Design and Analysis of Experiments

. . . or, my unfinished journey to become a better experimenter

Violet R. Syrotiuk

Arizona State University

The Second Chameleon User Meeting
February 6, 2019 in Austin, Texas

Outline

1. Design and Analysis of Experiments
2. Guidelines for Designing Experiments
3. Guidelines for Analyzing the Results of Experiments
4. Our Research on Screening Experiments
5. A Few Take-Aways

Table of Contents

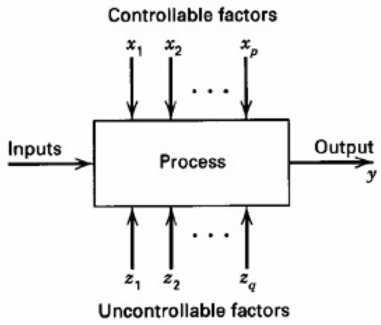
Our Interest

- Our interest is in engineered systems, such as Chameleon
  - $\Rightarrow$ A convergence of communication, computation, and storage!
- Complexity arises in such systems from their size, structure, operation, evolution over time, and human involvement.†

† NITRD Large Scale Networking (LSN) Workshop Report on Complex Engineered Networks, September 2012.

Formally, an experiment is

- a series of tests,
- in which purposeful changes are made to the input variables of a process or system,
- to observe and identify the reasons for changes that may be observed in the output response.



Test	Factors $x_1\ x_2 \ldots x_p$	Responses
1	$v_{x_1}\ v_{x_2} \ldots v_{x_p}$	$y_{1,1} \ldots y_{1,r}$
2	$v_{x_1}\ v_{x_2} \ldots v_{x_p}$	$y_{2,1} \ldots y_{2,r}$
$\vdots$	$\vdots\ \ \vdots \ldots \vdots$	$\vdots\ \ \vdots\ \ \vdots$
N	$v_{x_1}\ v_{x_2} \ldots v_{x_p}$	$y_{N,1} \ldots y_{N,r}$

Objectives of an Experiment

The objectives of an experiment include:

- Determine which x_i are most influential on the response y.
    - $\Rightarrow$ Screening.

Objectives of an Experiment

The objectives of an experiment include:

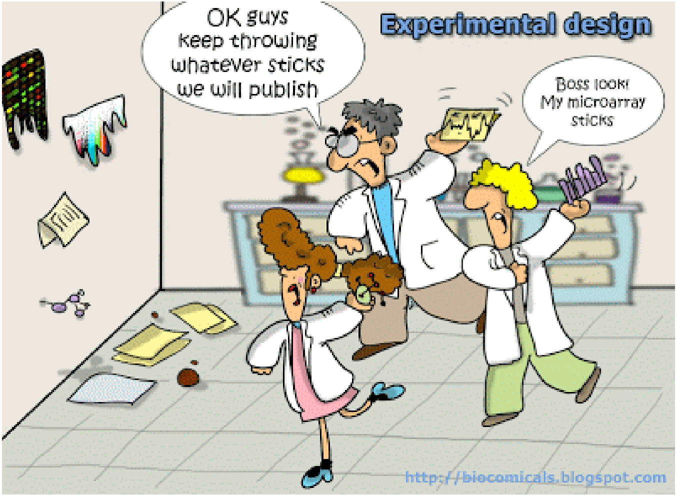
- Determine which x_i are most influential on the response y.
  - $\Rightarrow$ Screening.
- Determine where to set the influential x's so that y is almost always near the desired value.
  - $\Rightarrow$ Performance.

Objectives of an Experiment

The objectives of an experiment include:

- Determine which x_i are most influential on the response y.
  - $\Rightarrow$ Screening.
- Determine where to set the influential x's so that y is almost always near the desired value.
  - $\Rightarrow$ Performance.
- Determine where to set the influential x's so that the variability in y is small.
  - $\Rightarrow$ Robustness.

Among others!



"All experiments are designed experiments — some are poorly designed, some are well-designed." George E. P. Box

The Strategy of Experimentation

The general approach to planning and conducting an experiment is the strategy of experimentation.

Strategies include:

The Strategy of Experimentation

The general approach to planning and conducting an experiment is the strategy of experimentation.

Strategies include:

- The best-guess approach.

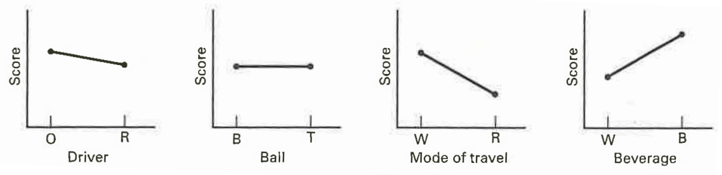
The Strategy of Experimentation

The general approach to planning and conducting an experiment is the strategy of experimentation.

Strategies include:

- The best-guess approach.
- The one-factor-at-a-time approach.

Consider a golf experiment:



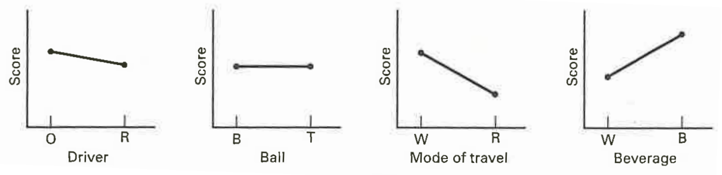
The Strategy of Experimentation

The general approach to planning and conducting an experiment is the strategy of experimentation.

Strategies include:

- The best-guess approach.
- The one-factor-at-a-time approach.

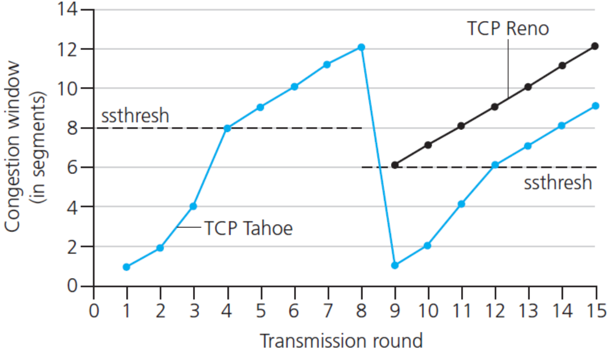
Consider a golf experiment:



This strategy may be useful when running a benchmark, but fails to consider any possible interaction between factors!

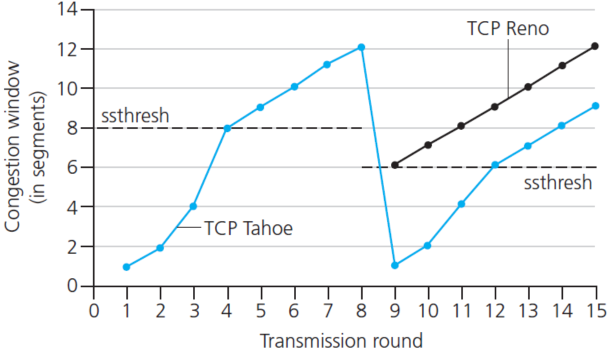
An Example Interaction

Congestion control in the TCP protocol:



An Example Interaction

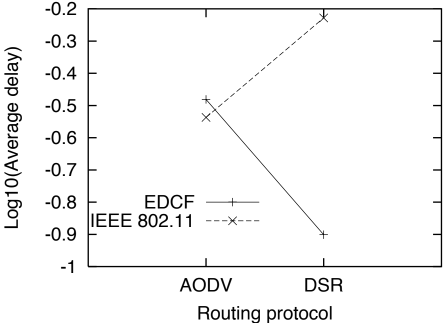
Congestion control in the TCP protocol:



- In wireless networks, contention manifests itself as congestion.
- But congestion control is the incorrect response to contention!

Interactions

- More formally, an interaction is the failure of a factor to produce the same effect at different levels of another factor.
- An example interaction graph for MAC/routing protocol interaction:

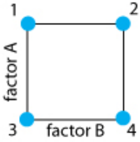


- Many examples of cross-layer interactions exist.

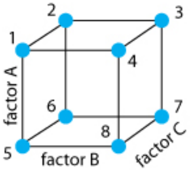
Other Experimental Strategies

Use an experimental strategy in which factors are varied together!

- A full-factorial experiment is one in which every possible combination of factor levels is tested.
  - In a system with k factors, each having two levels, the full factorial experiment has 2^k tests.



Test	Factor levels *A*	*B*
1	1	0
2	1	1
3	0	0
4	0	1



Test	Factor levels *A*	*B*	*C*
1	1	0	0
2	1	0	1
3	1	1	1
4	1	1	0
5	0	0	0
6	0	0	1
7	0	1	1
8	0	1	0

Other Experimental Strategies and Tools

There are many experimental strategies depending on the objective:

- Classical: screening, response surface, factorial, mixture, *etc.*
- Special purpose: covering arrays, space filling designs, nonlinear, balanced incomplete block designs, *etc.*

There are tools to help in experiment design and analysis:

Other Experimental Strategies and Tools

There are many experimental strategies depending on the objective:

- Classical: screening, response surface, factorial, mixture, *etc.*
- Special purpose: covering arrays, space filling designs, nonlinear, balanced incomplete block designs, *etc.*

There are tools to help in experiment design and analysis:

Make friends with a statistician!

Table of Contents

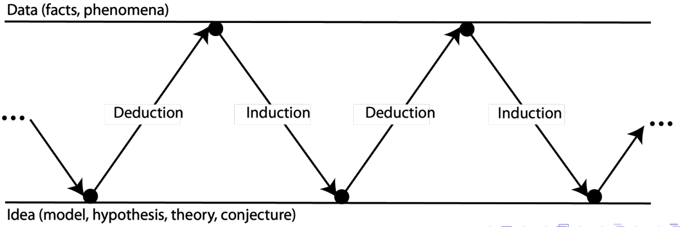
Guidelines for Designing Experiments

1. What is the problem?
2. Choose factors, levels, and range.
3. Select response variable(s).
4. Choose experimental design.
5. Perform the experiment.
6. Conduct a statistical analysis of the data.
7. What are the conclusions and/or recommendations?

Guidelines for Designing Experiments

1. What is the problem?
2. Choose factors, levels, and range.
3. Select response variable(s).
4. Choose experimental design.
5. Perform the experiment.
6. Conduct a statistical analysis of the data.
7. What are the conclusions and/or recommendations?

Steps 4–6 usually need to be repeated!



Values of Factors

The values a factor may take on can be:

- Discrete if there are a limited number of alternatives.

Values of Factors

The values a factor may take on can be:

- Discrete if there are a limited number of alternatives.
- Continuous if there an infinite number of values between any two values.
  - $\Rightarrow$ How to sample the range?

Values of Factors

The values a factor may take on can be:

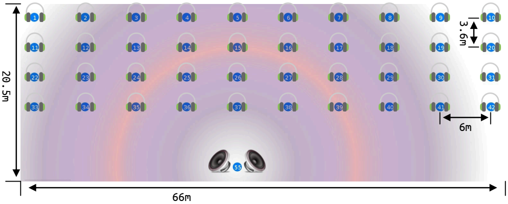
- Discrete if there are a limited number of alternatives.
- Continuous if there an infinite number of values between any two values.
  - $\Rightarrow$ How to sample the range?
- Categorical if there is no natural order between the categories (*e.g.*, eye colour).

Values of Factors

The values a factor may take on can be:

- Discrete if there are a limited number of alternatives.
- Continuous if there an infinite number of values between any two values.
  - $\Rightarrow$ How to sample the range?
- Categorical if there is no natural order between the categories (*e.g.*, eye colour).
- Ordinal if an ordering exists (*e.g.*, exam results, socio-economic status).

Parameters used in Wi-Fi Conferencing Scenario



Parameter	Identifier	Values
Band	band	2.4, 5 GHz
Channel	channel	1, 6, 11 (2.4 GHz); 36, 40, 44 (5 GHz)
Wi-Fi bitrate	bitrate	6, 9, 12, 24, 36 Mbps
Transmit power	txpower	1, 2, 4, 7, 10 dBm (2.4 GHz); 7, 8, 10, 13, **16** dBm (5 GHz)
MTU	mtu	256, 512, 1024, 1280, **1500** bytes
Transmit queue length	txqueuelen	10, 50, 100, 500, **1000** packets
Queuing discipline	qdisc	pfifo, bfifo, **pfifo_fast**
IP fragment low threshold	ipfrag_low_thresh	25%, 50%, **75%**, 100% of high threshold
IP fragment high threshold	ipfrag_high_thresh	16384, 65536, 262144, 1048576, **4194304** bytes
UDP receive buffer minimum	udp_rmem_min	**1.9231%**, 10%, 50% of maximum
UDP receive buffer default	rmem_default	0%, 25%, 50%, 75%, **100%** from minimum to maximum
UDP receive buffer maximum	rmem_max	2304, 10418, 47105, **212992** bytes
UDP transmit buffer minimum	udp_wmem_min	**1.9231%**, 10%, 50% of maximum
UDP transmit buffer default	wmem_default	0%, 25%, 50%, 75%, **100%** from minimum to maximum
UDP transmit buffer maximum	wmem_max	4608, 16537, 59349, **212992** bytes
UDP global buffer minimum	udp_mem_min	25%, **50%**, 75% of maximum
UDP global buffer pressure	udp_mem_pressure	0%, **33.338%**, 50%, 75%, 100% from minimum to maximum
UDP global buffer maximum	udp_mem_max	95, 949, 9490, **94896** pages
Robust header compression	ROHC	off, on (unimplemented)
Sensing	sensing	off, on (unimplemented)
Audio codec	codec	Opus, Speex
Audio codec bitrate	codecBitrate	7600, 16800, 24000, 34000 bit/s (or nearest allowed by codec)
Frame length aggregation	frameLen	20, 40, 60
Interference channel occupancy	intCOR	10%, 25%, 50%, 75%, 90%

Principles of Experimental Design

Three basic principles of experimental design are:

1. Replication: A repetition of the experiment.
   - Replication reflects sources of variability both between and (potentially) within tests.

Principles of Experimental Design

Three basic principles of experimental design are:

1. Replication: A repetition of the experiment.
    - Replication reflects sources of variability both between and (potentially) within tests.
2. Randomization: The order the individual tests of the experiment are to be performed are randomly determined.
    - Statistical methods require that observations (or errors) be independently distributed random variables.

Principles of Experimental Design

Three basic principles of experimental design are:

1. Replication: A repetition of the experiment.
   - Replication reflects sources of variability both between and (potentially) within tests.
2. Randomization: The order the individual tests of the experiment are to be performed are randomly determined.
   - Statistical methods require that observations (or errors) be independently distributed random variables.
3. Blocking: A design technique used to improve the precision with which comparison among the factors of interest are made.

Tools to Configure and Orchestrate the Experiment

- The `w-iLab.t` testbed uses OMF for resource allocation, hardware and software configuration, and the orchestration of experiments.
- Measurement data from each test is collected and stored for further processing.

Other Considerations

Other considerations in running the experiment:

- Resetting wireless interfaces for each test, *i.e.*, rebooting.
- Reinitialization, *e.g.*, flushing buffers cached by the OS.

Other Considerations

Other considerations in running the experiment:

- Resetting wireless interfaces for each test, *i.e.*, rebooting.
- Reinitialization, *e.g.*, flushing buffers cached by the OS.
- Collect measurements after a warm-up period.
  - $\Rightarrow$ Avoid transient effects (*e.g.*, avoid cold caches).
- Run the experiment long enough.
  - $\Rightarrow$ Ensure effects are observed (*e.g.*, changes to buffer sizes, queuing policies).

Table of Contents

Missing Data

Dirty data refers to data that are erroneous.

We encountered a problem with not all wireless nodes reporting a measurement in every test.

How to handle missing data?

Missing Data

Dirty data refers to data that are erroneous.

We encountered a problem with not all wireless nodes reporting a measurement in every test.

How to handle missing data?

- Ignore the missing value.
    - $\Rightarrow$ Effectively, a smaller sample size.

Missing Data

Dirty data refers to data that are erroneous.

We encountered a problem with not all wireless nodes reporting a measurement in every test.

How to handle missing data?

- Ignore the missing value.
  - $\Rightarrow$ Effectively, a smaller sample size.
- Estimate the missing value.
  - Substitute a mean, median, or mode.
  - Substitute a constant; add an indicator variable.
  - Impute a value using a model.

Missing Data

Dirty data refers to data that are erroneous.

We encountered a problem with not all wireless nodes reporting a measurement in every test.

How to handle missing data?

- Ignore the missing value.
  - $\Rightarrow$ Effectively, a smaller sample size.
- Estimate the missing value.
  - Substitute a mean, median, or mode.
  - Substitute a constant; add an indicator variable.
  - Impute a value using a model.

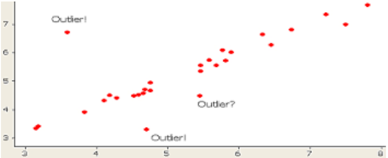
Use caution if the number of missing values is high!

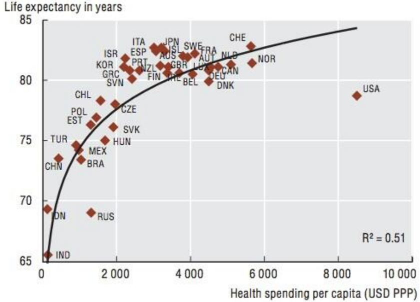
Outliers

What about outliers?

One of two situations could be true:

1. The actual value of the outlier is correct.
   - ⇒ Examine this observation further to understand why it occurred.
2. The value is incorrect.
   - ⇒ It may be possible to find out what is the actual value.





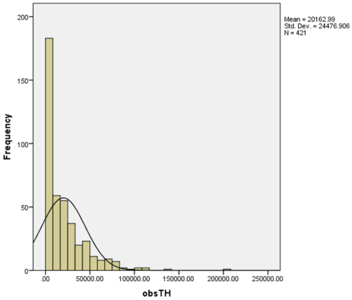
Source: OECD Health Statistics 2013, http://dx.doi.org/10.1787/health-data-en; World Bank for non-OECD countries.

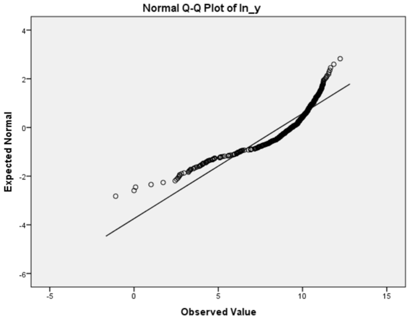
StatLink http://dx.doi.org/10.1787/888932916040

Plot your Data!

The first thing you should always do is plot your data!

- What is the distribution of your response?
  - A transformation of the data may be appropriate, otherwise the assumptions underlying any statistical tests used may be invalidated.

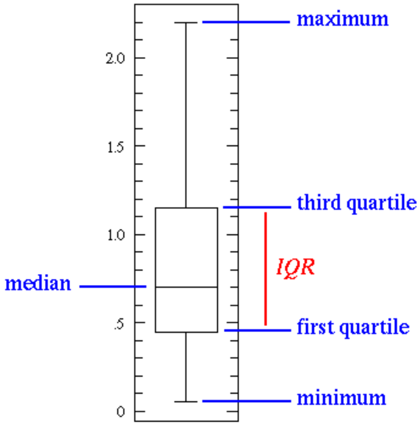




Box Plots: Display of Distribution

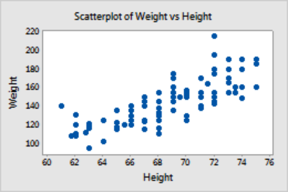
A box plot is a standardized way of displaying the distribution of data.

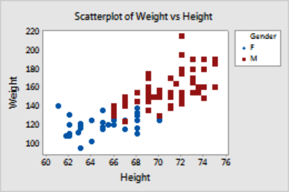
- The central rectangle spans the interquartile range (IQR).
- A segment inside the rectangle shows the median.
- The “whiskers” above and below the box show the locations of the minimum and maximum.

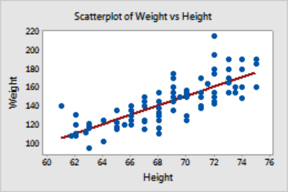


Scatter Plots

A scatter plot may be useful if the interesting feature is the pattern or clustering (or lack thereof) in the data.



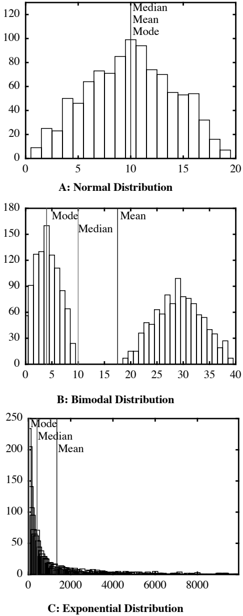




Mean, Median, Mode and Distribution

The relationship between the mean, median, and mode give hints about the data distribution.

- In a normal distribution, the mean is representative of the data set.
- In an exponential distribution, the mode and median are more representative.
- In a bimodal distribution, no single metric accurately describes the data.

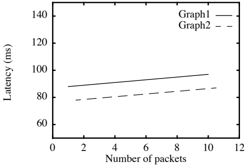


Margin of Error

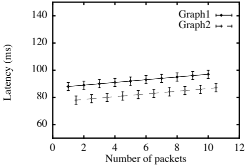
The margin of error expresses a range of values about the mean in which there is a high level of confidence that the true value falls.

Claim: A new technique reduces latency by 10%.

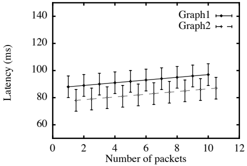
- A: no indication of the margin of error.
- B: it is reasonable to conclude that latency has been reduced.
- C: the margin of error isas large as the stated improvement. The 10% reduction in latency falls within the error bars, and might have arisen from experimental error.



A: Latency Improvement without error margins



B: Latency Improvement with small error margins

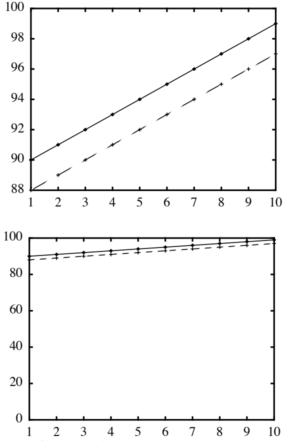


C: Latency Improvement with large error margins

Presentation of Results

Misleading vs. well-labelled *y* axes:

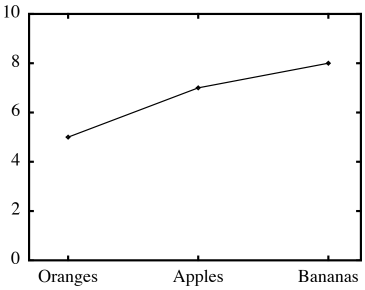
- By limiting the *y* axis to a narrow range of values, there appears to be a large difference between the two data sets in the top figure.
- The same data is shown in the lower figure with a better *y* axis selection. The data sets differ by only a small amount.



Presentation of Results (2)

Improper graph selection:

- In this example, the three data points represent three unrelated values that are implicitly being compared.
- By representing this data as a line graph, it suggests that *y* is presented as a function of *x*.



Presentation of Results (3)

Some common sense rules:

- Use zero-based axes when data is plotted on a linear scale.
- Use $\log$ scales to depict values that range over several orders of magnitude.
- Label all axes clearly, noting the units and scale if it is not linear.
- Use consistent graphic representation throughout (*i.e.*, colour, shape).

Presentation of Results (3)

Some common sense rules:

- Use zero-based axes when data is plotted on a linear scale.
- Use $\log$ scales to depict values that range over several orders of magnitude.
- Label all axes clearly, noting the units and scale if it is not linear.
- Use consistent graphic representation throughout (*i.e.*, colour, shape).

(Some of these "rules" were broken in some of the earlier figures on purpose!)

Table of Contents

Why do we need a New Screening Design?

The DOE community suggests using domain expertise to limit the number of factors used in experimentation to about ten.

But the complex engineered systems of interest have one to two orders of magnitude more factors!

- Conventional screening designs are only useful to screen main effects efficiently.
- Our interest is also in screening two-way interactions.
  - $\Rightarrow$ This motivates a new screening design, a locating array.
  - Under what conditions can we find one?

Why do we need a New Screening Design?

The DOE community suggests using domain expertise to limit the number of factors used in experimentation to about ten.

But the complex engineered systems of interest have one to two orders of magnitude more factors!

- Conventional screening designs are only useful to screen main effects efficiently.
- Our interest is also in screening two-way interactions.
  - $\Rightarrow$ This motivates a new screening design, a locating array.
  - Under what conditions can we find one?
    - Another day, a different talk!

Example Locating Array

- A (d, t)-locating array is a set of tests that ensure that every set of d distinct t-factor interactions appears in a different set of tests.
- This enables locating the causes of outcomes, such as an interaction most strongly influencing a response.
- Example:
  - Three 2-value factors (A-C).
  - One 3-value factor (D).
  - $(1, 2)$-locating.
  - The full design space has 24 tests.

Test	A	B	C	D
1	0	0	0	0
2	0	0	1	1
3	0	0	1	2
4	0	1	0	1
5	0	1	0	2
6	0	1	1	0
7	1	0	0	1
8	1	0	0	2
9	1	0	1	0
10	1	1	0	0
11	1	1	1	1
12	1	1	1	2

Example Locating Array

- A (d, t)-locating array is a set of tests that ensure that every set of d distinct t-factor interactions appears in a different set of tests.
- This enables locating the causes of outcomes, such as an interaction most strongly influencing a response.
- Example:
  - Three 2-value factors (A-C).
  - One 3-value factor (D).
  - $(1, 2)$-locating.
  - The full design space has 24 tests.

Test	A	B	C	D
1	0	0	0	0
2	0	0	1	1
3	0	0	1	2
4	0	1	0	1
5	0	1	0	2
6	0	1	1	0
7	1	0	0	1
8	1	0	0	2
9	1	0	1	0
10	1	1	0	0
11	1	1	1	1
12	1	1	1	2

Example Locating Array

- A (d, t)-locating array is a set of tests that ensure that every set of d distinct t-factor interactions appears in a different set of tests.
- This enables locating the causes of outcomes, such as an interaction most strongly influencing a response.
- Example:
  - Three 2-value factors (A-C).
  - One 3-value factor (D).
  - $(1, 2)$-locating.
  - The full design space has 24 tests.

Test	A	B	C	D
1	0	0	0	0
2	0	0	1	1
3	0	0	1	2
4	0	1	0	1
5	0	1	0	2
6	0	1	1	0
7	1	0	0	1
8	1	0	0	2
9	1	0	1	0
10	1	1	0	0
11	1	1	1	1
12	1	1	1	2

Locating Arrays

- Very little is known about locating arrays.
  - Fortunately, similar to covering arrays, their size is logarithmic in the number of factors!

- In the Wi-Fi audio streaming conferencing scenario, we control 24 potentially-relevant factors:
  - The full-factorial design is infeasible, with $> 10^{12}$ tests!
  - A $(1, 2)$-locating array has only 109 tests.

Locating Arrays

- Very little is known about locating arrays.
  - Fortunately, similar to covering arrays, their size is logarithmic in the number of factors!

- In the Wi-Fi audio streaming conferencing scenario, we control 24 potentially-relevant factors:
  - The full-factorial design is infeasible, with $> 10^{12}$ tests!
  - A $(1, 2)$-locating array has only 109 tests.

- In another mobile wireless scenario, we controlled 75 parameters spanning the MAC to the transport layer:
  - Again, the full-factorial design is infeasible, with $> 10^{43}$ tests!
  - A $(1, 2)$-locating array has only 421 tests.

Locating Arrays

- Very little is known about locating arrays.
  - Fortunately, similar to covering arrays, their size is logarithmic in the number of factors!
- In the Wi-Fi audio streaming conferencing scenario, we control 24 potentially-relevant factors:
  - The full-factorial design is infeasible, with $> 10^{12}$ tests!
  - A $(1, 2)$-locating array has only 109 tests.
- In another mobile wireless scenario, we controlled 75 parameters spanning the MAC to the transport layer:
  - Again, the full-factorial design is infeasible, with $> 10^{43}$ tests!
  - A $(1, 2)$-locating array has only 421 tests.
- Trade-off: Analysis is more complex because LA's are often highly unbalanced.

Table of Contents

The Three R's

Reproducibility is a fundamental part of the scientific method.

- It is different from repeatability where researchers repeat their own experiment to verify their results, and

The Three R's

Reproducibility is a fundamental part of the scientific method.

- It is different from repeatability where researchers repeat their own experiment to verify their results, and
- replicability where an independent group of researchers uses the original experimental set-up to verify results.

The Three R's

Reproducibility is a fundamental part of the scientific method.

- It is different from repeatability where researchers repeat their own experiment to verify their results, and
- replicability where an independent group of researchers uses the original experimental set-up to verify results.
- Reproducibility consists of a replication study performed by an independent group of researchers using their own experimental set-up to confirm the results and conclusions of an earlier experiment.

Tools to Help with the Three R's

Tools are starting to emerge to help:

- The Chameleon precís.
- Install/execute scripts.
- Snapshot system after it is configured and boot from VM, *e.g.*, Docker.
- Sysadmin tools such as Ansible.
- GENI-lib.
- Follow a DevOps approach, *e.g.*, using Popper, Jupyter.
- Use and/or generation of traces, both of data and the system.

Among others!

ANSIBLE

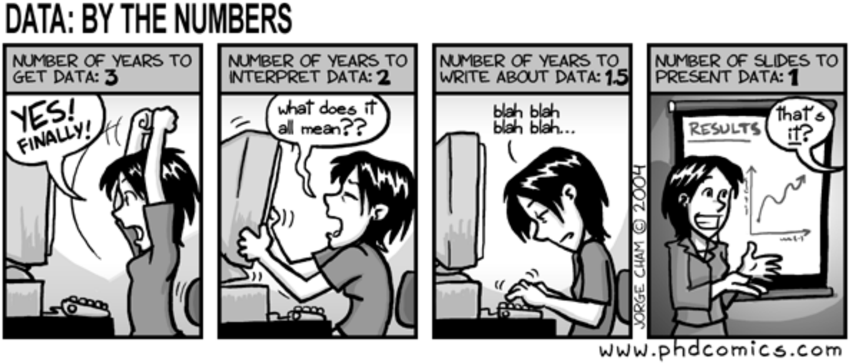
docker

jupyter

A Few Suggestions

1. Follow the guidelines for designing experiments. In particular.
   - Choose an experimental strategy.
   - Collect statistically sound data (remember the principles of replication and randomization).
   - Analyze the results properly.
     - Plot the data!
   - Present the data in a coherent and illustrative manner.
2. Familiarize yourself with tools to help!

Questions?



ARIZONA STATE UNIVERSITY

geni Exploring Networks of the Future

NSF