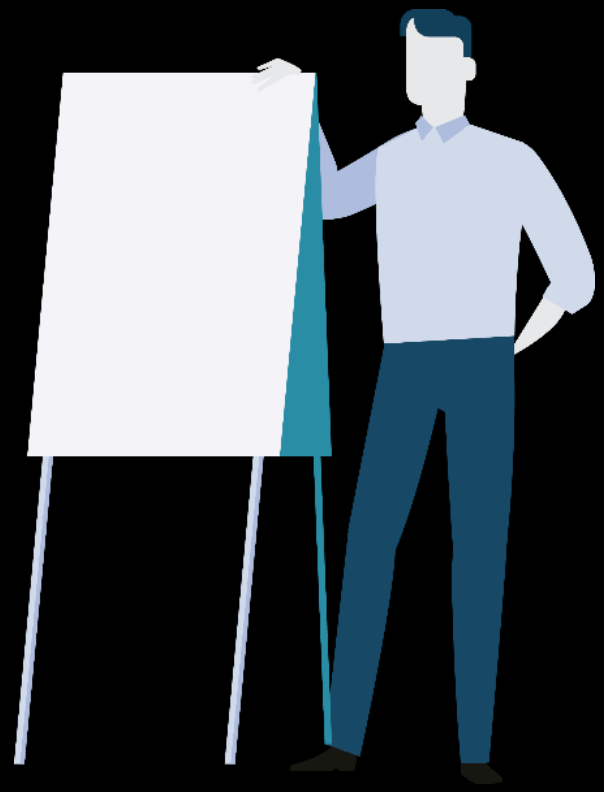
Get The Data

- DSCI100 is a prerequisite for this class.
- In a future version of My MSU, we WILL enforce prerequisites.



Mass Street University
PER EDUCATIONEM PROGRESSUS

Get The Data

- Getting your own data is the best way to get data!
- Getting data can be like a game of telephone.
- Knowing SQL is an absolute requirement to properly function in a self-service data environment.

Mass Street University
PER EDUCATIONEM PROGRESSUS

Get The Data

- Data is stored EVERYWHERE!
- File formats and databases are the most common sources of data.

Mass Street University
PER EDUCATIONEM PROGRESSUS

Get The Data

- Common types of file formats
  - CSV
  - Excel
  - JSON
  - XML

Mass Street University
PER EDUCATIONEM PROGRESSUS

Get The Data

- Uncommon sources of data
  - Super old mainframe file formats

Mass Street University
PER EDUCATIONEM PROGRESSUS

Timeout for a quick digression

Mass Street University

PER EDUCATIONEM PROGRESSUS

Ok maybe not so quick…..

Mass Street University

PER EDUCATIONEM PROGRESSUS

Yes. I’m blabbing. Pay attention.
This is important.

Mass Street University

PER EDUCATIONEM PROGRESSUS

FINALLY!

Mass Street University

PER EDUCATIONEM PROGRESSUS

Get The Data

- Uncommon sources of data
  - Super old mainframe file formats
  - Scientific datasets that don’t conform to common standards
  - HTML files
    - Getting this data is called webscraping
    - Use Beautiful Soup to help parse HTML

Mass Street University
PER EDUCATIONEM PROGRESSUS