
Before class...

Take one of each:

- *Sanity Checks for Saliency Maps*
- Piece of scrap paper

Sit together in groups of 1-3. There will be an exercise during the second half of class.

You may find it beneficial to talk through the exercise with others.

CS 295B/CS 395B
Systems for Knowledge
Discovery

Lecture 2:
Research Methods
Basics

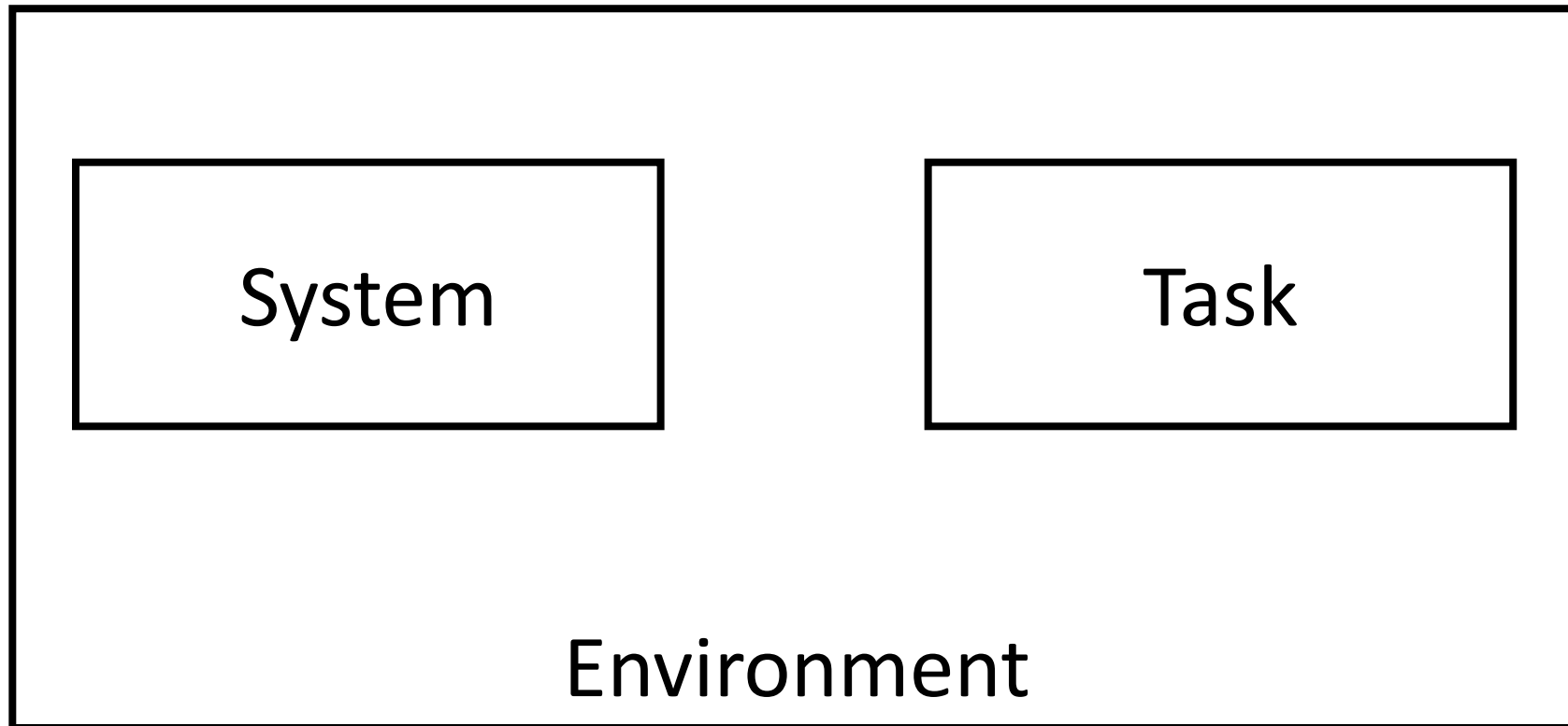


The University of Vermont

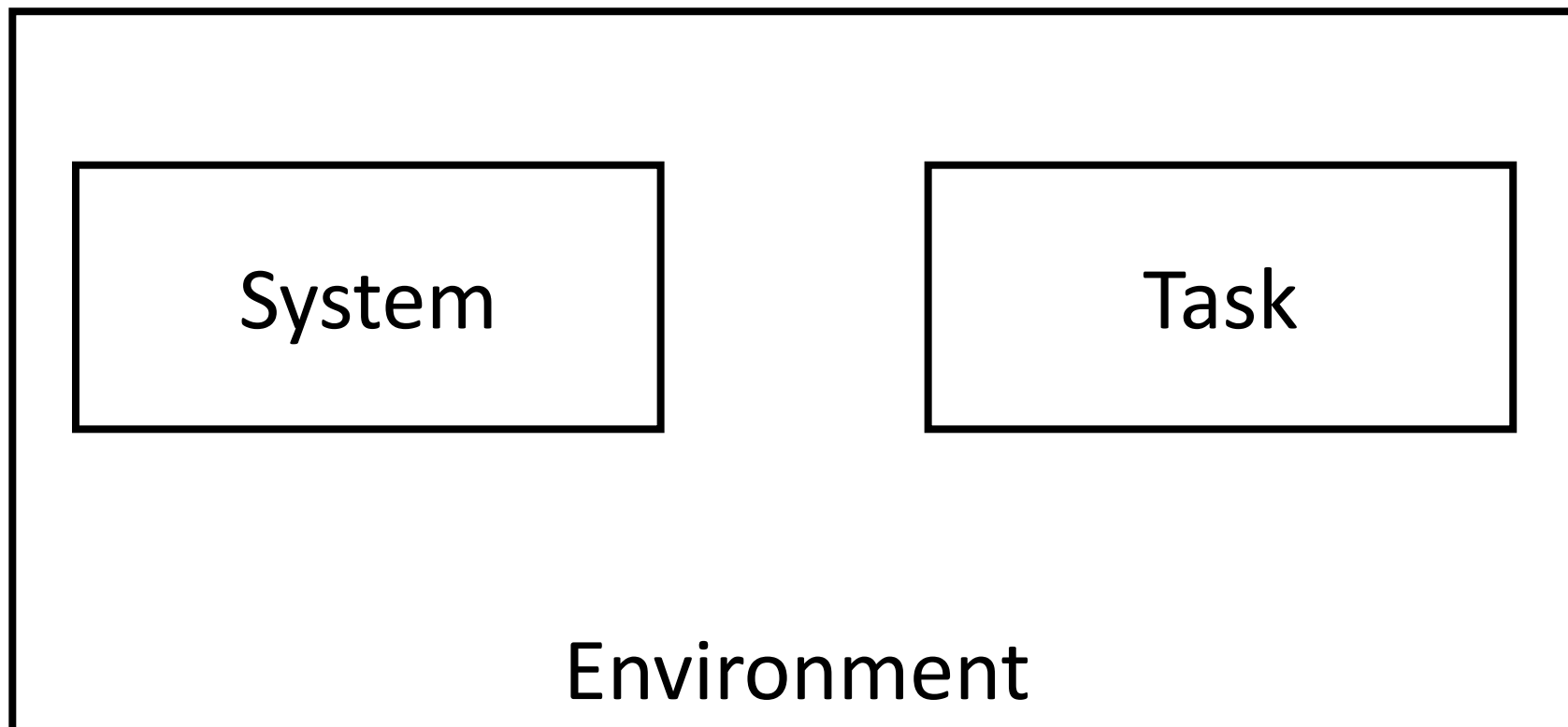
Outline

- Phenomena
- Research Questions
- Hypotheses
- Methods
- Findings
- Contributions and Authorship

Phenomena



Phenomena

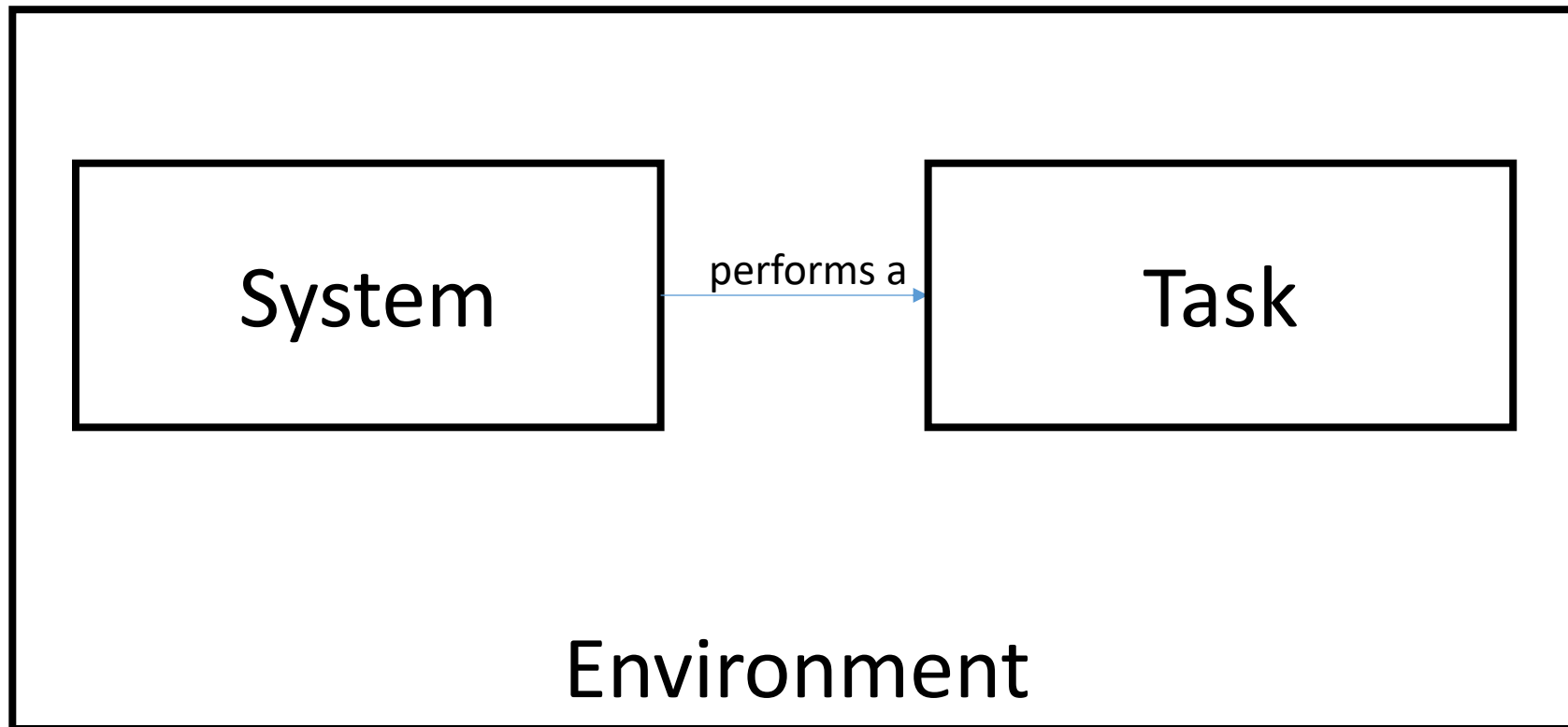


Empirical Methods for
Artificial Intelligence

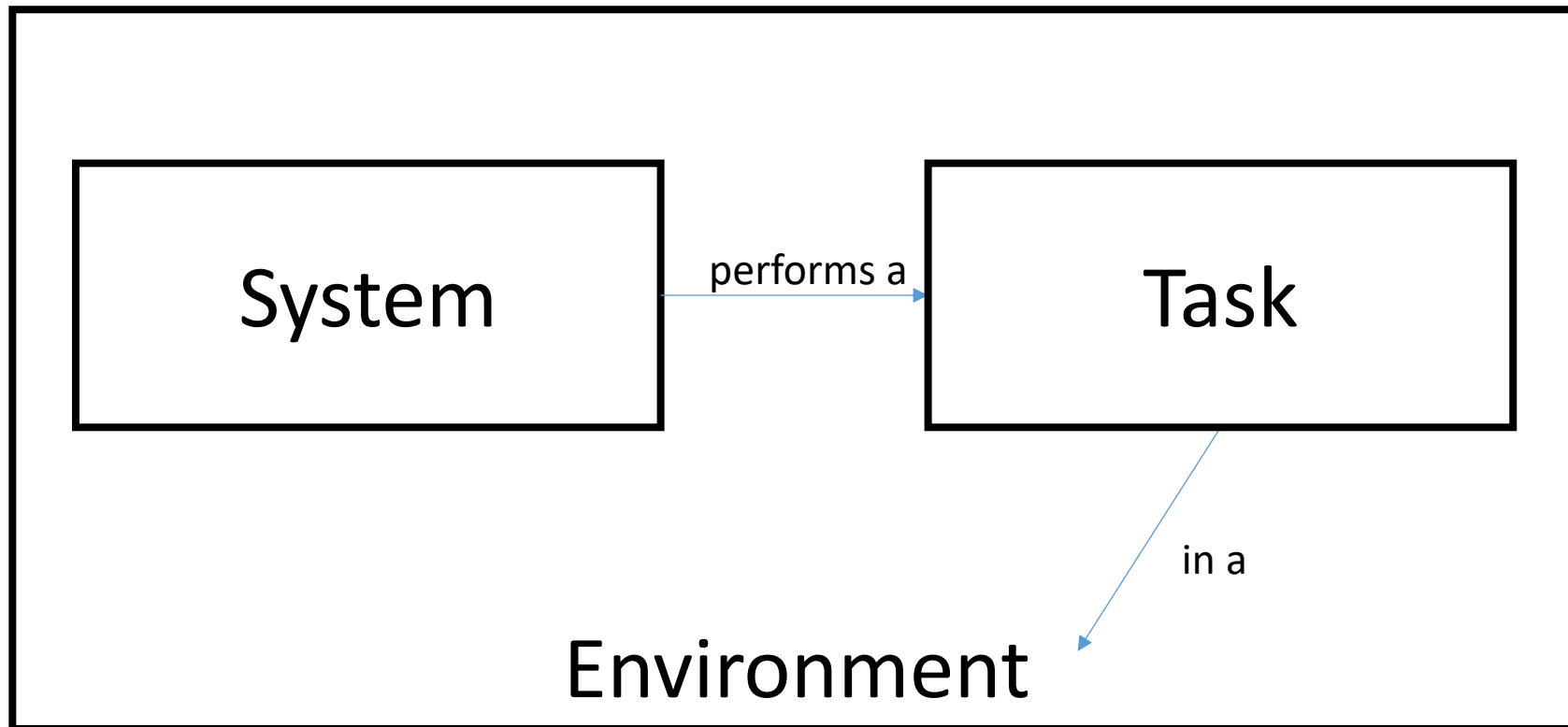
Paul R. Cohen



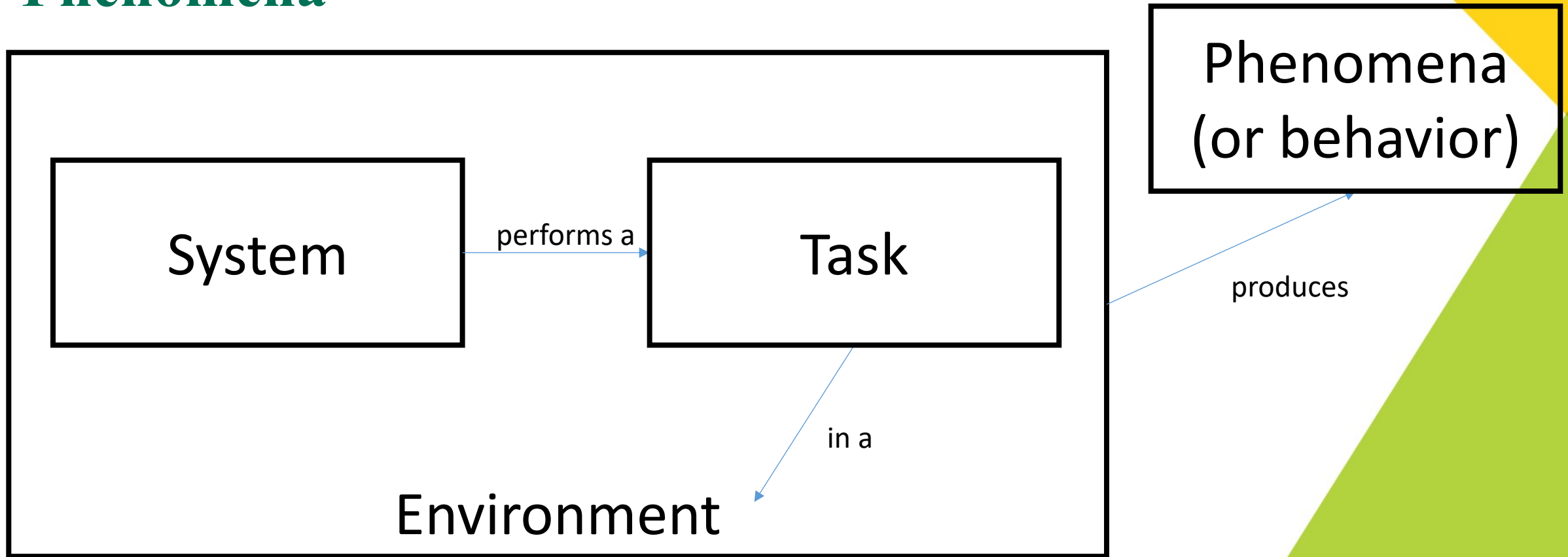
Phenomena



Phenomena



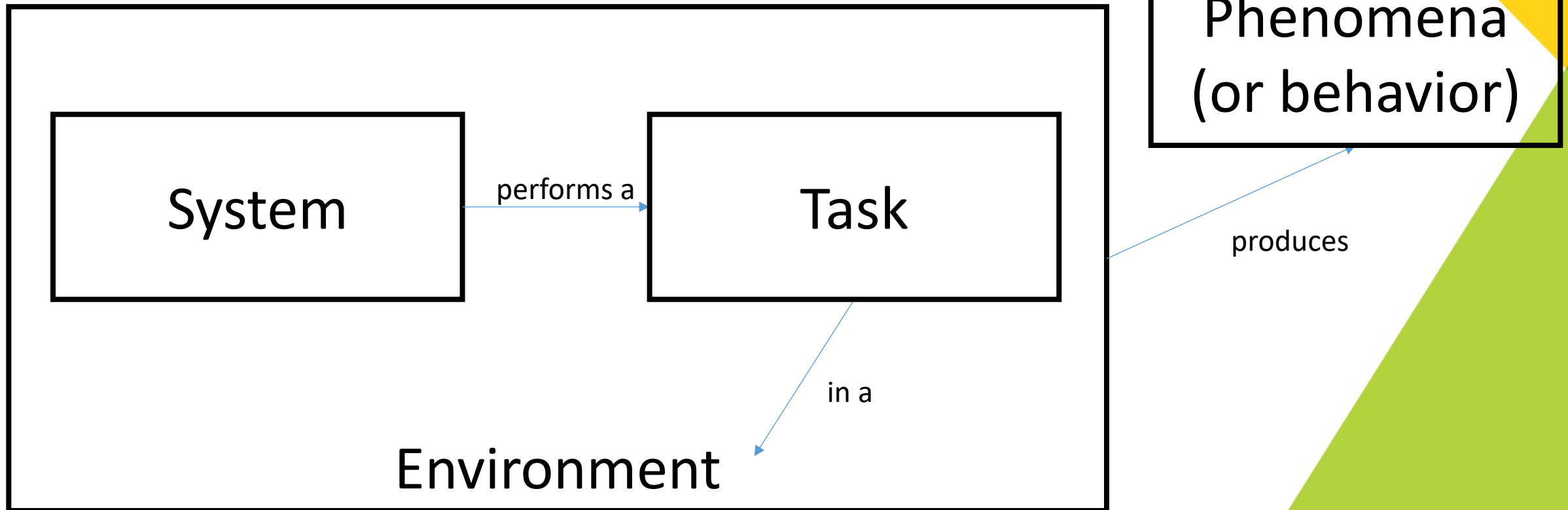
Phenomena



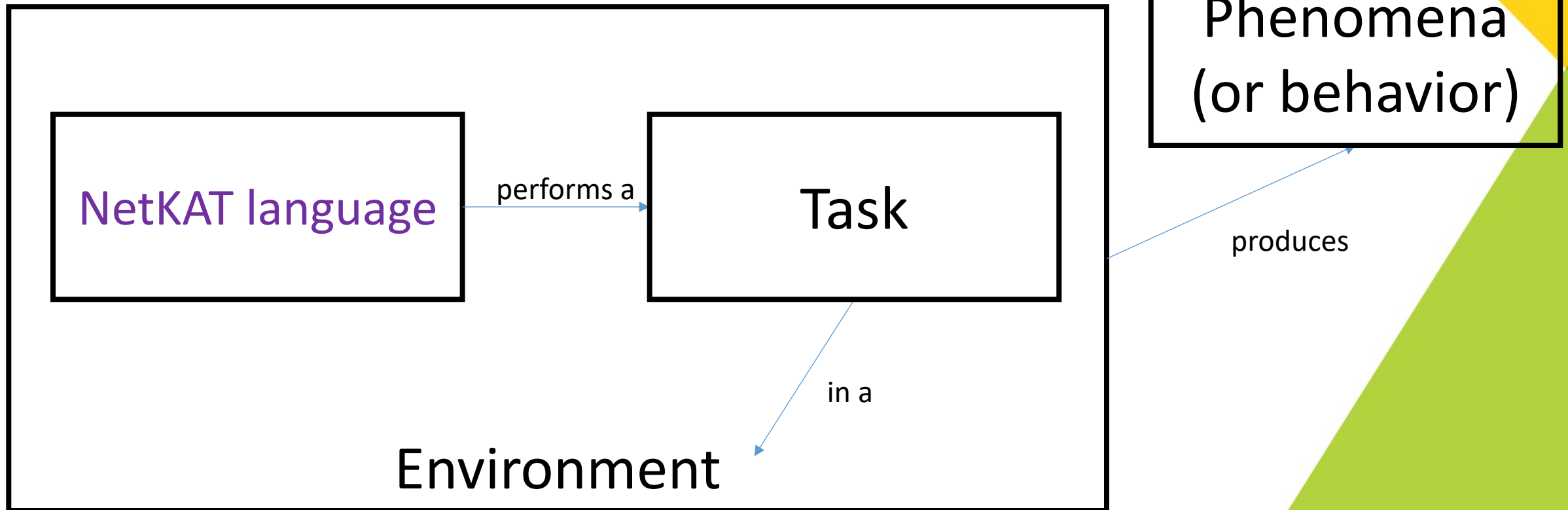
Phenomena: Example 1

**Domain: Software-Defined Networks (SDNs)
(Networks, Programming Languages, Systems Architecture)**

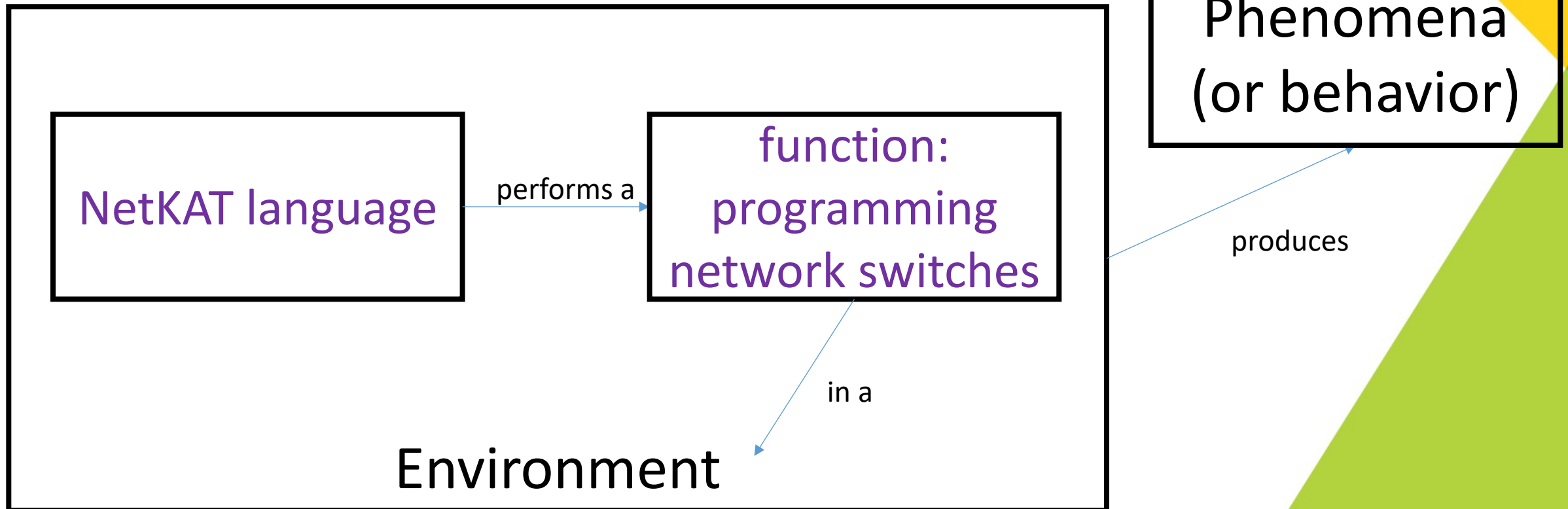
Example: Software-defined networks



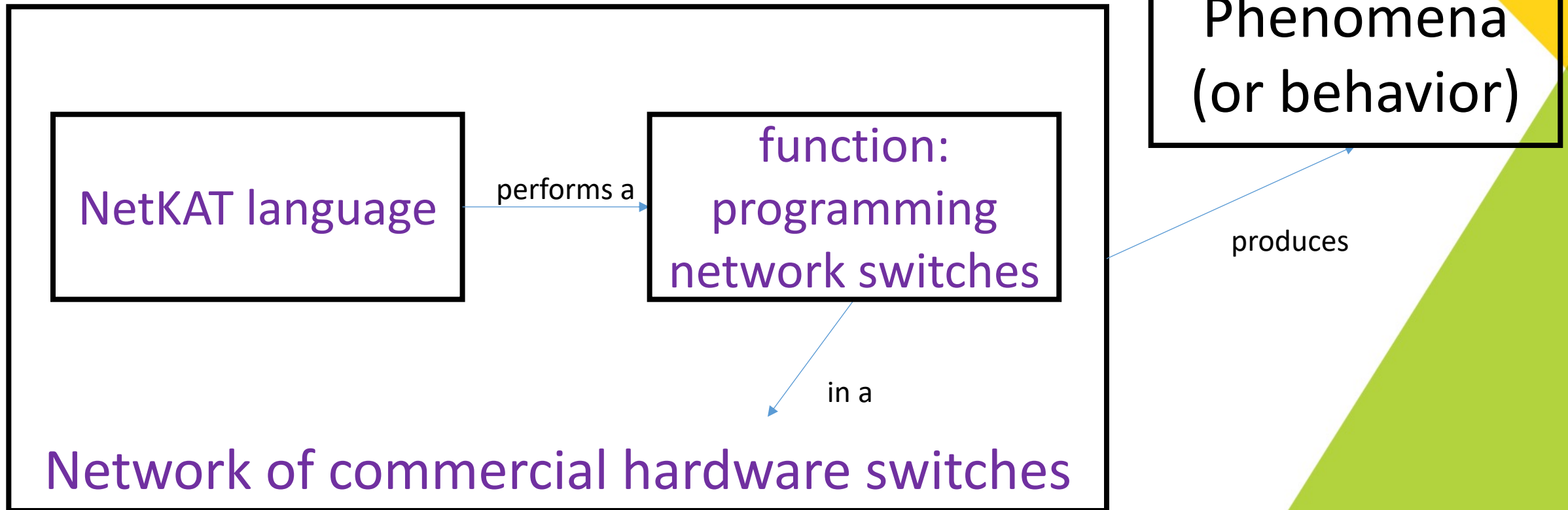
Example: Software-defined networks



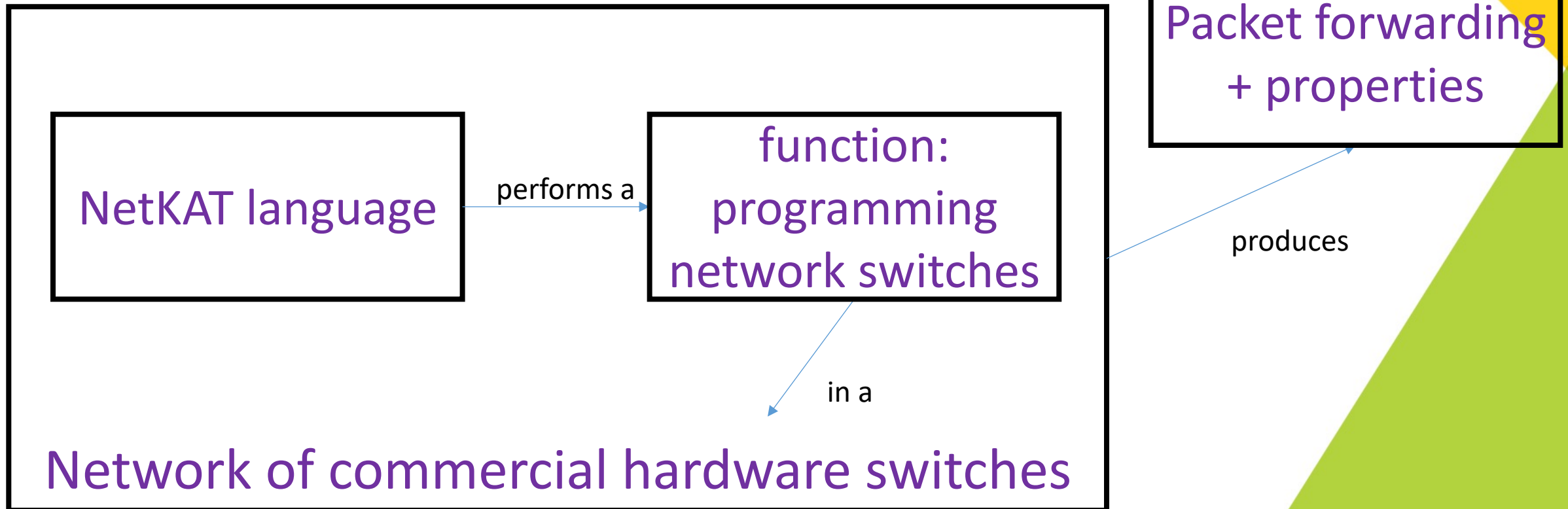
Example: Software-defined networks



Example: Software-defined networks



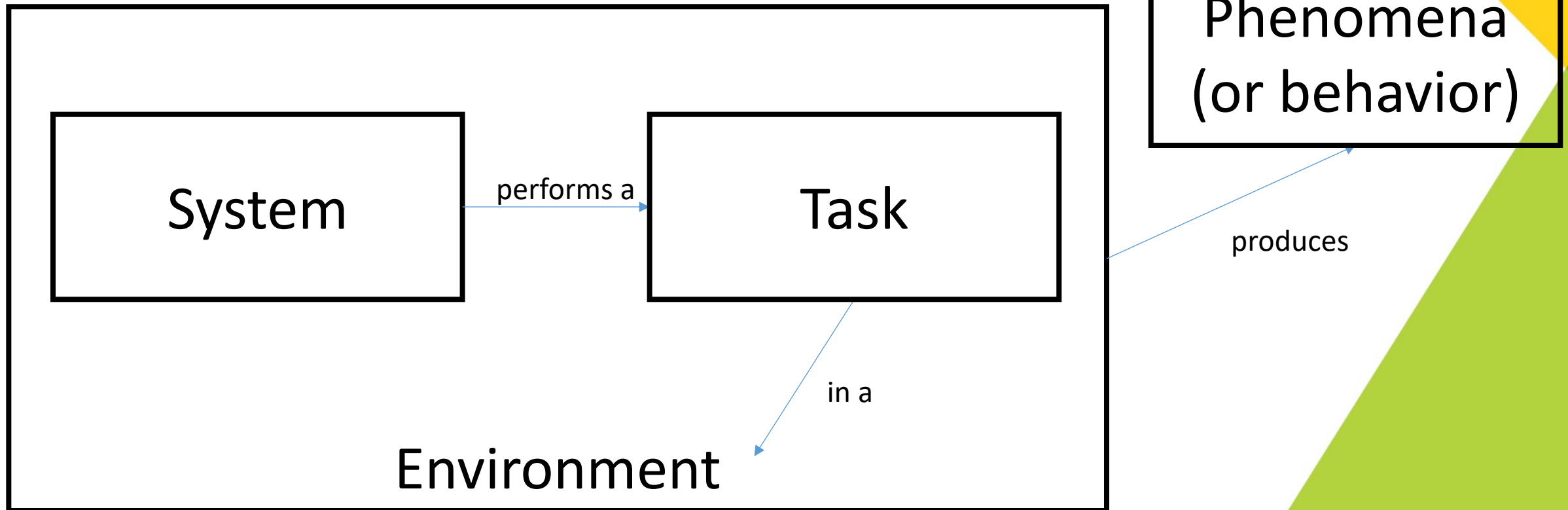
Example: Software-defined networks



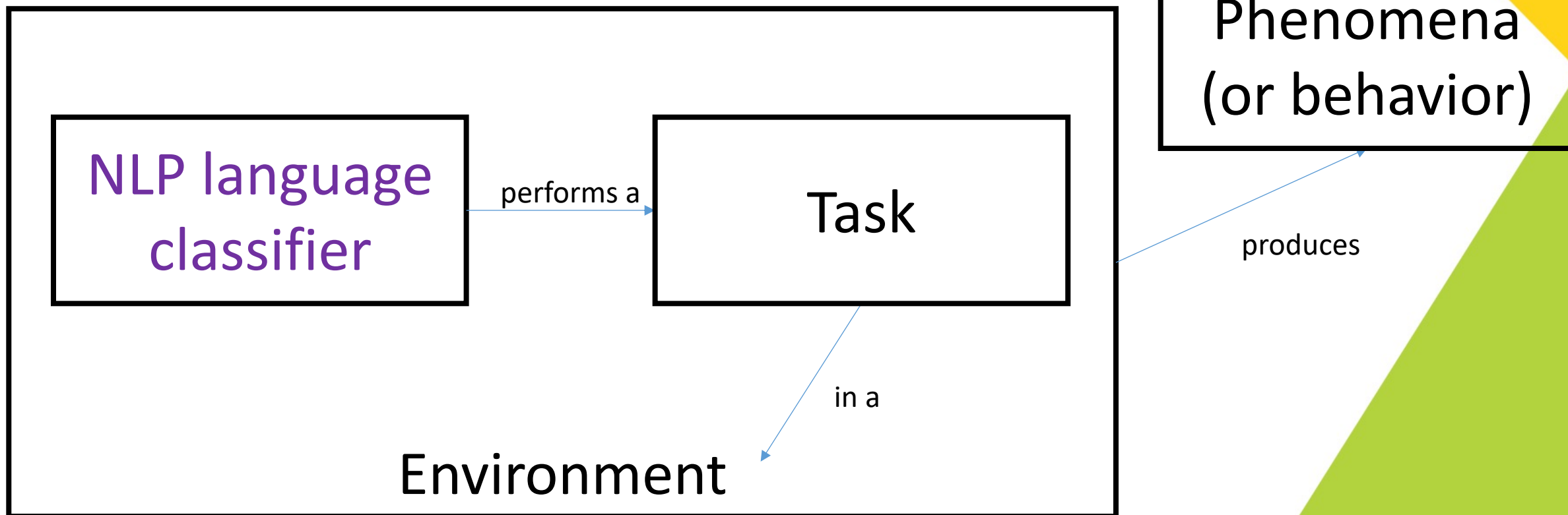
Phenomena: Example 2

Domain: Natural Language Processing (NLP)

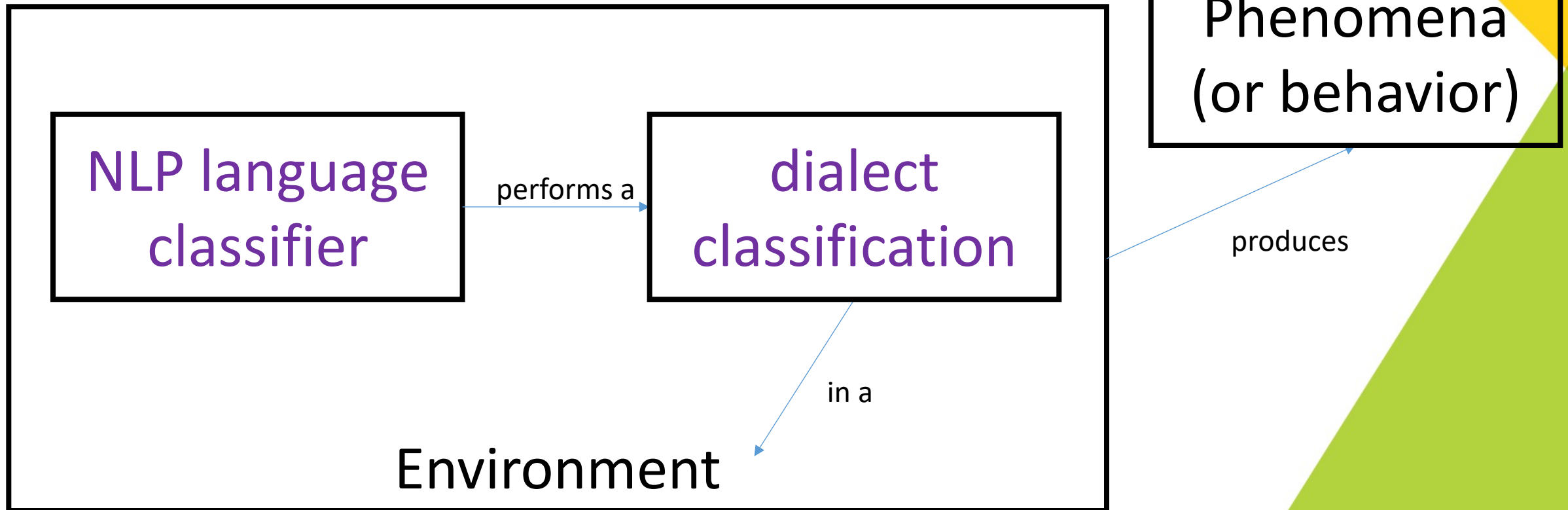
Example: Dialect identification on Twitter



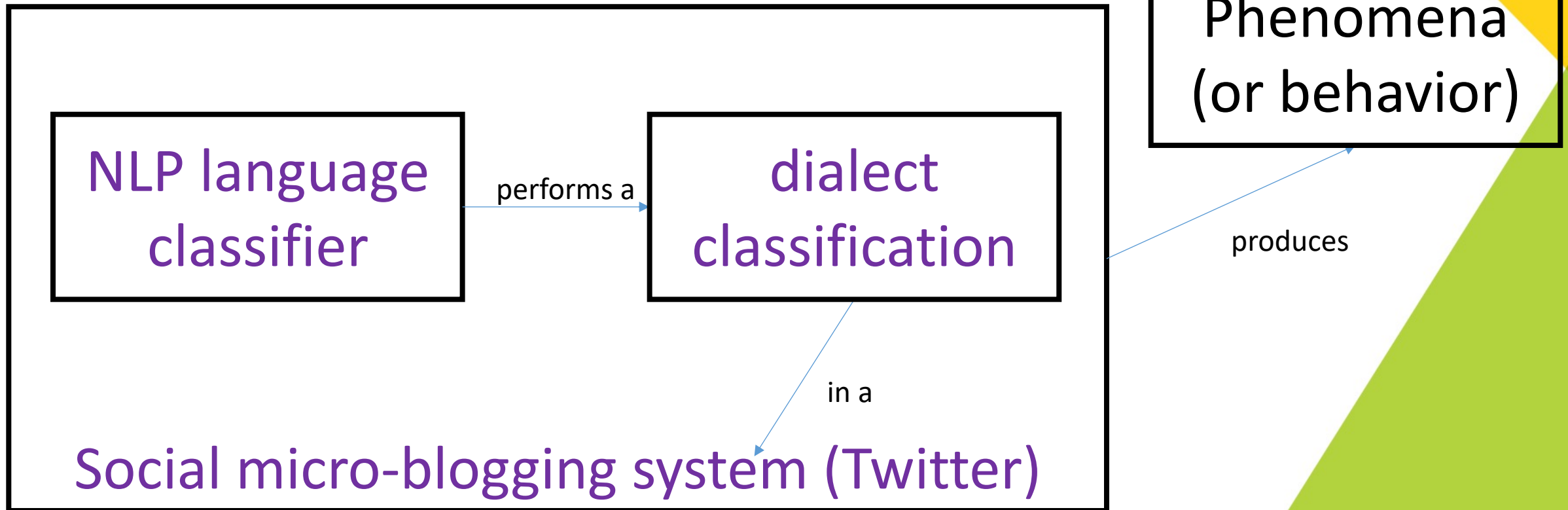
Example: Dialect identification on Twitter



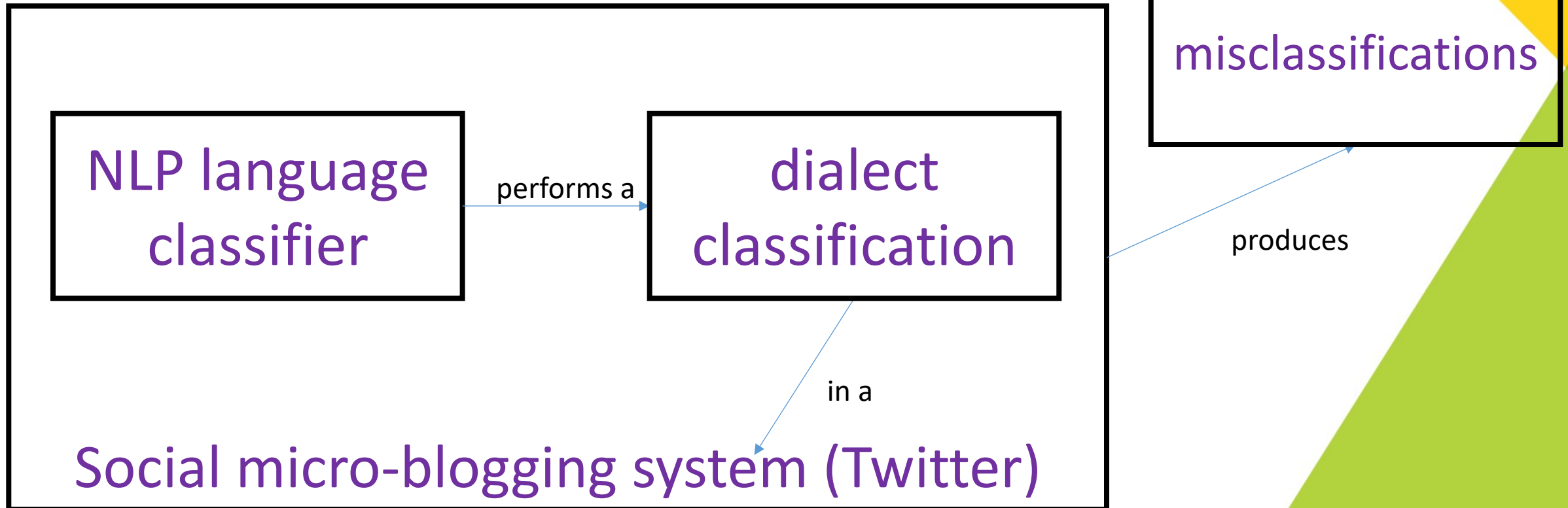
Example: Dialect identification on Twitter



Example: Dialect identification on Twitter



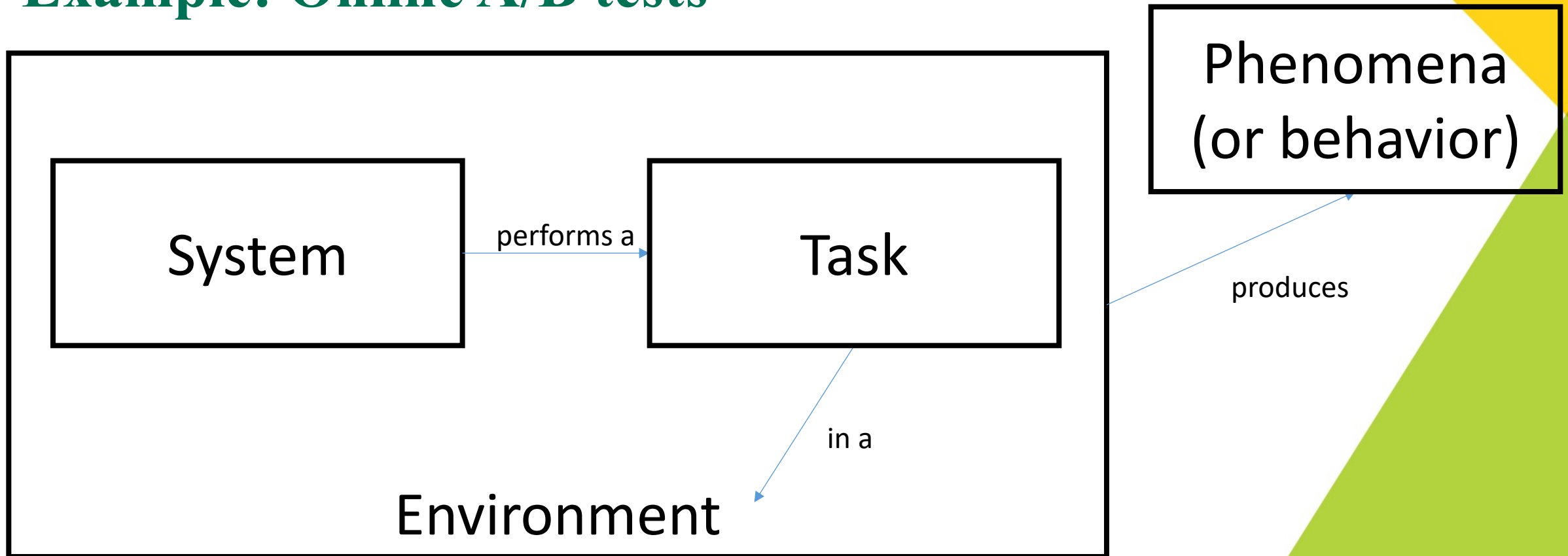
Example: Dialect identification on Twitter



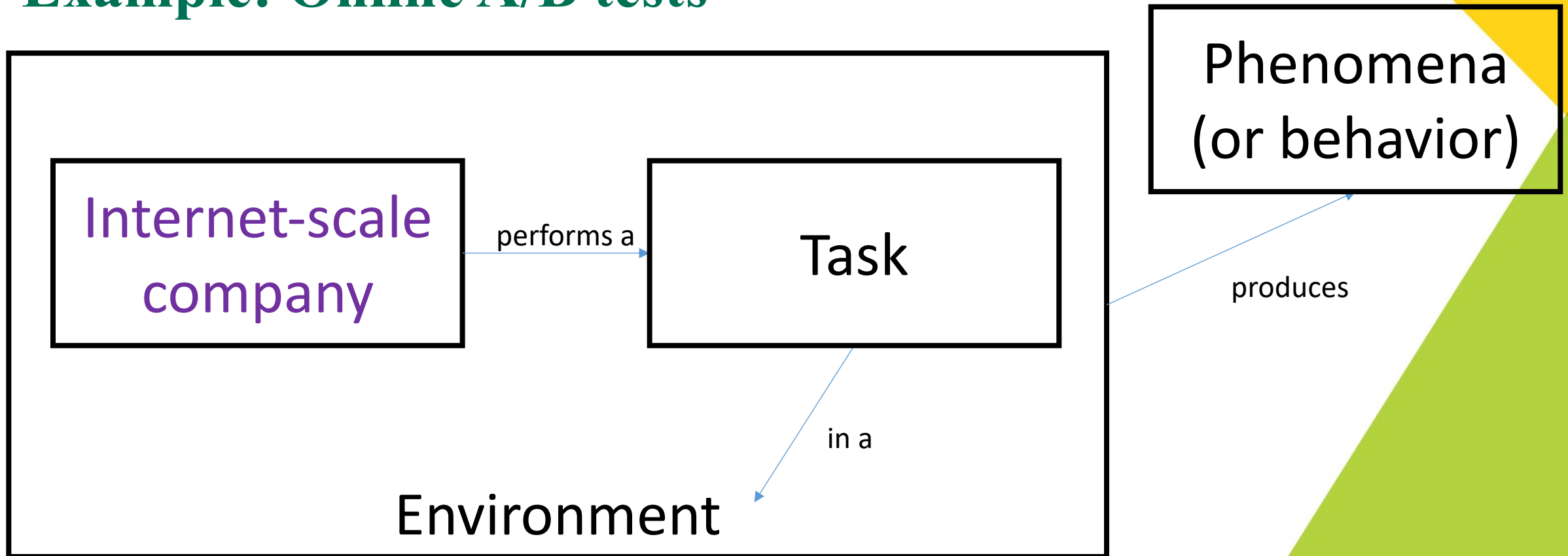
Phenomena: Example 3

**Domain: Online A/B tests
(Experimental design, software architecture)**

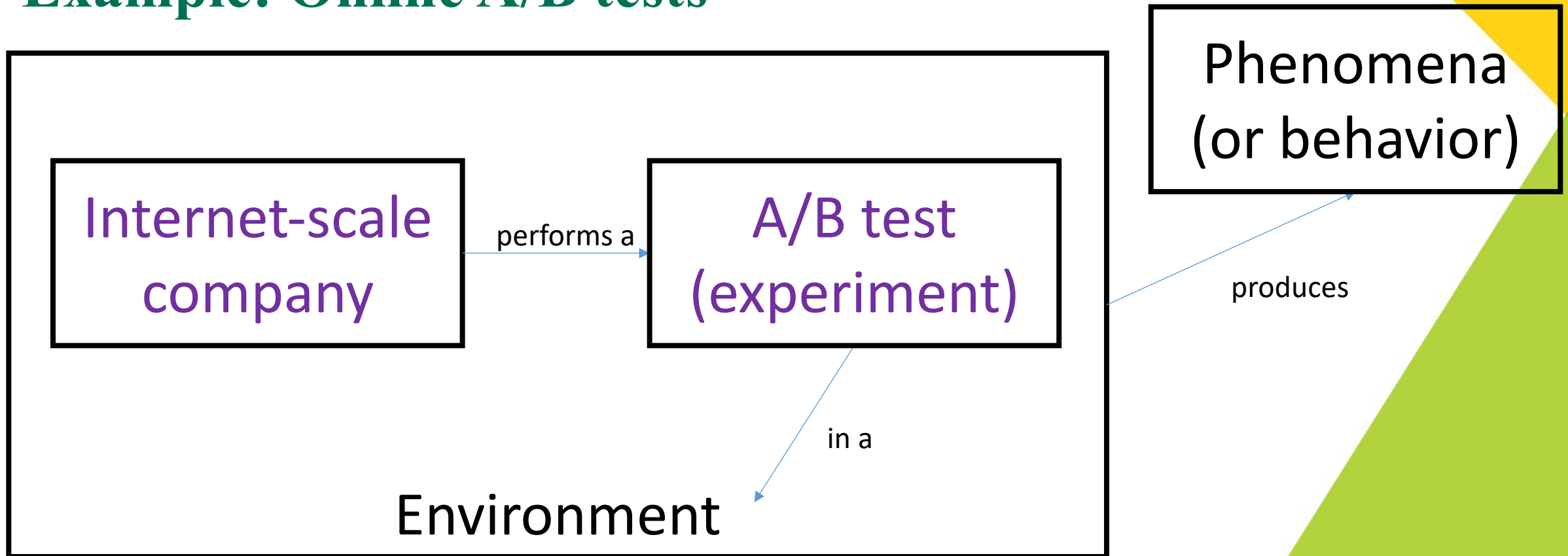
Example: Online A/B tests



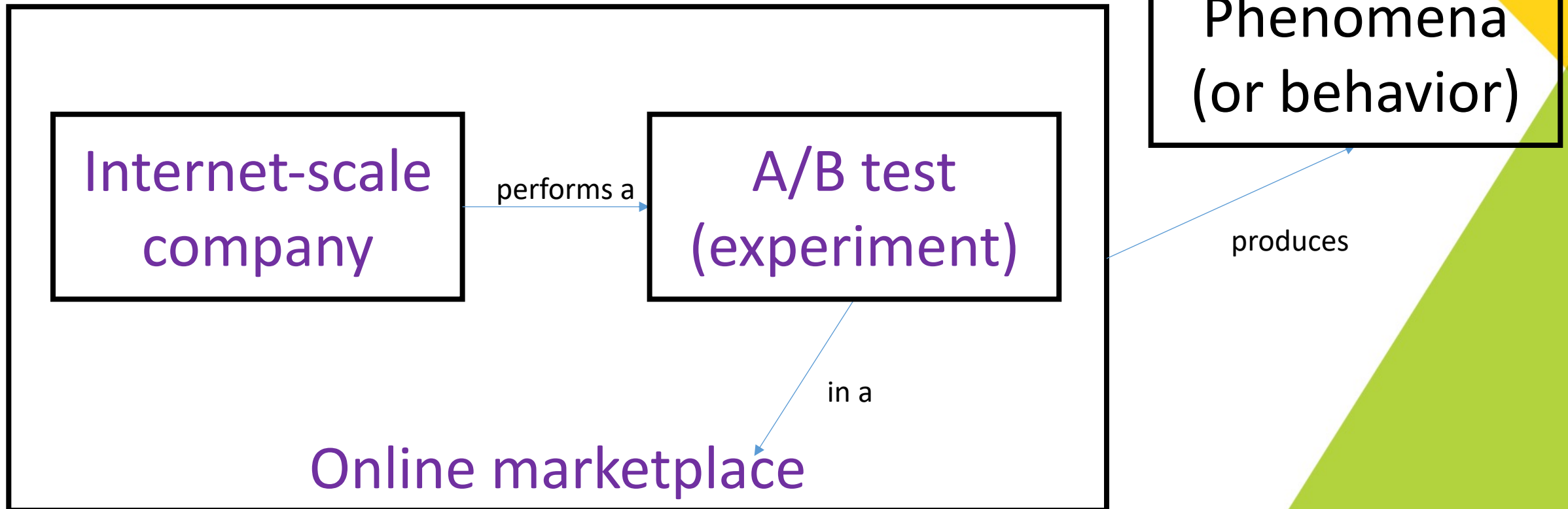
Example: Online A/B tests



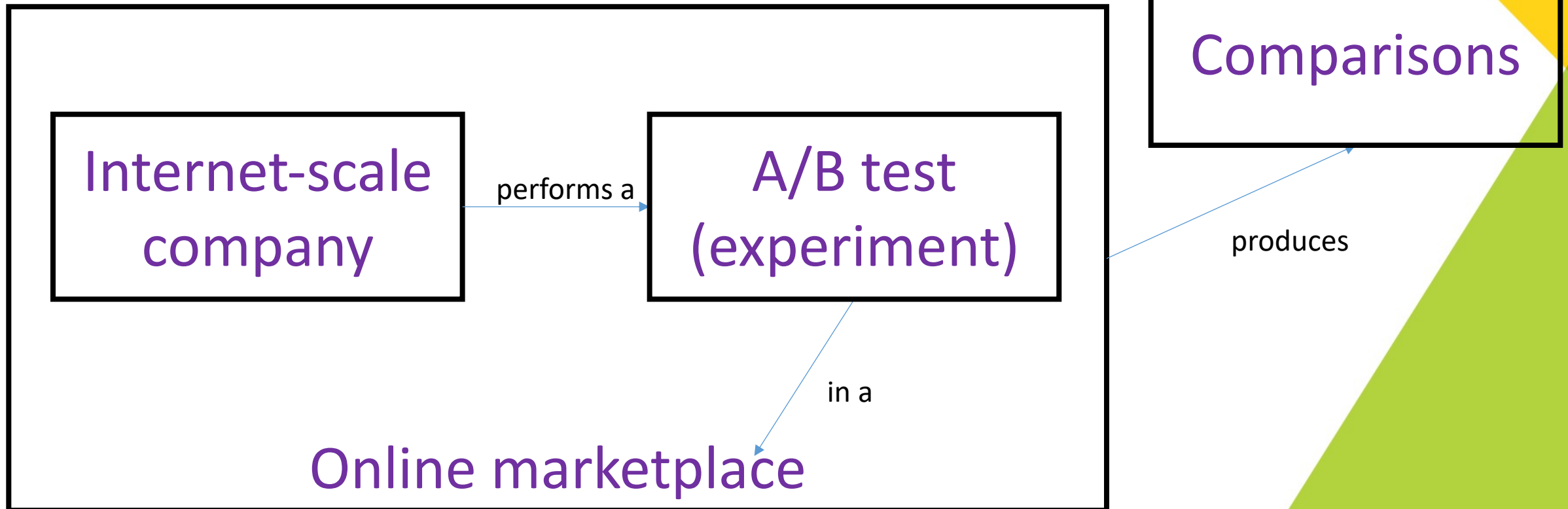
Example: Online A/B tests



Example: Online A/B tests



Example: Online A/B tests



Outline

- ~~Phenomena~~
- Research Questions
- Hypotheses
- Methods
- Findings
- Contributions and Authorship

Research Questions

- Many guides and acronyms out there
 - Can be specific to domain
 - Common themes:
 - Specific
 - Entail hypotheses
 - Scale with project scope; heuristic: length inversely proportional to the time-scale
- Remember: *must be in the form of a question!*

Research Questions: A loose progression

Descriptive

- What are the characteristics of <phenomena>?

Associative

- What is the relationship between...?
- Under what conditions...

Causal

- Can we <verb> <noun> such that <dependent clause>?
- Does X cause Y when Z...?

Research Questions: A loose progression

Descriptive

- What are the characteristics of <phenomena>?

Associative

- What is the relationship between...?
- Under what conditions...

Causal

- Can we <verb> <noun> such that <dependent clause>?
- Does X cause Y when Z...?

Research Questions: A loose progression

Descriptive


- What are the characteristics of <phenomena>?

Associative

- What is the relationship between...?
- Under what conditions...

Causal

- Can we <verb> <noun> such that <dependent clause>?
- Does X cause Y when Z...?



Identifies important features to use later.

More common in less mature topics.

Research Questions: A loose progression

Descriptive

- What are the characteristics of <phenomena>?

Associative

- What is the relationship between...?
- Under what conditions...

Causal

- Can we <verb> <noun> such that <dependent clause>?
- Does X cause Y when Z...?

Credit: Dr. David D. Jensen

Research Questions: A loose progression

Descriptive


- What are the characteristics of <phenomena>?

Associative

- What is the relationship between...?
- Under what conditions...

Causal

- Can we <verb> <noun> such that <dependent clause>?
- Does X cause Y when Z...?



Given some features or variables, look at correlations.

Research Questions: A loose progression

Descriptive

- What are the characteristics of <phenomena>?

Associative

- What is the relationship between...?
- Under what conditions...

Causal

- Can we <verb> <noun> such that <dependent clause>?
- Does X cause Y when Z...?

Research Questions: A loose progression

Descriptive


- What are the characteristics of <phenomena>?

Associative

- What is the relationship between...?
- Under what conditions...

Causal

- Can we <verb> <noun> such that <dependent clause>?
- Does X cause Y when Z...?

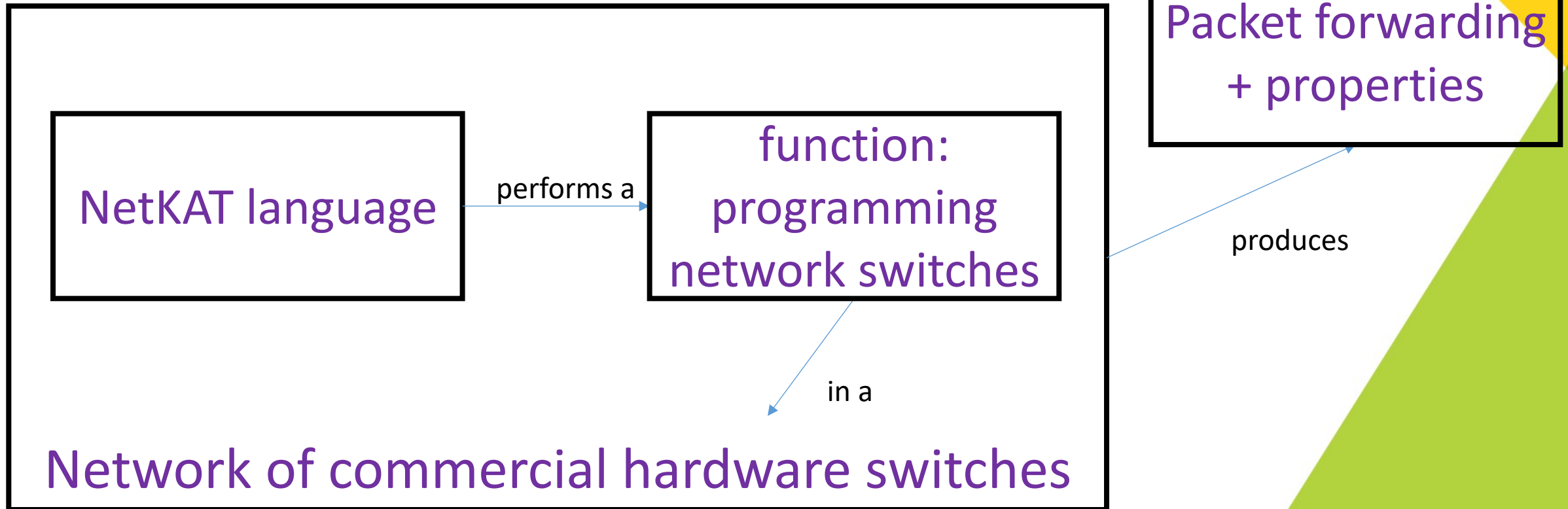


Ultimate goal: find
causal mechanism

Research Questions: Example 1

**Domain: Software-Defined Networks (SDNs)
(Networks, Programming Languages, Systems Architecture)**

Example: Software-defined networks



How to construct a research question...

Big ideas and problem spaces:

- Specialized hardware is expensive
- Existing protocol for cheap software is too hard to use (low-level) and thus error-prone
- Hard to make guarantees about existing protocol

How to construct a research question...

Big ideas and problem spaces:

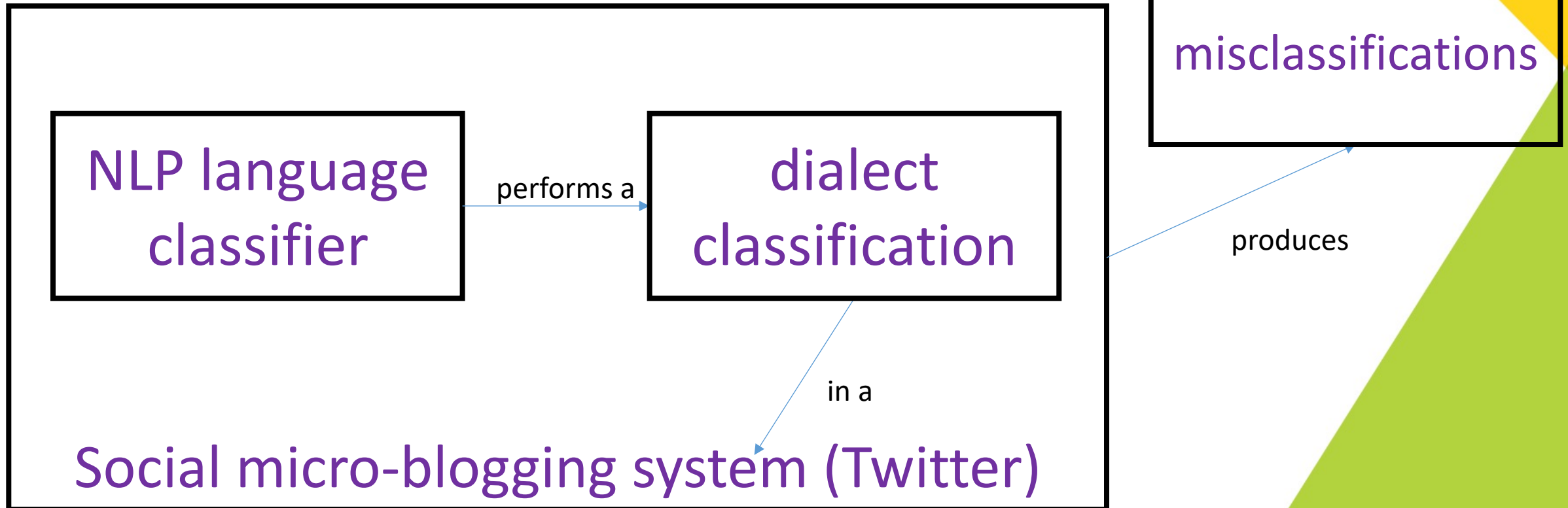
- Specialized hardware is expensive
- Existing protocol for cheap software is too hard to use (low-level) and thus error-prone
- Hard to make guarantees about existing protocol

Can we design a packet-forwarding language with a simple syntax that will compile to an existing protocol **and provide** provable guarantees about, e.g. load and reachability, and if so, **what is** the average performance cost **compared to** bespoke policies?

Research Questions: Example 2

Domain: Natural Language Processing (NLP)

Example: Dialect identification on Twitter



How to construct a research question...

Big ideas and problem spaces:

- Twitter classifies tweets according to language
- There is a paucity of data for many dialects
- Misclassification has disparate impact

How to construct a research question...

Big ideas and problem spaces:

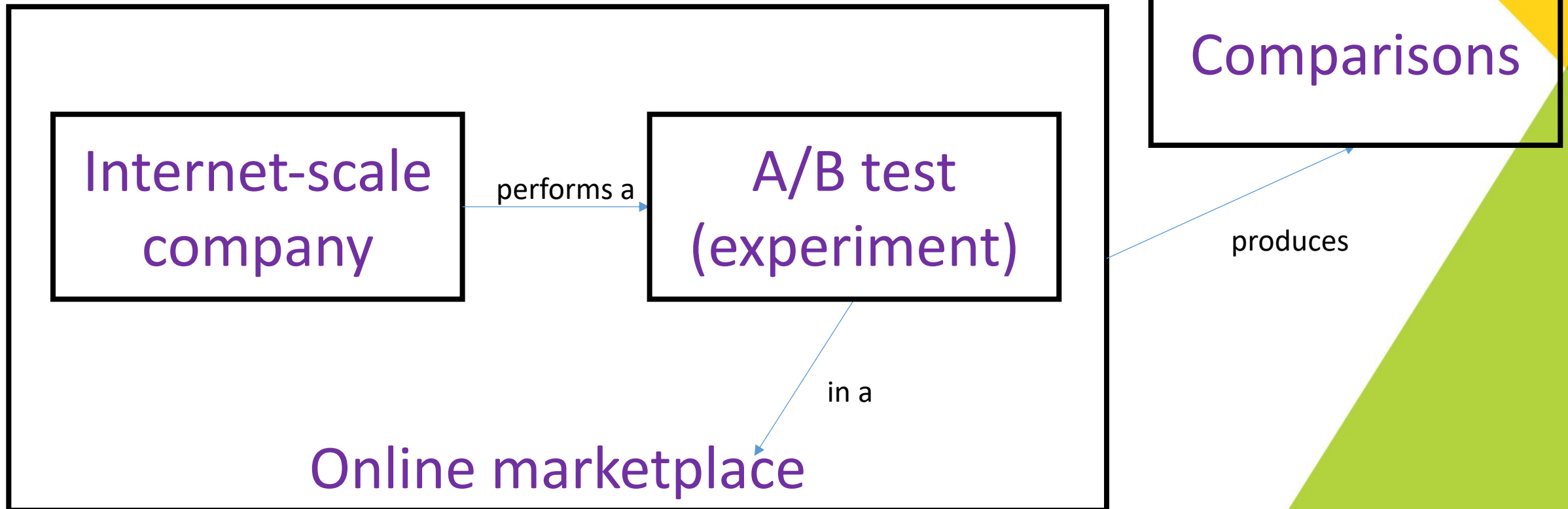
- Twitter classifies tweets according to language
- There is a paucity of data for many dialects
- Misclassification has disparate impact

What are the characteristics and consequences of the misclassification of African-American English on Twitter ***in terms of*** topic classification, and ***how can our findings be used to establish a general method*** for dialect classification?

Research Questions: Example 3

**Domain: Online A/B tests
(Experimental design, software architecture)**

Example: Online A/B tests



How to construct a research question...

Big ideas and problem spaces:

- Experimental design is well-established and understood...offline settings (expensive)
- Online settings: experiments can be cheap, fast, and numerous
- Online setting presents novel challenges

How to construct a research question...

Big ideas and problem spaces:

- Experimental design is well-established and understood...offline settings (expensive)
- Online settings: experiments can be cheap, fast, and numerous
- Online setting presents novel challenges

What are some necessary features in order to run a very large number of concurrent field experiments at Internet-scale firms, with little overhead and very low probability of catastrophic failure or loss of users/income?

Outline

- ~~• Phenomena~~
- ~~• Research Questions~~
- Hypotheses
- Methods
- Findings
- Contributions and Authorship

Hypotheses

- Fall out of the research question (RQ entails H)
- Falsifiable
 - Not always possible
 - Philosophical questions about what constitute evidence
- Hypotheses and methods strongly linked
- A hypothesis is a *model*

Hypotheses: Example 1

**Domain: Software-Defined Networks (SDNs)
(Networks, Programming Languages, Systems Architecture)**

How research questions entail hypotheses...

Can we design a packet-forwarding language with a simple syntax that will compile to an existing protocol **and provide** provable guarantees about, e.g. load and reachability, and if so, **what is** the average performance cost **compared to** bespoke policies?

How research questions entail hypotheses...

Can we design a packet-forwarding language with a simple syntax that will compile to an existing protocol **and provide** provable guarantees about, e.g. load and reachability, and if so, **what is** the average performance cost **compared to** bespoke policies?

Are these suitable hypotheses?

- “We can design such a language.”
- “The language we designed ensures load is balanced within...”
- “Our language produces more efficient policies and takes less time to write.”

Hypotheses: Example 2

Domain: Natural Language Processing (NLP)

How to construct a research question...

*What are the **characteristics and consequences** of the misclassification of African-American English (AAE) on Twitter **in terms of** topic classification, and **how can our findings be used to establish a general method** for dialect classification?*

How to construct a research question...

*What are the **characteristics and consequences** of the misclassification of African-American English (AAE) on Twitter **in terms of** topic classification, and **how can our findings be used to establish a general method** for dialect classification?*

Are these suitable hypotheses?

- “The rate of topic misclassification for AAE is higher than SAE.”
- “AAE speakers see fewer topical tweets than SAE speakers, even for their SAE tweets.”
- “Misclassification causes AAE speakers to see more disinformation.”

Hypotheses: Example 3

**Domain: Online A/B tests
(Experimental design, software architecture)**

How to construct a research question...

***What are some necessary features** in order to run a very large number of concurrent field experiments at Internet-scale firms, with little overhead and very low probability of catastrophic failure or loss of users/income?*

How to construct a research question...

What are some necessary features in order to run a very large number of concurrent field experiments at Internet-scale firms, with little overhead and very low probability of catastrophic failure or loss of users/income?

Are these suitable hypotheses?

- “We can place one person in several experiments at the same time.”
- “Client-side random assignment scales better without sacrificing power.”
- “The firm will lose users if they know they are in experiments.”

Outline

- ~~Phenomena~~
- ~~Research Questions~~
- ~~Hypotheses~~
- Methods
- Findings
- Contributions and Authorship

Methods

- Procedures you use to validate (or falsify) your hypotheses
- Usually a “methods” section of a paper
- Can be community-specific
- Most of the course

Methods

- Formal proofs
- Field experiments
- Simulation studies
- Surveys
- Interviews
- Case studies
- Statistical analysis and modeling
- Laboratory experiments
- Causal inference over observational data
- Prototypes

Outline

- ~~Phenomena~~
- ~~Research Questions~~
- ~~Hypotheses~~
- ~~Methods~~
- Findings
- Contributions and Authorship

Findings

- Outcomes that are in service of answering the research question.
- Usually generated from hypotheses
- Must be clear about whether and how they generalize
- Often contextualized (more narrative, less formal)
- Should be the easiest element to find in published work

Outline

- ~~Phenomena~~
- ~~Research Questions~~
- ~~Hypotheses~~
- ~~Methods~~
- ~~Findings~~
- Contributions and Authorship

Contributions and Authorship

- Paper's contributions \neq author's contributions
- Paper's contributions
 - Findings but also...
 - Software, models, methods, corpora, ...
 - Community-specific!
- Author's contribution \rightarrow Very community-specific, norms-based, evolves with time
 - *data collection: contribution in natural and social sciences, not in most computer science fields*
 - *software/programming: usually not enough in most CS disciplines (sometimes okay for undergraduates), unless the software/programming is itself a contribution*

For Authorship criteria in CS: <https://www.acm.org/publications/policies/authorship>

Sanity Checks for Saliency Maps

Julius Adebayo[‡], Justin Gilmer[‡], Michael Muelly[‡], Ian Goodfellow[‡], Moritz Hardt^{‡†}, Been Kim[‡]
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com

[‡]Google Brain

[†]University of California Berkeley

Abstract

Saliency methods have emerged as a popular tool to highlight features in an input deemed relevant for the prediction of a learned model. Several saliency methods have been proposed, often guided by visual appeal on image data. In this work, we propose an actionable methodology to evaluate what kinds of explanations a given method can and cannot provide. We find that reliance, solely, on visual assessment can be misleading. Through extensive experiments we show that some existing saliency methods are independent both of the model and of the data generating process. Consequently, methods that fail the proposed tests are inadequate for tasks that are sensitive to either data or model, such as, finding outliers in the data, explaining the relationship between inputs and outputs that the model learned, and debugging the model. We interpret our findings through an analogy with edge detection in images, a technique that requires neither training data nor model. Theory in the case of a linear model and a single-layer convolutional neural network supports our experimental findings[‡].

1 Introduction

As machine learning grows in complexity and impact, much hope rests on explanation methods as tools to elucidate important aspects of learned models [1, 2]. Explanations could potentially help satisfy regulatory requirements [3], help practitioners debug their model [4, 5], and perhaps, reveal bias or other unintended effects learned by a model [6, 7]. *Saliency methods*[‡] are an increasingly popular class of tools designed to highlight relevant features in an input, typically, an image. Despite much excitement, and significant recent contribution [8–21], the valuable effort of explaining machine learning models faces a methodological challenge: *the difficulty of assessing the scope and quality of model explanations*. A paucity of principled guidelines confound the practitioner when deciding between an abundance of competing methods.

We propose an actionable methodology based on randomization tests to evaluate the adequacy of explanation approaches. We instantiate our analysis on several saliency methods for image classification with neural networks; however, our methodology applies in generality to any explanation approach. Critically, our proposed randomization tests are easy to implement, and can help assess the suitability of an explanation method for a given task at hand.

In a broad experimental sweep, we apply our methodology to numerous existing saliency methods, model architectures, and data sets. To our surprise, *some widely deployed saliency methods are independent of both the data the model was trained on, and the model parameters*. Consequently,

[‡]Work done during the Google AI Residency Program.

²All code to replicate our findings will be available here: <https://goo.gl/hBmhDt>

³We refer here to the broad category of visualization and attribution methods aimed at interpreting trained models. These methods are often used for interpreting deep neural networks particularly on image data.

Apply your knowledge

- Phenomena
- Research Questions
- Hypotheses
- Methods
- Findings

Sanity Checks for Saliency Maps

Julius Adebayo[‡], Justin Gilmer[‡], Michael Muelly[‡], Ian Goodfellow[‡], Moritz Hardt^{‡†}, Been Kim[‡]
juliusad@mit.edu, {gilmer,muelly,goodfellow,mrtz,beenkim}@google.com

[‡]Google Brain

[†]University of California Berkeley

Abstract

Saliency methods have emerged as a popular tool to highlight features in an input deemed relevant for the prediction of a learned model. Several saliency methods have been proposed, often guided by visual appeal on image data. In this work, we propose an actionable methodology to evaluate what kinds of explanations a given method can and cannot provide. We find that reliance, solely, on visual assessment can be misleading. Through extensive experiments we show that some existing saliency methods are independent both of the model and of the data generating process. Consequently, methods that fail the proposed tests are inadequate for tasks that are sensitive to either data or model, such as, finding outliers in the data, explaining the relationship between inputs and outputs that the model learned, and debugging the model. We interpret our findings through an analogy with edge detection in images, a technique that requires neither training data nor model. Theory in the case of a linear model and a single-layer convolutional neural network supports our experimental findings[‡].

1 Introduction

As machine learning grows in complexity and impact, much hope rests on explanation methods as tools to elucidate important aspects of learned models [1, 2]. Explanations could potentially help satisfy regulatory requirements [3], help practitioners debug their model [4, 5], and perhaps, reveal bias or other unintended effects learned by a model [6, 7]. *Saliency methods*[‡] are an increasingly popular class of tools designed to highlight relevant features in an input, typically, an image. Despite much excitement, and significant recent contribution [8–21], the valuable effort of explaining machine learning models faces a methodological challenge: *the difficulty of assessing the scope and quality of model explanations*. A paucity of principled guidelines confound the practitioner when deciding between an abundance of competing methods.

We propose an actionable methodology based on randomization tests to evaluate the adequacy of explanation approaches. We instantiate our analysis on several saliency methods for image classification with neural networks; however, our methodology applies in generality to any explanation approach. Critically, our proposed randomization tests are easy to implement, and can help assess the suitability of an explanation method for a given task at hand.

In a broad experimental sweep, we apply our methodology to numerous existing saliency methods, model architectures, and data sets. To our surprise, *some widely deployed saliency methods are independent of both the data the model was trained on, and the model parameters*. Consequently,

^{*}Work done during the Google AI Residency Program.

[‡]All code to replicate our findings will be available here: <https://goo.gl/hBmhDt>

[‡]We refer here to the broad category of visualization and attribution methods aimed at interpreting trained models. These methods are often used for interpreting deep neural networks particularly on image data.

Apply your knowledge

- Findings
- Phenomena
- Methods
- Hypotheses
- Research Questions