The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Spring 2024

INTRODUCTION TO COMPUTER VISION

Atlas Wang
Associate Professor, The University of Texas at Austin

Visual Informatics Group@UT Austin
https://vita-group.github.io/

Overview of Course Logistics

- We meet on Tuesday & Thursday 2:00-3:30pm (EER 1.518)
  - Class format: in-person **(video record available on Canvas)**
  - Do I have to come to the classroom?
  - Can I audit?
  - After-class communication: **Slack (link sent) – IMPORTANT!**
  - Class materials are distributed on **Course Webpage (NOT Canvas)**: https://vita-group.github.io/spring_24.html
- We do not follow any textbook closely. Instead we will have many "recommended materials".

Overview of Course Logistics

- Instructor Office Hour： **Wednesday 10:00am - 11:00am, meet at Office EER 6.886**
- This class has two TAs:
  - **Wenyan Cong**, wycong@utexas.edu, Office Hour: **Monday 4:00-5:00pm**
  - **Runjin Chen**, chenrunjin@utexas.edu , Office Hour: **Friday 4:00-5:00pm**
  - Both TA slots: meet outside EER 3.854
- *Which office hour should I come to?*
- Online Q&A: ***anytime, just ask on Slack!***

Grading

- **Homework: 20%**
  - There will be 4 mandatory written or machine assignments, 5% each.
  - One 5th homework (5%) will be given as "optional" bonus
  - HOMEWORK 0 out **today**! Due next Wednesday (1/24) EOD
- **Mid-term exam: 30% (3/21 in class). No final exam**
- **Final Project: 50%**
  - Proposal (**10%**) Due by the end of **Week 5** (2/18 Sunday): **2-Page** report, including title, team member, problem description, preliminary literature survey, the proposed technical plan, and references
  - Mid-Report (**10%**) Due by the end of **Week 11** (4/02 Sunday): **4-Page** report
  - Presentation (**5%**): submitted in the form of 5-min precoded videos
  - Code review (**10%**): Write clean, well-documented and runnable codes, PLEASE
  - Final Report (**15%**): **8-page** report following the standard CVPR paper template
    - **Template file:** http://cvpr2020.thecvf.com/sites/default/files/2019-09/cvpr2020AuthorKit.zip

What is Computer Vision?

- An **interdisciplinary** field that deals with how computers can be made for gaining holistic understanding from digital images or videos.
- From the engineering perspective, it seeks to automate tasks that the human visual system can do.

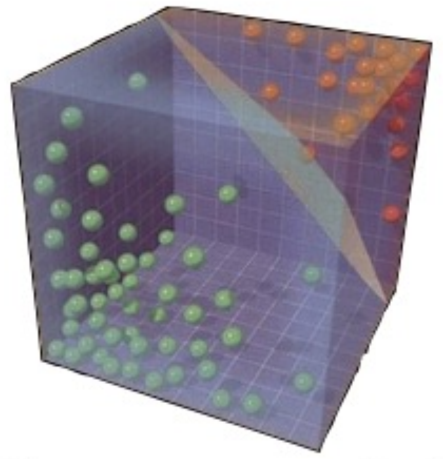
Computer Vision as Input-Output System:

- Input: images or video
- Output (ideally): description or understanding of the visual world, in a “human” way
- Outputs (practically): reconstructing, measuring, classifying, interpreting…

A Conceptual Visual Perception Pipeline

The parts that we are most interested in

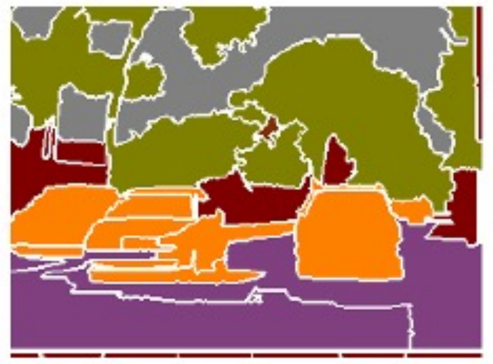
representation



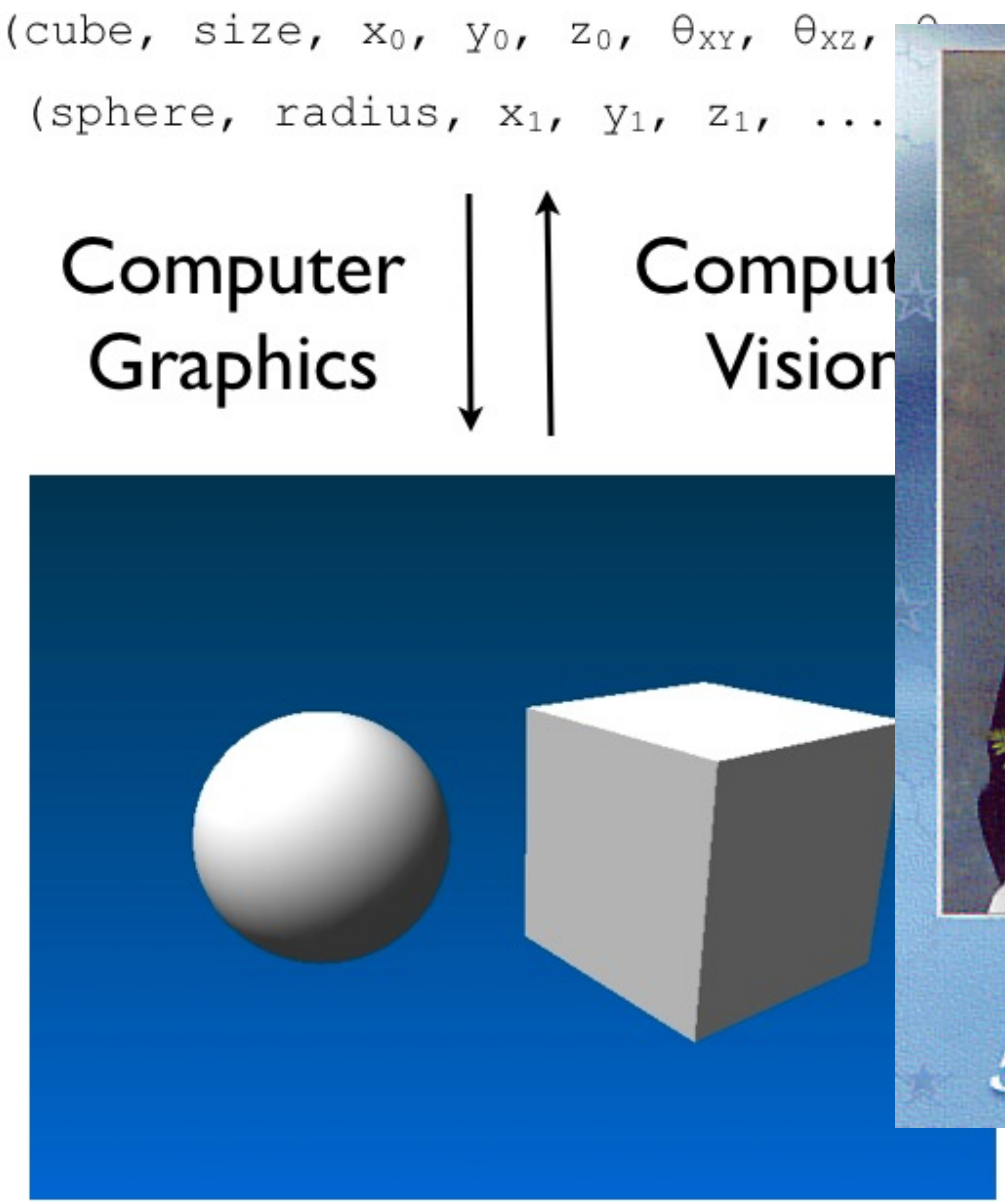
'fancy math'

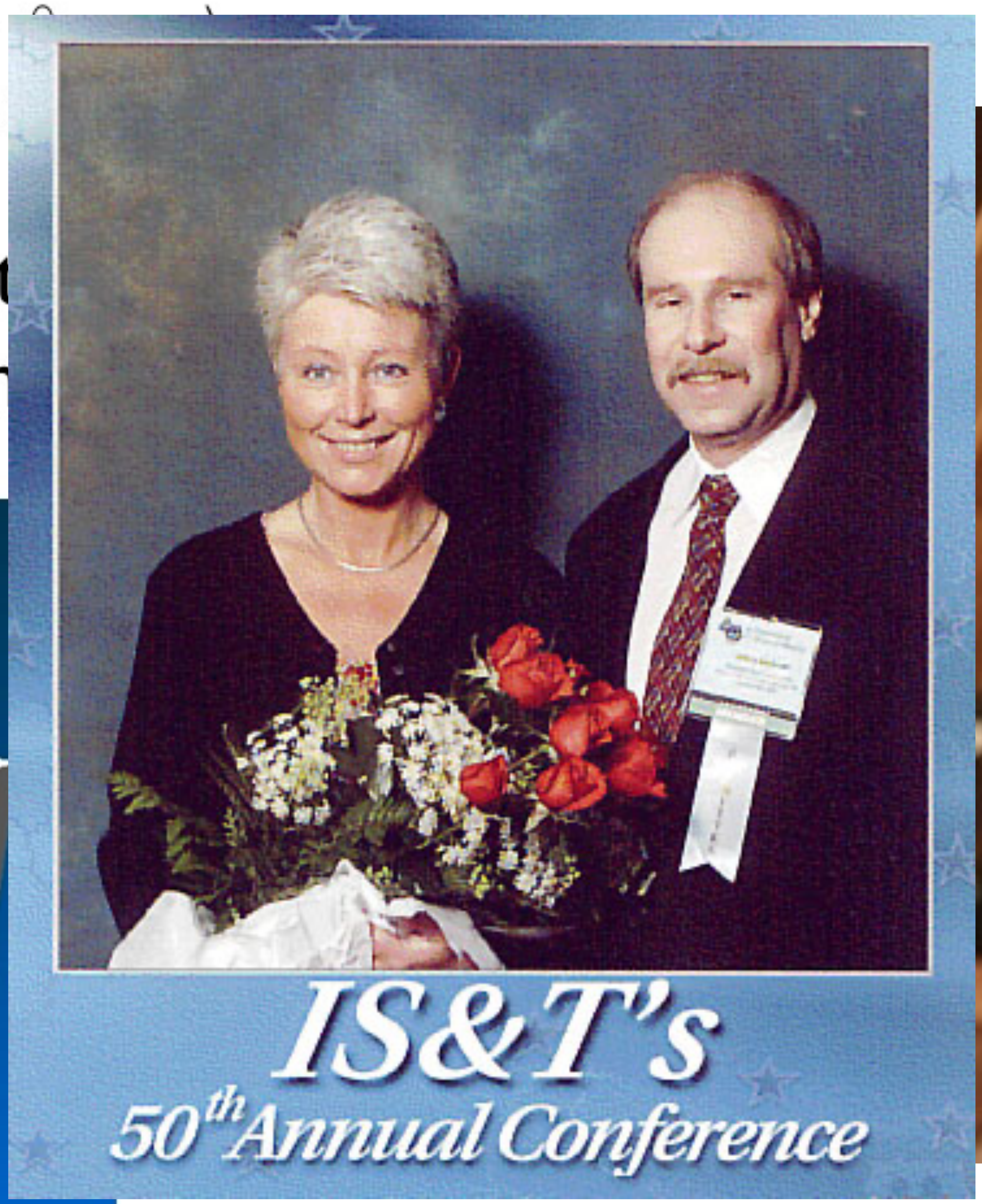
output

what should we look at?
(image features)



what can we understand?
(semantic segmentation)



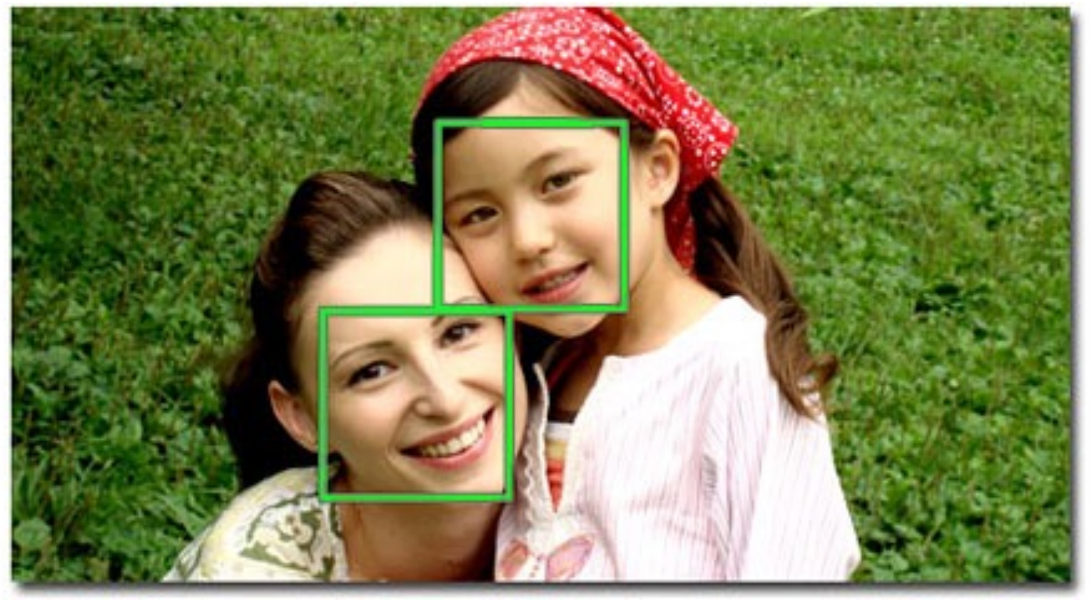


Computer Vision and Computer Graphics are often viewed as “inverse operations”

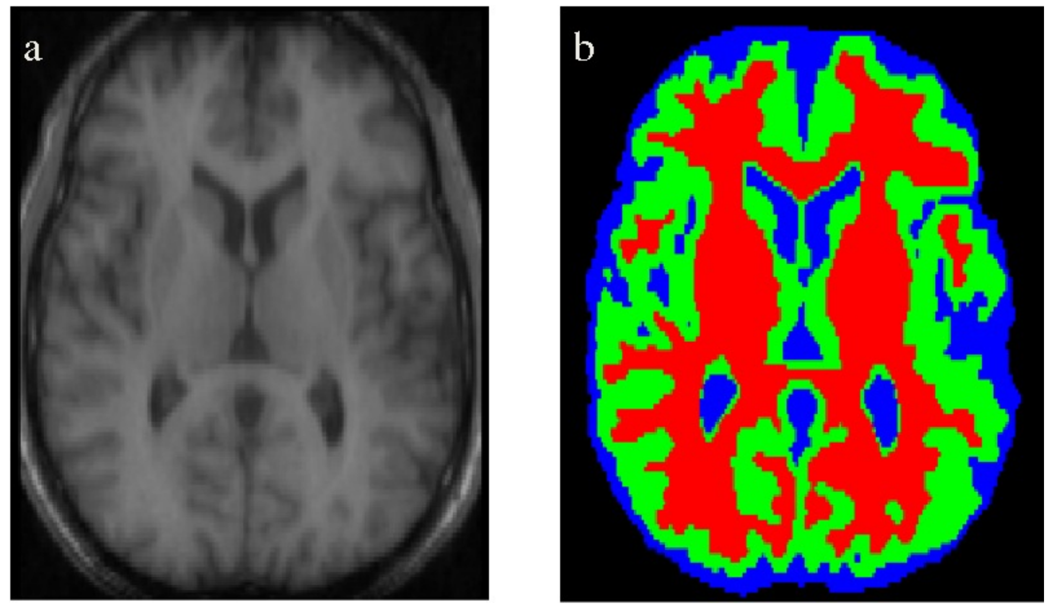
Computer Vision and Image Processing are significantly overlapped in their tools

(http://www.lenna.org/)

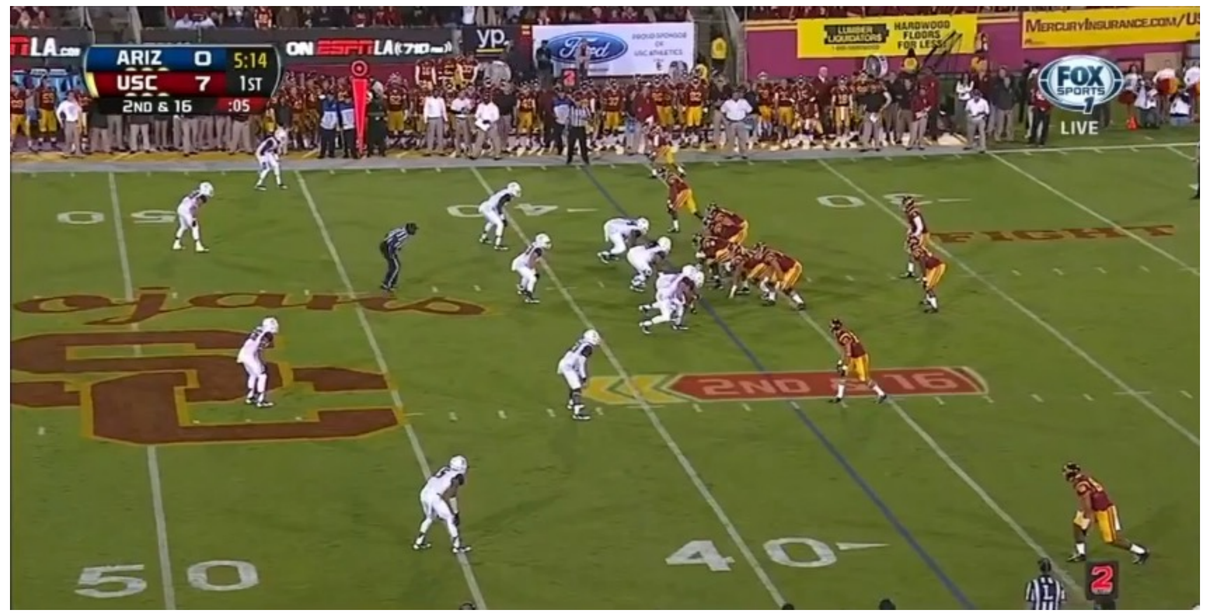
Computer Vision has SO MANY applications



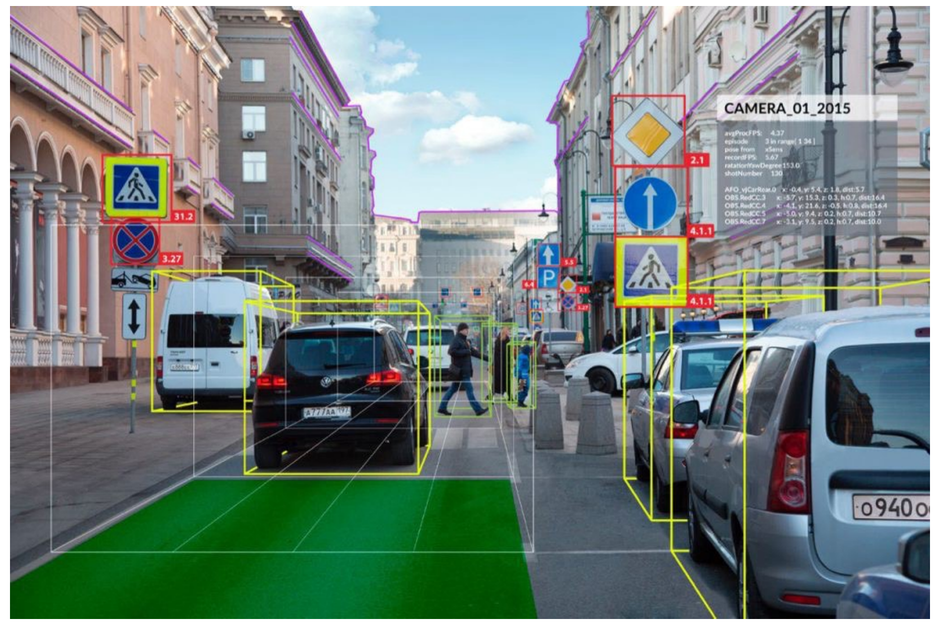
Face Detection/Smile recognition



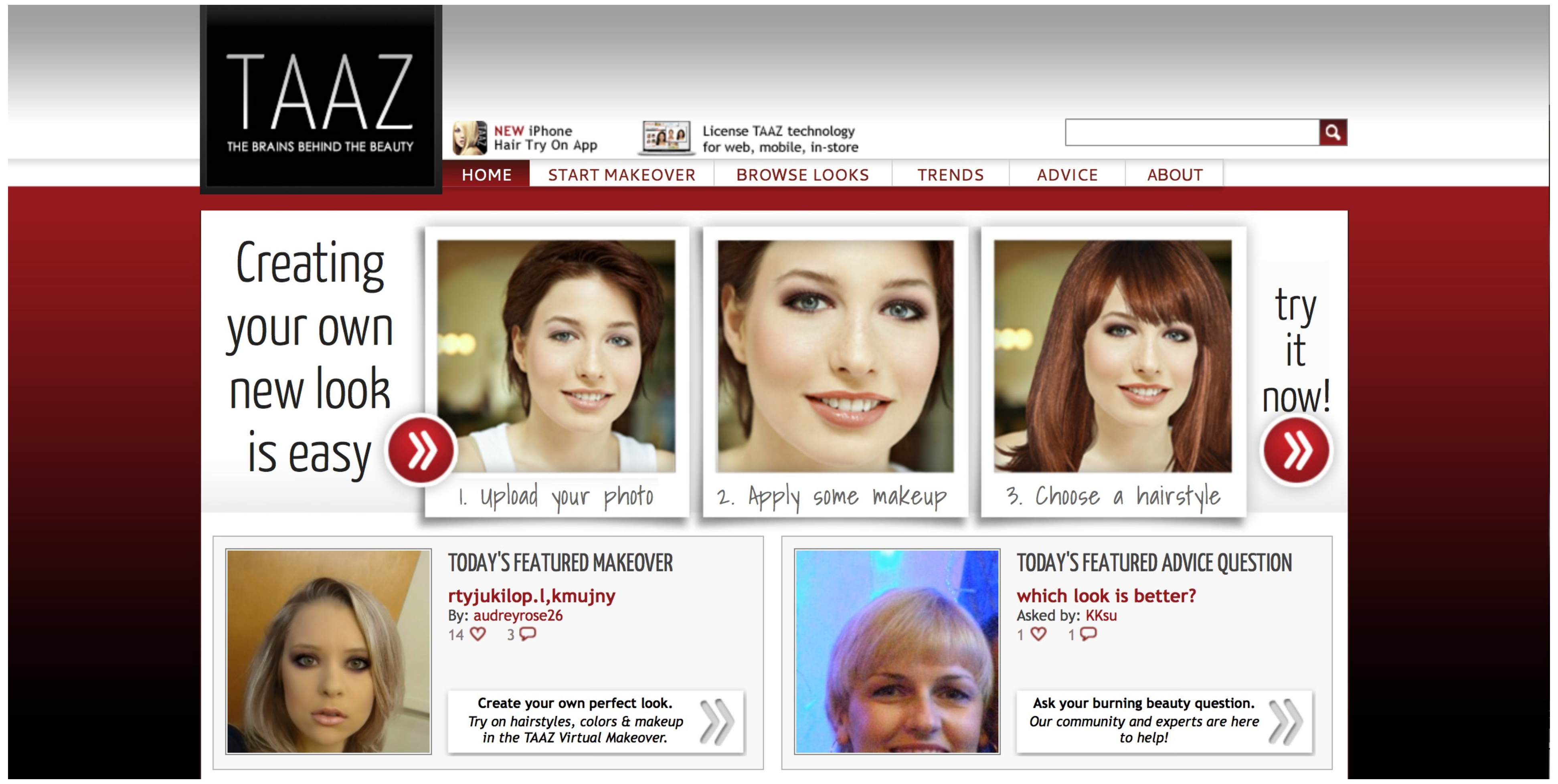
Medical Image Understanding



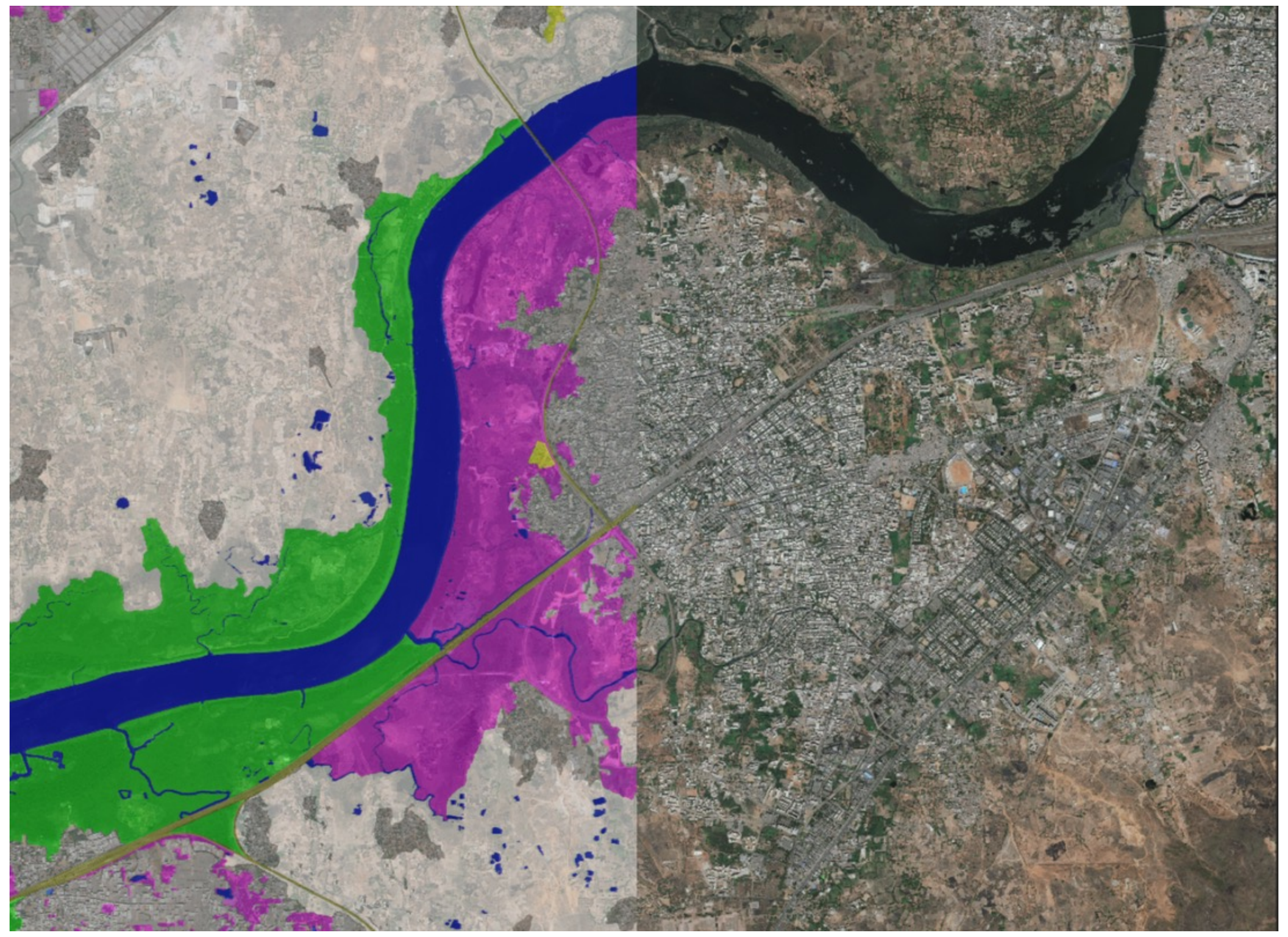
Tracking in Sports



Self-driving cars



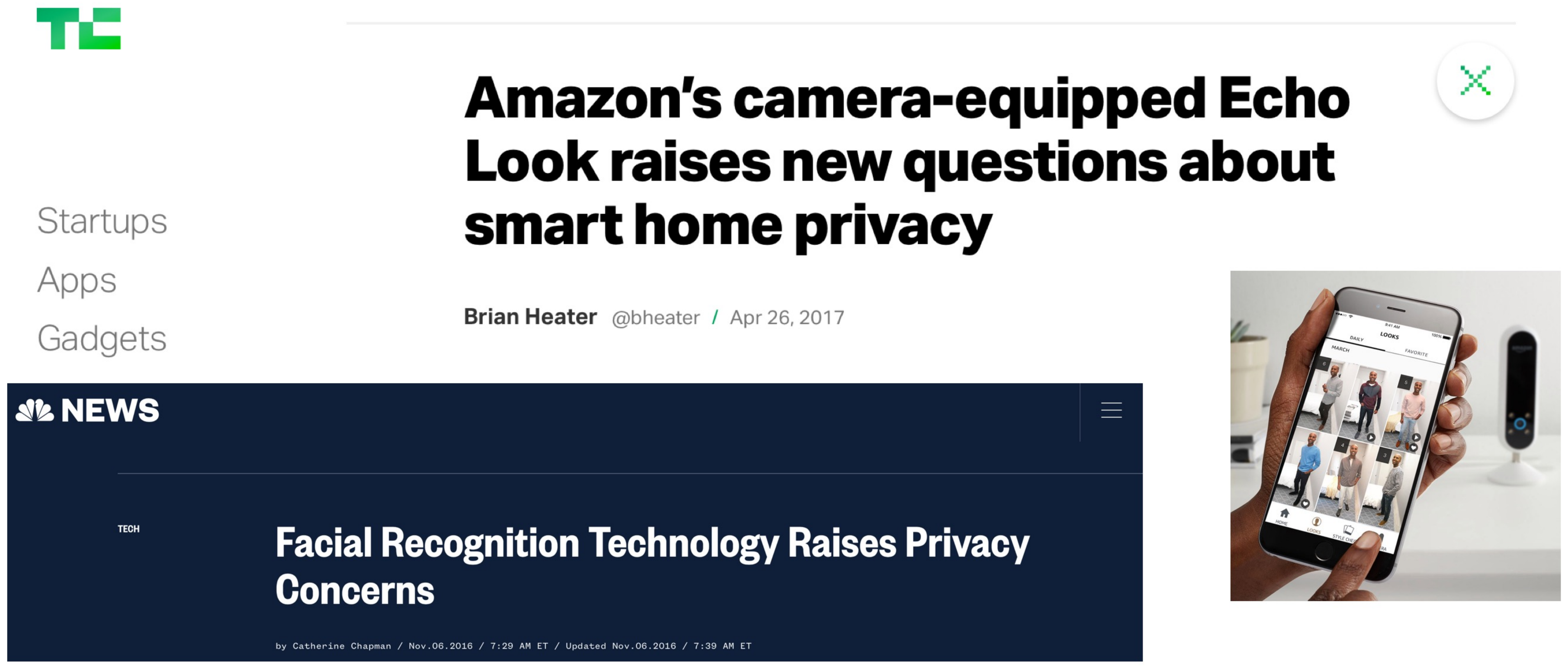
Face Makeover/Virtual try-on



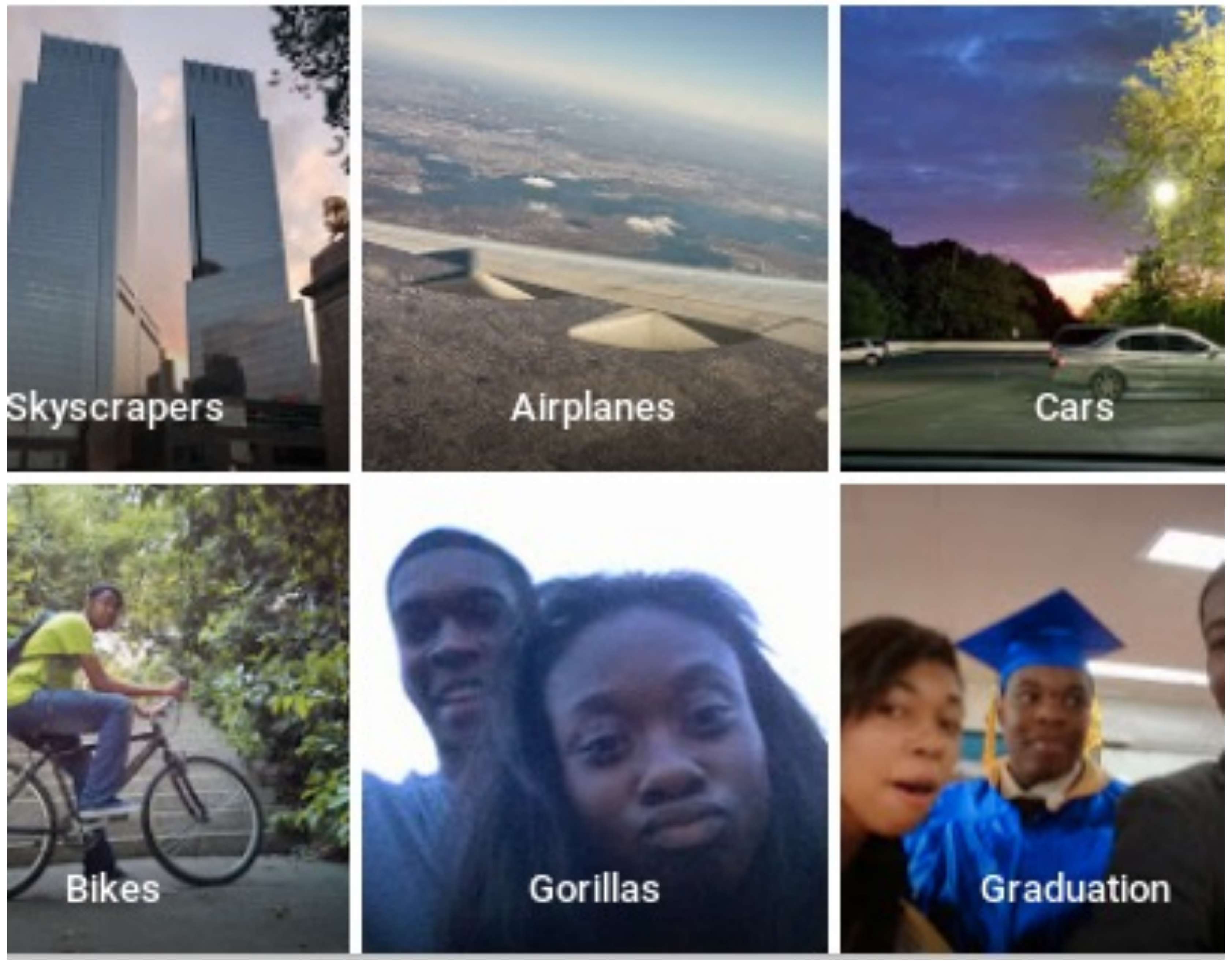
Remote sensing/earth mapping

Pose estimation (esp. fall detection)

… and Even More Open Challenges



Tesla autopilot failed to recognize white trailer against brightly lit sky [The Register]



https://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas/

Do you know?

The first "Computer Vision" work in this world was originally a summer project given to an MIT undergraduate student

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

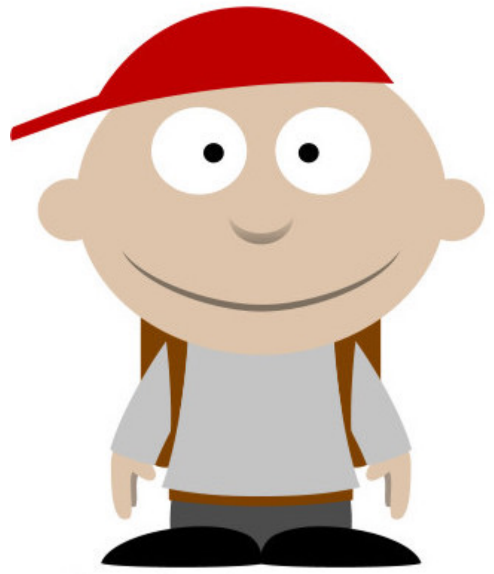
Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

Computer Vision Research History:
My (probably approximately correct) summary

- **Late 1960s**: CV was born = a branch of human vision and cognition research (*bio-inspired CV*)
- **1970s**: CV = estimate 3D structures from 2D images (*physically-grounded CV*)
- **1980s**: more rigorous math concepts such as scale space, texture analysis, contour models, as well as the emergence of optimization and inference methods
- **Early-to-mid 1990s:** camera calibration, multi-view stereo, scene reconstruction, image segmentation, the big boom of statistical learning methods
- **Late 1990s:** bridging CV and graphics: rendering, morphing, stitching…
- **2000s and after**: ML (graphical models, sparsity & low-rank), and finally Deep Learning ...

After near 60 Years... Computer Vision is Still Tough!

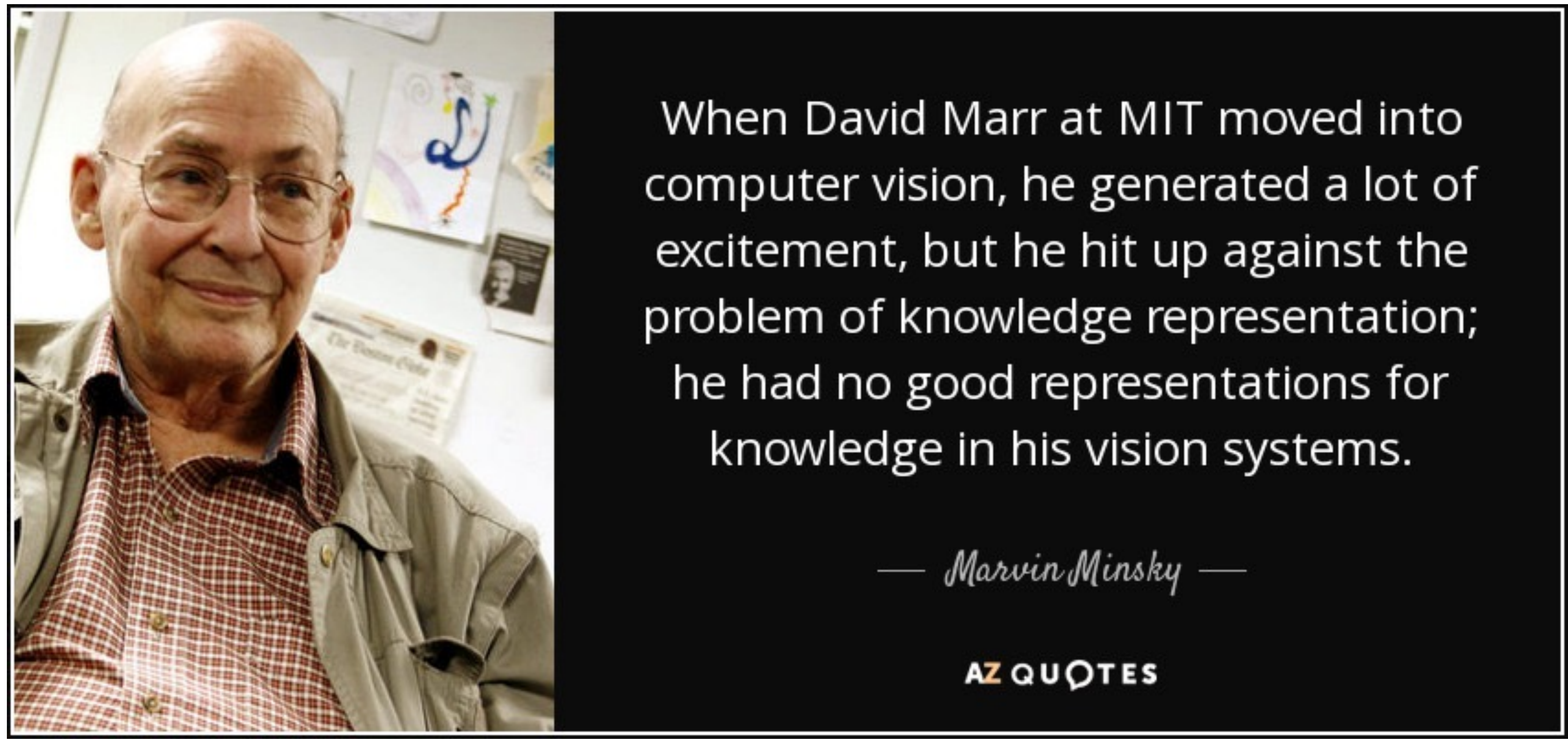


Atlas Wang

Hey Tom, What do you see as the biggest problem in computer vision?

One biggest problem of computer vision is – human never see in pixels!

Prof. Thomas S. Huang (1936 - 2020), ECE@UIUC
“A founding father in computer vision”



- Situation much the same as AI:
  - Some fundamental algorithms
  - Large collection of hacks / heuristics
- CV research is hard and “never ending”
  - Especially at high level, physiology unknown
  - Requires integrating many different methods
  - Requires reasoning and understanding: “AI completeness”

Overview of Class Structure & Agenda

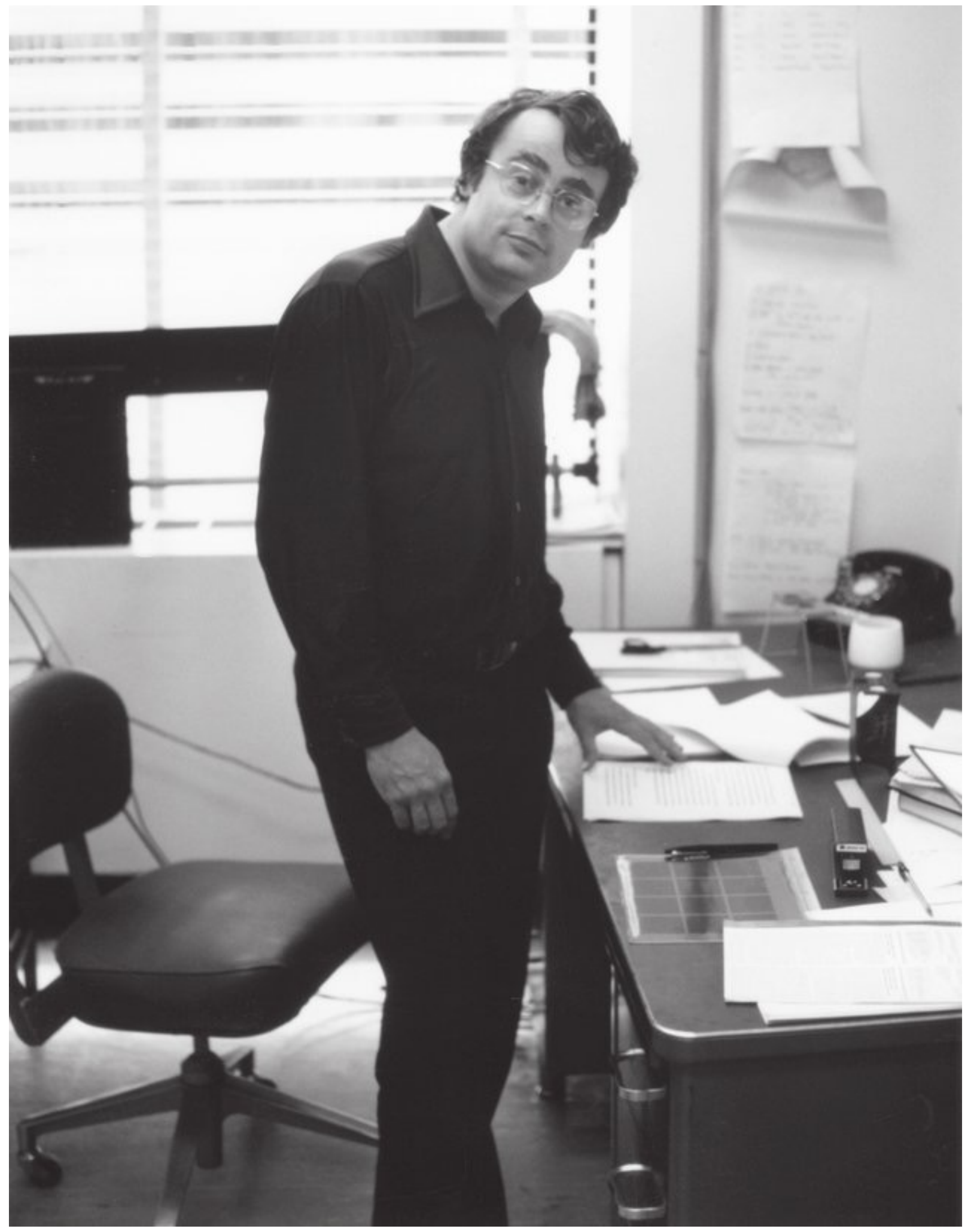
- **Section 1 (1/16 - 1/25):** Neuroscience, cognitive, and signal processing foundations of CV
- **Section 2 (1/30 – 2/15):** Extracting "good" features from 2D images *(keyword: describe & match)*
- **Section 3 (2/20 – 3/19):** From 2D to 3D vision and video *(keyword: geometry & motion)*
- **Section 4 (3/26 – 4/16):** Classical machine learning for CV tasks
- **Section 5 (4/18 - 4/25):** Modern deep learning for CV tasks

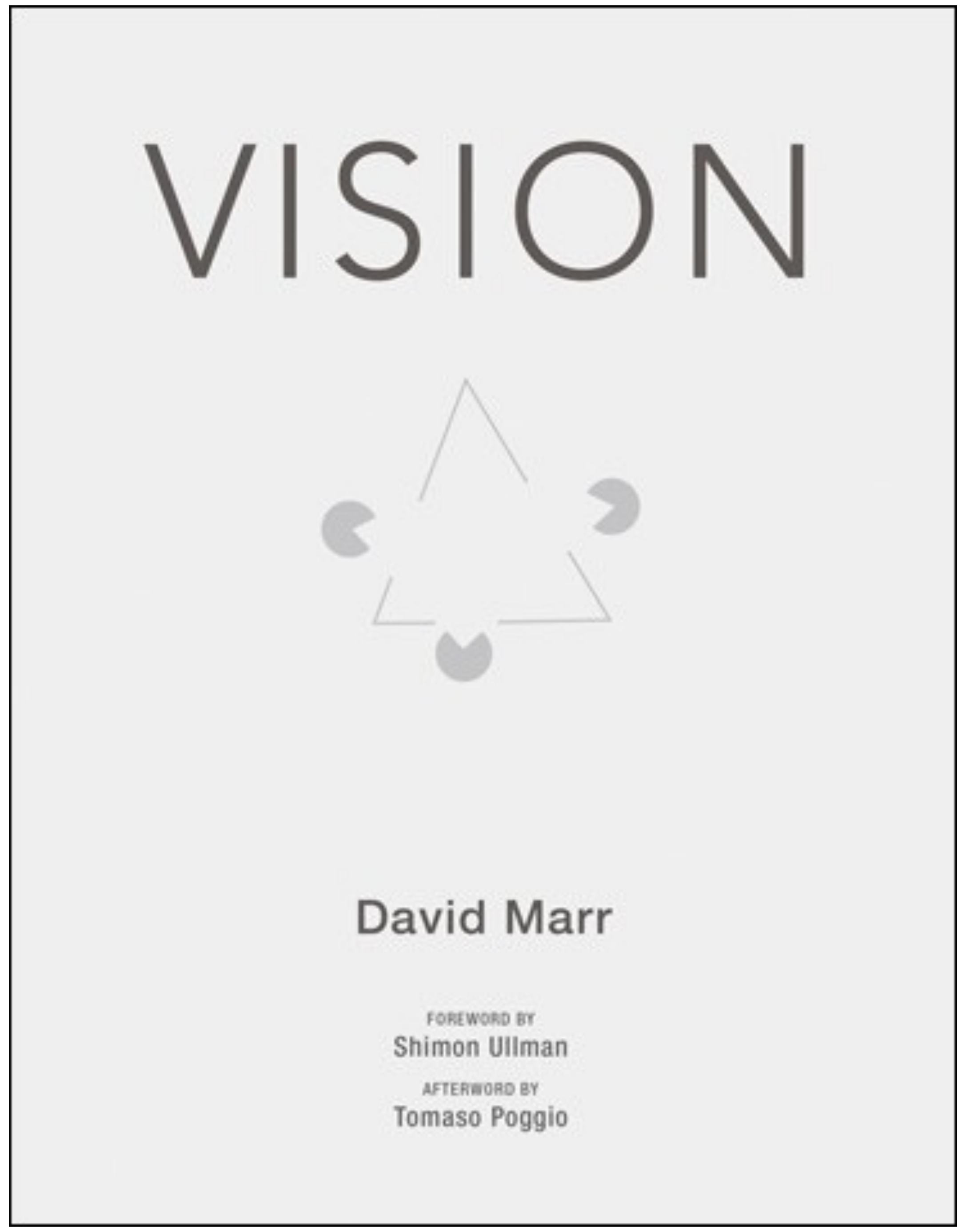
Lots of DSP and linear algebra await !

Marr's Tri-Level Hypothesis for Vision

David Marr integrated results from psychology, artificial intelligence, and neurophysiology into new models of visual processing, creating the field of Computer Vision.

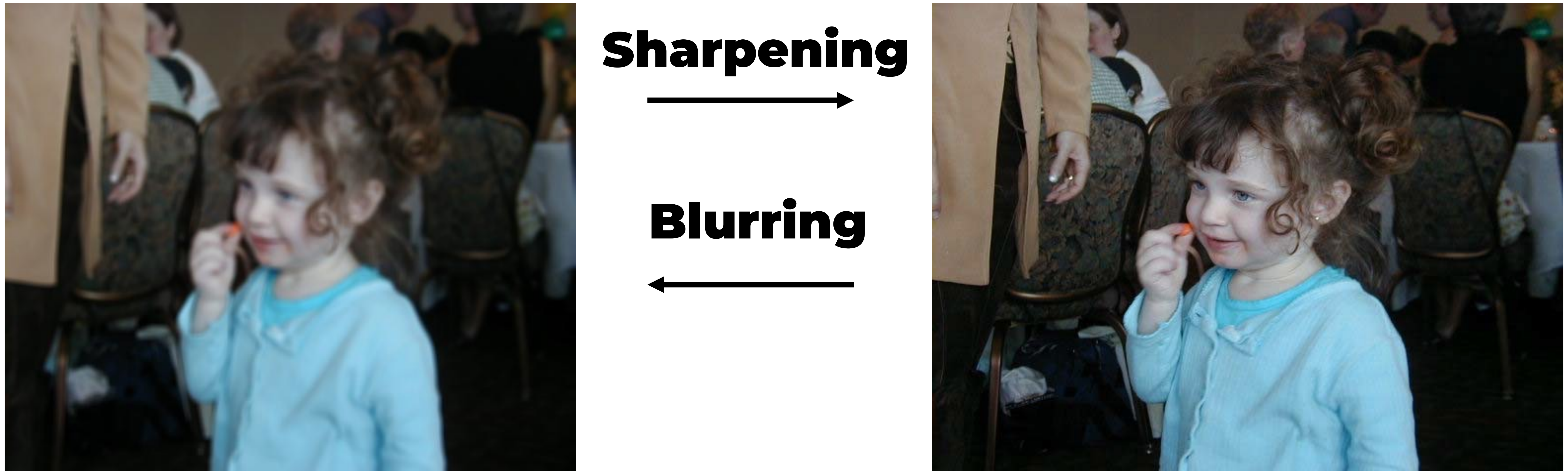
- **Computational level:** what does the system do (e.g.: what problems does it solve or overcome) and similarly, why does it do these things -- What is the problem?
- **Algorithmic level** (a.k.a. **representational level**): how does the system do what it does, specifically, what representations does it use and what processes does it employ to build and manipulate the representations -- How to solve the problem?
- **Implementational level** (a.k.a. **physics level**): how is the system physically realized (in the case of biological vision, what neural structures and neuronal activities implement the visual system) -- How the above are done in a computer or a brain?

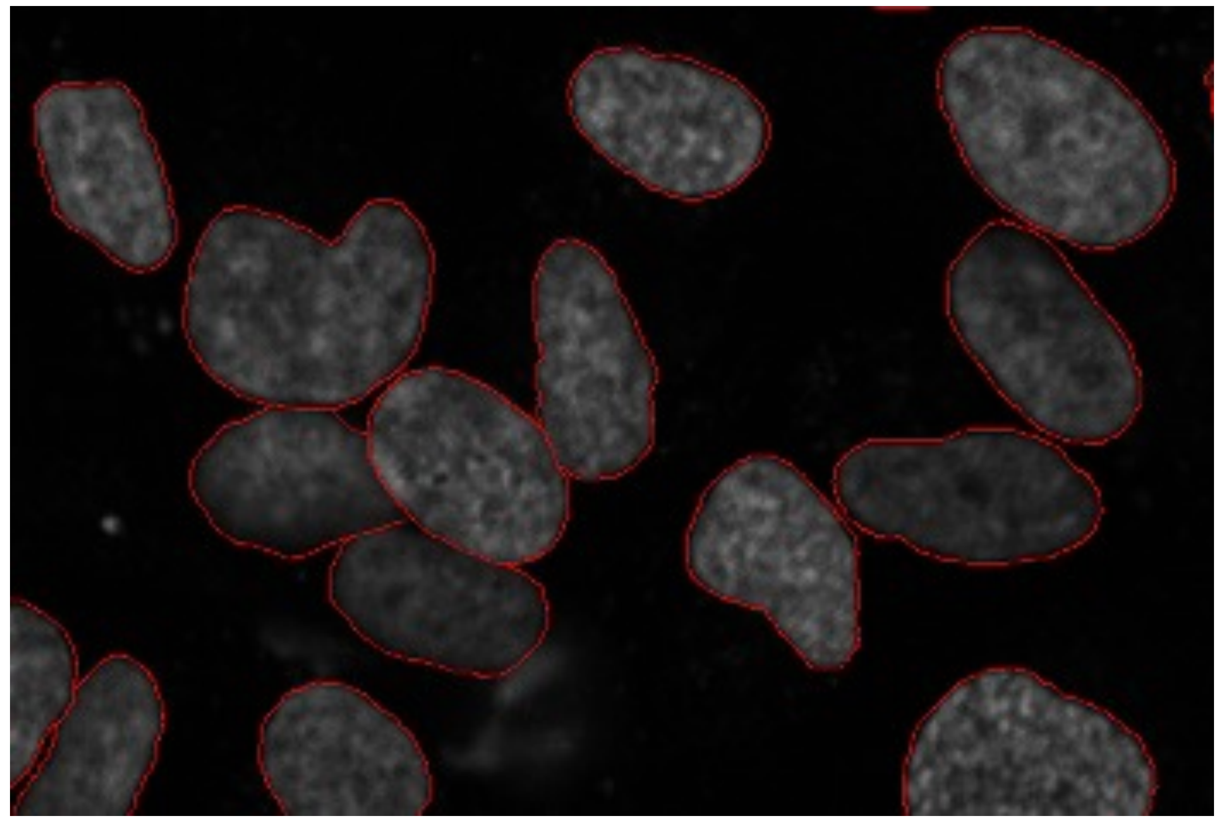




Three Stages in Computer Vision

- **Low-Level:** Image to image (enhancement, edge detection…)
  - Largely overlapped with signal or image “reconstruction” & “filtering”
  - Directly interface with image formulation, often considered as “pre-processing” for CV tasks



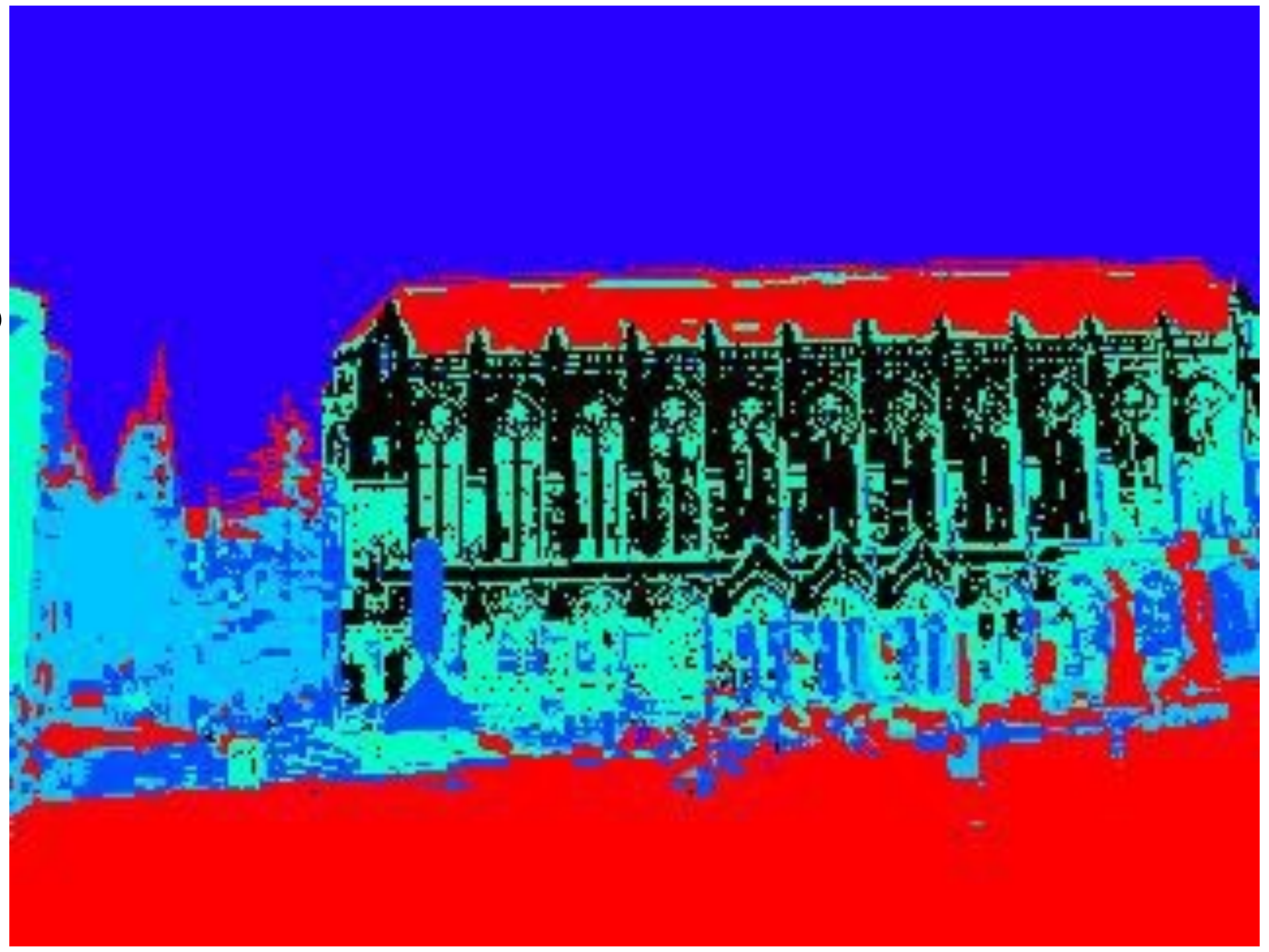


Three Stages in Computer Vision

- **Mid-Level:** Image to feature (classical segmentation, grouping…)
  - ***What's the criterion?*** Gestalt psychologists suggest an intermediate vision stage whose underlying processes are *grouping* mechanisms, which are essential for separating objects from background. Certain "*commonsense*" principles, such as closure, symmetry, or similarity guide how to group pieces of image and locate boundary.

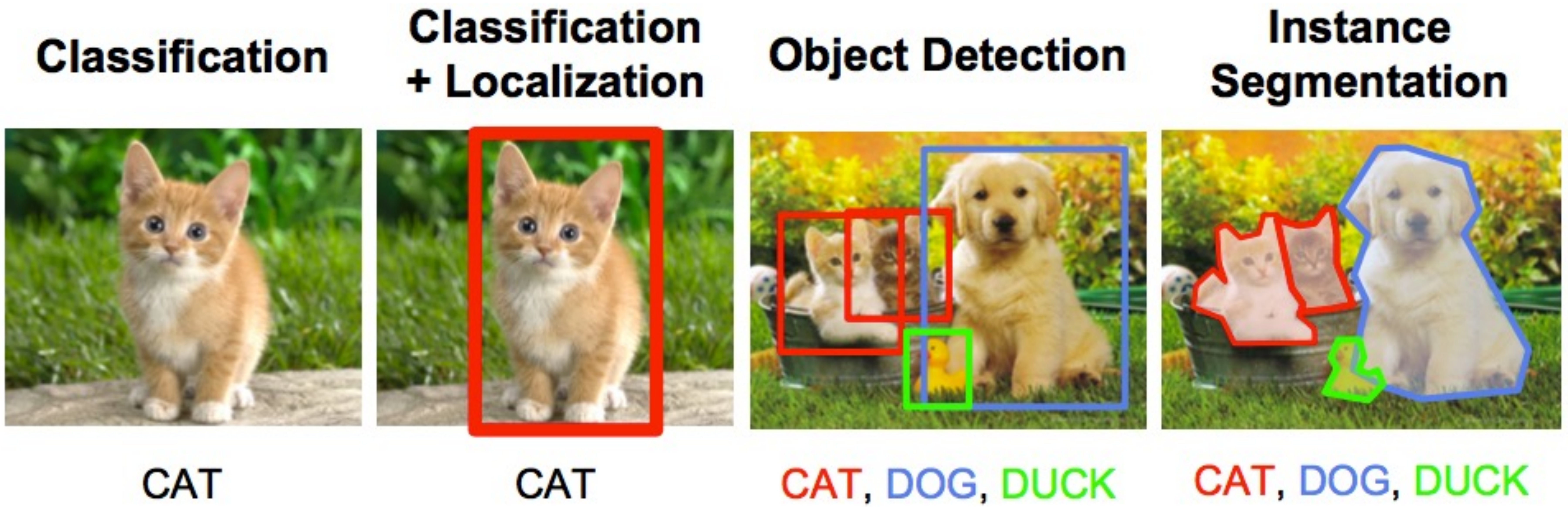
Clustering + connected component analysis

Object Structure



Three Stages in Computer Vision

- **High-Level:** Image to analysis (recognition, detection, semantic segmentation …)
  - Facilitating semantic interpretation of visual data, and required for numerous applications like robotics, driver assistance, multi-media retrieval, biometrics and surveillance …



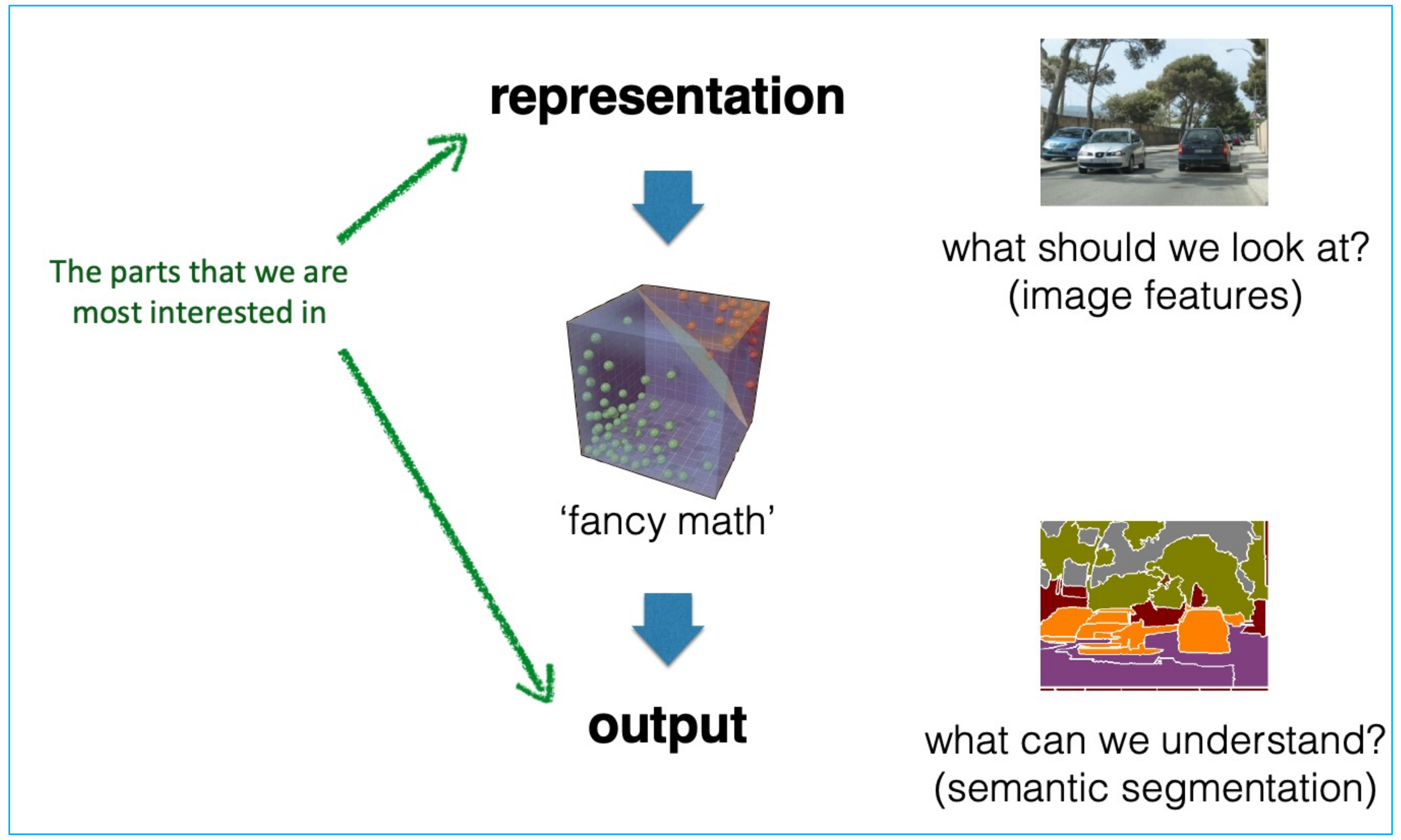
Three Levels: An Example

"There's an edge!"

"There's an object and a background!"

"There's a chair!"

Example: A Simple Computer Vision Pipeline (1990s)



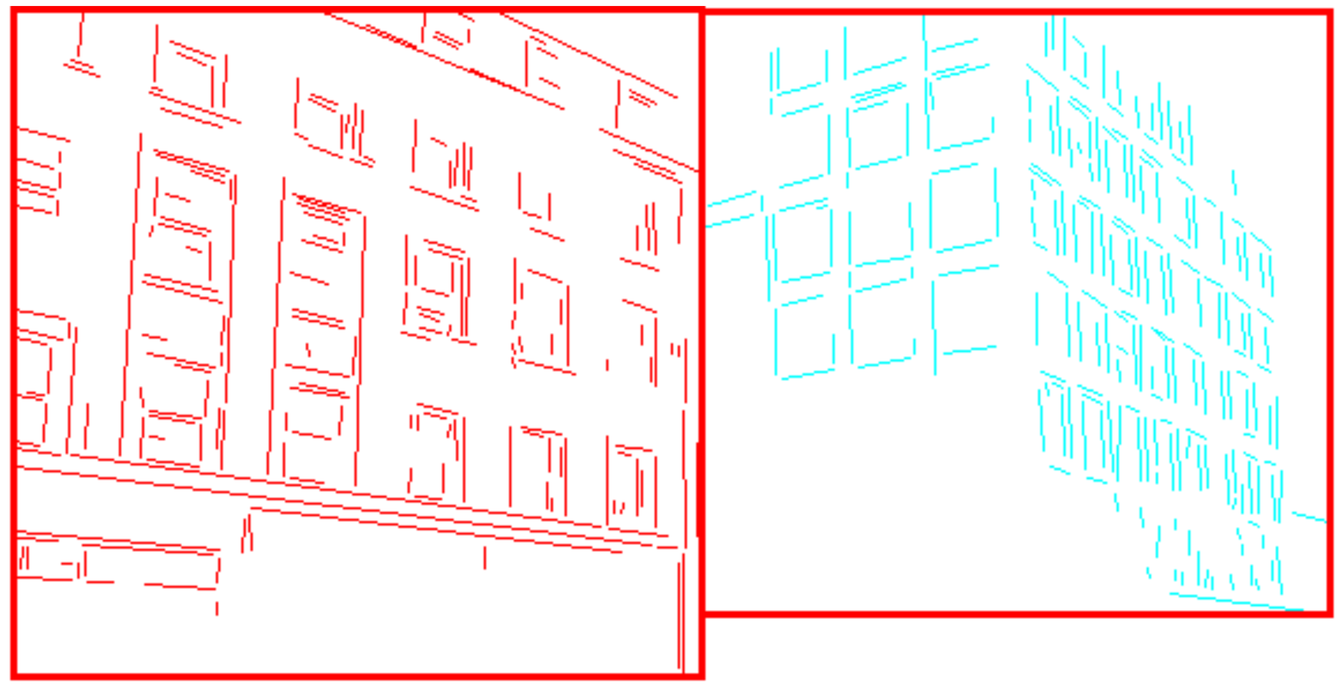
Low-Level

Edge Map

Mid-Level

Local Groups

High-Level



Building Recognition

The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering