Slides adapted from Henny Admoni

16-867

Experimental Design

Instructor: Andrea Bajcsy

Carnegie Mellon University

intent ROBOTICS LAB

Why Do HRI Studies?

- Validate that a system works as expected
- Compare two or more systems or algorithms
- Explore a phenomenon to develop a research question
- Collect training data for a model or train an algorithm

How to Conduct HRI User Studies

1. define the research question and hypothesis
2. design a study to address that question
3. execute the study
4. analyze data from the study
5. draw conclusions from the analysis

How to Conduct HRI User Studies

1. **define the research question and hypothesis**
2. design a study to address that question
3. execute the study
4. analyze data from the study
5. draw conclusions from the analysis

Basic vs Applied Research Questions

Basic	Applied
generalizable	context-specific
laboratory	real world
rigidly controlled	uncontrolled factors
artificial	ecologically valid
inferential analysis	descriptive analysis

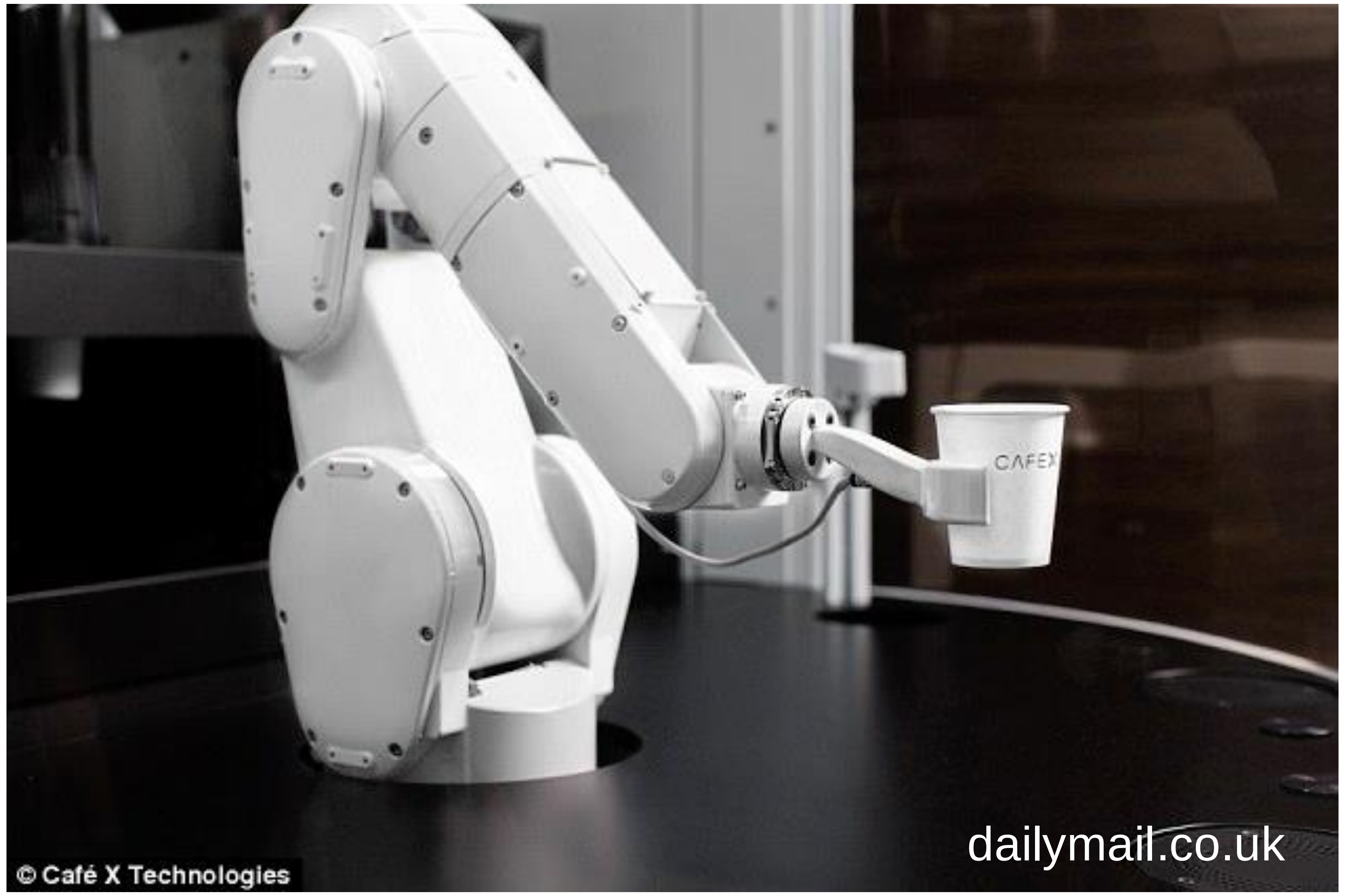
How applicable are experimental findings to the real world? (ecologically valid)

Inferential analysis: *Using data to infer / predict properties of a population, e.g. by comparing differences between treatment groups*

Descriptive analysis: *Summarizes observed data (e.g., mean, std. deviation)*

Example 1

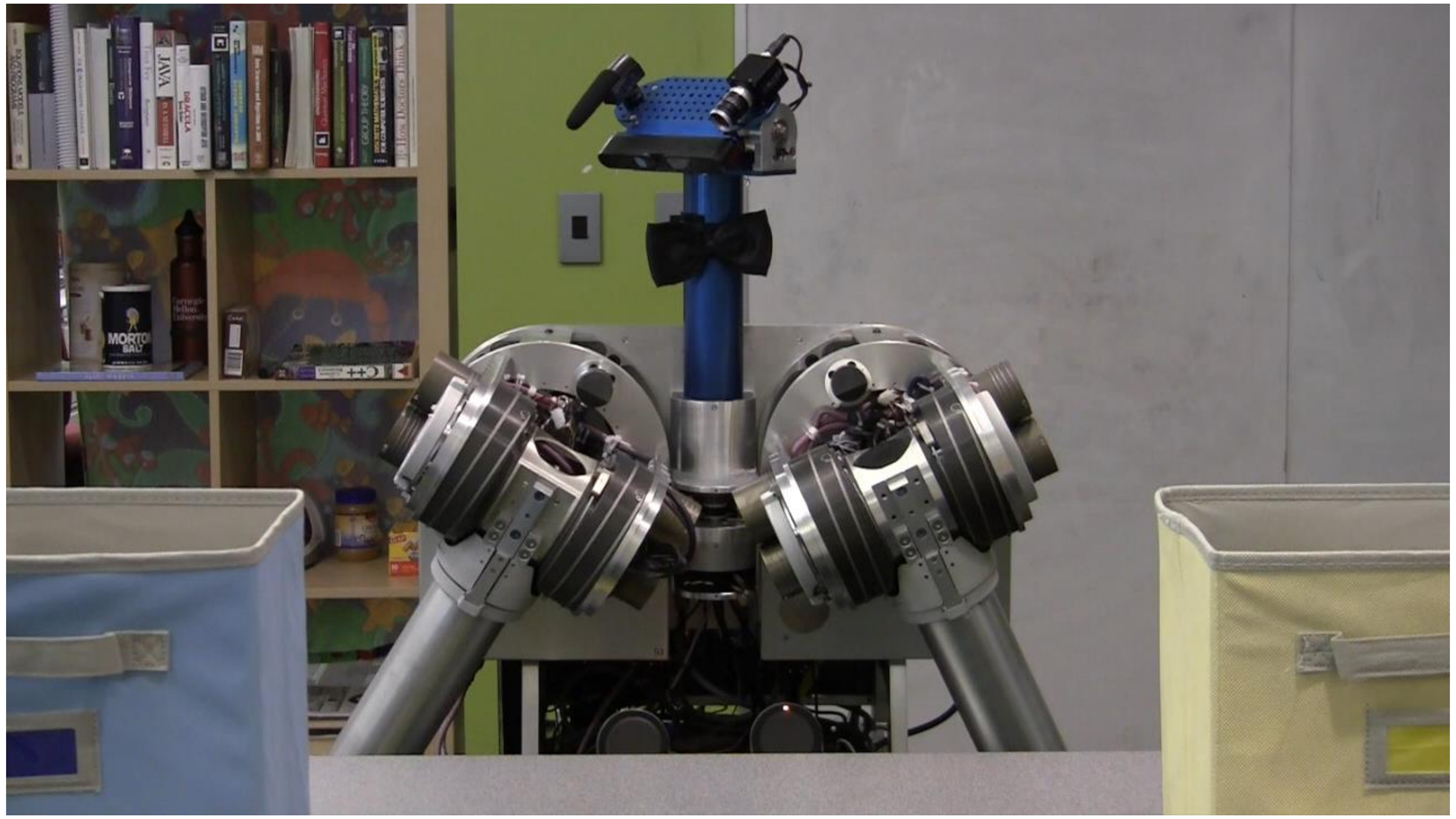
- My robot makes espresso in the PIT airport. I've updated the algorithm to monitor the water temperature.
- *Question*: How much more consistent is espresso temperature for the new robot algorithm compared to the old one?



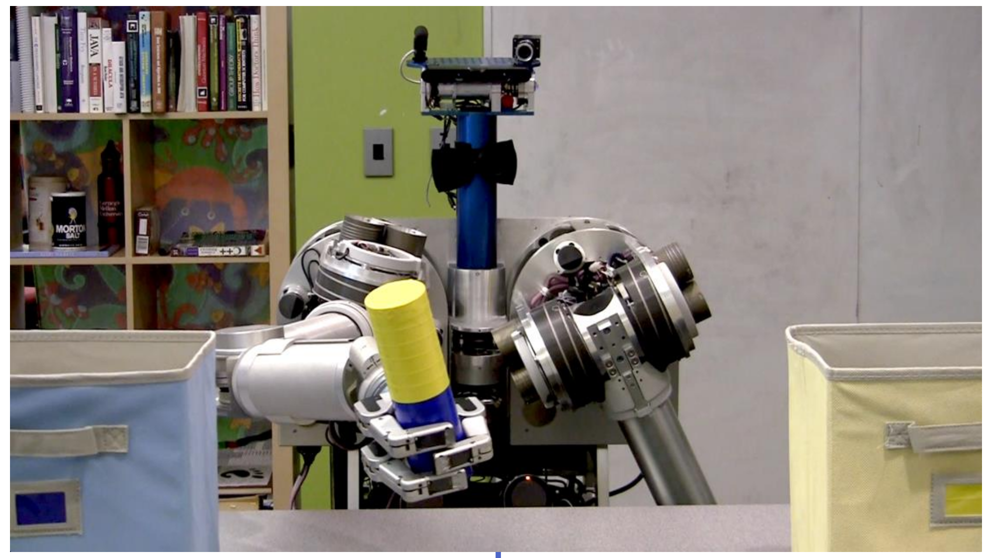
Example

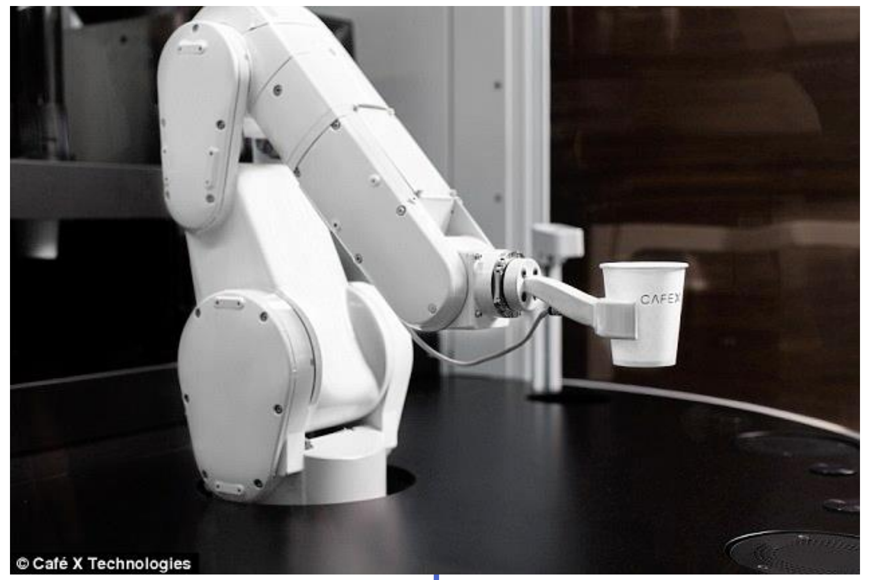
- My robot hands over objects. I've updated my algorithm to initiate and monitor human attention before the handover starts.

- *Question:* Does establishing joint attention before a handover help handovers occur more efficiently?



Basic vs Applied Research Questions





dailymail.co.uk

Basic ⟷ Applied

Basic	Applied
generalizable	context-specific
laboratory	real world
rigidly controlled	uncontrolled factors
artificial	ecological validity
inferential analysis	descriptive analysis

Variables and Metrics

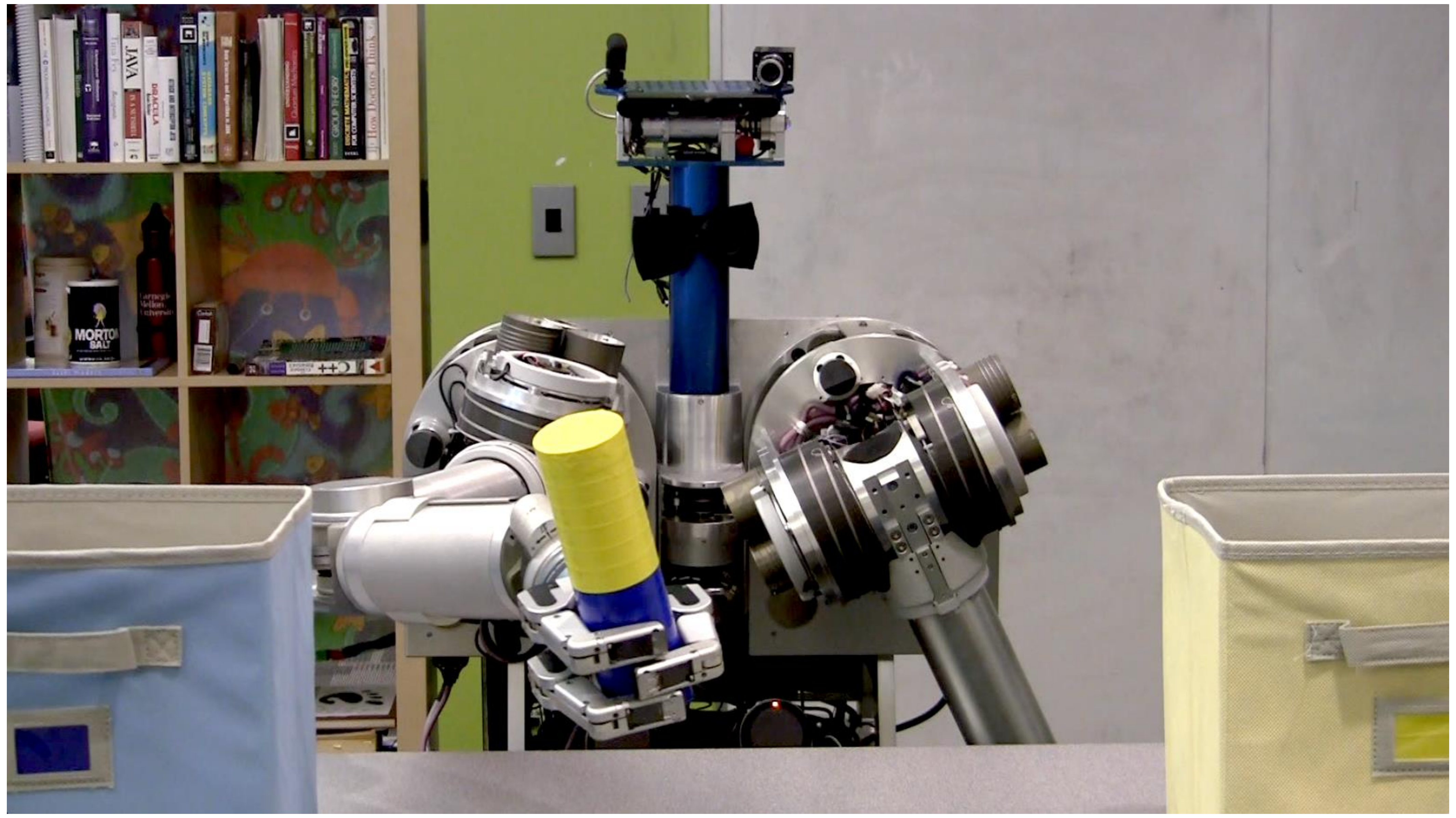
- Independent variable (IV) - causal factor that we can manipulate

  Joint attention during handover

- Dependent variable (DV) - factor that is influenced by the IV

  Handover efficiency

- Metric - how we measure the effect of IV on DV



Number of successful robot-to-human handovers

Hypotheses

Typical starting hypothesis formulation:

"The effect of ________ on ________ will be ________."

independent variable

dependent variable

as measured by metric

A hypothesis for our example:

The effect of joint attention on handover efficiency will be to increase the number of successful handovers.

Using joint attention will improve a robot's handover efficiency by increasing success rate.

Features of a Good Hypothesis

- Makes a specific prediction
  - a hypothesis is either *supported* or *not supported* by the data
- Is measurable
  - “better” isn’t measurable; “higher success rate” can be measured
- Addresses your research question
  - should be actually related to the problem, and not trivial

How to Conduct HRI User Studies

1. define the research question and hypothesis
2. **design a study to address that question**
3. execute the study
4. analyze data from the study
5. draw conclusions from the analysis

Design the Study

- Specify the study's *structure*
- Select *metrics*
- Define *procedure*
- Define *population*

Design the Study: *Structure*

- Specify the study's *structure*
  - how many IVs?
  - how many levels for each IV?
  - how will participants be assigned to conditions?

Independent Variables Have Levels

- Think of IVs as dials that can be turned to different settings

vector stock.com/20978533

Independent Variables Have Levels

- We can manipulate the level (setting) of an IV (dial)
  - two levels: no drug treatment vs. 100mg drug treatment
  - three levels: 100mg vs. 200mg vs. 300mg
  - also three levels: drug A vs. drug B vs. drug C

levels are *absence* or *presence* of IV

levels are *amount* of IV

levels are different *types* of a single IV

- A control is a special level which is the baseline against which other levels of IV are measured

Study Designs

Factorial Design (two IV, two levels each)

	IV 1 level 1	IV 1 level 2
IV 2 level 1	Condition 1	Condition 2
IV 2 level 2	Condition 3	Condition 4

IV 1 IV 2

2 x 2 design

two-way design

Study Designs

Factorial Design (two IV, differing number of levels each)

	IV 1 level 1	IV 1 level 2	IV 1 level 3
IV 2 level 1	Condition 1	Condition 2	Condition 3
IV 2 level 2	Condition 4	Condition 5	Condition 6

IV 1 IV 2

3 x 2 design

two-way design

Assigning Participants to Conditions

- When deciding how to split participants up across conditions, ask:
  - *would seeing more than one condition be problematic?*
  - *is there a lot of interpersonal variability in the metric?*
  - *how many participants are available?*

Assigning Participants

- Between subjects - each participant experiences only one level
  - useful if participant seeing multiple levels is a problem
  - good for large participant pools or for short study sessions

P1, P2, P3

Condition 1

P4, P5, P6

Condition 2

Assigning Participants

- Within subjects - each participant experiences all levels
    - accounts for interpersonal variability
    - efficiently uses your participant pool
    - susceptible to ordering effects

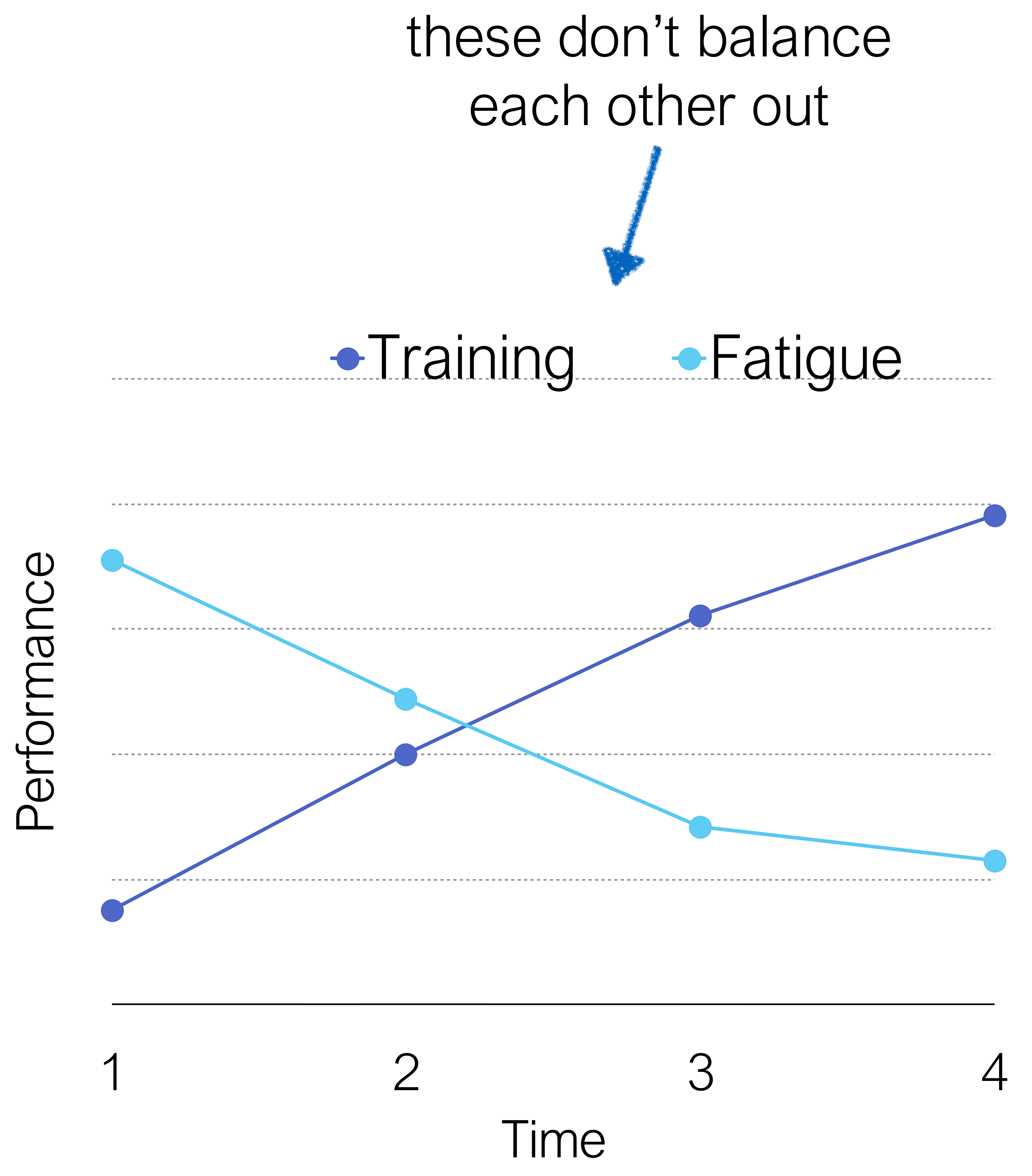
P1, P2, P3
P4, P5, P6
Condition 1

P4, P5, P6
P1, P2, P3
Condition 2

Confounds

- *Training effect* - performance improves because of practice

- *Fatigue effect* - performance declines because of tiredness



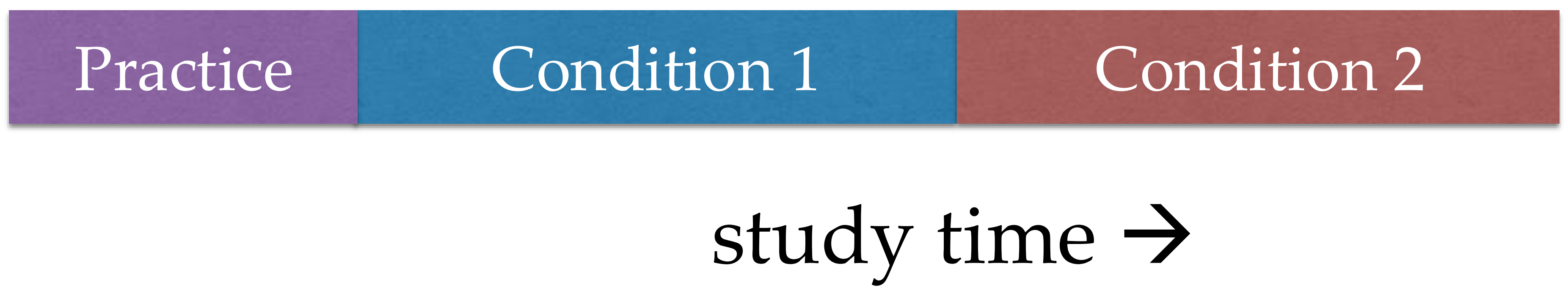
Confounds

- *Novelty effect* - performance/impression differs on first exposure to an HRI system
- *Personal characteristics* - performance influenced by age, gender, expertise, etc.

Joshua Ellingson for Willow Garage

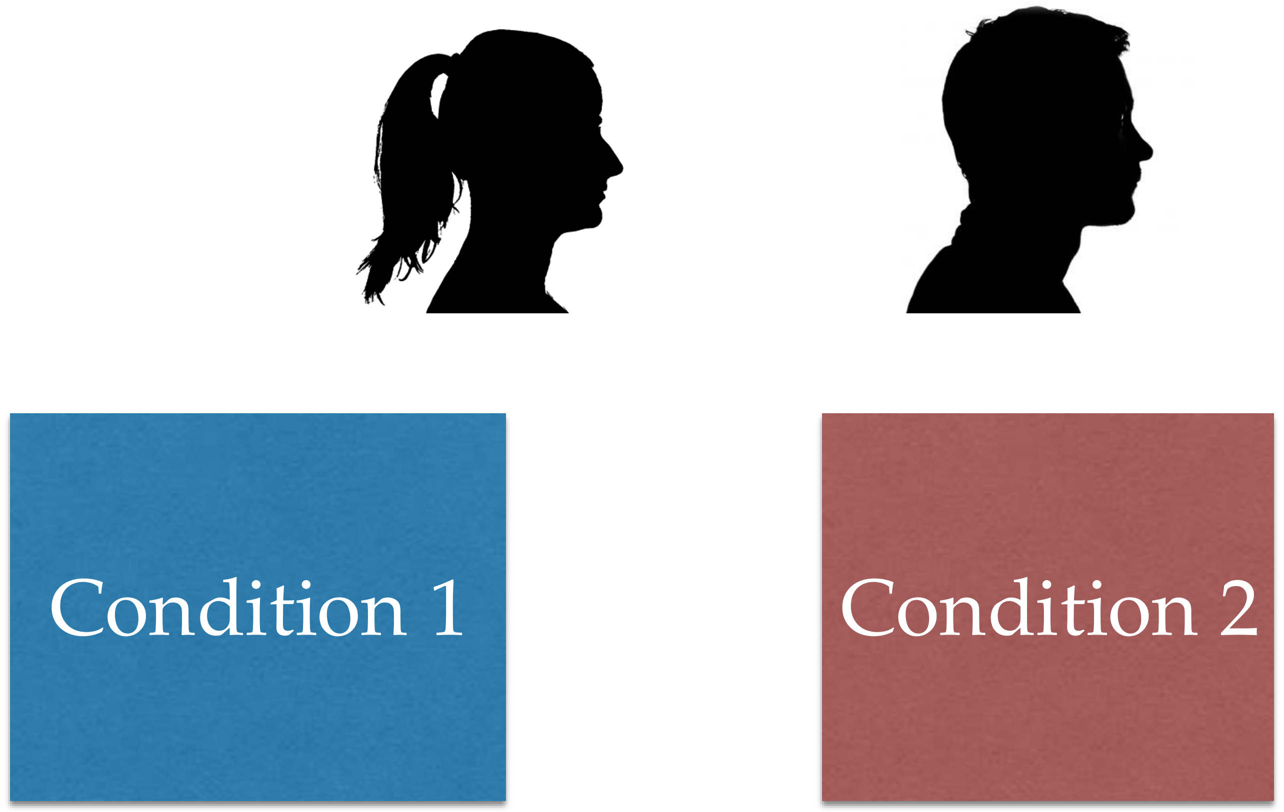
Addressing Confounds

- **Method 1:** *Modify the procedure*
    - Counterbalancing - balance order in which participants are exposed to conditions
    - Pre-study practice - provide unrecorded time to familiarize participants with the system



Addressing Confounds

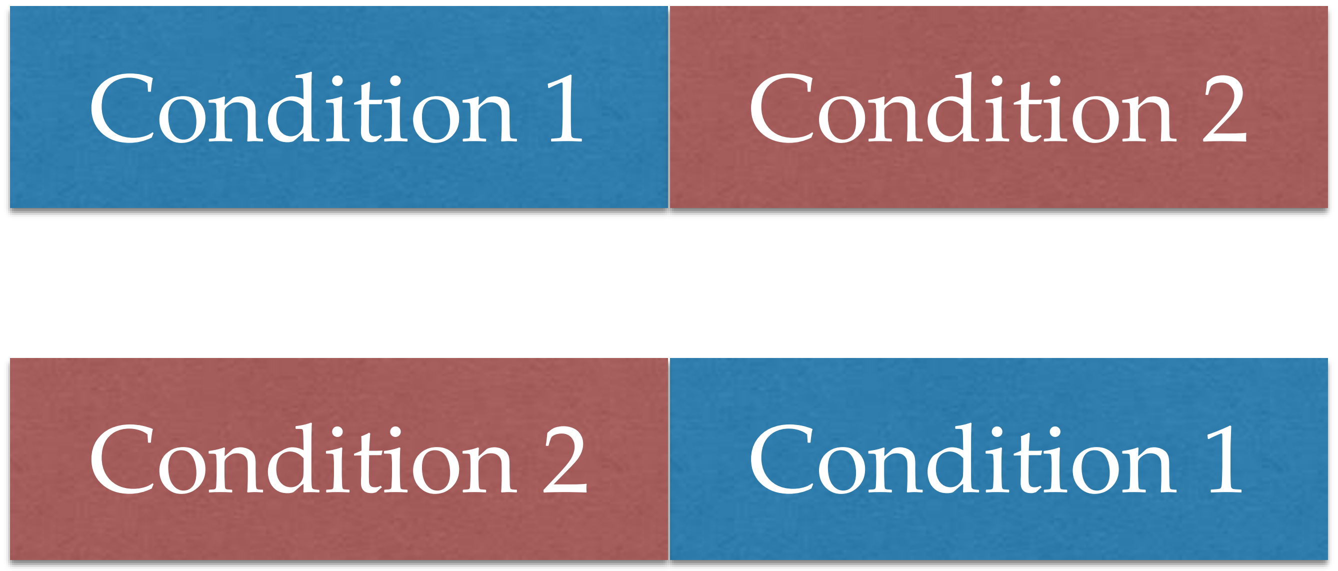
- **Method 2:** *Assign participants strategically*
  - Random group assignment - pre-assign groups to get approx. even distribution
  - Matched group assignment - recruit and assign specific people to specific groups based on characteristic to control, e.g., gender



Addressing Confounds Summary

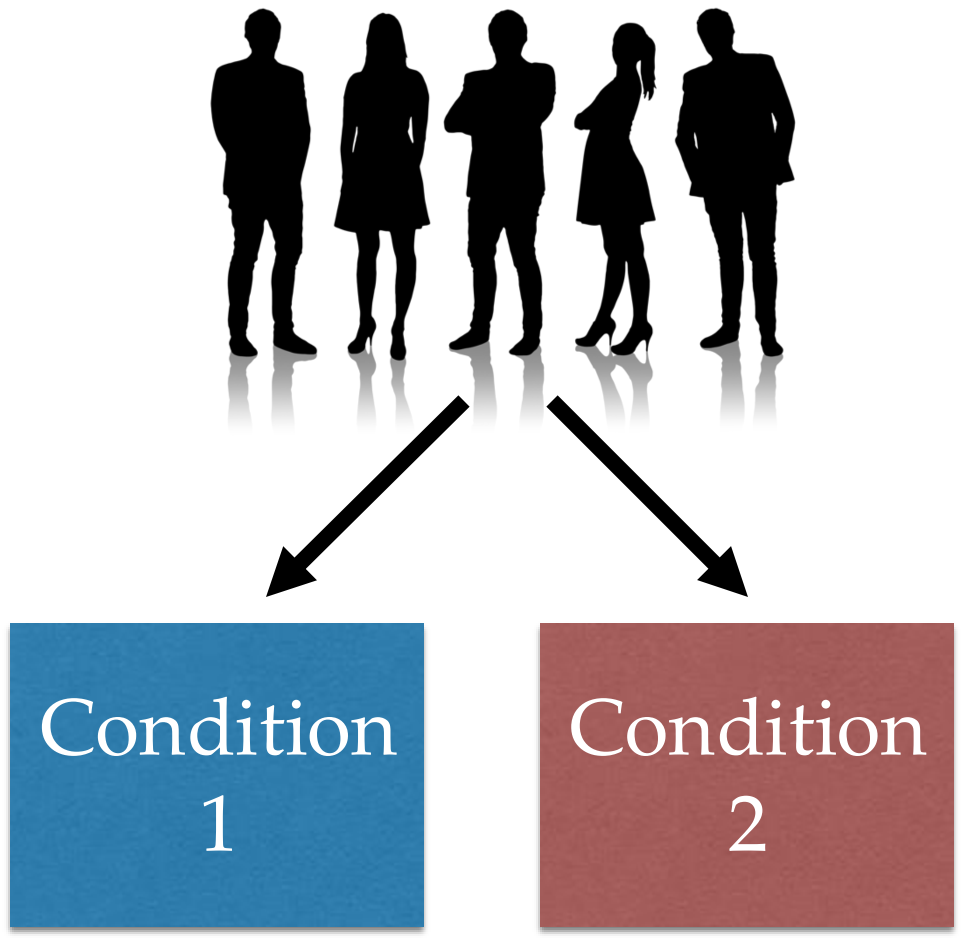
Within-subjects

- Counterbalancing



Between-subjects

- Random or matched group assignment



Both

- Pre-study practice

Design the Study: *Metrics*

- Specify the study's *structure*
- Select *metrics*
  - what dependent variables?
  - how will they be measured?
- Define *procedure*
- Define *population*

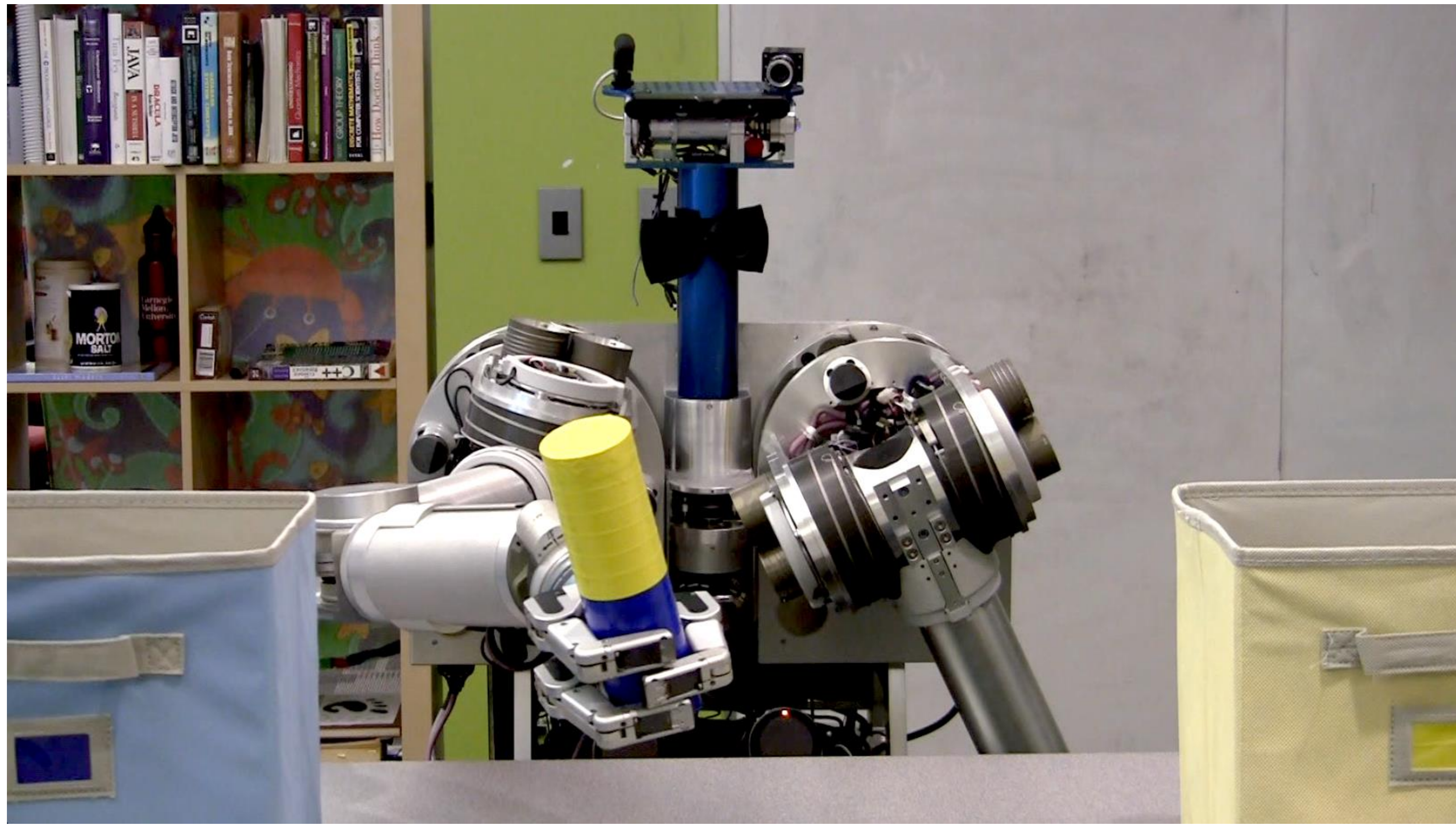
Selecting Metrics

- What do you want to measure?

  handover success rate
  (for each algorithm)

- How will you perform this measurement?

$$\frac{N\,successful}{N\,total}$$

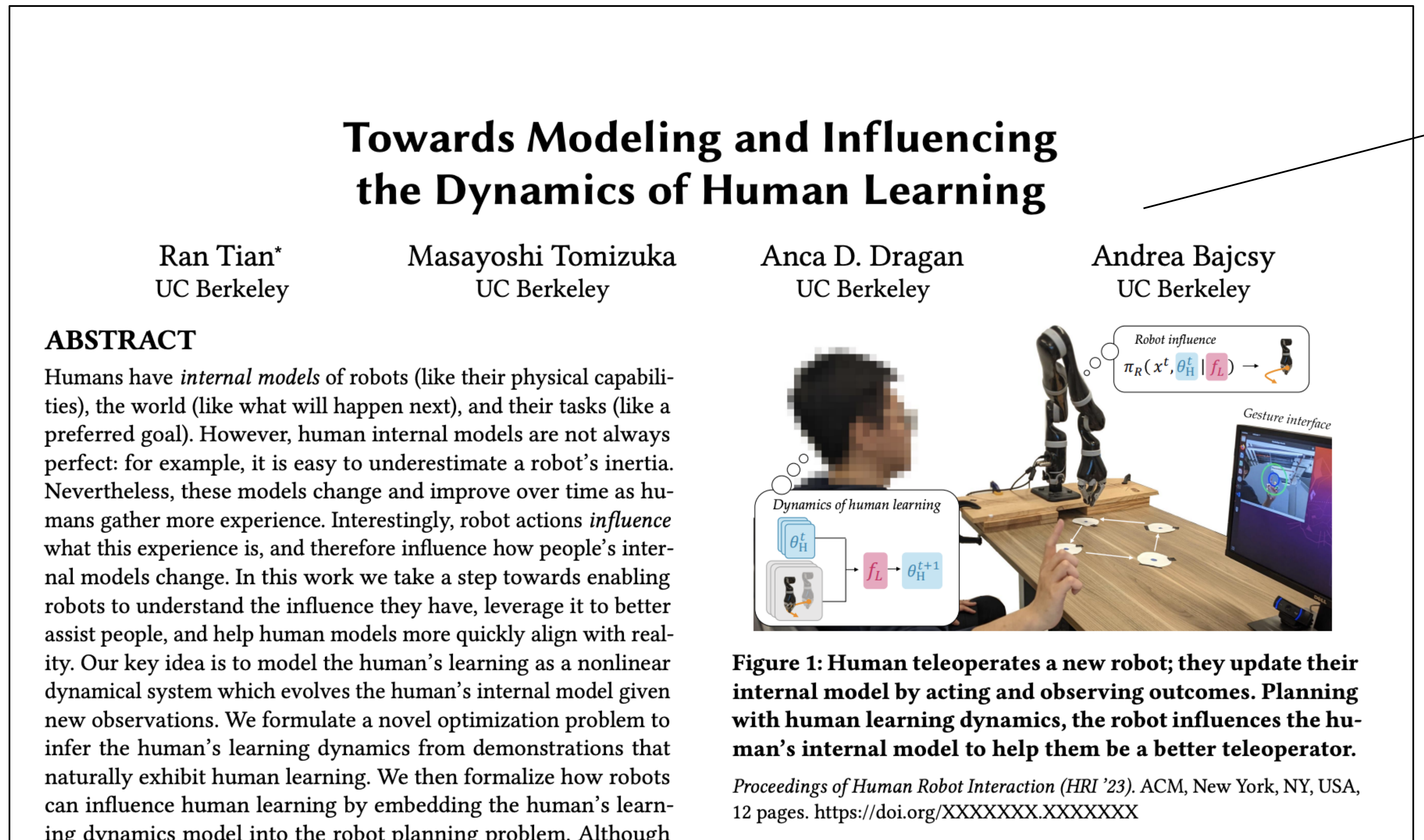


Types of Metrics

- Objective vs. Subjective
- Qualitative vs. Quantitative

Objective Metrics

- Objective - observable
  - usually task-based measures like success rate, completion time, etc.
  - can reveal difference between perceived and measured experience



Towards Modeling and Influencing the Dynamics of Human Learning

Ran Tian* — UC Berkeley
Masayoshi Tomizuka — UC Berkeley
Anca D. Dragan — UC Berkeley
Andrea Bajcsy — UC Berkeley

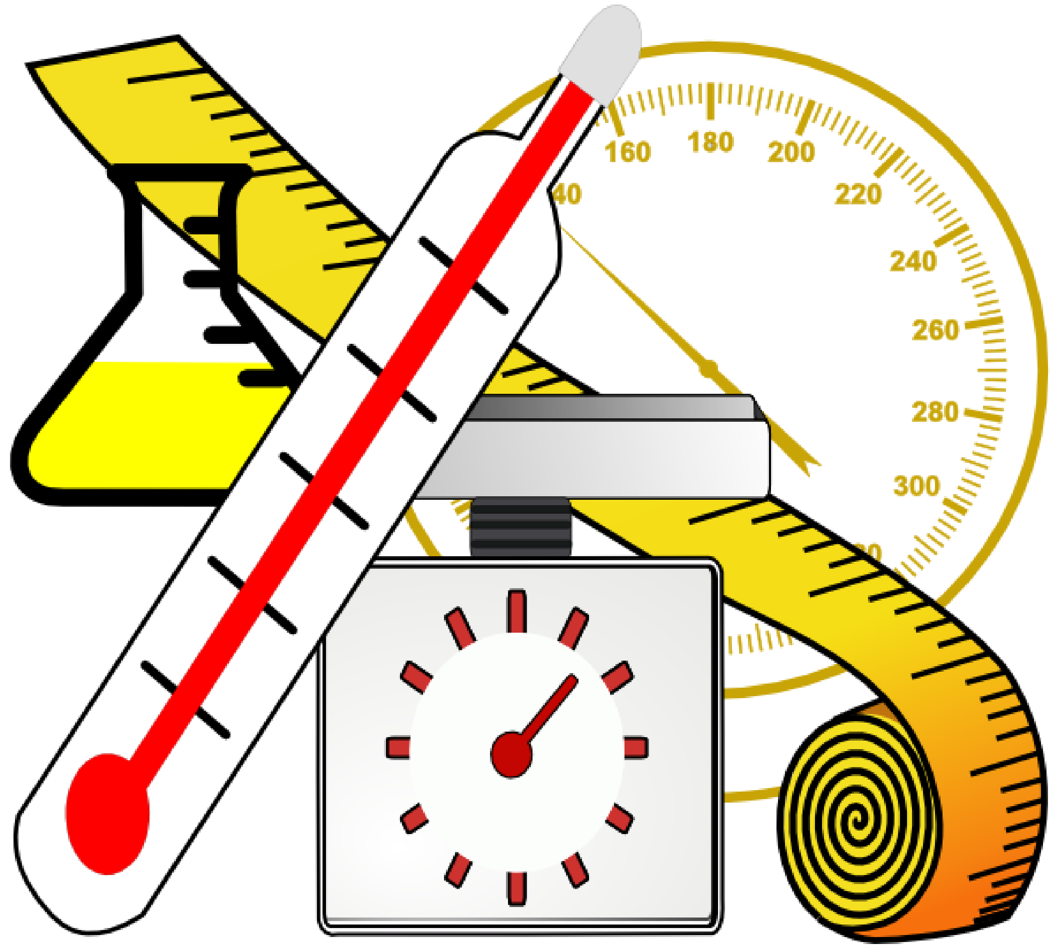
ABSTRACT

Humans have *internal models* of robots (like their physical capabilities), the world (like what will happen next), and their tasks (like a preferred goal). However, human internal models are not always perfect: for example, it is easy to underestimate a robot's inertia. Nevertheless, these models change and improve over time as humans gather more experience. Interestingly, robot actions *influence* what this experience is, and therefore influence how people's internal models change. In this work we take a step towards enabling robots to understand the influence they have, leverage it to better assist people, and help human models more quickly align with reality. Our key idea is to model the human's learning as a nonlinear dynamical system which evolves the human's internal model given new observations. We formulate a novel optimization problem to infer the human's learning dynamics from demonstrations that naturally exhibit human learning. We then formalize how robots can influence human learning by embedding the human's learning dynamics model into the robot planning problem. Although

Figure 1: Human teleoperates a new robot; they update their internal model by acting and observing outcomes. Planning with human learning dynamics, the robot influences the human's internal model to help them be a better teleoperator.

Proceedings of Human Robot Interaction (HRI '23). ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

(rejecting **H7**). What we found surprising was that even though participants were *quantitatively* performing better in the teaching condition, they did not *perceive* an improvement in performance (p = 0.689) nor in their understanding of the robot physics (p = 0.299). We hypothesize that this could be because participants only



https://www.clipartmax.com/max/m2i8K9i8Z5G6i8H7/

Subjective Metrics

- Subjective - based on personal interpretation
  - surveys, questionnaires, interviews
  - can be quantitative or qualitative
- Self-report is not always reliable, but sometimes it’s what you’ve got… and sometimes it’s the point of the research

Quantitative vs. Qualitative

- Quantitative - numerical
  - Categorical
  - Ordinal
  - Continuous
- Qualitative - descriptive

Quantitative Metrics

- Categorical - discrete categories, no inherent ordering
  - e.g., hair color, ice cream flavors
- Ordinal - discrete categories, ordered
  - e.g., low/medium/high economic status
- Continuous - measured along a continuum
  - e.g., age, income

Qualitative Metrics

- Non-numerical, descriptive format that enables deeper, more nuanced understanding
- **Methods**: interviews, free-response questions, observations, focus groups, etc.
- *Analysis techniques exist for qualitative data (see HCII classes for more!)*

Michael Chaplin via medium.com

Experiments: Learning Robot Objectives from Physical Human Robot Interactions

File Edit View Insert Format Tools Extensions Help

Menus 150% Normal text Arial 11

Experiments: Learning Robot Objective/Reward Function from Physical Human Robot Interactions

Hypotheses:

1. Learning the human's reward from pHRI will lead to objectively better human-robot interaction. In particular, the human will spend less time and effort interacting with the robot, and the robot's reward will better match the human's preference.
2. Human's will subjectively prefer interacting with a robot which learns their rewards through pHRI

Trajectory Features:

- Joint_velocity *[done]*
- Laptop_dist *[done]*
- Human_dist *[done]*
- Table_dist *[done]*
- EE_orientation (dot product EE-z with z-axis) *[done]*

Objective Metrics:

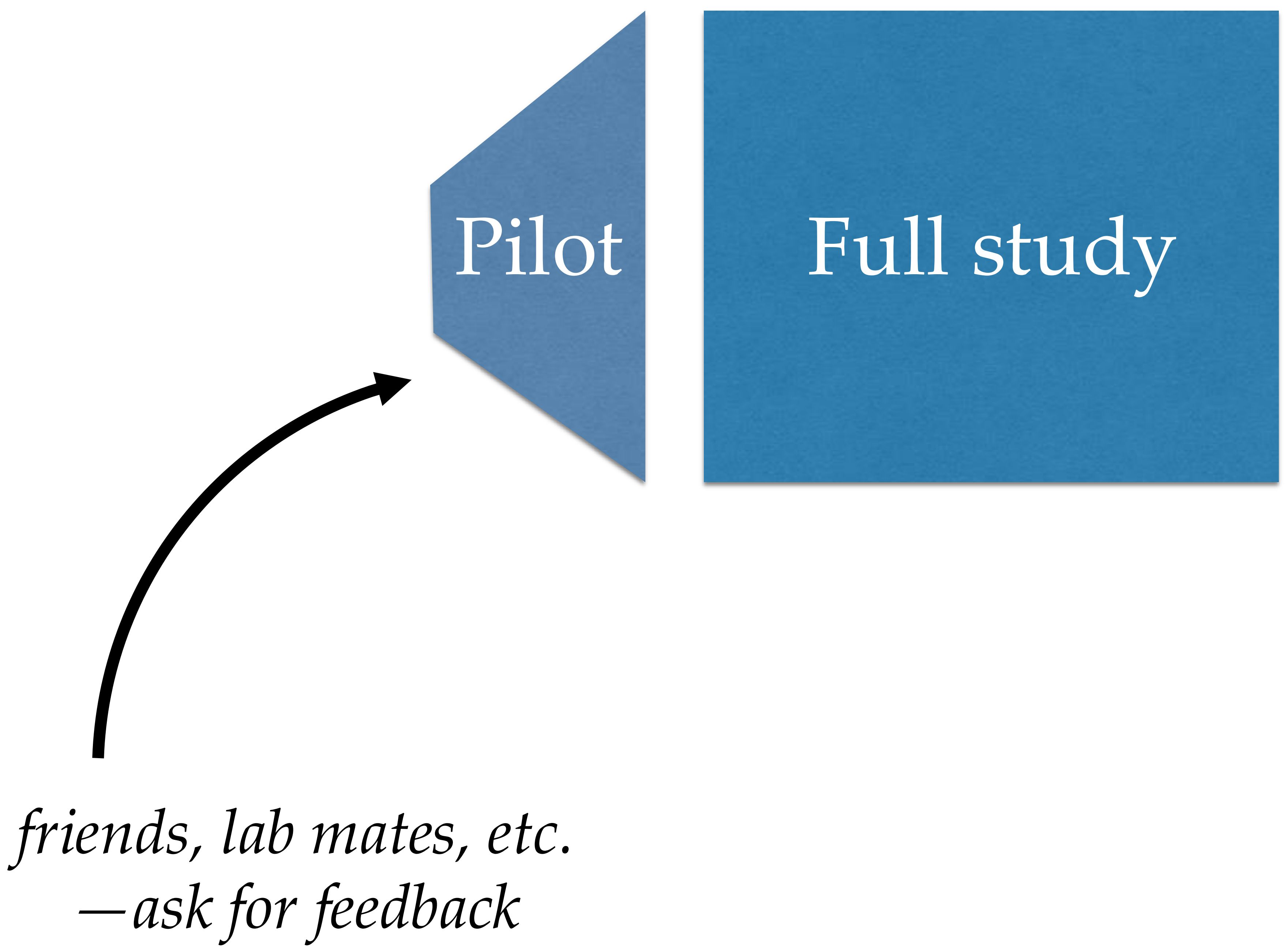
- Trajectory length (?)
- Trajectory execution time (sec) *[½ done]*
- Trajectory smoothness (?)

Design the Study: *Procedure*

- Specify the study's *structure*
- Select *metrics*
- Define *procedure*
  - what participants will actually do
- Define *population*

Defining the Task

- Pilot study - a small study that lets you refine your task, metrics, manipulations, etc.



Experimental Procedures

We use a *within-subjects* experimental design, where each participant is tested under both Method A and Method B. We *counterbalance* the methods by assigning participants to groups and to each group we present the methods in a different order. Aside from the order of the methods, the experimental procedure for Method A or Method B is as follows:

1. The subject is told that they are interacting with Method A (or Method B)
2. Familiarization
    a. The task feature is distance to person
    b. The human is instructed to stand in the first location
    c. The procedure is as follows:
        i. Demonstrate original trajectory (1x)
        ii. Demonstrate desired trajectory (1x)
        iii. Human performs task (2x)
        iv. Demonstrate original trajectory (1x)
        v. Demonstrate desired trajectory (1x)
        vi. Human performs task (2x)
3. Experimental Tasks
    a. Task 1: the feature is the orientation of the cup
        i. The procedure is as follows:
            1. Demonstrate original trajectory (1x)
            2. Demonstrate desired trajectory (1x)
            3. Human performs task (2x)
    b. The human is asked to stand in the second location
    c. Task 2: the feature is distance to table

Instructions to Participants

Thank you for participating, and welcome to the InterACT Lab. Before the experiment begins, please read and sign this consent form.

Today you will be performing household tasks with JACO, an assistive robot. Unfortunately, JACO does not currently perform these tasks in the way you'd like it to. For each task, JACO will start moving, and when you notice that it's not behaving correctly, you should physically intervene to correct it.

The goal today is to compare two methods that the robot can use to respond to your physical corrections. You will perform the same experimental procedure twice. First you will perform the experiment with Method A, and we will ask for your feedback. Then you will perform the same experiment with Method B, and we will again ask for your feedback. We won't tell you which method we prefer, we want to hear what you think.

As part of an experimental procedure, you will practice interacting with JACO in a familiarization task. Once you are used to interacting with the robot using the current method, you will assist the robot with three experimental tasks. We will ask for your feedback after all the experimental tasks are completed.

Familiarization

We will now begin the familiarization task, which allows you to become more comfortable with JACO and with Method A/B. You should view this task as practice, and not part of the

We will now repeat this procedure once more so you can better familiarize yourself with Method A/B.
[when you come back to this after the first run through the task and and they do the second method for the first time, ask them informally to comment on if they see any difference between them, and what is the difference]

Experimental Task 1

We will now begin Task 1 of the experiment.

In this task, the robot is taking a cup from a shelf and putting it down on the table. This cup is full of an imaginary liquid. **Your objective is to get the robot to keep the cup upright so that it does not spill the liquid..** Don't worry about correcting the distance to you this time around, just about keeping the cup upright. As before, try to not affect the speed during interaction.

Here is how the robot would normally perform the task.

Here is ***roughly*** how the you would like the robot to perform the task. ***The specifics of the trajectory are not important, what is important is that the robot keeps the cup upright.***

Now you will be given two experimental rounds. During each round, you should physically intervene to make the robot behave ***roughly*** like the desired motion that you saw before. Again, remember the two rules: 1) the robot should do the task correctly but also as independently as possible; 2) don't alter the speed.

Ok, please step away from the table for the reset. Ok, now you can resume your position, and be ready to correct the robot. We will start in 3-2-1. …

For the sake of the robustness in our data collection, we will do this once more. We will reset the robot and explicitly erase that interaction, so **you will start from scratch, as if the previous run never happened.**

Design the Study: *Population*

- Specify the study's *structure*
- Select *metrics*
- Define *procedure*
- Define *population*
  - who should be involved as participants?

Defining Population

- *How many?*
  - N = the size of your population
  - depends on *effect size*
  - generally, more people → higher likelihood of finding an effect, if one exists

Richard foster on Flickr

- *Who?*
  - general population vs. special group?
  - balance confounds of age, gender, technology experience, education, etc.

Research Ethics

- We protect participants from:
  - physical, mental, emotional harm
  - violations of privacy and confidentiality
  - feeling forced to start or continue with a study

https://gigaom.com/

Internal Review Board

- IRBs "protect the rights and welfare of humans participating in research"
  - https://www.cmu.edu/research-compliance/human-subjects-research/index.html

- Getting IRB approval is *critical for sponsored research*, but is good practice always

Institutional Review Board (IRB)

SPARCS

http://www.cmu.edu/research-office/sparcs/

- Human subjects study process at CMU:
  - Get certified by taking an online course (and complete 3 year renewals)
  - Submit a protocol for review
  - Renew protocol every 1-3 years based on type of study
  - Close protocol when study is done

Informed Consent

- Informed consent - giving voluntary permission with full knowledge of possible risks and benefits
    - informs participants about the task, risk, and benefits
    - acquires written confirmation of their voluntary, knowledgable participation
- A *process*, not a piece of paper!

University of California Berkeley

Consent Form for Participation in Research

Study Title: Learning Human Preferences from Physical Human-Robot Interaction

Anca Dragan
Assistant Professor
Electrical Engineering and Computer Science
776 Sutardja Dai Hall
University of California Berkeley
Berkeley, CA 94720-1758
anca@berkeley.edu

Other Investigator(s):
Andrea Bajcsy
Graduate Student

Participant's Name: ______________________________
Participant's ID Number: ________________

You may be eligible to take part in a research study. This form gives you important information about the study. It describes the purpose of the research, the risks and possible benefits of participating in the study.

Purpose of this Study
Lightweight, personal robots are increasingly being developed to work with humans in the home, on wheelchairs, and in social settings. A human's preference when closely interacting with a robot can vary across users, environment, and tasks, and generally cannot be manually encoded into a robot---instead, a robot should learn these preferences in real time. Physical interaction, which communicates intent through a sense of touch, provides a natural, human-like means to convey preferences between human and robot. This research is designed to advance the understanding of human-robot interactions, which have many applications in fields such as assistive robotics, teleoperation, and other fields that have significant human-robot interactions. In this study, we will test algorithms for learning human preferences for robotic motion from physical human-robot interactions.

Procedures / What will happen to me in this study?
For our experiments, we utilize a Kinova JACO^2 seven degree-of-freedom robotic arm, which is a lightweight assistive robot designed for safe grasping and manipulation in human-robot environments.

How to Conduct HRI User Studies

1. define the research question and hypothesis
2. design a study to address that question
3. **execute the study**
4. analyze data from the study
5. draw conclusions from the analysis

Finding Participants

- Where do I get participants?
  - Word of mouth ← *for this class*
  - Flyers, emails, Facebook messages
  - Online study sites (Mechanical Turk, Prolific)
  - CMU's Center for Behavioral and Decision Research Participant Pool (https://www.cbdr.cmu.edu/)

mturk.com

Final Project Specifics

- All final projects in this class are **pilot studies**
  - No IRB
  - No external recruitment (word of mouth only)
  - No paying participants
- If the pilot proves viable, we can discuss getting IRB approval to conduct a full study

Study Tips

- Record everything on video (so you know what happened later)
- Print out study instructions and read them each time (to avoid biasing your participants)
- Pilot with a variety of people (your labmates might not be representative of the general population)

How to Conduct HRI User Studies

1. define the research question and hypothesis
2. design a study to address that question
3. execute the study
4. **analyze data from the study**
5. draw conclusions from the analysis

Analyze the Data

- "Has the DV changed as a result of manipulating the IV?"
- Descriptive statistics - summarize the DV
- Inferential statistics - make conclusions beyond the current data

maxpixel.net/

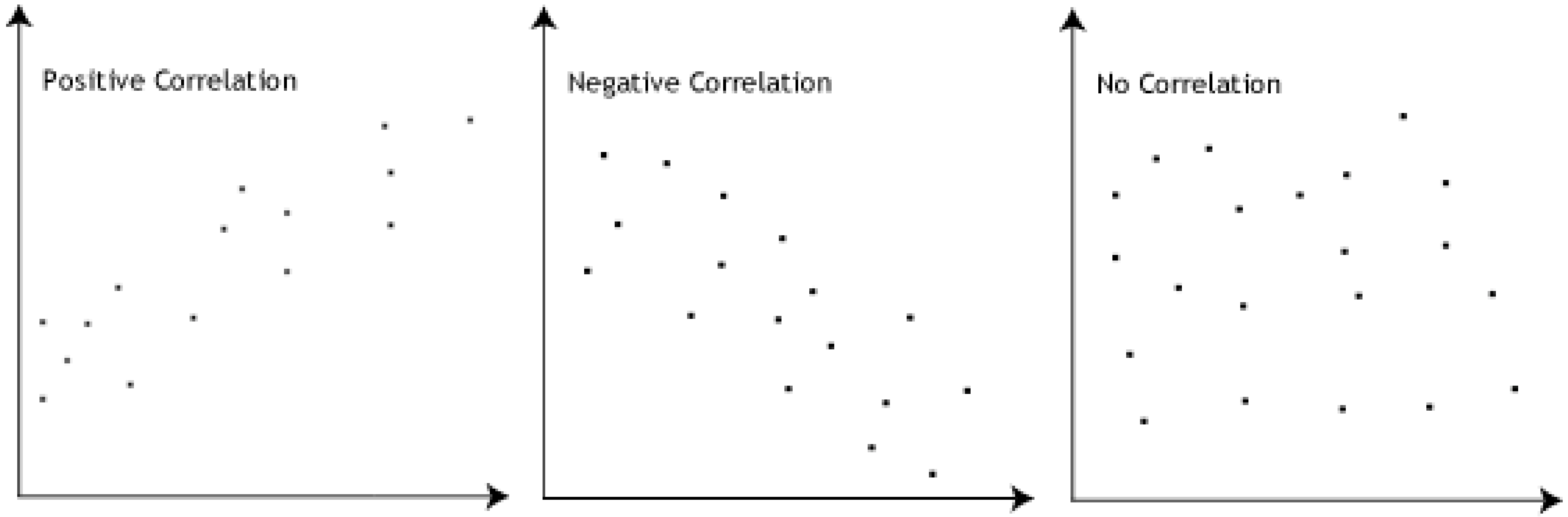
Questions That Statistical Tests Can Answer

- “What’s the probability that two variables are correlated?”
- “What’s the probability that two populations are actually different from each other on a certain measure?”
- “What’s the probability that a population is different than expected?”

Pearson's Correlation

"What's the probability that two variables are correlated?"

- Finds whether two variables are likely to have a ***linear*** association
- Use with: two continuous variables

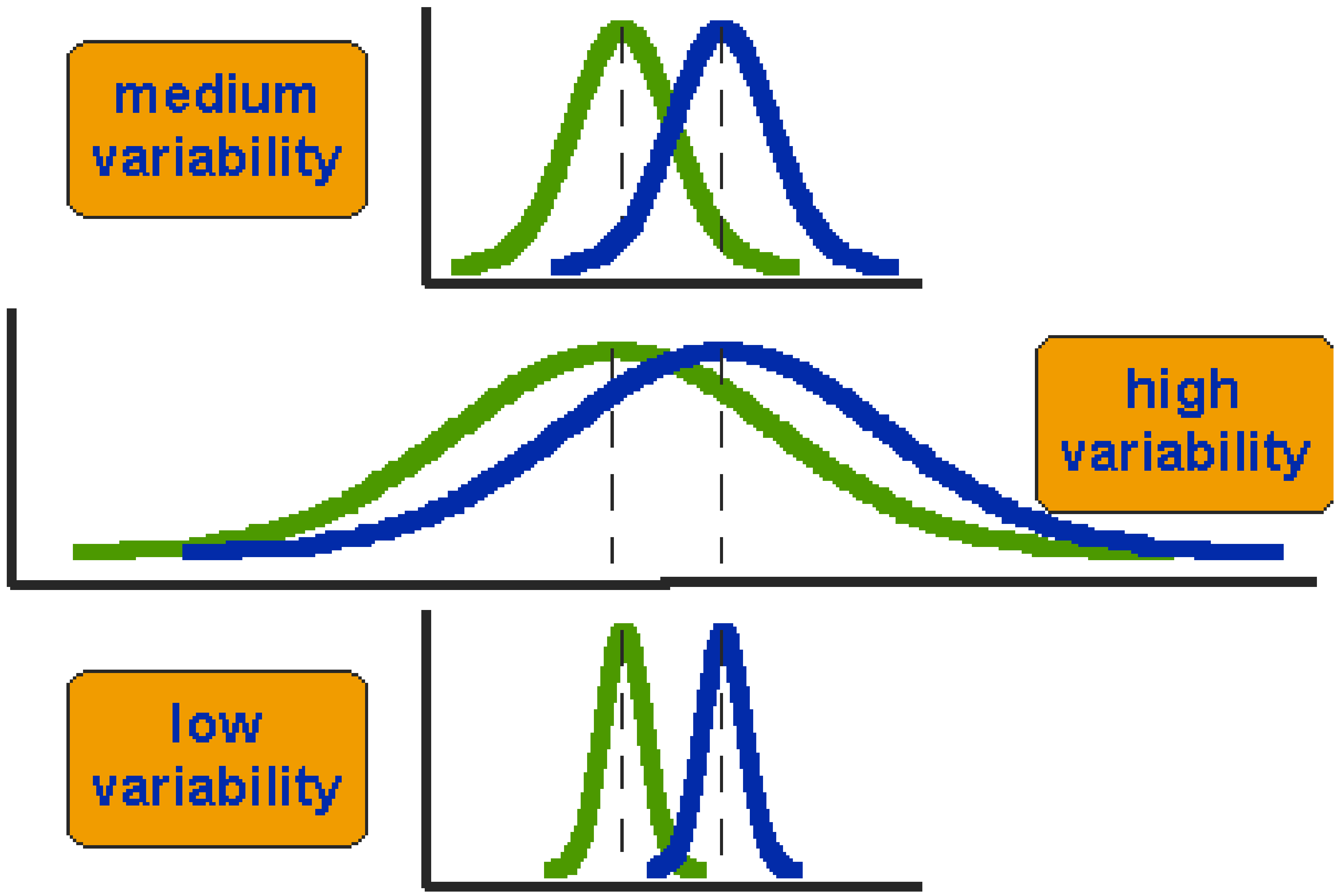


https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

Student's T-Test

"What's the probability that two populations are actually different from each other?"

- Finds whether data in 2 groups are likely to come from the same dataset
- Use with: 1 IV with 2 levels and a continuous DV that is normally distributed

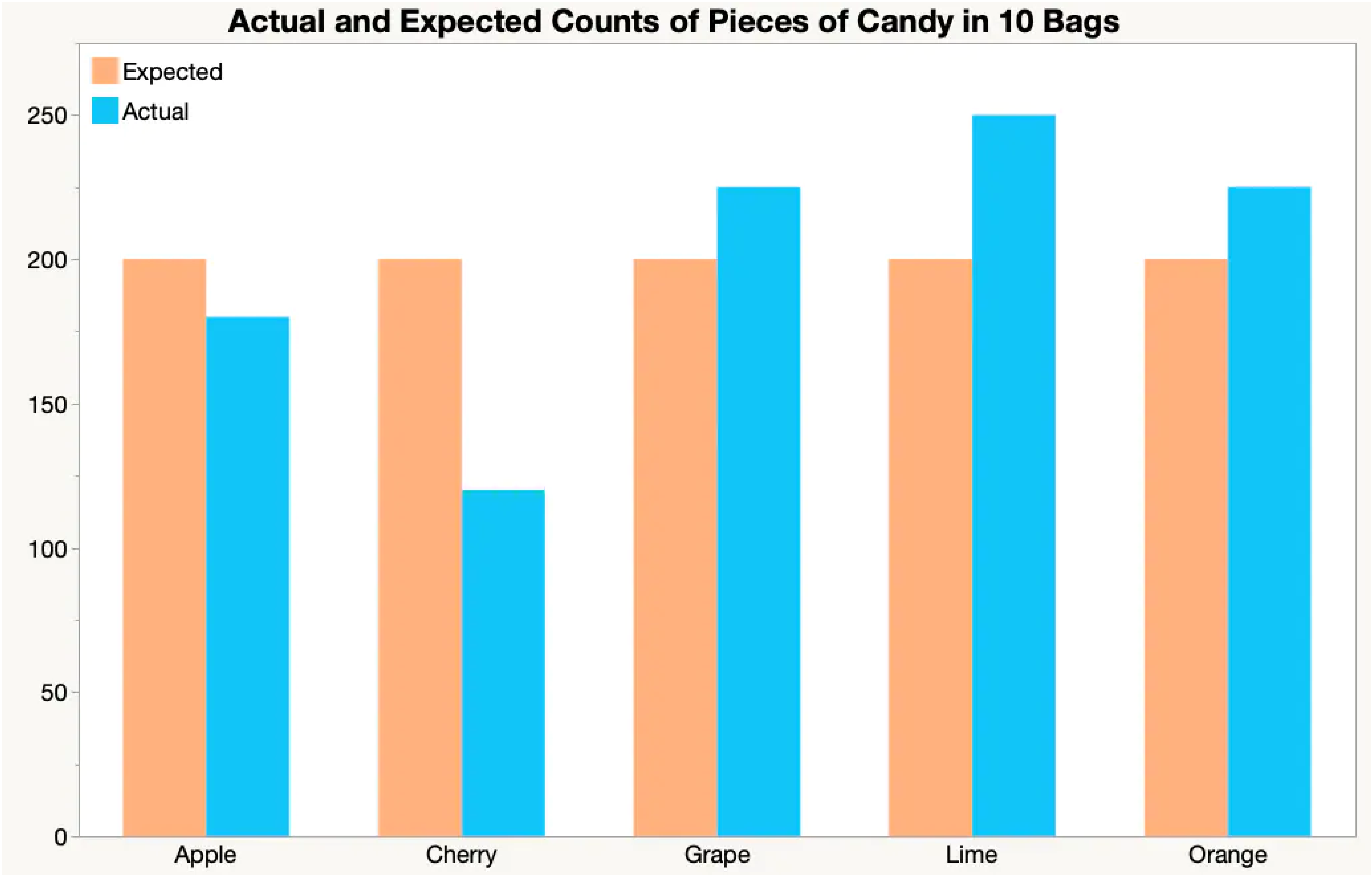


https://socialresearchmethods.net/kb/stat_t.php

Chi Squared Goodness of Fit

"What's the probability that a population is different than expected?"

- Finds whether observed data are consistent with some hypothesized distribution
  - Requires a hypothesized distribution over the categories
- Use with: 1 categorical/ordinal variable with 5+ observations in each category



jmp.com

Example Data Analysis

No Joint Attention	Joint Attention
11.92166184	16.32041702
11.32053311	13.92180794
16.75393131	13.55823062
9.596080771	11.23076697
14.20149488	17.16287452
10.65604112	12.11149578
13.97438489	12.57160394
14.42557754	15.99423833
8.778358321	9.312165405
10.44668764	10.0142651
5.1320481	7.497488059
8.841989234	14.20590455
8.992115131	13.15703192
10.94719434	12.97455817
9.84480963	17.88012696

Question: Will joint attention during a handover improve handover efficiency?

Hypothesis: Joint attention from a robot will improve handover efficiency as measured by speed of successful handovers in seconds.

Example Data Analysis

No Joint Attention	Joint Attention
11.92166184	16.32041702
11.32053311	13.92180794
16.75393131	13.55823062
9.596080771	11.23076697
14.20149488	17.16287452
10.65604112	12.11149578
13.97438489	12.57160394
14.42557754	15.99423833
8.778358321	9.312165405
10.44668764	10.0142651
5.1320481	7.497488059
8.841989234	14.20590455
8.992115131	13.15703192
10.94719434	12.97455817
9.84480963	17.88012696

No J-Attention:
Mean: 11.056
SD: 2.89
N = 15

J-Attention:
Mean: 13.194
SD: 2.93
N = 15

Example Data Analysis

- *Hypothesis*:
  - Joint attention from a robot will improve handover efficiency as measured by speed of successful handovers in seconds.
- *Descriptive Statistics:*
  - No joint attention: M = 13.2 s, SD = 2.9
  - Joint attention: M = 11.1 s, SD = 2.9
- *Inferential Statistics:*
  - Independent two-tailed t-test: $p = 0.0537$ → *If $P > 0.05$ then "no significant difference between groups"*

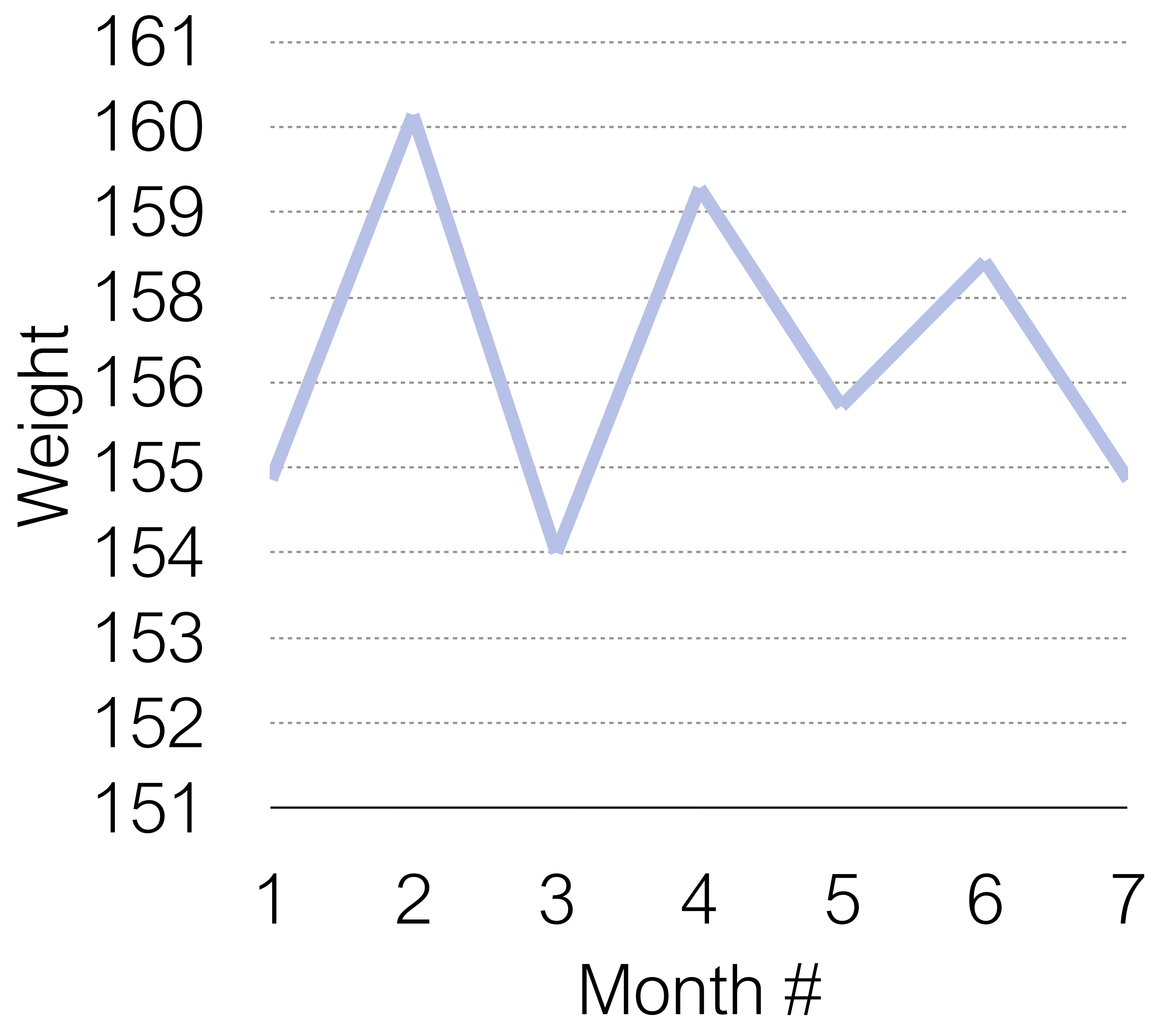
How to Conduct HRI User Studies

1. define the research question and hypothesis
2. design a study to address that question
3. execute the study
4. analyze data from the study
5. **draw conclusions from the analysis**

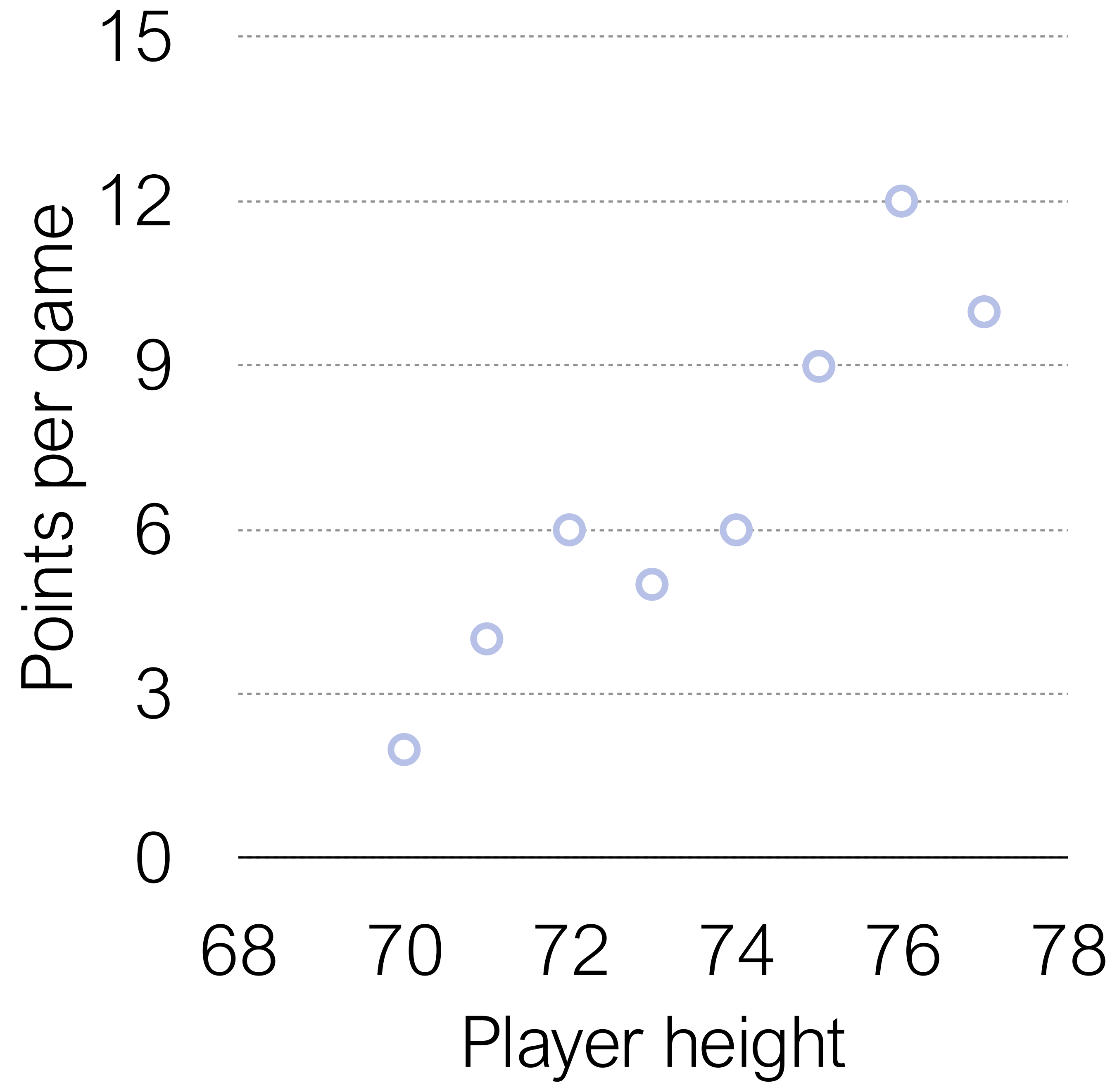
Visualizing Data

What kind of graph would you choose? (Bar graph, scatterplot, line graph)

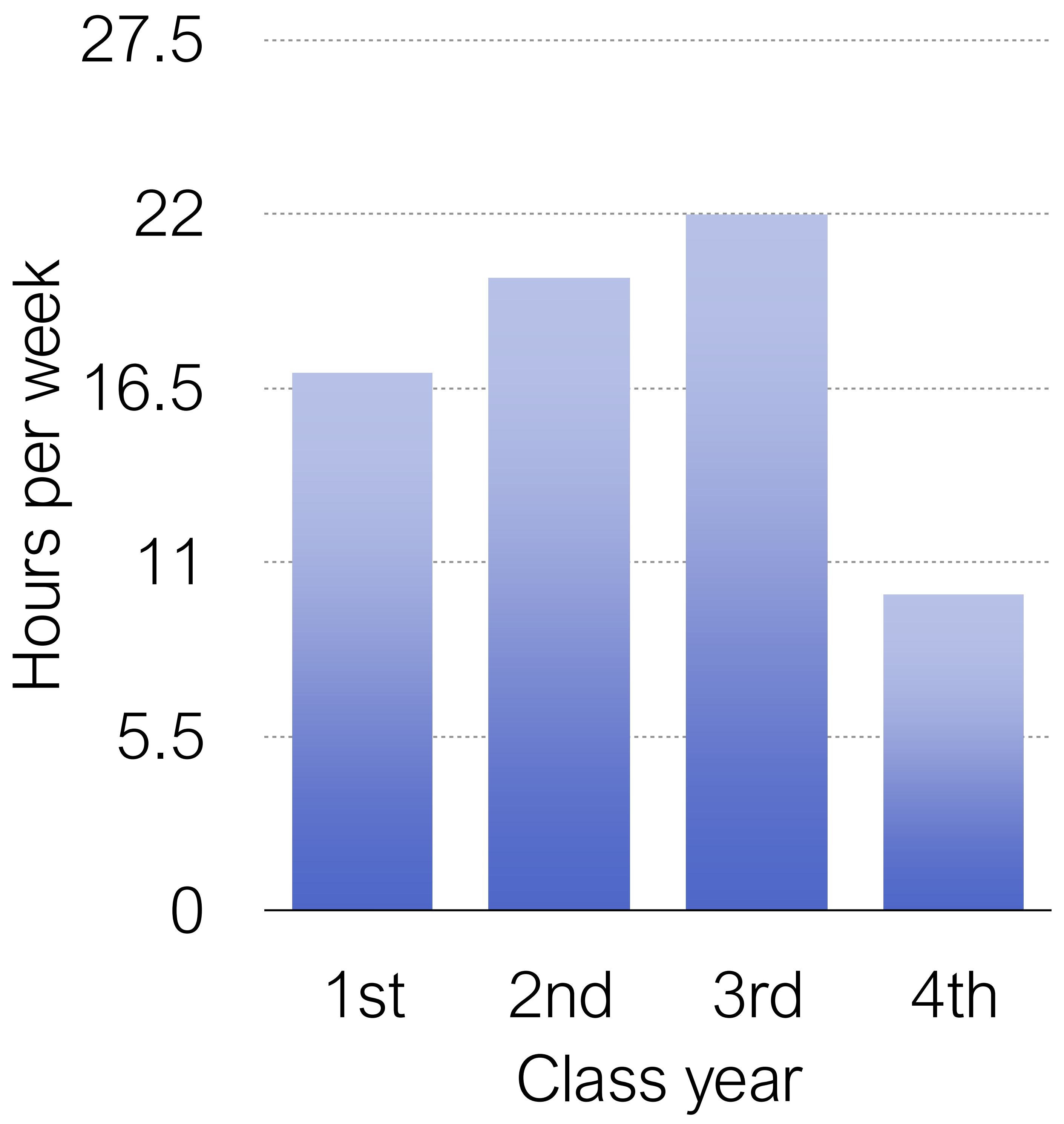
Change in weight over time



Points scored by player height



Hours in library by class year



Reporting Results

Each statistical analysis has a standard reporting format, e.g.

"To test Hypothesis H1, we ran an independent samples t-test. Consistent with our hypothesis, participants rated their trust in robots higher when the robot was running our adaptive algorithm ($M = 5.64$, $SD = 1.47$) compared to the baseline algorithm ($M = 4.86$, $SD = 1.62$), $t(102) = 2.54$, $p = .013$, $d = 0.50$."

descriptives **test statistic** **p value** **effect size**

Example from Hoffman & Zhao, THRI 2020

Resources for Statistical Tests

Nayak BK, Hazra A. How to choose the right statistical test? Indian Journal of Ophthalmology. 2011;59(2):85-86.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116565/

UCLA Institute for Digital Research and Education. “What statistical analysis should I use?”
https://stats.idre.ucla.edu/other/mult-pkg/whatstat/

Laerd Statistics.
https://statistics.laerd.com/

Final Thoughts

- Conducting studies involves six steps, and each is important:
  1. defining the research question and hypothesis
  2. designing a study to address that question
  3. executing the study
  4. analyzing data from the study
  5. drawing conclusions from the analysis
- Plan everything in the beginning to save headaches later!

Slides adapted from Henny Admoni

16-867

Experimental Design

Instructor: Andrea Bajcsy

Carnegie Mellon University

intent ROBOTICS LAB