

# RISK PREDICTION AND DIAGNOSTICS OF CARDIOVASCULAR DISEASES

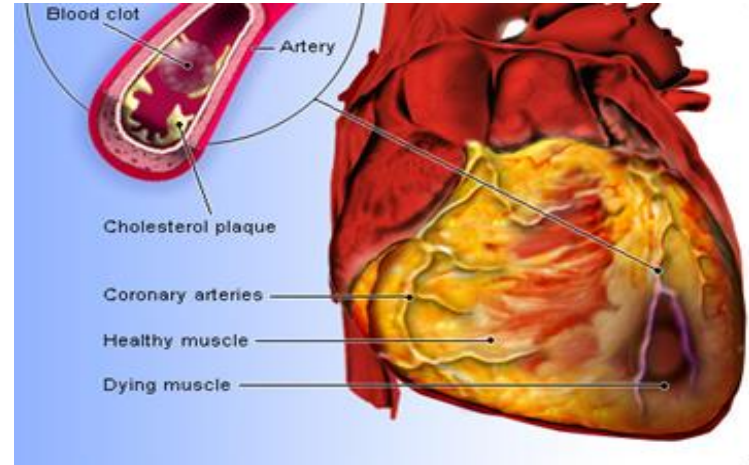
IGOR PUTRENKO

INSIGHT HEALTH DATA SCIENCE PROGRAM

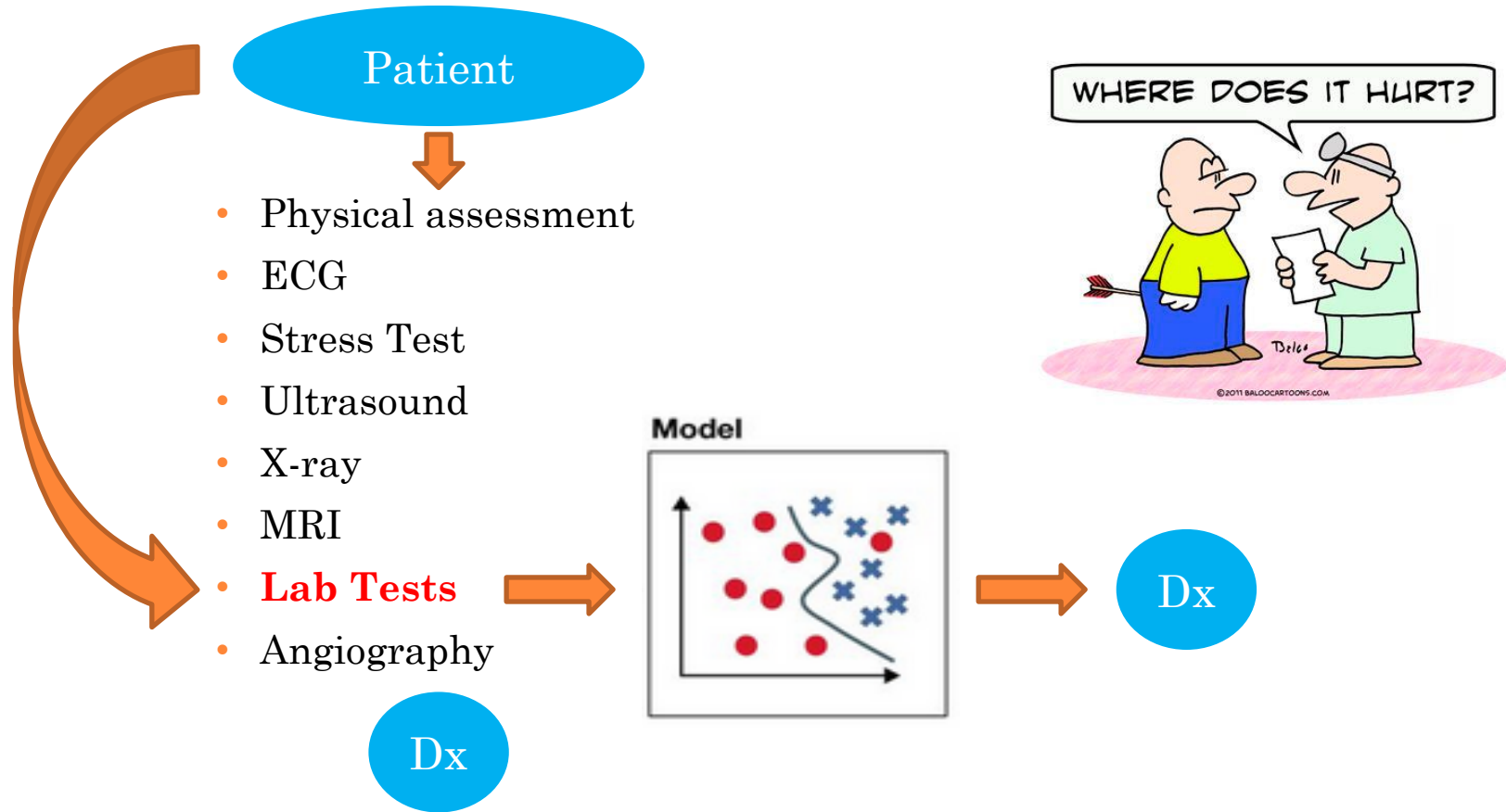
SAN FRANCISCO, CA  
OCTOBER 2017

# CARDIOVASCULAR DISEASES (CVD)

- Annual direct medical expenditures (by 2030) – **\$818 billion**
- **One in three** U.S. adults has one or more types of CVD
- Optimized diagnostic process can substantially decrease the costs



# CAN WE PREDICT CVD TYPE BASED ON LAB TESTS?

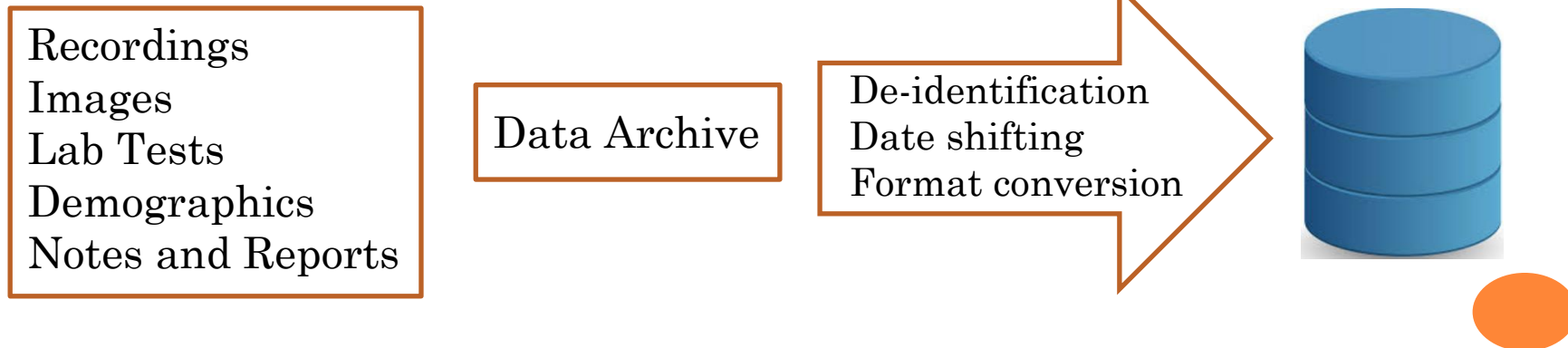


# DATA SOURCE



Beth Israel Deaconess  
Medical Center

- MIMIC-III, critical care database
- 38,597 patients
- Median age is 65.8 years (52.8–77.8)



# DATA PIPELINE

MIMIC-III  
csv files  
28 M records



SQL Server  
database



Dataset



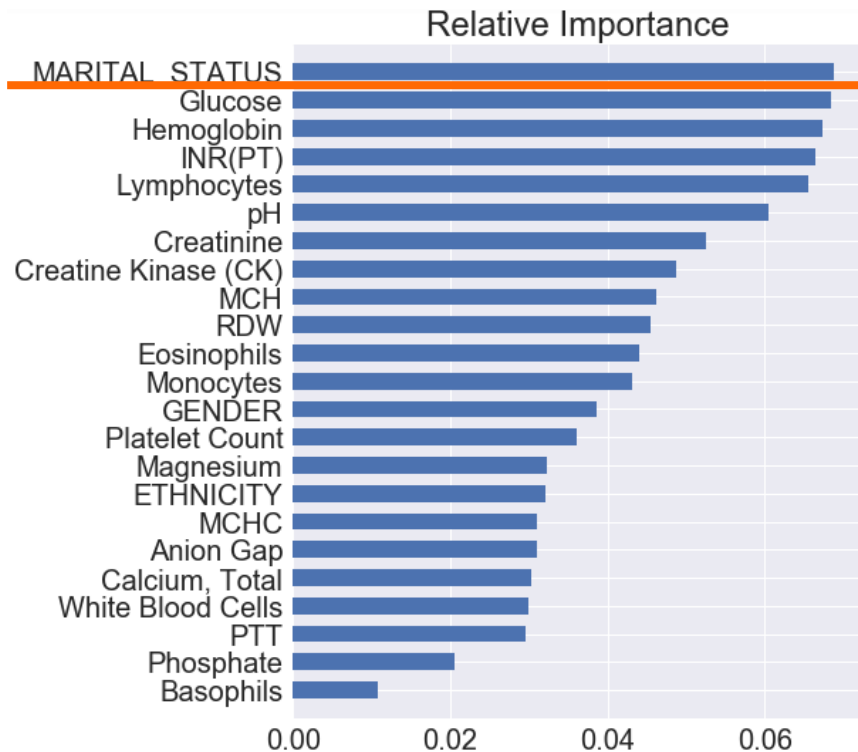
# PREDICTIVE MODELING

- 6 CVD categories
- Balanced classes
- Cross-validation (5-fold)
- Grid search
- Gradient Boosting
- Random Forest
- Artificial Neural Network

**Accuracy: 39% (vs 16.7% chance)**



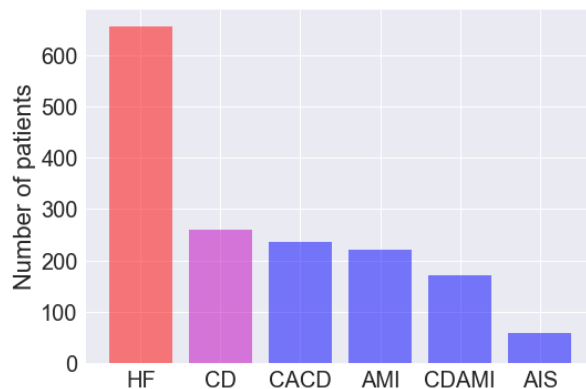
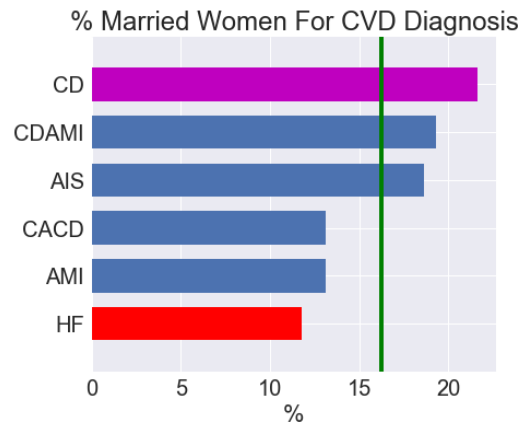
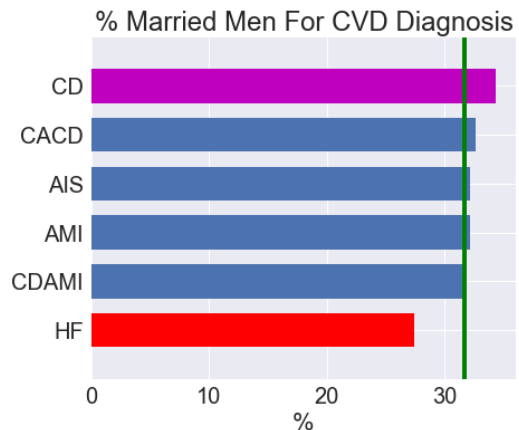
# FEATURE IMPORTANCE



Identified features are indicators of:

- psychological well-being
- status of immune system
- chronic diseases (kidneys, liver, heart)
- chronic infection
- genetic factors

# SHOULD YOU GET MARRIED?





# CONCLUSION

The developed model can improve CVD diagnostic process:

- Risk assessment
- Early detection
- Decision support
- Cost reduction
- Faster diagnostics



Clients: clinicians, health agencies, insurance companies



# ABOUT ME



- Ph.D. in Biochemistry
- Academia: ion channel electrophysiology, brain & pain research
- Biotech industry: pain drug discovery
- IT: business / systems analysis, infrastructure, network administration
- Concentrate on Data Science



# BACKUP SLIDES



# CVD CATEGORIES

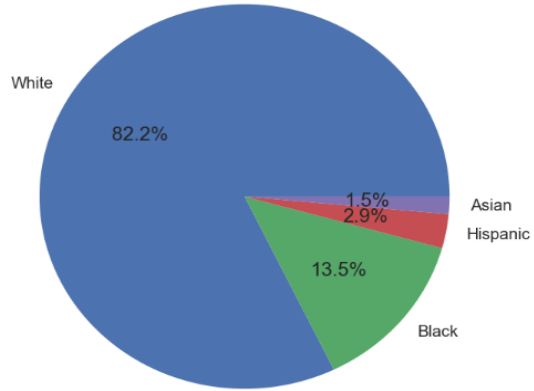
Patients selected with the following Diagnosis Related Groups (DRGCODES) used by the hospital for billing purposes

- ACUTE ISCHEMIC STROKE
- ACUTE MYOCARDIAL INFARCTION (AMI)
- CARDIAC ARRHYTHMIA & CONDUCTION DISORDERS
- HEART FAILURE
- CIRCULATORY DISORDERS EXCEPT AMI
- CIRCULATORY DISORDERS WITH AMI

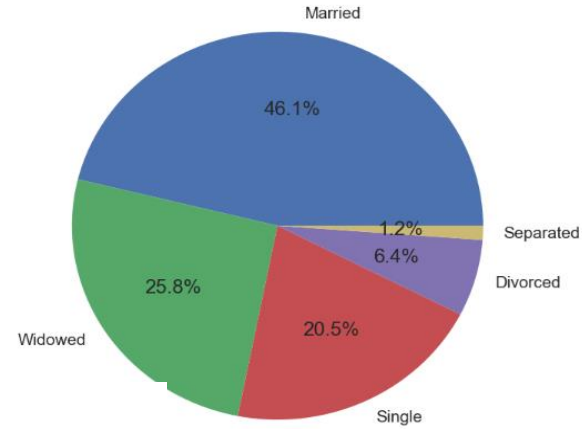


# DEMOGRAPHIC DATA

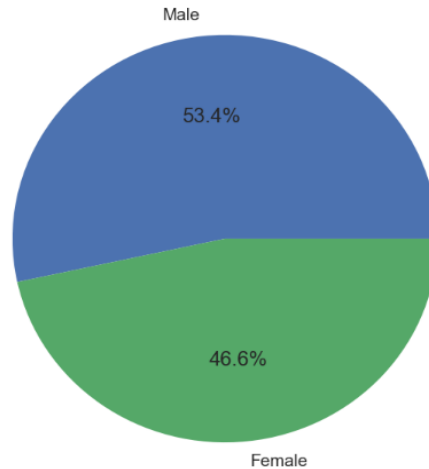
RACE / ETHNICITY



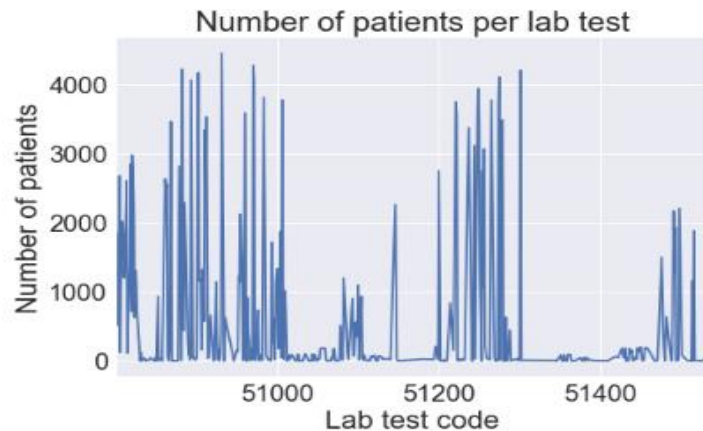
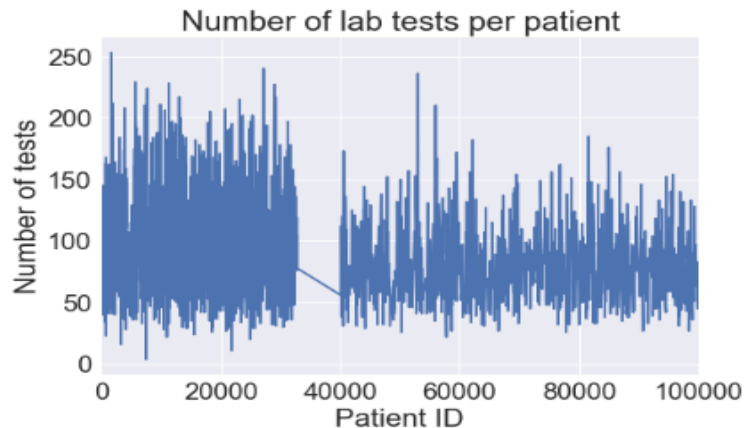
Marital Status



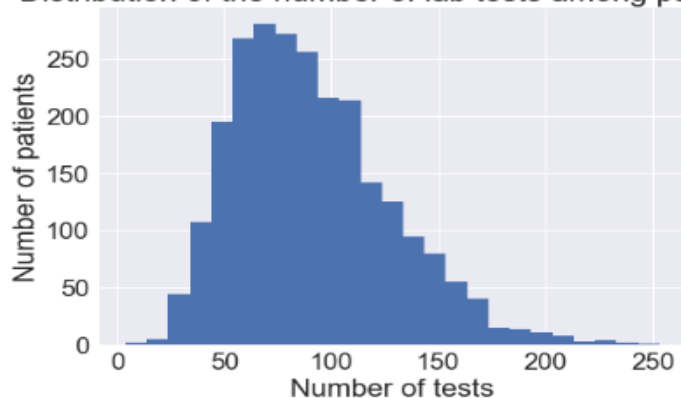
Gender



# DIMENSION REDUCTION – SELECTING LAB TESTS



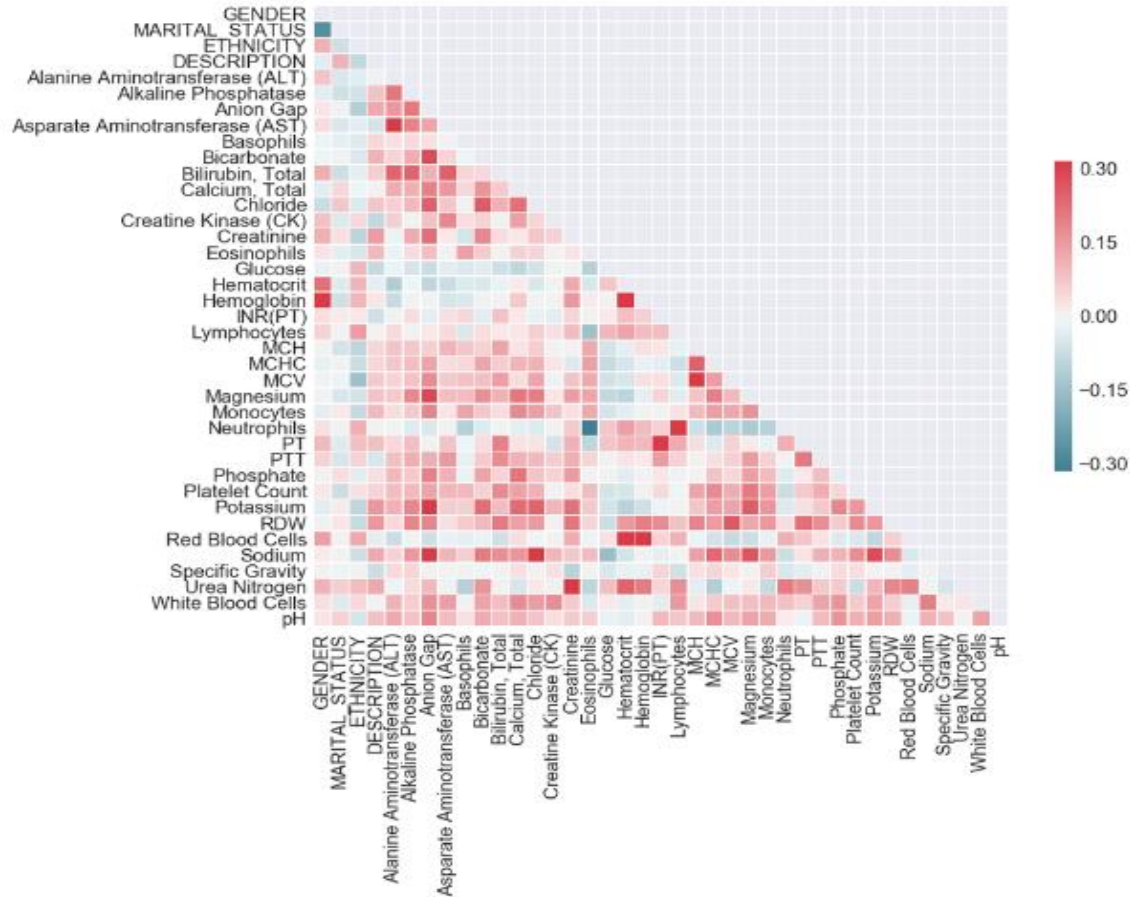
Distribution of the number of lab tests among patients



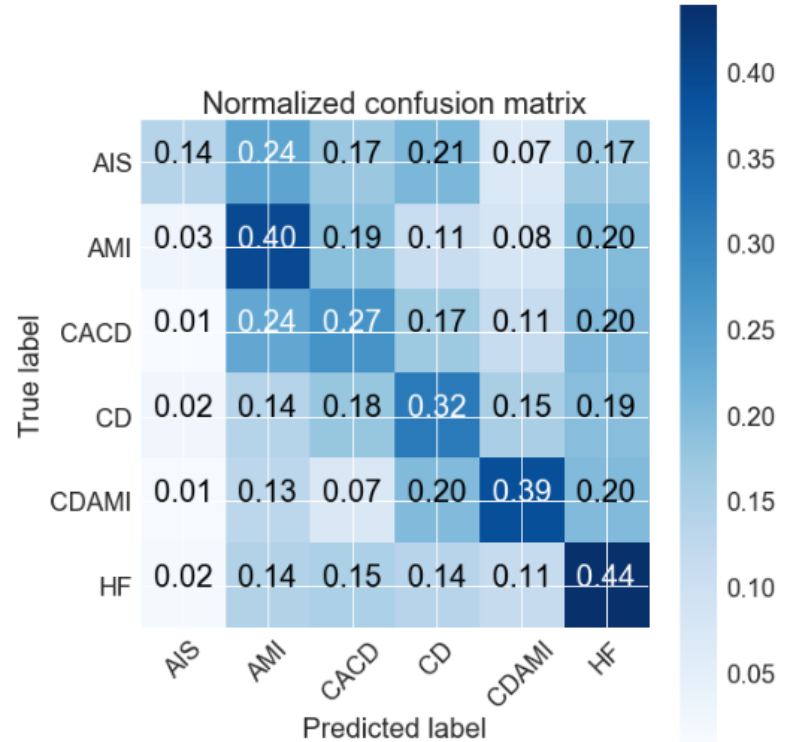
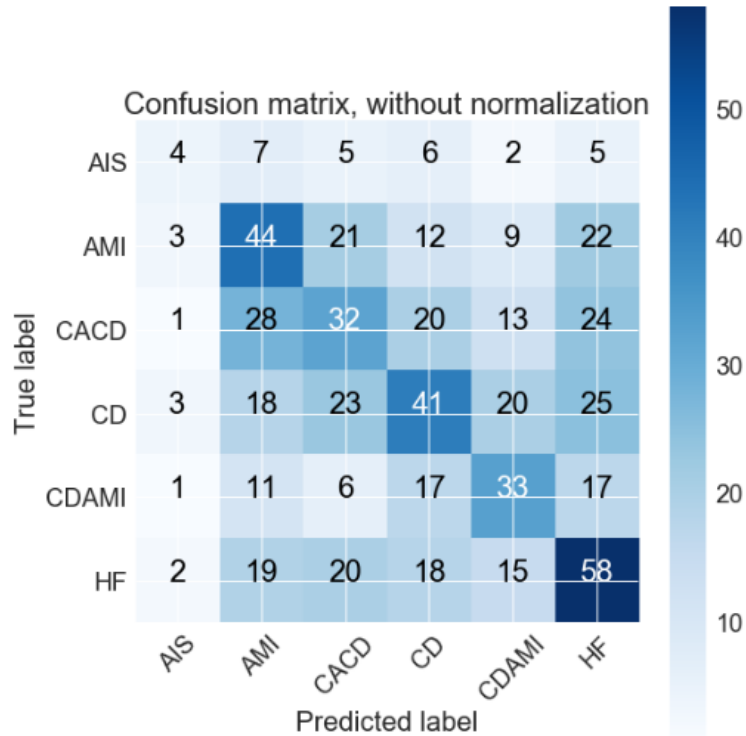
Total number of CVD patients	2455
Total number of lab tests	403
Feature selection – missing values threshold (patients/test)	2155
Lab tests (features) passed	35
Final number of selected patients	1209



# IDENTIFYING MULTICOLLINEARITY



# CONFUSION MATRIX





# PRECISION, RECALL, AND F1-SCORE

	precision	recall	f1-score	support
0	0.29	0.14	0.19	29
1	0.35	0.40	0.37	111
2	0.30	0.27	0.28	118
3	0.36	0.32	0.34	130
4	0.36	0.39	0.37	85
5	0.38	0.44	0.41	132
avg / total	0.35	0.35	0.35	605

