LA TROBE UNIVERSITY

Building AI – Module 5

Data and Stats Refresh
14 June 2022

Optus U

Agenda

1. Organisational Data
2. Data Taxonomy
3. Data Management
4. Statistical Techniques
5. Distributions And Tests

Optus U

Defining Data

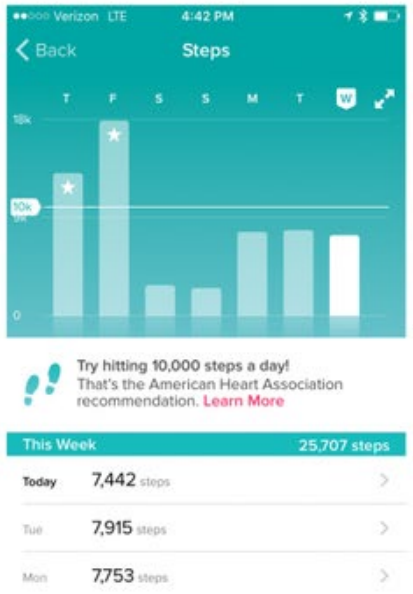
- Data is/are a digital representation of "……."
- Organisations – products/services, customers/stakeholders, staff, budgets, reputation, branding
- Individuals – Shopping list, Reading list, Netflix history, Fitbit tracking, Personal finances, CV, Social media (LinkedIn or Twitter) - network and interactions
- Governments – Public policy, Census data, Reserve bank data, Capital markets, Carbon emissions, Social welfare, Foreign policy
- Natural environment – BOM, land registry, local sensors, citizen science
- Social networks – hashtags, handles, tweets, retweets, likes, replies

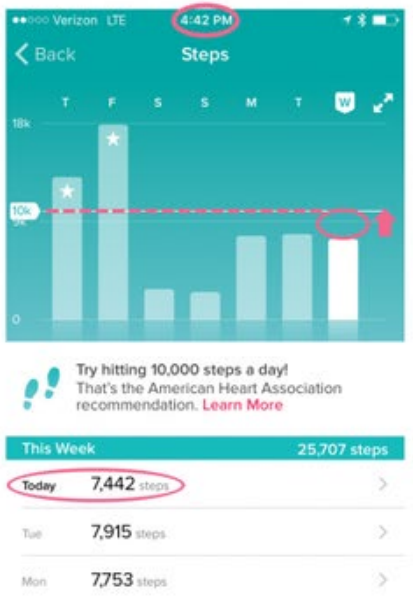
How do we use data?

- Transformation of data into insights that inform decisions
  - Data
  - Transformation – Analytics/AI/Automation
  - Insights
  - Decisions

A Relatable Example

DATA → INFORMATION → INSIGHT → ACTION

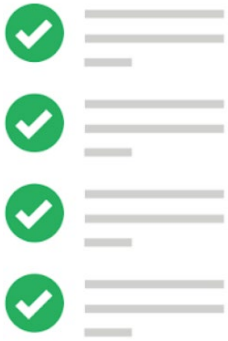




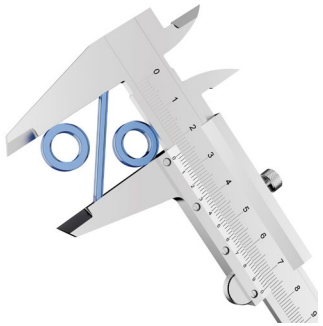
Forbes

Data to Decisions

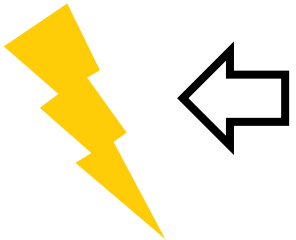
Strategic Goals



Organisational Objectives

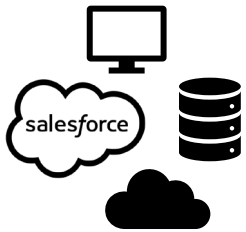


Operational Measures



Analytics/AI/ Automation

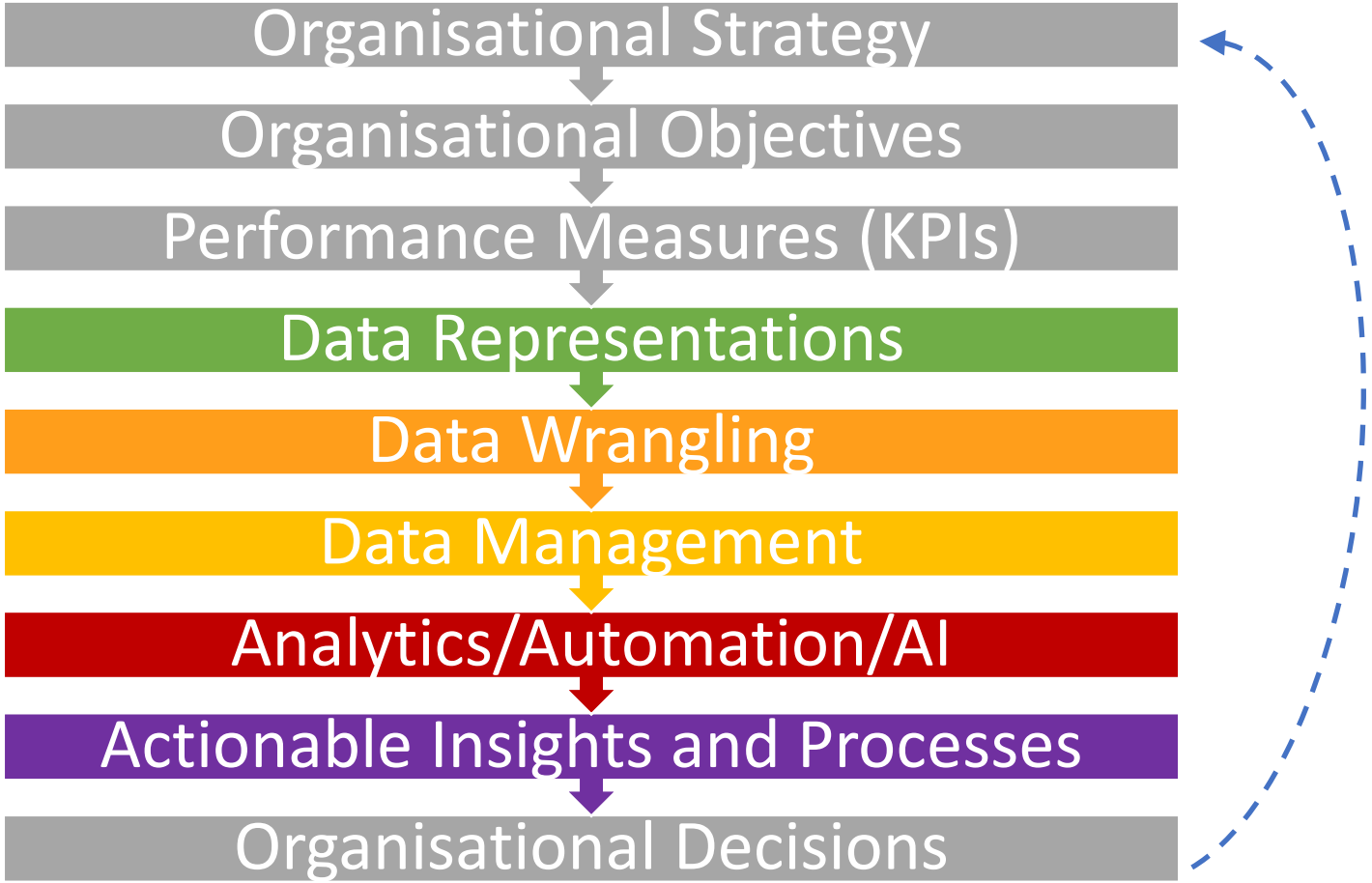
Social



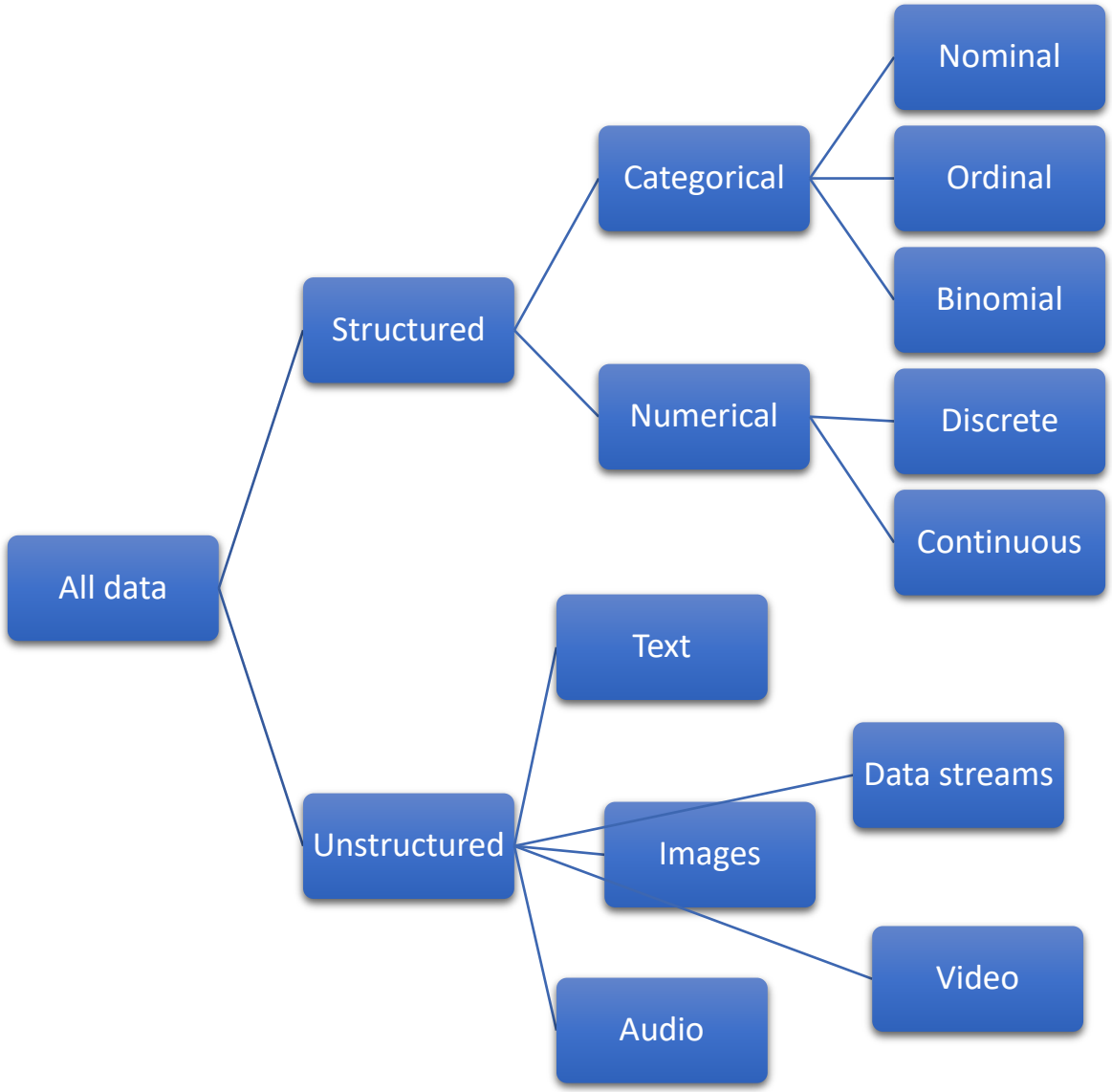
Internal and External Systems

Internal Data

Decisions to Strategy



A taxonomy of data types



Numerical Data (Quantitative)

- Discrete data
  - “How many?” – counting
  - Integer values, whole numbers but nothing in between
  - Units sold, hours worked, # of registrations, calories burned
- Continuous data
  - “How much?” – measuring
  - Real values, whole numbers and everything in between
  - Cost, revenue, GDP, duration, speed, weight, information, time
- Tricks
  - Discrete as categorical: age is more effective in age-groups
  - Continuous as categorical: cost represented as low, medium, high
  - Interval data: starts off as categorical but also provide a numerical output – date, time, temperature

Categorical Data (Qualitative)

- Each value belongs to one of a set of well-defined categories
- "What type?" – what is your occupation?
  - Nominal Data: purely used to identify, label, categorise
    - Service categories, gender, marital status, type of sport
    - Two service categories – 1) individual 2) families
    - "How many families do we serve?" 10 – this is numerical discrete
  - Ordinal Data: Nominal data + can be arranged in rank or specific order
    - Age group from the previous slide, in ascending order
    - level of education, sentiment score
  - Binomial (Binary) Data: one of two mutually exclusive categories
    - yes/no, true/false, accept/reject

Variables

- We record data - characteristics or measures of organisational interest.
  - revenue, customer satisfaction, plans, sentiment
- These are technically referred to as variables.
- Variables are also called attributes, data points, inputs, or columns.
- Observations - the actual data or values recorded for each variable
- Observations vary from one “entity” to another.
- A data record (or input vector or ‘rows’) - a collection of such observations
- Dataset (rows and columns) - a collection of data records relating to an entity

An Example: Winery

- We sell wine from different regions in various markets and measure our sales performance over time.

Variables (attributes, data points, inputs, or columns)

Date	Product	Country	Staff	SalesAmount	SalesQuantity
03/01/2018	Riesling	China	Jill	$300.00	30
06/02/2018	Riesling	Scandinavia	Jo	$440.00	20
08/02/2018	Riesling	Scandinavia	Jill	$440.00	20
09/02/2018	Riesling	China	Jo	$440.00	20
15/03/2018	Riesling	Scandinavia	Jill	$440.00	20
29/05/2018	CabSav	America	Jill	$778.00	20
20/04/2018	Prosecco	Scandinavia	Jill	$611.00	18
17/03/2018	Prosecco	South Africa	Jill	$699.00	18
17/03/2018	Prosecco	China	Jill	$601.00	18
25/05/2018	Prosecco	China	Jill	$537.00	17
19/04/2018	CabSav	South Africa	John	$739.00	17
23/04/2018	Prosecco	Scandinavia	Jo	$582.00	17
17/03/2018	Prosecco	Scandinavia	Jill	$637.00	17
23/04/2018	Prosecco	South Africa	John	$645.00	17
25/05/2018	Prosecco	China	John	$508.00	16

Variable names

Observations for variable 'SalesQuantity'
- All must be of the same data type
- Discrete numerical

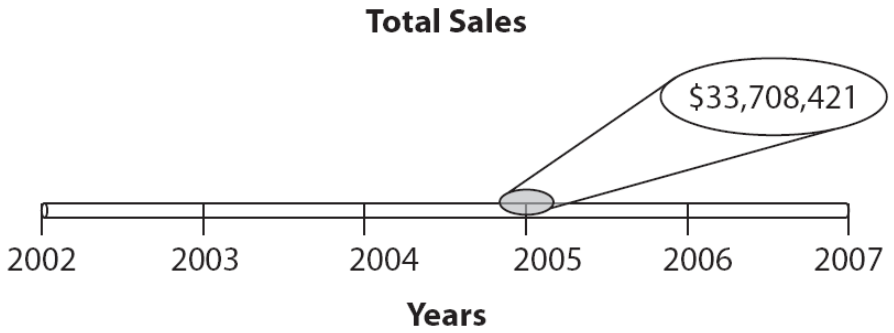
Data record, input vector
- Different data types
- Must contain a value for each variable
- If not, missing data

Dataset - a collection of data records relating to an entity?

Example: Numerical by Categorical

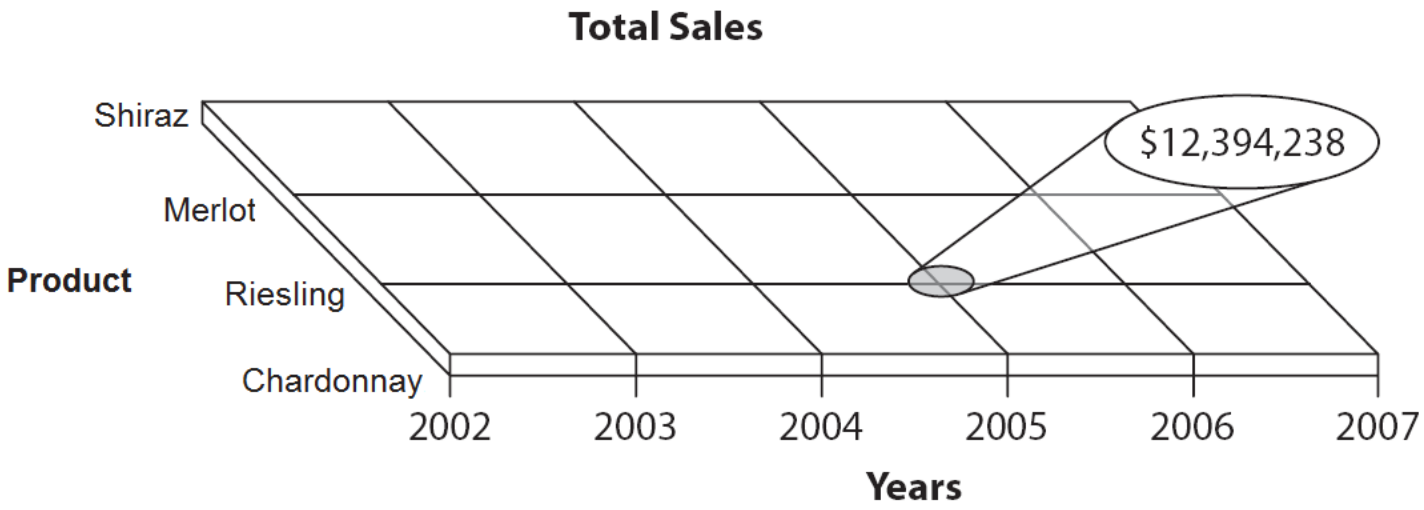
- We sell wine from different regions in various markets and measure our sales performance over time."
- We need to measure sales performance –

- Let's introduce a Date variable



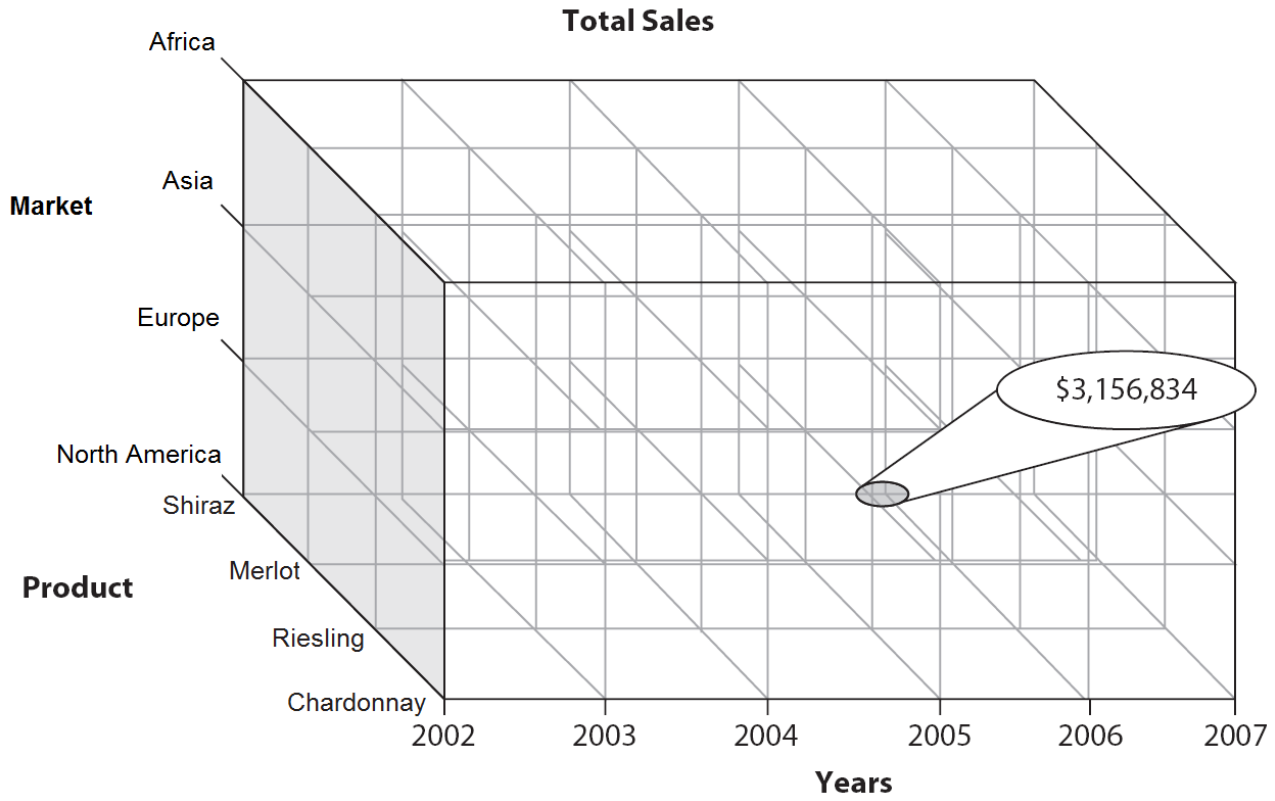
Example: Numerical by Categorical*2

- "We sell wine from different regions in various markets and measure our sales performance over time."
- Introduce a Product (wine) variable



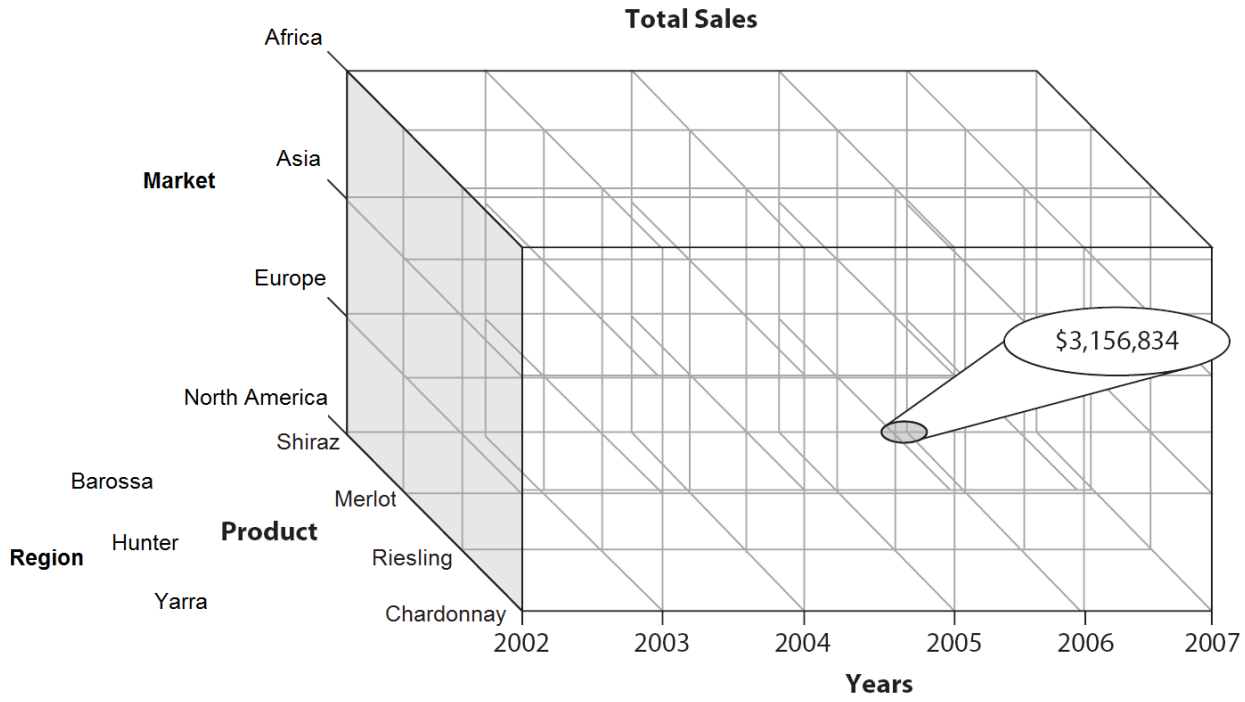
Example: Numerical by Categorical*3

- "We sell wine from different regions in various markets and measure our sales performance over time."
- And a Market variable



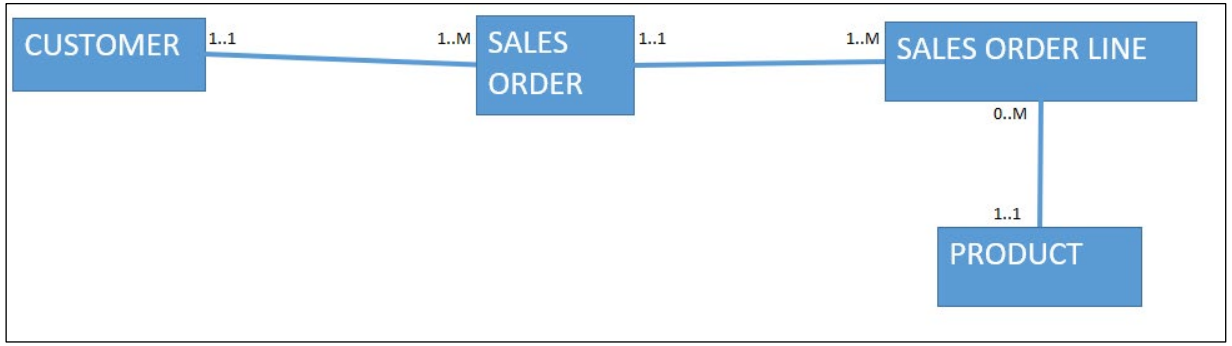
Example: Numerical by Categorical*4

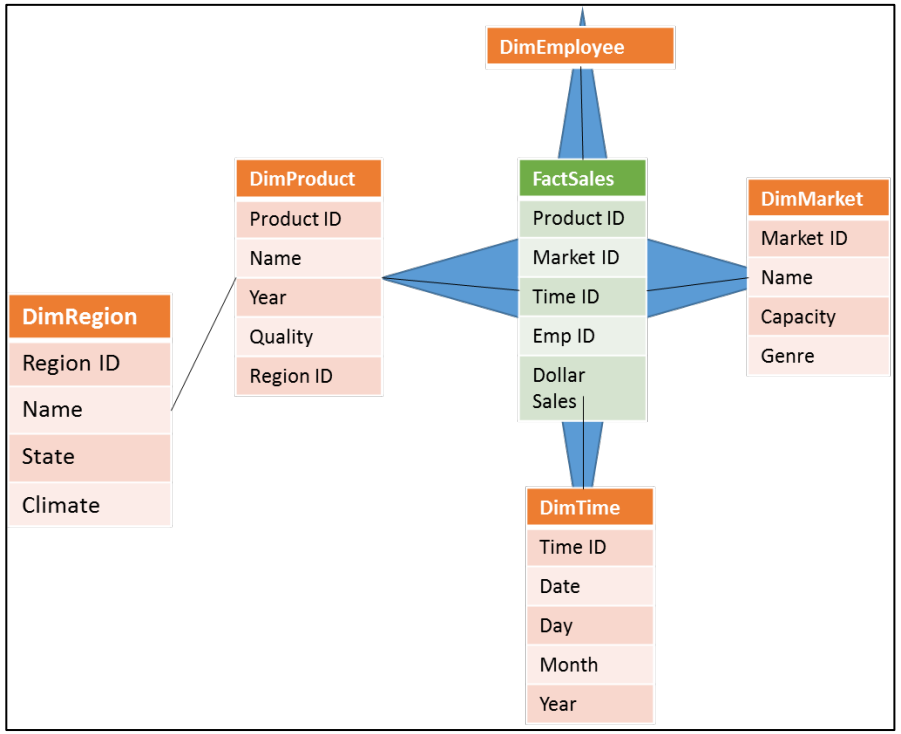
- What about Region?
- Not directly linked to sales performance, linked via product.
- More than three dimensions is not readily visualised in this format.



Data models

Date	Product	Country	Staff	SalesAmount	SalesQuantity
03/01/2018	Riesling	China	Jill	$300.00	30
06/02/2018	Riesling	Scandinavia	Jo	$440.00	20
08/02/2018	Riesling	Scandinavia	Jill	$440.00	20
09/02/2018	Riesling	China	Jo	$440.00	20
15/03/2018	Riesling	Scandinavia	Jill	$440.00	20
29/05/2018	CabSav	America	Jill	$778.00	20
20/04/2018	Prosecco	Scandinavia	Jill	$611.00	18
17/03/2018	Prosecco	South Africa	Jill	$699.00	18
17/03/2018	Prosecco	China	Jill	$601.00	18
25/05/2018	Prosecco	China	Jill	$537.00	17
19/04/2018	CabSav	South Africa	John	$739.00	17
23/04/2018	Prosecco	Scandinavia	Jo	$582.00	17
17/03/2018	Prosecco	Scandinavia	Jill	$637.00	17
23/04/2018	Prosecco	South Africa	John	$645.00	17
25/05/2018	Prosecco	China	John	$508.00	16





Data management

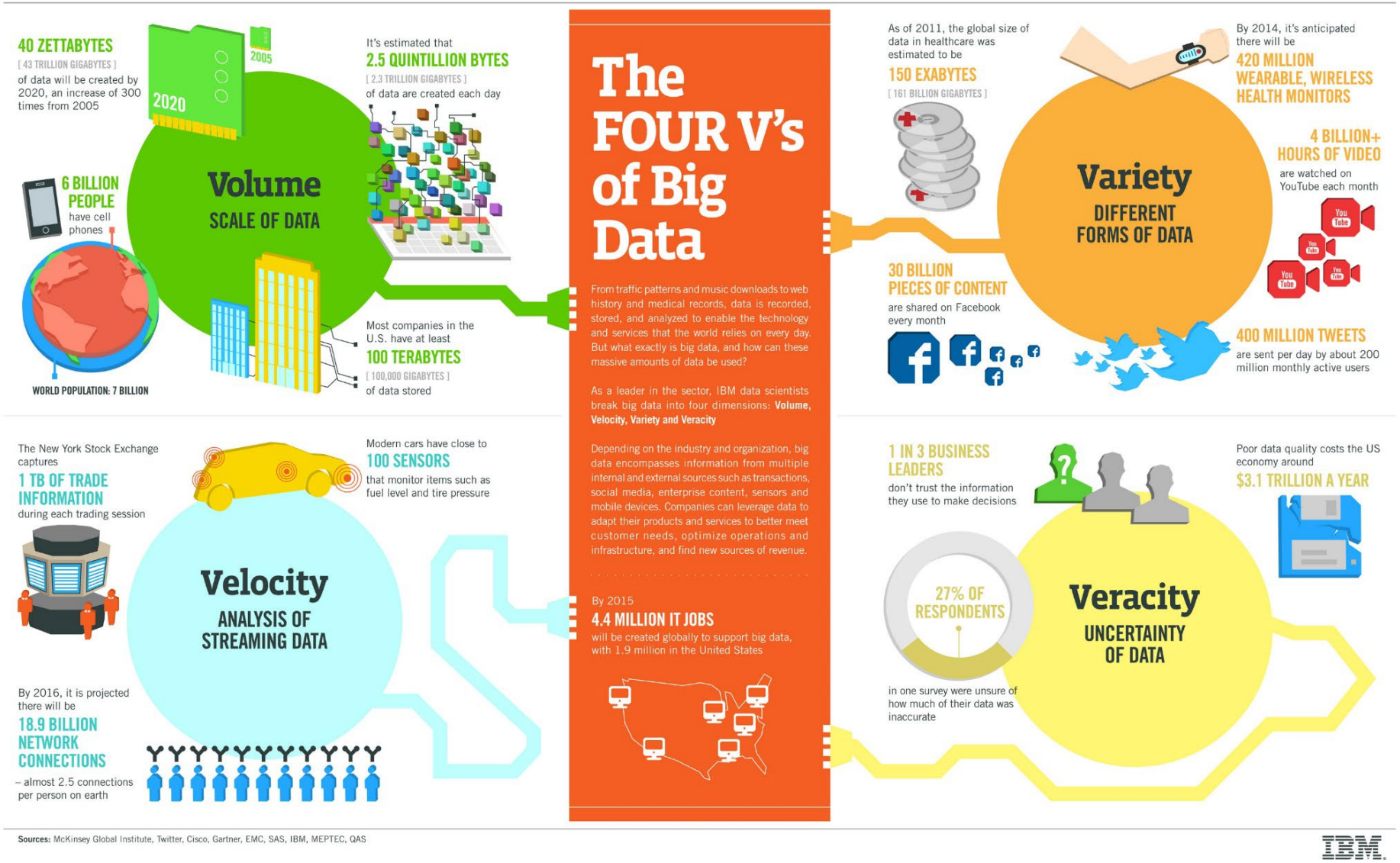
- Excel is useful..
- But what happens when the number of columns and rows increases?
  - Add new worksheets to the same spreadsheet.
  - Add new spreadsheets and/or text files
  - Pros and cons? “Excel hell”
- When not only columns and rows, but also relationships among datasets continue to increase, a more resilient structure/model is required.
- Occam's razor - “Entities should not be multiplied beyond necessity”
- Law of parsimony
- If one file works, stick to it. But more often it doesn’t.

Data management

- Database - the organisation of data into collections of two-dimensional structures called relations (tables).
- Data warehouse - A copy of transaction data, specifically structured for query and analysis
- Data lake - A centralized store of vast amounts of raw and transformed data (both structured or unstructured) described using metadata.

Data management system	Design approach	Query language
Database	Entity relationship (ER) model	SQL
Data warehouse	Dimensional model, ETL	MDX
Data lake	Key-value pairs, ELT	Hybrid (U-SQL)

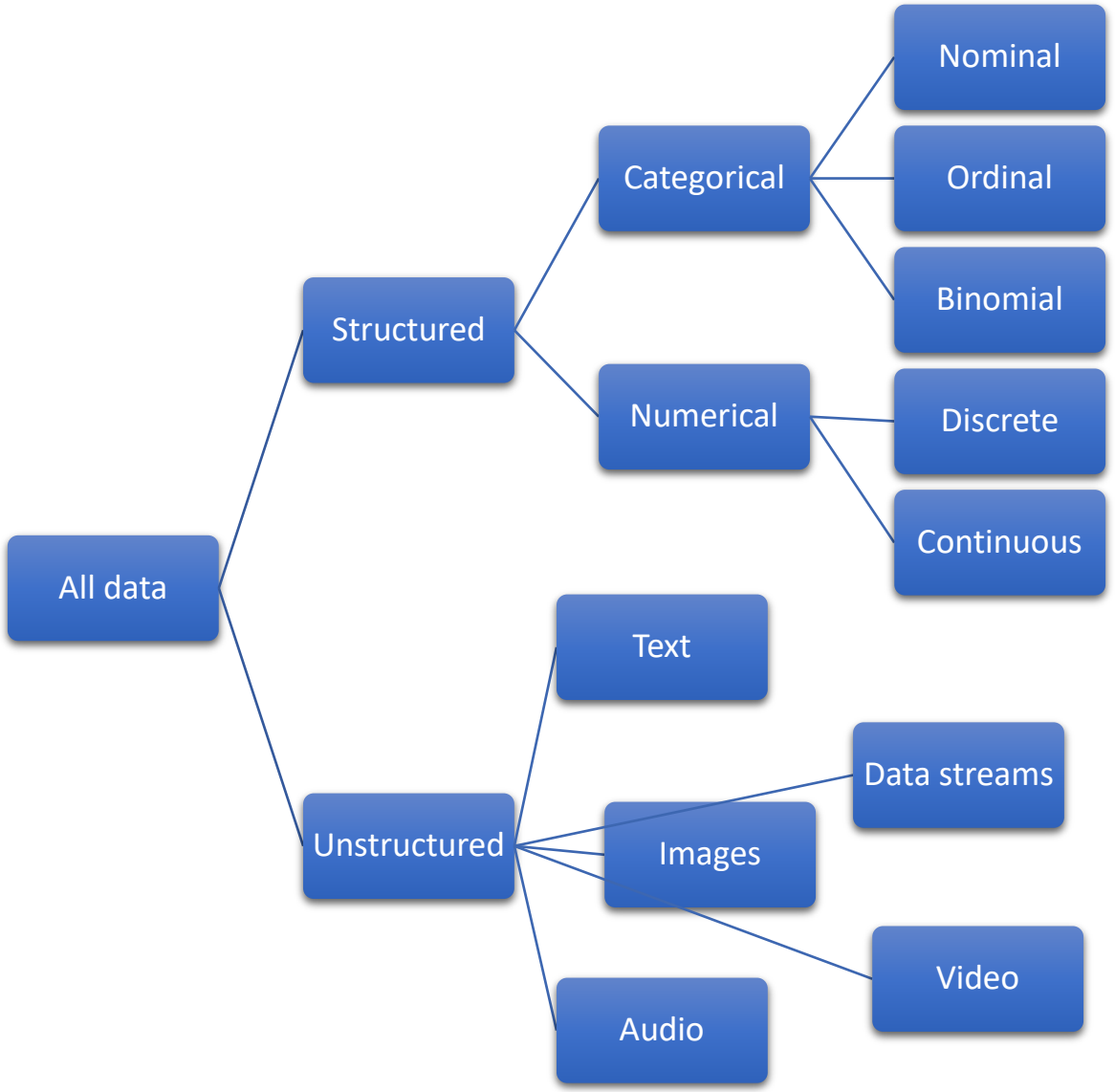
Big Data, visually..



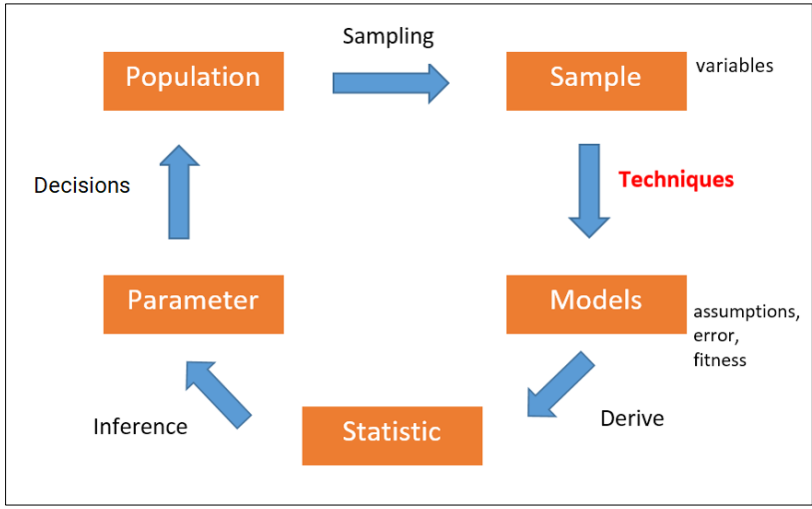
Defining Big Data

- “Big data is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making” – Doug Laney 2001.
- A high-level separation between Big Data and everything else,
  - **Structured data** – fits perfectly into rows and columns (rectangular data)
  - **Unstructured data** – everything else: documents, emails, twitter feeds, images, video, audio, speech, sensor data, variable length time series, networks of people (graphs)
- Wide data vs long data
  - Number of Columns vs number of rows

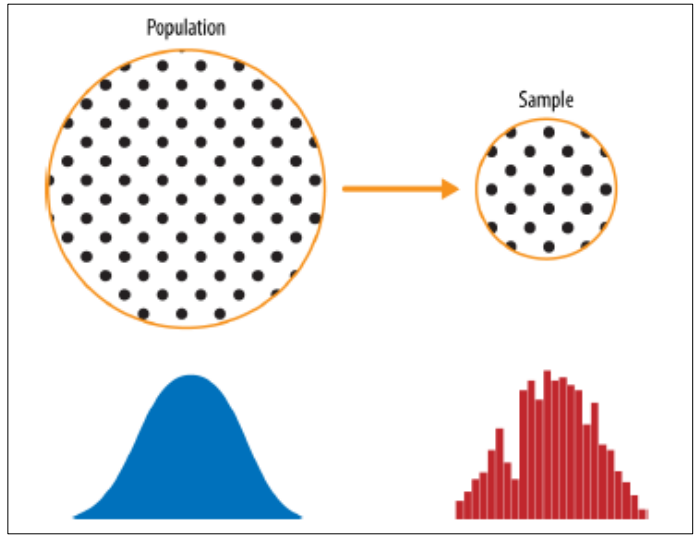
A taxonomy of data types



Data to Statistics



- A variable is an observed characteristic or measure.
  - And most variables vary (non-varying are called constants)
- We take actions/decisions by observing the variation of variables in groups.
  - A population is the entire group to be studied.
  - An individual is a person, entity or object that is a member of the population being studied.
  - A sample is a subset of the population that is being studied.
  - A statistic is a numerical summary of a sample.
  - A parameter is a numerical summary of a population.
- Variables are characteristics of individuals within a population.
  - Since we cannot observe entire populations, we take samples and derive statistics to draw conclusions on parameters, and take actions based on those conclusions.

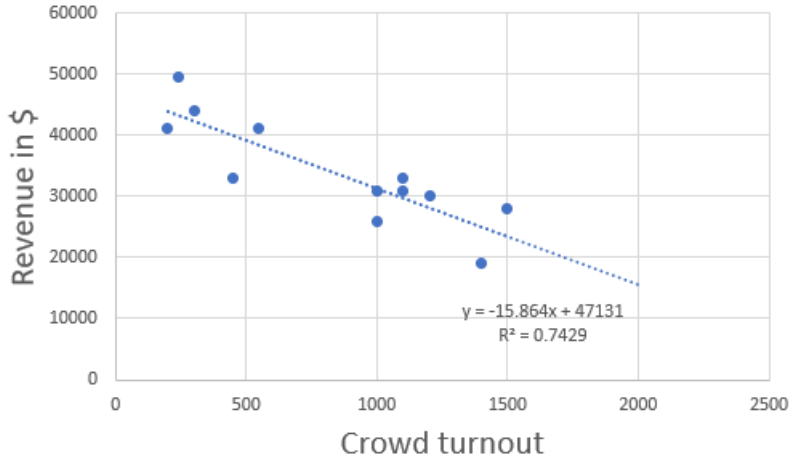


As Analytics

- What happened and why it happened?
  - Descriptive, visual, inferential, diagnostic analytics
- What is likely to occur?
  - Predictive analytics
- What actions can be taken?
  - Prescriptive analytics

An example: the trend line

- Descriptive Analytics:
  - When the crowd turnout is high the revenue generated is low.
  - Revenue is a decreasing function of crowd turnout.
- Predictive Analytics:
  - What revenue can be expected from a turnout of 2000 people?
- Prescriptive Analytics:
  - What are the revenue generating features of the smaller events?
  - How can they be replicated in large events?



Techniques by variables

- **Numerical**
  - **Univariate –** representing/aggregating a single variable
    - Centrality – mean, weighted mean, trimmed mean, median, mode
    - Distribution - standard deviation, variance, range, five number summary, boxplot, frequency table, histogram, density plot
    - Location – z-score, percentile, outliers
  - **Bivariate –** a relationship between two variables
    - Correlation - correlation coefficient, Correlation matrix, scatterplot, trend line
  - **Multivariate -** variables that are correlated but also vary together (co-vary)
    - (Machine learning)
- **Categorical** – frequency, mode, **contingency tables**

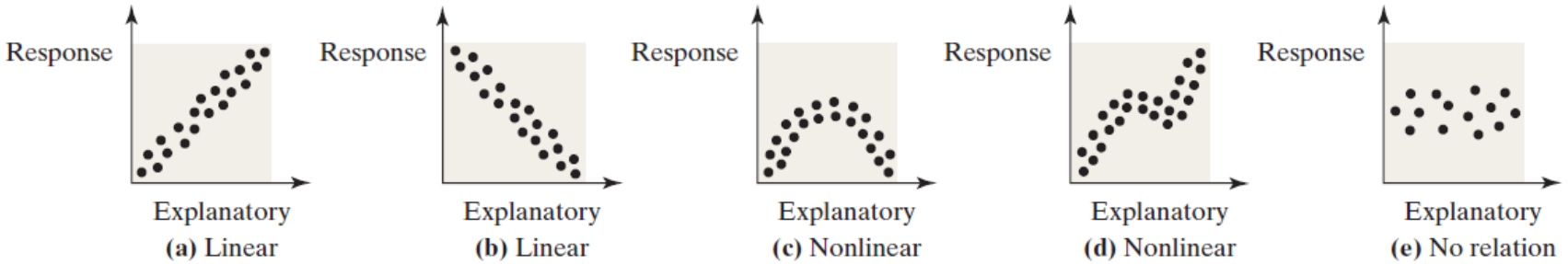
Contingency table

- **A contingency table (or two-way table)** is a display for two categorical variables. Its rows list the categories of one variable and its columns list the categories of the other variable. Each entry in the table is the number of observations in the sample at a particular combination of categories of the two categorical variables.
- Example: does level of education explain employment status?
- **Explanatory variable: level of education**
- **Response variable: employment status**

	Level of Education			
Employment Status	**Did Not Finish High School**	**High School Graduate**	**Some College**	**Bachelor's Degree or Higher**
Employed	9,993	34,130	34,067	43,992
Unemployed	1,806	3,838	3,161	2,149
Not in the labor force	19,969	30,246	18,373	16,290

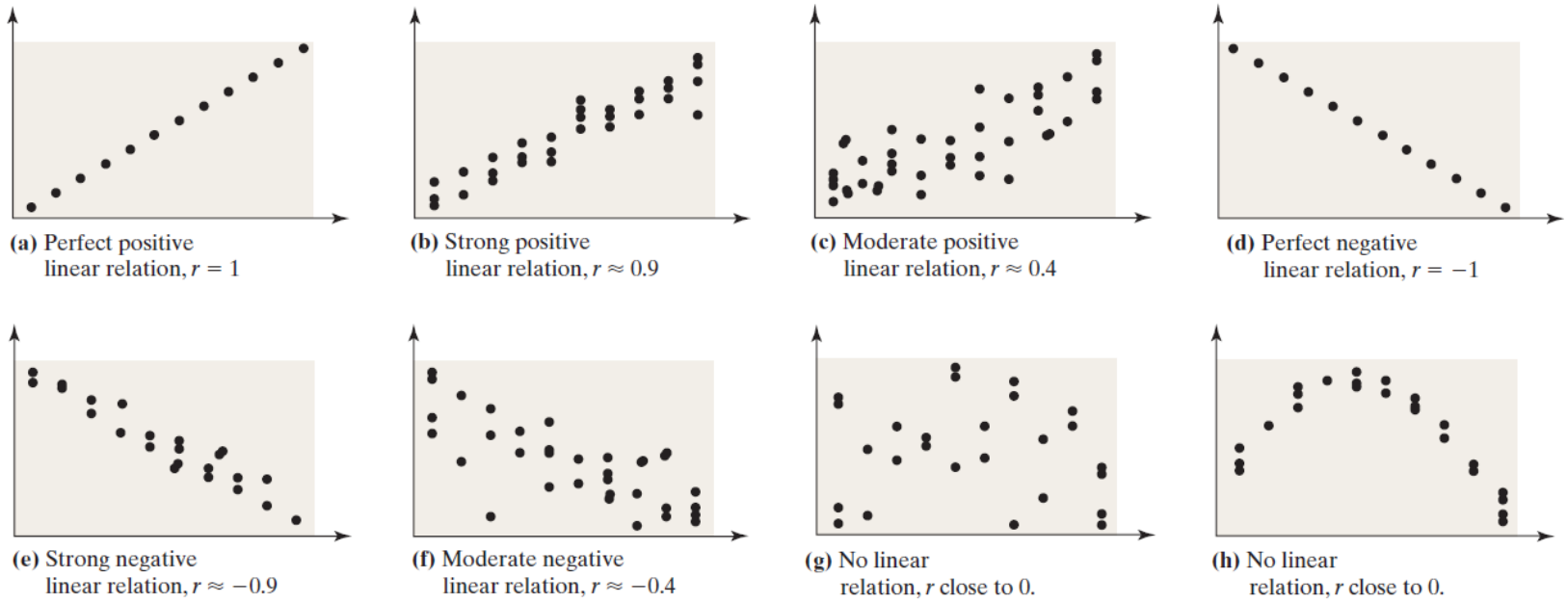
Scatterplot

- A graph that shows the relationship between two numerical variables measured on the same set of observations.
- Each observation is represented by a point in the scatter diagram (points are not connected).
- Explanatory variable is plotted on the x-axis (horizontal) and the response variable on the y-axis (vertical).
- Once plotted, we try to interpret the scatter diagram.
- The goal in interpreting is to distinguish scatter diagrams that imply a linear relation, a nonlinear relation, or no relation.



Correlation

- A positive correlation when high values of x tend to occur with high values of y, and when low values of x tend to occur with low values of y.
- And vice versa.
- A scatterplot > correlation coefficient > equation of the line, to further explore this relationship



(a) Perfect positive linear relation, $r = 1$

(b) Strong positive linear relation, $r \approx 0.9$

(c) Moderate positive linear relation, $r \approx 0.4$

(d) Perfect negative linear relation, $r = -1$

(e) Strong negative linear relation, $r \approx -0.9$

(f) Moderate negative linear relation, $r \approx -0.4$

(g) No linear relation, r close to 0.

(h) No linear relation, r close to 0.

Independent and dependent variables

- The response variable is the variable whose values can be explained by the values of the explanatory or predictor variables.
  - Explanatory variable is the input or independent variable.
  - Response variable is the output or dependent variable.
  - Does an increase in customer satisfaction explain or predict an increase in sales revenue?
  - Does vice versa make sense?
- A correlation exists between two variables if a particular value for one variable is more/less likely to occur with certain values of the other variable.
- How does this change for multivariate analysis?

Correlation does not imply causation

- Simply observing an association (or correlation) between two variables is not enough to imply a causal connection.
- Whenever two variables are associated, other variables may have influenced that association.
- A lurking variable is an unobserved variable that influences the association between the two variables of primary interest.
- Confounding is when two explanatory variables are both associated with a response variable but are also associated with each other.
- A lurking variable is not measured in the study. It has the potential for confounding.
- If it were included in the study and if it were associated both with the response variable and the explanatory variable, it would become a confounding variable.

Sampling Distributions and Significance Tests

- Central limit theorem
- Confidence intervals
- Normal distribution
- Long-tailed distribution
- Student's t distribution
- Binomial distribution
- Chi-Square distribution
- F distribution
- Poisson distribution
- A/B test
- Hypothesis test
- p-value, Type 1 error, Type 2 error
- t-test, t-distribution
- ANOVA, f-statistic
- Chi-square test
- Power and sample size

Statistical models vs Machine learning

- Isn’t regression a statistical model? - fairly dated debate, but you may still come across this.
  - “When we raise money it’s AI, when we hire it's machine learning, and when we do the work it's logistic regression” (by a statistician of course)
  - Short answer – just get the job done
- Model-based vs data-driven
  - In ML, training a linear regressor to predict = In SM, best fit line to minimise the squared error
  - SM starts with a hypothesis test and a set of rules (assumptions), ML is less rigid
  - SM takes all the data, ML separates training and test datasets – validation vs rigidity of modelling
  - But SM will also remove outliers or look for known distributions whereas ML doesn’t
  - SM focuses on causality through linearity, ML focuses on correlation through non-linearity
  - ML is more technology-ready (big data, cloud, pipelines)

A Universal Notation

- Descriptive, predictive and prescriptive analytics techniques can be represented using a universal notation:
- Findings = **model** + error
- y = -15.864x + 47131
- y = mean + standard deviation
- y = median + inter-quartile range
- Hypothesis test = test statistic + p-value
- Prediction = machine learning + accuracy
- Prediction = deep learning + F1 score
- Tagged photos = image classification AI + misclassification rate
- “Watch next” = recommender AI engine + false positives