Data 100

Lecture 4: Data Cleaning & Exploratory Data Analysis

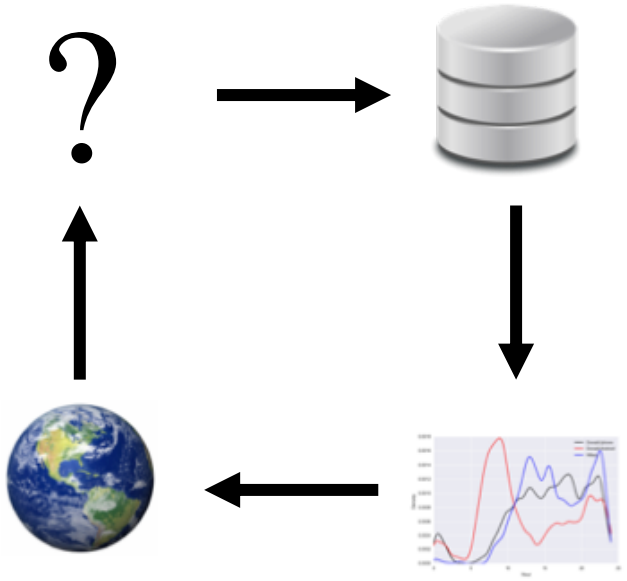
Slides by:

Joseph E. Gonzalez, Deb Nolan, & Joe Hellerstein

jegonzal@berkeley.edu

deborah_nolan@berkeley.edu

hellerstein@berkeley.edu



Congratulations!

Box of Data

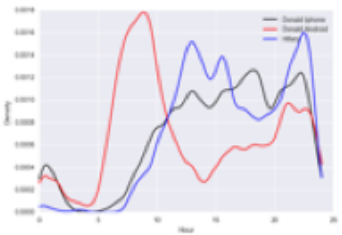
You have **collected** or **been given** a box of data?

What do you do next?

Question & Problem Formulation

?

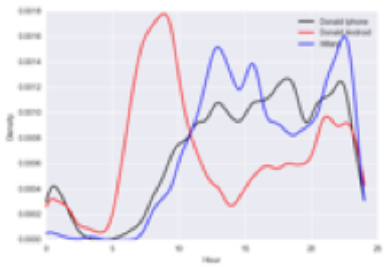
Data Acquisition



Exploratory Data Analysis

Prediction and Inference

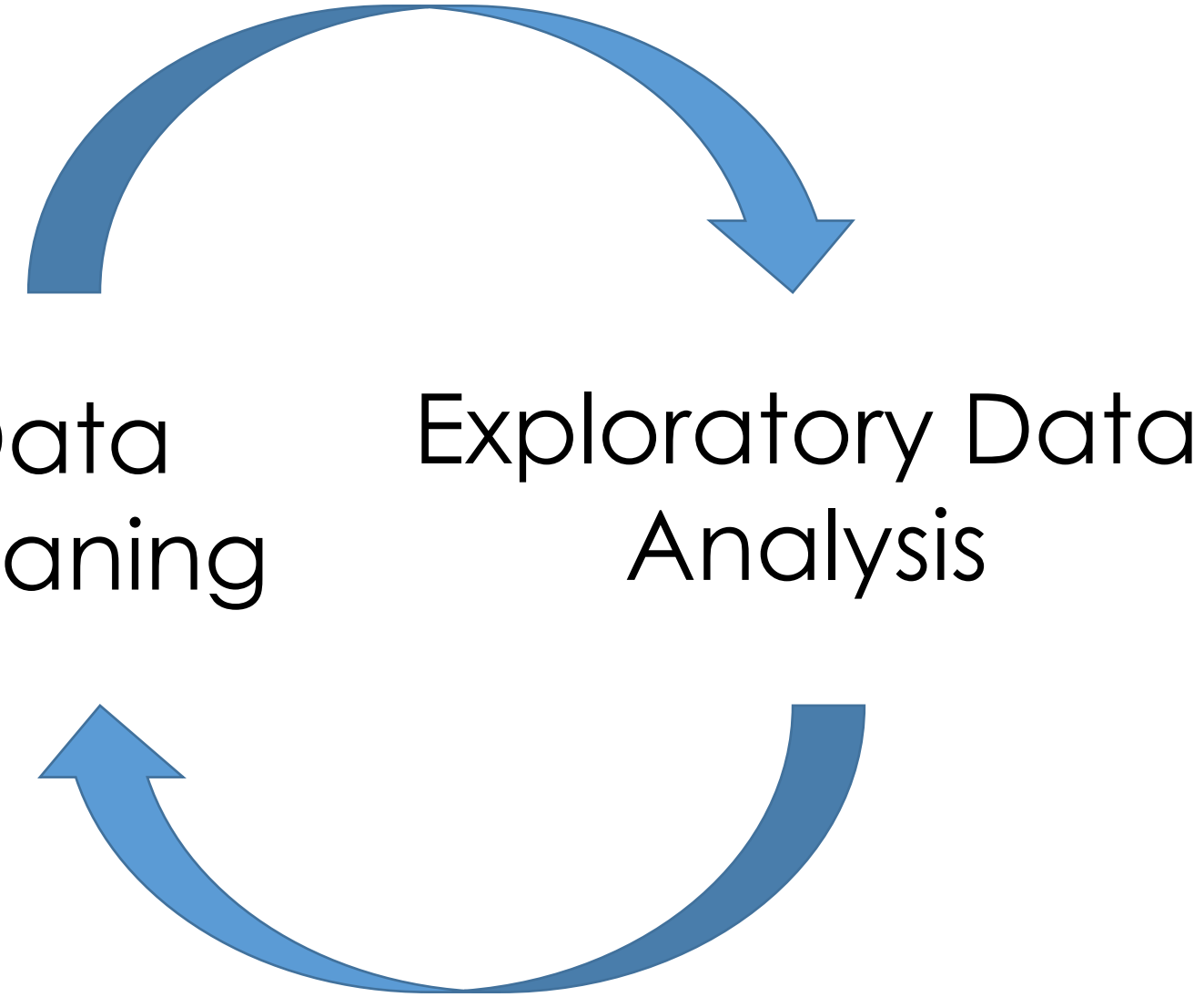
Data Acquisition



Exploratory Data Analysis

Topics For Lecture Today

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

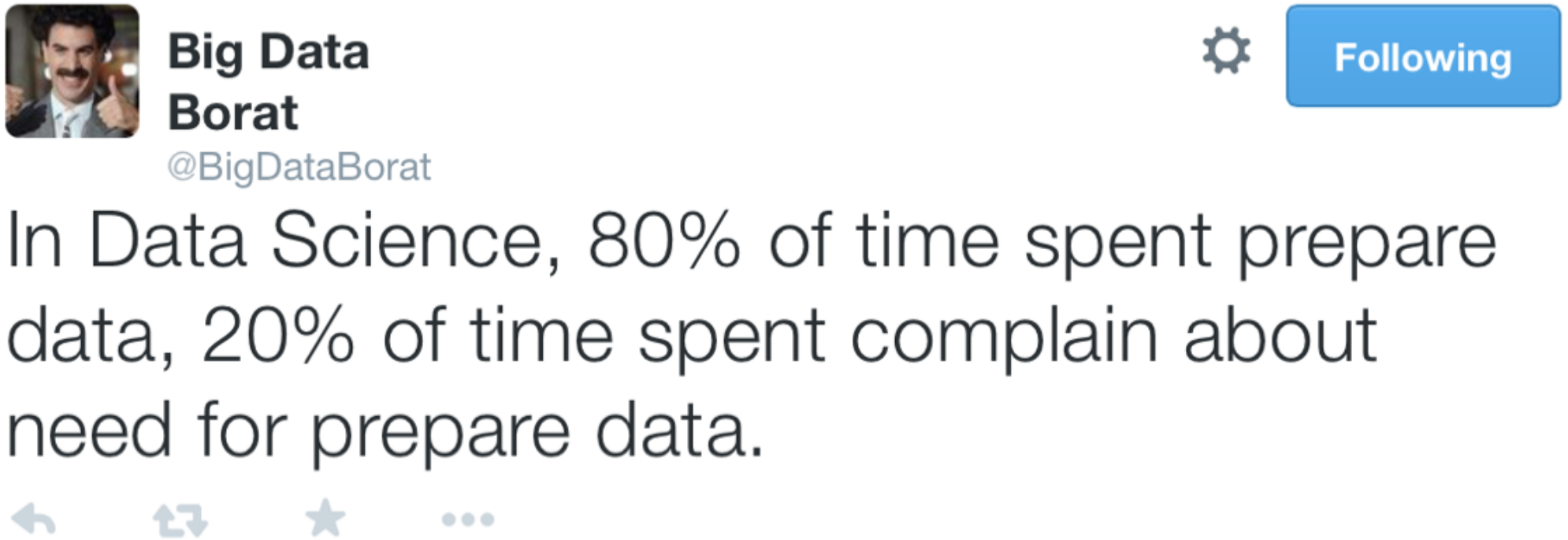


... the infinite loop of data science.

Data Cleaning

- The process of transforming raw data to facilitate subsequent analysis
- Data cleaning often addresses
  - structure / formatting
  - Missing or corrupted values
  - Units
  - Encoding text as numbers
  - …
- Sadly data cleaning is a big part of data science…

Another Statistic …



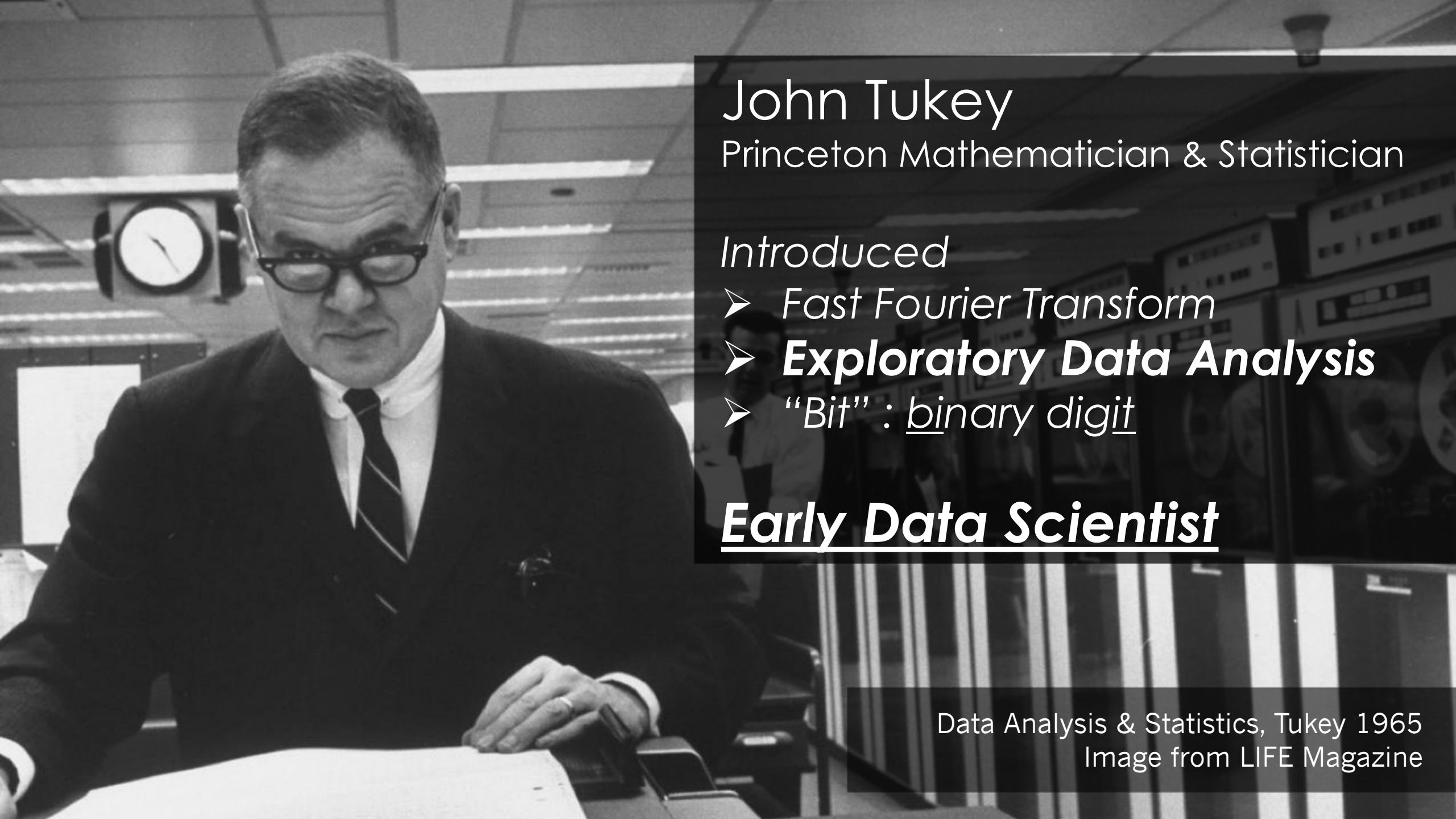
Exploratory Data Analysis (EDA)

"Getting to know the data"

A 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

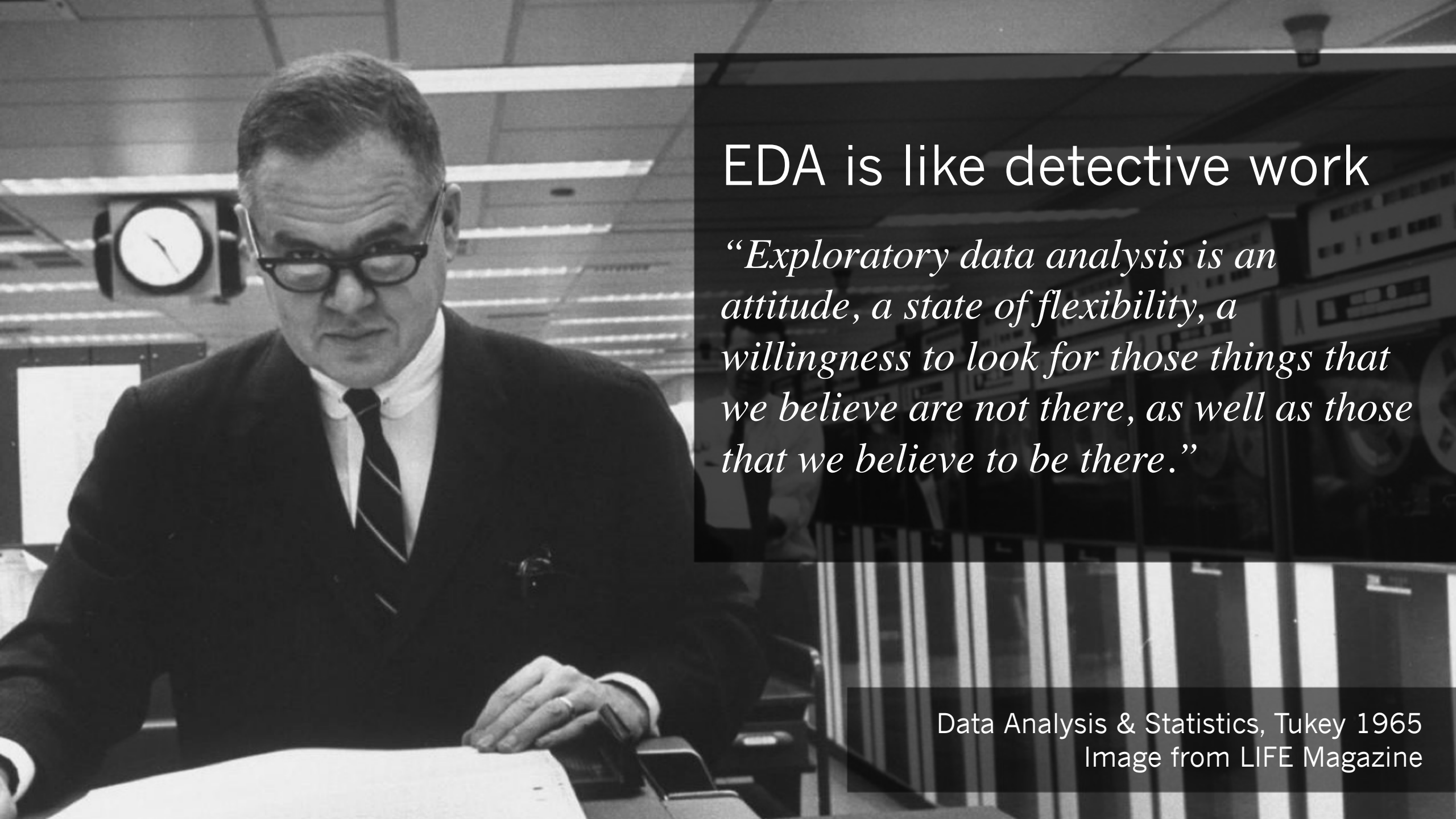
Princeton Mathematician & Statistician

Introduced

- *Fast Fourier Transform*
- ***Exploratory Data Analysis***
- *"Bit" : binary digit*

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

What should we look for?

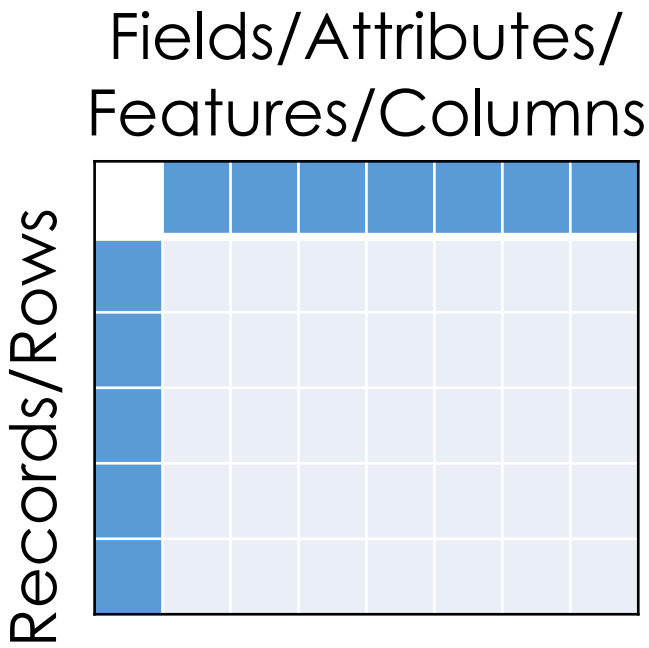
Key Data Properties to Consider in EDA

- **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?)

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?

TSV

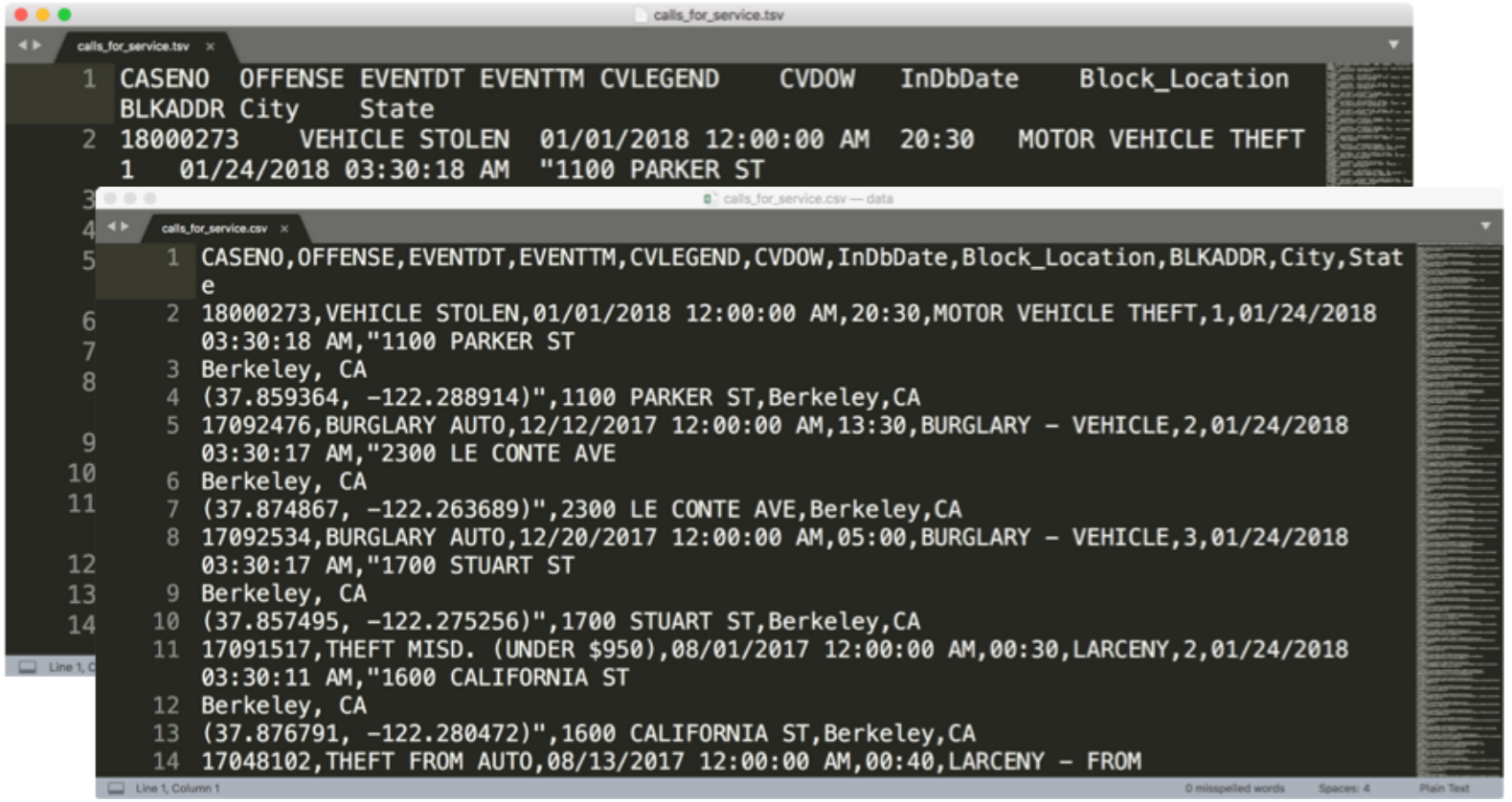
CSV

JSON

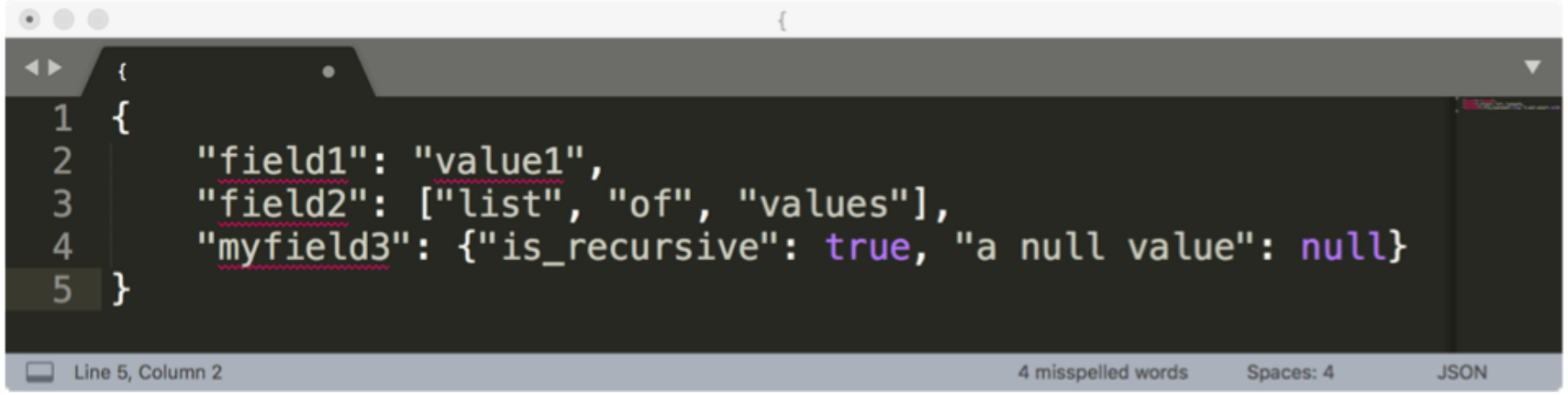
Which is the best?

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
  - …



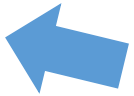
JavaScript Object Notation (JSON)



- Widely used file format for nested data and the web
  - Natural maps to python dictionaries (many tools for loading)
  - Strict formatting "quoting" addresses some issues in CSV/TSV
- Issues
  - Each record can have different fields
  - Nesting means records can contain records → complicated

XML 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

We will study XML later in the class

Real data -- Web log

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)”
```

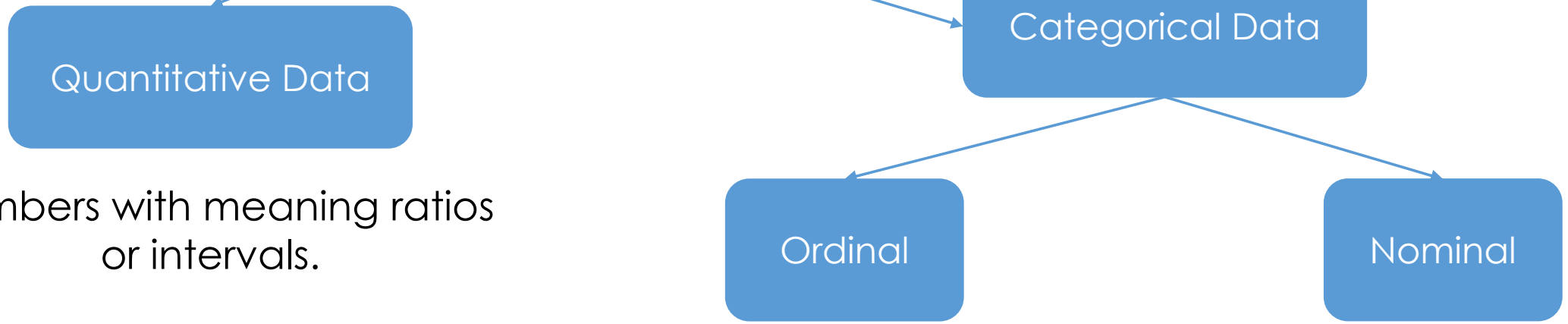
```
169.237.6.168 - - [8/Jan/2014:10:47:58 -0800] "GET
/stat141/Winter04/ HTTP/1.1" 200 2585
"http://anson.ucdavis.edu/courses/" "Mozilla/4.0 (compatible; MSIE
6.0; Windows NT 5.0; .NET CLR 1.1.4322)"
```

Questions to ask about ***Structure***

- Is the data in a standard format or encoding?
  - Tabular data: CSV, TSV, Excel, SQL
  - Nested data: JSON or XML
- Is the data organized in "records"?
  - Yes: How are the records organized?
  - No: (e.g., raw dialogue) Can we define records (e.g., conversations).
- Is the data nested? (Records contained within records…)
  - Yes: Can we reasonably un-nest the data?
- What are the fields in each record?
  - How are they encoded? (e.g., strings, numbers, binary, dates …)
  - Do all the records in the dataset contain the same fields?
    - No: Are values missing or are there different types of data?
  - What is the type of the data?

Kinds of Data

Note that data categorical data can also be numbers and quantitative data may be stored as strings.



Numbers with meaning ratios or intervals.

Examples:
- Price
- Quantity
- Temperature
- Date
- …

Categories with orders but no consistent meaning if magnitudes or intervals

Examples:
- Preferences
- Level of education
- …

Categories with no specific ordering.

Examples:
- Political Affiliation
- CalD number
- …

Field Types:

- Price in dollars of a product?
  - (A) Quantitative, (B) Ordinal, (C) Nominal
- Cal ID Numbers?
  - (A) Quantitative, (B) Ordinal, (C) Nominal
- Star Rating on Yelp?
  - (A) Quantitative, (B) Ordinal, (C) Nominal
- Date an item was sold?
  - (A) Quantitative, (B) Ordinal, (C) Nominal

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.

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 ..

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 as records
- If the data are coarse how was it aggregated?
  - Sampling, averaging, …
- What kinds of aggregation is possible/desirable?
  - From individual people to demographic groups?
  - From individual events to totals across time or regions?
  - Hierarchies (city/county/state, second/minute/hour/days)
- Understanding and manipulating granularity can help reveal patterns.

Granularity and Keys

- The primary key defines what the record represents → Granularity
- What is the granularity of these example tables?
  - Purchases.csv: PK=(OrderNum + ProdID) → Each Item in an order
  - Orders.csv: PK = OrderNum → an order
- How might we adjust the granularity?
  - Aggregation: count, mean, median, var, groupby, pivot …

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

Customers.csv

CustID	Addr
171345	Harmon..
281139	Main ..

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 …

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

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

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?
  - Examples?
    - 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 curb stoning (data falsification):
  - Repeated names, fake looking email addresses, repeated use of uncommon names or fields.

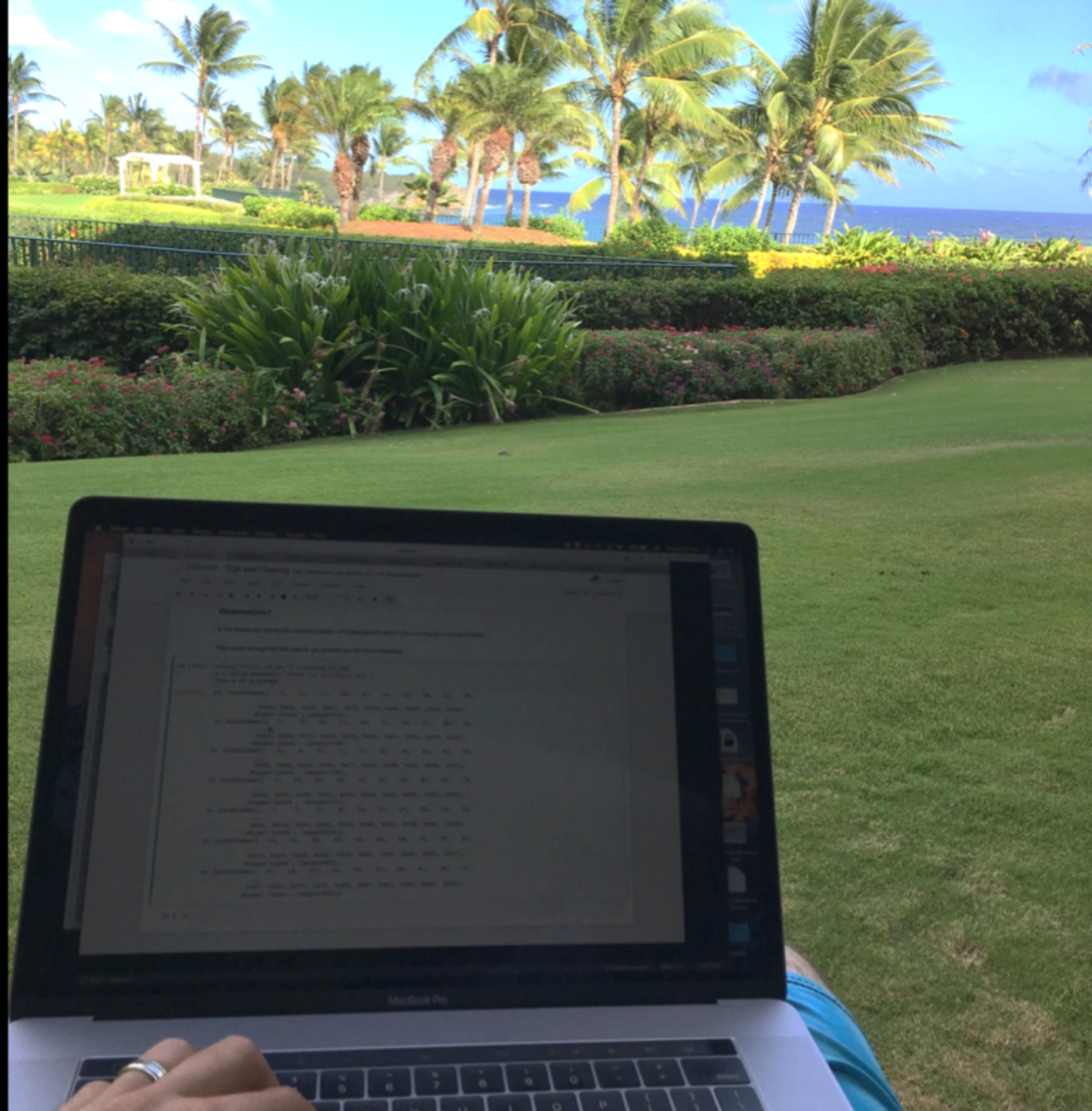
Signs that your data may not be faithful

- Missing Values/Default values: (0, -1, 999, 12345, NaN, Null, 1970, 1900, … others?)
  - **Soln 1:** Drop records with missing values → implications on your sample!
  - **Soln 2:** Impute missing values → Bias your conclusions
- 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
- Truncated data (early excel limits: 65536 Rows, 255 Columns)
  - **Soln:** be aware of consequences in analysis → how did truncation affect sample?
- Others…

Quick Break

Quick Break

Scope:
Do you have a full picture?



Berkeley Police Data Demo

Berkeley Police Public Datasets

- **Question:** For this analysis we will not begin with a detailed question but instead a rough goal of understanding Police activity.
- **Examine Two Data Sets:**
  - Call data
  - Stop data
- Today we will work through the basic process of data loading, some preliminary cleaning, and exploratory data analysis.

Call Data Description

Data pulled from Public Safety Server using data created for Berkeley's Crime View Community page. Displays **incidents reported** for **the last 180 days** along with **time**, **date**, **day of week** and **block level location information.**

The dataset reflects crimes as they have been reported to the BPD based on preliminary information **supplied by the reporting parties**. Preliminary crime classifications may change based on follow-up investigations. **Not all calls for police service are included (e.g. Animal Bite).** The information provided on this site is intended for use by the community to enhance their awareness of crimes occurring in their neighborhoods and the entire City. **The data should not be used for in-depth crime analysis** as the initial information is subject to change.

Stops Data Description

This data was extracted from the Department's Public Safety Server and covers the **data beginning January 26, 2015**. On January 26, 2015 the department began collecting data pursuant to General Order B-4 (issued December 31, 2014). Under that order, **officers were required to provide certain data after making all vehicle detentions** (including bicycles) and pedestrian detentions (up to five persons). This data **set lists stops by police** in the categories of traffic, suspicious vehicle, pedestrian and bicycle stops. Incident number, date and time, location and disposition codes are also listed in this data.

Address **data has been changed from a specific address**, where applicable, and listed as the block where the incident occurred. Disposition codes were entered by officers who made the stop. These codes included the person(s) race, gender, age (range), reason for the stop, enforcement action taken, and whether or not a search was conducted.

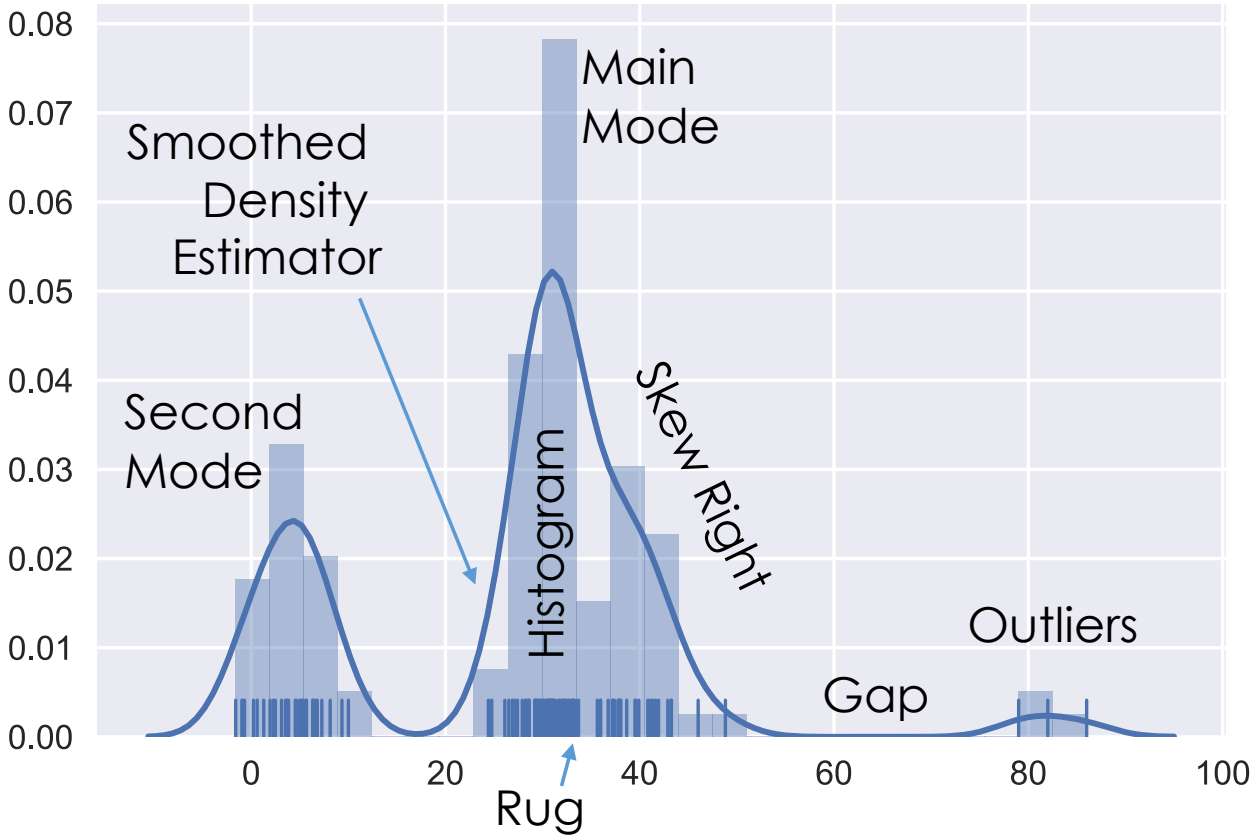
Visualizing Univariate Relationships

- **Quantitative Data**
  - Histograms, Box Plots, Rug Plots, Smoothed Interpolations (KDE – Kernel Density Estimators)
  - Look for spread, shape, modes, outliers, unreasonable values …
- **Nominal & Ordinal Data**
  - Bar plots (sorted by frequency or oridinal dimension)
  - Look for skew, frequent and rare categories, or invalid categories
  - Consider grouping categories and repeating analysis

Histograms, Rug Plots, and KDE Interpolation

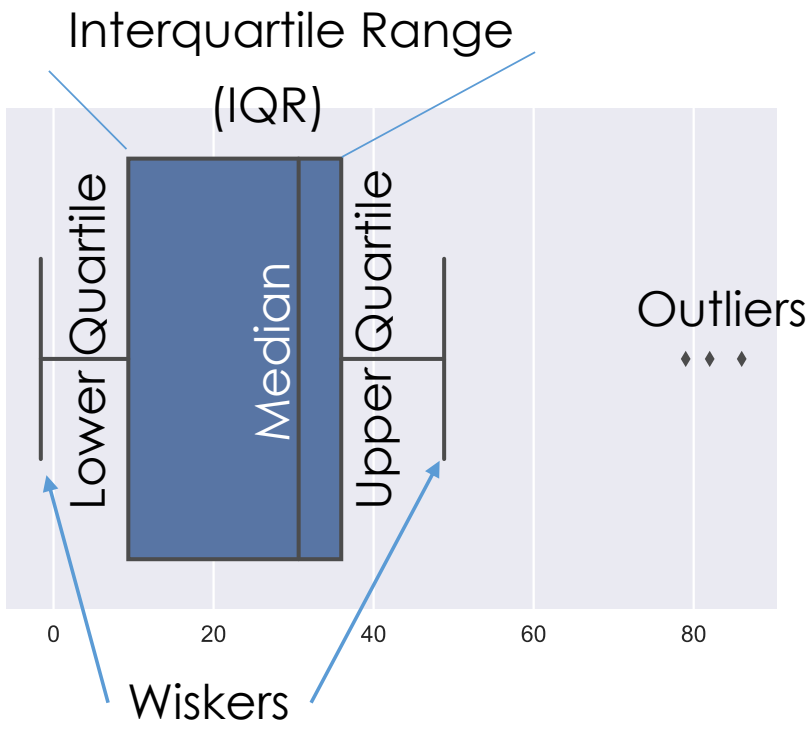
Describes distribution of data – relative prevalence of values

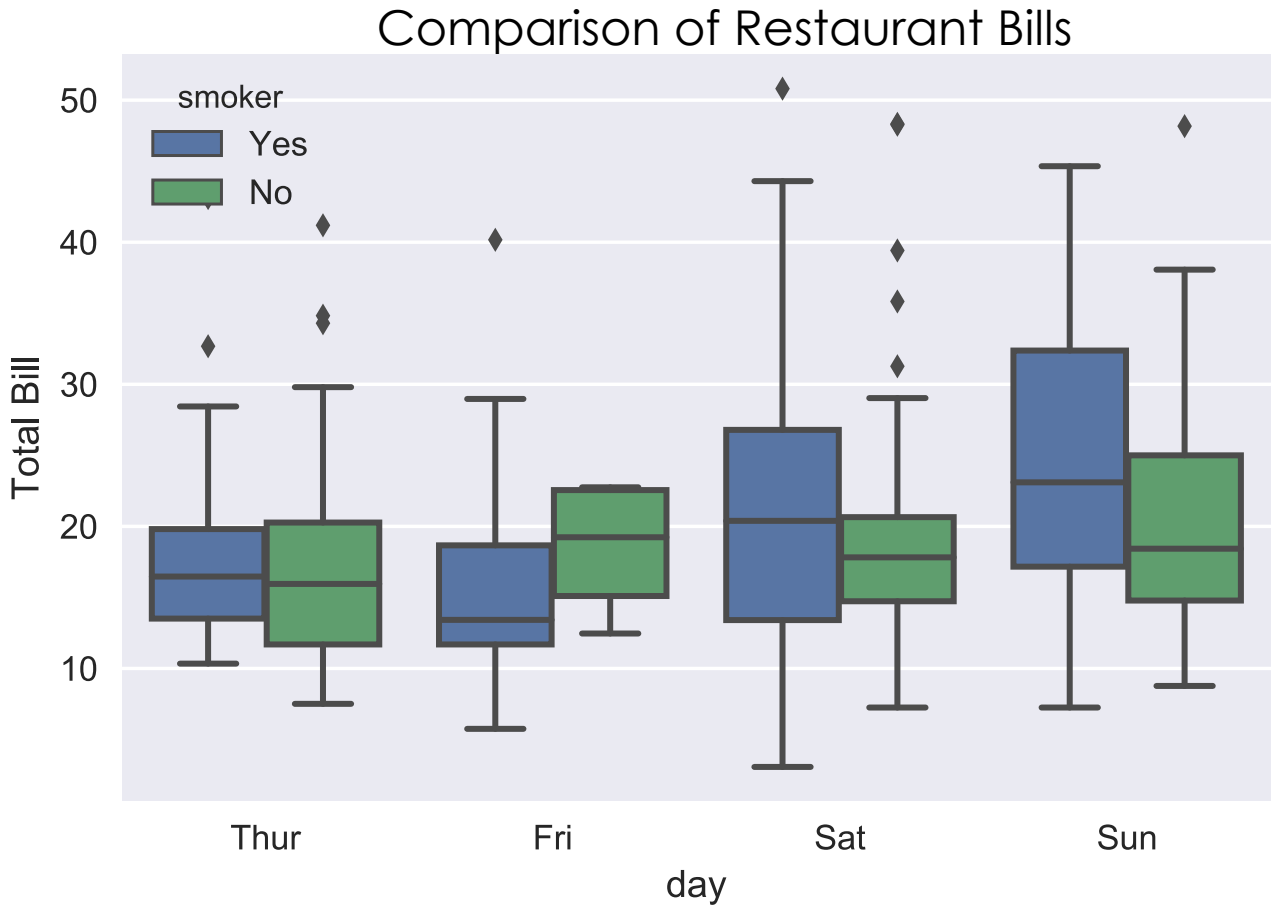
- Histogram
  - relative frequency of values in bins (ranges)
  - Tradeoff of bin sizes
- Rug Plot
  - Shows the actual data locations
- Smoothed density estimator
  - Tradeoff of "bandwidth" parameter (more on this later)



Box Charts

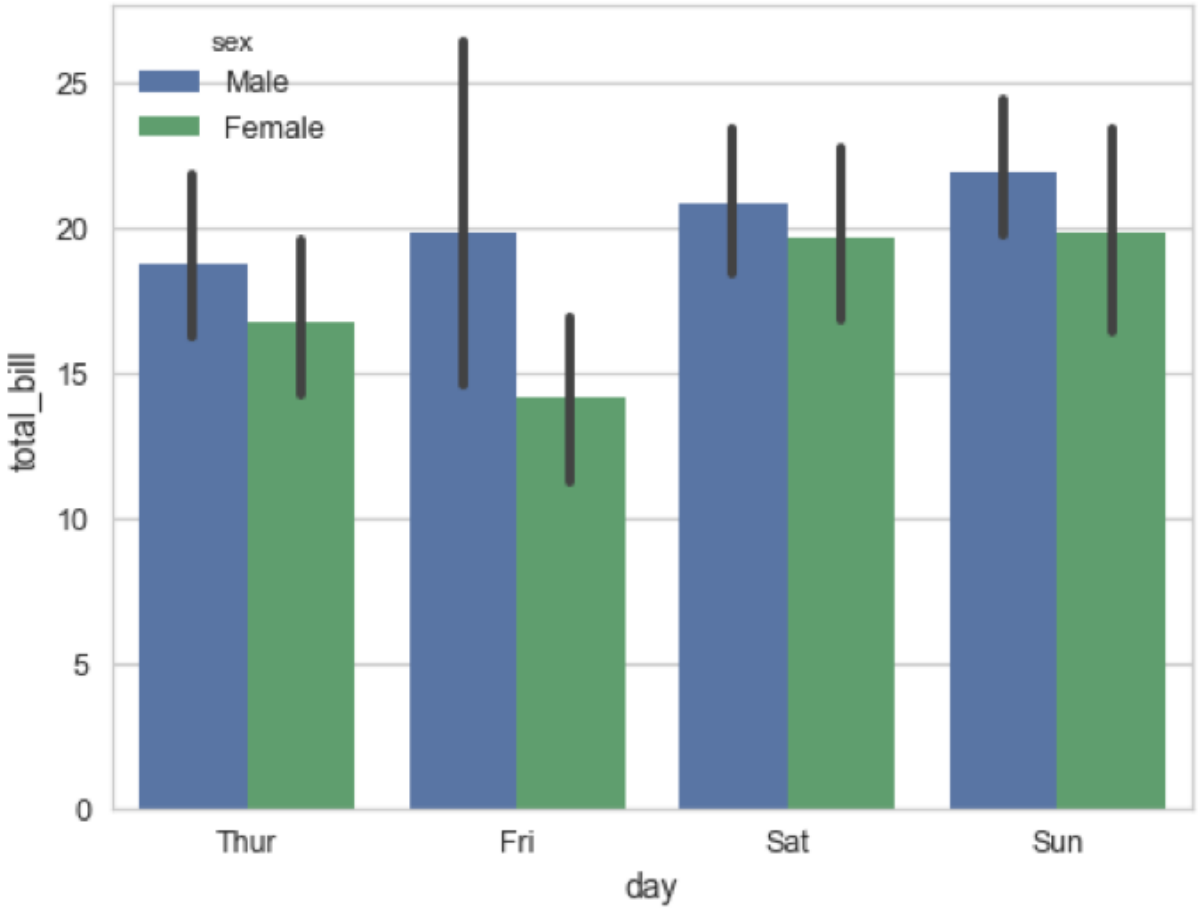
- Useful for summarizing distributions and comparing multiple distributions

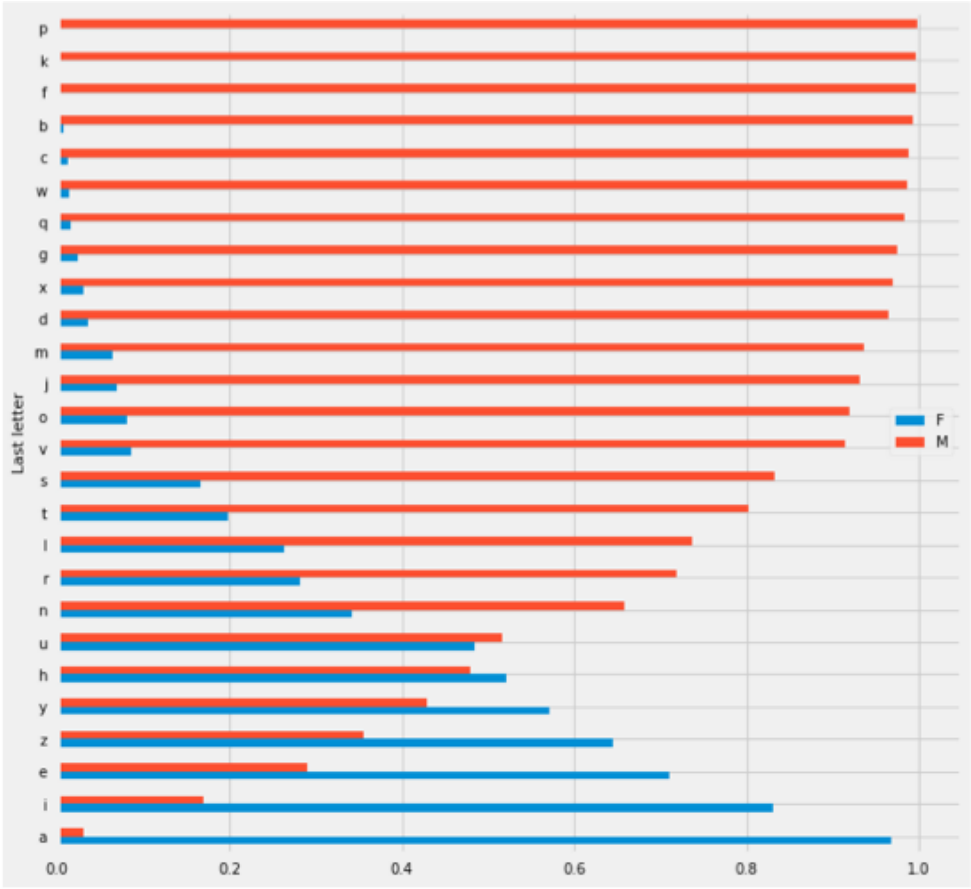




Bar Charts

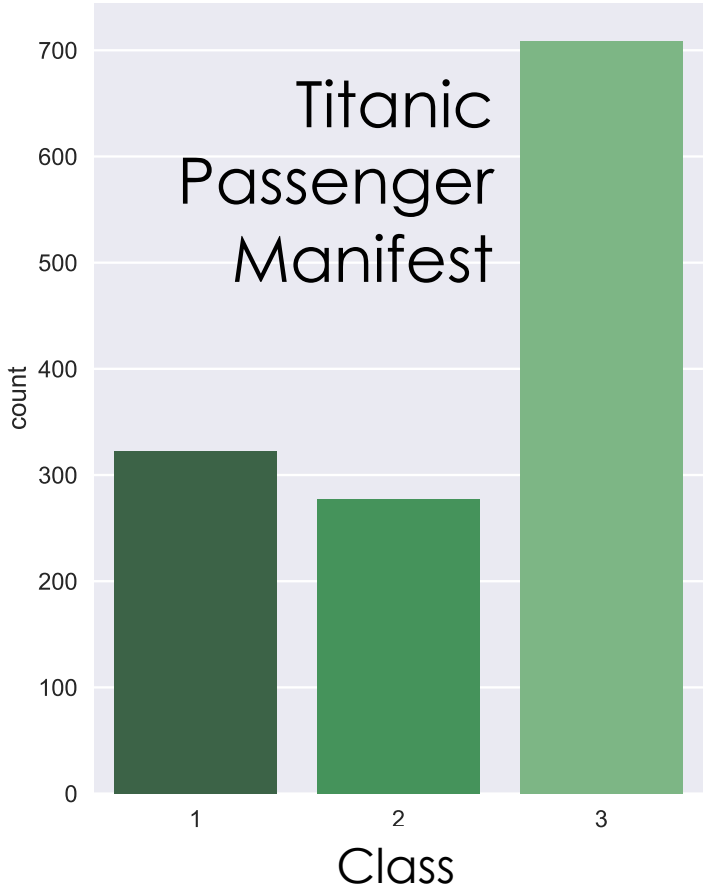
- Used to compare nominal and ordinal data.
  - Consider sorting by category or frequency

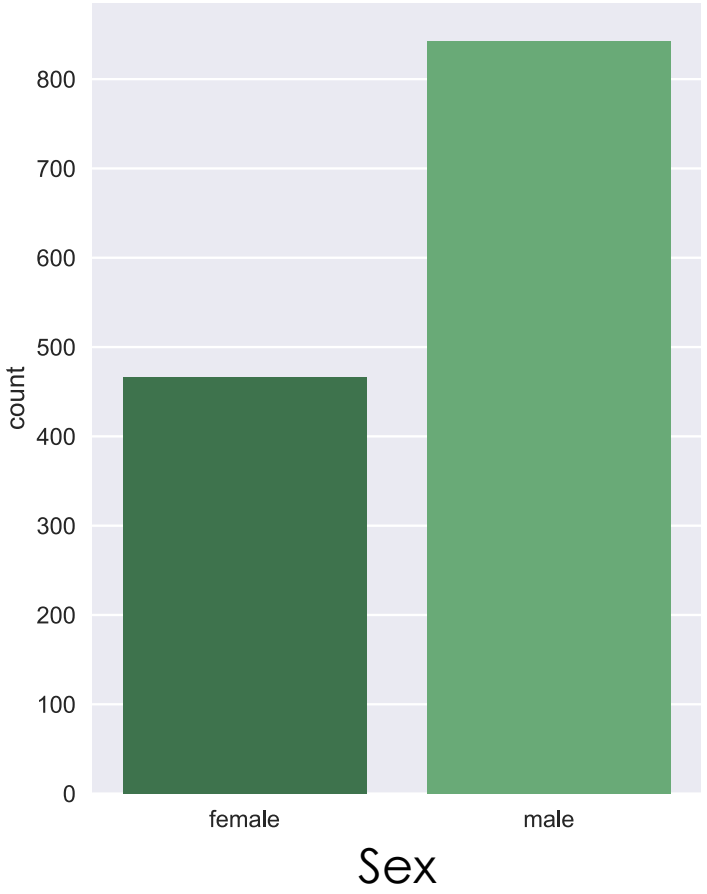


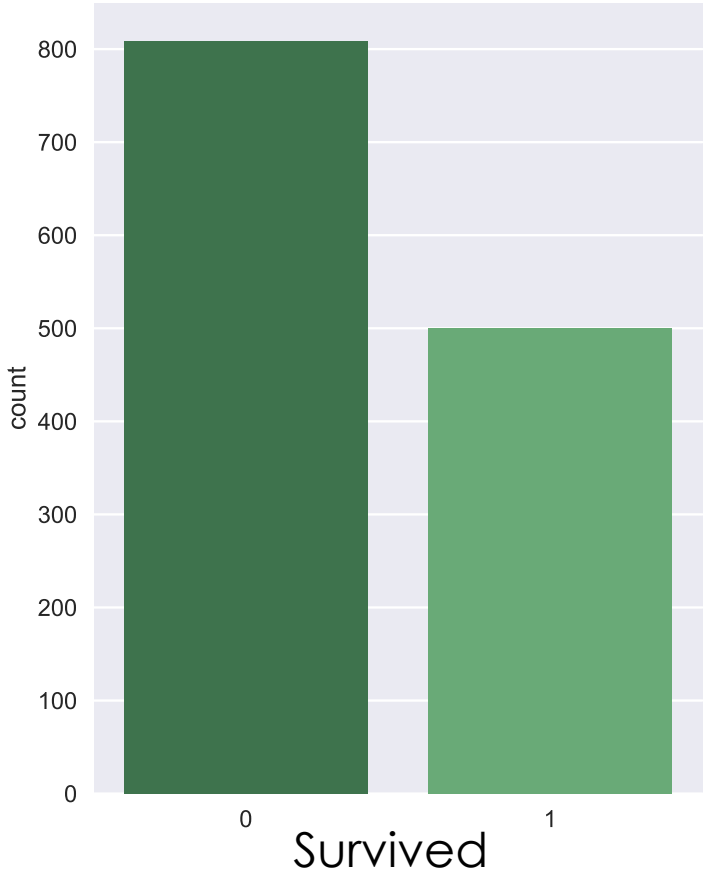


Bar Charts

- Used to compare nominal and ordinal data.
  - Consider sorting by category or frequency

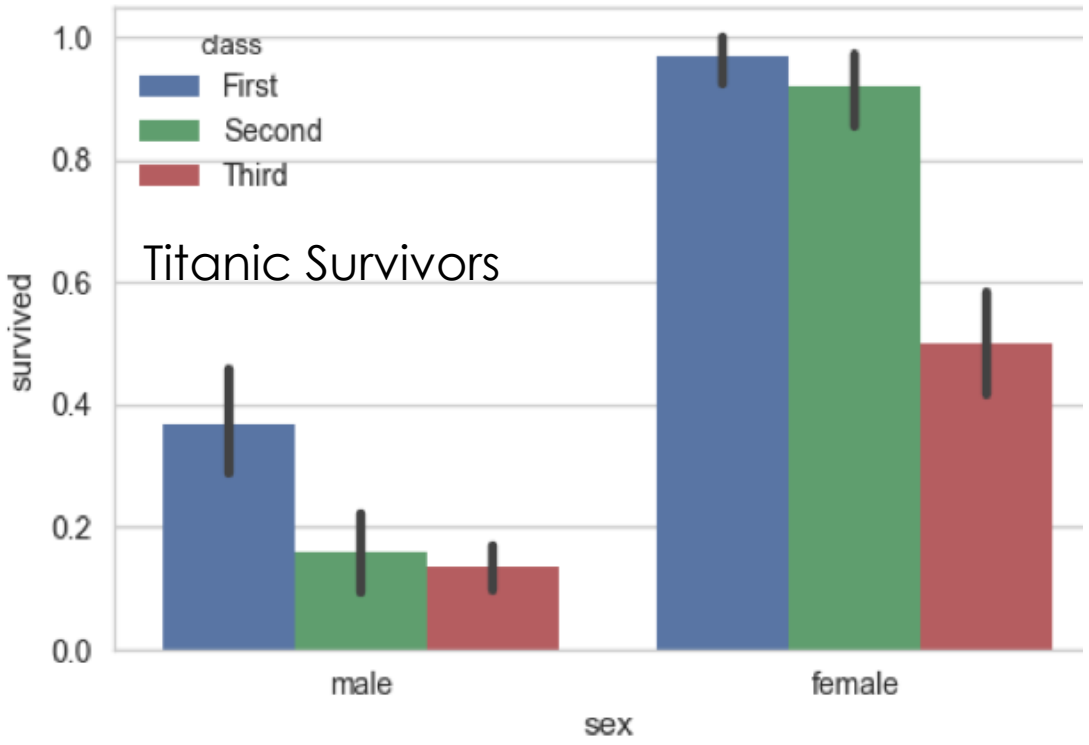


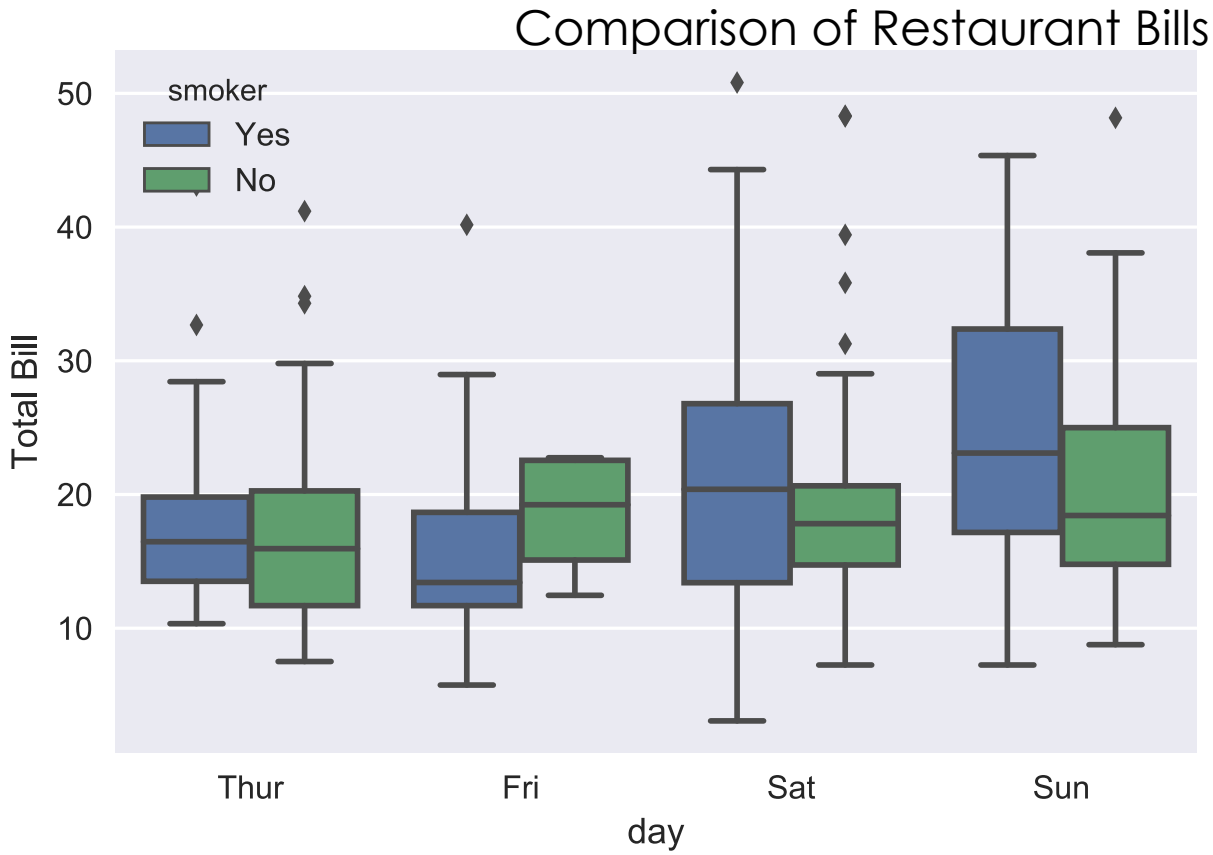




Visualizing Multivariate Relationships

- Conditioning on a range of values (e.g., ages in groups) and construct side by side box-plots or bar charts





Visualizing Multivariate Relationships

- Scatter Plots: try plotting variables against each other
  - Try to linearize relationships (eg., logs, exponents, square-roots)
  - More on transformations when we return to visualizations
- Conditioning on a range of values (e.g., ages in groups) and construct side by side box-plots or bar charts

Caution about EDA

With enough data, if you look hard enough you will find something ***"interesting"***

Important to differentiate **inferential conclusions** about world from **exploratory analysis of data**