

LECTURE 8

Data Cleaning and EDA

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

Today



Congratulations!



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

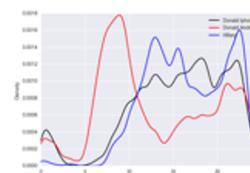
What do you do next?

Question &
Problem
Formulation



Data
Acquisition

Prediction
and
Inference



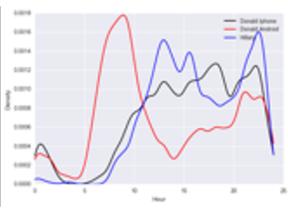
Exploratory
Data
Analysis



Data Acquisition

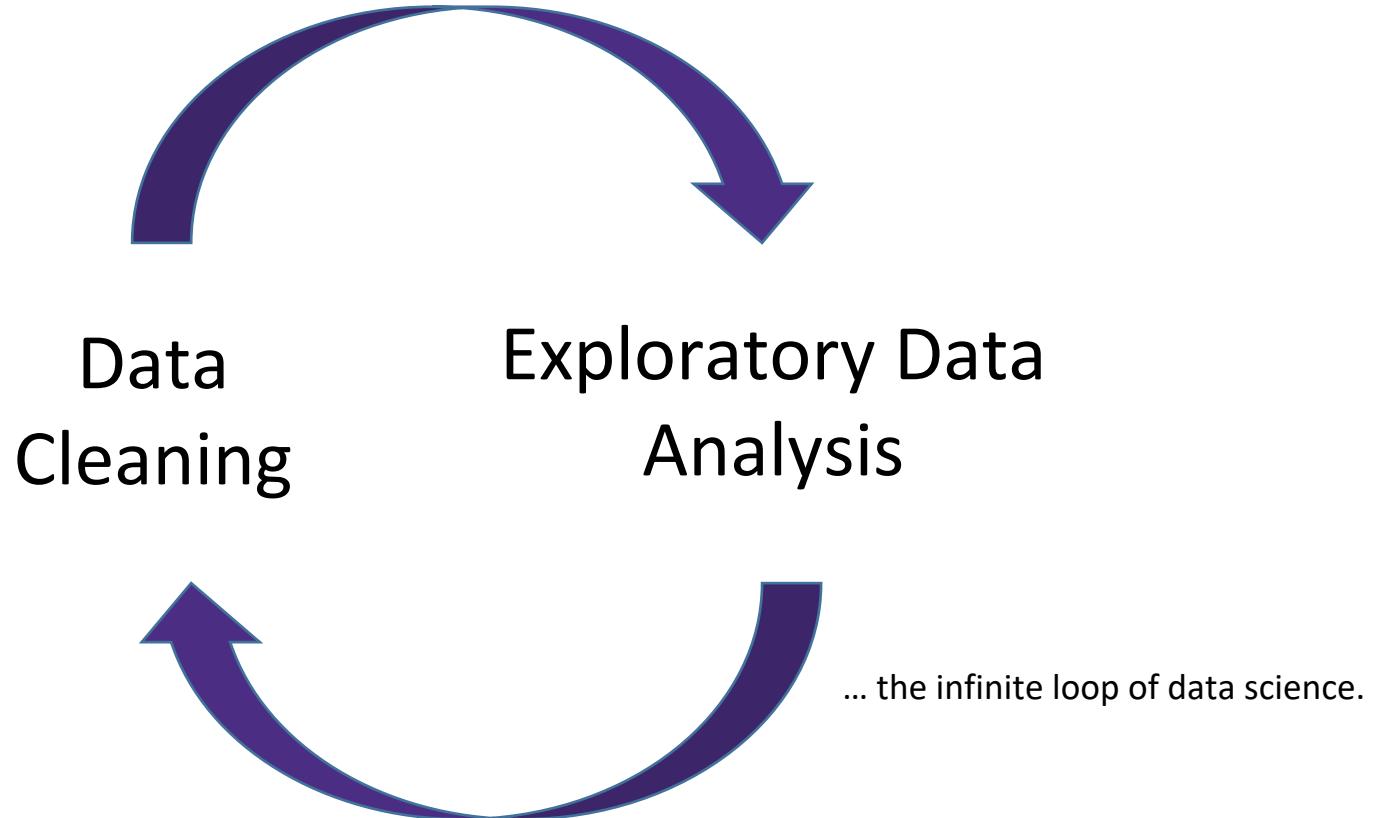


Exploratory Data Analysis



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

- 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...



Big Data
Borat

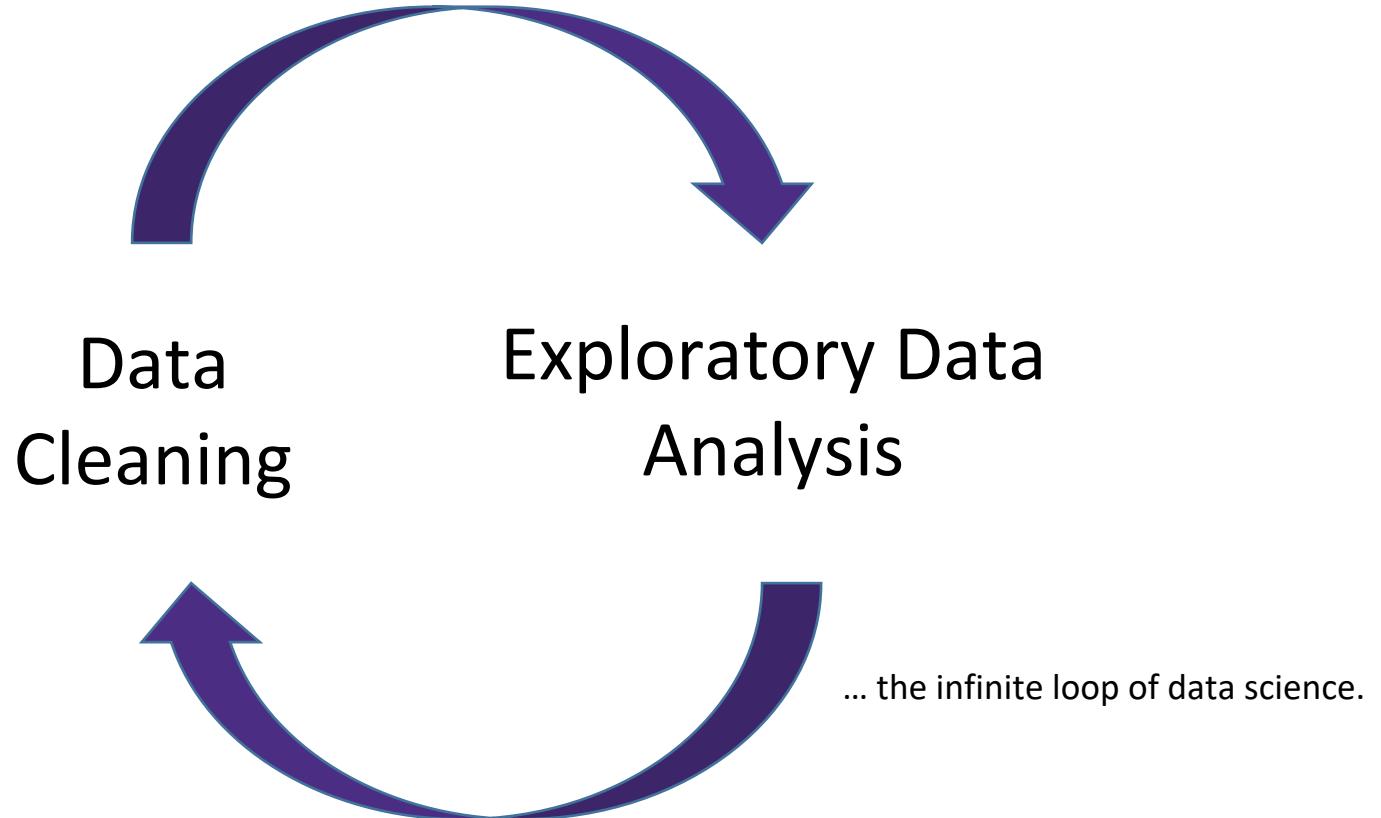
@BigDataBorat



Following

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

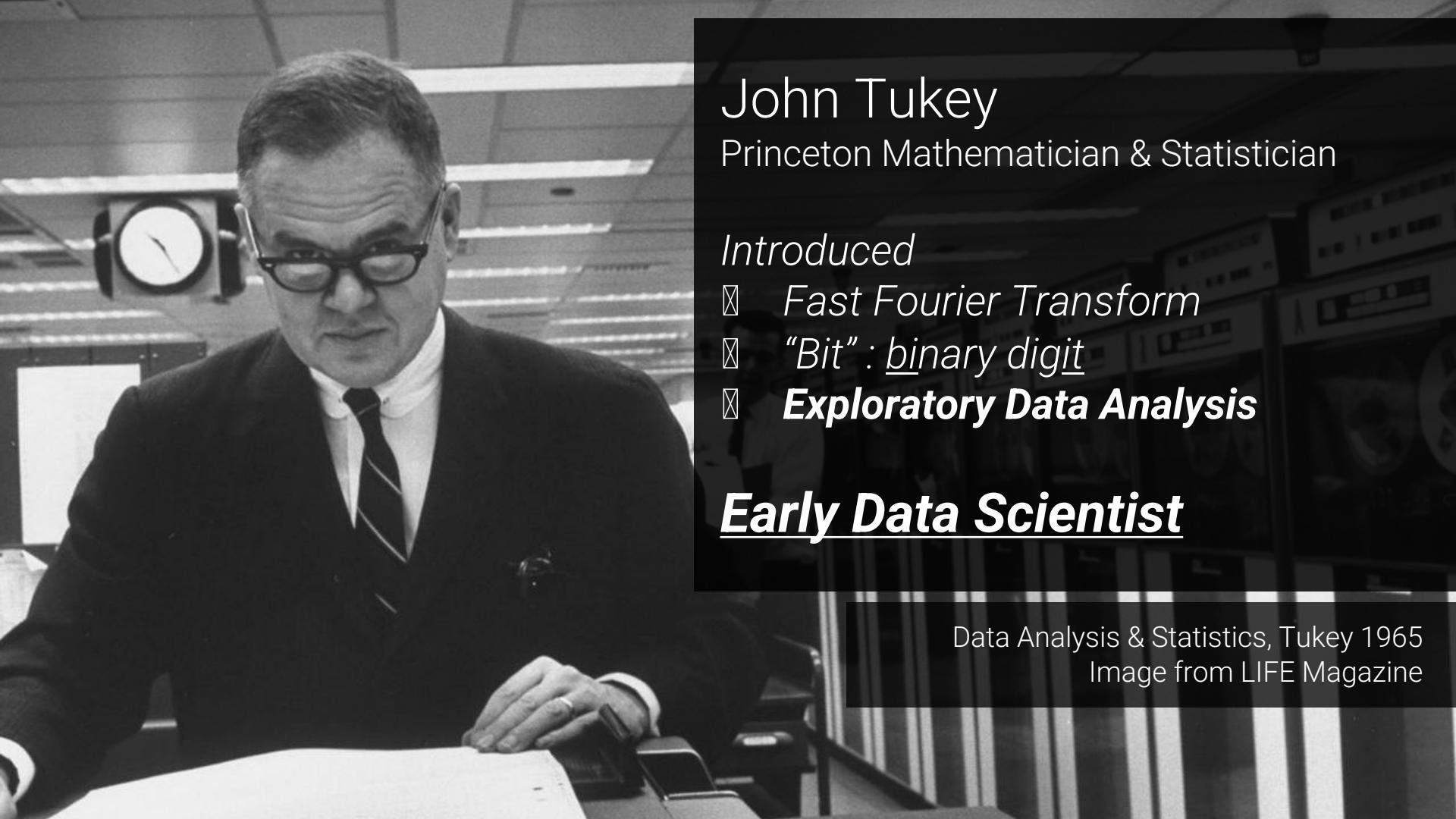




Exploratory Data Analysis (EDA)

“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
 - discover *potential* hypothesis ... (be careful)
- **EDA is an open-ended analysis**
 - Be willing to find something surprising



John Tukey

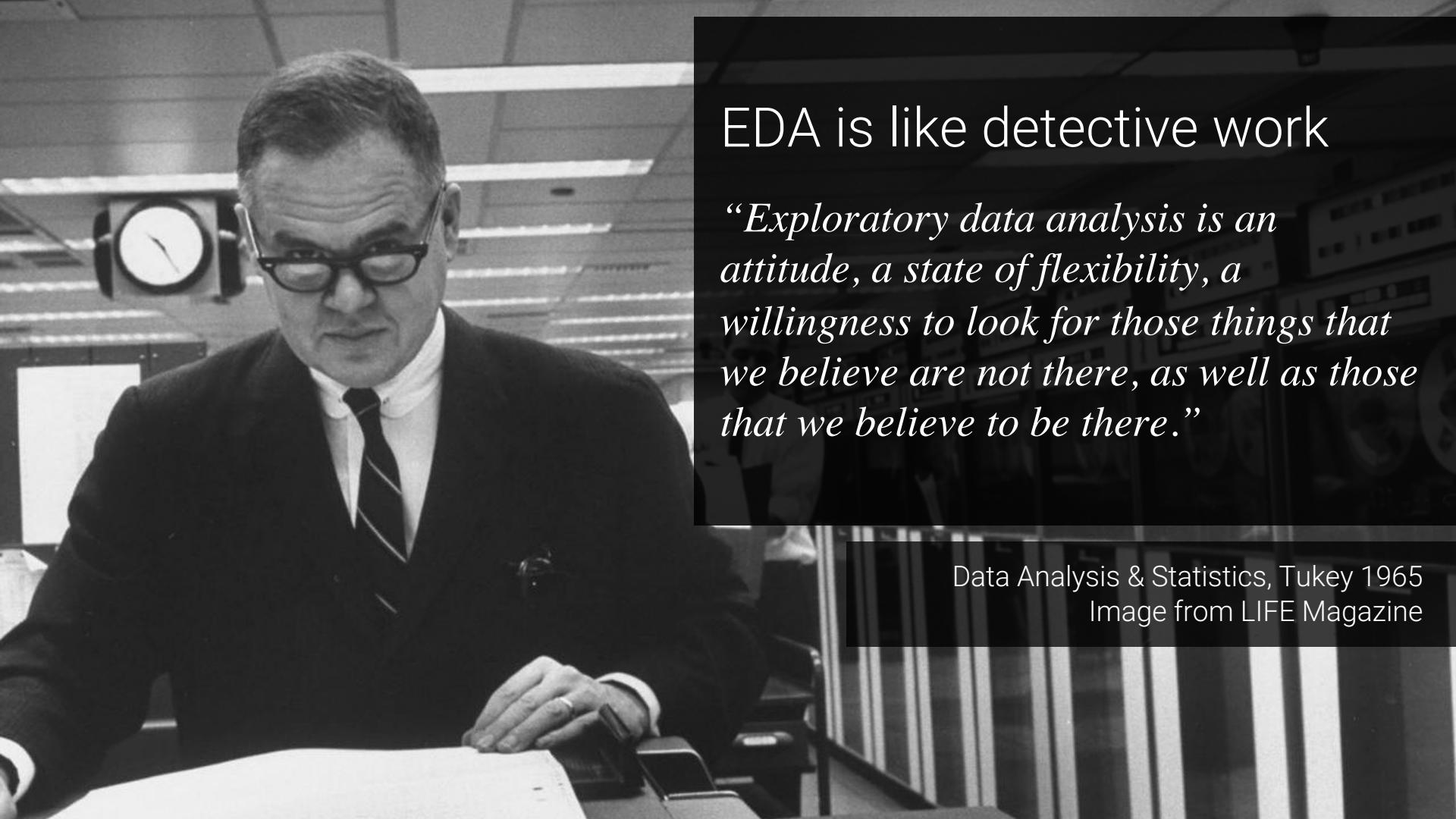
Princeton Mathematician & Statistician

Introduced

- ☒ *Fast Fourier Transform*
- ☒ “Bit” : *binary digit*
- ☒ ***Exploratory Data Analysis***

Early Data Scientist

Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine



EDA is like detective work

“Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those that we believe to be there.”

Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine

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”*

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?)

Fields/Attributes/
Features/Columns

Records/Rows	
--------------	--

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?)

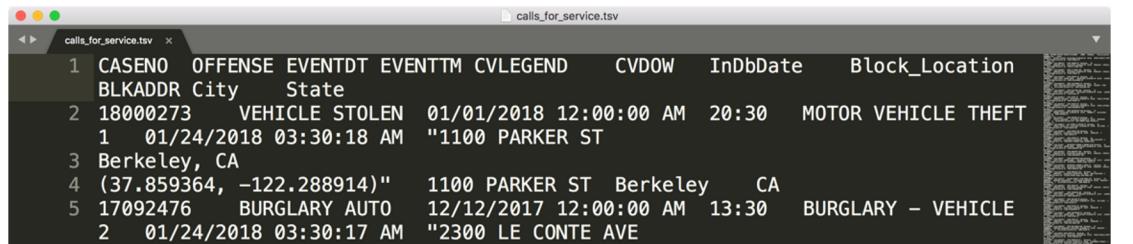
1. **Tables** (a.k.a. data-frames in R/Python and relations in SQL)
 - Named columns with different types
 - Manipulated using data transformation languages (map, filter, group by, join, ...)
 2. **Matrices**
 - Numeric data of the same type
 - Manipulated using linear algebra

Fields/Attributes/ Features/Columns

A 6x6 grid illustrating a repeating pattern of alternating dark purple and light gray cells. The first column consists of five dark purple cells. The second column has one light gray cell followed by five dark purple cells. This sequence of one light gray cell and five dark purple cells is repeated across all six columns of the grid.



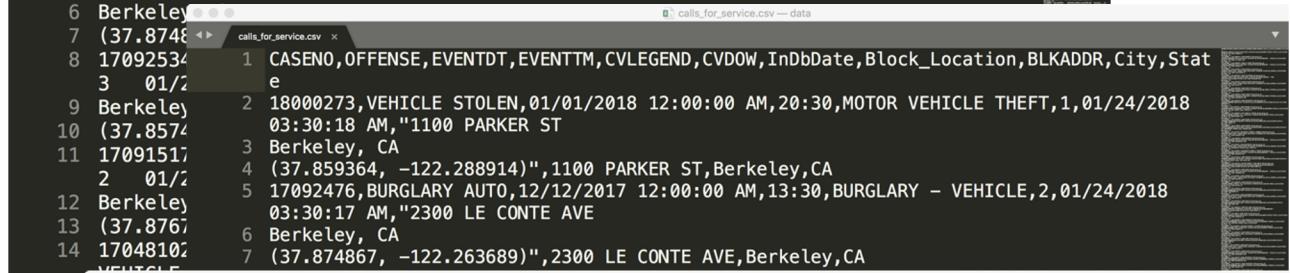
How are these data files formatted?



```
1 CASENO OFFENSE EVENTDT EVENTTM CVLEGEND CVDOW InDbDate Block_Location
BLKADDR City State
2 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
3 Berkeley, CA
4 (37.859364, -122.288914)" 1100 PARKER ST Berkeley CA
5 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
6 Berkeley
7 (37.874867, -122.263689)" 2300 LE CONTE AVE, Berkeley, CA
```

TSV

Tab separated values

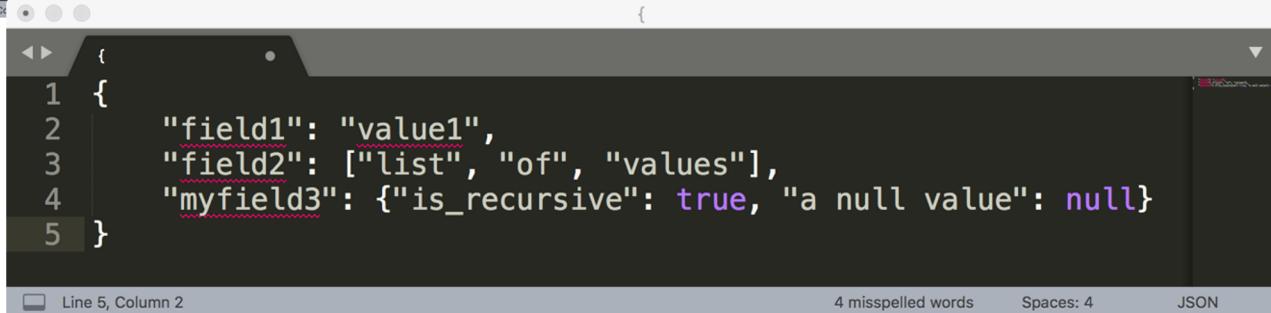


```
1 CASENO,OFFENSE,EVENTDT,EVENTTM,CVLEGEND,CVDOW,InDbDate,Block_Location,BLKADDR,City,Stat
2 18000273,VEHICLE STOLEN,01/01/2018 12:00:00 AM,20:30,MOTOR VEHICLE THEFT,1,01/24/2018
3 01/24/2018 03:30:18 AM,"1100 PARKER ST
4 Berkeley, CA
5 (37.859364, -122.288914)",1100 PARKER ST,Berkeley,CA
6 17092476,BURGLARY AUTO,12/12/2017 12:00:00 AM,13:30,BURGLARY - VEHICLE,2,01/24/2018
7 01/24/2018 03:30:17 AM,"2300 LE CONTE AVE
8 (37.874867, -122.263689)",2300 LE CONTE AVE,Berkeley,CA
```

CSV

Comma separated
values

Which is
the best?



```
1 {
2   "field1": "value1",
3   "field2": ["list", "of", "values"],
4   "myfield3": {"is_recursive": true, "a null value": null}
5 }
```

Line 5, Column 2 4 misspelled words Spaces: 4 JSON

JSON

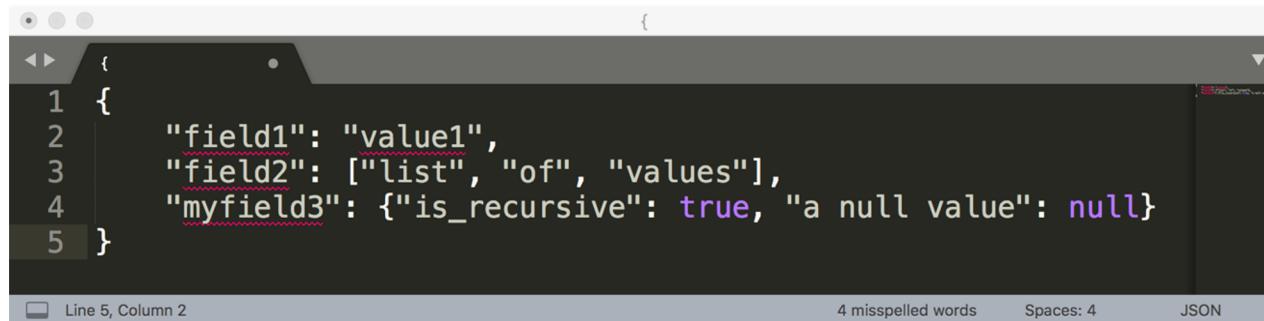
Comma and Tab Separated Values Files

- 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
 - ...

```
calls_for_service.tsv
1 CASENO OFFENSE EVENTDT EVENTTM CVLEGEND CVDOW InDbDate Block_Location
BLKADDR City State
2 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

calls_for_service.csv
1 CASENO,OFFENSE,EVENTDT,EVENTTM,CVLEGEND,CVDOW,InDbDate,Block_Location,BLKADDR,City,Stat
e
2 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
3 Berkeley, CA
4 (37.859364, -122.288914)",1100 PARKER ST,Berkeley,CA
5 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
6 Berkeley, CA
7 (37.874867, -122.263689)",2300 LE CONTE AVE,Berkeley,CA
8 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
9 Berkeley, CA
10 (37.857495, -122.275256)",1700 STUART ST,Berkeley,CA
11 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
12 Berkeley, CA
13 (37.876791, -122.280472)",1600 CALIFORNIA ST,Berkeley,CA
14 17048102,THEFT FROM AUTO,08/13/2017 12:00:00 AM,00:40,LARCENY - FROM
```

JavaScript Object Notation (JSON)



A screenshot of a code editor window displaying JSON code. The code is as follows:

```
{  
1 {  
2   "field1": "value1",  
3   "field2": ["list", "of", "values"],  
4   "myfield3": {"is_recursive": true, "a null value": null}  
5 }
```

The code editor interface includes a status bar at the bottom with the following information: Line 5, Column 2; 4 misspelled words; Spaces: 4; and JSON.

- 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

Primary Key



- 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: ID, ProductIDs, ...

<u>OrderNum</u>	<u>ProdID</u>	Quantity
1	42	3
1	999	2
2	42	1

Orders.csv

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

Products.csv

<u>ProdID</u>	Cost
42	3.14
999	2.72

Primary Key



Customers.csv

<u>CustID</u>	Addr
171345	Harmon..
281139	Main ..

Structure: Keys

Primary Key



- 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: ID, ProductIDs, ...
- **Foreign keys:** the column or sets of columns that reference primary keys in other tables.
- You will need to **join** across tables

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

Foreign Key



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

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?

Variable Types

Variable

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

Ratios and intervals have meaning.

Quantitative

Continuous

Could be measured to arbitrary precision.

Examples:

- Price
- Temperature

Discrete

Finite possible values

Examples:

- Number of siblings
- Yrs of education

Qualitative

Ordinal

Categories w/ levels but no consistent meaning to difference

Examples:

- Movie rating
- Level of education

Nominal

Categories w/ no specific ordering.

Examples:

- Gender/ Race
- ID

What is the type of variable?

	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				
Restaurant Rating				

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”*

Granularity



Fine
Grained

Rec. 1



Rec. 2



Rec. 3



Rec. 1



Coarse
Grained



- What does each record represent?
 - Examples: a purchase, a person, a group of users
- Do all records capture granularity at the same level?
 - Some data will include summaries (aka rollups) as records
- If the data are coarse how was it aggregated?
 - Sampling, averaging, ...

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”*

Scope

- Does my data cover my area of interest?
 - **Example:** *I am interested in studying crime in Shanghai but I only have Minhang District data.*
- Is my data too expansive?
 - **Example:** *I am interested in student grades for ECE 4710J but have student grades for all classes.*
 - **Solution:** *Filtering ⇒ Implications on sample?*
 - *If the data is a sample I may have poor coverage after filtering ...*
- Does my data cover the right time frame?
 - More on this in temporality ...

Sampling Frame

- The **sampling frame** is the **population** from which the data was **sampled**.
 - Note that this **may not be** the **population** of interest.
- How complete/incomplete is the frame (and its data)?
- How is the frame/data situated in place?
- How well does the frame/data capture reality?
- How is the frame/data situated in time?

Key Data Properties to Consider in EDA

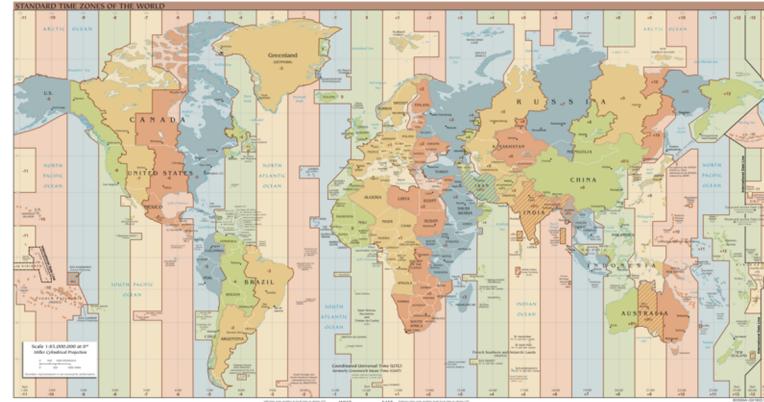
- **Structure** -- *the “shape” of a data file*
- **Granularity** -- *how fine/coarse is each datum*
- **Scope** -- *how (in)complete is the data*
- **Temporality** -- *how is the data situated in time*
- **Faithfulness** -- *how well does the data capture “reality”*

Temporality

- Data changes – when was the data collected?
- What is the meaning of the time and date fields?
 - When the “event” **happened**?
 - When the data was **collected** or was **entered** into the system?
 - Date the data was copied into a database (look for many matching timestamps)
- Time depends on where! (Time zones)
 - Learn to use **datetime** python library
 - Multiple string representation (depends on region): 07/08/09?
- Are there strange null values?
 - January 1st 1970 (Unix Epoch), January 1st 1900
- Is there periodicity? Diurnal patterns

Unix Time / POSIX Time

- Time **measured in seconds** since January 1st 1970
 - Minus leap seconds ...
- Unix time follows Coordinated Universal Time (UTC)
 - International time standard
 - Measured at 0 degrees latitude
 - Similar to Greenwich Mean Time (GMT)
 - No daylight savings
 - Time codes
- Time Zones:
 - Beijing (UTC+8)



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

Faithfulness and Missing Values

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”*

Faithfulness: *Do I trust this data?*

- Does my data contain **unrealistic** or “**incorrect**” values?
 - Dates in the future for events in the past
 - Locations that don’t exist
 - Negative counts
 - Misspellings of names
 - Large outliers
- Does my data violate **obvious dependencies**?
 - E.g., age and birthday don’t match
- Was the data **entered by hand**?
 - Spelling errors, fields shifted ...
 - Did the form require fields or provide default values?
- Are there obvious signs of **data falsification**:
 - Repeated names, fake looking email addresses, repeated use of uncommon names or fields.

Signs that your data may not be faithful

- Missing Values/Default values?
 - What do they look like?
 - “ ”,
 - 0,
 - -1, 999, 12345,
 - NaN, Null,
 - 1970, 1900

What to do with the Missing Values?

- **Drop records** with missing values
 - Probably most common
 - **Caution:** check for biases introduced by dropped values
 - Missing or corrupt records might be related to something of interest
- **Imputation:** (Inferring missing values)
 - **Mean Imputation:** replace with an average value
 - Which mean? Often use closest related subgroup mean.
 - **Hot deck imputation:** replace with a random value
 - Choose a random value from the subgroup and use it for the missing value.
- **Some Suggestion:**
 - Drop missing values **but check for induced bias (use domain knowledge)**
 - Directly **model missing values** during future analysis

Signs that your data may not be faithful

- Missing Values or **default** values
- Truncated data (early excel limits: 65536 Rows, 255 Columns)
 - **Soln:** be aware of consequences in analysis ⇒ how did truncation affect sample?
- Time Zone Inconsistencies
 - **Soln 1:** convert to a common timezone (e.g., UTC)
 - **Soln 2:** convert to the timezone of the location – useful in modeling behavior.
- Duplicated Records or Fields
 - **Soln:** identify and eliminate (use primary key) ⇒ implications on sample?
- Spelling Errors
 - **Soln:** Apply corrections or drop records not in a dictionary ⇒ implications on sample?
- Units not specified or consistent
 - **Solns:** Infer units, check values are in reasonable ranges for data
- Others...

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?)***