SI 568

Introduction to Applied Data Science

Welcome

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Remote class

Please keep your camera on if you are comfortable.

This is a new class

We have a wonderful teaching team.

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Neha Bhomia

Pronouns: She/Her

A former Oral Surgeon and MHI Alumna, Dr. Neha Bhomia currently works as a Programmer Analyst, for Michigan Medicine. Her professional interests lie at the intersection of technology and healthcare delivery. She is an avid reader and an amateur baker and likes to spend her free time..... what free time?

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Michael Hess

Pronouns: He/Him

Michael Hess is an adjunct lecturer and Solution Architect Lead at the University of Michigan, as well as the lead of the Drupal Security Team. He works with the School of Information and the U-M Medical Center, teaching several courses at UMSI. He also serves in a consulting and development role for many other University departments. He is a graduate of UMSI, with a master's degree in Information.

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Qingyi Wang

Pronouns: She/Her

Qingyi Wang is a second-year PhD student at the School of Information. Her fields of interest include behavioral economics and experimental economics. She is the GSI of this course and will explore the world of data science together with you.

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Jingcong Hu

Pronouns: She/Her

Jingcong Hu is a Master's student at the School of Information studying Digital Curation, webmaster for Ann Arbor Data Dive, and a GSI for this course. Her professional interests include web development and databases.

Calendar

Weeks 1-7

- Math notations
- Linear Algebra
- Probability
- Big-O notation
- Parameter Estimation
- Maximum Likelihood
- Regression
- Bayesian Estimation

Weeks 9-14

- Data Science Ethics
- Communication in Data Science
- The different fields within data science, taught by the faculty who teach the topics at UMSI.
  - Machine Learning
  - Data Mining
  - Information Retrieval
  - Natural Language Processing
  - Networks

Workload

Each week, you should expect

- A lecture
- Some type of reading
- Homework problems

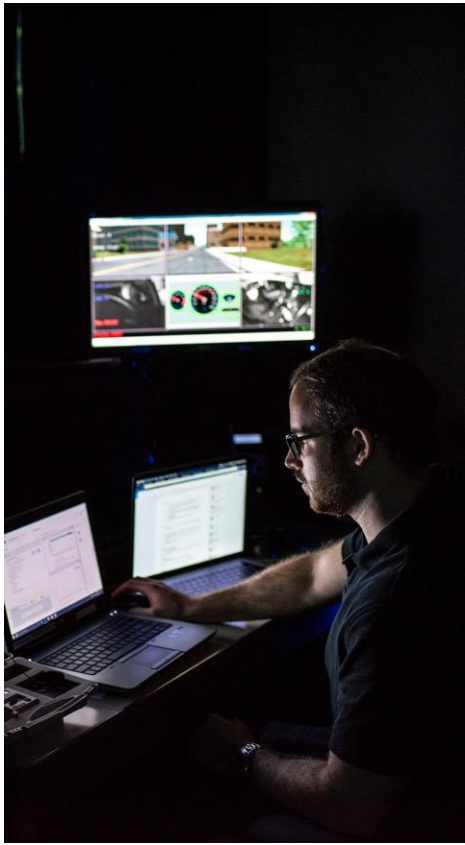
After week 7 you will have weekly data science related python problems to solve.

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Requests

- Join slack (slack.umich.edu)
- If emailing the teaching team use the email **si568-instructors@umich.edu**
- Do not post direct answers about homework in a public channel this is cheating. You can ask how to solve something, but don't show your work in the process.
- Please read the syllabus. All the details of class logistics and policies are clearly outlined there.
    - Any changes will be posted in the Announcement section of the course Canvas.

Becoming a Data Scientist



What is Data Science?

Please take 1 min and answer this question in zoom chat in under 30 words.

Please make sure your name in zoom chat, matches your name in wolverine access.

- What does it mean to be a Data Scientist?

Project

Always know the goal you are trying to solve?

Story time

Data Leakage

Data science is not magic.

Notebooks

Stats Review

Do you like to gamble?

Fair Coin flip. Heads you get $1, tails you get $0

Flip 1 => Tails

Flip 2 => Tails

Flip 3 => Tails

Flip 4 = Would you like to change your bet? If a heads is rolled in the next 3 flips, you get $50, if not, you lose $10

Coin Flip

$$P(H) = 0.5$$

$$P(T) = 0.5$$

But Math?

Well…Math is a four letter word, but so is good, neat, cool.

Data science = Math.

Why?

The background on math you get in this course is not going to make you a mathematician.

If you watch a movie in another language, you can read the subtitles, but knowing the language even parts of it, make the experience more enjoyable.

Math?

We have a wonderful team to cover the required topics in Math.

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Marisa Eisenberg

Pronouns: She/Her

Marisa Eisenberg is an Associate Professor of Epidemiology, Complex Systems, & Mathematics, and Interim Director of the Center for the Study of Complex Systems. Her research program is in mathematical epidemiology, and blends mathematics, statistics, and epidemiology to understand transmission dynamics, inform optimal intervention strategies, and improve forecasting.

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

Jeff Dunworth

Pronouns: He/Him

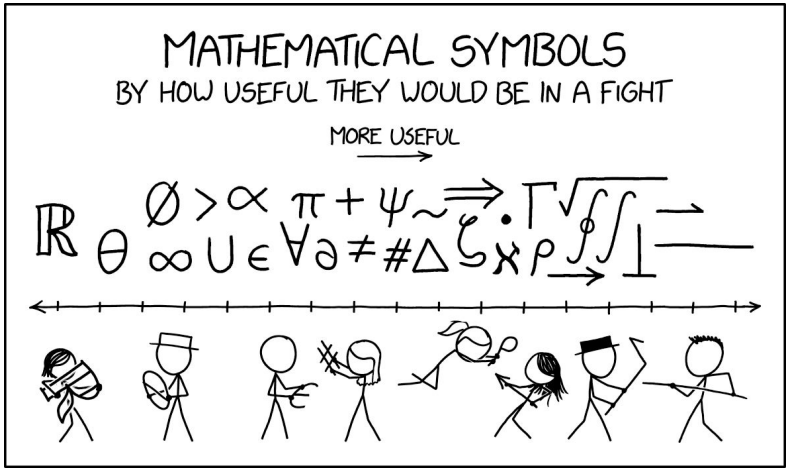
Jeff Dunworth is a lecturer in Mathematics and Complex Systems, with educational interests in continuing and post-secondary education with re-entry students. His research interests are in computational neuroscience, particularly focused on stochastic models.

Why do I need math for data science?

- **Math (including stats) and computation are the language of data science**

- Our goal in this crash-course is not to get you fully fluent in math, but more like 'order-at-a-restaurant' level of fluency, with the tools to pick up more as you need it

- In other words, our aim is to provide you with the tools to understand the notation, ways of thinking, and key ideas underlying data science



Source: xkcd

An example: linear regression

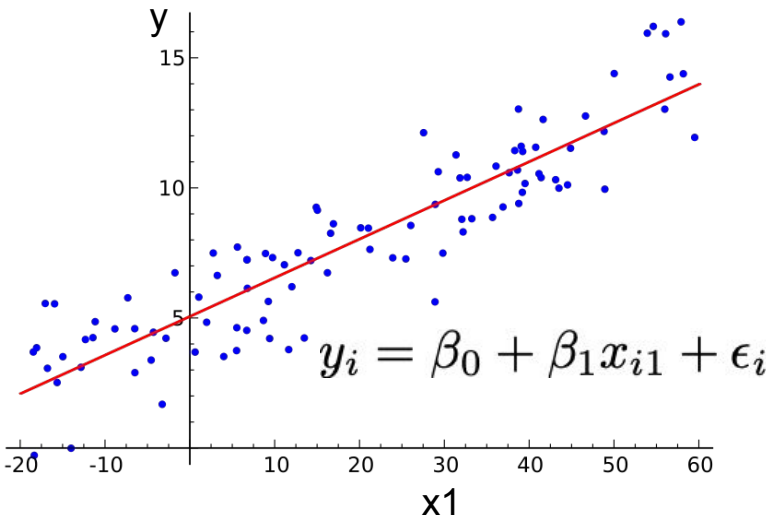
One of the most commonly used methods in data science/statistics/machine learning! How do we write down a linear regression model?

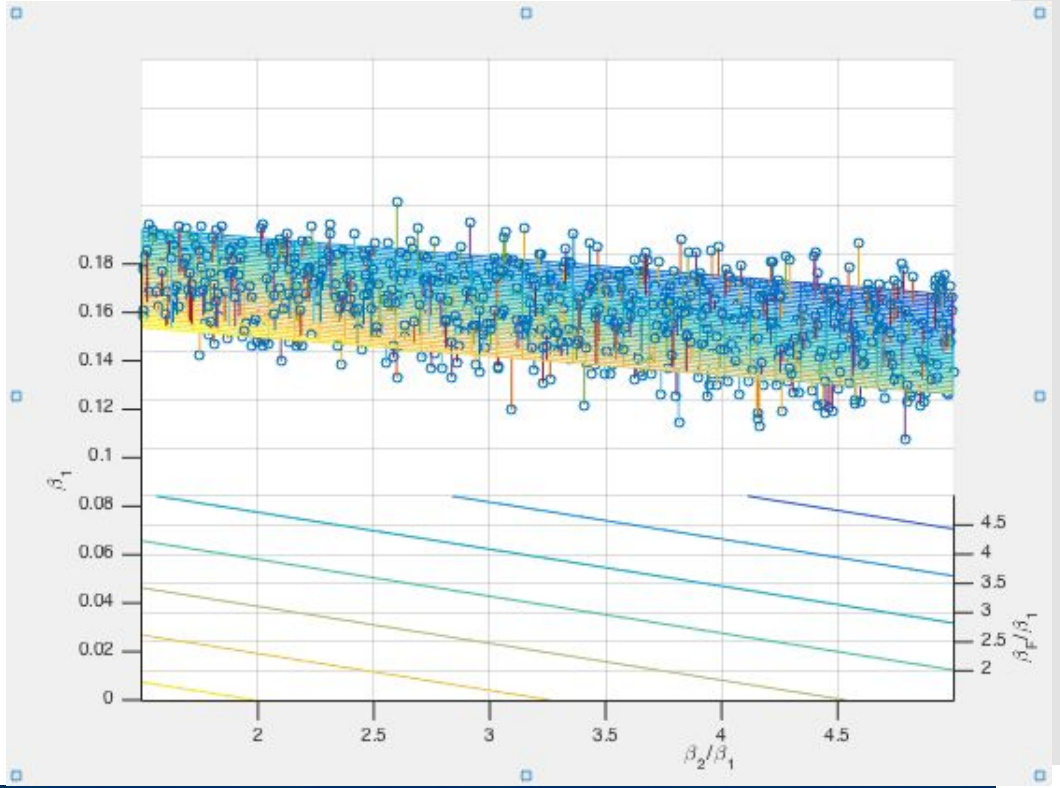
$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$y_i = \epsilon_i + \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \text{ for } i \in \{1, \ldots, n\}$$

Both equations above mean the same thing! Namely, we assume our observed data (y) is given by a linear function of our inputs (x), plus an error term (ε).

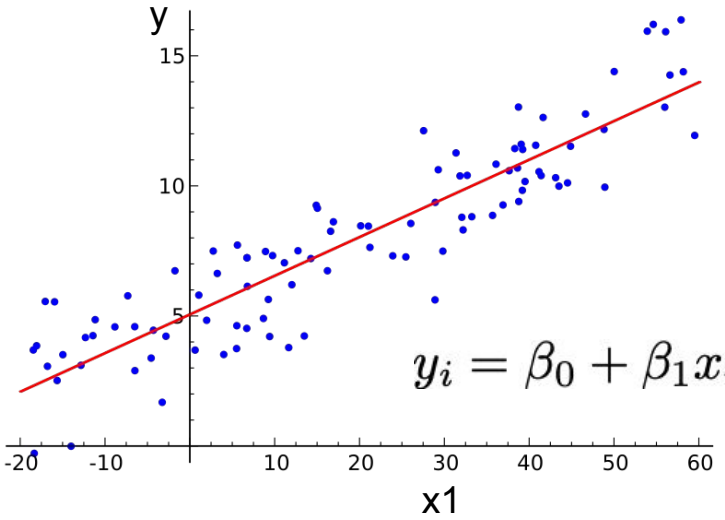
Visually, what does that mean? We are fitting a line or a plane to the data! (hyperplane in higher dimensions)



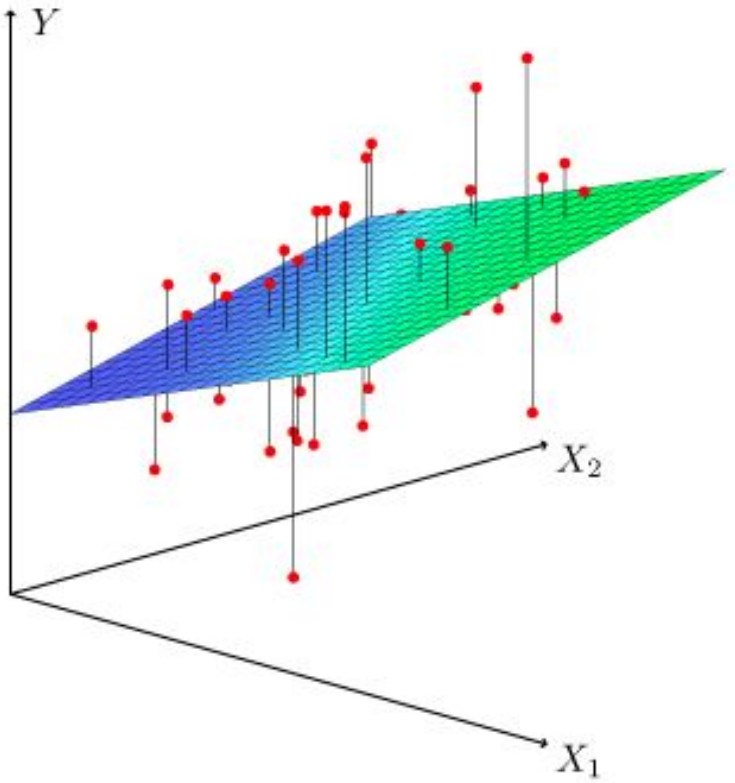


https://en.wikipedia.org/wiki/Linear_regression

Visually, what does that mean? We are fitting a line or a plane to the data! (hyperplane in higher dimensions)



$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$



https://en.wikipedia.org/wiki/Linear_regression

http://www.rnfc.org/courses/isl/Lesson%203/Summar

Sets

What is a set?

- A set is just a collection of objects—those objects can be numbers, points in space, websites, cats, anything! (usually mathematical objects though)

Set notation (doodle various examples, also reminder to introduce integers & reals)

$$A = \{1, 2, blue, x, \pi\}$$

$$A = \{1, 2, 3, 4, \dots\}$$

Roster notation

$$B = \{x | x \in \mathbb{Z}, x > 0\}$$

Set builder notation

$$\varnothing$$

De Morgan’s Laws

Jaccard similarity

Sums & products notation

- One of the most common elements of mathematical notation

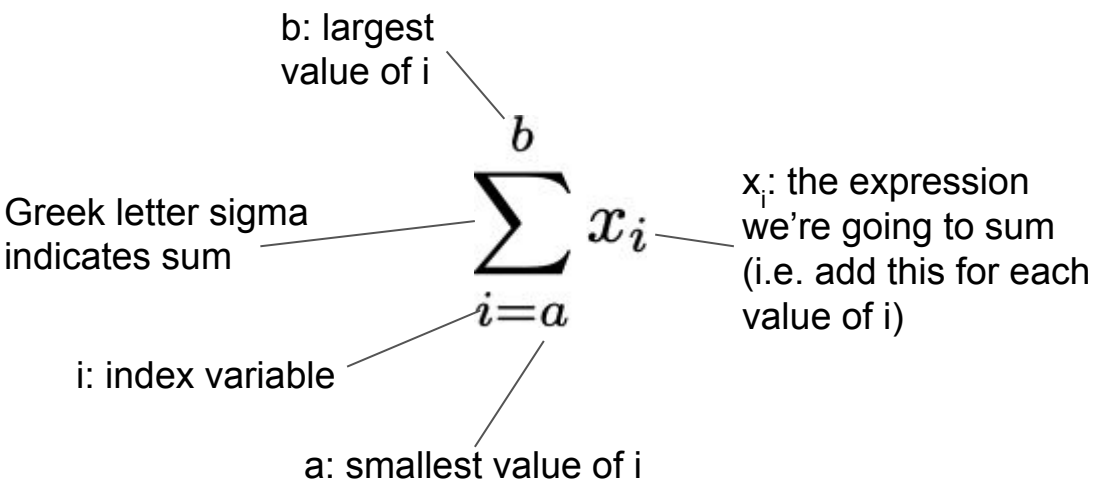
- Example: Add up 2i, for i = 1,2,3,4,5

$$\sum_{i=1}^{5} 2i = (2 \cdot 1) + (2 \cdot 2) + (2 \cdot 3) + (2 \cdot 4) + (2 \cdot 5)$$

- Code version:

```
x = 0
for i in [1,2,3,4,5]:
    x = x + 2*i
print(x)
```

= 30



$$\sum_{i=a}^{b} x_i$$

Greek letter pi indicates product—same idea but now we multiply the terms instead of add them

$$\prod_{i=a}^{b} x_i$$

$$\sum_{i \in A} x_i$$

Same idea, but now instead of upper and lower limits of the sum, we have a set A, and i takes on each value in A

$$A = \{1, 2, 3, 4, 5\}$$

Example

Math

$$\sum_{i=0}^{2}(i+x) = 0+x+1+x+2+x$$

$$= 0+1+2+x+x+x$$

$$= \sum_{i=0}^{2} i + \sum_{i=0}^{2} x$$

= 15

Code (setting x = 4)

```
x = 4
sum = 0
for i in range(3):
    sum = sum + i + x
print(sum)
```

How else could you code this?

Double Sum

Math

$$\sum_{i=3}^{4}\sum_{j=0}^{2}(i \cdot j) = ((3 \cdot 0) + (3 \cdot 1) + (3 \cdot 2)) + ((4 \cdot 0) + (4 \cdot 1) + (4 \cdot 2))$$

i = 3 i = 4

Code

Double sums become nested for loops

```
sum = 0
for i in [3,4]:
    for j in [0,1,2]:
        sum = sum + i*j
print(sum)
```

Infinite series

Math

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \cdots$$

$$= \frac{\pi^2}{6} \approx 1.644934\ldots$$

Code for the first 100 terms

```
sum = 0
for i in range(1,100):
    sum = sum + 1/(i**2)
print(sum)
```

```
1.6348839001848923
```

https://en.wikipedia.org/wiki/Basel_problem

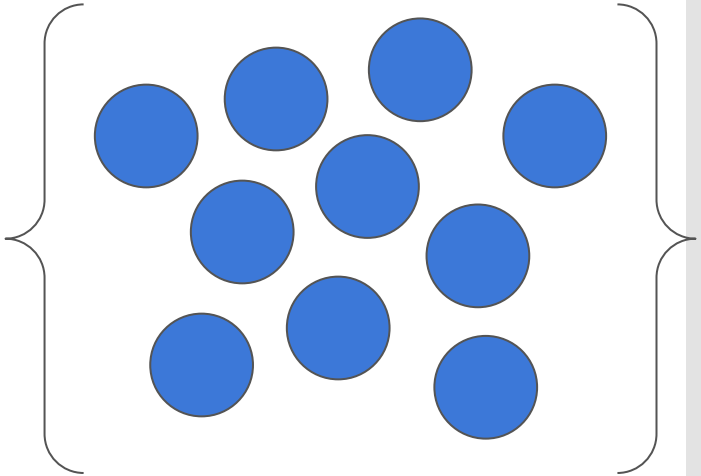
Big O notation

Problem: Subdividing sets!

Suppose you start with a single set of 10 objects (e.g. a pile of chips/coins/sticks)

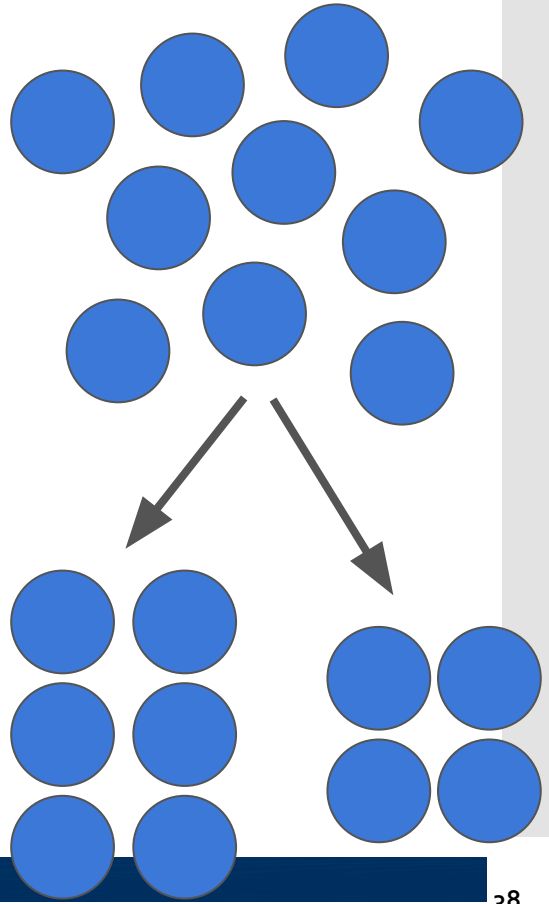
At each step of the game, you can choose a set and split it into two sets (they don't have to be equal sized!). Continue until all the sets have size = 1.

Your score is determined as follows: Your initial score is zero, and each time you divide a set into two, you add to your score the product of the sizes of the two new sets.



Subdividing sets: example game

Piles	Score	Move
10	0	$10 \rightarrow 6, 4$
6, 4	$0 + 6 \times 4 = 24$	$6 \rightarrow 3, 3$
3, 3, 4	$24 + 3 \times 3 = 33$	$3 \rightarrow 2, 1$
2, 1, 3, **4**	$33 + 2 \times 1 = 35$	$4 \rightarrow 2, 2$
2, 1, 3, **2**, 2	$35 + 2 \times 2 = 39$	$2 \rightarrow 1, 1$
2, 1, 3, 1, 1, 2	$39 + 1 \times 1 = 40$	$2 \rightarrow 1, 1$
1, 1, 1, **3**, 1, 1, 2	$40 + 1 \times 1 = 41$	$3 \rightarrow 2, 1$
1, 1, 1, 2, 1, 1, 1, **2**	$41 + 2 \times 1 = 43$	$2 \rightarrow 1, 1$
1, 1, 1, **2**, 1, 1, 1, 1, 1	$43 + 1 \times 1 = 44$	$2 \rightarrow 1, 1$
1, 1, 1, 1, 1, 1, 1, 1, 1, 1	$44 + 1 \times 1 = 45$	Game over



Subdividing sets: activity

Try out splitting the pile/set in a few different ways.

What are the highest and lowest scoring games you can find?

How many steps do your games take? What are the longest and shortest games you can find?

What did you find?

Why?

Does this pattern hold for other initial sizes?

Another way to think about this: multiplication & networks

Homework

2 parts

1) Why do you want to be a data scientist?
2) In your own words explain a math concept

Homework is due right before class next week. If you need help, setup office hours.

Information
changes
everything.

Thank You

SCHOOL OF
INFORMATION
UNIVERSITY OF MICHIGAN