university of groningen

Quantitative Text Analysis – Essex Summer School

New data

dr. Martijn Schoonvelde

University of Groningen

Today's class

- New data
  - Machine translation and multilingual data
  - Speech-to-text
  - Image data
- Lab session
- Flash talks: Aly, Alexandra

Machine Translation and Bag of Words models

Comparative social science

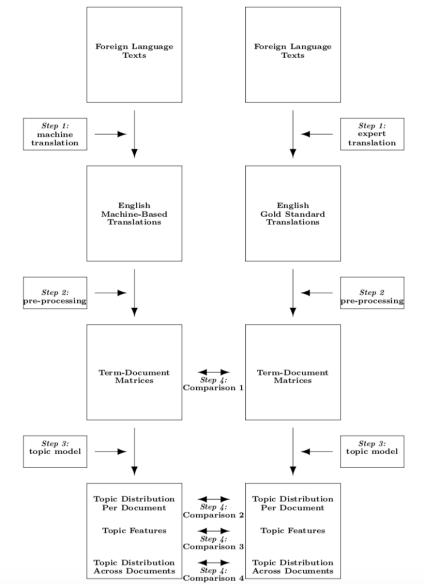
- Lots of data and theory for comparing institutions, behavior and public opinion
- But we know less about how to measure concepts across texts in different languages
  - Immense linguistic diversity: Some 6,000–7,000 languages are spoken globally today which vary in script, morphology and syntax

Research question (De Vries *et al.*, 2018): is machine-translated text of good enough quality for us to use it for BoW-based QTA research questions?

Data

- Europarl dataset (Koehn 2005)
- Transcriptions of European Parliament debates by official translators
  - Danish, German, Spanish, French and Polish for the period of January 2007 to November 2011
  - Required lots of cleaning to match individual chapters (i.e., debates, questioning, etc) per language pair – 11,469 documents in total
- Take source language debates and compare machine- and human-translations

Research design



Results at document level

Figure 3: Distribution of cosine similarity per language pair

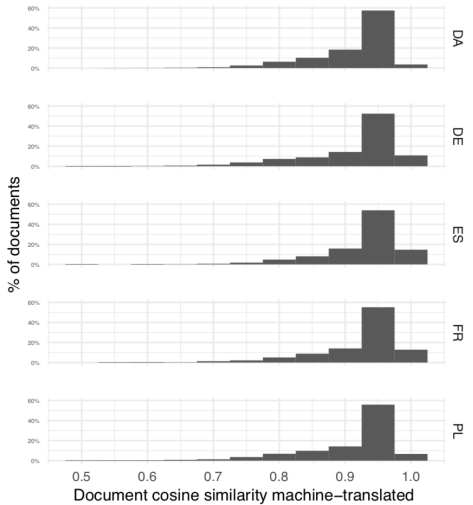
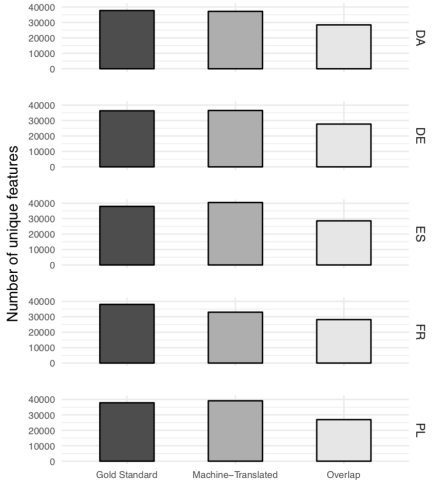


Table 2: Cosine similarity distribution per language

Language	N	Mean	St. Dev.	Min	Max
Danish	2,301	0.915	0.063	0.549	0.992
German	2,148	0.915	0.074	0.488	0.991
Spanish	2,335	0.929	0.059	0.483	0.991
French	2,347	0.925	0.064	0.564	0.989
Polish	2,338	0.913	0.073	0.475	0.989
Total:	11,469	0.919	0.066	0.475	0.992

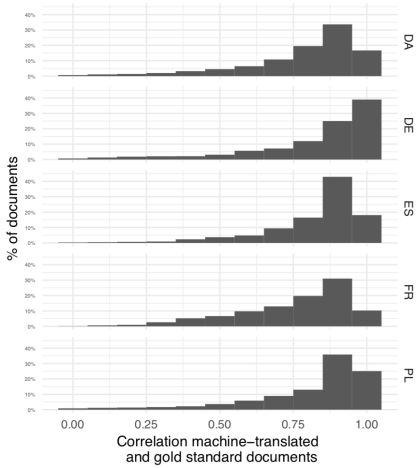
Results at document level

Figure 4: Unique TDM features for gold standard and machine-translated corpora



Results at document level

Figure 5: Similarity of document-level topical prevalence with equal number of topics



Take away

Machine translation works, at least when working with lots of text and bag of words models in some language pairs

- Paper contains further results on topical prevalence and topical content at corpus level

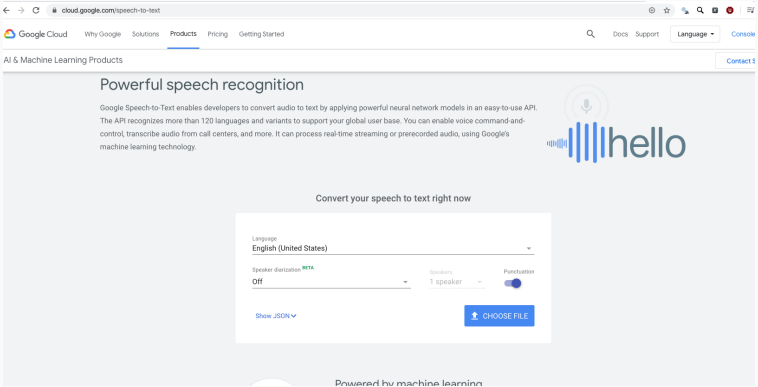
Many exciting developments on dealing with this Babel problem (Chan *et al.* 2020) since then

- At the data generation stage (e.g., large-scale multilingual language models, and machine translation) and at the modeling and research design stage (e.g., Lind *et al.* 2019; Lind *et al.* 2021a; Lind *et al.* 2021b)

Political speech

- Lots of political speech occurs in talk shows, on Youtube channels, etc.
- **Problem**: much of of it is not transcribed
- Proksch, Wratil, Wäckerle *et al.* (2019) test the quality of Automatic Speech Recognition (ASR) systems for bag of words models

Google Speech API



Corpus

57 speeches (French, German, English) from the EP's State of the Union plenary (SOTEU)

	Human transcriptions ("gold standard")	Google Speech API	Youtube Video
Speech 1			
Speech 2			
Speech 3			

Measurement

Measure: word error rate (WER):

$$WER = \frac{S + D + I}{N}$$

N = number of words in the gold standard transcriptions
S = number words with inaccurate ASR transcription ("substitutions")
D = number of words that are missing in the ASR transcription ("deletions")
I = number of words in the ASR transcription that are not in the gold standard text ("insertions")

WER

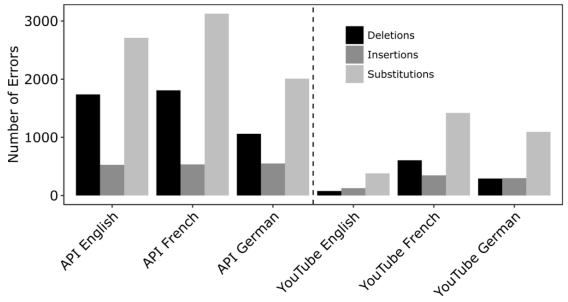
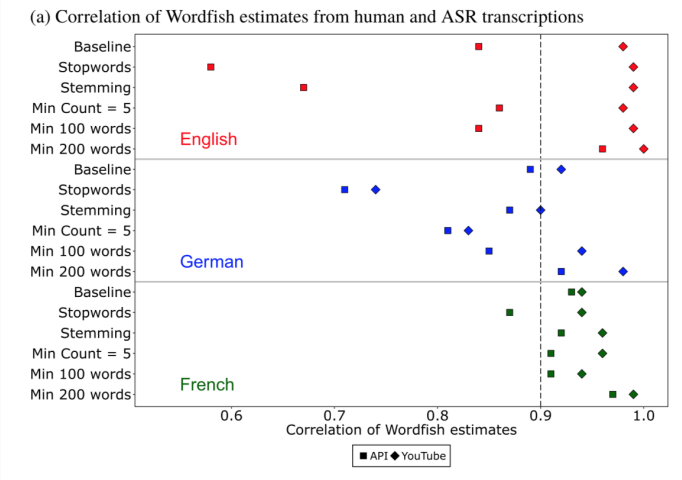


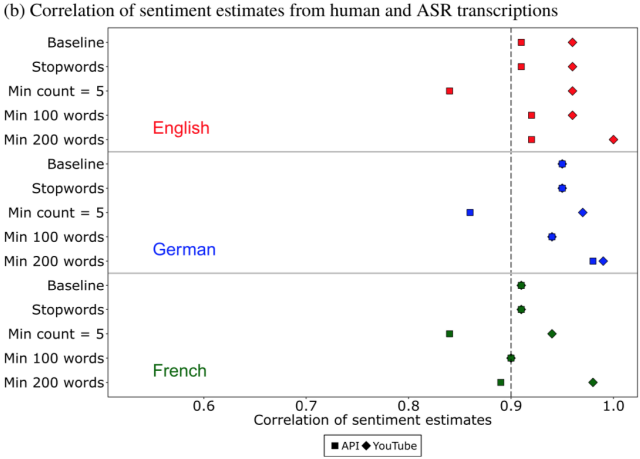
Figure 1. Word error structure in the State of the European Union debate.

Average WER Youtube: 0.03 (English), 0.12 (French), 0.10 (German)
Average WER Google API: 0.21 (English), 0.26 (French), 0.21 (German)

Applications: scaling



Applications: sentiment analysis



Take away

Automatic Speech Recognition generates meaningful data for bag-of-words text models

This will make transcription radically cheaper

Authors introduce a procedure (WERSIM) to simulate impact of increases in WER on quantities of interest

Images and political communication

When communicating with the public, politicians often rely on images

Images are powerful

Images are abundant

We're faced with an abundance of (political) images on social media and on the web:

- According to one estimate (Cross *et al.* 2021) 1 in 3 social media posts of politicians contain an image

This data abundance has already generated important social scientific insights:

- Displays of anger during in US Presidential debates are popular with the public (but only for male candidates) (Boussalis & Coan, 2020)
- Images posted on social media reveal information about depression and anxiety (Guntuku *et al.* 2019)

Studying images

But images are also difficult to study at large scale

- Still high technical barriers
  - See Webb Williams *et al.* (2020) for an overview of 'images-as-data'
- And much training data required
  - In order to train a computer to 'see' what is on a an image it needs a lot of **handholding and examples**

Alternative: image recognition services

For example, **Google Cloud Vision, Microsoft Azure Computer Vision, Amazon Rekognition, IBM Watson**

- offer **cheap and easy to use** alternatives for labeling images, object recognition, face recognition, emotion detection, etc.
- "the [service coded] 1,818 images in less than 5 min, whereas the human coder spent nearly 35 hours to complete the same task" (Bosch, Revilla, and Paura, 2019)

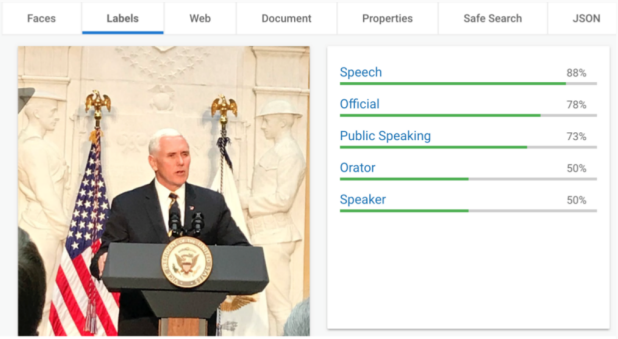
Google Cloud Vision

Image labels

Landmarks | Faces | Objects | **Labels** | Text | Properties | Safe Search

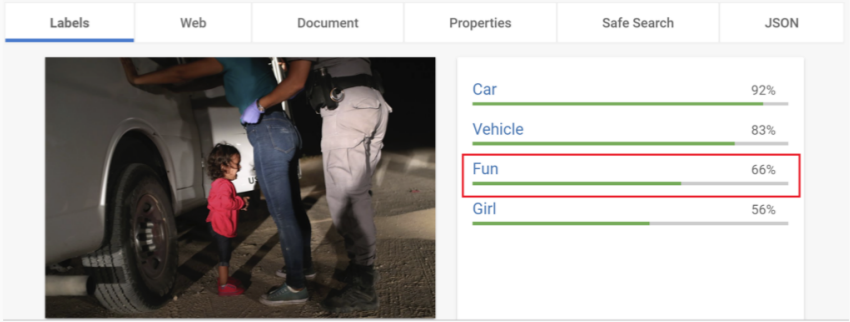
Sky	95%
Building	93%
Blue	90%
Hat	88%
Fan	86%
Gesture	85%
Crowd	83%
Flag	82%

Google Cloud Vision often works . . .



Speech by Vice President Mike Pence

. . . but sometimes fails



GCV and images of politicians

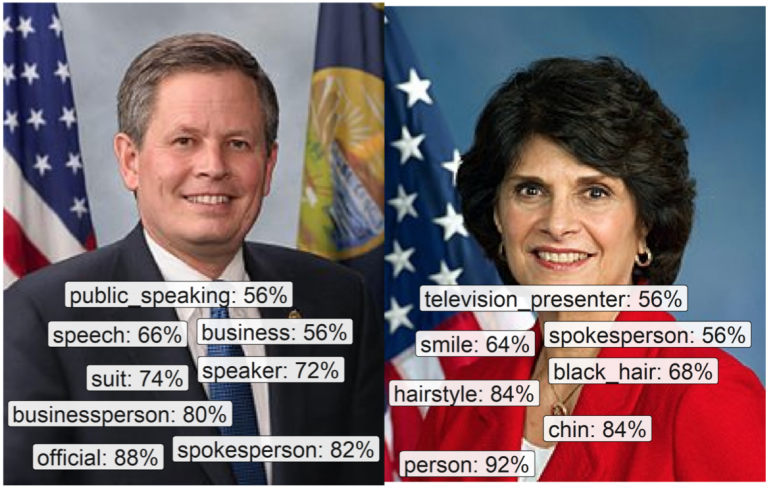


Fig. 4. Example of two images of U.S. Members of Congress with their corresponding labels as assigned by Google Cloud Vision. On the left, Steve Daines, Republican

So how biased are such data?

Our focus is on **Google Cloud Vision**

- Widely used in industry; shares technology with Google Image Search and Google Photos (integrated with every Android phone)

We are particularly interested in gender bias – the extent to which men and women politicians **are seen differently by GCV**

Two forms of bias

We examine **two forms of bias**:

- Bias in identification: does the algorithm "see" people with equal accuracy regardless of their gender?
- Bias in content: does the algorithm systematically return different types of (correct?) labels depending on who is in the picture?
  - **conditional demographic parity** (Corbett-Davies *et al.* 2017): "if men and women in a sample wear suits at equal rates, then an unbiased algorithm would return the label "suit" equally often for each gender"

Two datasets

To examine both biases we rely on two datasets:

- Found data $\approx 200,000$ images in tweets from US Members of Congress (Jan 2017 - June 2018)
- Controlled data professional portraits of all Members of Congress

On average, GCV returned 5.3 labels per image, and we selected only labels to **which GCV assigned high confidence** (≥ 0.75)

Crowd-coded validation

Crowd-coded validation on a **sample of found data** ($n = 9,250$): We presented crowd coders with 30 images and a set of five potential labels for each:

- some labels were assigned by GCV (**positive labels**)
- others were chosen at random from the set of GCV labels assigned to other images (**negative labels**)

Crowd-coded validation

Crowd-coded validation on a **sample of found data** ($n = 9,250$): We presented crowd coders with 30 images and a set of five potential labels for each:

- some labels were assigned by GCV (**positive labels**)
- others were chosen at random from the set of GCV labels assigned to other images (**negative labels**)

For each image, coders were presented with two tasks:

1. Select all labels that applied to the image they were seeing (i.e., **identify the positive labels**)
2. Indicate if they saw any **men, women, children**, or none in the image

Findings

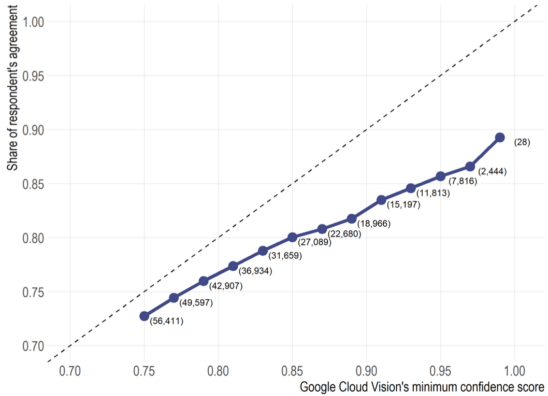


Fig. 2. Relationship between Google Cloud Vision confidence and human agreement with image labels. Numbers in parentheses denote observations for corresponding confidence score thresholds.

Bias in representation

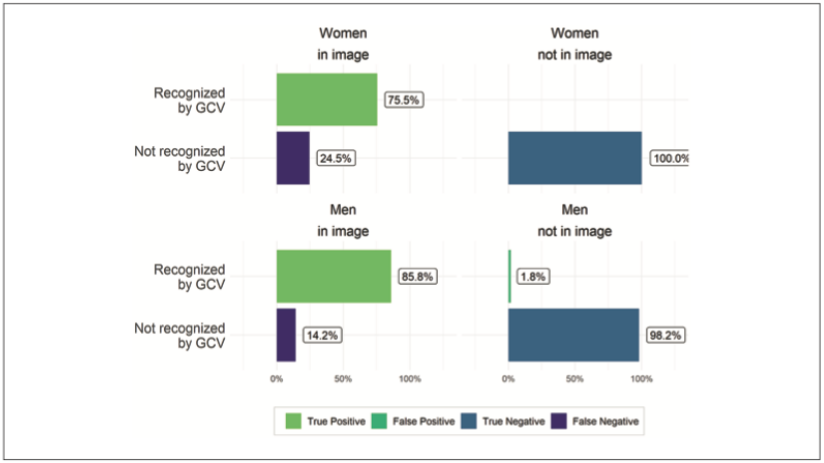


Figure 3. Accuracy of person detection of Google Cloud Vision (GCV). Percentages shown were determined by comparing gender of members of Congress depicted in uniform data (professional photographs) with annotations from the object recognition software.

GCV more likely to not see a woman than a man in an image

Bias in representation

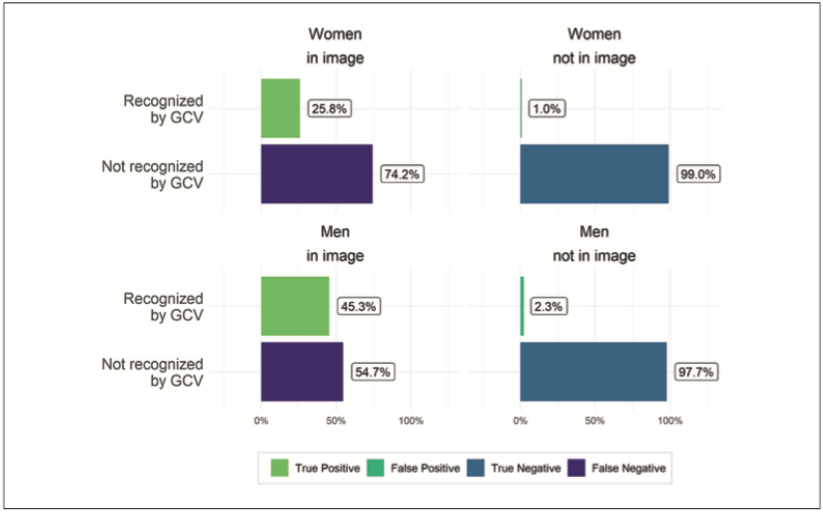


Figure 4. Accuracy of person detection of Google Cloud Vision (GCV). Percentages shown were determined by comparing human agreement about the presence of men or women in Twitter images with annotations from the object recognition software.

Bias in content

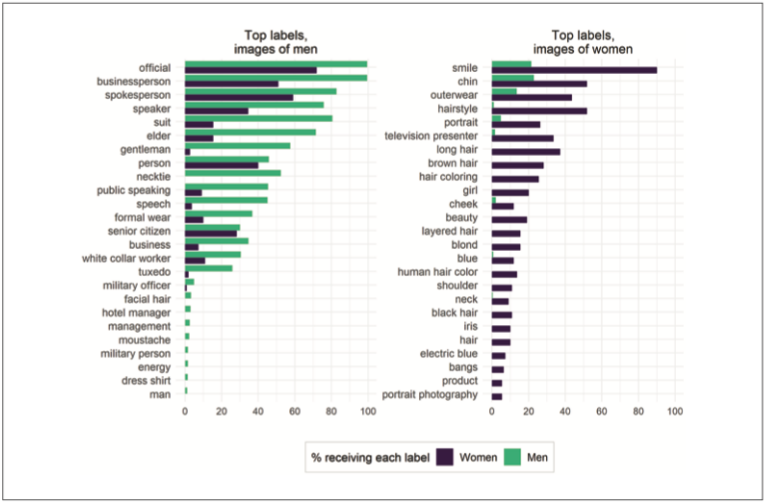
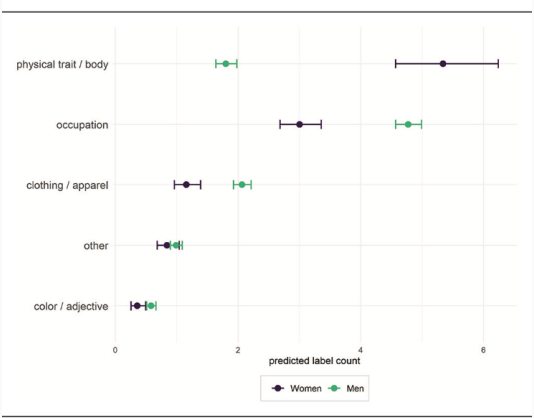


Figure 6. Google Cloud Vision labels applied to control dataset (professional photos). The 25 most gendered labels for men and women were identified with χ^2 tests ($p \leq .01$). Labels are sorted by absolute frequencies. Bars denote the percentage of images for a certain label by gender.

Bias in content



- Three of us coded the labels independently in 5 categories ($\alpha = 0.88$):
  - Images of women politicians receive about 3 times more labels categorized as "physical traits & body" (5.3 for women, 1.8 for men).
  - Images of men politicans receive about 1.5 times more labels categorized as "occupation" (3

Bias in content

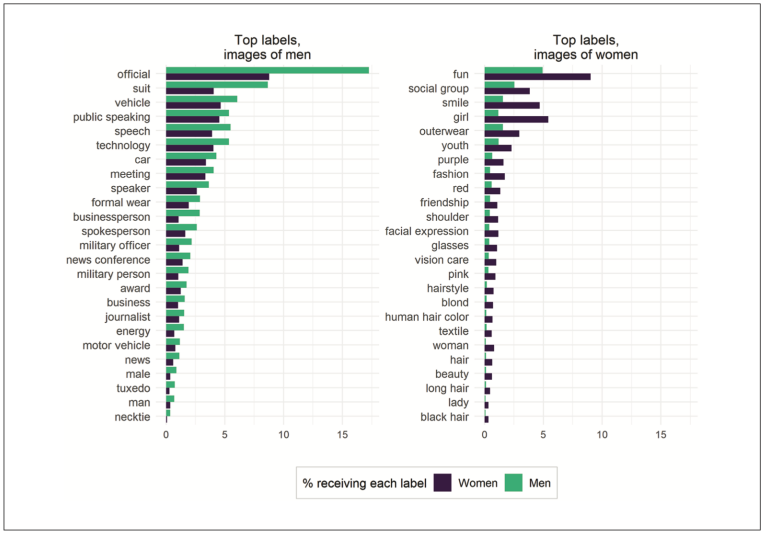


Figure 8. Google Cloud Vision labels applied to found data set (Twitter images). The 25 most gendered labels for men and women were identified using χ^2 tests ($p \leq .01$). Labels are sorted by absolute frequencies. Bars denote the percentage of images for a certain

Other image recognitions systems

We also examined Microsoft Azure Computer Vision and Amazon Rekognition and found similar patterns

Bias in content

If “a picture is worth a thousand words,” but an algorithm gives us only a handful, then which words it chooses matters a lot

- If labels that an algorithm picks to describe images are systematically biased along important social dimensions such as gender, they may retrench rather than shed light on patterns of social inequality

HOME > TECH

Google AI will no longer use gender labels like 'woman' or 'man' on images of people to avoid bias

Shona Ghosh Feb 20, 2020, 12:12 PM

Take away

Proceed with caution when interpreting such image data:

- These systems don't directly *see* what it is in a picture but infer content based on statistical patterns in training data
  - This training data may have unknown biases baked into them which are then reproduced by the system