Where art thou big data?

Identifying and Harnessing Sources of Data for MOOC Data Science

Kalyan Veeramachaneni

Joint work with:
Colin Taylor, Elaine Han, Franck Dernoncourt, Una-May O'Reilly

Any Scale Learning for All Group
CSAIL, MIT

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Why did I use this title ?

Where art thou big data ?

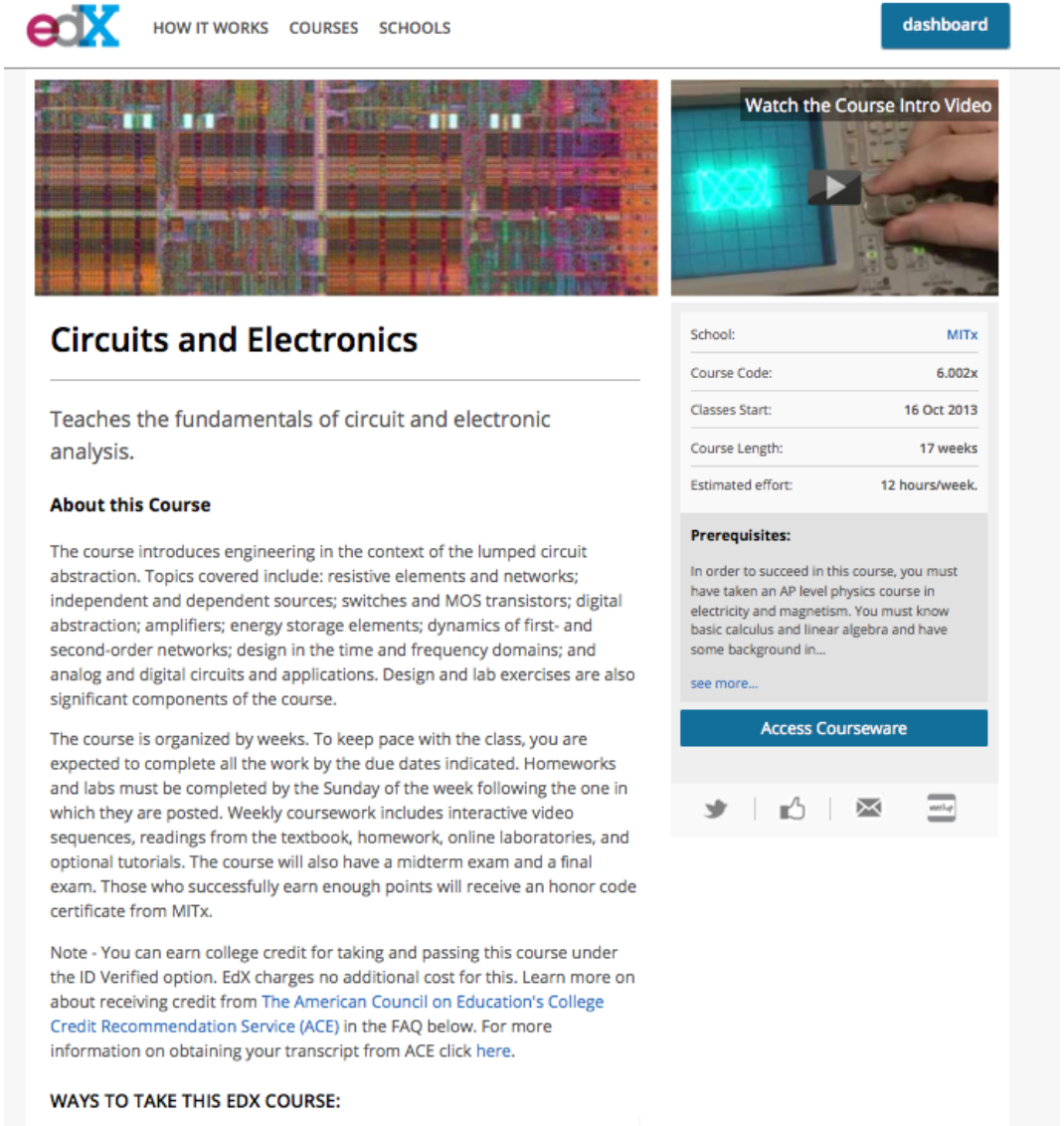
Overview

- Data collected in MOOCs
- Dataset we use
  - Properties
  - Size
  - Organization
- Challenges of Feature/Variable engineering
  - What does it entail?
  - How we are scaling it up? (or would like to)
- Stopout prediction problem
  - Definition
  - Machine learning models
  - How features/variables we extracted mattered
  - Results
- Conclusions

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

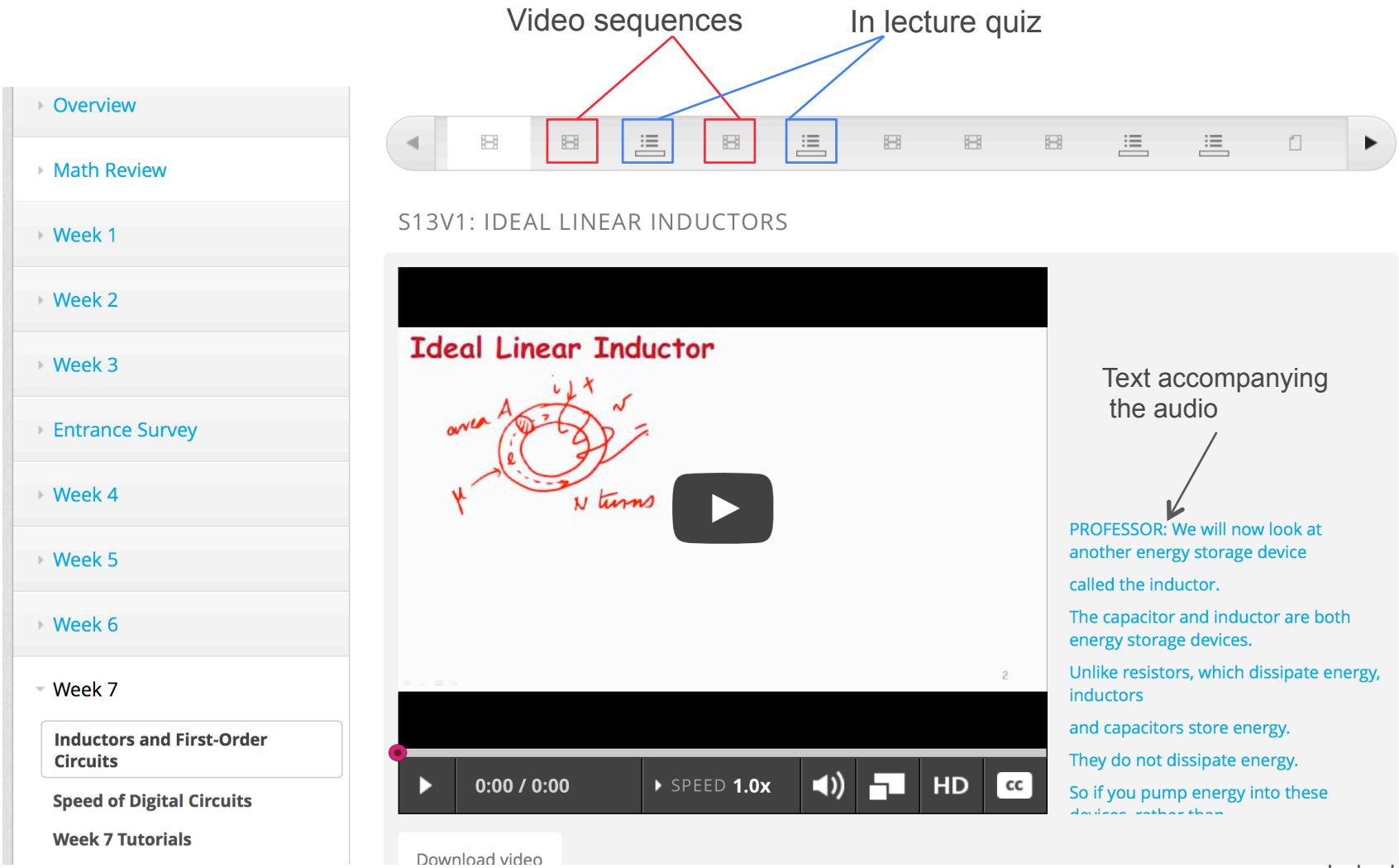
6.002x: Circuits and Electronics



ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Lectures (Inductors, 1st order circuits)

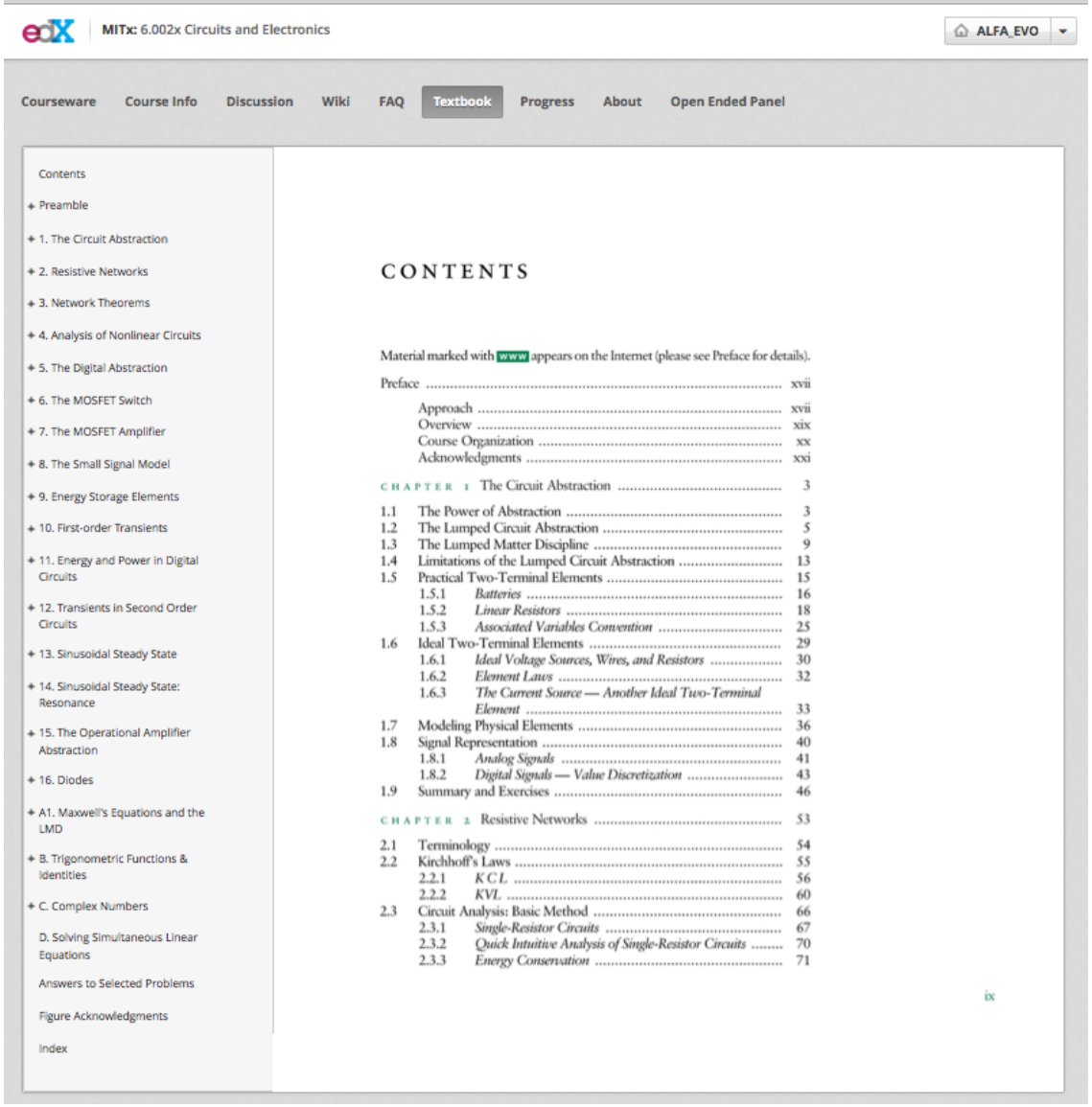


ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Textbook

Each page is uniquely identified by a URL



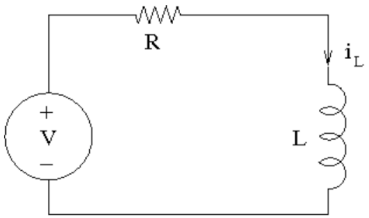
ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Problem Solution Entry/Retry

S13E3: THEVENIN INDUCTOR CIRCUIT (1/2 points)

For the circuit shown below,



write an algebraic expression for the rate of change of the inductor current in terms of V, R, L, and i_L.

Rate of change of inductor current $\frac{di_L}{dt}$ =

(V-R*iL)/L ✔

Let $R = 1.0\text{k}\Omega$ and $L = 10\mu\text{H}$. If at some time t we have $V(t) = 1.0\text{V}$ and $i_L(t) = 0.9\text{mA}$ what is the rate of change of inductor current? Express your answer in Amperes/millisecond.

9 ✖

Check | Show Answer(s)

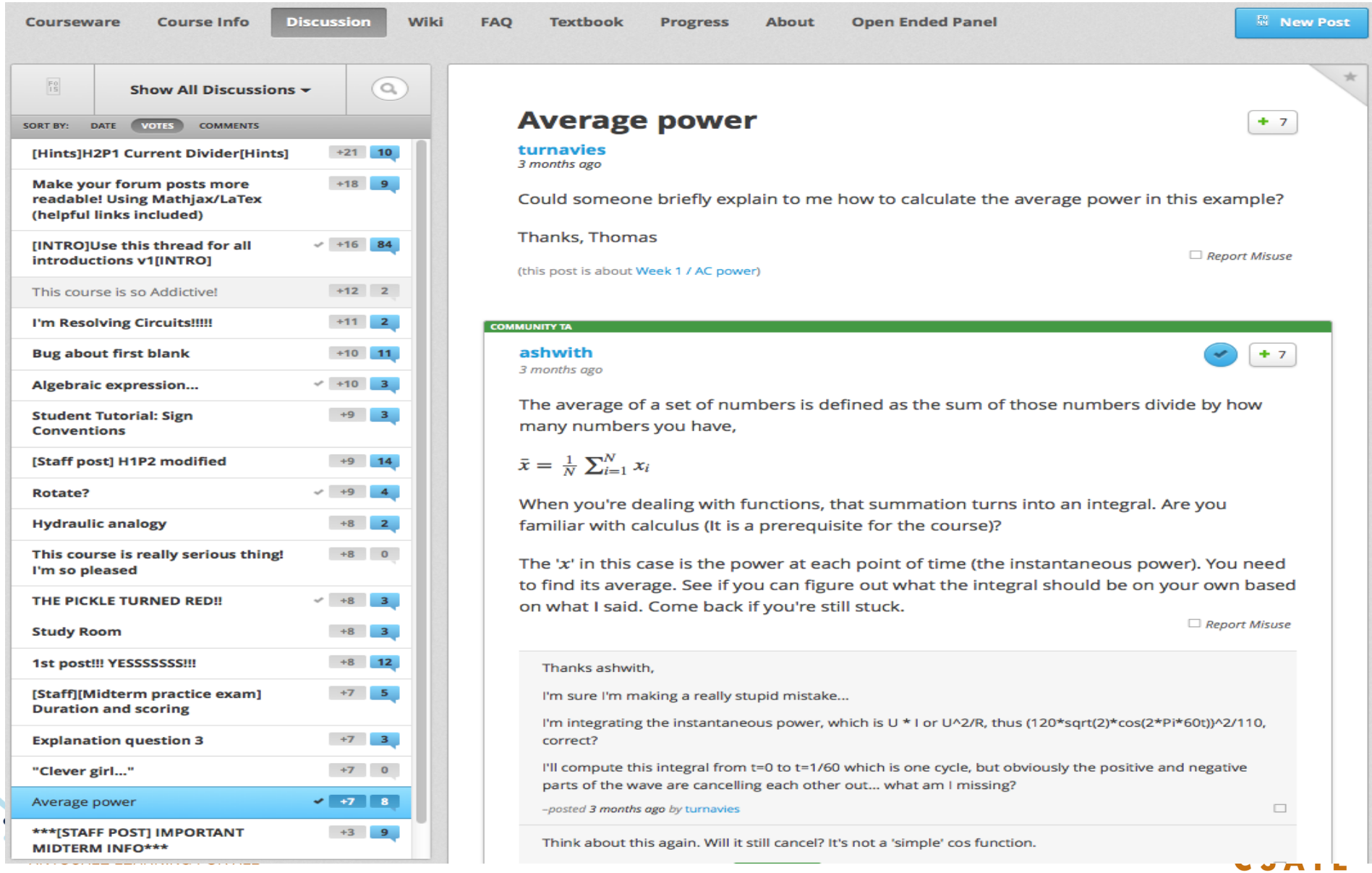
Staff Debug Info

Submission history

ALFA ANYSCALE LEARNING FOR ALL

CSAIL

Discussion board



Data: Students navigation on the website

{username: kalyan,
event_source: browser,
event_type: pause_video,
ip: 18.211.1.223,
agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_8_5),
page: https://6002x.mitx.mit.edu/courseware/6.002_Spring_2012/Week_3/Inside_the_Gate/,
time: 120913101}

	User_id	url_id	Timestamp	Duration (sec)	IP	OS
1	11023	22	8/15/2013 12:00:47	60	18.21.1.223	windows 7
2	344	34	8/15/2013 12:00:47	300	12.22.211.2	mac 10.5
3	3424	12	8/15/2013 12:00:47	420	311.3.23.1	windows xp
4	8982	1234	8/15/2013 12:00:47	60	120.12.11.1	windows 7
5	11023	333	8/15/2013 12:00:47	300	12.45.22.2	mac 10.5
6	889	232	8/15/2013 12:00:47	420	233.31.3.34	windows xp
.....	344	22	8/15/2013 12:00:47	60	122.21.11.1	windows 7
.....	11023	32	8/15/2013 12:00:47	300	123.23.21.2	mac 10.5
.....	8982	12	8/15/2013 12:00:47	420	132.33.1.31	windows xp

~133 Million

url_id	url
22	https://6002x.mitx.mit.edu/courseware/6.002_Spring_2012/Week_3/Inside_the_Gate/
32	https://6002x.mitx.mit.edu/courseware
1234	https://6002x.mitx.mit.edu/courseware/6.002_Spring_2012/Week_9/Homework
232	https://6002x.mitx.mit.edu/courseware/6.002_Spring_2012/Week_8/State_and_Memory
.....	.....

Url dictionary

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Data: Students submissions for problems

id	user_id	problem_id	timestamp	attempt number	answer	is_submitted	IP	OS
1	889	2	8/15/2013 12:00:47	1	123	yes	1.1.1.1	windows 7
2	11023	4	8/15/2013 12:00:47	2	1	yes	2.2.2.2	mac 10.5
3	8982	3	8/15/2013 12:00:47	4	23.4	yes	3.3.3.3	windows xp
4	6465	4	8/15/2013 12:00:47	1	33	yes	1.1.1.1	windows 7
5	11023	2	8/15/2013 12:00:47	5	123	yes	2.2.2.2	mac 10.5
6	344	3	8/15/2013 12:00:47	6	23.4	yes	3.3.3.3	windows xp
7	889	2	8/15/2013 12:00:47	7	232	yes	1.1.1.1	windows 7
8	989	2	8/15/2013 12:00:47	8	123	no	2.2.2.2	mac 10.5
9	11023	3	8/15/2013 12:00:47	9	33.6	yes	3.3.3.3	windows xp

Problem meta information is stored in another table

problem name - hw1p2

problem type- homework

problem release timestamp, deadline, max_submission, max_duration, etc

Assessment information is stored in another table

Feedback, grade, penalty, graded by who, when was it graded

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

6.002x data

- **154,763 registered students**
- **17.8 million submission events**
- **132.3 million navigational events**
- **~90,000 forum posts**

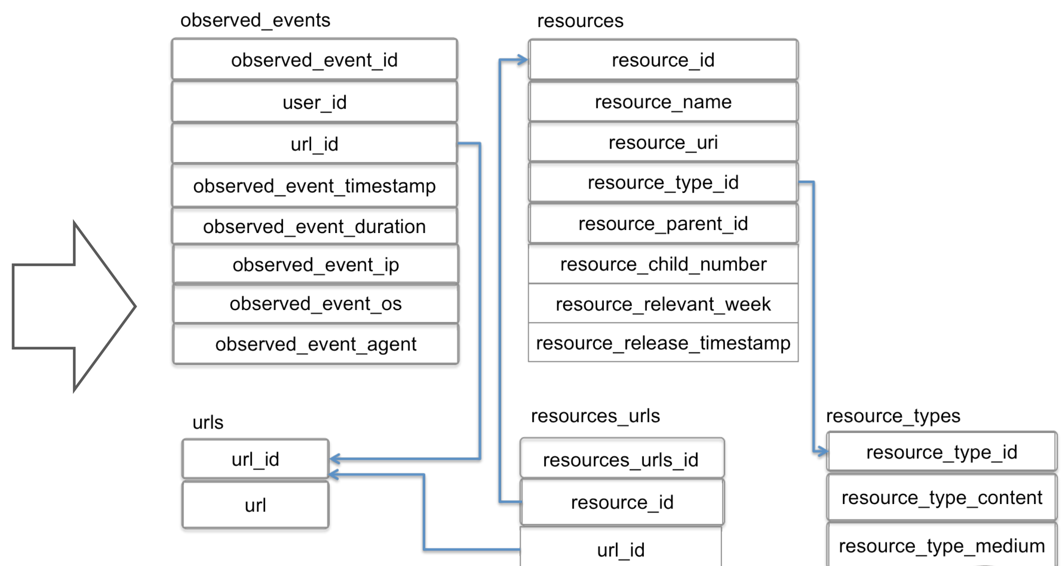
But is this big data?

{username: kalyan,
event_source: browser,
event_type: pause_video,
ip: 18.211.1.223,
agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_8_5),
page: https://6002x.mitx.mit.edu/courseware/6.002_Spring_2012/Week_3/Inside_the_Gate/,
time: 120913101}

{username: kalyan,
event_source: browser,
event_type: pause_video,
ip: 18.211.1.223,
agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_8_5),
page: https://6002x.mitx.mit.edu/courseware/6.002_Spring_2012/Week_3/Inside_the_Gate/,
time: 120913101}

{username: kalyan,
event_source: browser,
event_type: pause_video,
ip: 18.211.1.223,
agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_8_5),
page: https://6002x.mitx.mit.edu/courseware/6.002_Spring_2012/Week_3/Inside_the_Gate/,
time: 120913101}

~150 Million
~70 GB



~150 Million
~7 GB

Why did I use this title ?

Where art thou big data ?

It is not the size – 7GB

A typical Machine learning approach

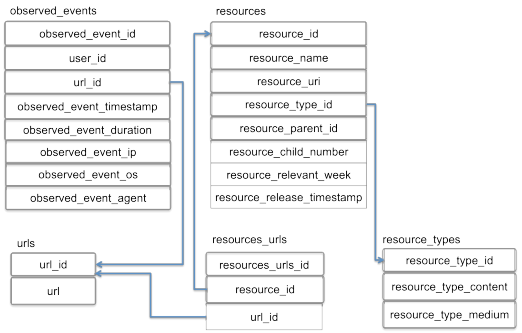
Raw Data

Cleaning
Conditioning
Organizing

JSON lines


```
{
   "status": "OK",
   "data": {
      "trends": {
        "uv": [
          {"date": "200906", "value": 90714948},
          {"date": "200907", "value": 98292793},
          {"date": "200908", "value": 103509116},
          ...
        ]
      },
      "trends_low_sample": false,
      "query_cost": 13,
      "trends_frequency": "monthly"
   }
}
```


MOOCdb



Feature Engineering

Goal: We want to predict who is going to drop out?

What could possibly predict a drop out?

- Maybe number of tries on problems
- Maybe the time spent on the course

We extract variables for students on a weekly basis

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Feature engineering - think-posit-extract

Feature Engineering

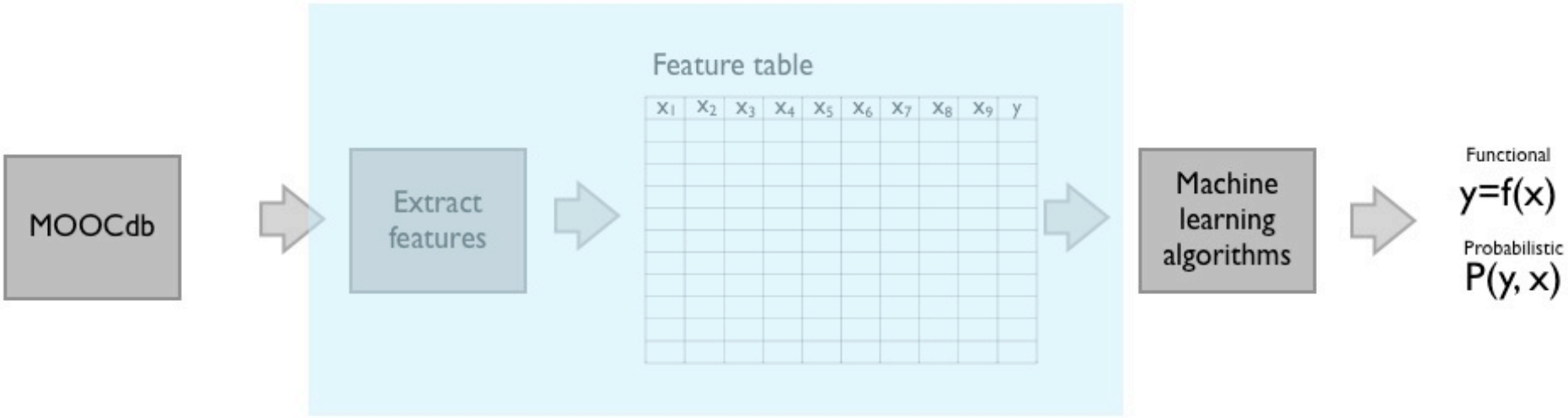
Goal: We want to predict who is going to drop out?

What could possibly predict a drop out?

- Maybe number of tries on problems
- Maybe the time spent on the course

Think

Posit



Extract

This is by-far represents the biggest challenge

ALFA
ANYSCALE LEARNING FOR ALL

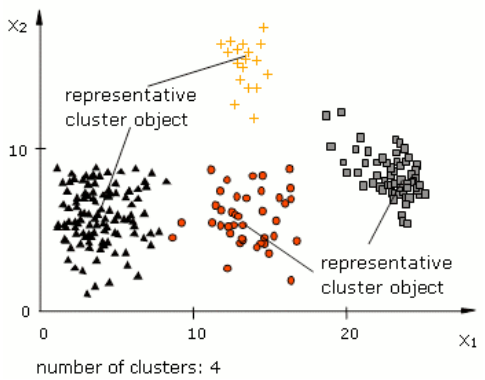
CSAIL

Supervised and Unsupervised learning given the features/variables

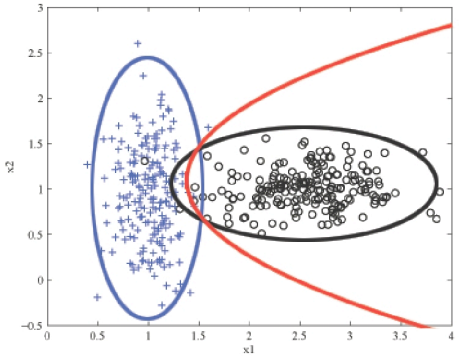
Feature table

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	y

We can change the label to be anything



Clustering the students



Classifying the students

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Types of variables

- Simple
- Complex
- Derived

Simple Variables

- Simple implies that we do some sort of counts or create aggregate for the student for a week

- Examples:
  - Time spent on the course during the week
  - Number of problem attempts made during the week
  - How many problems did the student get right ?
  - Amount of time spent on forums ?
  - Amount of time spent on videos?

Complex Variables

- Complex variables requires us to:
  - extract data at the intersection of two or more modes of student interactions
  - more complex scripting and processing
  - Curation and sometimes even manual processing
- Examples:
  - Number of times the student goes to forums while attempting problems
    - » For this, we have to go through every two consecutive attempts the student did for a problem and then between the two time stamps, we have to extract the events that correspond to his id in the data that stores his activity in forums.
  - On an average, how close to the deadline does this student start attempting problems
    - » For this we have to assemble the deadline timestamp for every problem during that week and then measure the difference between the time stamp of students first attempt and the time of his first attempt, then average over all problems.

Derived variables

- These involve:
  - combining two or more features to form a new feature
  - derive a feature based on a mathematical function
  - usually a domain expert is required or are designed based on human intuition
- Examples:
  - Ratios: ratio of number of correct submissions to total time spent in the course during that week
  - Trends: The difference in the number of attempts the student has during this week and the past week
  - Percentiles: Where does the student stand in terms of the whole class, with regards to his “total number of attempts”?
  - Many more

This is where the “big” value of this data is

How can we scale this up?

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Why did I use this title ?

Where art thou big data ?

It is in forming those complex variables
Scaling up the discovery process

So how do the ideas for the variables come about?

- **Self proposed (ALFA team came up with a few)**
- **Asked 6.MITx class**
- **Crowd sourcing**

What sort of input did we get ?
When we asked 6.MITx

- Feature idea:
  - *You can look at aggregate duration each student spent on resources. You can average the total time spent over all the students, and the further away from average they are, more likely to drop out, maybe?*
- Why this is useful?
  - *If one spends too less or too much time on resources, he might be not interested enough or struggling with the materials*

Properties of this feature:

Intuitive, combined and derived.

How can you participate?

- Please go to

  featurefactory.csail.mit.edu

1

Think and propose

2

Comment

3

Help us extract
by writing scripts

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

What did we assemble as variables so far?

Simple

Total time spent on the course
number of forum posts
number of wiki edits
average length of forum posts (words)
number of distinct problems attempted
number of submissions (includes all attempts)
number of distinct problems correct
average number of attempts
number of collaborations
max observed event duration
number of correct submissions

Complex

average time to solve problem
observed event variance (regularity)
total time spent on lecture
total time spent on book
total time spent on wiki
Number of forum responses
predeadline submission time (average)

Derived

attempts percentile
pset grade (approximate)
pset grade over time
lab grade
lab grade over time
time spent on the course per-correct-problem
attempts per correct problems
percent submissions correct

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

What did we assemble as variables so far?

Simple

Total time spent on the course
number of forum posts
number of wiki edits
average length of forum posts (words)
number of distinct problems attempted
number of submissions (includes all attempts)
number of distinct problems correct
average number of attempts
number of collaborations
max observed event duration
number of correct submissions

Complex

average time to solve problem
observed event variance (regularity)
total time spent on lecture
total time spent on book
total time spent on wiki
Number of forum responses
predeadline submission time (average)

Derived

attempts percentile
pset grade (approximate)
pset grade over time
lab grade
lab grade over time
time spent on the course per-correct-problem
attempts per correct problems
percent submissions correct

Predicting stopout

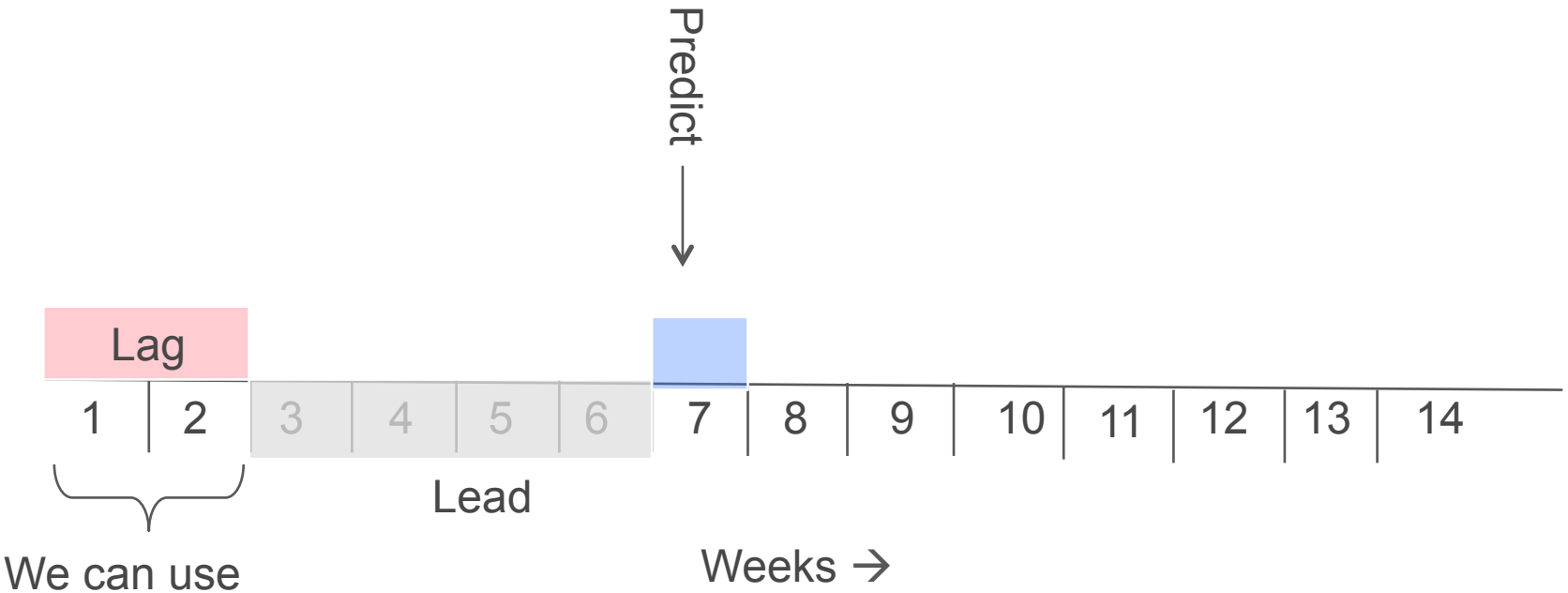
- Why predict Stopout?
  - Design of interventions/prevention
  - Identifying the reasons for Stopout
  - What does being able to predict Stopout for a student in advance tell us about the students intentions?
  - Logistics
- Definitions of Stopout
  - When student stops attempting problems
  - When student stops coming to the website (no trace of his/her activity)

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Defining the prediction problem

Given current student behavior if s/he will Stopout in the future?

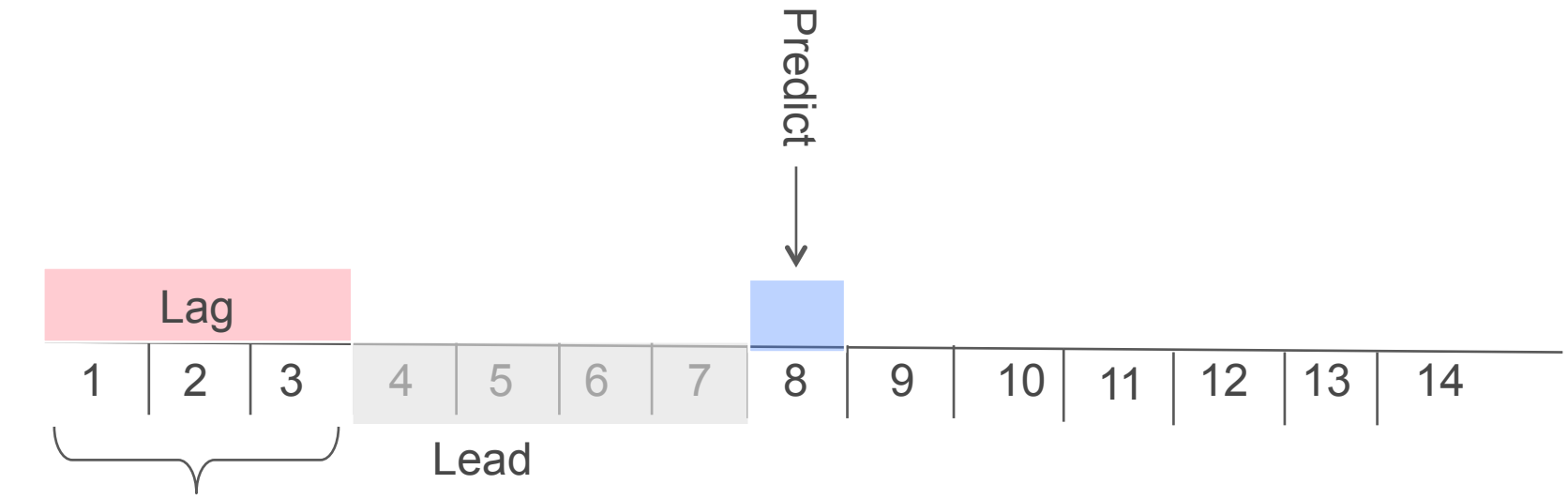


ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Defining the prediction problem

Given current student behavior if s/he will Stopout in the future?



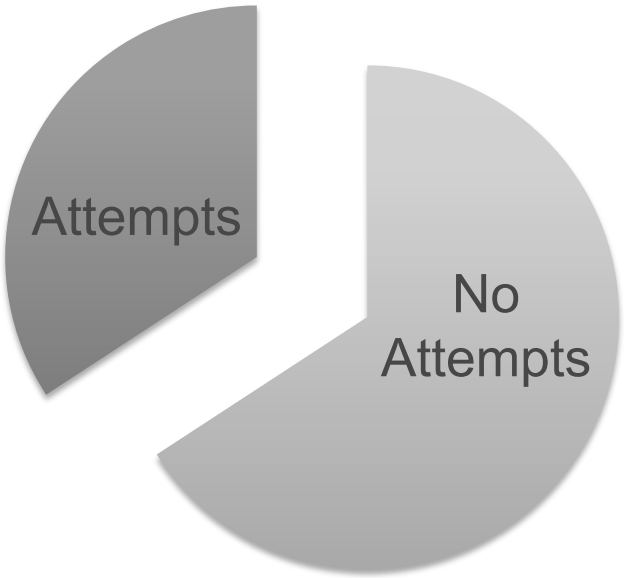
We can use students data during these weeks

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

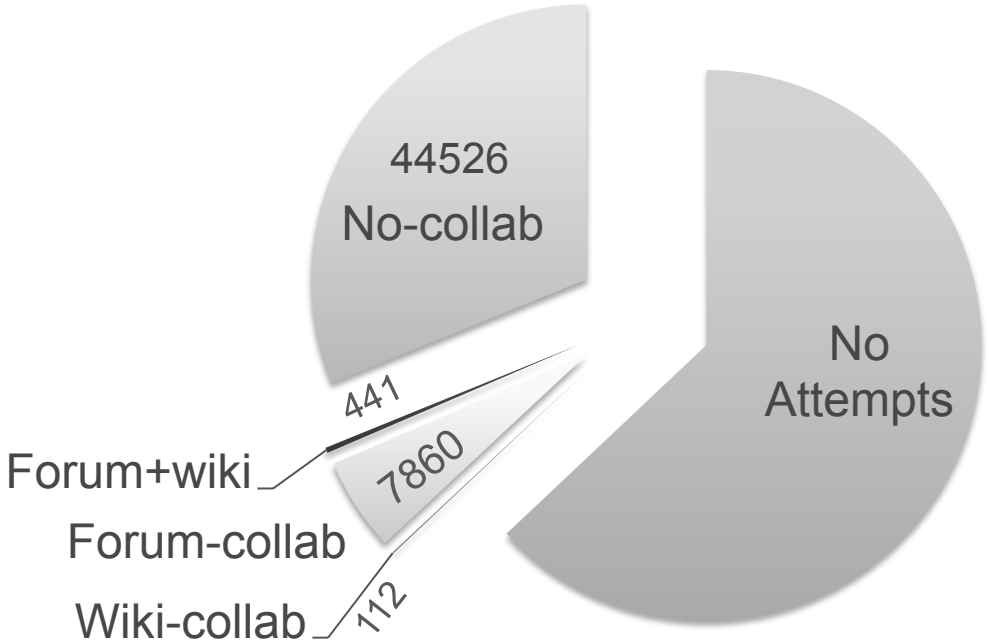
Our cohort

Splitting by attempts



Our cohort

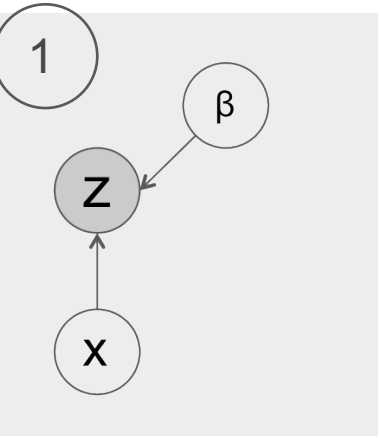
Further splitting by activity



ALFA
ANYSCALE LEARNING FOR ALL

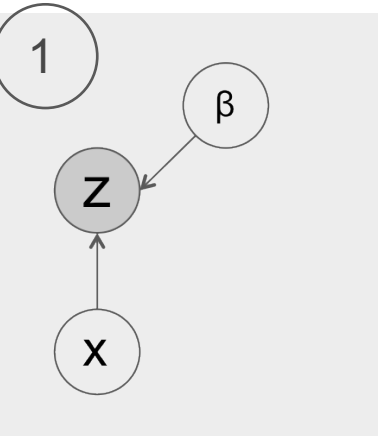
CSAIL

Models we build

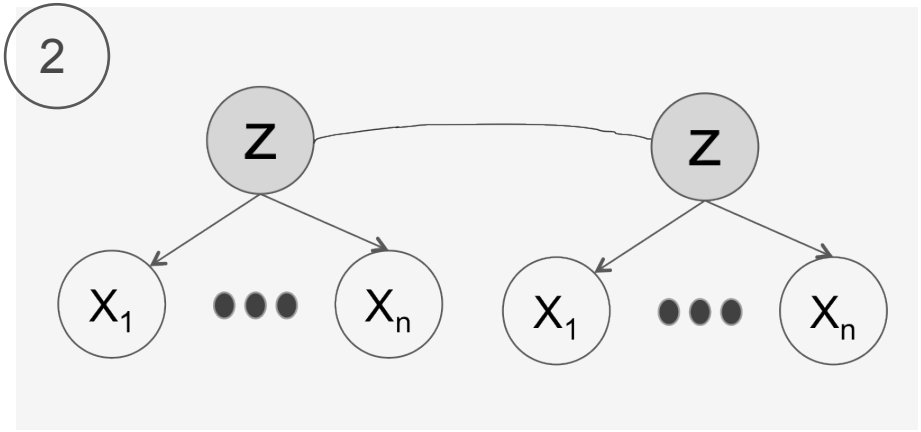


Logistic regression

Models we build

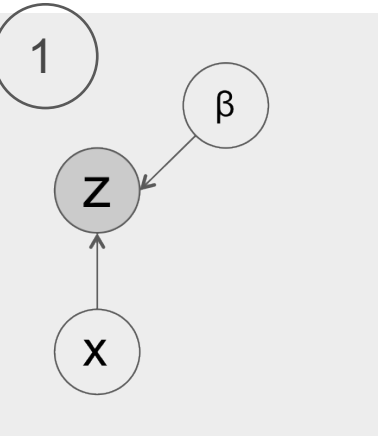


Logistic regression

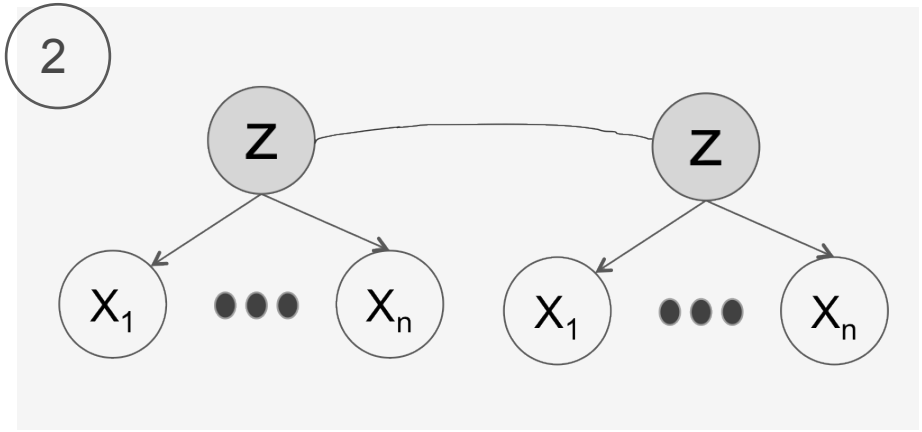


Hidden Markov Models

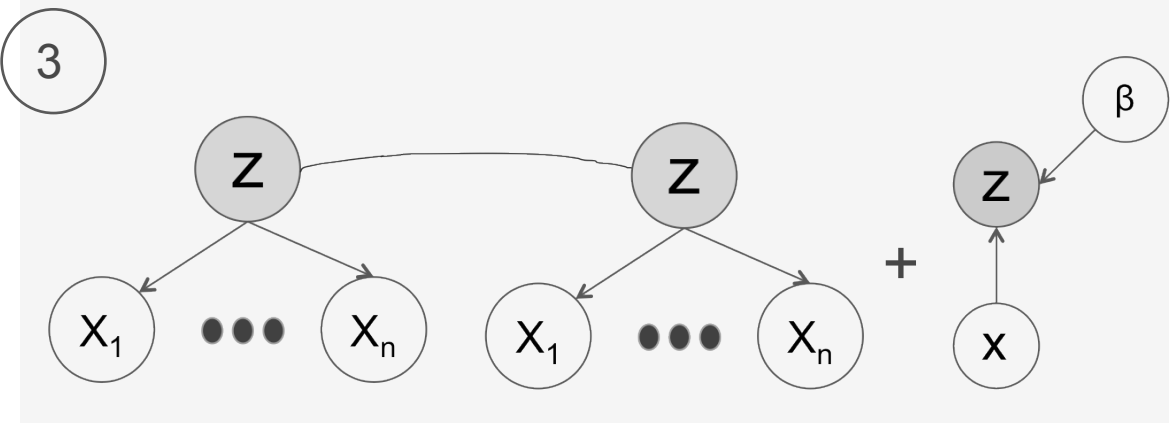
Models we build



Logistic regression



Hidden Markov Models

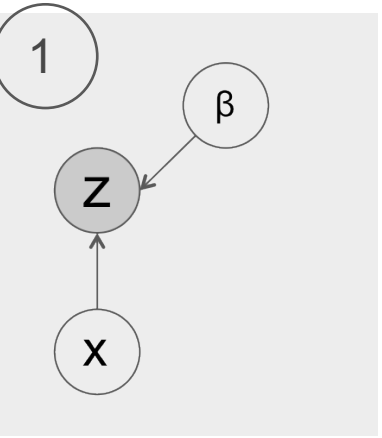


Hidden Markov Models + Logistic regression

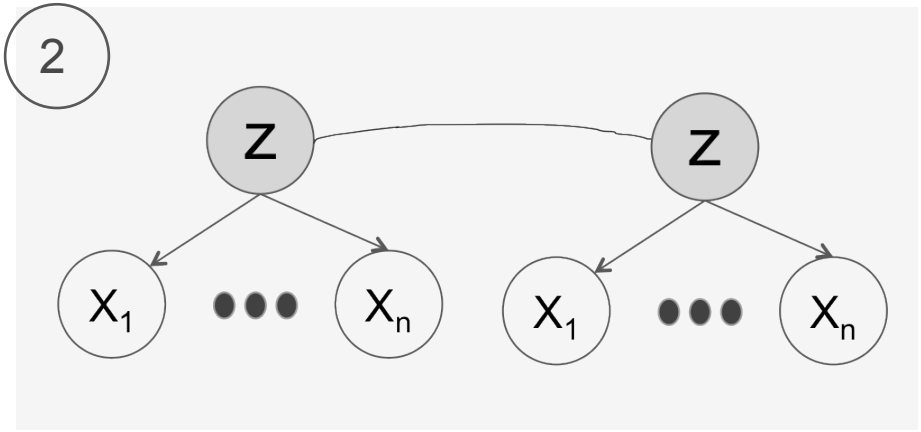
ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

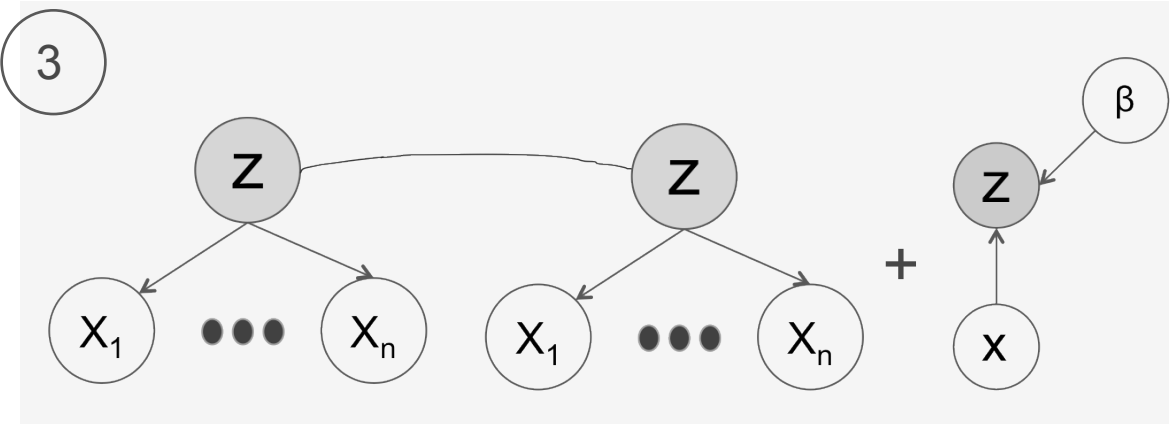
Models we build



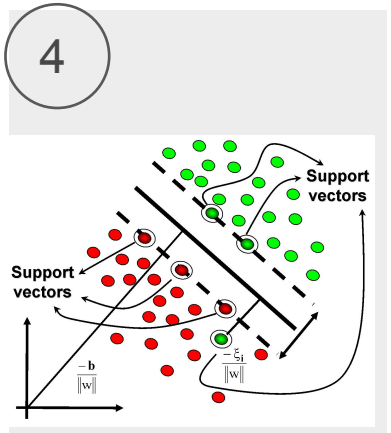
Logistic regression



Hidden Markov Models



Hidden Markov Models + Logistic regression

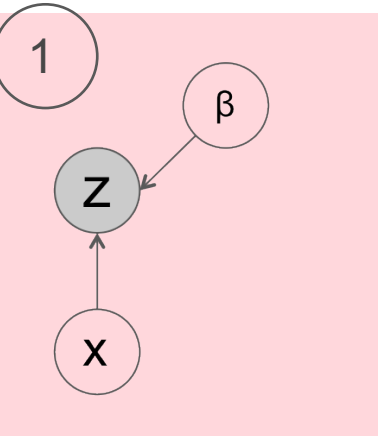


Discriminative models

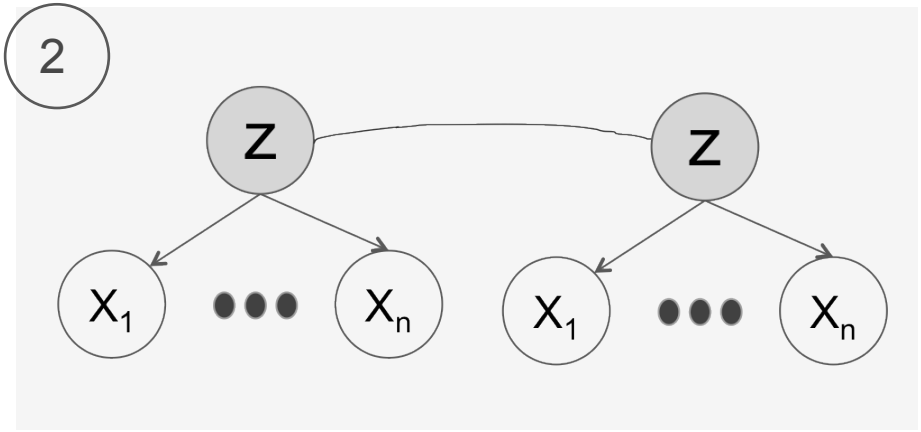
ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Models we build

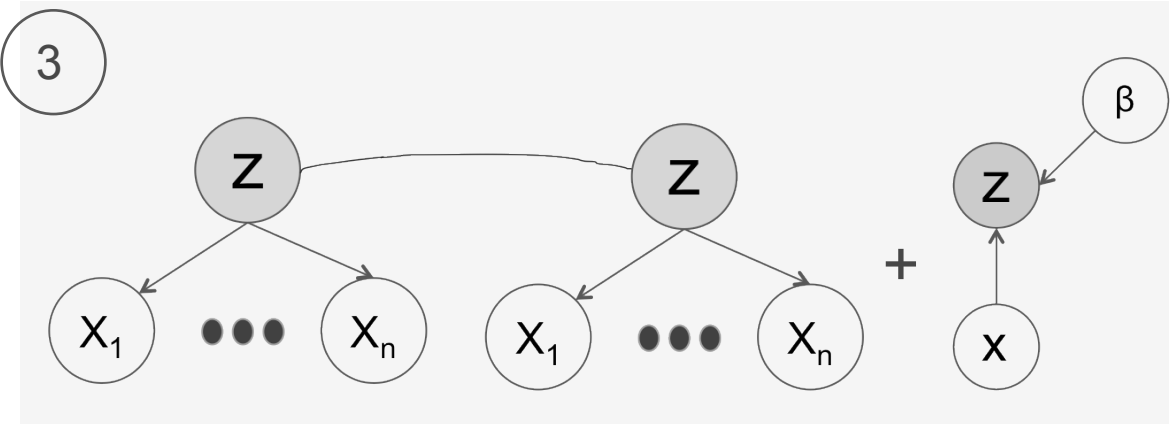


Logistic regression

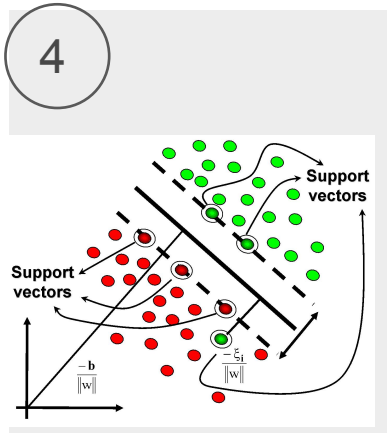


Hidden Markov Models

All results are presented
In Colin Taylors thesis



Hidden Markov Models + Logistic regression

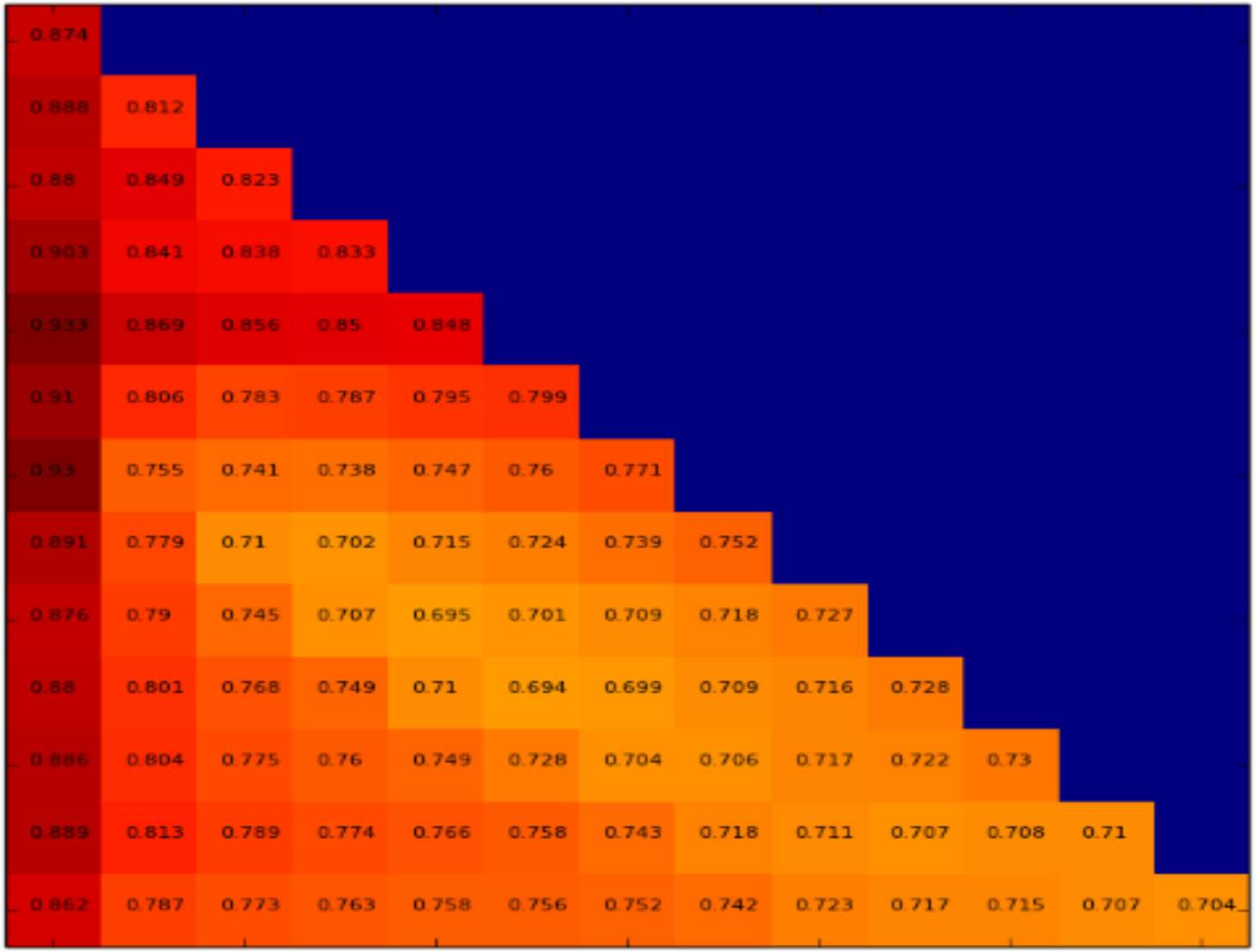


Discriminative models

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Results for students who do not collaborate



ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Defining the prediction problem

An extreme version of the problem

AUC – 0.71

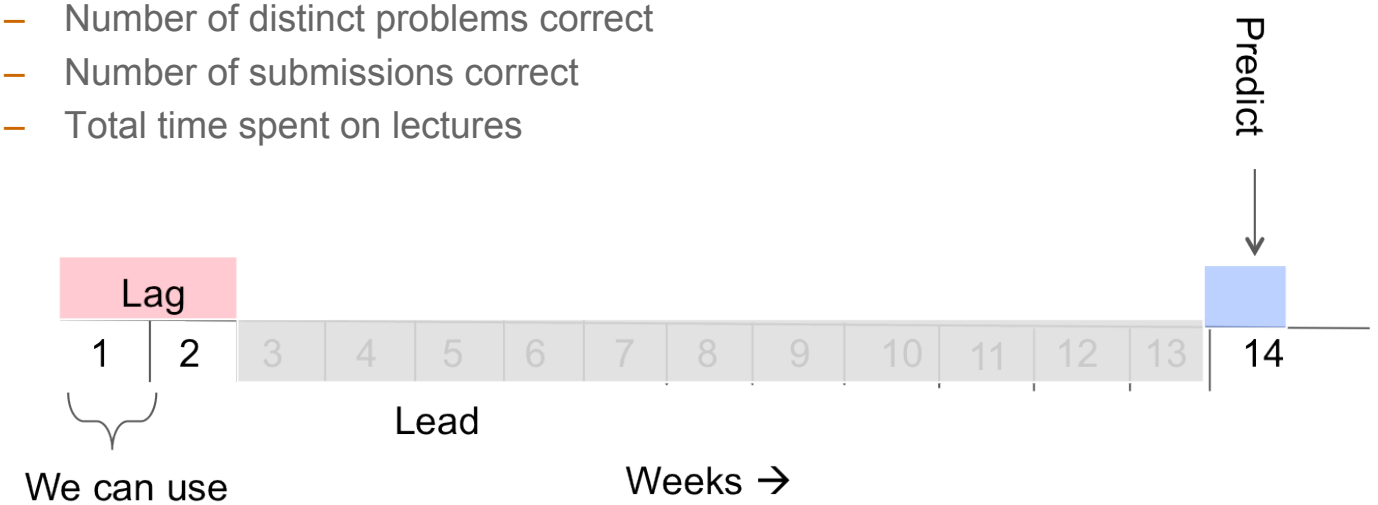
We can use students data during these weeks

ALFA
ANYSCALE LEARNING FOR ALL

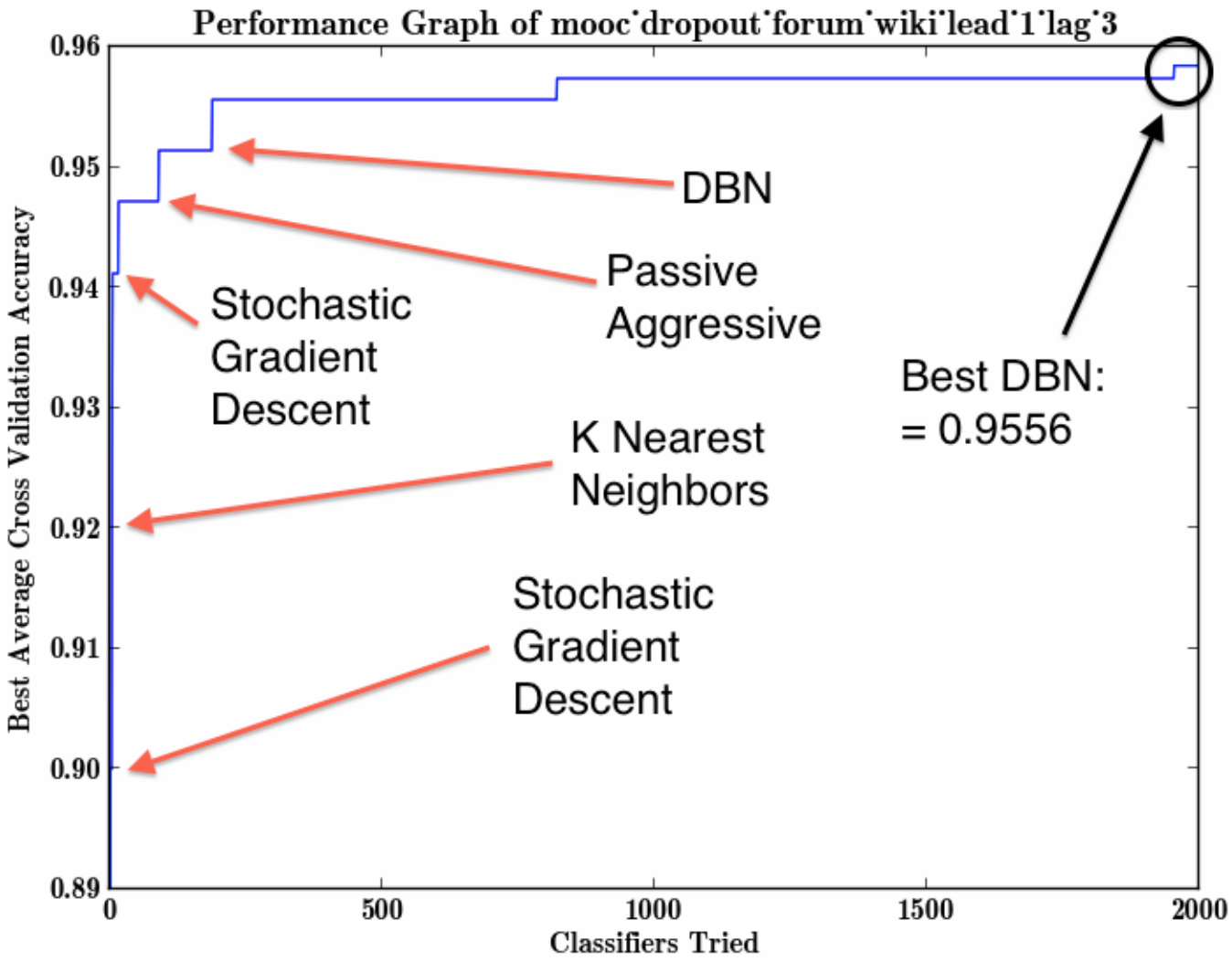
CSAIL

Top 10 features/variables that mattered

- We perform Randomized Logistic Regression to identify the variable importance
- Week 1
  - Number of distinct problems correct
  - Predeadline submission time
  - number of submissions correct
- Week 2
  - Lab grade
  - Attempts per correct problem
  - Predeadline submission time
  - Attempts percentile
  - Number of distinct problems correct
  - Number of submissions correct
  - Total time spent on lectures



Discriminatory model building on the cloud

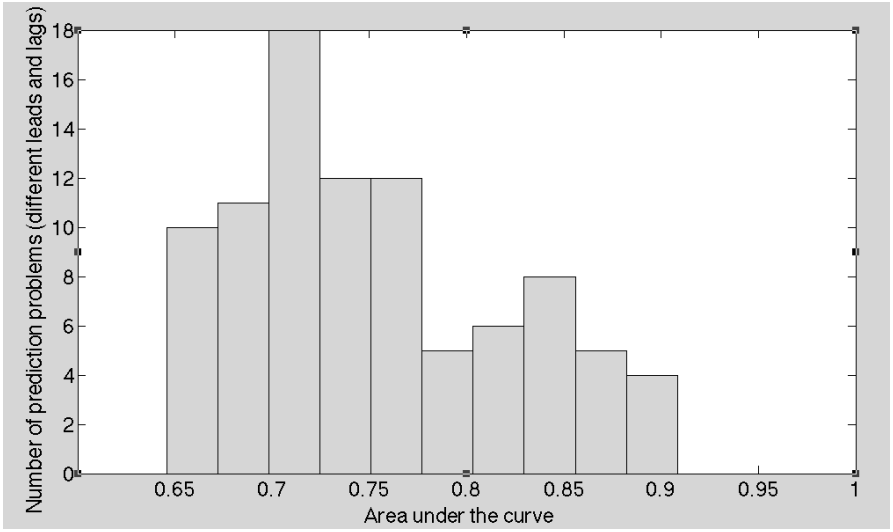


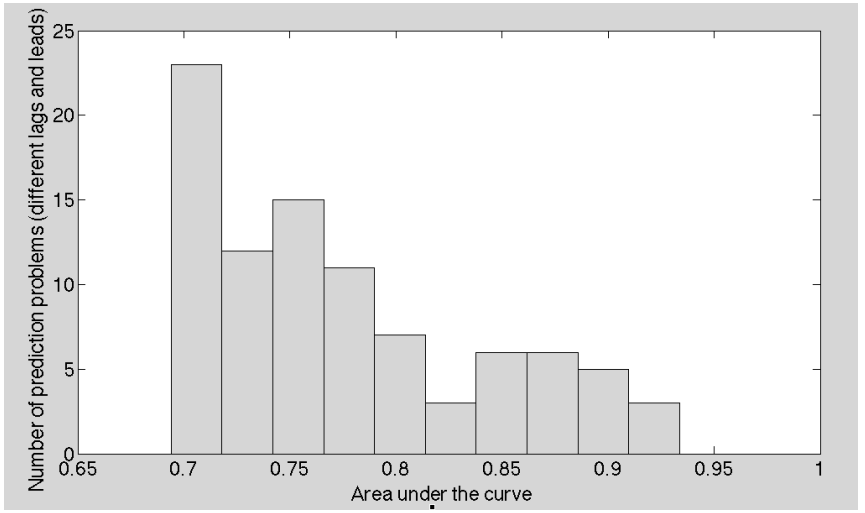
Please visit delphi.csail.mit.edu
if you would like to try this system

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

With and without crowd proposed features





ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Conclusions

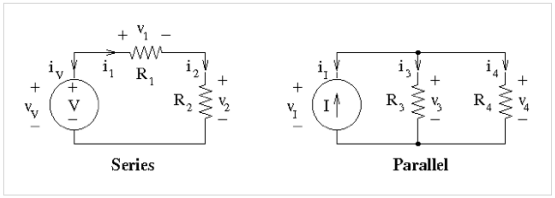
- We have to scale up feature engineering
- A lot of value in this data can only be derived
  - By defining complex, derived features
  - A number of scalable machine learning approaches

The digital student

Problem

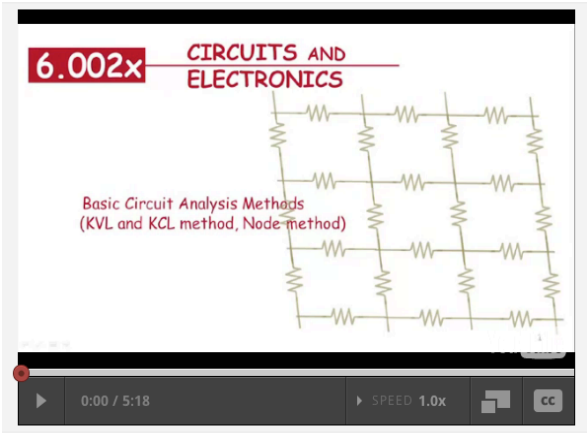
H1P2: DUALITY

In this problem we will investigate a fun idea called "duality."
Consider the series circuit in the diagram shown.

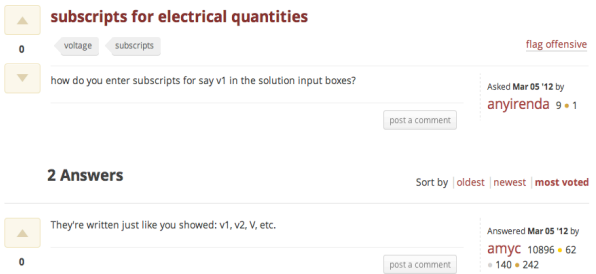


We are given device parameters $V = 15.3\text{v}$, $R_1 = 4.59\Omega$, and $R_2 = 3.06\Omega$. All of the unknown voltages and currents are labeled in associated reference directions. Solve this circuit for the unknowns and enter them into the boxes given.
The value (in Volts) of v_1 is:

Lecture video



Forum

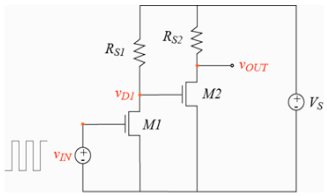


Wiki

Speed of Digital Circuits

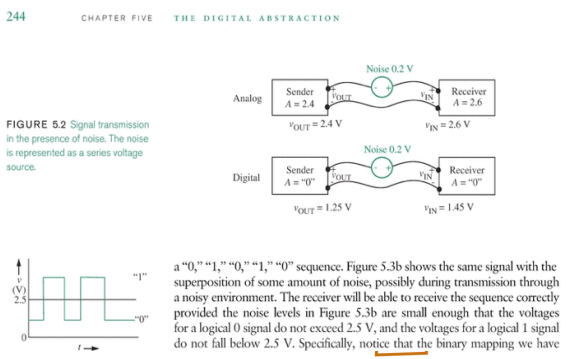
Why is the response of an inverter not instantaneous?

Let's consider two inverters driven by a square wave input:



Instead of ideal square waveforms, the actual waveforms in a short time scale looks like:

Book



ALFA
ANYSCALE LEARNING FOR ALL

CSAIL