



Cognitive Services: Computer vision

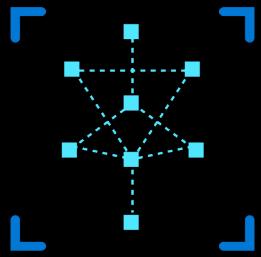
Rory Preddy
Senior Cloud Advocate

Twitter: @RoryPreddy



<https://aka.ms/java-vision>

Novel object captioning at scale (nocaps) challenge



Microsoft achieves
Image Captioning
Human Parity
July 2020

training

COCO (80 classes)

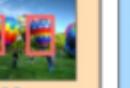


Two **pug** **dogs** sitting on a **bench** at the beach.



A **child** is sitting on a **couch** and holding an **umbrella**.

Open Images (600 classes)

		
goat	artichoke	accordion
		
dolphin	waffle	balloon

nocaps validation/test

in-domain: only COCO classes



The **person** in the **brown** **suit** is directing a **dog**.

near-domain: COCO & novel classes



A **person** holding a black **umbrella** and **accordion**.

out-of-domain: only novel classes



Some **dolphins** are swimming close to the base of the ocean.

The **nocaps** benchmark for novel object captioning (at scale).

Leaderboard

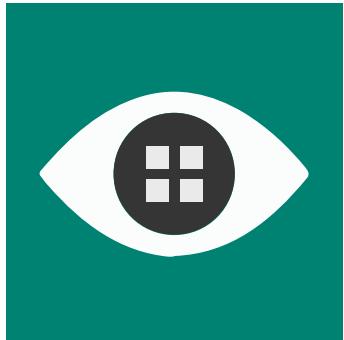
Phase: Test Phase, Split: Entire Dataset

B - Baseline submission * - Private submission

Rank	Participant team	B1	B2	B3	B4	ROUGE-L	METEOR	CIDEr	SPICE	Last submission at
1	Microsoft Cognitive Services team	79.44	61.66	40.65	22.32	54.65	26.30	86.58	12.38	3 months ago
2	nocaps Team (Human)	B	76.64	56.46	36.37	19.48	52.83	28.15	85.34	14.67



Vision



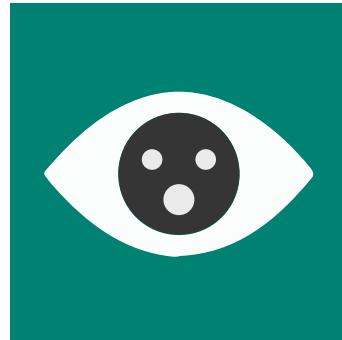
Computer Vision API

Distill actionable information from images



Face API

Detect, identify, analyze, organize, and tag faces in photos



Emotion API

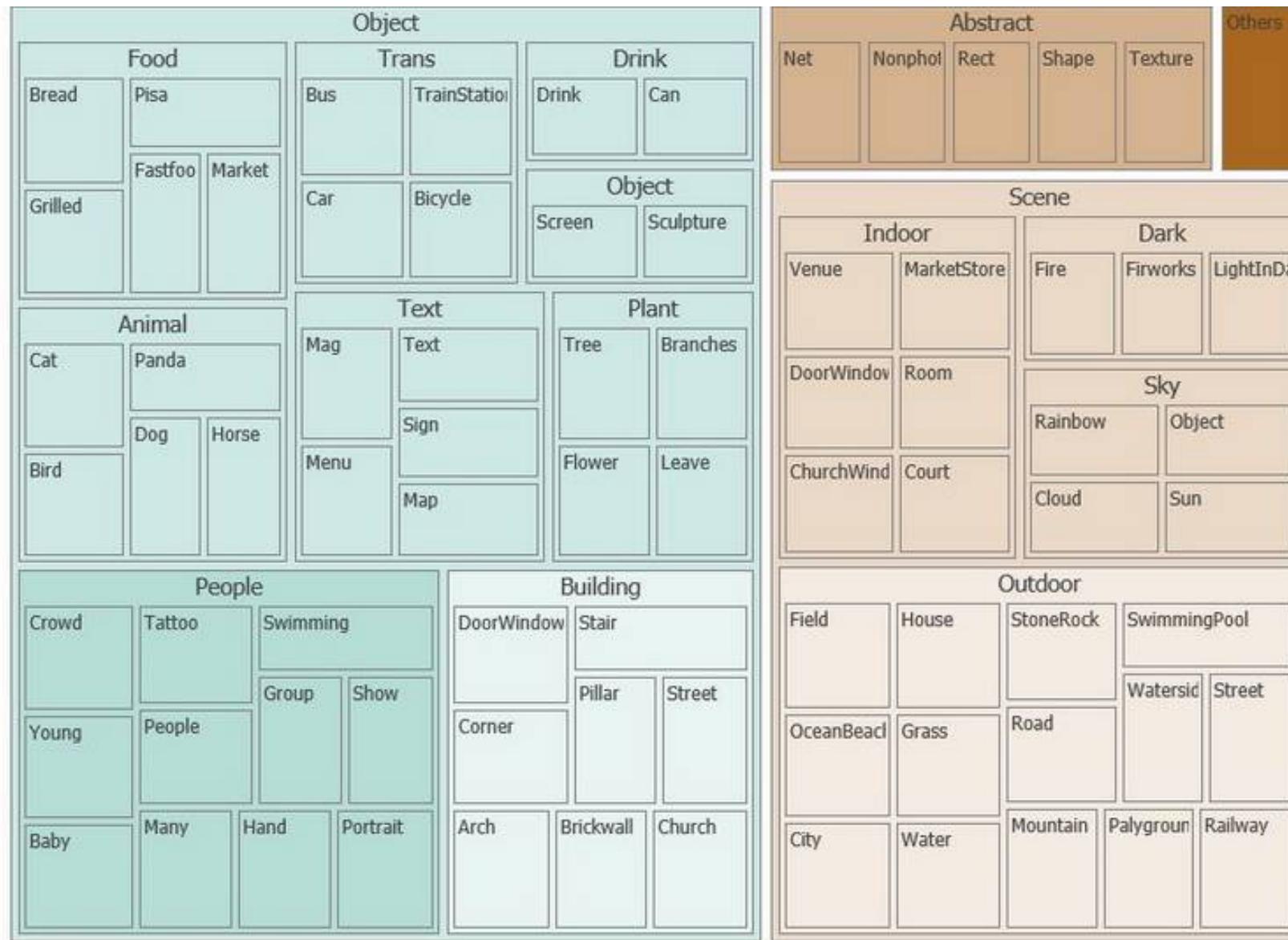
Personalize experiences with emotion recognition



Video API

Analyze, edit, and process videos within your app

Computer Vision: 86 Categories



Data

Computer Vision

Description, tags, clip art, line drawing, black & white, IsAdultContent/Score, IsRacy/Score, categories, faces, dominant colors, accent color

<https://www.microsoft.com/cognitive-services/en-us/computer-vision-api>

Emotions

Anger, contempt, disgust, fear, happiness, sadness, surprise, and neutral

<https://www.microsoft.com/cognitive-services/en-us/emotion-api>

Face

Bounding box, 27 facial landmarks, age, gender, head pose, smile, facial hair, glasses

<https://www.microsoft.com/cognitive-services/en-us/face-api>

Captioning

Problem Statement: Generate textual description (typically sentence) that is both adequate and fluent.



A person riding a motorcycle on a dirt road.



A group of young people playing a game of frisbee.



Two dogs play in the grass.



Two hockey players are fighting over the puck.

The long-held dream

- Image understanding by computers



Boy riding on
horse

Time for a Turing test



A man standing on a tennis court holding a racquet



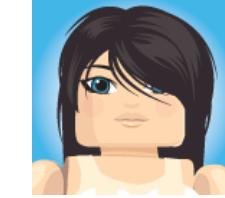
The man is on the tennis court playing a game



Another one



An ornate kitchen is designed with rustic wooden parts



A kitchen with wooden cabinets and a sink



Another one



A little girl in a pink shirt standing near a blue metal sculpture



A group of young people play soccer on a city street

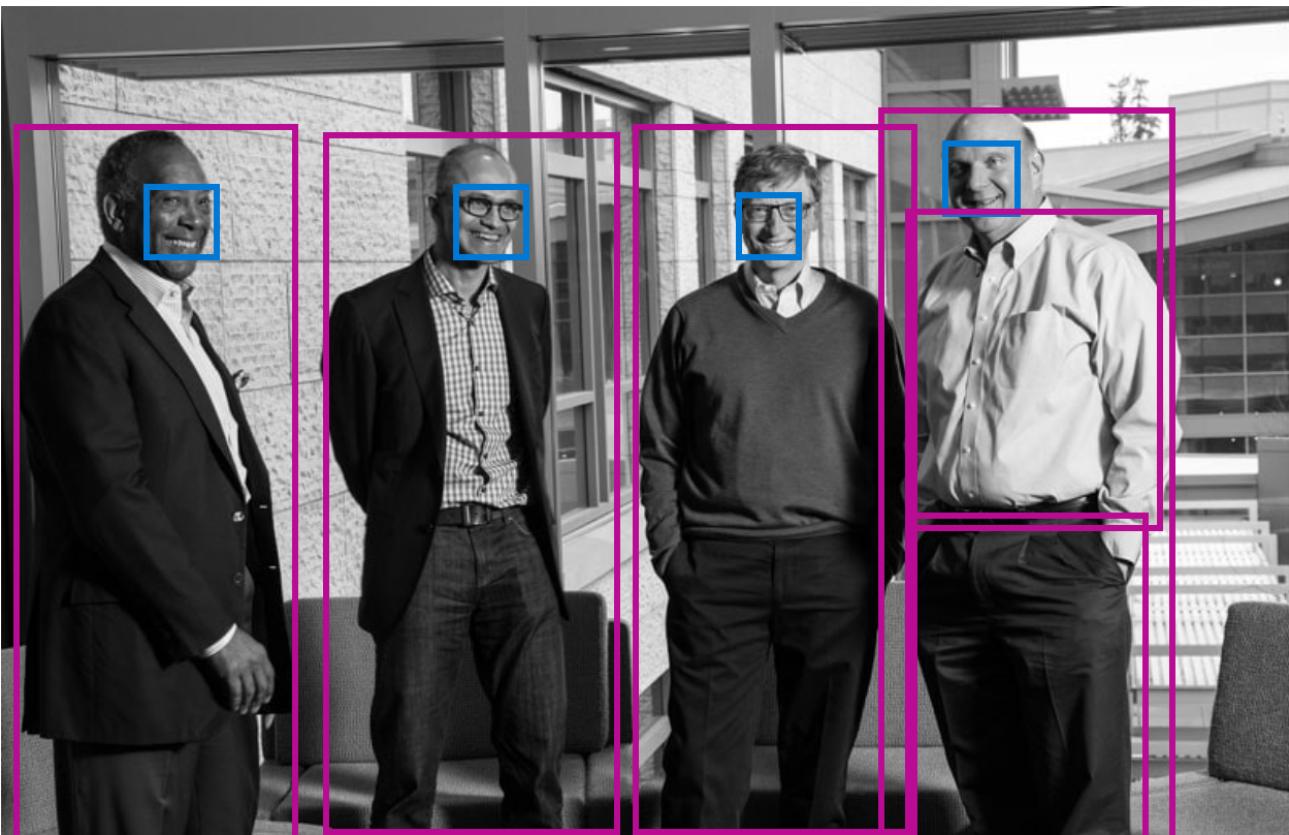


How did we do it?

- Language generation
 - Markov model, trained on caption, conditional on words detected from image
- Image analysis
 - Deep-learned features, applied to likely objects in the image, trained to produce words in captions
- Reranking
 - Hypothetical captions reranked by deep-learned model looking at entire image

<https://www.microsoft.com/en-us/research/project/azure-florence-vision-and-language/>

SDK results



0.967977047 }, { "name": "smile", "confidence": 0.9477451 }, { "name": "black and white", "confidence": 0.9309614 }, { "name": "man", "confidence": 0.926383138 }, { "name": "outdoor", "confidence": 0.8966964 }, { "name": "posing", "confidence": 0.834066331 }, { "name": "human face", "confidence": 0.7781277 }, { "name": "group", "confidence": 0.7551476 }, { "name": "suit", "confidence": 0.7084104 }, { "name": "clothes", "confidence": 0.172121257 }]

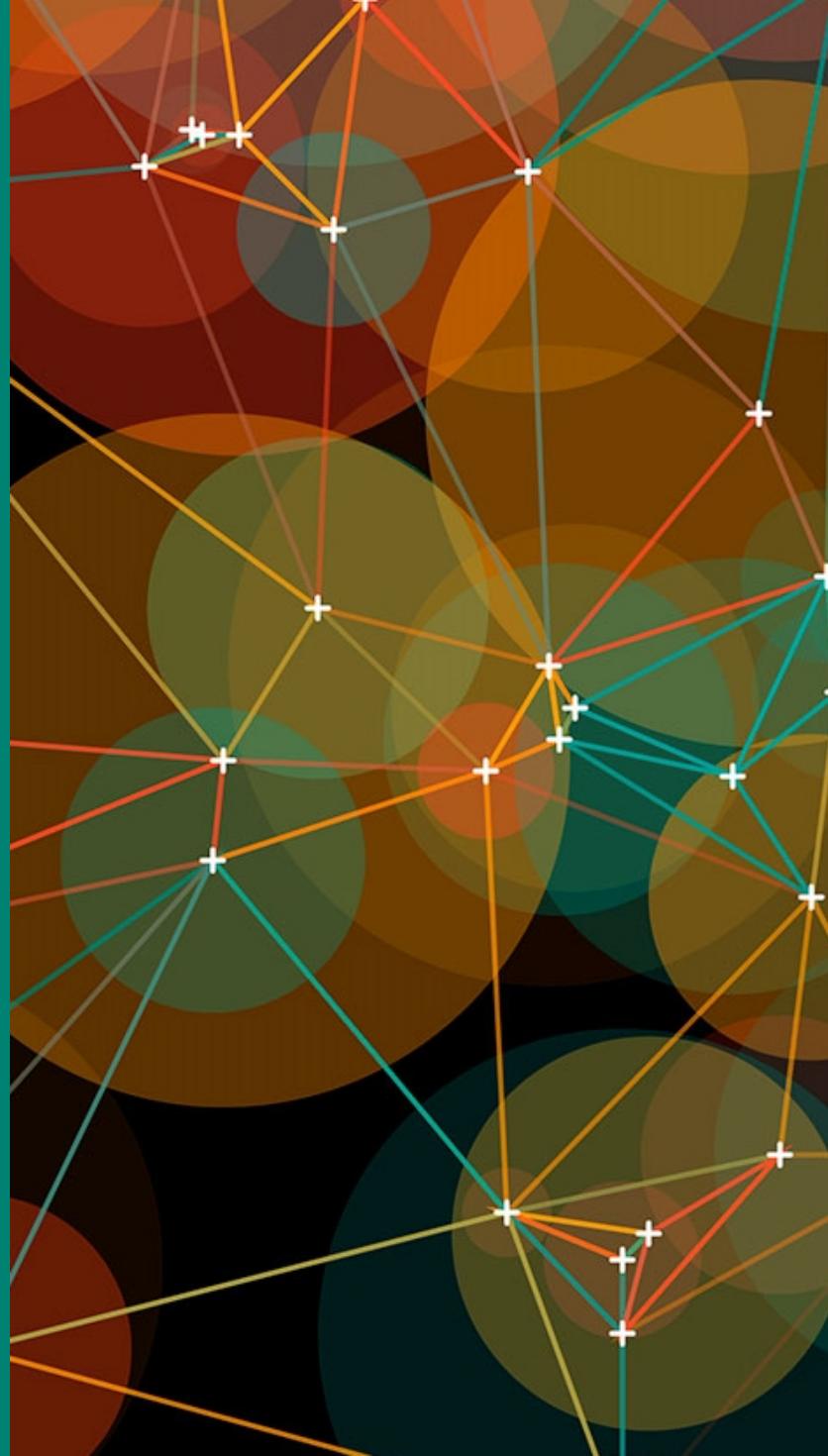
Description { "tags": ["person", "building", "standing", "photo", "man", "outdoor", "posing", "group", "suit", "wearing", "player", "front", "people", "old", "court", "woman", "holding", "dressed", "room", "baseball", "field"], "captions": [{ "text": "Satya Nadella, Bill Gates, Steve Ballmer posing for a photo", "confidence": 0.9753075 }] }

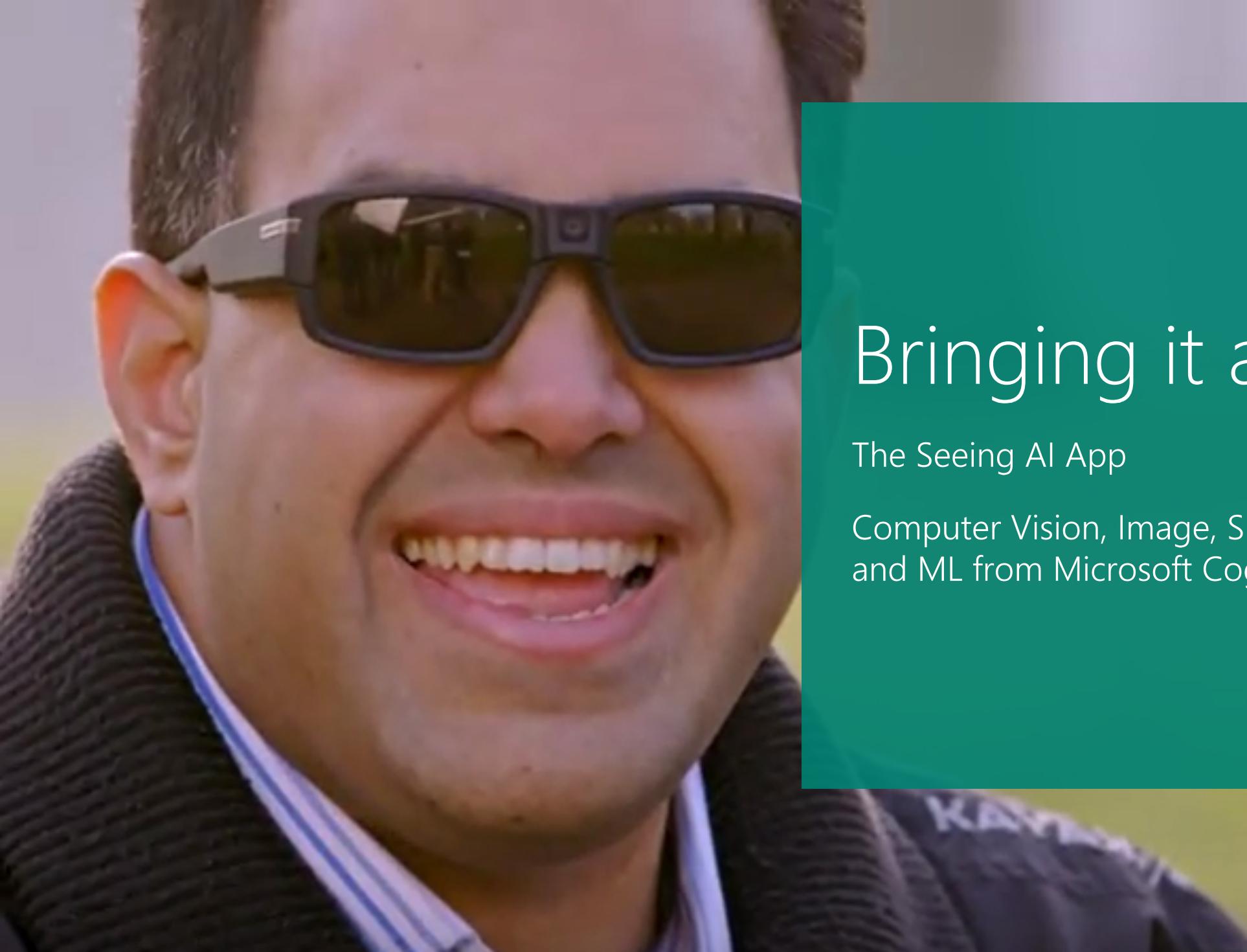
Image format "Jpeg"

Image 516 x 800

Demo

- Caption Bot – Computer Vision API
 - <https://github.com/Azure-Samples/cognitive-services-quickstart-code>





Bringing it all together

The Seeing AI App

Computer Vision, Image, Speech Recognition, NLP,
and ML from Microsoft Cognitive Services



<https://aka.ms/java-vision>

