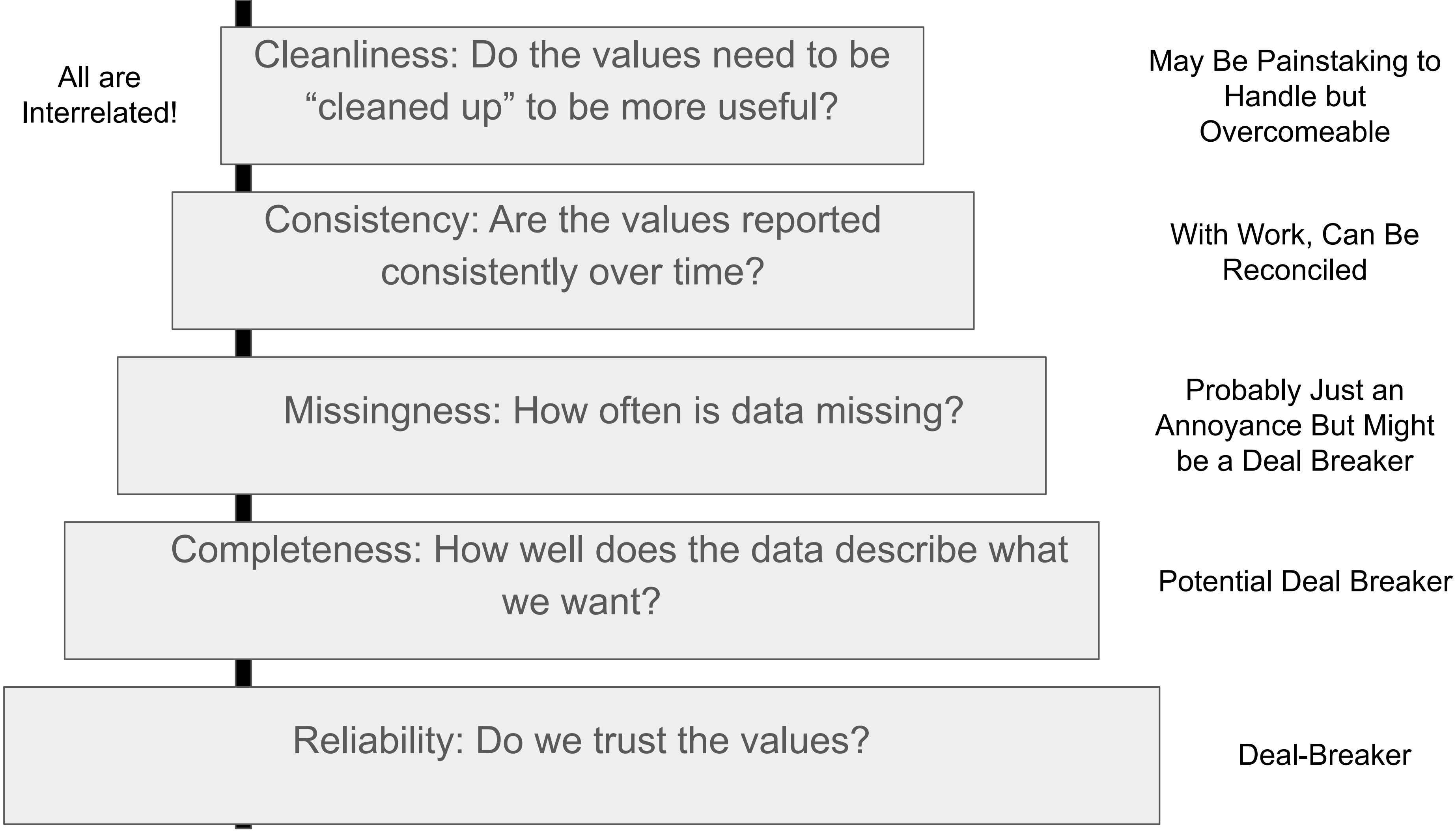
Data Analysis for Business Analytics

CIS621

Week 4 - Data Quality Issues

Lecture Goals

1. Understand the reasons for data quality issues along with the issues they cause and how to detect them.
2. Understand the causes of biased models and possible remedies.



All are Interrelated!

Cleanliness: Do the values need to be "cleaned up" to be more useful?

May Be Painstaking to Handle but Overcomeable

Consistency: Are the values reported consistently over time?

With Work, Can Be Reconciled

Missingness: How often is data missing?

Probably Just an Annoyance But Might be a Deal Breaker

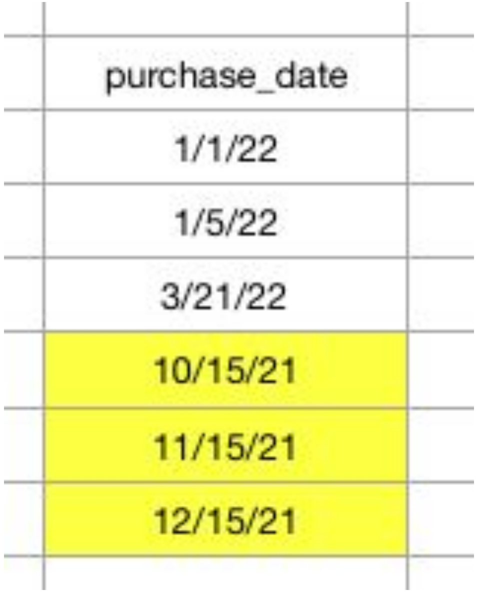
Completeness: How well does the data describe what we want?

Potential Deal Breaker

Reliability: Do we trust the values?

Deal-Breaker

Reliability - Example



purchase_date
1/1/22
1/5/22
3/21/22
10/15/21
11/15/21
12/15/21

Purchase Dates before 2022 were manually tracked. The data got corrupted, and all sales for a month were assigned to the 15th as a default. Therefore, a sale on 10/1/21 would now be coded as 10/15/21. We might not be told this - we might have to discover it!

Reliability - Why Might Data Not be Reliable?

- The biggest drivers of unreliable data are inconsistent collection and system bugs.
- These issues can be difficult to track down. They often require substantial research and talking with various areas of the business.

Reliability - Issues Caused

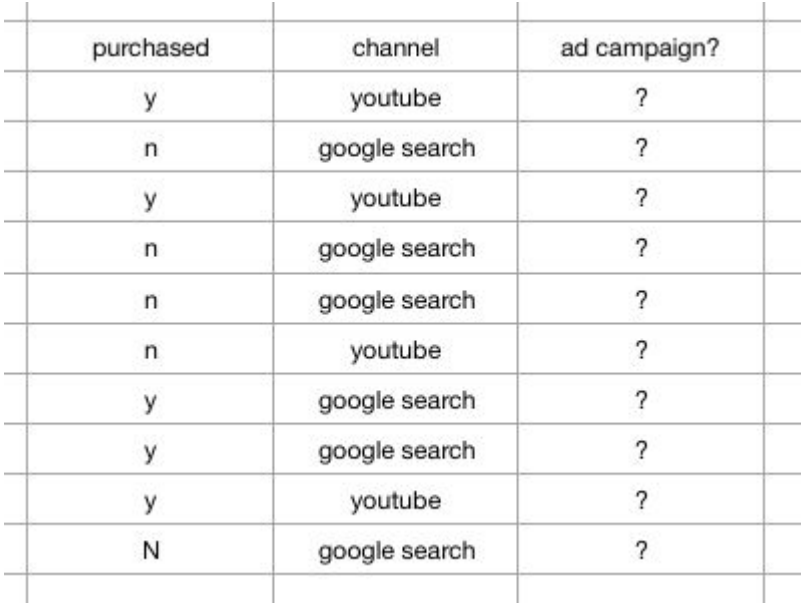
- We think we are right when we are not! We, therefore, develop a solution or analysis that will not generalize in the future.
- We try to understand and effectively handle the reliability issues…but end up in a never-ending loop of doing so!
- Reliability issues are pernicious, and they can silently kill your analytics work…and you only find out when you’re smacked in the face that your analysis didn’t hold up “in the real world”.

Reliability - Detection

- The best antidote is a healthy dose of skepticism. Keep your eyes peeled for anything that looks “off”.
- Exploratory data analysis - summary statistics and basic plots - can help reveal data that seems suspect.
- We can also test in a production-environment, called a shadow deployment.
  - We build a machine learning model and deploy it into production.
  - We make predictions and record them.
  - However, the predictions have no influence on any decision. They are simply recorded in the background.
  - We can then see how they model is performing. If the model has degraded notably from our expectations, we have a bad model. One reason could be that our training data was unreliable.

Completeness - Example

- We can always have more complete data. The question is if our analysis is *good enough* to surpass the status quo. If so, our data is *complete enough*.



purchased	channel	ad campaign?
y	youtube	?
n	google search	?
y	youtube	?
n	google search	?
n	google search	?
n	youtube	?
y	google search	?
y	google search	?
y	youtube	?
N	google search	?

Completeness - What Might Data Be Incomplete?

- Some data might be too invasive to collect.
- We didn’t think about capturing certain data.
- Data collection systems might have bugs.
- Point of Advice: Always capture as much data as you can, even if you don’t know if you will use it…because you *might* need it!

Completeness - Issues Caused

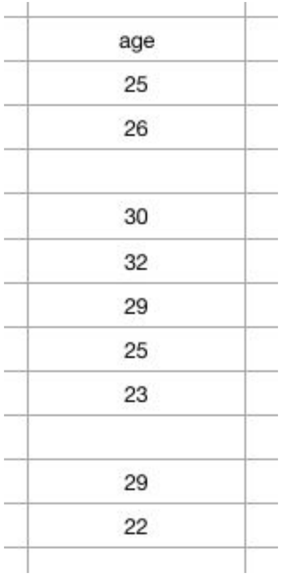
- Omitted variable bias - false attribution
  - Recall the example from Week 1?
- Omitted row bias - misleading distributions / representations
- We find the wrong thing because of false attribution or distributions that don’t generalize to broader applications. This is a major source of biased algorithms!
- We don’t find anything at all due to lack of signal.
- Our findings are more tentative than they otherwise would be.

Completeness - Detection

- Detection of incompleteness if mostly qualitative and accomplished via professional opinion and subject-matter expertise.
- Did we detect the wrong thing? Do we have a key missing piece? Did we think we could have found a better answer?

Missingness - Example

- Missing data is workaday in the real world!
- Word of Advice: Missing values may be hidden! Perhaps an upstream system imputes a value - sometimes 0 - when the original value is missing.



age
25
26
30
32
29
25
23
29
22

Missingness - Why Might Values Not Be Present?

- Same as completeness in addition to data not being required in all cases.

Missingness - Issued Caused

- May not be an issue - missingness may be an important and telling attribute! That is, the fact an observation has a certain value missing may tell us something useful.
- If that’s not the case, we might end up with weaker insights.
    - Some data may have so many missing values that it becomes unusable!
- If we handle missing values incorrectly, we may even end up with *more wrong* takeaways.

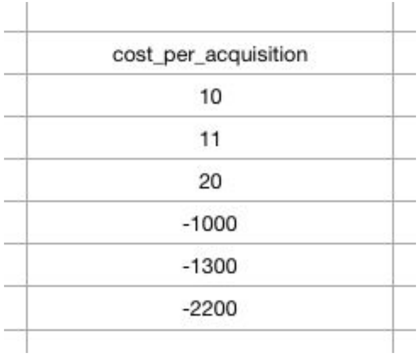
Missigness - Detection and Handling

- Python demo!
- For posterity, we broadly impute missing values with the following strategies:
  - Constant value (most appropriate when we can infer that being missing is a distinct category or when missigness could be interpreted as something like a zero).
  - Measure of central tendency (mode, mean, median).
  - Machine learning model to predict what the value likely could be (this is oftentimes overkill but could be worthwhile in certain cases).
  - Be aware of feature leakage! (Shown in demo).

Consistency - Example

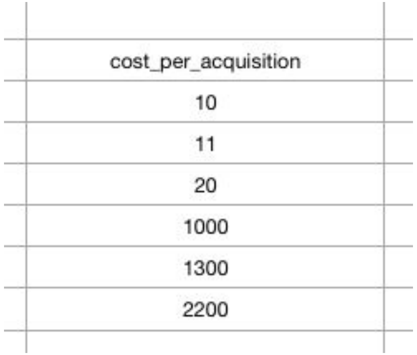
- Consistency is different from reliability. With reliability, we inherently cannot trust the data. Unreliable data may be consistent *or* inconsistent.
- With inconsistent but reliable data, we can make the data consistent again with some effort.

Inconsistent and unreliable - this value should never be negative (and started to be recorded in pennies)



cost_per_acquisition
10
11
20
-1000
-1300
-2200

Inconsistent but reliable - values started to get recorded in pennies



cost_per_acquisition
10
11
20
1000
1300
2200

Consistency - Why is Data Inconsistent?

- The biggest drivers of unreliable data are inconsistent collection and system bugs, much the same as Reliability.
- “We changed some process over time and stopped collecting X.”
- “A” now means “B”.

Consistency - Issues Caused

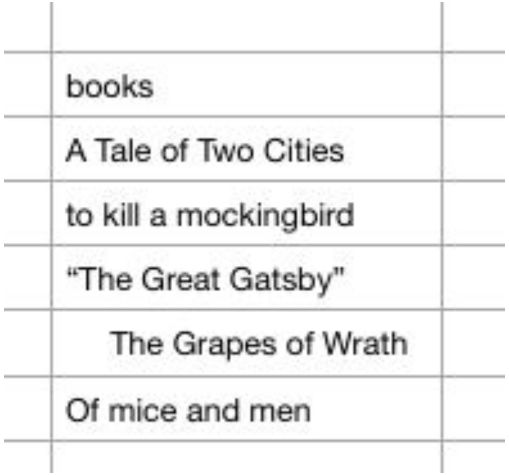
- Our analysis might not hold over time - we may rely on old data without realizing it.
- Our code may fail due to unexpected values.
- We may have unnecessary sparsity (e.g., “a” and “b” could be combined).

Consistency - Detection

- Python demo!

Cleanliness - Example

- Unclean data might have slightly different formats or nuisances. However, unclean data - on its own - is not unreliable or inconsistent. It simply needs to be massaged to be usable!



books
A Tale of Two Cities
to kill a mockingbird
“The Great Gatsby"
The Grapes of Wrath
Of mice and men

Cleanliness - Why Might Data Not be Clean?

- Oftentimes, unclean data is a function of messy data collection, poor normalization, or bugs in upstream systems.

Cleanliness - Issues Caused

- Increased and arbitrary sparsity of categorical data.
- Influence of inaccurate outliers when dealing with numeric data.

Cleanliness - Detection

- Exploratory data analysis is your best friend! Plotting distributions for numeric and categorical can unveil a ton.

Biased Models

- The underlying math is not biased. However, the data going into a model can create bias.
- Omitted variable bias - attribution goes to the incorrect features
- Omitted row bias - our rows are not reflective of the actual distributions; therefore, they fail to generalize to future applications

Biased Model

- We have tools that can help us identify bias, but the onus is on us - we have to be on the lookout for it!
- We might be tempted to not use any data that might create bias. This might help with omitted variable bias. However, if we have omitted row bias, we won’t have the requisite data to check if bias exists.
- We don’t have to model on features that can create bias. But we should at least append them post hoc to see if our model did create bias.

Python Demo: Fairlearn