From Human Explanation to Model Interpretability: A Framework Based on Weight of Evidence

David Alvarez-Melis (MSR)
Harmanpreet Kaur (UMich)
Hal Daumé III (MSR/UMD)
Hanna Wallach (MSR)
Jennifer Wortman Vaughan (MSR)

HCOMP
November 2021

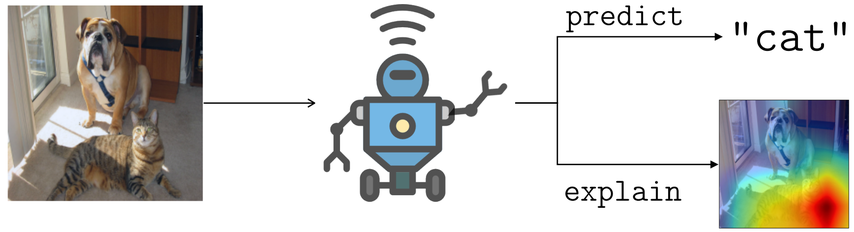
Interpretability
a human-centered approach

Interpretability
a human-centered approach

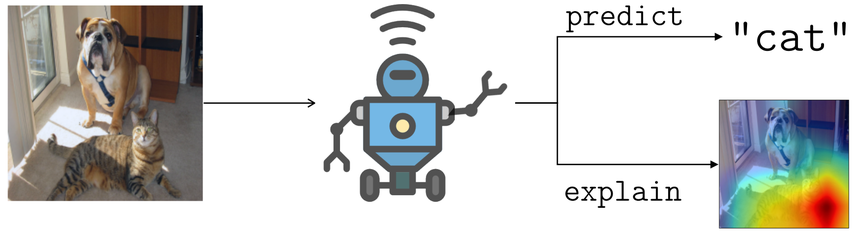
- *Local* interpretability: explain specific model predictions

Interpretability
a human-centered approach

- *Local* interpretability: explain specific model predictions

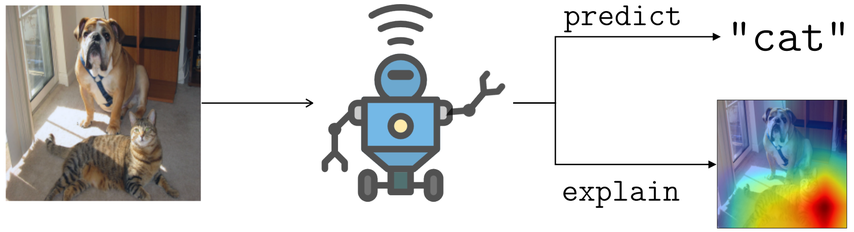


Interpretability
a human-centered approach



- *Local* interpretability: explain specific model predictions
- Typical approach: design and evaluate explanations based on abstract principles

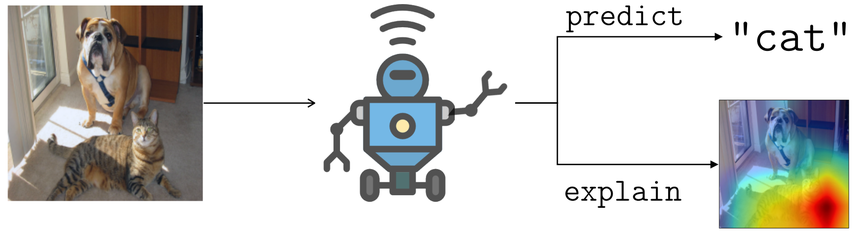
Interpretability
a human-centered approach



- *Local* interpretability: explain specific model predictions
- Typical approach: design and evaluate explanations based on abstract principles

Abstract Principles
(perturbation, axioms, robustness, etc)

Interpretability
a human-centered approach

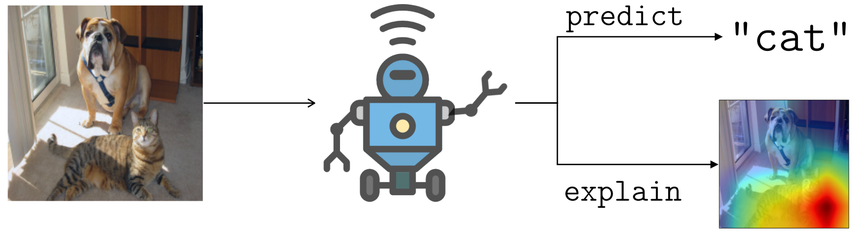


- *Local* interpretability: explain specific model predictions
- Typical approach: design and evaluate explanations based on abstract principles

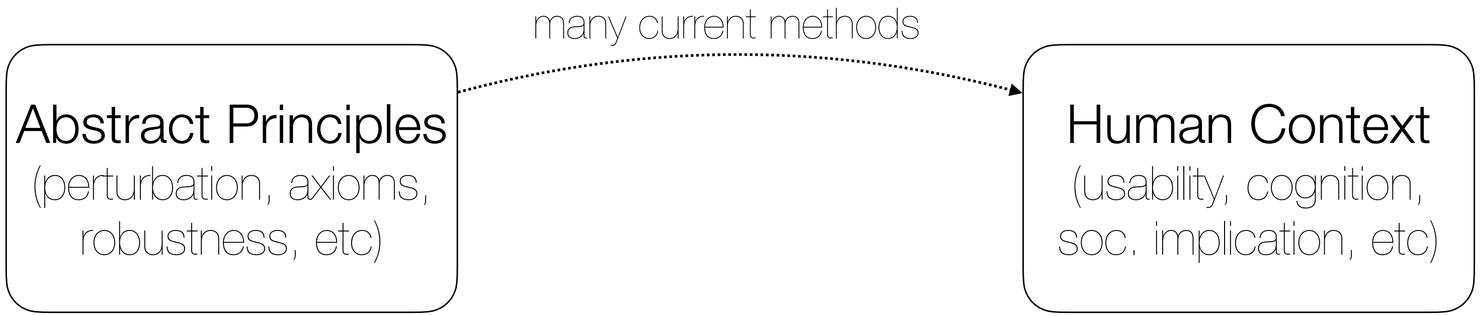
Abstract Principles
(perturbation, axioms, robustness, etc)

Human Context
(usability, cognition, soc. implication, etc)

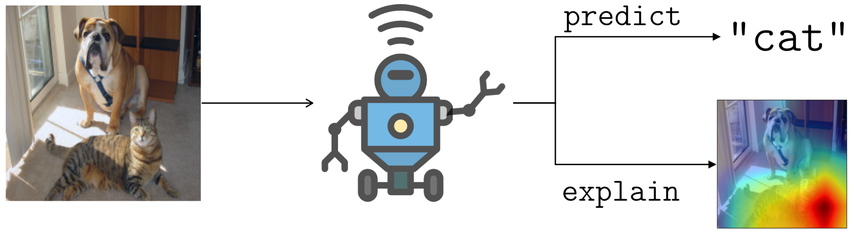
Interpretability
a human-centered approach



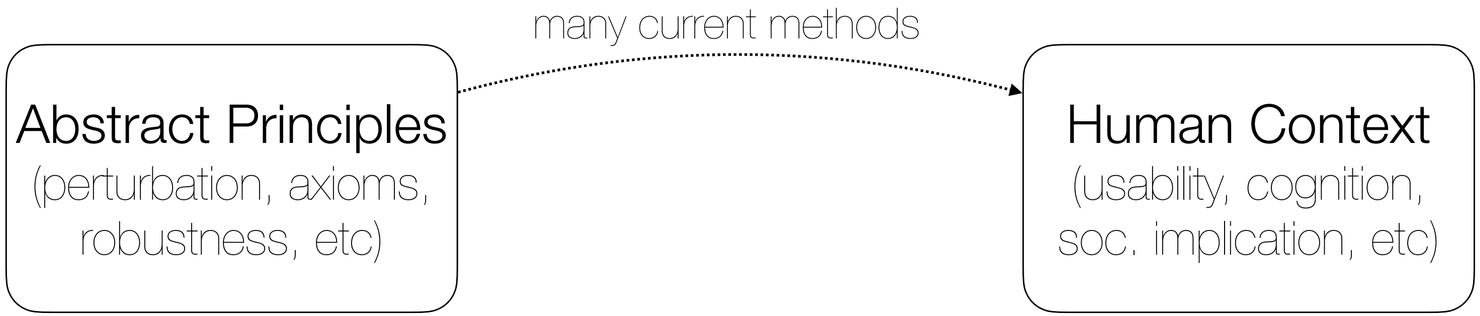
- *Local* interpretability: explain specific model predictions
- Typical approach: design and evaluate explanations based on abstract principles



Interpretability
a human-centered approach

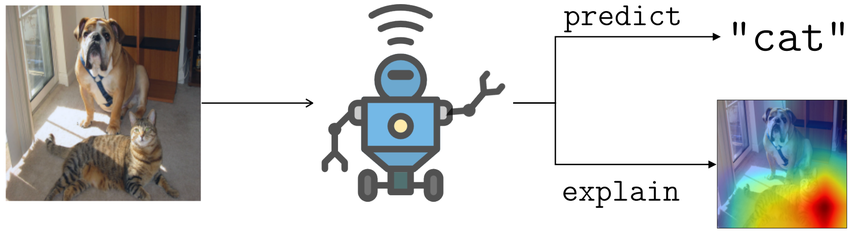


- *Local* interpretability: explain specific model predictions
- Typical approach: design and evaluate explanations based on abstract principles

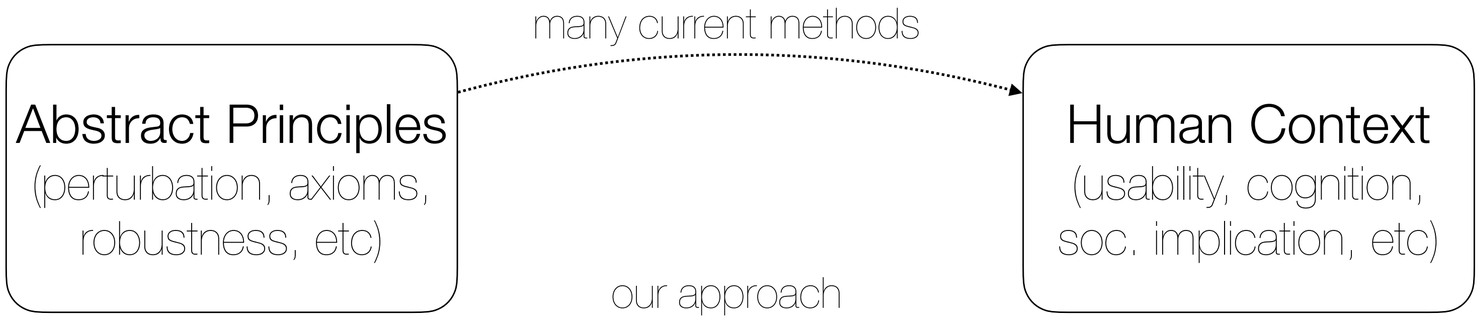


- The goal of ML interpretability: explain to humans

Interpretability
a human-centered approach

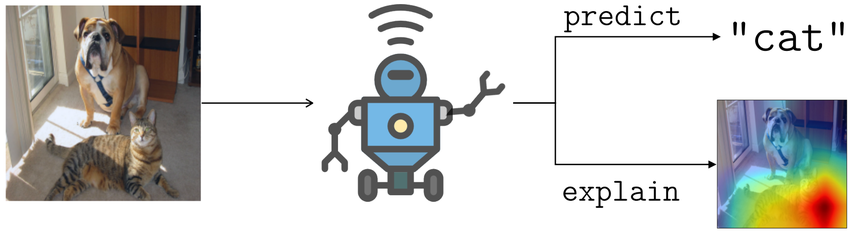


- *Local* interpretability: explain specific model predictions
- Typical approach: design and evaluate explanations based on abstract principles

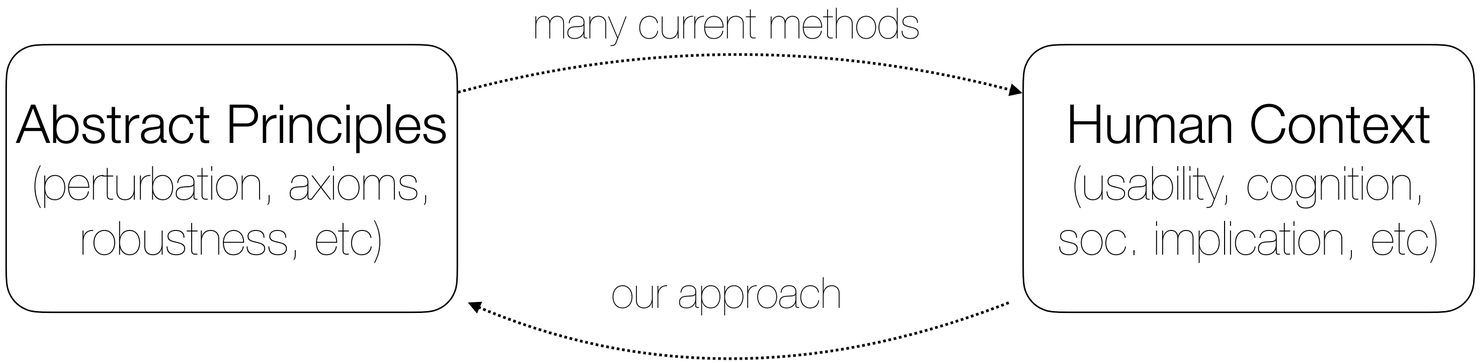


- The goal of ML interpretability: explain to humans

Interpretability
a human-centered approach

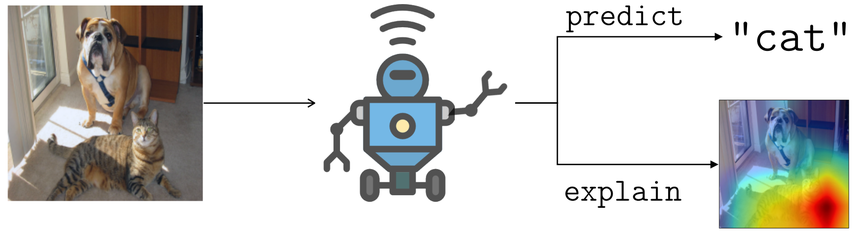


- *Local* interpretability: explain specific model predictions
- Typical approach: design and evaluate explanations based on abstract principles

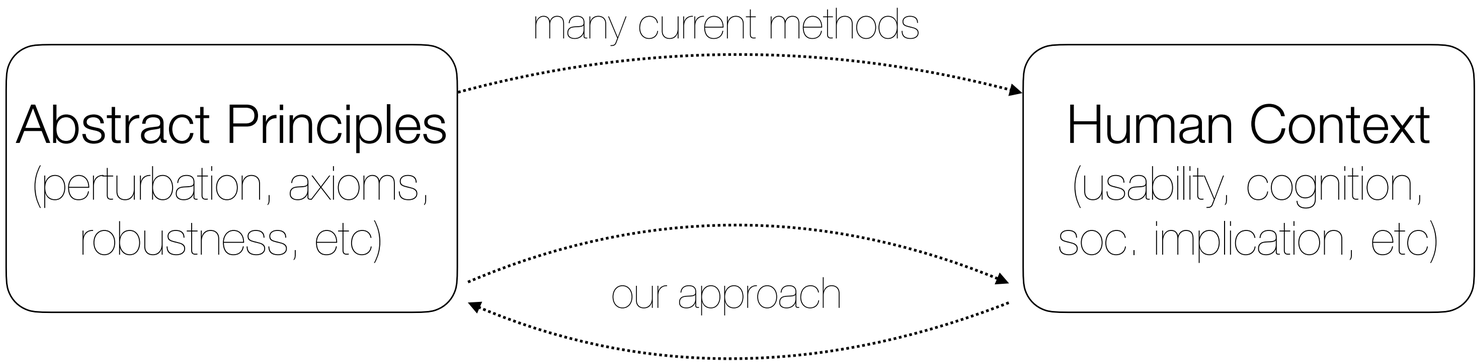


- The goal of ML interpretability: explain to humans

Interpretability
a human-centered approach

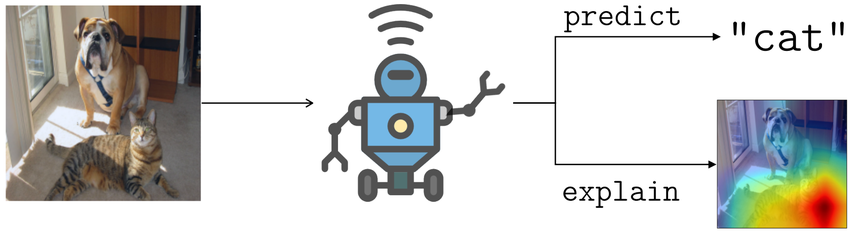


- *Local* interpretability: explain specific model predictions
- Typical approach: design and evaluate explanations based on abstract principles

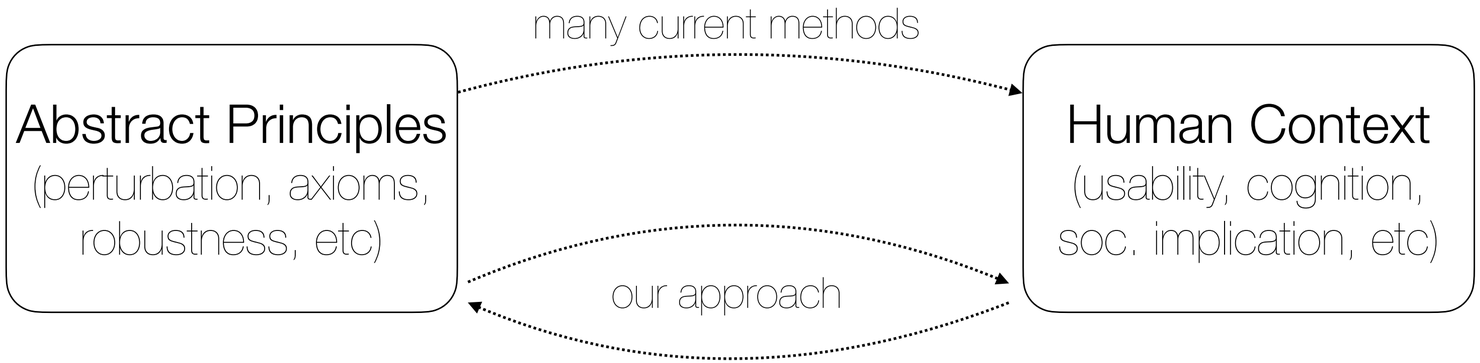


- The goal of ML interpretability: explain to humans

Interpretability
a human-centered approach



- *Local* interpretability: explain specific model predictions
- Typical approach: design and evaluate explanations based on abstract principles



- The goal of ML interpretability: explain to humans
- Starting point: *what does it mean to explain and how do humans do it*?

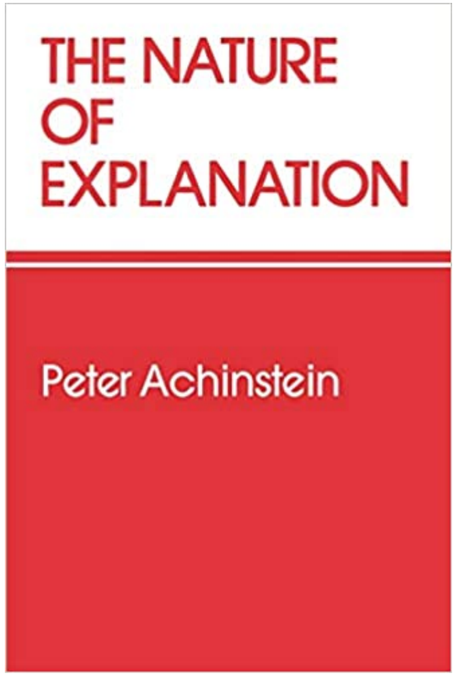
Explanation
a humanistic view

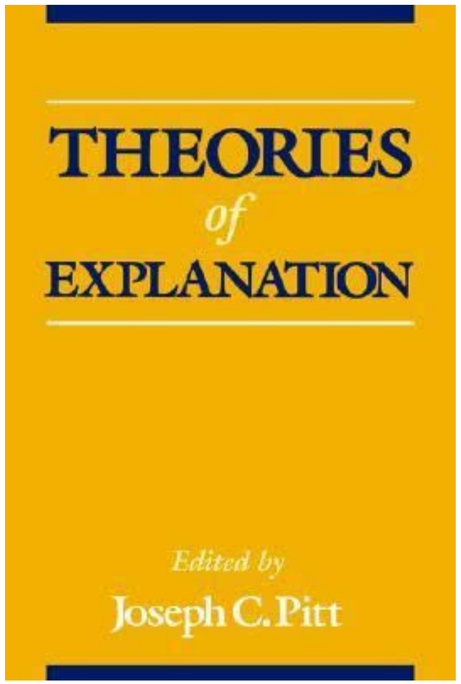
Explanation
a humanistic view

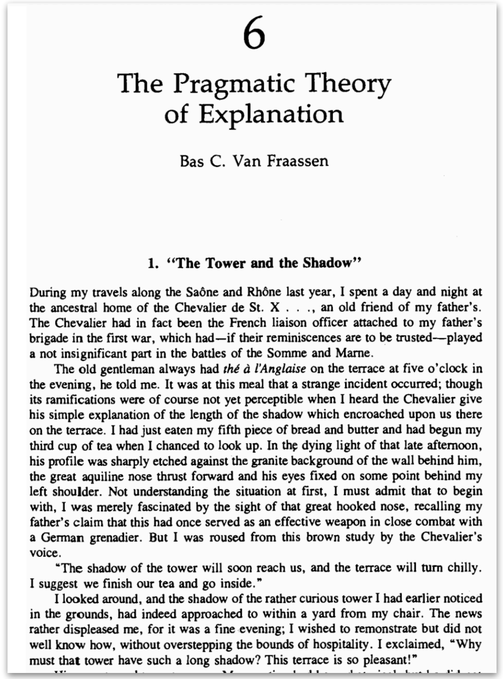
- Nature of human explanation long studied in philosophy, cognitive science, social sciences
  [e.g., Hempel, Achinstein, Pitt, van Fraassen, Lipton, Salmon, …] [survey by Miller '19]

Explanation
a humanistic view

- Nature of human explanation long studied in philosophy, cognitive science, social sciences
  [e.g., Hempel, Achinstein, Pitt, van Fraassen, Lipton, Salmon, …] [survey by Miller '19]







6

The Pragmatic Theory of Explanation

Bas C. Van Fraassen

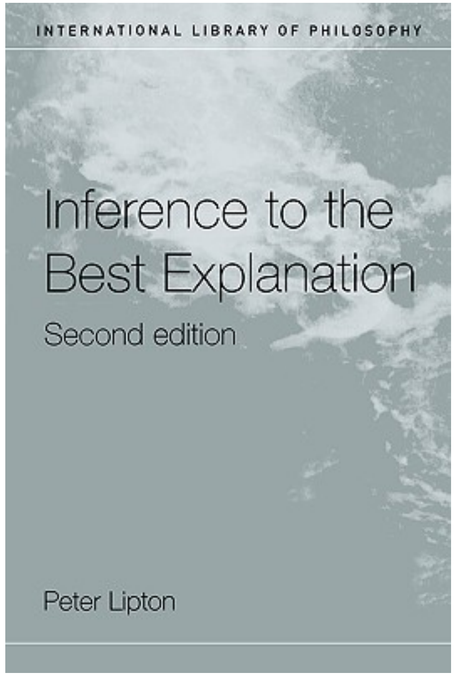
1. "The Tower and the Shadow"

During my travels along the Saône and Rhône last year, I spent a day and night at the ancestral home of the Chevalier de St. X . . ., an old friend of my father's. The Chevalier had in fact been the French liaison officer attached to my father's brigade in the first war, which had—if their reminiscences are to be trusted—played a not insignificant part in the battles of the Somme and Marne.

The old gentleman always had *thé à l'Anglaise* on the terrace at five o'clock in the evening, he told me. It was at this meal that a strange incident occurred; though its ramifications were of course not yet perceptible when I heard the Chevalier give his simple explanation of the length of the shadow which encroached upon us there on the terrace. I had just eaten my fifth piece of bread and butter and had begun my third cup of tea when I chanced to look up. In the dying light of that late afternoon, his profile was sharply etched against the granite background of the wall behind him, the great aquiline nose thrust forward and his eyes fixed on some point behind my left shoulder. Not understanding the situation at first, I must admit that to begin with, I was merely fascinated by the sight of that great hooked nose, recalling my father's claim that this had once served as an effective weapon in close combat with a German grenadier. But I was roused from this brown study by the Chevalier's voice.

"The shadow of the tower will soon reach us, and the terrace will turn chilly. I suggest we finish our tea and go inside."

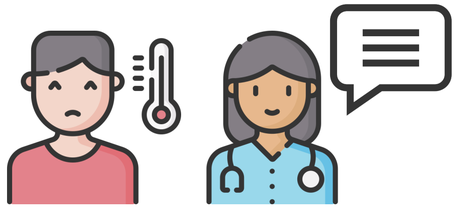
I looked around, and the shadow of the rather curious tower I had earlier noticed in the grounds, had indeed approached to within a yard from my chair. The news rather displeased me, for it was a fine evening; I wished to remonstrate but did not well know how, without overstepping the bounds of hospitality. I exclaimed, "Why must that tower have such a long shadow? This terrace is so pleasant!"



Interpretability:
lessons from the social sciences

Human Explanation

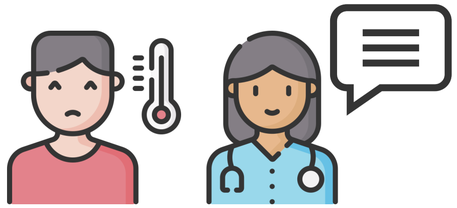
(according to philosophy & the social sciences)



Interpretability:
lessons from the social sciences

Human Explanation

(according to philosophy & the social sciences)

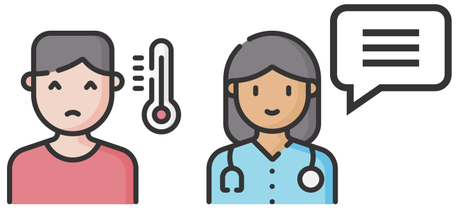


- Contrastive
  [van Fraassen, 1988; Lipton, 1990]

Interpretability:
lessons from the social sciences

Human Explanation

(according to philosophy & the social sciences)



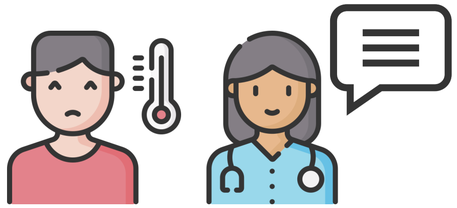
- Contrastive
  [van Fraassen, 1988; Lipton, 1990]

"fever is more consistent with pneumonia **than** with a cold"

Interpretability:
lessons from the social sciences

Human Explanation

(according to philosophy & the social sciences)



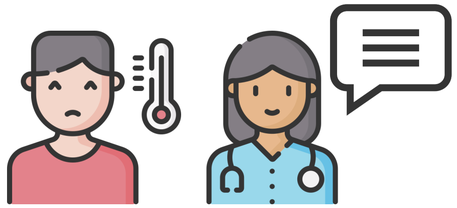
- Contrastive
  [van Fraassen, 1988; Lipton, 1990]
- Factual and counterfactual
  [van Fraassen, 1988]

"fever is more consistent with pneumonia **than** with a cold"

Interpretability:
lessons from the social sciences

Human Explanation

(according to philosophy & the social sciences)



- Contrastive
  [van Fraassen, 1988; Lipton, 1990]
- Factual and counterfactual
  [van Fraassen, 1988]

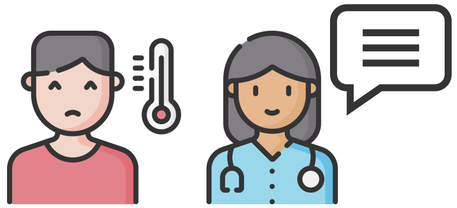
"fever is more consistent with pneumonia **than** with a cold"

"**had there been** chest pressure too, diagnosis would be bronchitis"

Interpretability:
lessons from the social sciences

Human Explanation

(according to philosophy & the social sciences)



- Contrastive
  [van Fraassen, 1988; Lipton, 1990]
- Factual and counterfactual
  [van Fraassen, 1988]
- Sequences of simple statements
  [Hempel, 1962]

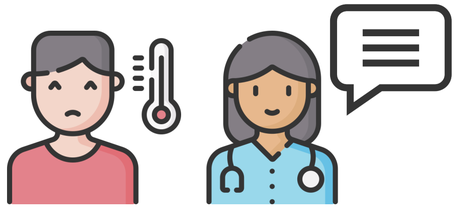
"fever is more consistent with pneumonia **than** with a cold"

"**had there been** chest pressure too, diagnosis would be bronchitis"

Interpretability:
lessons from the social sciences

Human Explanation

(according to philosophy & the social sciences)



- Contrastive
  [van Fraassen, 1988; Lipton, 1990]
- Factual and counterfactual
  [van Fraassen, 1988]
- Sequences of simple statements
  [Hempel, 1962]

"fever is more consistent with pneumonia **than** with a cold"

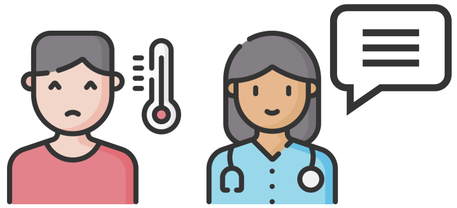
"**had there been** chest pressure too, diagnosis would be bronchitis"

"fever rules out a cold in favor of bronchitis or pneumonia;
among these, shortness of breath suggest the latter"

Interpretability:
lessons from the social sciences

Human Explanation

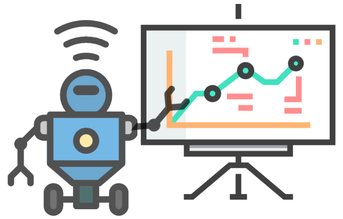
(according to philosophy & the social sciences)



- Contrastive
  [van Fraassen, 1988; Lipton, 1990]
- Factual and counterfactual
  [van Fraassen, 1988]
- Sequences of simple statements
  [Hempel, 1962]

Machine Explanation

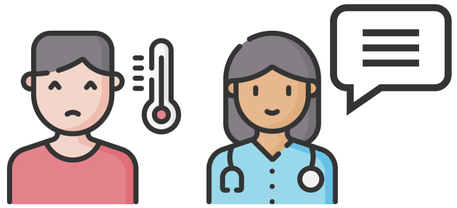
(as typically done in ML interpretability)



Interpretability:
lessons from the social sciences

Human Explanation

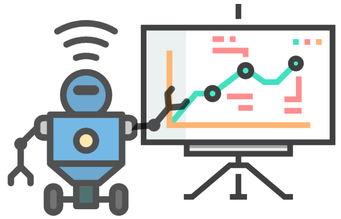
(according to philosophy & the social sciences)



- Contrastive
  [van Fraassen, 1988; Lipton, 1990]
- Factual and counterfactual
  [van Fraassen, 1988]
- Sequences of simple statements
  [Hempel, 1962]

Machine Explanation

(as typically done in ML interpretability)

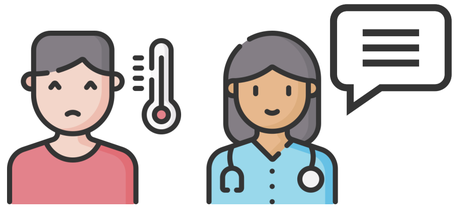


- Absolute, non-contrastive

Interpretability:
lessons from the social sciences

Human Explanation

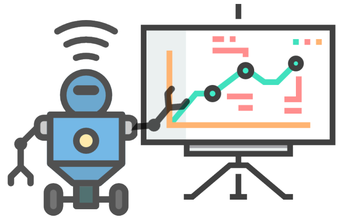
(according to philosophy & the social sciences)



- Contrastive
  [van Fraassen, 1988; Lipton, 1990]
- Factual and counterfactual
  [van Fraassen, 1988]
- Sequences of simple statements
  [Hempel, 1962]

Machine Explanation

(as typically done in ML interpretability)

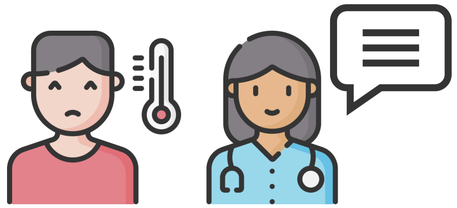


- Absolute, non-contrastive
- Only factual

Interpretability:
lessons from the social sciences

Human Explanation

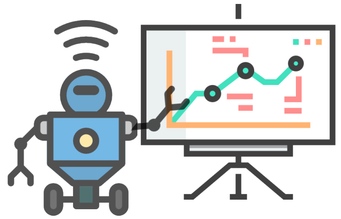
(according to philosophy & the social sciences)



- Contrastive
  [van Fraassen, 1988; Lipton, 1990]
- Factual and counterfactual
  [van Fraassen, 1988]
- Sequences of simple statements
  [Hempel, 1962]

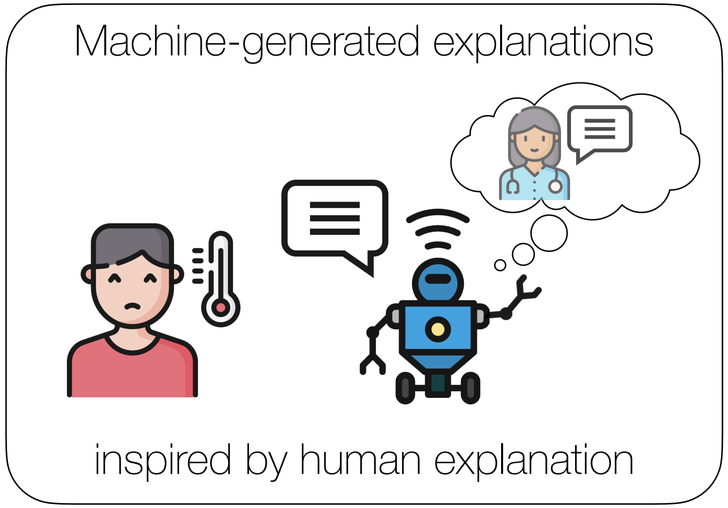
Machine Explanation

(as typically done in ML interpretability)



- Absolute, non-contrastive
- Only factual
- *Monolithic* (one-shot)

Human-centered interpretability:
design principles



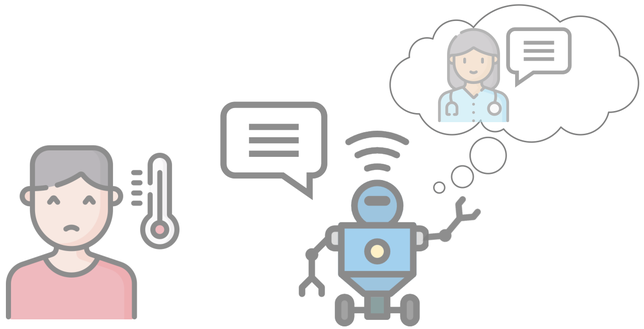
Human-centered interpretability: design principles



Contrastive

explain why y was predicted instead of y'

Machine-generated explanations



inspired by human explanation

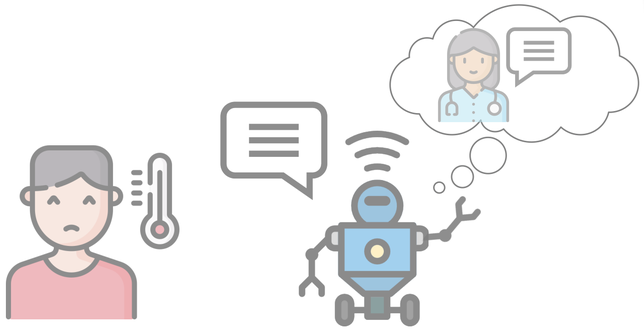
Human-centered interpretability:
design principles



Contrastive

explain why y was predicted instead of y'

Machine-generated explanations



inspired by human explanation



Exhaustive

be able to explain away all other alternatives y'

Human-centered interpretability: design principles

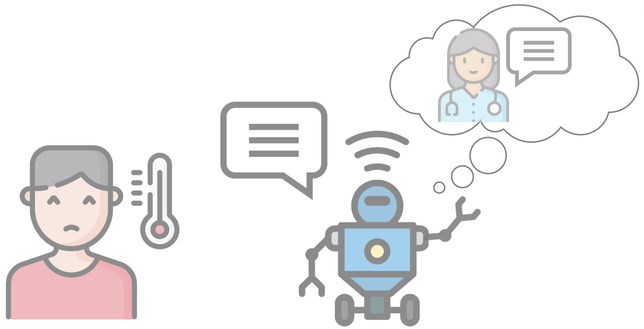
Contrastive

explain why y was predicted instead of y'

Modular

decompose into multiple simple atoms

Machine-generated explanations



inspired by human explanation

Exhaustive

be able to explain away all other alternatives y'

Human-centered interpretability: design principles

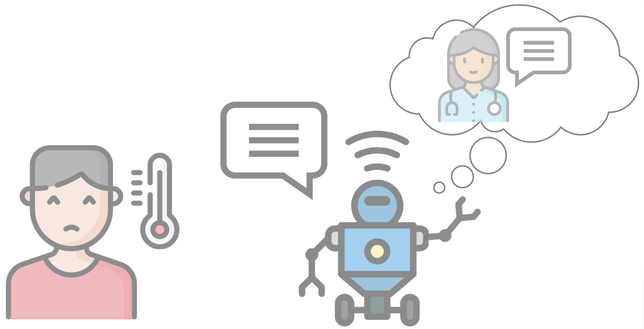
Contrastive

explain why y was predicted instead of y'

Modular

decompose into multiple simple atoms

Machine-generated explanations



inspired by human explanation

Exhaustive

be able to explain away all other alternatives y'

Intelligible

rely on user-friendly units and quantities

Human-centered interpretability: design principles

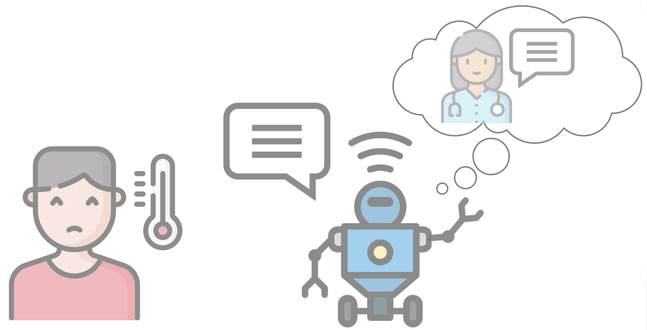
Contrastive

explain why y was predicted instead of y'

Modular

decompose into multiple simple atoms

Machine-generated explanations



inspired by human explanation

Exhaustive

be able to explain away all other alternatives y'

Intelligible

rely on user-friendly units and quantities

Parsimonious

only most relevant facts are provided to the user

The Weight-of-Evidence

[Pierce, 1878; Good, 1950; 1960; 1966; 1968; 1985; ...]

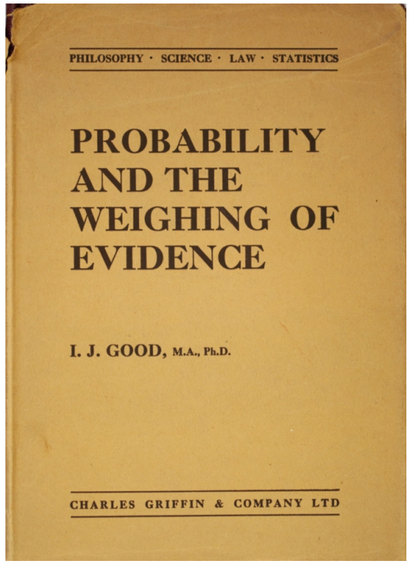
a notion of relevance to analyze variable effects



The Weight-of-Evidence

[Pierce, 1878; Good, 1950; 1960; 1966; 1968; 1985; ...]

a notion of relevance to analyze variable effects







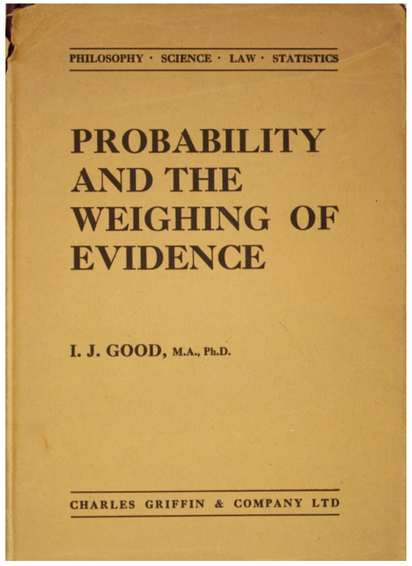
Irving J. Good (1916–2009)

Photo Credit: AMSTAT

The Weight-of-Evidence

[Pierce, 1878; Good, 1950; 1960; 1966; 1968; 1985; ...]

a notion of relevance to analyze variable effects







Irving J. Good (1916–2009)

Given hypothesis $\boldsymbol{h}$ and evidence $\boldsymbol{e}$,

how much does $\boldsymbol{e}$ *speak in favor* (or *against*) $\boldsymbol{h}$?

Photo Credit: AMSTAT

Weight-of-Evidence
in a nutshell

Weight-of-Evidence
in a nutshell

$$\text{woe}(h : e) = \log \frac{\text{Odds}(h \mid e)}{\text{Odds}(h)}$$

① The 'odds' interpretation

Weight-of-Evidence
in a nutshell

$$\text{woe}(h : e) = \log \frac{\text{Odds}(h \mid e)}{\text{Odds}(h)}$$

① The 'odds' interpretation

$$\text{woe}(h : e) = \log \frac{P(e \mid h)}{P(e \mid \bar{h})}$$

② The 'likelihood' interpretation

Weight-of-Evidence
in a nutshell

$$\text{woe}(h : e) = \log \frac{\text{Odds}(h \mid e)}{\text{Odds}(h)} \quad \equiv \quad \text{woe}(h : e) = \log \frac{P(e \mid h)}{P(e \mid \bar{h})}$$

① The 'odds' interpretation

② The 'likelihood' interpretation

Weight-of-Evidence
in a nutshell

$$\text{woe}(h : e) = \log \frac{\text{Odds}(h \mid e)}{\text{Odds}(h)} \quad \equiv \quad \text{woe}(h : e) = \log \frac{P(e \mid h)}{P(e \mid \bar{h})}$$

① The 'odds' interpretation

② The 'likelihood' interpretation

$$\text{woe}(h : e) > 0$$

Weight-of-Evidence
in a nutshell

$$\text{woe}(h : e) = \log \frac{\text{Odds}(h \mid e)}{\text{Odds}(h)} \equiv \text{woe}(h : e) = \log \frac{P(e \mid h)}{P(e \mid \bar{h})}$$

① The 'odds' interpretation

② The 'likelihood' interpretation

$$\text{woe}(h : e) > 0$$

odds of h are higher when taking into account evidence than marginally

Weight-of-Evidence
in a nutshell

$$\text{woe}(h : e) = \log \frac{\text{Odds}(h \mid e)}{\text{Odds}(h)} \equiv \text{woe}(h : e) = \log \frac{P(e \mid h)}{P(e \mid \bar{h})}$$

① The 'odds' interpretation

② The 'likelihood' interpretation

$$\text{woe}(h : e) > 0$$

odds of h are higher when taking into account evidence than marginally

e is more likely if h is true instead of its alternative

Weight-of-Evidence
in a nutshell

$$\text{woe}(h : e) = \log \frac{\text{Odds}(h \mid e)}{\text{Odds}(h)} \quad \equiv \quad \text{woe}(h : e) = \log \frac{P(e \mid h)}{P(e \mid \bar{h})}$$

① The 'odds' interpretation

② The 'likelihood' interpretation

$$\text{woe}(h : e) > 0$$

odds of h are higher when taking into account evidence than marginally

e is more likely if h is true instead of its alternative

"*The evidence speaks in favor of the hypothesis*"

Weight-of-Evidence
in a nutshell

$$\text{woe}(h : e) = \log \frac{\text{Odds}(h \mid e)}{\text{Odds}(h)} \quad \equiv \quad \text{woe}(h : e) = \log \frac{P(e \mid h)}{P(e \mid \bar{h})}$$

① The 'odds' interpretation

② The 'likelihood' interpretation

$\text{woe}(h : e) > 0$

odds of h are higher when taking into account evidence than marginally

e is more likely if h is true instead of its alternative

"*The evidence speaks in favor of the hypothesis*"

In ML context: hypothesis = model prediction (e.g., $y =$ "dog"), evidence = input features ($\mathbf{x}$)

Weight-of-Evidence
advantages

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

  woe$(h/h' : e)$ = WoE in favor of h and against h'

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

$\text{woe}(h/h' : e) =$ WoE in favor of h and against h' → Contrastive

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

woe$(h/h' : e)$ = WoE in favor of h and against h' → Contrastive

cog-sci: log-odd ratios are 'human-friendly'
[e.g., Phillips & Edwards, 1966; Gold & Shadlen, 2001; 2002; Maloney 2012]

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

woe$(h/h' : e)$ = WoE in favor of h and against h' → Contrastive

cog-sci: log-odd ratios are 'human-friendly'
[e.g., Phillips & Edwards, 1966; Gold & Shadlen, 2001; 2002; Maloney 2012] → Intelligible

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

woe$(h/h' : e) =$ WoE in favor of h and against h' → Contrastive

cog-sci: log-odd ratios are 'human-friendly'
[e.g., Phillips & Edwards, 1966; Gold & Shadlen, 2001; 2002; Maloney 2012]

concepts like 'evidence' are grounded in common language → Intelligible

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

woe$(h/h' : e) =$ WoE in favor of h and against h' → Contrastive

cog-sci: log-odd ratios are 'human-friendly'
[e.g., Phillips & Edwards, 1966; Gold & Shadlen, 2001; 2002; Maloney 2012]

concepts like 'evidence' are grounded in common language

→ Intelligible

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

woe$(h/h' : e)$ = WoE in favor of h and against h' → Contrastive

cog-sci: log-odd ratios are 'human-friendly'
[e.g., Phillips & Edwards, 1966; Gold & Shadlen, 2001; 2002; Maloney 2012]

concepts like 'evidence' are grounded in common language → Intelligible

$$\text{woe}(h/h' : \{e_i\}) = \sum_i \log \frac{P(e_i \mid e_{i-1}, \ldots, e_1, h)}{P(e_i \mid e_{i-1}, \ldots, e_1, h')}$$

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

woe$(h/h' : e)$ = WoE in favor of h and against h' → Contrastive

cog-sci: log-odd ratios are 'human-friendly'
[e.g., Phillips & Edwards, 1966; Gold & Shadlen, 2001; 2002; Maloney 2012]

concepts like 'evidence' are grounded in common language → Intelligible

$$\text{woe}(h/h' : \{e_i\}) = \sum_i \log \frac{P(e_i \mid e_{i-1}, \ldots, e_1, h)}{P(e_i \mid e_{i-1}, \ldots, e_1, h')}$$ → Modular

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

woe$(h/h' : e)$ = WoE in favor of h and against h' → Contrastive

cog-sci: log-odd ratios are 'human-friendly'
[e.g., Phillips & Edwards, 1966; Gold & Shadlen, 2001; 2002; Maloney 2012]

concepts like 'evidence' are grounded in common language → Intelligible

$$\text{woe}(h/h' : \{e_i\}) = \sum_i \log \frac{P(e_i \mid e_{i-1}, \ldots, e_1, h)}{P(e_i \mid e_{i-1}, \ldots, e_1, h')}$$ → Modular

$$\underbrace{\log \frac{P(Y=1 \mid X)}{P(Y=0 \mid X)}}_{\text{Posterior log odds}} = \underbrace{\log \frac{P(Y=1)}{P(Y=0)}}_{\text{Prior log odds}} + \underbrace{\log \frac{P(X \mid Y=1)}{P(X \mid Y=0)}}_{\text{Weight of evidence}}$$

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

$\text{woe}(h/h' : e) =$ WoE in favor of h and against h' → Contrastive

cog-sci: log-odd ratios are 'human-friendly'
[e.g., Phillips & Edwards, 1966; Gold & Shadlen, 2001; 2002; Maloney 2012]

concepts like 'evidence' are grounded in common language → Intelligible

$$\text{woe}(h/h' : \{e_i\}) = \sum_i \log \frac{P(e_i \mid e_{i-1}, \ldots, e_1, h)}{P(e_i \mid e_{i-1}, \ldots, e_1, h')}$$

→ Modular

$$\underbrace{\log \frac{P(Y=1 \mid X)}{P(Y=0 \mid X)}}_{\text{Posterior log odds}} = \underbrace{\log \frac{P(Y=1)}{P(Y=0)}}_{\text{Prior log odds}} + \underbrace{\log \frac{P(X \mid Y=1)}{P(X \mid Y=0)}}_{\text{Weight of evidence}}$$

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

$\text{woe}(h/h' : e) =$ WoE in favor of h and against h' → Contrastive

cog-sci: log-odd ratios are 'human-friendly'
[e.g., Phillips & Edwards, 1966; Gold & Shadlen, 2001; 2002; Maloney 2012]

concepts like 'evidence' are grounded in common language → Intelligible

$$\text{woe}(h/h' : \{e_i\}) = \sum_i \log \frac{P(e_i \mid e_{i-1}, \ldots, e_1, h)}{P(e_i \mid e_{i-1}, \ldots, e_1, h')}$$ → Modular

$$\underbrace{\log \frac{P(Y=1 \mid X)}{P(Y=0 \mid X)}}_{\text{Posterior log odds}} = \underbrace{\log \frac{P(Y=1)}{P(Y=0)}}_{\text{Prior log odds}} + \underbrace{\log \frac{P(X \mid Y=1)}{P(X \mid Y=0)}}_{\text{Weight of evidence}}$$

disentangles priors and likelihoods - prevents *base-rate fallacy*
[Tversky & Kahneman, 1974]

Weight-of-Evidence
advantages

- Pros: simple, satisfies our design principles (some immediately, some with simple extension)

$\text{woe}(h/h' : e) =$ WoE in favor of h and against h' → Contrastive

cog-sci: log-odd ratios are 'human-friendly'
[e.g., Phillips & Edwards, 1966; Gold & Shadlen, 2001; 2002; Maloney 2012]

concepts like 'evidence' are grounded in common language → Intelligible

$$\text{woe}(h/h' : \{e_i\}) = \sum_i \log \frac{P(e_i \mid e_{i-1}, \ldots, e_1, h)}{P(e_i \mid e_{i-1}, \ldots, e_1, h')}$$ → Modular

$$\underbrace{\log \frac{P(Y=1 \mid X)}{P(Y=0 \mid X)}}_{\text{Posterior log odds}} = \underbrace{\log \frac{P(Y=1)}{P(Y=0)}}_{\text{Prior log odds}} + \underbrace{\log \frac{P(X \mid Y=1)}{P(X \mid Y=0)}}_{\text{Weight of evidence}}$$

disentangles priors and likelihoods - prevents *base-rate fallacy*
[Tversky & Kahneman, 1974]

WoE explanation taxonomy

Feature Name	cough	dyspnea	aches	weakness	chills	fever
Feature Value	1	0	0	0	1	0

Prediction: 'flu'

WoE explanation taxonomy

Weight of Evidence for 'flu' (vs. other diseases)

Feature Name	cough	dyspnea	aches	weakness	chills	fever
Feature Value	1	0	0	0	1	0

WoE explanation taxonomy

Weight of Evidence for 'flu' (vs. other diseases)

Positive WoE:
the feature's value
speaks in favor
of the hypothesis

Feature Name	cough	dyspnea	aches	weakness	chills	fever
Feature Value	1	0	0	0	1	0

WoE explanation taxonomy

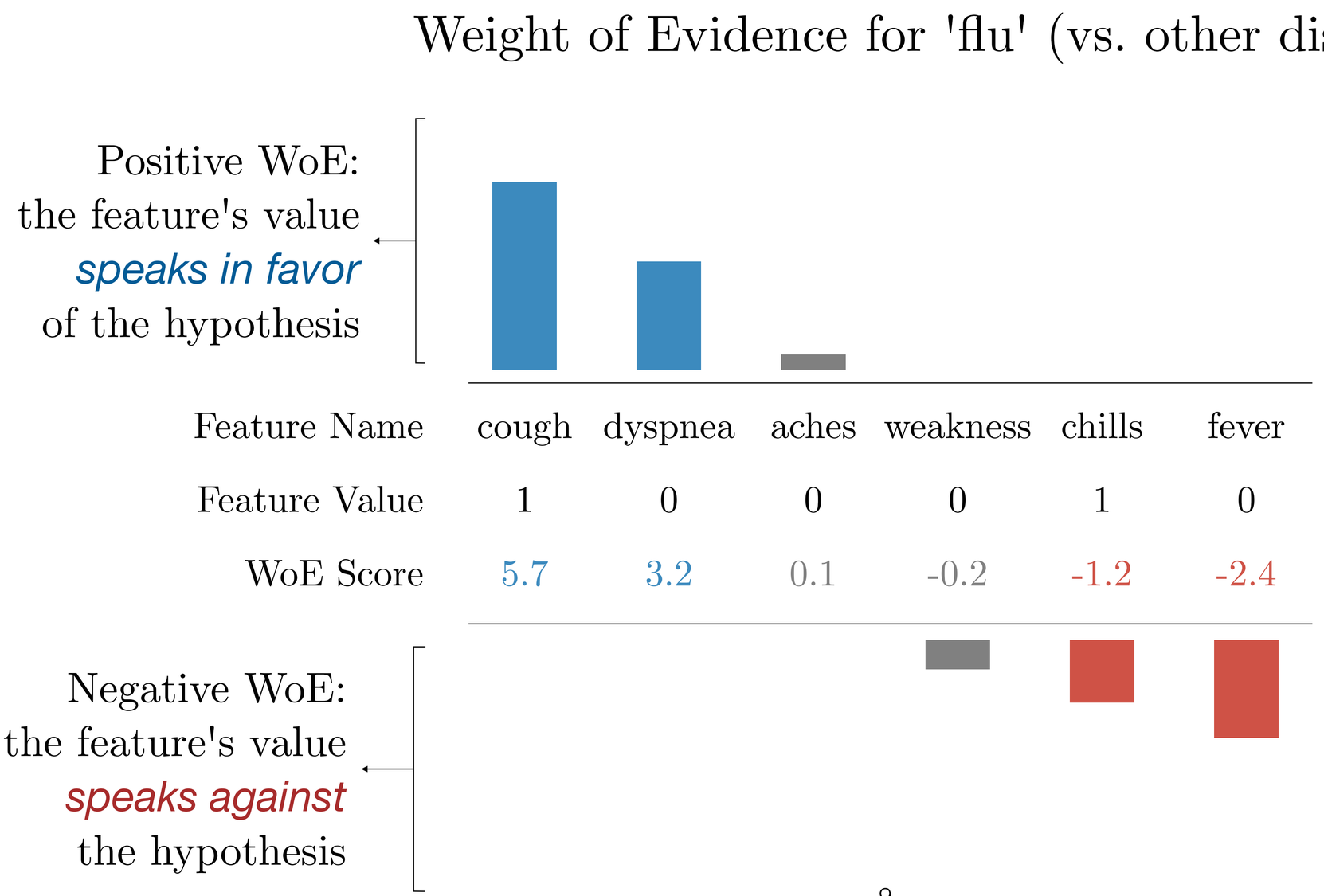
Weight of Evidence for 'flu' (vs. other diseases)

Positive WoE:
the feature's value
speaks in favor
of the hypothesis

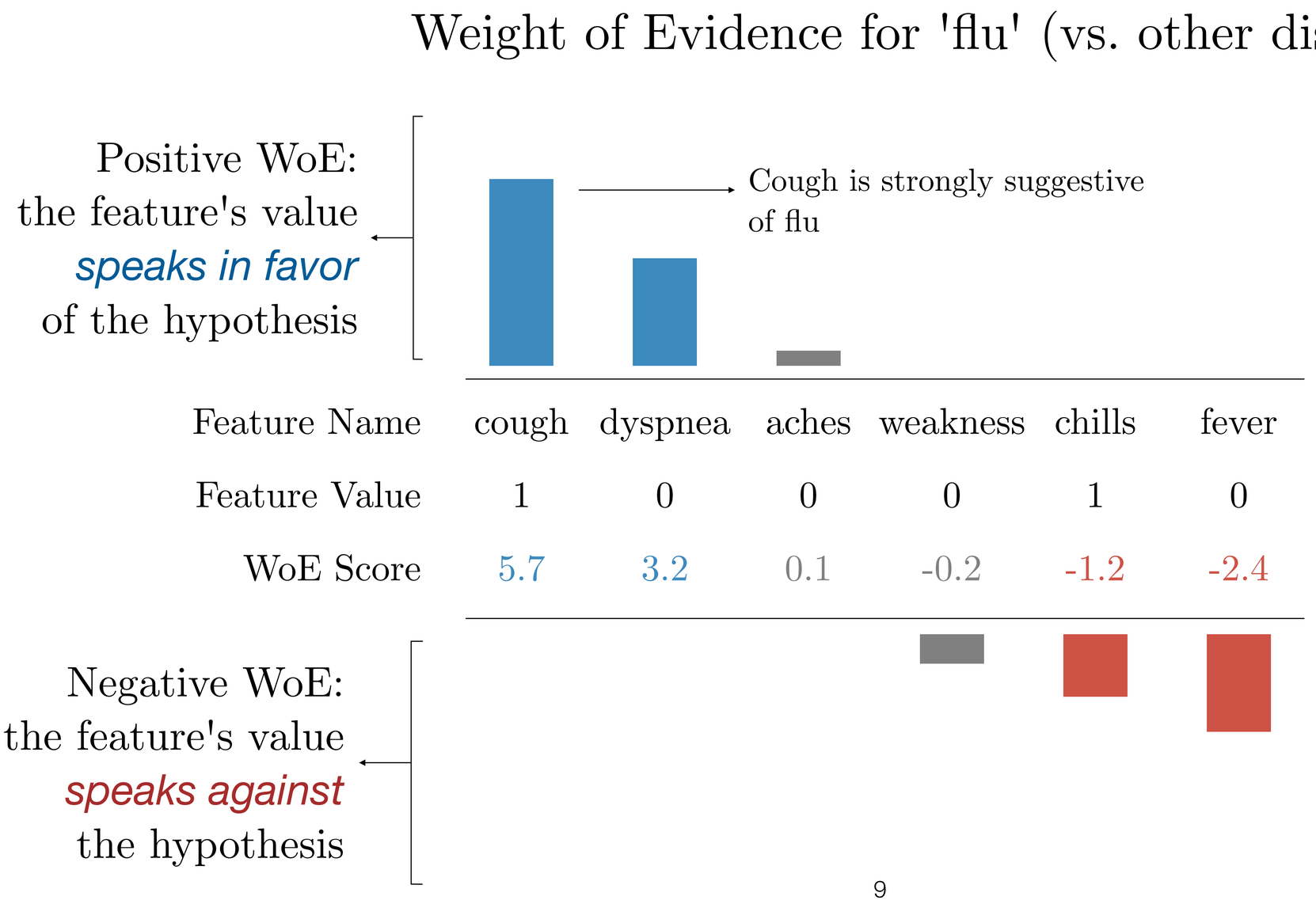
Feature Name	cough	dyspnea	aches	weakness	chills	fever
Feature Value	1	0	0	0	1	0

Negative WoE:
the feature's value
speaks against
the hypothesis

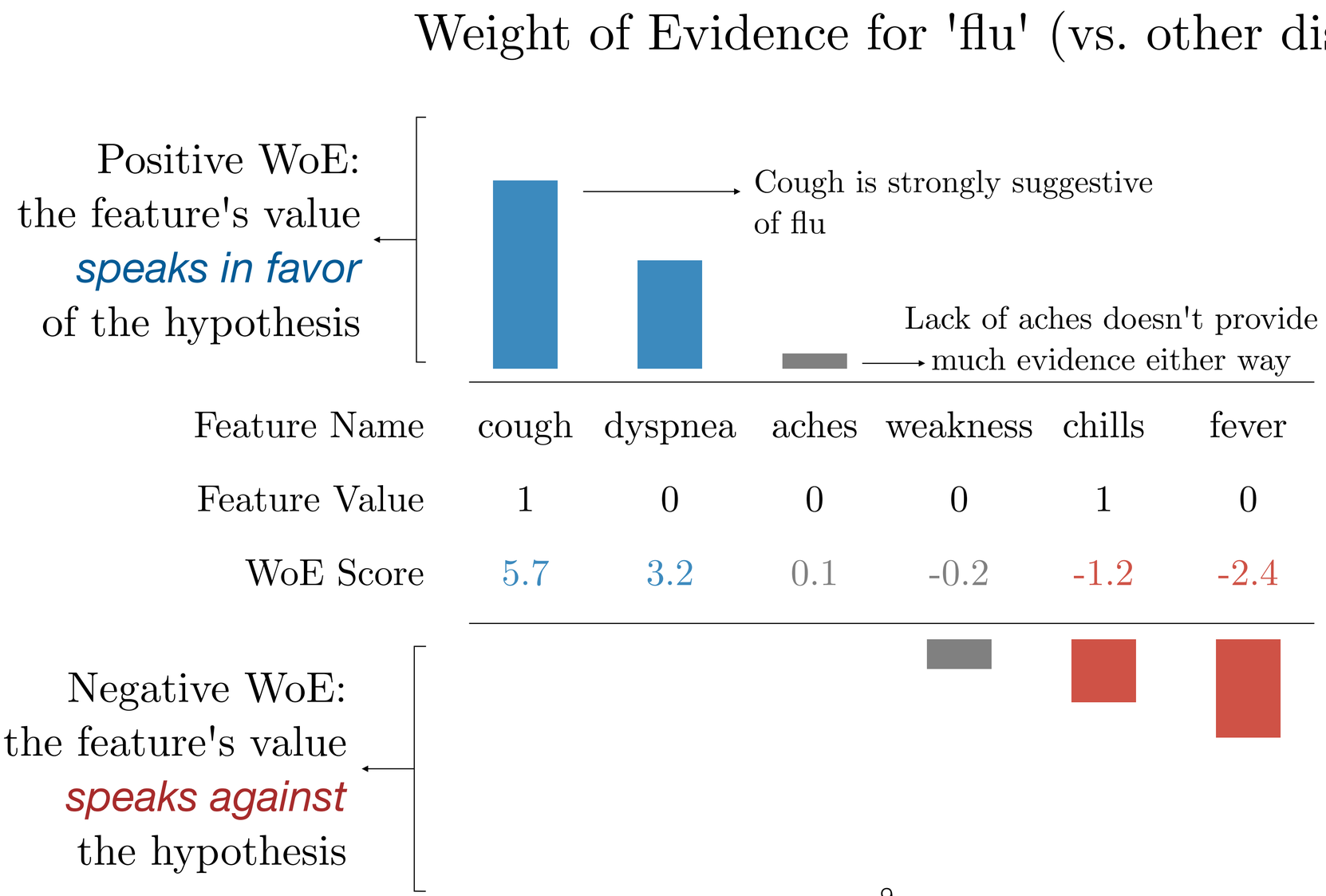
WoE explanation taxonomy



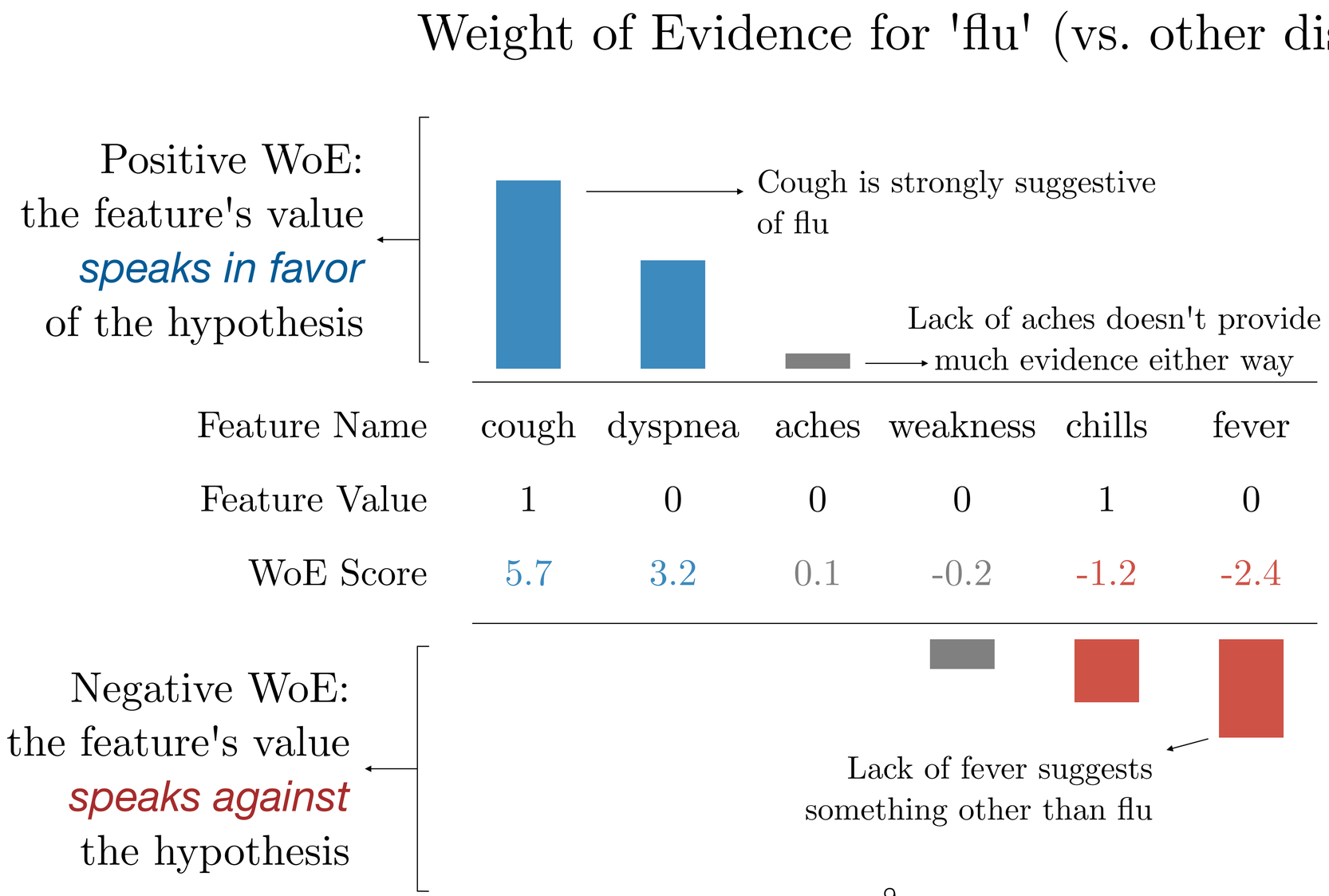
WoE explanation taxonomy



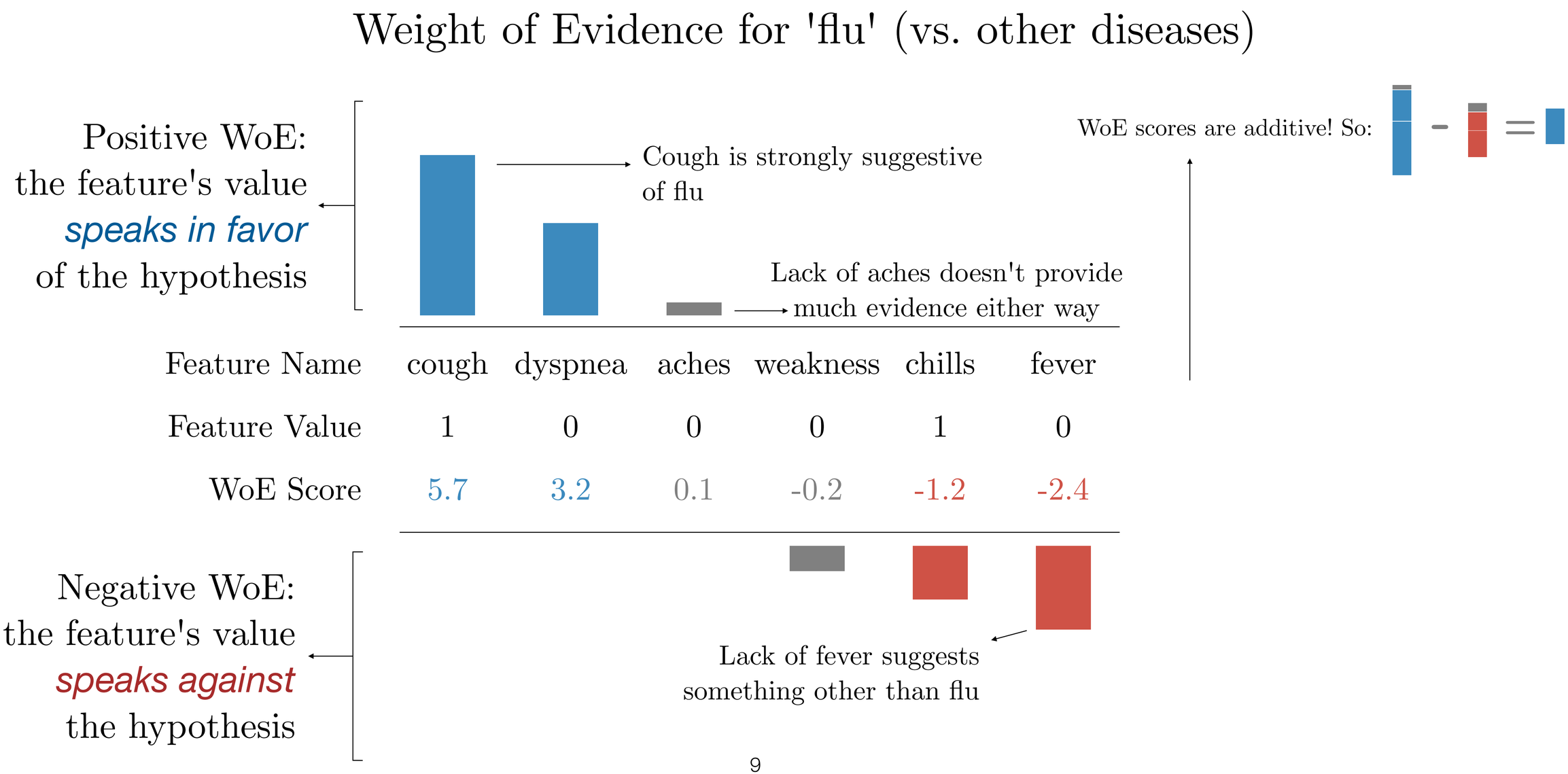
WoE explanation taxonomy



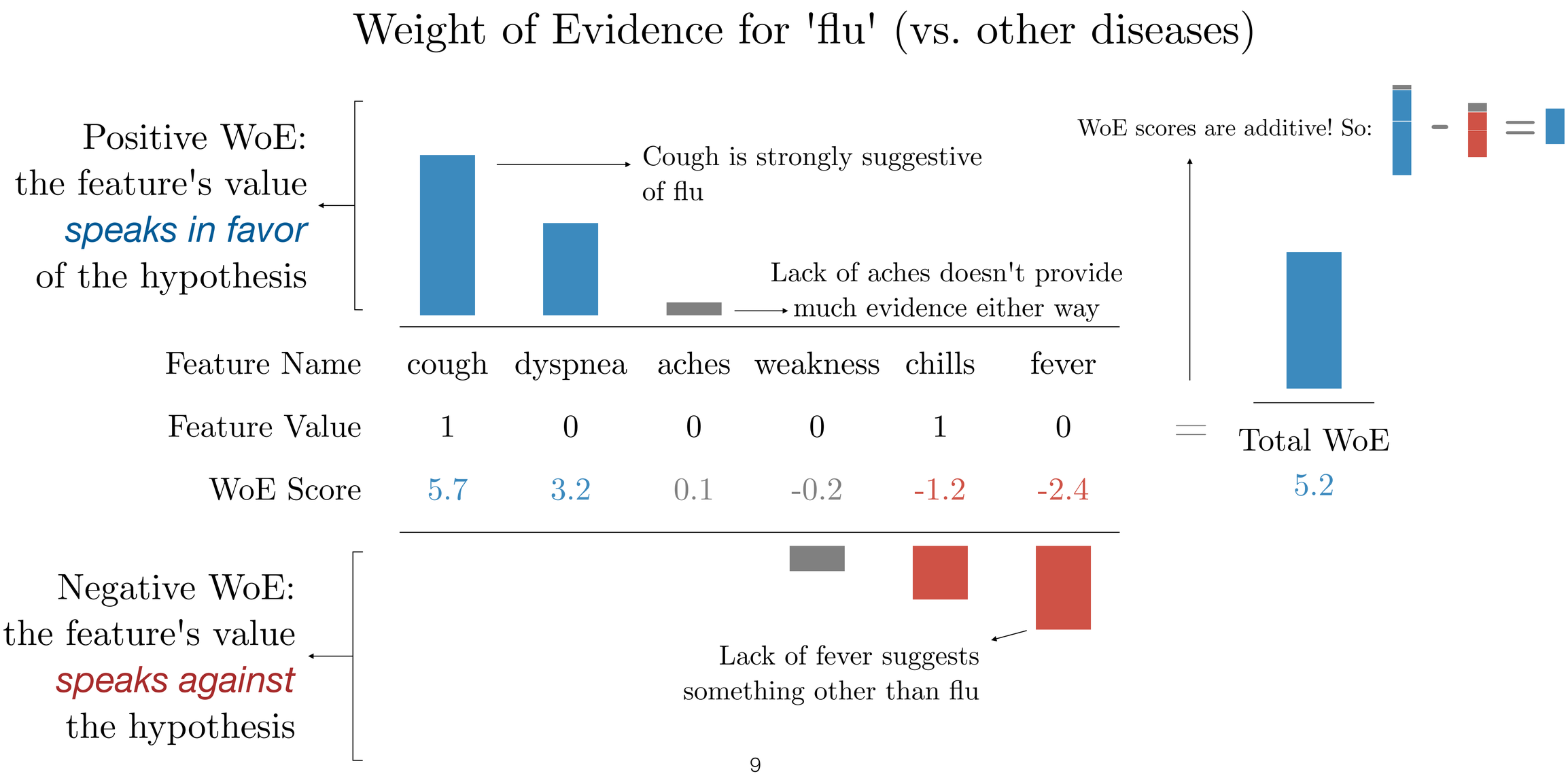
WoE explanation taxonomy



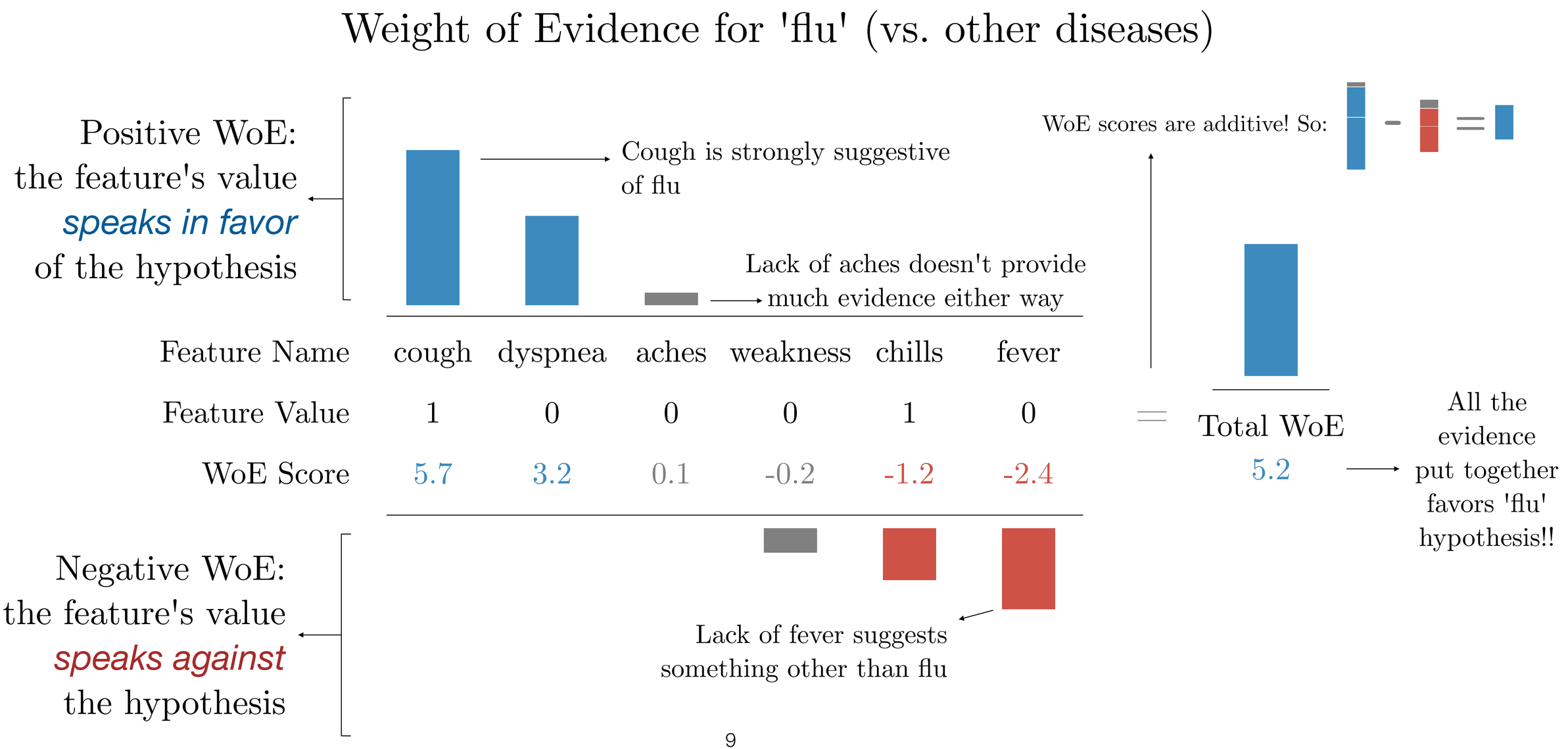
WoE explanation taxonomy



WoE explanation taxonomy



WoE explanation taxonomy



Weight-of-Evidence
challenges

Weight-of-Evidence
challenges

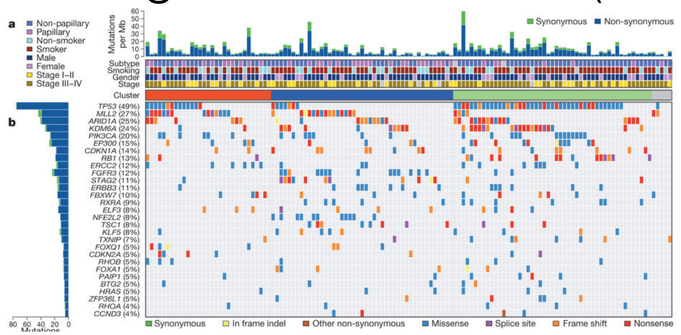
- WoE traditionally used in simple settings (binary classification, few features)

Weight-of-Evidence
challenges

- WoE traditionally used in simple settings (binary classification, few features)
- Challenge: how to extend (and maintain interpretability) in ML settings, e.g.:

Weight-of-Evidence challenges

- WoE traditionally used in simple settings (binary classification, few features)
- Challenge: how to extend (and maintain interpretability) in ML settings, e.g.:

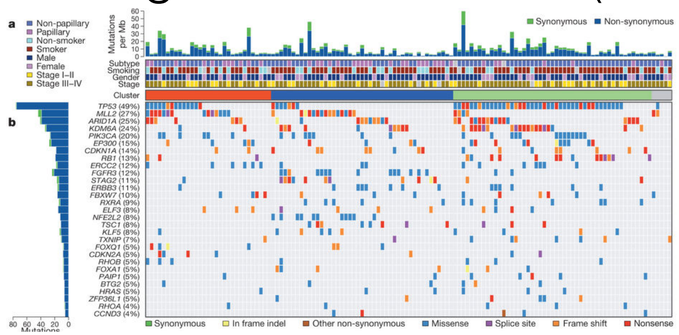


Ref.: Nature 507, 315–322 (20 March 2014) doi:10.1038/nature12965

high-dimensional inputs

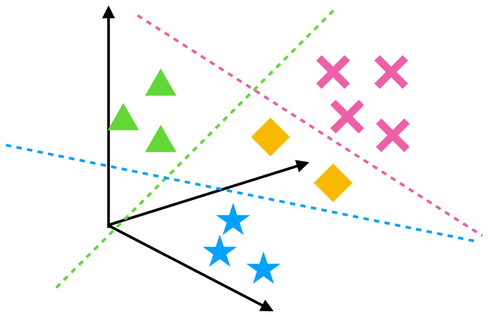
Weight-of-Evidence challenges

- WoE traditionally used in simple settings (binary classification, few features)
- Challenge: how to extend (and maintain interpretability) in ML settings, e.g.:



Ref.: Nature 507, 315–322 (20 March 2014) doi:10.1038/nature12965

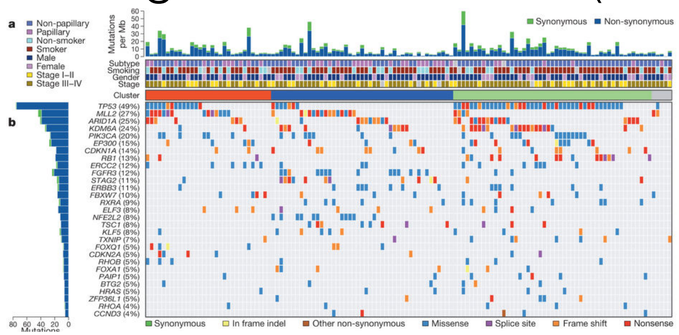
high-dimensional inputs



multi-class classification

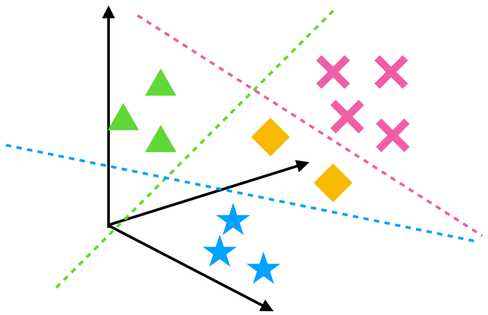
Weight-of-Evidence challenges

- WoE traditionally used in simple settings (binary classification, few features)
- Challenge: how to extend (and maintain interpretability) in ML settings, e.g.:



Ref.: Nature 507, 315–322 (20 March 2014) doi:10.1038/nature12965

high-dimensional inputs

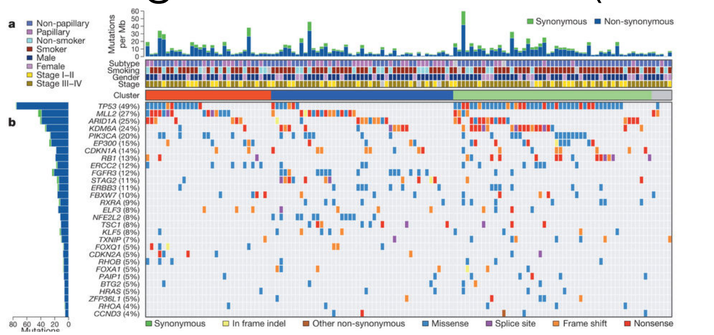


multi-class classification

grouped explanations:
explain in terms of aggregated 'atoms'
e.g., groups of features

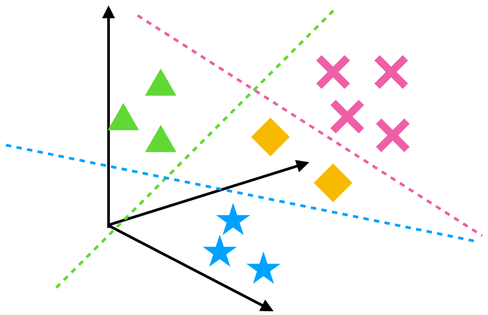
Weight-of-Evidence challenges

- WoE traditionally used in simple settings (binary classification, few features)
- Challenge: how to extend (and maintain interpretability) in ML settings, e.g.:



Ref.: Nature 507, 315–322 (20 March 2014) doi:10.1038/nature12965

high-dimensional inputs



multi-class classification

grouped explanations:
explain in terms of aggregated 'atoms'
e.g., groups of features

sequential explanations:
break up explanation into sequence
of contrastive comparisons

WoE explanation: feature grouping

What if there are too many features?

WoE explanation: feature grouping

What if there are too many features?

Weight of Evidence for 'flu' (vs. other diseases)

Feature Name	cough	dyspnea	aches	weakness	chills	fever
Feature Value	1	0	0	0	1	0

WoE explanation: feature grouping

What if there are too many features?

Weight of Evidence for 'flu' (vs. other diseases)

Feature Group	**Respiratory**		**Body**		**Temperature**	
Feature Name	cough	dyspnea	aches	weakness	chills	fever
Feature Value	1	0	0	0	1	0

WoE explanation: feature grouping

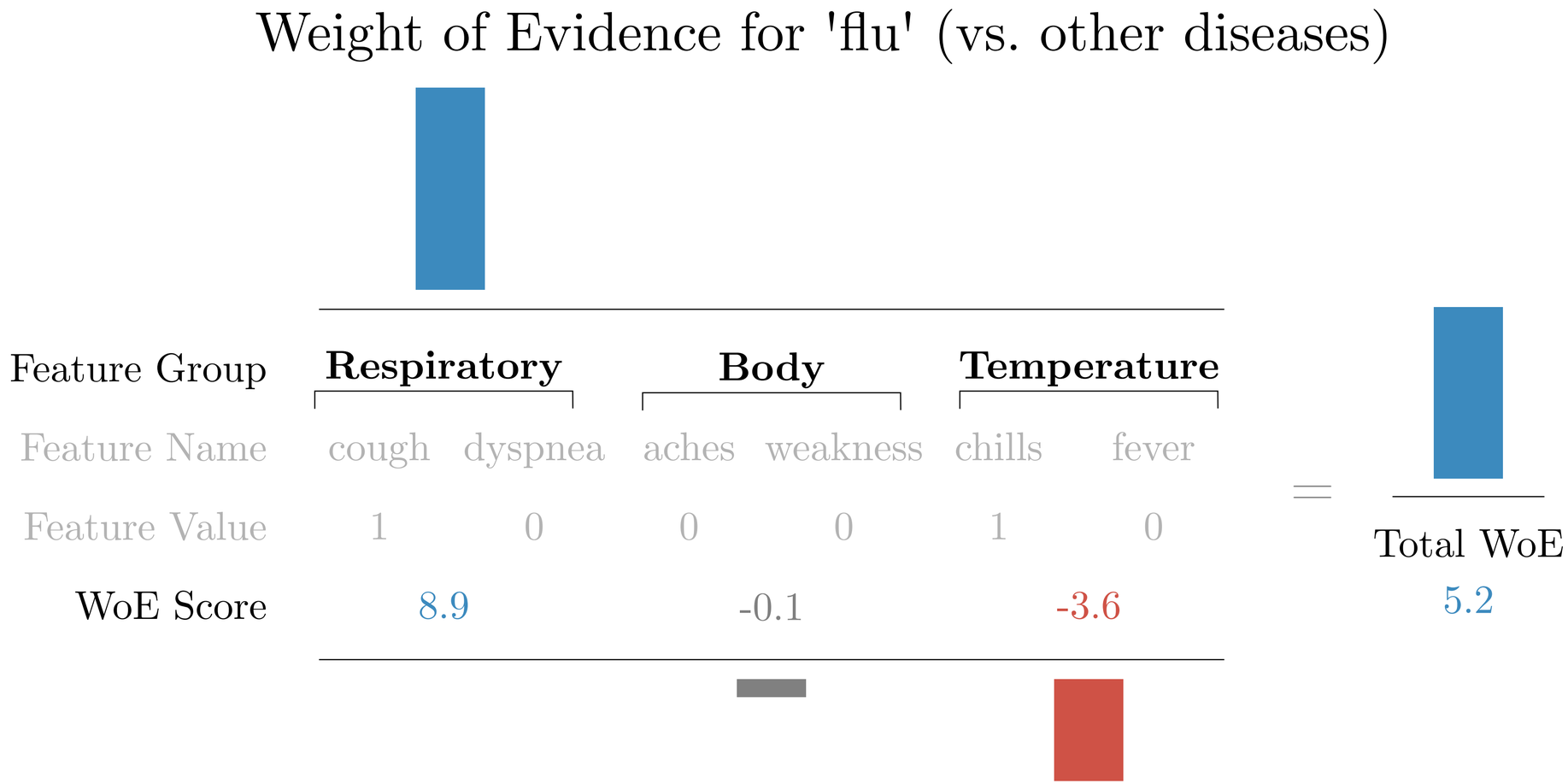
What if there are too many features?

Weight of Evidence for 'flu' (vs. other diseases)

Feature Group	**Respiratory**		**Body**		**Temperature**	
Feature Name	cough	dyspnea	aches	weakness	chills	fever
Feature Value	1	0	0	0	1	0
WoE Score	8.9		-0.1		-3.6	

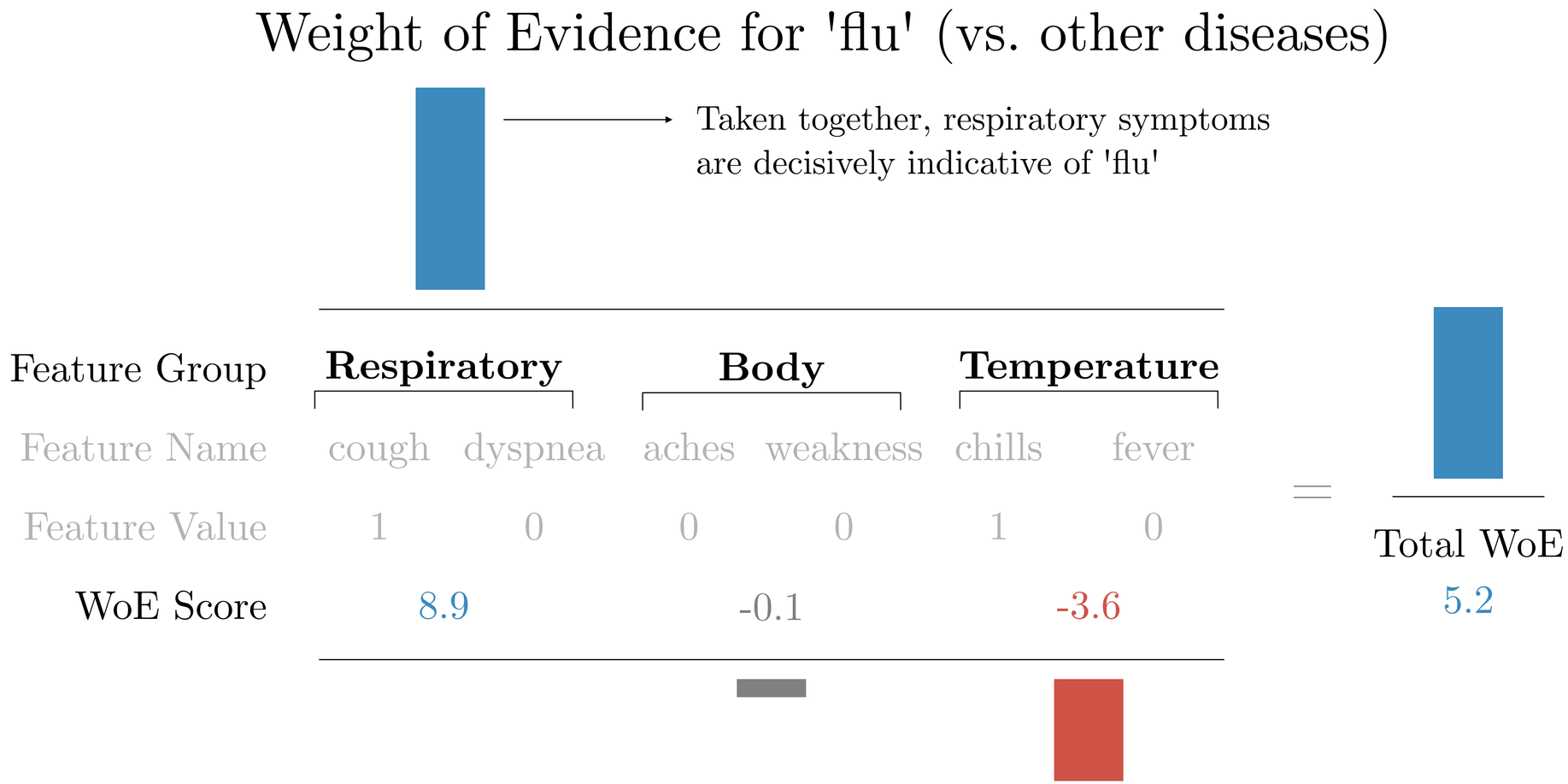
WoE explanation: feature grouping

What if there are too many features?



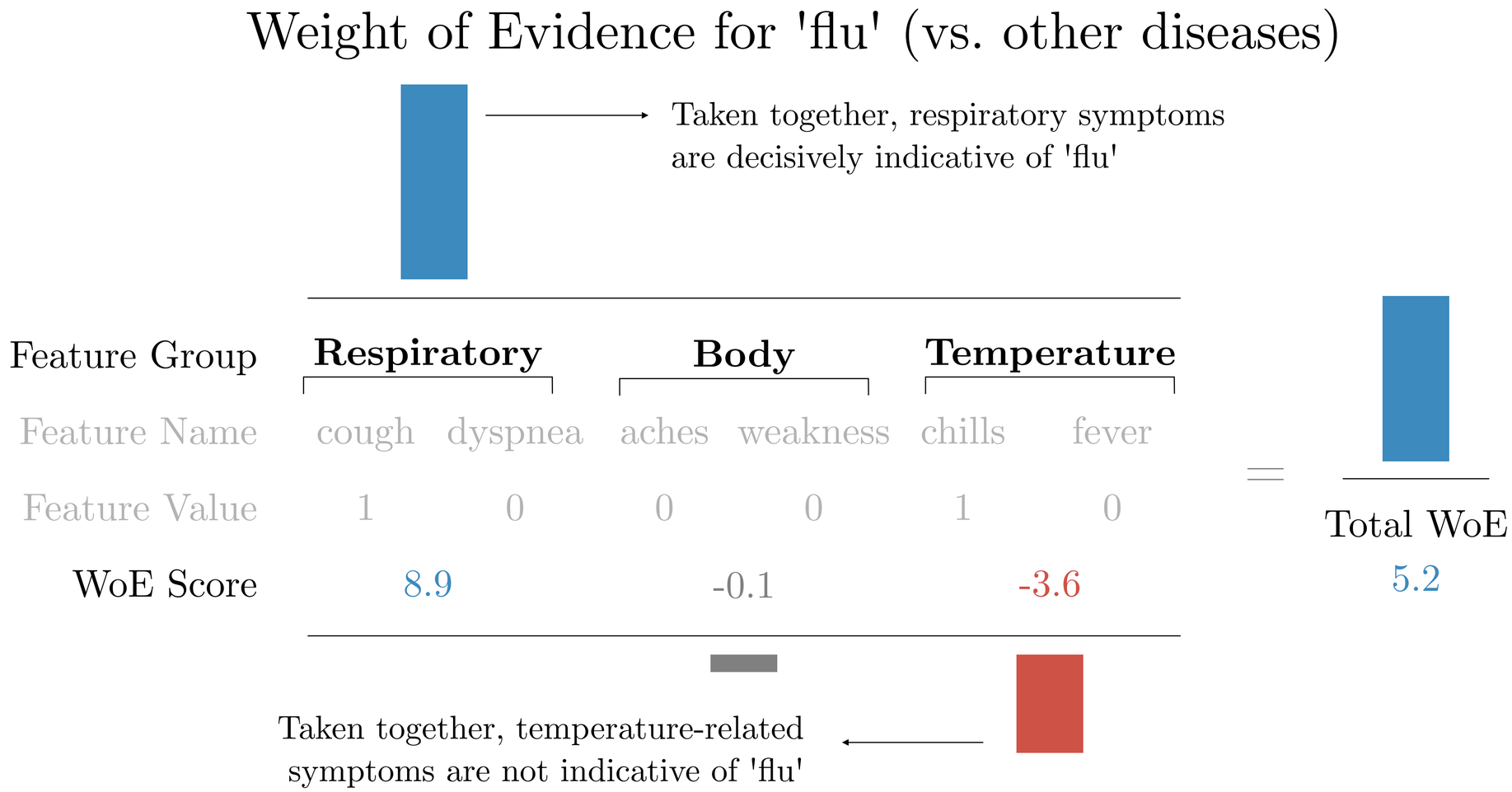
WoE explanation: feature grouping

What if there are too many features?



WoE explanation: feature grouping

What if there are too many features?



WoE explanation: sequential breakdown

What if there are too many possible classes?

WoE explanation: sequential breakdown

What if there are too many possible classes?

1. Flu
2. Parainfluenza
3. Adenovirus
4. Rhinovirus
5. SARS-CoV
6. Strep throat
7. Scarlet fever
8. Diphtheria
9. Pertussis
10. Tuberculosis

WoE explanation: sequential breakdown

What if there are too many possible classes?

1. Flu
2. Parainfluenza
3. Adenovirus
4. Rhinovirus
5. SARS-CoV
6. Strep throat
7. Scarlet fever
8. Diphtheria
9. Pertussis
10. Tuberculosis

viral diseases

WoE explanation: sequential breakdown

What if there are too many possible classes?

1. Flu
2. Parainfluenza
3. Adenovirus
4. Rhinovirus
5. SARS-CoV

viral diseases

6. Strep throat
7. Scarlet fever
8. Diphtheria
9. Pertussis
10. Tuberculosis

bacterial diseases

WoE explanation: sequential breakdown

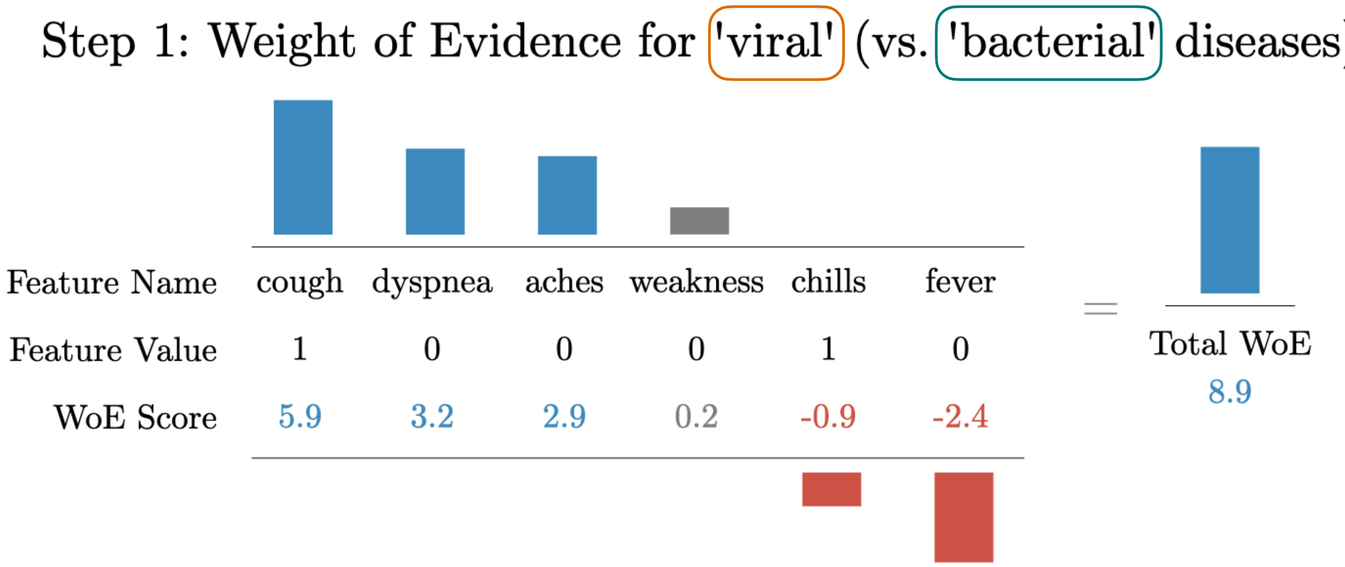
What if there are too many possible classes?

1. Flu
2. Parainfluenza
3. Adenovirus
4. Rhinovirus
5. SARS-CoV

viral diseases

6. Strep throat
7. Scarlet fever
8. Diphtheria
9. Pertussis
10. Tuberculosis

bacterial diseases



WoE explanation: sequential breakdown

What if there are too many possible classes?

1. Flu
2. Parainfluenza
3. Adenovirus
4. Rhinovirus
5. SARS-CoV

viral diseases

6. Strep throat
7. Scarlet fever
8. Diphtheria
9. Pertussis
10. Tuberculosis

bacterial diseases

WoE explanation: sequential breakdown

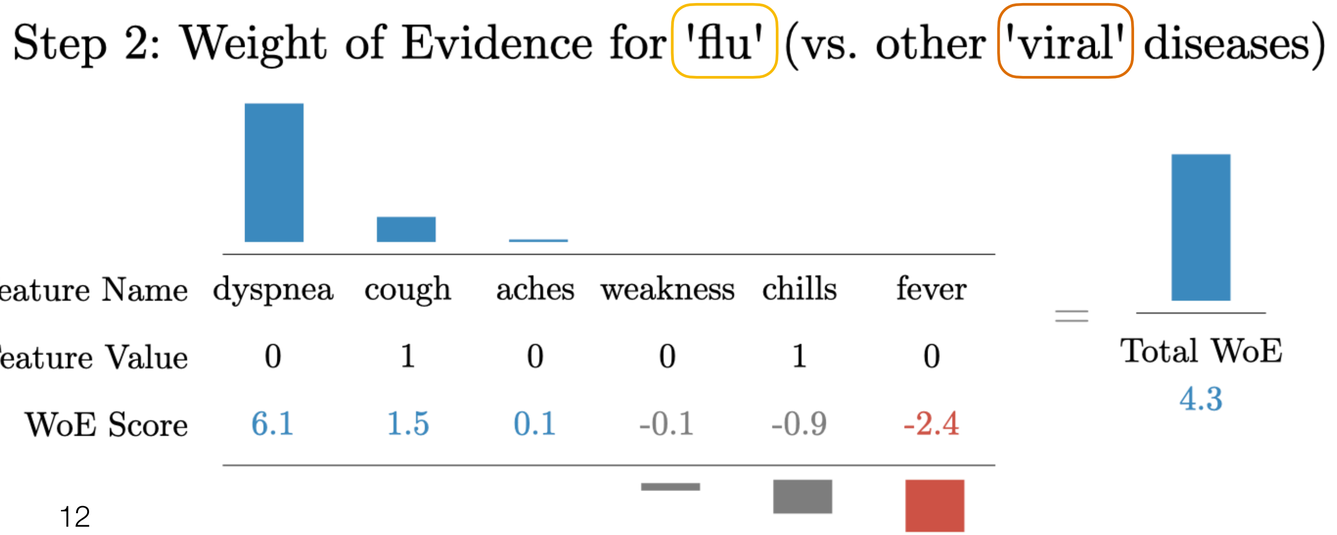
What if there are too many possible classes?

1. Flu
2. Parainfluenza
3. Adenovirus
4. Rhinovirus
5. SARS-CoV

viral diseases

6. Strep throat
7. Scarlet fever
8. Diphtheria
9. Pertussis
10. Tuberculosis

bacterial diseases



Weight-of-Evidence implementation

Algorithm 1 WoE meta-algorithm for complex models

Input: Instance $X \in \mathbb{R}^n$, prediction $y^* \in \{1, \ldots, k\}$
Parameters: Features $\mathcal{A} = \{\{1\}, \ldots, \{n\}\}$ or feature groups $\mathcal{A} = \{S_1, \ldots, S_m\}$
Initialize $\mathrm{U}_0 \leftarrow \{1, \ldots, k\}$
$t \leftarrow 0$
while $|\mathrm{U}_t| > 1$ **do**
    $t \leftarrow t + 1$
    $\mathrm{U}_t \leftarrow \mathrm{SelectHypothesis}(\mathrm{U}_{t-1}, y^*)$
    $\overline{\mathrm{U}}_t \leftarrow \mathrm{U}_{t-1} \setminus \mathrm{U}_t$ {relative complement}
    $\pi(\mathrm{U}_t) \leftarrow \log \frac{P(y \in \mathrm{U}_t)}{P(y \in \overline{\mathrm{U}}_t)}$ {prior log-odds}
    for $i = 1, \ldots, |\mathcal{A}|$ **do**
        $\omega_i^t \leftarrow \mathrm{woe}(y \in \mathrm{U}_t / y \in \overline{\mathrm{U}}_t : X_{\mathcal{A}_i} \mid X_{\mathcal{A}_{i-1}}, \ldots, X_{\mathcal{A}_1})$
    end for
    $\Omega_t \leftarrow \sum_{i=1}^{|\mathcal{A}|} \omega_i^t$
    $\mathrm{DisplayExplanation}(\mathrm{U}_t, \overline{\mathrm{U}}_t, \mathcal{A}, \pi(\mathrm{U}_t), \{\omega_i^t\}_i, \Omega_t)$
end while

Weight-of-Evidence implementation

Algorithm 1 WoE meta-algorithm for complex models

Input: Instance $X \in \mathbb{R}^n$, prediction $y^* \in \{1, \ldots, k\}$
Parameters: Features $\mathcal{A} = \{\{1\}, \ldots, \{n\}\}$ or feature groups $\mathcal{A} = \{S_1, \ldots, S_m\}$
Initialize $\mathsf{U}_0 \leftarrow \{1, \ldots, k\}$
$t \leftarrow 0$
while $|\mathsf{U}_t| > 1$ **do**
  $t \leftarrow t + 1$
  $\mathsf{U}_t \leftarrow \text{SelectHypothesis}(\mathsf{U}_{t-1}, y^*)$
  $\overline{\mathsf{U}}_t \leftarrow \mathsf{U}_{t-1} \setminus \mathsf{U}_t$ {relative complement}
  $\pi(\mathsf{U}_t) \leftarrow \log \frac{P(y \in \mathsf{U}_t)}{P(y \in \overline{\mathsf{U}}_t)}$ {prior log-odds}
  for $i = 1, \ldots, |\mathcal{A}|$ **do**
    $\omega_i^t \leftarrow \text{woe}(y \in \mathsf{U}_t / y \in \overline{\mathsf{U}}_t : X_{\mathcal{A}_i} \mid X_{\mathcal{A}_{i-1}}, \ldots, X_{\mathcal{A}_1})$
  end for
  $\Omega_t \leftarrow \sum_{i=1}^{|\mathcal{A}|} \omega_i^t$
  $\text{DisplayExplanation}(\mathsf{U}_t, \overline{\mathsf{U}}_t, \mathcal{A}, \pi(\mathsf{U}_t), \{\omega_i^t\}_i, \Omega_t)$
end while

U_t : classes remaining to be explained after t steps

select main and alternative hypotheses to contrast

compute per-atom weight of evidence values

Weight-of-Evidence implementation

Algorithm 1 WoE meta-algorithm for complex models

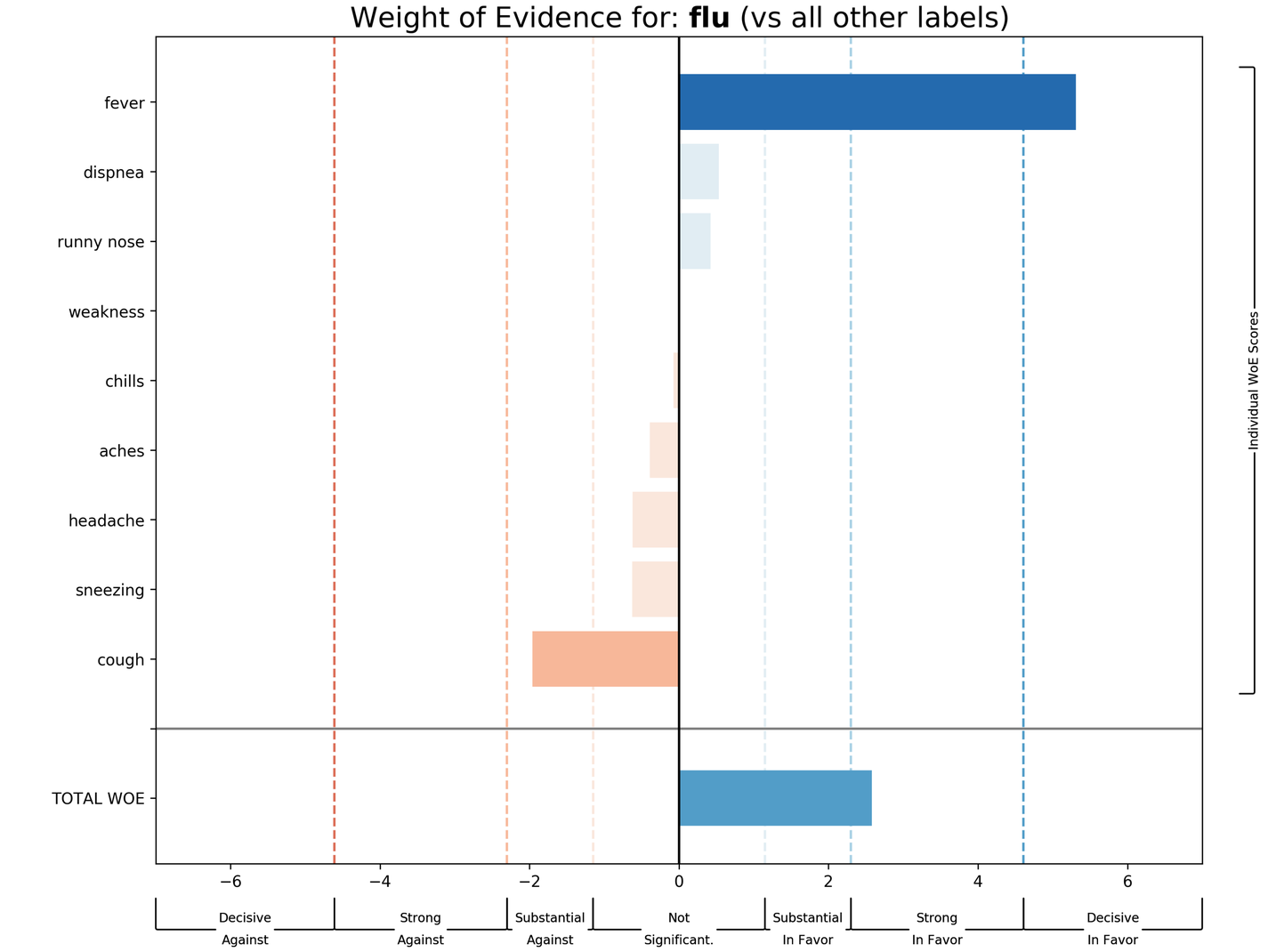
Input: Instance $X \in \mathbb{R}^n$, prediction $y^* \in \{1, \ldots, k\}$
Parameters: Features $\mathcal{A} = \{\{1\}, \ldots, \{n\}\}$ or feature groups $\mathcal{A} = \{S_1, \ldots, S_m\}$
Initialize $\mathsf{U}_0 \leftarrow \{1, \ldots, k\}$
$t \leftarrow 0$
while $|\mathsf{U}_t| > 1$ **do**
  $t \leftarrow t + 1$
  $\mathsf{U}_t \leftarrow \text{SelectHypothesis}(\mathsf{U}_{t-1}, y^*)$
  $\overline{\mathsf{U}}_t \leftarrow \mathsf{U}_{t-1} \setminus \mathsf{U}_t$ {relative complement}
  $\pi(\mathsf{U}_t) \leftarrow \log \frac{P(y \in \mathsf{U}_t)}{P(y \in \overline{\mathsf{U}}_t)}$ {prior log-odds}
  for $i = 1, \ldots, |\mathcal{A}|$ **do**
    $\omega_i^t \leftarrow \text{woe}(y \in \mathsf{U}_t / y \in \overline{\mathsf{U}}_t : X_{\mathcal{A}_i} \,|\, X_{\mathcal{A}_{i-1}}, \ldots, X_{\mathcal{A}_1})$
  end for
  $\Omega_t \leftarrow \sum_{i=1}^{|\mathcal{A}|} \omega_i^t$
  $\text{DisplayExplanation}(\mathsf{U}_t, \overline{\mathsf{U}}_t, \mathcal{A}, \pi(\mathsf{U}_t), \{\omega_i^t\}_i, \Omega_t)$
end while

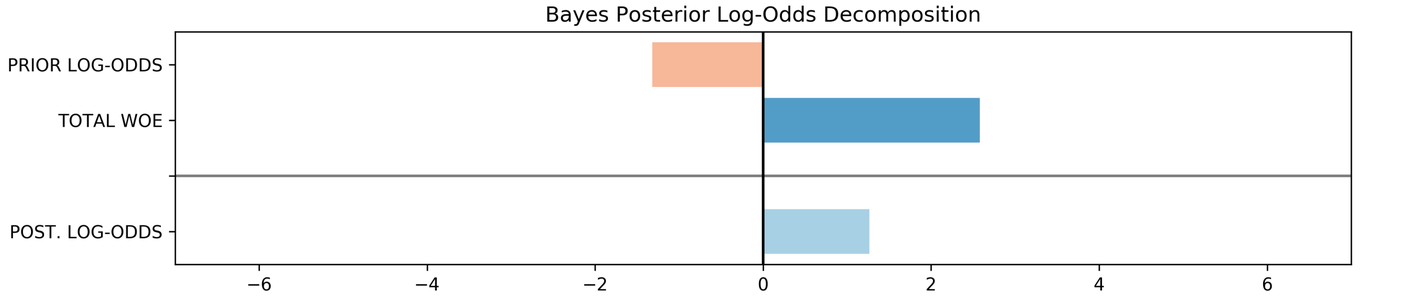
U_t : classes remaining to be explained after t steps

select main and alternative hypotheses to contrast

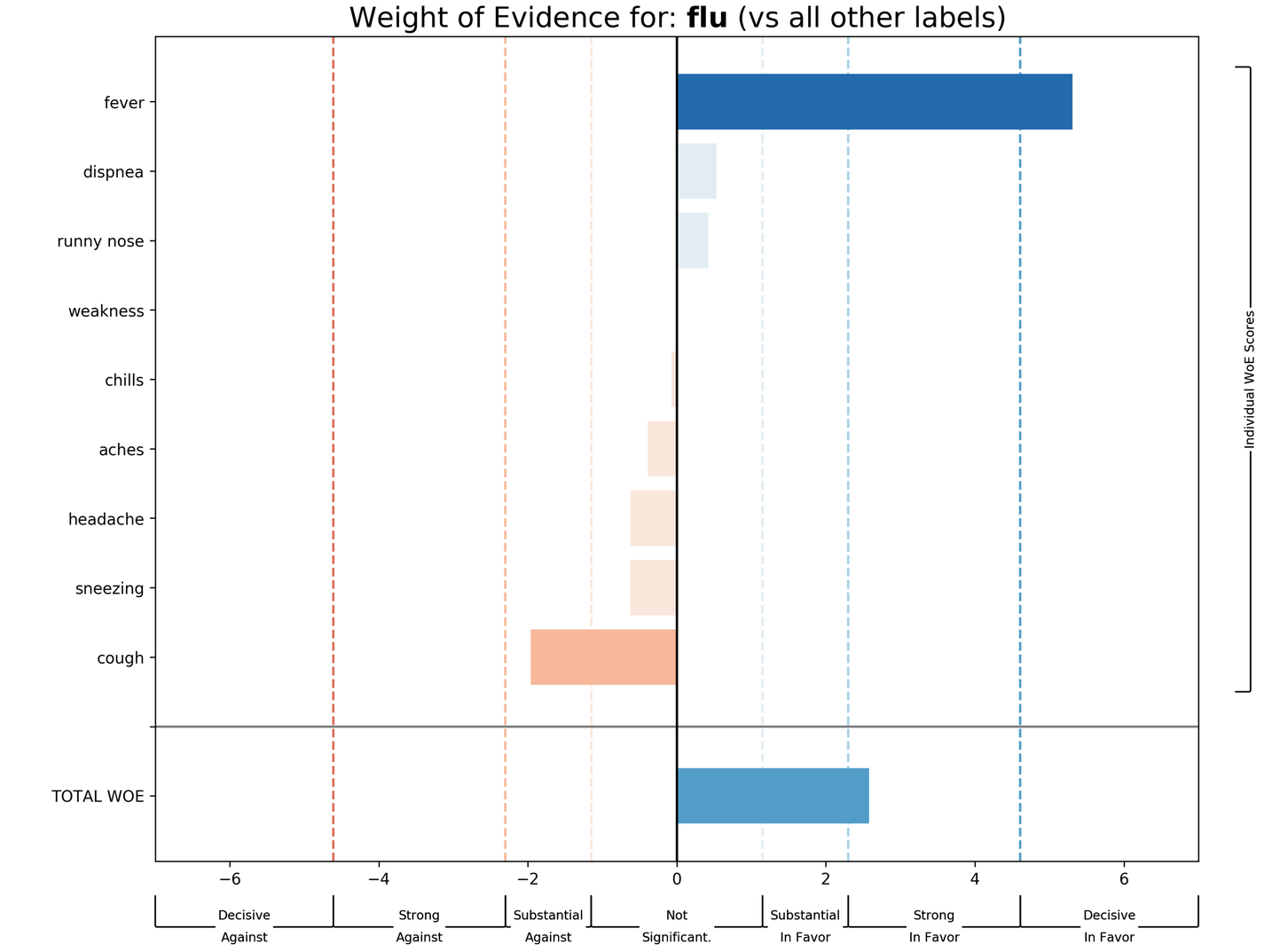
compute per-atom weight of evidence values

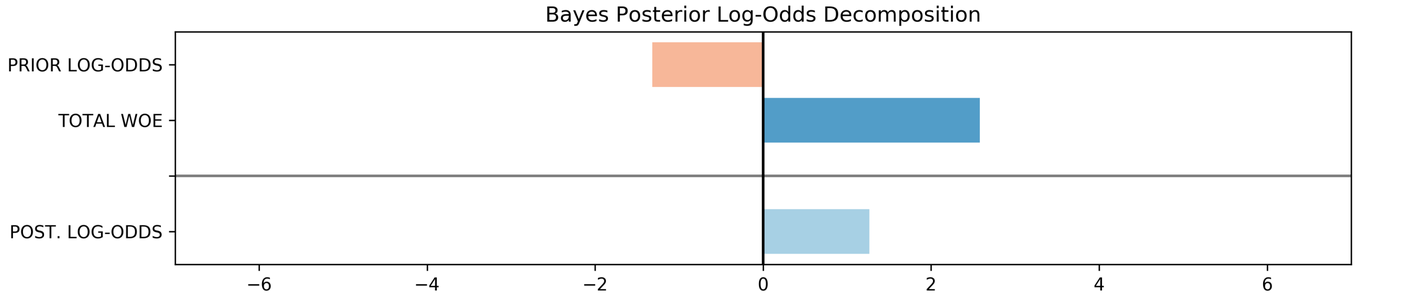
total weight of evidence is sum of individual ones





actual explanation generated by our tool





actual explanation generated by our tool

'vertical addition' of WoE values

color-coding with rule-of-thumb significance bands

WoE disentangles base-rates from posteriors:

prior odds + woe = posterior odds

WoE Interpretability
user study - setup

WoE Interpretability
user study - setup

- Qualitative in-context user study with 10 MSFT ML practitioners, various levels of experience

WoE Interpretability
user study - setup

- Qualitative in-context user study with 10 MSFT ML practitioners, various levels of experience
- "Think-aloud" protocol - participants were asked to verbalize thought process throughout

WoE Interpretability
user study - setup

- Qualitative in-context user study with 10 MSFT ML practitioners, various levels of experience
- "Think-aloud" protocol - participants were asked to verbalize thought process throughout
- Tutorial (~45 min): introduce weight-of-evidence, example explanations

WoE Interpretability
user study - setup

- Qualitative in-context user study with 10 MSFT ML practitioners, various levels of experience
- "Think-aloud" protocol - participants were asked to verbalize thought process throughout
- Tutorial (~45 min): introduce weight-of-evidence, example explanations
- Practitioners often use interpretability tools without fully understanding them [Kaur et al., 2020]

WoE Interpretability
user study - setup

- Qualitative in-context user study with 10 MSFT ML practitioners, various levels of experience
- "Think-aloud" protocol - participants were asked to verbalize thought process throughout
- Tutorial (~45 min): introduce weight-of-evidence, example explanations
- Practitioners often use interpretability tools without fully understanding them [Kaur et al., 2020]
- Main Study (~40 min): real news classification dataset, questions on predictions

WoE Interpretability
user study - setup

- Qualitative in-context user study with 10 MSFT ML practitioners, various levels of experience
- "Think-aloud" protocol - participants were asked to verbalize thought process throughout
- Tutorial (~45 min): introduce weight-of-evidence, example explanations
- Practitioners often use interpretability tools without fully understanding them [Kaur et al., 2020]
- Main Study (~40 min): real news classification dataset, questions on predictions
- Follow-up Interview (~10 min): feedback

Thank you for participating in our study!

The study is structured as follows:

1. **Tutorial**: Overview of background concepts that will be used in the rest of the study.
2. **Setup Description**: Presentation of the dataset used, descriptive statistics, etc.
3. **Main Component**: Using the interpretability tool to answer questions about an ML model.
4. **Follow-Up**: Questionnaire and interview.

Please keep this tutorial open and handy while you complete the study. **You're welcome (and encouraged!) to refer back to it at any point.**

Concepts covered in this tutorial

Weight of Evidence:

- Definition
- Interpretation
- Binary vs. Multiclass
- Grouping Features
- Sequential Explanation
- Example

Tutorial

Weight Of Evidence



Setup

The study will focus on machine learning classifiers, so we will introduce the Weight of Evidence using a (made up) classification example.

Suppose we are trying to predict the variable Y, where:

User Study on Interpretability - Main Sections

BEFORE YOU BEGIN:

- Make sure this is running on the **Python 3.6 Kernel** (not Python 3). This can be changed in the 'Kernel' menu above.
- Go to "Cell" -> "Run All" to start executing preliminary commands in the background while you read the instructions below.

Description:

Throughout this study we will be using the 'Online News Popularity' dataset.

Each instance in this datas length, contents, polarity and some metadata. We will provide s training/testing).

The task is to predict the c/ news article was published.

The 'Preliminaries' section b

Main Study

After the model is trained, the interpretabilty tool will be instantatied and used to explain the predictions of this model.

Instructions:

- Please read carefully and **execute all cells** (if you did "Run All", the first part will already be executed, no need to run those again).
- At the end of each section you will find a some questions, which you can answer in the empty cells provided below them.
- If you have any questions, please let the researcher now.
- Feel free to refer to the tutorial if you need a reminder of any of the concepts introduced there.

Preliminaries: Data, Features, Meta-Features & Models

In [ ]:
```
import sys
import importlib
import numpy as np
import sklearn
import pandas as pd
import matplotlib.pyplot as plt
sys.path.append('../')
```

The target variable is the 'channel', which has 6 classes, not evenly distributed:

In [ ]:
```
from src.data import load_online_news
X, Y, df, feature_groups, feature_desc = load_online_news(target='channel', transform='log')
```

This dataset has 53 features, which could be hard to analyze simultaenously. Fortunately, there's many variables that encode similar aspects of the input, like length or polarity. A faily simple and natural grouping of features is shown below.

Note: there is no need to read the description of all features. Should you need them, you can scroll back here and read those that might be relevant for questions later on.

In [ ]:
```
for i,idxs in enumerate(feature_groups.idxs):
    print('\nFeatue Group {}: {}'.format(i, feature_groups.names[i].upper()))
    for j in idxs:
```

Thank you for participating in our study!

The study is structured as follows:

1. **Tutorial**: Overview of background concepts that will be used in the rest of the study.
2. **Setup Description**: Presentation of the dataset used, descriptive statistics, etc.
3. **Main Component**: Using the interpretability tool to answer questions about an ML model.
4. **Follow-Up**: Questionnaire and interview.

Please keep this tutorial open and handy while you complete the study. **You're welcome (and encouraged!) to refer back to it at any point.**

Concepts covered in this tutorial

Weight of Evidence:

- Definition
- Interpretation
- Binary vs. Multiclass
- Grouping Features
- Sequential Explanation
- Example

Tutorial

Weight Of Evidence



Setup

The study will focus on machine learning classifiers, so we will introduce the Weight of Evidence using a (made up) classification example.

Suppose we are trying to predict the variable Y, where:

User Study on Interpretability - Main Sections

BEFORE YOU BEGIN:

- Make sure this is running on the **Python 3.6 Kernel** (not Python 3). This can be changed in the 'Kernel' menu above.
- Go to "Cell" -> "Run All" to start executing preliminary commands in the background while you read the instructions below.

Description:

Throughout this study we will be using the 'Online News Popularity' dataset.

Each instance in this datas length, contents, polarity and some metadata. We will provide s training/testing).

The task is to predict the c/ news article was published.

The 'Preliminaries' section b

Main Study

After the model is trained, the interpretabilty tool will be instantatied and used to explain the predictions of this model.

Instructions:

- Please read carefully and **execute all cells** (if you did "Run All", the first part will already be executed, no need to run those again).
- At the end of each section you will find a some questions, which you can answer in the empty cells provided below them.
- If you have any questions, please let the researcher now.
- Feel free to refer to the tutorial if you need a reminder of any of the concepts introduced there.

Preliminaries: Data, Features, Meta-Features & Models

In [ ]:
```
import sys
import importlib
import numpy as np
import sklearn
import pandas as pd
import matplotlib.pyplot as plt
sys.path.append('../')
```

The target variable is the 'channel', which has 6 classes, not evenly distributed:

In [ ]:
```
from src.data import load_online_news
X, Y, df, feature_groups, feature_desc = load_online_news(target='channel', transform='log')
```

This dataset has 53 features, which could be hard to analyze simultaenously. Fortunately, there's many variables that encode similar aspects of the input, like length or polarity. A faily simple and natural grouping of features is shown below.

Note: there is no need to read the description of all features. Should you need them, you can scroll back here and read those that might be relevant for questions later on.

In [ ]:
```
for i,idxs in enumerate(feature_groups.idxs):
    print('\nFeatue Group {}: {}'.format(i, feature_groups.names[i].upper()))
    for j in idxs:
```

WoE Interpretability
user study - main findings

- Participants successfully used the tool, despite some struggling with background concepts
- Sequential explanations useful for '**breaking down**' prediction - led to more detailed answers
  "[they were like] a *story of how the prediction was made*" (P3)
- Users want **more interactivity** - consistent with cognitive science [Lombrozo, 2012] and HCI [Amershi et al., 2019; Hohmanet al., 2019b; ...]
- **Tutorial was crucial**, time spent on it correlated with quality of answers
  "I can't imagine doing this [study] without the tutorial." (P5)

WoE Interpretability
Key Takeaways + Future Work

- **Ethos**: ML interpretability should put humans at the front and center!
- **Method**: 5 design principles from cog/soc sciences, operationalize via Weight-of-Evidence
- **Findings**: modularity, interactivity + tutorials matter!
- **Improvements**: more participants, contextualize study to user's familiar settings
- **Code + Notebooks**: GitHub.com/dmelis/interpretwoe