Data Mining

CS4821-CS5831-s24

Laura E. Brown

Jan. 10, 2024

Some slides from G. Piatetsky-Shapiro; Han, Kamber, & Pei; M. Hahsler
P. Smyth; C. Volinsky; Tan, Steinbach, & Kumar; J. Taylor; G. Dong; L. Hannah

Today’s Agenda

- Examples of Data Mining (continue from 01.intro.part1)

- Multidimensional View of Data Mining

- Tools for Data Mining

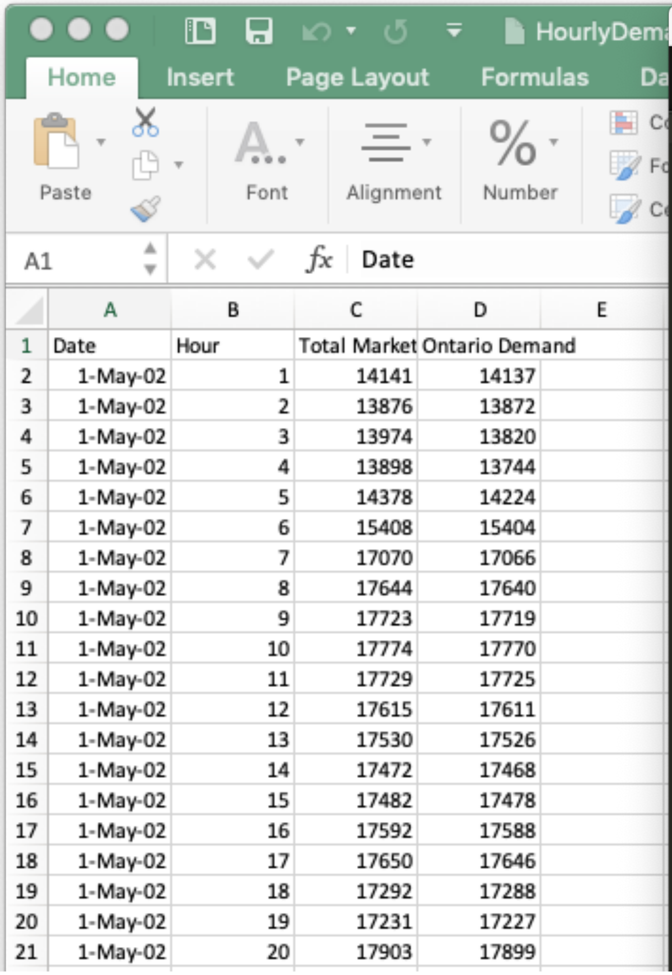
- Data Mining vs. Privacy

Multi-dimensional View of Data Mining

- What kinds of **data** can be mined?
- What kinds of **patterns** can be mined?
- What kinds of **techniques** are used?
- What kinds of **applications** are targeted?

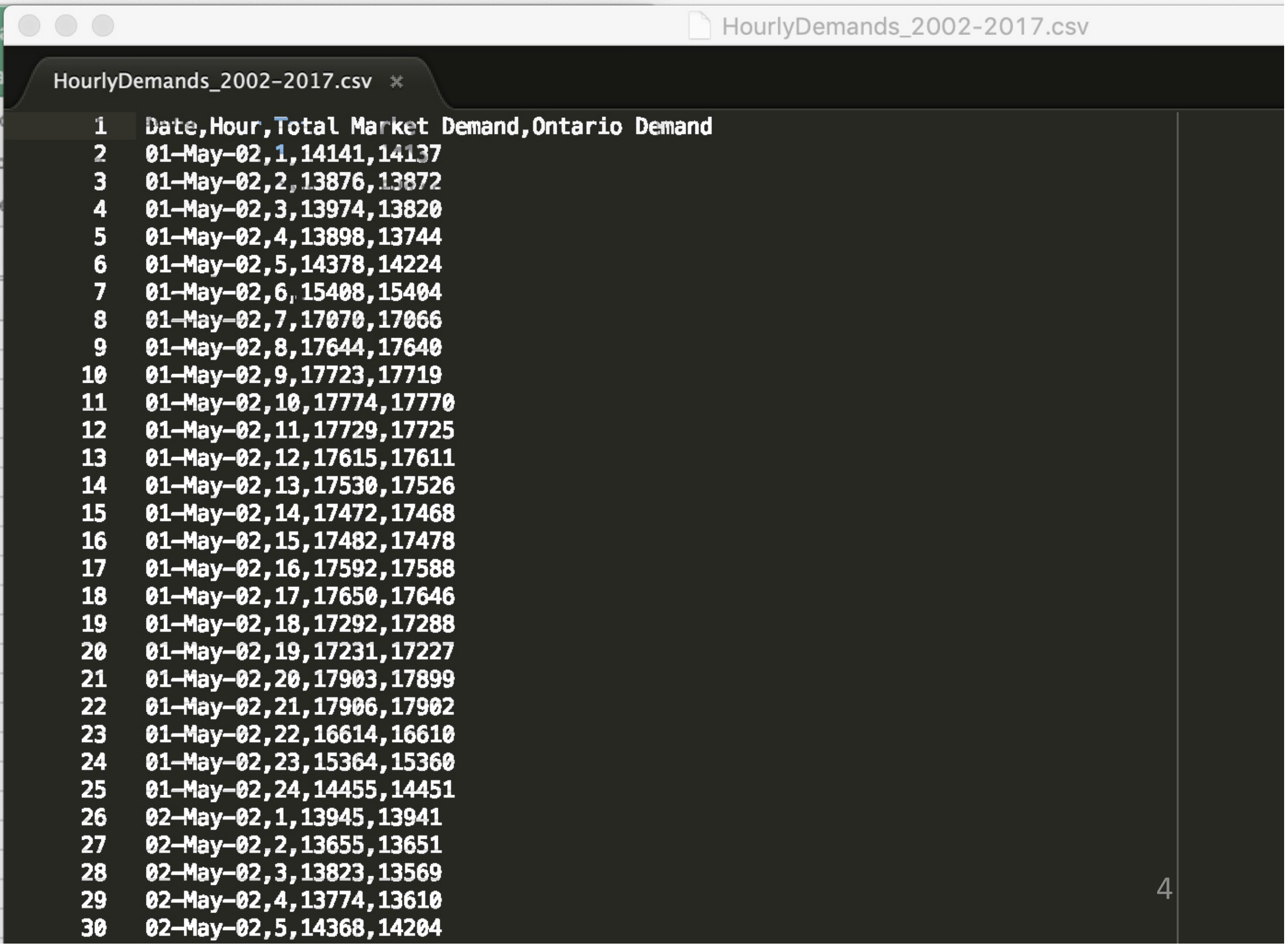
What kind of data?

- Spreadsheets
- Flat file, vector data



Date

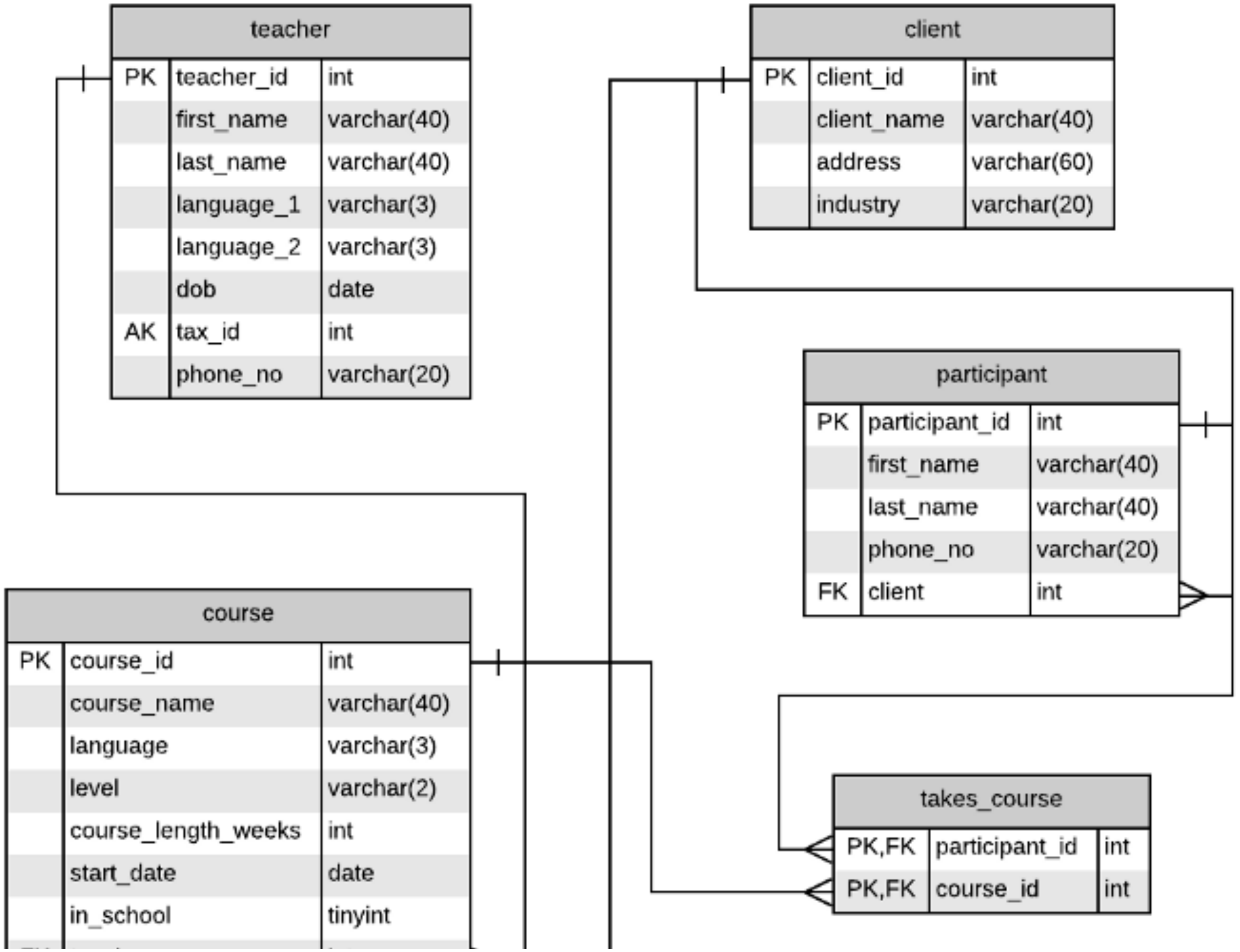
	A	B	C	D	E
1	Date	Hour	Total Market	Ontario Demand	
2	1-May-02	1	14141	14137	
3	1-May-02	2	13876	13872	
4	1-May-02	3	13974	13820	
5	1-May-02	4	13898	13744	
6	1-May-02	5	14378	14224	
7	1-May-02	6	15408	15404	
8	1-May-02	7	17070	17066	
9	1-May-02	8	17644	17640	
10	1-May-02	9	17723	17719	
11	1-May-02	10	17774	17770	
12	1-May-02	11	17729	17725	
13	1-May-02	12	17615	17611	
14	1-May-02	13	17530	17526	
15	1-May-02	14	17472	17468	
16	1-May-02	15	17482	17478	
17	1-May-02	16	17592	17588	
18	1-May-02	17	17650	17646	
19	1-May-02	18	17292	17288	
20	1-May-02	19	17231	17227	
21	1-May-02	20	17903	17899	



```
Date,Hour,Total Market Demand,Ontario Demand
01-May-02,1,14141,14137
01-May-02,2,13876,13872
01-May-02,3,13974,13820
01-May-02,4,13898,13744
01-May-02,5,14378,14224
01-May-02,6,15408,15404
01-May-02,7,17070,17066
01-May-02,8,17644,17640
01-May-02,9,17723,17719
01-May-02,10,17774,17770
01-May-02,11,17729,17725
01-May-02,12,17615,17611
01-May-02,13,17530,17526
01-May-02,14,17472,17468
01-May-02,15,17482,17478
01-May-02,16,17592,17588
01-May-02,17,17650,17646
01-May-02,18,17292,17288
01-May-02,19,17231,17227
01-May-02,20,17903,17899
01-May-02,21,17906,17902
01-May-02,22,16614,16610
01-May-02,23,15364,15360
01-May-02,24,14455,14451
02-May-02,1,13945,13941
02-May-02,2,13655,13651
02-May-02,3,13823,13569
02-May-02,4,13774,13610
02-May-02,5,14368,14204
```

What kind of data?

- Relational data - Databases



teacher

PK	teacher_id	int
	first_name	varchar(40)
	last_name	varchar(40)
	language_1	varchar(3)
	language_2	varchar(3)
	dob	date
AK	tax_id	int
	phone_no	varchar(20)

client

PK	client_id	int
	client_name	varchar(40)
	address	varchar(60)
	industry	varchar(20)

participant

PK	participant_id	int
	first_name	varchar(40)
	last_name	varchar(40)
	phone_no	varchar(20)
FK	client	int

course

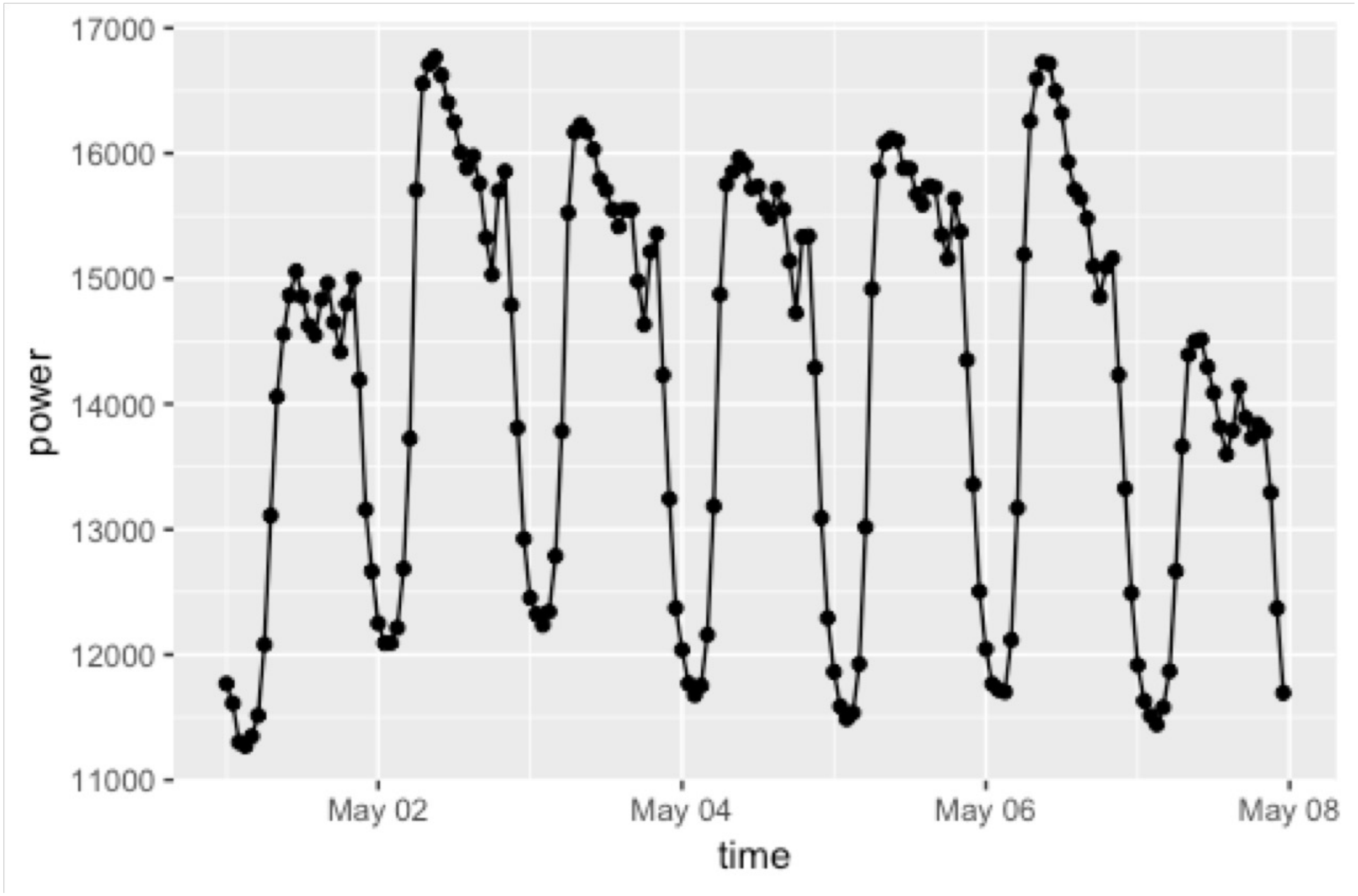
PK	course_id	int
	course_name	varchar(40)
	language	varchar(3)
	level	varchar(2)
	course_length_weeks	int
	start_date	date
	in_school	tinyint

takes_course

PK,FK	participant_id	int
PK,FK	course_id	int

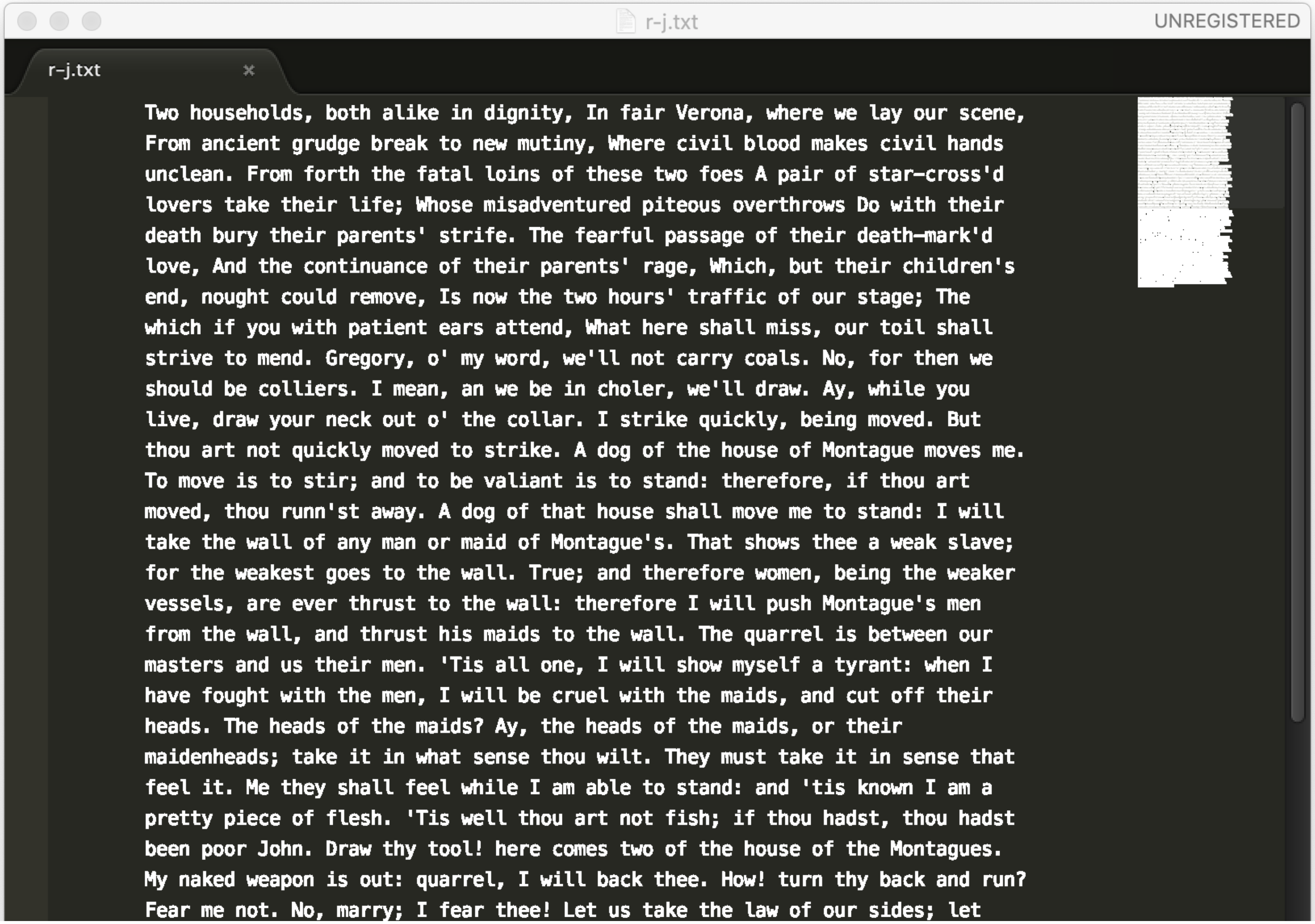
What kind of data?

- Time Series Data



What kind of data?

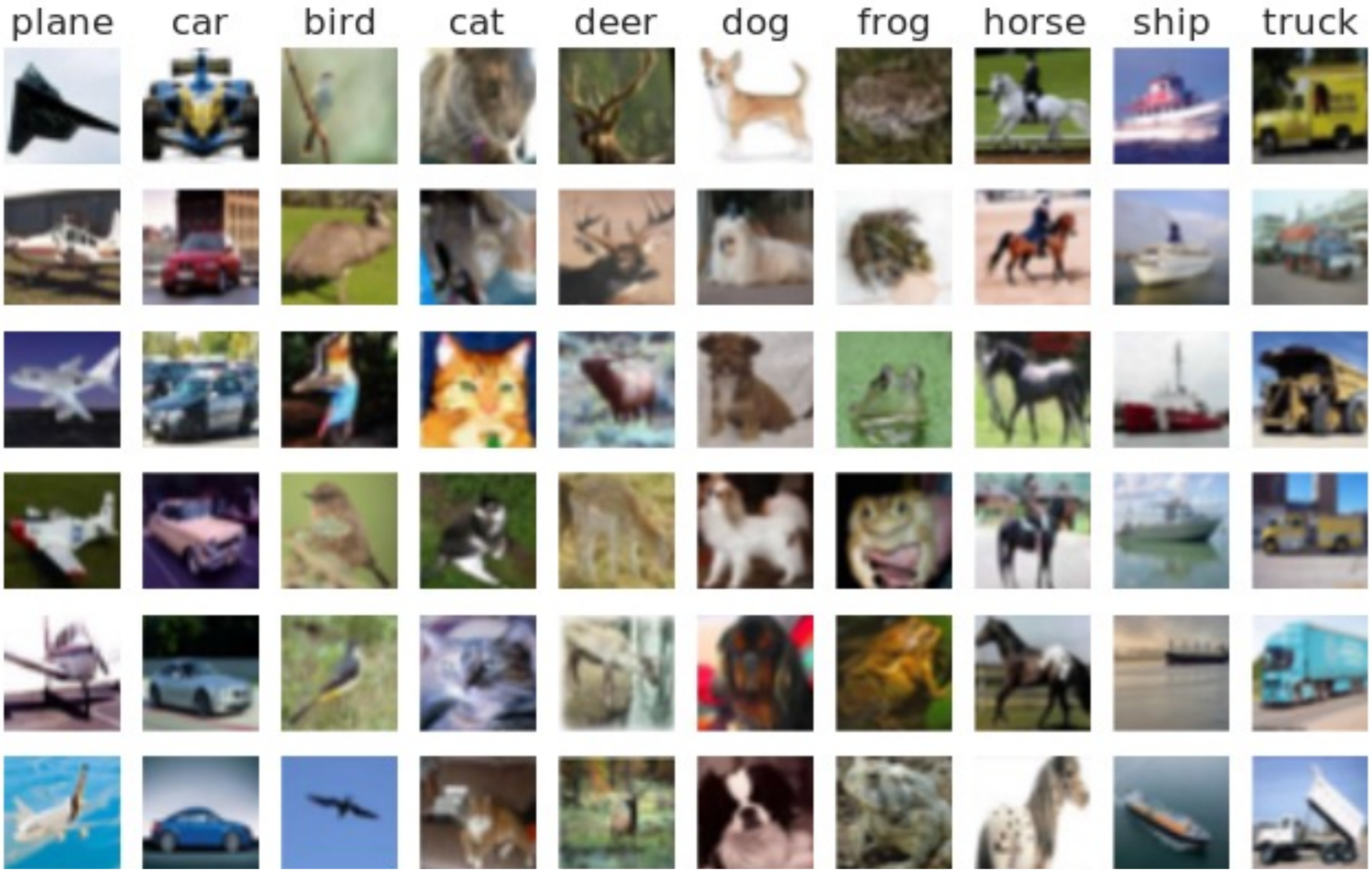
- Text files, document collections



Two households, both alike in dignity, In fair Verona, where we lay our scene, From ancient grudge break to new mutiny, Where civil blood makes civil hands unclean. From forth the fatal loins of these two foes A pair of star-cross'd lovers take their life; Whose misadventured piteous overthrows Do with their death bury their parents' strife. The fearful passage of their death-mark'd love, And the continuance of their parents' rage, Which, but their children's end, nought could remove, Is now the two hours' traffic of our stage; The which if you with patient ears attend, What here shall miss, our toil shall strive to mend. Gregory, o' my word, we'll not carry coals. No, for then we should be colliers. I mean, an we be in choler, we'll draw. Ay, while you live, draw your neck out o' the collar. I strike quickly, being moved. But thou art not quickly moved to strike. A dog of the house of Montague moves me. To move is to stir; and to be valiant is to stand: therefore, if thou art moved, thou runn'st away. A dog of that house shall move me to stand: I will take the wall of any man or maid of Montague's. That shows thee a weak slave; for the weakest goes to the wall. True; and therefore women, being the weaker vessels, are ever thrust to the wall: therefore I will push Montague's men from the wall, and thrust his maids to the wall. The quarrel is between our masters and us their men. 'Tis all one, I will show myself a tyrant: when I have fought with the men, I will be cruel with the maids, and cut off their heads. The heads of the maids? Ay, the heads of the maids, or their maidenheads; take it in what sense thou wilt. They must take it in sense that feel it. Me they shall feel while I am able to stand: and 'tis known I am a pretty piece of flesh. 'Tis well thou art not fish; if thou hadst, thou hadst been poor John. Draw thy tool! here comes two of the house of the Montagues. My naked weapon is out: quarrel, I will back thee. How! turn thy back and run? Fear me not. No, marry; I fear thee! Let us take the law of our sides; let

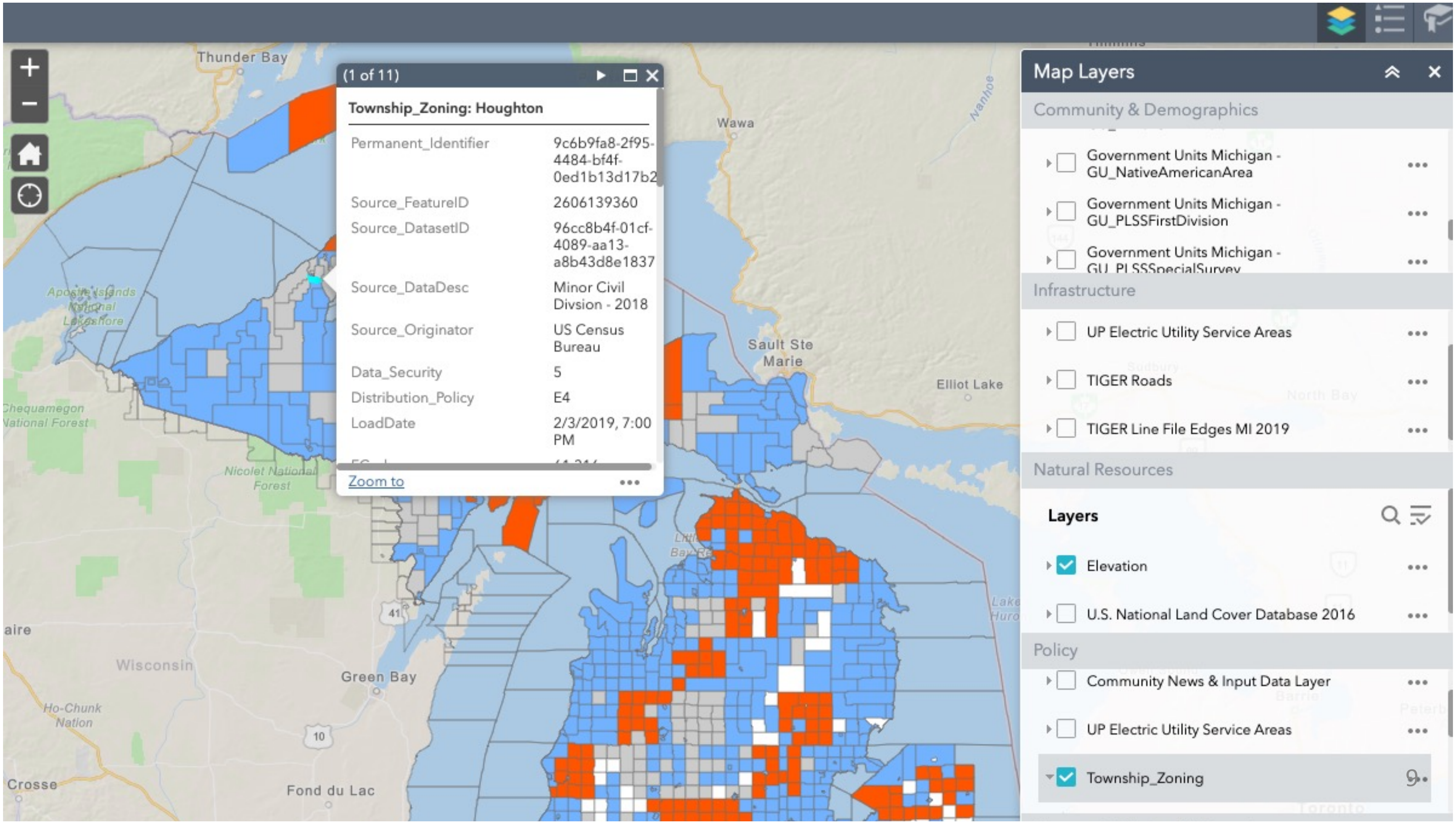
What kind of data?

- Image data



What kind of data?

- Spatial Data



(1 of 11)

Township_Zoning: Houghton

Permanent_Identifier	9c6b9fa8-2f95-4484-bf4f-0ed1b13d17b2
Source_FeatureID	2606139360
Source_DatasetID	96cc8b4f-01cf-4089-aa13-a8b43d8e1837
Source_DataDesc	Minor Civil Divsion - 2018
Source_Originator	US Census Bureau
Data_Security	5
Distribution_Policy	E4
LoadDate	2/3/2019, 7:00 PM

Map Layers

Community & Demographics

- □ Government Units Michigan - GU_NativeAmericanArea
- □ Government Units Michigan - GU_PLSSFirstDivision
- □ Government Units Michigan - GU_PLSSSpecialSurvey

Infrastructure

- □ UP Electric Utility Service Areas
- □ TIGER Roads
- □ TIGER Line File Edges MI 2019

Natural Resources

Layers

- ☑ Elevation
- □ U.S. National Land Cover Database 2016

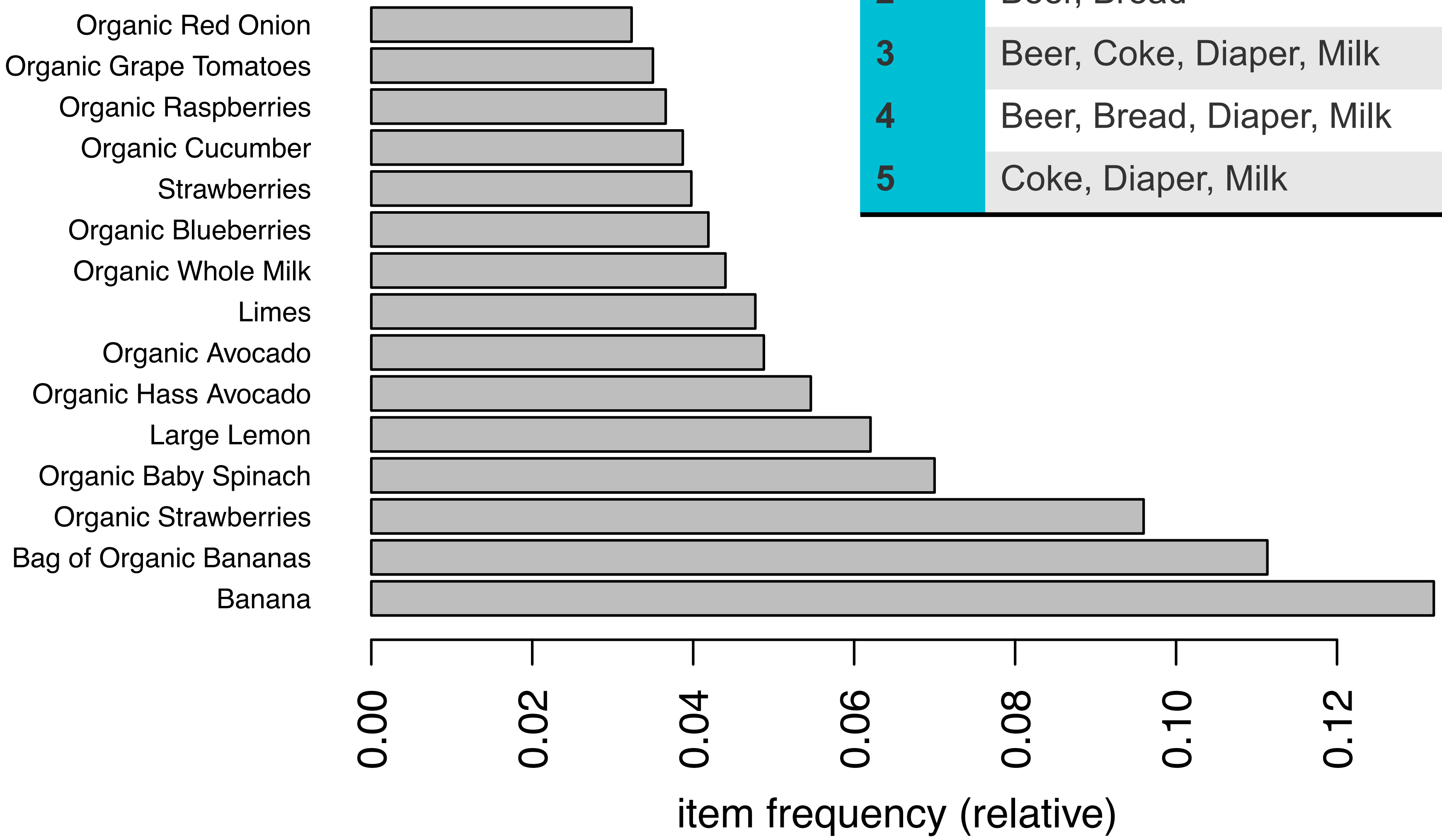
Policy

- □ Community News & Input Data Layer
- □ UP Electric Utility Service Areas
- ☑ Township_Zoning

What kind of data?

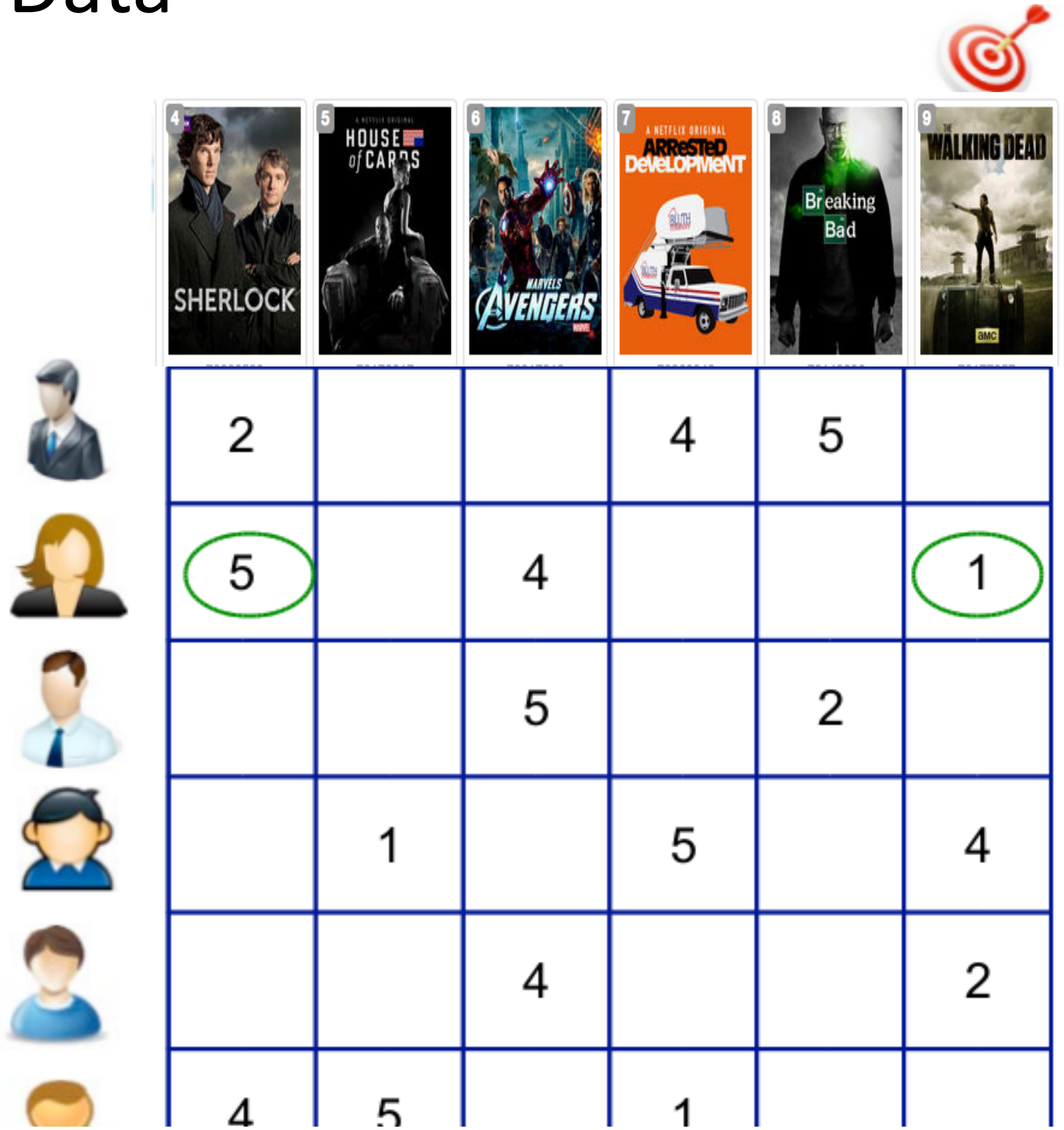
- Transaction Data

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



What kind of data?

- Ratings Data



	SHERLOCK	HOUSE of CARDS	AVENGERS	ARRESTED DEVELOPMENT	Breaking Bad	WALKING DEAD
	2			4	5	
	5		4			1
			5		2	
		1		5		4
			4			2
	4	5		1		

4
5
6
7
8
9
A NETFLIX ORIGINAL
MARVEL'S
AMC

What kinds of data?

- Flat File, Vector data
- Relational data
- Text files, document collections
- Time series data
- Spatial data
- Spatio-temporal data
- Transactional data
- Ratings data
- Network data
- Image data
- Custom data for particular application
- …, and many more

What types of patterns?

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables

- Descriptive Methods
  - Find human-interpretable patterns that describe the data

What type of patterns?

- Generalization / Characterization
- Classification
- Association and Correlation Analysis
- Cluster Analysis
- Recommender Systems
- Structure / Network Analysis
- Outlier Analysis
- Sequential Pattern Analysis
- …

What kinds of techniques?

- Confluence of techniques and disciplines
  - Statistics
  - Machine Learning
  - Database technology
  - Algorithms
  - Pattern Recognition
  - Parallel Computing
  - Visualization

What kinds of applications?

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Example: Fraud Detection

- Which credit card transactions are fraudulent?
- Goal: Predict fraudulent cases in credit card transactions.
  - Data: credit card transactions, information on account-holder
  - Pattern: classification
    - Labeled transaction data: fair or fraud
    - Learn a model to prediction this label
  - Technique: machine learning

Example: Fraud Detection (2)

- Which credit card transactions are fraudulent?
- Goal: Predict fraudulent cases in credit card transactions.
  - Data: credit card transactions, information on account-holder
  - Pattern: outlier analysis
    - Fraud events are rare (hopefully!)
    - Detect transaction as unusual from prior historical data
  - Technique: machine learning, statistics

Example: Clustering

- How should a set of images be placed into groups?
    - Data: image files
    - Pattern: cluster analysis
        - Group images by similarity
        - How to measure similarity?
    - Techniques: machine learning

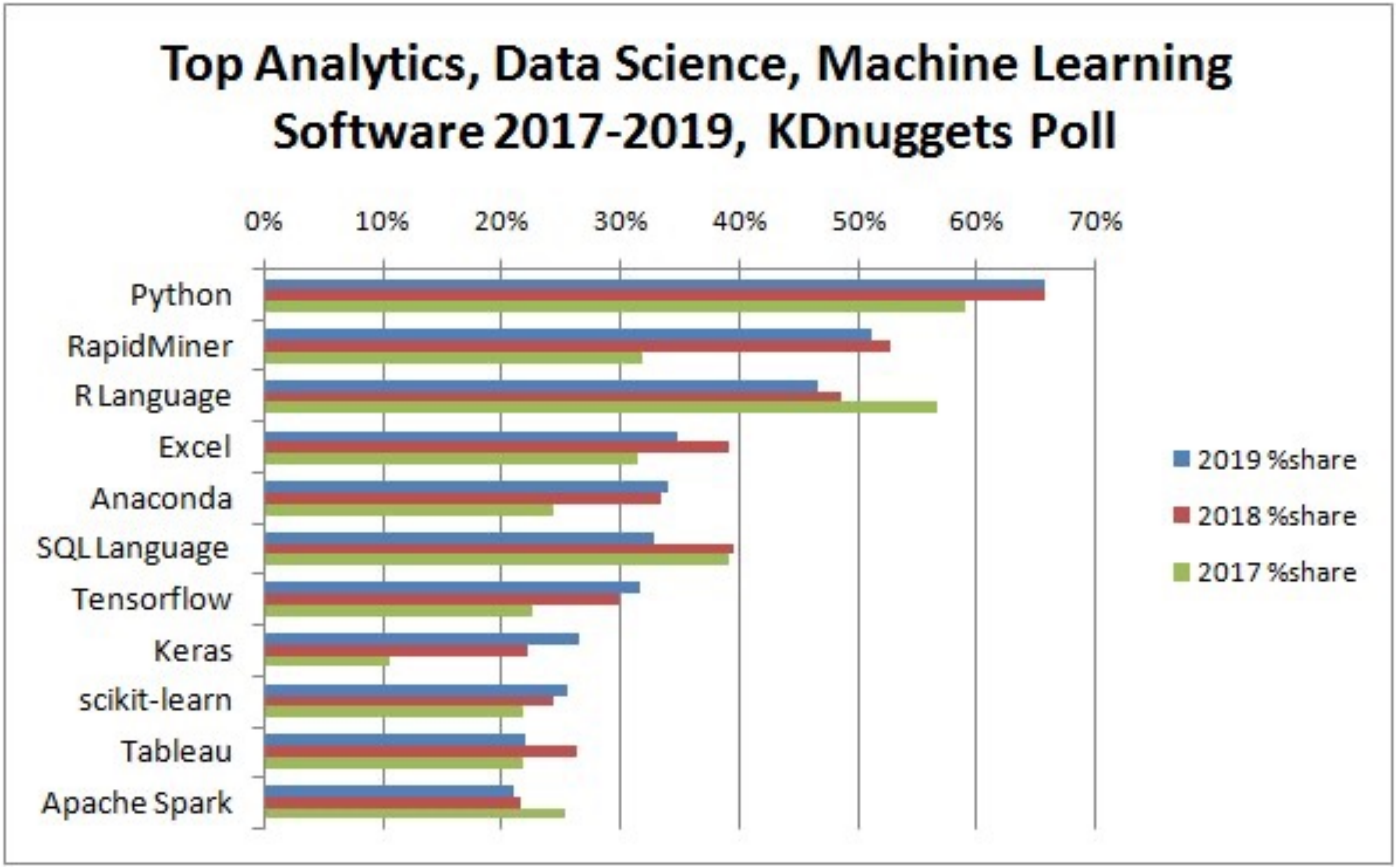
In this class …

- You will study algorithms that exploit and reveal patterns in data.
- Goal:
    - You will learn how to think about problems in data mining
    - You will learn about a set of data analysis tools:
        - How to use them
        - What their assumptions are
        - The capabilities and limitations

Methods to Be Examined

- Supervised learning – Classification
- Unsupervised learning – Data Reduction
- Text Mining
- Unsupervised learning – Clustering
- Association Mining
- Recommendation Systems
- Web Mining (if time permits)

Popular Tools for Data Mining



https://www.kdnuggets.com/polls/

Types of Tools

Simple graphical user interface

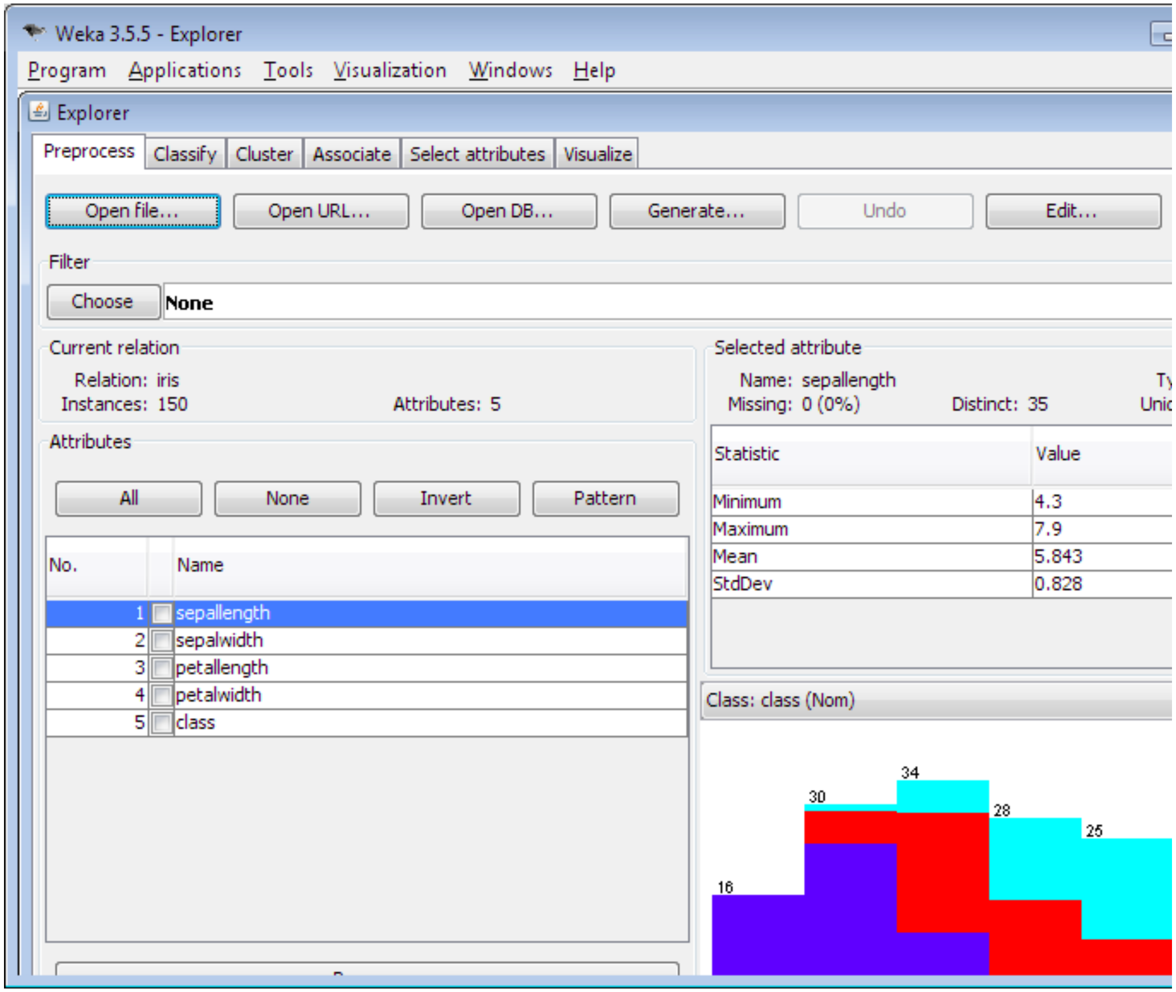
Process oriented

Programming oriented

Tools: Simple GUI

- Weka: Waikato Environment for Knowledge Analysis (Java API)

- Rattle: GUI for Data Mining using R


Weka 3.5.5 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit...
Filter
Choose
All None Invert Pattern
16
30
34
28
25

Filter: None

Current relation

Relation: iris
Instances: 150 Attributes: 5

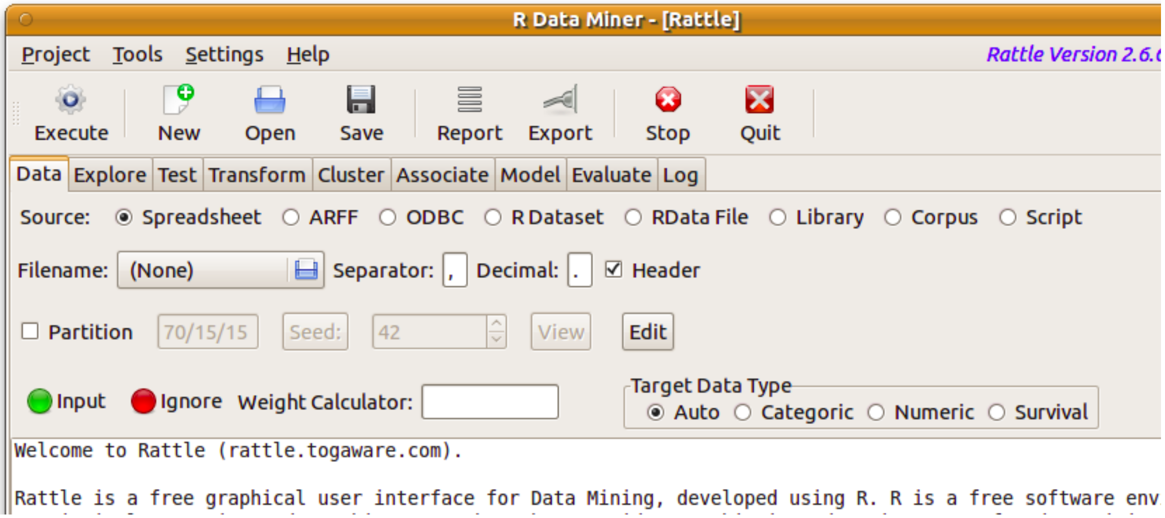
Selected attribute

Name: sepallength
Missing: 0 (0%) Distinct: 35

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom)


R Data Miner - [Rattle]
Project Tools Settings Help
Rattle Version 2.6.
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
View Edit

Source: ● Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: (None) Separator: , Decimal: . ☑ Header

□ Partition 70/15/15 Seed: 42

● Input ● Ignore Weight Calculator:

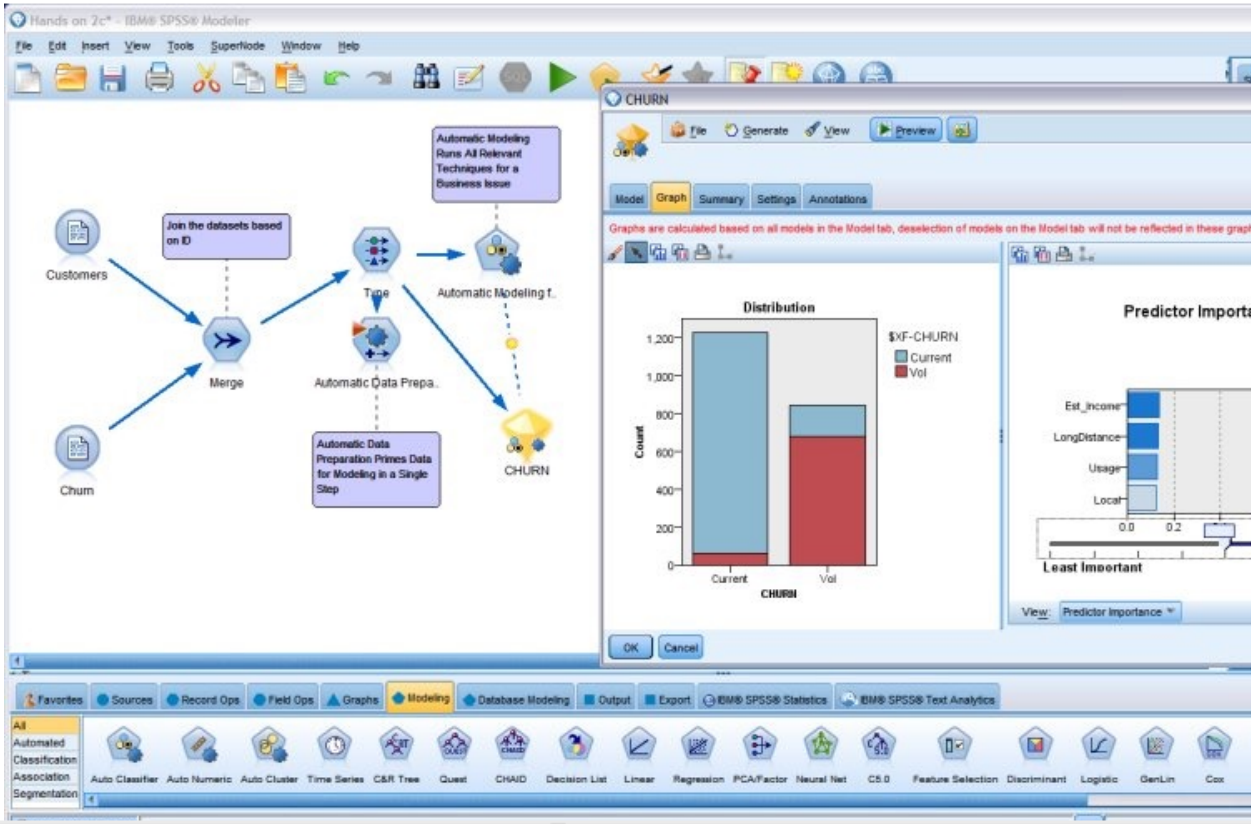
Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

```
Welcome to Rattle (rattle.togaware.com).

Rattle is a free graphical user interface for Data Mining, developed using R. R is a free software env
```

Tools: Process oriented

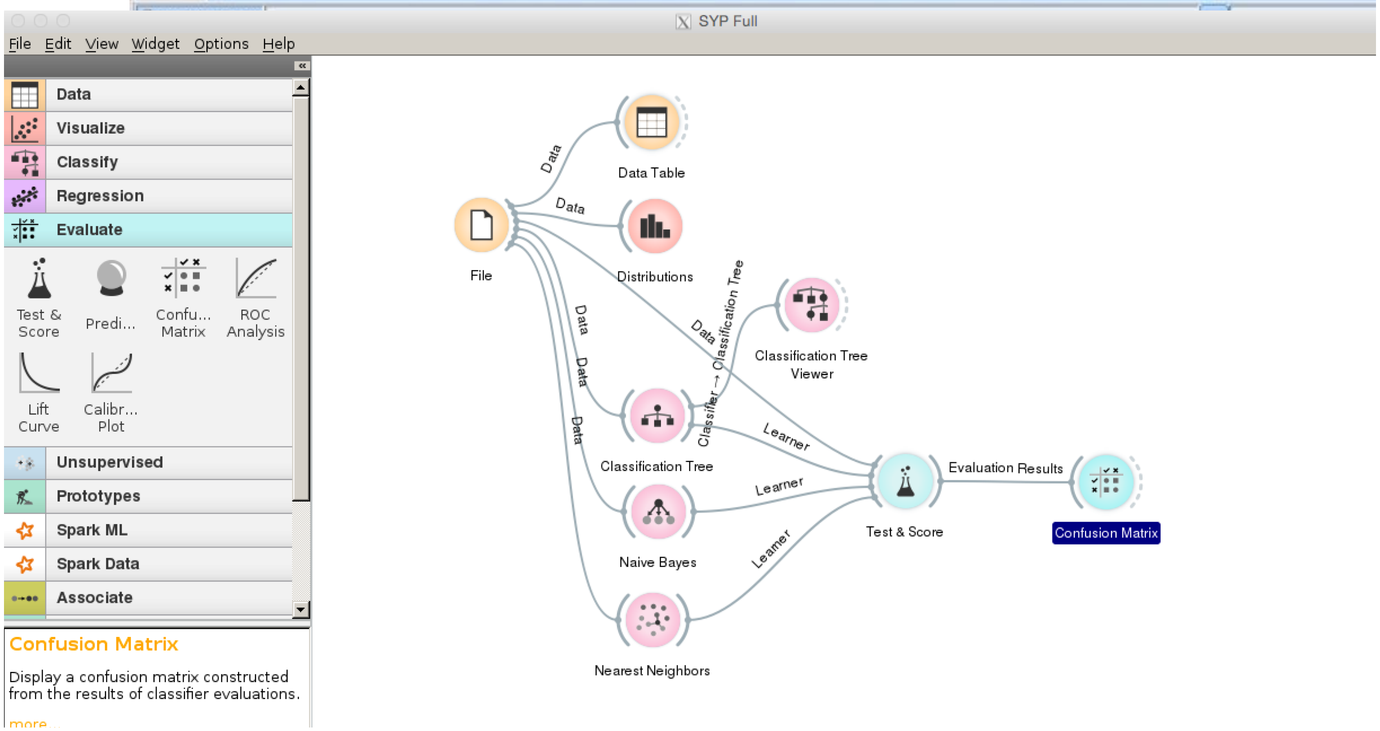
- SAS Enterprise Miner
- IBM SPSS Modeler
- RapidMiner
- Knime
- Orange


Join the datasets based on ID

Automatic Modeling Runs All Relevant Techniques for a Business Issue

Automatic Data Preparation Primes Data for Modeling in a Single Step

Graphs are calculated based on all models in the Model tab, deselection of models on the Model tab will not be reflected in these graph


Confusion Matrix

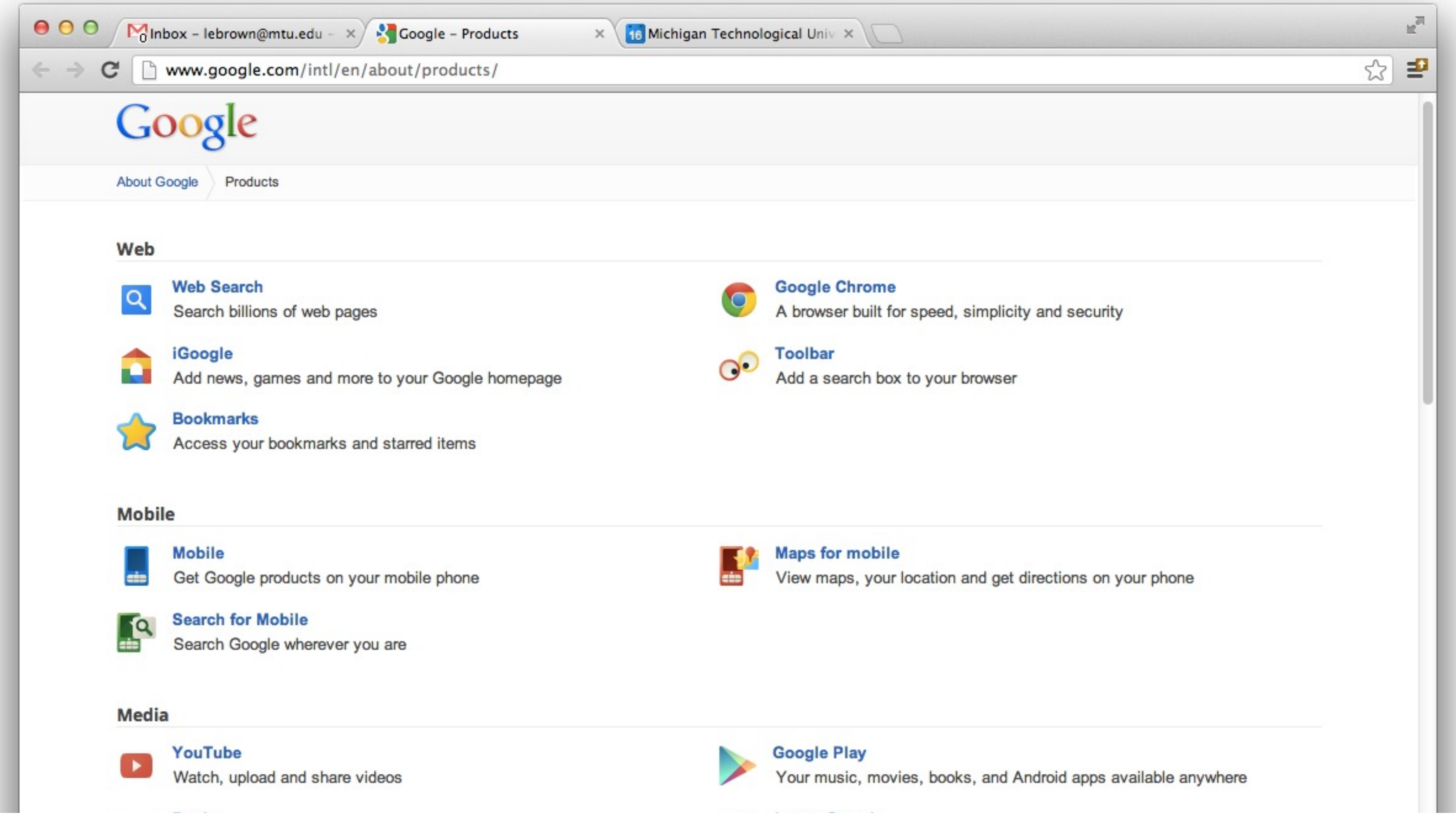
Display a confusion matrix constructed from the results of classifier evaluations.

Tools: Programming oriented

- R
  - RStudio IDE
    - Caret
    - Tidyverse
- Python
  - Jupyter notebooks
    - Numpy
    - Pandas
    - Scikit-learn

Data Mining vs. Privacy

- Tension between data mining and personal privacy



Web

Web Search
Search billions of web pages

Google Chrome
A browser built for speed, simplicity and security

iGoogle
Add news, games and more to your Google homepage

Toolbar
Add a search box to your browser

Bookmarks
Access your bookmarks and starred items

Mobile

Mobile
Get Google products on your mobile phone

Maps for mobile
View maps, your location and get directions on your phone

Search for Mobile
Search Google wherever you are

Media

YouTube
Watch, upload and share videos

Google Play
Your music, movies, books, and Android apps available anywhere

Data Mining vs. Privacy

- How can we leverage sensitive personal data for research / commercial purposes?

- 3 cases
  - AOL search data set
  - Netflix prize
  - Barabasi mobile study

Examples from C. Volinsky

Case 1: AOL Search Data

- Aug. 4, 2006 – AOL releases 20M search terms by anonymized users “for research purposes”
- Within hours, uproar on blogs
  - “The utter stupidity of this is staggering” – TechCrunch
- Aug. 7, 2006 – AOL removes data, issues apology
  - “this was a screw-up, and we are angry”
  - “an innocent enough attempt to reach out the the research community”
- Aug. 9, 2006 – NYT front page story
  - Identifies user
- Aug. 21, 2006 – CTO resigns

Case 1: AOL Search Data

- What’s the big deal?
  - How and why people search is often personal and may contain information they do not want released to public
- What went wrong?
  - Not well thought out
  - Poor internal controls
  - Lack of understanding on anonymizing
- Fallout
  - CTO + at least two others fired
  - Data is still out there

Case 2: Netflix Prize

- Oct. 2006: Netflix released anonymized movie ratings from its customer database
  - 100M ratings, 500K customers (<10% of data)
  - Random integer for user ID
  - “some of the rating data for some customers in the training and qualifying sets have been deliberately perturbed in one or more of the following ways: deleting ratings; inserting alternative ratings and dates; and modifying rating dates”
- 2007, Paper claiming to de-anonymize Netflix data

Case 2: Netflix Prize

- Narayanan and Shmatikov
  - “The adversary with a small amount of background knowledge about an individual … can identify with high probability that individual’s record in the data and learn … sensitive attributes”
  - Claim Netflix’s data sanitization not relevant
  - Basic Idea:
    - With aux info on 8 movies, where 2 can be wrong, and dates are known within 14 days, 99% de-anonymization
    - Aux info can come from other web-sites (IMDB), personal contact, etc.

Case 2: Netflix Prize

- Much ado about nothing
  - Paper is technically correct, but dates are key
  - Without dates, you must know 8 movies, all outside the top 500 to get over 80% chance of de-anonymization
  - Aux info is not easy to come by for many people
  - No identities released
- Netflix did it right
  - Consulted with top machine learning experts
  - Gained new knowledge in machine learning and also privacy fields
- Fallout
  - Netflix was planning another challenge was canceled to due privacy concerns

Case 3: Barabasi Mobile Study

Gonzalez, Hidalgo, and Barabasi (2008)

- Article in Nature outlines study on human mobility patterns
  - 100000 individuals selected randomly from dataset of 6 million
  - Unidentified country (unclear if researchers knew)
  - Cell tower location at start of call
  - 206 individuals were “pinged” every two hours for a week
- Findings
  - “humans follow simple, reproducible patterns”
  - Sample finding: Nearly three-quarters of those studied mainly stayed in a 20-mile circle for half a year.
  - Results “could impact all phenomena driven by human mobility, from epidemic prevention to emergency response and urban planning”

Case 3: Barabasi Mobile Study

- Uproar over “secret tracking” of cell phone users
  - Blowback of negative feedback to Nature and scientists
  - Study would be “illegal in the US”
  - Approval from ONR review board and Northeastern review board. Barabasi did not check with an “ethics panel”
- Response
  - Hidalgo: “the data could be misused”, but we were “not trying to do evil things. We are trying to make the world a little better.”
  - Northeastern and Nature backed the research
  - Continues to be referenced as an example of dangerous research
  - Risk and reward both very high

Data Mining and Ethics

- Privacy is not the only issue in data mining
  - Selling of Data
  - Transparency of Data Collection
  - Security
  - Discrimination and Bias

- We will explore these topics periodically throughout the semester

Data Mining Introduction Summary

Data Mining is interdisciplinary and overlaps significantly with many fields

- Statistics
- CS (machine learning, AI, data bases)
- Optimization

Data Mining requires a team effort with members who have expertise in several areas

- Data management
- Statistics
- Programming
- Communication
- + Application domain