LECTURE 6

Data Cleaning and EDA

Exploratory data analysis and its role in the data science lifecycle.

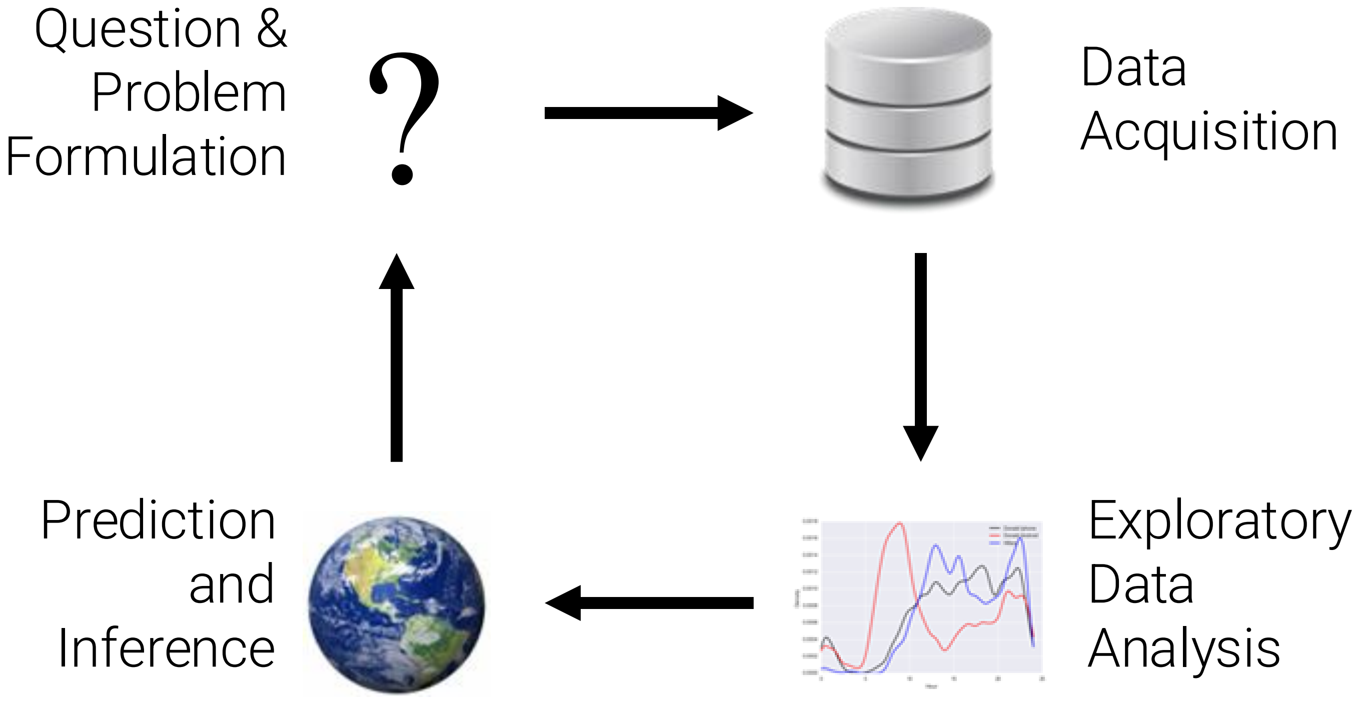
Sean Kang

Quick Recap Pandas and Jupyter Notebooks

- Reviewing DataFrame concepts
  - **Series**: A named column of data with an index
  - **Indexes**: The mapping from keys to rows
  - **DataFrame**: collection of series with common index
- Dataframe access methods
  - **Filtering** on predicts and **slicing**
  - **df.loc**: location by index
  - **df.iloc**: location by integer address
  - **groupby** & **pivot** aggregating data

You have **collected** or **been given** a box of data?

What do you do next?

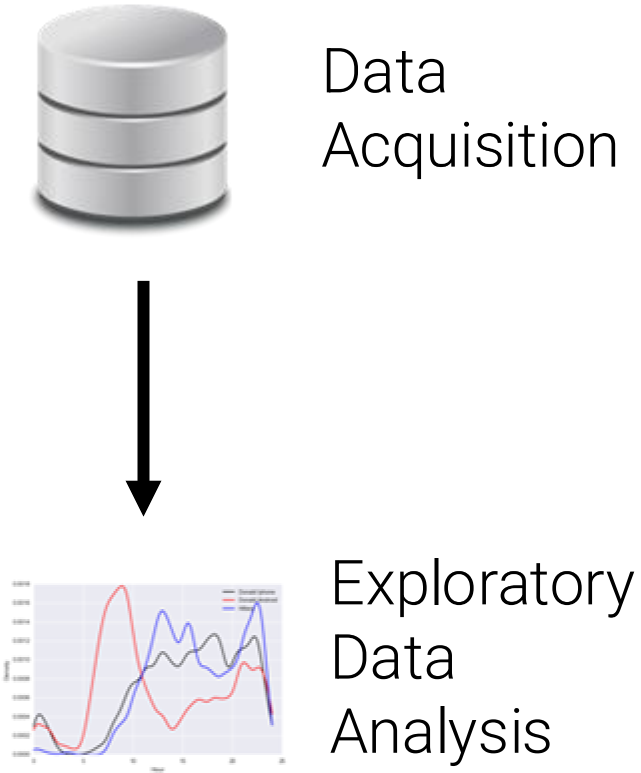


Question & Problem Formulation

Data Acquisition

Exploratory Data Analysis

Prediction and Inference

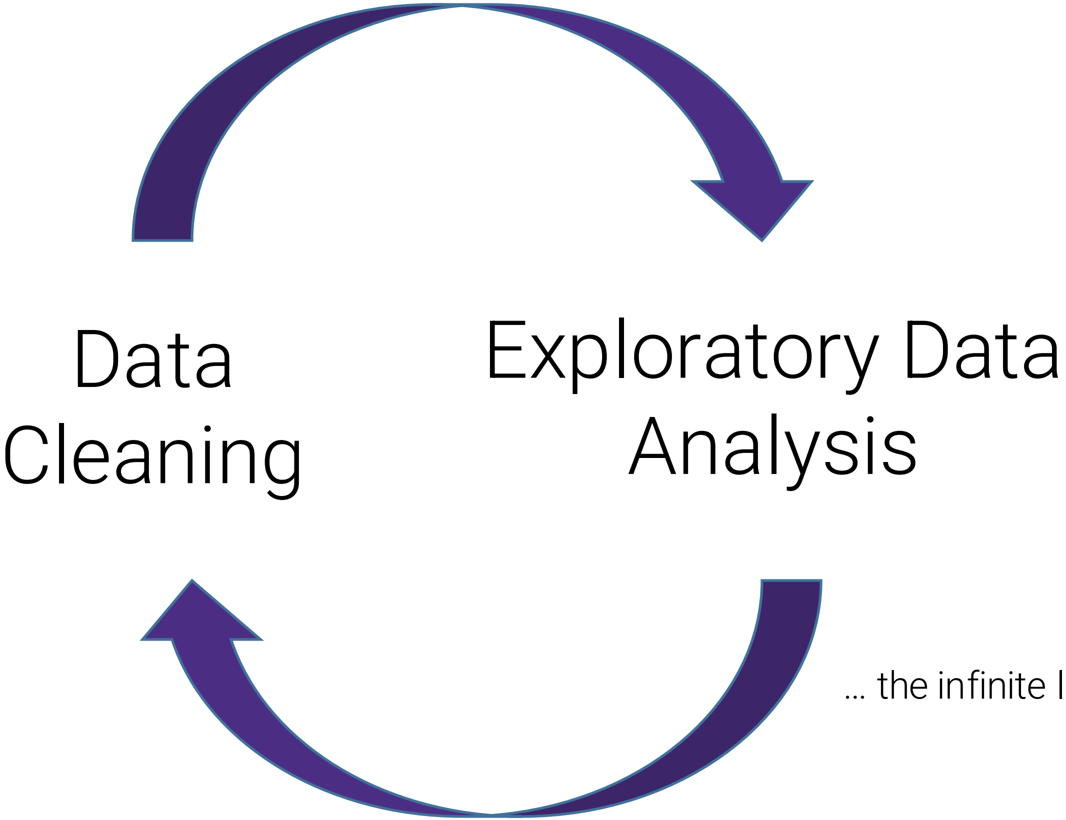


Data Acquisition

Exploratory Data Analysis

Topics For This Lecture

- Understanding the Data
  - Data Cleaning
  - Exploratory Data Analysis (EDA)
  - Basic data visualization
- Common Data Anomalies
  - … and how to fix them



Data Cleaning

Exploratory Data Analysis

… the infinite loop of data science.

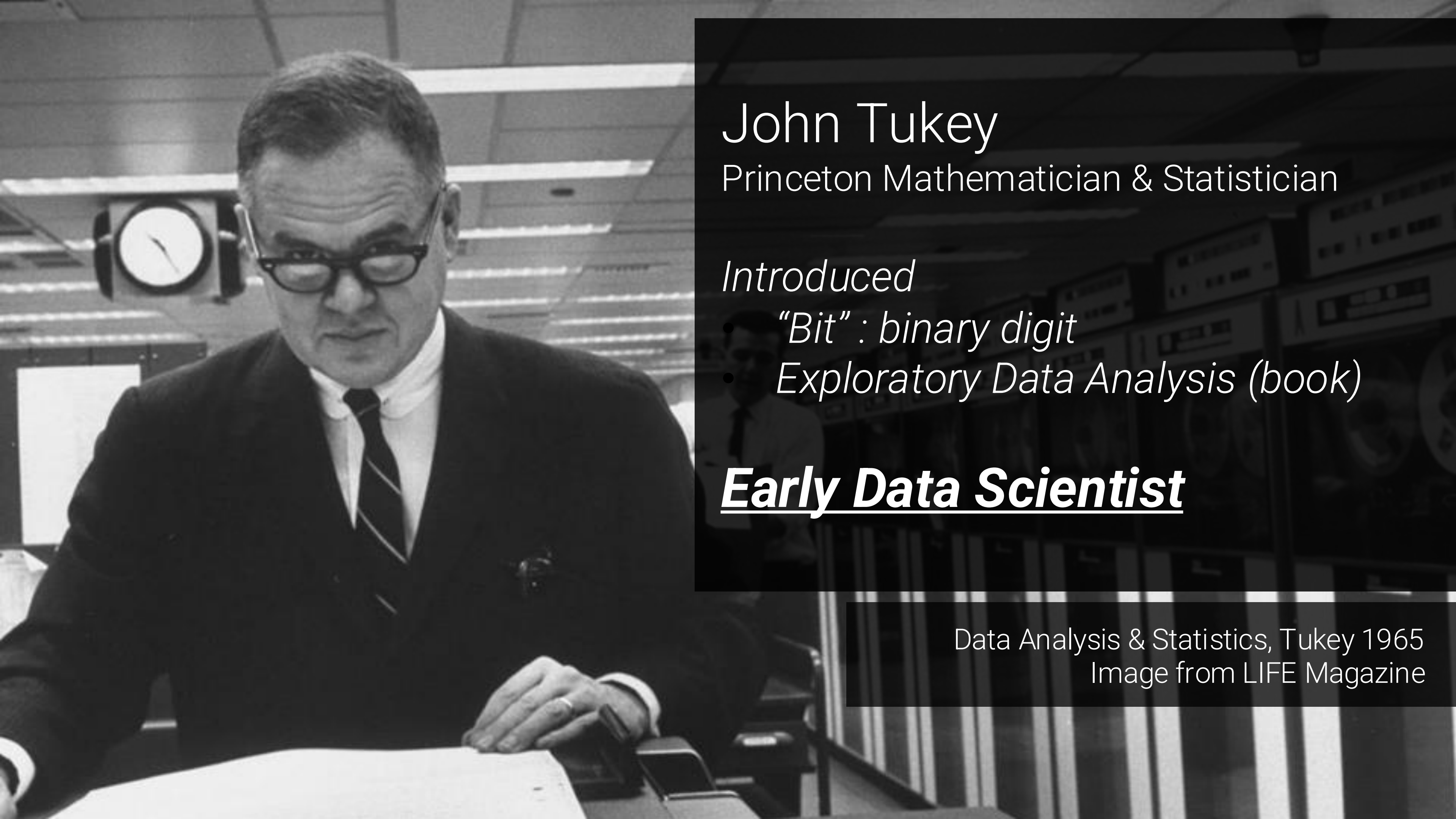
Data Cleaning

- The process of transforming **raw data** to facilitate subsequent analysis
- Data cleaning often addresses **issues**
  - structure / formatting
  - missing or corrupted values
  - unit conversion
  - encoding text as numbers
  - …
- Sadly, data cleaning is a big part of data science… (Large part of the time is spent here)

Exploratory Data Analysis (EDA)

"Getting to know the data"

- The process of **transforming**, **visualizing**, and **summarizing** data to:
  - Build/confirm understanding of the data and its provenance
  - Identify and address potential issues in the data
  - Inform the subsequent analysis
    - Journaling the data changes, or the way the data has been reformatt is an important part of the data transformation.
  - discover *potential* hypothesis … (be careful)
- **EDA is an open-ended analysis**
  - Be willing to find something surprising



John Tukey

Princeton Mathematician & Statistician

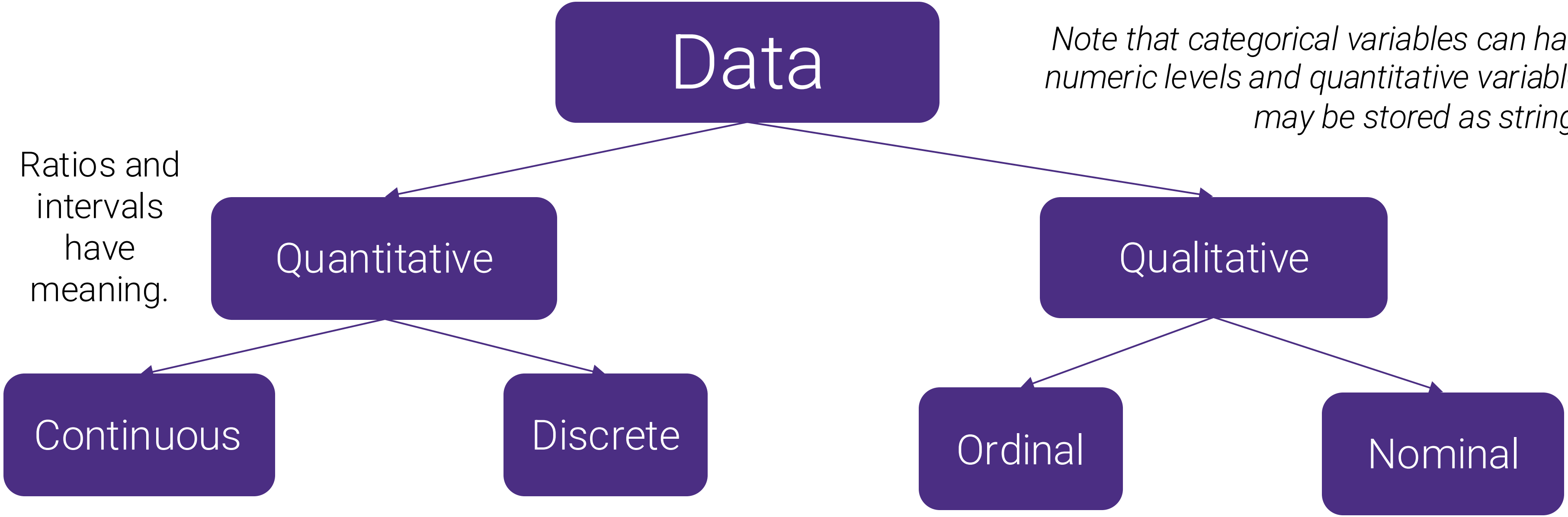
Introduced

- *“Bit” : binary digit*
- *Exploratory Data Analysis (book)*

Early Data Scientist

Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine

Data Types



Data

Note that categorical variables can have numeric levels and quantitative variables may be stored as strings.

Ratios and intervals have meaning.

Quantitative

Qualitative

Continuous

Discrete

Ordinal

Nominal

Could be measured to arbitrary precision.

Examples:

- Price
- Temperature

Finite possible values

Examples:

- Number of siblings
- Yrs of education
- Number of Lanes on Freeway

Categories w/ levels but no consistent meaning to difference

Examples:

- Preferences
- Level of education

Categories w/ no specific ordering.

Examples:

- Political Affiliation (Demo, Republic, Inc etc)
- CalID number

What is the type of data?

	Quantitative Continuous	Quantitative Discrete	Qualitative Ordinal	Qualitative Nominal
CO_2 level (PPM)	🟧			
Number of siblings		🟧		
GPA	🟧			
Income bracket (low, med, high)			🟧	
Race				🟧
Number of years of education		🟧		
Yelp Rating			🟧	

File Formats and Structure

What should we look for?

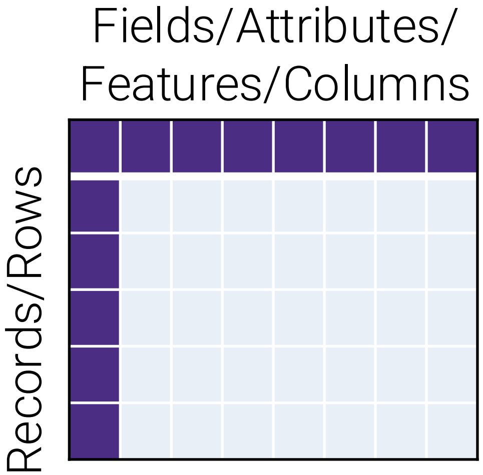
Key Data Properties to Consider in EDA

- **Structure --** *the “shape” of a data file*
- **Granularity --** *how fine/coarse is each datum*
- **Scope --** *how (in)complete is the data*
- **Temporality --** *how is the data situated in time*
- **Faithfulness --** *how well does the data capture “reality”*

Rectangular Data

We prefer rectangular data for data analysis (why?)

- Regular structures are easy manipulate and analyze
- A big part of data cleaning is about transforming data to be more rectangular

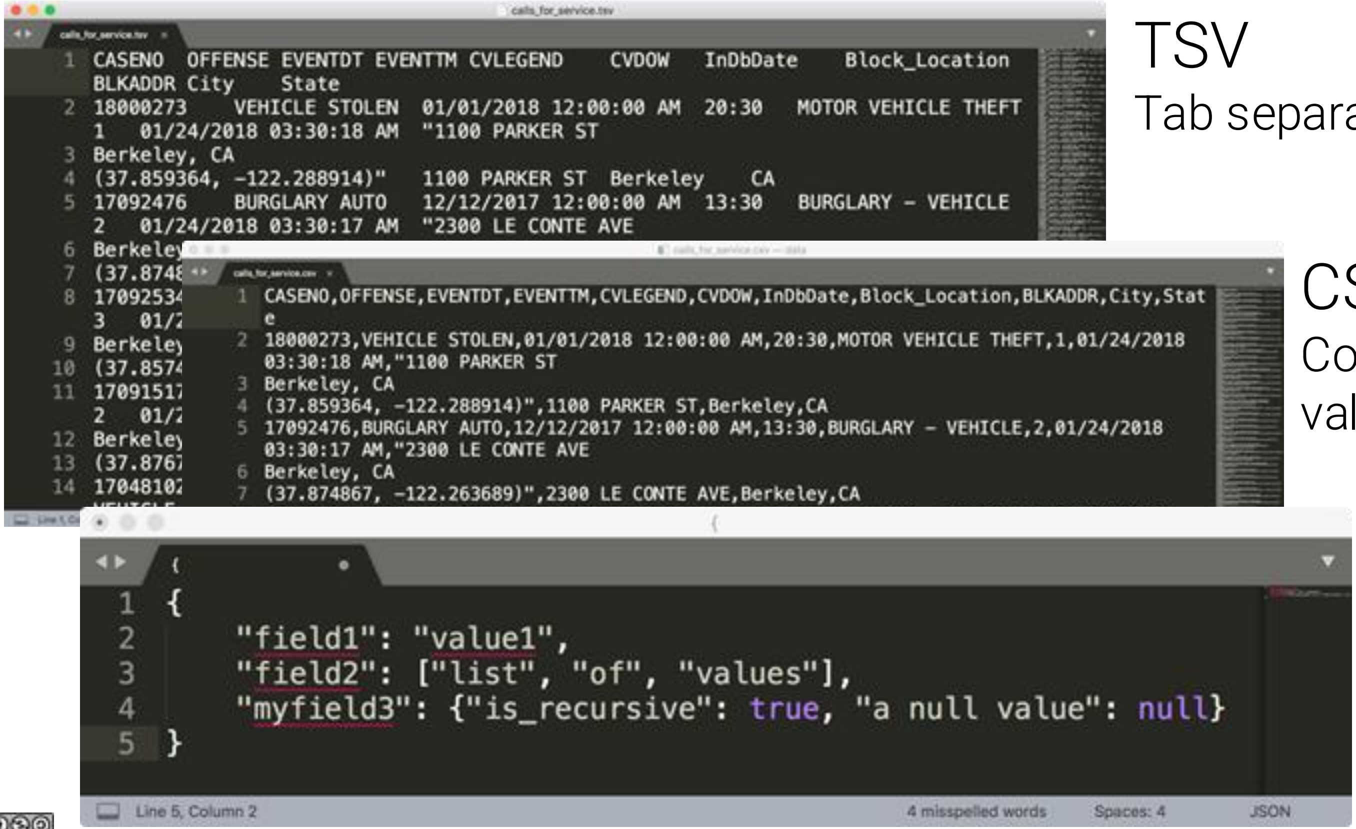


Two kinds of rectangular data: *Tables and Matrices*

(what are the differences?)

1. **Tables** (a.k.a. data-frames in R/Python and relations in SQL)
   - Named columns with different types
   - Manipulated using data transformation languages (map, filter, group by, join, …)
2. **Matrices**
   - Numeric data of the same type
   - Manipulated using linear algebra
   - Will discuss this in the future for more in-depth data science, but in this lecture.

How are these data files formatted?



```
CASENO	OFFENSE	EVENTDT	EVENTTM	CVLEGEND	CVDOW	InDbDate	Block_Location
BLKADDR	City	State
18000273	VEHICLE STOLEN	01/01/2018 12:00:00 AM	20:30	MOTOR VEHICLE THEFT
1	01/24/2018 03:30:18 AM	"1100 PARKER ST
Berkeley, CA
(37.859364, -122.288914)"	1100 PARKER ST	Berkeley	CA
17092476	BURGLARY AUTO	12/12/2017 12:00:00 AM	13:30	BURGLARY - VEHICLE
2	01/24/2018 03:30:17 AM	"2300 LE CONTE AVE
Berkeley
(37.8748
1709253
3	01/2
Berkeley
(37.857
1709151
2	01/2
Berkeley
(37.876
1704810
```

```
CASENO,OFFENSE,EVENTDT,EVENTTM,CVLEGEND,CVDOW,InDbDate,Block_Location,BLKADDR,City,State
18000273,VEHICLE STOLEN,01/01/2018 12:00:00 AM,20:30,MOTOR VEHICLE THEFT,1,01/24/2018 03:30:18 AM,"1100 PARKER ST
Berkeley, CA
(37.859364, -122.288914)",1100 PARKER ST,Berkeley,CA
17092476,BURGLARY AUTO,12/12/2017 12:00:00 AM,13:30,BURGLARY - VEHICLE,2,01/24/2018 03:30:17 AM,"2300 LE CONTE AVE
Berkeley, CA
(37.874867, -122.263689)",2300 LE CONTE AVE,Berkeley,CA
```

```
{
    "field1": "value1",
    "field2": ["list", "of", "values"],
    "myfield3": {"is_recursive": true, "a null value": null}
}
```

TSV

Tab separated values

Which is the best?

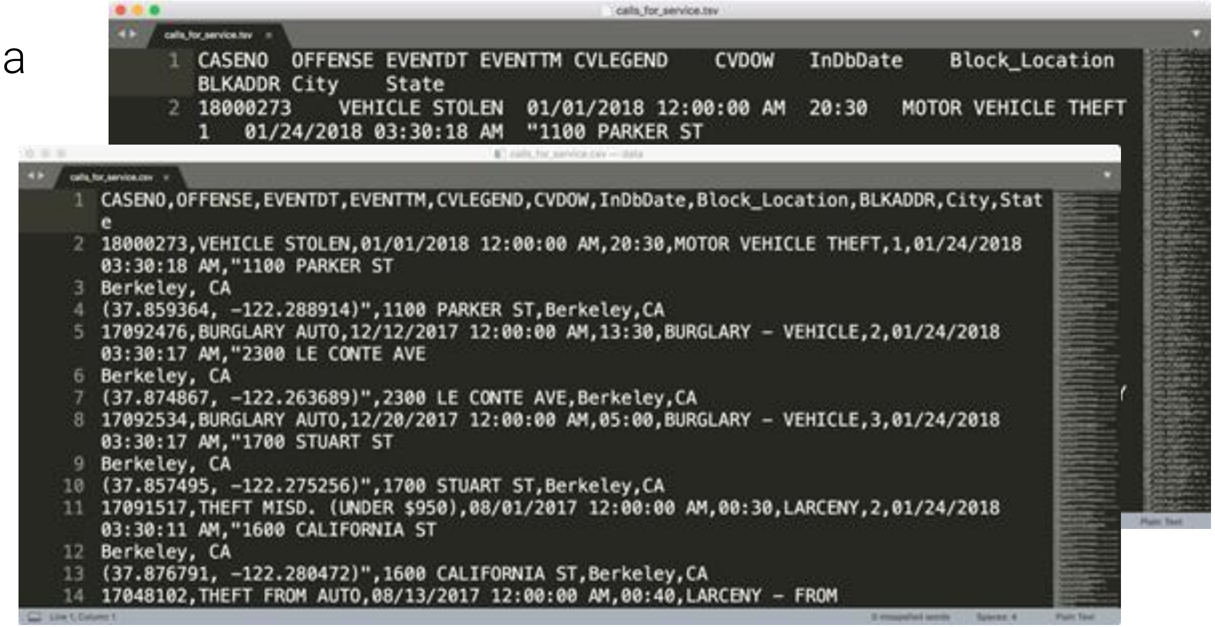
CSV

Comma separated values

JSON

Comma and Tab Separated Values Files

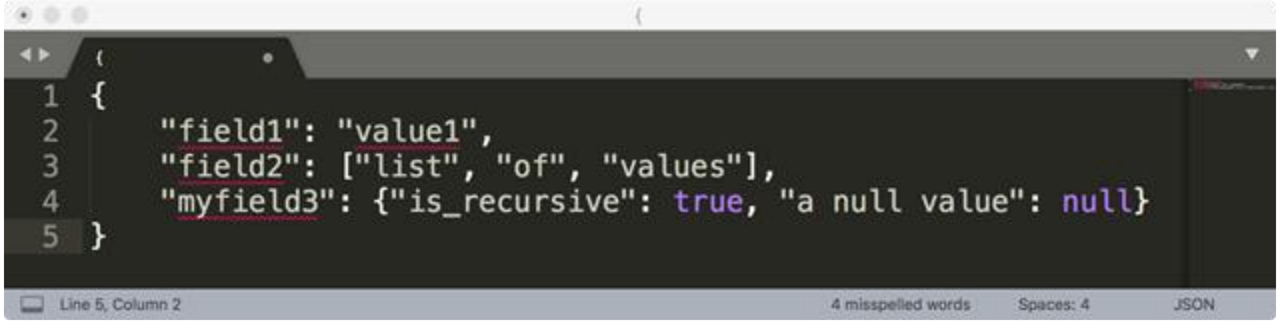
- Tabular data where
  - Records are delimited by a *newline*: “\n”, “\r\n”
  - Fields are delimited by ‘,’ (comma) or ‘\t’ (tab)
- Very Common!
- Issues?
  - Commas, tabs in records
  - Quoting
  - …



```
CASENO  OFFENSE EVENTDT EVENTTM CVLEGEND    CVDOW   InDbDate    Block_Location BLKADDR City    State
18000273    VEHICLE STOLEN  01/01/2018 12:00:00 AM  20:30   MOTOR VEHICLE THEFT 1   01/24/2018 03:30:18 AM  "1100 PARKER ST
```

```
CASENO,OFFENSE,EVENTDT,EVENTTM,CVLEGEND,CVDOW,InDbDate,Block_Location,BLKADDR,City,State
18000273,VEHICLE STOLEN,01/01/2018 12:00:00 AM,20:30,MOTOR VEHICLE THEFT,1,01/24/2018 03:30:18 AM,"1100 PARKER ST
Berkeley, CA
(37.859364, -122.288914)",1100 PARKER ST,Berkeley,CA
17092476,BURGLARY AUTO,12/12/2017 12:00:00 AM,13:30,BURGLARY - VEHICLE,2,01/24/2018 03:30:17 AM,"2300 LE CONTE AVE
Berkeley, CA
(37.874867, -122.263689)",2300 LE CONTE AVE,Berkeley,CA
17092534,BURGLARY AUTO,12/20/2017 12:00:00 AM,05:00,BURGLARY - VEHICLE,3,01/24/2018 03:30:17 AM,"1700 STUART ST
Berkeley, CA
(37.857495, -122.275256)",1700 STUART ST,Berkeley,CA
17091517,THEFT MISD. (UNDER $950),08/01/2017 12:00:00 AM,00:30,LARCENY,2,01/24/2018 03:30:11 AM,"1600 CALIFORNIA ST
Berkeley, CA
(37.876791, -122.280472)",1600 CALIFORNIA ST,Berkeley,CA
17048102,THEFT FROM AUTO,08/13/2017 12:00:00 AM,00:40,LARCENY - FROM
```

JavaScript Object Notation (JSON)



```
{
    "field1": "value1",
    "field2": ["list", "of", "values"],
    "myfield3": {"is_recursive": true, "a null value": null}
}
```

- Widely used file format for nested data
  - Very similar to python dictionaries
  - Strict formatting ”quoting” addresses some issues in CSV/TSV
- Issues
  - Not rectangular
  - Each record can have different fields
  - Nesting means records can contain tables – complicated

Extensible Markup Language - XML (another kind of nested data)

```
<catalog>
  <plant type='a'>
    <common>Bloodroot</common>
    <botanical>Sanguinaria canadensis</botanical>
    <zone>4</zone>
    <light>Mostly Shady</light>
    <price>2.44</price>
    <availability>03/15/2006</availability>
    <description>
          <color>white</color>
          <petals>true</petals>
    </description>
    <indoor>true</indoor>
  </plant>
…
</catalog>
```

Nested structure

Log Data

Is this a csv file? tsv? JSON/XML?

```
169.237.46.168 - - [26/Jan/2014:10:47:58 -0800] "GET
/stat141/Winter04 HTTP/1.1" 301 328
"http://anson.ucdavis.edu/courses/"  "Mozilla/4.0 (compatible; MSIE
6.0; Windows NT 5.0; .NET CLR 1.1.4322)”
```

```
169.237.6.168 - - [8/Jan/2014:10:47:58 -0800] "GET
/stat141/Winter04/ HTTP/1.1" 200 2585
"http://anson.ucdavis.edu/courses/" "Mozilla/4.0 (compatible; MSIE
6.0; Windows NT 5.0; .NET CLR 1.1.4322)"
```

Keys and Joins

Structure: Keys

- Often data will reference other pieces of data
- **Primary key**: *the column or set of columns in a table that determine the values of the remaining columns*
  - Primary keys are unique
  - Examples: SSN, ProductIDs, …

Primary Key

Purchases.csv

OrderNum	ProdID	Quantity
1	42	3
1	999	2
2	42	1

Orders.csv

OrderNum	CustID	Date
1	171345	8/21/2017
2	281139	8/30/2017

Products.csv

ProdID	Cost
42	3.14
999	2.72

Primary Key

Customers.csv

CustID	Addr
171345	Harmon..
281139	Main ..

Structure: Keys

- Often data will reference other pieces of data
- **Primary key**: *the column or set of columns in a table that determine the values of the remaining columns*
  - Primary keys are unique
  - Examples: SSN, ProductIDs, …
- **Foreign keys**: the column or sets of columns that reference primary keys in other tables.
- You will need to **join** across tables

Primary Key

Purchases.csv

OrderNum	ProdID	Quantity
1	42	3
1	999	2
2	42	1

Foreign Key

Orders.csv

OrderNum	CustID	Date
1	171345	8/21/2017
2	281139	8/30/2017

Products.csv

ProdID	Cost
42	3.14
999	2.72

Primary Key

Customers.csv

CustID	Addr
171345	Harmon..
281139	Main ..

Questions to ask about *Structure*

- Are the data in a standard format or encoding?
  - **Tabular data**: CSV, TSV, Excel, SQL
  - **Nested data**: JSON or XML
- Are the data organized in “records”?
  - No: Can we define records by parsing the data?
- Are the data nested? (records contained within records…)
  - Yes: Can we reasonably un-nest the data?
- Does the data reference other data?
  - Yes: can we join/merge the data
- What are the fields in each record?
  - How are they encoded? (e.g., strings, numbers, binary, dates …)
  - What is the **type** of the data?

Concepts and Terminology of Tables in Data Mining

- There are two types of tables in Data Mining: Fact Table and Dimensional Table
    - **Fact table: contains the main data.**
    - **Dimensional table: supports the fact table**

- Examples of Fact Table
    - Call Service Record
    - Medical Tests

- Examples of Dimensional Table
    - Types of service calls
    - Referenced by Fact table by foreign keys

Summary

Summary: How do you do EDA/Data Cleaning?

- Examine data and metadata:
    - What is the date, size, organization, and structure of the data?
- Examine each field/attribute/dimension individually
- Examine pairs of related dimensions
    - Stratifying earlier analysis: break down grades by major …
- Along the way:
    - Visualize/summarize the data
    - Validate assumptions about data and collection process
    - Identify and address anomalies
    - Apply data transformations and corrections
    - ***Record everything you do! (why?)***