

Visual Question Answering and Dialog

Danna Gurari

The University of Texas at Austin

Fall 2019



Review

- Last week:
 - Object tracking applications
 - Evaluating object tracking methods
 - Strategies for crowdsourcing object tracking
 - Object tracking datasets & challenges
- Assignments (Class Website & Canvas)
 - Project outline and prototype demo due next week
- Questions?

Today's Topics

- Visual question answering and dialog applications
- Evaluation for visual question answering
- Crowdsourcing for visual question answering
- Crowdsourcing for visual dialog
- Guest: General Manager Warren Barkley at Amazon Web Services

Today's Topics

- Visual question answering and dialog applications
- Evaluation for visual question answering
- Crowdsourcing for visual question answering
- Crowdsourcing for visual dialog
- Guest: General Manager Warren Barkley at Amazon Web Services

Task: Answer Visual Questions (VQs)



Is my monitor on?



Hi there can you please tell me what flavor this is?



Does this picture look scary?



Which side of the room is the toilet on?

Visual Assistance for People with Vision Loss; e.g.,

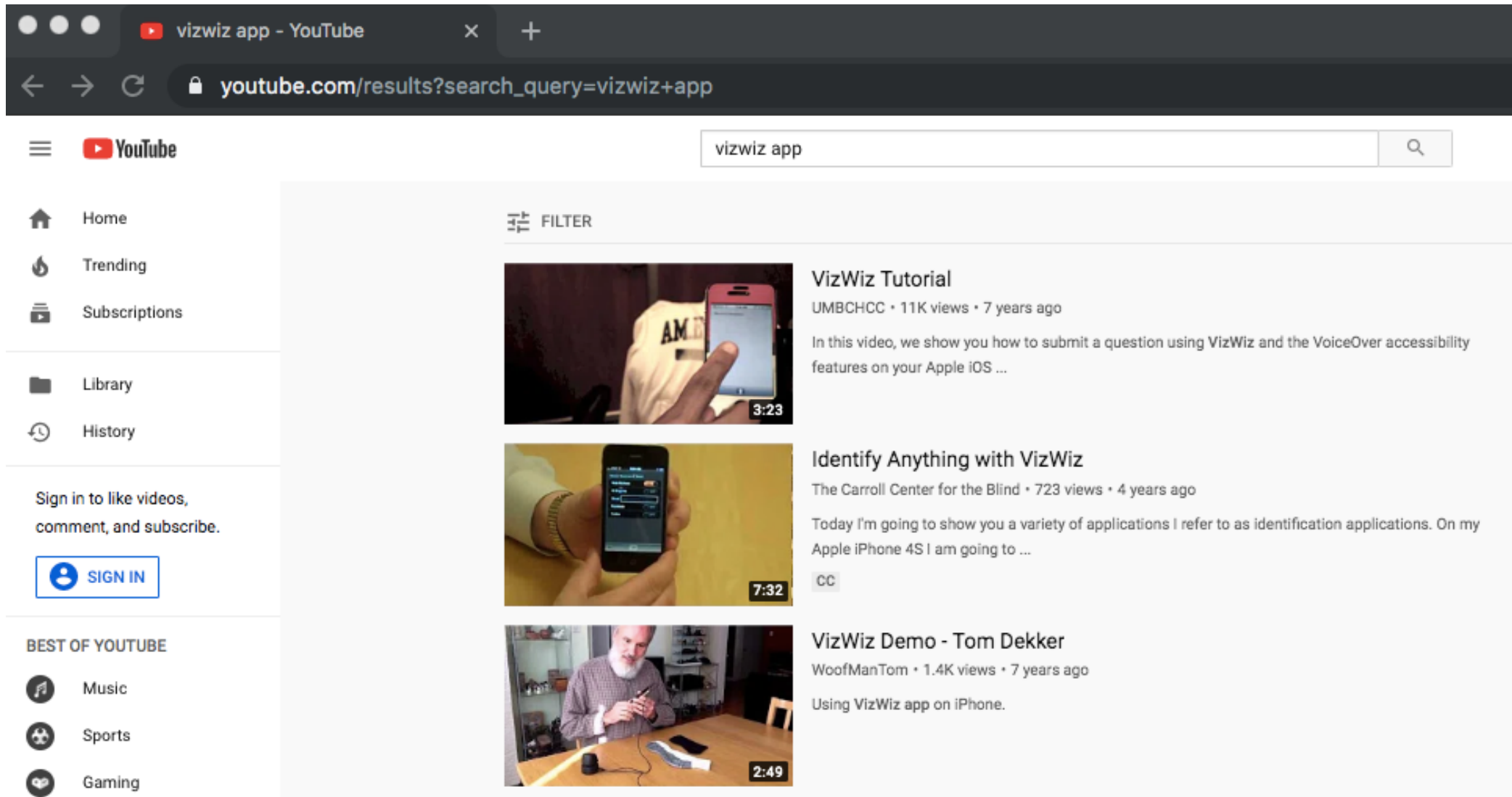


BeSpecular
Through the world's eyes



Be My Eyes

Visual Assistance for People with Vision Loss; e.g.,



The image shows a browser window displaying a YouTube search results page for the query "vizwiz app". The browser's address bar shows the URL "youtube.com/results?search_query=vizwiz+app". The YouTube interface includes a search bar with the query "vizwiz app", a left-hand navigation menu with options like Home, Trending, Subscriptions, Library, and History, and a main content area with three video results. The first result is "VizWiz Tutorial" by UMBCHCC, with 11K views and posted 7 years ago. The second is "Identify Anything with VizWiz" by The Carroll Center for the Blind, with 723 views and posted 4 years ago. The third is "VizWiz Demo - Tom Dekker" by WoofManTom, with 1.4K views and posted 7 years ago. The video thumbnails show a person using the app on an iPhone.

Browser tabs: vizwiz app - YouTube

Address bar: youtube.com/results?search_query=vizwiz+app

Search bar: vizwiz app

Navigation menu: Home, Trending, Subscriptions, Library, History

Sign in to like videos, comment, and subscribe. [SIGN IN](#)

BEST OF YOUTUBE

- Music
- Sports
- Gaming

FILTER

VizWiz Tutorial
UMBCHCC • 11K views • 7 years ago
In this video, we show you how to submit a question using VizWiz and the VoiceOver accessibility features on your Apple iOS ...
3:23

Identify Anything with VizWiz
The Carroll Center for the Blind • 723 views • 4 years ago
Today I'm going to show you a variety of applications I refer to as identification applications. On my Apple iPhone 4S I am going to ...
7:32

VizWiz Demo - Tom Dekker
WoofManTom • 1.4K views • 7 years ago
Using VizWiz app on iPhone.
2:49

Visual Assistance for People with Vision Loss; e.g.,



 1-800-835-1934

Connecting you to real people instantly to simplify daily life

Get started for free today. Type in your number and hit submit to have a link sent to your phone to download the app.

Aira is supported in the US, Canada, Australia, and New Zealand.

Or contact us: [1-800-835-1934](tel:1-800-835-1934)



Visual Question Answering (VQA)

For what other applications could visual question answer and dialog systems be useful?


Visual Question Answering (VQA)

How is VQA different than the image captioning task?

VQA Dialog

“hold a meaningful dialog with humans in natural language about visual content”

Visual Dialog



A cat drinking water out of a coffee mug.

What color is the mug?

White and red

Are there any pictures on it?

No, something is there can't tell what it is

Is the mug and cat on a table?

Yes, they are

Are there other items on the table?

Yes, magazines, books, toaster and basket, and a plate

Start typing question here ...

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra; Visual Dialog. CVPR 2017.

VQA Dialog vs VQA and Image Descriptions



VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

Captioning

Two people are in a wheelchair and one is holding a racket.

Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman



Visual Dialog

Q: What is the gender of the one in the white shirt ?

A: She is a woman

Q: What is she doing ?

A: Playing a Wii game

Q: Is that a man to her right

A: No, it's a woman

Involves

- Memory
- Follow-up questions

Today's Topics

- Visual question answering and dialog applications
- **Evaluation for visual question answering**
- Crowdsourcing for visual question answering
- Crowdsourcing for visual dialog
- Guest: General Manager Warren Barkley at Amazon Web Services

Class Task: Answer Visual Question



Is my monitor on?

(1)



Hi there can you please tell me what flavor this is?

(2)



Does this picture look scary?

(3)



Which side of the room is the toilet on?

(4)

Crowdsourced Answers



Is my monitor on?

- (1) yes
- (2) yes
- (3) yes
- (4) yes
- (5) yes
- (6) yes
- (7) yes
- (8) yes
- (9) yes
- (10) yes



Hi there can you please tell me what flavor this is?

- (1) sweet pepper
- (2) sweet pepper
- (3) sweet pepper
- (4) sweet pepper
- (5) sweet pepper
- (6) sweet pepper
- (7) sweet pepper
- (8) sweet pepper
- (9) sweet pepper
- (10) sweet pepper



Does this picture look scary?

- (1) yes
- (2) no
- (3) no
- (4) yes
- (5) no
- (6) yes
- (7) yes
- (8) no
- (9) no
- (10) no



Which side of the room is the toilet on?

- (1) right
- (2) left
- (3) right
- (4) right
- (5) right
- (6) right
- (7) right side
- (8) right
- (9) center
- (10) right

Evaluating Automated Predictions

VQA: Ask any question about this image



Is this man thirsty?

Answer

Answer	Confidence
yes	0.8778
no	0.1211
6	0.0001
5	0.0001
pink	0.0001

<https://vqa.cloudcv.org/>

Evaluating Automated Predictions



Is my monitor on?

(1) yes



Hi there can you please tell me what flavor this is?

(2) chocolate



Does this picture look scary?

(3) yes



Which side of the room is the toilet on?

(4) right

Evaluating Automated Predictions

$$\text{accuracy} = \min\left(\frac{\# \text{ humans that provided that answer}}{3}, 1\right)$$

Evaluation: Example



Does this picture
look scary?

- (1) yes
- (2) no
- (3) no
- (4) yes
- (5) no
- (6) yes
- (7) yes
- (8) no
- (9) no
- (10) no

What is the accuracy of an algorithm prediction of

- “yes”?
- “no”?
- “maybe”?

$$\text{accuracy} = \min\left(\frac{\# \text{ humans that provided that answer}}{3}, 1\right)$$

Evaluation: Example



Which side of the room is the toilet on?

- (1) right
- (2) left
- (3) right
- (4) right
- (5) right
- (6) right
- (7) right side
- (8) right
- (9) center
- (10) right

What is the accuracy of an algorithm prediction of

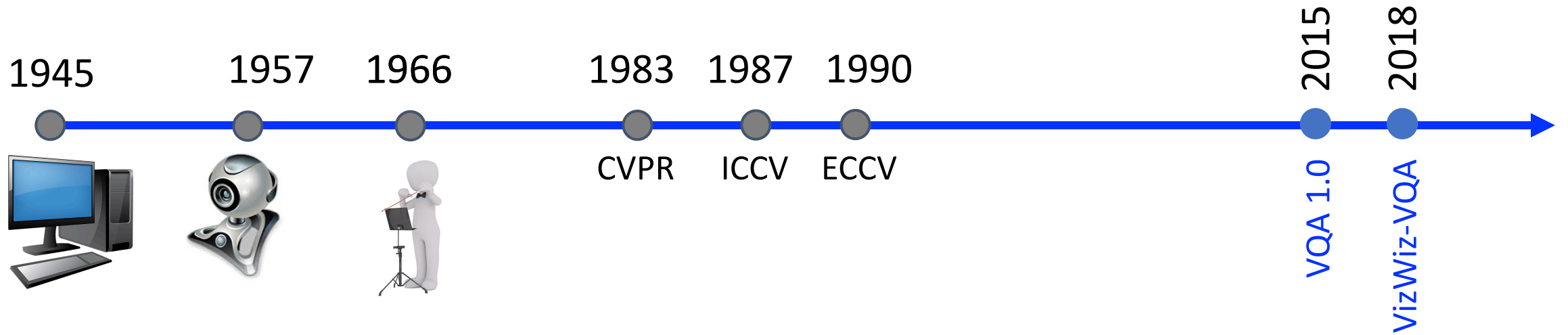
- “right”?
- “left”?
- “right side”?
- “center”?
- “bottom”?

$$\text{accuracy} = \min\left(\frac{\# \text{ humans that provided that answer}}{3}, 1\right)$$

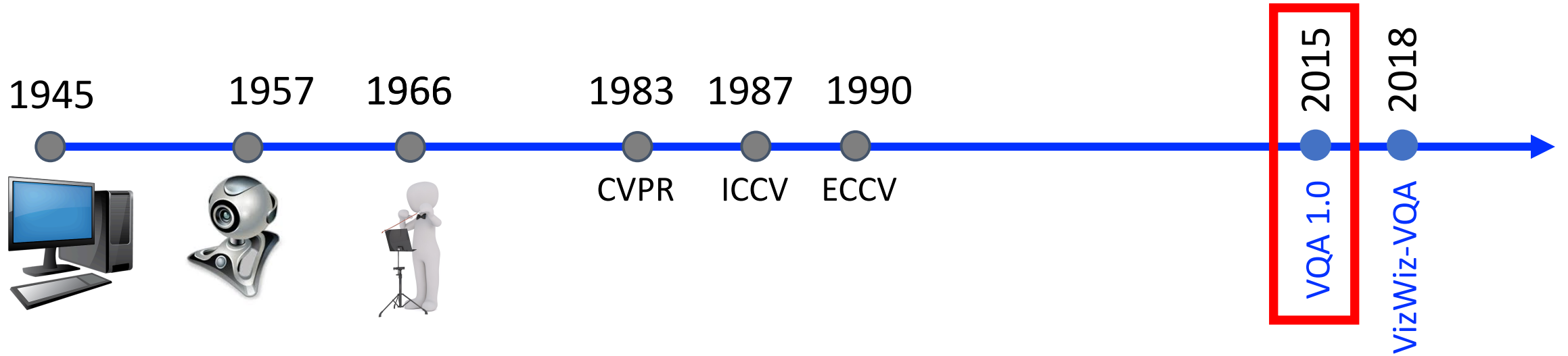
Today's Topics

- Visual question answering and dialog applications
- Evaluation for visual question answering
- **Crowdsourcing for visual question answering**
- Crowdsourcing for visual dialog
- Guest: General Manager Warren Barkley at Amazon Web Services

Visual Question Answering Datasets



Visual Question Answering Datasets



VQA 1.0

1. Collect Images

* MS-COCO
- 204,721

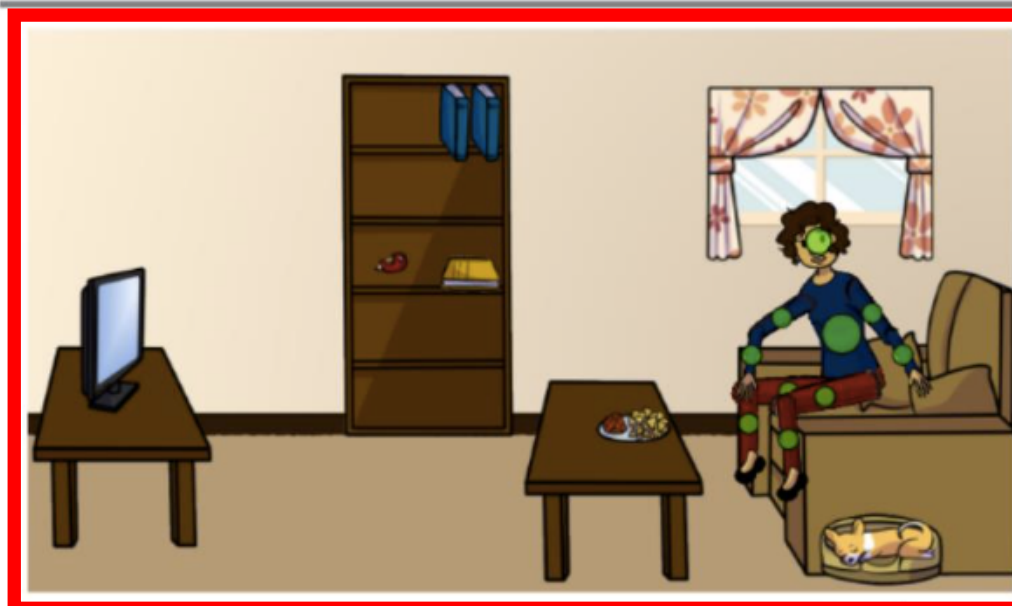
* Abstract
- 90,000

Help Us Create Clipart Illustrations! (Living/Dining Room)



Scene 4/4

Prev Next



Scene Depth

Flip

Expression



VQA 1.0

1. Collect Images 2. Collect Questions

* MS-COCO
- 204,721

* Abstract
- 90,000

* Pilot study:
compare
questions
designed to be
too difficult to
answer for a:
- "toddler"
- "alien"

- "smart robot"

* Collect 3
questions per
image

Stump a smart robot! Ask a question about this scene that a human can answer, but a smart robot probably can't!

Updated instructions: Please read carefully

Hide

Show

We have built a smart robot. It understands a lot about scenes. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene type (e.g., kitchen, beach), people's expressions and poses, and properties of objects (e.g., the color of objects, their texture). Your task is to stump this smart robot! **In particular, it already knows answers to some questions about this scene. We will tell you what these questions are.**

Ask a question about this scene that this SMART robot probably can not answer, but any human can easily answer while looking at the scene in the image. **IMPORTANT:** The question should be about this scene. That is, the human should need the image to be able to answer the question – the human should not be able to answer the question without looking at the image.



Your work **will get rejected** if you do not follow the instructions below:

- Do not ask questions that are similar to the ones listed below each image. As mentioned, the robot already knows the answers to those questions for the scene in this image. Please ask about something different.
- Do not repeat questions. Do not ask the same questions or the same questions with minor variations over and over again across images. Think of a **new question each time** specific to the scene in each image.
- Each question should be a **single question**. Do not ask questions that have multiple parts or multiple sub-questions in them.
- Do not ask generic questions that can be asked of many other scenes. Ask questions **specific to the scene in each image**.

Below is a list of questions the smart robot can already answer. Please ask a different question about this scene that a human can answer "if" looking at the scene in the image (and not otherwise), but would stump this smart robot:

Q1: What is unusual about this mustache? (The robot already knows the answer to this question.)

Q2: What is her facial expression? (The robot already knows the answer to this question.)

Q3: Write your question, different from the questions above, here to stump this smart robot.

prev

next

Page 2/3

VQA 1.0

1. Collect Images 2. Collect Questions

3. Collect Answers



VQA 1.0

Help Us Answer Questions About Images!

Updated instructions: Please read carefully

Hide

Show

Please answer some questions about images **with brief answers**. Your answers should be how most other people would answer the questions. If the question doesn't make sense, please try your best to answer it and indicate via the buttons that you are unsure of your response.

If you don't follow the following instructions, your work will be rejected.



Your work **will get rejected** if you do not follow the instructions below:

- Answer the question based on what is going on in the **scene depicted in the image**.
- Your answer should be a **brief phrase** (not a complete sentence).
 - "It is a kitchen." -> "kitchen"
- For yes/no questions, please **just say yes/no**.
 - "You bet it is!" -> "yes"
- For numerical answers, please use **digits**.
 - "Ten." -> "10"
- If you need to speculate (e.g., "What just happened?"), provide an answer **that most people would agree on**.
- If you don't know the answer (e.g., specific dog breed), provide **your best guess**.
- Respond matter-of-factly and **avoid using conversational language or inserting your opinion**.

3. Collect Answers

* Collect 10 answers per visual question

Please answer the question using as few words as possible:

Q1: What is unusual about this mustache?

A1:

Do you think you were able to answer the question correctly?

(Clicking an option will take you to the next question.)

no

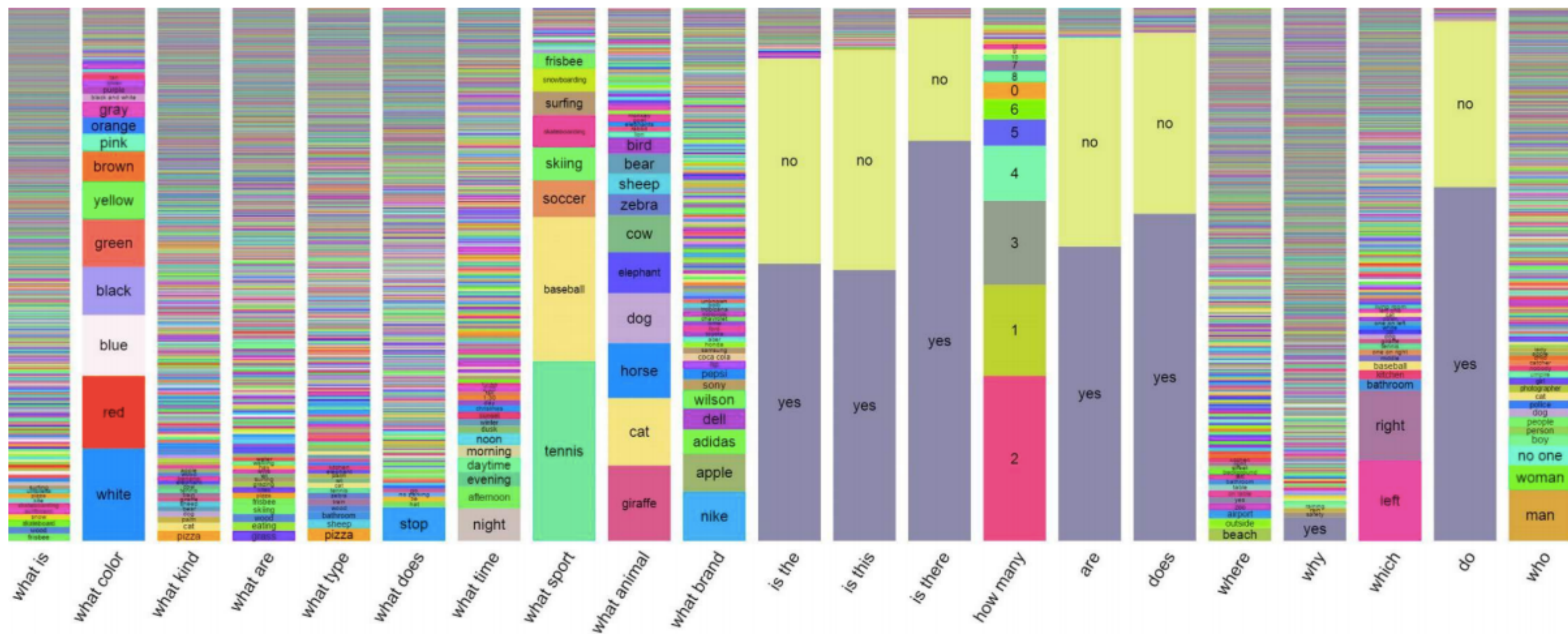
maybe

yes

Page 1/2

VQA 1.0 Answers

Answers with Images

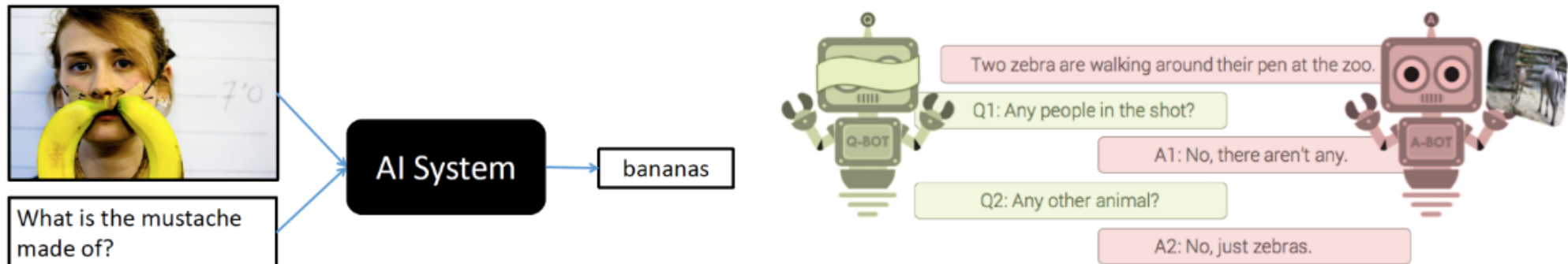


VQA Annual Challenge (held for 4 years now)

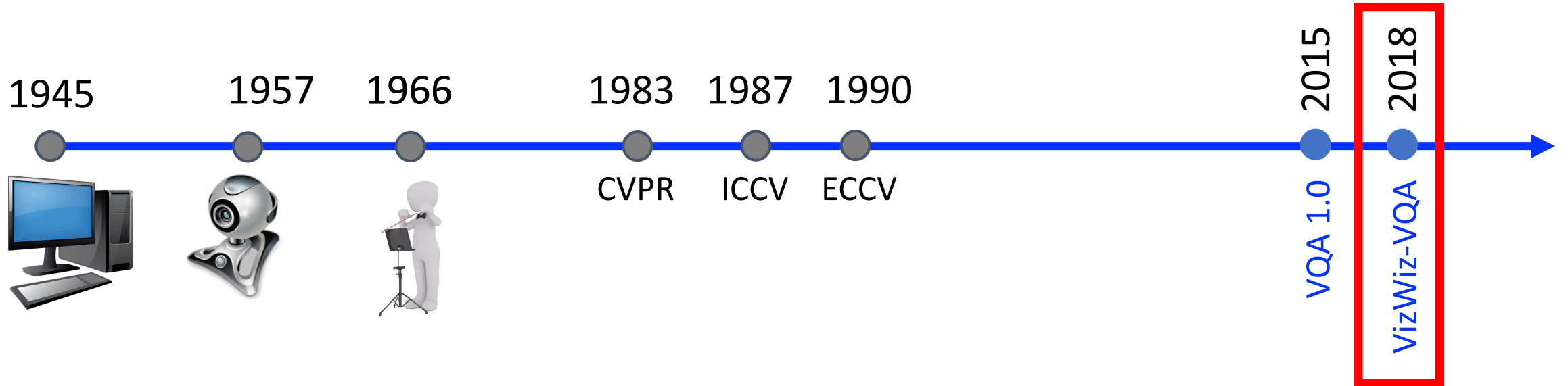
Visual Question Answering and Dialog Workshop

Location: **Seaside Ballroom B, Long Beach Convention & Entertainment Center**

at CVPR 2019, June 17, Long Beach, California, USA



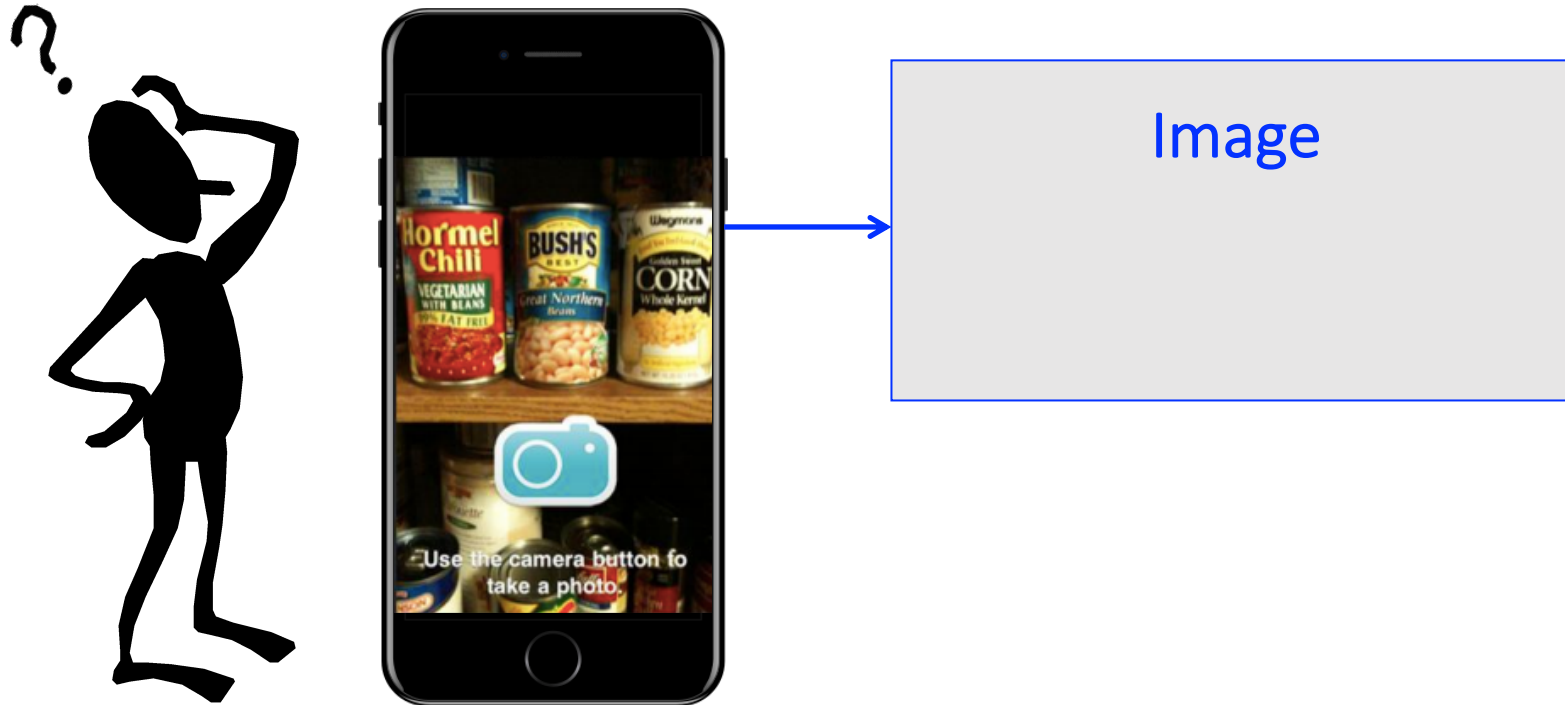
Visual Question Answering Datasets



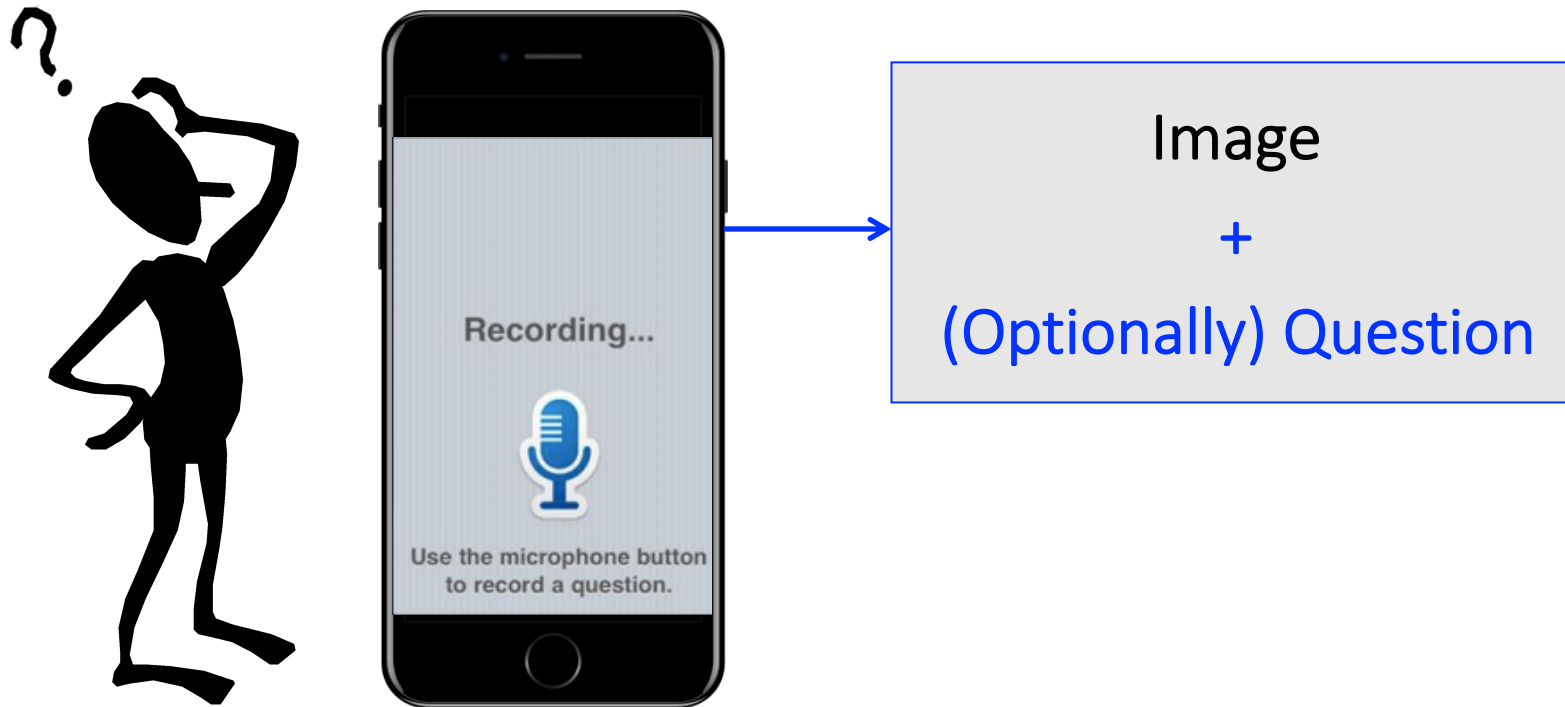
Data: Real Users of On-Demand Visual Assistance



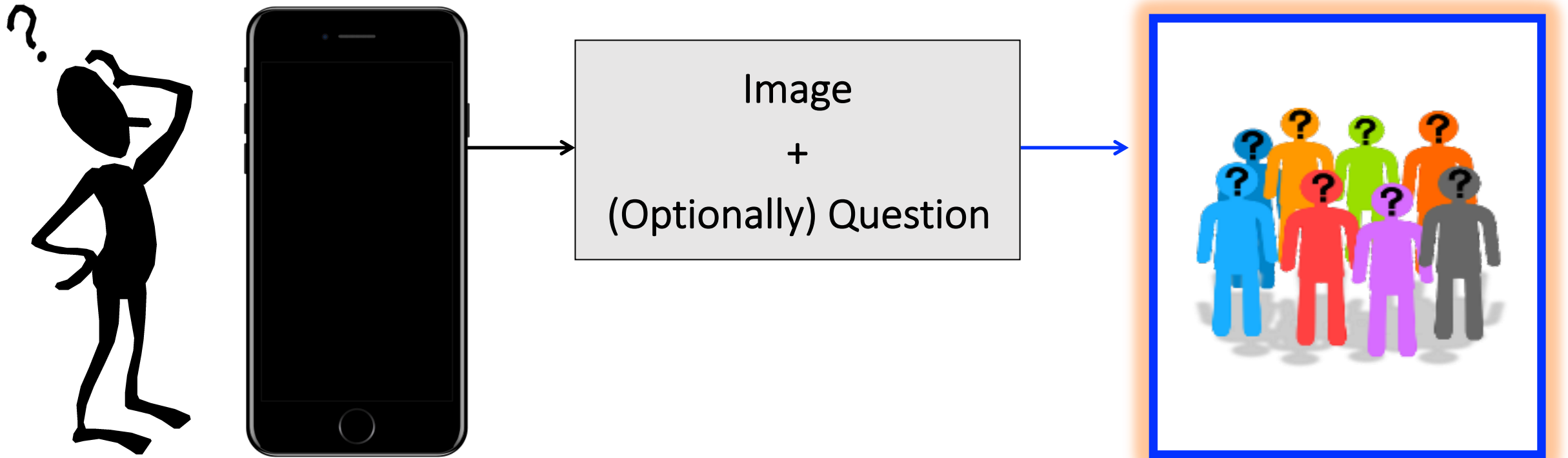
Data: Real Users of On-Demand Visual Assistance



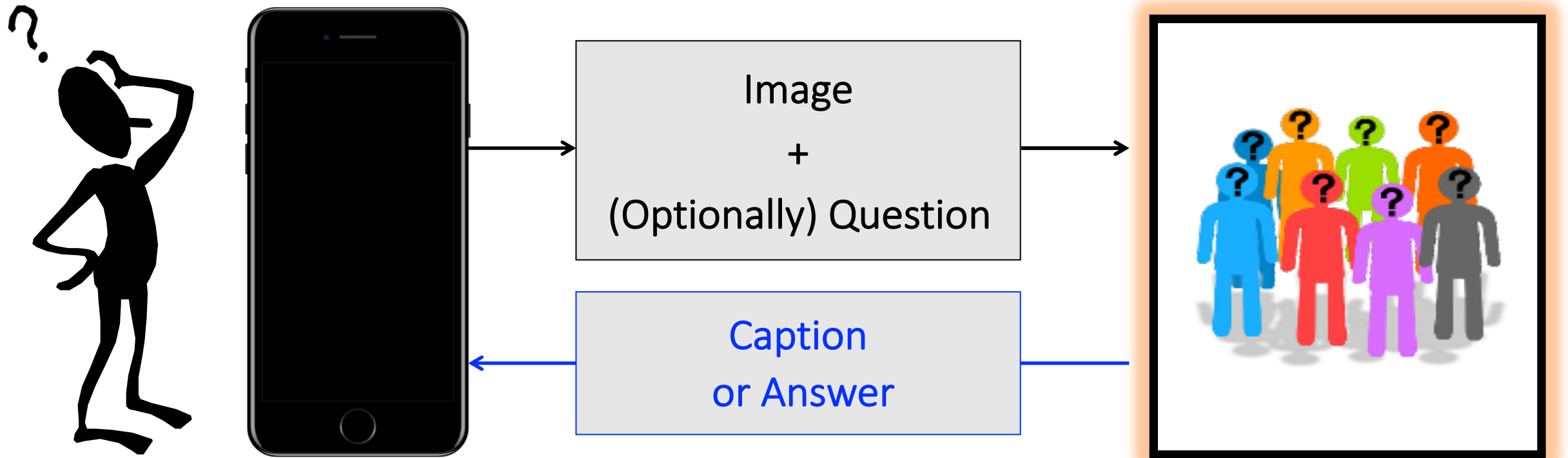
Data: Real Users of On-Demand Visual Assistance



Data: Real Users of On-Demand Visual Assistance



Data: Real Users of On-Demand Visual Assistance



Data: Real Users of On-Demand Visual Assistance

(Distribution for 31,173 questions)



VQA Datasets: VizWiz

When working with data from real users of a real-world application, what must be done differently when creating a visual question answering dataset?

VQA Datasets: VizWiz

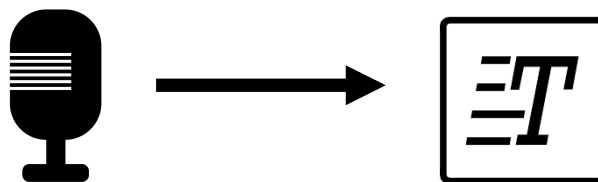
1. Collect Images & Questions

* For 44,799 of the visual questions that the users agreed to share:

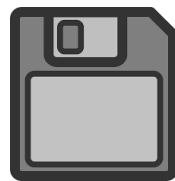
1. Anonymize the data
2. Flag all images with private information
3. Obscure all private content in images using various image inpainting algorithms

Anonymization

1. Transcribe questions



2. Re-save images



Flag private images



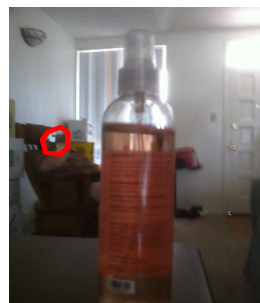
VQA Datasets: VizWiz

1. Collect Images & Questions

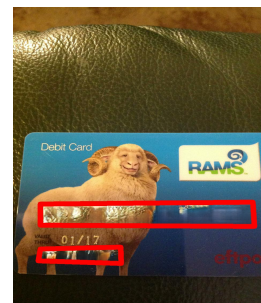
* For 44,799 of the visual questions that the users agreed to share:

1. Anonymize the data
2. Flag all images with private information
3. Obscure all private content in images using various image inpainting algorithms

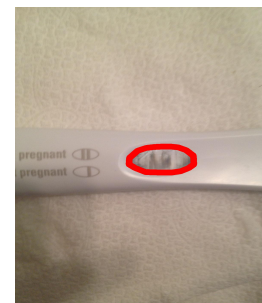
12% of 44,799 images show private content!



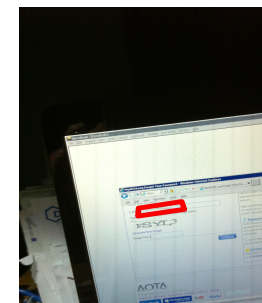
Face



Credit
Card



Pregnancy
Test



Computer
Screen



Prescription
Pills

VQA Datasets: VizWiz

1. Collect Images & Questions

* For 44,799 of the visual questions that the users agreed to share:

1. Anonymize the data
2. Flag all images with private information
3. Obscure all private content in images using various image inpainting algorithms

3. Collect Answers

* Collect 10 answers per visual question

VQA Datasets: VizWiz

Help Us Answer Questions About Images!

Your work will help to build machines that automatically answer questions asked by blind people about the visual world. In particular, you will work with images taken by blind people paired with questions they asked about the images.

Please answer the questions about the images **with brief answers**. Your answers should be how most other people would answer the questions. If the question doesn't make sense, please try your best to answer it and indicate via the buttons that you are unsure of your response.

Warning: It is possible that some images and/or questions could be inappropriate or offensive. This is because we cannot control what pictures are taken and what questions are asked.

If you don't follow the following instructions, your work will be rejected.



Your work **will get rejected** if you do not follow the instructions below:

- Answer the question based on what is going on in the **scene depicted in the image**.
- Your answer should be a **brief phrase** (not a complete sentence).
 - "It is a kitchen." -> "kitchen"
- For yes/no questions, please **just say yes/no**.
 - "You bet it is!" -> "yes"
- For numerical answers, please use **digits**.

- If an image is too poor in quality to answer the question (i.e., all white, all black, or too blurry), please say **"Unsuitable Image"**.
 - **Insufficient image quality** -> "Unsuitable Image"
- If the question cannot be answered from the image, please say **"Unanswerable"**.
 - **Question unrelated to image** -> "Unanswerable"

- provide an answer **that most people would agree on**.
- If you don't know the answer (e.g., specific dog breed), provide **your best guess**.
- Respond **matter-of-factly** and **avoid using conversational language** or **inserting your opinion**.

Please answer the question using as few words as possible:

Q1: What color socks are these?

A1: Answer question 1 here.

Do you think you were able to answer the question correctly?

(Clicking an option will take you to the next question.)

Prev

no

maybe

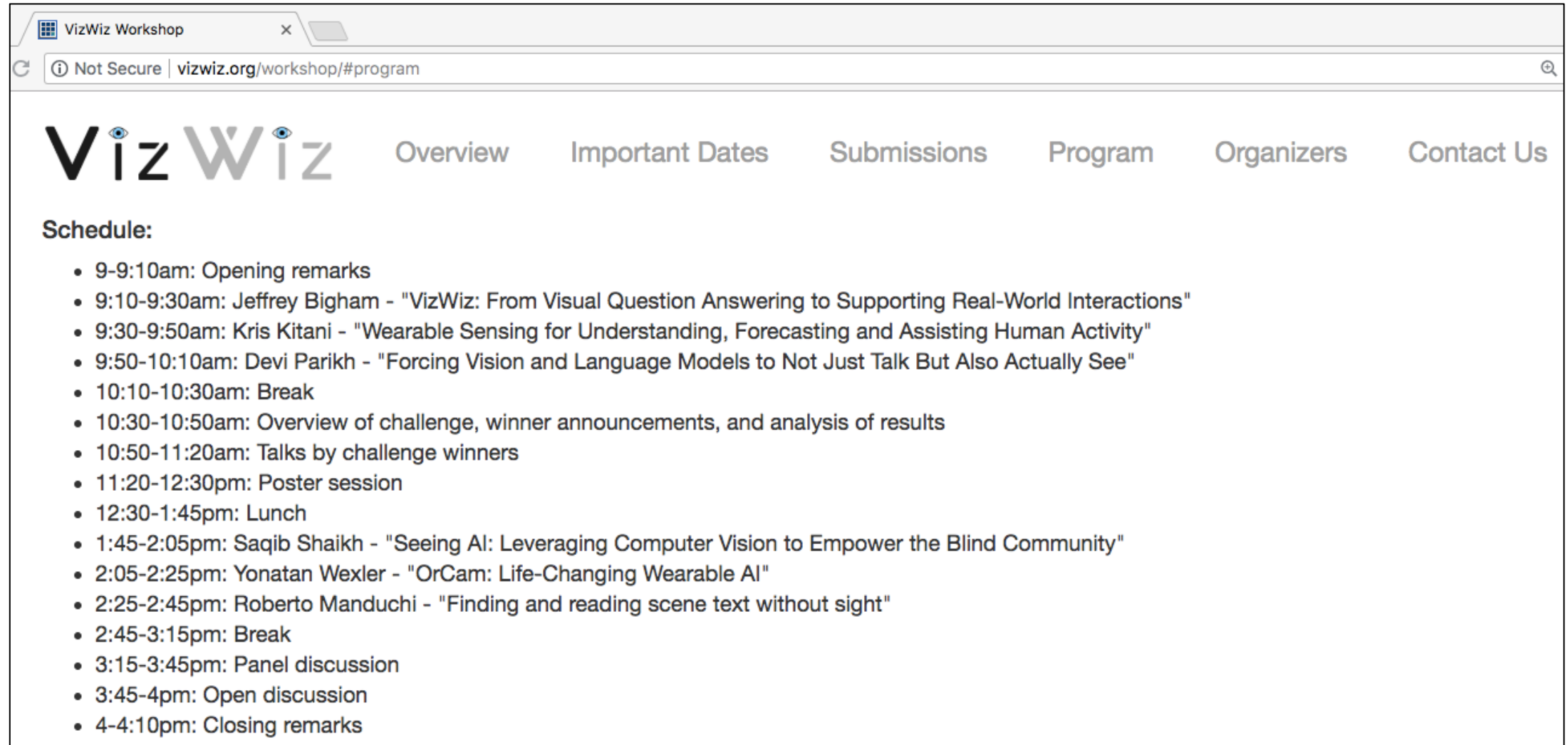
yes

Page 2/10

3. Collect Answers

* Collect 10 answers per visual question

VizWiz Grand Challenge

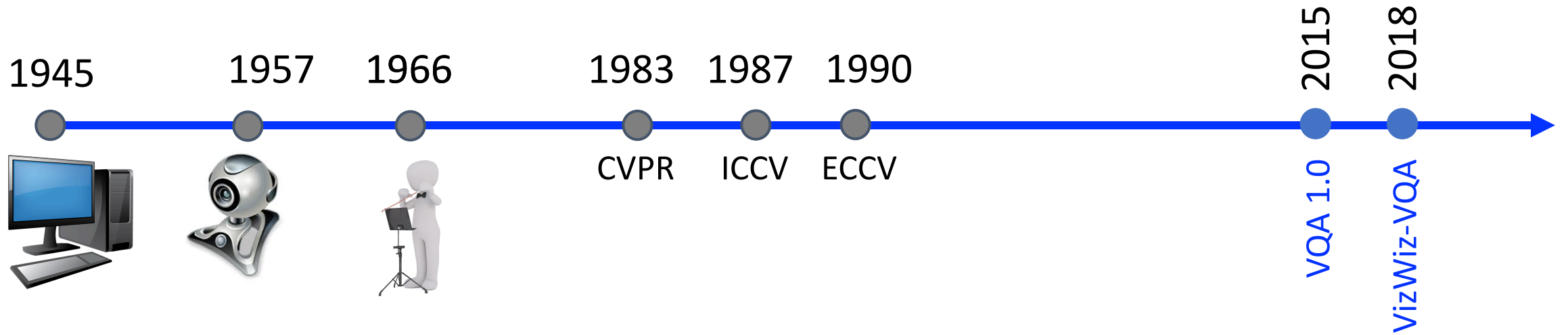


The screenshot shows a web browser window with the address bar displaying "vizwiz.org/workshop/#program". The page features the VizWiz logo and a navigation menu with links for Overview, Important Dates, Submissions, Program, Organizers, and Contact Us. Below the navigation, the "Schedule:" section lists the following events:

- 9-9:10am: Opening remarks
- 9:10-9:30am: Jeffrey Bigham - "VizWiz: From Visual Question Answering to Supporting Real-World Interactions"
- 9:30-9:50am: Kris Kitani - "Wearable Sensing for Understanding, Forecasting and Assisting Human Activity"
- 9:50-10:10am: Devi Parikh - "Forcing Vision and Language Models to Not Just Talk But Also Actually See"
- 10:10-10:30am: Break
- 10:30-10:50am: Overview of challenge, winner announcements, and analysis of results
- 10:50-11:20am: Talks by challenge winners
- 11:20-12:30pm: Poster session
- 12:30-1:45pm: Lunch
- 1:45-2:05pm: Saqib Shaikh - "Seeing AI: Leveraging Computer Vision to Empower the Blind Community"
- 2:05-2:25pm: Yonatan Wexler - "OrCam: Life-Changing Wearable AI"
- 2:25-2:45pm: Roberto Manduchi - "Finding and reading scene text without sight"
- 2:45-3:15pm: Break
- 3:15-3:45pm: Panel discussion
- 3:45-4pm: Open discussion
- 4-4:10pm: Closing remarks

<https://vizwiz.org/workshops/vizwiz-workshop-2018/>

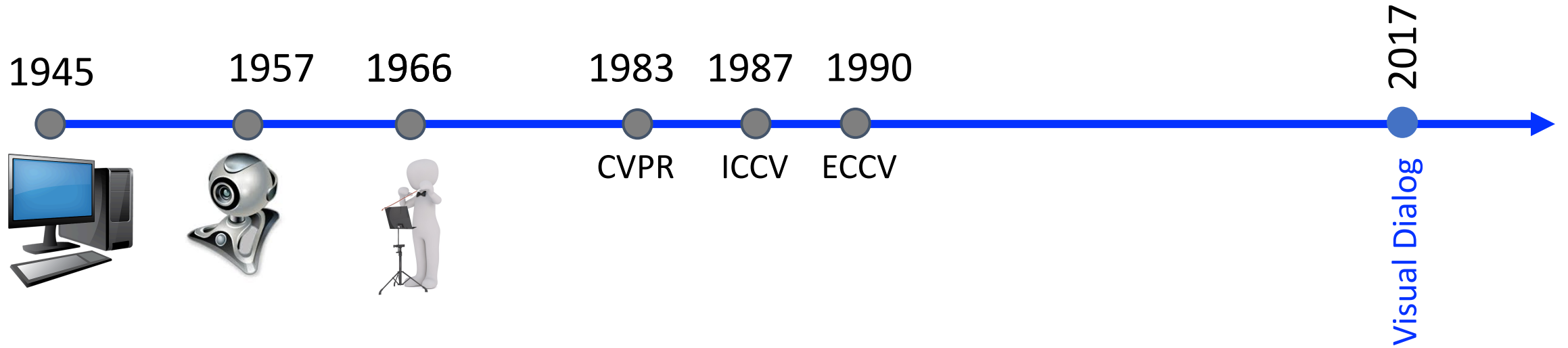
Visual Question Answering Datasets



Today's Topics

- Visual question answering and dialog applications
- Evaluation for visual question answering
- Crowdsourcing for visual question answering
- **Crowdsourcing for visual dialog**
- Guest: General Manager Warren Barkley at Amazon Web Services

Visual Question Answering Datasets



VQA: Crowdsourcing Instructions

Live Question/Answering about an Image.

▼ Instructions

In this task, you will be talking to a fellow Turker. You will either be asking questions or answering questions about an image. You will be given more specific instructions once you are connected to a fellow Turker.

Stay tuned. A message and a beep will notify you when you have been connected with a fellow Turker.

Please keep the following in mind while chatting with your fellow Turker:

- 1 Please directly start the conversation. Do not make small talk.
- 2 Please do not write potentially offensive messages.
- 3 Please do not have conversations about something other than the image. Just either ask questions, or answer questions about an image (depending on your role).
- 4 Please do not use chat/IM language (e.g, "r8" instead of "right"). Please use professional and grammatically correct English.
- 5 **Please have a natural conversation. Unnatural sounding conversation including awkward messages and long silences will be rejected.**
- 6 Please note that you are expected to complete and submit the hit in one go (once you have been connected with a partner). You cannot resume hits.
- 7 **If you see someone who isn't performing HITS as per instructions or is idle for long, do let us know. We'll make sure we keep a close watch on their work and reject it if they have a track record of not doing HITS properly or wasting too much time. Make sure you include a snippet of the conversation and your role (questioner or answerer) in your message to us, so we can look up who the other worker was.**
- 8 **Do not wait for your partner to disconnect to be able to type in responses quickly, or your work will be rejected.**

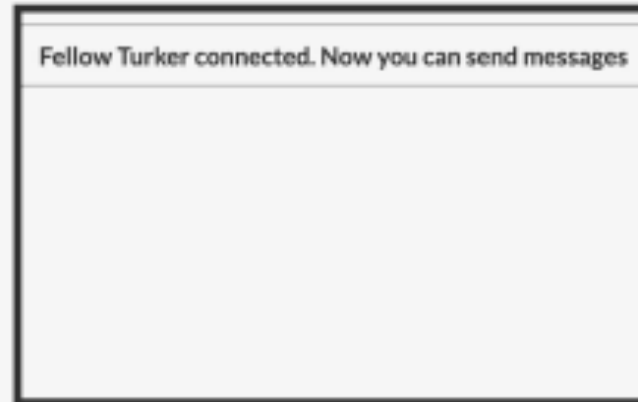
Please complete one hit before proceeding to the other. Please don't open multiple tabs, you cannot chat with yourself.

VQA: Asking Crowdsourcing Interface

Caption: A man, wearing goggles and a backpack on skis pulls a girl on skis behind him.

You have to ASK Questions about the image.

Fellow Turker connected. Now you can send messages



Type Message Here:

Message

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra; Visual Dialog. CVPR 2017.

VQA: Answering Crowdsourcing Interface

Caption: A man, wearing goggles and a backpack on skis pulls a girl on skis behind him.

You have to ANSWER questions about the image.



Fellow Turker connected. Now you can send messages

Type Message Here:

Send

End Conversation And Finish Hit

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra; Visual Dialog. CVPR 2017.

Crowdsourcing Task

Caption: A sink and toilet in a small room.

You have to ASK questions about the image.

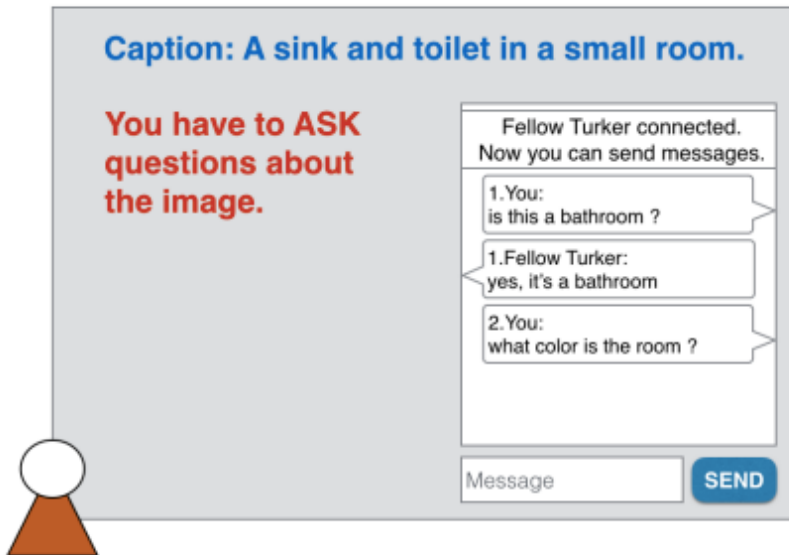
Fellow Turker connected.
Now you can send messages.

1.You:
is this a bathroom ?

1.Fellow Turker:
yes, it's a bathroom

2.You:
what color is the room ?

Message **SEND**



(a) What the 'questioner' sees.

Caption: A sink and toilet in a small room.

You have to ANSWER questions about the image.

Fellow Turker connected.
Now you can send messages.

1.Fellow Turker:
is this a bathroom ?

1.You:
yes, it's a bathroom

2.Fellow Turker:
what color is the room ?

2.You:
it looks cream colored

Message **SEND**



(b) What the 'answerer' sees.



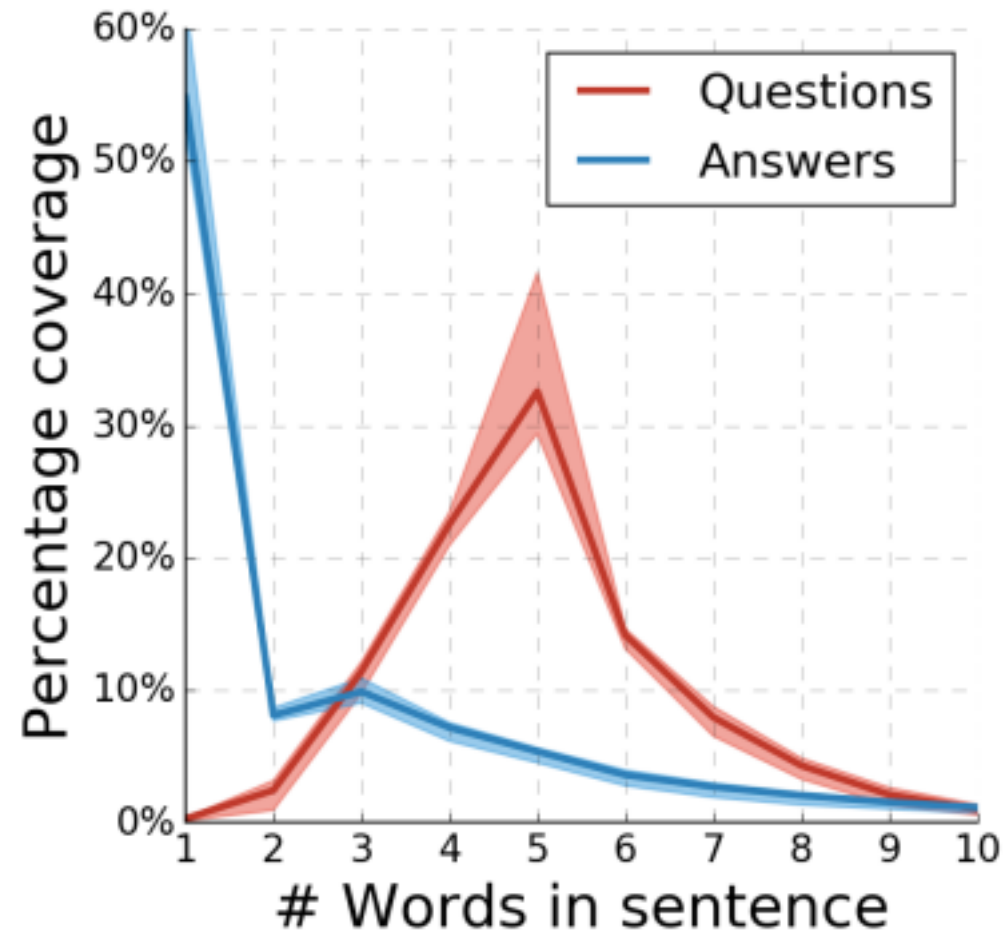
Caption:
A sink and toilet in a small room.

- Q3: can you see anything else ?
A3: there is a shelf with items on it
Q4: is anyone in the room ?
A4: nobody is in the room
Q5: can you see on the outside ?
A5: no, it is only inside
Q6: what color is the sink ?
A6: the sink is white
Q7: is the room clean ?
A7: it is very clean
Q8: is the toilet facing the sink ?
A8: yes the toilet is facing the sink
Q9: can you see a door ?
A9: yes, I can see the door
Q10 what color is the door ?
A10 the door is tan colored

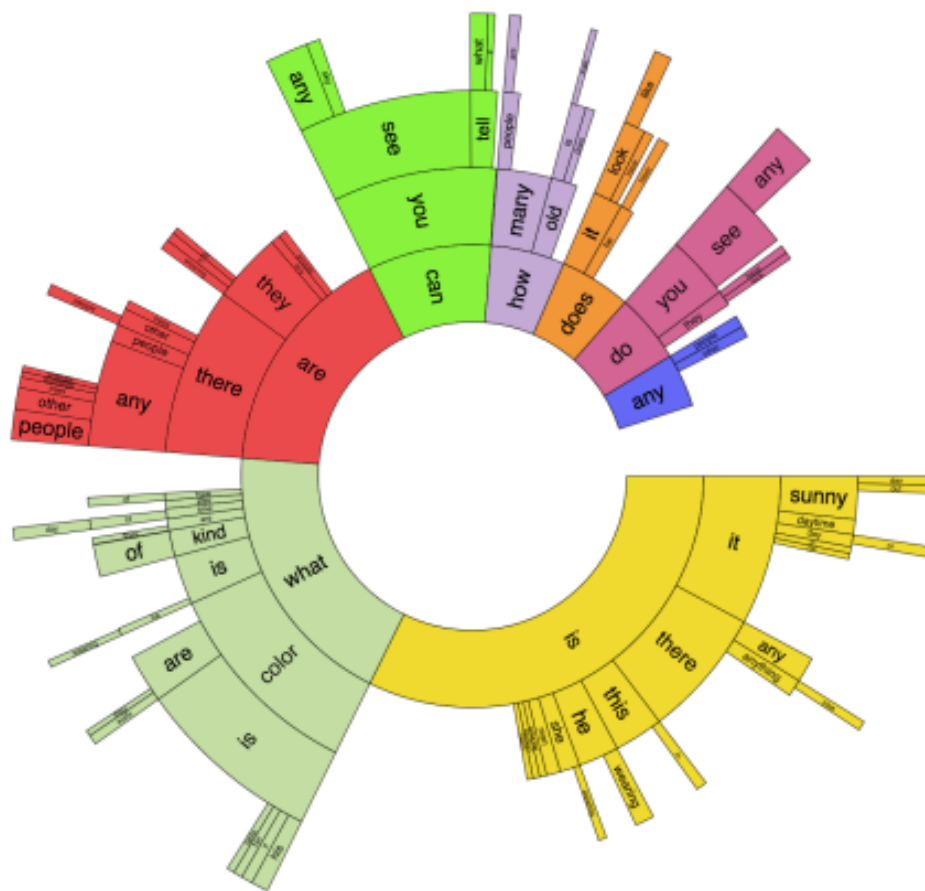
(c) Example dialog from our VisDial dataset.

Workers can end a conversation after 20 messages are exchanged (10 question-answer pairs)

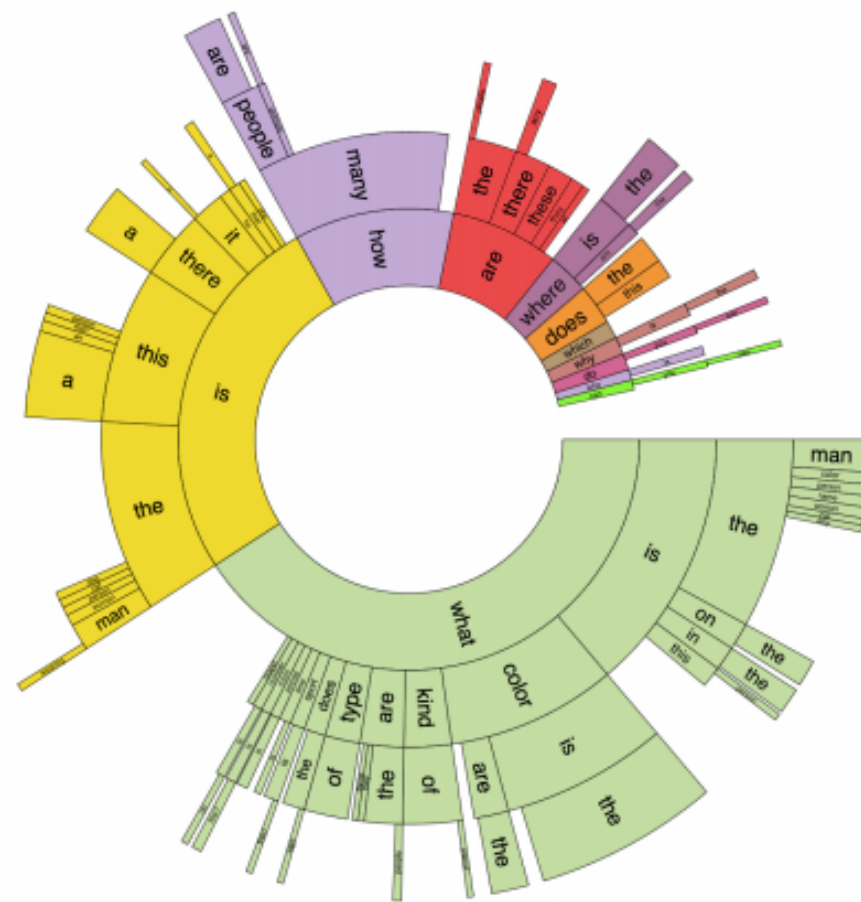
Distribution of Question & Answer Lengths



Popular Question Words/Phrases



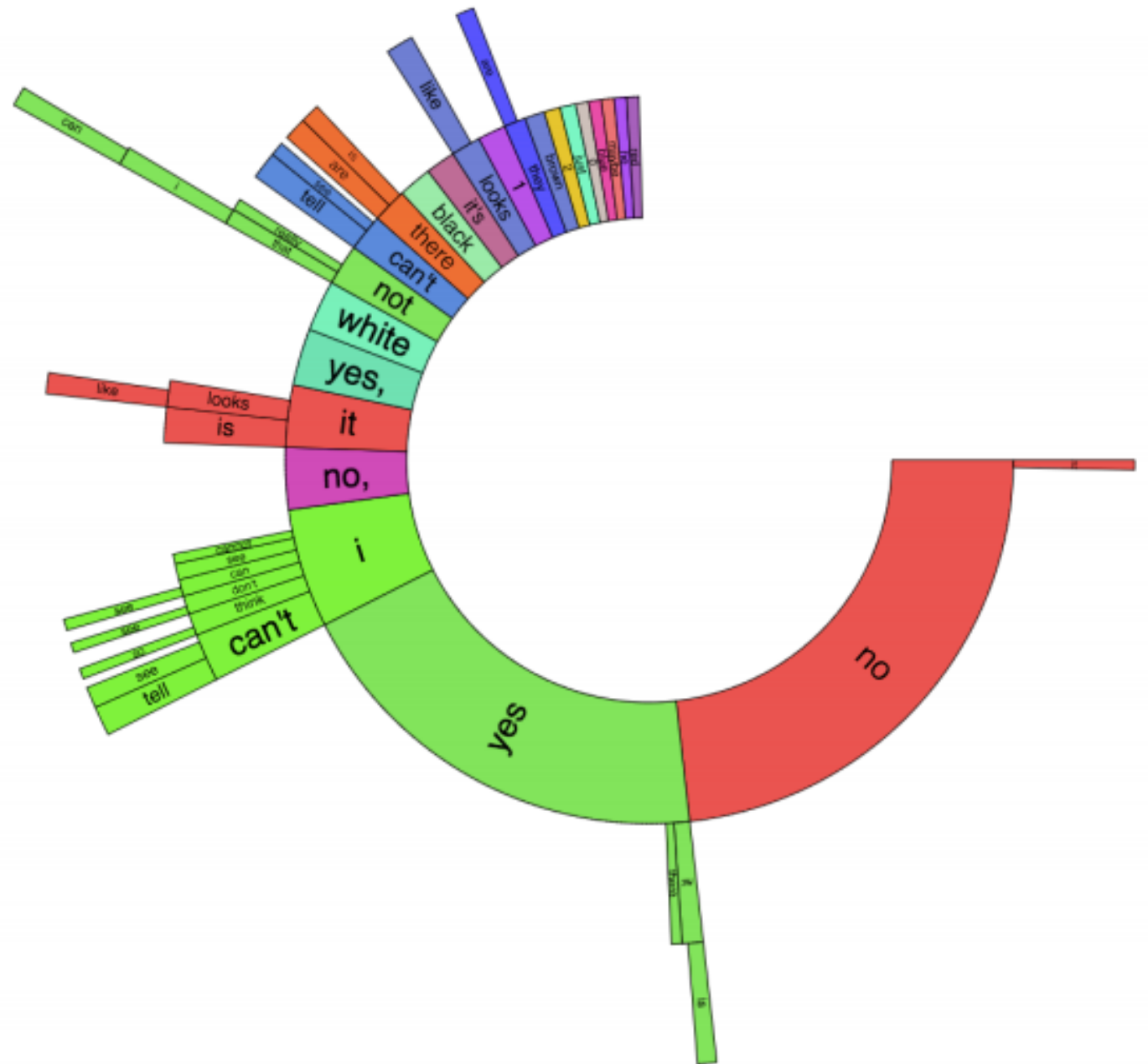
(a) VisDial Questions



(b) VQA Questions

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra; Visual Dialog. CVPR 2017.

VisDial Answers



Class Discussion

1. Why do different crowd workers' answers differ for a visual question?
2. How would you decide what answer you use when different crowd workers provide different answers to a visual question? Please note your method must scale to efficiently support large datasets.
3. All crowd workers were restricted to US locations. How might different cultural backgrounds affect VQA datasets?

Today's Topics

- Visual question answering and dialog applications
- Evaluation for visual question answering
- Crowdsourcing for visual question answering
- Crowdsourcing for visual dialog
- **Guest: General Manager Warren Barkley at Amazon Web Services**