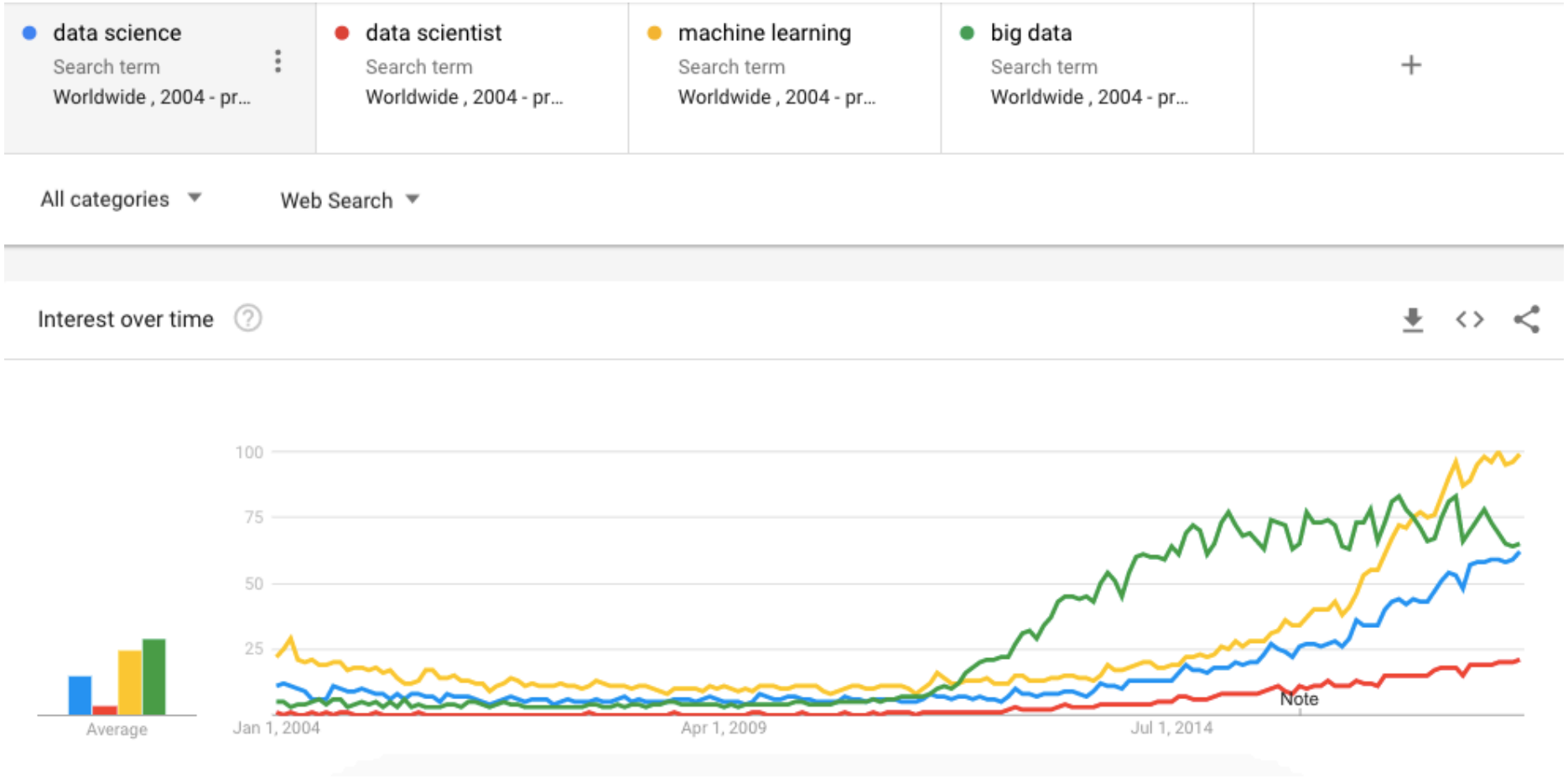
Data 601 @ UMBC

Welcome to Data 601



- Course Logistics
- What is Data Science?
- Software tools
- What are we not covering?
- Data Formats
- Soft skills
- Homework

Components of a good class

- How do you promote a safe classroom environment?
- How do you address the variety of backgrounds of participants?
- How do you encourage participation? Enable rest?
- What behavior fosters growth?

Activity: Think, then write

Ground rules

- Schedule: 7:10 – 9:40 (we may have breaks)
- I value being punctual (start of class, breaks, end of class)

- Don't apologize for asking a question or for not knowing something
- I find it acceptable for you to occasionally not participate

- Tell me if you cannot hear me or if you cannot understand me
- Slides/notes will be provided after lecture (Github)

- I value your feedback:
  - Direct: verbal
  - Indirect: anonymous question/comment sheets on your desk

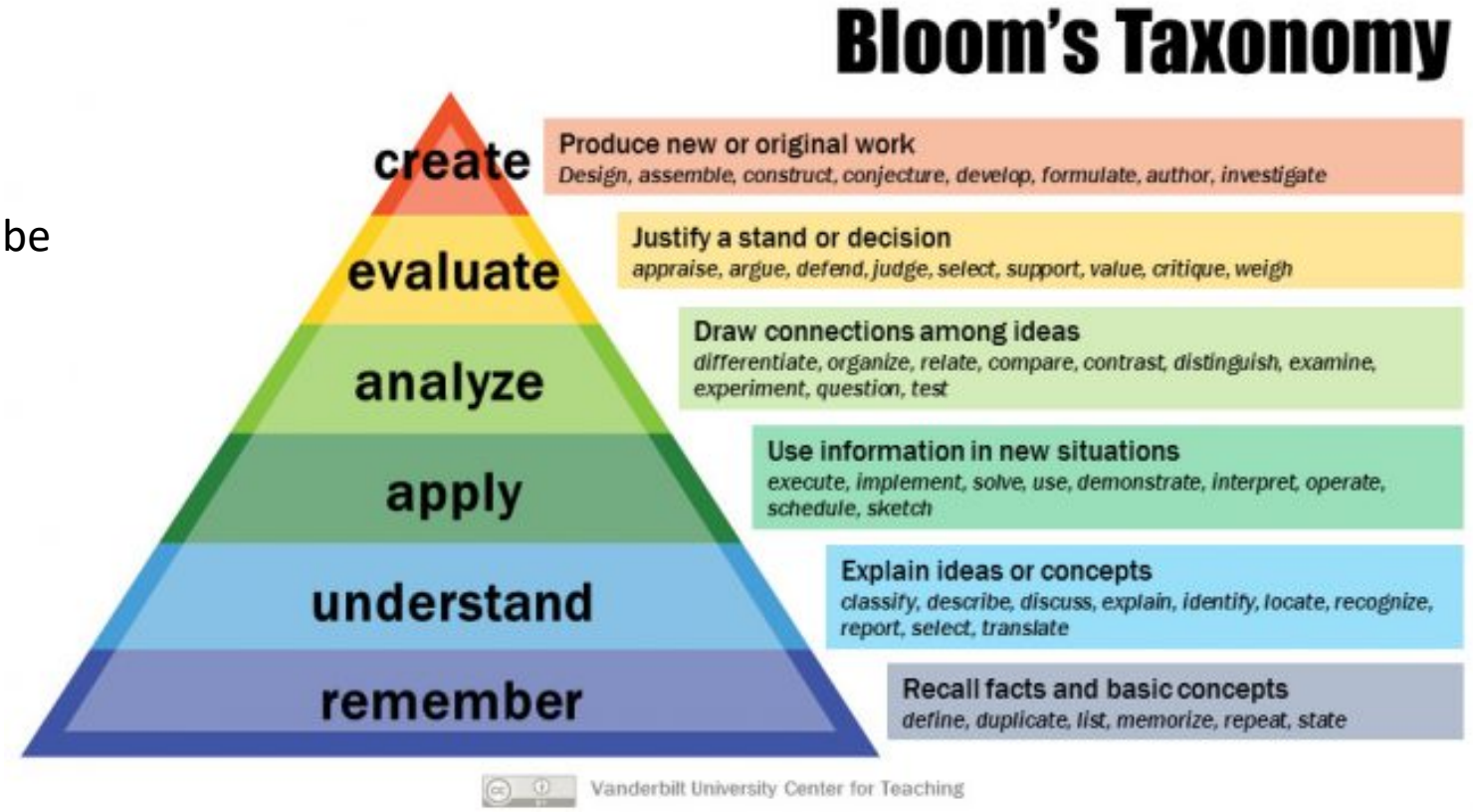
Grading

What you may care about for evaluation

- Attendance: 10%
  - Show up for class
  - Participate in class exercises and surveys
- Homework: 30%
  - When homework is assigned, the due date will be provided
- Midterm Project: %30
- Final Project: 30%

Learning

What I care about conveying

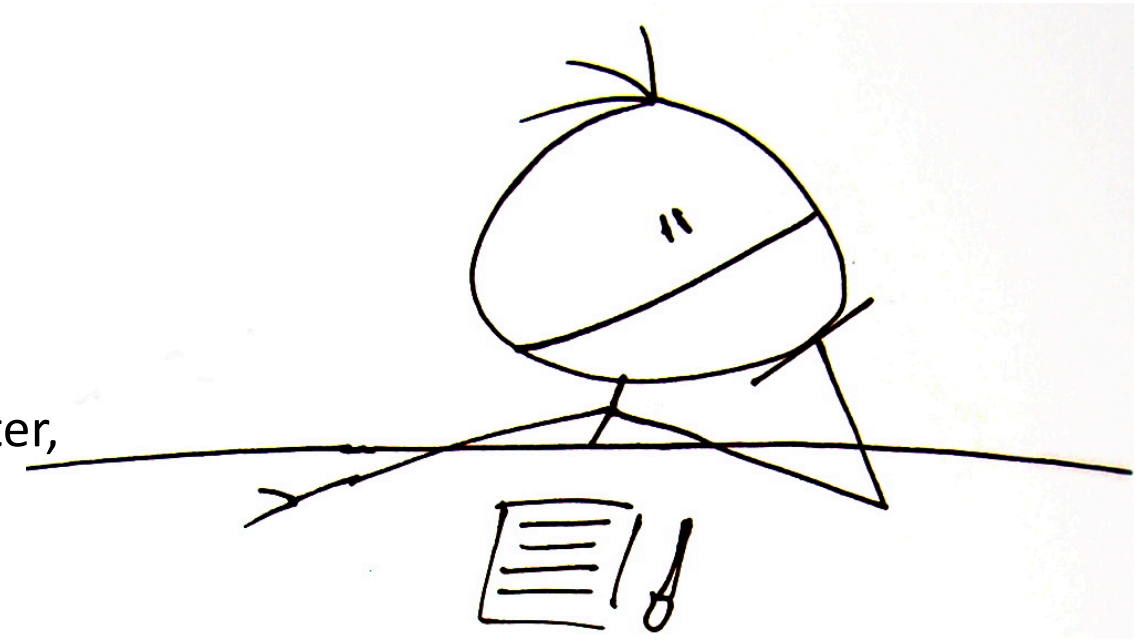


Log your assumptions and expectations

In-class exercise:

- Open a notepad(or a word document) on your computer, record the date.

- Write down your assumptions about this class
- Write down your expectations for this class



Activity: think and write

Store assumptions and expectations

We will revisit these notes later in the semester

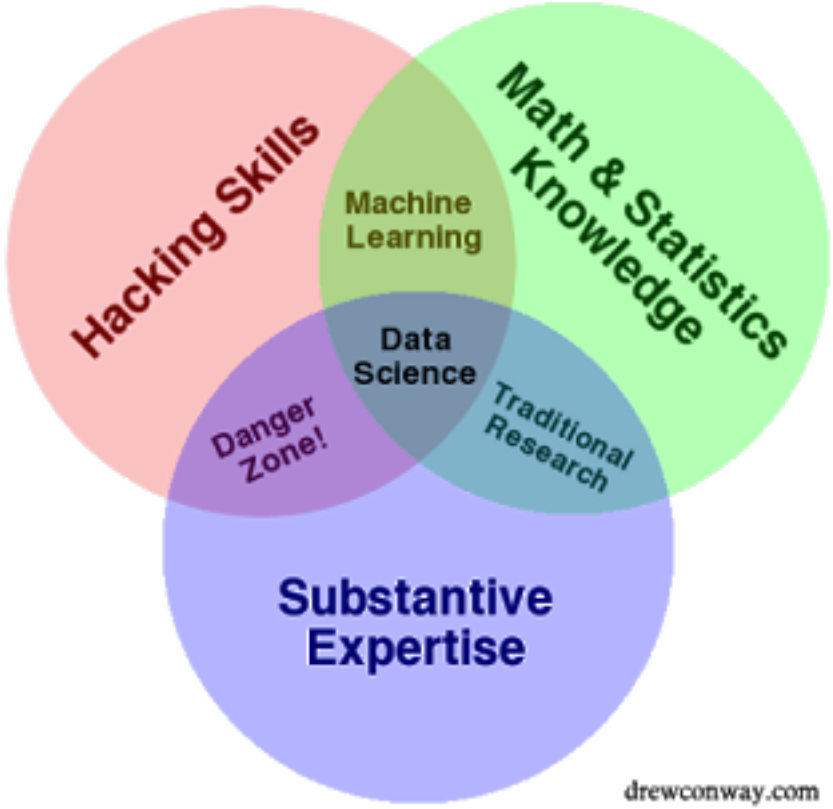
Store your note where it can be accessed later in the course

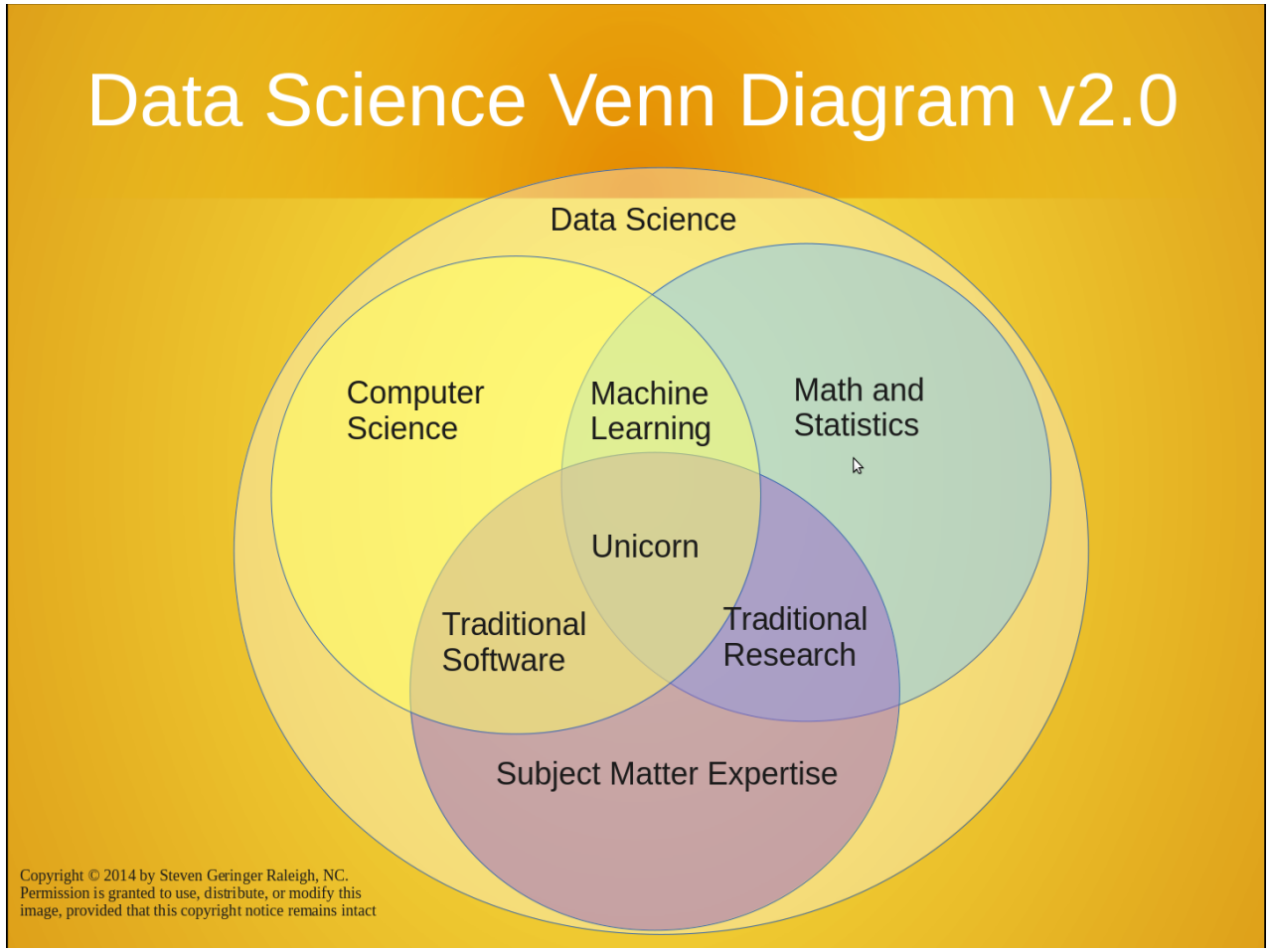
What do you want to learn in this class?

Activity: verbal popcorn; record answers on board

- ~~Course Logistics~~
- What is Data Science?
- Software tools
- What are we not covering?
- Data Formats
- Soft skills
- Homework

There's a lot to cover

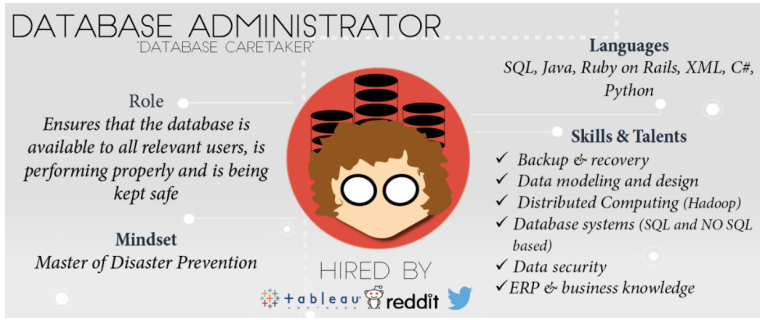


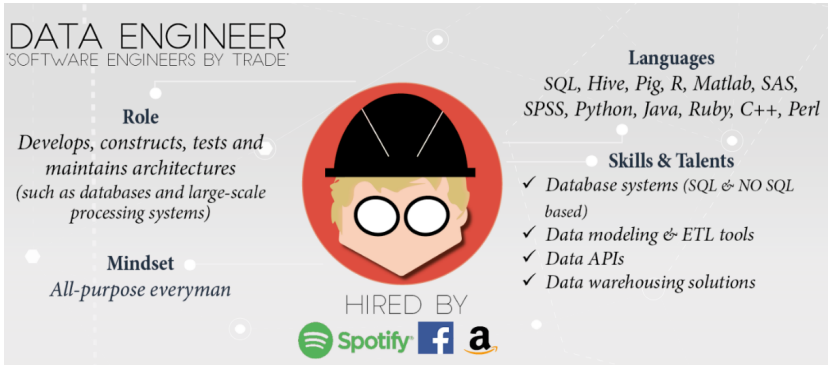


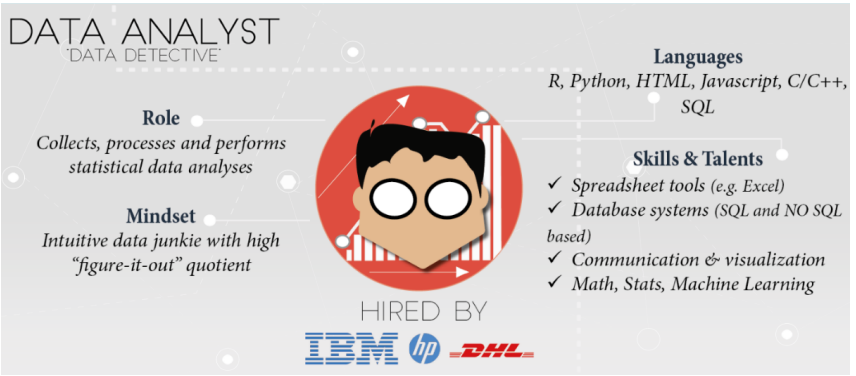
Suggested reading:

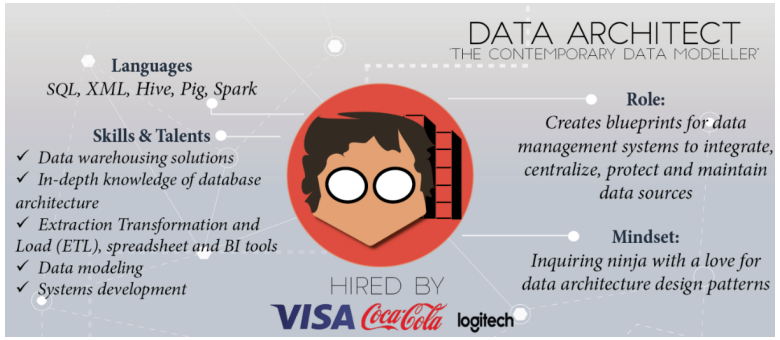
https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html

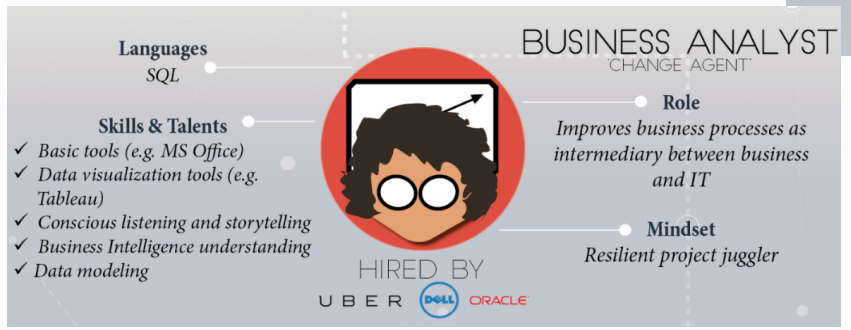
Skills and experience matter more than title and labels

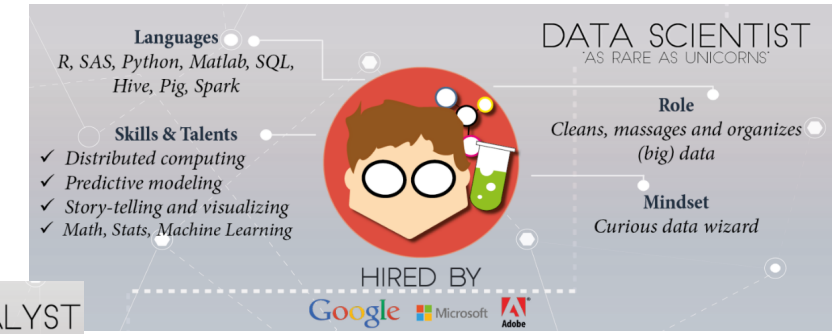


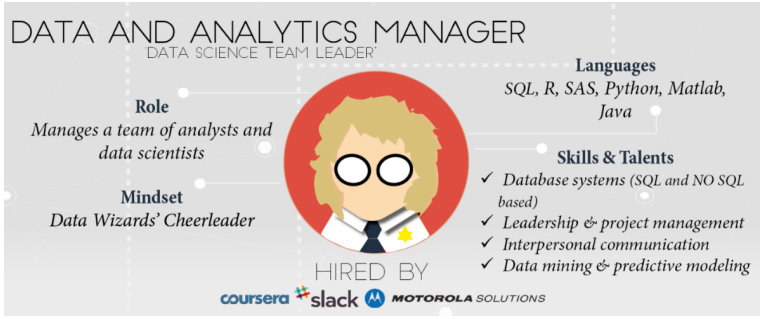












https://www.datacamp.com/community/tutorials/data-science-industry-infographic

Historical progression: data grooming, data mining, data scientist

Data science is an active field with lots of jargon

There will always be something you haven't heard of before.

- Know enough to be conversant with peers
- Be curious about new topics
- Research concepts and labels before using them

Reference: http://www.datascienceglossary.org/

Why learn data science?

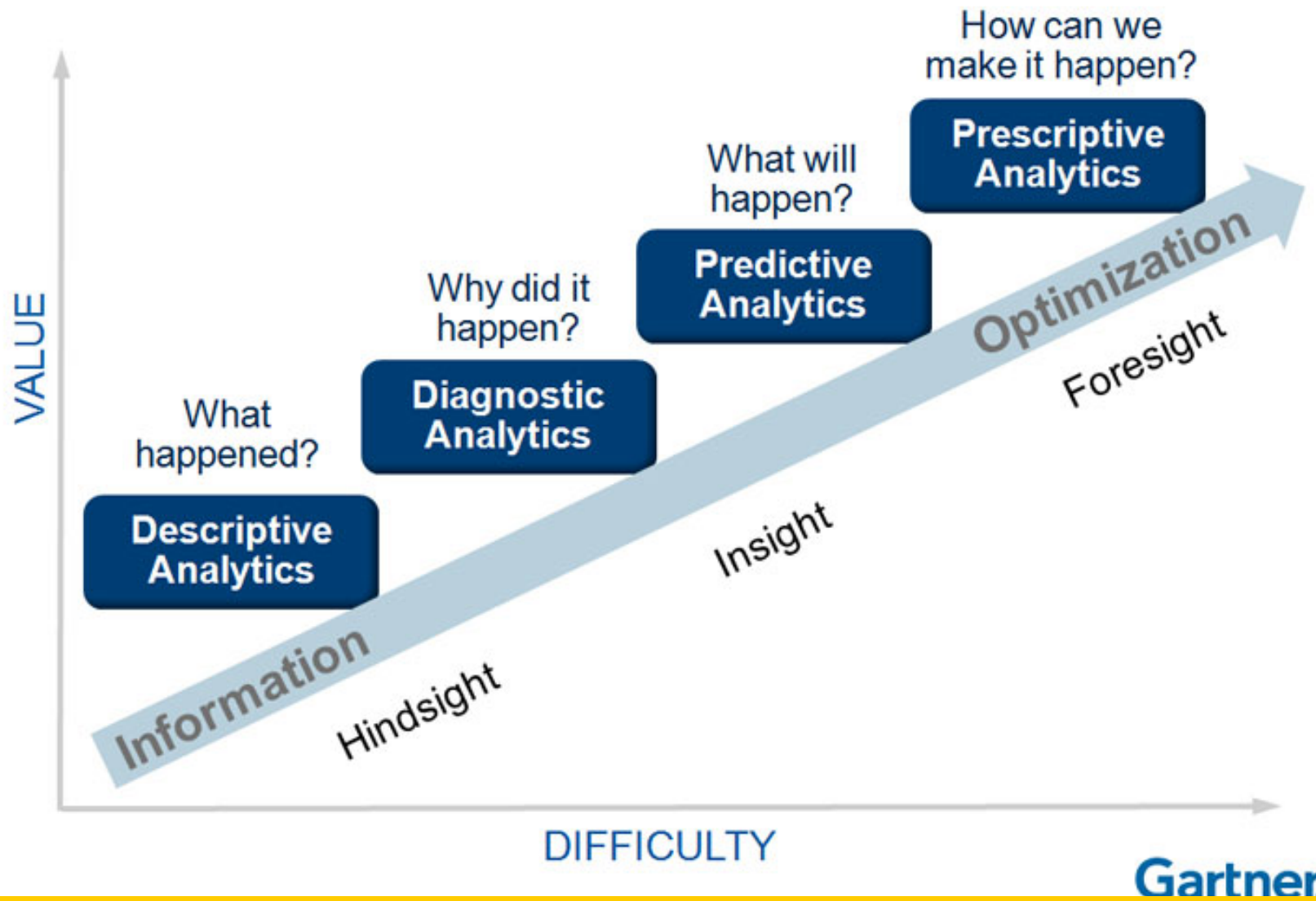
Explore: **identify patterns**

Predict: **make informed guesses**

Infer: **quantify what you know**

Motives:

- Make money
  - Employment
  - Promotion
- Help people
- Gain new knowledge



Large scale use cases with lots of data

- Google's search engine
- Recommendations from Amazon and Netflix
- Bank and Credit Card fraud detection
- Logistics (DHL, UPS) of fleet management
- Healthcare records from patients

Each depends on availability of compute and data

Assumption in this class

- In class we will assume you are a lone data scientist on an island with an internet connection.
- This is not the typical case -- you'll have coworkers, customers, bosses, competitors, collaborators, peers.

Example of how class ≠ real world

- This class will not use competitive grading. (Imagine if it were.)
- As an employee at a company, you may be competing for a bonus or promotion

--> consequence: personal and organizational politics factor into the work environment

- ~~Course Logistics~~
- ~~What is Data Science?~~
- Software tools
- What are we not covering?
- Data Formats
- Soft skills
- Homework

Most popular tool in data science

- what do you think the most popular software tool in data science is?

https://www.trippinsights.com/2018/01/04/milling-about/

Activity: interactive survey

Most popular tool in data science

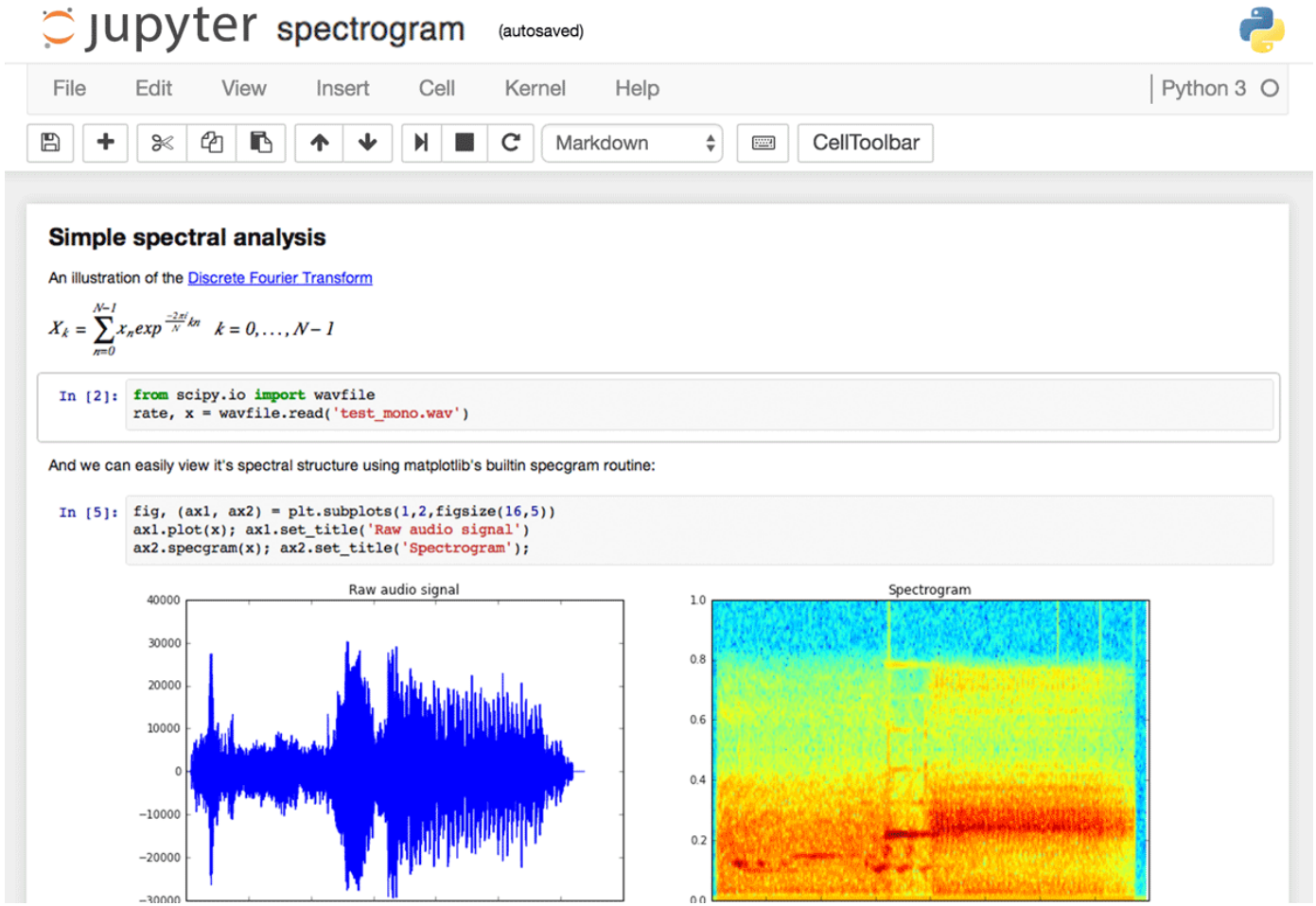
The most popular tool in data science is **Excel**.

- Excel is attractive because it's flexible and capable -- it can store data, the transform, and the visualization (no context switch).
- However, as an analytic ages in Excel and is increasingly tailored, it becomes increasingly brittle.
- Also, in the list of capabilities I didn't include documentation.

Most popular tool in Data 601: *Python*

Interface to Python in Data 601: *Jupyter*

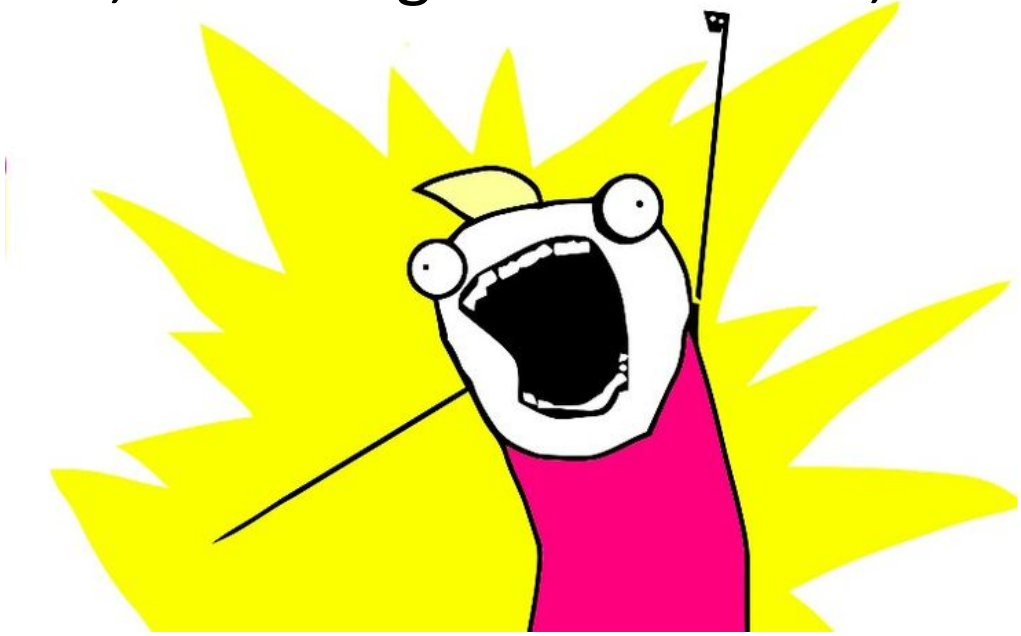
Write in the chat box if you have **used** a Jupyter notebook



Why Jupyter + Python for Data 601?

Jupyter is useful for

- Exploration of data (*jargon*: EDA = exploratory data analysis)
- Documenting your activities (to enable reproducibility)
- Figuring out which software is relevant, which algorithms to use, which software libraries are useful
- Visualizing results



And both Jupyter and Python are free!
And both are widely used!

Python and Jupyter do not cover every use case

- For sufficiently large data sets, Jupyter and Python are not the right tool
- For sufficiently complex analytics, Jupyter and Python are not the right tool

Speed and security are typically not your priority during exploration

Knowing when to invest in switching tools is a skill

Evaluate trade-offs of flexibility and security and speed for a given scale

Relevance of infrastructure to data science

Usual explanation when replicating analysis:

1. Get this data
2. (*Documentation*) Apply this transformation to get result

No explanation of

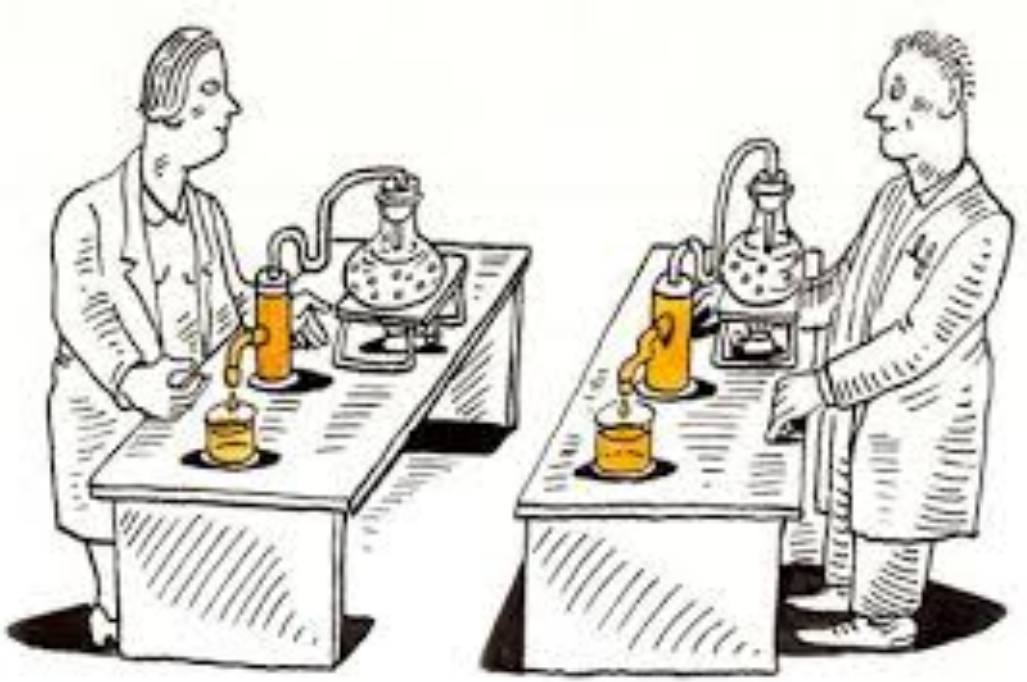
- software used
- software versions
- configurations
- Implementation details

Digital archeology:
Suppose you are to diagnose why someone else's approach doesn't yield same results
Suppose they did their work 20 years ago

Infrastructure in data science to enable Reproducibility and Portability

In addition to data and analysis, implementation and environment matters

1. Use this Operating System
2. Install this software
3. Configure software this way
4. Add these packages
5. Get this data in this format
6. Run analysis against data
7. Create plots
8. Generate report



Best practices: Version control

- Reproducibility applies to your own attempts (not just other people)
- Regardless of how you develop analytics, you'll be creating or editing software and documents.
- [*lesson*] Regardless of how you implement best practices, avoid inventing solutions for which someone else already provided a path.

Suggested resource: https://try.github.io/

Have you used software for version control?

Activity: Install Anaconda

Examples: git, svn, hg

Activity: Install git

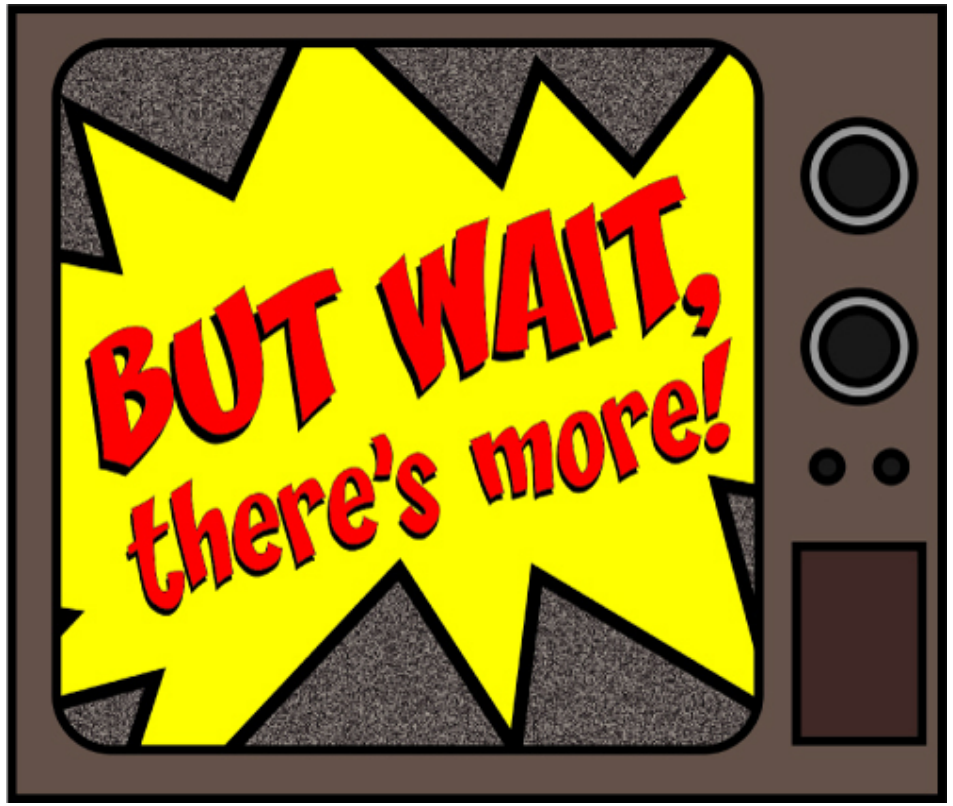
- ~~Course Logistics~~
- ~~What is Data Science?~~
- ~~Software tools~~
- What are we not covering?
- Data Formats
- Soft skills
- Homework

Processes external to data science

Exploration may be enough and the effort terminates

Processes external to data science

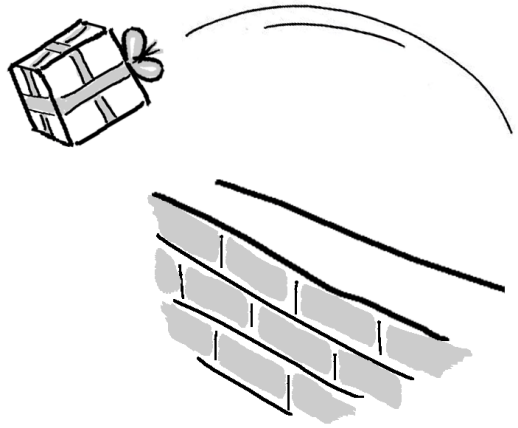
~~Exploration may be enough and the effort terminates~~



Additional refinement is often needed; data science is often merely the start of an investment

Not covered: machine learning

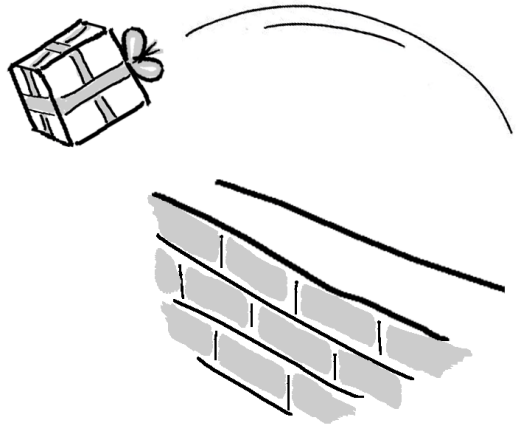
- 602 Introduction to Machine Learning covers Machine Learning
  - If you want to try out, we can do it in one of the last sessions



See http://dev2ops.org/2010/02/what-is-devops/

Not covered: product integration

- There's a complex network of dependencies (i.e. software engineers, managers) of which data science is one component.
- Downstream consumers of your output are likely to be software developers who use containers and support users.
- This class is focused on the data science; not with integration.



See http://dev2ops.org/2010/02/what-is-devops/

Not covered: security

We focus on data science techniques; these do not emphasize secure design of software.

- Let’s Look at some Python code

Figure 1

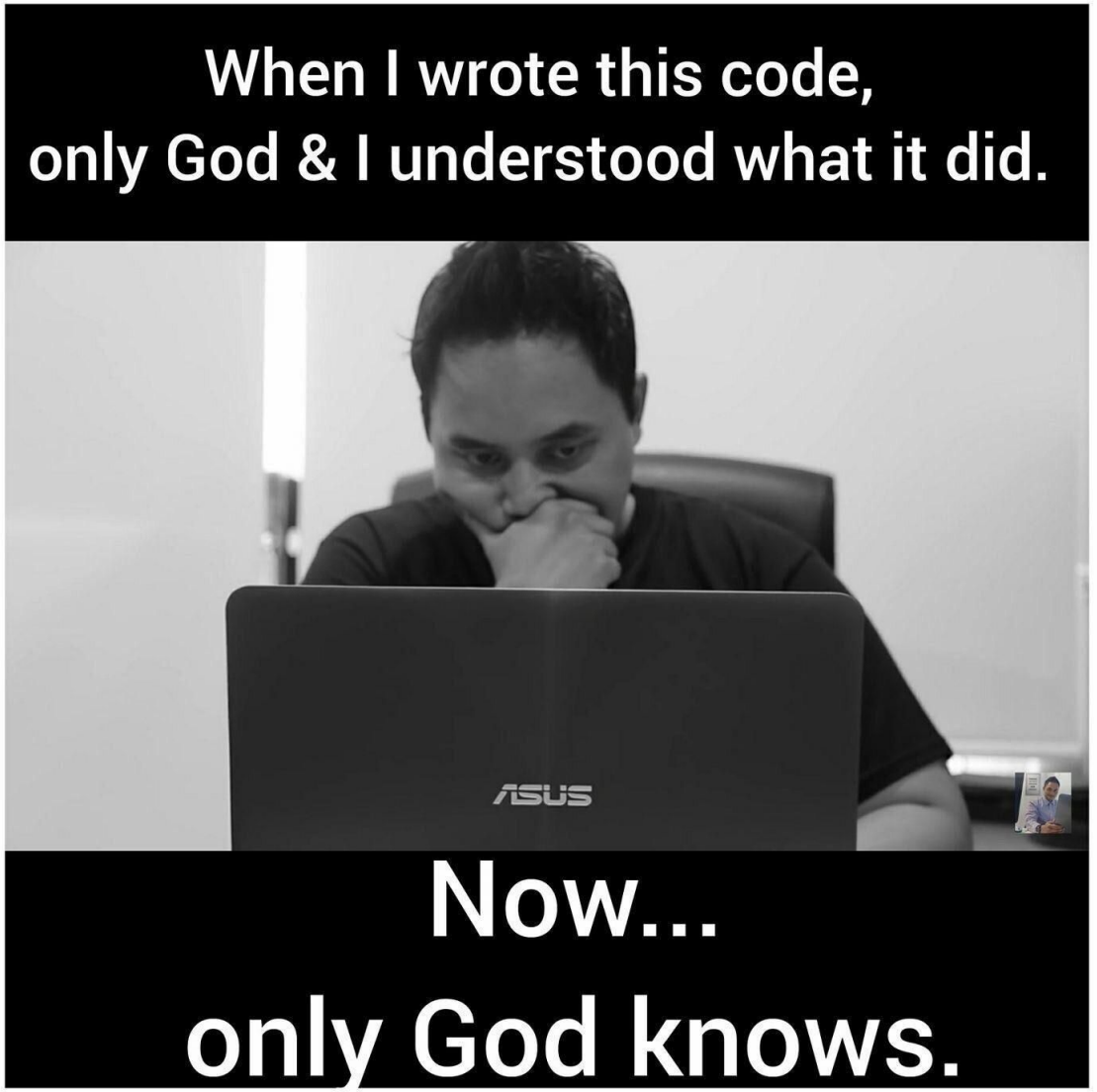
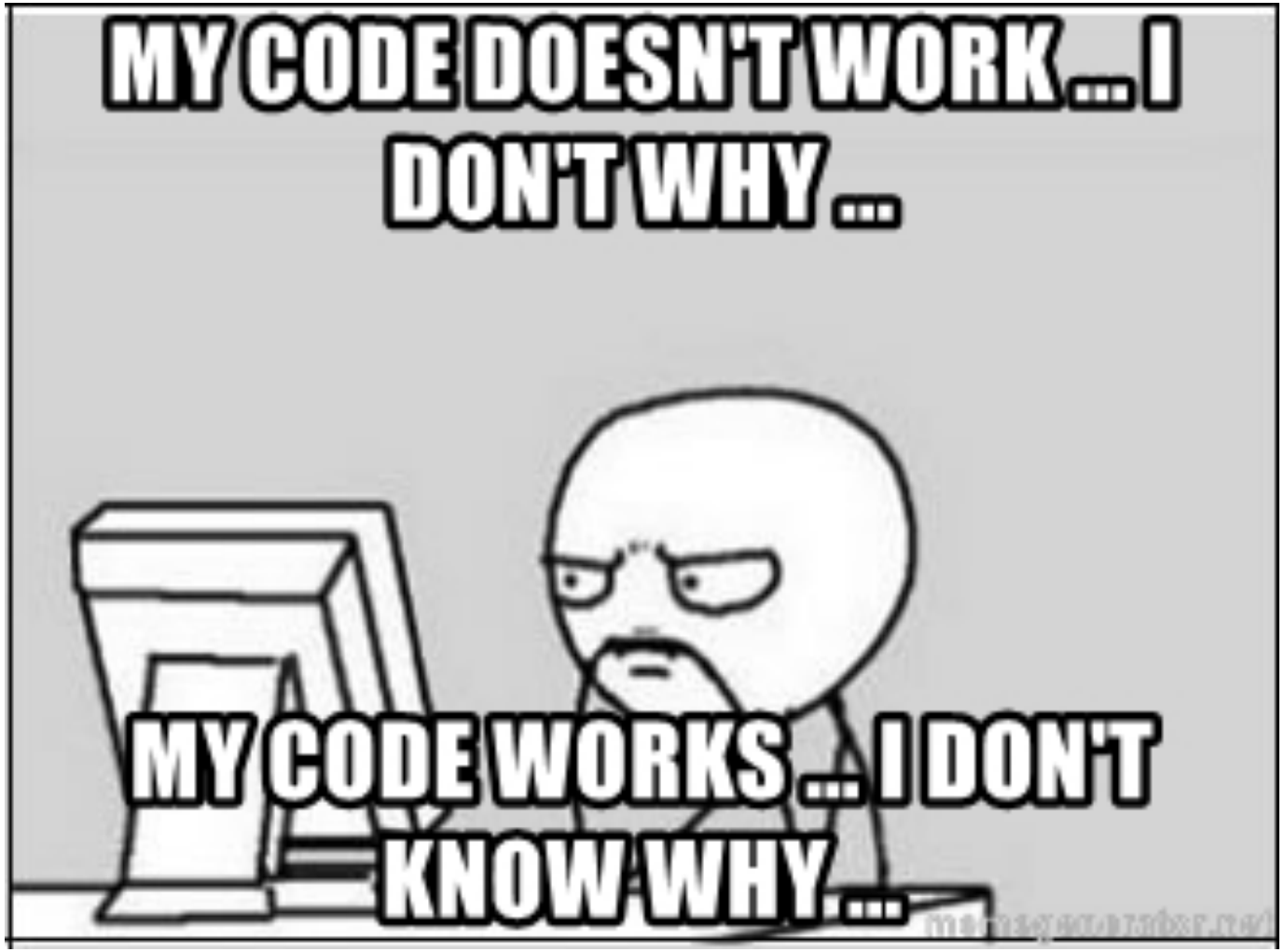
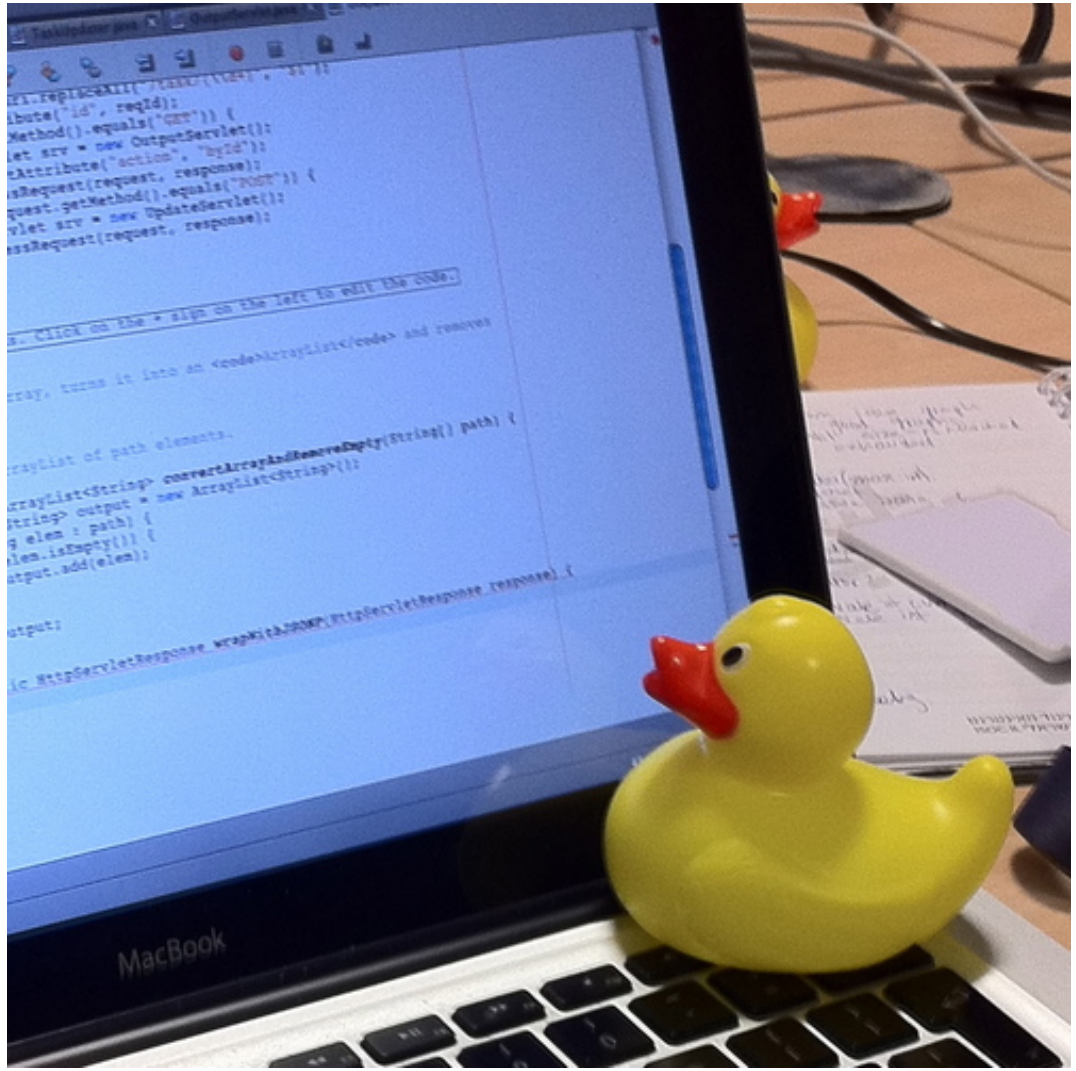

When I wrote this code,
only God & I understood what it did.
ASUS
Now...
only God knows.

Figure 2



Rubber Duck Debugging



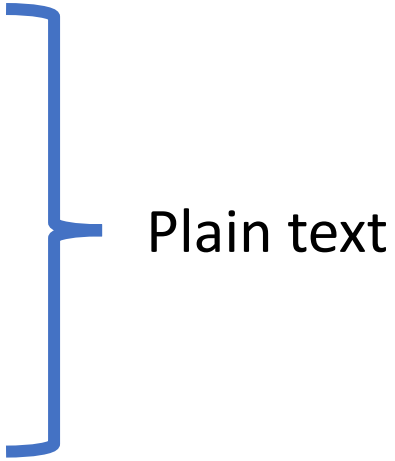
https://en.wikipedia.org/wiki/Rubber_duck_debugging

- ~~Course Logistics~~
- ~~What is Data Science?~~
- ~~Software tools~~
- ~~What are we not covering?~~
- Data Formats
- Soft skills
- Homework

We will discuss: data formats

Examples of data formats

- CSV
  - Excel
- JSON
- XML
- Unstructured text (i.e. txt)
- Images (i.e. png, jpg, bmp, svg)
- Sound (i.e. mp3, wav)
- Images + sound: video



Data formats

- CSV
- JSON
- XML
- Excel
- Unstructured text
- Images
- Sound
- Video

Plain text
Platform independent
Not tied to specific software

--> Recommended data formats by the Library of Congress
See http://www.loc.gov/preservation/resources/rfs/data.html

unstructured text

- LDA - topic modeling
- TF-IDF - document summarization
- word2vec - word meaning

image, sound, video

- OCR - image to text
- captioning - text for image
- Object detection

Data 601 focus is on numerical data and semi-structured text

Tables: what are they good for?

- How would you represent course grades for Data 601 as a table?

(question is intentionally underspecified)

Activity: Raise your hand/Write in chatbox if you have a suggestion

Example of a table for course grades

Last Name	First Name	Homework 1 grade	Homework 2 grade	Exam grade
rgiasg	igmign	79	84	92
qgimzf	gvdvig	73	86	96
asgmi	ybgngfj	58	42	71

Tables: what are they good for?

- How would you represent a corpus of emails?

(question is intentionally underspecified)

Activity: Raise your hand/Write in chatbox if you have a suggestion

Example of a table for emails

Subject	From	To	Date	Has Attachment?
Doggy	Bob@gmail.com	Sam@yahoo.com	2018-08-30	N
Meeting time	Sam@yahoo.com	Alice@hotmail.com	2018-09-01	N
syllabus	Alex@gmail.com	Ari@gmail.com	2018-09-06	Y

CSV

No set standard; see https://tools.ietf.org/html/rfc4180
https://en.wikipedia.org/wiki/Comma-separated_values

- A table with rows and columns composed of text

data-set - Notepad

File Edit Format View Help

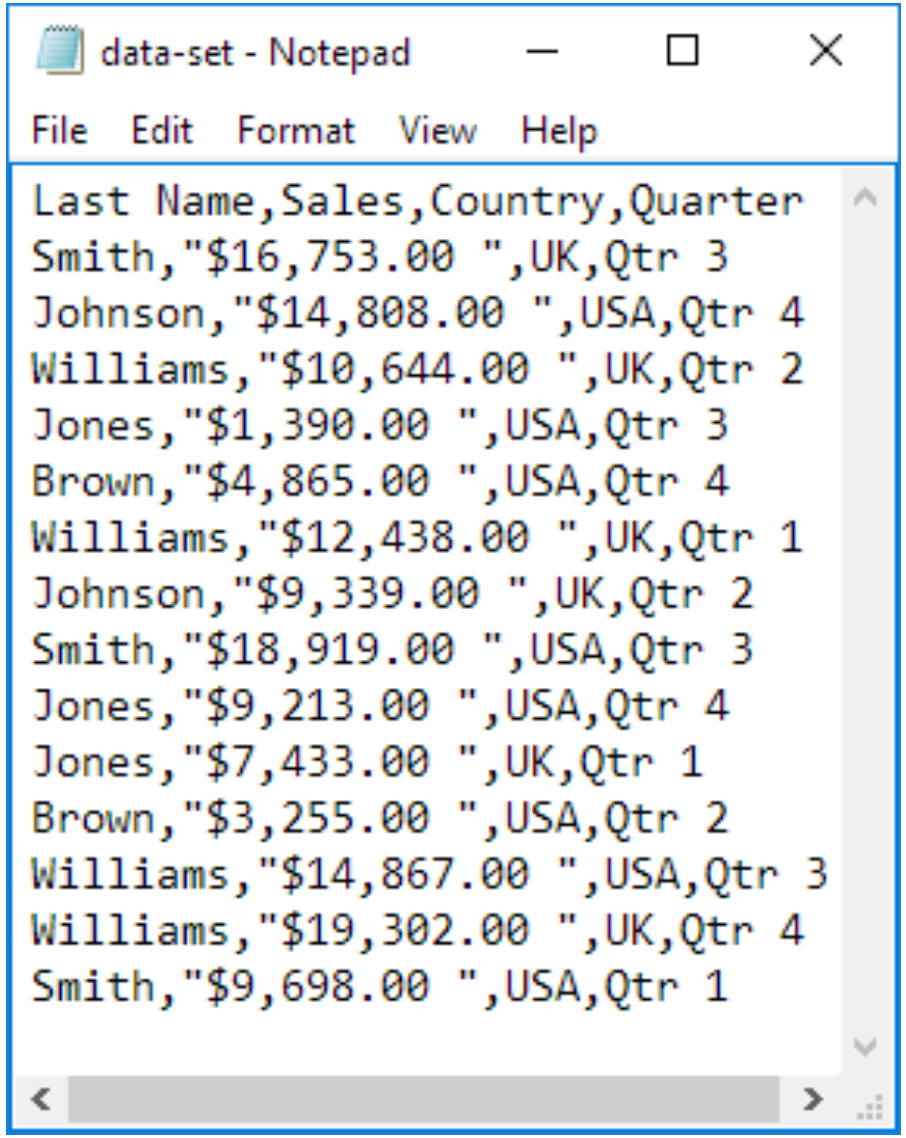
```
Last Name,Sales,Country,Quarter
Smith,"$16,753.00 ",UK,Qtr 3
Johnson,"$14,808.00 ",USA,Qtr 4
Williams,"$10,644.00 ",UK,Qtr 2
Jones,"$1,390.00 ",USA,Qtr 3
Brown,"$4,865.00 ",USA,Qtr 4
Williams,"$12,438.00 ",UK,Qtr 1
Johnson,"$9,339.00 ",UK,Qtr 2
Smith,"$18,919.00 ",USA,Qtr 3
Jones,"$9,213.00 ",USA,Qtr 4
Jones,"$7,433.00 ",UK,Qtr 1
Brown,"$3,255.00 ",USA,Qtr 2
Williams,"$14,867.00 ",USA,Qtr 3
Williams,"$19,302.00 ",UK,Qtr 4
Smith,"$9,698.00 ",USA,Qtr 1
```

CSV

No set standard; see https://tools.ietf.org/html/rfc4180
https://en.wikipedia.org/wiki/Comma-separated_values

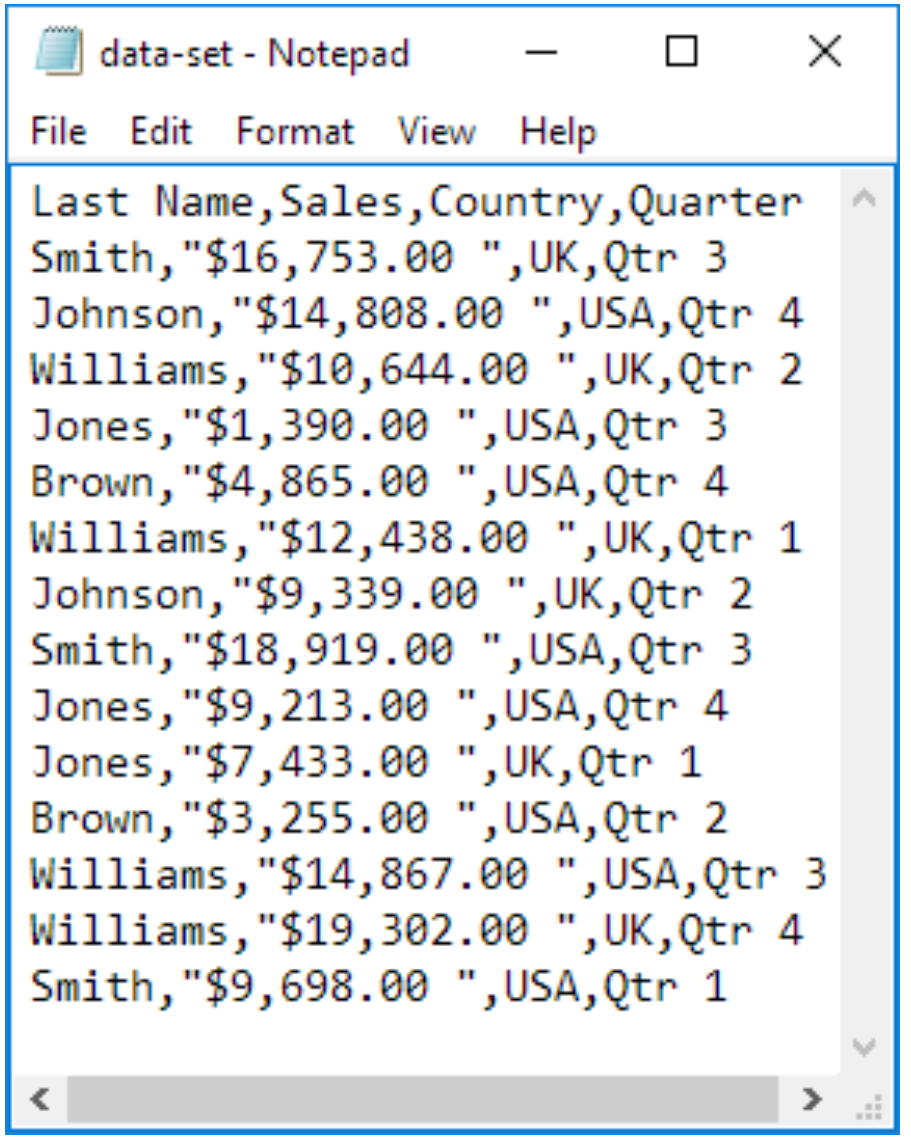
- A table with rows and columns composed of text
- delimiters:
  - Line break (separates rows)
  - Comma (separates columns)
- May or may not have a header:

  First row is descriptions of columns



CSV caveats

- Text (not images, audio, video)
- Each column of same type
  - Name or word or description (text)
  - Number
  - Category label (text)
- Delimiters within a value
  - Enclosed within a pair of double quotes: " "
- Quotes for quotes, ie 24"



CSV variations

Delimiters vary:

- Tab – see https://en.wikipedia.org/wiki/Tab-separated_values
- | *(pipe)*

```
Month|count|location
Jan|1332|here
June|5,593|there
Feb|953,24|everywhere, CA
Oct|592|my home town, MD
```

Can you see why using pipes would be desirable?

CSV alternative: fixed width

- Consistent number of characters per column

```
ID      l_name     f_nametuition
0585822Potter, JrHarry 204
0485572Weasley    Ron    958
5924245Granger    Hermio422
4724926Diggory    Cedric1042
4938243Weasley    Fred   394
```

Can you see why fixed width would be desirable?

CSV alternative: fixed width

- Consistent number of characters per column

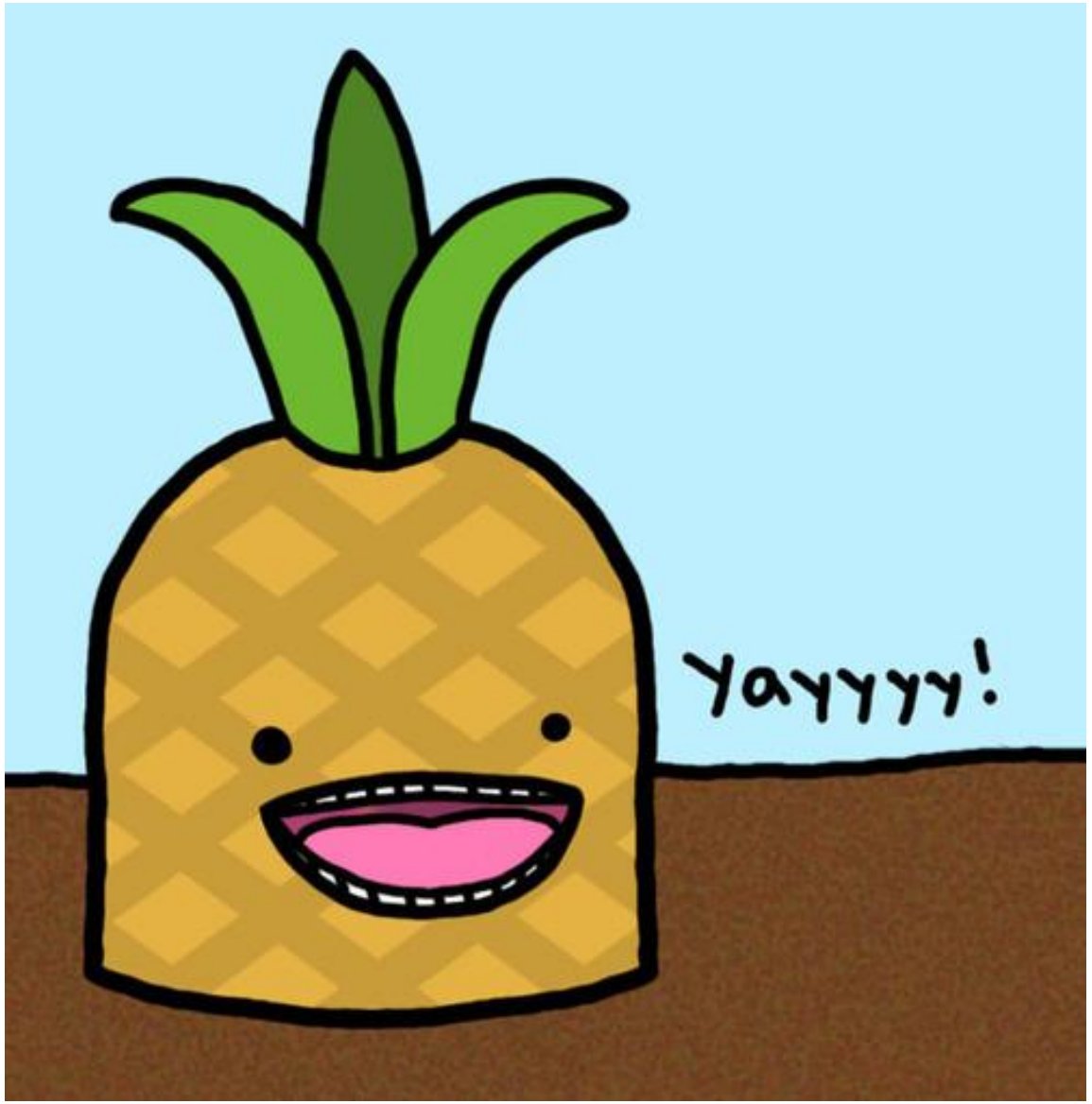
ID	l_name	f_name	tuition
0585822	Potter, Jr	Harry	204
0485572	Weasley	Ron	958
5924245	Granger	Hermio	422
4724926	Diggory	Cedric	1042
4938243	Weasley	Fred	394

Caveat: each column needs to be at least as long as the longest entry

CSV quiz

When you see the question, **don't** shout out the answer.

Use Blackboard for voting on answers; the colored cards are a backup.



Name, year, class, grade
Ming, 2013, Data 601, B
Imgb, 2015, Data 601, B
Rimgw, 2012, Data 601, A
Wemf, 2014, Data 602, C

Name, year, class, grade
Ming, 2013, Data 601, B
Imgb, 2015, Data 601, B
Rimgw, 2012, Data 601, A
Wemf, 2014, Data 602, C

Q: Does this CSV have a header?

(Don't shout out the answer)

Name, year, class, grade
Ming, 2013, Data 601, B
Imgb, 2015, Data 601, B
Rimgw, 2012, Data 601, A
Wemf, 2014, Data 602, C

Q: Does this CSV have a header?

Backup voting:

YES

NO

NOT SURE

Name, year, class, grade
Ming, 2013, Data 601, B
Imgb, 2015, Data 601, B
Rimgw, 2012, Data 601, A
Wemf, 2014, Data 602, C

Q: Does this CSV have a header?

Answer -- YES

~~NO~~

~~NOT SURE~~

Name, year, class, grade
Ming, 2013, Data 601, B
Imgb, 2015, Data 601, B
Rimgw, 2012, Data 601, A
Wemf, 2014, Data 602, C

Q: How many rows does this CSV have?

(Don't shout out the answer)

Name, year, class, grade
Ming, 2013, Data 601, B
Imgb, 2015, Data 601, B
Rimgw, 2012, Data 601, A
Wemf, 2014, Data 602, C

Q: How many rows does this CSV have?

5 rows
4 rows
NOT SURE

Row 1: Name, year, class, grade
Row 2: Ming, 2013, Data 601, B
Row 3: Imgb, 2015, Data 601, B
Row 4: Rimgw, 2012, Data 601, A
Row 5: Wemf, 2014, Data 602, C

Col 1, Col 2, Col 3, Col 4

Q: How many rows does this CSV have?

Answer -- 5 rows

~~4 rows~~

~~NOT SURE~~

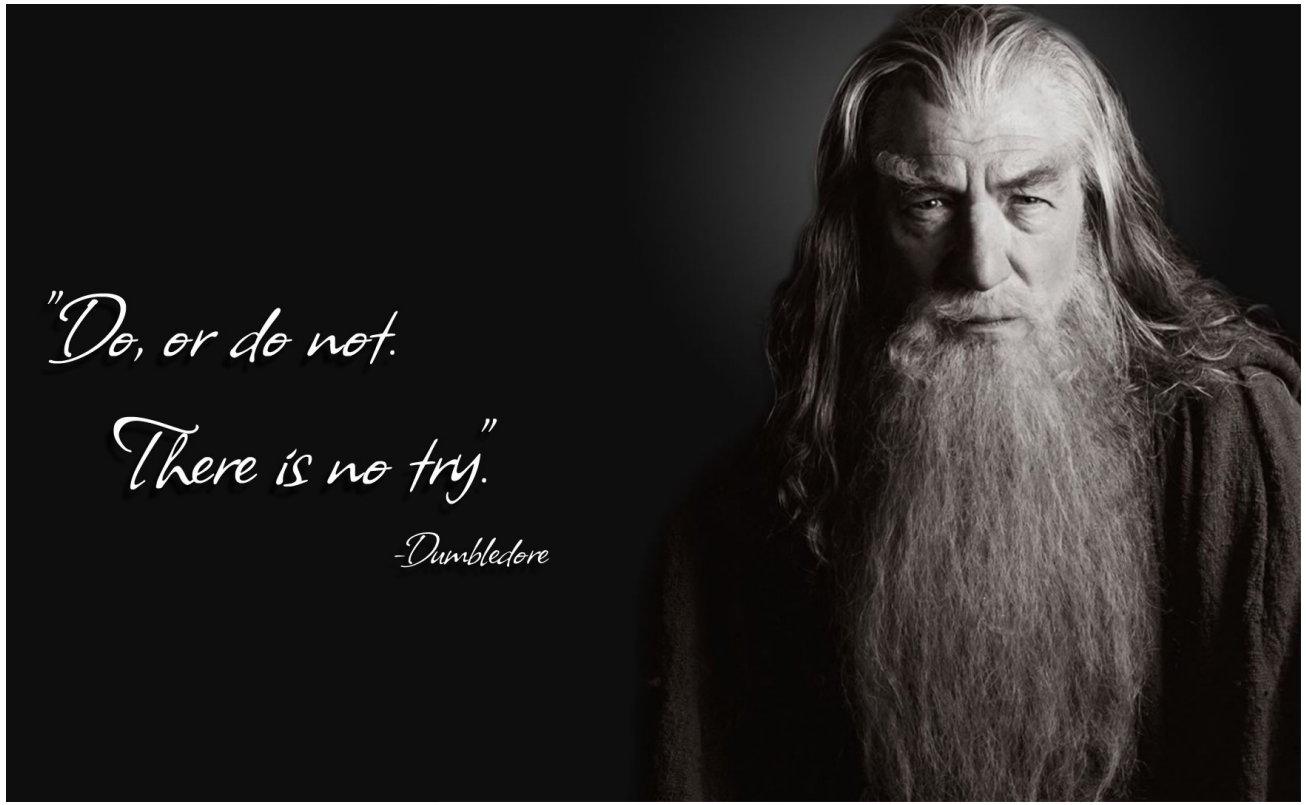
Why not 4?

- Header is a row
- Rows (5) versus columns (4)

ID, last name, first name, tuition
058582, Potter, Jr, Harry, $204
0485572, Weasley, Ron, $958
592424, Granger, Hermione, $422
472492, Diggory, Cedric, $1,042
493824, Weasley, Fred, $394

Q: What is wrong with this CSV?

Activity: Raise your hand/Write in chatbox if you have an answer



ID, last name, first name, tuition
058582, "Potter, Jr", Harry, $204
0485572, Weasley, Ron, $958
592424, Granger, Hermione, $422
472492, Diggory, Cedric, "$1,042"
493824, Weasley, Fred, $394

Q: What is wrong with this CSV?

A: Two mistakes with unquoted commas

Jupyter and CSV

`load_csv.ipynb`

reasons not to use CSV data format

- data type of columns undefined

--> string versus float versus integer

This makes loading data from file to memory expensive

- consistency of data types per row is not guaranteed

Row 1 could be "Bob, 52, 485-2949" while

Row 2 could be "859492, mark@yahoo.com, $95"

- Format consistency is not enforced

Header could be comma separated, data is pipe separated, except fourth and fifth column are separated by semicolon

See Falsehoods Programmers Believe About CSVs

Additional common data formats: JSON and XML

- Semi-structured_data for flexibility

JSON jargon: Key-value pairs

XML jargon: Tags, Elements, Attributes

- Attributes contain data related to a specific element

Data format: JavaScript Object Notation (JSON)

```
{"menu": {
  "id": "file",
  "value": "File",
  "popup": {
    "menuitem": [
      {"value": "New", "onclick": "CreateNewDoc()"},
      {"value": "Open", "onclick": "OpenDoc()"},
      {"value": "Close", "onclick": "CloseDoc()"}
    ]
  }
}}
```

source: https://json.org/example.html

https://tools.ietf.org/html/rfc4627

Data format: JavaScript Object Notation (JSON)

```
{"menu": {
  "id": "file",
  "value": "File",
  "popup": {
    "menuitem": [
      {"value": "New", "onclick": "CreateNewDoc()"},
      {"value": "Open", "onclick": "OpenDoc()"},
      {"value": "Close", "onclick": "CloseDoc()"}
    ]
  }
}}
```

source: https://json.org/example.html

https://tools.ietf.org/html/rfc4627

Data format: JavaScript Object Notation (JSON)

```
{"menu": {
  "id": "file",
  "value": "File",
  "popup": {
    "menuitem": [
      {"value": "New", "onclick": "CreateNewDoc()"},
      {"value": "Open", "onclick": "OpenDoc()"},
      {"value": "Close", "onclick": "CloseDoc()"}
    ]
  }
}}
```

source: https://json.org/example.html

https://tools.ietf.org/html/rfc4627

How many key-value pairs does this JSON feature?

```
{"menu": {
  "id": "file",
  "value": "File",
  "popup": {
    "menuitem": [
      {"value": "New", "onclick": "CreateNewDoc()"},
      {"value": "Open", "onclick": "OpenDoc()"},
      {"value": "Close", "onclick": "CloseDoc()"}
    ]
  }
}}
```

source: https://json.org/example.html

https://tools.ietf.org/html/rfc4627

How many key value pairs does this JSON feature?

```
{"menu": {
  "id": "file",
  "value": "File",
  "popup": {
    "menuitem": [
      {"value": "New", "onclick": "CreateNewDoc()"},
      {"value": "Open", "onclick": "OpenDoc()"},
      {"value": "Close", "onclick": "CloseDoc()"}
    ]
  }
}}
```

There are 11 key value pairs

source: https://json.org/example.html

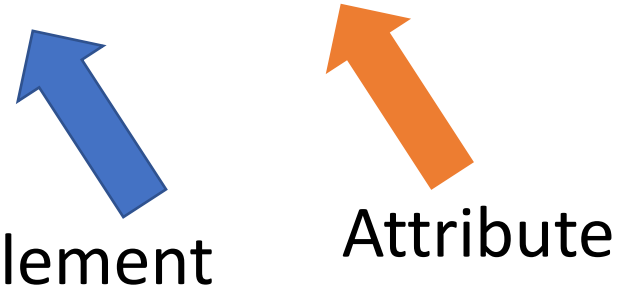
https://tools.ietf.org/html/rfc4627

Reading JSON in Jupyter

`load_json.ipynb`

Data format: Extensible Markup Language (XML)

```
<menu id="file" value="File">
  <popup>
    <menuitem value="New" onclick="CreateNewDoc()" />
    <menuitem value="Open" onclick="OpenDoc()" />
    <menuitem value="Close" onclick="CloseDoc()" />
  </popup>
</menu>
```



attributes cannot contain multiple values (elements can)
attributes cannot contain tree structures (elements can)
Source: https://www.w3schools.com/xml/xml_attributes.asp

source: https://json.org/example.html

https://tools.ietf.org/html/rfc2376

Multiple representations in XML

```
<person>
  <firstName>John</firstName>
  <lastName>Smith</lastName>
  <age>25</age>
  <address>
    <street>21 2nd Street</street>
    <city>New York</city>
    <state>NY</state>
    <zip>10021</zip>
  </address>
</person>
```

There are no rules about when to use attributes or when to use elements in XML.

Source: https://en.wikipedia.org/wiki/JSON#XML_samples

```
<person firstName="John" lastName="Smith" age="25">
  <address street="21 2nd Street" city="New York" state="NY" zip="10021" />
</person>
```

Reading XML in Jupyter

`load_xml.ipynb`

There are more data formats used in different situations

- HDF5 – for extremely large and complex data collections.

- pickle - data serialization for Python

See https://www.benfrederickson.com/dont-pickle-your-data/

- ~~Course Logistics~~
- ~~What is Data Science?~~
- ~~Software tools~~
- ~~What are we not covering?~~
- ~~Data Formats~~
- Soft skills
- Homework

Data Science is more than Math and Software

Human interaction in data science

- Discovering stakeholders
- Negotiating with data owners
- Customer engagement

https://hbr.org/2017/01/the-best-data-scientists-get-out-and-talk-to-people

Iterating with customers

- As a data scientist, you'll often be working for someone other than yourself.
- Expect under-specified requirements from customers. Iterate.
- Provide incomplete solutions rather than waiting until the product is perfect.

https://en.wikipedia.org/wiki/Minimum_viable_product

When to persist,
When to change course,
When to seek help

Try attacking the challenge for 30 minutes

Then seek help or do something else for a while

https://en.wikipedia.org/wiki/Pomodoro_Technique

Pro-tip when seeking help

How to ask well-formed questions:

https://stackoverflow.com/help/how-to-ask

[Intentional sidetrack to StackOverflow.]

Ask technical questions:

- *Poor*: "I don't understand Python dictionaries" (--> online tutorials)
- *Better*: "When is it appropriate to use a key-value pair?"

- *Poor*: If I submitted this assignment as is, what score would I get?
- *Better*: I am planning to submit the attached assignment, but currently there's an error in the third cell. I've searched online but don't find any references to the error message. Can you provide guidance?

Emotions in Data Science

- As a data scientist, most of your time will be spent in a desert of uncertainty, frustration, and doubt.

- There will be rare short-lived interspersed spikes of excitement and happiness due to events like getting a new dataset, creating a new analytic, getting a new result, or being thanked by a stakeholder.

This experience is normal and does not go away.

See also the psychology of slot machines

- ~~Course Logistics~~
- ~~What is Data Science?~~
- ~~Software tools~~
- ~~What are we not covering?~~
- ~~Data Formats~~
- ~~Soft skills~~
- Homework

Task: Jupyter with Python 3 kernel

- *Recommended*: Local installation on your computer
  - https://www.anaconda.com/

- *Backup*: Web-based (free, no guarantee of availability, features vary)
  - https://colab.research.google.com/
  - https://mybinder.org/
    - Depends on reference to github repository

Homework

1. Read: First 10 pages of "50 years of data science"
   Write a half page summary of the text
   Submit essay via Blackboard

2. Complete Homework-week1.ipynb
   Submit it as ipython notebook to Blackboard

Action: Read, write, tell

Online resources (in addition to books)

- Meetups
  - https://www.meetup.com/topics/data-science/
  - https://www.meetup.com/BigDataBaltimore/
  - https://www.meetup.com/Statistical-Seminars-DC/events/254200651/
- News and blogs
  - https://www.kdnuggets.com/
  - https://news.ycombinator.com/
  - https://hackernoon.com/
  - https://www.reddit.com/r/datascience/
  - https://dataelixir.com/newsletters/
- Online courses
  - Coursera
    - https://www.coursera.org/learn/machine-learning/home/welcome