Interpretability for complex models in Machine Learning and NLP

David Alvarez-Melis
(joint work with Tommi Jaakkola)

Guest Lecture, April 18th, 2018

Roadmap

- **Intro: why interpretability?**
- Part 1: Interpretability for black-box structured models
  - Background and Motivation
  - Approach
  - Experiments
  - Summary and extensions
- Part 2: Self-explaining neural networks
  - Motivation
  - Model
  - Results

Intro: Why interpretability?

Interpretability - Why?

Interpretability - Why?

- Lack of transparency limits adoption in decision-critical domains

Interpretability - Why?

- Lack of transparency limits adoption in decision-critical domains
- Algorithmic decision making - models that impact lives should come with explanations!

Interpretability - Why?

- Lack of transparency limits adoption in decision-critical domains

- Algorithmic decision making - models that impact lives should come with explanations!

- EU's GDPR law (2018) guarantees a "right to explanation"

Interpretability - Why?

- Lack of transparency limits adoption in decision-critical domains
- Algorithmic decision making - models that impact lives should come with explanations!
- EU's GDPR law (2018) guarantees a "right to explanation"
- A means to satisfy other criteria (e.g., fairness, privacy, causality [Doshi-Velez and Kim, 2018])

Interpretability - Challenges

Interpretability - Challenges

Interpretability - Challenges

- Emergent sub-field of AI, suffers from:

Interpretability - Challenges

- Emergent sub-field of AI, suffers from:
  - Ill-defined goals

Interpretability - Challenges

- Emergent sub-field of AI, suffers from:
    - Ill-defined goals
    - No universally agreed-upon definition

Interpretability - Challenges

- Emergent sub-field of AI, suffers from:
  - Ill-defined goals
  - No universally agreed-upon definition
  - Few formalisms - existing ones sometimes contradictory

Interpretability - Challenges

- Emergent sub-field of AI, suffers from:
    - Ill-defined goals
    - No universally agreed-upon definition
    - Few formalisms - existing ones sometimes contradictory
    - Under-appreciation among many in the community

CSAIL

A controversial topic

Ian Goodfellow
@goodfellow_ian

Following

One of my main concerns about machine learning interpretability tools is that they will make people think they understand ML when they don't. People seem to think linear models are interpretable, but no one looks at them and has the intuition that they have adversarial examples

Pedro Domingos
@pmddomingos

Following

Given the choice between an AI doctor that's 80% accurate and can explain its diagnoses and one that's 90% accurate but can't, I'd pick the latter.

7:17 PM - 25 Jan 2018

56 Retweets **173** Likes

28 56 173

CSAIL

Demystifying interpretability

Demystifying interpretability

"The objective of interpretability is to

let us understand exactly how a complex model works"

Demystifying interpretability

"The objective of interpretability is to

~~let us understand exactly how a complex model works"~~

Demystifying interpretability

"The objective of interpretability is to

~~let us understand exactly how a complex model works"~~

provide useful abstractions that summarize the model's behavior

Demystifying interpretability

"The objective of interpretability is to

~~let us understand exactly how a complex model works"~~

provide useful abstractions that summarize the model's behavior

Implies a concrete objective, e.g. debugging, auditing, verifying model properties

Demystifying interpretability

"The objective of interpretability is to

~~let us understand exactly how a complex model works"~~

provide useful abstractions that summarize the model's behavior

By definition incomplete

Implies a concrete objective, e.g. debugging, auditing, verifying model properties

Demystifying interpretability

"The objective of interpretability is to

~~let us understand exactly how a complex model works"~~

provide useful abstractions that summarize the model's behavior

By definition incomplete

Implies a concrete objective, e.g. debugging, auditing, verifying model properties

"All models are wrong, some are useful" - George E.P. Box

CSAIL

Demystifying interpretability

"The objective of interpretability is to

~~let us understand exactly how a complex model works"~~

provide useful abstractions that summarize the model's behavior

By definition incomplete

Implies a concrete objective, e.g. debugging, auditing, verifying model properties

All explanations are deficient, some are useful

"All models are wrong, some are useful" - George E.P. Box

CSAIL

Demystifying interpretability

Demystifying interpretability

"All explanations are glorified heatmaps on the input"

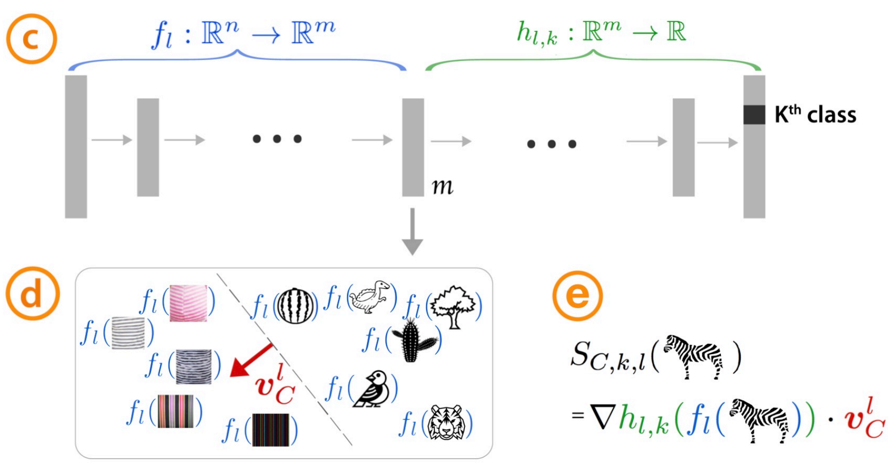
Demystifying interpretability

"~~All explanations are glorified heatmaps on the input"~~

Demystifying interpretability

"~~All explanations are glorified heatmaps on the input~~"

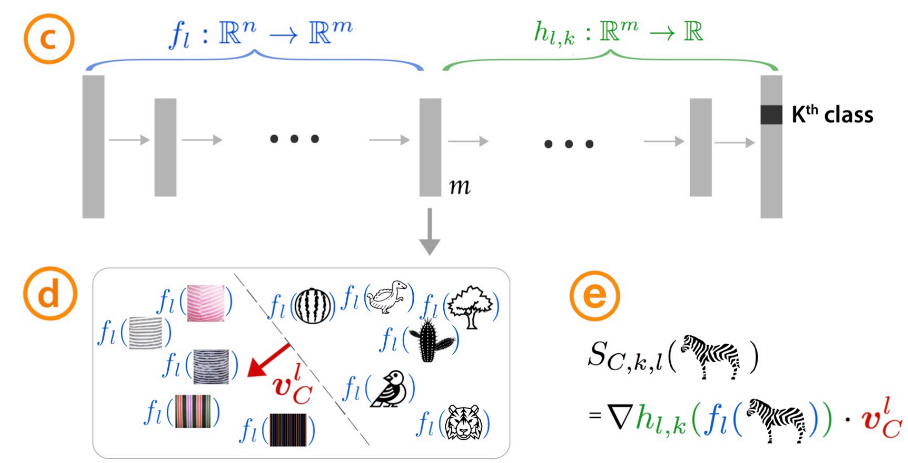
- Higher level concepts (instead if inputs) [Kim et al. 2017]



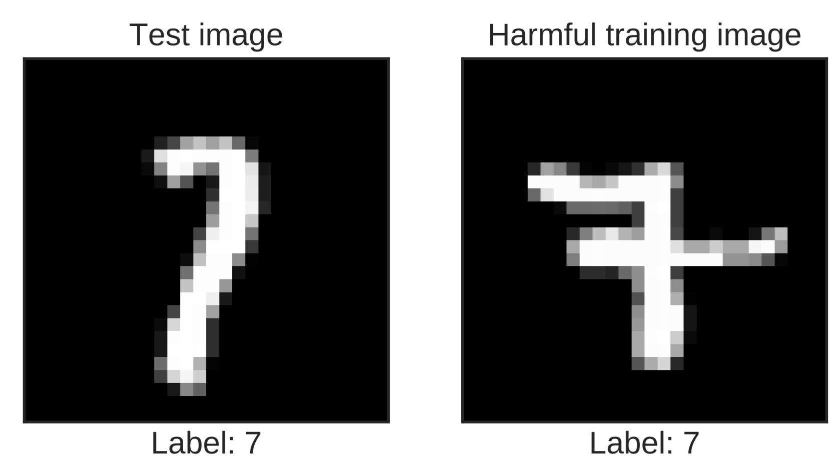
Demystifying interpretability

~~"All explanations are glorified heatmaps on the input"~~

- Higher level concepts (instead if inputs) [Kim et al. 2017]



- Explanations in terms of training data [Koh & Liang, 2017]

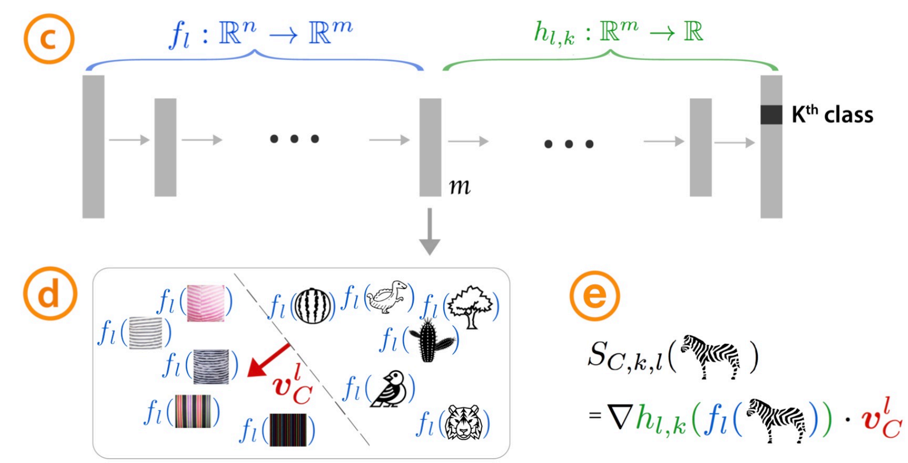


CSAIL

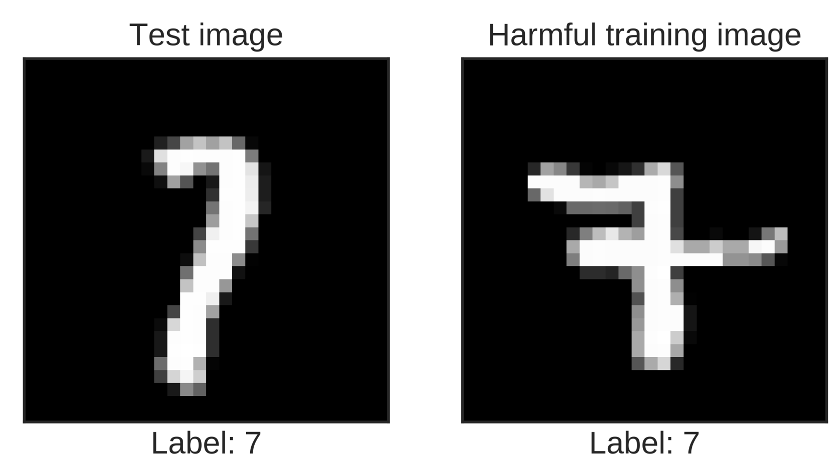
Demystifying interpretability

"~~All explanations are glorified heatmaps on the input~~"

- Higher level concepts (instead if inputs) [Kim et al. 2017]



- Explanations in terms of training data [Koh & Liang, 2017]



- Causal rules (instead of relevance scores)

Demystifying interpretability

Demystifying interpretability

"It's impossible to evaluate interpretability methods"

Demystifying interpretability

~~"It's impossible to evaluate interpretability methods"~~

Demystifying interpretability

~~"It's impossible to evaluate interpretability methods"~~

- Task-driven*:

*(see section on "Taxonomy of Interpretability of Evaluation" in [Doshi-Velez & Kim, 2017] for more details)

Demystifying interpretability

~~"It's impossible to evaluate interpretability methods"~~

- Task-driven*:
  - ✦ Functionally-grounded Evaluation on Proxy Tasks

*(see section on "Taxonomy of Interpretability of Evaluation"
in [Doshi-Velez & Kim, 2017] for more details)

Demystifying interpretability

~~"It's impossible to evaluate interpretability methods"~~

- Task-driven*:
  - ✦ Functionally-grounded Evaluation on Proxy Tasks
  - ✦ Human-grounded evaluation on Simple Tasks

*(see section on "Taxonomy of Interpretability of Evaluation" in [Doshi-Velez & Kim, 2017] for more details)

Demystifying interpretability

~~"It's impossible to evaluate interpretability methods"~~

- Task-driven*:
  - ✦ Functionally-grounded Evaluation on Proxy Tasks
  - ✦ Human-grounded evaluation on Simple Tasks
  - ✦ Application-grounded evaluation on Real Tasks

*(see section on "Taxonomy of Interpretability of Evaluation" in [Doshi-Velez & Kim, 2017] for more details)

Demystifying interpretability

~~"It's impossible to evaluate interpretability methods"~~

- Task-driven*:
  - ✦ Functionally-grounded Evaluation on Proxy Tasks
  - ✦ Human-grounded evaluation on Simple Tasks
  - ✦ Application-grounded evaluation on Real Tasks
- Intrinsic:

*(see section on "Taxonomy of Interpretability of Evaluation" in [Doshi-Velez & Kim, 2017] for more details)

Demystifying interpretability

~~"It's impossible to evaluate interpretability methods"~~

- Task-driven*:
  - ✦ Functionally-grounded Evaluation on Proxy Tasks
  - ✦ Human-grounded evaluation on Simple Tasks
  - ✦ Application-grounded evaluation on Real Tasks
- Intrinsic:
  - ✦ Robustness/stability of explanations

*(see section on "Taxonomy of Interpretability of Evaluation" in [Doshi-Velez & Kim, 2017] for more details)

CSAIL

Demystifying interpretability

~~"It's impossible to evaluate interpretability methods"~~

- Task-driven*:
  - ✦ Functionally-grounded Evaluation on Proxy Tasks
  - ✦ Human-grounded evaluation on Simple Tasks
  - ✦ Application-grounded evaluation on Real Tasks
- Intrinsic:
  - ✦ Robustness/stability of explanations
  - ✦ Consistency with actual prediction

*(see section on "Taxonomy of Interpretability of Evaluation" in [Doshi-Velez & Kim, 2017] for more details)

Demystifying interpretability

~~"It's impossible to evaluate interpretability methods"~~

- Task-driven*:
  - ✦ Functionally-grounded Evaluation on Proxy Tasks
  - ✦ Human-grounded evaluation on Simple Tasks
  - ✦ Application-grounded evaluation on Real Tasks
- Intrinsic:
  - ✦ Robustness/stability of explanations
  - ✦ Consistency with actual prediction
  - ✦ Information-theoretic notions

*(see section on "Taxonomy of Interpretability of Evaluation" in [Doshi-Velez & Kim, 2017] for more details)

Demystifying interpretability

Demystifying interpretability

"Interpretability is always necessary / useful "

Demystifying interpretability

~~"Interpretability is always necessary / useful."~~

Demystifying interpretability

~~"Interpretability is always necessary / useful."~~

Demystifying interpretability

~~"Interpretability is always necessary / useful."~~

[DVK17]: "Need for interpretability stems from an *incompleteness* in the problem formalization"

Demystifying interpretability

~~"Interpretability is always necessary / useful."~~

[DVK17]: "Need for interpretability stems from an *incompleteness* in the problem formalization"

- It's necessary in:

Demystifying interpretability

~~"Interpretability is always necessary / useful."~~

[DVK17]: "Need for interpretability stems from an *incompleteness* in the problem formalization"

- It's necessary in:
    - Decision-critical domains with human intervention (e.g., medical)

Demystifying interpretability

~~"Interpretability is always necessary / useful."~~

[DVK17]: "Need for interpretability stems from an *incompleteness* in the problem formalization"

- It's necessary in:
  - Decision-critical domains with human intervention (e.g., medical)
  - Settings where law protects right to explanation (e.g., legal)

Demystifying interpretability

~~"Interpretability is always necessary / useful."~~

[DVK17]: "Need for interpretability stems from an *incompleteness* in the problem formalization"

- It's necessary in:
  - Decision-critical domains with human intervention (e.g., medical)
  - Settings where law protects right to explanation (e.g., legal)
- Less so for fully automatic systems with no human intervention, not critical domain (e.g. postal code sorting)

Interpretability: two paradigms

Model-based
~ make the model itself interpretable

Prediction-based
~ explain *specific predictions*

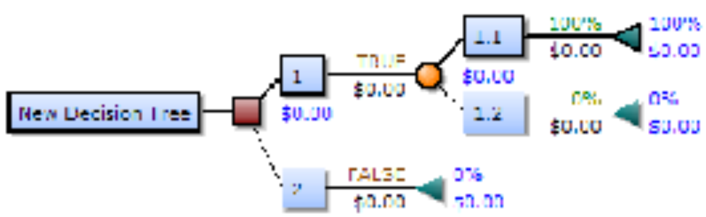
Interpretability: two paradigms

Model-based
~ make the model itself interpretable

Prediction-based
~ explain *specific predictions*

E.g.

- Sparse models, decision trees



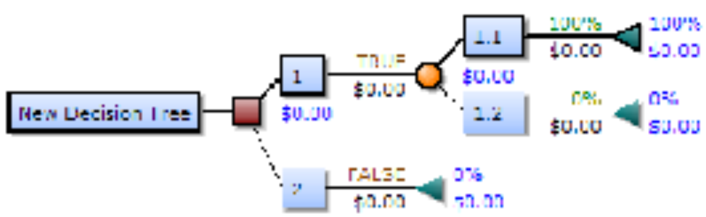
Interpretability: two paradigms

Model-based
~ make the model itself interpretable

Prediction-based
~ explain *specific predictions*

E.g.

- Sparse models, decision trees



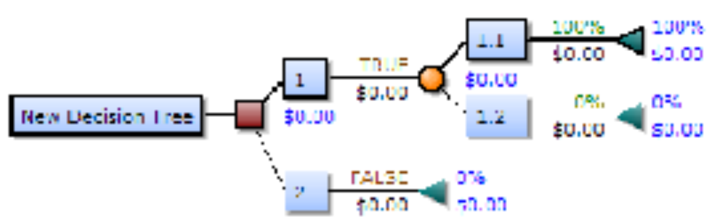
Interpretability: two paradigms

Model-based
~ make the model itself interpretable

Prediction-based
~ explain *specific predictions*

E.g.

- Sparse models, decision trees

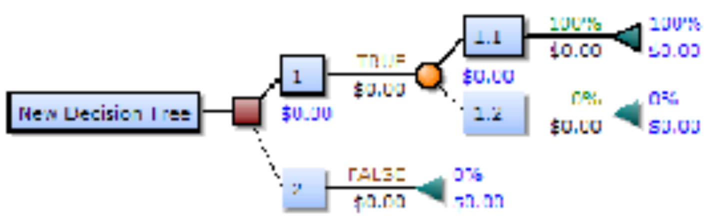
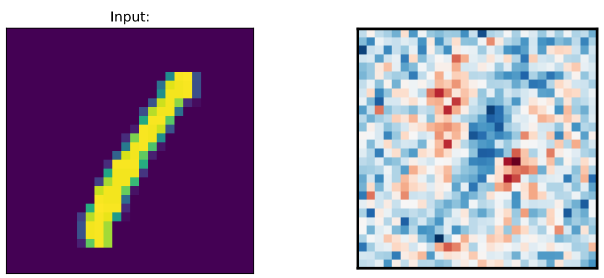


Advantages

✔ Intuitive

✔ No additional estimation needed for interpretability

Interpretability: two paradigms

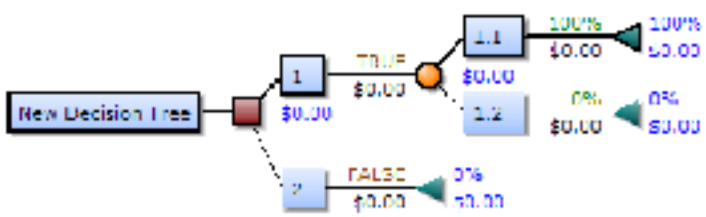
	Model-based ~ make the model itself interpretable	**Prediction-based** ~ explain *specific predictions*
E.g.	• Sparse models, decision trees	• Feature relevance
		
Advantages	✔ Intuitive ✔ No additional estimation needed for interpretability	

Interpretability: two paradigms

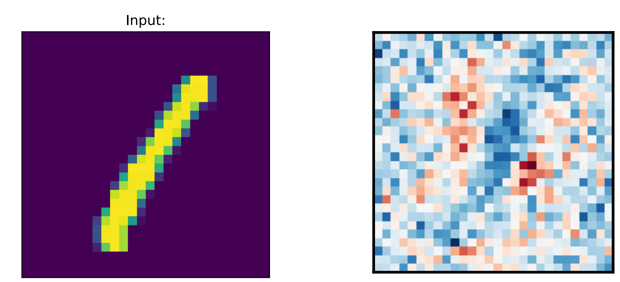
Model-based	**Prediction-based**
~ make the model itself interpretable	~ explain *specific predictions*

E.g.

- Sparse models, decision trees



- Feature relevance



Advantages

Model-based:

- ✔ Intuitive
- ✔ No additional estimation needed for interpretability

Prediction-based:

- ✔ Does not restrict model capacity
- ✔ Can be done for black-box / already-trained models
- ✔ Targeted: why was ***this*** predicted?

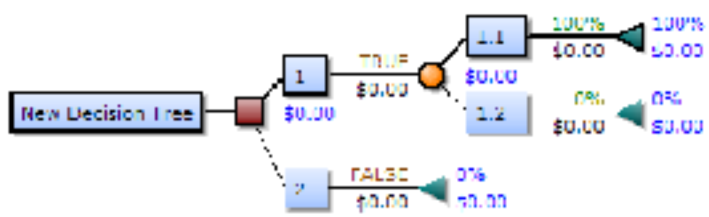
Interpretability: two paradigms

Model-based
~ make the model itself interpretable

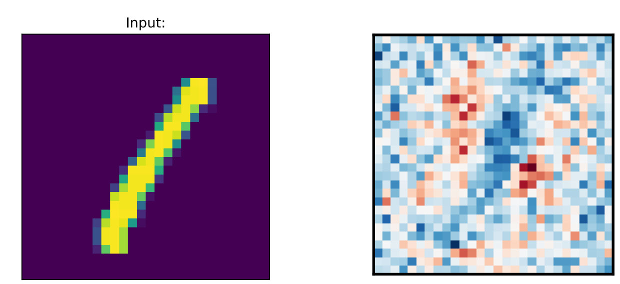
Prediction-based
~ explain *specific predictions*

E.g.

- Sparse models, decision trees



- Feature relevance



Advantages

✔ Intuitive

✔ No additional estimation needed for interpretability

✔ Does not restrict model capacity

✔ Can be done for black-box / already-trained models

✔ Targeted: why was ***this*** predicted?

What is an "explanation" anyways?

What is an "explanation" anyways?

- A justification for a particular prediction

What is an "explanation" anyways?

- A justification for a particular prediction
- Should be:

What is an "explanation" anyways?

- A justification for a particular prediction
- Should be:
    - small

What is an "explanation" anyways?

- A justification for a particular prediction
- Should be:
  - small
  - self-contained

What is an "explanation" anyways?

- A justification for a particular prediction
- Should be:
  - small
  - self-contained
  - sufficient

What is an "explanation" anyways?

- A justification for a particular prediction
- Should be:
    - small
    - self-contained
    - sufficient
- Simplest approach:

What is an "explanation" anyways?

- A justification for a particular prediction
- Should be:
  - small
  - self-contained
  - sufficient
- Simplest approach:

"what parts of the input led to a particular prediction"

Input-based explanations

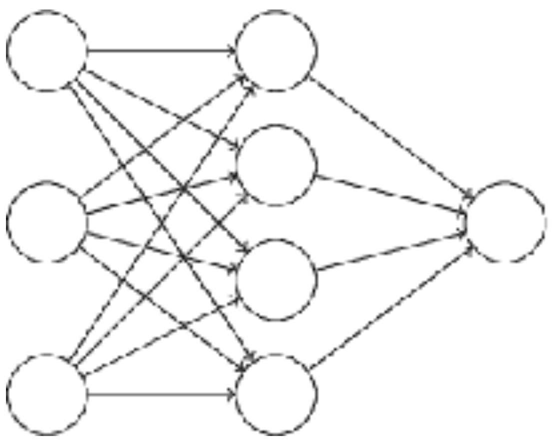
- Example: for image classification

- Example: text-based prediction

[Image Credit: Selvaraju et al.]

Input-based explanations

- Example: for image classification

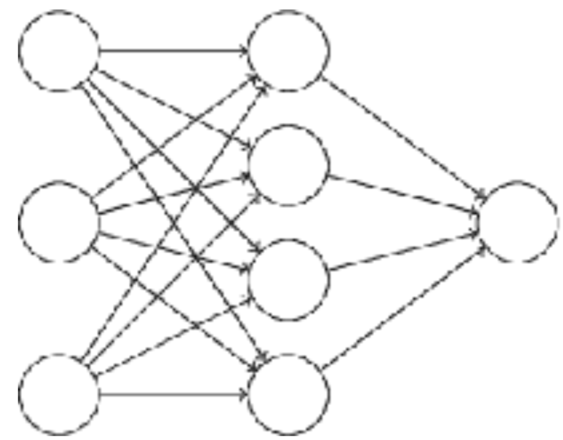


- Example: text-based prediction

[Image Credit: Selvaraju et al.]

Input-based explanations

- Example: for image classification

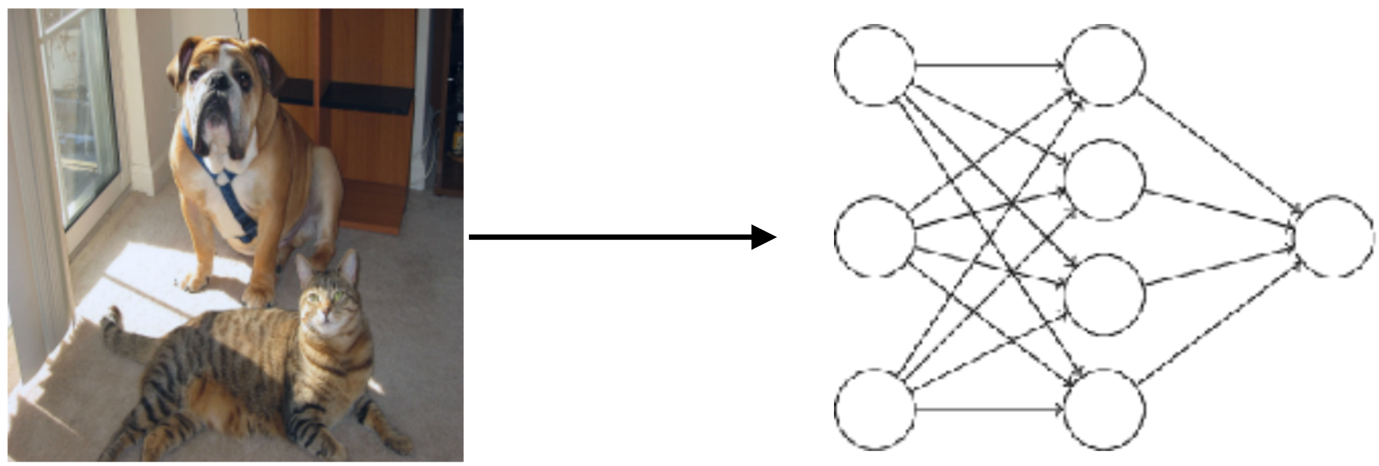


- Example: text-based prediction

[Image Credit: Selvaraju et al.]

Input-based explanations

- Example: for image classification

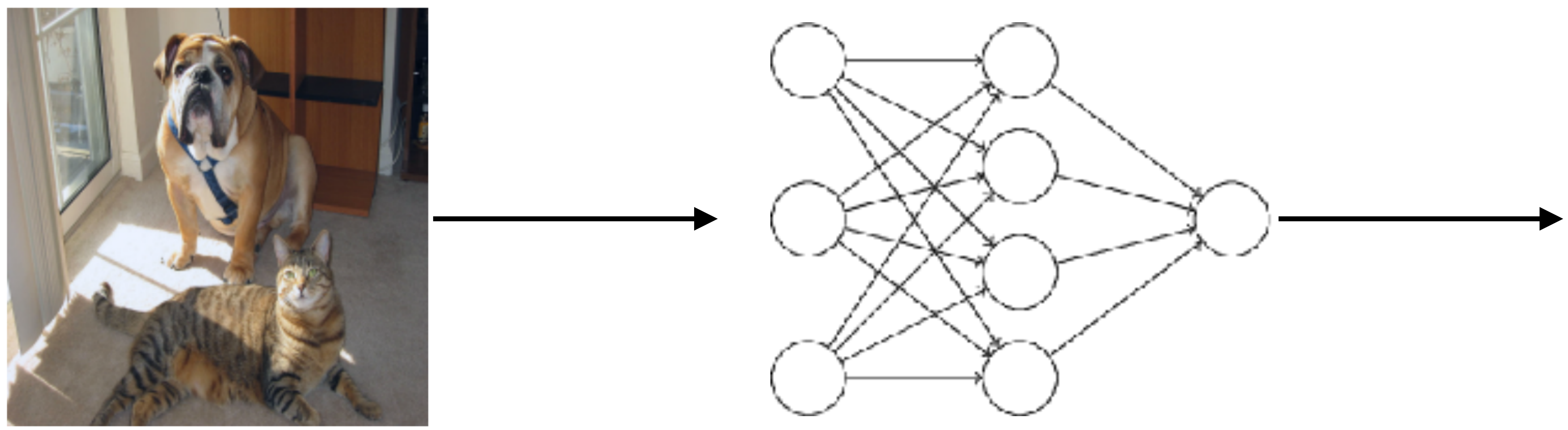


- Example: text-based prediction

[Image Credit: Selvaraju et al.]

Input-based explanations

- Example: for image classification

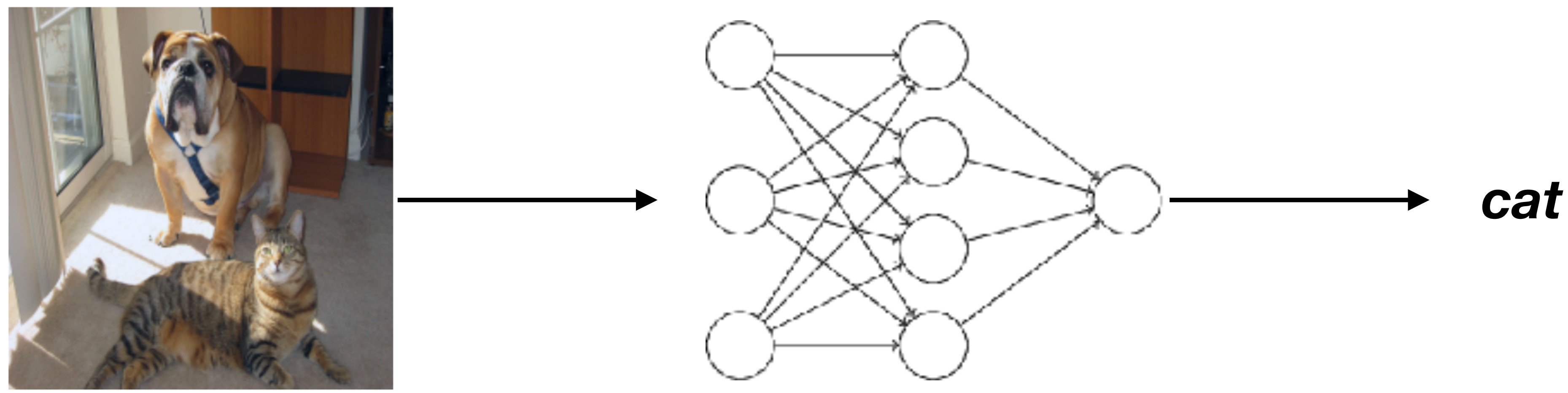


- Example: text-based prediction

[Image Credit: Selvaraju et al.]

Input-based explanations

- Example: for image classification

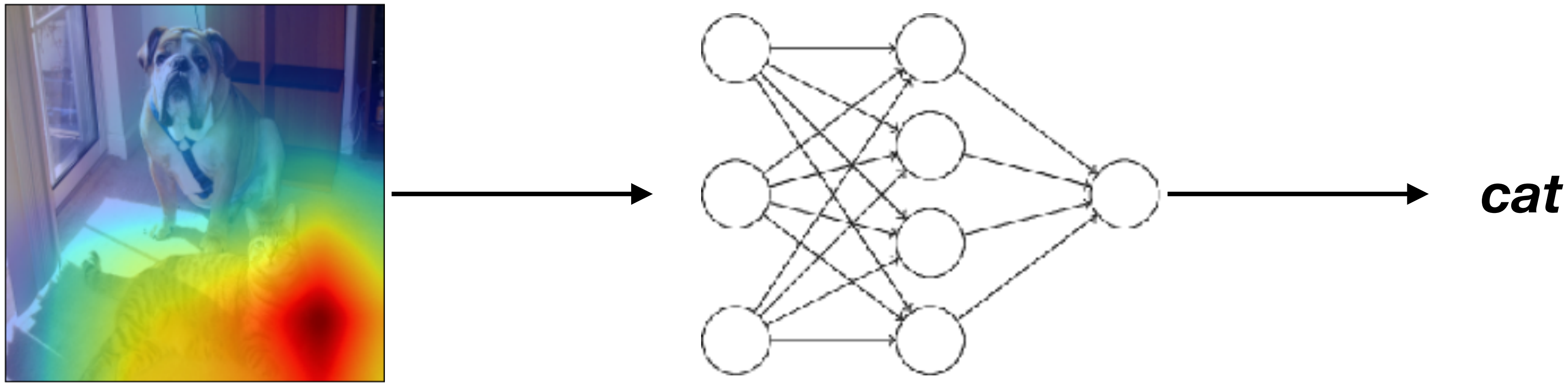


- Example: text-based prediction

[Image Credit: Selvaraju et al.]

Input-based explanations

- Example: for image classification

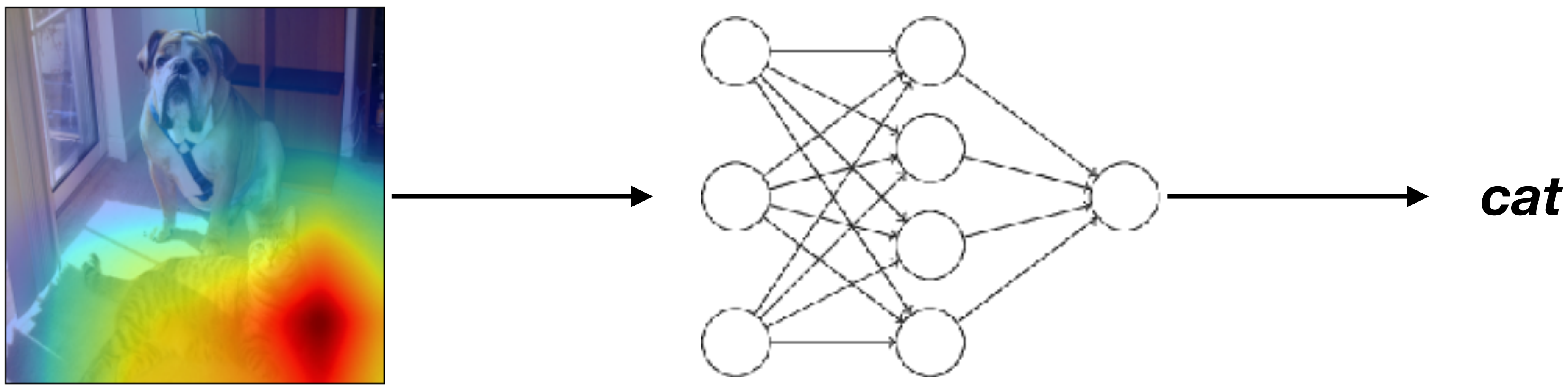


- Example: text-based prediction

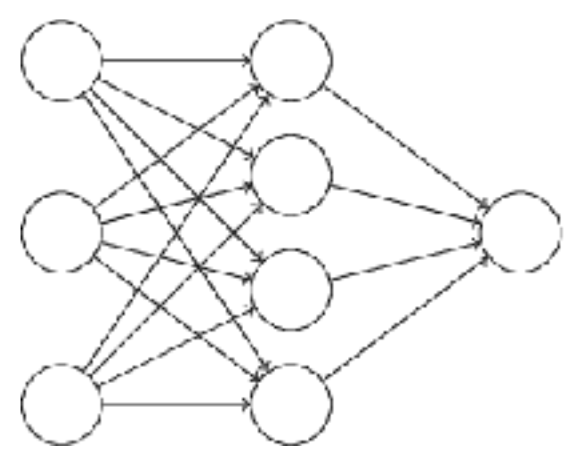
[Image Credit: Selvaraju et al.]

Input-based explanations

- Example: for image classification



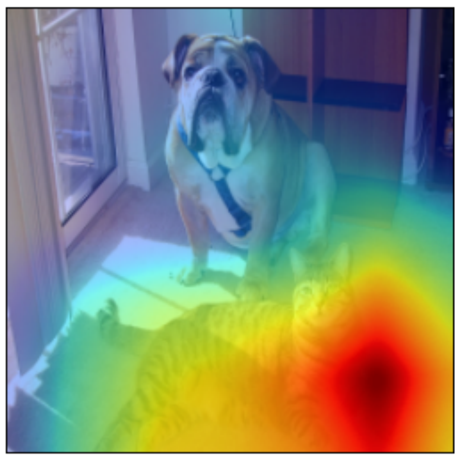
- Example: text-based prediction

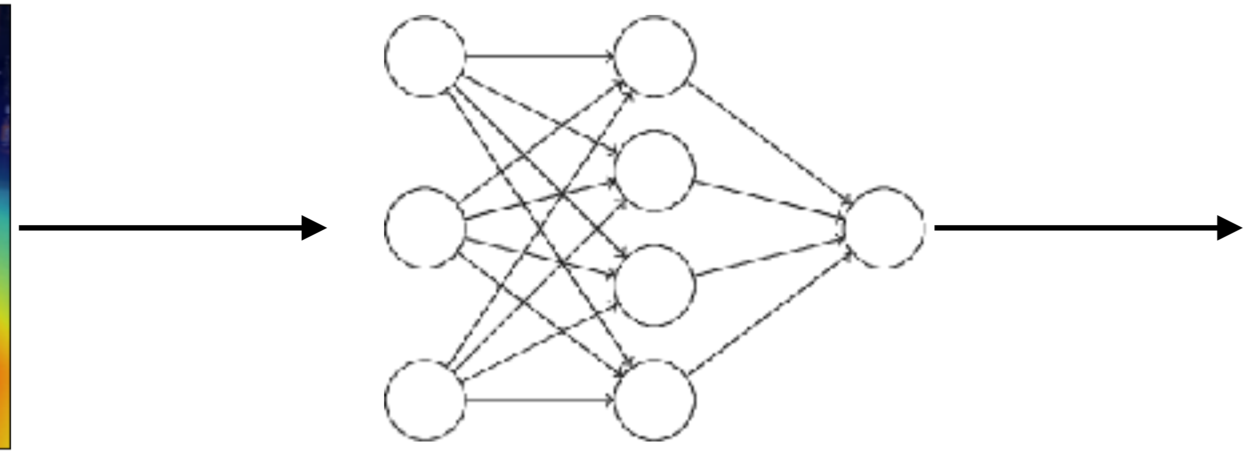


[Image Credit: Selvaraju et al.]

Input-based explanations

- Example: for image classification

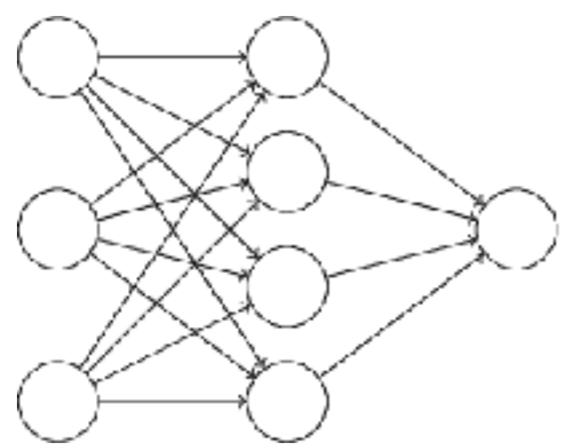




cat

- Example: text-based prediction

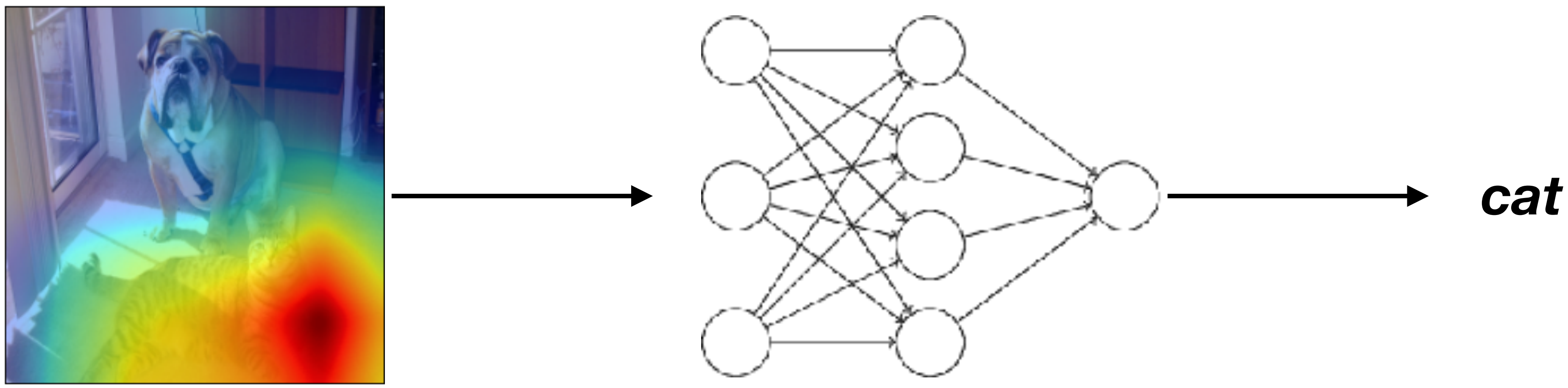
Accession Number <unk> ***Report Status*** Final
Type Surgical Pathology ... ***Pathology Report***:
LEFT BREAST ULTRASOUND GUIDED CORE NEEDLE BIOPSIES ... INVASIVE DUCTAL CARCINOMA poorly differentiateied modified Bloom Richardson grade III III measuring at least 0 7cm in this limited 98% specimen Central hyalinization is present within the tumor mass but no necrosis is noted No lymphovascular invasion is identified No in situ carcinoma is present Special studies were performed at an outside institution with the following results not reviewed ESTROGEN RECEPTOR NEGATIVE PROGESTERONE RECEPTOR NEGATIVE



[Image Credit: Selvaraju et al.]

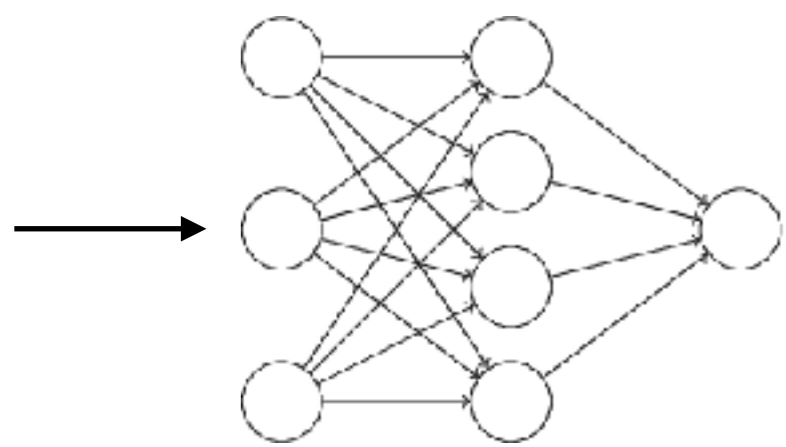
Input-based explanations

- Example: for image classification



- Example: text-based prediction

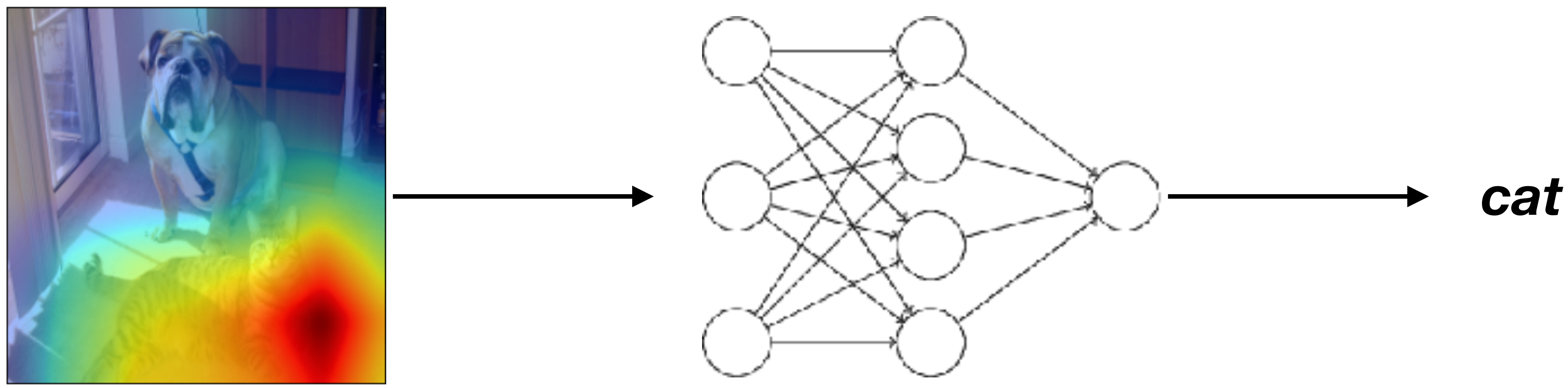
Accession Number <unk> ***Report Status*** Final
Type Surgical Pathology ... ***Pathology Report***:
LEFT BREAST ULTRASOUND GUIDED CORE NEEDLE BIOPSIES ... INVASIVE DUCTAL CARCINOMA poorly differentiateied modified Bloom Richardson grade III III measuring at least 0 7cm in this limited 98% specimen Central hyalinization is present within the tumor mass but no necrosis is noted No lymphovascular invasion is identified No in situ carcinoma is present Special studies were performed at an outside institution with the following results not reviewed ESTROGEN RECEPTOR NEGATIVE PROGESTERONE RECEPTOR NEGATIVE



[Image Credit: Selvaraju et al.]

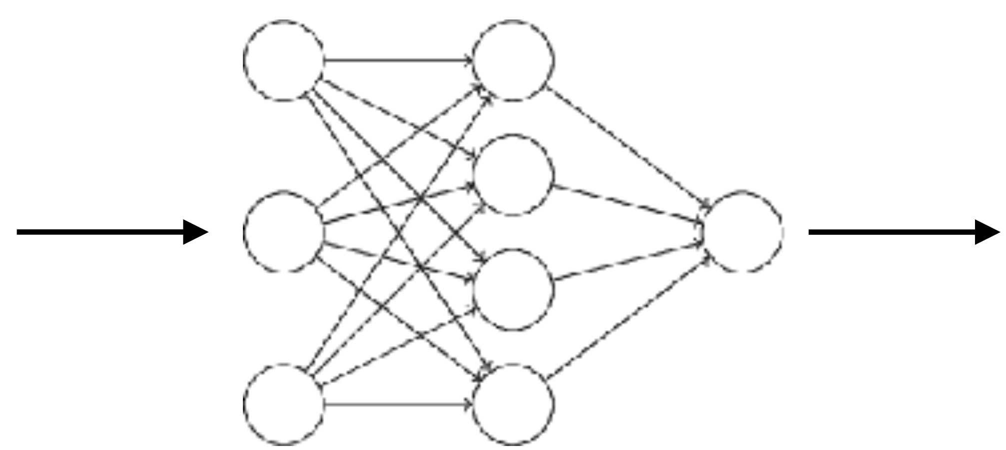
Input-based explanations

- Example: for image classification



- Example: text-based prediction

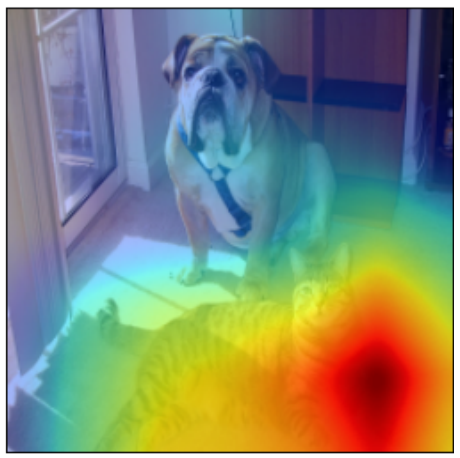
Accession Number <unk> ***Report Status*** Final
Type Surgical Pathology ... ***Pathology Report***:
LEFT BREAST ULTRASOUND GUIDED CORE NEEDLE BIOPSIES ... INVASIVE DUCTAL CARCINOMA poorly differentiateied modified Bloom Richardson grade III III measuring at least 0 7cm in this limited 98% specimen Central hyalinization is present within the tumor mass but no necrosis is noted No lymphovascular invasion is identified No in situ carcinoma is present Special studies were performed at an outside institution with the following results not reviewed ESTROGEN RECEPTOR NEGATIVE PROGESTERONE RECEPTOR NEGATIVE

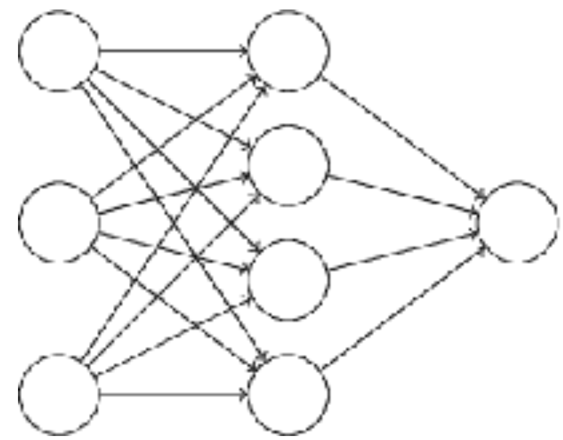


[Image Credit: Selvaraju et al.]

Input-based explanations

- Example: for image classification

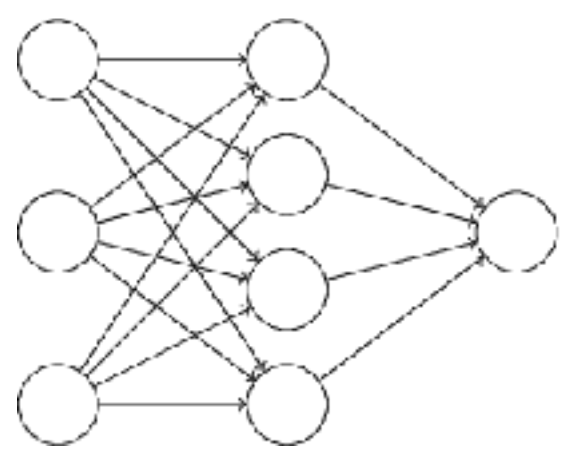




cat

- Example: text-based prediction

Accession Number <unk> ***Report Status*** Final
Type Surgical Pathology ... ***Pathology Report***:
LEFT BREAST ULTRASOUND GUIDED CORE NEEDLE BIOPSIES ... INVASIVE DUCTAL CARCINOMA poorly differentiateied modified Bloom Richardson grade III III measuring at least 0 7cm in this limited 98% specimen Central hyalinization is present within the tumor mass but no necrosis is noted No lymphovascular invasion is identified No in situ carcinoma is present Special studies were performed at an outside institution with the following results not reviewed ESTROGEN RECEPTOR NEGATIVE PROGESTERONE RECEPTOR NEGATIVE

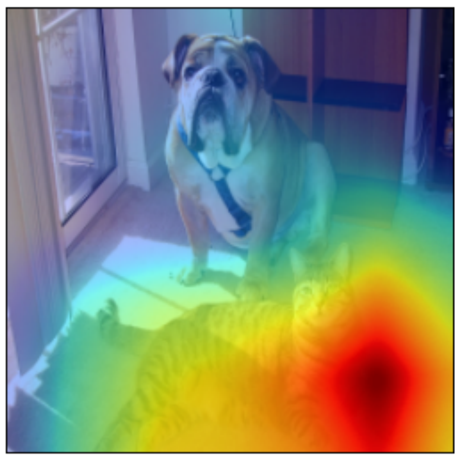


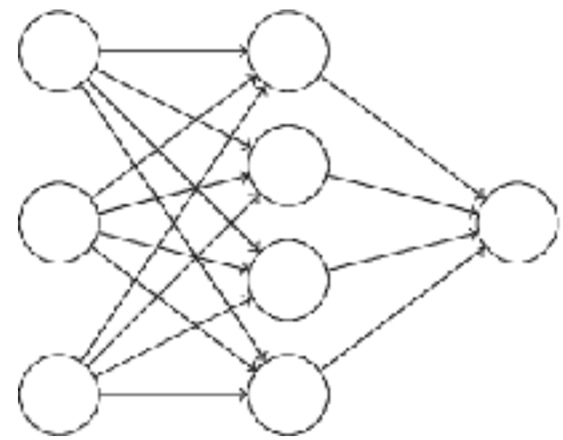
$p(\mathrm{IDC}) = 0.9$

[Image Credit: Selvaraju et al.]

Input-based explanations

- Example: for image classification

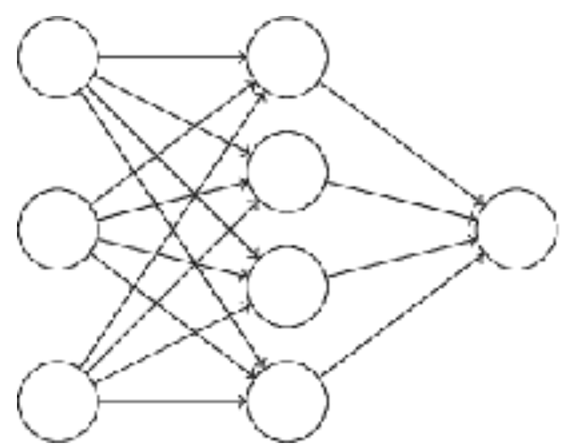




cat

- Example: text-based prediction

Accession Number <unk> ***Report Status*** Final
Type Surgical Pathology ... ***Pathology Report***:
LEFT BREAST ULTRASOUND GUIDED CORE NEEDLE BIOPSIES ... **INVASIVE DUCTAL CARCINOMA poorly differentiateied modified Bloom** Richardson grade III III measuring at least 0 7cm in this limited 98% specimen Central hyalinization is present within the tumor mass but no necrosis is noted No lymphovascular invasion is identified No in situ carcinoma is present Special studies were performed at an outside institution with the following results not reviewed ESTROGEN RECEPTOR NEGATIVE PROGESTERONE RECEPTOR NEGATIVE



$p(\mathrm{IDC}) = 0.9$

[Image Credit: Selvaraju et al.]

Part I: Interpretability for black-box sequence-to-sequence models

[A-M & Jaakkola, EMNLP 2017]

Driving example: Machine Translation

Input

"*Mary did not slap the green witch*"

Output

"*Mary hat die grüne Hexe nicht geschlagen*"

Driving example: Machine Translation

Input

"Mary did not slap the green witch"

→

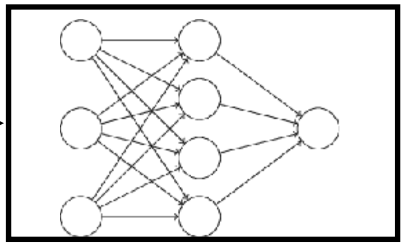
Output

"Mary hat die grüne Hexe nicht geschlagen"

Driving example: Machine Translation

Input

"Mary did not slap the green witch"



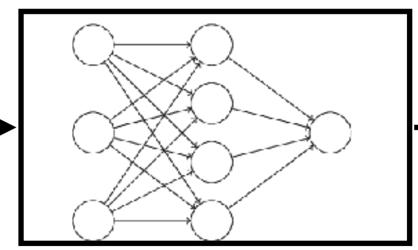
Output

"Mary hat die grüne Hexe nicht geschlagen"

Driving example: Machine Translation

Input

"*Mary did not slap the green witch*"



Output

"*Mary hat die grüne Hexe nicht geschlagen*"

Driving example: Machine Translation

Input

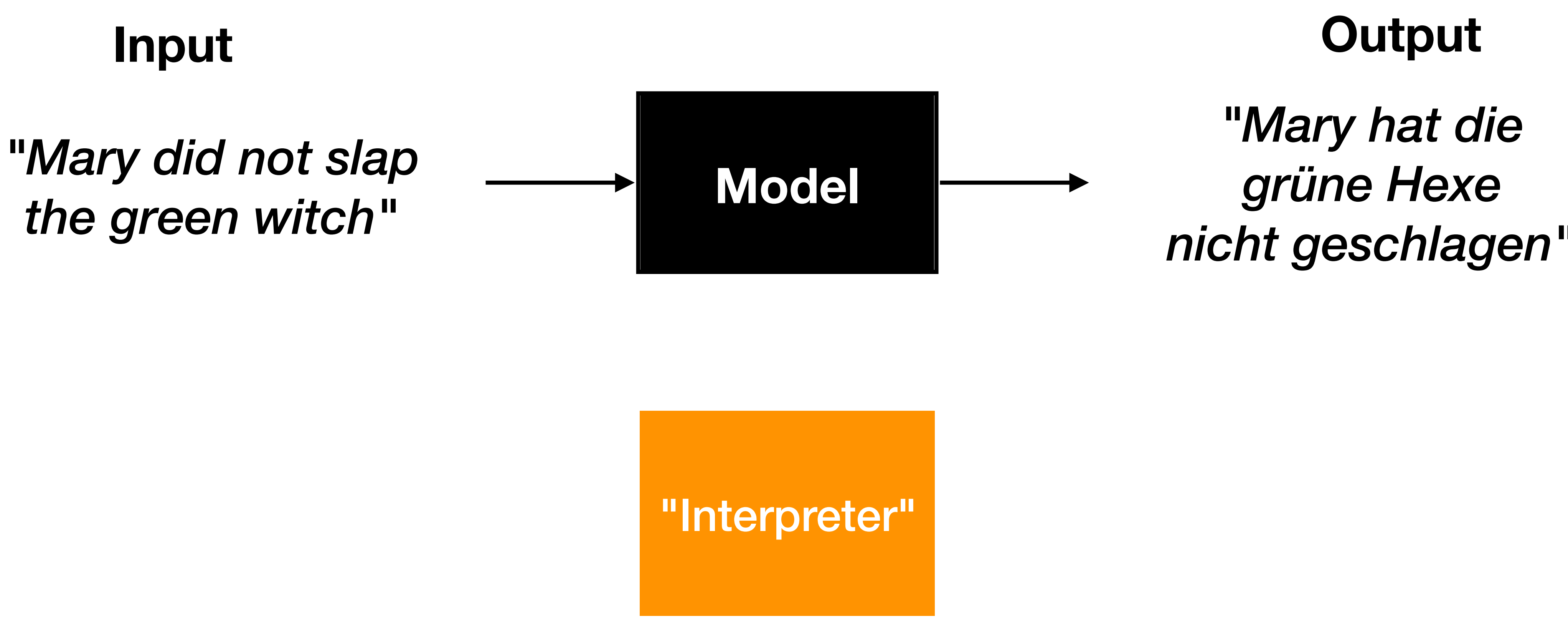
"Mary did not slap the green witch"

Model

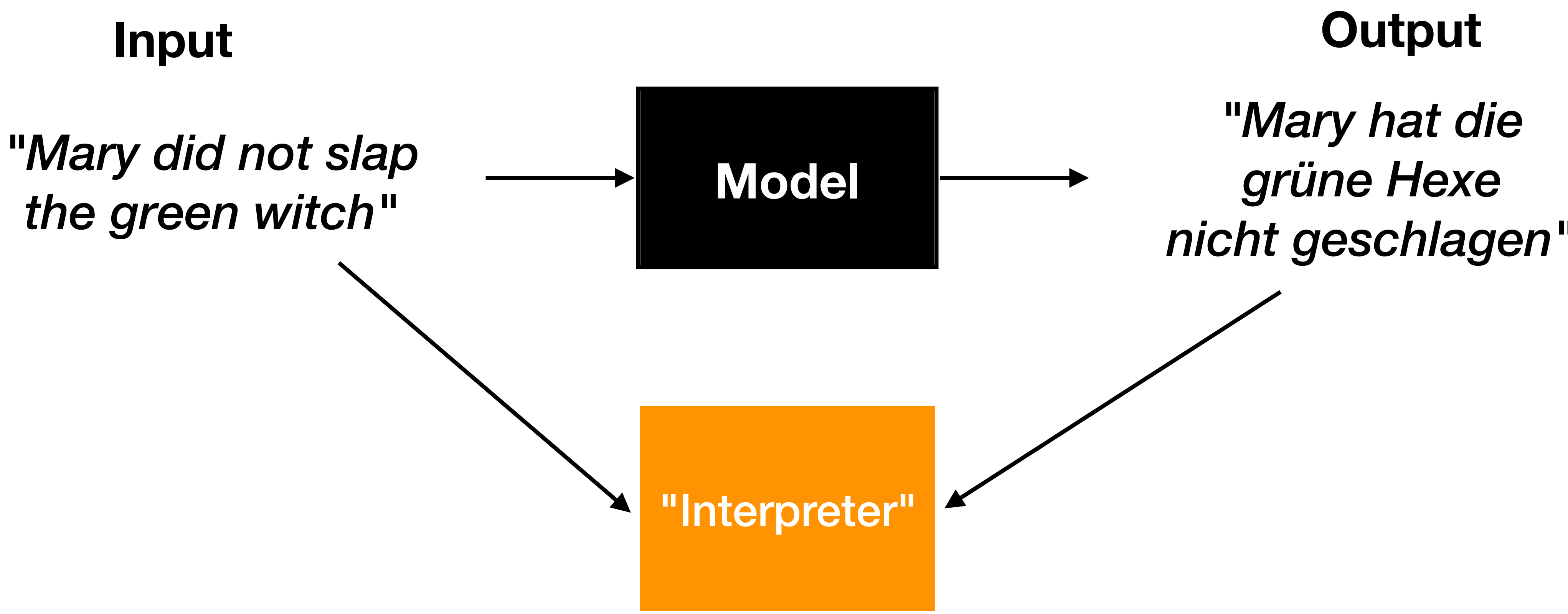
Output

"Mary hat die grüne Hexe nicht geschlagen"

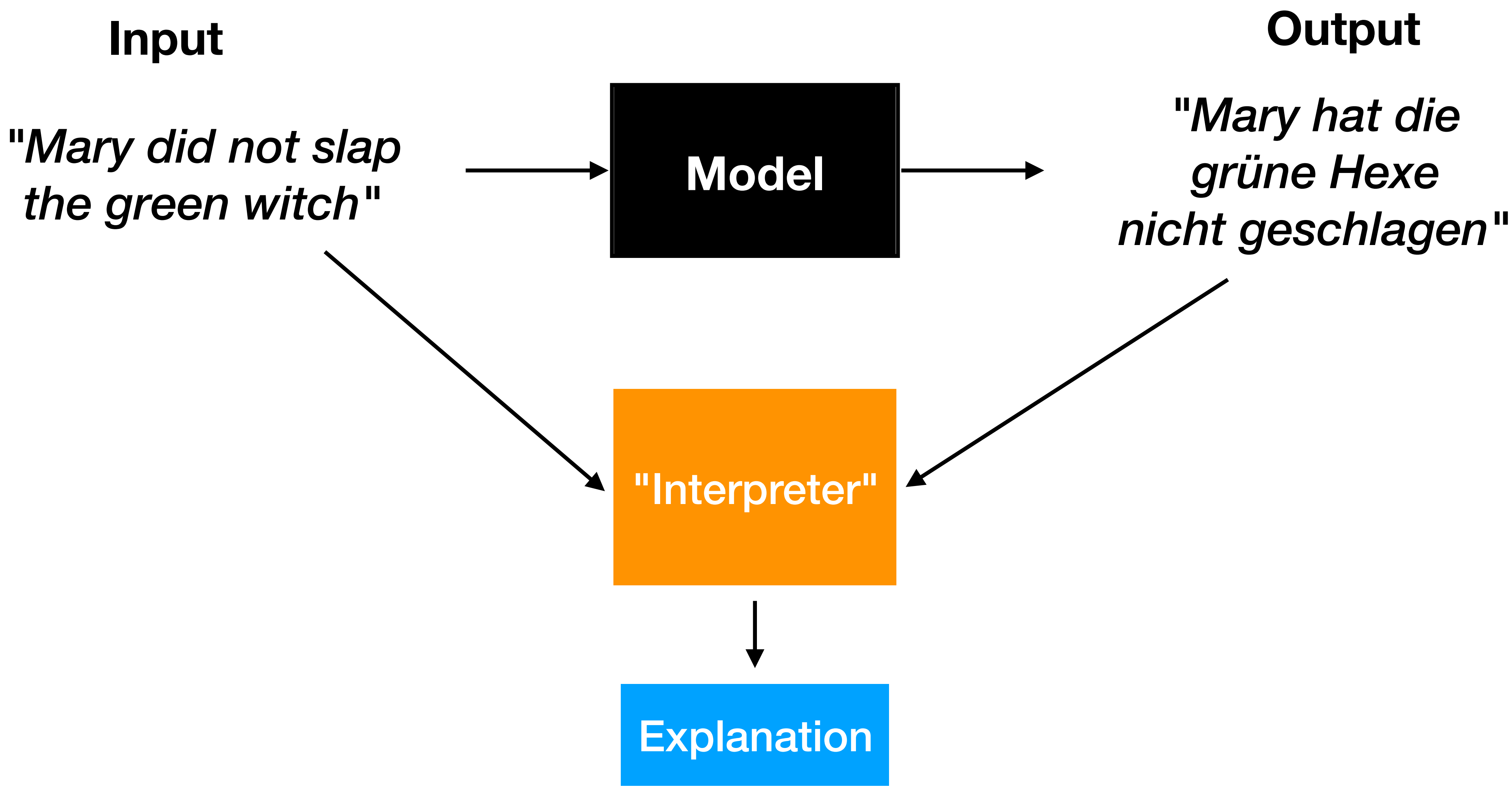
Driving example: Machine Translation



Driving example: Machine Translation



Driving example: Machine Translation



Motivation

Motivation

- SOTA structured prediction methods in NLP tasks are essentially black-boxes

Motivation

- SOTA structured prediction methods in NLP tasks are essentially black-boxes
- Most interpretability work focuses on image classification

Motivation

- SOTA structured prediction methods in NLP tasks are essentially black-boxes
- Most interpretability work focuses on image classification
- Concrete uses of interpretability in NLP:

Motivation

- SOTA structured prediction methods in NLP tasks are essentially black-boxes

- Most interpretability work focuses on image classification

- Concrete uses of interpretability in NLP:

    - Error analysis + model refinement

Motivation

- SOTA structured prediction methods in NLP tasks are essentially black-boxes
- Most interpretability work focuses on image classification
- Concrete uses of interpretability in NLP:
  - Error analysis + model refinement
  - Diagnose undesired behaviors (biases, etc.)

Motivation

- SOTA structured prediction methods in NLP tasks are essentially black-boxes
- Most interpretability work focuses on image classification
- Concrete uses of interpretability in NLP:
  - Error analysis + model refinement
  - Diagnose undesired behaviors (biases, etc.)
  - Trust: "why did you say that"

Motivation

Motivation

- Most methods assume a "simple" (scalar/categorical) output

Motivation

- Most methods assume a "simple" (scalar/categorical) output
- What if inputs/outputs are structured (sentences, graphs)?

Motivation

- Most methods assume a "simple" (scalar/categorical) output
- What if inputs/outputs are structured (sentences, graphs)?
- What if we don't have access to the model?

Motivation

- Most methods assume a "simple" (scalar/categorical) output
- What if inputs/outputs are structured (sentences, graphs)?
- What if we don't have access to the model?
- Can we avoid additional computation?

Related Work

Related Work

- Various work spanning various fields on "interpretability"

Related Work

- Various work spanning various fields on "interpretability"
- Explanations through gradients [Bach et al., 2015; Selvaraju et al., 2017]:

Related Work

- Various work spanning various fields on "interpretability"
- Explanations through gradients [Bach et al., 2015; Selvaraju et al., 2017]:
  - Not Black-Box, expensive computation, no structured output 🙁

CSAIL

Related Work

- Various work spanning various fields on "interpretability"
- Explanations through gradients [Bach et al., 2015; Selvaraju et al., 2017]:
    - Not Black-Box, expensive computation, no structured output 🙁
- Learning Rationales [Lei et al., 2016]:

Related Work

- Various work spanning various fields on "interpretability"
- Explanations through gradients [Bach et al., 2015; Selvaraju et al., 2017]:
  - Not Black-Box, expensive computation, no structured output 🙁
- Learning Rationales [Lei et al., 2016]:
  - Not Black-Box 🙁, no structured output 🙁

Related Work

- Various work spanning various fields on "interpretability"
- Explanations through gradients [Bach et al., 2015; Selvaraju et al., 2017]:
    - Not Black-Box, expensive computation, no structured output 🙁
- Learning Rationales [Lei et al., 2016]:
    - Not Black-Box 🙁, no structured output 🙁
- LIME [Ribeiro et al, 2016]: locally-faithful interpretable models

Related Work

- Various work spanning various fields on "interpretability"
- Explanations through gradients [Bach et al., 2015; Selvaraju et al., 2017]:
    - Not Black-Box, expensive computation, no structured output 🙁
- Learning Rationales [Lei et al., 2016]:
    - Not Black-Box 🙁, no structured output 🙁
- LIME [Ribeiro et al, 2016]: locally-faithful interpretable models
    - Black-box 😀, no structured input nor output 🙁

Related Work

- Various work spanning various fields on "interpretability"
- Explanations through gradients [Bach et al., 2015; Selvaraju et al., 2017]:
  - Not Black-Box, expensive computation, no structured output 🙁
- Learning Rationales [Lei et al., 2016]:
  - Not Black-Box 🙁, no structured output 🙁
- LIME [Ribeiro et al, 2016]: locally-faithful interpretable models
  - Black-box 😀, no structured input nor output 🙁

Setting: Black Box Interpretability

[Image Credit: Ribeiro et al.]

Setting: Black Box Interpretability

- No information about the model. How do we explain?

[Image Credit: Ribeiro et al.]

Setting: Black Box Interpretability

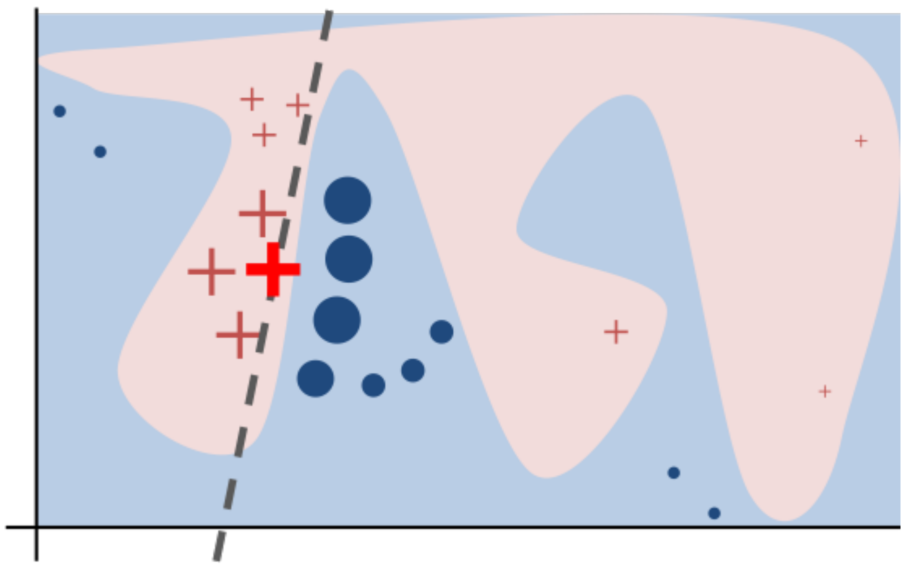
- No information about the model. How do we explain?

- LIME [Ribeiro et al. 2016]: Characterize model locally around prediction by perturbing input + querying model

[Image Credit: Ribeiro et al.]

Setting: Black Box Interpretability

- No information about the model. How do we explain?
- LIME [Ribeiro et al. 2016]: Characterize model locally around prediction by perturbing input + querying model

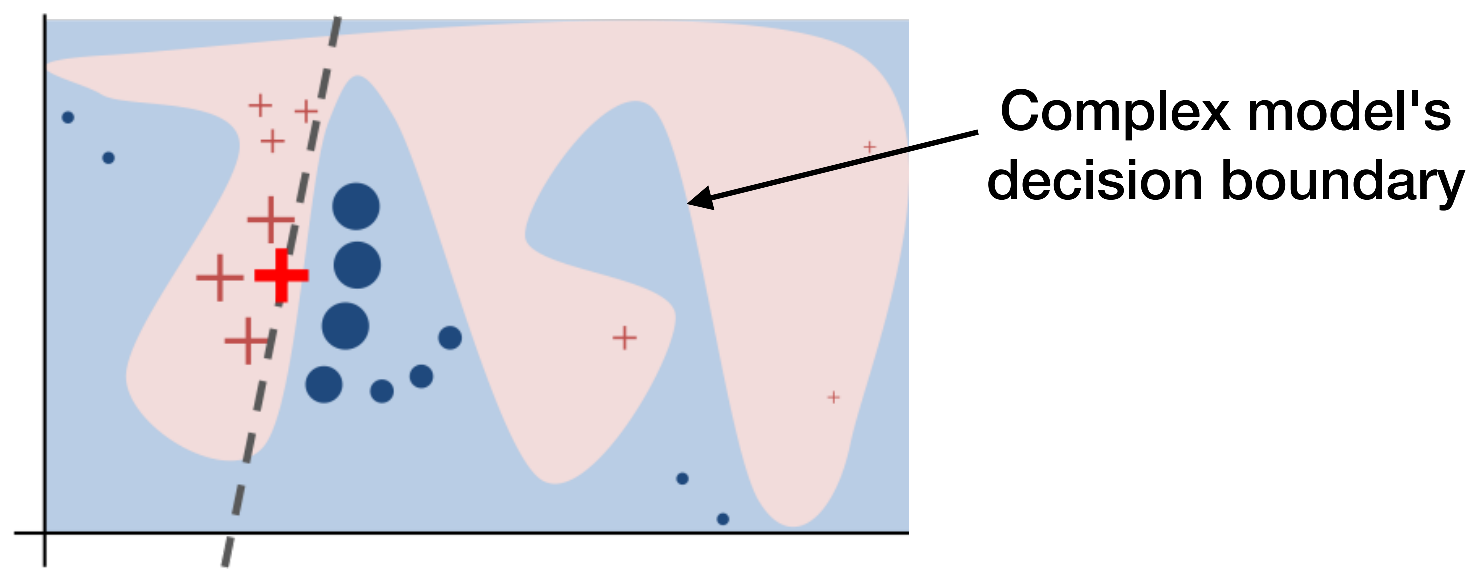


[Image Credit: Ribeiro et al.]

Setting: Black Box Interpretability

- No information about the model. How do we explain?

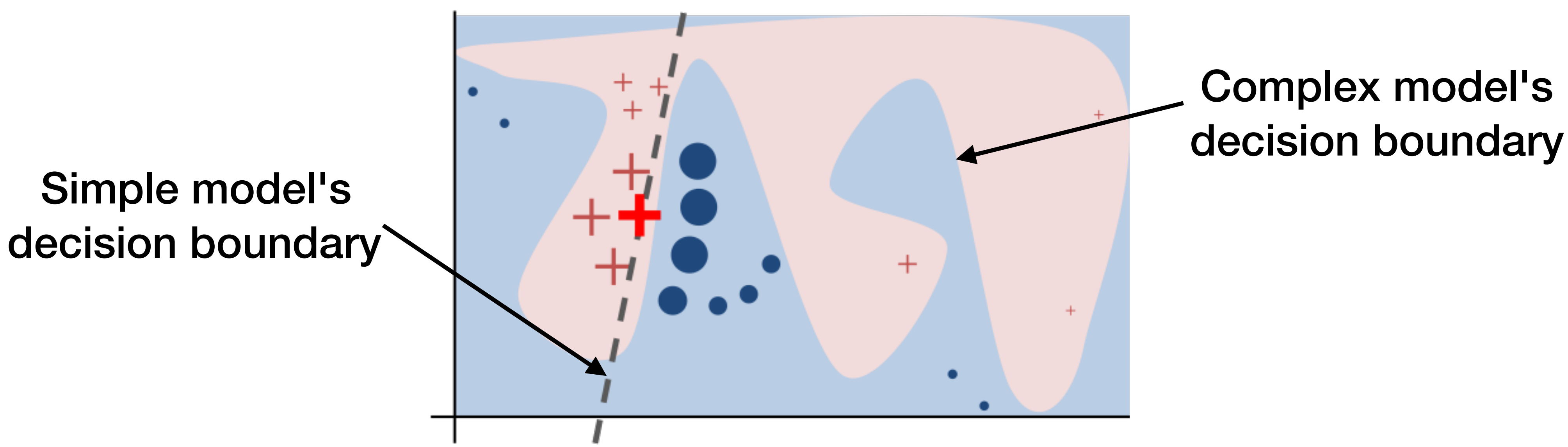
- LIME [Ribeiro et al. 2016]: Characterize model locally around prediction by perturbing input + querying model



[Image Credit: Ribeiro et al.]

Setting: Black Box Interpretability

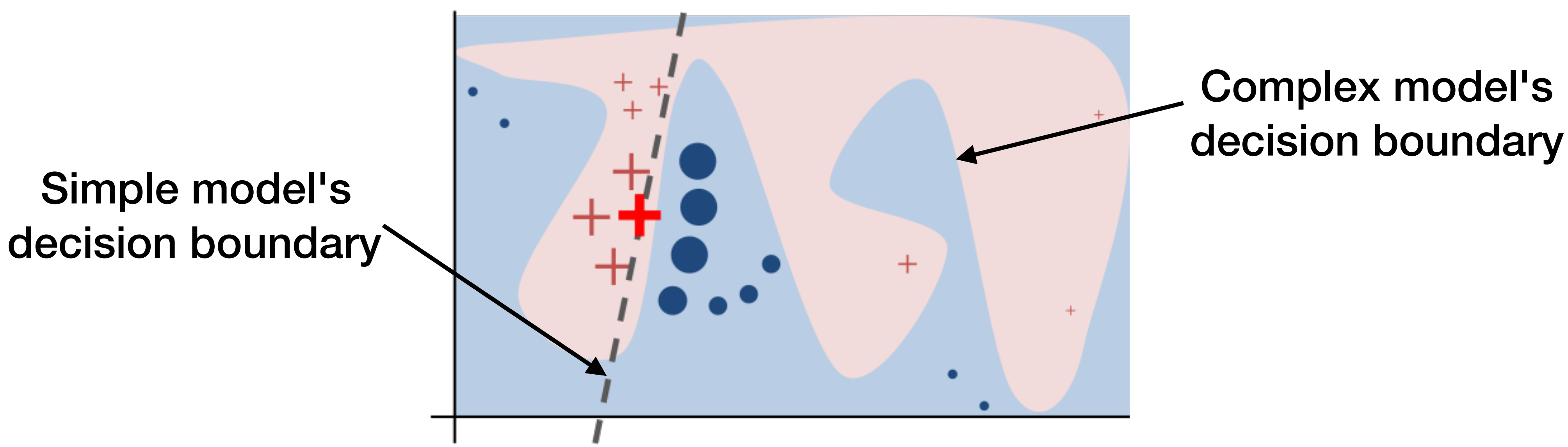
- No information about the model. How do we explain?
- LIME [Ribeiro et al. 2016]: Characterize model locally around prediction by perturbing input + querying model



[Image Credit: Ribeiro et al.]

Setting: Black Box Interpretability

- No information about the model. How do we explain?
- LIME [Ribeiro et al. 2016]: Characterize model locally around prediction by perturbing input + querying model

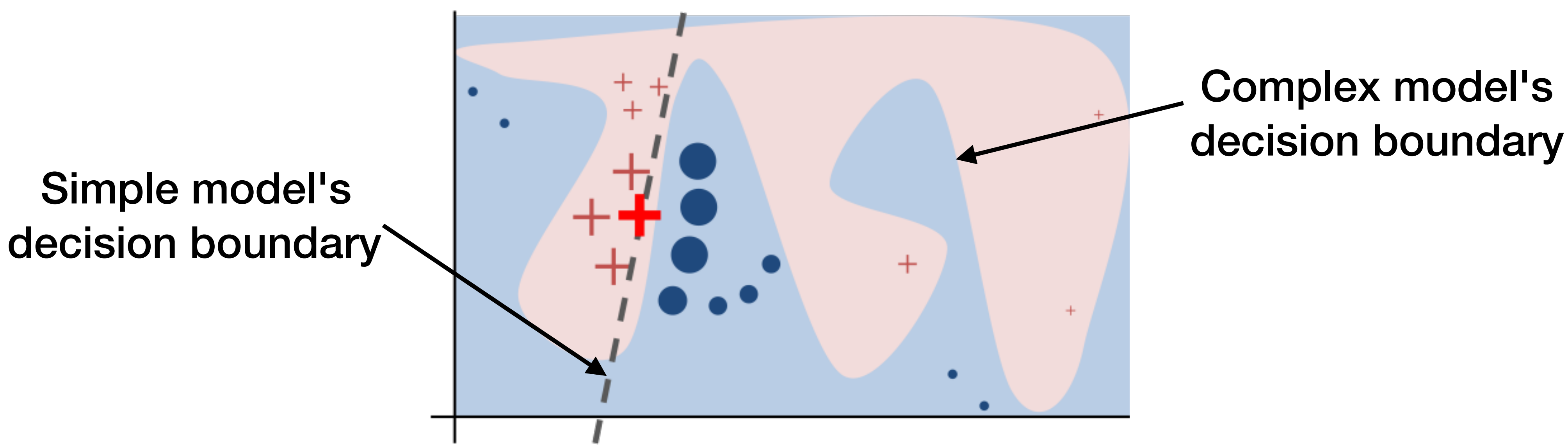


- Assumes input is continuous, output is a a single value.

[Image Credit: Ribeiro et al.]

Setting: Black Box Interpretability

- No information about the model. How do we explain?
- LIME [Ribeiro et al. 2016]: Characterize model locally around prediction by perturbing input + querying model



- Assumes input is continuous, output is a a single value.
- Can we extend this to structured data?

[Image Credit: Ribeiro et al.]

Explanations of structured objects

Explanations of structured objects

- Structured predictions vary in size and complexity

Explanations of structured objects

- Structured predictions vary in size and complexity
- What parts of the input/output to explain?

Explanations of structured objects

- Structured predictions vary in size and complexity
- What parts of the input/output to explain?
- How to keep explanations interpretable regardless of input/output size?

Explanations of structured objects

- Structured predictions vary in size and complexity
- What parts of the input/output to explain?
- How to keep explanations interpretable regardless of input/output size?
- What does "local" mean for a structured input?

Setting

- Black-box: $F : \mathcal{X} \to \mathcal{Y}$

- Elements $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$ admit feature-set representation

$$\mathbf{x} = \{x_1, x_2, \ldots, x_n\}, \quad \mathbf{y} = \{y_1, y_2, \ldots, y_m\}$$

- Goal: explain output $\mathbf{y}$ in terms of input $\mathbf{x}$

- Requirements: locally faithful, model agnostic

Explaining with graphs

Explaining with graphs

- Weighted bipartite graph summarizes local behavior of F

Explaining with graphs

- Weighted bipartite graph summarizes local behavior of F

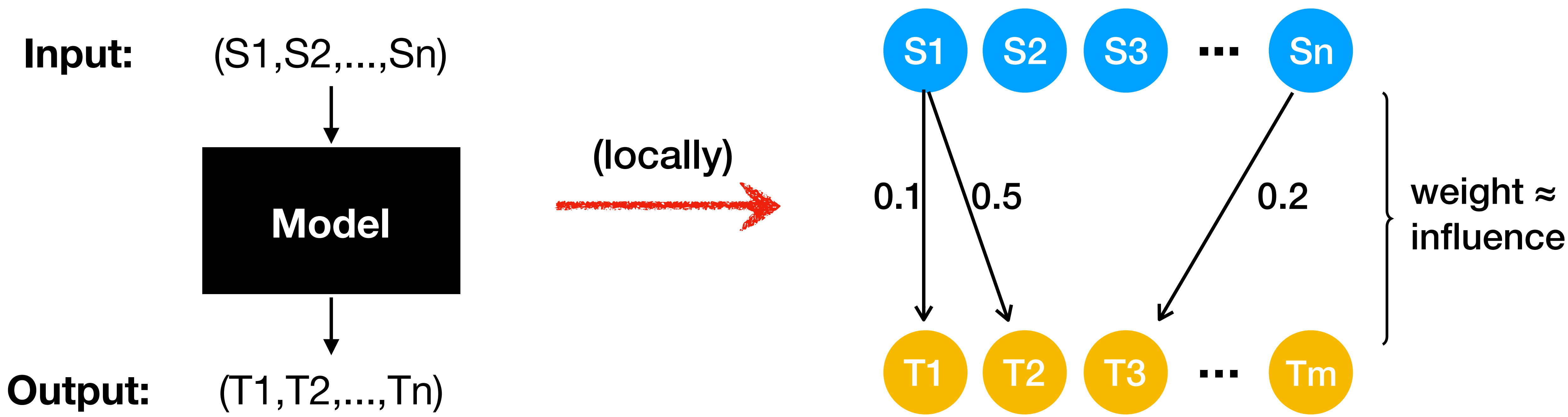
Input: (S1,S2,...,Sn)

Model

Output: (T1,T2,...,Tn)

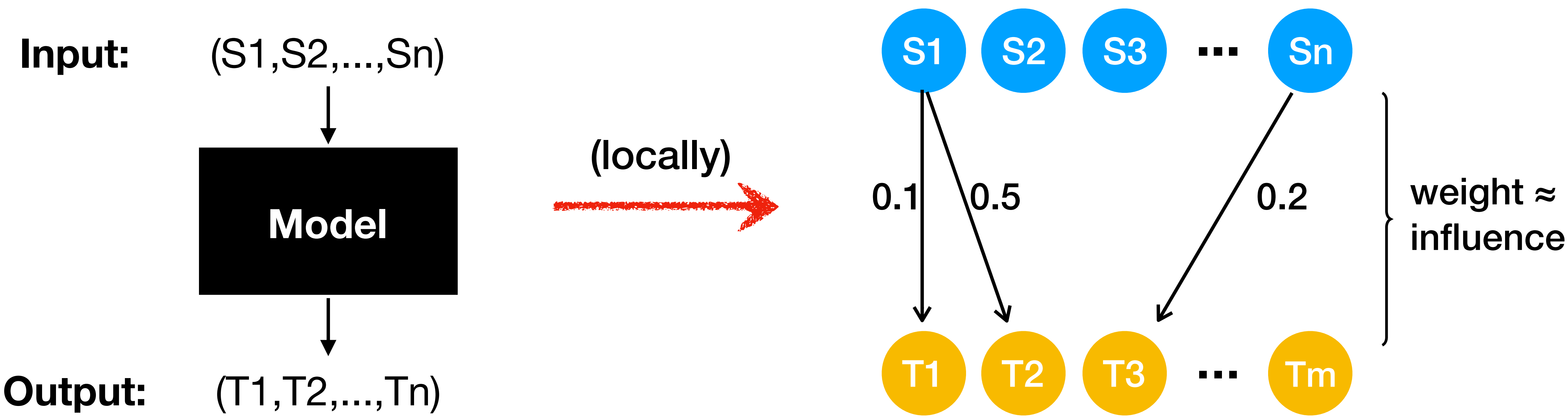
Explaining with graphs

- Weighted bipartite graph summarizes local behavior of F



Explaining with graphs

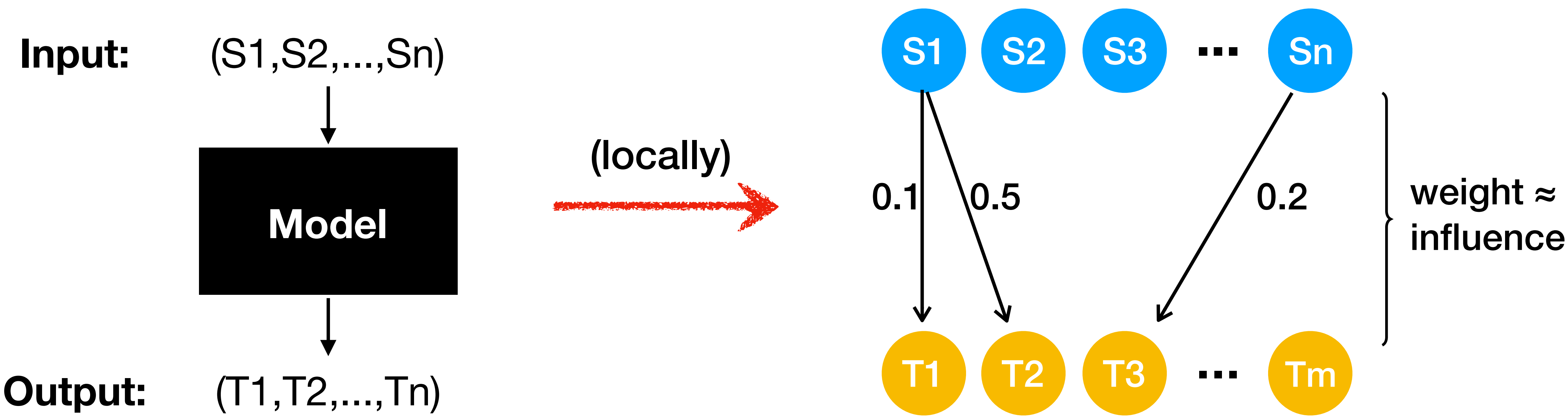
- Weighted bipartite graph summarizes local behavior of F



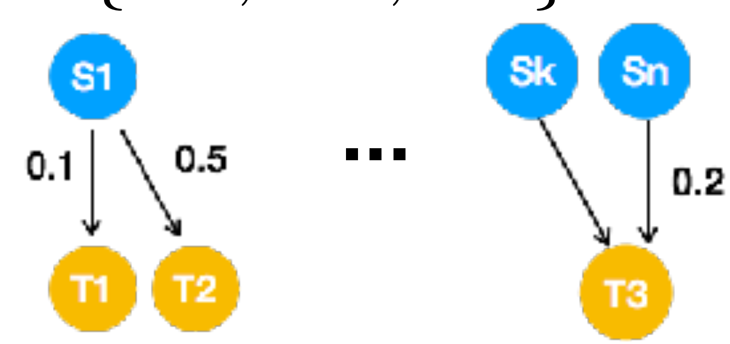
- Explanation: $E_{x \to y} = \{G^1, \ldots, G^k\}$

Explaining with graphs

- Weighted bipartite graph summarizes local behavior of F

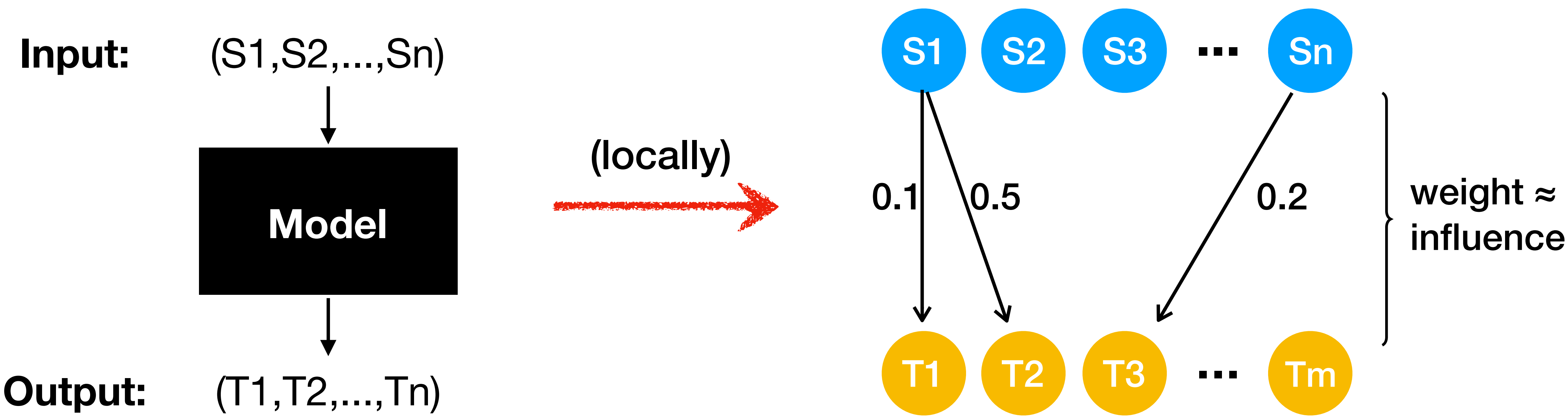


- Explanation: $E_{x \to y} = \{G^1, \ldots, G^k\}$



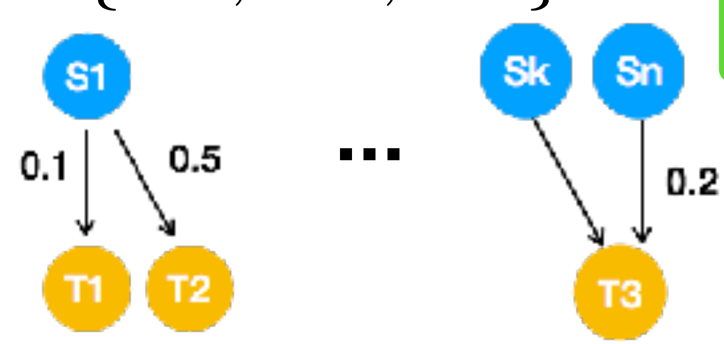
Explaining with graphs

- Weighted bipartite graph summarizes local behavior of F

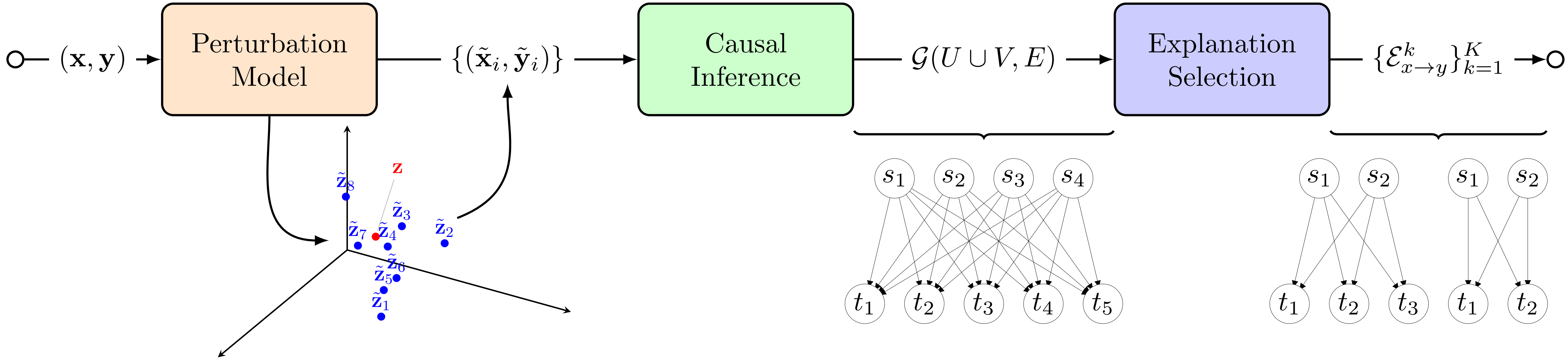


- Explanation: $E_{x \to y} = \{G^1, \ldots, G^k\}$

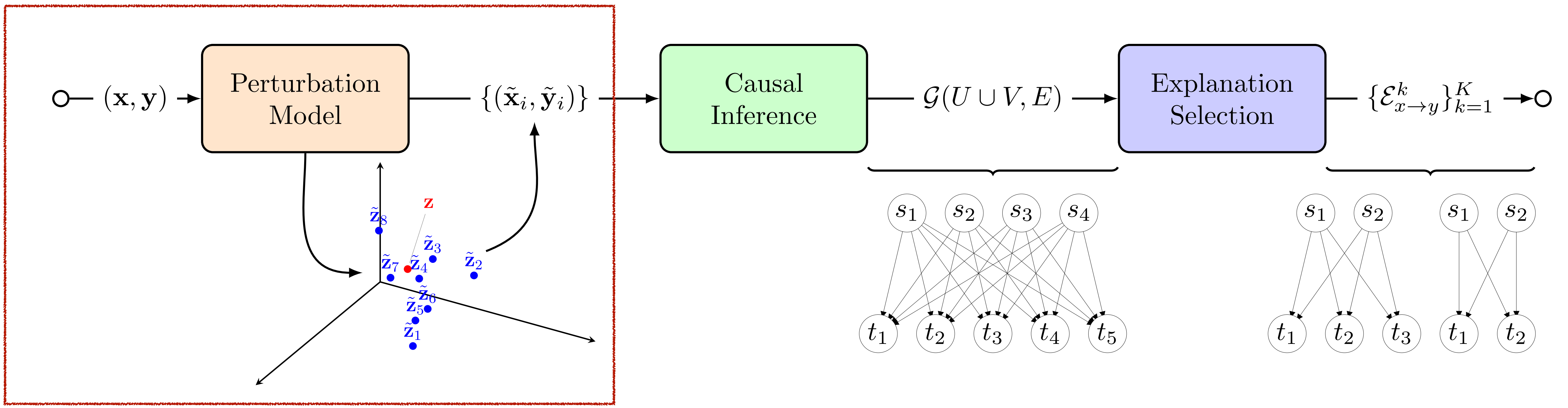
Might need to partition if dense



SOCRat: Structured-Output Causal Rationalizer

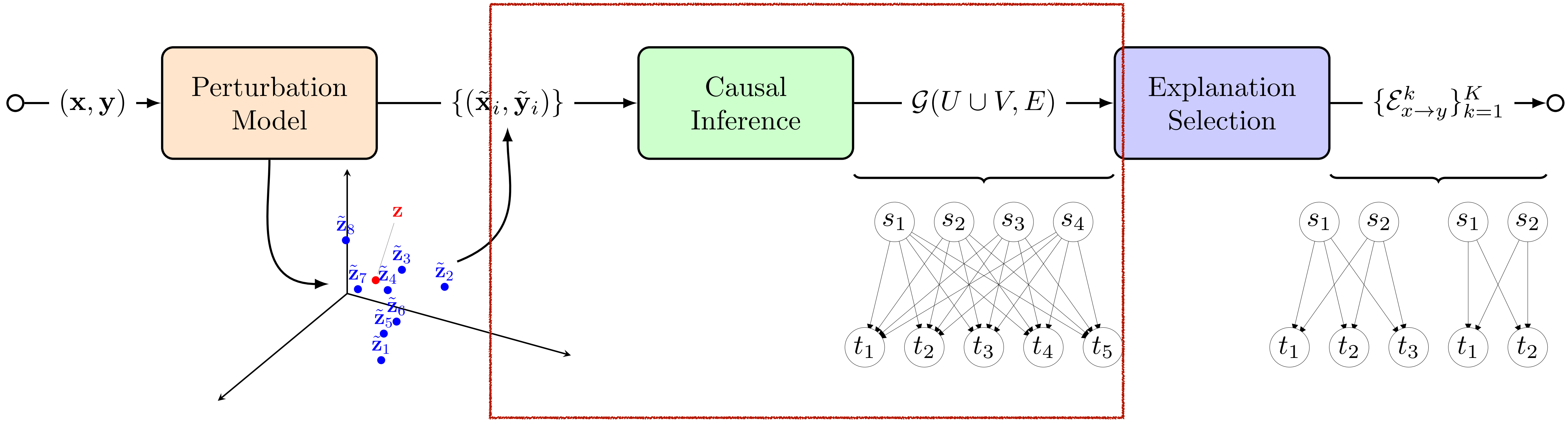


SOCRat: Structured-Output Causal Rationalizer



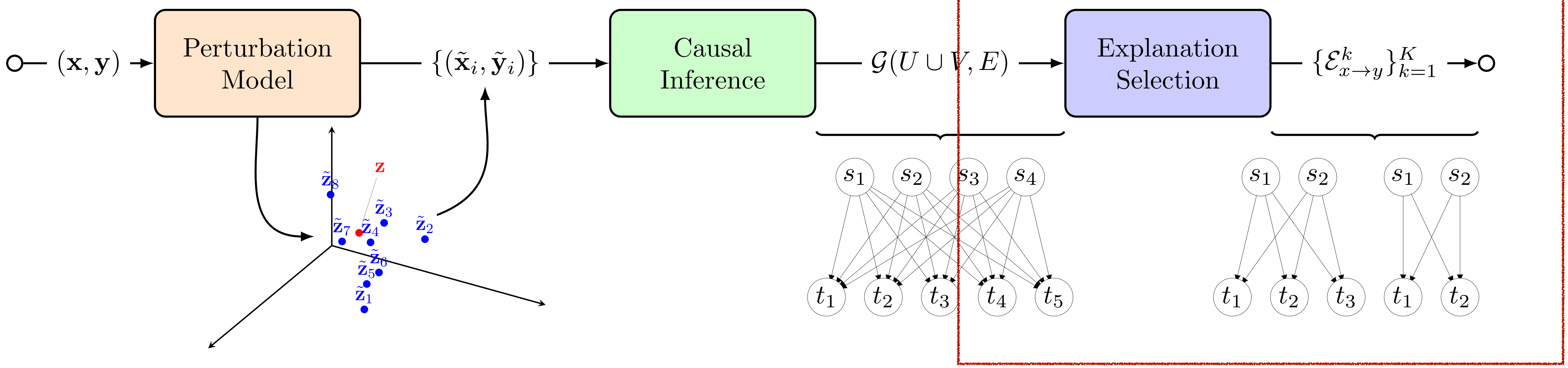
- **Perturb:** Encode -> perturb vector representation -> decode

SOCRat: Structured-Output Causal Rationalizer



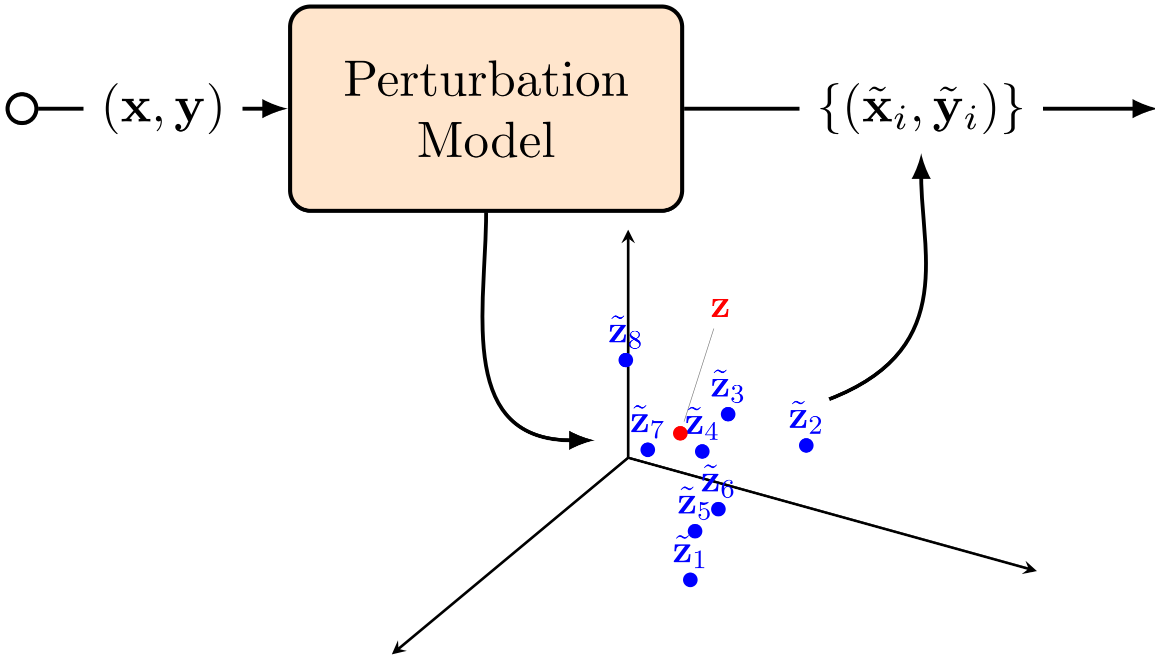
- **Perturb:** Encode -> perturb vector representation -> decode
- **Infer:** Logistic regression to infer causal dependencies

SOCRat: Structured-Output Causal Rationalizer



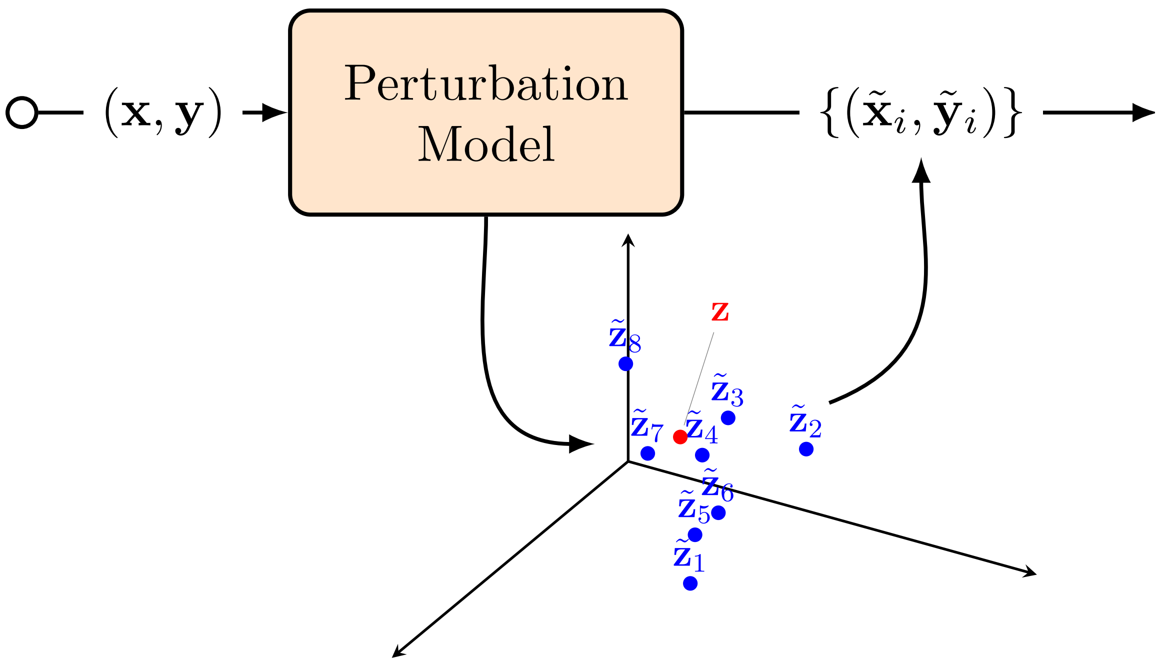
- **Perturb:** Encode -> perturb vector representation -> decode
- **Infer:** Logistic regression to infer causal dependencies
- **Select:** Partition dependency graph into *explanation chunks*

Perturbation Model



x: input
y: output (prediction)

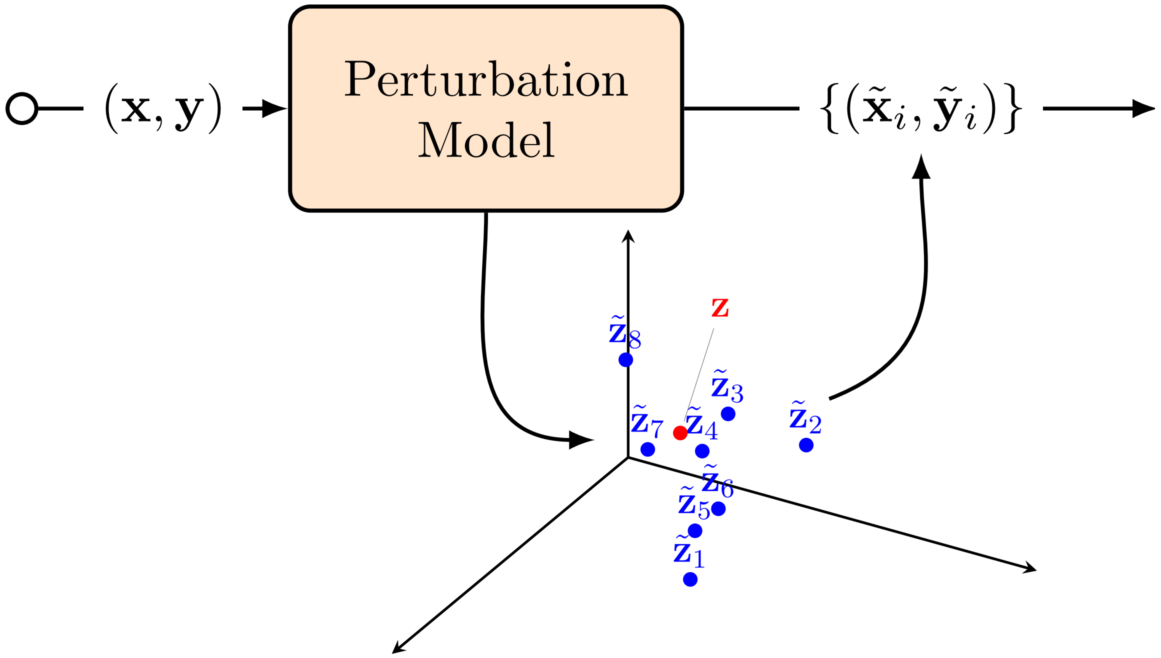
Perturbation Model



x: input
y: output (prediction)

1. Encode input to vector representation *z*

Perturbation Model

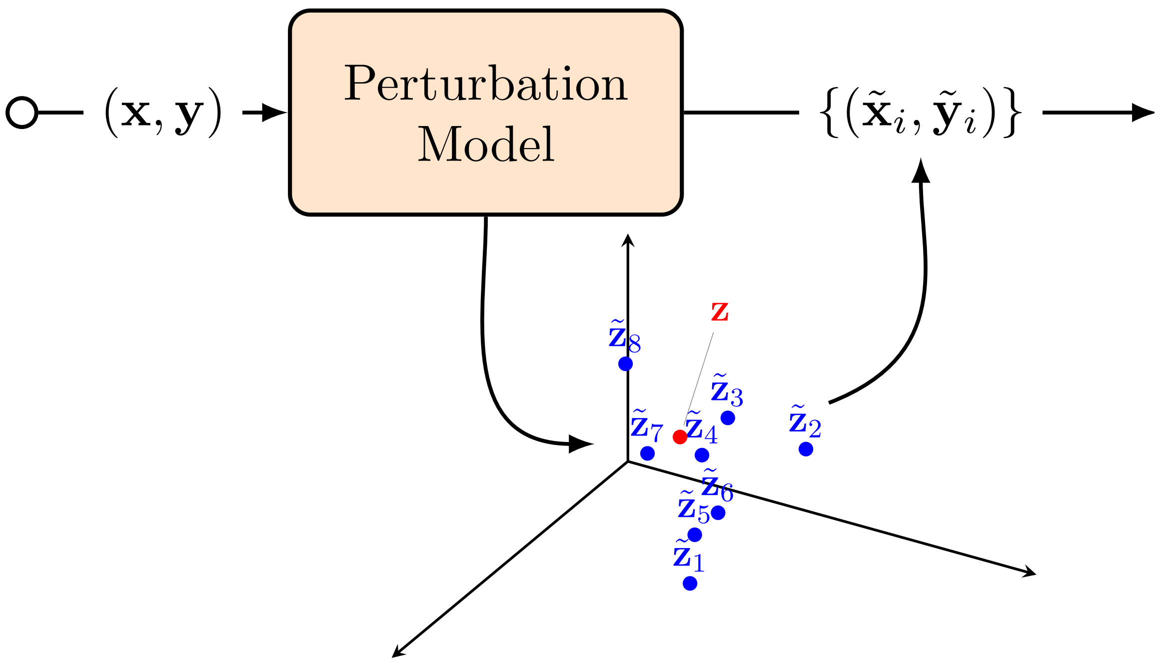


x: input
y: output (prediction)

1. Encode input to vector representation *z*

2. Generate samples $\tilde{z}$ around *z*

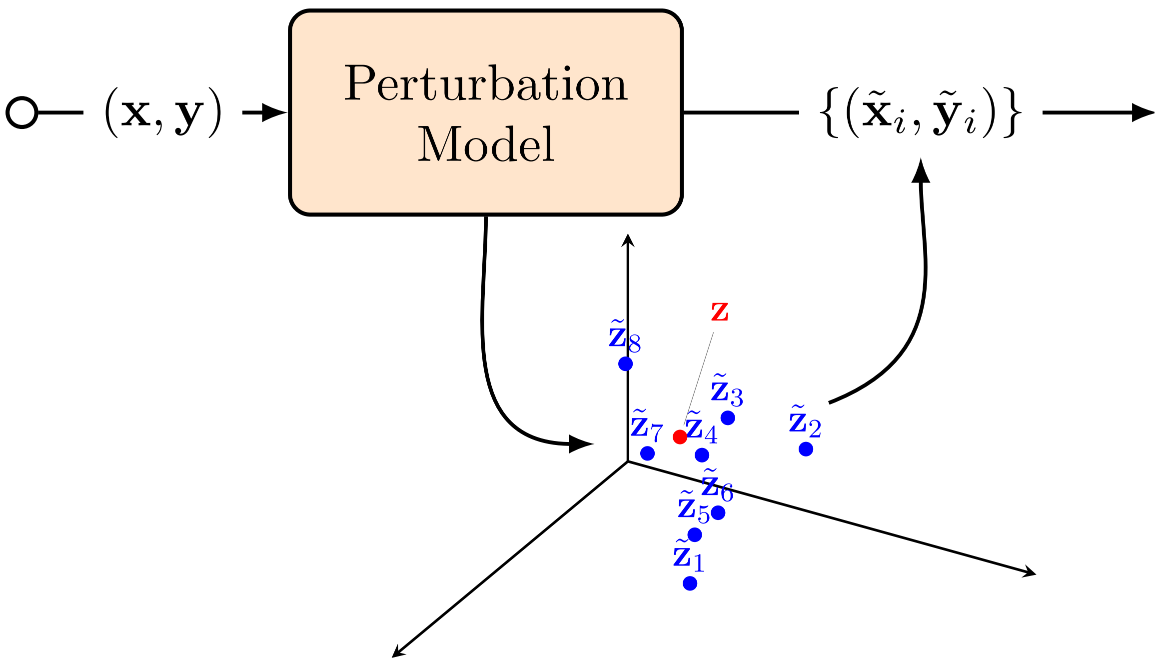
Perturbation Model



x: input
y: output (prediction)

1. Encode input to vector representation *z*
2. Generate samples $\tilde{z}$ around *z*
3. Decode samples $\tilde{z}$ into sequences

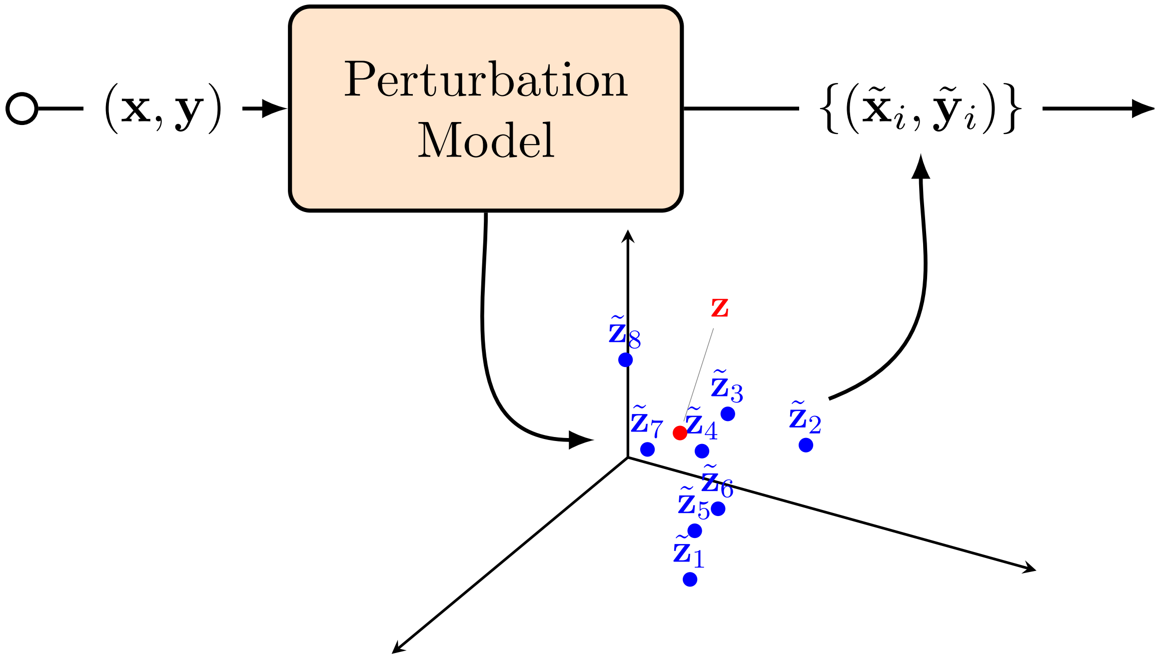
Perturbation Model



x: input
y: output (prediction)

1. Encode input to vector representation *z*
2. Generate samples $\tilde{z}$ around *z*
3. Decode samples $\tilde{z}$ into sequences
4. Map perturbed sequences using F

Perturbation Model

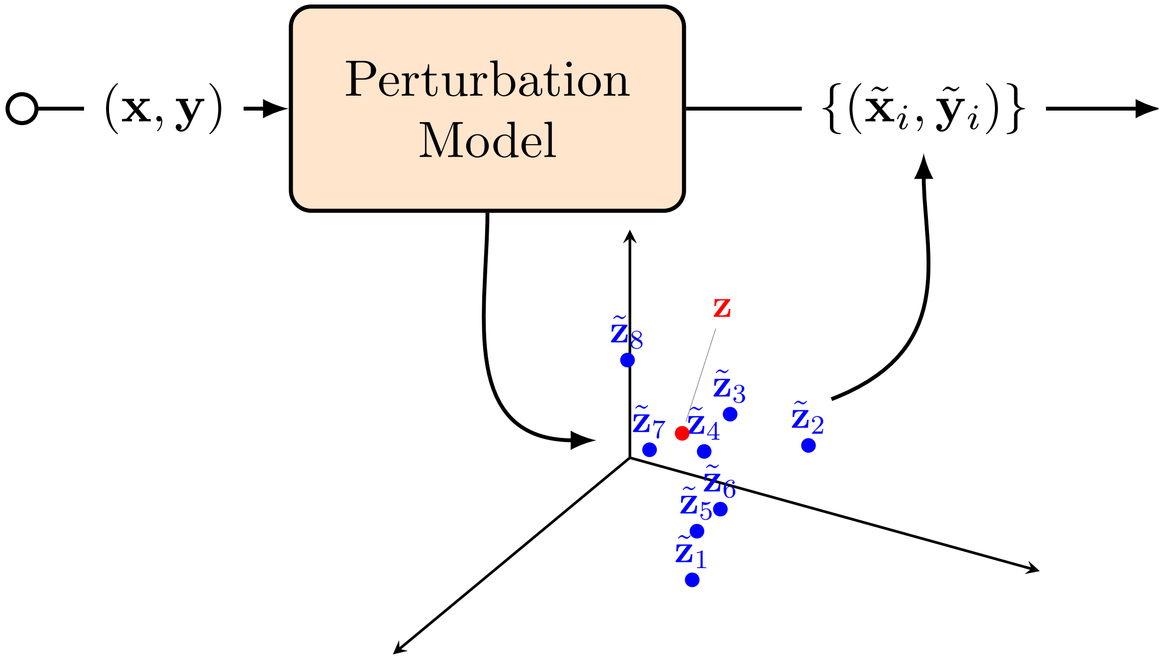


x: input
y: output (prediction)

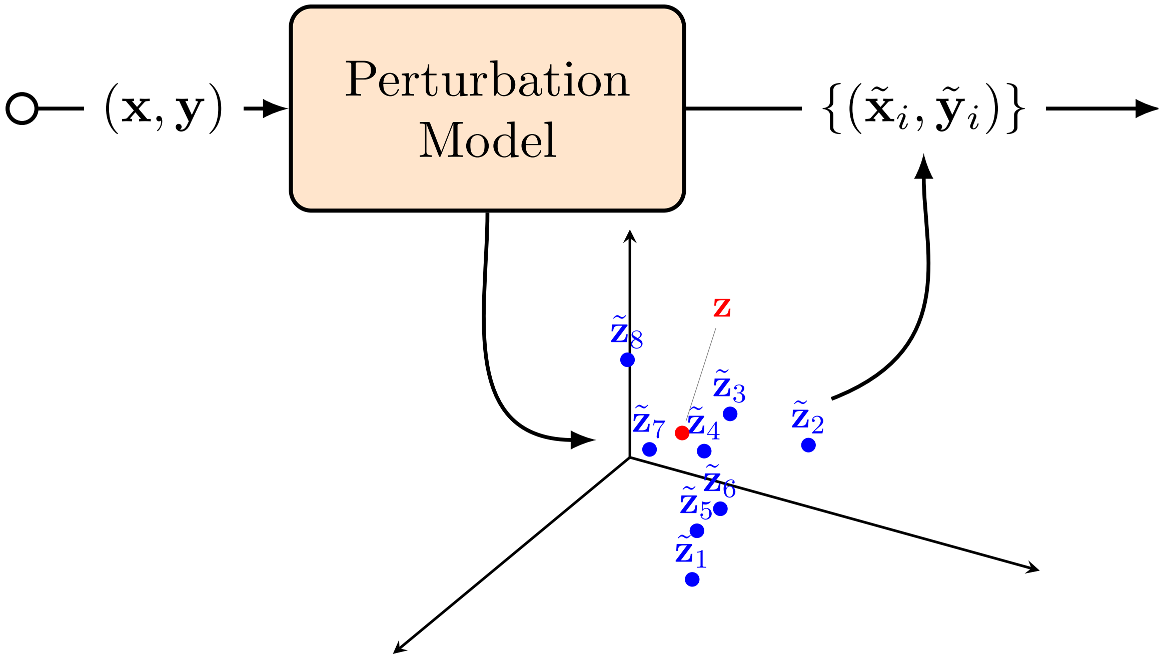
VAE

1. Encode input to vector representation *z*
2. Generate samples $\tilde{z}$ around *z*
3. Decode samples $\tilde{z}$ into sequences
4. Map perturbed sequences using F

Perturbation Model

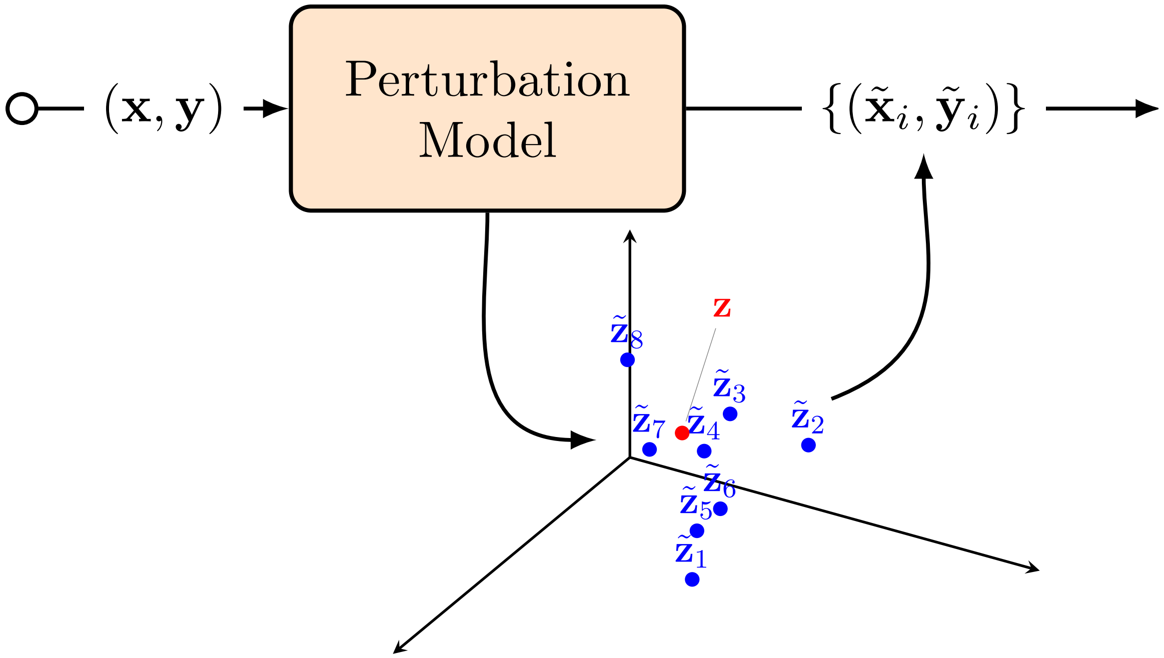


Perturbation Model



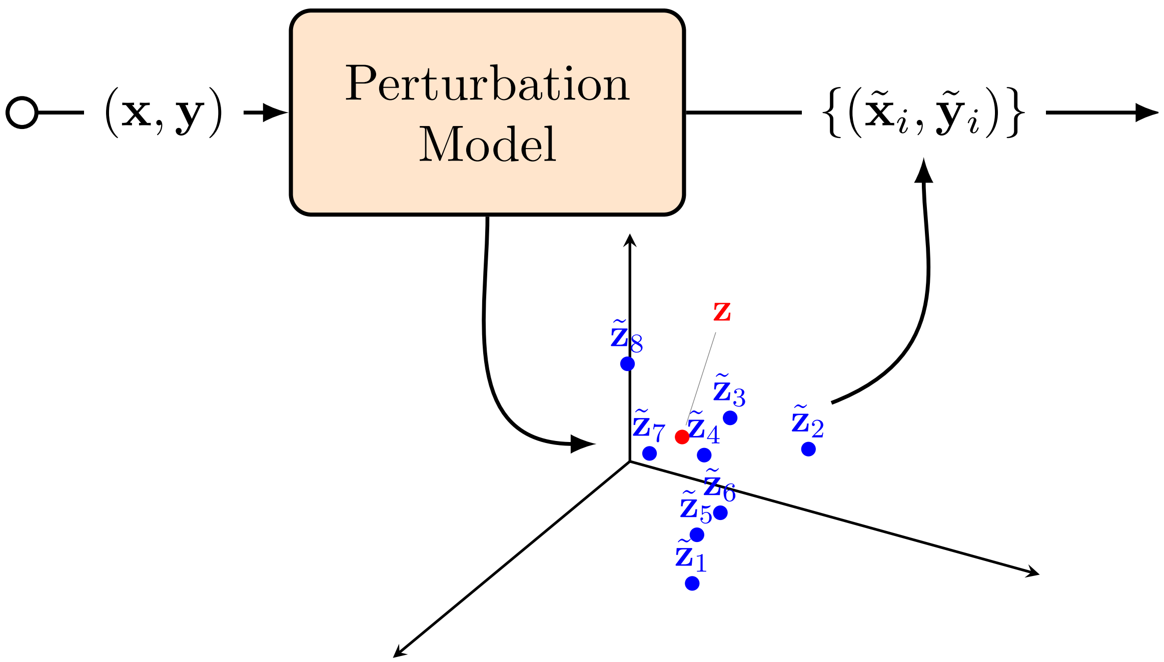
- Notion of "locality" here is **semantic**

Perturbation Model



- Notion of "locality" here is **semantic**
- After this step: list of pairs of perturbed inputs and outputs

Perturbation Model



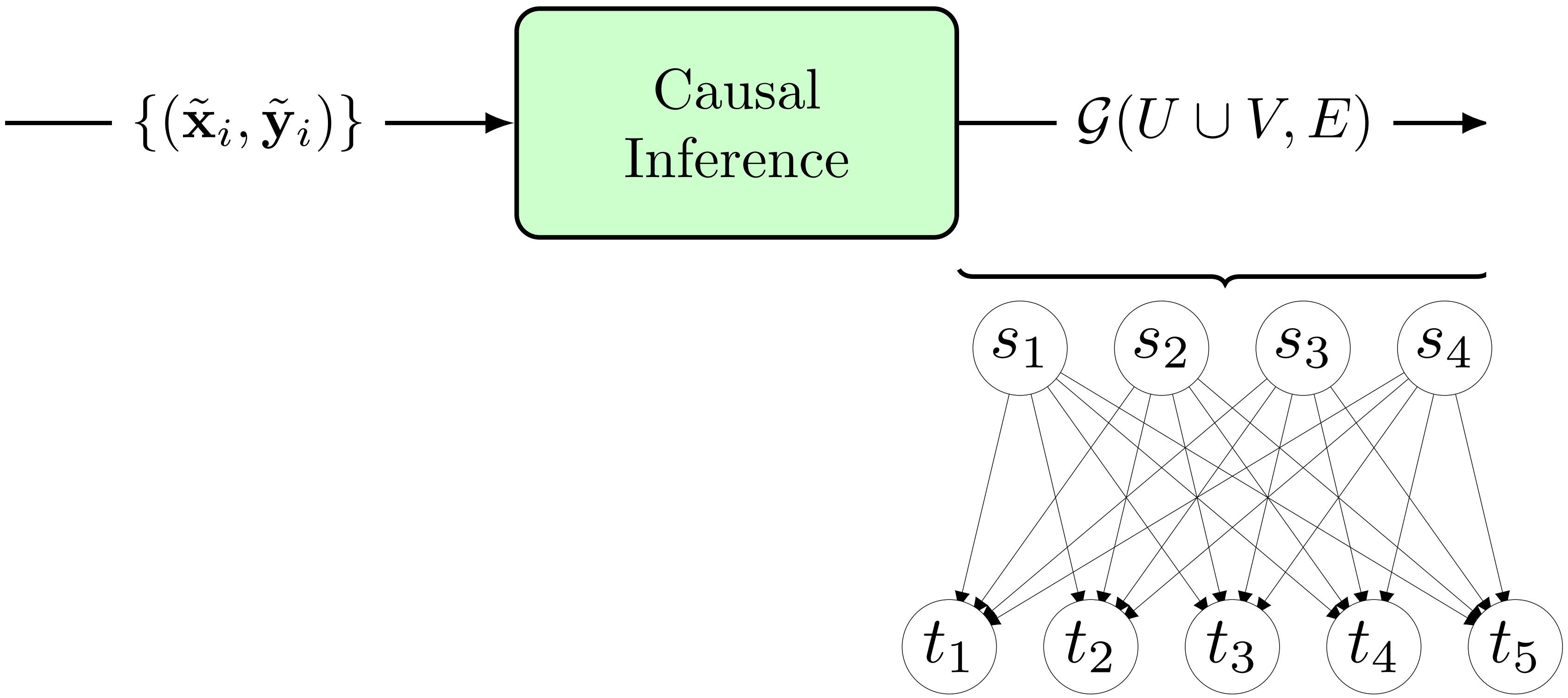
- Notion of "locality" here is **semantic**
- After this step: list of pairs of perturbed inputs and outputs

"The house is red" $\longrightarrow$ "La maison est rouge"
"The apartment is red" $\longrightarrow$ "L'appartement est rouge"
"The house is brown" $\longrightarrow$ "La maison est brune"
... etc ... ... etc ...

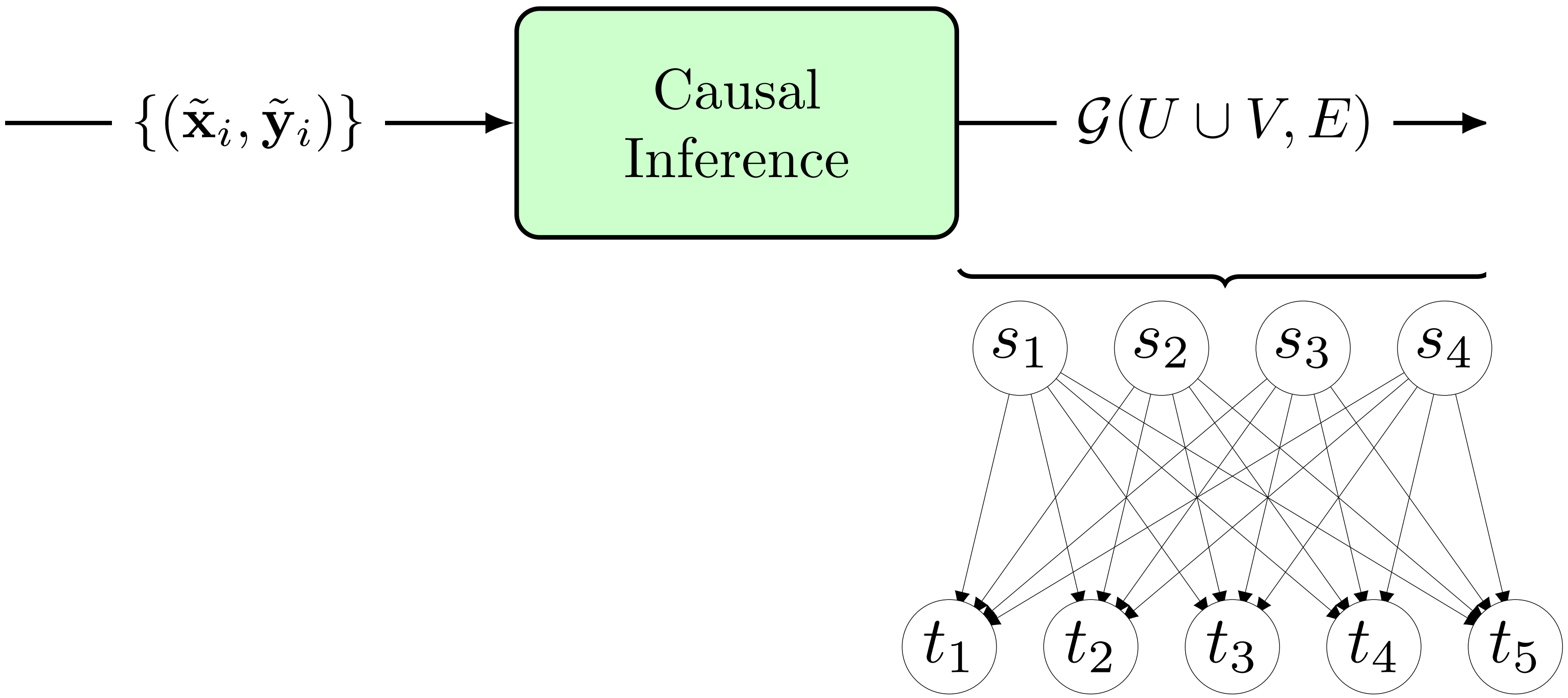
$$\{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)\}_{i=1}^{N}$$

Causal Model

Causal Model

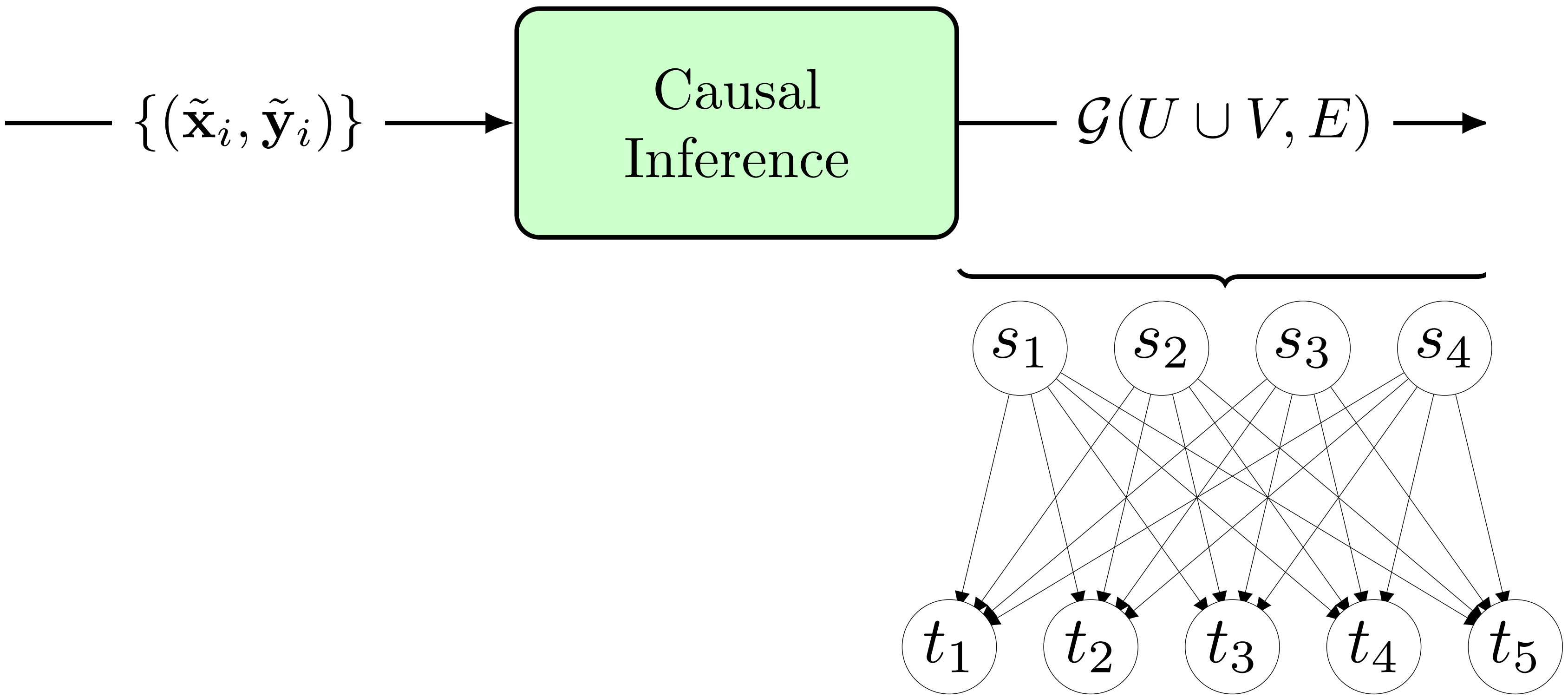


Causal Model



- Given perturbed input-output pairs, infer dependencies between original input/output tokens

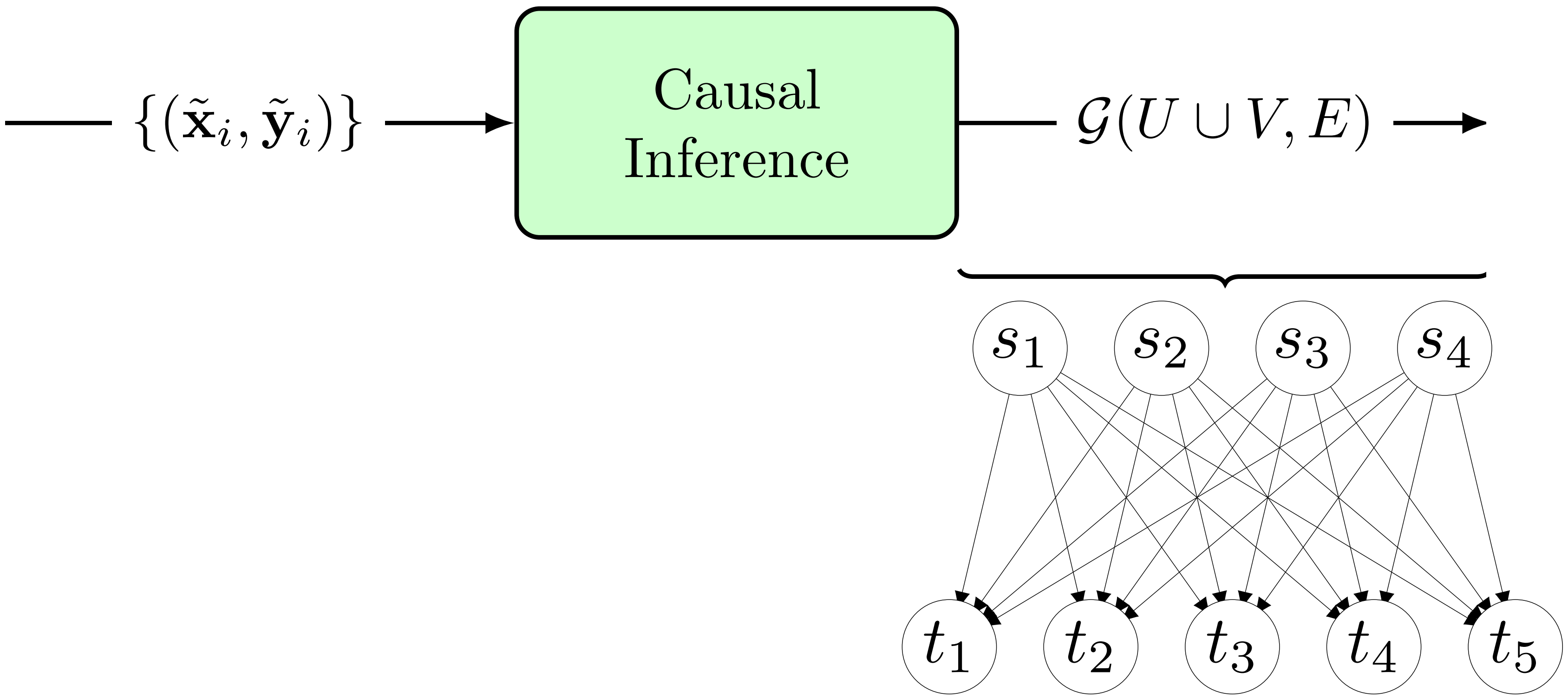
Causal Model



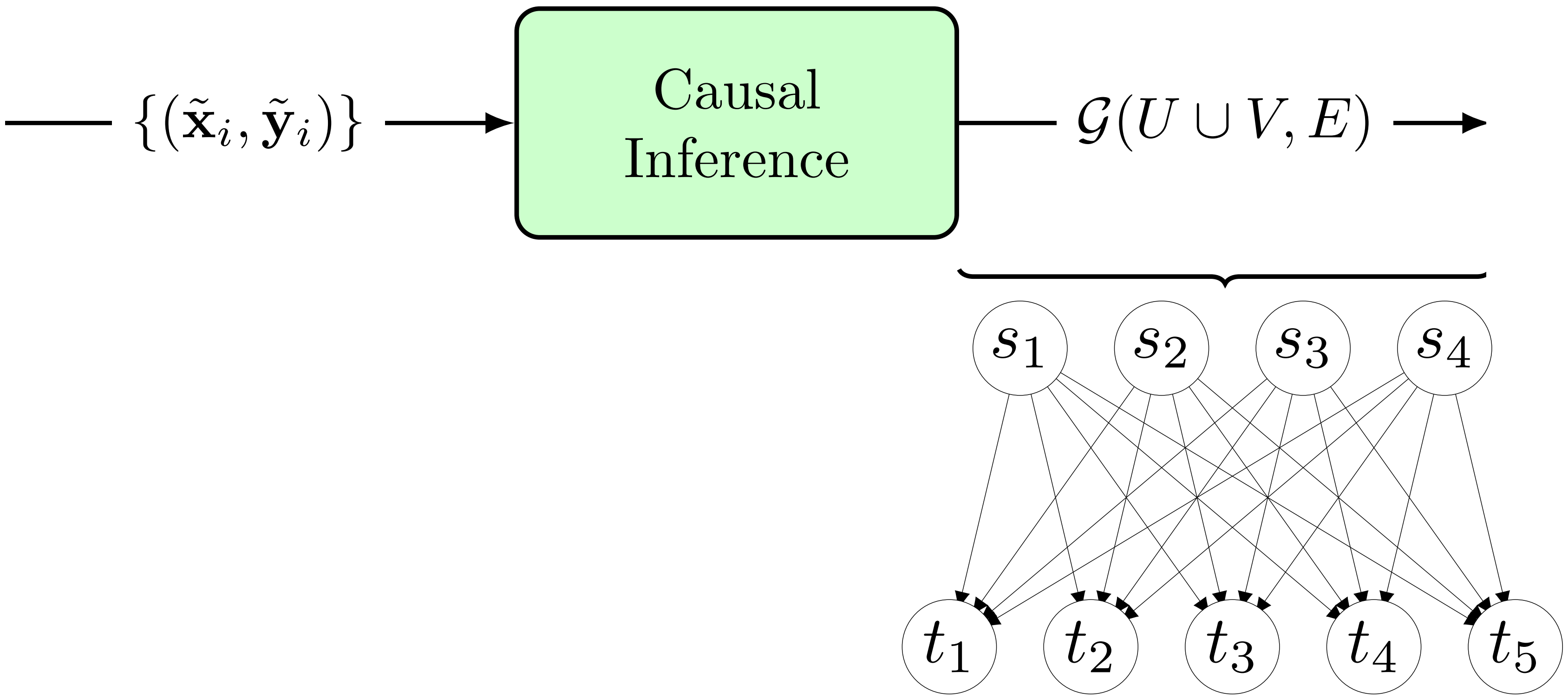
- Given perturbed input-output pairs, infer dependencies between original input/output tokens

- Simplest approach: logistic regression

Causal Model

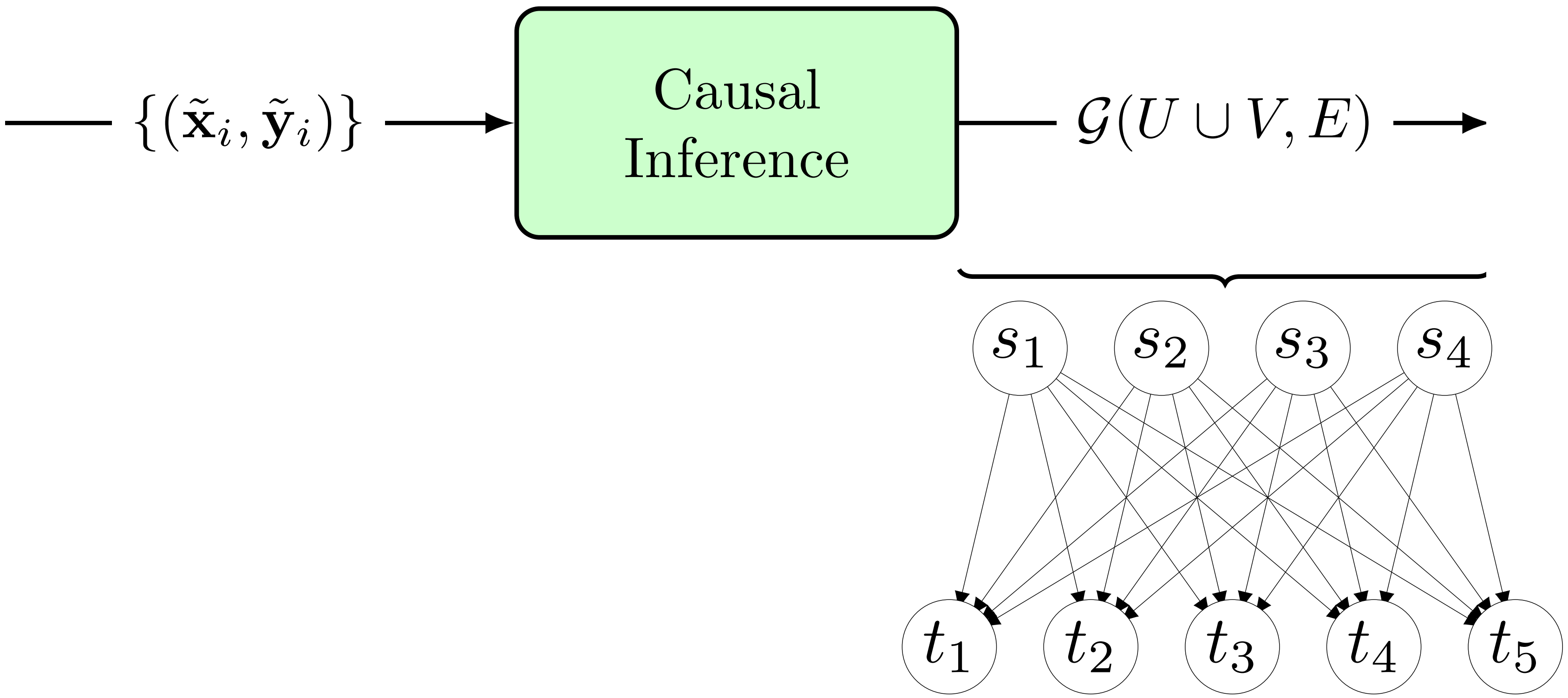


Causal Model



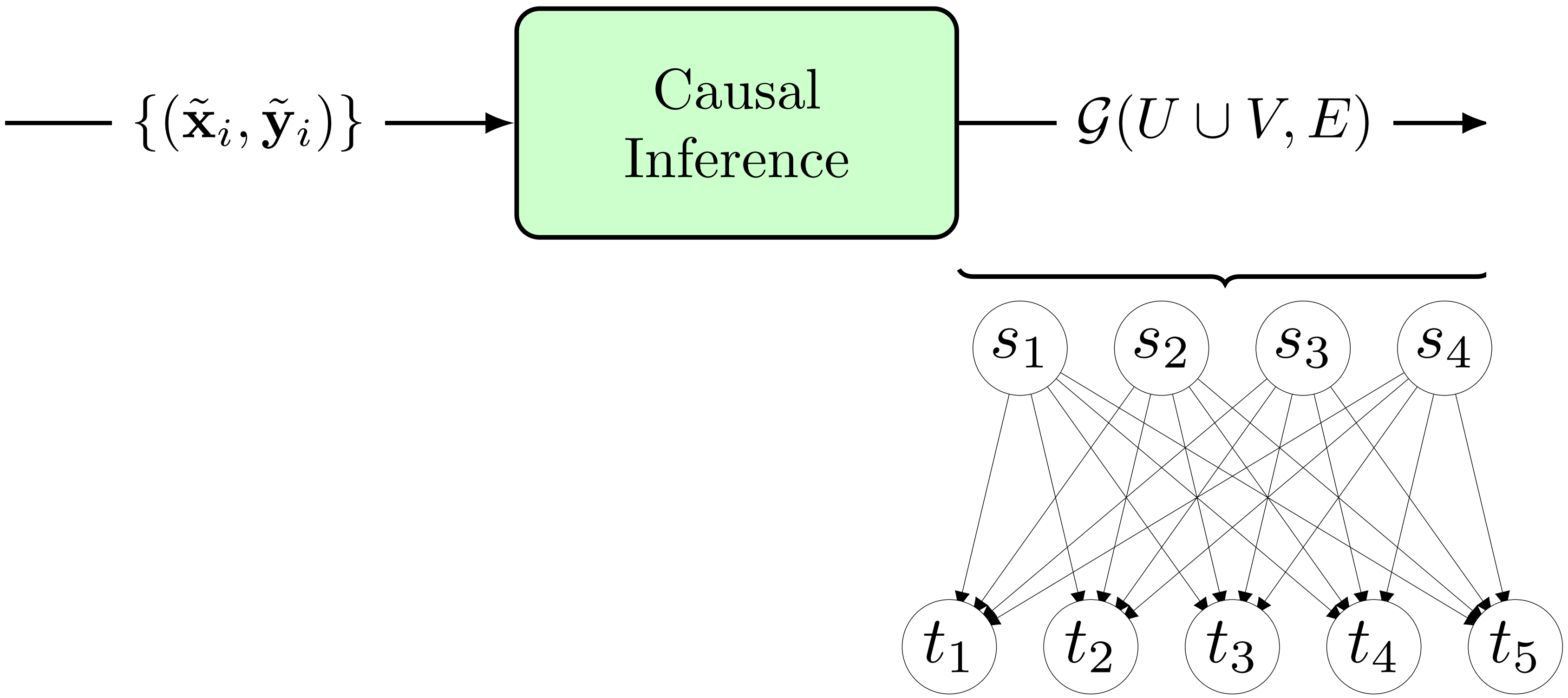
- We want our estimations to take into account **uncertainty**

Causal Model



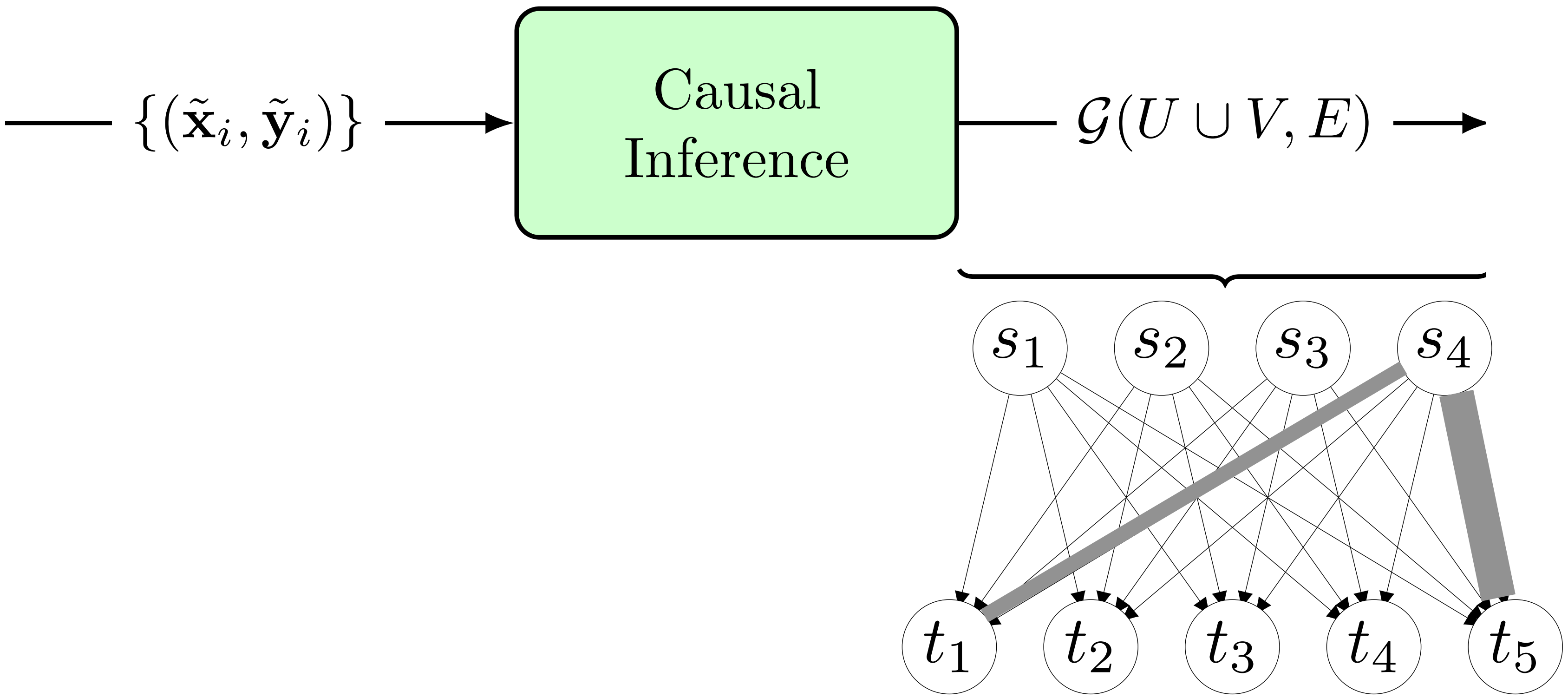
- We want our estimations to take into account **uncertainty**
- Bayesian logistic regression: $P(y_j \in \tilde{y} \mid \tilde{\mathbf{x}}) = \sigma(\boldsymbol{\theta}_j^T \phi_{\mathbf{x}}(\tilde{x}))$

Causal Model



- We want our estimations to take into account **uncertainty**
- Bayesian logistic regression: $P(y_j \in \tilde{y} \mid \tilde{\mathbf{x}}) = \sigma(\boldsymbol{\theta}_j^T \phi_{\mathbf{x}}(\tilde{x}))$
- Result: posterior mean, covariance

Causal Model



edge width = (mean) strength of dependency

- We want our estimations to take into account **uncertainty**
- Bayesian logistic regression: $P(y_j \in \tilde{y} \mid \tilde{\mathbf{x}}) = \sigma(\boldsymbol{\theta}_j^T \phi_{\mathbf{x}}(\tilde{x}))$
- Result: posterior mean, covariance

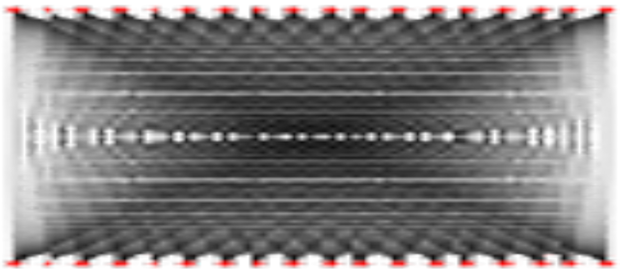
Explanation Selection

Explanation Selection

- For large inputs/outputs, dense graph might not be interpretable

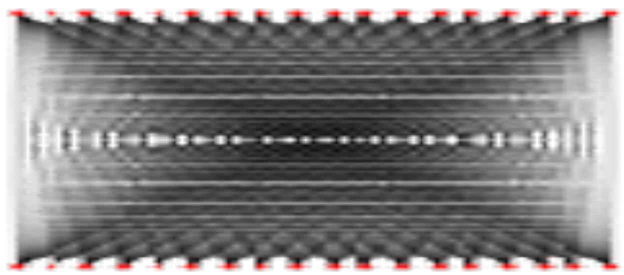
Explanation Selection

- For large inputs/outputs, dense graph might not be interpretable



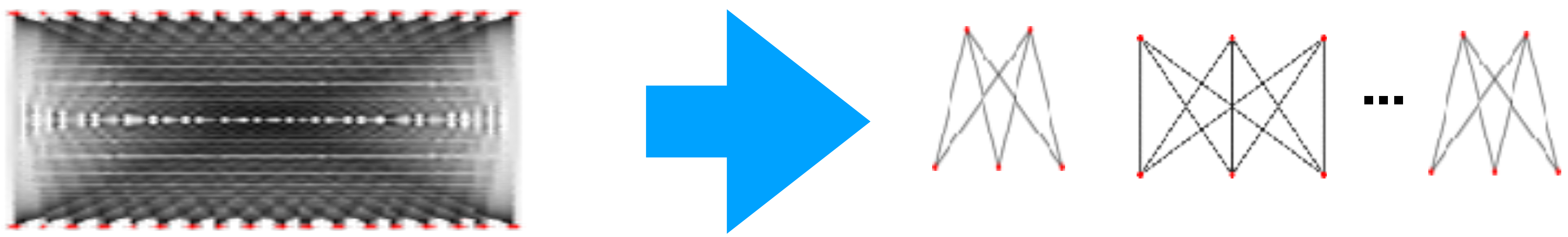
Explanation Selection

- For large inputs/outputs, dense graph might not be interpretable



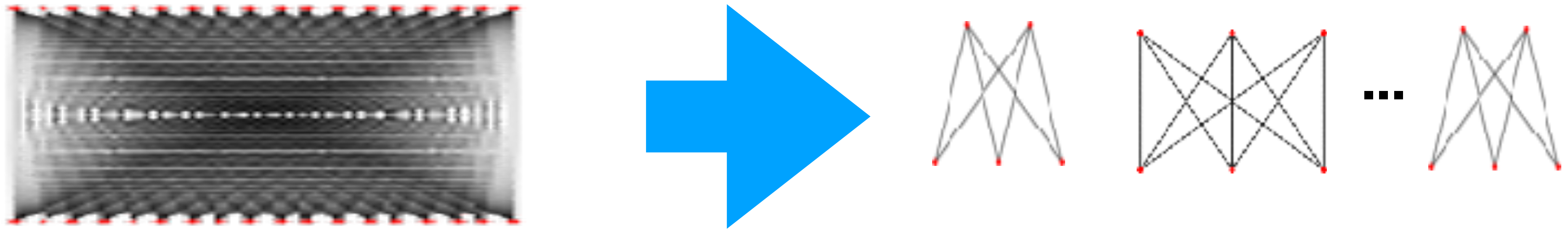
Explanation Selection

- For large inputs/outputs, dense graph might not be interpretable



Explanation Selection

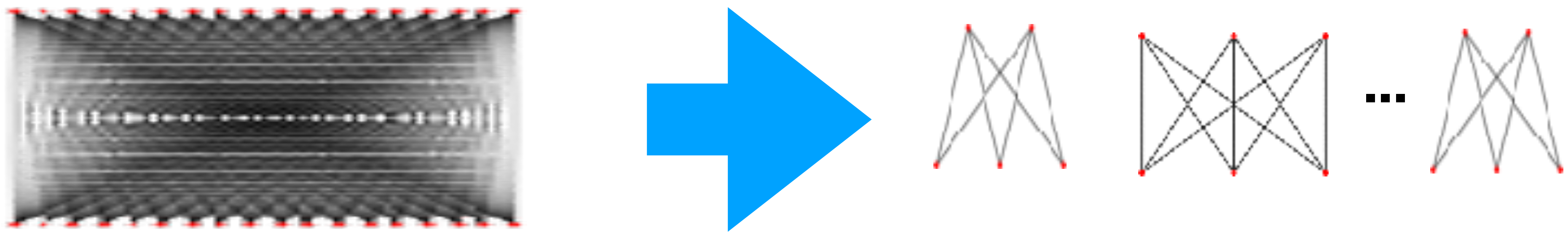
- For large inputs/outputs, dense graph might not be interpretable



- We cast the problem as k-cut graph partitioning

Explanation Selection

- For large inputs/outputs, dense graph might not be interpretable

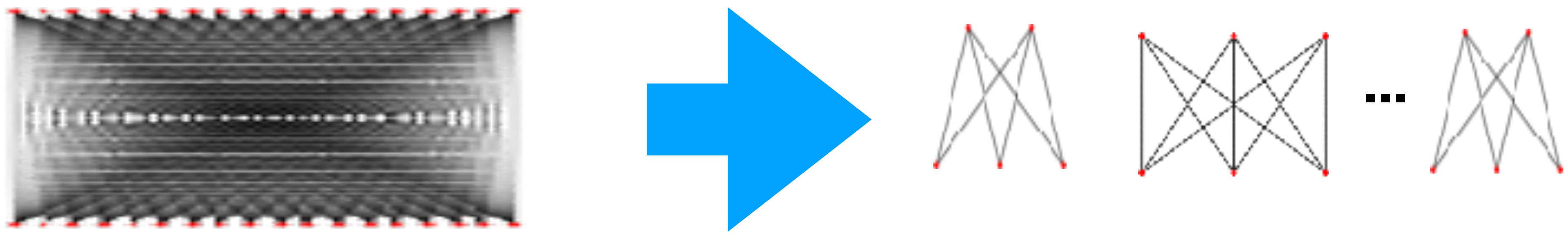


- We cast the problem as k-cut graph partitioning

- Traditional methods (coclustering, biclustering) don't take into account uncertainty

Explanation Selection

- For large inputs/outputs, dense graph might not be interpretable



- We cast the problem as k-cut graph partitioning
- Traditional methods (coclustering, biclustering) don't take into account uncertainty
- **Graph partitioning with uncertainty** [Fan et al. 2012]

Explanation Selection

$$y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$$

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

$$y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$$

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

  Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

$$\min_{(x_{ik}^u, x_{jk}^v, y_{ij}) \in Y} \sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij} + \max_{\substack{S: S \subseteq J, |S| \leq \Gamma \\ (i_t, j_t) \in J \setminus S}} \sum_{(i,j) \in S} \hat{\theta}_{ij} y_{ij} + (\Gamma - \lfloor \Gamma \rfloor) \hat{\theta}_{i_t, j_t} y_{i_t, j_t}$$

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

$$\min_{(x^u_{ik}, x^v_{jk}, y_{ij}) \in Y} \sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij}$$

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

$$\min_{(x^u_{ik}, x^v_{jk}, y_{ij}) \in Y} \sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij}$$

partition size
constraints

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

$$\min_{(x^u_{ik}, x^v_{jk}, y_{ij}) \in Y} \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij}}_{\text{Mean Total Cost}}$$

partition size constraints

Mean Total Cost

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

$$\min_{(x^u_{ik}, x^v_{jk}, y_{ij}) \in Y} \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij} +}_{\text{Mean Total Cost}} \max_{\substack{S: S \subseteq J, |S| \leq \Gamma \\ (i_t, j_t) \in J \setminus S}} \sum_{(i,j) \in S} \hat{\theta}_{ij} y_{ij} + (\Gamma - \lfloor \Gamma \rfloor) \hat{\theta}_{i_t, j_t} y_{i_t, j_t}$$

partition size constraints

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

$$\min_{(x^u_{ik}, x^v_{jk}, y_{ij}) \in Y} \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij}}_{\text{Mean Total Cost}} + \max_{\substack{S: S \subseteq J, |S| \leq \Gamma \\ (i_t, j_t) \in J \setminus S}} \sum_{(i,j) \in S} \hat{\theta}_{ij} y_{ij} + (\Gamma - \lfloor \Gamma \rfloor) \hat{\theta}_{i_t, j_t} y_{i_t, j_t}$$

edges allowed to deviate

partition size constraints

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

edges allowed to deviate

robustness control

$$\min_{(x^u_{ik}, x^v_{jk}, y_{ij}) \in Y} \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij}}_{\text{Mean Total Cost}} + \max_{\substack{S: S \subseteq J, |S| \leq \Gamma \\ (i_t, j_t) \in J \setminus S}} \sum_{(i,j) \in S} \hat{\theta}_{ij} y_{ij} + (\Gamma - \lfloor \Gamma \rfloor) \hat{\theta}_{i_t, j_t} y_{i_t, j_t}$$

partition size constraints

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

edges allowed to deviate

robustness control

$$\min_{(x^u_{ik}, x^v_{jk}, y_{ij}) \in Y} \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij}}_{\text{Mean Total Cost}} + \underbrace{\max_{\substack{S: S \subseteq J, |S| \leq \Gamma \\ (i_t, j_t) \in J \setminus S}} \sum_{(i,j) \in S} \hat{\theta}_{ij} y_{ij} + (\Gamma - \lfloor \Gamma \rfloor) \hat{\theta}_{i_t, j_t} y_{i_t, j_t}}_{\text{Cost of worst-case deviation}}$$

partition size constraints

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

edges allowed to deviate

robustness control

$$\min_{(x^u_{ik}, x^v_{jk}, y_{ij}) \in Y} \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij}}_{\text{Mean Total Cost}} + \underbrace{\max_{\substack{S: S \subseteq J, |S| \leq \Gamma \\ (i_t, j_t) \in J \setminus S}} \sum_{(i,j) \in S} \hat{\theta}_{ij} y_{ij} + (\Gamma - \lfloor \Gamma \rfloor) \hat{\theta}_{i_t, j_t} y_{i_t, j_t}}_{\text{Cost of worst-case deviation}}$$

partition size constraints

Mean Total Cost

Cost of worst-case deviation

- Can be cast as Mixed Integer Programming problem

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

edges allowed to deviate

robustness control

$$\min_{(x^u_{ik}, x^v_{jk}, y_{ij}) \in Y} \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij}}_{\text{Mean Total Cost}} + \underbrace{\max_{\substack{S : S \subseteq J, |S| \leq \Gamma \\ (i_t, j_t) \in J \setminus S}} \sum_{(i,j) \in S} \hat{\theta}_{ij} y_{ij} + (\Gamma - \lfloor \Gamma \rfloor) \hat{\theta}_{i_t, j_t} y_{i_t, j_t}}_{\text{Cost of worst-case deviation}}$$

partition size constraints

- Can be cast as Mixed Integer Programming problem

- Each partition -> an explanation *chunk*

Explanation Selection

- **Graph partitioning with uncertainty** [Fan et al. 2012]

Edge weight intervals: $\theta_{ij} \pm \hat{\theta}_{ij}$ $\quad y_{ij} = \begin{cases} 1 & \text{if } v_i, u_j \text{ in different components} \\ 0 & \text{ow} \end{cases}$

edges allowed to deviate

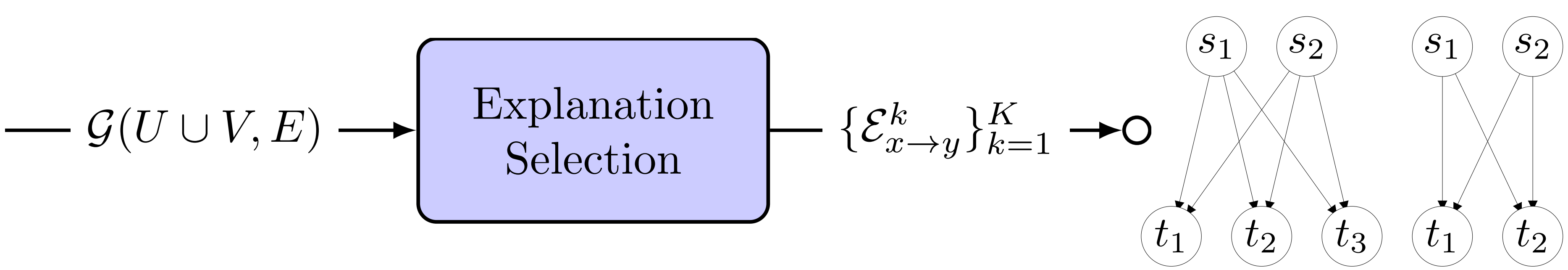
robustness control

$$\min_{(x^u_{ik}, x^v_{jk}, y_{ij}) \in Y} \underbrace{\sum_{i=1}^{n} \sum_{j=1}^{m} \theta_{ij} y_{ij}}_{\text{Mean Total Cost}} + \underbrace{\max_{\substack{S: S \subseteq J, |S| \leq \Gamma \\ (i_t, j_t) \in J \setminus S}} \sum_{(i,j) \in S} \hat{\theta}_{ij} y_{ij} + (\Gamma - \lfloor \Gamma \rfloor) \hat{\theta}_{i_t, j_t} y_{i_t, j_t}}_{\text{Cost of worst-case deviation}}$$

partition size constraints

- Can be cast as Mixed Integer Programming problem

- Each partition -> an explanation *chunk*



SocRat - Pseudocode

Algorithm 1 Structured-output causal rationalizer

procedure SOCRAT($\mathbf{x}, \mathbf{y}, F$)
  $(\boldsymbol{\mu}, \boldsymbol{\sigma}) \leftarrow \text{ENCODE}(\mathbf{x})$
  for $i = 1$ **to** N **do**
    $\tilde{\mathbf{z}}_i \leftarrow \text{SAMPLE}(\boldsymbol{\mu}, \boldsymbol{\sigma})$
    $\tilde{\mathbf{x}}_i \leftarrow \text{DECODE}(\tilde{\mathbf{z}}_i)$
    $\tilde{\mathbf{y}}_i \leftarrow F(\tilde{\mathbf{x}}_i)$
  end for
  $G \leftarrow \text{CAUSAL}(\mathbf{x}, \mathbf{y}, \{\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i\}_{i=1}^{N})$
  $E_{x \mapsto y} \leftarrow \text{BIPARTITION}(G)$
  $E_{x \mapsto y} \leftarrow \text{SORT}(E_{x \mapsto y})$ ▷ By cut capacity
  return $E_{x \mapsto y}$
end procedure

Perturbation Model.

Experiments

How good are the explanations?

How good are the explanations?

- Gold human explanations are hard to obtain

How good are the explanations?

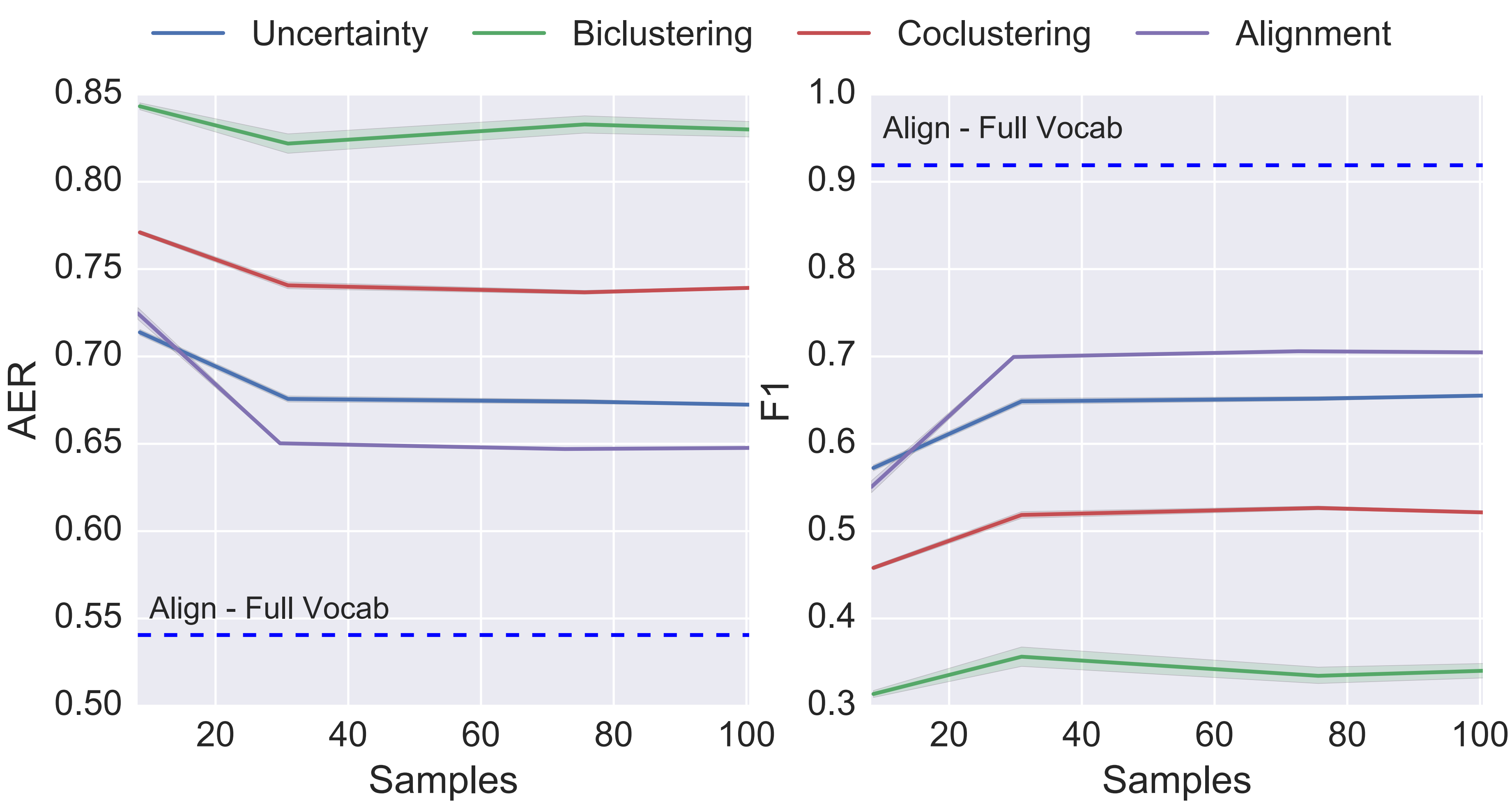
- Gold human explanations are hard to obtain
- Instead: toy task with known alignments

How good are the explanations?

- Gold human explanations are hard to obtain
- Instead: toy task with known alignments
- Word-to-phoneme mapping (e.g. *vowels* -> `V AW1 AHO L Z`)

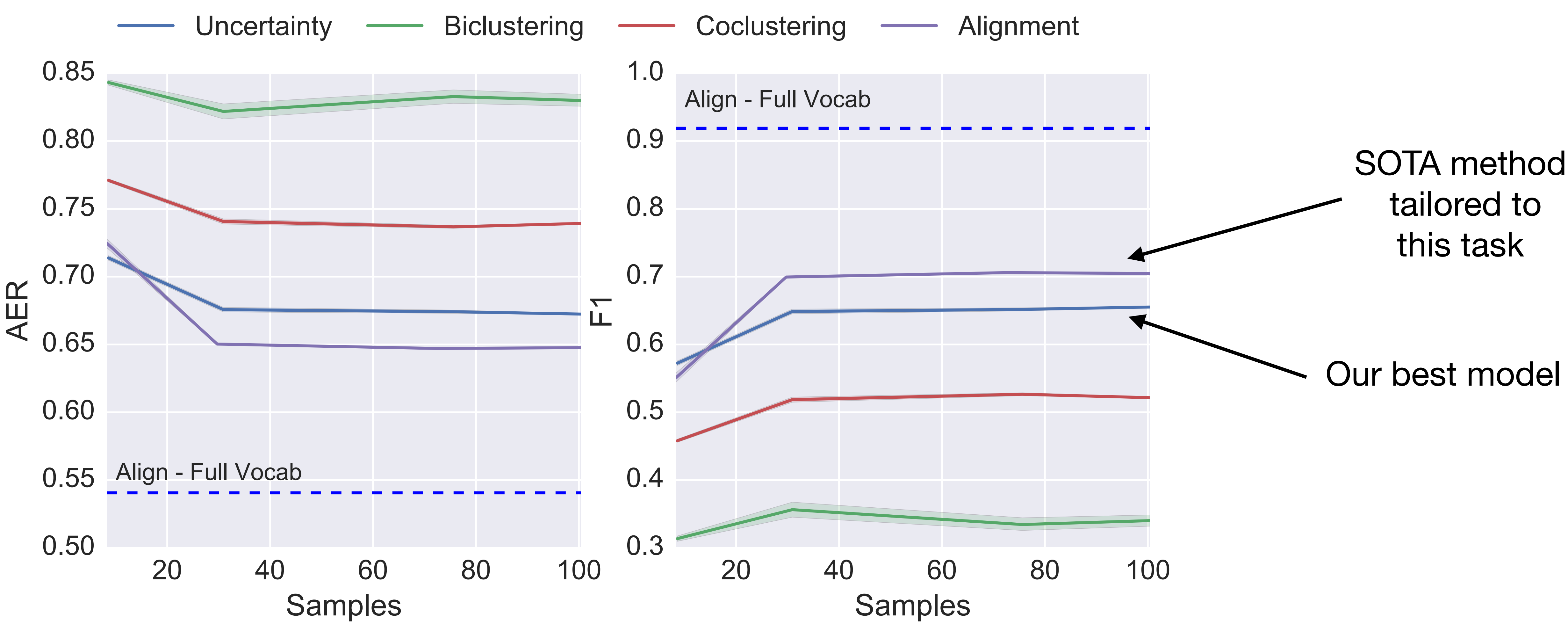
How good are the explanations?

- Gold human explanations are hard to obtain
- Instead: toy task with known alignments
- Word-to-phoneme mapping (e.g. *vowels* -> `V AW1 AHO L Z`)



How good are the explanations?

- Gold human explanations are hard to obtain
- Instead: toy task with known alignments
- Word-to-phoneme mapping (e.g. *vowels* -> `V AW1 AH0 L Z`)



Word-to-phoneme explanations

Word-to-phoneme explanations

- **Input:** b o o l e a n

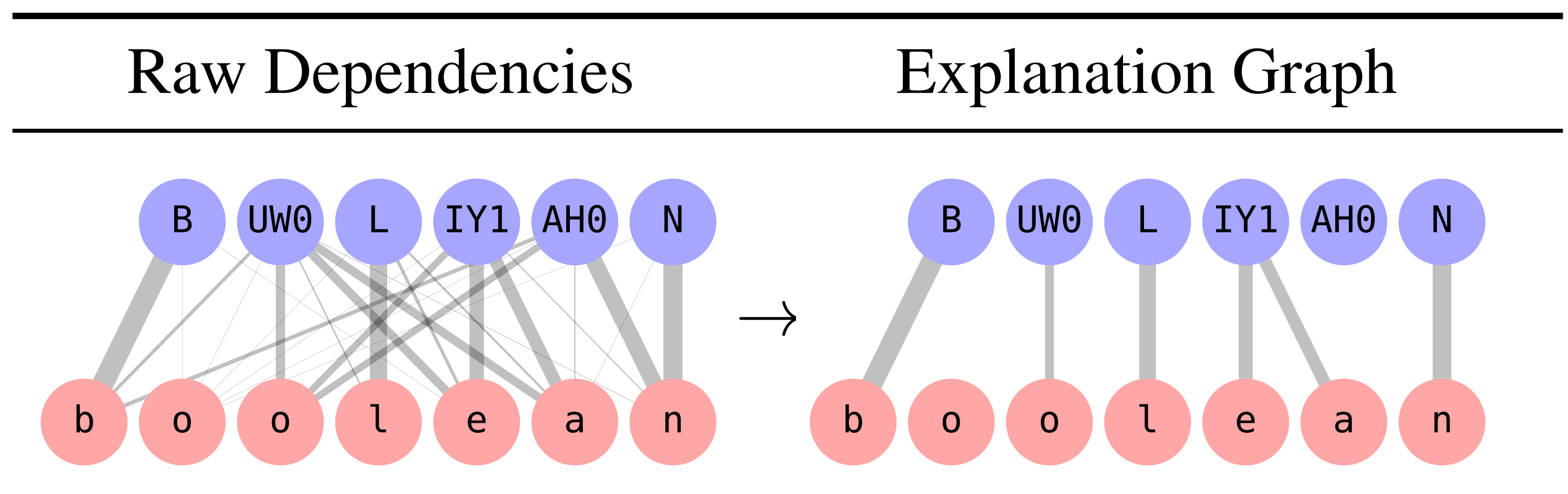
Word-to-phoneme explanations

- **Input:** b o o l e a n
- **Output:** `B UW0 L IY1 AH0 N`

Word-to-phoneme explanations

- **Input:** b o o l e a n
- **Output:** `B UW0 L IY1 AH0 N`

Large k
(more clusters)



(before partitioning)

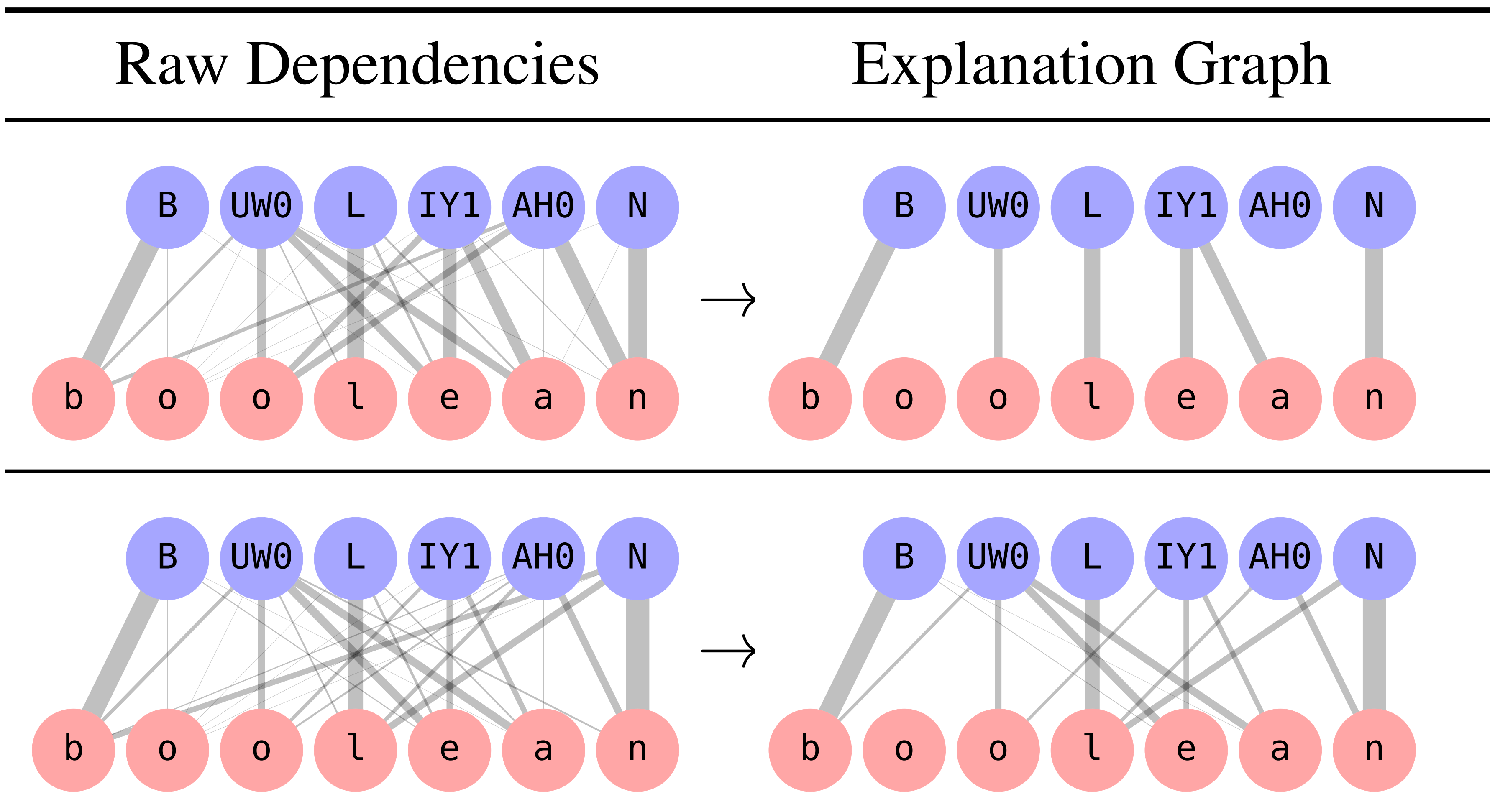
(after partitioning)

Word-to-phoneme explanations

- **Input:** b o o l e a n
- **Output:** B UW0 L IY1 AH0 N

Large k
(more clusters)

Small k
(fewer clusters)



(before partitioning) (after partitioning)

Application: Machine Translation

Application: Machine Translation

- MT is arguably the most popular sequence-to-sequence task

Application: Machine Translation

- MT is arguably the most popular sequence-to-sequence task
- SOTA models are very complex: 50-200M params, >4 layers, hierarchical self-attention

Application: Machine Translation

- MT is arguably the most popular sequence-to-sequence task
- SOTA models are very complex: 50-200M params, >4 layers, hierarchical self-attention
- **Task**: English -> German

Application: Machine Translation

- MT is arguably the most popular sequence-to-sequence task
- SOTA models are very complex: 50-200M params, >4 layers, hierarchical self-attention
- **Task**: English -> German
- **Black-box translators:**

Application: Machine Translation

- MT is arguably the most popular sequence-to-sequence task
- SOTA models are very complex: 50-200M params, >4 layers, hierarchical self-attention
- **Task**: English -> German
- **Black-box translators:**
    - Azure's MT system

Application: Machine Translation

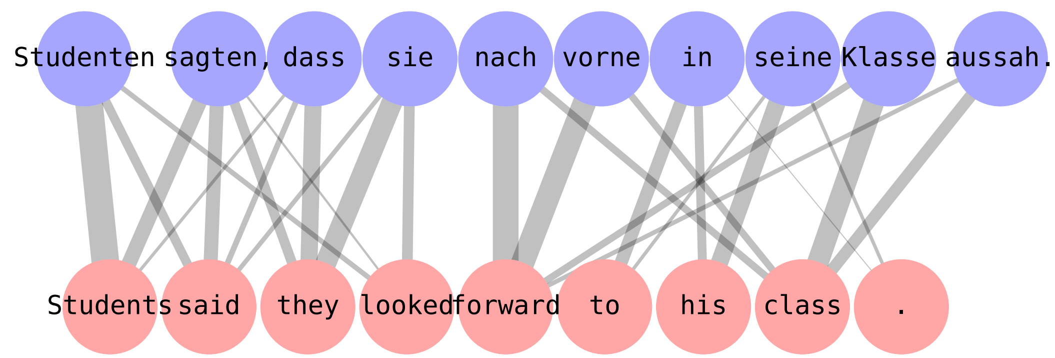
- MT is arguably the most popular sequence-to-sequence task
- SOTA models are very complex: 50-200M params, >4 layers, hierarchical self-attention
- **Task**: English -> German
- **Black-box translators:**
    - Azure's MT system
    - Neural MT system (trained by us)

Application: Machine Translation

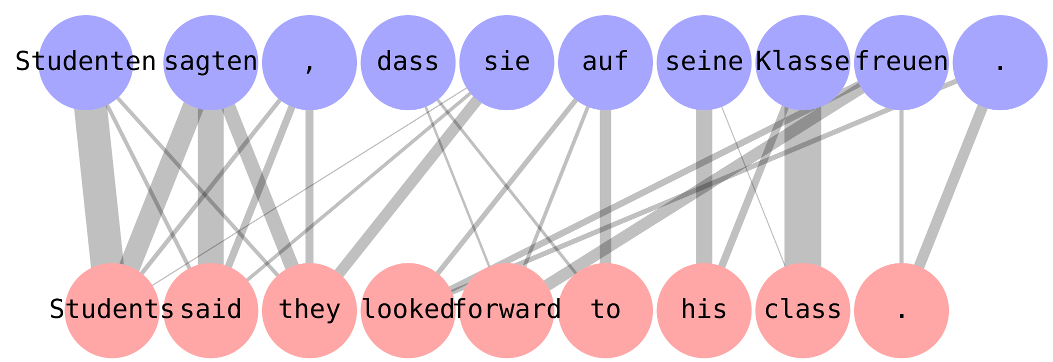
- MT is arguably the most popular sequence-to-sequence task
- SOTA models are very complex: 50-200M params, >4 layers, hierarchical self-attention
- **Task**: English -> German
- **Black-box translators:**
    - Azure's MT system
    - Neural MT system (trained by us)
    - A human (native speaker of German)

Application: Machine Translation

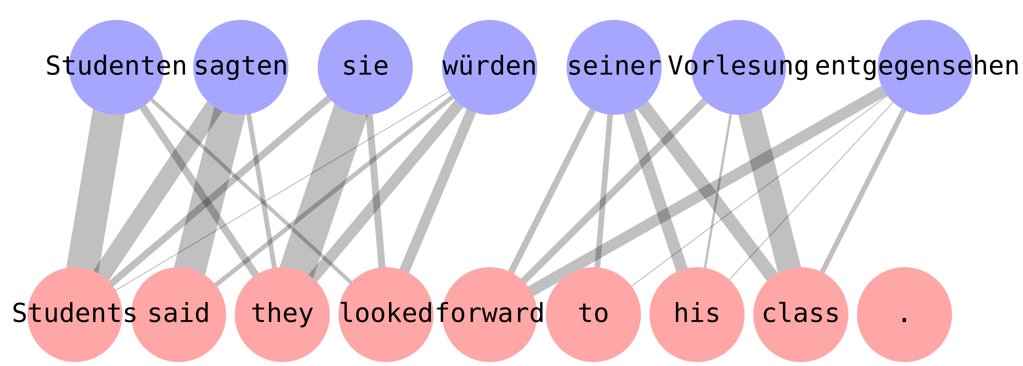
- **Input:** "Students say they looked forward to his class"
- **Explanations:**
  - Azure:



  - NMT:



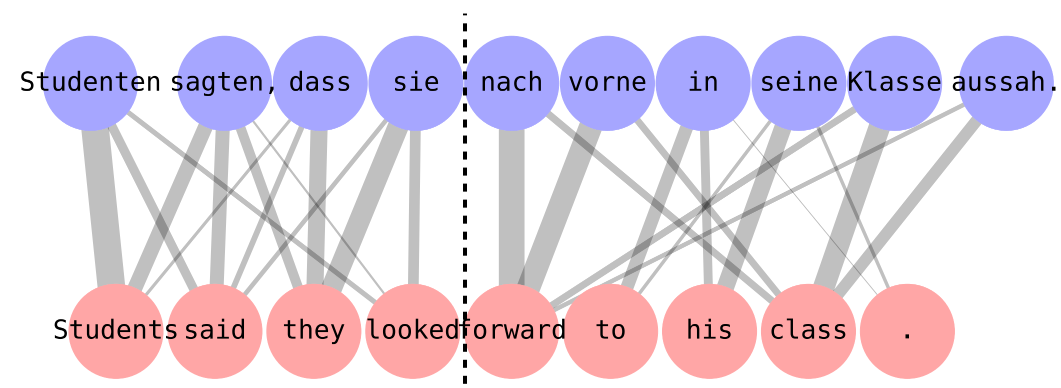
  - Human:



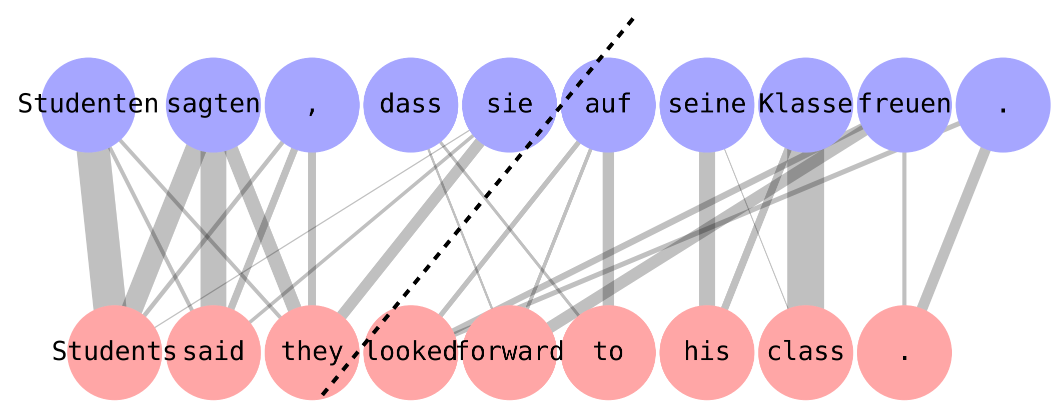
Application: Machine Translation

- **Input:** "Students say they looked forward to his class"

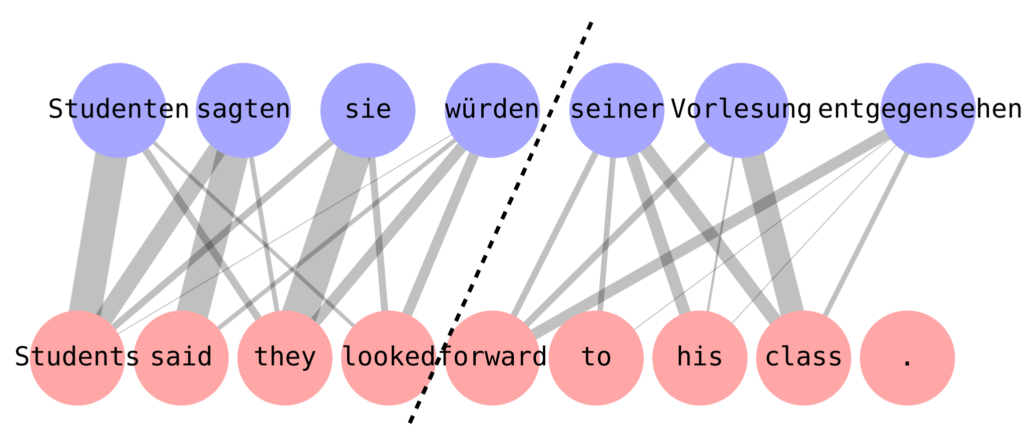
- **Explanations:**
  - Azure:



  - NMT:



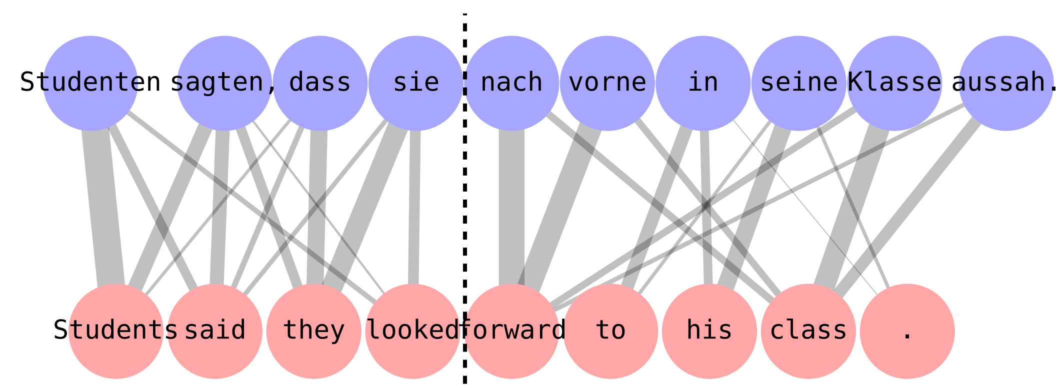
  - Human:



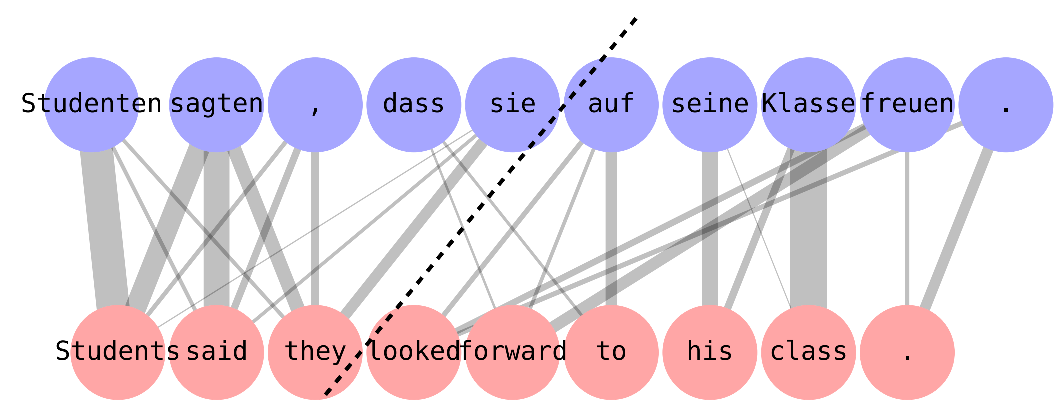
Application: Machine Translation

- **Input:** "Students say they looked forward to his class"

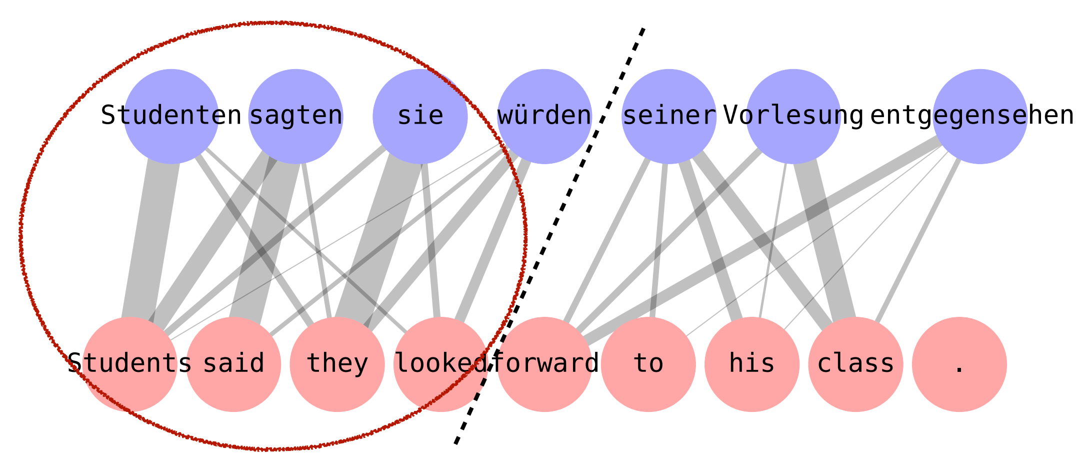
- **Explanations:**
  - Azure:



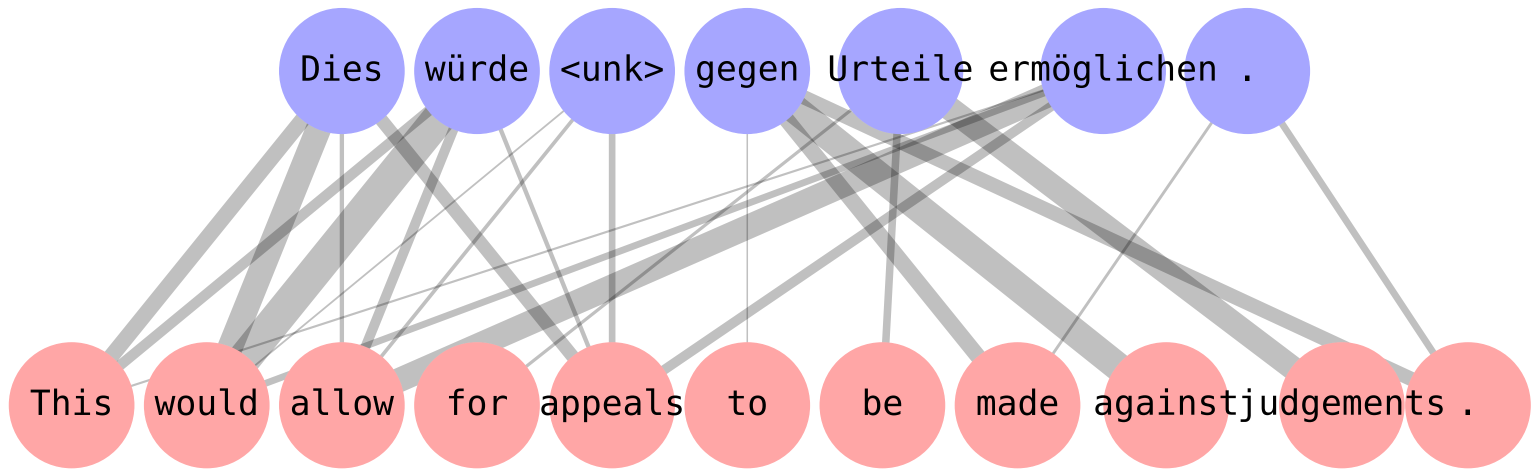
  - NMT:



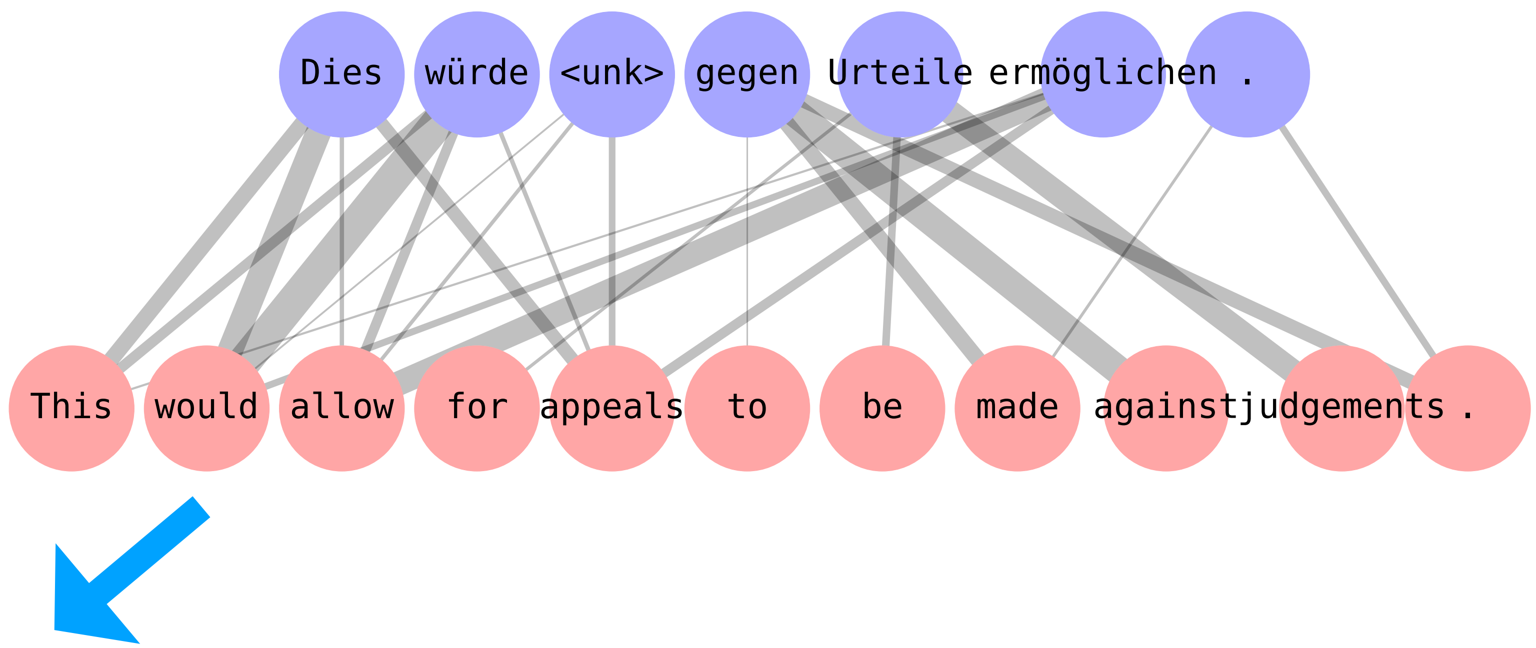
  - Human:



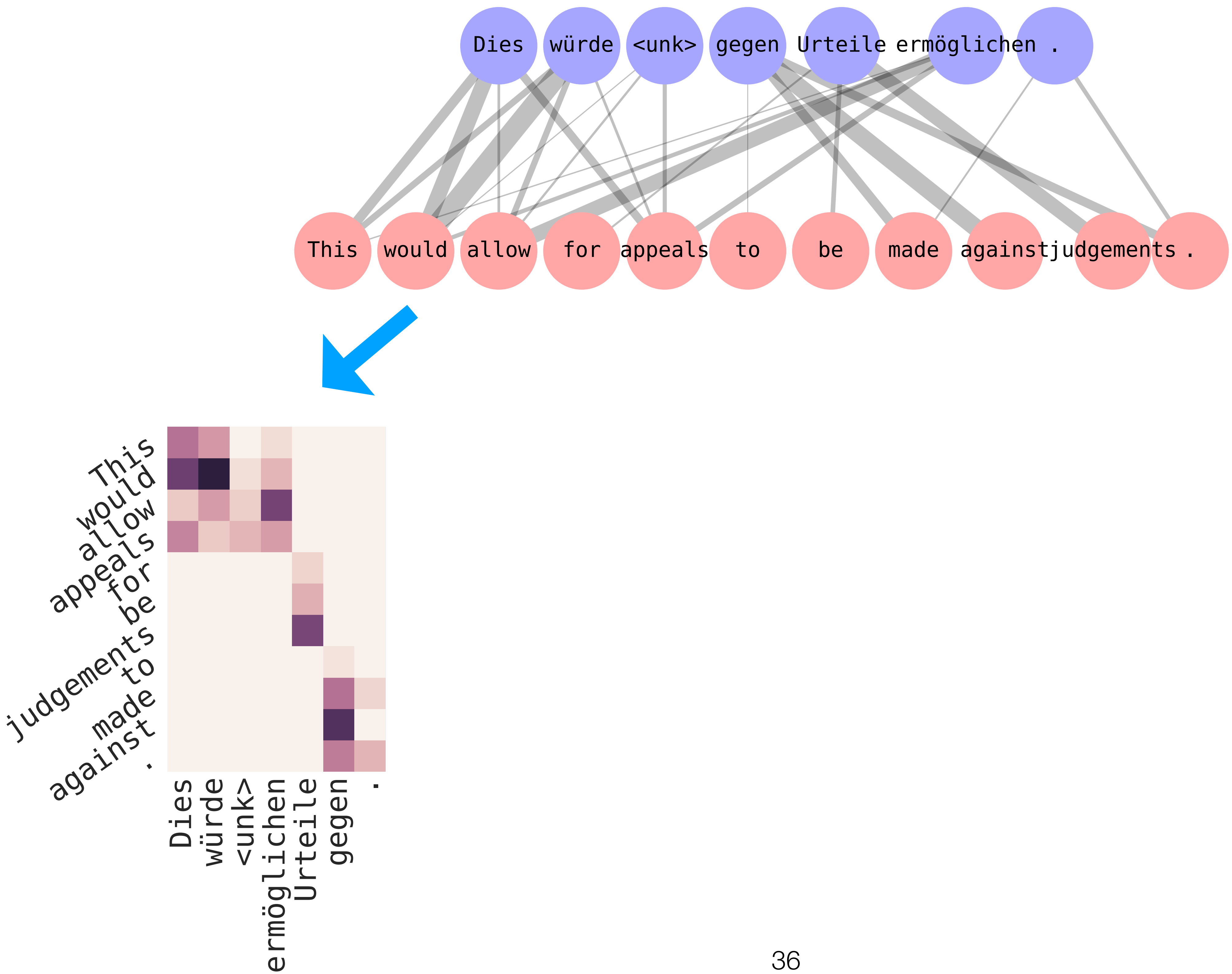
Dependency weights vs attention



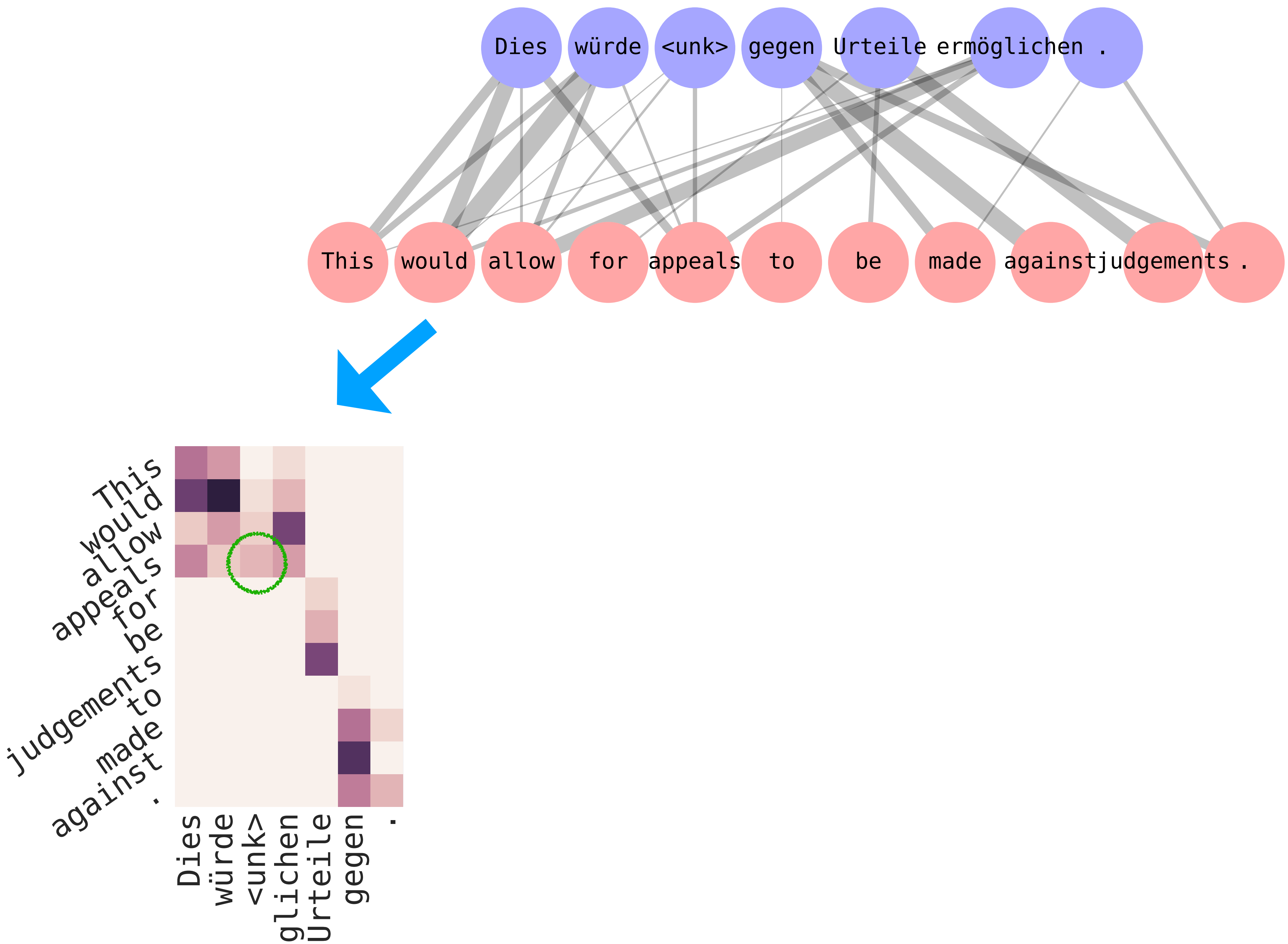
Dependency weights vs attention



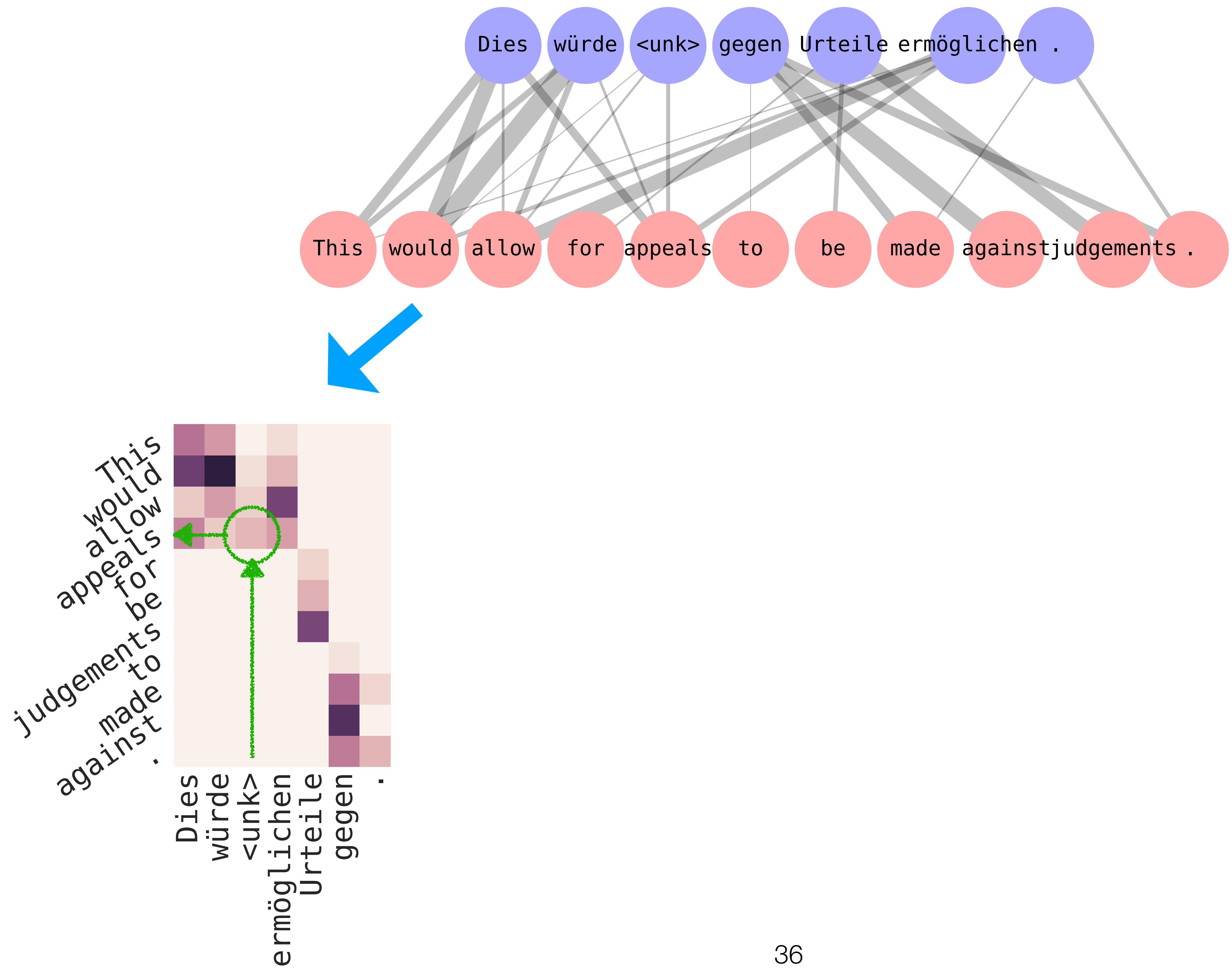
Dependency weights vs attention



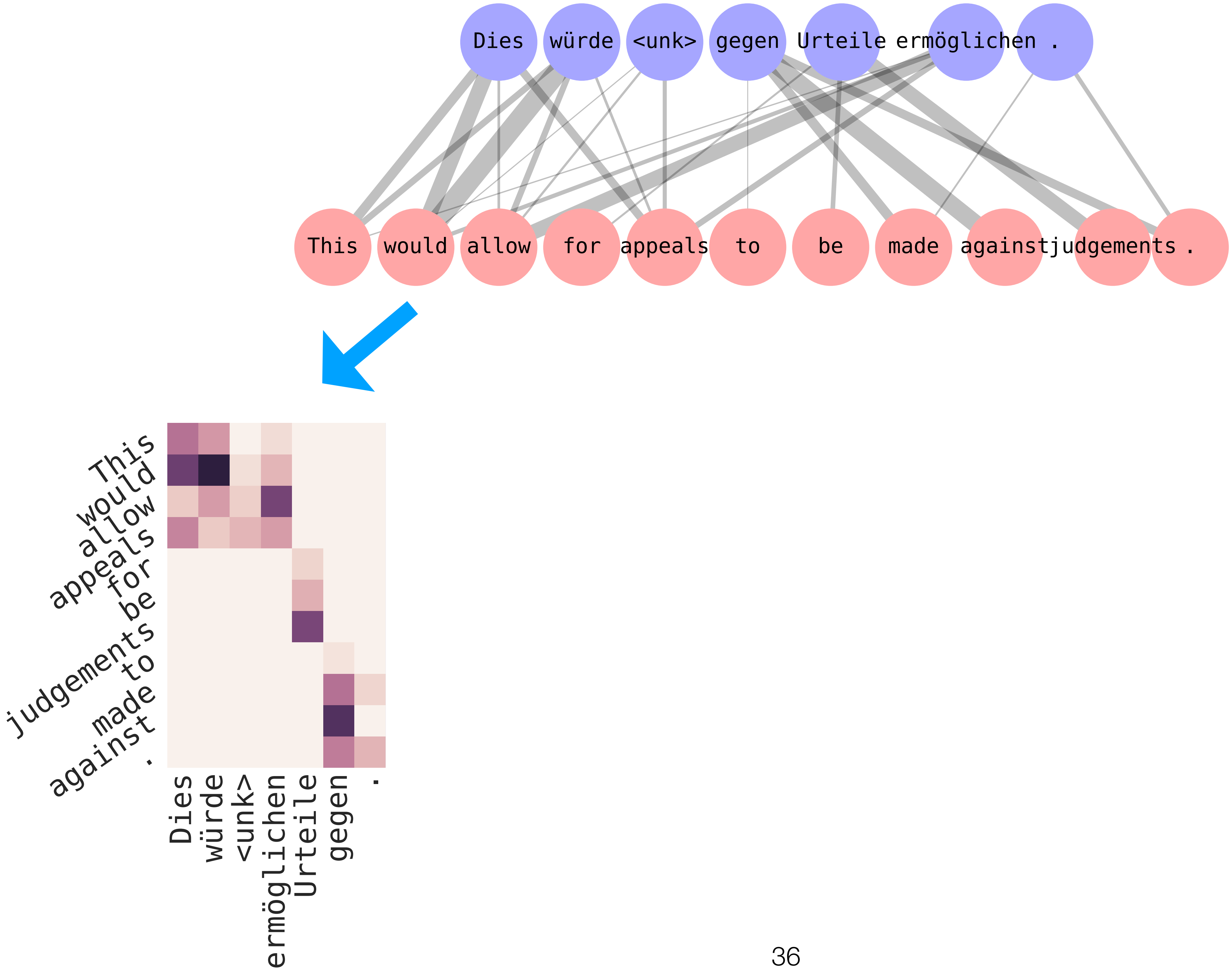
Dependency weights vs attention



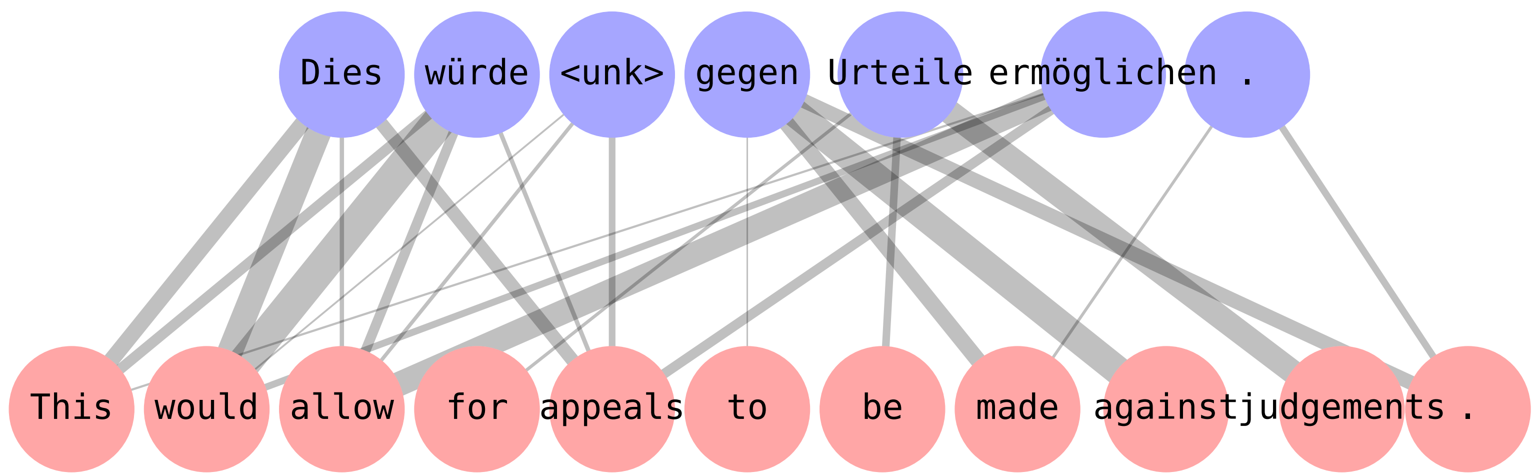
Dependency weights vs attention

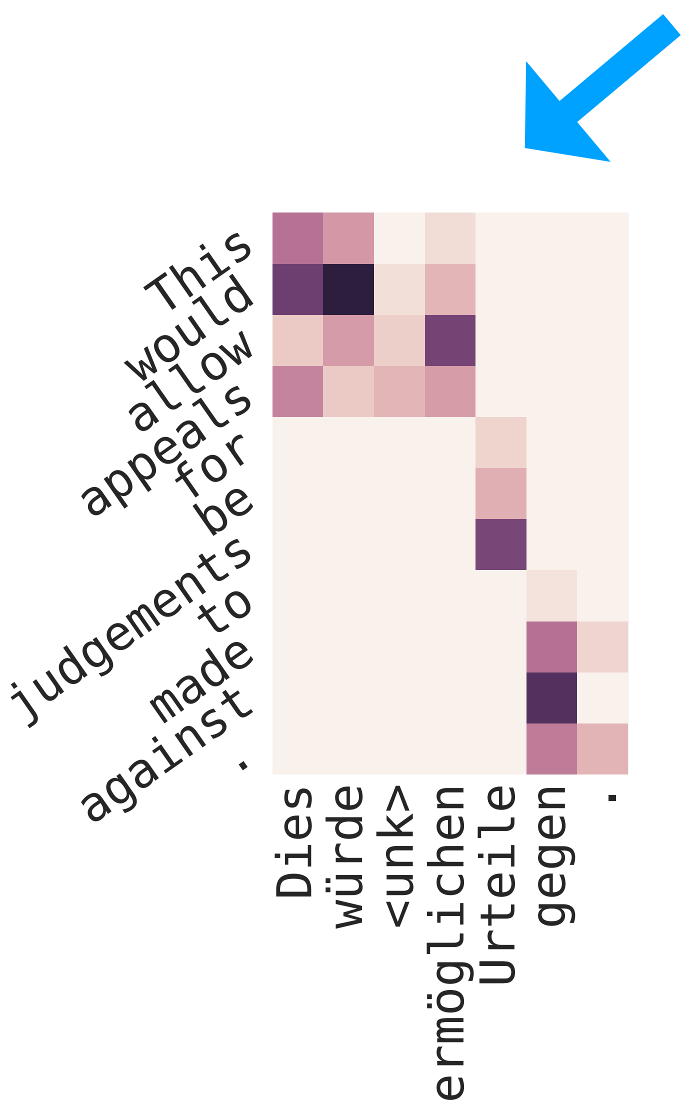


Dependency weights vs attention



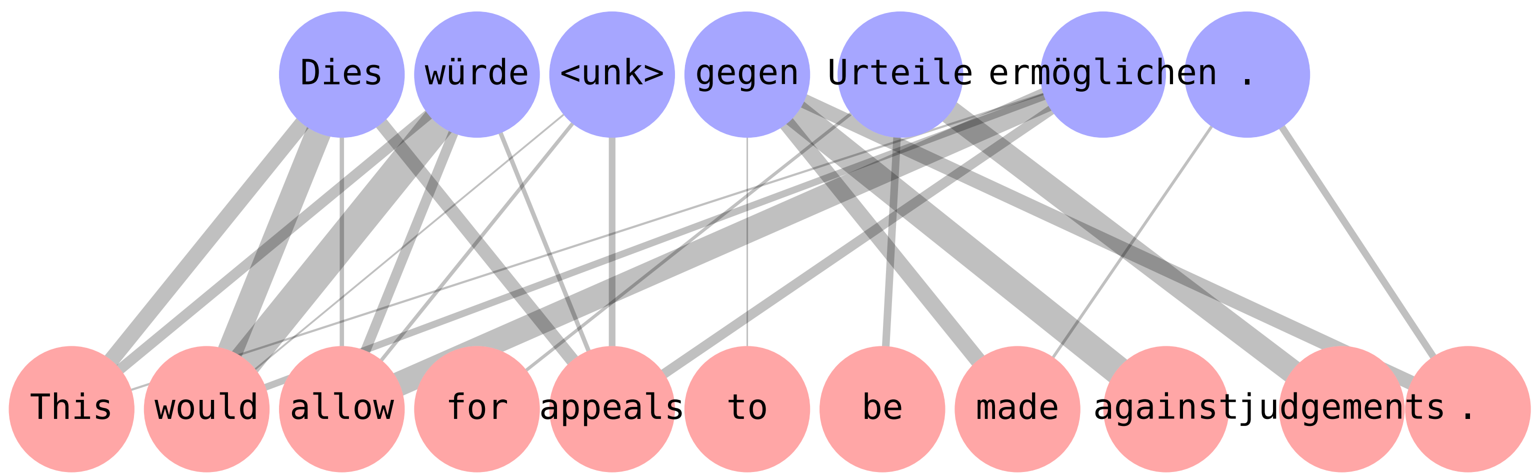
Dependency weights vs attention

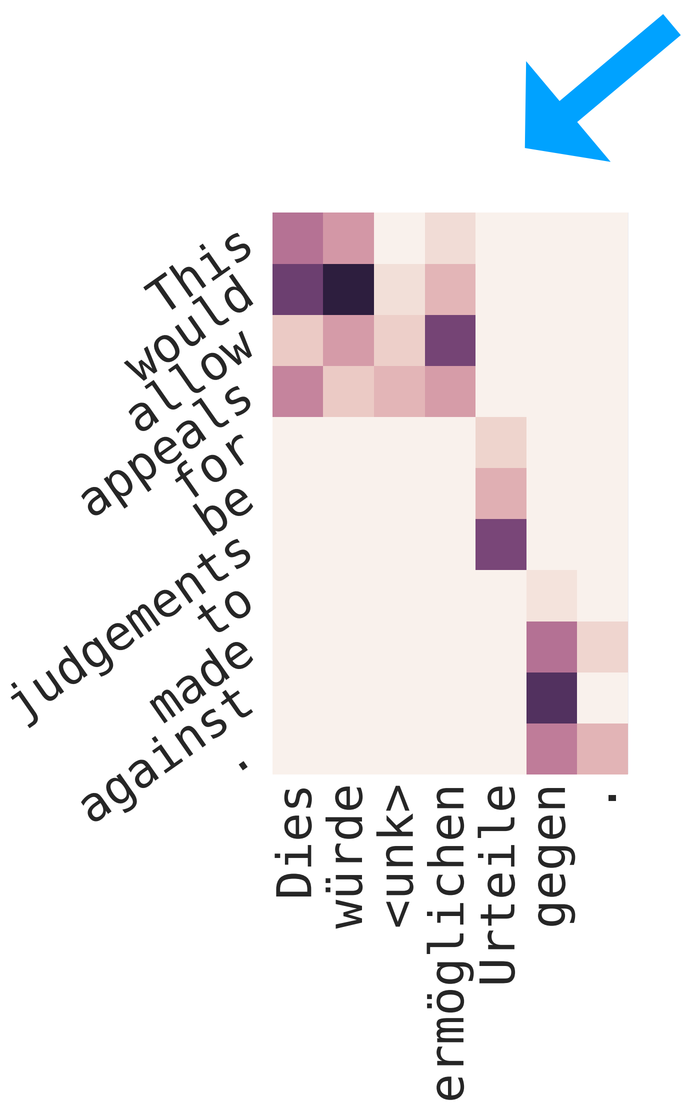




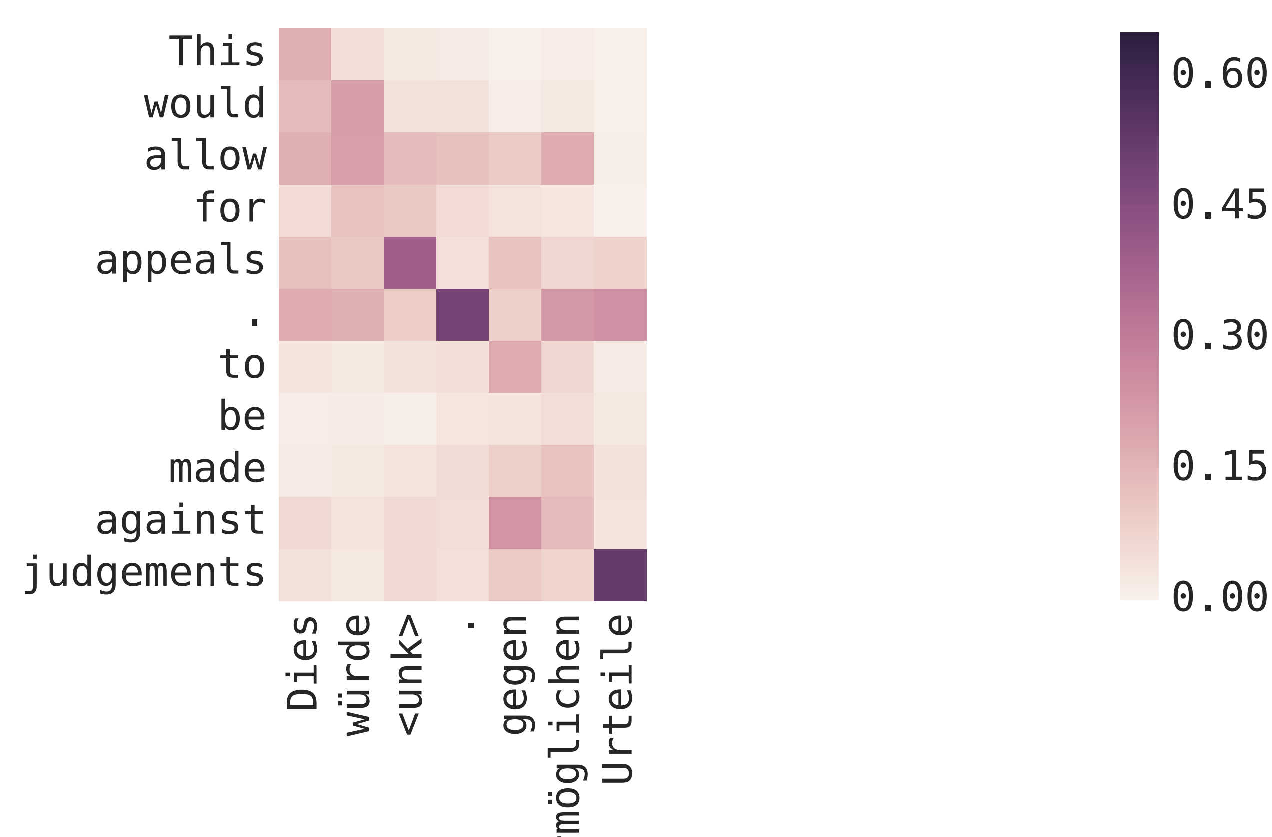
Actual Model's Attention

Dependency weights vs attention

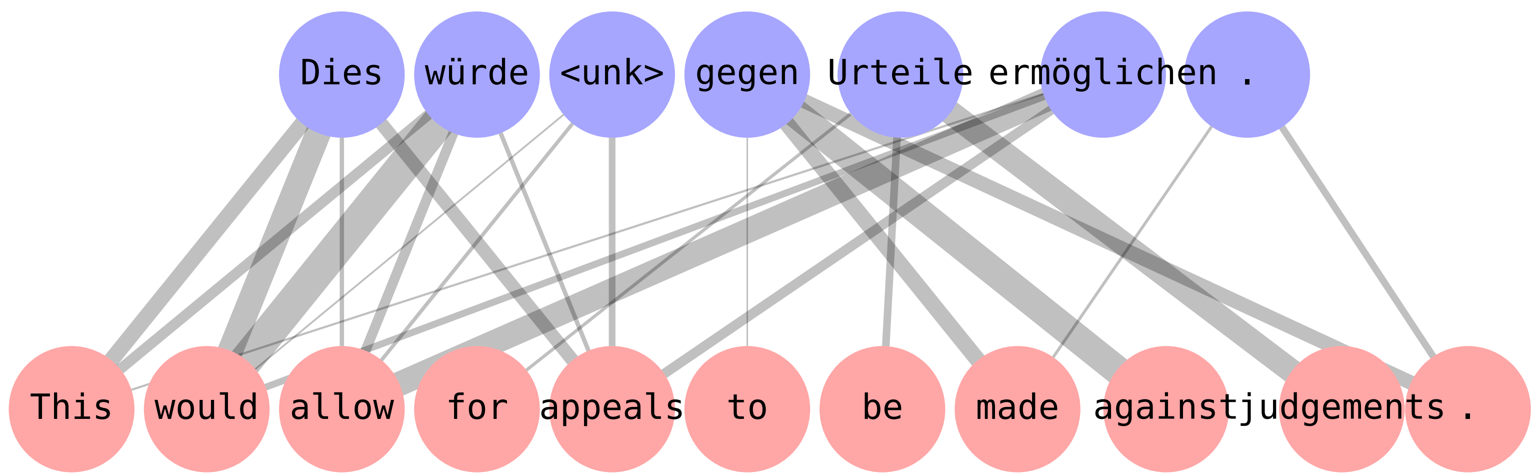


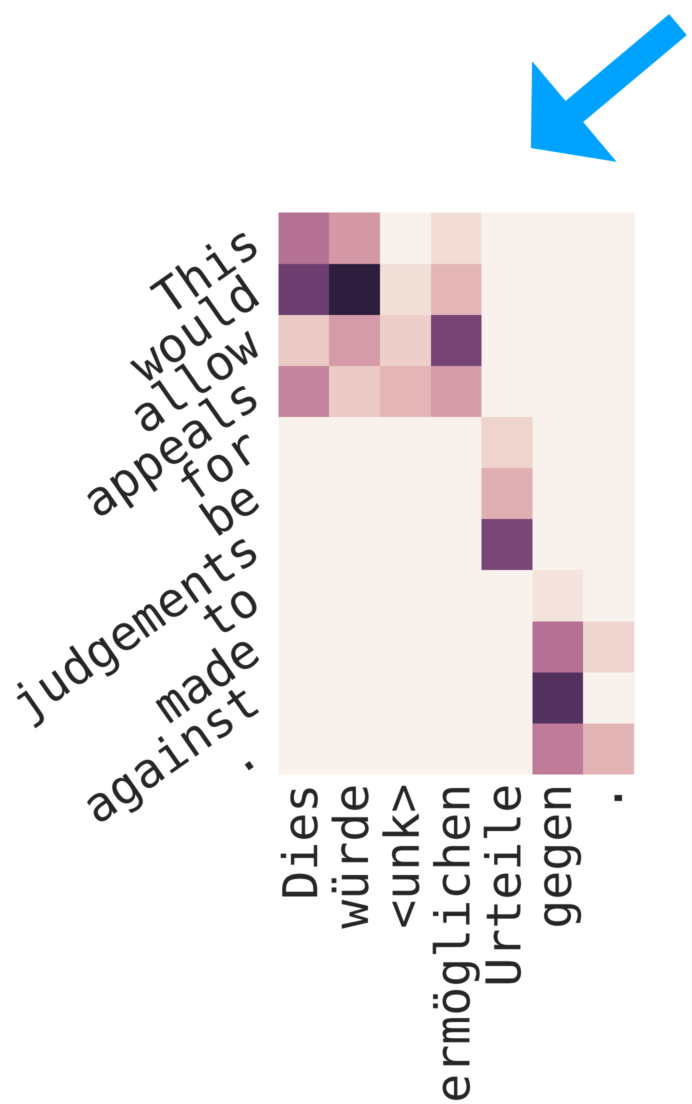


Actual Model's Attention

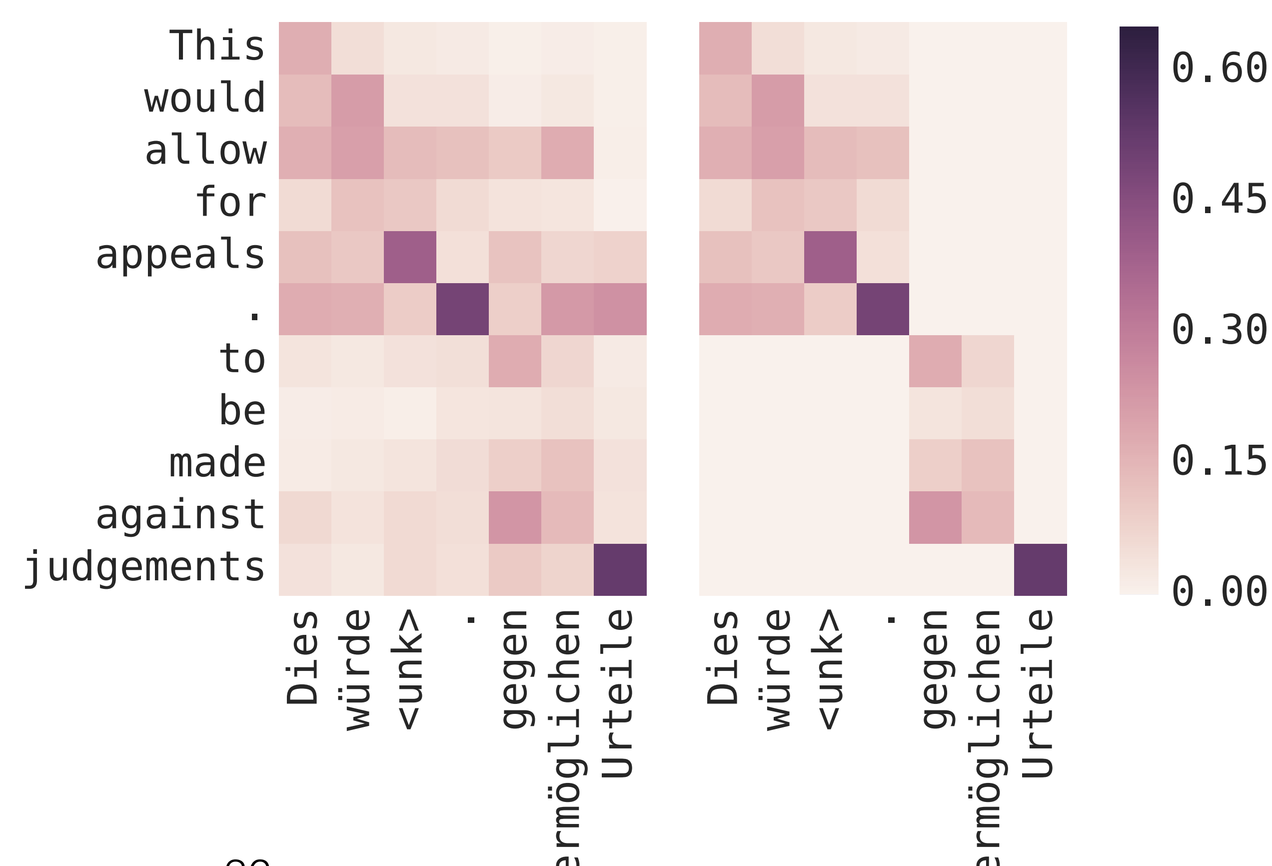


Dependency weights vs attention





Actual Model's Attention



Application: Bias detection in MT

Application: Bias detection in MT

- NLP methods tend to incorporate biases present in their training data
  - Archaic gender ↔ occupation stereotypes [Caliskan et al. 2017]
  - Sexist adjective associations [Bolukbasi et al. 2016]

Application: Bias detection in MT

- NLP methods tend to incorporate biases present in their training data
  - Archaic gender ↔ occupation stereotypes [Caliskan et al. 2017]
  - Sexist adjective associations [Bolukbasi et al. 2016]
- Can we use our interpretability framework to detect and understand these biases?

Application: Bias detection in MT

Application: Bias detection in MT

- **Black-box:** MSFT Azure's MT service, English $\longrightarrow$ French

Application: Bias detection in MT

- **Black-box:** MSFT Azure's MT service, English $\longrightarrow$ French
- **Inputs:** sentences containing bias-prone words

Application: Bias detection in MT

- **Black-box:** MSFT Azure's MT service, English $\longrightarrow$ French
- **Inputs:** sentences containing bias-prone words
- Our findings: model exhibits strong grammatical gender preferences

CSAIL

Application: Bias detection in MT

- **Black-box:** MSFT Azure's MT service, English ⟶ French
- **Inputs:** sentences containing bias-prone words
- Our findings: model exhibits strong grammatical gender preferences
- Chooses **masculine** in sentences containing *doctor*, *professor, smart, talented*

Application: Bias detection in MT

- **Black-box:** MSFT Azure's MT service, English ⟶ French
- **Inputs:** sentences containing bias-prone words
- Our findings: model exhibits strong grammatical gender preferences
- Chooses **masculine** in sentences containing *doctor*, *professor, smart, talented*
- Chooses **feminine** in sentences containing *dancer, nurse, charming, compassionate*

Application: Bias detection in MT

Application: Bias detection in MT

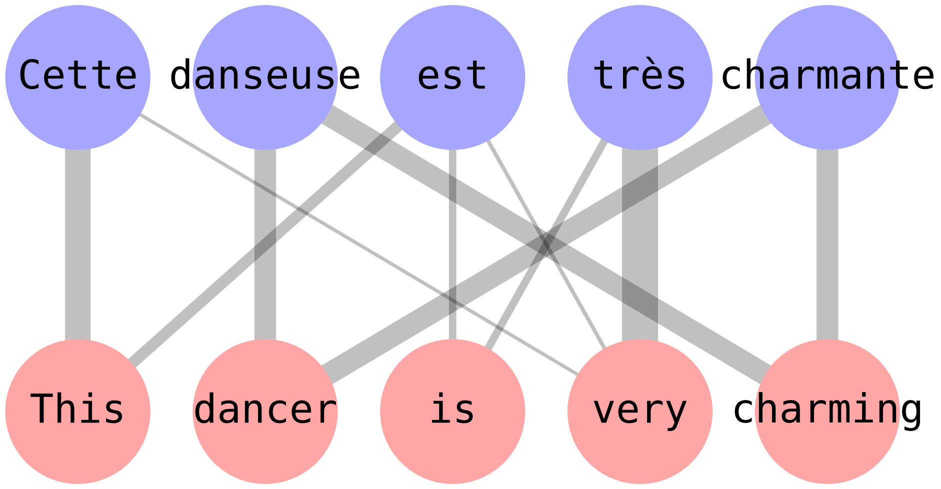
- **Black-box:** MSFT Azure's MT service, English $\longrightarrow$ French

Application: Bias detection in MT

- **Black-box:** MSFT Azure's MT service, English $\longrightarrow$ French
- **Inputs:** sentences containing bias-prone words

Application: Bias detection in MT

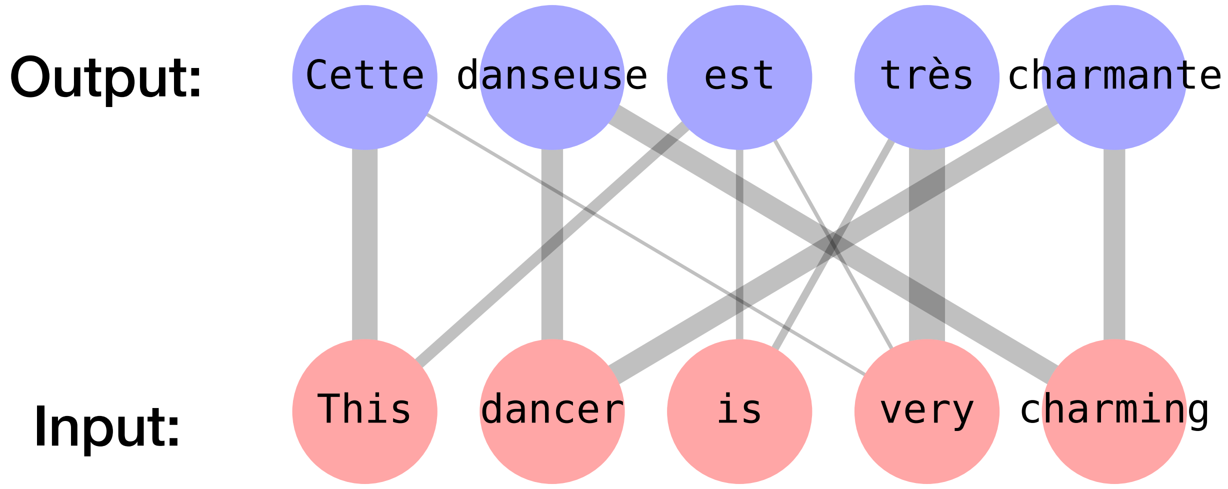
- **Black-box:** MSFT Azure's MT service, English ⟶ French
- **Inputs:** sentences containing bias-prone words



Application: Bias detection in MT

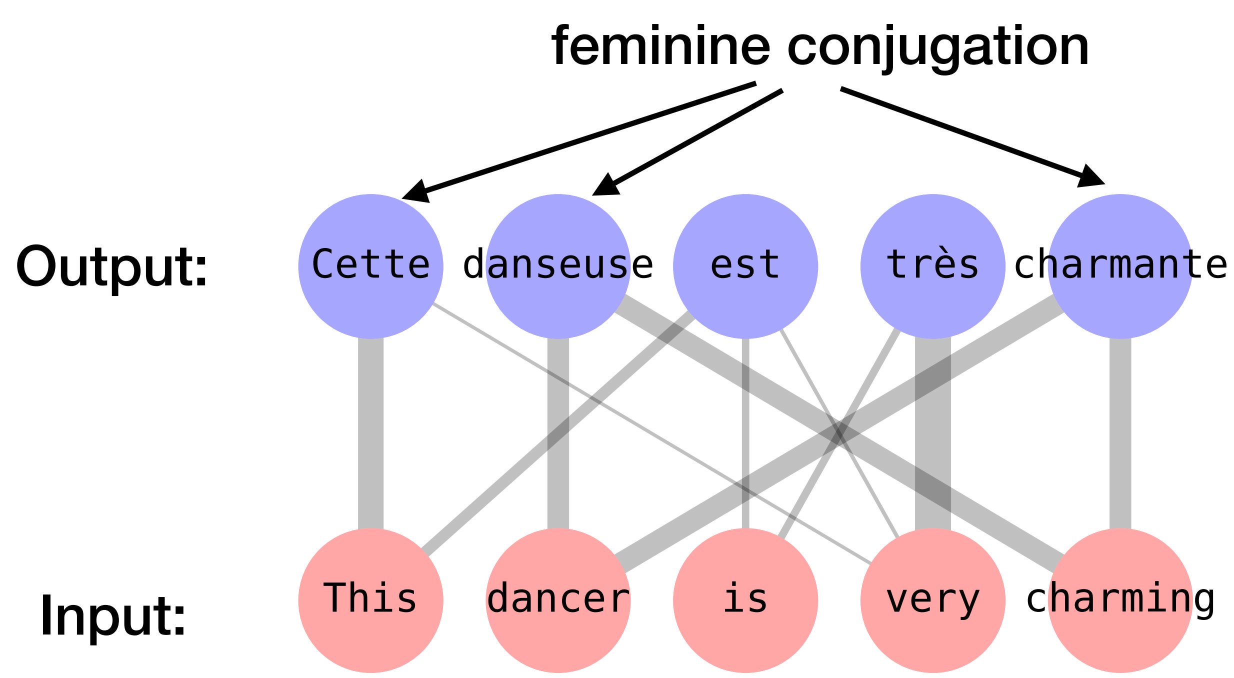
- **Black-box:** MSFT Azure's MT service, English ⟶ French

- **Inputs:** sentences containing bias-prone words



Application: Bias detection in MT

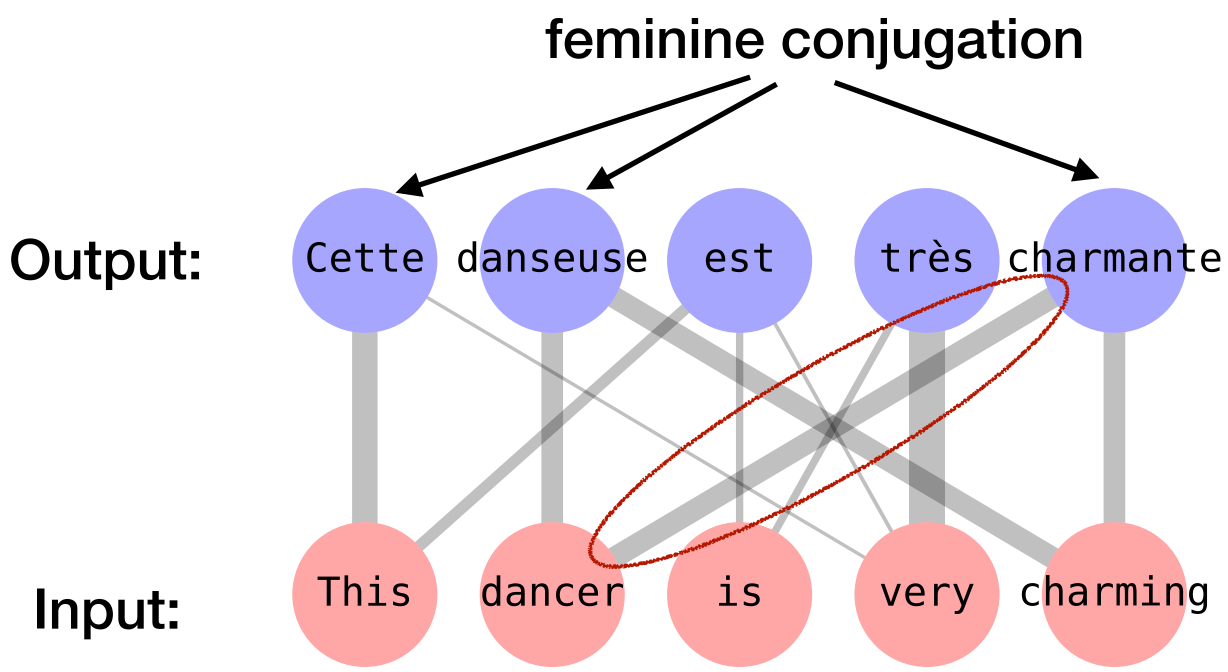
- **Black-box:** MSFT Azure's MT service, English ⟶ French
- **Inputs:** sentences containing bias-prone words



Application: Bias detection in MT

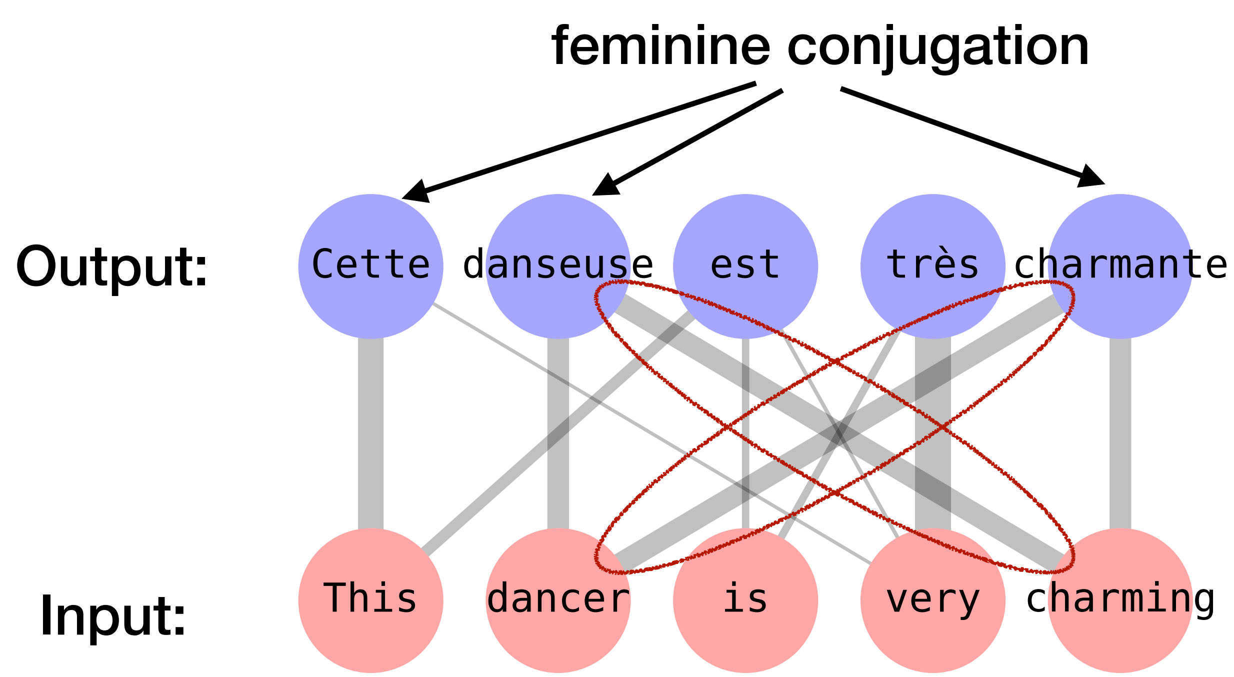
- **Black-box:** MSFT Azure's MT service, English ⟶ French

- **Inputs:** sentences containing bias-prone words



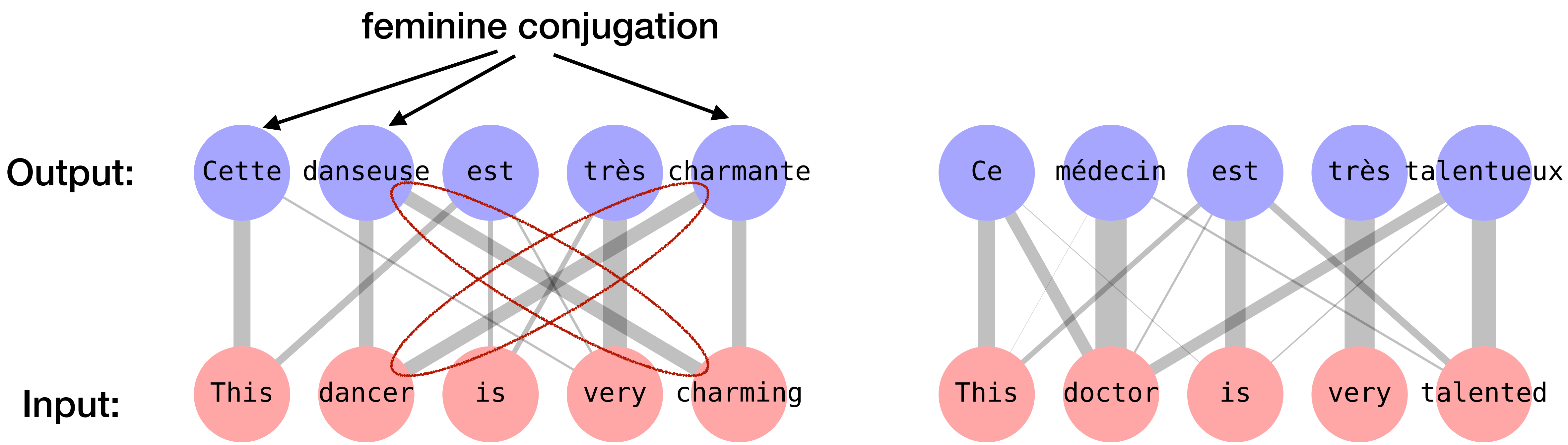
Application: Bias detection in MT

- **Black-box:** MSFT Azure's MT service, English ⟶ French
- **Inputs:** sentences containing bias-prone words



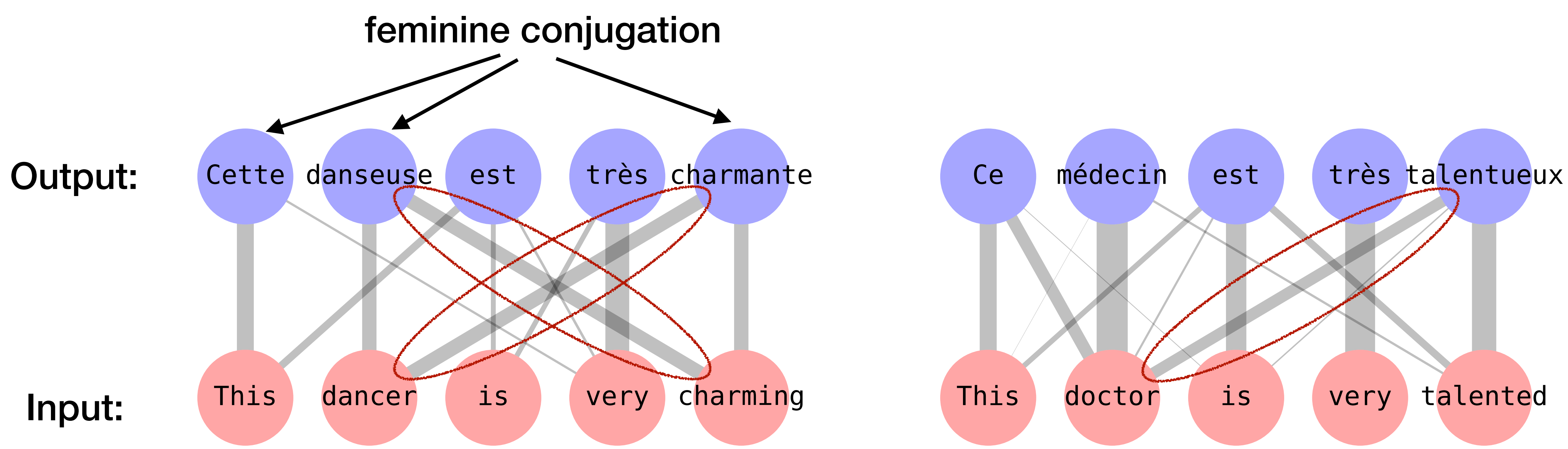
Application: Bias detection in MT

- **Black-box:** MSFT Azure's MT service, English ⟶ French
- **Inputs:** sentences containing bias-prone words



Application: Bias detection in MT

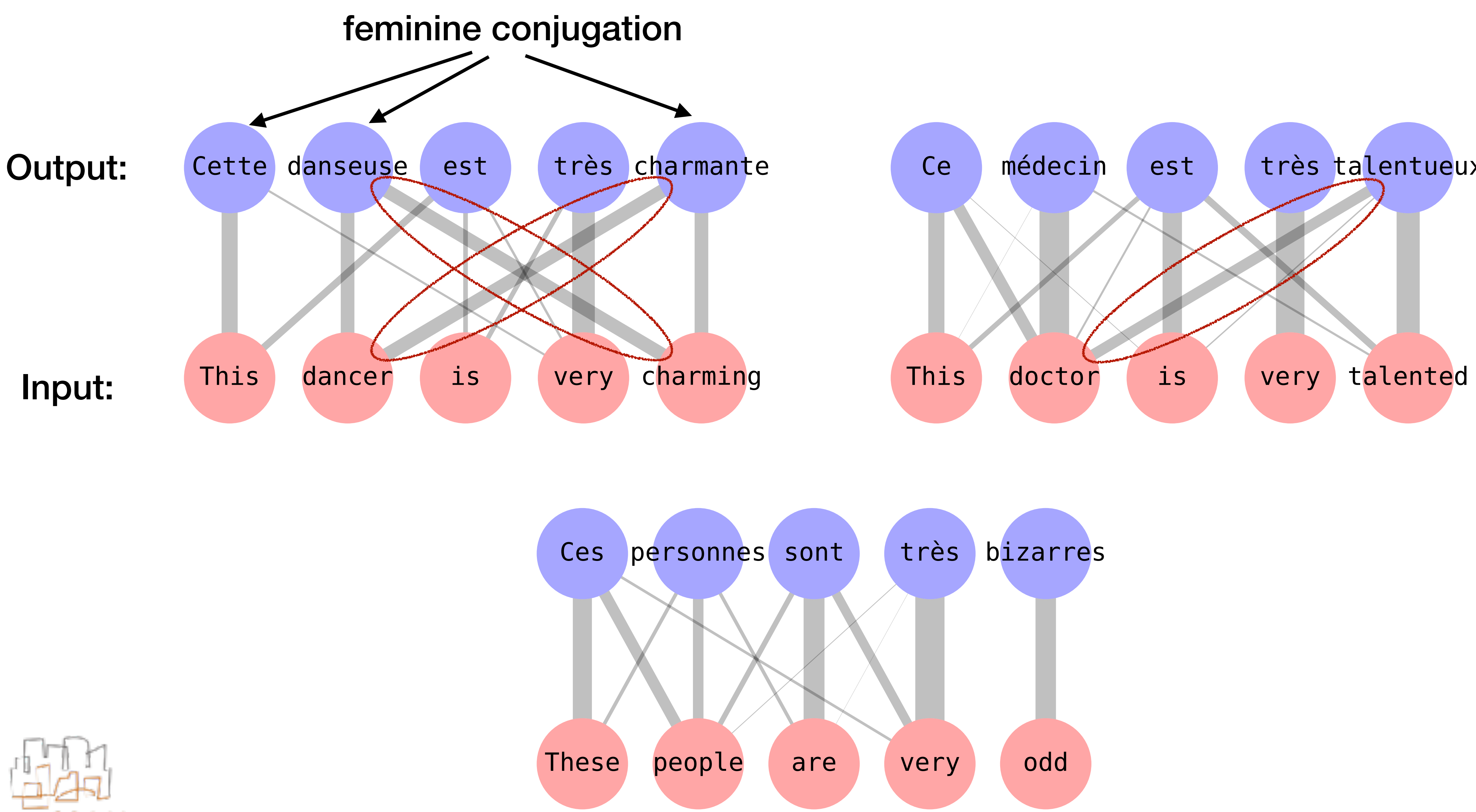
- **Black-box:** MSFT Azure's MT service, English ⟶ French
- **Inputs:** sentences containing bias-prone words



Application: Bias detection in MT

- **Black-box:** MSFT Azure's MT service, English ⟶ French

- **Inputs:** sentences containing bias-prone words



Application: Flaw detection in dialogue systems

Application: Flaw detection in dialogue systems

- Interpretability can help us verify if model has learned "properly"

Application: Flaw detection in dialogue systems

- Interpretability can help us verify if model has learned "properly"
- We train a simple dialogue system on the OpenSubtitle corpus

Application: Flaw detection in dialogue systems

- Interpretability can help us verify if model has learned "properly"
- We train a simple dialogue system on the OpenSubtitle corpus
- ~14M two-step movie dialogues

Are you the son of Vito Corleone? → *Yes. I am.*

Input (query) → **Output (response)**

Application: Flaw detection in dialogue systems

- Interpretability can help us verify if model has learned "properly"
- We train a simple dialogue system on the OpenSubtitle corpus
- ~14M two-step movie dialogues

Are you the son of Vito Corleone? → *Yes. I am.*

Input (query) **Output (response)**

- **Black-box:** seq2seq with attention, 2 layers, dim 100, no tuning

Application: Flaw detection in dialogue systems

Application: Flaw detection in dialogue systems

- **Black-box:** seq2seq with attention, 2 layers, dim 100, no tuning

Application: Flaw detection in dialogue systems

- **Black-box:** seq2seq with attention, 2 layers, dim 100, no tuning

Application: Flaw detection in dialogue systems

- **Black-box:** seq2seq with attention, 2 layers, dim 100, no tuning

- Example actual predictions:

Input	Prediction
What do you mean it doesn't matter?	*I don't know*
Perhaps have we met before?	*I don't think so*
Can I get you two a cocktail?	*No, thanks.*

Application: Flaw detection in dialogue systems

- **Black-box:** seq2seq with attention, 2 layers, dim 100, no tuning

- Example actual predictions:

Input	Prediction
What do you mean it doesn't matter?	*I don't know*
Perhaps have we met before?	*I don't think so*
Can I get you two a cocktail?	*No, thanks.*

- Is the model good?

Application: Flaw detection in dialogue systems

Application: Flaw detection in dialogue systems

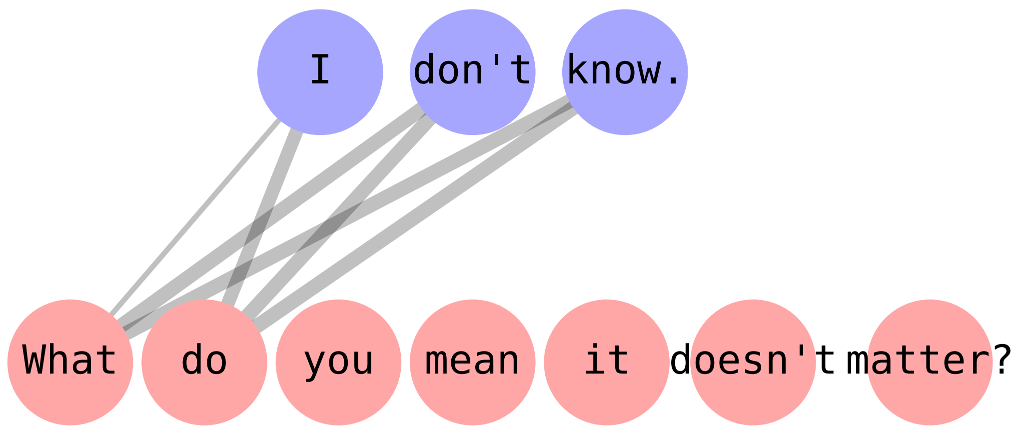
- **Input:** *What do you mean it doesn't matter?*

Application: Flaw detection in dialogue systems

- **Input:** *What do you mean it doesn't matter?*
- **Output:** *I don't know*

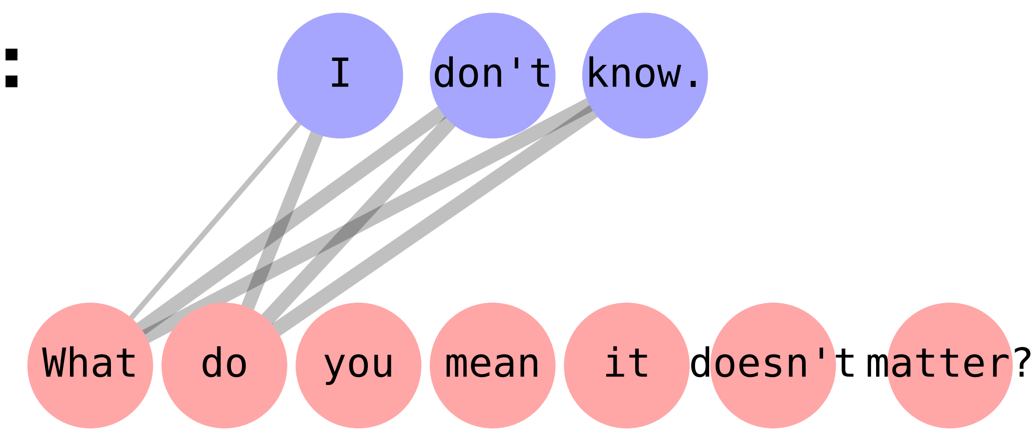
Application: Flaw detection in dialogue systems

- **Input:** *What do you mean it doesn't matter?*
- **Output:** *I don't know*
- **Explanation:**

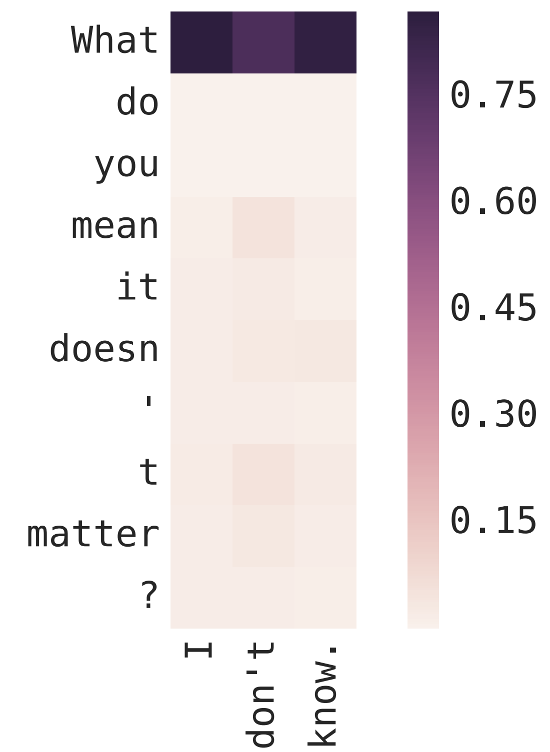


Application: Flaw detection in dialogue systems

- **Input:** *What do you mean it doesn't matter?*
- **Output:** *I don't know*
- **Explanation:**

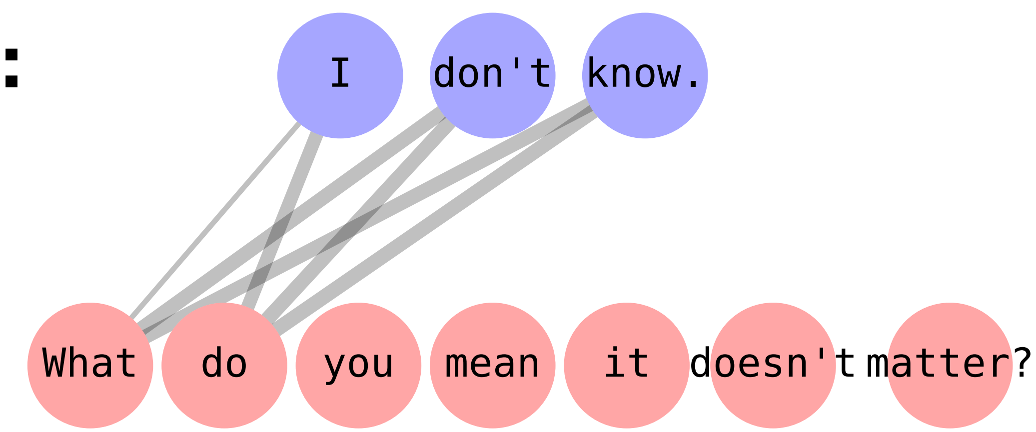


- **Actual attention scores:**

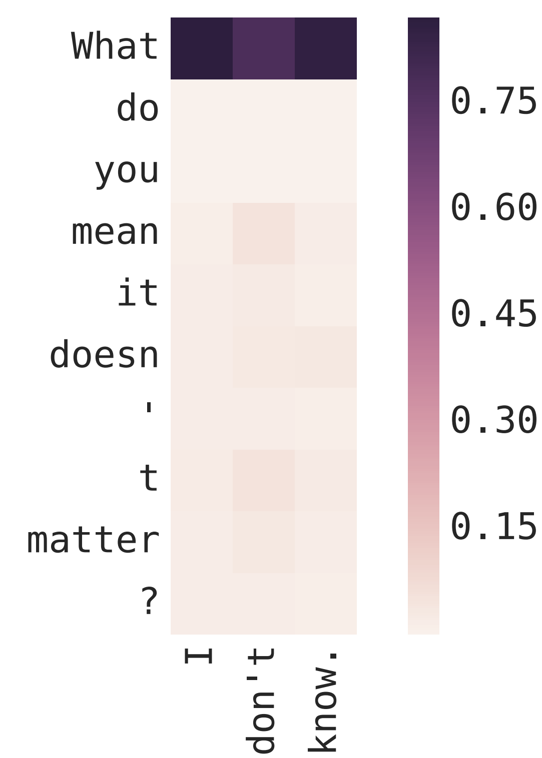


Application: Flaw detection in dialogue systems

- **Input:** *What do you mean it doesn't matter?*
- **Output:** *I don't know*
- **Explanation:**



- **Actual attention scores:**



The model is flawed!

Summary

- Interpretability framework for structured-data models (not only sentences!)

- Works directly on inputs/outputs, model-agnostic

- Experiments show how explanations yield partial view into inner workings of black-box systems

Discussion

- As with most interpretability frameworks, assumes uncorrelated inputs - strong assumption
- Can we enhance the probabilistic modeling to account for this?
- Can we prove reconstruction guarantees in some form?

Epilogue

- Various approaches to interpretability in NLP in the last year:
  - [Arras et al. 2017]: uses Layer-wise Relevance Propagation
  - [Sundararajan et al 2017]: integrated gradients, applications to MT
  - [Murdoch et al. 2018]: decompose nonlinearities in LSTM via telescoping sums, analyze "focalized" contributions of subsets of the input

References

- Arras et al."Explaining Predictions of Non-Linear Classifiers in NLP", ACL Workshop on Representation Learning for NLP, 2016.

- Arras et al. "Explaining Recurrent Neural Network Predictions in Sentiment Analysis", EMNLP Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2017.

- Murdoch, Liu and Yu. "Beyond Word Importance: Contextual Decomposition To Extract Interactions From LSTMs", ICLR 2018

Self-explaining neural networks

[A-M & Jaakkola, in progress]

Self explaining neural nets

Self explaining neural nets

- Current gradient-based methods require additional computation / optimization

Self explaining neural nets

- Current gradient-based methods require additional computation / optimization

- Can we get explanations as a *byproduct* of computation?

Self explaining neural nets

- Current gradient-based methods require additional computation / optimization
- Can we get explanations as a *byproduct* of computation?
- ... with minimal architectural modification?

Self explaining neural nets

- Current gradient-based methods require additional computation / optimization
- Can we get explanations as a *byproduct* of computation?
- ... with minimal architectural modification?
- Our approach: hybrid simple-complex models

CSAIL

Interpretability: linear and beyond

Interpretability: linear and beyond

- The archetypical interpretable model:

$$f(x) = \sum_i^n \theta_i x_i + \theta_0$$

Interpretability: linear and beyond

- The archetypical interpretable model:

$$f(x) = \sum_{i}^{n} \theta_i x_i + \theta_0$$

- What makes it interpretable?

Interpretability: linear and beyond

- The archetypical interpretable model:

$$f(x) = \sum_i^n \theta_i x_i + \theta_0$$

- What makes it interpretable?

    1. Inputs are clearly anchored - interpretable quantities

Interpretability: linear and beyond

- The archetypical interpretable model:

$$f(x) = \sum_{i}^{n} \theta_i x_i + \theta_0$$

- What makes it interpretable?

1. Inputs are clearly anchored - interpretable quantities
2. Parameters -> (signed) contribution of each feature

Interpretability: linear and beyond

- The archetypical interpretable model:

$$f(x) = \sum_{i}^{n} \theta_i x_i + \theta_0$$

- What makes it interpretable?

  1. Inputs are clearly anchored - interpretable quantities

  2. Parameters -> (signed) contribution of each feature

  3. Simple aggregation function (sum)

Interpretability: linear and beyond

- The archetypical interpretable model:

$$f(x) = \sum_{i}^{n} \theta_i x_i + \theta_0$$

- What makes it interpretable?

  1. Inputs are clearly anchored - interpretable quantities

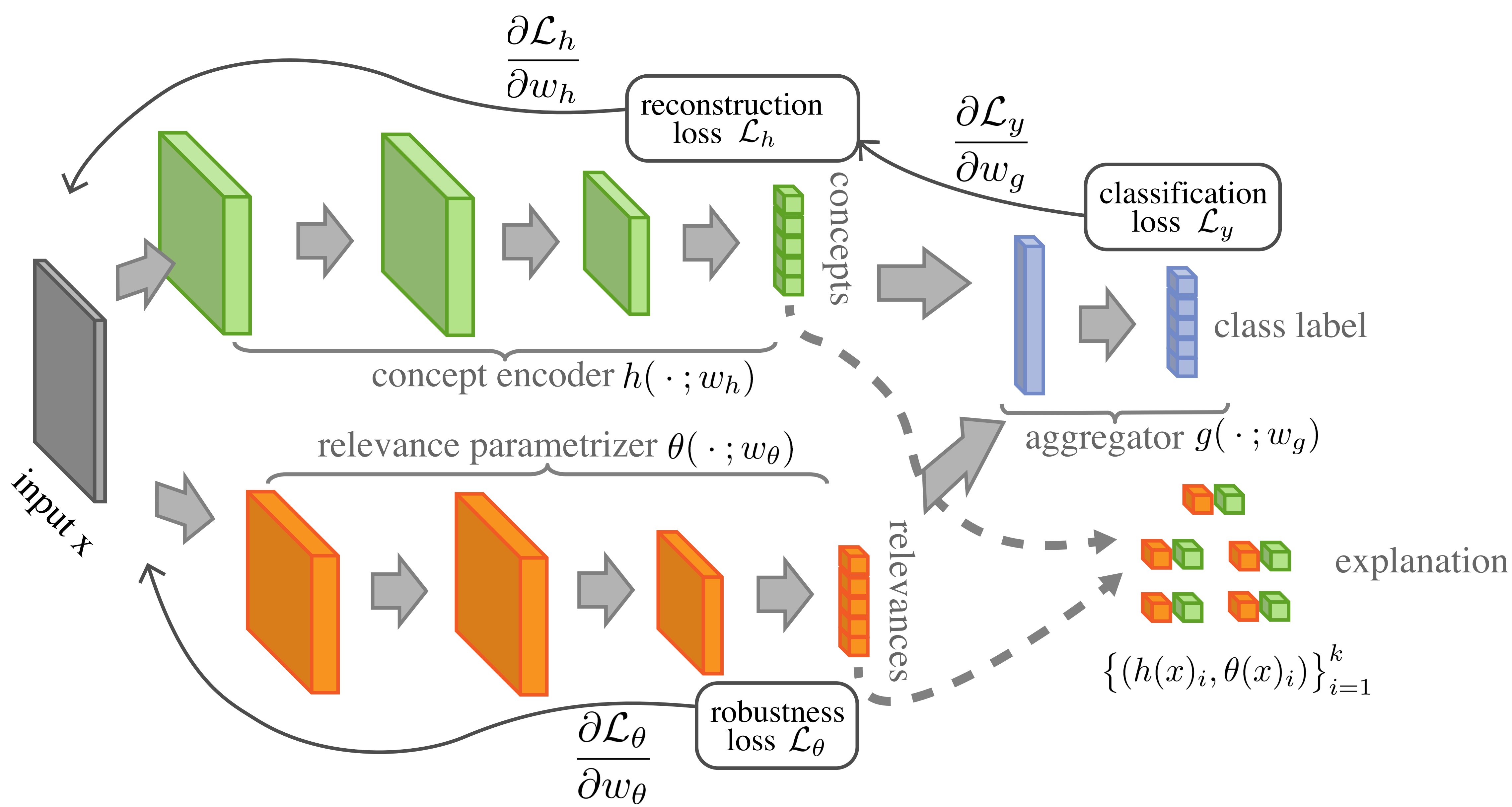
  2. Parameters -> (signed) contribution of each feature

  3. Simple aggregation function (sum)

- How much can we generalize the model without losing (1)-(3)?

Self-explaining models

$$f(\mathbf{x}) = g(\theta_1(x)h_1(x), \ldots, \theta_k(x)h_k(x))$$



Explaining MNIST via inputs

- Surface model: linear, parameter model: CNN
- MNIST dataset

Explaining MNIST via inputs

- Surface model: linear, parameter model: CNN

$$f(x) = \text{softmax}(\theta(x)^T x)$$

- MNIST dataset

Explaining MNIST via inputs

- Surface model: linear, parameter model: CNN

$$f(x) = \text{softmax}(\theta(x)^T x)$$

- MNIST dataset

CNN (LeNet)

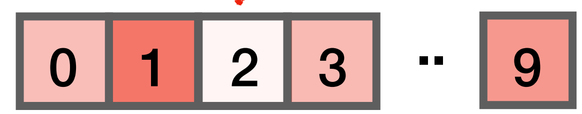
Explaining MNIST via inputs

- Surface model: linear, parameter model: CNN

$$f(x) = \text{softmax}(\theta(x)^T x)$$

CNN (LeNet)

- MNIST dataset



Explaining MNIST via inputs

- Surface model: linear, parameter model: CNN

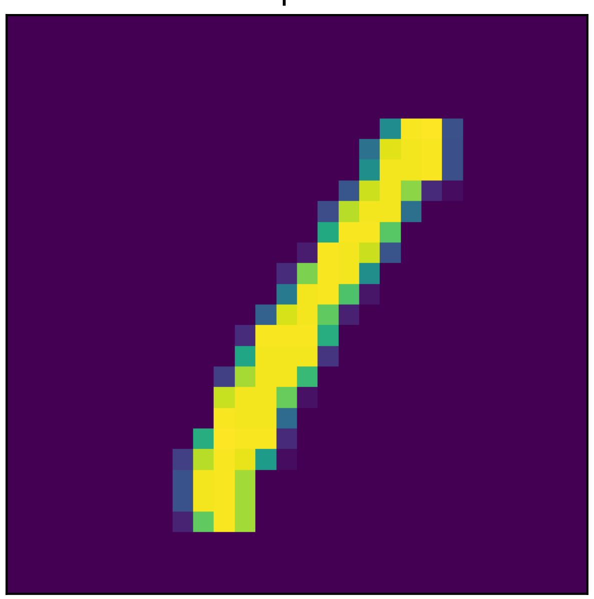
$$f(x) = \text{softmax}(\theta(x)^T x)$$

CNN (LeNet)

- MNIST dataset

0	1	2	3	..	9

Input:



Explaining MNIST via inputs

- Surface model: linear, parameter model: CNN

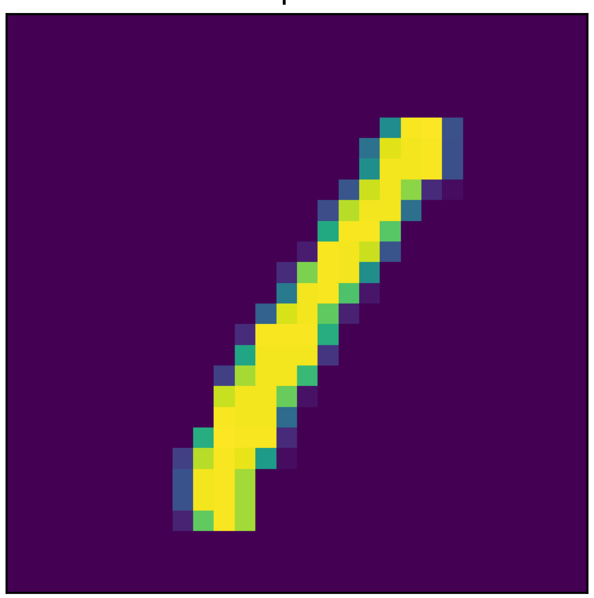
$$f(x) = \text{softmax}(\theta(x)^T x)$$

CNN (LeNet)

0	1	2	3	··	9

- MNIST dataset

Input:



Predicted class: 1

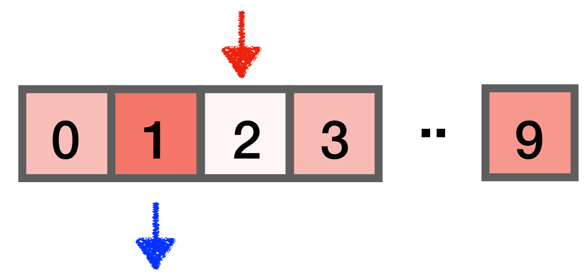
Explaining MNIST via inputs

- Surface model: linear, parameter model: CNN

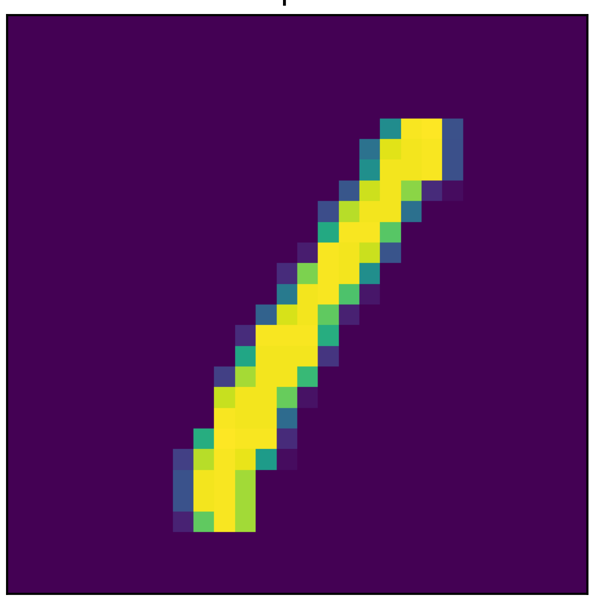
$$f(x) = \mathrm{softmax}(\theta(x)^T x)$$

CNN (LeNet)

- MNIST dataset



Input:



Predicted class: 1

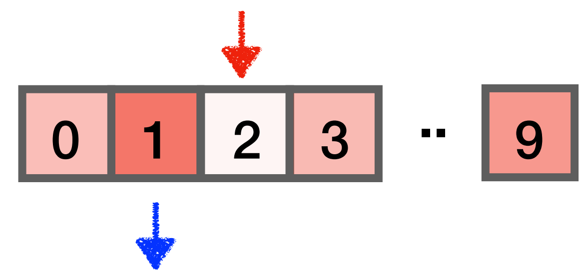
Explaining MNIST via inputs

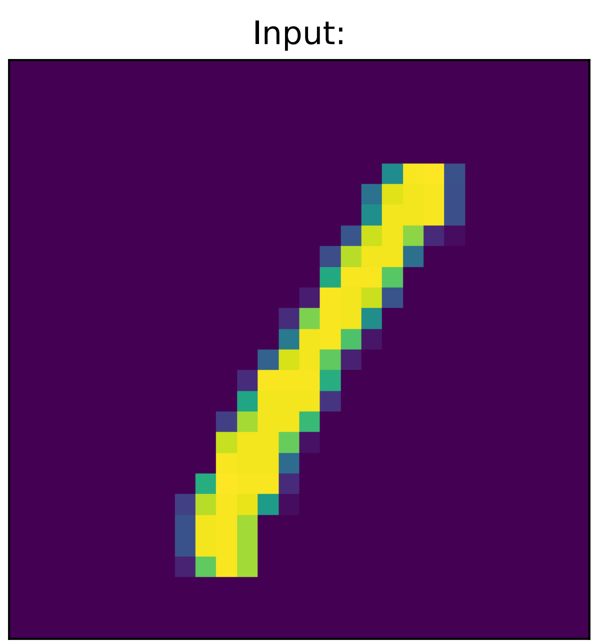
- Surface model: linear, parameter model: CNN

$$f(x) = \text{softmax}(\theta(x)^T x)$$

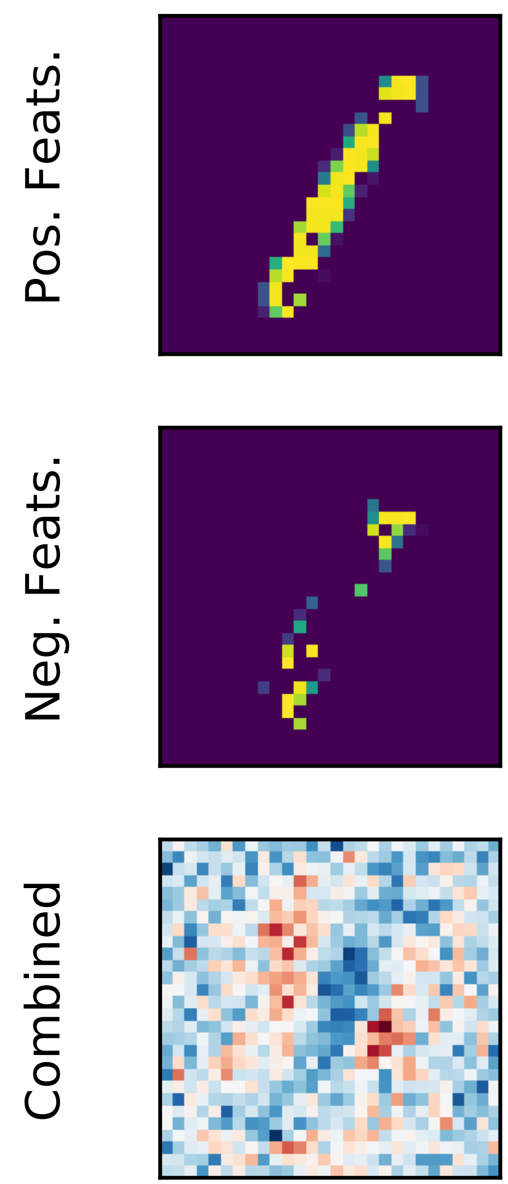
CNN (LeNet)

- MNIST dataset





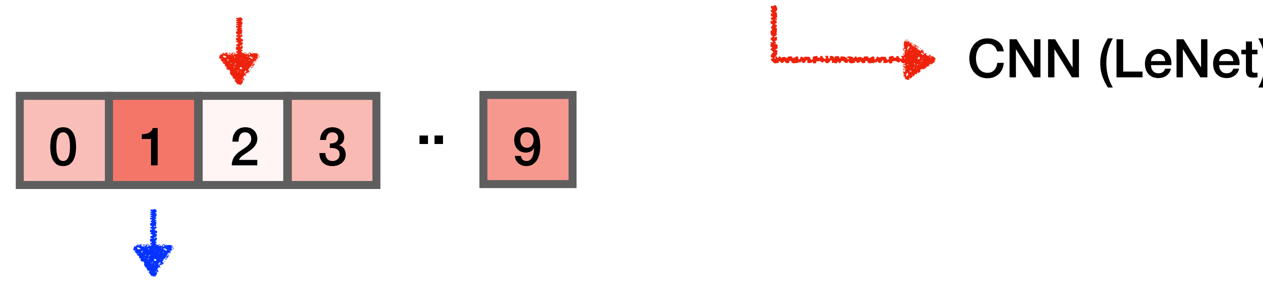
Predicted class: 1



Explaining MNIST via inputs

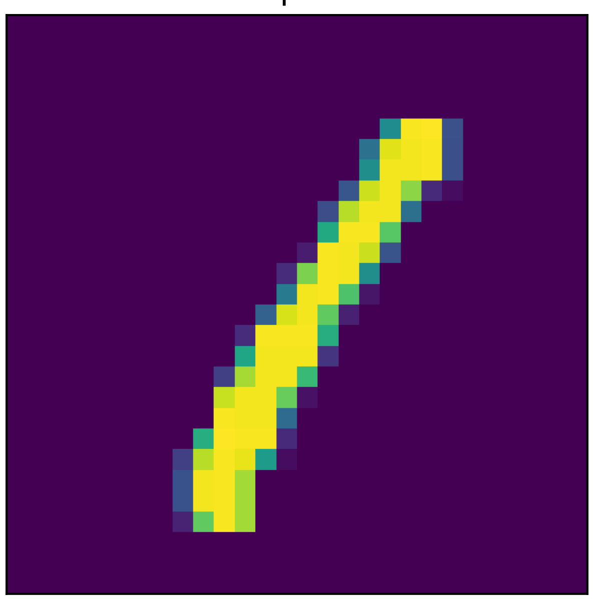
- Surface model: linear, parameter model: CNN

$$f(x) = \text{softmax}(\theta(x)^T x)$$

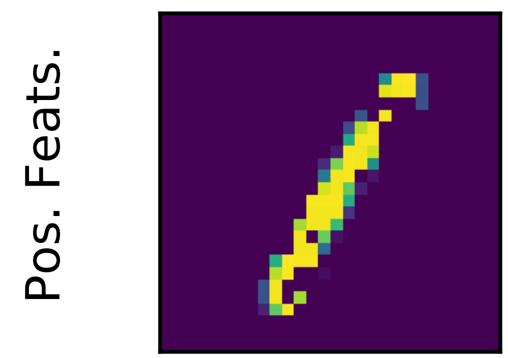


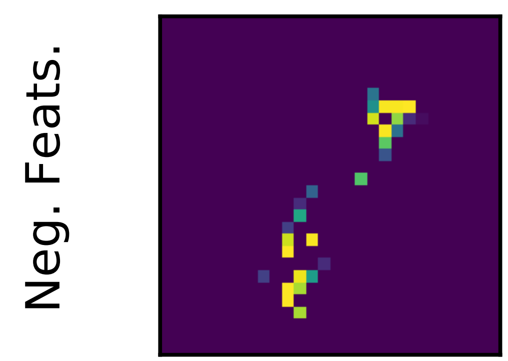
- MNIST dataset

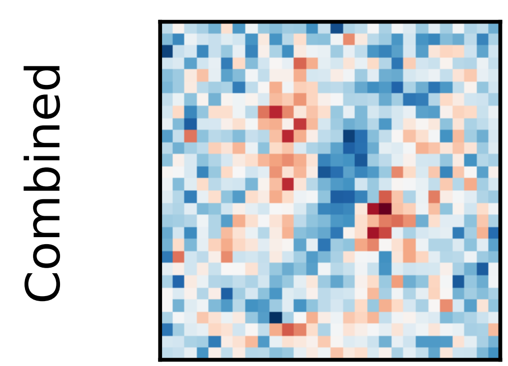
Input:



Predicted class: 1







Pixels supporting the prediction

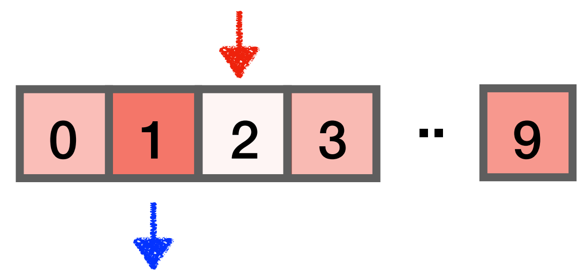
Explaining MNIST via inputs

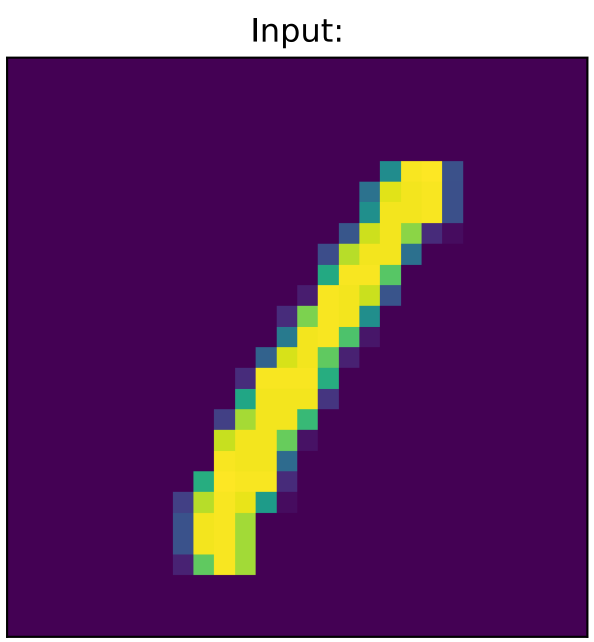
- Surface model: linear, parameter model: CNN

$$f(x) = \text{softmax}(\theta(x)^T x)$$

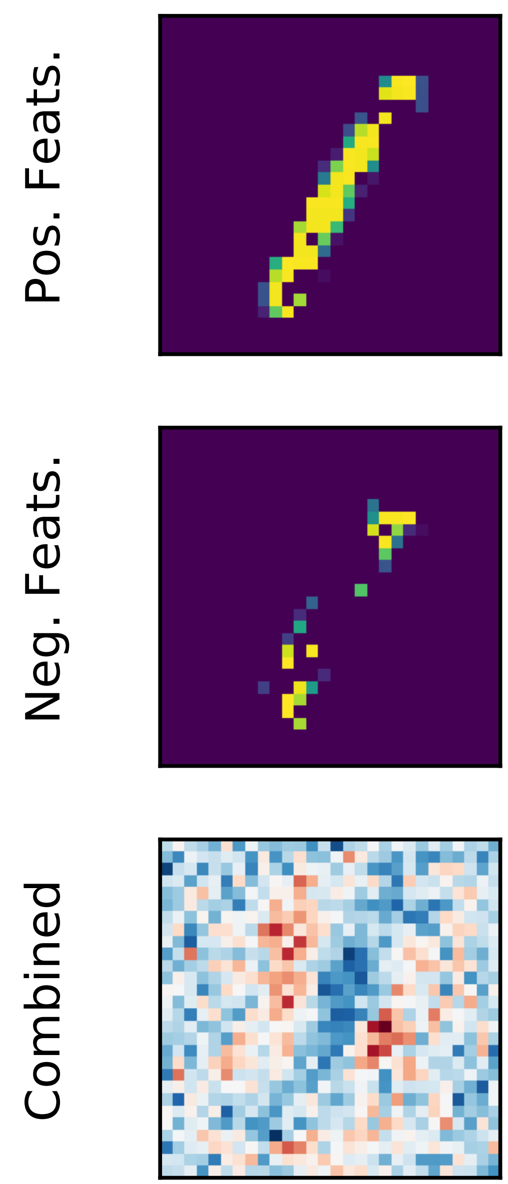
CNN (LeNet)

- MNIST dataset





Predicted class: 1



Pixels supporting the prediction

Pixels contradicting the prediction

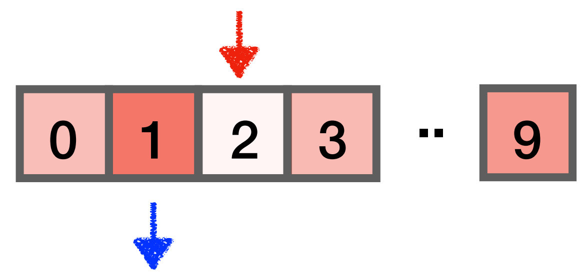
Explaining MNIST via inputs

- Surface model: linear, parameter model: CNN

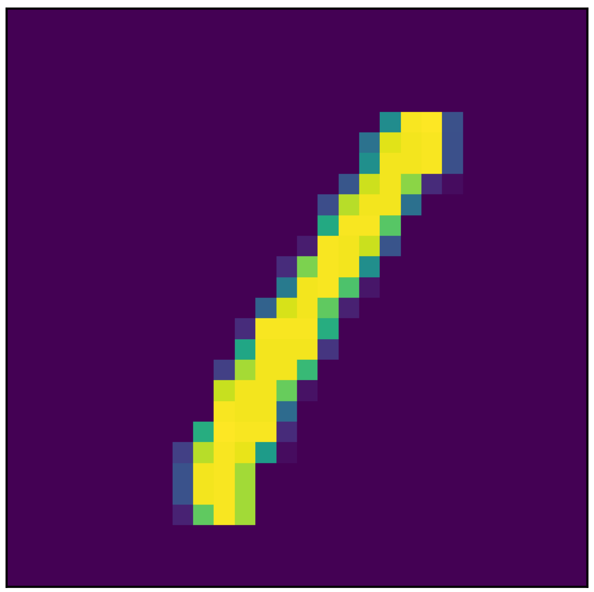
$$f(x) = \text{softmax}(\theta(x)^T x)$$

CNN (LeNet)

- MNIST dataset

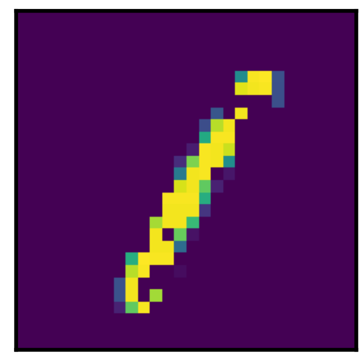


Input:



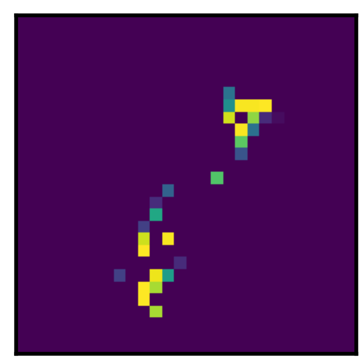
Predicted class: 1

Pos. Feats.



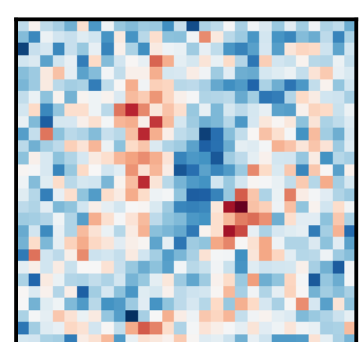
Pixels supporting the prediction

Neg. Feats.



Pixels contradicting the prediction

Combined



"Continuous" explanation

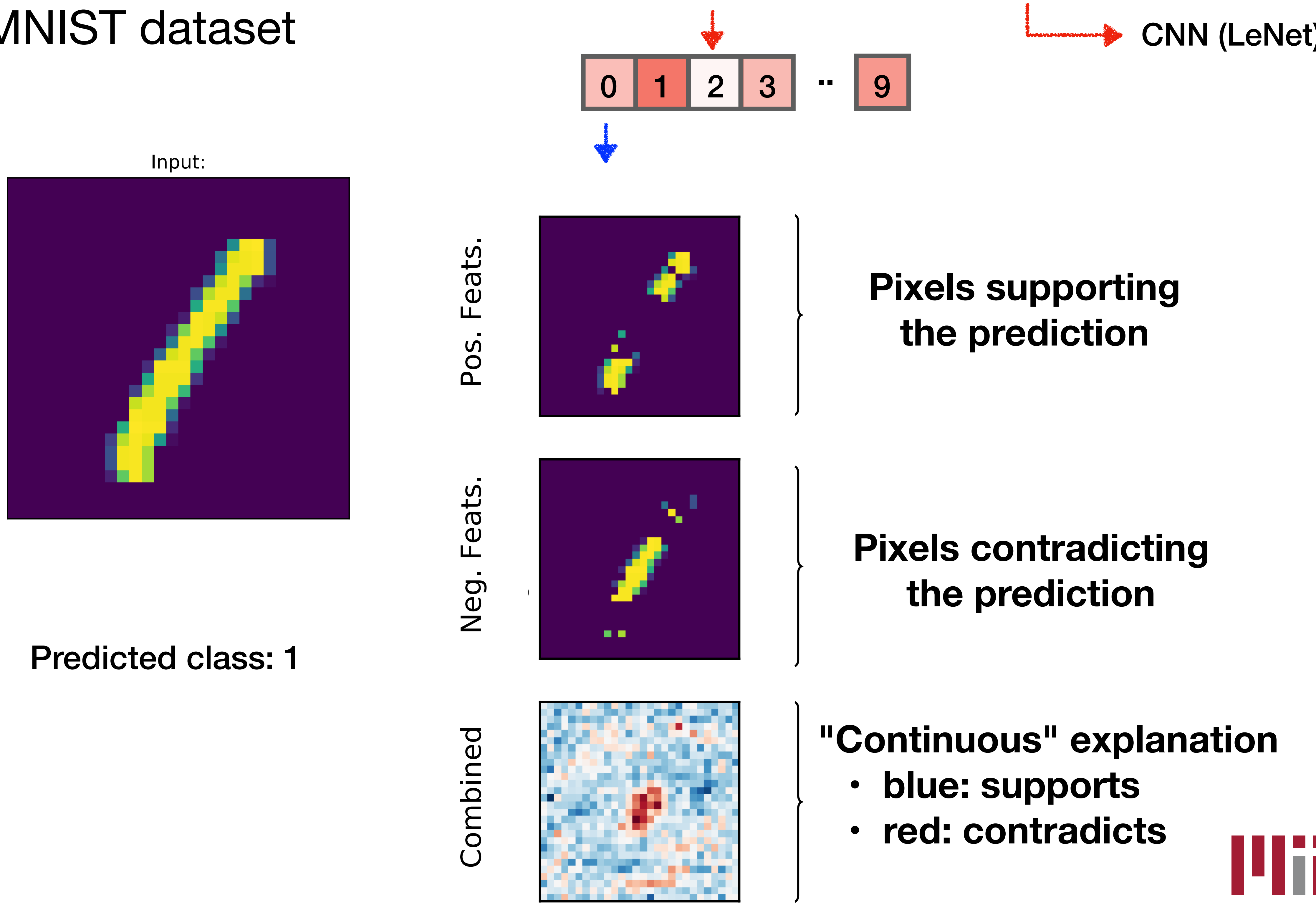
- **blue: supports**
- **red: contradicts**

Explaining MNIST via inputs

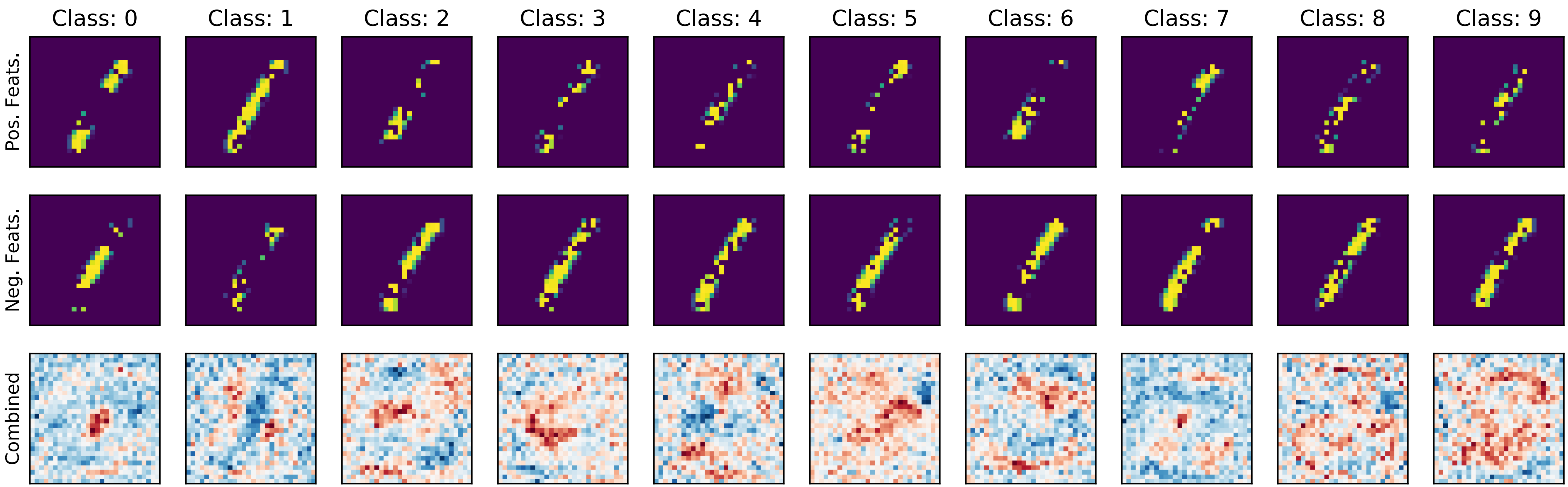
- Surface model: linear, parameter model: CNN

$$f(x) = \text{softmax}(\theta(x)^T x)$$

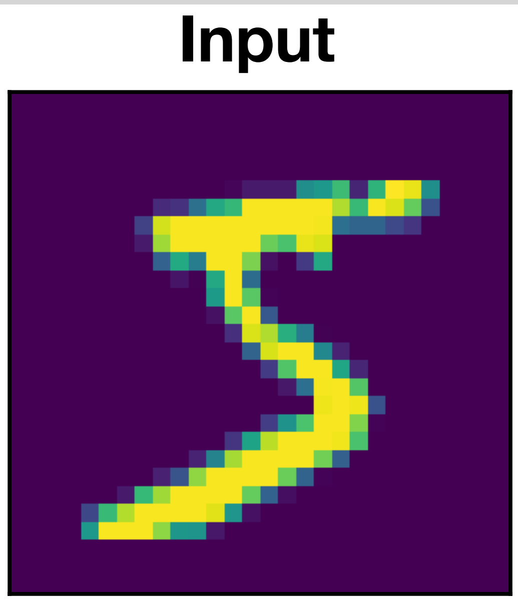
- MNIST dataset

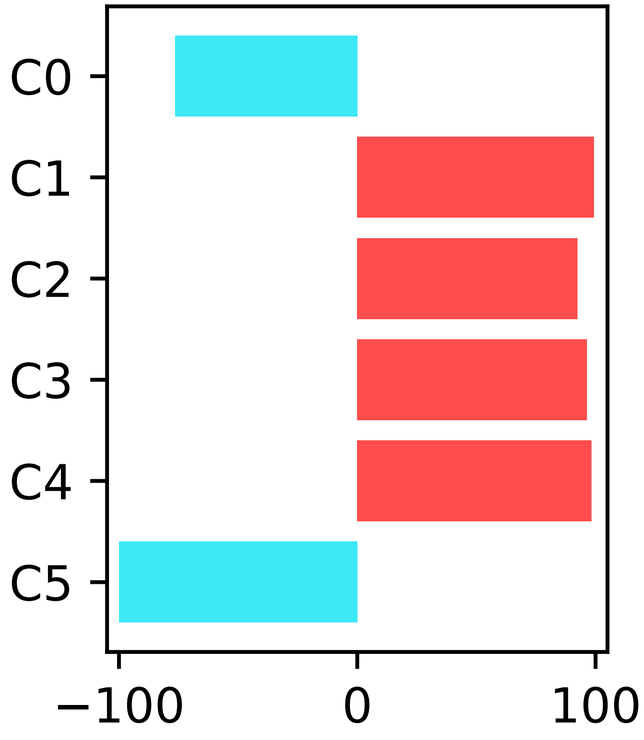


Explaining MNIST via inputs



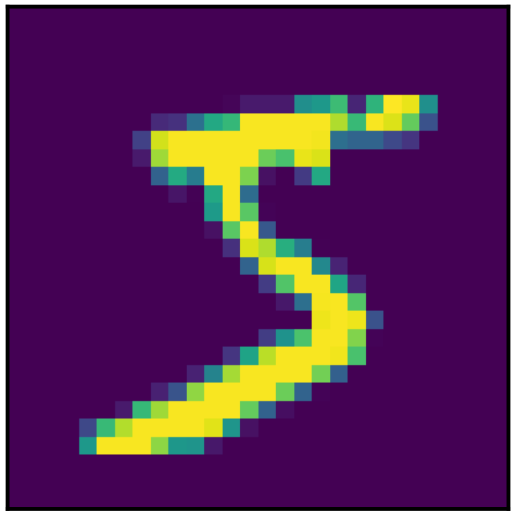
Explaining MNIST via concepts

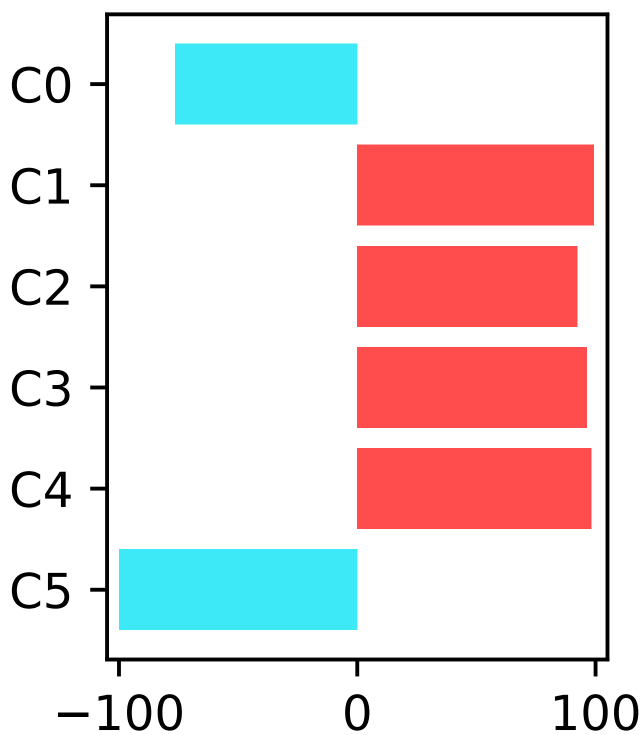




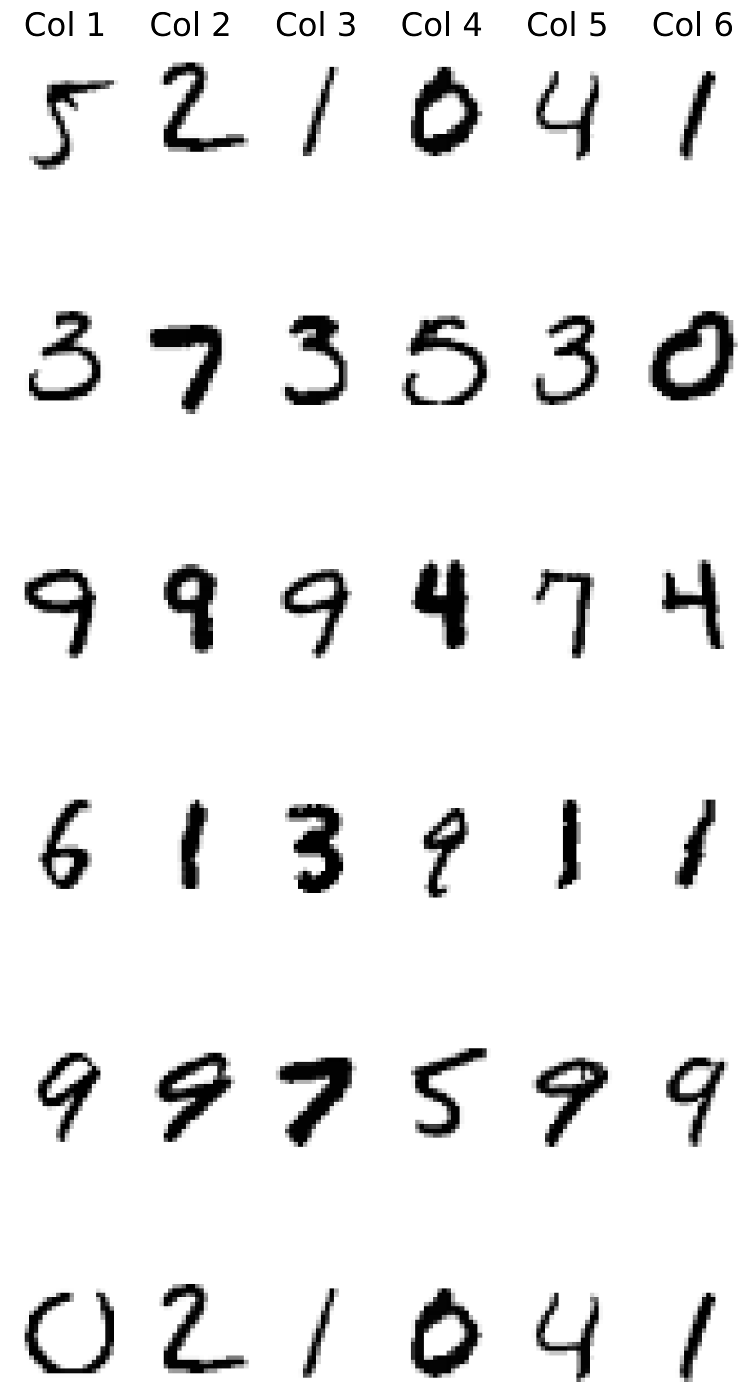
Explaining MNIST via concepts

Input



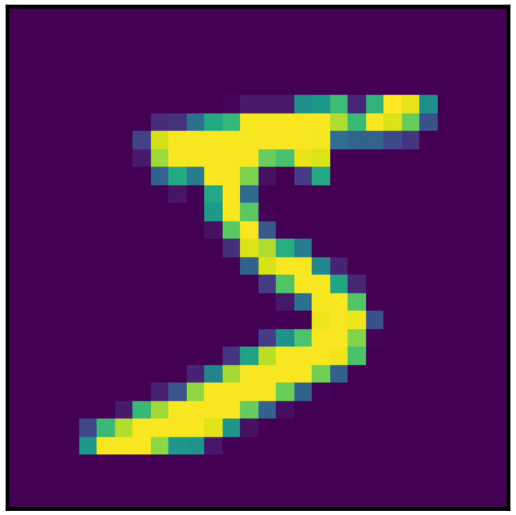


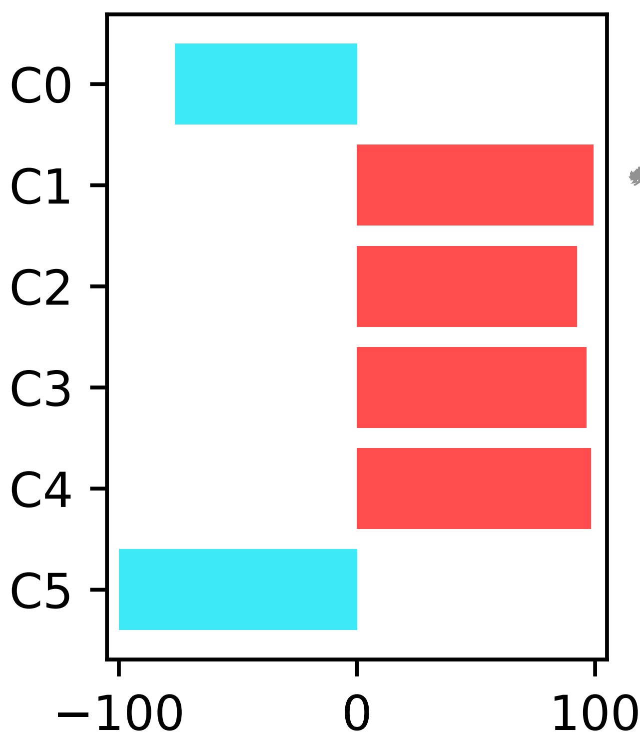
Concept prototypes



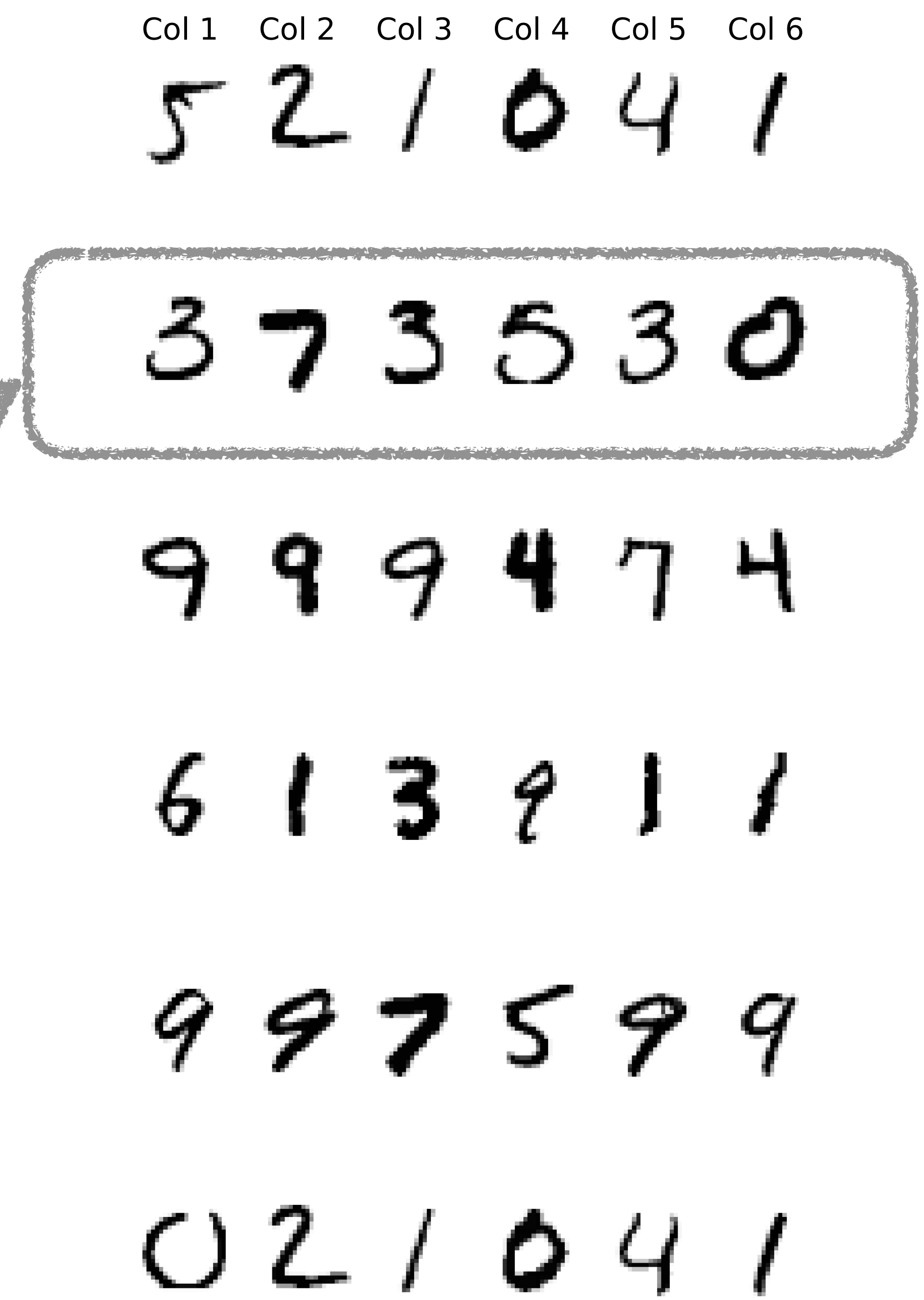
Explaining MNIST via concepts

Input



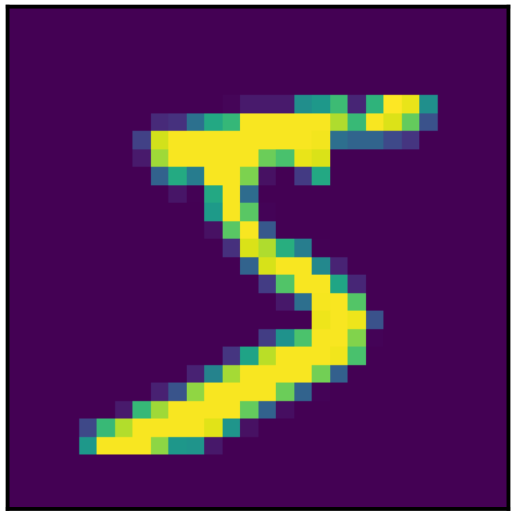


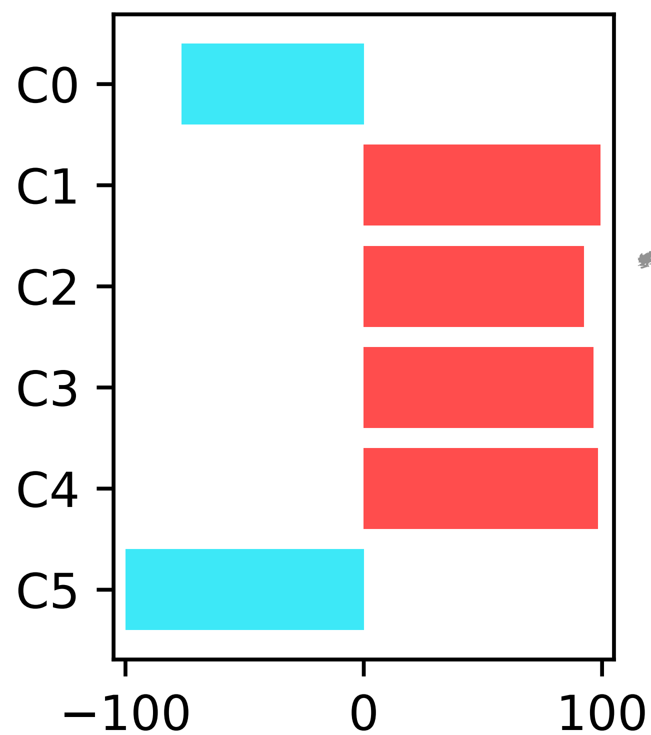
Concept prototypes



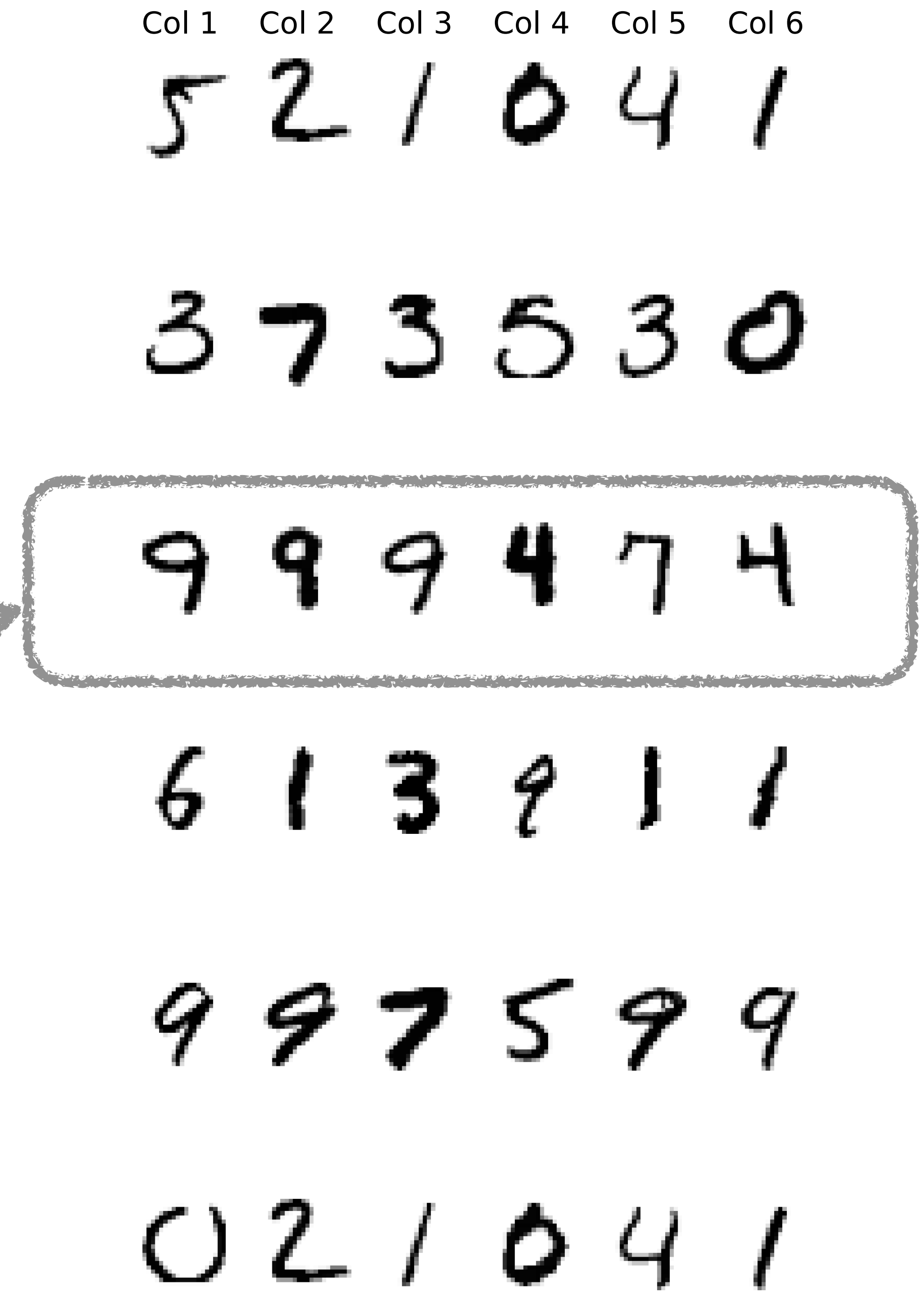
Explaining MNIST via concepts

Input



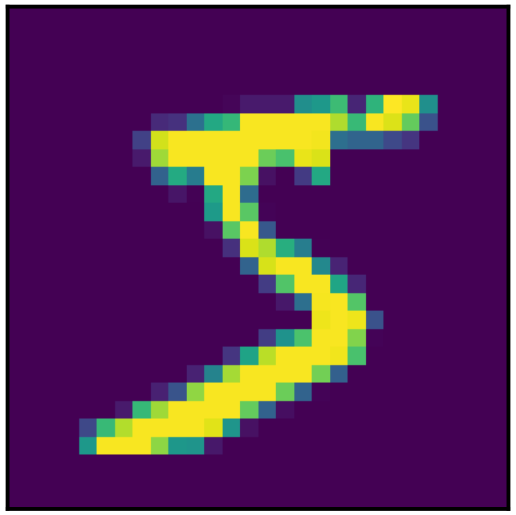


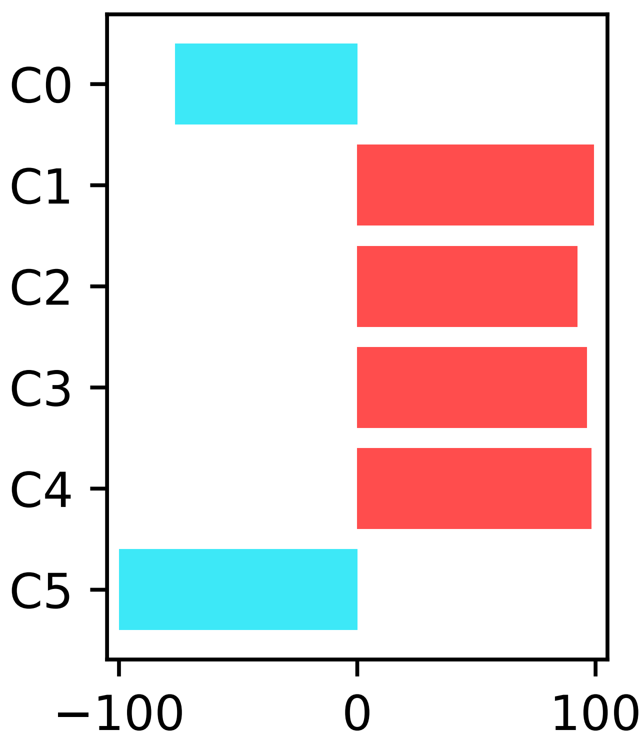
Concept prototypes



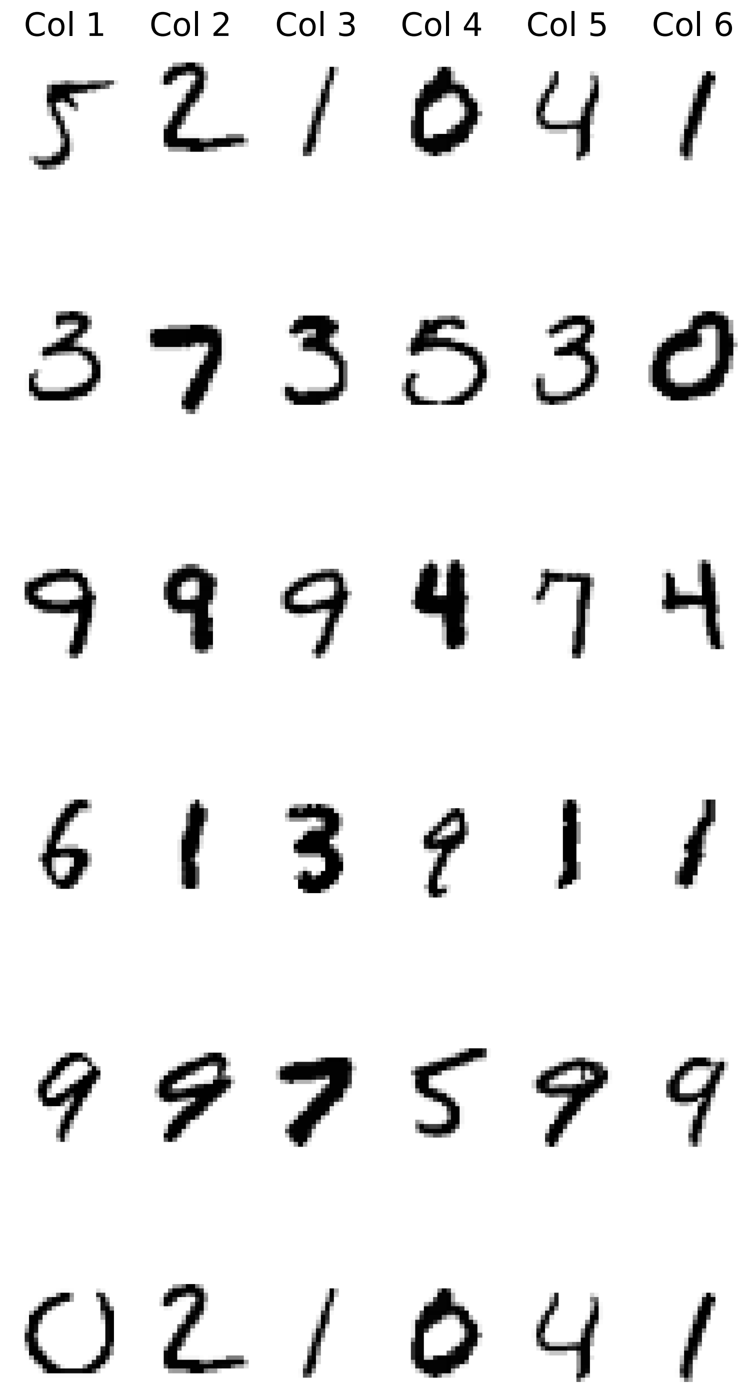
Explaining MNIST via concepts

Input



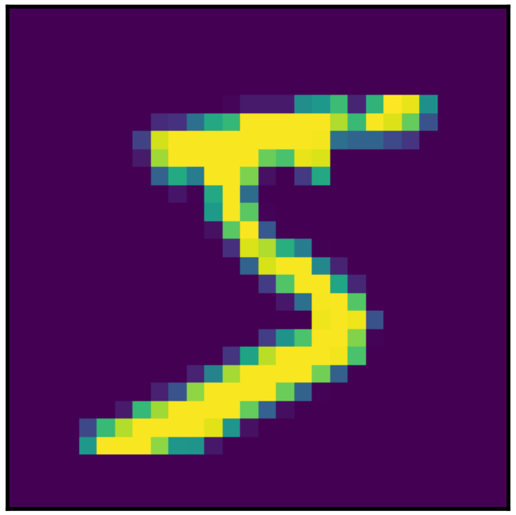


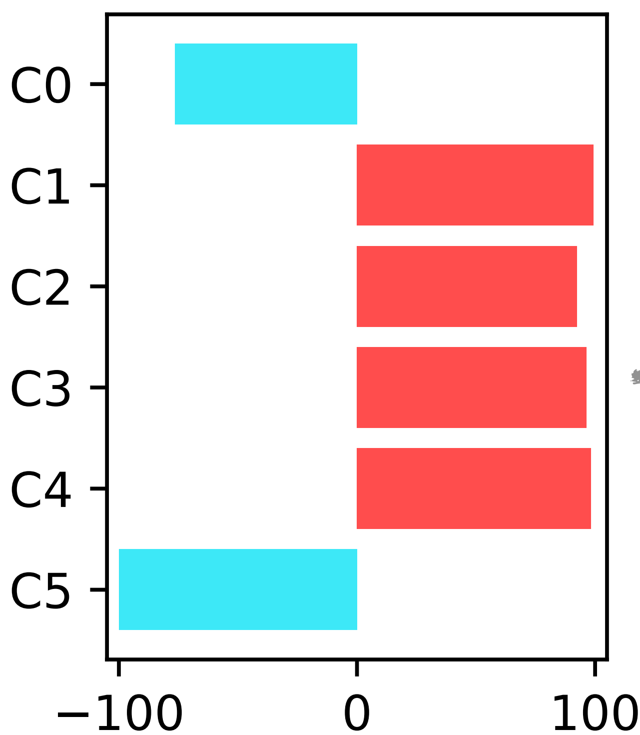
Concept prototypes



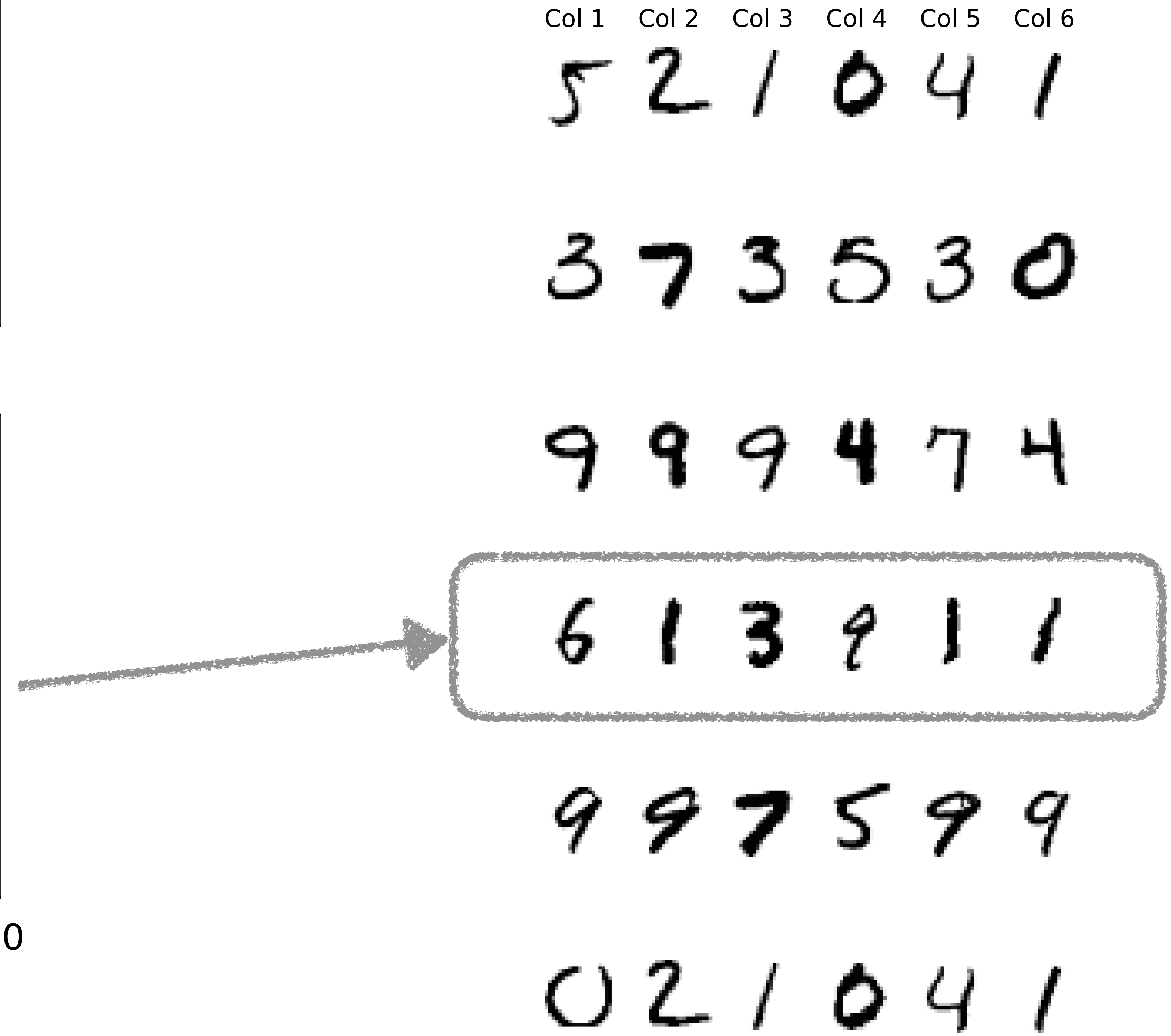
Explaining MNIST via concepts

Input



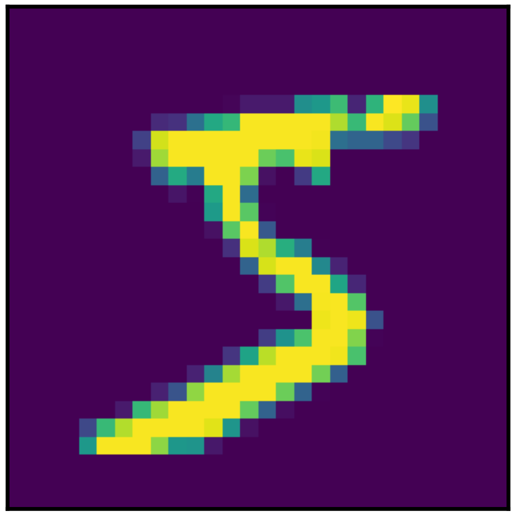


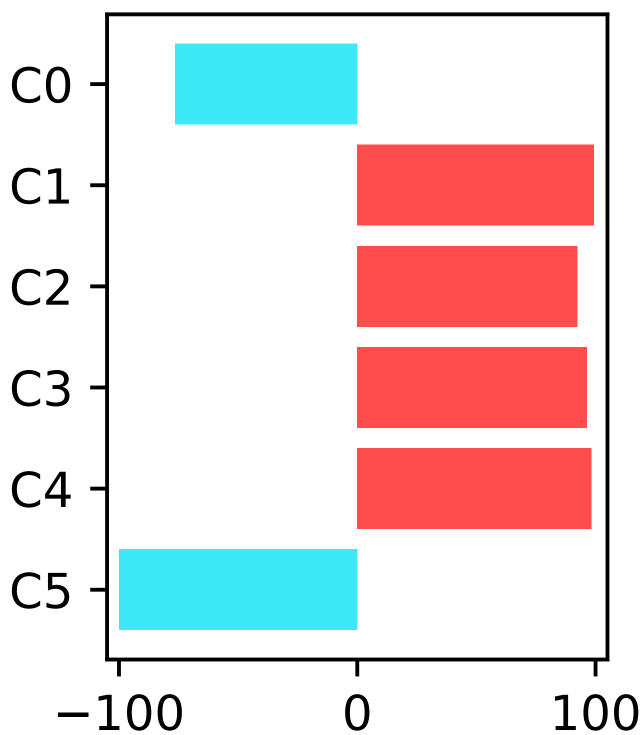
Concept prototypes



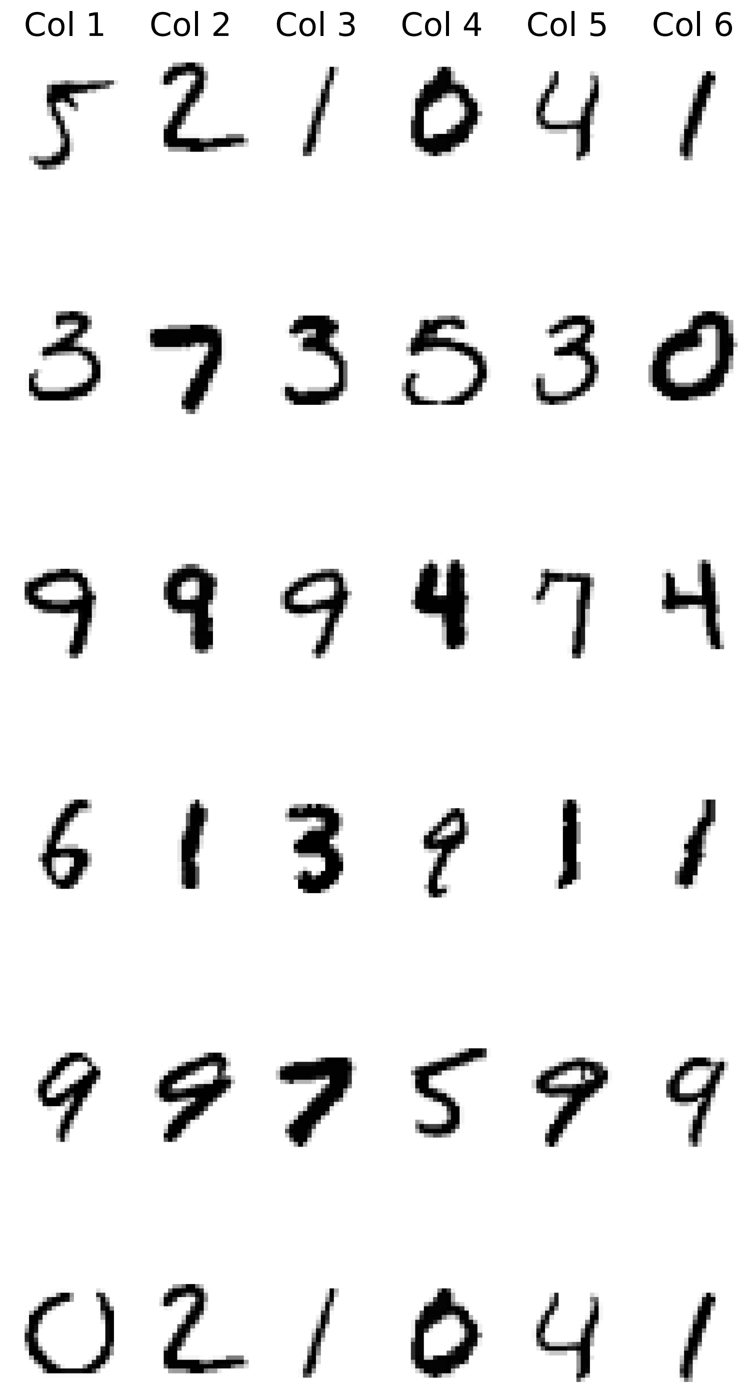
Explaining MNIST via concepts

Input



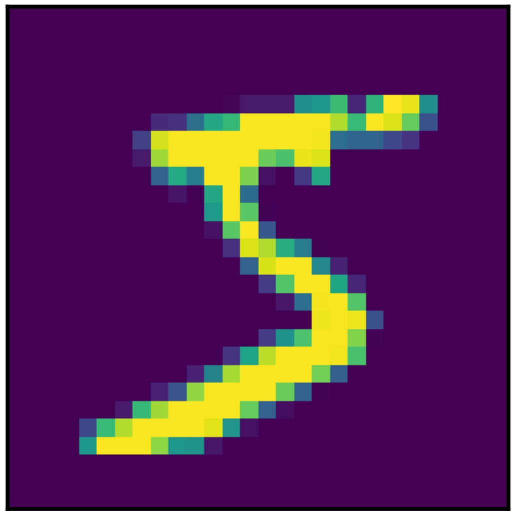


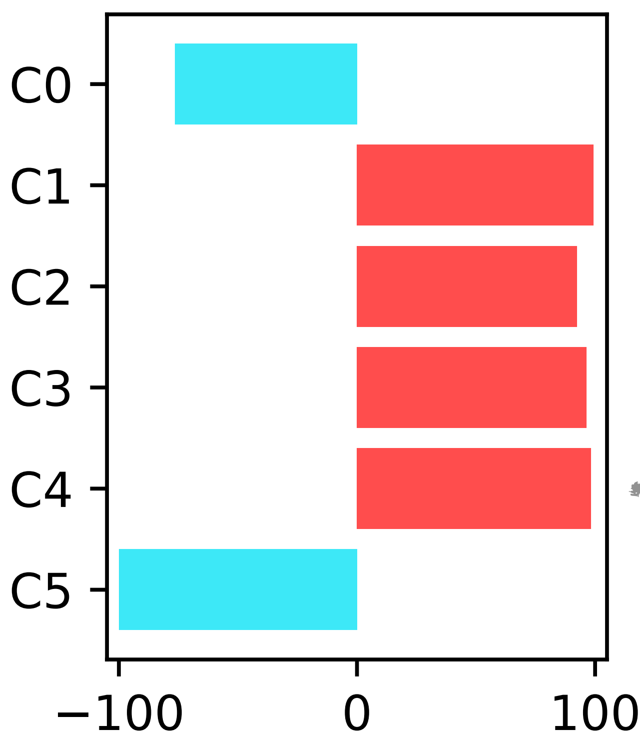
Concept prototypes



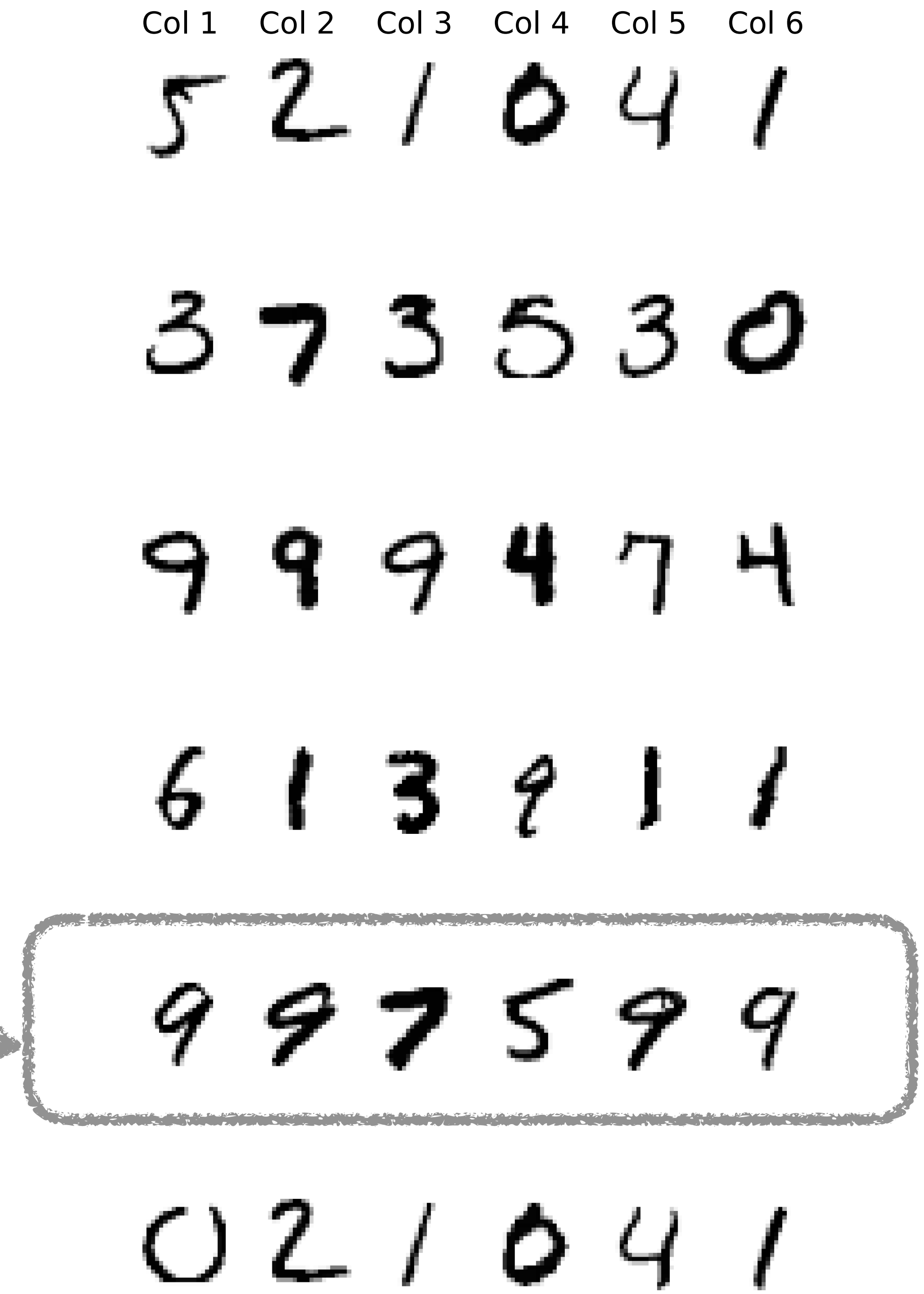
Explaining MNIST via concepts

Input





Concept prototypes



MNIST: Quantitative Evaluation

MNIST: Quantitative Evaluation

- **Consistency**. Does $\theta(x)$ really behave as importance?

MNIST: Quantitative Evaluation

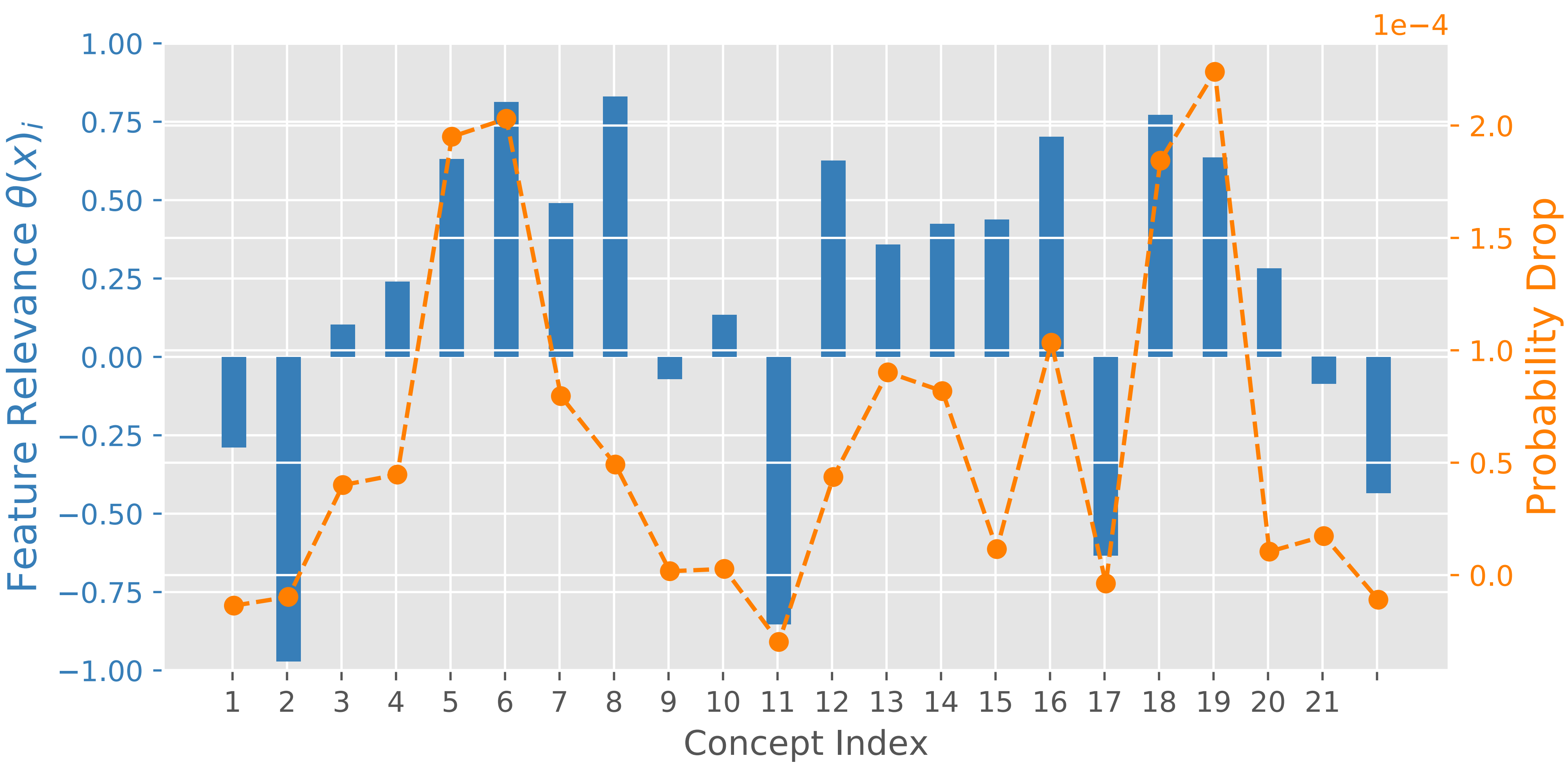
- **Consistency**. Does $\theta(x)$ really behave as importance?
- Set $\theta_i(x) \leftarrow 0$. How does class probability change?

MNIST: Quantitative Evaluation

- **Consistency**. Does $\theta(x)$ really behave as importance?
- Set $\theta_i(x) \leftarrow 0$. How does class probability change?
- Compare original $\theta_i(x)$ and drop in class probability. Should be similar!

MNIST: Quantitative Evaluation

- **Consistency**. Does $\theta(x)$ really behave as importance?
- Set $\theta_i(x) \leftarrow 0$. How does class probability change?
- Compare original $\theta_i(x)$ and drop in class probability. Should be similar!

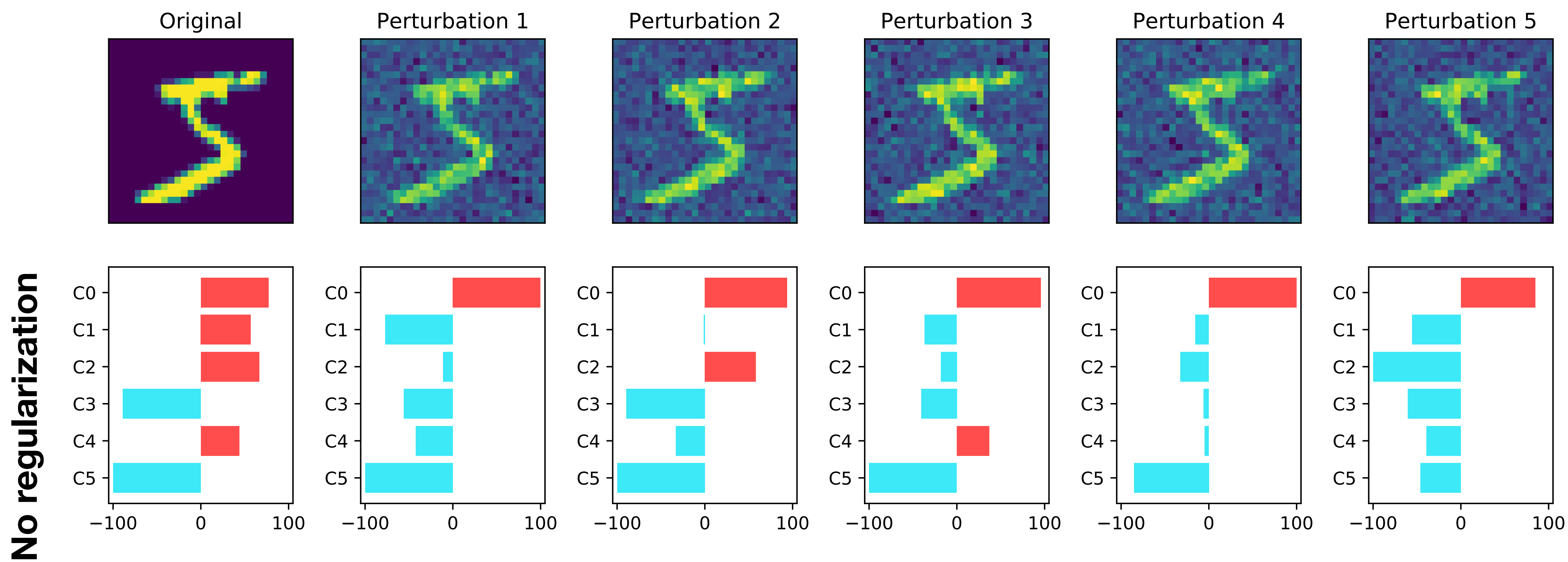


MNIST: Explanations via concepts

- **Stability**. How coherent are the explanations of similar examples?

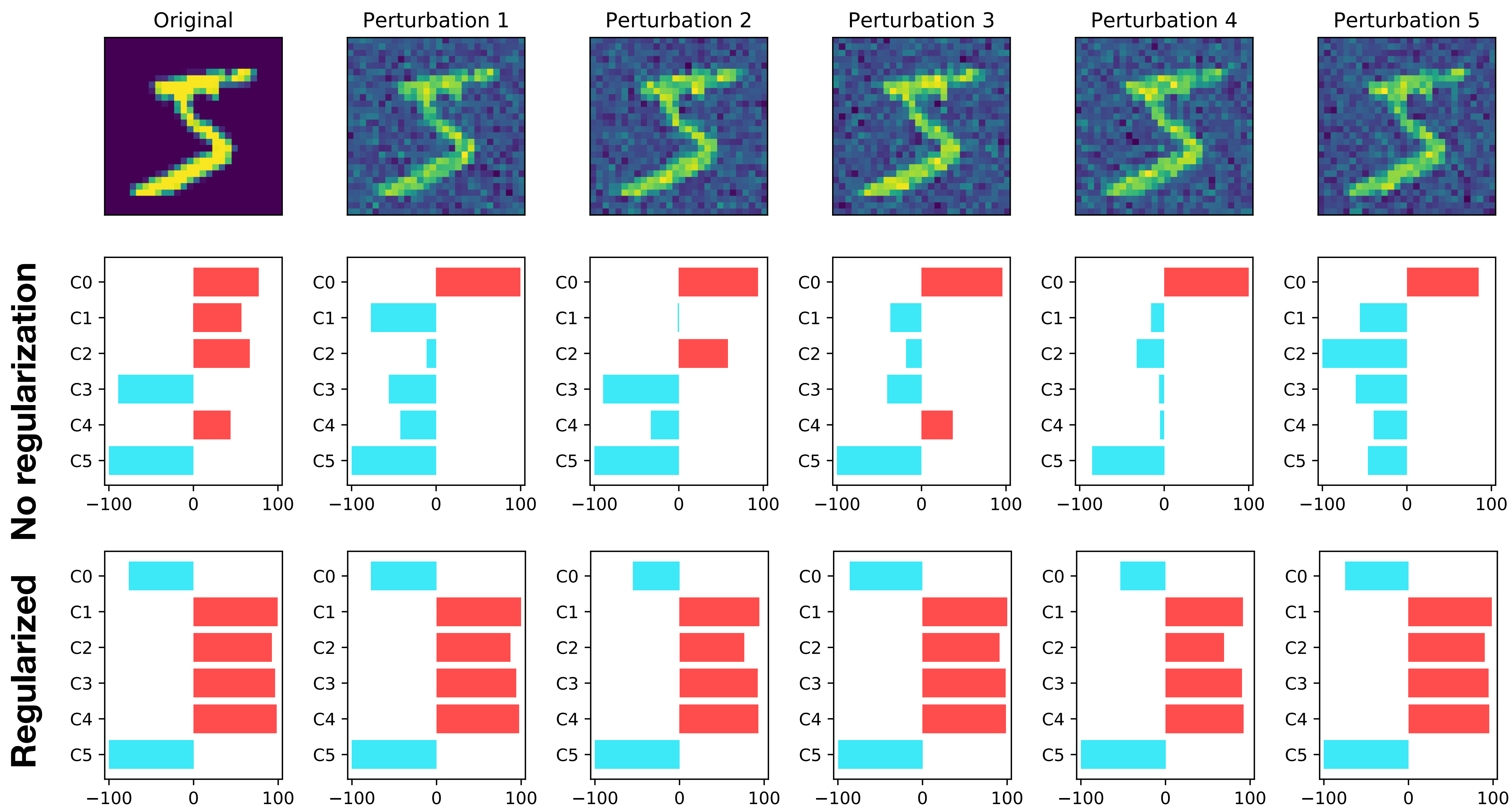
MNIST: Explanations via concepts

- **Stability**. How coherent are the explanations of similar examples?



MNIST: Explanations via concepts

- **Stability.** How coherent are the explanations of similar examples?

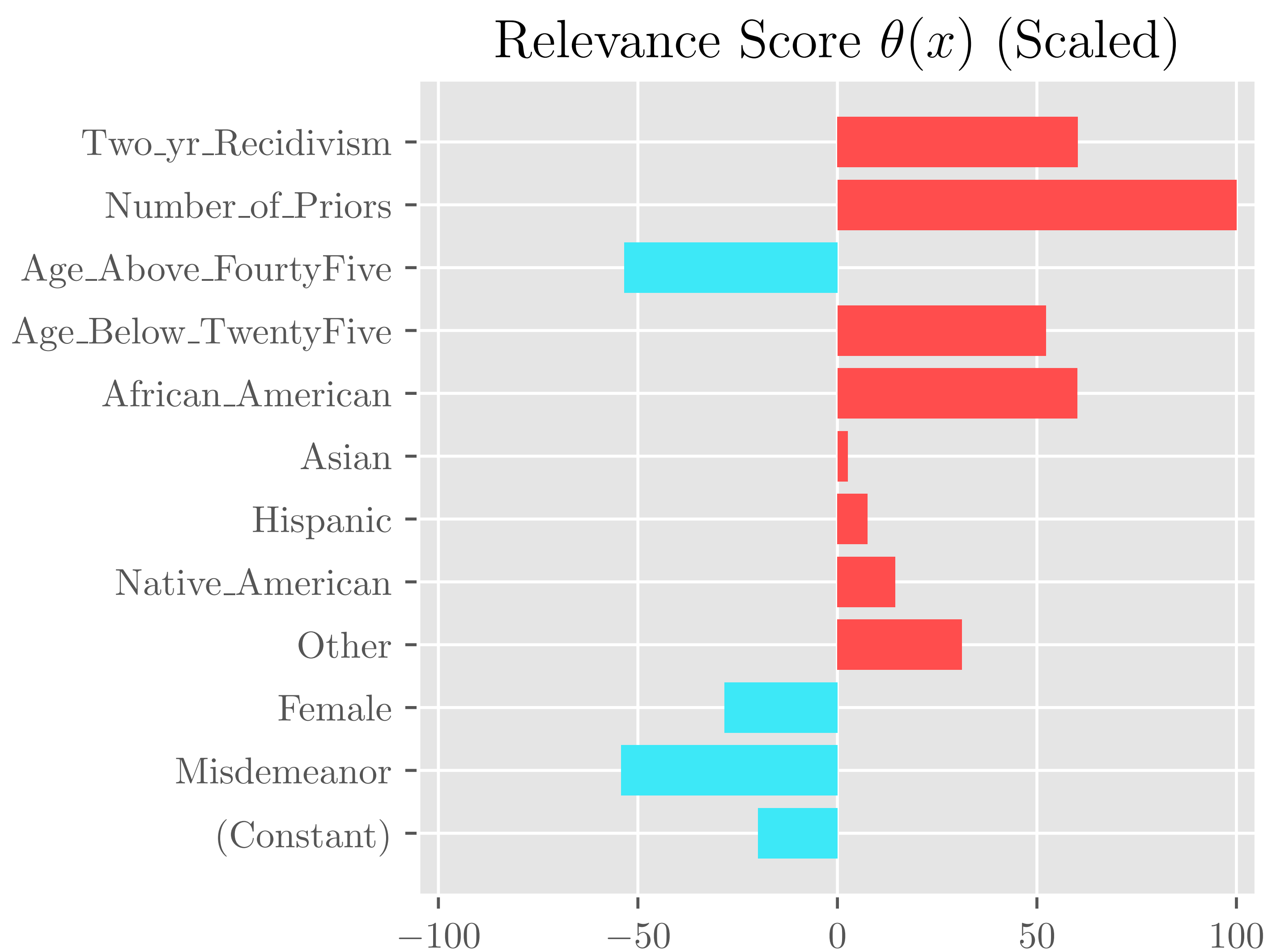


Application: COMPAS dataset

- COMPAS recidivism risk score dataset (ProPublica)
- "Relapse" scores produced by COMPAS - private proprietary algorithm
- Used in criminal justice system to aid in bail granting decisions
- Various works analyzing its fairness [Grgic-Hlaca et al., 2018, Zafar et al., 2017]

Application: COMPAS dataset

- Task: train model to reproduce COMPAS scores
- SENN model achieves 4% improvement over baseline
- Example explanation:



Effect of gradient regularization

Effect of gradient regularization

- How to evaluate **stability** of explanations?

Effect of gradient regularization

- How to evaluate **stability** of explanations?
- Continuous notion of stability:

$$\hat{L}_i = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|\theta(x_i) - \theta(x_j)\|_2}{\|h(x_i) - h(x_j)\|_2}$$

Effect of gradient regularization

- How to evaluate **stability** of explanations?
- Continuous notion of stability:

$$\hat{L}_i = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|\theta(x_i) - \theta(x_j)\|_2}{\|h(x_i) - h(x_j)\|_2}$$

Ball of radius eps around x_i

Effect of gradient regularization

- How to evaluate **stability** of explanations?

- Continuous notion of stability:

$$\hat{L}_i = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|\theta(x_i) - \theta(x_j)\|_2}{\|h(x_i) - h(x_j)\|_2}$$

Ball of radius eps around x_i

- Discrete analogue:

$$\hat{L}_i = \underset{x_j \in \mathcal{N}_\epsilon(x_i) \leq \epsilon}{\operatorname{argmax}} \frac{\|\theta(x_i) - \theta(x_j)\|_2}{\|h(x_i) - h(x_j)\|_2}$$

Effect of gradient regularization

- How to evaluate **stability** of explanations?

- Continuous notion of stability:

$$\hat{L}_i = \underset{x_j \in B_\epsilon(x_i)}{\operatorname{argmax}} \frac{\|\theta(x_i) - \theta(x_j)\|_2}{\|h(x_i) - h(x_j)\|_2}$$

Ball of radius eps around x_i

- Discrete analogue:

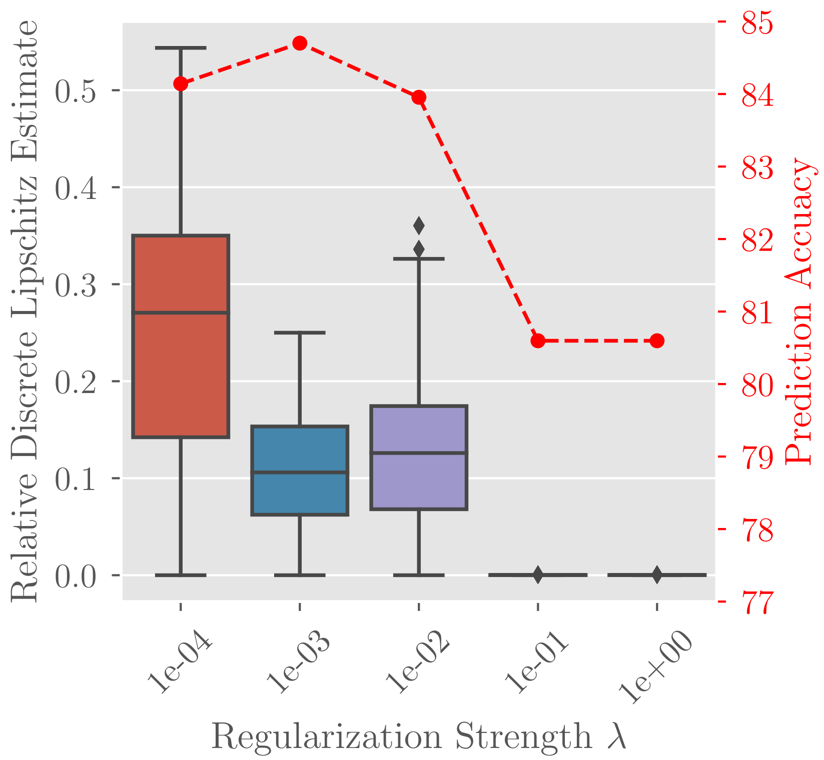
$$\hat{L}_i = \underset{x_j \in \mathcal{N}_\epsilon(x_i) \leq \epsilon}{\operatorname{argmax}} \frac{\|\theta(x_i) - \theta(x_j)\|_2}{\|h(x_i) - h(x_j)\|_2}$$

Set of points in dataset at most distance eps away from x_i

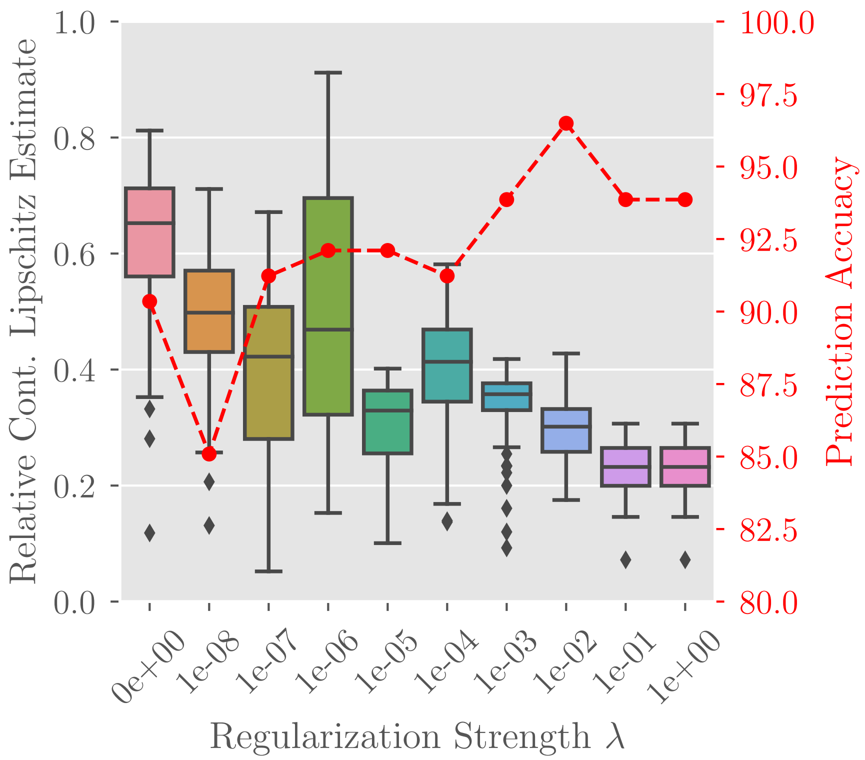
Effect of gradient regularization

- Stronger gradient regularization -> more stability (and often better accuracy!)

COMPAS dataset



Breast Cancer dataset



Next Steps

- Larger, more complex datasets
- Alternative approaches to learn interpretable concepts
- Can we use explanations during training to improve performance?

Summary

- Inject interpretability into rich neural network models
- Framework draws inspiration from classic notions of interpretability
- Directly enforces stability and consistency of explanations