



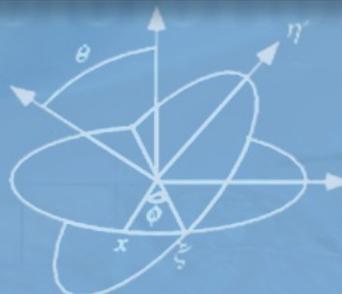
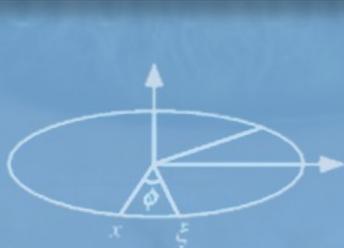
JHU vision lab

# Computer Vision: History, the Rise of Deep Networks, and Future Vistas

Panel on Perception and Cognition, MORS Meeting on Artificial Intelligence and Autonomy

**René Vidal**

Herschel Seder Professor of Biomedical Engineering,  
Director of the Mathematical Institute for Data Science, Johns Hopkins University



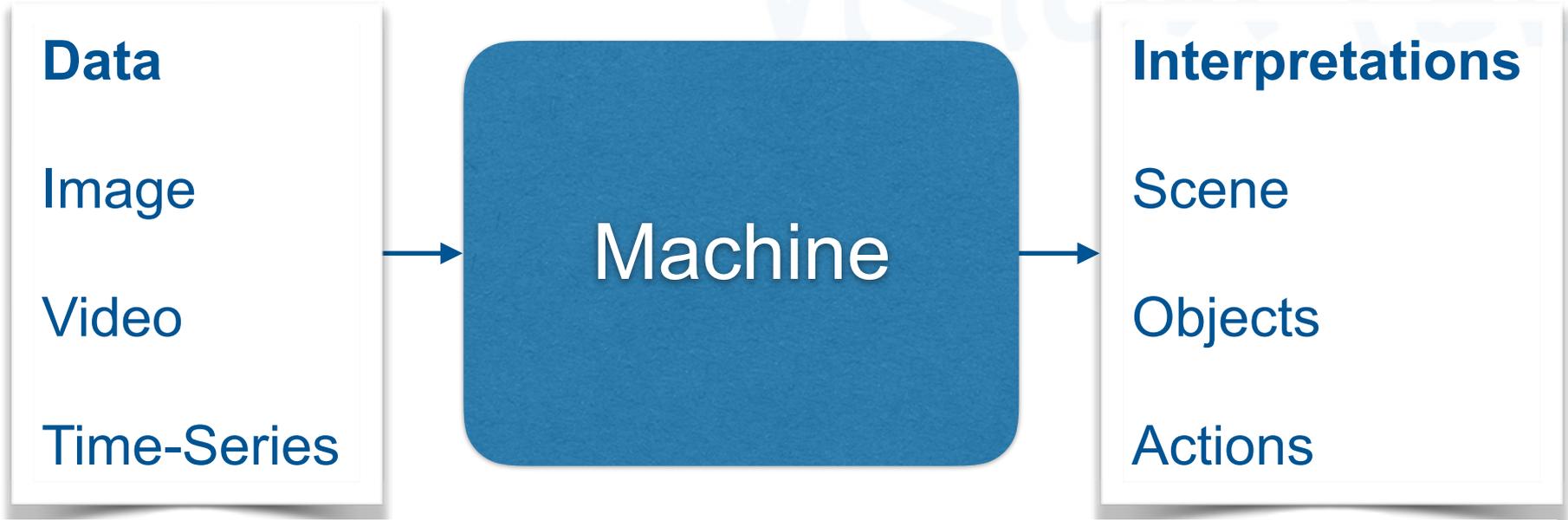
THE DEPARTMENT OF BIOMEDICAL ENGINEERING

The Whitaker Institute at Johns Hopkins



JOHNS HOPKINS  
MATHEMATICAL INSTITUTE  
for DATA SCIENCE

# What is Computer Vision About?



- Body
- Void
- Grass
- Tree
- Dog
- Face
- Sign
- Mixed

# Gilman Hall Atrium



# Scene Classification

What is this scene about?



# Object Verification



Is this a table?

# Object Detection

Is there a table?  
Where?



# Object Classification

What objects are there in the scene?

Window

Bag

Flowers

Shoes

Table

Person

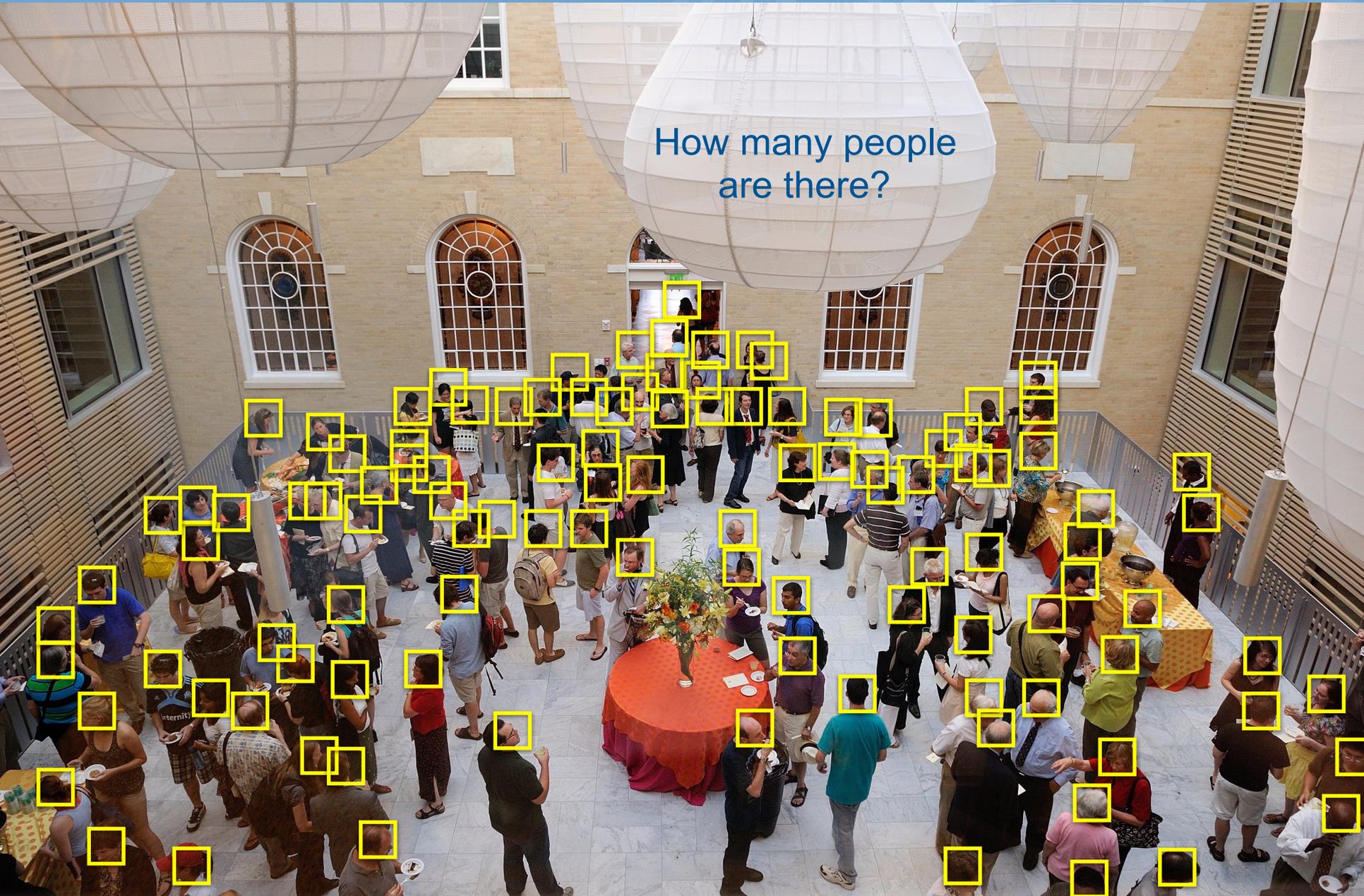
Plate

Floor



# Object Counting

How many people are there?

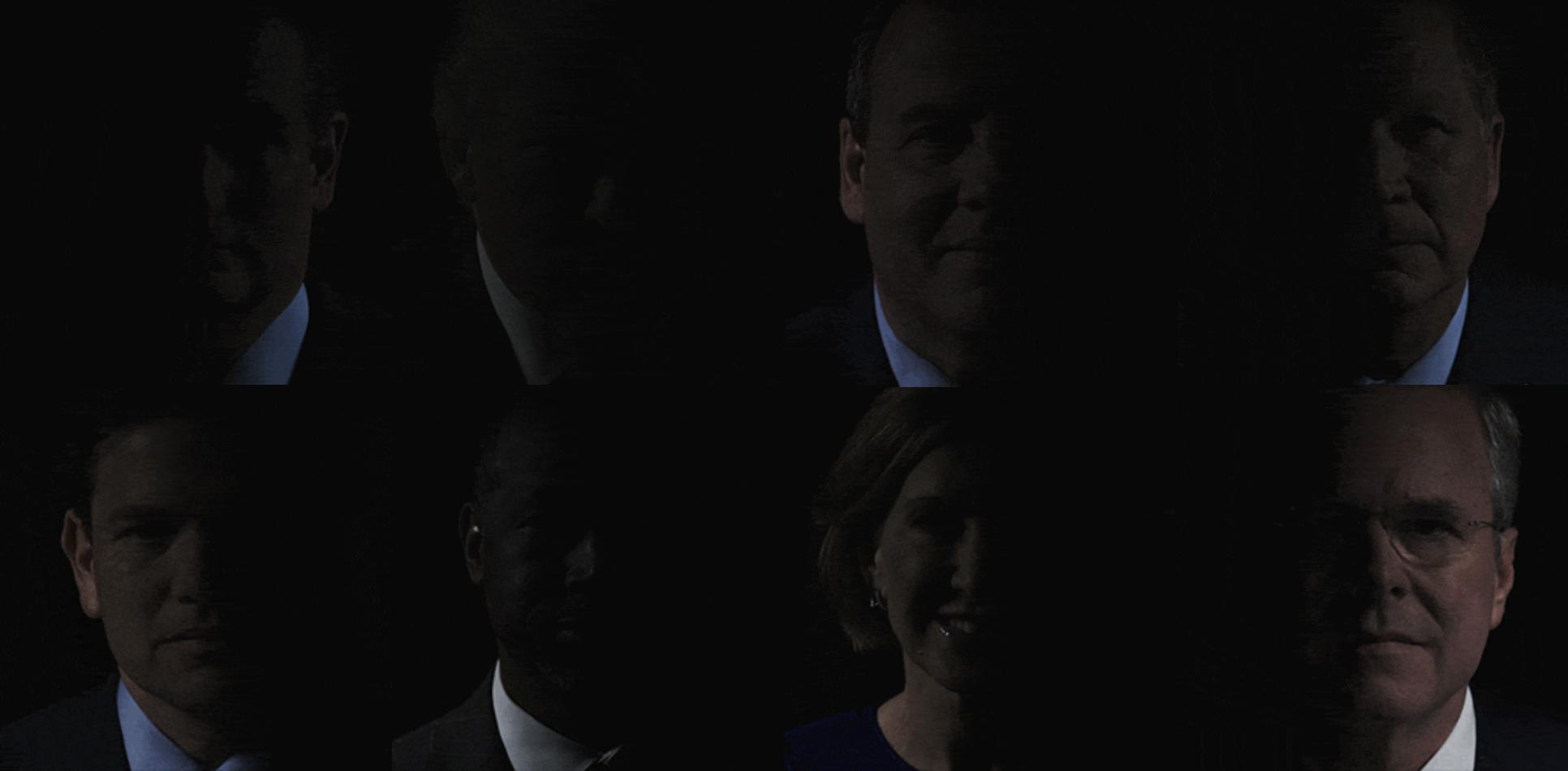
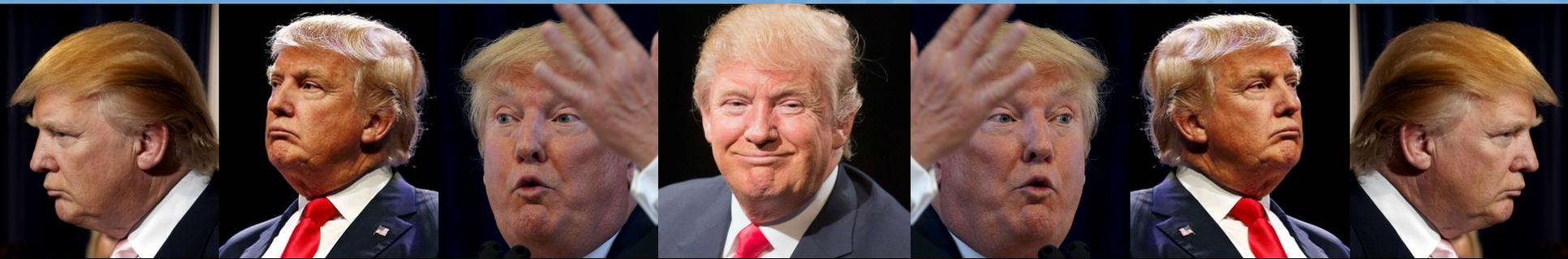


# Scene Understanding

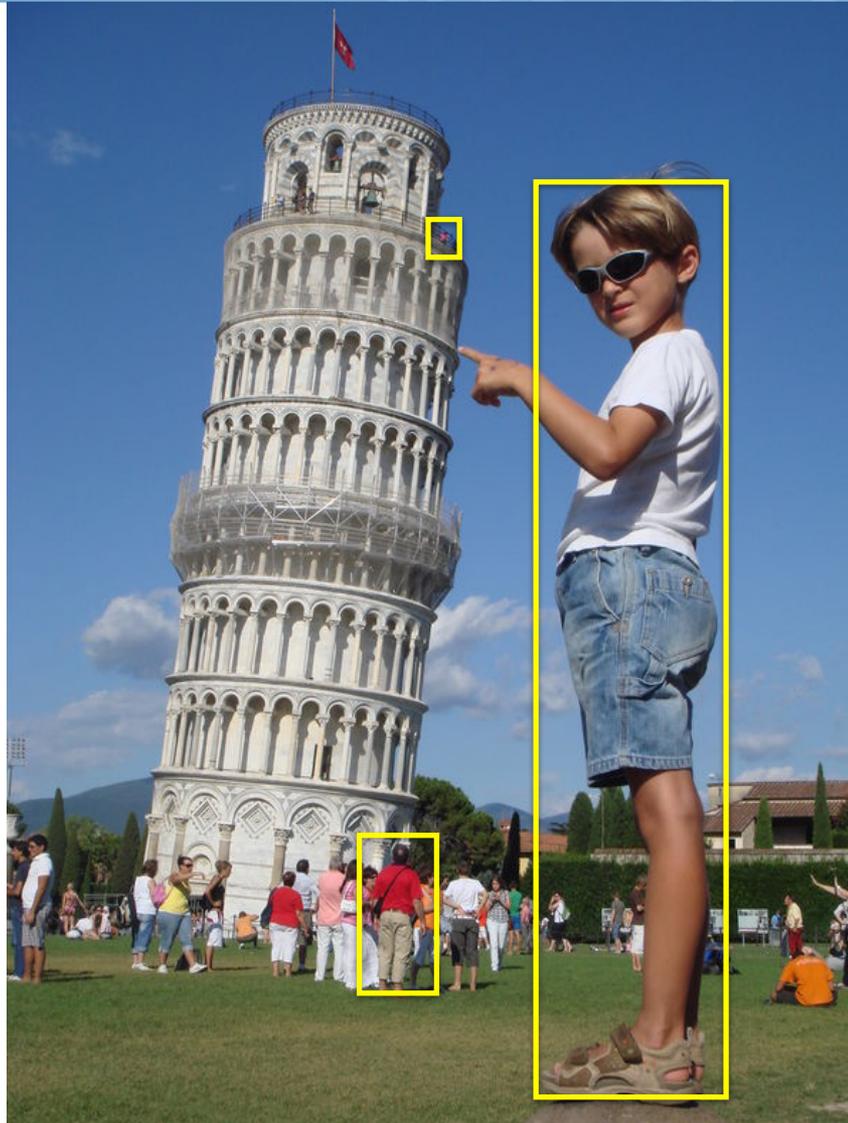
A party with lots of people in a beautiful atrium. Daytime, probably in the afternoon in a warm day.



# Fundamental Challenges: Viewpoint Lighting



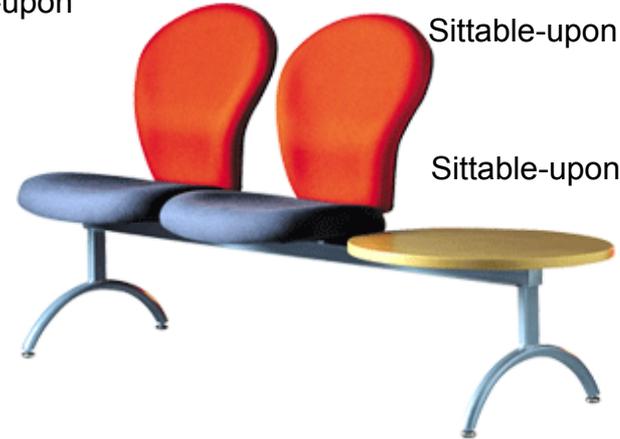
# Fundamental Challenges: Scale



# Fundamental Challenges: What's an Object?



Sittable-upon



Sittable-upon

Sittable-upon

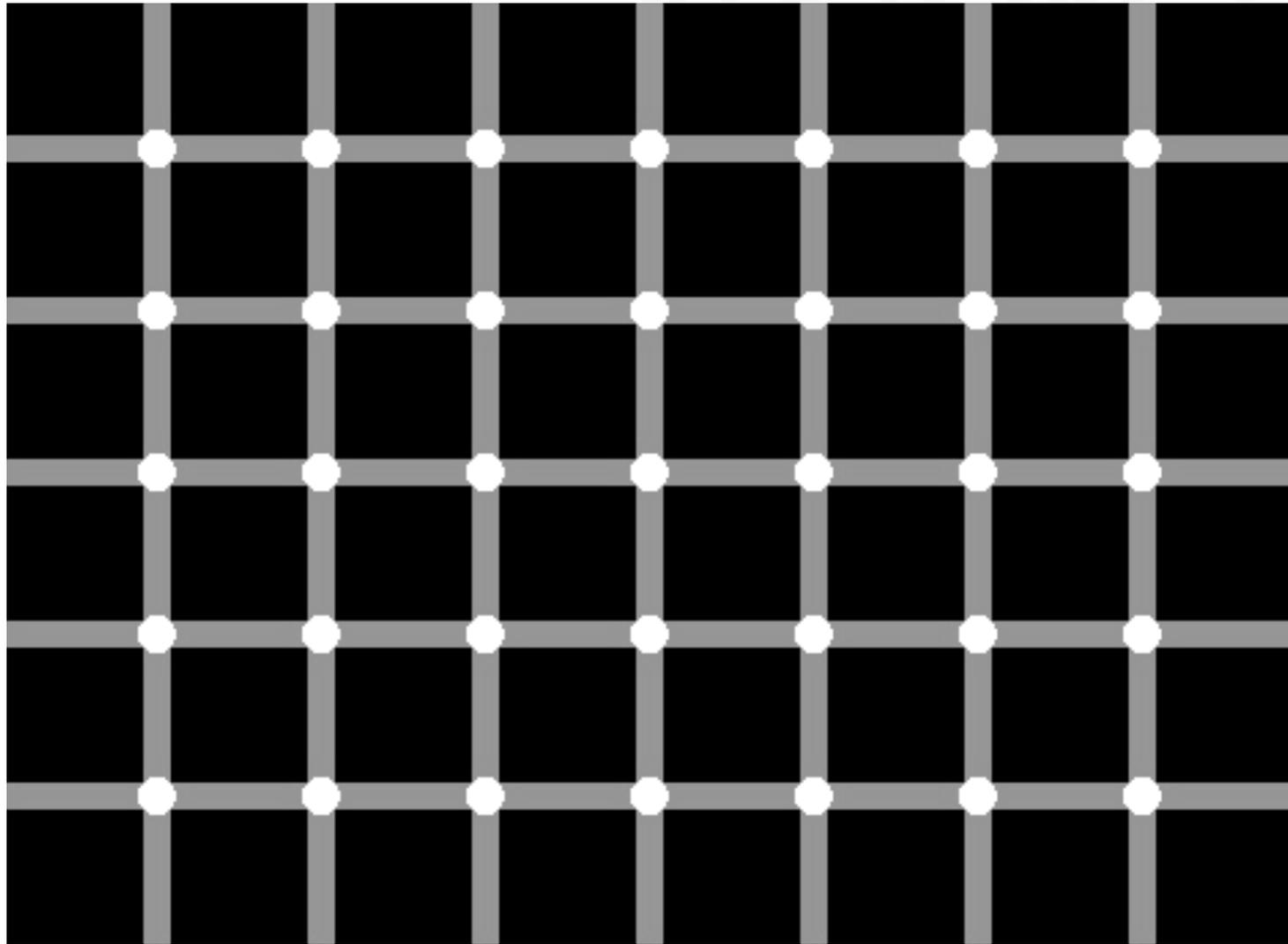
It does not seem easy to sit-upon this...



# Fundamental Challenges: Occlusion, Clutter

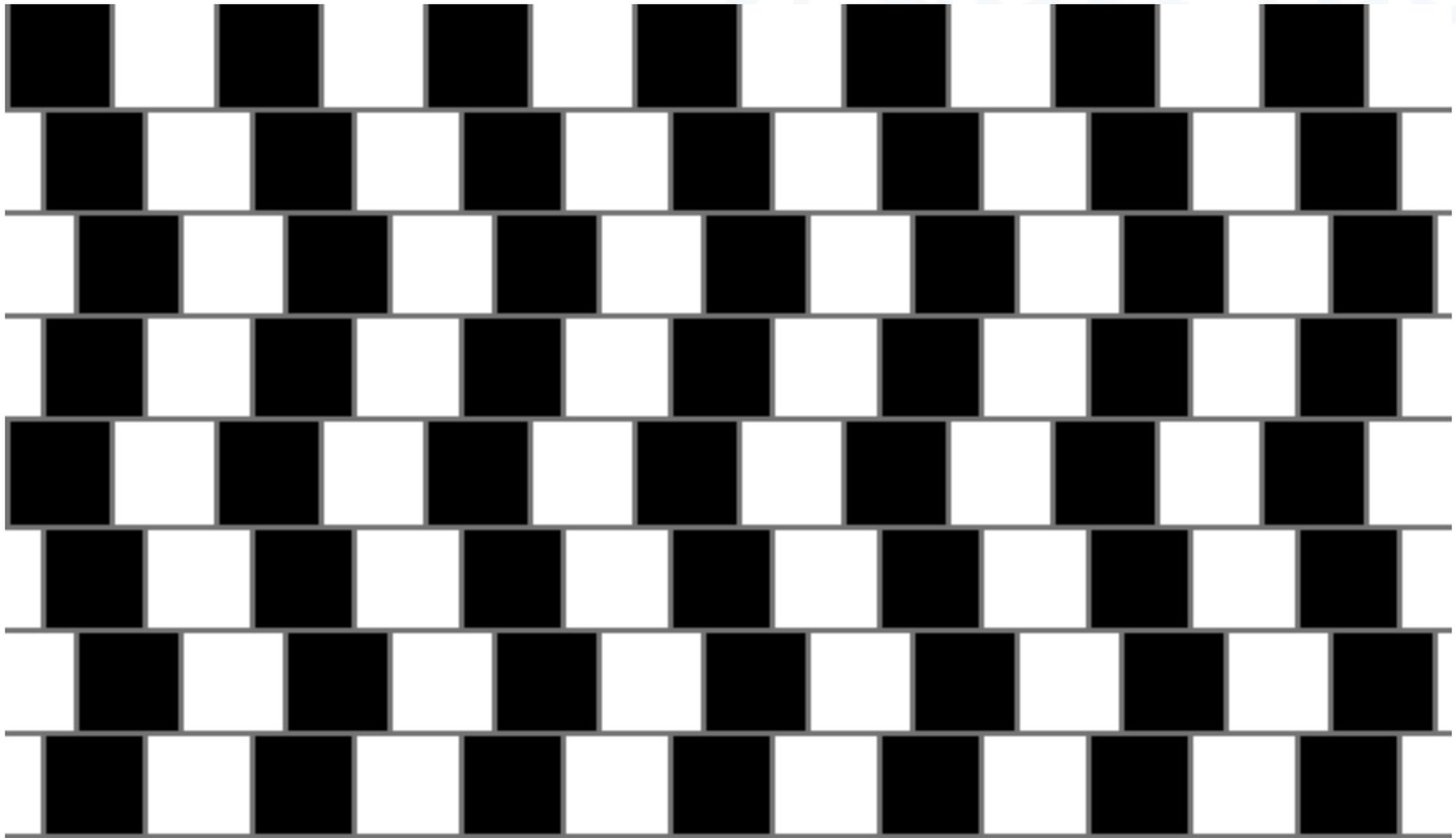


# Illusions

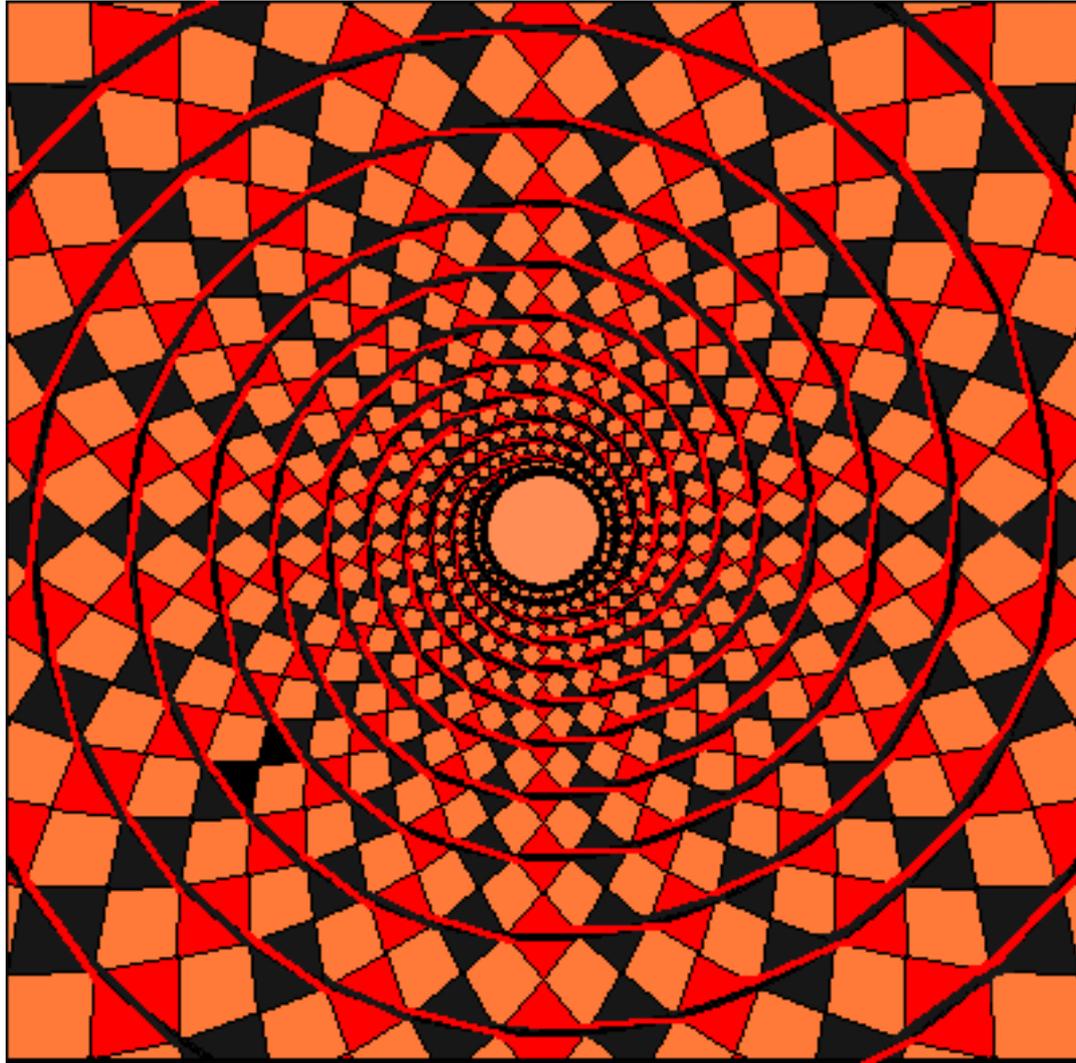


# Illusions

Vision Labs



# Illusions

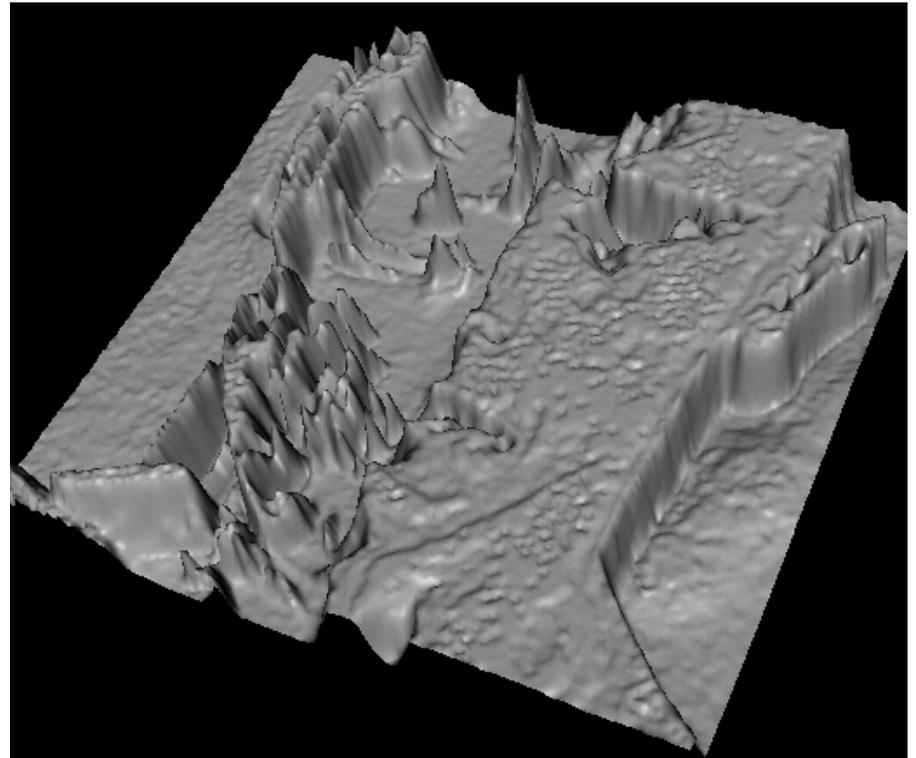


# Fundamental Challenges: Representation

**What we see**



**What a computer sees**



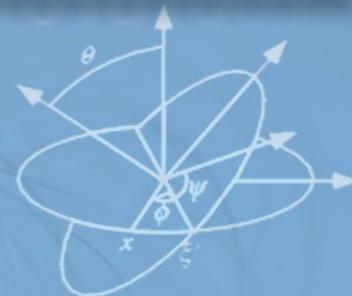
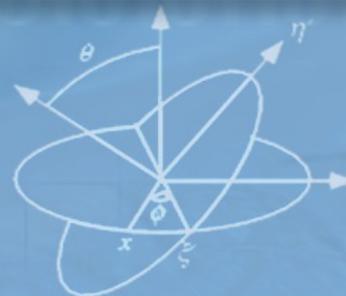
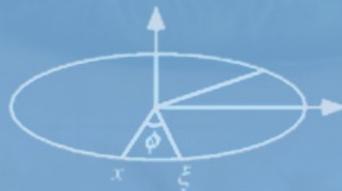


JHU vision lab

# Computer Vision: Historical Overview

**René Vidal**

Herschel Seder Professor of Biomedical Engineering,  
Director of the Mathematical Institute for Data Science, Johns Hopkins University



THE DEPARTMENT OF BIOMEDICAL ENGINEERING

The Whitaker Institute at Johns Hopkins



JOHNS HOPKINS  
MATHEMATICAL INSTITUTE  
for DATA SCIENCE

# Minsky's Summer Vision Project: July 1966

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group  
Vision Memo. No. 100.

July 7, 1966

## THE SUMMER VISION PROJECT

Seymour Papert

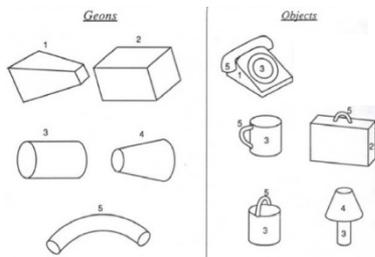
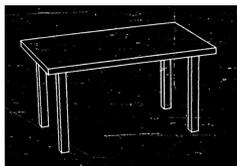
The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the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



# Object Recognition: Historical Overview

2D: View  
Centered

3D: Object  
Centered



1960

1970

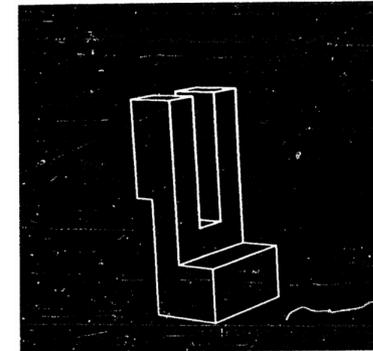
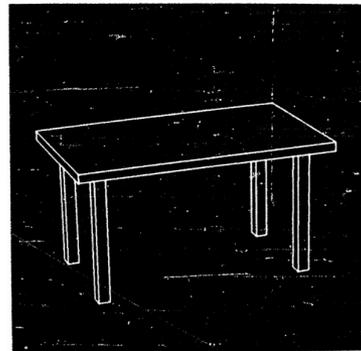
1980

1990

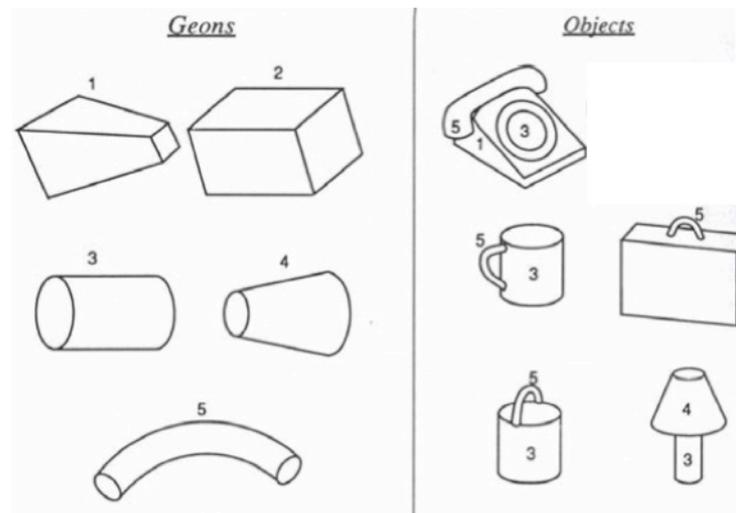
2000

2010

Blocks World



Geometric Ions (Geons)



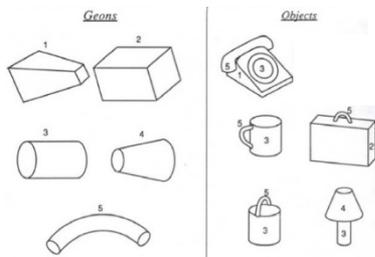
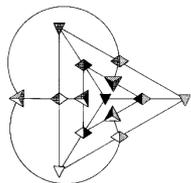
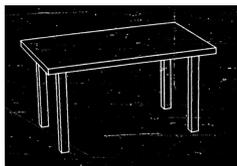
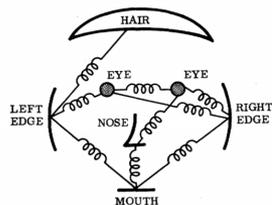
[1] Roberts, Machine Perception of Three-Dimensional Solids, 1963.

[2] Biederman, Recognition-by-components: A theory of human image understanding, 1987.

# Object Recognition: Historical Overview

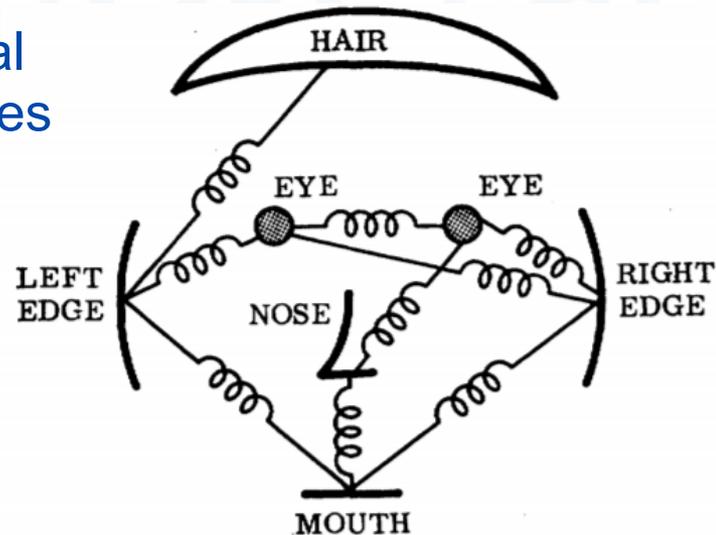
2D: View  
Centered

3D: Object  
Centered



1960

Pictorial  
Structures



1970

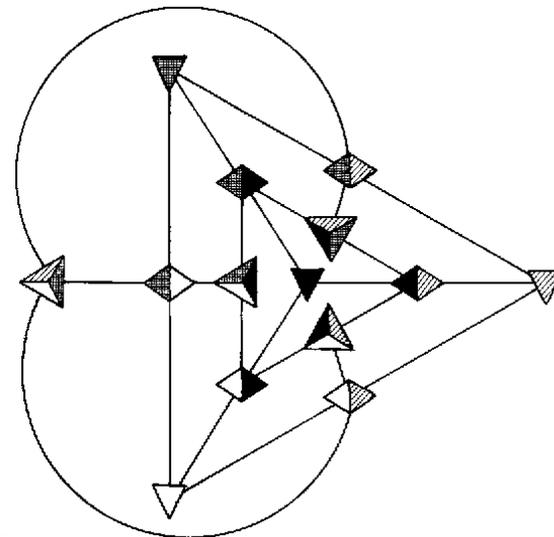
1980

1990

2000

Aspect  
Graphs

2010



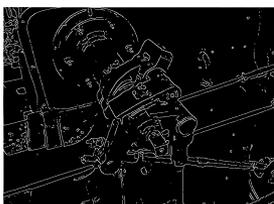
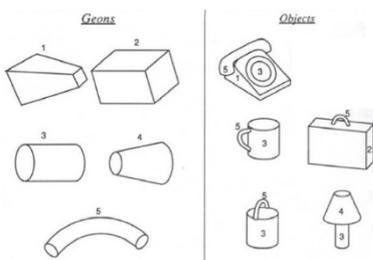
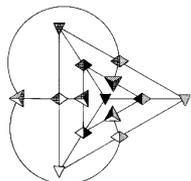
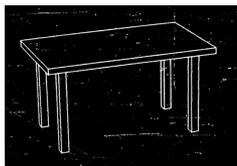
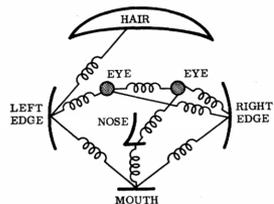
[1] Fischler & Elschlager, The Representation and Matching of Pictorial Structures, 1973.

[2] Koenderink & van Doorn, The internal representation of solid shape with respect to vision, 1979.

# Object Recognition: Historical Overview

2D: View  
Centered

3D: Object  
Centered



1960

1970

1980

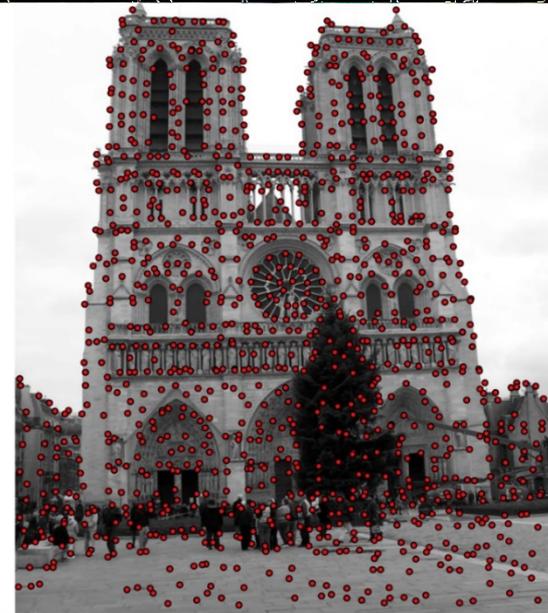
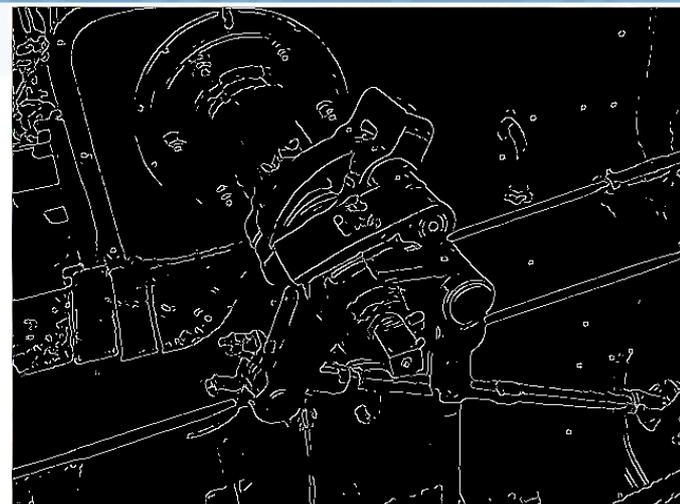
1990

2000

2010

Canny  
Edges

Harris  
Corners

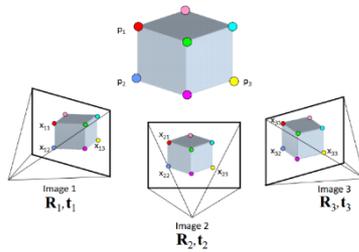
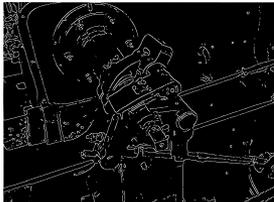
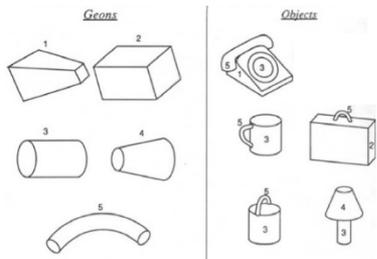
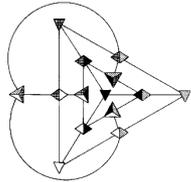
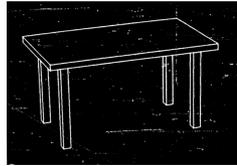
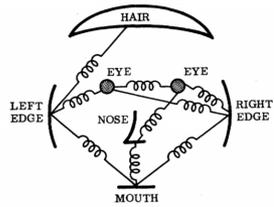


[1] Canny, A Computational Approach To Edge Detection, 1986.  
[2] Harris & Stephens, A Combined Corner and Edge Detector, 1988.

# Object Recognition: Historical Overview

2D: View Centered

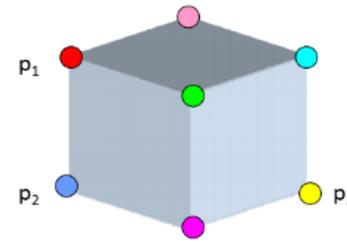
3D: Object Centered



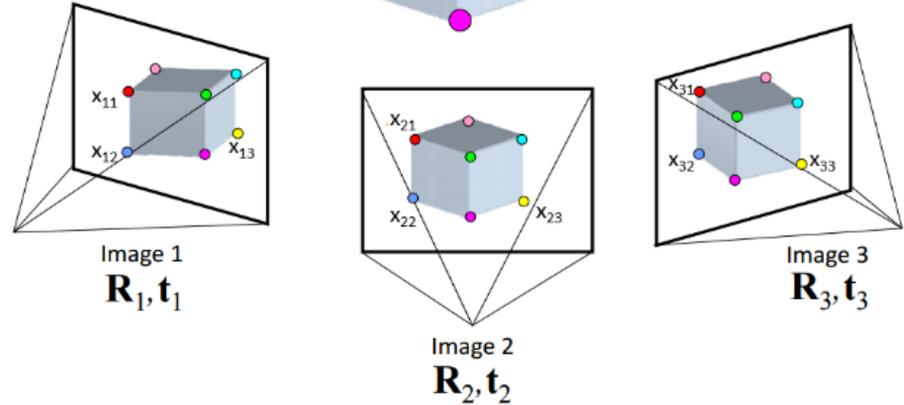
1960

Structure from Motion

1970



1980



1990

2000

2010

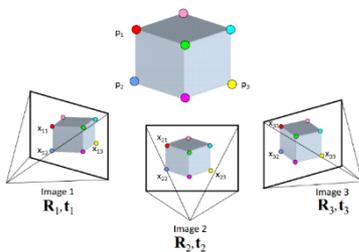
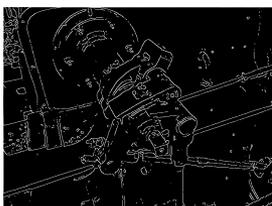
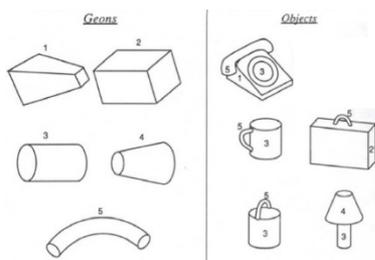
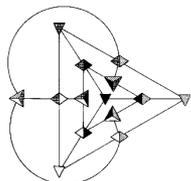
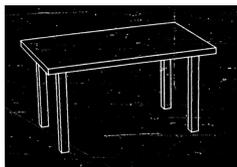
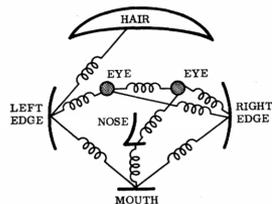
Use Multiple 2D Images to Reconstruct and Recognize a 3D Object

[1] Tomasi-Kanade. Shape and motion from image streams under orthography: a factorization method, 1992.  
 [2] Sturm-Triggs. A factorization based algorithm for multi-image projective structure and motion, 1996

# Object Recognition: Historical Overview

2D: View Centered

3D: Object Centered



1960

SIFT

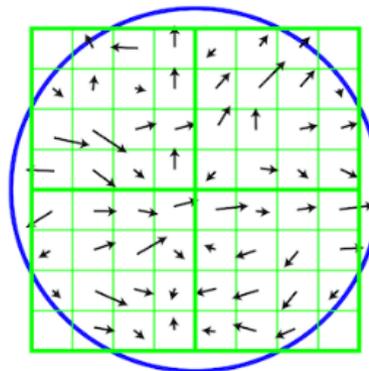
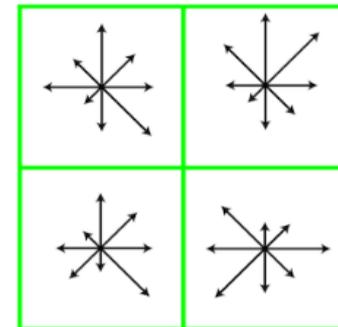


Image gradients

SURF



Keypoint descriptor

1970

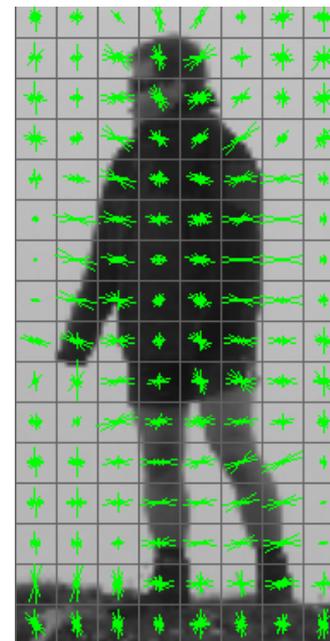
1980

HOG

1990

2000

2010

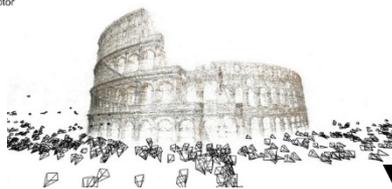
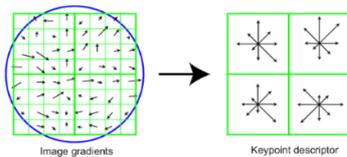
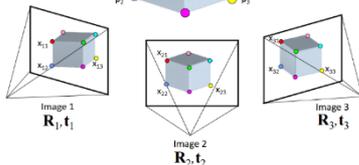
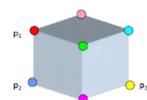
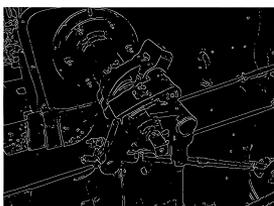
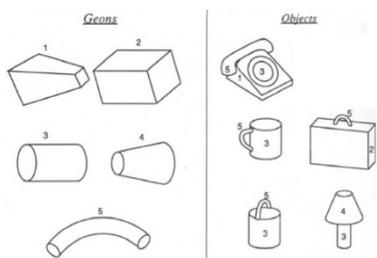
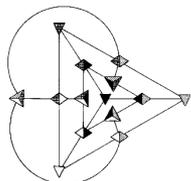
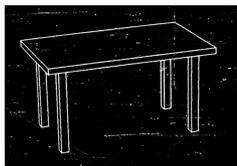
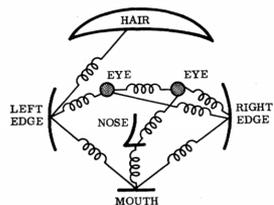


[1] Dalal & Triggs, Histograms of oriented gradients for human detection, 2005.  
 [2] Lowe, Distinctive image features from scale-invariant keypoints, 2005.  
 [3] Bay, Ess, Tuytelaars & Van Gool, Speeded Up Robust Features, 2004.

# Object Recognition: Historical Overview

2D: View Centered

3D: Object Centered



1960

1970

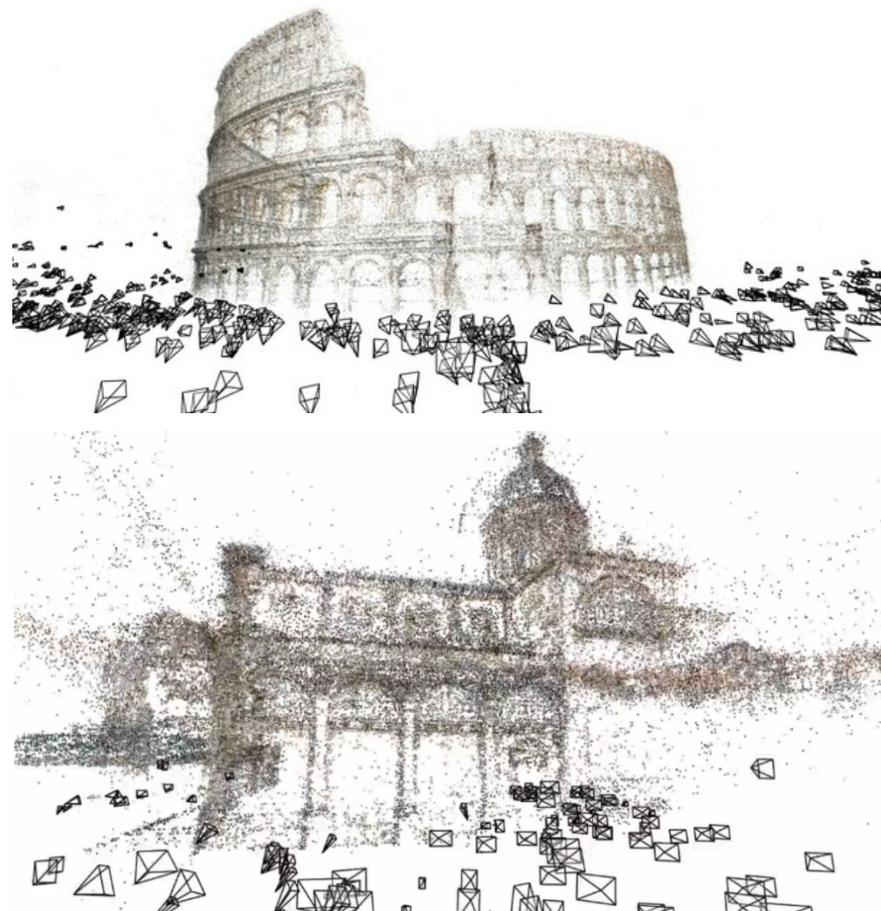
1980

1990

2000

2010

Building Rome in a Day

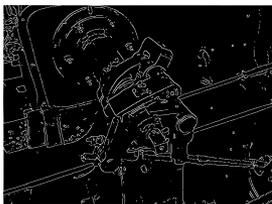
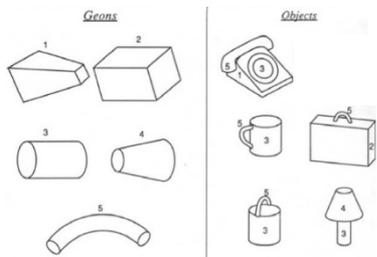
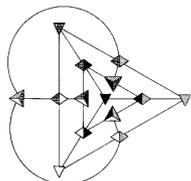
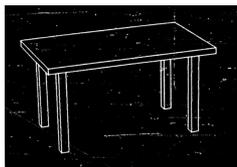
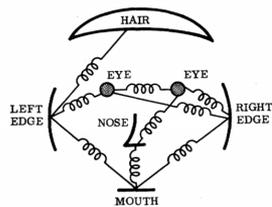


[1] Agarwal et al., Building Rome in a Day, ICCV 2009.  
 [2] Agarwal et al., Reconstructing Rome, CVPR 2010.

# Object Recognition: Historical Overview

2D: View Centered

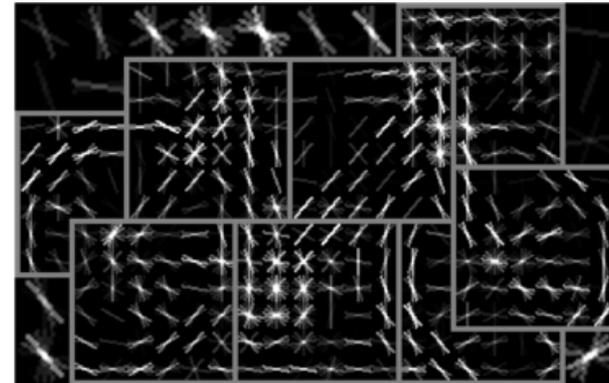
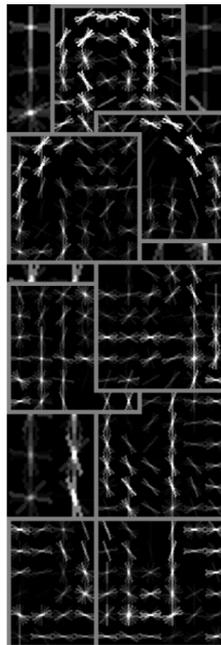
3D: Object Centered



1960

Deformable part models

1970

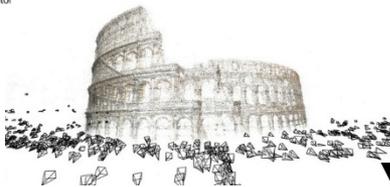
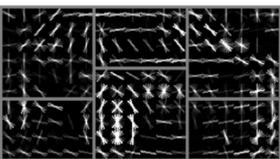
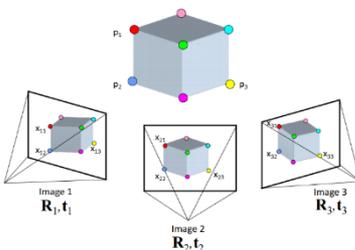
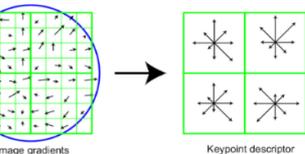
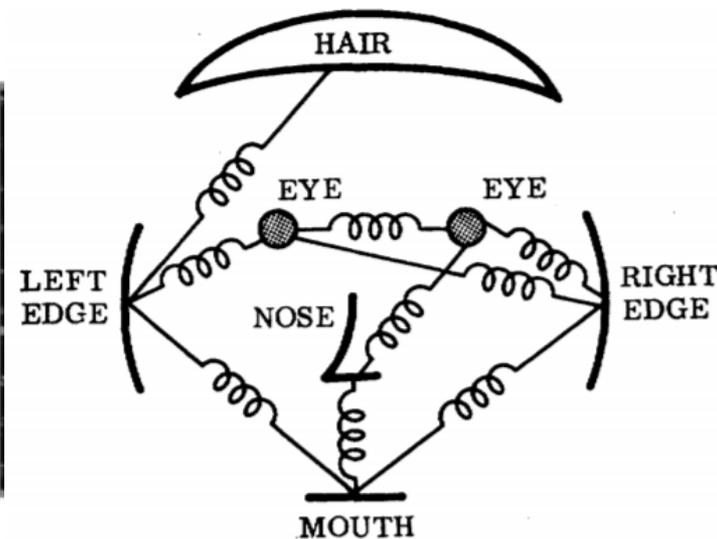


1980

1990

2000

2010



[1] Felzenszwalb et al., A discriminatively trained, multiscale, deformable part model, CVPR 2008.

[2] Felzenszwalb et al., Object detection with discriminatively trained part-based models, PAMI 2010.

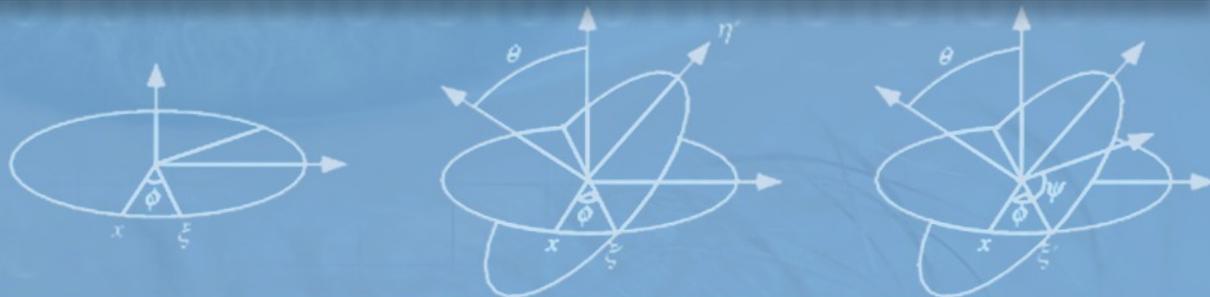


JHU vision lab

# Computer Vision: The Rise of Deep Networks

**René Vidal**

Herschel Seder Professor of Biomedical Engineering,  
Director of the Mathematical Institute for Data Science, Johns Hopkins University



THE DEPARTMENT OF BIOMEDICAL ENGINEERING

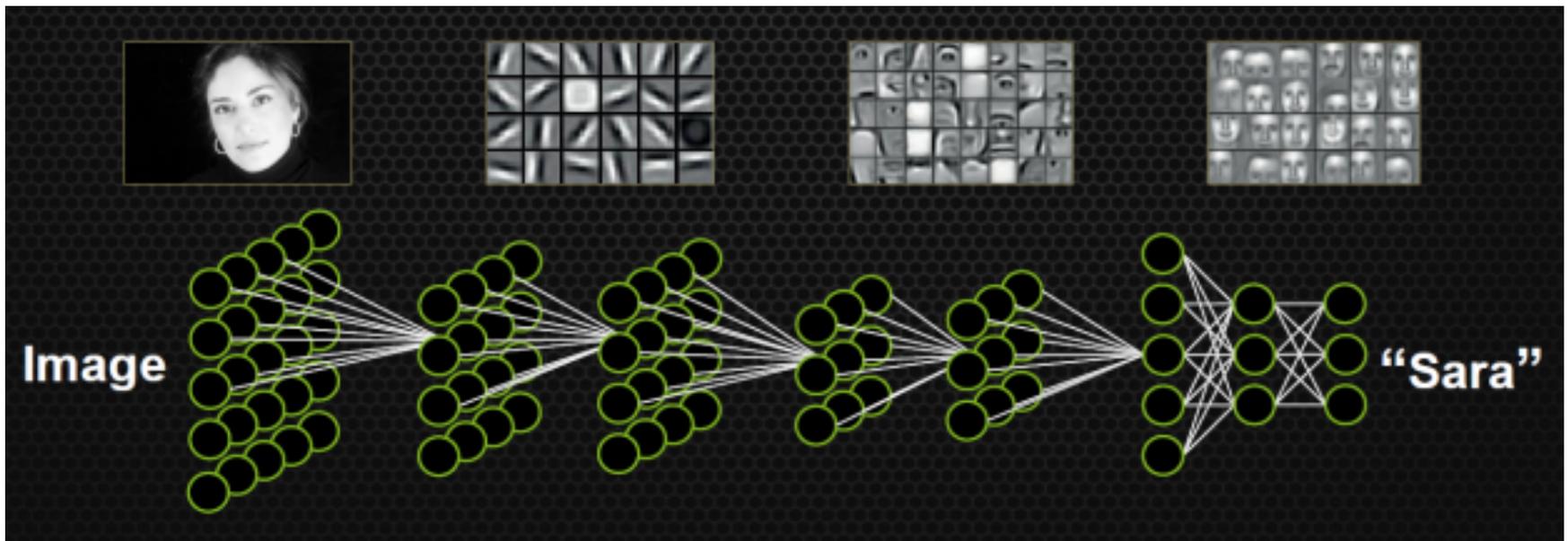
The Whitaker Institute at Johns Hopkins



JOHNS HOPKINS  
MATHEMATICAL INSTITUTE  
for DATA SCIENCE

# Impact of Deep Learning in Computer Vision

- Deep learning gives ~ 10% improvement on ImageNet
  - 1.2M images
  - 1000 categories
  - 60 million parameters



[1] Krizhevsky, Sutskever and Hinton. ImageNet classification with deep convolutional neural networks, NIPS'12.

[2] Sermanet, Eigen, Zhang, Mathieu, Fergus, LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. ICLR'14.

[3] Donahue, Jia, Vinyals, Hoffman, Zhang, Tzeng, Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. ICML'14.



# Impact of Deep Learning in Computer Vision

- 2012-2014 classification results in ImageNet

CNN  
non-CNN

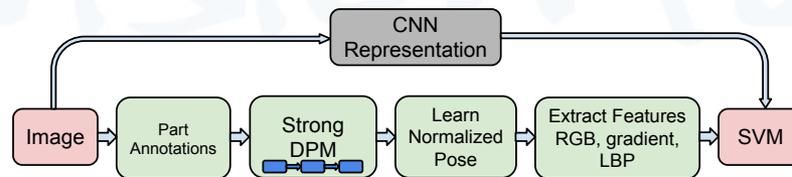
2012 Teams	%error	2013 Teams	%error	2014 Teams	%error
Supervision (Toronto)	15.3	Clarifai (NYU spinoff)	11.7	GoogLeNet	6.6
ISI (Tokyo)	26.1	NUS (singapore)	12.9	VGG (Oxford)	7.3
VGG (Oxford)	26.9	Zeiler-Fergus (NYU)	13.5	MSRA	8.0
XRCE/INRIA	27.0	A. Howard	13.5	A. Howard	8.1
UvA (Amsterdam)	29.6	OverFeat (NYU)	14.1	DeeperVision	9.5
INRIA/LEAR	33.4	UvA (Amsterdam)	14.2	NUS-BST	9.7
		Adobe	15.2	TTIC-ECP	10.2
		VGG (Oxford)	15.2	XYZ	11.2
		VGG (Oxford)	23.0	UvA	12.1

- 2015 results: ResNet under 3.5% error using 150 layers!

# Transfer from ImageNet to Other Datasets

## • CNNs + SMVs [1]

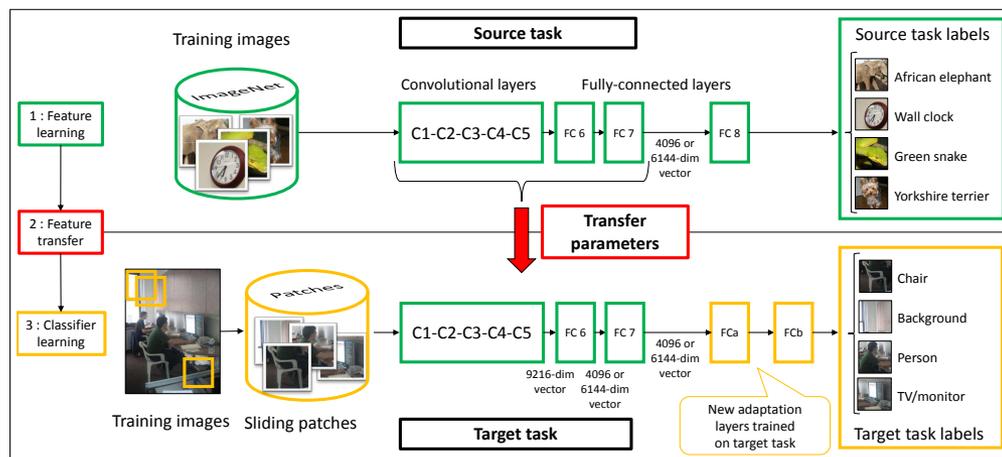
Pascal VOC 2007	mAP
GHM[8]	64.7
AGS[11]	71.1
NUS[39]	70.5
CNN-SVM	73.9
CNNaug-SVM	<b>77.2</b>



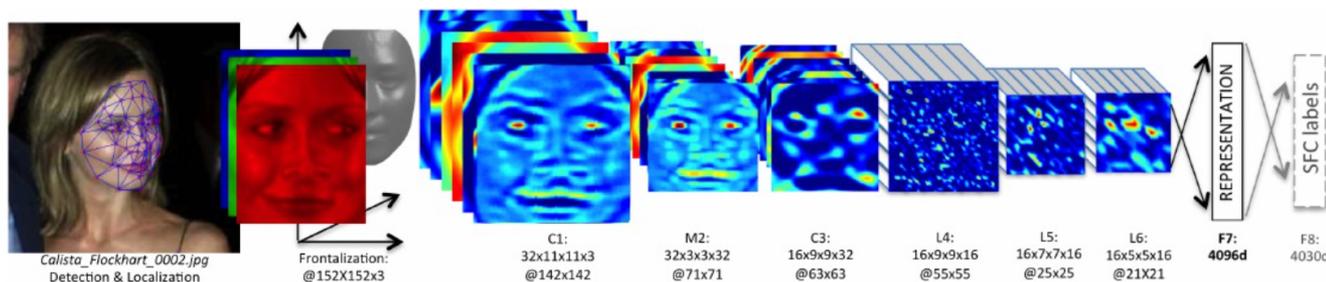
## • Retrain top-layer [2]

Pascal VOC 2007	mAP
INRIA [32]	59.4
NUS-PSL [44]	70.5
PRE-1000C	<b>77.7</b>

Pascal VOC 2012	mAP
NUS-PSL [49]	82.2
NO PRETRAIN	70.9
PRE-1000C	78.7
PRE-1000R	76.3
PRE-1512	<b>82.8</b>



## • Deep Face [3]



[1] Razavian, Azizpour, Sullivan, Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition. CVPRW'14.

[2] Oquab, Bottou, Laptev, Sivic. Learning and transferring mid-level image representations using convolutional neural networks CVPR'14

[3] Taigman, Yang, Ranzato, Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. CVPR'14

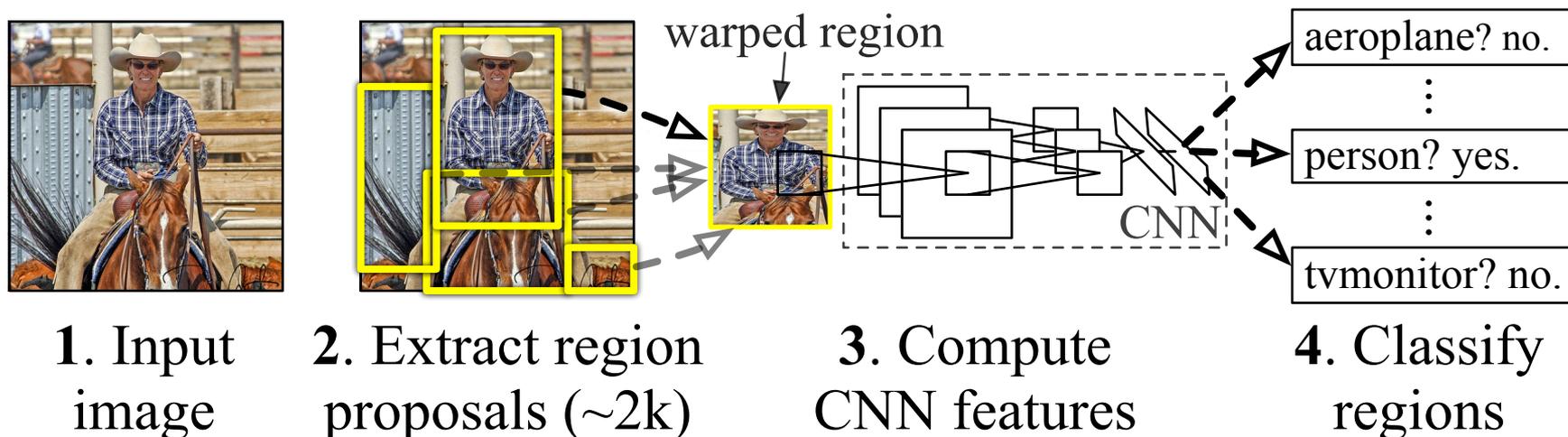
# Transfer from Classification to Detection

- R-CNN, OverFeat, SPPNets, MultiBox, YOLO

- Extract region proposals
- Compute CNN features
- Classify proposal features
- Detect by using regression to refine proposal

VOC 2010 test	mAP
DPM v5 [20] <sup>†</sup>	33.4
UVA [39]	35.1
Regionlets [41]	39.7
SegDPM [18] <sup>†</sup>	40.4
R-CNN	50.2
R-CNN BB	<b>53.7</b>

## R-CNN: *Regions with CNN features*



[1] Girshick, Donahue, Darrell, Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR'14

[2] Sermanet et al. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. ICLR

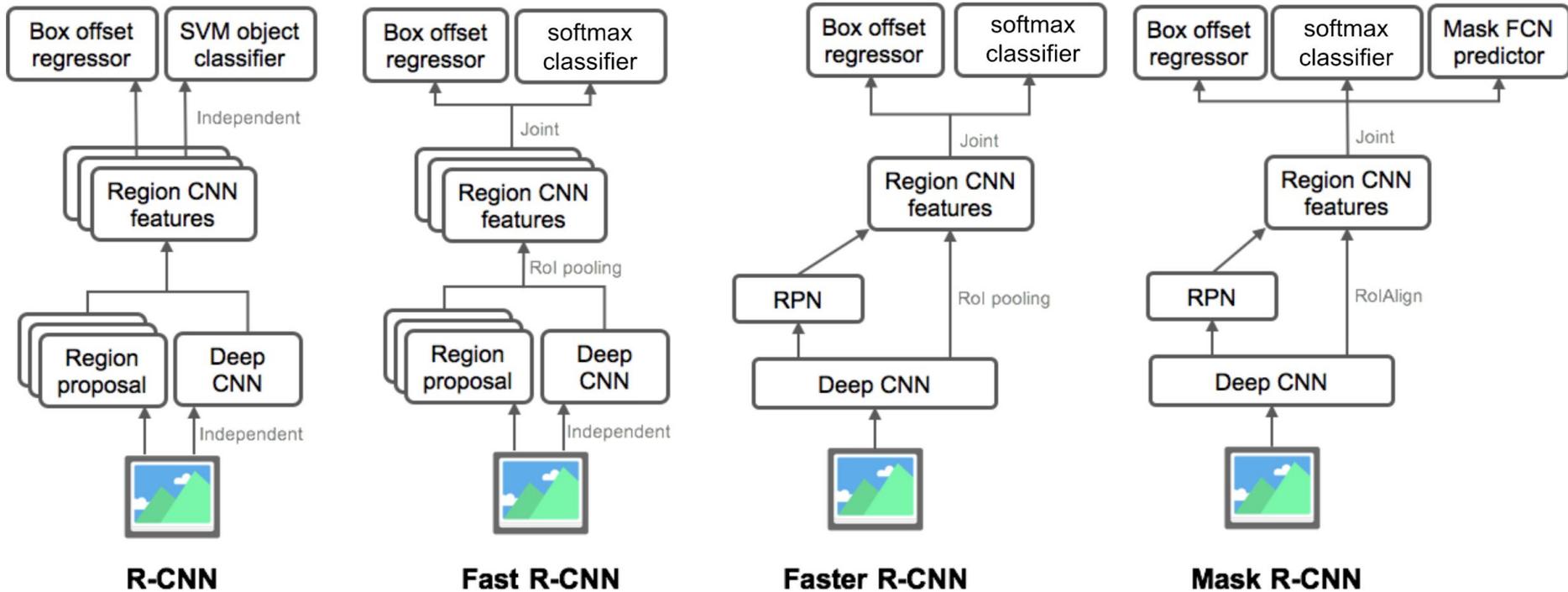
[3] He, Zhang, Ren, Sun. Spatial Pyramid Pooling in deep convolutional networks for visual recognition. ECCV 2004.

[4] Liu, Anguelov, Erhan, Szegedy, Reed, Fu, Berg. SSD: Single Shot MultiBox Detector. ECCV 2016.

[5] Redmon, Divvala, Girshick, Farhadi. You Only Look Once: Unified, Real-Time Object Detection. CVPR 2016.

# Transfer from Classification to Detection

- RCNN Family

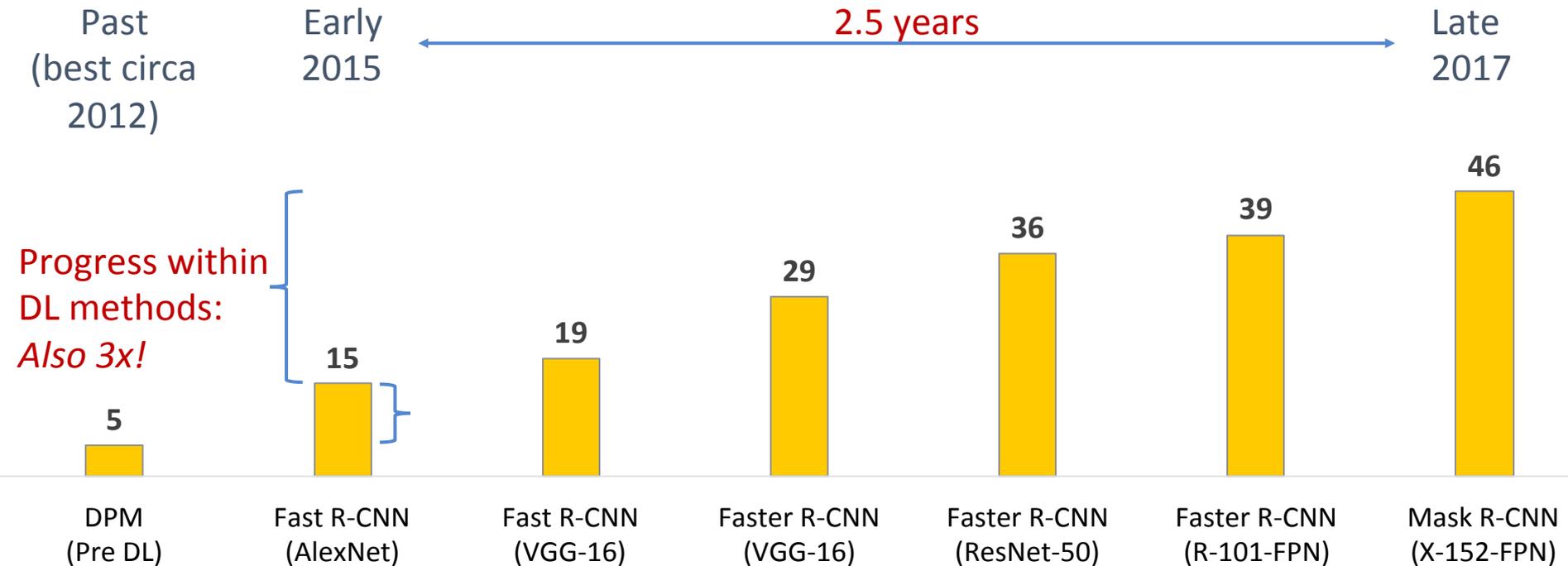


Slide Courtesy of Jiageng Zhang, Jingyao Zhang, Yanhan Ma

[1] Girshick, Donahue, Darrell and Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, CVPR'14  
[2] Girshick. Fast R-CNN. ICCV 2015.  
[3] Ren, He, Girshick, Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. NIPS 2015.  
[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick. Mask R-CNN. ICCV 2017.

# Transfer from Classification to Detection

## COCO Object Detection Average Precision (%)



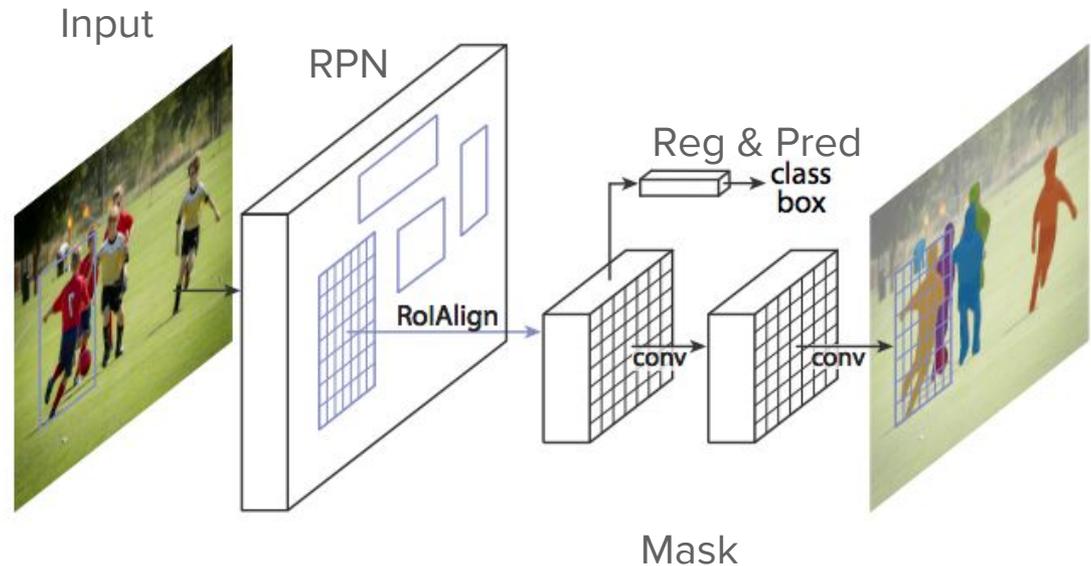
Slide Courtesy of Ross Girshick, ECCV18

# Transfer from Classification to Other Tasks

- CNNs for pose estimation [1] and semantic segmentation [2]



- Mask RCNNs [3]

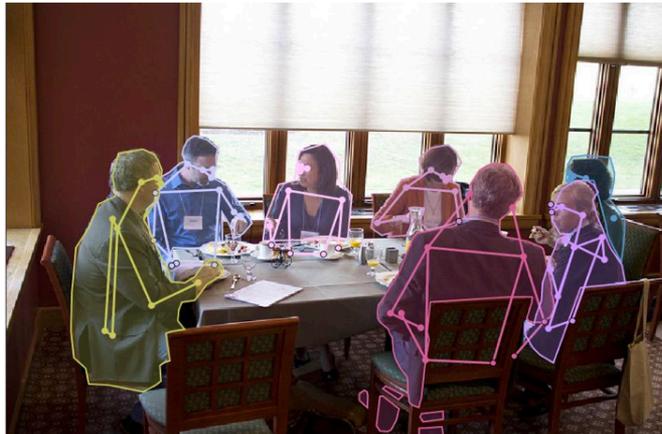


[1] Tompson, Goroshin, Jain, LeCun, Bregler. Efficient Object Localization Using Convolutional Networks. CVPR'15

[2] Pinheiro, Collobert, Dollár. Learning to Segment Object Candidates. NIPS'15

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick. Mask R-CNN. ICCV 2017.

# Transfer from Classification to Keypoints



Slide Courtesy of Ross Girshick, ECCV18

# Transfer from Classification to Surfaces



Slide Courtesy of Ross Girshick, ECCV18

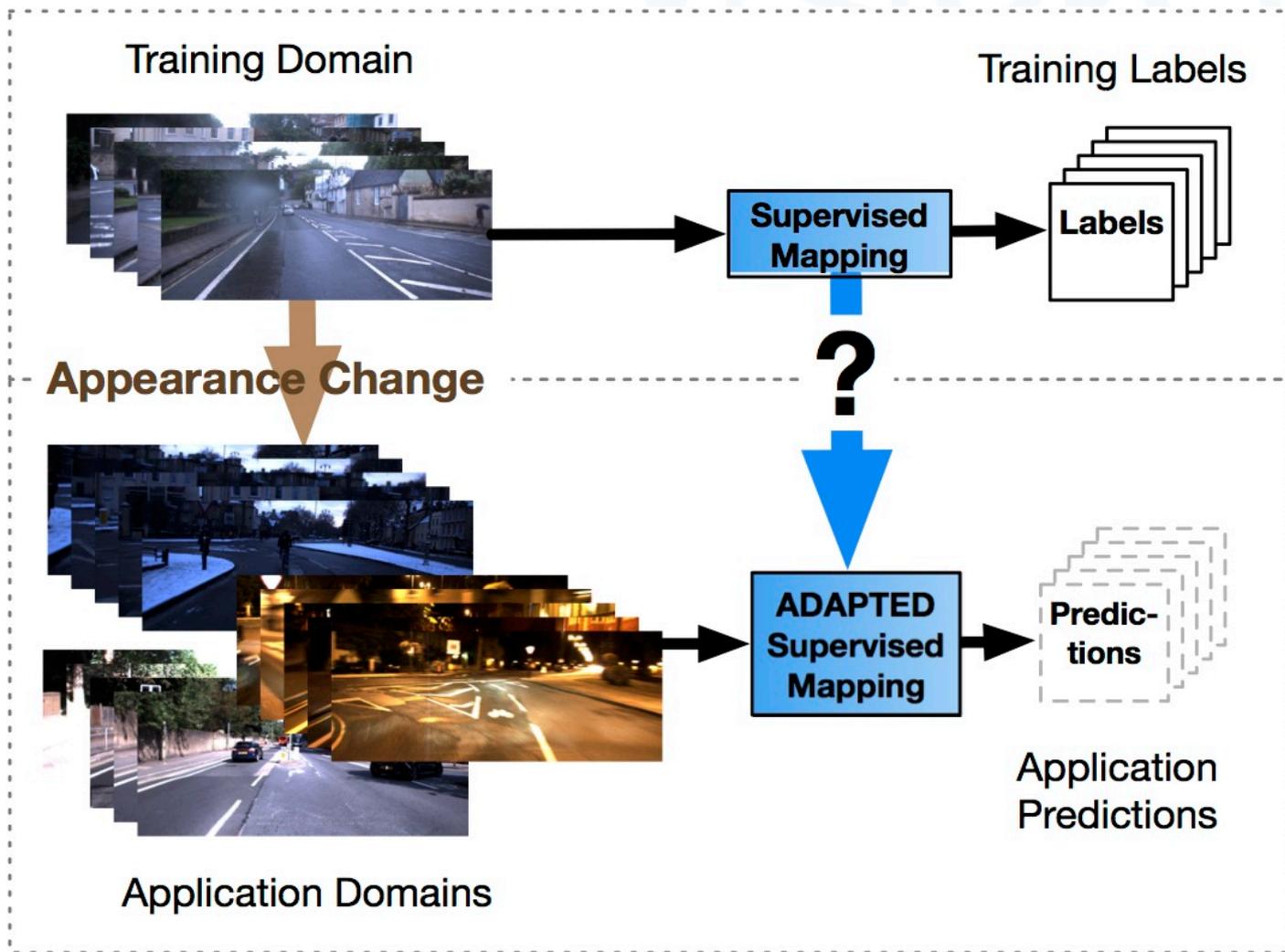
Güler, Neverova, Kokkinos DensePose: Dense Human Pose Estimation In The Wild, CVPR 2018.

# Transfer from Classification to 3D Shape



Slide Courtesy of Ross Girshick, ECCV18

# Transfer to Other Domains



# Generative Adversarial Networks

- “the most interesting idea in the last 10 years in ML.” (LeCun)

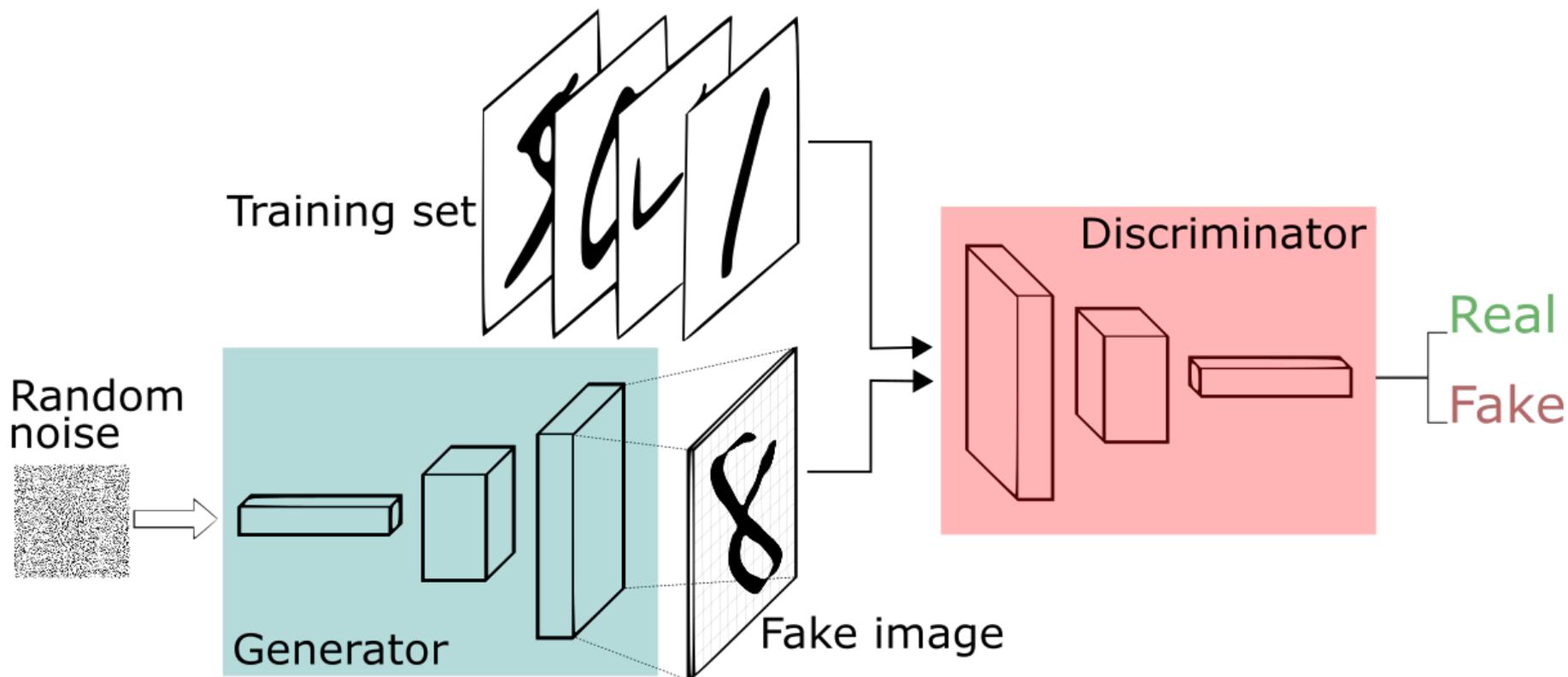
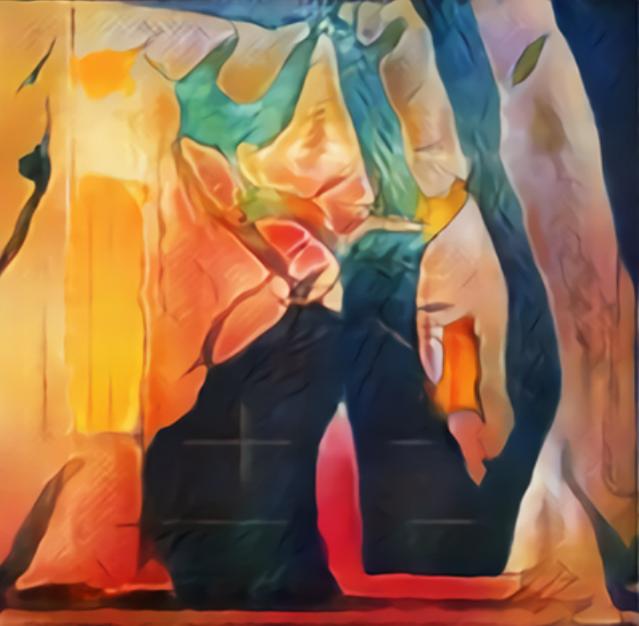


Image credit: Thalles Silva

# Generative Adversarial Networks



# GANs for Style Generation



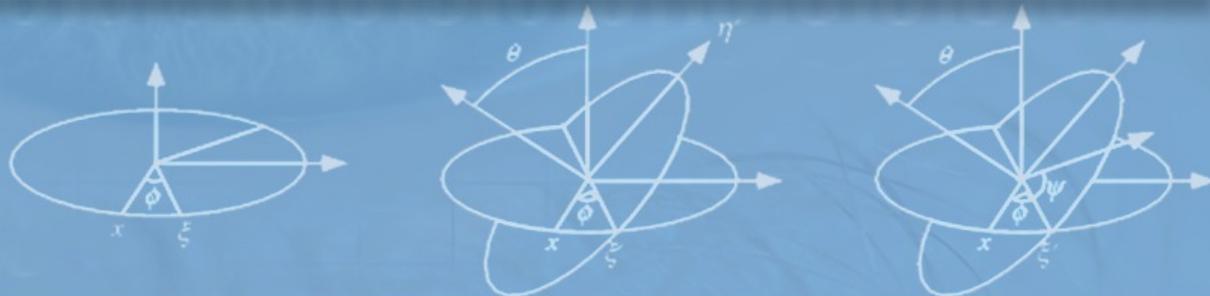


JHU vision lab

# Computer Vision: Future Vistas

**René Vidal**

Herschel Seder Professor of Biomedical Engineering,  
Director of the Mathematical Institute for Data Science, Johns Hopkins University



THE DEPARTMENT OF BIOMEDICAL ENGINEERING

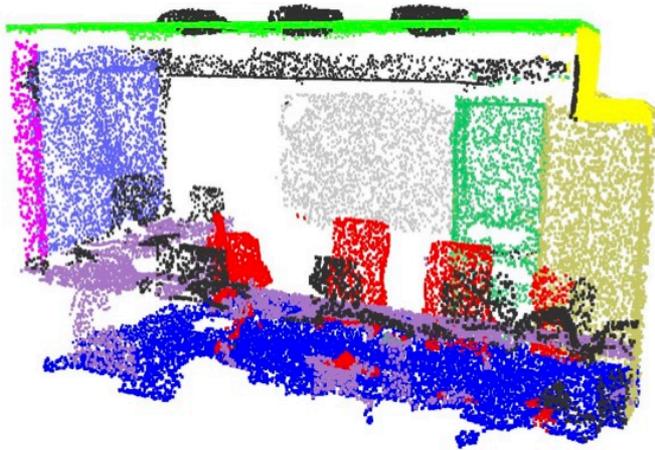
The Whitaker Institute at Johns Hopkins



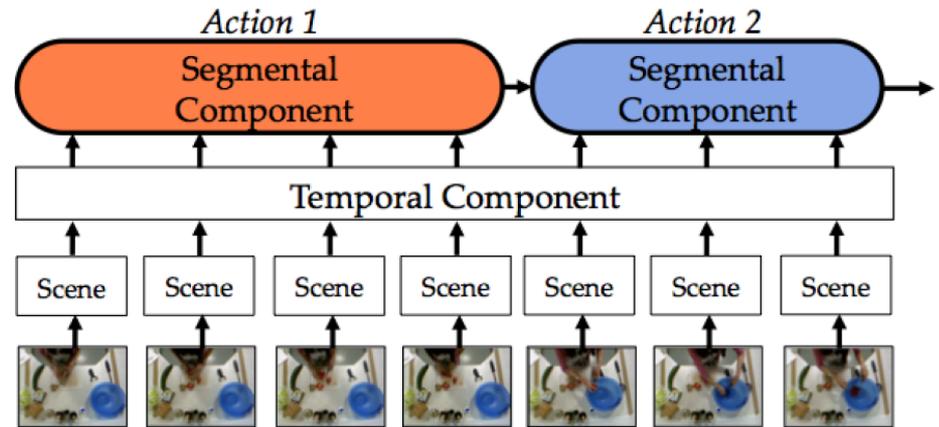
JOHNS HOPKINS  
MATHEMATICAL INSTITUTE  
for DATA SCIENCE

# Computer Vision: Future Vistas

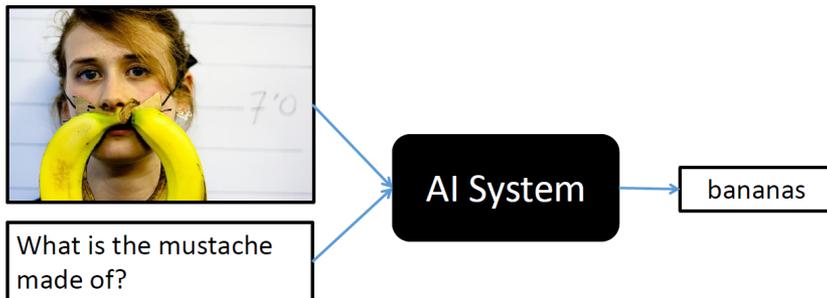
## Geometric Deep Learning



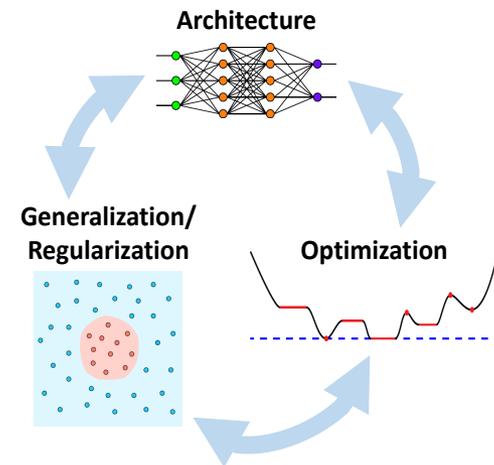
## Action Recognition: RNNs



## Vision and Language: Scene Parsing



## Deep Learning Theory



# Geometric Deep Learning

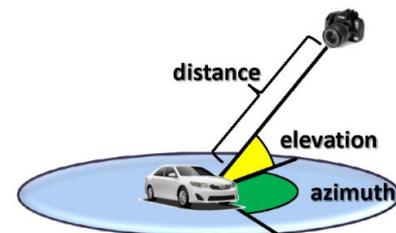
Renewed interest on joint object reconstruction and recognition

CVPR 2014



*Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models*, M. Aubry, D. Maturana, A. Efros, B. Russell and J. Sivic

WACV 2014

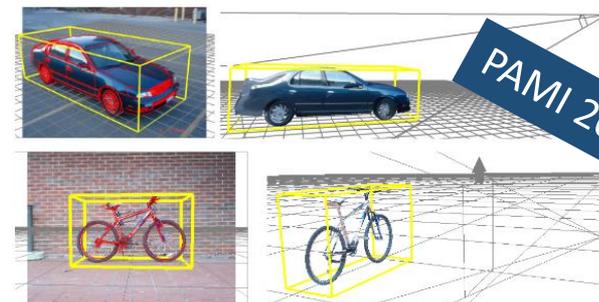


*Beyond PASCAL: A Benchmark for 3D Object Detection in the Wild*, Y. Xiang, R. Mottaghi and S. Savarese



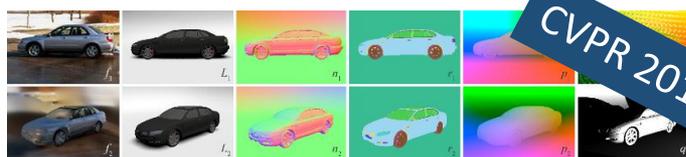
Siggraph 2014

*Estimating Image Depth Using Shape Collections*, H. Su, Q. Huang, N. Mitra, Y. Li and L. Guibas



PAMI 2013

*Detailed 3D Representations for Object Recognition and Modeling*, Z. Zia, M. Stark, B. Schiele and K. Schindler



CVPR 2014

*Image-based Synthesis and Re-Synthesis of Viewpoints Guided by 3D Models*. K. Rematas, T. Ritschel, M. Fritz, and T. Tuytelaars



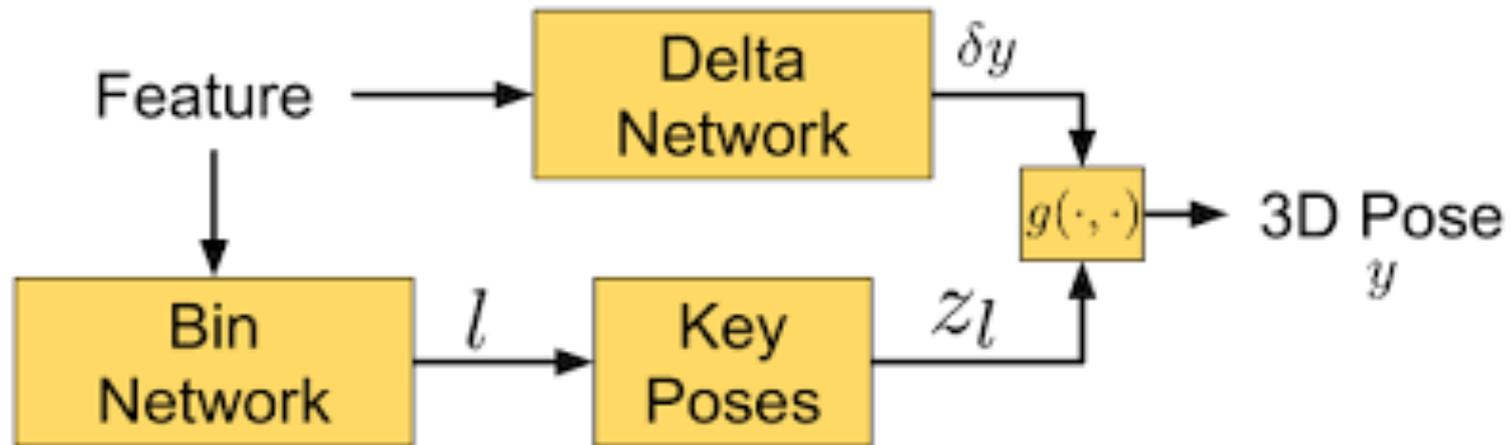
ICCV 2013

*Parsing IKEA objects: Fine Pose Estimation*. J. Lim, H. Pirsivash and A. Torralba

# Geometric Deep Learning: 3D Pose



Bounding box  
+  
Category label  
+  
3D orientation



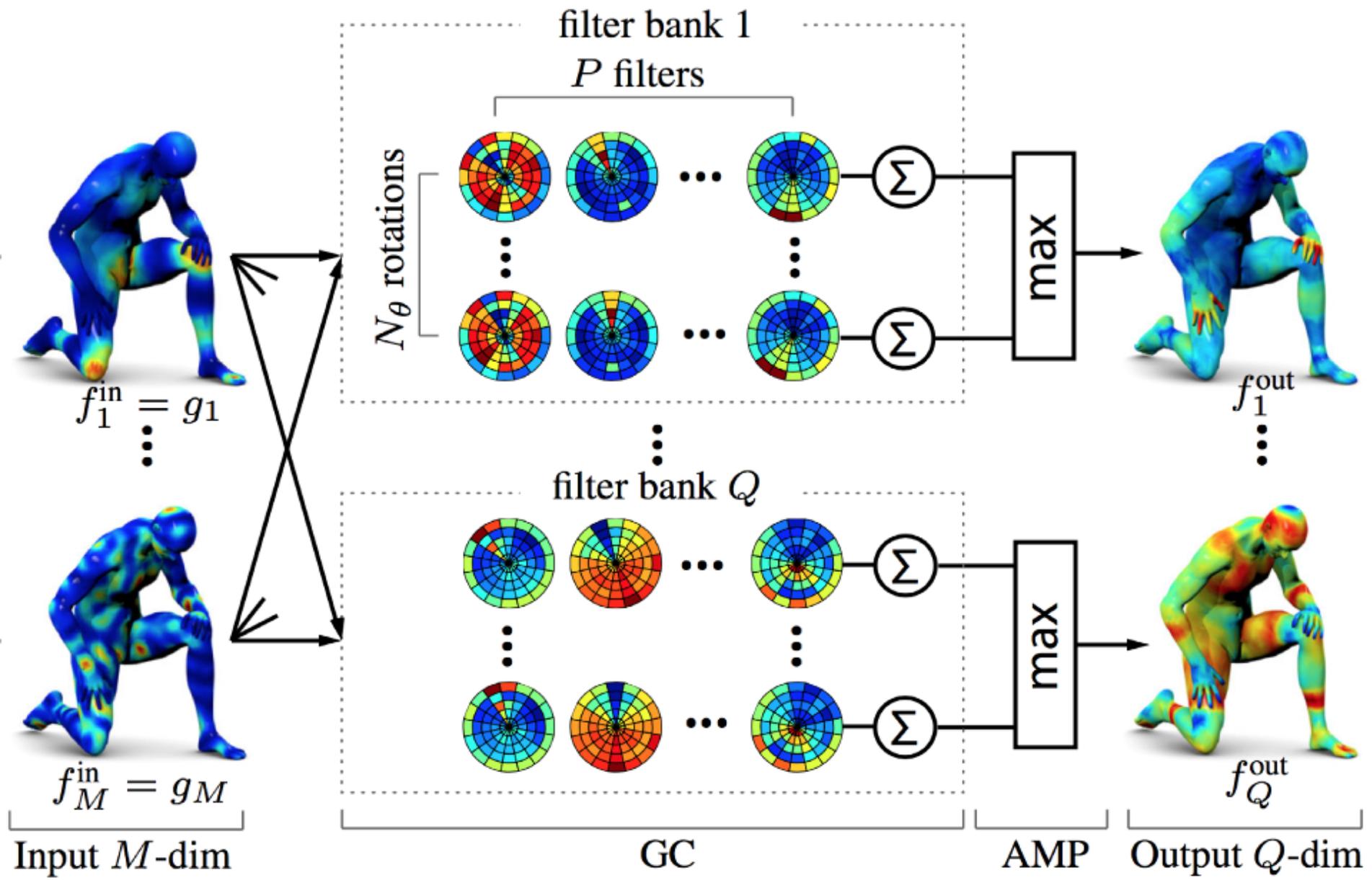
# Geometric Deep Learning: 3D Pose/Shape



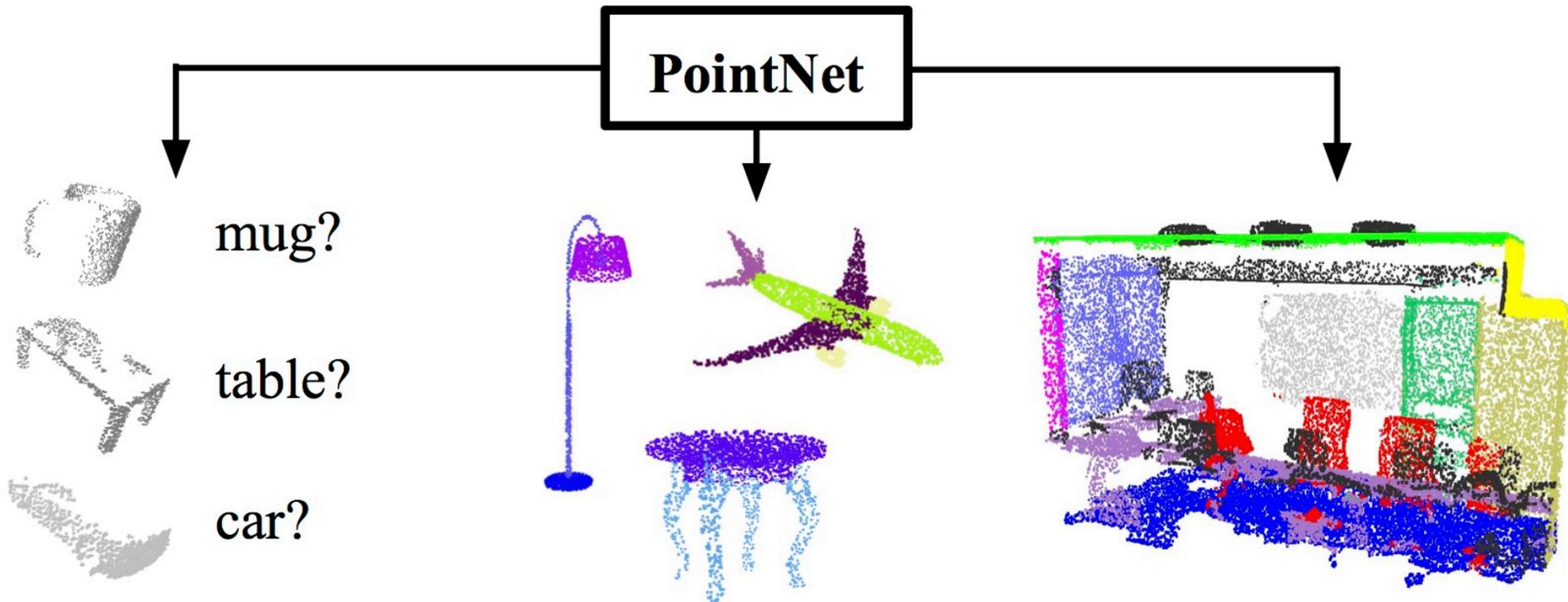
Bounding box  
+  
Category label  
+  
3D orientation



# Geometric Deep Learning: 3D Shape



# Geometric Deep Learning: 3D Point Clouds



# Geometric Deep Learning: Graph CNNs

## How Graph Convolutions work

CNN on image

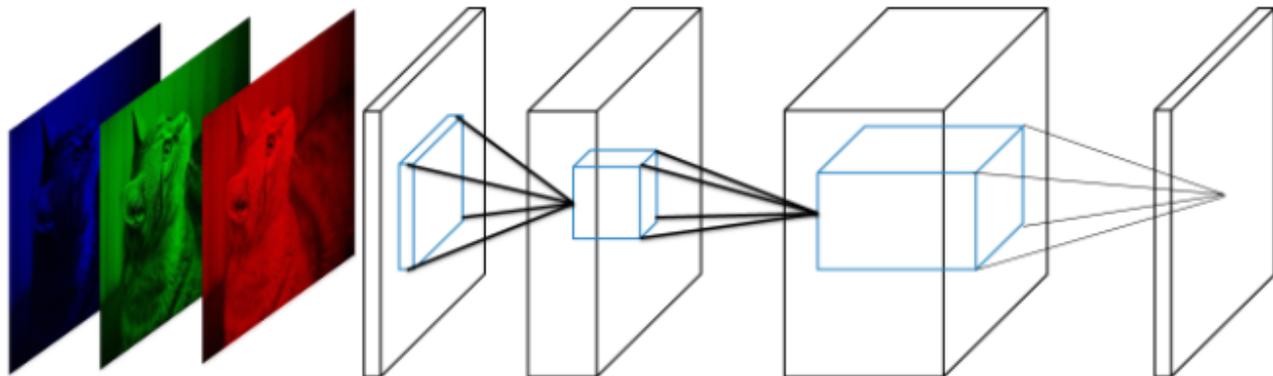
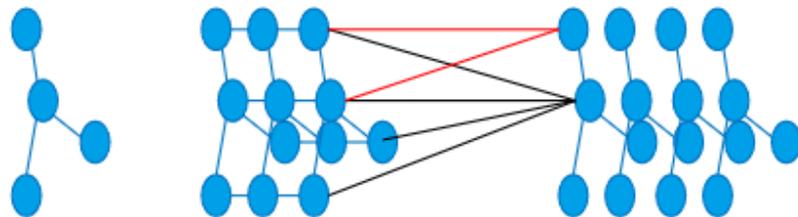
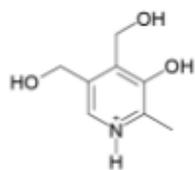


Image  
class label

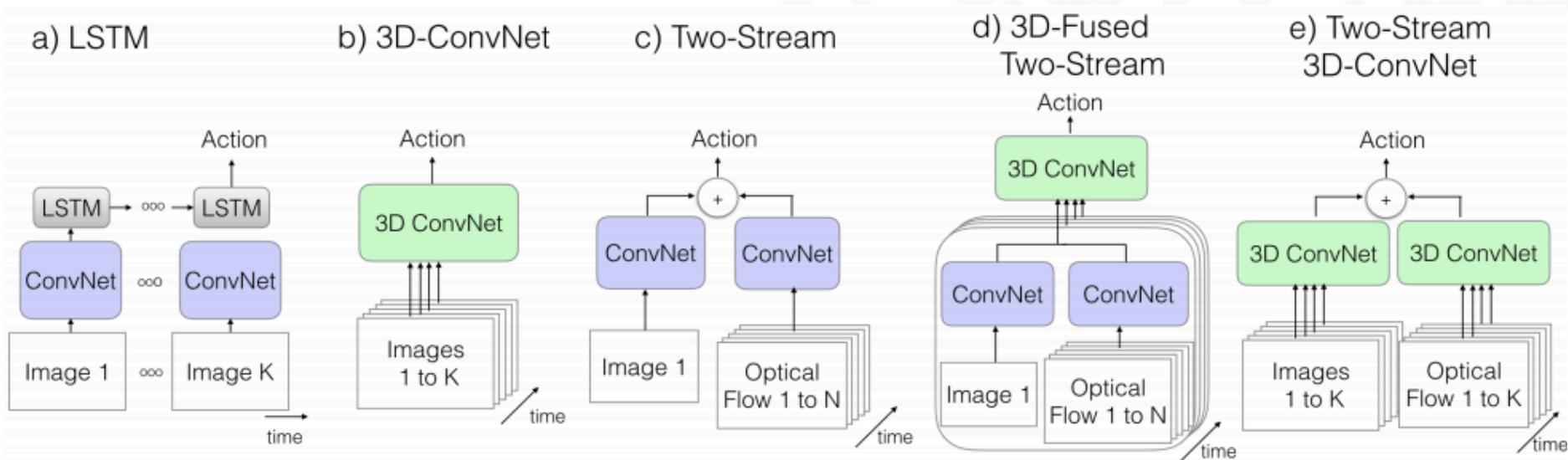
Graph convolution



Chemical  
property

Convolution "kernel" depends on Graph structure

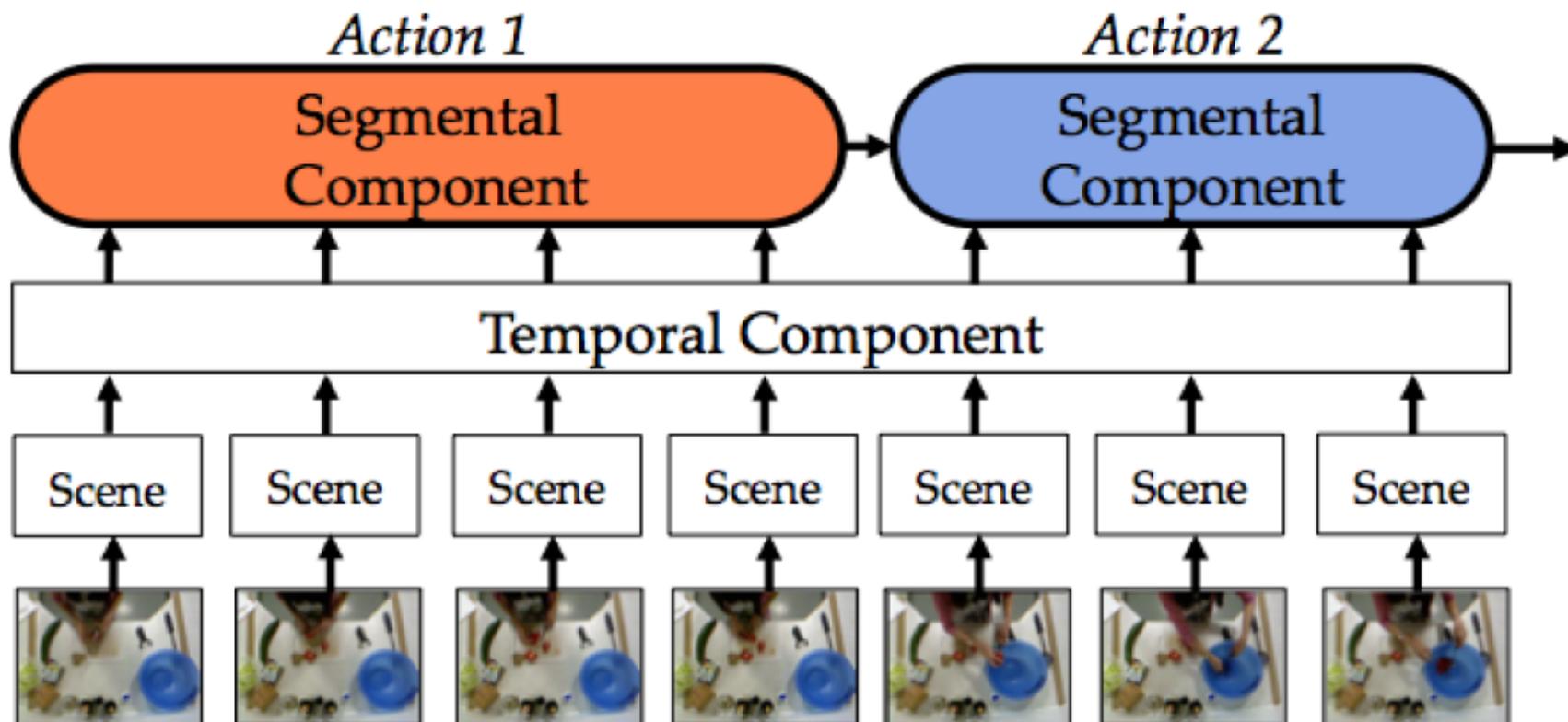
# Action Recognition



Architecture	UCF-101			HMDB-51			Kinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	63.3	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	56.1	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	62.2	52.4	65.6
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	–	–	67.2
<b>(e) Two-Stream I3D</b>	<b>84.5</b>	<b>90.6</b>	<b>93.4</b>	<b>49.8</b>	<b>61.9</b>	<b>66.4</b>	<b>71.1</b>	<b>63.4</b>	<b>74.2</b>

# Action Segmentation

- State-of-the-art methods for action classification, detection and segmentation rely on spatio-temporal deep networks.



[1] C. Lea, G. Hager, R. Vidal. An Improved Model for Segmentation and Recognition of Fine-Grained Activities. WACV 2015.

[2] C. Lea, R. Vidal, G. Hager. Learning Convolutional Action Primitives for Fine-grained Action Recognition. ICRA 2016.

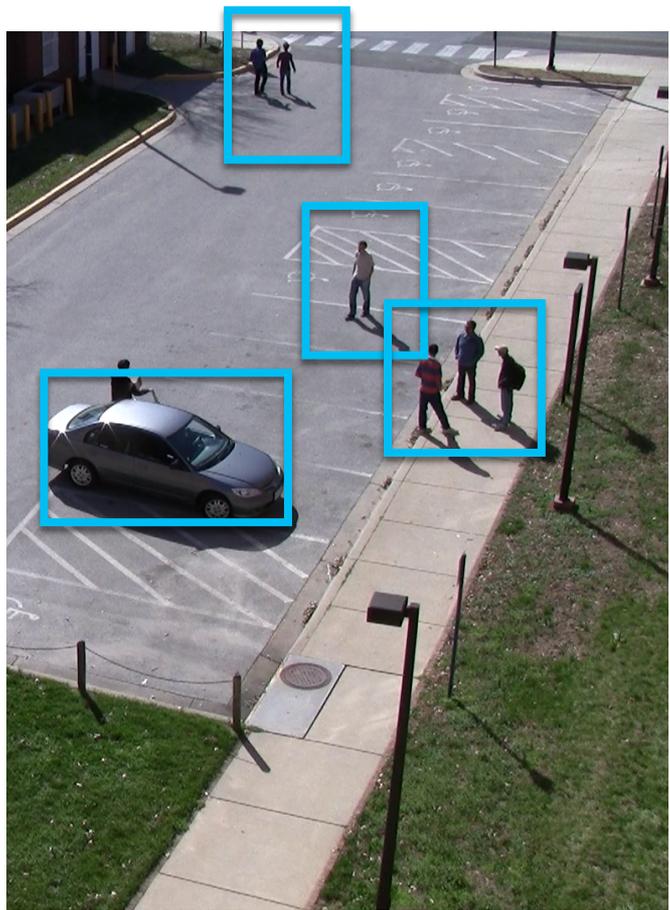
[3] C. Lea, A. Reiter, R. Vidal, G. Hager. Segmental Spatiotemporal CNNs for Fine-grained Action Segmentation. ECCV 2016.

[4] Tao, Vidal. Moving Poselets: A Discriminative and Interpretable Skeletal Motion Representation for Action Recognition. ICCVW 2015.

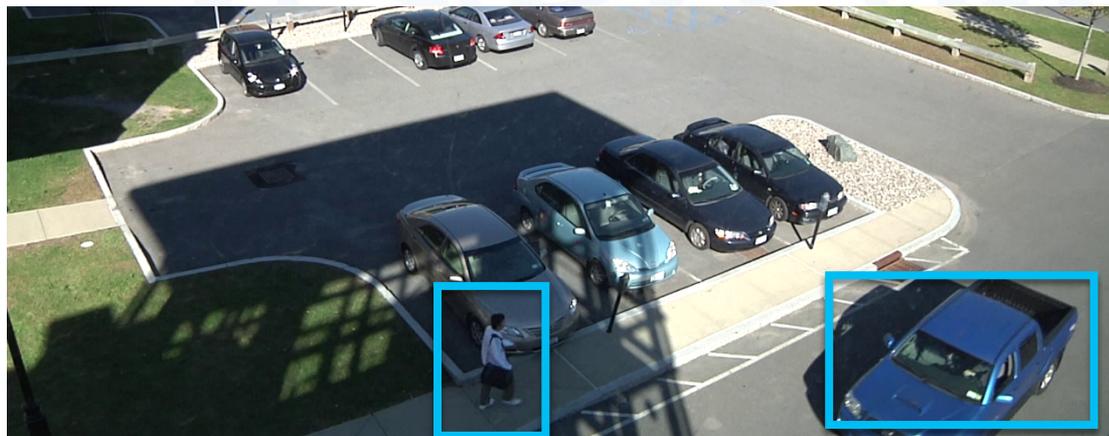
[5] Mavroudi, Tao, Vidal. Deep Moving Poselets for Video Based Action Recognition. WACV 2017.

[6] Mavroudi, Bhaskara, Sefati, Ali, Vidal. End-to-End Fine-Grained Action Segmentation and Recognition Using Conditional Random Field Models and Discriminative Sparse Coding. WACV, 2018.

# IARPA: Deep Intermodal Video Analytics (DIVA)



Multiple Activities  
at Multiple Scales

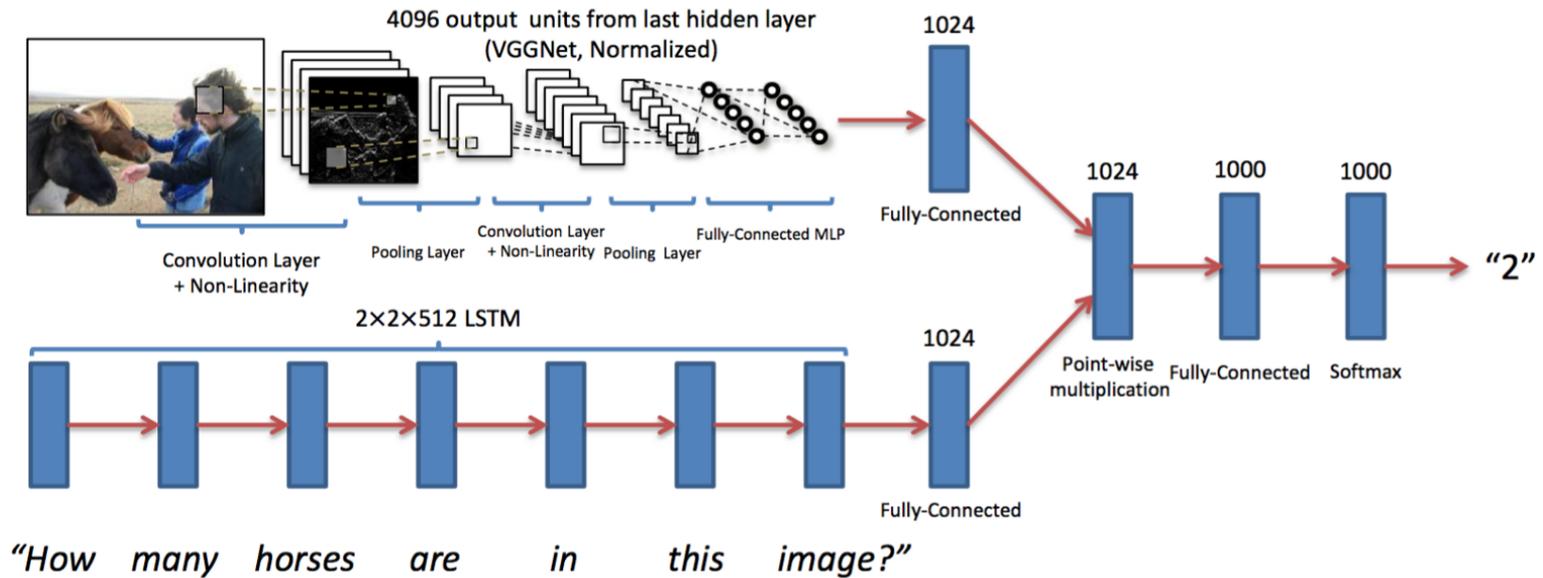
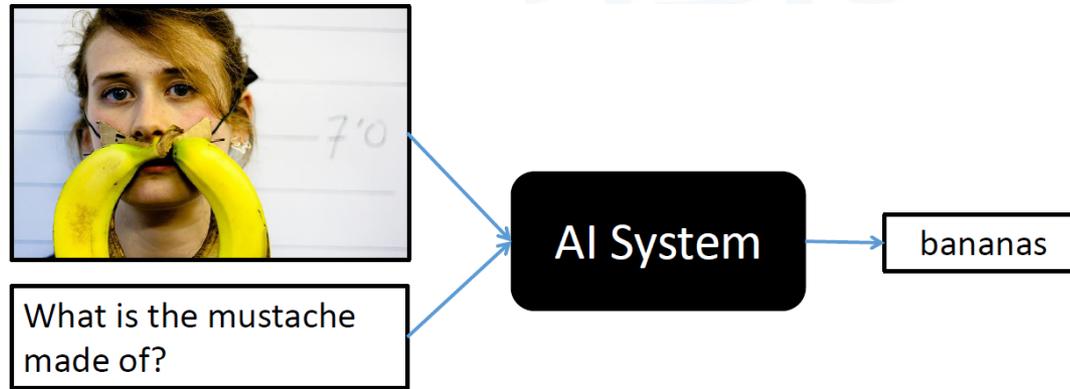


Multiple Actors, People and Vehicles



Varying Illumination

# Scene Parsing | Visual Question Answering



# MURI on Semantic Information Pursuit

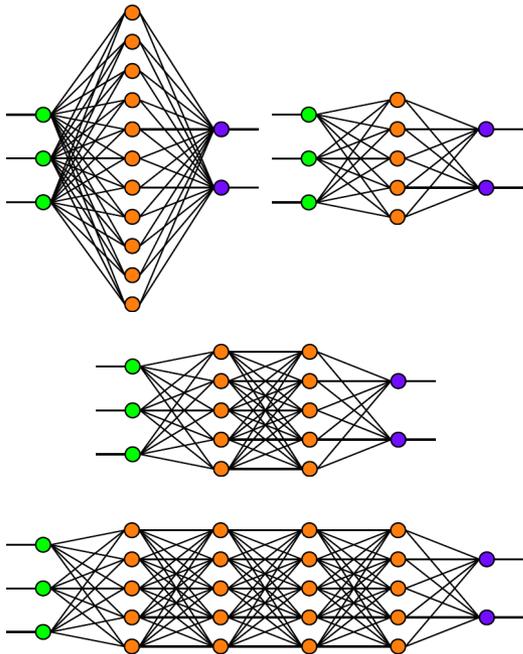
- Develop an information-theoretic framework for characterizing semantic information content in complex multimodal data.



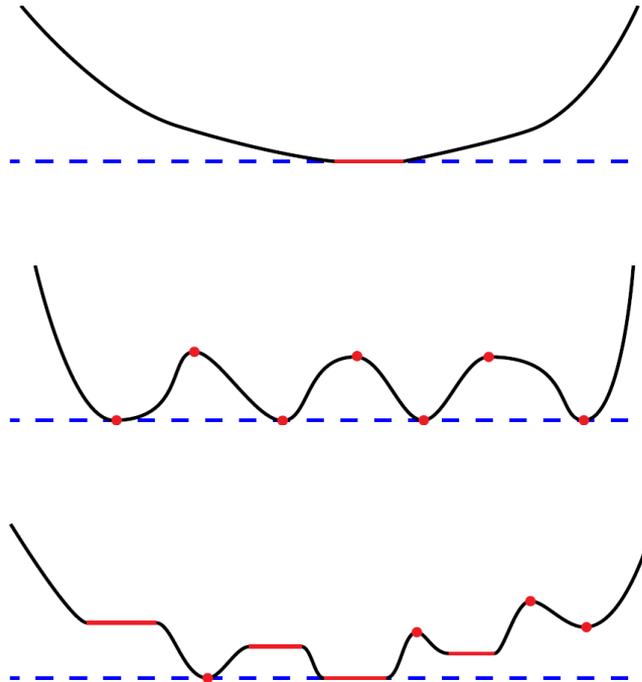
1. Q: Is there a person in the blue region? A: yes
2. Q: Is there a unique person in the blue region?  
(Label this person 1) A: yes
3. Q: Is person 1 carrying something? A: yes
4. Q: Is person 1 female? A: yes
5. Q: Is person 1 walking on a sidewalk? A: yes
6. Q: Is person 1 interacting with any other object? A: no
- :

# Key Theoretical Questions in Deep Learning

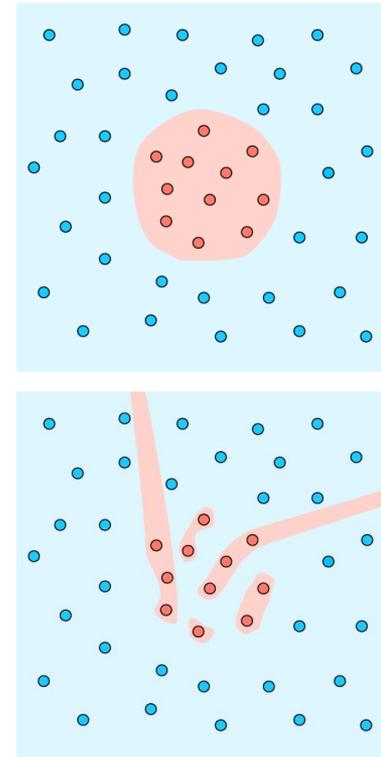
## Architecture Design



## Optimization

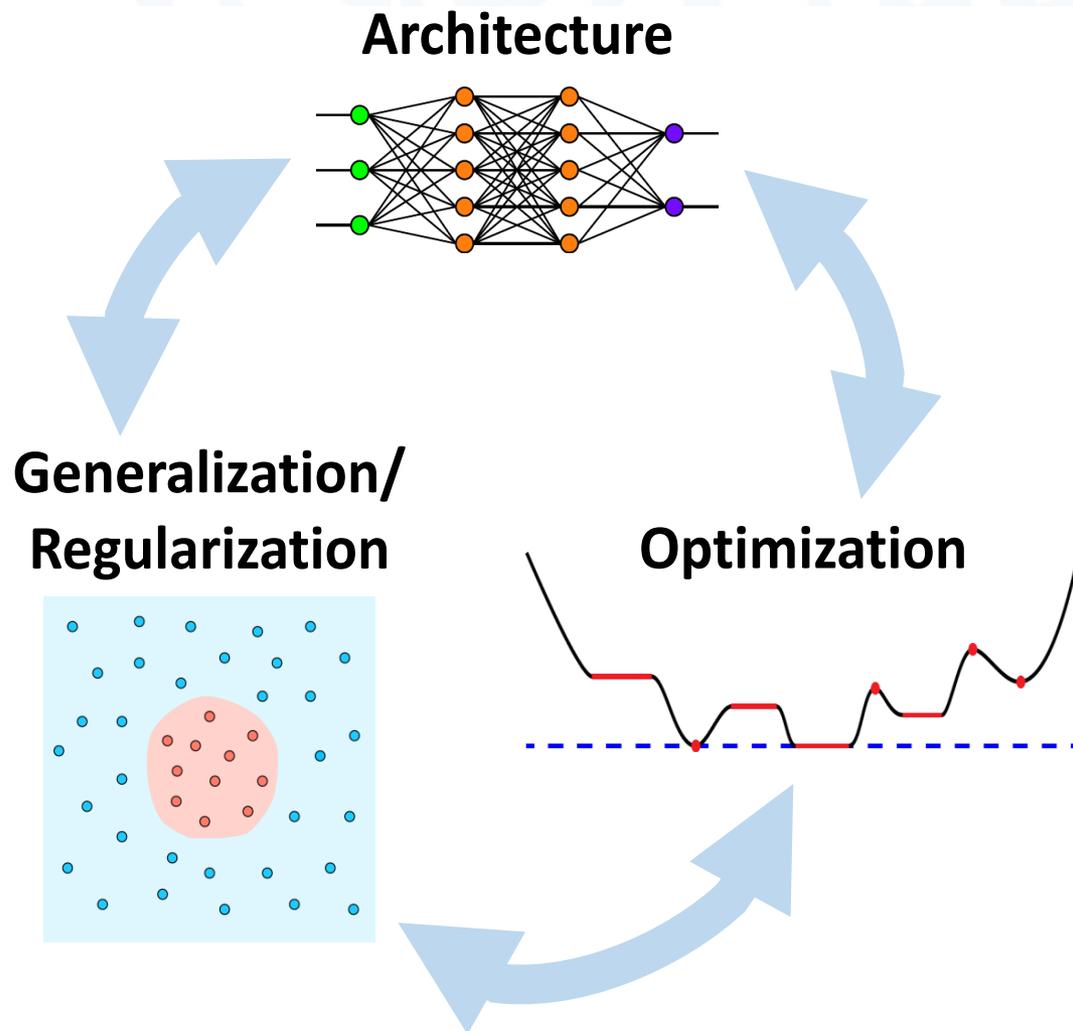


## Generalization



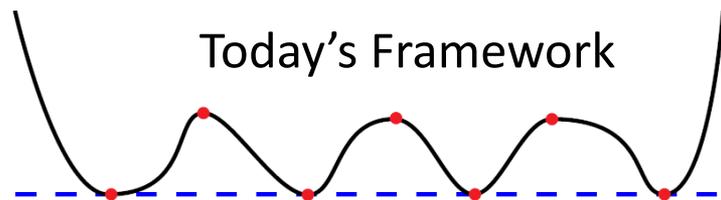
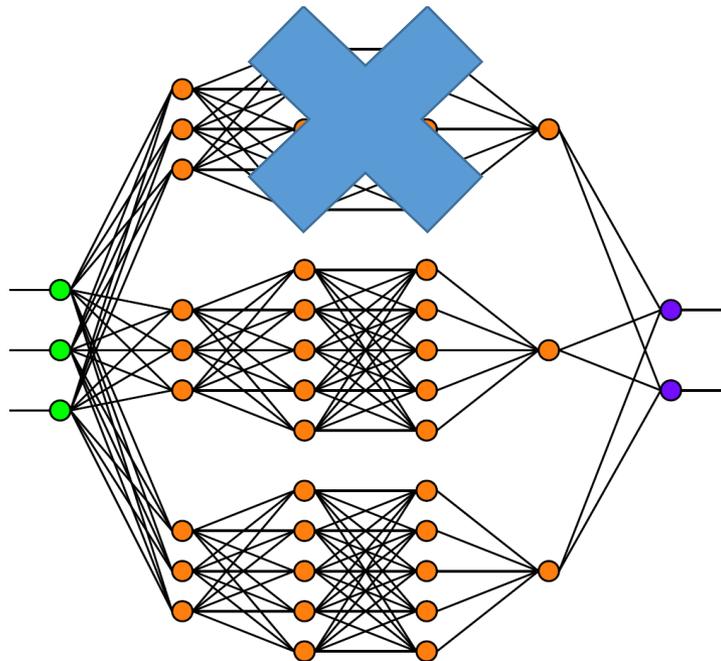
# Key Theoretical Questions are Interrelated

- Optimization can impact generalization [1,2]
- Architecture has strong effect on generalization [3]
- Some architectures could be easier to optimize than others [4]



[1] Neyshabur et. al. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning." ICLR workshop. (2015).  
[2] P. Zhou, J. Feng. The Landscape of Deep Learning Algorithms. 1705.07038, 2017  
[3] Zhang, et al., "Understanding deep learning requires rethinking generalization." ICLR. (2017).  
[4] Haeffele, Vidal. Global optimality in neural network training. CVPR 2017.

# Analysis of Optimization: Main Results



- **Questions:** What properties of the architecture and regularization function facilitate optimization?

- **Assumptions:**

- Parallel network structure.
- Positively homogeneous activations.
- Positively homogeneous regularizers.

- **Theorem 1:** A local minimum such that all weights from one subnet are zero is a global minimum.

- **Theorem 2:** If network size is large enough local descent can find global minimum from any initialization.

[1] Haeffele, Young, Vidal. Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing, ICML '14

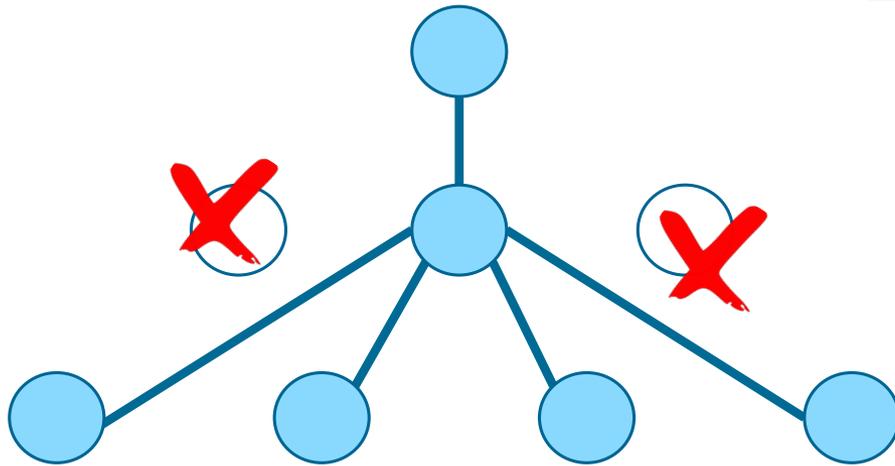
[2] Haeffele, Vidal. Global Optimality in Tensor Factorization, Deep Learning and Beyond, arXiv, '15

[3] Haeffele, Vidal. Global optimality in neural network training. CVPR 2017.

[4] Haeffele, Vidal. Structured Low-Rank Matrix Factorization: Global Optimality, Algorithms, and Applications. TPAMI 2018.



# Analysis of Dropout Regularization: Main Results



- **Question:** What type of regularization is induced by dropout?
- **Theorem 4:** Dropout induces explicit low-rank regularization (nuclear norm squared).

- **Question:** What objective function is being minimized by dropout?
- **Theorem 3:** Dropout is SGD applied to a stochastic objective.

- **Question:** What are the properties of the optimal weights?
- **Theorem 5:** Dropout induces balanced weights.

# Conclusions and Future Directions

- **Computer vision has rich a history of model-based and data-driven methods**
  - Object and view centered representations
  - Handcrafted and learned features
- **Recently remarkable progress of data driven methods**
  - Object and image classification, object detection, pose estimation
  - Semantic segmentation, generative adversarial networks
- **But still far from intelligence: need model-based + data driven methods**
  - Geometric deep learning, action recognition, scene parsing
  - Lifelong learning
  - Theory of CNNs, RNNs, GANs

# More Information,

JHU Vision Lab

<http://www.vision.jhu.edu/>

Mathematical Institute for Data Science @ JHU

<http://www.minds.jhu.edu>

# Thank You!



JOHNS HOPKINS  
MATHEMATICAL INSTITUTE  
*for* DATA SCIENCE