MANNHEIM
BUSINESS SCHOOL

DATA ETHICS

PROF. DR. FLORIAN STAHL

UNIVERSITY OF MANNHEIM

AACSB ACCREDITED

EQUIS ACCREDITED

ASSOCIATION OF MBAS ACCREDITED

Overview – Data Ethics

What are ethics?

Data ownership

Privacy

Anonymity

Data validity

Algorithmic fairness

Validity

As there are a lot of possible interactions between two variables, we need to make sure that **validity** is given.

Absence of Validity Leads to Data Error

Bad data and **bad models** lead to **bad decisions**.

If **decision-making is non-transparent**, results can be bad on an aggregated level, and **catastrophic for an individual**.

What if someone **is denied a loan** because of an **error in the analyzed data?** Or in the analysis method design?

Poor Data in Organizations

Wrong Business Strategy

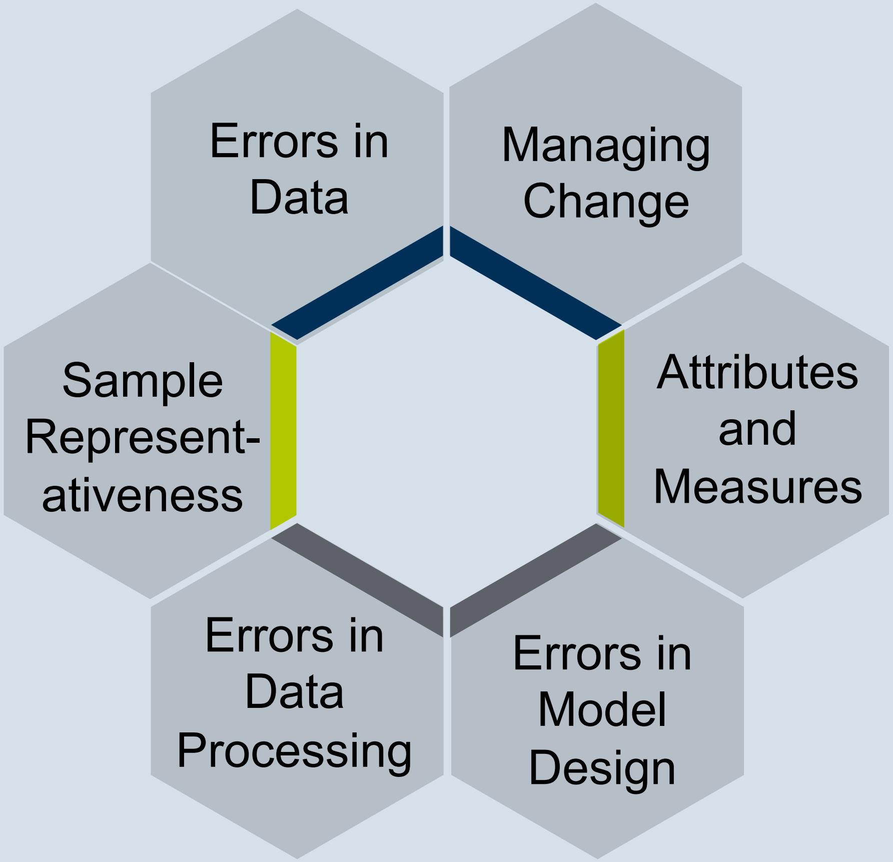
Damaged Reputation

Increase in Financial Costs

Missed Opportunities

https://commence.com/blog/2021/01/16/bad-data-in-decision-making-process/

Sources of Error



Sample Representativeness

The Streetlight Effect

Drunk people look for their keys under a lamppost, because this is where they can see.

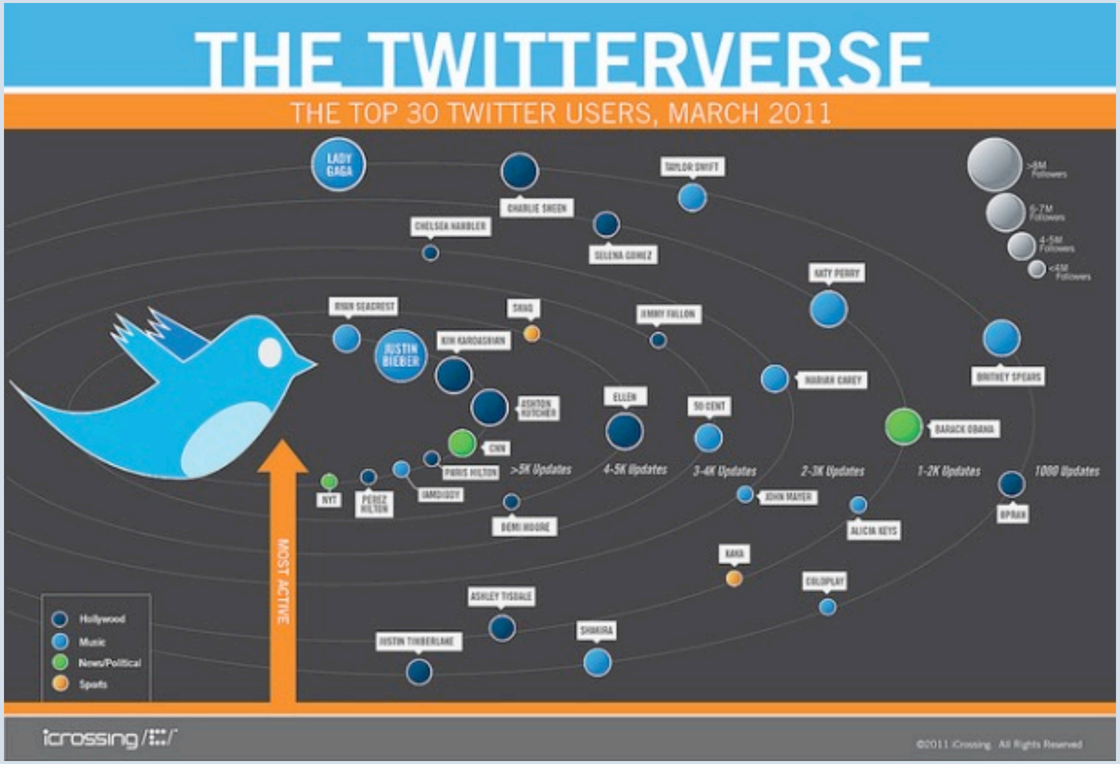
We are often limited by what data we have.

We just analyze what we have and hope for the best.

Sample Representativeness
Example: Twitterverse

Are twitter users **representative of the population** as a whole?

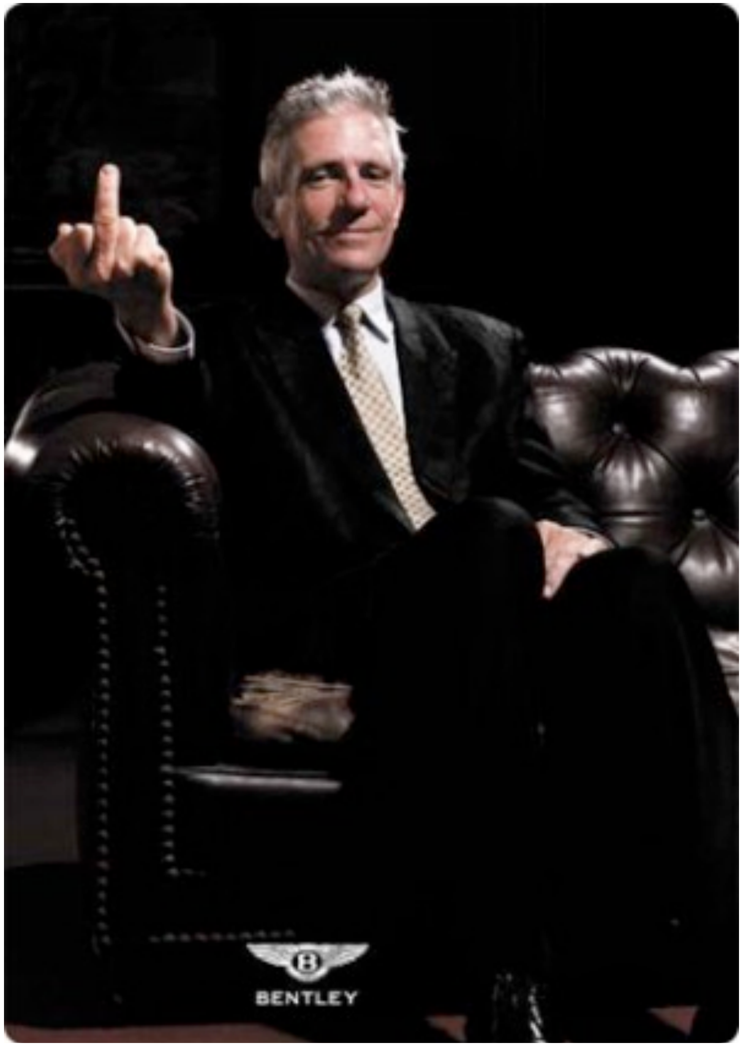
Are tweets **representative of the opinions** of twitter users?



Sample Representativeness: Opinionated Customers on Forums

Sometimes, it **may not matter** whether the **opinion is representative** of the population.

It may be **enough if it is representative of a segment** of the population.



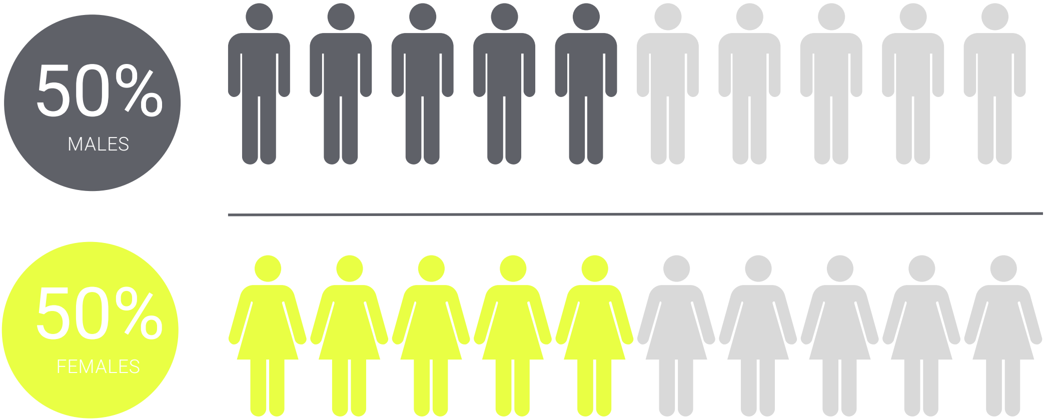
Sample Representativeness: Counting Variables

> “Not everything that can be counted counts,
> and not everything that counts can be counted.”

William Cameron, 1963

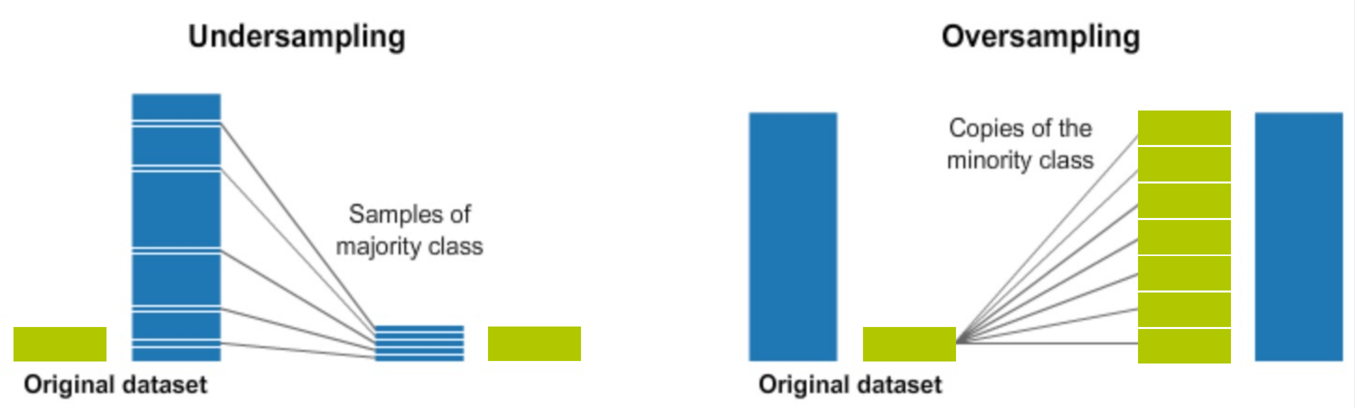
Sample Representativeness: Balance Important Attributes

If a variable (e.g., race, gender, age) is likely to matter, you need to make sure **the sample is well balanced** in these **attributes**.



Sample Representativeness: Accuracy Paradox as a Problem of Resampling

When working with data, you need samples of the **same** size, which you can achieve by **resampling**

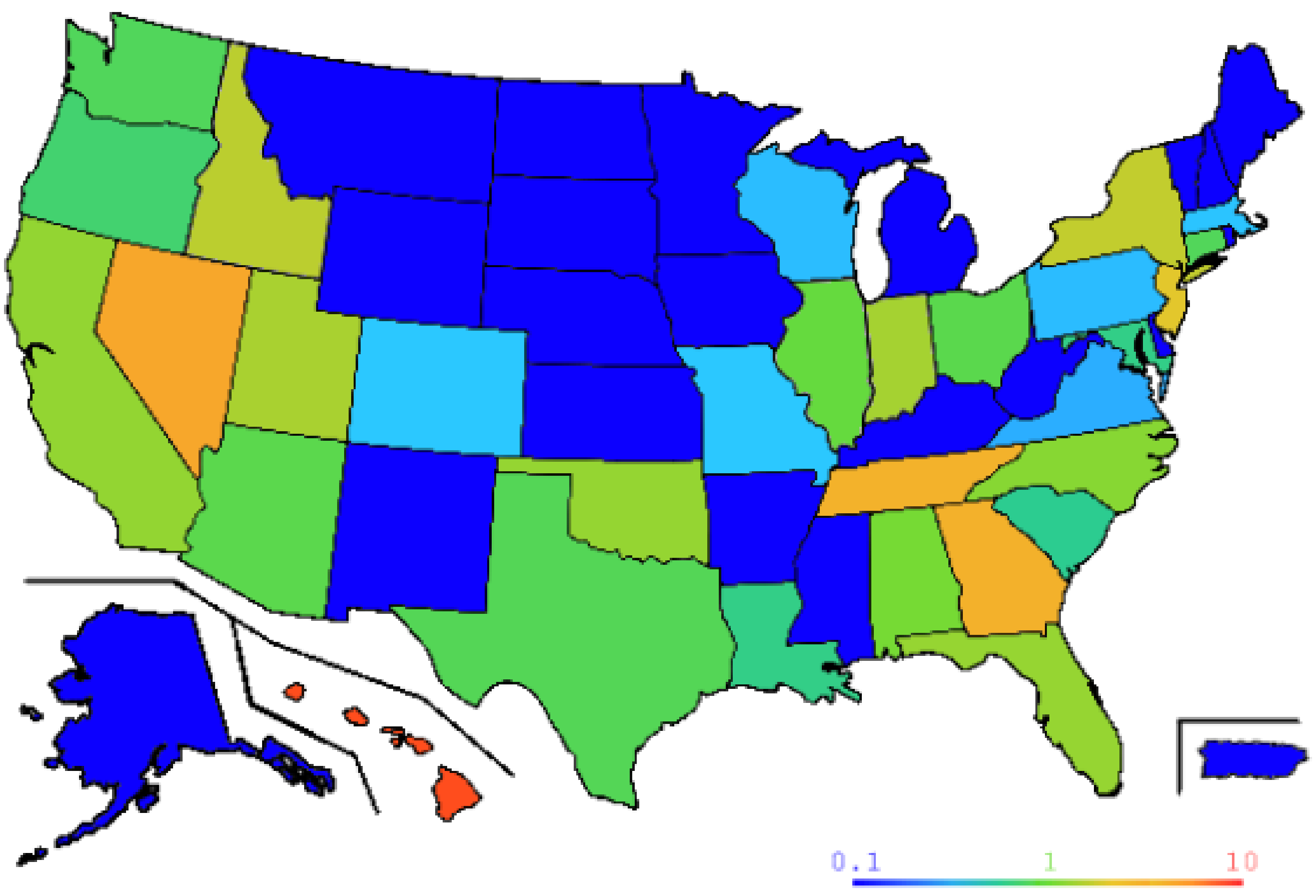


But this can lead to a **lack of accuracy** because it does **not clearly distinguish between the numbers of correctly classified examples**

https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/

Sample Representativeness: American Idol Semi-finalists

Overrepresentation of semifinalists by state, seasons 1-4



Sample Representativeness: Project Future Population

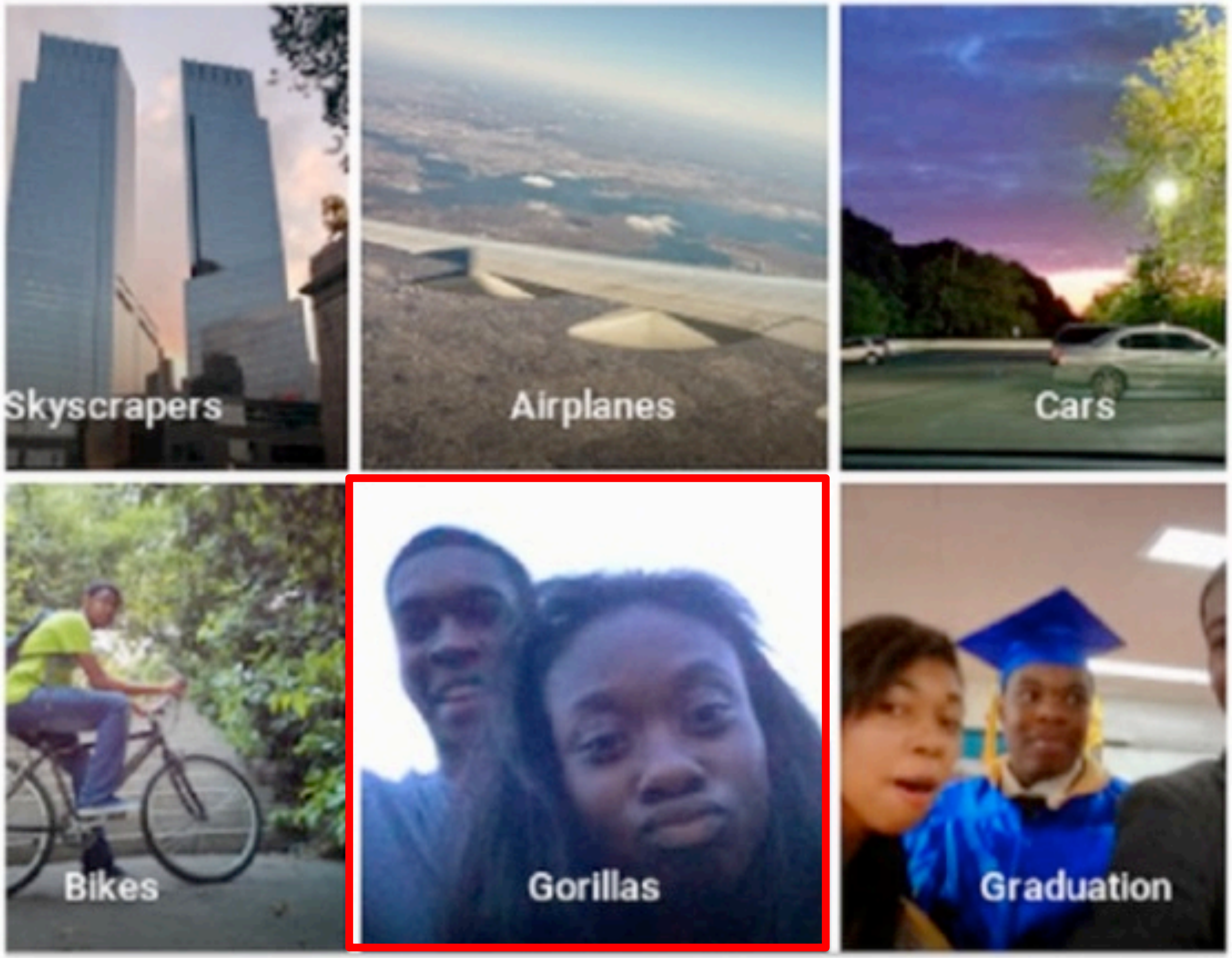
Past population is not the same as the future population.

Analysis based on the past will work in the future only to the extent to which the future resembles the past.

Watch out for singularities, but also worry about gradual drift.

Errors in Data

Example: Google Labeling Error



Errors in Data: Algorithms and Race

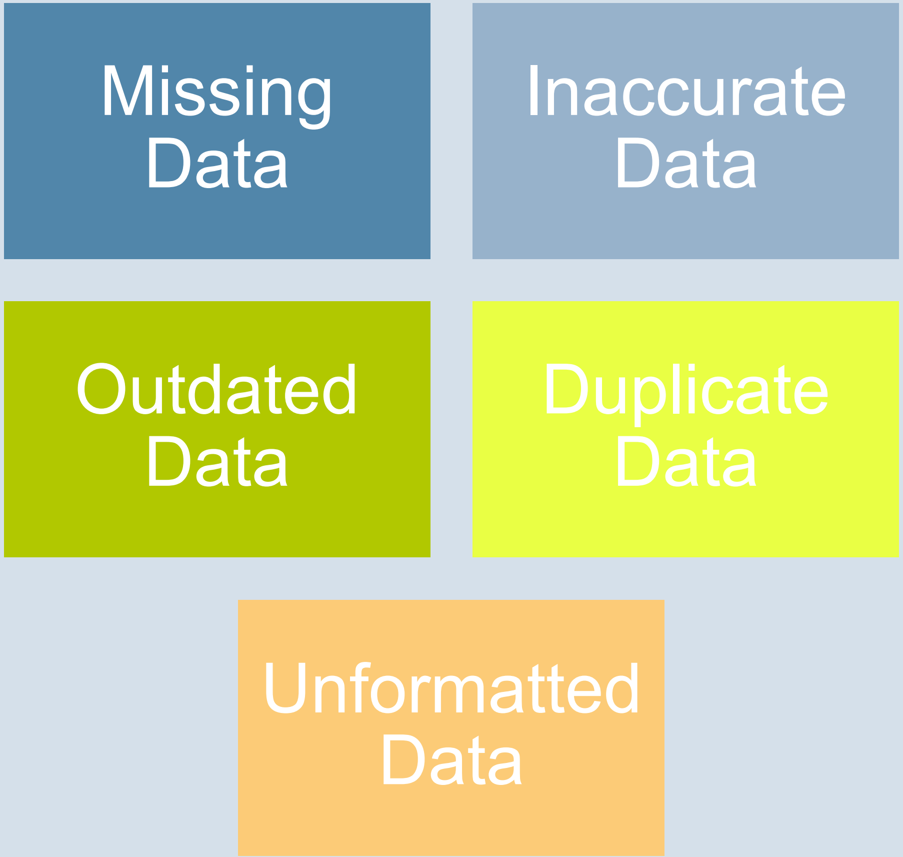
CAN COMPUTERS BE RACIST?

Big data, the internet, and the law

FORDFOUNDATION

@fordfoundation

Errors in Data



Missing Data

Inaccurate Data

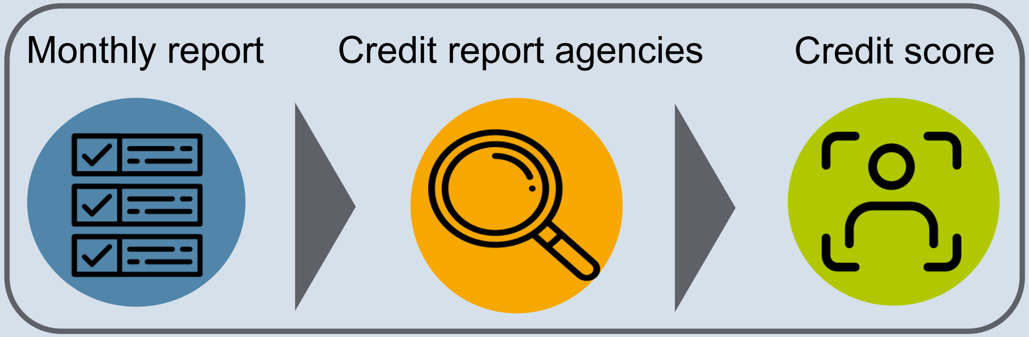
Outdated Data

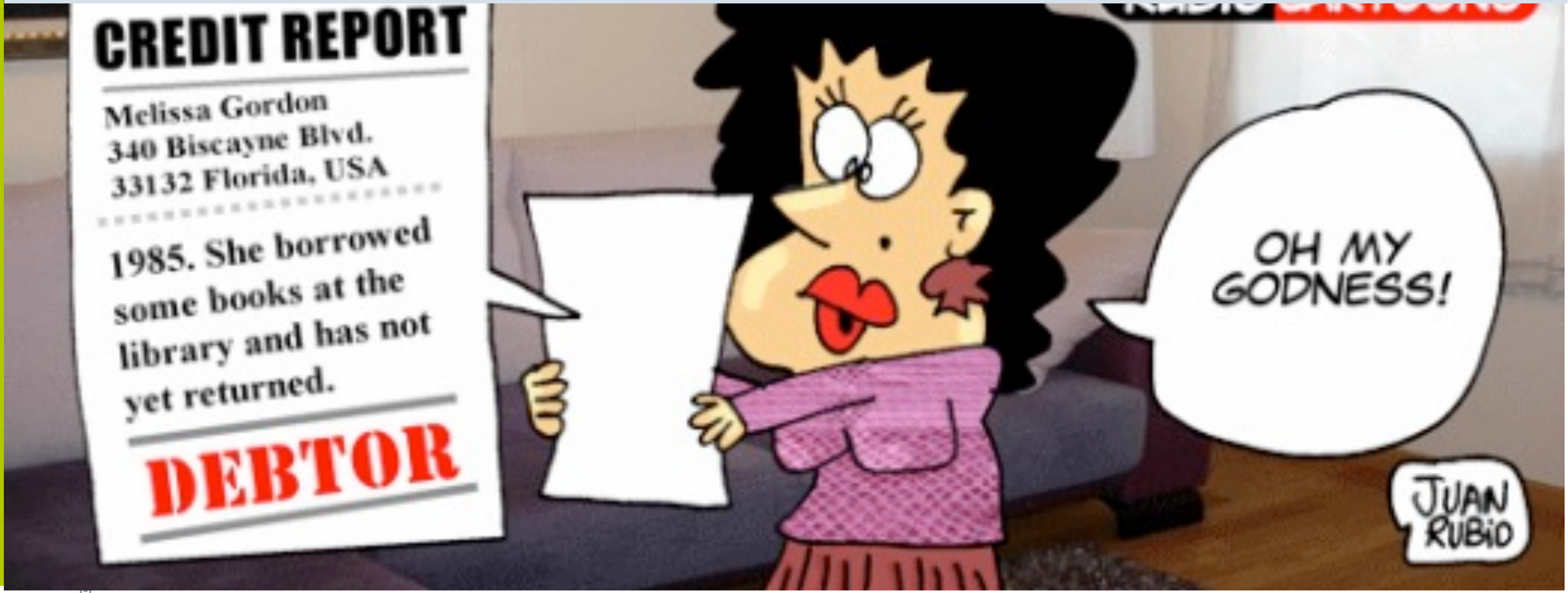
Duplicate Data

Unformatted Data

https://commence.com/blog/2021/01/16/bad-data-in-decision-making-process/

Errors in Data
Example: Credit Reports


Monthly report
Credit report agencies
Credit score



Attributes and Measures: What Attributes to Choose

Attributes decide on the research we can conduct

Usually **limited** by **what is available**.

Additional attributes can sometimes be **purchased** or **collected**.

Still, we need to think about **missing attributes.**

Attributes and Measures: What Attributes to Leave Out

May be **limited by law.** For example, in many cases, race can and should **not** be considered.

Attributes and Measures: Paid Ads Based on Followers

Kim Kardashian West has **70 million** followers on **Twitter**.

Company X paid her to **tweet about its products.**

- 50 million saw the tweet
- 2 million visited Company X's web site
- 30,000 orders ($30 each, on average)

→ **$900,000 in sales**

Are these the sales based on this tweet?

Attributes and Measures: Paid Ads Based on Followers

50 million saw the tweet.

At $0.003 per view = $150,000

Pay per new customer

Associated sales of $900,000.

At 10% profit margin = $90,000

Pay per view

2 million visited Company X’s website.

At $0.05 per new visitor = $100,000

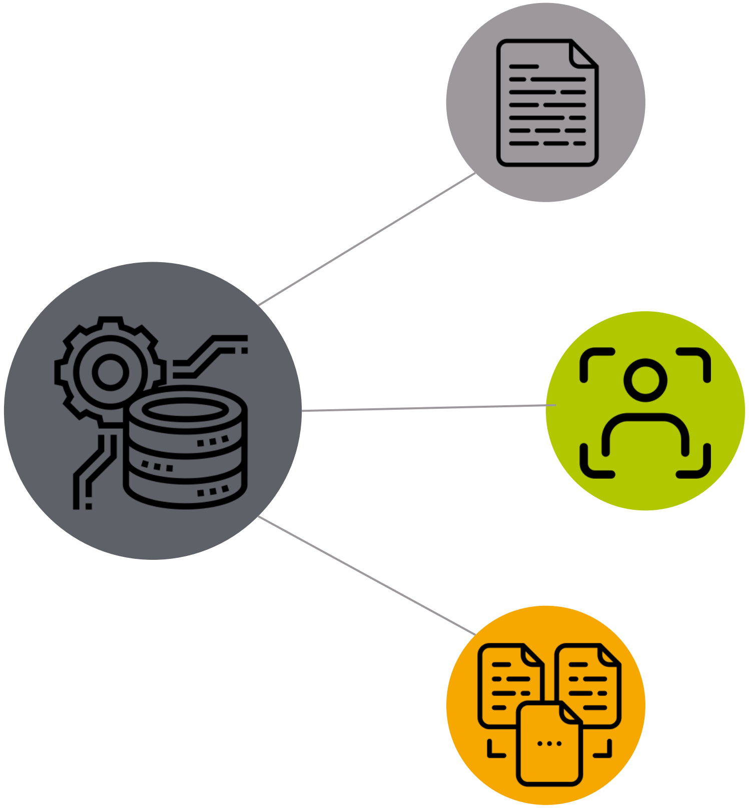
Pay profit margin

Errors in Data Processing: “Fancy Data Processing”

Extracting sentiment from text.

Recognizing faces from photos.

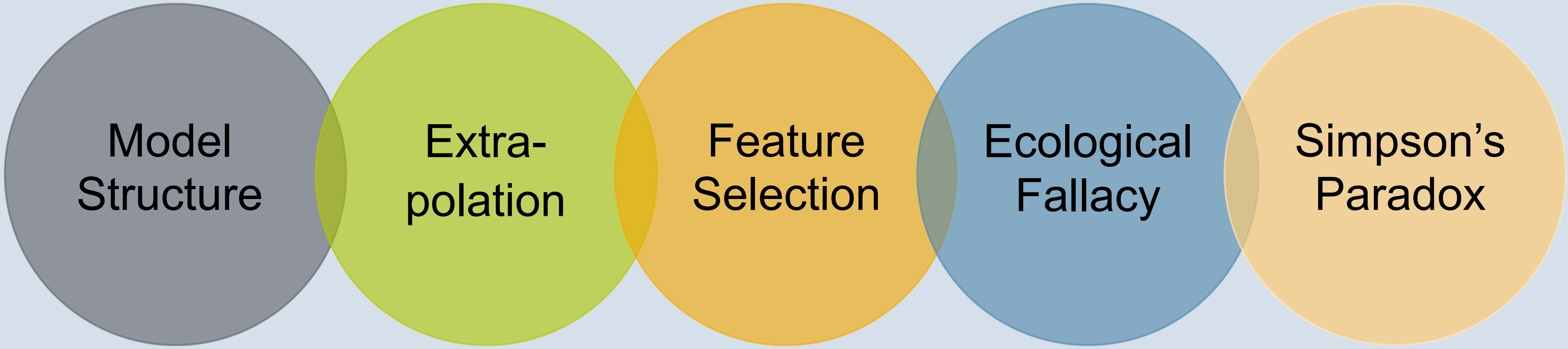
Merging records for the “same” person.



Combination of Errors: Algorithms on Social Media

https://assets.deutschlandfunk.de/FILE_40d4b5c766050371d5ea859d130de380/512x384.jpg?t=1597525089324
https://i.computer-bild.de/imgs/5/1/8/3/8/5/6/Facebook-1024x576-e343b150b7305a53.jpg
https://news.stanford.edu/wp-content/uploads/2021/10/GettyImages-1298158450-555x370.jpg

Errors in Model Design: Different Cases



Model Structure

Extra-polation

Feature Selection

Ecological Fallacy

Simpson’s Paradox

Errors in Model Design: Model Structure

Most machine learning just **estimates parameters** to fit a **pre-determined model**.

Do you know the model is appropriate?

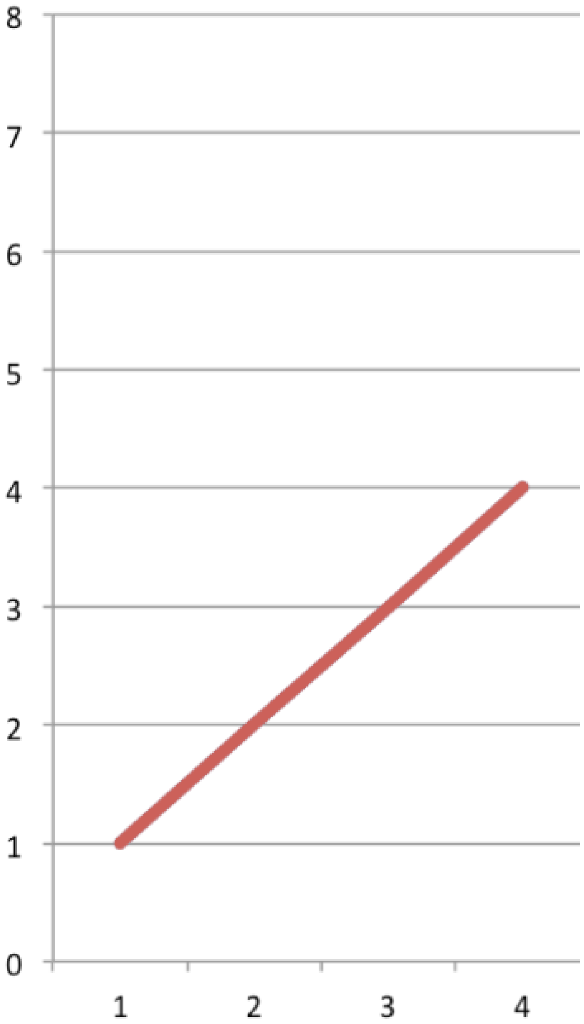
Are you trying to fit a linear model to a complex nonlinear reality?

Errors in Model Design: Extrapolation

We have a perfect linear graph in the range 1-4.

What can you say about X=7?

This is a simple linear structure. So, it needs to go on and on and on...



Errors in Model Design: Feature Selection

Did you know that **taller people** are more likely to **grow beards**?

Women generally are shorter. Women don’t grow beards. **This doesn’t tell us anything about taller vs shorter men!**

Errors in Model Design: Ecological Fallacy

Analyzing results for a group and ascribing results to the individual.

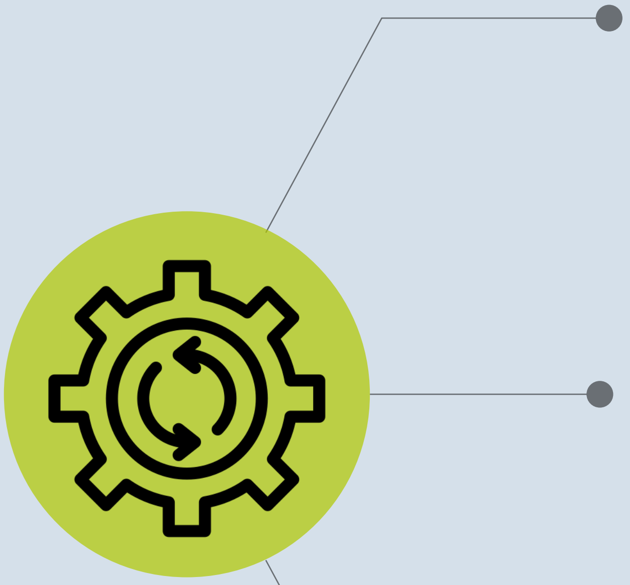
Example: District with high income + Very low crime rate ≠ A certain wealthy individual is not a criminal

Errors in Model Design: Simpson’s Paradox

Women are **accepted more often** by both Easy U and Hard U. But they are **accepted less** often by the **two combined**. Because **more women than men apply to Hard U**.

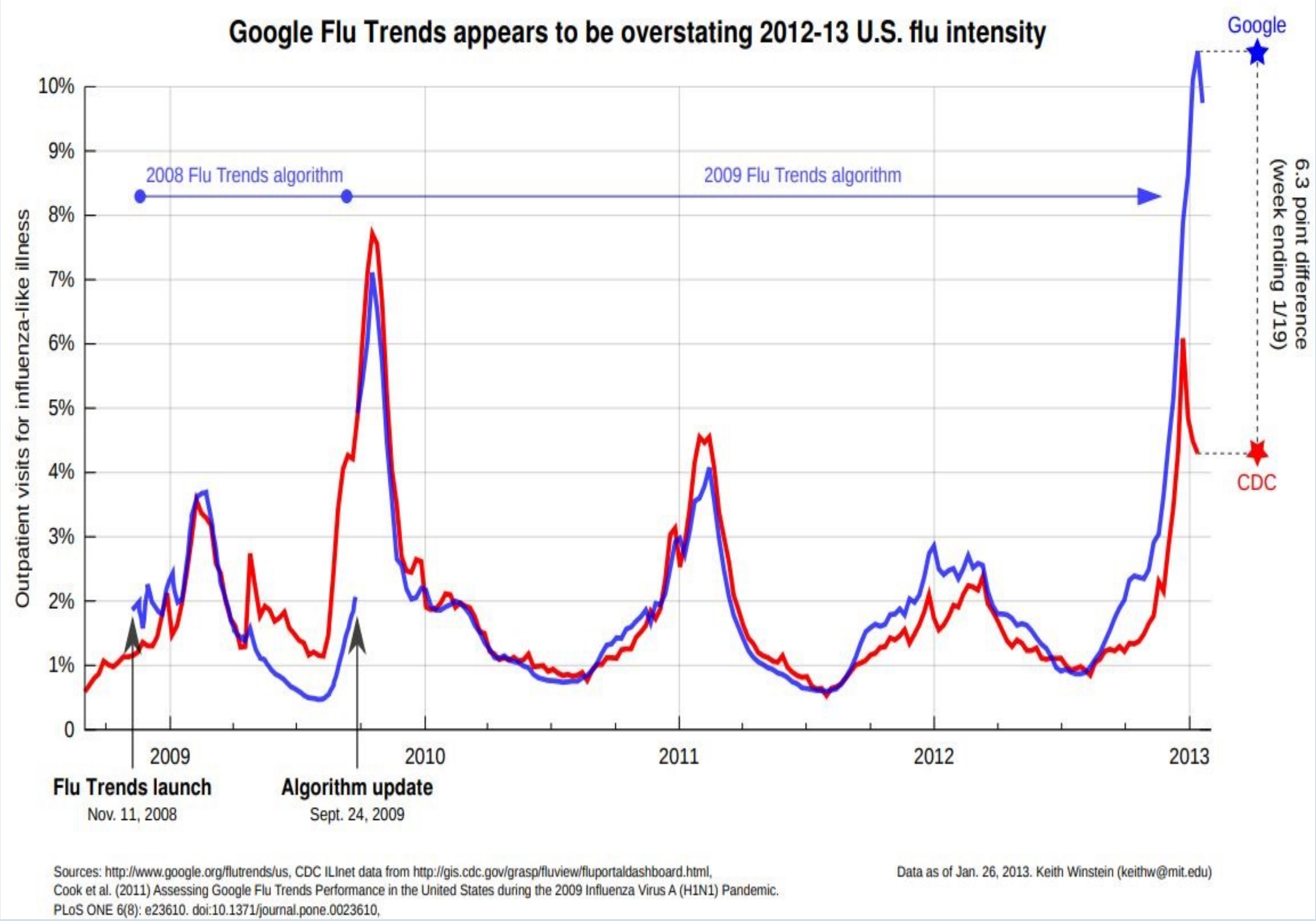
	Men	Women
Easy	7/10 = 0.7	4/5 = 0.8
Hard	3/10 = 0.3	5/15 = 0.33
All	10/20 = 0.5	9/20 = 0.45

Managing Change: Analysis of Complex System



- System **changes continuously**.
- Is the analysis **still valid** then? Most changes do not impact the analysis.
- But **some do**, and we may **not know which ones**!

Managing Change
Example: Google Flu Trends



Google Flu Trends appears to be overstating 2012-13 U.S. flu intensity

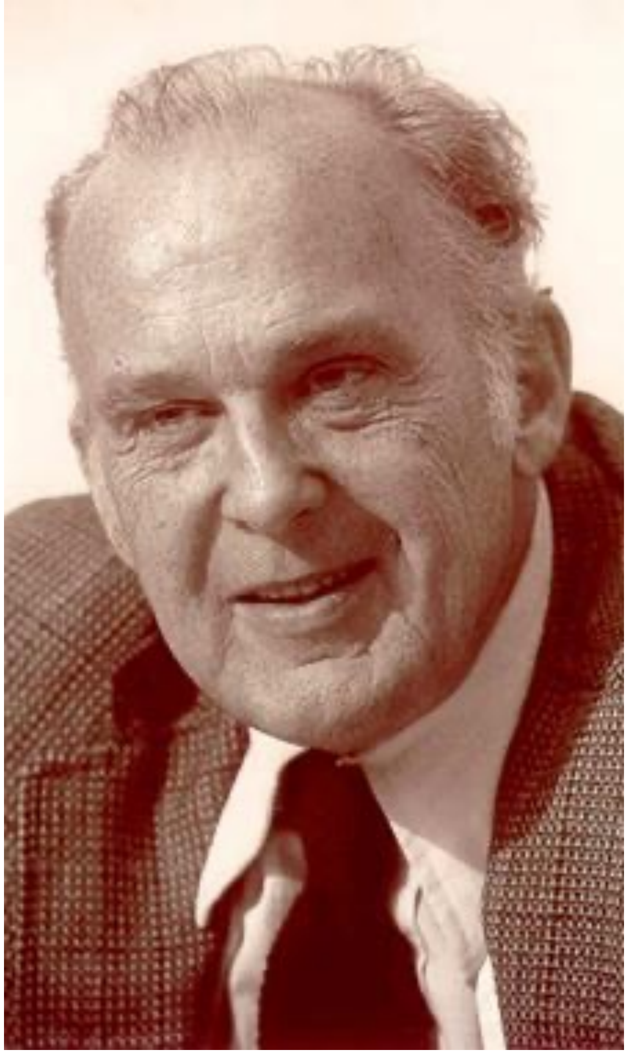
Sources: http://www.google.org/flutrends/us, CDC ILInet data from http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html, Cook et al. (2011) Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. PLoS ONE 6(8): e23610. doi:10.1371/journal.pone.0023610,

Data as of Jan. 26, 2013. Keith Winstein (keithw@mit.edu)

Managing Change: Campbell's Law

> The more any quantitative social indicator (or even some qualitative indicator) is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.

Donald Campbell, 1979



Managing Change: Campbell's Law
Example: Crime Rate

Assume there is a **decrease in a city's crime rate** (= social quant. indicator).

People likely attribute this to a **reduction in the actual number of crimes**.

However, it may reflect a **change in how the crime rate is recorded** or which police encounters are **classified as crimes**

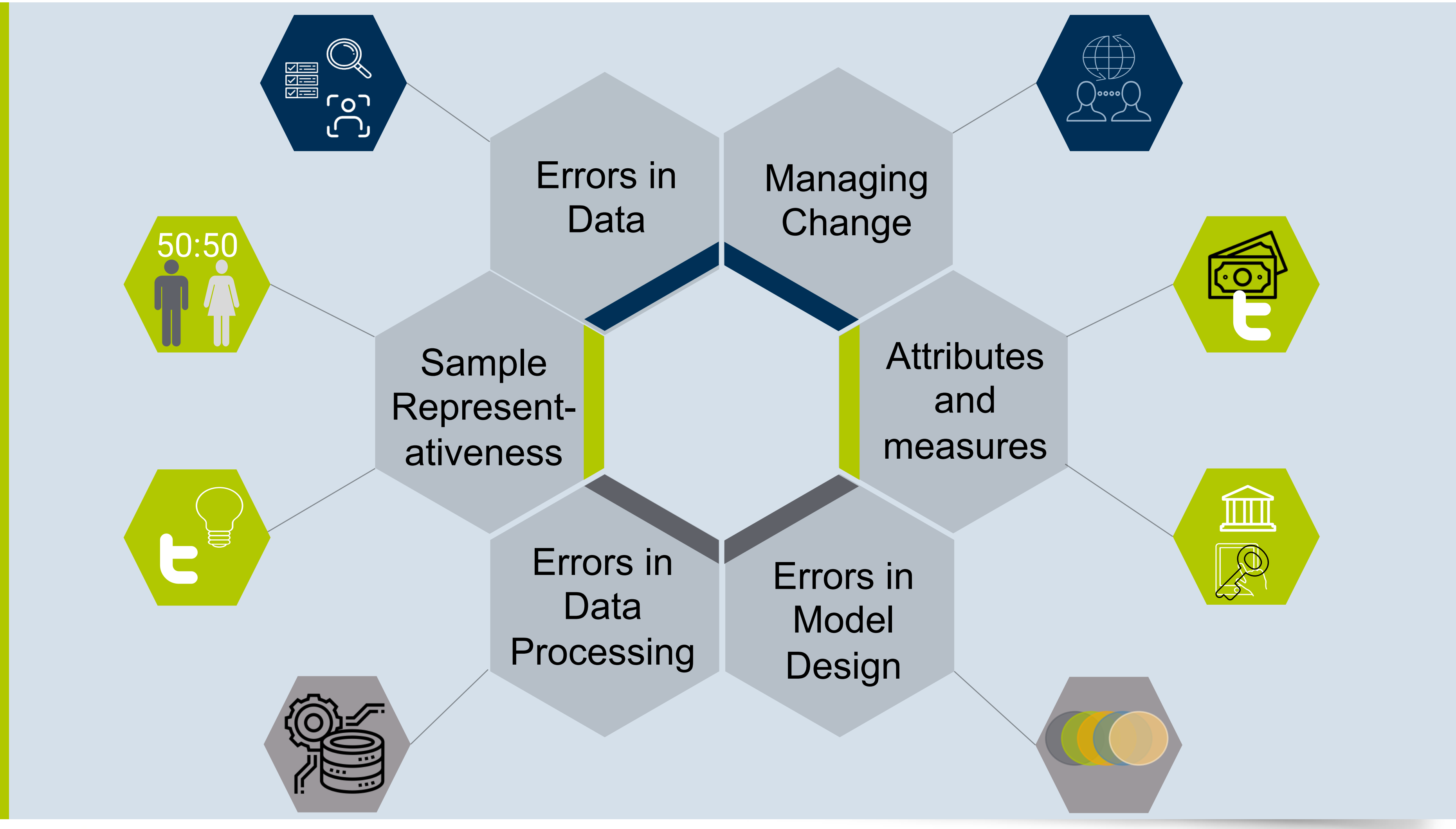
https://searchbusinessanalytics.techtarget.com/definition/Campbells-Law

Picture: https://3er1viui9wo30pkxh1v2nh4w-wpengine.netdna-ssl.com/wp-content/uploads/sites/68/2017/09/GettyImages-93192527-1600x1067.jpg

Managing Change: Campbell's Law
Example: The Facebook Case

Metric Obsession Weakens UX: The Facebook Case

Sources of Error



Conclusion

It is crucial that we pay careful attention to the validity of our data, and of the model.

Otherwise, we will get bad results.

Which can cause real harm.