

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 Deroncourt, Una-May O'Reilly

Any Scale Learning for All Group
CSAIL, MIT

Why did I use this title ?

Where art thou big data ?

Overview

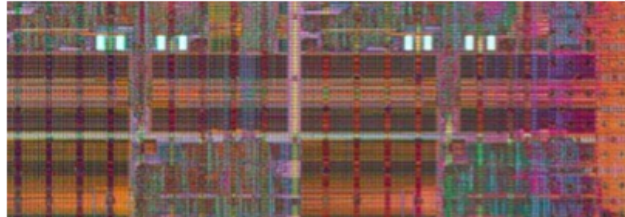
- Data collected in MOOCs
- Dataset we use
 - Properties
 - Size
 - Organization
- Challenges of Feature/Variable engineering
 - What does it entail?
 - How we are scaling it up? (or would like to)
- Stopout prediction problem
 - Definition
 - Machine learning models
 - How features/variables we extracted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

6.002x: Circuits and Electronics



HOW IT WORKS COURSES SCHOOLS

dashboard



Circuits and Electronics

Teaches the fundamentals of circuit and electronic analysis.

About this Course

The course introduces engineering in the context of the lumped circuit abstraction. Topics covered include: resistive elements and networks; independent and dependent sources; switches and MOS transistors; digital abstraction; amplifiers; energy storage elements; dynamics of first- and second-order networks; design in the time and frequency domains; and analog and digital circuits and applications. Design and lab exercises are also significant components of the course.

The course is organized by weeks. To keep pace with the class, you are expected to complete all the work by the due dates indicated. Homeworks and labs must be completed by the Sunday of the week following the one in which they are posted. Weekly coursework includes interactive video sequences, readings from the textbook, homework, online laboratories, and optional tutorials. The course will also have a midterm exam and a final exam. Those who successfully earn enough points will receive an honor code certificate from MITx.

Note - You can earn college credit for taking and passing this course under the ID Verified option. EdX charges no additional cost for this. Learn more on about receiving credit from [The American Council on Education's College Credit Recommendation Service \(ACE\)](#) in the FAQ below. For more information on obtaining your transcript from ACE click [here](#).

WAYS TO TAKE THIS EDX COURSE:



School:	MITx
Course Code:	6.002x
Classes Start:	16 Oct 2013
Course Length:	17 weeks
Estimated effort:	12 hours/week.

Prerequisites:

In order to succeed in this course, you must have taken an AP level physics course in electricity and magnetism. You must know basic calculus and linear algebra and have some background in...

[see more...](#)

Access Courseware



Lectures (Inductors, 1st order circuits)

Video sequences

In lecture quiz

► Overview

► Math Review

► Week 1

► Week 2

► Week 3

► Entrance Survey

► Week 4

► Week 5

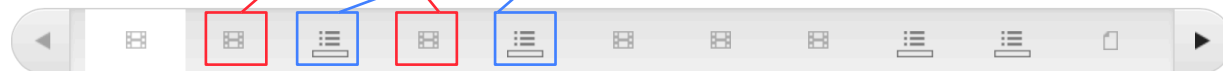
► Week 6

▼ Week 7

Inductors and First-Order Circuits

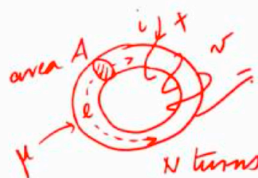
Speed of Digital Circuits

Week 7 Tutorials



S13V1: IDEAL LINEAR INDUCTORS

Ideal Linear Inductor



Text accompanying the audio

PROFESSOR: We will now look at another energy storage device called the inductor.

The capacitor and inductor are both energy storage devices.

Unlike resistors, which dissipate energy, inductors

and capacitors store energy.

They do not dissipate energy.

So if you pump energy into these devices, rather than

Download video

Textbook

Each page is uniquely identified by a URL

 MITx: 6.002x Circuits and Electronics

ALFA_EVO

Courseware | Course Info | Discussion | Wiki | FAQ | **Textbook** | Progress | About | Open Ended Panel

Contents

+ Preamble

+ 1. The Circuit Abstraction

+ 2. Resistive Networks

+ 3. Network Theorems

+ 4. Analysis of Nonlinear Circuits

+ 5. The Digital Abstraction

+ 6. The MOSFET Switch

+ 7. The MOSFET Amplifier

+ 8. The Small Signal Model

+ 9. Energy Storage Elements

+ 10. First-order Transients

+ 11. Energy and Power in Digital Circuits

+ 12. Transients in Second Order Circuits

+ 13. Sinusoidal Steady State

+ 14. Sinusoidal Steady State: Resonance

+ 15. The Operational Amplifier Abstraction

+ 16. Diodes

+ A1. Maxwell's Equations and the LMD

+ B. Trigonometric Functions & Identities

+ C. Complex Numbers

D. Solving Simultaneous Linear Equations

Answers to Selected Problems

Figure Acknowledgments

Index

CONTENTS

Material marked with **www** appears on the Internet (please see Preface for details).

Preface xvii

 Approach xvii

 Overview xix

 Course Organization xx

 Acknowledgments xxi

CHAPTER 1 The Circuit Abstraction 3

 1.1 The Power of Abstraction 3

 1.2 The Lumped Circuit Abstraction 5

 1.3 The Lumped Matter Discipline 9

 1.4 Limitations of the Lumped Circuit Abstraction 13

 1.5 Practical Two-Terminal Elements 15

 1.5.1 Batteries 16

 1.5.2 Linear Resistors 18

 1.5.3 Associated Variables Convention 25

 1.6 Ideal Two-Terminal Elements 29

 1.6.1 Ideal Voltage Sources, Wires, and Resistors 30

 1.6.2 Element Laws 32

 1.6.3 The Current Source — Another Ideal Two-Terminal Element 33

 1.7 Modeling Physical Elements 36

 1.8 Signal Representation 40

 1.8.1 Analog Signals 41

 1.8.2 Digital Signals — Value Discretization 43

 1.9 Summary and Exercises 46

CHAPTER 2 Resistive Networks 53

 2.1 Terminology 54

 2.2 Kirchhoff's Laws 55

 2.2.1 KCL 56

 2.2.2 KVL 60

 2.3 Circuit Analysis: Basic Method 66

 2.3.1 Single-Resistor Circuits 67

 2.3.2 Quick Intuitive Analysis of Single-Resistor Circuits 70

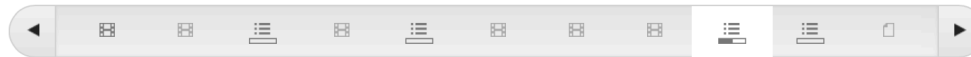
 2.3.3 Energy Conservation 71

 ANYSCALE LEARNING FOR ALL



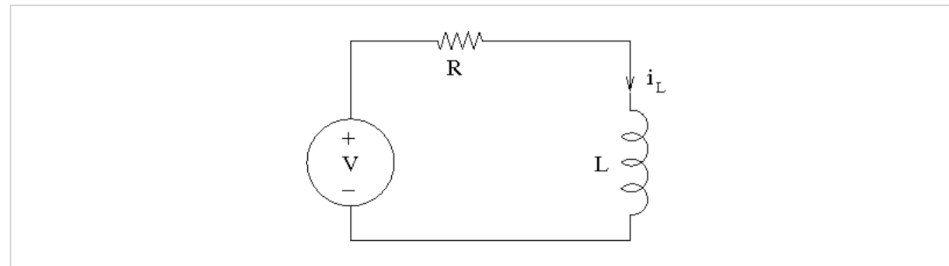
Problem Solution Entry/Retry

- Overview
- Math Review
- Week 1
- Week 2
- Week 3
- Entrance Survey
- Week 4
- Week 5
- Week 6
- Week 7
 - Inductors and First-Order Circuits
 - Speed of Digital Circuits
 - Week 7 Tutorials
 - Homework 7
 - Homework due Dec 17, 2013 at 23:59 UTC
 - Lab 7
 - Lab due Dec 17, 2013 at 23:59 UTC
- Fall 2012 Midterm Exam



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 $\frac{di_L}{dt} =$

$(V - R \cdot i_L) / L$



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](#)

Discussion board

Courseware

Course Info

Discussion

Wiki

FAQ

Textbook

Progress

About

Open Ended Panel

For New Post

FD 15

Show All Discussions ▾

🔍

SORT BY:

DATE

VOTES

COMMENTS

[Hints]H2P1 Current Divider[Hints]

+21

10

Make your forum posts more readable! Using Mathjax/LaTeX (helpful links included)

+18

9

[INTRO]Use this thread for all introductions v1[INTRO]

✓ +16

84

This course is so Addictive!

+12

2

I'm Resolving Circuits!!!!

+11

2

Bug about first blank

+10

11

Algebraic expression...

✓ +10

3

Student Tutorial: Sign Conventions

+9

3

[Staff post] H1P2 modified

+9

14

Rotate?

✓ +9

4

Hydraulic analogy

+8

2

This course is really serious thing! I'm so pleased

+8

0

THE PICKLE TURNED RED!!

✓ +8

3

Study Room

+8

3

1st post!!! YESSSSSSS!!!

+8

12

[Staff][Midterm practice exam] Duration and scoring

+7

5

Explanation question 3

+7

3

"Clever girl..."

+7

0

Average power

✓ +7

8

[STAFF POST] IMPORTANT MIDTERM INFO

+3

9

Average power

turnavies

3 months ago

+ 7

Could someone briefly explain to me how to calculate the average power in this example?

Thanks, Thomas

(this post is about [Week 1 / AC power](#))

☐ Report Misuse

COMMUNITY TA

ashwith

3 months ago

✓ + 7

The average of a set of numbers is defined as the sum of those numbers divide by how many numbers you have,

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

When you're dealing with functions, that summation turns into an integral. Are you familiar with calculus (It is a prerequisite for the course)?

The 'x' in this case is the power at each point of time (the instantaneous power). You need to find its average. See if you can figure out what the integral should be on your own based on what I said. Come back if you're still stuck.

☐ Report Misuse

Thanks ashwith,

I'm sure I'm making a really stupid mistake...

I'm integrating the instantaneous power, which is $U \cdot I$ or U^2/R , thus $(120 \cdot \sqrt{2} \cdot \cos(2\pi \cdot 60t))^2 / 110$, correct?

I'll compute this integral from $t=0$ to $t=1/60$ which is one cycle, but obviously the positive and negative parts of the wave are cancelling each other out... what am I missing?

–posted 3 months ago by turnavies

☐

Think about this again. Will it still cancel? It's not a 'simple' cos function.

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

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

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: {20913101}}

{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: {20913101}}



{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: {20913101}}



observed_events

observed_event_id
user_id
url_id
observed_event_timestamp
observed_event_duration
observed_event_ip
observed_event_os
observed_event_agent

resources

resource_id
resource_name
resource_uri
resource_type_id
resource_parent_id
resource_child_number
resource_relevant_week
resource_release_timestamp

urls

url_id
url

resources_urls

resources_urls_id
resource_id
url_id

resource_types

resource_type_id
resource_type_content
resource_type_medium



CSAIL

~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



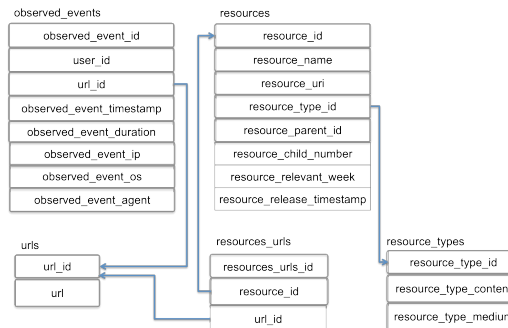
Cleaning
Conditioning
Organizing



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

Feature engineering - think-posit-extract

Feature Engineering

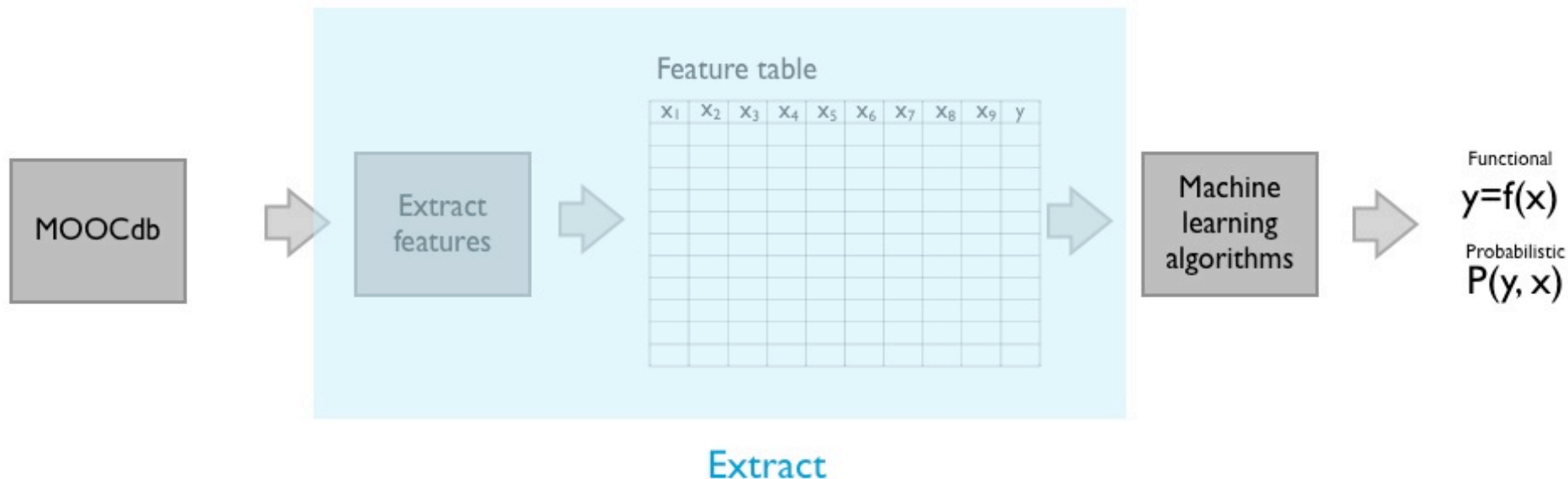
Goal: We want to predict who is going to drop out?

What could possibly predict a drop out?

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

Think

Posit



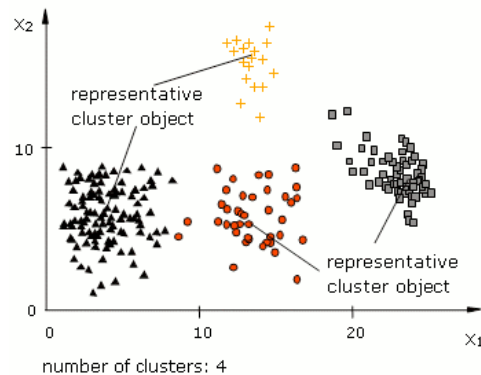
This is by-far represents the biggest challenge

Supervised and Unsupervised learning given the features/variables

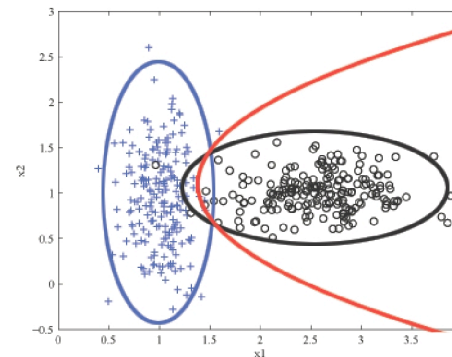
Feature table

[illegible]

We can change the label to be anything



Clustering the students



Classifying the students

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?

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

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

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

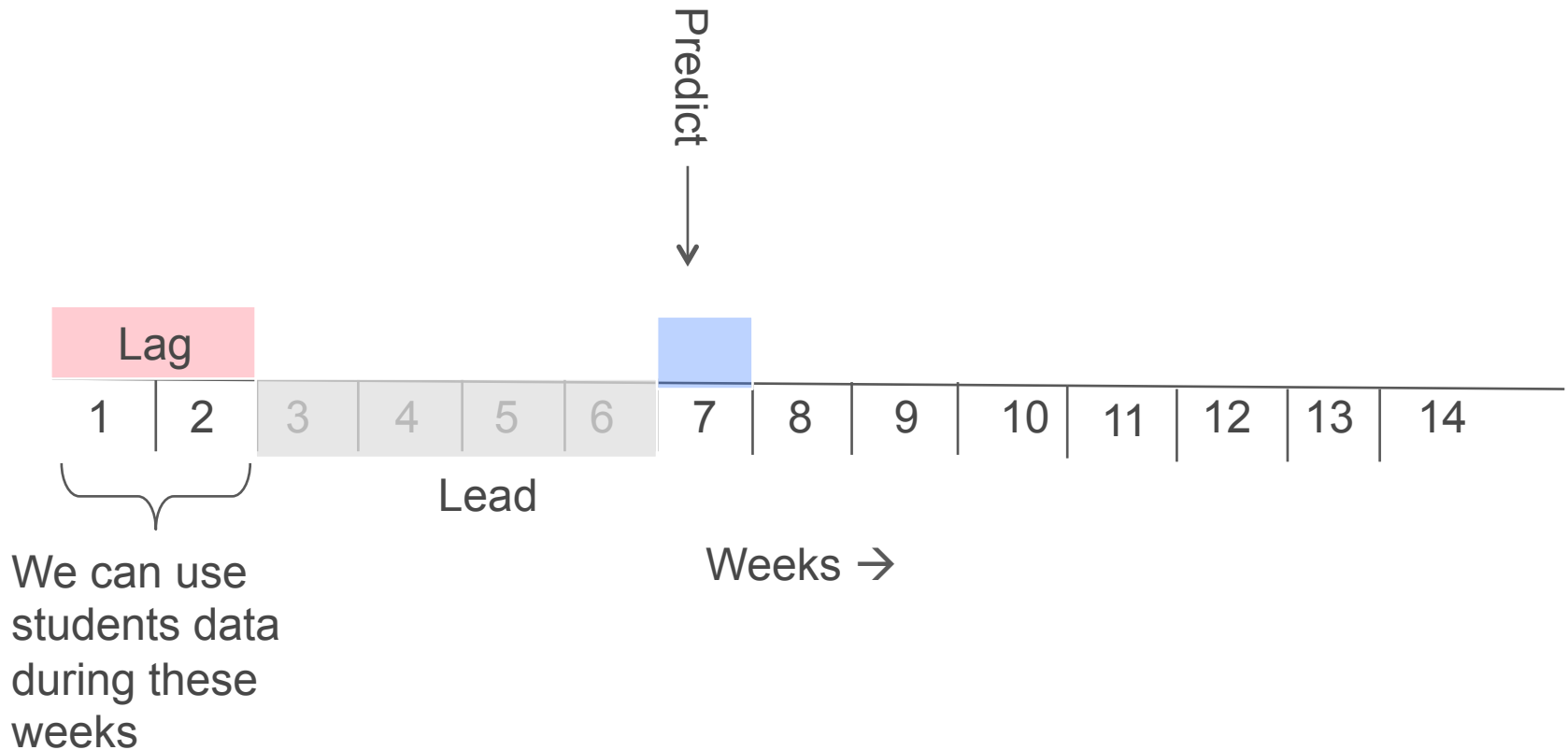
- attempts percentile
- pset grade (approximate)
- pset grade over time
- lab grade
- lab grade over time
- time spent on the course per-correct-problem
- attempts per correct problems
- percent submissions correct

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)

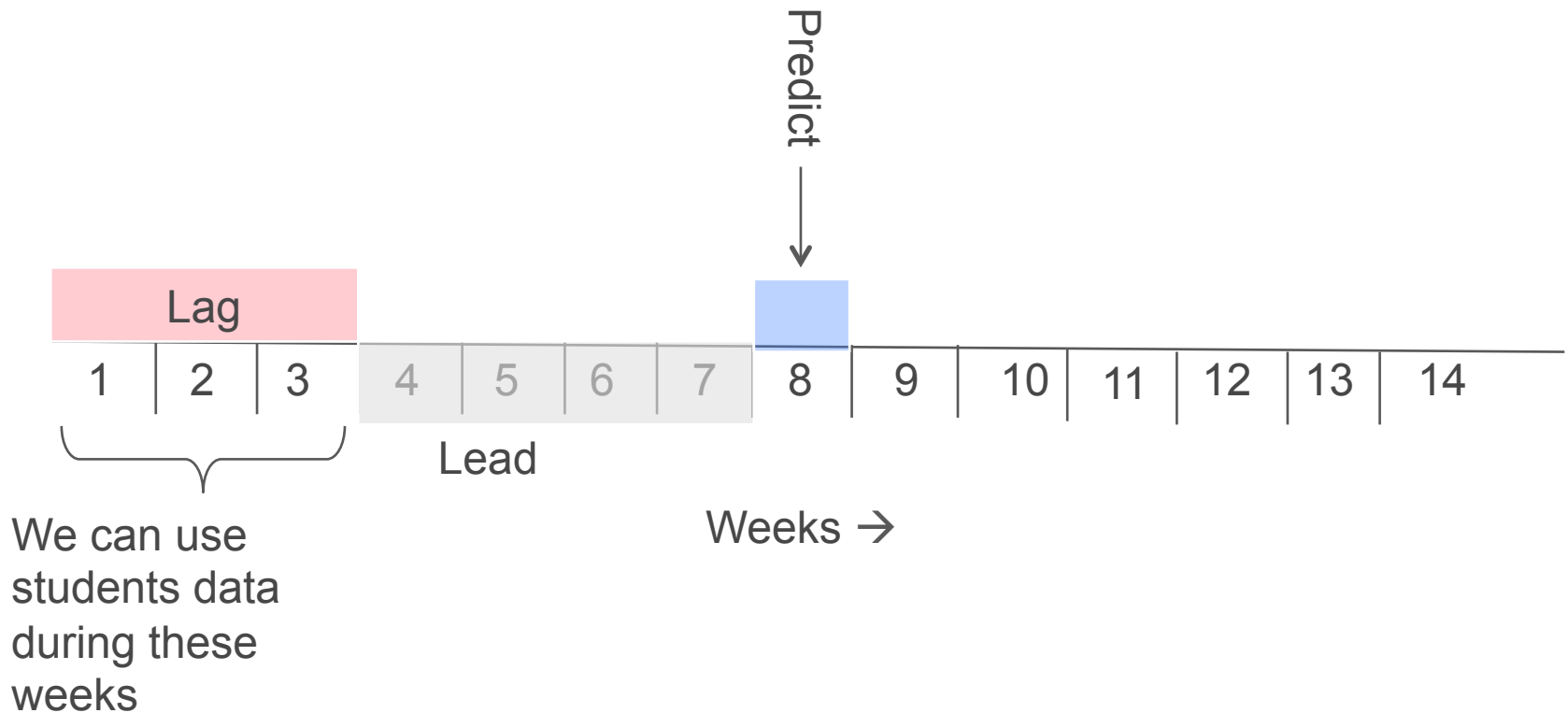
Defining the prediction problem

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



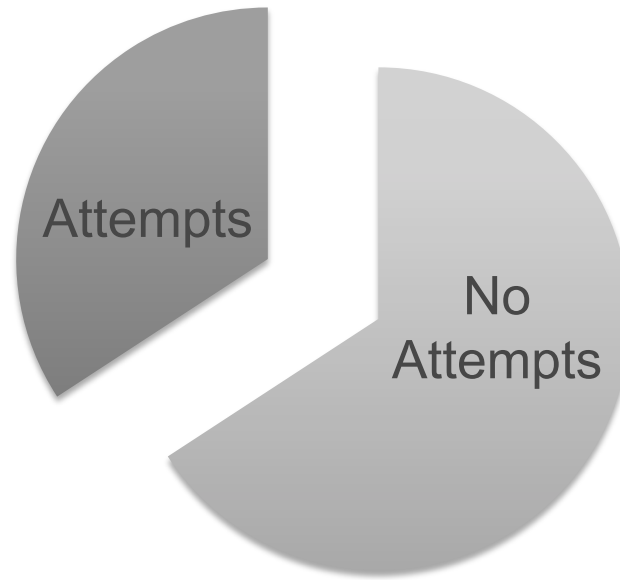
Defining the prediction problem

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



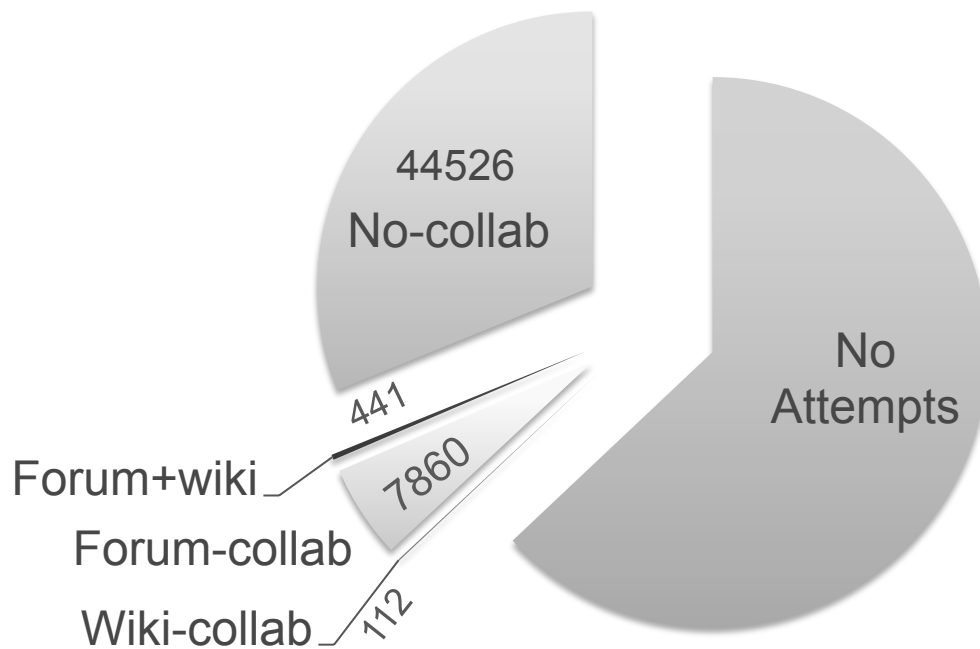
Our cohort

Splitting by attempts

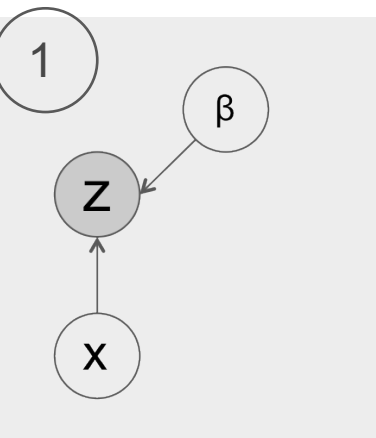


Our cohort

Further splitting by activity

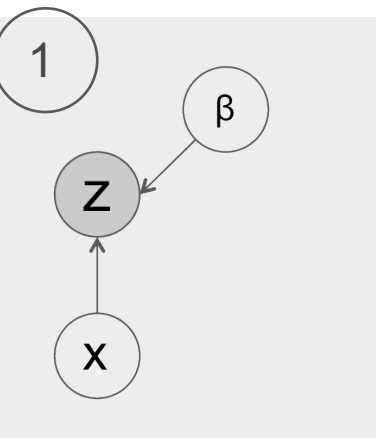


Models we build

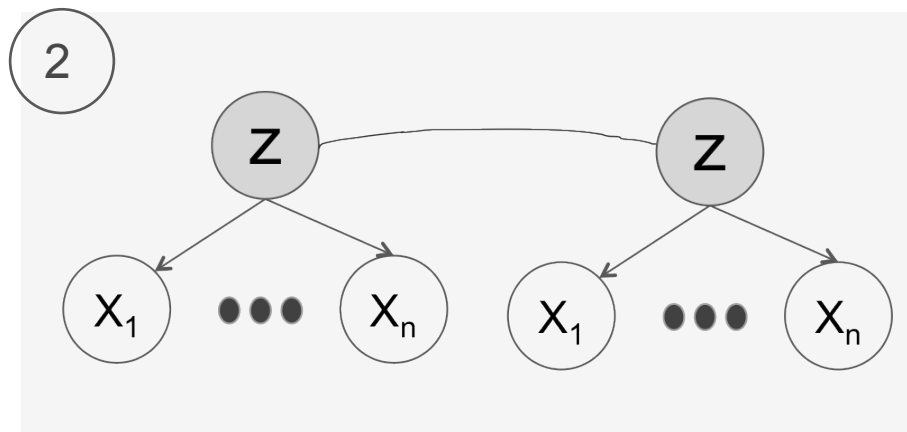


Logistic regression

Models we build

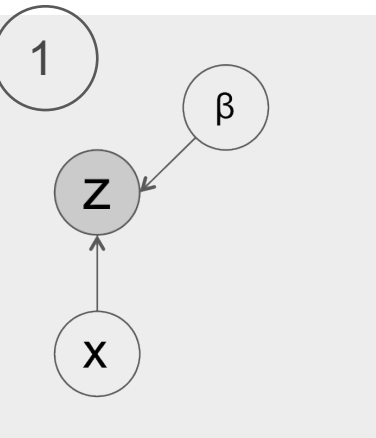


Logistic regression

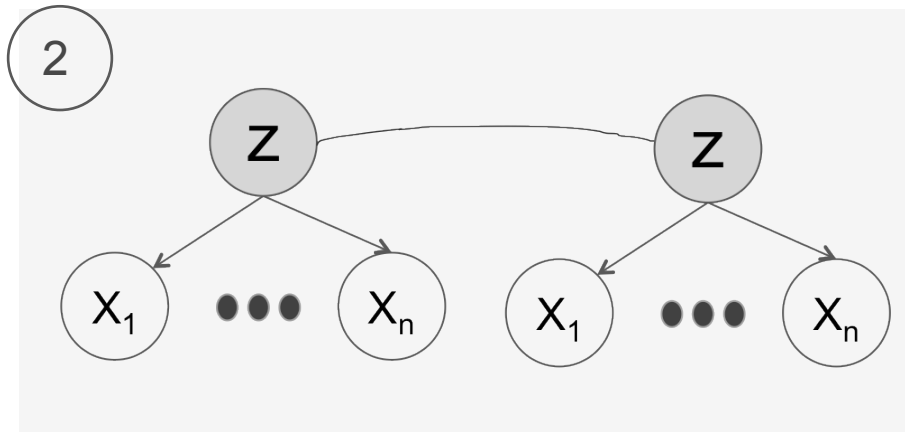


Hidden Markov Models

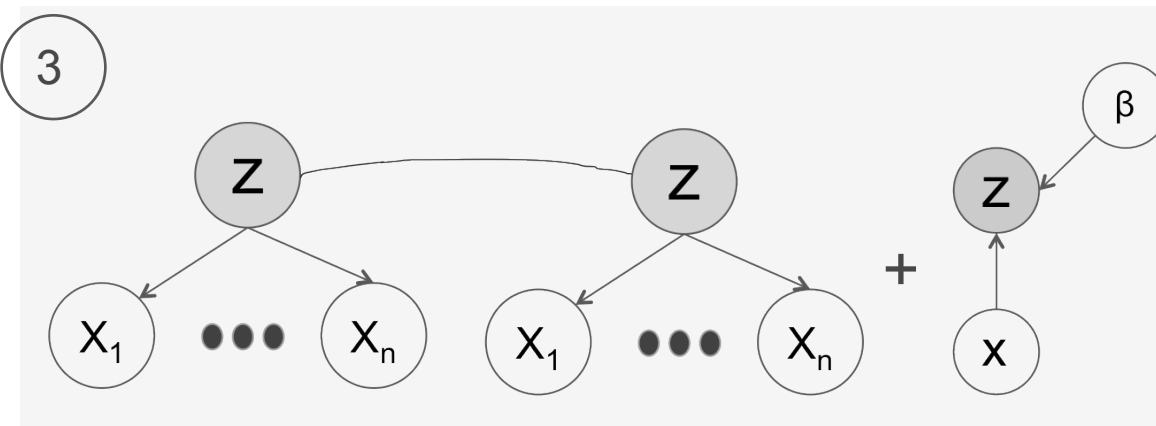
Models we build



Logistic regression

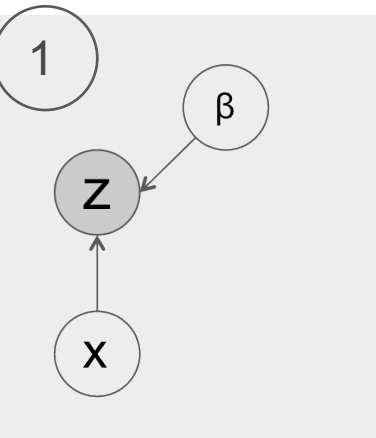


Hidden Markov Models

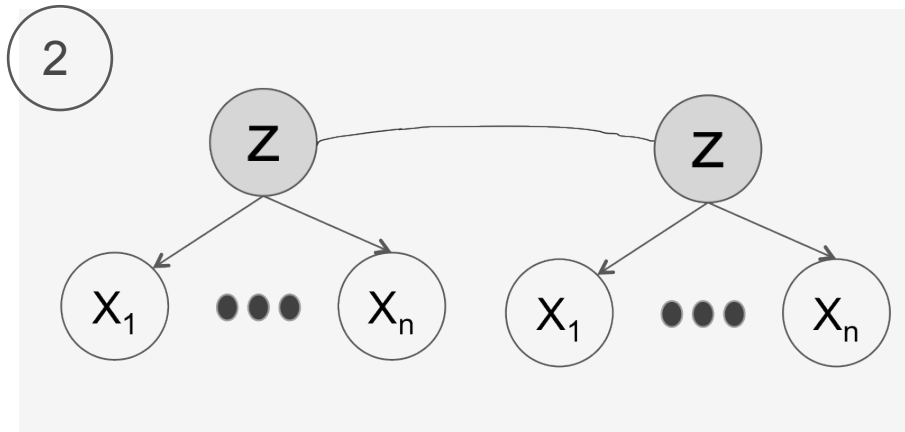


Hidden Markov Models + Logistic regression

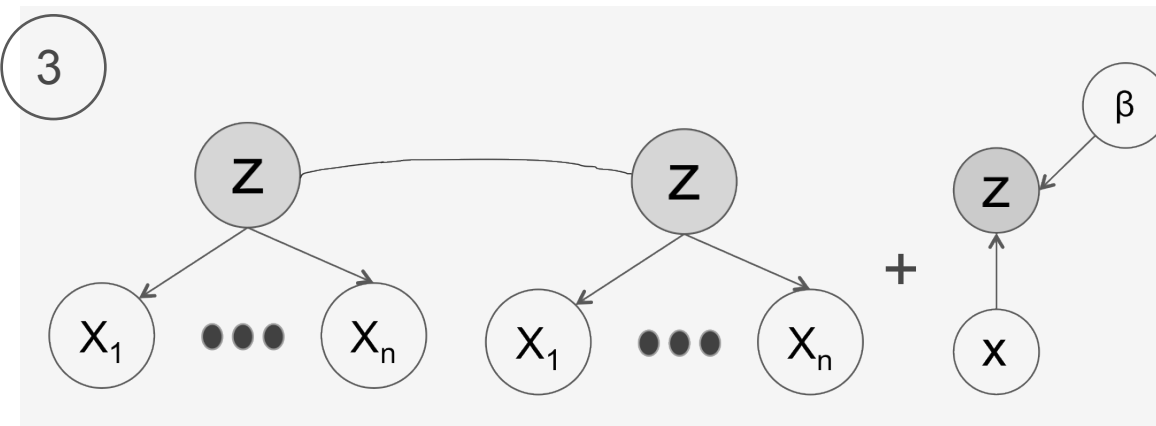
Models we build



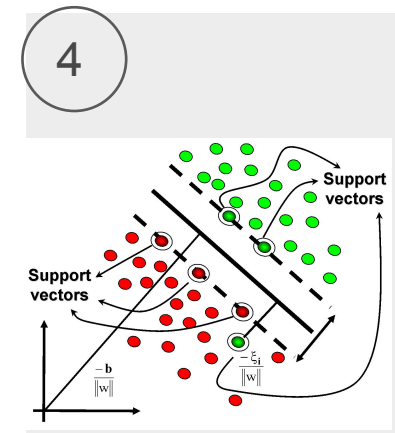
Logistic regression



Hidden Markov Models

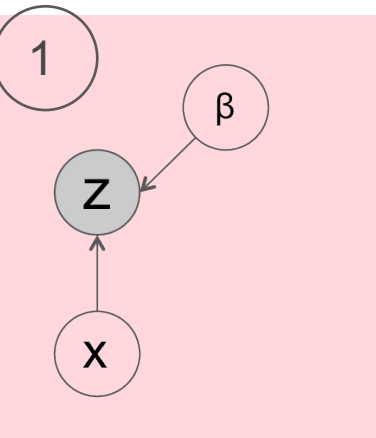


Hidden Markov Models + Logistic regression

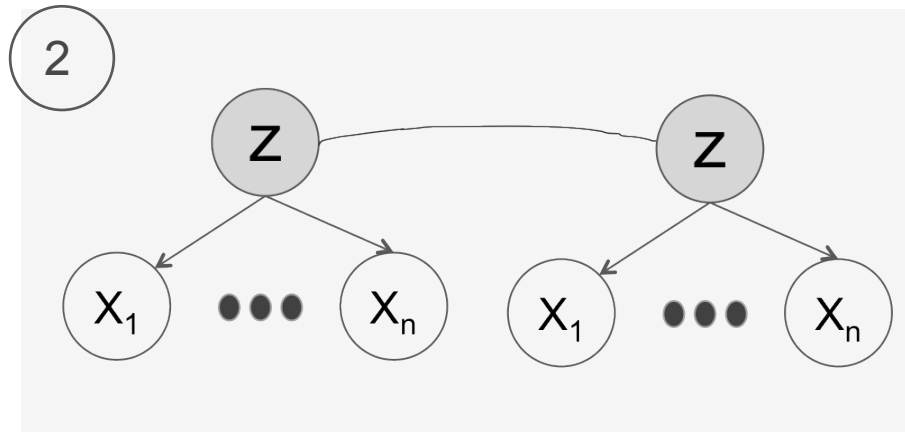


Discriminative models

Models we build

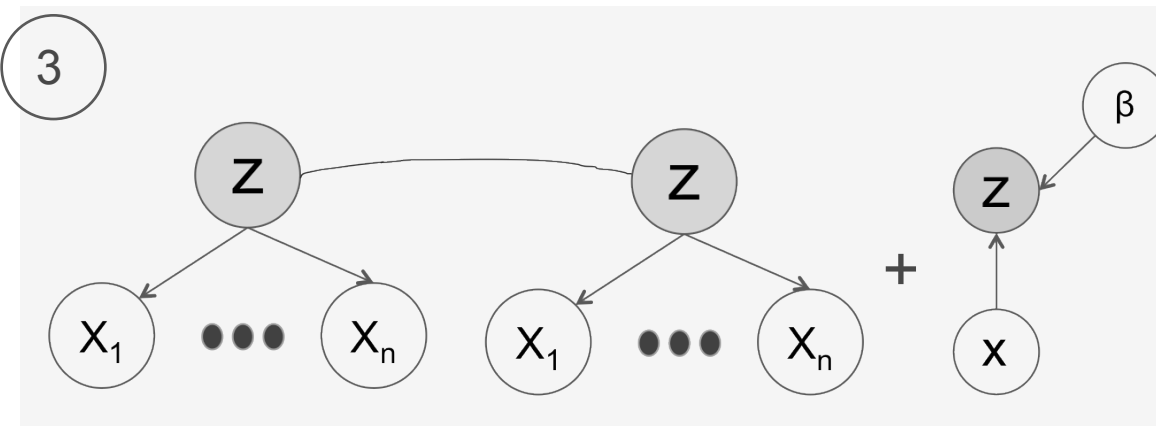


Logistic regression

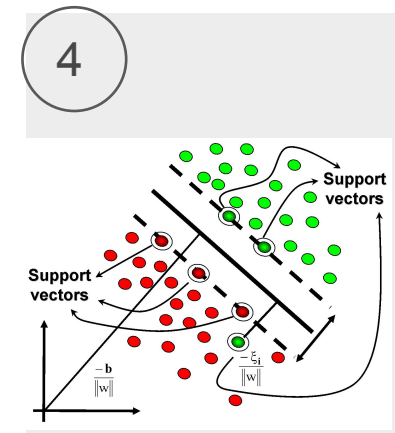


Hidden Markov Models

All results are presented
In Colin Taylors thesis

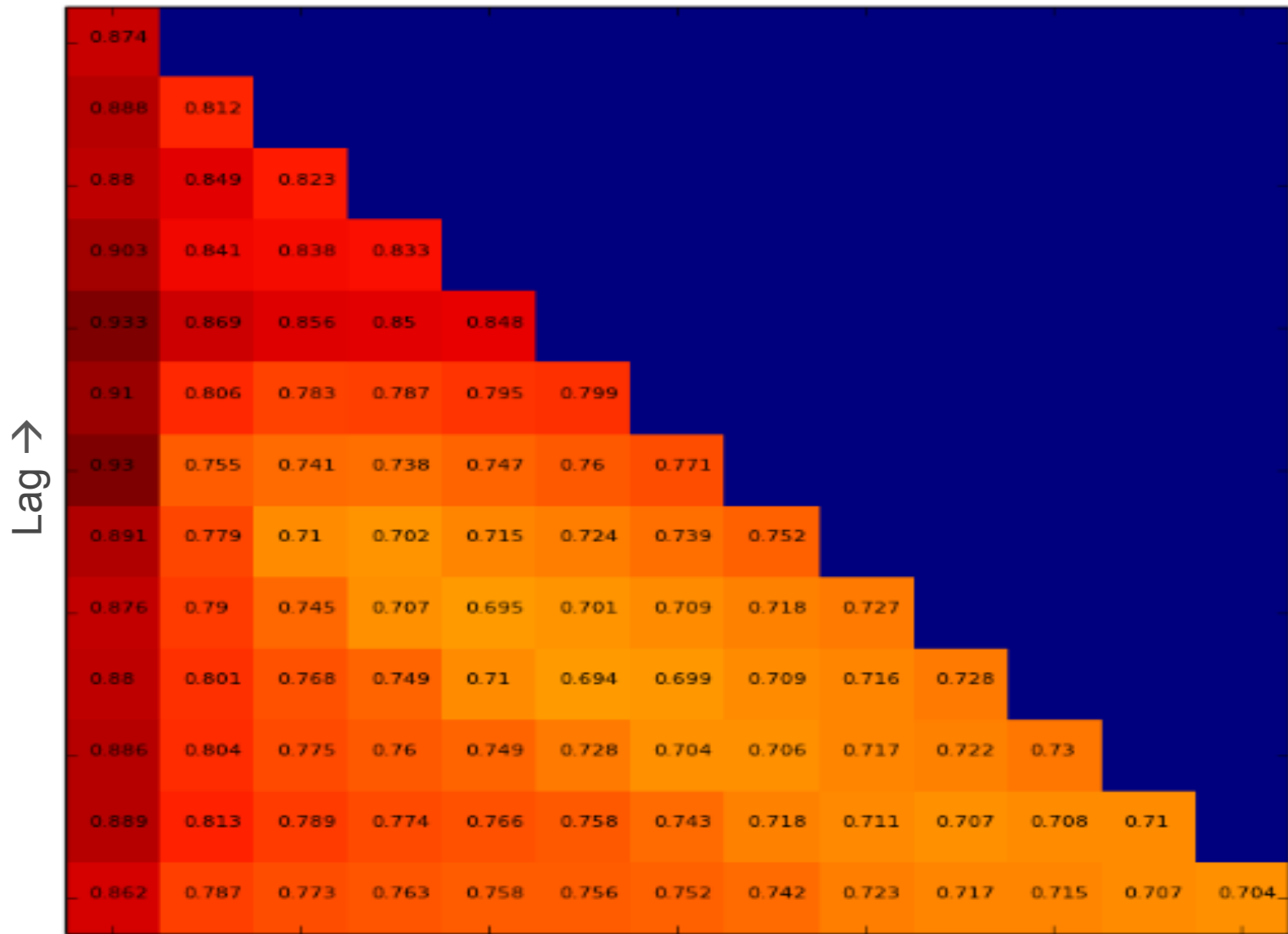


Hidden Markov Models + Logistic regression



Discriminative models

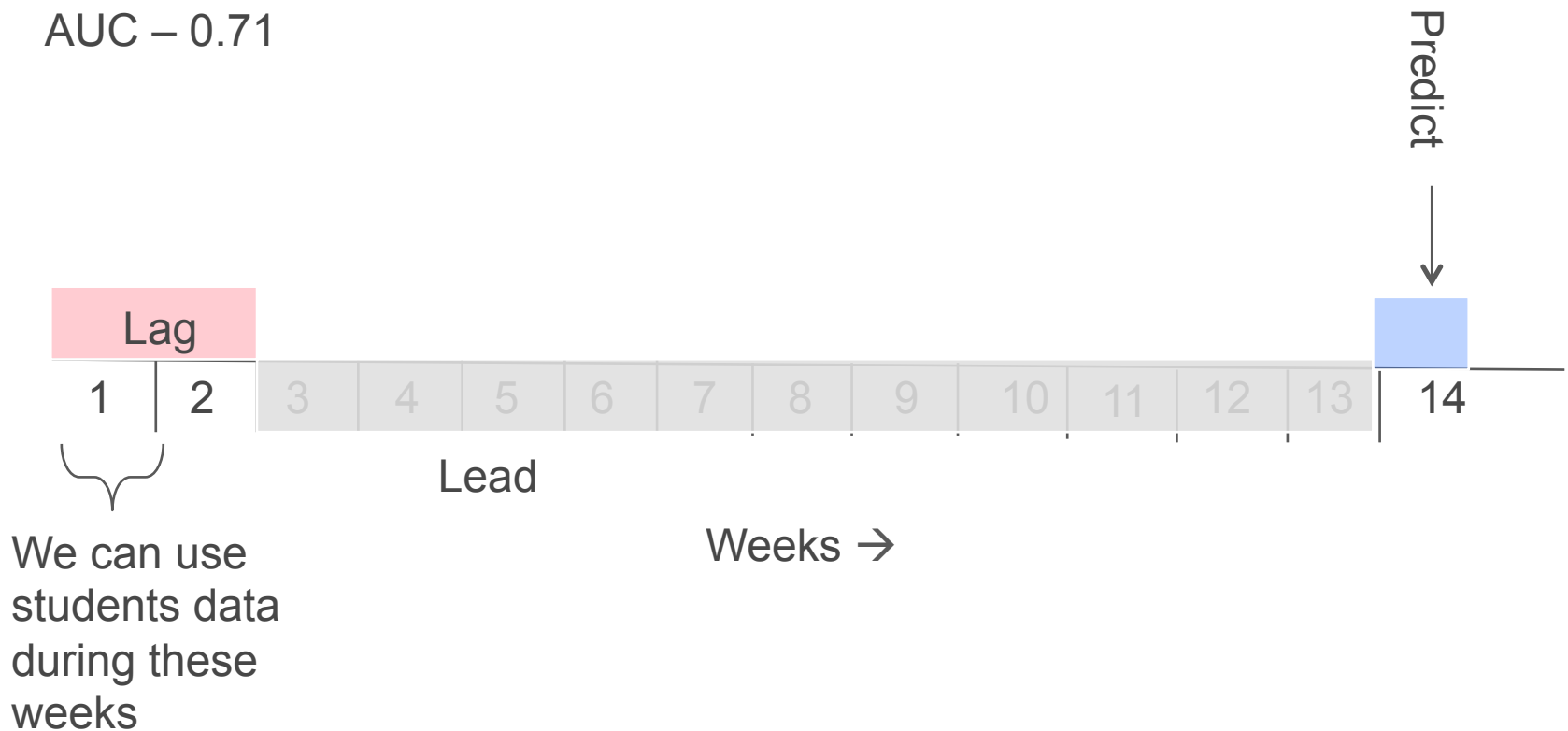
Results for students who do not collaborate



Defining the prediction problem

An extreme version of the problem

AUC – 0.71



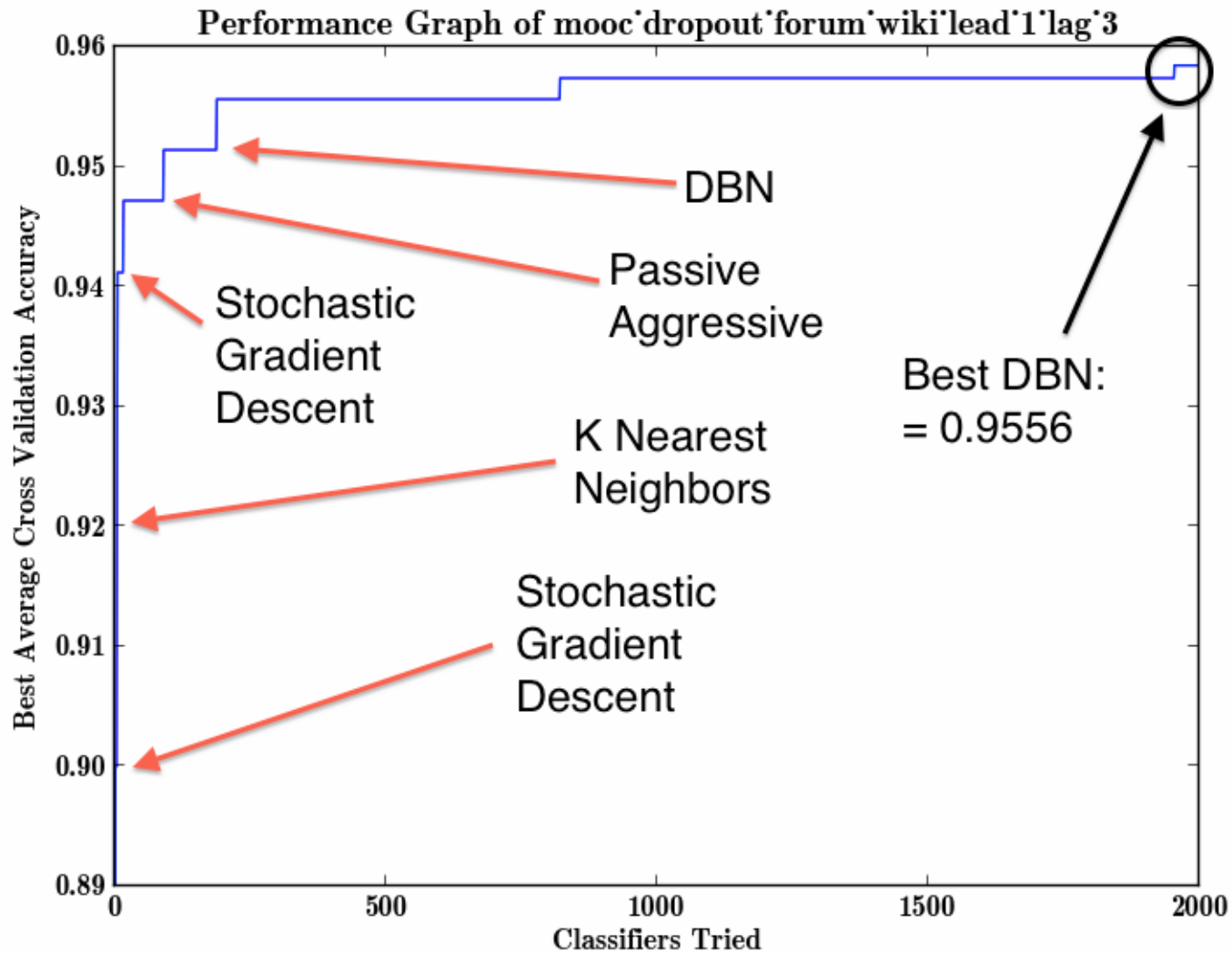
Top 10 features/variables that mattered

- 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



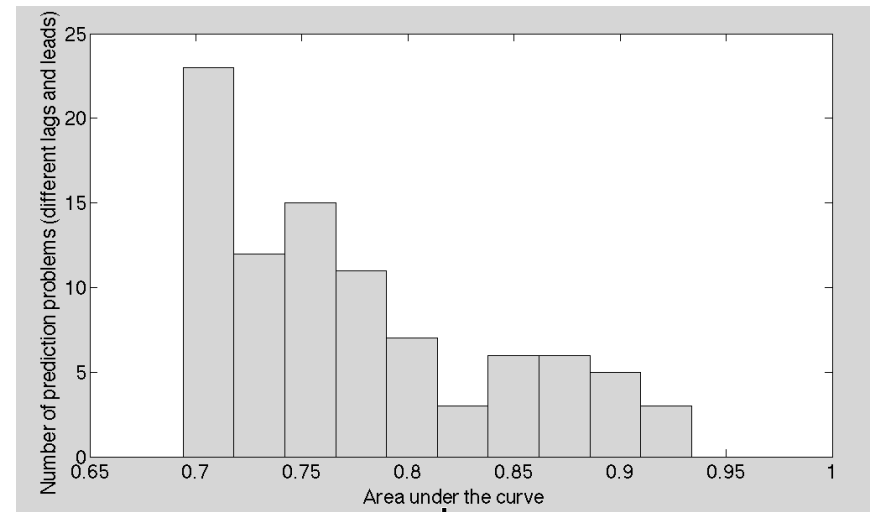
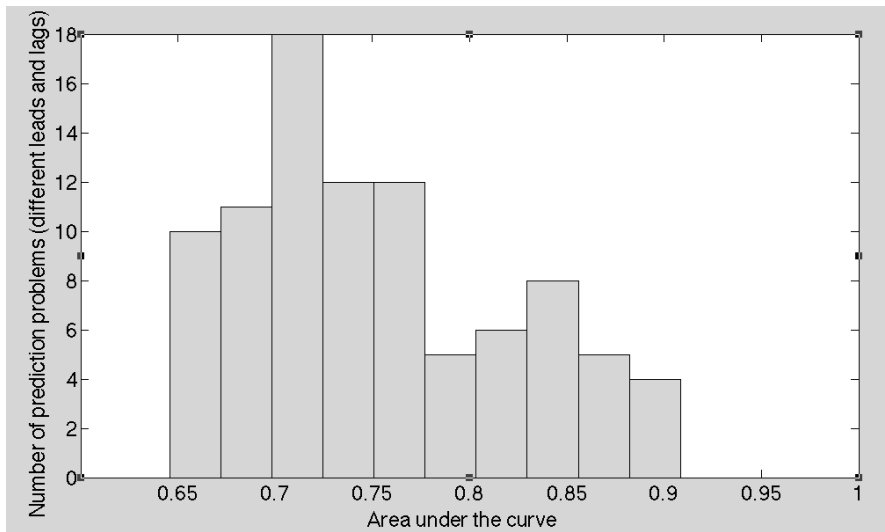
We can use
students data
during these
weeks

Discriminatory model building on the cloud



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

With and without crowd proposed features



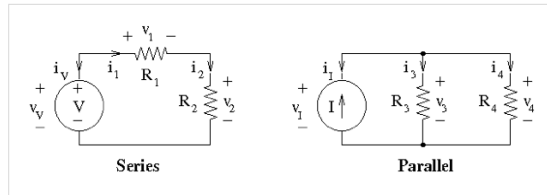
Conclusions

- 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

Problem

H1P2: DUALITY

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

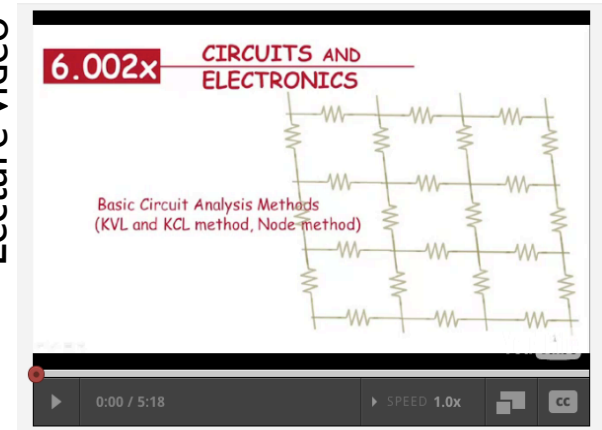


We are given device parameters $V = 15.3\text{V}$, $R_1 = 4.59\Omega$, and $R_2 = 3.06\Omega$. All of the unknown voltages and currents are labeled in associated reference directions. Solve this circuit for the unknowns and enter them into the boxes given.

The value (in Volts) of v_1 is:



Lecture video



Forum

subscripts for electrical quantities

voltage subscripts

flag offensive

how do you enter subscripts for say v_1 in the solution input boxes?

Asked Mar 05 '12 by
anyirenda 9 • 1

[post a comment](#)

2 Answers

Sort by [oldest](#) [newest](#) [most voted](#)

They're written just like you showed: v_1 , v_2 , V , etc.

Answered Mar 05 '12 by
amyc 10896 • 62
• 140 • 242

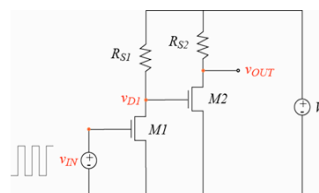
[post a comment](#)

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

244 CHAPTER FIVE THE DIGITAL ABSTRACTION

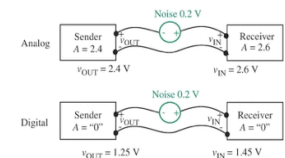
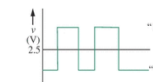


FIGURE 5.2 Signal transmission in the presence of noise. The noise is represented as a series voltage source.



a “0”, “1”, “0”, “1”, “0” sequence. Figure 5.3b shows the same signal with the superposition of some amount of noise, possibly during transmission through a noisy environment. The receiver will be able to receive the sequence correctly provided the noise levels in Figure 5.3b are small enough that the voltages for a logical 0 signal do not exceed 2.5 V, and the voltages for a logical 1 signal do not fall below 2.5 V. Specifically, notice that the binary mapping we have