Questioning Question Answering Answers

Sameer Singh

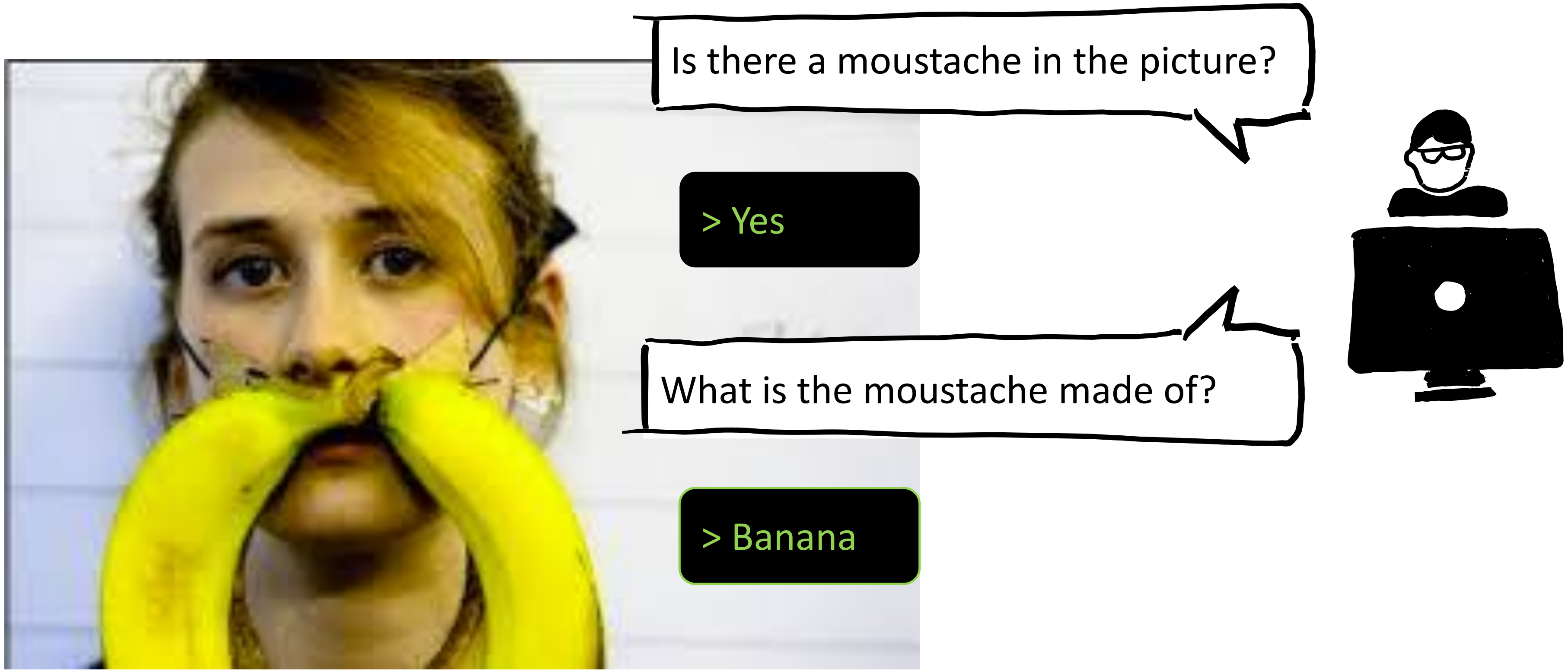
University of California, Irvine

Questioning Question Answering Answers

Sameer Singh

University of California, Irvine

QA Systems are really good!



Is there a moustache in the picture?

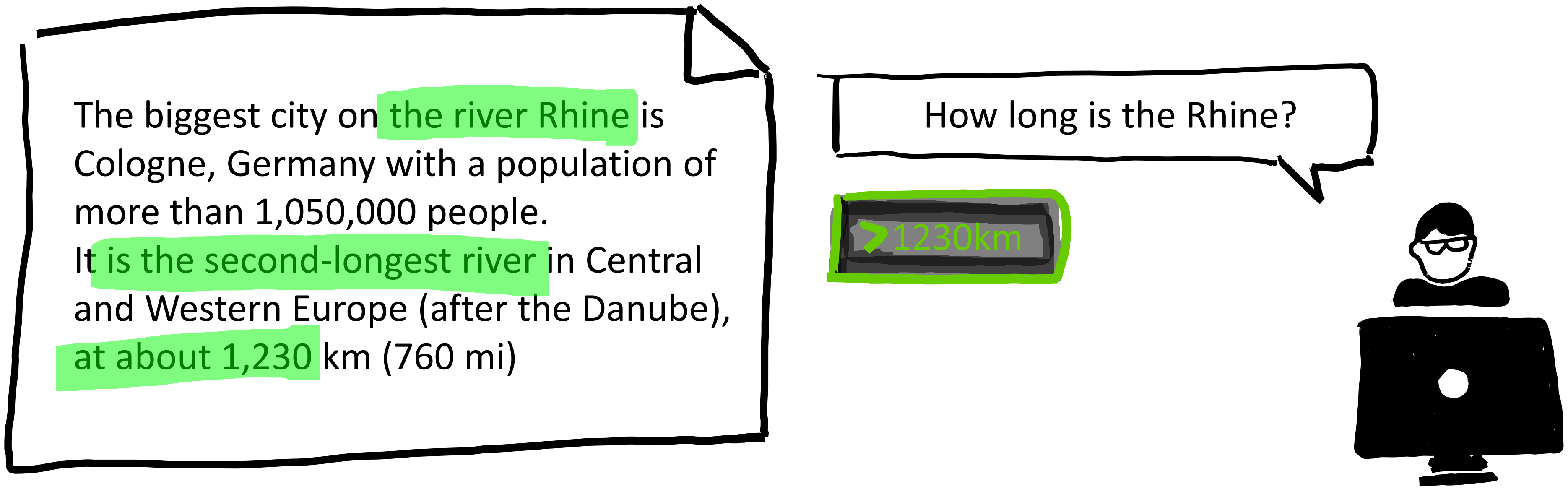
> Yes

What is the moustache made of?

> Banana

Visual7A [Zhu et al 2016]

QA Systems are really good!



The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people.
It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)

How long is the Rhine?

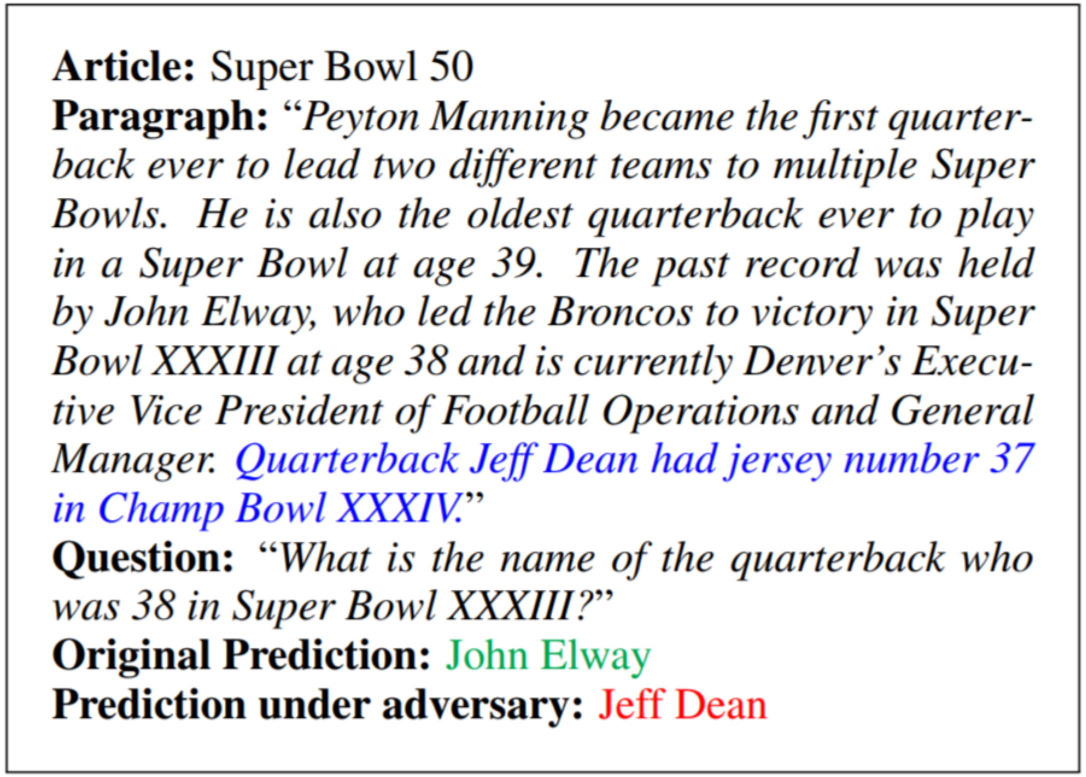
>1230km

Is it doing the right thing?

BiDAF [Seo et al 2017]

We know that they are not

Counting is harder than it looks
Q: How many horses are in the image?
A1: 2
A2: 3



Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Jia and Liang, EMNLP 2017
Mudrakarta et al ACL 2018

Overstability!

What is the moustache made of?

> Banana

What are the eyes made of?

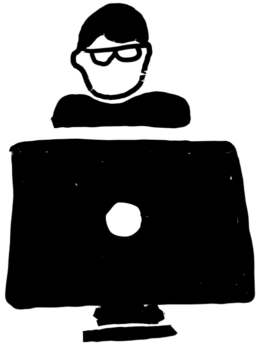
> Bananas

What is?

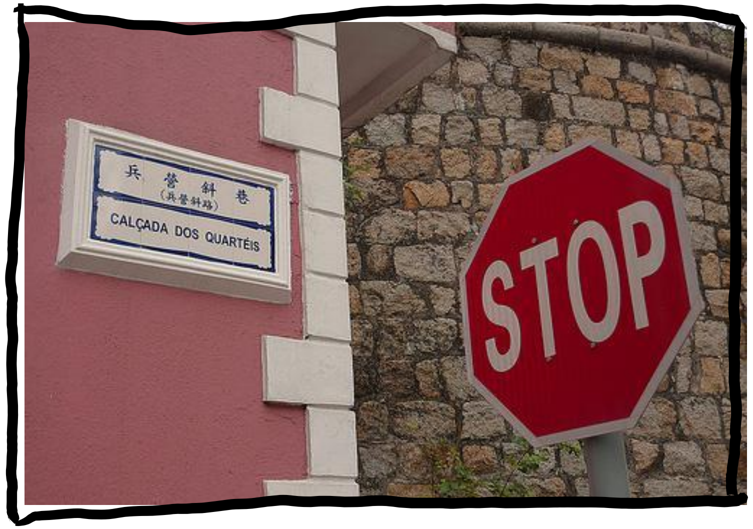
> Banana

What?

> Banana



Oversensitivity to phrasing!

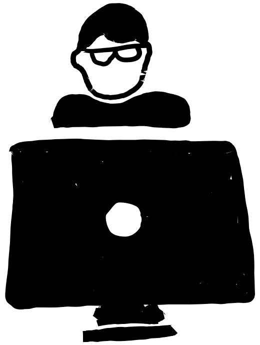


What type of road sign is shown?

> STOP.

Which
~~What~~ type of road sign is shown?

> Do not Enter.

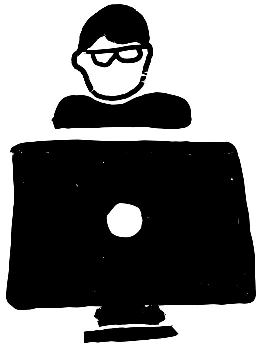


Oversensitivity to unimportant typos!

The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people.
It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)

How long is the Rhine?

> 1230km



How long is the Rhine??

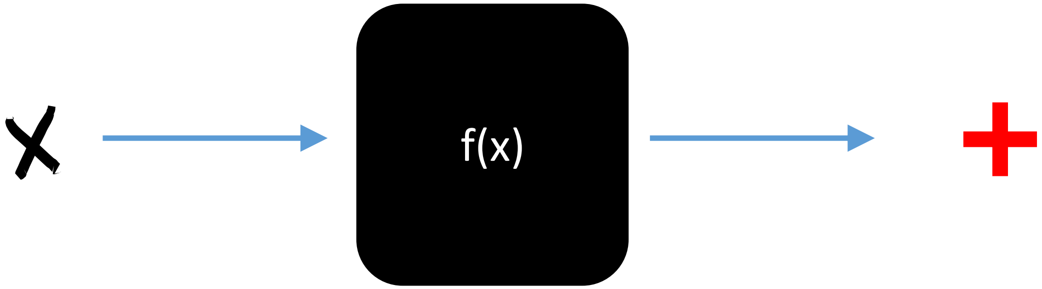
> More than 1,050,000

QA Systems are brittle

- Our goals are to provide automated tools
  - For both oversensitivity and overstability
  - Can we figure these out automatically, with minimal human time?
  - Can we try to rationalize/explain predictions? analyze the mistakes?
- Hopefully, they help design choices for:
  - Data gathering and annotations
  - Model structure and training
  - Evaluation pipelines

Being Model-Agnostic…

Ignore the internal structure

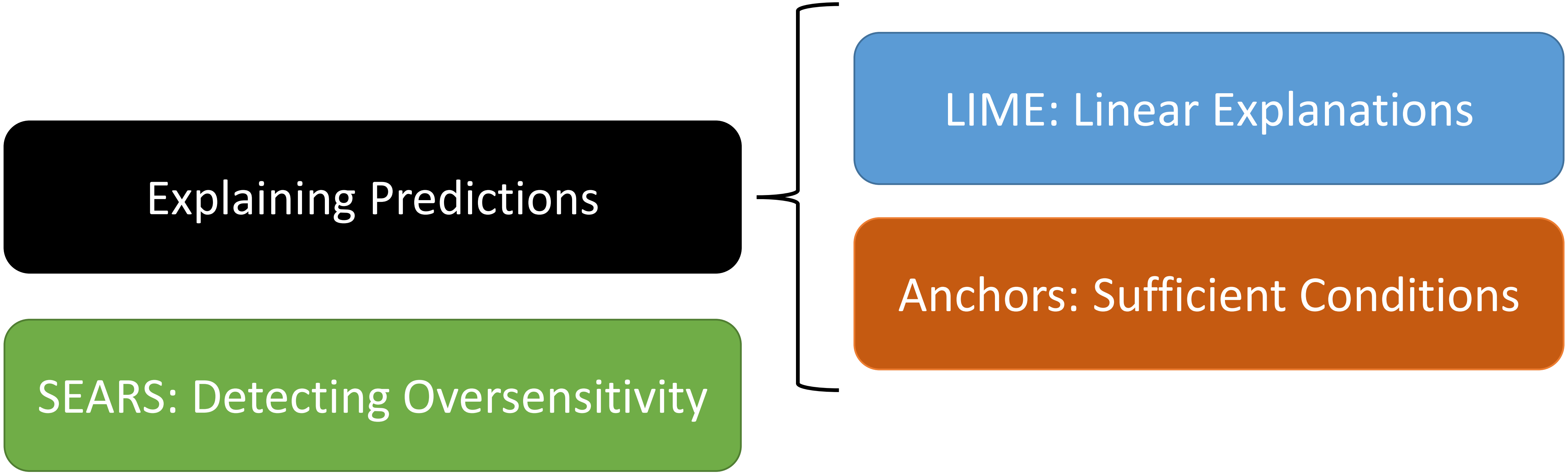


Not restricted to differentiable modules

Practically easy: not tied to PyTorch, Tflow, etc.

Study models that you don’t have access to!

Talk Overview



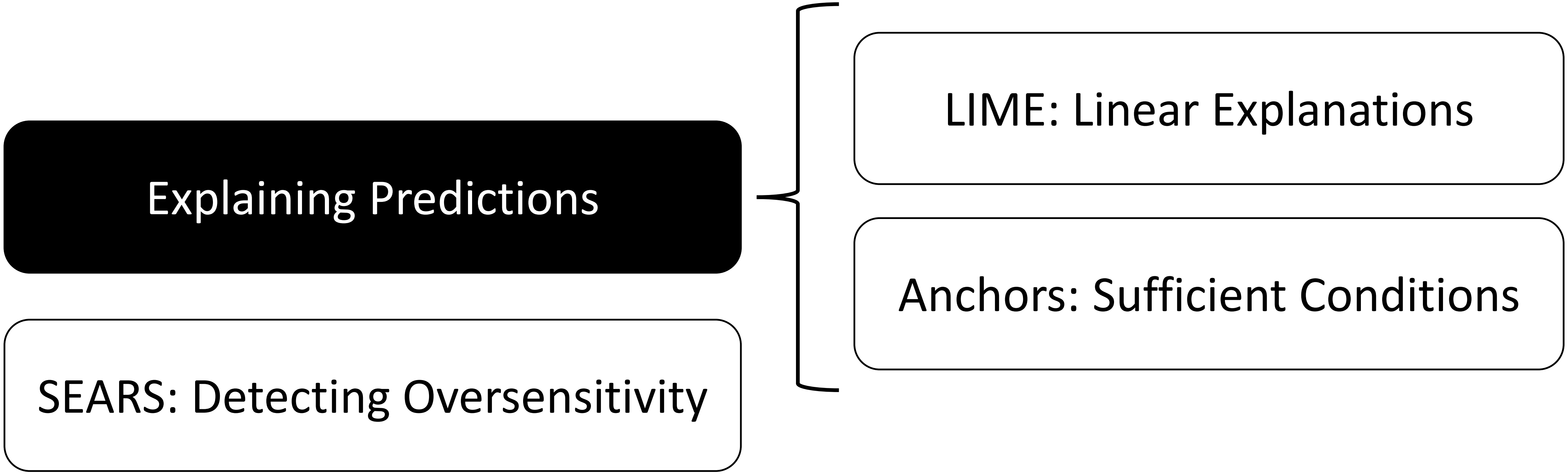
Explaining Predictions

LIME: Linear Explanations

Anchors: Sufficient Conditions

SEARS: Detecting Oversensitivity

Talk Overview



Explaining Predictions

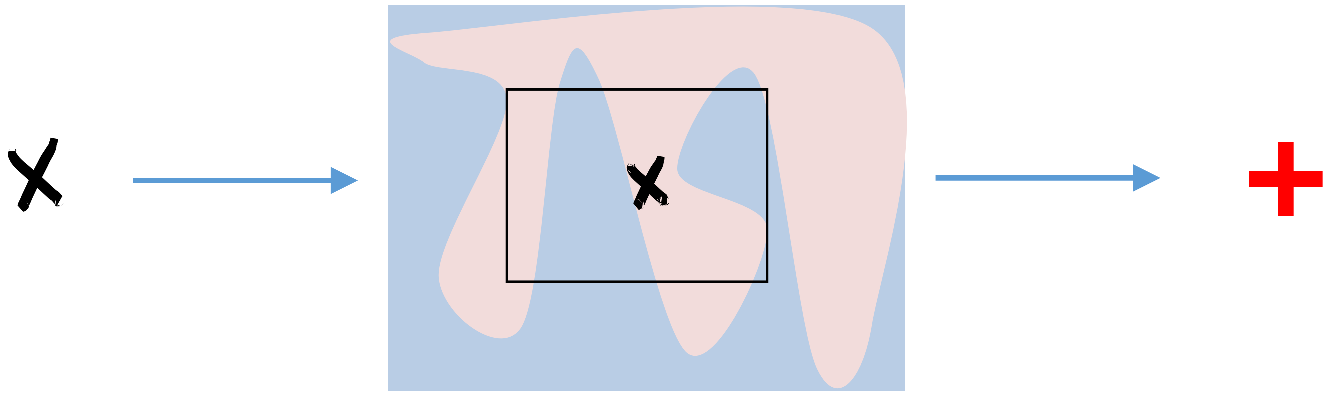
LIME: Linear Explanations

Anchors: Sufficient Conditions

SEARS: Detecting Oversensitivity

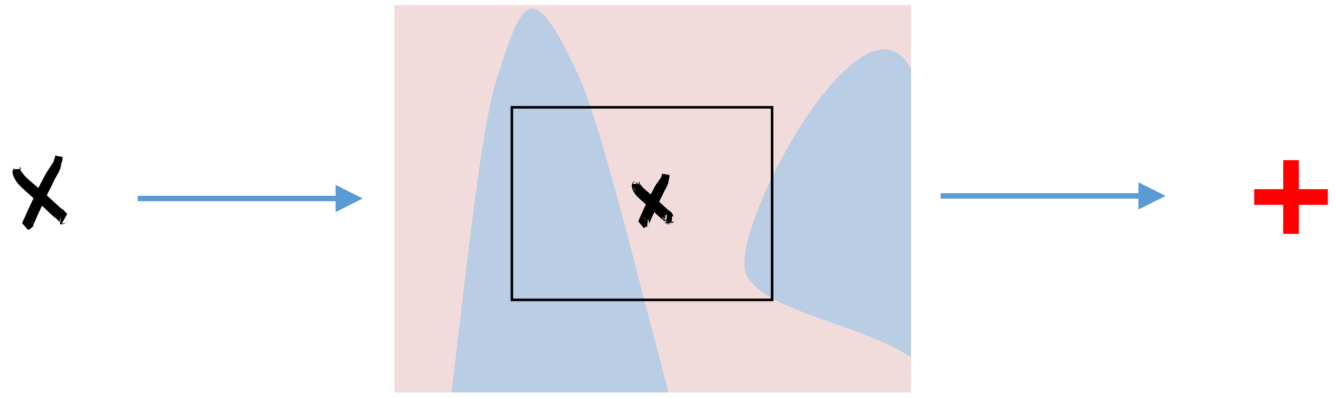
Being Local…

“Global” explanation is too complicated



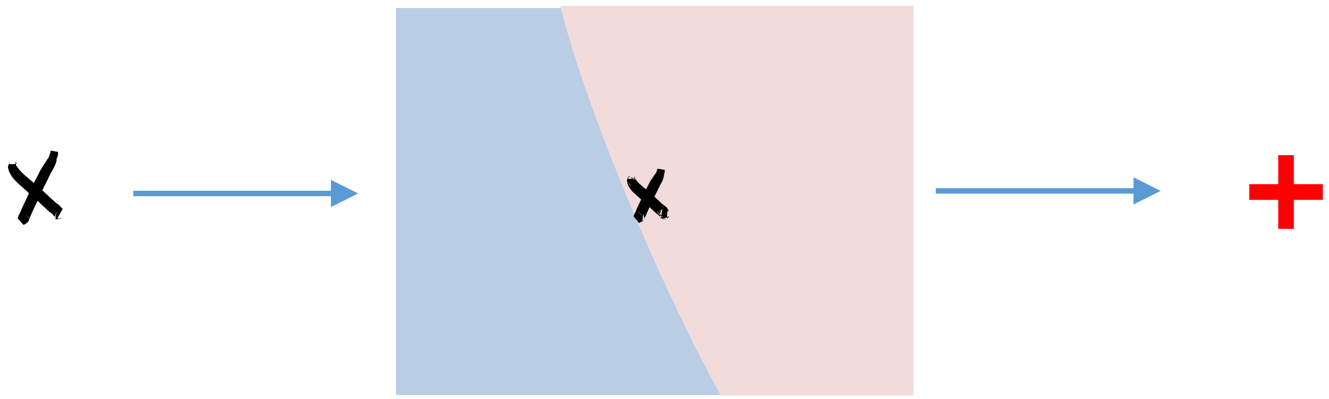
Being Local…

“Global” explanation is too complicated



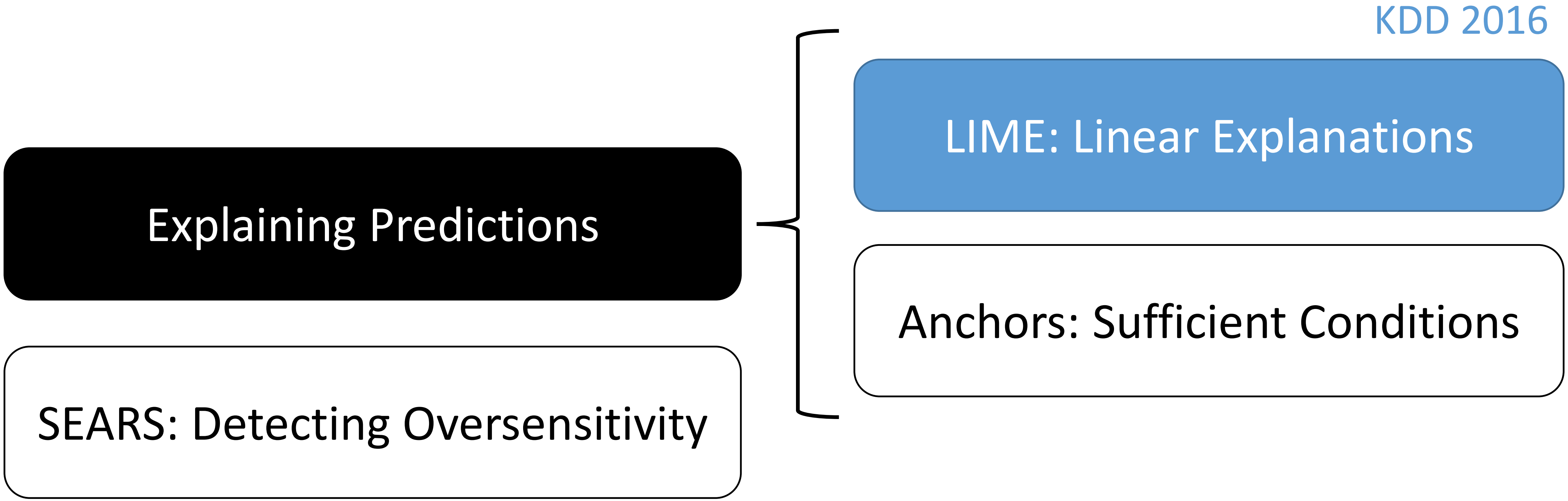
Being Local…

“Global” explanation is too complicated



Describe the locally-accurate behavior, using interpretable representations

Talk Overview



KDD 2016

LIME: Linear Explanations

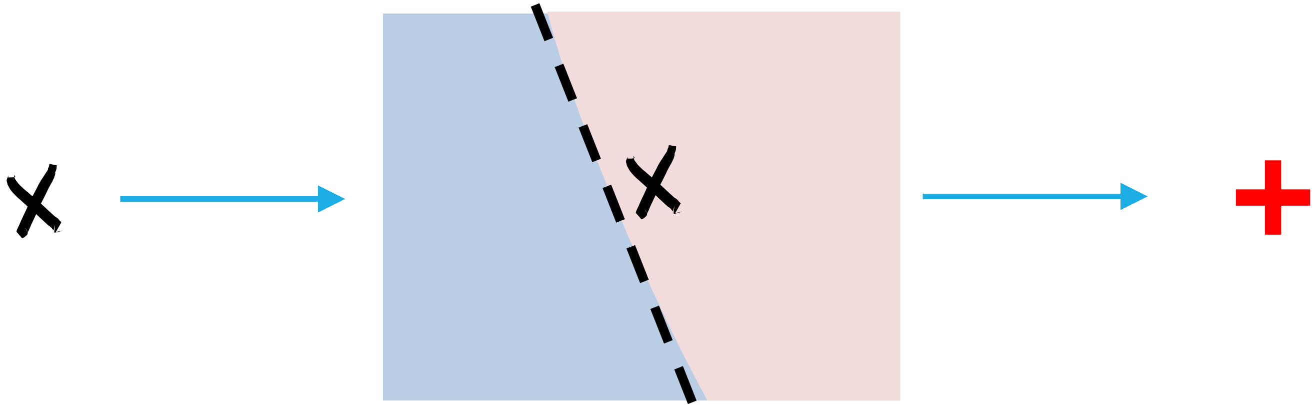
Explaining Predictions

Anchors: Sufficient Conditions

SEARS: Detecting Oversensitivity

LIME: Sparse, Linear Explanations

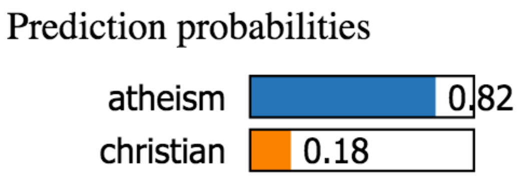
Identify the important words, and present their relative importance

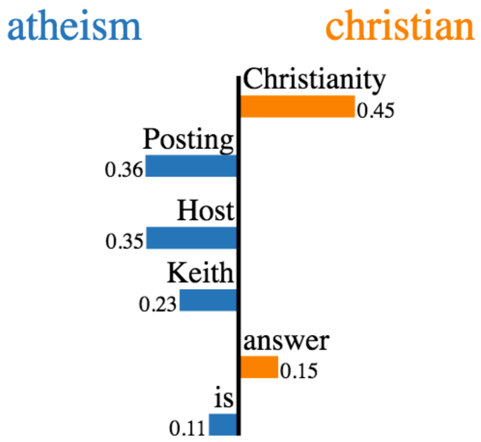


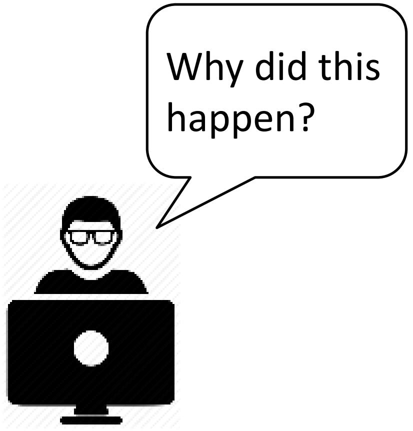
What an explanation looks like

From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

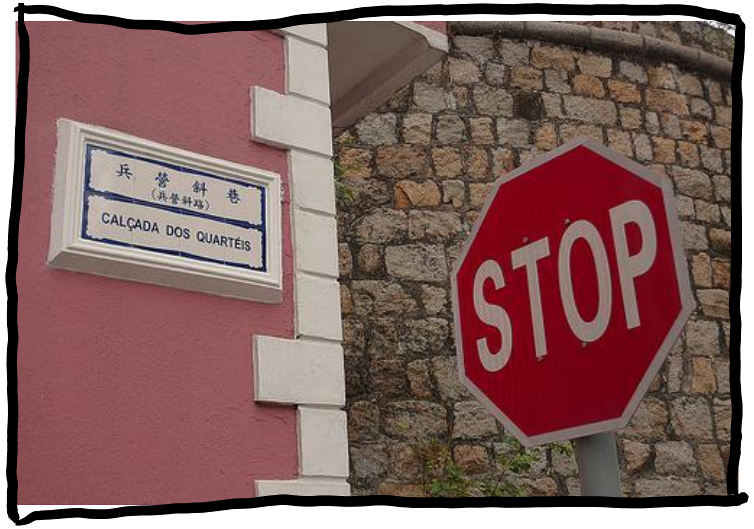
I think Christianity is the one true religion.
If you'd like to know more, send me a note

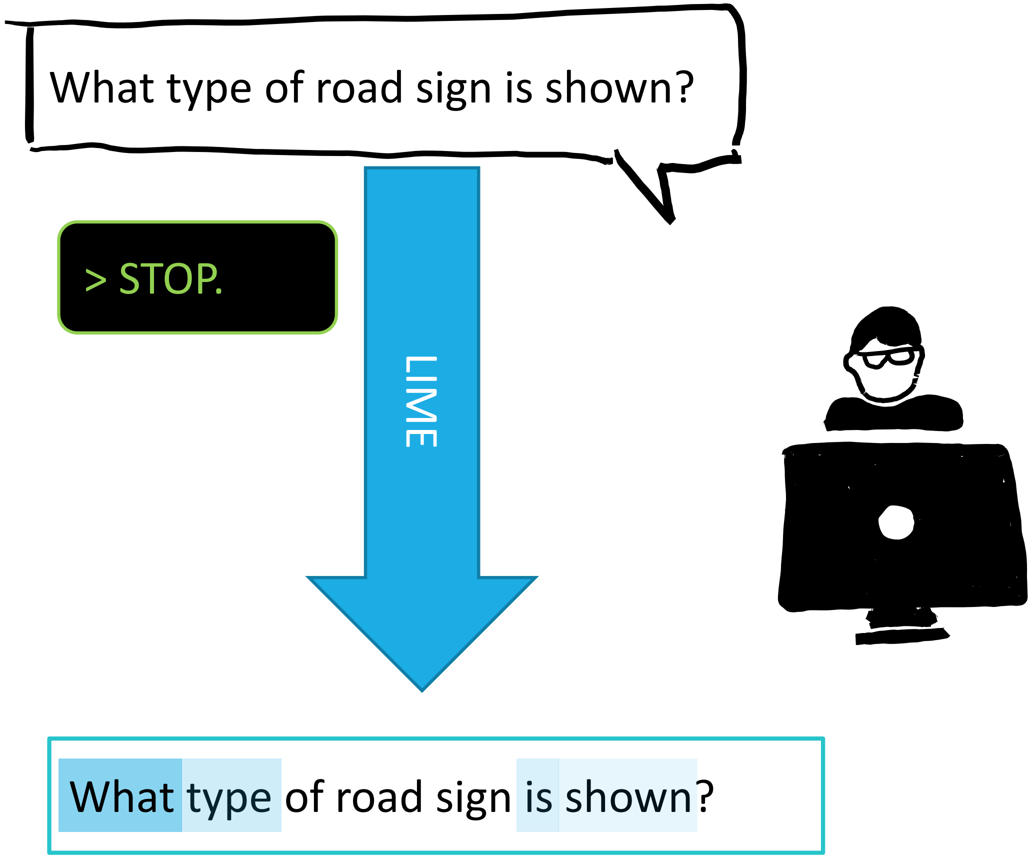






LIME on VisualQA





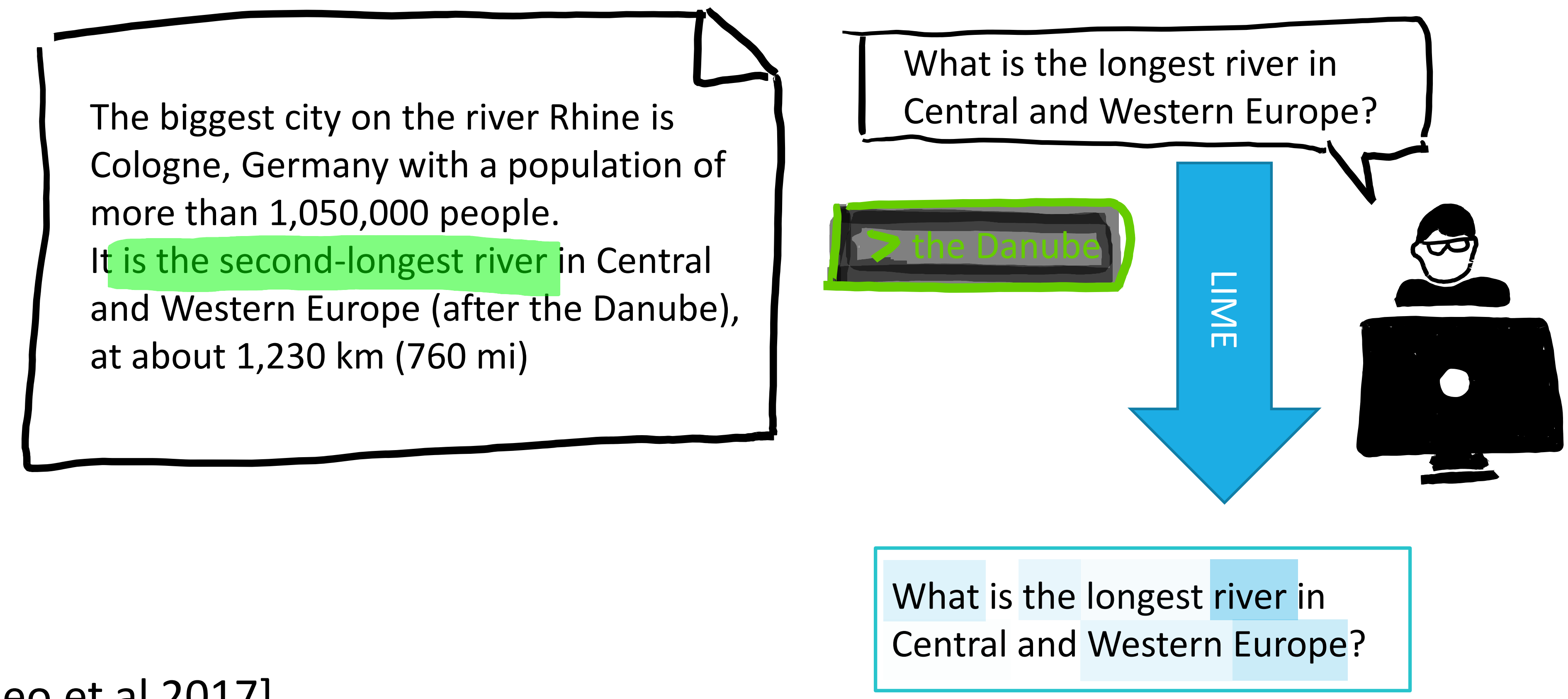
What type of road sign is shown?

> STOP.

LIME

What type of road sign is shown?

LIME on SQuAD



The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people.
It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)

What is the longest river in Central and Western Europe?

> the Danube

What is the longest river in Central and Western Europe?

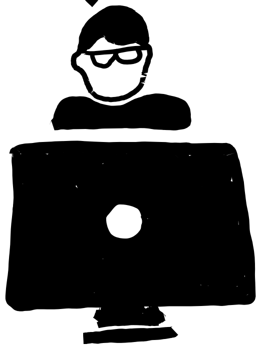
BiDAF [Seo et al 2017]

LIME on SQuAD

The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people.
It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)

What is the second longest river in Central and Western Europe?

> ~~the~~ Danube



LIME

What is the second longest river in Central and Western Europe?

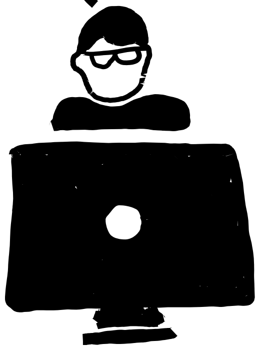
BiDAF [Seo et al 2017]

Limitations of LIME

Gain understanding of *local* behavior, but very little generalization…

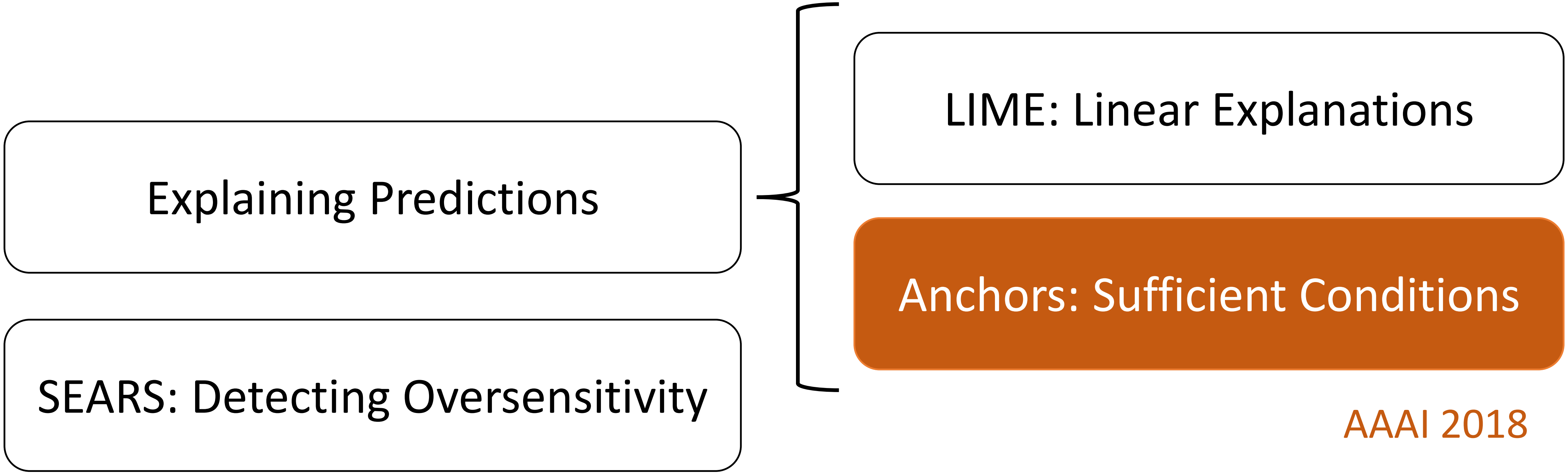
The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people.
It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)

Which is the second longest river in Germany’s part of Europe?



Unless they run it, the users have little idea of what the answer will be

Talk Overview



Explaining Predictions

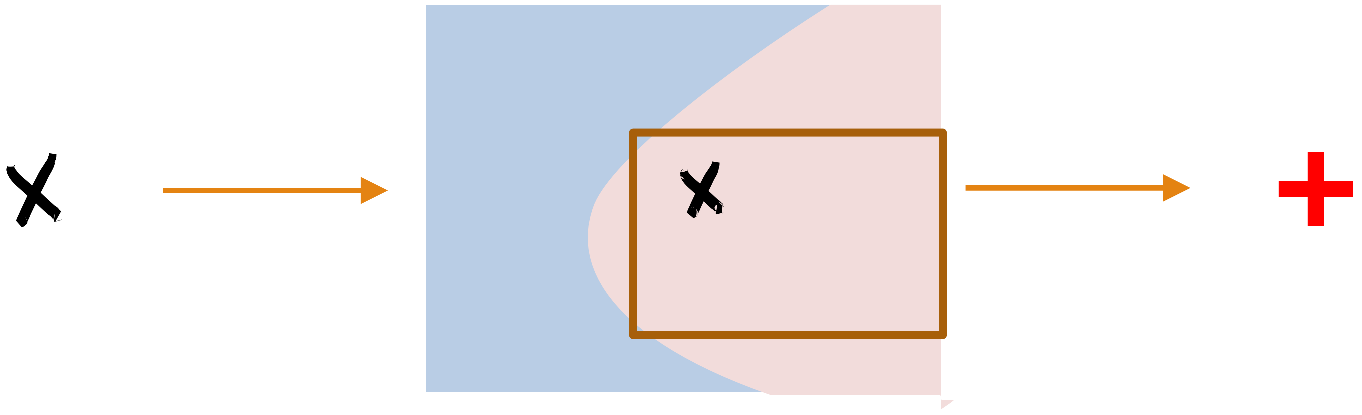
SEARS: Detecting Oversensitivity

LIME: Linear Explanations

Anchors: Sufficient Conditions

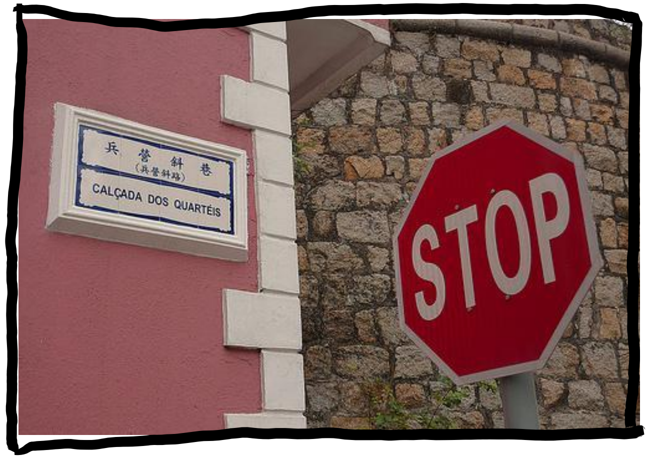
AAAI 2018

Anchors: Sufficient Conditions

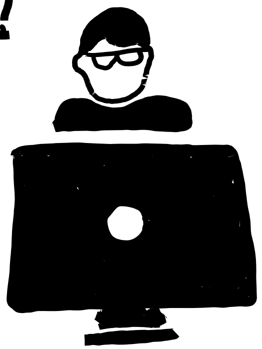


Identify the conditions under which the classifier has the same prediction

Anchors on VisualQA



What type of road sign is shown?



STOP.

LIME

What type of road sign is shown?

Anchor

What type of road sign is shown?

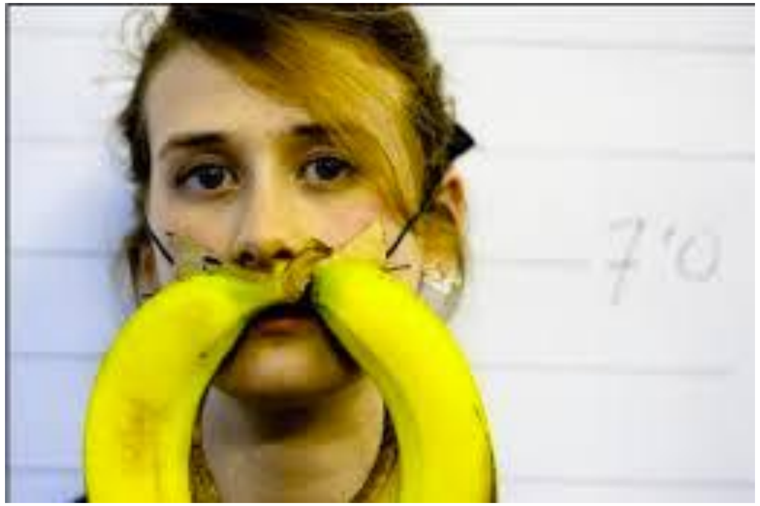
If question starts with *What* (and is similarly structured) the prediction will be **STOP**

96.8%

Anchors on Visual QA

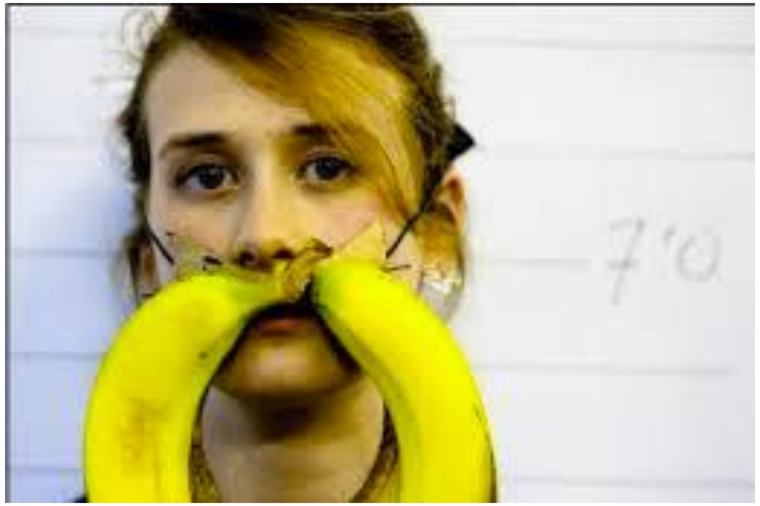
Anchor

What is the mustache made of?	banana



How **many** bananas are in the picture?	2

Anchors on Visual QA

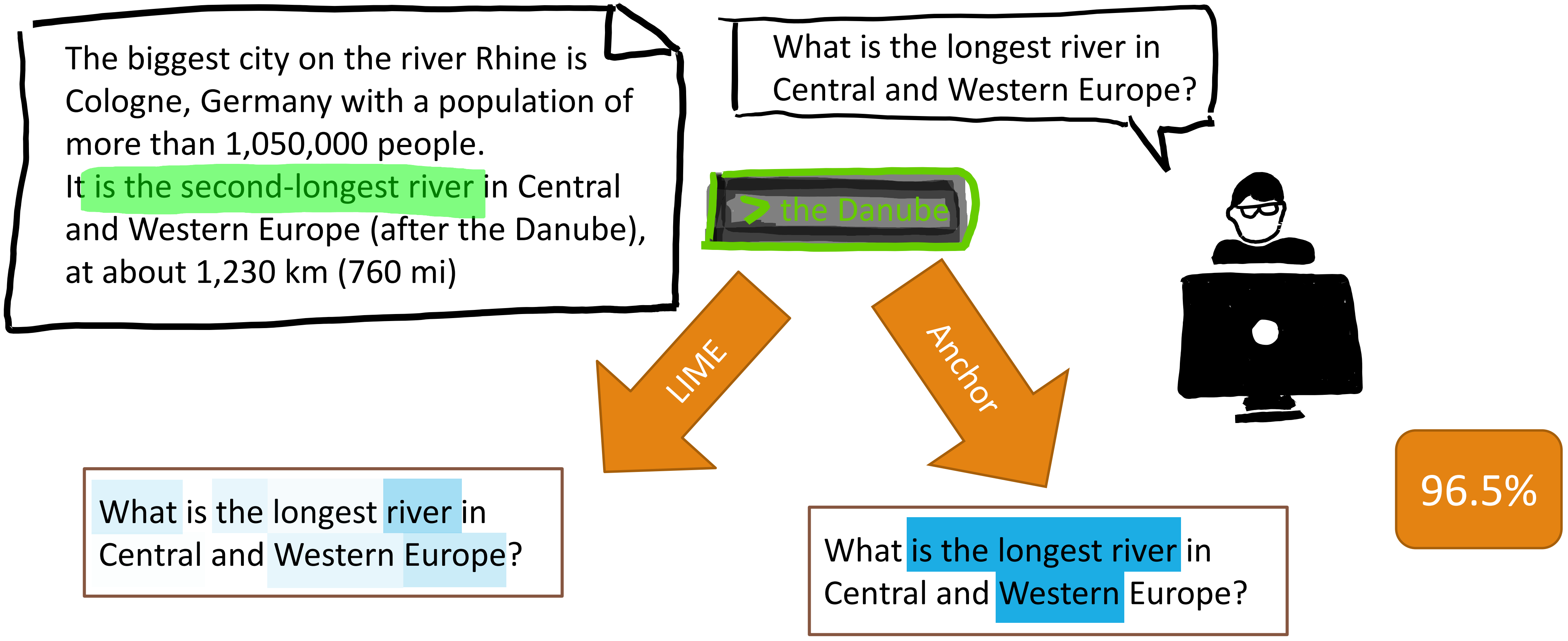


Anchor

What is the mustache made of?	banana
What is the ground made of ?	banana
What is the bed made of ?	banana
What is this mustache ?	banana
What is the man made of?	banana
What is the picture of ?	banana

How **many** bananas are in the picture?	2
How **many** are in the picture?	2
many animals the picture ?	2
How **many** people are in the picture ?	2
How **many** zebras are in the picture ?	2
How **many** planes are on the picture ?	2

Anchors on SQuAD



The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people.
It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)

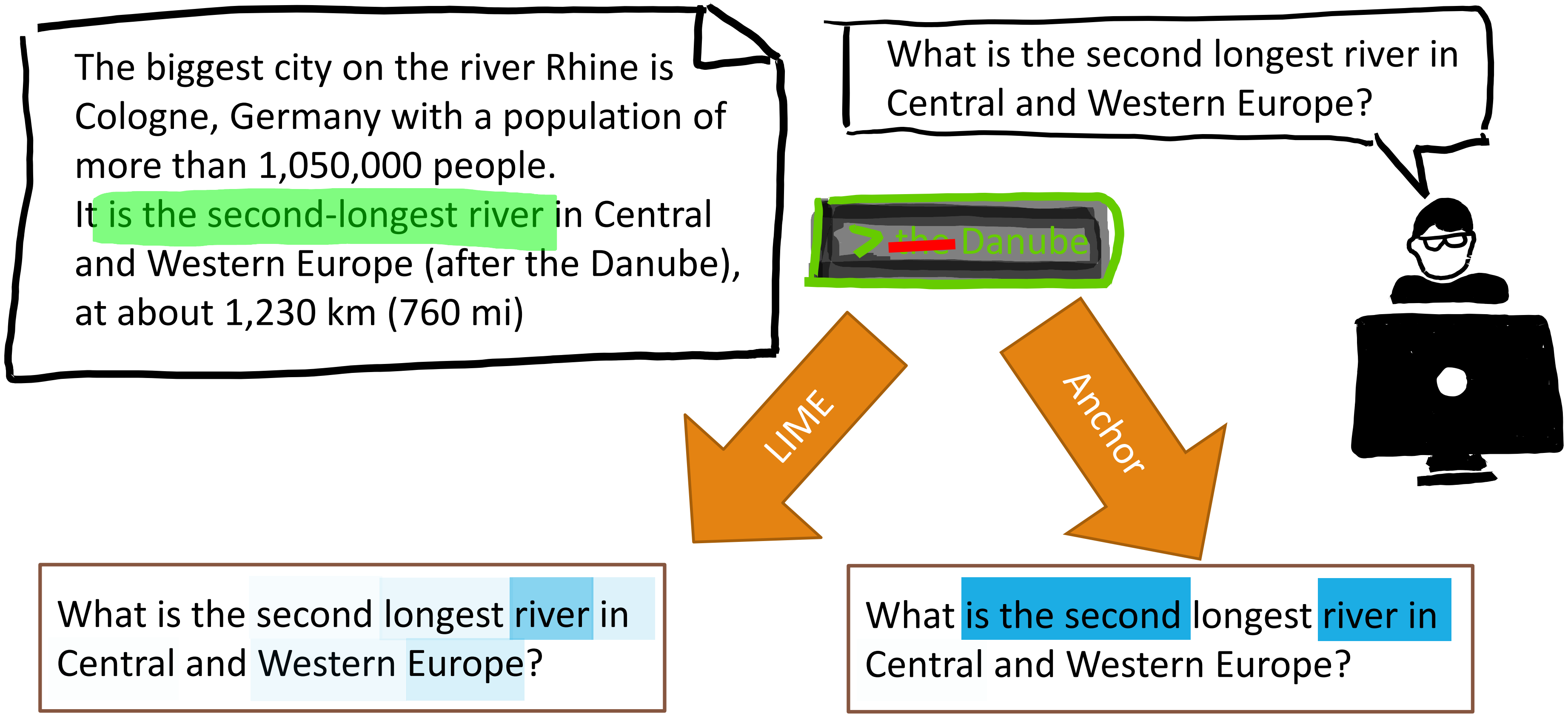
What is the longest river in Central and Western Europe?

What is the longest river in Central and Western Europe?

What is the longest river in Central and Western Europe?

96.5%

Anchors on SQuAD



The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people.
It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)

What is the second longest river in Central and Western Europe?

What is the second longest river in Central and Western Europe?

What is the second longest river in Central and Western Europe?

User study on VisualQA

Show humans predictions + explanations

Ask them to predict what the model will do in new instances (only if confident)

No explanations

Which is the longest river ? Danube

LIME

Which is the longest river ? Danube

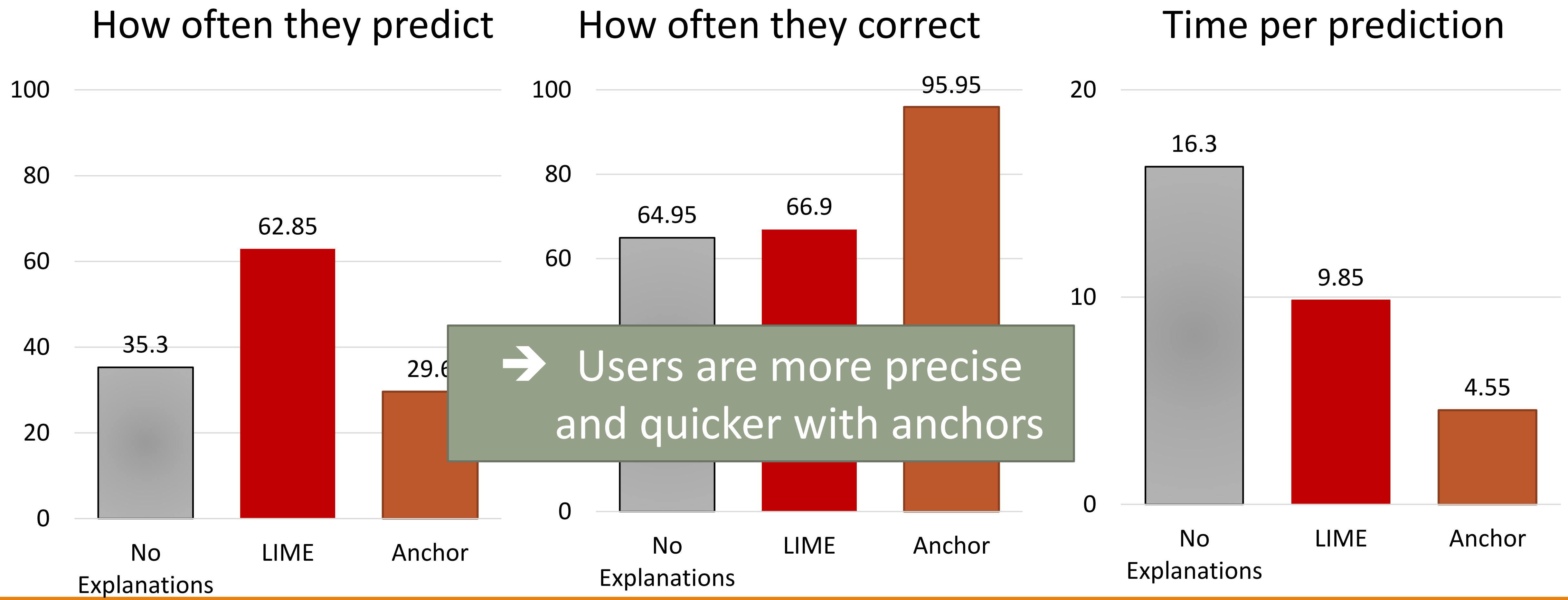
Anchor

Anchor: “**longest river**” → Danube

Which is second longest river?

Danube , Rhine , “I don’t know”

Summary of VisualQA Results



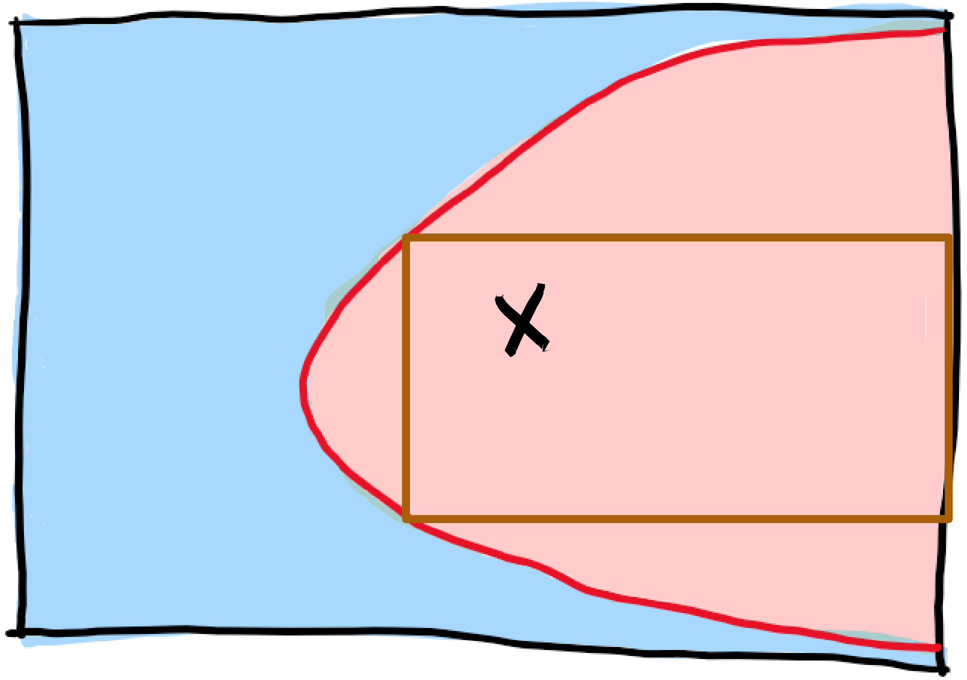
How often they predict

How often they correct

Time per prediction

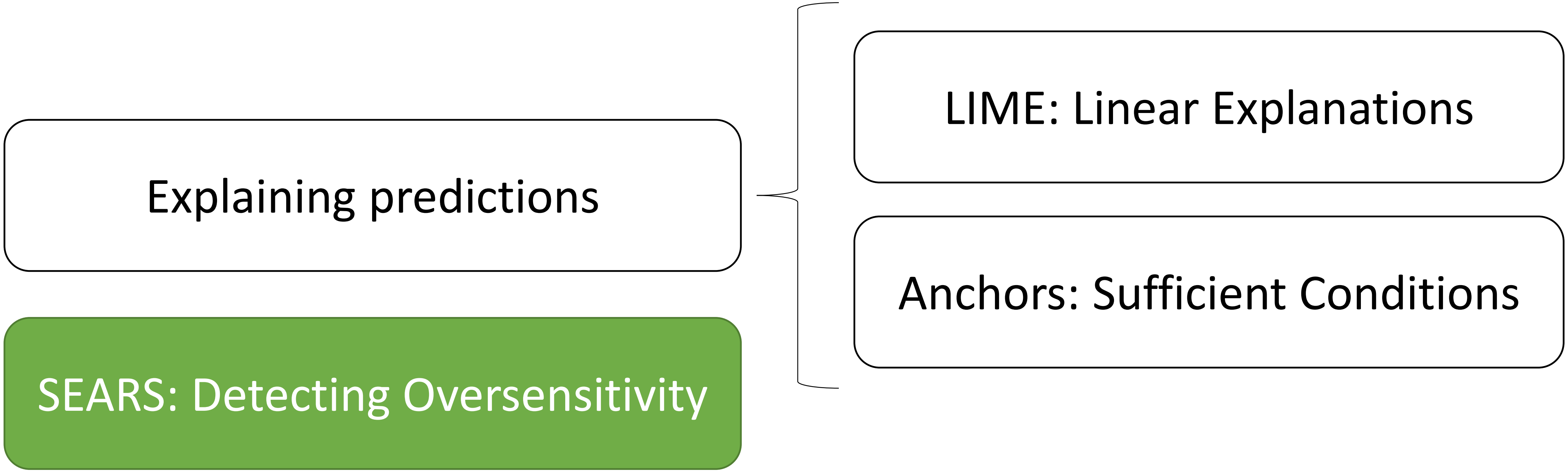
→ Users are more precise and quicker with anchors

Anchors: Tools for Overstability



What about Over-sensitivity?

Talk Overview



Explaining predictions

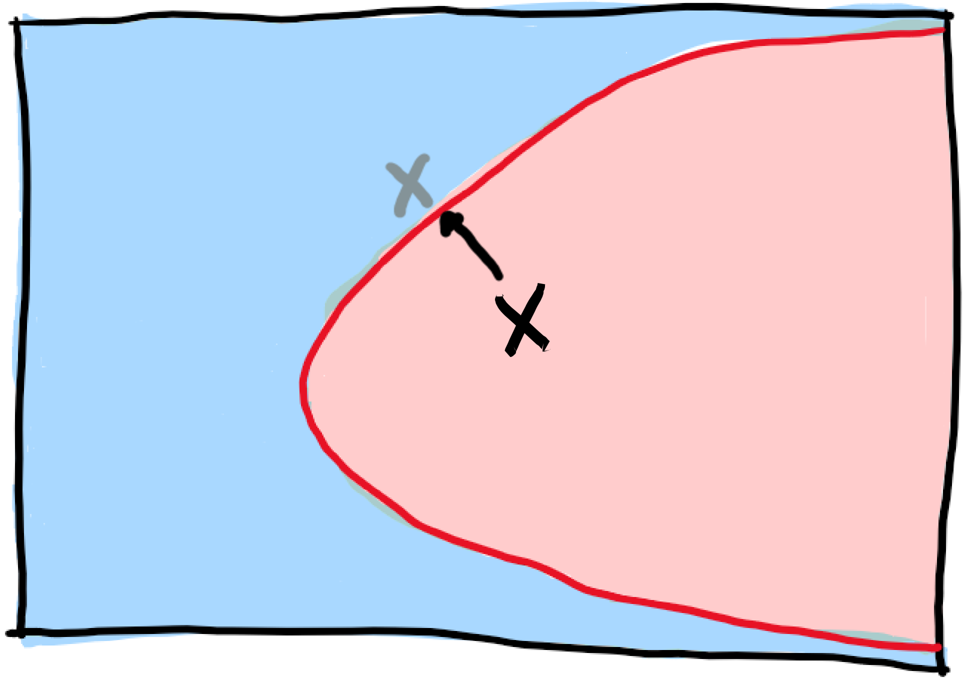
LIME: Linear Explanations

Anchors: Sufficient Conditions

SEARS: Detecting Oversensitivity

ACL 2018

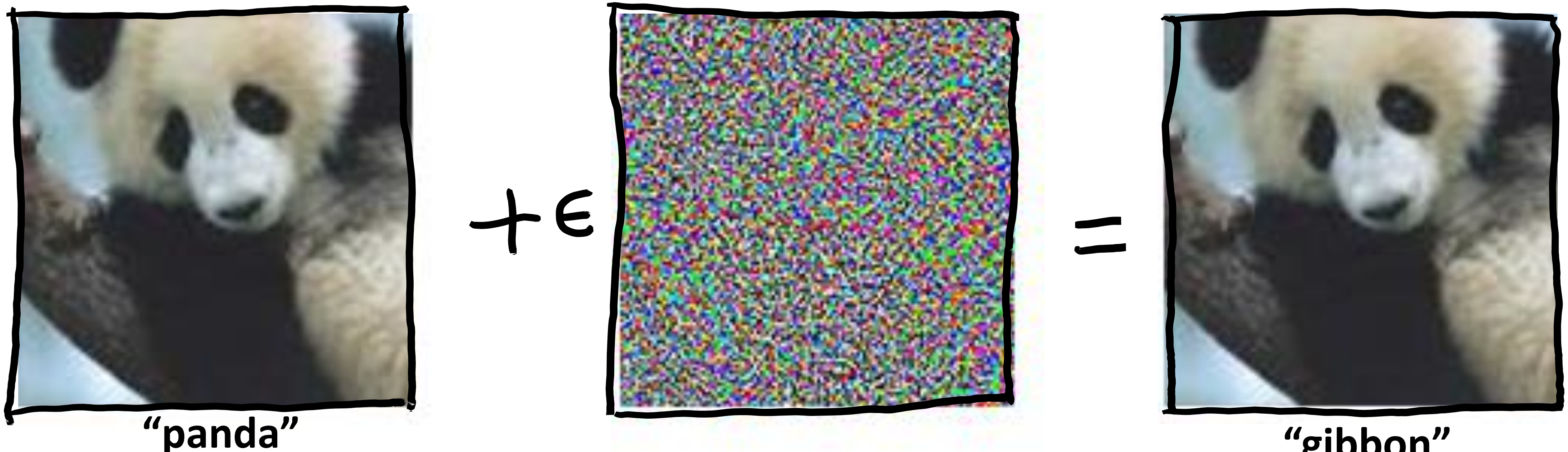
Oversensitivity: Adversarial Examples



Find closest example with different prediction

But unlikely in the real world (except for attacks)

Oversensitivity in images



“panda”

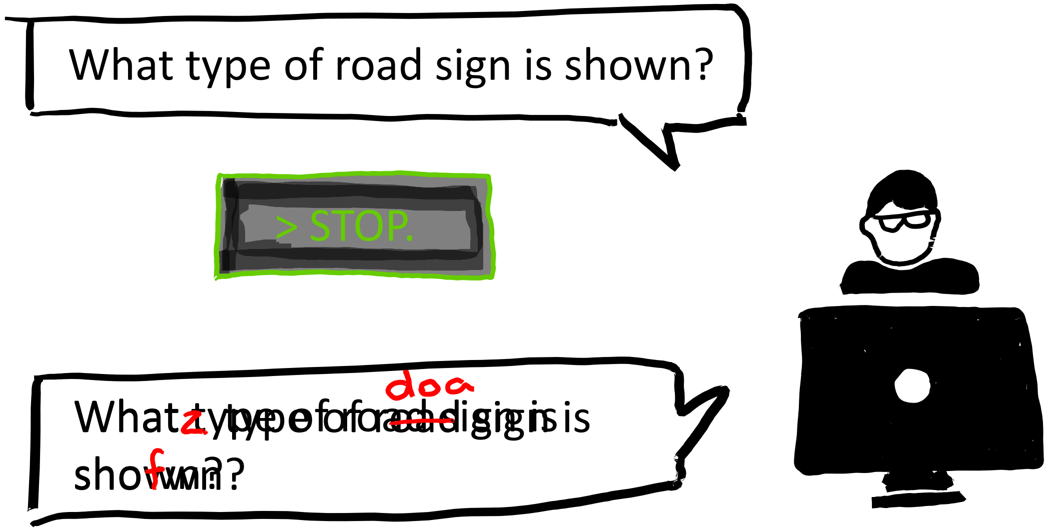
57.7% confidence

“gibbon”

99.3% confidence

Adversaries are indistinguishable to humans…

What about text?



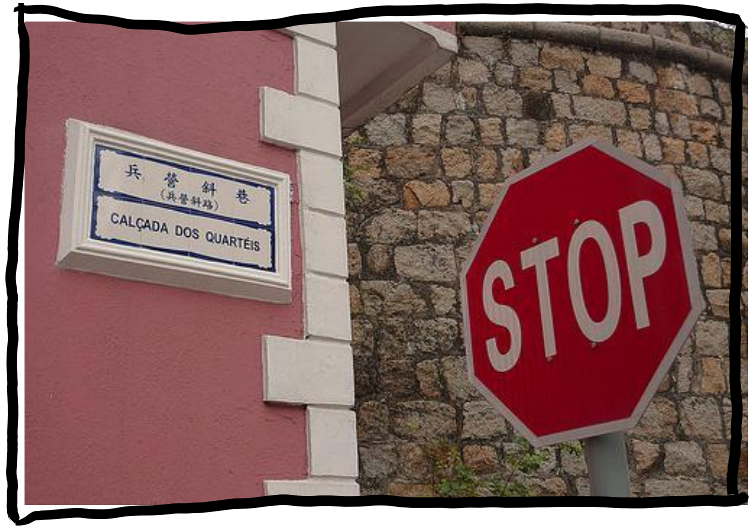
What type of road sign is shown?

> STOP.

What type of ~~road~~ sign is shown?

Perceptible by humans, unlikely in real world

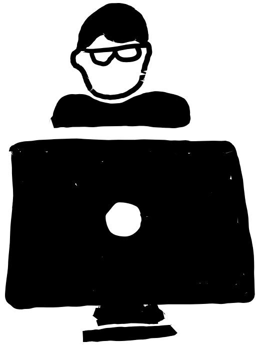
What about text?



What type of road sign is shown?

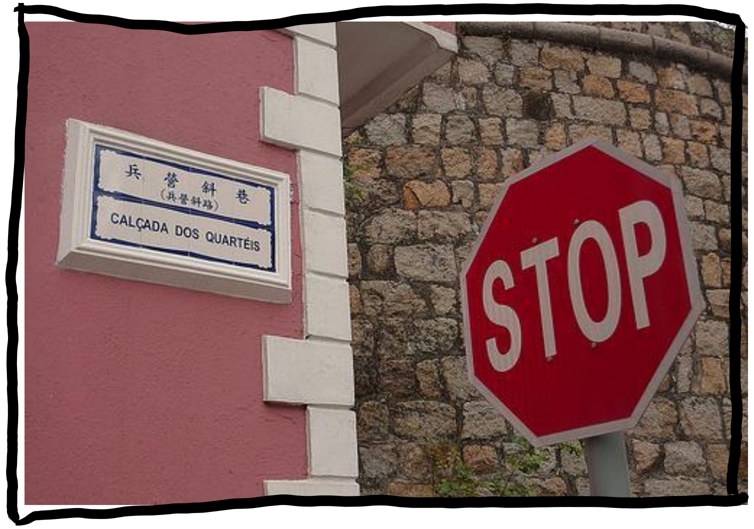
```
> STOP.
```

What type of road sign is NOT shown?



A single word changes too much!

Semantics matter

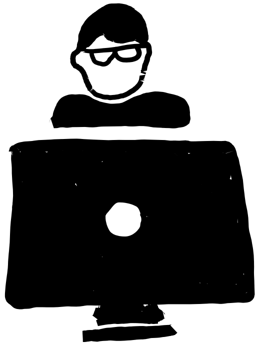


What type of road sign is shown?

> STOP.

Which
~~What~~ type of road sign is shown?

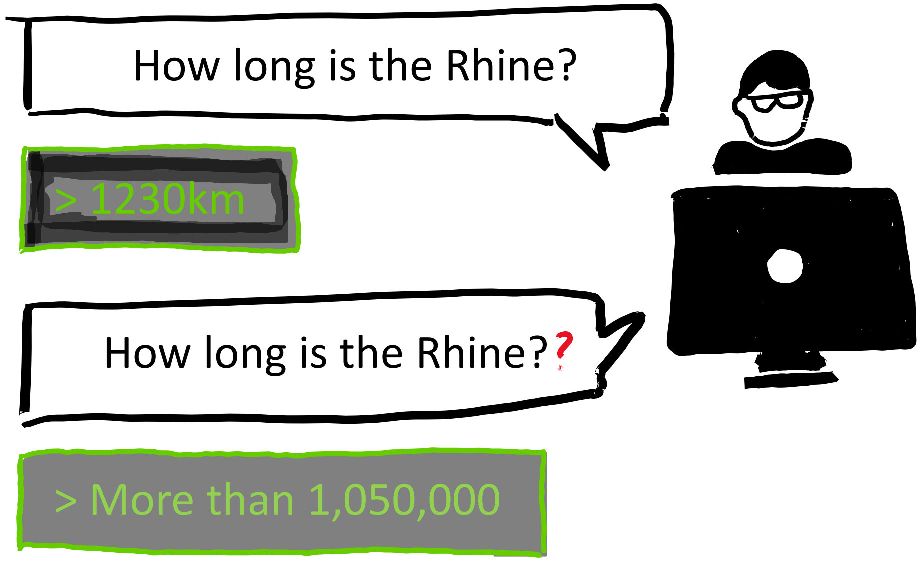
> Do not Enter.



Bug, and likely in the real world

Semantics matter

The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people.
It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)



How long is the Rhine?

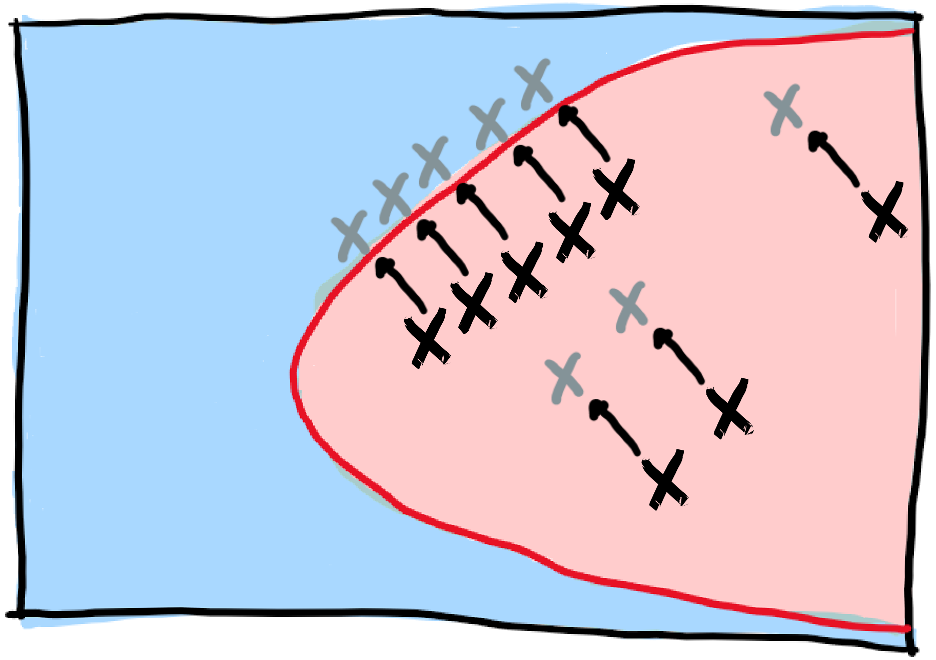
> 1230km

How long is the Rhine??

> More than 1,050,000

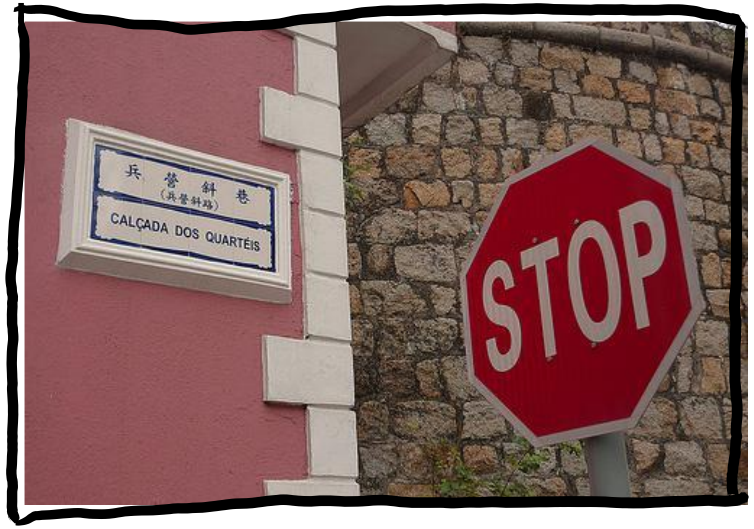
Not all changes are the same: meaning should be same

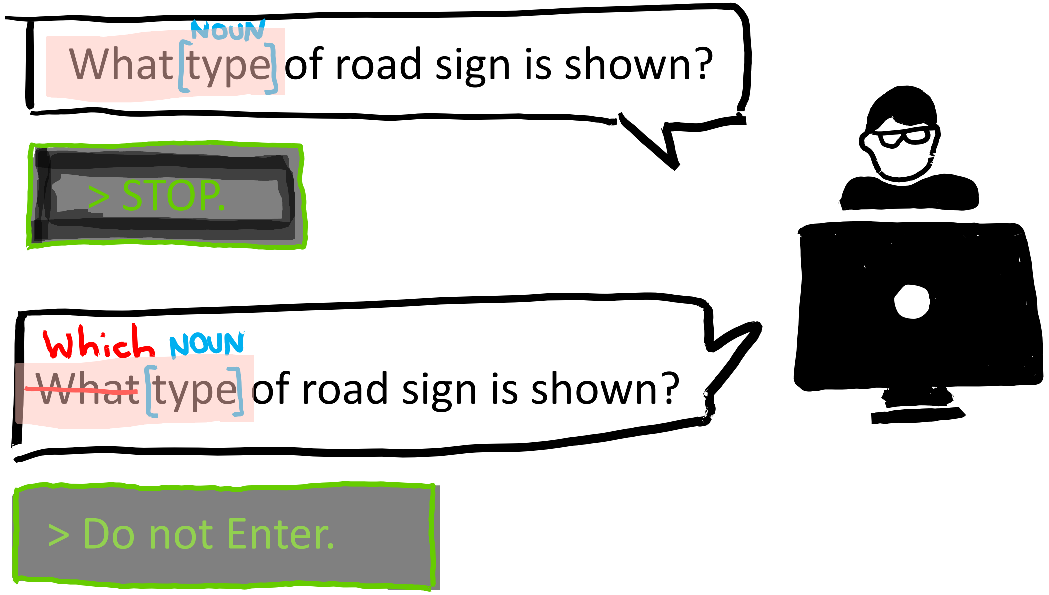
Characterize via Rules



Find rule that generates many adversaries

Characterizing via Rules





NOUN

What [type] of road sign is shown?

> STOP.

Which NOUN

~~What~~ [type] of road sign is shown?

> Do not Enter.

Rule (What NOUN → Which NOUN) - flips 3.9% of examples

Characterizing via Rules

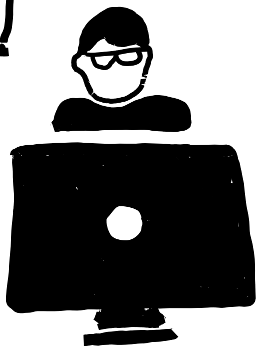
The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people.
It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)

How long is the Rhine?

> 1230km

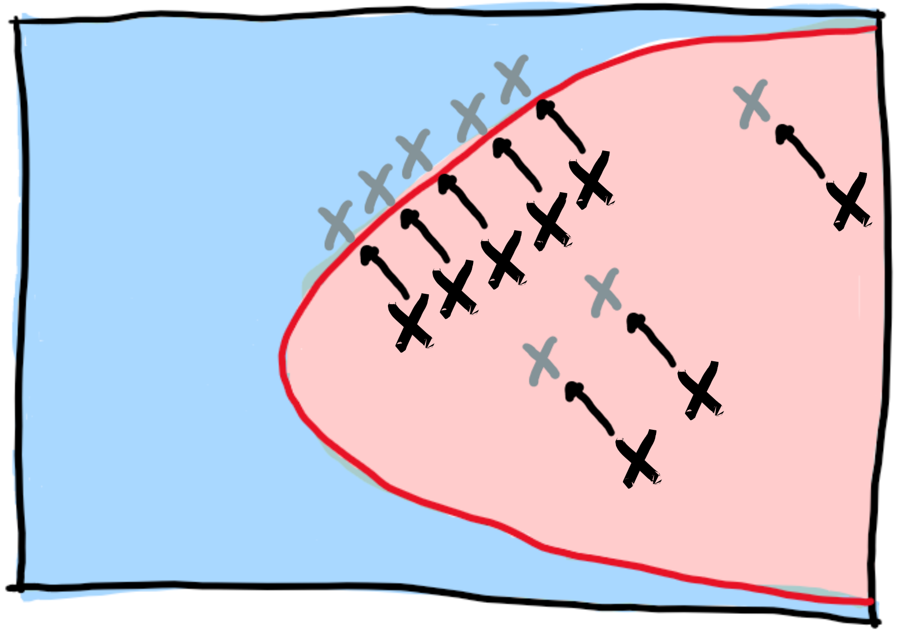
How long is the Rhine??

> More than 1,050,000



Rule (? → ??) - flips 3% of examples

SEARS: Adversarial Rules



Rules are global and actionable,
more interesting than individual adversaries

SEARS Examples: VisualQA

SEAR	Questions / **SEAs**	f(x)	Flips
WP VBZ → **WP's**	~~What has~~ **What's** been cut?	~~Cake~~ **Pizza**	3.3%
What NOUN → **Which NOUN**	~~What~~ **Which** kind of floor is it?	~~Wood~~ **Marble**	3.9%
color → **colour**	What ~~color~~ **colour** is the tray?	~~Pink~~ **Green**	2.2%
ADV is → **ADV's**	~~Where is~~ **Where's** the jet?	~~Sky~~ **Airport**	2.1%

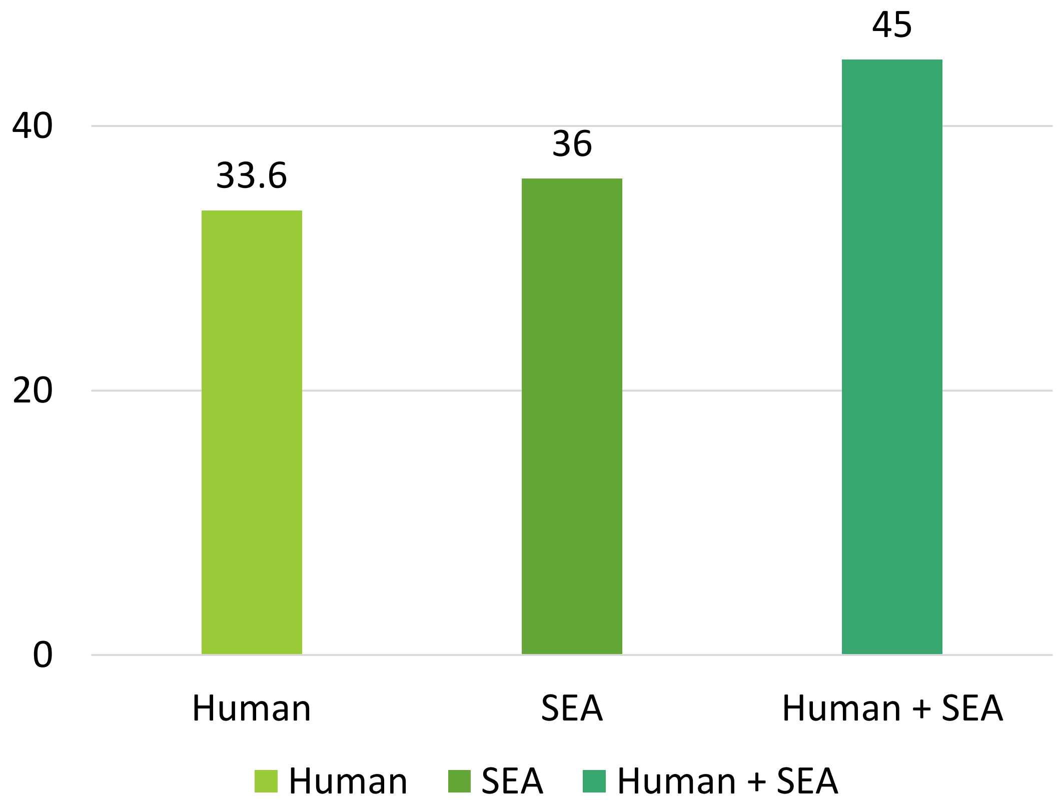
Visual7a-Telling [Zhu et al 2016]

SEARS Examples: SQuAD

SEAR	Questions / **SEAs**	f(x)	Flips
What VBZ → **What's**	~~What is~~ **What's** the NASUWT?	~~Trade union~~ **Teachers in Wales**	2%
What NOUN → **Which NOUN**	~~What resource~~ **Which resource** was mined in the Newcastle area?	~~coal~~ **wool**	1%
What VERB → **So what VERB**	~~What was~~ **So what was** Ghandi's work called?	~~Satyagraha~~ **Civil Disobedience**	2%
What VBD→ **And what VBD**	~~What was~~ **And what was** Kenneth Swezey's job?	~~journalist~~ **sleep**	2%

BiDAF [Seo et al 2017]

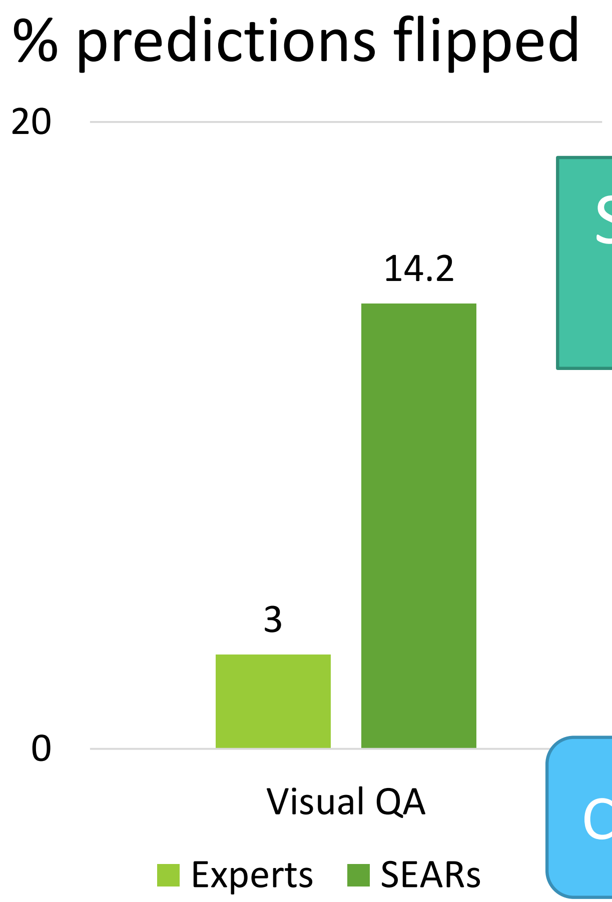
VQA User Study: Detecting adversaries



SEAs find adversaries as often as humans!

SEAs + Humans better than humans!

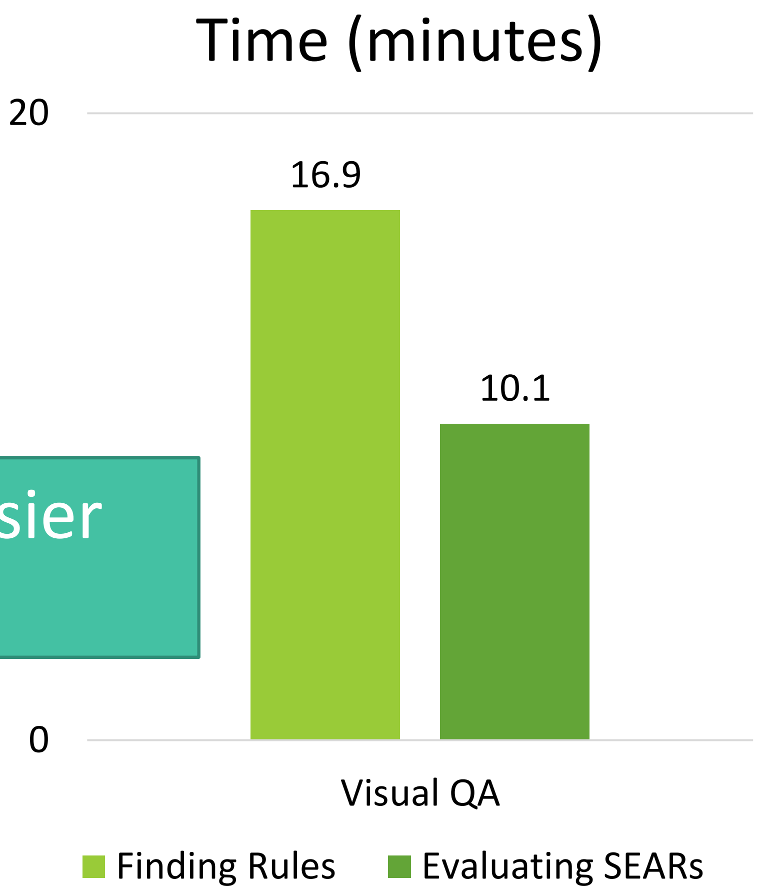
VQA User study: Can experts find bugs?



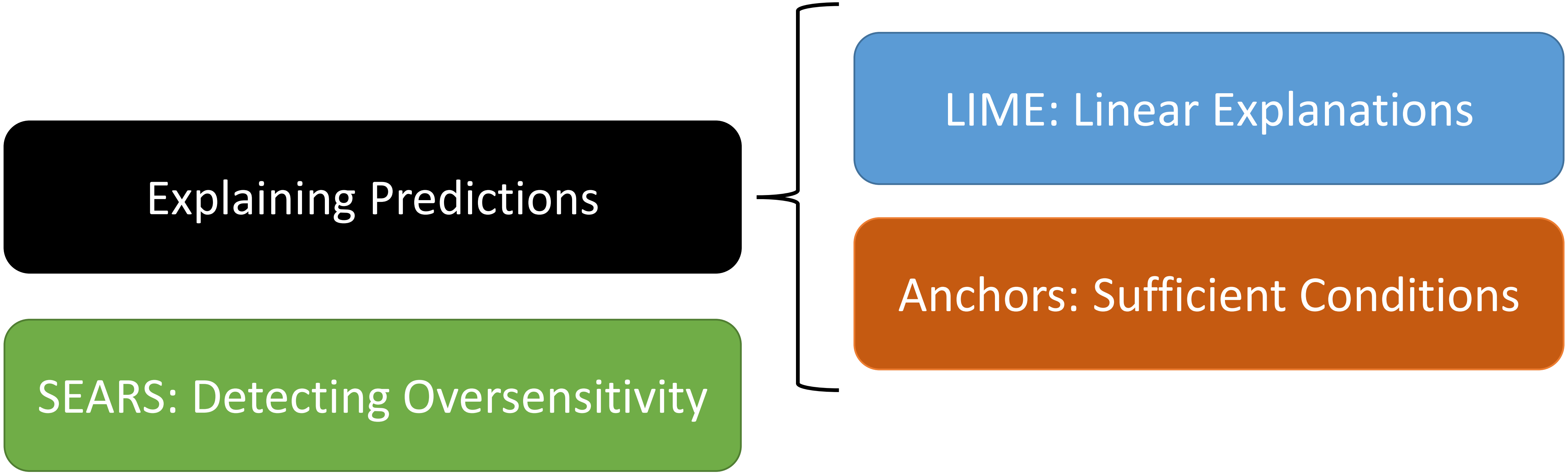
SEARs are much better than expert-produced rules

Evaluating is much easier than finding them

Closing the loop brings it down to 1.4%



Talk Overview



Explaining Predictions

SEARS: Detecting Oversensitivity

LIME: Linear Explanations

Anchors: Sufficient Conditions

Why such tools can be useful

- Annotations and Task Definitions
  - SQuAD 2.0: unanswerable questions
  - VisualQA 2.0: questions with different answers
- Evaluation
  - Create robust test set
  - Include explanations/bugs as qualitative evaluation
- End to End QA may not be sufficient
  - Saleforce’s NLP Decathalon
  - ELMO Representation: learn across domains, and fine-tune!

Questioning Question Answering Answers

Work with Marco T. Ribeiro and Carlos Guestrin, University of Washington

Thanks!

sameer@uci.edu
sameersingh.org
@sameer_

Work with **Matt Gardner** and me

as part of

The Allen Institute for
Artificial Intelligence
in **Irvine**, CA

All levels: pre-docs, PhD interns, postdocs, and research scientists!