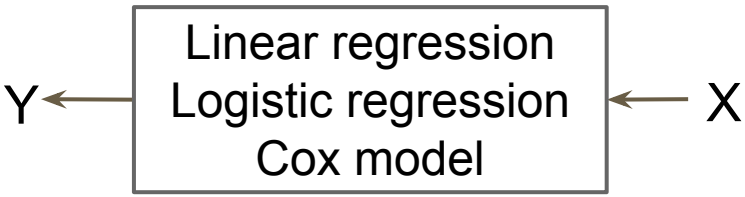
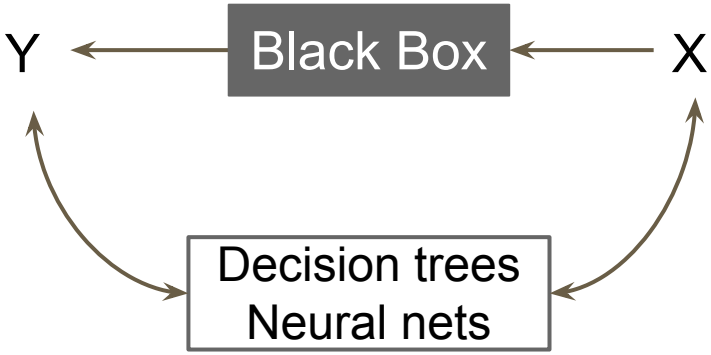
The Conflict of Two Cultures in Statistical Modeling

Yuran Liang, Daiwei Zhang,
Yongwen Zhuang, Yuqi Zhai

Introduction

	Data Modeling Culture	Algorithmic Modeling Culture
Model		
Validation	Goodness-of-fit tests Residual examination	Predictive accuracy
Estimated culture population	98%	2%

Data Models: Problems

- Model fitting
  - Goodness-of-fit
    - Rejects non-linearity only when extreme
    - Not applicable when variables deleted or non-linear terms added
  - Residual analysis
    - Unreliable in more than 4-5 dimensions
    - Too many ways to analyze residuals

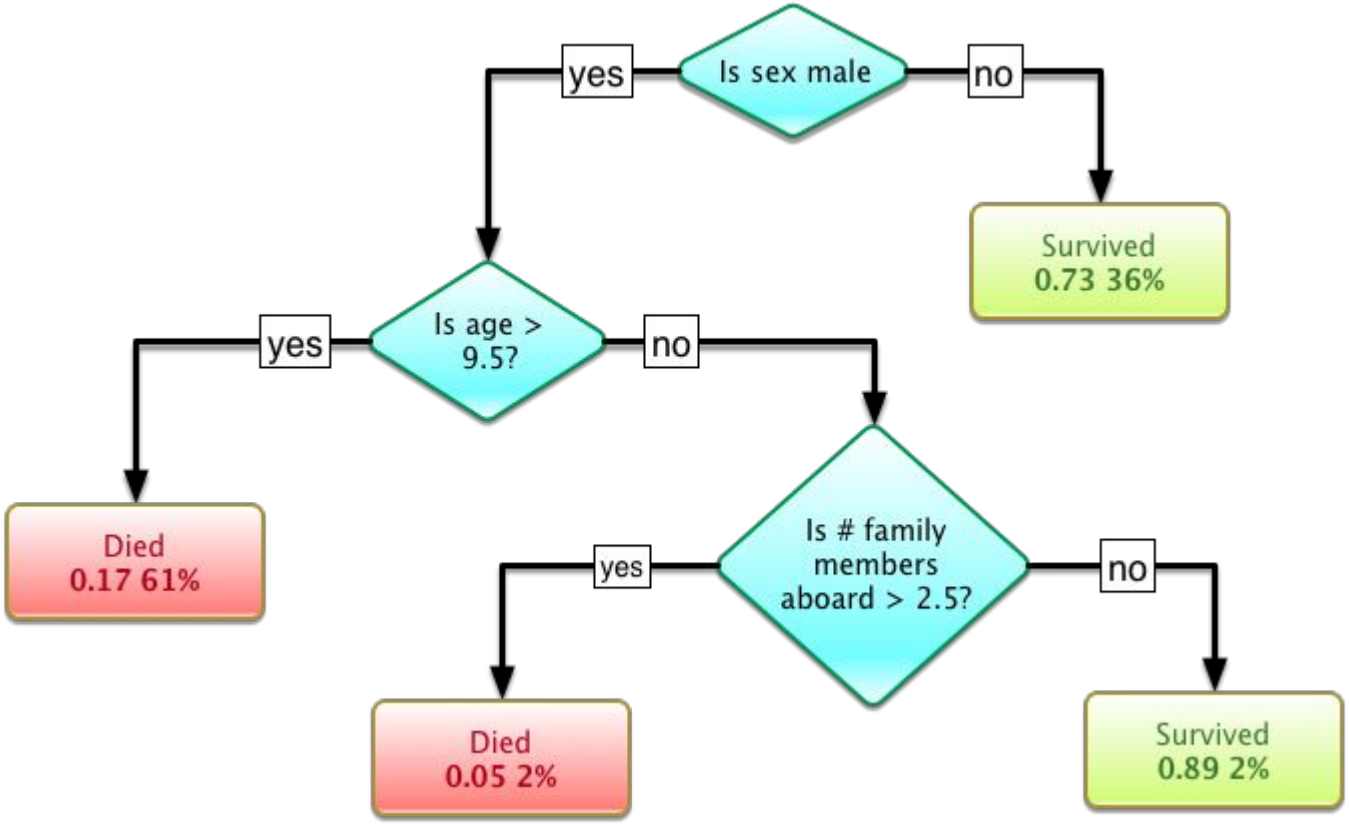
Data Models: Problems

- Multiplicity
  - Multiple models can fit the same data
- Predictive Accuracy
  - “Noisy” variables are unmeasured
    - Overfitting can be resolved by cross-validation
- Complex systems
  - Data are not multivariate normal
  - Simple parametric models are not enough

Algorithmic Modeling: Introduction

- Philosophy
  - “Nature produces data in a black box whose insides are complex, mysterious, and at least, partly unknowable.”
- Assumption
  - Data is drawn iid from a multivariate distribution
- Goal
  - Find an algorithm **f(x)** that predicts **y**
    - **x**: predictor data
    - **y**: response data
  - Includes but is not limited to linear models

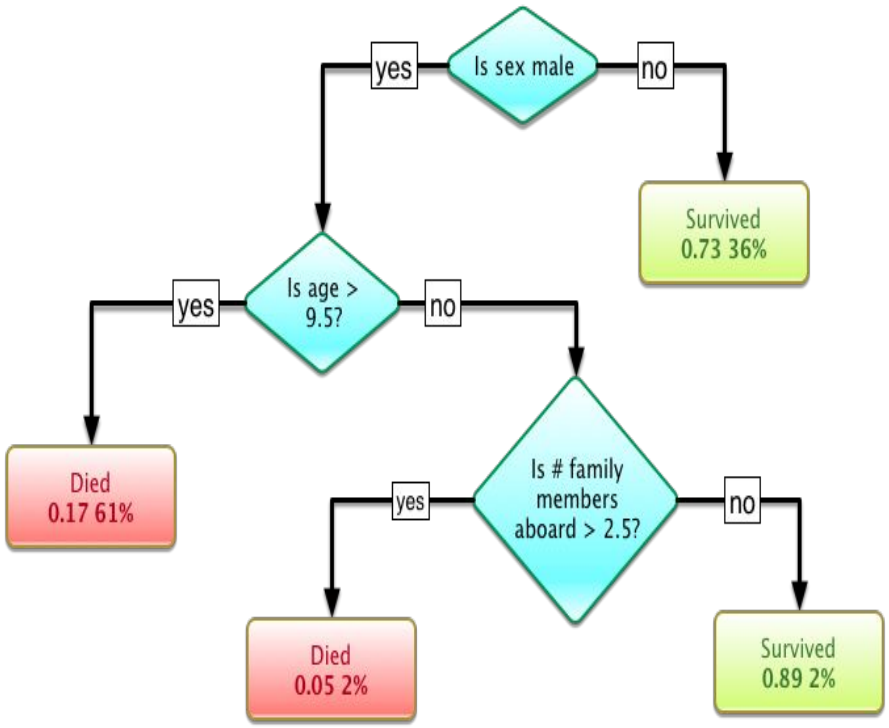
Algorithm Example: Decision Tree



Survival of passengers on the Titanic (Source: Wikipedia)

Algorithm Example: Decision Tree

1. Randomly select a sequence of predictors as nodes
2. Determine the decision at the end of each path by using the training data
3. Find the tree's prediction error rate when applied to the testing data
4. Repeat Step 1-3 for many times and select the tree with the lowest prediction error rate



Data Modeling

- Pros
  - **Simple** to interpret
- Cons
  - Strong assumptions about models
  - Model assumptions are not falsifiable
  - Ineffective for handling high-dim data
  - Many models can fit the same data equally well

Algorithmic Modeling

- Pros
  - **Accurate** in making predictions
  - More relaxed assumptions about models
  - Takes advantage of high-dim data
- Cons
  - Algorithms are difficult to interpret
  - Many models can fit the same data equally well

Breiman's Argument for Algorithmic Modeling

- "Simplicity vs Accuracy" is an incorrect way to understand the goal of statistical analysis.
- The purpose of a model is to retrieve useful information about the relation between the response and the predictor.
- Interpretability is only *one way* of getting information
- A model does not have to be simple to be reliable
- There is no reason to restrict our options to a certain subset of models

Comment & Rejoinder: Cox

- **Starting point:** **Data** vs **Issue**
  - It is the question/hypothesis that is unknown
  - Hard to understand data without probabilistic modeling
- **Objective:** **Prediction** vs **Understanding**
  - Stability of the predictor comes from clarifying mechanisms
  - Prediction for different conditions:
    - Cannot rely solely on data set available
  - Other objectives need emphasis on understanding

Comment & Rejoinder: Cox

- **Starting point: Data** vs **Issue**
  - It is the question/hypothesis that is unknown
  - Hard to understand data without probabilistic modeling
- **Objective: Prediction** vs **Understanding**
  - Stability of the predictor comes from clarifying mechanisms
  - Prediction for different conditions: ← **Yes, with insufficient data, data models can be useful**
    - Cannot rely solely on data set available
  - Other objectives need emphasis on understanding

Many successful algorithmic applications extract useful information from data

Comment & Rejoinder: Cox

- Ideal method for statistical analysis
  - Descriptively appealing
  - Transparent
  - Firm model base
  - **NOT** a mechanical process

picture source: xkcd.com

Comment & Rejoinder: Cox

- Ideal method for statistical analysis
  - Descriptively appealing
  - Transparent
  - Firm model base
  - **NOT** a mechanical process

Cannot rely on data models alone, due to vast extensions & changes in problems

Comment & Rejoinder: Cox

Form the right question

Subject-matter → a form for interpretation

"Provide the many people working in applications outside of academia with

useful, reliable, and accurate

analysis tools ..."

Comment & Rejoinder: Efron

- **Changes** in new era
  - More noise in data
  - Less distinct goals
  - New & complex methods w/ little supporting theory
- **Role of Prediction**
  - Prediction is not sufficient in many settings
  - Can lead to potential inferential innovations
- "The whole point of science is to **open up black boxes**, understand their insides, and build better boxes for the purpose of mankind"

Comment & Rejoinder: Efron

- **Changes** in new era
  - More noise in data
  - Less distinct goals
  - New & complex methods w/ little supporting theory

Need for answering questions with complex and accurate models

- **Role of Prediction**
  - Prediction is not sufficient in many settings
  - Can lead to potential inferential innovations

Can be useful in model assessment

- "The whole point of science is to **open up black boxes**, understand their insides, and build better boxes for the purpose of mankind"

Be pragmatic

Comment & Rejoinder: Hoadley

- Breiman's conclusions are **consistent** with practice of statistics in **business**
  - Credit scoring
    - Input variables **x**: monthly bills and payments over the last 12 months (24-dimensional)
    - Output variable **y**: indicator of no severe delinquency over the next 6 months (binary)
    - **Goal**: estimate **f(x)=log(Pr{ y=1|x } / Pr{ y=0|x })**

Data modeling culture → ~~**Simple logistic regression**~~

Algorithmic modeling culture → **Fair, Isaac (scorecard)**

Comment & Rejoinder: Hoadley

- Performance on the test sample
  - Breiman emphasizes it; however, it can be **overdone**
    - A random sample of current population
    - High performance $\not\Rightarrow$ High performance on future samples
    - **Things do change** → Protect against change:
      Monitor the performance of models over time
      Sufficient degradation of performance → new model
    - Example: Fair, Isaac -- redevelopment cycle is about 18-24 months -- make the models more robust over time
- Rejoinder by Breiman
  - **Agree** -- models must be modified to stay accurate
  - **NOT necessary** to alter the way of model construction

Comment & Rejoinder: Hoadley

- It is possible to have both accuracy and interpretability
  - Segmented palatable scorecards -- interpretable by the customer and very accurate
- Rejoinder by Breiman
  - In either stock prediction or credit scoring, the **priority is accuracy**
  - Interpretability is a secondary goal that can be finessed

Comment & Rejoinder: Hoadley

- **Challenges** for the algorithmic modeling approach
  - Tuning dilemma
    - How to set the tuning parameters → optimize the results
  - Measuring importance -- Is it really possible?
    - A variable and its relationships will change
    - “Ping-Pong theorem”

Comment & Rejoinder: Hoadley

- Do algorithmic modeling with data modeling tools
  - Ignore most textbook advice
  - Embrace the blessing of dimensionality
  - Use constraints in the fitting optimizations
  - Use regularization
  - Validate the results

Comment & Rejoinder: Parzen

- Open to the issue raised in Leo Breiman's paper
  - Statistician, AVOID doing harm
  - Two goals in analyzing the data
    - **Management** (prediction) seeks **profit** -- practical answers → decision making **in the short run**
    - **Science** (information) seeks **truth** -- fundamental knowledge about nature → understanding and control **in the long run**
  - Methods of algorithmic modeling are important contributions to the tool kit of Statisticians
    - **Better** predictive accuracy, **better** information about the underlying mechanism, **perfect** separation and discrimination between two classes...

Comment & Rejoinder: Parzen

- Hypotheses to test to avoid blunders of statistical modeling (generic deviations from standard assumptions)
  - Bivariate dependence (correlation)
    - Between independent (input) variables
  - Two-sample conditional clustering
    - Arises in the distributions of independent (input) variables to discriminate between two classes
    - (conditional) P(class 1 | X) = P(class 1) (pooled)

Comment & Rejoinder: Parzen

- NOT **two**, BUT **many** modeling cultures
  - Robust methods, Bayesian methods...
  - Eclectic philosophy of statistical modeling
    - **Statistical methods mining** seeks to provide a framework to synthesize and apply the past methodological progress in computationally intensive methods for statistical modeling
  - "Data mining" is a special case of "data modeling"
    - Statistical data modeling done in a systematic way -- SIEVE/PPDAC
  - Rejoinder by Breiman
    - Not an issue he wants to fiercely contest
    - Pretty clear cut -- are you modeling the inside of the box or not?

Comment & Rejoinder: Parzen

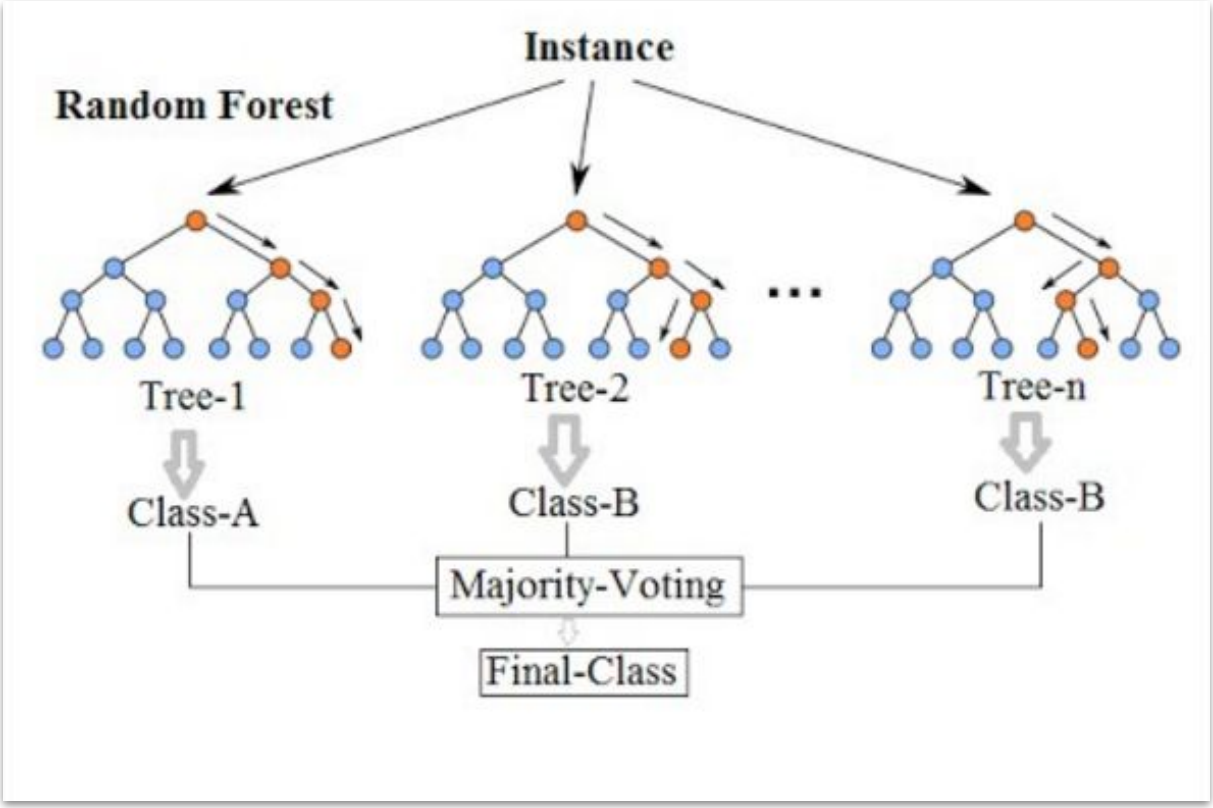
- Quantile culture
- Quantile ideas for HIGH dimensional data analysis

Questions & Discussion

Questions for discussion

- Which type of models is more appealing to you, data models or algorithmic models?
- If you can only choose one, are you more interested in **1)** understanding the mechanism behind natural processes, OR **2)** being able to predict and manipulate natural processes?
- If a skeptic reads your report for the first project in Biostat 699 and asks you “How do you know the model you use is true”, how would you respond?
- What are the possible future paths for both cultures?

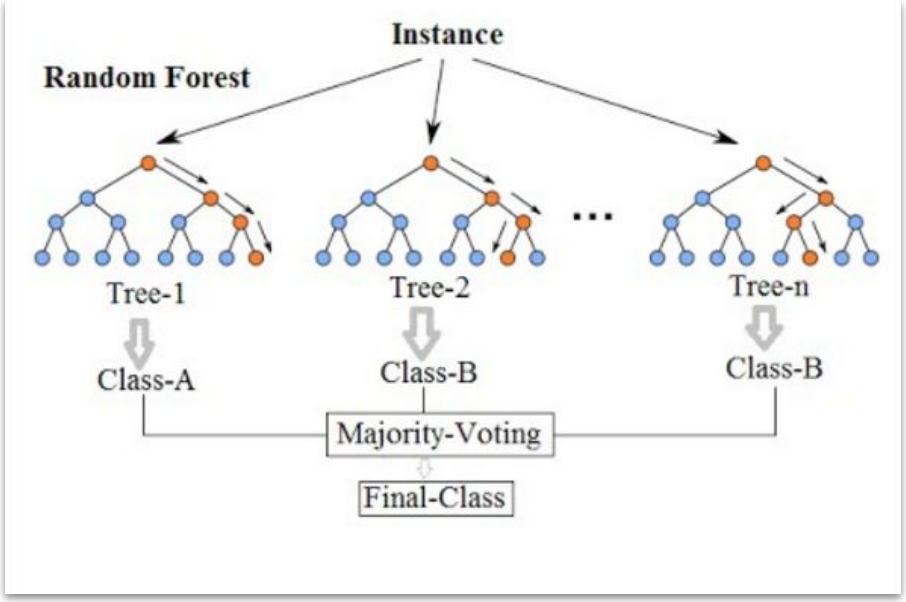
Algorithm Example: Random Forest



Source: TIBCO

Algorithm Example: Random Forest

1. Randomly select a subset of the training data
2. Find the best decision tree for this training subset
3. Repeat Step 1-2 for many times and obtain many trees
4. Everytime a new input comes, path it down to every tree and combine the results by majority vote



Thoughts (maybe move to discussion?)

Questions for discussion

- Different purposes (e.g. Netflix recommendation vs. medical study) may be the underlying reason for adopting different culture
- What does statistics mean (do)?
  - “Mainstream serious statistics” ?- Cox
  - Build methodology foundation?
  - Applied analytics?
  - Embrace computation & data mining?