

Guidelines for Human-AI Interaction

A decorative graphic consisting of seven horizontal bars of different colors (light blue, dark blue, purple, red, orange, yellow, and green) that transition into arrows pointing to the right. The bars are stacked vertically, with the top bar being the widest and the bottom bar being the narrowest.

Saleema Amershi, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Derek DeBellis, Ruth Kikin-Gil, Shamsi Iqbal, Paul Bennett, Dan Weld, Jina Suh, Kori Inkpen, Jaime Teevan, and Eric Horvitz

<https://aka.ms/aiguideelines>

INITIALLY

1
Make clear
what the
system
can do.

2
Make clear
how well the
system can
do what it
can do.

Guidelines for Human AI Interaction

Learn more: <https://aka.ms/aiguideelines>



DURING INTERACTION

3
Time services
based on
context.

4
Show
contextually
relevant
information.

5
Match
relevant
social norms.

6
Mitigate
social biases.

WHEN WRONG

7
Support
efficient
invocation.

8
Support
efficient
dismissal.

9
Support
efficient
correction.

10
Scope
services when
in doubt.

11
Make clear
why the
system did
what it did.

OVER TIME

12
Remember
recent
interactions.

13
Learn from
user behavior.

14
Update and
adapt
cautiously.

15
Encourage
granular
feedback.

16
Convey the
consequences
of user
actions.

17
Provide
global
controls.

18
Notify users
about
changes.

Agenda

Intro to the guidelines

Findings and impact

Engineering and AI implications

Challenges for Intelligible AI

Agenda

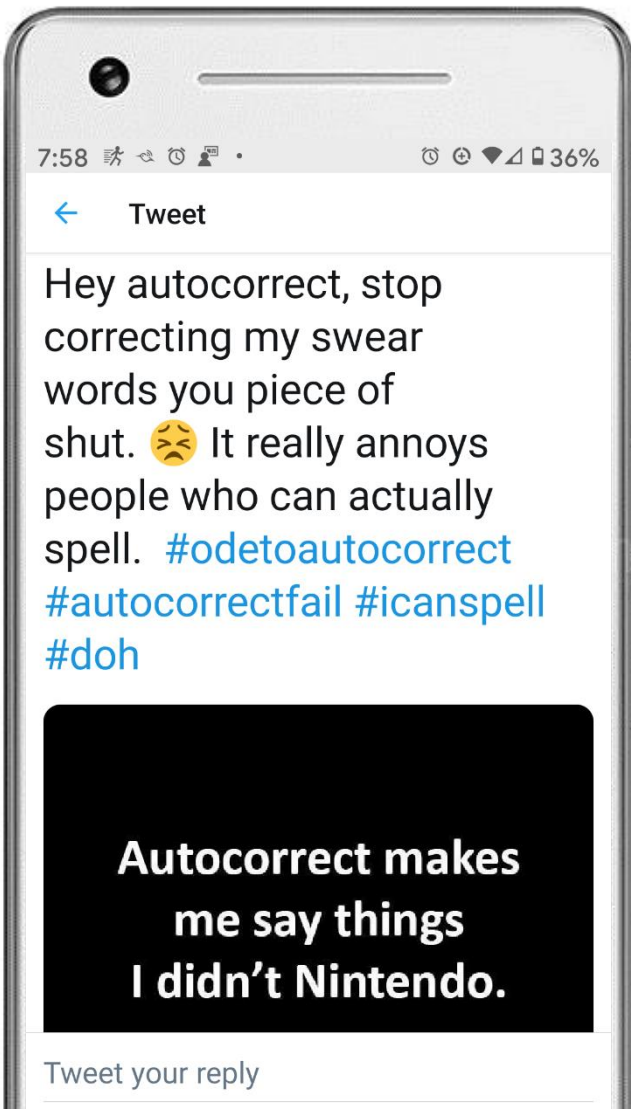
Intro to the guidelines

Findings and impact

Engineering and AI implications

Challenges for Intelligible AI

Creating good AI user experiences is hard



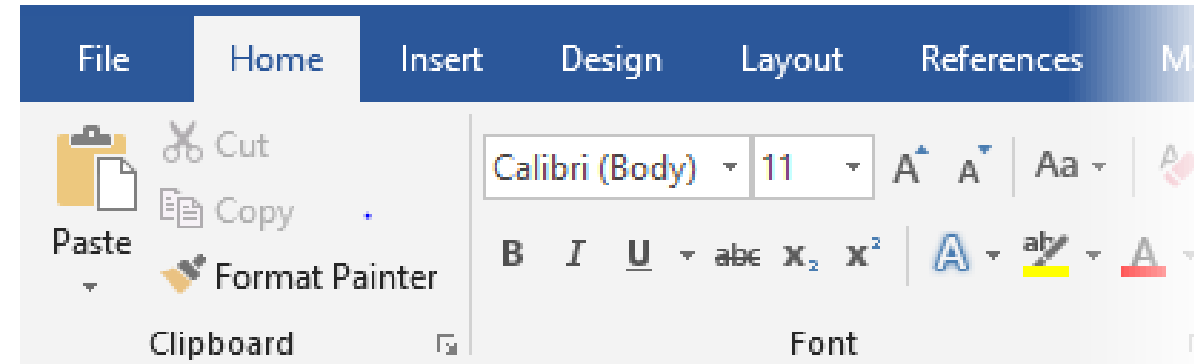


AI is fundamentally changing how we interact with computing systems

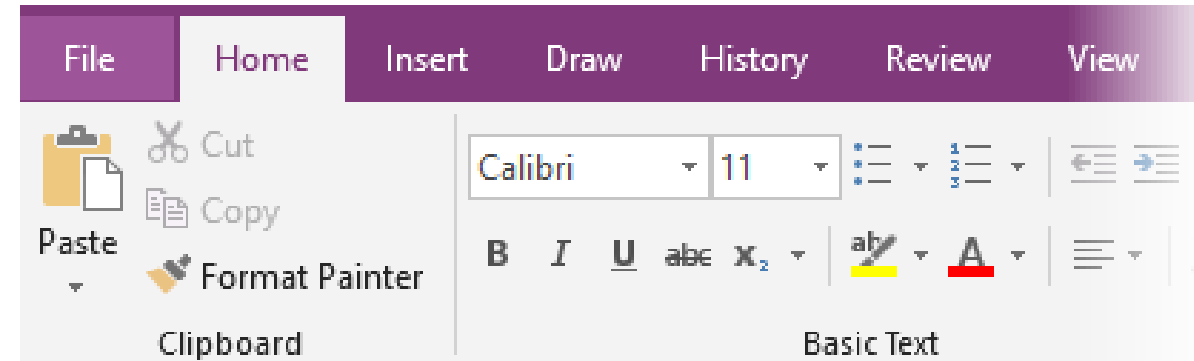
The Consistency Principle

Consistent interfaces and predictable behaviors saves people time and reduces errors.

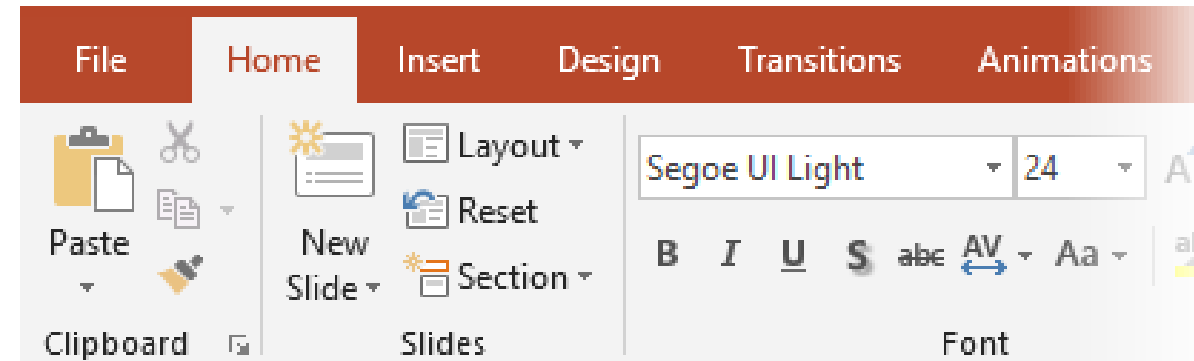
MS Word



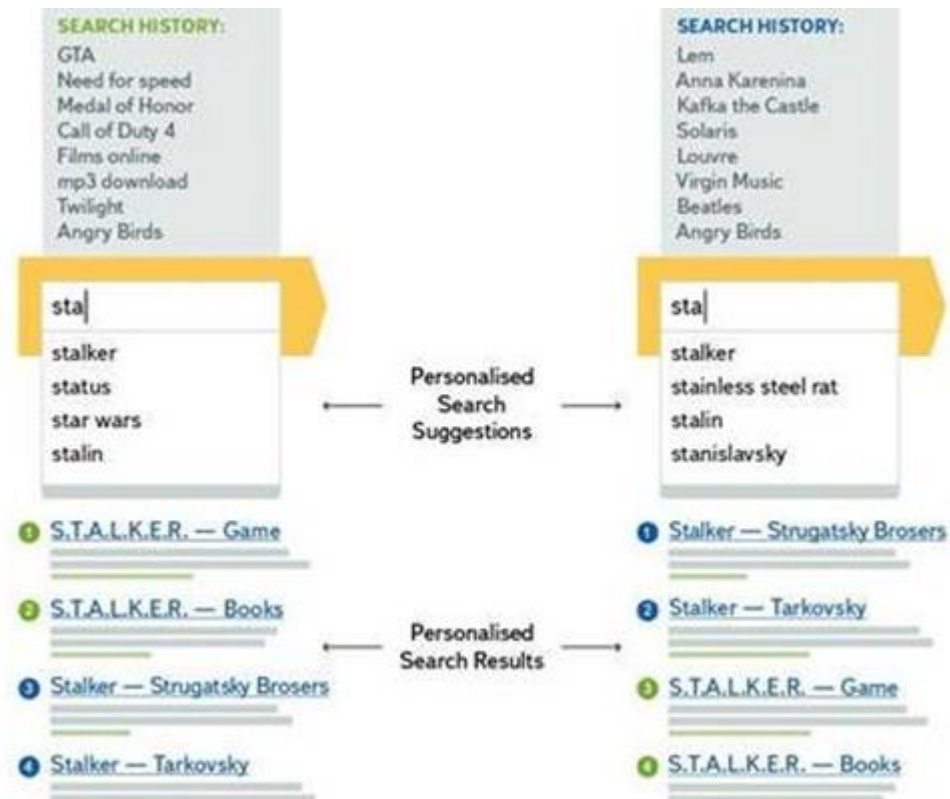
MS OneNote



MS PowerPoint



AI systems are probabilistic and can change over time



Behaviors may change over time



Behaviors may differ in subtly different contexts

Disclaimers

The guidelines are not a checklist

Additional guidelines may be needed in some scenarios

You are using them “the right way” if you consider them during development

INITIALLY

1
Make clear
what the
system
can do.

2
Make clear
how well the
system can
do what it
can do.

Guidelines for Human AI Interaction

Learn more: <https://aka.ms/aiguideelines>



DURING INTERACTION

3
Time services
based on
context.

4
Show
contextually
relevant
information.

5
Match
relevant
social norms.

6
Mitigate
social biases.

WHEN WRONG

7
Support
efficient
invocation.

8
Support
efficient
dismissal.

9
Support
efficient
correction.

10
Scope
services when
in doubt.

11
Make clear
why the
system did
what it did.

OVER TIME

12
Remember
recent
interactions.

13
Learn from
user behavior.

14
Update and
adapt
cautiously.

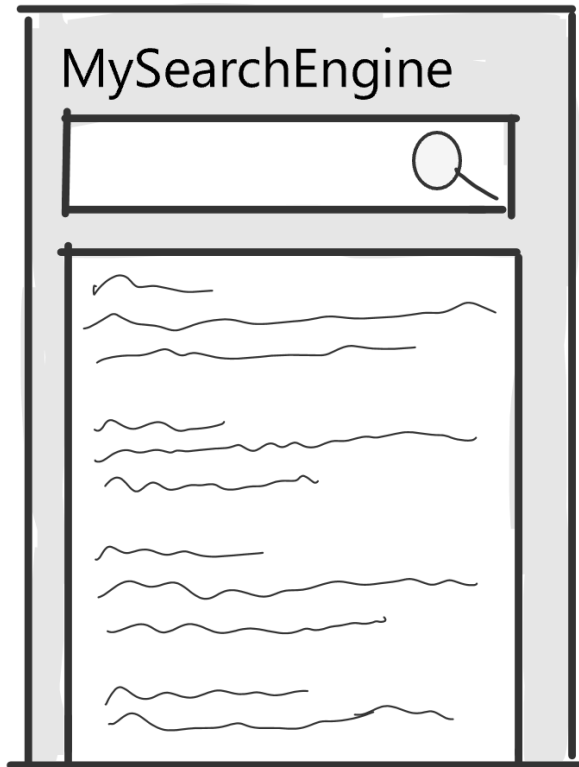
15
Encourage
granular
feedback.

16
Convey the
consequences
of user
actions.

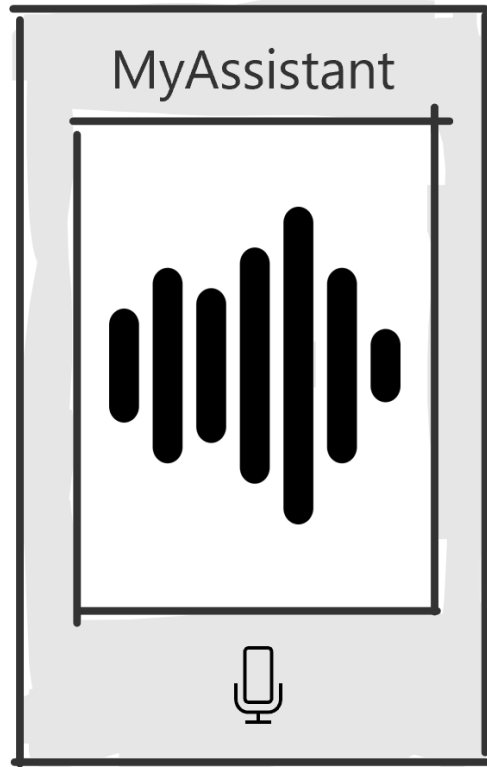
17
Provide
global
controls.

18
Notify users
about
changes.

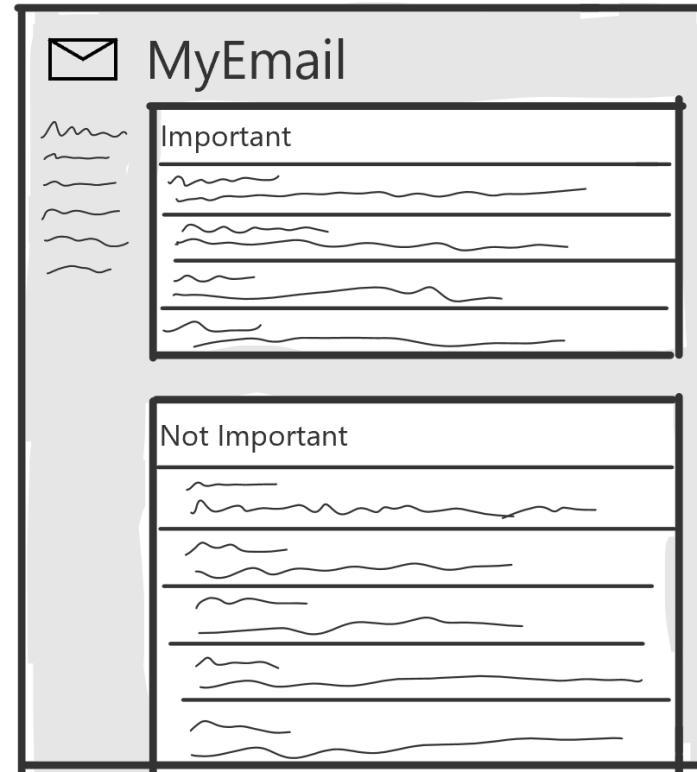
Examples from common AI-based products



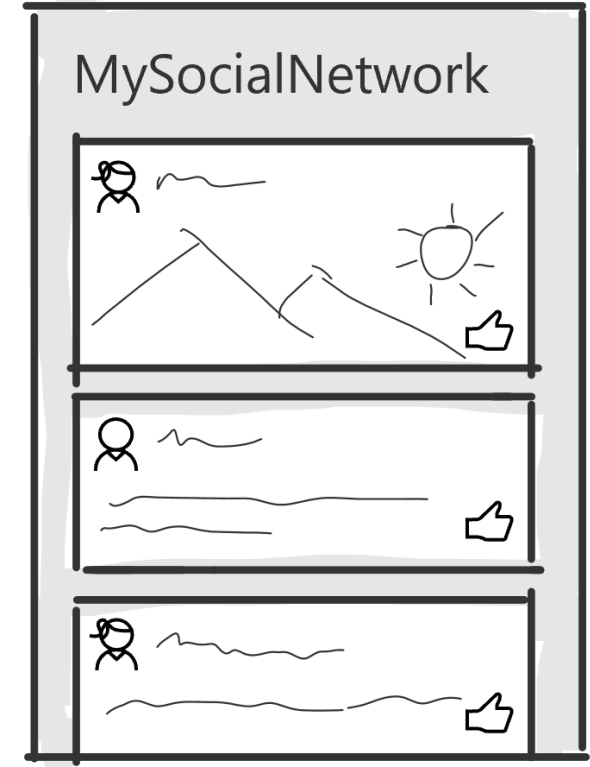
AI used for query processing, ranking results, filtering spam...



AI used for speech processing, task support....



AI used for email sorting, entity detection, response generation...



AI used for filtering feed, recommending ads...

INITIALLY

1
Make clear
what the
system
can do.

2
Make clear
how well the
system can
do what it
can do.

Guidelines for Human AI Interaction

Learn more: <https://aka.ms/aiguideelines>



DURING INTERACTION

3
Time services
based on
context.

4
Show
contextually
relevant
information.

5
Match
relevant
social norms.

6
Mitigate
social biases.

WHEN WRONG

7
Support
efficient
invocation.

8
Support
efficient
dismissal.

9
Support
efficient
correction.

10
Scope
services when
in doubt.

11
Make clear
why the
system did
what it did.

OVER TIME

12
Remember
recent
interactions.

13
Learn from
user behavior.

14
Update and
adapt
cautiously.

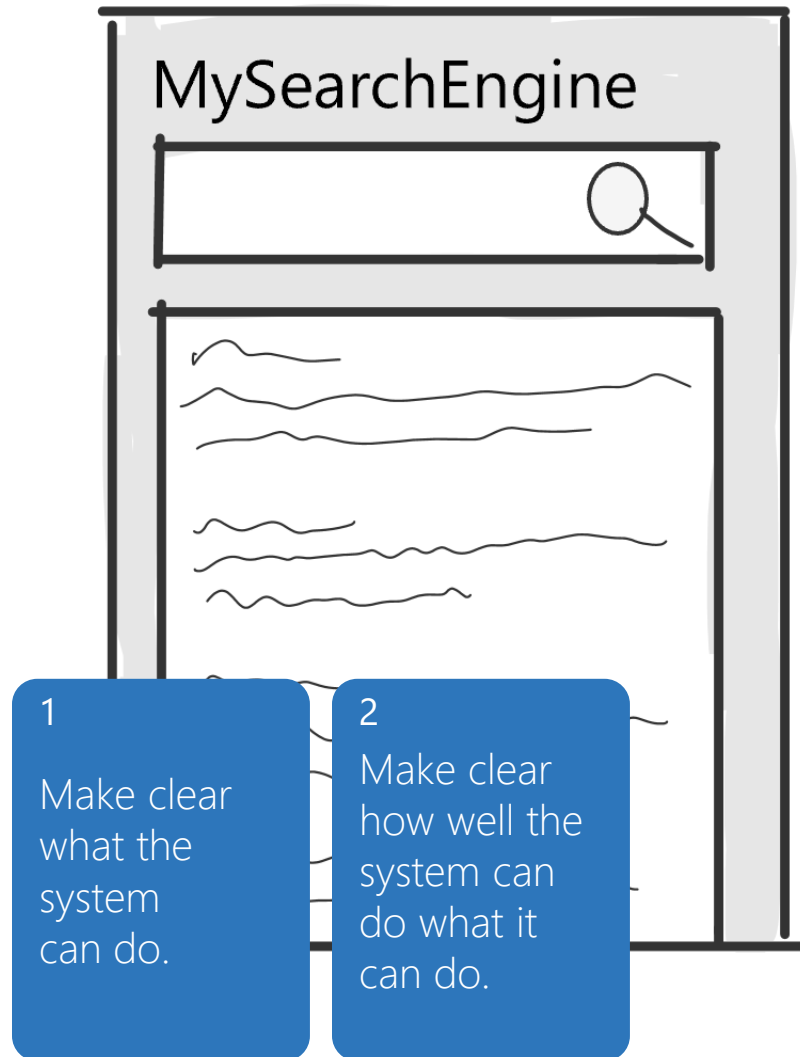
15
Encourage
granular
feedback.

16
Convey the
consequences
of user
actions.

17
Provide
global
controls.

18
Notify users
about
changes.

Set the right expectations



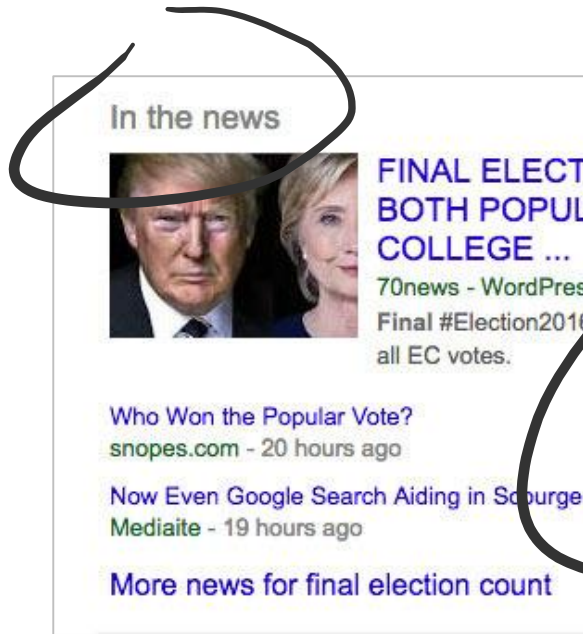
Coverage: Many people think "everything is on the web"*

Quality: 33% of people use the term "magic" when explaining how search works*

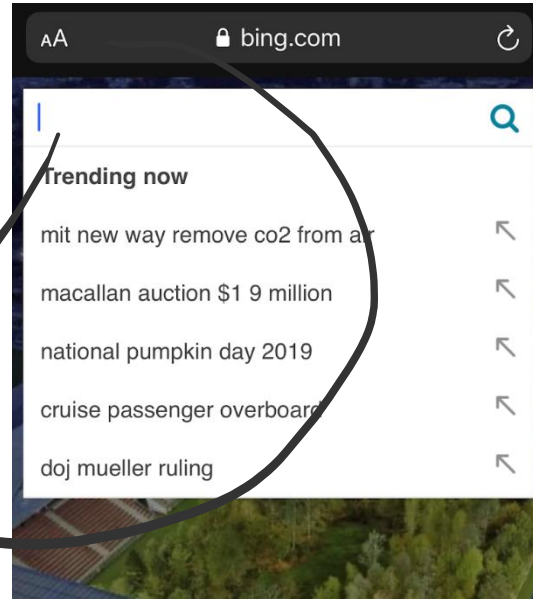
Can be problematic when people overestimate search capabilities for high-stakes tasks

*Dan Russell. The Joy of Search.

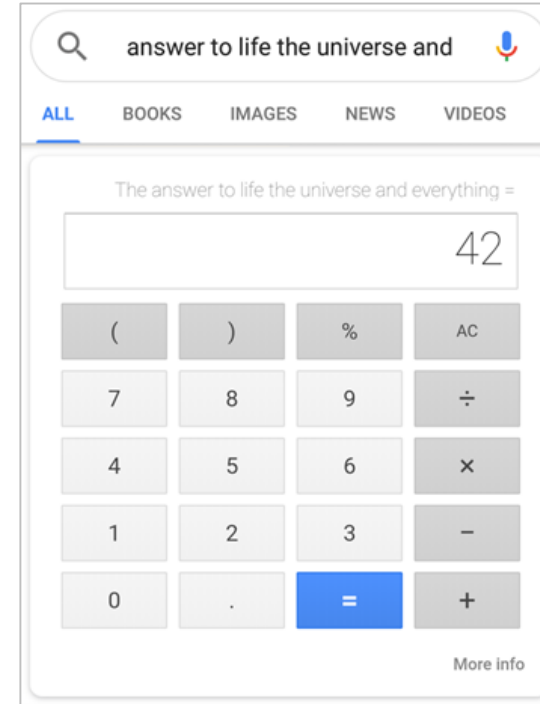
Set the right expectations – What can you do?



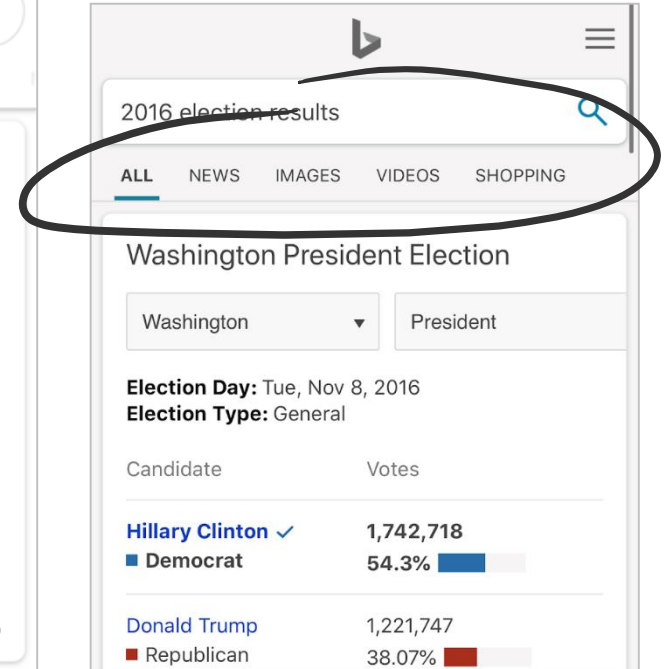
Provide documentation (use sparingly)



Show examples



Introduce features at appropriate times



Give people controls

INITIALLY

1
Make clear
what the
system
can do.

2
Make clear
how well the
system can
do what it
can do.

Guidelines for Human AI Interaction

Learn more: <https://aka.ms/aiguideelines>



DURING INTERACTION

3
Time services
based on
context.

4
Show
contextually
relevant
information.

5
Match
relevant
social norms.

6
Mitigate
social biases.

WHEN WRONG

7
Support
efficient
invocation.

8
Support
efficient
dismissal.

9
Support
efficient
correction.

10
Scope
services when
in doubt.

11
Make clear
why the
system did
what it did.

OVER TIME

12
Remember
recent
interactions.

13
Learn from
user behavior.

14
Update and
adapt
cautiously.

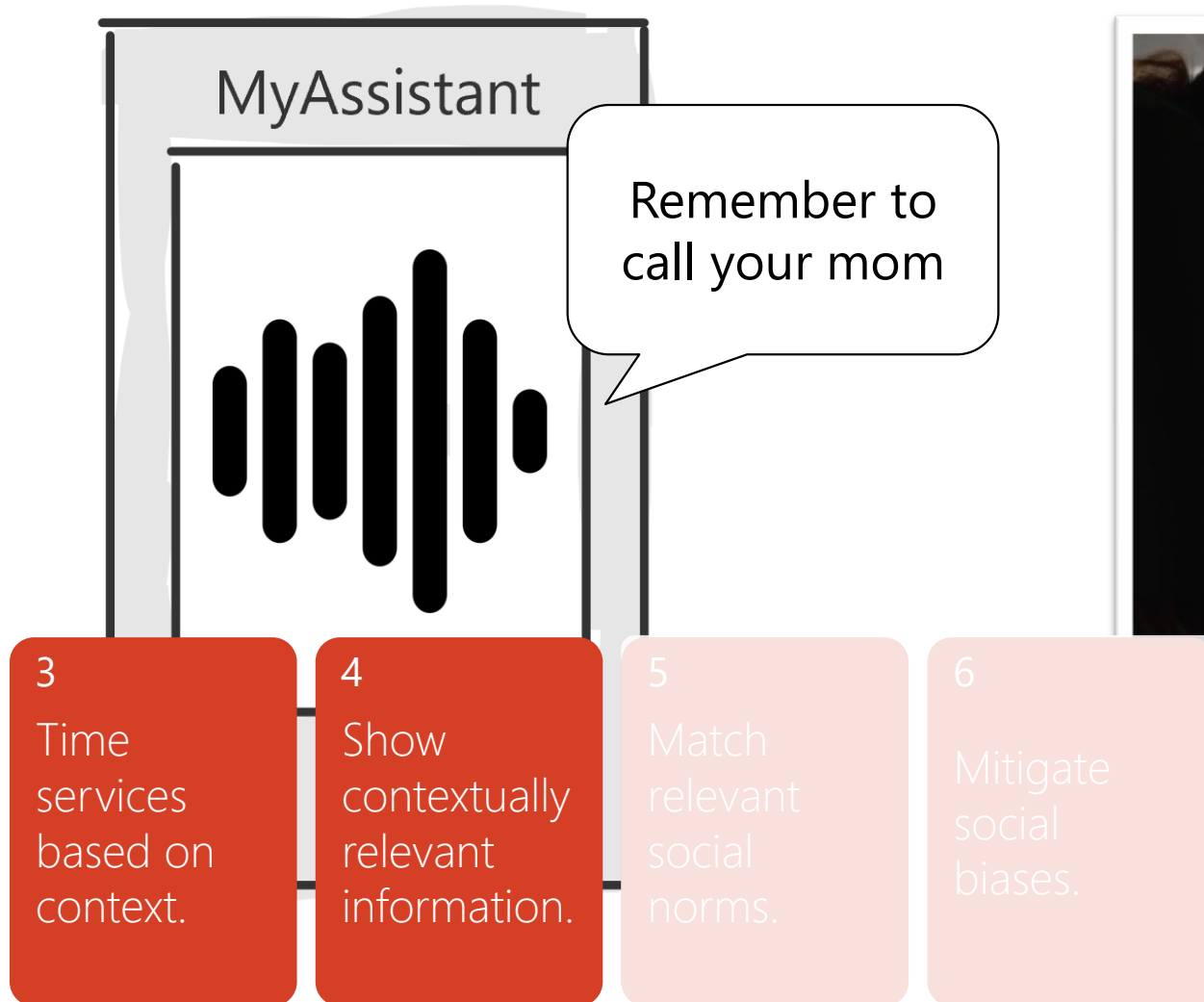
15
Encourage
granular
feedback.

16
Convey the
consequences
of user
actions.

17
Provide
global
controls.

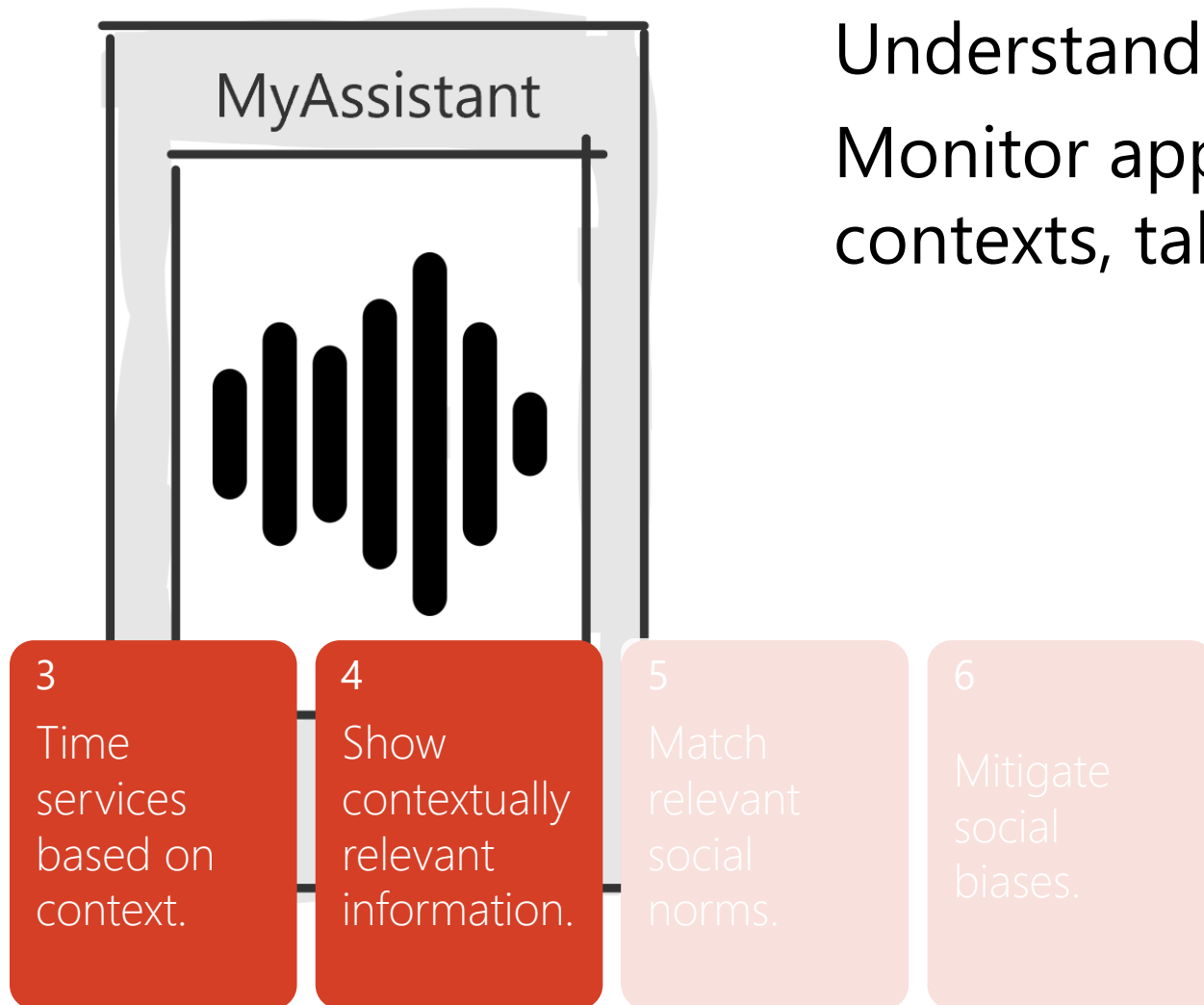
18
Notify users
about
changes.

Contextual Mismatches



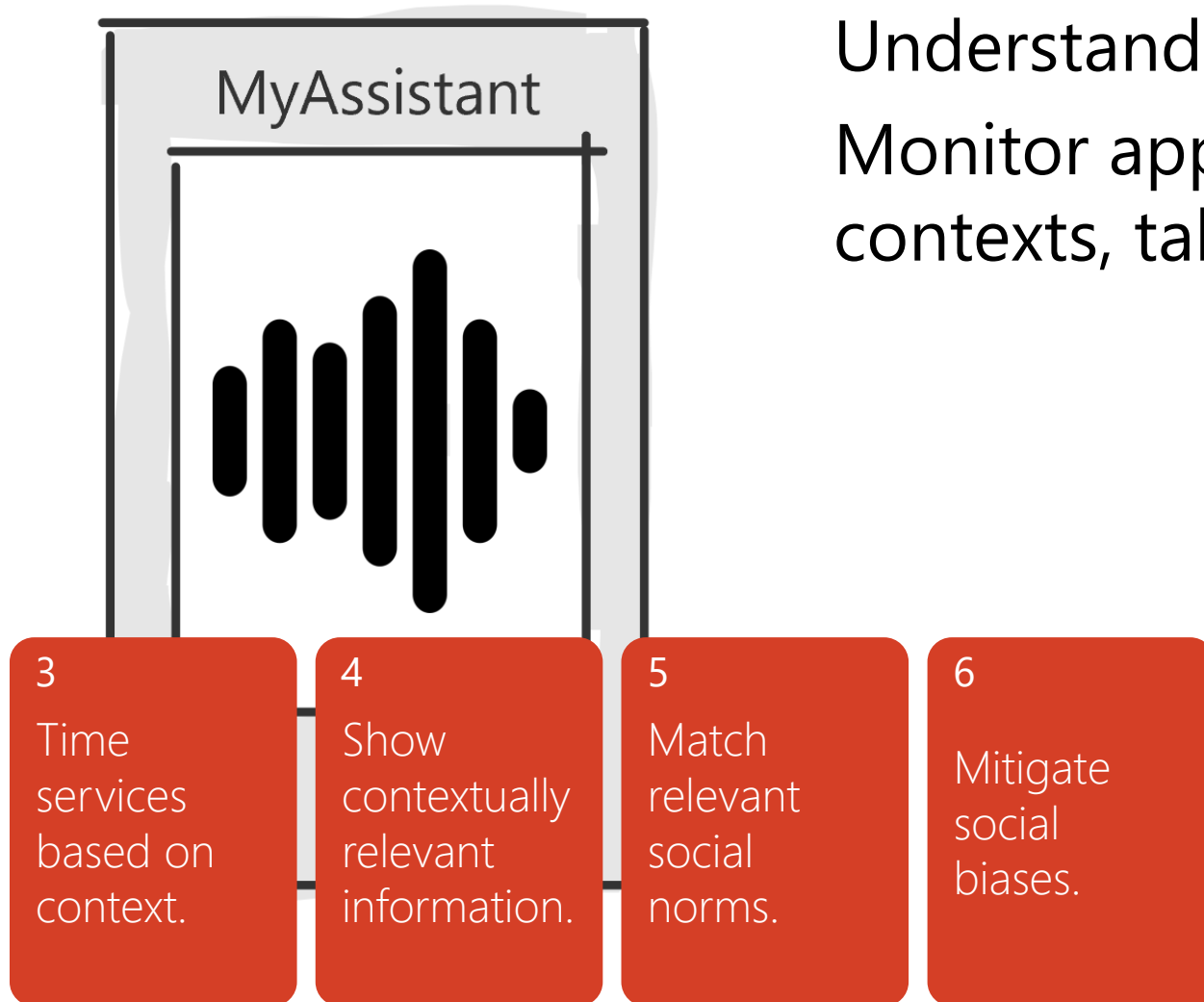
Contextual Mismatches – What can you do?

Understand and infer critical contexts
Monitor appropriate signals, model critical contexts, take appropriate actions

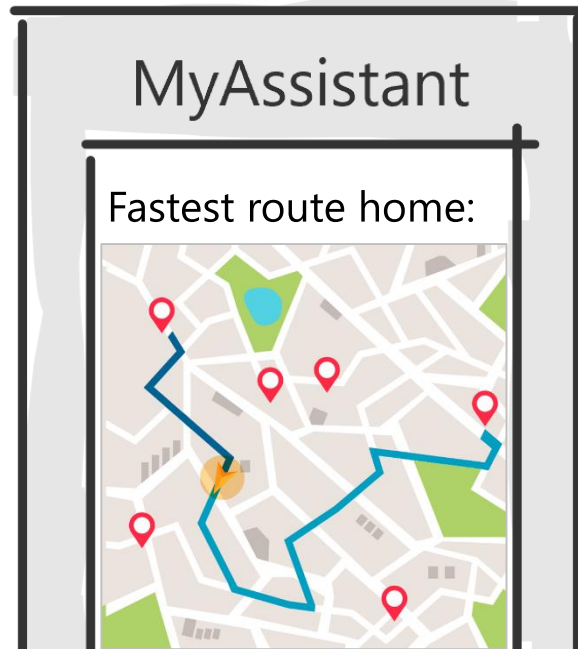


Contextual Mismatches – What can you do?

Understand and infer critical contexts
Monitor appropriate signals, model critical contexts, take appropriate actions



Contextual Mismatches – What can you do?



Understand and infer critical contexts
Monitor appropriate signals, model critical contexts, take appropriate actions
Develop and test with diversity in mind

3

Time services based on context.

4

Show contextually relevant information.

5

Match relevant social norms.

6

Mitigate social biases.

"Information is not subject to biases, unless users are biased against fastest routes"

"There's no way to set an avg walking speed. [The product] assumes users to be healthy"

INITIALLY

1
Make clear
what the
system
can do.

2
Make clear
how well the
system can
do what it
can do.

Guidelines for Human AI Interaction

Learn more: <https://aka.ms/aiguideelines>



DURING INTERACTION

3
Time services
based on
context.

4
Show
contextually
relevant
information.

5
Match
relevant
social norms.

6
Mitigate
social biases.

WHEN WRONG

7
Support
efficient
invocation.

8
Support
efficient
dismissal.

9
Support
efficient
correction.

10
Scope
services when
in doubt.

11
Make clear
why the
system did
what it did.

OVER TIME

12
Remember
recent
interactions.

13
Learn from
user behavior.

14
Update and
adapt
cautiously.

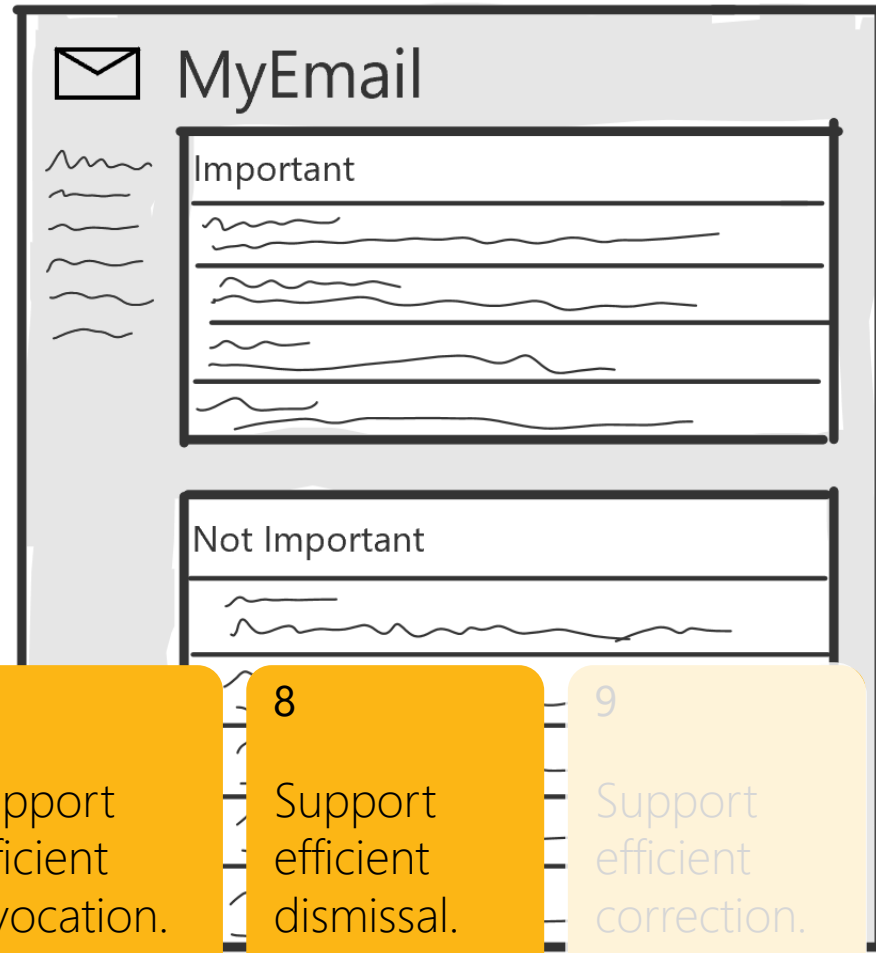
15
Encourage
granular
feedback.

16
Convey the
consequences
of user
actions.

17
Provide
global
controls.

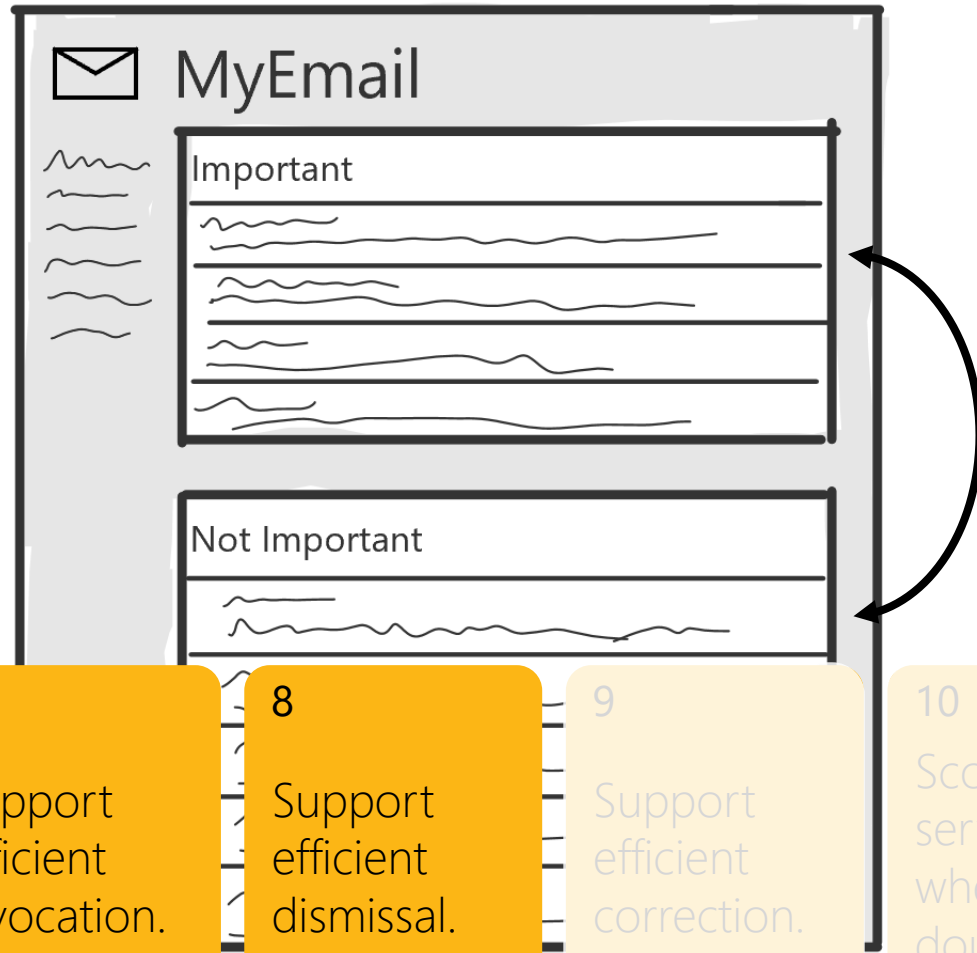
18
Notify users
about
changes.

Model Errors



Common errors: false positives, false negatives, partially correct, uncertain...

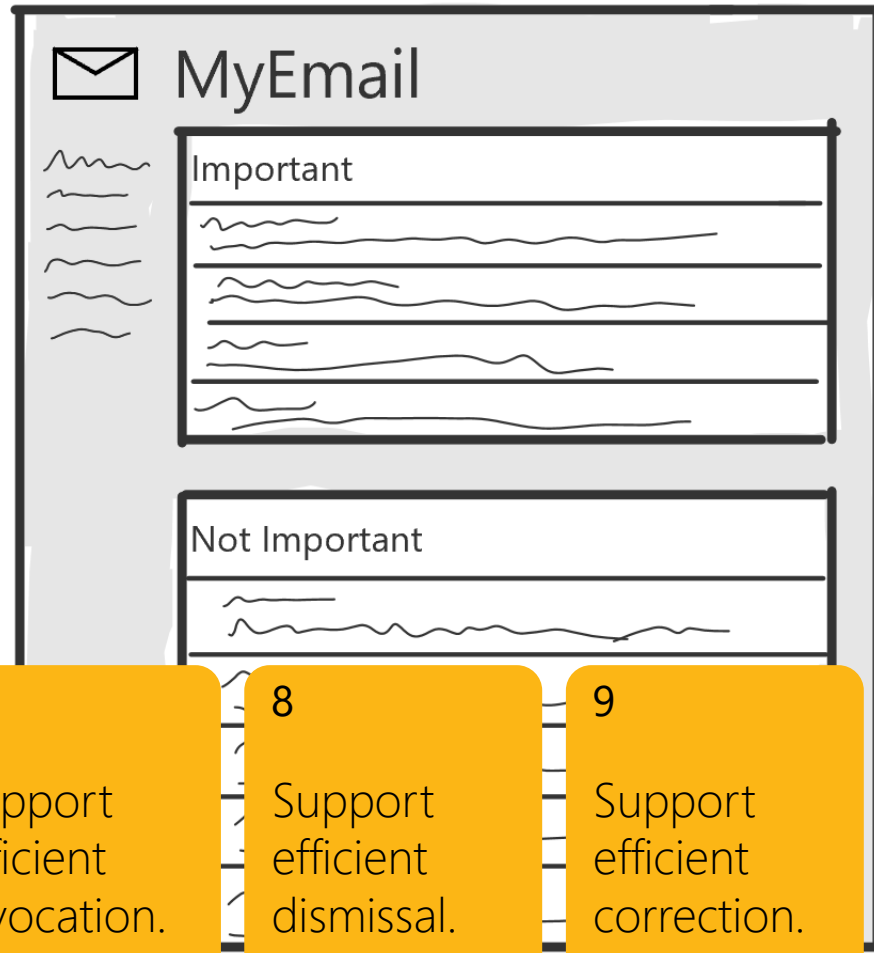
Model Errors – What can you do?



Common errors: false positives, false negatives, partially correct, uncertain...

Consider the costs of errors and provide appropriate mitigation strategies

Model Errors – What can you do?



Common errors: false positives, false negatives, partially correct, uncertain...

Consider the costs of errors and provide appropriate mitigation strategies (or explanations)

INITIALLY

1
Make clear
what the
system
can do.

2
Make clear
how well the
system can
do what it
can do.

Guidelines for Human AI Interaction

Learn more: <https://aka.ms/aiguideelines>



DURING INTERACTION

3
Time services
based on
context.

4
Show
contextually
relevant
information.

5
Match
relevant
social norms.

6
Mitigate
social biases.

WHEN WRONG

7
Support
efficient
invocation.

8
Support
efficient
dismissal.

9
Support
efficient
correction.

10
Scope
services when
in doubt.

11
Make clear
why the
system did
what it did.

OVER TIME

12
Remember
recent
interactions.

13
Learn from
user behavior.

14
Update and
adapt
cautiously.

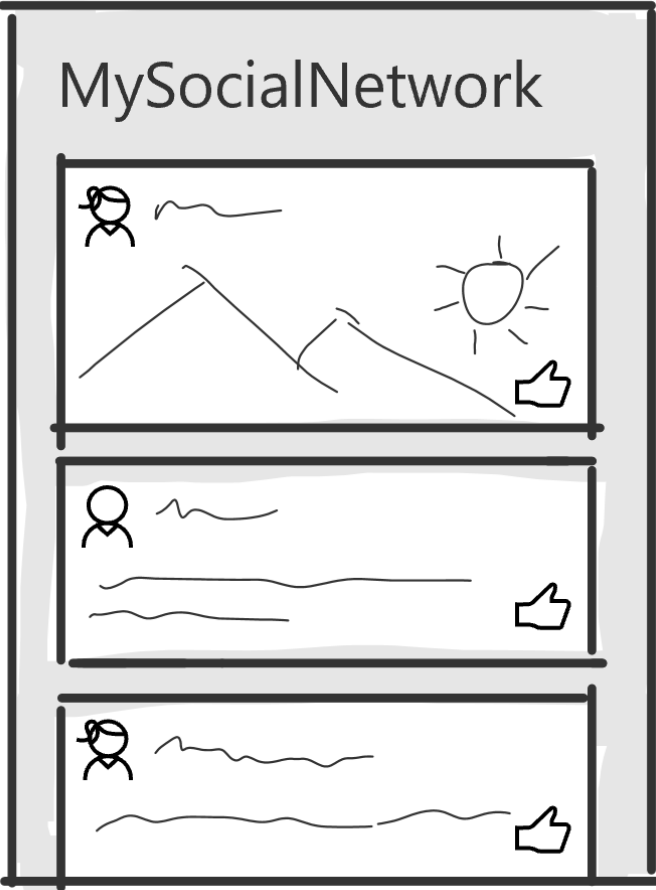
15
Encourage
granular
feedback.

16
Convey the
consequences
of user
actions.

17
Provide
global
controls.

18
Notify users
about
changes.

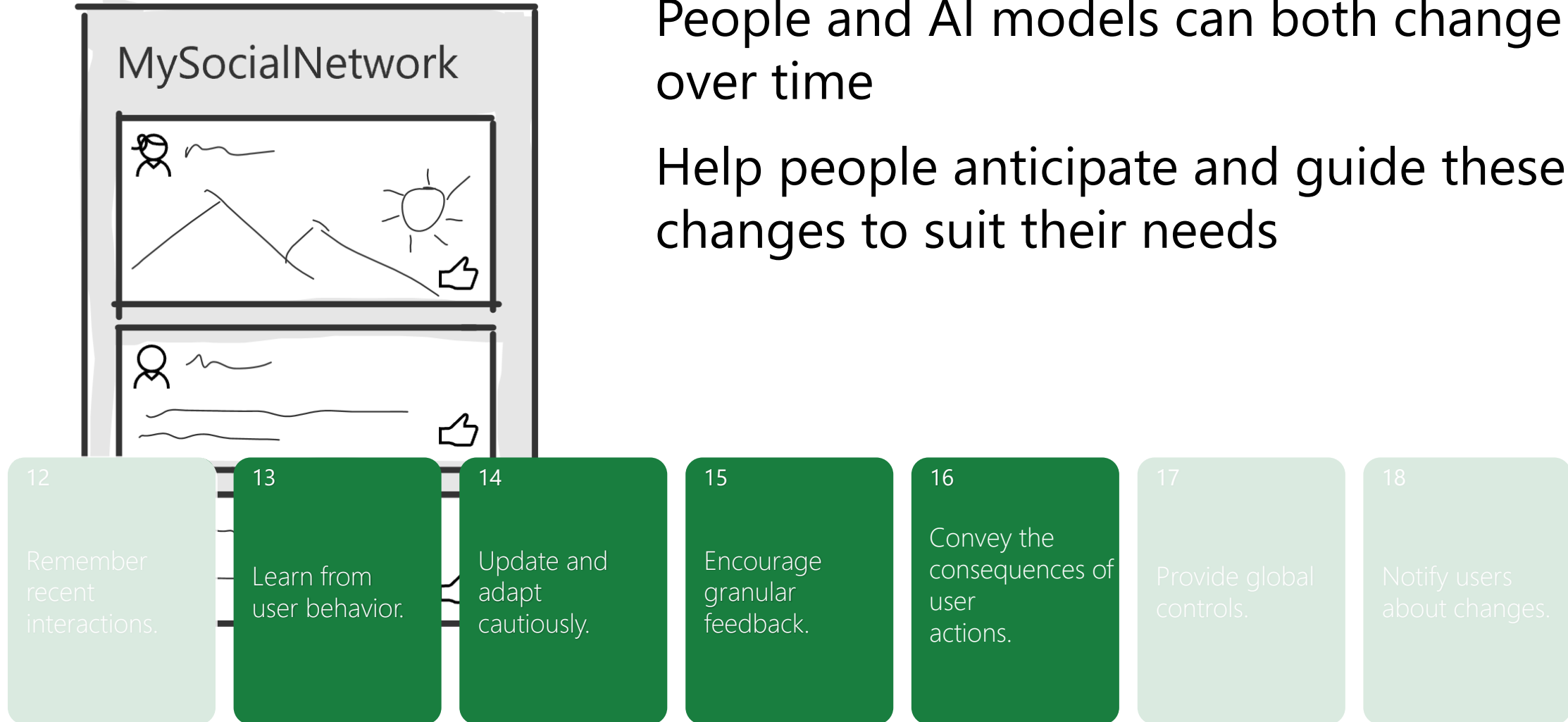
Consider changes over time



Consider changes over time – What can you do?

People and AI models can both change over time

Help people anticipate and guide these changes to suit their needs



Agenda

Intro to the guidelines

Findings and impact

Engineering and AI implications

Challenges for Intelligible AI

Agenda

Intro to the guidelines

Findings and impact

Engineering and AI implications

Challenges for Intelligible AI

Findings & Impact

Initial Impact

Opportunity Analysis

Engagements with Practitioners

Findings & Impact

Initial Impact

Opportunity Analysis

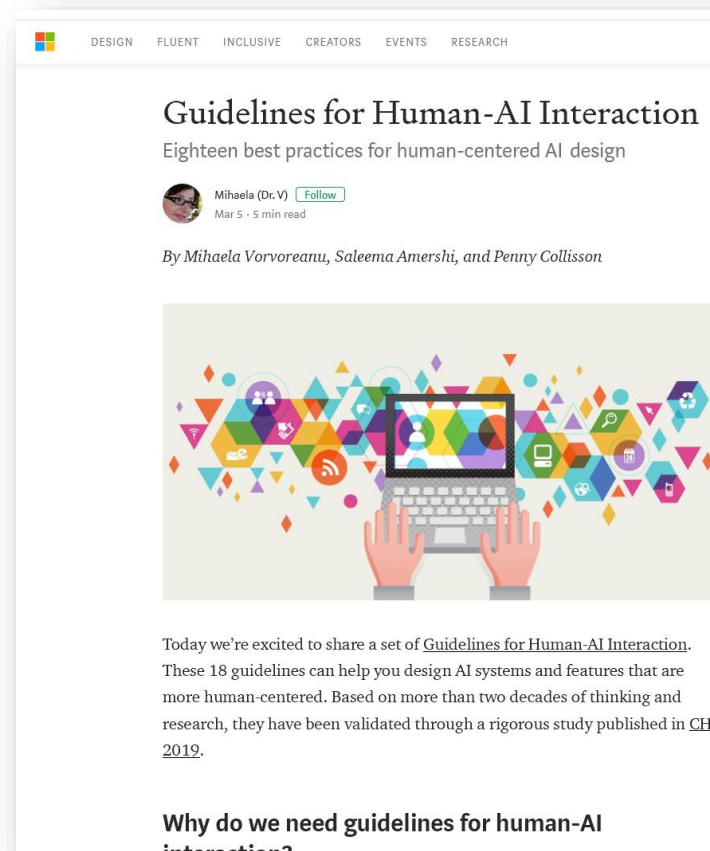
Engagements with Practitioners

Academia



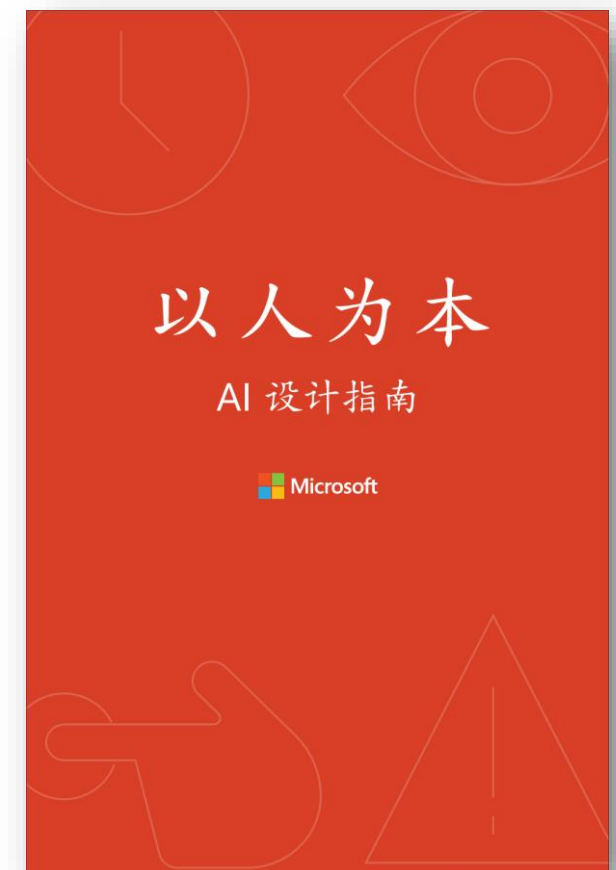
CHI 2019 Best Paper Honorable Mention

Practitioners



Being leveraged by product teams across the company throughout the design and development process

Industry



Cited and used in related organizations
Translated to other languages

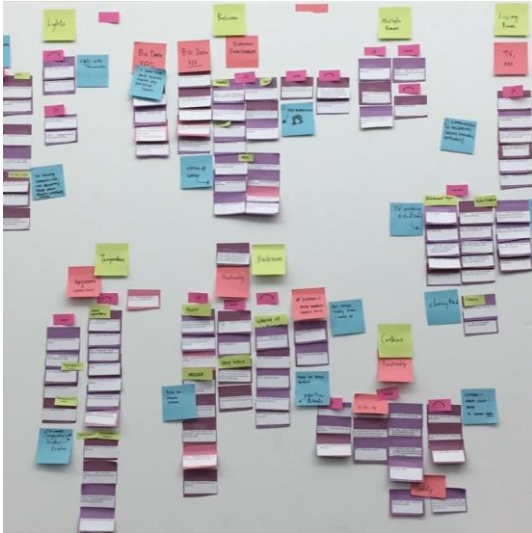
Findings & Impact

Initial Impact

Opportunity Analysis

Engagements with Practitioners

Developing the Guidelines for Human-AI Interaction



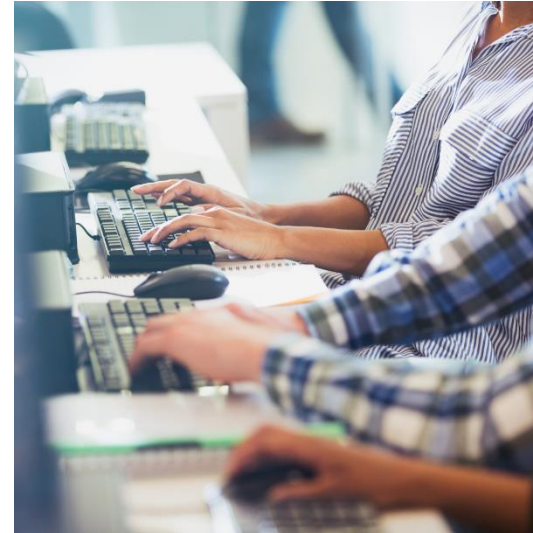
Phase 1. Consolidation

150+ recommendations



Phase 2. Team Evaluation

13 common AI products



Phase 3. User Evaluation

49 UX practitioners,
20 AI products



Phase 4. Expert Review

11 UX practitioners

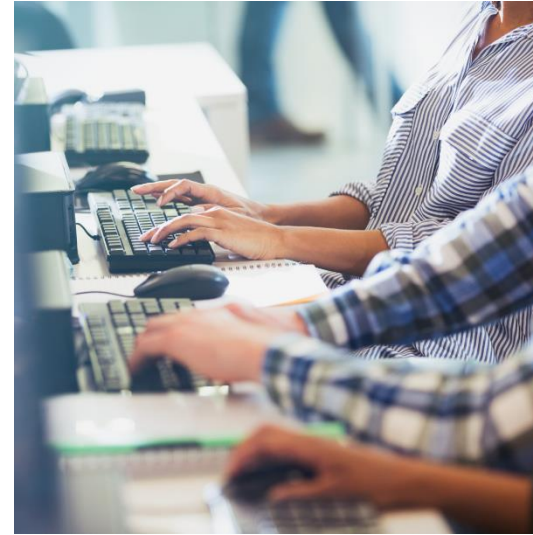
Developing the Guidelines for Human-AI Interaction



Phase 1.
Consolidation
150+ recommendations



Phase 2.
Team Evaluation
13 common AI products

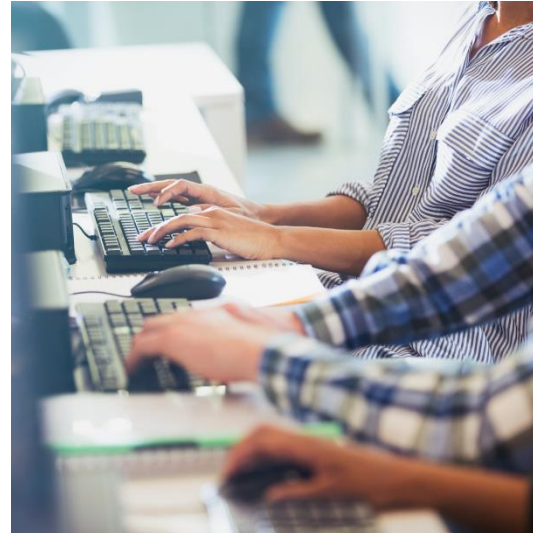


Phase 3.
User Evaluation
49 UX practitioners,
20 AI products



Phase 4.
Expert Review
11 UX practitioners

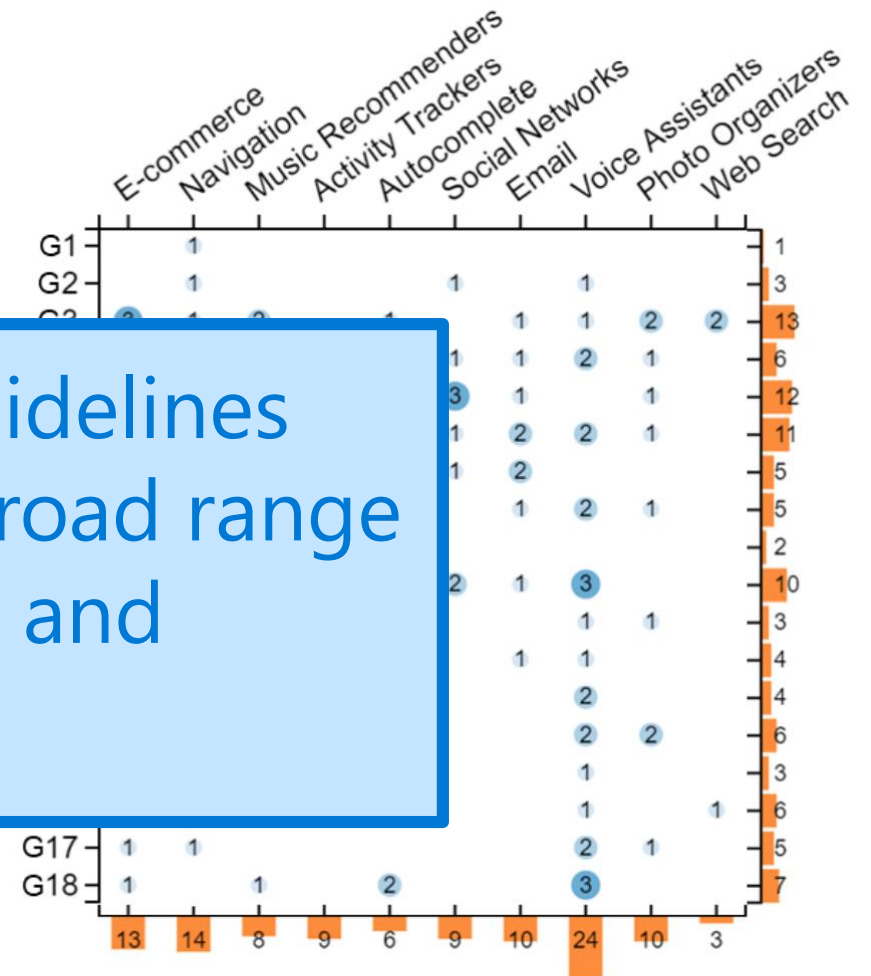
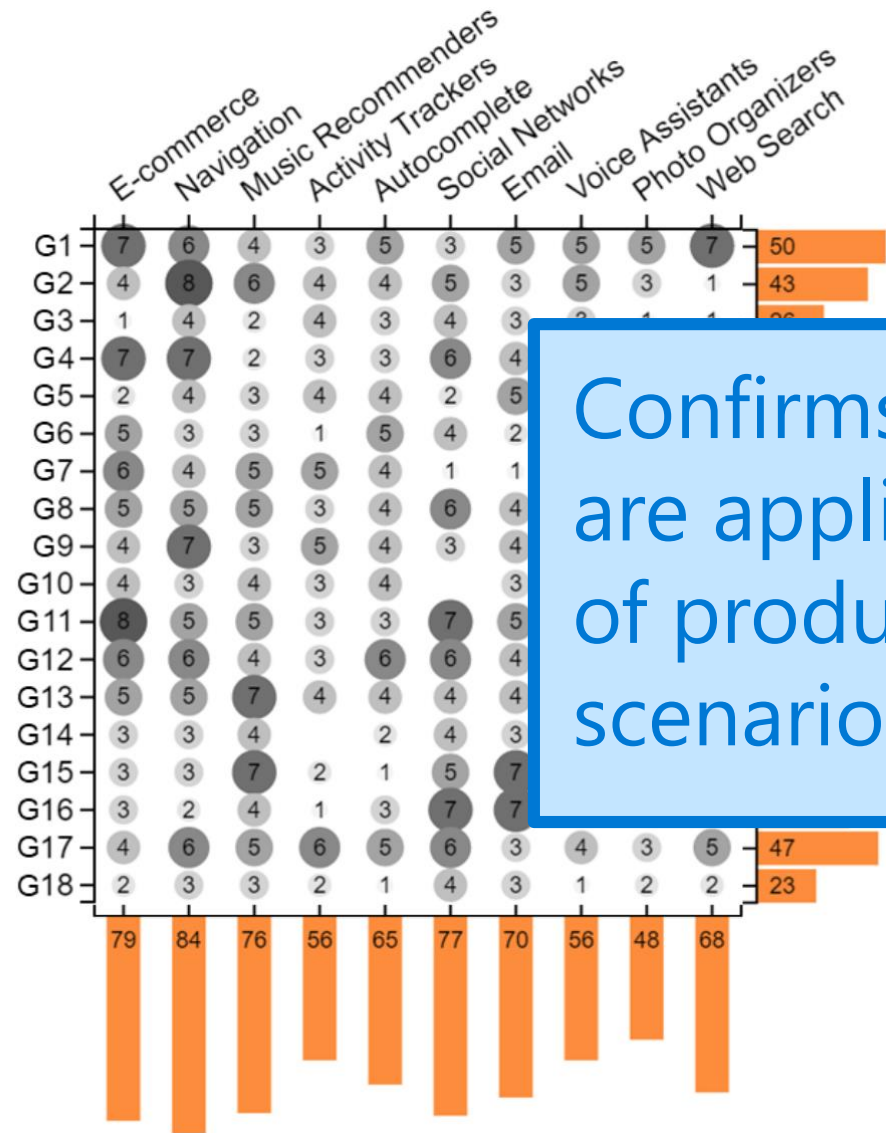
- Collected of **700+** examples of the guidelines being applied or violated
- **20** different products (both Microsoft and 3rd-party)
- **10** product categories (from fitness trackers to music recommenders)



Phase 3.
User Evaluation
49 UX practitioners,
20 AI products

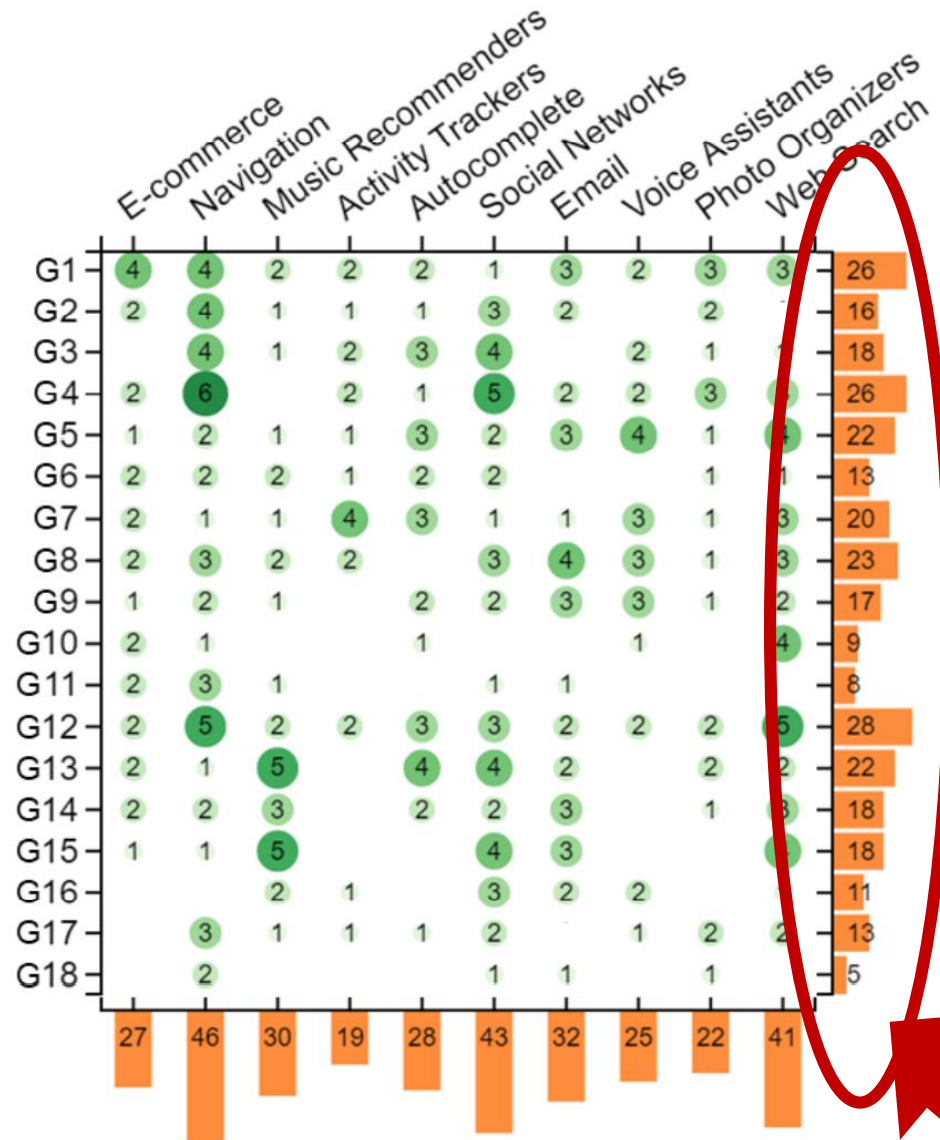


Phase 4.
Expert Review
11 UX practitioners

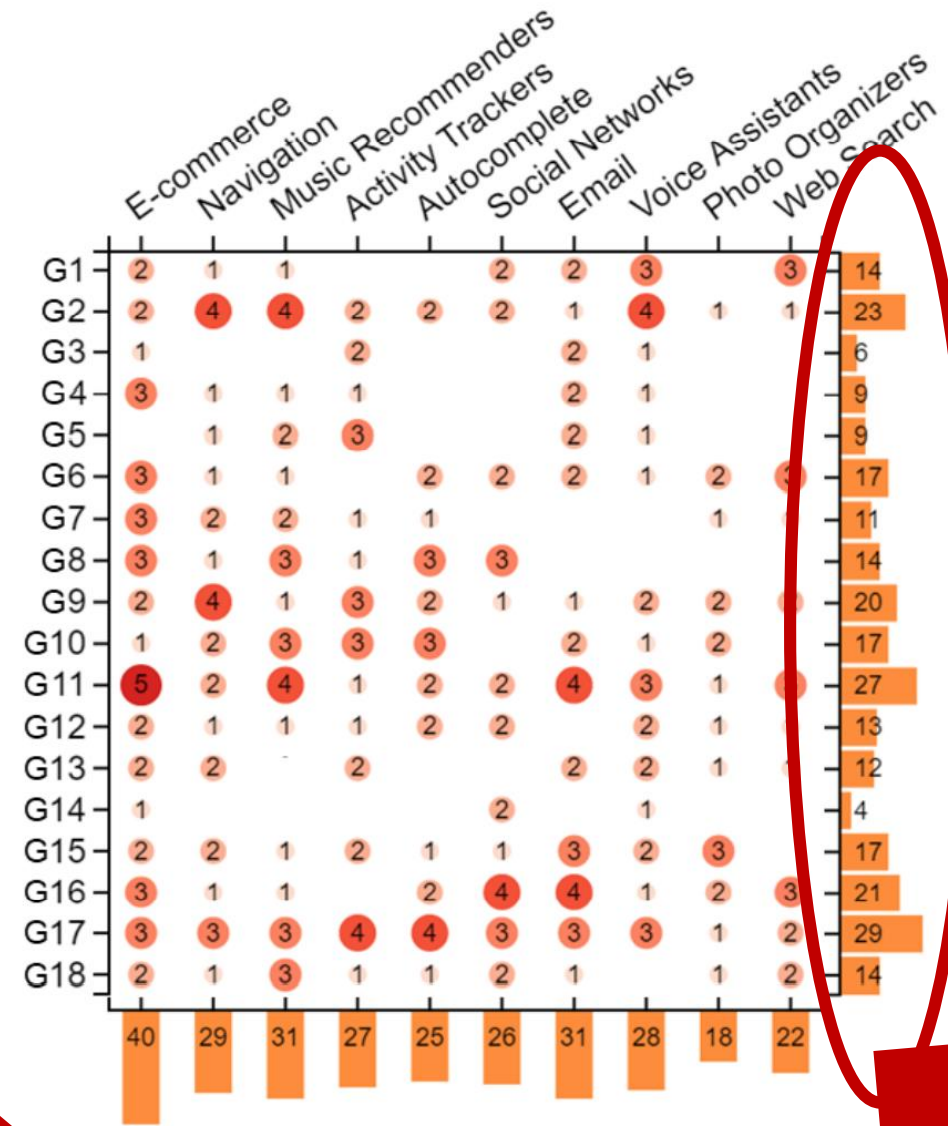


Guideline Applies to Scenario

Guideline Does Not Apply



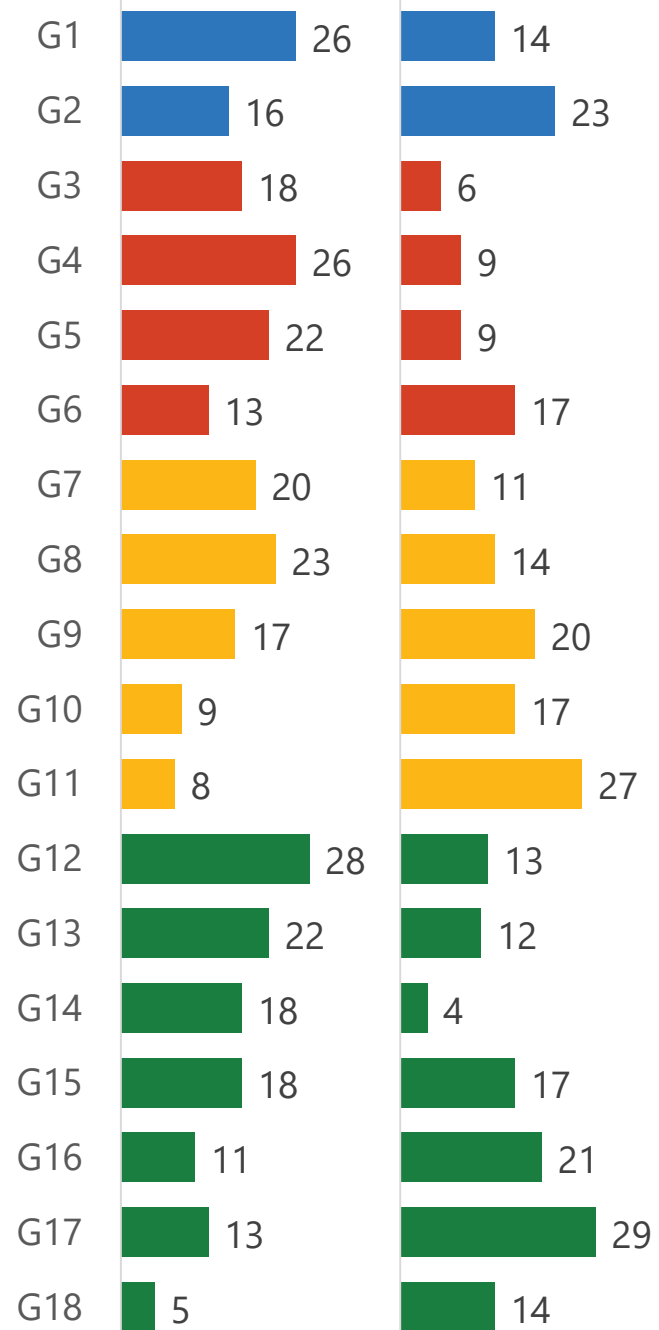
Product/Feature
Applies Guideline

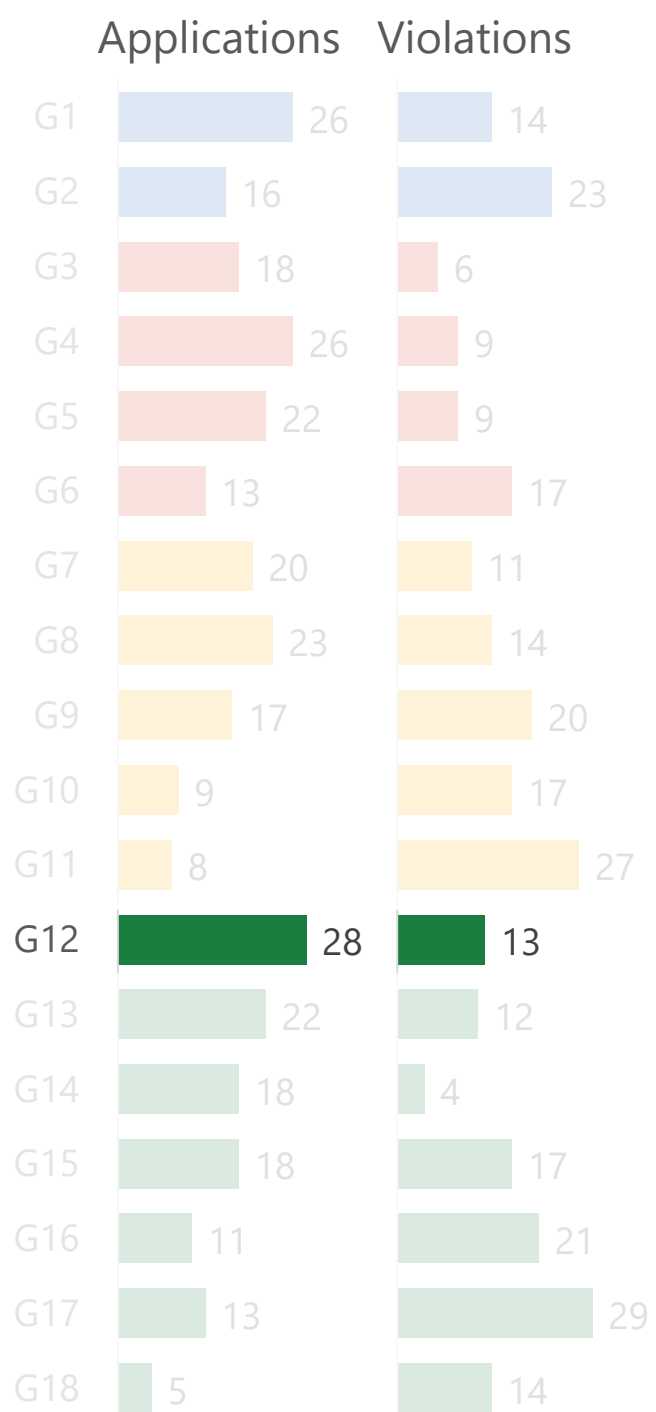


Product/Feature
Violates Guideline

Applications

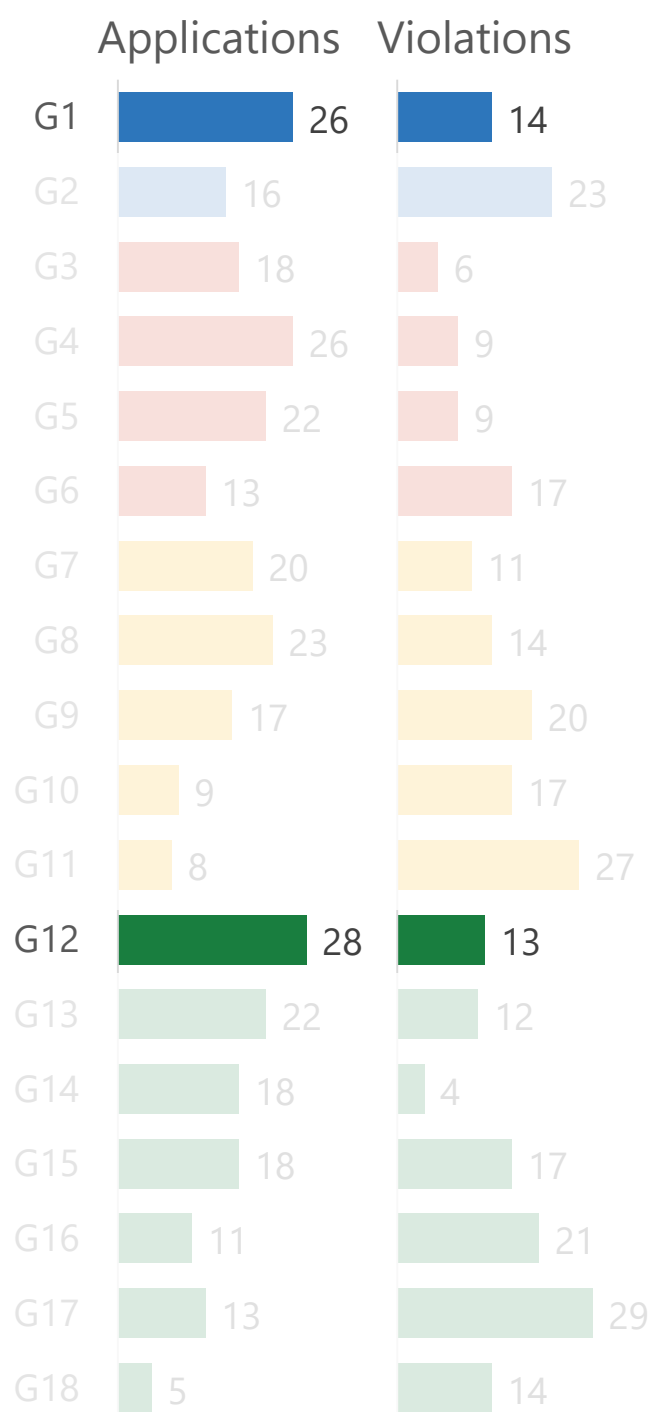
Violations





12

Remember
recent
interactions.

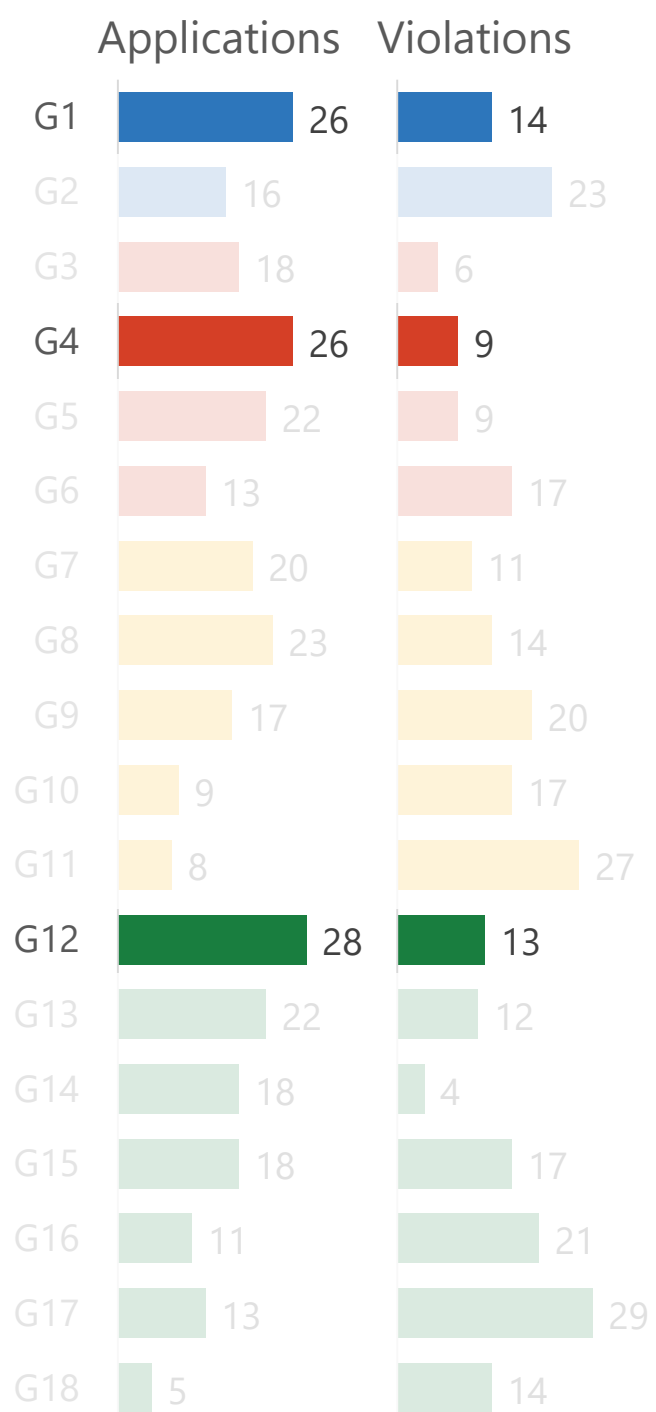


12

Remember
recent
interact

1

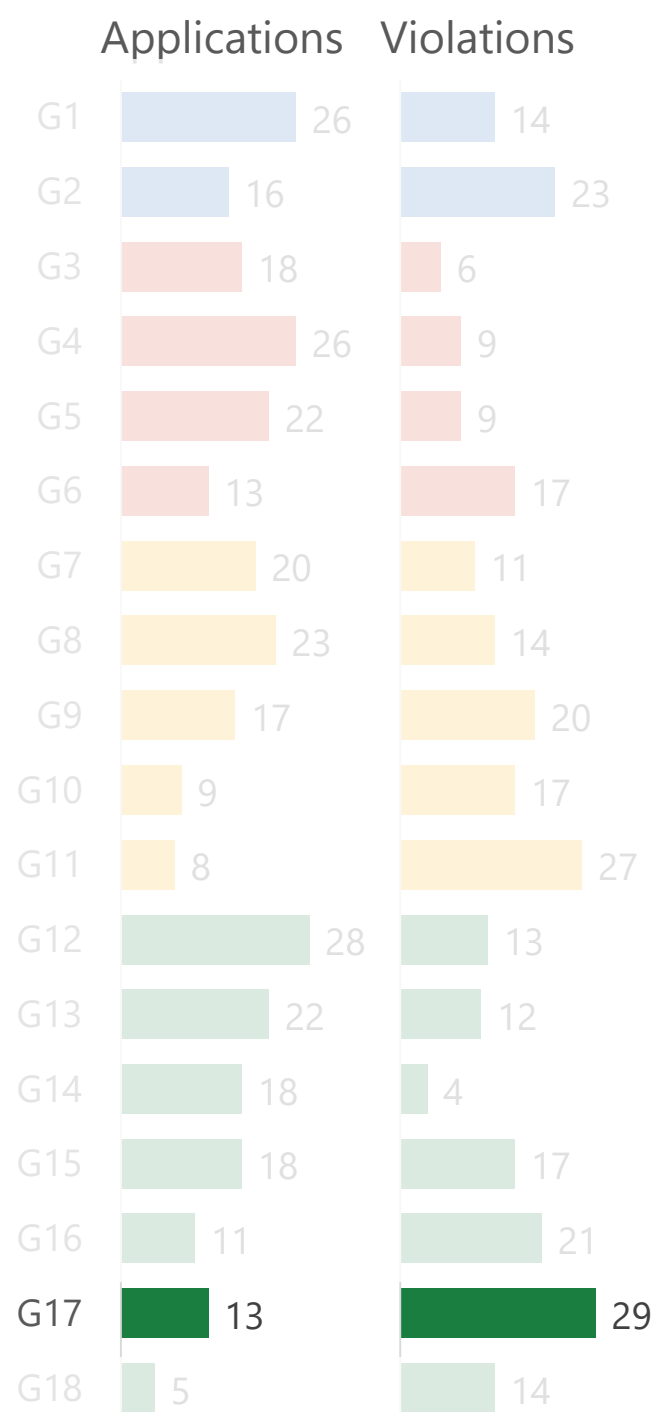
Make clear
what the
system
can do.



12
Remember
recent
interact

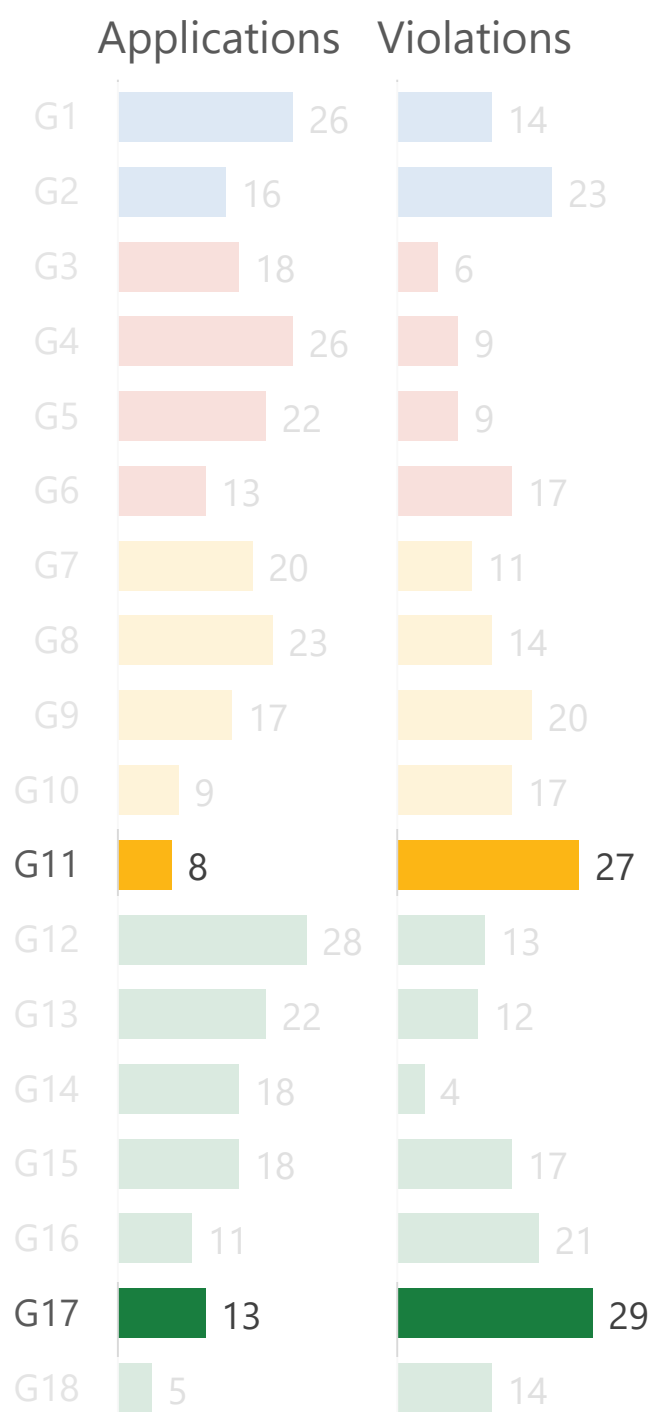
1
Make clear
what the
system
can do.

4
Show
contextually
relevant
information.



17

Provide
global
controls.

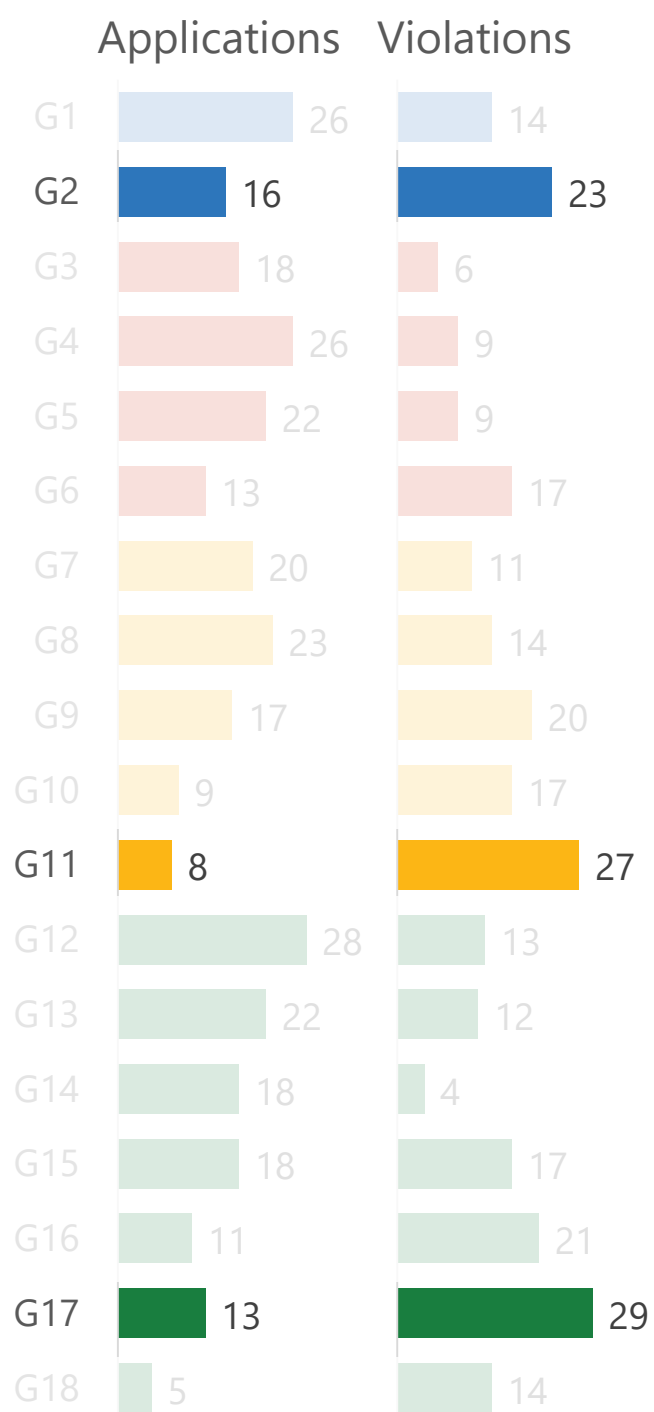


17

Provide
global
controls.

11

Make clear
why the
system did
what it did.



17
Provide
global
controls.

11
Make clear
why the
system did
what it did

2
Make clear
how well the
system can
do what it
can do.

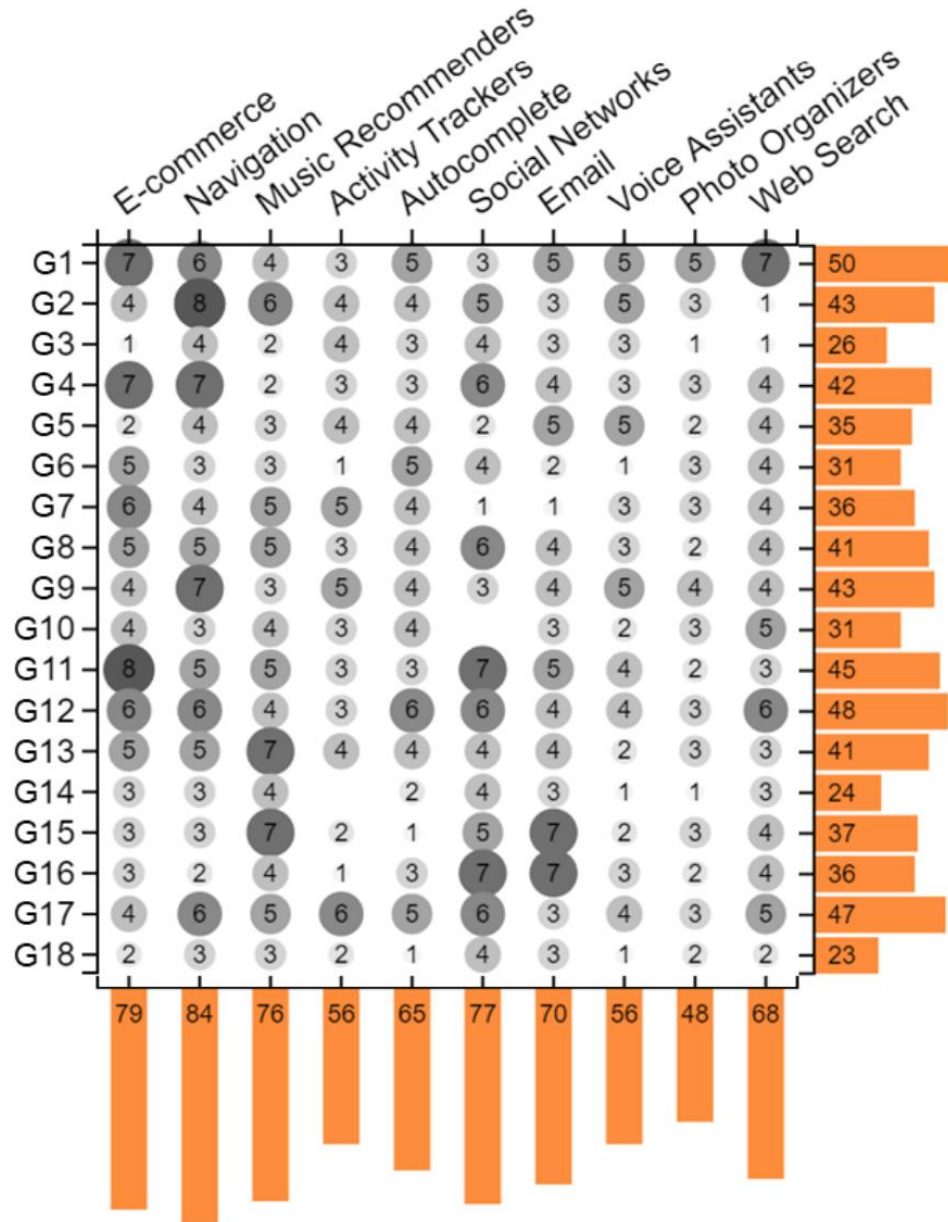
Consolidate into a Library (Work in Progress)

Types of content: examples, patterns,
research, code

Tagged by guideline and scenario
with faceted search and filtering

Comments and ratings to support
learning

Grow with examples and case studies
submitted by practitioners



Findings & Impact

Initial Impact

Opportunity Analysis

Engagements with Practitioners



Workshops and Courses

Q & A Break

Agenda

Intro to the guidelines

Findings and impact

Engineering and AI implications

Challenges for Intelligible AI

Agenda

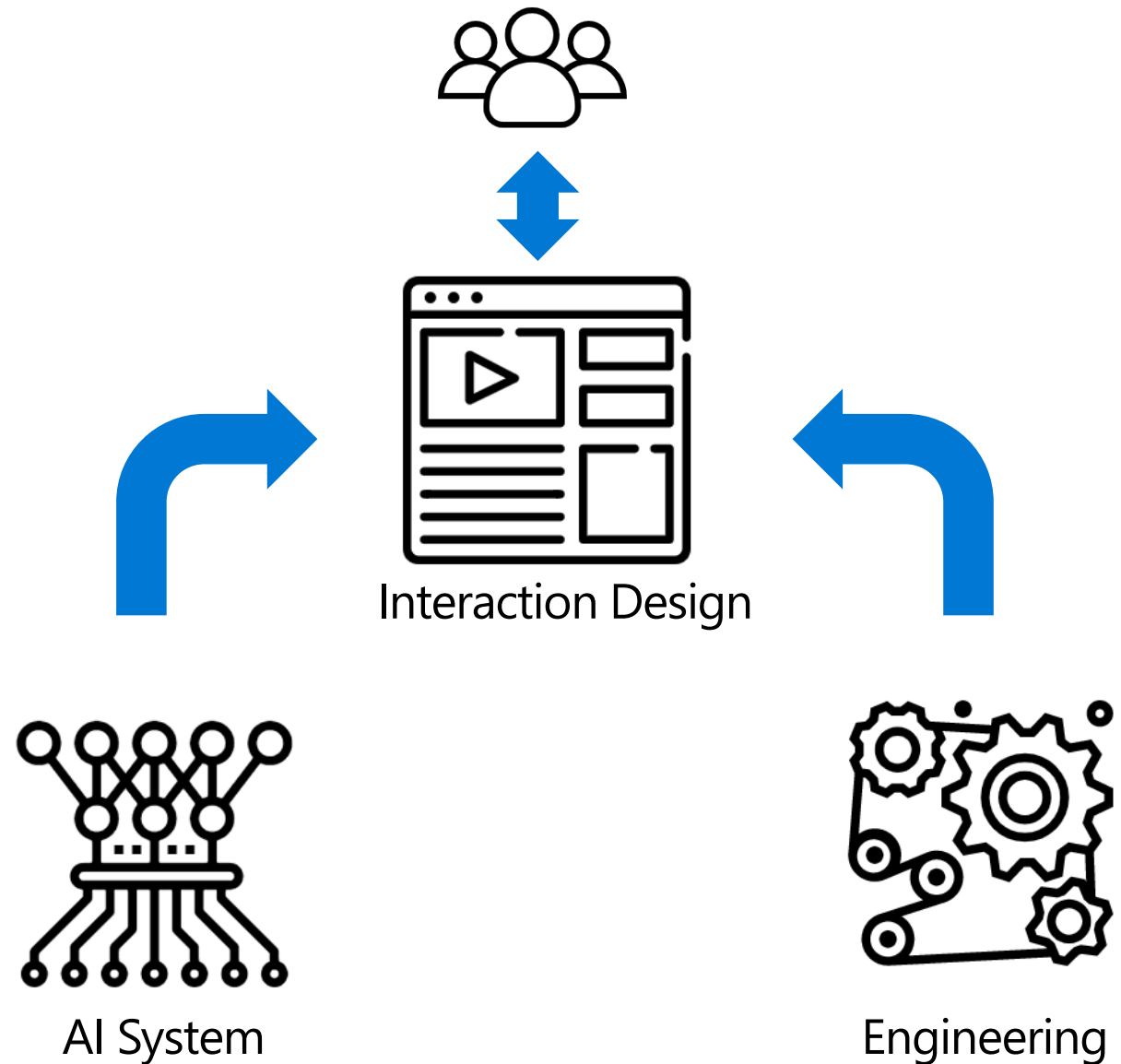
Intro to the guidelines

Findings and impact

Engineering and AI implications

Challenges for Intelligible AI

How can I implement the HAI Guidelines?



Interaction Design for AI requires ML & Eng Support

3

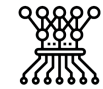
Time services
based on
context.



Hard to implement if the logging infrastructure is oblivious to context.

10

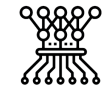
Scope
services when
in doubt.



Does the ML algorithm know or state that it is "in doubt"?

11

Make clear
why the
system did
what it did.



Is the ML algorithm explainable?

Setting expectations right – Performance reports

1

Make clear what the system can do.

2

Make clear how well the system can do what it can do.

AI-powered scans can identify people at risk of a fatal heart attack almost a **DECADE in advance 'by looking at the entire iceberg and not just the tip'**

- The AI predicted heart risk with **90% accuracy**, according to data
- Current medical scans are only able to see 'the tip of the iceberg'
- It could benefit around 350,000 in Britain, cardiologists believe
- Government funding will fast track the tech into the NHS in two years

Setting expectations right – Performance reports

1

Make clear what the system can do.

2

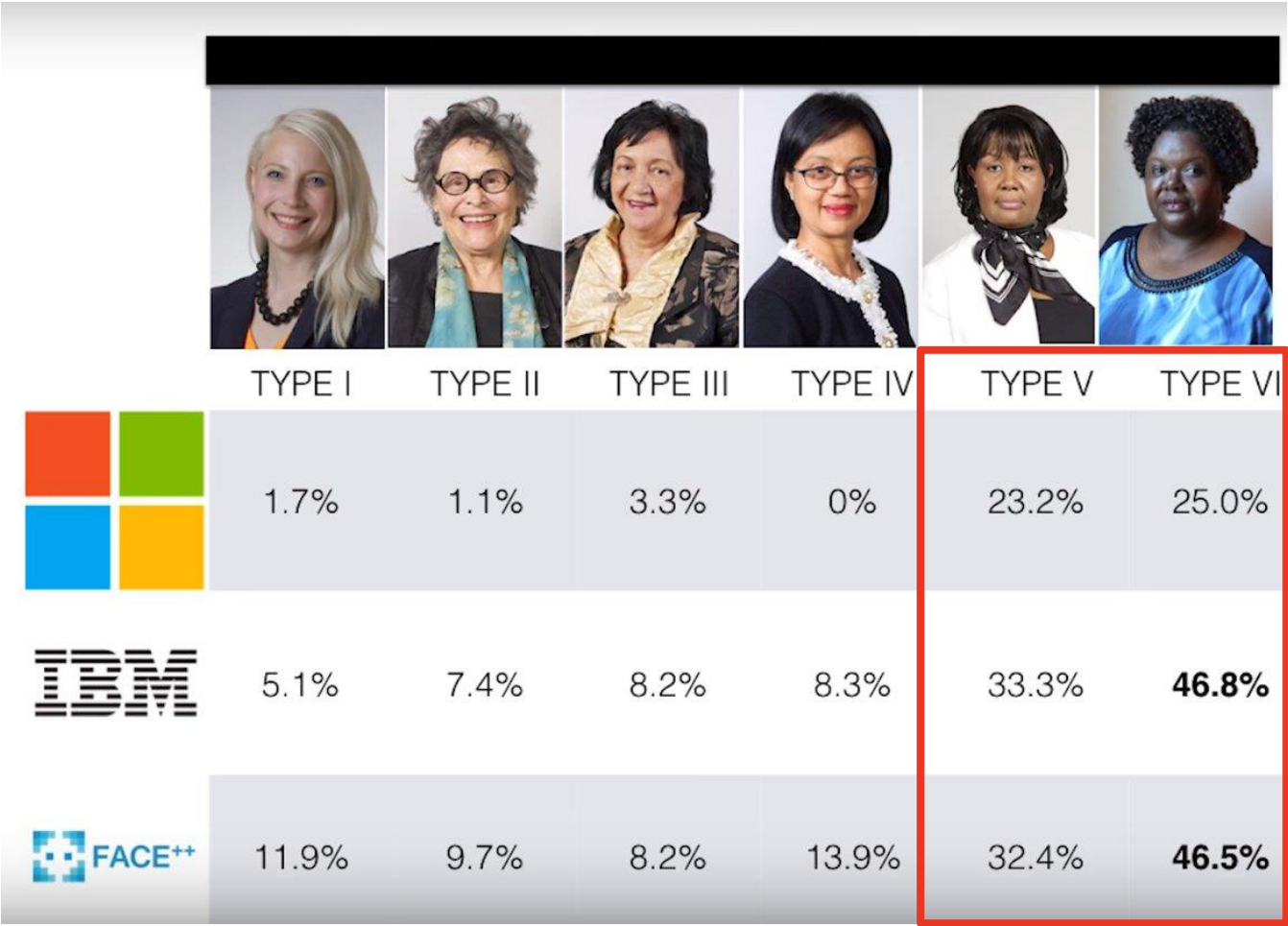
Make clear how well the system can do what it can do.

In the money Gold Silver Bronze								
#	Team Name	Notebook	Team Members	Score ?	Entries	Last		
1	PFDet		   +3	0.62882	49	1y		
2	Avengers		   	0.62161	48	1y		
3	kivajok		  	0.61707	102	1y		
4	XJTU			0.61559	22	1y		
5	ikciting		   +5	0.59472	39	1y		
6	Sogou_MM		  	0.57936	105	1y		
7	QLearning		   	0.56688	20	1y		
8	[RingUkraine] CloudResearch		 	0.53742	50	1y		
9	Res101+SoftNMS		 	0.53413	29	1y		
10	Kyle L.			0.51464	53	1y		

Setting expectations right – Gender Shades study

1
Make clear what the system can do.

2
Make clear how well the system can do what it can do.



90% accuracy



?	?	?
?	?	?

[Buolamwini, J. & Gebru, T. 2018]

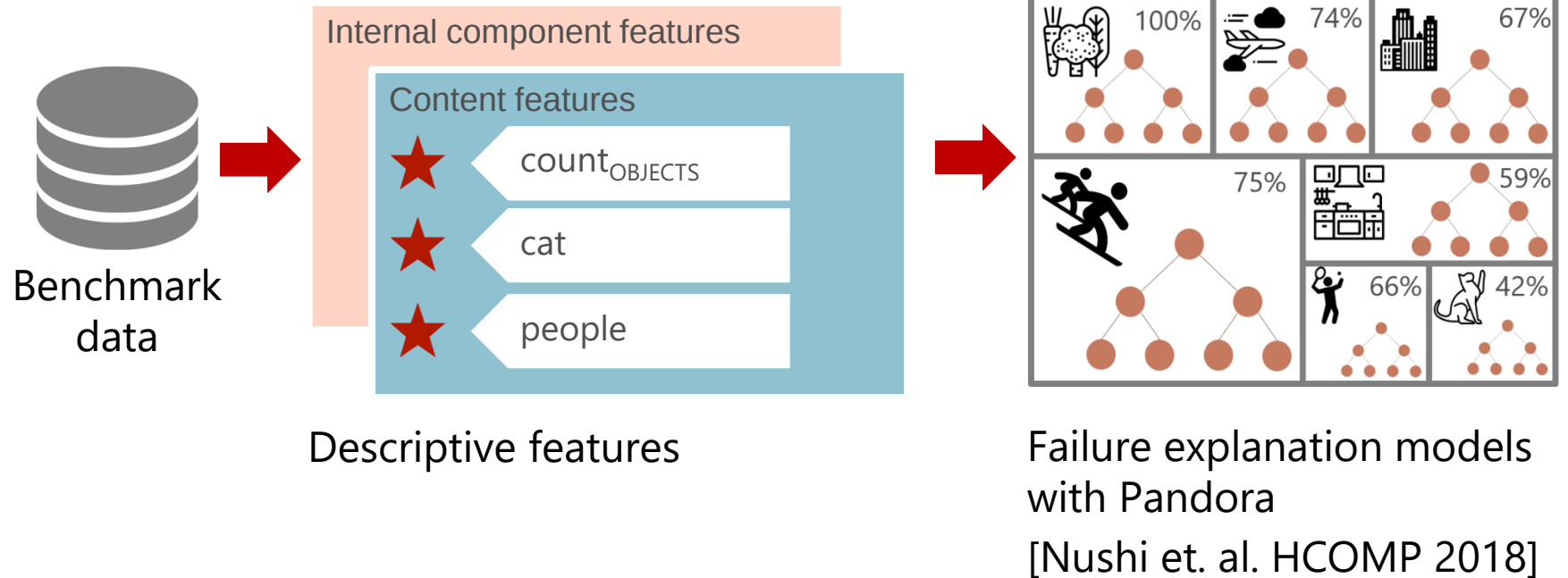
Setting expectations right – Error Terrain Analysis

1

Make clear
what the
system
can do.

2

Make clear
how well the
system can
do what it
can do.



Decision Tree

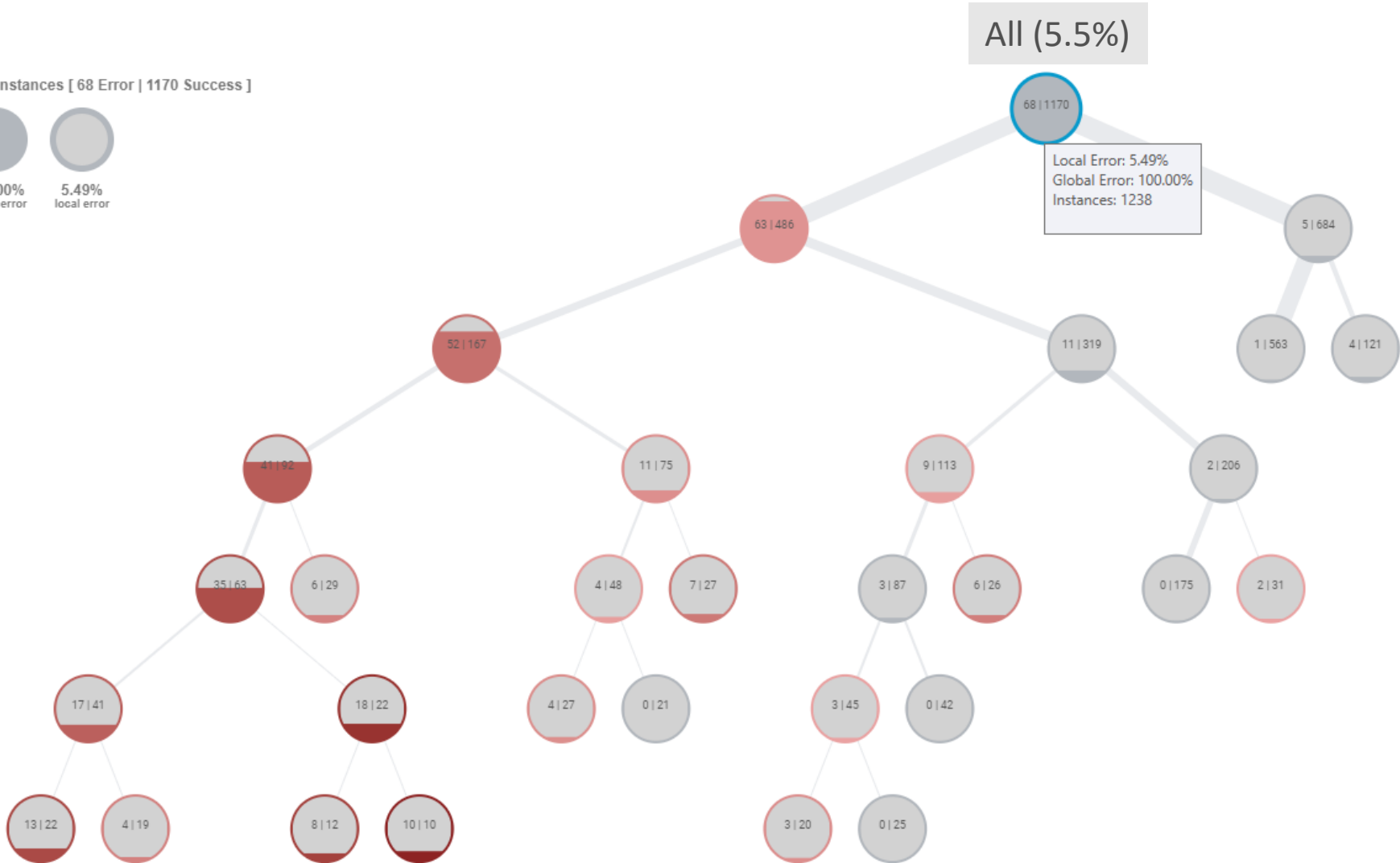
🔍 Type to filter

NAME	GAIN
☒ gender_gt	
☒ facialHair_sid	
☒ facialHair_mc	
☒ facialHair_be	
☒ skin_type_gt	
☒ hair_length_g	
☒ accessories_	
☒ age	
☒ hair_bald	
☒ smile	
☒ noise_noiseL	
☒ makeup_eyel	
☒ glasses_gt	
☒ glasses	
☒ hair_invisible	
☒ occlusion_for	
☒ exposure_exp	

APPLY



1238 Instances [68 Error | 1170 Success]



Decision Tree

🔍 Type to filter

NAME

GAIN

☒ gender_gt

☒ facialHair_sid

☒ facialHair_mc

☒ facialHair_be

☒ skin_type_gt

☒ hair_length_g

☒ accessories_

☒ age

☒ hair_bald

☒ smile

☒ noise_noiseL

☒ makeup_eyel

☒ glasses_gt

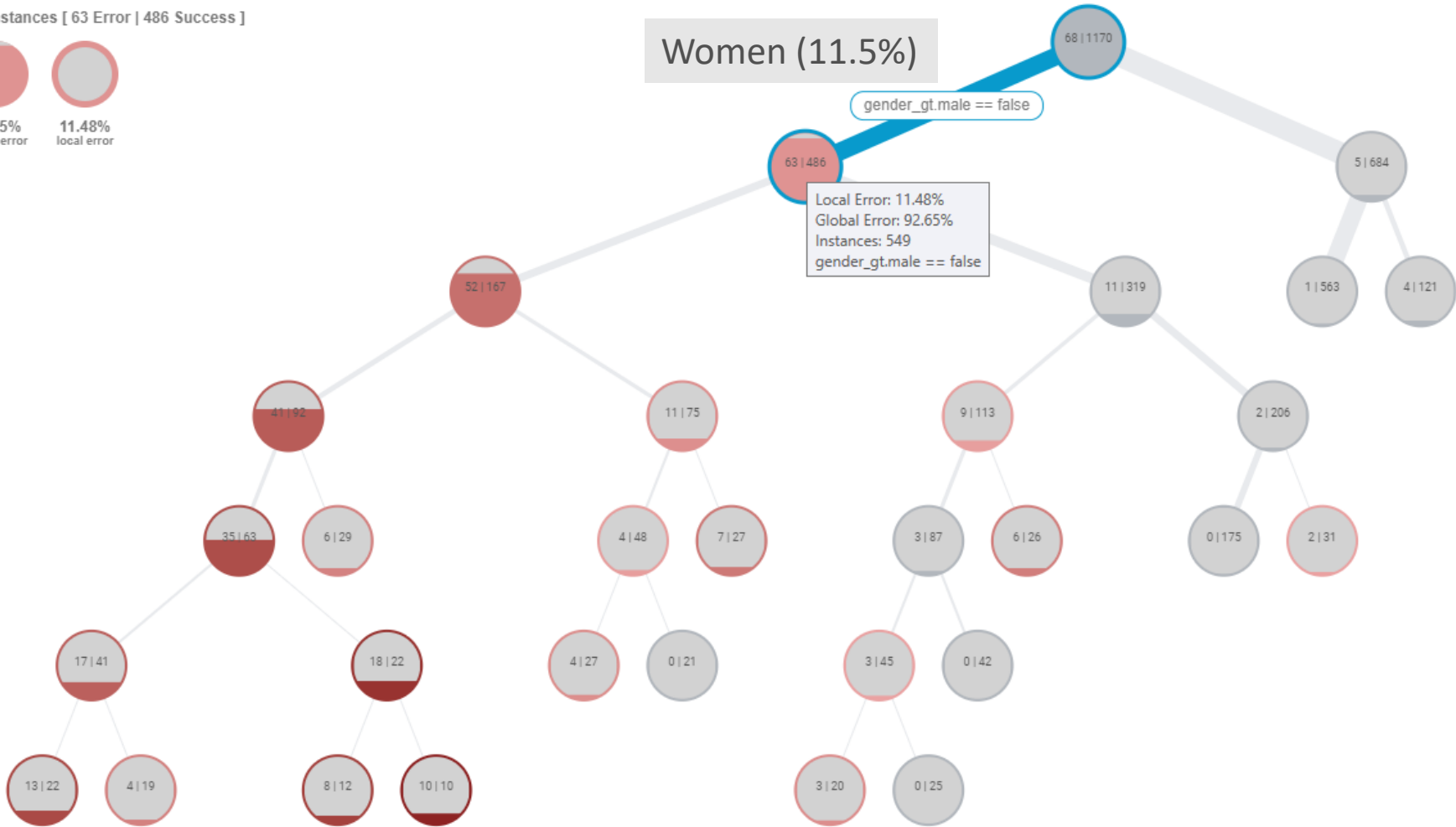
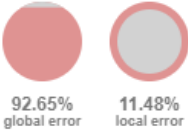
☒ glasses

☒ hair_invisible

☒ occlusion_for

☒ exposure_exp

549 Instances [63 Error | 486 Success]



Local Error: 11.48%
Global Error: 92.65%
Instances: 549
gender_gt.male == false

APPLY



Decision Tree

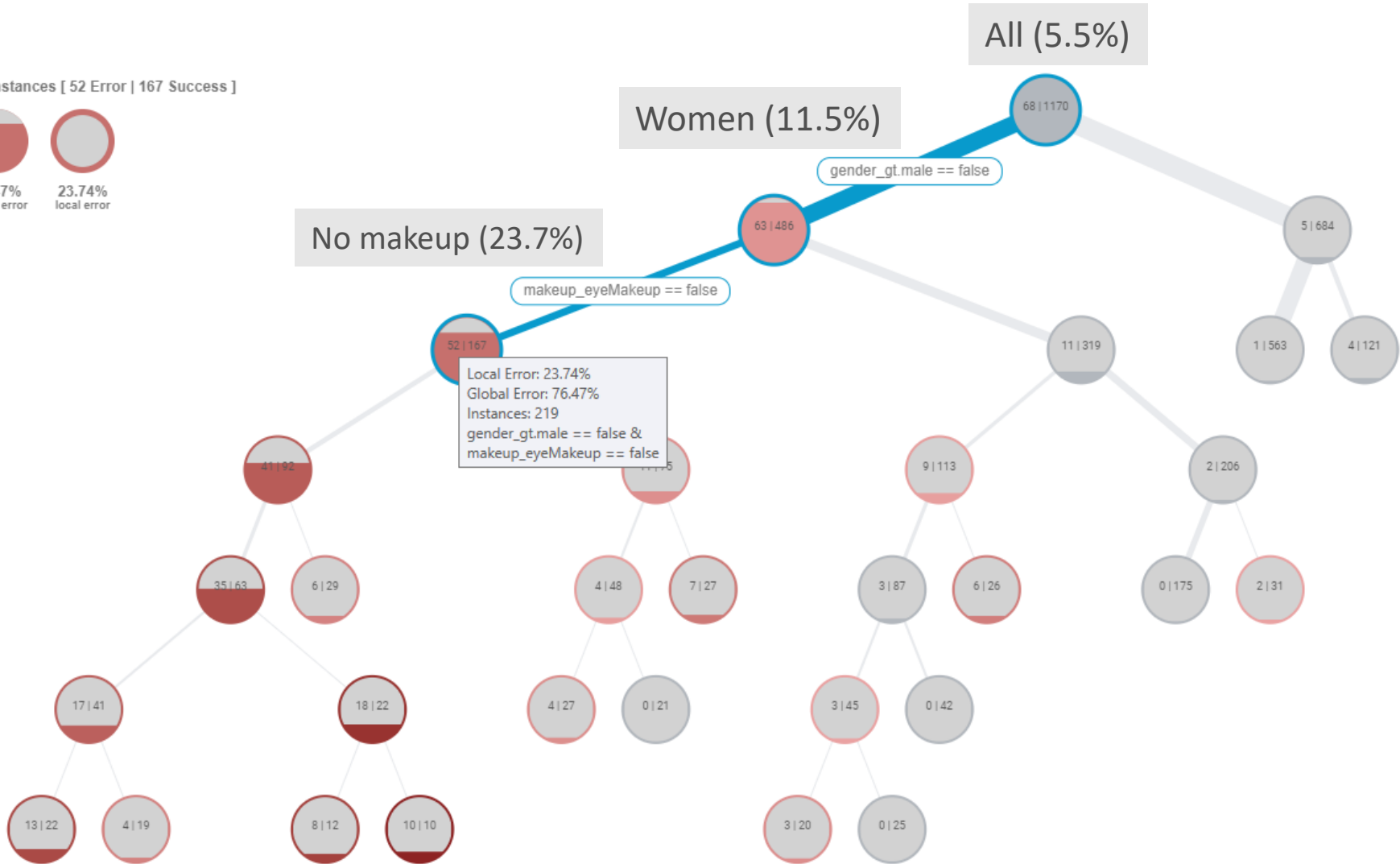
🔍 Type to filter

NAME	GAIN
☒ gender_gt	
☒ facialHair_sid	
☒ facialHair_mc	
☒ facialHair_be	
☒ skin_type_gt	
☒ hair_length_g	
☒ accessories_	
☒ age	
☒ hair_bald	
☒ smile	
☒ noise_noiseL	
☒ makeup_eyeM	
☒ glasses_gt	
☒ glasses	
☒ hair_invisible	
☒ occlusion_for	
☒ exposure_exp	

APPLY



219 Instances [52 Error | 167 Success]



Decision Tree

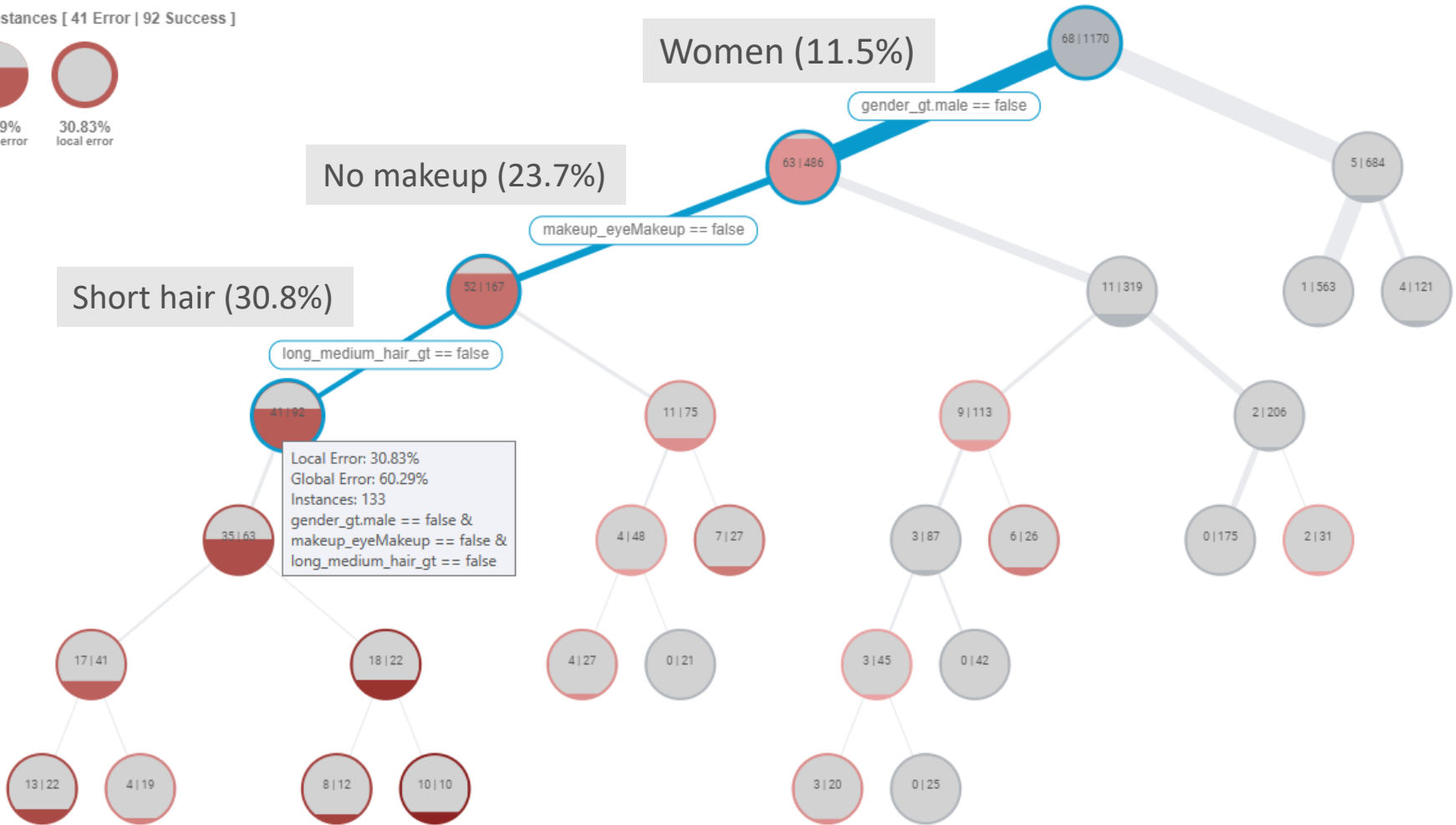
🔍 Type to filter

NAME	GAIN
☒ gender_gt	
☒ facialHair_sid	
☒ facialHair_mc	
☒ facialHair_be:	
☒ skin_type_gt	
☒ hair_length_g	
☒ accessories_	
☒ age	
☒ hair_bald	
☒ smile	
☒ noise_noiseL	
☒ makeup_eyeM	
☒ glasses_gt	
☒ glasses	
☒ hair_invisible	
☒ occlusion_for	
☒ exposure_exp	

APPLY



133 Instances [41 Error | 92 Success]

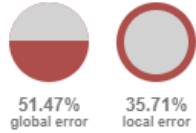


Decision Tree

🔍 Type to filter

NAME	GAIN
☒ gender_gt	
☒ facialHair_sid	
☒ facialHair_mc	
☒ facialHair_be	
☒ skin_type_gt	
☒ hair_length_g	
☒ accessories_	
☒ age	
☒ hair_bald	
☒ smile	
☒ noise_noiseL	
☒ makeup_eyeM	
☒ glasses_gt	
☒ glasses	
☒ hair_invisible	
☒ occlusion_for	
☒ exposure_exp	

98 Instances [35 Error | 63 Success]



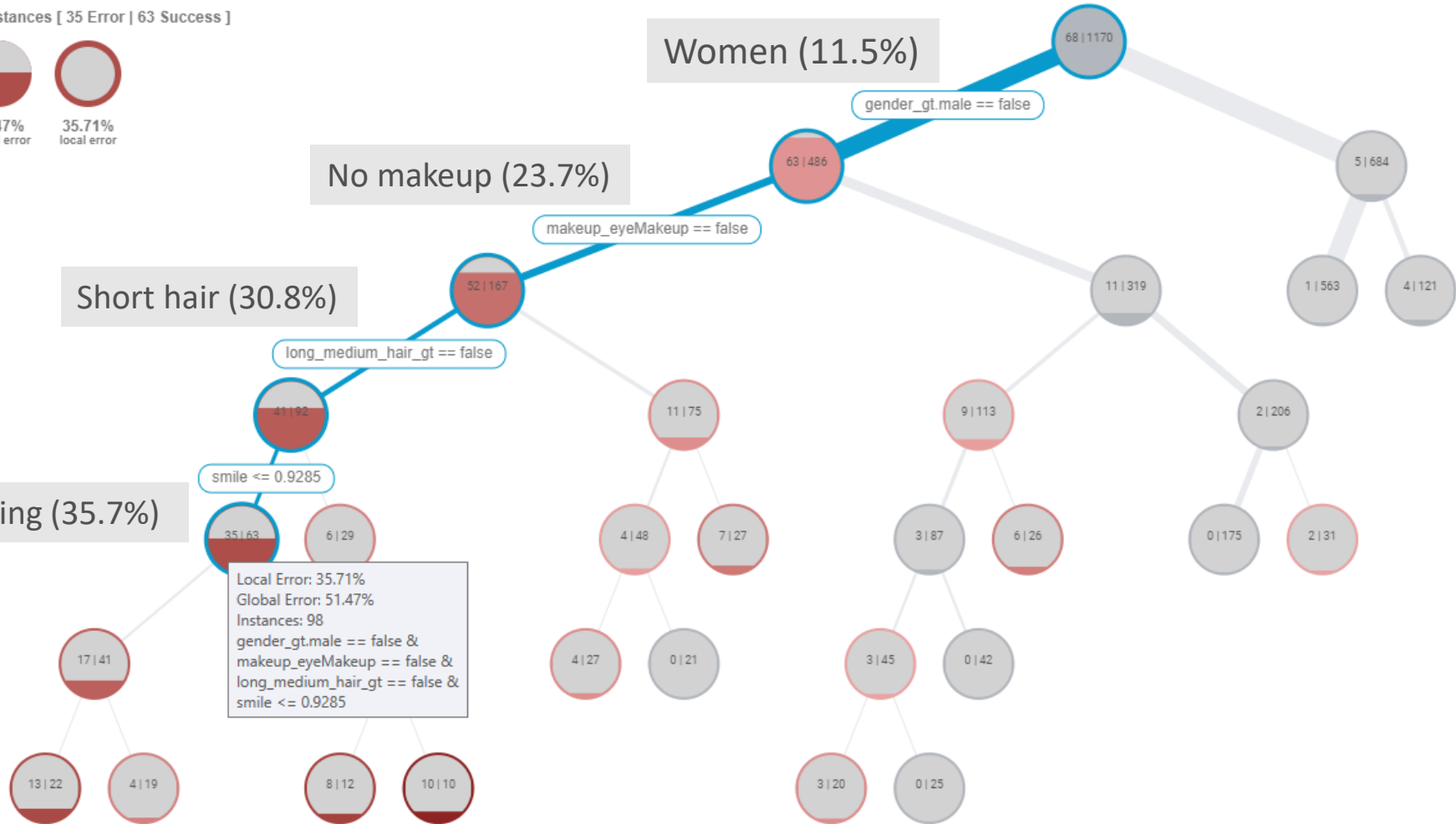
Not smiling (35.7%)

Short hair (30.8%)

No makeup (23.7%)

Women (11.5%)

All (5.5%)



Local Error: 35.71%
Global Error: 51.47%
Instances: 98
gender_gt.male == false &
makeup_eyeMakeup == false &
long_medium_hair_gt == false &
smile <= 0.9285

APPLY



Setting expectations right – Error Analysis

1

Make clear
what the
system
can do.

2

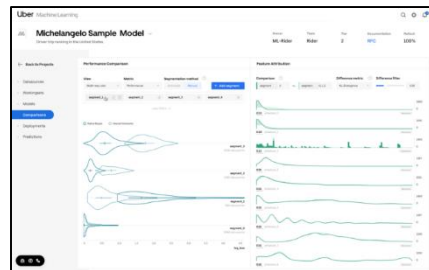
Make clear
how well the
system can
do what it
can do.



Error Terrain Analysis \ Pandora
[Nushi et. al. HCOMP 2018]



Errudite
[Wu et. al. ACL 2019]



Manifold
[Zhang et. al. IEEE TVCG 2018]

Setting expectations right: other implications

1

Make clear what the system can do.

2

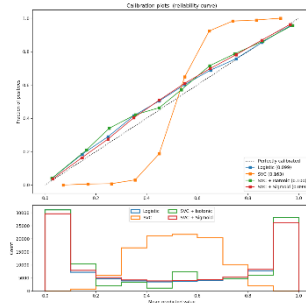
Make clear how well the system can do what it can do.



Use multiple and realistic benchmarks

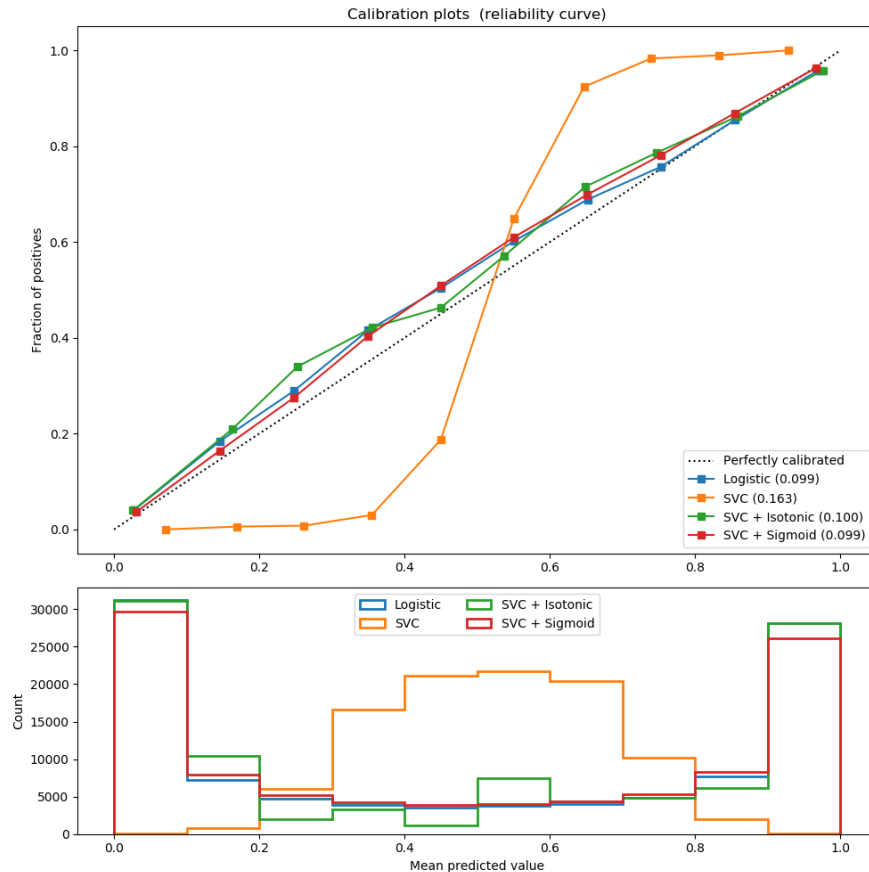


Estimate the cost and risk of mistakes



Calibrate and explain uncertainty

Setting expectations right – Uncertainty Calibration



Post-hoc calibration:

Platt scaling, Isotonic regression
[Platt et al., 1999; Zadrozny & Elkan, 2001]

In-built model uncertainty

Bayesian DNNs, Ensemble methods
[Gal & Ghahramani, 2016; Osband et al., 2016]

https://scikit-learn.org/stable/auto_examples/calibration/plot_calibration_curve.html

Setting expectations right – Uncertainty explanation

Explaining "Probability of Precipitation"

Forecasts issued by the National Weather Service routinely include a "PoP" (probability of precipitation) statement, which is often expressed as the "chance of rain" or "chance of precipitation".

EXAMPLE

ZONE FORECASTS FOR NORTH AND CENTRAL GEORGIA
NATIONAL WEATHER SERVICE PEACHTREE CITY GA
119 PM EDT THU MAY 8 2008

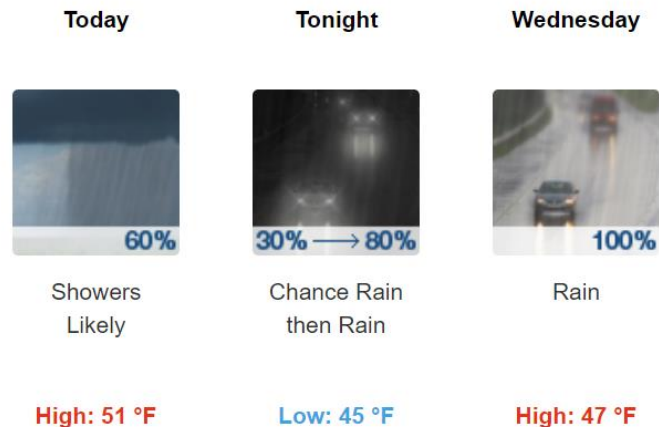
GAZ021-022-032034-044046-055-057-090815-
CHEROKEE-CLAYTON-COBB-DEKALB-FORSYTH-GWINNETT-HENRY-NORTH FULTON-
ROCKDALE-SOUTH FULTON-
INCLUDING THE CITIES OF...ATLANTA...CONYERS...DECATUR...
EAST POINT...LAWRENCEVILLE...MARIETTA
119 PM EDT THU MAY x 2008

.THIS AFTERNOON...MOSTLY CLOUDY WITH A 40 PERCENT CHANCE OF
SHOWERS AND THUNDERSTORMS. WINDY. HIGHS IN THE LOWER 80S. NEAR
STEADY TEMPERATURE IN THE LOWER 80S. SOUTH WINDS 15 TO 25 MPH.
.TONIGHT...MOSTLY CLOUDY WITH A CHANCE OF SHOWERS AND
THUNDERSTORMS IN THE EVENING...THEN A SLIGHT CHANCE OF SHOWERS
AND THUNDERSTORMS AFTER MIDNIGHT. LOWS IN THE MID 60S. SOUTHWEST
WINDS 5 TO 15 MPH. CHANCE OF RAIN 40 PERCENT.

What does this "40 percent" mean? ...will it rain 40 percent of of the time? ...will it rain over 40 percent of the area?

The "Probability of Precipitation" (PoP) simply describes **the probability that the forecast grid/point in question will receive at least 0.01" of rain**. So, in this example, there is a 40 percent probability for at least 0.01" of rain at the specific forecast point of interest!

Extended Forecast for Downtown Seattle WA



<https://forecast.weather.gov/>

Setting expectations right – Uncertainty explanation



Probably a yellow school bus **driving** down a street

English (detected) ▾	🔊 🎤	Italian ▾	🔊 📄
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness		↔	E 'stata la migliore delle volte, è stata la peggiore delle volte, era l'età della saggezza, era l'età della follia
110/5000			

Context, Invocation, Dismissal

3

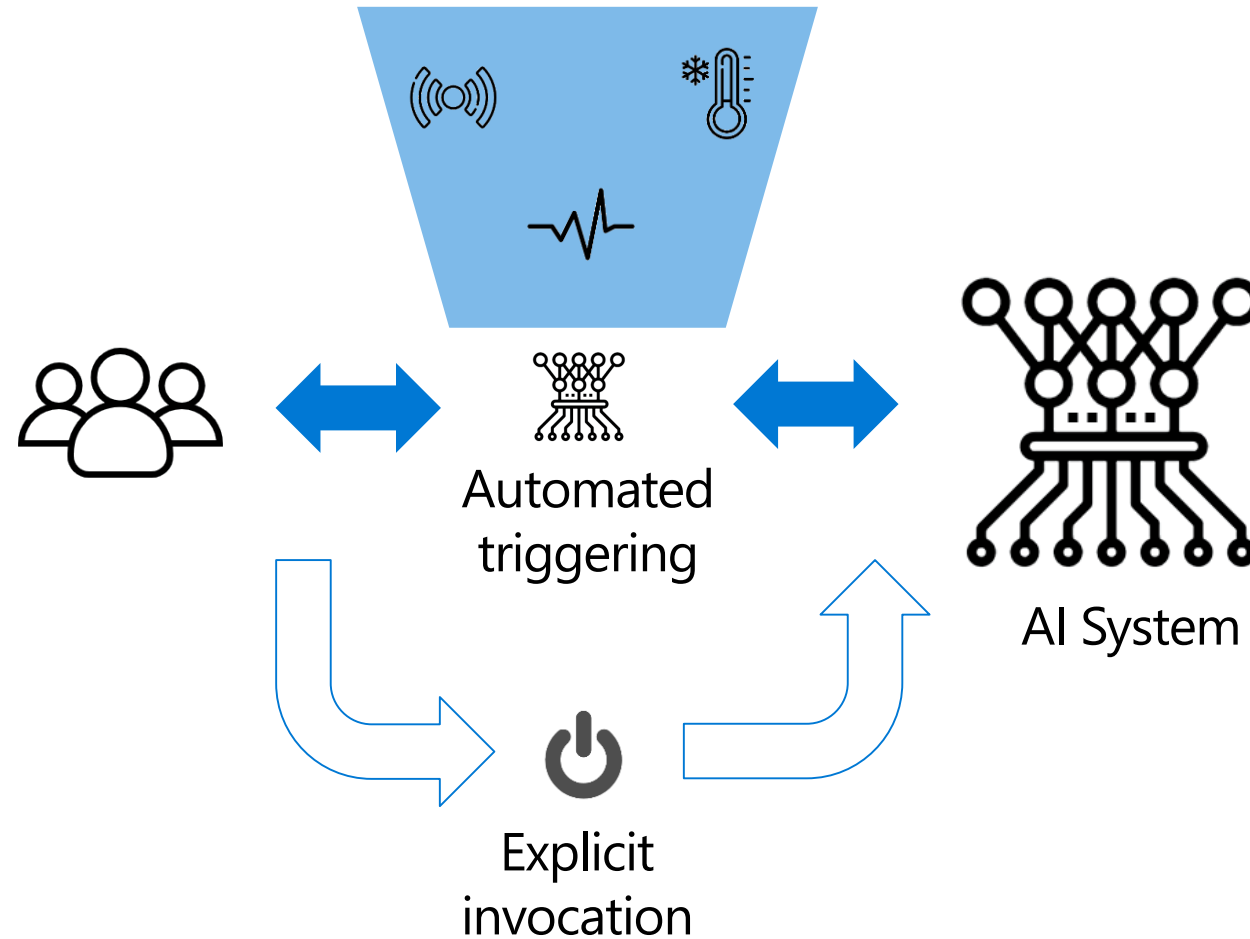
Time services
based on
context.

7

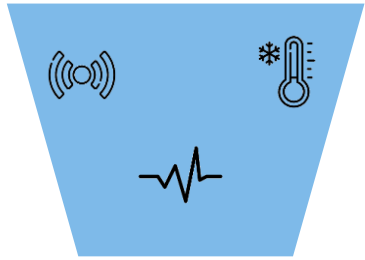
Support
efficient
invocation.

8

Support
efficient
dismissal.

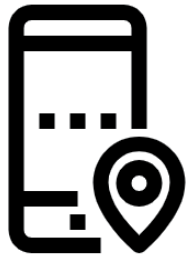


Context inference



Sensor Data
Infrastructure

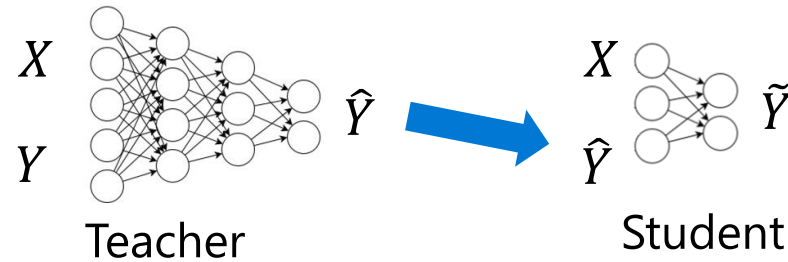
Privacy Concerns



ML on the Edge

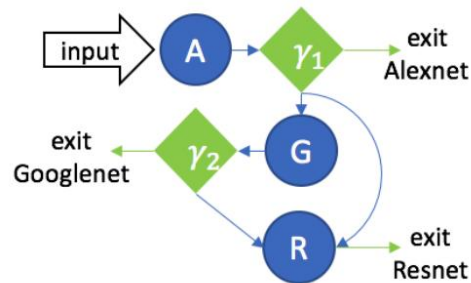
Model compression

[Ba and Caruana 2014; Hinton 2015]



Adaptive networks for inference

[Bolutkasi et. al 2017]



Context, Invocation, Dismissal

3

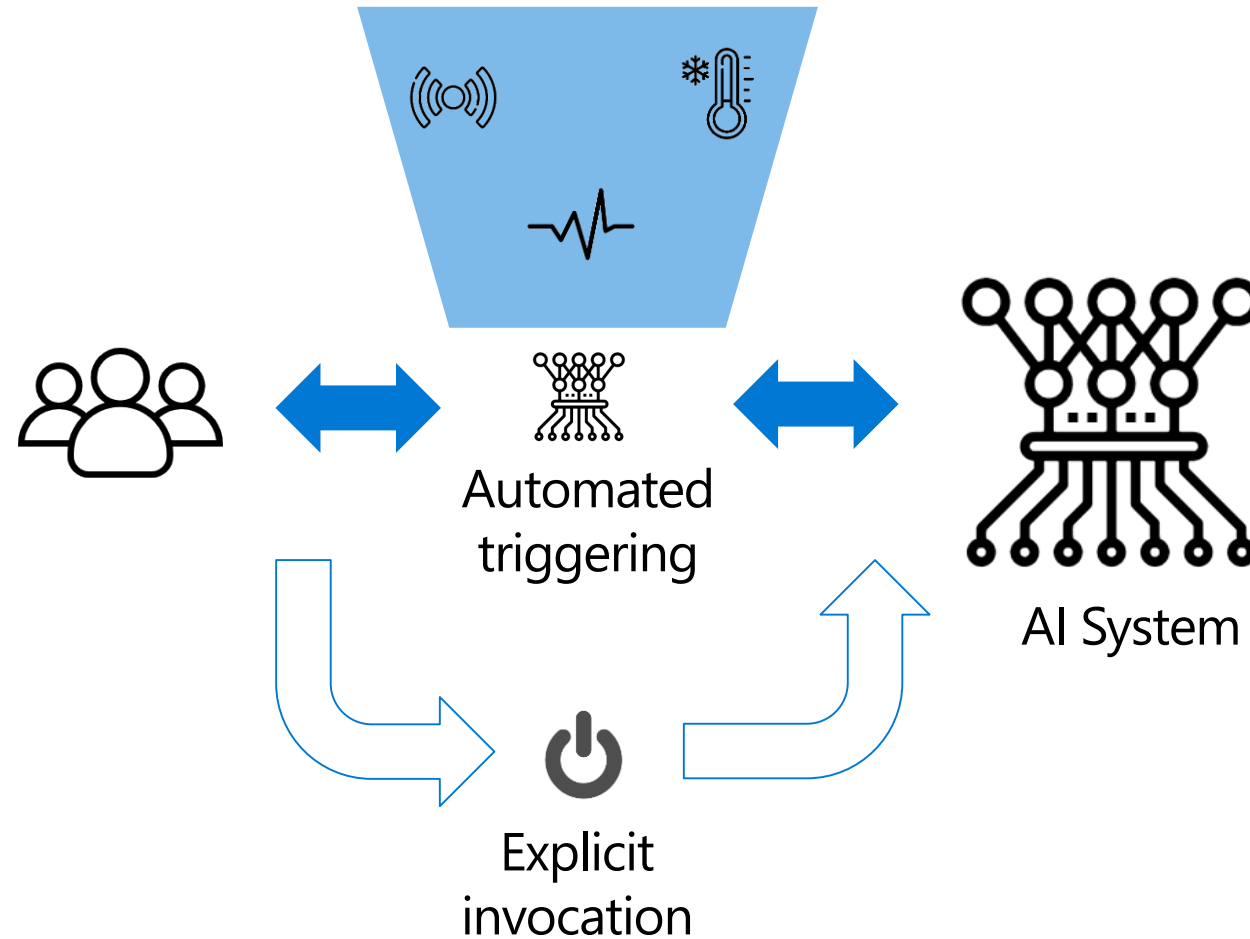
Time services
based on
context.

7

Support
efficient
invocation.

8

Support
efficient
dismissal.



Tuning automated triggering

3

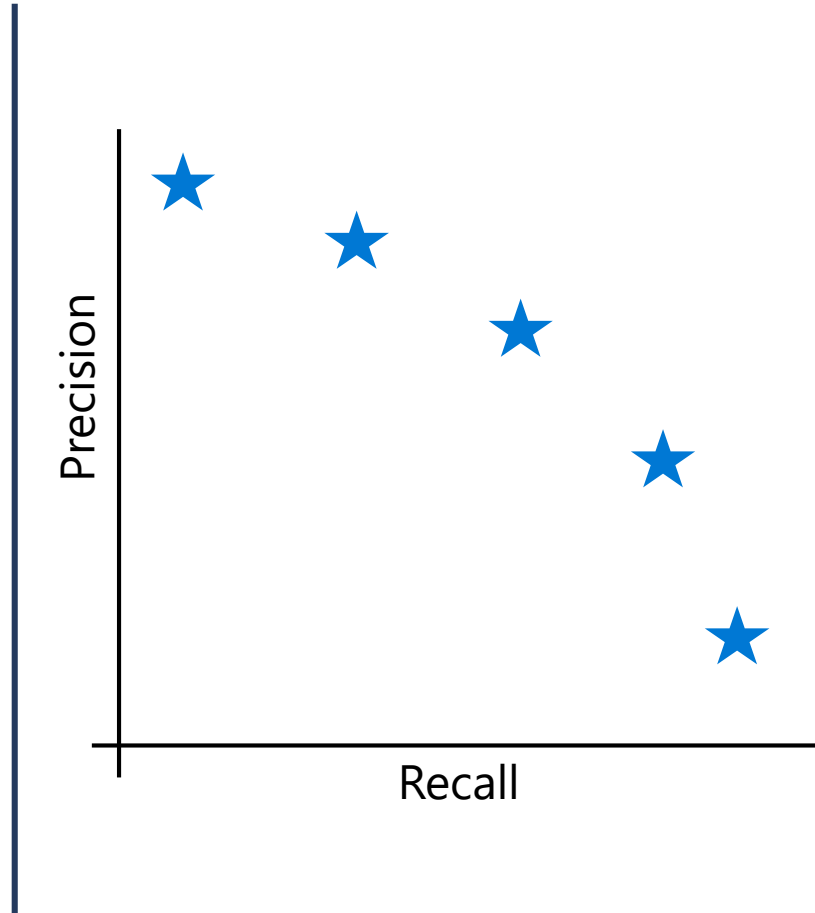
Time services based on context.

7

Support efficient invocation.

8

Support efficient dismissal.



Cost of explicit invocation
user time, accessibility

Cost of wrong invocation
cognitive load, dismissal time

Cost of wrong AI prediction
risk mitigation

Incorporating user feedback over time

13

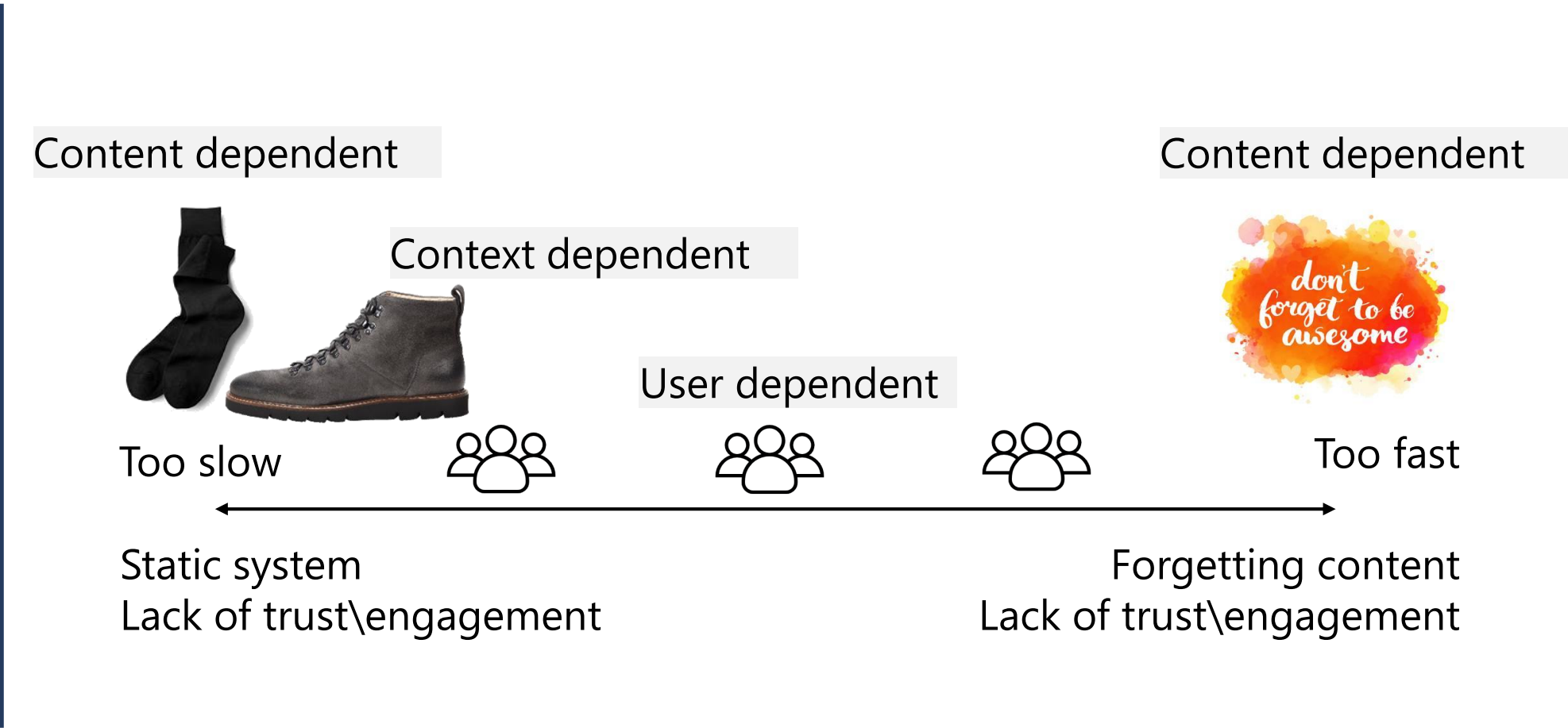
Learn from user behavior.

15

Encourage granular feedback.

14

Update and adapt cautiously.



Feature engineering

13

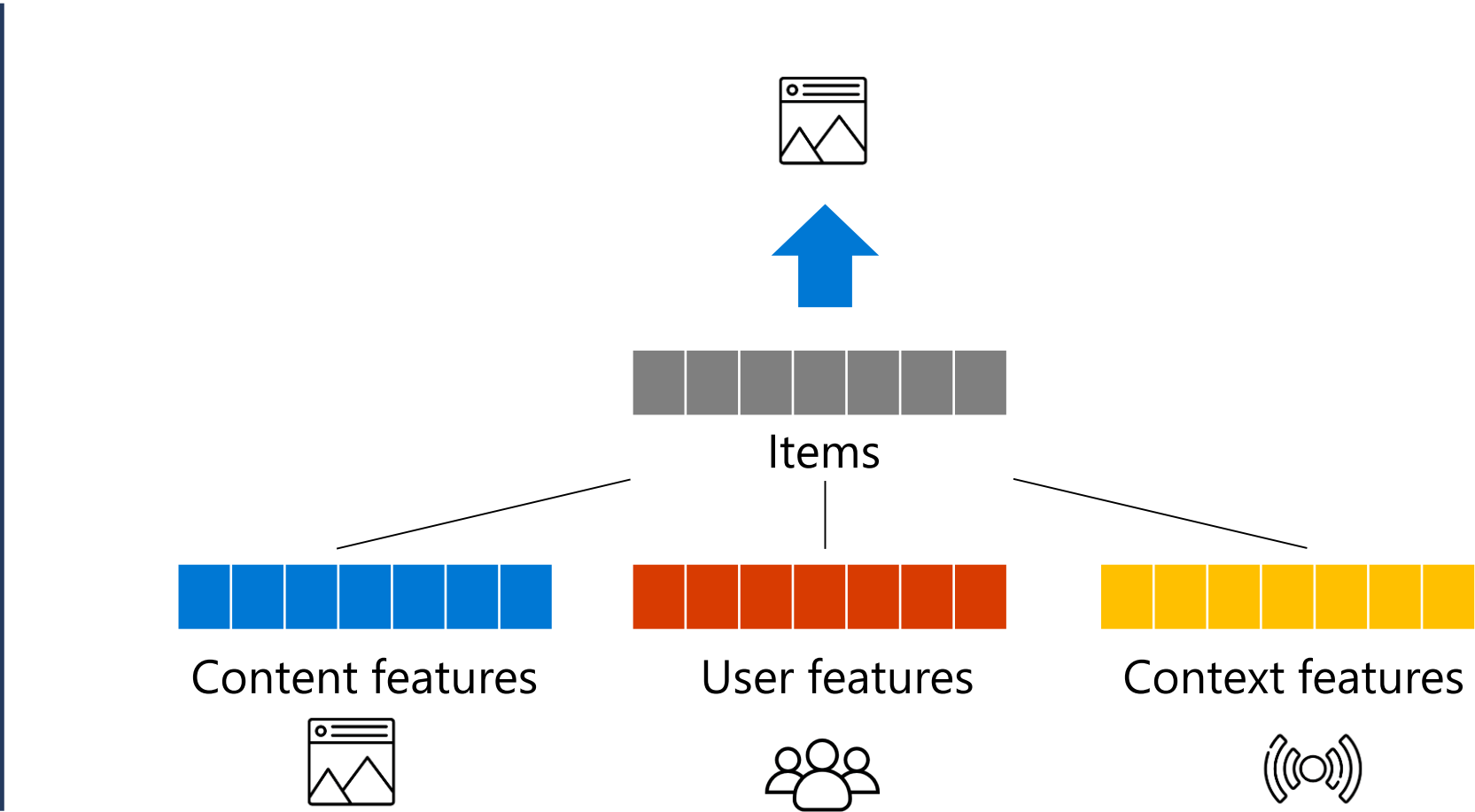
Learn from user behavior.

15

Encourage granular feedback.

14

Update and adapt cautiously.



Dealing with sparse data

13

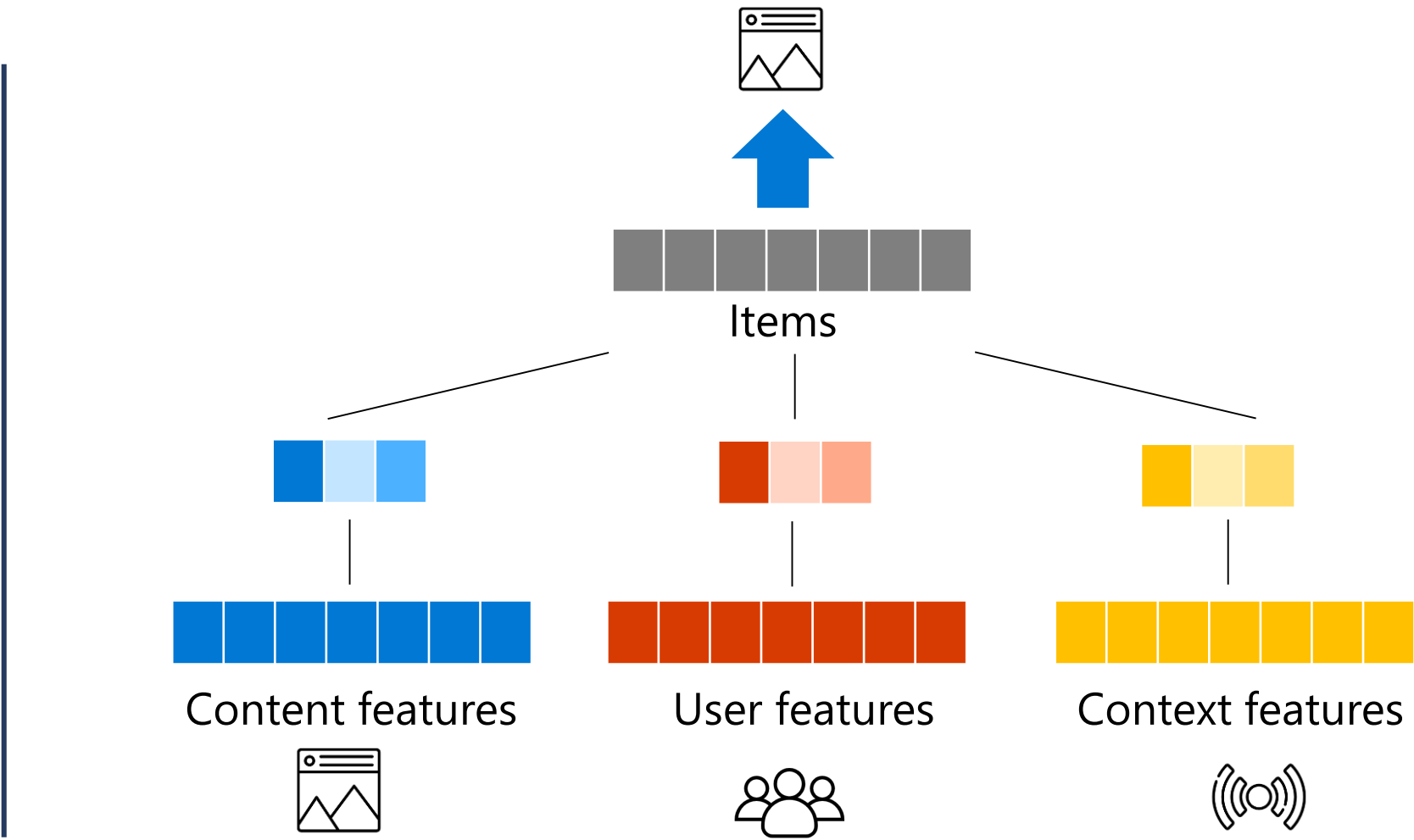
Learn from user behavior.

15

Encourage granular feedback.

14

Update and adapt cautiously.



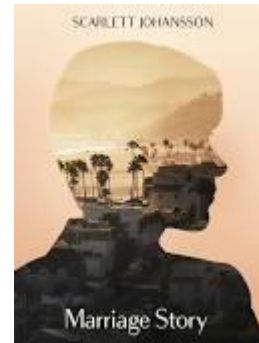
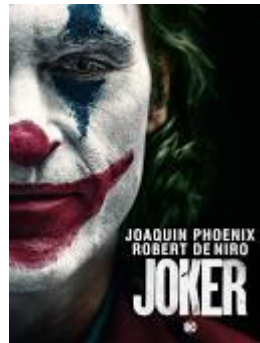
Global control support: feedback generalization

15

Encourage
granular
feedback.



Sci-fi



Drama

17

Provide
global
controls.



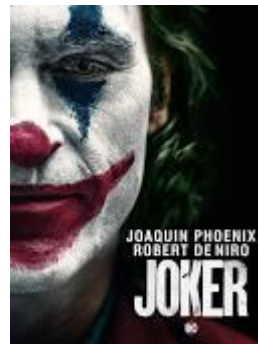
Global control support: feedback generalization

15

Encourage
granular
feedback.



Disney



Hollywood

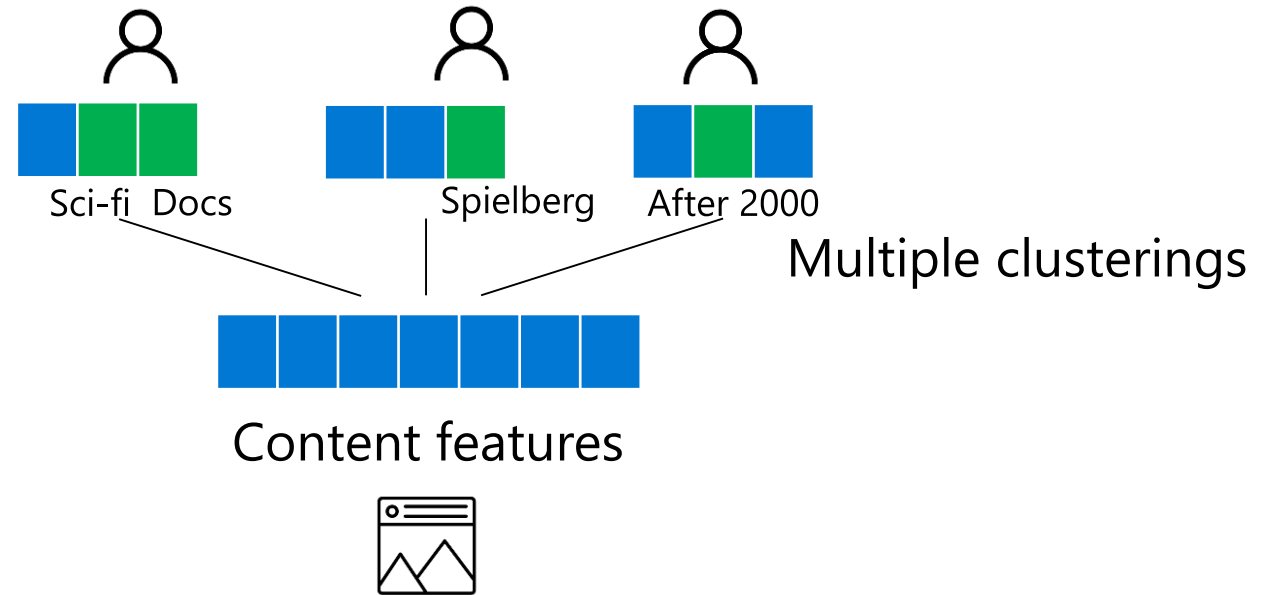
17

Provide
global
controls.

Global control support: feedback generalization

15

Encourage
granular
feedback.



17

Provide
global
controls.

Q & A

Is there any other functionality you know of or you wish you had in ML & Eng that could simplify Human-AI Interaction?

How much do interaction considerations impact ML & Engineering decisions?

What else do you (or your colleagues) do to support better Human-AI interaction?

Agenda

Intro to the guidelines

Findings and impact

Engineering and AI implications

Challenges for Intelligible AI

Agenda

Intro to the guidelines

Findings and impact

Engineering and AI implications

Challenges for Intelligible AI

Machine Learning Everywhere

11

Make clear
why the
system did
what it did.

12

Remember
recent
interactions.

13

Learn from
user behavior.



Intelligible, Transparent, Explainable AI

Terminology

Caveat: My take – No consensus here

• Predictable $\sim$ (Human) Simulate-able

Predict exactly what it will do

$\cap$

• Intelligible $\sim$ Transparent

Answer counterfactual
predict how a **change** to model's input
will **change** its output

$\cap$

• Explainable $\sim$ Interpretable

Construct rationalization for why
(maybe) it did what it did

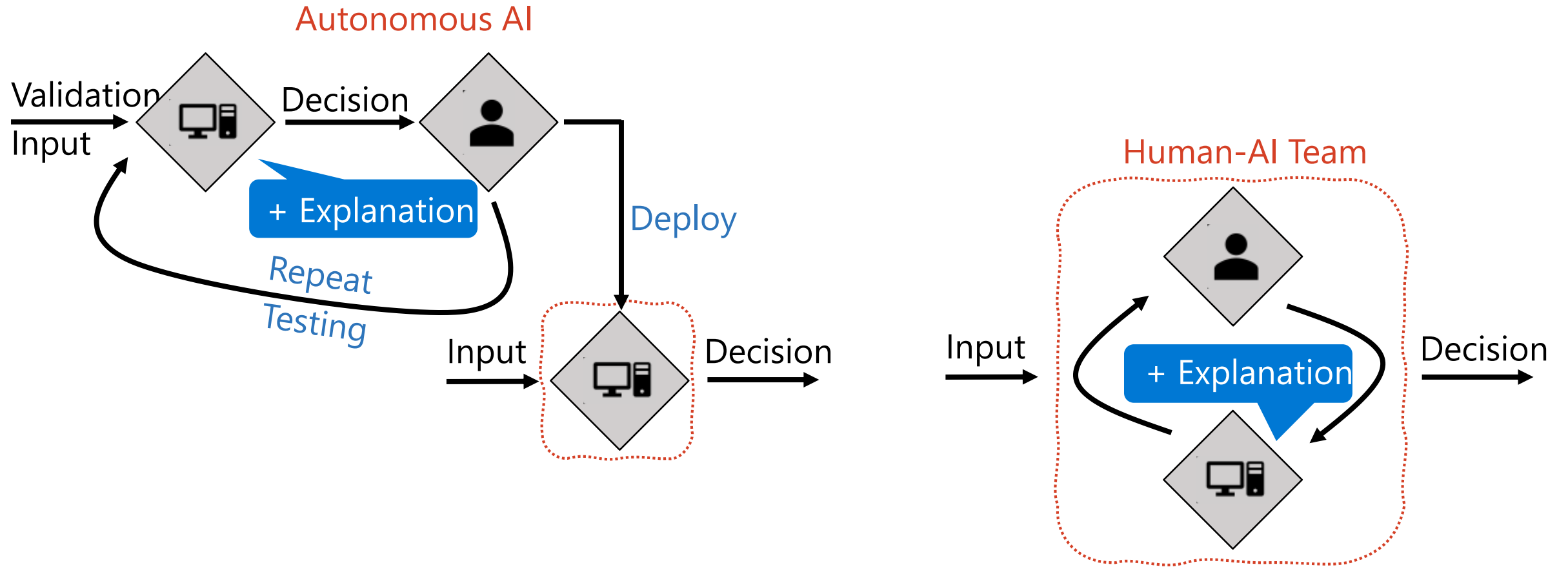
• Inscrutable $\supseteq$ Blackbox

Inscrutable: too complex to understand
Blackbox: know **nothing** about it

Reasons for Wanting Intelligibility

1. The AI May be Optimizing the Wrong Thing
2. Missing a Crucial Feature
3. Distributional Drift
4. Facilitating User Control in Mixed Human/AI Teams
5. User Acceptance
6. Learning for Human Insight
7. Legal Requirements

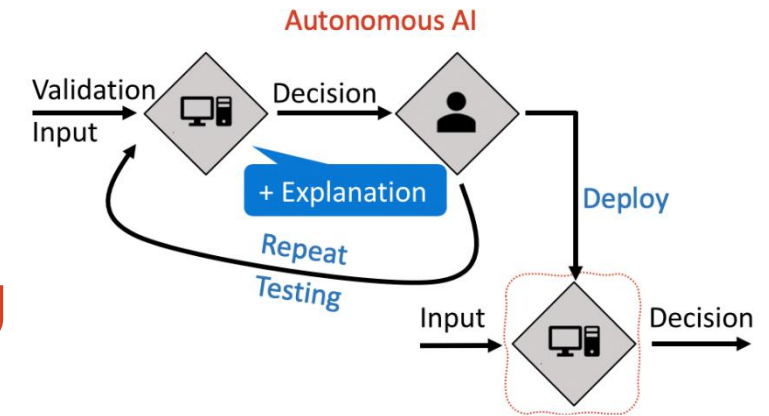
AI Deployments



Intelligibility Useful in Both Cases

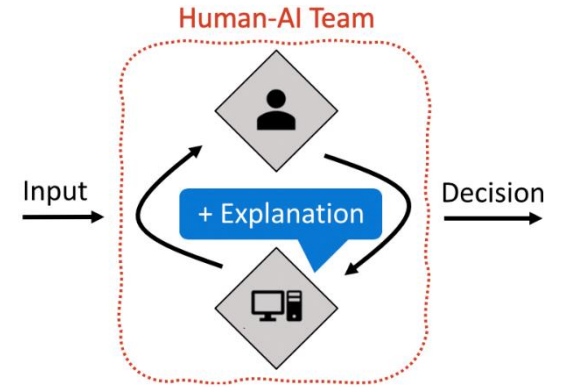
Reasons for Wanting Intelligibility

1. **The AI May be Optimizing the Wrong Thing**
2. **Missing a Crucial Feature**
3. **Distributional Drift**
4. Facilitating User Control in Mixed Human/AI Teams
5. User Acceptance
6. Learning for Human Insight
7. Legal Requirements



Reasons for Wanting Intelligibility

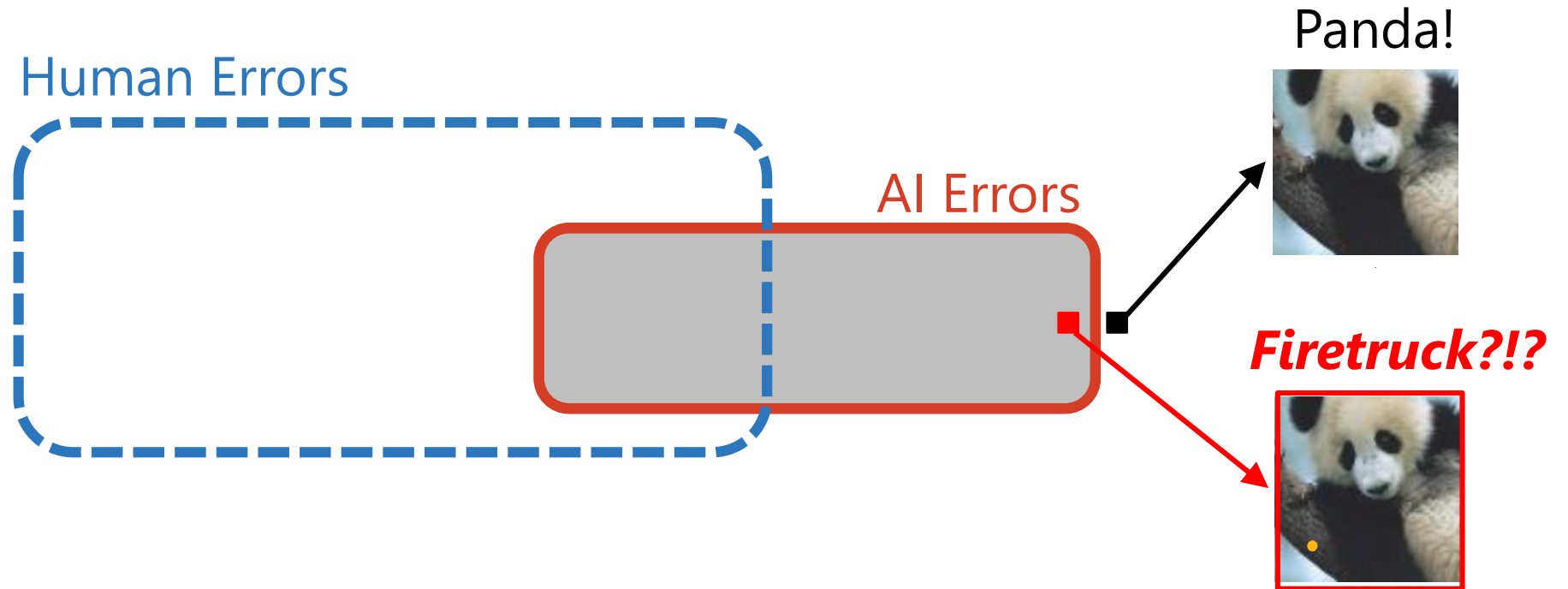
1. The AI May be Optimizing the Wrong Thing
2. Missing a Crucial Feature
3. Distributional Drift
- 4. Facilitating User Control in Mixed Human/AI Teams**
5. User Acceptance
6. Learning for Human Insight
7. Legal Requirements



The Growing Era of Human-AI Teams

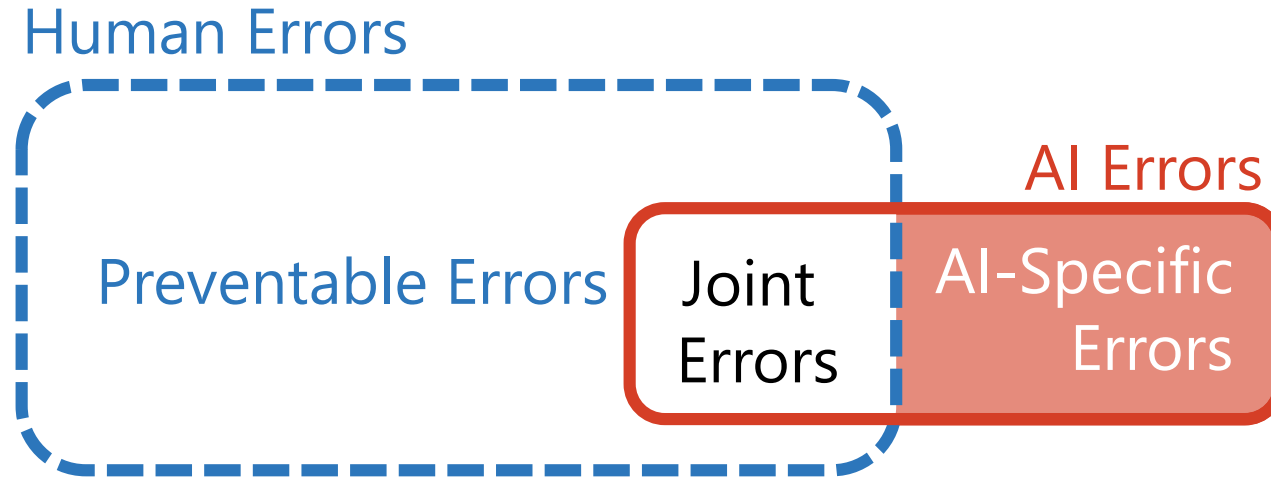


Artificial Intelligence Often Isn't

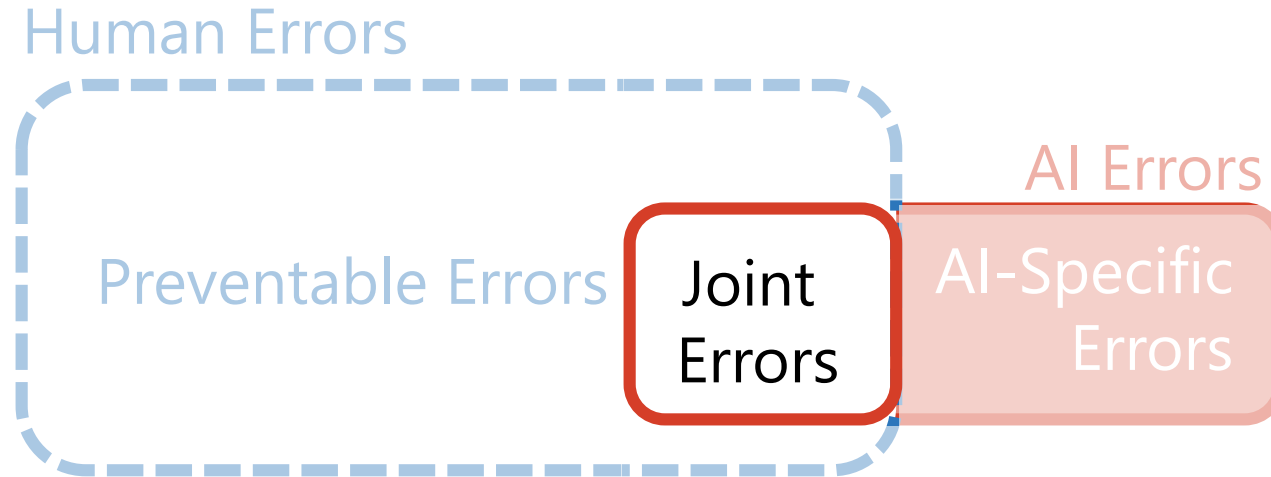


But Humans Err as Well

The Space of Errors

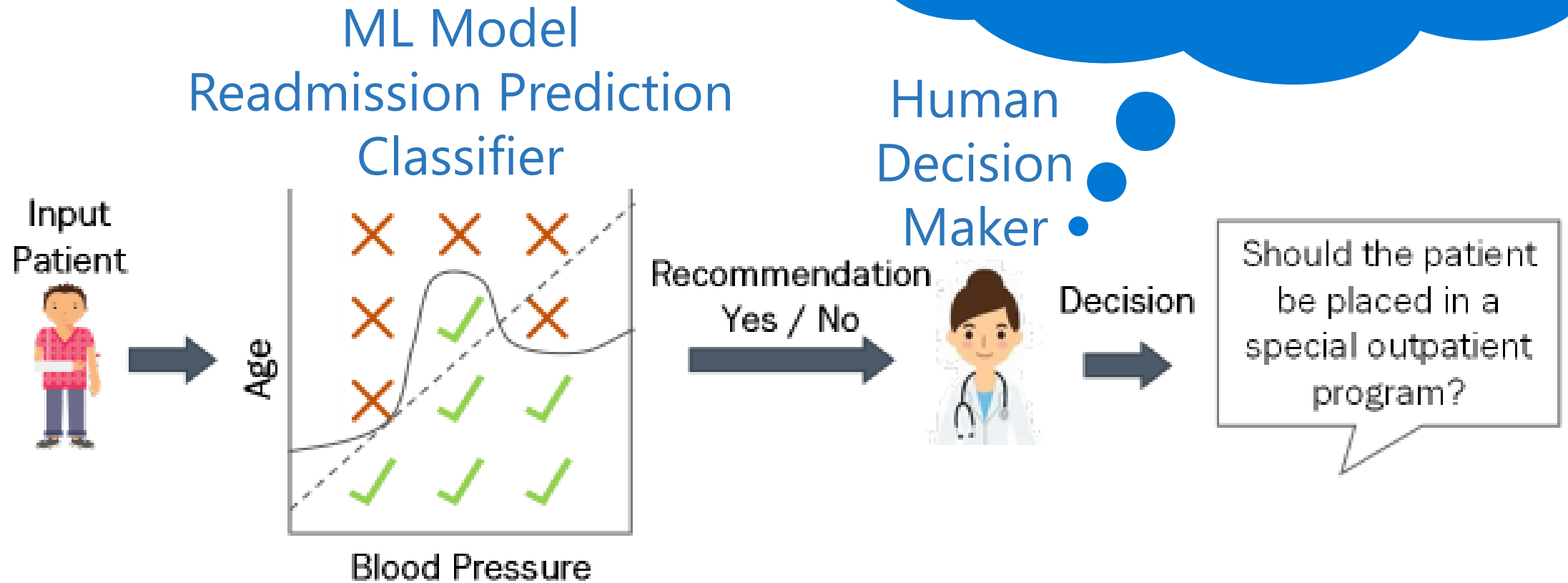


The Dream Team



Intelligible AI → Better Teamwork

A Simple Human-AI Team



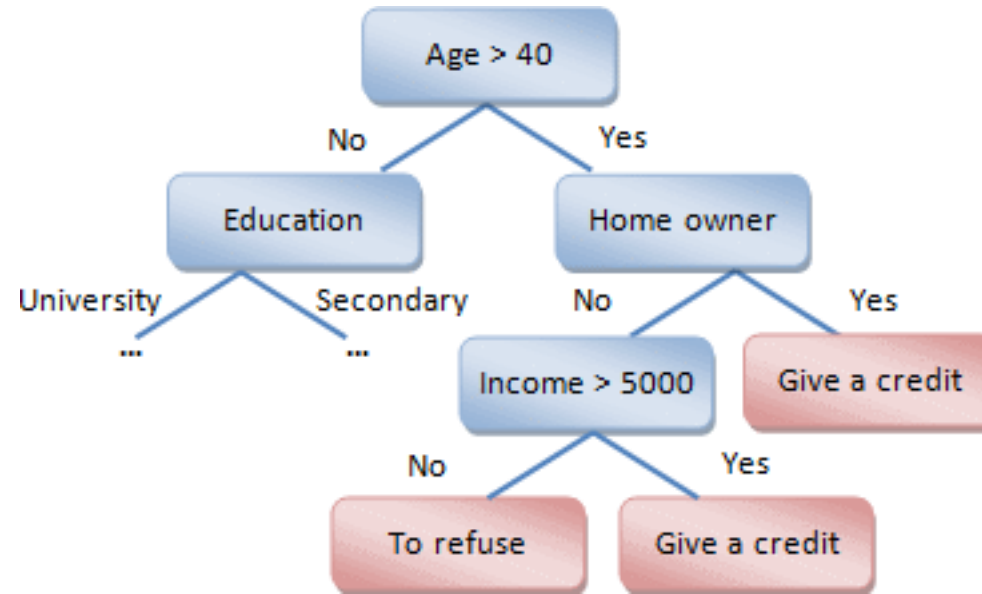
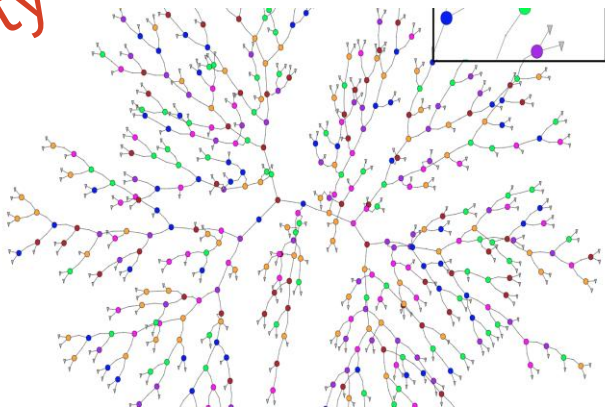
Inherently Intelligible ML – Example 1

When can I trust it?
How can I adjust it?

Small decision tree over semantically meaningful primitives



Intelligibility threatened if tree grows big



Inherently Intelligible ML – Example 2

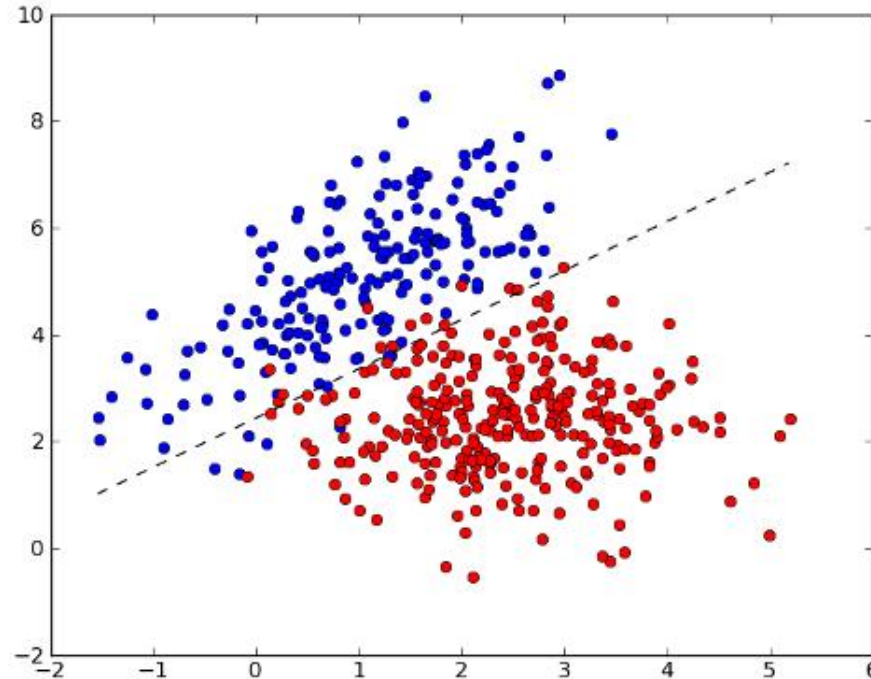
Linear model over
meaningful primitives

semantically

When can I trust it?
How can I adjust it?



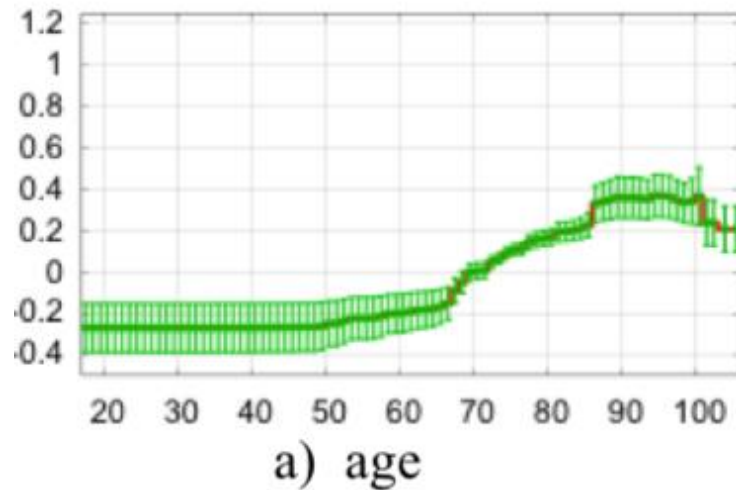
Other models often perform much better



Inherently Intelligible ML – Example 3

GA²M model over semantically meaningful primitives

$$y = \beta_0 + \sum_j f_j(x_j)$$



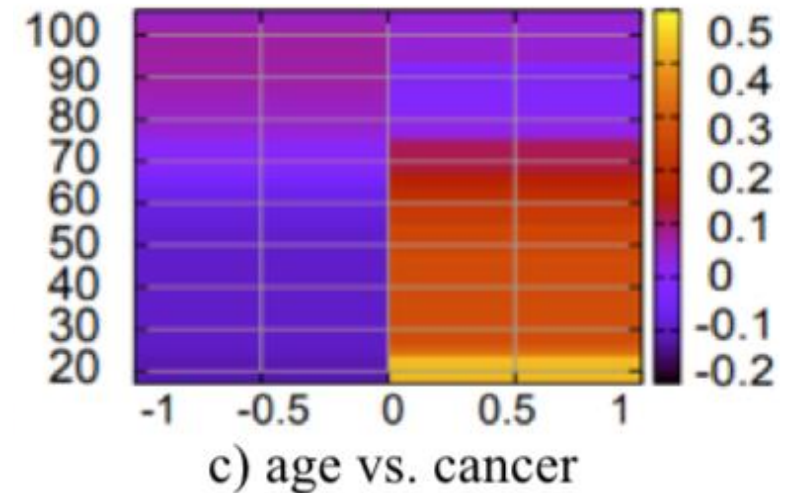
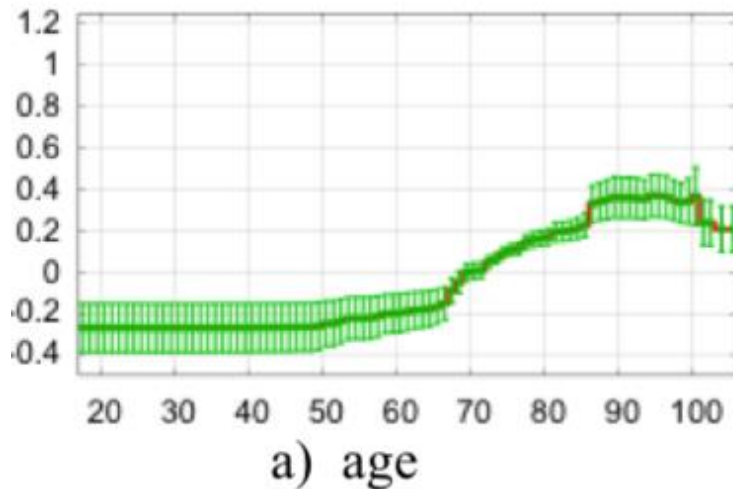
1 (of 56) components of learned GA²M: risk of pneumonia death

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In KDD 2015.

Inherently Intelligible ML – Example 3

GA²M model over semantically meaningful primitives

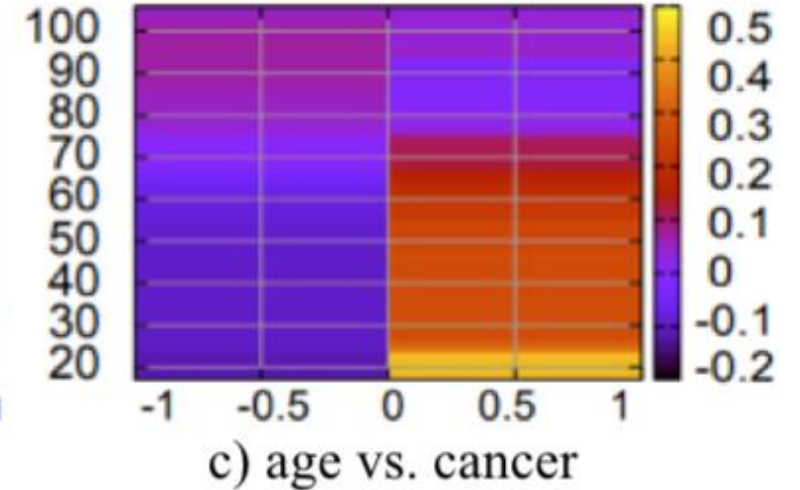
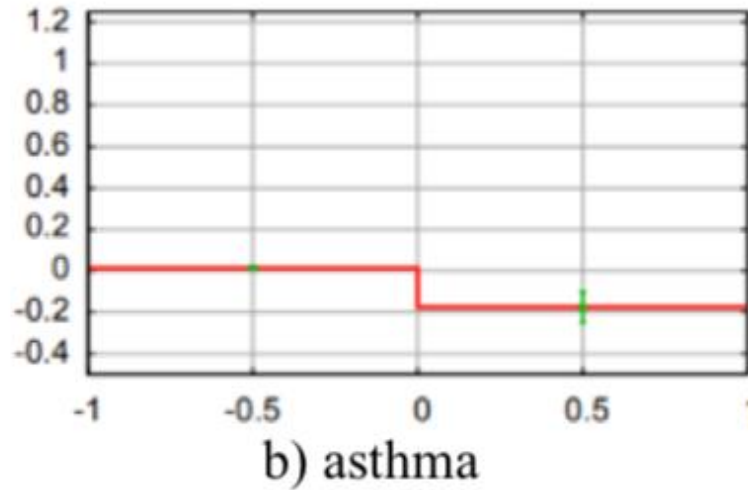
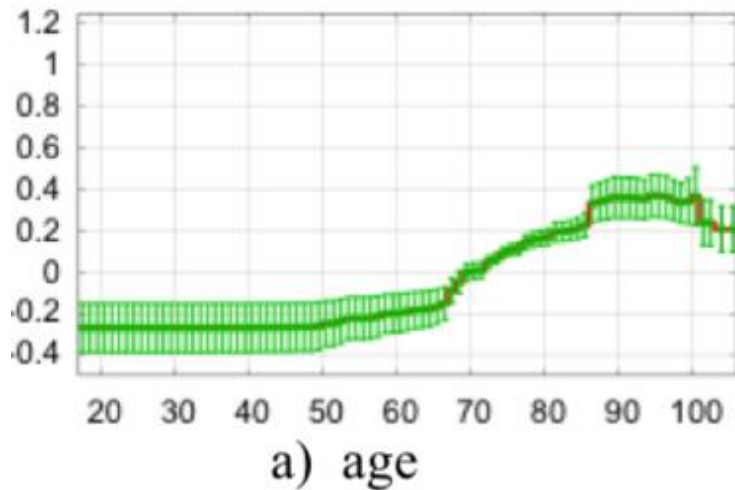
$$y = \beta_0 + \sum_j f_j(x_j) + \underbrace{\sum_{i \neq j} f_{ij}(x_i, x_j)}_{\text{pairwise terms}}$$



2 (of 56) components of learned GA²M: risk of pneumonia death

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In KDD 2015.

When can I trust it?
How can I adjust it?



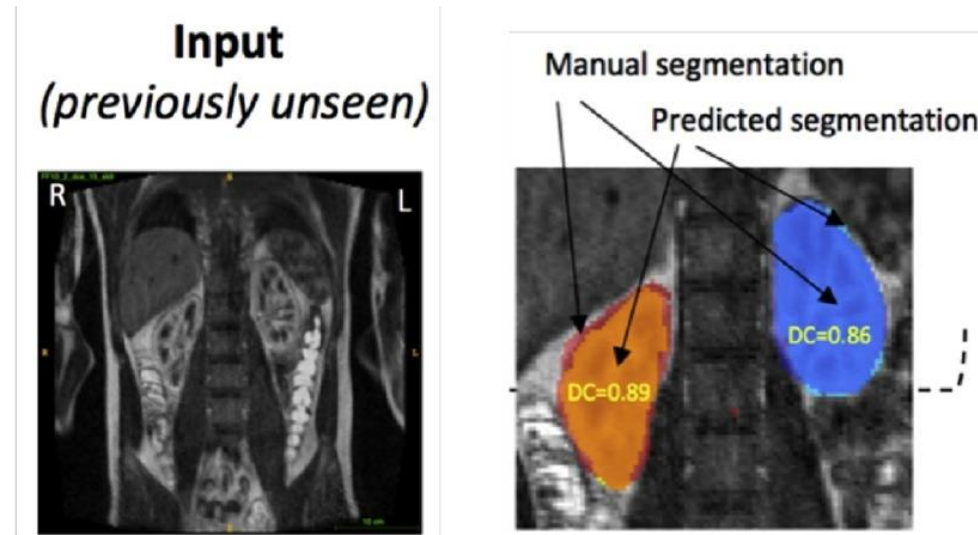
3 (of 56) components of learned GA²M: risk of pneumonia death

Part of Fig 1 from R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In KDD 2015.

Sometimes you just *need* an inscrutable model

E.g., Medical image analysis

- Deep cascade of CNNs
- Variational networks
- Transfer learning
- GANs



Input: Pixels

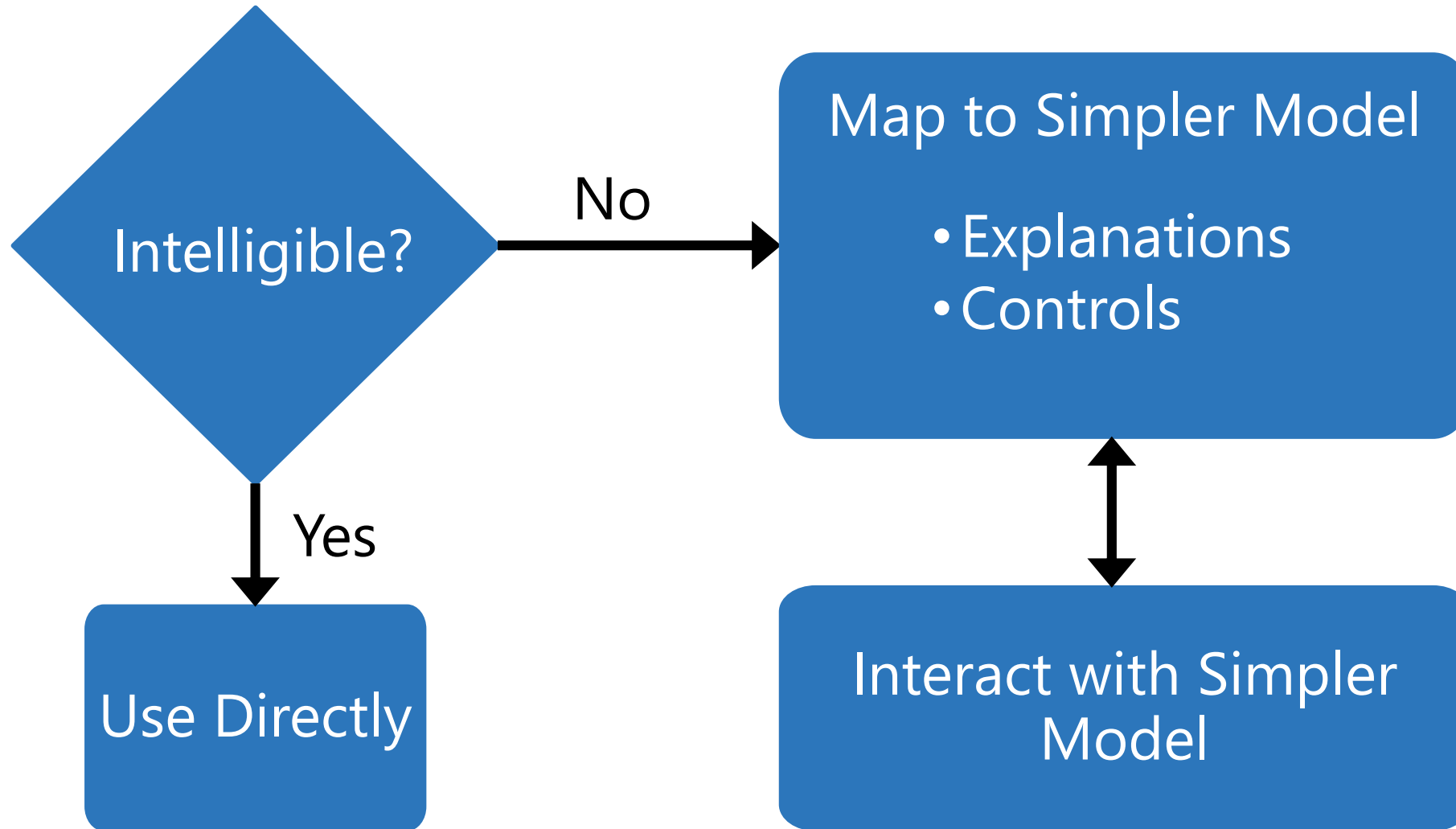
Features are not semantically meaningful

Kidney MRI

From [Lundervold & Lundervold 2018]

<https://www.sciencedirect.com/science/article/pii/S09393889183011>

Roadmap for Intelligibility



Reasons for Inscrutability



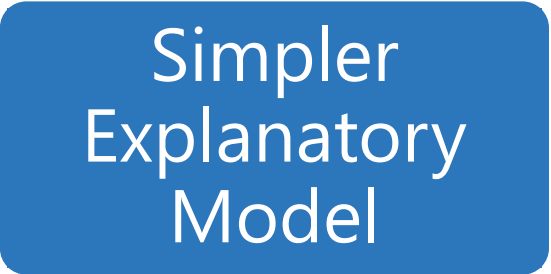
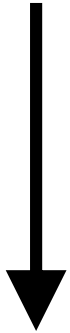
Inscrutable
Model

- Too Complex
- Features not Semantically Meaningful

Explaining Inscrutable Models



Inscrutable
Model



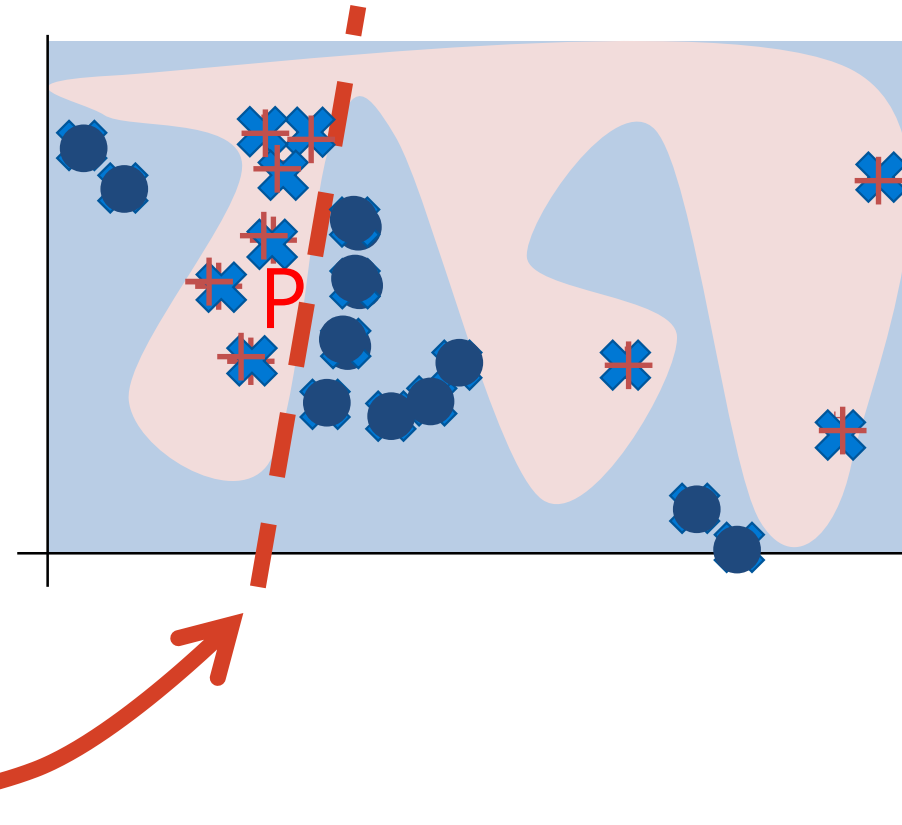
Simpler
Explanatory
Model

- Too Complex
 - Simplify by currying → instance-specific explanation
 - Simplify by approximating
- Features not Semantically Meaningful
 - Map to new vocabulary
- Usually have to do all of these!

LIME - Local Approximations

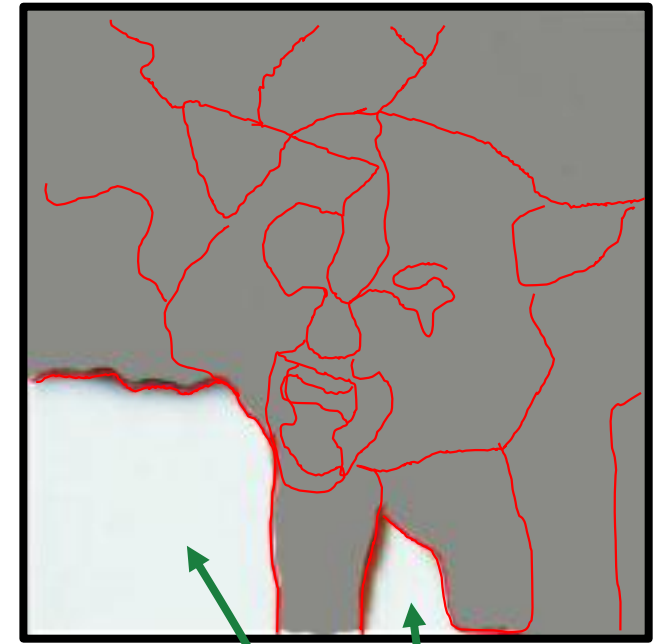
To explain prediction for point p ...

1. Sample points around p
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance from p
4. Learn new simple model on weighted samples
(possibly using different features)
5. Use simple model as explanation



Semantically Meaningful Vocabulary?

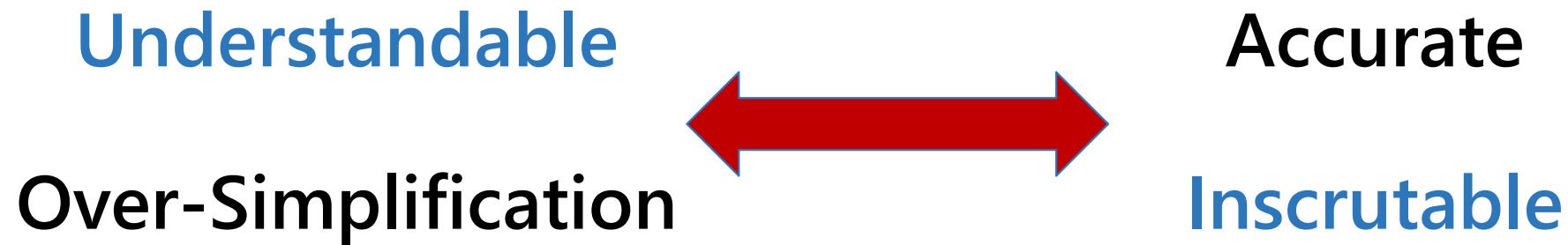
To create **features** for explanatory classifier,
Compute 'superpixels' using off-the-shelf image segmenter
Hope that feature/values are semantically meaningful



To **sample** points around p , set some superpixels to grey
Explanation is set of superpixels with high coefficients...

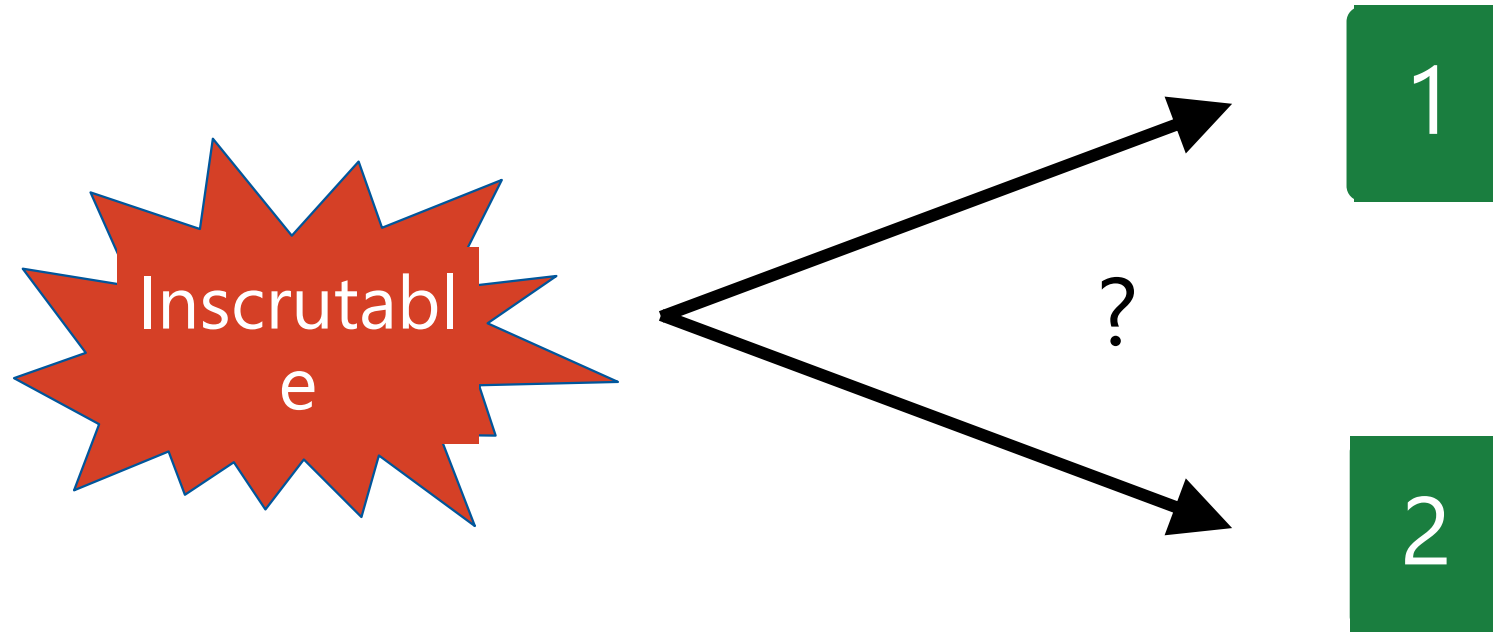
"It's just
looking
for
noise!"

Central Dilemma



Any model simplification is a
Lie

What Makes a Good Explanation?



Need Desiderata

Psychology Experiments → Ranking

If you can't include ***all*** details, humans prefer

- Details distinguishing fact & foil
- Necessary causes >> sufficient ones
- Intentional actions >> actions taken w/o deliberation
- Proximal causes >> distant ones
- Abnormal causes >> common ones
- Fewer conjuncts (regardless of probability)
- Explanations consistent with listener's prior beliefs

Tversky & Kahneman
Cognitive Biases

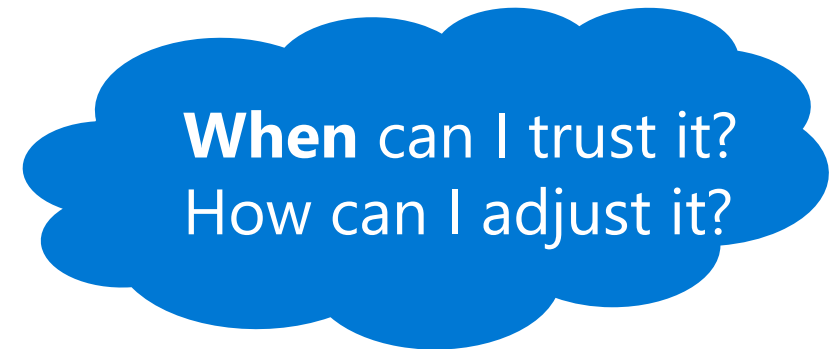
Presenting an explanation made people believe P was true
If explanation ~ previous, effect was strengthened

Trust

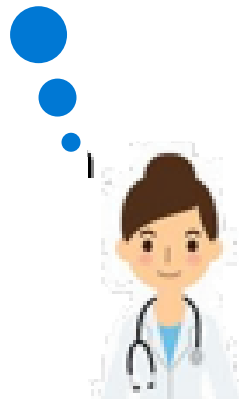


- Everybody talks about ***increasing trust...***
- The psychology literature shows explanations increase trust
[Miller AIJ-18]

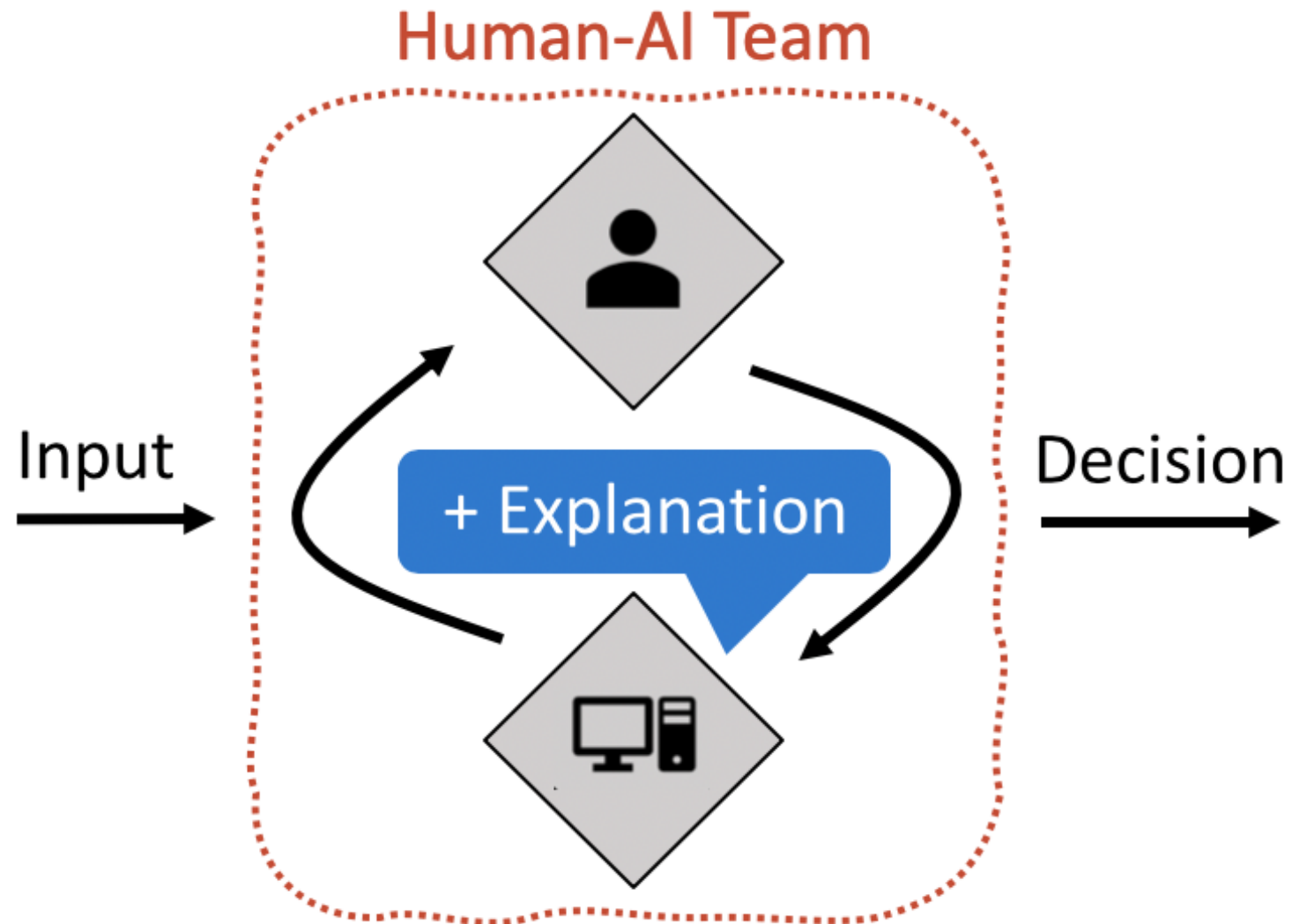
... Even when the explainer is ***wrong...***



- We ***shouldn't*** seek or measure trust...
- We should seek to show the human ***when not to trust***



Do Explanations Help *Team* Performance?



Yes!

- Medical Diagnosis

[Lundberg et al. *Nature biomedical engineering*. 2018]

- Annotation

[Schmidt & Biessmann. *AAAI Workshop* 2019]

- Deception Detection

[Lai & Tan FAT* 2019]

Except...

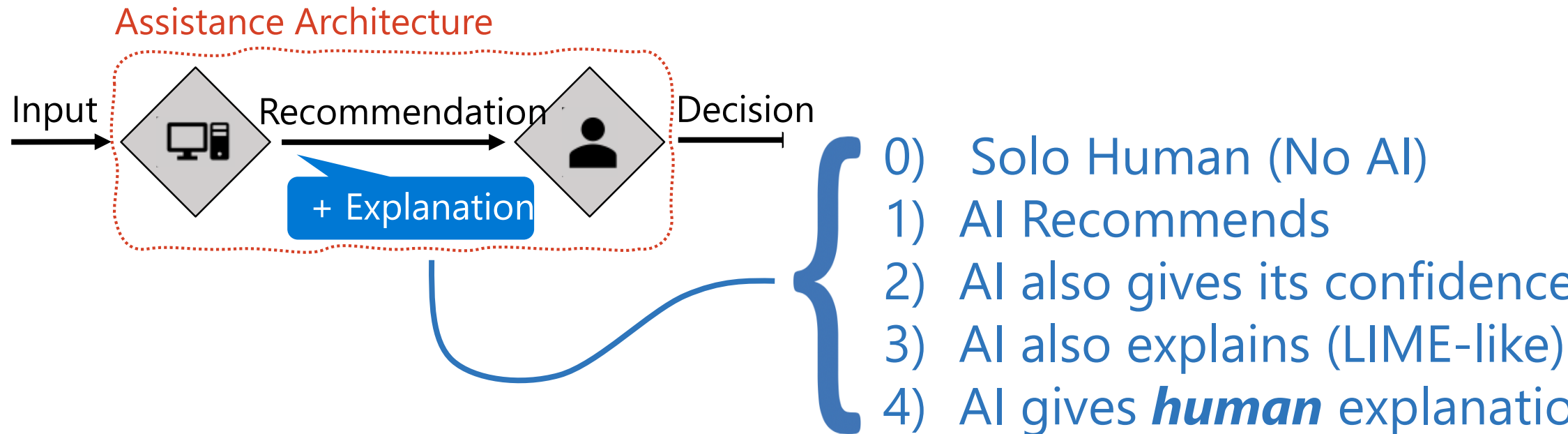
In these papers, $\text{Accuracy}(\text{Humans}) \ll \text{Accuracy}(\text{AI})$

So... the rational decision is to **omit** the humans (not explain)

Are Explanations Helpful??

We studied a simple human-AI team where

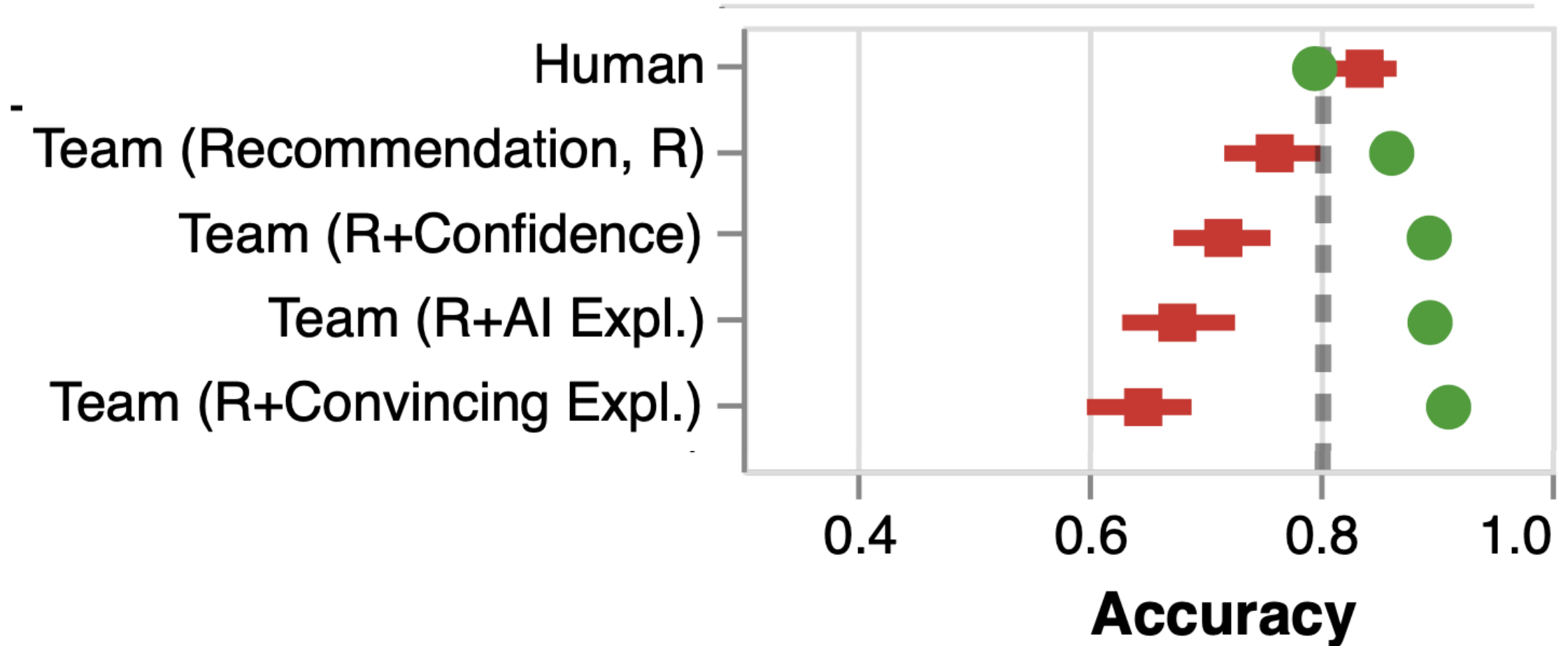
$$\text{Accuracy(Human)} = \text{Accuracy(AI)} = 0.8$$



Not Necessarily...

Explanations are Convincing

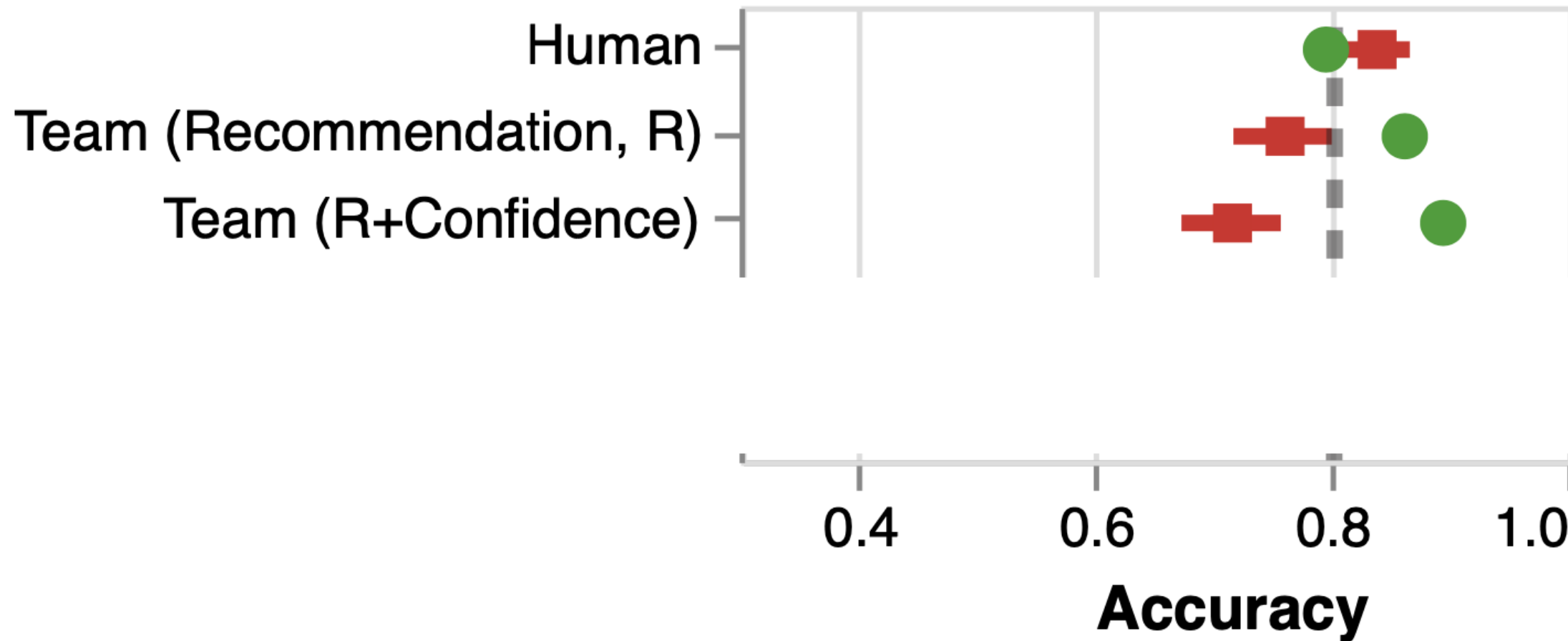
● AI Correct
■ AI Incorrect



Not Necessarily...

Explanations are Convincing

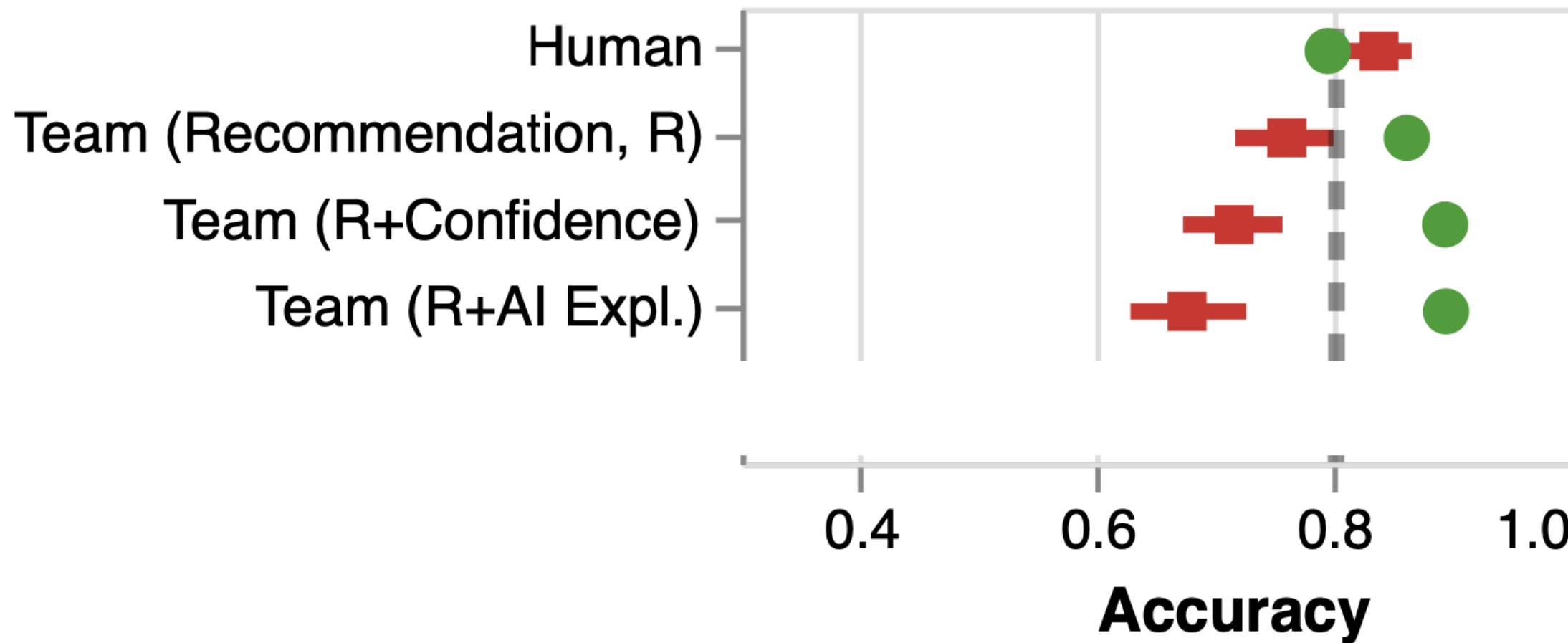
● AI Correct
■ AI Incorrect



Not Necessarily...

Explanations are Convincing

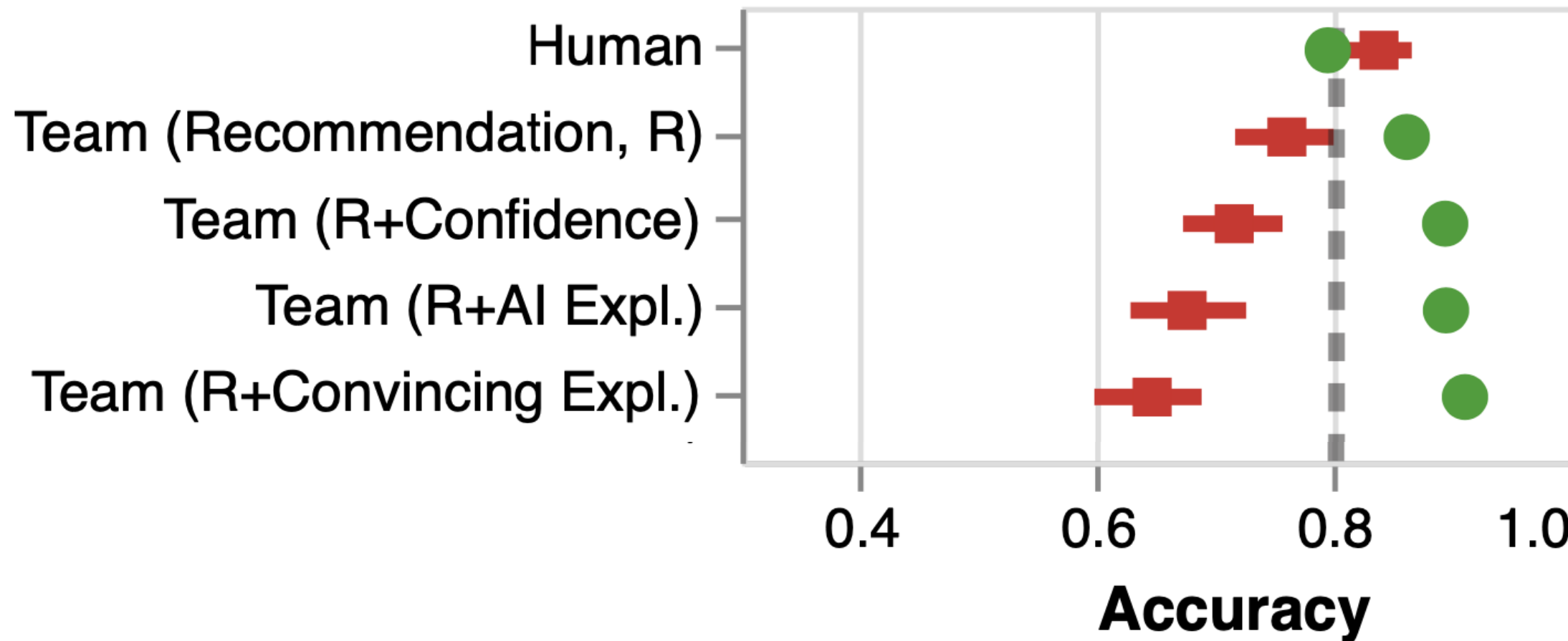
● AI Correct
■ AI Incorrect



Not Necessarily...

Explanations are Convincing

● AI Correct
■ AI Incorrect



Better Explanations are More Convincing

Coming Soon...

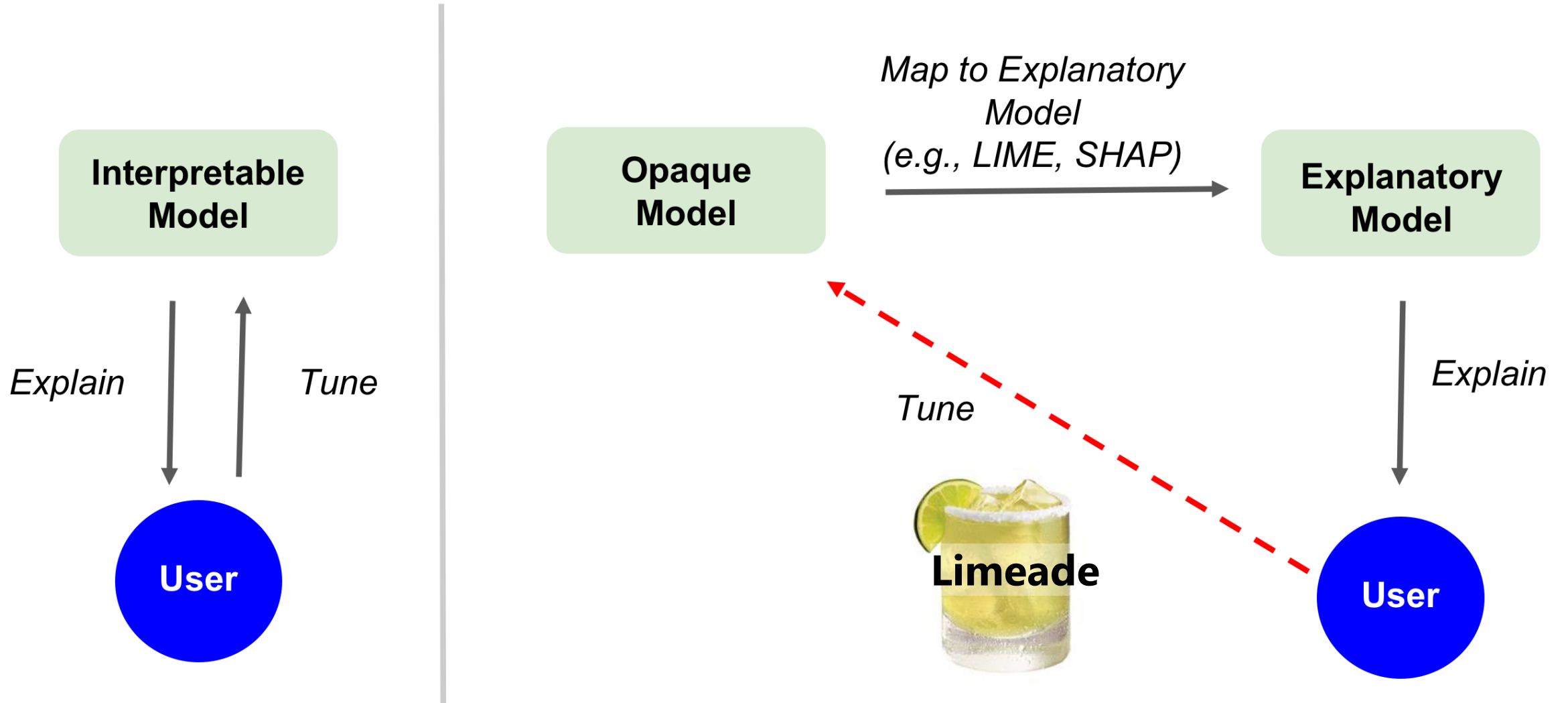
- Adaptive Explanations...

That Other Question...

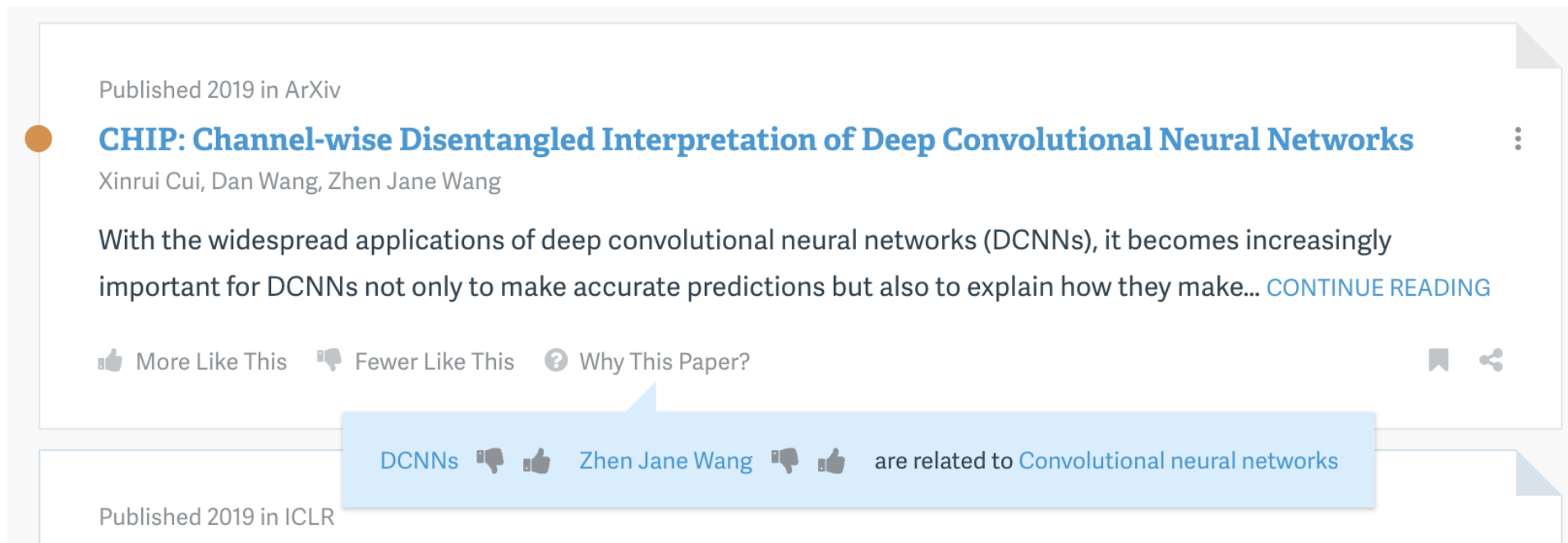
How can I adjust it?



Tuning

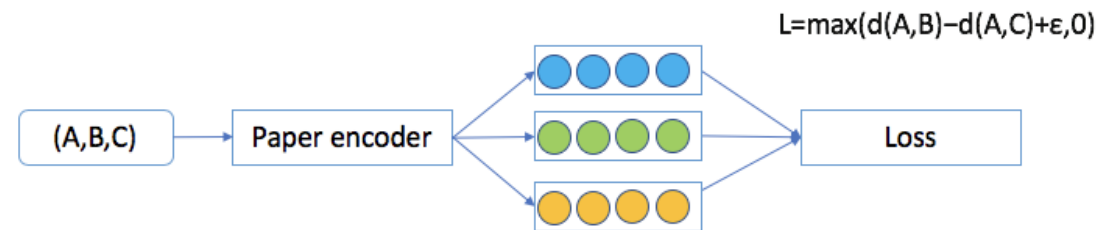


Adaptive Research-Paper Recommendations



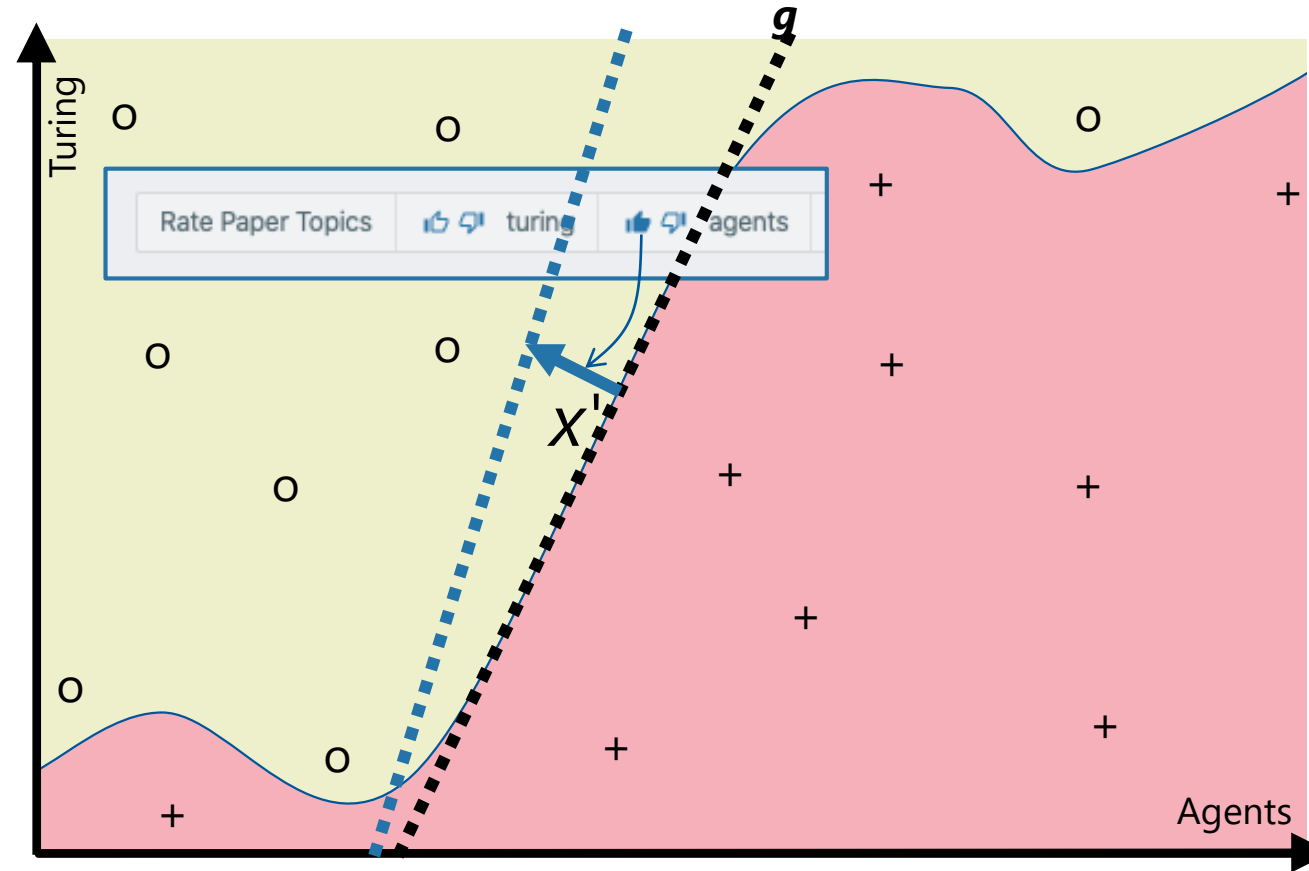
Beta: s2-sanity.apps.allenai.org

- Deep neural paper embeddings
- Explain with linear bigrams



[Cohen *et al.* Submitted]

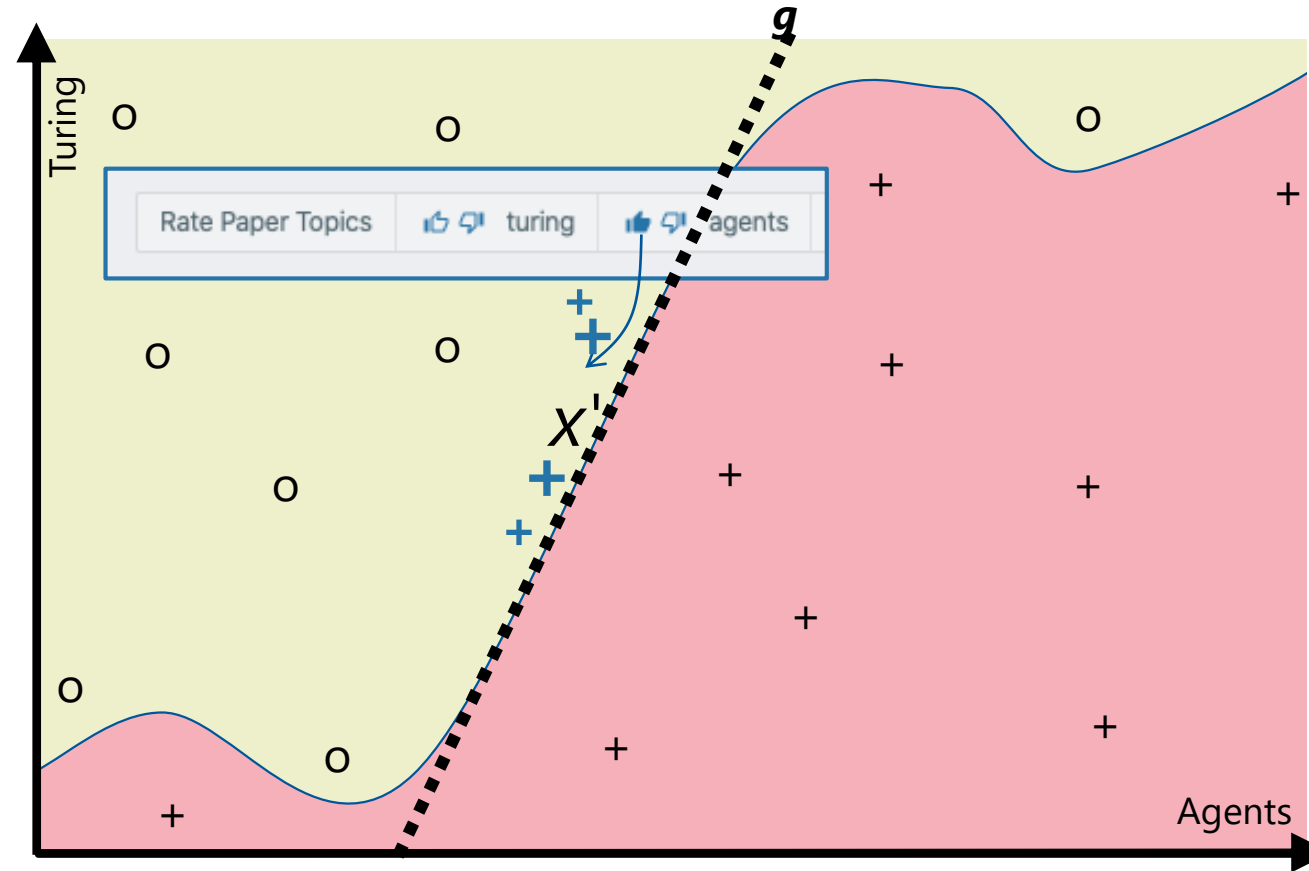
Tuning with Limeade



If all one cared about was the explanatory model,
one could change this parameters... but not even the
features are shared with the neural model!

[Lee *et al.* Submitted]

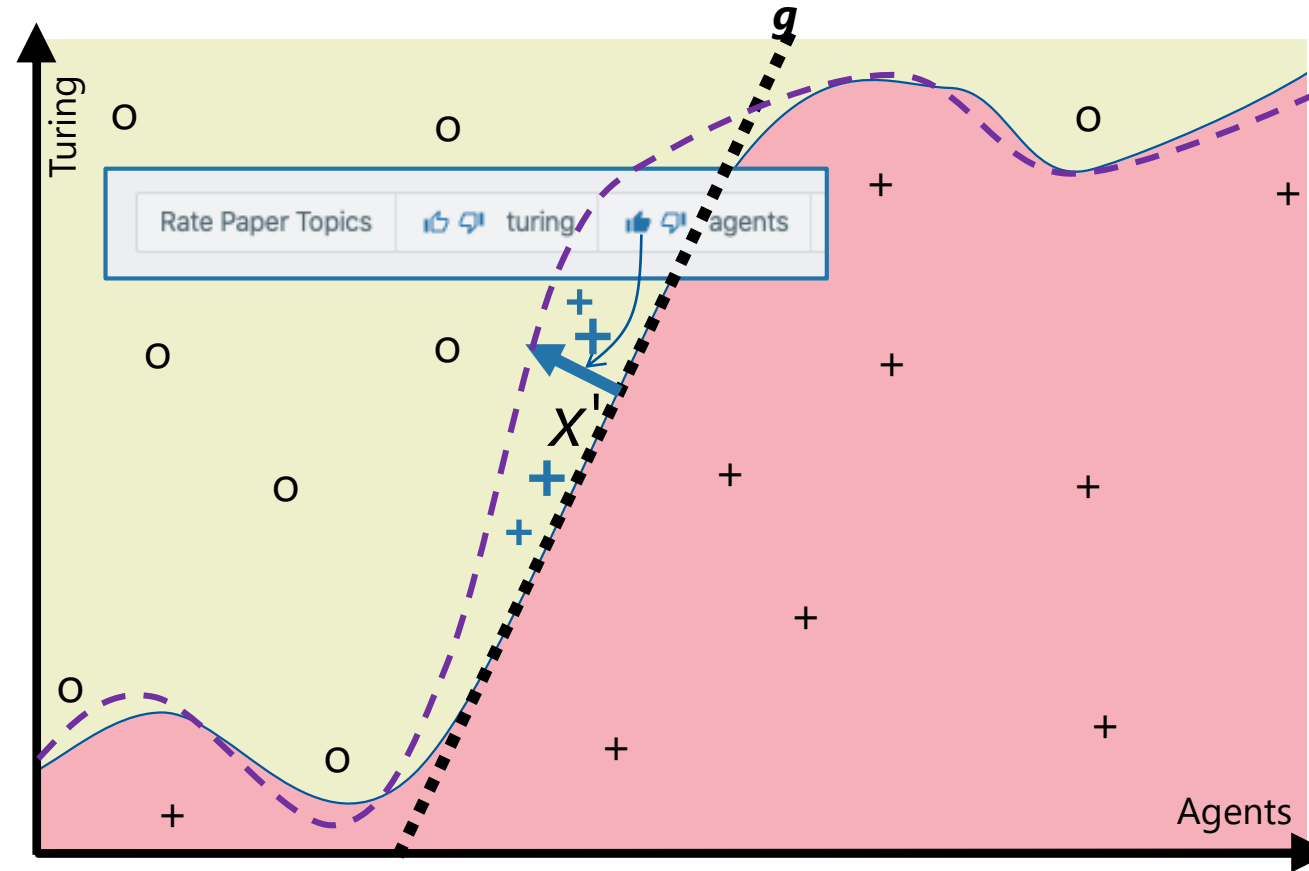
Tuning with Limeade



Instead... We generate new training instances by varying the feedback feature, weight by distance to x' ...

[Lee *et al.* Submitted]

Tuning with Limeade



Instead... We generate new training instances by varying the feedback feature, weight by distance to ~~and~~ **Retrain**.

[Lee *et al.* Submitted]

Evaluation

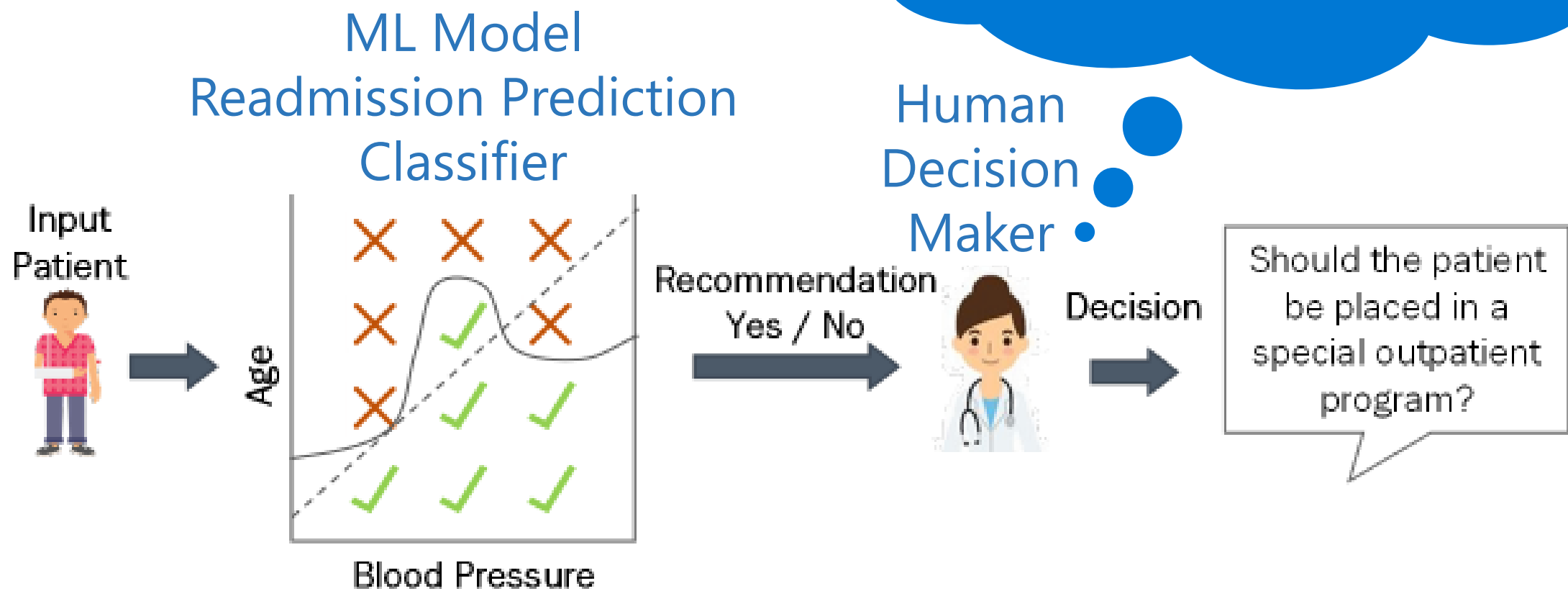
Good News:

Which system...	Baseline	Ours	<i>p</i> -value
...trust more?	4	17	0.043
...more control?	0	21	≈ 0
...more transparent?	3	18	0.012
...more intuitive?	12	9	0.664
...not missing relevant papers?	3	18	0.012

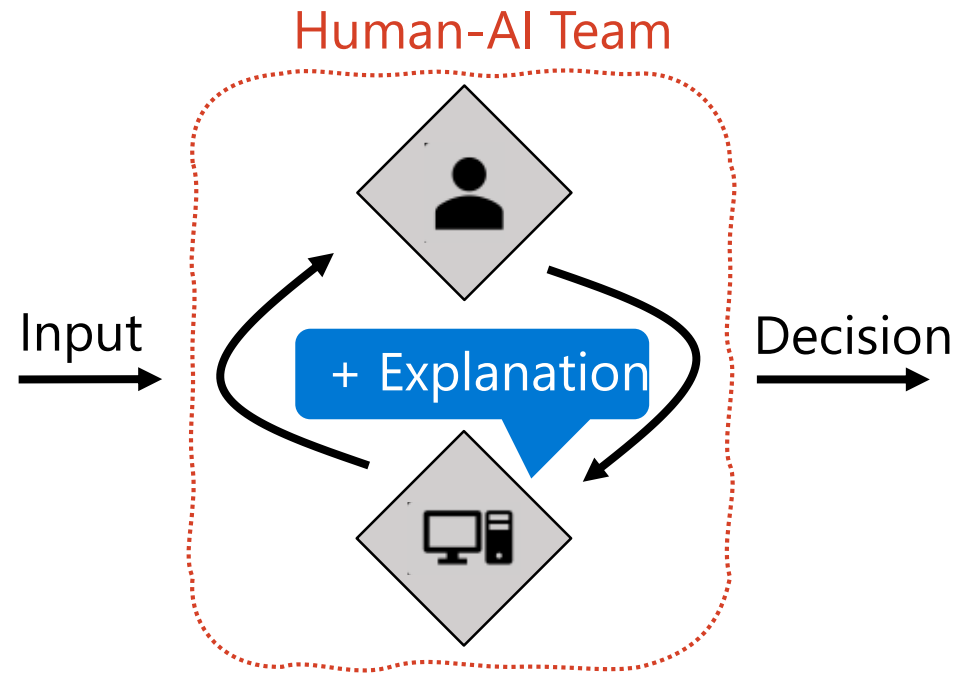
Less Good News:

No significant improvement on feed quality (team performance) as measured by clickthru

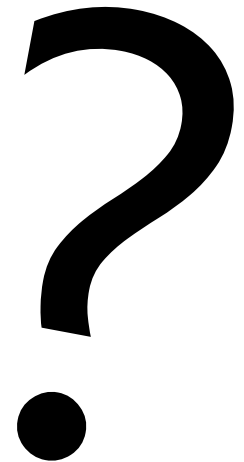
Summary



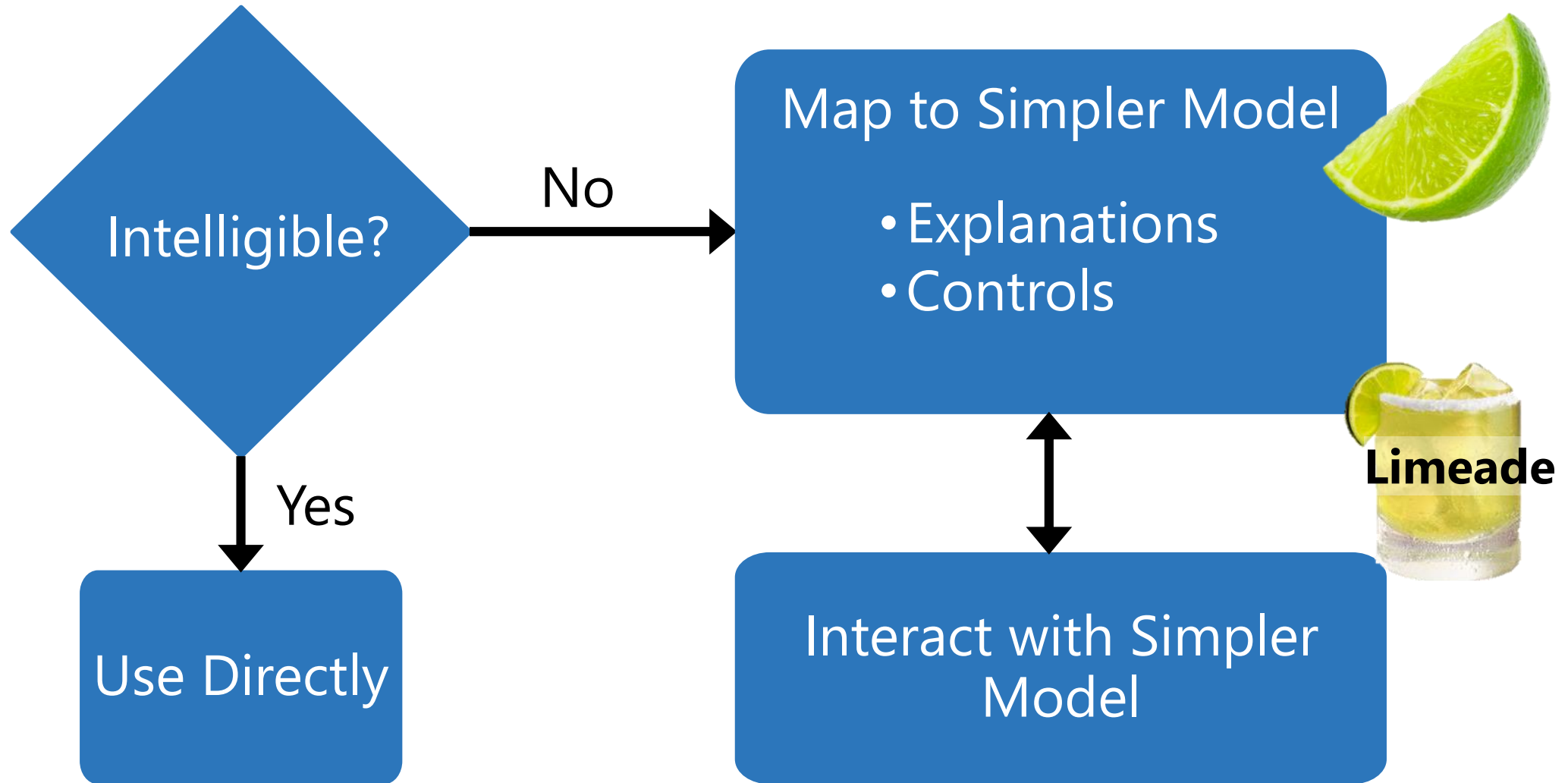
Summary



Helpful



Summary



Guidelines for Human AI Interaction

Learn more: <https://aka.ms/aiguideelines>



Thanks! Questions?

INITIALLY

1
Make clear
what the
system
can do.

2
Make clear
how well the
system can
do what it
can do.

DURING INTERACTION

3
Time services
based on
context.

4
Show
contextually
relevant
information.

5
Match
relevant
social norms.

6
Mitigate
social biases.

WHEN WRONG

7
Support
efficient
invocation.

8
Support
efficient
dismissal.

9
Support
efficient
correction.

10
Scope
services when
in doubt.

11
Make clear
why the
system did
what it did.

OVER TIME

12
Remember
recent
interactions.

13
Learn from
user behavior.

14
Update and
adapt
cautiously.

15
Encourage
granular
feedback.

16
Convey the
consequences
of user
actions.

17
Provide
global
controls.

18
Notify users
about
changes.

Resources

Tutorial website: <https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/articles/aaai-2020-tutorial-guidelines-for-human-ai-interaction/>

Learn the guidelines

[Introduction to guidelines for human-AI interaction](#)

[Interactive cards with examples of the guidelines in practice](#)

Use the guidelines in your work

[Printable cards \(PDF\)](#)

[Printable poster \(PDF\)](#)

Find out more

[Guidelines for human-AI interaction design](#), Microsoft Research Blog

[AI guidelines in the creative process: How we're putting the human-AI guidelines into practice at Microsoft](#), Microsoft Design on Medium

[How to build effective human-AI interaction: Considerations for machine learning and software engineering](#), Microsoft Research Blog