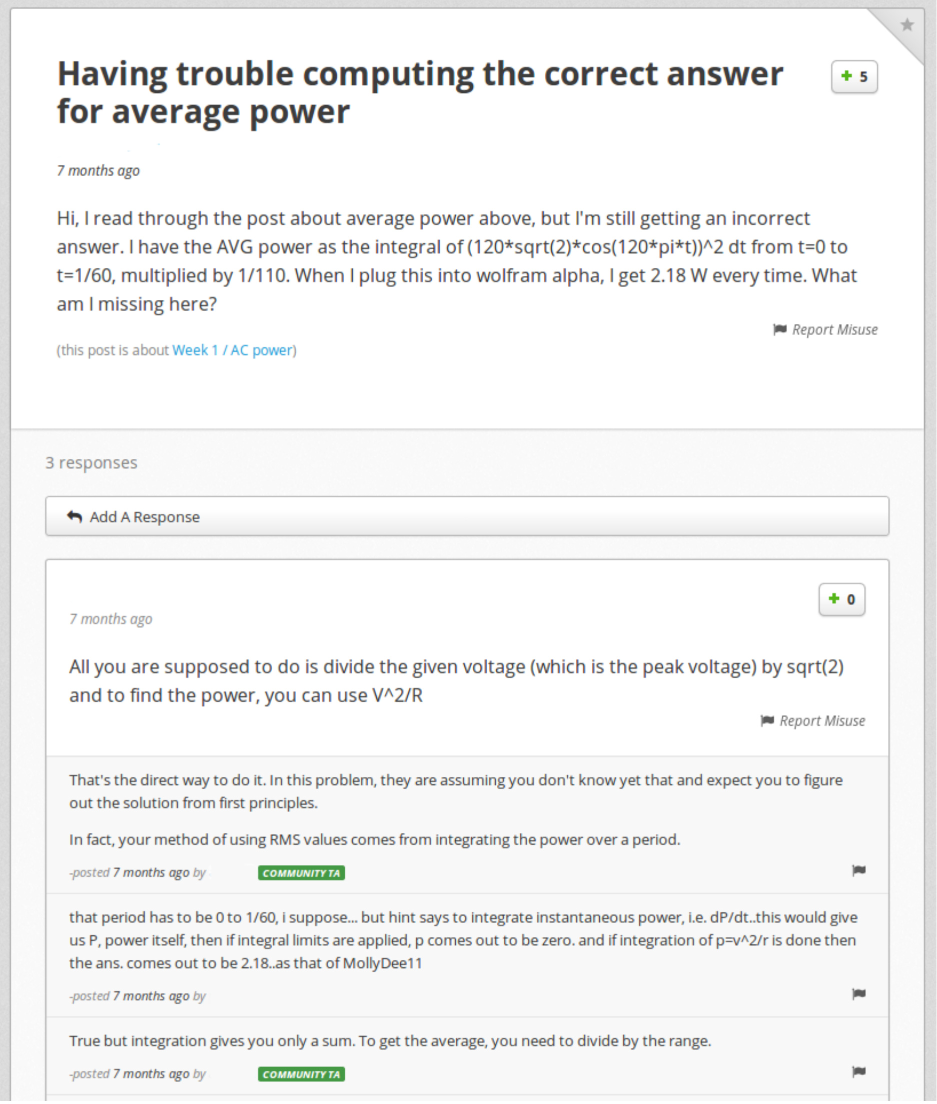
Delphi
Towards a Recommender System That Suggests Models and Parameters for Data

Kalyan Veeramachaneni

Joint work with
Will Drevo, Una-May O'Reilly

Any Scale Learning for All Group
CSAIL, MIT

An example data science project
Automatic tagging of MOOC forum posts



Having trouble computing the correct answer for average power

✚ 5

7 months ago

Hi, I read through the post about average power above, but I'm still getting an incorrect answer. I have the AVG power as the integral of (120*sqrt(2)*cos(120*pi*t))^2 dt from t=0 to t=1/60, multiplied by 1/110. When I plug this into wolfram alpha, I get 2.18 W every time. What am I missing here?

Report Misuse

(this post is about Week 1 / AC power)

3 responses

Add A Response

✚ 0

7 months ago

All you are supposed to do is divide the given voltage (which is the peak voltage) by sqrt(2) and to find the power, you can use V^2/R

Report Misuse

That's the direct way to do it. In this problem, they are assuming you don't know yet that and expect you to figure out the solution from first principles.

In fact, your method of using RMS values comes from integrating the power over a period.

-posted *7 months ago* by COMMUNITY TA

that period has to be 0 to 1/60, i suppose... but hint says to integrate instantaneous power, i.e. dP/dt..this would give us P, power itself, then if integral limits are applied, p comes out to be zero. and if integration of p=v^2/r is done then the ans. comes out to be 2.18..as that of MollyDee11

-posted *7 months ago* by

True but integration gives you only a sum. To get the average, you need to divide by the range.

-posted *7 months ago* by COMMUNITY TA

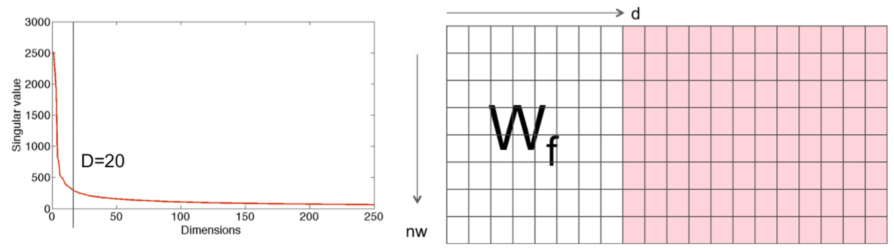
ALFA
ANYSCALE LEARNING FOR ALL

Automatic tagging of MOOC forum posts

CSAIL

Steps involved in data preparation

4. Select a subset form feature matrix



3. Latent semantic analysis

Posts → np=95K

Words

course	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	0	1	0	1	0
one	0	0	0	0	0	1	0	1	1	1	1	0	0	0	1	0	1	1	1	1
voltage	0	1	1	0	1	0	0	0	0	1	1	1	0	0	1	0	0	1	1	1
time	0	1	0	0	1	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0
answer	1	1	1	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	1	1
current	1	0	0	0	0	1	1	1	1	0	1	1	1	0	1	1	1	1	0	0
question	0	0	0	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	0	1
sub	1	0	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	0	1	0
nw=1000																				

Diagonal matrix

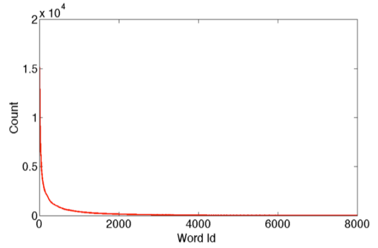
$$M = PQW$$

np X nw　　nw X np

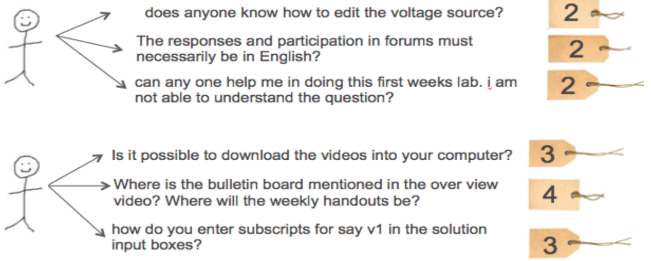
M – co-occurrence matrix
Q- Singular values

2. Bag of Words analysis

Top 10 Words	Counts
course	15171
one	14123
voltage	14050
time	12543
answer	11735
current	11484
question	11365
sub	11048
circuit	10762
out	10078



1. Get annotations from humans



does anyone know how to edit the voltage source? 2

The responses and participation in forums must necessarily be in English? 2

can any one help me in doing this first weeks lab. i am not able to understand the question? 2

Is it possible to download the videos into your computer? 3

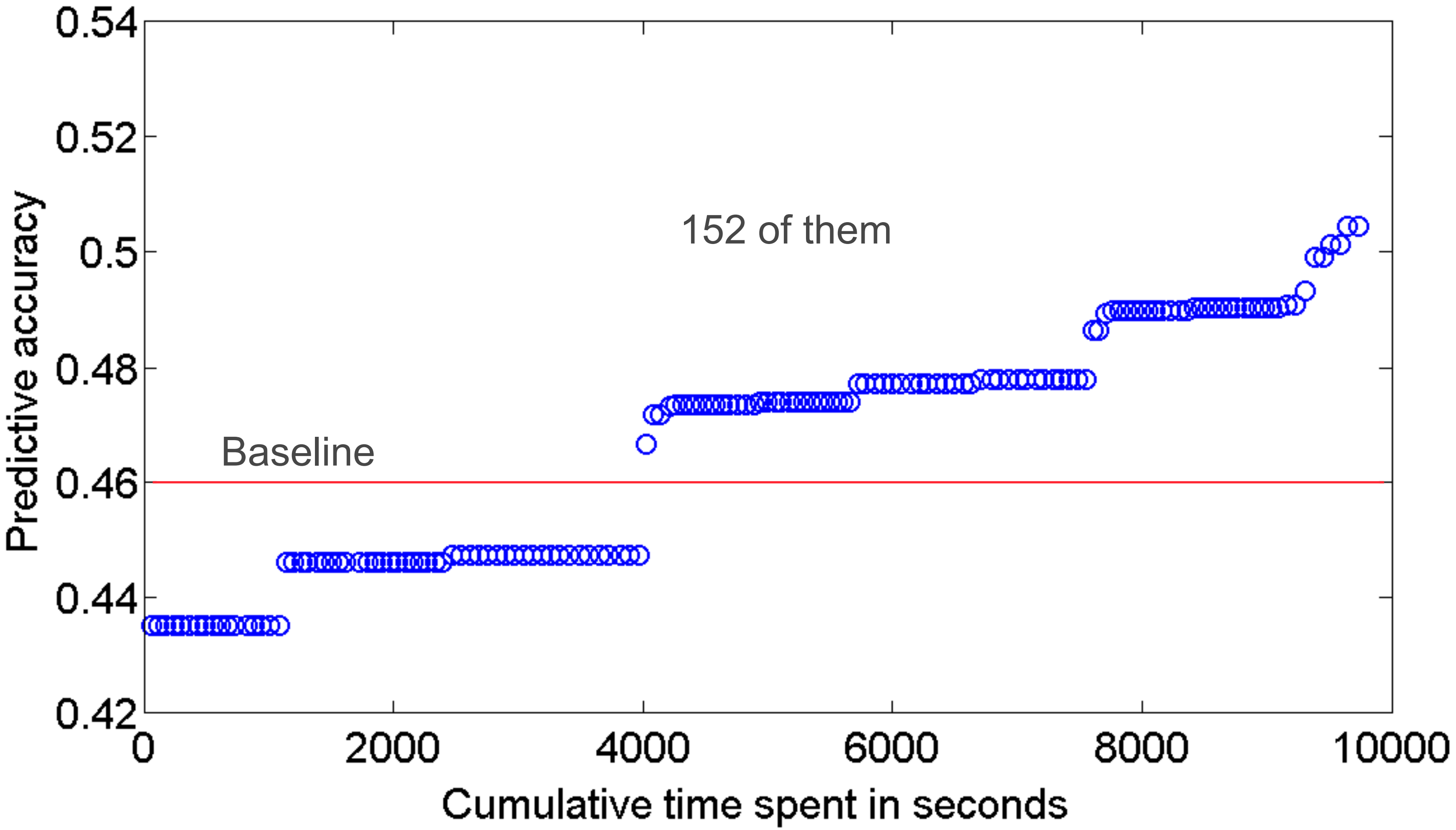
Where is the bulletin board mentioned in the over view video? Where will the weekly handouts be? 4

how do you enter subscripts for say v1 in the solution input boxes? 3

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

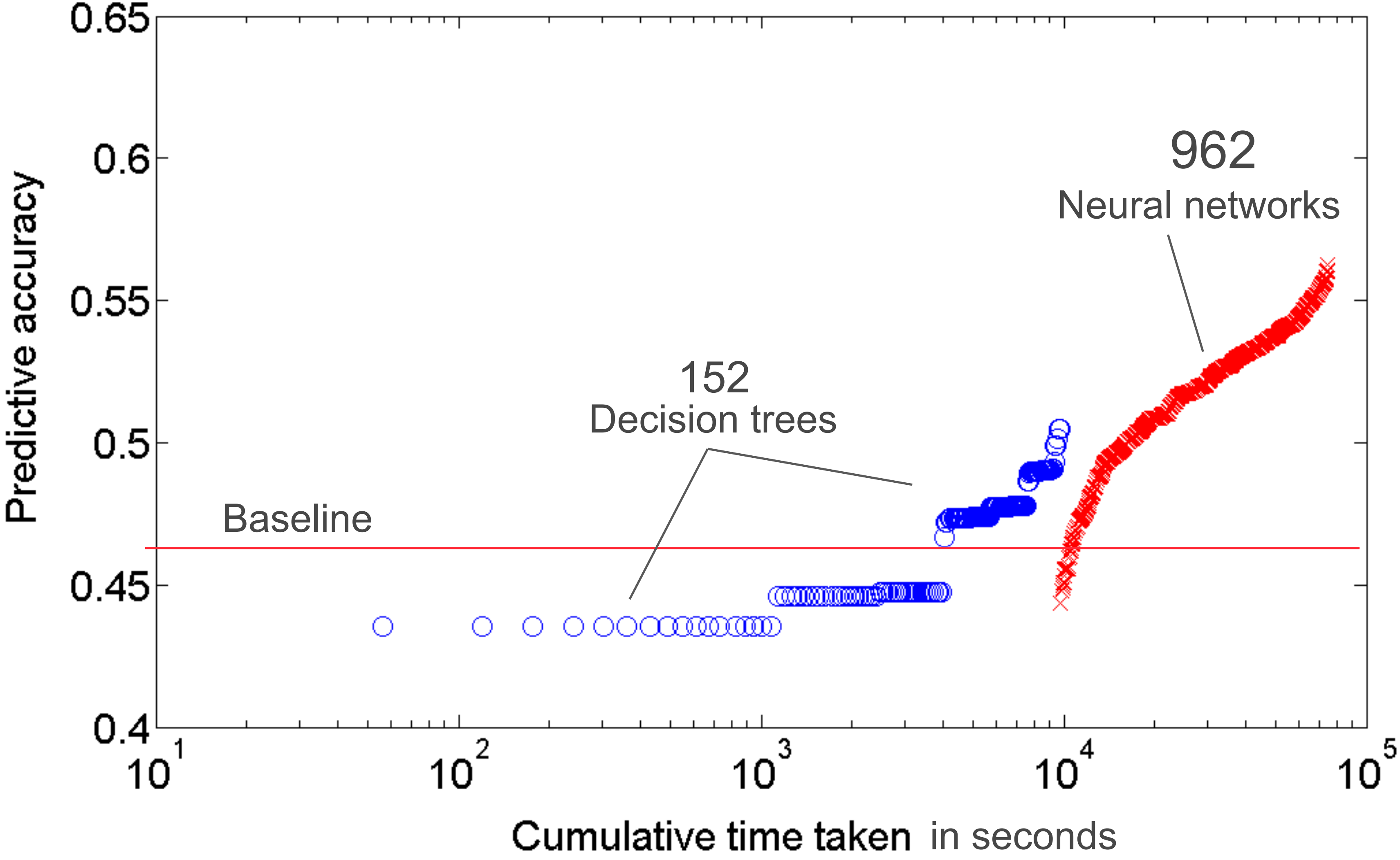
First try – Decision trees



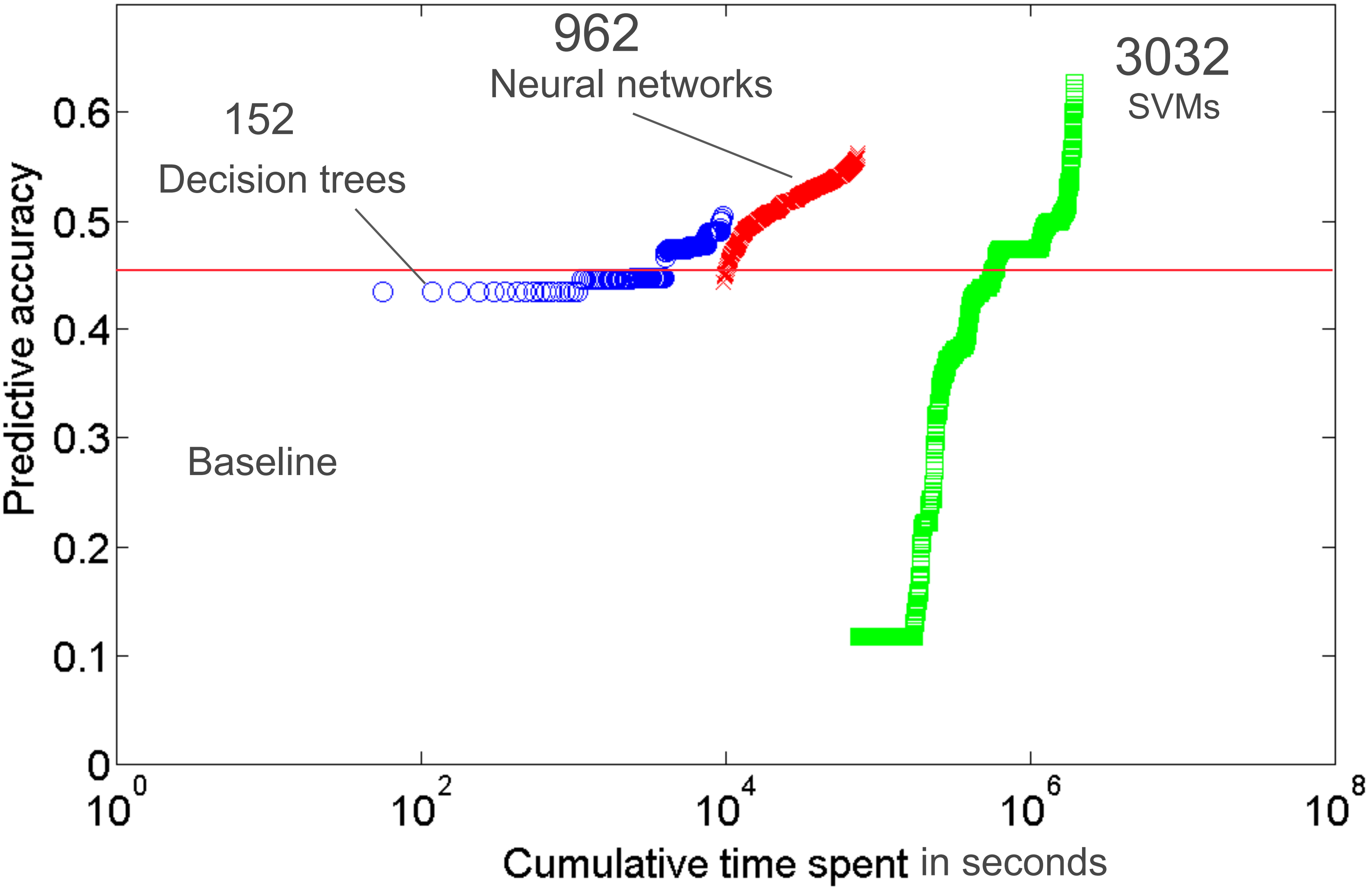
ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

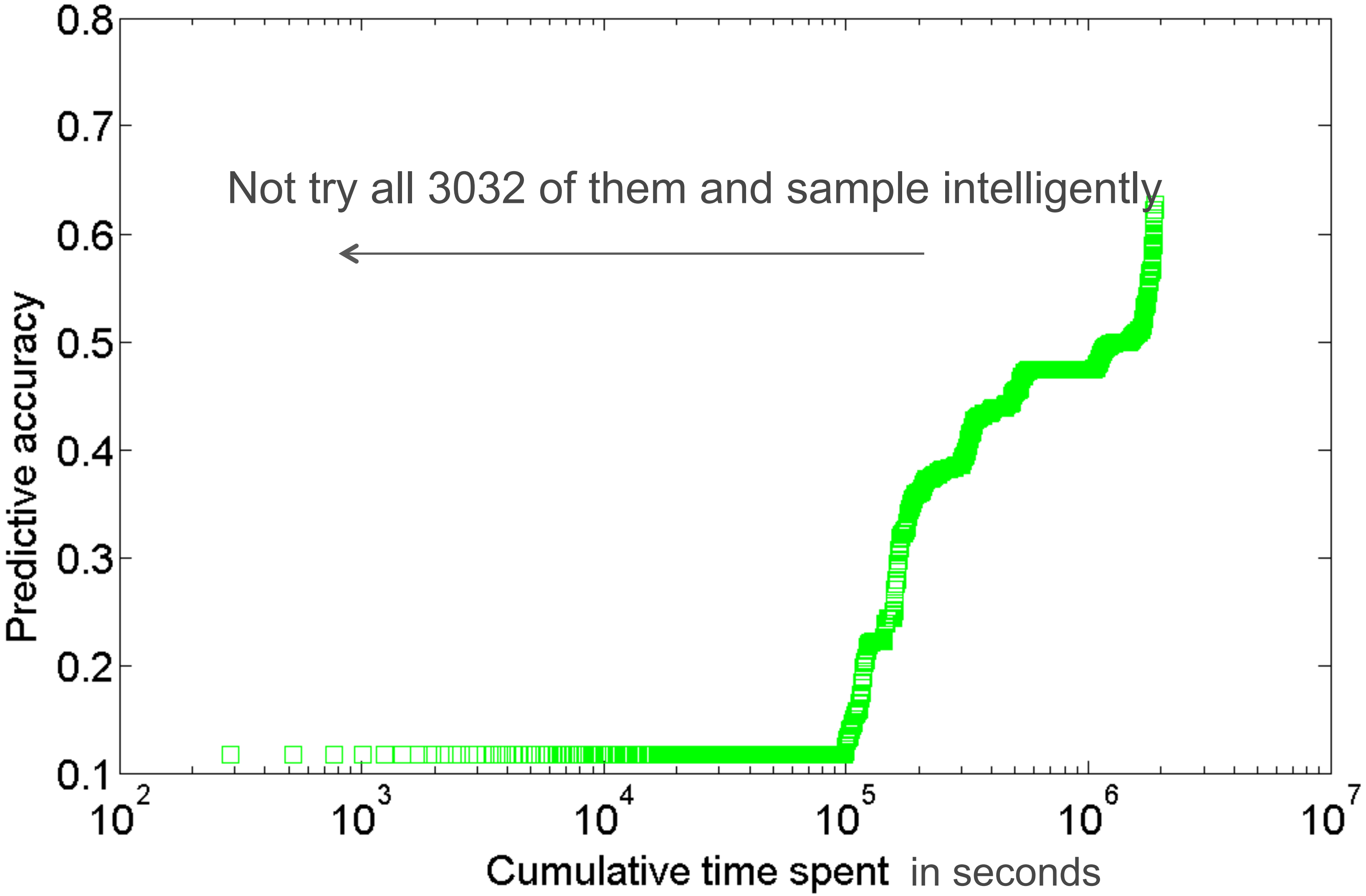
Then.. try Neural networks



Then.. try Support vector machines



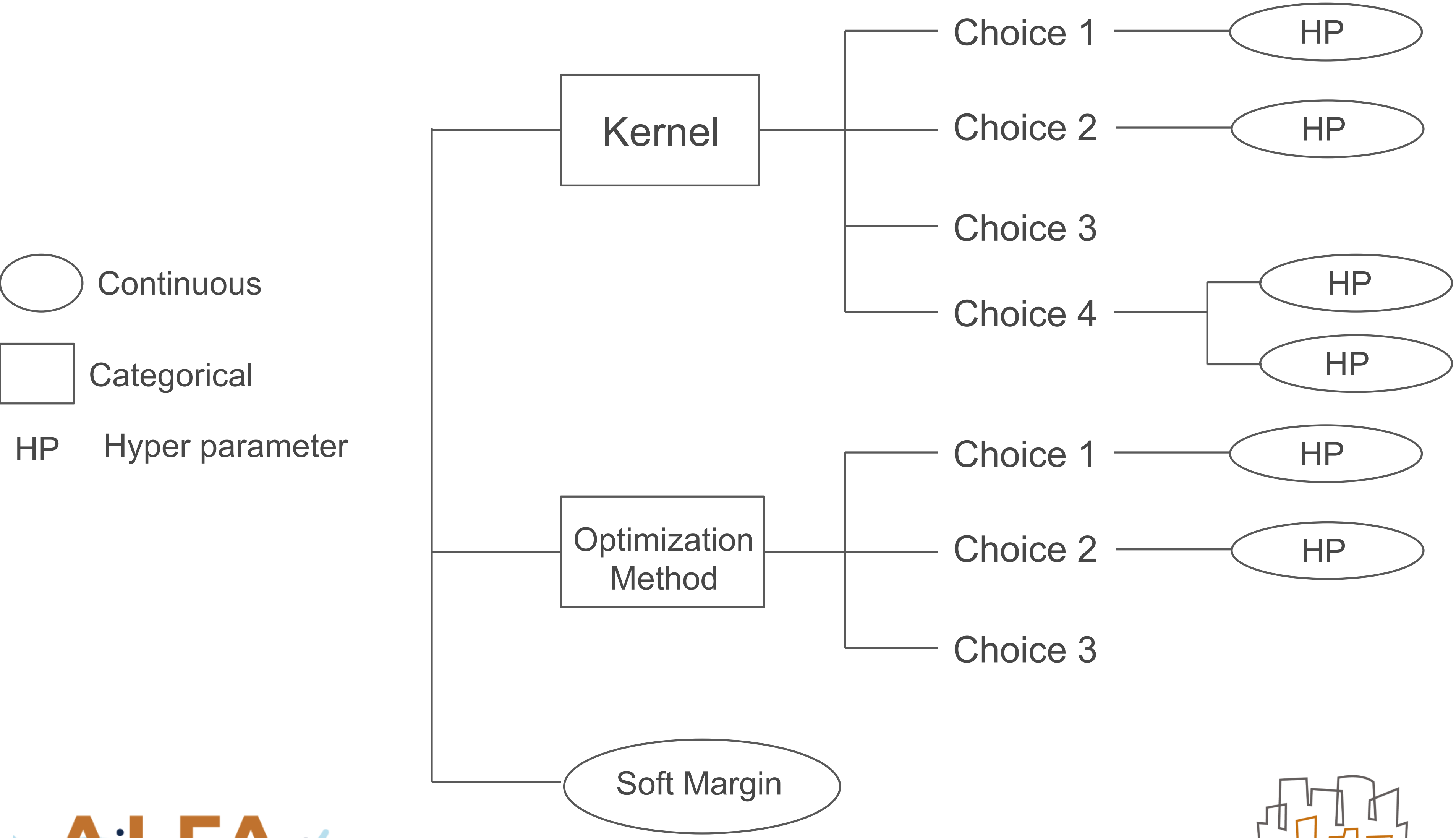
Perhaps I can do this intelligently ?



Not try all 3032 of them and sample intelligently

in seconds

Where do these 3032 possibilities for SVMs come from?



Kernel

Choice 1 — HP

Choice 2 — HP

Choice 3

Choice 4 — HP, HP

Optimization Method

Choice 1 — HP

Choice 2 — HP

Choice 3

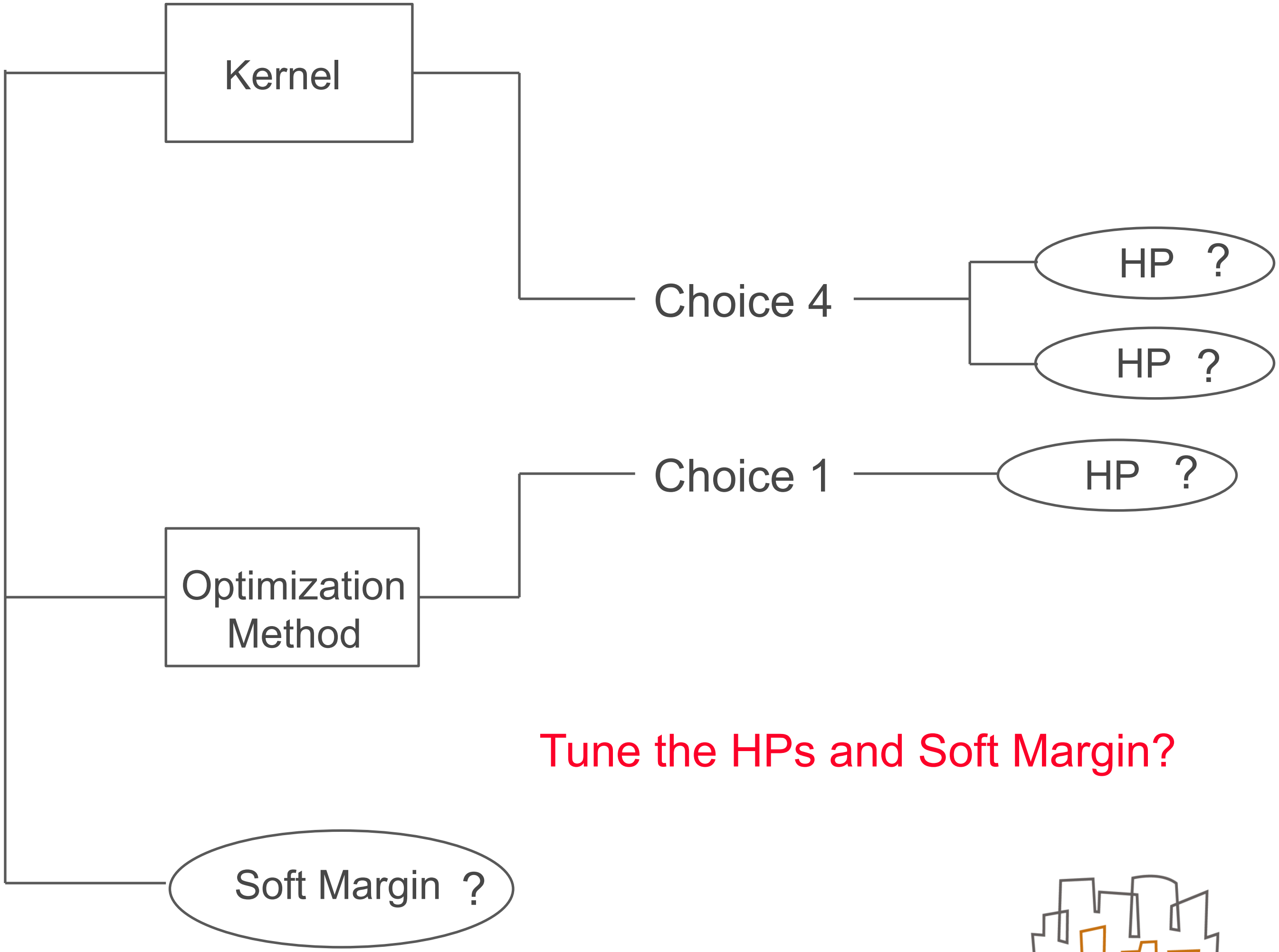
Soft Margin

○ Continuous

□ Categorical

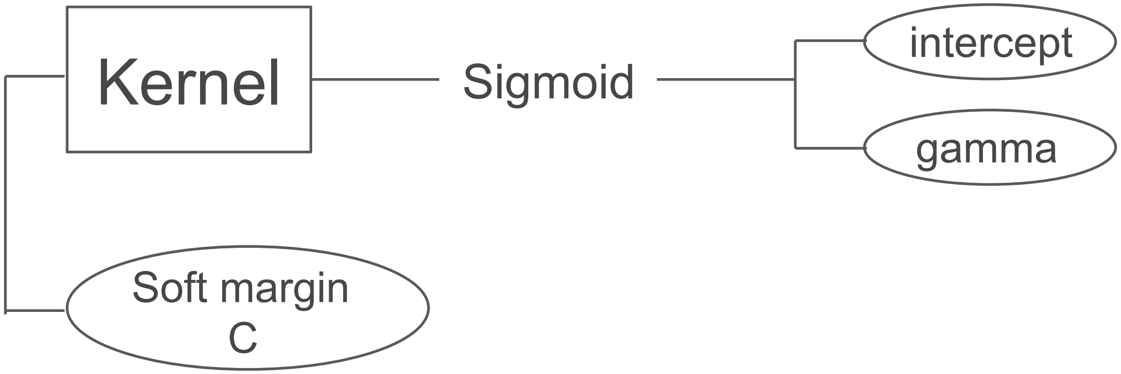
HP Hyper parameter

What is a Hyper partition within those choices?



Tune the HPs and Soft Margin?

Tuning hyper parameters



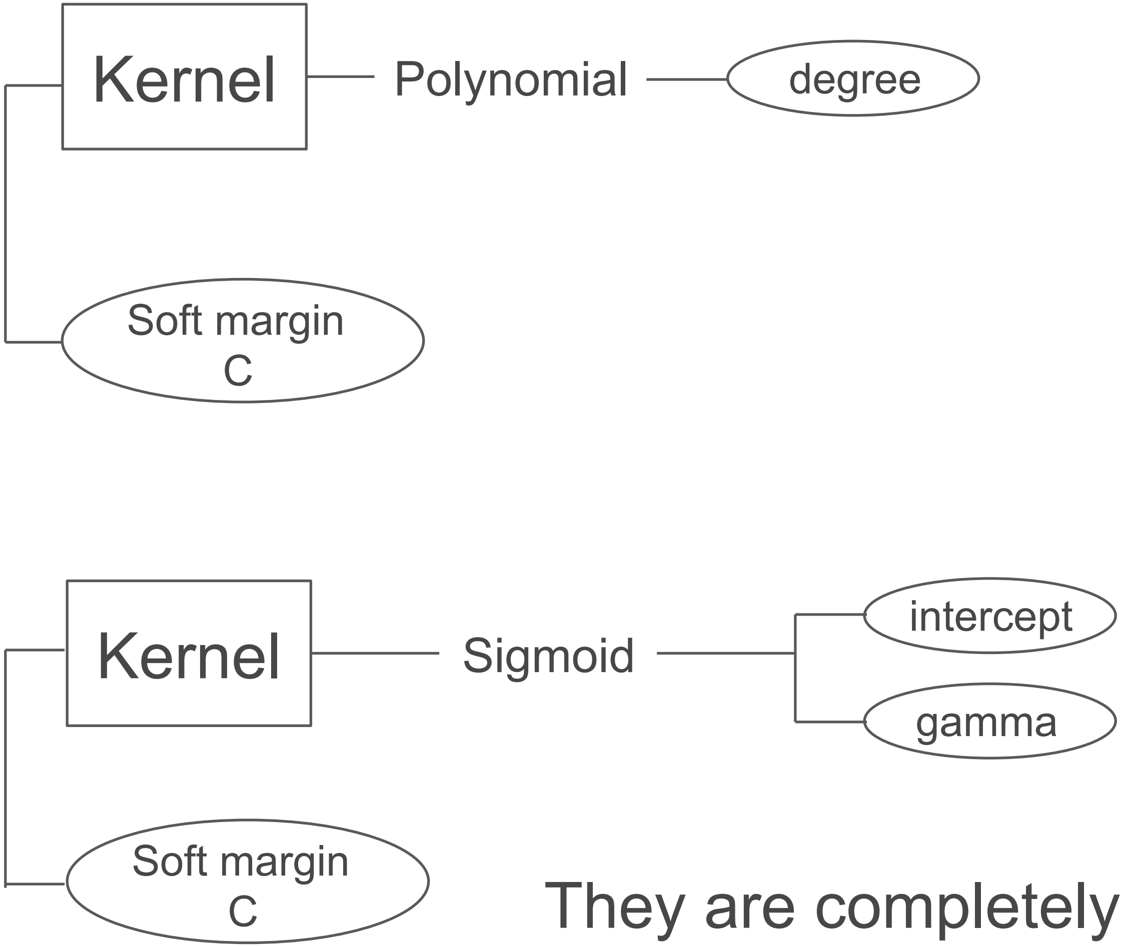
1. Sample a few combinations (*C*,*i*, *g*)
2. Model using Gaussian process

$$a = f_{GP}(C,i,g)$$

3. Predict using model for other parameters choices and propose the best

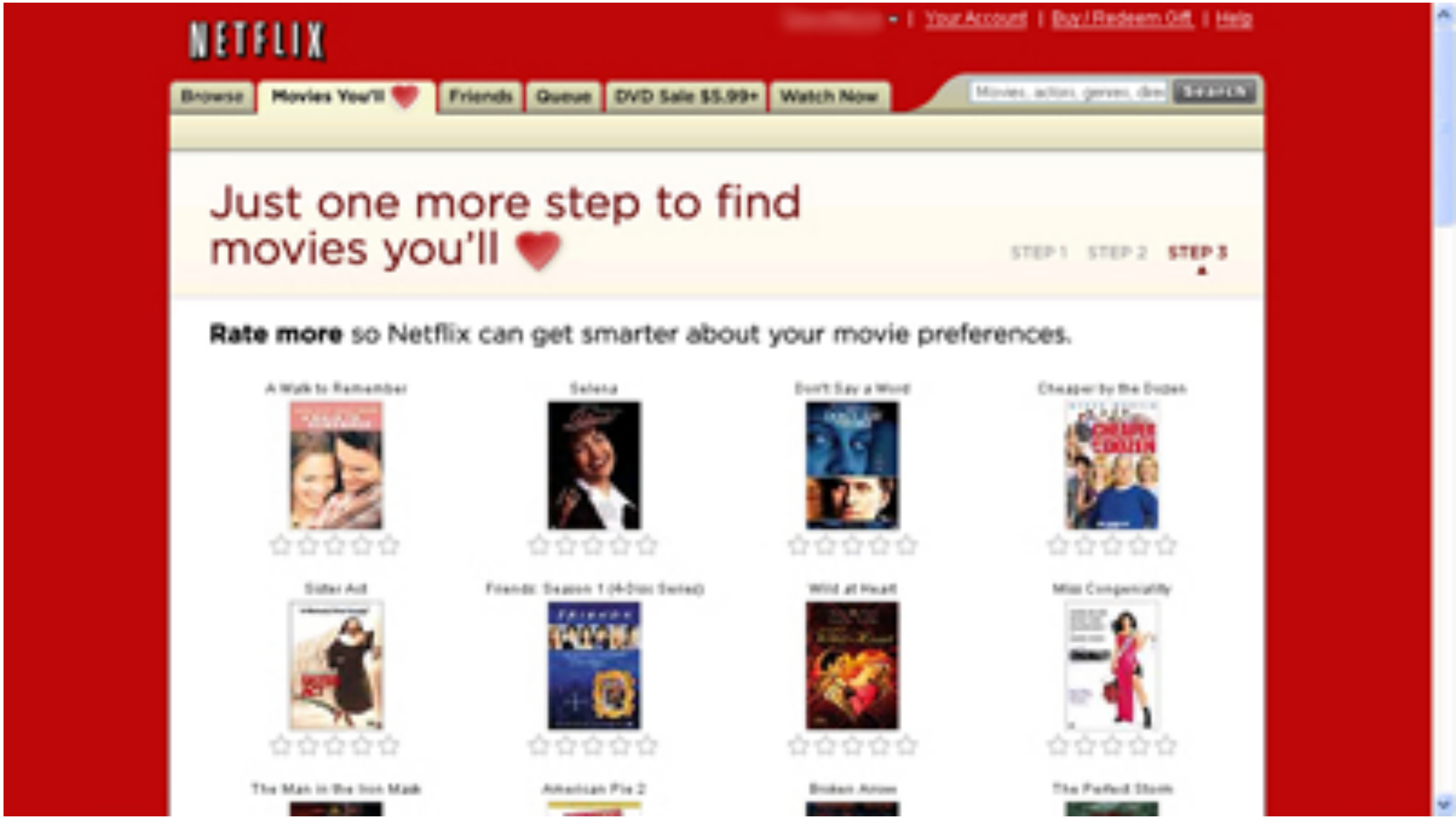
$$\{C^{new}, i^{new}, g^{new}\} = \underset{C,i,g}{\operatorname{argmax}} f_{GP}(C,i,g)$$

Two hyper partitions



They are completely different search spaces !

Can we?



Just one more step to find movies you'll ♥

Rate more so Netflix can get smarter about your movie preferences.

Generate recommendations for datasets?

A critical ingredient for making recommendations

- User item matrix stores for each user the rating for the items

	i_1	i_2	i_3	i_4	i_5	...	i_M
u_1	2	0	3	2	5	...	1
u_2	0	4	0	0	0	...	5
u_3	0	2	0	0	0	...	4
u_4	1	0	4	2	4	...	2
...	...	...	...	...	...	...	...
u_K	2	...	4	...	4	...	1

Predicting unknown ratings

So for us …

Datasets →

Approaches

	Skin	Adult	Bank Credit	Spam	Fetal	Car Safty	Banknote	Diabates	Transfusion	Breast	Credit	WDBC	Climate	Congress	Unnary	ABP	MOOC
SVM1																	
⋮																	
SVM4000																	
DT1																	
⋮																	
DT2355																	
DBN1																	
⋮																	
DBN4400																	
NN1																	
⋮																	
NN8411																	
RF1																	
⋮																	
RF10461																	
LR																	

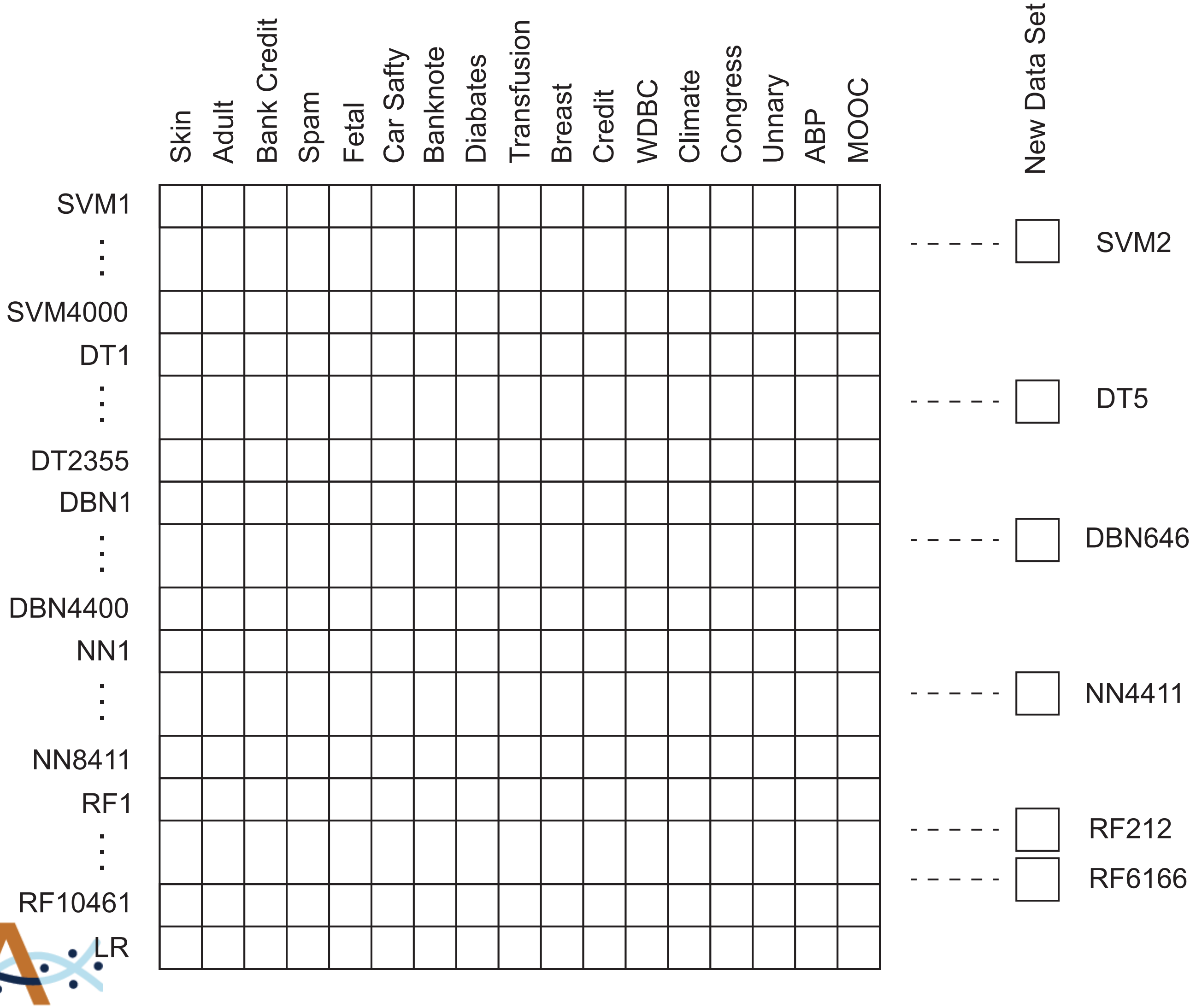
ALFA

ANYSCALE LEARNING FOR ALL

CSAIL

How would we use such a matrix?

Step 1: For a new dataset try a few models randomly



How would we use such a matrix?

Step 2: Correlate with the other datasets in the “approach-performance” space

MOOC dataset	New Data Set			MOOC dataset	New Data Set
0.66	0.32	SVM2		3	6
0.71	0.63	DT5		2	2
0.82	0.61	DBN646	Rank →	1	3
0.64	0.71	NN4411		4	1
0.48	0.48	RF212		5	4
0.89	0.38	RF6166		6	5

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

How could we use such a matrix?

Step 3: Identify the dataset that is correlated to this one most

Skin	Adult	Bank Credit	Spam	Fetal	Car Safty	Banknote	Diabates	Transfusion	Breast	Credit	WDBC	Climate	Congress	Unnary	ABP	MOOC
0.22	0.43	0.61	0.12	0.11	0.18	0.84	0.43	0.66	0.77	0.13	0.64	0.23	0.88	0.14	0.21	0.74

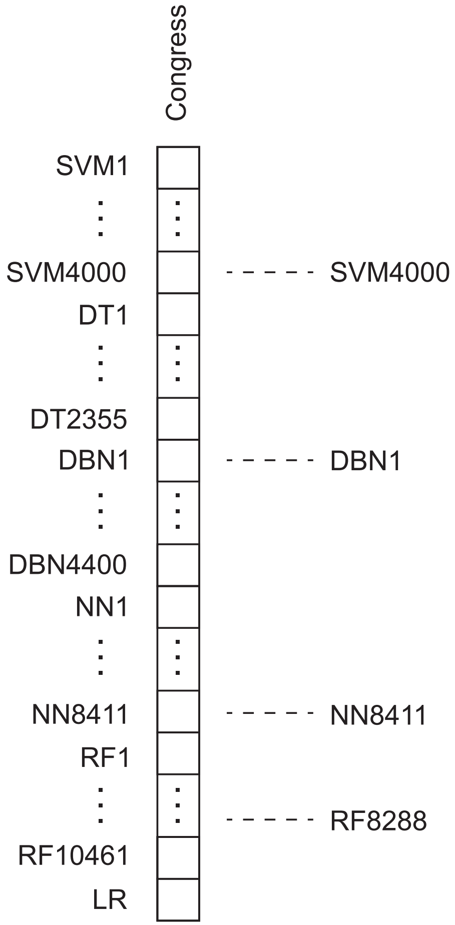
Most correlated with

ANYSCALE LEARNING FOR ALL

CSAIL

How would we use such a matrix?

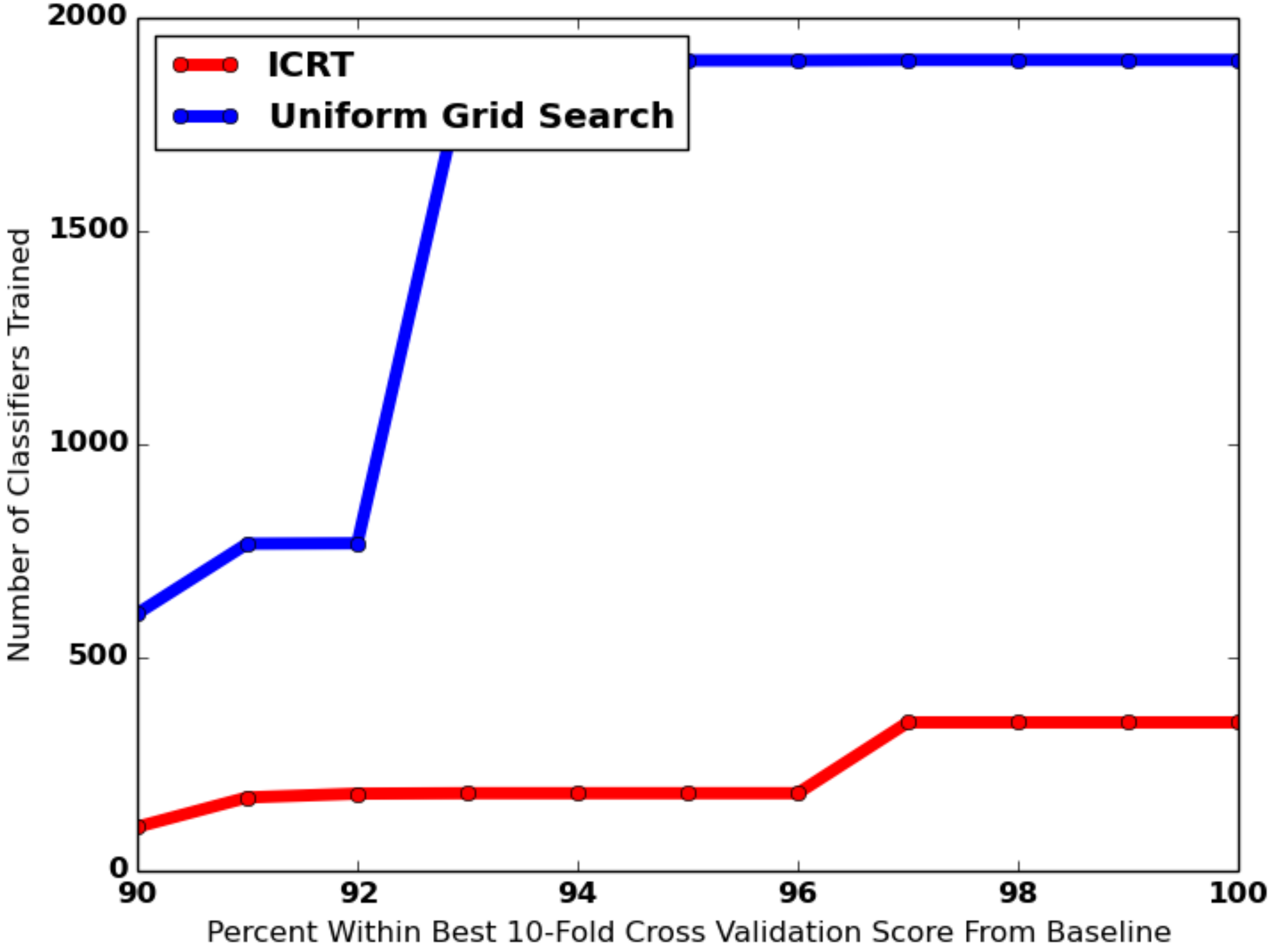
Step 4:Choose the best approaches for that dataset and propose for the new dataset



We compiled such a matrix

- **With 30 different datasets**
- **And around 5000 modeling approaches (different models, different parameters and hyper parameters)**
- **We learnt a total of 1.5 Million models**
- **Still accumulating more**

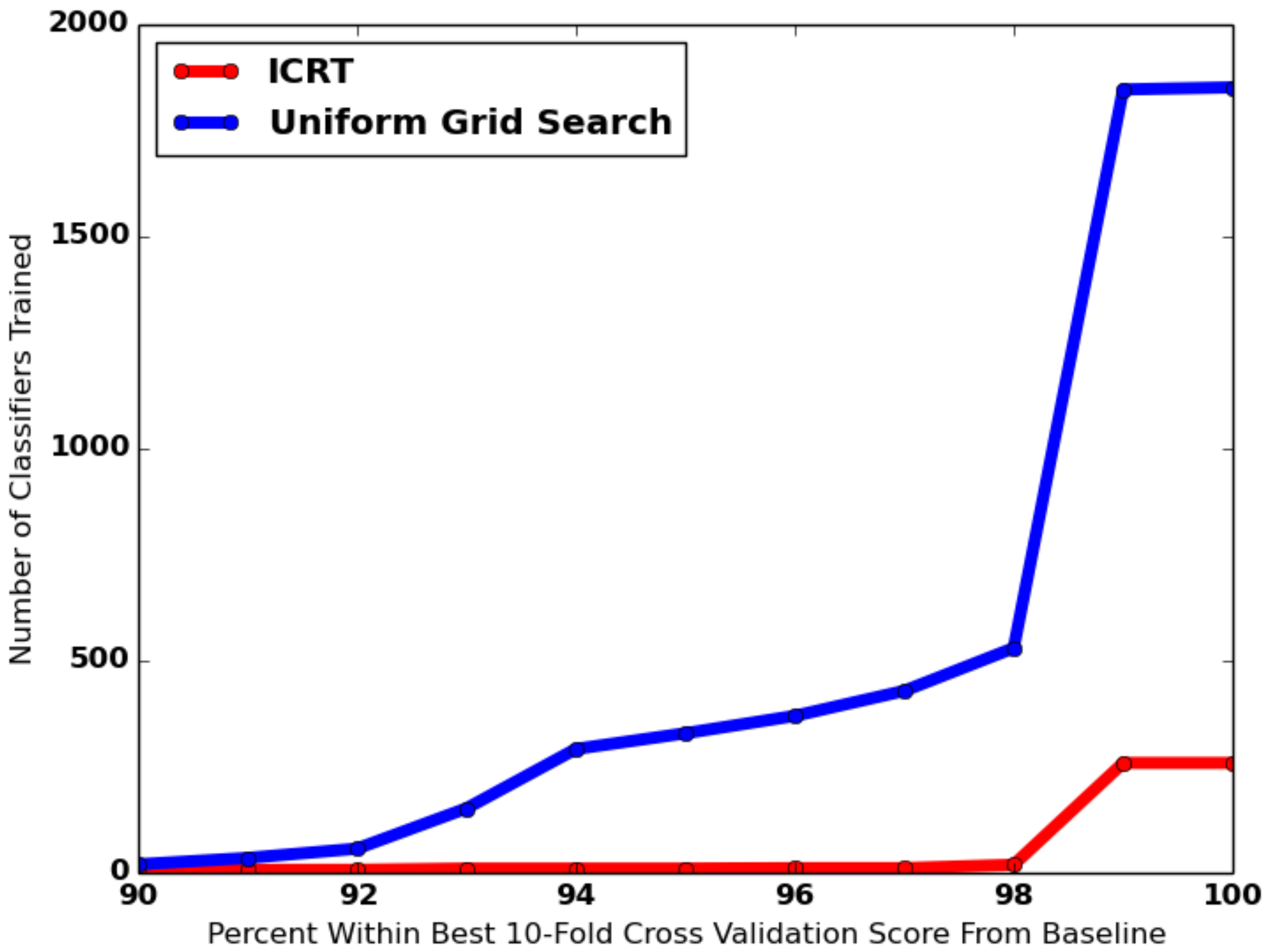
For a real case study
MOOC student Stopout prediction



ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Another dataset



ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

More datasets will help

Datasets →

Approaches

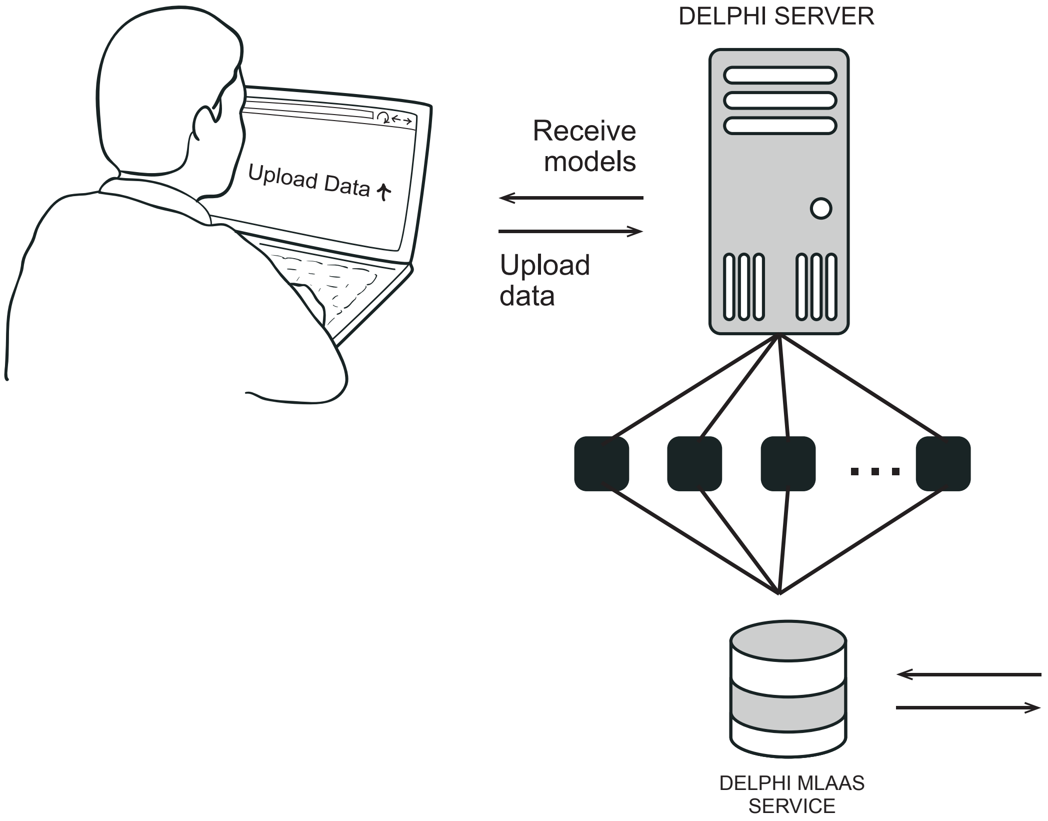
	Skin	Adult	Bank Credit	Spam	Fetal	Car Safty	Banknote	Diabates	Transfusion	Breast	Credit	WDBC	Climate	Congress	Unnary	ABP	MOOC
SVM1																	
⋮																	
SVM4000																	
DT1																	
⋮																	
DT2355																	
DBN1																	
⋮																	
DBN4400																	
NN1																	
⋮																	
NN8411																	
RF1																	
⋮																	
RF10461																	
LR																	

More datasets

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Bring your data

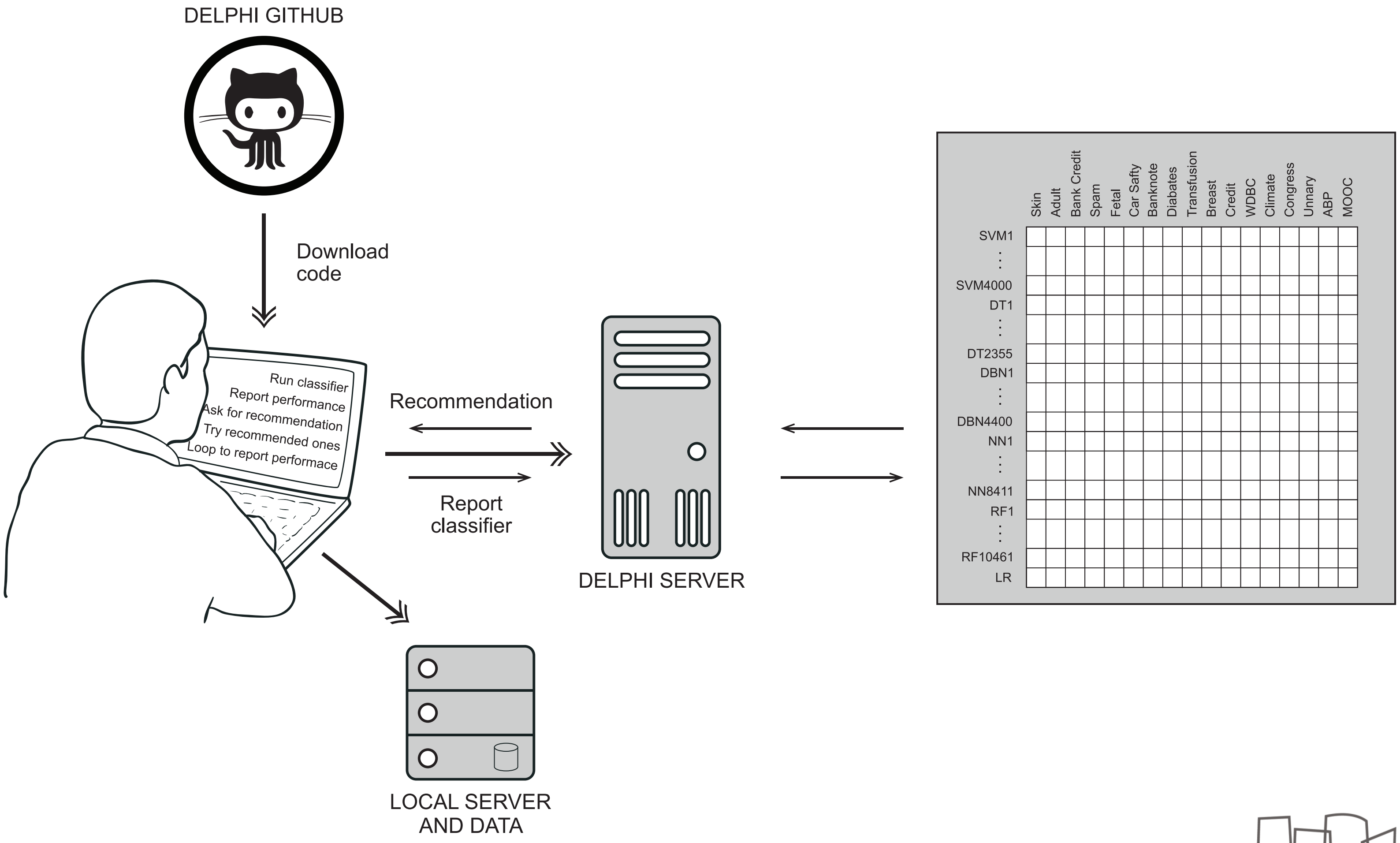


	Skin	Adult	Bank Credit	Spam	Fetal	Car Safty	Banknote	Diabates	Transfusion	Breast	Credit	WDBC	Climate	Congress	Unnary	ABP	MOOC
SVM1																	
⋮																	
SVM4000																	
DT1																	
⋮																	
DT2355																	
DBN1																	
⋮																	
DBN4400																	
NN1																	
⋮																	
NN8411																	
RF1																	
⋮																	
RF10461																	
LR																	

ANYSCALE LEARNING FOR ALL

CSAIL

Query the recommender system



Conclusions

- We can now use experiences from previous data science projects to help inform the new projects
- This will require
  - A systematic way of storing the data pertaining to the data science projects in your entity and history of modeling approaches tried on those
  - Build infrastructure and approaches to make recommendations and accumulate more experience as a result.
- Extend this to the entire pipeline and not just modeling
  - Data preparation
  - Cleaning
  - Feature extraction

Data scavenger

Questions ?

Thank you

ANYSCALE LEARNING FOR ALL

CSAIL