Natural Language Processing

- A Survey Talk @ MSR UW Summer Institute 2016 -

Margaret Mitchell

Cognition Group

Microsoft Research

Yejin Choi

Computer Science & Engineering

UNIVERSITY of WASHINGTON

Accessibility in the Past…

Reading Machine to OCR

Kurzweil Reading Machine (1976)
$30,000-$50,000

OCR & Flatbed Scanners (Today)

Text to speech (Today)

Accessibility in the Past…

Talking books to Audio Books

Talking Books (1930s)

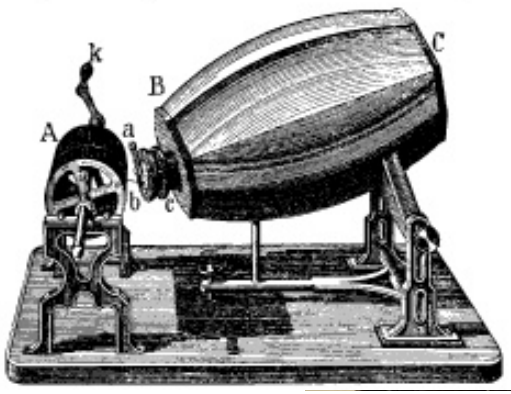
audible.com

Audio Books (Today)

http://www.silksoundbooks.com/history-of-audiobooks.html

Accessibility in the Past…

Telephone



Abraham Graham Bell's experiments with hearing devices (phonautograph) for deaf individuals (1876)

Phones (Today)

What is Novel today will become Standard tomorrow

NLP Today ...

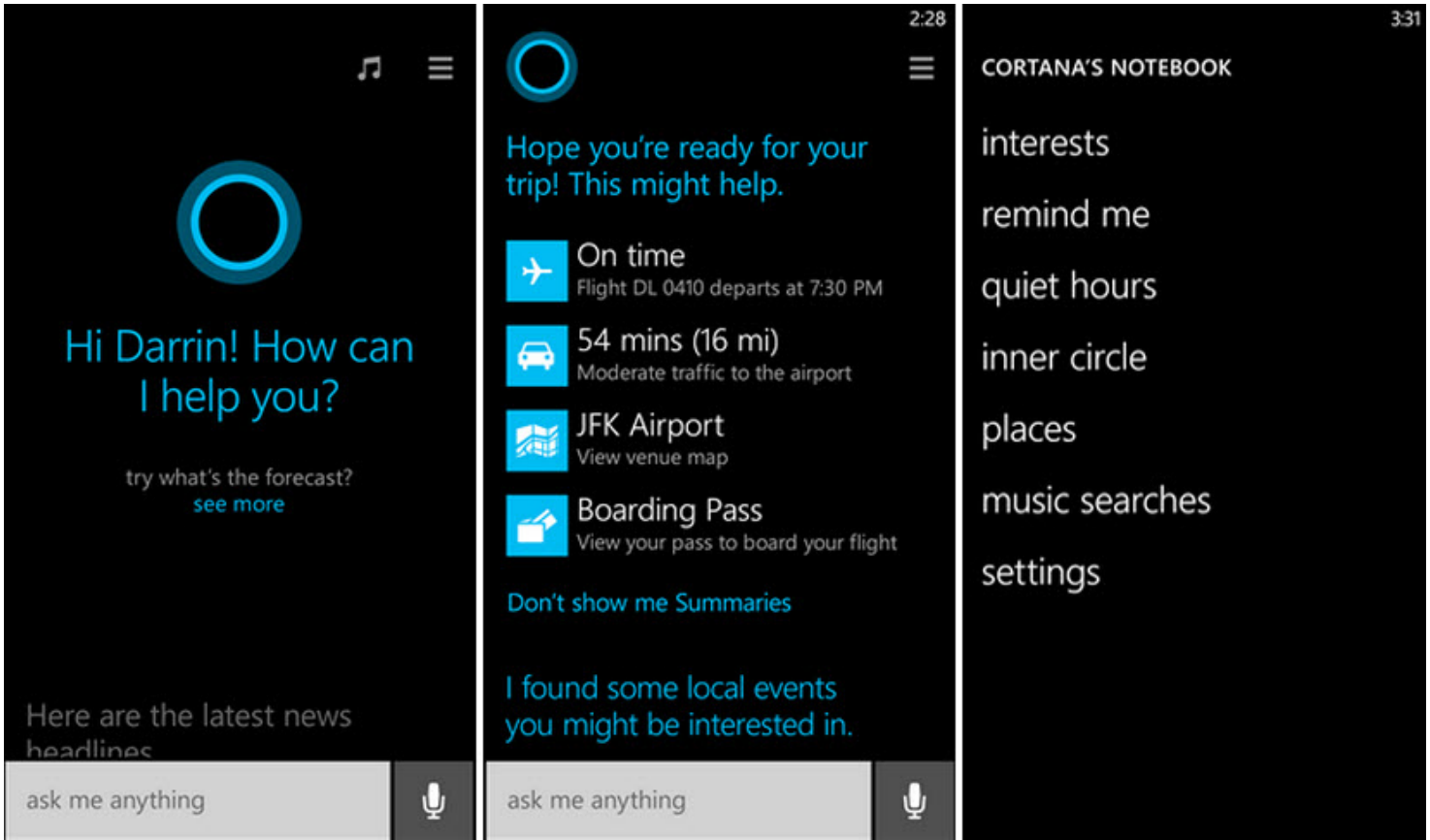
- Web search
  - -- can handle natural language queries better
  - -- often presents us structured knowledge
  - -- Research on web search started in 90's (or maybe 80's)

NLP Today ...

- Automatic translation
  - -- not perfect, but good enough for people to use
  - -- real time translation with audio
  - -- first statistical model (IBM model 1) came out in 1993
  - -- first MT service based on statistical model in 2007

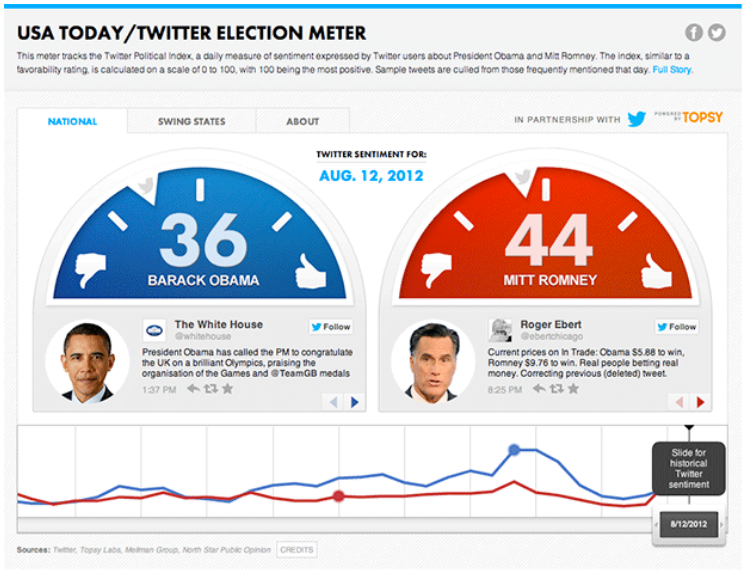
NLP Today …

- Everyday devices with natural language interface
  -- simple question answering
  -- simple natural language commands



NLP Today

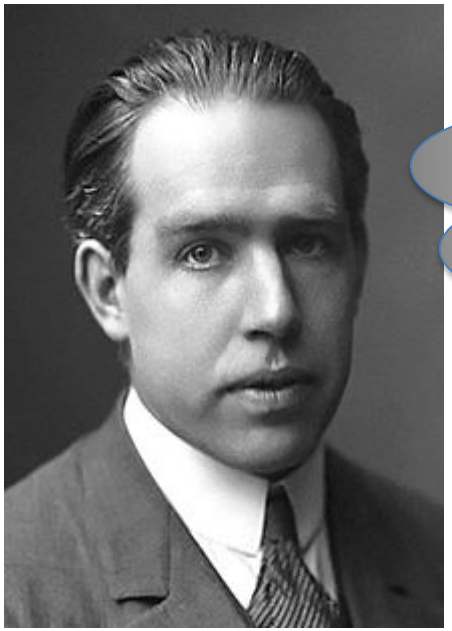
- Sentiment Analysis
  - Domains: twitter, product reviews, news trend…
  - Works well for some types of data
  - Used in 2012 election
  - Challenge: sarcasm, subtext, nuanced messages
  - First research on sentiment analysis started in 2002



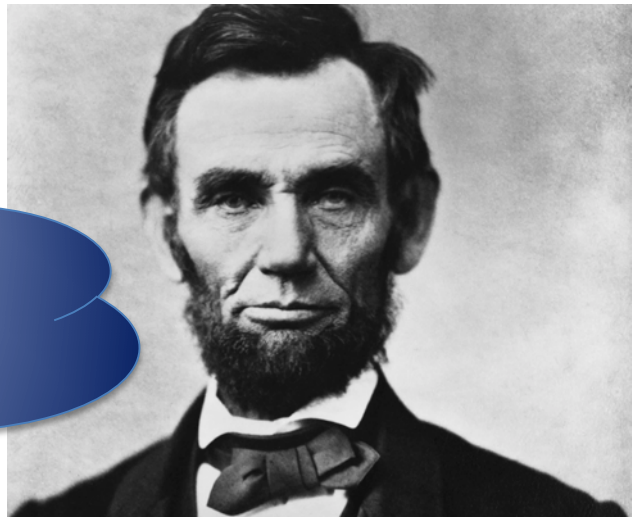
A bit of Introspection…

- A lot of exciting results
  - After a (very) long period of hard work!
- Computers do surprisingly well on some hard tasks
  - Machine translation
  - IBM Watson winning Jeopardy
- Computers do poorly on many tasks that are trivial for us.
- Computers seem to do well when
  - Lots of data for training (where low hanging fruits are)
  - The task can be solved by primarily based on pattern matching
  - (no complex reasoning, especially not commonsense reasoning)
- What are challenges that seem just impossible today but might become possible after 10 year of hard work?

What’s in the Future?

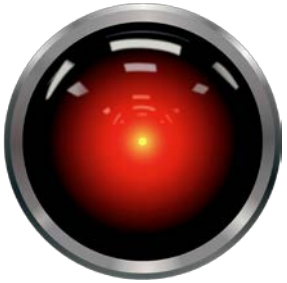


It's hard to make predictions, especially about the future.



The best way to predict your future is to invent it.

Intelligent Communication

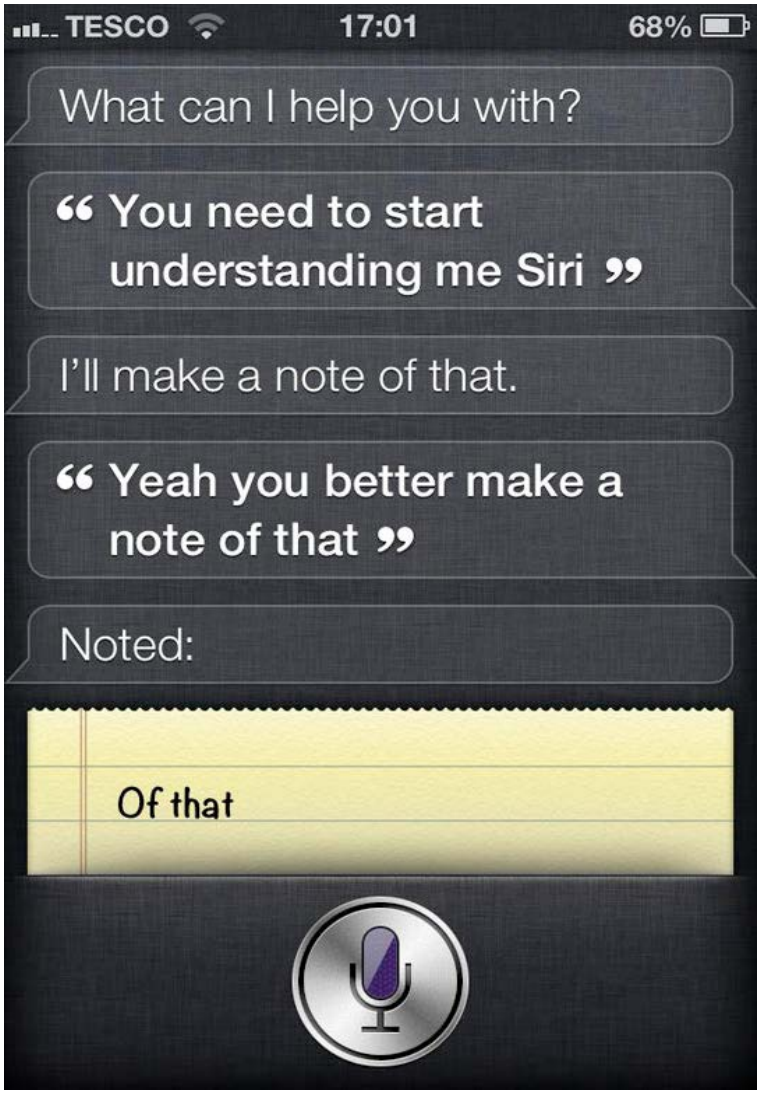


Intelligent Communication

Reading between the lines

- **Content**: *what* is said

  + *what* is not said



language in physical context

Blueberry Muffins

Ingredients
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Need **knowledge about the cooking world**!

Bake what? where?

Procedure
1. Preheat oven to 400 degrees F. Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
3. **Bake for 20 minutes**. Serve hot.

http://allrecipes.com/Recipe/Blueberry-Muffins-I/

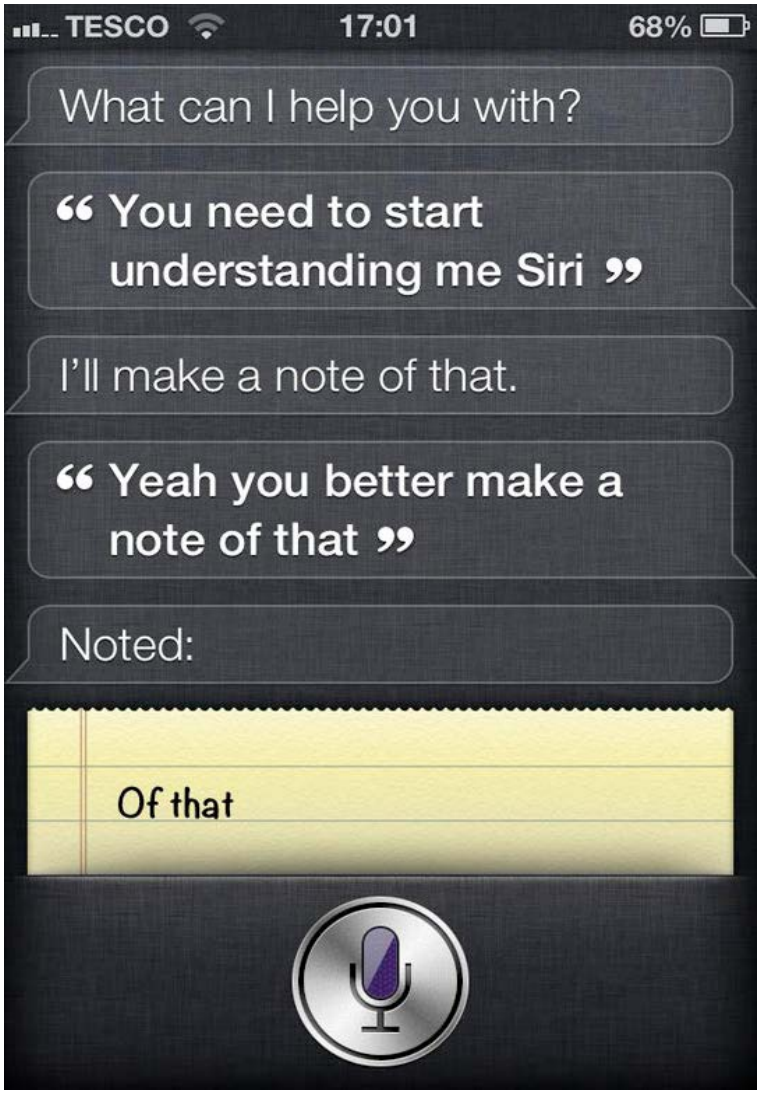
Intelligent Communication

Reading between the lines

- **Content**: *what* is said
- + *what* is not said
- **Intent**: *why* it is said

Language in the world

- in the physical / visual context
- in the social / cognitive context



NLP for Acting AI:

Natural Language Instructions

Smart devices and personal robots

executing commands in natural language instructions

Instructional Recipes

Blueberry Muffins

Ingredients
1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Procedure
1. Preheat oven to 400 degrees F. Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
3. **Bake for 20 minutes**. Serve hot.

http://allrecipes.com/Recipe/Blueberry-Muffins-I/

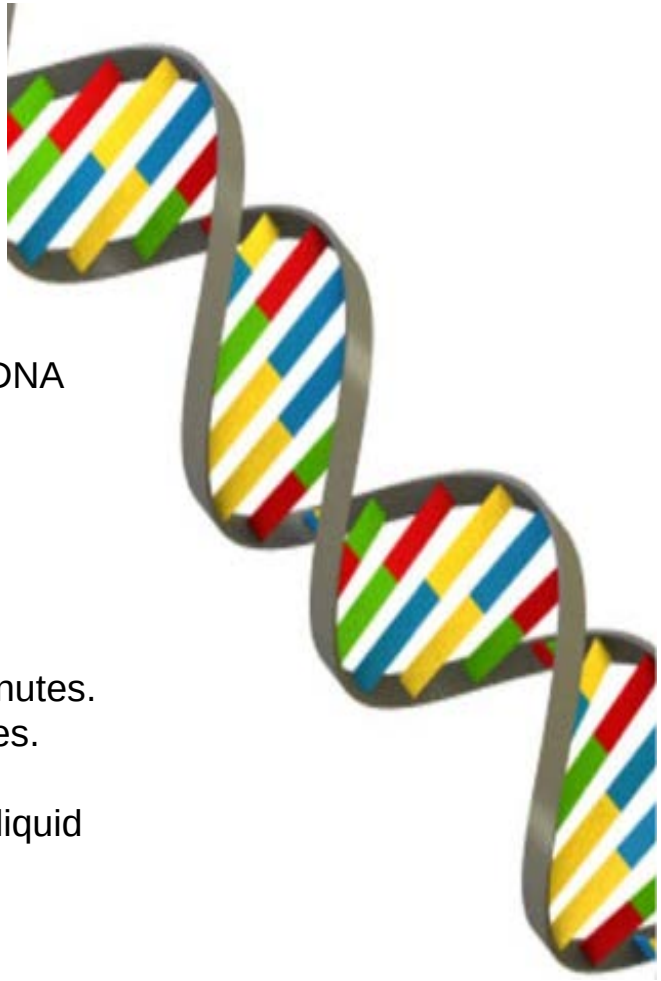
From Kitchen to Biology Labs

DNA Precipitation

Materials
3M NaOAc pH 5.2
EtOH 95%
Glycogen (optional)

Procedure
1. Add 0.1 volumes of 3M Sodium Acetate solution to 1 volume of DNA sample.
2. Add 1ul Glycogen to the DNA sample.
3. Add 2 volumes of 95% EtOH to the DNA Sample.
4. Store the solution overnight at -20°C or for 30 minutes at -80°C.
5. Centrifuge the solution at maximum speed for least 15 minutes.
6. Decant and discard the supernatant.
7. (Optional) Add 1 ml of 70% EtOH to the pellet and let sit for 5 minutes.
8. (Optional) Centrifuge the sample at maximum speed for 5 minutes.
9. (Optional) Decant and Discard the supernatant.
10. Air-dry the pellet for 10-15 minutes at room temperature until all liquid is gone.
11. Resuspend in desired volume of water or buffer

http://openwetware.org/wiki/DNA_Precipitation



Translation of Language to Action Diagrams

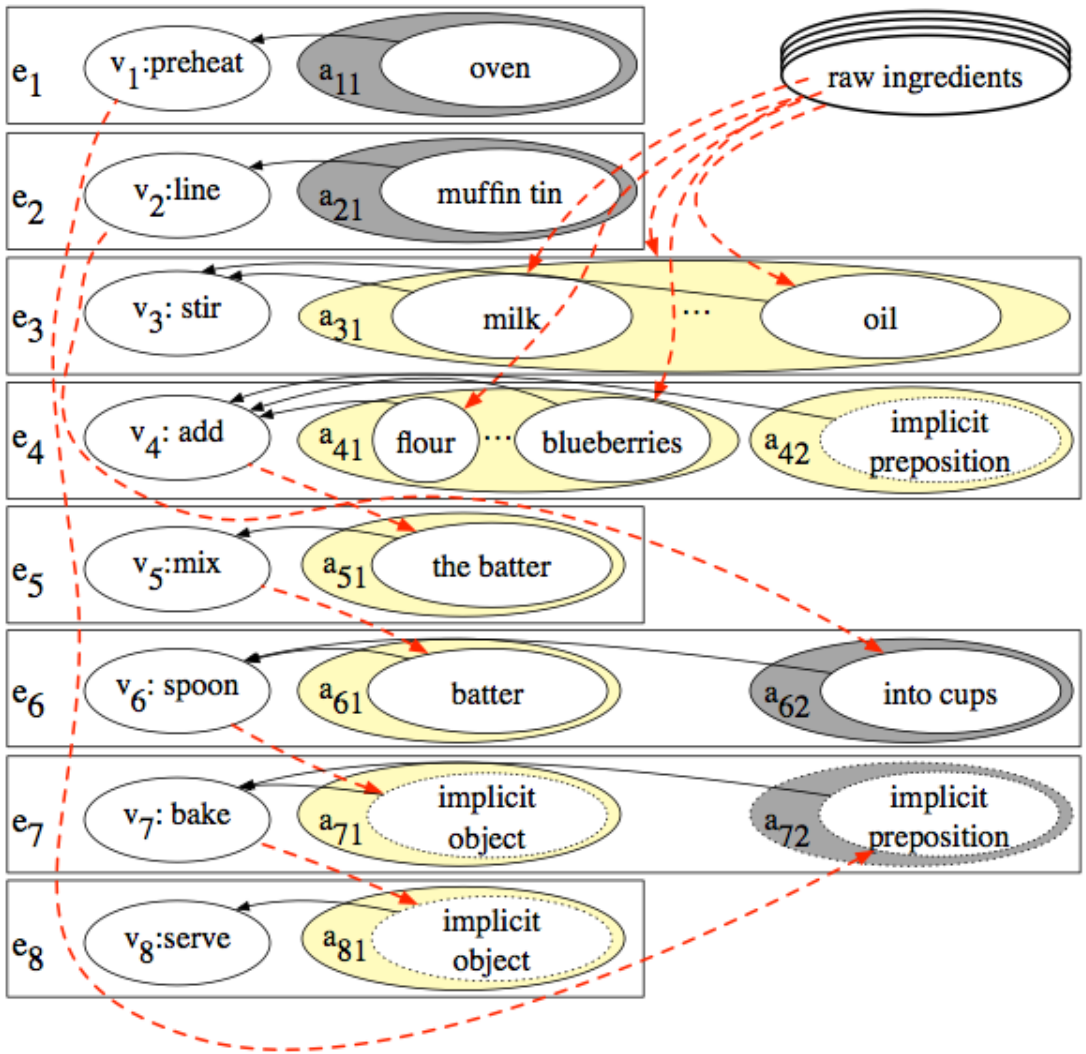
Blueberry Muffins I (http://allrecipes.com/Recipe/Blueberry-Muffins-I/)

Ingredients

1 cup milk
1 egg
1/3 cup vegetable oil
2 cups all-purpose flour
2 teaspoons baking powder
1/2 cup white sugar
1/2 cup fresh blueberries

Directions

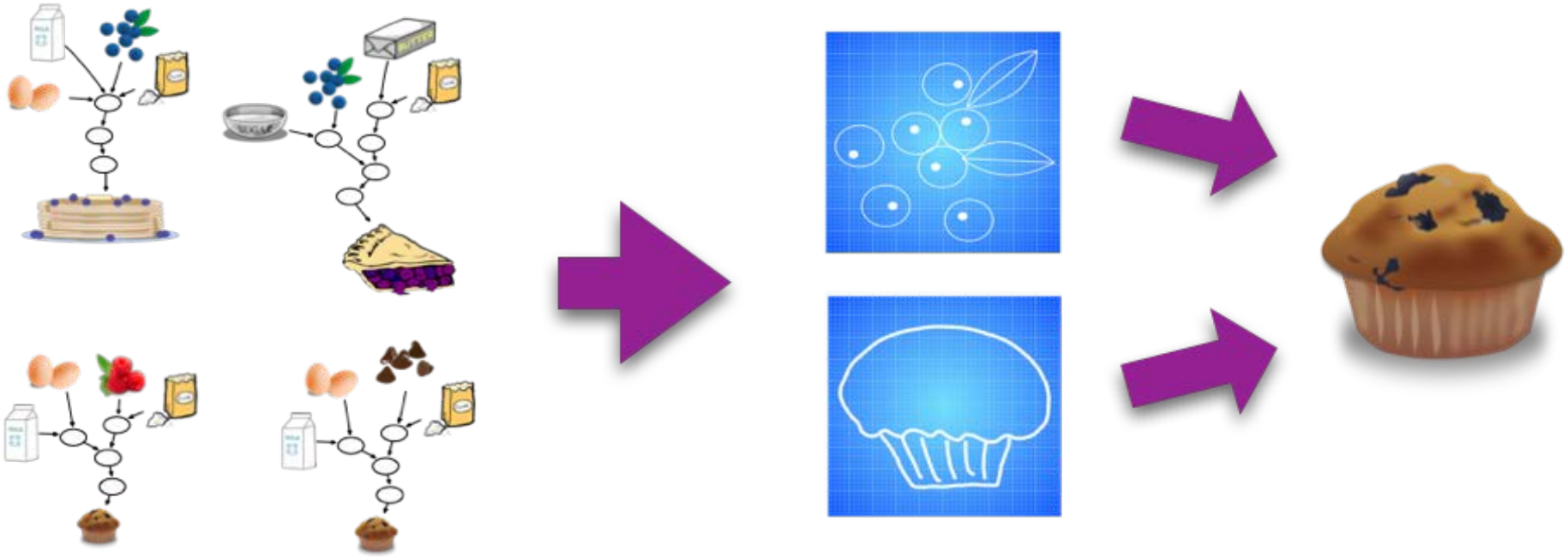
1. Preheat oven to 400 degrees F (205 degrees C). Line a 12-cup muffin tin with paper liners.
2. In a large bowl, stir together milk, egg, and oil. Add flour, baking powder, sugar, and blueberries; gently mix the batter with only a few strokes. Spoon batter into cups.
3. Bake for 20 minutes. Serve hot.



What’s Next: Composing a New Recipe

Compose new recipes given a recipe title (or what’s in the fridge)!

Grounding instructions with multimodal perception



NLP for Seeing AI

Seeing AI : How Deep Learning can Empower the Blind Community

Margaret Mitchell, Anirudh Koul, Saqib Shaikh
Microsoft

Points of Contact: seeingai@microsoft.com

Margaret Mitchell

memitc

Microsoft Research

Anirudh Koul

akoul

Bing

Saqib Shaikh

saqibs

Bing

Plan

- Accessibility Accelerates Innovation
- Backend Research Technology
  - Cognitive Services
    - Object Detection and Image Captioning

Microsoft

Accessibility Accelerates Innovation

Our Mission

Empower every person and every organization on the planet to achieve more

Approach to Advancing AI

1. Make AI vastly capable
2. Make vastly capable AI beneficial

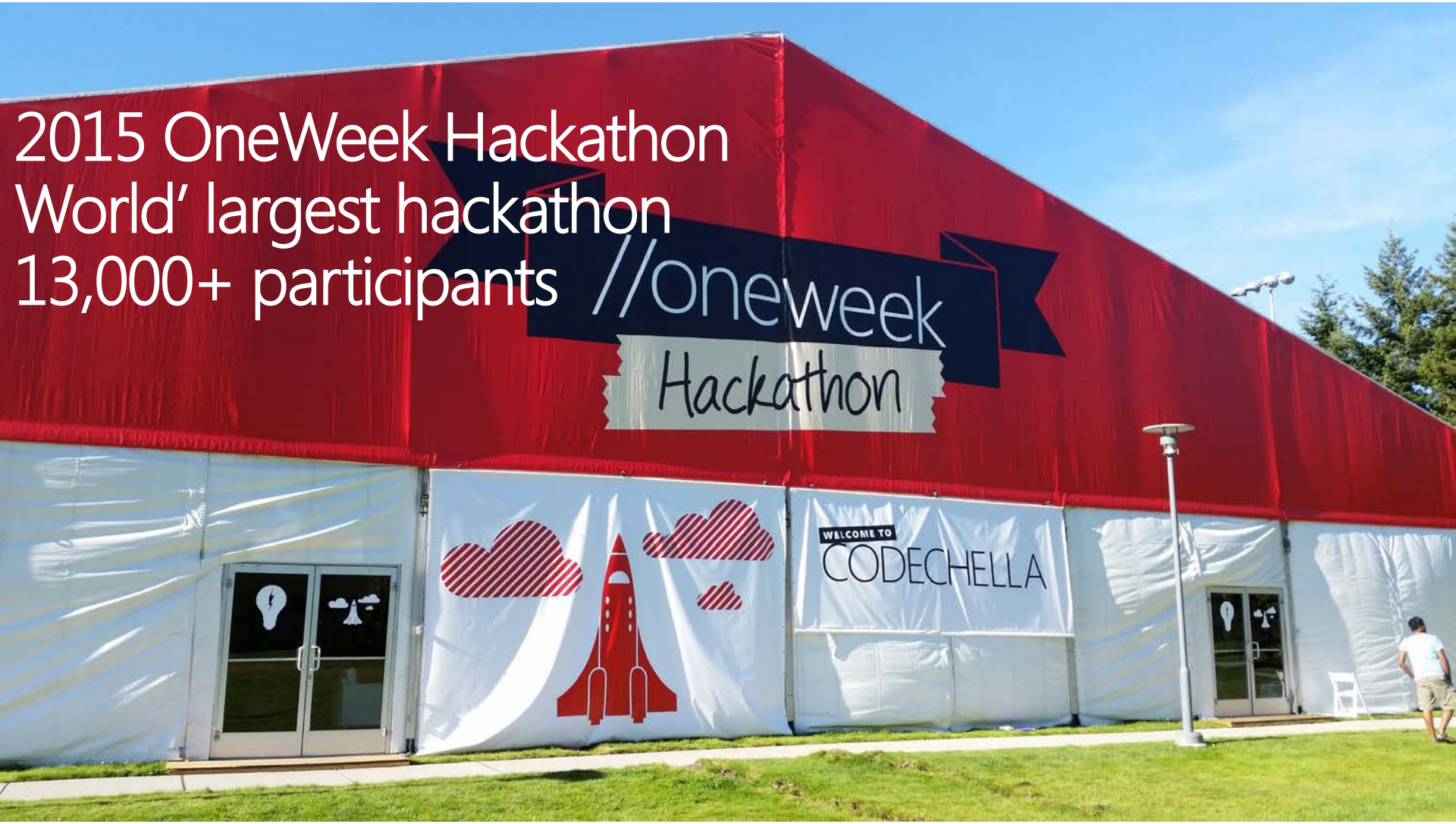
Microsoft

Approach to Advancing AI

~~1. Make AI vastly capable~~
~~2. Make vastly capable AI beneficial~~

1. Make AI beneficial
2. Make beneficial AI vastly capable

Microsoft



2015 OneWeek Hackathon
World' largest hackathon
13,000+ participants

Hack Team

Silicon Valley

Redmond

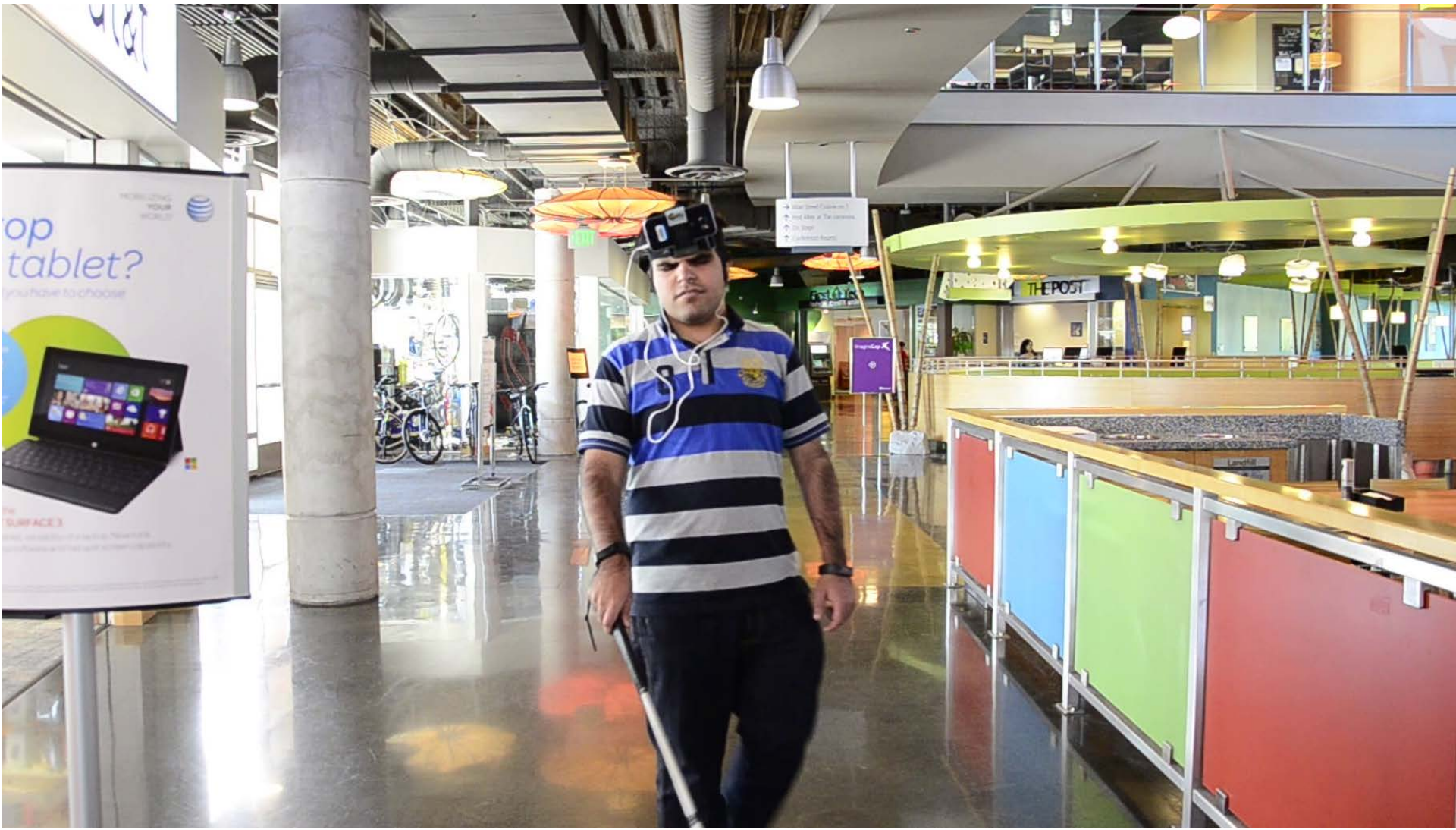
London

TT

Tanya Tran – July 23, 2015 at 5:09pm

Just saw someone at SVMT Cafe working on his hack, something with his phone strapped to his head. Can't wait to see what it's all about next Thursday at the Science Fair!

LIKE REPLY SHARE ...


op
tablet?
THE POST

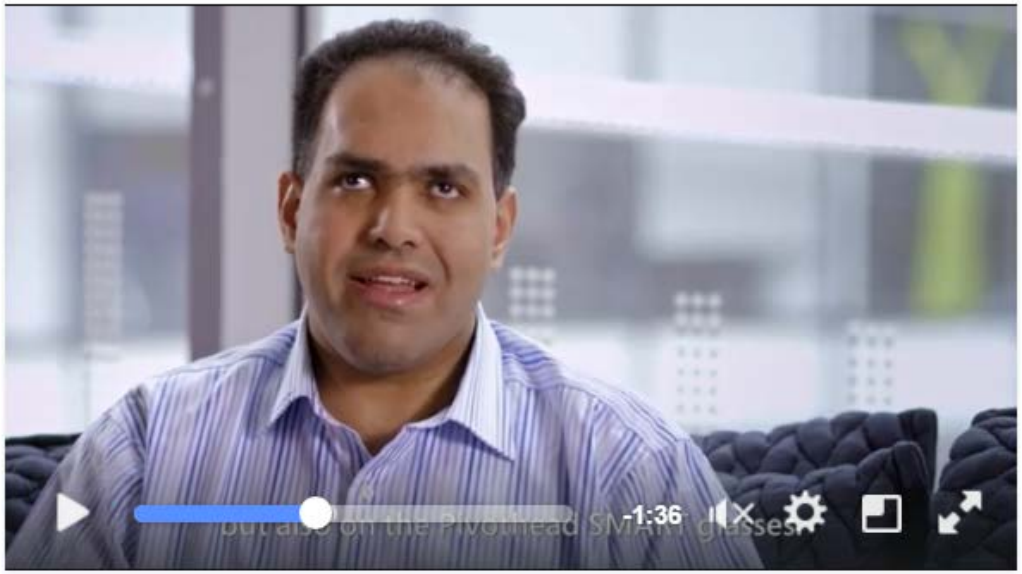
Build Conference 2016

"We as developers have this tremendous opportunity and tremendous responsibility, because not only do we get to dream of the future, we get to build the future"

Video

BuzzFeed Tech
March 30 at 3:00pm ·

By far the coolest thing Microsoft announced at its Build conference: Using artificial intelligence to help the blind see.



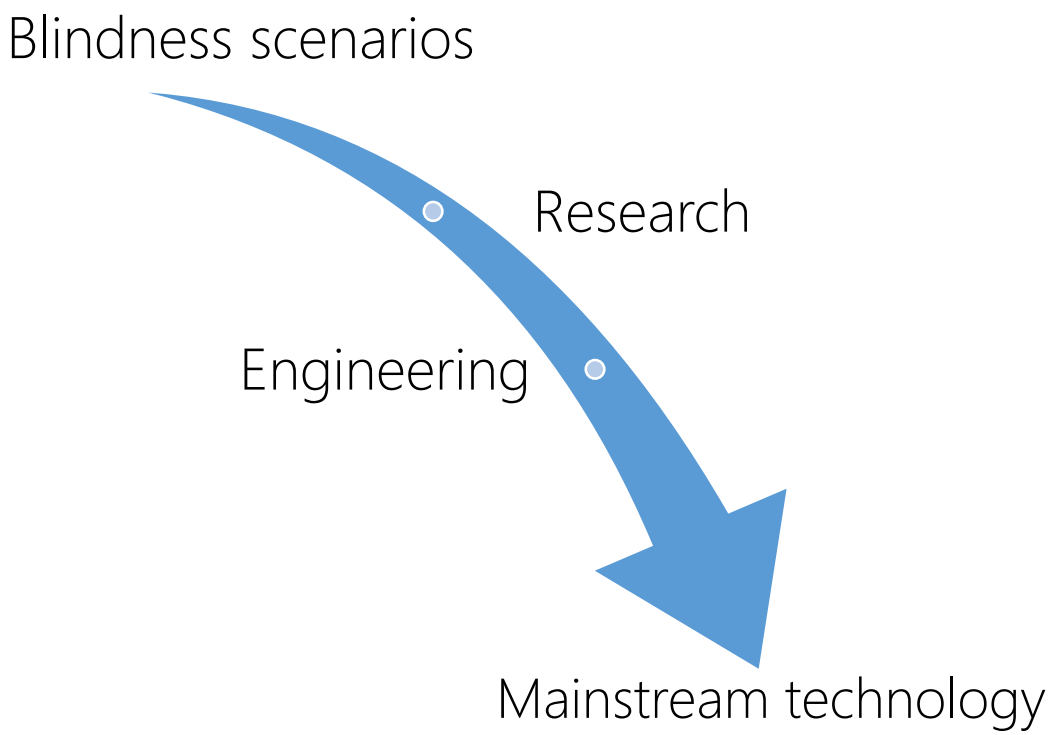
1.2M Views

Like Comment Share

You, Arbind Koul, Ranjitha Kulkarni and 11K others

Top Comments

Our team cycle



Seeing AI

A research project at Microsoft

Microsoft

Cognitive Services

The Connectors:

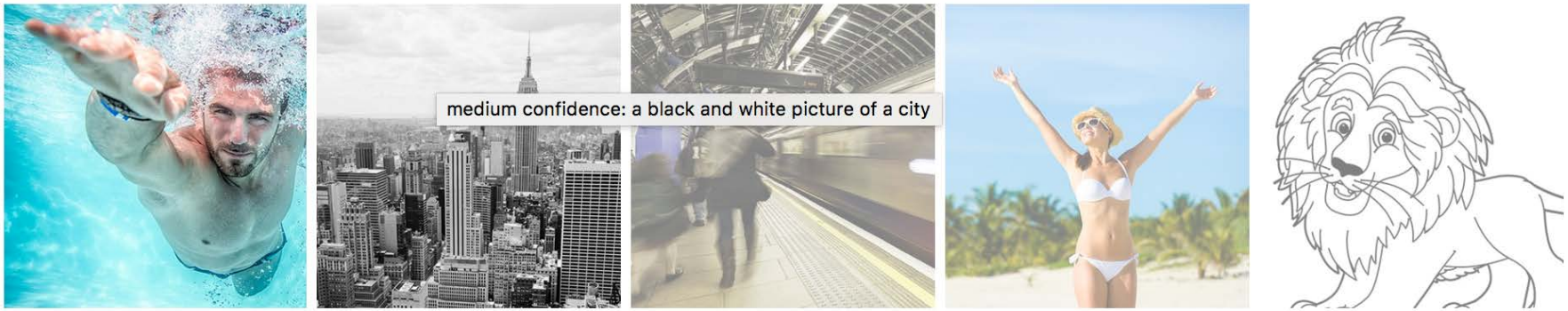
- Cornelia Carapcea,
  - Cognitive Services Vision APIs
- Margaret Mitchell
  - Microsoft Research

Seeing AI Experiences

IOT Devices like
Pivothead Live Streaming glasses

Phone app

Seeing AI Experiences for Screen Readers



Console Sources Network Timeline Profiles Resources Security Audits Adblock Plus HTTPS Everywhere

```
▼ <div id="Analysis" class="vision-demo scroll_item recaptcha-demo recaptcha-visionAnalysis-demo" style="position: relative;">
  ▼ <div data-demo-type="detection" class="detection-demo">
      <input type="hidden" data-name="demoId" value="vision_computervision_featureanalysis">
      <input type="hidden" data-name="demoResult" value>
    ▶ <div class="demo-content">…</div>
    ▶ <div class="detection-demo-right">…</div>
      <div class="clear"></div>
    ▼ <div class="image-list">
      ▶ <div class="image-item" data-selected="selected" style="opacity: 1">…</div>
      ▼ <div class="image-item" style="opacity: 1;">
          <img data-large-src="https://projectoxfordportal.azureedge.net/vision/Analysis/2.jpg" src="https://
          projectoxfordportal.azureedge.net/vision/Analysis/2-1.jpg" alt="medium confidence: a black and white picture of a city" title=
          "medium confidence: a black and white picture of a city">
        </div>
      ▶ <div class="image-item" style="opacity: 0.5;">…</div>
      ▶ <div class="image-item">…</div>
      ▶ <div class="image-item">…</div>
      ▶ <div class="image-item">…</div>
    </div>
```

Backend Technology

Object Detection and Image Captioning

Object Detection and Image Captioning

- Vision: Deep Convolutional Neural Network (CNN)
  - Object detection accuracy has shot up in past few years
    - Due to neural networks
  - Groups all over the world beginning to connect:

    Newly accurate vision detections

    . . . to . . .

    Language models that can use the visual representations
- Language: Several models with strong performance

  1. Multimodal recurrent neural network language model
  2. Multiple instance learning w/ maximum entropy language model

Microsoft

Vision Model

- Vision model: Multiple-layer network (16; recently 152)
  - Pre-trained on 1.2 million image, 1000-class ImageNet data set
  - Fine-tuned on the MS COCO training data
  - State-of-the-art accuracy

4096-dim final feature vector
fc7

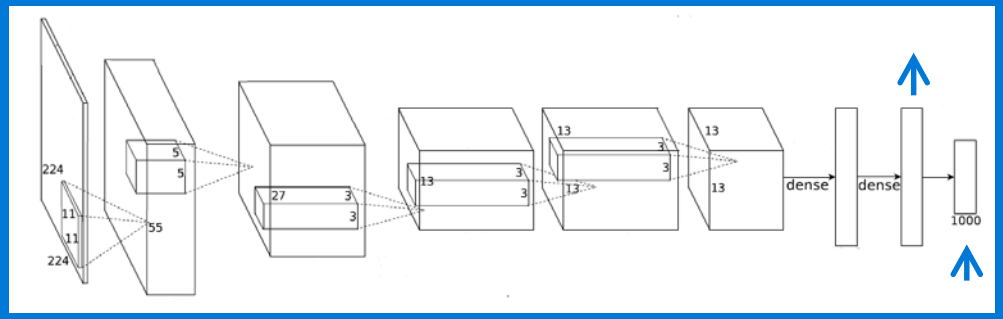
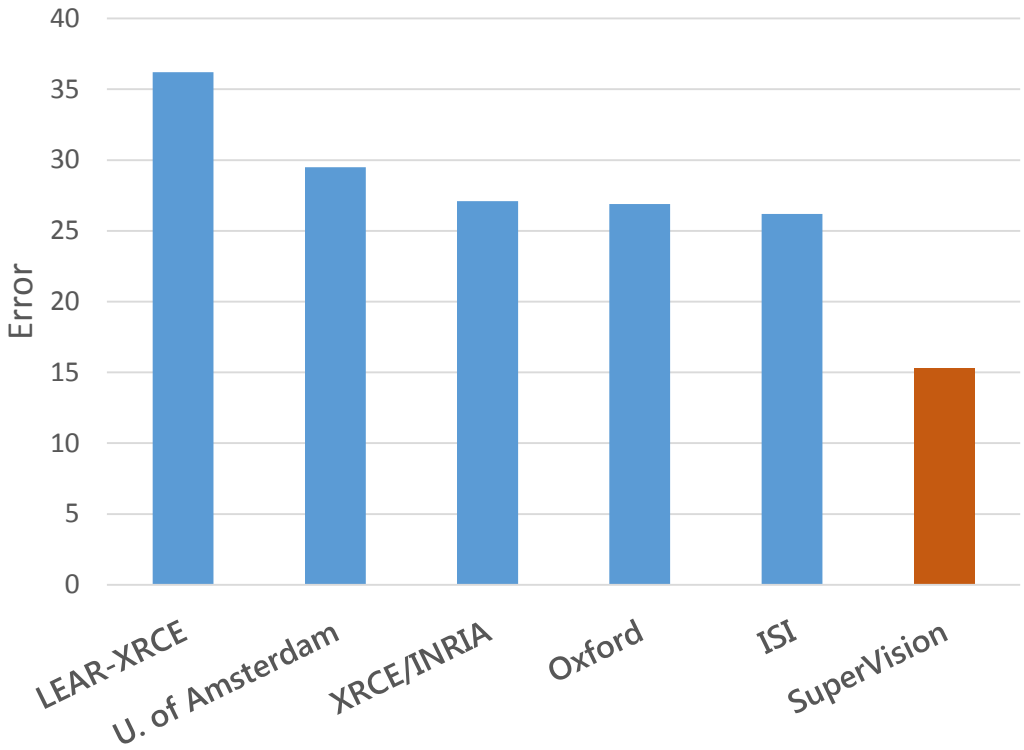


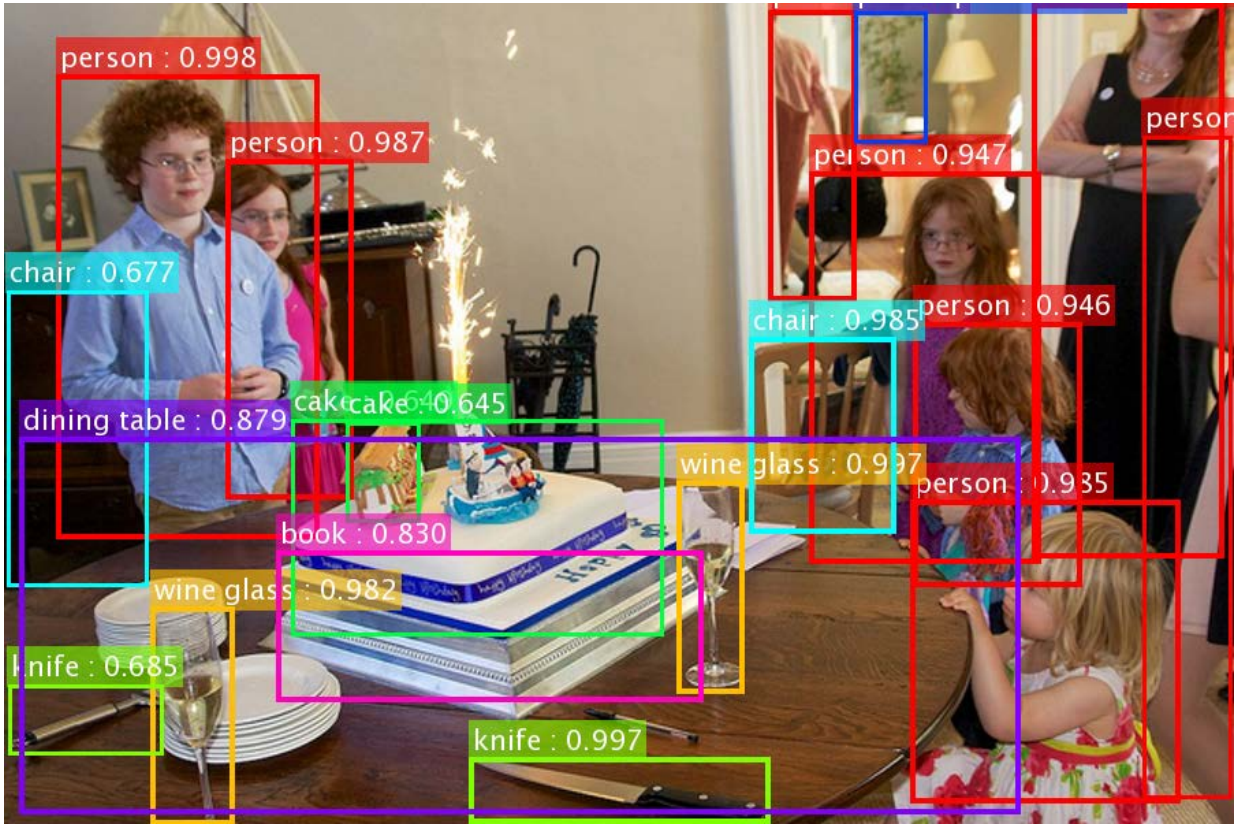
Image labels

Image Credit: Krizhevsky et al. 2012

Microsoft

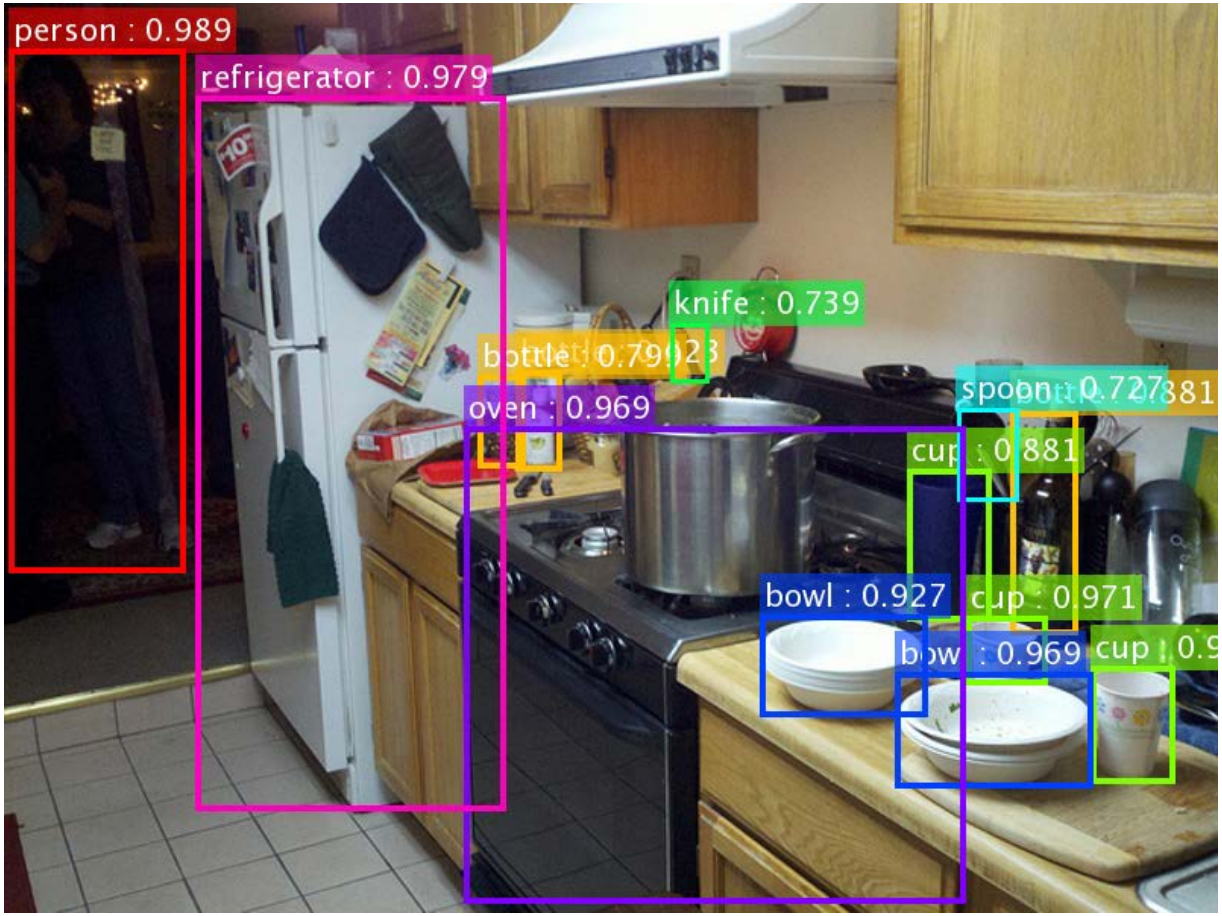
2012 ImageNet 1K Challenge





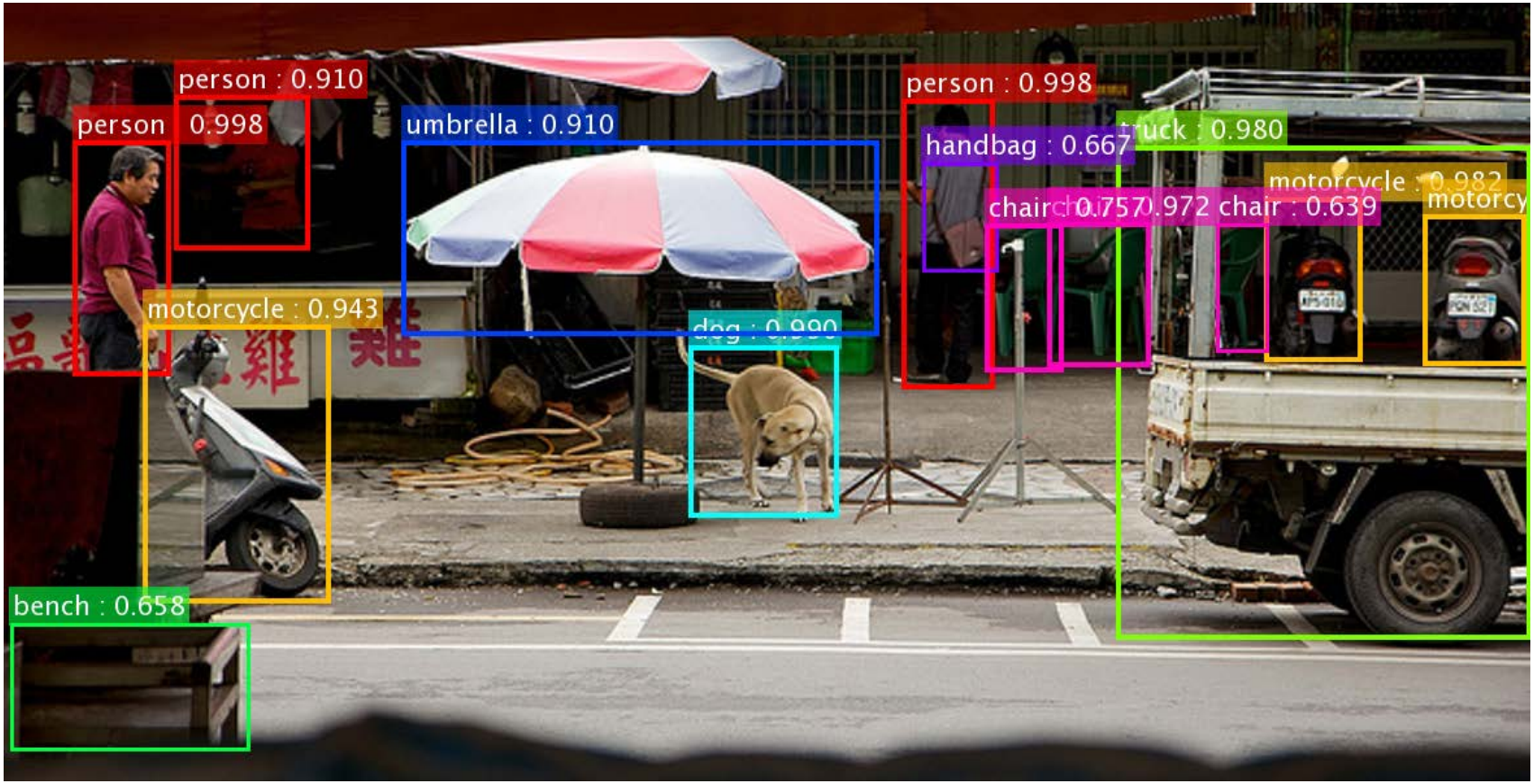
*the original image is from the COCO dataset

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. NIPS 2015.



*the original image is from the COCO dataset

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. NIPS 2015.



*the original image is from the COCO dataset

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. NIPS 2015.

Microsoft Captioning Model

Image Captioning

- Vision: Deep Convolutional Neural Network (CNN)
  - Object detection accuracy has shot up in past few years
    - Due to neural networks
  - Groups all over the world beginning to connect:
    Newly accurate vision detections
    . . . to . . .
    Language models that can use the visual representations
- Language: Several models with strong performance
  1. Multiple instance learning with maximum entropy language model
  2. Multimodal recurrent neural network language model
  3. Nearest neighbor

Microsoft

Vision Model

- Vision model: Multiple-layer network (16; recently 152)
  - Pre-trained on 1.2 million image, 1000-class ImageNet data set
  - Fine-tuned on the MS COCO training data
  - State-of-the-art accuracy

4096-dim final feature vector
fc7

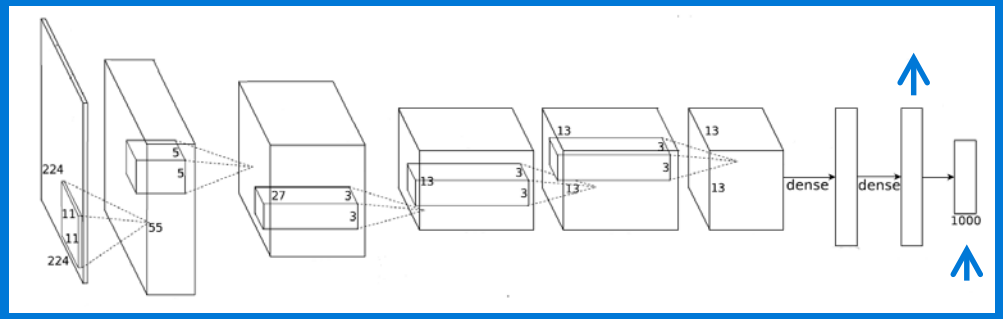


Image labels

Image Credit: Krizhevsky et al. 2012

Microsoft

Captioning Model 1

1: Multimodal Recurrent Neural Network

- Use fc7 as initial state in recurrent neural network language model
- Words output in sequence

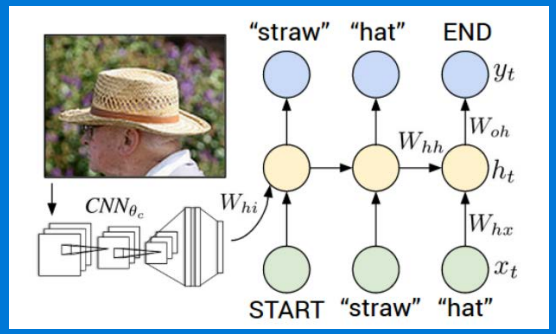


Image Credit: Karpathy and Fei-Fei 2015

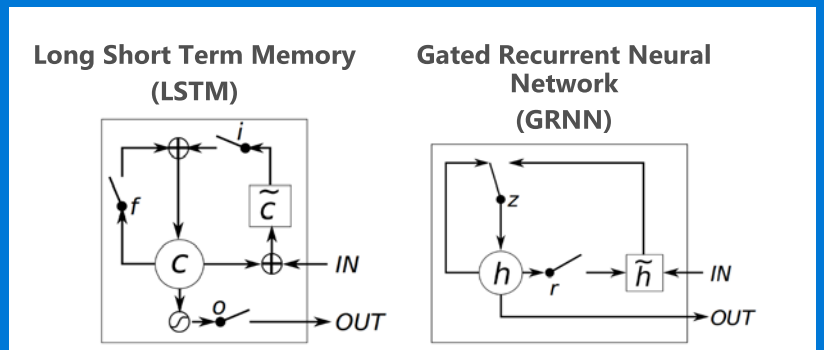


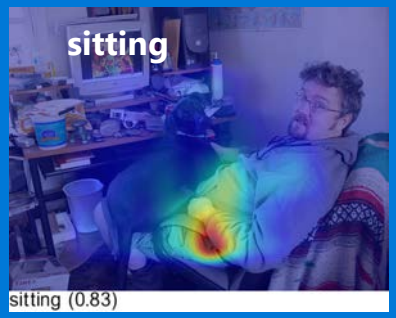
Image Credit: Cho et al. 2015

- Used by: Google, Baidu, Stanford, Berkeley

Microsoft

Captioning Model 2:
The Microsoft System

1: Multiple Instance Learning (MIL)



$$p_i^w = 1 - \prod_{j \in r_i} (1 - \sigma(f_{ij} \cdot v_w))$$

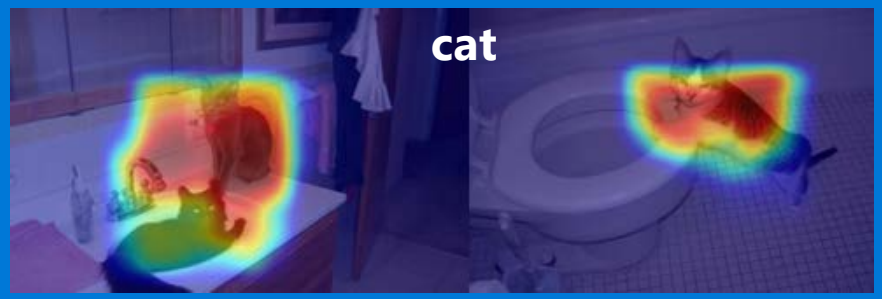
i = image id
f_{ij} = fc7 vector
$\sigma(x)$ = sigmoid
r_i = regions
v_w = learned classifier weights

- Treat training caption as bag of image labels
- Train one binary classifier per label on all images
- Noisy-Or classifier
  - Image divided into 12x12, 6x6, 3x3 label for whole image
  - fc7 overlapping regions
  - Votes from several regions used to predict vector used for image features

Microsoft

2: Map features to likely image words

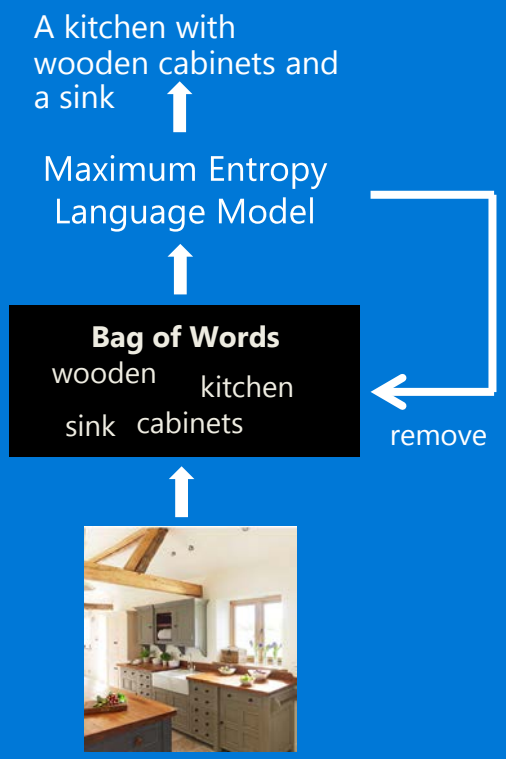
- Compute global precision threshold on held-out training subset
- Output all words $\tilde{V}$ with precision of τ or higher



	Metric	Noun	Verb	Adjective
Human Agreement	PHR	63.8	35.1	35.9
Classification	PHR	45.3	31.0	37.1
MIL NOR	PHR	**51.6**	**33.3**	**44.3**

	Metric	Noun	Verb	Adjective
Chance	AP	2.0	2.3	2.5
Classification	AP	37.0	19.4	22.5
MIL NOR	AP	**41.4**	**20.7**	**24.9**

3: Maximum Entropy Language Model and Reranking



- Use MIL classifier to generate bag of detected words
- Generate many captions using 4-gram maximum entropy LM
- Then re-rank using Deep Multimodal Similarity Model within MERT

→ Turns out this works **really well**.

- 1st place at CVPR image captioning challenge
- Evaluated by humans

Metric	Description
M1	Percentage of captions that are evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing Test.
M3	Average correctness of the captions on a scale 1-5 (incorrect - correct).
M4	Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).
M5	Percentage of captions that are similar to human description.

Microsoft

→ Turns out this works **really well**.

- 1st place at CVPR image captioning challenge
- Evaluated by humans

Metric	Description
M1	Percentage of captions that are evaluated as better or equal to human caption.
M2	**Percentage of captions that pass the Turing Test.**
M3	Average correctness of the captions on a scale 1-5 (incorrect - correct).
M4	Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).
M5	**Percentage of captions that are similar to human description.**

Microsoft

→ Turns out this works **really well**.

- Also, 2nd place at CVPR image captioning challenge
  - When we add in GRNN forced decoding

Metric	Description
M1	Percentage of captions that are evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing Test.
M3	**Average correctness of the captions on a scale 1-5 (incorrect - correct).**
M4	Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).
M5	Percentage of captions that are similar to human description.

Microsoft

baseball ball player throwing

1. Image Processing

plate baseball holding man on to batter player up pitcher throwing his game field

2. Caption Generation

A baseball player swinging a bat at a game.
A man holding a baseball bat during a game.
...
A pitcher throwing the ball.
A close up of a baseball game.

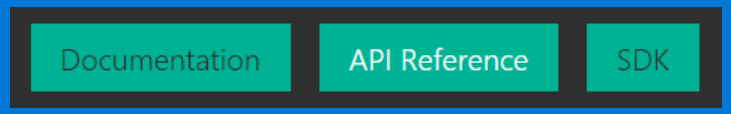
3. Global Re-ranking

A baseball player throwing a ball.

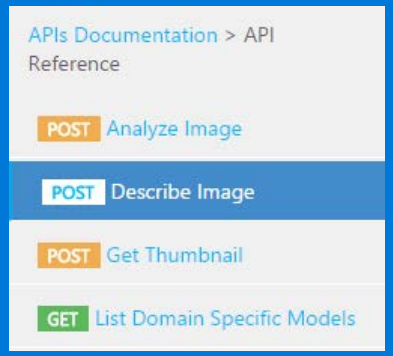
Now You Try!

Image Captioning: Now You Try!

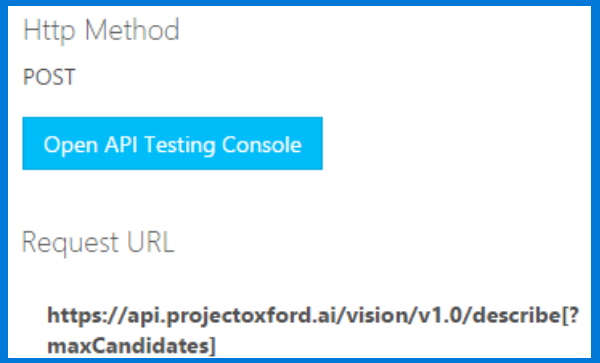
- Go to https://www.microsoft.com/cognitive-services/en-us/computer-vision-api
- Click **API Reference**



- Click **Describe Image**



- Then **Open API Testing Console**



Microsoft

Image Captioning: Now You Try!

Cognitive Services – Image Captioning

- **Request URL:**
  - https://api.projectoxford.ai/vision/v1.0/describe?3 ← return 3 captions
- **Image URL:** http://visionandlanguage.net/captioning.jpg
- **Response:**

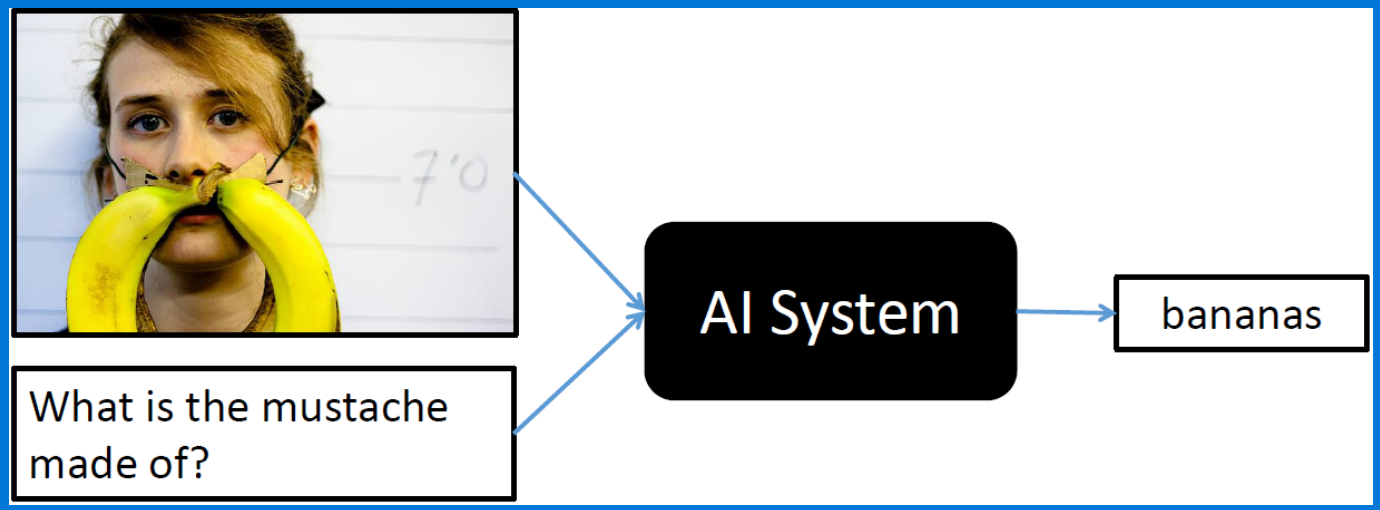

```
description: {        tags: [man, outdoor, sitting, window, bench, building, tree],
                captions: [{              text: A man sitting on a bench,
                                  confidence: 0.483},
                           {              text: A man is sitting on a bench,
                              confidence: 0.400},
                           {              text: A man sitting in front of a building,
                              confidence: 0.380 }]}
```


Microsoft

New Challenges
informed by
visually impaired accessibility

Visual Question Answering

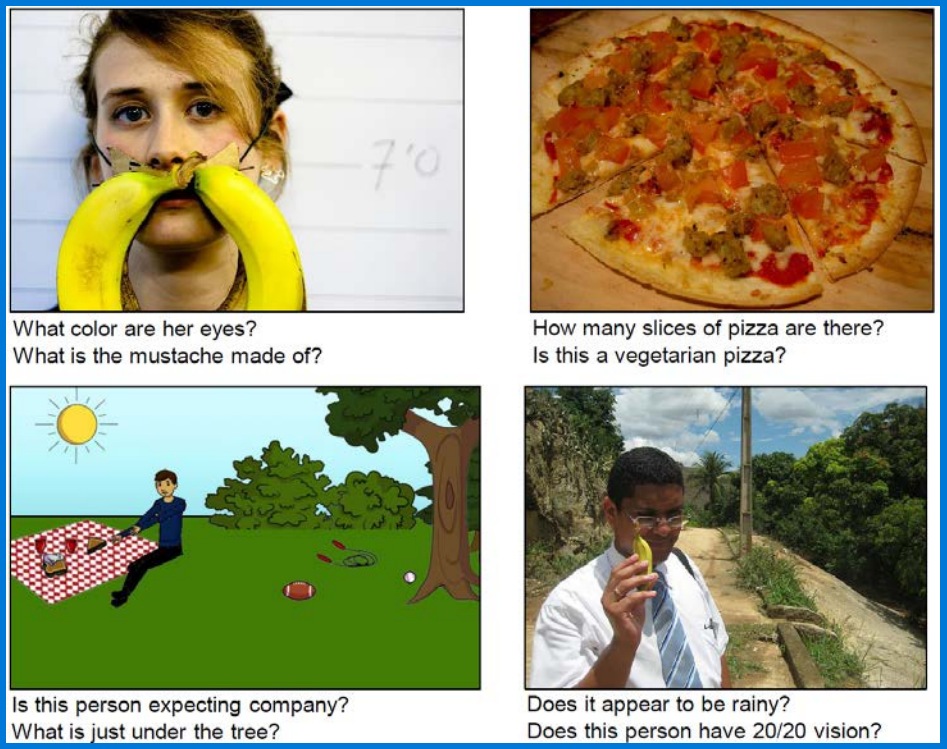
VQA: Visual Question Answering



visualqa.org

Microsoft

VQA: Visual Question Answering



visualqa.org

Microsoft

VQA: Visual Question Answering

254,721 images

764,163 questions (3 questions per image)

764,163 answers

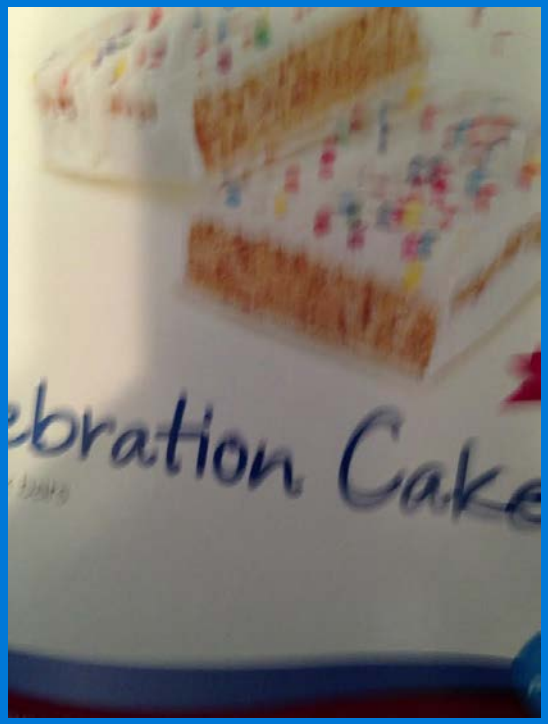
- 10 ground truth answers per question
- 3 plausible (but likely incorrect) answers per question

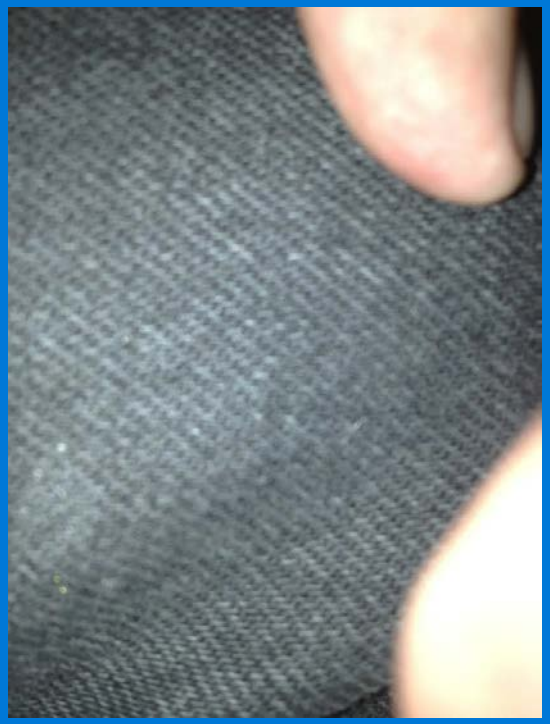
Open-ended and multiple-choice answering tasks

visualqa.org

Photo Complexity

Pictures Taken By People Who Are Visually Impaired Are Visually Complex



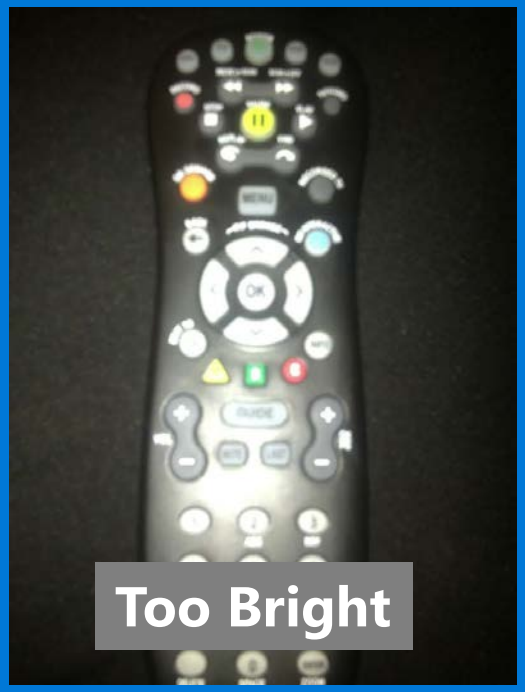


Pictures Taken By People Who Are Visually Impaired Are Visually Complex

- Partnership with Jeff Bigham, **VizWiz**
- Visually Impaired users take picture, ask question, answered by crowdsourcing.
  - E.g., What is this? Is this a lemon or a lime? What do the instructions say?
- Annotating images for Illumination, blur, cropping, clutter

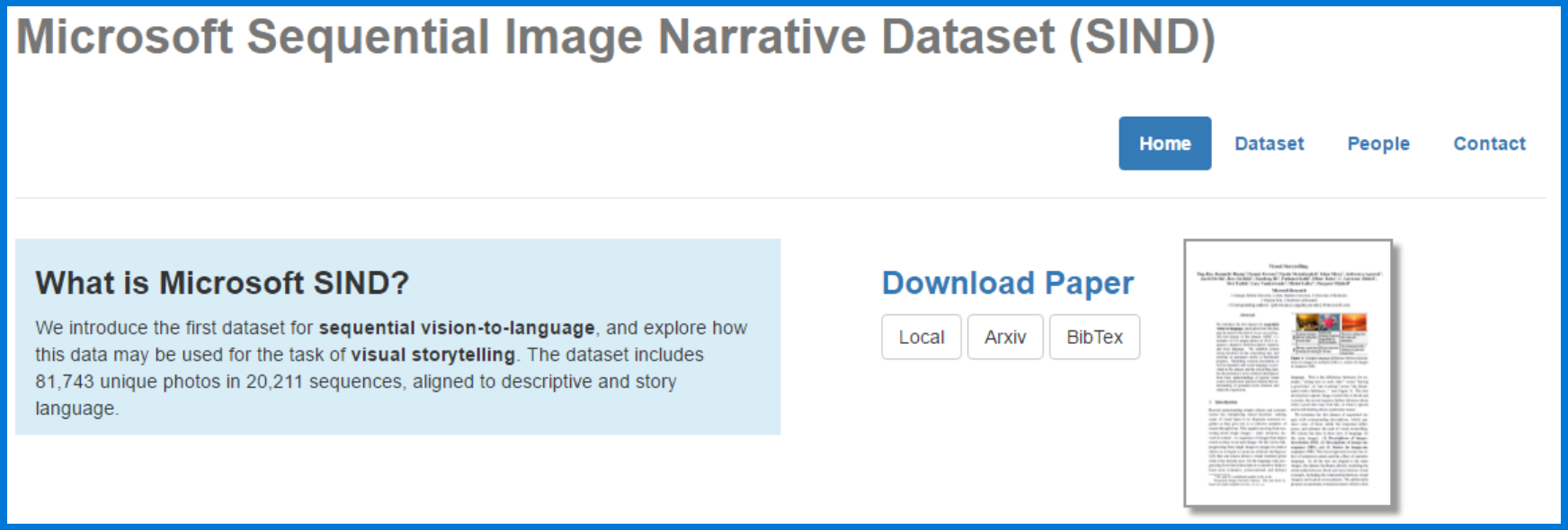
We Can Tackle This Problem!

- Need: Engineers dedicated to making this work.



Expressive Language

Abstract Conceptual Thoughts in Sequence



sind.ai

Microsoft

Abstract Conceptual Thoughts in Sequence

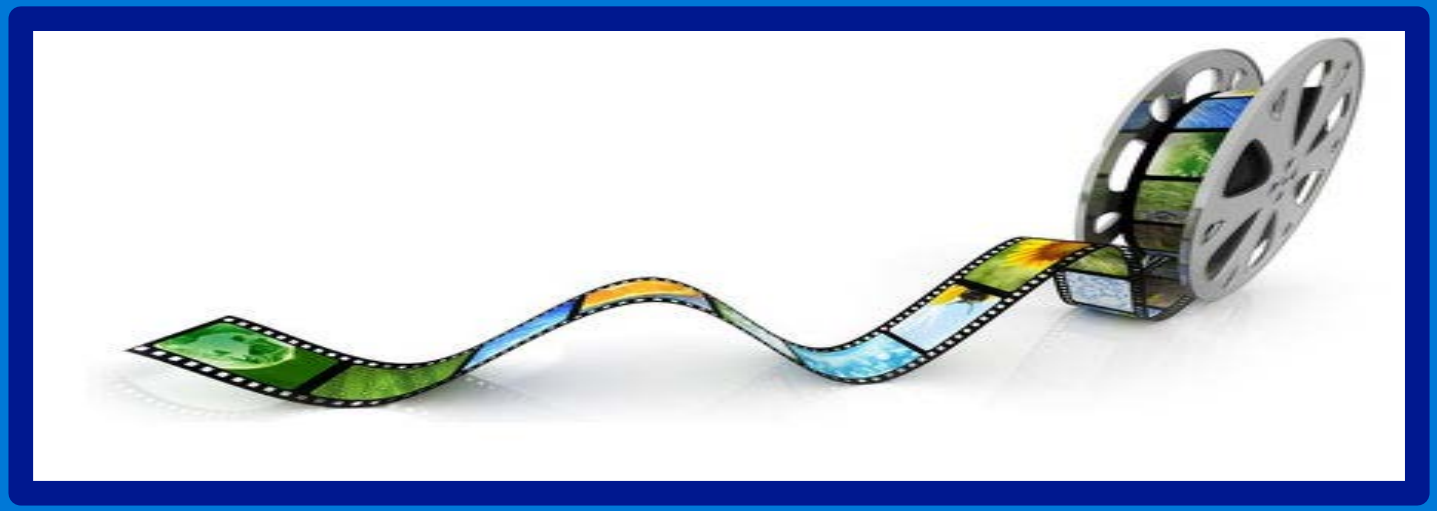
Example Generated Story

1	2	3	4	5
				
The dog was ready to go.	He had a great time on the hike.	And was very happy to be in the field.	His mom was so proud of him.	It was a beautiful day for him.

Microsoft

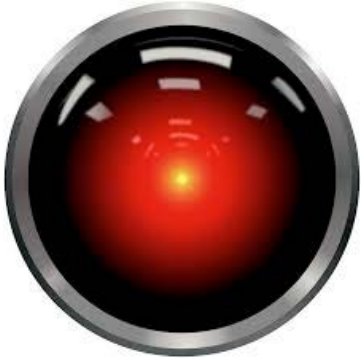
Looking further…

- Descriptions from video



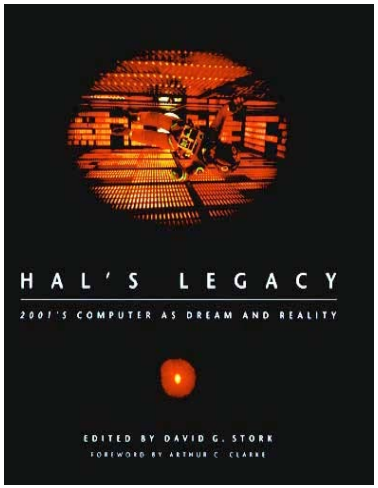
- Video is images in sequence . . .

  . . . with serious **compute needs**. (Hololens HPU?)



HAL (A space odyssey, 1968)

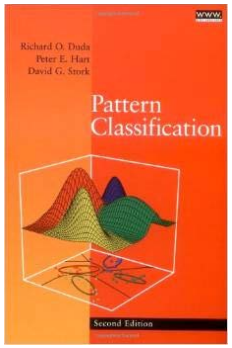
- David Stork (HAL's Legacy, 1998)



"Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description.

This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them.

I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers"



Web in 1995

MONEY & INVESTING UPDATE

from THE WALL STREET JOURNAL.

Front Page | STOCKS: U.S. | Small U.S. | Americas | Asia | Europe | Heard on the Street | Credit Markets | Foreign Exchange | Commodities | Mutual Funds

Wednesday, September 6, 1995

What's News —

• • •

Business and Finance

MARKETS DIARY 5 p.m. EDT

DJIA	4683.81	+ 13.73
S&P 500		+ 0.18%
Nasdaq Composite		+ 0.48%
Tokyo (Nikkei 225)		- 0.98%
London (FT 100)		+ 0.72%
30-Yr Treasury Yield		6.59%
Japanese yen (per US$)		98.82
German mark (per US$)		1.4768

Computer Shares Lift Stocks Again; Bonds Are Weak

By DAVE PETTIT
Money & Investing Update

amazon.com
Earth's biggest bookstore

Welcome to Amazon.com Books!

One million titles, consistently low prices.

(If you explore just one thing, make it our personal notification service. We think it's very cool!)

Spotlight! -- August 16th

These are the books we love, offered at Amazon.com low prices. The spotlight moves **EVERY** day so please come often.

One Million Titles

Search Amazon.com's million title catalog by author, subject, title, keyword, and more... Or take a look at the books we recommend in over 20 categories... Check out our customer reviews and the award winners from the Hugo and Nebula to the Pulitzer and Nobel... and bestsellers are 30% off the publishers list...

Eyes & Editors, a personal notification service

Like to know when that book you want comes out in paperback or when your favorite author

Web Today: Increasingly Visual

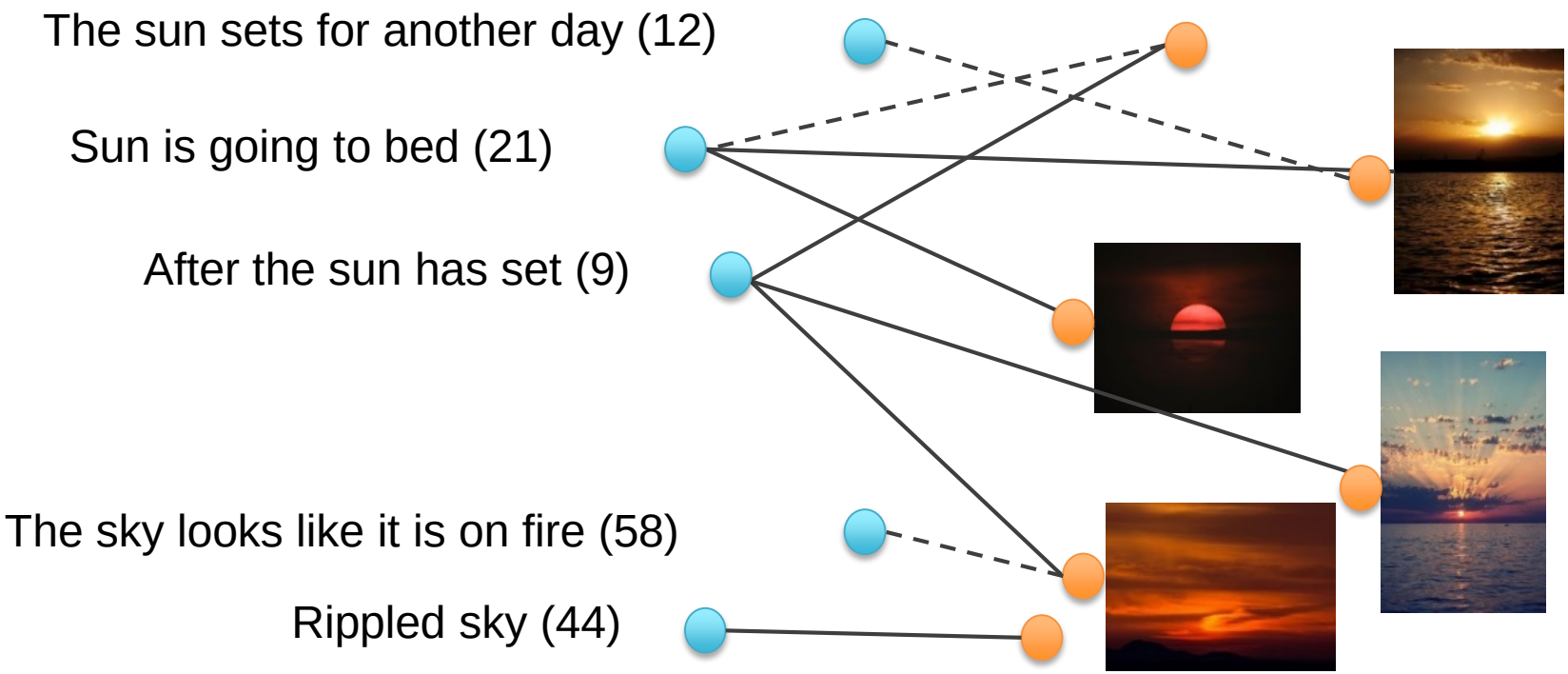
-- social media, news media, online shopping

- Facebook.com has over 250 billion images uploaded as of Jun 2013
- 1.15 billion users uploading 350 million images a day on average

Learning from data in the wild

Deja Image-caption corpus (NAACL 2015):

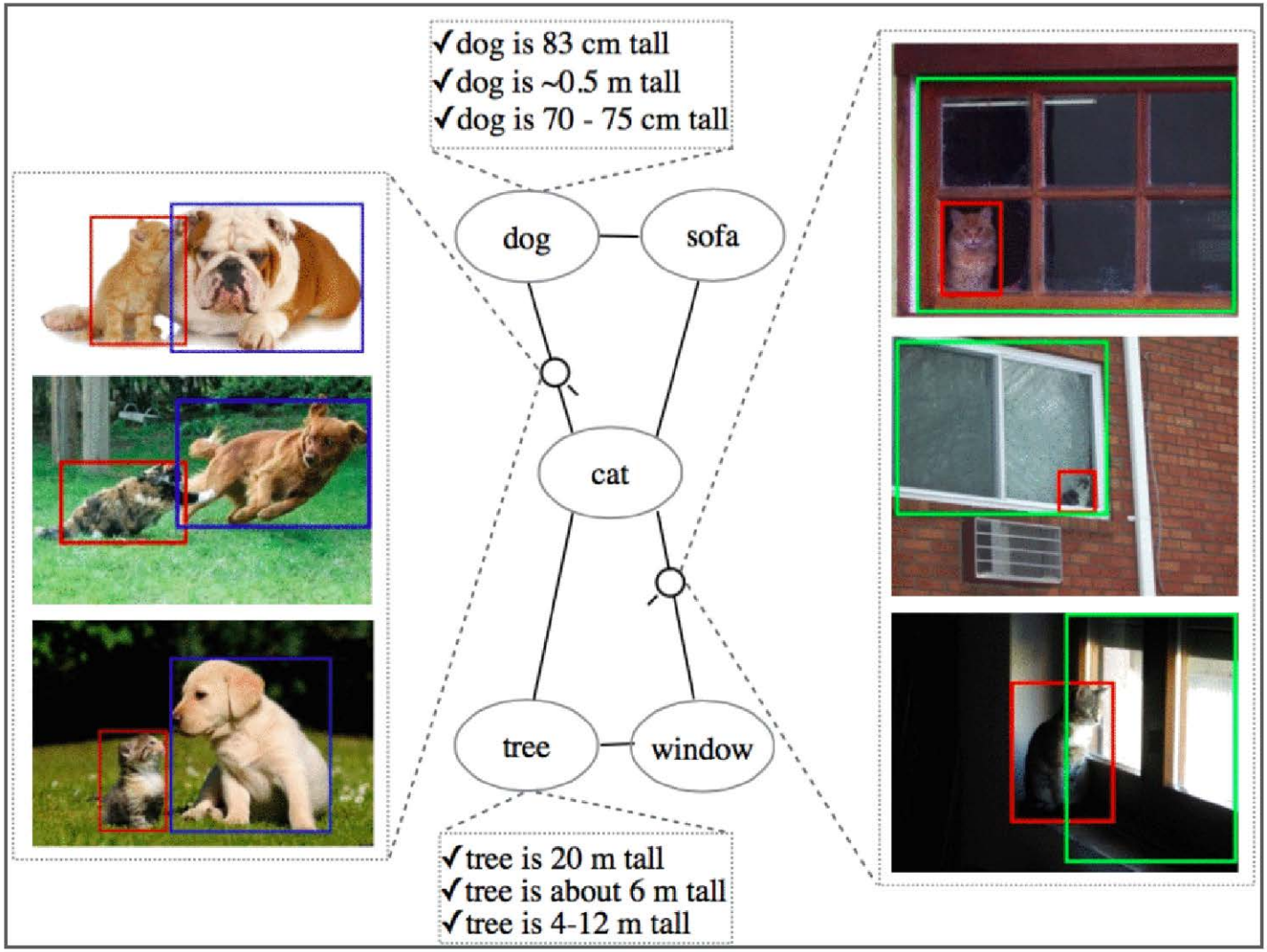
- Of 750 million image-caption pairs from Flickr
- Retain only those captions that are repeated verbatim by more than one user
- Yielding 4 million images with 180K unique captions



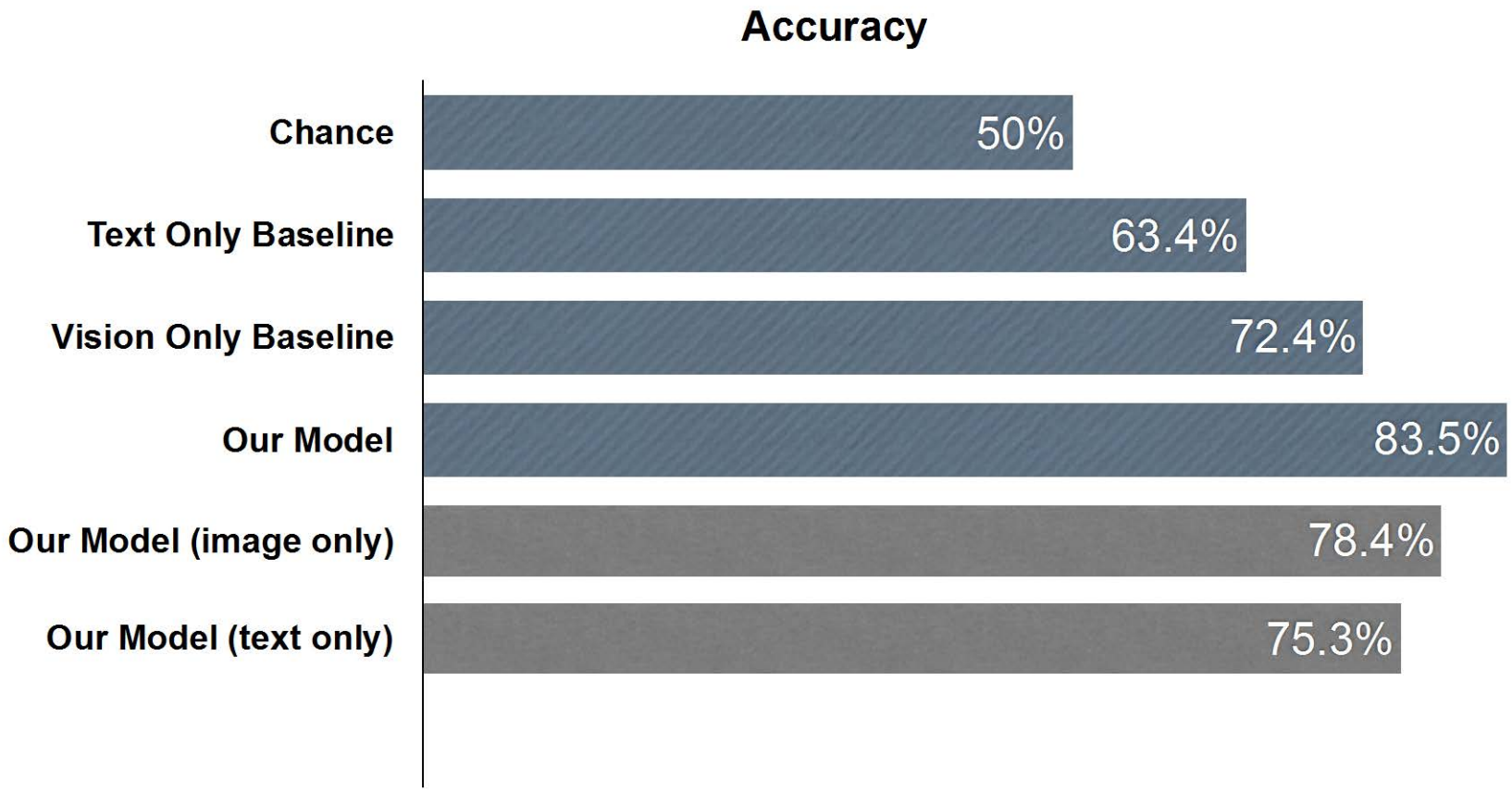
Seeing beyond What's in the Image

- What's happening?
- How / why did this happen?
- What are the intent / goal of the participants?
- Sentiment: are they happy?
- Reaction: do we need to act on them (e.g., dispatching help)?

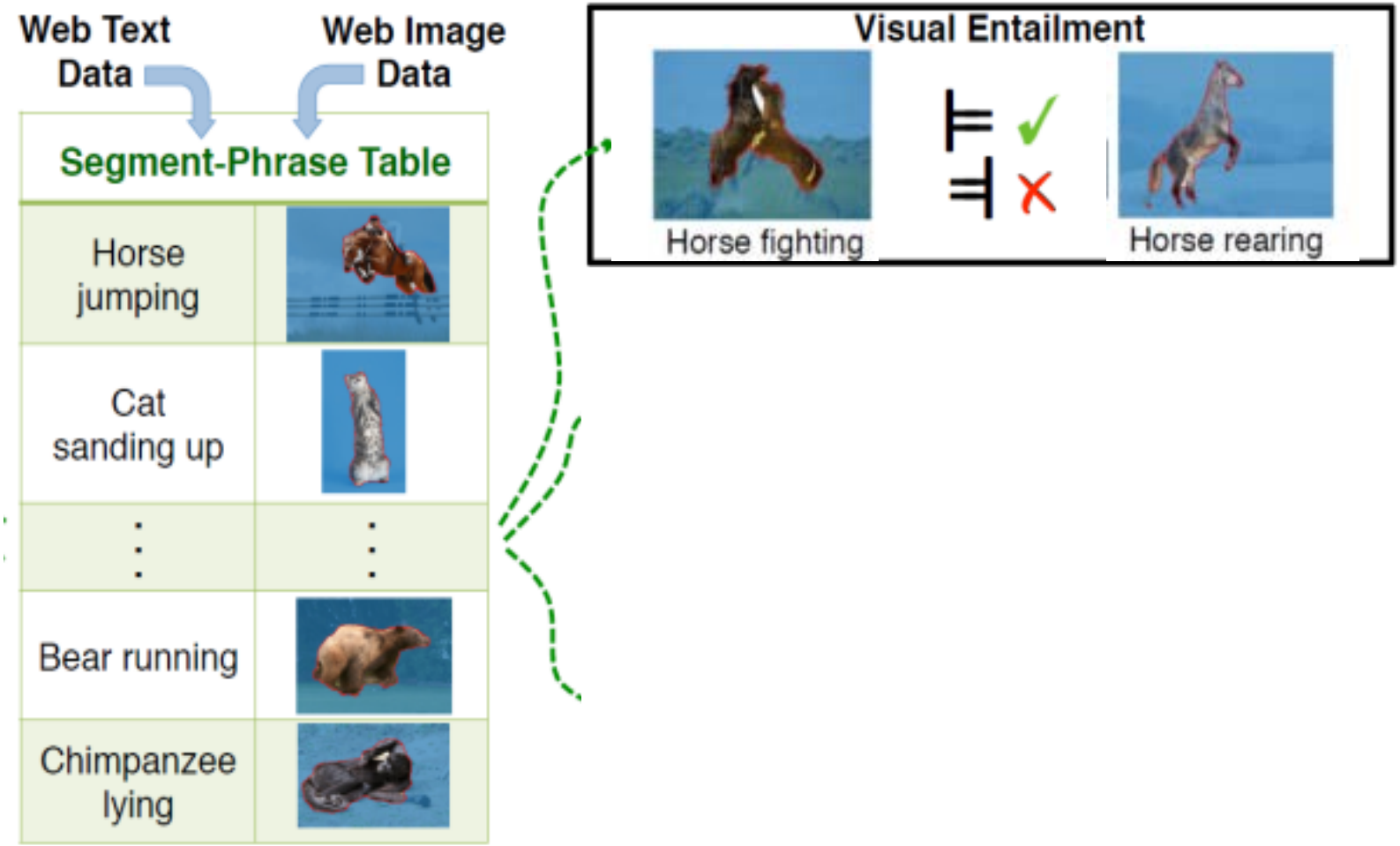
Size Graph


✓dog is 83 cm tall
✓dog is ~0.5 m tall
✓dog is 70 - 75 cm tall
dog
sofa
cat
tree
window
✓tree is 20 m tall
✓tree is about 6 m tall
✓tree is 4-12 m tall

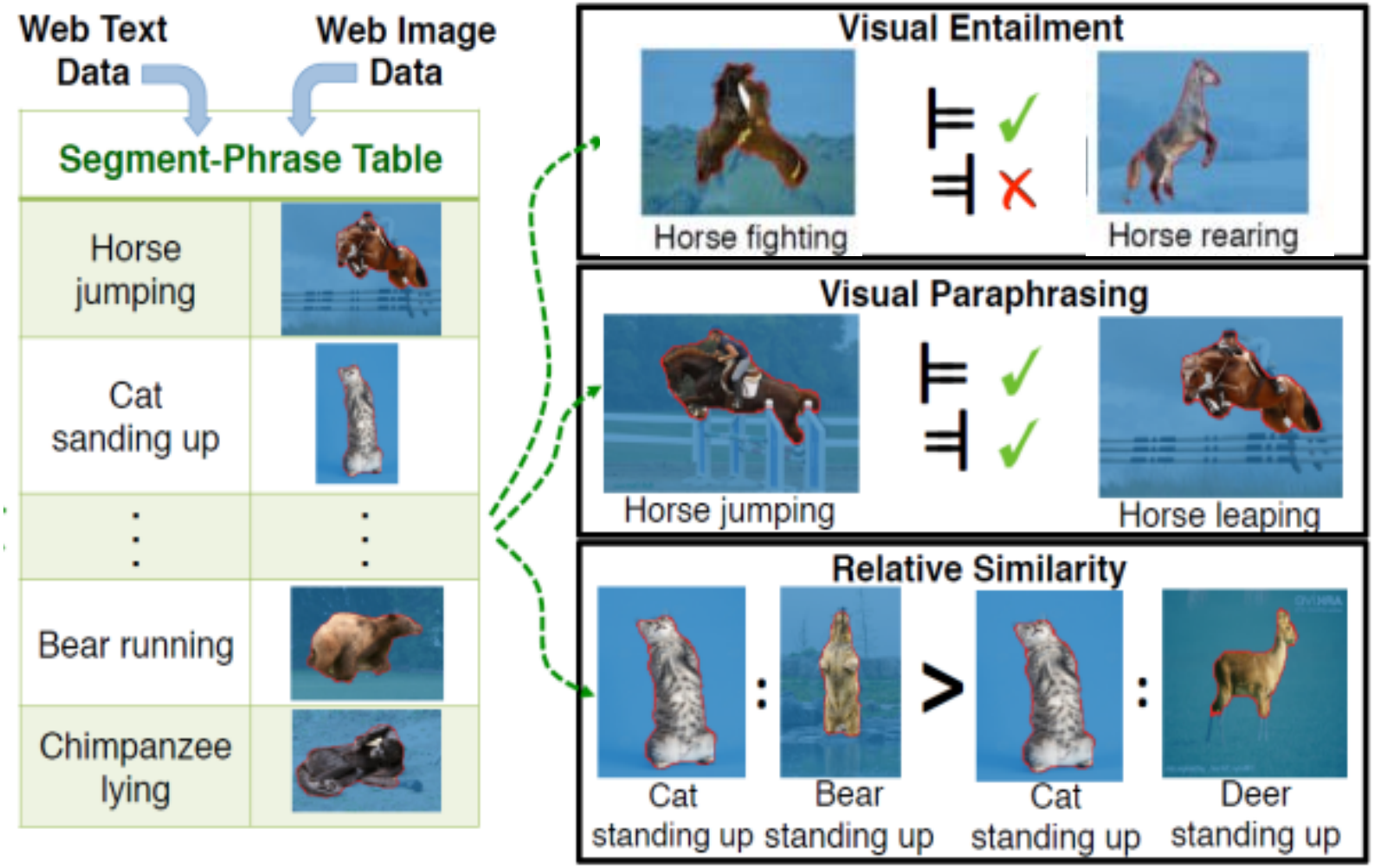
Which of objects A or B is bigger?



Segment-Phrase Table for Visual Entailment (ICCV 2015)



Segment-Phrase Table for Visual Entailment (ICCV 2015)



Thanks!!!

- Yejin Choi: yejin@cs.washington.edu

- Margaret Mitchell: memitc@Microsoft.com