Experimentation

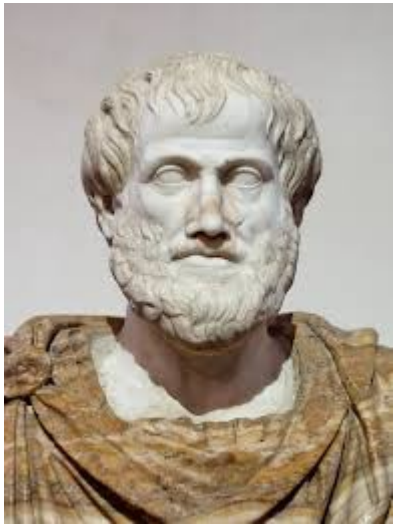
Peter Stuckey

UPM workshop

Why do we run experiments?

"*The fundamental principle of science, the definition almost, is this: the sole test of the validity of any idea is experiment*"

– Richard P. Feynman



"There are two possible outcomes:

- *if the result confirms the hypothesis, then you've made a measurement.*
- *If the result is contrary to the hypothesis, then you've made a discovery*."

– Enrico Fermi

An experiment answers questions

To design an experiment, first ask:

- What are you trying to establish?
- What is the biggest concern or question that someone might have?
- Is the experiment for you (you don't know the answer) or for other people (you want to convince them)?

Know your question & state it clearly

A “fishing expedition” can be great for exploratory work

It is not useful as an experiment

Make your experiment:

- specific enough to be feasible and testable
- broad enough to generalize
- **Recall:** write the related work first

Types of experiments

- Controlled experiments (quantitative evidence)
- Case studies (qualitative evidence)
  - Done by the researchers themselves
  - Done by other people

When is a case study better than a controlled experiment?

- **Example:** learning from the first users of a tool
- In HCI, 5 users is considered adequate

When not to do an experiment

Other types of evidence:

- Surveys
- Proofs
- Theoretical result
- New cryptographic construction
- natural experiment (observational experiment)

An experiment is a comparison

Observation *O1* : process **P** in environment **E1**

Observation O2 : process **P** in environment **E2**

If *O1* = *O2*, the environmental differences are irrelevant to the process

If *O1* ≠ *O2*, the difference is caused by the environmental differences

It is not enough to report, “my technique does well”
You must compare to the state of the art

Minimize differences

E1 and **E2** should be as similar as possible

If many differences, which one caused O1≠O2?

E1 and **E2** should be realistic of actual practice

- real traces/logs, real development practices, ...

Benchmarks

- Microbenchmarks

- Macrobenchmarks

Aside: comparing enhancements

Suppose you implement 3 enhancements, which improves results

full = baseline+e1+e2+e3 > baseline

Which enhancement is best?

Wrong approach

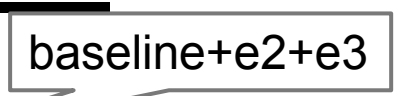
compare:

baseline+e1 to baseline,
baseline+e2 to baseline,
baseline+e3 to baseline

Right approach

compare:

full - e1 to full,
full - e2 to full,
full - e3 to full



Treatment and effect

Treatment = input = independent variables

- We called this the “environment” earlier
- Minimize the number!

Effect = output = dependent variables

Subjects

Experimental subjects:

- in social sciences, people
- in computer science, can be programs, etc.
    - it’s better to experiment on people when possible, especially if working on usability
    - its harder to experiment on people

Ethical considerations

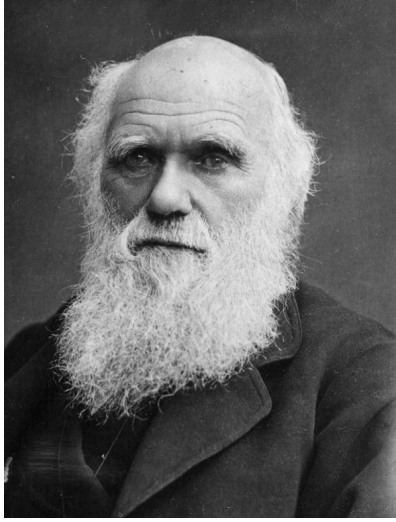
(when experimenting on people)

- informed consent
- potential harm

Experiment is reviewed by the HSC (human subjects committee) or IRB (institutional review board)

Long turnaround: submit your application early!

Problem: People differ a *lot*



Medicine has it easy:
height differs by about a factor of 2

Programming skill differs by *orders of magnitude*

Individual ability/knowledge/motivation is an independent variable

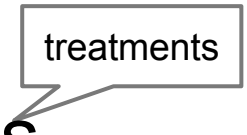
Non-human subjects also have variation

College students vs. practitioners

You can learn a lot from students

- available
- homogeneous
- uncharacteristic? (No evidence of this...)

Example experiment



Programmers fix bugs, using 2 tools

s1: t1

s2: t2

If s2 was faster, then t2 is better

How can we fix this experiment?

Improvement 1:replication

Programmers fix bugs, using 2 tools

s1: t1	s3: t1	s5: t1	...
s2: t2	s4: t2	s6: t2	...

If on average subjects using tool2 are faster, then tool2 is better

(How many programmers do we need? Lots.)

Improvement 2:paired design

Programmers fix bugs, using 2 tools

s1: t1xp1 t2xp2

s2: t1xp1 t2xp2

s3: t1xp1 t2xp2

s4: t1xp1 t2xp2

...

If tool2 trials are faster on average, then tool2 is better

Needs fewer subjects:~40 as a rule of thumb

Improvement 2b:paired design

Programmers fix bugs, using 2 tools

s1:	t1xp1	t2xp2
s2:	t1xp2	t2xp1
s3:	t1xp1	t2xp2
s4:	t1xp2	t2xp1

...

If tool2 trials are faster on average, then tool2 is better

Avoids confounding effect, or tool-program interaction

Improvement 2c:paired design

(blocked/counterbalanced)

Programmers fix bugs, using 2 tools

s1:	t1xp1	t2xp2
s2:	t1xp2	t2xp1
s3:	t2xp1	t1xp2
s4:	t2xp2	t1xp1

...

If tool2 trials are faster on average, then tool2 is better

Avoids another confound:learning/fatigue effects

Many other conflating factors exist

Example: self-selection

Between vs. within subjects

- Within subjects
- All participants try all conditions
  - + Can isolate effect of individual differences
  - + Requires fewer participants
  - - Ordering and fatigue effects
- Between subjects
- Each participant tries one condition
  - + No ordering effects, less fatigue.
  - - Cannot isolate effects due to individual differences.
  - - Need more participants

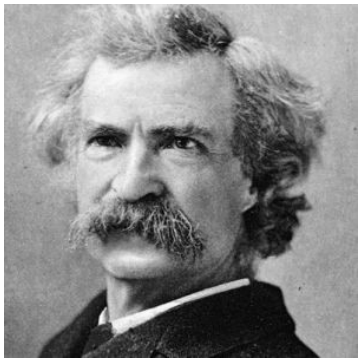
Combating individual variation

- Replication
  - In a population, individual variation averages away (central limit theorem)
- Randomization
  - Avoid conflating effects
- Statistics
  - Indicates when a difference is large enough to matter

Statistics

“There are three types of lies:
lies, damned lies, and statistics.”

- Mark Twain



Choosing a statistical test

It’s best to consult an expert or take a course in experimental design or statistical methods (≠ a course in statistics)

When in doubt, use ANOVA

- “ANalysis Of VAriance”

Statistics

Tests of Proportions

Samples	Response Categories	Tests
1	2	One-sample χ^2 test, binomial test
1	>2	One-sample χ^2 test, multinomial test
>1	≥2	*N*-sample χ^2 test, *G*-test, Fisher's exact test

Analyses of Variance

Factors	Levels	(B)etween or (W)ithin	Parametric Tests	Nonparametric tests
1	2	B	Independent-samples *t*-test	Mann-Whitney *U* test
1	>2	B	One-way ANOVA	Kruskal-Wallis test
1	2	W	Paired-samples *t*-test	Wilcoxon signed-rank test
1	>2	W	One-way repeated measures ANOVA	Friedman test
>1	≥2	B	Factorial ANOVA Linear Models (LM)	Aligned Rank Transform (ART) Generalized Linear Models (GLM)
>1	≥2	W	Factorial repeated measures ANOVA Linear Mixed Models (LMM)	Aligned Rank Transform (ART) Generalized Linear Mixed Models (GLMM)

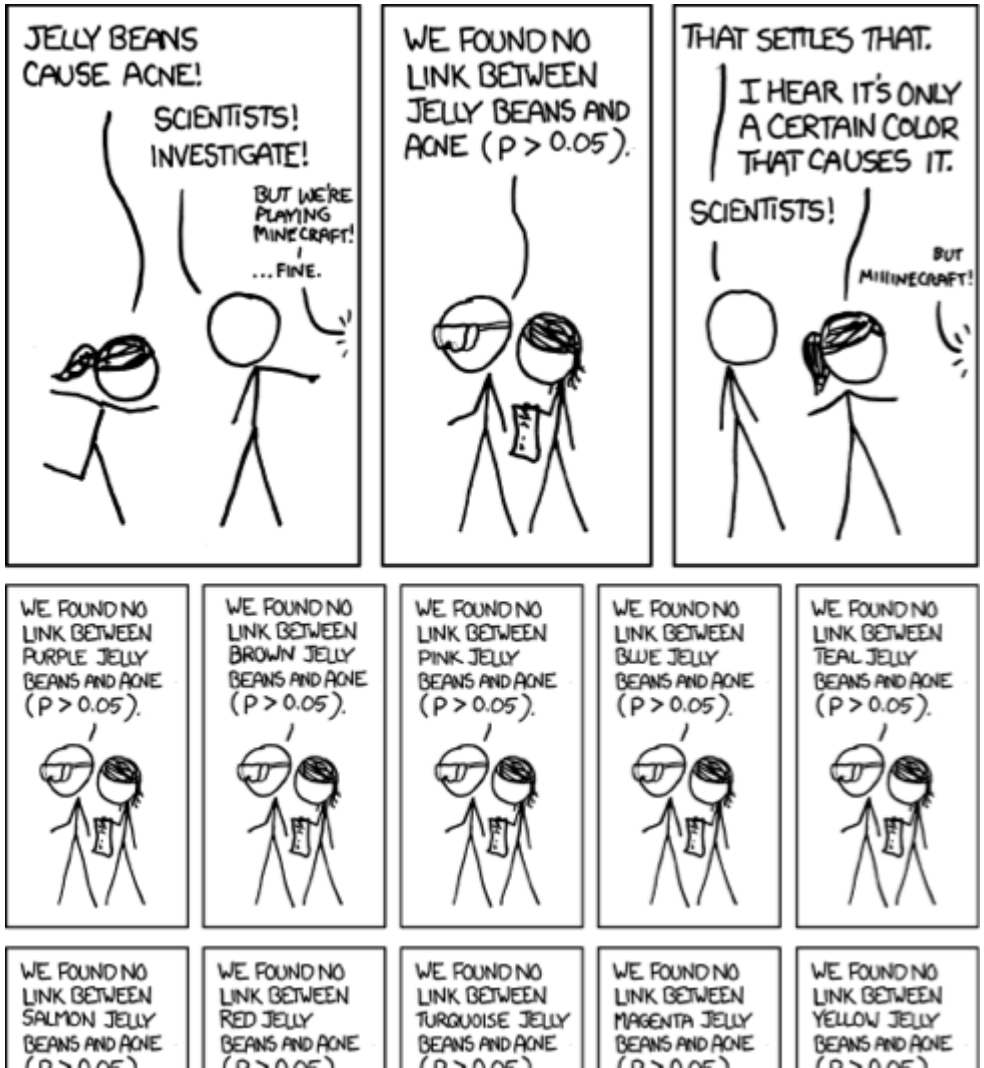
False positive errors

False positive (or false alarm or Type I error): no real effect, but report an effect (through good/bad luck or coincidence)

– If no real effect, a false positive occurs about 1 time in 20
    - 5% is a convention; there is nothing magic about it

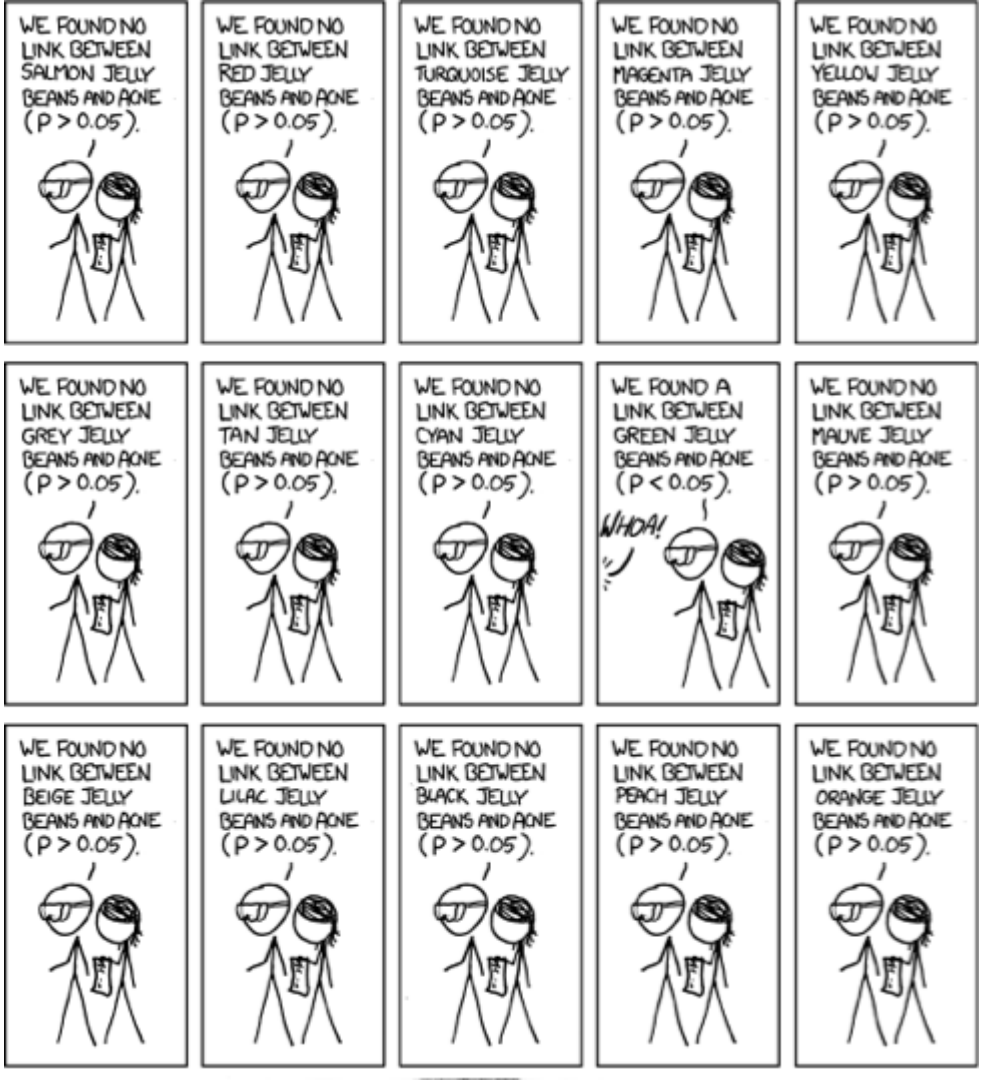
– If there is a real effect, a false positive occurs less often

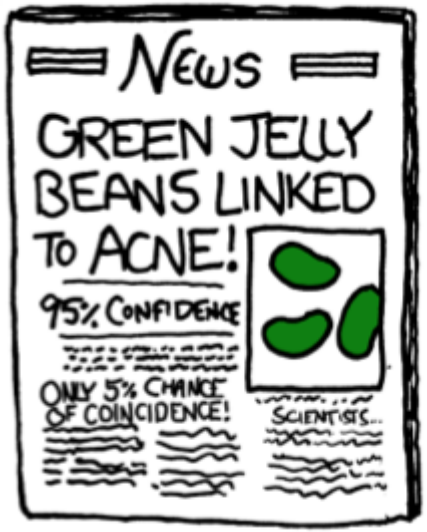
– **Caveat:** Different fields (e.g. security) have different standards



A false positive

http://xkcd.com/882/





Lesson: don't test too many factors

http://xkcd.com/882/

False negative errors

False negative (or miss or Type II error): real effect, but report no effect (through good/bad luck or coincidence)

–The smaller the effect, the more likely a false negative is

–How many die rolls to detect a die that is only slightly loaded?

The larger the sample, the less the likelihood of a false positive or negative

Some Experimental Terms

- False positive rate
- False negative rate
- True positive rate (sensitivity)
- True negative rate (specificity)

False Positive Paradox

Imagine you are testing for an unlikely property

- test 95% accurate for not property (true negative rate)
- test 100% accurate for property (true positive rate)

In 1000 tests 69 events return positive results

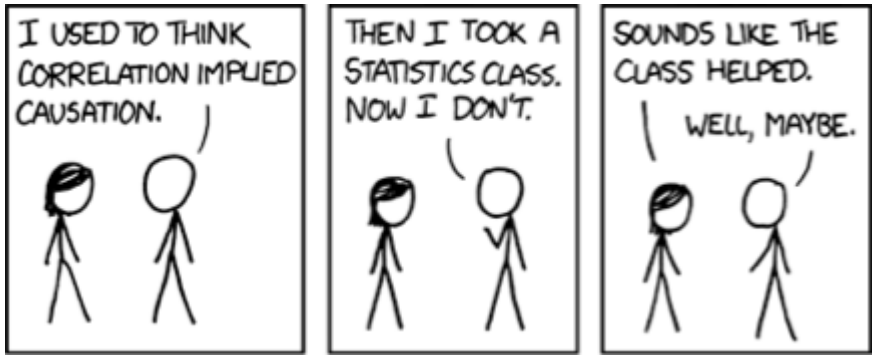
What percentage of the events have the property?

Events	Have Property	Don't have	Total
Test positive	20	49	69
Test negative	0	931	931
Total	20	980	1000

Correlation ≠ causation

Ice cream sales and murder rates are correlated

Lesson: you should always have an explanation for an effect



http://xkcd.com/552/

Statistical significance ≠ practical importance

After 1,000,000 rolls of a die, we find it is biased by 1 part in 10,000

Measurements (metrics to gather)

Decide methodology and measurements before you see the data

A common error:

1. Observe what you see in the real world
2. Decide on a metric (bigger value = better)

For any observation, there is something unique about it. Example: dice roll

Don’t trust your intuition

- People have very bad statistical intuition
- It’s much better to follow the methodology and do the experiments

Digits of precision

2 digits of precision is usually enough:
1.2 34 2500 3.7×10^6

A difference of less than 1% doesn't matter to readers

- The extra digits just distract

This is *not* consistent: 3.1 34.7 2594.6

Don’t report irrelevant measures

If a measurement is not relevant to your experimental questions, don’t report it.

- Don’t snow the reader with extra raw data

Some More Terminology

- Training test data
- K-fold cross validation
- Leave out testing
- ….

Test Sets

When building a software model of something we need at least two test sets

- training test set: used to train the model
- testing test set: used to evaluate the model

Danger of overfitting

- we learn a model specific to training set

Avoiding Overfitting: Exhaustive

Repeatedly apply the training/testing to different sets

Cross validation

D items, |D| = n

- leave out p testing
  - choose each set $V \subset D$ where $|V| = p$, train on $D - V$, test on V
  - problem: n choose p is very large
- leave out 1 testing:
  - efficient, but only makes sense when testing 1 is meaningful

Avoiding Overfitting:

k-fold cross validation

- break D in to k equal size parts P1,…,Pk
- train on D – Pi, test on Pi

Advantages

- each item used for training and testing
- only k tests (typically k = 10)

Biases

Your research is destined to succeed

- Hawthorne effect
- Friendly users, underestimate effort
- Sloppiness
- Fraud
  - (Compare to sloppiness)

Be as objective as possible

The Hawthorne Effect

Trials to determine effect on productivity of working conditions in the Hawthorne factory

- high light environment — increased
- low light environment — increased
- cleaning work spaces — increased
- clearing floors of obstacles — increased
- tiny increase in light — increased

Threats to validity

Discuss them in the paper

- Internal validity

whether an experimental treatment/condition makes a difference or not, and whether there is sufficient evidence to support the claim.

- External validity

generalizibility of the treatment/condition outcomes.

Another classification of threats to validity

- construct (correct measurements)
- internal (alternative explanations)
- external (generalize beyond subjects)
- reliability (reproduce)

Pilot studies (= prototypes)

Always do a pilot study

An experiment is costly

- your time, limited pool of subjects

You learn most from the first users

- you are certain to make mistakes

If you change anything, do another pilot study

Reproducibility

Your experiments should be reproducible

- treat them like other software
  - version control, tests, ...

If there are subjective decisions, have them cross-checked

Publish your data!

Publish your software!

Recomputation Manifesto

“Computational experiments should be recomputable for all time”

recomputation.org

Create a virtual machine including

- all the code you used
- all the data you used
- they will test it and make it available!

Results Presentation

Presenting results is critical to convincing readers that what you have done is significant

- tables
- charts
- graphs

Ideally results should be presented to be repeatable

Machine Details

For any experimentation you should give

- CPU details
  - 3.1 GHz Intel Core I7
- Operating Systems details
  - e.g. MAC OS X 10.10.5
- System details (mem, disk)
  - e.g. 16 GB RAM
- Important software component details
  - e.g. CPLEX 12.0.1 used as underlying MIP solver

Benchmark Selection

Which examples should you compare against?

- real world instances
- benchmark suites used in early work
- random instances created by you
- toy instances

Benchmark Instance Selection

Only presenting some instances is fraud

- If you don't present them all argue why
  - lack of time to do experiments (weak)
  - omitted all instances where no method solved in time limit (fine)
  - omitted all instances which the baseline solved in under 1 second (medium)
  - omitted all instances which could not be solved by new method (yikes)
- If you only have limited examples
  - argue why they are representative
- Better to have some negative results
  - than positive results on all instances, but very few

Presentation Methods

The results in the paper must be

- presented concisely
  - no unnecessary information presented
  - summarize results
  - give enough information to justify your discussion of experiments
- online appendices
  - give all the instances and raw data in an online appendix
  - allows people to compare against your instances
  - allows people to check your summarization/statistics

Tables

- The most basic form of presentation of results
- Positives
  - Allows the reader to make many comparisons
  - (Often) Allows other authors to compare their work with yours
  - Dense information
- Negatives
  - Dense information
  - Hard to see what the author is talking about

Tables Basics

Bench	CPLEX	Chuffed	Gecode	Gurobi
georege	123	100	232	233
enere	2342	1233	1233	1801
hablle	423	9090	2343	423
asdas	2343	TO	2342	22312
podsa	15654	11893	14543	15402
riods	0903	700	1230	780
zoiad	2344	1232	2442	2344

Tables Basics

- Have a meaningful caption
- Don’t make the numbers to small to read
- Highlight entries you want to focus on
- Sort the benchmarks meaningfully
- Place related columns next to each other
- Use correct justification

Improved Tables

Bench	CPLEX	Gurobi	Chuffed	Gecode
asdas	2343	**2312**	TO	22342
enere	2342	1801	**1233**	**1233**
georege	123	233	**100**	232
hablle	**423**	**423**	9090	2343
podsa	15654	15402	**11893**	14543
riods	903	780	**700**	1230
zoiad	2344	2344	**1232**	2442
sum	24132	**23295**	49248	44365
geom mean	**1379**	1419	2245	2524

Comparison of solvers, times in seconds, time limit 25000s, TO = timeout

Shrinking Tables

- summarize instances into benchmark suites
  - arithmetic mean?
  - geometric mean?
  - median?
  - standard deviation?

Outliers

What do you do for instances where

- a method crashed
- a method timed out
- a method ran out of memory
- a method gave a wrong answer

For full tables: give code for instance

- e.g. CR / MO / TO / WR

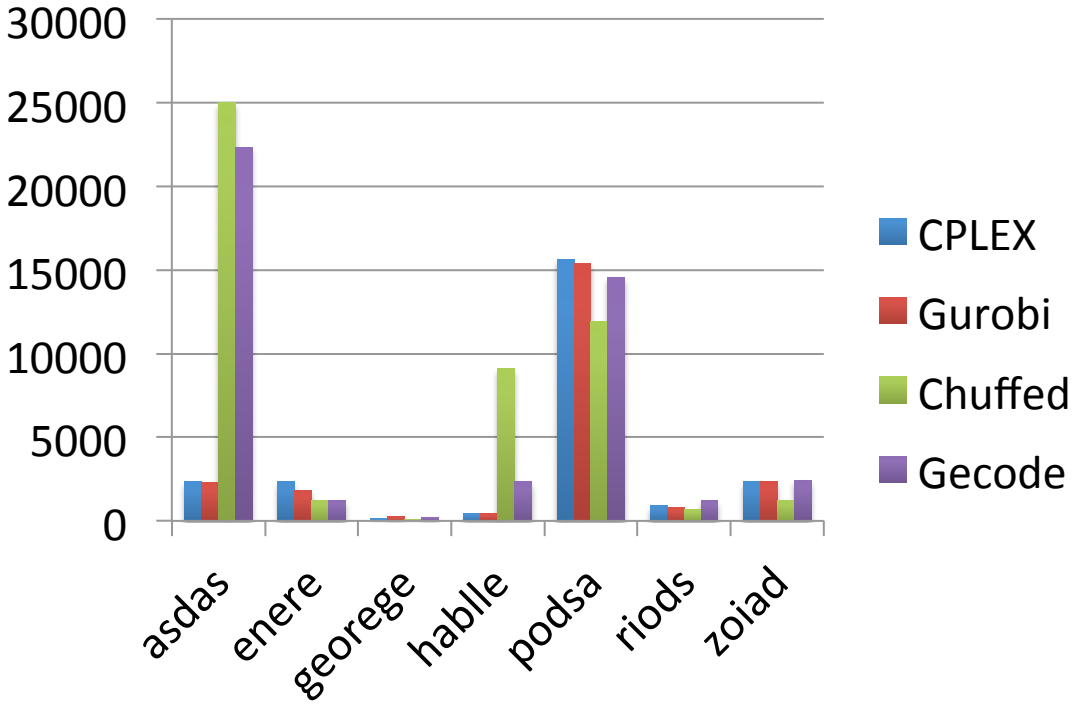
Probably need some discussion of crashes/wrong answers

Outliers and Summarization

How do you summarize with outliers

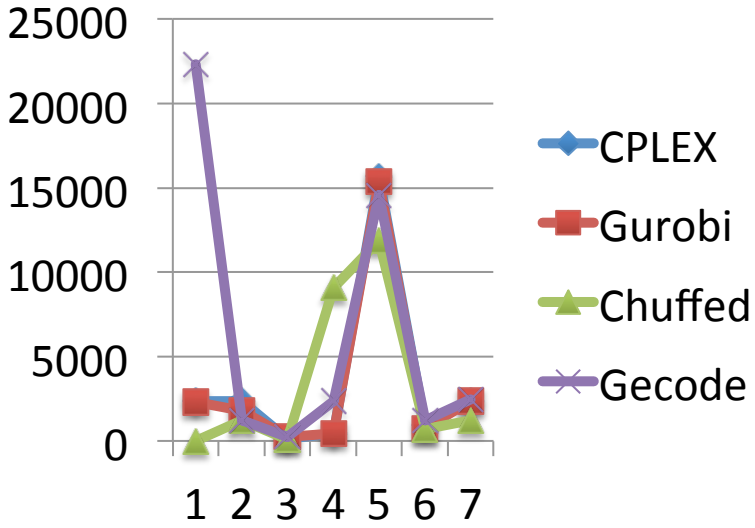
- remove them and add count of outliers
  - e.g. 100, 60, 70, MO, TO, 90, TO, 80 = 80^3 (mean 80 with 3 outliers)
- treat them as time out
  - e.g. 100, 60, 70, MO, TO, 90, TO, 80 = 125 (3 outliers treated as 200 = TO)
- treat them as penalized time out (PAR10)
  - e.g. 100, 60, 70, MO, TO, 90, TO, 80 = 425 (3 outliers as 2000 = TO*10)
- combine 1 and 2
  - e.g. 100, 60, 70, MO, TO, 90, TO, 80 = 125^3

Charts


30000
25000
20000
15000
10000
5000
0
asdas
enere
georege
hablle
podsa
riods
zoiad
CPLEX
Gurobi
Chuffed
Gecode

Charts

Which chart type shows what you want

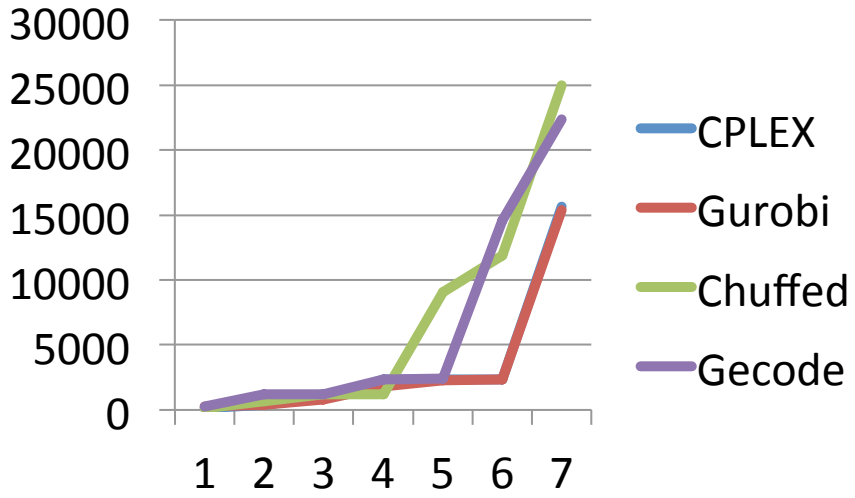


Charts Considerations

- Meaningful axes
  - does the order make sense
  - is the scale sensible, can we see the results
- Colors
  - make sure the chart is readable when printed in black and white
- Right form of the data
  - cumulative data
  - scatter plots
  - area under the curve

Cumulative Charts

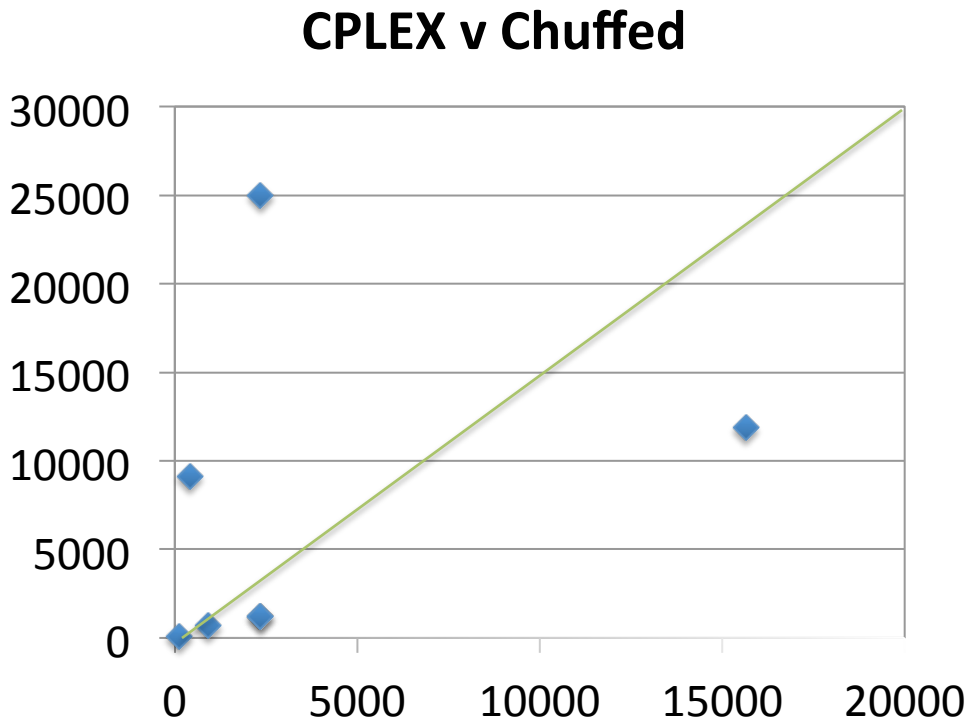
- How much time required for n successes
- Good for showing
  - many examples with
  - differing characteristics
- Also inverse
  - time required for % success



Scatter Plots

Comparing two data sets

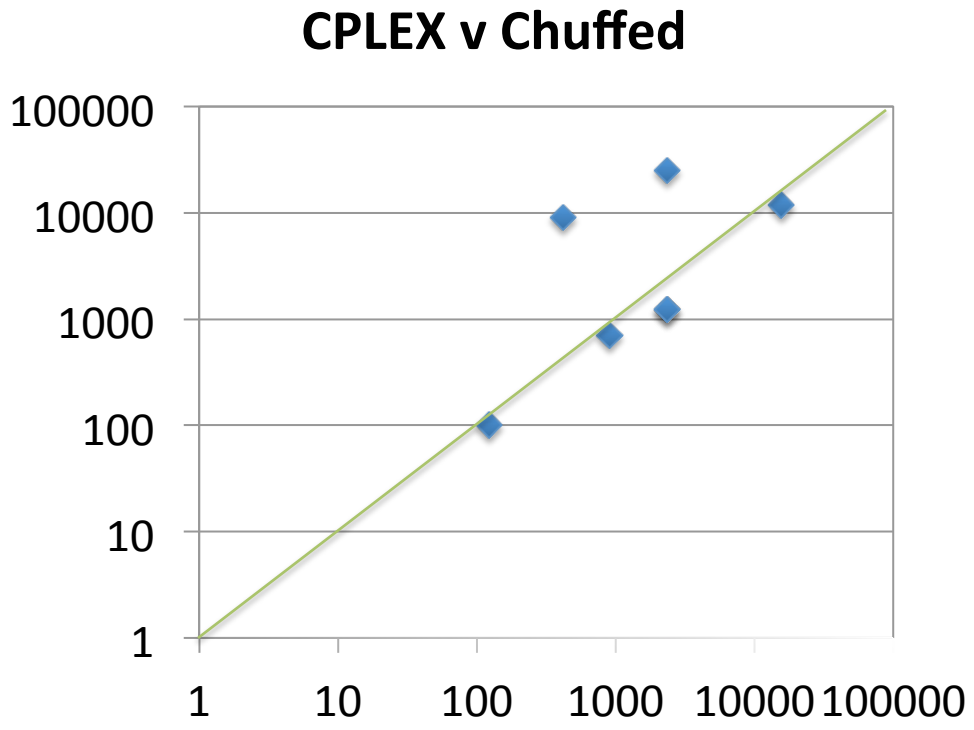
- not so good when numbers differ greatly



Scatter Plots

Comparing two data sets

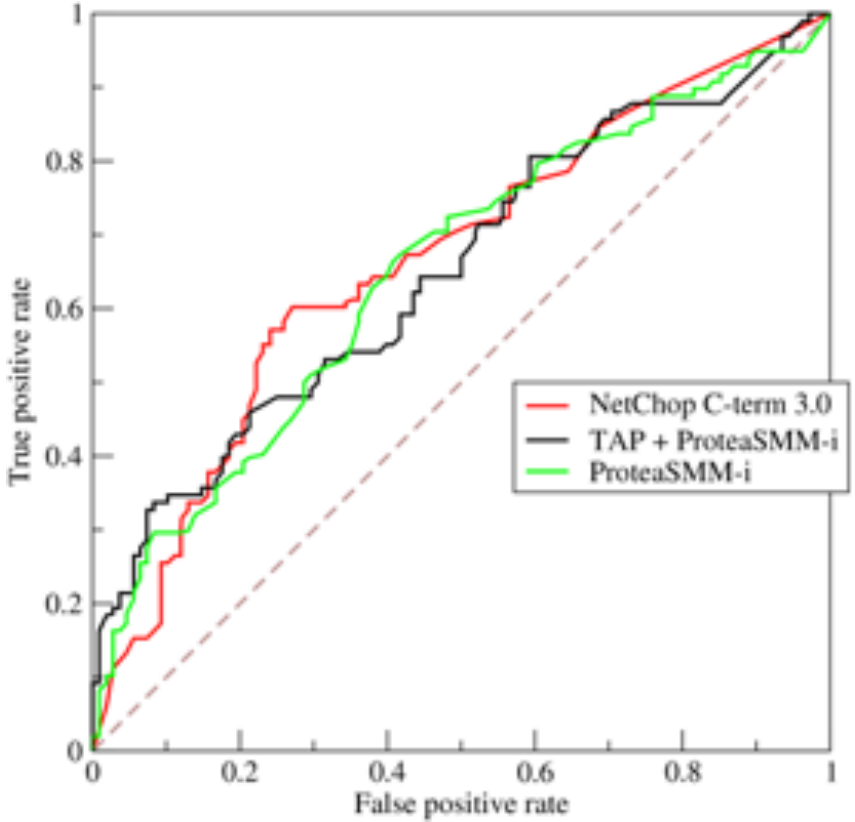
- Logarithmic scale
- makes big differences clearer



Special Charts

Some areas have special graphical forms of presentation

- binary classification
- receiver operating characteristic (ROC) curves



https://en.wikipedia.org/wiki/Receiver_operating_characteristic#/media/File:Roccurves.png

Results Presentation Summary

- Make the results concise and relevant
  - a new figure that makes an important point is worthwhile
- Use the method that makes differences you want to highlight clear
- Know the usual presentation methods for your research area