Data Exploration, Cleaning, and Integration for Data Science

AnHai Doan

What We Will Discuss

- **What is data science?**
- **The rise of data science**
- **Data exploration, cleaning, and integration**
  - Addressing the two Vs of Big Data: variety and veracity
- **Why do you need these even if you don't do data science**
- **Course coverage, goals, syllabus**
- **What you should do in the next few weeks**

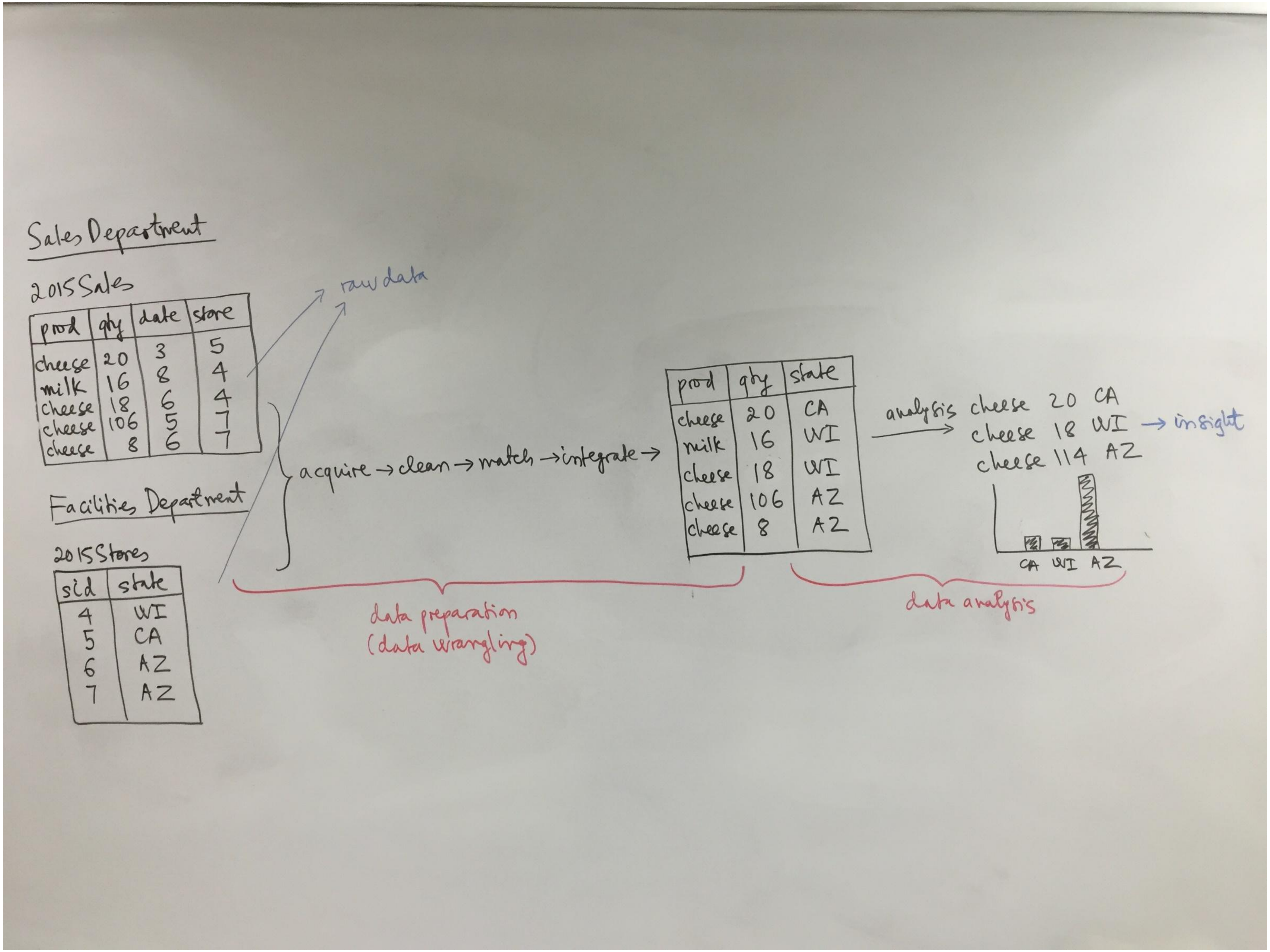
What Is Data Science?

- **No one really knows**
  - the field has not been around long enough so that people converge to a single definition
- **There is a popular joke about this**

- **Our definition: data science is a new interdisciplinary field that develops principles, algorithms, and best practices to manage data, focusing on**
  - Infer insights from raw data (the raw data to insight pipeline)
  - Build data driven artifacts (e.g., knowledge graphs, recommender systems)
  - Design data driven experiments to answer questions
- **The field draws on CS, stat, math, operation research, optimization, information science, etc.**

- **This is a broad definition encompassing most current definitions**
  - DS definition may still change in the future
  - But this concrete definition will be sufficient for us to get started

The Raw Data to Insight Pipeline



Sales Department

2015 Sales

prod	qty	date	store
cheese	20	3	5
milk	16	8	4
cheese	18	6	4
cheese	106	5	7
cheese	8	6	7

raw data

Facilities Department

2015 Stores

sid	state
4	WI
5	CA
6	AZ
7	AZ

acquire → clean → match → integrate →

prod	qty	state
cheese	20	CA
milk	16	WI
cheese	18	WI
cheese	106	AZ
cheese	8	AZ

anal

data preparation
(data wrangling)

d

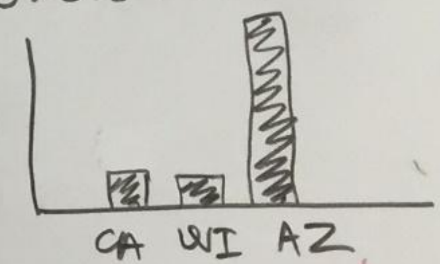
raw data

acquire → clean → match → integrate →

prod	qty	state
cheese	20	CA
milk	16	WI
cheese	18	WI
cheese	106	AZ
cheese	8	AZ

$\xrightarrow{\text{analysis}}$

cheese 20 CA
cheese 18 WI → insight
cheese 114 AZ

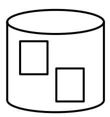


data preparation
(data wrangling)

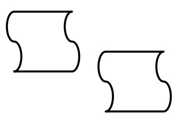
data analysis

Another Example

- **The raw data to insight pipeline**



id	name	loc
x_1	Apple	CA
x_2	IBM	NY



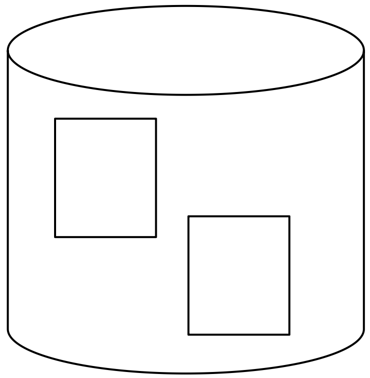
id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

name	loc	rev
Apple Inc	CA	51
IBM Corp	NY	25
GE		35

Data analysis

is there any correlation between location and revenue?

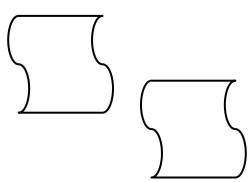
Data integration, data wrangling, …



X

id	name	loc
x_1	Apple	CA
x_2	IBM	NY

news
articles



data
extraction

Y

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

X

id	name	loc
x_1	Apple	CA
x_2	IBM	NY

Y

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

data cleaning: GE revenue: 351 $\longrightarrow$ 35.1

schema matching: name = cname
loc = address

schema merging:
$$\left.\begin{array}{l}\text{X(name, loc)}\\ \text{Y(cname, address, rev)}\end{array}\right\}\ \text{Z(name, loc, rev)}$$

data matching:

M

xid	yid
x_1	y_2
x_2	y_1

data merging: for name, return the longer string from X.name and Y.cname
for loc, return X.loc

schema mapping: Z = select merge_name(X.name, Y.cname), X.loc, Y.rev
from X, Y, M
where X.id = M.xid and Y.id = M.yid

X

id	nam e	loc
x_1	Apple	CA
x_2	IBM	NY

Y

id	cname	addres s	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

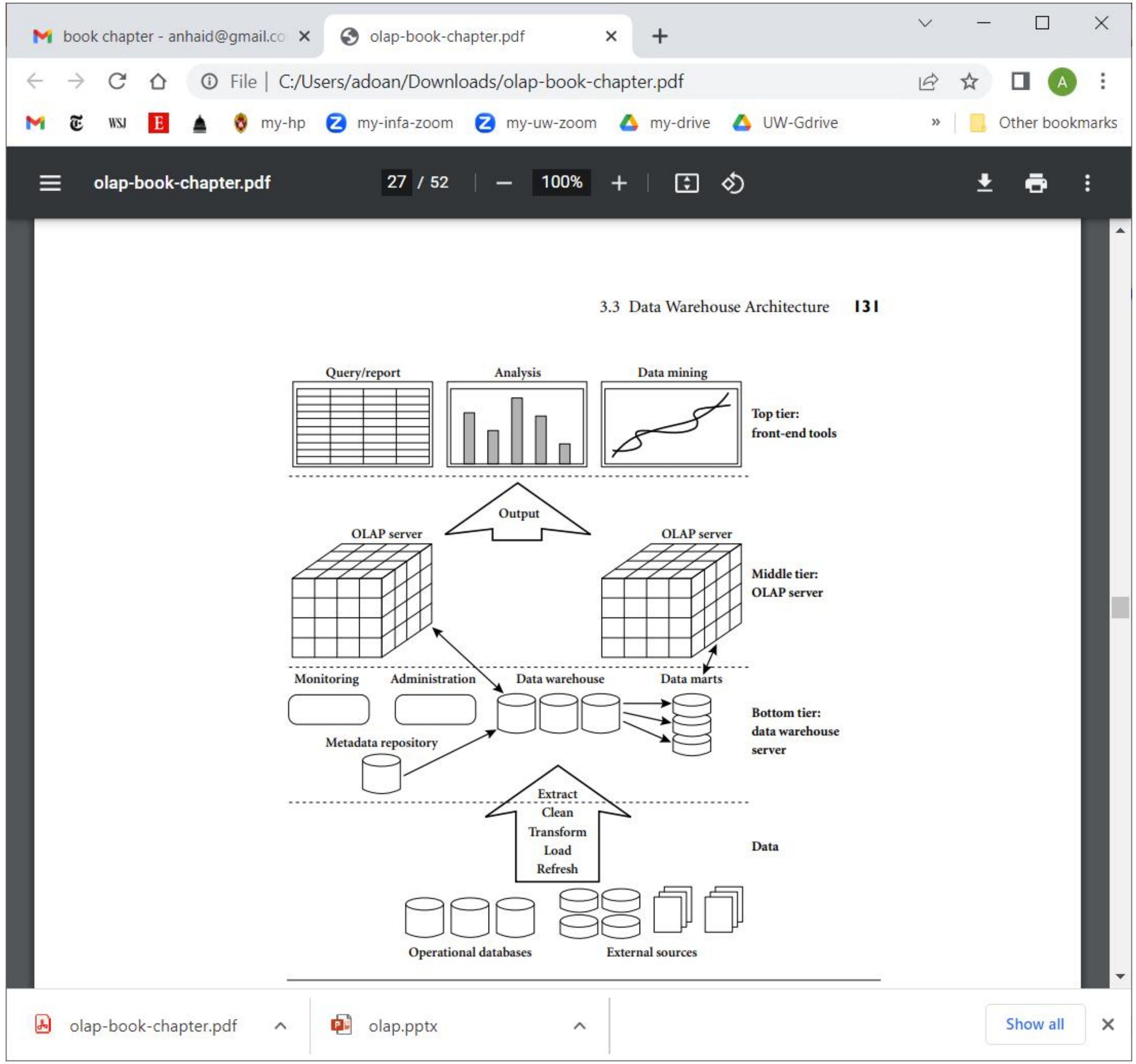
$\longrightarrow$

Z

name	loc	rev
Apple Inc	CA	51
IBM Corp	NY	25

Building Data Driven Artifacts

- Data warehouse



Fact Table

Sales

Dimension attributes: tid, pid, lid

Measurement attributes: qty

tid	pid	lid	qty
t_1	p_1	l_1	3
t_1	p_1	l_2	4
t_2	p_2	l_1	6
...	...	...	...

Dimension Tables

Times

tid	day	quarter	year
t_1	3	1	2016
t_2	10	2	2016
...	...	...	...

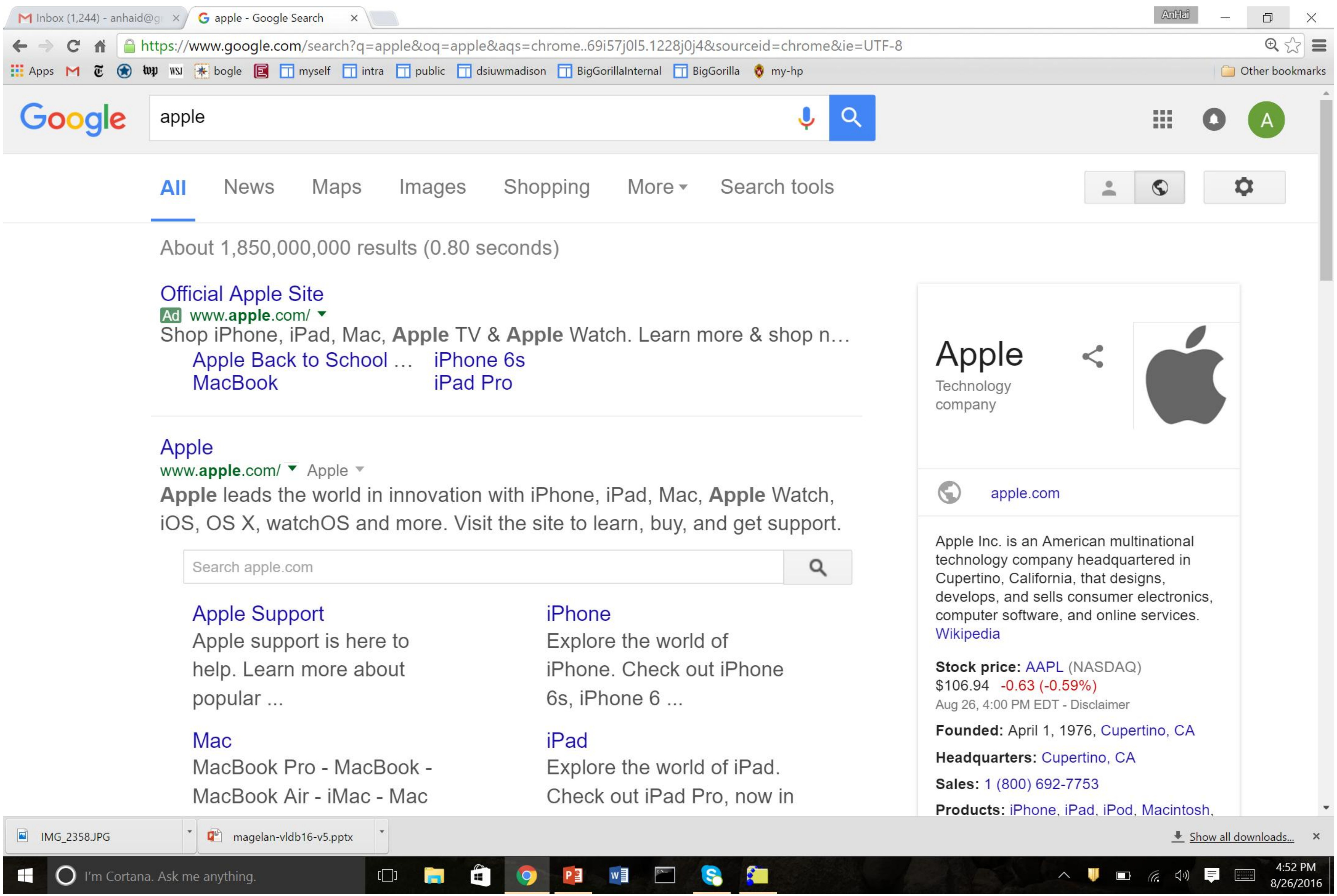
Products

pid	name	color	brand
p_1	TV	black	Sony
p_2	Phone	white	LG

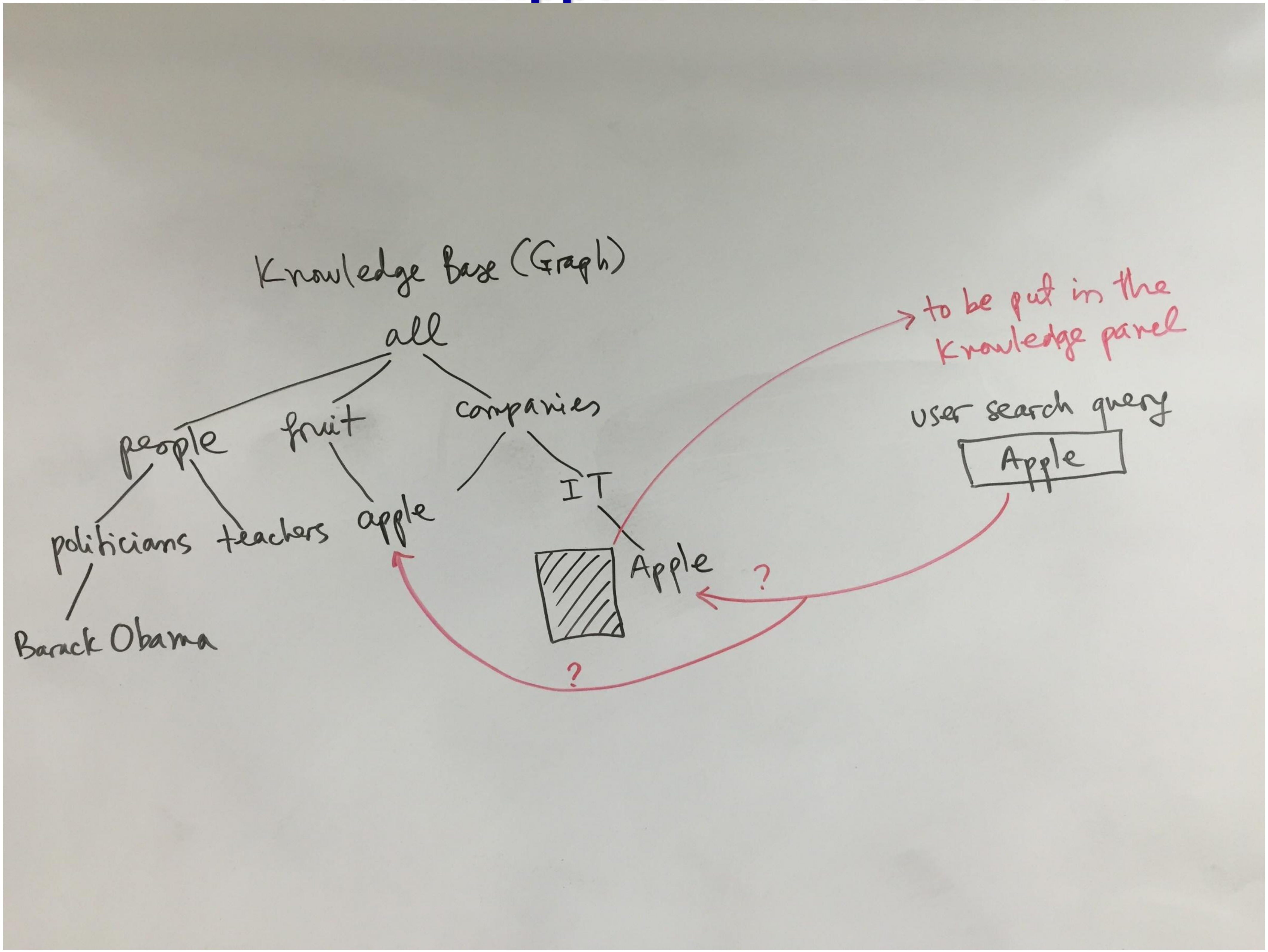
Locations

lid	city	state	country
l_1	Madison	WI	USA
l_2	Milwaukee	WI	USA
l_3	San Jose	CA	USA
l_4	Toronto	ON	Canada

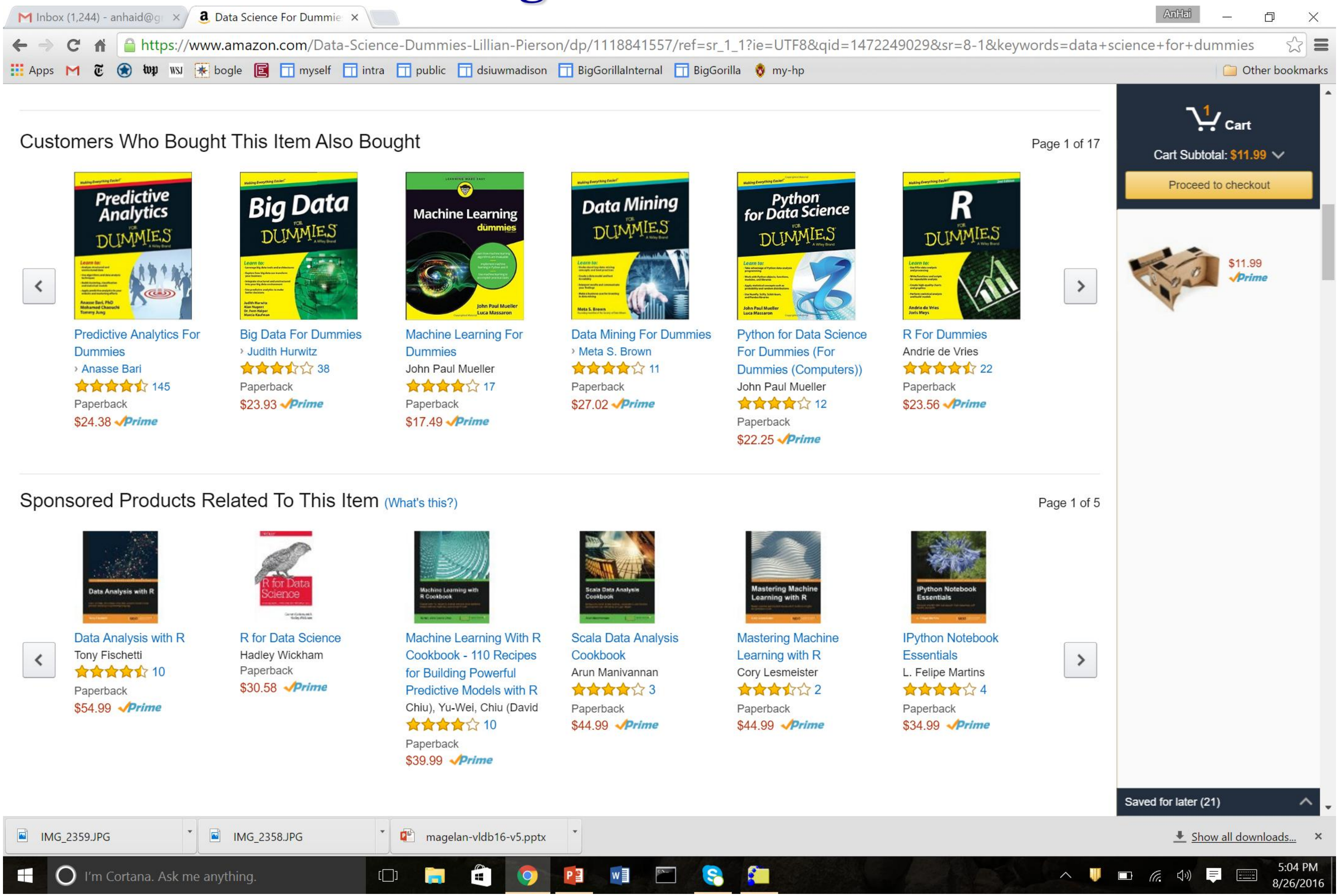
Building Data Driven Artifacts



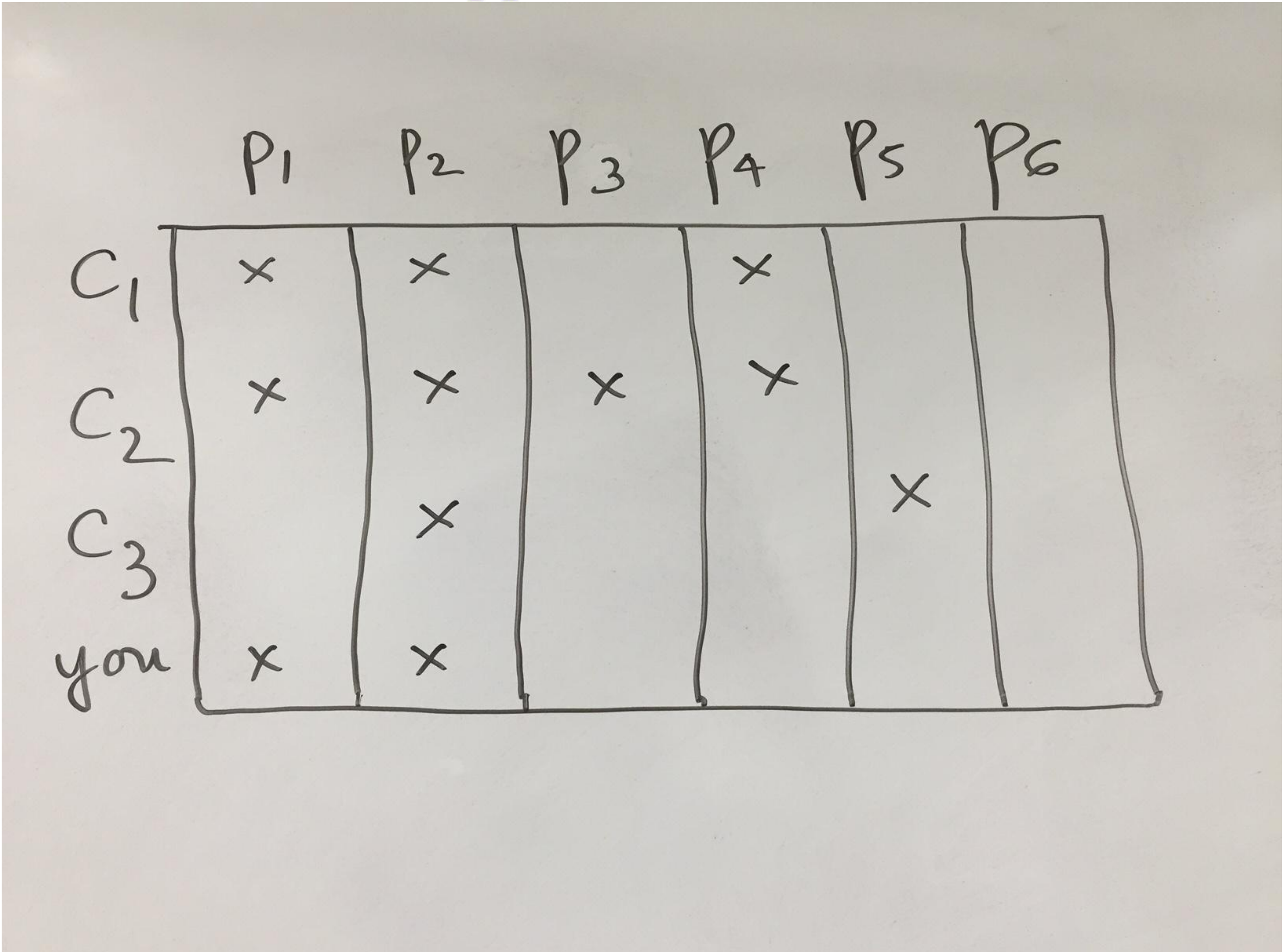
What Happens in the Backend?



Building Data Driven Artifacts



What Happens in the Backend?



What Kinds of Data-Driven Artifacts?

- **We have discussed data warehouses, knowledge graphs, and recommender systems**
- **Would a relational database fit here? After all, it is also a data-driven artifact**

- **DS focuses on data-driven artifacts that can be used to infer insights or provide more semantic information (in the "discover something new" sense)**
  - For example, a data warehouse is used to infer insights (aka actionable knowledge)
  - A knowledge graph can be used to provide more semantic information to a search query; it can also be used to help the process of inferring insights from raw data
  - A recommender system offers insights about what kinds of book (say) you may be interested in
  - A relational database, in contrast, is typically used for querying and manage user transactions.

Design Data Driven Experiments

- **For example A/B testing**

- **Problem**
  - Have two new ads A and B, don't know which one would be better, i.e., users would be more likely to click on
  - Show A to a fraction of incoming user traffic, show B to a fraction of incoming user traffic
  - Collect data, use statistical analysis to decide if the difference in user reactions to A and B is statistically significant

- **Example**
  - If 990 out of 1000 users who view A click on the ad, while only 10 out of 1000 users who view B click on the ad ➔ A is better
  - If 200 out of 1000 and 180 out of 1000?
  - If 200 out of 1000 and 150 out of 1000?

- **The previous two directions deal with existing data, here we generate data**

The Rise of Data Science

- **RDBMSs**
  - transactional data management, belong to the CIO, no one else cares about data
  - Data is not at the heart of the enterprise

- **Web => Google, other Web companies**
- **Social media: Twitter, Facebook, blogs, …**
- **Cloud computing (e.g., at Amazon), crowdsourcing (Wikipedia, Mturk)**

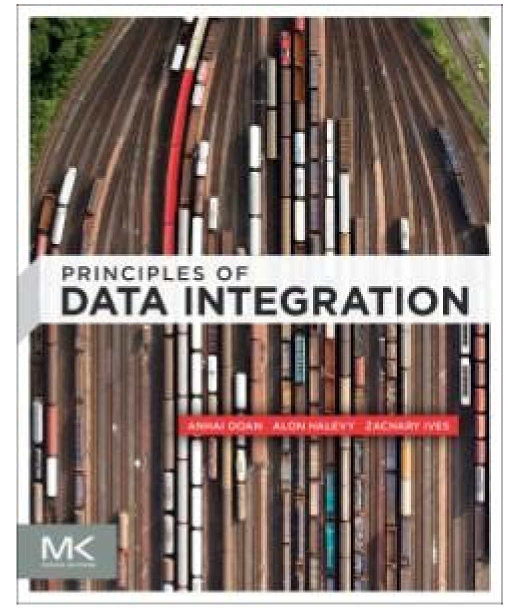
- **In recent years three trends emerged**
  - much easier to generate and capture data
  - much easier to process data (eg using open source software, cloud computing)
  - many more people become involved (e.g., Wikipedia, Facebook face tagging)
- **Lead to a major change in perception ➔ the Big Data trend**
  - data is now at the heart of enterprises, at the heart of everything
  - people want to capture as much data as possible, process it, infer insights
  - **Everything is becoming increasingly data driven**
- **Data science emerged to respond to this need**

How is DS Different From …

- **RDBMSs**
  - Is concerned with tabular data
  - A major focus is on transaction processing
  - Not concerned with the three tasks that we discussed
  - But they are often used to store and query tabular data
- **Machine learning, visualization, optimization, etc**
  - these are the techniques that DS often use
- **Big Data**
  - Two meanings: (a) the Big Data phenomenon, (b) systems that can process large amounts of data
- **Data mining**
  - Historically is concerned with the first task: infer insights from raw data
  - Can be viewed as “subsumed” or “being continued“ by data science (a controversial point)
- **Statistics**
  - Historically is concerned with (a) infer properties of a whole population based on those of a sample, and (b) design data intensive experiments to answer questions
  - But typically has not been concerned with a very large volume of data, has not been concerned with data wrangling
  - Typically does not deal with “existing” data
- **All of the above are necessary for data science**

Data Exploration, Cleaning, and Integration

- **When doing DS, users often must do data exploration, cleaning, integration**
- **Examples**
- **Often take an enormous amount of time**
  - often quoted number is 80%
- **These activities have been studied for 40+ years**
- **The field is known under many names**
  - data integration, data wrangling, data preparation, data curation, etc.
  - I have been working on it for 20+ years



- **It is becoming increasingly critical**
  - initially, a lot of work in DS focus on the analysis step
  - now people increasingly realize that “garbage data in, garbage results out”
  - so a lot of work is increasingly devoted to this field
- **This course covers this field**
  - we say data exploration, cleaning, integration
  - but we do refer to the entire field

Data Exploration, Cleaning, and Integration

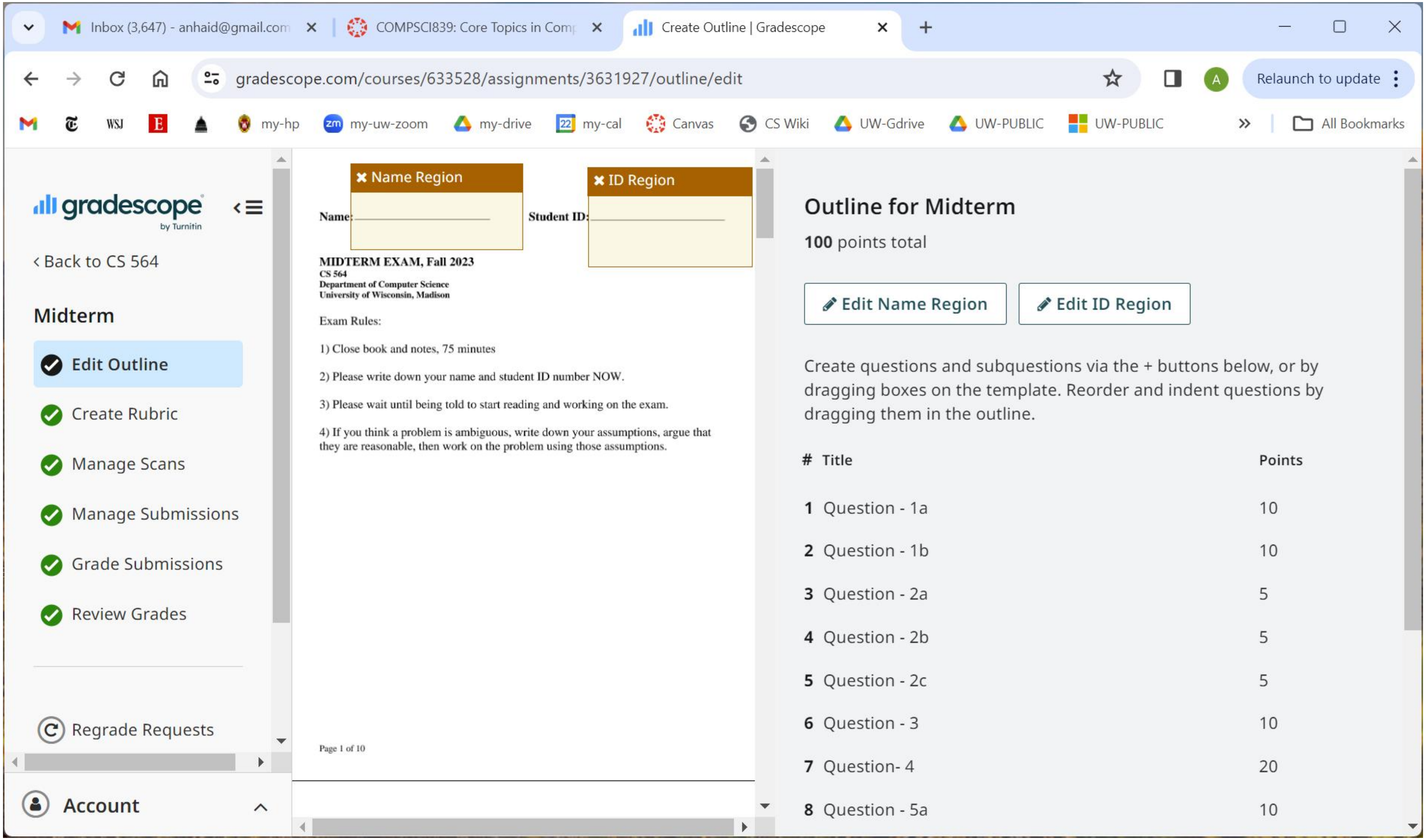
- **Solving these problems often requires multiple techniques**
  - databases, ML, big data scaling (e.g., Spark), effective user interaction, crowdsourcing, etc.

What We Will Discuss

- What is data science?
- The rise of data science
- Data exploration, cleaning, and integration
  - Addressing the two Vs of Big Data: variety and veracity
- Why do you need these even if you don’t do data science
- Course coverage, goals, syllabus
- What you should do in the next few weeks

Why Do You Need These Even without DS?

- **Data used to be isolated in a corner of a company**
- **Now data is at the heart of a company**
  - Treated as a major asset ➔ Big Data
  - What other assets companies have? Human resources, tech know-how, products, etc.
- **When dealing with data, what are major challenges?**
  - Volume, velocity ➔ CS 544, CS 744
  - Variety, veracity ➔ this course
- **Variety and veracity challenges come up even if you don't do DS**
  - If the company is moderately large, you WILL have these challenges
  - E.g., moving/integrating data between Canvas and Gradescope


Inbox (3,647) - anhaid@gmail.com
COMPSCI839: Core Topics in Comp
Create Outline | Gradescope
gradescope.com/courses/633528/assignments/3631927/outline/edit
Relaunch to update
my-hp
my-uw-zoom
my-drive
my-cal
Canvas
CS Wiki
UW-Gdrive
UW-PUBLIC
UW-PUBLIC
All Bookmarks
gradescope
by Turnitin
Back to CS 564
Midterm
Edit Outline
Create Rubric
Manage Scans
Manage Submissions
Grade Submissions
Review Grades
Regrade Requests
Account
Name Region
ID Region
Name:
Student ID:
MIDTERM EXAM, Fall 2023
CS 564
Department of Computer Science
University of Wisconsin, Madison
Exam Rules:
1) Close book and notes, 75 minutes
2) Please write down your name and student ID number NOW.
3) Please wait until being told to start reading and working on the exam.
4) If you think a problem is ambiguous, write down your assumptions, argue that they are reasonable, then work on the problem using those assumptions.
Page 1 of 10
Outline for Midterm
100 points total
Edit Name Region
Edit ID Region
Create questions and subquestions via the + buttons below, or by dragging boxes on the template. Reorder and indent questions by dragging them in the outline.
Title Points
1 Question - 1a 10
2 Question - 1b 10
3 Question - 2a 5
4 Question - 2b 5
5 Question - 2c 5
6 Question - 3 10
7 Question- 4 20
8 Question - 5a 10

Course Syllabus and Misc Issues

- **Let's discuss class homepage in Canvas**

- **You should start learning**
  - Python, pandas, machine learning, scikit-learn
- **Start thinking about project teams**