CSCI 6443: Data Mining

Instructor: Paul Melby

Outline

Block 1 (6:10-7:20)

- Whirlwind tour of data mining: what is it, why it is important
- Course logistics
- Programming environment setup

Break (7:20-7:30)

Block 2: (7:30-8:40)

- Complete programming environment setup
- Exploratory Data Analysis, part 1

Introduction to Data Mining

Outline

- We are awash with data
- What is data mining?
- Data Mining Process
- Techniques and Skills
  - Data Wrangling
  - Stats and ML Methods
  - Monitoring/Validation
- Summary

So much data, so little time...

12B transactions/year[1]

Purchase Volume on Capital One Credit Cards, 2023

(Based on $620B purchase volume and $50 average transaction)

88B radar hits per year[2]

Flights handled by the FAA per year.

Estimated from 16.5M flights per year with a 1.5 hour average duration and 1 radar hit per second

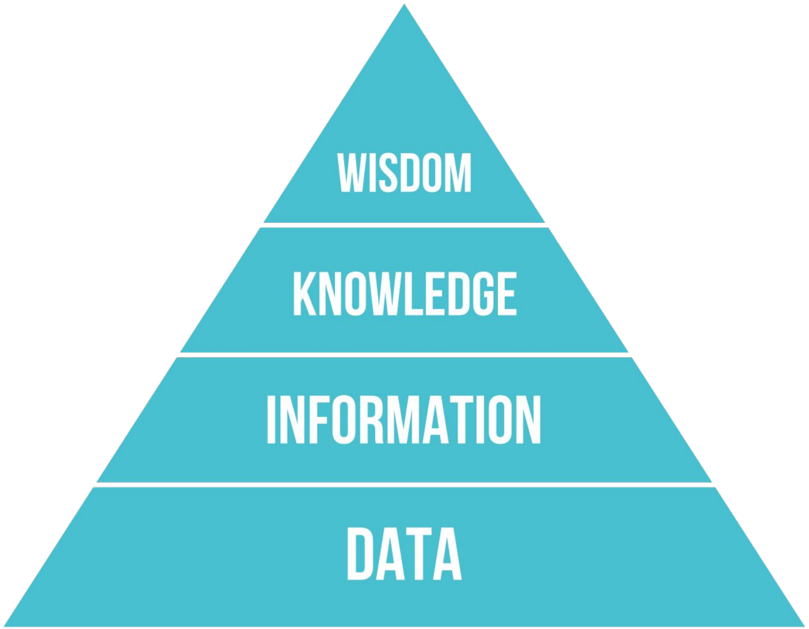
16B Webpages
2.1PB Data[3]

Common Crawl data set from just the past 1 year of crawls

We are faced with massive amounts of data, but want *knowledge* to make decisions

1. Capital One 2023 Annual Report
2. FAA by the numbers
3. Common Crawl Blog

How do we derive Knowledge from Data?



DIKW pyramid

When faced with massive data sets, one answer is the process of **Data Mining,** also sometimes referred to as **Knowledge Discovery from Data (KDD)**

Defining Data Mining[1]:

> *"Data Mining is the process of discovering interesting patterns, models, and other kinds of knowledge in large data sets"*

1.Han, Pei & Tong, (2022). "Data Mining: Concepts and Techniques", Morgan Kaufmann (link)

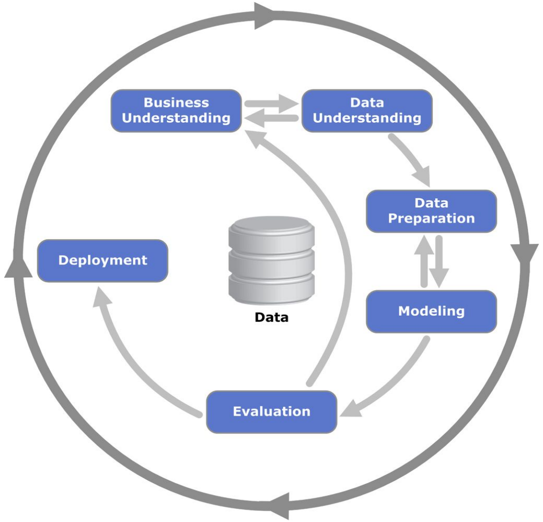
How does Data Mining fit into the larger picture?

Data Mining:

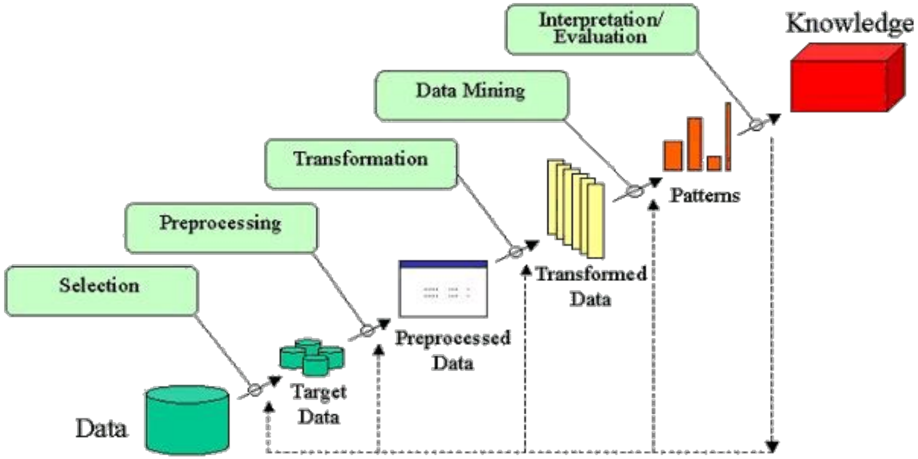
- Uses tools for processing large amounts of data, such as **databases**, **data warehouses**, or **“Big Data” systems**
- Uses **statistics**, **machine learning, visualization**, and other analytical methods as tools for deriving knowledge from data

Can be viewed as a ***process*** or ***collection of practices*** that are part of the ***field*** of **Data Science**

Data Mining Process Models



CRISP-DM



KDD Process

There are a number of models that describe the data mining process.

- These can be useful guidelines to provide structure
- However, just like with the “Scientific Method”, the work is often “messier” or iterative in practice

“All data is dirty, some data is useful”

Getting access to the data, understanding it, cleaning it, processing it and getting it ready for analysis is a very large part of the job

- Getting good at this requires skills (SQL, programming, analysis) and a healthy dose of skepticism (typically built on experience)

The data can be anywhere:

- Flat files, Cloud Storage, Data Lake, Database, Data Warehouse, APIs, Streaming system, or even the internet

The data can be dirty in all kinds of different ways:

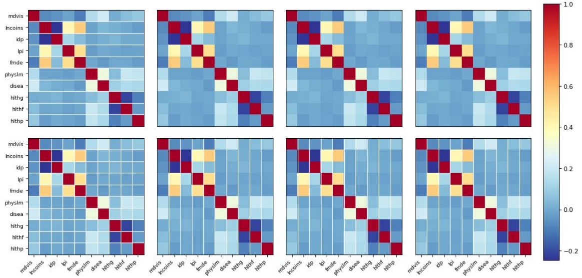
- Missing data, corrupted values, time lags, inconsistent or no schema, wrong data-type

The stakes can be high:

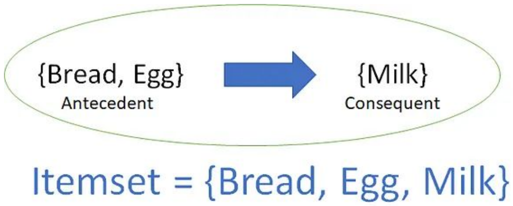
- “Data” related errors are the #1 source of “model errors” in most production systems

Finding relationships within your data

Correlation[1]



Association Rules[2]



Many times, analysis needs to start with

- **Correlation Analysis**: finds (linear) relationships between variables in your data
- **Co-occurrence Analysis:** finds how often events/items occur together. Includes methods such as frequent pattern mining or association rules

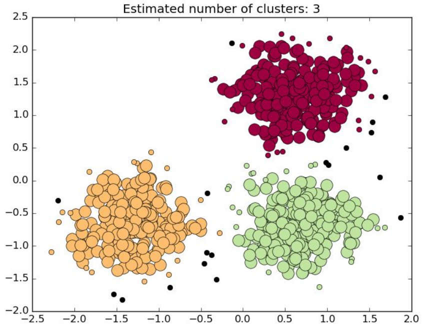
Motivating Examples:

- Finding redundant variables or identifying variables with a strong relationship to a target variable (Correlation Analysis)
- Finding products that people frequently buy together - perhaps as a building block for a recommendation engine (Association Rules)

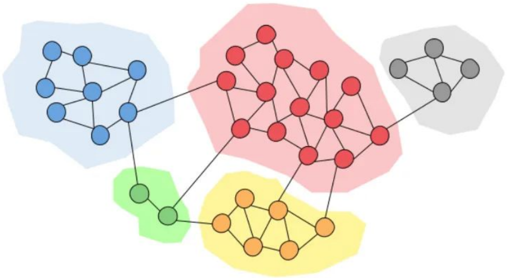
1. https://www.statsmodels.org/dev/generated/statsmodels.graphics.correlation.plot_corr_grid.html
2. https://towardsdatascience.com/association-rules-2-aa9a77241654

Dividing your data into groups

Clustering



Community Detection



There are many classes of methods to try to discover the most natural groupings or partitions of your data, such as:

- **Clustering** (for record oriented data)
- **Community Detection** (for graph oriented data)
- **Topic Modeling** (for record oriented data, esp. text)

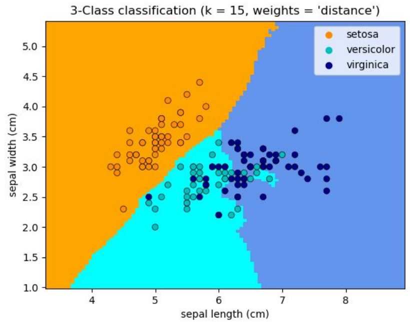
Motivating Examples:

- Finding "customer archetypes" (clustering)
- Discovering "fraud rings" (community detection)

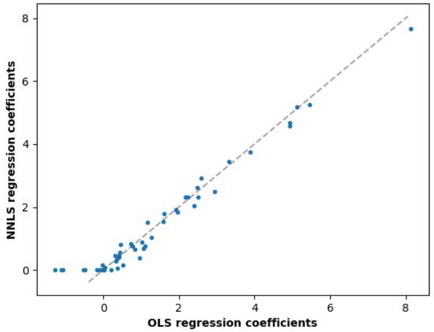
1. https://scikit-learn.org/0.17/auto_examples/cluster/plot_dbscan.html
2. https://towardsdatascience.com/community-detection-algorithms-9bd8951e7dae

Labeling your data and making predictions

Classification[1]



Regression[2]



Supervised methods help you make predictions or label data according to “known” categories

- **Classification** (for discrete predictions/categories)
- **Regression** (for graph oriented data)

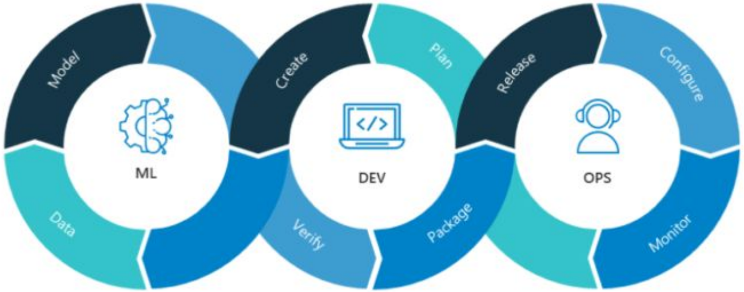
Motivating Examples:

- Predicting if a transaction is fraud (classification)
- Labeling documents by category (multi-class classification)
- Forecasting customer lifetime value (regression)

1. https://scikit-learn.org/1.2/auto_examples/neighbors/plot_classification.html
2. https://scikit-learn.org/stable/auto_examples/linear_model/plot_nnls.html#sphx-glr-auto-examples-linear-model-plot-nnls-py

Validation and Monitoring:
ML Ops and Model Risk Management

ML Ops Lifecycle[1]



Validation:

- How do you know that the insights discovered are "real"?
- How can you trust a model to make automated decisions?

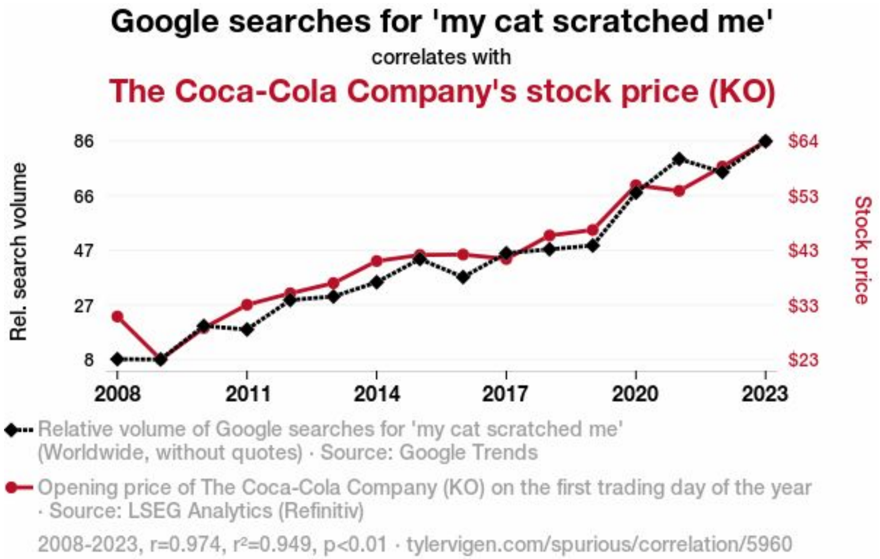
Monitoring:

- What happens to your model the data goes bad?
- What if the fraudsters change their tactics?

These questions are addressed in the emerging set of practices called **ML Ops**, but the principles have existed as part of a larger framework of "**model risk management**" for much longer[2].

1. https://blogs.nvidia.com/blog/what-is-mlops/
2. See, for example: https://www.occ.treas.gov/news-issuances/bulletins/2011/bulletin-2011-12a.pdf

Caveat emptor: Data Mining done “wrong”



https://www.tylervigen.com/spurious/correlation/5960

“p-value hacking”, “data dredging”

- Doing enough comparisons, you can almost always find *some* pattern, but it may meaningless
- See also: https://xkcd.com/882/

Even more reason to ***understand the methods*** you are using and to ***draw sound conclusions*!**

Ethical considerations

Privacy concerns:

- Our search and browsing history, social media posts and all kinds of personal information can be available to “data miners” in various industries
- Thoughtful consideration of the collection, storage and use of personal information is required

Discrimination & Bias:

- Data reflects the process that created it
- Because of a history of discrimination in society, we might find patterns that indicate that race or gender are correlated with creditworthiness, but we must guard against propagating these discriminatory practices by ‘repeating history’ with our models

This goes much deeper!

Summary

Data Mining is the process of extracting knowledge, patterns or models from data

The results can be incredibly valuable to organizations:

- Commercial: Increase safety, improve customer satisfaction, drive revenue, decrease fraud or other costs
- Non-commercial: Analyze large-scale scientific data (astronomical, biological, etc.), form hypotheses

It requires a diverse set of skills to perform well, including Data Wrangling, Stats & ML, and a healthy dose of skepticism

Course Logistics

Goals of the Course

After completing this course, you should be able to independently execute a data mining project, including:

1. **Matching the goals** of the project to appropriate data, algorithms and metrics/objective-functions
2. **Executing** the project, using an appropriate choice of technology & tools in a manner that is reproducible by other "data miners"
3. **Present** your results to both technical and non-technical audiences in a manner that communicates the key points and relative success of the project
4. Be aware of common **ethical concerns** that may be of concern for the project

Grading

- Homework Assignments: 60%
- Final Project Presentations: 20%
- Final Project Paper and Materials: 20%

Homework will be due at midnight on Wednesdays. Late assignments will receive a 20% penalty.

Turning Homework In

The majority of homework assignments should be written in Python (and SQL), in a Jupyter Notebook format. An important aspect of data mining is reproducibility. I should be able to reproduce any results that you provide. Therefore, when submitting homework, the following should be included:

- The .ipynb file
- An encapsulated version of the notebook with results (tables, plots, executed code), in PDF or HTML format
- A list of required dependencies to independently reproduce the notebook, including:
- Data or link to source of data
- Python version, package dependencies

Note: all notebooks must be run “top to bottom” using the “Restart and Run All” functionality prior to both the creation of HTML/PDF and submission

Programming Environment

- Each lecture will include code-based examples of the methods discussed, in the form of Jupyter notebooks written in Python.
- Homework will also be in the form of Jupyter notebooks.
- Reproducibility is a big concern, so all homework submissions should include:
    - The .ipynb file
    - An encapsulated version of the notebook with results (tables, plots, executed code), in PDF or HTML format
    - A list of required dependencies to independently reproduce the notebook, including:
        - Data or link to source of data (if too large to attach)
        - Python version, package dependencies
- My initial plan is for homework to be submitted through GitHub

Environment Setup Overview

For programming environments, you may choose to use your personal computer. If you would prefer using outside resources, Google Colab provides some amount of free computing resources, with a Jupyter notebook interface. Amazon SageMaker Studio Lab has a similar service.

No matter what environment you choose to use, please remember to take the reproducibility requirements into account. When managing local environments, using an environment manager such as conda (or mini-forge) or PyEnv plus virtualenv to support reproducibility is strongly recommended.

Let’s set a few things up! ➜ optional

1. Setting up a GW GitHub account:
    a. https://ithelp.gwu.edu/en-us/article/1533994
    b. You should also set up SSH keys
2. Let me know your GH username, so I can add you to the team for the course:
    a. https://github.com/orgs/gwuniversity/teams/csci_6443
3. ~~Create a private repo for your homework submissions and add me as a collaborator, so that I can see the code and comment on pull requests~~
    a. My GWU GH username is: paulmelby-gwu

GitHub setup continued

To check out and check in code from GWU's github, you need to set up ssh keys and tie them to your single-sign-on information:

- Set up an ssh key:
  - https://docs.github.com/en/authentication/connecting-to-github-with-ssh/generating-a-new-ssh-key-and-adding-it-to-the-ssh-agent
- Connecting ssh key to your single sign on:
  - https://docs.github.com/en/enterprise-cloud@latest/authentication/authenticating-with-saml-single-sign-on/authorizing-an-ssh-key-for-use-with-saml-single-sign-on

Installing miniforge

Miniforge is the open-source version of the conda package/environment manager. Installers are located here: https://github.com/conda-forge/miniforge

To create an environment for this class, you can use the following command at your terminal:

```
> conda create -n cs6443 python=3.11
```

Once your environment is created, you need to activate it:

```
> conda activate cs6443
```

If you are on a Mac and getting an error with activating an environment, you may need to run `conda init zsh`

You can install the required files to run the class notebooks with:

```
> pip install -r requirements_class1.txt
```

Depending on your system, before pip works, you may need to:

```
>  conda install pip
```

Launching Jupyter

```
> jupyter notebook
```