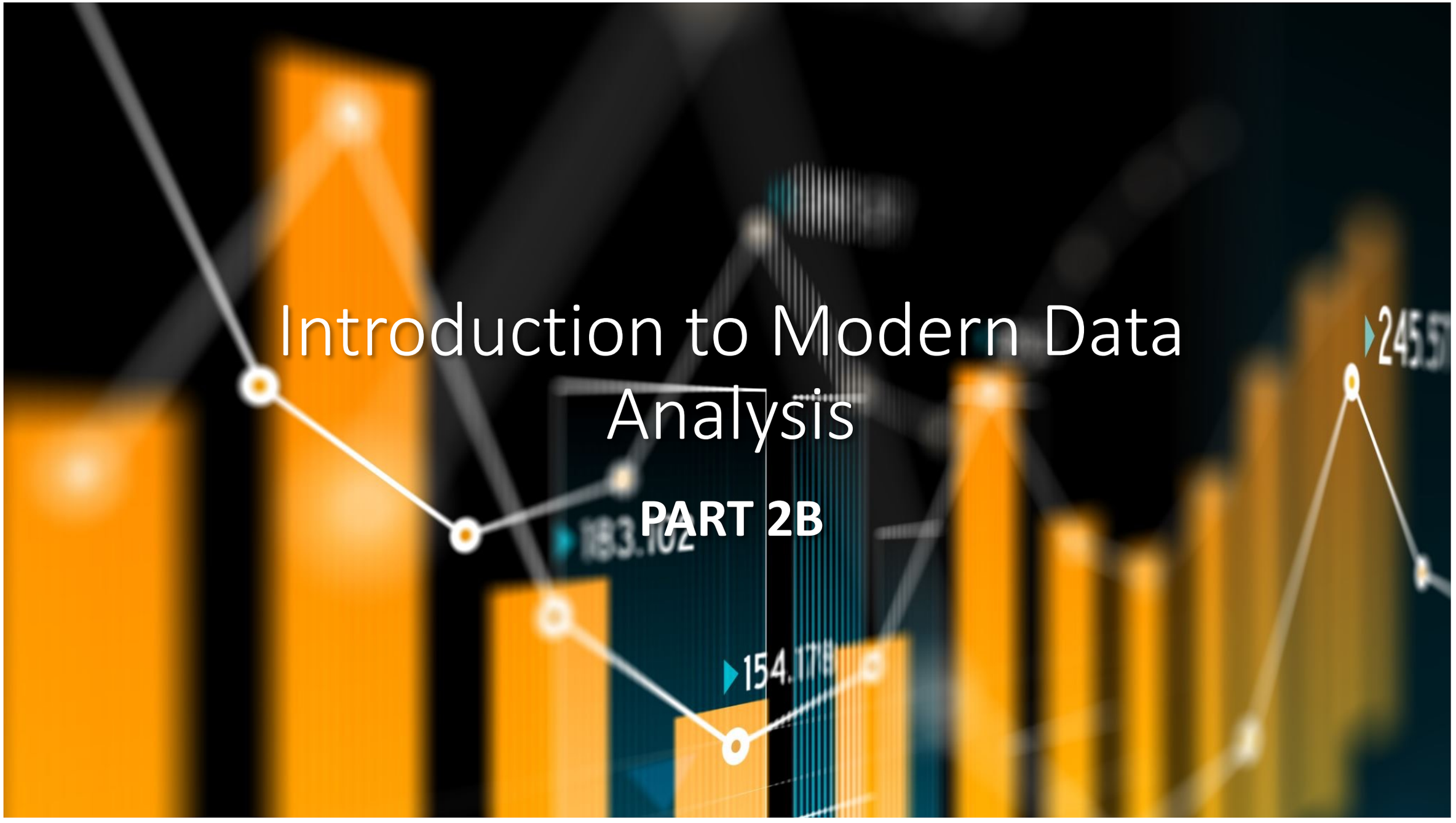


Introduction to Modern Data Analysis

PART 2B

Analysis

Outline For Analysis

Machine Learning vs Statistics vs Business Intelligence vs …

Business Intelligence

- Data Analytics
- Comparison
- Relationships

Statistics

- A tools discussion
- Modern Statistics – Controversies and Conversations
- Some Relevant Statistical Concepts and Techniques

Machine Learning/AI

- A quick tools discussion
- Relevant Techniques Overview
  - Supervised, Unsupervised, Reinforcement
- Text Mining

BI vs ML/AI vs STATISTICS/DS vs OTHER

Business Intelligence

- Idea has been around for a while but term was popularized by Dresden (1989). Think data warehouses + data reports.
- Uses whatever tools and techniques come in handy to provide an understanding of (business) operations (past, present, future)

Artificial Intelligence/Machine Learning

- Research project that tries to create autonomous intelligent machines – that's the end goal.
- Machine learning is a type of artificial intelligence that originally focused on finding ways for machines gathering sensor data to learn from this sensor data

Statistics (Data Science?)

- The study and theory of using data to generate information and knowledge
- Typically a focus on inference from a sample of data to a population
- Data Science is maybe just applied statistics?

Other Analysis Techniques: simulations, network analysis, mathematical models

Parallel evolution of techniques across these disciplines!

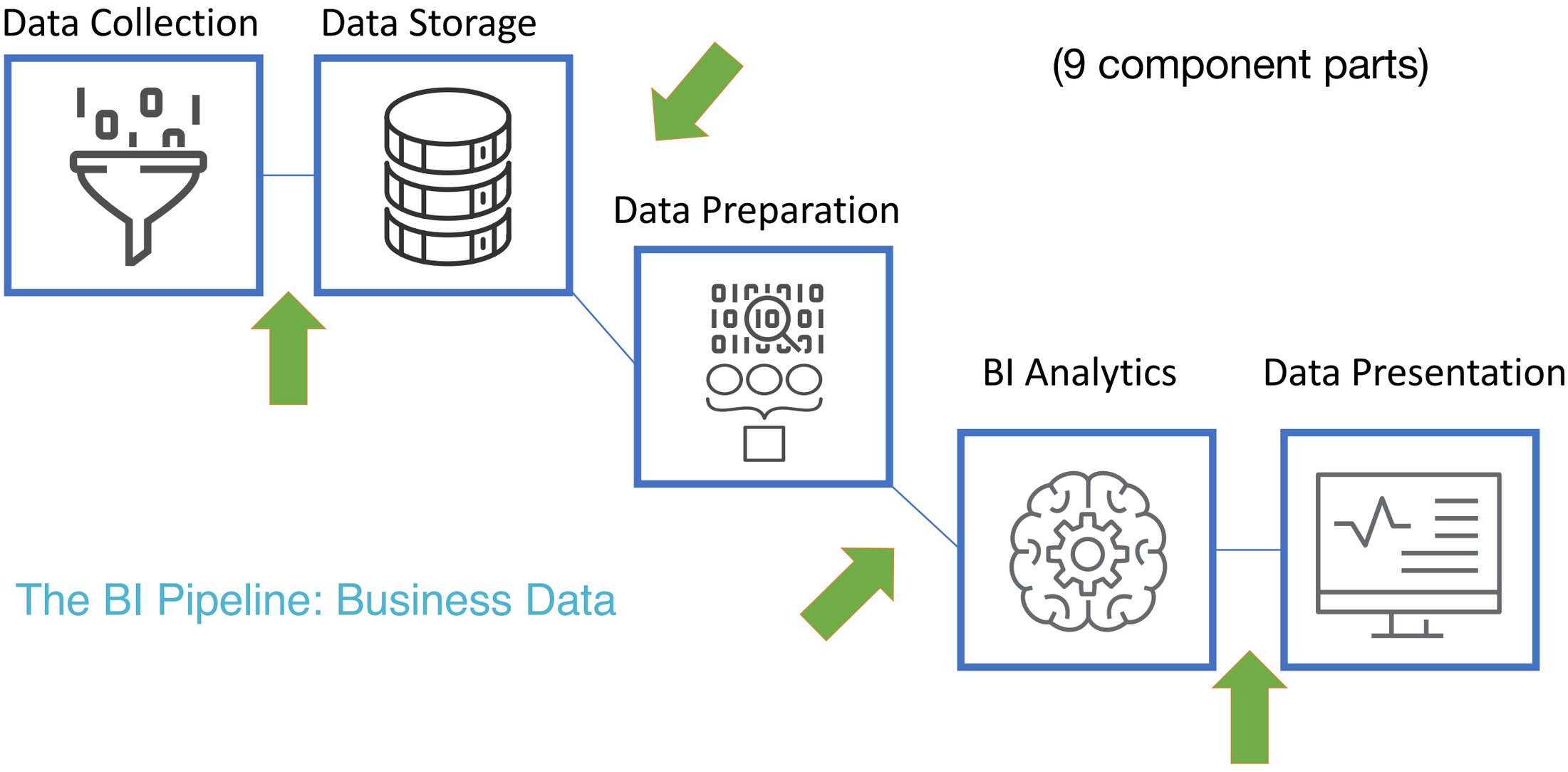
ML and Statistics – Different Approaches

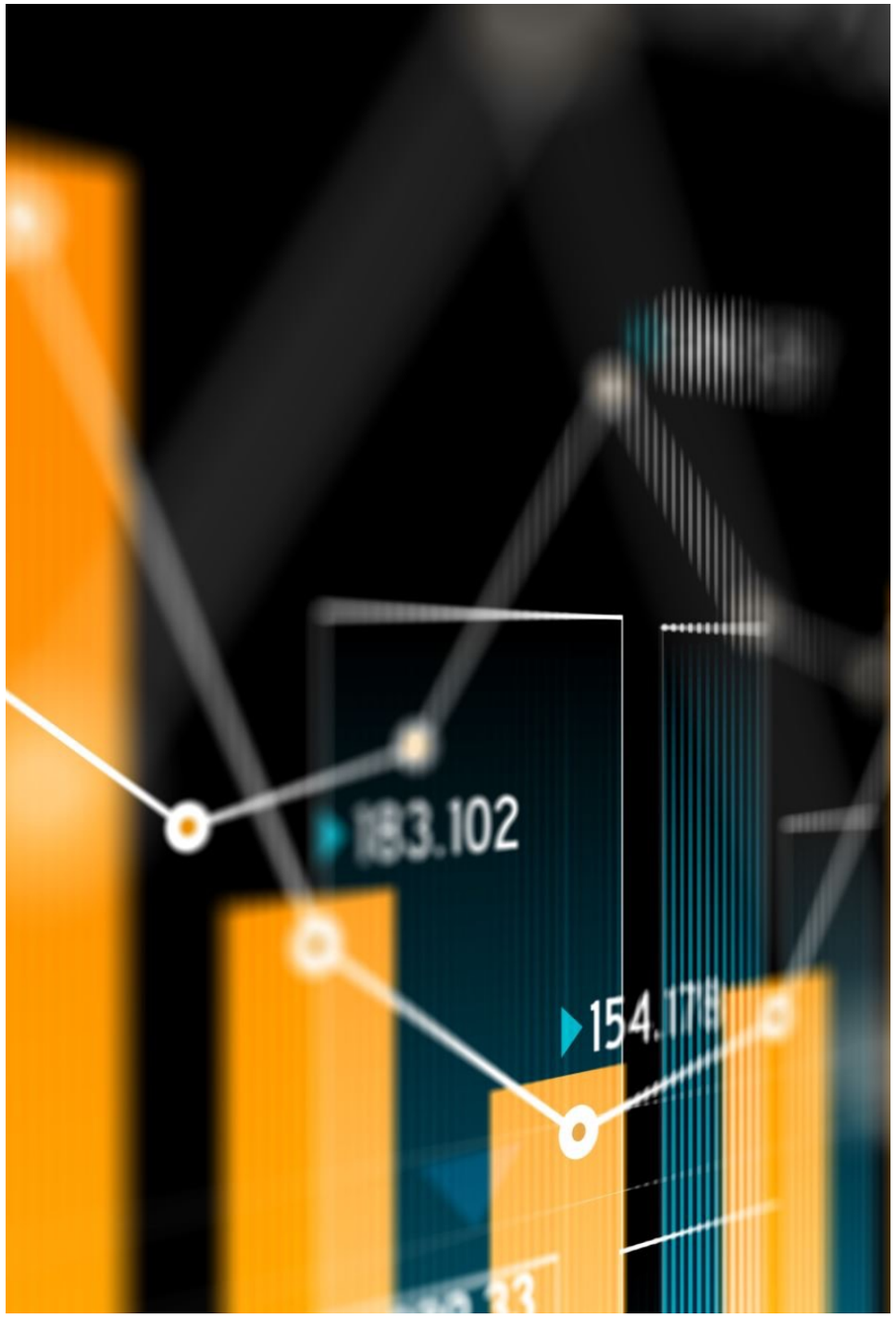
Practically speaking:

- **Machine learning**:
  - is about the output. For example, many ML techniques focus on prediction, so in these cases the output is a specific **prediction**.
  - is typically not explanation focused. The attitude is: If it works it works!
  - Underlying mechanisms of both the ML model and the system itself are irrelevant
- **Statistics**:
  - Is about understanding relationships and patterns in data
  - Isn't directly explanation focused, in terms of mechanics, but can shed light on connections and say "focus here"
  - Makes a serious attempt to be rigorous – wants to be a source of true information and knowledge, and quantify level of certainty

In reality, these days people typically combine both approaches.

Business/Organization Intelligence


Data Collection
Data Storage
(9 component parts)
Data Preparation
BI Analytics
Data Presentation
The BI Pipeline: Business Data



Data Analytics

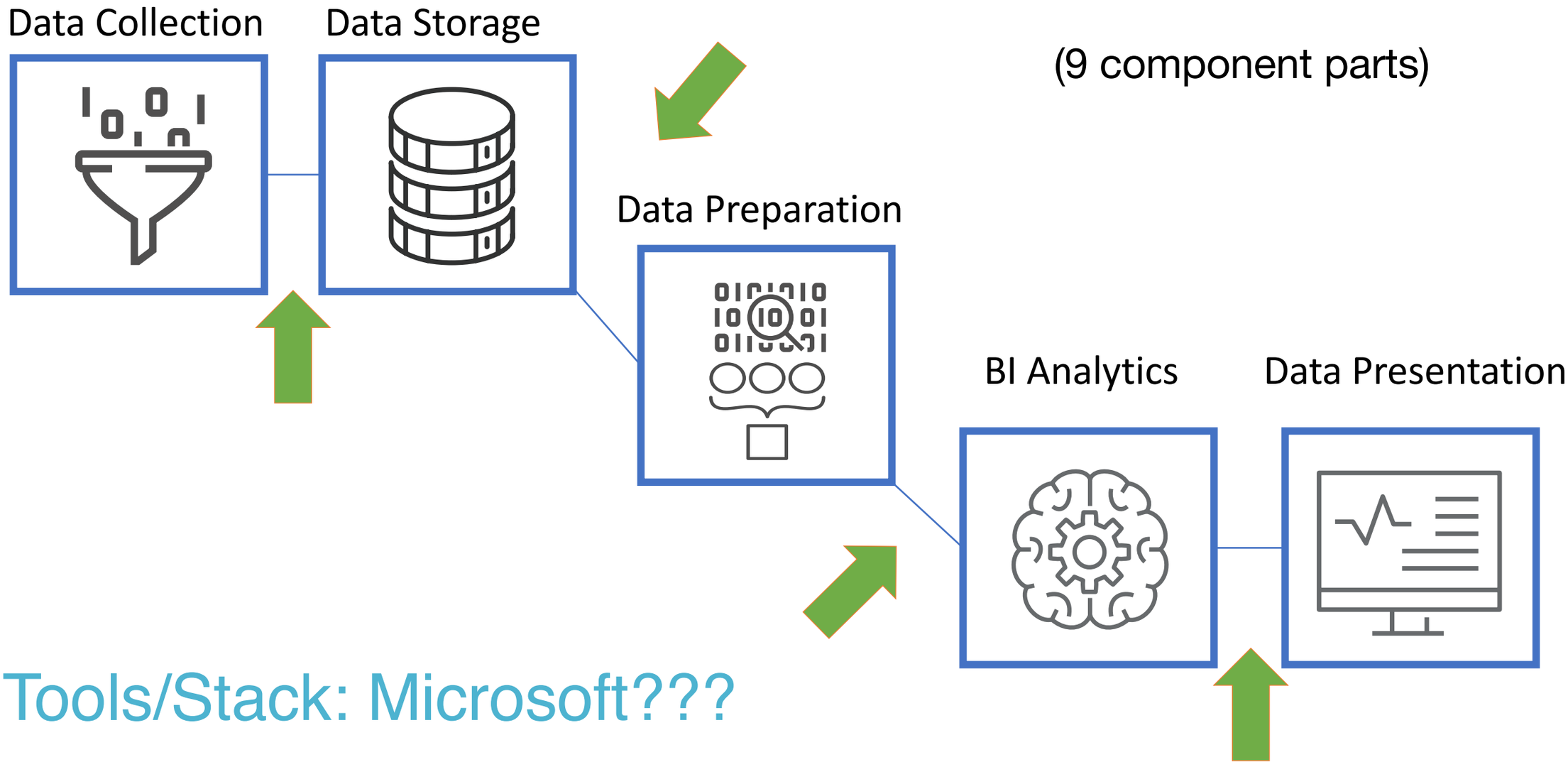
- **Data analytics** is sometimes used as an umbrella term for analysis in a *business intelligence context*.
- Importantly, this particular umbrella **includes analysis** focusing on:
  - Raw values – comparisons, part whole relationships
  - Summaries and roll ups
  - Measures and Metrics
- With BI the **process of inference** is often less formal or structured – often driven or supported by data visualization
- Caution required, **but not necessarily bad**, *when scope is kept in mind*
- Still evidence-based, data-driven!

Desktop Data Analysis

- Business Intelligence needs are pushing the development of **desktop data analysis** tools and pipelines:
  - PowerBI
  - Tableau
- Democratization of data + increase in data/digital literacy
- This is likely going to push organizations forward as well
- Not ***necessarily*** a substitute for ‘industrial’ or ‘professional’ data pipelines

BI Gateway to AI/ML

- To some extent getting a solid professional/industrial BI pipeline up and running is a major stepping stone in an organization
- **BUT** – the data architecture and tools you need for AI/ML/DS analysis may not be the same as those for BI
- You will MAY need to redesign some parts of your BI pipeline to support AI/ML/DS
- In particular – your database architecture: Data Lakes vs DataMart vs NoSQL



Tools/Stack: Microsoft???

Data Analytics

Analysis in a Business Context

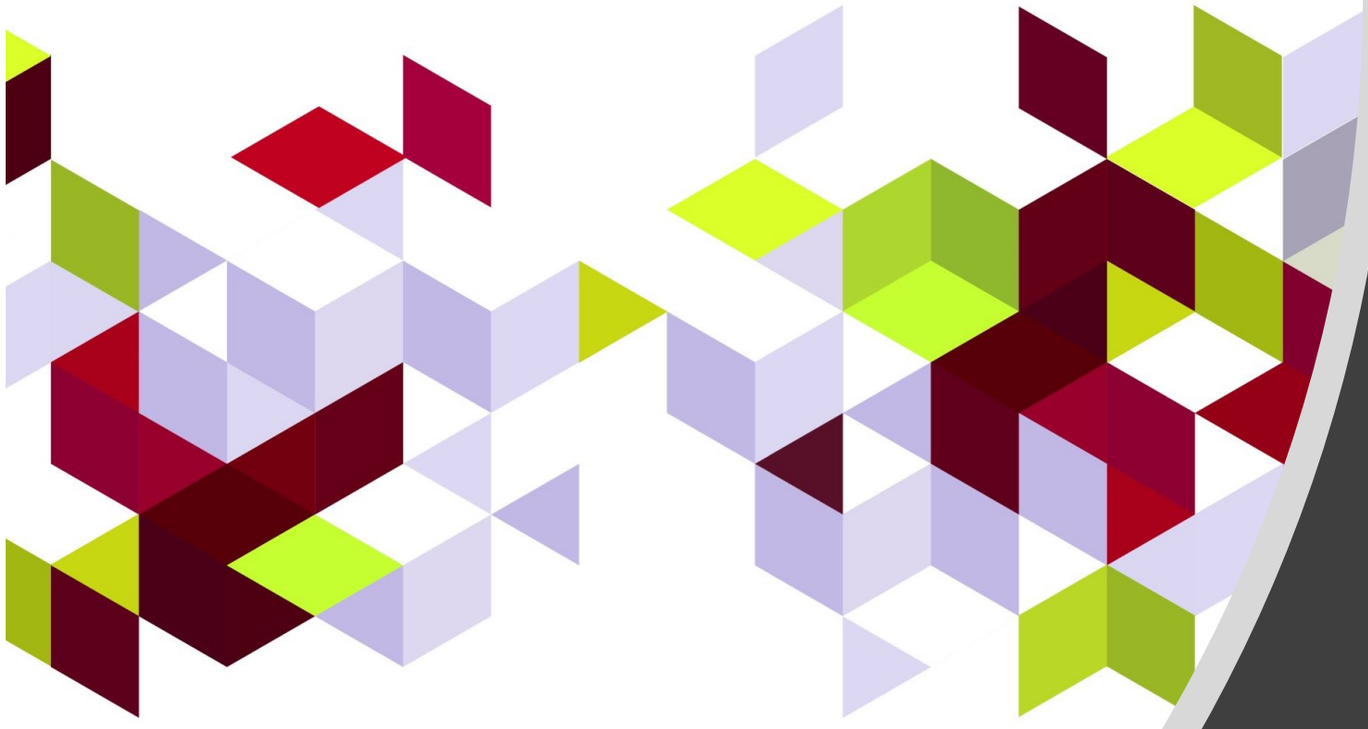
simple patterns

=

simple analysis

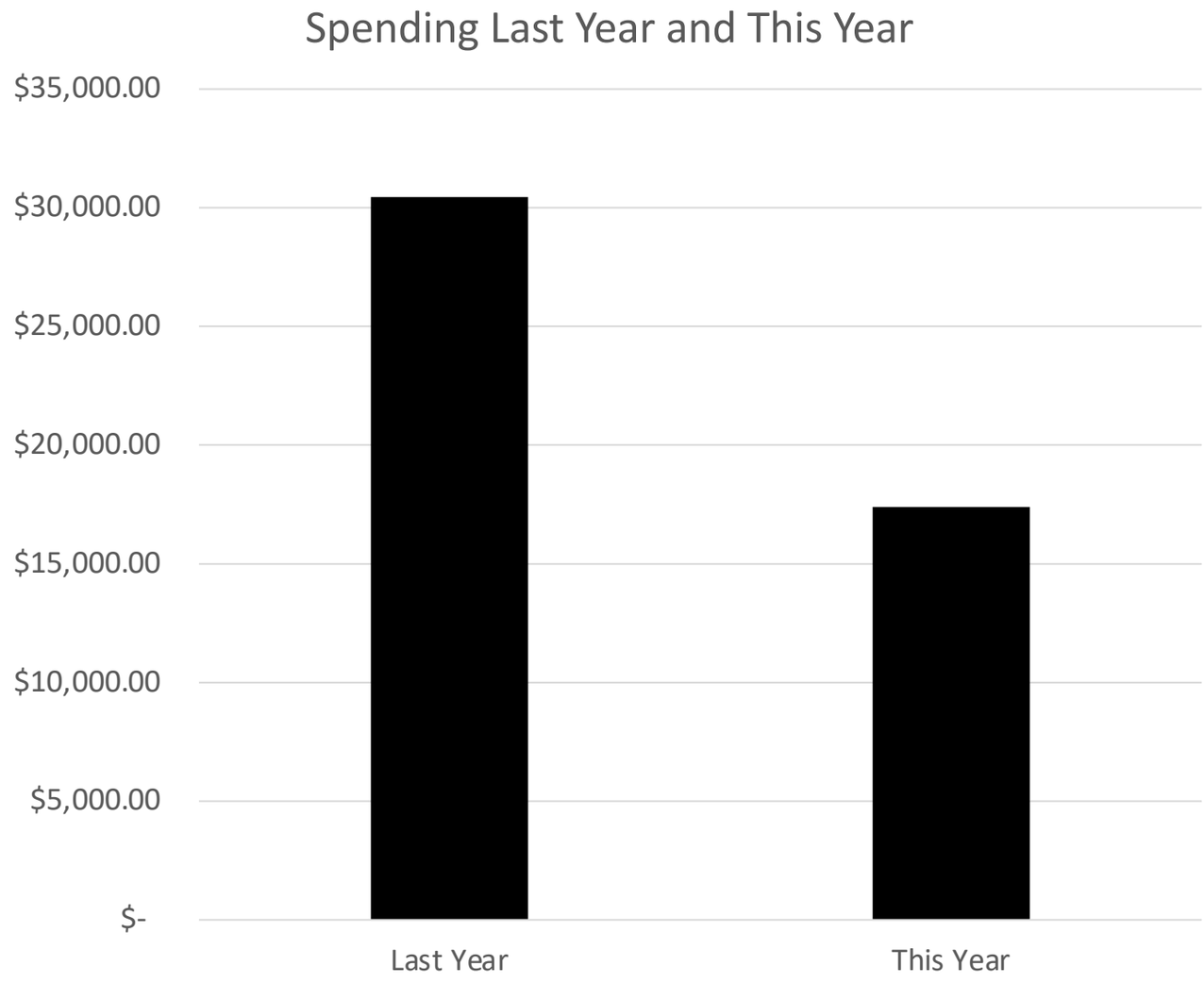
(and that's okay!)

Getting every last drop of value from your data!

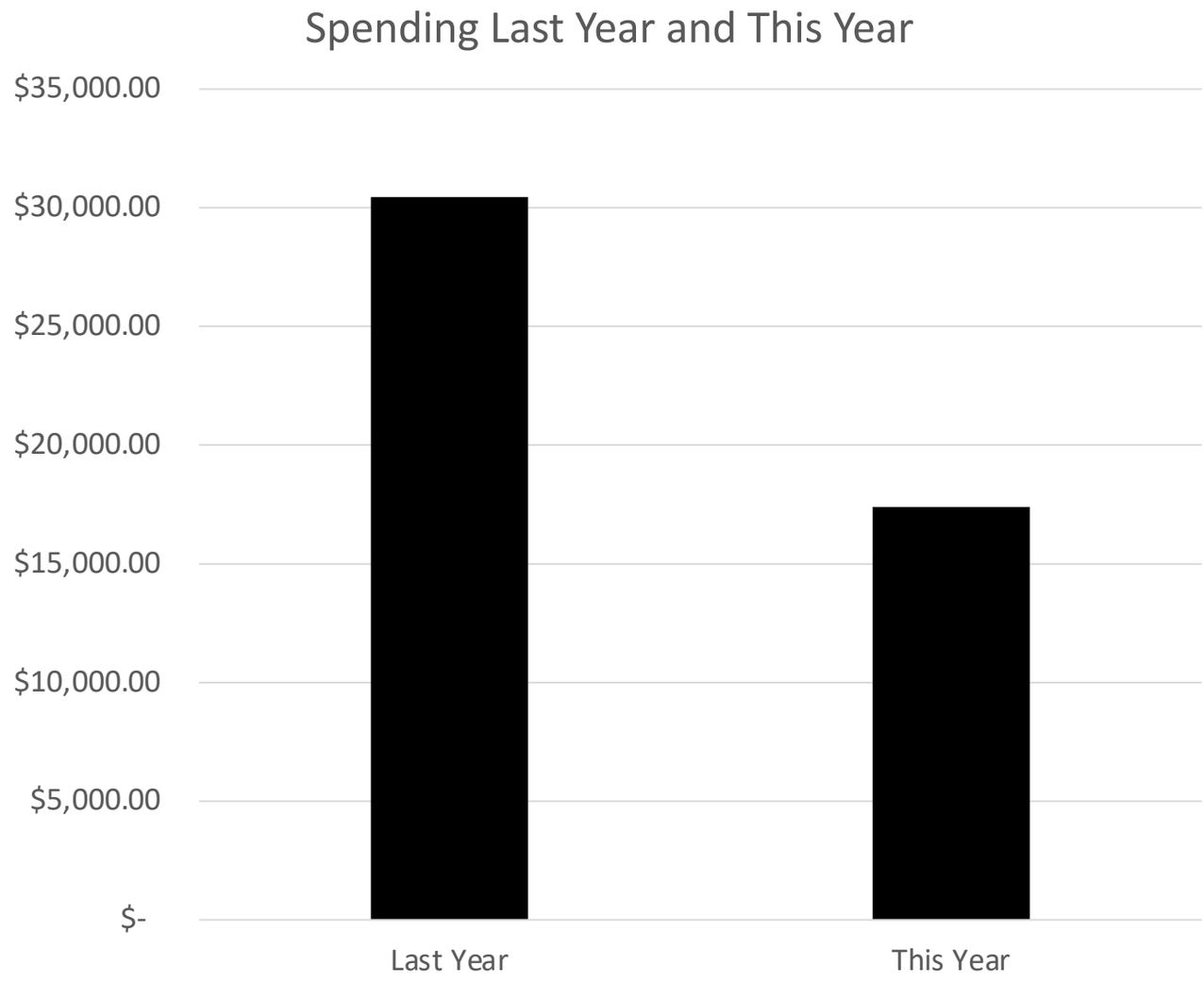


Analysis: Comparison

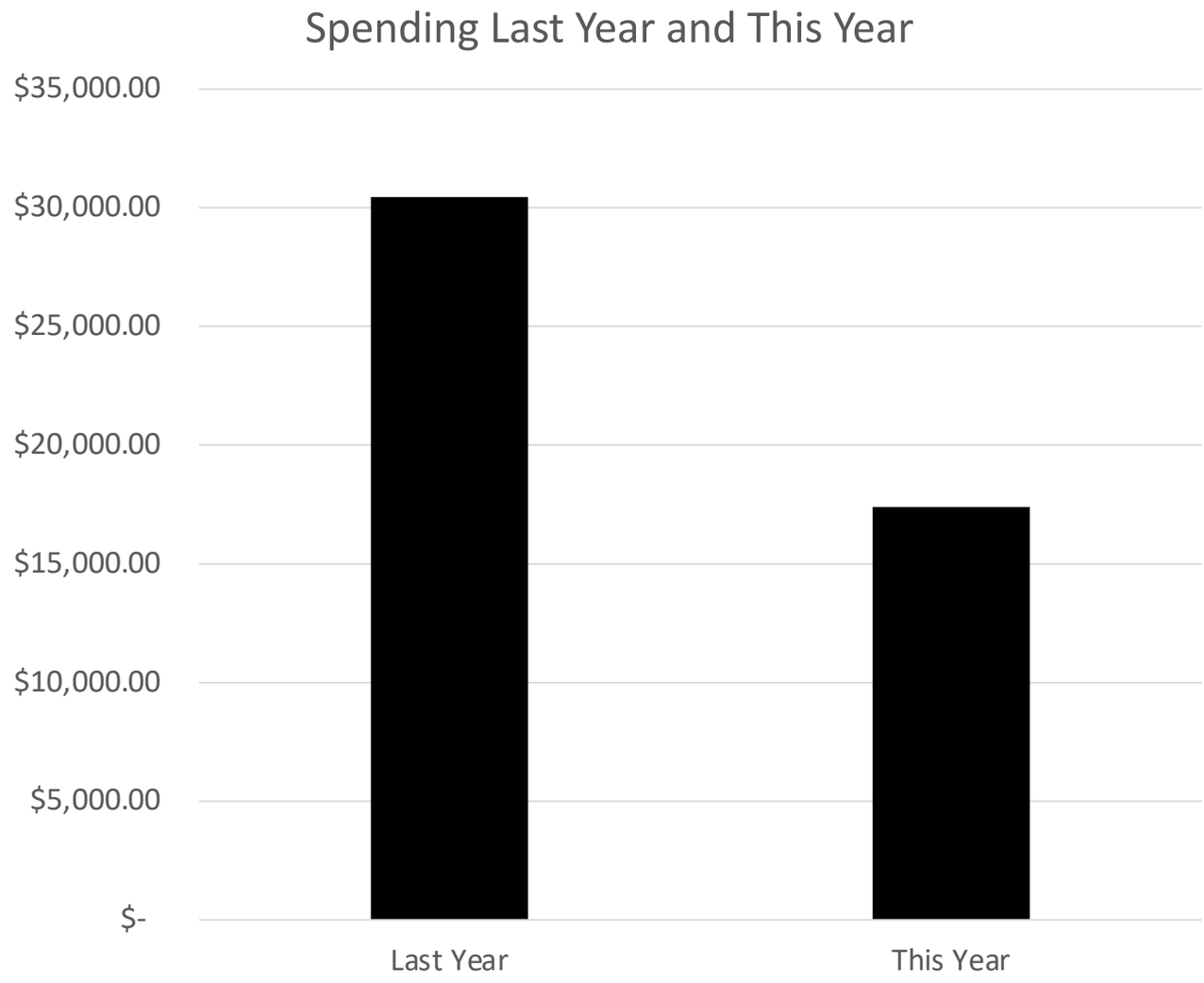
What can we say?



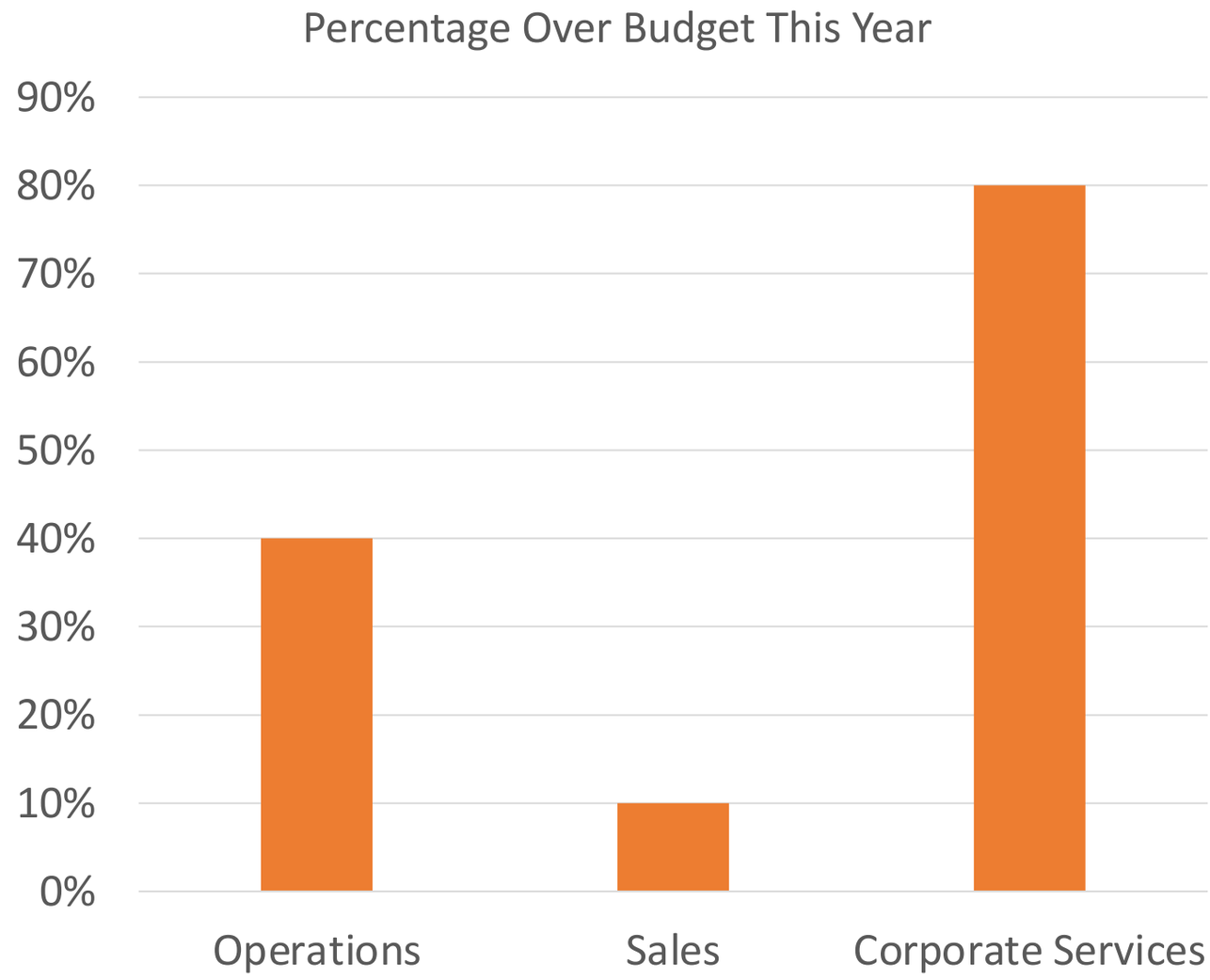
What are some hypotheses?

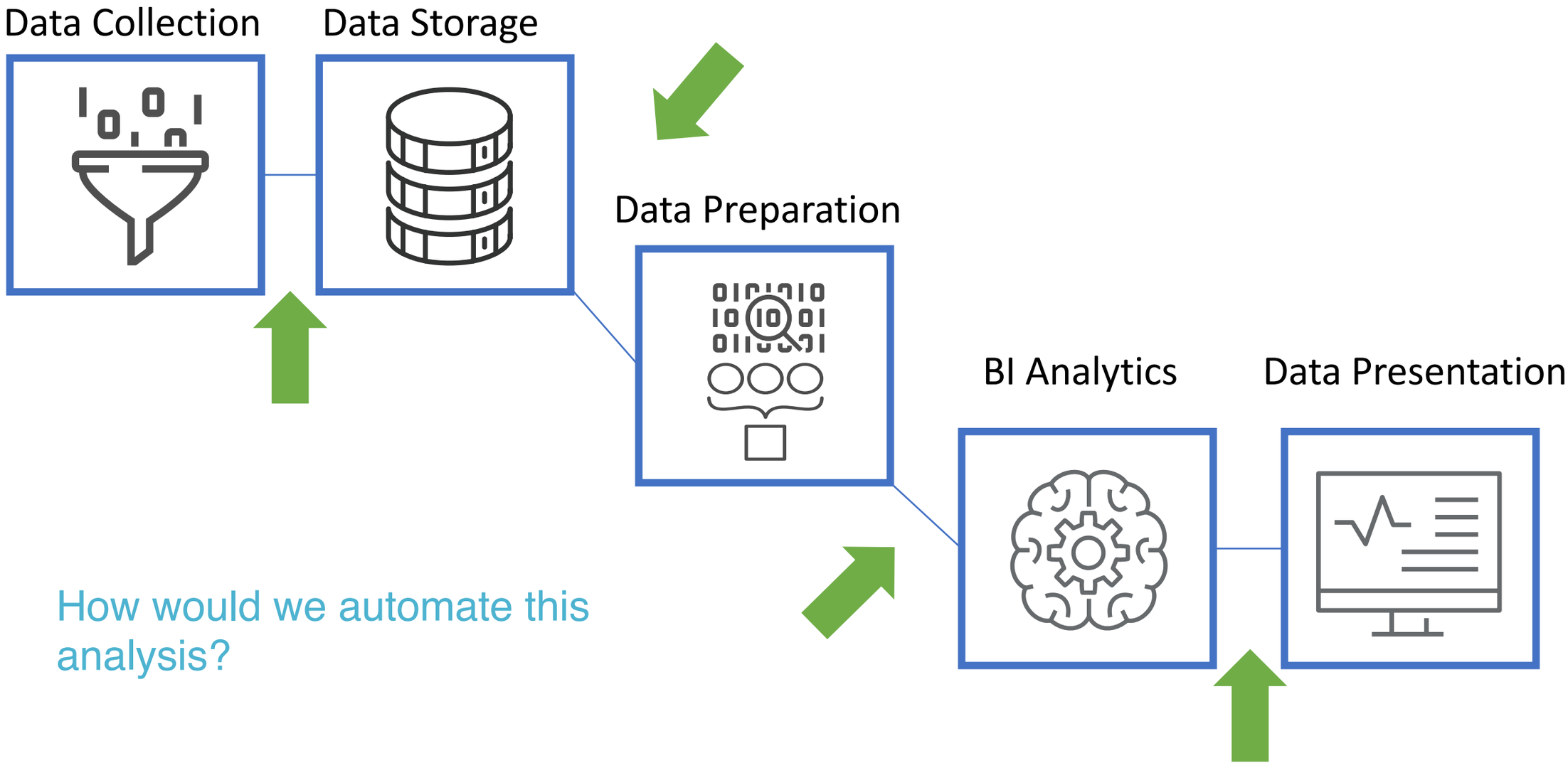


What can we infer?

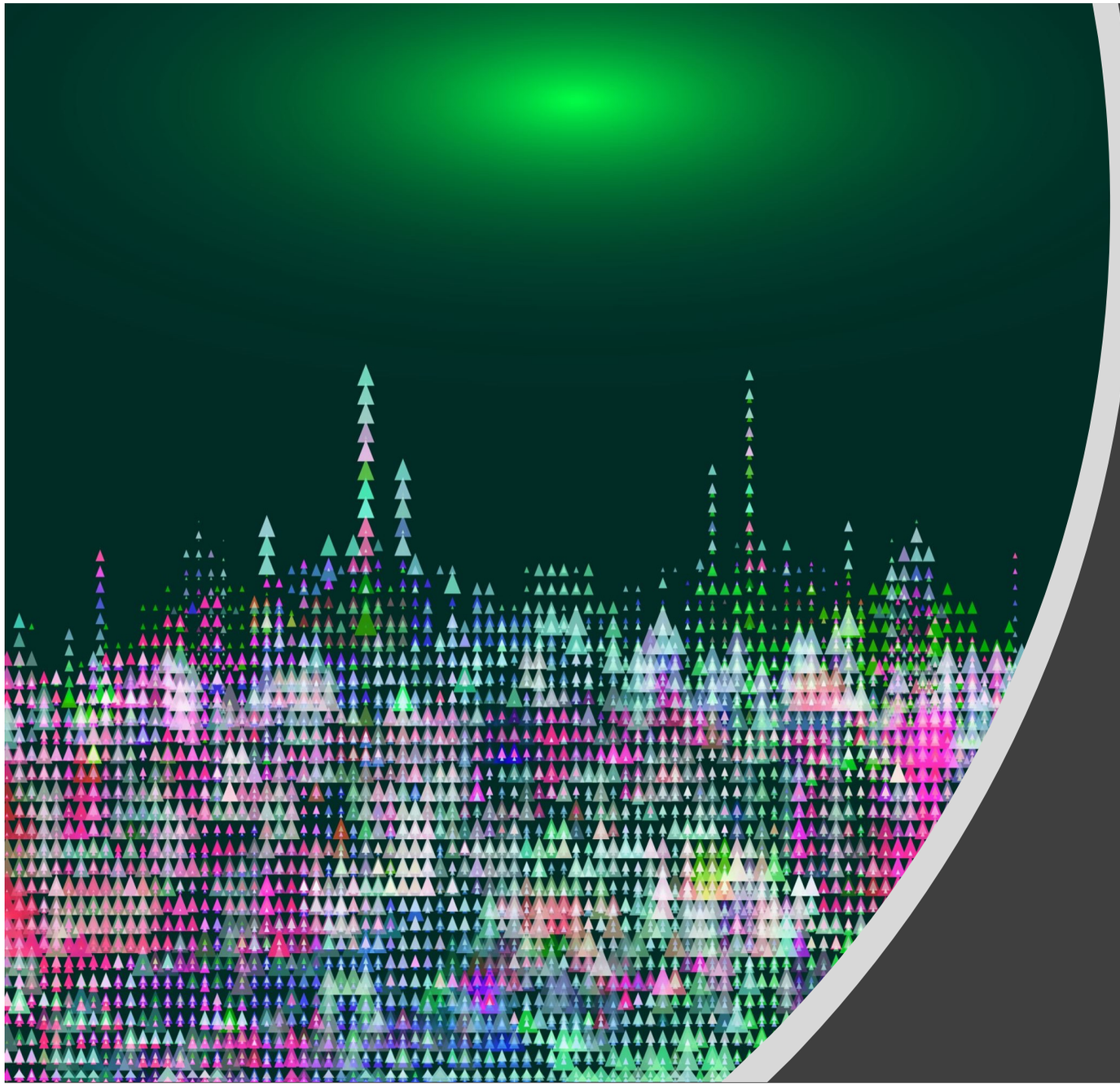


What can we infer in this case?



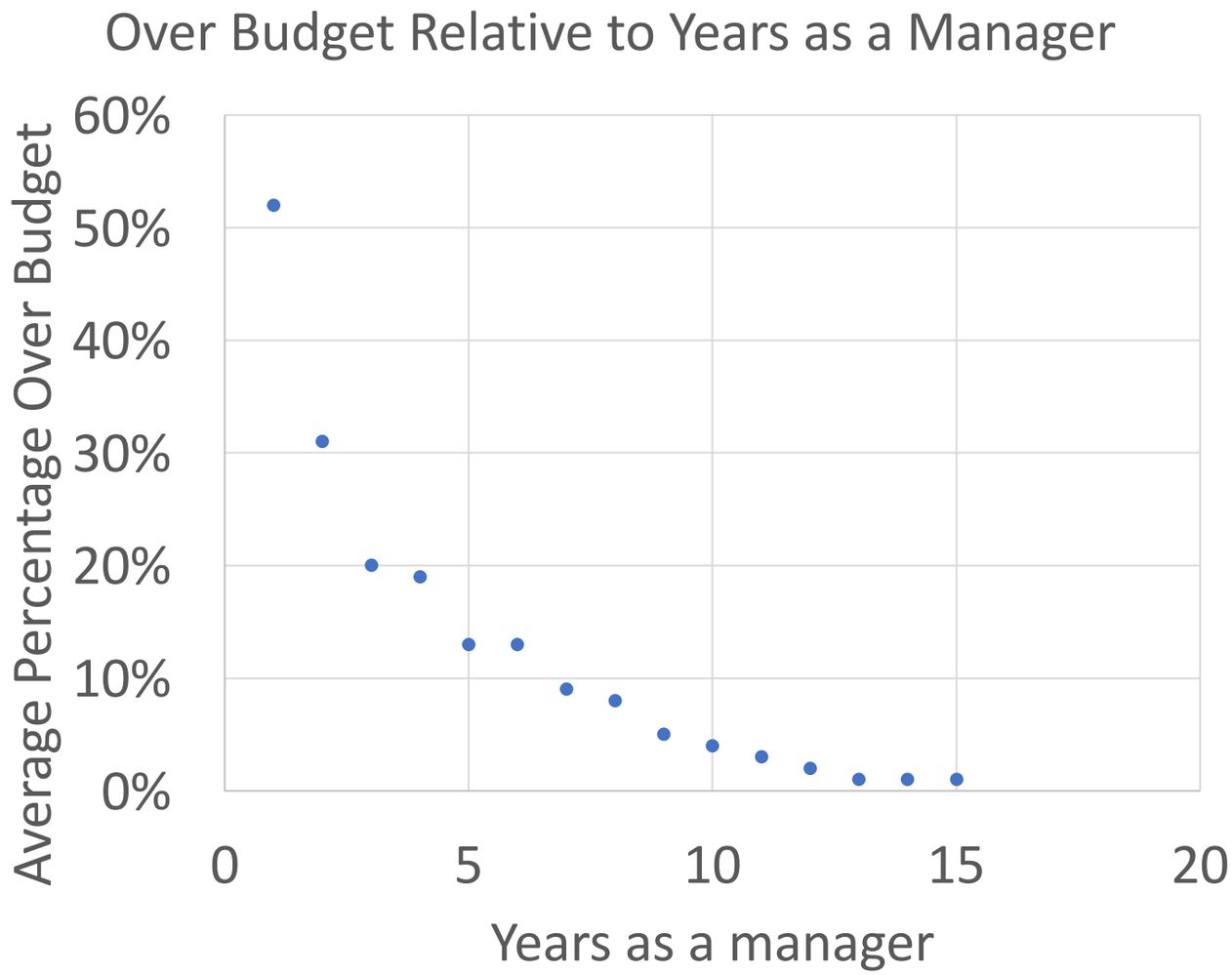


How would we automate this analysis?

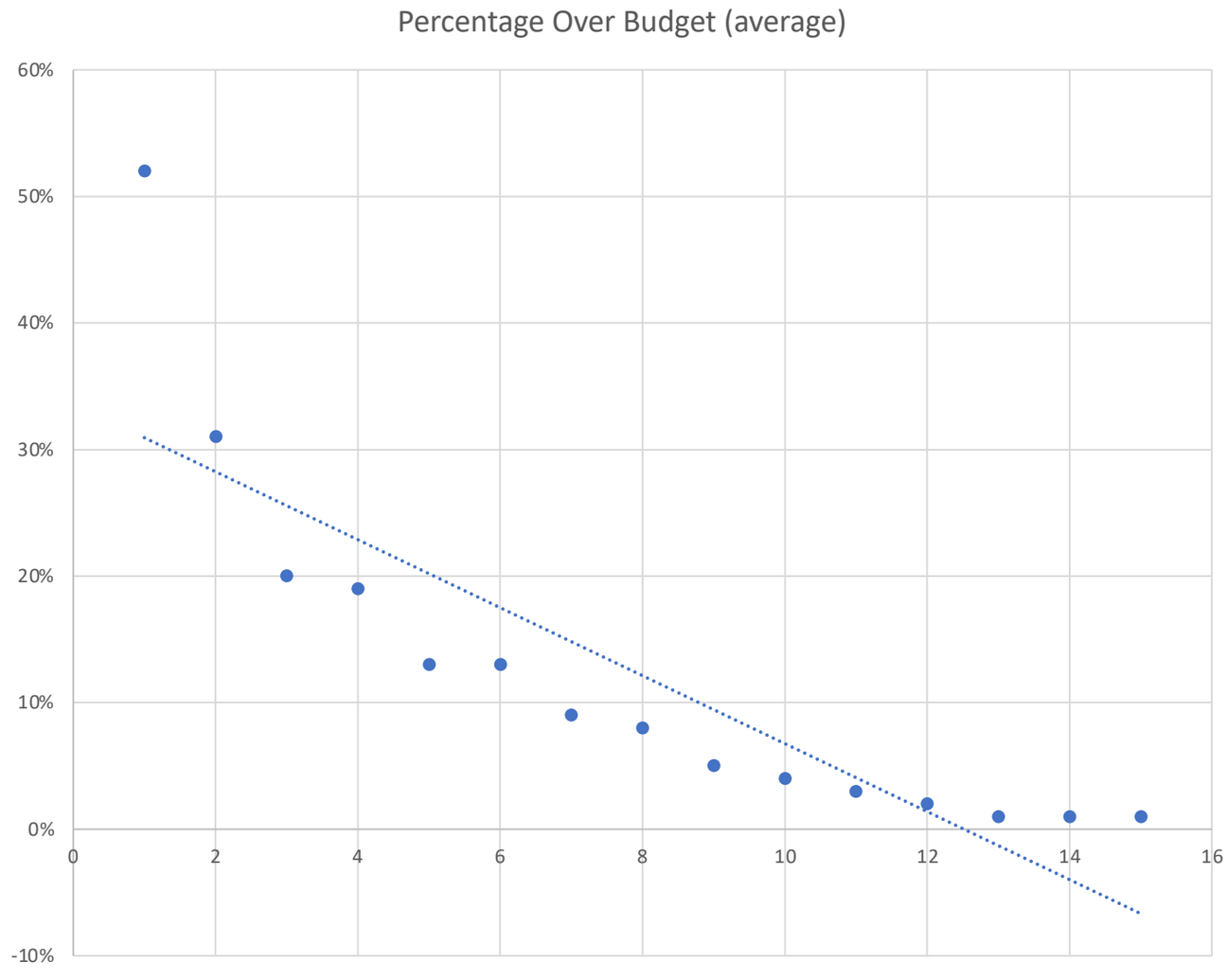


Analysis: Numerical Relationship

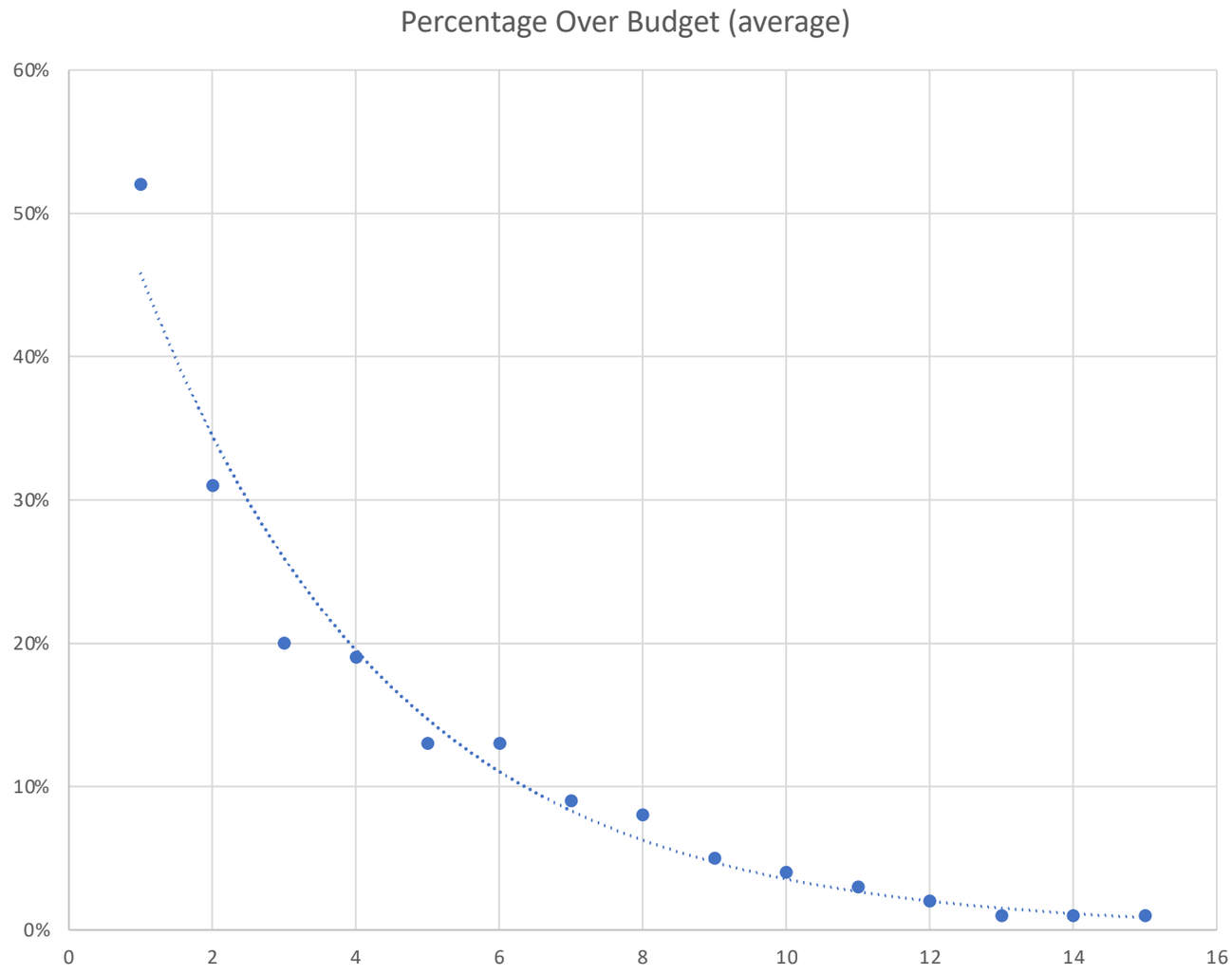
Analysis of Relationships



Adding a Trendline

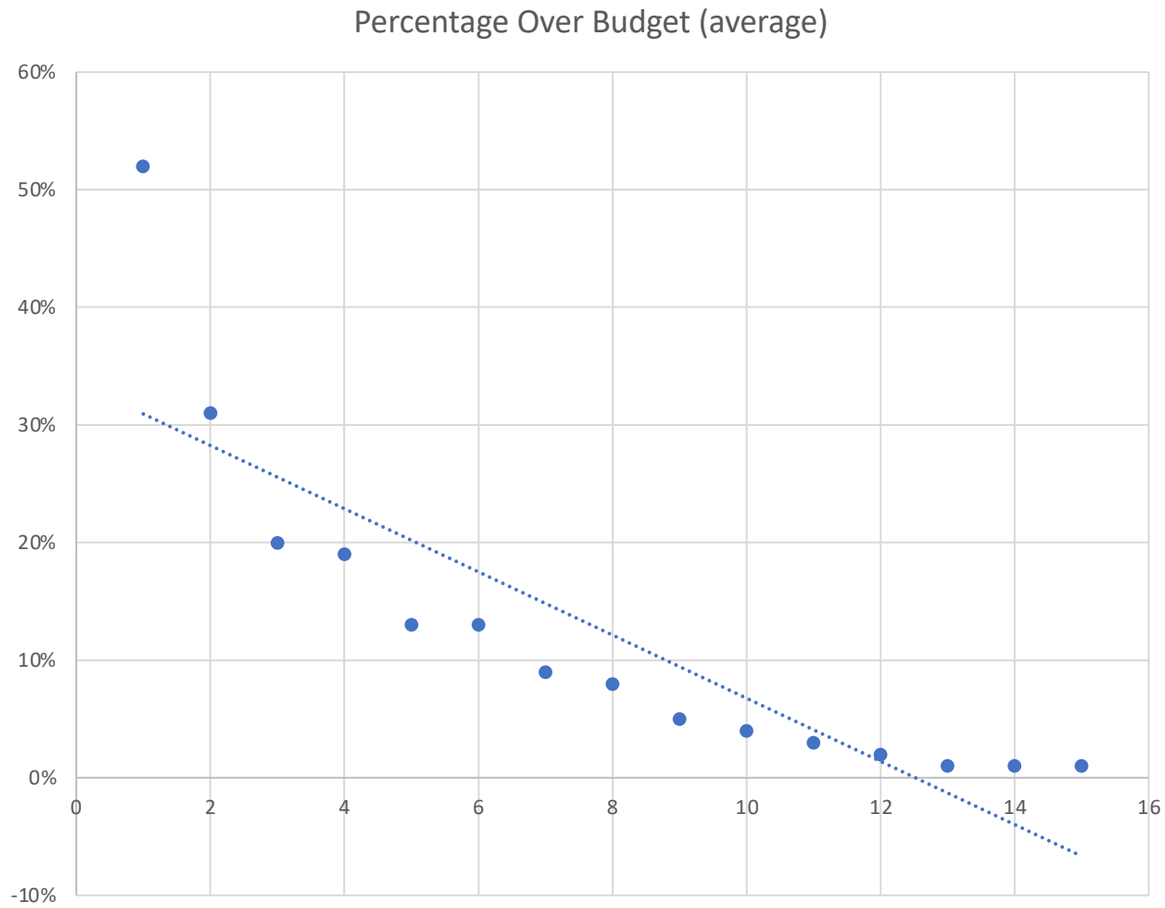


Adding a **Better** Trendline



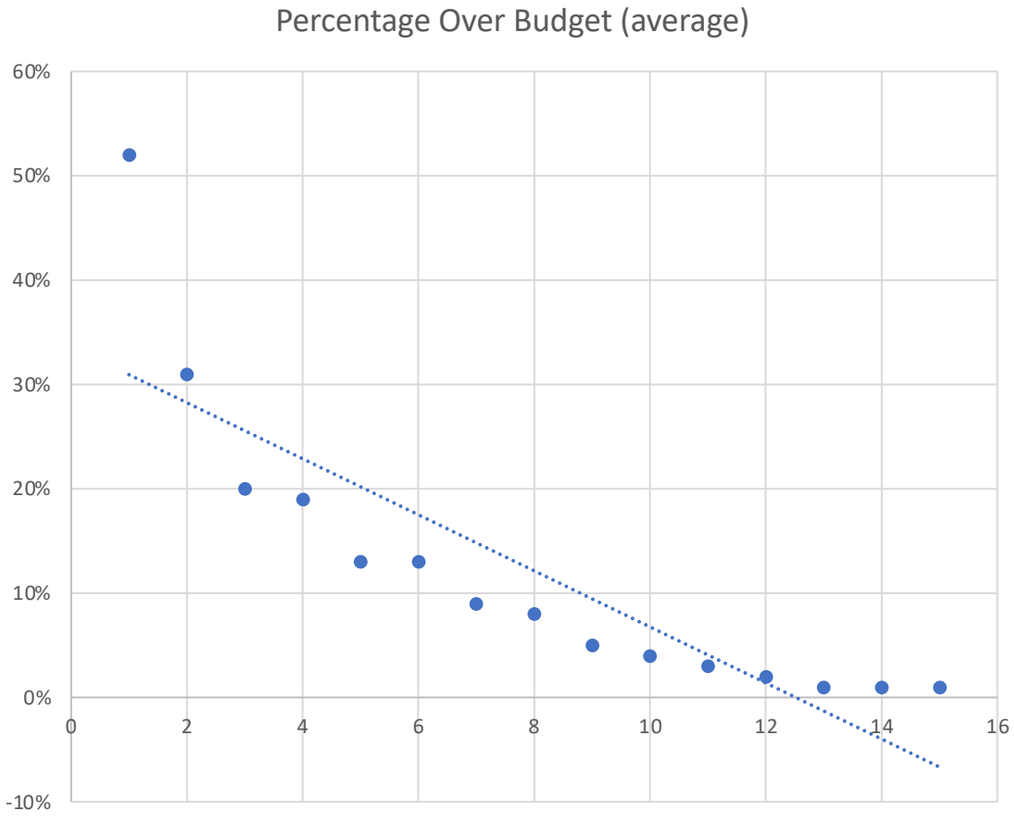
How To Quantify Trendline Fit?

- We could use something called the Mean Absolute Error (MAE)*:
- Determine the distance of the points to the line.
- Take the absolute values
- Add them up and take the average (divide by number of points)
- This give a measure of how well the model fits
- (Hint: Just use some R code, which will do this sort of thing for you automatically!)

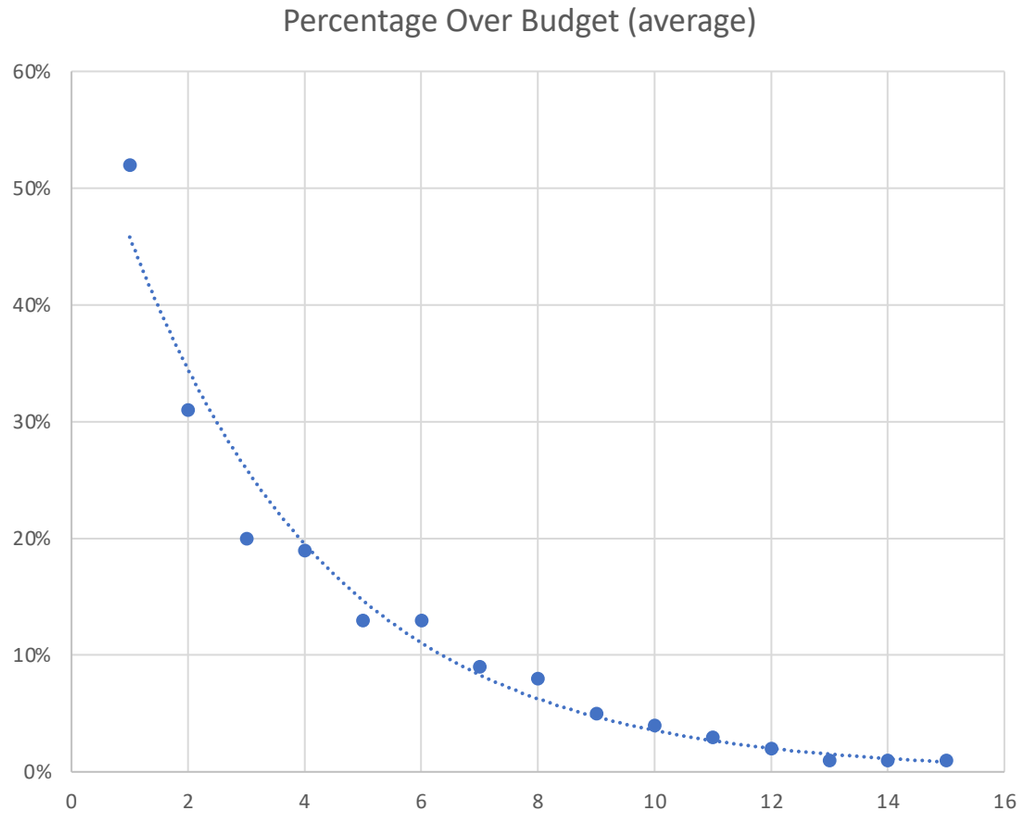


*There are more sophisticated strategies for measuring fit, but this as a starting point.

Can we prove which is the better trendline?



MAE = 5.25



MAE = 0.56

Analysis: Categorical Relationship

Consider A Categorical Hypothesis

- Suppose we have the following hypothesis:
  - Across all departments, there is the same percentage of managers and non-managers
  - Within any department, this percentage is 20% managers and 80% non-managers
- (Maybe we want to take this even further and say this is what the breakdown **should** be like…)

What does this look like numerically?

	Operations	Sales	Corporate Services
Managers	20%	20%	20%
Non-Managers	80%	80%	80%

Suppose we have 135 people in the organization…

	Operations	Sales	Corporate Services	
Managers	10	3	14	27
Non-Managers	40	12	56	108
	50	15	70	135

What is the actual breakdown?

	Operations	Sales	Corporate Services	Totals
Managers	8	22	10	40
Non-Managers	32	3	60	95
Totals	42	25	70	135

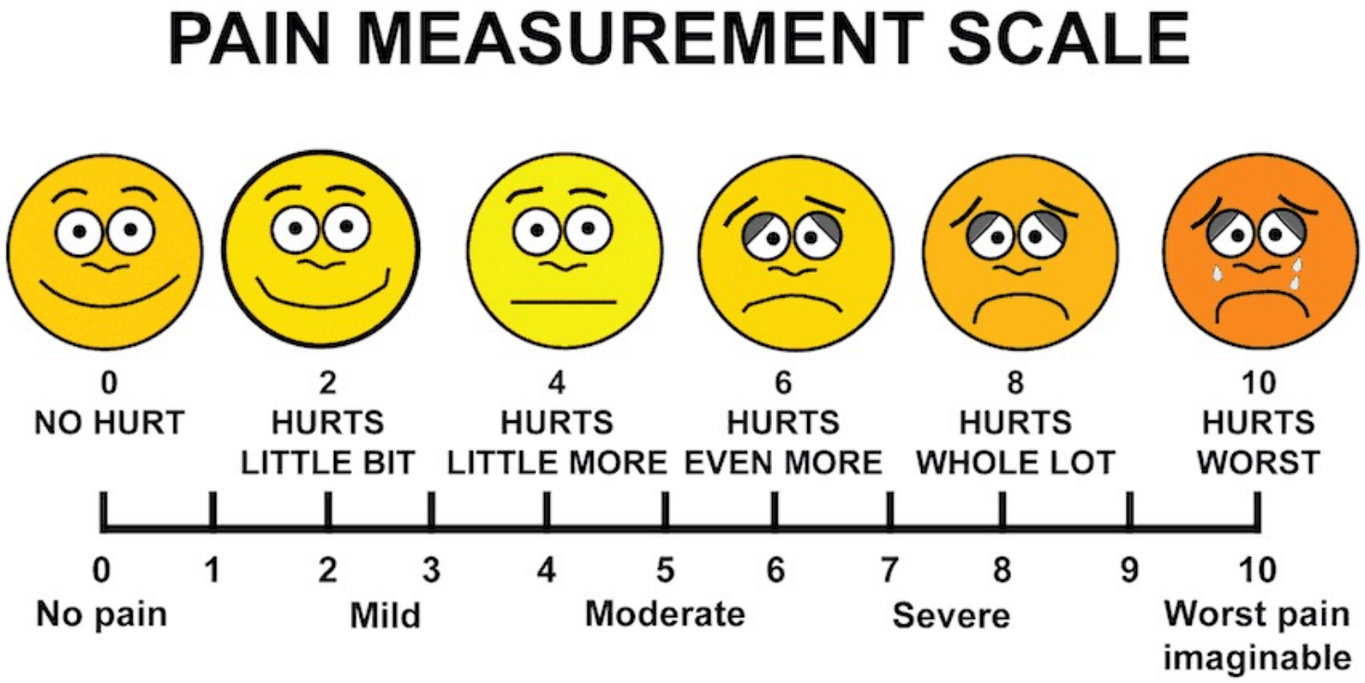
	Operations	Sales	Corporate Services	Totals
Managers	20%	88%	14%	30%
Non-Managers	80%	12%	86%	70%

	Operations	Sales	Corporate Services	Total
Managers	20%	20%	20%	20%
Non-Managers	80%	80%	80%	80%

	Operations	Sales	Corporate Services	Total
Managers	20%	88%	14%	30%
Non-Managers	80%	12%	86%	70%

Results: Anticipated vs Actual

A Note About Ordinal Data



Ordinal Data Best Practices

- Try to avoid it! But that's probably not realistic…
- Could just treat as categorical, then do proportions
- Does it make sense to take the mean of ordinal data values? It's certainly possible to do, but…
- Some argument in social science literature has been made (on the basis of some evidence) that it is can be acceptable to treat ordinal data as if it were numeric for the purposes of analysis, under some circumstances.
- The more fine-grained, the closer you are to numerical (but… heaping!)
- Non-parametric tests. Or use a Lickert scale/Likert item approach.