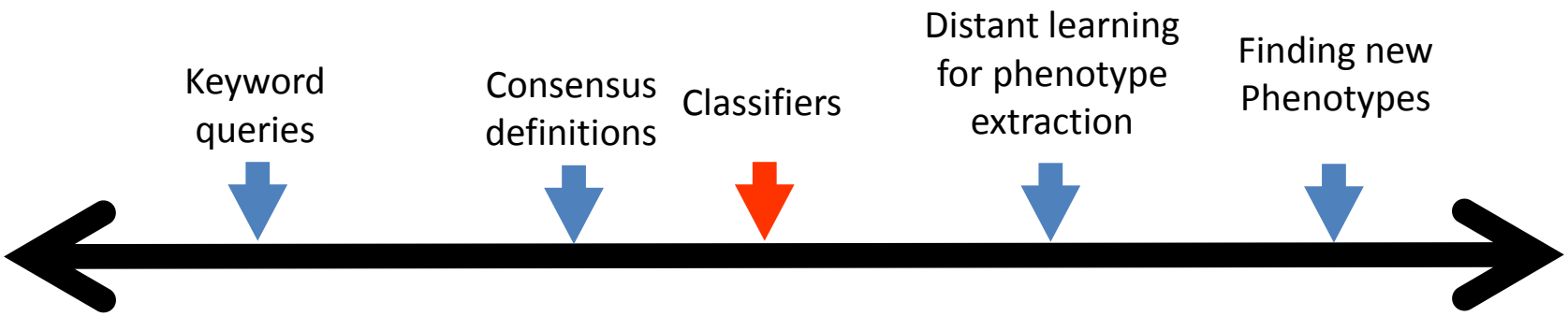
Phenotyping WG update

Nigam Shah

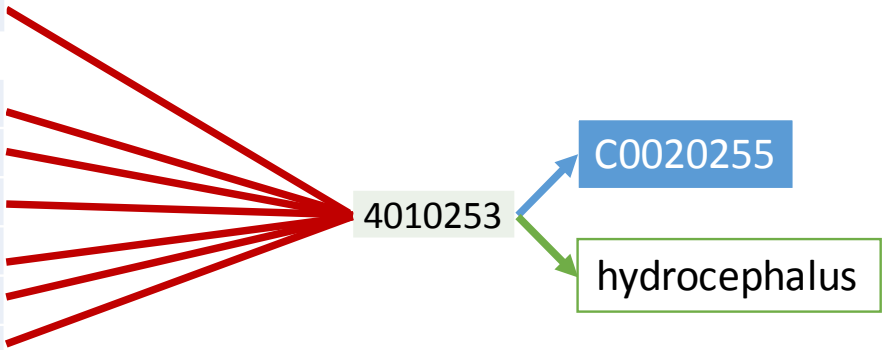
STANFORD
SCHOOL OF MEDICINE

Electronic Phenotyping



Keyword queries for “noisy labeling”

2933	hydrocephalus
90773	water on the brain
3107785	hydrocephalus [disease/finding]
42612	hydrocephaly
2889104	hydrocephalus (disorder)
1936514	hydrocephalus, unspecified
2001177	hydrocephalus nos

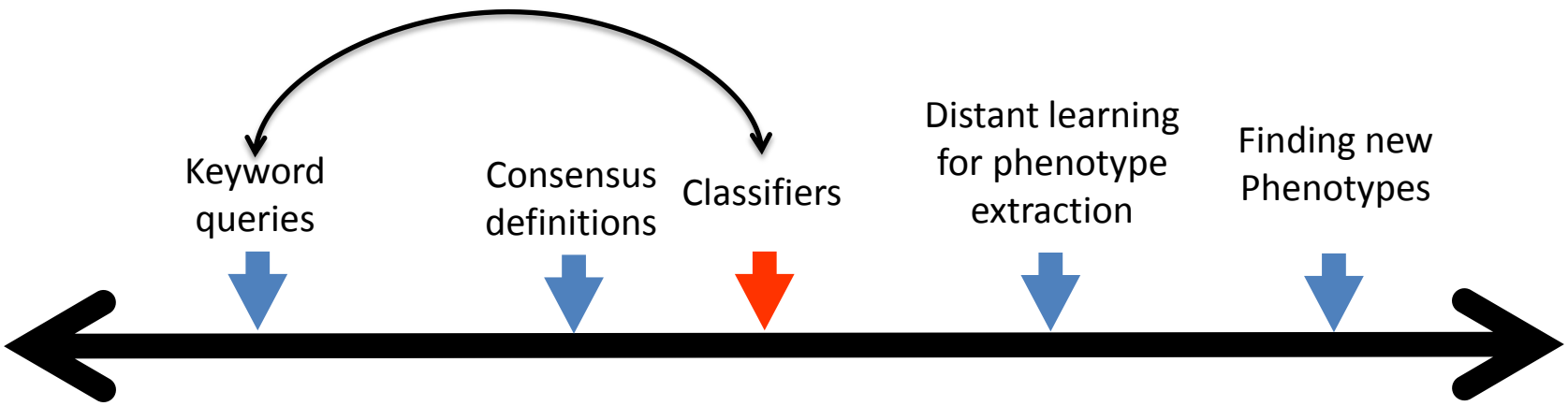


tid	cui	str	Note freq	syn	Medline freq	% noun
2933	C0020255	hydrocephalus	29,634	NNS	19,541	64.61
42612	C0020255	hydrocephaly	113	NN	275	49.81
90773	C0020255	water on the brain	8	ROOT	1	50

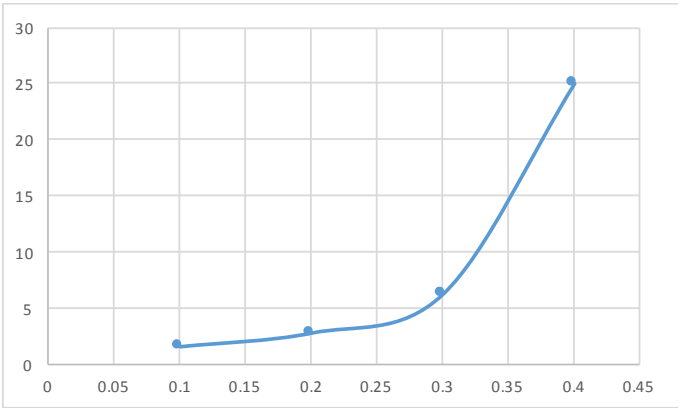
Assumption: “long mention” is a reliable indicator of presence

Electronic Phenotyping

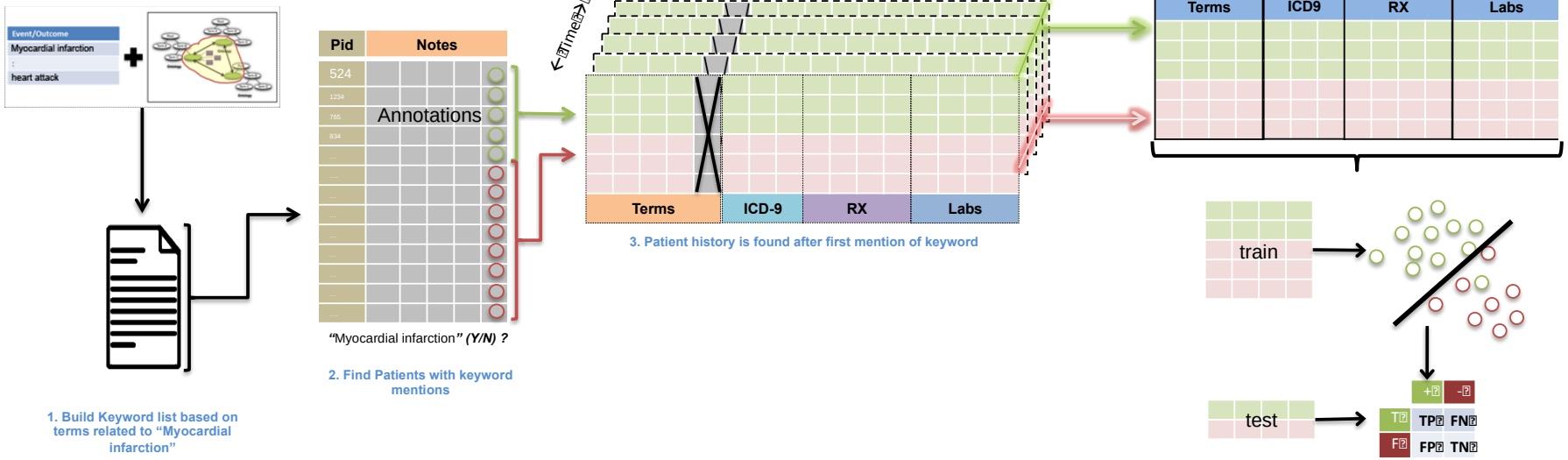
$$(1-2t)^2$$



Error rate in labeling	Sample size
10 %	1.56 x
20 %	2.77 x
30 %	6.25 x
40 %	25 x



XPRESS- EXtraction of Phenotypes from clinical Records using Silver Standards



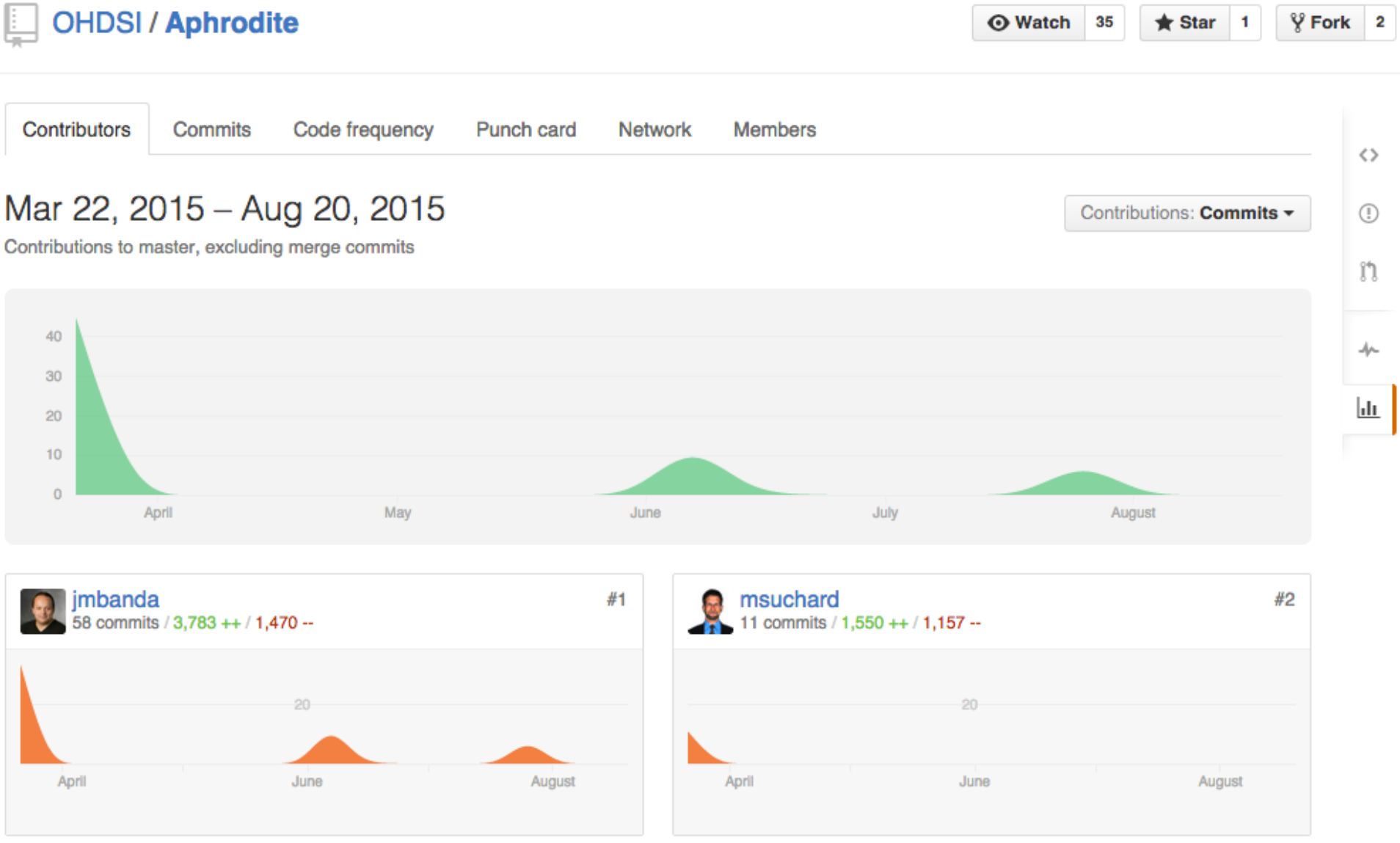
Input: getPatients.R -- config.R, keywords.tsv, ignore.tsv
Output: feature_vectors.Rda

Input: config.R – with term search settings
Output: keywords.tsv and ignore.tsv

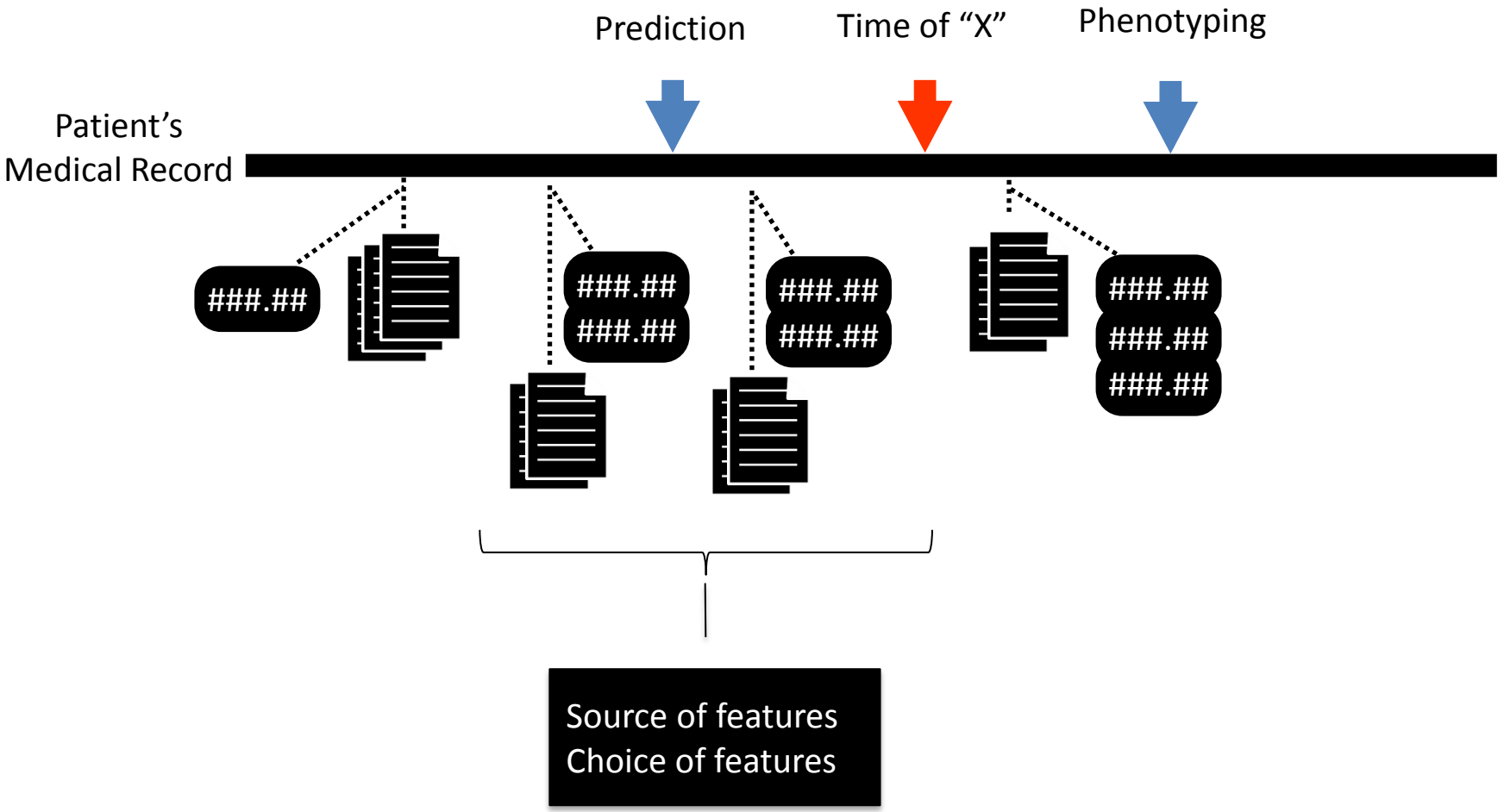
Input: buildModel.R -- config.R, feature_vectors.Rda
Output: model.Rda

APHRODITE

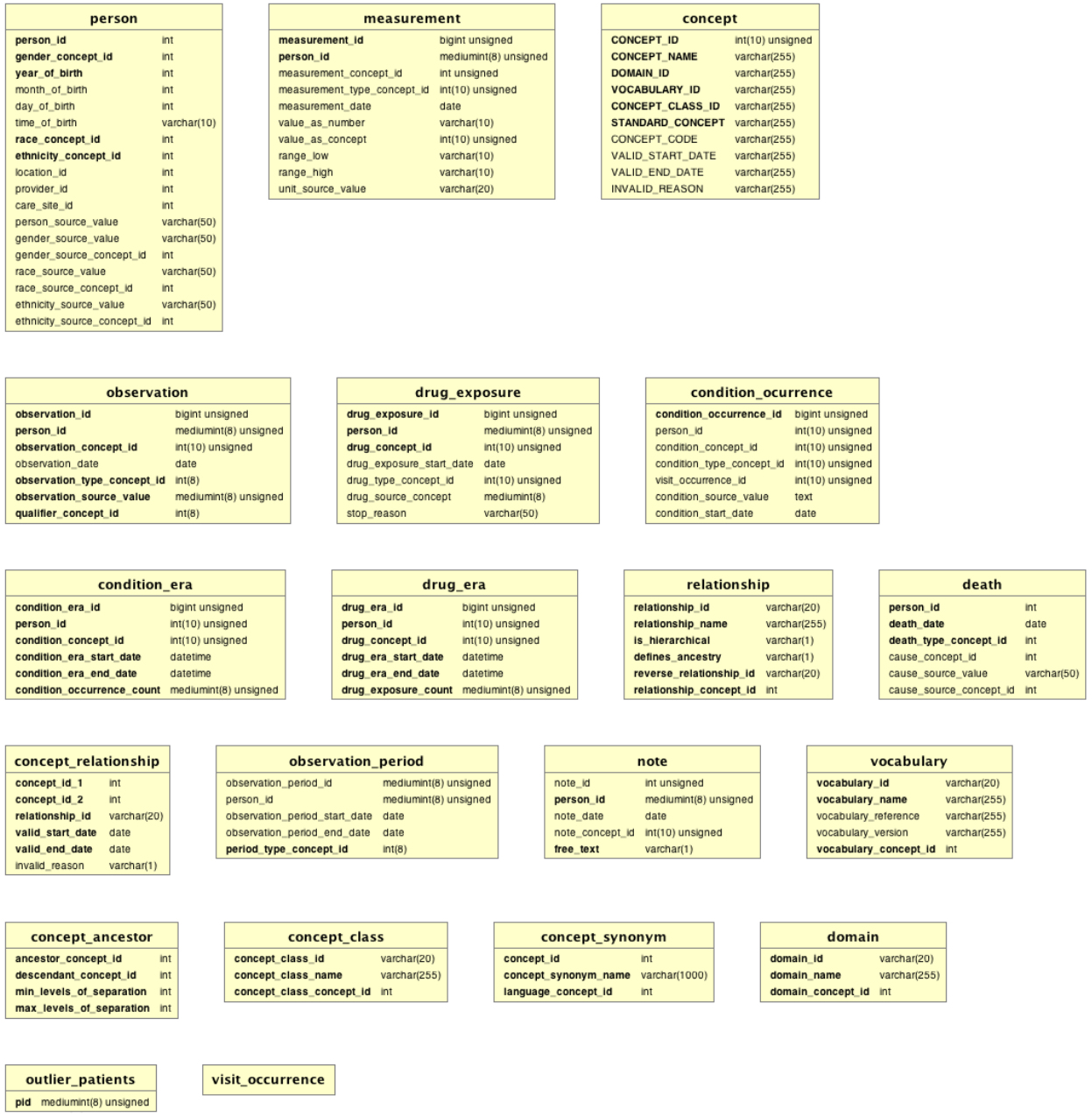
Automated **PH**enotype **R**outine for **O**bservational **D**efinition, **I**dentification, **T**raining and **E**valuation

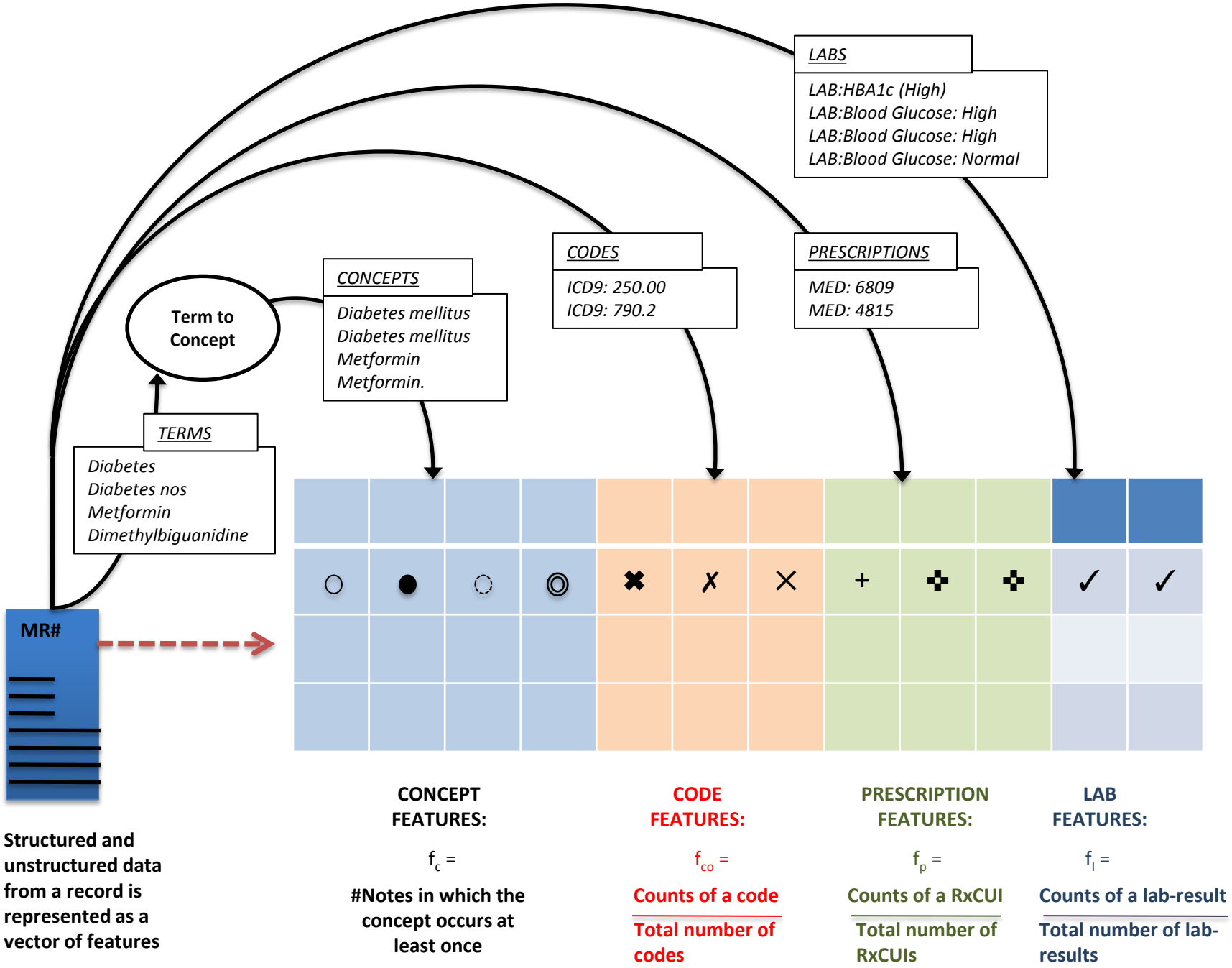


Phenotyping or Prediction



The source of features




LABS
LAB:HBA1c (High)
LAB:Blood Glucose: High
LAB:Blood Glucose: High
LAB:Blood Glucose: Normal
CODES
ICD9: 250.00
ICD9: 790.2
PRESCRIPTIONS
MED: 6809
MED: 4815
CONCEPTS
Diabetes mellitus
Diabetes mellitus
Metformin
Metformin.
Term to Concept
TERMS
Diabetes
Diabetes nos
Metformin
Dimethylbiguanidine
MR#
Structured and unstructured data from a record is represented as a vector of features
CONCEPT FEATURES:
f_c =
#Notes in which the concept occurs at least once
CODE FEATURES:
f_{co} =
Counts of a code / Total number of codes
PRESCRIPTION FEATURES:
f_p =
Counts of a RxCUI / Total number of RxCUIs
LAB FEATURES:
f_l =
Counts of a lab-result / Total number of lab-results

Models built using APHRODITE

- Diabetes Mellitus
- Myocardial Infarction
- Familial Hyperlipidemia
- Celiac disease

AUC	Sens.	Spec.	PPV
0.95	91 %	83 %	83 %
0.91	89 %	91 %	91 %
0.90	76.5%	93.6%	~20%
0.75	40 %	90 %	~4 %

Multi-class learning

- Diabetes Mellitus
- Myocardial Infarction

AUC	Sens.	Spec.	PPV
0.96	52 %	99 %	99 %
0.97	94 %	94 %	90 %

Learning multiple models using neural nets

[very preliminary results – see slide notes]

	AUC	Sens.	Spec.	PPV
1. Diabetes Mellitus	0.97	93 %	92 %	93 %
2. Myocardial Infarction	0.96	89 %	93 %	89 %
3. Familial Hyperlipidemia	0.83	7 %	99 %	3 %
4. Celiac disease	0.90	28 %	99 %	28 %
5. Acute liver injury	0.98	6 %	99 %	5 %
6. Acute renal injury	0.97	25 %	99 %	25 %
7. Congestive heart failure	0.95	48 %	99 %	48 %
8. Gastrointestinal complic.	0.90	10 %	99 %	9 %
9. Hepatitis C	0.90	36 %	99 %	36 %
10. Peripheral artery disease	0.95	25 %	99 %	25 %
11. Pancreatitis	0.91	23 %	99 %	23 %
12. Seizures	0.91	58 %	97 %	58 %

Discussion items

- Do we share models or the model building workflow (and retrain at each site)?
- Very few sites have data in CDM v5
- Storage of processed clinical notes is not standardized
- Incorporating ‘Anchor’ based learning (can help build predictive models)
- How do we use such models
  - cohort building
  - outcome ascertainment
  - clinician alerting (FIND FH example)
- How do such classifiers relate to consensus definition building?