❀ Honorable Mention at 25th International Conference on Intelligent User Interfaces (IUI '20)

How Do Visual Explanations Foster End Users' Appropriate Trust In Machine Learning?

Fumeng Yang[1], Zhuanyi (Yi) Huang[2], Jean Scholtz[3], and Dustin L. Arendt[2]

[1] Fumeng Yang is with Brown University. She conducted this research as a Ph.D. Intern at Pacific Northwest National Laboratory.
[2] Zhuanyi Huang and Dustin L. Arendt are with Pacific Northwest National Laboratory.
[3] Jean Scholtz retired from Pacific Northwest National Laboratory in September 2018.

Brown Visual Computing Seminar | May 11, 2020

Highlights

- **Visual explanations improve** end users' **trust** in an automated system.
  - Such **trust** must be **appropriate**.
    - The **design** of visual explanations affects users' **appropriate trust.**

> **“Human-computer Trust** is defined in this study to be, the extent to which **a user is confident in**, and **willing to act on** the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid. ”

——— Madsen and Gregor

Madsen, M., & Gregor, S. (2000, December). Measuring human-computer trust. In *11th australasian conference on information systems* (Vol. 53, pp. 6-8).

Appropriate Trust is the alignment between the perceived and actual performance of the system.

McBride, M., & Morgan, S. (2010). Trust calibration for automated decision aids. Institute for Homeland Security Solutions.[Online]. Available: https://www. ihssnc. org/portals/0/Documents/VIMSDocuments/McBride_Research_Brief. pdf.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Human factors, 48(4), 656-665.

Marsh, S., & Dibben, M. R. (2005, May). Trust, untrust, distrust and mistrust–an exploration of the dark (er) side. In International conference on trust management (pp. 17-33). Springer, Berlin, Heidelberg.

de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014, June). A design methodology for trust cue calibration in cognitive agents. In International conference on virtual, augmented and mixed reality (pp. 251-262). Springer, Cham.

Appropriate Trust

		System Recommendation	
		Correct	*Incorrect*
User Decision	*Follow*	Appropriate trust	Overtrust
	Not follow	Undertrust	Appropriate trust

Marsh, S., & Dibben, M. R. (2005, May). Trust, untrust, distrust and mistrust–an exploration of the dark (er) side. In International conference on trust management (pp. 17-33). Springer, Berlin, Heidelberg.

Example: My trust in an iRobot

My **confidence** in that it could clean the floor, my **willingness** to get it do the work;
overtrust is when I think it would avoid hitting the wall, but it does not;
undertrust is when I think it would hit the wall, but it makes a turn.

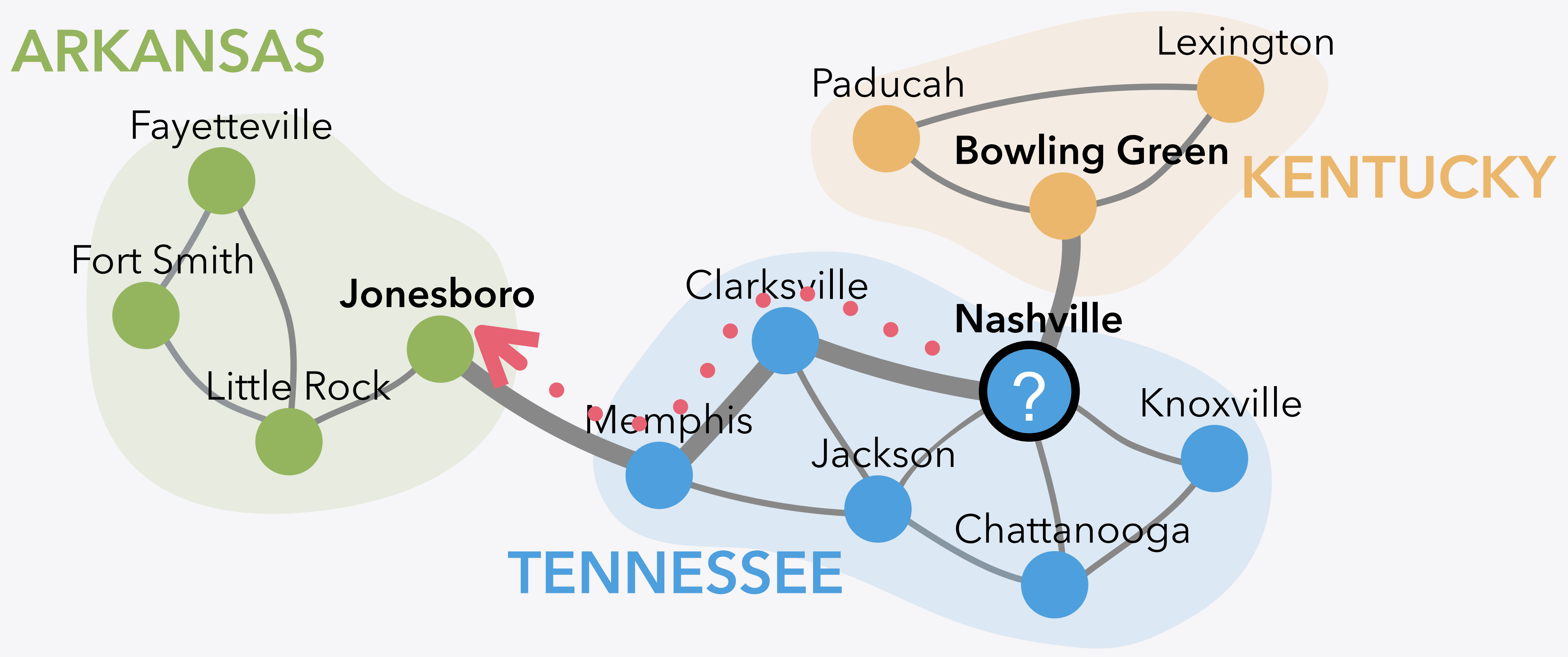
Goals

- The relationship between users' trust in a system and visual explanations;
- The effects of different visualization designs on users' trust in machine learning;
- An understanding of users' appropriate trust for proper usage of an automated system.

Experiment

- **Materials** Example-based explanation
- **Experimental variables** Instance representation, Spatial layout
- **Measures** Appropriate trust metrics, usability, individual differences
- **Task** Assistant botanists and classify leaves aided by classifiers with or without visual explanations

Example-based Explanation



"Escape Routes"

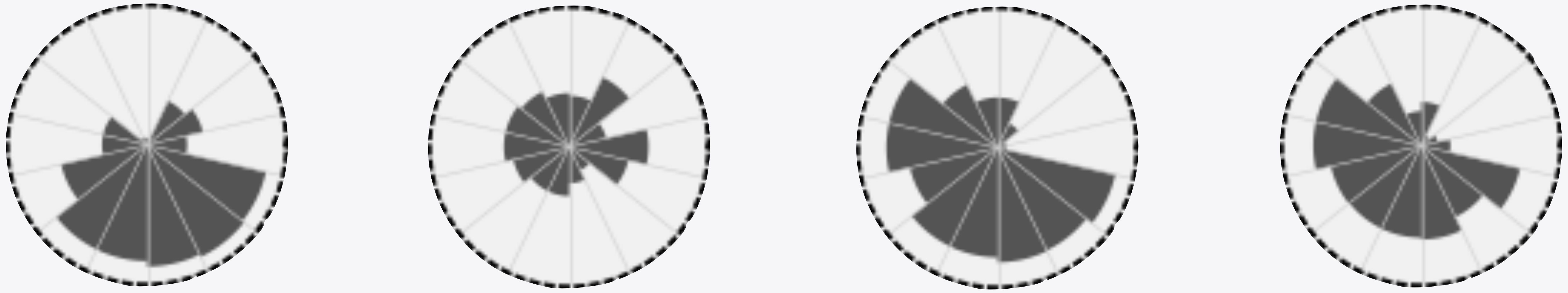
The shortest paths to travel to another state (class)

- k-nearest neighbors graph
  - Internal representation of the training set
  - Minkowski distance
- A shortest path tree rooted at the input node
- Prune until only leaves may have a different class from the input node

Instance Representation

To represent each instance in a dataset

Images

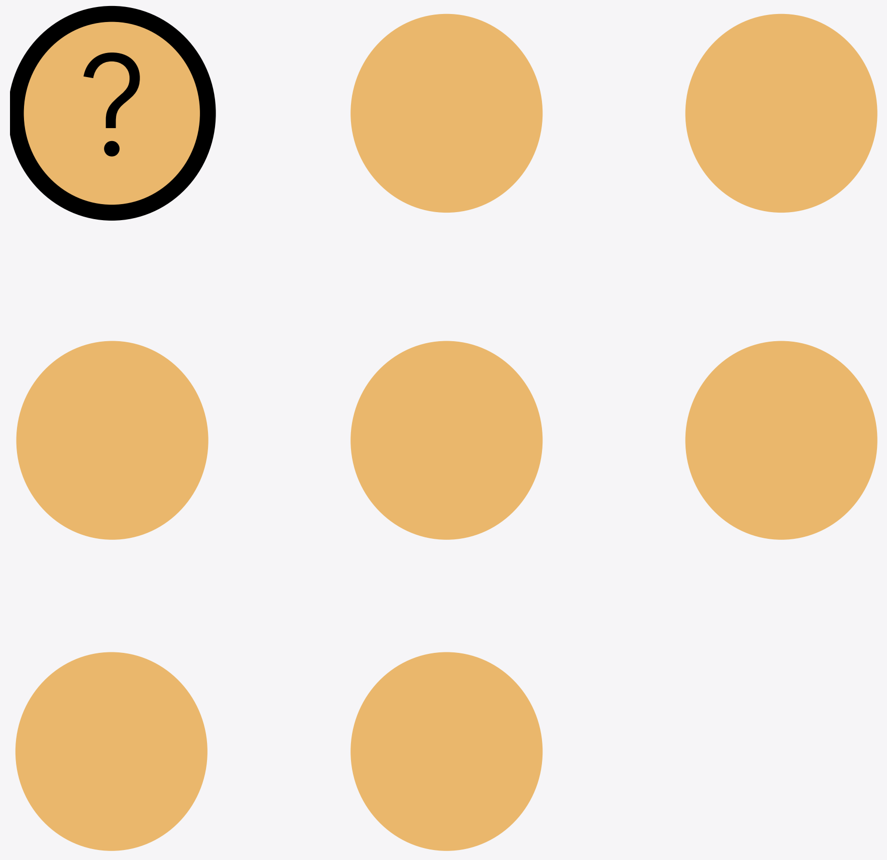


Rose charts (Roses) for feature vector

Spatial Layout

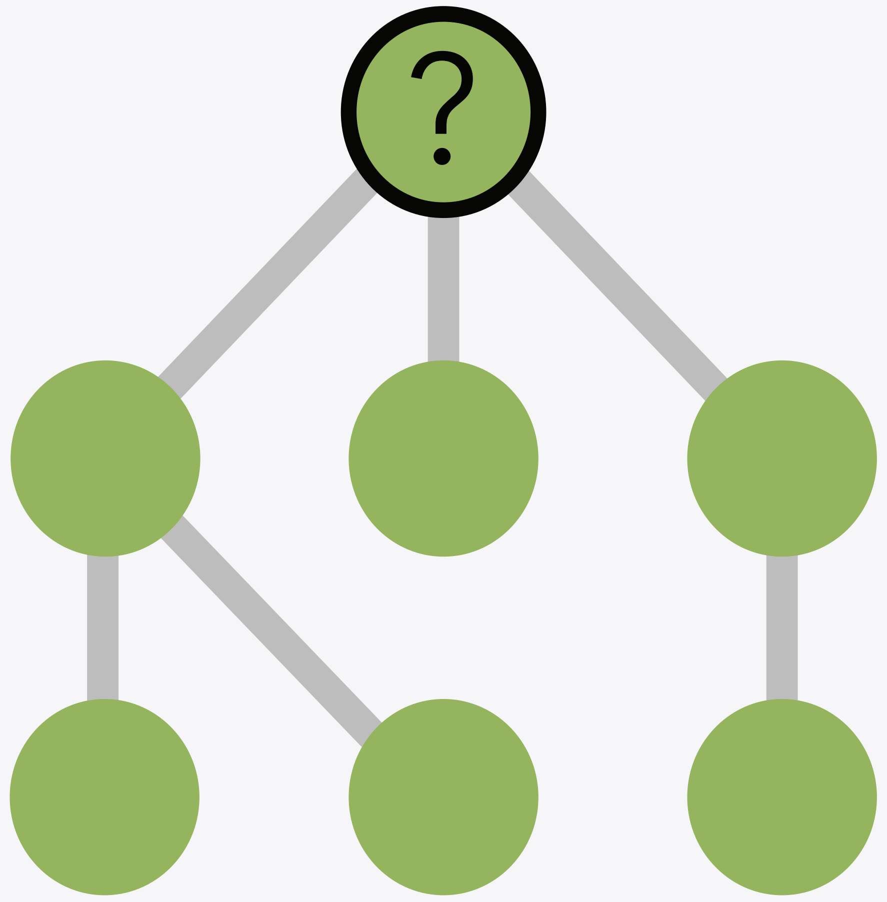
To arrange instances and illustrate the relationship between them

Grid



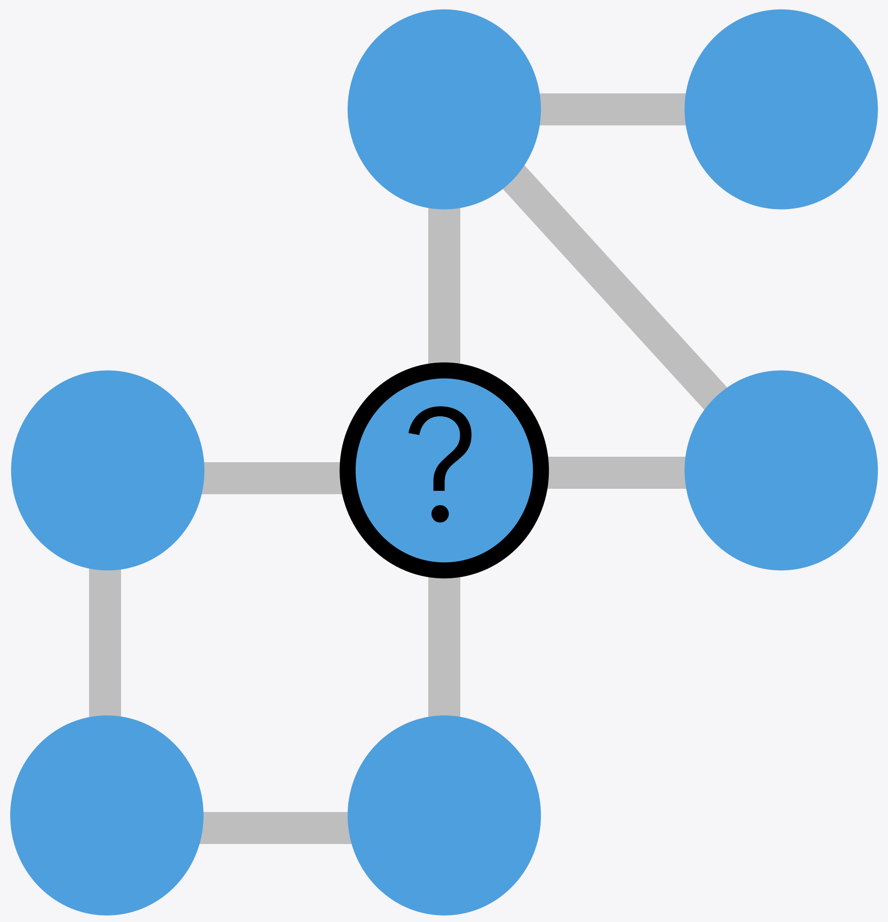
Sort instances within a column by their weighted geodesic distance to the input node

Tree



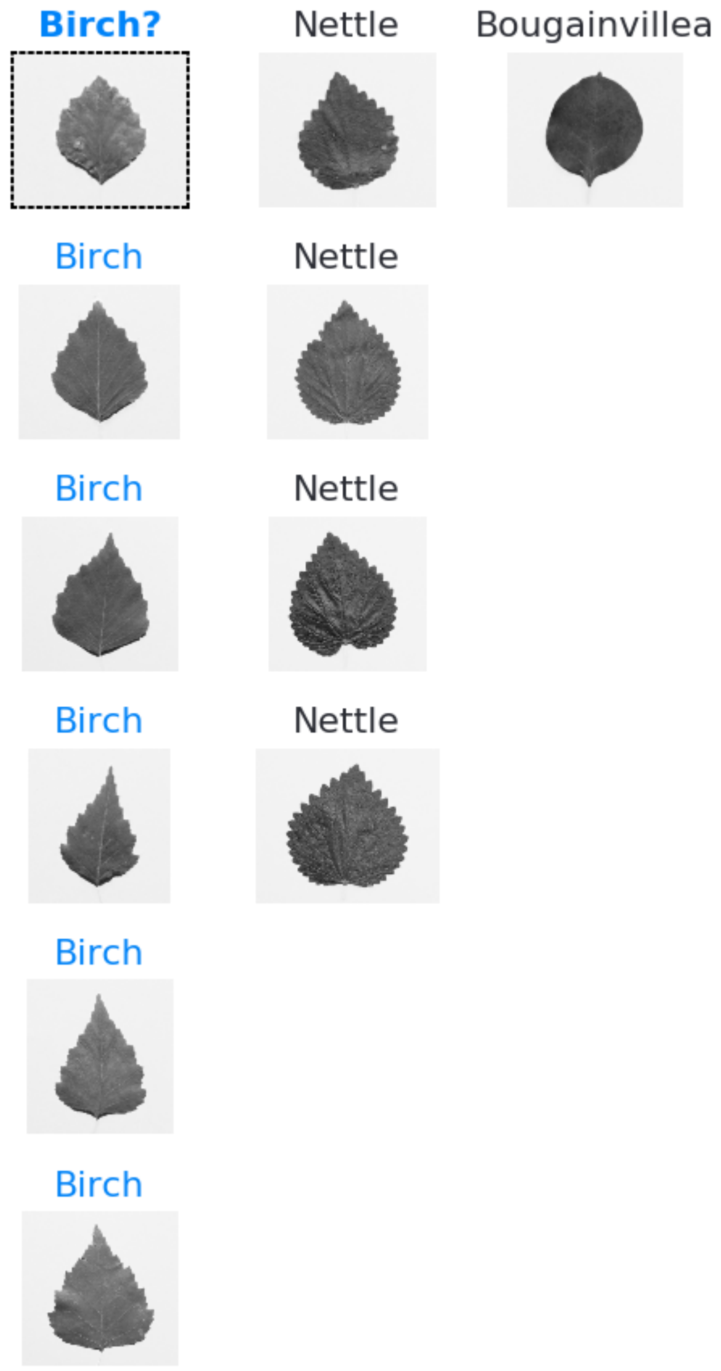
Use a layered graph layout of the pruned shortest path tree

Graph

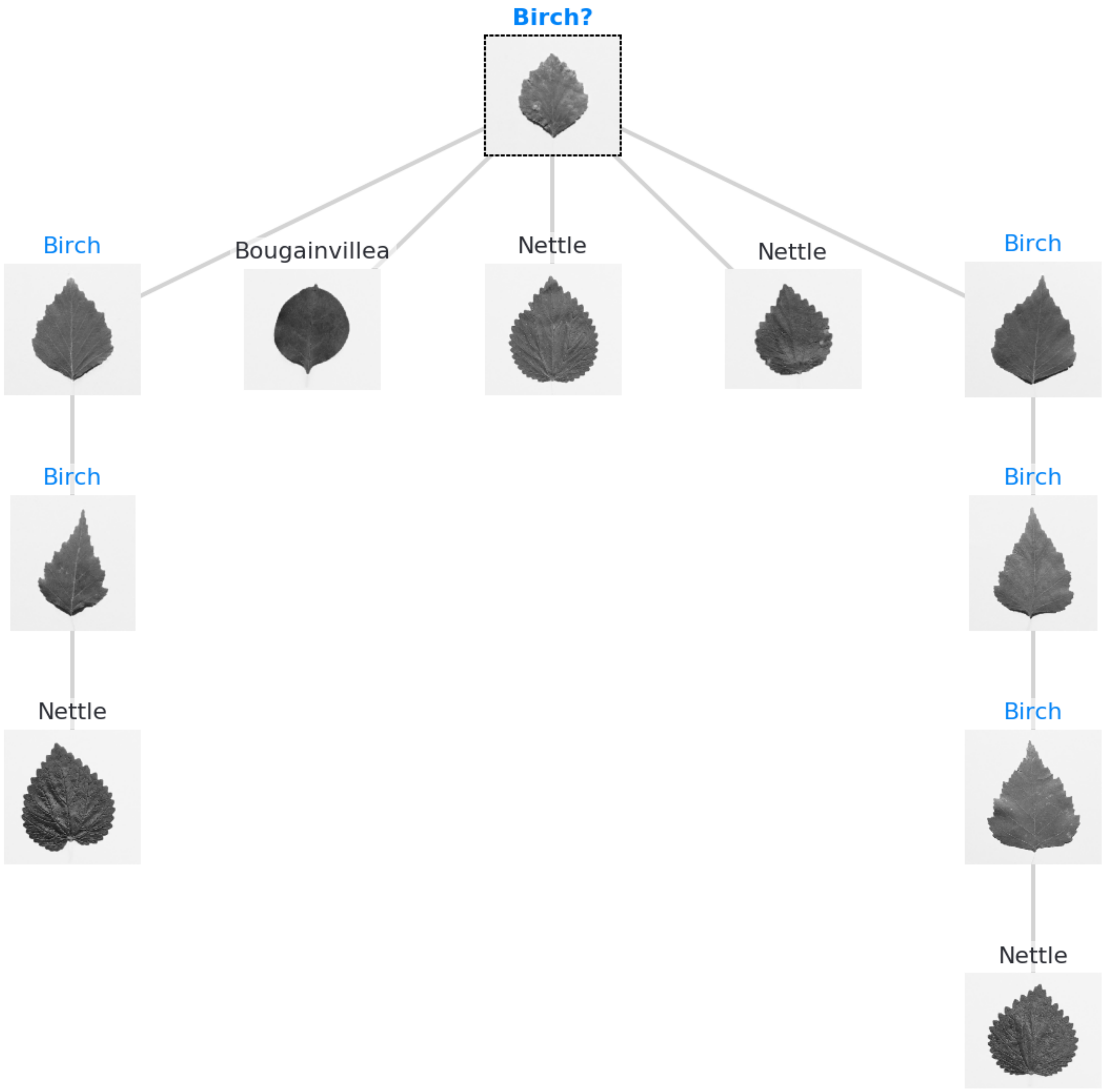


Use a force-directed layout algorithm to arrange instances based on their connections

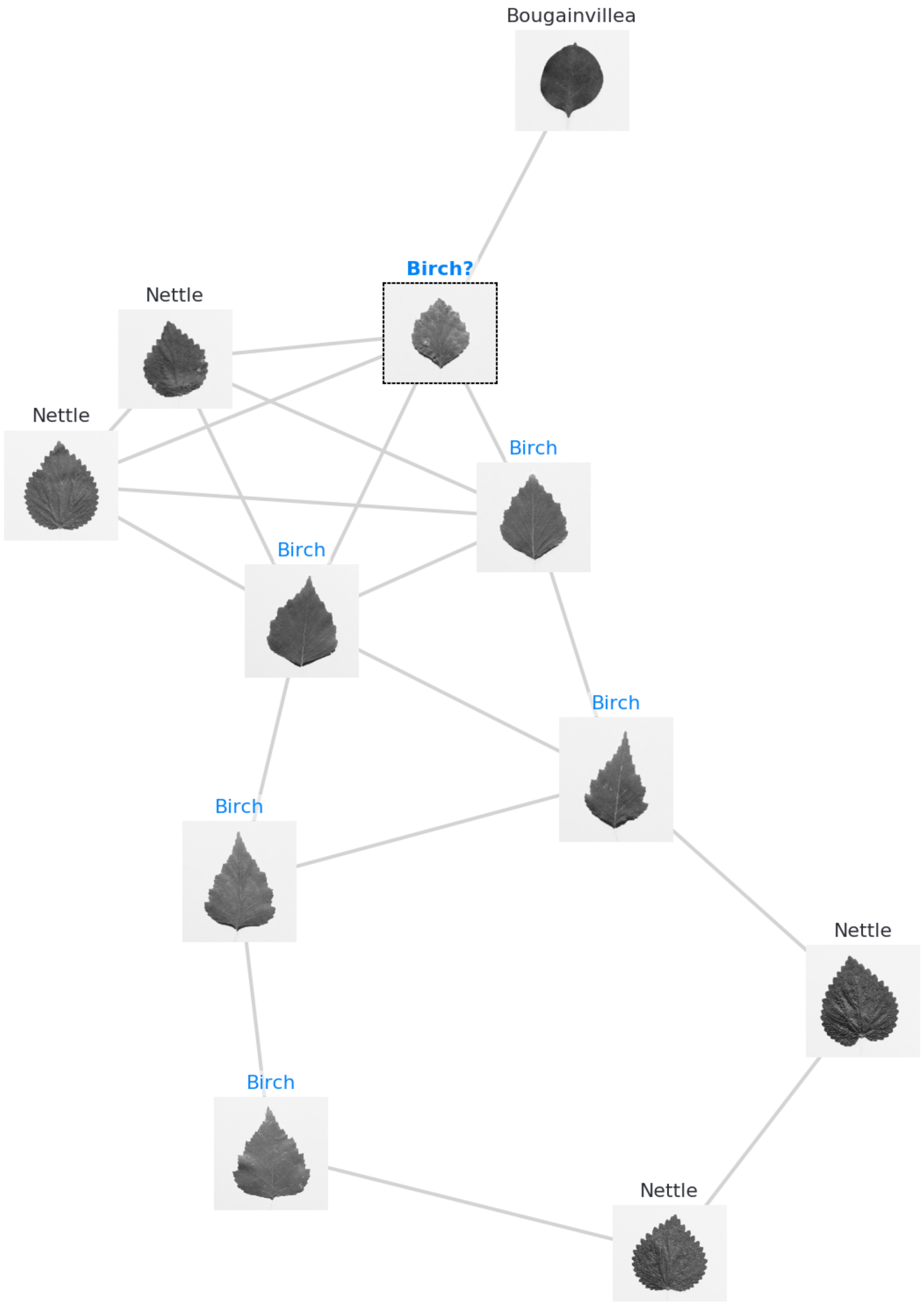
Examples



Grid

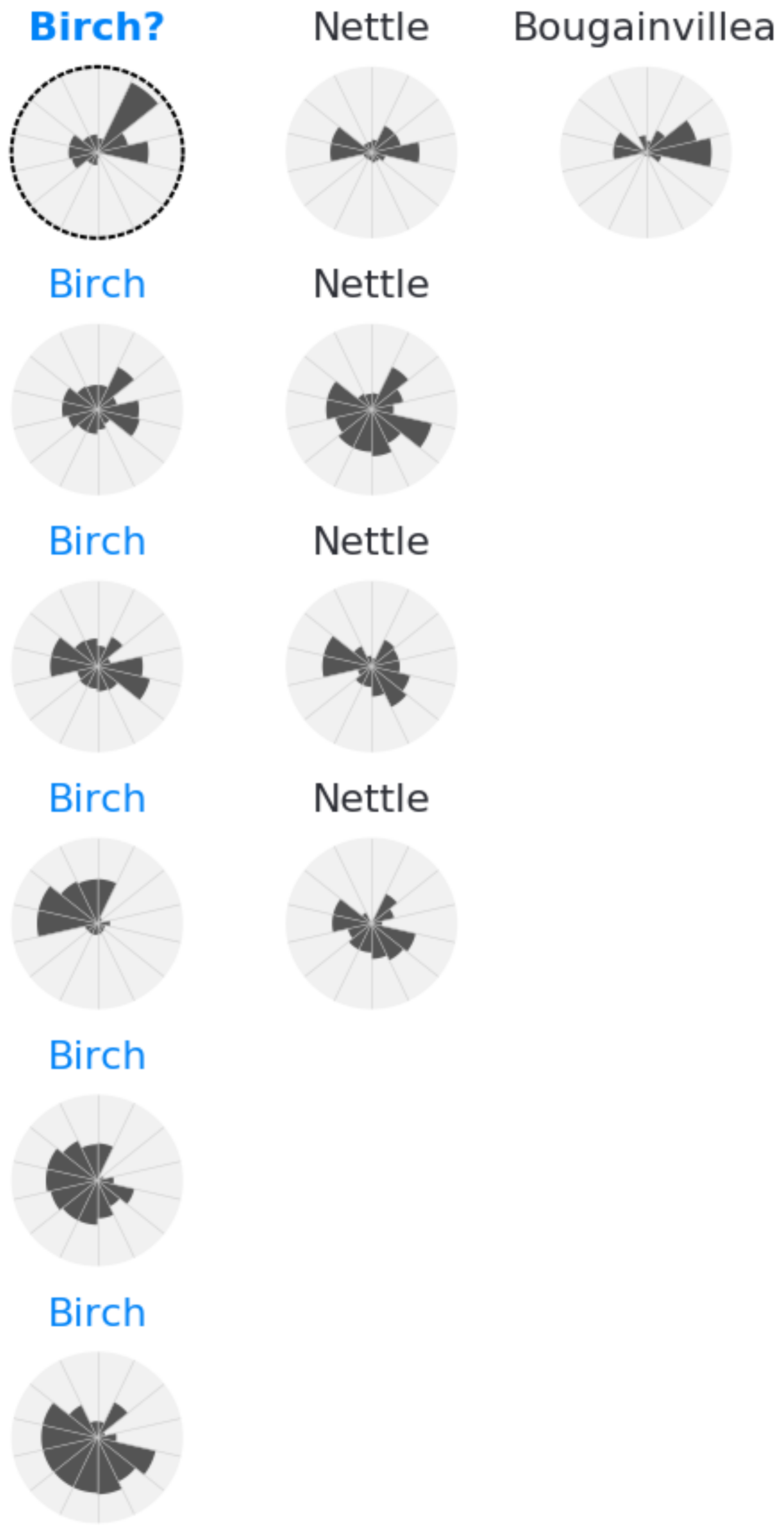


Tree

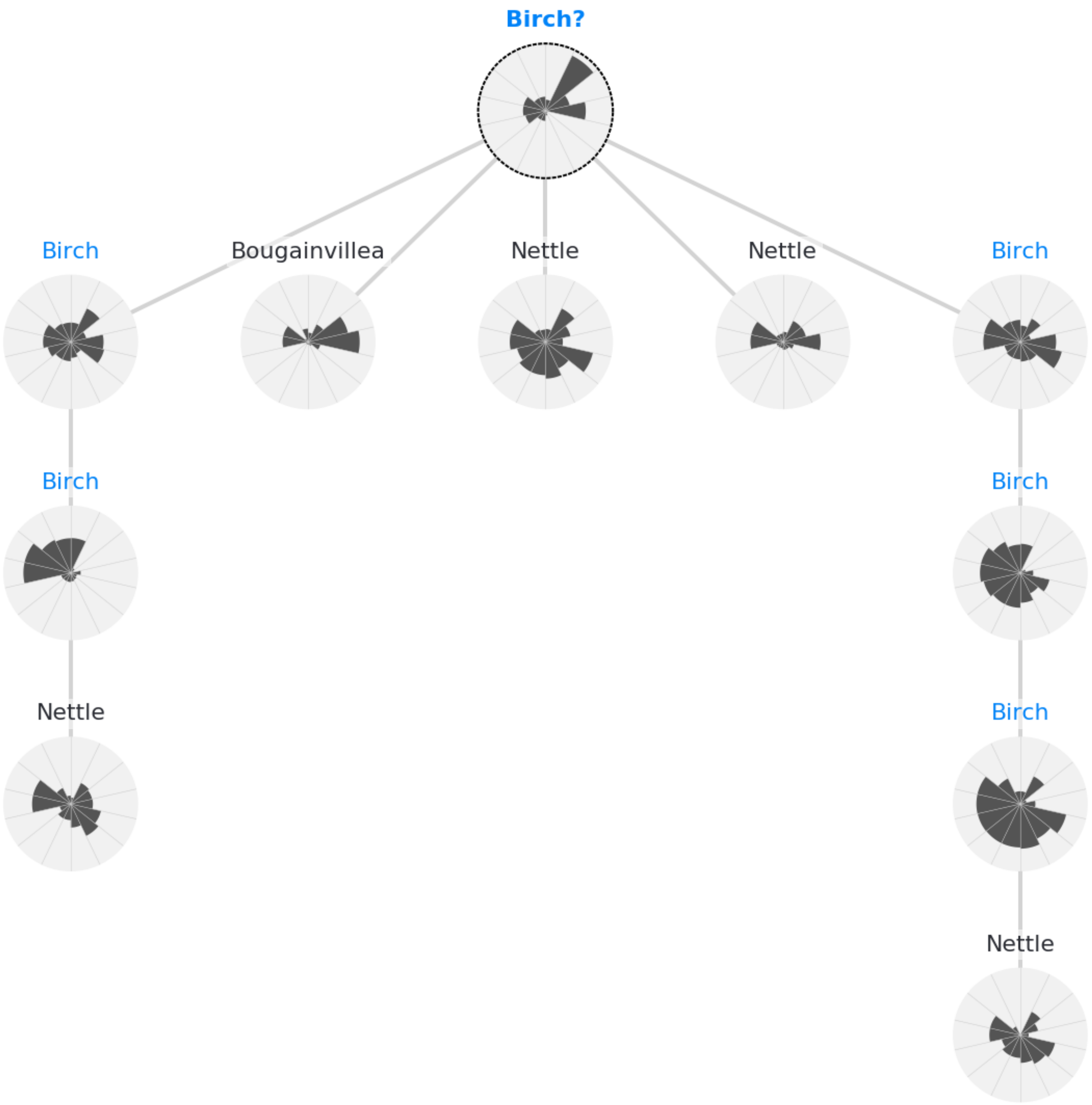


Graph

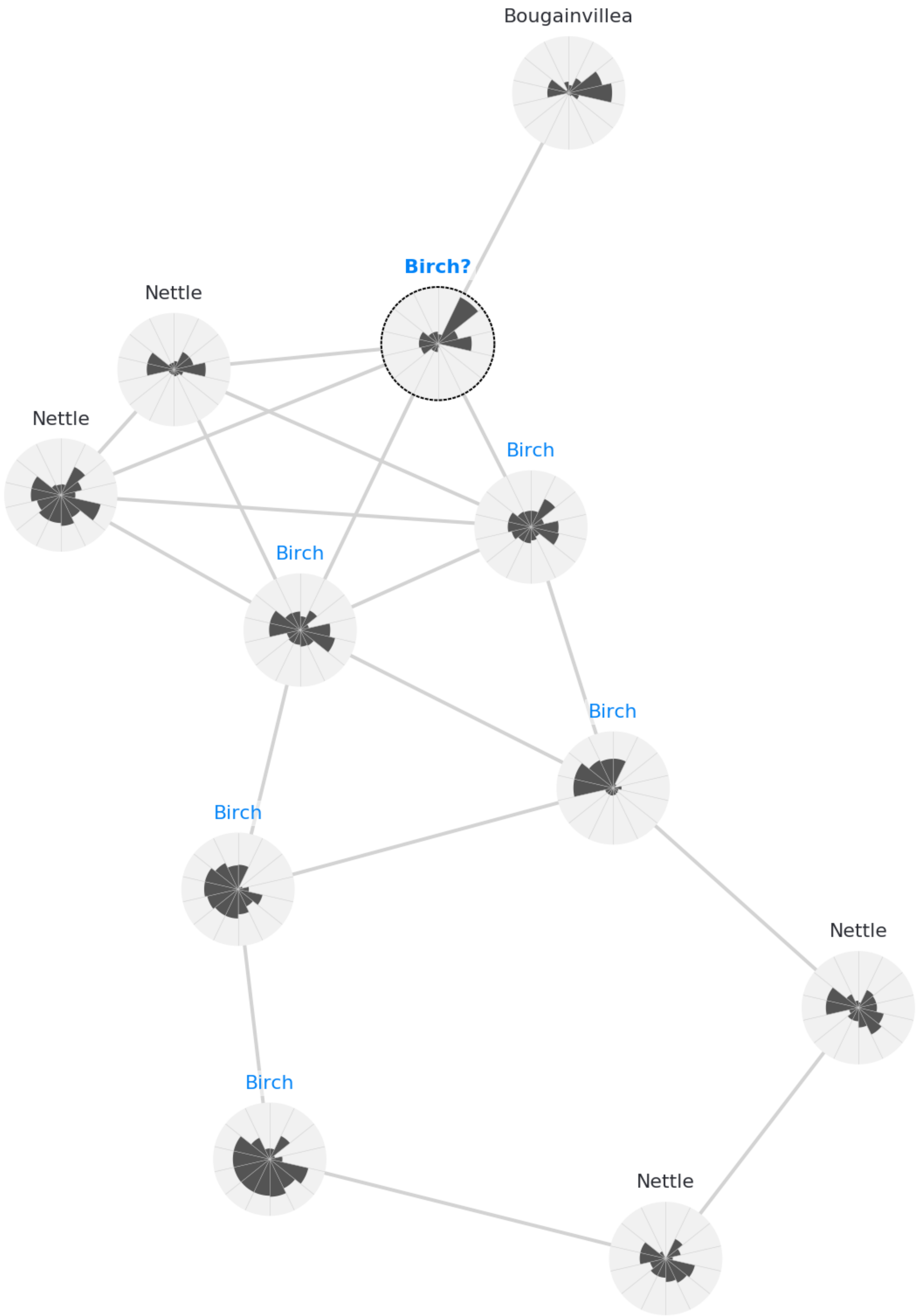
Examples



Grid

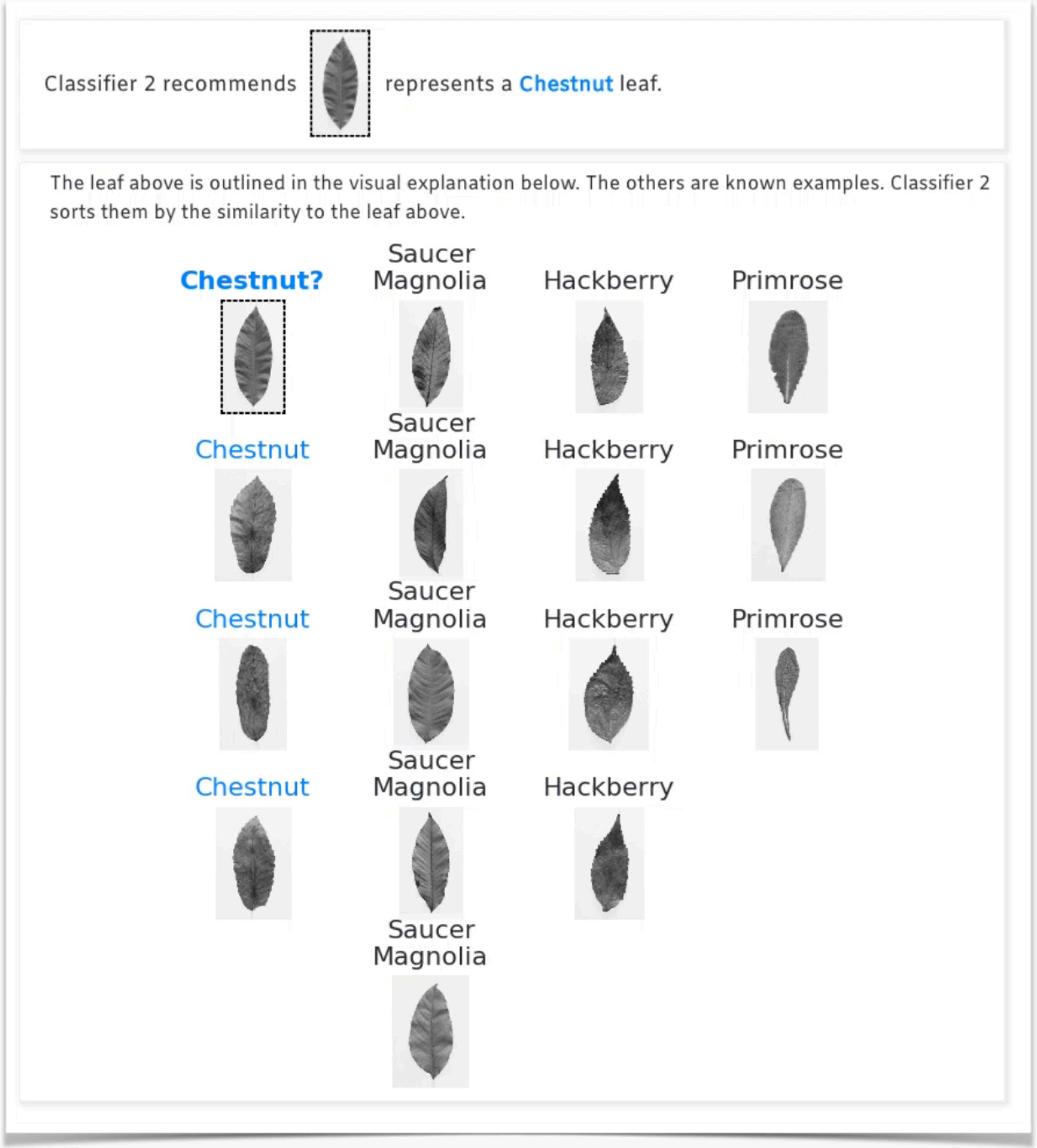


Tree



Graph

Interface & Task



Classifier 2 recommends represents a **Chestnut** leaf.

The leaf above is outlined in the visual explanation below. The others are known examples. Classifier 2 sorts them by the similarity to the leaf above.

Measuring Trust in the Classifier

“Participants’ willingness to follow the recommendation and their self-confidence in the decision.”

- Will you follow this recommendation?
- How do you feel about your decision above?
- Was the explanation helpful in making the decision above?
- A linear ''Trust Meter'' ranged from -100 to +100

Experimental Design

A complete within-subjects design

Each participant finished

two instance representations on two different days

three layouts and a control condition (no explanation)

e.g., tree + roses, none + images

A series of trials

27 trials for each condition

20 correct, 7 incorrect = 74% vs. classifier 71%

a fixed sequence by MC with randomized instances

33 participants from PNNL

19 female, 14 male

16 data scientists, 17 others

Data Collected

Trust Measures

- **Appropriate trust** - correct decision rate
- **Overtrust** - follow an incorrect recommendation
- **Undertrust** - not follow a correct recommendation
- **Self-confidence**

Perceived helpfulness

Trust meter

8,184 / **7,128** trials

= (3+1) layout conditions
x 2 representations
x 27 trials
x 33 participants

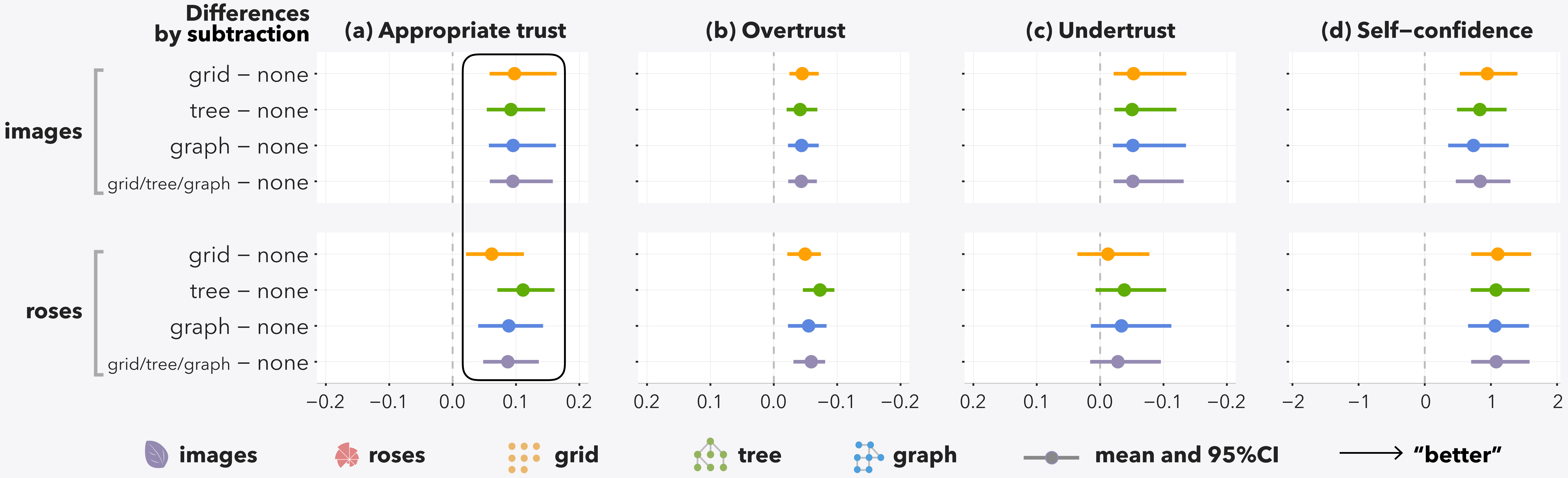
Analyses and Results

Research Questions Five research questions (four for this talk)

Methods bootstrapped 95% CIs, effect sizes,
mixed-effects models for individual differences,
aggregated each participant, and subtracted within participants

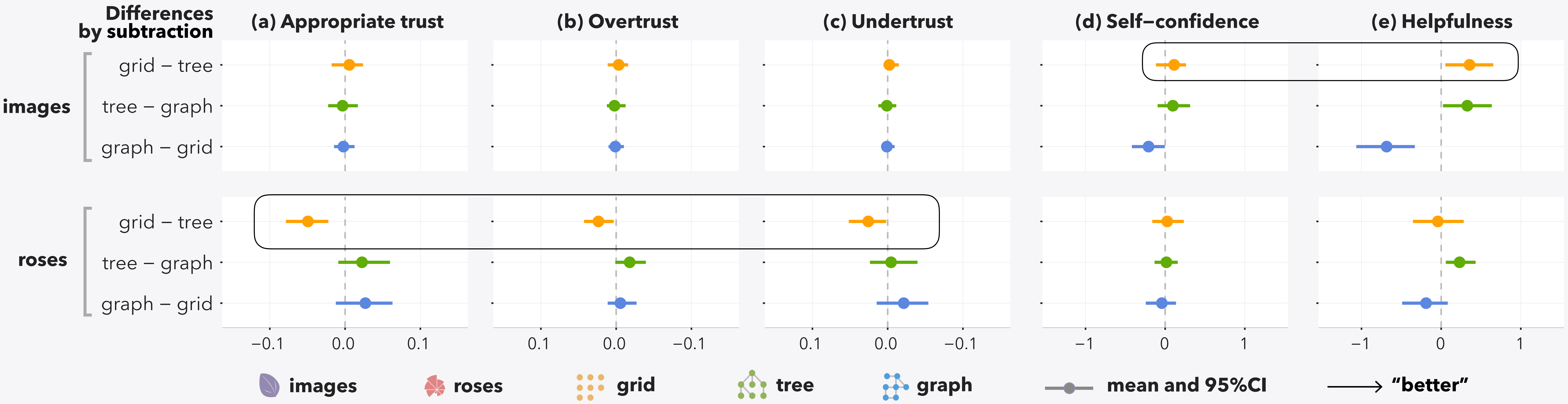
Interpretation Summarizing all confidence intervals

RQ1 Do our visual explanations foster more appropriate trust?



All our visual explanations largely increase appropriate trust, decrease overtrust and underthrust, and improve self-confidence.

RQ2 How did the three spatial layouts (grid, tree, and graph) affect users’ trust?



***Images**: grid explanations are slightly more helpful than tree explanations, which are slightly more helpful than graph explanations.*

***Roses**: tree and graph explanations, especially tree, lead to more appropriate trust than grid explanations.*

RQ3 How did the two instance representations (images and roses) affect users' trust?

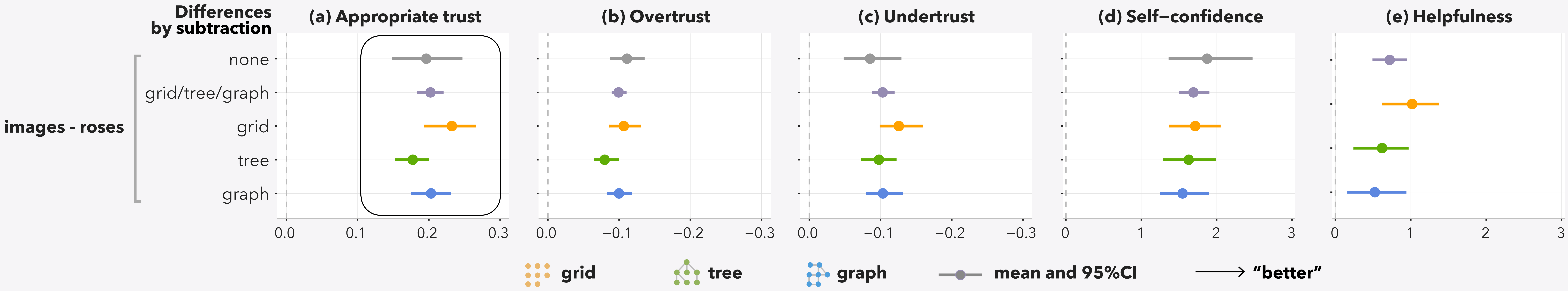
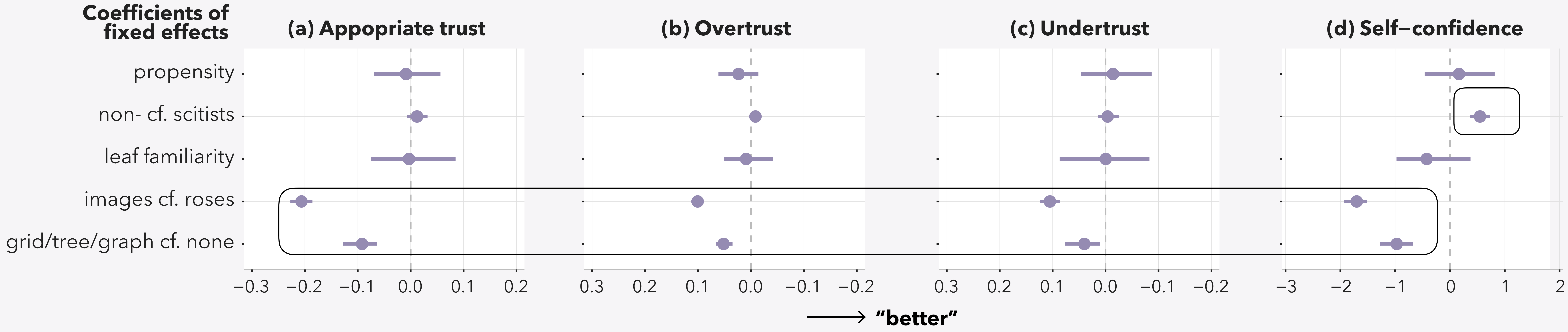


Image-based explanations outperform rose-based explanations on all the dimensions.

RQ4 How did individual differences (e.g., expert users vs. non-expert users, prior knowledge, and propensity to trust) affect users' trust?



The strongest effects come from the two experimental variables:

images outperform roses;

having a visual explanation outperforms no explanation.

The only exception is that non-expert users seem to have more confidence in their decisions.

Summary & Takeaways

- Use a **grid** layout if the representation is easy to understand;
  Use a **tree** layout if the representation is difficult to read or its usability is unknown.

- Understanding and trust are **relevant but different.**

- Future research should consider **appropriate trust,**
  instead of simply measuring an increase in users' trust.
  Overtrust and undertrust should be avoided.

Thank You

“How Do Visual Explanations Foster End Users’ Appropriate Trust in Machine Learning?”

Fumeng Yang	fy@brown.edu
Zhuanyi (Yi) Huang	zhuanyi.huang@pnnl.gov
Jean Scholtz	jean.scholtz@pnnl.gov
Dustin L. Arendt	dustin.arendt@pnnl.gov

✿ **http://www.fmyang.com/projs/ml-trust** ✿