LECTURE 6

# Data Cleaning and EDA

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

**Data 100/Data 200, Fall 2021 @ UC Berkeley**
Fernando Pérez and Alvin Wan
(content by Joseph Gonzalez, Deborah Nolan, and Joseph Hellerstein)

# Previously …

http://abcnews.go.com/Lifestyle/silly-baby-panda-falls-flat-face-public-debut/story?id=42481478

# Pandas and Jupyter Notebooks

- Introduced 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

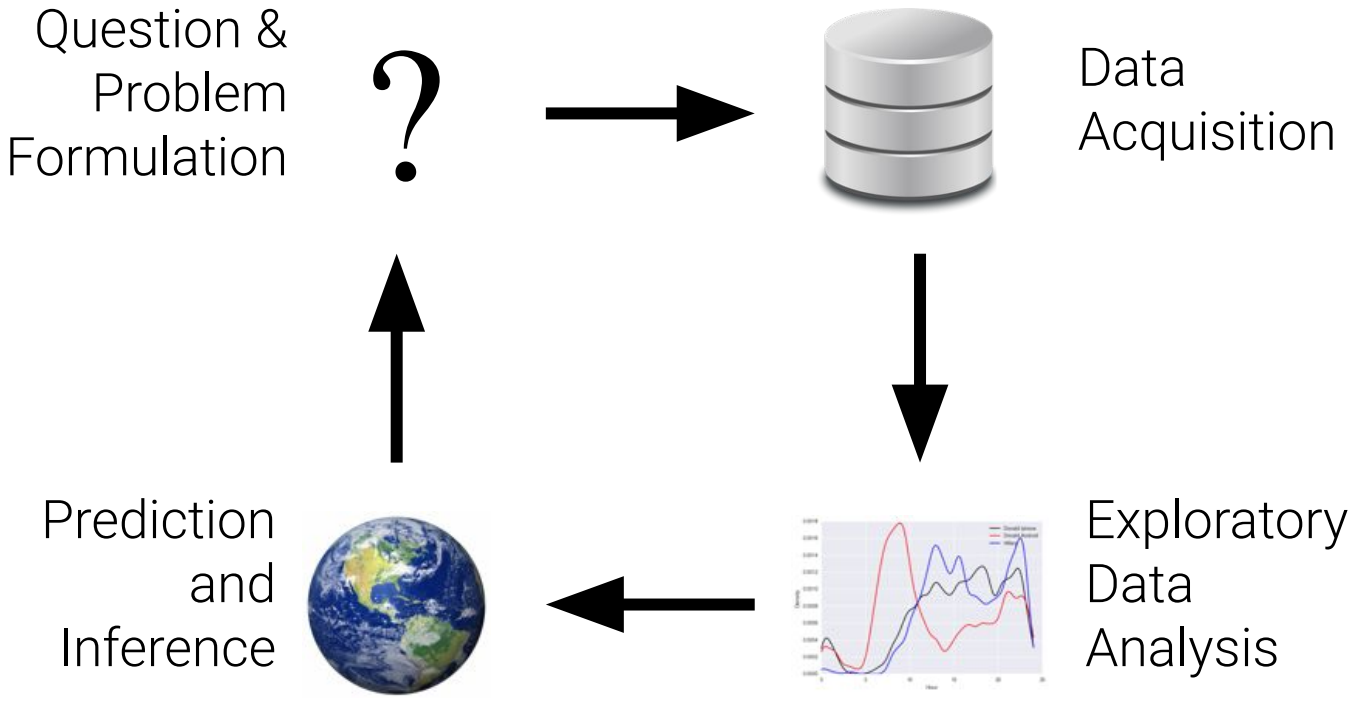
- Dataframe access methods
  - **Filtering** on predicts and **slicing**
  - **df.loc**: location by index
  - **df.iloc**: location by integer address
  - **groupby** & **pivot** aggregating data

# Today

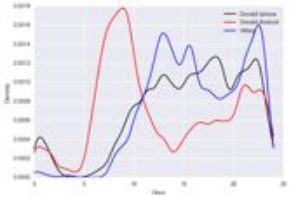
# Congratulations!


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

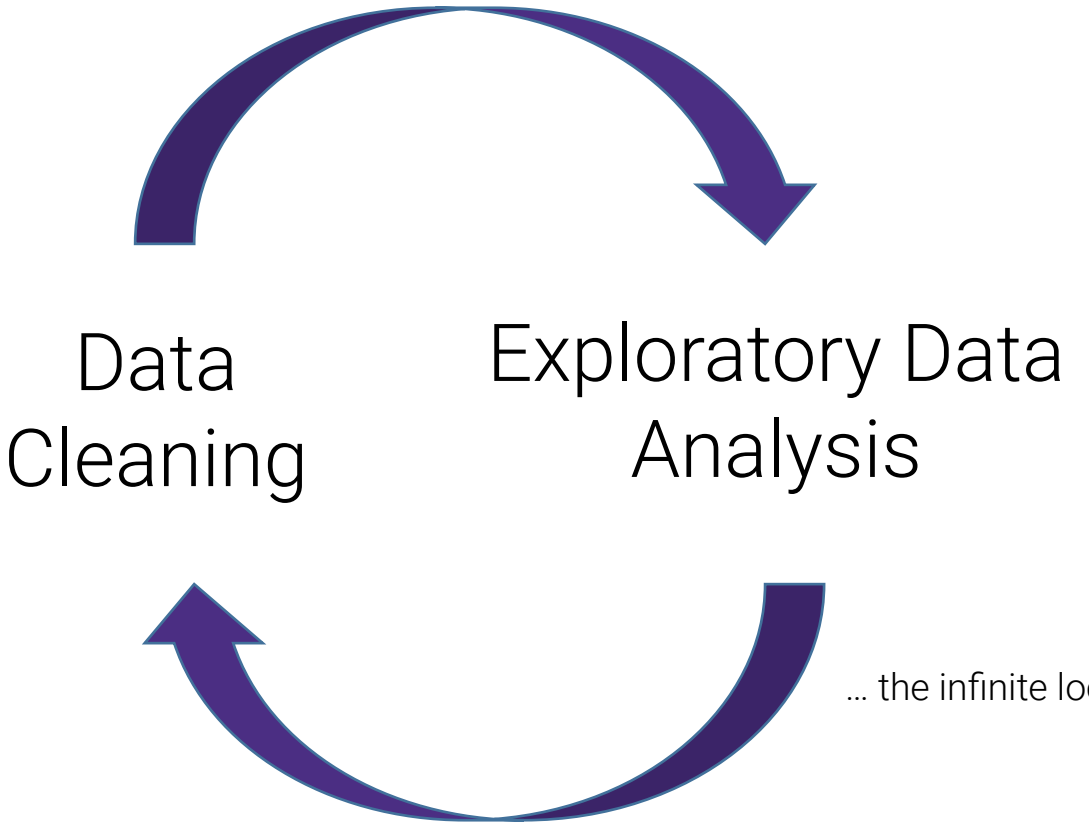
Data Acquisition



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




… the infinite loop of data science.

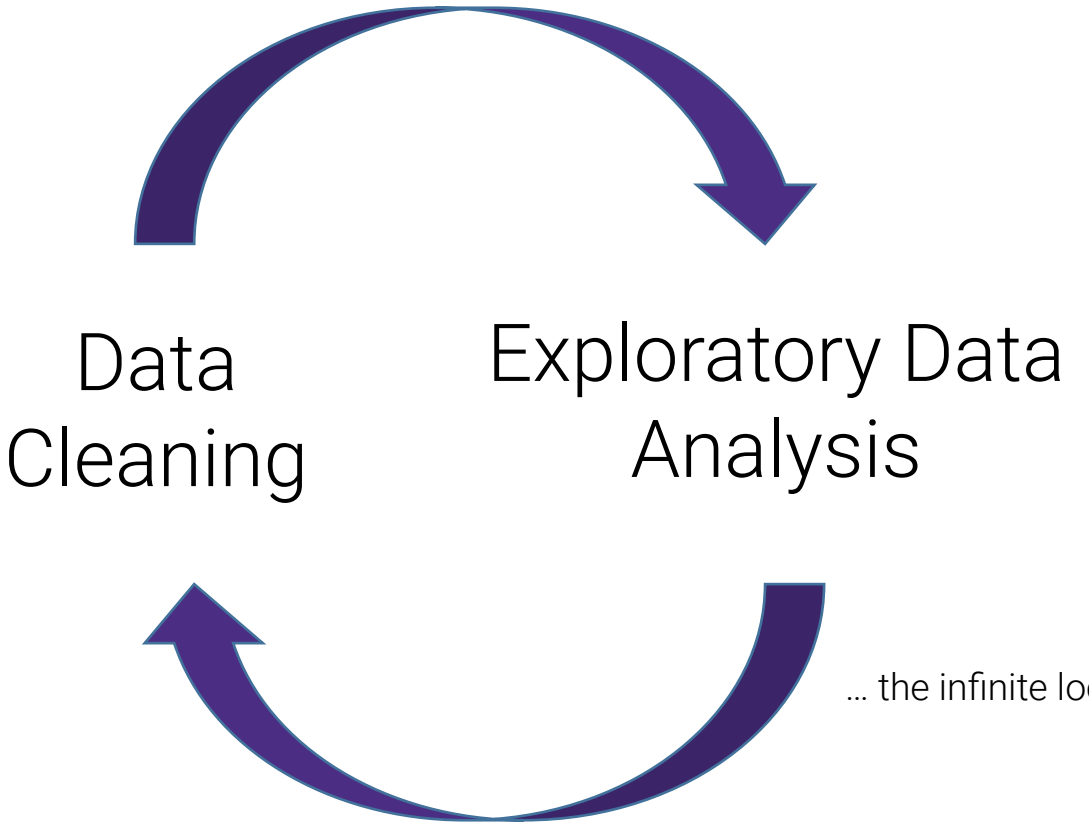
# Data Cleaning

- 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…

**Big Data Borat**
@BigDataBorat

Following

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

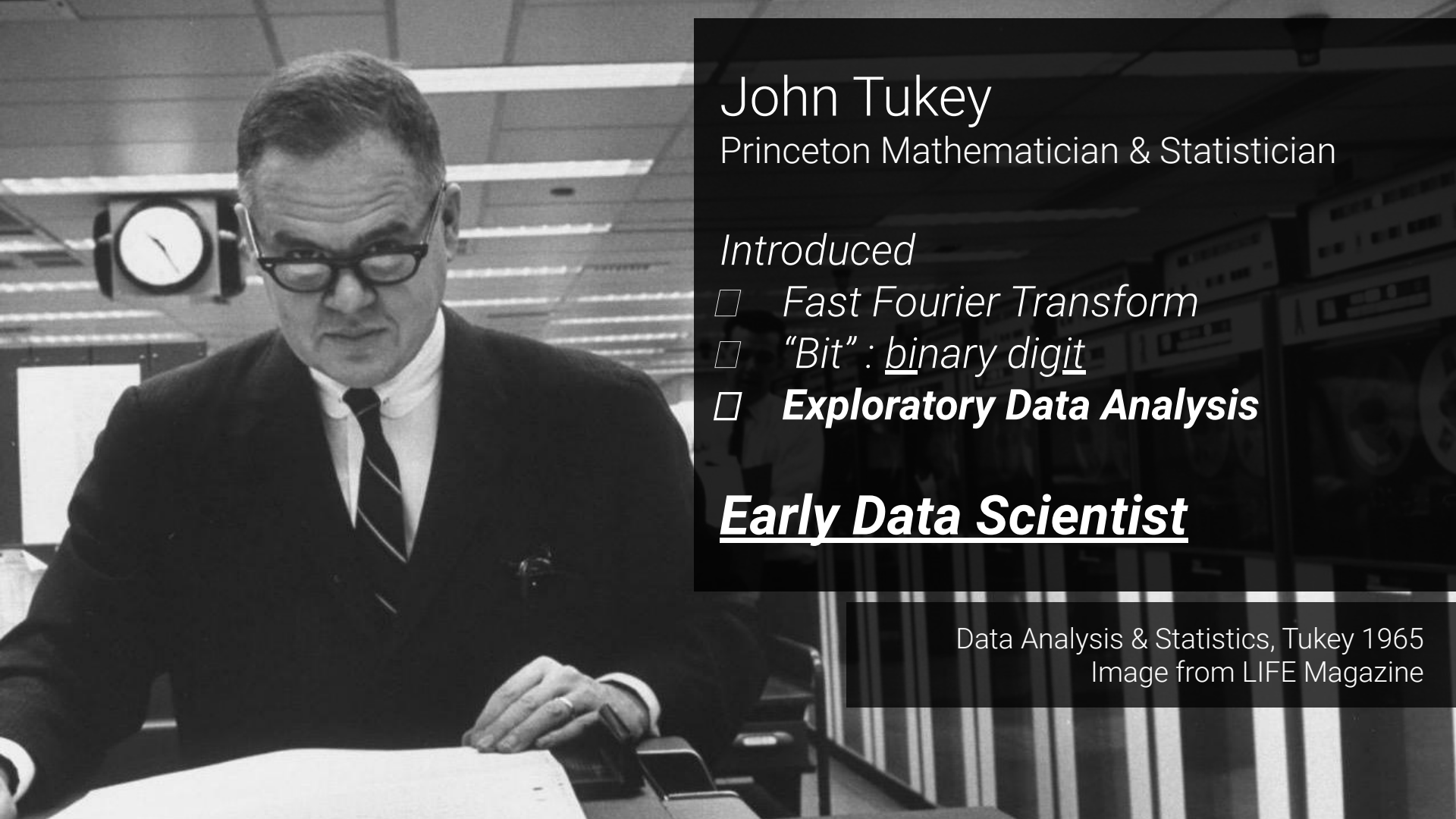



… the infinite loop of data science.

# 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

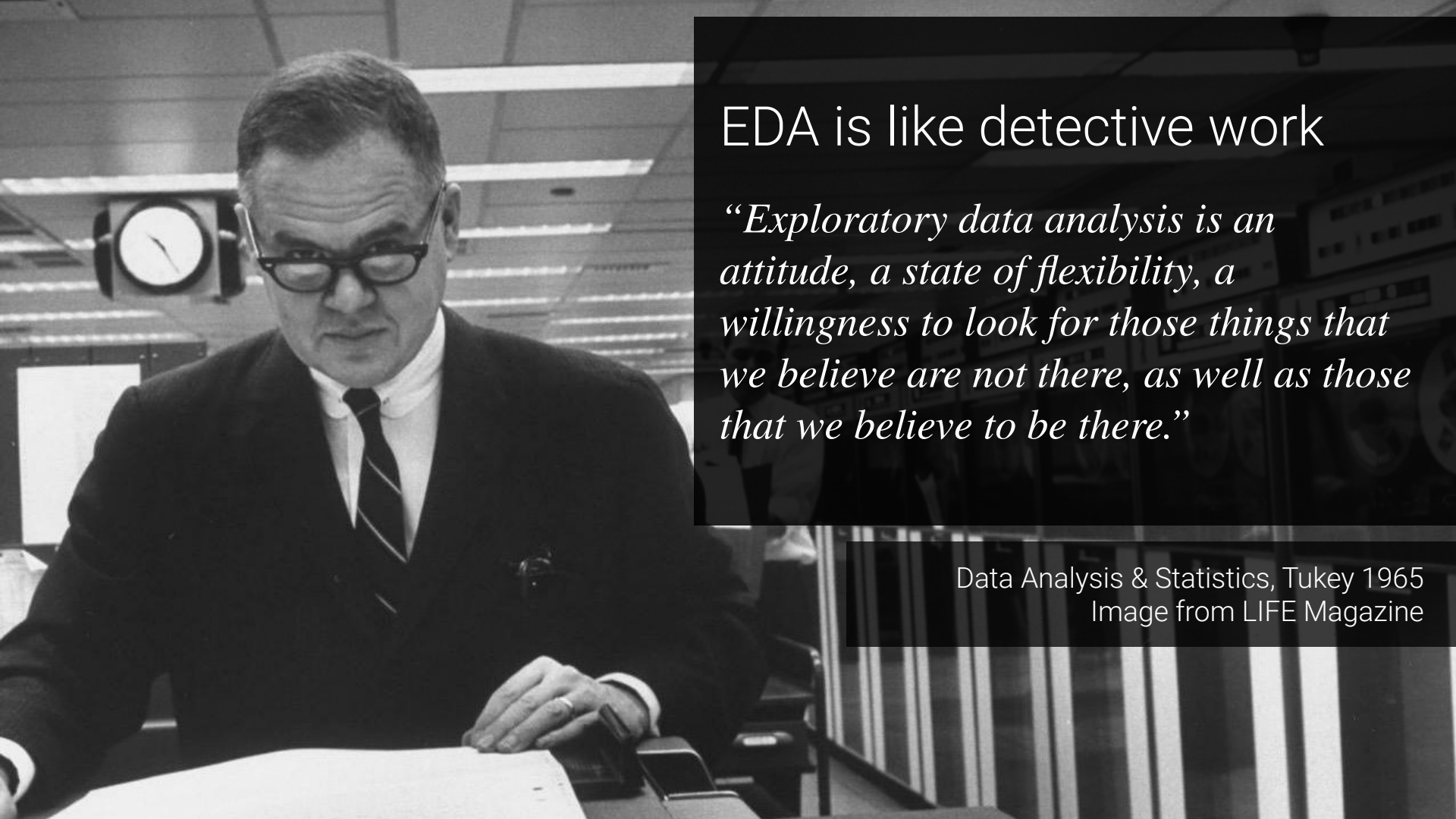
Princeton Mathematician & Statistician

*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"*

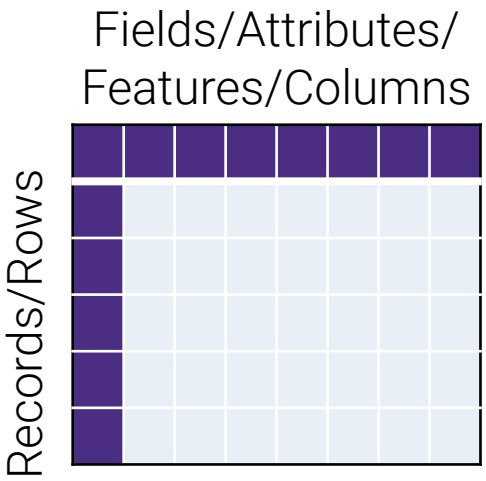
# Rectangular Data

We prefer rectangular data for data analysis (why?)

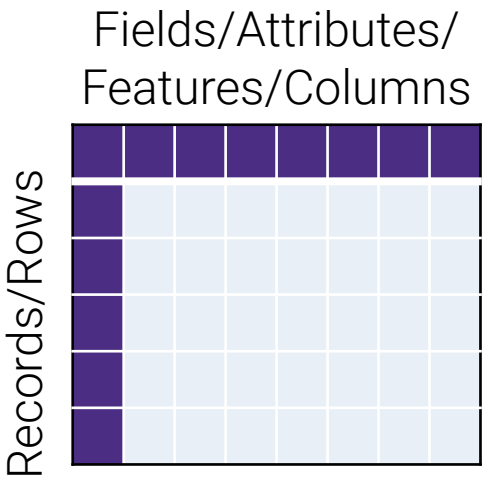
- Regular structures are easy manipulate and analyze
- A big part of data cleaning is about transforming data to be more rectangular

Two kinds of rectangular data: *Tables and Matrices*

(what are the differences?)



# 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

# How are these data files formatted?

calls_for_service.tsv

```
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
17092534
3   01/2
Berkeley
(37.8574
17091517
2   01/2
Berkeley
(37.8767
17048102
```

## TSV

Tab separated values

calls_for_service.csv

```
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
```

## CSV

Comma separated values

Which is the best?

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

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

```
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
```

calls_for_service.csv — data

```
CASENO,OFFENSE,EVENTDT,EVENTTM,CVLEGEND,CVDOW,InDbDate,Block_Location,BLKADDR,City,Stat
e
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}
}
```

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

- 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?

# 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:**

- Preferences
- Level of education

### Nominal

Categories w/ no specific ordering.

**Examples:**

- Political Affiliation
- CallD number

# 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 | | | | |
| Yelp Rating | | | | |

# Granularity, Scope, and Temporality

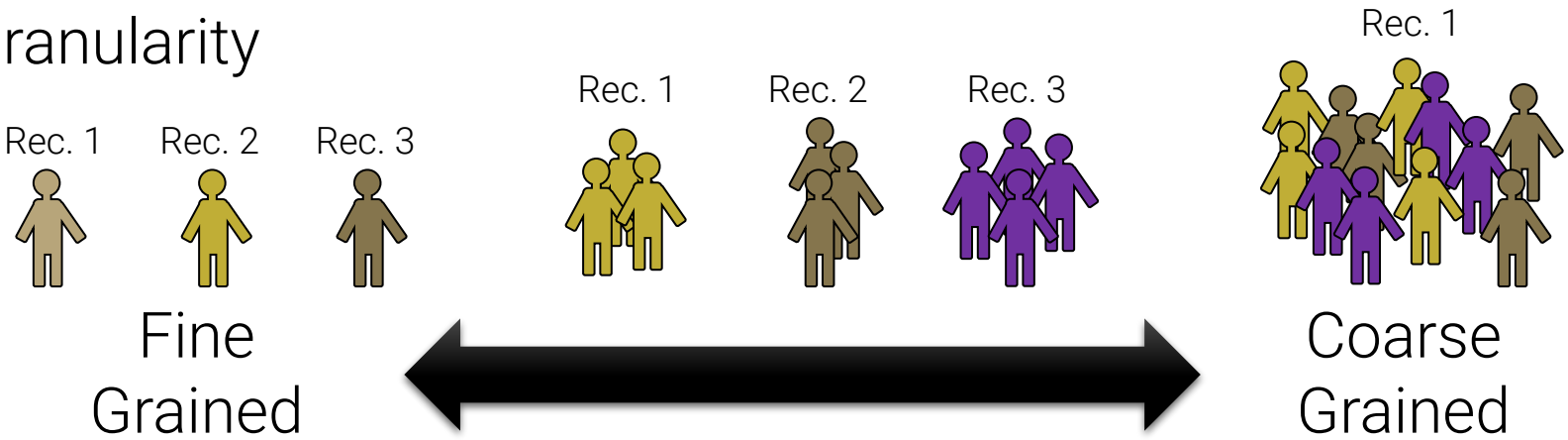
# Key Data Properties to Consider in EDA

- **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"*

# Granularity




- 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"*

# 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 California but I only have Berkeley crime data.*
- Is my data too expansive?
  - **Example:** *I am interested in student grades for DS100 but have student grades for all statistics 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 …

# Revisiting the 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"*

# 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 & daylight savings)
  - Learn to use **datetime** python library
  - Multiple string representation (depends on region): 07/08/09?

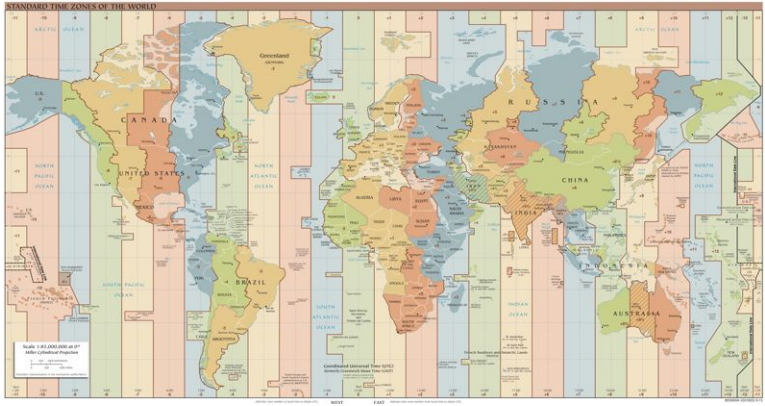
- Are there strange null values?
  - January 1st 1970, January 1st 1900

- 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:
  - San Francisco (UTC-8)
    without daylight savings

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"*

# 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.

- **Prof. Gonzalez 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?)***