Introduction to Computer Vision
CS 280

Professors: Jitendra Malik & Angjoo Kanazawa
GSIs: Jathushan Rajasegaran, Rahul Ravishankar, Ryan Tabrizi

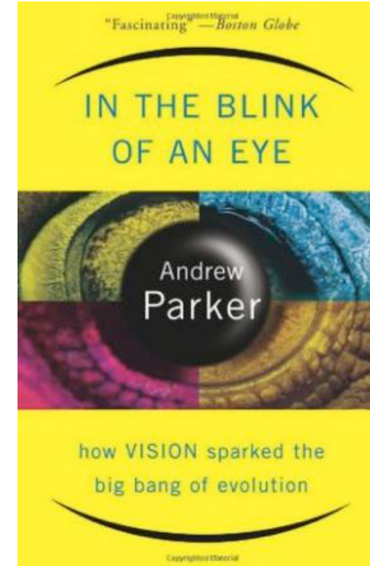
Course Website: https://cs280-berkeley.github.io/

Phylogeny of Intelligence

Cambrian Explosion
540 million years ago

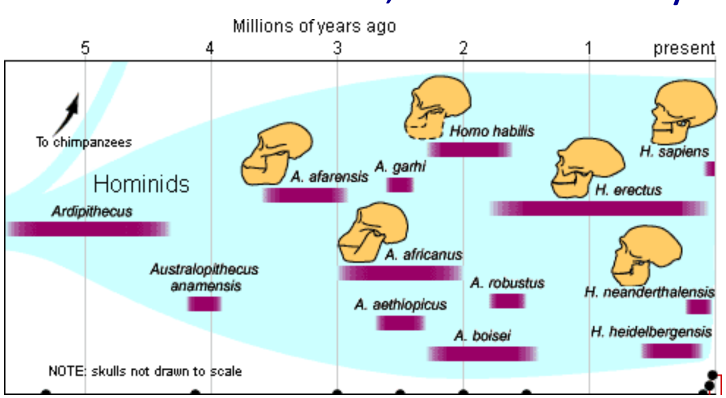
Variety of life forms,
almost all phyla emerge

Animals that could
see and move



Gibson: we see in order to move and we move in order to see

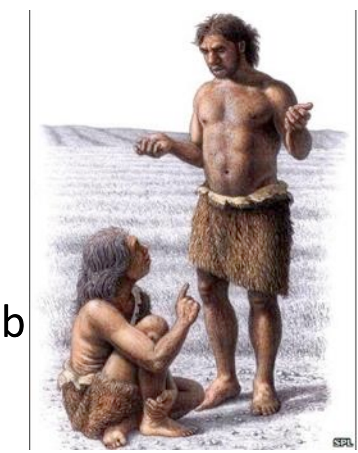
Hominid evolution, last 5 million years



Bipedalism
Opposable thumb
Tool use

Anaxogaras: It is because of his being armed with hands that man is the most intelligent animal

Modern humans, last 50 K years



Language
Abstract thinking
Symbolic behavior

The evolutionary progression

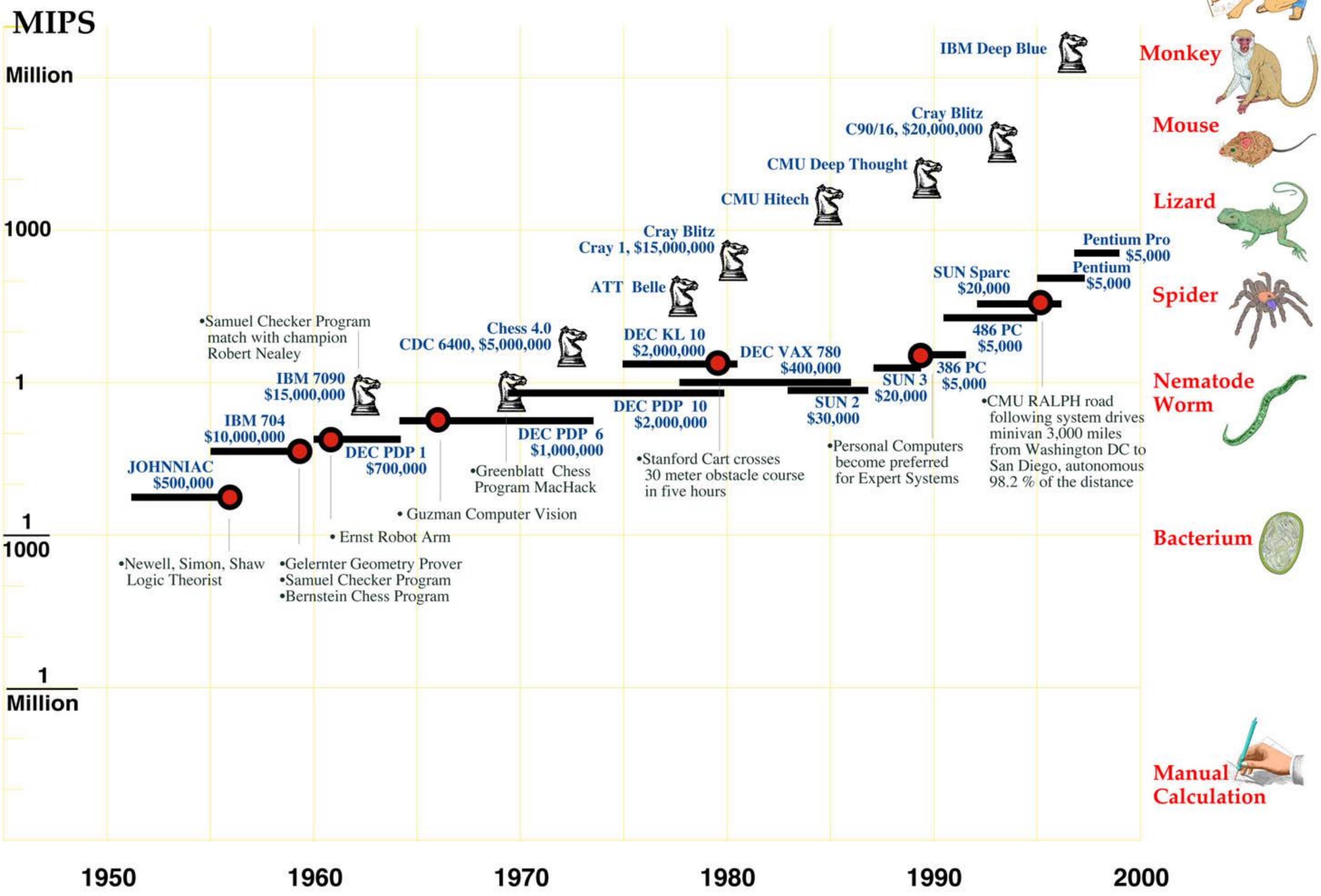
- Vision and Locomotion
- Manipulation
- Language

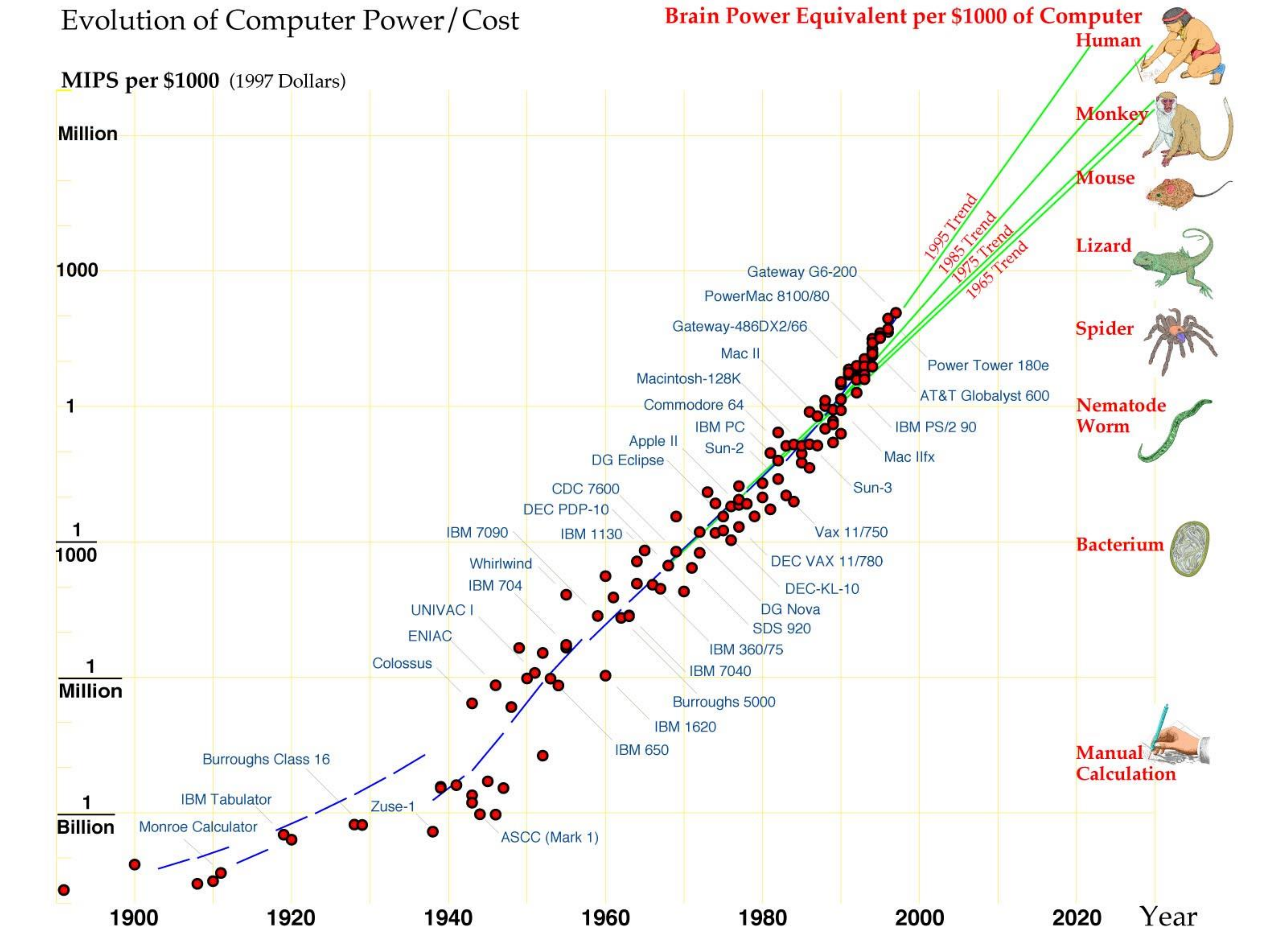
Moravec's argument(1998)

ROBOT: Mere Machine To Transcendent Mind

- 1 neuron = 1000 instructions/sec
- 1 synapse = 1 byte of information
- Human brain then processes 10^14 IPS and has 10^14 bytes of storage
- In 2000, we have 10^9 IPS and 10^9 bytes on a desktop machine
- Assuming Moore's law we obtain human level computing power in 2025.

Computer power available to AI and Robot programs





Evolution of Computer Power/Cost
Brain Power Equivalent per $1000 of Computer
MIPS per $1000 (1997 Dollars)
Million
1000
1
1/1000
1/Million
1/Billion
1900
1920
1940
1960
1980
2000
2020
Year
Human
Monkey
Mouse
Lizard
Spider
Nematode Worm
Bacterium
Manual Calculation
1995 Trend
1985 Trend
1975 Trend
1965 Trend
Gateway G6-200
PowerMac 8100/80
Gateway-486DX2/66
Mac II
Macintosh-128K
Commodore 64
IBM PC
Sun-2
Apple II
DG Eclipse
CDC 7600
DEC PDP-10
IBM 7090
IBM 1130
Whirlwind
IBM 704
UNIVAC I
ENIAC
Colossus
Power Tower 180e
AT&T Globalyst 600
IBM PS/2 90
Mac IIfx
Sun-3
Vax 11/750
DEC VAX 11/780
DEC-KL-10
DG Nova
SDS 920
IBM 360/75
IBM 7040
Burroughs 5000
IBM 1620
IBM 650
Burroughs Class 16
IBM Tabulator
Zuse-1
Monroe Calculator
ASCC (Mark 1)

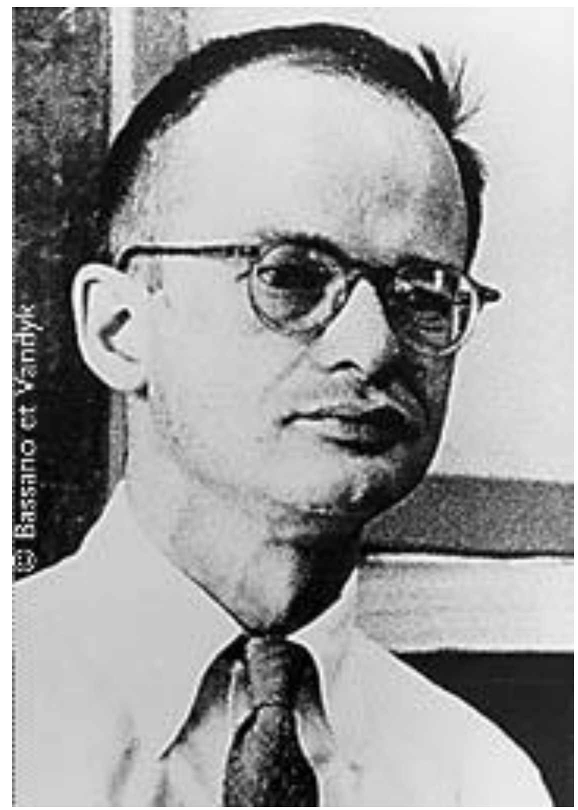
Moravec was right!

- Human brain processes 10^14 IPS and has 10^14 bytes of storage
- The NVIDIA H100 GPU has a computing power of approximately 67 TeraFLOPs (TFLOPs) in FP32 precision, meaning it can perform 67 trillion floating-point operations per second; in TF32 Tensor Core, it can reach up to 989 TeraFLOPs.

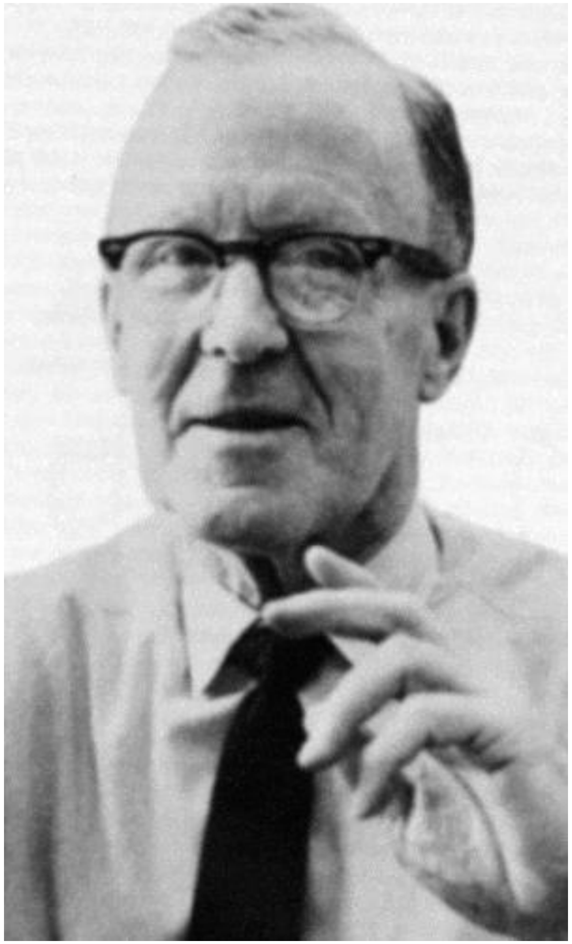
Some early history…

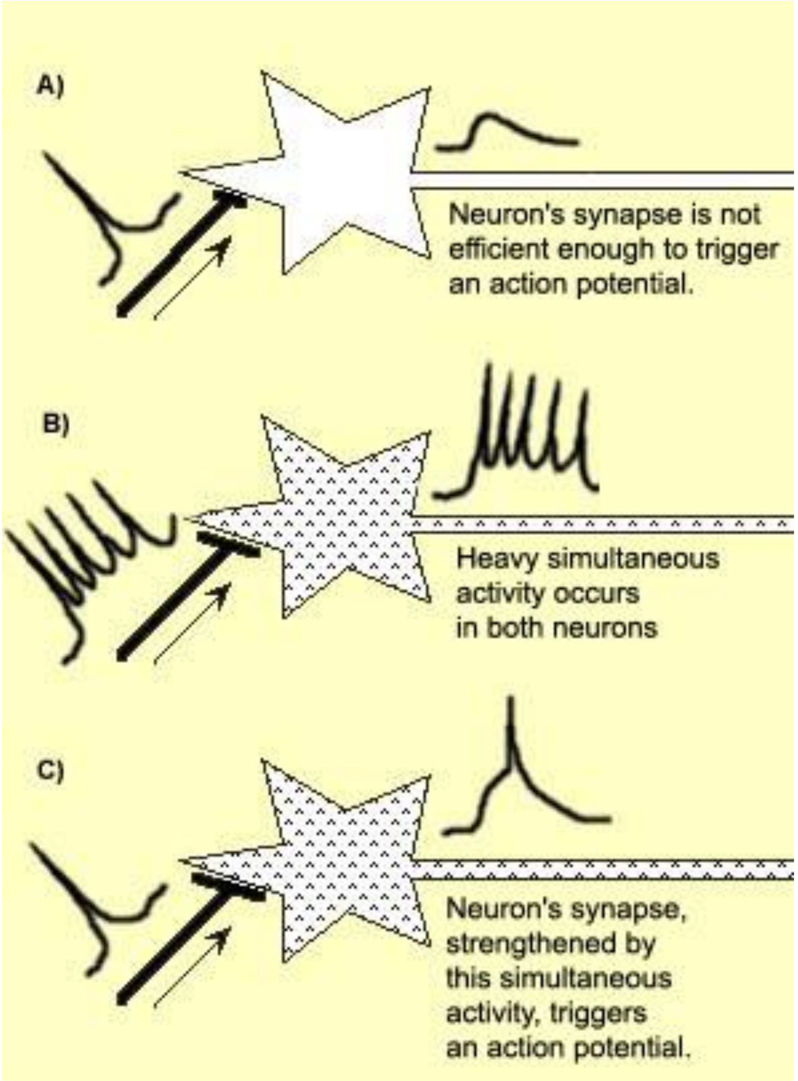
McCulloch & Pitts (1943)

A logical calculus of the ideas immanent in nervous activity



D. Hebb and Synaptic Learning





Turing's suggestion

Perception and Interaction

Language

456 A. M. TURING :

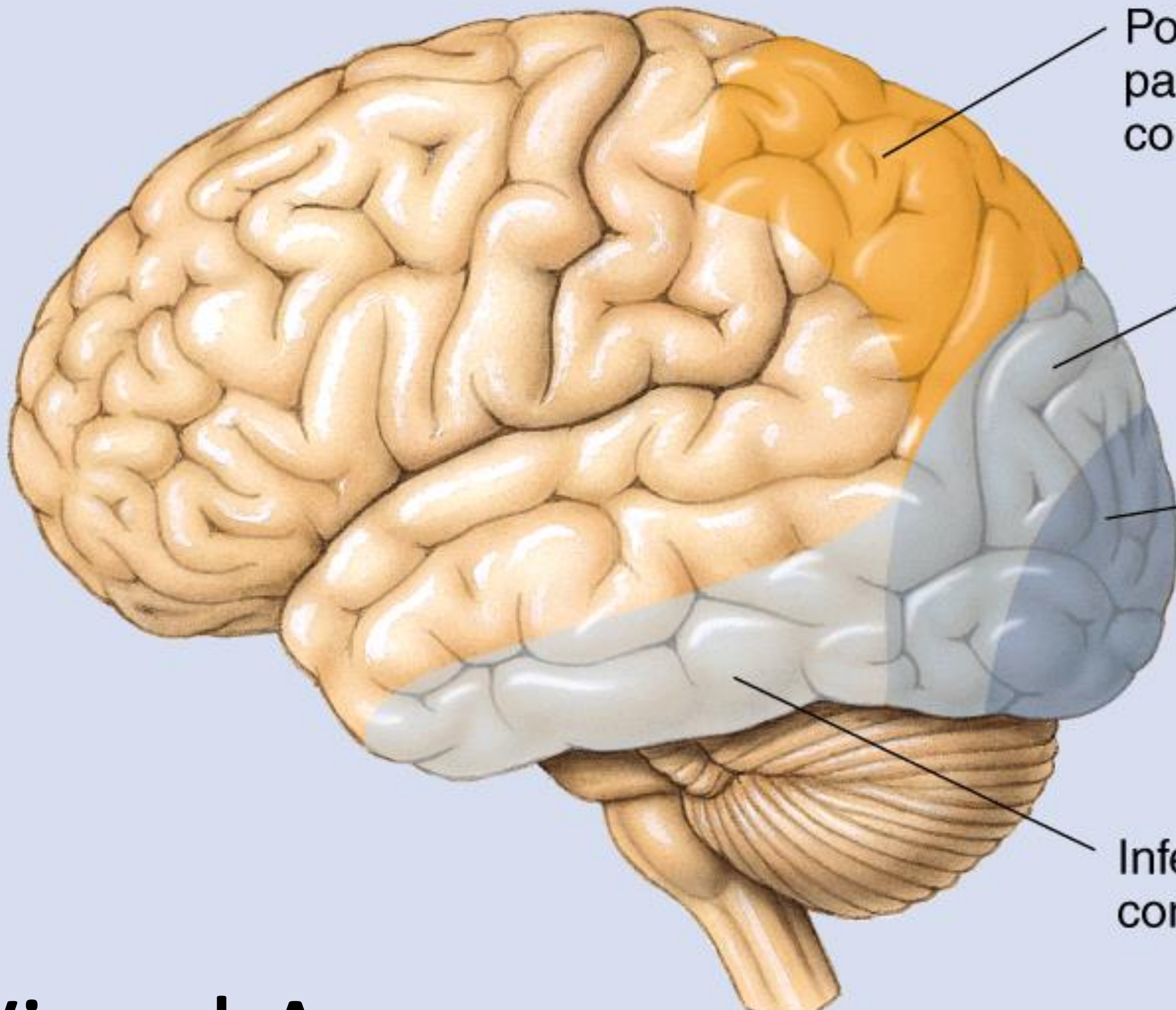
Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's ? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.

Turing (1950)
Computing Machinery
And Intelligence

Paradigms for mechanizing intelligence ~1960

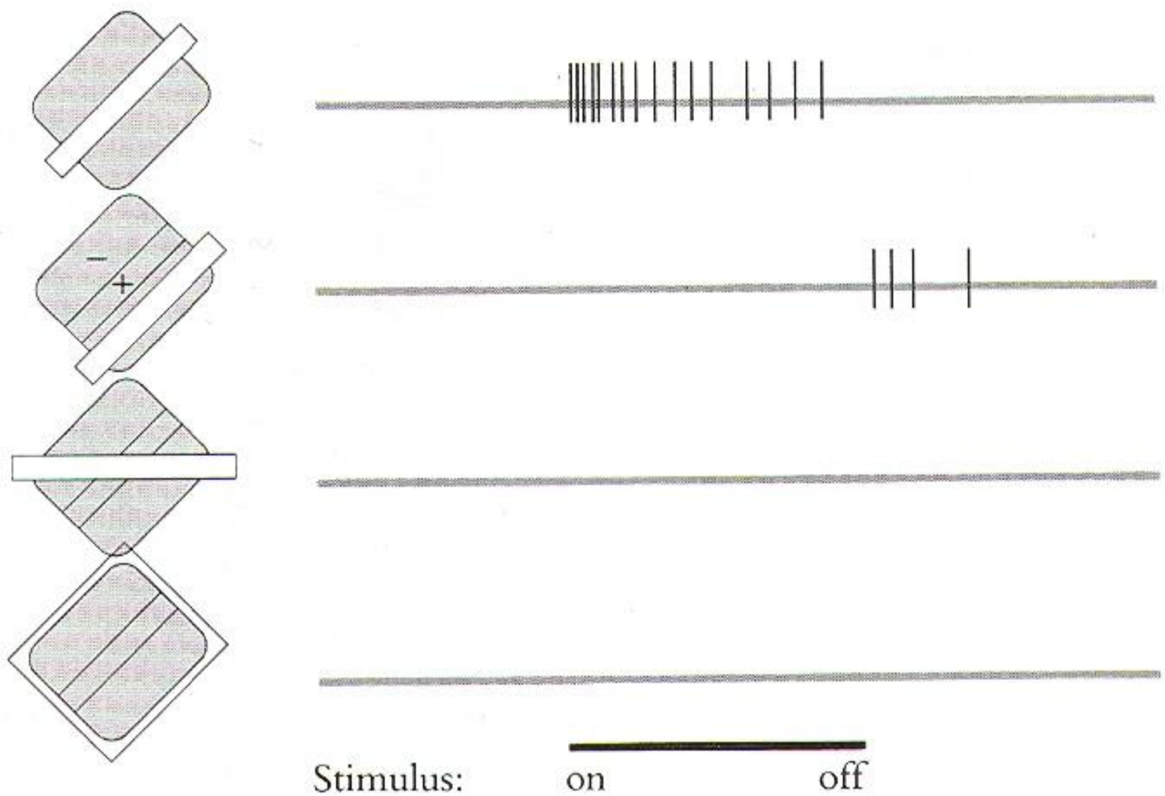
- Classic AI (McCarthy, Minsky, Newell, Simon)
  - Games, theorem-proving, reasoning
  - Search, represent and reason in first-order logic
- Pattern Recognition (Rosenblatt, Widrow)
  - Classification, Associative memory
  - Learning (Perceptrons …)
- Estimation and Control (Bellman, Kalman)
  - Decide action in uncertain, time-varying environment
  - Markov Decision Processes, adaptive control …

► **Visual Areas of the Human Cerebral Cortex**



Visual Areas

Hubel and Wiesel (1962) discovered orientation sensitive neurons in V1



Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position

Kunihiko Fukushima

NHK Broadcasting Science Research Laboratories, Kinuta, Setagaya, Tokyo, Japan

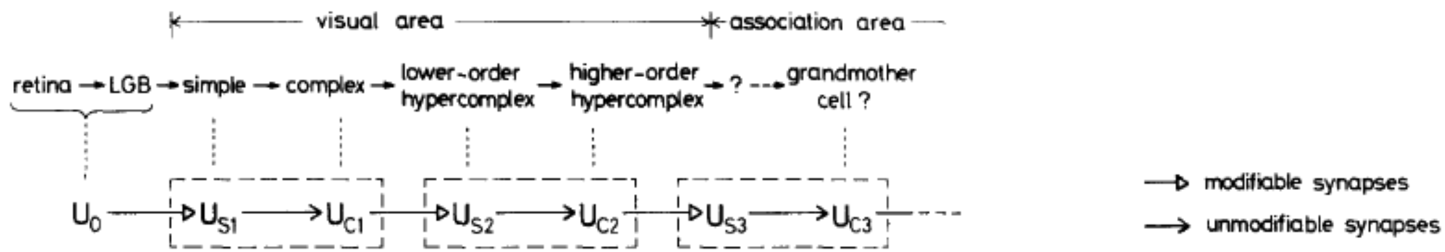


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

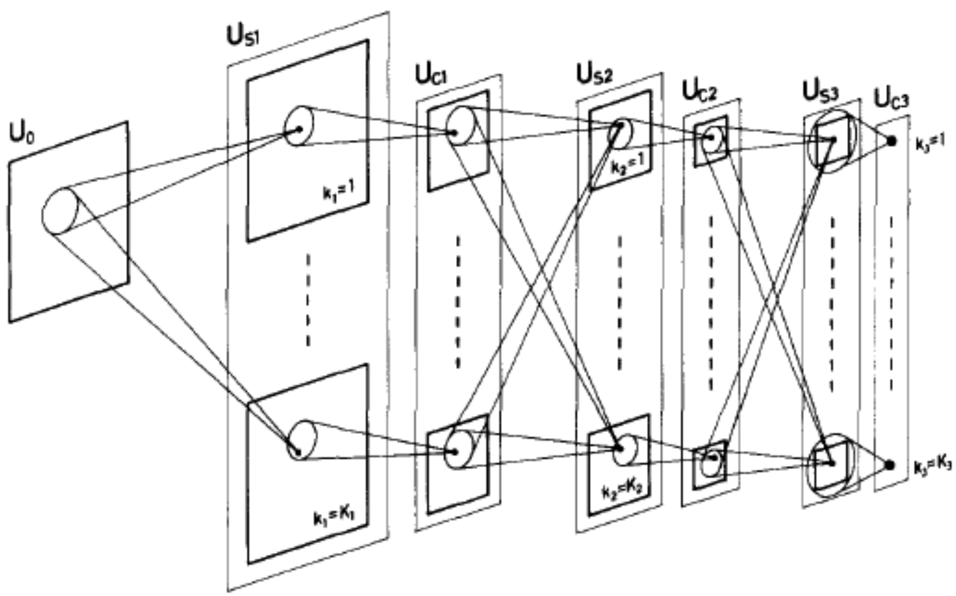


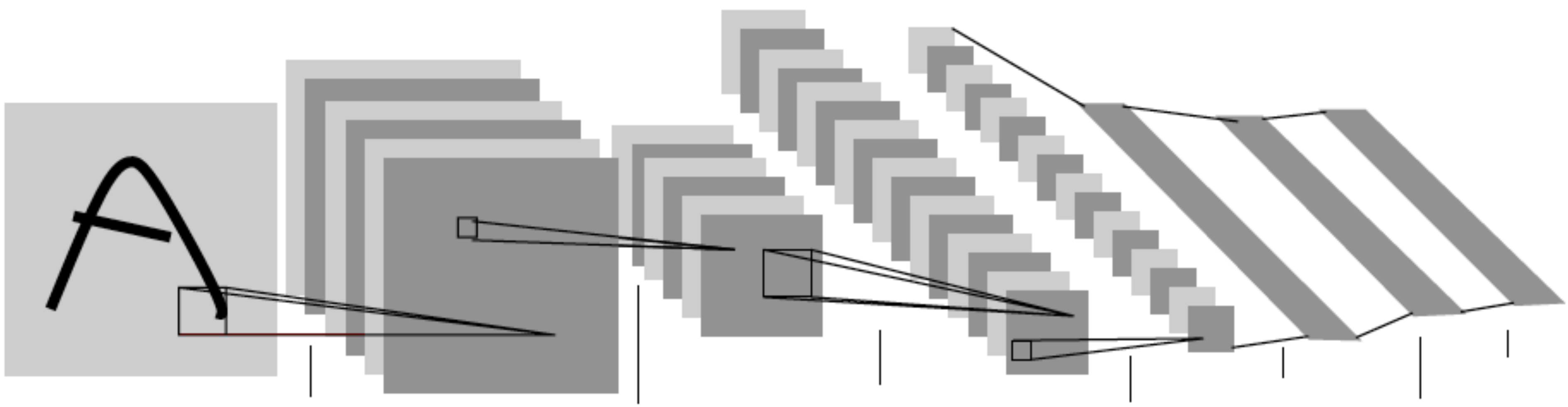
Fig. 2. Schematic diagram illustrating the interconnections between layers in the neocognitron

Biol. Cybernetics 36, 193–202 (1980)

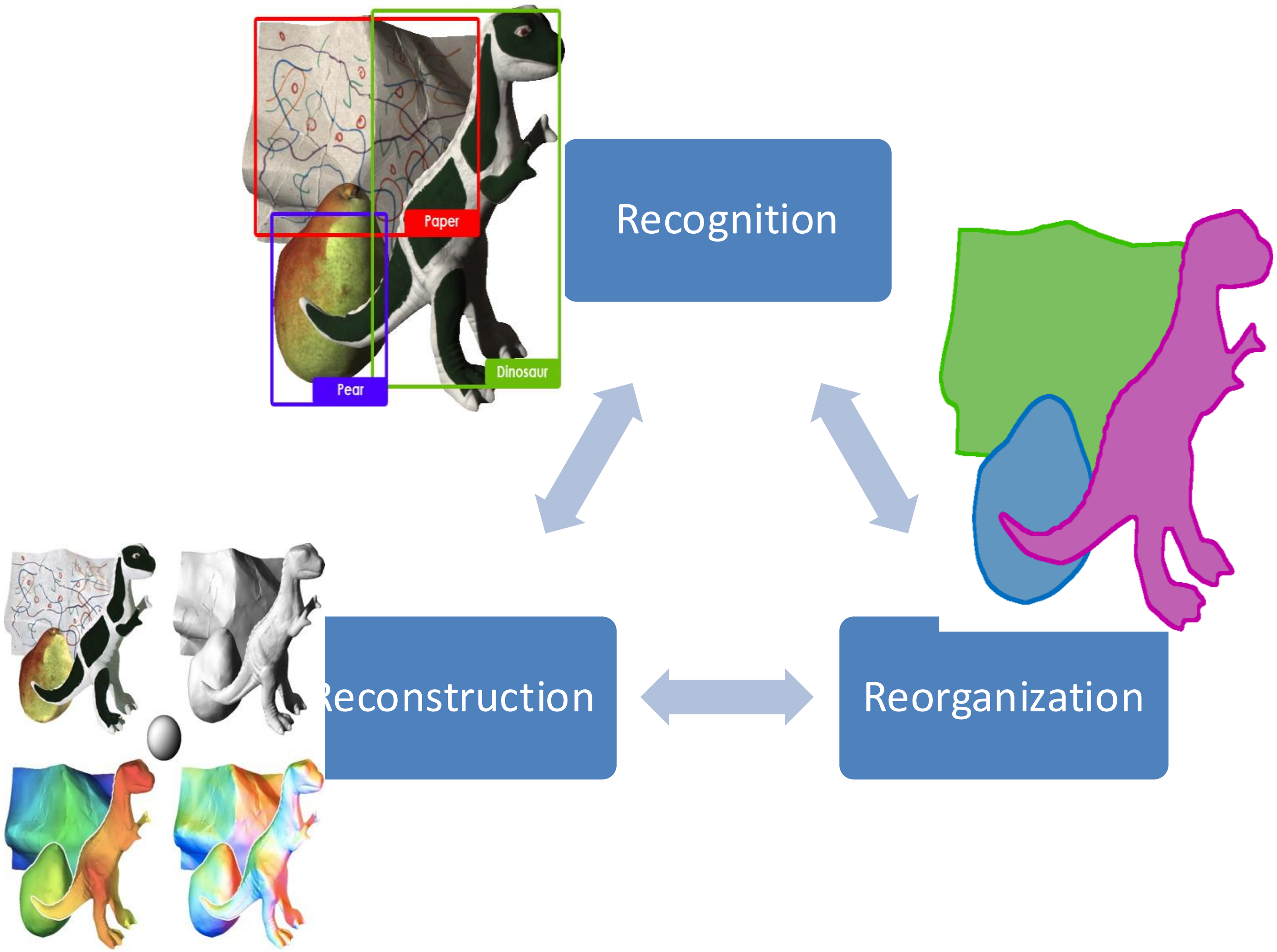
Convolutional Neural Networks (LeCun et al)

Used backpropagation to train the weights in this architecture

- First demonstrated by LeCun et al for handwritten digit recognition(1989)
- Applied in sliding window paradigm for tasks such as face detection in the 1990s.
- However was not competitive on standard computer vision object detection benchmarks in the 2000s.
- Thanks to availability of faster computing (GPUs) and large amounts of labeled data (Imagenet) we have seen an amazing renaissance led by Krizhevsky, Sutskever & Hinton (2012)

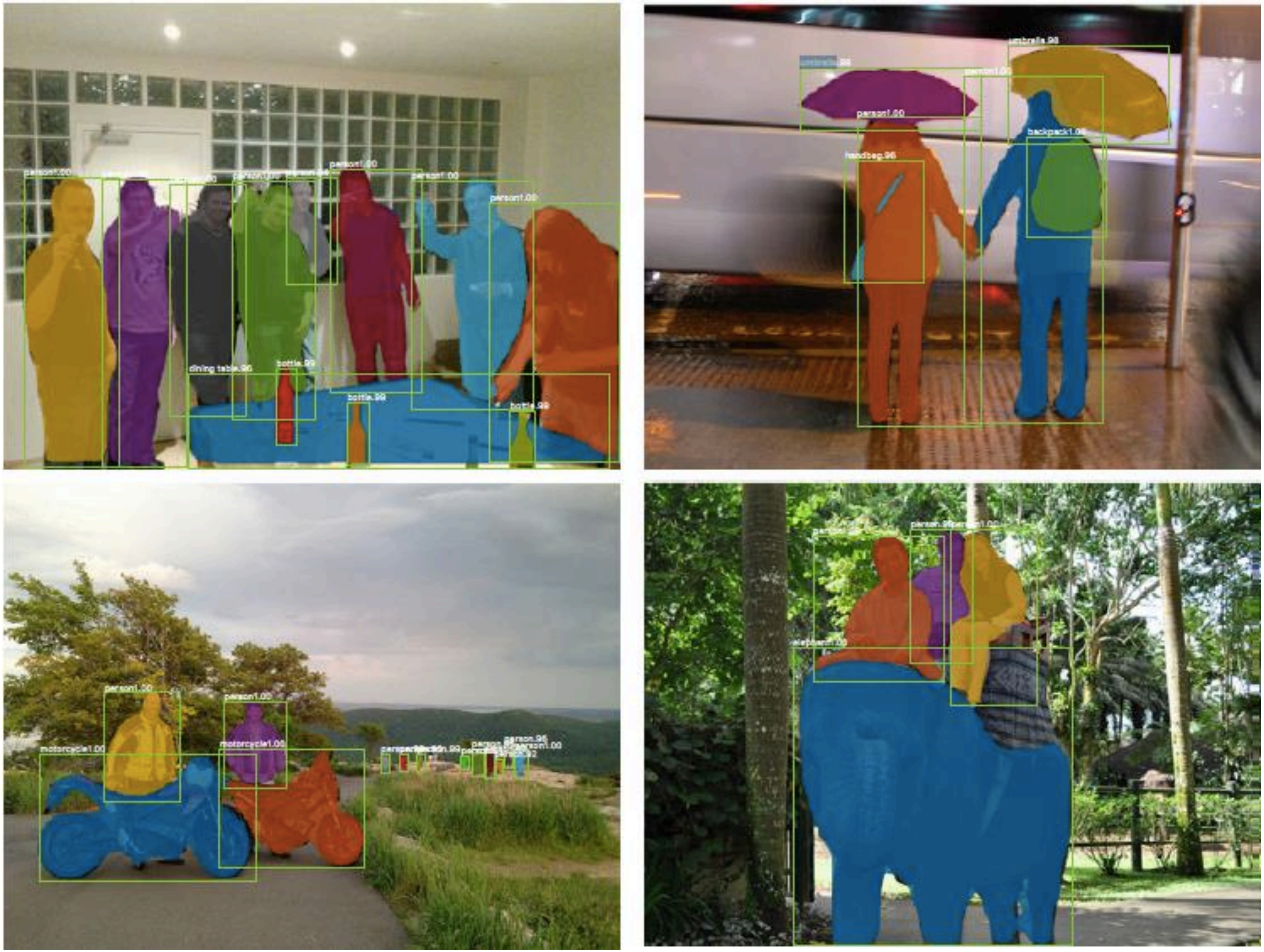


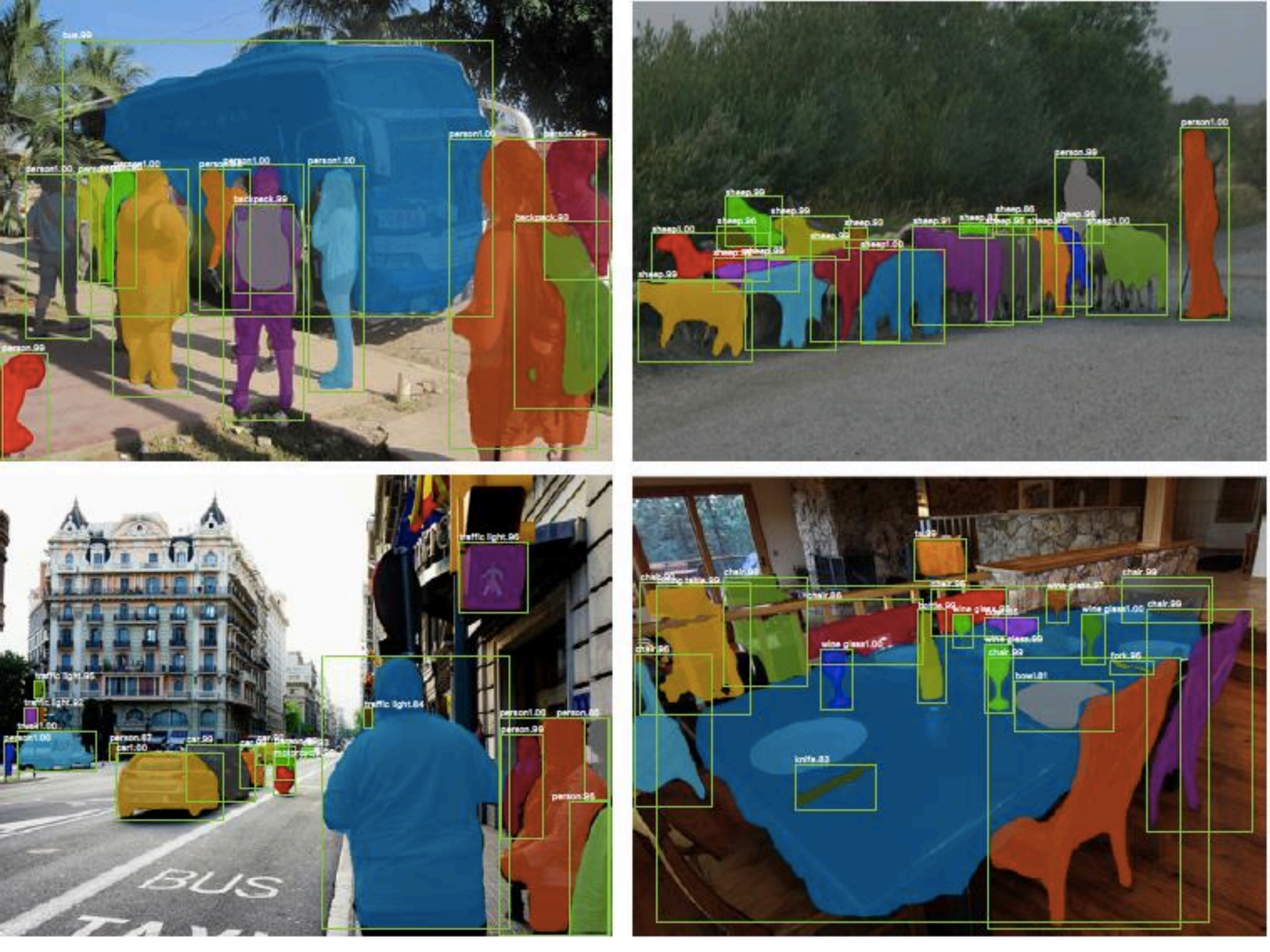
The 3R's of Vision:
Recognition, Reconstruction & Reorganization



Talk at POCV Workshop, CVPR 2012

Mask R-CNN : He, Gkioxari, Dollar & Girshick (2017)




bus.99
person1.00
backpack.99
backpack.99
person.99
person.99
sheep.99
sheep1.00
sheep.96
sheep.91
sheep.95
sheep.98
sheep1.00
person1.00
person.99
traffic light.96
traffic light.95
traffic light.92
traffic light.84
truck1.00
person1.00
car.99
car1.00
person1.00
person.99
person.95
BUS
tv.99
chair.99
chair.96
dining table.99
bottle.99
wine glass.99
wine glass.97
wine glass1.00
chair.99
fork.95
bowl.81
knife.83

SAM-1

SAM-1

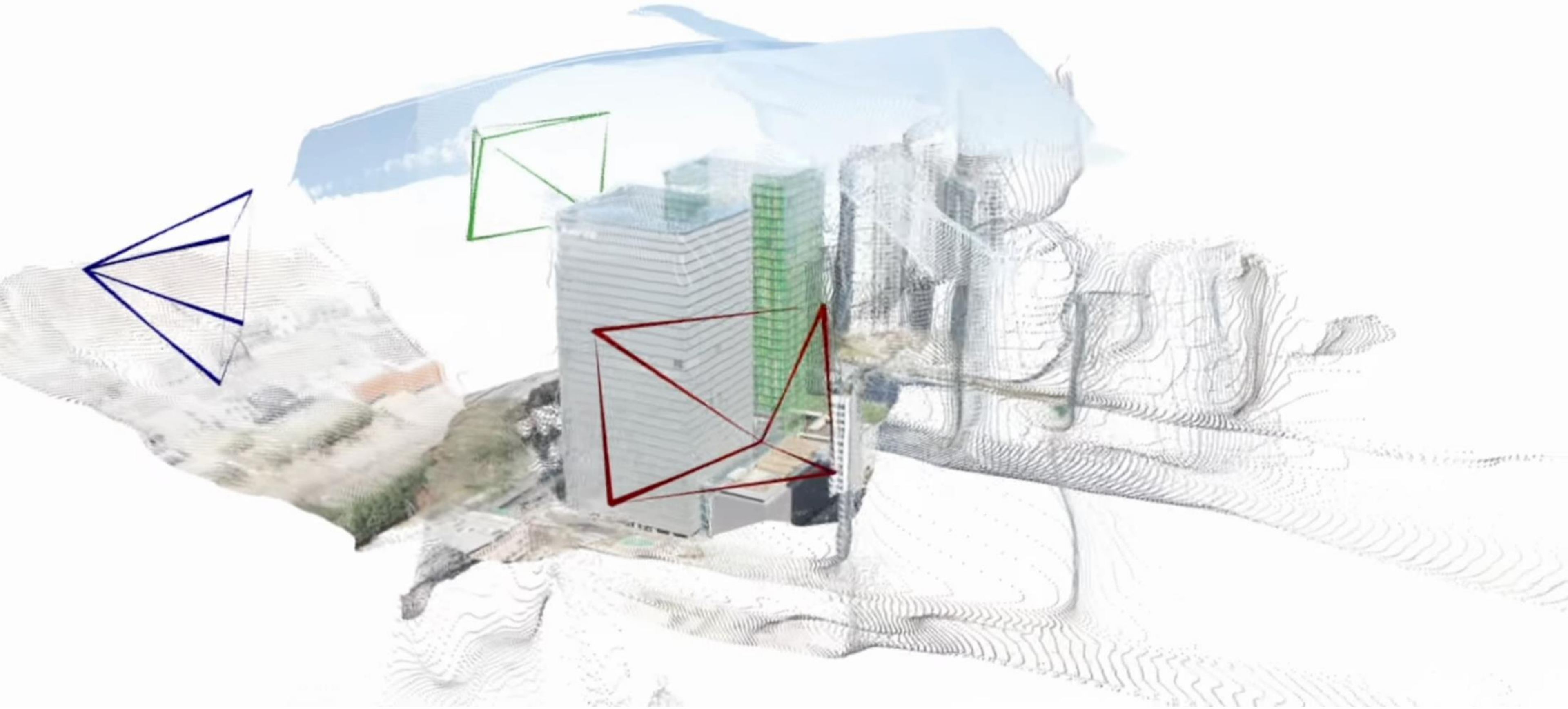
SAM-1

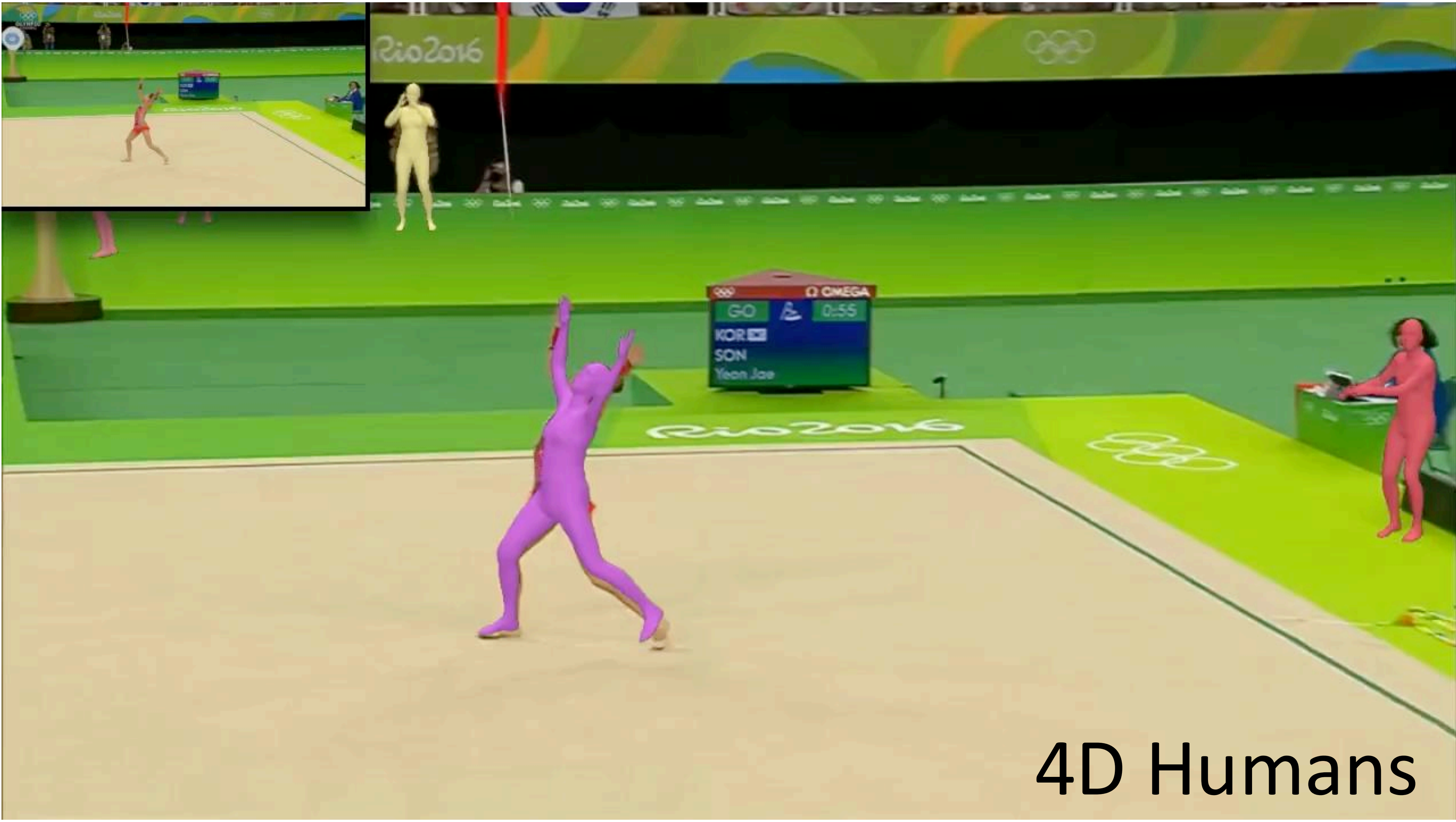
SAM-2

SAM-1

DUST3R

Output



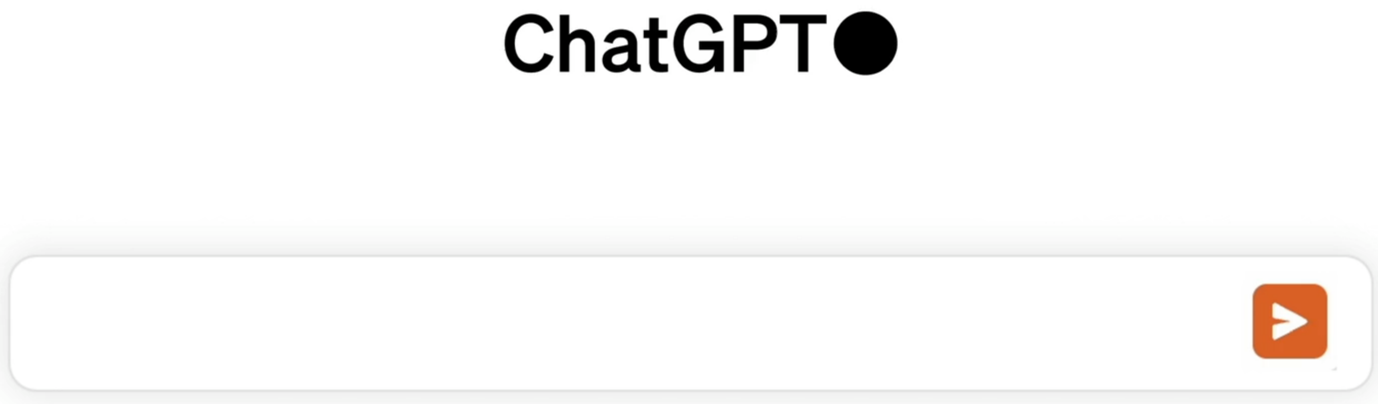


4D Humans

Gemini-2.0


Build AI agents
with Gemini 2.0
December 2024
Native audio output
Native image output
Native tool use
Spatial understanding
Video understanding
Multimodal live streaming

DALLE-3



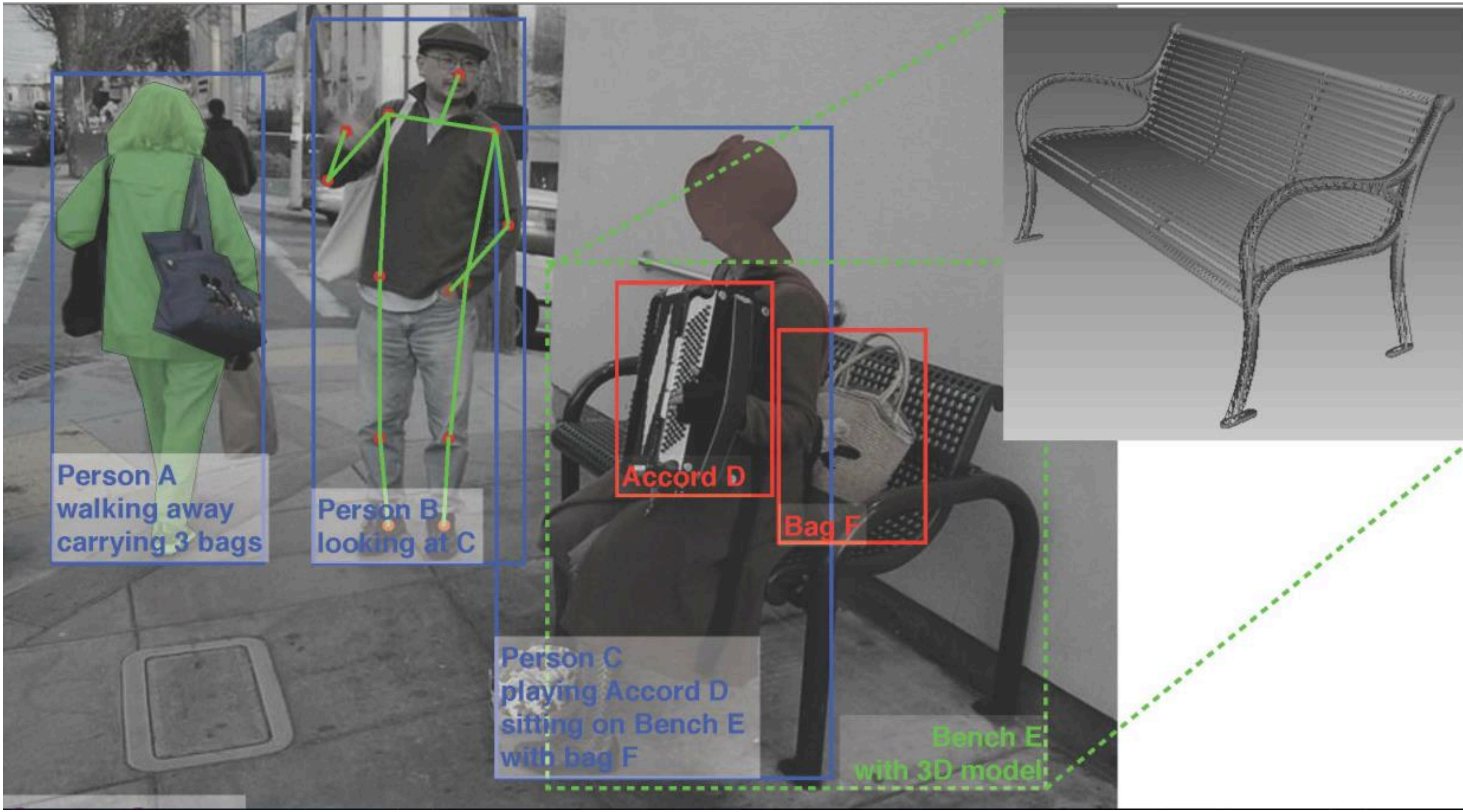

SORA

SORA

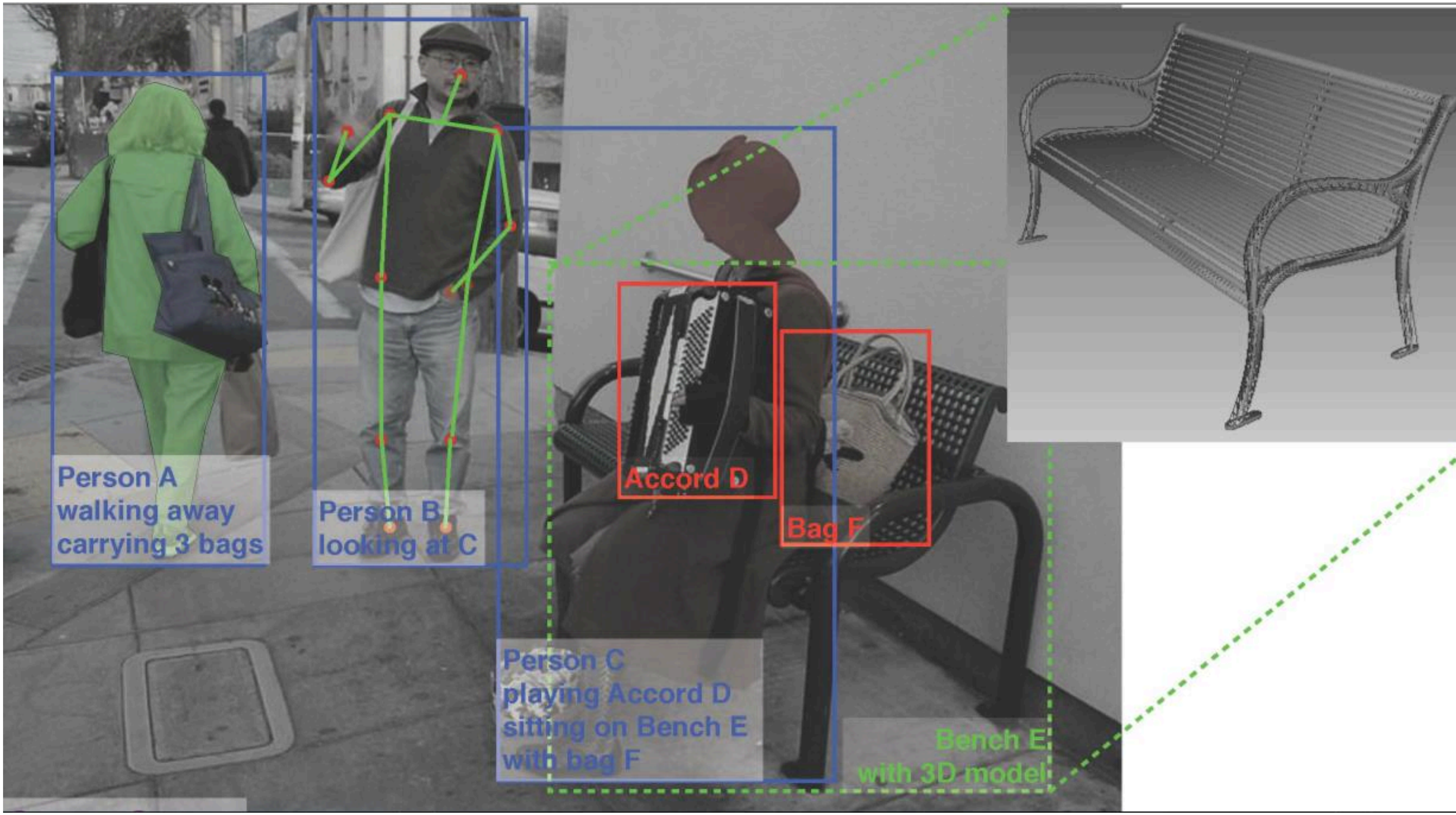
SORA


VEO2

What we can infer...

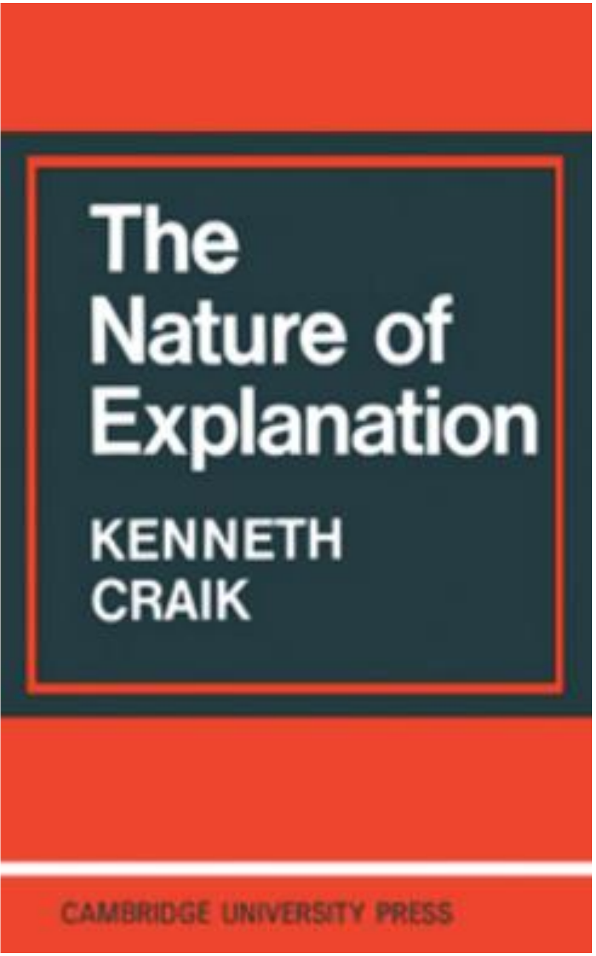


What we would like to infer…



Will person B put some money into Person C’s tip bag?

AI systems need to build "mental models"



If the organism carries a `small-scale model' of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it (Craik, 1943,Ch. 5, p.61)

Commonsense is not just facts, it is a collection of models

Where should we go next?

- Turing's Baby

Ontogeny of Intelligence

Perception and Interaction

Language

456 A. M. TURING :

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's ? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed. The amount of work in the education we can assume, as a first approximation, to be much the same as for the human child.

Turing (1950)
Computing Machinery
And Intelligence

The Development of Embodied Cognition: Six Lessons from Babies

Linda Smith & Michael Gasser

Abstract. The embodiment hypothesis is the idea that intelligence emerges in the interaction of an agent with an environment and as a result of sensorimotor activity. In this paper we offer six lessons for *developing* embodied intelligent agents suggested by research in developmental psychology. We argue that starting as a baby grounded in a physical, social and linguistic world is crucial to the development of the flexible and inventive intelligence that characterizes humankind.

The Six Lessons

- Be multi-modal
- Be incremental
- Be physical
- Explore
- Be social
- Use language

- I think this provides the right structure for viewing the stages of inbuilt, supervised by observation, supervised by interaction, supervised by culture

We can only see a short distance ahead, but we can see plenty there that needs to be done.

-Alan Turing

Fundamentals of Image Formation

Jitendra Malik

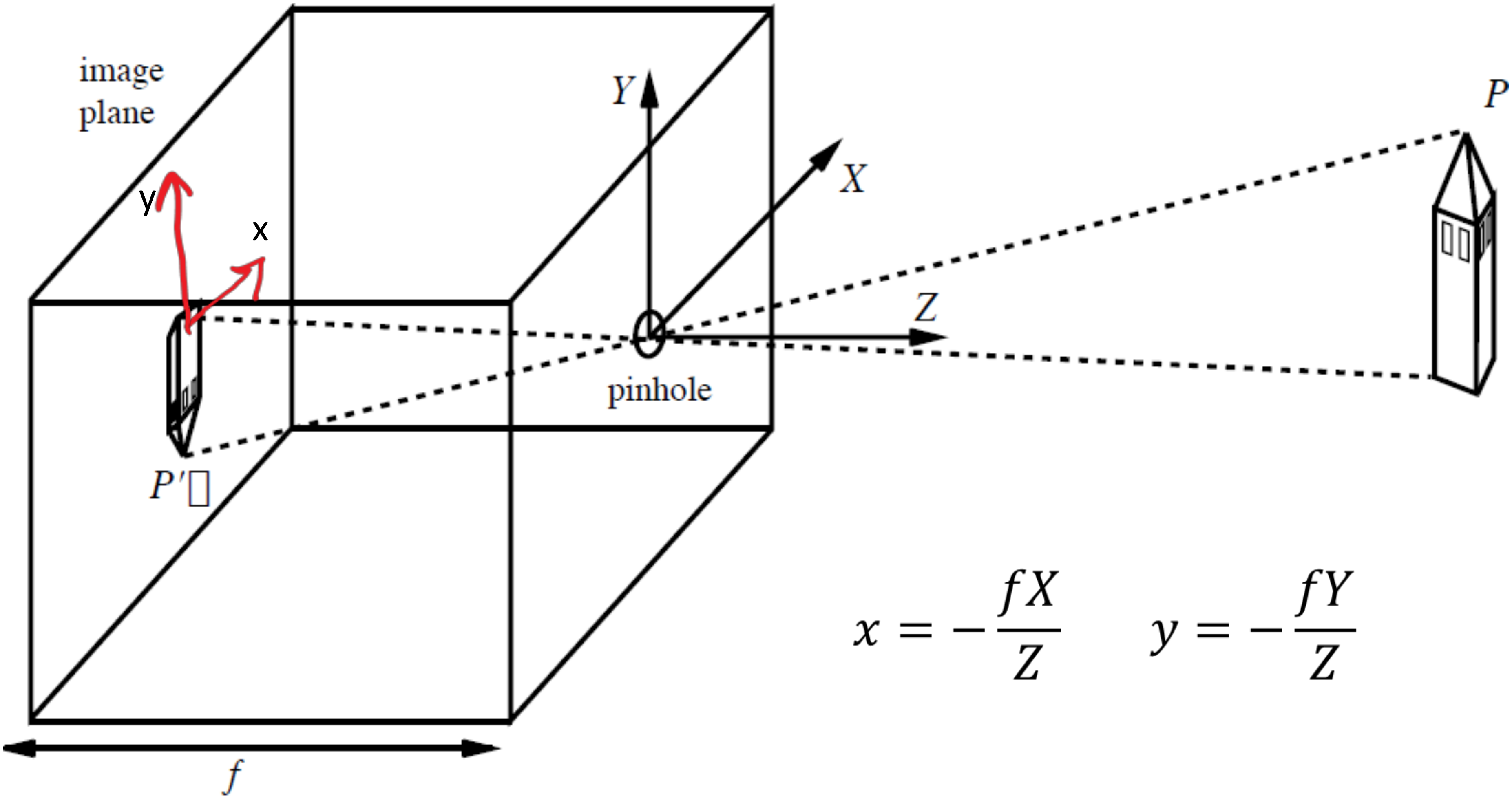
A camera creates an image …

The image I(x,y) measures how much light is captured at pixel (x,y)

We want to know

- Where does a point (X,Y,Z) in the world get imaged?
- What is the brightness at the resulting point (x,y)?

The Pinhole Camera

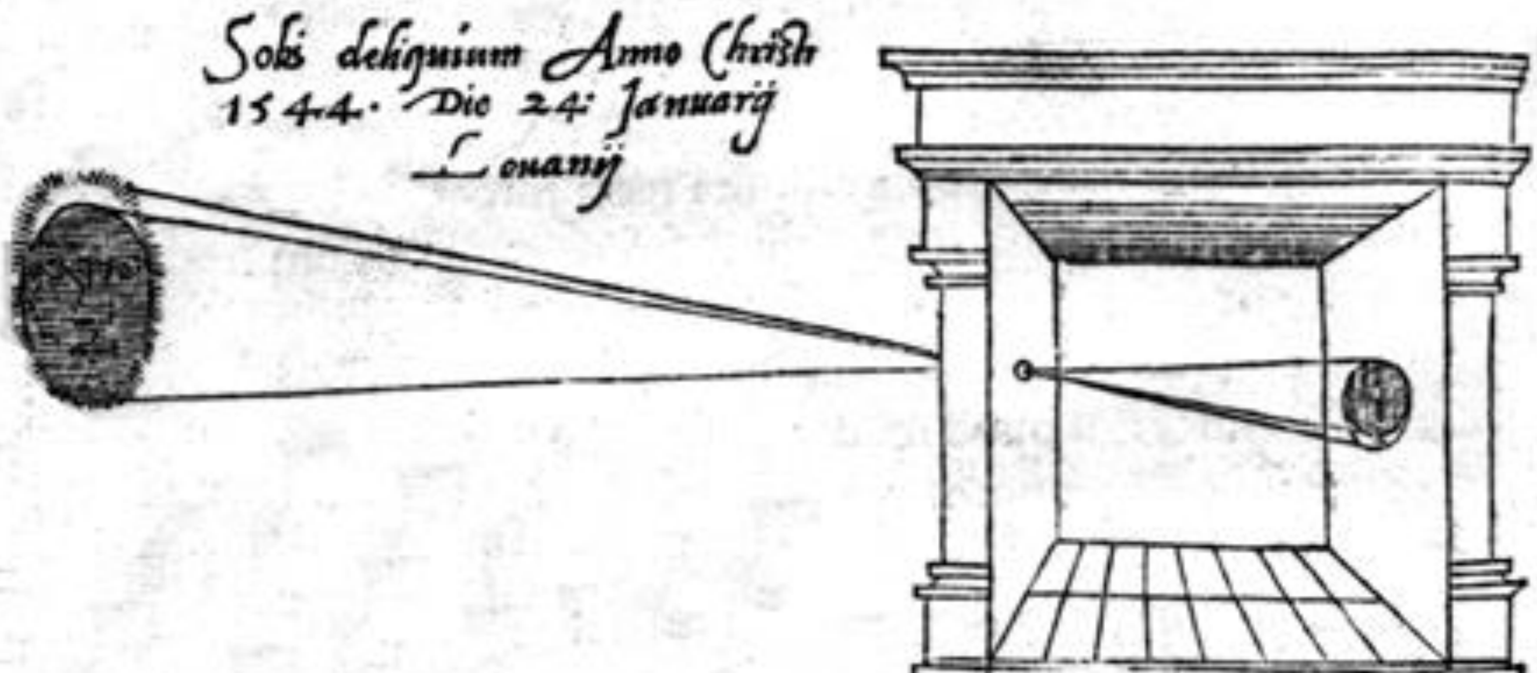


$$x = -\frac{fX}{Z} \qquad y = -\frac{fY}{Z}$$

Camera Obscura

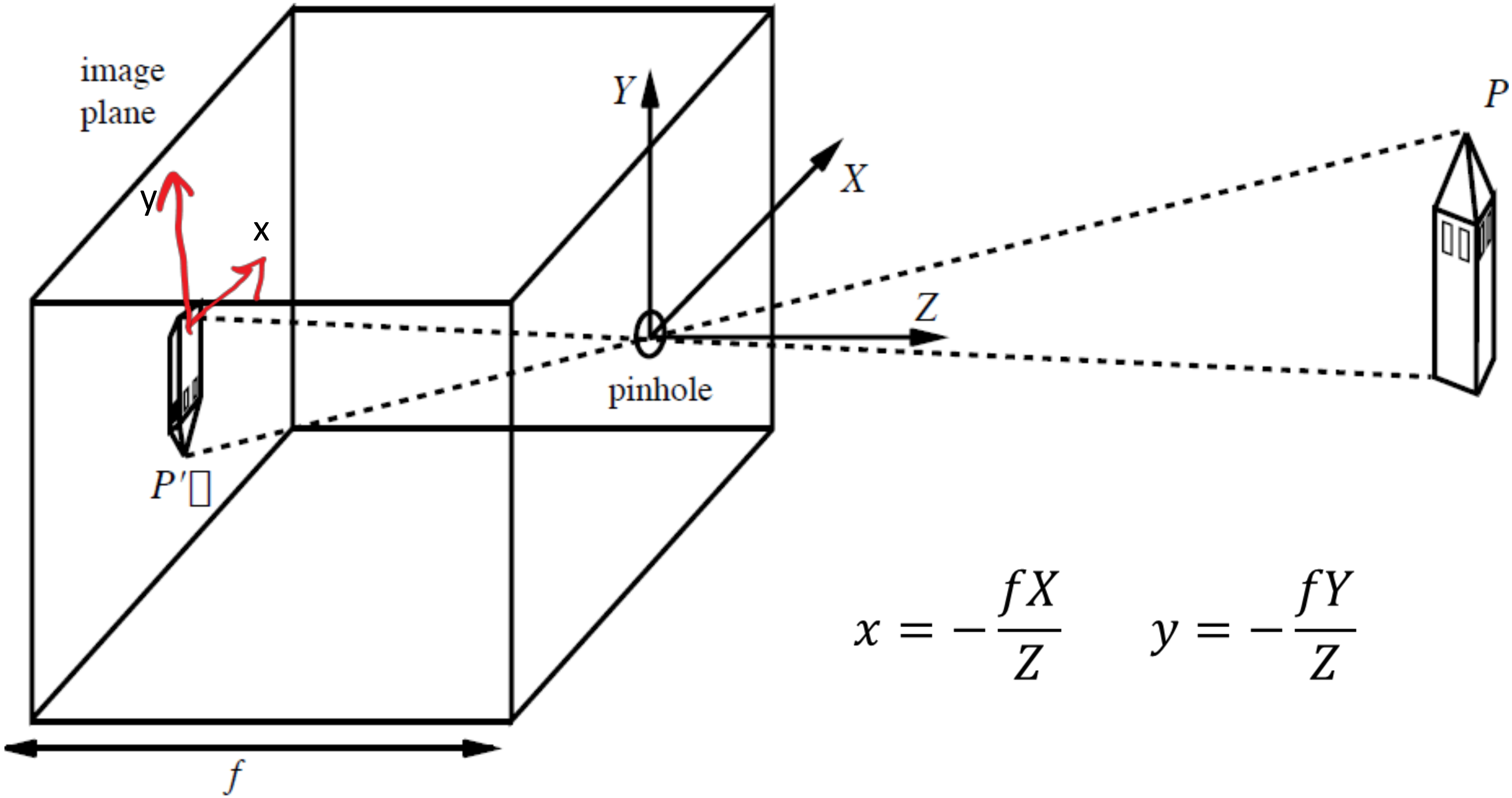
(Reinerus Gemma-Frisius, 1544)

illum in tabula per radios Solis, quàm in cœlo contin-
git: hoc eſt, ſi in cœlo ſuperior pars deliquiũ patiatur, in
radiis apparebit inferior deficere, vt ratio exigit optica.



Sic nos exactè Anno .1544. Louanii eclipſim Solis
obſeruauimus, inuenimusq; deficere paulò plus q̃ dex-

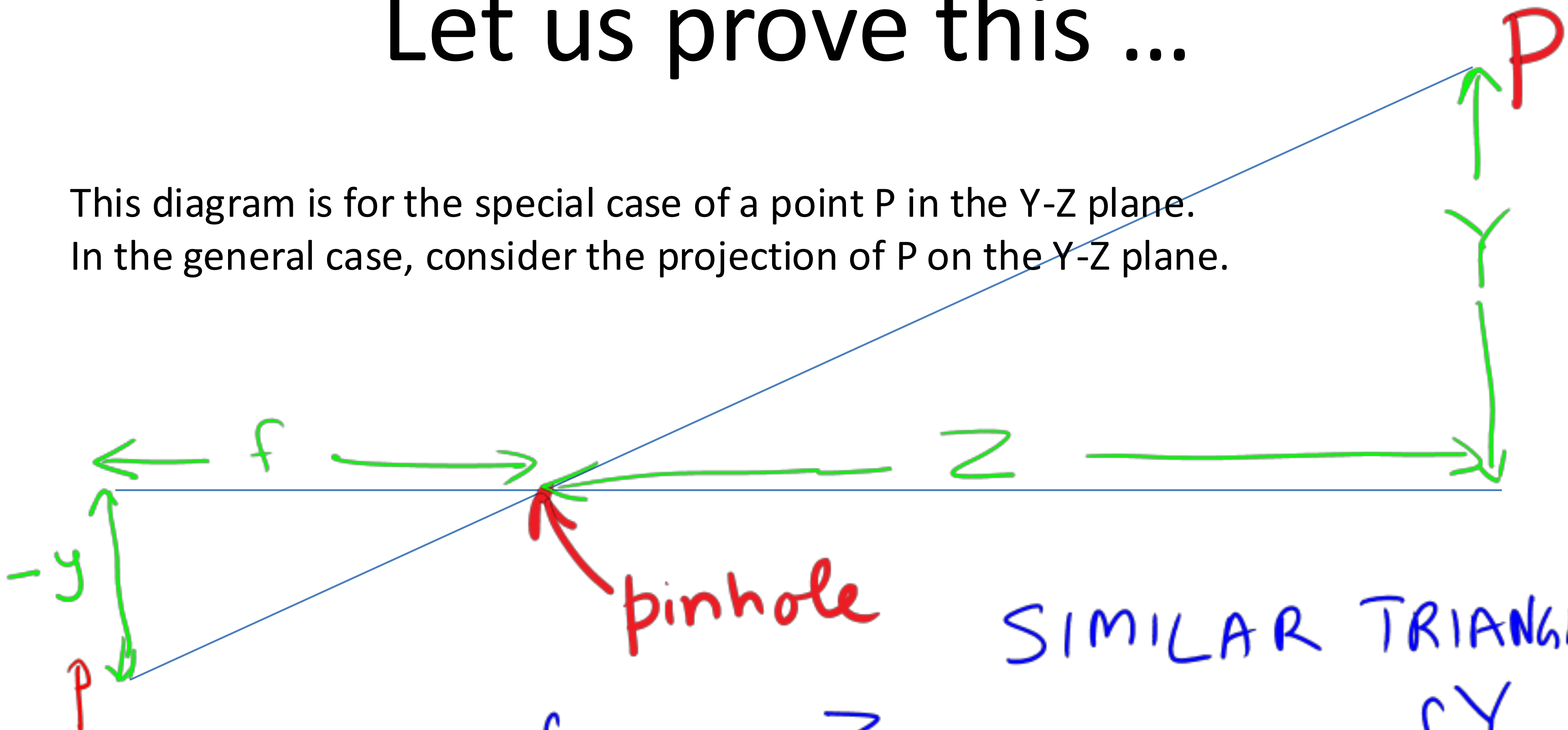
The Pinhole Camera



$$x = -\frac{fX}{Z} \qquad y = -\frac{fY}{Z}$$

Let us prove this …

This diagram is for the special case of a point P in the Y-Z plane.
In the general case, consider the projection of P on the Y-Z plane.



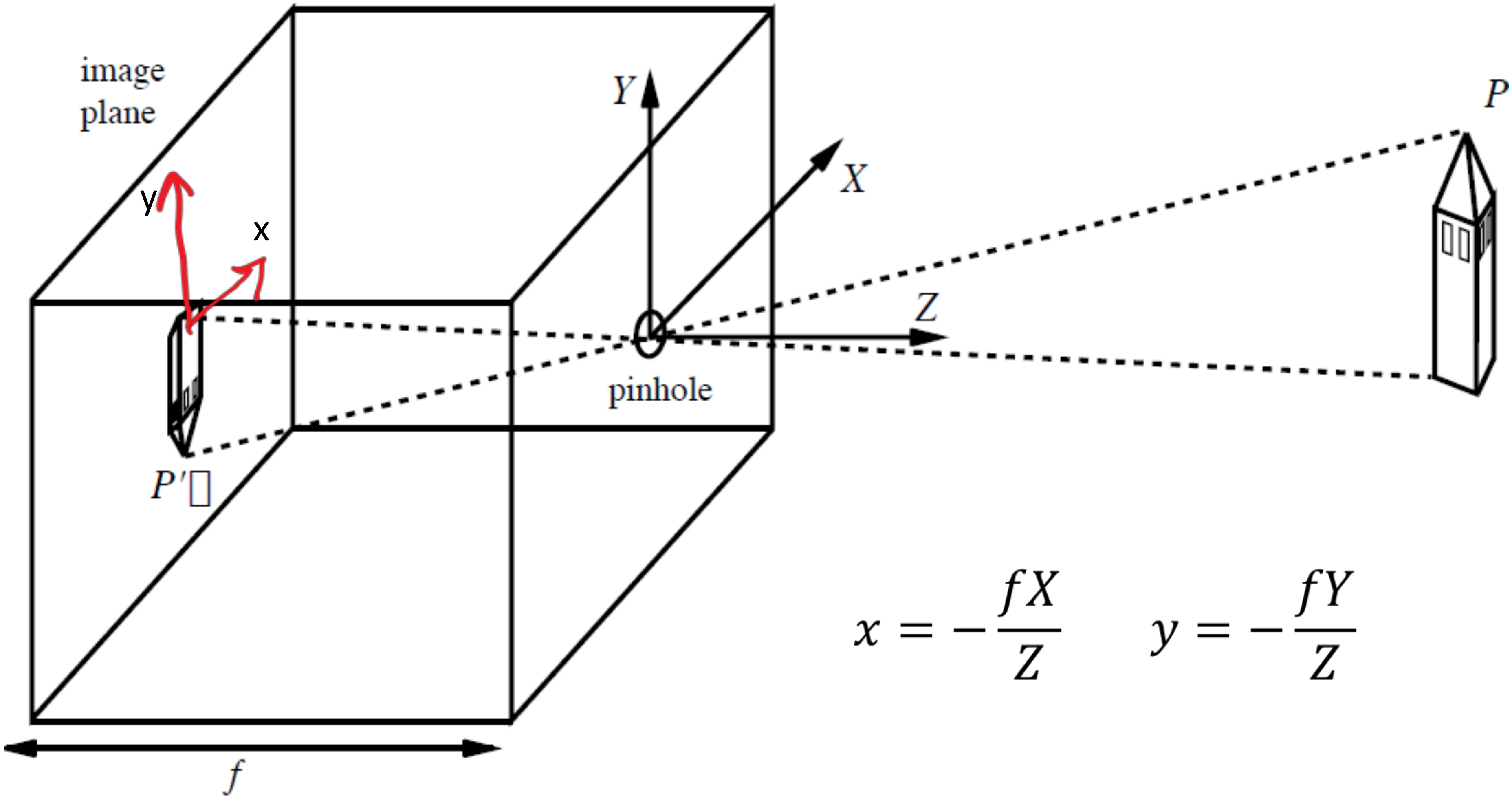
SIMILAR TRIANGLES

$$\frac{f}{-y} = \frac{Z}{Y} \Rightarrow y = \frac{-fY}{Z}$$

This is true even if the point P is not in the YZ plane.

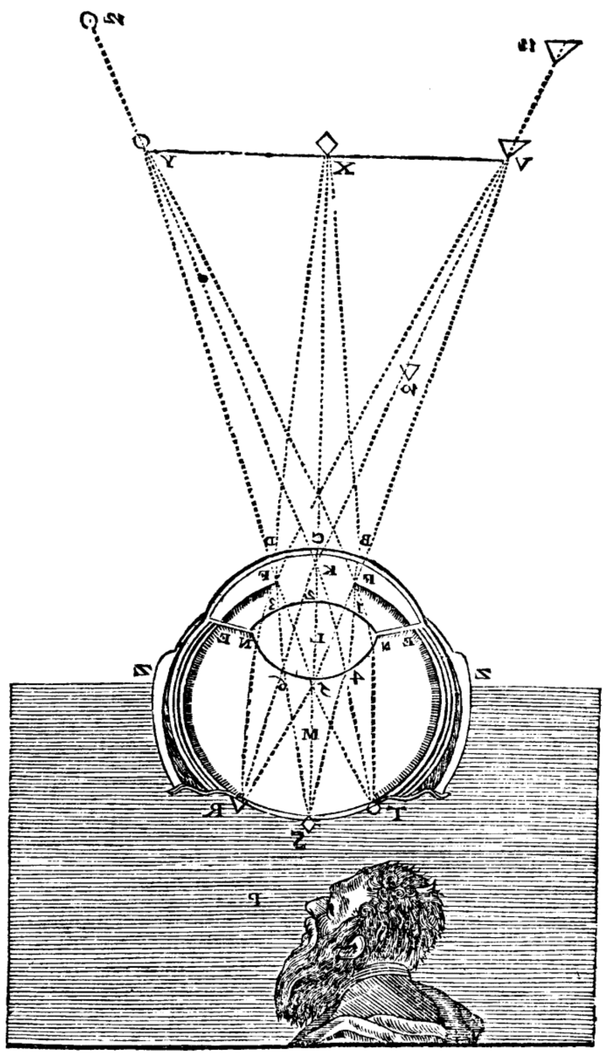
By similar reasoning $x = \frac{-fX}{Z}$

The Pinhole Camera



$$x = -\frac{fX}{Z} \qquad y = -\frac{fY}{Z}$$

The image is inverted



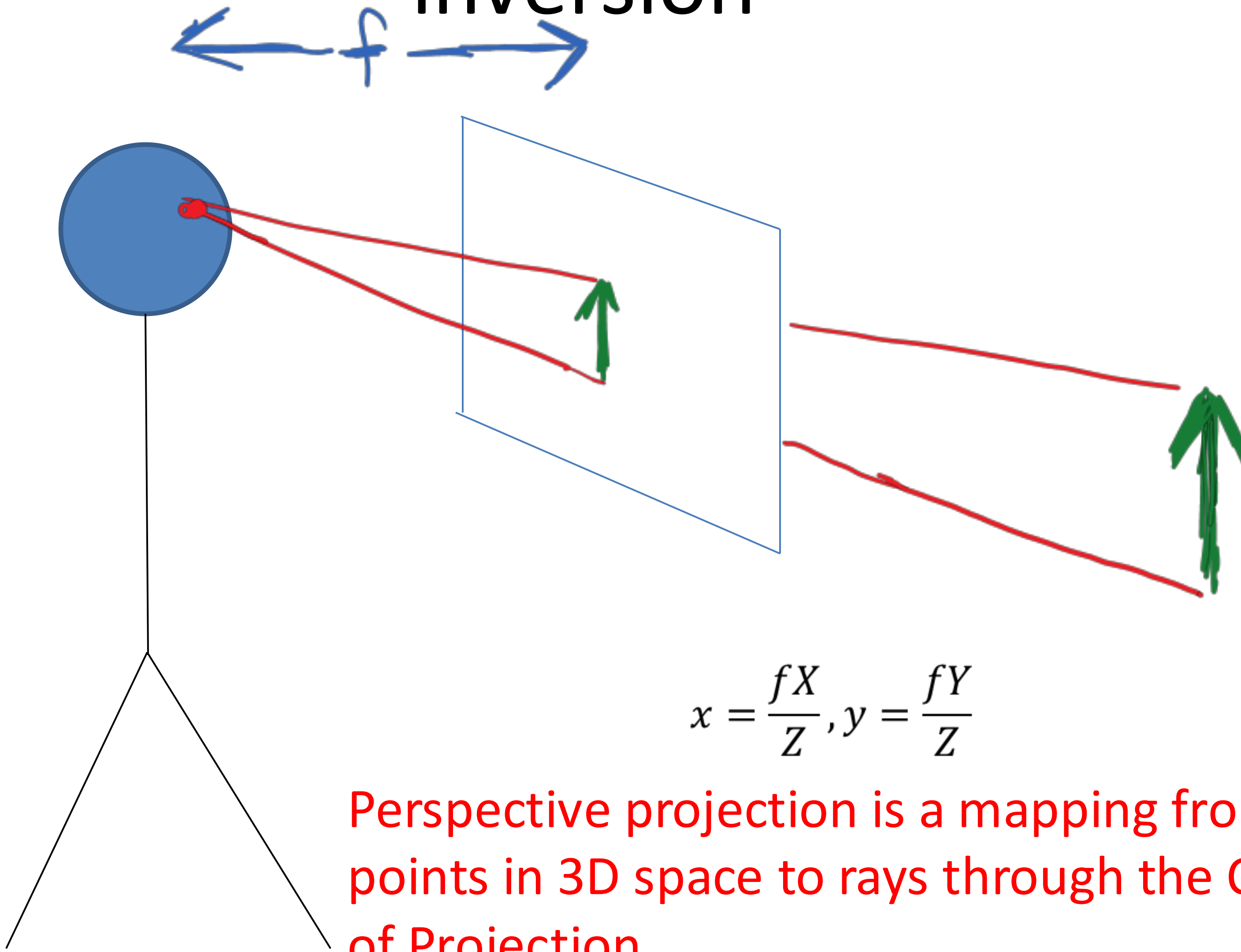
From Descartes(1637), La Dioptrique

This was pointed out by Kepler in 1604

But this is no big deal. The brain can interpret it the right way. And for a camera, software can simply flip the image top-down and right-left. After this trick, we get

$$x = \frac{fX}{Z} \qquad y = \frac{fY}{Z}$$

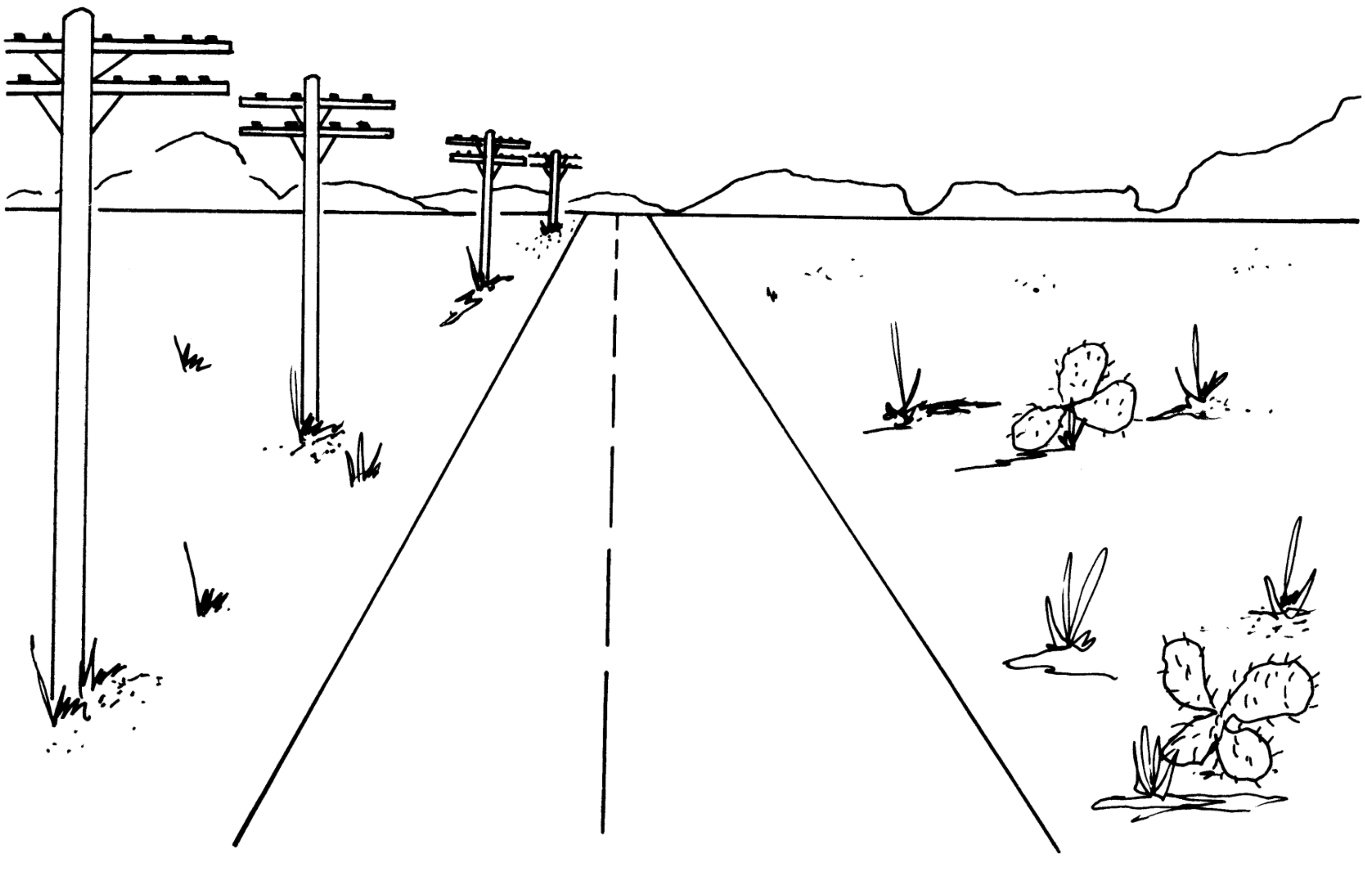
A projection model that avoids inversion



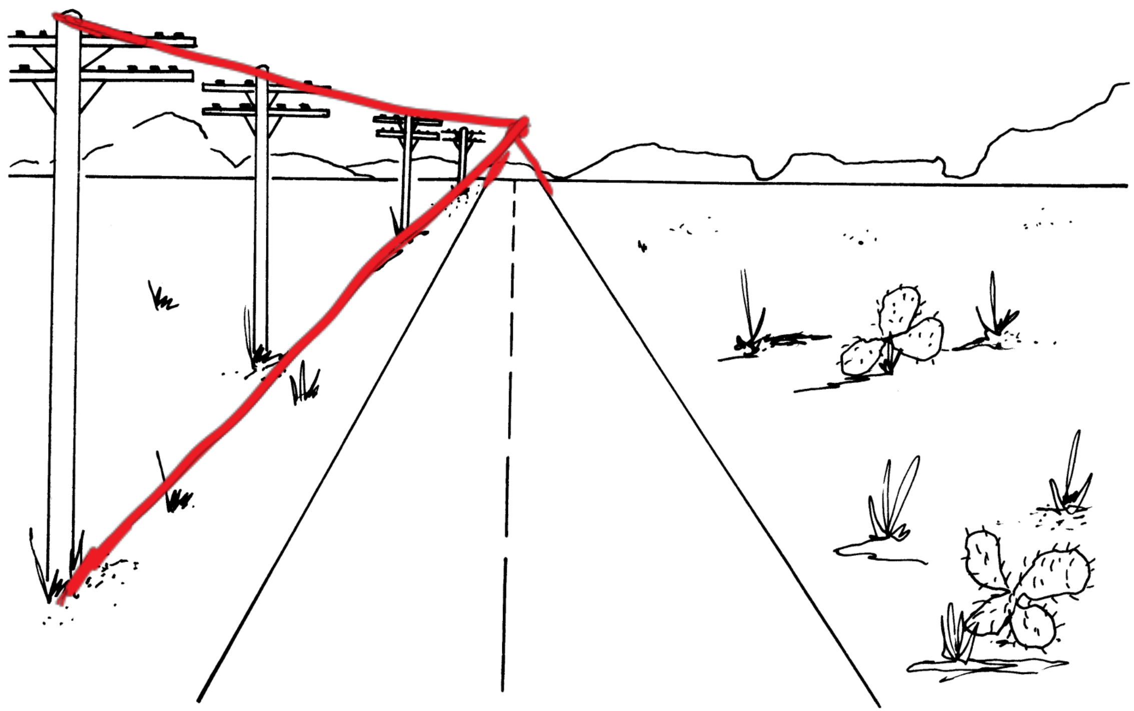
$$x = \frac{fX}{Z}, y = \frac{fY}{Z}$$

Perspective projection is a mapping from points in 3D space to rays through the Center of Projection

Some perspective phenomena…

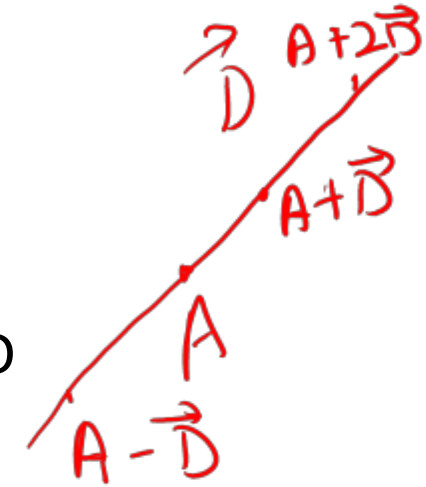


Parallel lines converge to a vanishing point



Each family of parallel lines has its own vanishing point

Proof



Let there be a point A and a direction vector D in three dimensional space.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} A_x \\ A_y \\ A_z \end{bmatrix} + \lambda \begin{bmatrix} D_x \\ D_y \\ D_z \end{bmatrix} \quad -\infty \leftarrow \lambda \rightarrow \infty$$

$$x = \frac{fX}{Z} = \frac{f(A_x + \lambda D_x)}{A_z + \lambda D_z}$$

Let us consider $\lambda \rightarrow \infty$

$$x = \frac{f \lambda D_x}{\lambda D_z} = \frac{f D_x}{D_z}$$

This expression does not depend on A

Coordinates of the projected point are

$$\frac{f D_x}{D_z}$$ for the x-coordinate

(and by doing the same process for y-coordinate

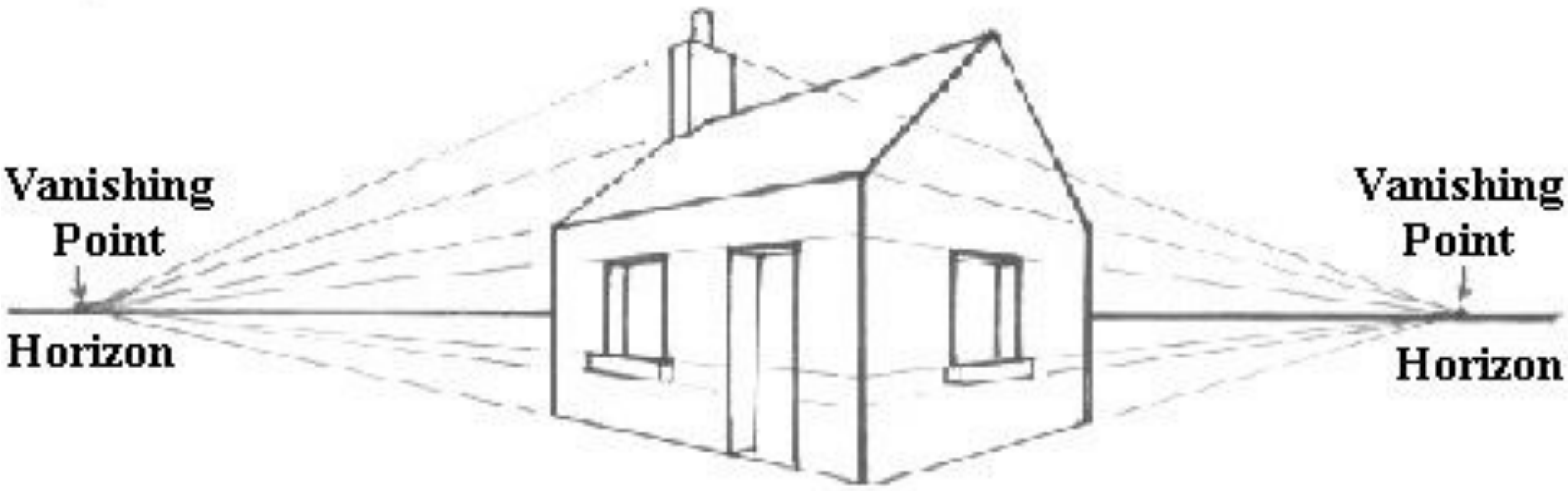
$$\frac{f D_y}{D_z}$$ for the y-coordinate.

Thus $\left(\frac{f D_x}{D_z}, \frac{f D_y}{D_z} \right)$ are

the coordinates of the vanishing point

..

Each family of parallel lines has its own vanishing point



But this isn't true of the vertical lines. They stay parallel. Why?

Vanishing point in vector notation

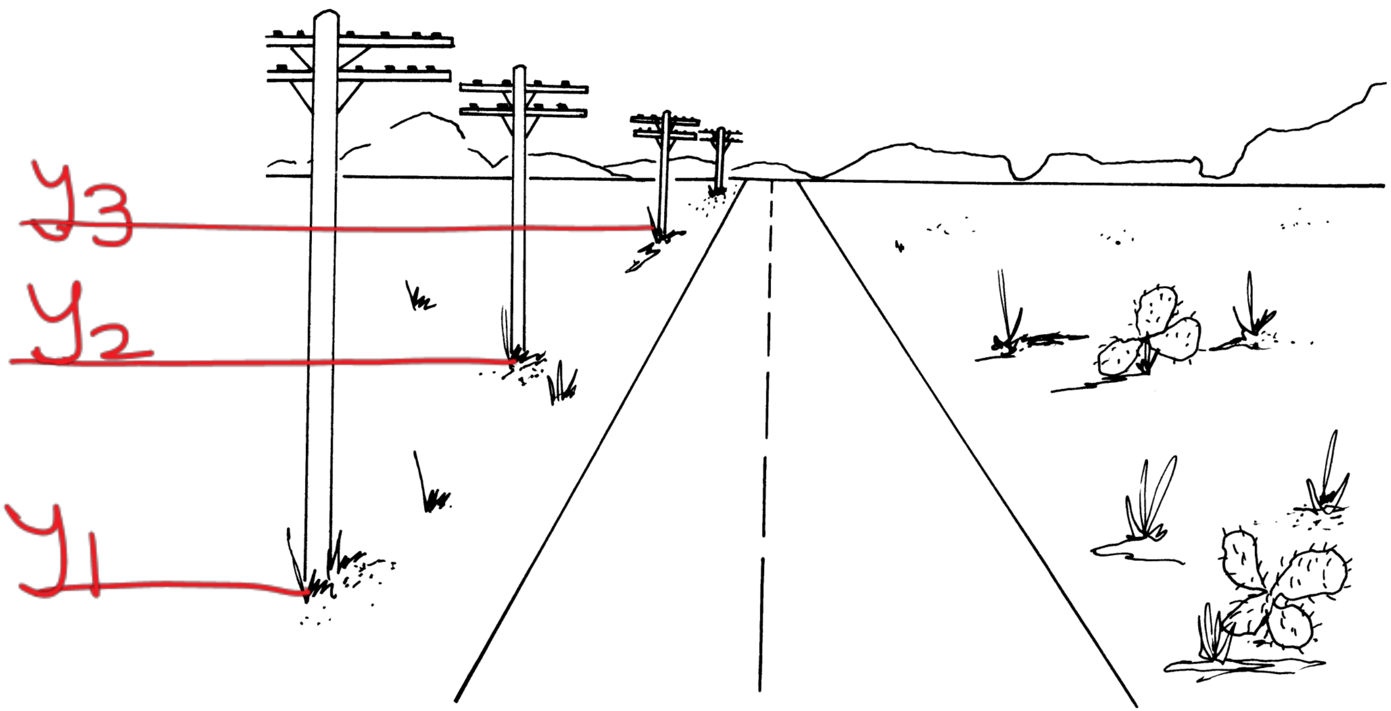
$$\mathbf{p} = f\frac{\mathbf{X}}{Z}$$

A line of points in 3D can be represented as $\mathbf{X} = \mathbf{A} + \lambda\mathbf{D}$, where $\mathbf{A}$ is a fixed point, $\mathbf{D}$ a unit vector parallel to the line, and λ a measure of distance along the line. As λ increases points are increasingly further away and in the limit:

$$\lim_{\lambda\to\infty} \mathbf{p} = f\frac{\mathbf{A} + \lambda\mathbf{D}}{A_Z + \lambda D_Z} = f\frac{\mathbf{D}}{D_Z}$$

i.e. the image of the line terminates in a *vanishing point* with coordinates $(fD_X/D_Z, fD_Y/D_Z)$, unless the line is parallel to the image plane ($D_Z = 0$). Note, the vanishing point is unaffected (invariant to) line position, $\mathbf{A}$, it only depends on line orientation, $\mathbf{D}$. Consequently, the family of lines parallel to $\mathbf{D}$ have the same vanishing point.

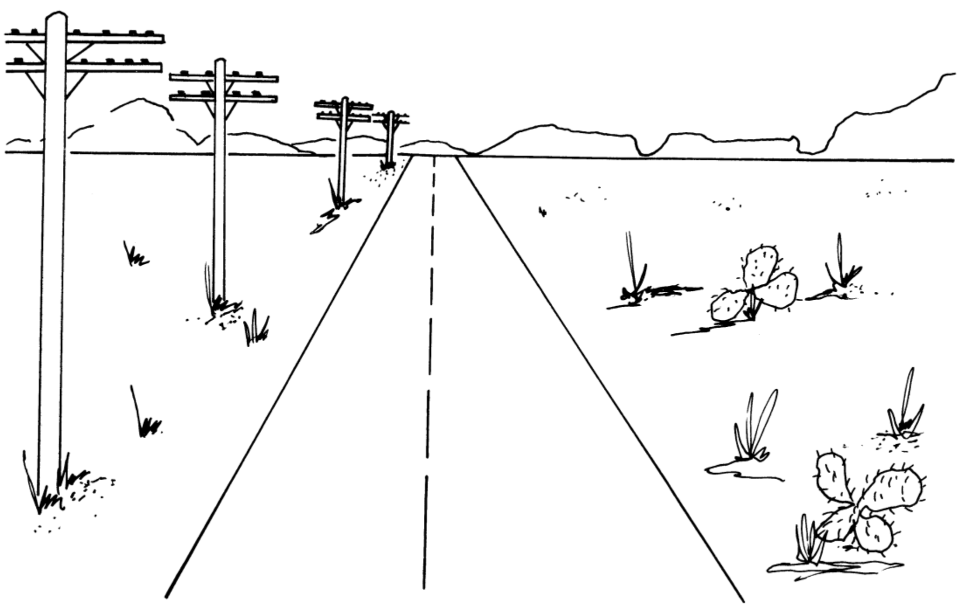
Nearer objects are lower in the image



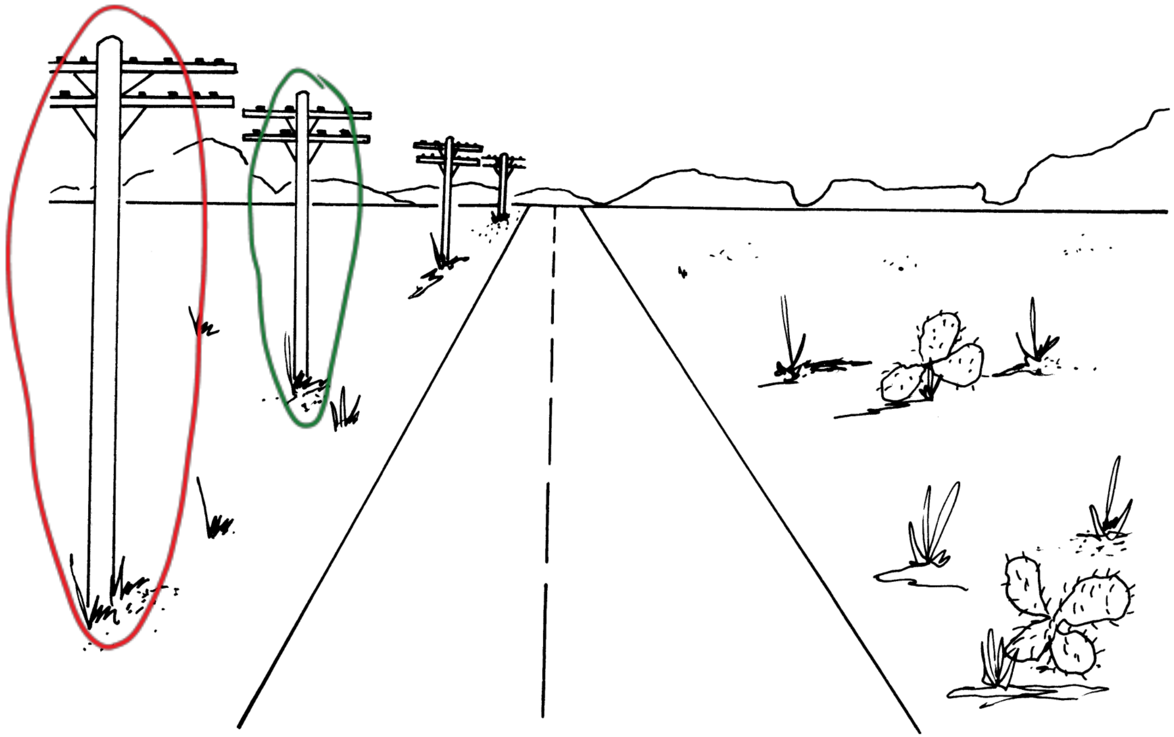
Proof

The equation of the ground plane is $Y = -h$

A point on the ground plane will have y-coordinate $y = -fh/Z$

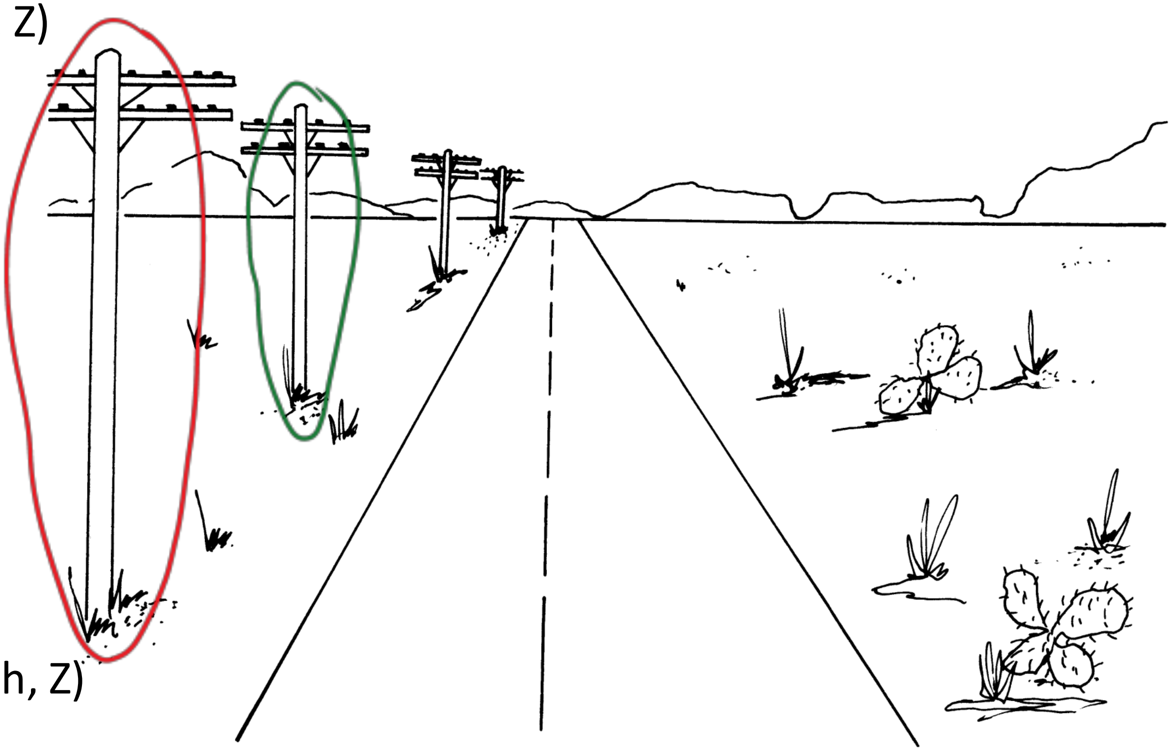


Nearer objects look bigger



Nearer objects look bigger

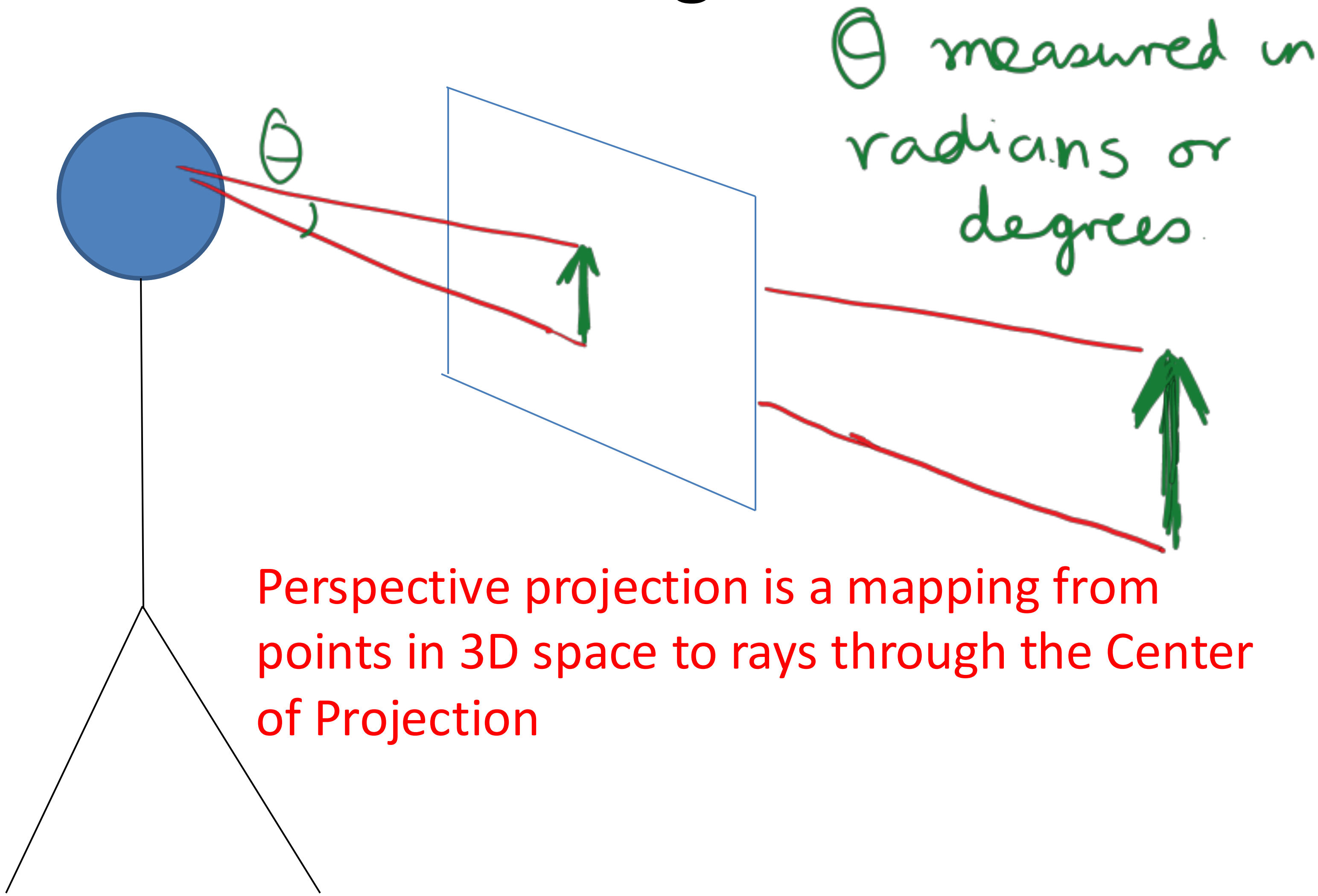
Top at (X, L – h, Z)



Bottom at (X, - h, Z)

It is straightforward to calculate the projection of the top & bottom of the pole. The difference is the “apparent height”

The natural measure of image size is visual angle



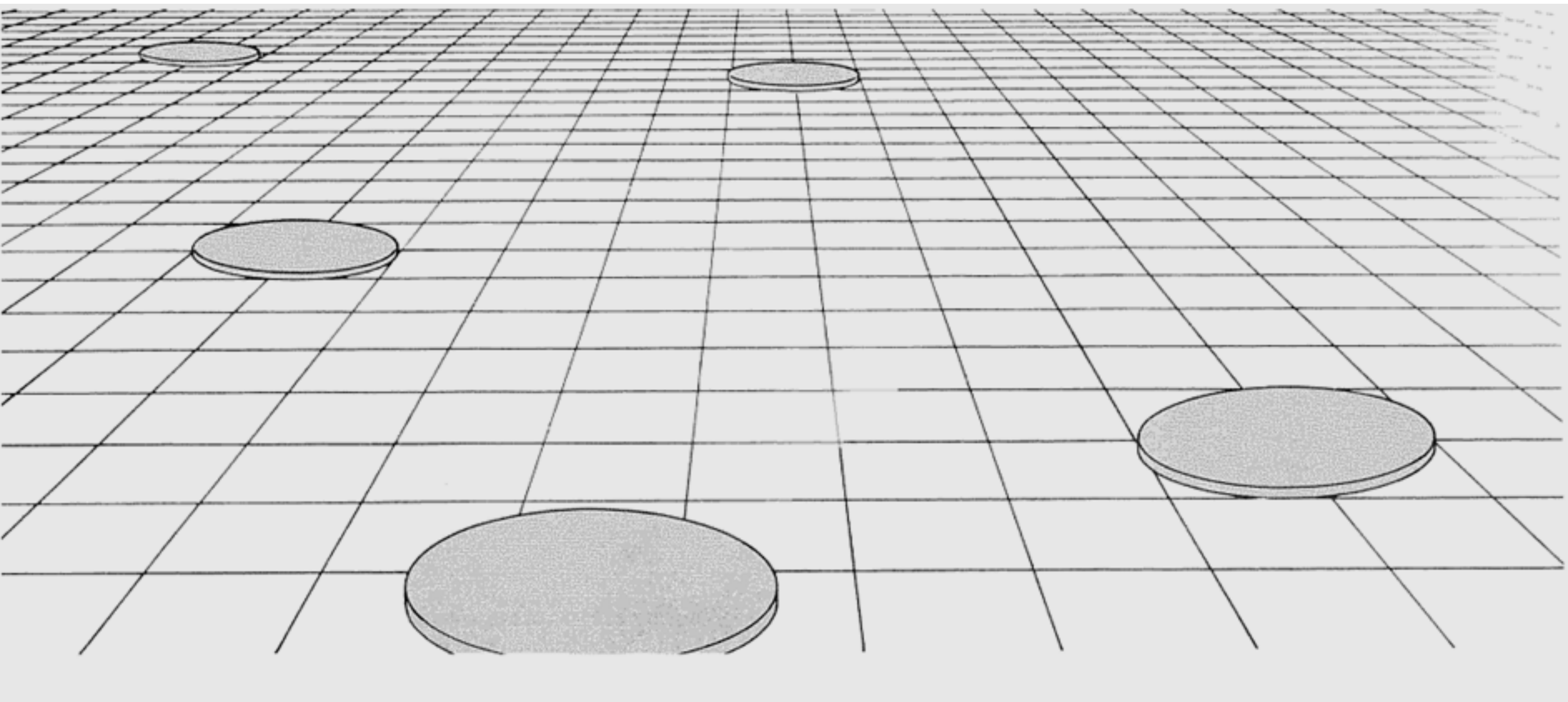
Perspective projection is a mapping from points in 3D space to rays through the Center of Projection

Two main effects of perspective projection

1. Distance – farther objects project to smaller sizes on the image plane. The scaling factor is 1/Z
2. Foreshortening – objects that are slanted with respect to the line of sight project to smaller sizes on the image plane. The scaling factor is cos σ

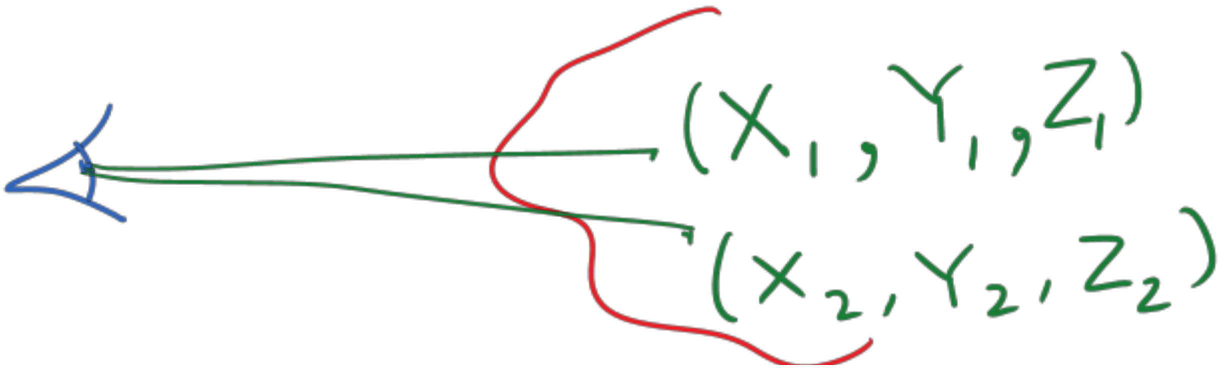
σ is the angle between the line of sight and the surface normal

The slabs that are far away not only look smaller, but also more foreshortened



Orthographic projection

Approximation to perspective when the object is relatively far away compared to the depth variation in it



The idea is as follows: If the depth Z of points on the object varies within some range $Z_0 \pm \Delta Z$, with $\Delta Z \ll Z_0$, then the perspective scaling factor f/Z can be approximated by a constant $s = f/Z_0$. The equations for projection from the scene coordinates (X, Y, Z) to the image plane become $x = sX$ and $y = sY$. Note that scaled orthographic projection is an approximation that is valid only for those parts of the scene with not much internal depth variation;

Cartoon. (Drawing by S. Harris; © 1975 The New Yorker Magazine, Inc.)