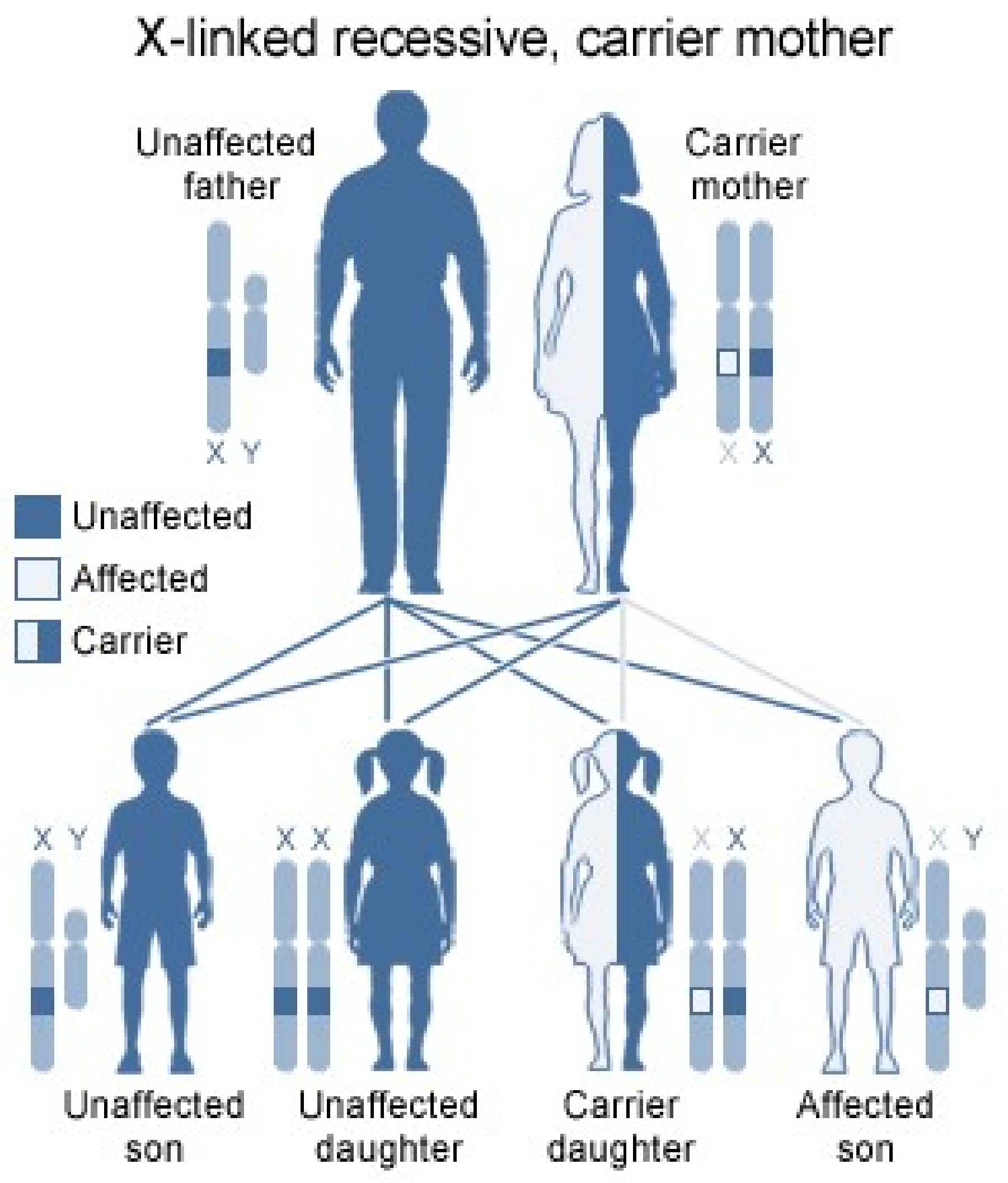
Predicting Chronic Diseases with Machine Learning

Vlad Korolev, Anupam Joshi, Yelena Yesha, Michael Grasso

Focus

1. Use machine learning techniques aid in determining predisposition to chronic decease based on individual's genetic information and clinical data.
2. Use all available data from SNP profile
3. Ensure reproduciblity of experiments and keep track of data provenance

Genetic Causes of Chronic Diseases



U.S. National Library of Medicine

Single Gene Disorders

- Depend on a single-gene mutation
- Have been suspected for long time
- Have been proven for quite some time
- Notable examples
  - Sickle cell anemia
  - Cystic fibrosis
  - Hemophilia
- Easily determined by Mendelian methods, looking at family history

Multi Gene Disorders

- Depend on two or more mutations
- Well studied mutation : Horse color
- Two gene conditions
  - Lactose intolerance
- Polygenic complex mutations
  - Asthma
  - Diabetes
  - Cancers
  - Hypertension
  - Autoimmune diseases such as multiple sclerosis
- Very hard to determine through Mendelian methods
- Suspected to be genetic based on tendencies to run in families
- No clear pattern of inheritance

Previous Work

1. de Miguel-Yanes JM, Shrader P, Pencina MJ, Fox CS, Manning AK, Grant RW. Genetic risk reclassification for type 2 diabetes by age below or above 50 years using 40 type 2 diabetes risk single nucleotide polymorphisms. Diabetes Care. 2011 Jan;34(1):121-5.
2. Lanktree M, Oh J, Hegele RA. Genetic testing for atherosclerosis risk: inevitability or pipe dream? Can J Cardiol. 2008 Nov;24(11):851-4.

- Combine clinical and genetic information
- Used statistical models
- Did not show benefit when including genetic information

Darshana Dalvi, Aniket Bochare

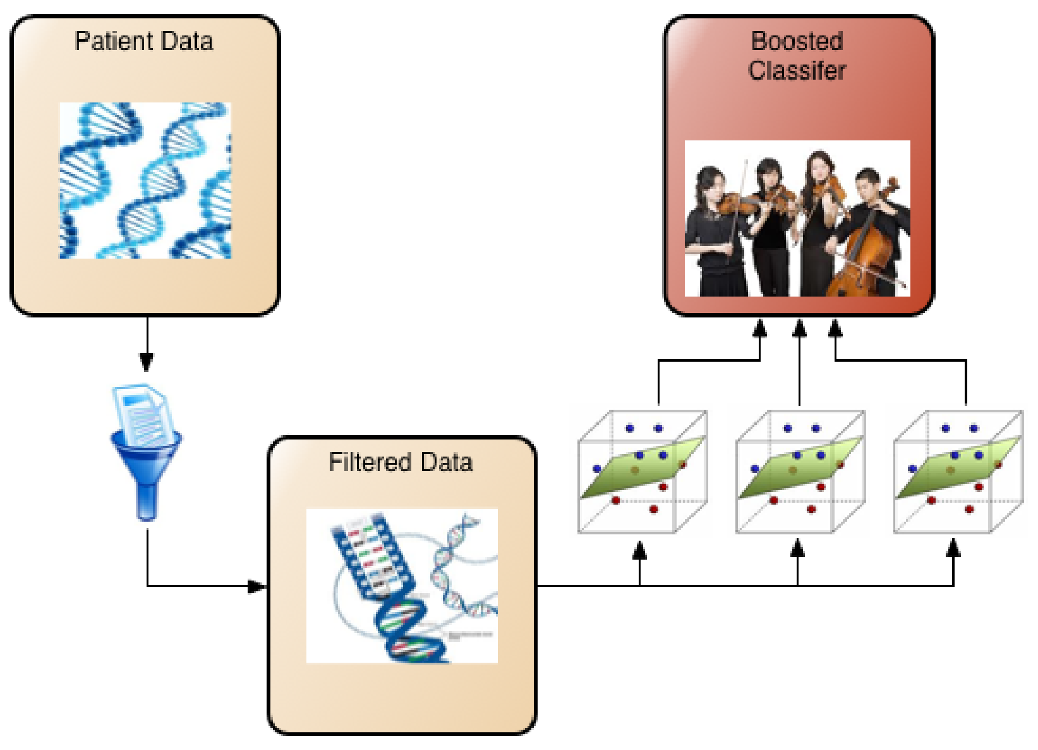
- Extracted subset of SNPs that are known to cause the disease
- Combined SNPs with clinical data
- Trained decision tree algorithm to build a classifier
- Cross-validated the classifier to obtain accuracy of the method
- Showed slight improvement over pure statistical methods. But amount of improvement was not that great.

Approach

Challenges

- Large Datasets (500 GB)
- Too many attributes
- Repeatability of experiments

Method



Patient Data

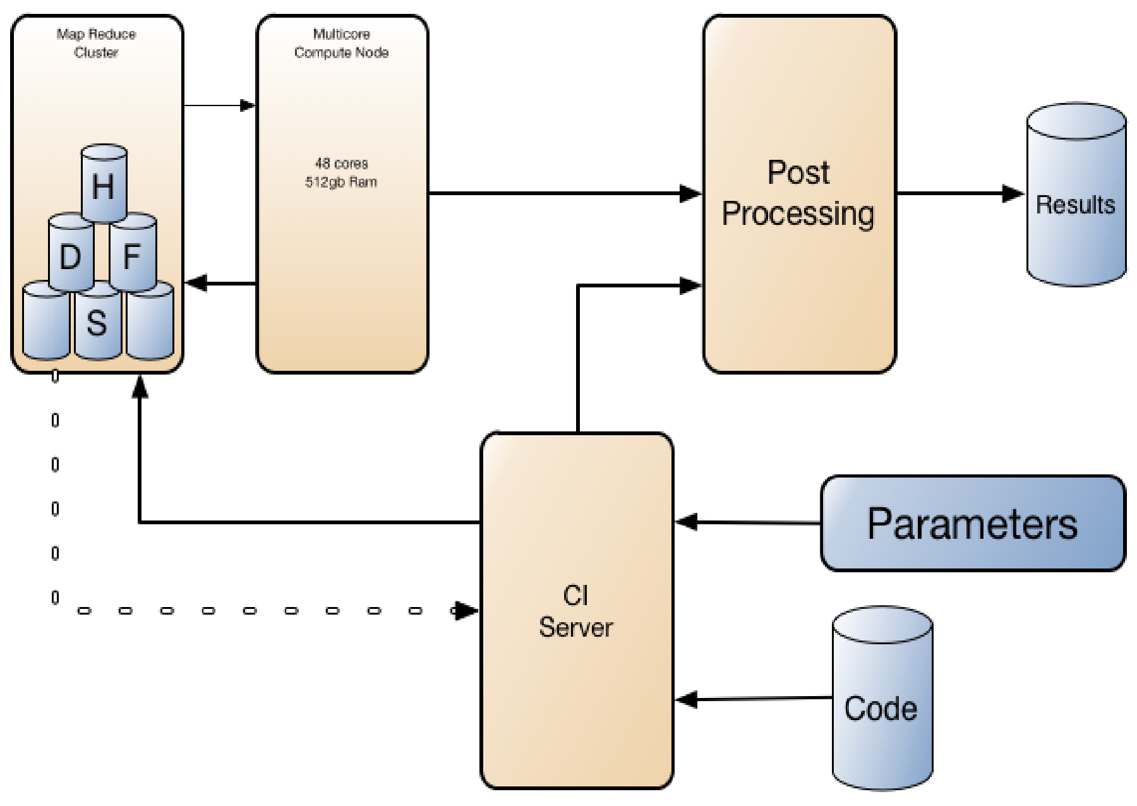
Boosted Classifier

Filtered Data

Objectives

Capacity Performance
Repeatability Automation

Platform



Map Reduce Cluster

H D F S

Multicore Compute Node

48 cores 512gb Ram

Post Processing

Results

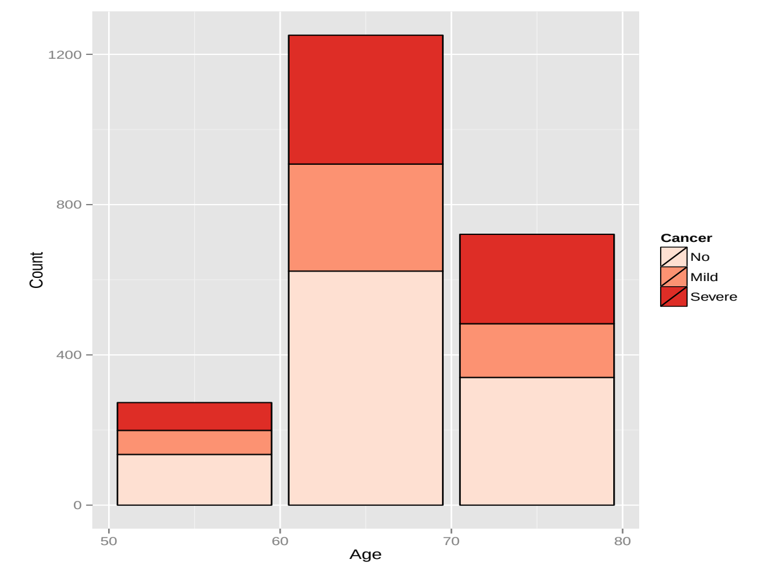
CI Server

Parameters

Code

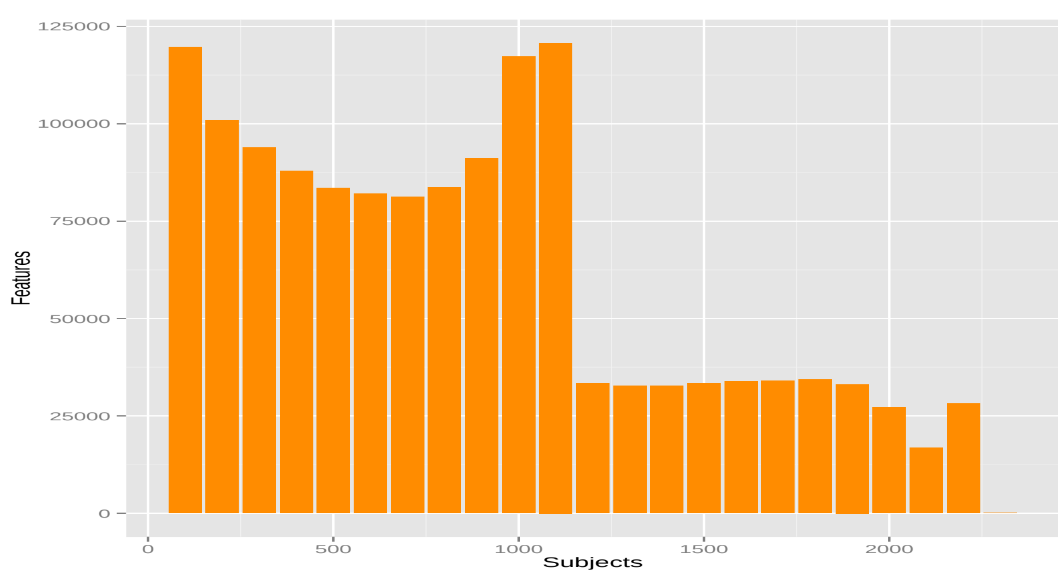
Initial Results

Prostate Cancer Study



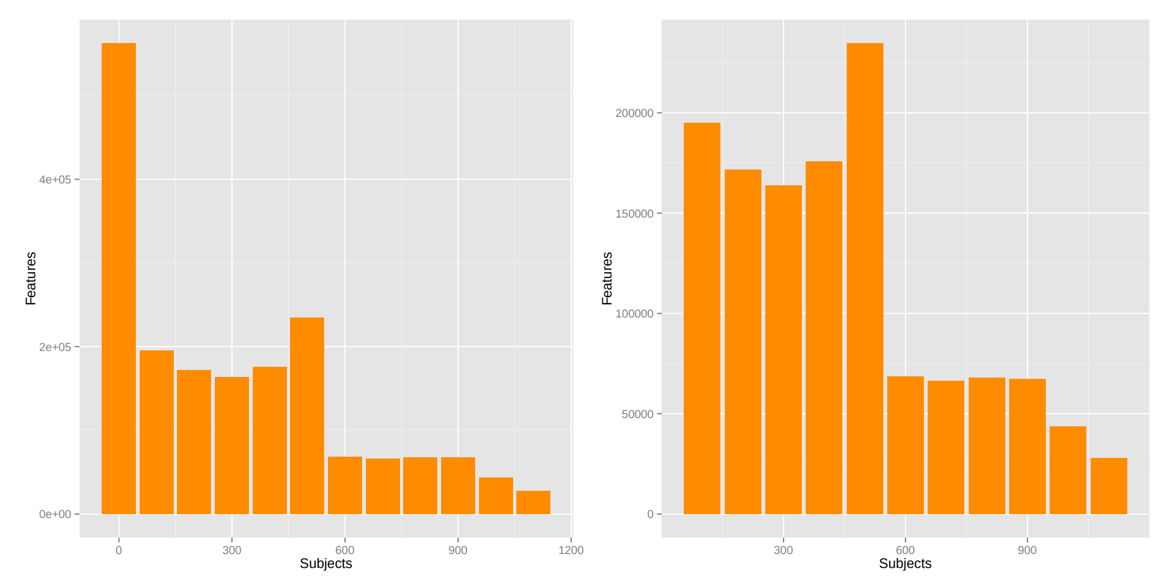
No Cancer	1098
Cancer	1147

Combined Feature Frequency



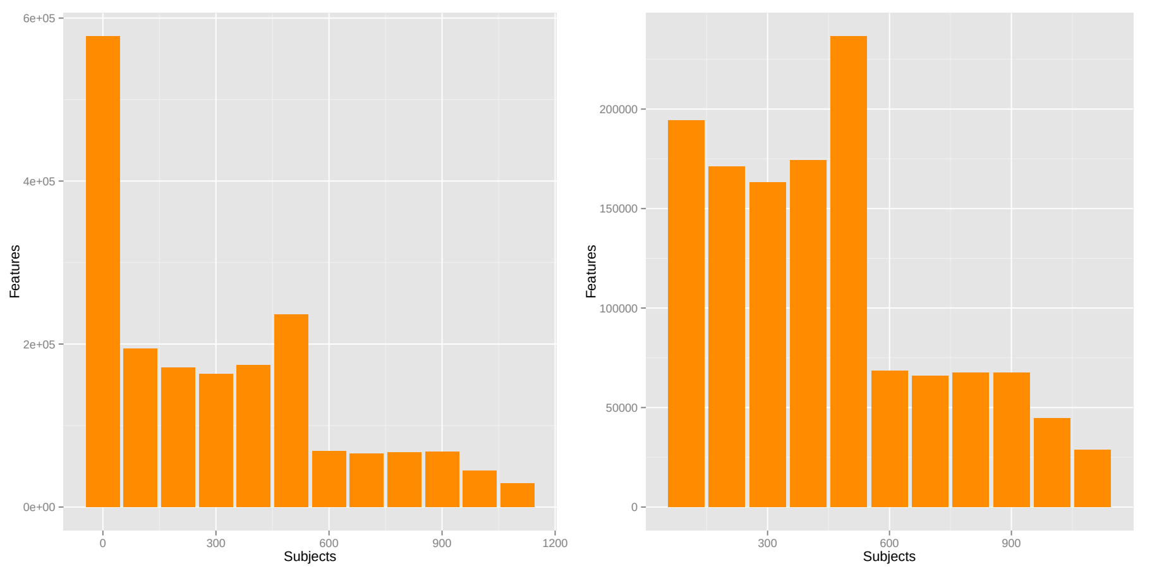
Freq	Count
0	516673
100	119780
200	100841
300	93909
400	87831
500	83527
600	82016
700	81296
800	83610
900	91181
1000	117347
1100	120769
1200	33409
1300	32734
1400	32756
1500	33398
1600	33787
1700	34016
1800	34374
1900	33121
2000	27135
2100	16808
2200	28155
2300	65
Total	1918538

Feature Frequency : Controls



Freq	Count
0	562257
100	194963
200	171566
300	163795
400	175656
500	234558
600	68646
700	66275
800	67841
900	67354
1000	43590
1100	27930

Feature Frequency : Cases



Freq	Count
0	577407
100	194297
200	171074
300	163058
400	174393
500	236466
600	68488
700	65814
800	67487
900	67649
1000	44748
1100	28778
Total	1859659

Work in progress

University of Maryland Baltimore Coounty

University of Marlyand School of Medicine