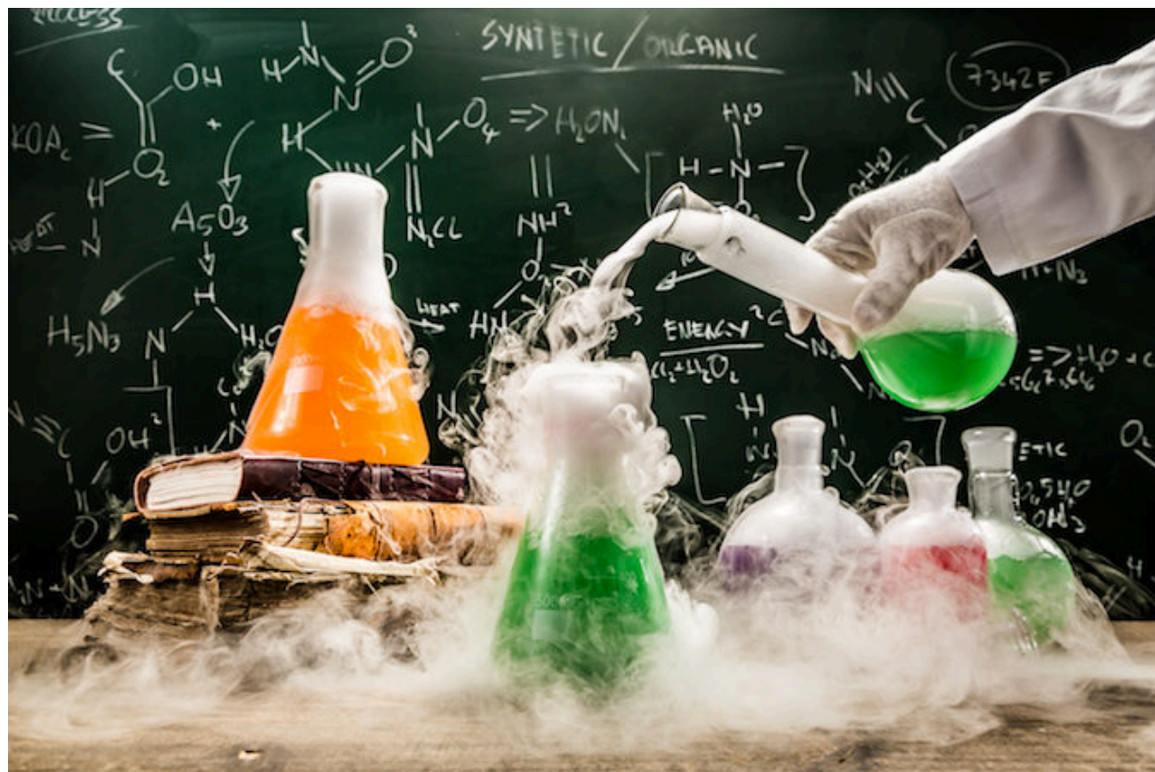
Project Evaluation

F21RP - Research Methods and Project Planning

(based on slides by Oliver Lemon, Judy Robertson, Helen Hastie, Frank Broz and Tom Methven)

Learning Outcomes

- Why do we need to evaluate?
- How can we evaluate?
  - Case Studies, Testing
  - Experimental Design
  - Empirical Evaluation
  - Human Evaluation



Q: Why bother evaluating?

A: The stakeholders (aka you/supervisor/clients/users) of the project need to know if your method/algorithm/model/system

- is successful
- is *comparable/better/worse* compared to the competition

$\mathcal{Q}$: How do you formulate your evaluation?

$\mathcal{A}$: By defining your research questions (usually in the Introduction).

- These are the main things you want to investigate in your project. For example:
  - *How do compressed representations improve performance on a machine learning task?*
  - *How do users respond to different dialogue flows?*
  - *How does technology effect the consumption of visual art?*
  - *How do different visualisations of uncertainty affect the trust of the user of an autonomous system?*

$\mathcal{Q}$: How do you answer your Research Questions?

$\mathcal{A}$-1: By stating a thesis.

- A thesis statement is a short, direct sentence that summarises the main points or claims of your project.
  - For example:
    - *Using transfer learning based on large pre-trained models, we can achieve reliable text generation systems using significantly less data.*

AND optionally

$\mathcal{A}$-2: By formulating one or more hypotheses.

- A hypothesis is a statement that can be proved or disproved.

Jablonski, Judith. What is the difference between a thesis statement and an hypothesis statement? Online Library. American Public University System. Jun 16, 2014.

$\mathcal{Q}$: How do you support your thesis?

$\mathcal{A}$: By developing, supporting, and explaining it via examples and evidence.

- Examples include:
  - Case studies of a system
  - Qualitative analysis
  - Software Evaluation

Jablonski, Judith. What is the difference between a thesis statement and an hypothesis statement? Online Library. American Public University System. Jun 16, 2014.
Carlos Andujar, Viola Schiaffonati, Fabio A. Schreiber, Letizia Tanca, Matti Tedre‖ Kees van Hee and Jan van Leeuwe. The Role and Relevance of Experimentation in Informatics. ECSS 2012.

$\mathcal{Q}$: How do you prove/disprove your hypotheses?

$\mathcal{A}$-1: By designing and conducting a controlled experiment

- Examples of experimentation include:
  - Empirical Evaluation (automatic metrics)
  - Human Evaluation

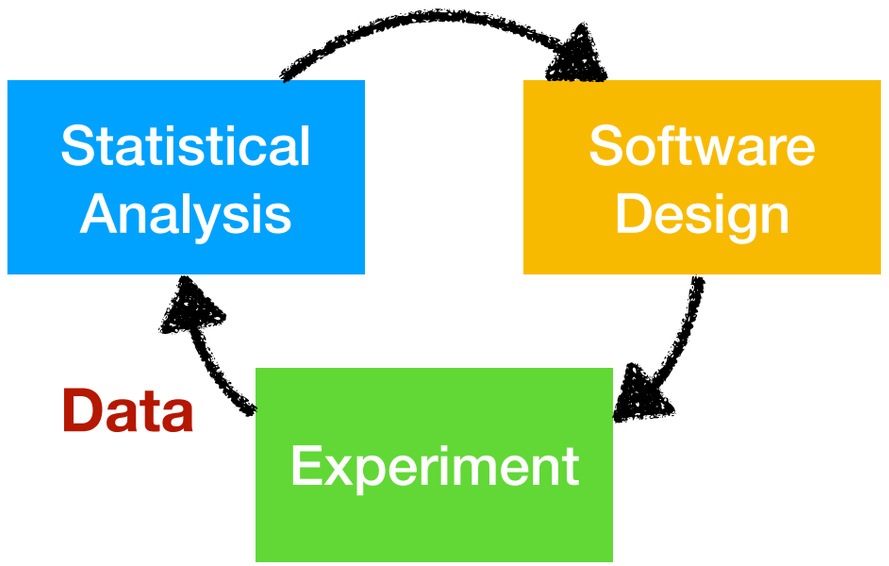
$\mathcal{A}$-2: By designing and conducting a usability study

Jablonski, Judith. What is the difference between a thesis statement and an hypothesis statement? Online Library. American Public University System. Jun 16, 2014.
Carlos Andujar, Viola Schiaffonati, Fabio A. Schreiber, Letizia Tanca, Matti Tedre‖ Kees van Hee and Jan van Leeuwe. The Role and Relevance of Experimentation in Informatics. ECSS 2012.

Experimental Design

Experiment and Design

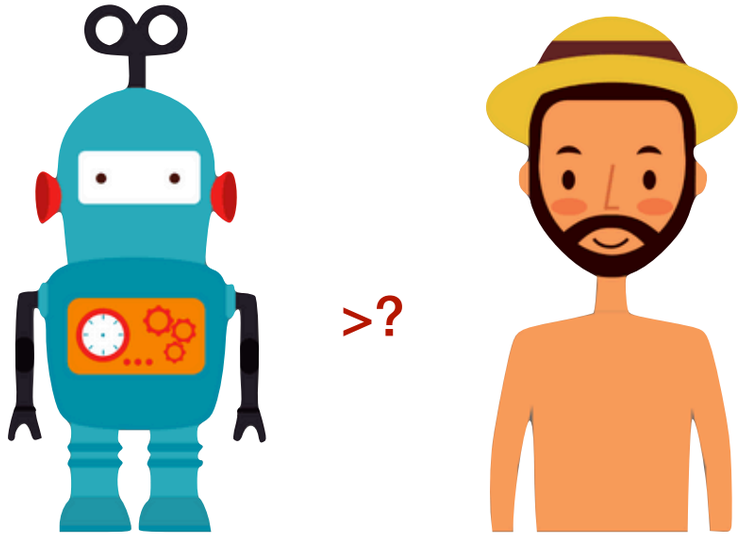
- Conventional software design cycle:



Examples:
- System
- Algorithm
- Model
- User Interface

Claims require 2 ingredients

- “My system / method / design / algorithm is good”
  - …good compared to what?
- “My system ... X is better than system ... Y”
- Is Y a strong competitor?
  - Not a ‘straw man’ but a sensible alternative approach
  - Ideally the current state-of-the-art approach
  - We call this a “baseline” for comparison
- What does “better than” mean?
  - Need to define metrics that measure “goodness”
  - E.g., speed, user satisfaction, task completion, energy usage, BLEU score ....



2 ingredients generate data

- Baselines
  - Comparison systems / methods / techniques control condition
- Metrics
  - Ways of measuring performance of your approach and the baseline(s)
- When you run experiments, you will apply the metrics to your system and the baseline(s)
  - …this will generate data that you need to analyse!

Why does it matter?

- Interpreting claims that you read:
  - “Method B is better than method A”
  - “Average task completion with system B is higher than with system A”
  - E.g., “Users of System A preferred it to System B 23% of the time, $p < 0.05$”
  - E.g., “System A is significantly faster than System B”
  - “Our results are significant”
- Making your own experimental claims:
  - E.g., “My new algorithm X means that users can complete tasks in only 52% of the time that they needed previously ($p<0.001$)”

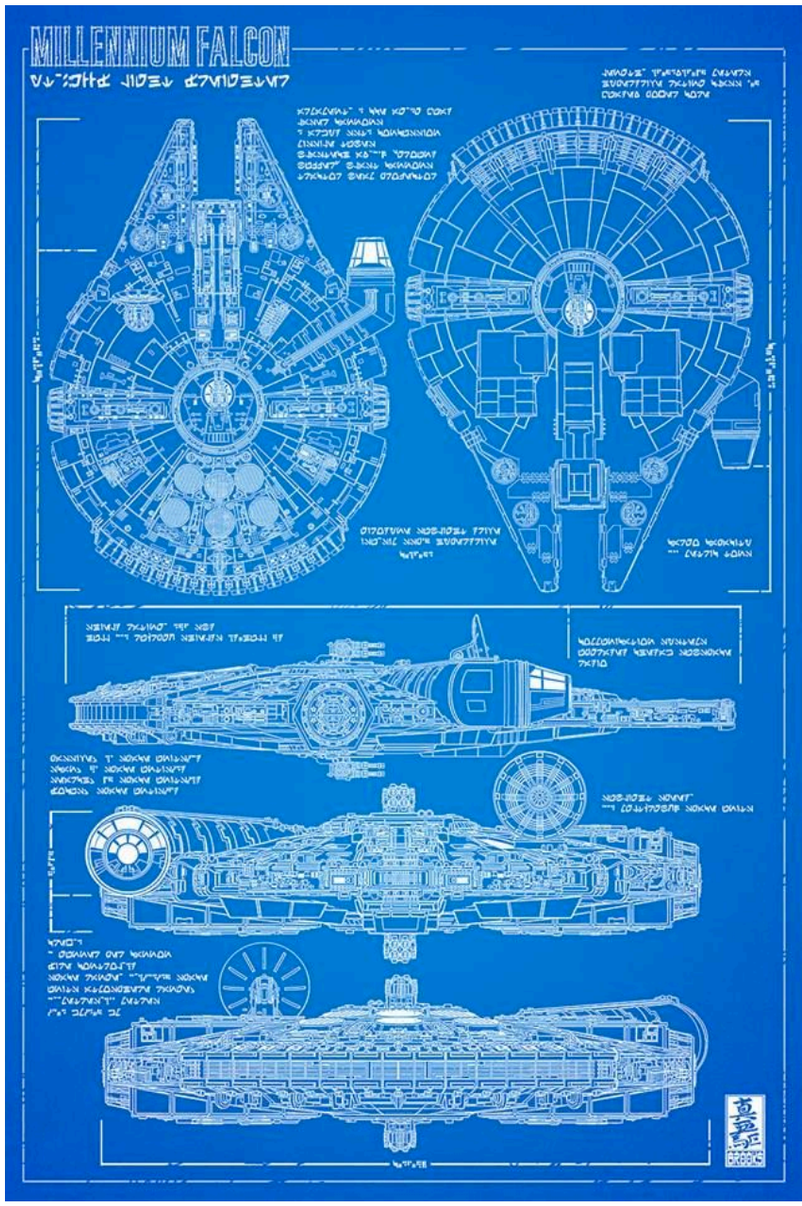
Experimental Method

- Experiments are required where:
  - You want to know the difference between two or more features or methods
    - *Which screen layout is easier to use?*
    - *How does my algorithm affect speed and accuracy?*
  - You want to see what happens when you make a change or series of changes to a design parameter
    - *At what font size does text become unreadable ?*
  - You want to see how the changes you've made to a method affect overall performance
    - *Do your changes make the system faster/more usable?*

Working artefact

- An experiment enables the researcher to observe the effect of manipulating one or more variables

- Requires an artefact: a simulation, prototype, or a fully implemented system



Experimental Design

1. Determine

   (a) the research questions of the project

   (b) the hypotheses being tested

2. Design the experiment(s)

3. Run the experiment AND/OR user study [takes a lot of time!]

4. Analyse measured data AND/OR questionnaires using statistical tests

5. Summarise and present the result

1(b). Hypothesis

???

- A statement of fact that you are going to try to prove/disprove

- An educated guess as to why something happens

Hypothesis: Examples

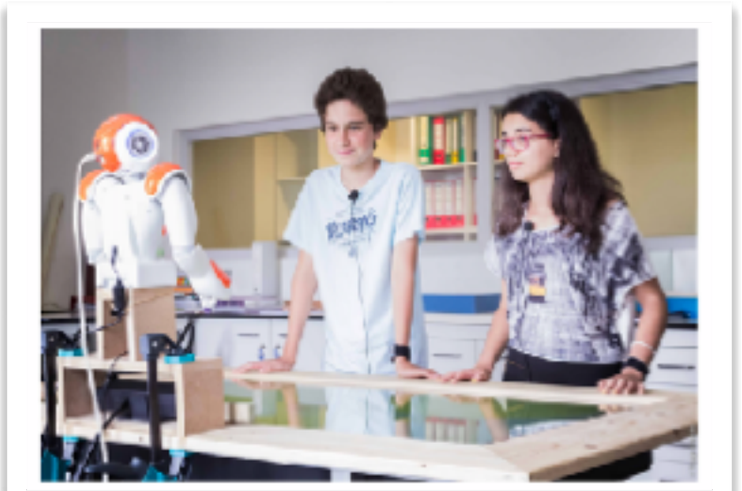
- Most of the time they are written like this:
- If _____ [I do this] ___ then ___ [this will happen]

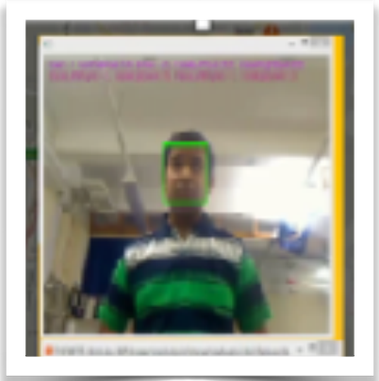
- For example:
  - *If I change the robot voice from male to female, then this makes the system perceived as more friendly*
  - *If users use voice instead of text input, then this will be more efficient for texting messages*

Hypotheses: more examples

- *Empathic robots are better teachers than non- empathic robots*

- *Robots with memory make better teachers than robots without memory*





Hypotheses: even more examples

- *Representation X allows faster convergence of the SARSA algorithm than representation Y, on task T.*

- *E.g. Users prefer dialogue flow A to standard flow B, measured by the PARADISE metrics.*

2. Design the Experiment: Principles

- Three basic principles of experimental design

(a) Replication

(b) Importance

(c) Control/Validity

2(a). Replication

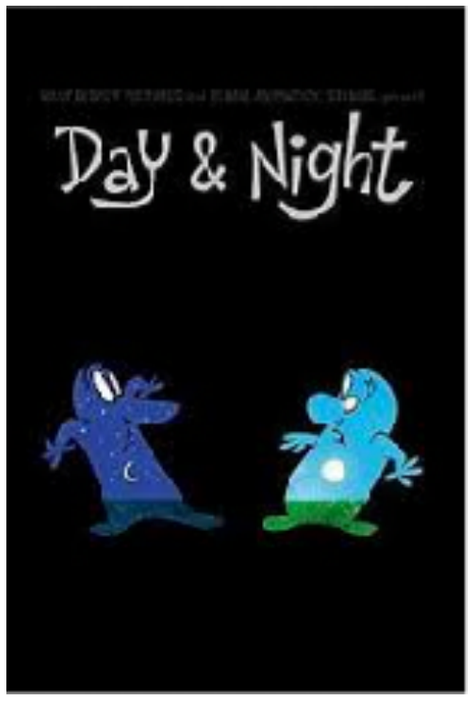
- Careful measurements allow you to replicate your experiments

2(b). Importance

- An experiment can be replicated and controlled but not worth it if it's not important

2(c). Controlled Environment

- Some variables correlate but don't necessarily cause effects
- Need to isolate the variables that cause effect



Designing Experiments: Recipe (Automatic Evaluation)

1. Create a controlled environment
2. Define parameters and hyperparameters
3. Decide about measurements of metrics
4. Execute model

Designing Experiments: Recipe (Human Evaluation)

1. Create a controlled environment

2. Create scenarios and tasks

3. Decide about measurements of metrics

4. Prepare questionnaires

5. Recruit subjects (5 - 10000000000000)

Ensuring a Controlled Environment

- Isolating the cause: control condition

Everything else must be the same

Ensuring a Controlled Environment

- Hypothesis: "*If mobile phones are used then more tumours will occur*"
- Cause: mobile phones
- Effect: tumours

One condition where **mobile phones** are **present**

One condition where **mobile phones** are **absent**

What else do you want to keep constant?

???

Everything else must be the same

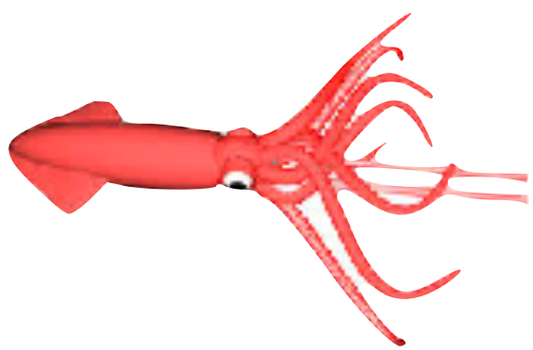
Ensuring a Controlled Environment

- Isolating the cause: control condition

Everything else must be the same

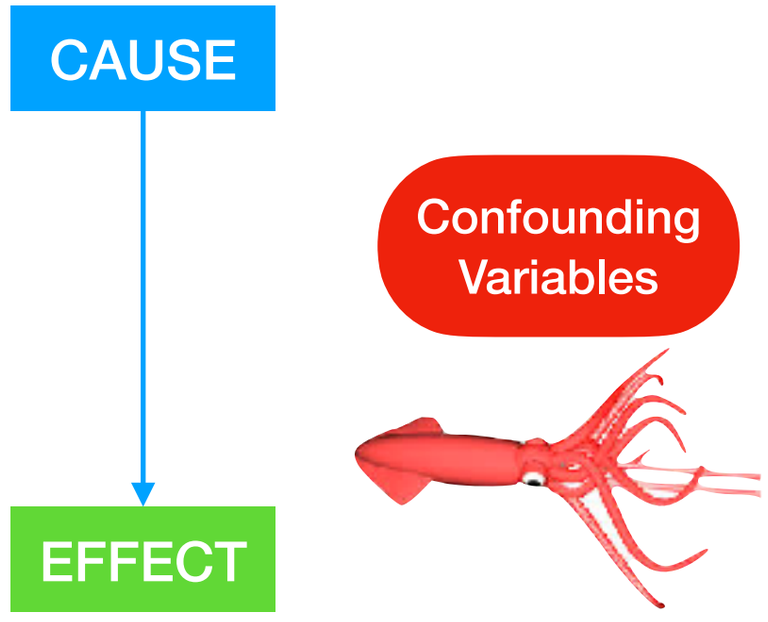
Confounding Variables: Tertium Quid

- Examples of confounding variables
  - Type of phone
  - Brain health
- Killing the Tertium Quid
  - Well thought-out and well executed experiment
  - Balanced experiment
  - Randomisation



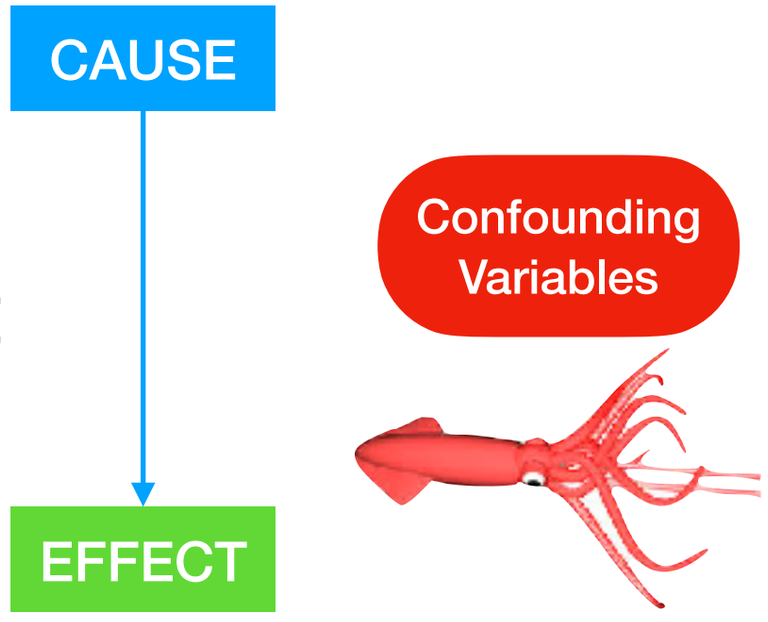
Independent/Dependent Variables (Automatic Evaluation)

- Independent Variable
  - Variable that is manipulated
    - e.g., type of training loss, encryption mechanism
- Dependent Variable
  - Variable that is measured
    - e.g., F-1 score



Independent/Dependent Variables (Human Evaluation)

- Independent Variable
  - Variable that is manipulated
    - e.g., phone use, type of diet
- Dependent Variable
  - Variable that is measured
    - e.g., brain tumours, blood pressure



Experimental Hypotheses

- A Hypothesis is:
  - Your prediction of outcome, in terms of the IV and DV

- Null hypothesis: that there is no difference between experimental conditions (i.e., changing the IV does not lead to any real changes in DV)

  - The aim is to disprove the null hypothesis

Hypotheses Examples

- ***If*** *task success* *is* ***related*** *to website structure*, ***then*** *changing the website's structure will result in changes in task success.*

- ***If*** *user preference* *is* ***related*** *to speed of spoken output*, ***then*** *people who experience different speed of speech synthesiser will give different preference ratings.*

  - Blue is dependent variable, red is independent variable
  - Blue is what you measure, red is what you manipulate/ change

Attributions

- https://static.makeuseof.com/wp-content/uploads/2017/04/experiment_lab-670x447.jpg
- https://thepolymathproject.com/wp-content/uploads/2018/07/steel-man.png
- https://webgnomes-webgnomesllc.netdna-ssl.com/wp-content/uploads/2012/07/seo-analysis.jpg
- https://i.pinimg.com/originals/d7/ff/6c/d7ff6cfc680189370e446d2a7f18c09b.jpg
- https://image.shutterstock.com/image-illustration/abc-building-blocks-on-white-260nw-109234082.jpg
- https://www.dictionary.com/e/wp-content/uploads/2018/04/hmm.jpg
- https://i.pinimg.com/236x/d7/a3/ae/d7a3ae5506817d1ef60dabde37150fe9--grumpy-cat-humor-grumpy-cats.jpg
- http://clipart-library.com/clipart/832162.htm