IND E 498 Special Topics on Data Analytics

Instructor: Prof. Shuai Huang

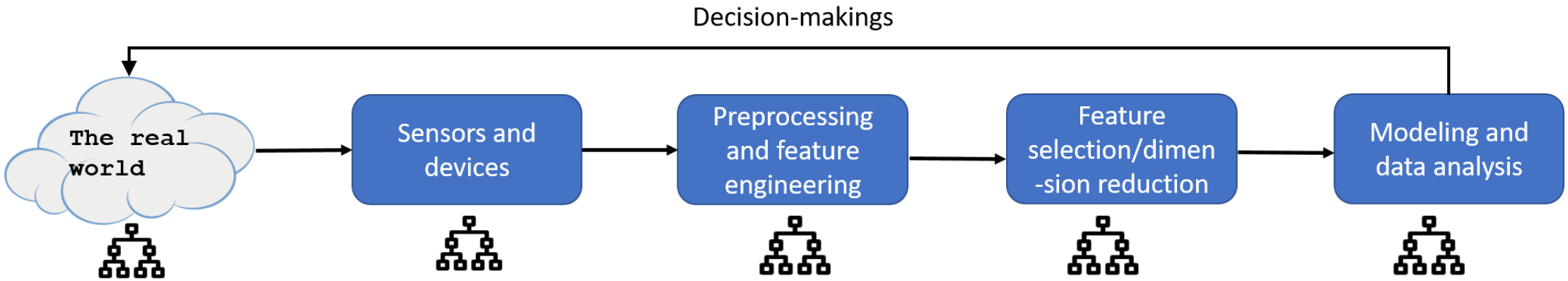
Industrial and Systems Engineering

University of Washington

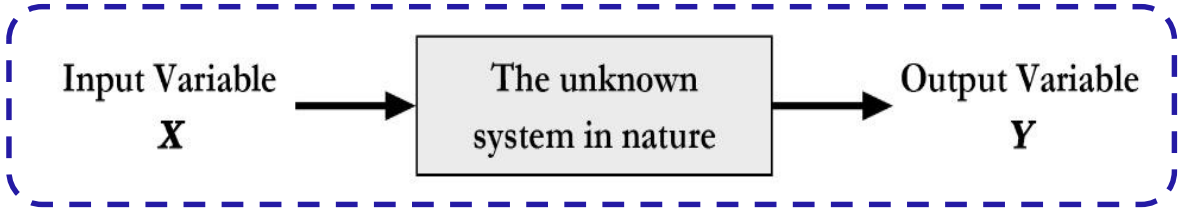
Overview of the course

- Course website (http://analytics.shuaihuang.info/)
- Syllabus
- Study group
- Data sources/R/stackoverflow/github
- Project meetings

A typical data analytics pipeline



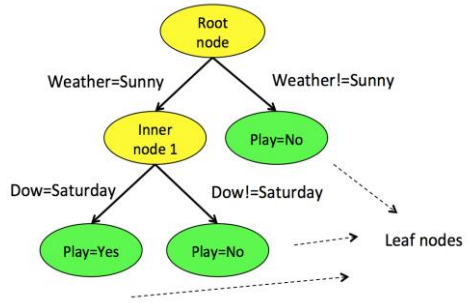
The two cultures of statistical modeling



$$y \stackrel{?}{=} f(\boldsymbol{x}) + \epsilon$$

	$f(x)$	ϵ	"Cosmology"
Data Modeling	Explicit form (e.g., linear regression)	Statistical distribution (e.g., Gaussian)	Imply Cause and effect; articulate uncertainty
Algorithmic Modeling	Implicit form (e.g., tree model)	Rarely modeled as structured uncertainty; only acknowledged as meaningless noise	Look for accurate surrogate for prediction; to fit the data rather than to explain the data

$$f(x) = \beta_0 + \beta_1 x$$



Key topics in regression models

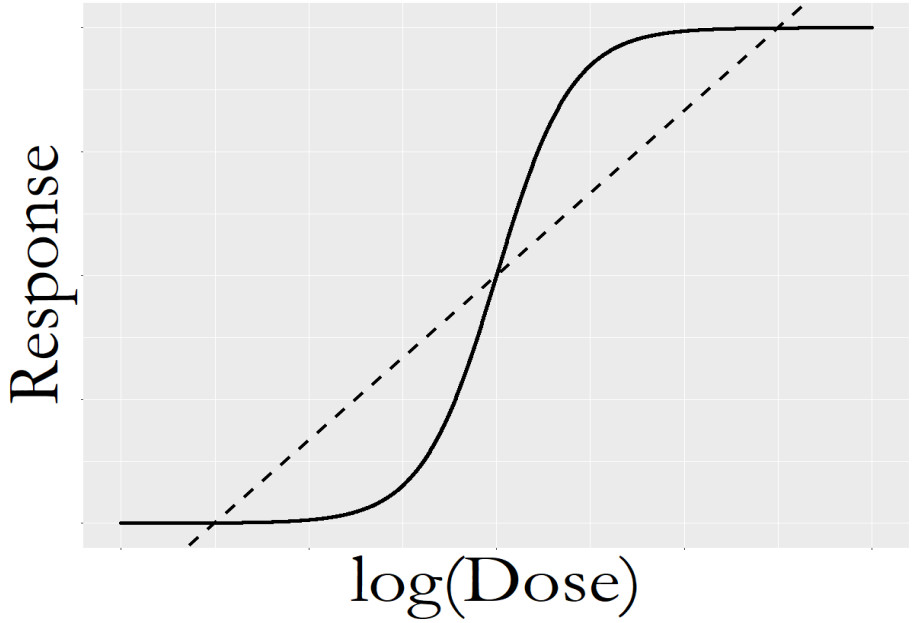
- Chapter 2: Linear regression, least-square estimation, hypothesis testing, why normal distribution, its connection with experimental design, R-squared.
- Chapter 3: Logistic regression, generalized least square estimation, iterative reweighted least square (IRLS) algorithm, approximated hypothesis testing, Ranking as a linear regression
- Chapter 4: Bootstrap, data resampling, nonparametric hypothesis testing, nonparametric confidence interval
- Chapter 5: Overfitting and underfitting, limitation of R-squared, training dataset and testing dataset, random sampling, K-fold cross validation, the confusion matrix, false positive and false negative, and Receiver Operating Characteristics (ROC) curve
- Chapter 6: Residual analysis, normal Q-Q plot, Cook's distance, leverage, multicollinearity, subset selection, heterogeneity, clustering, gaussian mixture model (GMM), and the Expectation-Maximization (EM) algorithm
- Chapter 7: Support Vector Machine (SVM), generalize data versus memorize data, maximum margin, support vectors, model complexity and regularization, primal-dual formulation, quadratic programming, KKT condition, kernel trick, kernel machines, SVM as a neural network model
- Chapter 8: LASSO, sparse learning, L1-norm and L2-norm regularization, Ridge regression, feature selection, shooting algorithm, Principal Component Analysis (PCA), eigenvalue decomposition, scree plot
- Chapter 9: Kernel regression as generalization of linear regression model, kernel functions, local smoother regression model, k-nearest regression model, conditional variance regression model, heteroscedasticity, weighted least square estimation, model extension and stacking

Key topics in tree models

- Chapter 2: Decision tree, entropy gain, node splitting, pre- and post-pruning, empirical error, generalization error, pessimistic error by binomial approximation, greedy recursive splitting
- Chapter 4: Random forest, Gini index, weak classifiers, probabilistic mechanism why random forest works
- Chapter 5: Out-of-bag (OOB) error in random forest
- Chapter 6: Importance score, partial dependency plot, residual analysis
- Chapter 7: Ensemble learning, Adaboost, sampling with (or without) replacement
- Chapter 8: Importance score in random forest, regularized random forests (RRF), guided regularized random forests (GRRF)
- Chapter 9: System monitoring reformulated as classification, real-time contrasts method (RTC), design of monitoring statistics, sliding window, anomaly detection, false alarm
- Chapter 10: Integration of tree models, feature selection, and regression models in inTrees, random forest as a rule generator, rule extraction, pruning, selection, and summarization, confidence and support of rules, variable interactions, rule-based prediction

Key concepts – significance versus truth

- Statistical modeling is to pursue statistical significance
- In other words, it may not be true, but it is significant



Key concepts – The rhetoric of “what if”

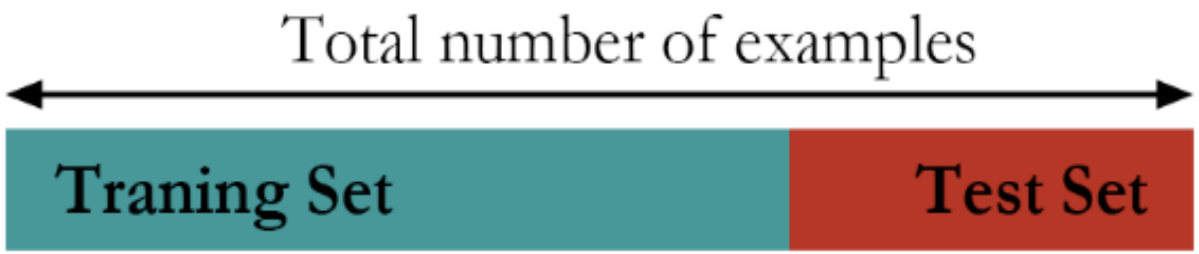
- “Luckily, the data is not contradictory with our hypothesis/theory”
- You will rarely hear statisticians say that, “luckily, we accept the null hypothesis”

Hypothesis testing: Pr(data | Null hypothesis is true)
Truth seeking: Pr(Null hypothesis is true | data)

This mentality, the “negative” reading of data, is one foundation of classic statistics

Key concepts – Training/testing data

- Instead of establishing the significance of the model by hypothesis testing, modern machine learning models establish the significance of the model by, roughly speaking, the paradigm of “training/testing data”



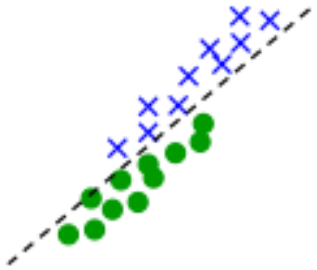
Key concepts – feature

"Good" features

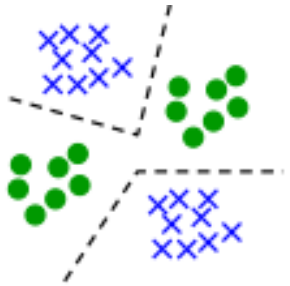
"Bad" features

Linear separability

Non-linear separability



Highly correlated features



Multi-modal

A side story about features

Predictive Segmentation of Populations

TAGS: Computer Science/Information Technology, Math/Statistics, Life Sciences, RTP

STATUS: **Awarded** | ACTIVE SOLVERS: 865 | POSTED: 8/31/10

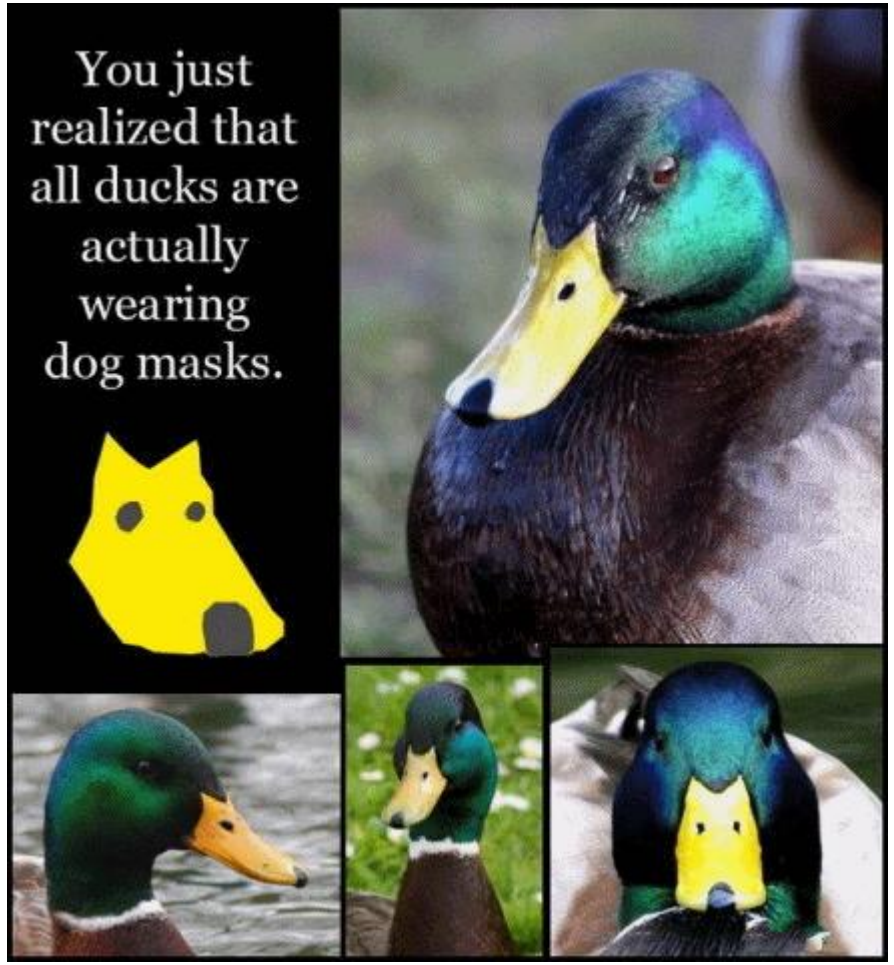
The Seeker is looking for novel means to detect and construct population segmentation that is relevant and predictive. More details and data are provided in this Challenge.

Source: InnoCentive **Challenge ID:** 9667082

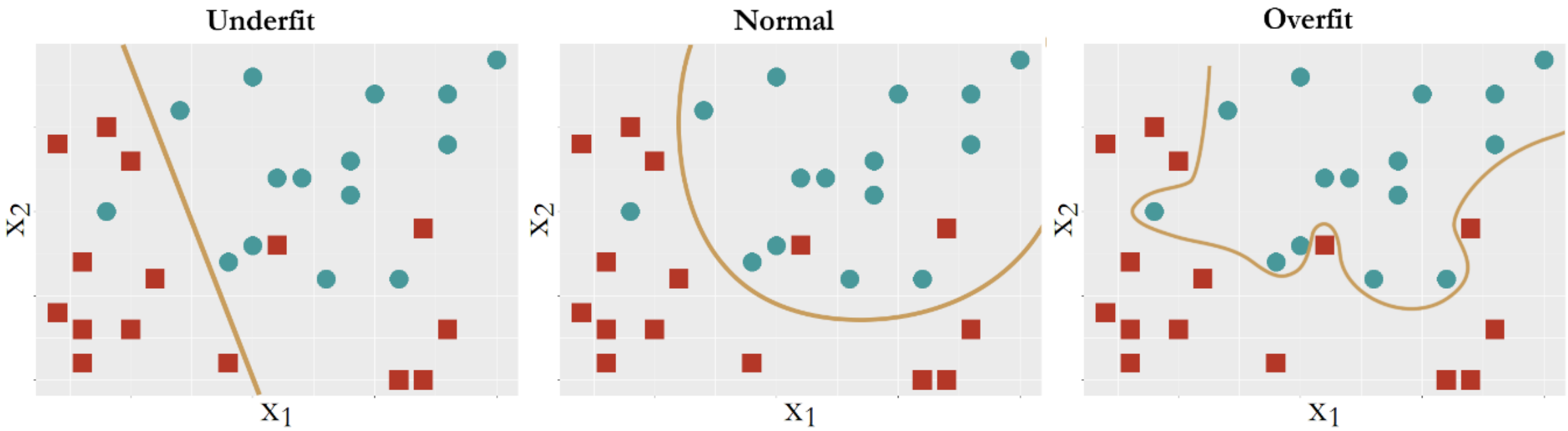
Challenge Overview

The Seeker is looking for novel means to detect and construct population segmentation that is relevant and predictive. Though this problem originally stems from applications in clinical trials for drug development, no prior experience in that specific domain is needed in order to solve this Challenge. More details and data are provided in this Challenge.

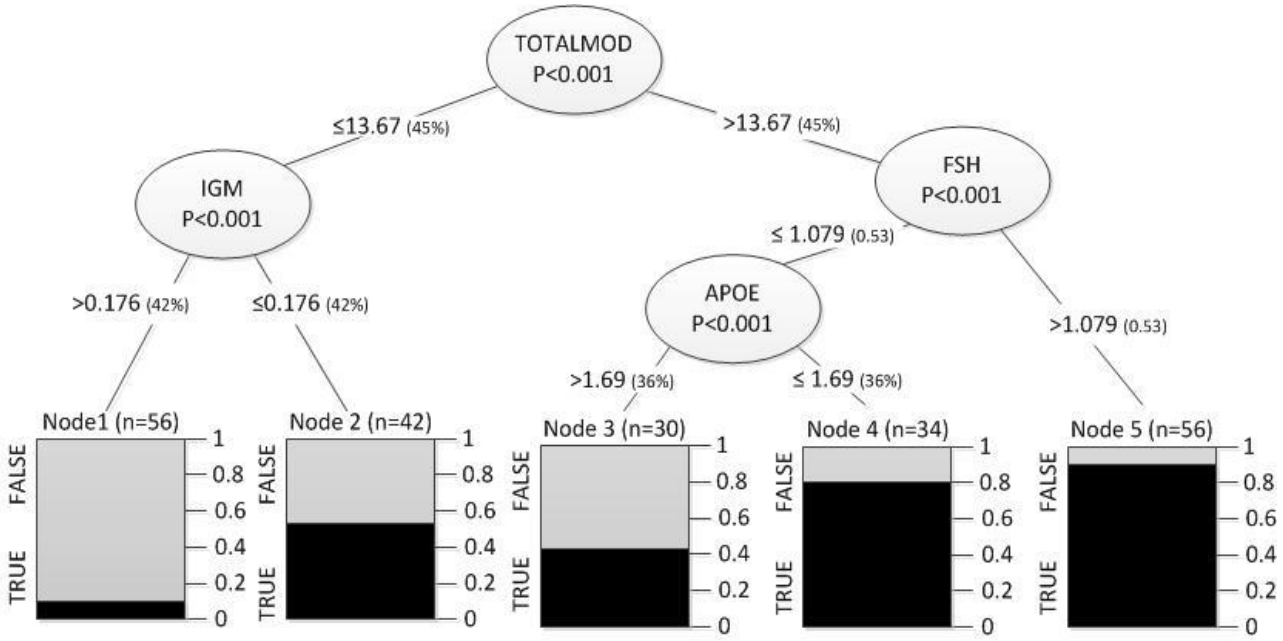
Another story about features …



Key concepts – overfitting/generalization



Key concepts – context



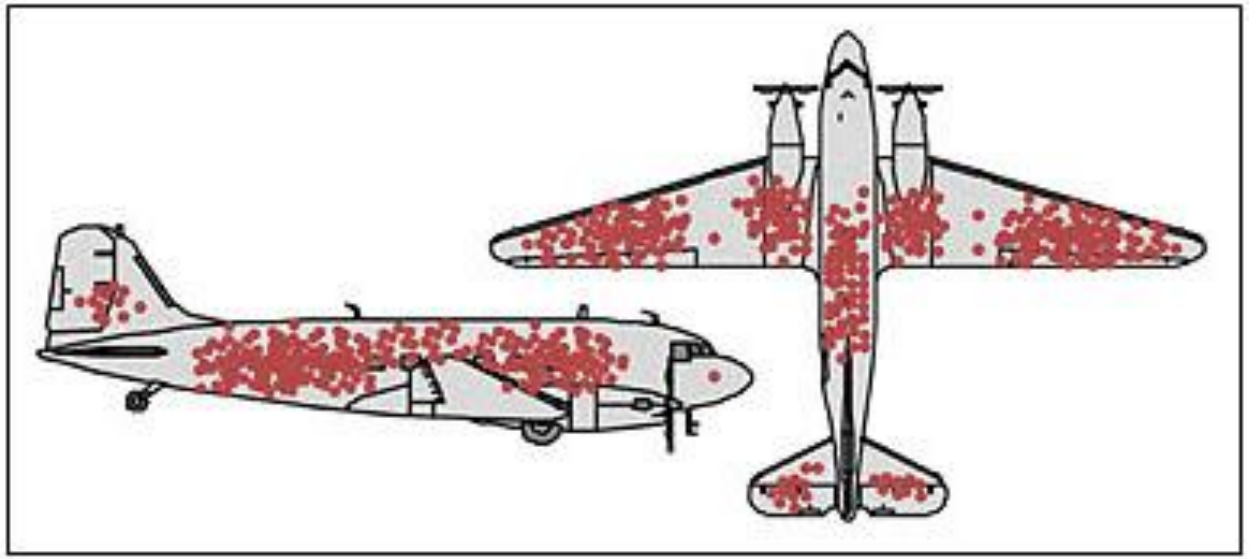
Why 60% accuracy is still very valuable

- **Anti-amyloid clinical trials need large-scale screening: $3,000 per PET scan**
- **If the PET scan shows negative result, $3,000 is a waste**
- **Blood measurements cost $200 per visit**
- **Question: can we use blood measurements to predict the amyloid?**
- **Benefit: enrich the cohort pool with more amyloid positive cases**

Key concepts – insight

The story of the statistician Abraham Wald in World War II

- The Allied AF lost many aircrafts, so they decided to armor their aircrafts up
- However, limited resources are available – which parts of the aircrafts should be armored up?
- Abraham Wald stayed in the runaway, to catalog the bullet holes on the returning aircrafts



Credit: Cameron Moll