BASIC DATA ANALYTICS TECHNIQUES

SETTING THE STAGE

OUTLINE

1. Background and Process
2. Insight *via* Number Crunching: Some Core Concepts
3. Insight *via* Number Crunching: Some Core Techniques
4. Next Week: A Different Context

WHAT WE'VE COVERED SO FAR…

- Data modeling and conceptual analysis
- Data collection
- Data transformation
- Data storage
- Data exploration
- Data presentation

WHAT WE'VE COVERED SO FAR…

- Data modeling and conceptual analysis
- Data collection
- Data transformation
- Data storage
- Data exploration
- Data presentation

Today:
Putting it all together in the context of business intelligence and data analysis for business intelligence.

BUSINESS INTELLIGENCE – BUSINESS ANALYTICS

Use data (and information) about internal operations and the state of the market to support **informed decision making** about **business operations** and **business strategy**.

No firmly agreed upon definition of these terms – is one a subset of the other?

Goals: increased situational awareness + improved foresight

HISTORY OF BUSINESS INTELLIGENCE

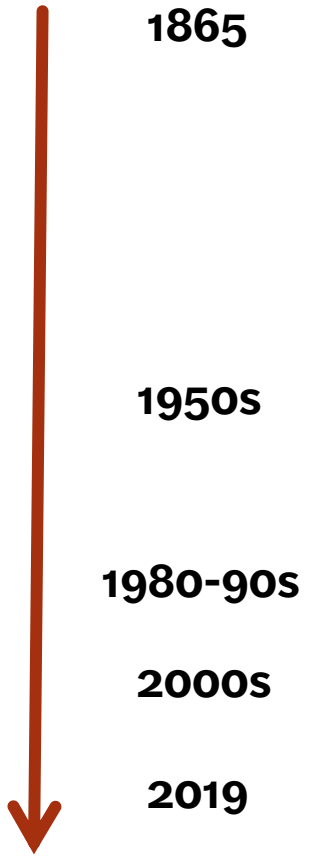
Late 1800s: people started to recognize that they could use data to gain a competitive advantage

1950s: advent of the first business database for decision support

1980s -1990s: computers and data becoming increasingly available - data warehouses, data mining – still very technical and specialized

2000s: trying to take business analytics out of the hands of data miners and other specialists and more into the hands of domain experts

Now: big data and specialized techniques have arrived on the scene, but so has data visualization, dashboards, software as a service



BUSINESS INTELLIGENCE AND DATA SCIENCE

Historically, one of the streams contributing to modern day Data Science

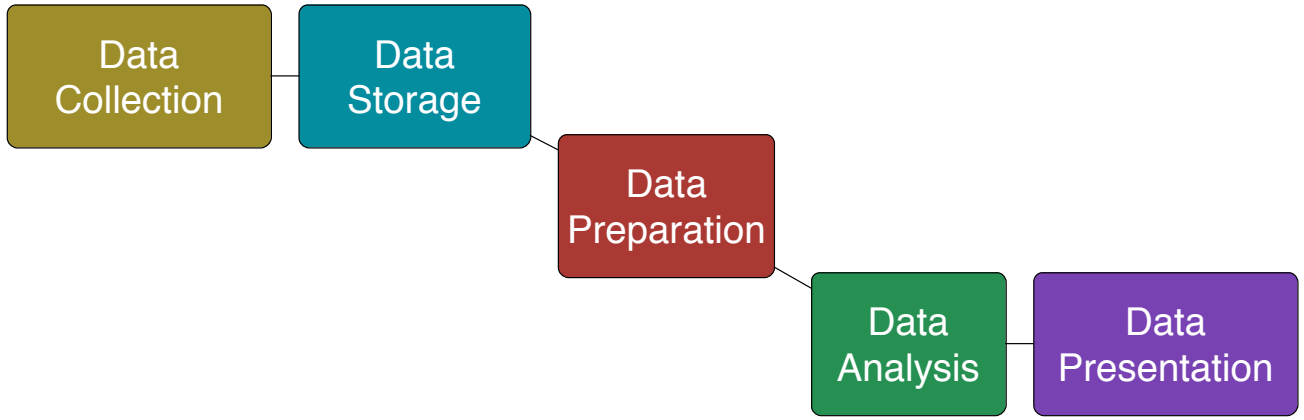
- **System of interest**: the commercial realm - the market in which you are involved
- **Sources of data**: transaction data, financial data, sales data, organizational data
- **Goals**: provide awareness of competitors, consumers and internal activity and use this to support decision making
- **Culture and preferred techniques**: datamarts, key performance indicators, consumer behaviour, slicing and dicing, business 'facts'

The ultimate goal is still the same - insight into your system of interest.

BUSINESS INTELLIGENCE AND THE DATA PIPELINE

Our general data pipeline model also works for business intelligence.

What are some aspects of the business intelligence pipeline that might distinguish it from a more generic analysis pipeline?

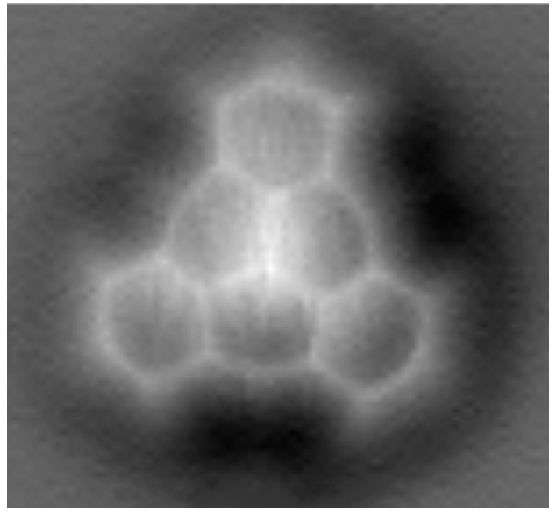


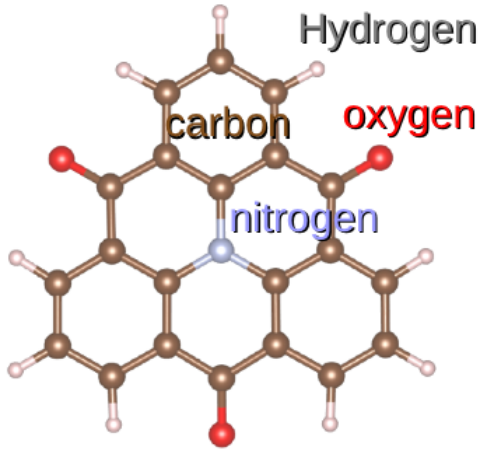
FINDING PATTERNS, GENERALIZATIONS AND STRUCTURE

- **Pattern**: A predictable, repeating regularity
- **Structure**: An organization of elements in a system
- **Generalization**: Creation of more general or abstract concepts from more specific concepts or instances

Underlying goal during analysis - find patterns or structure in our data, draw conclusions via these patterns or structures.

Finding patterns and structure is not bad or wrong, per se, it's how you use these discoveries – the conclusions that you draw - that is important.





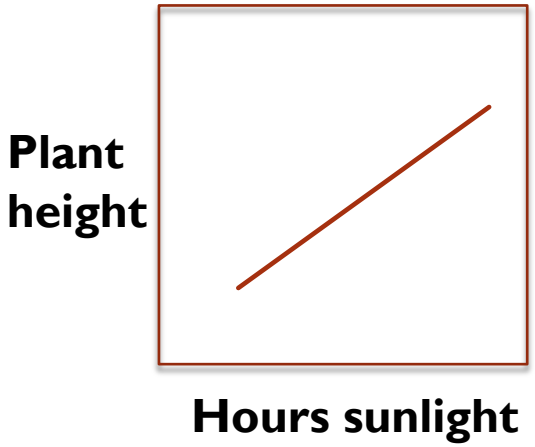
https://upload.wikimedia.org/wikipedia/commons/d/d0/TOAT_AFM.png

INDEPENDENT VS DEPENDENT VARIABLES

In an experimental setting:

- **Control/Extraneous Variables**: We do our best to keep these controlled and unchanging while other variables are changed
- **Independent**: We control the values of the variable. We suspect that they influence the dependent variables.
- **Dependent**: We don't control the values - they are generated in some way during the experiment, and presumably are dependent on everything

How do these translate over to other datasets?



TYPES OF DATA

Numerical Data: integers or continuous numbers

- 1, 7, 34.654, 0.000004

Text Data: strings of text – may be restricted to a certain number of characters

- “Welcome to the park”, “AAAAA”, “345”, “45.678”

Categorical Data: a fixed number of values, may be numeric or represented by strings. **There is no specific or inherent ordering**

- (‘red’,’blue’,’green’),(‘1’,’2’,’3’)

Ordinal Data: Categorical data with an inherent ordering. Unlike integer data, the spacing between values is **not** defined

- (very cold, cold, tepid, warm, super hot)

TURNING CATEGORICAL DATA INTO NUMERICAL (COUNT) DATA

We can useful turn categorical data into numeric data by generating frequency counts of the different values of the categorical variable.

This in turn allows us to apply numerical analysis techniques.

House Colour	Frequency
red	40
blue	13
green	2

THE SPECIAL ROLE OF CATEGORICAL DATA

Categorical data play a special role:

- In *data science*, we talk about a categorical variable with a pre-defined set of values
- In *experimental science*, a factor is an independent variable with the levels of the variable defined - it may also be viewed as a category of treatment
- In *business analytics,* people talk about dimensions (with members) vs measures

However we label these types of variables, we can use these them to **subset** our data, or **roll up/summarize** our data.

HIERARCHICAL / NESTED / MULTILEVEL DATA / MODELS

If a categorical variable has multiple levels of abstraction, we can create levels out of this variable.

We can view these levels as new categorical variables, in a sense.

The 'new' categorical variable has a pre-defined relationship with the more detailed level.

This is common with time and space variables – we can 'zoom' in or out.

This lets us talk about the **granularity** of the data – what is the 'maximum zoom'?

Year	Quarter	Count
2012	1	34
2012	2	12
2012	3	52
2012	4	0
2013	1	21
2013	2	9
2013	3	112
2103	4	8

DATA SUMMARIZING

Min: Smallest value of variable

Max: Largest value of variable

Median: Middle value of variable

Mode: Most frequent value

Unique Values: List of unique values

Signal	Type
4.31	Blue
5.34	Orange
3.79	Blue
5.19	Blue
4.93	Green
5.76	Orange
3.25	Orange
7.12	Orange
2.85	Blue

ROLLING UP YOUR DATA

We can perform an operation over a set (or subset) of the data - typically over a column of the data.

When we do this, we can think of this as compressing or 'rolling up' the many data values into a single representative value.

Typical roll up functions are 'mean', 'sum' and 'count'.

If we apply the same roll up function to many different columns we can think of this as **mapping** (a list of) columns to functions.

Signal	Type
4.31	Blue
5.34	Orange
3.79	Blue
5.19	Blue
4.93	Green
5.76	Orange
3.25	Orange
7.12	Orange
2.85	Blue

CONTINGENCY TABLES / PIVOT TABLES

Contingency Table: A table used to examine the relationship between two categorical variables - specifically the frequency of one variable relative to a second variable (cross tabulation).

Pivot Table: A table generated in a software application by applying operations (e.g. sum, count, mean) to variables, possibly based on another (categorical) variable. Can be used to create a contingency table.

	Large	Medium	Small
Blue	1	32	31
Orange	14	11	0
Green	5	5	5

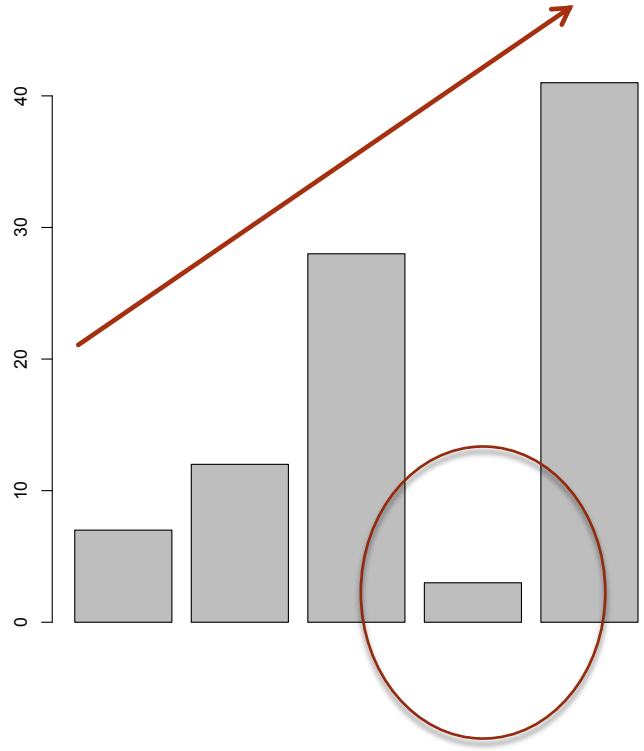
ANALYSIS THROUGH VISUALIZATION

Analysis broadly defined:

- identifying patterns or structure
- adding meaning to these patterns or structure by **interpreting** them in the context of your system.

Option 1: use analysis techniques to do this.

Option 2: visualize the data and use the analytic power of our (perceptual) brain to come to meaningful conclusions about these patterns.



SOME SIMPLE VISUALIZATIONS TO REVEAL PATTERNS

Scatter plot: best suited for two numeric variables

Line chart: numeric variable and categorical variable

Bar chart: best suited for one categorical and one numeric - or multiple categorical/nested categorical data and