Literate Models for Computer Vision: Combining vision, language and reading

Dimosthenis Karatzas (*dimos@cvc.uab.es*)

Centre de Visió per Computador

MCV 2023


Intelligent Reading Systems
Joint visual and textual modelling
Natural language interfaces
“Purchase some chocolate cookies without gluten”

“Purchase some chocolate cookies without gluten”

“Purchase some chocolate cookies without gluten”

Nutrition Facts

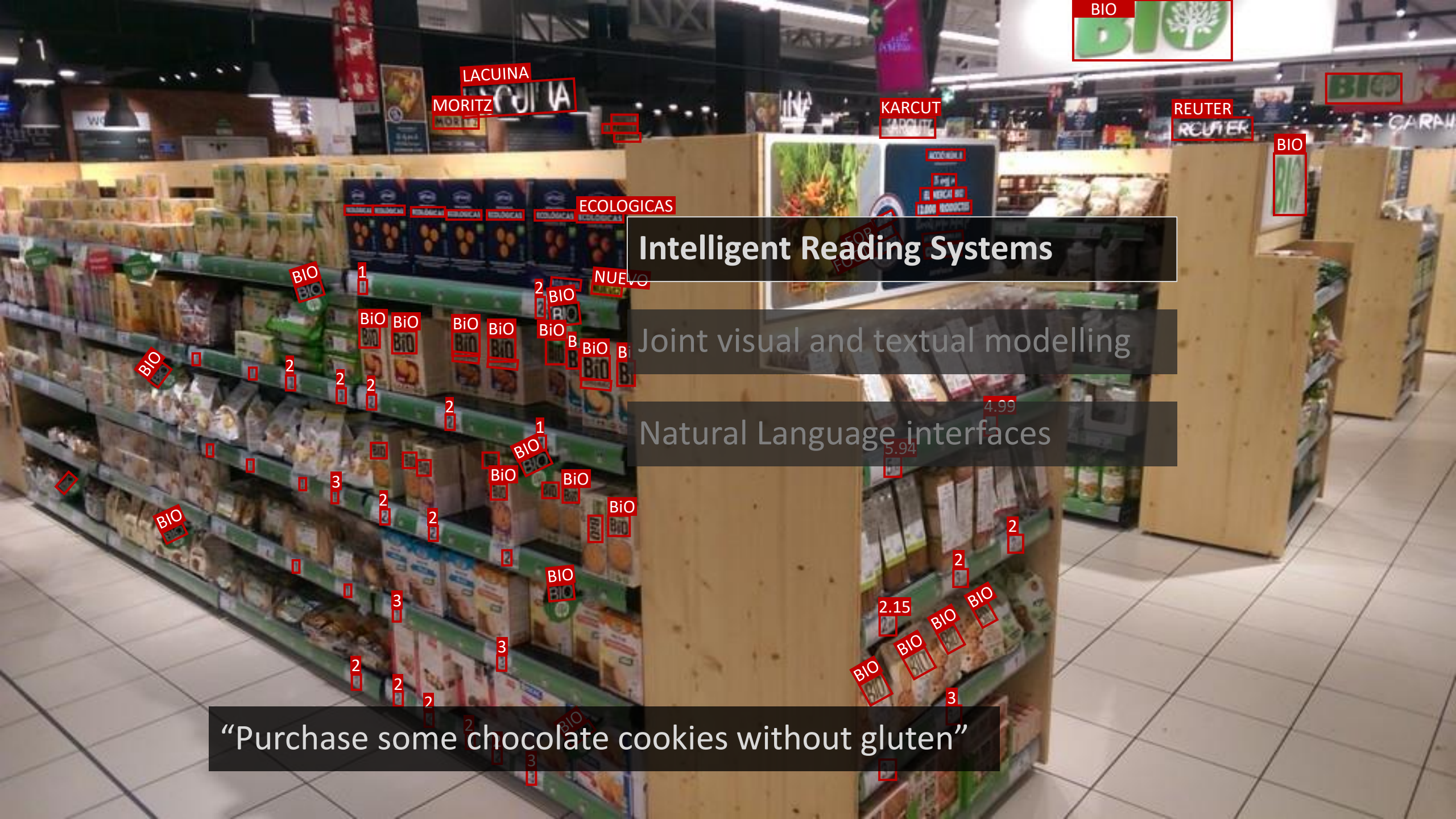
Serving Size 3 Cookies (32g/1.1oz)
Servings Per Container About 5

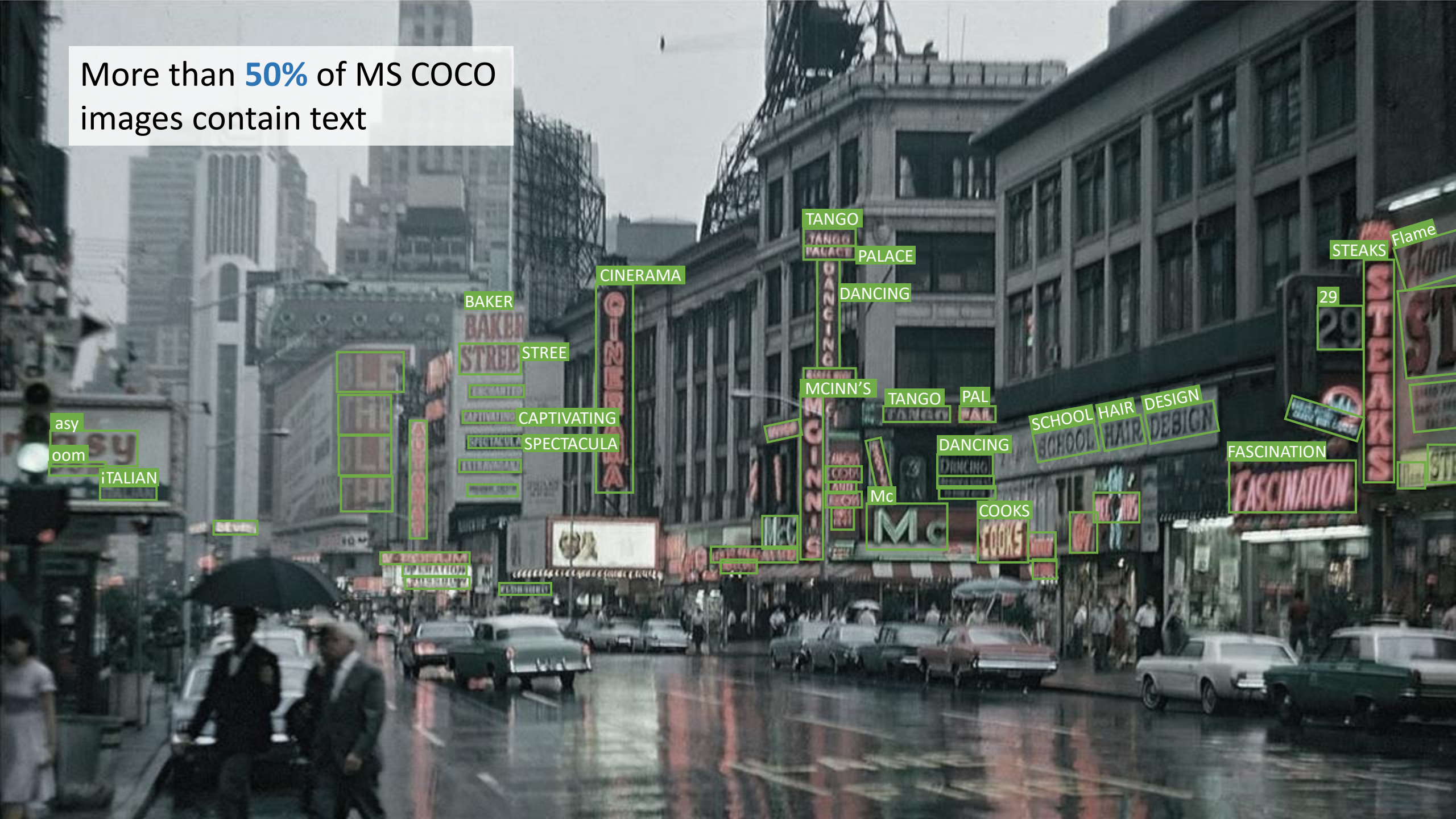
Amount Per Serving	
Calories 130	Calories from Fat 40
	% Daily Value*
Total Fat 4.5g	7%
Saturated Fat 1.5g	6%
Trans Fat 0g	0%
Cholesterol 0mg	0%
Sodium 180mg	7%
Total Carbohydrate 21g	7%
Dietary Fiber 2g	6%
Sugars 13g	
Protein 2g	

*Percent Daily Values are based on a 2,000 calorie diet. Your daily values may be higher or lower depending on your calorie needs. Not a significant source of Vitamin A, Vitamin C, Calcium and Iron.

***Serving* contains 1/44 cup gluten-free oats**


Intelligent Reading Systems
Joint visual and textual modelling
Natural Language interfaces
“Purchase some chocolate cookies without gluten”


Intelligent Reading Systems
Joint visual and textual modelling
Natural Language interfaces
“Purchase some chocolate cookies without gluten”


More than 50% of MS COCO images contain text
asy
oom
iTALIAN
BAKER
STREE
CINERAMA
CAPTIVATING
SPECTACULA
TANGO
PALACE
DANCING
MCINN'S
TANGO
PAL
DANCING
Mc
COOKS
SCHOOL
HAIR
DESIGN
FASCINATION
29
STEAKS
Flame

Text Understanding in the Wild

S.R. Battu, M. Mathew, L. Gomez, M. Russinyol, D. Karatzas, C.V. Jawahar, "RoadText-1K : A Dataset for Text Detection and Recognition in Driving Videos", ICRA 2020

R. Gomez, A. Biten, L. Gomez, J. Gibert, D. Karatzas, M. Rusiñol. "Selective style transfer for text". ICDAR 2019

L. Gómez, A. Mafla, M. Rusinol, D. Karatzas. "Single Shot Scene Text Retrieval". ECCV 2018

L. Gómez, & D. Karatzas, "Textproposals: a text-specific selective search algorithm for word spotting in the wild". Pattern Recognition, 2017

L. Gomez, et al, "Improving patch-based scene text script identification with ensembles of conjoined networks", Pattern Recognition, 2017

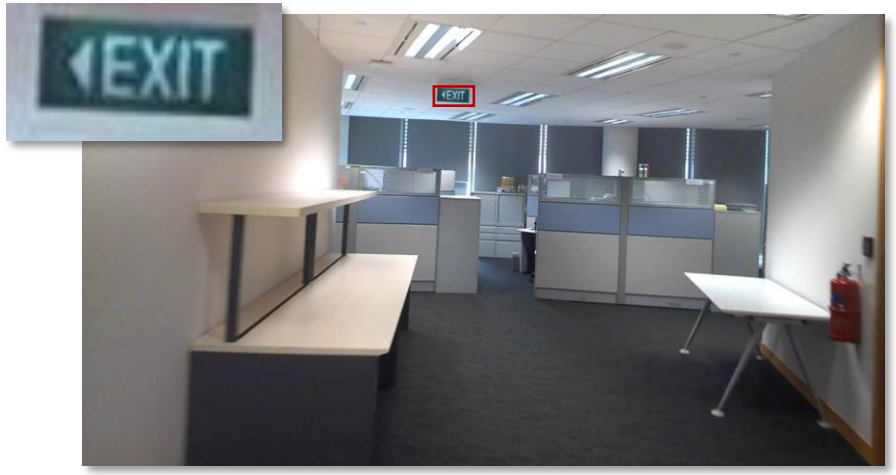
Typical Tasks

- Detection / Localization
- Recognition
- Word spotting
- Information spotting
- Scene text retrieval
- Script identification
- Video text / text tracking
- Text segmentation / replacement
- Text style transfer

Open Challenges

- Video text / text tracking
- Multi-script / multi-lingual scenarios
- Uncontrolled conditions (night, rain, etc)
- etc

Not a solved problem…



Small text instances

Out of vocabulary text

Text in motion

Multiple scripts

CV models that cannot read...

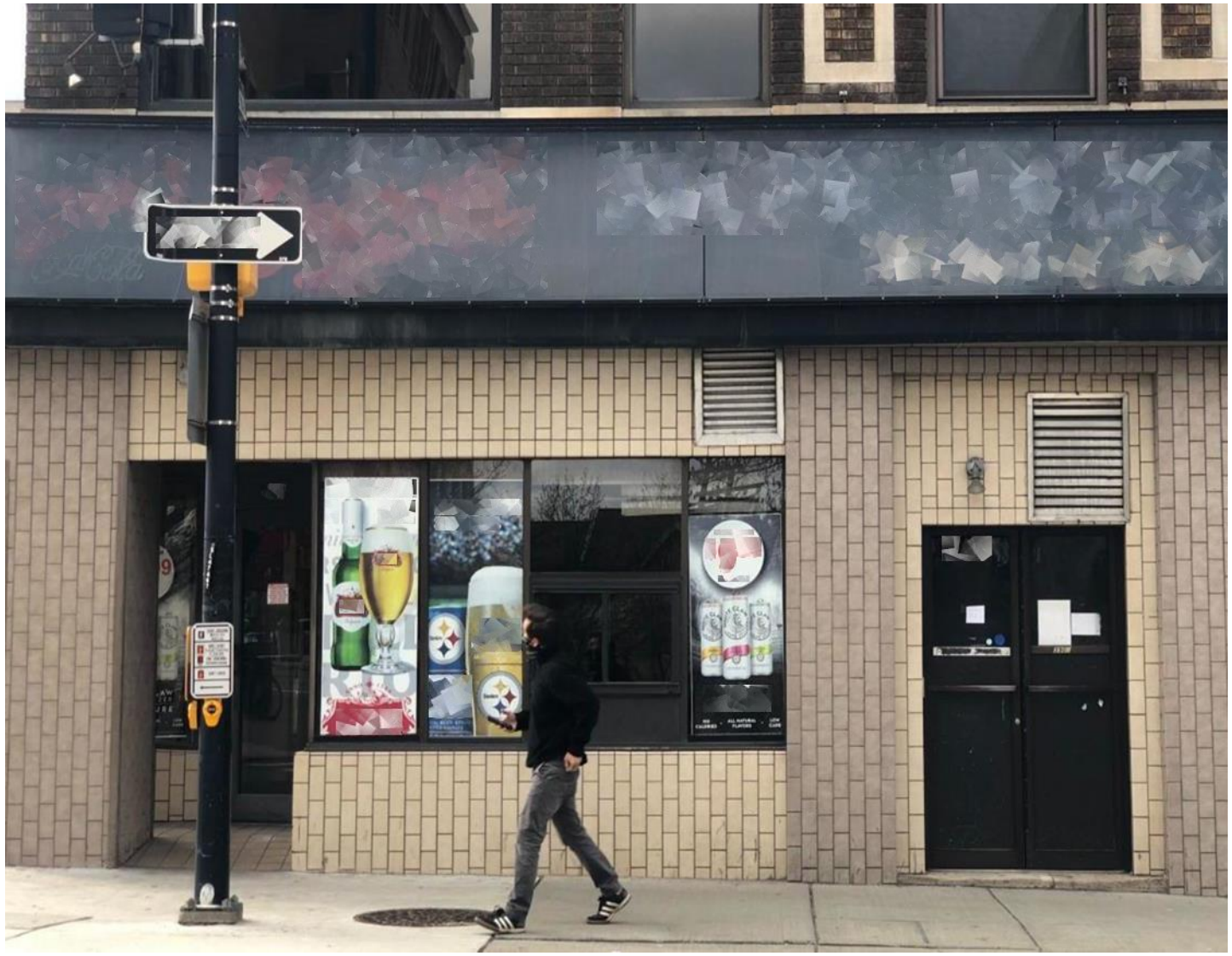


Image classification and retrieval

Building

Visual Question Answering

Q: What is the name of the restaurant?

UNKNOWN

Image Captioning

A man walking down the street.

CV models that can read

Image classification and retrieval

- Building
- Hot Dog Restaurant

Visual Question Answering

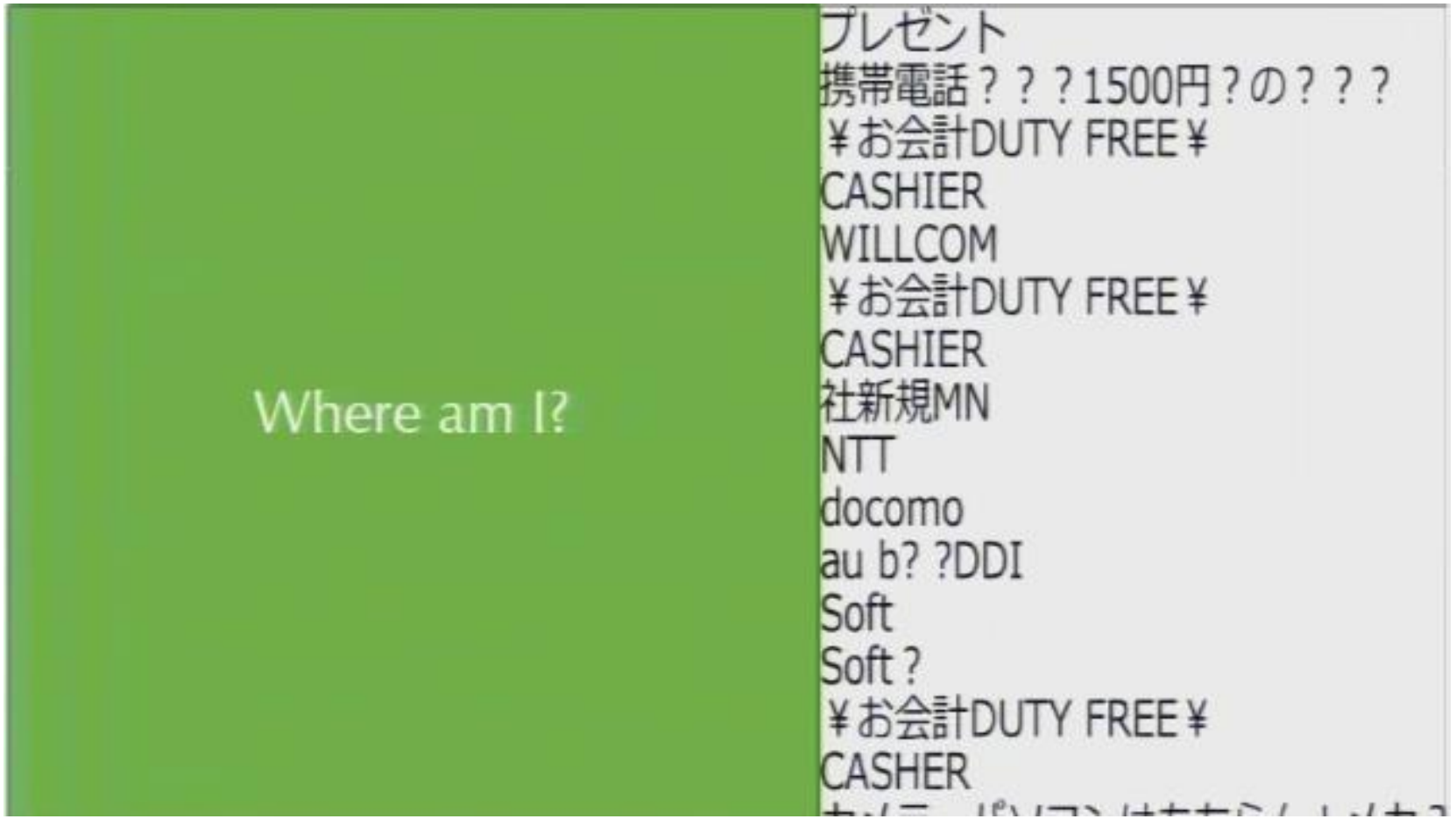
Q: What is the name of the restaurant?

- UNKNOWN
- Essie's Original

Image Captioning

- A man walking down the street.
- A man walking in front of Essie's hot-dog shop.

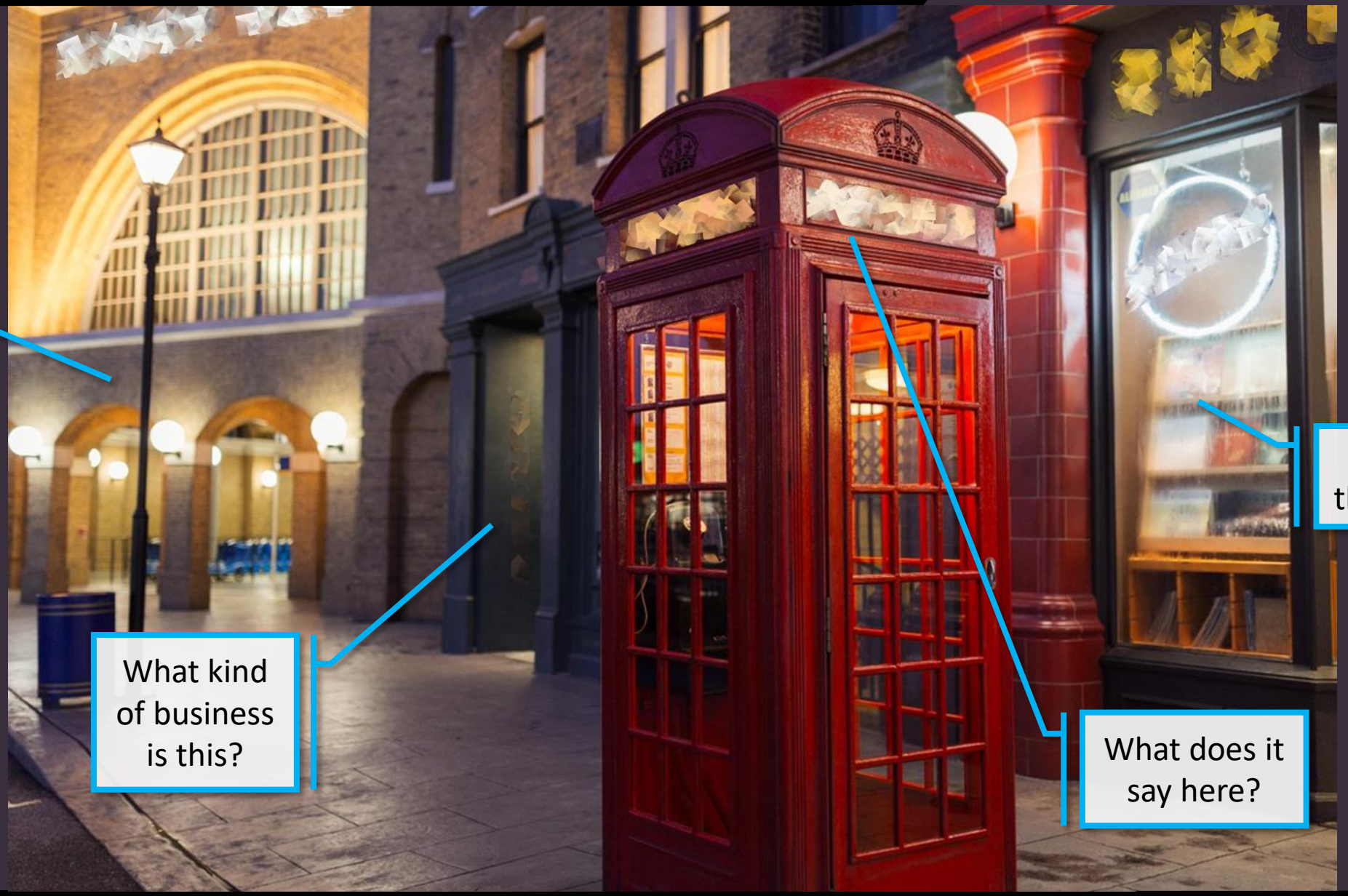
How much does scene text reveal about our environment?



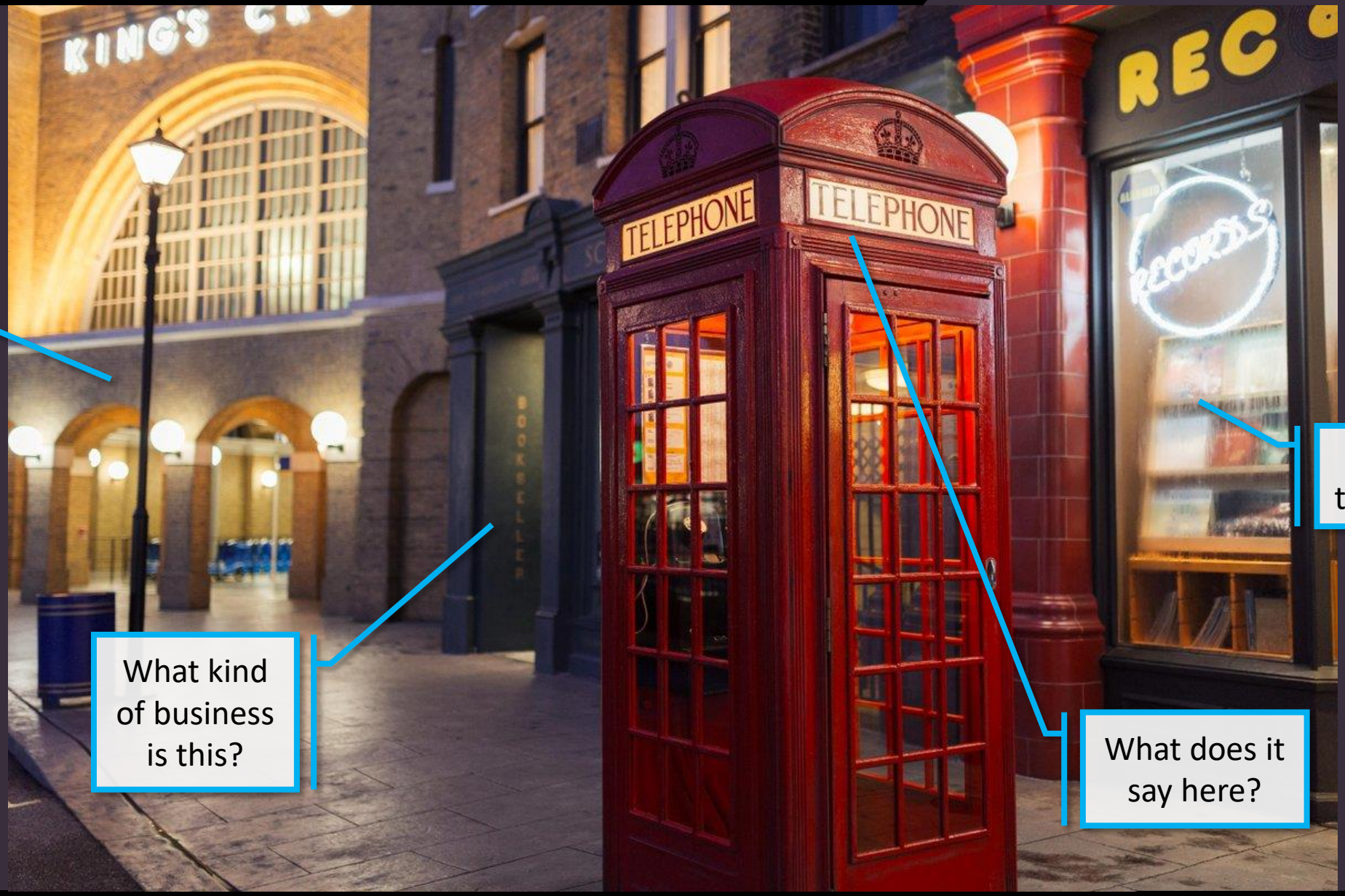
Video from Seiichi Uchida, CVC Seminar, 2017


Reading Systems
Joint visual and textual modelling
Natural language interfaces
“Purchase some chocolate cookies without gluten”

Scene understanding



Scene understanding



The interplay of textual and visual info

Correlated

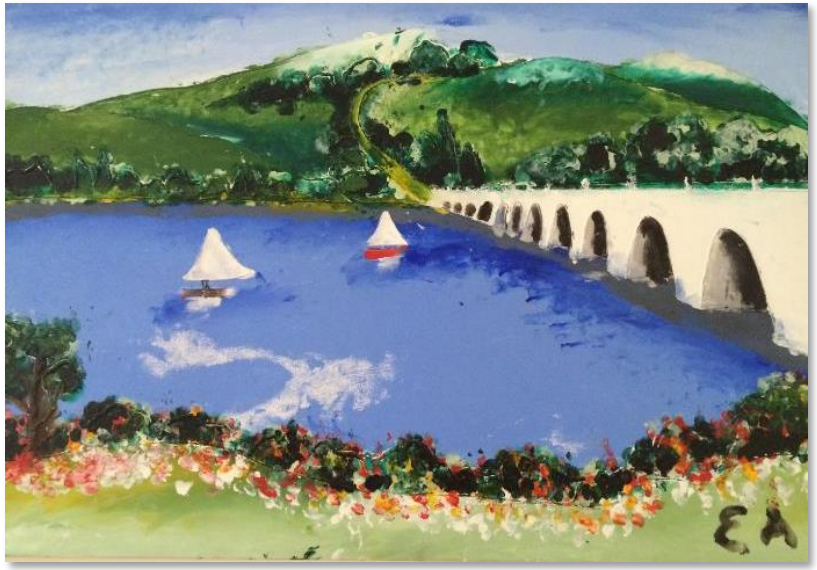
Complementary

Reinterpreting each other

Orthogonal, distractor


Reading Systems
Joint visual and textual modelling
Natural language interfaces
“Purchase some chocolate cookies without gluten”

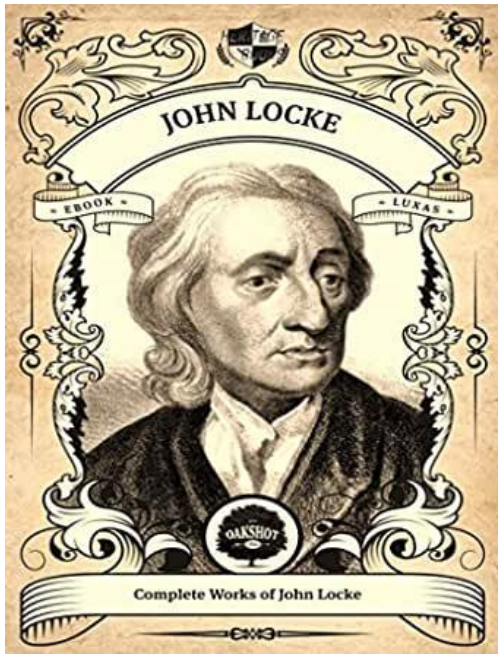
How is appearance information acquired in the absence of sensory access?



Paintings by **born-blind** artist
Eserf Armagan (1953-present)

How is appearance information acquired in the absence of sensory access?

The learn-from-description hypothesis



Blind individuals learn from sighted people's **verbal descriptions**

Learn-from-kind hypothesis

"In the absence of direct sensory access, knowledge of appearance is acquired **primarily through inference**, rather than through memorization of verbally stipulated facts."

"In other words, **language serves as an indirect source of information** about appearance by providing information about ontological kind."

Learning from verbal descriptions and learning via inferences are deeply intertwined

J. Kim et.al. "Knowledge of animal appearance among sighted and blind adults", PNAS 2021

M. Ostarek et al, "Sighted people's language is not helpful for blind individuals' acquisition of typical animal colors", PNAS 2019

Computer Vision models learn how to “read” scene text anyway…

Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%

Granny Smith	0.1%
iPod	99.7%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.0%

The “Spiderman” multi-modal Neuron
Neuron 244 from penultimate layer in CLIP RN50x4

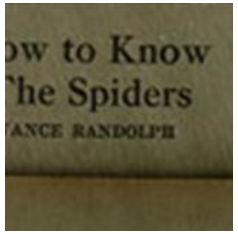
***Literal**: photos of Spider-Man in costume and spiders*




***Conceptual**: comics or drawings of Spider-Man and spider-themed icons*




***Symbolic**: text “spiders”*



G. Goh et.al. “Multimodal Neurons in Artificial Neural Networks”, Distill, 6(3), e30, 2021

Multimodal neurons

Unit 122

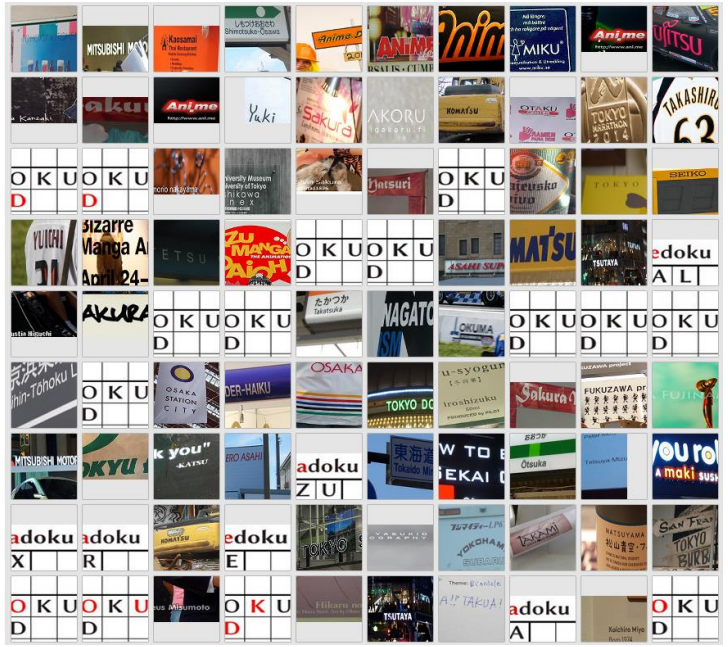
“Things from Japan”

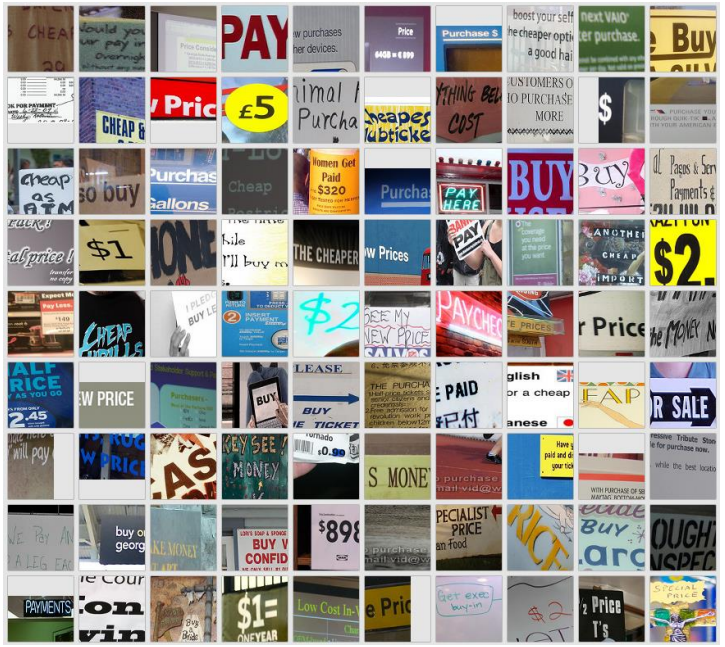
Unit 1330

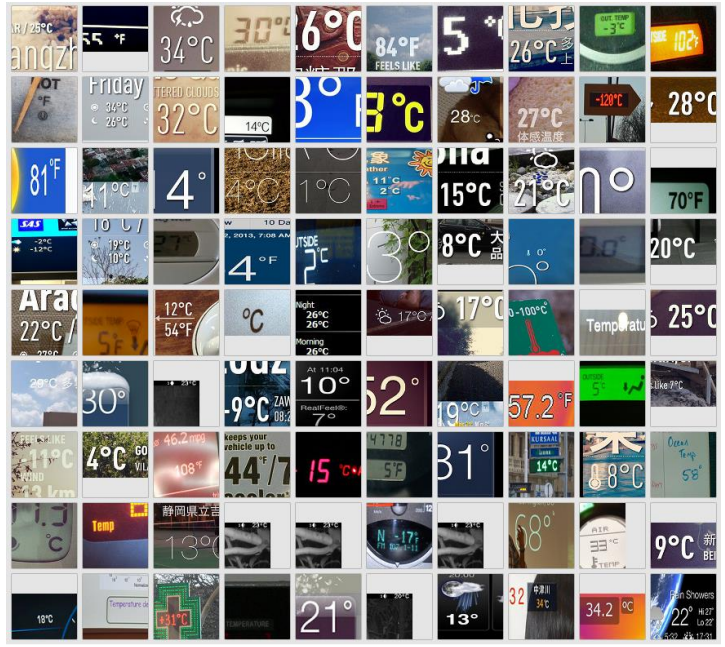
“Cash and buying”

Unit 73

“Weather”

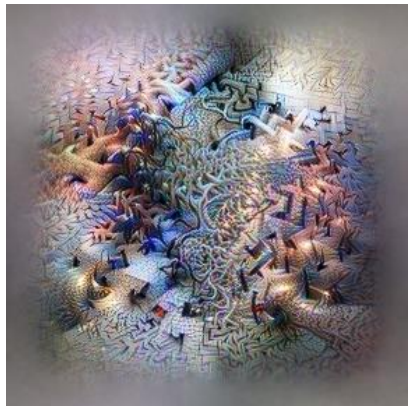






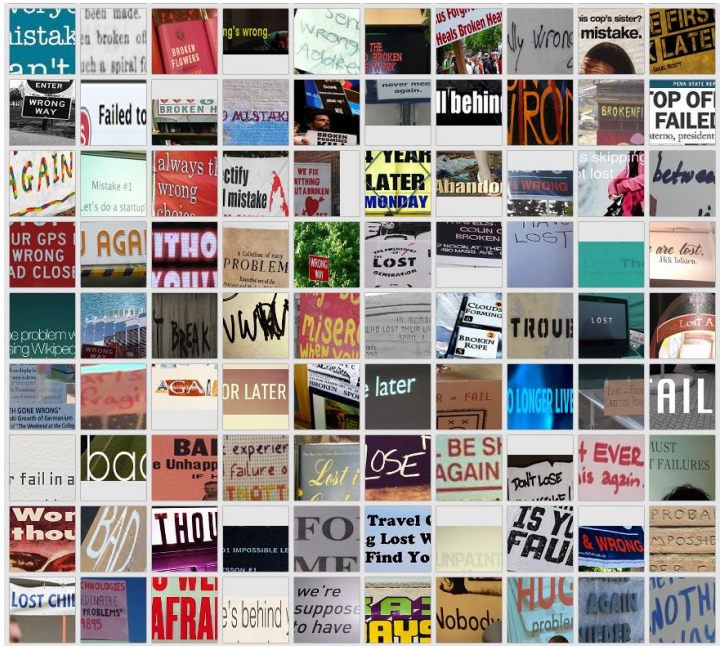
Try it yourself with “OpenAI Microscope”: https://microscope.openai.com/

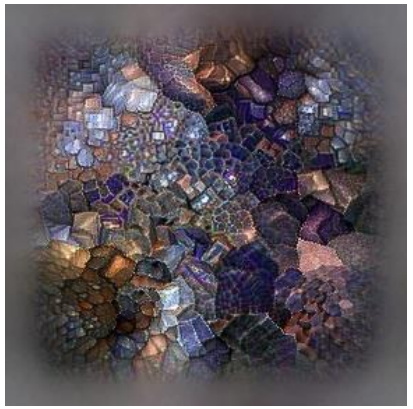
Multimodal neurons



Unit 76

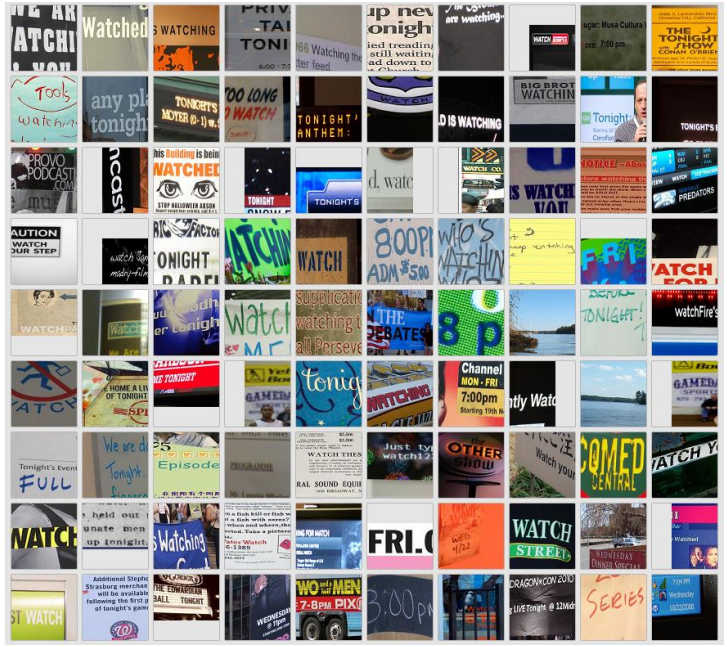
“Pesimistic”





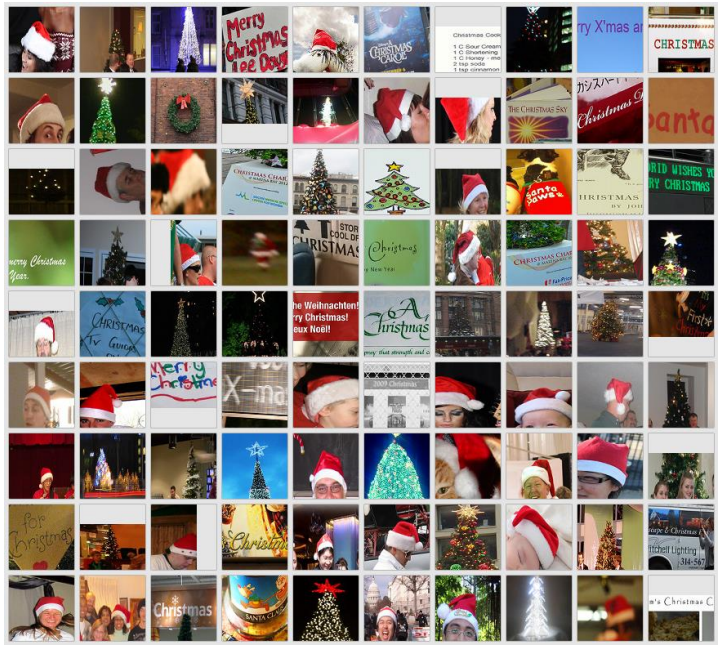
Unit 142

“Watch TV tonight”



Unit 1326

“Christmas”



Try it yourself with “OpenAI Microscope”: https://microscope.openai.com/

Multimodal neurons

Unit 1384

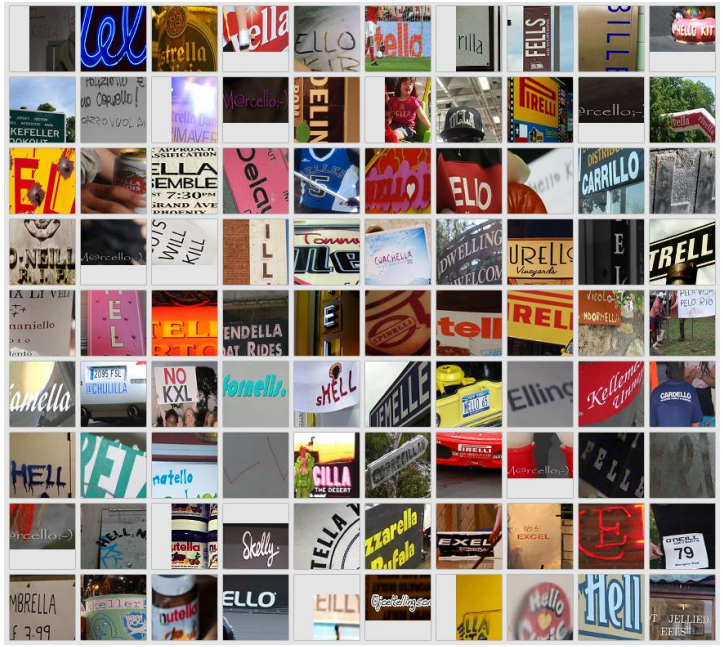
“ELL”

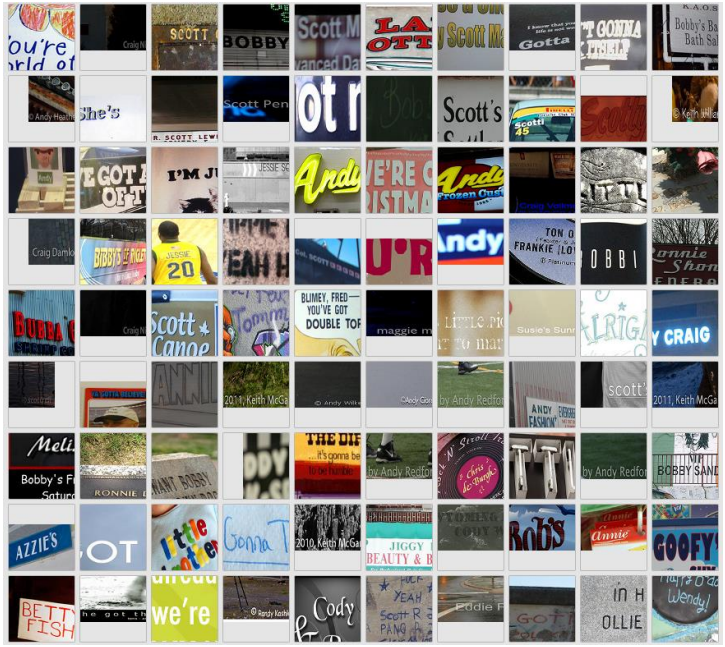
Unit 1371

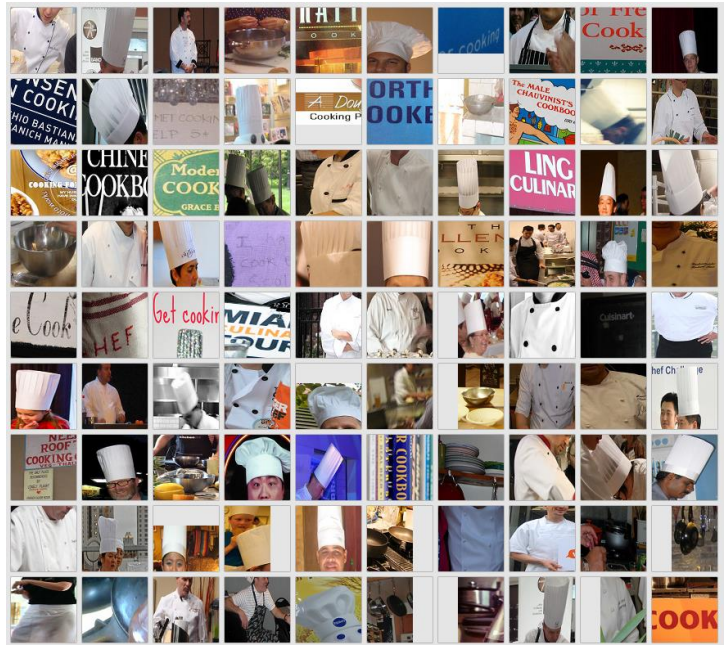
“First Names”

Unit 1379

“Cooking”

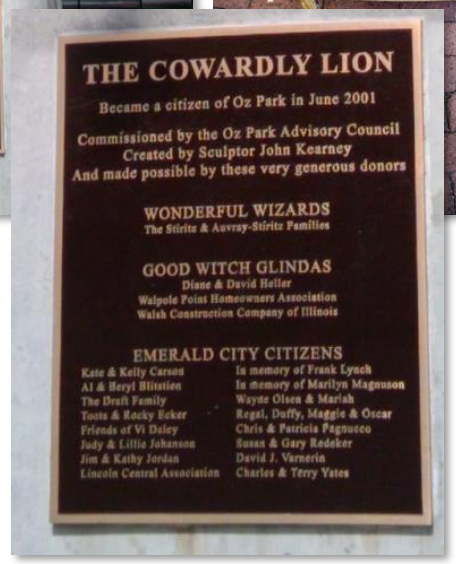


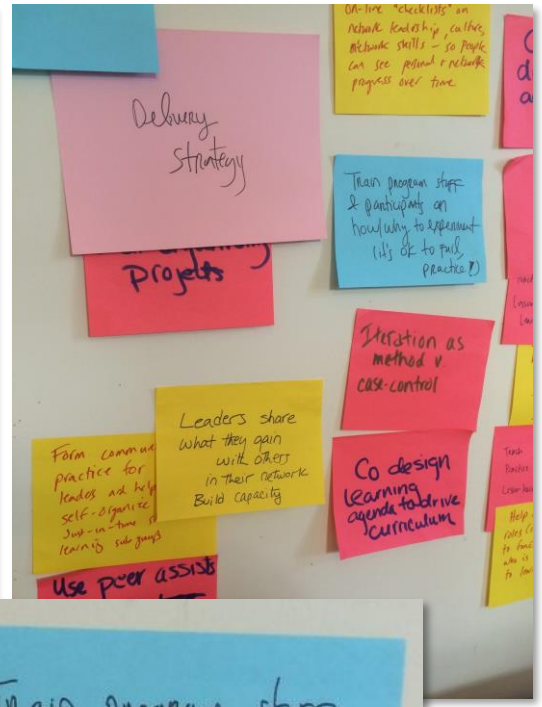


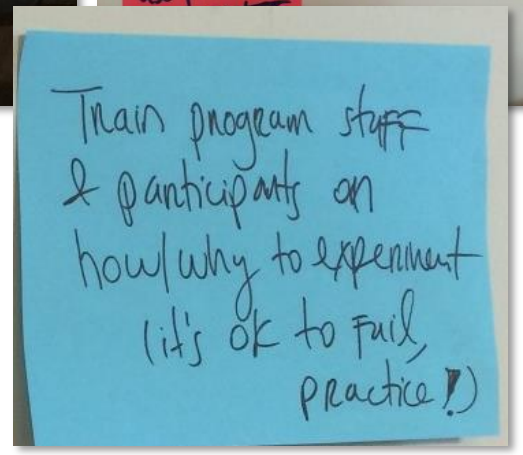


Try it yourself with “OpenAI Microscope”: https://microscope.openai.com/

The fuzzy frontier between scene text and documents







A. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusiñol, E. Valveny, C.V. Jawahar, D. Karatzas, "Scene Text Visual Question Answering", ICCV 2019
S. Long, et. al. "Towards End-to-End Unified Scene Text Detection and Layout Analysis", CVPR 2022

Understanding complex written communication