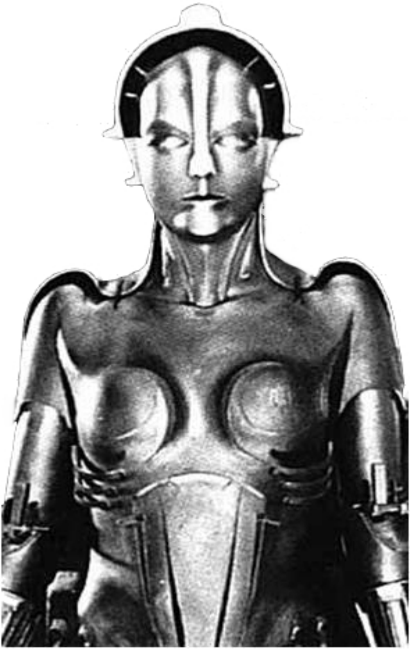
Welcome to machine learning

A brief introduction to an enormous topic

Steven Struhl



Converge Analytic

In brief: Examples, definitions, practical pointers, what has worked

1. A couple of strong examples
   - Two instances of what we can do before we even try to describe this field
2. What it is
   - A brief look at definitions and key ideas
3. Practical pointers
   - Anybody can afford some powerful applications (they are, in fact, free)
4. Looking forward from what has worked
   - A few names to keep in mind

Brevity can be better

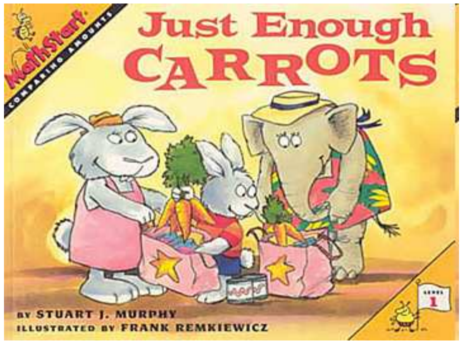
A conceptual approach (no statistics and notation)

- We will touch selected highlights at a conceptual level[1]
- We will skim over a truly vast field
  - Describing some newer methods, approaches and programs
  - Newer because developments are emerging rapidly
    - This field truly is a moving target
  - We will use non-technical language
- If you do not use statistics every day, you should still get the gist
- We guarantee it poses no more challenge than the book to the right on relativity, which also uses no equations (we are told)

This presentation is guaranteed easier to follow than this book [2]

[1] *Those hoping for lots of Greek letters and subscripts, please contain your disappointment*

[2] *For English speakers*



Perhaps enough output

1. Learning about importances via Adaptive boosting

2. Seeing how variables work together via a Bayes Net

Most roads to be left untraveled

A different take on importances from Adaptive boosting

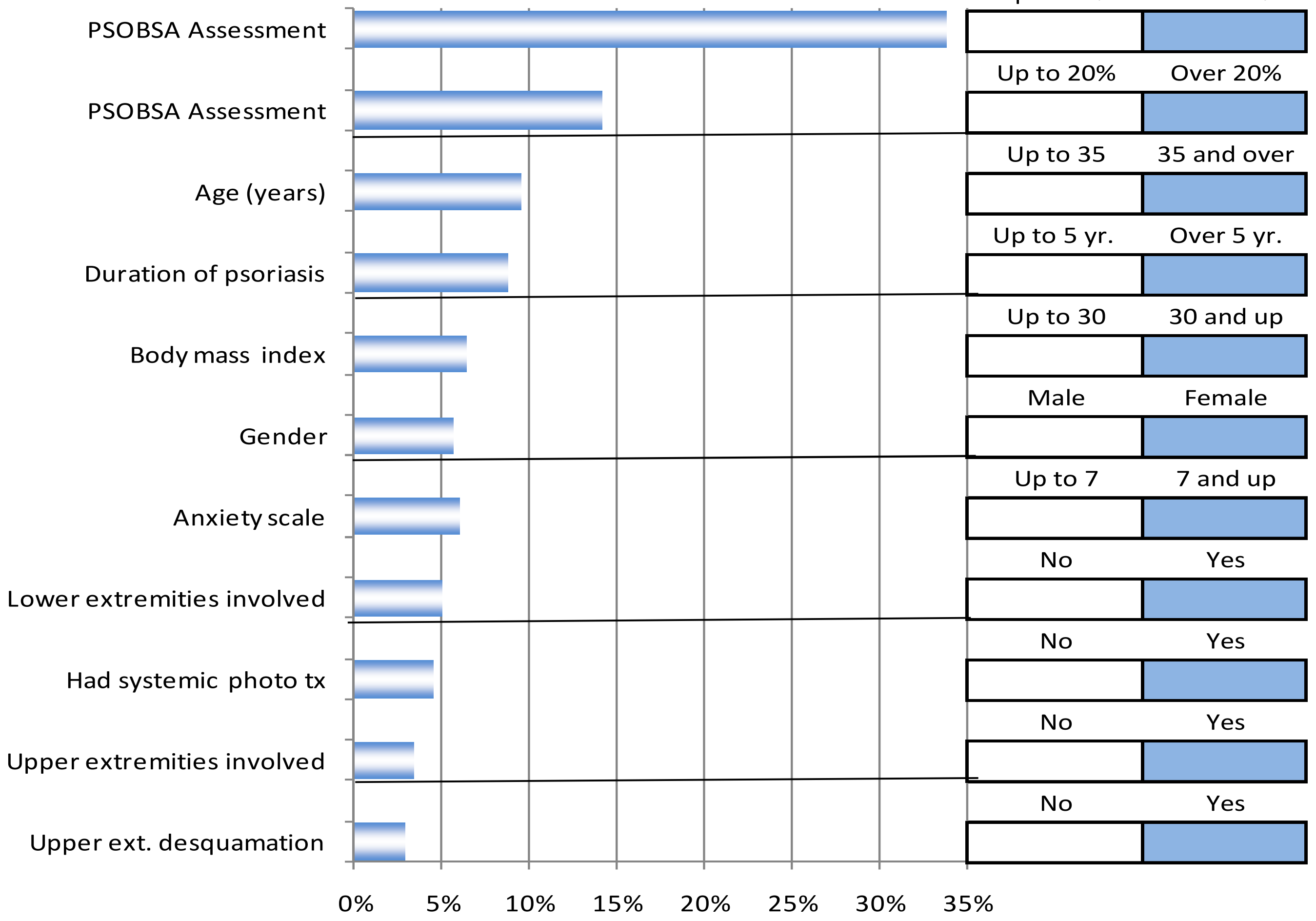
- What it does (we will explain how later)
  - This type of boosting shows not just a weight for each variable, but also where each variable has **break points** or **thresholds**
- Another novel and valuable feature
  - Each variable can show up more than once, with different **breakpoints** given different weights.
- Full name:
  - **AdaBoost MI** using **decision stumps** (one level classification trees)
- Output follows (next page)

Wrong kind of stump
and owls not included

Adaptive boosting output: Importances and break points

- From a study of psoriasis sufferers, using measurements taken in examinations to forecast whether the patient is "at risk" for serious depression

Attributes and their relative importances



Critical values

PSOBSA Assessment	Up to 10%	Over 10%
PSOBSA Assessment	Up to 20%	Over 20%
Age (years)	Up to 35	35 and over
Duration of psoriasis	Up to 5 yr.	Over 5 yr.
Body mass index	Up to 30	30 and up
Gender	Male	Female
Anxiety scale	Up to 7	7 and up
Lower extremities involved	No	Yes
Had systemic photo tx	No	Yes
Upper extremities involved	No	Yes
Upper ext. desquamation	No	Yes

Note that this found two breaking points on the same variable with different importances in predicting "at risk" scores

The group associated with higher "at risk of serious depression" scores is shaded

Correct classification = 77%

Bayes Nets show how variables group and even can show causation

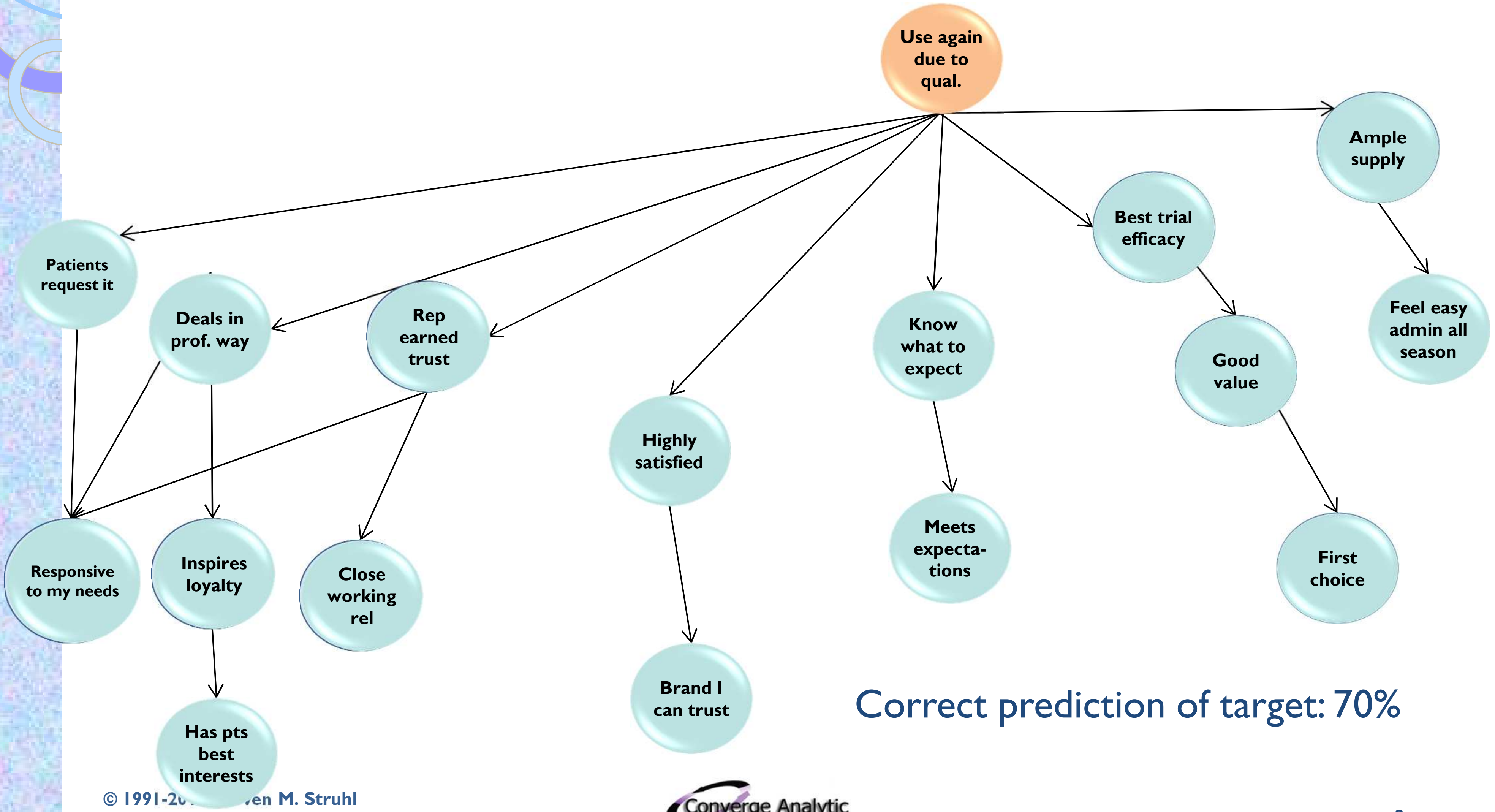
- These are networks or groupings of variables that arrange themselves
  - They can take variables you define then show how they fit together vs. a dependent variable
    - These groupings typically make a great deal of intuitive sense
    - The data shows which variables fit together and which not
- Longer names: Bayesian networks, belief networks, or Bayesian belief networks.
- These are now used in hard sciences (such as cancer research)
- With the right conditions they even show cause and effect [1]
  - Could this be getting close to the researcher's version of the holy grail?

Holy grail
(non-research version)

[1] *Disclaimer: These conditions unfortunately rarely exist in the types of data we use*

This Bayesian network organized itself

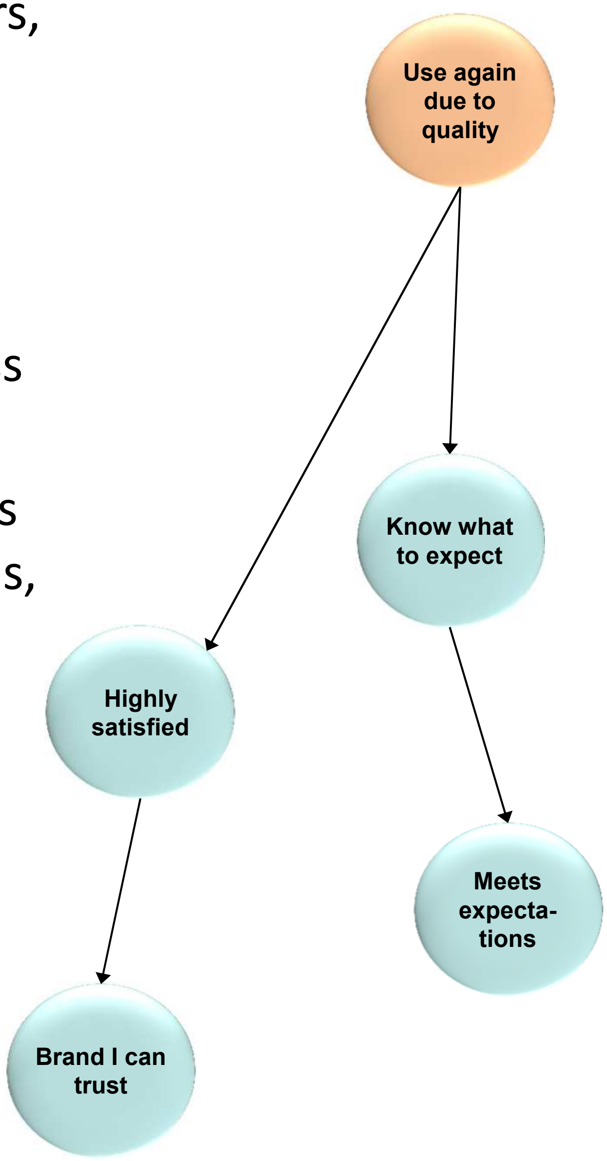
- The network predicted the dependent remarkably well and provides many insights
- Note that we colored in the target variable so it stands out



Correct prediction of target: 70%

Satisfaction relates most strongly to trust

- Satisfaction and trust are next door neighbors, meaning that they are very closely tied
  - We cannot have satisfaction without trust and vice versa
  - Even though the arrow points one way, effects travel in both directions—closeness largely determines the strength of effects
- Knowing what to expect logically enough falls close to its complement among the questions, whether the product meets expectations
  - Note that neither of these latter two connect to directly to satisfaction, but do have their own direct path to the target variable



Use again due to quality

Know what to expect

Highly satisfied

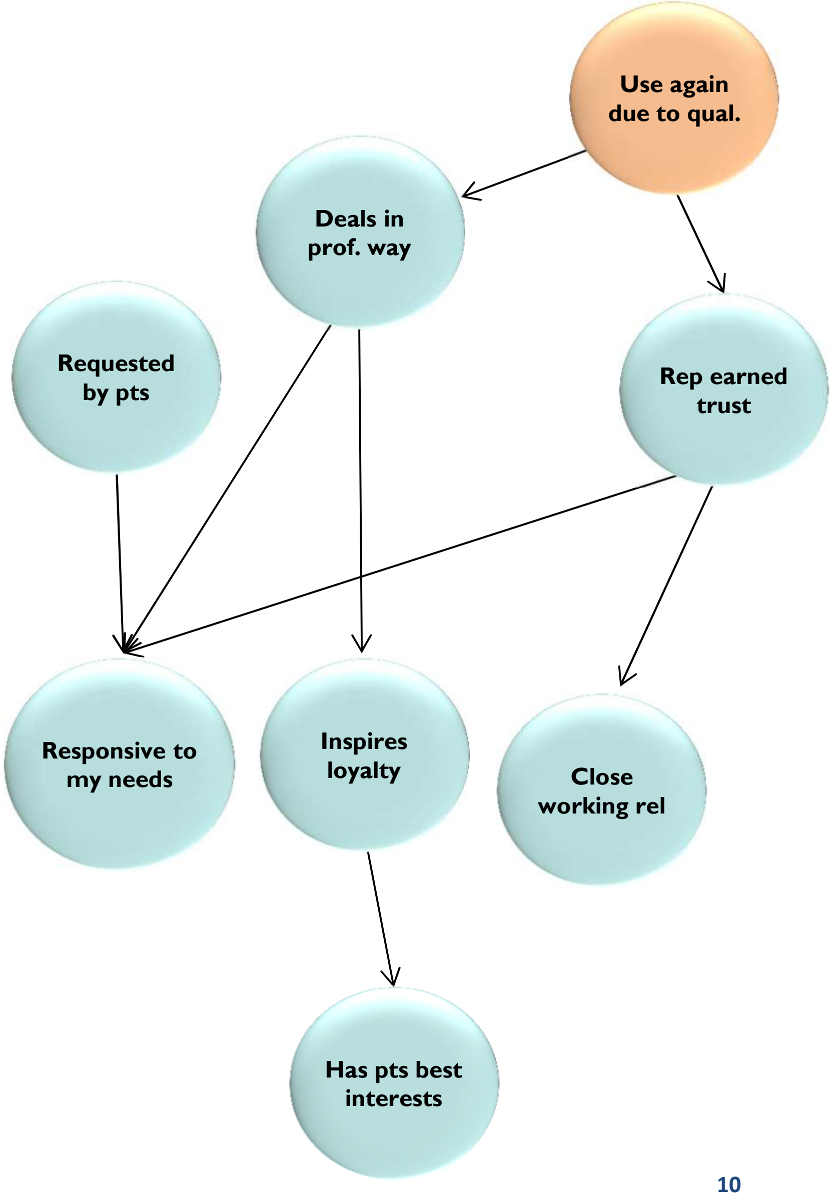
Meets expecta-tions

Brand I can trust

We are here

Earning trust and professional dealings have pivotal roles

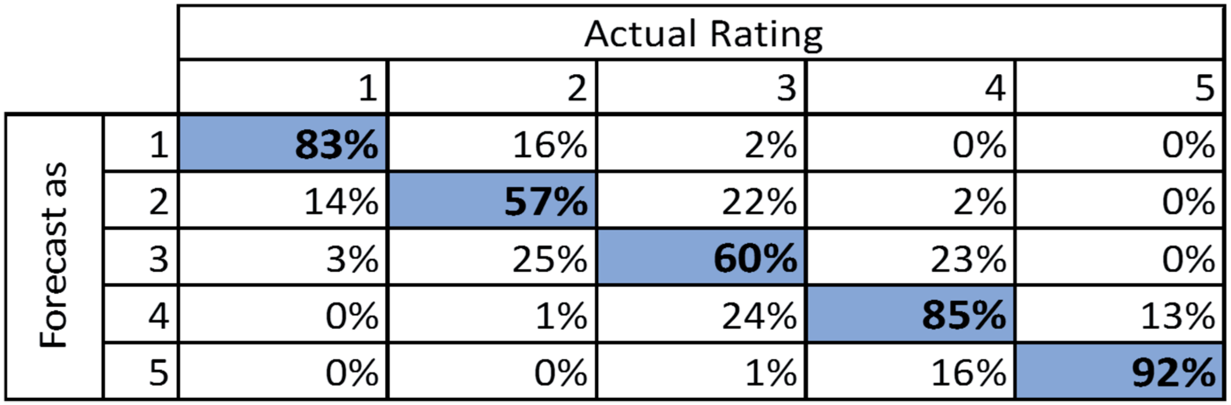
- Perceptions of **responsiveness to needs** connect directly to the **rep earning trust**, and to **dealing in a professional way**
- Dealing in a professional way also links directly to **inspires loyalty**
  - Loyalty also has a direct tie to having **the patient's best interests at heart**
- Having a **close working relationship** links most closely to the rep earning trust
  - We cannot say the rep earning trust causes a close working relationship, but this does seem like a logical inference from the diagram



Converge Analytic

This Net does exceptionally well predicting intent to use

- The overall correct prediction rate of 70% is very good, and even stronger than this number alone would suggest
- The extreme ratings are captured at much higher levels
  - 92% of top ratings correct
  - 85% of next highest correct
  - 83% of lowest correct
- Nearly all incorrect predictions are within one point of the actual rating, with none off by more than one point for the highest rating and no more than 3% off by one point anywhere



		Actual Rating				
		1	2	3	4	5
Forecast as	1	**83%**	16%	2%	0%	0%
	2	14%	**57%**	22%	2%	0%
	3	3%	25%	**60%**	23%	0%
	4	0%	1%	24%	**85%**	13%
	5	0%	0%	1%	16%	**92%**

Next steps: Networks that can show cause and effect

- As Pearl showed in his remarkable work,[1] with the right data this method indeed can find causal relationships
- Networks can find the right number of variables to include in the network and the right way to structure those variables
- For machine learning, these networks alone are proving to be worth the price of admission

Disneyland
JUNIOR
$2.50
35
ADMIT ONE
000001
Be a Guest of Walt Disney Productions and spend a few . . .
GREAT MOMENTS WITH MR. LINCOLN
ADMIT ONE JUNIOR (12 thru 17)
000001

Disclaimer: Mr. Lincoln not included

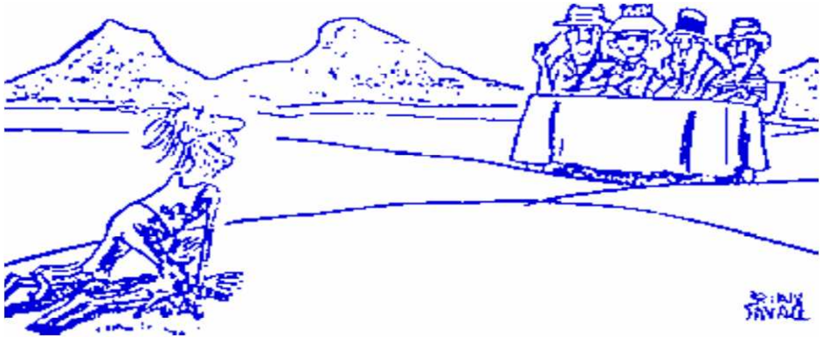
[1] *See for instance,* Causality: Models, Reasoning and Inference *(2009)*

What is in machine learning

(a little about how it works)

What is machine learning? Experts and near experts speak

- Several expert opinions, fortunately not 100% contradictory
  - “Computer programs able to induce patterns, regularities, or rules”[1]
  - “Subspecialty of artificial intelligence. . . developing methods for software to learn from experience or extract knowledge from examples”[2]
  - “Given training data. . . select the most probable hypothesis generating the data”[3]
  - “Overlaps heavily with statistics. . . but unlike statistics, machine learning is concerned with the algorithmic complexity of computational implementations”[4]
- Let’s put this together. . .



Thank God! A panel of experts

[1] amsglossary.allenpress.com/glossary/browse
[2] library.ahima.org/xpedio/groups/public/documents/ahima/pub_bok1_025042.html
[3] www.idsia.ch/~juergen/loconet/node2.html
[4] en.wikipedia.org/wiki/Machine_learning

Machine learning includes traditional methods

- Machine learning starts with established methods and adds new approaches
- Some relatively familiar methods:
  - Classification trees in particular
    - Trees' ability to create simple if-then "rules" lends itself well to machine learning
    - A good number of methods also grow out of classification trees
  - Clustering
  - Regression-based methods
- Some newer machine learning methods incorporate and extend more traditional approaches
  - (More on how this happens to follow.)

Not clear what this machine is learning

New: Taking many passes through the data and more

- New and useful
  - Taking many passes through, or looks at, a data set
  - Methods that actually "learn" from earlier passes through the data
  - New ways to determine the usefulness of information
    - e.g., weighing the cost of describing data vs. the information gained
  - Methods that supplement standard statistics
    - Methods based on graphical or spatial analysis
    - New methods to select useful variables and test conclusions

"New and useful" according to the Web

Machine learning at the far reaches adaptively teaches itself

- At the far end
  - "Applications we need Google or Apple to program," e.g.—
    - Speech recognition
      - "Siri, where can I find coffee?"
      - "Want a late afternoon, snack, Steven?"[1]
      - Autonomous driving
      - Google car: hundreds of thousands of miles with one human-driver caused accident
  - Self-customizing programs, e.g.—
    - Newsreader that learns preferences
    - Spam filter that learns what to exclude
- Nice as these may be. . .
  - We can get terrific insights without reaching this far

We finally are doing better

[1] *I am not making this up*

Why take many passes? Many weak estimates → stronger one

- A key insight: The average or consensus of many indifferent results can be better than any of them
  - Sometimes, the more weak estimates, the better

- Noise, or at least some uncertain values, even can help if you combine enough estimates
  - This might actually work out with the kinds of data we typically gather

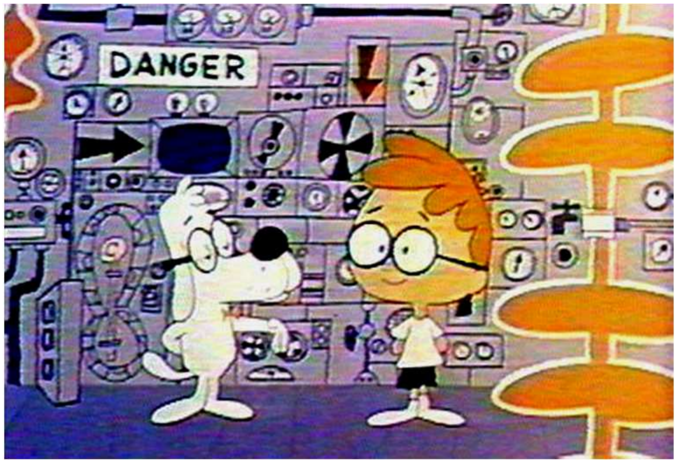
- All we need to do is know how, when and what to combine, while mastering some highly-inscrutable-seeming methods

On the Web, under the heading of "helpful noise"

We are here

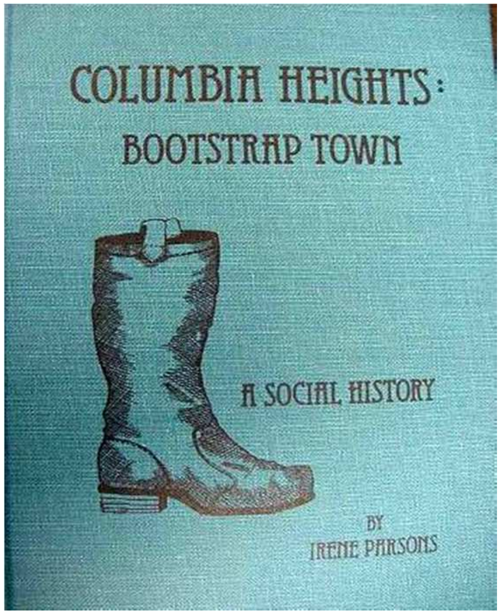
We have several ways to take many passes through data

- Passes can happen in any of several ways; these seem to be key
  - Taking many samples of the data
    - **Bagging** and many other related methods
  - Re-run the problem many times, learning from earlier runs and weighting results (a.k.a., **boosting**)
  - Randomly adding noise and re-running the problem
  - Randomly sampling possible predictors and rerunning the problem
  - Mixing several methods and averaging
  - “Experiments” that do not seem to be experiments (running and comparing many methods)



Bagging or repeated sampling can improve results

- **Bagging** combines many estimates
  - From multiple similar models, or
  - From the same model repeatedly
  - Most often based on random samples drawn repeatedly from the data set (or **bootstrapping**)
  - Helps reduce instability in complex models
- At the end, may use "voting" for classification, averaging for regression-type problems
- Another bad-sounding name—it comes from **bootstrap aggregating**
  - Mathematical types sometimes have a true fondness for ugly names



Bootstrapping from simpler times

Boosting runs many times and learns from earlier runs

- **Boosting** gets more accurate predictions by combining many estimates and weighting the results.
  - It can apply nearly any method to the data
  - First assigns each observation equal weight
    - Then computes predicted classifications.
    - Then, applies more weight to misclassified observations
    - Lower weight to classified correctly
  - Runs again with the reweighted data
  - Does this again, and again
  - Gets "votes" from all the runs
  - Weights them for an overall estimate
- A name to remember
  - **AdaBoost**, subject of much work

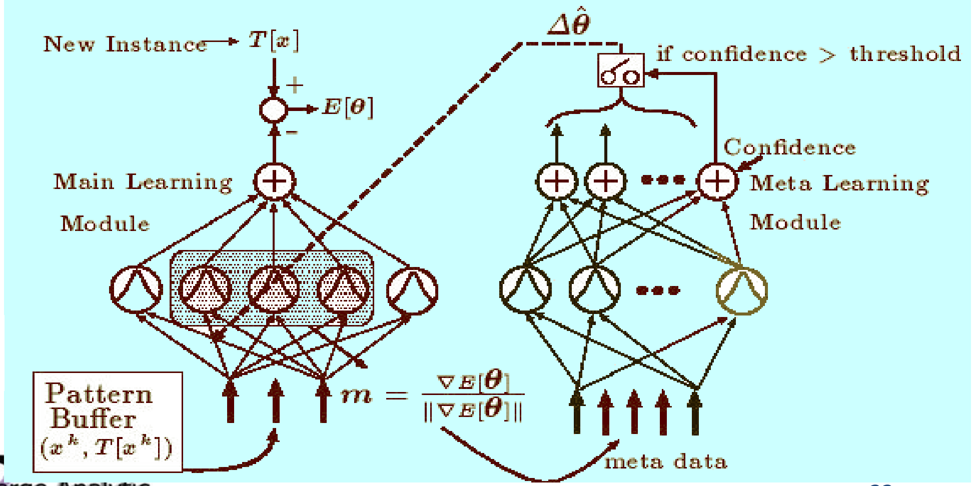
CLEARCHANNEL
4'9" Is The Magic Number!
Until Then, Kids Need A Booster Seat
visit www.boosterseat.gov

Not our kind of boosting
And 4'9"? What about grandma?

Meta-learners average or combine many estimates

- **Meta learner**
  - Anything that combines estimates from different methods, even wildly different ones
    - Includes bagging and boosting, most commonly thought to combine estimates from one method (or similar ones)
    - Also called **stacking**
    - Some methods are quite abstruse, although they seem to perform well.
      - e.g., the **decorate** method (or Diverse Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) chooses which methods go in the **ensemble** by testing first on artificial data sets it constructs
      - It may work well but it is hard to explain

Complex? Profound? Possibly both?



Validation, again with many passes through the data

- **Validation** tries to make realistic estimates of how well models work
- Traditional validation **holds out** part of the sample.
  - The model gets built on the rest of the sample, then tried on the part not used
- **N-fold cross-validation** gets wide use in machine learning[1]
  - First, we draw “n” random samples of all the data (usually 10);
    - Each of these samples is divided 90%/10%
  - The model gets built on the 90% portion and tested on the remaining 10%
    - This gets repeated for all 10 random samples
  - Accuracy gets averaged across all 10 runs
  - Very demanding, because the model gets tested repeatedly on small portions of the data set
- This discourages **over-fitting of results**

Runs\Folds	1	2	3	4	5	6	7	8	9	10
1	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟦
2	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟦	🟩
3	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟦	🟩	🟩
4	🟩	🟩	🟩	🟩	🟩	🟩	🟦	🟩	🟩	🟩
5	🟩	🟩	🟩	🟩	🟩	🟦	🟩	🟩	🟩	🟩
6	🟩	🟩	🟩	🟩	🟦	🟩	🟩	🟩	🟩	🟩
7	🟩	🟩	🟩	🟦	🟩	🟩	🟩	🟩	🟩	🟩
8	🟩	🟩	🟦	🟩	🟩	🟩	🟩	🟩	🟩	🟩
9	🟩	🟦	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
10	🟦	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

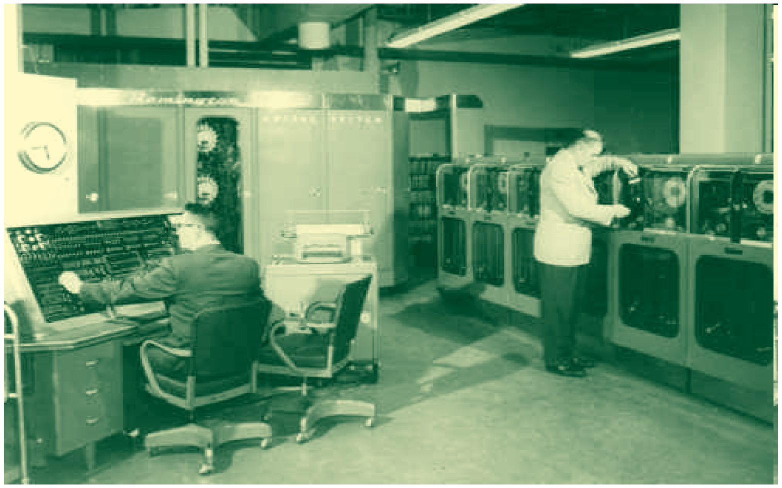
🟩 <-- Training set
🟦 <-- Validation set

[1] This is where we remind you not to panic if you don’t do statistics

Practical considerations

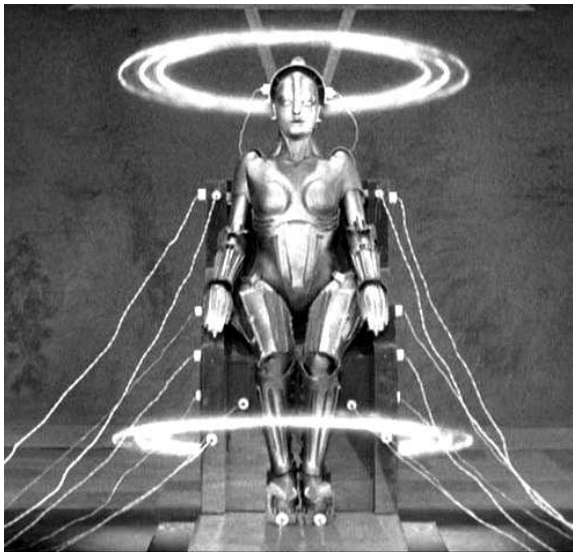
Examples: Two programs that work and are free

How the “open source” movement has helped



Open source has made machine learning affordable

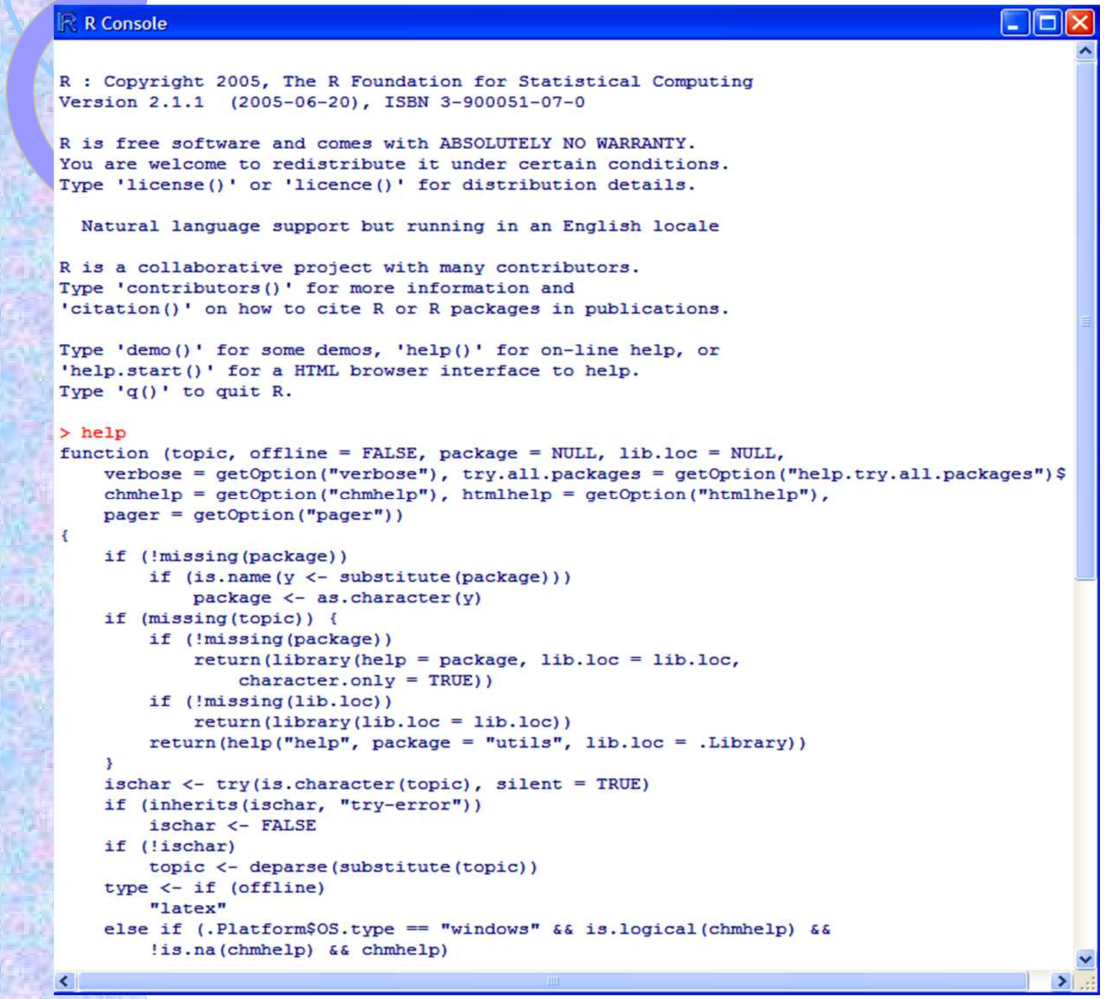
- In the old days, we had the **enterprise software** pay model[1]
  - Big and costly; just a few examples—
    - SPSS Clementine
    - SAS Enterprise Miner
    - Salford's commercial strength random forests and random trees
  - A growing countermovement contested this, building and distributing useful software—
    - Free and in the public domain
    - Open code
    - At times, less polished
    - Extensively supported by academia
  - These are real, and powerful, programs



May require enterprise level software

[1] **Enterprise** is secret code for costly (mostly very costly)

R was first: highly powerful and at times daunting



```
R : Copyright 2005, The R Foundation for Statistical Computing
Version 2.1.1  (2005-06-20), ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

> help
function (topic, offline = FALSE, package = NULL, lib.loc = NULL,
    verbose = getOption("verbose"), try.all.packages = getOption("help.try.all.packages")$
    chmhelp = getOption("chmhelp"), htmlhelp = getOption("htmlhelp"),
    pager = getOption("pager"))
{
    if (!missing(package))
        if (is.name(y <- substitute(package)))
            package <- as.character(y)
    if (missing(topic)) {
        if (!missing(package))
            return(library(help = package, lib.loc = lib.loc,
                character.only = TRUE))
        if (!missing(lib.loc))
            return(library(lib.loc = lib.loc))
        return(help("help", package = "utils", lib.loc = .Library))
    }
    ischar <- try(is.character(topic), silent = TRUE)
    if (inherits(ischar, "try-error"))
        ischar <- FALSE
    if (!ischar)
        topic <- deparse(substitute(topic))
    type <- if (offline)
        "latex"
    else if (.Platform$OS.type == "windows" && is.logical(chmhelp) &&
        !is.na(chmhelp) && chmhelp)
```

Your presenter's first session with R, preserved for posterity

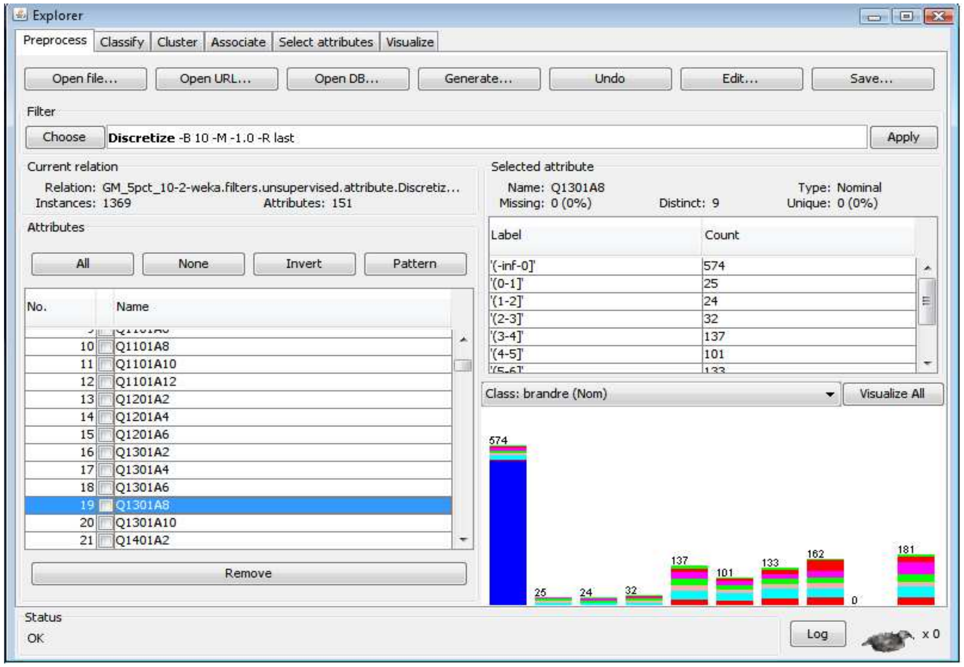
- R may be the oldest and best known[1] open source statistical programming environment
- Extensible if you master its language
- Intimidating for some of us
    - So huge it may be hard to know where to start
- There is a steep starting curve
- Some members of the R elite seem to find making it hard part of the fun

> "R syntax is sufficiently complex that it is difficult to write directly into the command window without making numerous syntax errors."
>
> *An Introduction to HB Modeling in R*

[1] *This obviously is highly relative*

Weka provides a start by organizing and stressing usability

- **Weka** takes a different path than standard R by offering accessible graphical user interfaces **(GUIs)**
  - Constantly evolving;
  - Extensive academic support;
  - You could write code but do not need to
- Built-in emphasis on data visualization and validation



Filter

Discretize -B 10 -M -1.0 -R last

Current relation

Relation: GM_5pct_10-2-weka.filters.unsupervised.attribute.Discretiz...
Instances: 1369 Attributes: 151

Selected attribute

Name: Q1301A8 Type: Nominal
Missing: 0 (0%) Distinct: 9 Unique: 0 (0%)

Label	Count
'(-inf-0]'	574
'(0-1]'	25
'(1-2]'	24
'(2-3]'	32
'(3-4]'	137
'(4-5]'	101
'(5-6]'	133

Attributes

No.		Name
10	□	Q1101A8
11	□	Q1101A10
12	□	Q1101A12
13	□	Q1201A2
14	□	Q1201A4
15	□	Q1201A6
16	□	Q1301A2
17	□	Q1301A4
18	□	Q1301A6
19	□	Q1301A8
20	□	Q1301A10
21	□	Q1401A2

Class: brandre (Nom)

Status

OK

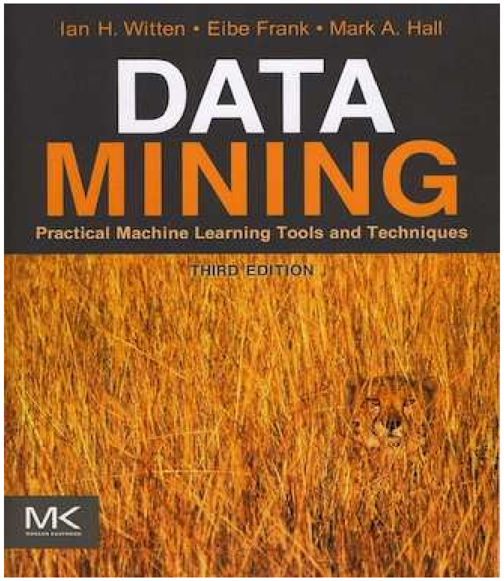
Not our type of "gooey"
(plastic ducks optional)

WEKA
The University of Waikato

No relative of the duck

Converge Analytic

Weka works and is explained by an excellent book


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

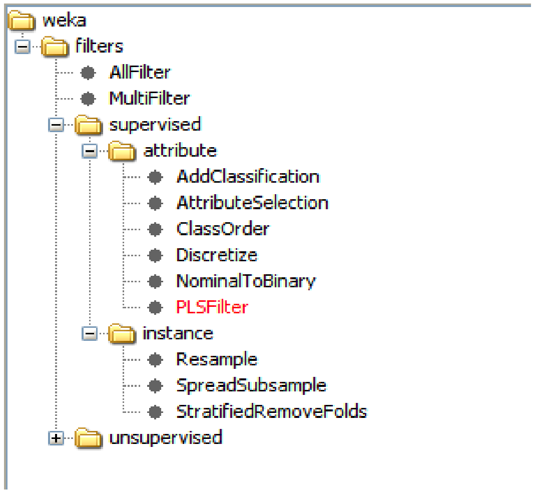
- Weka has a truly great book about its methods and theories (Witten & Frank), running 525 pages
  - It at least touches on many program features and approaches (as of 2011)[1]
  - Actually readable and affordable, unlike many academic publications
- It is possible to get Weka to work on the first try (especially if you read part of the book)
  - Good for highly analytical and more time-pressured people
- The book and program encourage experimentation and comparison of methods

[1] Data Mining: Practical Machine Learning Tools and Techniques , 3rd Edition (2011)

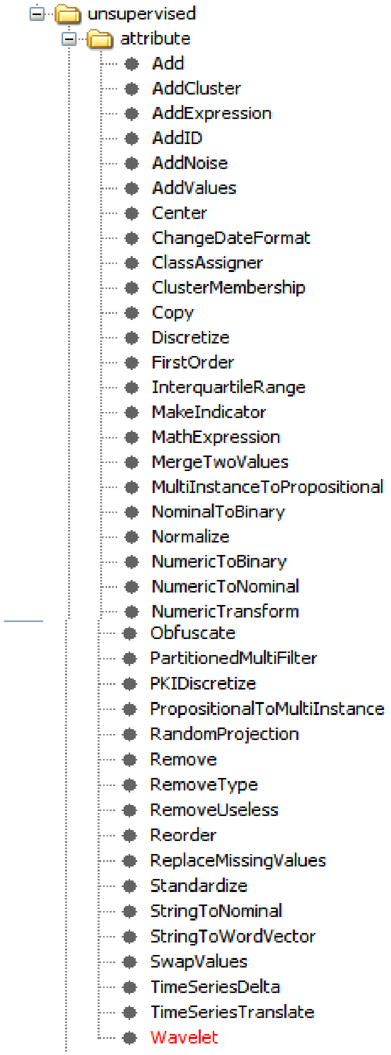
We are here

Weka has become vast

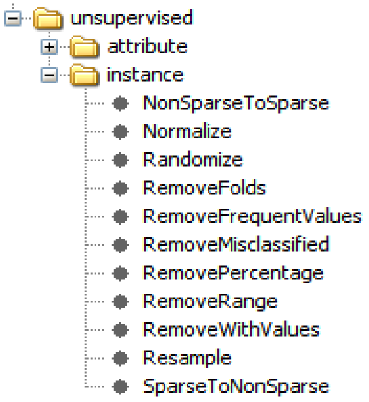
- These are just the ways in which Weka can process data before analysis
- We won’t discuss them, other than pointing out that this is an amazing number of options



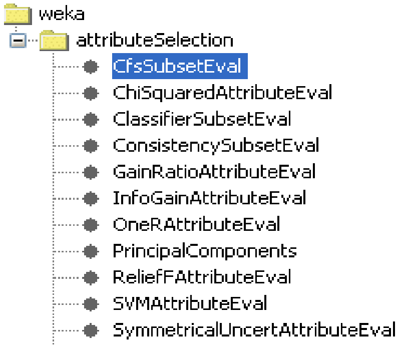
- weka
  - filters
    - AllFilter
    - MultiFilter
    - supervised
      - attribute
        - AddClassification
        - AttributeSelection
        - ClassOrder
        - Discretize
        - NominalToBinary
        - PLSFilter
      - instance
        - Resample
        - SpreadSubsample
        - StratifiedRemoveFolds
    - unsupervised



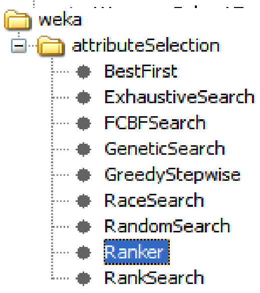
- unsupervised
  - attribute
    - Add
    - AddCluster
    - AddExpression
    - AddID
    - AddNoise
    - AddValues
    - Center
    - ChangeDateFormat
    - ClassAssigner
    - ClusterMembership
    - Copy
    - Discretize
    - FirstOrder
    - InterquartileRange
    - MakeIndicator
    - MathExpression
    - MergeTwoValues
    - MultiInstanceToPropositional
    - NominalToBinary
    - Normalize
    - NumericToBinary
    - NumericToNominal
    - NumericTransform
    - Obfuscate
    - PartitionedMultiFilter
    - PKIDiscretize
    - PropositionalToMultiInstance
    - RandomProjection
    - Remove
    - RemoveType
    - RemoveUseless
    - Reorder
    - ReplaceMissingValues
    - Standardize
    - StringToNominal
    - StringToWordVector
    - SwapValues
    - TimeSeriesDelta
    - TimeSeriesTranslate
    - Wavelet



- unsupervised
  - attribute
  - instance
    - NonSparseToSparse
    - Normalize
    - Randomize
    - RemoveFolds
    - RemoveFrequentValues
    - RemoveMisclassified
    - RemovePercentage
    - RemoveRange
    - RemoveWithValues
    - Resample
    - SparseToNonSparse



- weka
  - attributeSelection
    - CfsSubsetEval
    - ChiSquaredAttributeEval
    - ClassifierSubsetEval
    - ConsistencySubsetEval
    - GainRatioAttributeEval
    - InfoGainAttributeEval
    - OneRAttributeEval
    - PrincipalComponents
    - ReliefFAttributeEval
    - SVMAttributeEval
    - SymmetricalUncertAttributeEval

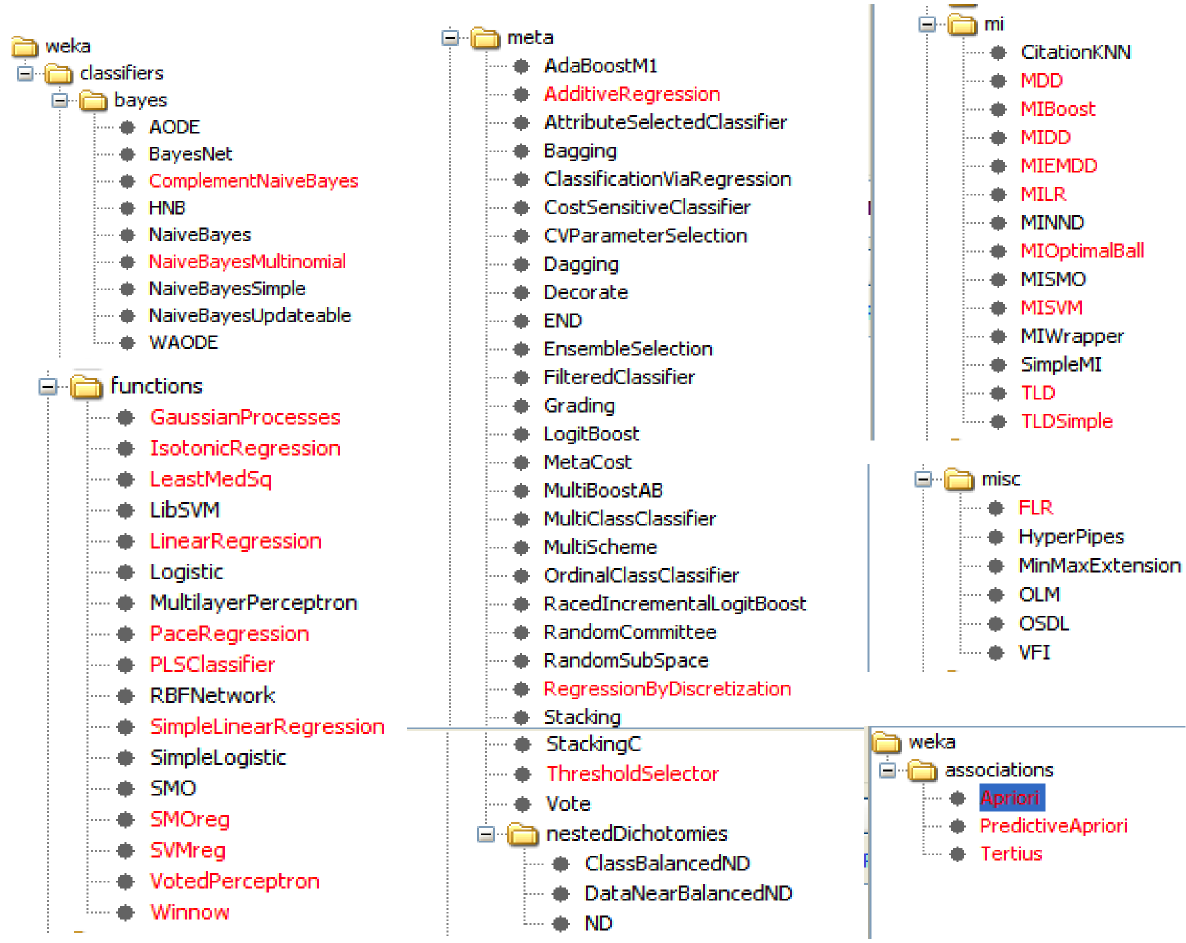


- weka
  - attributeSelection
    - BestFirst
    - ExhaustiveSearch
    - FCBFSearch
    - GeneticSearch
    - GreedyStepwise
    - RaceSearch
    - RandomSearch
    - Ranker
    - RankSearch

Red = new since the second edition 2001 book

A few more Weka menus (for analysis)

- Again, this just shows the wide variety of methods--many really new



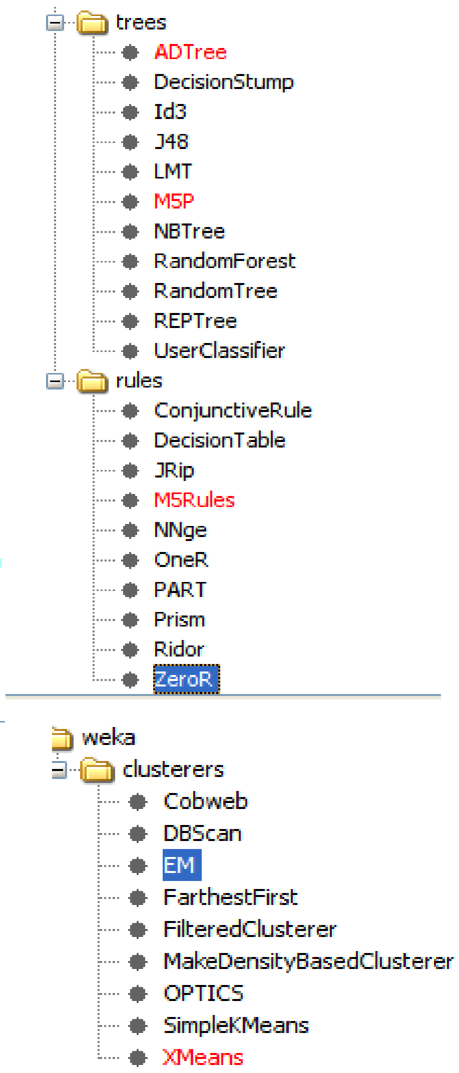
- weka
  - classifiers
    - bayes
      - AODE
      - BayesNet
      - ComplementNaiveBayes
      - HNB
      - NaiveBayes
      - NaiveBayesMultinomial
      - NaiveBayesSimple
      - NaiveBayesUpdateable
      - WAODE
    - functions
      - GaussianProcesses
      - IsotonicRegression
      - LeastMedSq
      - LibSVM
      - LinearRegression
      - Logistic
      - MultilayerPerceptron
      - PaceRegression
      - PLSClassifier
      - RBFNetwork
      - SimpleLinearRegression
      - SimpleLogistic
      - SMO
      - SMOreg
      - SVMreg
      - VotedPerceptron
      - Winnow

- meta
  - AdaBoostM1
  - AdditiveRegression
  - AttributeSelectedClassifier
  - Bagging
  - ClassificationViaRegression
  - CostSensitiveClassifier
  - CVParameterSelection
  - Dagging
  - Decorate
  - END
  - EnsembleSelection
  - FilteredClassifier
  - Grading
  - LogitBoost
  - MetaCost
  - MultiBoostAB
  - MultiClassClassifier
  - MultiScheme
  - OrdinalClassClassifier
  - RacedIncrementalLogitBoost
  - RandomCommittee
  - RandomSubSpace
  - RegressionByDiscretization
  - Stacking
  - StackingC
  - ThresholdSelector
  - Vote
  - nestedDichotomies
    - ClassBalancedND
    - DataNearBalancedND
    - ND

- mi
  - CitationKNN
  - MDD
  - MIBoost
  - MIDD
  - MIEMDD
  - MILR
  - MINND
  - MIOptimalBall
  - MISMO
  - MISVM
  - MIWrapper
  - SimpleMI
  - TLD
  - TLDSimple

- misc
  - FLR
  - HyperPipes
  - MinMaxExtension
  - OLM
  - OSDL
  - VFI

- weka
  - associations
    - Apriori
    - PredictiveApriori
    - Tertius



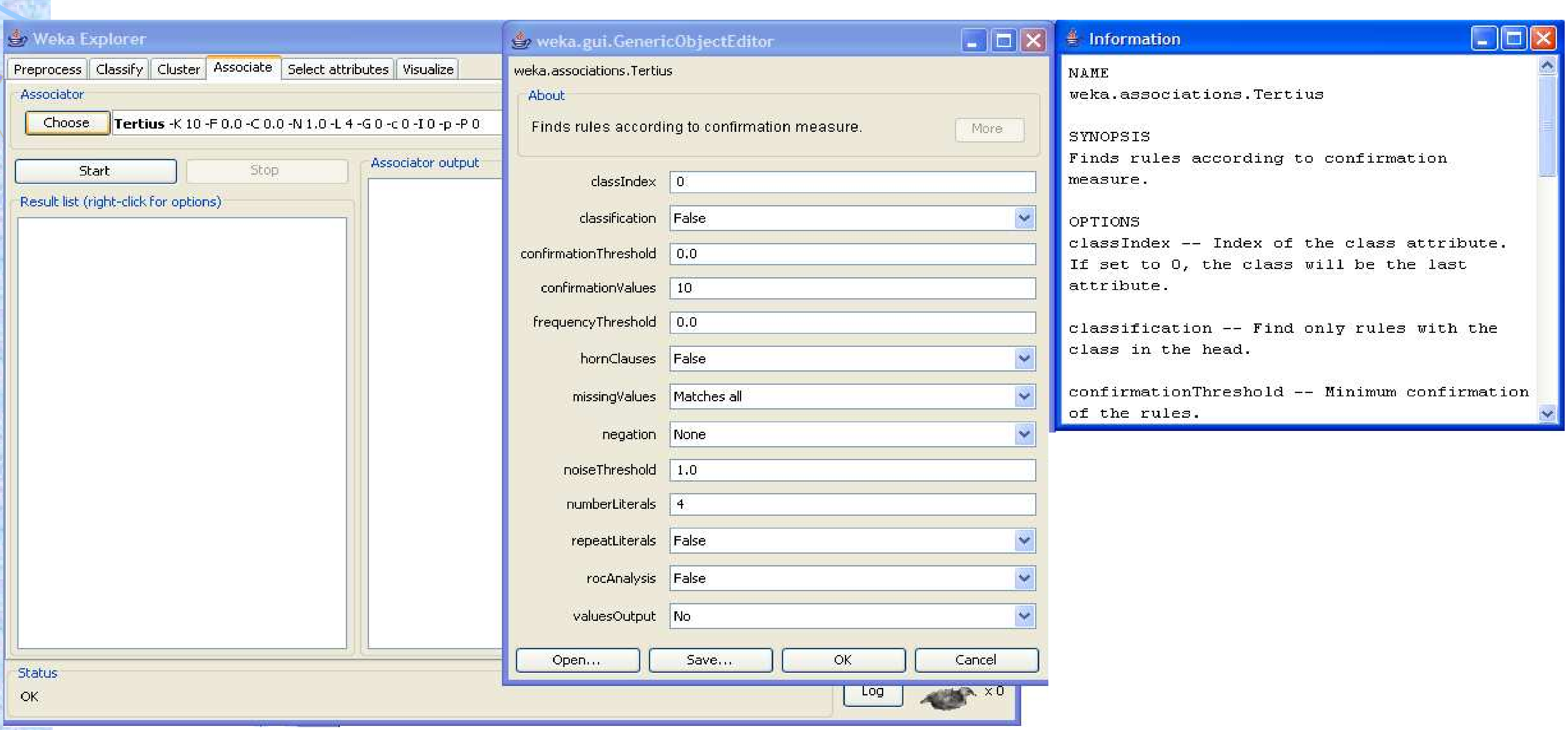
- trees
  - ADTree
  - DecisionStump
  - Id3
  - J48
  - LMT
  - M5P
  - NBTree
  - RandomForest
  - RandomTree
  - REPTree
  - UserClassifier
- rules
  - ConjunctiveRule
  - DecisionTable
  - JRip
  - M5Rules
  - NNge
  - OneR
  - PART
  - Prism
  - Ridor
  - ZeroR

- weka
  - clusterers
    - Cobweb
    - DBScan
    - EM
    - FarthestFirst
    - FilteredClusterer
    - MakeDensityBasedClusterer
    - OPTICS
    - SimpleKMeans
    - XMeans

Red = new since the second edition 2001 book

It helps to know the methods (we have a lot to learn)

- Tertius, anyone?



Tertius -K 10 -F 0.0 -C 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -p -P 0

weka.associations.Tertius

Finds rules according to confirmation measure.

classIndex	0
classification	False
confirmationThreshold	0.0
confirmationValues	10
frequencyThreshold	0.0
hornClauses	False
missingValues	Matches all
negation	None
noiseThreshold	1.0
numberLiterals	4
repeatLiterals	False
rocAnalysis	False
valuesOutput	No

```
NAME
weka.associations.Tertius

SYNOPSIS
Finds rules according to confirmation
measure.

OPTIONS
classIndex -- Index of the class attribute.
If set to 0, the class will be the last
attribute.

classification -- Find only rules with the
class in the head.

confirmationThreshold -- Minimum confirmation
of the rules.
```

© 1991-2013 Steven M. Struhl

Converge Analytic

Not pretty as is, but powerful output from AdaBoostMI

Begin output from AdaBoostMI (trimmed)

```
Decision Stump Classifications
DermAsmtPSOBSA Weight: 1.63
    Up to 10%/Over 10%
DurationPsoriasis Weight: 0.42
    Up to 5 yrs/Over 5 yrs
BodyMassInd Weight: 0.31
    Up to 30/Over 30
Sex Weight 0.27
    Male (0)/Female (1)
DermAsmtPSOBSA Weight: 0.68
    Up to 20%/Over 20%
Anxiety Weight: 0.29
    Scale to 7/Scale 7 and up
LowerExtremitiesInvolve Weight: 0.24
    No (0)/Yes (1)
PriorSystemicPhotoTx Weight: 0.22
    No (0)/Yes (1)
Age(Years) Weight: 0.46
    Up to 35 (0)/ Over 25 (1)
UpperExtremitiesInfiltration Weight: 0.16
    No (0)/Yes (1)
UpperExtremitiesDesquamation Weight: 0.14
    No (0)/Yes (1)
```

Note
This shows the relative importances and where variables split to create differences

End of output from AdaBoostMI

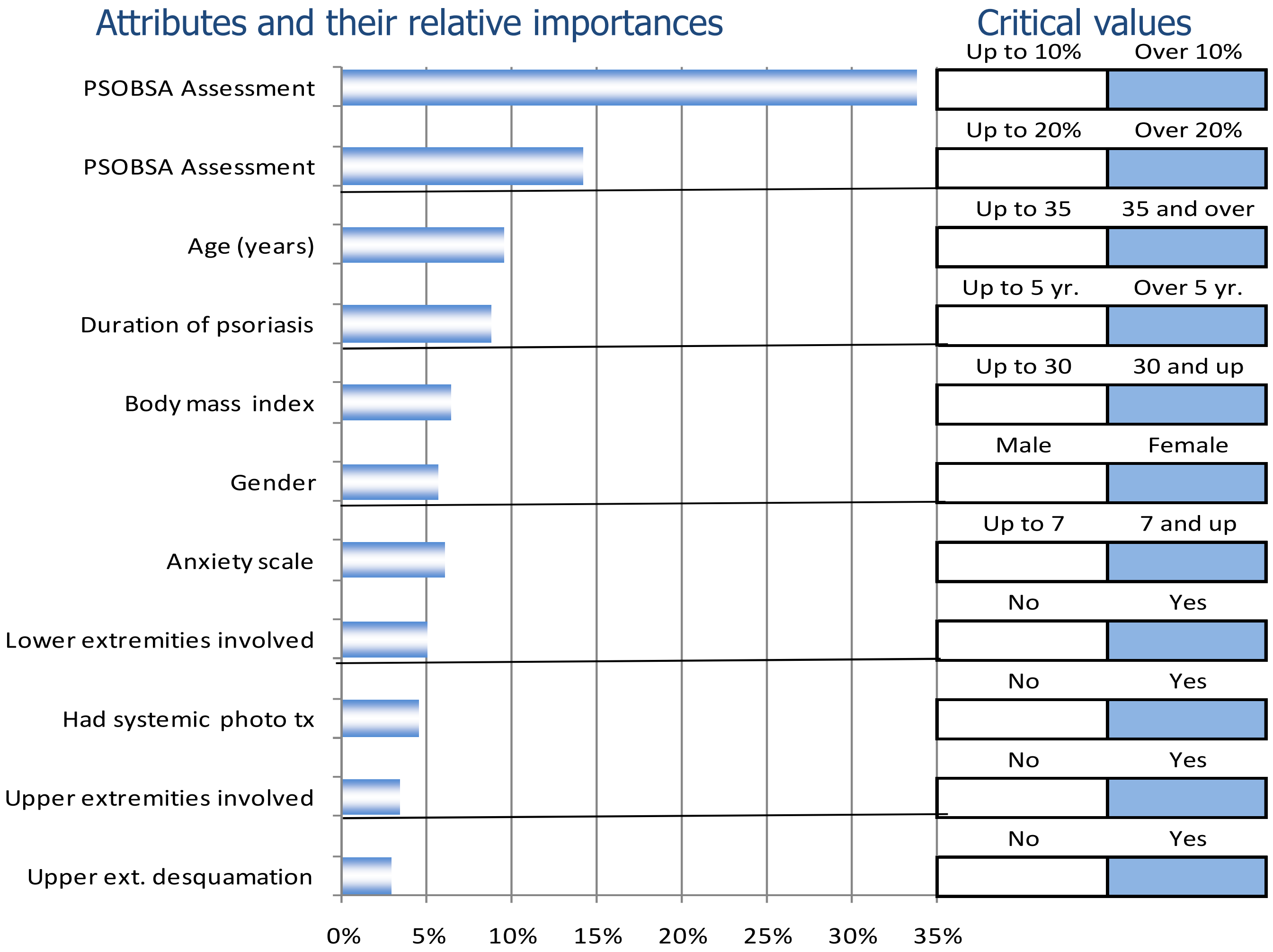
```
Number of performed Iterations: 32
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2512      77.4 %
Incorrectly Classified Instances       733      22.6 %
Total Number of Instances             3245
Ignored Class Unknown Instances                  15

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.966     0.811     0.784     0.966     0.866     '0'
  0.189     0.034     0.643     0.189     0.292     '1'

=== Confusion Matrix ===

    a    b   <-- classified as
 2361   84 |    a = '0'
  649  151 |    b = '1'
```

Notes
77% overall is noticeably better than a carefully hand-tuned classification tree, and correct classification of the smaller "1" class at 19% (in the TP or true positive column) also is far better.

A person is correctly classified only if her likelihood of being in a group is over 50% so even 49% for group 1 (2.45 times as likely as the average) does not count as correct

How boosting output becomes the slide we saw earlier

- The raw output is put into user-accessible form
- As a reminder, from a study of psoriasis sufferers, using measurements taken in exams to forecast whether the patient is at risk for serious depression

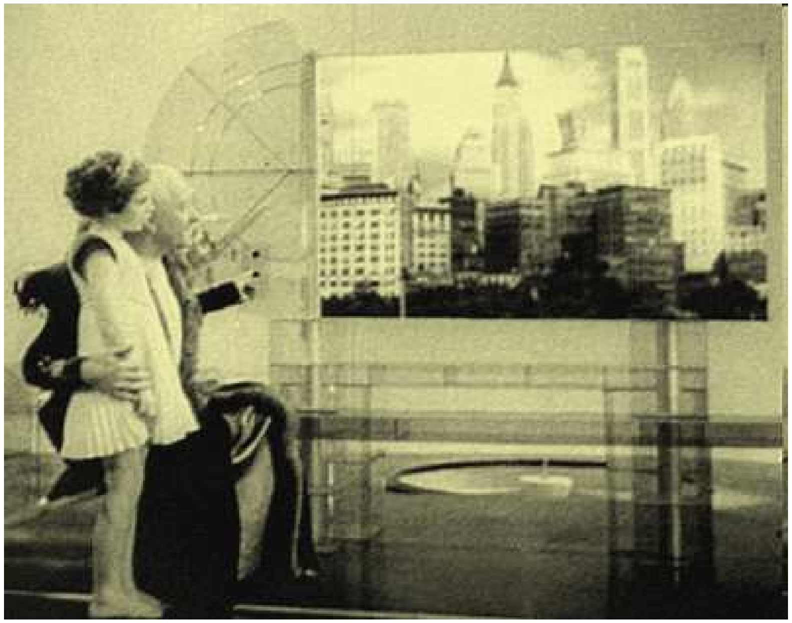


Note that this found two breaking points on the same variable with different importances in predicting "at risk" scores

The group associated with higher "at risk of serious depression" scores is shaded

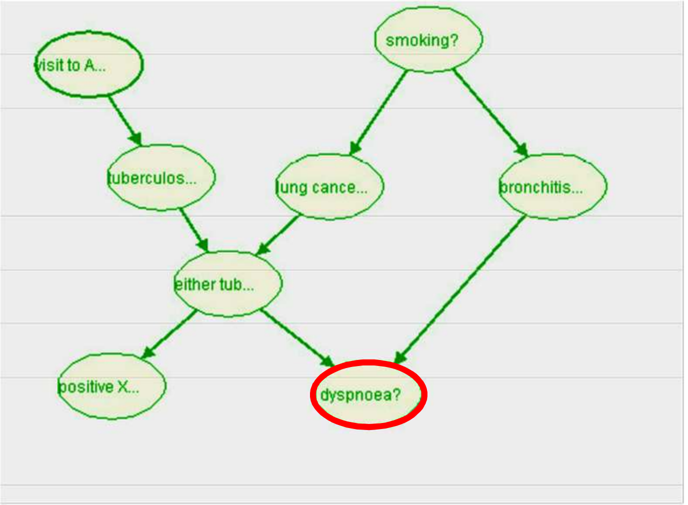
Looking forward

Starting with some methods that have worked already



A sampling of methods that provide useful output (so far)

- **Bayes Nets** (we saw these)
  - Really remarkable
  - These can start by organizing themselves
  - Provide deep insights into complex structures
  - Even used now to show cause and effect in the hard sciences



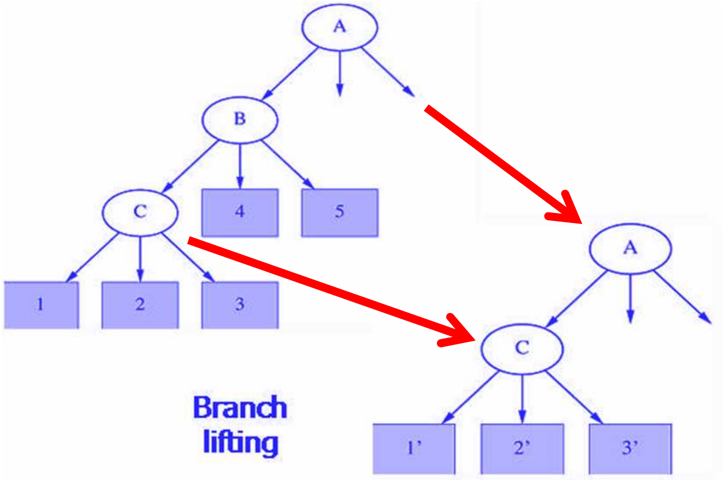
Dependent variable highlighted

- **Random forests/random trees**
  - Can do very well making and applying models
  - However, models typically are too complex to understand well
  - Programs could do better explaining whatever we might understand
- **Model trees**
  - Regression models at the ends of short trees
    - Splits sample first, then builds regression models on the split groups

More sampling of methods provide useful output (so far)

- **AdaBoost or boosting** (we saw this)
  - Highly adaptable to many problems, using many methods
  - Can give new insights into variable importances
- **EM clustering**
  - Has done really well grouping respondents using mixes of nominal and continuous variables
- **AODE (and WAODE) classification**
  - Better readings of performance with categorical and continuous variables than discriminant analysis or multi-nomial logit
- **C4.5 (J4.8) CHAID program**
  - Can do **pruning** and **branch lifting**
    - The ultimate in making compact trees
- **New forms of variable selection**
  - Many non-linear (e.g., **genetic algorithms**) can help winnow variables for analysis

C is moved up one row, eliminating B



Conclusions: Results now and the future looks productive

- Useful now but still very much in progress
  - Some real improvements mixed with others that now appear novel but useless
  - Much that needs evaluation for worth
- Methods that require new thinking
  - Some that we now must rely on a computer to understand
  - Some still perplexing approaches
- However—
  - We have started seeing real analytical gains
  - This holds tremendous promise for further developments

Coming soon?

Questions? Comments? Need more information?

Dr. Steven Struhl

smstruhl@convergeanalytic.com

smstruhl@gmail.com

☎ (847) 624-2268

Converge Analytic