Introduction to Data Science

UNDERSTANDING YOUR SAMPLE

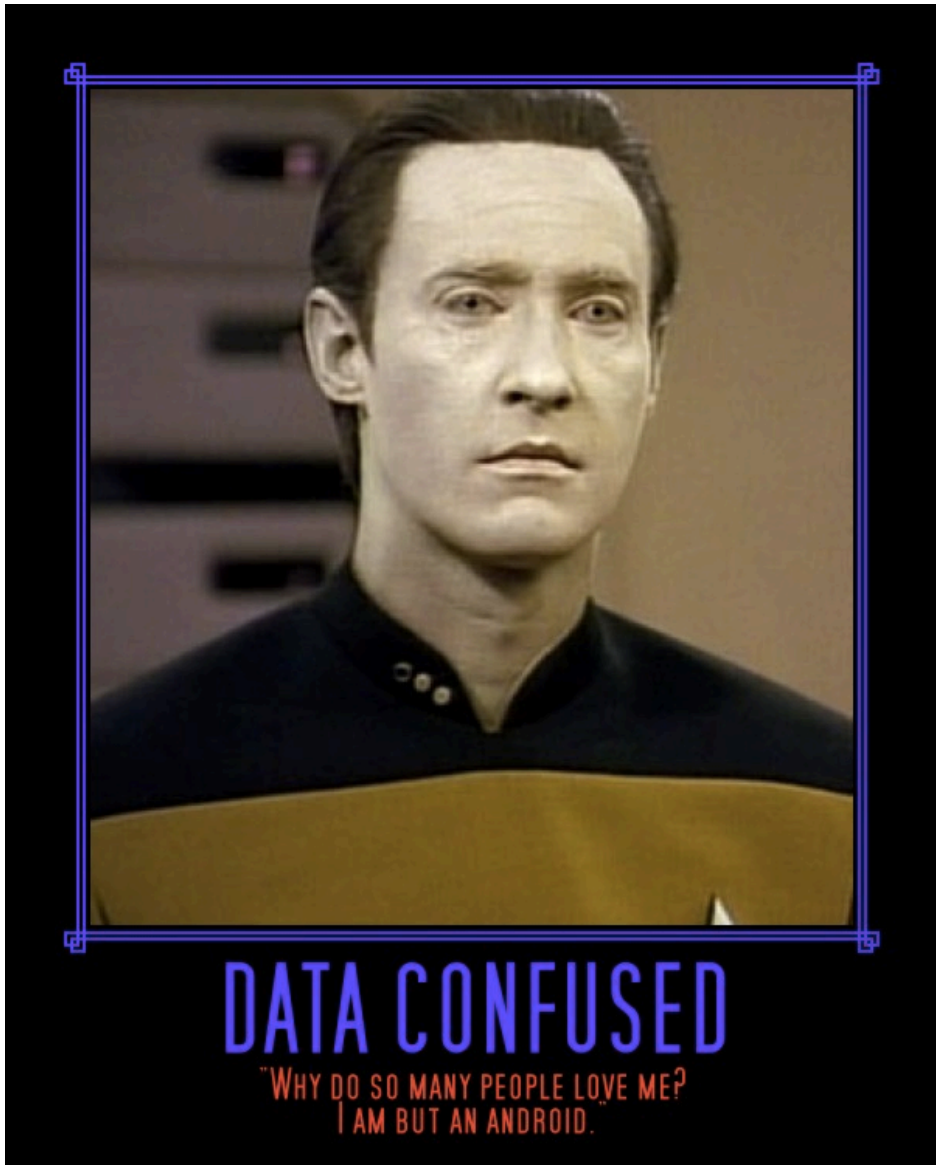
BRIAN D'ALESSANDRO

ADJUNCT PROFESSOR, NYU

FALL 2019

DATA

WHAT IS DATA ANYWAYS?

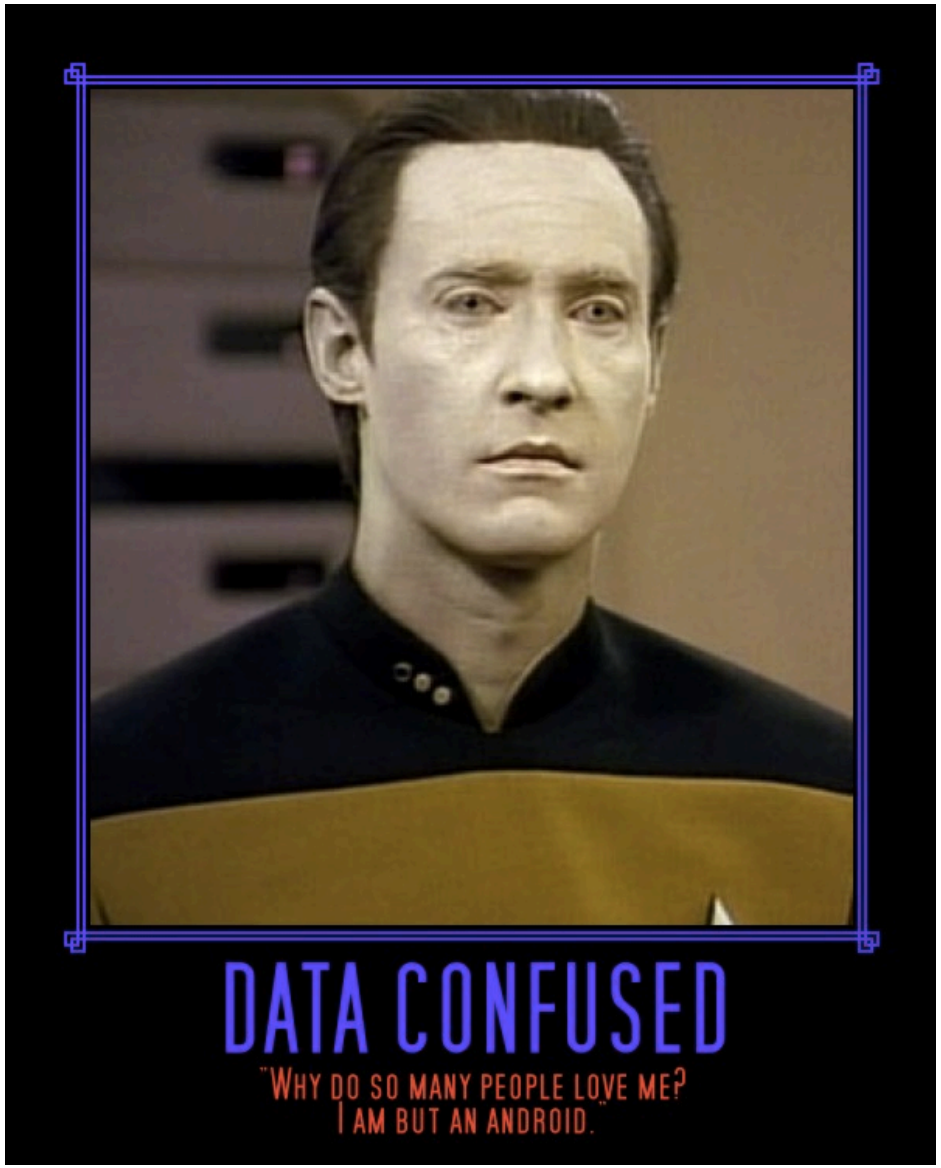


Files
(i.e. .txt, .binary, .csv)?

Programming Objects
(i.e. tuples, arrays, hashMaps, dicts)?

Variables and their distribution
(i.e. X, P(X))?

WHAT IS DATA ANYWAYS?



For now we're going to focus on this meaning of data.

Variables and their distribution (i.e. X, P(X))?

SOMETHING YOU SHOULD KNOW

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

> Yet far too much handcrafted work — what data scientists call "data wrangling," "data munging" and "data janitor work" — is still required. Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

Source:http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0

GARBAGE HEADLINES

Don't think of data preparation as a nuisance or pejorative

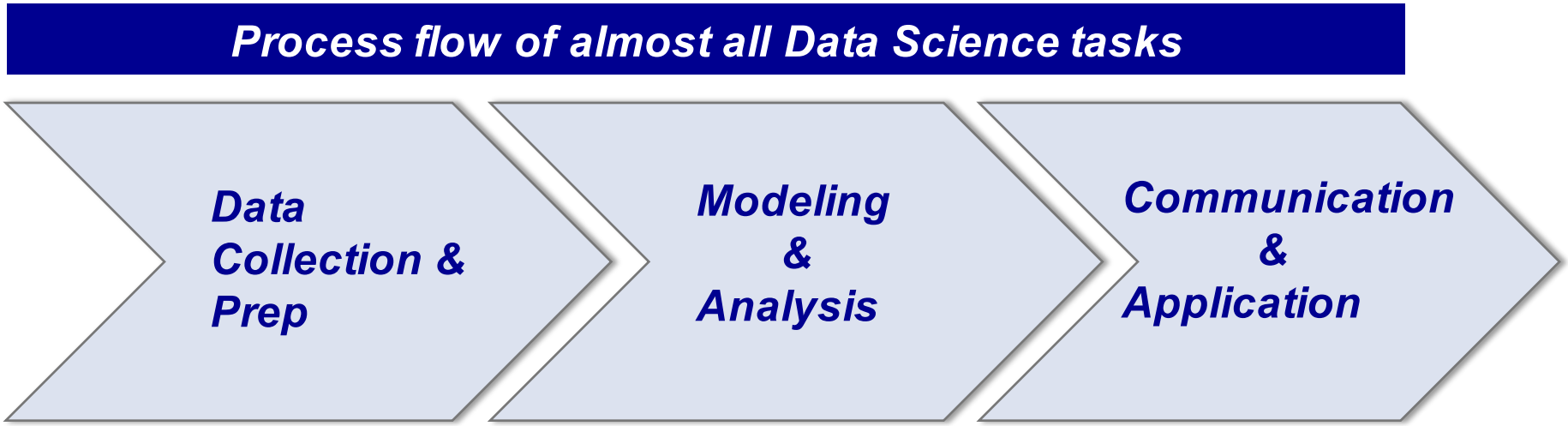
TECHNOLOGY

For Big-Data Scientis s, 'Janito Work' Is ey Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

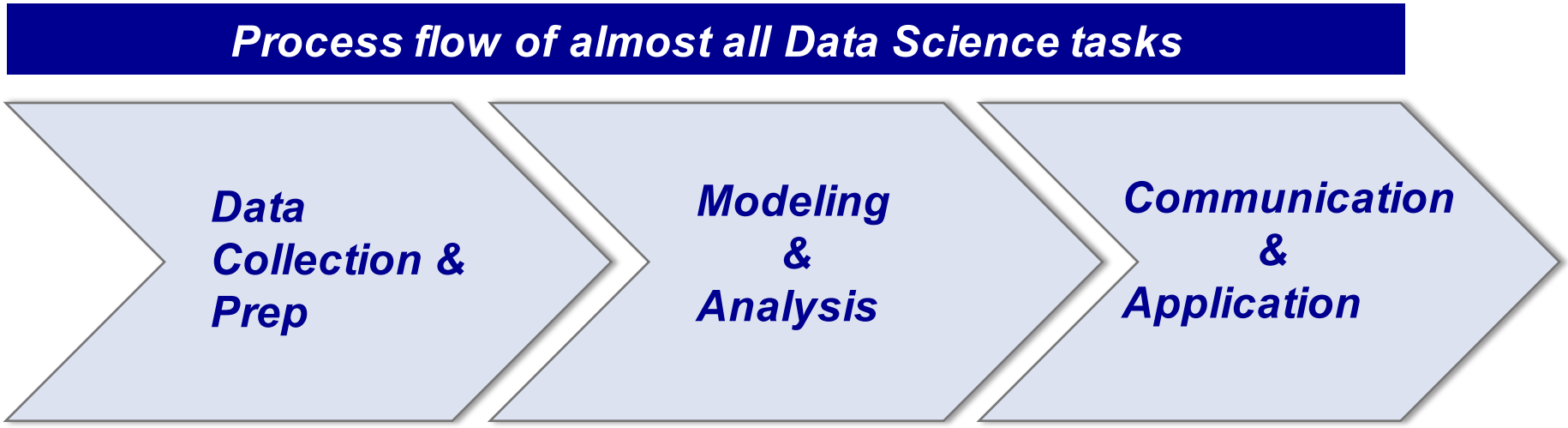
Data preparation and understanding is integral to great analysis. It is well worth spending 80% of your time getting it right!

DATA PREP => SUCCESS



This is the *sexy* part, getting all the research and attention from competitions …
and it is indeed fun!

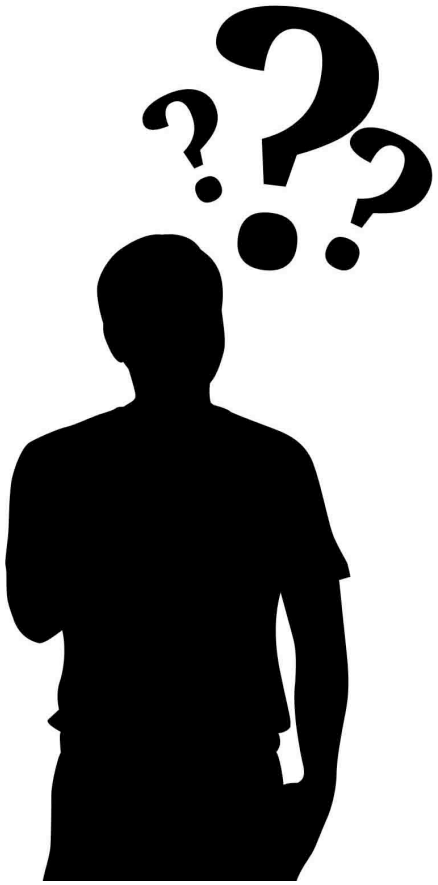
DATA PREP => SUCCESS



These are what make you a good and effective data scientist however

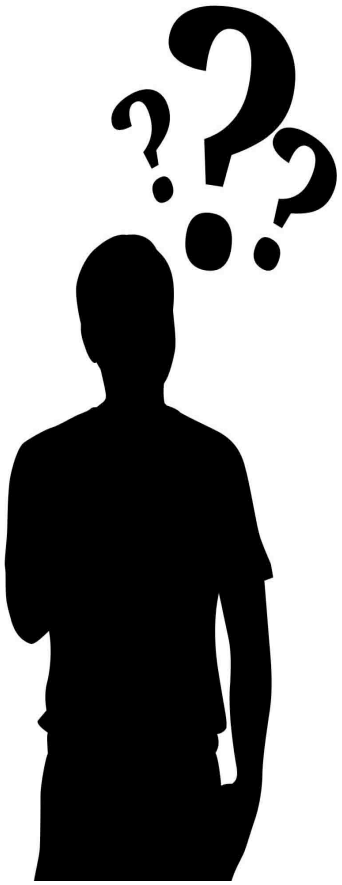
2 KEY QUESTIONS

1. Where did my data come from?
2. What does it look like?



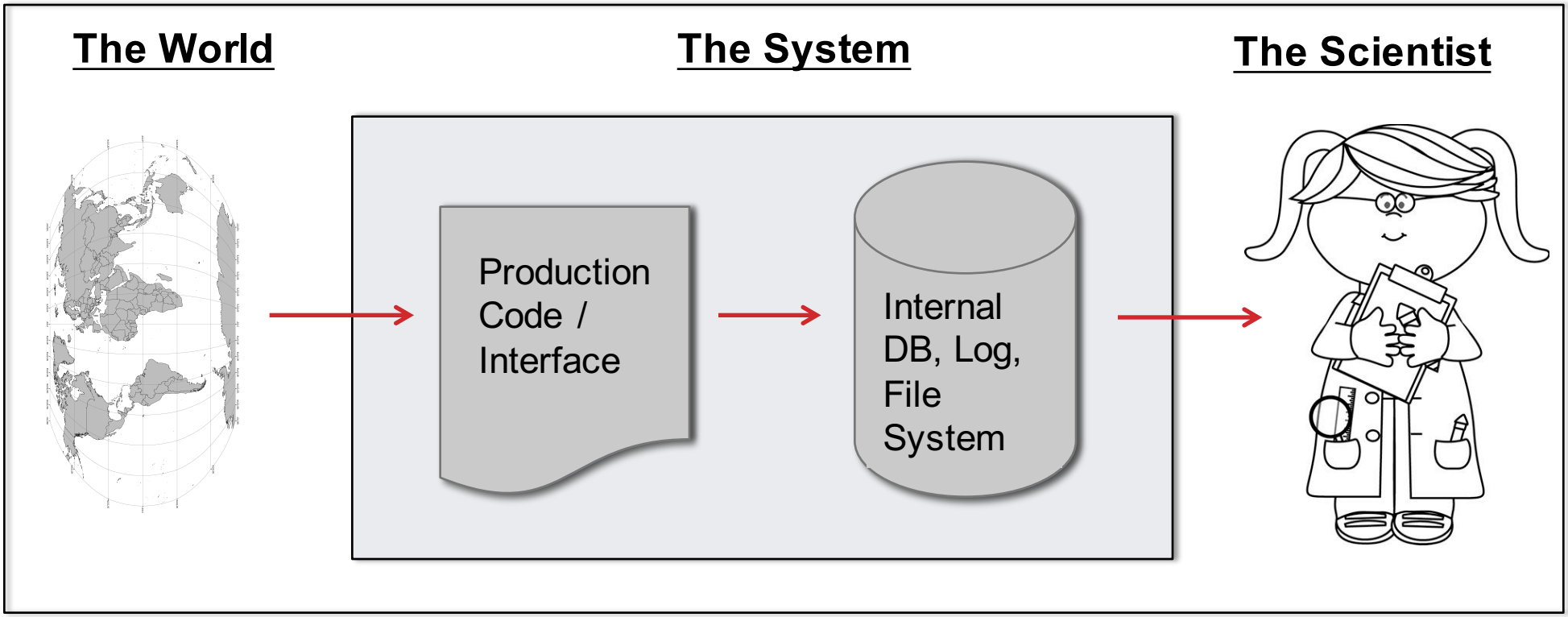
ALWAYS KNOW THE SOURCE

1. Where did my data come from?
   - Internal ETL processes
   - Production logging/sampling
   - Web scraping/ API
   - Survey/Panel



Rule of thumb: If you did not pull your own data, always be skeptical (that it is indeed what you need), and spend extra time validating it.

COMMON DATA FLOW



Challenges we often face:

- Often the data scientist doesn't write/own the code that produces the data. (Selection Bias, Unknown Unknowns)
- The system intervening on the world changes the nature of the data we collect from it. (Selection Bias, Negative Feedback Loop)
- We have no control often of data that streams into the system. (Concept Drift & Selection Bias)

source: http://www.nytimes.com/1989/08/22/science/on-landing-like-a-cat-it-is-a-fact.html
https://www.youtube.com/watch?v=TGGGDpb04Yc

OCCAM'S RAZOR QUIZ

Conclusion derived from the data…

Even more surprising, the longer the fall, the greater the chance of survival.

Explanation 1 (per the article):

"Cats may be behaving like well-trained paratroopers," Dr. Jared Diamond, who teaches physiology at the University of California at Los Angeles Medical School, wrote in the August issue of the magazine Natural History.

Explanation 2 (per a reasonable data scientist):

Nobody brings their dead cat to the hospital, therefore this data suffers from selection bias. Cats that are clearly alive after a fall are more likely to be brought to the hospital, and more likely to survive. It is likely that the ambiguity of a cat's condition decreases with the height of the fall. Cat's are either obviously dead or obviously alive, which explains the trend in the data.

ANALYZING THE FALLING CAT ANALYSIS

It's always good to think through an analysis and sampling using the language and tools of probability.

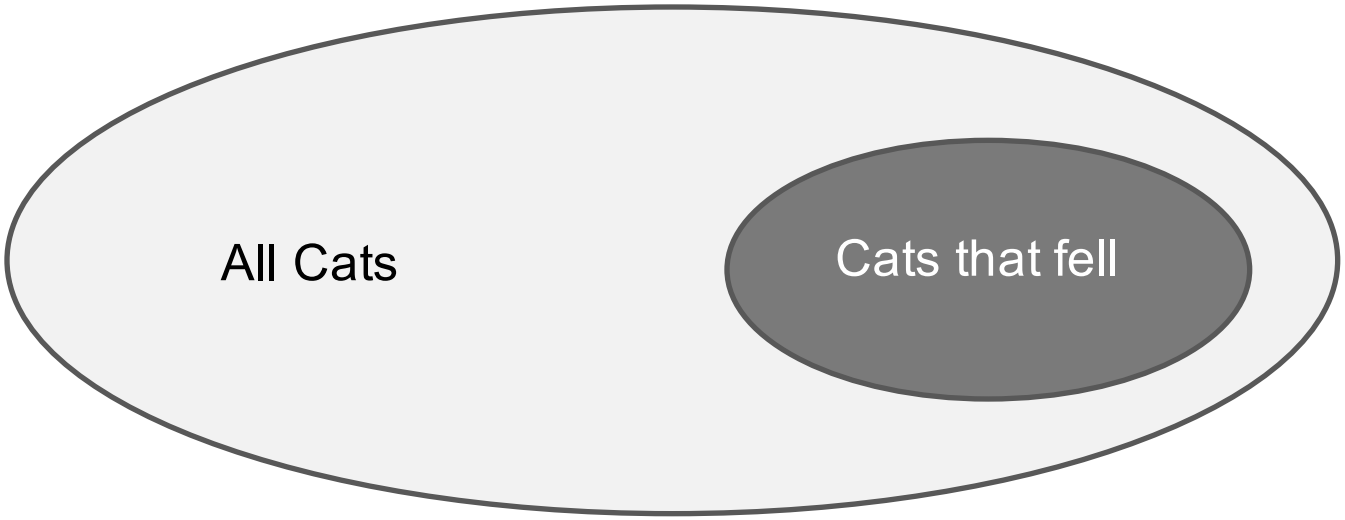
I.e.,

As a starting point, we're interested in cat survival. So let's say the we want to estimate using some data: *P(Survive)*. Note that specifying this probability essentially defines the problem, so being precise is incredibly important!

But is this precise enough? How might we change this to reflect the analysis? Should we modify event of interest ("Survive" vs something else)? Should we make it a conditional probability?

ANALYZING THE FALLING CAT ANALYSIS

Let's instead focus on *P(Survive 1 week)*. Let's also be more precise about the sub-population. We don't want all cats, we want all cats that fall from a window.



So let's reformulate the problem as estimating *P(Survive 1 Week | Fell)*.

ANALYZING THE FALLING CAT ANALYSIS

Can we estimate *P(Survive 1 Week | Fell)* from data collected by the vet?

Questions to ask:

1. Does the vet data represent all cats that fell?
2. If not, is the data missing at random?

MAR in this case is equivalent to saying:

$$P(\text{Go to vet} \mid \text{Fell}) = P(\text{Go to vet} \mid \text{Fell, State of cat at Fall})$$

ANALYZING THE FALLING CAT ANALYSIS

We can use the law of total probability to work this out.

P(Survive 1 Week | Fell) =

*P(Survive 1 Week | Fell, ! Clearly Dead) *P(! Clearly Dead | Fell) +*

*P(Survive 1 Week | Fell, Clearly Dead) * P(Clearly Dead | Fell)*

*= P(Survive 1 Week | Fell, ! Clearly Dead) *P(! Clearly Dead | Fell)*

Let's assume the sampling mechanism is: *If ! Clearly Dead => Go to vet*

Thus:

P(Survive 1 Week | Fell, Sampled) = P(Survive 1 Week | Fell, ! Clearly Dead)

ANALYZING THE FALLING CAT ANALYSIS

So the reporting implies the following:

P(Survive 1 Week | Fell, Sampled) = P(Survive 1 Week | Fell)

But in reality they are actually measuring:

P(Survive 1 Week | Fell, Sampled) =

P(Survive 1 Week | Fell) / P(! Clearly Dead | Fell)

The degree of bias depends on how low *P(! Clearly Dead | Fell) is*.

SELECTION BIAS

Every analysis starts by drawing a data sample ***S*** from a population ***D.***

Each instance is characterized by a set of features ***(X,Y)***

If being in the sample ***S*** is independent of ***X*** and ***Y***, the sample is unbiased:

i.e. ***P(S|X,Y)=P(S)***

Else the sample is biased: i.e. ***P(S|X,Y)≠P(S)***

Source:
Zadrozny, Bianca. "Learning and evaluating classifiers under sample selection bias." *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004.

REFINING OUR DEFINITION

We can actually consider two types of selection bias:

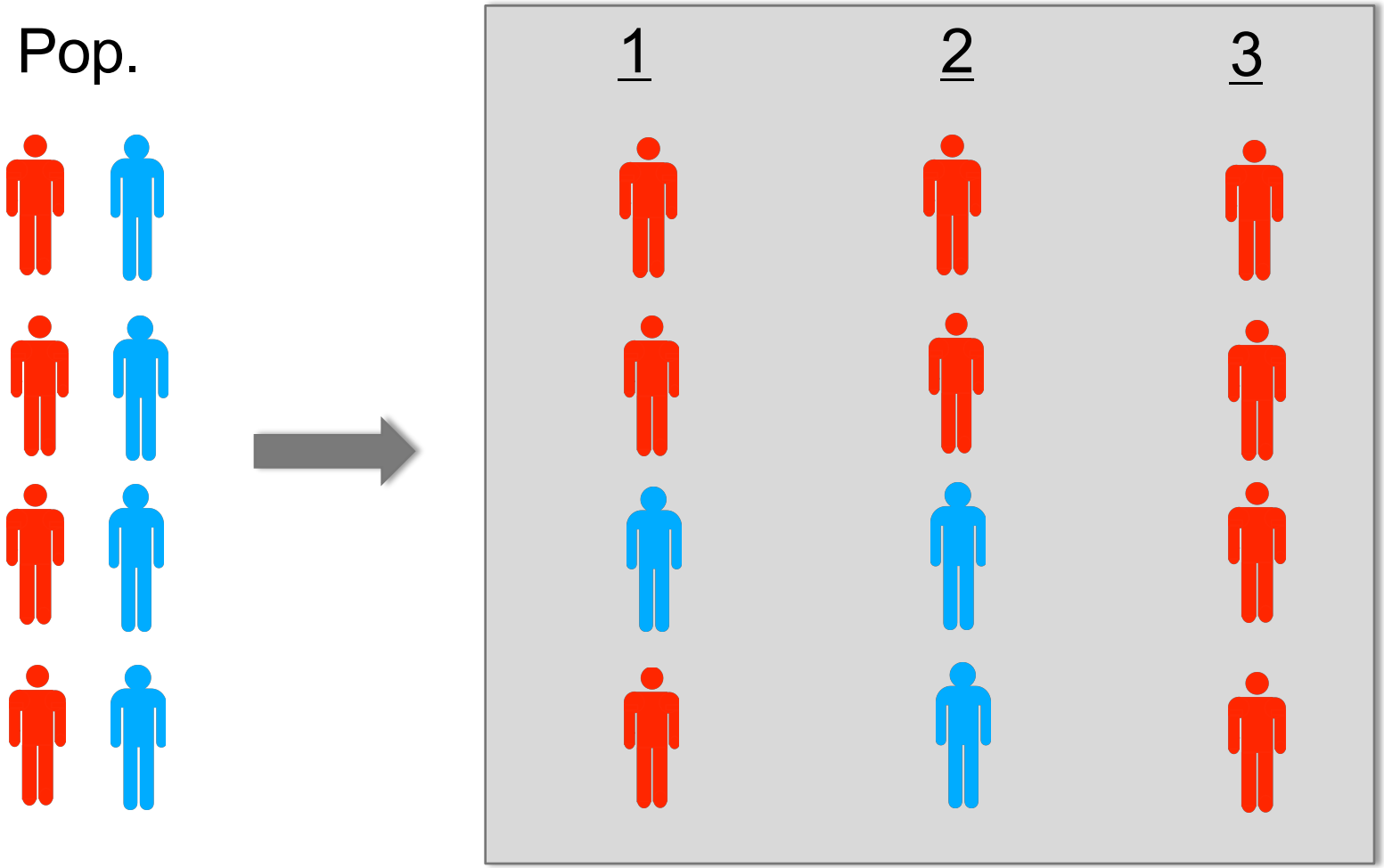
P(S|X,Y) = P(S|Y) *-> Bias only depends on Target variable*


- This type of bias is common, intentional and often justified
- A bi-product of stratified sampling, up/down sampling
- Is usually done to improve learning
- Impacts prior probability (base rate) and is easily corrected
- Needs to be corrected when running evaluation

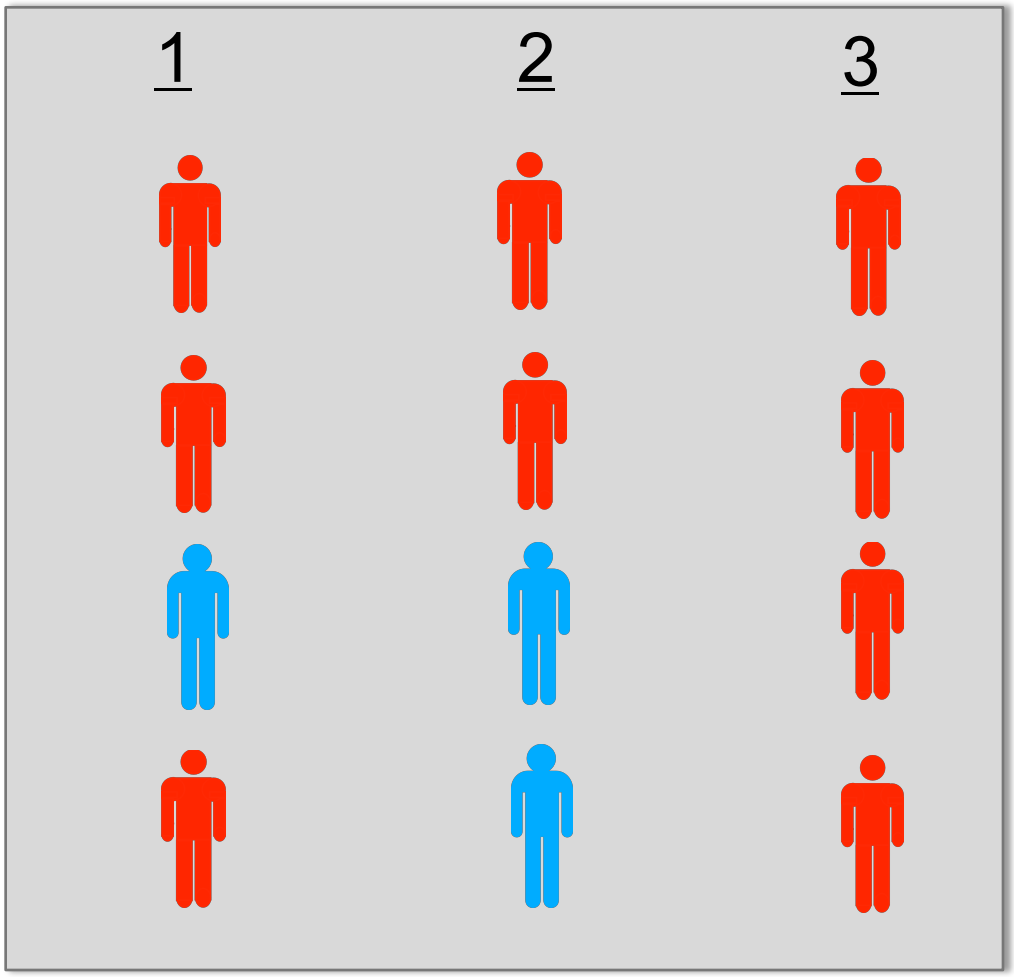
P(S|X,Y) = P(S|X) *-> Bias only depends on feature vector*

Source:
Zadrozny, Bianca. "Learning and evaluating classifiers under sample selection bias." *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004.

SELECTION BIAS – TOY EXAMPLES



SELECTION BIAS – TOY EXAMPLES



P(S1) = 0.5
P(S1|R) = 0.75
P(S1|B) = 0.25

P(S2) = 0.5
P(S2|R) = 0.5
P(S2|B) = 0.5

P(S3) = 0.5
P(S3|R) = 1
P(S3|B) = 0

SELECTION BIAS – IMPLICATIONS

Selection bias within data affects generalizability of results and potentially the identifiability of model parameters.

Generalizability:
Does your model represent the population at large?
Does your prediction match the production results?
Is your statistic representative of the greater population?

Identifiability:
Can you learn a model, parameter or statistic given the data at hand?

e.g,
in previous example, sample 3. Let's assume we want to know the average sales for blue figures, i.e., E[Sales|B]. Because P(Samp3|B)=0, we can not learn this parameter from the data.

SELECTION BIAS – EXAMPLE 1

This is a case of selection bias happening because of technical constraints.

1. A user comes into the system  *, uid=10000*

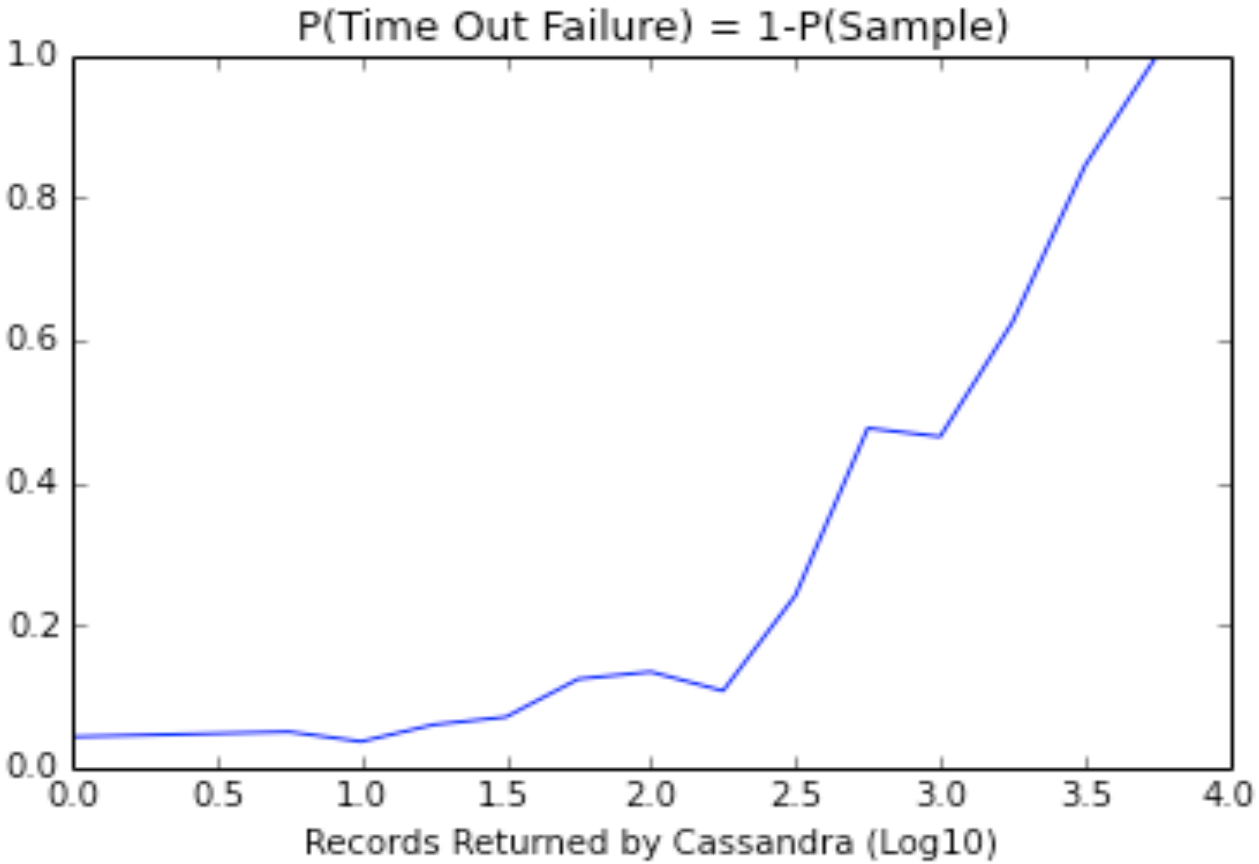
2. Query Cassandra (production database) to get user's data:

3. Log user event and Cassandra output to a research database

Now what could go wrong?

SELECTION BIAS – EXAMPLE 1

It turns out, when you have a user with a lot of data attached to it, the Cassandra output is too big and it clutters up the network. The result is that the system is more likely to time out and kill the query, and the user never gets logged.



Its pretty clear that $P(S|X) \neq P(S)$, where X is the number of records attached to the user.

SELECTION BIAS – EXAMPLE 2

Sometimes, selection bias is almost intentional, and is a rational decision caused by business and economic factors.

Credit Risk Modeling

Goal: Predict P(Default in 6 mo's | Application Data)

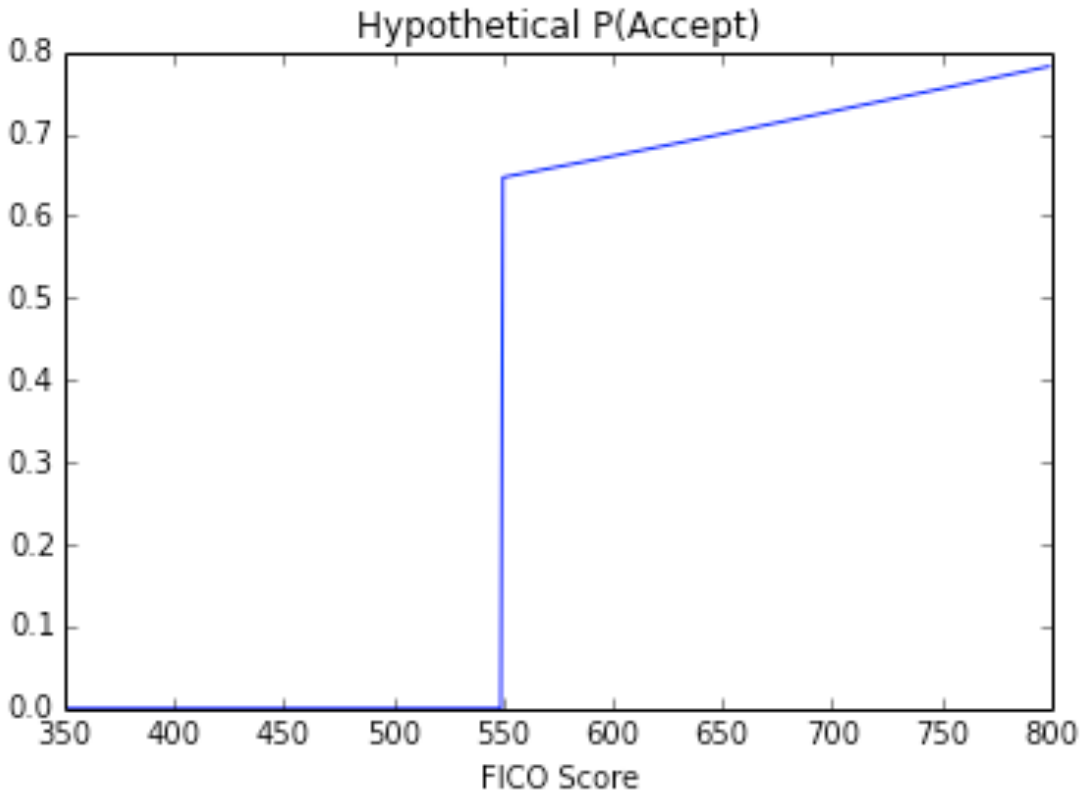
Method:

1. Sample new credit card users, log app data
2. Observe 6 months, log if user defaults
3. Build a predictive model on sampled observations

Now what could go wrong?

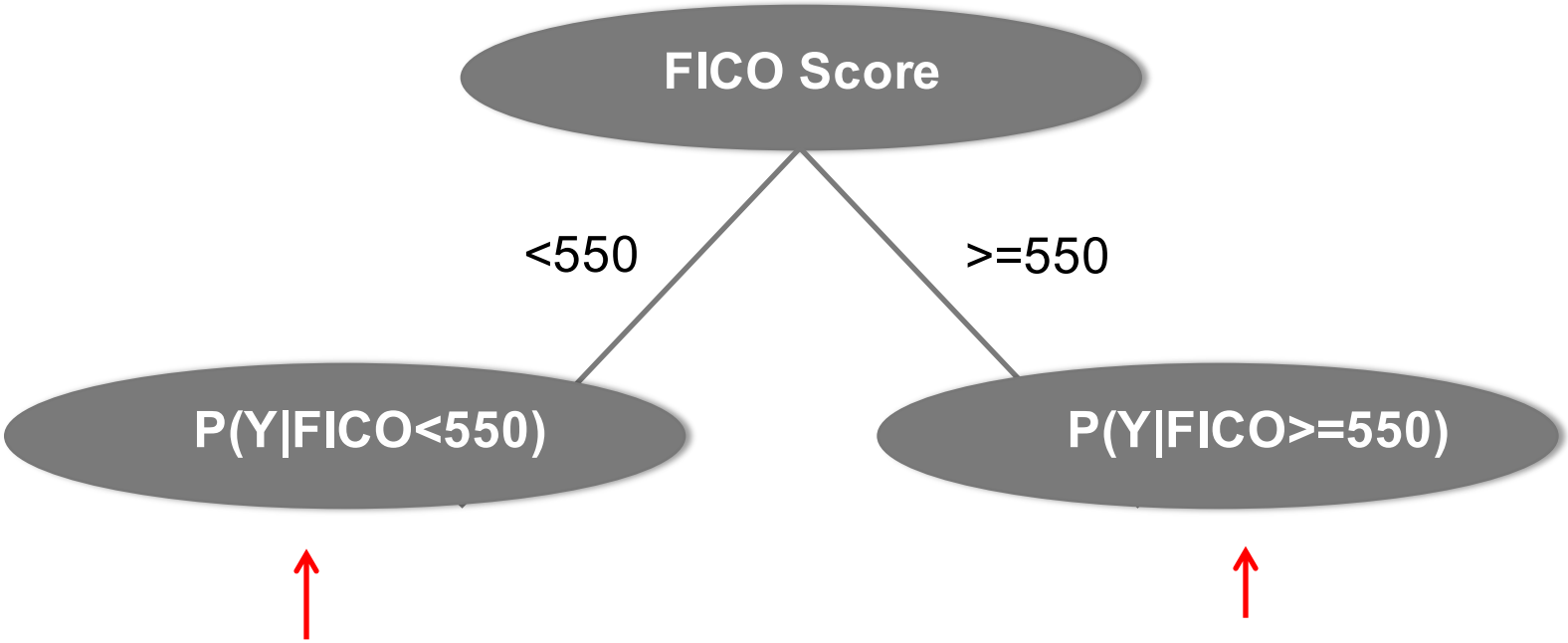
SELECTION BIAS – EXAMPLE 2

Q&A:
What is P(Sample|Fico<550)?
What implication does this have on model building?
Is it worth it?



SELECTION BIAS – EXAMPLE 2

The scenario here is similar to a Logistic Model Tree, but where a portion of the data is missing:



We have no data in this region, so it is technically not identifiable.

If the model is truly linear, we can extrapolate to this region, but that could be an expensive assumption!

We have the data to estimate this,…

but we can't generalize to the entire population.

SELECTION BIAS – WHAT TO DO

Selection bias is one of the biggest realities of production system data collection. What can you do about it?

1. Avoid It.

Design and use random sampling schemes as much as possible.

2. Adjust It.

In many cases you can statistically adjust for selection bias by weighting examples by 1/P(Samp|X) or Heckman Correction. In some cases your models will be fine (i.e., Logistic Regression w/ full data support).

3. Expect It.

Whether by design or accident, selection biases are likely to occur. Its always important to anticipate it and prepare for how it might affect your analysis.

AND ON TO CONCEPT DRIFT

aka. Non - Stationarity

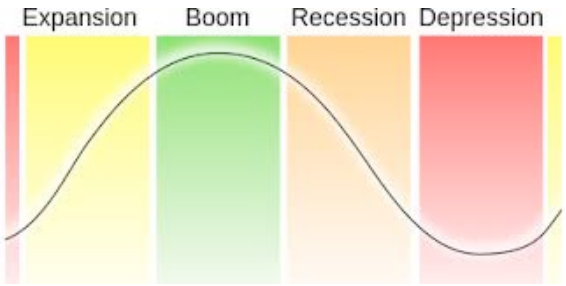
Defined simply as ***P(X)*** or ***P(Y|X)*** that changes over time, and is almost a fact Of life.

Example causes are…

Seasonality

Promotions (Exogenous Shocks)

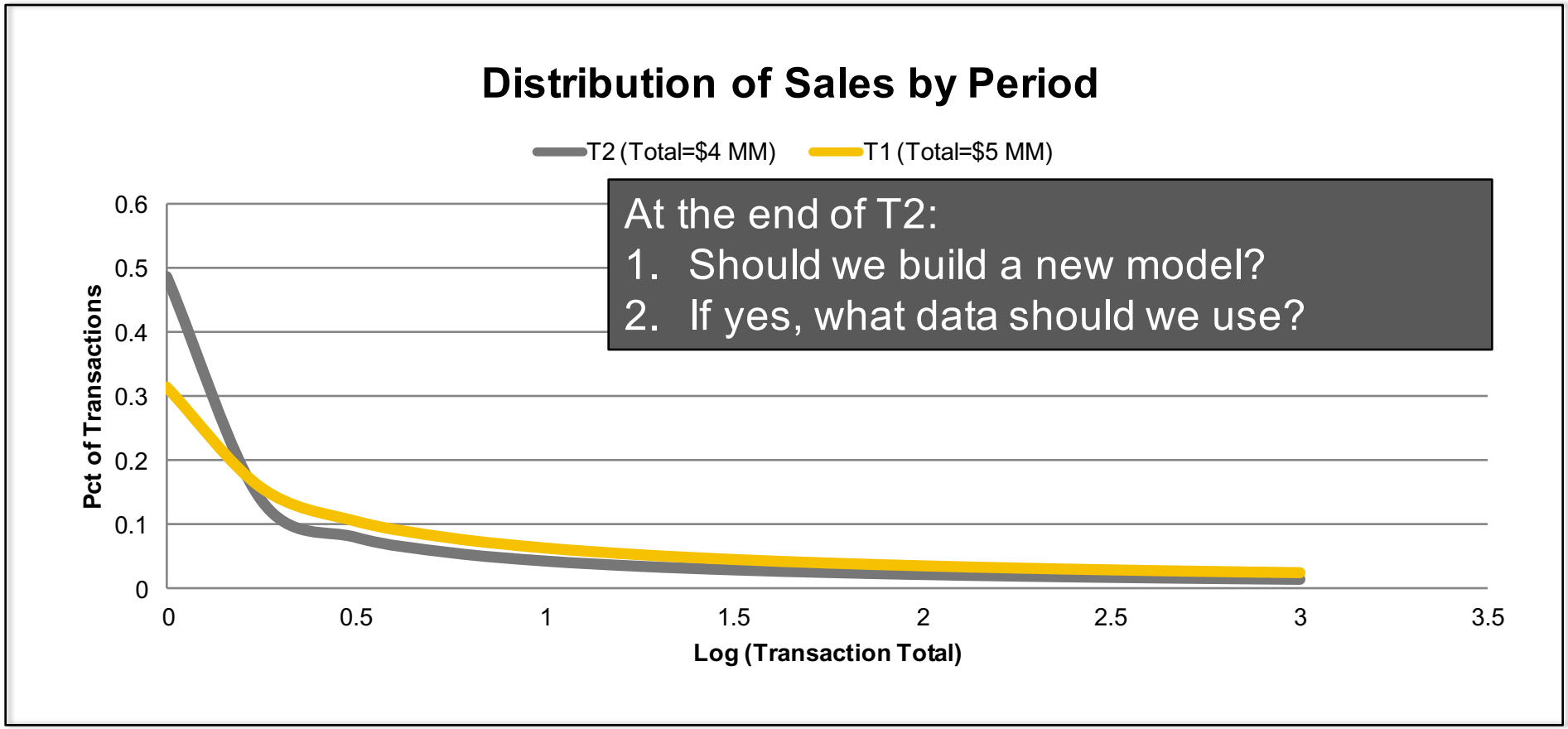
Economic Cycles



SIMPLE ILLUSTRATION

T1 (w/ TV Campaign)	T2 (No Campaign)	T3 (No Campaign)

We built a model here, to predict sales as a function of a user's history and demos.

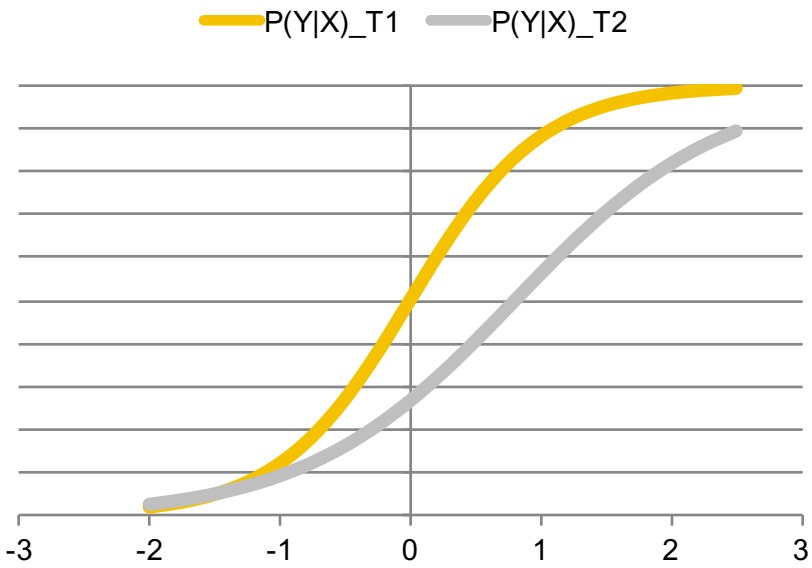


SIMPLE ILLUSTRATION

1. **Should we build a new model?**

We have proof that the distribution of transactions has changed, suggesting the underlying driver of purchase is different without the campaign. We know that T3 will look more like T2 than T1, so we probably should rebuild.

2. **If yes, what data should we use?**



We can pull both datasets and compare models. If models are very similar, pool data.

Otherwise, it's a balance between having more data and having the right data.

CONCEPT DRIFT TAKEAWAYS

Monitor predictive performance

You don't know future distributions, but model predictive performance should tell you when changes are happening.

Retrain as often as possible

The simplest way (in theory) to deal with an out-of-date model is to build a new one.

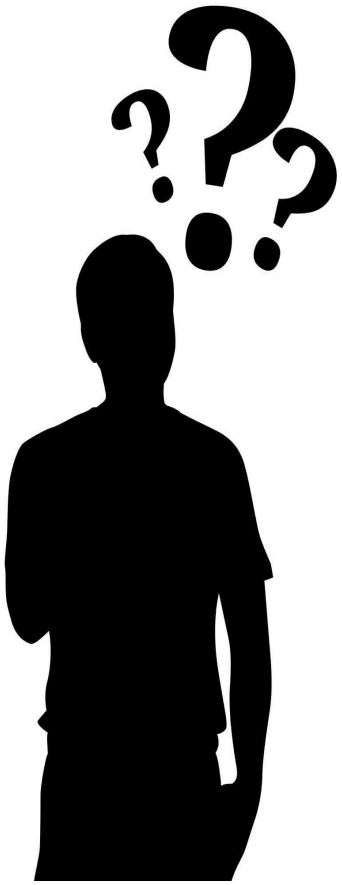
Test balance between data recency and data volume

This ties back to the classic bias-variance tradeoff, which is at the heart of many design decisions in data science.

THE SHAPE OF YOUR DATA

2. What does it look like?

- Distribution
- Feature types
- Missing values/outliers



A ROSE BY ANY OTHER NAME

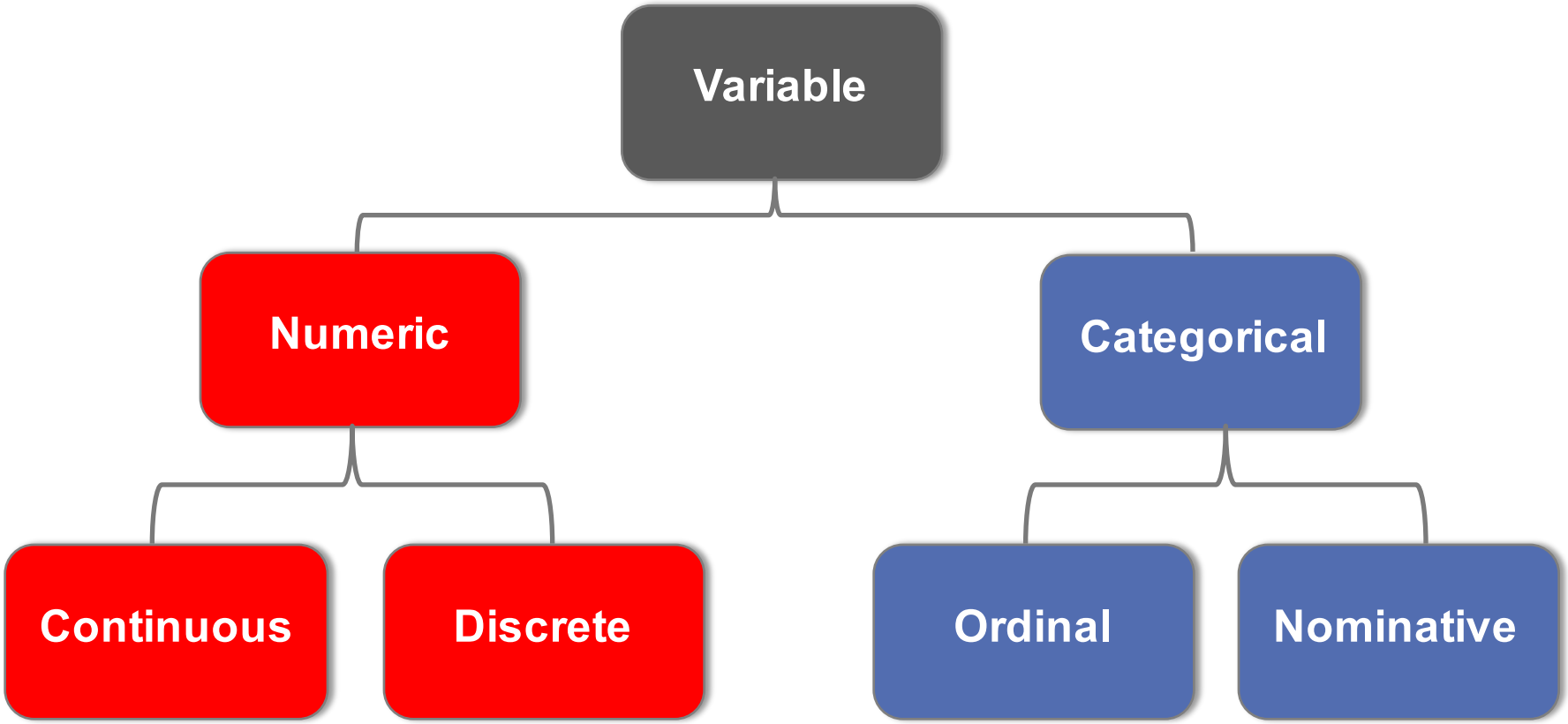
Independent of type, it is also important to classify a variable according to how it will be used or analyzed.

	Statistics	Machine Learning	Common Symbols
Thing you want to predict	Dependent Variable	Target Variable	Y, f(X)
Things you use to predict	Independent Variable	Feature	X

NB: Admittedly, I often use these terms interchangeably.

TYPES OF FEATURES

Most variables fall under four basic categories



IMPLICATIONS OF FEATURE TYPE

Modeling

- Every algorithm has requirements on what type of data can be used as an input. I.e., regression based methods require numeric or binary variables, while tree methods can accept any type.

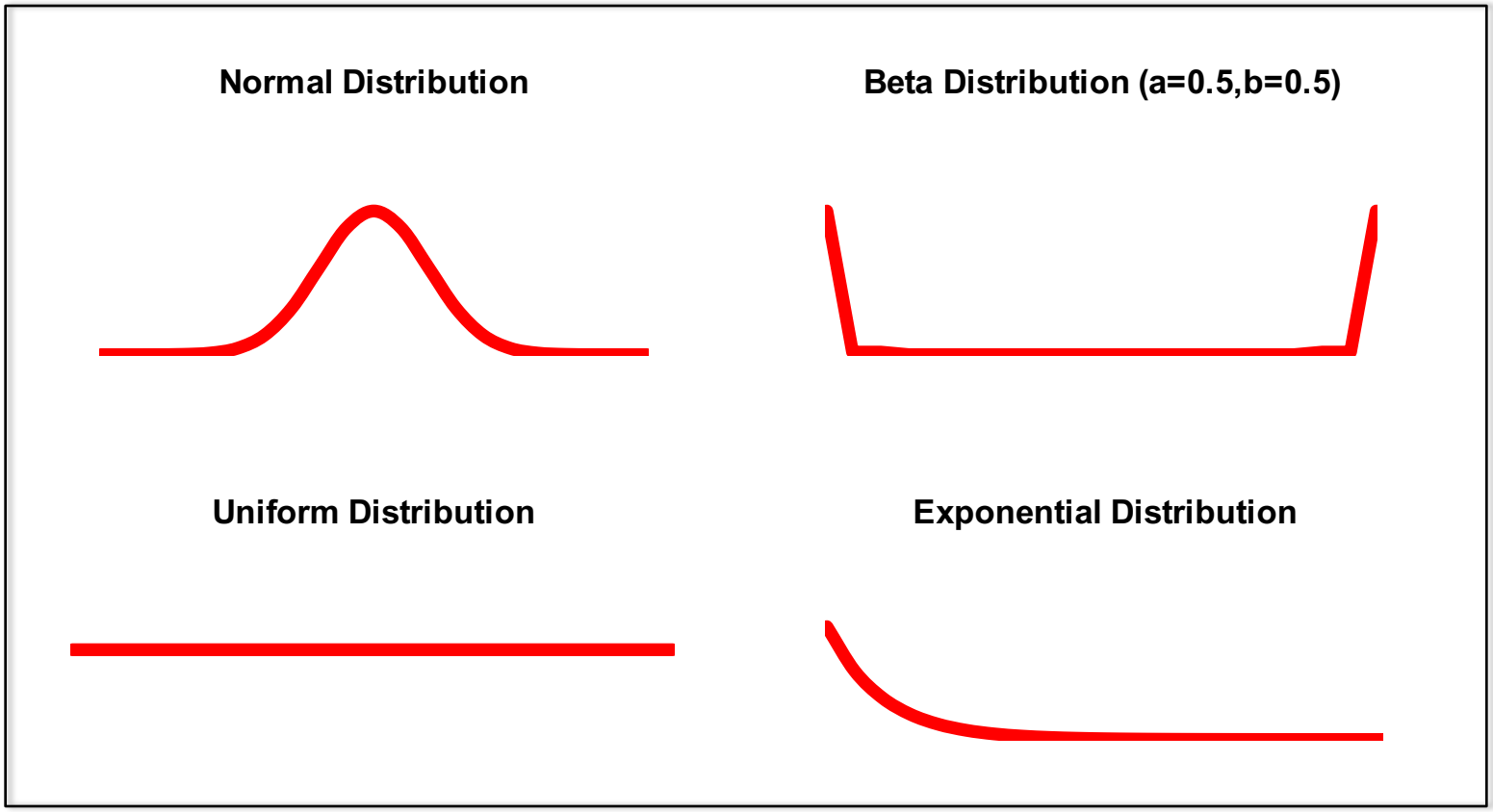
- Often times you can 'cheat' linear models by transforming the data correctly (more on this in later lectures).

Analysis/Exploration

- Distributional statistics aren't defined on categorical data, but you can use categorical data to compare statistics across category groupings.

DISTRIBUTIONS

This isn't a probability course, but data can usually be characterized by its distribution and associated statistics.

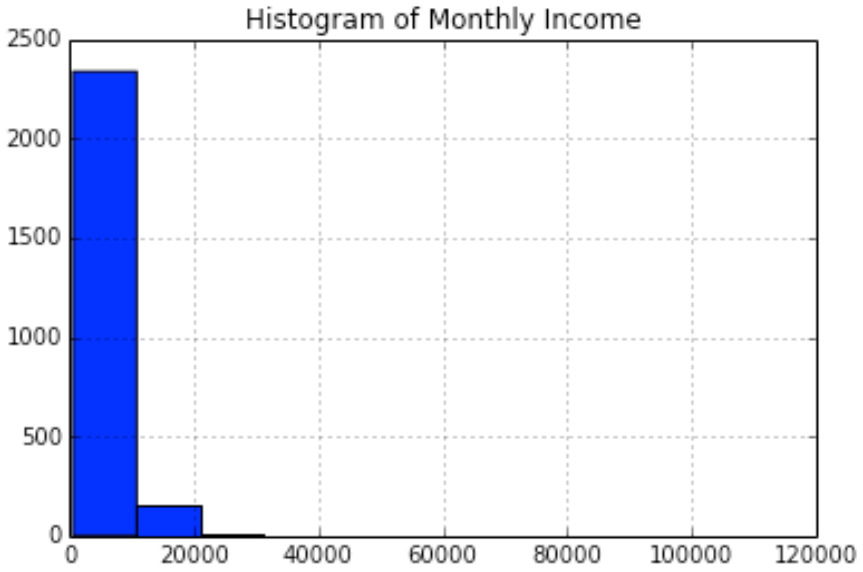


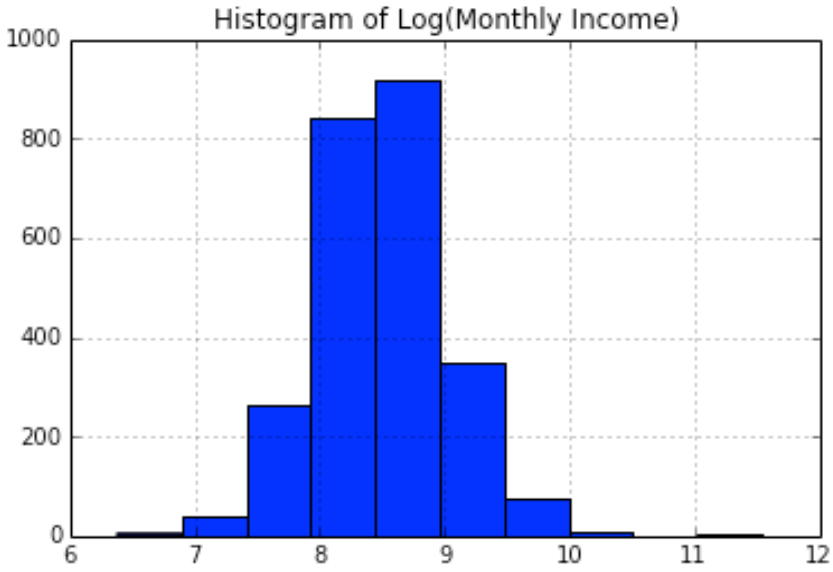
In reality though, empirical data rarely conforms to standard probability distributions. These are still useful for simulations for smoothing/regularization using various Bayesian methods.

EMPIRICAL DISTRIBUTIONS

Most data comes in as is and it doesn't always matter what you call its distribution…

but its still useful to understand its shape. The histogram is a great tool for this.



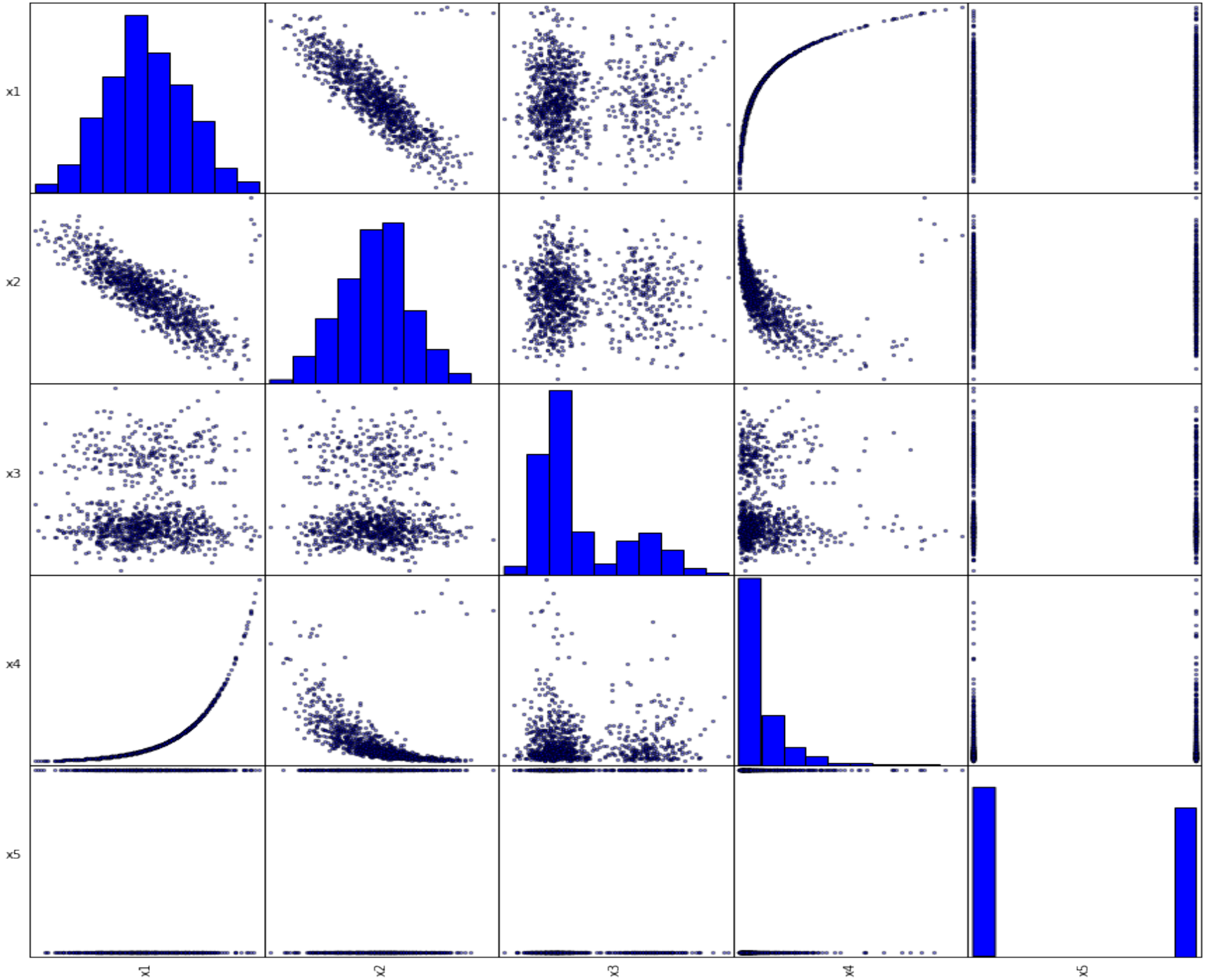


EMPIRICAL DISTRIBUTIONS

It is often very useful to know the distributional statistics of your data…

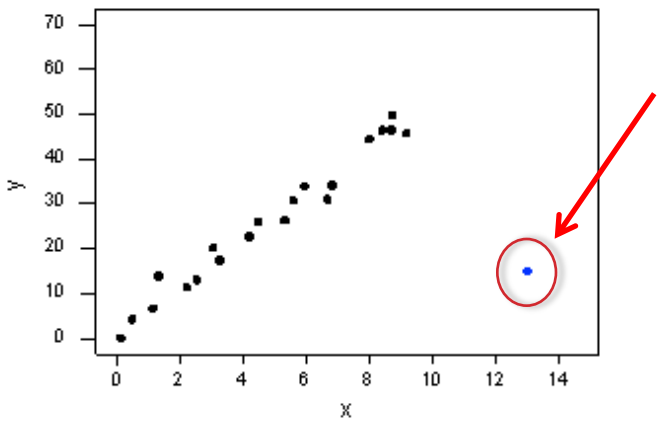
```
In [7]: loansData['Monthly.LogIncome'].describe()

Out[7]: count    2499.000000
        mean        8.501915
        std         0.523019
        min         6.377577
        25%         8.160518
        50%         8.517193
        75%         8.824678
        max        11.540054
        dtype: float64
```



DATA CLEANING

Beyond general data exploration and transformations, one usually wants some degree of data cleanup, looking for outliers and missing values.



Outliers are often the result of erroneous data collection or processing. Not only are they a good clue for data processing QA, but they can have severe influence on model estimates or summary statistics.

Missing values can indicate processing/collection errors. When present in the data (as either null, NA or “”), they can break a lot of modeling algorithms.

HANDLING OUTLIERS/NULLS

What should one do when faced with these issues? The easiest answer in data science is *it depends…*

If missing/bad at random…

- If the occurrence is rare, delete observations (most extreme)
- Can also impute/replace with average for that feature.

Otherwise…

- Impute with some constant (usually the average), create a dummy variable to indicate missing/bad
- Exploit multi-collinearity, i.e., use a model to estimate E[Missing Val|X].