data, add NLP unstructured data → structured, add Op Res

The Value of Forecasting on a Population

Just Reach out to a Targeted Subset of Customers

"Fishing in a Lake" analogy

- Fishing
  - –Fish in a subset area. It is not practical to take action on all the area
  - –Don't expect to catch all the fish, given practical amount of time
- Without a model


Dreamworks logo

  - –Use judgement or a best guess to pick a spot to go fishing
  - –**Casually catch some fish (which fine for a hobby)**

The Value of Forecasting on a Population

Just Reach out to a Targeted Subset of Customers

"Fishing in a Lake" analogy

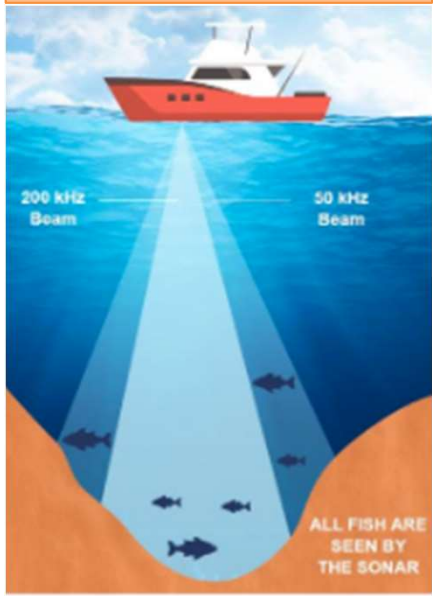
- Fishing
  - –Fish in a subset area. It is not practical to take action on all the area
  - –Don't expect to catch all the fish, given practical amount of time

- Without a model

  - –Use judgement or a best guess to pick a spot to go fishing
  - –Casually catch some fish (which fine for a hobby)

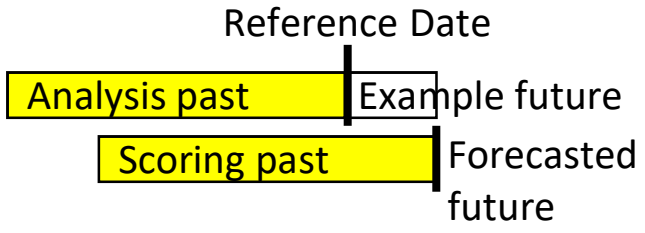
- With a Model **(Professional Fishing Business)**
  - –Like using SONAR to find the highest concentration of fish
  - –SONAR determines which 5% of the lake to target
  - –Catch many more fish per hour (per fixed cost)

Build a "Sonar map" or model in March, use in spring to fall (generalization)



Data Science Project & Model Loop Sprints

Define business objectives and project plan during the Knowledge Discovery Workshop



Select the **"Analysis Universe"** data

Include holdout verification data (~10 days)

Repeat through **model loop** (1-3 times, ~2 weeks each)

Days per sprint		
2	1	1
5	4	4
2	4	3
1	1	2

Exploratory Data Analysis (EDA)

Transformation (Preprocessing)

Build Model – dozens or 100's of models **(Data Mining)**

Evaluate and explain the model – use business metric

Score or deploy the model on "Forecast Universe"

Track results, refresh or rebuild model, subdivide or refine as needed

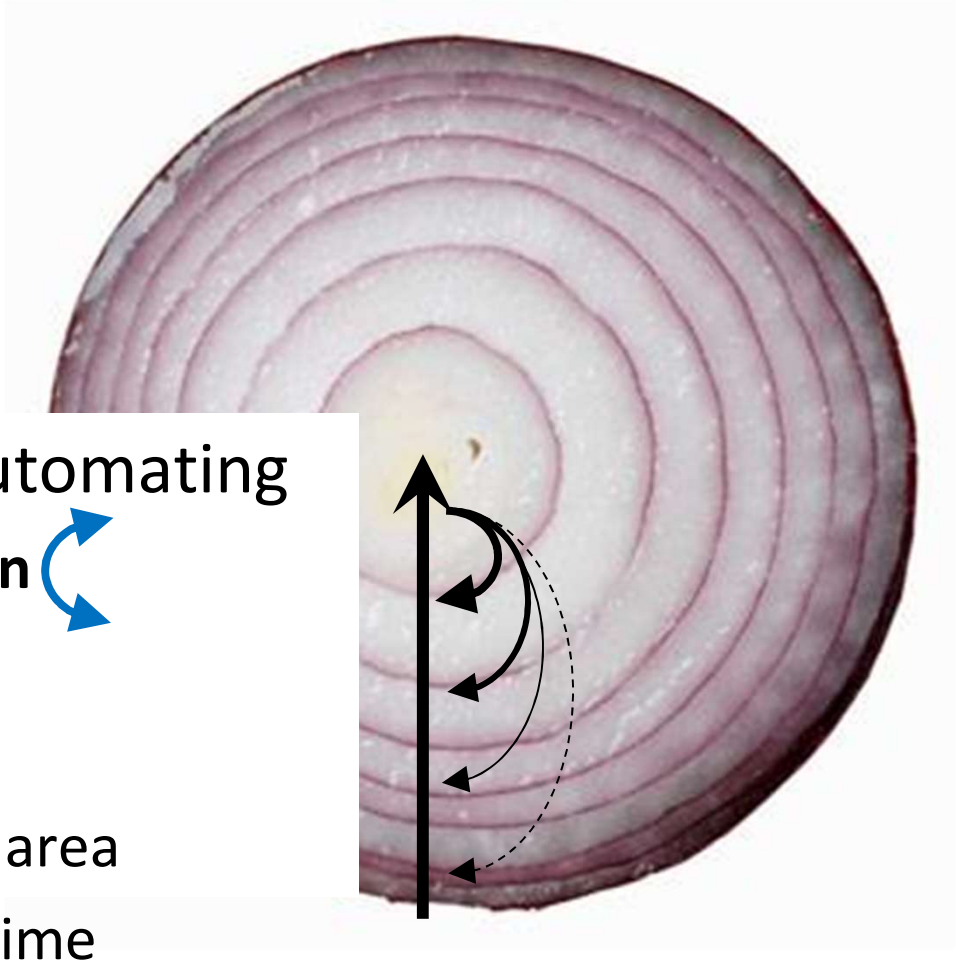
From Data Mining to Knowledge Discovery in Databases, 1996
https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131

Easiest to Automate From the Core

Experience from Embedding Automatic Data Mining in Many Enterprise Applications

Cycle through

- **Measure**
- **Manipulate**



Go through full process, automating

- **model building & evaluation**
- EDA & Preprocessing

The “core” has the smallest area

… and the smallest project time

Knowledge Discovery in Databases (KDD)

"The mere formulation of a problem is far more often essential than its solution, which may be merely a matter of mathematical or experimental skill.

To raise new questions, new possibilities, to regard old problems from a new angle requires creative imagination and marks real advances in science." –

Albert Einstein

Consulting Opportunity Assessment

When Picking Projects to Propose, Brainstorm a Dozen+

Think of Queries to Assess the Size of the Problem

Value Sizing = (number affected) * (value of effect)

- Look at rough upper bound
  - Like assessing the market size for a new product
  - Rank brainstorm ideas, are some $0.5B or $10B market size?
  - Not that you will sell to everybody, or solve every problem (maybe impact 10%)
  - Usually better to plan to influence millions of decisions a small amount
    - Resulting in a large, cumulative effect

  - What problems have the best potential upside to play in?
  - What is the size of the sandbox (of possible value)?
  - https://www.marsdd.com/mars-library/how-to-estimate-market-size-business-and-marketing-planning-for-startups/
    - one example of extensive literature to read (i.e. VC investment assessment)

- Not the same as market sizing, but an analogy to assess the upside for possible projects within a prospective company

Consulting Opportunity Assessment

When Picking Projects to Propose, Brainstorm a Dozen+

Think of Queries to Assess the Size of the Problem

Value Sizing = (number affected) * (value of effect)

- A large % of business problems involve **discrete actions**
  - –Contact customer / prospect vs. don't contact
  - –Offer incentive A or B or C or none
  - –Investigate fraud risk or not
  - –Show banner A or B or C or …
  - –Lifetime value of customer / business is High, Medium or Low
  - –**FREQUENTLY ACT AT A DIFFERENT % THAN NATURAL %**

- Some problems do need accuracy over the full range of target values
  - –Future customer spending
  - –Value of a business prospect
  - –Risk probability (used to multiply times value of transaction or loss pain)

Consulting Opportunity Assessment

- Gap Analysis
  - –Have strong vertical domain experience or do heavy research
  - –Understand best practices in that application, vertical or job function
    - –Who are industry leaders, keynote speakers, main conferences?
    - –Anything on www.Slideshare.com ?
    - –Industry White Papers? ACM Digital Library? www.CiteceerX.com ?
  - –Figure out assessment questions to rank (i.e. 1..5 scale)
  - –Investigate those questions at the prospect, ask them to self assess
  - –Find the GAP per assessment area, between the current client state
    - –This can lead to future projects beyond the current project

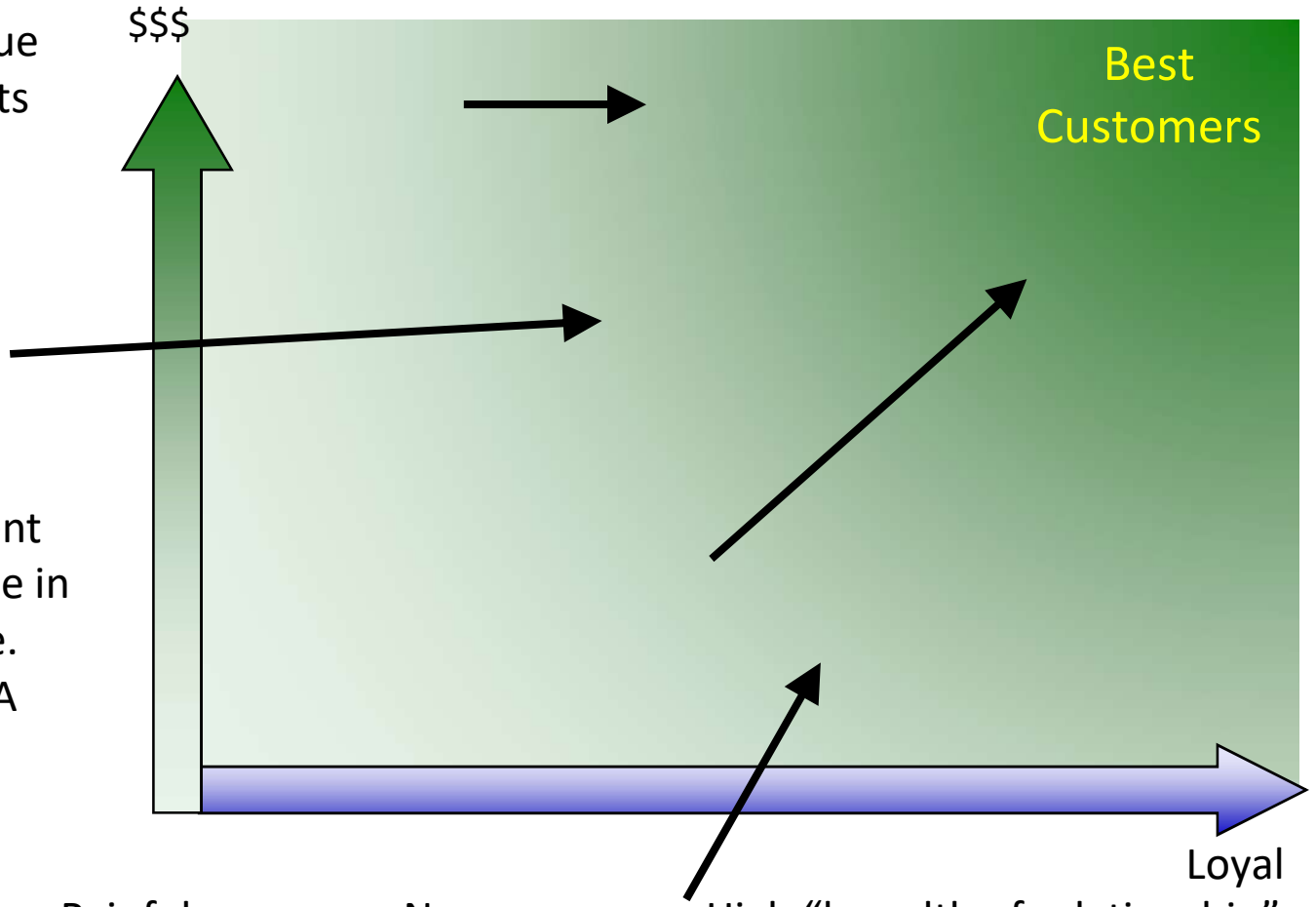
- SWOT Analysis (i.e. market position, http://www.marketingteacher.com/swot-analysis/)
  - –Strength, Weakness, Opportunity, Threats
  - –What are areas for improvement?
  - –How can Data Science (or your products / services) help?

Business Problems to Solutions

Desired CRM Strategies

CRM is different for:
- Marketing,
- Sales or
- Customer Support

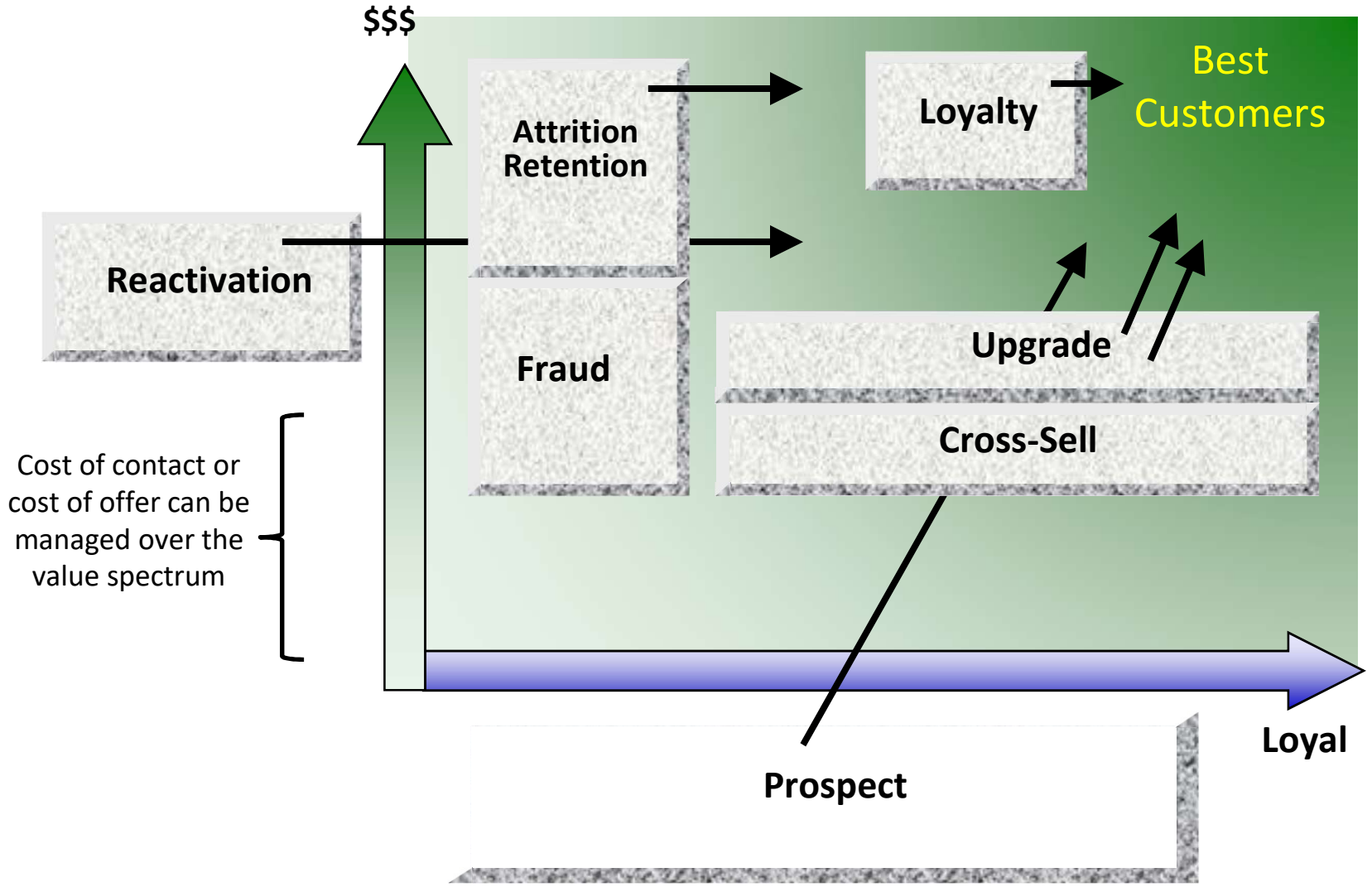
Life Time Value
Net Present Value
Over all Products



May have segment with negative value in one product, i.e. checking or DDA

Business Problems to Solutions

"Who" to talk to (Supervised, Prediction)



Business Problems to Solutions

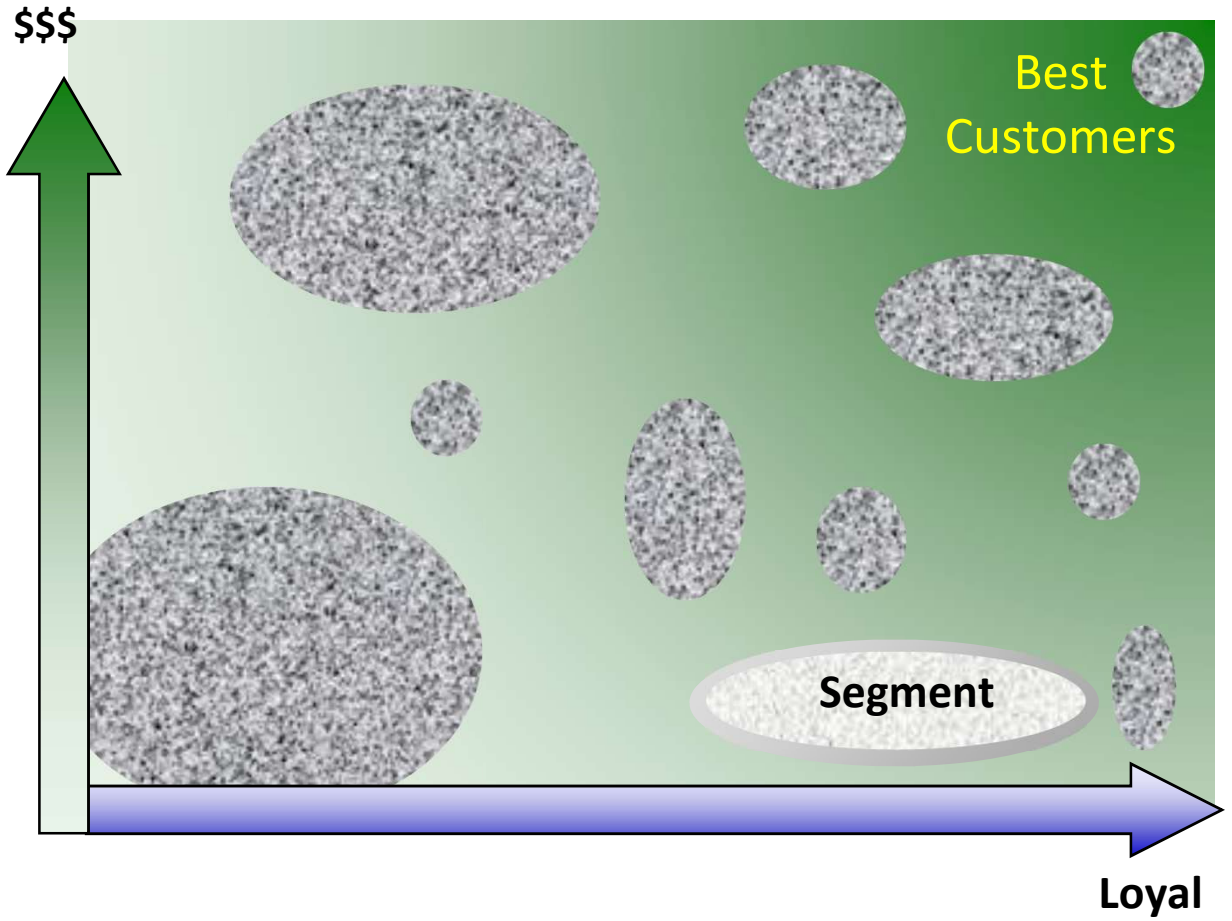
"How" to talk to (UNsupervised)

Intelligence for "creative advertising group"

Segments or clusters may be driven by:

- Spending behavior
- Time using applic.
- Demographics, income, education, rural vs. urban, life stage

The selection of fields used in cluster analysis drives the "voting" in how people get grouped

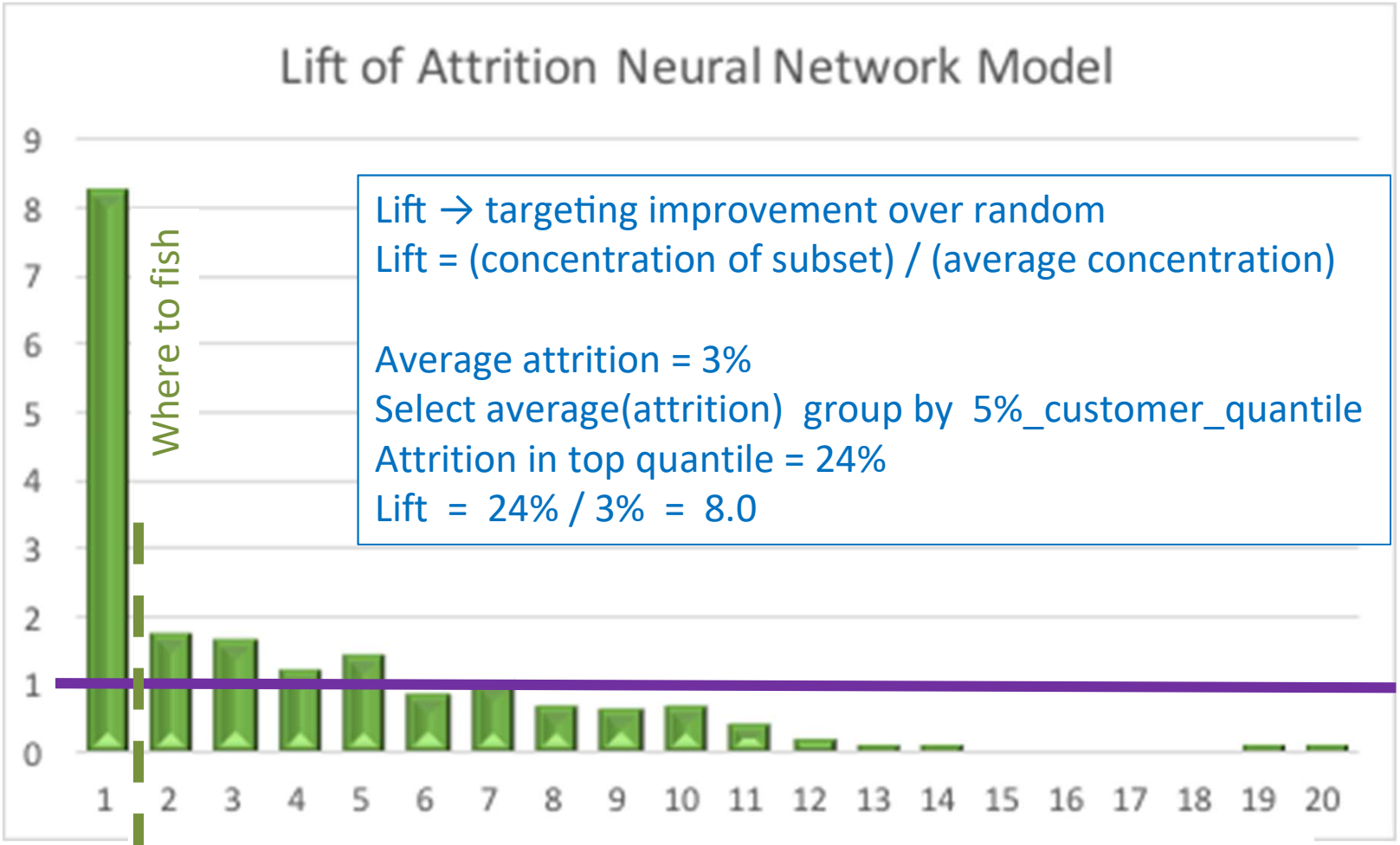


It is very common to decompose a problem, using multiple solutions

The Value of Forecasting on a Population

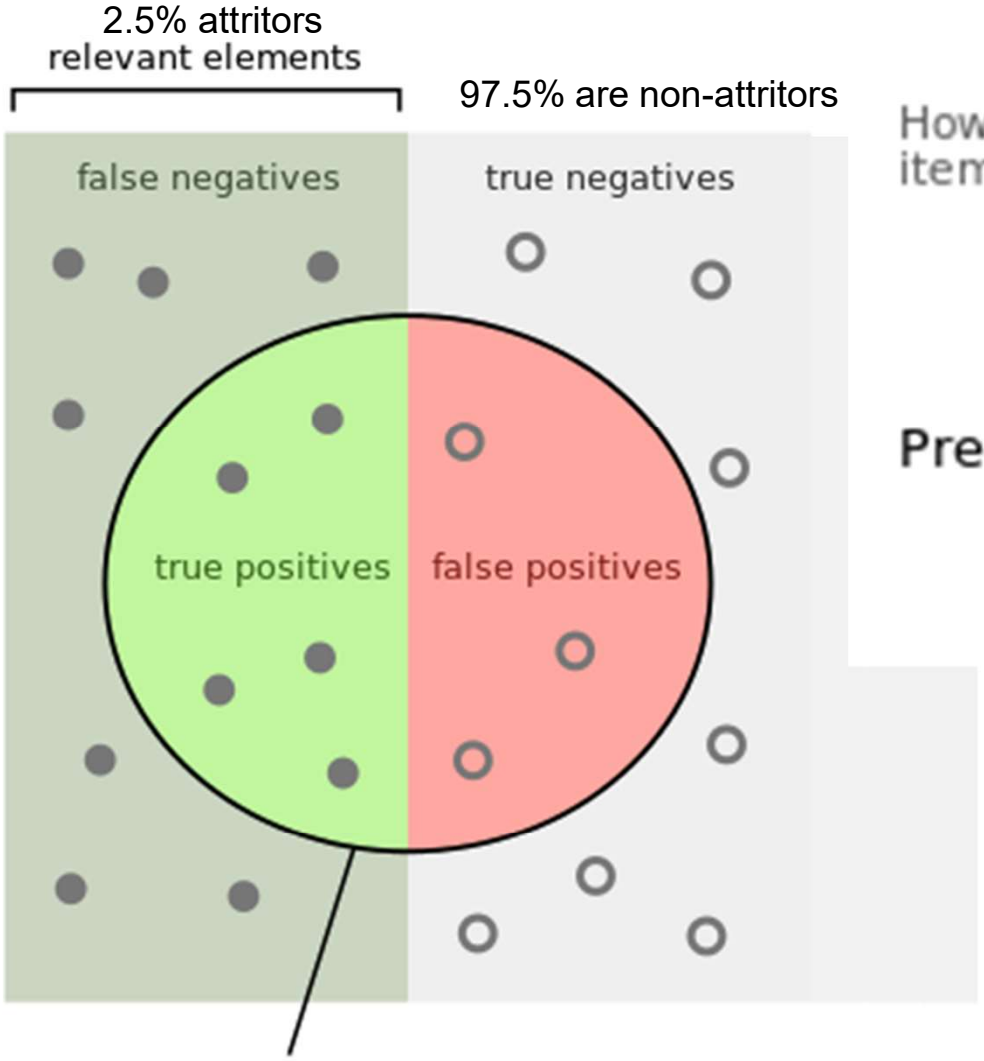
"Fishing in a Lake" analogy

Lift Chart shows concentration by 5% of the Lake



Precision and Recall

Lift Table Selects the Deployment Boundary to Optimize Revenue



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

CAUTION: Precision and Recall technical metrics may not be the best fit to the business
max(profit = revenue – cost)

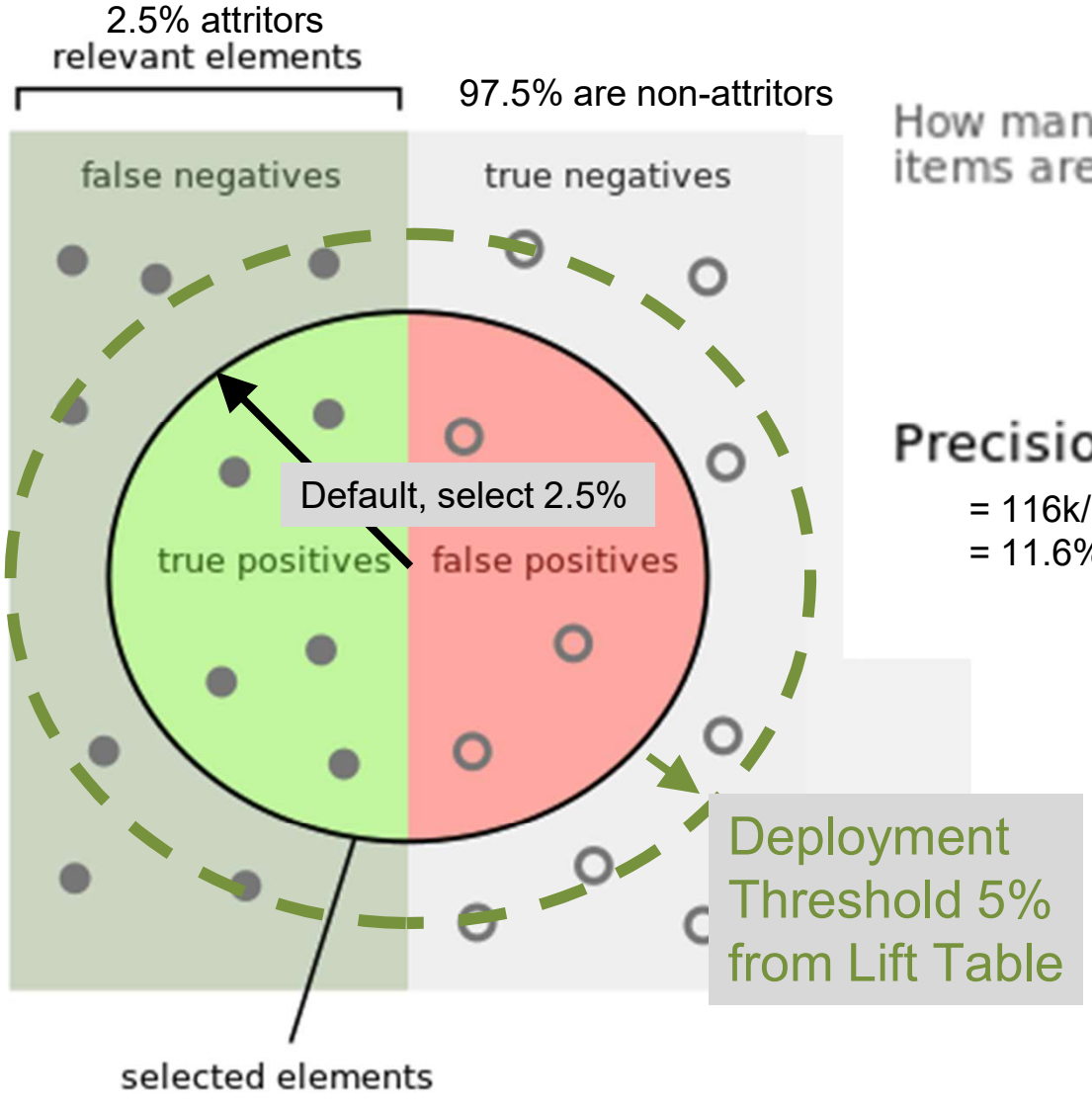
How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

https://en.wikipedia.org/wiki/Precision_and_recall

Precision and Recall

Lift Table Selects the Deployment Boundary to Optimize Revenue



How many selected items are relevant?

CAUTION: Precision and Recall technical metrics may not be the best fit to the business
max(profit = revenue – cost)

Precision =

= 116k/1mm
= 11.6%

Deploy to top 5%
(not 2.5%)

$75 revenue per found
$ 3 cost per false positive

How many relevant items are selected?

Recall =

= 116k/500k
= 23.2%

https://en.wikipedia.org/wiki/Precision_and_recall

Use Lift to Estimate Value for an Attrition Project

Total fish in lake

Number of times cast the fishing net

Category	Input or Assumption	Calculation	Description
Business	20,000,000		Number of Customers
	$ 52.00		Cost per customer acquired
Attrition	2.5%		Attrition Rate per Year (of customers leaving)
Problem		500,000	Number of Cust leaving / year (attritors)
	$75		Annual profit/ customer / year
		$ 37,500,000	Attrition Loss / year (**PROBLEM VALUE**)
Model	5%		Decide to deploy model on top X%

Use Lift to Estimate Value for an Attrition Project

Total fish in lake

Number of times cast the fishing net

Fish caught

Category	Input or Assumption	Calculation	Description
Business	20,000,000		Number of Customers
	$ 52.00		Cost per customer acquired
Attrition	2.5%		Attrition Rate per Year (of customers leaving)
Problem		500,000	Number of Cust leaving / year (attritors)
	$75		Annual profit/ customer / year
		$ 37,500,000	Attrition Loss / year (**PROBLEM VALUE**)
Model	5%		Decide to deploy model on top X%
	9.3		Lift over random in top X%
		23.3%	Forecasted attrition rate in top X%
		116,250	Found Attritors / year
		$ 8,718,750	Value of attritors found / year (**REVENUE**)
Deployment			Cost of targeted communication
costs	$3.00		Marketing + offer cost to top X% / customer
		$ 3,000,000	Cost to top X% / year (**COSTS**)
Net Profit		$ 5,718,750	**(Profit = Revenue - Cost) / year**
		$ 25.81	Cost per customer saved (incl contacting all)
		2.02	Can save X cust, per cost of acquiring 1 cust

Knowledge Discovery Workshop

Objectives: a) Assess Opportunities, b) End up with Sprint Plan, Roles, Responsibilities and Cost

- First, a salesperson qualifies an opportunity
- Or a boss approves an investigation process for a proposal
  - Many times, focus on profit-center projects rather than cost-center projects.
    - Exceptions: legal compliance, cyber security, fraud detection, insider threat
  - Build infrastructure or data lake = cost center, no revenue is expected
  - Profit center projects are in alignment with the company's revenue stream
  - Find out if there is a budget preference for CapEx vs. OpEx (see Investopedia)
  - Capital Expenditures are one time. Operating expenditures are ongoing
  - These buckets of budgetary money can be very important to an executive

- If they are a public company, check their annual report
  - Look for the "CEO Letter to Investors" to talk about KEY STRATEGIES
  - i.e. "grow this division/product line of the company…" "reduce fraud/cost X"
  - Prioritize opportunity brainstorming in alignment with stated strategies and values – it helps later on when resource contention comes up, or when your project is the 29th priority by other departments
    - **"But we all agree to the importance of this CEO strategy…."**

”Begin With the End”
Understand how to Put the Model in Production

Cut out extra preprocessed variables not used in final model
Minimize passes of the data

Many situations, I have had to RECODE prep and/or model to meet production system requirements

- BAD: recode to Oracle, move SAS to mainframe & create JCL
  Could take 2 months for conversion & full QA

- GOOD:
  - Generate PMML code for model (Predictive Modeling Markup Language)
  - http://dmg.org/pmml/v4-3/GeneralStructure.html
  - Deploy in your model building language (R, Python, TensorFlow,…)
  - Spark MLib uses JSON instead of PMML, but a similar principle

Knowledge Discovery Workshop

Objectives: a) Assess Opportunities, b) End up with Sprint Plan, Roles, Responsibilities and Cost

- If you are an external consultant, Always Listen First !!!
  - –Catch up to the business and system knowledge of the employees
  - –Start off with learning by phone, one-on-one
    - Minimize wasting people's time in meetings for initial learning
    - Be a good listener to their experience/advice/brainstorm
  - –Ask management for support time from a "data god or goddess", who has lived with the data for awhile and knows how it lives and breathes
  - –Talk to mid-level business decision makers around the candidate projects
    - People who benefit from various deployments, profit center responsibility
  - –Talk to Dev-Ops or other technical resources involved in deployments
  - –**GAP ANALYSIS of their processes vs BEST PRACTICES**

  - –Start to prioritize the opportunities
    - Be sure to avoid past project landmines that have occurred at the company
    - Engage the data god/ess to query the data to size the problems
    - Their results will be presented to executives in the Discovery Workshop
    - Be sure they can join to describe their findings
    - Work together on assumptions
    - Iterate with business contacts to check assumptions

Knowledge Discovery Workshop

Objectives: a) Assess Opportunities, b) End up with Sprint Plan, Roles, Responsibilities and Cost

- Once you (the consultant/proposer) has
  - Baseline knowledge
  - Initial candidate brainstormed projects
  - Initial sizing assessments
  - Work on draft project plans with your consulting team
  - Identify named people on the client side who own tasks in the plan
  - Build a foundation on what they are doing now or what works
    - **enables incremental improvement** (i.e. preprocessing w/ rules)
- Hold the Knowledge Discovery Workshop (~3-4 hrs)
  - Invite: check writers, those with veto power, mid-business, data god/ess, dev-ops
  - "This is what we have learned so far with one-on-one conversations…"
  - Present draft findings, GAP analysis results
  - "Now we can brainstorm together to prioritize opportunities…"
  - If uncertainty remains, it may be valid to propose start 10% of a project to assess between the 2-3 best options, and then have a fixed date check point to present and decide
  - "Give me 1-3 days to come back with a plan and scope" (big value on draft prep)

During the Project &
Project Wrap-Up or Deployment Checkpoint

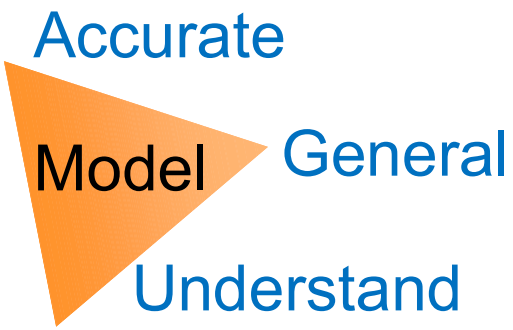
- During the project
  - Some client contacts are in daily or weekly status meetings (not those…)
  - **On occasion, some strategy decisions come up during the project**
  - It is good to reach out, ask & update check writers, veto-power, dev-ops & strategy roles (sparingly, be sensitive and not a burden)
    - **Ask for advice on incremental decisions** (and evaluate with data as possible)

- In the Project Wrap up Presentation
  - Stage: Analysis is finished, checkpoint on deployment – may need consensus
    - Not if small project, but may be needed if the cost is perceived as high
  - **If the strategy questions come up, let the senior clients talk about how they arrived at certain decisions**
    - It is much better and more credible than coming from an external consultant
    - Engagement in making decisions → ownership

Can you Meet All 3 Objectives?

YES!

Constantly ask “how to stay general”



1. Accuracy
   - Focus on the right problem positioning (subdivide & specialize)
   - Preprocessing
   - Model building
   - Incorporate business revenue and deployment costs
   - Later sprints improve on most predictive vars (found

2. Generalization
   - Focus on the right problem positioning (hold out ov
   - Preprocessing (evaluate each detector for false alerts)
   - Model building
   - Track most predictive inputs for changes over time

3. Understandable
   - Sensitivity analysis to rank input vars impact on model ENSEMBLE
   - Create record level reason codes by default – helps with deployment
   - Model building

Add Acc, Gen, Und tags per slide

Design Training Set to Represent the Production Scoring Data

(illustrate with 2 input dimensions vs. target)

Accurate
General

Think about what you want the model to be general on, capture behavior VARIETY:

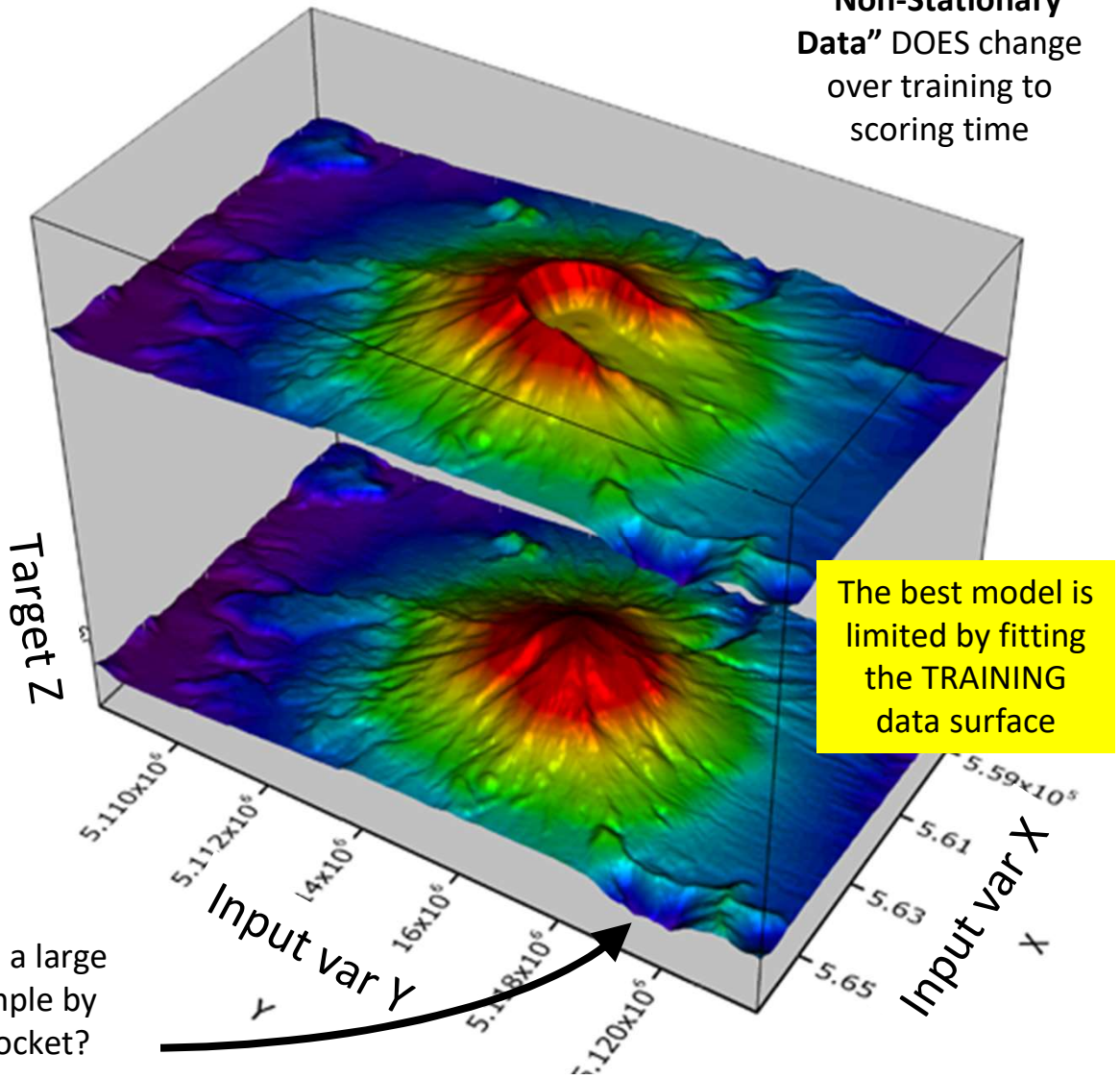
satellite images only during afternoon

Christmas or vacation spending spikes

Current Scoring Data

Training Data

"Non-Stationary Data" DOES change over training to scoring time



The best model is limited by fitting the TRAINING data surface

Do you have a large enough sample by behavior pocket?

Design Training Set to Represent the Production Scoring Data
(illustrate with 2 input dimensions vs. target)

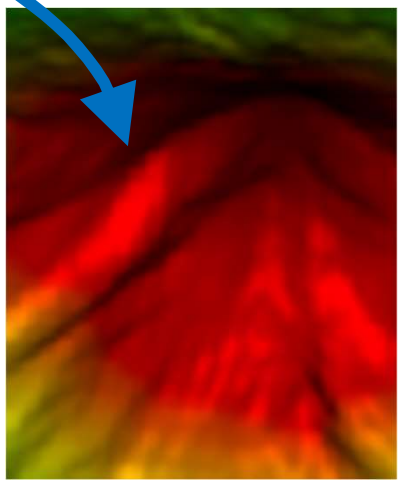
Accurate
General

Accurate vs General

- Note in the training set, the river valley in the high value area, with large target values
- The valley is an example of a key variable interaction
- If accuracy of the model on the training set is very close to the validation data, the valley may be in both
- If the accuracy of the model stays consistent on a "hold out in time" data set, the training data is good
  - Look at "Gap" between train & val or train & holdout

- If there is a big gap in performance, look for ways to generalize the preprocessing to be more independent of features that change over time, that you want to be robust to
- I.e. if the river valley moves to the North slope
- I.e. robust to Christmas credit card spending

"Non-Stationary Data" DOES change over training to scoring time



The best model is limited by fitting the TRAINING data surface

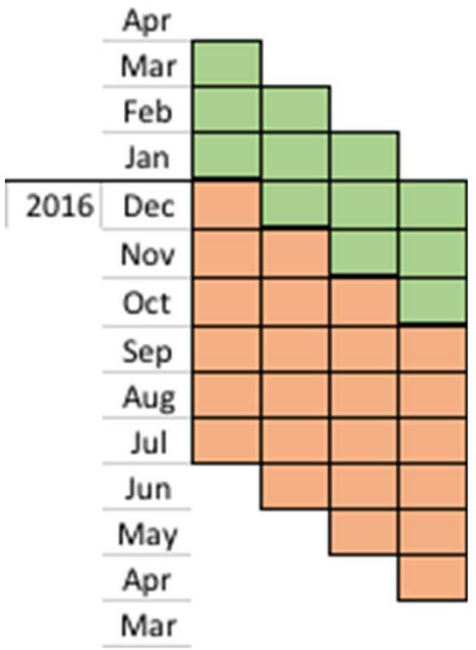
Reference Dates
in Forecasting

Accurate
General

- If scoring now, the ”reference date” = today
  - –Separates input fields in the “analysis past” from
  - –Target fields, defined by the “analysis future”
  - –If the target is “behavior B in the future 90 days”,
  - –May use ”Jan 2 to April 1” as the analysis future
  - –“Jan 1” as the reference date
  - –Select customers with a minimum 6 months history before the reference date

- Captures variety & generalization over seasonality
- Greatly increases the training set size
  - –Enables capturing more interaction details
- Lengthens the life span of the model
- Use most recent as “holdout in time”
  - –Perform a 75% vs. 25% split for training/validation

Same idea
for months or
minutes



Choose
2+ as much past time,
compared to future time

Accurate
General

Pull Time Series Fields

- Time Series (may be closely related to reference date)
  - –Mix time series and non-time series data is good
  - –Get multiple time series
    - •Count of transactions / month (or time unit)
    - •$ spending / month
    - •Web page viewing / month / category
    - •Later preprocessing can fit a line and extrapolate
    - •Get time series per IoT sensor, combine with other data
    - •Watch out for any seasonality issues for future scoring
  - –RFM = Recency + Frequency + Monetary (or time, …)
    - •3 bins of each dimension
    - •Average count of behavior, $, viewing time

Time Series and RFM models used to be stand-alone models.

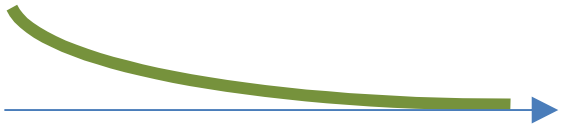
Now part of preprocessing, or ensemble model

Focus on “tool for the job”

Target and Weight Field Variations

Accurate
General
Business

- Target Variable variations
  - –Future spending



    - •May have a ”long tail” or “high skew” (over 3.0)
  - –safeLog(Future spending), then convert back in post-processing
  - –May decompose f(value, business deployment cost) to a simpler thing to be forecasted. Then reconstruct on the back end
  - –Forecast HOW FAR into the future? Farther has larger errors.

- Weight variables
  - –Emphasize stratified random sampling
  - –Emphasize “most valuable records” with larger weight value
  - –May need as integer values, 1+

    (i.e. like number of copied records, more copies → more important)

Business Metric to Optimize

- Does the business act on all values of the target?

- With equal value (or pain) over any target value?
  - –Then continuous measures like R^2, correlation or ROC are valid
  - –This is infrequent → these measures are less useful

- Does the business just act on the top 0.01% or 20% of the riskiest / best / most valuable records?
  - –Then use Lift tables
  - –Sort validation data descending by forecast
  - –Split in 20 (or N) bins with an equal number of records (i.e. 5%)
  - –Lift(binX) = avg(forecast in binX) / avg(actual over all)
  - –Lift = 4% / .5% = 8.0 (8 times higher concentration in binX)

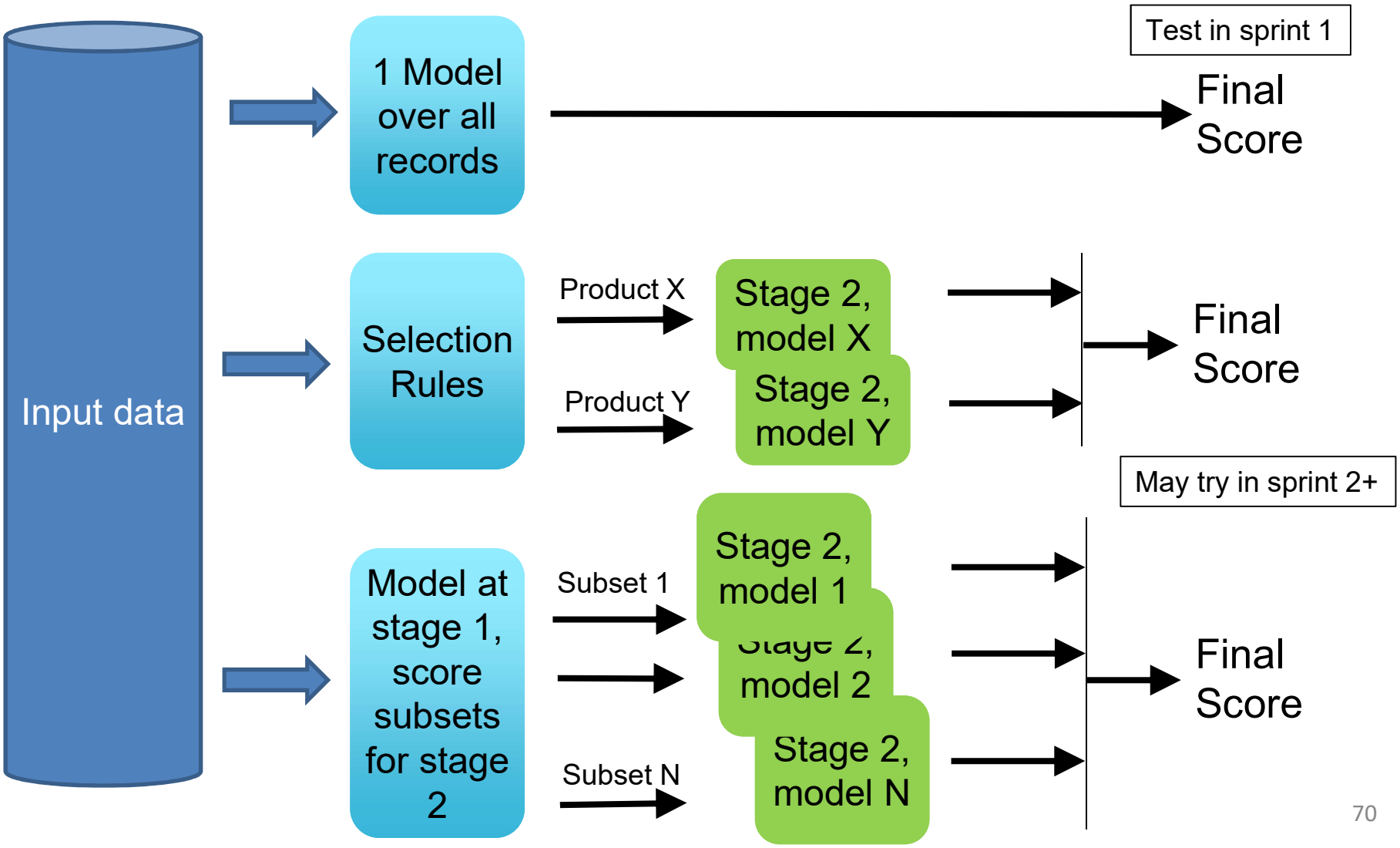
Don't care about error size in bottom 25-99% of records!

Design for the Training Set

What to solve as different problems?

3 Alternatives to Test

Accurate



Design for the Training Set
What to solve as different problems?

Accurate
General

- Just because you are given a problem to solve, don't take it for granted that it is solved with one training set, or one type of algorithm
- Examples of problems to break up:
  - There is a fundamentally different population, or decision process
  - Bank Product: next likely product: checking, savings, CD, credit card, mortgage, …
  - 150 fields, 70% records with no missing, vs. 195 fields 30% rec
  - Mortgage prepayment per month
    - Flat, people who never prepay
    - People who prepay 0…$75
    - People who prepay a steady $100 – 300 / month
    - People who refinance or move
  - Web banner ad vs. in app mobile ad

Doesn't come up much in academic classes

Accurate
General

Design for the Training Set

What to solve as different problems?

Why? Divide and Conquer, may weight diff fields

- Similar to divisions of human expertise, or professions
  - **Narrow focus - enables deeper focus** on the subtle
  - Narrow context (record variety) enables deeper focus

- Have frequently tested this hypothesis in the past.
- You can test on your problems
  - Data version 1:
    - 100% of the data with 4 targets
  - Data version 2 (testing divide and conquer)
    - 60% subset of data with 2 targets (business behaviors)
    - 40% subset of data with 2 targets
  - Compare overall accuracy over 100%

Cluster or Decision Trees can be used for first level splits

Then use other algorithms and vars within subsets

Lab 2 setup

R data.tables (vs. data.frames)

https://www.analyticsvidhya.com/blog/2016/05/data-table-data-frame-work-large-data-sets/
https://www.kaggle.com/general/22444
http://blog.revolutionanalytics.com/2015/05/random-walks-and-datatable.html 6* faster query on 6GB

data.frames()	**data.tables()**
OLDER, see much more commonly	2008 Much more in use in last 5 yrs
readcsv() sequential	fread() 20-30* faster Parallel implementation
Modify column – copy all data	By reference, just add the col Much less memory usage Fast to drop columns
No index	setkey(DT, v4) to create index
	CON: a new grammar, different to learn

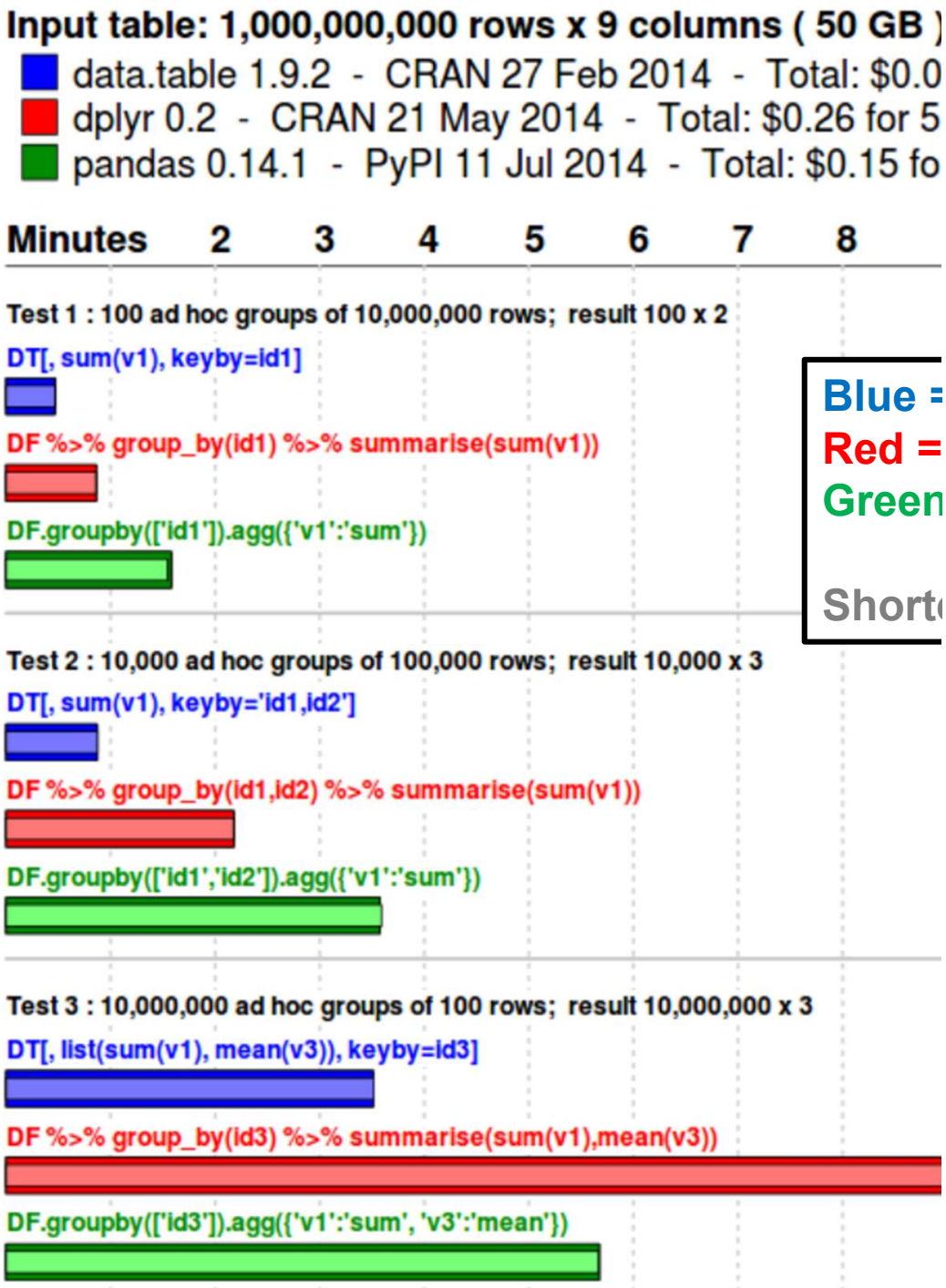
3 Using data.table

Scales on 50GB files,
Billion record files

Amazon EC2 r3.8large
237 GB free in benchmark

https://github.com/Rdatatable/data.table/wiki/Benchmarks-:-Grouping

9 pages of benchmark detail



Lab 2 setup

R data.tables (vs. data.frames)

https://www.analyticsvidhya.com/blog/2016/05/data-table-data-frame-work-large-data-sets/

Walk through like the web site

See also the data.table cheet sheet

https://s3.amazonaws.com/assets.datacamp.com/blog_assets/datatable_Cheat_Sheet_R.pdf

R data.table() demo / lab

See the “data table cheat sheet”

```
      dt[ i  j  k ]     (positions are called ijk)
R:          i         j         k
SQL: where  select  group
```

```
DT = data.table( x=c("b","b","b","a","a"), v=rnorm(5) )
   x          v
1: b  0.96995435
2: b -1.05973768
3: b  1.25886668
4: a  0.29097810
5: a -0.05242568
```

```
# query the records with x containing b
DT[x=="b", ]
   x         v
1: b  0.9699543
2: b -1.0597377
3: b  1.2588667


DT[, .(sumbyx=sum(v), avgbyx=mean(v) ), by=x]
    x    sumbyx    avgbyx
1: b 1.1690833 0.3896944
2: a 0.2385524 0.1192762
```

Outline

Part 1: Get started with R, play with data

Part 2: Data Science Project Design

Part 3: Preprocessing Design

- Review math requirements of algorithms on data preparation
- Missing data handling (simple to complex)
- Convert rules or functions to “detector fields”, “risk”, “rare”
- **Lab 3:** preprocessing your HMEQ data
- Mix time series & non time series, fit linear models within rec
- DBC: categorical vars with 20+ cat, hierarchies, 4-way interact

Part 4: Modeling Design

"More data beats clever algorithms,
But BETTER DATA beats more data"

- Peter Norvig

Director of Research at Google
Fellow of Association for the
Advancement of Artificial Intelligence (AAAI)

1. Algorithm Requirements

Two broad kinds of preprocessing

1. Knowledge representation, try to capture human understanding
2. Technical: from math requirements of algorithms

Want one preprocessed data version to broadly support algorithms

Alg family	Input variable constraints		category	missing	Target var
regression, logistic	numeric (contin or categorical)	similar range (standardize)	requires 1 of N encoding	handle in preprocess	1, binary
regression, continuous	(same)	(same)	(same)	(same)	1, continuous
tree based algorithms	numeric or character	any mix of ranges	deals with N categories (ordered or not)	tree supports missing	1, cat or contin
neural net	numeric	[0..1] or [-1..1] dep on sigmoid	requires 1 of N encoding	handle in preprocess	1 to millions
SVM	numeric	[0..1]	requires 1 of N encoding	handle in preprocess	1, continuous
MOST GENERAL	numeric	[0..1]	requires 1 of N encoding	handle in preprocess	

2. Missing Data Handling

- In all cases, read the meaning of the variable.

- If continuous, a common low effort is to use “mean”,
  - –“median” if the data is skewed,
  - –sometimes 0 for (was not seen or delivered,..)
- If categorical, may use most frequent category, mode, or create “other”

- Ask how it became missing
  - –Self reported income tends to be unreported among wealthier people.
  - –Estimate: select top 66% highest age (not retired), mean(income)
    - •Could sub-divide by geography, education, industry…

2. Missing Data Handling, Sophisticated

- Is there a set of 20+ fields that are always missing at the same time?
  - –Are these half of the top 20 most predictive fields?
  - –Then maybe train model on records with – and another on records without the key missing fields

- If one or a few fields have a high missing rate (33%+) and they seem to be very valuable, use non-missing inputs to:
  - –30 to 70 clusters, calculate average missing var by cluster https://www.omicsonline.org/open-access/a-comparison-of-six-methods-for-missing-data-imputation-2155-6180-1000224.php?aid=54590
  - –Create a quick predictive tree or regression

3. Rules or Queries to Detectors

Accurate
General

- Other systems may make use of business rules, or expert system rules.
- The same functionality can be easily integrated into preprocessing, creating new vars.

- Think of a SQL query, with selected result records tagged with a ‘1’, and ignored records tagged with ‘0’
- How can we go from 0/1 to grey shades using continuous numbers?
- Use the range of field values in the select conditions

- Define Detector as ranging [0..1], 1=business problem, high risk high value, fraud..

Rules or Queries to Detectors

Simple Example,
General Approach to Data Interaction,
Can be Explained

Accurate
General
Descriptive

Select 1 as detect_prospect (result field has 0 or 1 values)
where (.6 < recency) and
(.7 < frequency) and
(.3 < time)

Select (recency + frequency + time)/1.4 as detect_prospect
where (.6 < recency) and (has 100's of values
(.7 < frequency) and (or rescale to [0...1])
(.3 < time)

Develop "fuzzy" detectors, result in [0..1]

Accurate

Existing Compound Detectors

Examples from your Gym or Treadmill

Returned measure 1:
% of max heart rate

Add time + weight (if high)
Returned measure 2:
calories burned

Compound Detectors

Accurate
General

Implemented as a Lookup Table (in this case, same for all people)

- This illustrates the process of creating a detector
- Lets not debate now about specific values
- Don't need perfection
- Dozens of reasonable detectors are powerful

- If user is failing login attempts over more applications, that is more suspicious (virus intrusion?)
- Joe failed logging in over 3 applications, 8 times in 5 minutes

  → failed_log_risk = 0.6

Example Detector Relating Variables to Risk

Failed Login Attempt Risk (Risk is cell value)

		Time Window				
		min	min	day	day	day
Num of apps	Attmp logins	1	5	1	7	30
1	1 - 3	0	0	0	0	0
	4 - 6	0.2	0.1	0	0	0
	7 - 12	0.6	0.5	0.4	0.2	0
	12 +	0.8	0.7	0.5	0.3	0.1
2 - 3	2 - 3	0	0	0	0	0
	4 - 6	0.2	0.1	0	0	0
	7 - 12	0.7	0.6	0.5	0.3	0.1
	12 +	0.9	0.8	0.6	0.4	0.2
4+	4 - 6	0	0	0	0	0
	7 - 12	0.8	0.7	0.6	0.4	0.2
	12 +	0.98	0.9	0.7	0.5	0.3

4. Convert distribution of "normal" categories to "rareness" detectors

Accurate
General

Consider the analysis was of normal vs. unusual behavior at work – to detect if you account as been taken over

What % of your work day involves:

- Phone 10%
- Text 2%
- Email 30%
- MS Office 4%
- Web browser 53%
- Mobile games 1%

This becomes a lookup table used in preprocessing

To calculate "how unusual" is a behavior

Personal Behavior Detectors, Dynamically Updated

Accurate
General

(ANY categorical behavior → Rareness or surprise)

Detect insider threat from changing behavior as a component

Weekly Computer Use of Applications

	Joe (Finance)	Sue (HR)	Tom (Sales)
Word	7%	7%	7%
Excel	25%	2%	2%
PowerPoint	2%	1%	7%
email	10%	24%	9%
messageing	1%	2%	4%
browser	2%	19%	11%
Webex/Zoom	2%	5%	13%
inactive (phn)	12%	24%	40%
HR database	0%	14%	0%
accounting sw	37%	1%	0%
Salesforce	0%	0%	7%

	Joe	Sue	Tom
If all apps equal, avg	11% 1 of 9	10% 1 of 10	9% 1 of 9

Design Principals:

1) No surprise (0) if something is used above average
2) Most surprise if NEVER used (surprise = 1)
3) Scale surprise between 0..1

```
IF (use == 0) then          surprise = 1
Else if (avg < use) then    surprise = 0
Else                        surprise = (avg - use)/use
```

Use	Joe Avg	Surprise = (avg - use)/use
0%	11%	100%
1%	11%	91%
8%	11%	27%
10%	11%	9%
11%	11%	0%
15%	11%	0%

Constantly update counts, weight on most recent weeks, to adapt to current project mix (skip vaca).

Implement with behavior lookup tables per person

Higher Level Detectors

Accurate
General

Illustrated as rules, but typically functions for a continuous score

”Higher Level” or compound detectors

–Group one of many to an overall behavior issue **(using NLP tags)**

```
if (hide communications identity with email alias) or
   (hide communication subject with code phrase) then
hiding_comm on date_time X = 0.2
```

–Group many low level alerts in a short time

```
if (5 <= failed login attempts) and (3 minutes <= time window) then
      Possible password guessing = 0.3
else if (20 <= failed login attempts) and (5 minutes <= time) then
      Possible password guessing = 0.7
```

Analogy

- **Defense attorney** debating plausible innocence
- **Prosecuting attorney** debating guilt

- Detectors seeing the plausible **“best case”** (to reduce false alerts)
- Other detectors seeing the **“worst case”** in each record

–Compare different levels of context **(possibly from different source systems)**

```
if (4 <= sum(over=week, event=hiding_comm) and   # sum smaller detector over time
   (3 <= comm network size(hiding_comm)) and      # network analysis
   (manager not in(network(hiding_comm)))         # reporting hierarchy
escalating comm secrecy = 0.8                     # thresholds distance increases score
```

Lab 3 setup
Preprocessing your HMEQ data

5. Fit a item series with a linear model, then extrapolate

Accurate
General

Name series like:

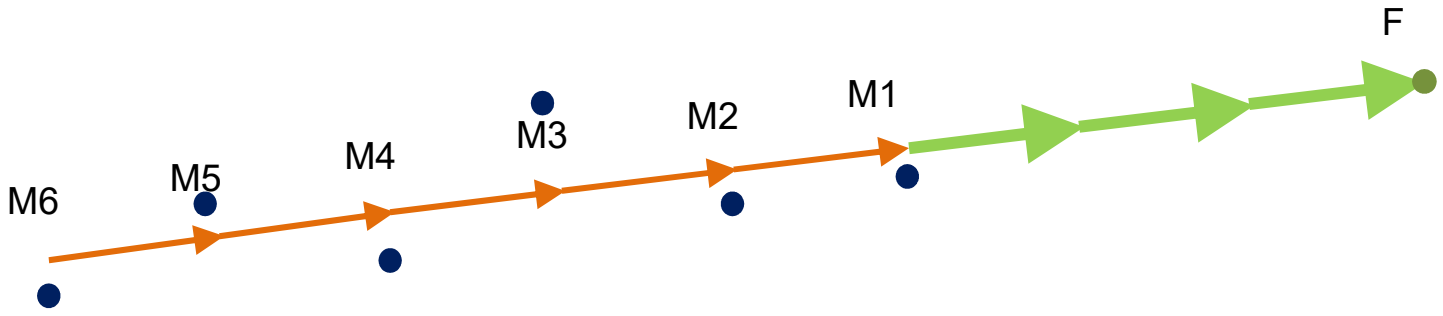
spend_val_m1 … m6 (m1 is closest to ref date)

spend_cnt_m1 … m6 (consistent naming helps)

create new fields extrapolated 90 days like:

spend_val_f spend_cnt_f

(worry about robustness to time or seasonality)



6. Time Series: Detect Individual Past Behaviors to Adapt Future Estimate

- Describe with a use case
  - –Grocery stores want to order X (sauce) cases 3 days in advance
  - –Manufacturers want 7 days notice. Manufacturer wants forecast
  - –Daily order patterns look like: 1700, 0, 0, 0, 1500, 0,0, 1600, 0,0, 0, 0,

- Default analytics
  - –Holt-Winters time series, forecast gap & amount
  - –Average amount = 1600, average gap = 3

    1600, 0, 0, 0, 1600, 0, 0, 0, 1600, 0, 0, 0, 1600,
  - –Is their an amount trend?

    1600, 0, 0, 0, 1625, 0, 0, 0, 1650, 0, 0, 0, 1675, ….

6. Time Series: Detect Individual Past Behaviors to Adapt Future Estimate

- "Fix" est with heuristics by grocer's individual patterns
  - –1'st (or D'th) day of week
  - –1st and 15th of the month
  - –Every 8 (or N) work days

- Develop a series of convolution "detectors" for each type
  - –A fuzzy match with a FIR filter, multiplied by past orders to score

- If (detector X) and (ord est within ~2 day radius) then
  - –Shift holt winters to better fit the individual behavior

	Mon	Tue	Wed	Thr	Fri	**Mon**	Tue	Wed	Thr	Fri	**Mon**
Before:	0	1600	0	0	0	1600	0	0	0	1600	0
After:	**1600**	0	0	0	0	1600	0	0	0	0	**1600**

7. Don't Ignore Input Variables with 20+ Categories (they are valuable)

- Common Handling of Categorical Fields
  - –Such as in SAS Enterprise Miner or other packages
    - •IGNORES variables with 20+ categories !
  - –With 3 to 12-15 categories
  - –Create "dummy variables" or 1 of N encoding

color →	color_red	color_green	color_blue
"red"	1	0	0
"blue"	0	0	1

- Calculate "average target value by category"
  - –Dependent by Category (DBC)
  - –Examples: Merchant category, zip, phone, product category, server, data source, application …

Don't Ignore Input Variables with 20+ Catg

Accurate
General

Use DBC (Dependent by Category)

- Training, create DBC lookup tables, then apply
  - –Support DBC Lookup Table (assume the **average target = 0.5%**)

color →	DBC_color	CNT_color	Lift	
"red"	0.3%	2,341	.6	0.6 = 0.3 / 0.5
"blue"	1.9%	1,208	3.8	(much stronger!)
"tan"	8.2%	6	(count is too small, delete)	

  - –Training records (usually drop the character or categorical variable)

ID	color	DBC_color	
1438	"black"	0.3%	
8347	"blue"	1.9%	
4867	"green"	2.7%	
3486	"black"	0.3%	
9384	"tan"	0.5%	(substitute with file average, 0.5%)

  - –Get a HUGE VALUE going from 100k to 10mm training records
  - –**May provide half of the top 10 best variables**

Accurate
General

Hierarchical Categories

Use DBC (Dependent by Category)

- Hierarchical: geography (zip), products, within company, NLP or topic hierarchies
  - –Higher level (fewer categories) stay closer to average
  - –More granular levels have bigger variation
    - •can be increasingly predictive
    - •But you have to start worrying about “thin ice”, not being stable, not being general
    - •If you flip a coin 3 times and it comes up heads all times, you don’t want to generalize that all coins have only heads
  - –**Can ”use the most detailed, that is stable”**

8. Variable Interactions
Not A*B, DBC

Accurate
General

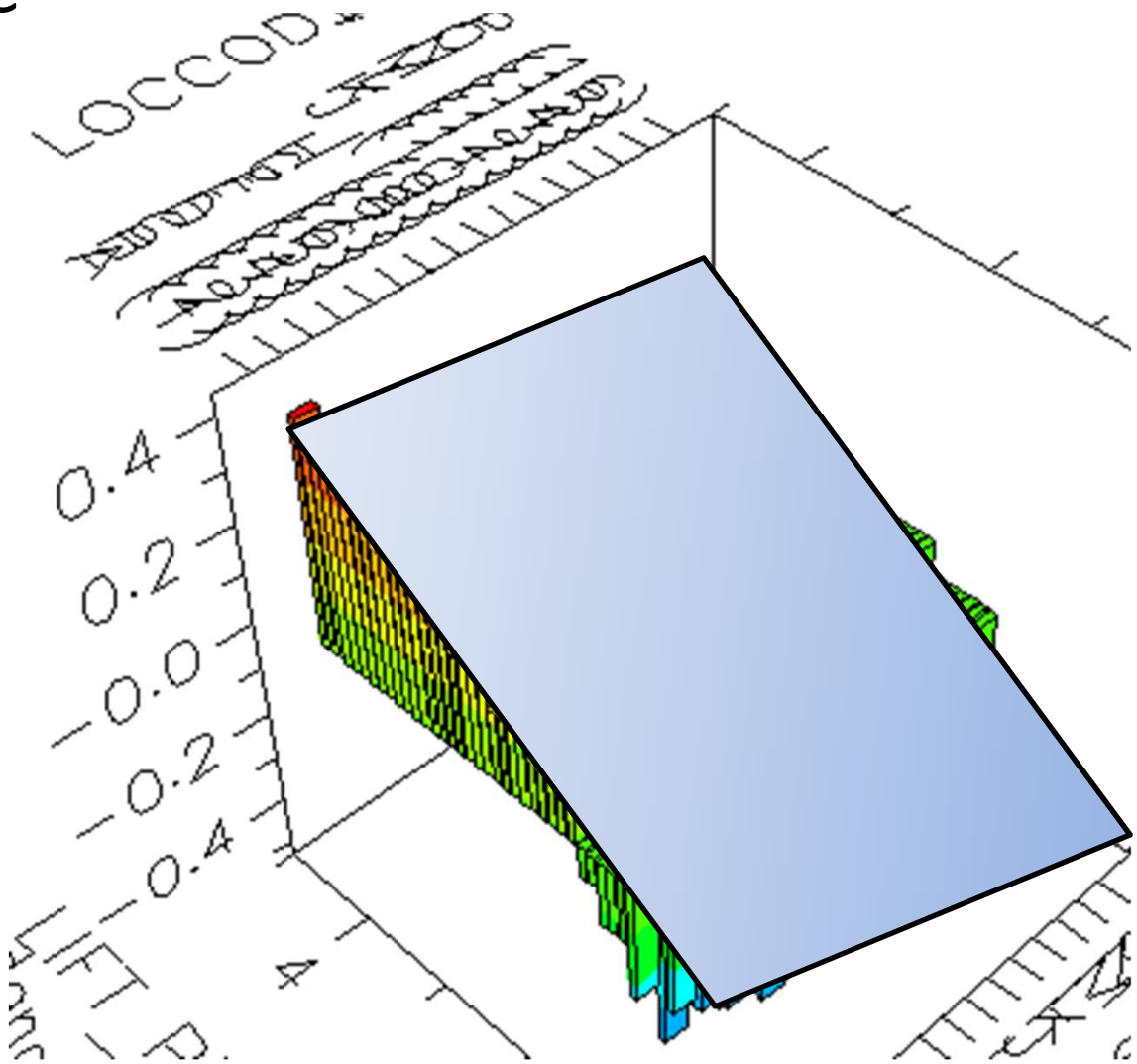
- Common Handling of Interactions (from statistical regression training)
  - –Input variables: A, B, C, A*B, A*C, B*C, A^2, B^2, C^2
  - –Problem: fits target with linear weight times the input interaction
    - •Income is lowest for youngest and retired people
    - •Some things are popular in pockets of geography

- Can get a ”more natural fit”

```
color   tenure →    DBC_color_tenure    CNT_color_tenure
”blue”   0-6           1.1%   (within blue,    1,486
”blue”   7-15          1.6%    find variation  1,248
”blue”   16-25         2.3%    by tenure)      1,967
   :
”green”  0-6           2.8%                    1,208
```

Fitting a flat surface with any slope will be a poor fit

Flat surface = A*B*C

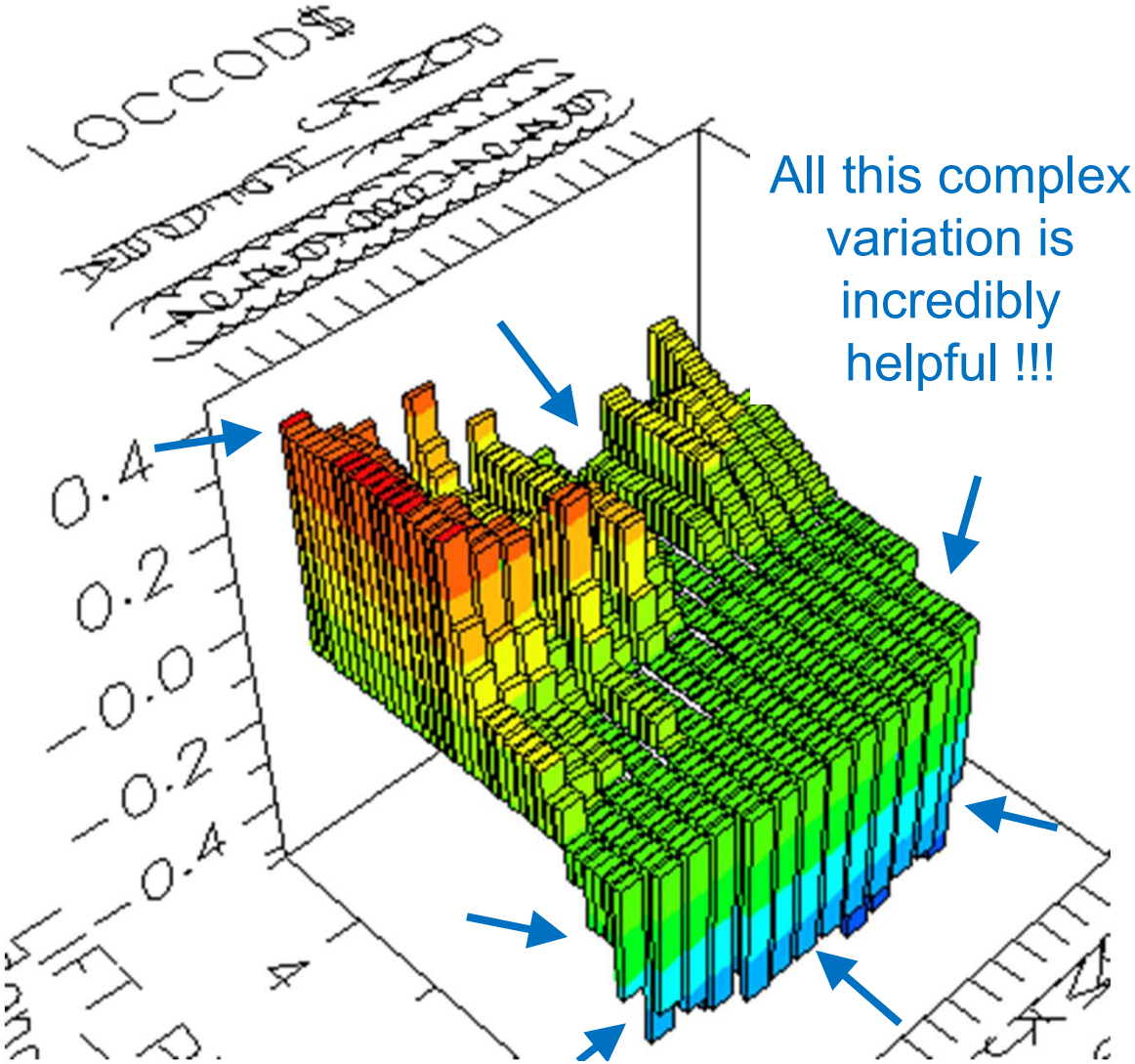


Accurate
General

DBC Captures COMPLEX Interactions
A single sophisticated variable captures this

1) Pre-calculate a lookup table, with the past avg target for each set of prior conditions

2) Apply by looking up conditions for a given store-item returning the target estimate

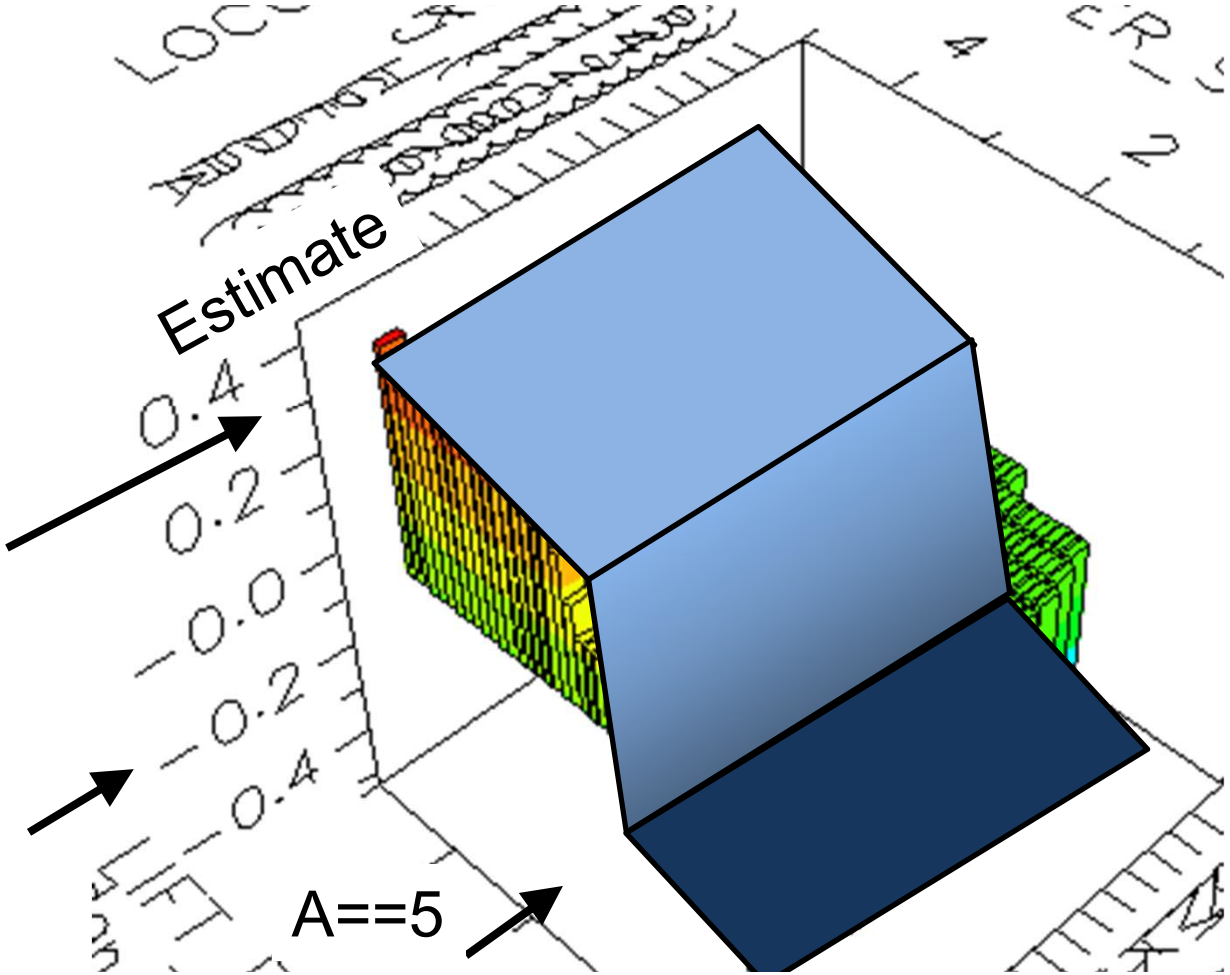


All this complex variation is incredibly helpful !!!

Decision Tree surface fit – less detailed fit

Accurate
General

IF (A < 5) then estimate = **0.3**
ELSE estimate = **-0.2**



9. Variable Interactions
Use Clustering

Accurate
General

- Development Steps
  - –Train a series of models, pick the best one to date
  - –Pick the best 8-20 variables (use sensitivity analysis)

  - –Manually eliminate some, because of strong correlations
    - •Correlate all to each other,
    - •Find the highest correlations
    - •Consider eliminating one of the correlating pair
  - –Create a cluster analysis, such as K-Means, with K like 25 to 40
    - •Drop clusters with under 1% of records
  - –Score the records with a cluster ID, create a 1 of N encoding

  - –Train a new predictive model with “best model param settings”
  - –Evaluate (with sensitivity), and maybe drop all but 1-5 clusters
  - –Now have some powerful, new interaction variables!

Outline

Part 1: Get started with R, play with data

Part 2: Data Science Project Design

Part 3: Preprocessing Design

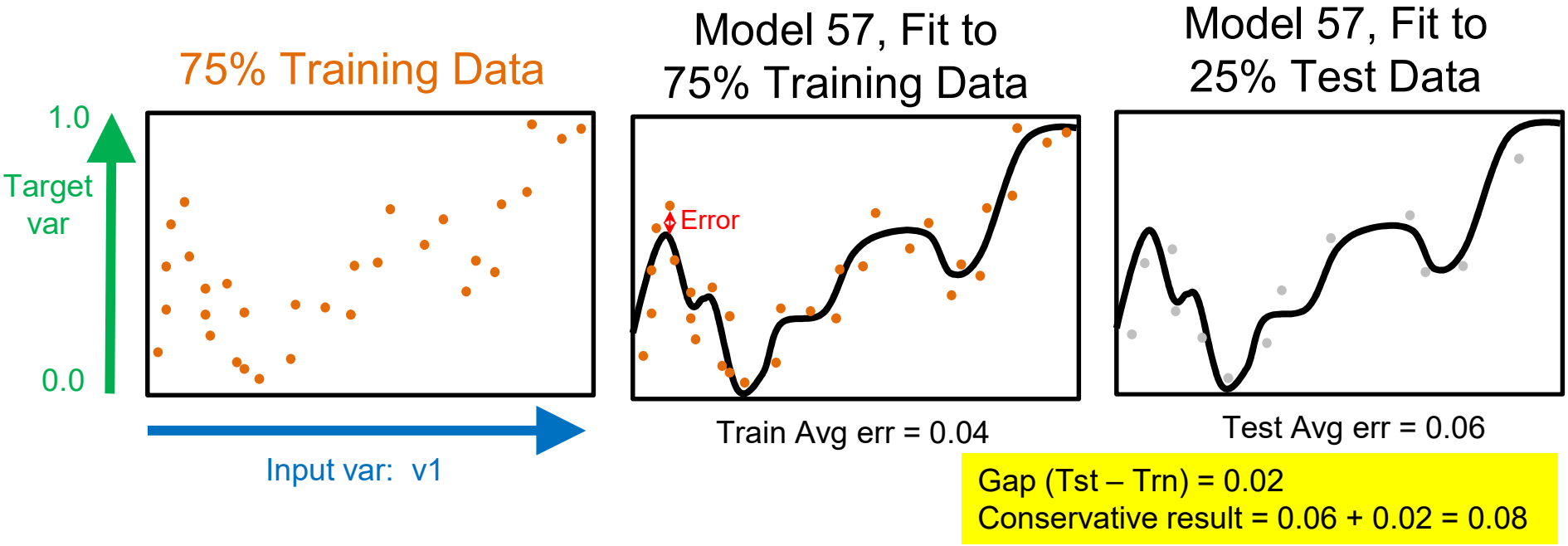
Part 4: Modeling Design

- Model notebook to track, plan design of experiments
- Sensitivity analysis to describe models and record scores
- Review the R caret library
- **Lab 4:** train and evaluate models in caret
- Review available R + Spark combinations
- Example of Shiny Dashboard web GUI

Add overfitting – SHOW tide pool, changing land-water boundary

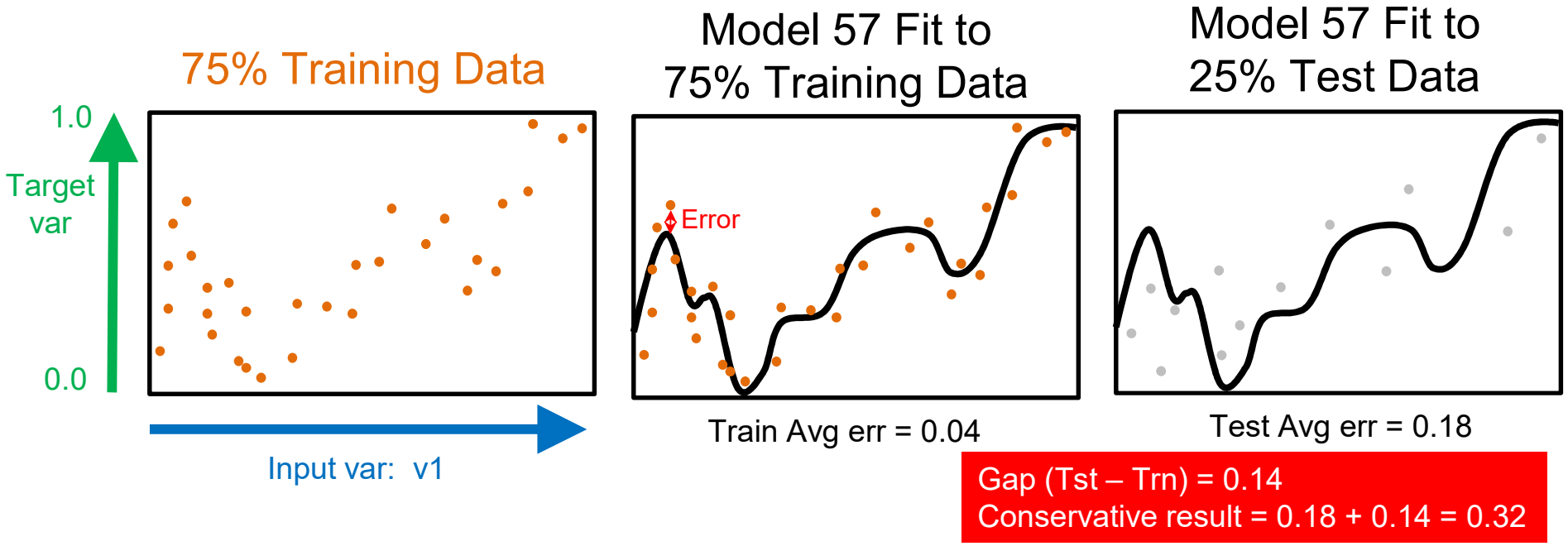
- Give example of simple model vs. complex model
- Validation data that shows when to choose the simple model
- Validation data when to choose the complex model
  - If complex models validates, do not use Occam's razor
  - The simplest model is the average target – we don't consider using that "model"
- Calculate gap between training & test

Avoid Overfitting with Test / Validation Data
Use to Pick Between Models



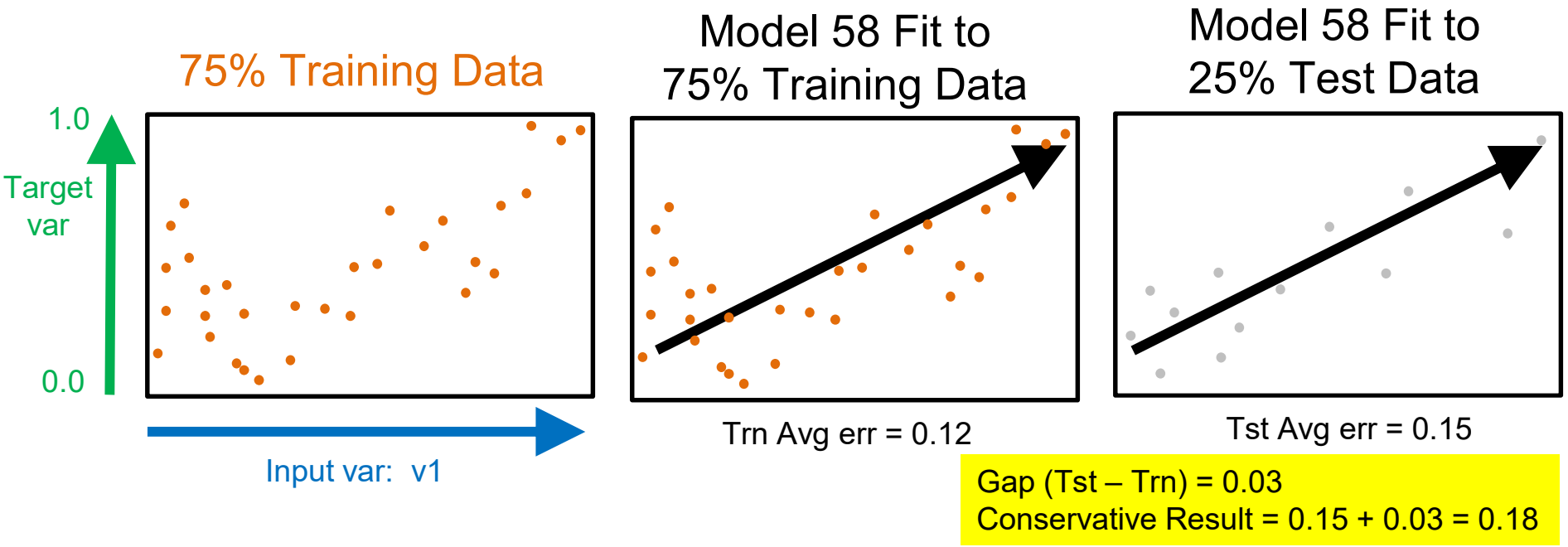
Conclusion: This model consistently fits the training and test data

Avoid Overfitting with Test / Validation Data
Use to Pick Between Models



Conclusion: BIG GAP or INCONSISTENCY
This model consistently fit the training,
BUT NOT THE TEST data
(Detected Overfit)

Avoid Overfitting with Test / Validation Data
Use to Pick Between Models



Conclusion: Can balance
Both ACCURACY and GENERALIZATION
in a metric to minimize

Tracking Model Drift

A trained model is only as general as

the variety of behavior in the training data

the artifacts abstracted out by preprocessing

Good KDD process and variable designs the analysis universe like the general scoring universe

Over time, there is “drift” from the behavior represented in the scoring data, and the original training data

Stock market cycles

Bull → Bear → Bull → …

Tracking Model Drift

MODEL DRIFT DETECTOR in N dimensions

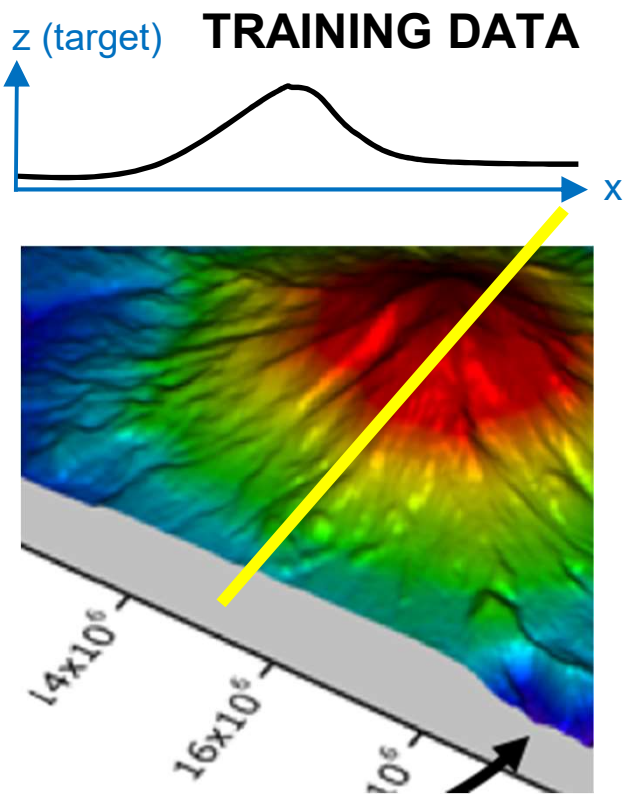
- Change in distribution of target (alert over threshold)
  - During training, find thresholds for 10 or 20 equal frequency bins of the score
  - During scoring, look at key thresholds around business decisions (act vs not)
  - Has the % over the fixed threshold changed much?

- Change in distribution of **most important input fields**
  - Diagnose CAUSES, what is changing, how much…
  - Out of the top 25% of the most important input fields…
  - Which had the largest change in distribution, or a contingency table metric?

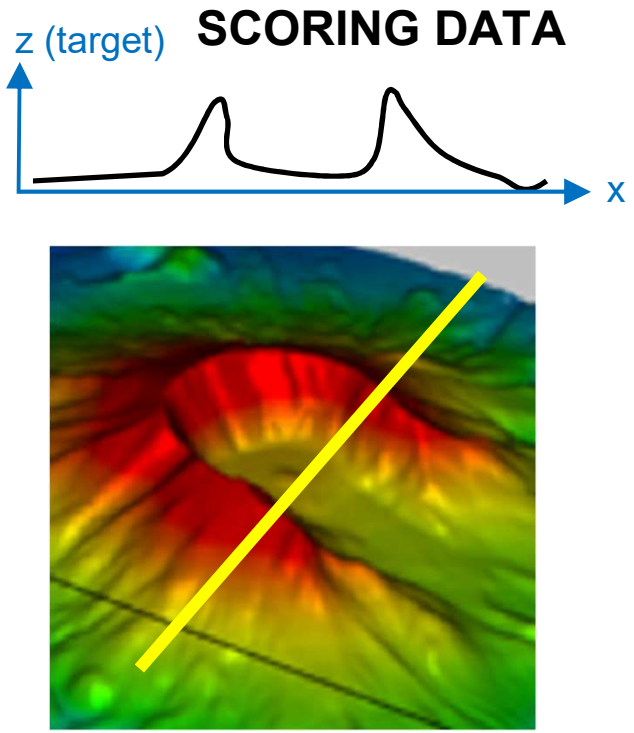
Tracking Model Drift

MODEL DRIFT DETECTOR in N dimensions

- Change in distribution of **most important input fields**
  - Diagnose CAUSES, what is changing, how much…
  - Out of the top 25% of the most important input fields…
  - Which had the largest change?



Distribution of important variable X (where Y=15) changes from one peak to two



4. Retraining Frequency

Is Retraining a) nightly or b) 3-30 months?

- If banner ads, social networking → may want nightly
  - –Assumption: "I will just keep it simple and not have as much to worry about"
  - –Caution: you loose all older experience that can re-emerge
  - –On "5th of July" scoring, is the "4th of July Holiday" data the best training?
    - If the 5th is a workday, a Thursday – may be better to skip Wed and use Mon and Tue (2nd and 3rd) to use to train for the 5th
    - Want to choose training data that best matches the scoring period
    - Don't blindly use only the previous one day's data for training
    - Think about work days, weekends, holidays

Retraining Frequency
Is Retraining a) nightly or b) 3-30 months?

- If most other projects → build models that last
  - –Think of "fundamental behavior changes"
  - –If working on a fraud problem, and normally use the last 6 months
    - •Worry about loosing the example of the fraud attack 8 months ago
    - •May want to pull out just that fraud example, and re-add to training
  - –Think of "fundamental changes in driving variables"
  - –How often do they change?
  - –Like stock market "bull" → "bear"
    - •One model for bull. One for bear.
    - •One for detecting transitions to each one.
  - –May want to select much more data, to capture variation

General
Describe

Retrain or Refresh

Model ReTrain (just like Training)

- Brute force, **most effort, most expense, most reliable**
- Repeat the full data mining model training project
- Re-evaluate all algorithms, preprocessing, ensembles

1-2 months

Model Refresh (like automatic re-training)

- **"Minimal human effort for retraining"**
- Just run the final 1-3 model trainings **on "fresher" data**
- Do not repeat exploring all algorithms and ensembles
- Keep all the same preprocessing programs (no or minimal EDA)
- Retrain model algorithms
- Assume the "structure" is a reasonable solution
- Update DBC and other preprocessing support lookup tables
- Go back to your prior **Model Notebook – choose the best as a short cut**

1-5 days

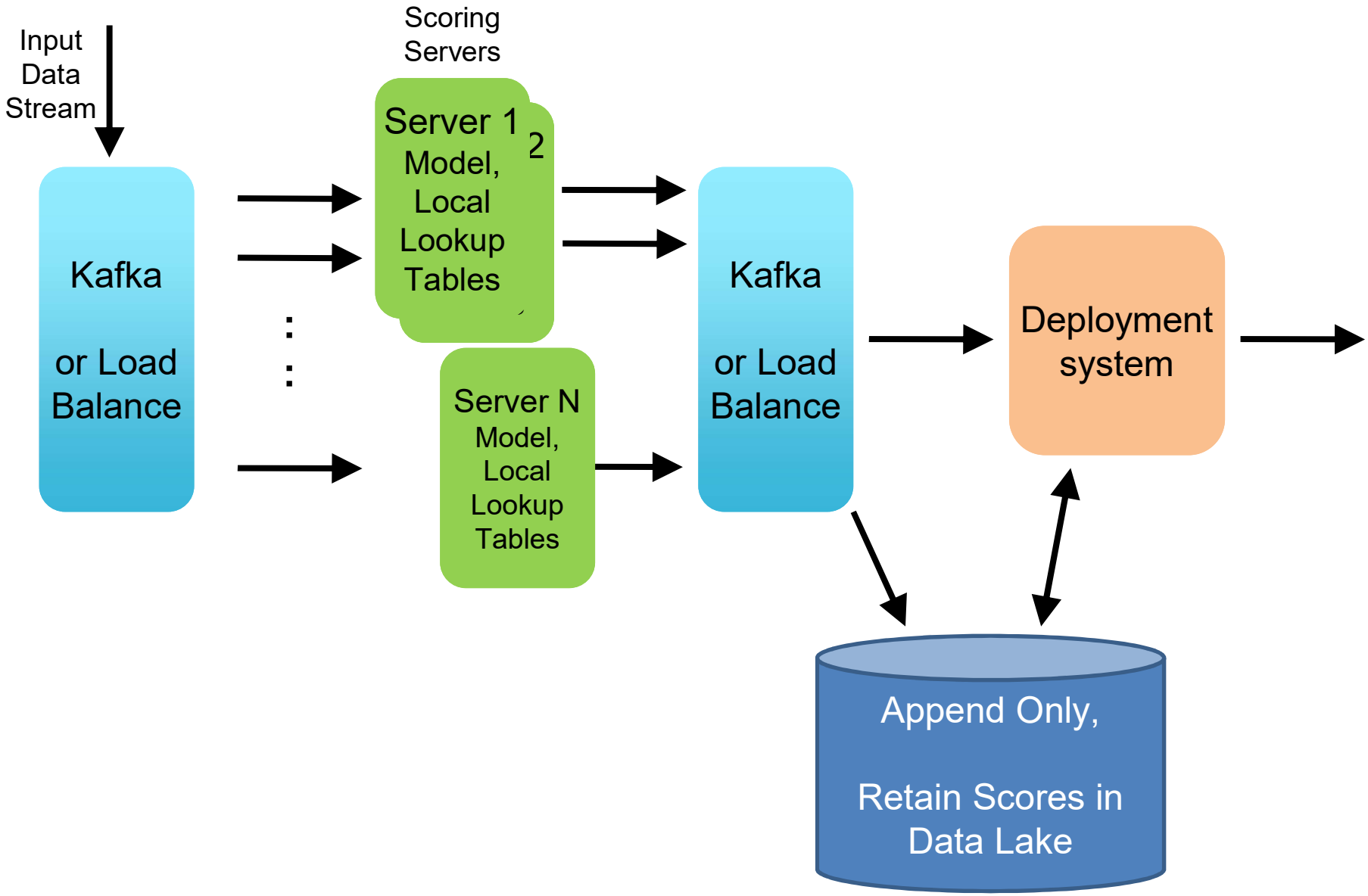
Scalable

9. Big Data Production, Kappa Architecture

- Design Objectives:
  - –Support complexity (enabling accuracy and generalization)
  - –Design efficient preprocessing and scoring
    - •To preprocess one record, don't allow queries of past data to capture "state"
    - •Support "state" of record (i.e. customer, employee) with lookup tables
    - •Lookup: distribution of past behavior, trends, context
    - •Lookup: variable interactions, how they relate to the target variable
  - –Distributed stream processing for scoring
    - •Hash by record key (i.e. customer ID, employee email, …)
    - •Cache local lookup tables (assume persons A..C → Server 1)
  - –Distributed stream processing for lookup table updates, retraining
    - •Can update counts and distributions once, as records are processed
    - •Can trigger retraining when a significant amount of new data is available

Big Data Production, Kappa Architecture

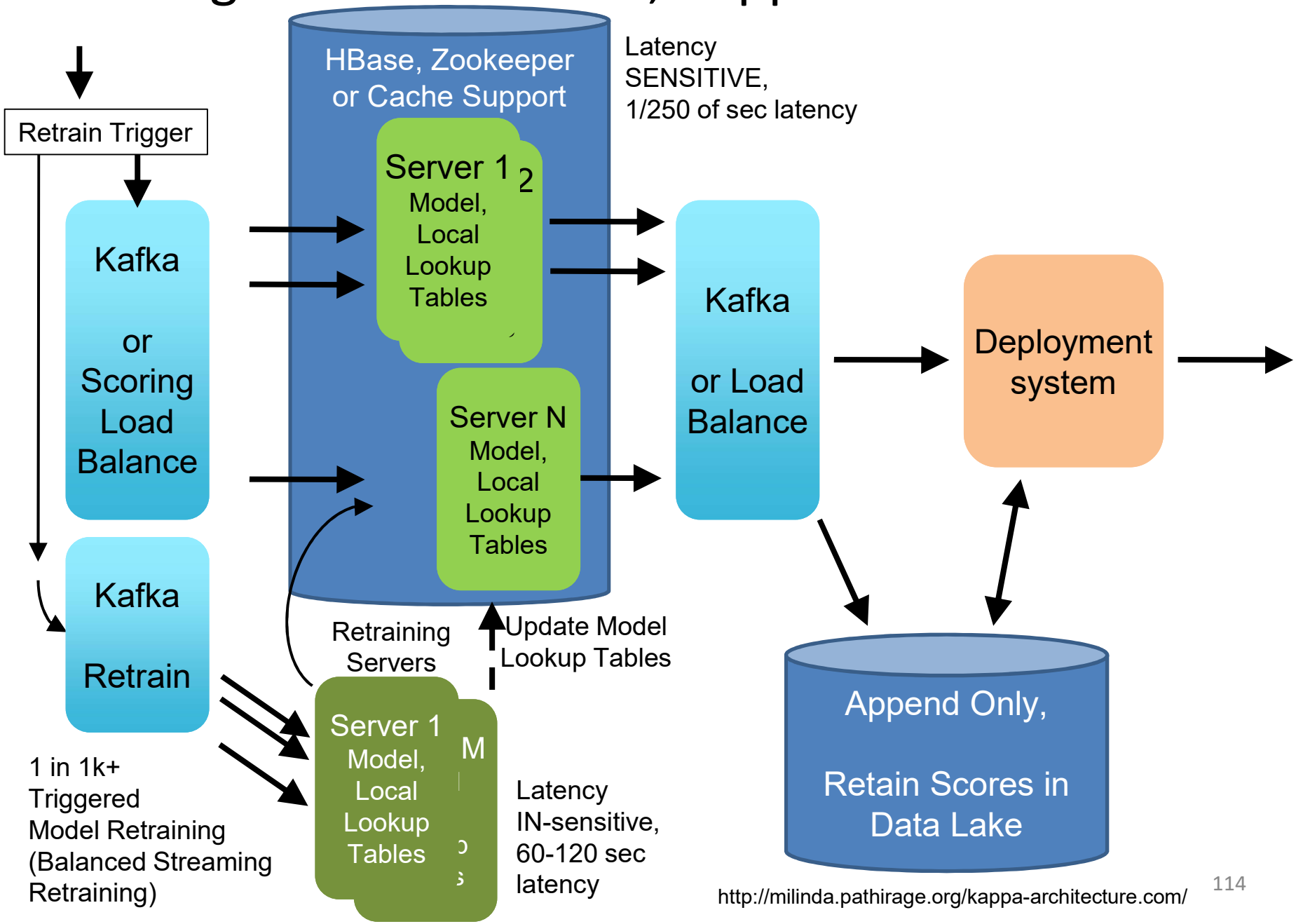
Scalable



http://milinda.pathirage.org/kappa-architecture.com/

Big Data Production, Kappa Architecture

Scalable



http://milinda.pathirage.org/kappa-architecture.com/

Model Notebook

Accurate
General

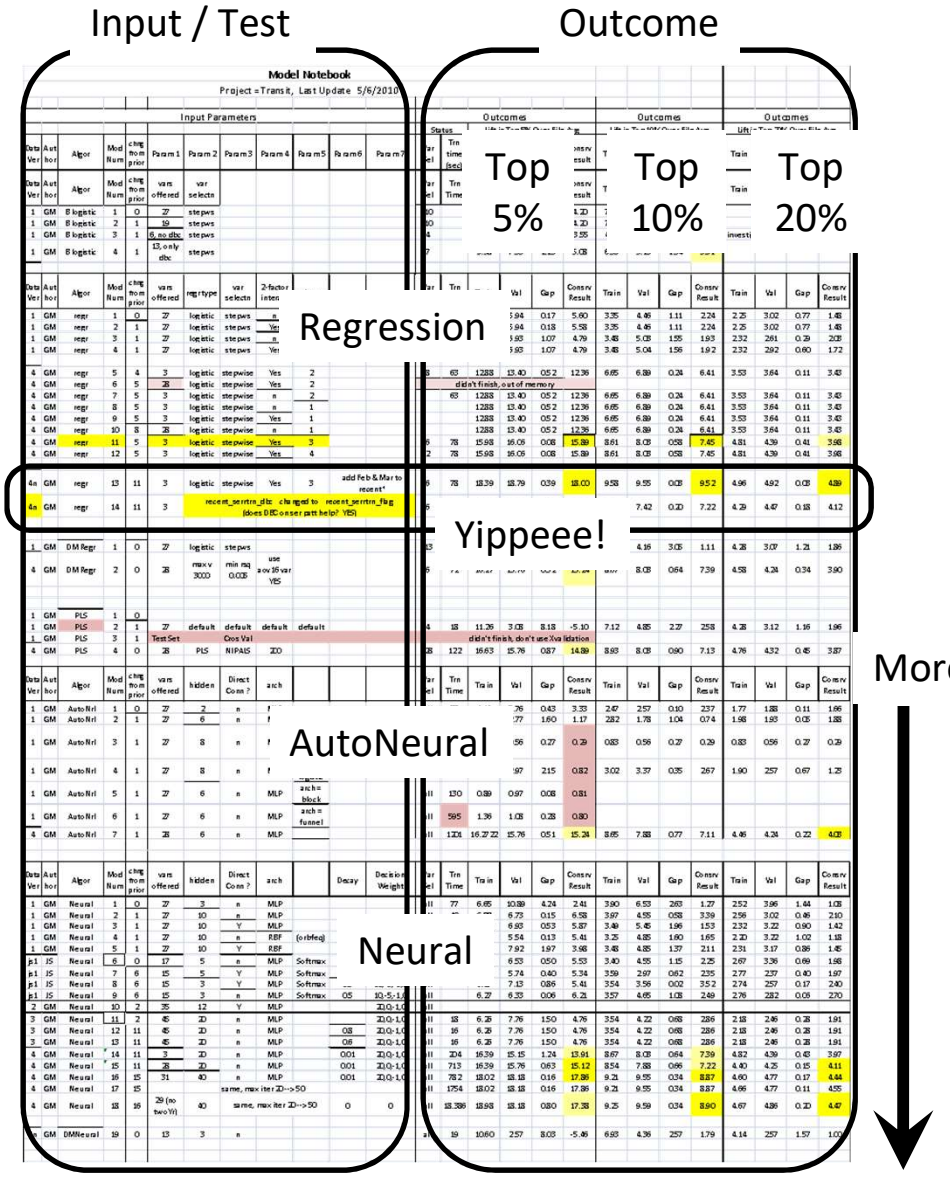
Input Parameters							Outcomes					
Data Ver	Algor	Mod Num	Param 1	Param 2	Param 3	...	Vars Seltd	Trn Time	Lift in Top **10%**			
									Train	Val	Gap = Abs(Trn-Val)	Consrv Result
			vars offerd	var selct								
1	Regrsn	1	27	stepw			9	12	5.77	5.94	0.17	5.60
			vars offerd	Hidn Nodes	Direct Conn	Arch						
1	Neural	1	27	3	n	MLP	all	77	6.65	10.89	4.24	2.41
1	Neural	2	27	10	n	MLP	all	40	6.88	6.73	0.15	6.58
1	Neural	3	27	10	Y	MLP	all	36	6.40	6.93	0.53	5.87
1	Neural	4	27	10	n	RBF	all	34	5.67	5.54	0.13	5.41
1	Neural	5	27	10	Y	RBF	all	35	5.95	7.92	1.97	3.98

Bad vs. Good

Model Notebook Process

Tracking Detail

Accurate
General



Heuristic Strategy:

- Try a few models of many algorithm types (seed the search)
- Opportunistically spend more effort on what is working (invest in top stocks)
- Still try a few trials on medium success (diversify, limited by project time-box)
- Try ensemble methods, combining model forecasts & top source vars w/ model

When Rejecting Credit – Law Requires 4 Record Level Reasons

The law does not care how complex the model or ensemble was..

i.e. NOT sex, age, marital status, race, ….

i.e. ”over 180 days late on 2+ bills”

There are solutions to this constraint, for an arbitrary black box

The solutions have broad use in the model lifecycle

“I understand how a bike works, but I drive a car to work”

“I can explain the model, to the level of detail needed to drive your business”

Set Client Expectations

I understand completely how a bicycle works….

However, I still drive a car to work

A certain level of detail is NOT needed

Do you find out why the automotive engineer picked X mm for the diameter of the cylinders?

You can learn enough detail to let the model drive your business

Desired Benefits of Model Description

Describe the most important data inputs to a model

- What is driving the forecast?
- Good Communication is a Competitive Advantage

For each record, what are reasons for the forecast?

Use to detect data drift – when model refresh is needed

During model building – use to improve the model

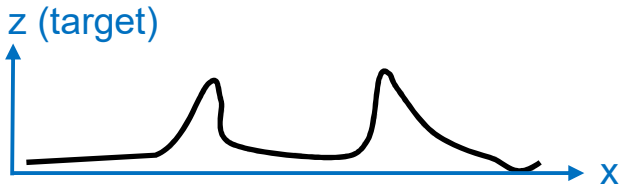
Description Algorithm Design Objectives

1. Describe the model in terms of variables understandable to the target audience

2. Be independent of the algorithm (i.e. Neural Net, SVM, Xtreame Gradient Boosting, Random Forests…)

3. Support describing an arbitrary ensemble of models

4. Pick up non-linearities in the vars

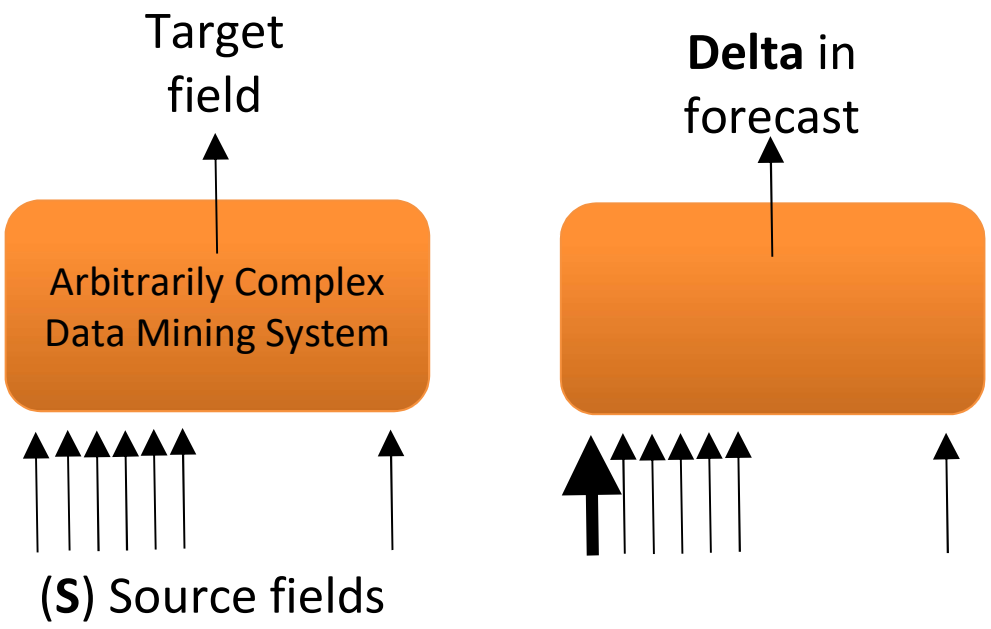


5. Pick up interaction effects

6. Understand the model system in a very local way

Description Solution – Sensitivity Analysis

(OAT) One At a Time



For source fields with binned ranges, sensitivity tells you importance of the range, i.e. “low”, …. “high”

Can put sensitivity values in Pivot Tables or Cluster

Record Level **“Reason codes”** can be extracted from the most important bins that apply to the given record

Present record **N**, **S** times, each input 5% bigger (fixed input delta)
Record **delta** change in output, **S** times per record
Aggregate: average(abs(**delta**)), target change per input field delta

https://en.wikipedia.org/wiki/Sensitivity_analysis

5 Example Sensitivity Records
Intermediate Table of Sensitivities /rec /var

Changes from the target variable, after multiplying each input by 1.05, One At a Time (OAT)

Forecasted Target Variable

Delta 1

Delta 2

Name	BAD forecast	LOAN	MORT DUE	YOJ	CLAGE	NINQ	CLNO	DEBTINC	JOB Mgr	JOB missing
Jacob	0.77	0.003	0	-0.007	0.019	0.011	-0.026	0.009	0	0
Sophia	0.998	0	0	0	0	0	0	0	0	0
Jayden	0.002	0	0	-0.002	0.001	0.001	0	0	0	0.001
Isabella	0.372	0.003	0.015	-0.351	0.038	0	0.079	-0.001	-0.15	0
Daniel	0.929	0.004	0.021	-0.047	0.008	-0.013	0.013	0.011	0	0

Delta N

JOB Office	JOB Other	JOB ProfExe	JOB Sales	JOB Self	REASON DebtCon	REASON HomeImp	value_lg	derog_lg	delinq_lg
0	-0.021	0	0	0	0	-0.560	0.022	0	0
0	0	0	0	0	0	-0.206	0.001	0	0
0	0	0	0	0	-0.002	0	0.002	0	0
0	0	0	0	0	0	0.031	0.194	0	0
0	0	-0.135	0	0	-0.217	0	0.003	-0.023	0

Both Positive and Negative Effects

Changes within Variable Range (Neural Net model 3)

Example Raw Values for Top 12 Variables

Abs = (Total Width over neg and pos)

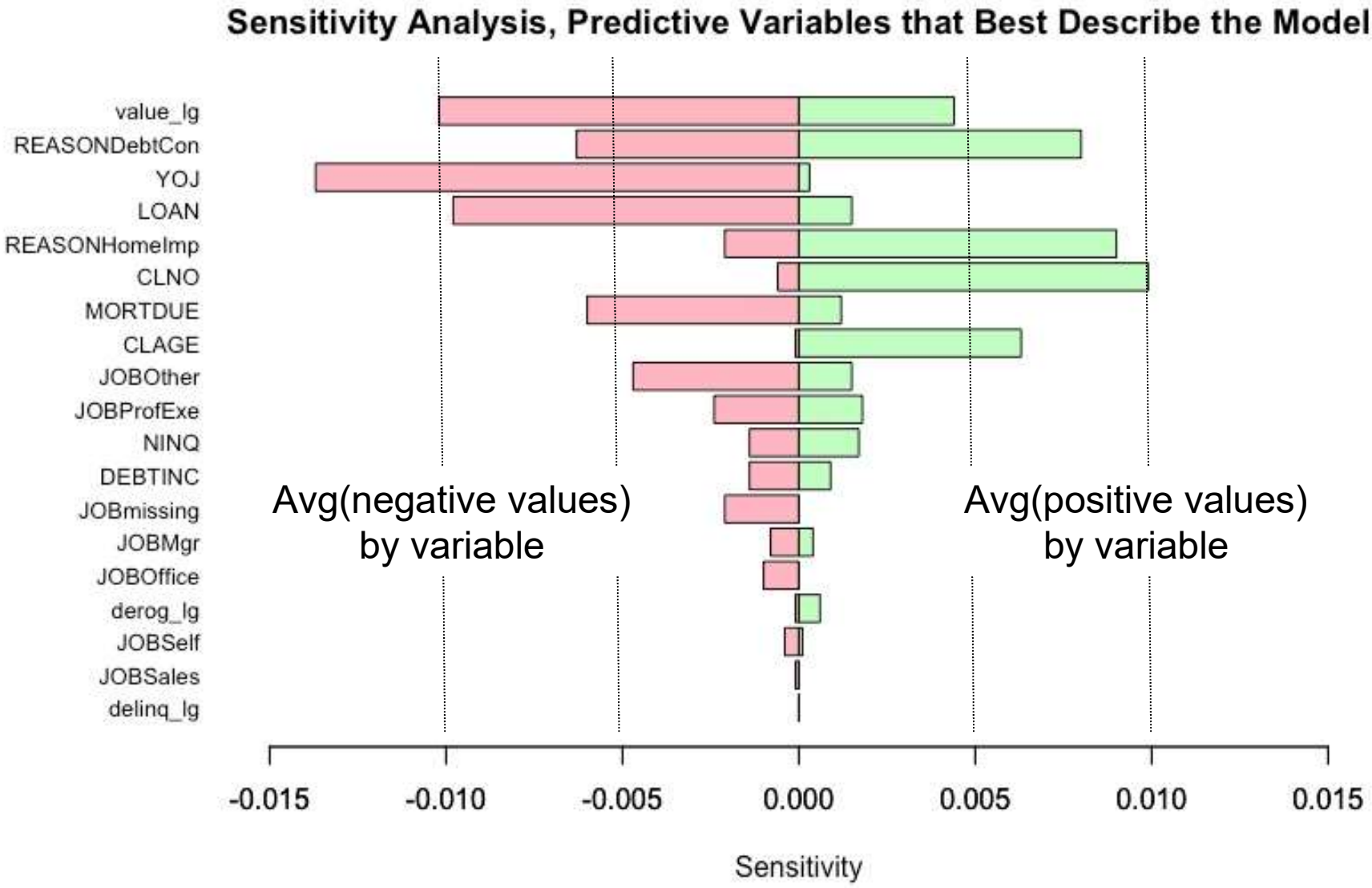
vars	abs	neg	pos	sd
value_lg	0.015	-0.010	0.004	0.050
REASONDebtCon	0.014	-0.006	0.008	0.052
YOJ	0.014	-0.014	0.000	0.046
LOAN	0.011	-0.010	0.002	0.038
REASONHomeImp	0.011	-0.002	0.009	0.067
CLNO	0.011	-0.001	0.010	0.034
MORTDUE	0.007	-0.006	0.001	0.025
CLAGE	0.006	0.000	0.006	0.023
JOBOther	0.006	-0.005	0.002	0.033
JOBProfExe	0.004	-0.002	0.002	0.024
NINQ	0.003	-0.001	0.002	0.016
DEBTINC	0.002	-0.001	0.001	0.008

Standard Deviation

Can be another ranking metric

Both Positive and Negative Effects

Changes within Variable Range (Neural Net model 3)



Description Per Record

Need ”reasons” that apply to some people (records) but not others

A given variable has some value for everybody

Need “sub-ranges” that only apply to some people, i.e.

- Very Low, Low, Medium, High, Very High
- Create 5 “bins”, with a roughly equal number of records per bin
- Focus on the sub-ranges or bins that have the highest sensitivity

Description Solution – Sensitivity Analysis

Applying Reasons per record (independent of var ranking)

- Reason codes are specific to the model **and record**

- Ranked predictive fields

	record 1	record 2
	Mr. Smith	Mr. Jones
max_late_payment_120d	0	**1**
max_late_payment_90d	**1**	0
bankrupt_in_last_5_yrs	**1**	**1**
max_late_payment_60d	0	0

- Mr. Smith's reason codes include:

max_late_payment_90d	**1**
bankrupt_in_last_5_yrs	**1**

Model Notebook: Example of Describing Models

|||||||||||||||||||||||||||||||	Top 1/6 of most expensive items, $5.30+
||||||||||	Past lift by store, sub-dept, dept, front page
||||||||||	Average daily sales per item over prior events
|||	Average price
		Item is located on the front page of the flyer
		Number of Saturday & Sundays in the event
		Item comes from the Health and Beauty dept
		Item in the Stationary department
		Avg # items sold / day

Sensitivity analysis finds what inputs have the biggest impact on the output

Sensitivity Analysis for both Most Predictive Fields and Most Important Modeling Algorithm Params

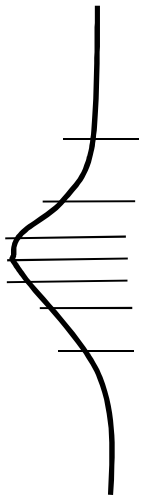
Capture “Data Drift” Over Time
Behavior Changes (pricing, competition)

Use “Training Data” as the baseline

- Create 20 equal frequency bins of the forecast variable (5.0% / bin)
- Save the original, Training, bin thresholds

Check the Scored data over time (i.e. daily, monthly)

Training Split, min	Bin	% Trn Rec	Time 1 Score %	Time 2 Score %	....	Time N Score %
0.000000	1	10.00%	9.70%	9.70%		9.10%
0.000018	2	10.00%	9.90%	9.90%		9.50%
0.001398	3	10.00%	10.00%	10.00%		9.81%
0.013841	4	9.91%	9.91%	9.91%		9.91%
0.118576	5	10.00%	10.01%	10.01%		10.01%
0.341671	6	10.07%	10.07%	10.07%		10.07%
0.635022	7	10.00%	10.20%	10.20%		10.50%
0.986523	8	10.01%	10.01%	10.01%		10.30%
0.996450	9	10.00%	10.10%	10.10%		10.40%
0.999900	10	10.00%	10.10%	10.10%		10.40%



Chi-Sqare or KS-Statistic To measure The slow changes

238 Predictive R Algorithms in caret

Model	method Argument Value	Type	Pac
Boosted Classification Trees	ada	Classification	ada, pl
Bagged AdaBoost	AdaBag	Classification	adaba
AdaBoost.M1	AdaBoost.M1	Classification	adaba
AdaBoost Classification Trees	adaboost	Classification	fastAd
Adaptive Mixture Discriminant Analysis	amdai	Classification	adaptD
Adaptive-Network-Based Fuzzy Inference System	ANFIS	Regression	frbs
Model Averaged Neural Network	avMxnet	Classification	mxnet
Model Averaged Neural Network	avNNet	Dual Use	nnet
Naive Bayes Classifier with Attribute Weighting	awnb	Classification	bnclas
Tree Augmented Naive Bayes Classifier with Attribute Weighting	awtan	Classification	bnclas
Bagged Model	bag	Dual Use	caret
Bagged MARS	bagEarth	Dual Use	earth
Bagged MARS using gCV Pruning	bagEarthGCV	Dual Use	earth
Bagged Flexible Discriminant Analysis	bagFDA	Classification	earth,
Bagged FDA using gCV Pruning	bagFDAGCV	Classification	earth
		Dual Use	bartMa
		Dual Use	arm
		Dual Use	kohon
		Classification	binda
		Dual Use	party, plyr
		Regression	monon
		Regression	monon
		Dual Use	Boruta randor
		Regression	monon
		Regression	brnn
Boosted Linear Model	BstLm	Dual Use	bst, ply
Boosted Smoothing Spline	bstSm	Dual Use	bst, ply
Boosted Tree	bstTree	Dual Use	bst, ply
C5.0	C5.0	Classification	C50, p
Cost-Sensitive C5.0	C5.0Cost	Classification	C50, p
Single C5.0 Ruleset	C5.0Rules	Classification	C50
Single C5.0 Tree	C5.0Tree	Classification	C50
Conditional Inference Random Forest	cforest	Dual Use	party
CHi-squared Automated Interaction Detection	chaid	Classification	CHAID
Stochastic Gradient Boosting	gbm	Dual Use	gbm, plyr	n.trees, interaction.depth, shrinkage, n.minobsinnode
Multivariate Adaptive Regression Splines	gcvEarth	Dual Use	earth	degree
Fuzzy Rules via MOGUL	GFS.FR.MOGUL	Regression	frbs	max.gen, max.iter, max.tune
Logic Regression	logreg	Dual Use	LogicReg	treesize, ntrees
Least Squares Support Vector Machine	lssvmLinear	Classification	kernlab	None
Principal Component Analysis	pcr			
Penalized Discriminant Analysis	pda			
Penalized Discriminant Analysis	pda2			
Non-Convex Penalized Quantile Regression	rqnc	Regression	rqPer
Regularized Random Forest	RRF	Dual Use	rando RRF
Regularized Random Forest	RRFglobal	Dual Use	RRF
Support Vector Machines with Radial Basis Function Kernel	svmRadialSigma	Dual Use	kernlab	sigma, C
Support Vector Machines with Class Weights	svmRadialWeights	Classification	kernlab	sigma, C, Weight
Support Vector Machines with Spectrum	svmSpectrumString	Dual Use	kernlab	length, C

Do you really want a one-off solution?

- Experimenting with Algorithms
- Experimenting with Algorithm Parameters
- Variable description → refine preprocessing

238 Predictive R Algorithms in caret

http://topepo.github.io/caret/available-models.html

Suggestion: Use Some Naming Convention for

data versions, train/test data, models and model versions

Plan to manage complexity, consistency over project team members

d1_trn	data version 1, training input data
d1_val	data ver 1, test or valiation data
d1_hld	data ver 1, on another holdout data subset (if used)
d1_tree_m1	trained model, data ver = 1, alg = tree, tree model ver = 1
d1_treec_m3	similar to above, the tree was trained with caret
d1_trn_tree_m1	apply tree m1 to score the d1 training data
d1_val_tree_m1	apply tree m1 to score the d1 validation data

One Algorithm of Note: Xgboost

See Jerome Friedman's 1999 paper

"Greedy Function Approximation: A Gradient Boosting Machine"

- See also Salford Systems commercial Treenet software (won many competitions in 2000's)
- Xgboost is a good open source implementation, has recently won a number of competitions (X = eXtreme)

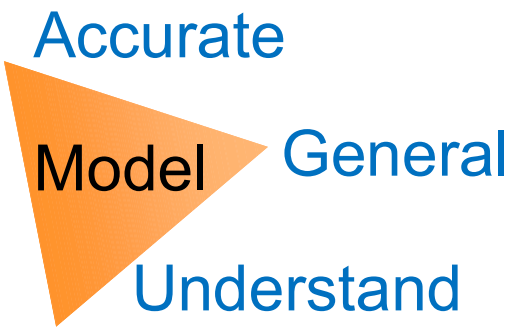
- Initial forecast = average target
- Target 1 = (initial forecast – actual target) a "residual model" forecasts prior errors
- Model 1:
  - Take a random 50% sample of the training data different subsets helps gener.
  - Train a small tree (i.e. limited to 6 leaves) small models, helps gen.
  - Forecast 1 = (forecast so far) + (model 1 forecast)
- Repeat 200 to 2,000 times
  - Avg targ + forc1 + forc2 + … forcN

The 2015 ImageNet competition winner was a Deep Learning residual model, 152 layers deep.
4 people from Microsoft using TensorFlow

5. HMEQ demo/lab

Model Training

Summary



- You can have it all: **accurate, general** & **describable**
  - You may fully understand a bike – but drive a car to work (level of detail)
- Control and plan complexity: track in a model notebooks
  - Reuse notebook when you need to retrain
  - Balance **accuracy and generalization** in the notebook outcomes
  - Track business net value per model (be more competitive)
- Model and record level **description** helps model lifecycle
  - Helps during model building, to improve preprocessing, DBC
  - Helps gain trust
  - Helps track model drift and degradation
- R has a rich library of data science algorithms

Questions?

APPENDIX