

# Upcoming Classes

March 6 - Presenting?

- Report 2 Experimental Plan due

Spring Break!

March 20 - Presenting?

- Report 3 Literature Survey due

March 27 - Statistics in Experiments by Pablo

April 3 - Personal Productivity

- Report 4 Experimental Results due

# Mini Tutorials

Plan a 30 minute tutorial on a tool/topic

- 1-3 students
- Mixture of slides and hands-on examples
  - Try to involve class participation
- Practice presenting clearly and interacting with students
- Audience: 1st year CS PhD student

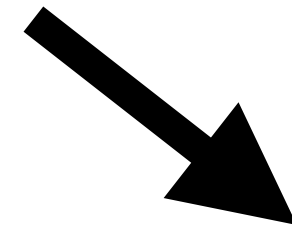
Topics?

- Jupiter notebooks
- Advanced Latex
- Tensorflow 101
- Shortest Paths algorithms (something more exciting than this)

Will start after Spring Break

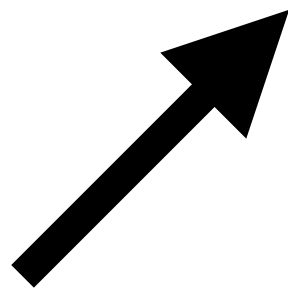
# Recipe: Introduction

**But imagine how wonderful  
it could be if we could figure  
out how to do X!**



**My work helps us get  
one step closer to the  
magical dream world!**

**The world is a  
terrible, terrible  
place.**



# Recipe: Abstract

This is a mini version of an intro

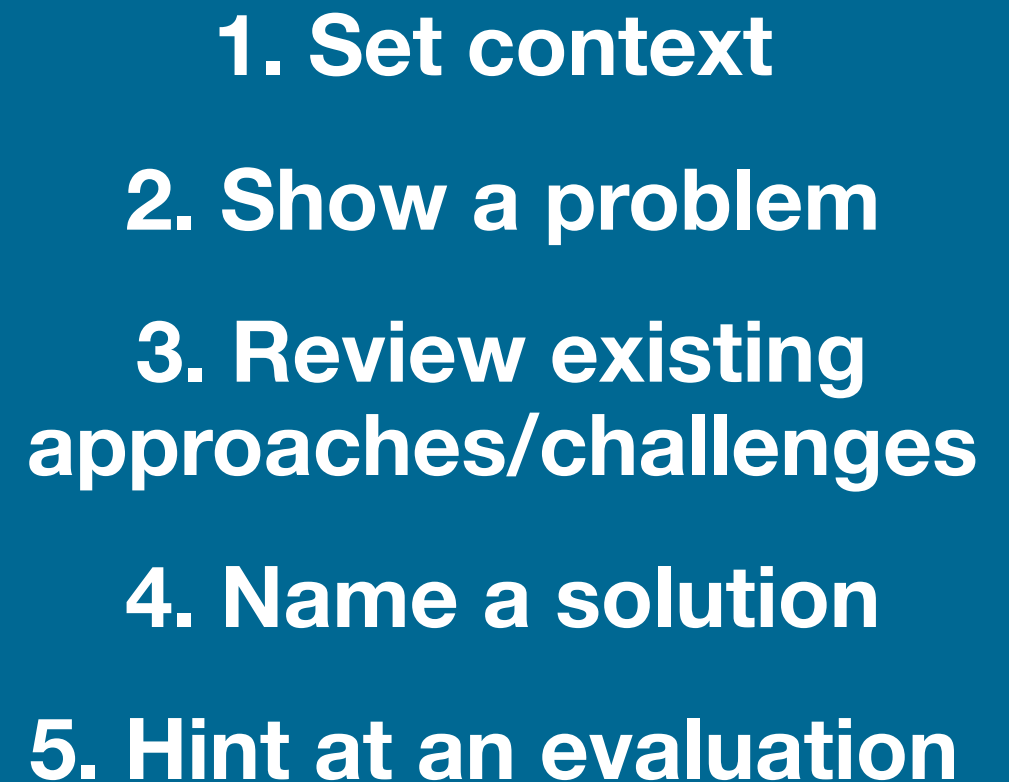
1. Set context
2. Show a problem
3. Review existing approaches/challenges
4. Name a solution
5. Hint at an evaluation

# Recipe: Writing a Paper

1. Write a 2 paragraph abstract
  - High level brain dump of problem and goals
2. Write a title for the paper
3. Add titles for all sections and subsections
4. Outline key sections
  - One bullet point per paragraph
5. Sketch key figures
  - System design, algorithm flow
  - Predicted experimental results

Do steps 1-2 now!

- Use [overleaf.com](https://overleaf.com) or your favorite paper writing tool

- 
- 1. Set context**
  - 2. Show a problem**
  - 3. Review existing approaches/challenges**
  - 4. Name a solution**
  - 5. Hint at an evaluation**

# Books!

# Experimental Design

CSCI 8901:  
Research & Evaluation Methods

Prof. Tim Wood  
GWU



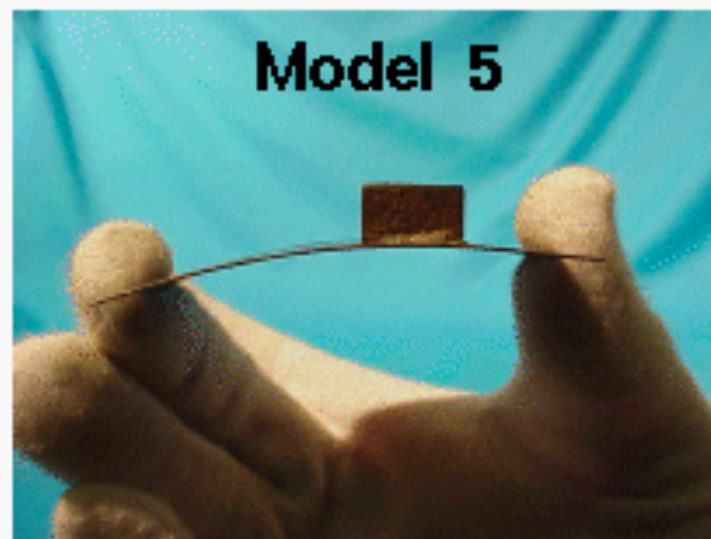
# Why Experiment?

Experiments allow us to evaluate a system or theory

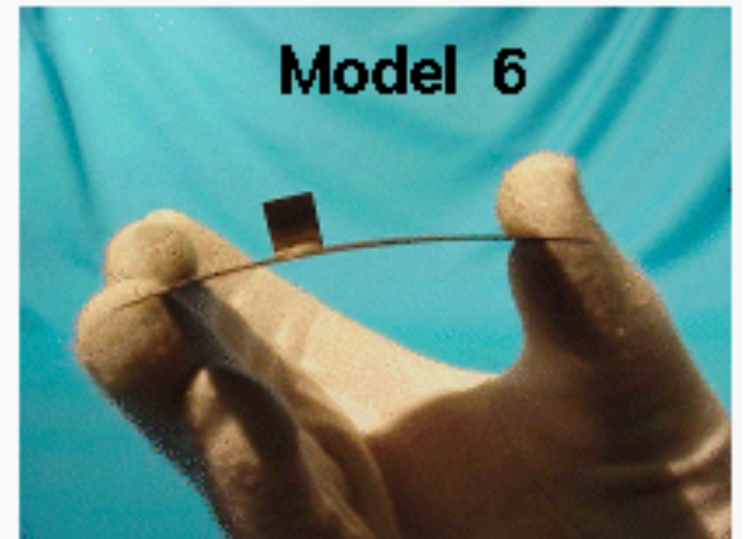
- Validate claims
- Comparison



**Model 4**



**Model 5**

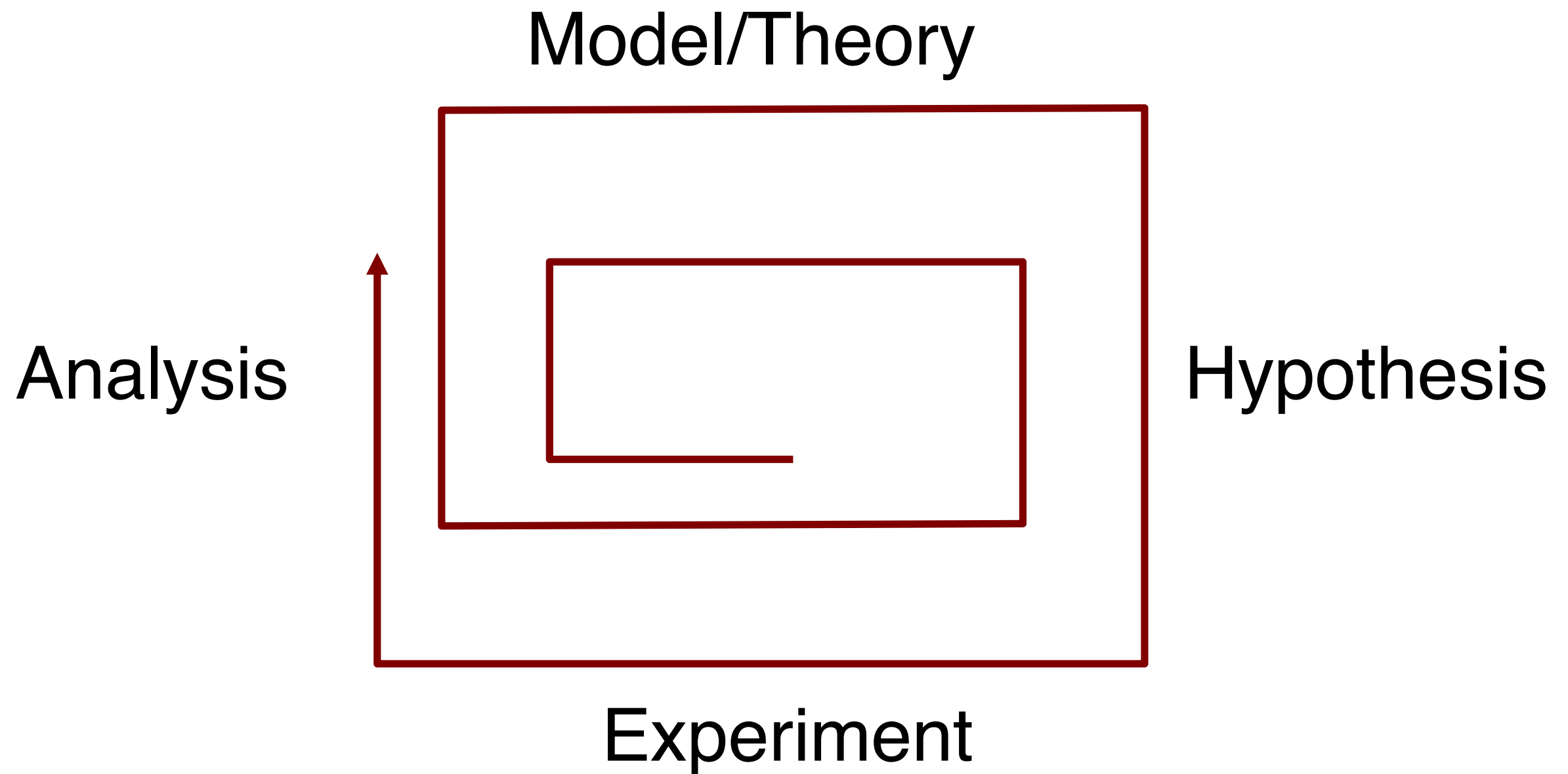


**Model 6**

**Camber test, arc foils, low aspect ratio**



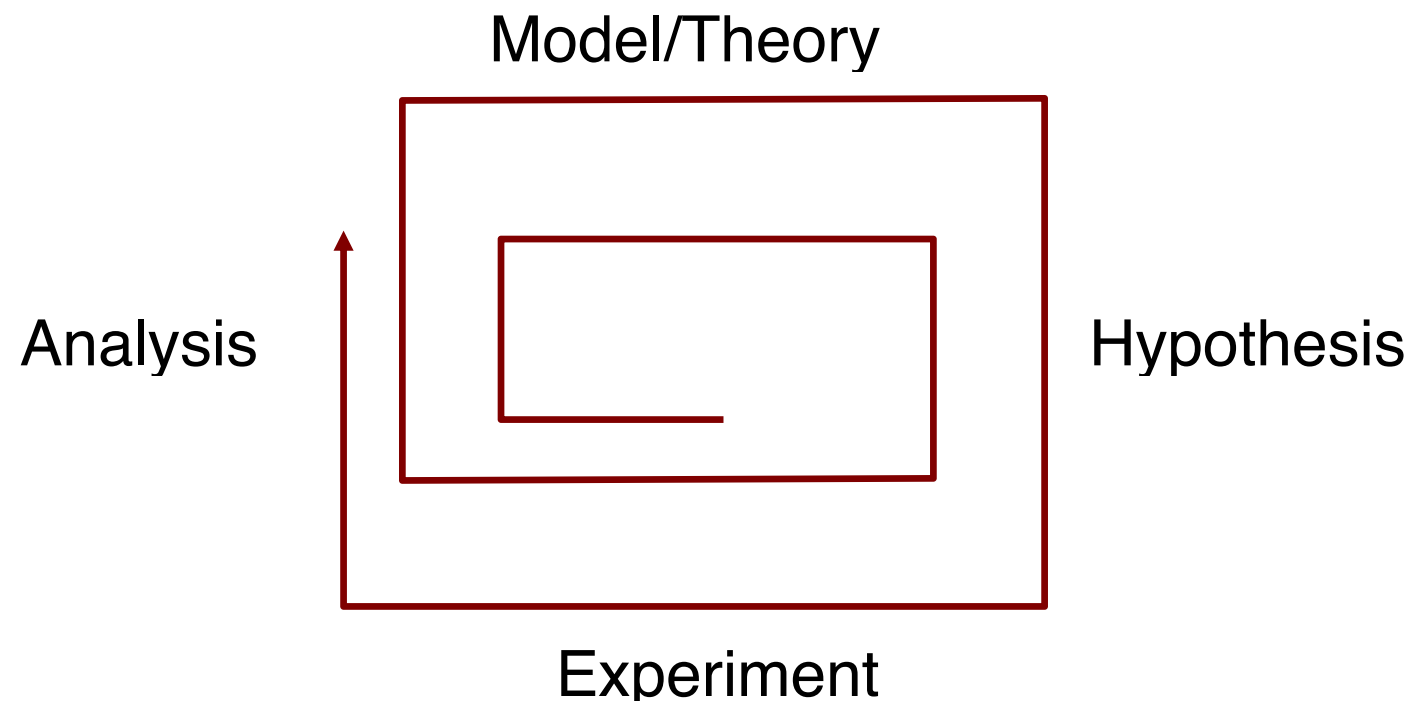
# The Loop



From Ido Dagan, The Basis for Experimental Design lecture

# Hypotheses

1. Design a model/theory/system
2. Predict how it will work
  - These predictions are our hypotheses
3. Run experiments to validate the hypotheses
4. Analyze the results and update our approach



# Testability

A hypothesis must be **testable**

- We can design an experiment to observe the accuracy of the hypothesis

Hypothesis: *There are parallel universes to ours that we can never contact or interact with.*

By the definition of the hypothesis, we will not be able to test if it is true.

- If we can interact with a parallel universe it has nothing to do with this hypothesis

# Falsifiability

A hypothesis must be **falsifiable**

- An experiment could produce some result which would disprove the hypothesis

Hypothesis: *There are planets in our universe inhabited by alien life.*

We can test this by going to every planet. But what observation could we make to show it is false?

- We would need to enumerate all possible planets in the universe, but that is not feasible

# Swans

**Hypothesis:** *All swans are white.*

Testable:

- Go to a pond and look at a swan. Is it white?

Falsifiable:

- If we ever find a red swan, then we know the hypothesis is false





# Swans

**Hypothesis:** *All swans are white.*

Testable:

- Go to a pond and look at a swan. Is it white?



Falsifiable:

- If we ever find a red swan, then we know the hypothesis is false



Thought to be true!

- Until Willem de Vlamingh explored Australia

# Truth?

**Hypothesis:** *All swans are white or black.*

Can we prove this hypothesis is true?

# Truth?

**Hypothesis:** *All swans are white or black.*

Can we prove this hypothesis is true?

No!

- We can gather a lot of evidence to support our claim, but often we cannot prove our claim is correct

That's OK.

Note: Induction doesn't work in experimental science

# Bottom Line

Hypotheses in CS are helpful to ensure you are being rigorous and unbiased

But, they don't always fit the structure of how we think about evaluating the work that we do

## Key Idea:

- You must have a mental model of how your system/approach works
- Design your experiments to evaluate if that mental model is correct
- Use hypotheses to guide experimental design and sanity check your results

# Planning Your Experiments





# Importance

Experiments will give evidence to support your hypotheses

- Typically the claimed contributions of your work

Top level hypotheses are easy to define for your primary research claim:

- *Deep neural networks improve event detection and classification accuracy.*
- *Pseudo-Linear cryptanalysis can break the Speck cipher more effectively than linear cryptanalysis.*

# Comparisons

My system can analyze your face and predict if you are left or right handed.

- I tested on 100 people and it was right 90% of the time!

# Comparisons

My system can analyze your face and predict if you are left or right handed.

- I tested on 100 people and it was right 90% of the time!

Actually, it just always predicts you are right handed and 90% of people are right handed

**Performance results are only meaningful with a baseline to compare against!**

- Also, performance must account for different types of errors!

**VERY important to reviewers!**

# Micro vs Macro

Typically you want a mixture of experiments:

**Macro behavior:** How well does it work?

- Overall performance/effectiveness under realistic inputs
- High level comparison against other approaches
- Better/worse, linear/nonlinear

**Micro behavior:** Why does it work so well?

- Finer grain analysis of behavior in carefully tuned settings
- Constrain **environment** or **task** to explain the **behavior**
- Identifies causes, explains behavior

# Sketching

Before running experiments:

- Design several experiments you want to run
- Predict the behavior you expect to see
- Sketch out the graphs you plan to generate

Use this to check for:

- Good comparisons
- Diversity of graph types
- (In)Dependent variables
- Space estimate
- Sanity check experimental result accuracy



# Sketching

Before running experiments:

- Design several experiments you want to run
- Predict the behavior you expect to see
- Sketch out the graphs you plan to generate

Use this to check for:

- Good comparisons
- Diversity of graph types
- (In)Dependent variables
- Space estimate
- Sanity check experimental result accuracy

**Expect to change your plans! Expect the unexpected!**

# Types of Experiments

## Motivational

- Show problem is difficult and current approaches don't work well

## Behavior Comparison

- Compare against state of the art approaches and/or variants of proposed approach

## Microbenchmarks/Error Analysis

- Breakdown performance/cost/accuracy in a very constrained environment

## Case studies

- Illustrative examples of behavior in a realistic setting

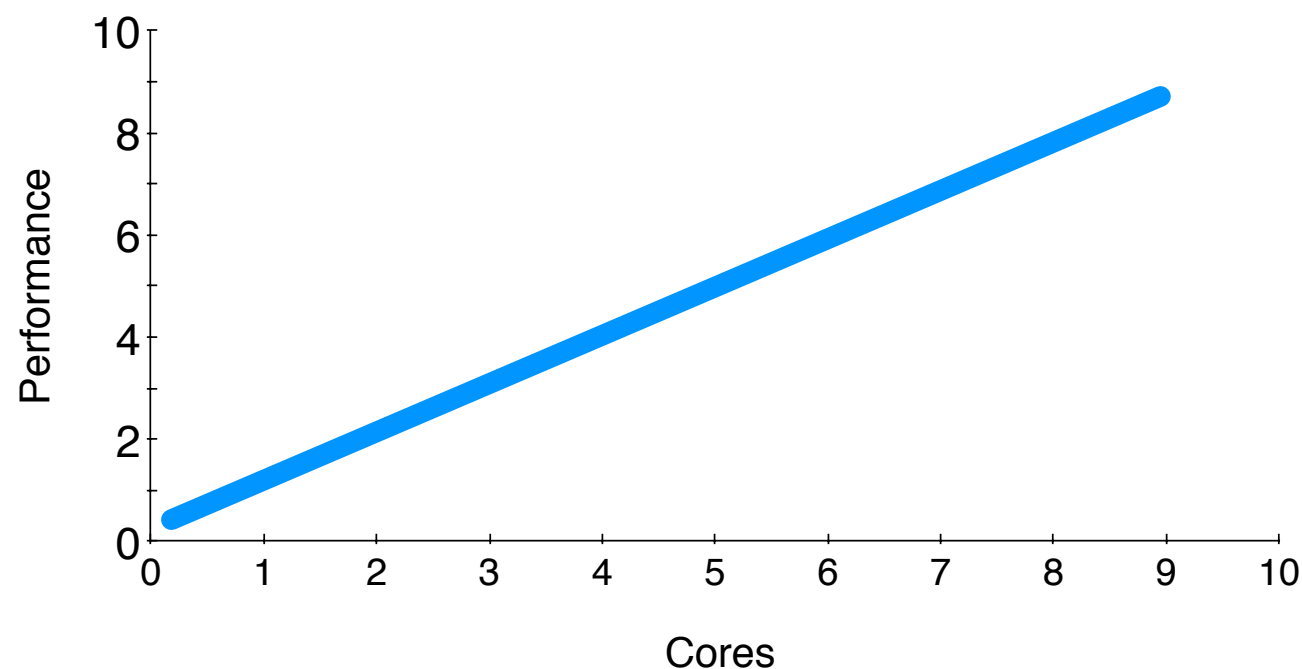
# Predicting Results

Before running an experiment, predict what will happen

- This is related to your hypotheses, but more specific

Examples:

- If I increase the number of CPU cores assigned to my algorithm, performance will increase linearly



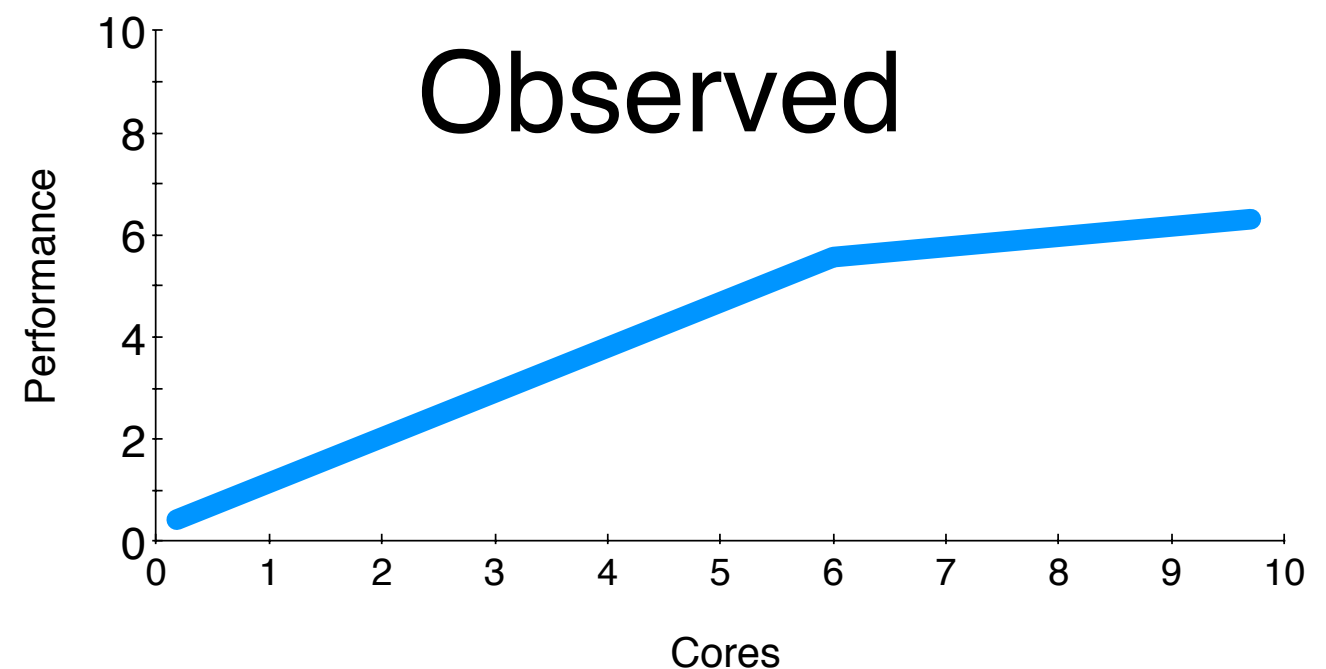
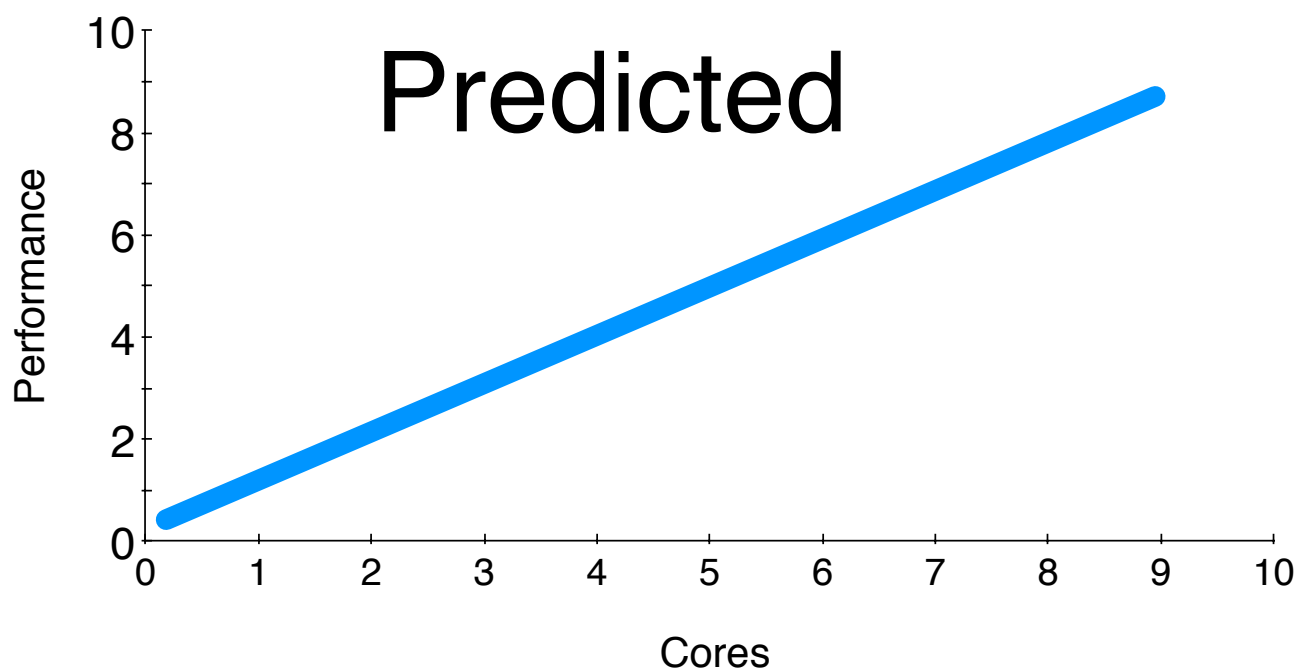
# Predicting Results

Before running an experiment, predict what will happen

- This is related to your hypotheses, but more specific

Examples:

- If I increase the number of CPU cores assigned to my algorithm, performance will increase linearly



# Variable Types

Task, Environment, and System all have variables under your control

- These are **independent** variables you control

Observe the output behavior

- These are **dependent** variables you are trying to understand
- Hopefully they depend on something under your control!

Endogenous variables: on the causal path between your independent and dependent variables

Exogenous variables: external variables also affecting dependent variables



# Isolating Variables

Experiments should have a single “treatment”

- Only manipulate one variable and see its effect relative to control

Avoid confounding variables

- Inability to determine which change produces the result
- (Algorithm 1 + Data Set 1) vs (Algorithm 2 + Data Set 2)

Exogenous factors may be outside our control

- Solution: Random sampling, repeated evaluation to average out their effect

# Factorial Variable Problem

Hypothesis: Hash table lookups are faster than finding elements in arrays

System variables: 3

- Hash table size, hash function, collision algorithm

Task variables: 2

- Number of elements, Key distribution

Environment: 2

- System memory, CPU cores

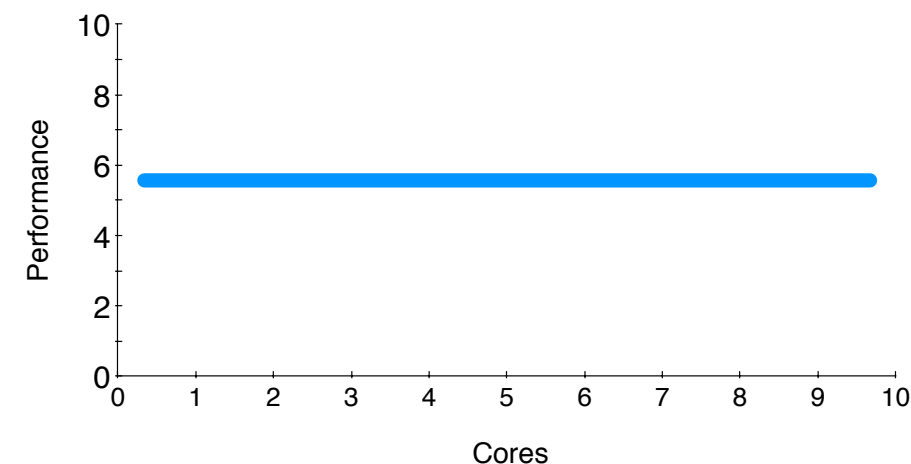
Suppose 10 possibilities for each...  $10^7$  different variations!

# Factorial Variable Problem

Need to carefully pick which variables to focus on and what ranges to test

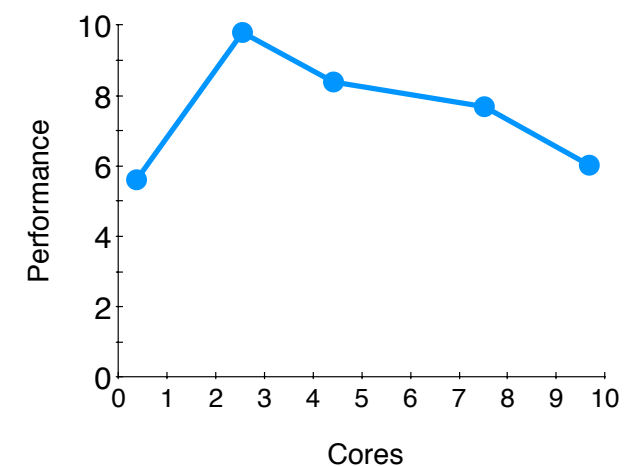
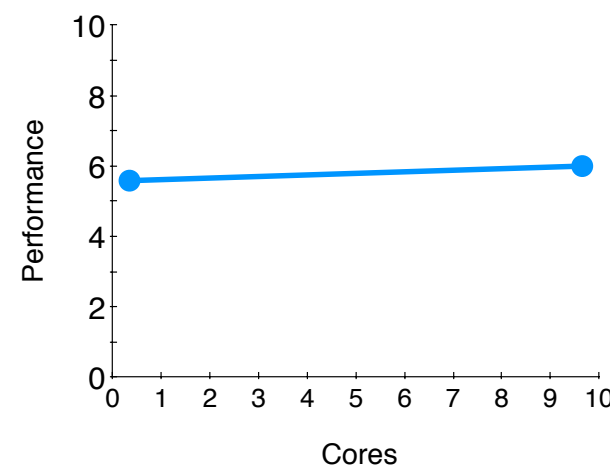
Use intuition to predict relationship

- Rule out variables that should not affect behavior (flat lines)



Test extremes, skip values

- Need at least 3 values to understand trends
- Try to search parameter space efficiently (binary)



# Causality

Good science shows **causal relationships**

- The output behavior occurs because of specific changes to the system, task, or environment

Correlation isn't causality!

Bird have feathers. Birds fly. Therefore, flight requires feathers.

- Is there a causal link between feathers and flight?

Need to consider:

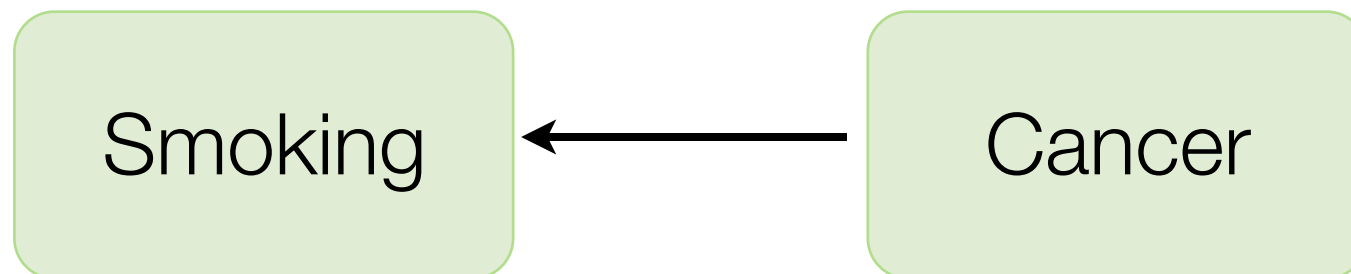
- Association
- Direction, timing
- Exogenous factors

# Causality

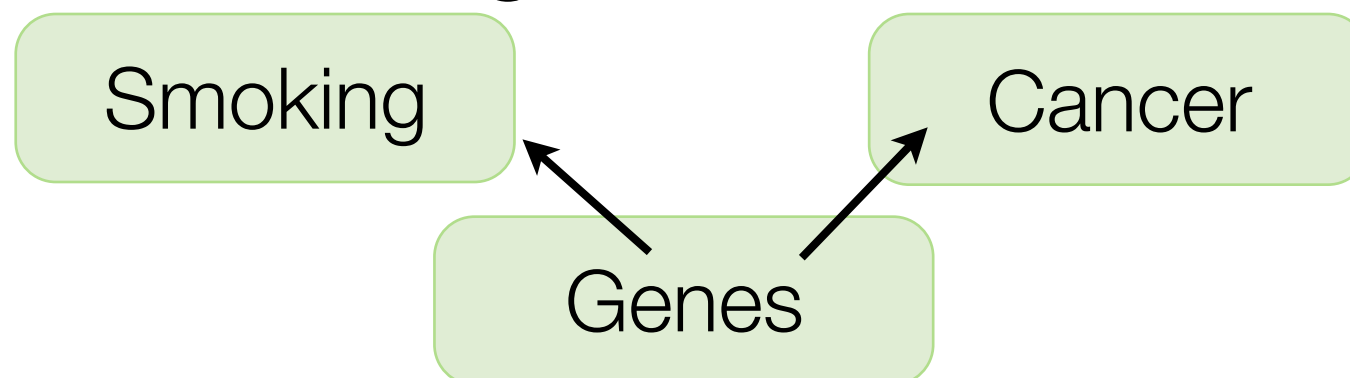
Smoking causes cancer



Cancer causes smoking



Some people have genes that make them want to smoke and those genes also cause cancer





# Causality

## **Good news:**

In CS we control more aspects of the system, task, and environment

- Makes it easier to determine “direction”

## **Bad news:**

Computers are complex systems with many interacting parts

- Difficult to account for all external factors

# Threats to Validity

We want our experiments to be accurate!

Internal validity:

- Experiment shows a real relationship between treatment and dependent variables

External validity:

- The relationship demonstrated in the experiments is generalizable to other scenarios (e.g., tasks, environments)

Need to avoid conditions that affect these

# Internal Validity

*Do changes of independent variables truly cause changes in dependent variables?*

History: unanticipated event during experiment that disrupts output behavior

- Virus scanner starts in background and slows down analysis

Maturation: Changes in output is affected by timing

- CPU cache warms up over time, causing early experiments to be slower and later experiments to be faster

Regression: Extreme scores tend to regress to mean

- Change in behavior of outlier data sets

# Internal Validity

*Do changes of independent variables truly cause changes in dependent variables?*

Selection: groups being studied not identical prior to experiment

- Control runs on server A, my system on server B... what if they aren't identical? (Worse when dealing with humans)

Mortality: some inputs fail/drop out

- Web server with admission control shows very low latency, but only because it drops many connections. (Worse with humans)

Instrumentation: act of measurement affects results

- Additional logging during experiment might slow down system

# External Validity

*Do the relationships shown in an experiment generalize to other tasks and environments?*

Population Validity: Is the input task representative of real world tasks?

- iPhone face ID works great... unless you are Asian. Training data set was not representative of real world users

Ecological Validity: Does the experiment account for factors from the real world environment?

- Image recognition algorithm fails when paired with lower resolution camera in low-light conditions



# Recipe: Experimental Design

1. Have something to compare against
  2. Consider and isolate the most important variables
  3. Plan experiments to show:
    - How well your system does **compared** to a baseline
    - **Why** your system does well
  4. Predict results and sketch graphs before starting
  5. Run experiments
  6. Ensure results are repeatable and significant
    - Think about threats to internal and external validity
- (Throughout) Iterate and feedback as needed

# Resources/Acknowledgements

## Internal/External Validity:

- [http://www.indiana.edu/~educy520/sec5982/week\\_9/520in\\_ex\\_validity.pdf](http://www.indiana.edu/~educy520/sec5982/week_9/520in_ex_validity.pdf)

## Statistical Methods in CS by Ido Dagan:

- <http://u.cs.biu.ac.il/~shey/362-2010/Lectures.htm>

## Testable/Falsifiable Hypotheses:

- <http://www.batesville.k12.in.us/physics/PhyNet/AboutScience/Hypotheses.html>