CSCI 3022

Class Introduction

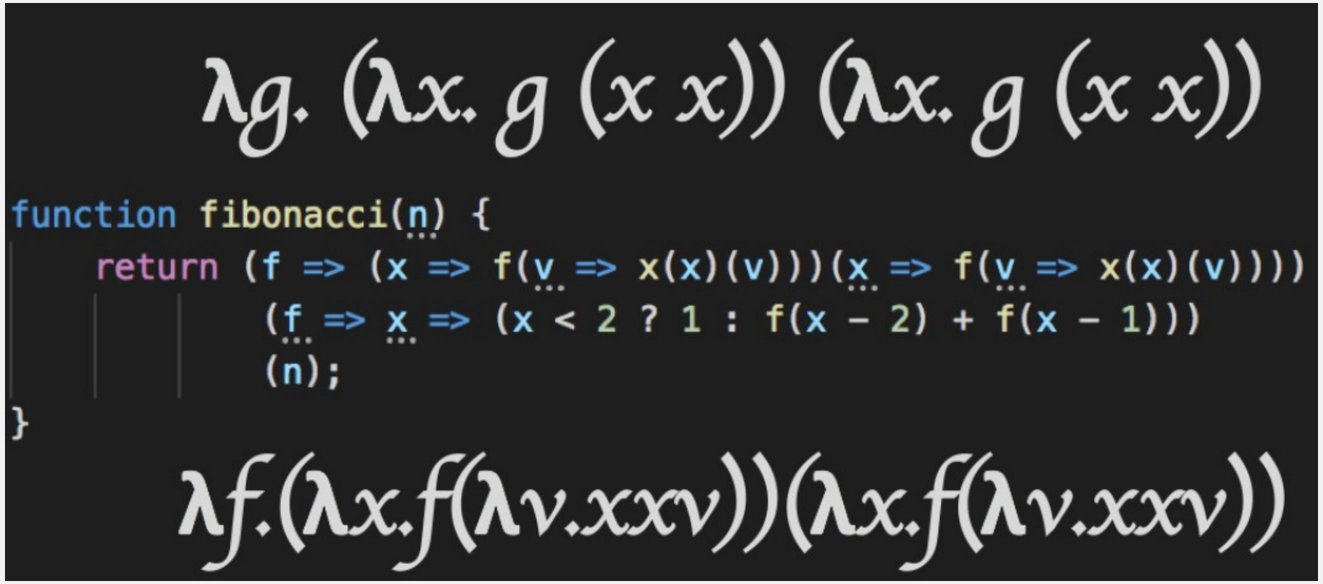
- This is CSCI 3022

  Also known as:

Intro to Data Science with Probability and Statistics

You are here because you are comfortable with coding, data structures, discrete mathematics, and Calculus II.

Now you are going to use those skills to study probability and statistics with Python as the primary tool in that investigation.



$$\lambda g.\ (\lambda x.\ g\ (x\ x))\ (\lambda x.\ g\ (x\ x))$$

```
function fibonacci(n) {
    return (f => (x => f(v => x(x)(v)))(x => f(v => x(x)(v))))
            (f => x => (x < 2 ? 1 : f(x - 2) + f(x - 1)))
            (n);
}
```

$$\lambda f.(\lambda x.f(\lambda v.xxv))(\lambda x.f(\lambda v.xxv))$$

My name is Murray Cox
We will be meeting three days per week:
M, W, and F from 4:10 – 5:00

When you have questions or need help you can look me up during my office hours which are posted in Canvas.

Also, you might find the most efficient source of help to be found in this order:
1] Read the book, the syllabus and/or look at the schedule and Canvas
2] Ask a classmate
3] Post your thoughts on Piazza. Many minds are better than one.
4] We pay the TA's $$$ to be at your service; use them.
5] Lastly, set up an appointment to talk with me

The job of a University is to ‘certify’ their product.
A University wants to be known as having a quality product.
Therefore, we (CU, your instructors, your TA’s, etc) want/need you to pass each class with a high grade and lots of knowledge.

Here is how this class will certify its product (you):
Each student will be assessed with…

12 Quizzes (20 points each)
5 Homeworks (100 points each)
2 written exams (100 points each)
2 coding exams (80 points each)

The more points that you earn, the higher your grade will be.
In December, we will add up your points and divide by 1000.
This percentage (letter grade equiv.) will be sent to the registrar at CU.

I try to keep class due dates constant to help scheduling

6:00 PM MT is an important time for this class.

Every Wednesday there is a quiz at 6:00.

Every other Friday HW is due at 6:00.

99.99% of what you need to know can be found on Canvas.

Your textbook is found there.

Your schedule is found there.

Your HW, Quizzes, and exams are found there.

Anything you hand in is started on Canvas.

LOOK AT CANVAS AND SEE WHAT QUESTIONS ARISE.

CLICK THE LINKS AND EXPLORE THE SITE

CANVAS

Now, Data Science…

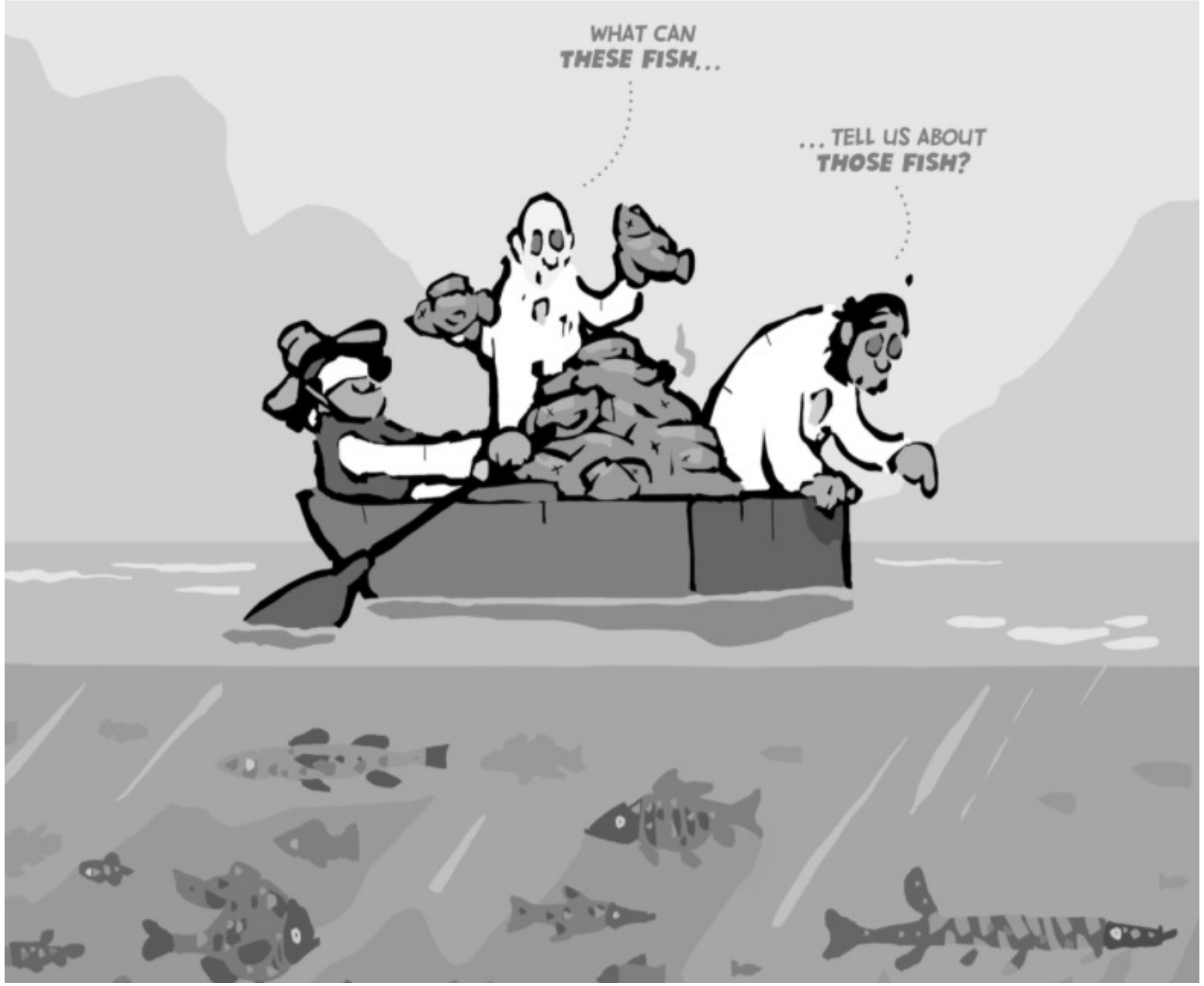
WHAT IS DATA SCIENCE?

Recovering insights & trends hiding within the data

Using data to answer interesting questions

Using data to understand the world around us

Probability & Statistics (Math!)



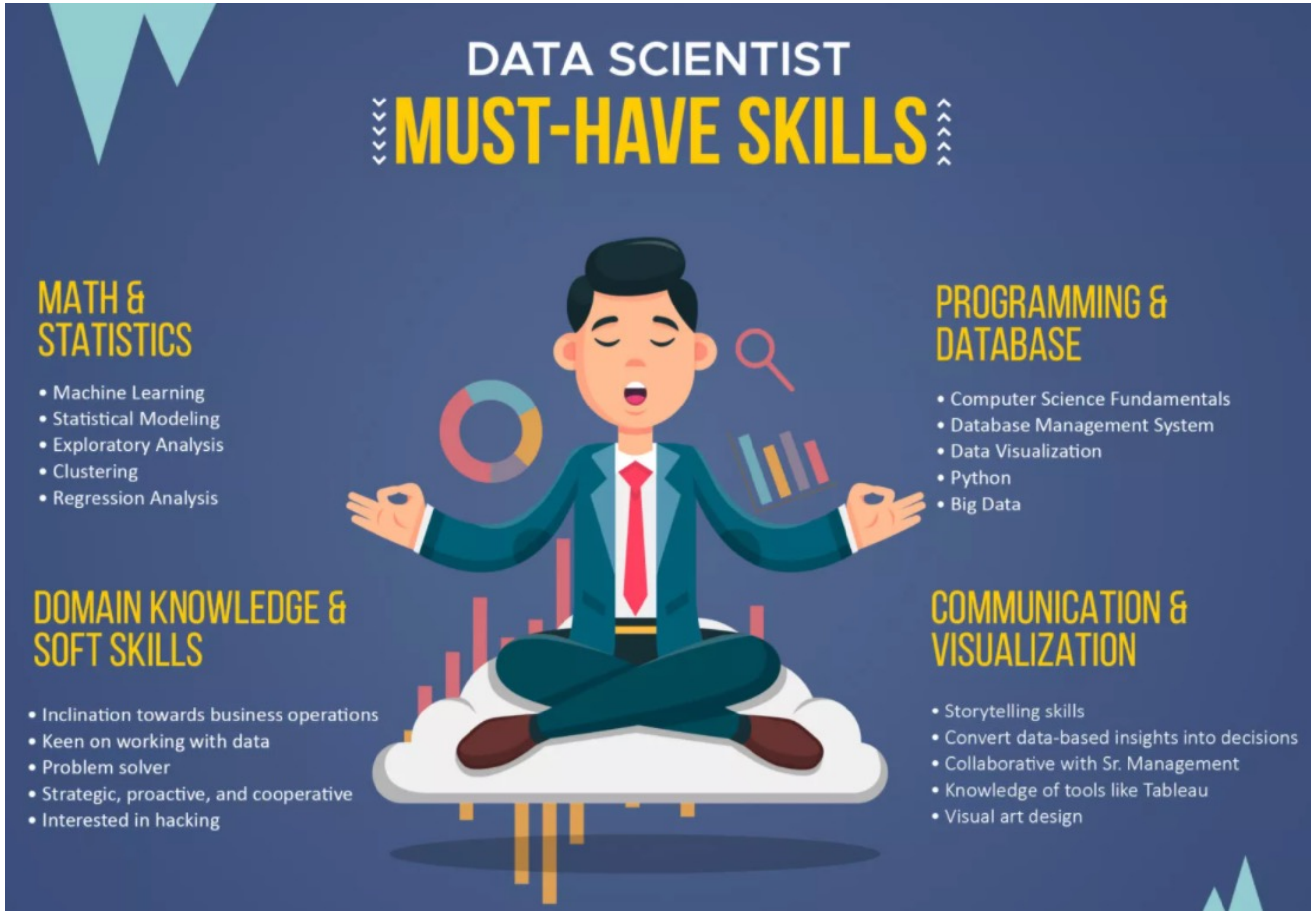
ASK AN INTERESTING QUESTION

GET THE DATA

★ EXPLORE THE DATA

★ MODEL THE DATA

COMMUNICATE AND VISUALIZE THE RESULTS



DATA SCIENTIST

MUST-HAVE SKILLS

MATH & STATISTICS

- Machine Learning
- Statistical Modeling
- Exploratory Analysis
- Clustering
- Regression Analysis

PROGRAMMING & DATABASE

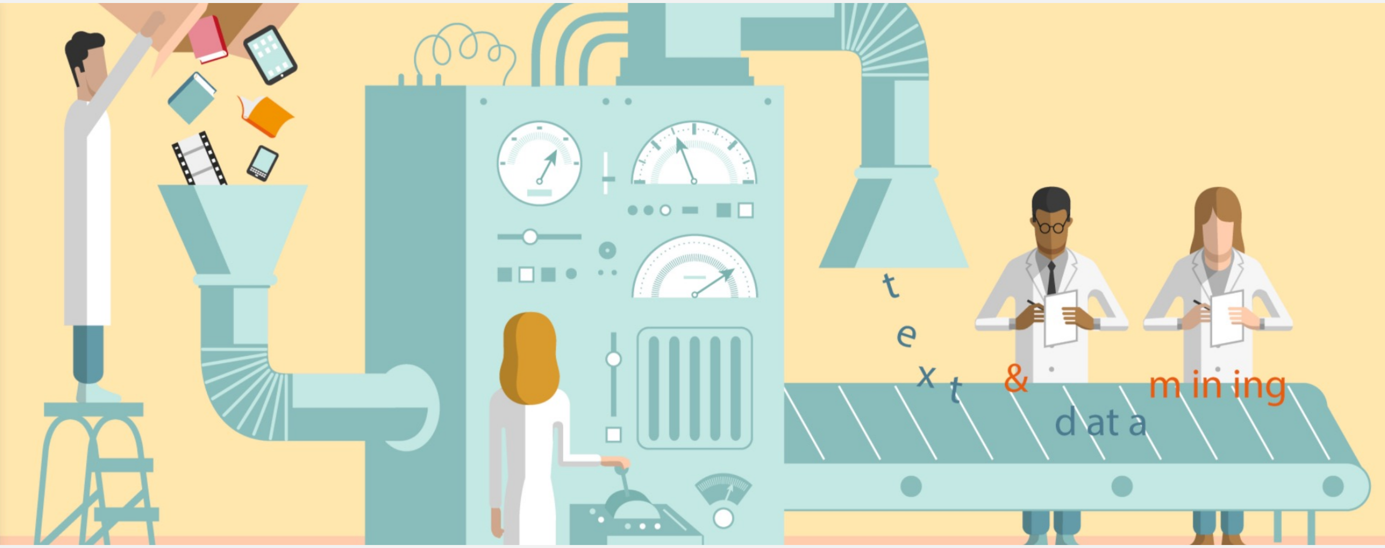
- Computer Science Fundamentals
- Database Management System
- Data Visualization
- Python
- Big Data

DOMAIN KNOWLEDGE & SOFT SKILLS

- Inclination towards business operations
- Keen on working with data
- Problem solver
- Strategic, proactive, and cooperative
- Interested in hacking

COMMUNICATION & VISUALIZATION

- Storytelling skills
- Convert data-based insights into decisions
- Collaborative with Sr. Management
- Knowledge of tools like Tableau
- Visual art design



- In other words:

  Hypothesis → Observations → Analysis → Conclusions → Refinement (do it again)

- In this class we will focus largely on Exploration (Data Mining) and Modeling (Statistical analysis).

- We want to discover something, understand it, and predict the future (rudimentary machine learning.)

- Class topics will be focused on:
- **Exploratory data analysis**
  - Dig into your data, compute simple statistics,
  - draw pictures and look for insight.
-

Cleaning, munging, and wrangling data

  - Some data scientists report they spend
  - up to 80% of their time cleaning data sets.

- I**Probability theory and simulation**
  - Different probability distributions model different types of events.
-

Hypothesis testing and inferential statistics

  - What is the data really telling us and how confident should we be in our conclusions?

-

Modeling and Classification

  - Linear regression, multiple linear regression, logistic regression

When you leave this class, it is my goal that you will have a fluency in the theoretical and computational aspects of data analysis.

You will be able to:

Clean, munge, and wrangle data in Python and perform Exploratory Data Analysis.
Draw insight from data by computing and interpreting classic summary statistics.
Know the ins-and-outs of probability and how to use it to solve real-world problems.
Construct and analyze simple models to make predictions and inferences about data.
Perform statistical tests to determine if your conclusions are real or due to chance.
Tell compelling stories about data using modern visualization and presentation tools.

```
 *
 * @var boolean
 */
define('PSI_INTERNAL_XML', false);

if (version_compare("5.2", PHP_VERSION, ">")) {
    die("PHP 5.2 or greater is required!!!");
}
if (!extension_loaded("pcre")) {
    die("phpSysInfo requires the pcre extension to php in order to w
        properly.");
}

require_once APP_ROOT.'/includes/autoloader.inc.php';

// Load configuration
require_once APP_ROOT.'/config.php';

if (!defined('PSI_CONFIG_FILE') || !defined('PSI_DEBUG')) {
    $tpl = new Template("/templates/html/error_config.html");
    echo $tpl->fetch();
```

Munge

/mənj/

verb

INFORMAL•COMPUTING

verb: **munge**

The act of consolidating 2 or more mutually exclusive data sets or sections of computer code, circumventing the requirement to write a shit load of complex code.

WIKIPEDIA
The Free Encyclopedia

Mung is computer jargon for a series of potentially destructive or irrevocable changes to a piece of data or a file.[1] It is sometimes used for vague data transformation steps that are not yet clear to the speaker.[2] Common munging operations include removing punctuation or HTML tags, data parsing, filtering, and transformation.[2]

The term was coined in 1958 in the Tech Model Railroad Club at the Massachusetts Institute of Technology.[1] In 1960 the backronym "Mash Until No Good" was created to describe Mung, and by 1976 it was revised to "Mung Until No Good", making it one of the first recursive acronyms.[3] It lived on as a recursive command in the editing language TECO.[4]

It differs from the very similar term munge, because munging usually implies destruction of data, while mungeing usually implies modifying data (simple passwords) in order to create protection related to that data.

Munging may also describe the constructive operation of tying together systems and interfaces that were not specifically designed to interoperate (also called 'duct-taping'). Munging can also describe the processing or filtering of raw data into another form.[2]

Data wrangling, sometimes referred to as **data munging**, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. The goal of data wrangling is to assure quality and useful data. Data analysts typically spend the majority of their time in the process of data wrangling compared to the actual analysis of the data.

The process of data wrangling may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses. Data wrangling typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.[1]

Software Specifics

We will use **Python 3** and in particular:
Numpy and **Pandas**

Why? Lots of great data science libraries and decent plotting

We'll work exclusively in Jupyter Notebooks

Easiest way to get Jupyter is **Anaconda Python 3.6**; we strongly recommend you install local copy

If not, you can use **Microsoft Azure** or **Google Colab** notebooks

We will often work on problems during class.
You should attempt to follow along during class, although this may be difficult via distance.
You can also practice before lecture, but the key is to actually do it yourself.
Do not think that learning is watching someone else type on the keys!

Mandatory

1] Homework assignments will be done through Jupyter Notebooks

Install Jupyter Notebook on your computer

2] Back your work up!

Github, Google Drive, SOMEthing

Make the repo **private** (collaboration policy)

Before we meet again:

Be ready for Wednesday's quiz. It covers syllabus material and class logistics.

Get on Canvas and click around, investigate, explore.

Check out the Piazza page, post a question, post an answer, be known.

Install Anaconda (or another reliable Jupyter notebook method)

Review and complete the Numpy/Pandas tutorial

Explore NB00 and NB01