Data, Environment and Society

Lecture 6:
Exploratory Data Analysis
and
Data Cleaning

September 17, 2018

Instructor: Duncan Calllaway

GSI: Salma Elmallah

Seminar tomorrow, 88 Dwinelle, 12:10-1:30

Katherine Meckel

UC San Diego

Homepage

“Are Inspections Going to Waste? Using Machine Learning to Improve EPA Inspection Targeting of Hazardous Waste Facilities”

Abstract: Machine learning (ML) algorithms are increasingly used to model and predict economic outcomes. Using 15 years of data and nearly 10,000 variables, we build an ML model to predict the likelihood that manufacturing facilities will violate EPA regulations on hazardous waste. Given that the EPA can inspect a limited number of these facilities per year, we simulate the case in which the EPA's inspection choices are replaced by facilities predicted to be high risk by our model. The results suggest that our model's predictions improve on the EPA's rate of finding violations by 50%. To validate our estimates of the model's efficacy at improving targeting, we run a multi-year field test in which the EPA and the model each choose half of the facilities to be inspected. A field test that incorporates real world implementation challenges is critical for agency adoption. Ours is the first direct test of potential for machine learning to improve on the decision based targeting of government resources in the U.S.

Announcements

- Reminders
  - Late policy
- Lab 2 due yesterday, Lab 3 due next Monday
- HW2 due Thursday
- Next Thursday: Dan Kammen (ERG Professor) will speak about different populations’ access levels to photovoltaics in the US.

Upcoming reading

- Today:
  - Hino *et al*, Pelletier *et al*
  - We will be using material from DS100 Ch4 and 5
- Thursday: DS100 Ch6 textbook (visualization)
- Next Tuesday: Ch 10 of DS100, Ch 2 of ISLR
- Next Thursday: Sunter *et al* (Kammen will present this paper)

Today

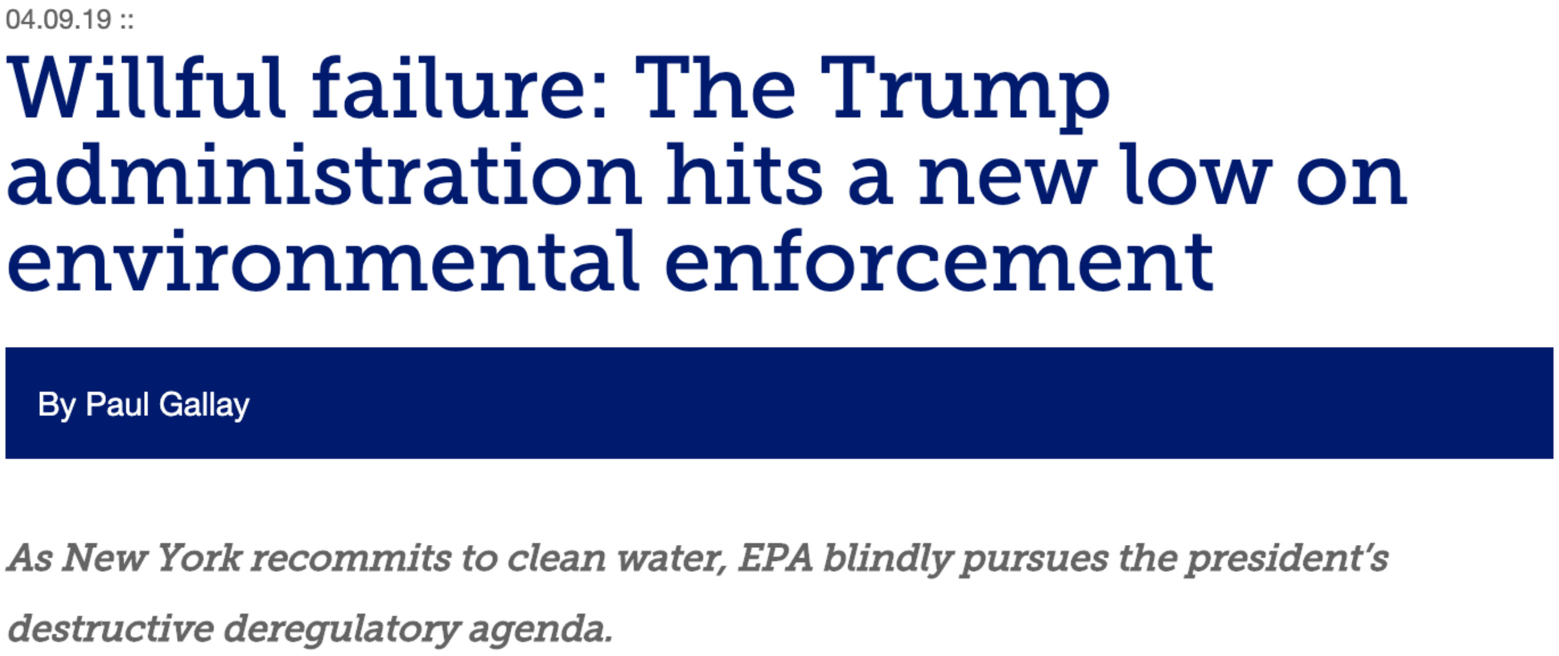
- Understanding how prediction tools get applied to policy problems (Hino et al and Pelletier et al)
  - Objective: that you have more fuel for scoping project questions and understanding the uses and abuses of prediction tools
- Exploratory data analysis and data cleaning – basic guidelines
  - Objective: acquire some guiding principles for working with new data
- Exploratory data analysis and data cleaning – in practice
  - We'll use a local air quality data set
  - Objective: see some specific issues that come up with a new data set.

First, the reading

- Pelletier *et al* and Hino *et al*
- Questions to focus on:
  - What is the prediction question? Are the authors making predictions across space, time, or some other dimension?
  - What is the key policy application?
  - Name two or more factors readers should take into consideration as they interpret and implement the results. These could relate to error propogation, causal inference or other factors.
- Discuss both papers. How are they similar and different?

Hino *et al*

- "Here, we predict the likelihood of a facility failing a water- pollution inspection and propose alternative inspection allocations that would target high-risk facilities."
  - Predictions across facilities
  - Application: Increase identification of polluters for given budget
- They seek at address Athey's concerns by adding constraints in the resource allocation rules.
  - State level inspection budget differences
  - Minimum inspection probability
- Concerns
  - What if different facilities cost more to inspect?
  - "External validity": are the inspected facilities (the training data) representative of all facilities?
  - Strategic response still possible
  - What do the inspectors know that we don't? What might they think of this paper?



04.09.19 ::

Willful failure: The Trump administration hits a new low on environmental enforcement

By Paul Gallay

As New York recommits to clean water, EPA blindly pursues the president's destructive deregulatory agenda.

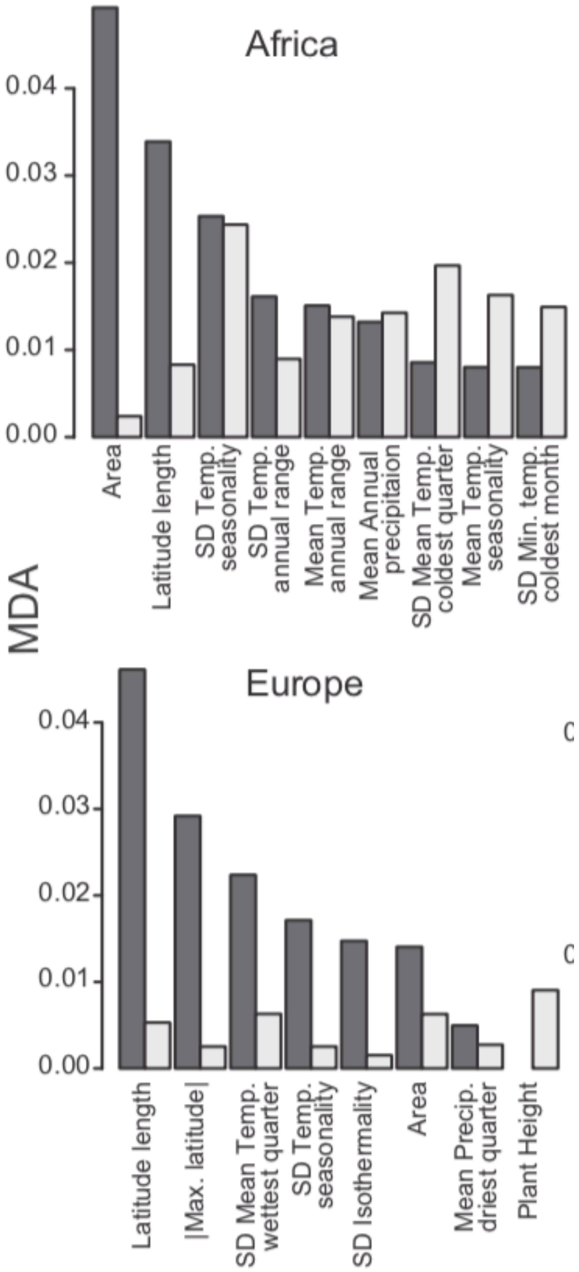
- "The total number of compliance inspections by EPA has **dropped by one-third**, since 2015.
- Total enforcement cases resolved — **down by 50%**, since 2017.
- Compliance spending by polluters has also **plunged by half**, since 2015.
- Significant violations of the Clean Water Act at major facilities and high priority violations of the Clean Air Act have collectively **spiked by 19%**, in the past four years. And,
- The total number of enforcement office staff at EPA **fell by 16%**, in the first 18 months of Trump's presidency."

Pelletier *et al*

- **Objective**: Predict conservation status of land plant species based on publicly available data.
  - "Our results indicate that a large number of unassessed species have a high probability of being at risk, and these probabilities can be used to establish assessment prioritization."
  - Predictions across space
  - Application: create guide for identifying species of concern; preserve biodiversity
- They are careful to talk about variables' importance for prediction without invoking the language of inference.
  - "Although we did identify trends in the variables that contribute the most to at-risk classifiers across continents, there is no one single global variable that predicts conservation status."
  - They use the language "explanatory variable" which makes me a little uncomfortable.

Pelletier *et al*, ctd

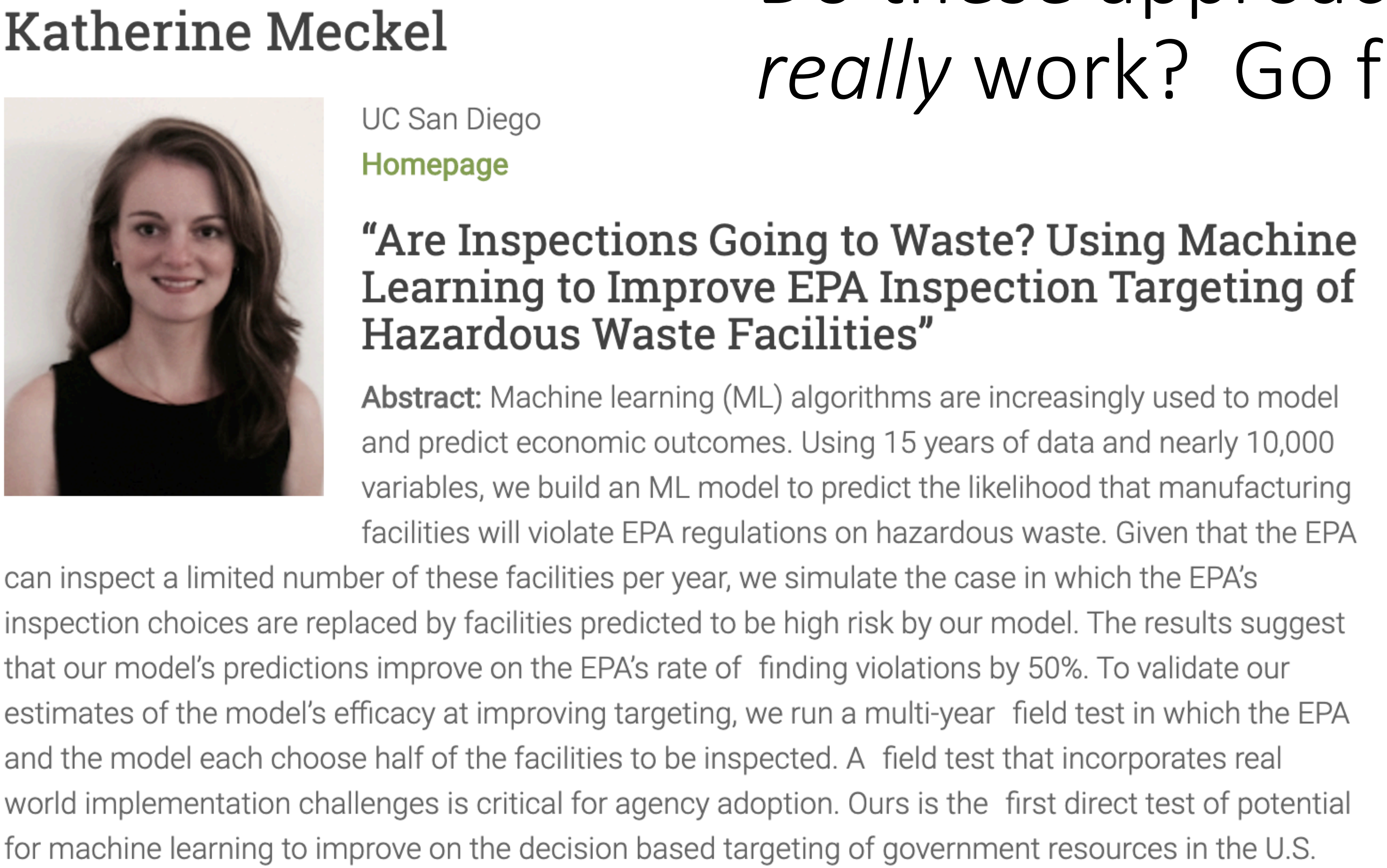
- Some points of caution
  - Be careful interpreting the error rate
  - “Mean decrease in accuracy” metrics need to be interpreted with caution
  - We will get to both issues later in the semester.
- Strikingly different non-LC probabilities for different training data sets.



Comparing the papers

- Hino *et al* have more obvious justice and fairness implications. Bias, it could be argued, matters more here.
- Both have a common objective of predicting something we have not yet measured

Do these approaches *really* work? Go find out!



Katherine Meckel

UC San Diego

Homepage

“Are Inspections Going to Waste? Using Machine Learning to Improve EPA Inspection Targeting of Hazardous Waste Facilities”

Abstract: Machine learning (ML) algorithms are increasingly used to model and predict economic outcomes. Using 15 years of data and nearly 10,000 variables, we build an ML model to predict the likelihood that manufacturing facilities will violate EPA regulations on hazardous waste. Given that the EPA can inspect a limited number of these facilities per year, we simulate the case in which the EPA’s inspection choices are replaced by facilities predicted to be high risk by our model. The results suggest that our model’s predictions improve on the EPA’s rate of finding violations by 50%. To validate our estimates of the model’s efficacy at improving targeting, we run a multi-year field test in which the EPA and the model each choose half of the facilities to be inspected. A field test that incorporates real world implementation challenges is critical for agency adoption. Ours is the first direct test of potential for machine learning to improve on the decision based targeting of government resources in the U.S.

Questions for data cleaning

- Do you see missing values?
- Are there cells where missing values were obviously filled in?
- Are there cells where values are clearly wrong?
- Are there values where two entries could mean the same thing? Often human-entered values, e.g.:
  - canine and k9;
  - recommend and recommend,
  - Zürich and Zurich etc

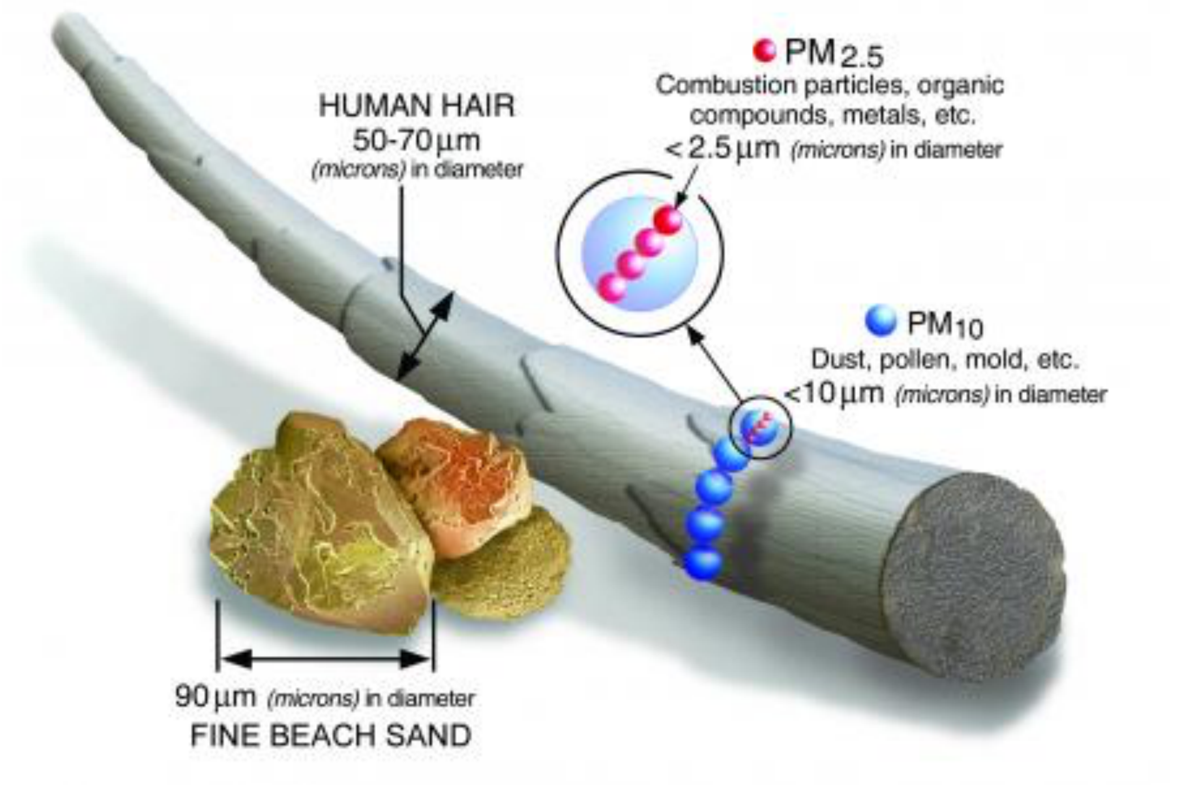
Data merging

- How many rows and columns do you have when you start?
- How many do you have after the merge?
- What’s missing? Is it acceptable to you if you’ve lost some data?
    - We’ll return to this when we talk about faithfulness and scope.
- I wrote a script for the class to use in the upcoming homework that helps decipher what data gets lost.

As we proceed, let's play with the PurpleAir data set.

First, a little more on PM. Sources:

- Many sizes and shapes and can be made up of hundreds of different chemicals.
- Some directly from a source: construction sites, unpaved roads, fields, smokestacks or fires.
- Most form in the atmosphere by complex reactions of chemicals such as sulfur dioxide and nitrogen oxides



PM2.5
Combustion particles, organic compounds, metals, etc.
<2.5µm (microns) in diameter
HUMAN HAIR
50-70µm
(microns) in diameter
PM10
Dust, pollen, mold, etc.
<10µm (microns) in diameter
90µm (microns) in diameter
FINE BEACH SAND

More on PM2.5…

- (text adapted from epa.gov) Health effects:
  - premature death in people with heart or lung disease
  - nonfatal heart attacks
  - irregular heartbeat
  - aggravated asthma
  - decreased lung function
  - increased respiratory symptoms, such as irritation of the airways
- Environmental effects
  - makes lakes and streams acidic
  - changes the nutrient balance in coastal waters and large river basins
  - depletes soil nutrients
  - damages sensitive forests and farm crops
  - affects the diversity of ecosystems
  - contributes to acid rain

Exploratory Data Analysis (EDA)

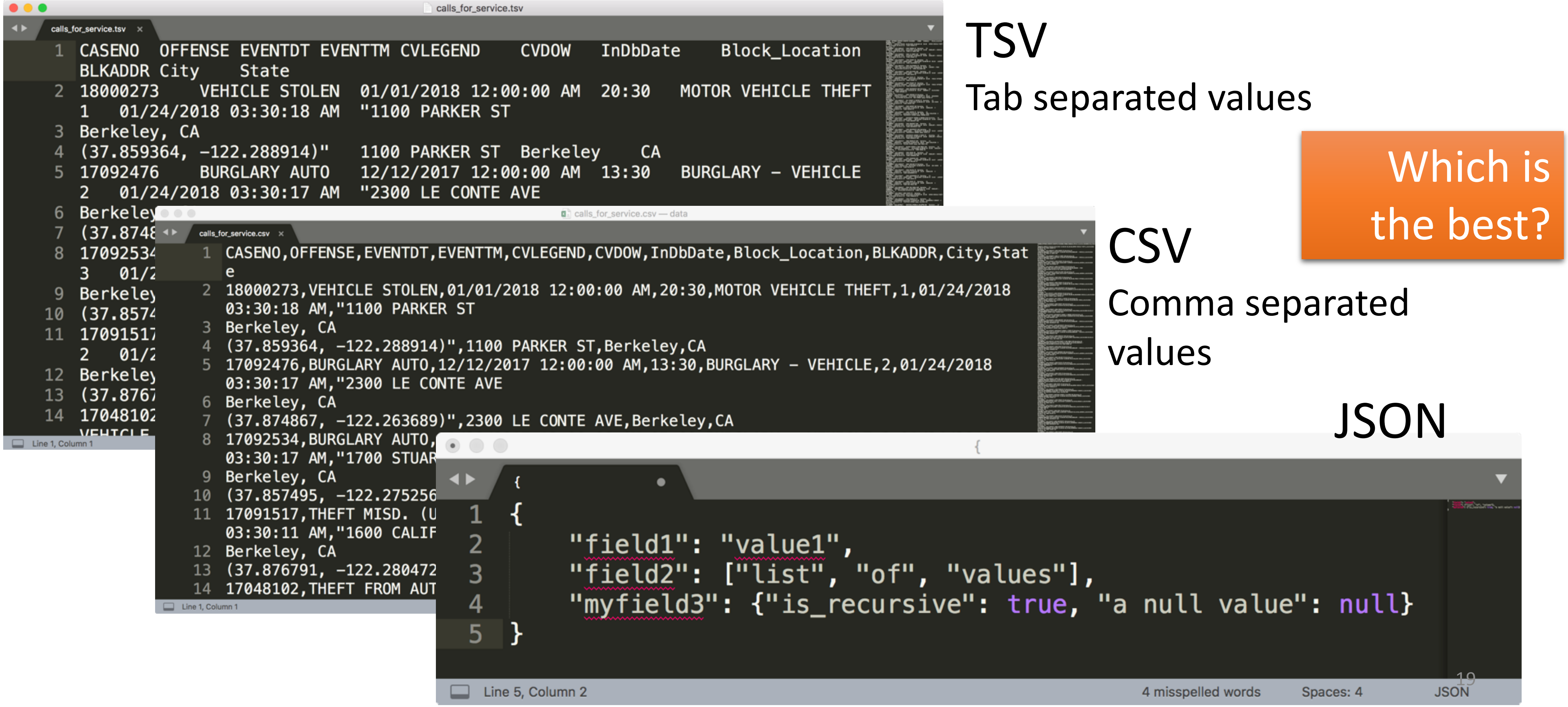
One can approach EDA by asking questions about the data:

- Structure
- Granularity
- Scope
- Temporality
- Faithfulness

Structure – how are the data stored?

- Are the data in a standard format or encoding?
    - Tabular data: CSV, TSV, Excel, SQL
    - Nested data: JSON, XML
- Are the data organized in records (e.g. rows)? If not, can we define records by parsing the data?
- Are the data nested? If so, can we reasonably un-nest the data?
- Do the data reference other data? If so, can we join the data?
- What are the fields (e.g. columns) in each record? What is the type of each column?

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
(37.8574
1709151
2    01/2
Berkeley
(37.8767
1704810
VEHICLE
```

TSV

Tab separated values

```
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,
03:30:17 AM,"1700 STUAR
Berkeley, CA
(37.857495, -122.275256
17091517,THEFT MISD. (U
03:30:11 AM,"1600 CALIF
Berkeley, CA
(37.876791, -122.280472
17048102,THEFT FROM AUT
```

CSV

Comma separated values

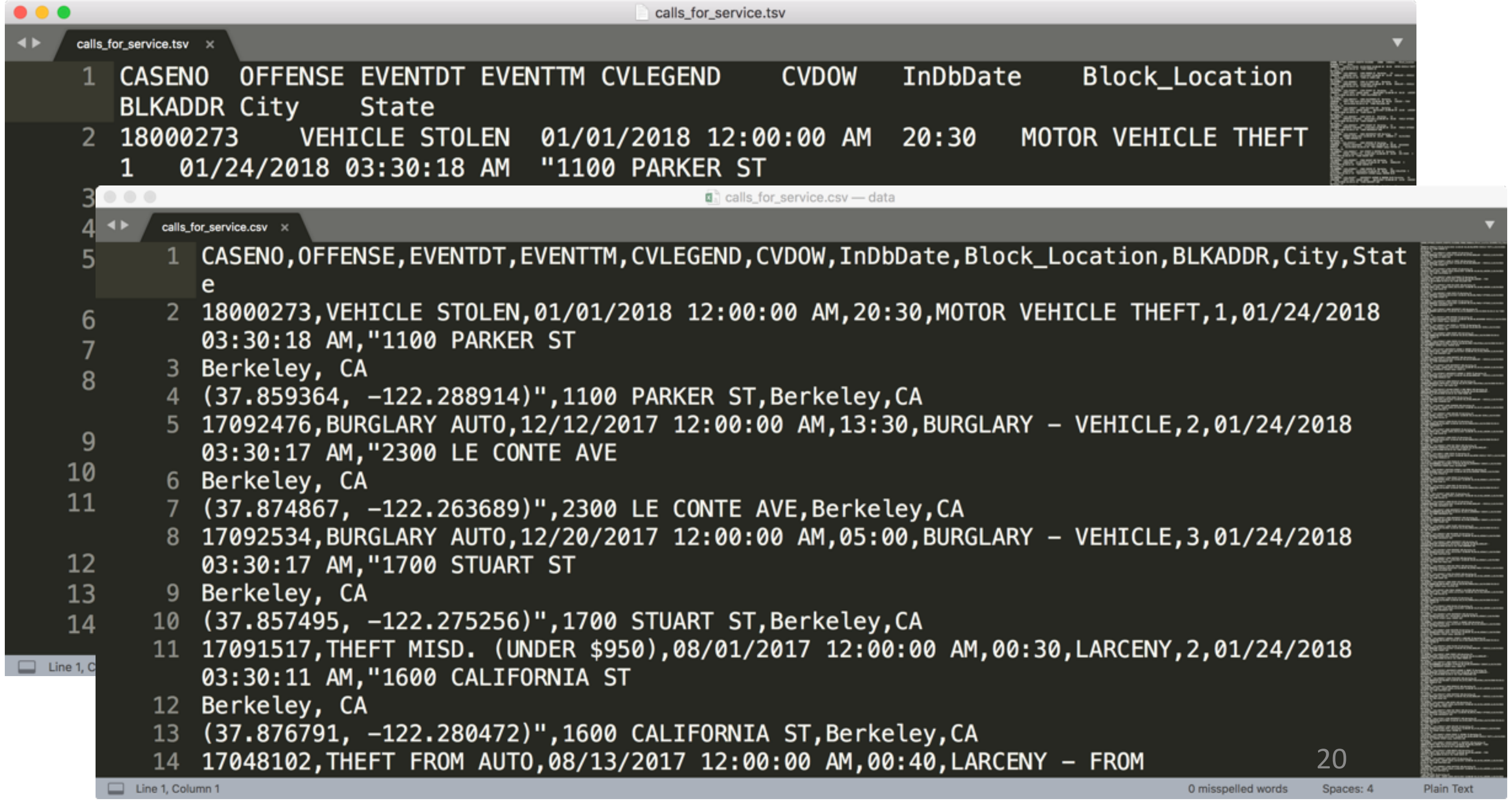
Which is the best?

JSON

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

Comma and Tab Separated Values Files

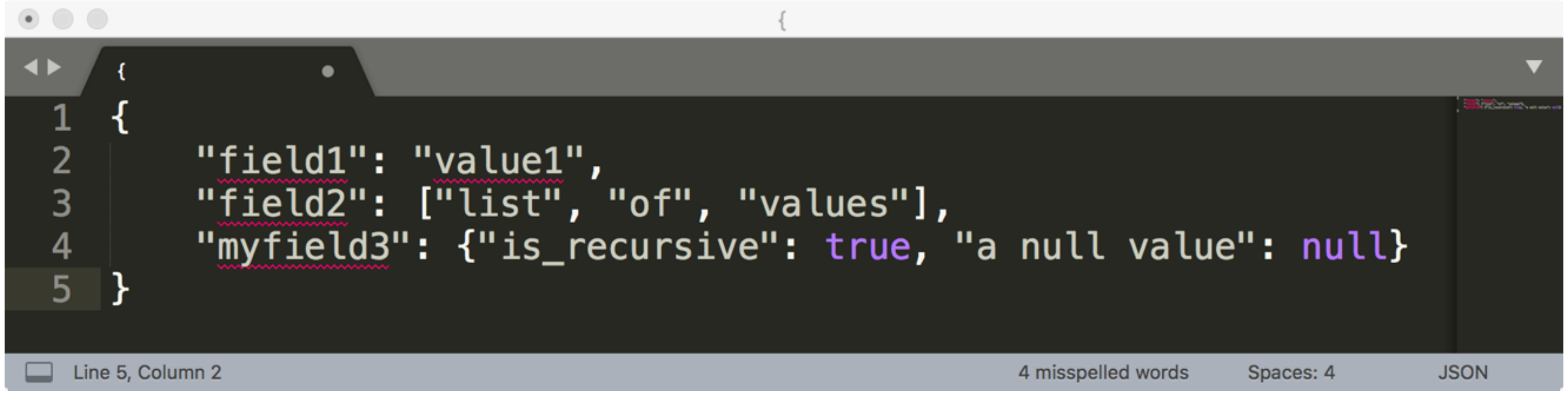
- Tabular data where
  - records are delimited by a *newline*: “\n”, “\r\n”
  - Fields are delimited by ‘,’ (comma) or ‘\t’ (tab)
- Very Common!
- Main issue?
  - Some things that are part of the record get interpreted as delimiters: Commas, tabs, quotation marks



```
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)

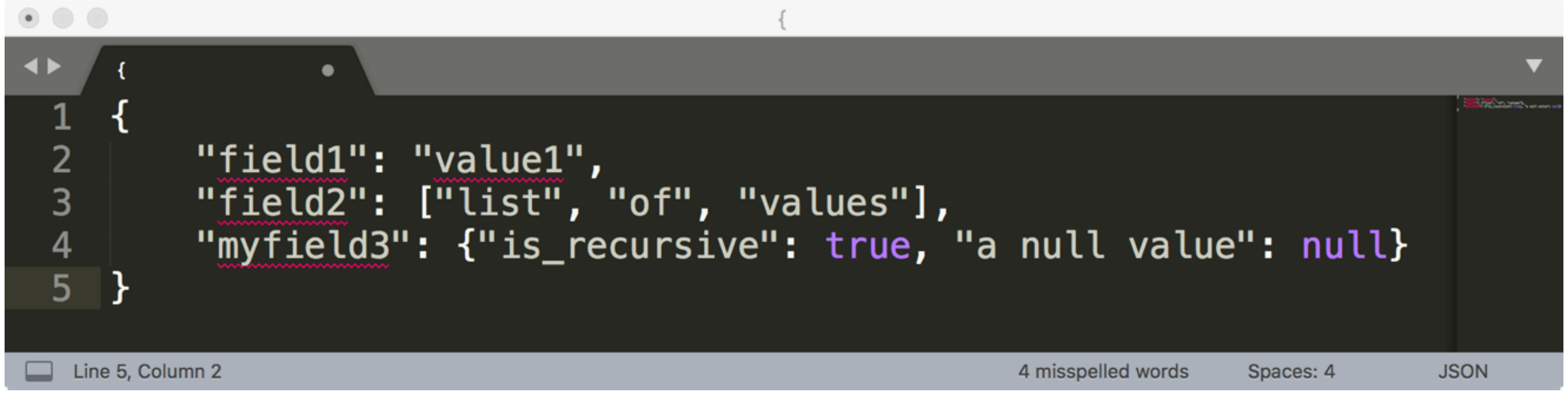


```
{
    "field1": "value1",
    "field2": ["list", "of", "values"],
    "myfield3": {"is_recursive": true, "a null value": null}
}
```

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

- Squiggly brackets act as 'containers’
- Square brackets holds arrays
- Names and values are separated by a colon.
- Array elements are separated by commas

JavaScript Object Notation (JSON)



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

- Widely used file format for nested data
  - 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
- Side note: look at the “raw” form of the .ipynb files! (They’re JSON.)

Granularity – how are the data aggregated?

- What does each record represent?
  - a purchase, a person, a group of users?
  - A home, a city, a country?
  - A minute, an hour, a year?
- Do all records capture granularity at the same level?
  - Data sometimes includes summaries as records
- If the data are coarse how was it aggregated?
  - Sampling, averaging, summing…
- What additional 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)

Scope – how much time, how many people, what spatial area?

- Does the data cover the topic of interest?
  - Subset of a population?
  - Specific range in time
  - Specific location
- How complete are the data?
  - Are countries missing?
  - Are periods of time missing?

Lecture 6 stopped here (and covered up to Scope in the ipynb as well)

Temporality: How is time represented in the data?

- What is the meaning of the date and time fields in the dataset?
  - Beware of time zones, daylight savings!
- What representation do the date and time fields have in the data?
- Are there funky timestamps that might represent null values or cloud your interpretation?

Faithfulness: are the data trustworthy?

Faithful data lack:

- Unrealistic or incorrect values
- Violations of obvious dependencies
  - E.g. age and birthday for individuals don’t match
  - E.g. sorting by record ID gives different result than sorting by time in PurpleAir data
- Hand-entered data
  - Spelling errors, etc.
- Clear signs of falsified data
  - E.g. repeated names, fake looking email addresses, or repeated use of uncommon names or fields.

Summary: How do you “do” EDA?

- Examine data and meta-data:
  - 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 (next time!)
  - Test your assumptions about the data, for example
    - “The range should be…”
    - “Sudden changes should not occur...”
    - Identify anomalies and either change update your assumptions or modify the data.
- ***Record everything you do! (why?)***