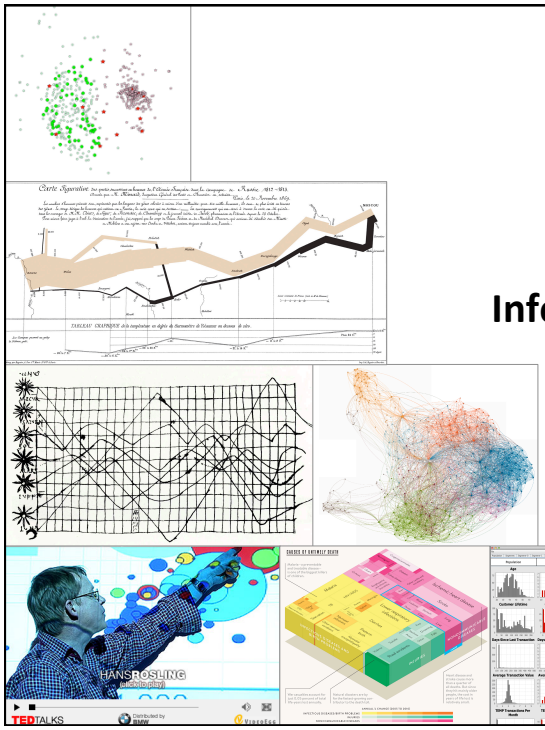


Evaluation Experiments

Information Visualisation
COMP40610

Dr. Brian Mac Namee

Origins

This course curates material from multiple online and published sources

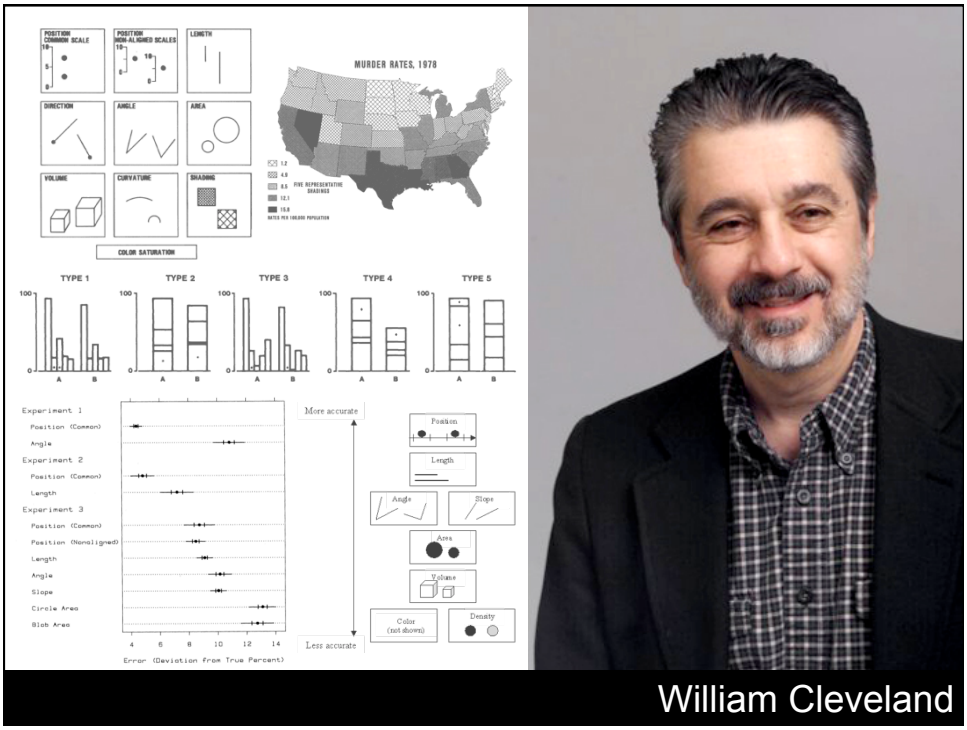
When this is the case full citations will be given

Agenda

In this lecture we will cover

- Visualisation experiment examples
- Designing visualisation experiments
- Case study

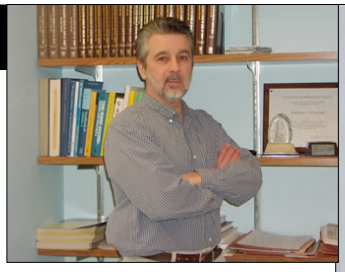
VISUALISATION EXPERIMENTS

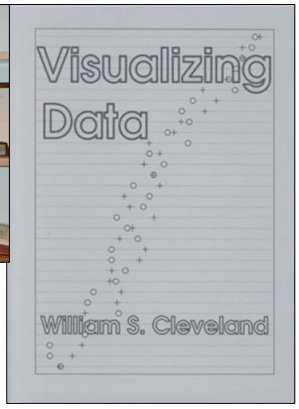


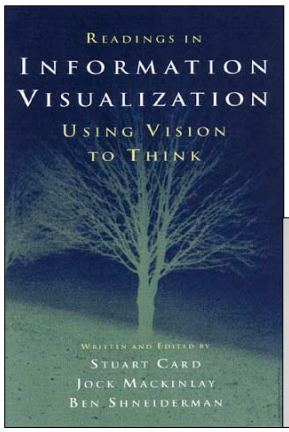
William Cleveland

Jock MacKinlay

Cleveland

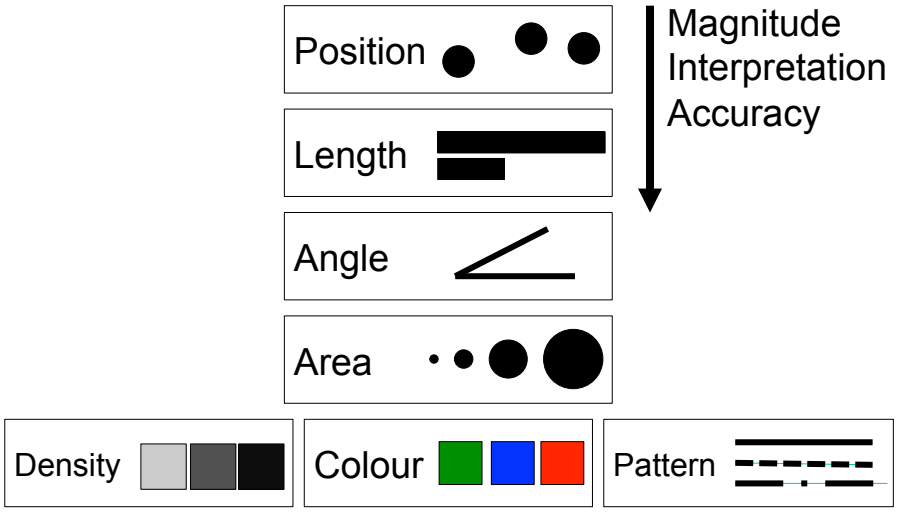




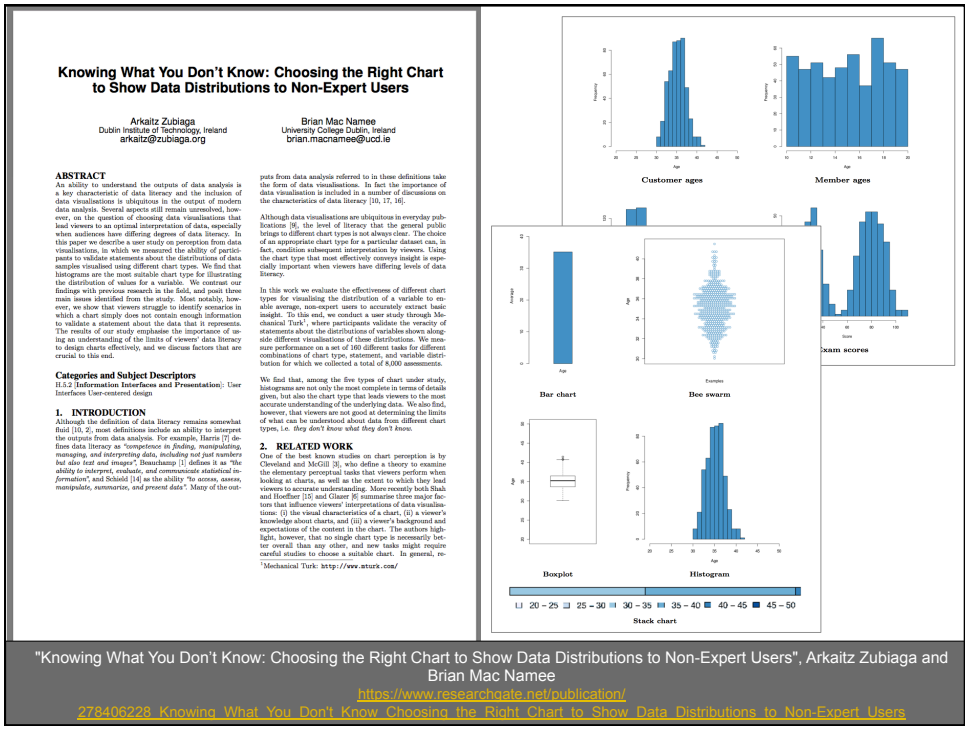


Mackinlay

Ease of Perception Of Encodings



Graphical perception: Theory, experimentation, and application to the development of graphical methods, W S Cleveland, R McGill, Journal of the American Statistical Association 01/1984; 79:531-554.
https://www.researchgate.net/publication/
229099907_Graphical_perception_Theory_experimentation_and_application_to_the_development_of_graphical_methods



Knowing What You Don't Know: Choosing the Right Chart to Show Data Distributions to Non-Expert Users

Arkaitz Zubiaga
Dublin Institute of Technology, Ireland
arkaitz@zubiaga.org

Brian Mac Namee
University College Dublin, Ireland
brian.macnamee@ucd.ie

ABSTRACT

An ability to understand the outputs of data analysis is a key characteristic of data literacy and the inclusion of data visualisations is ubiquitous in the output of modern data analysis. Several aspects still remain unresolved, however, on the question of choosing data visualisations that lead viewers to an optimal interpretation of data, especially when audiences have differing degrees of data literacy. In this paper we describe a user study on perception from data visualisations, in which we measured the ability of participants to validate statements about the distributions of data samples visualised using different chart types. We find that histograms are the most suitable chart type for illustrating the distribution of values for a variable. We contrast our findings with previous research in the field, and posit three main issues identified from the study. Most notably, however, we show that viewers struggle to identify scenarios in which a chart simply does not contain enough information to validate a statement about the data that it represents. The results of our study emphasise the importance of using an understanding of the limits of viewers' data literacy to design charts effectively, and we discuss factors that are crucial to this end.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces User-centered design

1. INTRODUCTION

Although the definition of data literacy remains somewhat fluid [10, 2], most definitions include an ability to interpret the outputs from data analysis. For example, Harris [7] defines data literacy as *"competence in finding, manipulating, managing, and interpreting data, including not just numbers but also text and images"*, Beauchamp [1] defines it as *"the ability to interpret, evaluate, and communicate statistical information"*, and Schield [14] as the ability *"to access, assess, manipulate, summarize, and present data"*. Many of the outputs from data analysis referred to in these definitions take the form of data visualisations. In fact the importance of data visualisation is included in a number of discussions on the characteristics of data literacy [10, 17, 16].

Although data visualisations are ubiquitous in everyday publications [9], the level of literacy that the general public brings to different chart types is not always clear. The choice of an appropriate chart type for a particular dataset can, in fact, condition subsequent interpretation by viewers. Using the chart type that most effectively conveys insight is especially important when viewers have differing levels of data literacy.

In this work we evaluate the effectiveness of different chart types for visualising the distribution of a variable to enable average, non-expert users to accurately extract basic insight. To this end, we conduct a user study through Mechanical Turk[1], where participants validate the veracity of statements about the distributions of variables shown alongside different visualisations of these distributions. We measure performance on a set of 100 different tasks for different combinations of chart type, statement, and variable distribution for which we collected a total of 8,000 assessments.

We find that, among the five types of chart under study, histograms are not only the most complete in terms of details given, but also the chart type that leads viewers to the most accurate understanding of the underlying data. We also find, however, that viewers are not good at determining the limits of what can be understood about data from different chart types, i.e. *they don't know what they don't know*.

2. RELATED WORK

One of the best known studies on chart perception is by Cleveland and McGill [3], who define a theory to examine the elementary perceptual tasks that viewers perform when looking at charts, as well as the extent to which they lead viewers to accurate understanding. More recently both Shah and Hoeffner [15] and Glaser [6] summarise three major factors that influence viewers' interpretations of data visualisations: (i) the visual characteristics of a chart, (ii) a viewer's knowledge about charts, and (iii) a viewer's background and expectations of the content in the chart. The authors highlight, however, that no single chart type is necessarily better overall than any other, and new tasks might require careful studies to choose a suitable chart. In general, re-

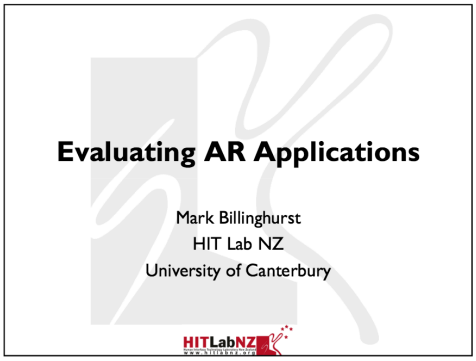
[1]Mechanical Turk: http://www.mturk.com/

"Knowing What You Don't Know: Choosing the Right Chart to Show Data Distributions to Non-Expert Users", Arkaitz Zubiaga and Brian Mac Namee
https://www.researchgate.net/publication/278406228_Knowing_What_You_Don't_Know_Choosing_the_Right_Chart_to_Show_Data_Distributions_to_Non-Expert_Users

DESIGNING VISUALISATION EXPERIMENTS

Evaluating AR Applications

A lot of the material in this presentation is based “Evaluating AR Applications” by Mark Billinghurst



Evaluation

Evaluation is concerned with gathering data about the usability of a design or product, by a specified group of users, for a particular activity, within a specified environment or work context

Why Evaluate?

Evaluations allow us to:

- Compare the effectiveness of approaches
- Explore aspects of visual perception
- Test usability (learnability, efficiency,...)
- Get feedback from users
- Better understand users
- Refine the design of visualisations

Types Of Evaluation

There are two key types of evaluations:

- **Controlled experiments**
  - Performed in a fully controlled, lab environment
  - Allows us to be very specific about what we test
- **Field studies**
  - Performed in a natural settings
  - Very good for understanding how users will actually interact with visualisations

Types Of Evaluation

There are two key types of evaluations:

- **Controlled experiments**
  - Performed in a fully controlled, lab environment
  - Allows us to be very specific about what we test
- **Field studies**
  - Performed in a natural settings
  - Very good for understanding how users will actually interact with visualisations

Designing Controlled Experiments

To design a controlled experiment we need to consider:

- Proposed hypothesis
- Measured variables
- Experimental methods
- Selected subjects
- Data collection
- Data analysis
- Managing the experiment

Designing Controlled Experiments

To design a controlled experiment we need to consider:

- **Proposed hypothesis**
- Measured variables
- Experimental methods
- Selected subjects
- Data collection
- Data analysis
- Managing the experiment

"Evaluating AR Applications", Mark Billinghurst
http://www.slideshare.net/marknb00/426-lecture7final

Proposed Hypothesis

What is the hypothesis being tested within an experiment?

- Easiest to think about as what is the question that I want to ask in this experiment
- Should be well defined in advance of the experiment
- Should be defined in terms of the variables to be used in the experiment

Designing Controlled Experiments

To design a controlled experiment we need to consider:

- Proposed hypothesis
- **Measured variables**
- Experimental methods
- Selected subjects
- Data collection
- Data analysis
- Managing the experiment

Measured Variables

There are two types of variables that are important in a controlled experiment:

- **Independent:** variables that are manipulated to create different experimental conditions
  - e.g. visual encodings used, colours used, interactions available, algorithm used
- **Dependent:** variables that are measured to find out the effects of changing the independent variables
  - e.g. speed of question answering, accuracy of question answering

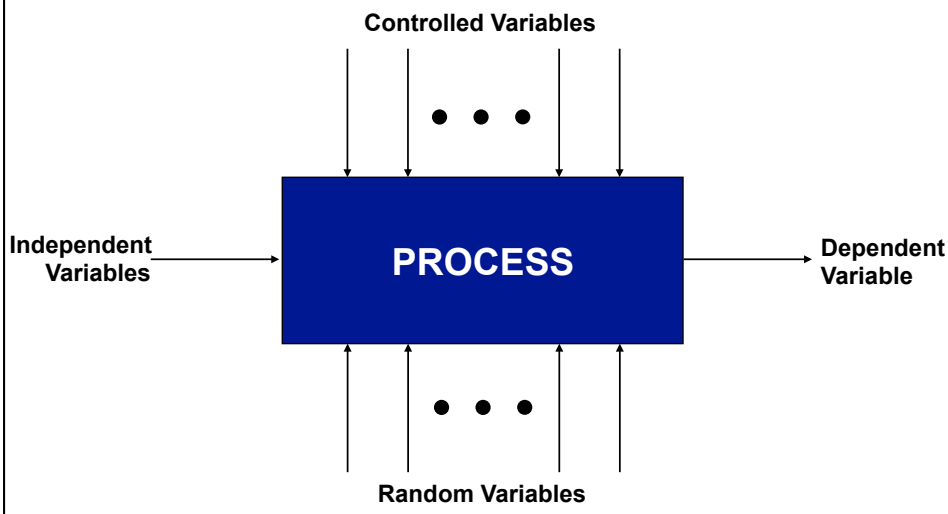
Measures Variables

Other Variables

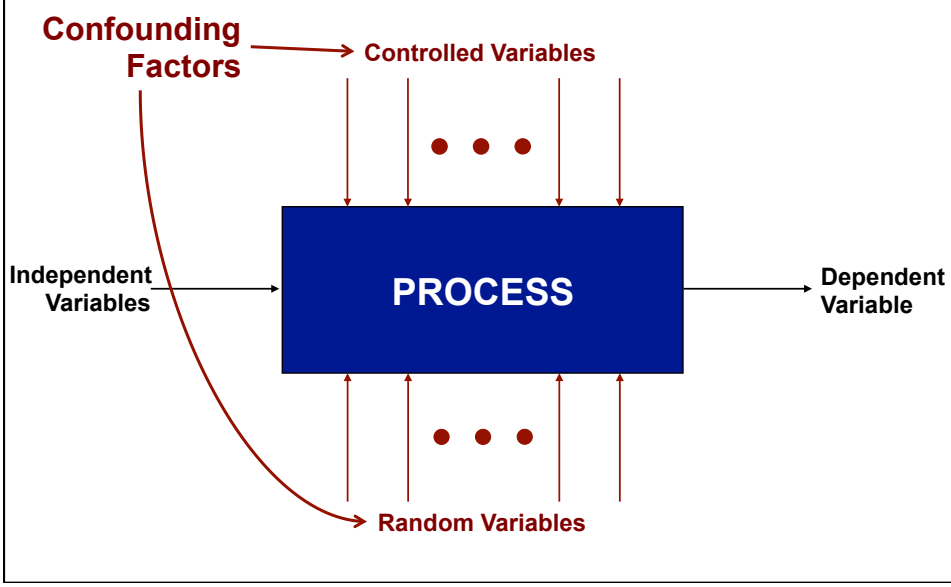
Other variables in an experiment that we do not measure, but which could have influence:

- **Controlled variables:** variables that we can control in an experiment but which do not relate to the hypothesis
  - e.g. room light, noise, …
- **Random variables:** variables that we can't control within an experiment, and do not measure
  - e.g. fatigue, …
- **Confounding variables:** variables that can interfere with measurement of measured variables
  - e.g. learning effects, previous experience, …

Measures Variables



Measures Variables



Designing Controlled Experiments

To design a controlled experiment we need to consider:

- Proposed hypothesis
- Measured variables
- **Experimental methods**
- Selected subjects
- Data collection
- Data analysis
- Managing the experiment

Experiments & Variables

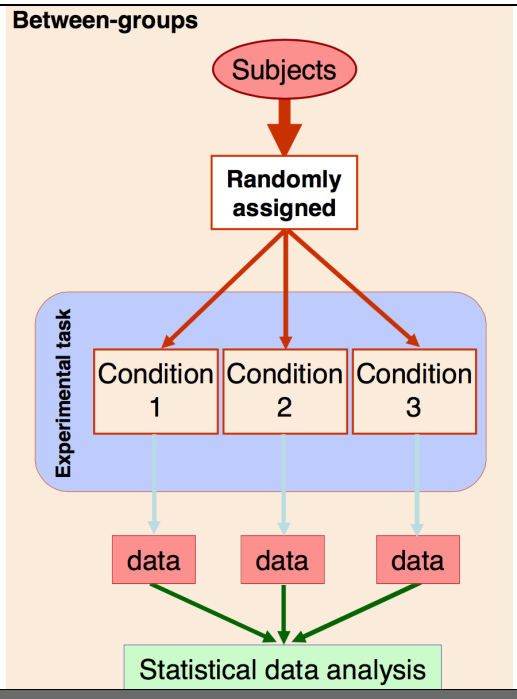
Experiments fundamentally explore the effect of different levels/values of independent variables on dependent variables

The **experiment conditions** are the levels/ values of independent variables used in an experiment

Experimental Methods

There are two key experimental methods we use:

- **Between-groups:** each subject is assigned to one experimental condition
- **Within-groups:** each subject performs under all the different conditions



"Evaluating AR Applications", Mark Billinghurst http://www.slideshare.net/marknb00/426-lecture7final

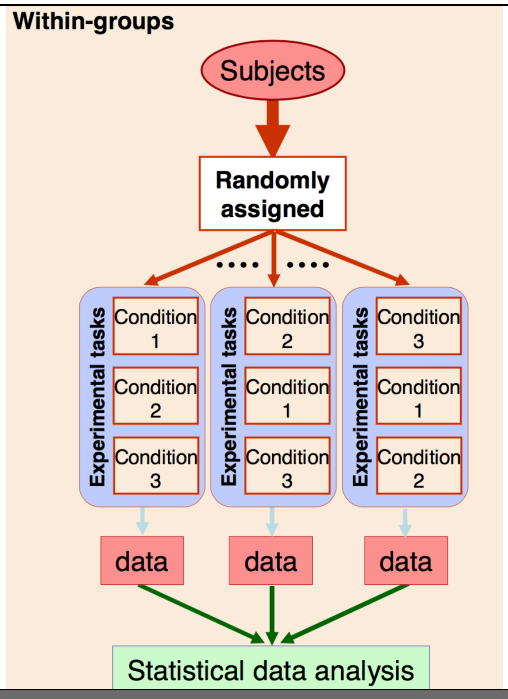
Between Groups

Advantages: Avoids interference effects (e.g. learning effects)

Disadvantages: Requires large number of subjects

Important: randomised assignment to conditions

Sometimes a factor must be tested between groups, e.g. gender, age



"Evaluating AR Applications", Mark Billinghurst http://www.slideshare.net/marknb00/426-lecture7final

Within Groups

Advantages: Less participants needed

Disadvantages: Requires thinking about the order of conditions to avoid learning effects

Sometimes a factor must be tested within groups, e.g. measuring learning effects

Within Groups

The **Latin Square** is a good way to order conditions in within group experiments to control for learning effects

3 X 3 Latin Square

A	B	C
B	C	A
C	A	B

4 x 4 Latin Square

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

4 x 4 Balanced Latin Square

A	B	C	D
B	D	A	C
D	C	B	A
C	A	D	B

In a balanced Latin Square each condition both precedes and follows each other condition an equal number of times

Designing Controlled Experiments

To design a controlled experiment we need to consider:

- Proposed hypothesis
- Measured variables
- Experimental methods
- **Selected subjects**
- Data collection
- Data analysis
- Managing the experiment

"Evaluating AR Applications", Mark Billinghurst
http://www.slideshare.net/marknb00/426-lecture7final

Selected Subjects

The subjects selected for a visualisation experiment should be **representative** of the population the hypothesis refers to

- Think about things like age, gender, level of education, …
- But balance this with who is available

Selected Subjects

Sample size should be large enough to show any effects of interest

- Always > 10
- Large enough for smallest interesting facet to be > 10
- Smaller effects require larger sample sizes

Selected Subjects

Mechanical Turk (www.mturk.com) offers a convenient, quick way to source subjects

- Only works for certain hyoptheses
- Parallel in-person experiments reinforce results
- Be very careful of representativeness of participants
- Validate performance of participants

"Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design", Jeffrey Heer & Michael Bostock
http://vis.stanford.edu/files/2010-MTurk-CHI.pdf

Designing Controlled Experiments

To design a controlled experiment we need to consider:

- Proposed hypothesis
- Measured variables
- Experimental methods
- Selected subjects
- **Data collection**
- Data analysis

Data Collection

Before an experiment starts we need to make decisions about all the data that will be collected

Observations are gathered in two ways

- **Manually** (human observers)
- **Automatically** (cameras, sensors, etc.)

A measurement is a recorded observation

- **Objective** measurements
- **Subjective** measurements

Objective Measurements

Typical objective measurements include:

- task completion time
- errors (number, percent,...)
- percent of task completed
- ratio of successes to failures
- number of repetitions
- number of commands used
- number of failed commands
- physiological data (heart rate,...)

Subjective Measurements

Typical subjective measurements include

- user satisfaction
- ease of use
- intuitiveness
- judgments

Designing Controlled Experiments

To design a controlled experiment we need to consider:

- Proposed hypothesis
- Measured variables
- Experimental methods
- Selected subjects
- Data collection
- **Data analysis**
- Managing the experiment

"Evaluating AR Applications", Mark Billinghurst
http://www.slideshare.net/marknb00/426-lecture7final

Data Analysis

Once collected we can use statistical techniques to analyse data

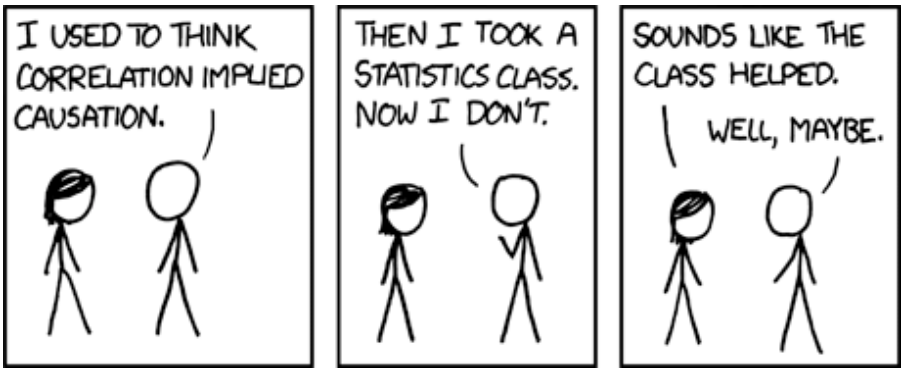
- Comparing between two results
  - Unpaired T-Test (for between groups – assumes normal distribution)
  - Paired T-Test (for within groups – assumes normal distribution)
  - Mann–Whitney U (independent samples)
- Comparing between > two results
  - Analysis of Variance – ANOVA
  - Followed by post-hoc analysis – Bonferroni Test
  - Kruskal– Wallis (does not assume normal distribution)

Choosing A Statistical Test

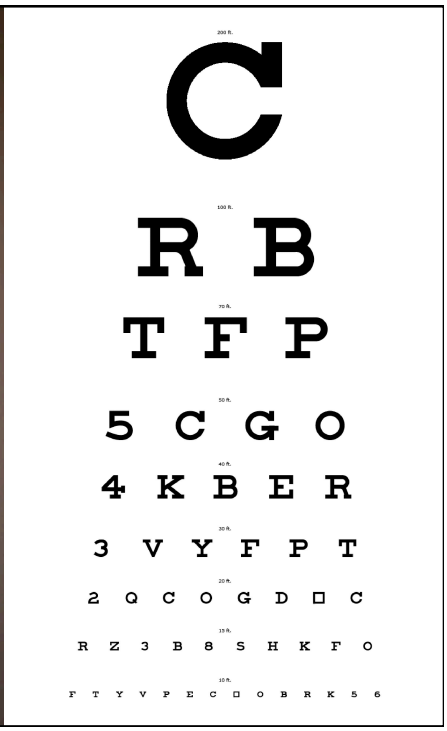
What do you want to do?	Types of your dependant variables		
	Interval/Ratio (normality assumed)	Interval/Ratio (normality not assumed), ordinal	Dichotomy (binomial)
Compare two unpaired groups	Unpaired t test	Mann-Whitney test	Fisher's test
Compare two paired groups	Paired t test	Wilcoxon test	McNemar's test
Compare more than two unmatched groups	ANOVA	Kruskal-Wallis test	Chi-square test
Compare more than two matched groups	Repeated-measures ANOVA	Friedman test	Cochran's Q test
Find relationships between two variables	Pearson correlation	Spearman correlation	Cramer's V
Predict a value with one independent variable	Linear/Non-linear regression	Non-parametric regression	Logistic regression
Predict a value with multiple independent variables or binomial variables	Multiple Linear /Non-linear regression		Multiple logistic regression

"Statistics for HCI Research", GUSTAVO ROVELO
https://iira.ai2.upv.es/confluence/download/attachments/13631509/WGM+38+-+Statistics+for+HCI+Research+I.pdf

Correlation & Causation

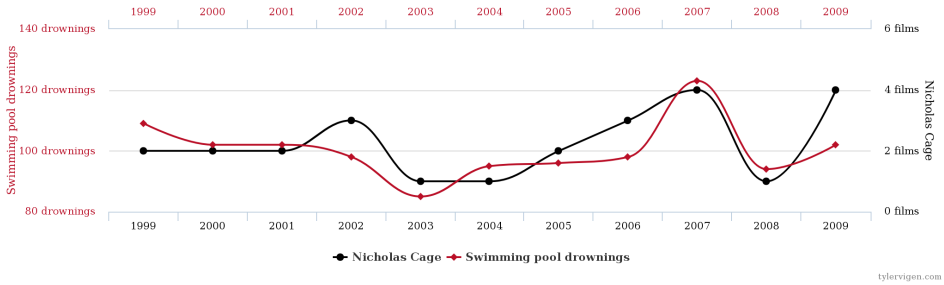


"Correlation", xkcd
www.xkcd.com/552/



Correlation & Causation

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in



Spurious Correaltion
http://www.tylervigen.com/spurious-correlations

Designing Controlled Experiments

To design a controlled experiment we need to consider:

- Proposed hypothesis
- Measured variables
- Experimental methods
- Selected subjects
- Data collection
- Data analysis
- **Managing the experiment**

"Evaluating AR Applications", Mark Billinghurst
http://www.slideshare.net/marknb00/426-lecture7final

Pilot Studies

A **pilot study** is a small version of an experimental study used to test the design of the study

- Is the plan viable?
- Are the questions correct?
- Are you collecting the correct data?
- Will statistical analysis work?

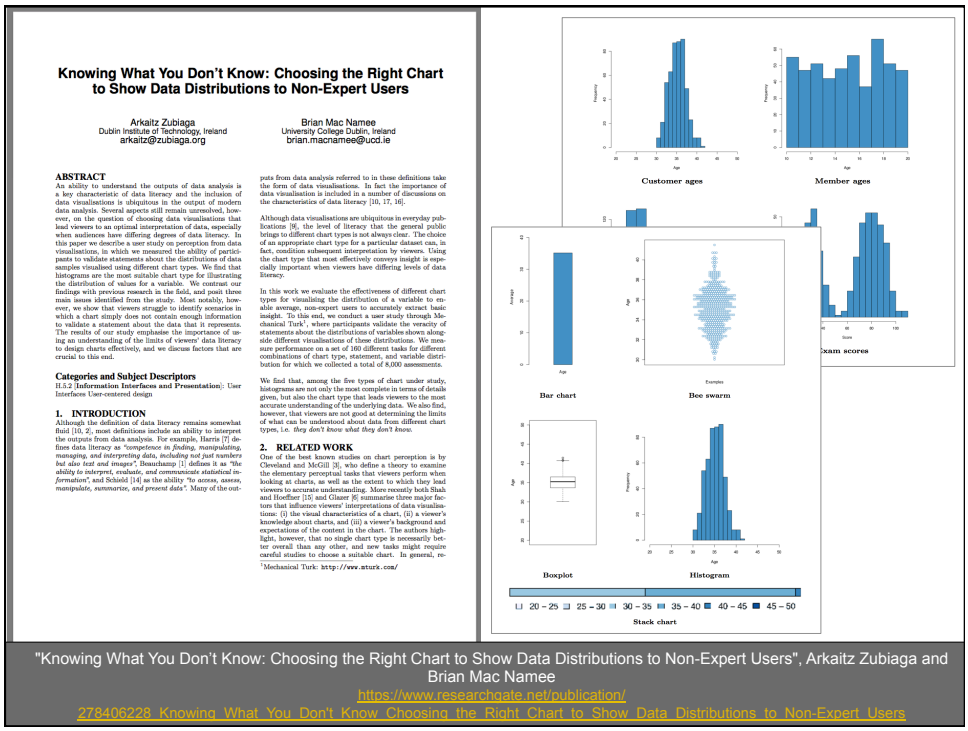
Should be run with 5-10 participants before the main study

Managing An Experiment

When running an experiment it is important that every participanbt has as consistent an experience as possible

- Write scripts for everything
- Use checklists
- Treat participants nicely
- Take notes about any strange behaviour
- Take the role of a friendly waiter

CASE STUDY



Knowing What You Don't Know: Choosing the Right Chart to Show Data Distributions to Non-Expert Users

Arkaitz Zubiaga
Dublin Institute of Technology, Ireland
arkaitz@zubiaga.org

Brian Mac Namee
University College Dublin, Ireland
brian.macnamee@ucd.ie

ABSTRACT

An ability to understand the outputs of data analysis is a key characteristic of data literacy and the inclusion of data visualisations is ubiquitous in the output of modern data analysis. Several aspects still remain unresolved, however, on the question of choosing data visualisations that lead viewers to an optimal interpretation of data, especially when audiences have differing degrees of data literacy. In this paper we describe a user study on perception from data visualisations, in which we measured the ability of participants to validate statements about the distributions of data samples visualised using different chart types. We find that histograms are the most suitable chart type for illustrating the distribution of values for a variable. We contrast our findings with previous research in the field, and posit three main issues identified from the study. Most notably, however, we show that viewers struggle to identify scenarios in which a chart simply does not contain enough information to validate a statement about the data that it represents. The results of our study emphasise the importance of using an understanding of the limits of viewers' data literacy to design charts effectively, and we discuss factors that are crucial to this end.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces User-centered design

1. INTRODUCTION

Although the definition of data literacy remains somewhat fluid [10, 2], most definitions include an ability to interpret the outputs from data analysis. For example, Harris [7] defines data literacy as *"competence in finding, manipulating, managing, and interpreting data, including not just numbers but also text and images"*, Beauchamp [1] defines it as *"the ability to interpret, evaluate, and communicate statistical information"*, and Schield [14] as the ability *"to access, assess, manipulate, summarize, and present data"*. Many of the outputs from data analysis referred to in these definitions take the form of data visualisations. In fact the importance of data visualisation is included in a number of discussions on the characteristics of data literacy [10, 17, 16].

Although data visualisations are ubiquitous in everyday publications [9], the level of literacy that the general public brings to different chart types is not always clear. The choice of an appropriate chart type for a particular dataset can, in fact, condition subsequent interpretation by viewers. Using the chart type that most effectively conveys insight is especially important when viewers have differing levels of data literacy.

In this work we evaluate the effectiveness of different chart types for visualising the distribution of a variable to enable average, non-expert users to accurately extract basic insight. To this end, we conduct a user study through Mechanical Turk[1], where participants validate the veracity of statements about the distributions of variables shown alongside different visualisations of these distributions. We measure performance on a set of 100 different tasks for different combinations of chart type, statement, and variable distribution for which we collected a total of 8,000 assessments.

We find that, among the five types of chart under study, histograms are not only the most complete in terms of details given, but also the chart type that leads viewers to the most accurate understanding of the underlying data. We also find, however, that viewers are not good at determining the limits of what can be understood about data from different chart types, i.e. *they don't know what they don't know*.

2. RELATED WORK

One of the best known studies on chart perception is by Cleveland and McGill [3], who define a theory to examine the elementary perceptual tasks that viewers perform when looking at charts, as well as the extent to which they lead viewers to accurate understanding. More recently both Shah and Hoeffner [15] and Glaser [6] summarise three major factors that influence viewers' interpretations of data visualisations: (i) the visual characteristics of a chart, (ii) a viewer's knowledge about charts, and (iii) a viewer's background and expectations of the content in the chart. The authors highlight, however, that no single chart type is necessarily better overall than any other, and new tasks might require careful studies to choose a suitable chart. In general, re-

[1]Mechanical Turk: http://www.mturk.com/

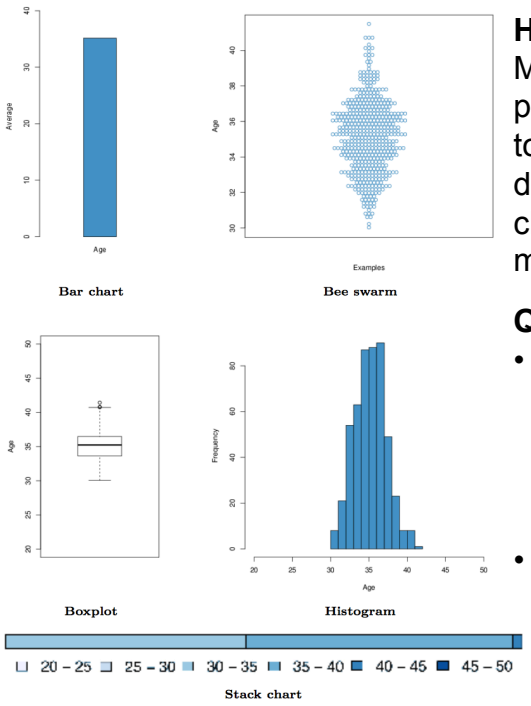
"Knowing What You Don't Know: Choosing the Right Chart to Show Data Distributions to Non-Expert Users", Arkaitz Zubiaga and Brian Mac Namee
https://www.researchgate.net/publication/278406228_Knowing_What_You_Don't_Know_Choosing_the_Right_Chart_to_Show_Data_Distributions_to_Non-Expert_Users

Designing Controlled Experiments

To design a controlled experiment we need to consider:

- Proposed hypothesis
- Measured variables
- Experimental methods
- Selected subjects
- Data collection
- Data analysis

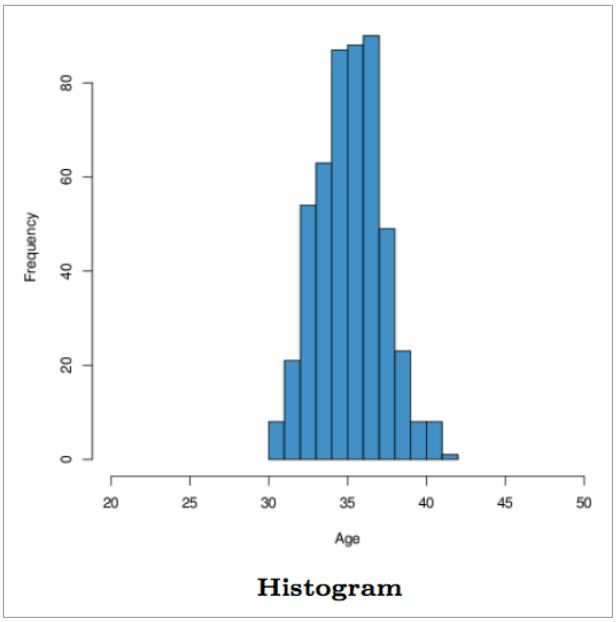
"Evaluating AR Applications", Mark Billinghurst
http://www.slideshare.net/marknb00/426-lecture7final

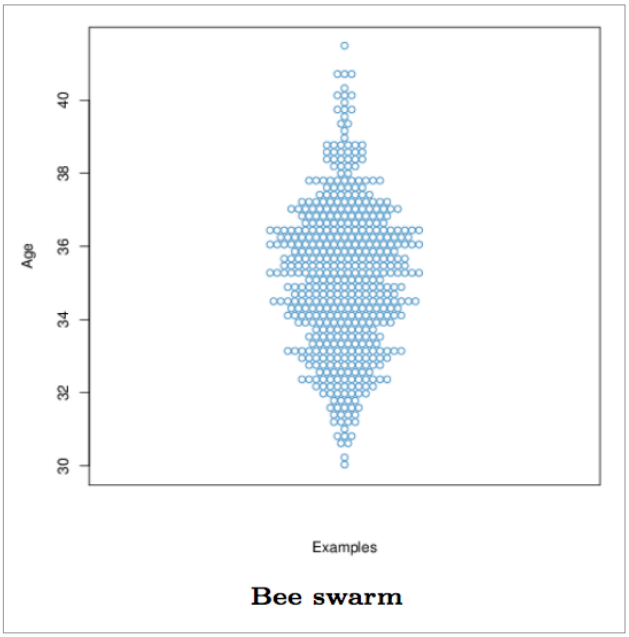


Hypothesis:
Members of the general public have differing ability to read details of data distributions from different chart types - some are more effective than others

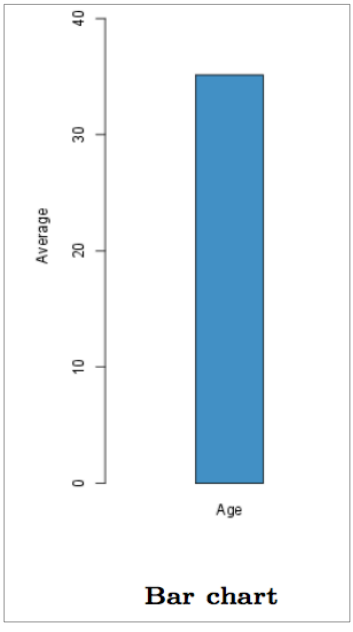
Questions:

- What chart types are best for displaying distributions of variables to members of the general public?
- Can people identify when enough data is not present in a chart to answer a question?

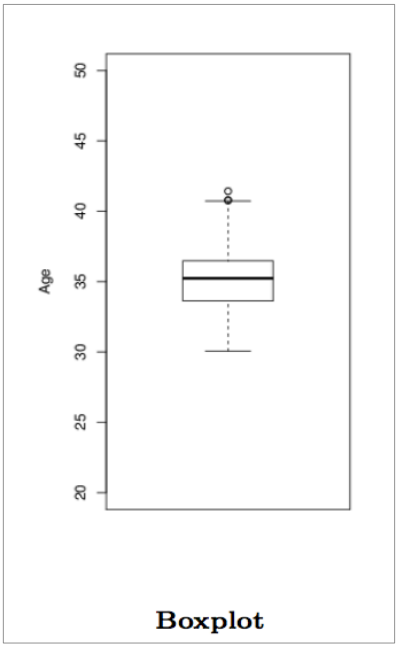




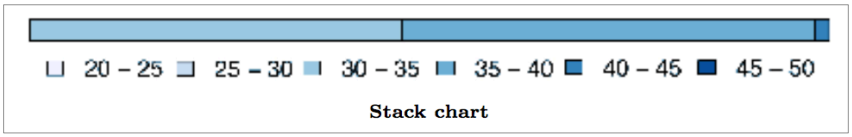
Bee swarm



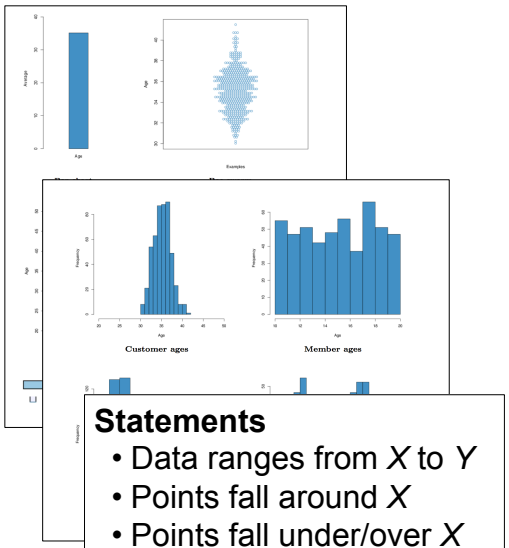
Bar chart



Boxplot



Stack chart



Statements

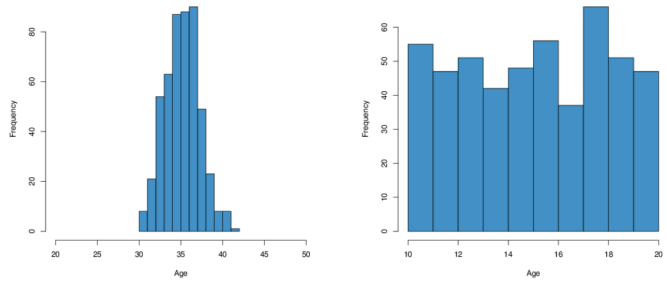
- Data ranges from *X* to *Y*
- Points fall around *X*
- Points fall under/over *X*
- Points clustered to either side of *X*

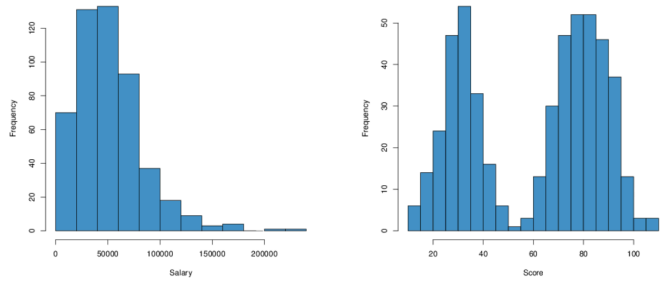
Dependent Variables:

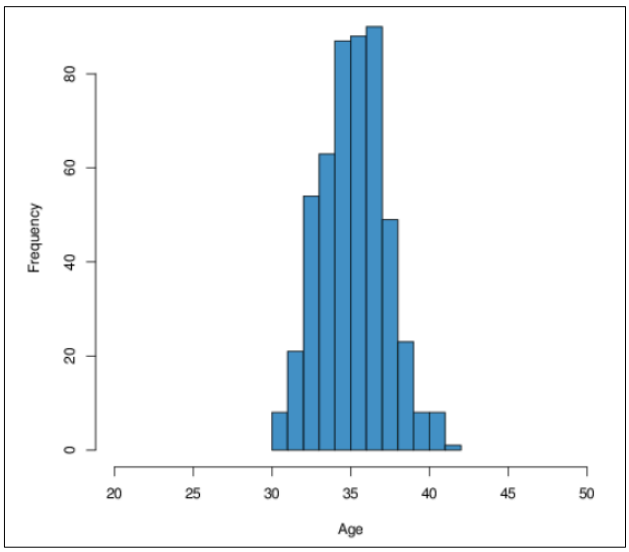
- People's ability to read the details of a variable's distribution from a chart

Independent Variables:

- Different chart types: bar chart, bee swarm, boxplot, histogram, stack chart
- Different data distributions: normal, skewed, uniform, bimodal
- Different questions: range, central tendency, upper/lower limit, clustering

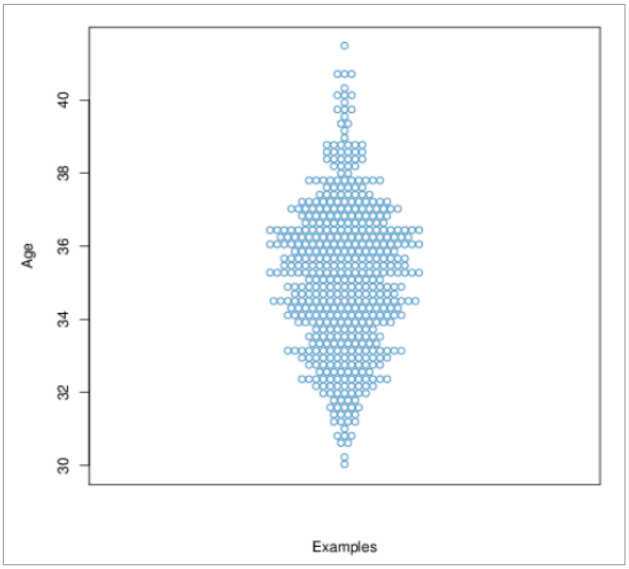






The data ranges from 25 to 45

strongly disagree, disagree, neutral, agree, strongly agree

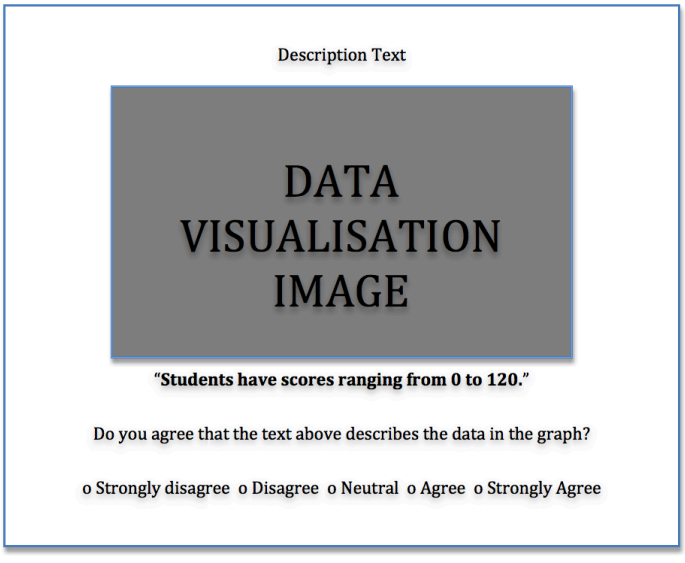


Customers have an average age of 45

strongly disagree, disagree, neutral, agree, strongly agree

Experimental Method

- The combinations of variables (4), chart types (5), and statements (8) amounted to a total of 160 different experimental conditions.
- Participants in our experiments were shown one task at a time
  - Rated the accuracy of the statement shown on a 5 point Likert scale
  - Additionally, could opt for an alternative choice *impossible to tell from this chart*.
- Within-groups design - every participant performed every experimental condition.
  - Tasks were presented in random order to control for learning effects.



Selected Subjects

- Sourced subjects from Amazon Mechanical Turk
- Participants did not need to have any prior expertise in data analytics
- Restricted participation to US-based participants to control for English language capability
- Restricted to participants with at least a 95% HIT acceptance rate
- Used 50 participants
- With 50 ratings collected for each of the 160 tasks, we gathered a total of 8,000 ratings.

Statements	
Data ranges from X to Y	0.416 (moderate)
Points fall around X	0.304 (fair)
Points fall under/over X	0.440 (moderate)
Points clustered to either side of X	0.360 (fair)
Charts	
Bar chart (average)	0.232 (fair)
Bee swarm	0.495 (moderate)
Boxplot	0.313 (fair)
Stack chart	0.211 (fair)
Histogram	0.479 (moderate)
Variables	
Online movie customer ages	0.442 (moderate)
Youth sports centre ages	0.413 (moderate)
Salaries	0.391 (fair)
Student scores	0.288 (fair)
Overall Inter-rater agreement	**0.390 (fair)**

Statements	
Data ranges from X to Y	0.900
Points fall around X	0.550
Points fall under/over X	0.750
Points clustered to either side of X	0.525
Charts	
Bar chart (average)	0.531
Bee swarm	0.906
Boxplot	0.563
Stack chart	0.438
Histogram	0.969
Variables	
Online movie customer ages	0.700
Youth sports centre ages	0.700
Salaries	0.750
Student scores	0.575
Overall accuracy	**0.681**

		Responses			
		Imp.	**False**	**Neutral**	**True**
Ground Truth	**Imp.**	**23.8**	45.9	5.4	24.9
	False	6.9	**72.1**	6.4	14.6
	True	4.7	14.0	5.5	**75.8**
Precision		67.2	54.6	-	65.7

SUMMARY

Summary

Controlled evaluation experiments are a great way to understand the effectiveness of different types of visualisation

- Plan carefully
- Record lots of data
- Objective measures work better than subjective ones
- Use pilot experiments