Apprentissage continu de représentations visuelles

ENSIMAG
2023-2024

Karteek Alahari & Diane Larlus

https://project.inria.fr/bigvisdata/

Informations

- Site web : https://project.inria.fr/bigvisdata

- Intervenants :
  - Karteek Alahari, Directeur de recherche, Inria
    karteek.alahari@inria.fr
  - Diane Larlus, Principal scientist, Naver Labs
    diane.larlus@naverlabs.com

- 8 x 1h30 + 2 x 3h = 18h de cours

Informations

- s42 – Jeudi 19/10/2023 – 11h15 à 12h45 – Karteek – H206 – cours
- s43 – Jeudi 26/10/2023 – 9h45 à 12h45 – Diane – H203 – cours (***attention cours de 3h***)
- s45 – Jeudi 09/11/2023 – 11h15 à 12h45 – Diane – D213 – Présentations Article 1 & 2 + Quizz
- s46 – Jeudi 16/11/2023 – 11h15 à 12h45 – Karteek – D211 – cours + présentation Article 3 + Quizz
- s47 – Jeudi 23/11/2023 – 11h15 à 12h45 – Diane – D211 – cours
- s49 – Jeudi 07/12/2023 – 11h15 à 12h45 – Diane – D211 – cours
- s50 – Jeudi 14/12/2023 – 9h45 à 12h45 – Diane + Karteek – D211 – Présentations Article 4 & 5 & 6 + Quizz + cours (***attention cours de 3h***)
- s51 – Jeudi 21/12/2023 – 11h15 à 12h45 – Karteek – cours + présentation Article 7 + Quizz
- s2 – Jeudi 11/01/2024 – 11h15 à 12h45 – Karteek – cours + présentation Article 8 + Quizz
- s3 – Jeudi 18/01/2024 – 11h15 à 12h45 – Karteek – cours + révisions

Informations

- Évaluation
  - Examen final écrit
  - Quizz sur des articles de recherche
  - https://project.inria.fr/bigvisdata/grading
- Points bonus
  - Presentation de papier
  - Voir la liste : https://project.inria.fr/bigvisdata/presentations
  - Votre choix par email

Informations

- Article 1 – **Unsupervised Domain Adaptation by Backpropagation** (UDA). ICML 2015 [pdf]
- Article 2 – **Unsupervised Representation Learning by Predicting Image Rotations** (RotNet). ICLR 2018 [pdf]
- Article 3 – **Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset** (Quo Vadis). CVPR 2017 [pdf]
- Article 4 – **Momentum Contrast for unsupervised visual representation learning** (MoCo). CVPR 2020 [pdf]
- Article 5 – **Learning Transferable Visual Models From Natural Language Supervision** (CLIP). ICML 2021 [pdf] (only sections 1, 2, 3.1.1, 3.1.2, 3.1.3)
- Article 6 – **Masked Autoencoders Are Scalable Vision Learners** (MAE). CVPR 2022 [pdf]
- Article 7 – **Learning without Forgetting** (LwF). ECCV 2016 [pdf]
- Article 8 – **Incremental Learning of Object Detectors without Catastrophic Forgetting** (IncDet). ICCV 2017 [pdf]

Les données à grande échelle

Wikipedia

Le *big data*

- littéralement « grosses données »
- **méga données**
- **données massives**

désigne un ensemble de données qui sont tellement volumineuses qu'elles en deviennent difficiles à travailler avec des outils classiques de gestion de base de données

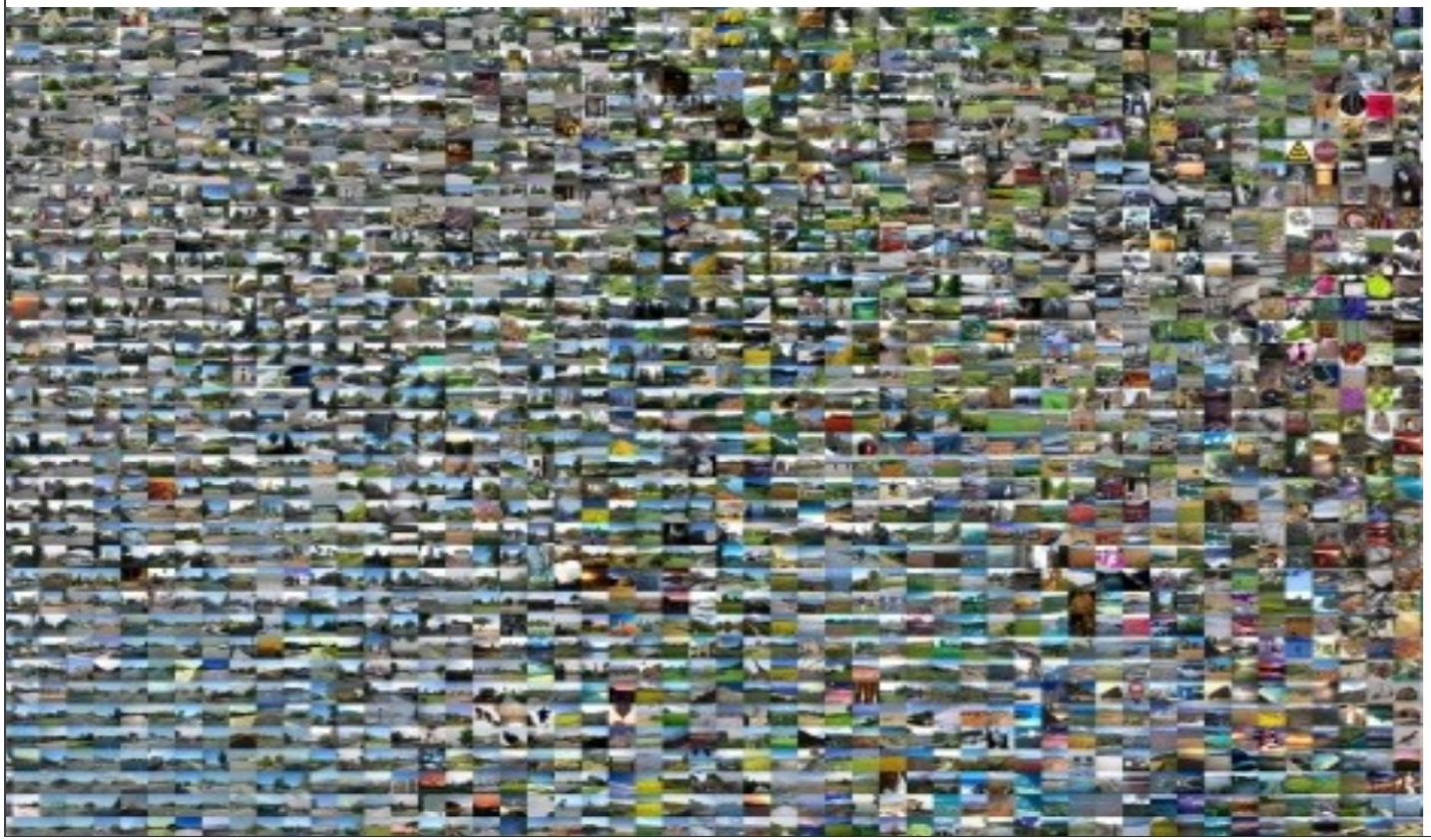
- Nécessite le développement d'outils spécifiques

Slide courtesy : D. Larlus

Les données visuelles à grande échelle

ou ***Big Visual Data***

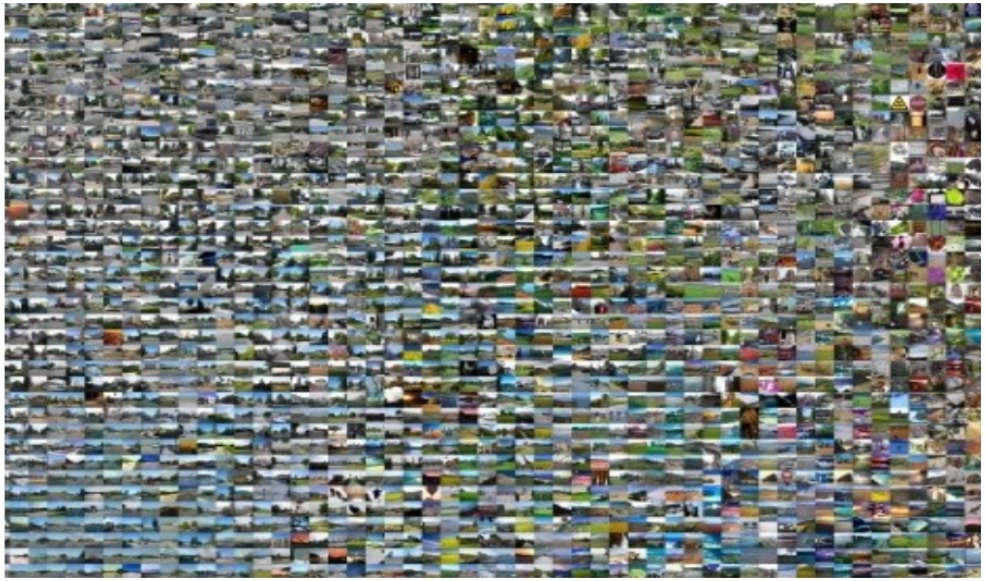
sont devenues une façon majeure de transférer l'information



Slide courtesy : D. Larlus

Les données visuelles… qui évoluent

- Croissance très importante, en raison de l'accumulation des contenus numériques auto-produits par le grand public
  - **ImDb** recense plus de 400 000 films
  - Images (semi-)pro : **Corbis**, **Getty**, **Fotolia**
    - centaines de milliers d' images
  - **Facebook** – chiffres de 2013
    - 350 millions de nouvelles photos / jour
    - 250 milliards de photos stockées
    - 3,000 années de vidéo générées chaque jour
  - **Flickr** – novembre 2016
    - 13 milliards de photos
  - **Youtube**
    - 35h / min uloaded in 2011
    - 100h / min uploaded in 2014



2022 : Flickr – 25 millards, YouTube – 400h / min

Slide courtesy : D. Larlus

Exemples d'applications du Big Visual Data

- News/Films à la demande
- Commerce électronique
- Informations médicales
- Systèmes d'informations géographiques
- Architecture/Design
- Protection du copyright / traçage de contenu
- Géolocalisation, système de navigation
- Enquêtes policières
- Militaire
- Expérimentations scientifiques
- Enseignement
- Archivage, gestion des bases de données de contenu (personnelles ou professionnelles)
- Moteur de recherche (Internet, collections personnelles)
- Voitures et autres véhicules autonomes
- Robotique, plateformes d'intelligence ambiante
- Autres applications industrielles
- Etc.

Slide courtesy : D. Larlus

Chaîne du Big Visual Data

1. Génération
   - Outils de production et de création
2. Représentation
   - Utilisation de formats de représentation différents
3. Stockage
4. Transmission
   - Problème de réseaux, architecture
5. Recherche d'information
   - Taches d'analyse d'images
6. Distribution
   - Conception de serveur de streaming, interfaces d'application, etc.

Slide courtesy : D. Larlus

Chaîne du Big Visual Data

1. Génération
   - Outils de production et de création
2. Représentation
   - Utilisation de formats de représentation différents
3. Stockage
4. Transmission
   - Problème de réseaux, architecture
5. **Recherche d'information**
   - Taches d'analyse d'images
6. Distribution
   - Conception de serveur de streaming, interfaces d'application, etc.

Slide courtesy : D. Larlus

Dans ce cours

- Apprentissage supervisé (*supervised learning*)
  - Variantes, ex. Semi-supervisé (*semi-supervised*)
- Apprentissage auto-supervisé (*self-supervised*)
- L'adaptation de domaine (*domain adaptation*)
- Apprentissage continu (*continual learning*)
- Problèmes en vidéos

Dans ce cours

- **Apprentissage supervisé** (*supervised learning*)
  - Variantes, ex. Semi-supervisé (*semi-supervised*)
- Apprentissage auto-supervisé (*self-supervised*)
- L'adaptation de domaine (*domain adaptation*)
- Apprentissage continu (*continual learning*)
- Problèmes en vidéos

Supervised Learning

- Input: x (images, text, emails…)
- Output: y (spam or non-spam…)
- (Unknown) Target Function
  - f: X → Y (the "true" mapping / reality)
- Data
  - (x_1,y_1), (x_2,y_2), …, (x_N,y_N)
- Model / Hypothesis Class
  - g: X → Y
  - $y = g(x) = \text{sign}(w^T x)$
- Learning = Search in hypothesis space
  - Find best g in model class

Slide courtesy : D. Batra

Basic Steps of Supervised Learning

- **Set up** a supervised learning problem

- **Data collection**
  - Start with training data for which we know the correct outcome provided by a teacher or oracle

- **Representation**
  - Choose how to represent the data

- **Modeling**
  - Choose a hypothesis class: $H = \{g: X \rightarrow Y\}$

- **Learning/Estimation**
  - Find best hypothesis you can in the chosen class

- **Model Selection**
  - Try different models. Picks the best one. (More on this later)

- If happy stop
  - Else refine one or more of the above

Slide courtesy : D. Batra

Basic Steps of Supervised Learning

- **Set up** a supervised learning problem

- **Data collection**
  - Start with training data for which we know the correct outcome provided by a teacher or oracle

- **Representation**
  - Choose how to represent the data

- **Modeling**
  - Choose a hypothesis class: $H = \{g: X \rightarrow Y\}$

- **Learning/Estimation**
  - Find best hypothesis you can in the chosen class

- **Model Selection**
  - Try different models. Picks the best one. (More on this later)

- If happy stop
  - Else refine one or more of the above

Slide courtesy : D. Batra

Annotation d'images : difficulté et ambiguïté

- Contenu de métadonnées
  - Les données "brutes" (fichier image, fichier son) contiennent des informations sémantiques = directement compréhensibles pour l'utilisateur

- Ces métadonnées proviennent
  - Soit de propriétés de descripteurs d'objets (ex. Couleur moyenne d'une image)
  - Soit de données d'autres médias (ex. GPS)
  - Soit d'annotations manuelles (ex. Tags)

Slide courtesy : D. Larlus

Exemple : Exchangeable image file format (Exif)

- Spécification pour les formats d'images des appareils numériques
  - non uniformisé, mais largement utilisé
- Pour JPEG, TIFF, RIFF, ne supporte pas, PNG ou GIF
- Le format supporte souvent
  - Date et heure, enregistrés par l'appareil
  - Les paramètres de l'appareil
    - Dépendent du modèle : inclus la marque et des informations diverses telles que le temps d'ouverture, l'orientation, la focale, l'ISO, etc.
  - Une vignette de prévisualisation
  - La description et les informations de copyright
  - Les coordonnées GPS
- Ce format est supporté par de nombreuses applications

Slide courtesy : D. Larlus

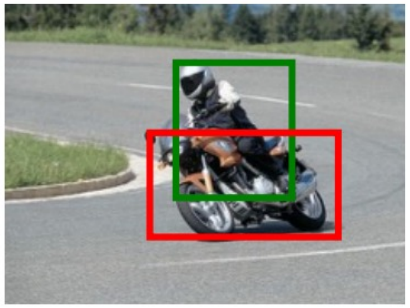
Contenu et métadonnées
… vs acquisitions d'annotations spécifiques

- Il est parfois nécessaire d'acquérir des annotations spécifiques à la tâche qui nous intéresse
- On parle de données labélisées
- Elles sont souvent associées à l'apprentissage supervisé
  - Dans ce cas elles sont utilisées comme données d'apprentissage pour apprendre un modèle de prédiction utile à la tâche à résoudre
    (Des exemples de tâches automatisables seront données juste après)
- L'apprentissage supervisé est un exemple d'apprentissage machine

Slide courtesy : D. Larlus

Annotation d'images

- Il faut choisir un type d'annotation
  - Ensemble de *tags* / étiquettes (une ou plusieurs étiquettes par image)
  - Position approximative de tous les objets (boites englobantes)
  - Position précise de tous les objets (masques de segmentation)
  - Phrases descriptives
- Il faut une cohérence des annotations sur toute une base



motorcyclist turning right

Slide courtesy : D. Larlus

Ambiguïté de l'annotation d'images

- Difficile de se mettre d'accord sur les annotations
- Exemple: Instructions pour la création d'une vérité terrain pour la compétition PASCAL 2009

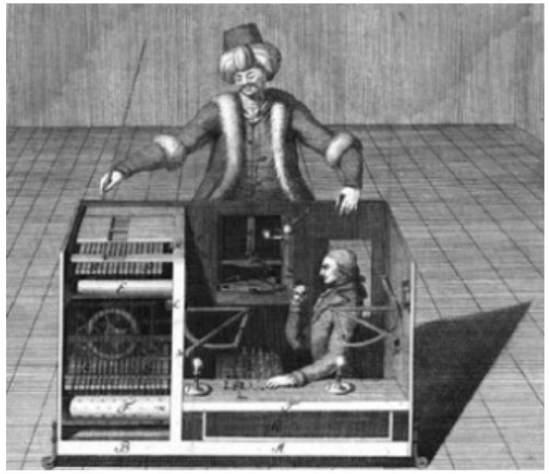
What to label	*All objects of the defined categories*, unless: you are unsure what the object is. the object is very small (at your discretion). less than 10-20% of the object is visible. If this is not possible because too many objects, mark image as bad.
Viewpoint	Record the viewpoint of the 'bulk' of the object e.g. the body rather than the head. Allow viewpoints within 10-20 degrees. If ambiguous, leave as 'Unspecified'. Unusually rotated objects e.g. upside-down people should be left as 'Unspecified'.
Bounding box	Mark the bounding box of the visible area of the object (*not* the estimated total extent of the object). Bounding box should contain all visible pixels, except where the bounding box would have to be made excessively large to include a few additional pixels (<5%) e.g. a car aerial.
Truncation	If more than 15-20% of the object lies outside the bounding box mark as Truncated. The flag indicates that the bounding box does not cover the total extent of the object.
Occlusion	If more than 5% of the object is occluded within the bounding box, mark as Occluded. The flag indicates that the object is not totally visible within the bounding box.

➢ Malgré ces instructions, on observe tout de même des incohérences dans les annotations

Slide courtesy : D. Larlus

Ambiguïté de l'annotation d'images

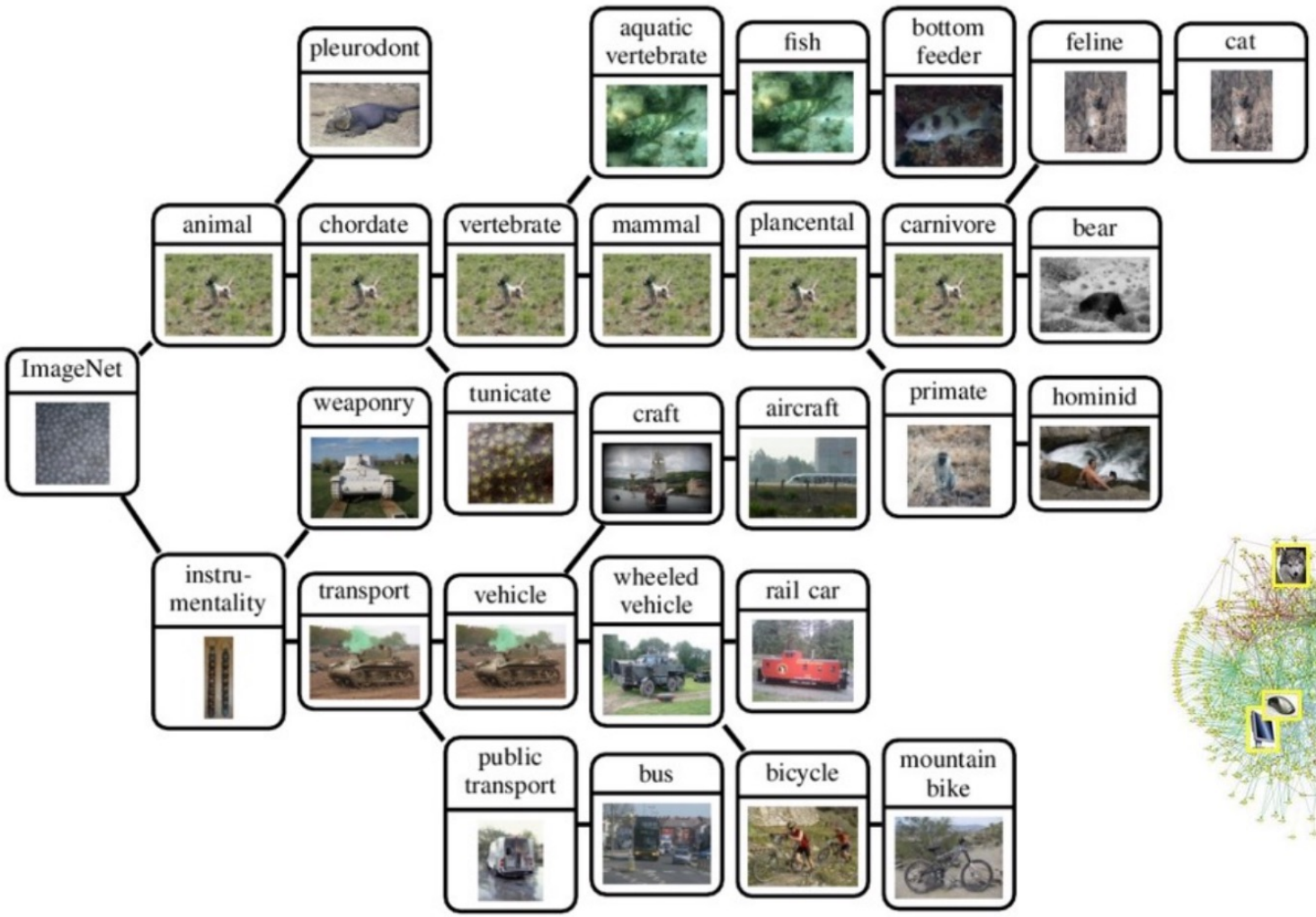
- Obtenir des annotations précises est une tâche fastidieuse
- Une solution courante: plateformes de ***crowdsourcing***, comme Amazon Mechanical Turk
  - *Crowdsourcing* : **externalisation ouverte** ou **production participative** en français
- *Amazon Mechanical Turk* (AMT)
  - Origine du nom : le **Turc mécanique** ou l'**automate joueur d'échecs** est un célèbre canular construit à la fin du XVIII^{e} siècle : il s'agissait d'un prétendu automate doté de la faculté de jouer aux échecs



- L'utilisation d'*Amazon Mechanical Turk* a permis la construction de bases d'images annotées qui ont grandement participé à l'avancée de la recherche en vision par ordinateur, par exemple:
  - ImageNet (http://image-net.org/)
  - MS Coco (http://mscoco.org/)
  - Visual Genome (https://visualgenome.org/)
  - Base VQA (http://www.visualqa.org/)

Slide courtesy : D. Larlus

ImageNet

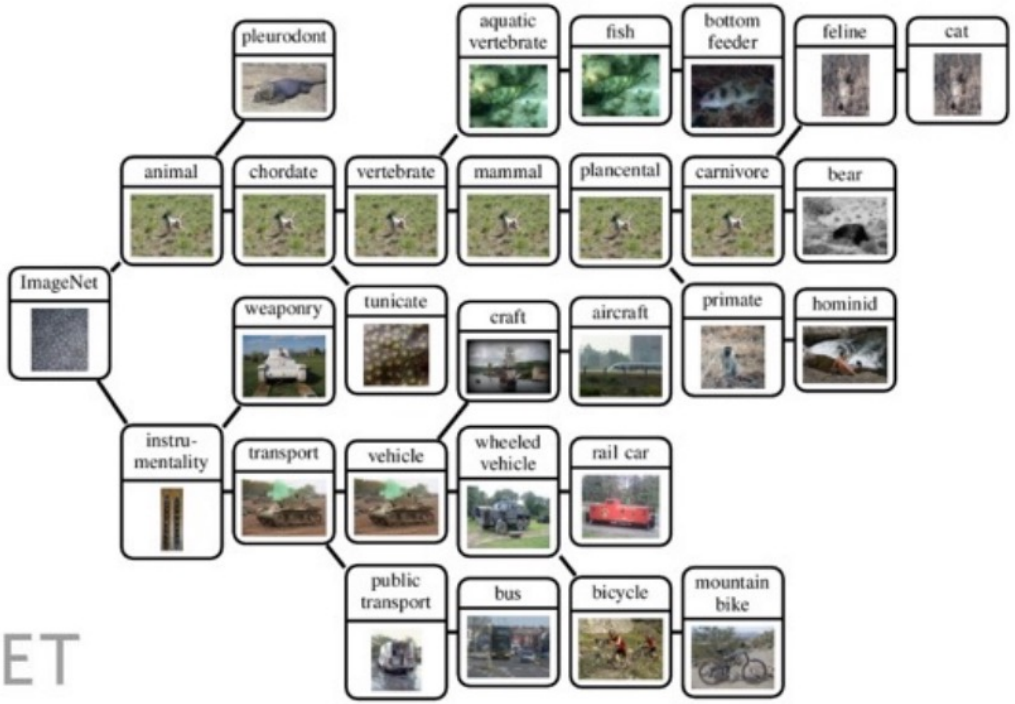


Slide courtesy : D. Larlus

ImageNet : pushing the state of the art

- ImageNet challenge:
  - Task: Categorize images into 1000 classes

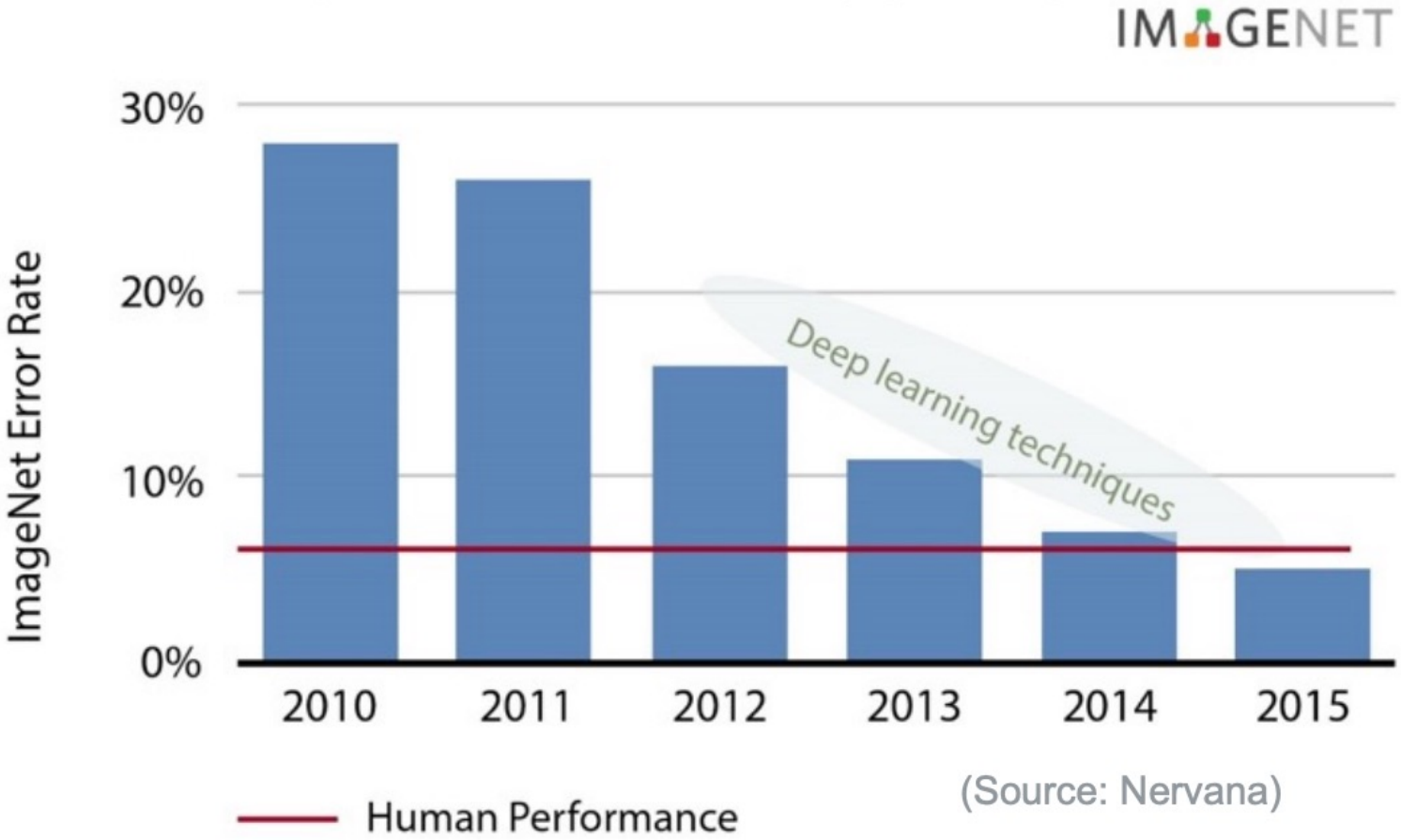
IMAGENET



Slide courtesy : D. Larlus

ImageNet : pushing the state of the art

- ImageNet challenge:
  - Task: Categorize images into 1000 classes
  - Until 2011: “standard” techniques (Fisher Vector)
  - Starting 2012: deep learning (CNNs)



Slide courtesy : D. Larlus

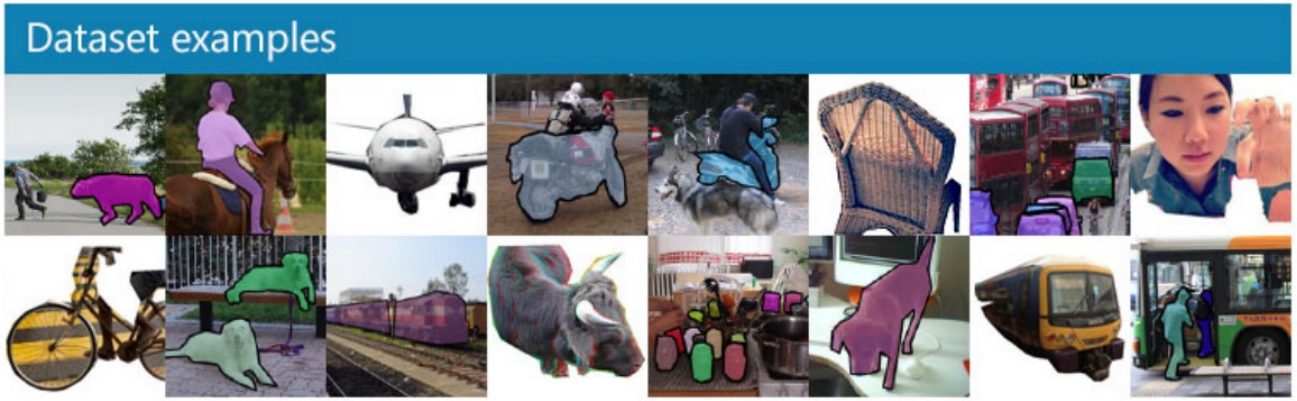
MS COCO

- *Common Objects in Contexts*

"What is COCO?" from its webpage

COCO is a large-scale object detection, segmentation, and captioning dataset
COCO has several features:

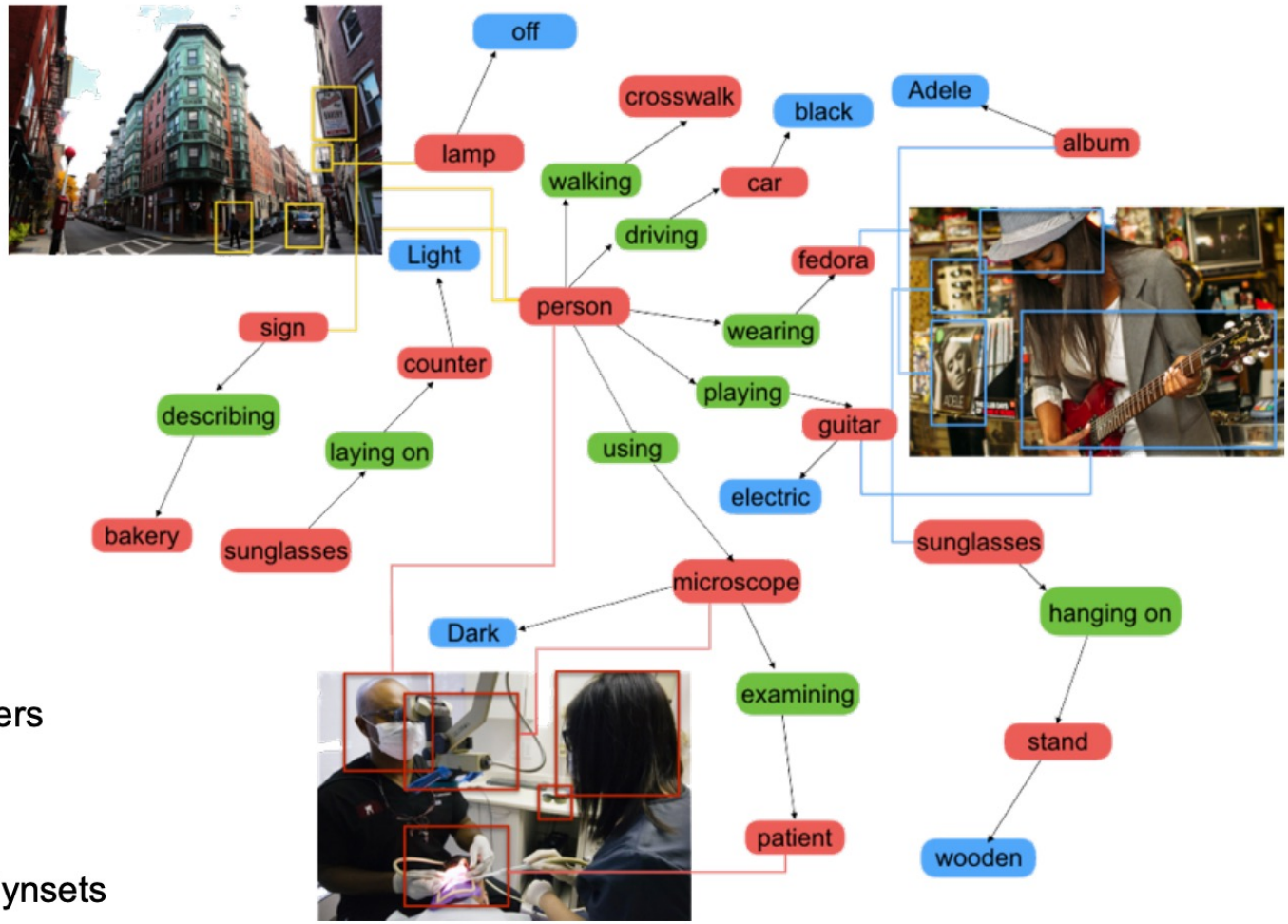
- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints



http://cocodataset.org

Slide courtesy : D. Larlus

Visual Genome



- 108,077 Images
- 5.4 Million Region Descriptions
- 1.7 Million Visual Question Answers
- 3.8 Million Object Instances
- 2.8 Million Attributes
- 2.3 Million Relationships
- Everything Mapped to Wordnet Synsets

Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li Jia-Li, David Ayman Shamma, Michael Bernstein, Li Fei-Fei

Slide courtesy : D. Larlus

Basic Steps of Supervised Learning

- **Set up** a supervised learning problem

- **Data collection**
  - Start with training data for which we know the correct outcome provided by a teacher or oracle

- **Representation**
  - Choose how to represent the data

- **Modeling**
  - Choose a hypothesis class: $H = \{g: X \rightarrow Y\}$

- **Learning/Estimation**
  - Find best hypothesis you can in the chosen class

- **Model Selection**
  - Try different models. Picks the best one. (More on this later)

- If happy stop
  - Else refine one or more of the above

Slide courtesy : D. Batra

Basic Steps of Supervised Learning

- **Set up** a supervised learning problem

- **Data collection**
  - Start with training data for which we know the correct outcome provided by a teacher or oracle

- **Representation**
  - Choose how to represent the data

- **Modeling**
  - Choose a hypothesis class: $H = \{g: X \rightarrow Y\}$

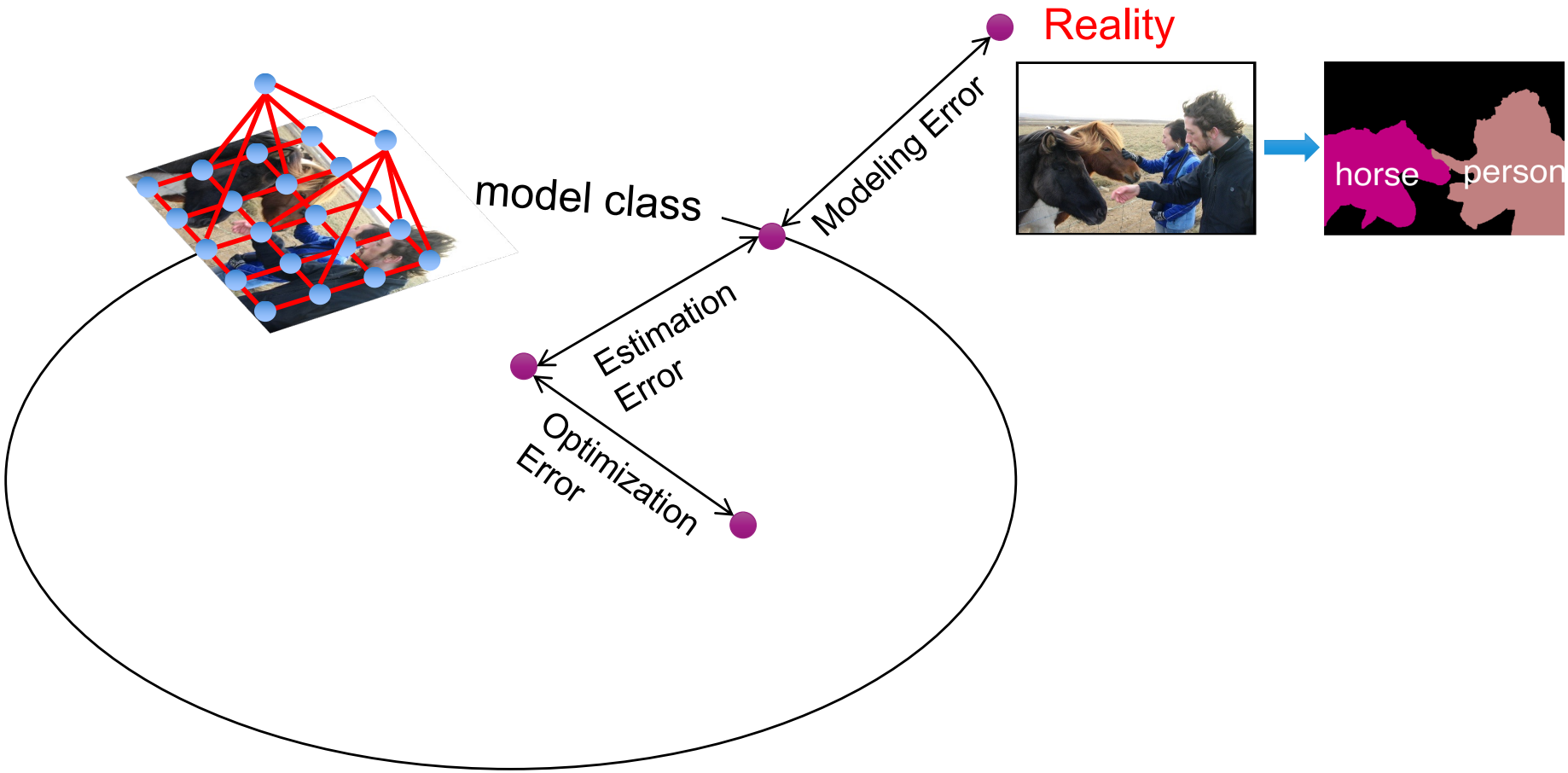
- **Learning/Estimation**
  - Find best hypothesis you can in the chosen class

- **Model Selection**
  - Try different models. Picks the best one. (More on this later)

- If happy stop
  - Else refine one or more of the above

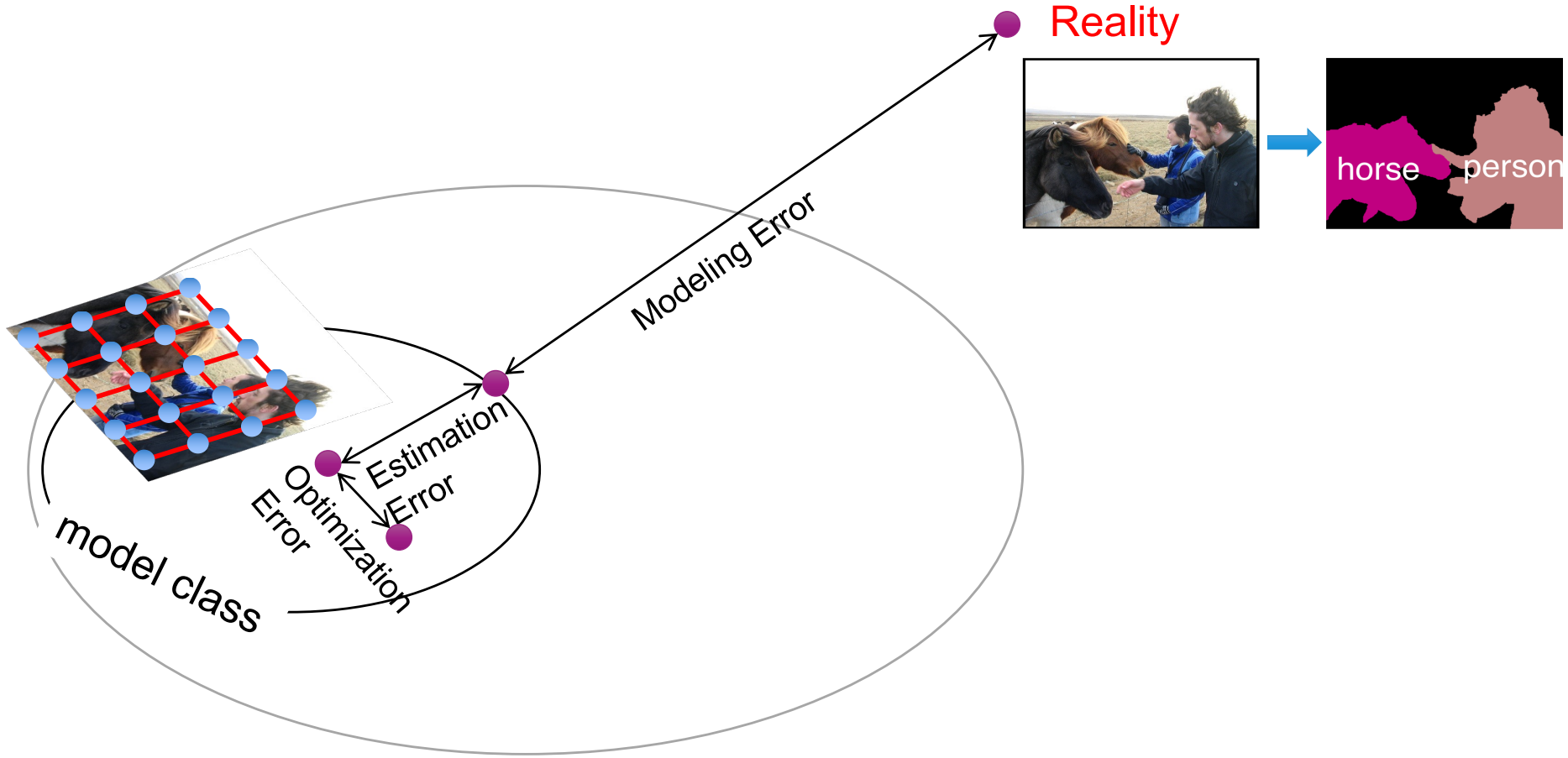
Slide courtesy : D. Batra

Error Decomposition



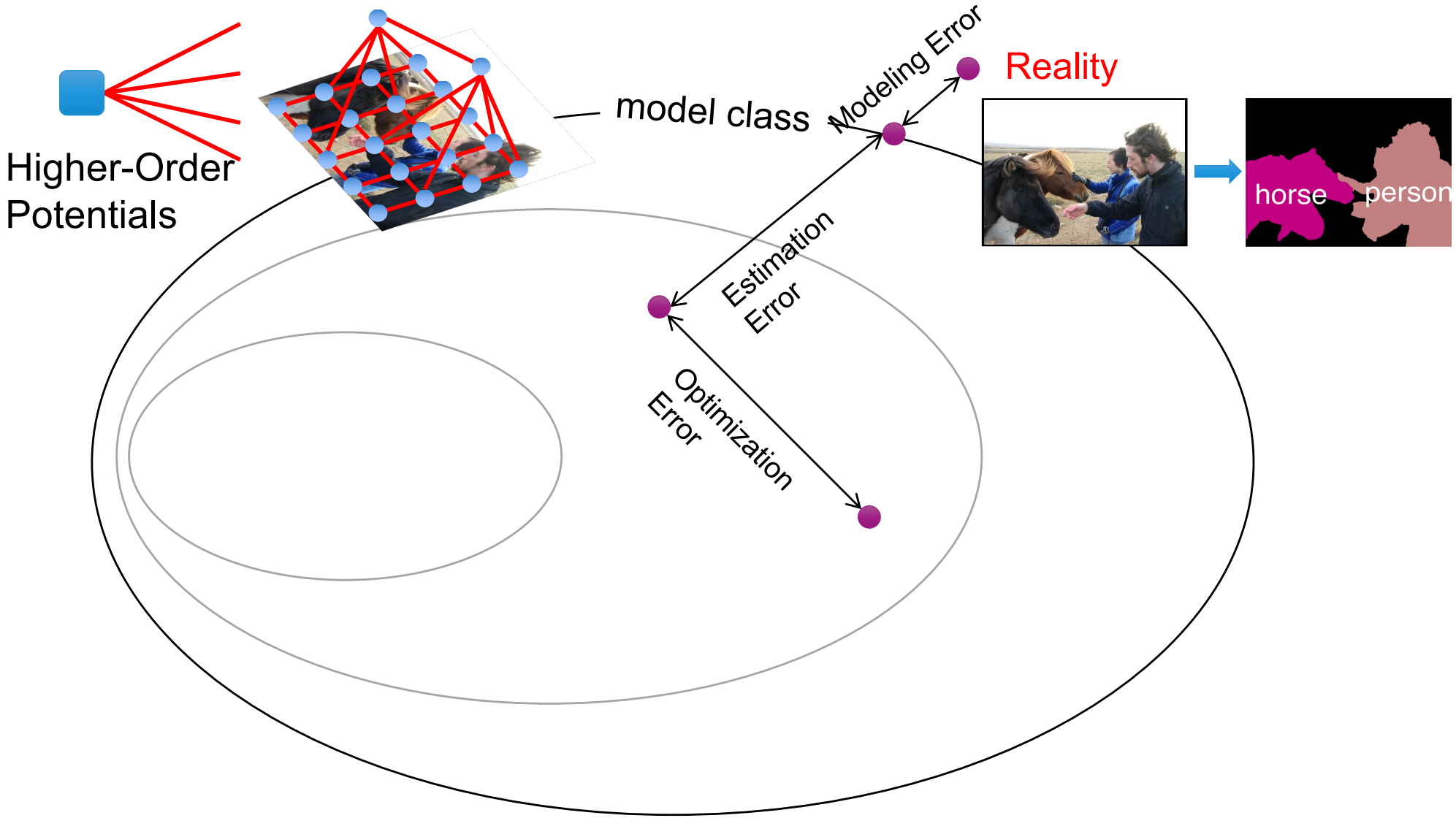
Slide courtesy : D. Batra

Error Decomposition



Slide courtesy : D. Batra

Error Decomposition



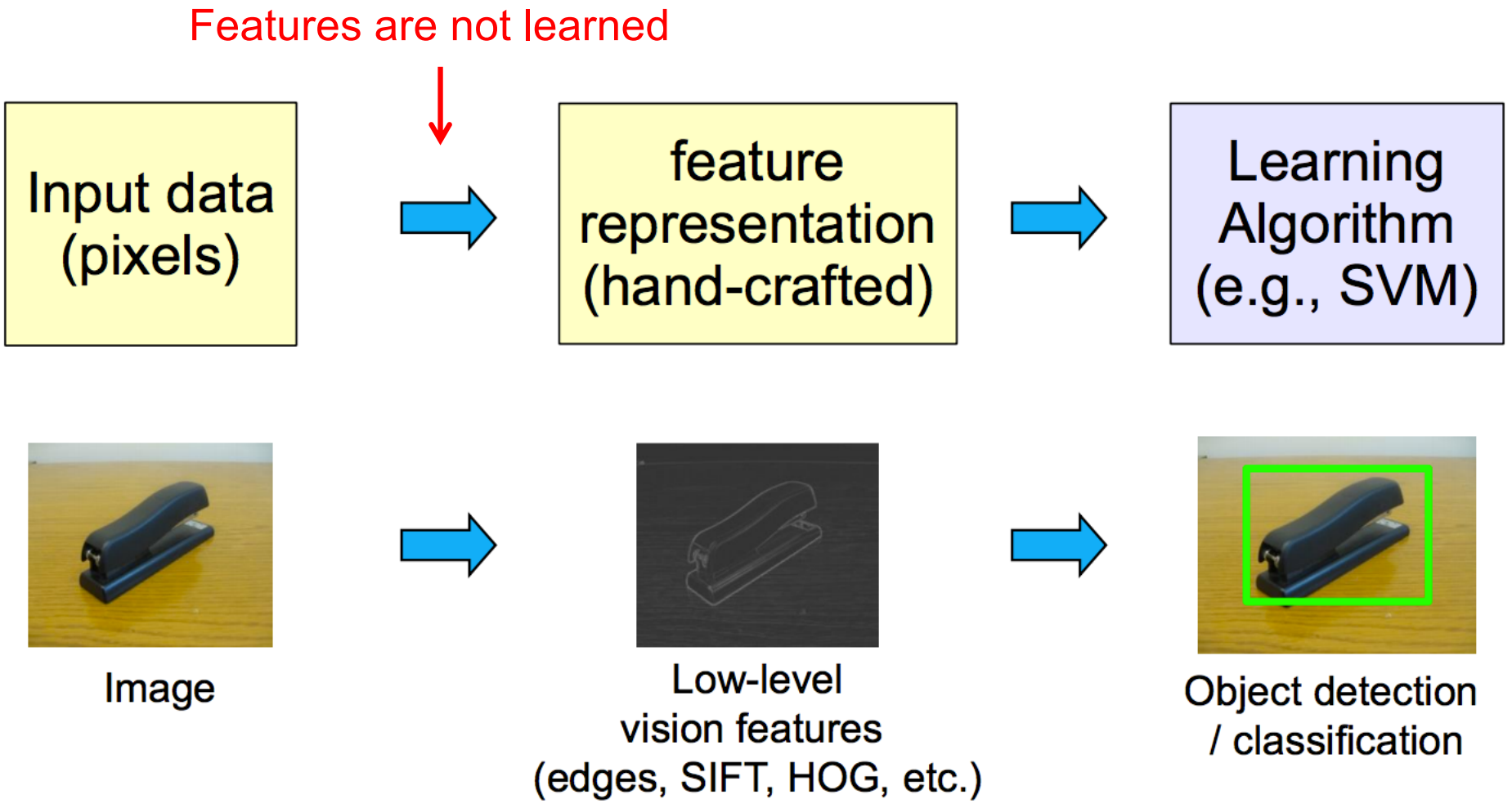
Slide courtesy : D. Batra

Recall: Basic Steps of Supervised Learning

- **Set up** a supervised learning problem
- **Data collection**
  - Start with training data for which we know the correct outcome provided by a teacher or oracle
- **Representation**
  - Choose how to represent the data
- **Modeling**
  - Choose a hypothesis class: H = {g: X → Y}
- **Learning/Estimation**
  - Find best hypothesis you can in the chosen class
- **Model Selection**
  - Try different models. Picks the best one. (More on this later)
- If happy stop
  - Else refine one or more of the above

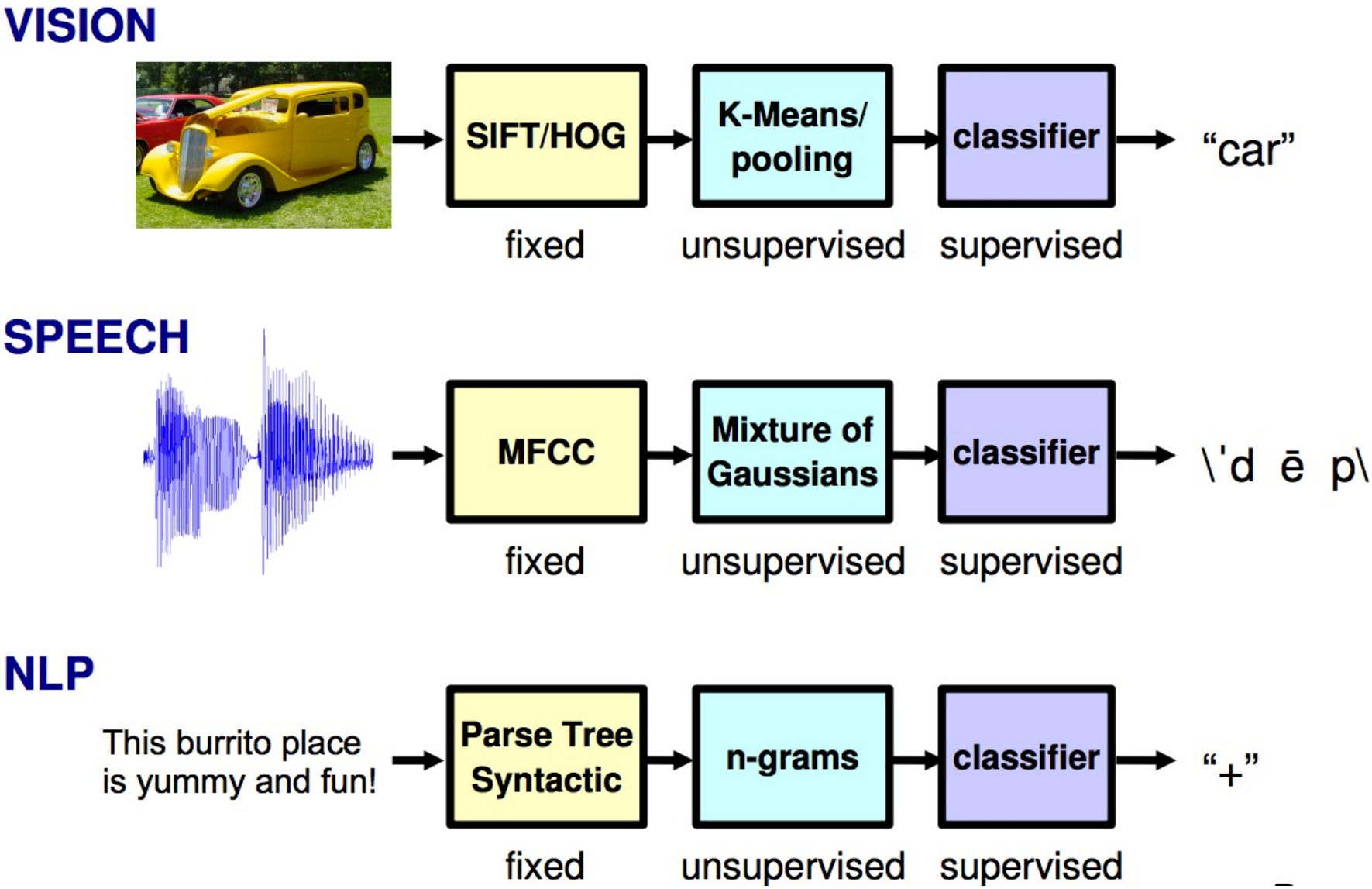
Slide courtesy : D. Batra

Traditional Approaches for Recognition



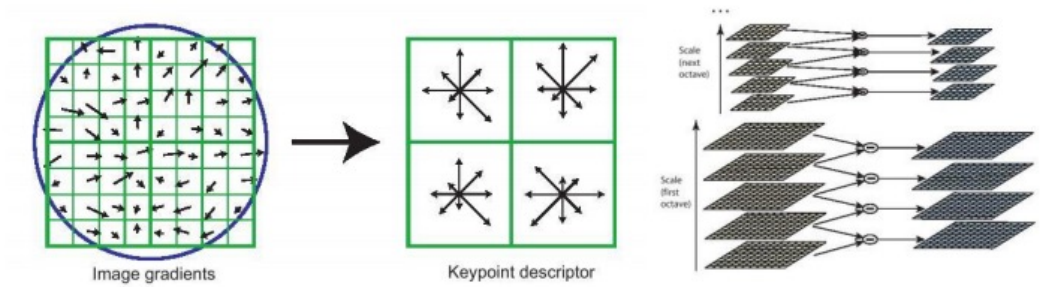
Slide courtesy: R. Fergus, Y. LeCun, M. Renzato, HKUST

Traditional Approaches for Recognition

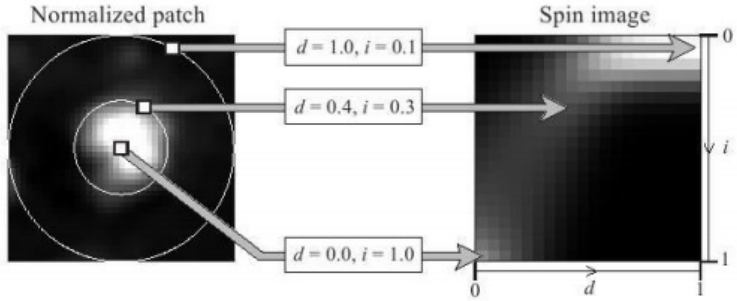


Slide courtesy: R. Fergus, Y. LeCun, M. Renzato, HKUST

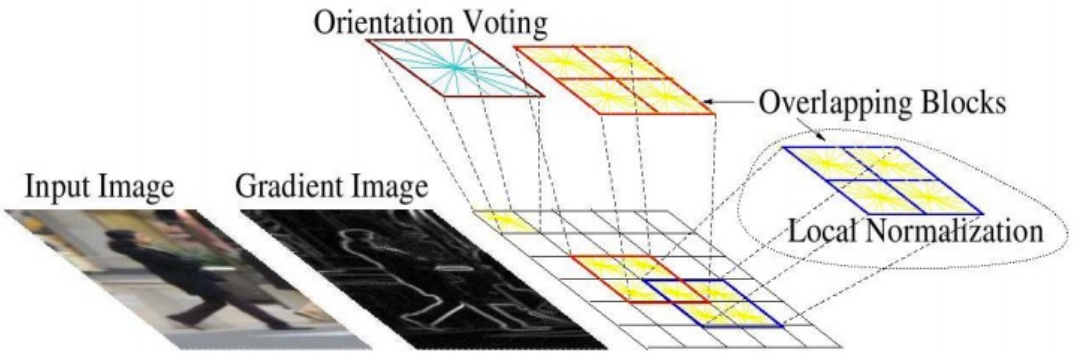
Computer Vision Features



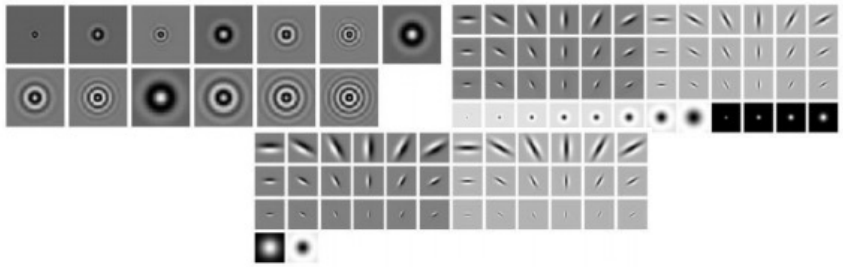
SIFT



Spin image



HoG



Textons

and many others:

SURF, MSER, LBP, Color-SIFT, Color histogram, GLOH, …..

Slide courtesy: R. Fergus, Y. LeCun, M. Renzato, HKUST

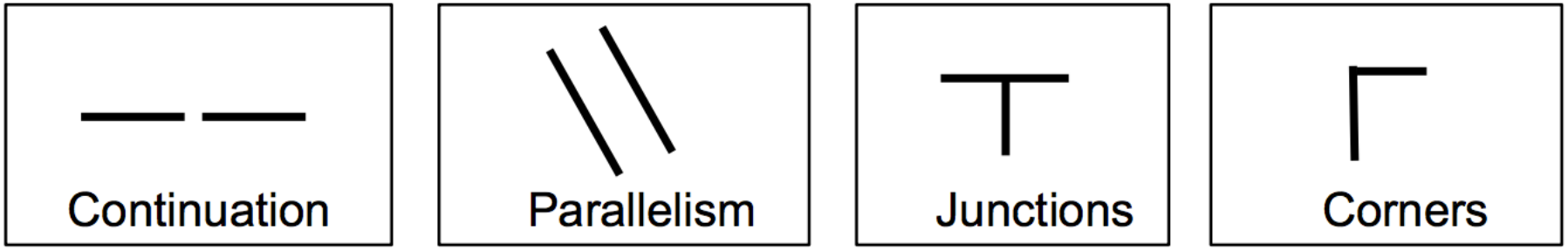
Computer Vision Features

- Features are key to progress
- Have led to impressive results in various competitions (e.g., PASCAL VOC)
- Where do we go from here? Better features? Better classifiers?

Slide courtesy: R. Fergus, Y. LeCun, M. Renzato, HKUST

Mid-level Representations

- Mid-level cues



"Tokens" from Vision by D.Marr:

- Object parts:

Slide courtesy: R. Fergus, Y. LeCun, M. Renzato, HKUST

Mid-level Representations

VISION

pixels → edge → texton → motif → part → object

SPEECH

sample → spectral band → formant → motif → phone → word

NLP

character → word → NP/VP/.. → clause → sentence → story

Difficult to hand-engineer → What about learning them?

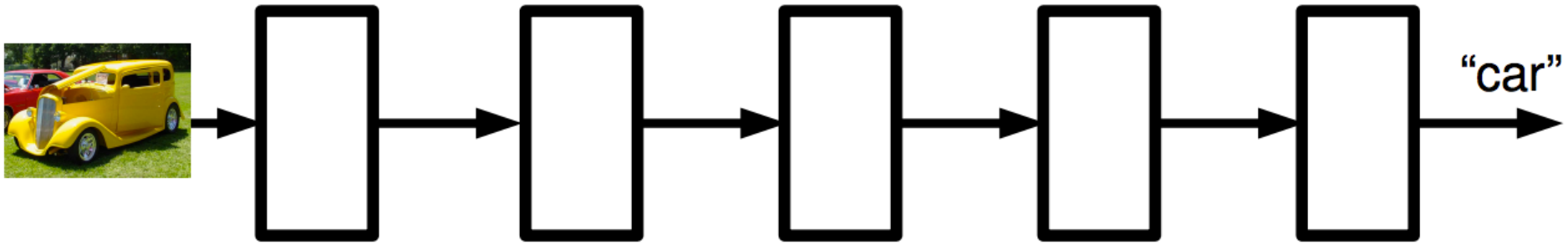
Learning Feature Hierarchy

- Learn hierarchy
- All the way from pixels → classifier
- One layer extracts features from output of previous layer

- Train all layers jointly

Slide courtesy: R. Fergus, Y. LeCun, M. Renzato, HKUST

Deep Learning

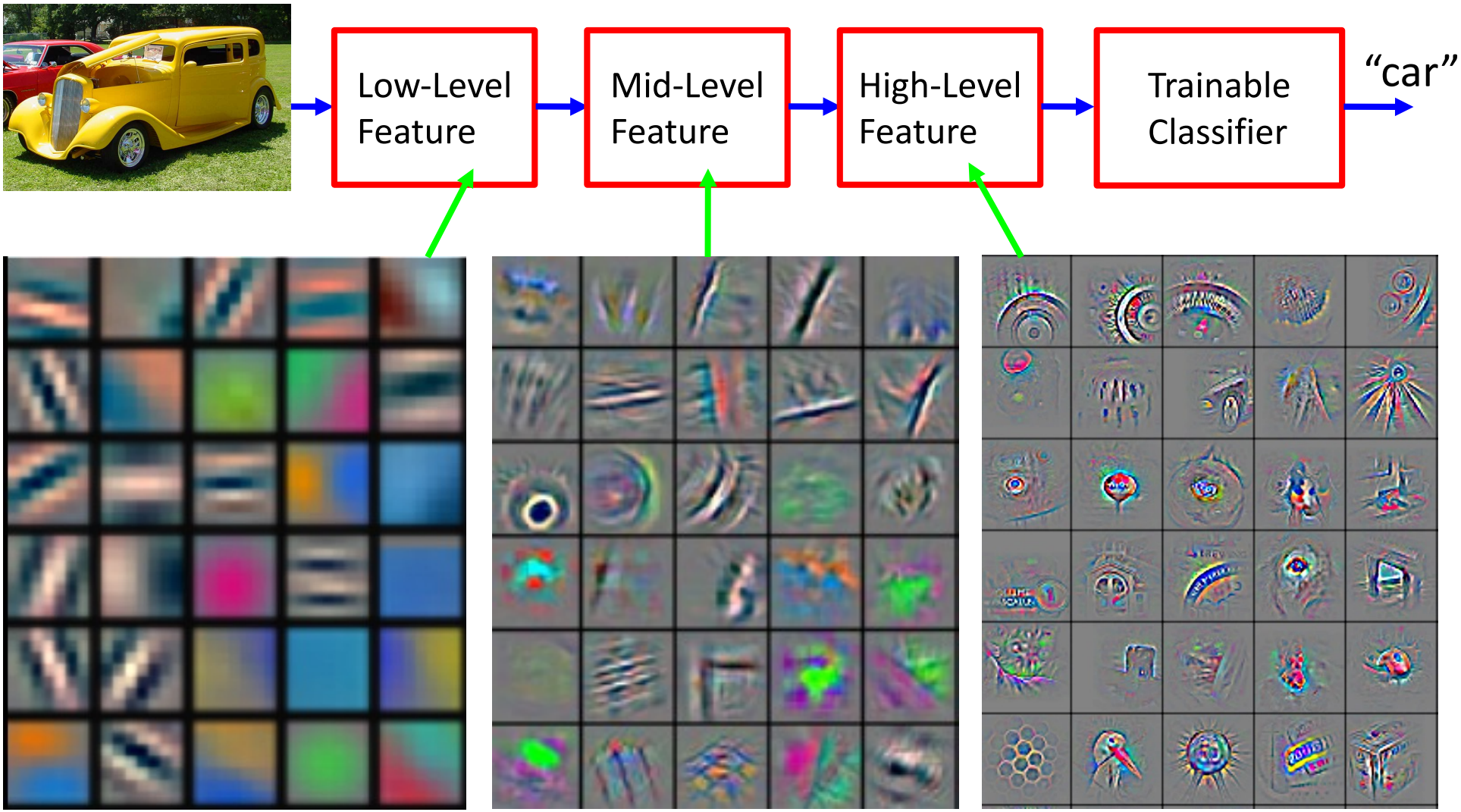


What is Deep Learning

- Cascade of non-linear transformations
- End to end learning
- General framework (any hierarchical model is deep)

Slide courtesy: R. Fergus, Y. LeCun, M. Renzato, HKUST

Deep Learning = Hierarchical Compositionality



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Reconnaissance d’actions