A Quintillion Live Pixels:

The Challenge of Continuously Interpreting, Organizing, and Generating the World's Visual Information

Kayvon Fatahalian
Carnegie Mellon University

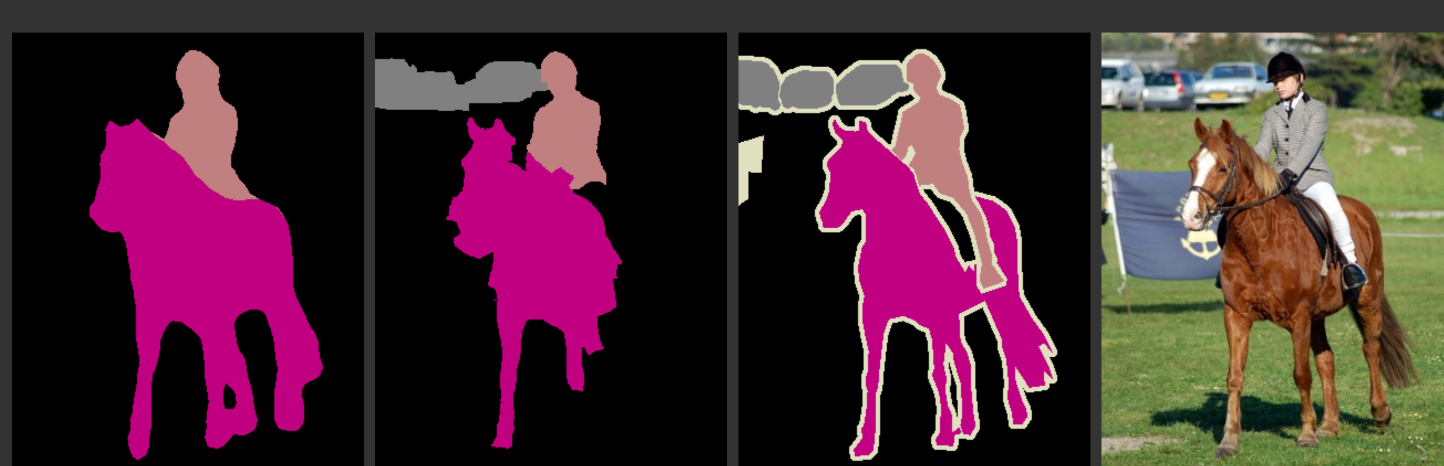
Thanks to Yaser Sheikh, Srinivas Narasimhan, and Ren Ng for various slide credits

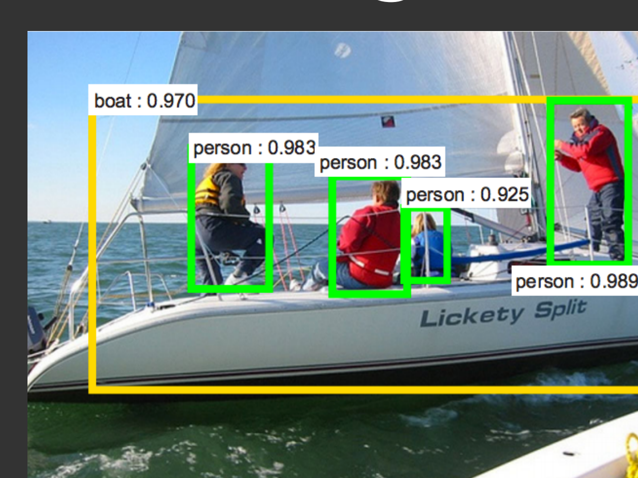
Visual computing

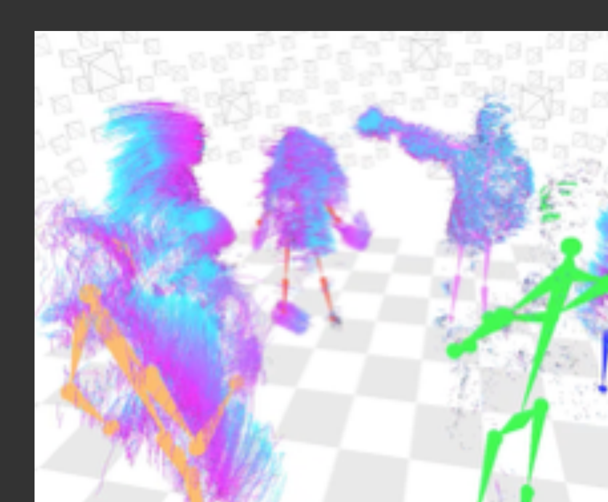
2D/3D graphics

Image processing / computational photography

Computer vision (visual scene understanding)



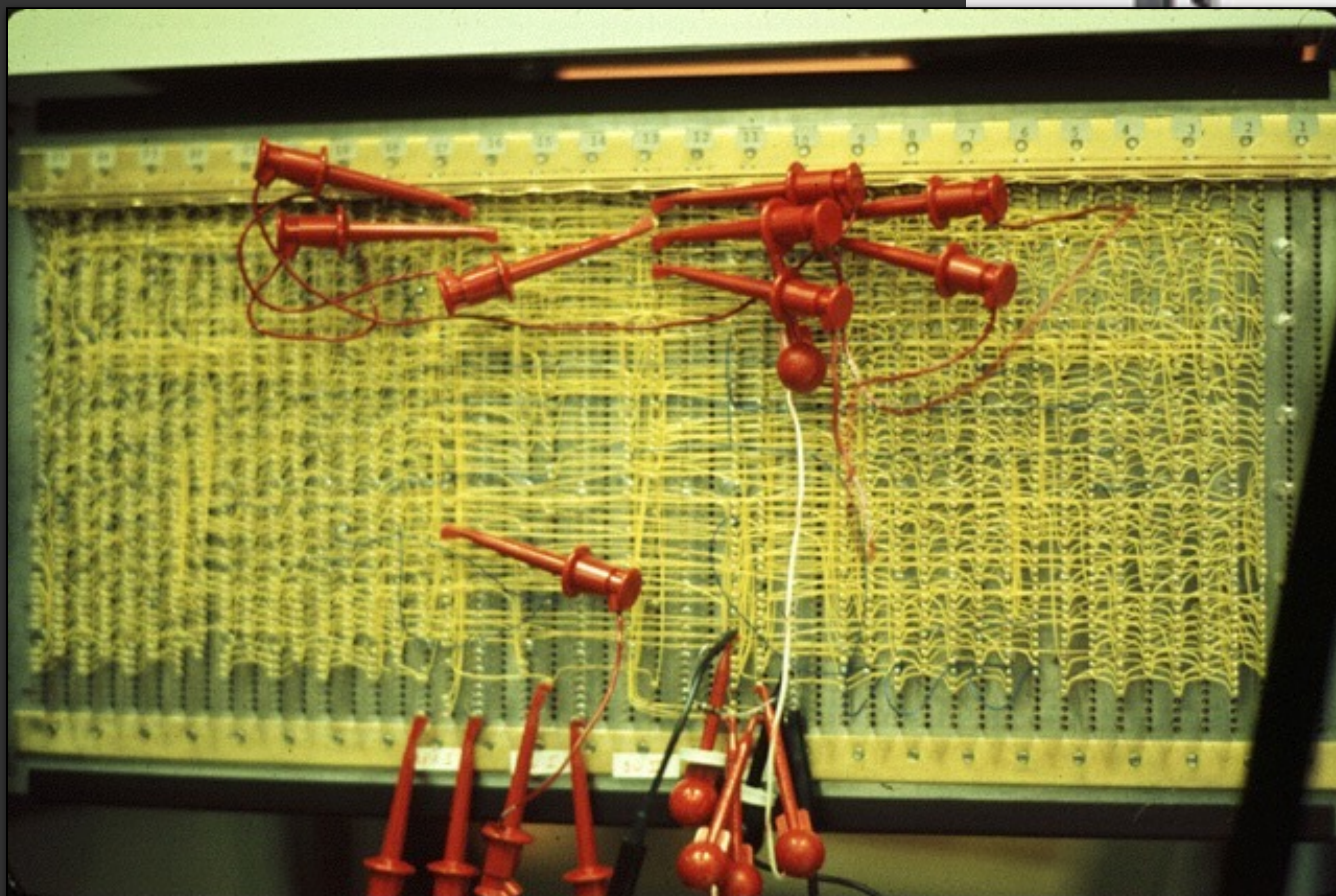




Ivan Sutherland's Sketchpad on MIT TX-2 (1962)

The frame buffer

Shoup's SuperPaint (PARC 1972-73)

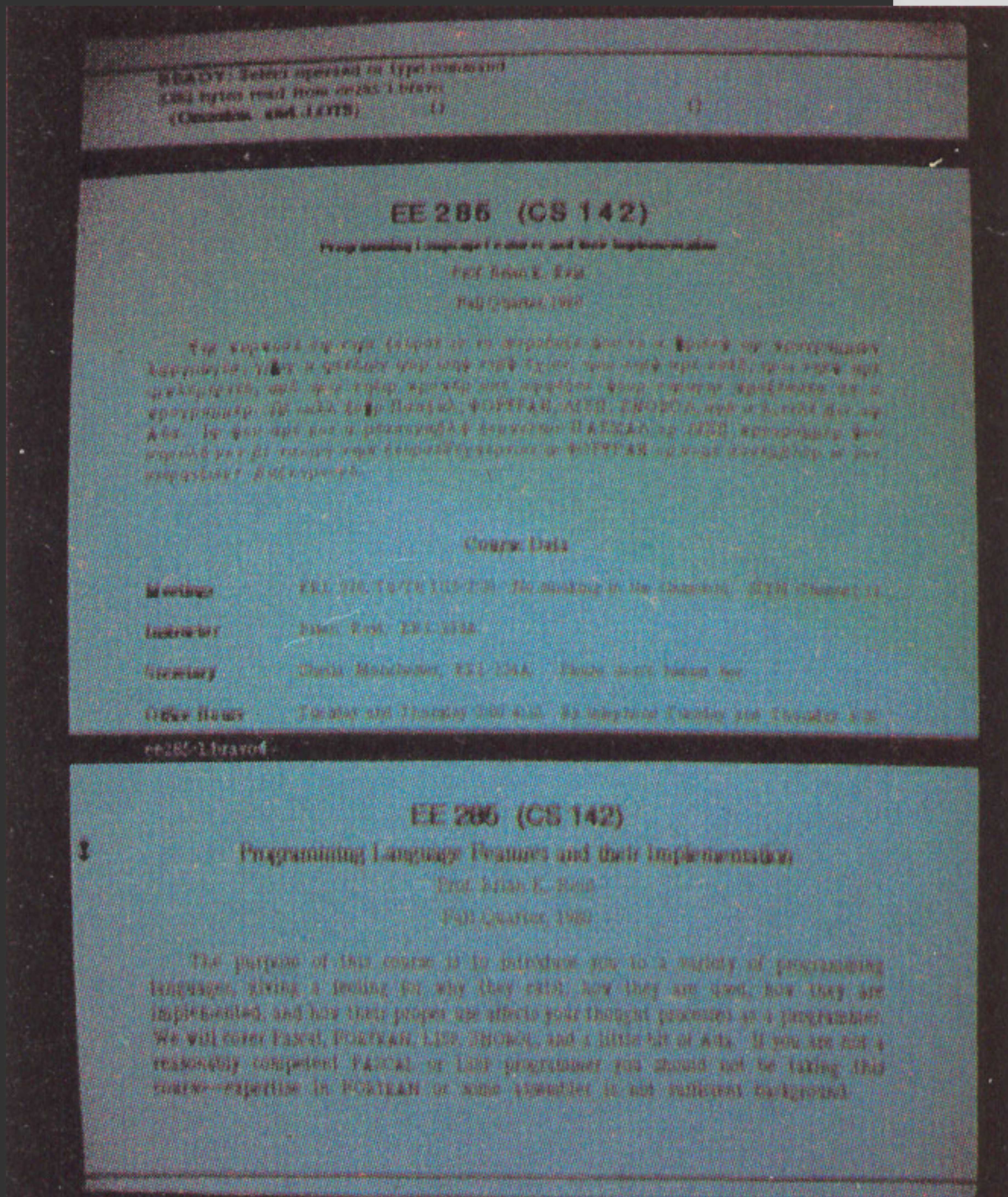


The frame buffer

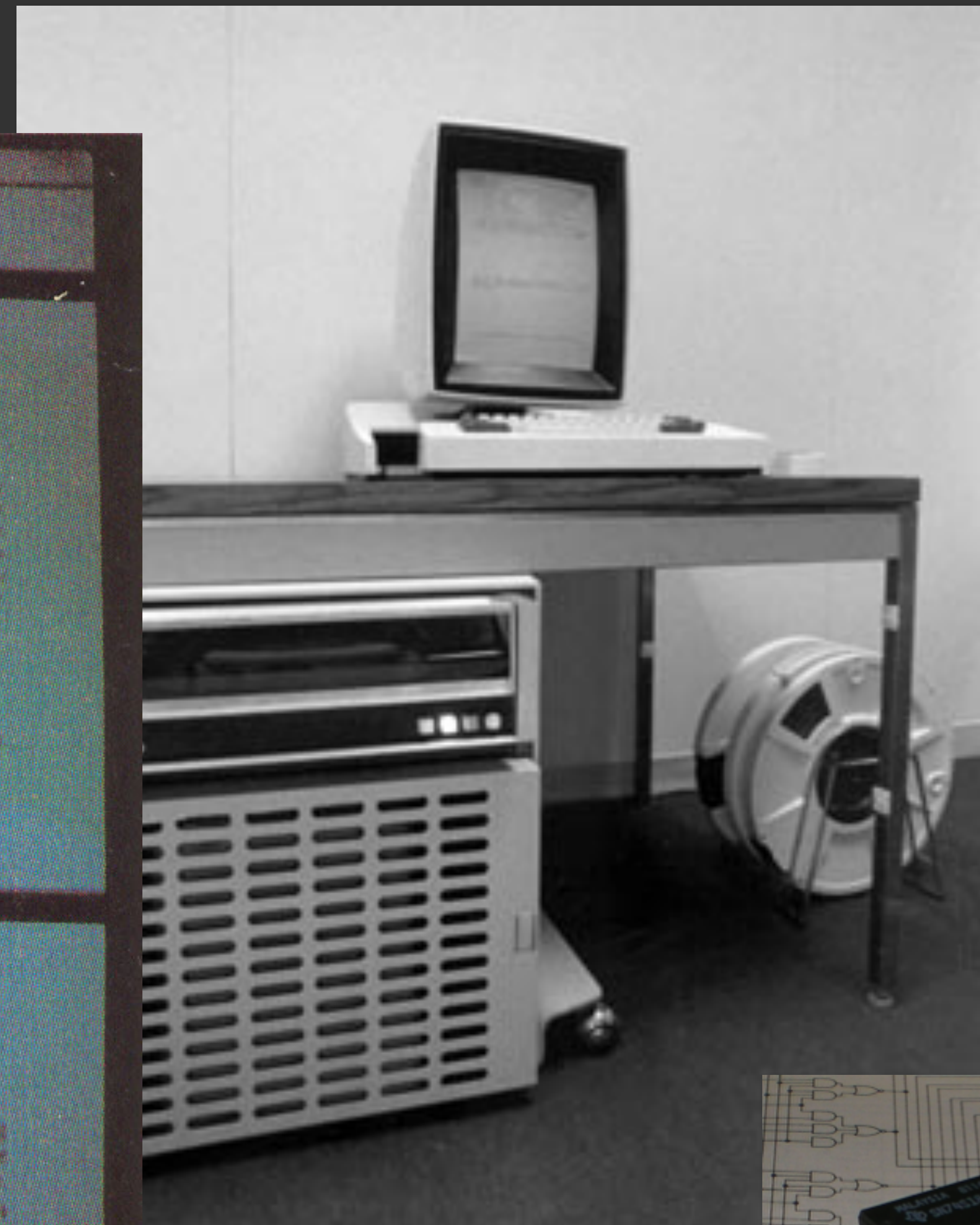
Shoup's SuperPaint (PARC 1972-73)

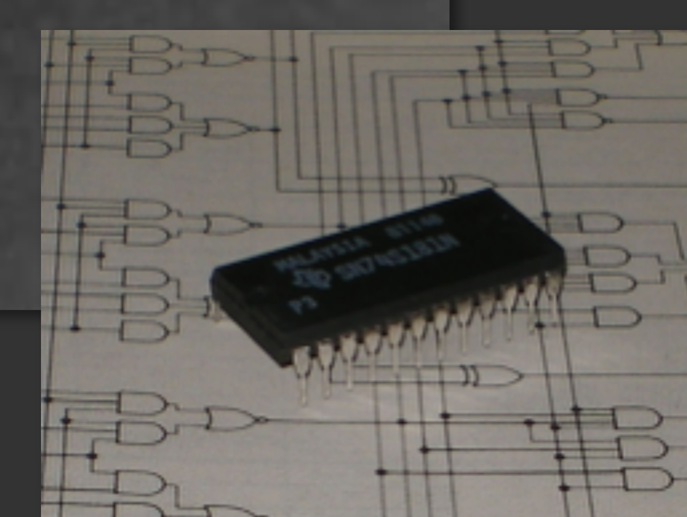
16 2K shift registers (640 x 486 x 8 bits)

Xerox Alto (1973)



Bravo (WYSIWYG)





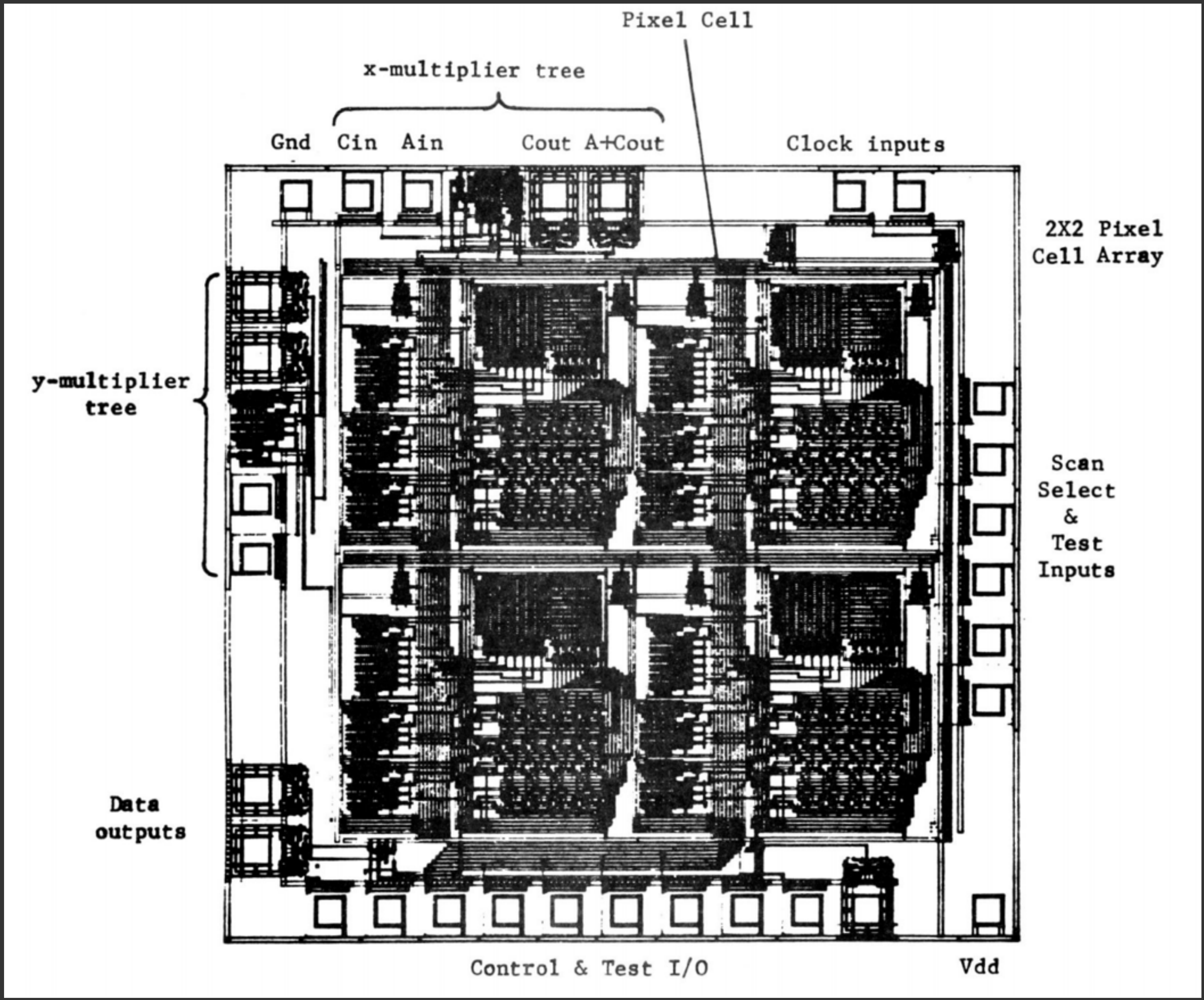
TI 74181 ALU

Goal: render everything you've ever seen

"Road to Pt. Reyes"
LucasFilm (1983)

Pixar's Toy Story (1995)

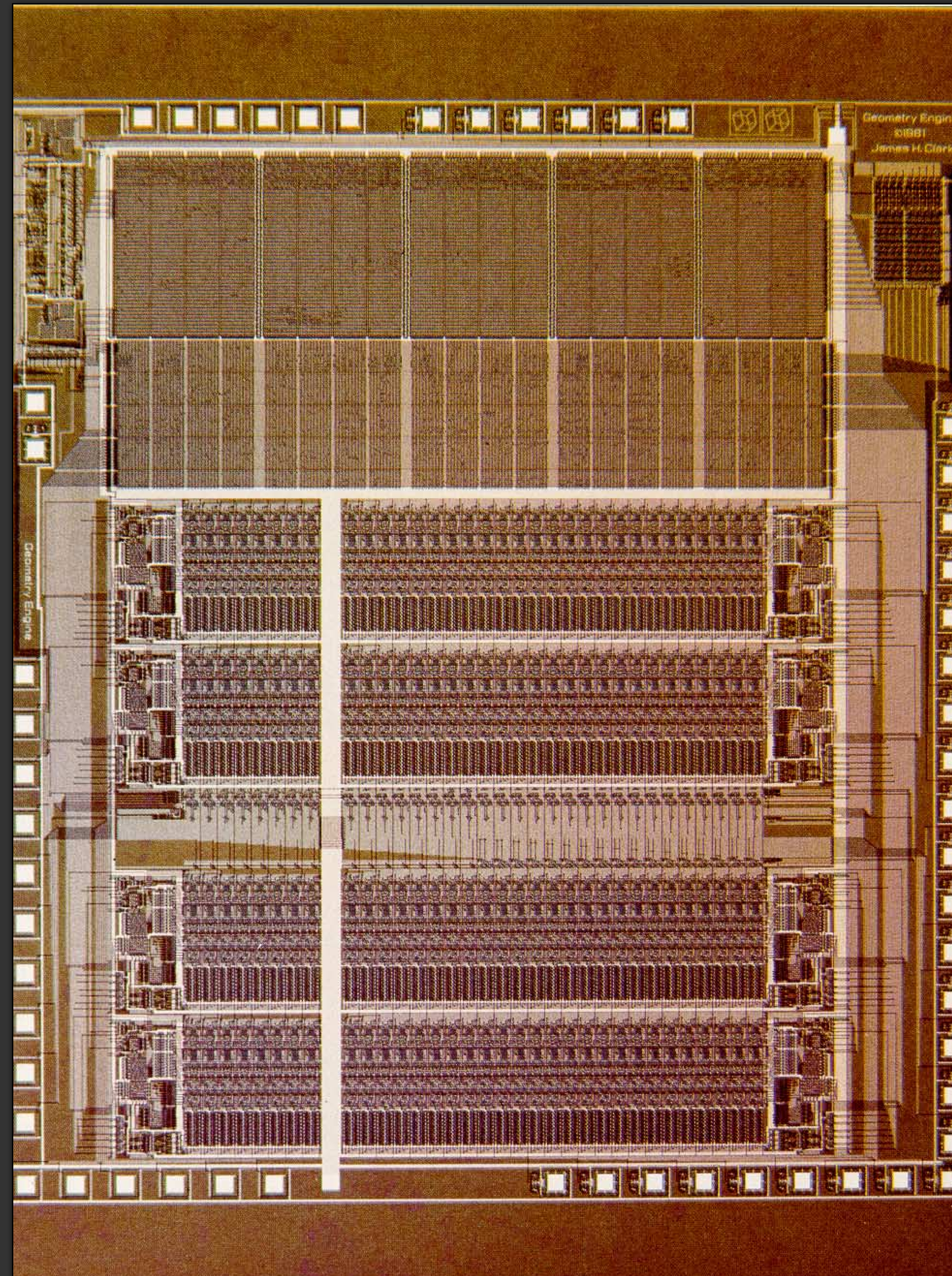
"We take an average of three hours to draw a single frame on the fastest computer money can buy."
- Steve Jobs

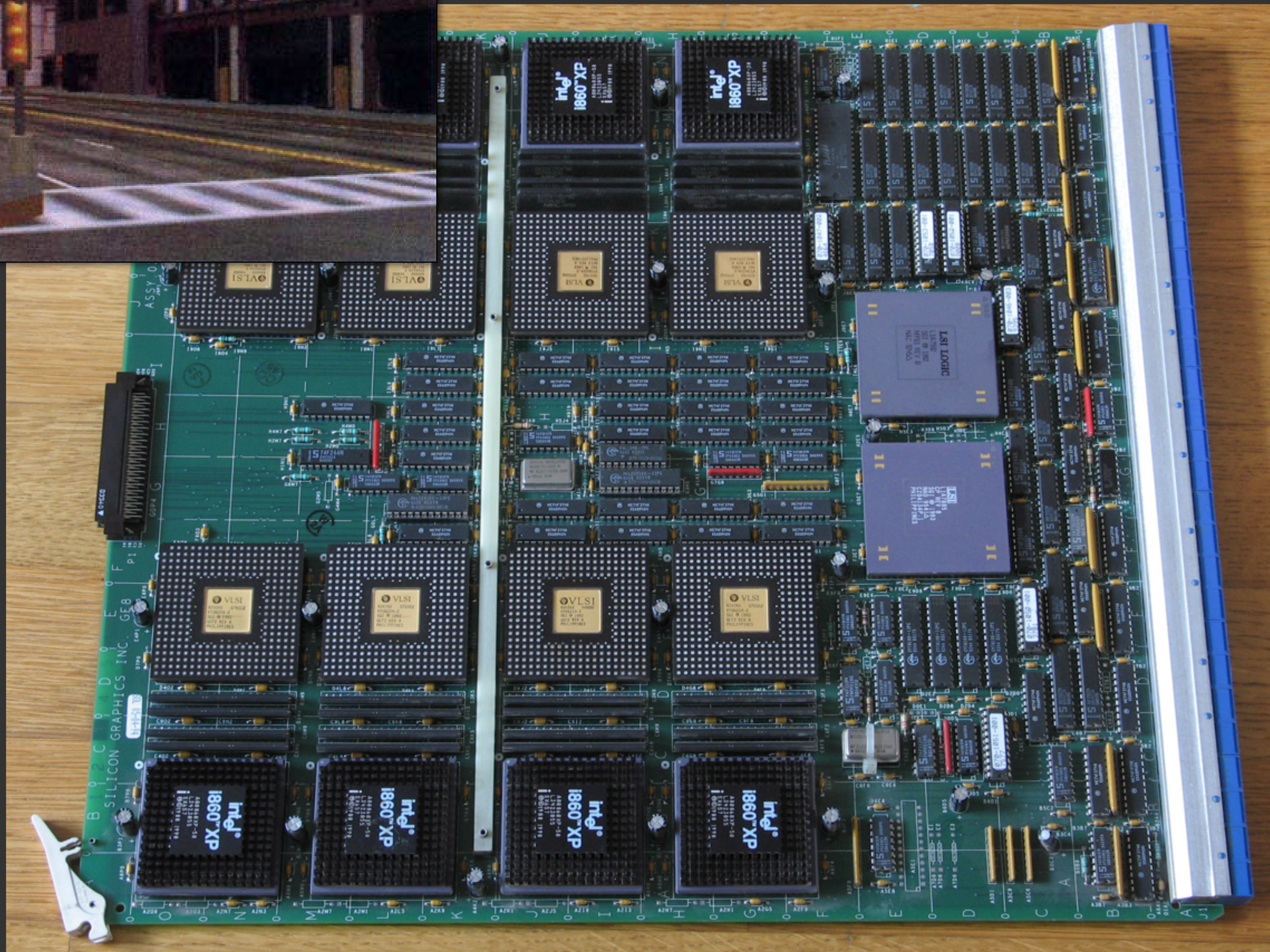


UNC Pixel Planes (1981), computation-enhanced frame buffer

Ed Clark's Geometry Engine (1982)

ASIC for geometric transforms used in real-time graphics.

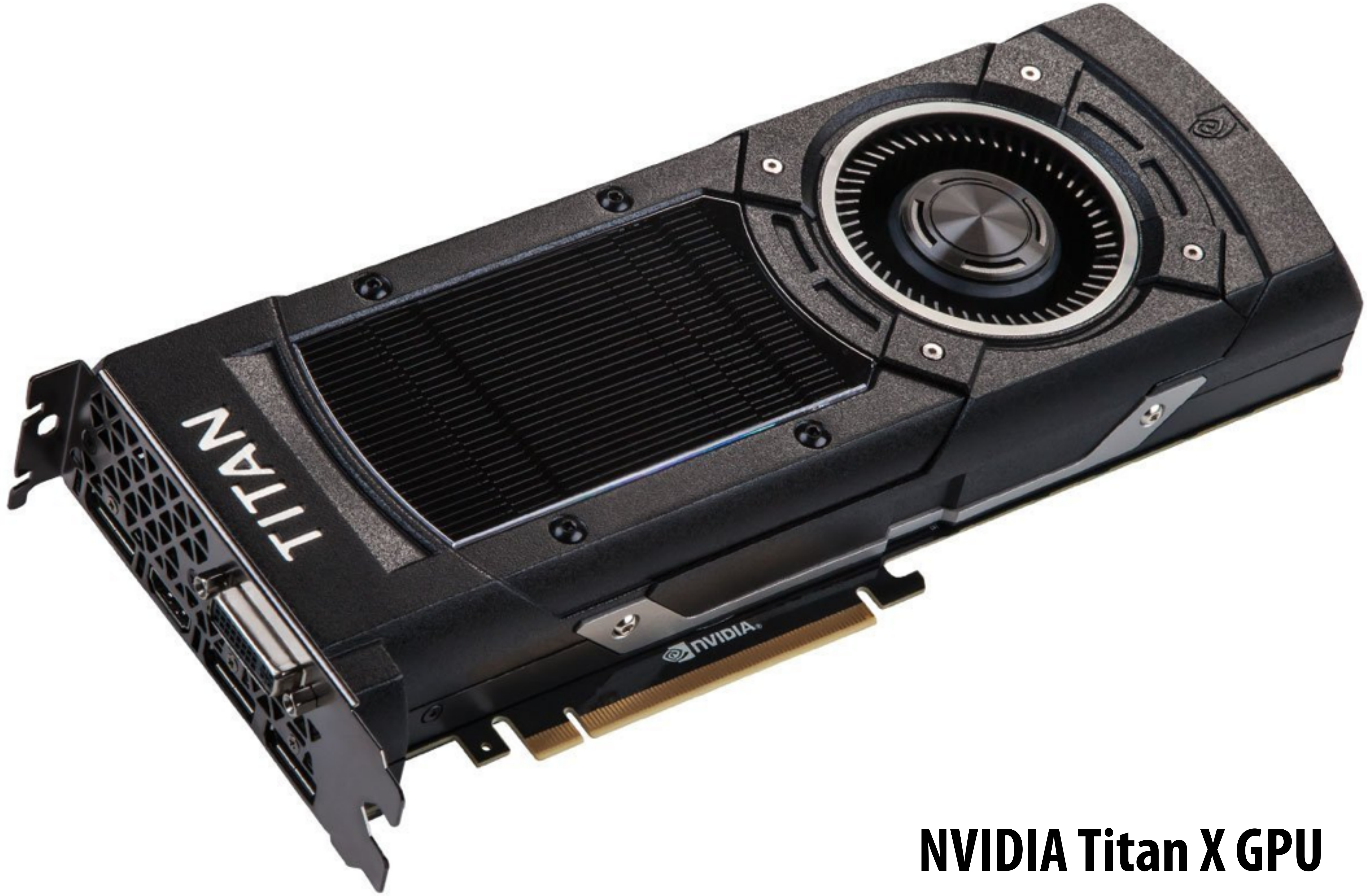




SGI RealityEngine GE8 board (1993)

Real-time (30 fps) on a NVIDIA Titan X

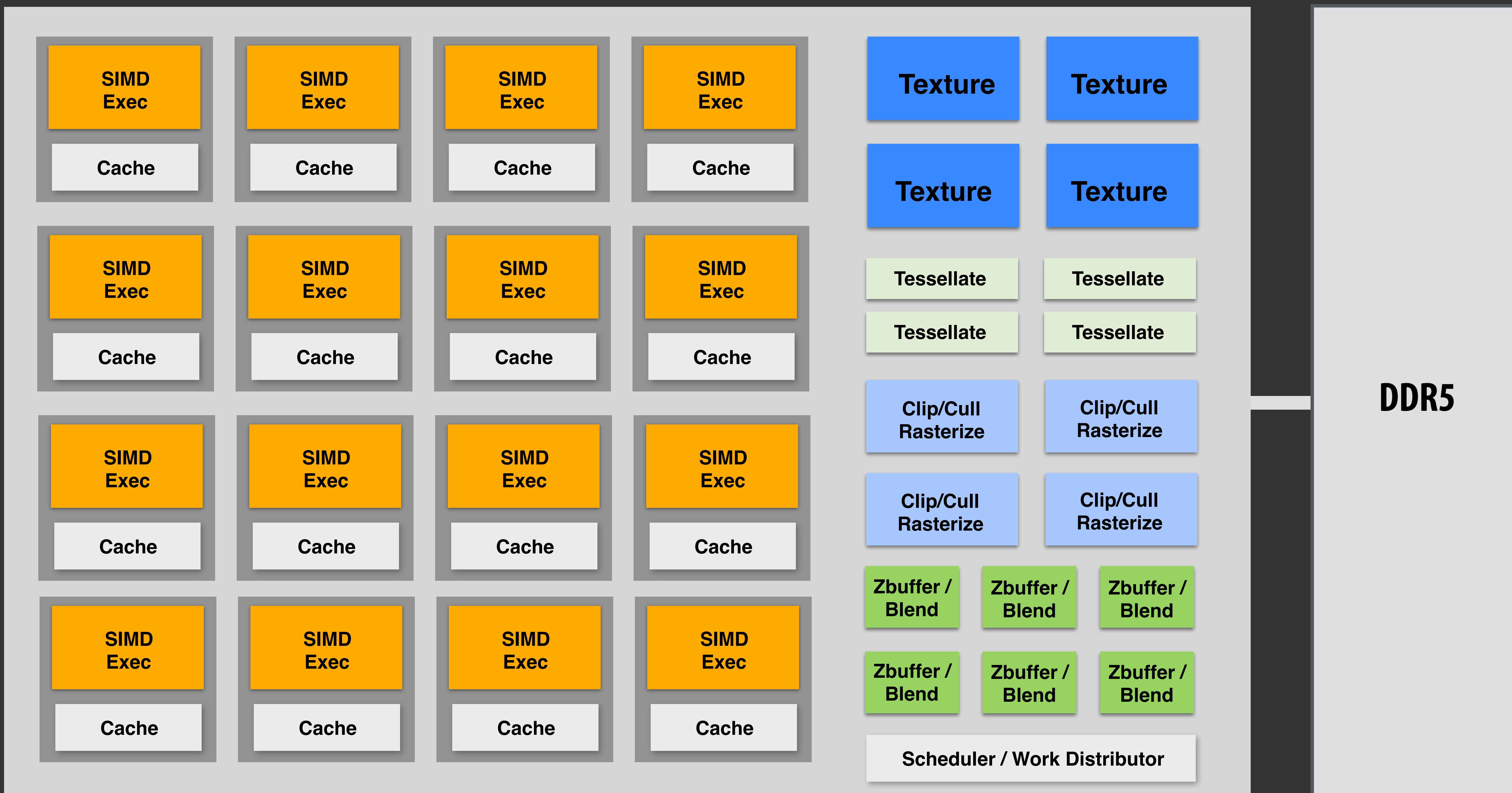
Unreal Engine Kite Demo (Epic Games 2015)



NVIDIA Titan X GPU
(~ 7 TFLOPs fp32)

Tesla generation NV chip ~ ASCI Red

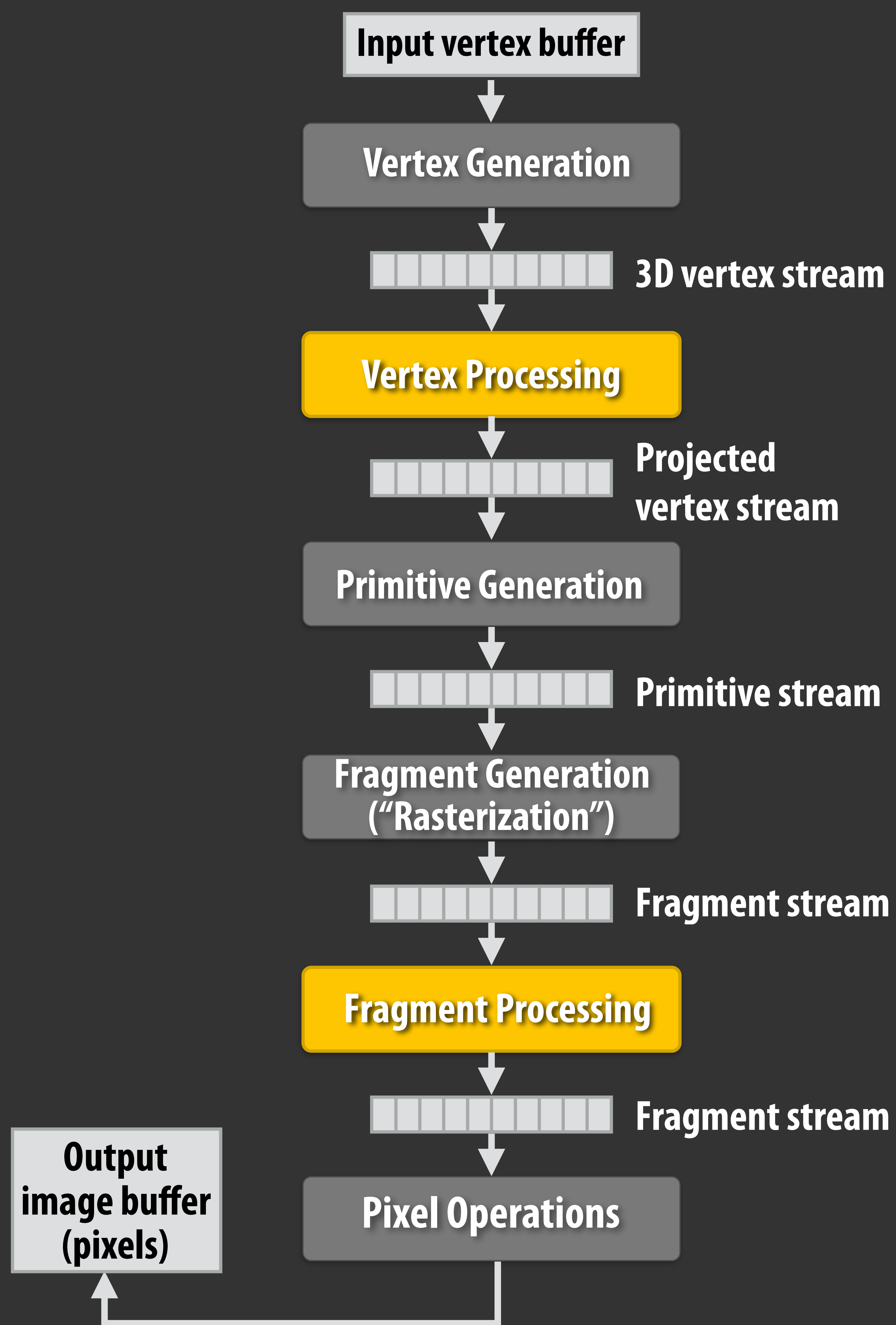
Modern GPU: heterogeneous multi-core

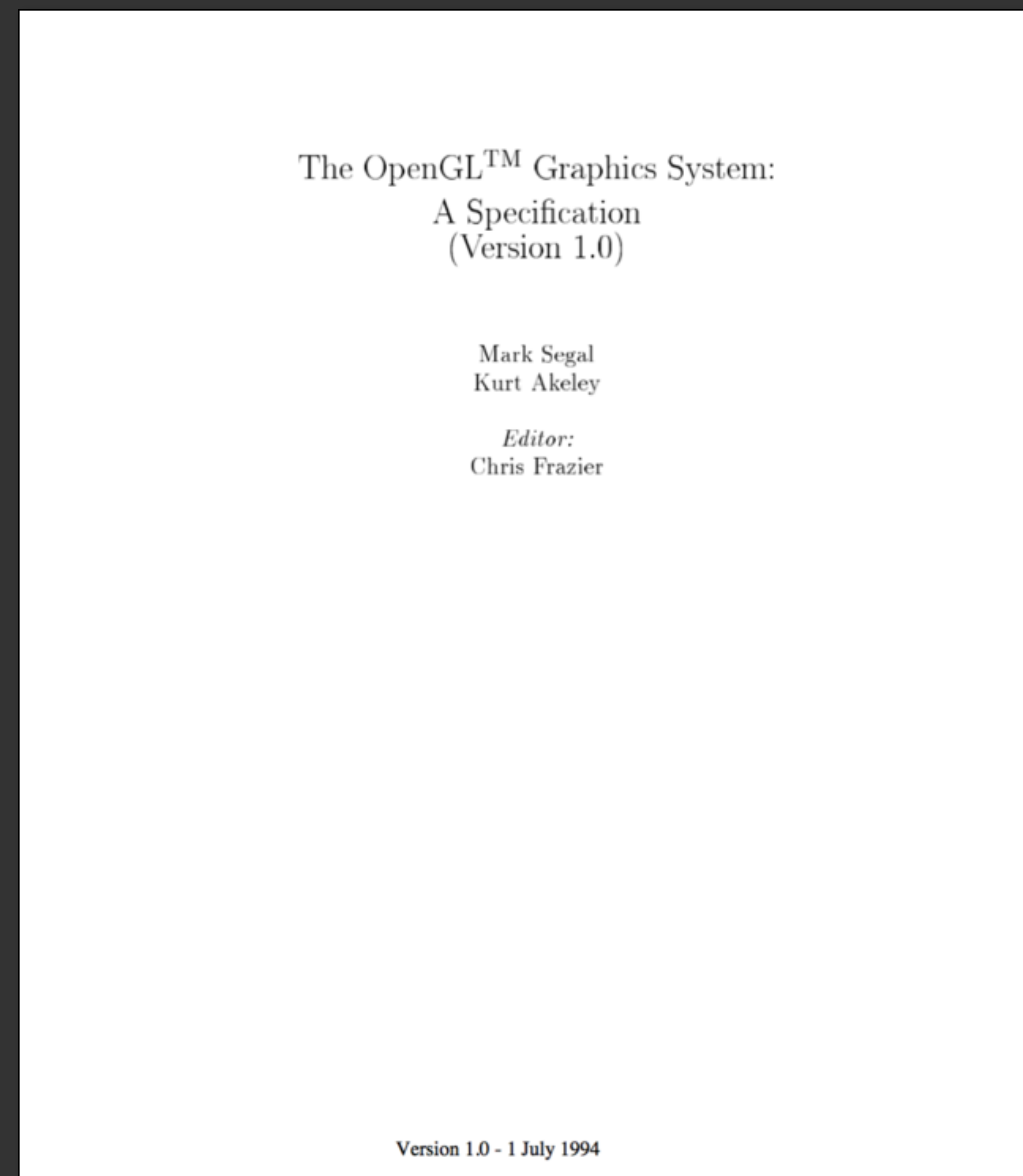


Multi-threaded, SIMD cores
Custom circuits for key graphics arithmetic
Custom circuits for HW-assisted graphics-specific DRAM compression
HW logic for scheduling work onto these resources

Domain-specific languages for heterogeneous computing

OpenGL Graphics Pipeline (circa 2007)





The OpenGL™ Graphics System:
A Specification
(Version 1.0)

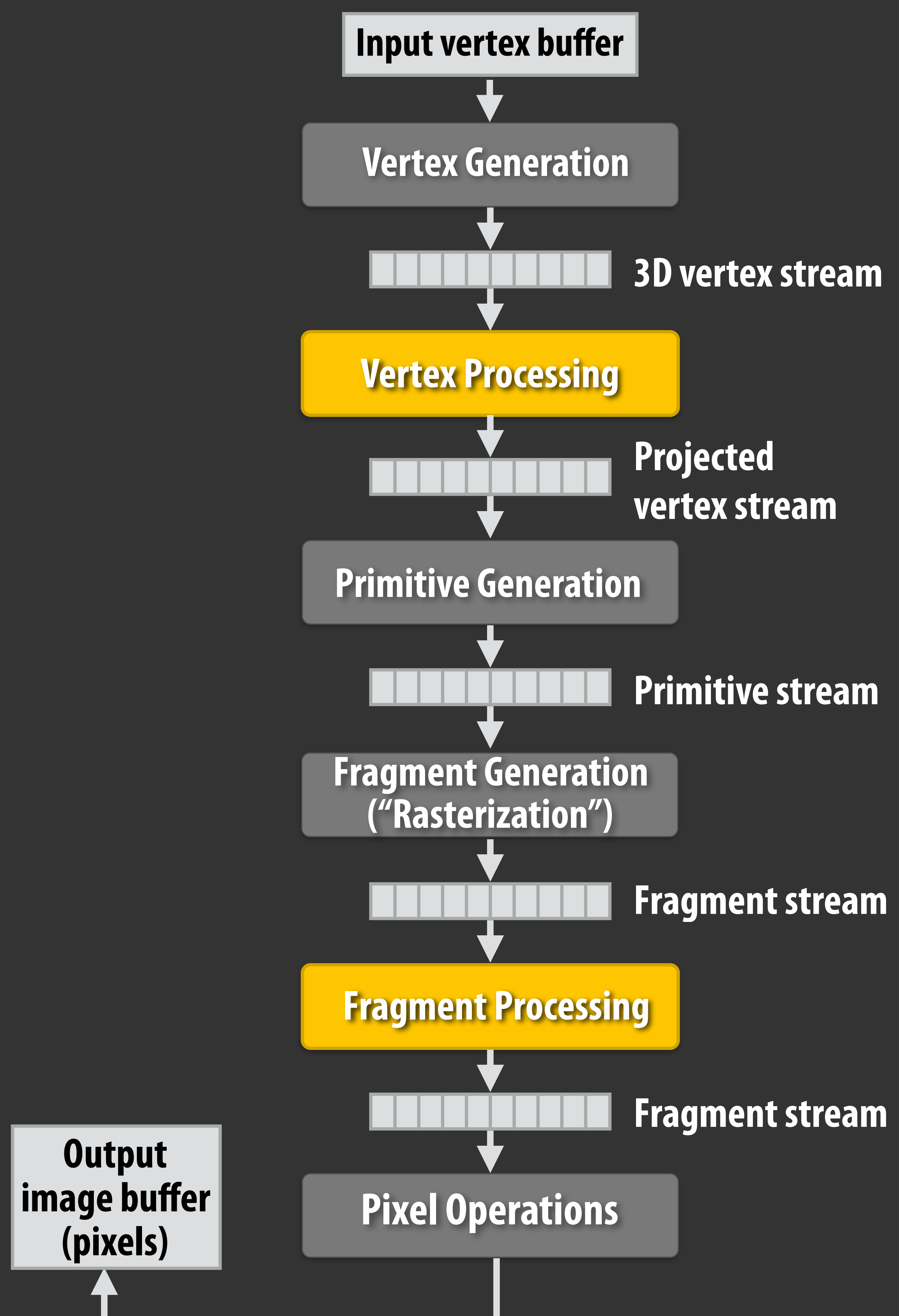
Mark Segal
Kurt Akeley

Editor:
Chris Frazier

Version 1.0 - 1 July 1994

Domain-specific languages for heterogeneous computing

OpenGL Graphics Pipeline (circa 2007)




```
uniform sampler2D myTexture;
uniform float3 lightDir;
varying vec3 norm;
varying vec2 uv;
```

read-only global variables (uniform lines)

"per-element" inputs (varying lines)

```
void myFragmentShader()
{
  vec3 kd = texture2D(myTexture, uv);
  kd *= clamp(dot(lightDir, norm), 0.0, 1.0);
  return vec4(kd, 1.0);
}
```

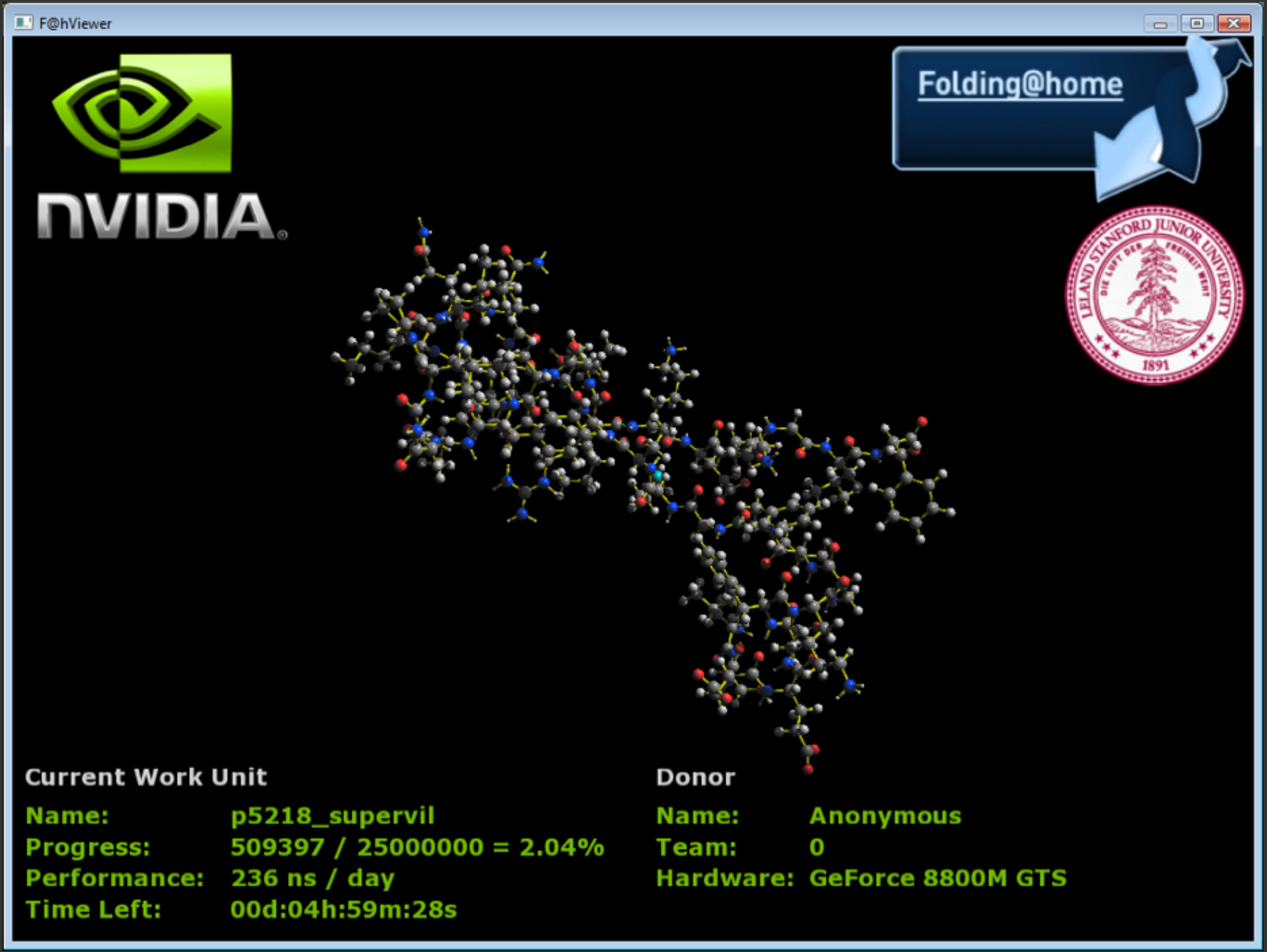
per-element output: RGBA surface color at pixel

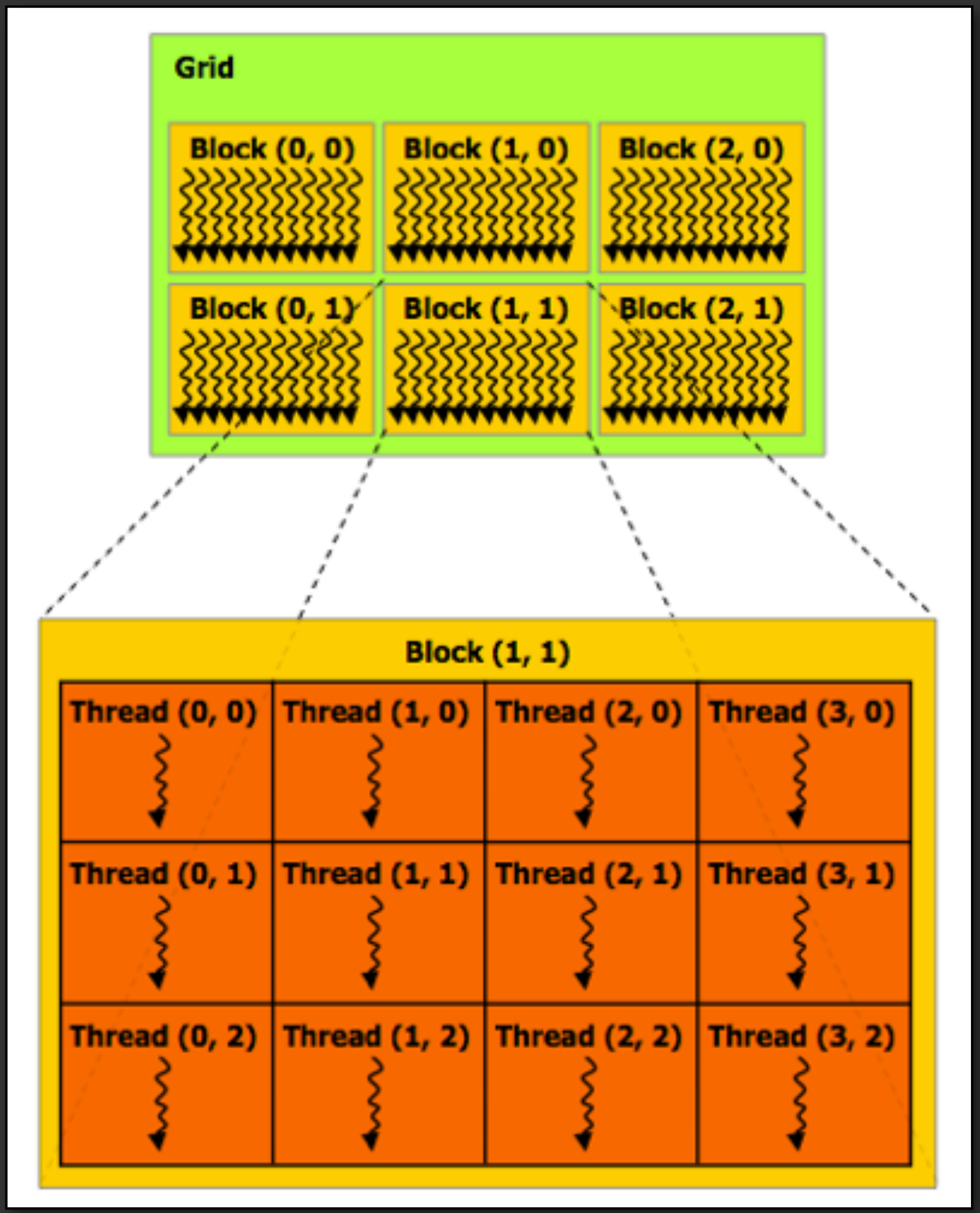
"fragment shader" (a.k.a kernel function mapped onto input fragment stream)

Generalization beyond graphics: commodity parallel computing

Brook for GPUs (Buck 2004)

NVIDIA CUDA (2007)





Goals of visual computing (to date)

Modeling the real-world in increasingly rich detail: so we can simulate it ("render everything you've ever seen")

Depict and organize information to augment human thought: enable humans to effectively use computing to create/analyze/interpret/communicate

Key characteristics of visual computing

Requires exceptional levels of efficiency

- Applications turn more ops/watt into new value
- Pack chips full of ALUs (parallel, heterogeneity/specialization are fundamental)
- Applications utilize hardware pipelines very well

Embrace domain-specific programming frameworks

- Achieve high efficiency/productivity
- Today: OpenGL, Halide, game engine frameworks, deep learning frameworks

Aspects of computation are fundamentally approximate

- Manifests as willingness to change algorithms (not approximate HW)

Visual computing — what’s next?

Goals of visual computing (present — future)

To capture everything that can be seen

To enable humans to communicate more effectively

To record and analyze the world's visual information so that computers can understand and reason about it

The immediate future: capturing rich visual information to enhance communication

Capturing pixels to communicate

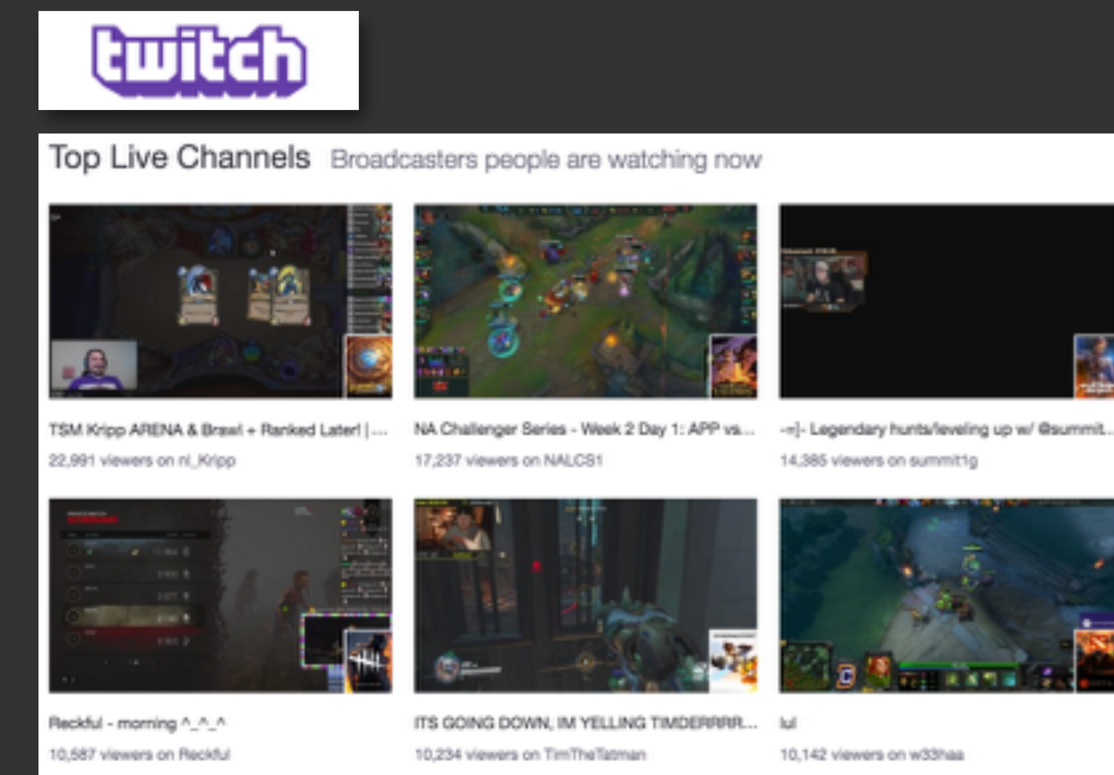
Ingesting/serving
the world's photos

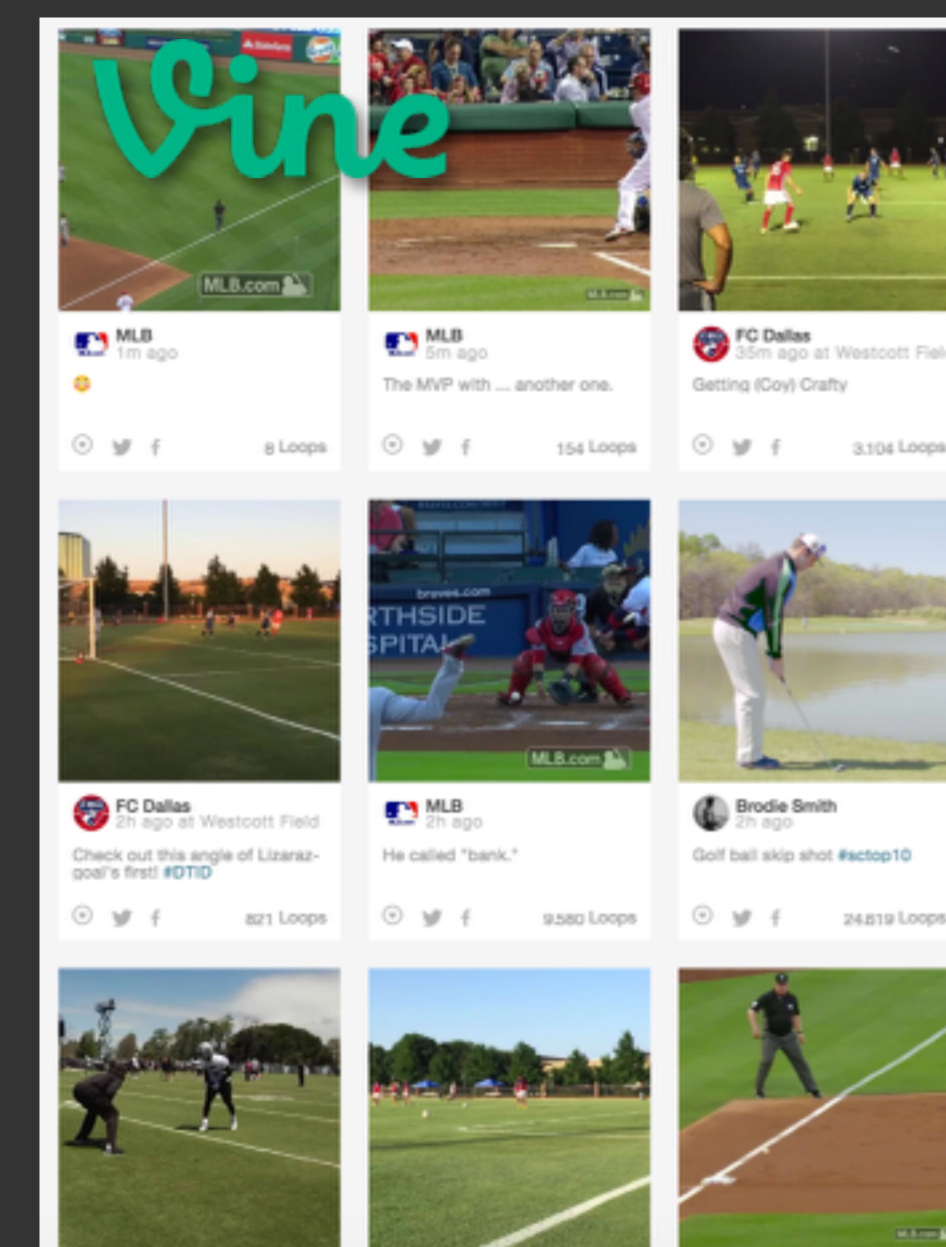
2B photo uploads and shares per day across Facebook sites (incl. Instagram+WhatsApp) [FB2015]

Ingesting/streaming
world's video

Youtube 2015: 300 hours uploaded per minute [Youtube]

Cisco VNI projection:
80-90% of 2019 internet traffic will be video.
(64% in 2014)





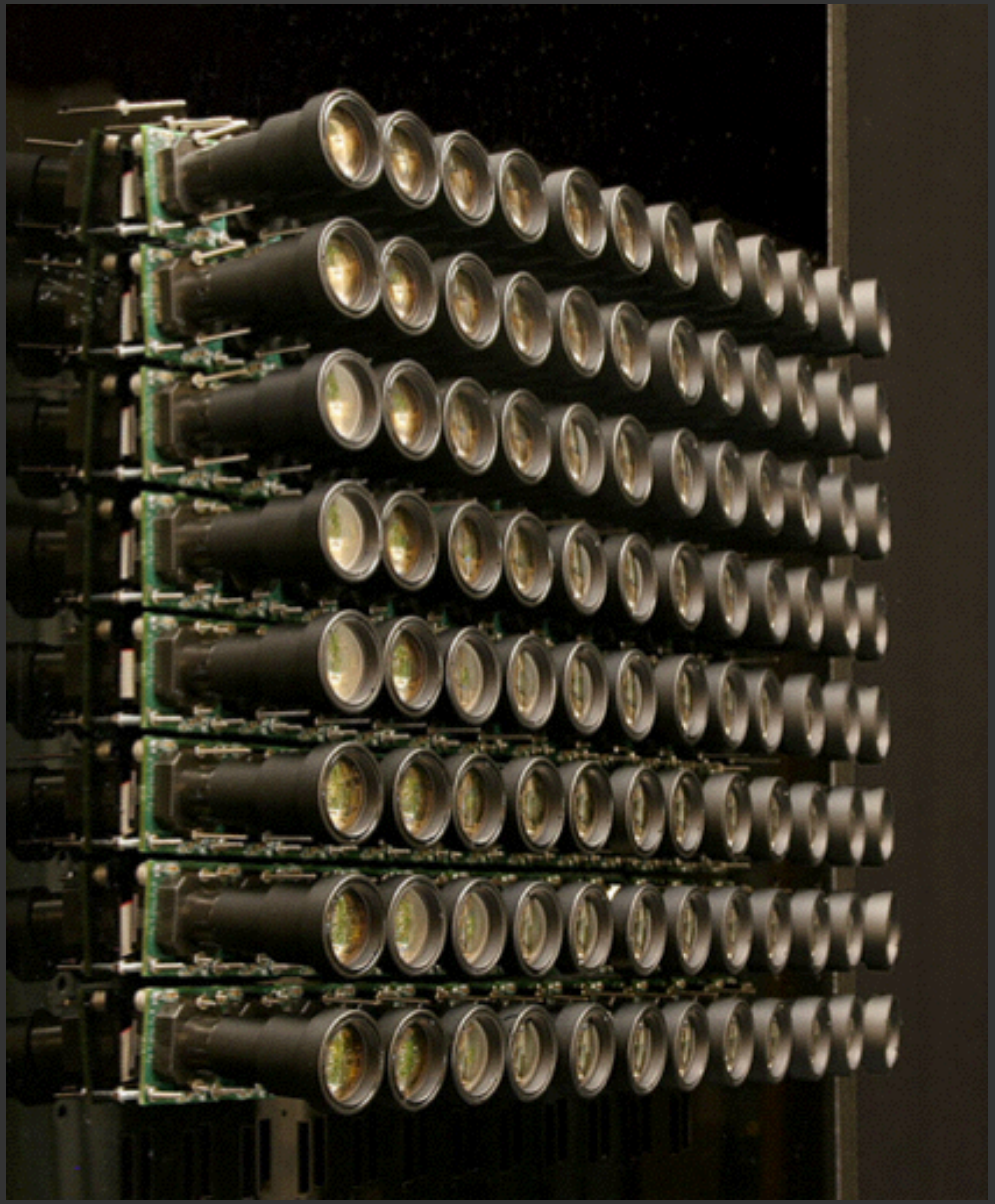
Richer content: beyond a single image

- Example: Apple's "Live Photos"
- Each photo is not only a single frame, but a few seconds of video before and after the shutter is clicked

Facebook Live

Acquiring richer content: light fields

Stanford camera array
Wilburn [2005]

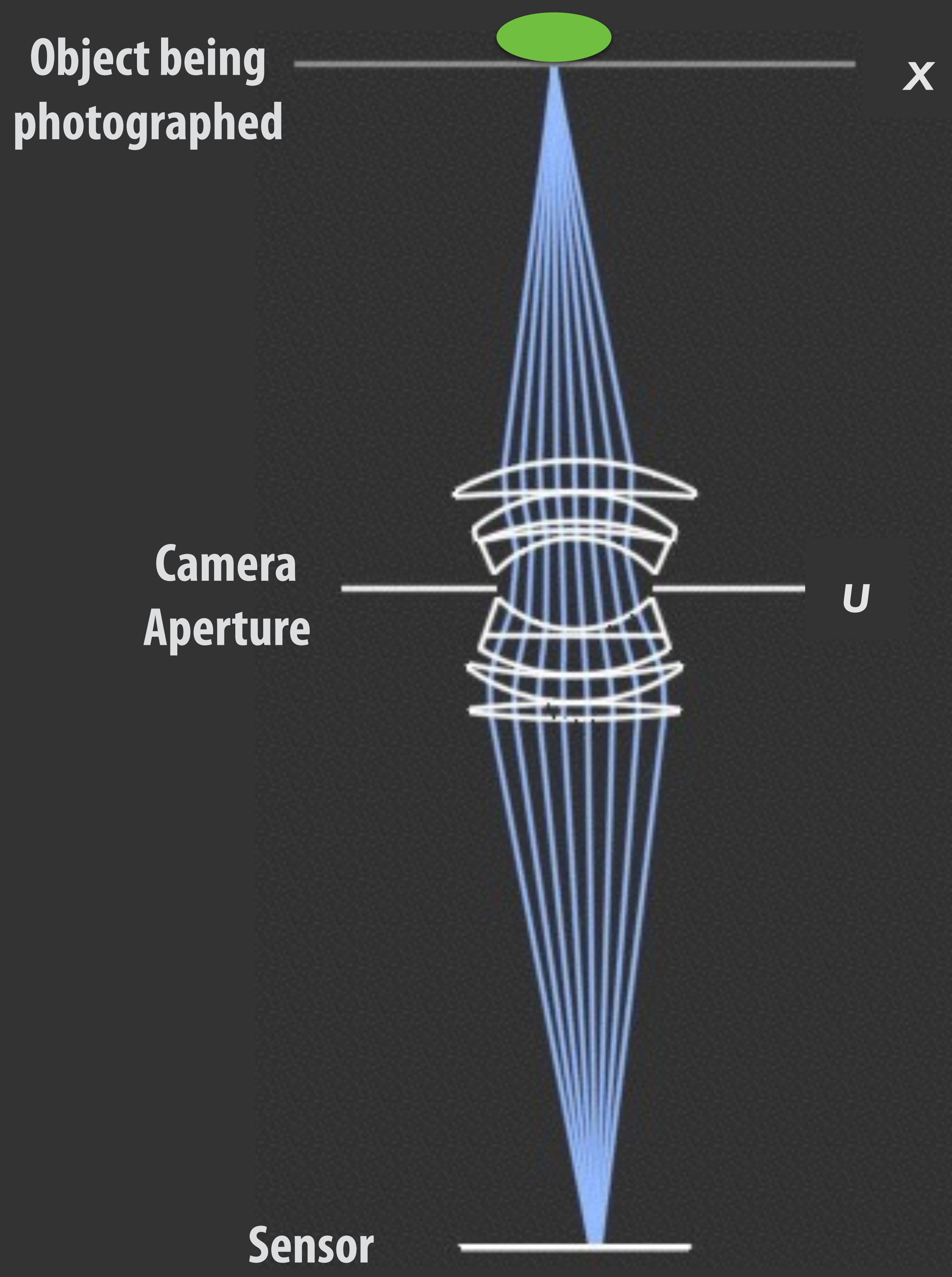


Richer content: light fields

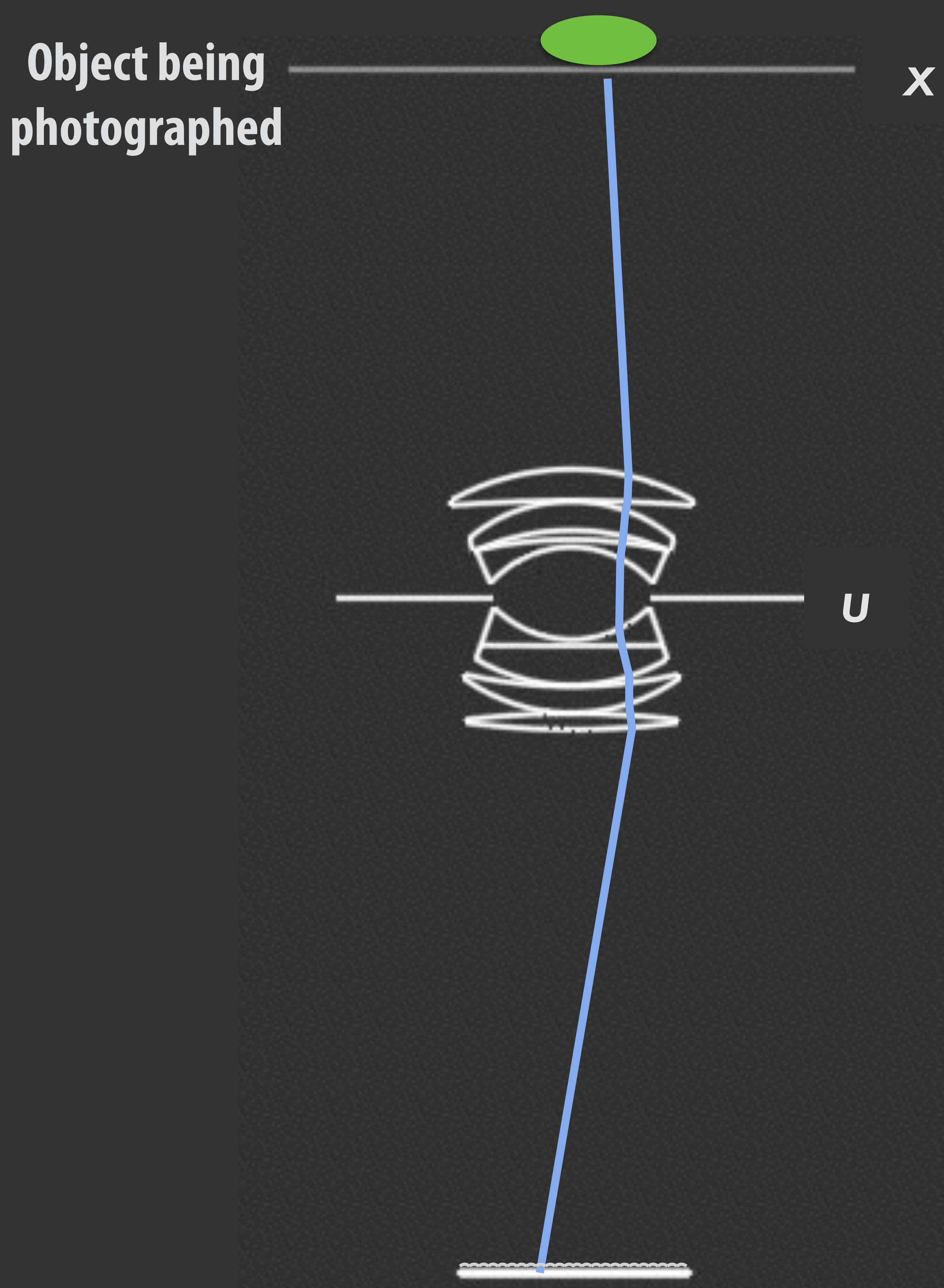
Light L16

Lytro Illum

Light field camera: capturing a light field



2D traditional camera:
measures how much light hits a
point on sensor



"4D" light field camera:
measures how much light hits point
on sensor **from a particular direction**

[Slide courtesy Ren Ng]

[Slide courtesy Ren Ng]

[Slide courtesy Ren Ng]

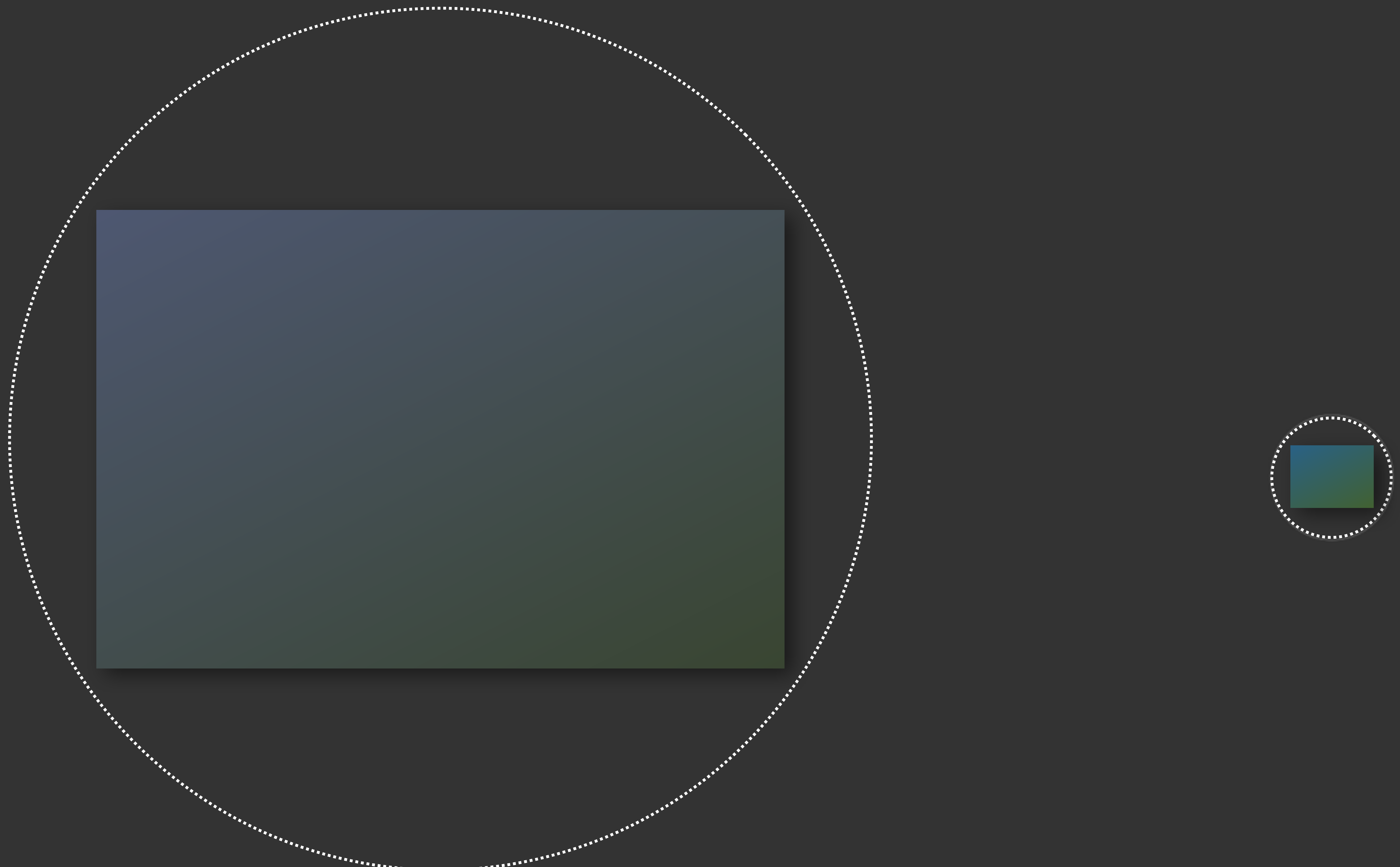
[Slide courtesy Ren Ng]

[Slide courtesy Ren Ng]

[Slide courtesy Ren Ng]

[Slide courtesy Ren Ng]

Sensor industry has large untapped resolution



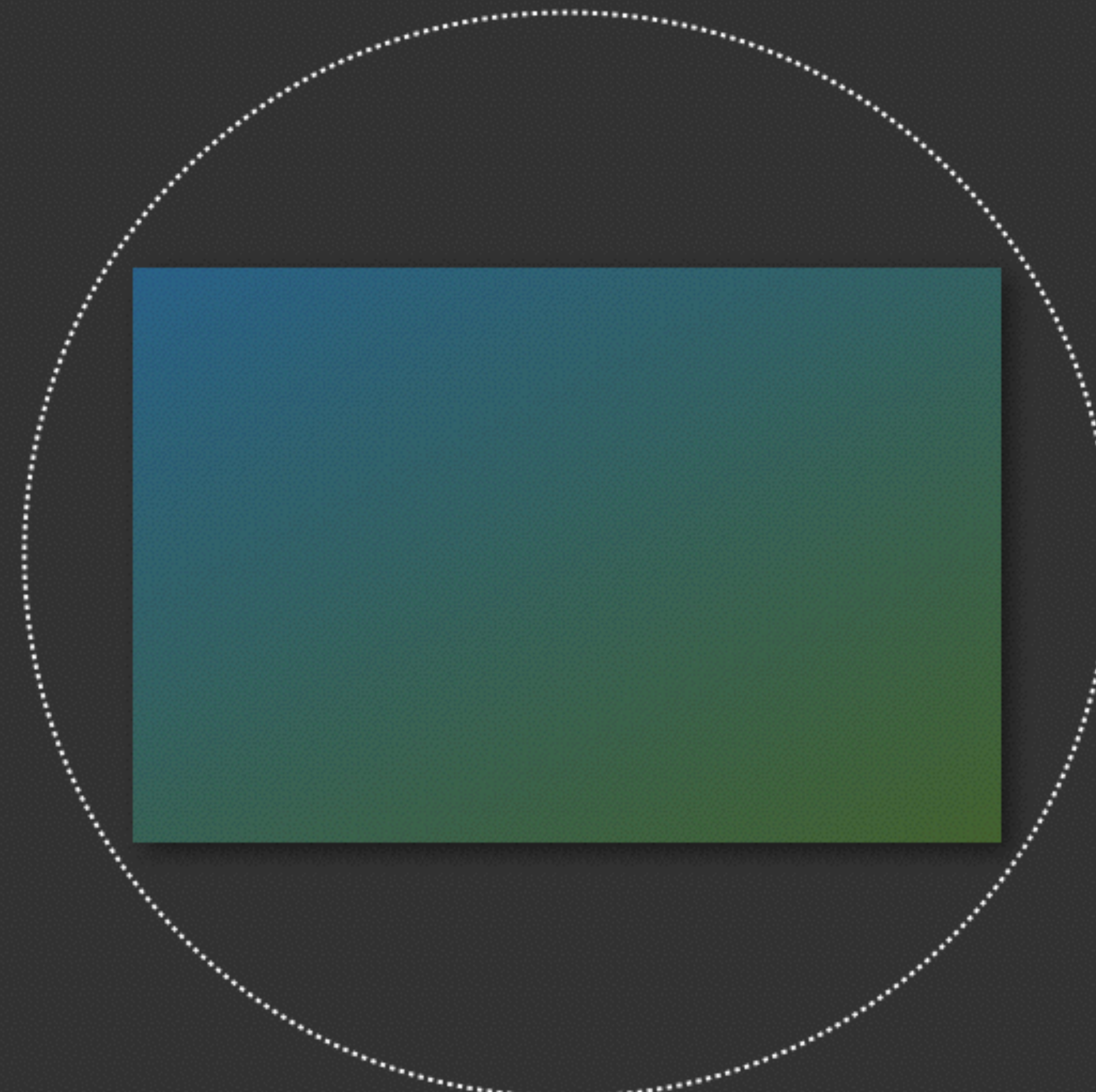
Full-Frame Sensor
36 x 24 mm
Up to 36 MP
4.9 micron pixel

1/3" Sensor
4.8 x 3.6 mm
Up to 13 MP
1.12 micron pixel

[Slide courtesy Ren Ng]

Sensor industry has large untapped resolution

Full-Frame Sensor
36 x 24 mm
Up to 36 MP
4.9 micron pixel



Full-Frame Sensor
36 x 24 mm
688 MP
1.12 micron pixel

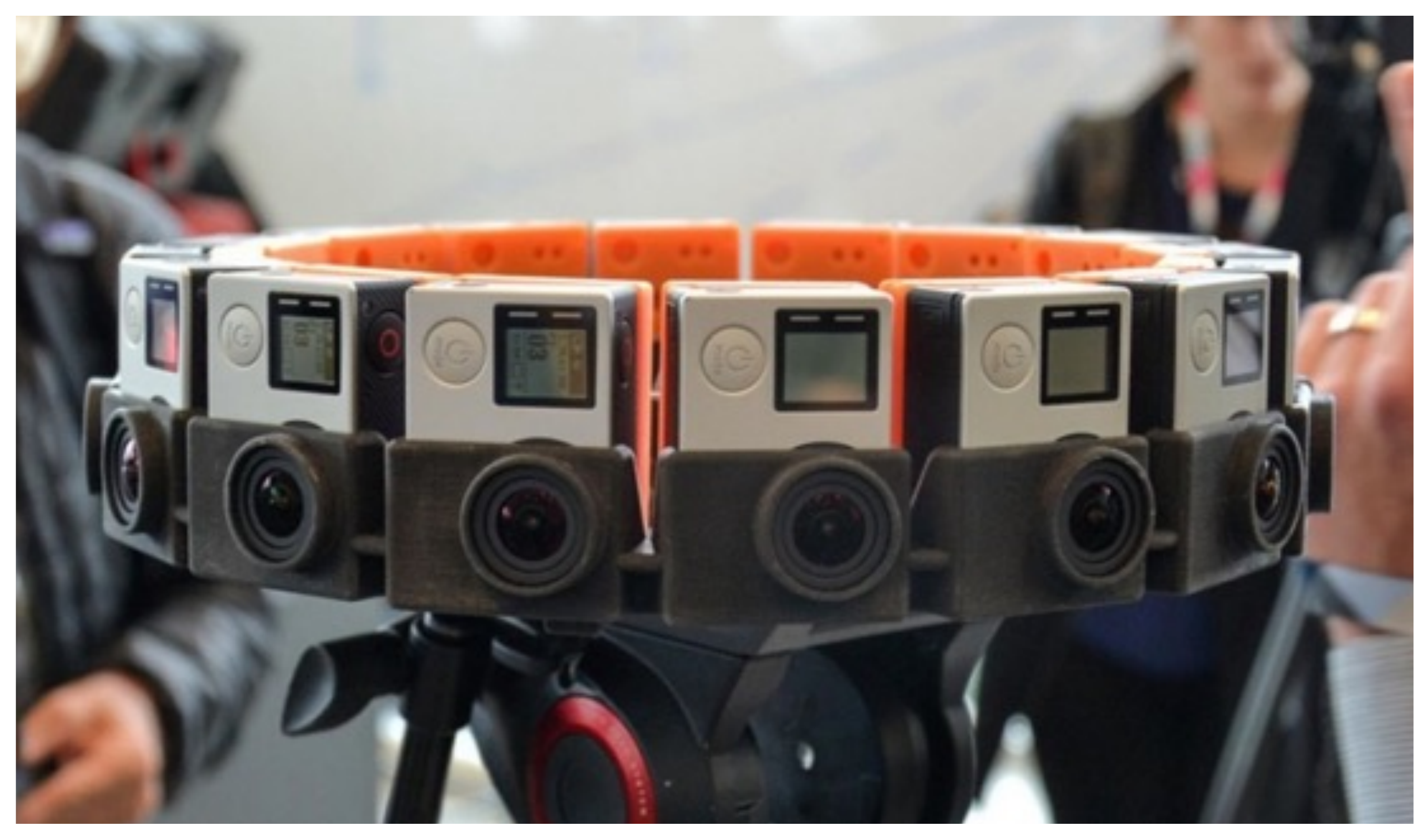
[Slide courtesy Ren Ng]

Lytro Cinema

755 Mpixel camera

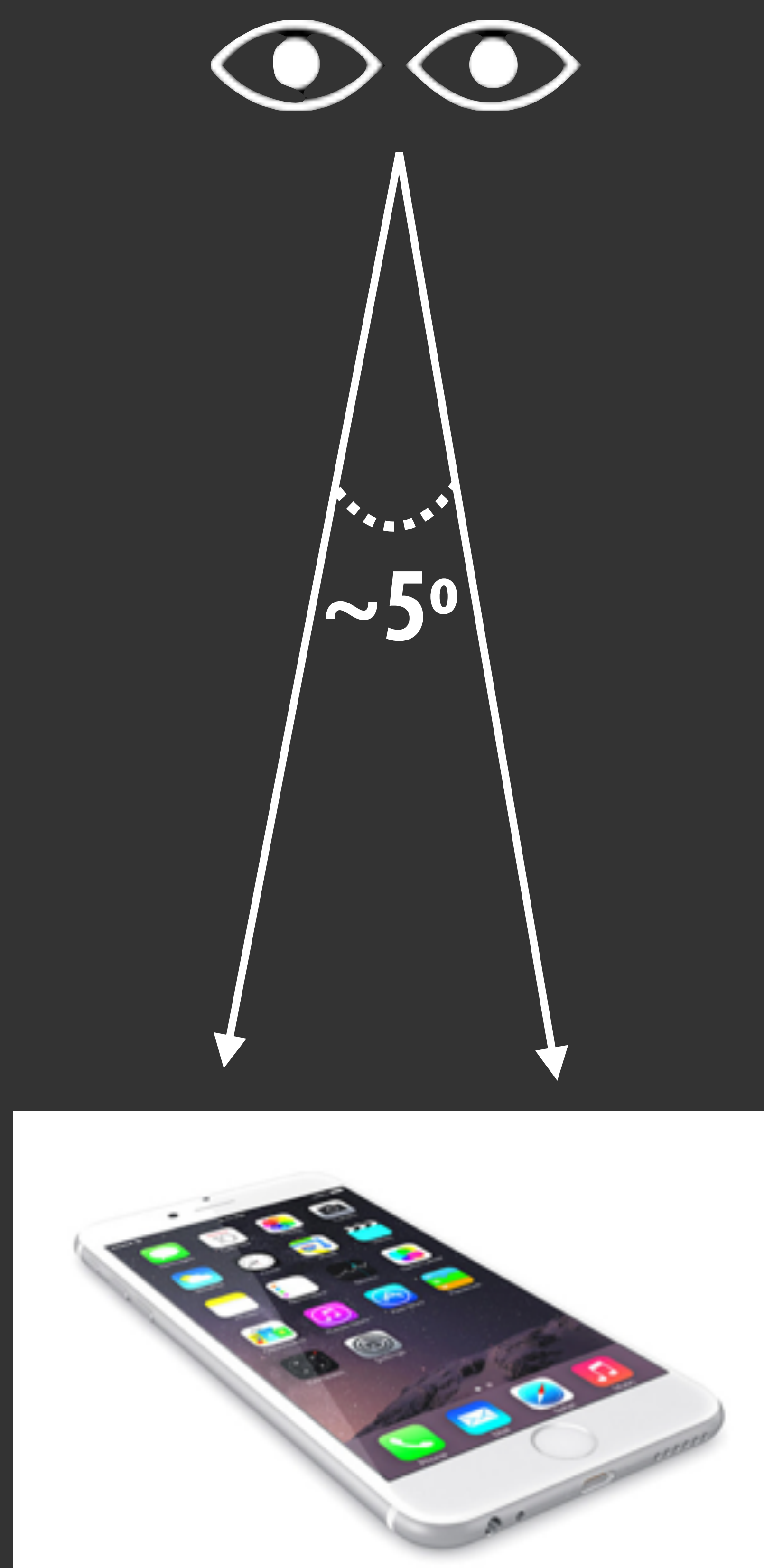
VR output

Example: Google's JumpVR video
Input stream: 16 4K GoPro cameras

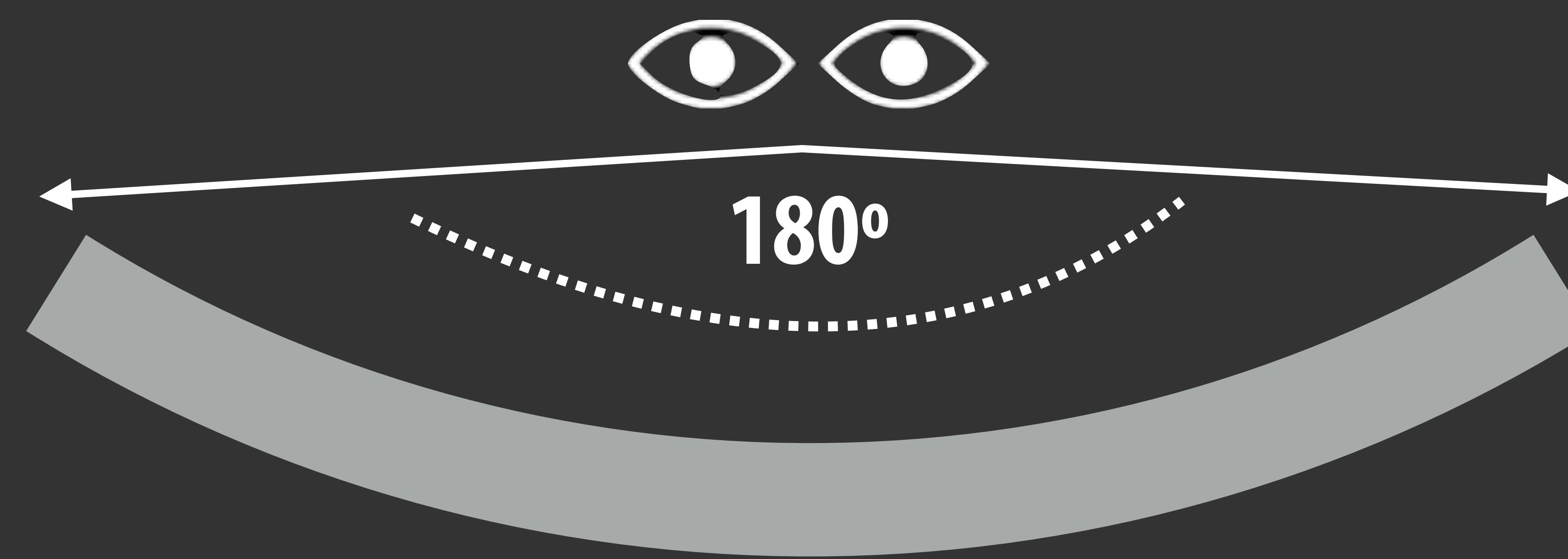


Register + 3D align video stream (on edge device)
Broadcast encoded video stream across the country to millions of viewers

VR creates high resolution requirements



iPhone 6: 4.7 in "retina" display:
1.3 MPixel
326 ppi → 57 ppd

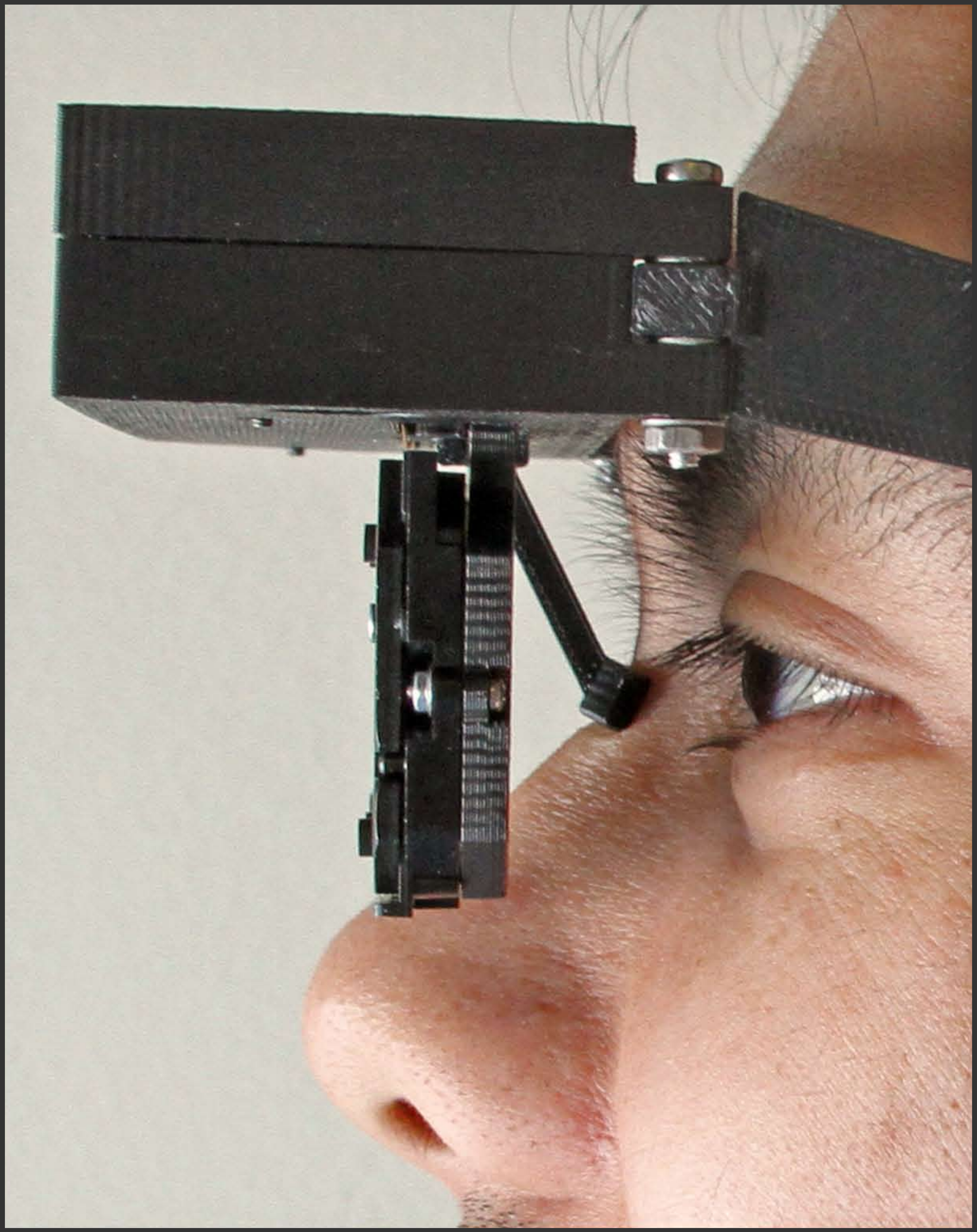


Future "retina" VR display:
57 ppd covering 180°
= 10K x 10K display per eye
= 200 MPixel

RAW data rate @ 120Hz ≈ 72 GB/sec

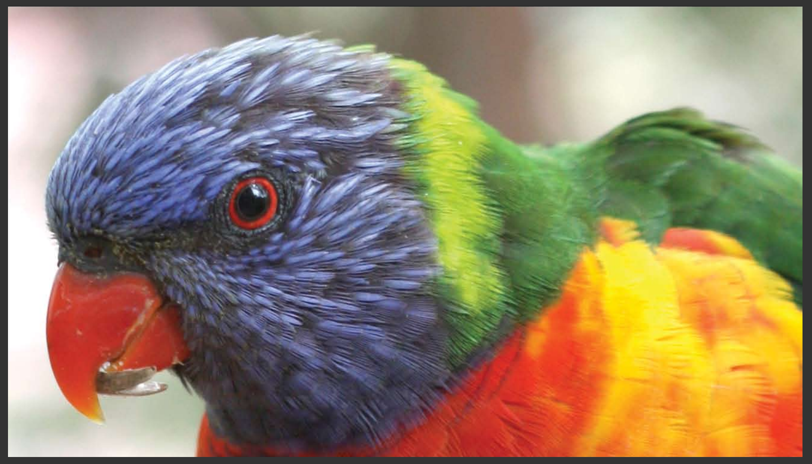
[Lanman 2013]

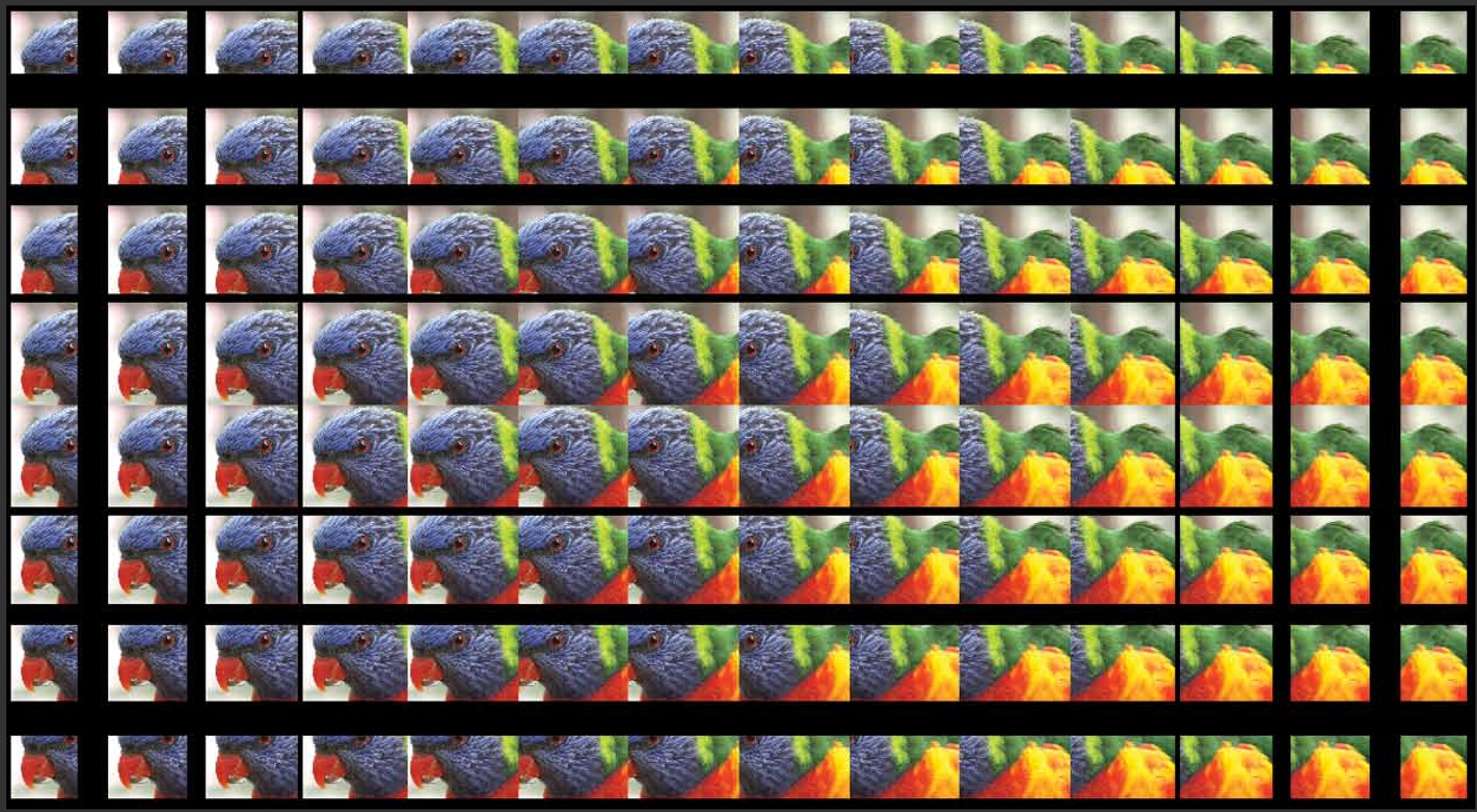
VR: Light field display



146 x 78 spatial resolution
Using 1MP microdisplay

Simple idea:
Recreate the same light field that was present in the scene when it was captured





Output of display (prior to optics)

Enhancing communication: understanding images to improve acquired content

AutoEnhance:

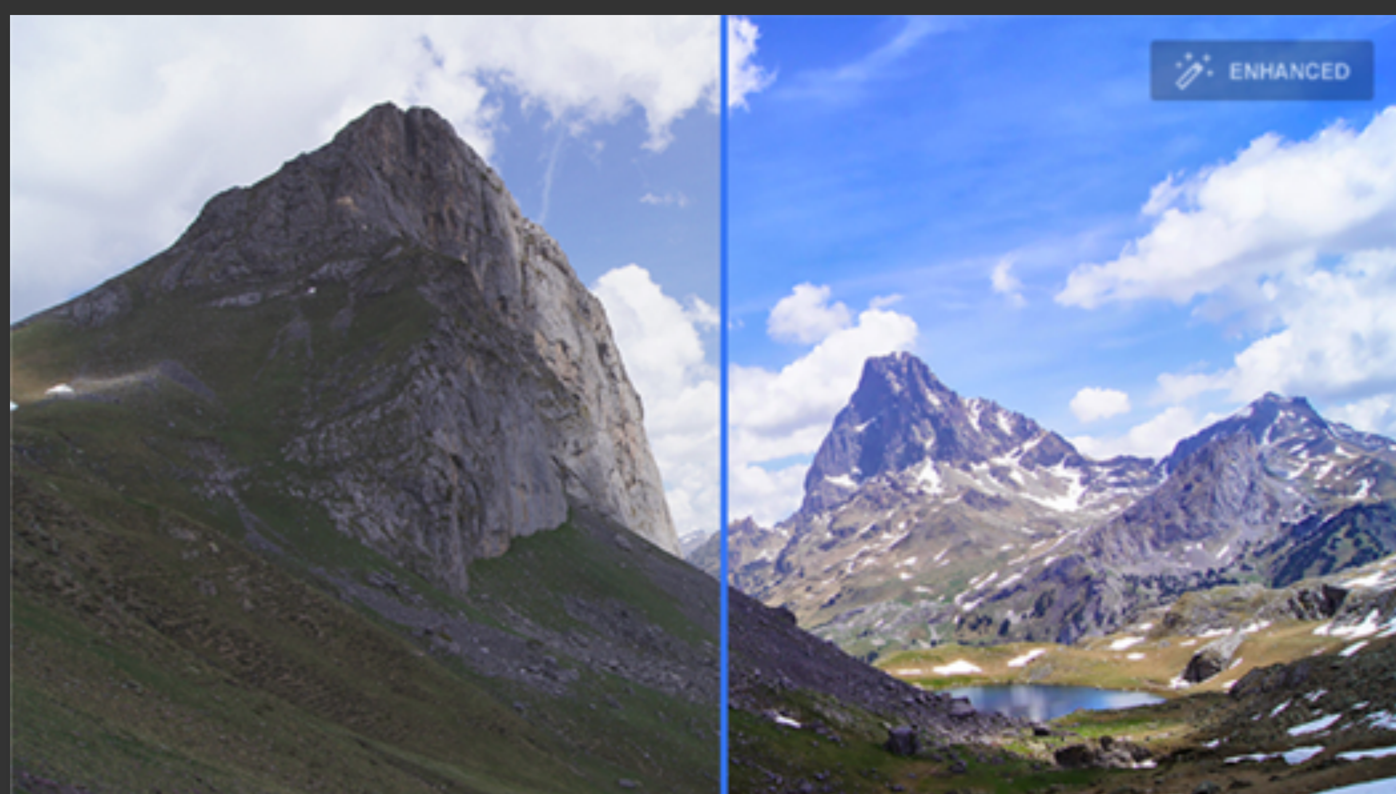
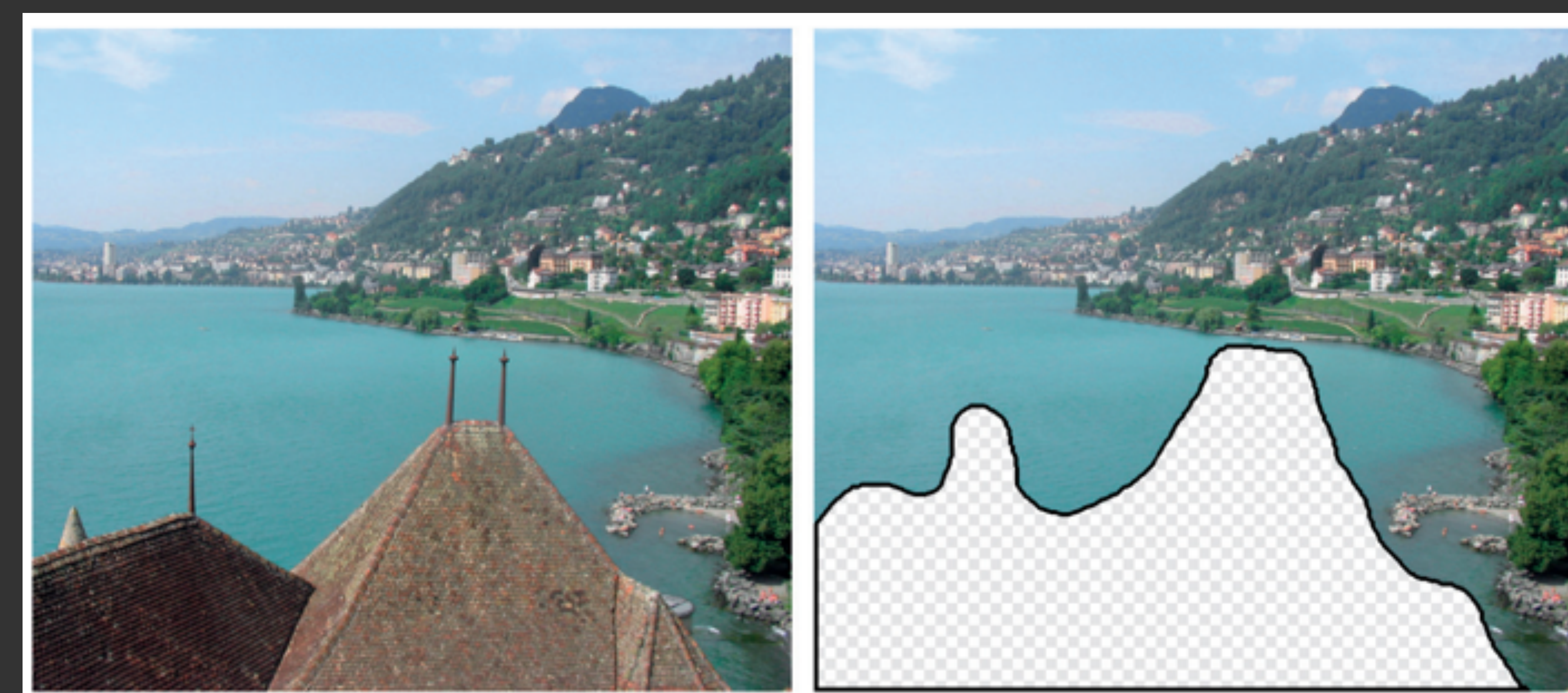
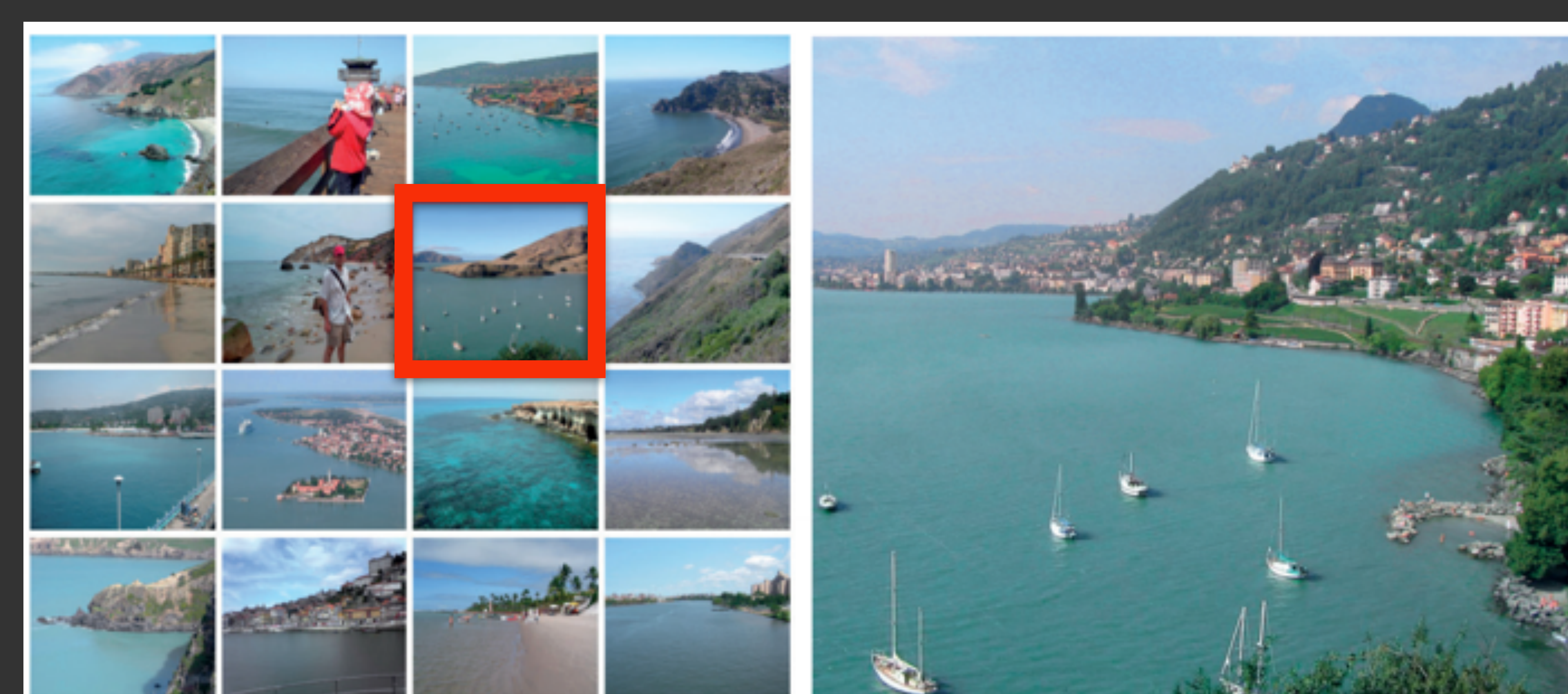


Photo “fix up” [Hayes 2007]



My bad vacation photo

Part to fix



Similar photos others have taken

Fixed!

Summary

We are observing rapid growth in the richness of visual communication

Sensing the world with higher fidelity to deliver improved content to humans

2030 challenge: recording and analyzing the world's visual information, so computers can understand and reason about it

Capturing everything about the visual world

To understand people

To understand the world around vehicles/drones

To understand cities

Mobile

Continuous (always on)

Exceptionally high resolution

Capture for computers to analyze, not humans to watch

Capturing images to understand humans

(why there will be high-resolution camera(s) always on, on every human)

Google Glass

Gwangjiang Market (Seoul)

What does this say?

What is this?


동부A
59-1
호
순희네 빈대떡
2273-5057
동부A
59-1
호
순희네
빈대떡
2273-5057
58
호
빈대떡,고기전
MG새마을금고
phenontay.com

Is it okay for me to sit there?

Is this woman annoyed that I sat down beside her? (Am I offending anyone?)

Why is she staring at me?

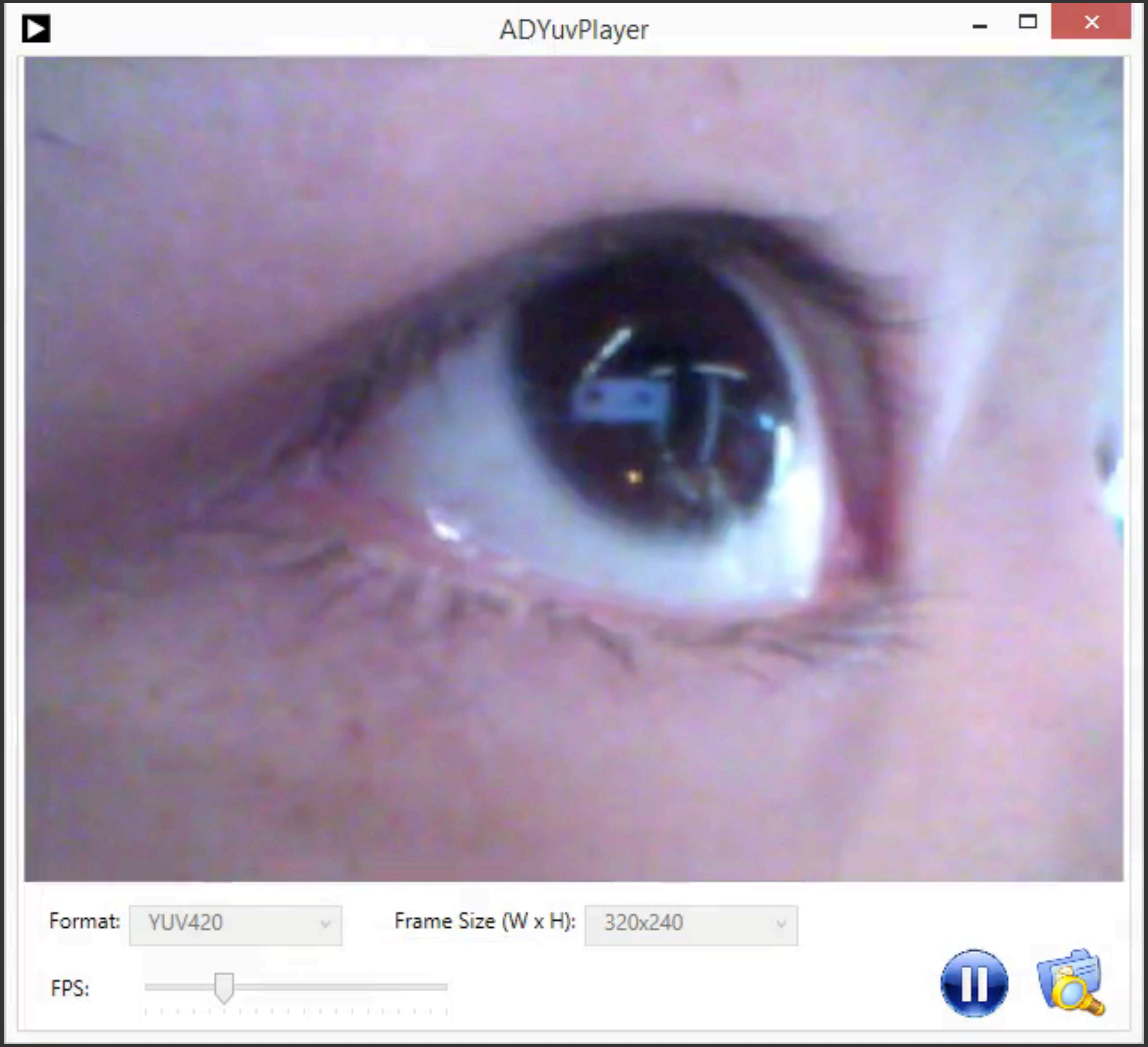
Should I attempt to greet the individuals at my table? (are they in a conversation that should not be interrupted)

When is a socially appropriate time to interrupt?

A future digital assistant must capture and comprehend extremely subtle aspects of human social behavior

Body language
Eye movement
Social context

Capturing / tracking eye movement



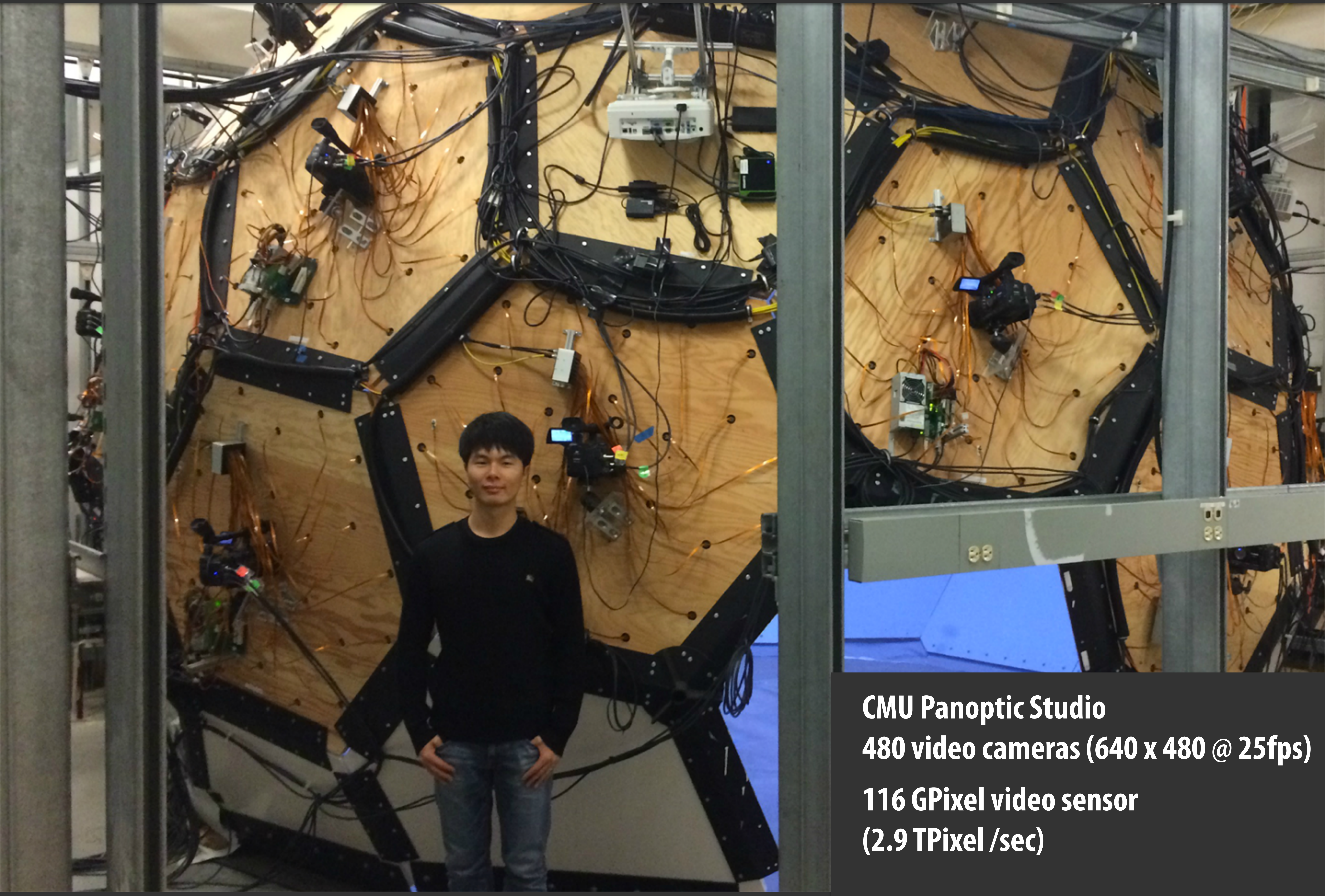
[Courtesy Yaser Sheikh]

Capturing subtle facial expressions

[Gotardo, Simon 2015]

Sensing human social interactions

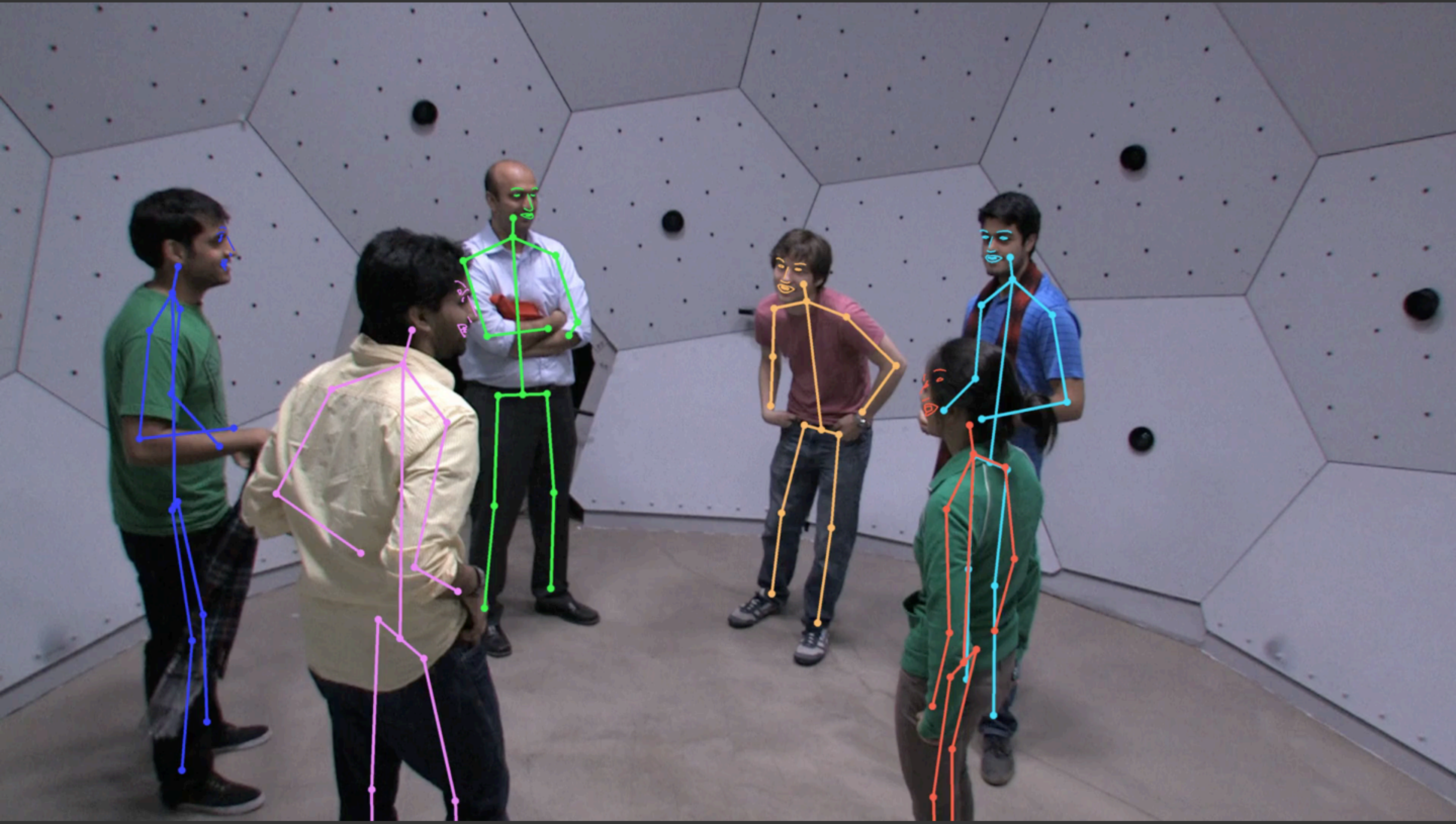
[Joo 2015]



CMU Panoptic Studio
480 video cameras (640 x 480 @ 25fps)

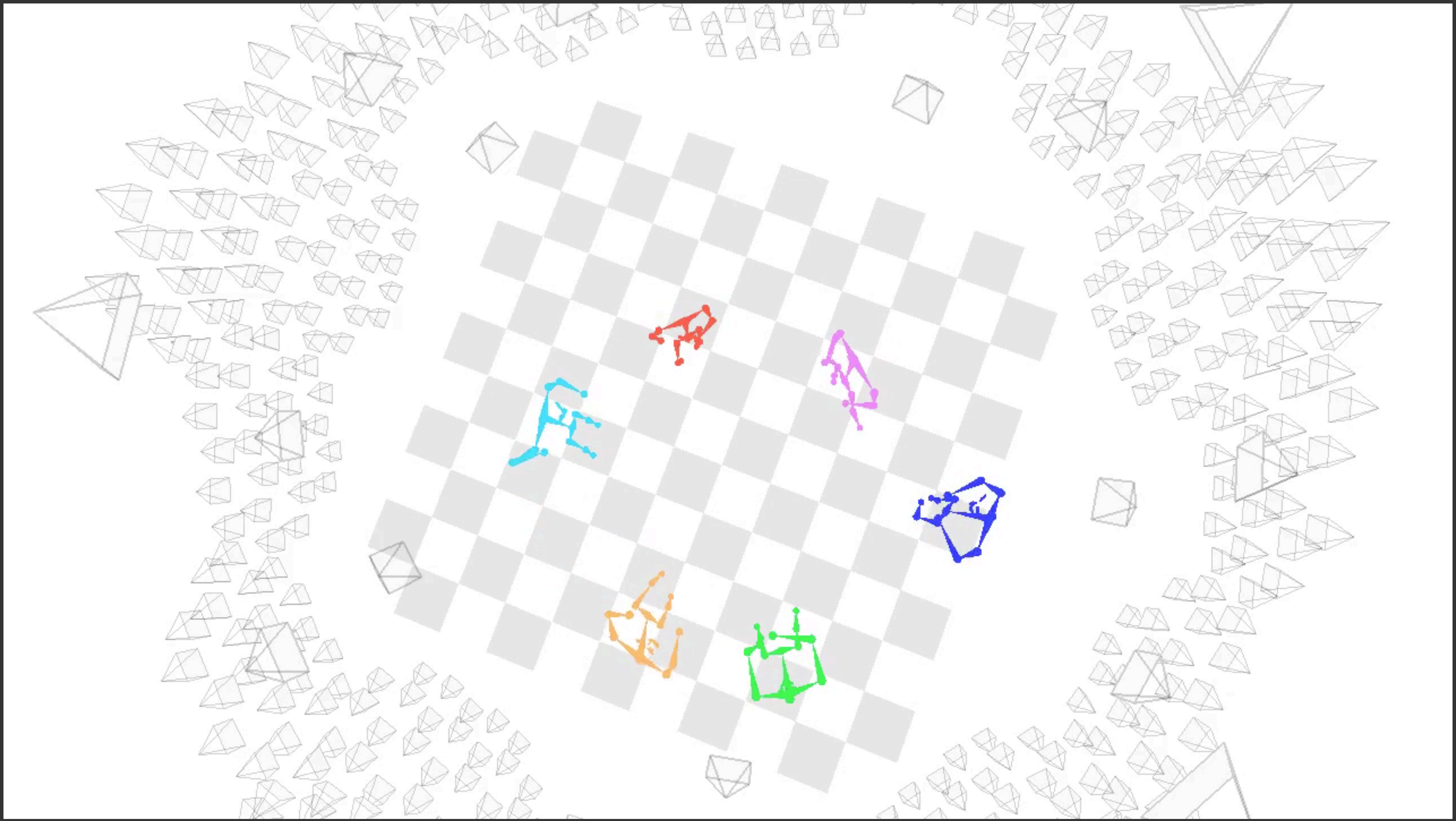
116 GPixel video sensor
(2.9 TPixel /sec)

Capturing social interactions



[Courtesy Yaser Sheikh, Tomas Simon, Hanbyul Joo]

Capturing social interactions



[Courtesy Yaser Sheikh, Tomas Simon, Hanbyul Joo]

What is the latent dimensionality of social signals?

of people

BRDF
(surface appearance)

5 million vertices[1] x (12 + 100)[2] x 300 Hz x 3 = 500 GB/sec

3-space coordinates

Sampling rate[3]

1 Based on USC-ICT Scan Resolution of Faces

2 Kautz et al., "Fast Arbitrary BRDF Shading for Low-Frequency Lighting Using Spherical Harmonics," 2002.

3 Andersson et al., "Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more", Journal of Eye Movement Research, 2010.

[Courtesy Yaser Sheikh]

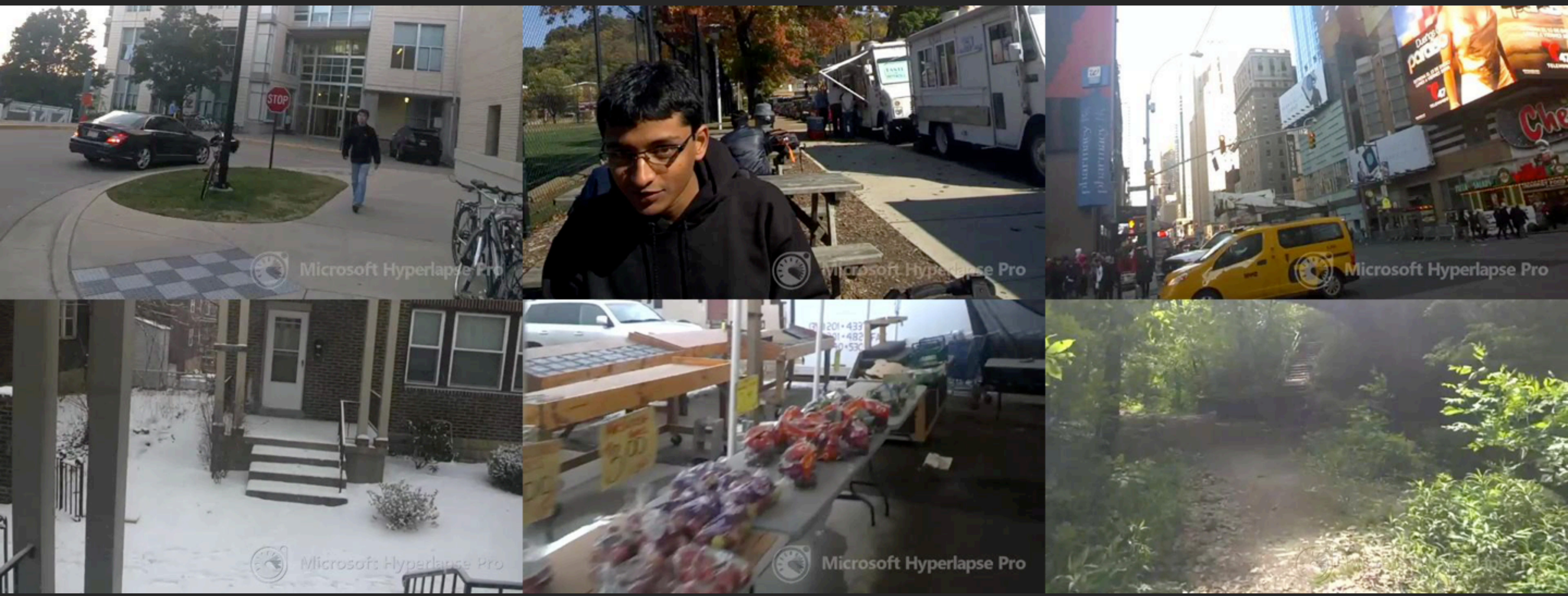
[Singh 2016]

Context is learned over time
"KrishnaCam" egocentric video dataset

72 hours of recording over nine months: (Sep 2014 – May 2015)

Google Glass

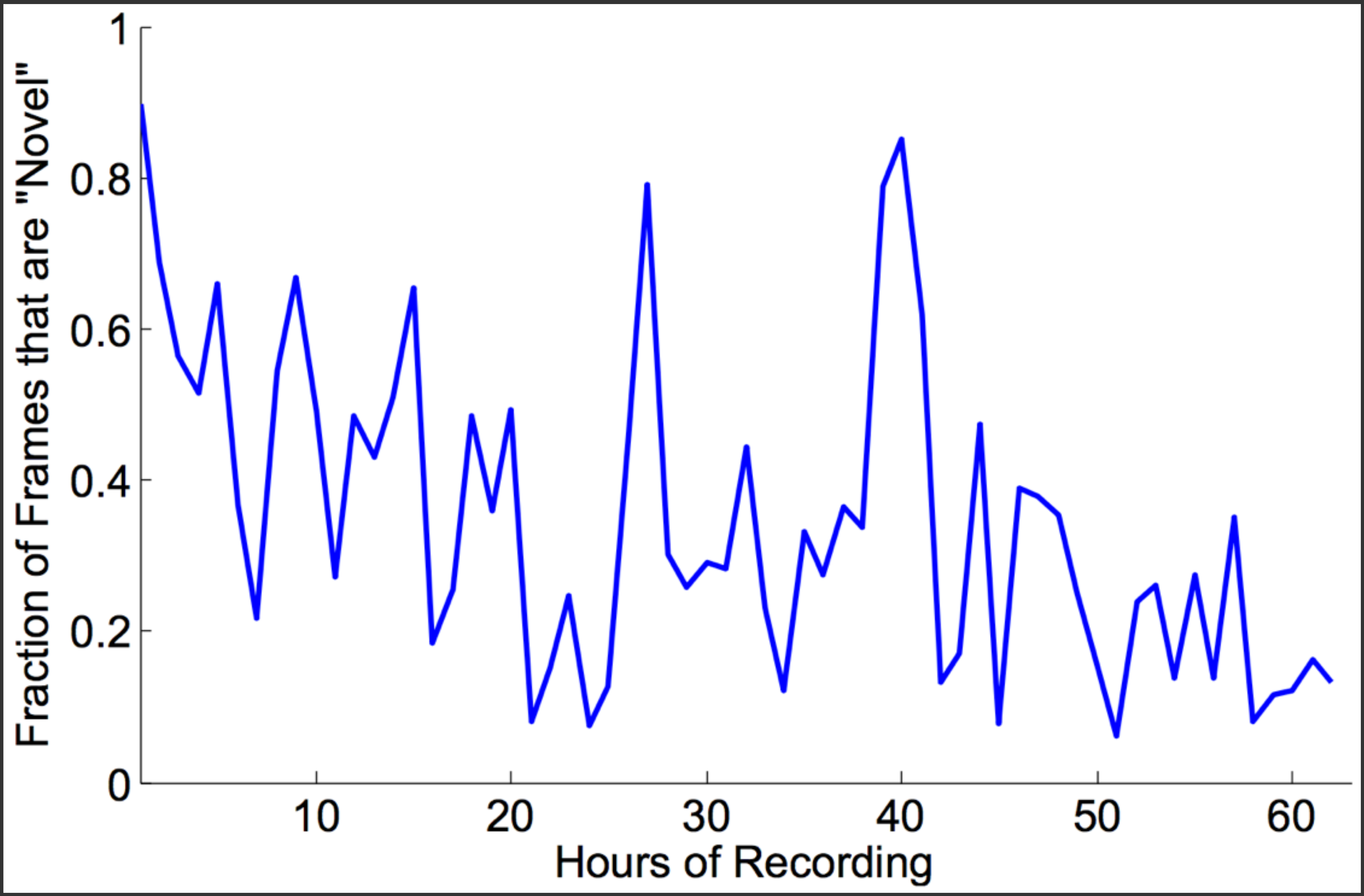
[Singh 2016]


STOP
Microsoft Hyperlapse Pro
Microsoft Hyperlapse Pro
Microsoft Hyperlapse Pro
Microsoft Hyperlapse Pro
Microsoft Hyperlapse Pro

Novel data growth

[Singh 2016]

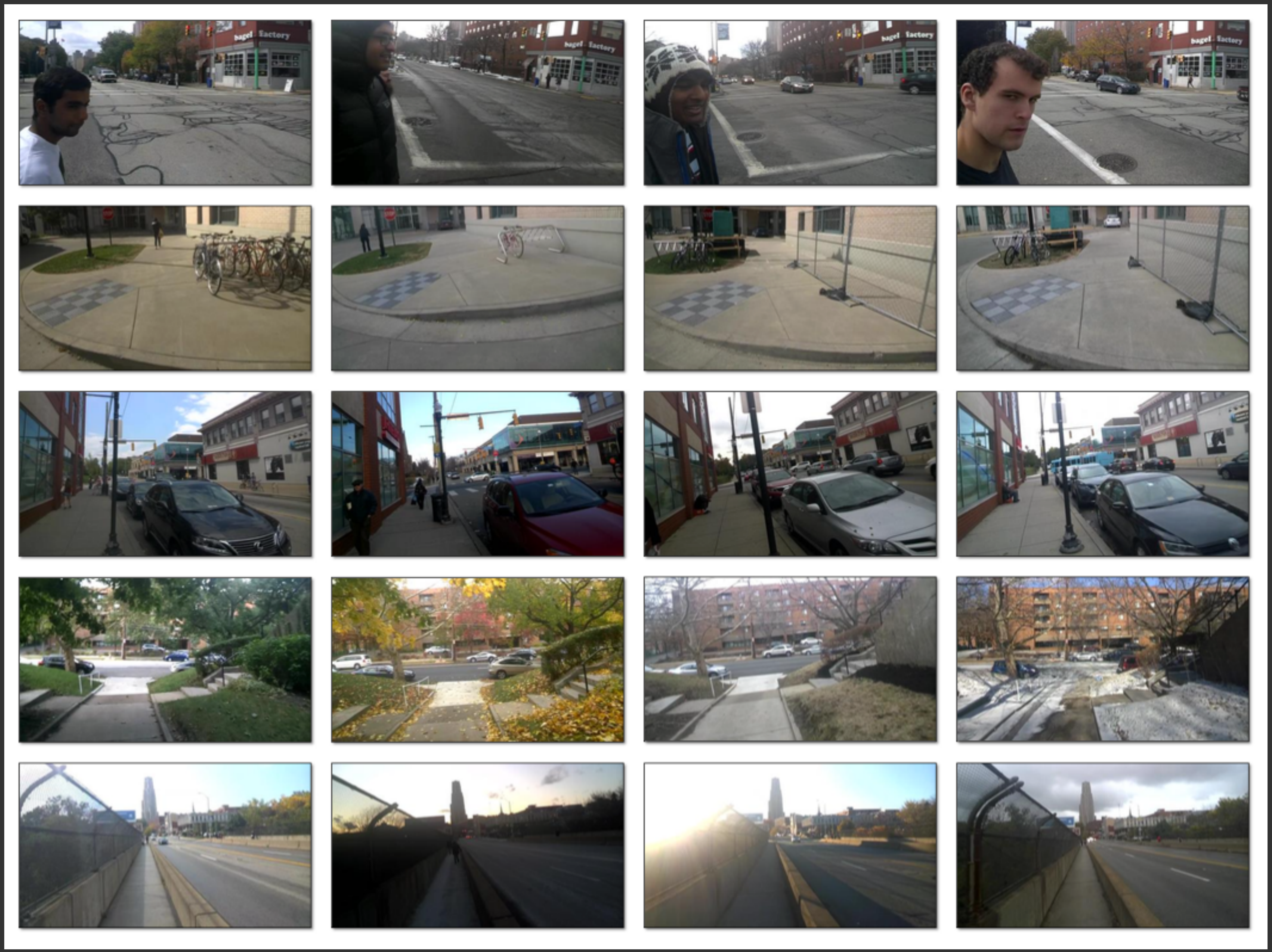
How much new visual data is seen as recording continues?



Similarity = cos distance of MIT Places layer 5 responses (full scene)
“Novel frames” = average distance to top-5 nearest neighbors greater than threshold

[Singh 2016]

How does the world evolve?

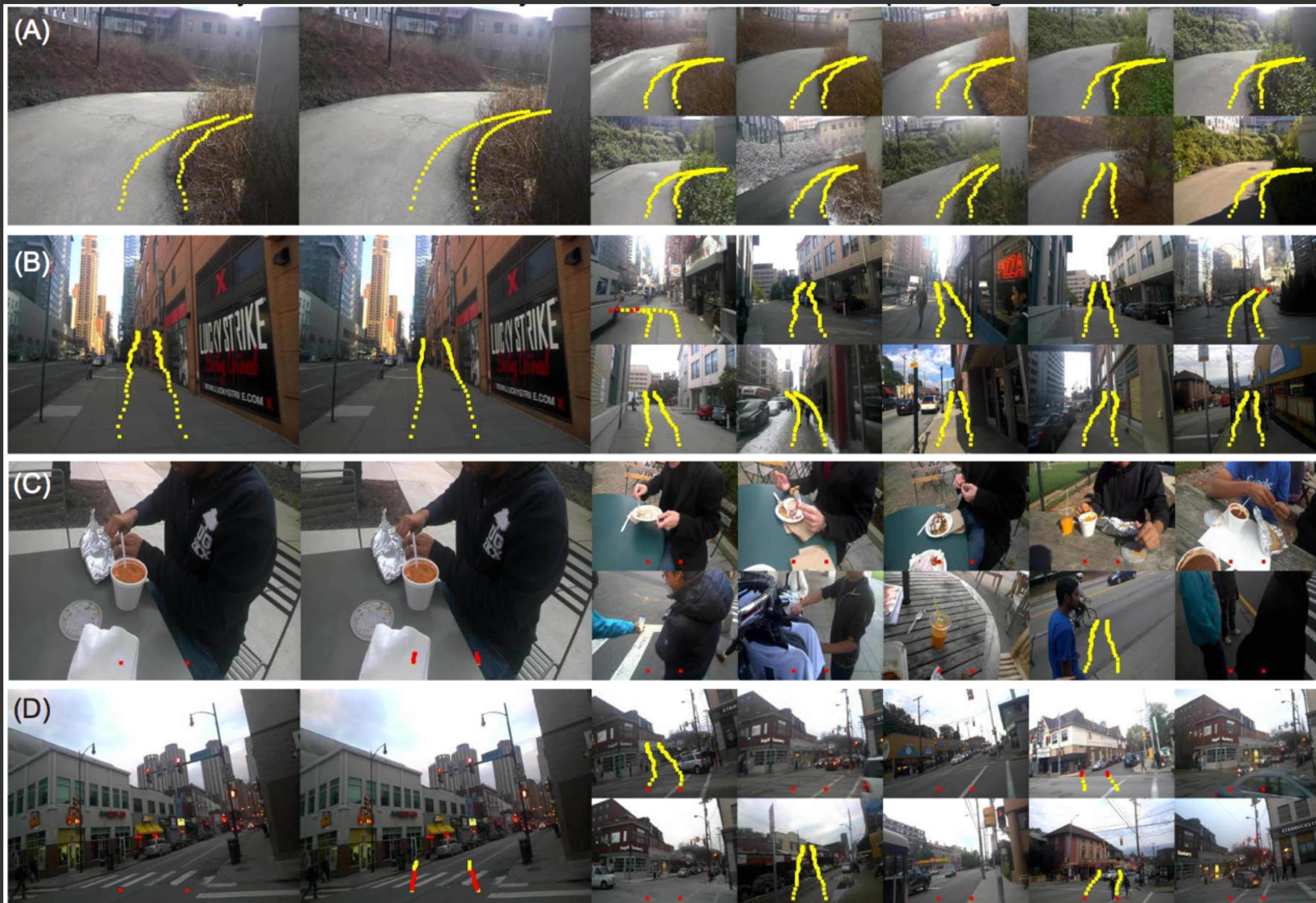


1. Change in companion
2. Change in object location (bike rack moved)
3. Change in object (different parked cars)
4. Change in season
5. Change in time of day (lighting conditions)

Predicting where Krishna will walk next

[Singh 2016]

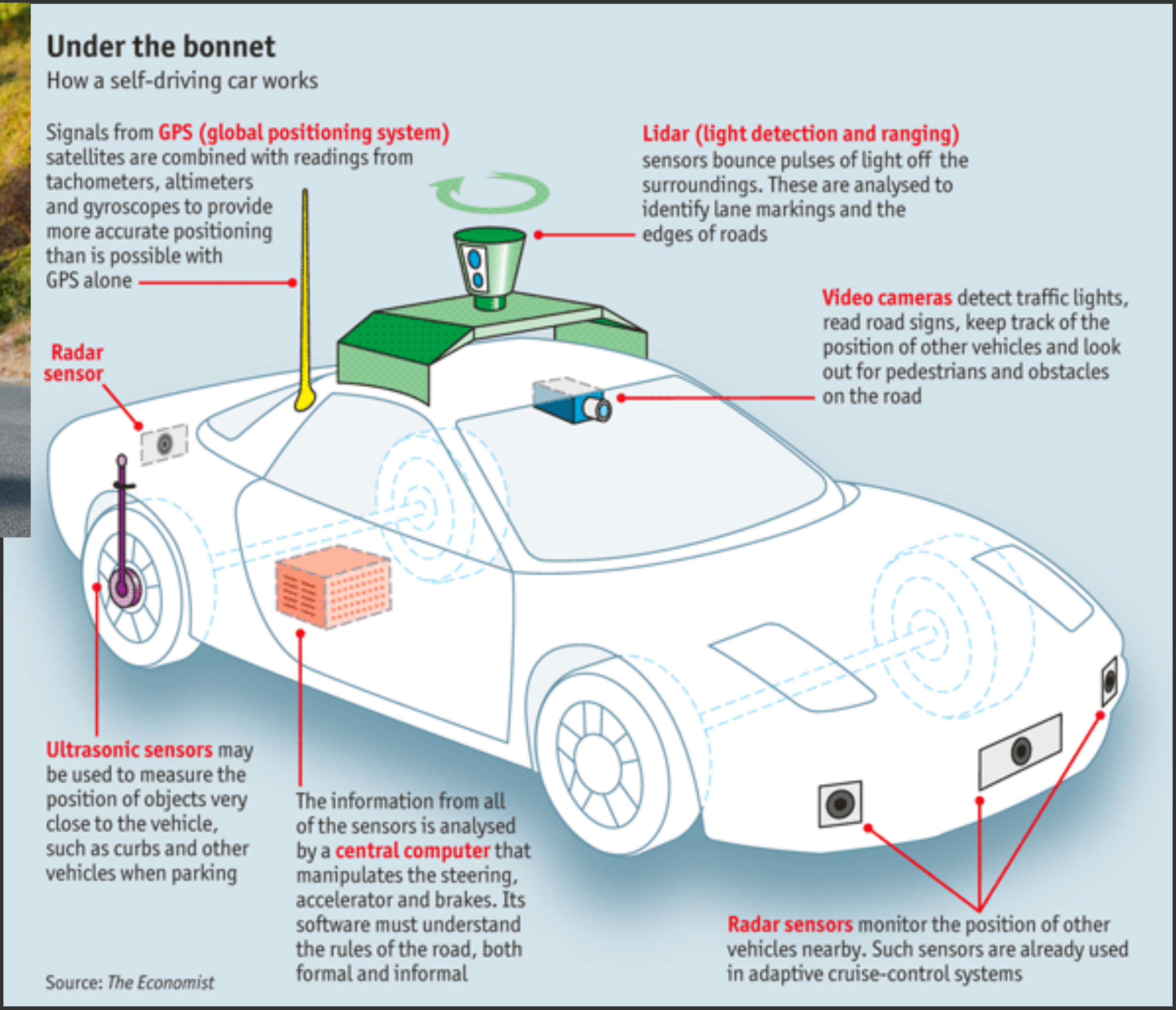
Current Frame: Ground Truth Traj | **Predicted Traj** | **Top 10 nearest neighbors to current frame**



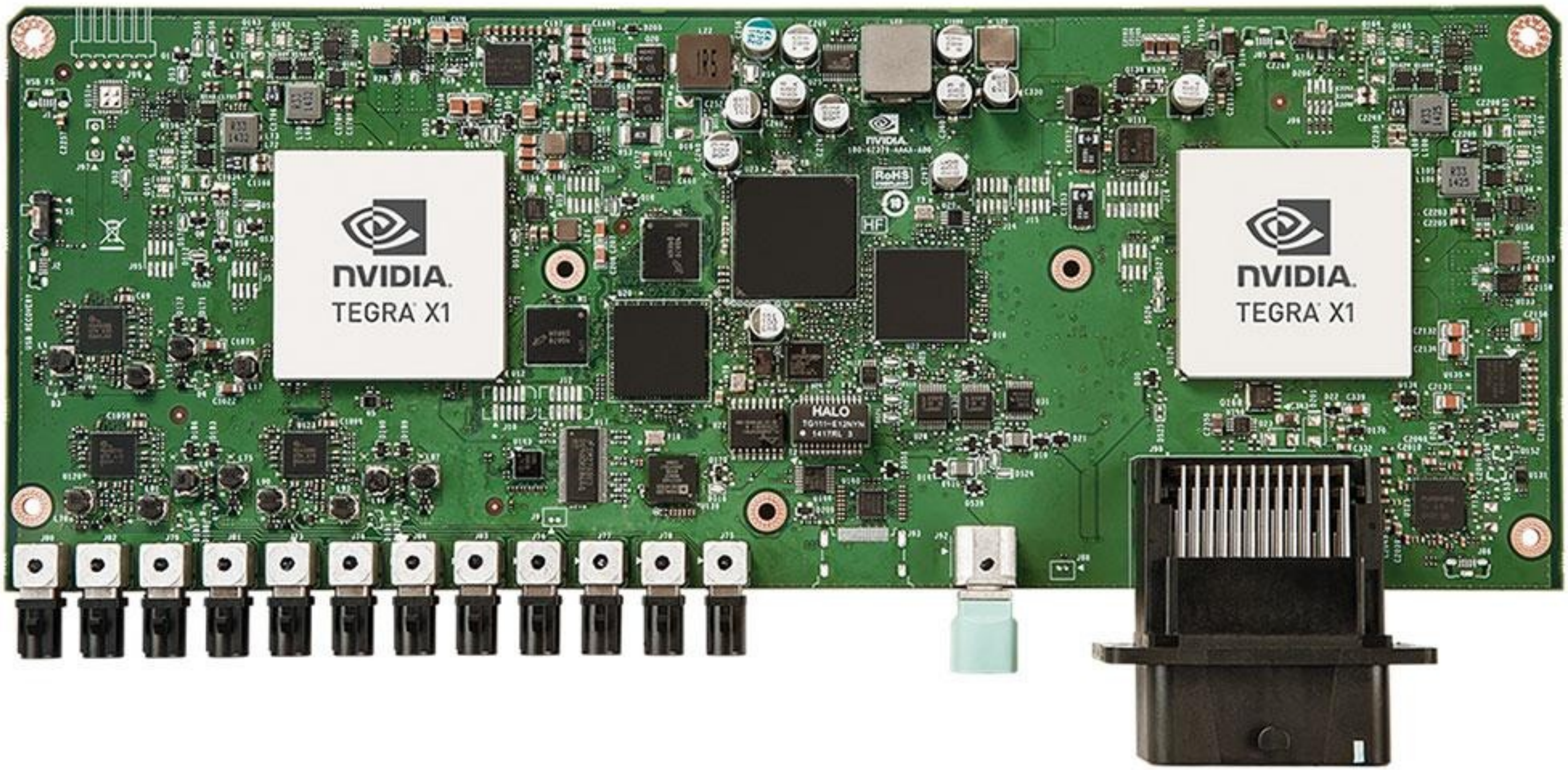
Capturing to localize and navigate

(cameras on every vehicle and robot)

Robot navigation depends on low-latency localization and surrounding object recognition



NVIDIA Drive PX



Tegra X1 (1 TFlop fp16 at 1GHz)

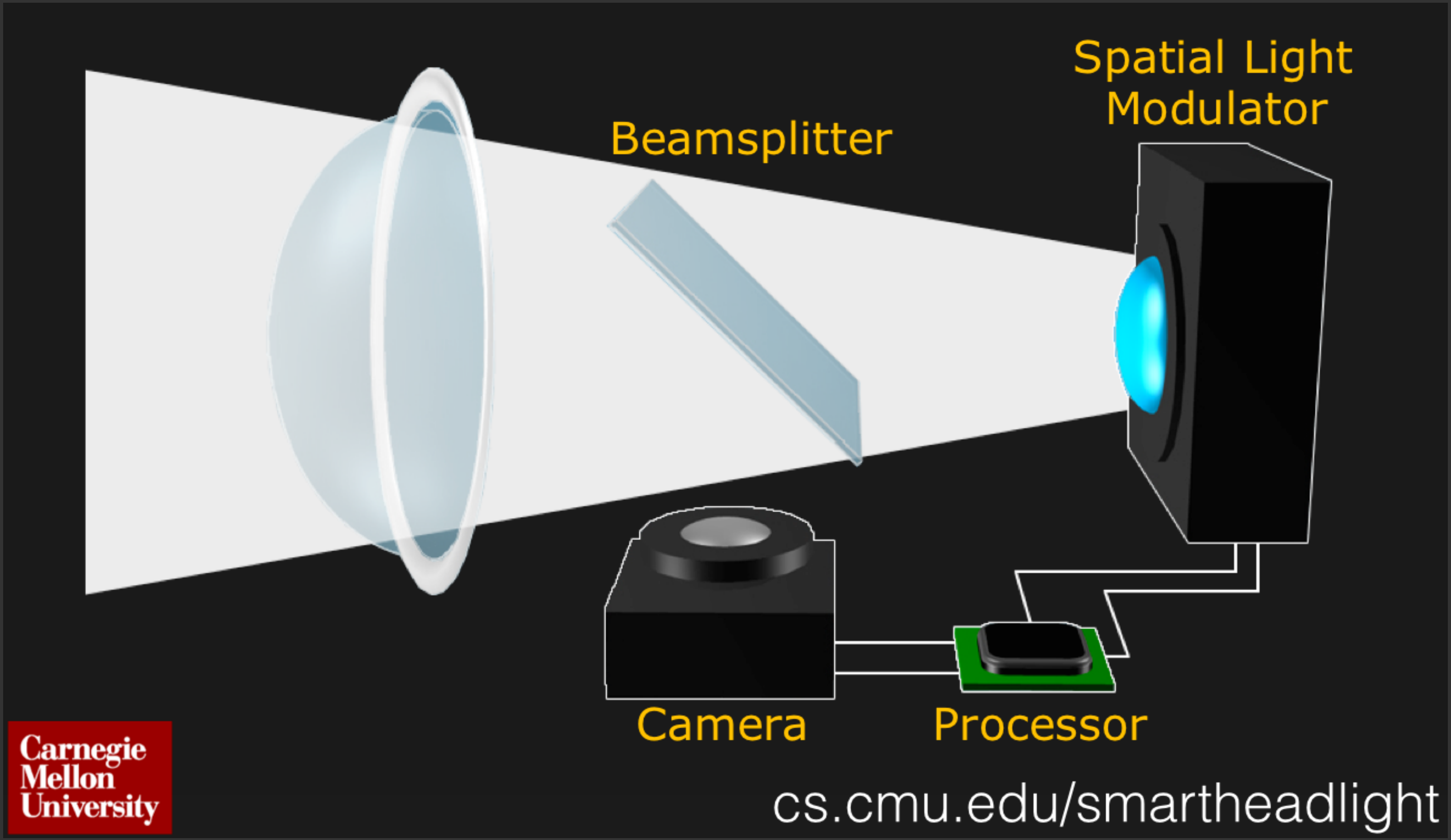
AR requires low-latency localization and scene object recognition

Making "maps": pervasive 3D construction


NISSAN
Carnegie Mellon University
cs.cmu.edu/smartheadlight

[Tamburo 2016]

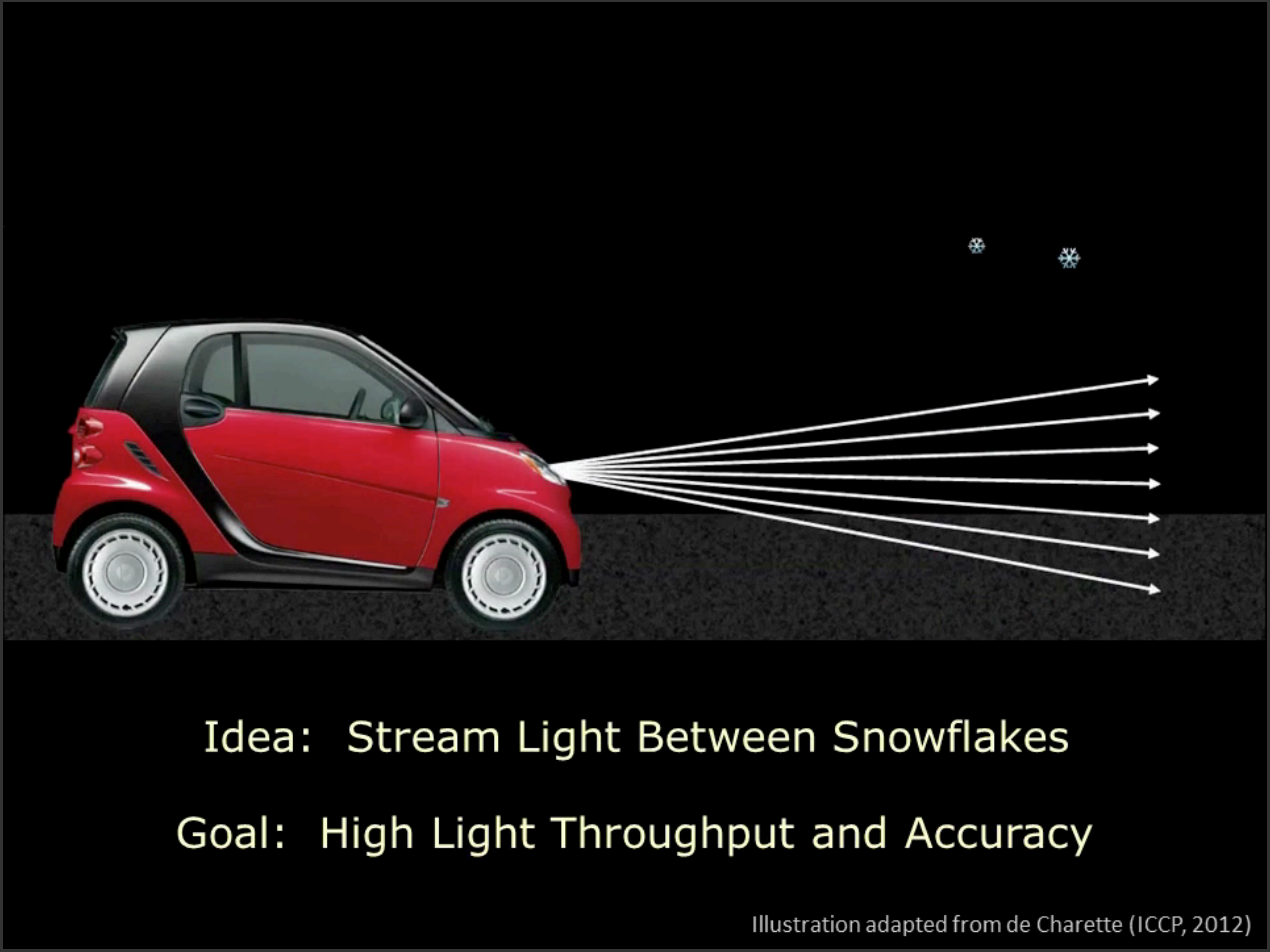
Smart headlight system



~1000 Hz (1 - 1.5 ms latency)

[Tamburo 2016]

Seeing clearly through precipitation



Capturing to understand cities

(Cameras on every street)

(The megacity as the distributed compute/sensing platform of the future)

"Managing urban areas has become one of the most important development challenges of the 21st century. Our success or failure in building sustainable cities will be a major factor in the success of the post-2015 UN development agenda." - UN Dept. of Economic and Social Affairs

Urban video command center

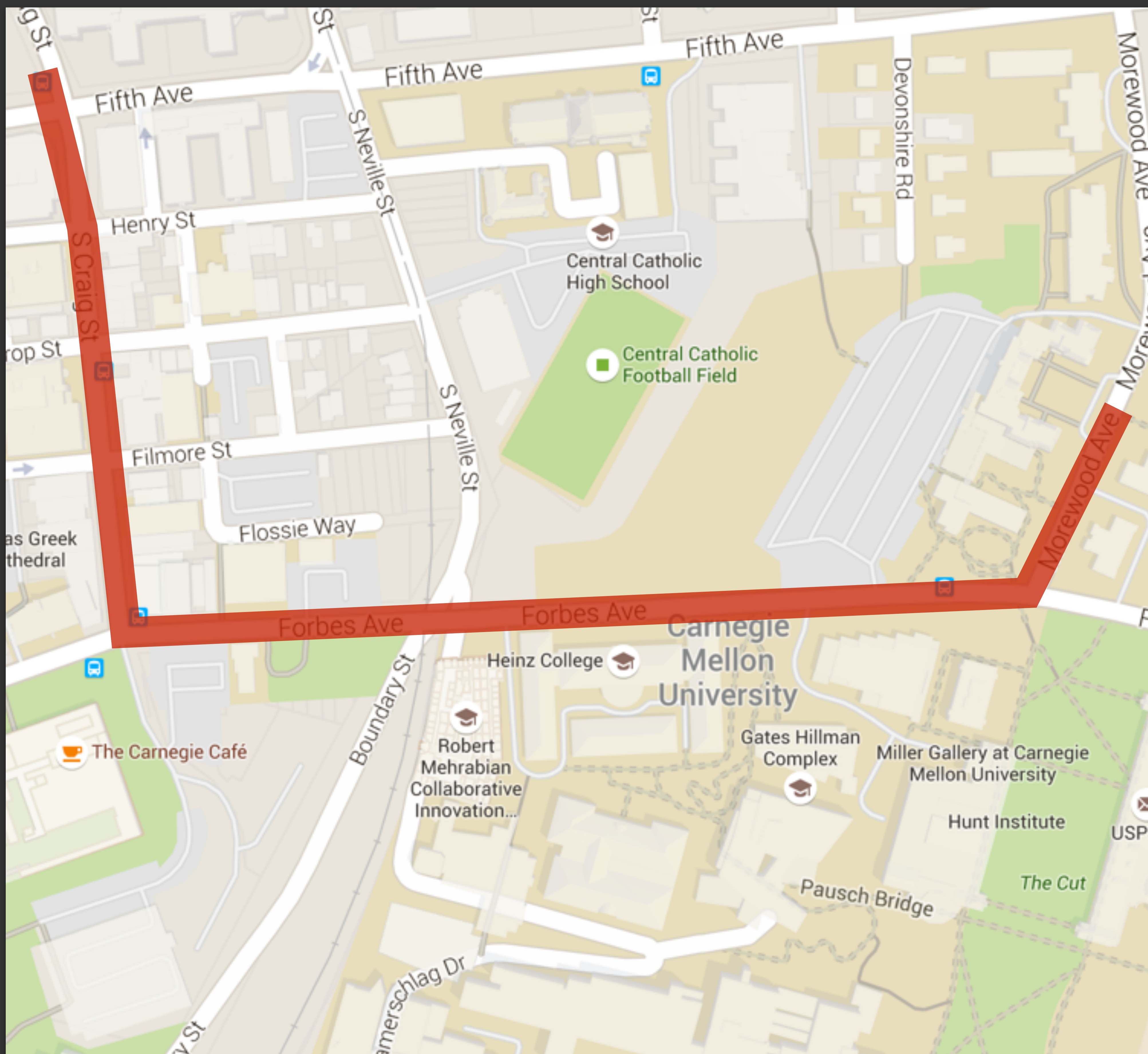
(Centro de Operações Preifetura do Rio de Janeiro)

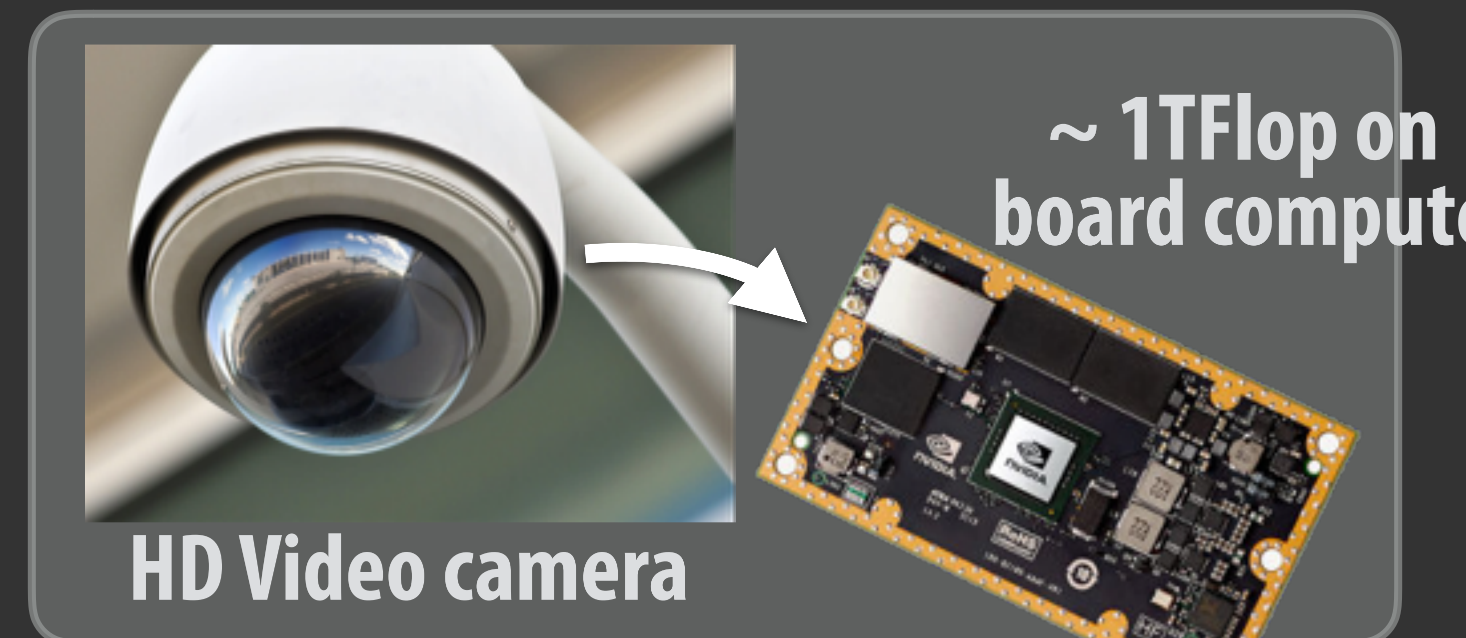
Urban camera deployments today

- 245M security cameras deployed worldwide (this number includes government owned and private)
  - 6,000 networked cameras in NYC
  - ~500,000 in Beijing [100% public area coverage]
  - 6M in UK, 20M in China

- Purpose is largely to observe and achieve for human query
  - Some ability to perform face / license plate detection, motion detection

Distributed software platform for Pittsburgh-scale video-based data mining and analytics





High speed link

On-campus Parallel Data Lab Datacenter

1. Use sensing infrastructure to actuate. How can video-based analytics improve city efficiency?
2. How do we build an platform that supports analytics application development for "all cameras in a city"?

Goal: establish viability of city instrumentation to deliver applications that improve efficiency and quality-of-life

~5 sec resolution query-able map of all cars, pedestrians, bicycles, etc.

Open parking spot detection and routing (eliminate circling for parking in greater Pittsburgh)

Postmortem analytics for city planning (How many times was a bike near a bus? Did pedestrians hold up traffic?)

Tracking/localization for autonomous vehicles

Accident or (near accident) detection

Hit-and-run detection (work with insurance companies)

Infrastructure monitoring: pot-hole detection, frozen street detection (salt truck allocation)

Air-quality monitoring

Watch my kids walk home alone after school…

Testbed for addressing interrelated technical, political, and privacy issues

Edge-to-datacenter distribution of computation (scheduling applications across the datacenter and to the edge)

Multi-tenancy near the image sensor (multiple applications must share sensor feeds)

First-class DBMS support for visual computing data

Programming systems for expressing video analysis applications for this infrastructure ("how to program a city")

New computer vision models for attention and compression (leveraging history and priors to reduce datacenter ingest)

New representations for images and videos that preserve privacy (what information is acceptable to leave the camera? Blurred faces? Features?)

Working with local city government to establish policy and protocol as a research output.

The world in 2030

The world in 2030

- 8.5 billion people [UN estimate]
- 61% urban (41 “megacities” of 10M people or more) [UN estimate]
- 2 billion cars [Sperling 2009]
- 700 - 1.12B streaming security video cameras
  — Extrapolation from 245M in 2014, for growth between 7-10% [IHS]
- Assume 8K (7680 x 4320) stereo sources (2 x 33 MPixel image)
- **Total continuous capture capability of the world:**
  - **25.6B video streams**
  - **1.7×10^{18} pixels ≈ 2 quintillion pixels (2 exapixels)**

The world in 2030

- Total continuous capture capability of the world:
  - 25.6B streams (assumed 8K stereo)

- Consider evaluating a modern object-detection deep neural network (GoogLeNet) on every frame from these streams at 30 fps $\approx 10^{12}$ images/sec
  - Today: Tegra X1 fp16: 12 images/sec/watt on tiny 224x224 images [NVIDIA]
  - Let's (naively) multiply per frame cost by 100 to account for larger image size

  - With today's technology: 10^{13} Watts
  - Estimated world's power consumption in 2013: 10^{13} Watts

Final thoughts

- Computer graphics has always involved a healthy interaction between architecture, programming systems, and algorithms
  - Domain focus has been exceptionally useful for vertical thought
  - Willing to throw out old and re-engineer software (new hardware enables programs that haven't been written yet!)
  - Architects should know the algorithms well, and influence them!

- Visual computing has always challenged computer systems by its desire to simulate/synthesize complex visual information

- Next 1-2 decades: interpreting the worldwide visual signal
  - Acquiring and modeling everything humans would see, to enable computers to interpret and analyze
  - We will continue to take every op (op/Watt) you can give us

Thank you

Thanks to Yaser Sheikh, Srinivas Narasimhan, and Ren Ng for various slide credits