Experimental design and data analysis

Arsenii, Feyza & Samarth

Content Overview

- Scientific Method
- Data Collection
- Experiment Design
- Data analysis

Scientific Method

Forming a hypothesis

What: “An educated guess about a particular problem or idea”.
Example: Data augmentation improves performance in image classification tasks.

Support the hypothesis

How: Collect data, design and perform experiments.
Example: Train an image classifier on both the raw and augmented data.

Prove or refute

Collate your results and use statistics to prove or refute your hypothesis.

Ref. https://www.dummies.com/education/science/designing-experiments-using-the-scientific-method/

Data collection

Data is an empirical basis of scientific finding.

Main principle: **Data sample should statistically represent population**.

Common problems:

-Data sample is too small

-Data sample is biased, because samples with some properties are easier to collect (e.g survival bias)

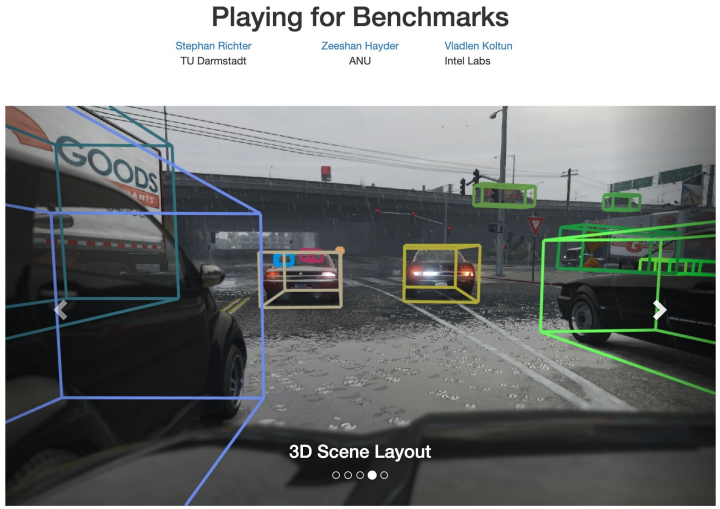
Data collection

Questions to be asked to foster data integrity (before you start!):

1) How will the data be collected?
2) How should records be kept and stored?
3) How, if at all, will data be backed up?
4) How long should data be kept?
5) Who owns the data?
6) When and with whom should data be shared?
7) Are you going to collect private data?

Data ownership example

GTA V dataset is video images dataset where all images are taken in the GTA V game. The dataset is very good for various computer vision tasks and, in general, publicly available, but has unusual requirement.



"...we request that you buy Grand Theft Auto V if you use data from the provided benchmark suite."

How to design good experiments?

Independent Variable
E.g. image rotation.

Dependent Variable
E.g. image classification accuracy.

Detailed and easily-accessible *records* and *visuals* of *data* and experiment *results*. With some hypotheses, experiments are not necessary and visuals are sufficient!

Control is a must so that you have something to compare (test your hypothesis against) e.g. image classification without data augmentation.

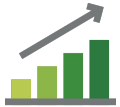
Results *as well as* the errors must be reported for the sake of scientific progress.

What should drive the experiments? Question(s) first, then everything else!

Things to consider and common mistakes

Which metric to use?

E.g. in ML, accuracy is not a proper metric to use when data is imbalanced, go for f1-score instead.



Not controlling for confounding variables. E.g. an additional preprocessing step that's not included in the control.

Also the classic: correlation doesn't imply causation.

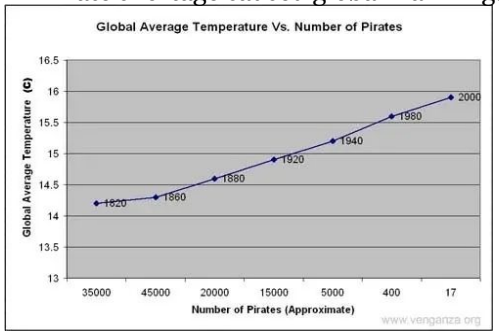
Data related errors, including bias in data, errors in data collection stage which is another setting for experimental design.

Violation of ethics and privacy e.g. collecting and analyzing personal data without consent. Overlooking license restrictions.

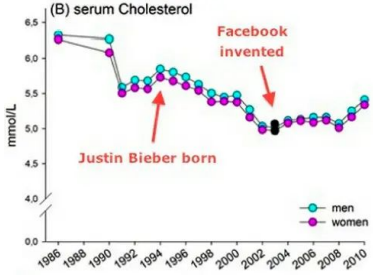
What are the performance metrics you use in your respective fields?

Tragic, comic or both?

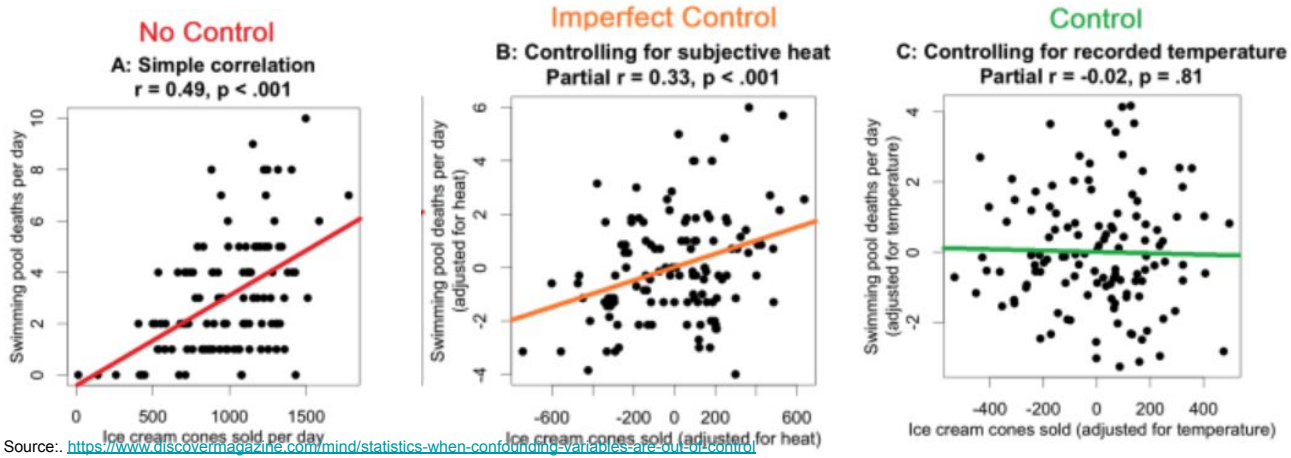
Pirate shortage caused global warming.



Facebook cancelled out the cholesterol-lowering effects of Justin Bieber.



Source: https://www.buzzfeednews.com/article/kjh2110/the-10-most-bizarre-correlations



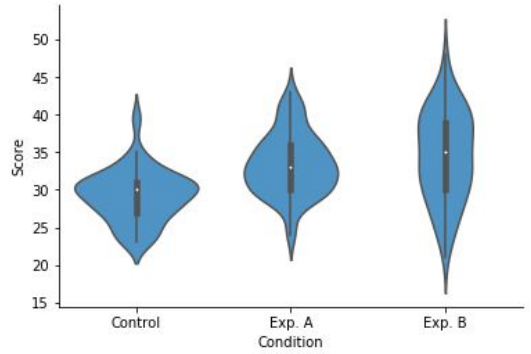
Source:. https://www.discovermagazine.com/mind/statistics-when-confounding-variables-are-out-of-control

Any data analysis or experiment design flaws or scandals you remember of?

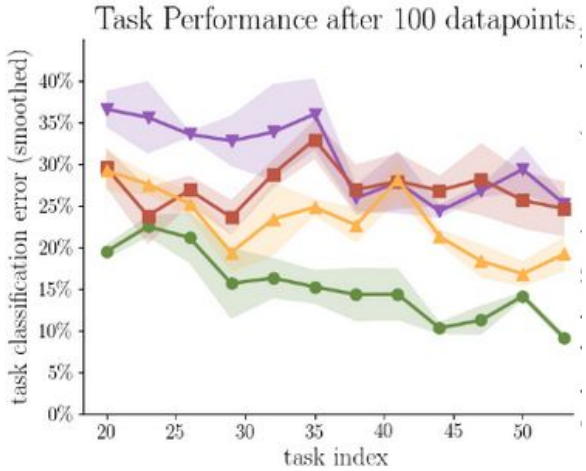
Data Analysis and Presentation

Use appropriate visualizations. Do not skimp or overdo.

Highlight smallest set of experiments/visualizations for an argument.



Source: https://towardsdatascience.com/violin-plots-explained-fb1d115e023d



Source: https://arxiv.org/pdf/1902.08438.pdf

Some other good experimental practices

Keep track of experiments and results. Useful tools : version control, config files. In case randomness is involved save random states too.

Long experiments? Log intermediate results. Checkpoint if you can.

```
args:
  dropout: 0.5
  eval_only: false
  lam: 0.01
  log_step: 1
  lr: 0.001
  model_save_freq: 1000
  num_epoch: 1000
  resume: ''
  save_dir: path/to/expt/results
  test_split: false
githash: d076d124oj6874ce295f849d41ef0bd8facf3e29
```

```
------------------Expt Log------------------
Time                        : 13:38 22-02-2020
Epoch                       : 1
Train loss                  : 0.6871
Val accuracy                : 56.34
--------------------------------------------
------------------Expt Log------------------
Time                        : 13:49 22-02-2020
Epoch                       : 2
Train loss                  : 0.7146
Val accuracy                : 66.25
--------------------------------------------
```

Any preferred methods for tracking experiments or creating visualizations?

Keeping Track

- Long experiments?
  - Hiccups with compute, errors hard to foresee
  - Logs/checkpoints to resume experiments from if possible

- “Fire and forget”

Common dos and don’ts

- Version control
  - Commit any experiment code (on a branch that does not get deleted)
  - Can save commit hash with experiment results
  - Good option : git
- Configs
  - Experimenting with different parameters
  - Frequent code changes needed?
  - Good option : yaml files
- Randomness
  - Save random state/seed for reproducibility
- Never lose track by saving above 3 with experiment results

Intro + higher levels points

- Experimental design [Feyza - How does the process begin? How to come up with good experiments?]
-
- Invest time in learning as much as possible from current experiment before moving on

- Hypothesize before experiments (Ask question)
- Dataset construction
- Model selection
- Evaluation
- Metrics
- Use plots/visualization to be thorough

Some opinionated guidelines

- Git/version control to not lose track of code
- Use configs. Save configs and git hashes for tracking
- Long experiments : Save checkpoints (ML e.g.)
- Some common mistakes
  - Example in ML is accuracy on an imbalanced dataset
  - Biased dataset
- Matplotlib : easier said than done