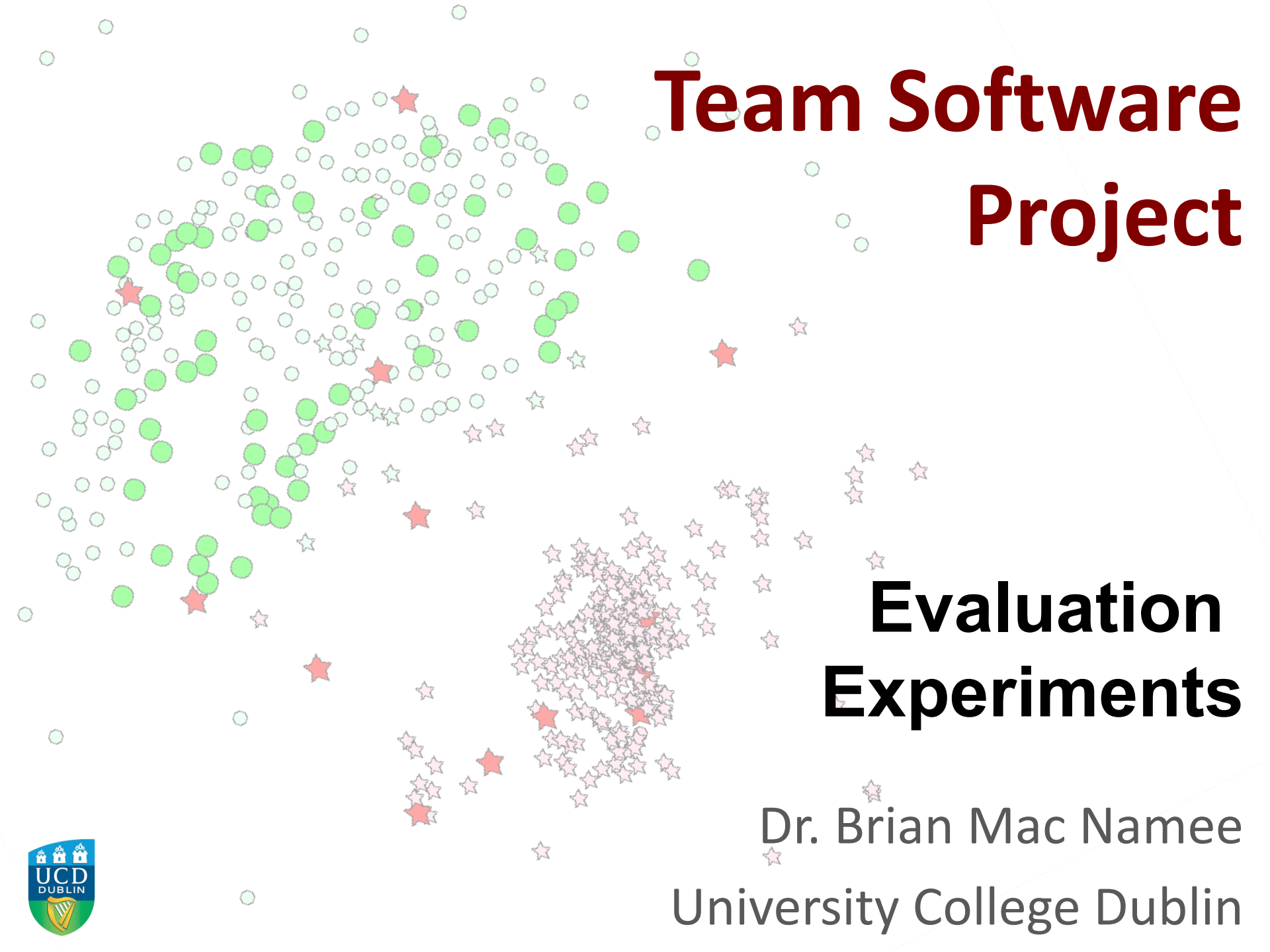


Team Software Project

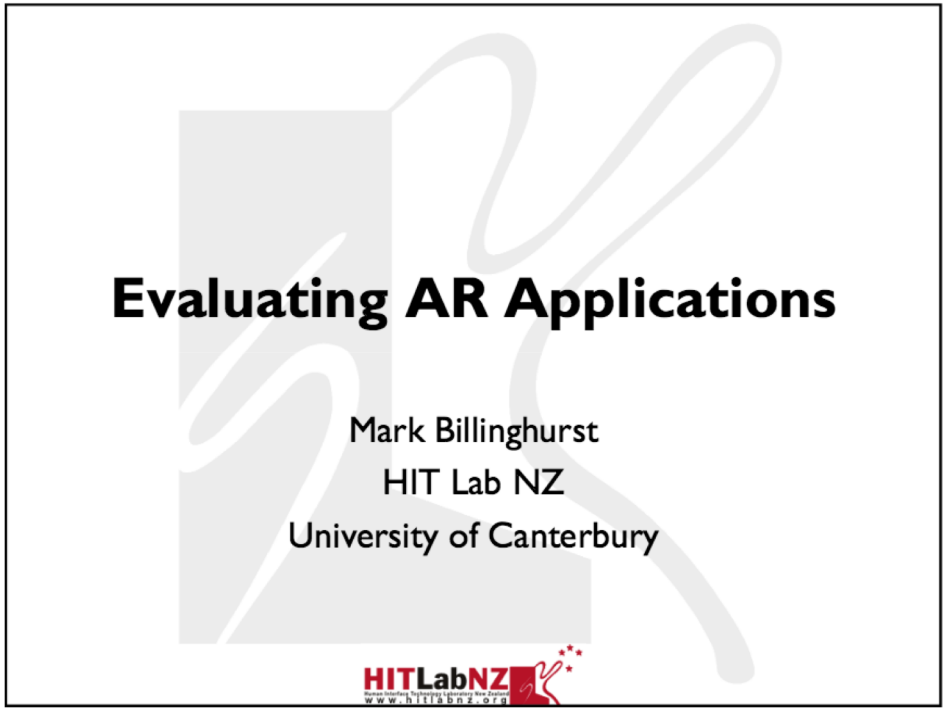
Evaluation Experiments

Dr. Brian Mac Namee

University College Dublin

Evaluating AR Applications

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



INTRODUCTION

Testing

≠

Evaluation

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
- Test usability (learnability, efficiency,...)
- Get feedback from users
- Better understand users
- Refine the design of the user interface for an application

Types Of Evaluation

There are different types of evaluations:

- **Desk experiments**
  - Evaluate the performance of a key part of your system
    - For example the accuracy of a classifier, the speed of an algorithm, the ability of a system to retrieve data
  - Take careful design and typically require some kind of ground truth dataset
  - Doesn't tell us anything about usability or the downstream impact of the performance of this component

Types Of Evaluation

There are different types of evaluations:

- **Surveys**
  - Good for getting feedback from large groups on things like required features, need for a system etc
  - Notoriously unreliable
    - “If I asked people what they wanted they would have said a faster horse!” Henry Ford[1]
  - Surveys take very careful design or can be a waste of time[2]

[1] Henry ford probably never said this!
https://hbr.org/2011/08/henry-ford-never-said-the-fast

[2] Yes Minister, Opinion Poll Design
https://www.youtube.com/watch?v=G0ZZJXw4MTA

Types Of Evaluation

There are different 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 an application

Types Of Evaluation

There are different 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 an application

Simple Comparisons In Controlled Experiments

Two key types of comparison experiment

- Test against an **incumbunt**
  - Test a new approach against the current standard approach to something
- Test against a **straw man**
  - Test a new approach against the simplest, yet realistic way that something could be done

CONTROLLED EXPERIMENTS

Designing Controlled Experiments

To design a controlled experiment we need to consider:

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

Designing Controlled Experiments

To design a controlled experiment we need to consider:

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

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:

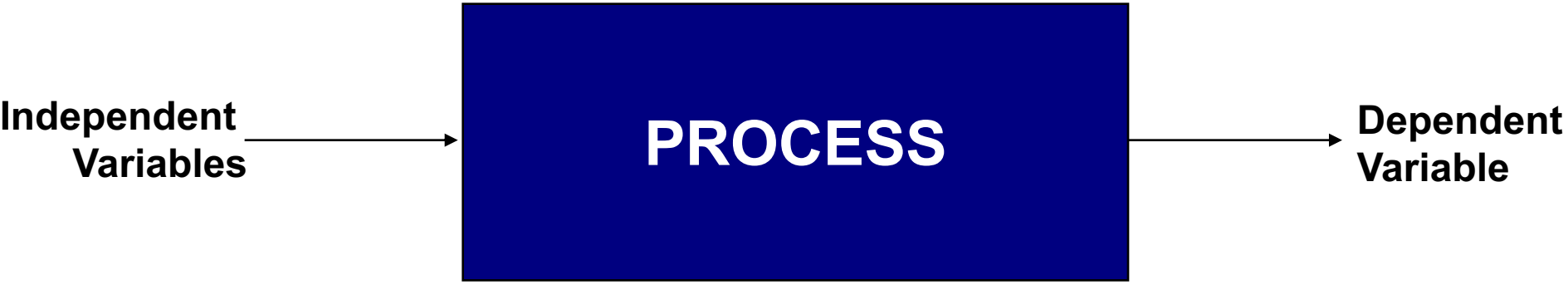
- Proposed hypothesis
- **Measured variables**
- Selected subjects
- Data collection
- Data analysis

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. interface used, algorithm used, features provided, …
- **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

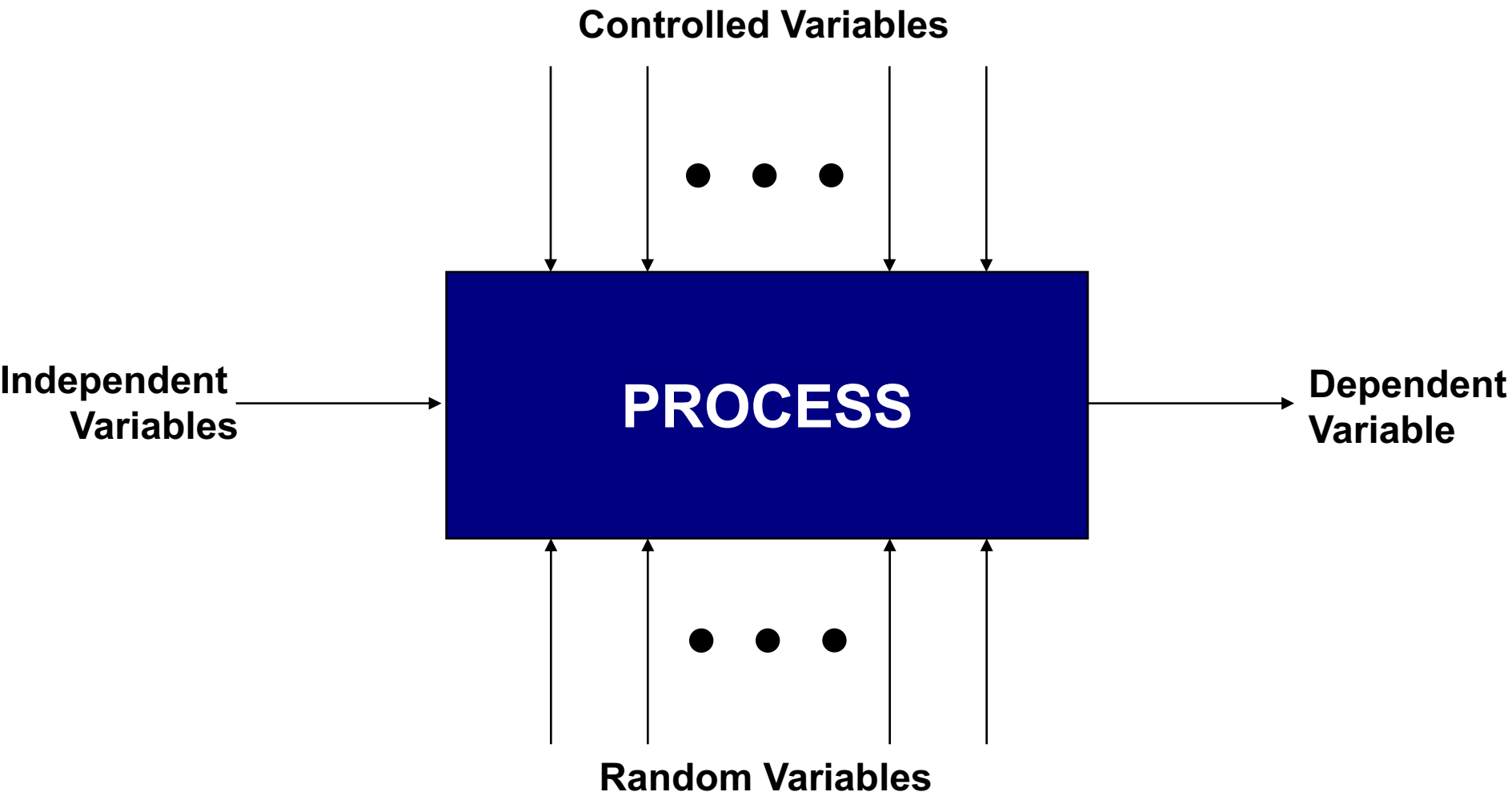


Other Variables

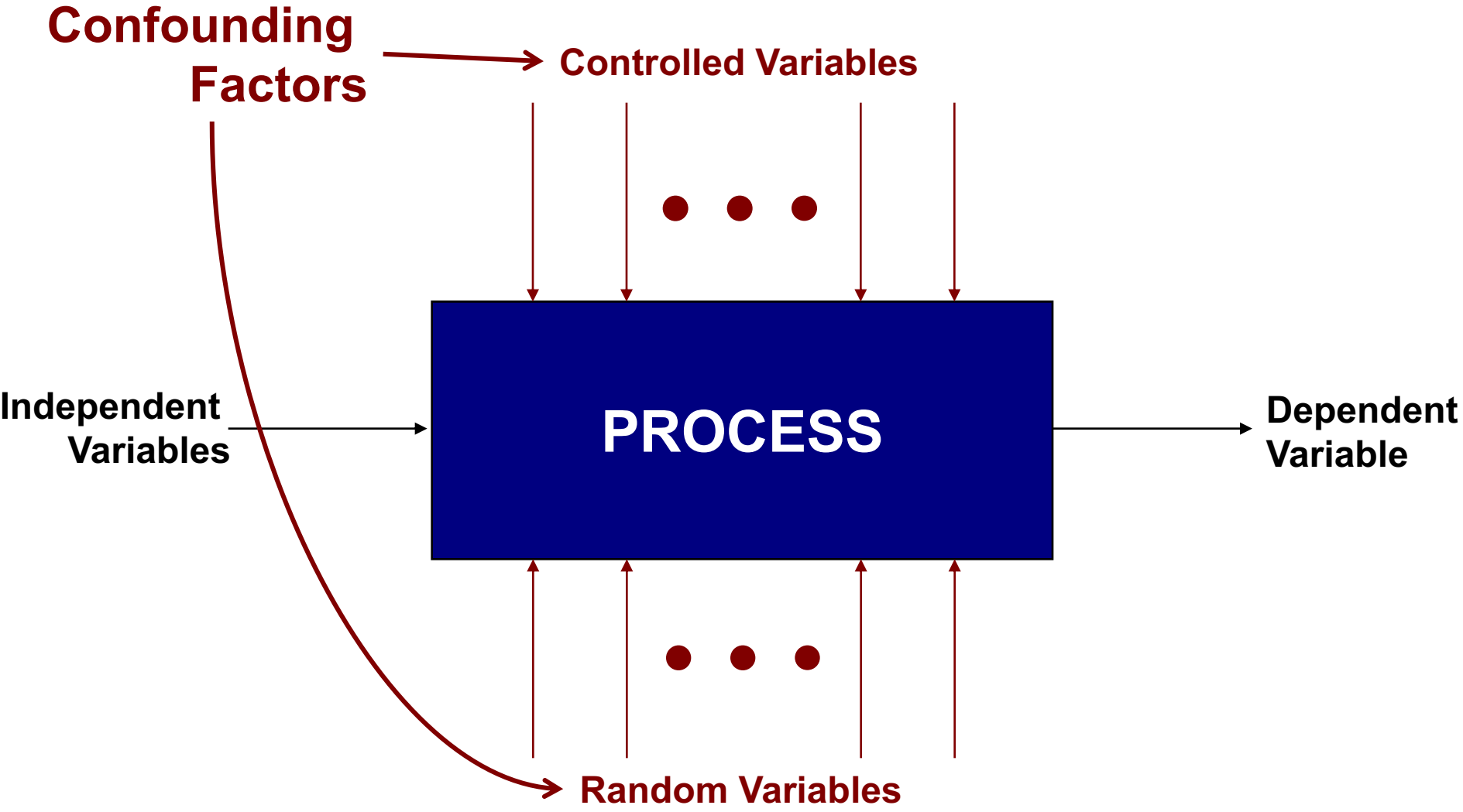
Other variables in an experiment that we do not measure, but which could have influence:

- **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
- **Selected subjects**
- Data collection
- Data analysis

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

Designing Controlled Experiments

To design a controlled experiment we need to consider:

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

Data Analysis

Once collected we can use statistical techniques to analyse data

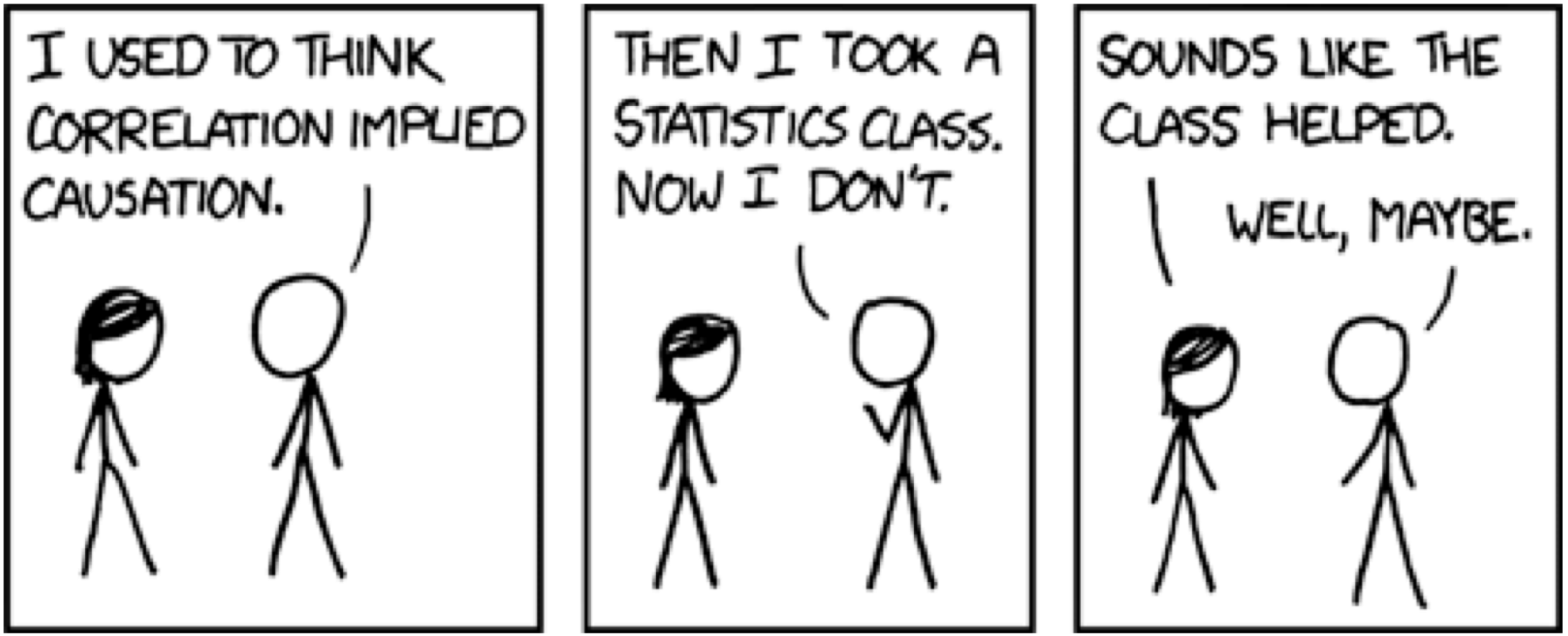
- Comparing between two results
  - Unpaired T-Test (for between groups – assumes normal distribution)
  - Paired T-Test (for within groups – assumes normal distribution)
  - Mann–Whitney U (independent samples)
- Comparing between > two results
  - Analysis of Variance – ANOVA
  - Followed by post-hoc 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://jira.ai2.upv.es/confluence/download/attachments/13631509/WGM+38+-+Statistics+for+HCI+Research+I.pdf

Correlation & Causation



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

200 ft.

C

100 ft.

R B

70 ft.

T F P

50 ft.

5 C G O

40 ft.

4 K B E R

30 ft.

3 V Y F P T

20 ft.

2 Q C O G D □ C

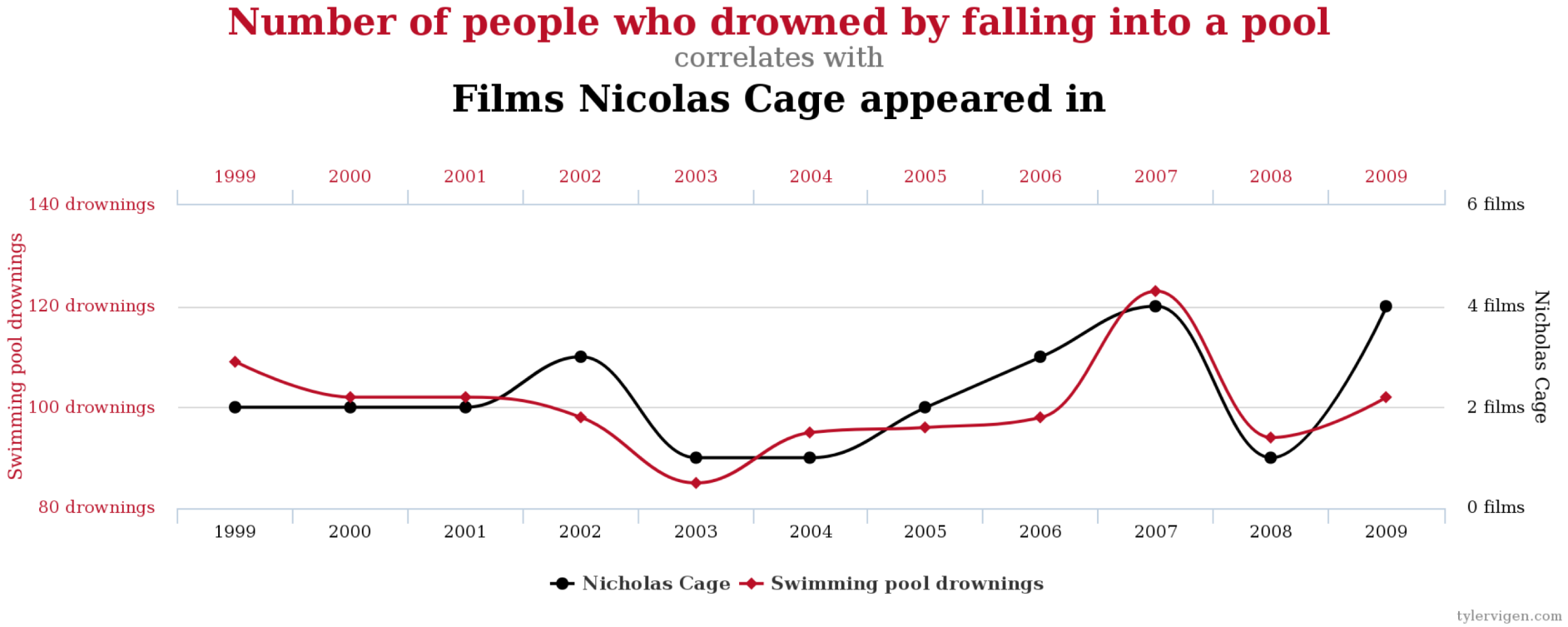
15 ft.

R Z 3 B 8 S H K F O

10 ft.

F T Y V P E C □ O B R K 5 6

Correlation & Causation



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

SUMMARY

Summary

Controlled evaluation experiments are a great way to understand the effectiveness of applications

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

Previous Evaluation Experiments

XXXNews

- "Is there a statistically difference in the average completion time and mean accuracy levels between news readers who use XXXNews and users using other news delivery platforms (BBC news, CNN news, ABC news and New York Times)".
- Controlled experiment with ~20 participants
  - Independent variable: systems used
  - Dependent variable: time taken to find information

Previous Evaluation Experiments

YYYYNews

- Desk evaluation of sentiment analysis technique
  - Required ground truth dataset – manually labelled by users
- Controlled experiment with ~30 participants
  - Independent variable: sentiment filtering used
  - Dependent variable: choice between three conditions