Comprendre les données visuelles à grande échelle

ENSIMAG
2019-2020

Karteek Alahari & Diane Larlus

17 octobre 2019

Comprendre les données visuelles à grande échelle

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

- Intervenants:
  - **Karteek Alahari**, Chargé de Recherche @ INRIA Grenoble <karteek.alahari@inria.fr>
  - **Diane Larlus**, Senior Scientist @ NAVER LABS <diane.larlus@naverlabs.com>
- 12 x 1h30 = 18h de cours

- **Évaluation**
  - Examen final écrit
  - A partir de janvier: toutes les semaines
    - **Quizz** sur des articles de recherche
    - Présentation de ces articles (**bonus!**)

NAVER LABS EUROPE

Organisation du cours

- **17/10/19** cours Diane
- **24/10/19** cours Karteek
- **07/11/19** cours Karteek
- **14/11/19** cours Diane
- **28/11/19** cours Karteek
- **05/12/19** cours Karteek
- **12/12/19** cours Diane
- **19/12/19** cours Diane

Vacances d'hiver

- **09/01/20** cours Diane + présentation articles 1 & 2 + quizz
- **16/01/20** cours Diane + présentation articles 3 & 4 + quizz
- **23/01/20** cours Karteek + présentation articles 5 & 6 + quizz
- **30/01/20** cours Karteek + présentation articles 7 & 8 + quizz

Attention: la salle change régulièrement

Cours 1: Introduction: définitions, données, descripteurs locaux

Comprendre les données visuelles à grande échelle
17 octobre 2019

Les données à grande échelle

Comprendre les données visuelles à grande échelle

Cours 1: Introduction, 17 octobre 2019

NAVER LABS EUROPE

Les données à grande échelle

Wikipedia

Le *big data*

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

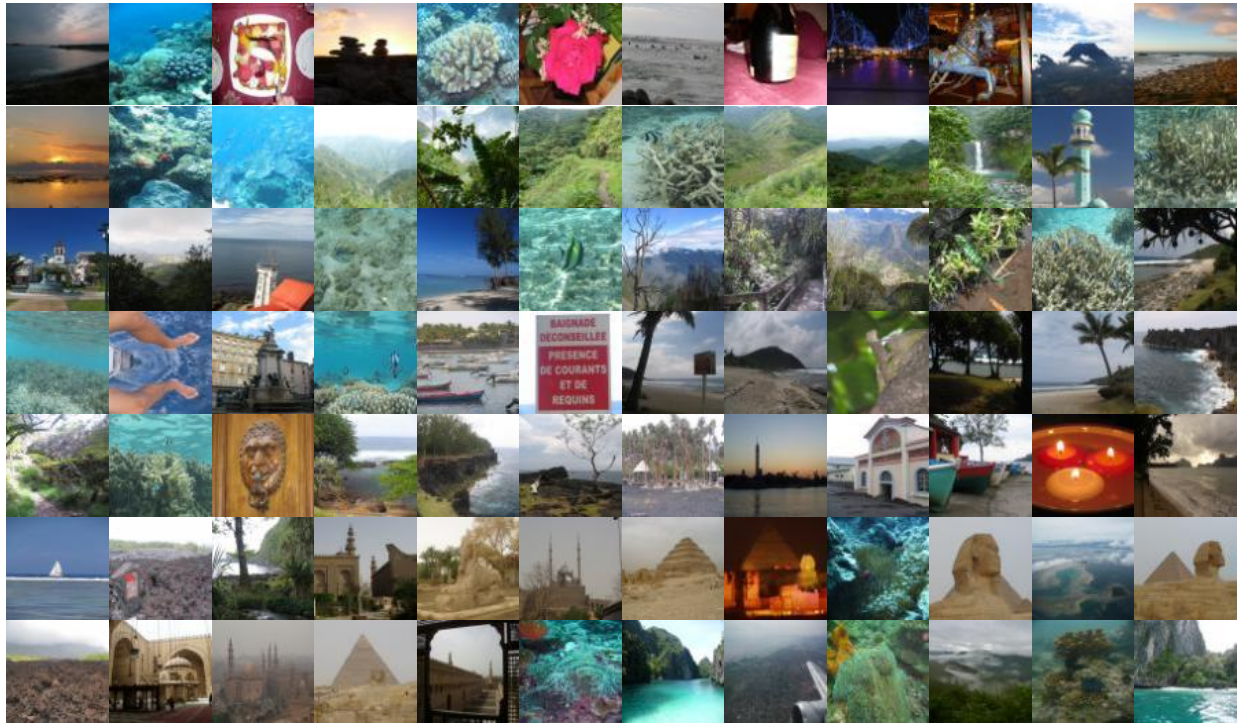
désigne un ensemble de données qui deviennent tellement volumineuses qu'elles en deviennent difficiles à travailler avec des outils classiques de gestion de base de données.

- Nécessite le développement d'outils spécifiques

Les données **visuelles** à grande échelle

ou ***Big Visual Data***

sont devenues une façon majeure de transférer l'information



facebook

gettyimages

YouTube

NAVER

LINE@

flickr

Pinterest

Signal Private Messenger

Instagram

WhatsApp

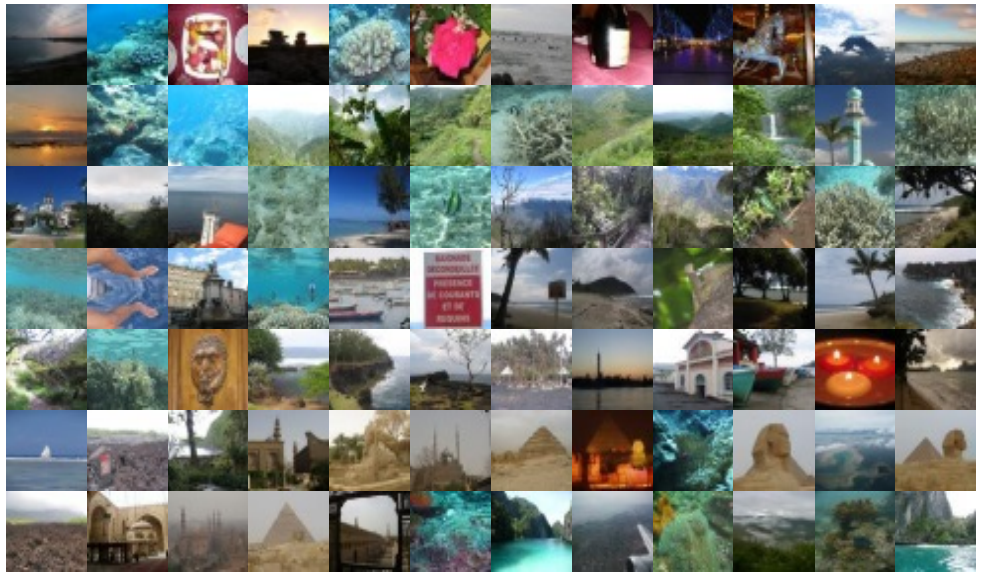
SnapChat

NAVER LABS EUROPE

Quelques chiffres

- Croissance très importante, en raison de l'accumulation des contenus numériques auto-produits par le grand public

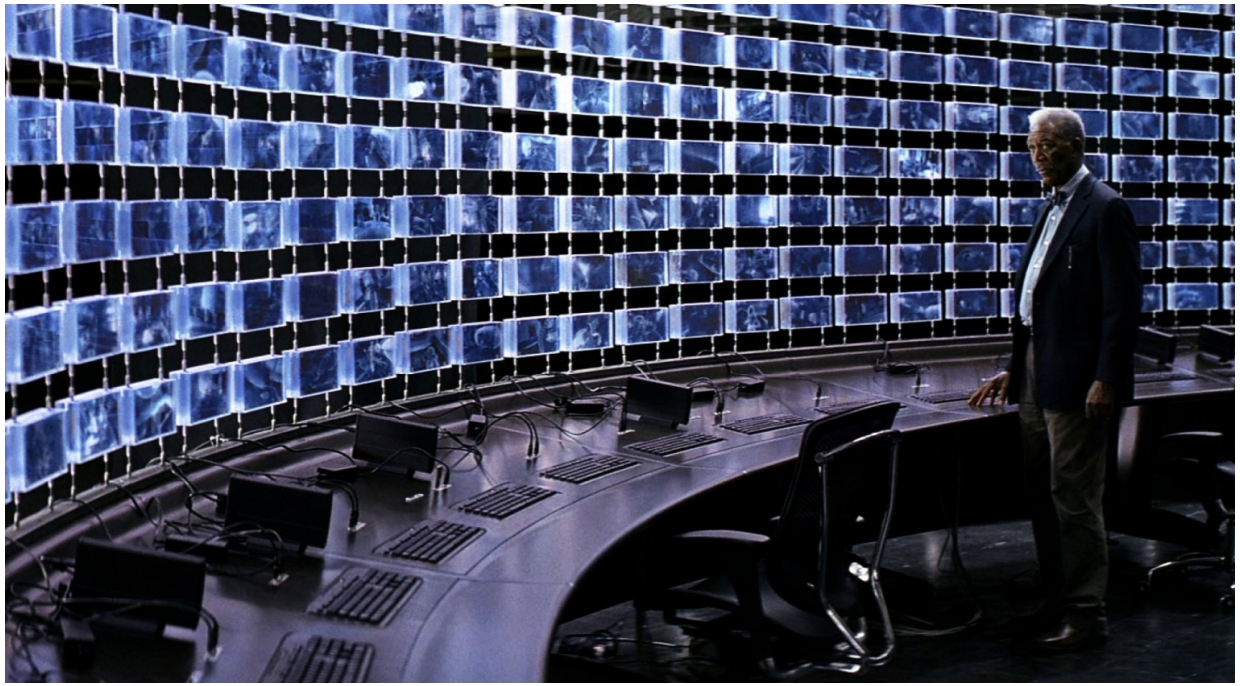
  - **ImDb** recense plus de 400 000 films
  - Images (semi-)pro : **Corbis**, **Getty**, **Fotolia**
    - centaines de milliers d' images
  - **Facebook** – chiffres de 2013
    - 350 millions de nouvelles photos / jour
    - 250 milliards de photos stockées
  - **Flickr** – novembre 2016
    - 13 milliards de photos
  - **Youtube**
    - 35h / min uloaded in 2011
    - 100h / min uploaded in 2014



Les données visuelles à grande échelle

ou ***Big Visual Data***

sont devenues une façon majeure de transférer l'information



Selon la *British Security Industry Authority (BSIA)*, il y aurait autour de 5M de CCTV cameras en GB, soit une pour 14 habitants (source: telegraph-2013)

Les données visuelles à grande échelle

ou ***Big Visual Data***

sont devenues une façon majeure de transférer l'information

Seule une partie de ces données est exploitée, et une bonne partie du processus est fait à la main

→ opportunités de simplification de tâches fastidieuses

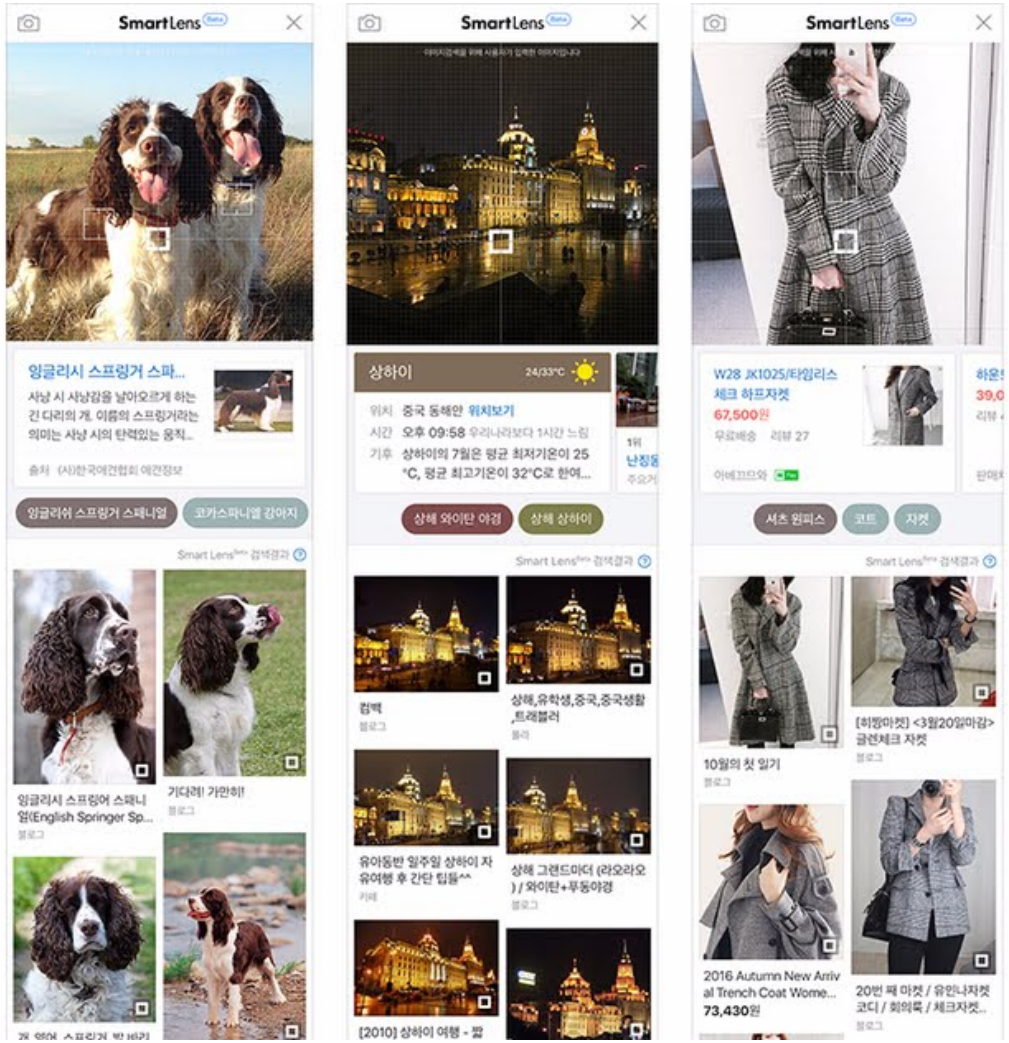
→ opportunités de nouvelles applications et de nouvelles solutions commerciales

→ développement de **l'intelligence ambiante**

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.

Exemples d’applications à NAVER LABS



LABS
NAVER LABS EUROPE

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**
   - **recherche d'images basée sur le contenu (*image retrieval* ou *image search*)**
   - **Autres taches de vision par ordinateur / d'analyse d'images (voir exemples)**
6. Distribution
   - conception de serveur de streaming, interfaces de l'application, etc.

Annotations d’images: difficulté et ambiguïté

Comprendre les données visuelles à grande échelle

Cours 1: Introduction, 17 octobre 2019

NAVER LABS EUROPE

Contenu et métadonnées

- Les données “brutes” (fichier image, fichier son) contiennent des informations sémantiques = directement compréhensibles pour l'utilisateur

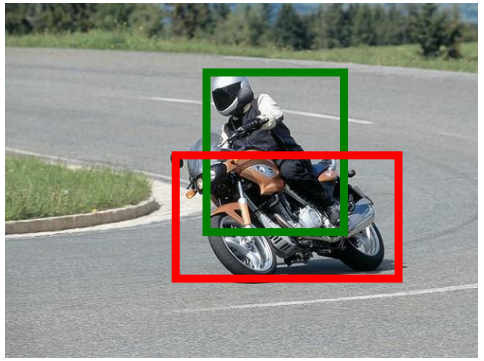
- Ces **métadonnées** proviennent
  - Soit de propriétés de descripteurs des objets (ex: couleur moyenne d’une image, métadonnée surexposé)
  - Soit de données d’autres médias (ex: GPS)
  - Soit d’annotations manuelles (ex. Tags)

Exemple: *Exchangeable image file format* (Exif)

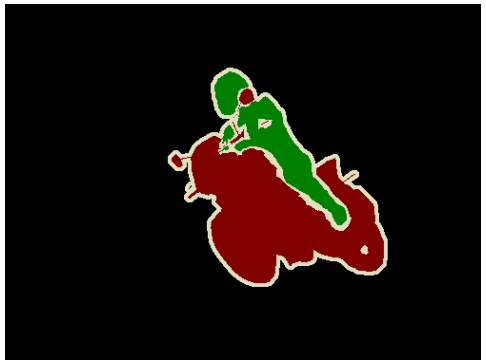
- Spécification pour les formats d'images des appareils numériques
  - non uniformisé, mais largement utilisé
- Pour JPEG, TIFF, RIFF, ne supporte pas, PNG ou GIF
- Le format supporte souvent
  - Date et heure, enregistrés par l'appareil
  - Les paramètres de l'appareil
    - Dépendent du modèle : inclus 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

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



people motorbike



people motorbike

motorcyclist turning right

LABS
NAVER LABS EUROPE

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

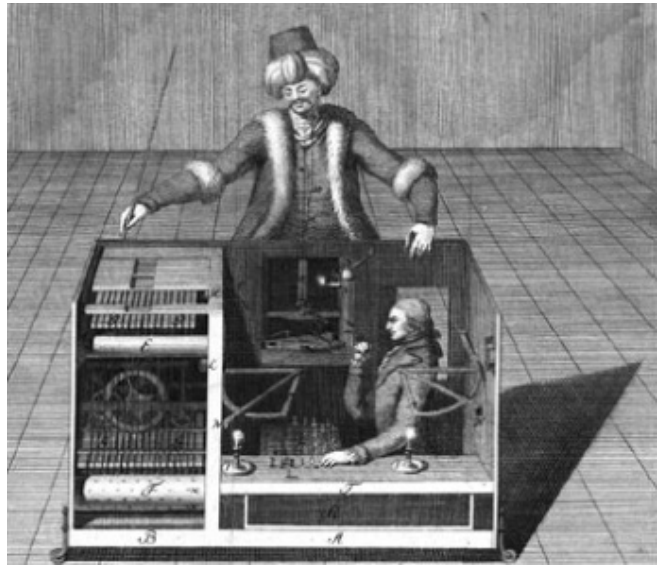
What to label	*All objects of the defined categories*, unless: you are unsure what the object is. the object is very small (at your discretion). less than 10-20% of the object is visible. If this is not possible because too many objects, mark image as bad.
Viewpoint	Record the viewpoint of the 'bulk' 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

NAVER LABS EUROPE

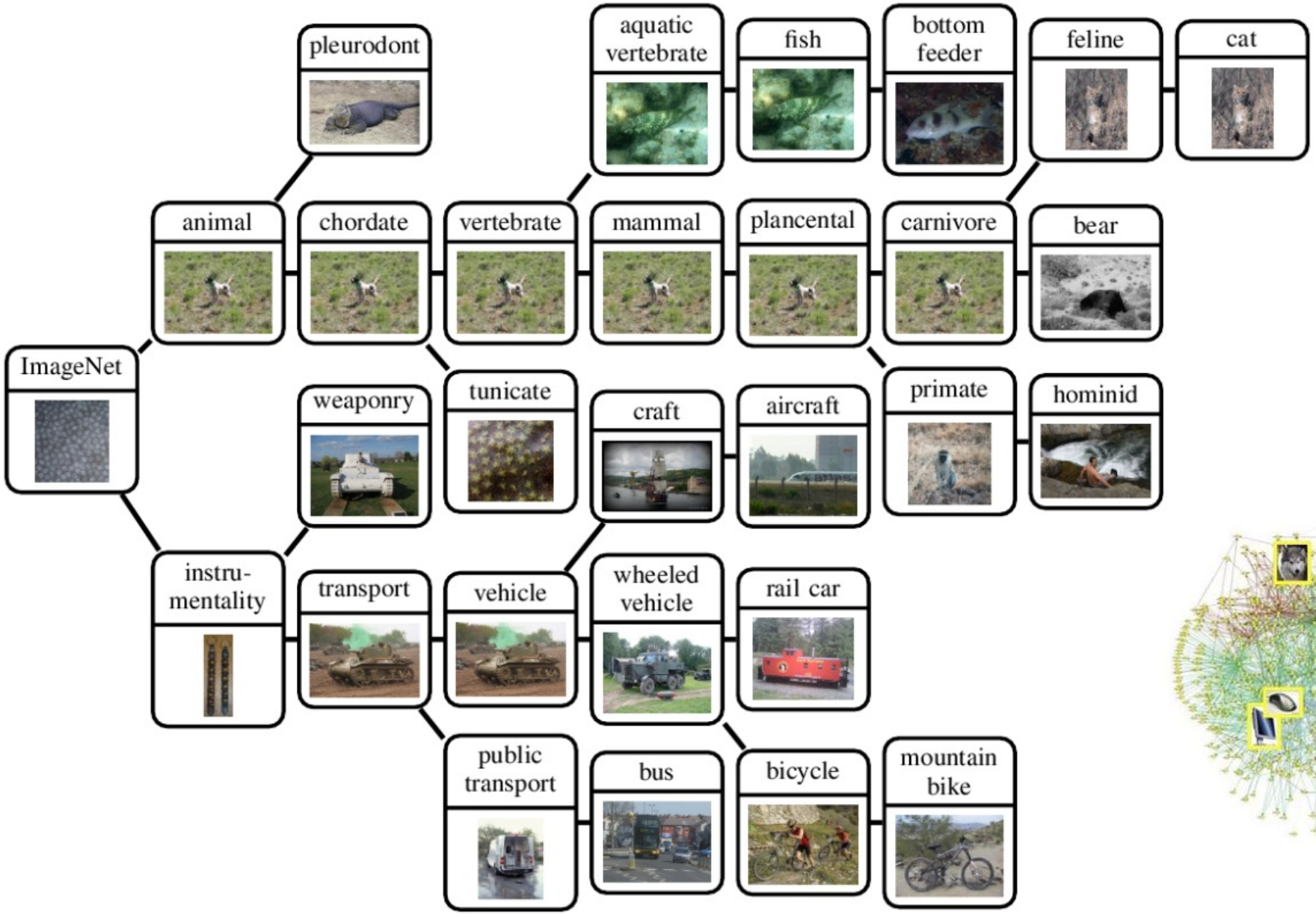
Ambiguïté de l'annotation d'images

- Obtenir des annotations précises est une tâche fastidieuse
- Une solution récente : 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/)

ImageNet



Article à lire pour janvier

IMAGENET

NAVER LABS EUROPE

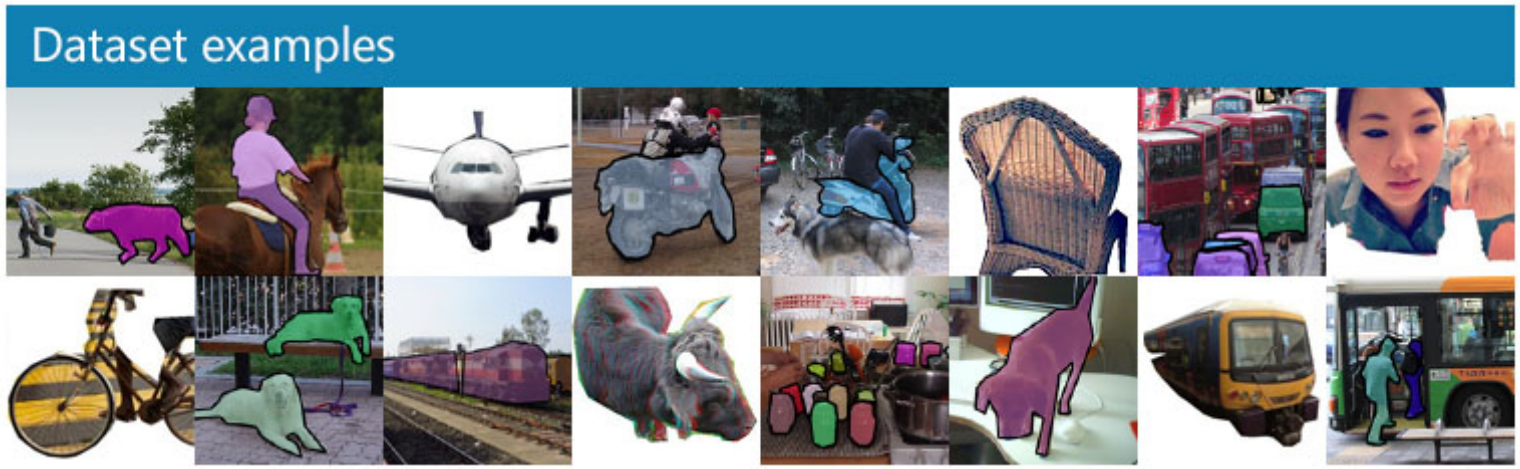
MS COCO

- *Common Objects in Contexts*

“What is COCO?” from its webpage

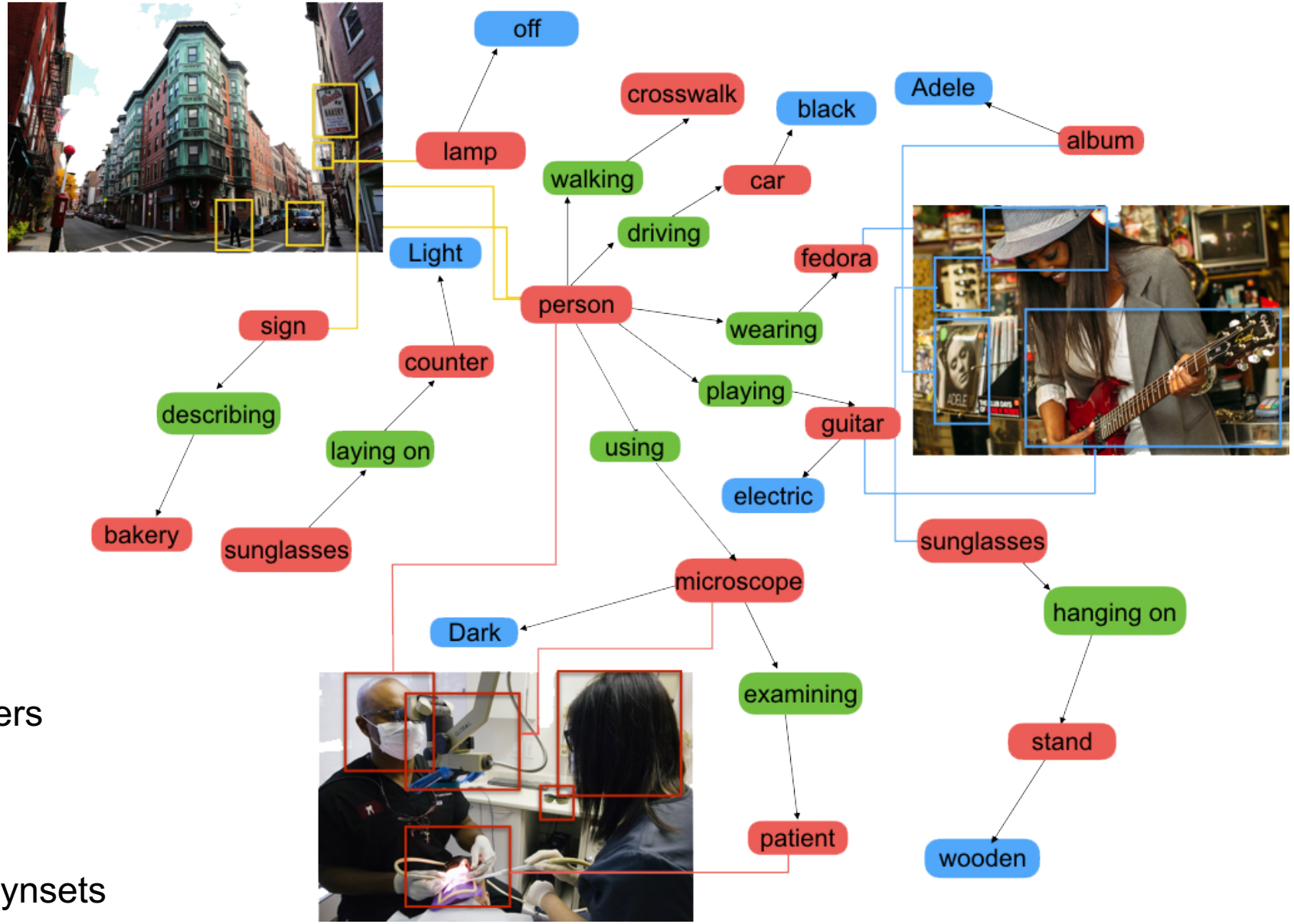
COCO is a large-scale object detection, segmentation, and captioning dataset
COCO has several features:

- 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

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

NAVER LABS EUROPE

Quels sont les processus automatisables en analyse de données visuelles ?

Comprendre les données visuelles à grande échelle

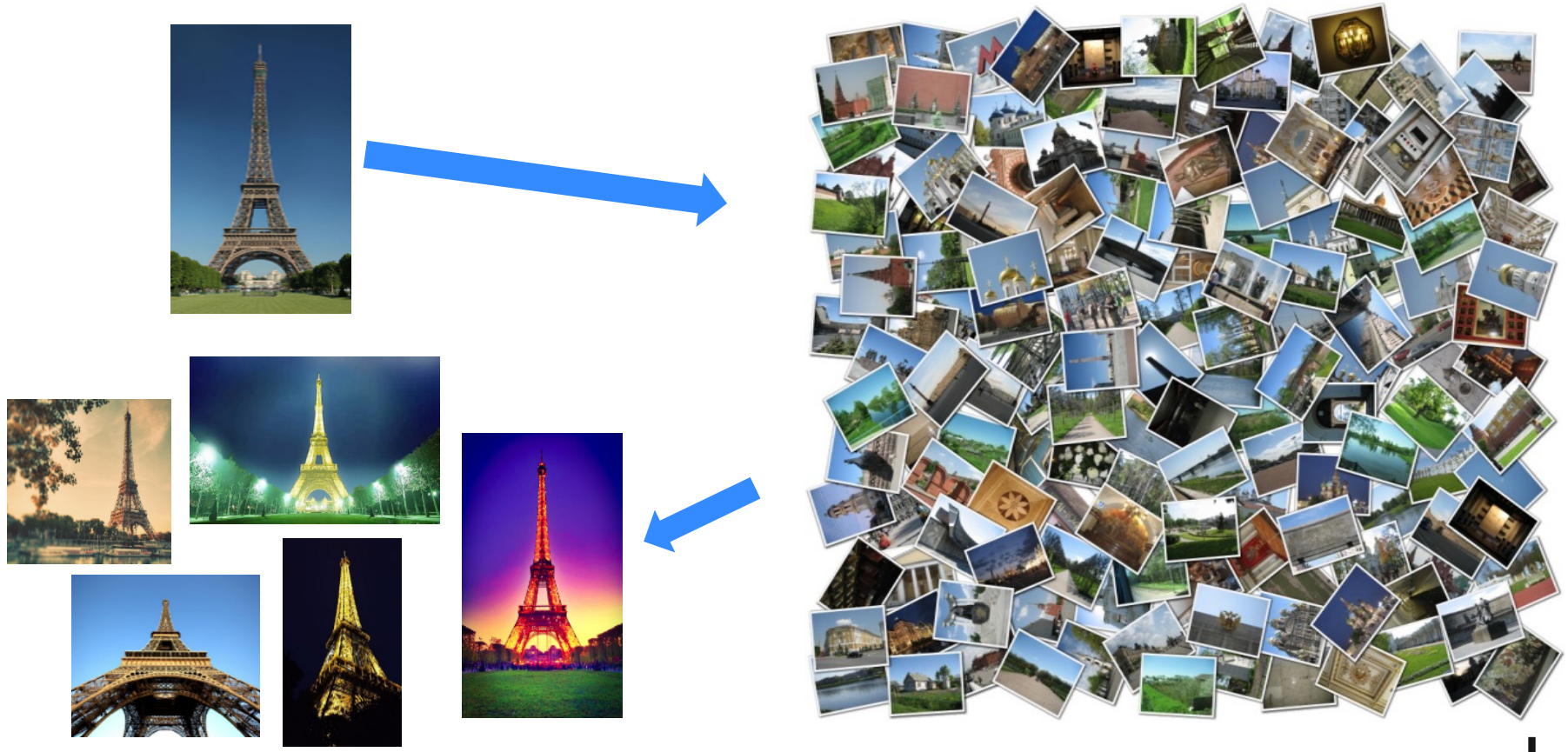
Cours 1: Introduction, 17 octobre 2019

NAVER LABS EUROPE

1) Recherche d’images par similarité

Principe

- Etant donné une image, retrouver les images similaires dans une grande base visuelle



NAVER LABS EUROPE

1) Recherche d'images par similarité

- Utilise la notion de proximité, de similarité, ou de distance entre images
- Généralement, la requête est exprimée sous la forme d'un ou de plusieurs **vecteurs** dans un espace multidimensionnel.
  - définition d'une distance (ou mesure de similarité) sur cet espace
  - recherche des objets dont la distance est minimale
- Les **vecteurs** sont extraits du contenu de l'image

Recherche par le contenu

ou CBIR: *content based information retrieval*

- Possibilité d'utiliser un **retour de pertinence**
  - L'utilisateur spécifie quels résultats sont les plus pertinents pour sa requête
  - Le système raffine les résultats en fonction de ce retour

Variation d'apparence d'une **instance** d'objet donné

NAVER LABS EUROPE

Autres processus automatisables: la reconnaissance d'objet

2) Catégorisation d'image

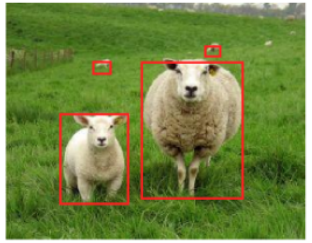
- Catégorie principale associée à l'image, ou réponse oui/non à une liste de catégories connues à l'avance

Sheep ?

yes

3) Détection d'objet

- Boite englobante pour toutes les instances d'une catégorie d'intérêt



Sheep ?

4

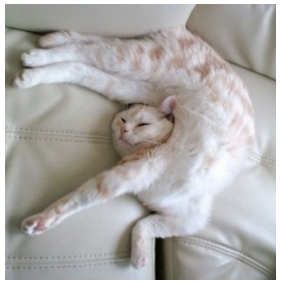
4) Segmentation d'objet, segmentation sémantique

- Localisation précise des objets au niveau du pixel

Sheep ?

Difficultés de la modélisation des **catégories** d'objet

- Illumination, ombres
- Orientation et pose
- Fond texturé, distracteurs
- Occultations
- Variations intra-classe



5) Etiquetage d’image

ou *Image tagging*

- Catégorisation de l’objet principal contenu dans l’image ou des différents objets
  - lien avec la reconnaissance d’objet et notamment la classification d’image
- Catégorisation de la scène
  - intérieur vs extérieur, paysage urbain ou campagnard, etc.
- Tout autre type d’attributs possible
  - Couleur, forme, texture, etc.

Sheep

Tokyo tower
Landscape
Sunset
Urban

NAVER LABS EUROPE

Exemple d’outil d’étiquetage

Démo proposée par Clarifiai
(https://www.clarifai.com/demo)

- lake
- wood
- water
- fall
- nature
- no person
- reflection
- outdoors
- landscape
- scenic
- mountain
- wild
- tree
- river

Annotations avec des critères subjectifs

- 6) Qualités esthétiques
  - Prédire si le plus grand nombre va trouver une image agréable à regarder, visuellement esthétique
- 7) Iconicité d'une image
  - Prédire si une image est un bon représentant d'un concept, par exemple dans un but d'enseignement
- 8) Mémorabilité
  - Prédire si une image sera plus facilement remarquée, mémorisée (*remembered*)

[AVA: Murray et al. CVPR12]
[What makes an image iconic? Zhang et al. Arxiv 14]
[Image memorability. Khosla et al. ICCV15]

Annotations plus complexes

- 9) Le sous-titrage automatique d'image (*image captioning*)

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

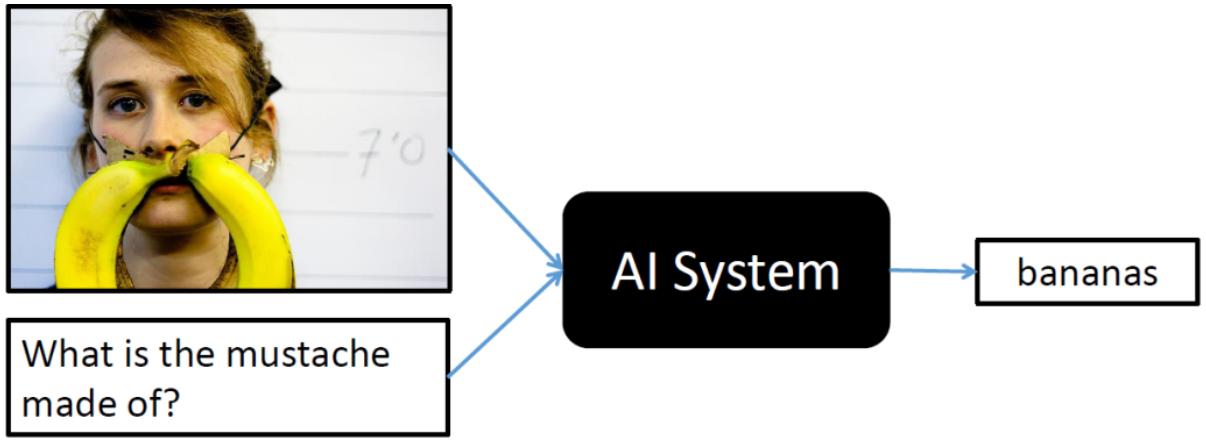
Automatic Image Caption Generation
Sample taken from the work of
Andrej Karpathy and
Li Fei-Fei

More recent demo:
https://www.captionbot.ai/

NAVER LABS EUROPE

Annotations plus complexes

- 10) La réponse à des questions visuelles (*Visual question answering* ou *VQA*)



http://www.visualqa.org/

More recent demo:
http://vqa.cloudcv.org/

Apprentissage automatique et intelligence artificielle

Comprendre les données visuelles à grande échelle

Cours 1: Introduction, 17 octobre 2019

NAVER LABS EUROPE

Vocabulaire

Wikipedia

- Apprentissage automatique (*Machine learning*)
  champ d'étude de l'intelligence artificielle qui se base sur des approches statistiques pour donner aux ordinateurs la capacité d' « apprendre » à partir de données [..].
  Il comporte généralement deux phases.
  - La première consiste à estimer un modèle à partir de données, appelées observations [..] Cette phase dite « d'apprentissage » ou « d'entraînement » est généralement réalisée préalablement à l'utilisation pratique du modèle.
  - La seconde phase correspond à la mise en production : le modèle étant déterminé, de nouvelles données peuvent alors être soumises afin d'obtenir le résultat correspondant à la tâche souhaitée.

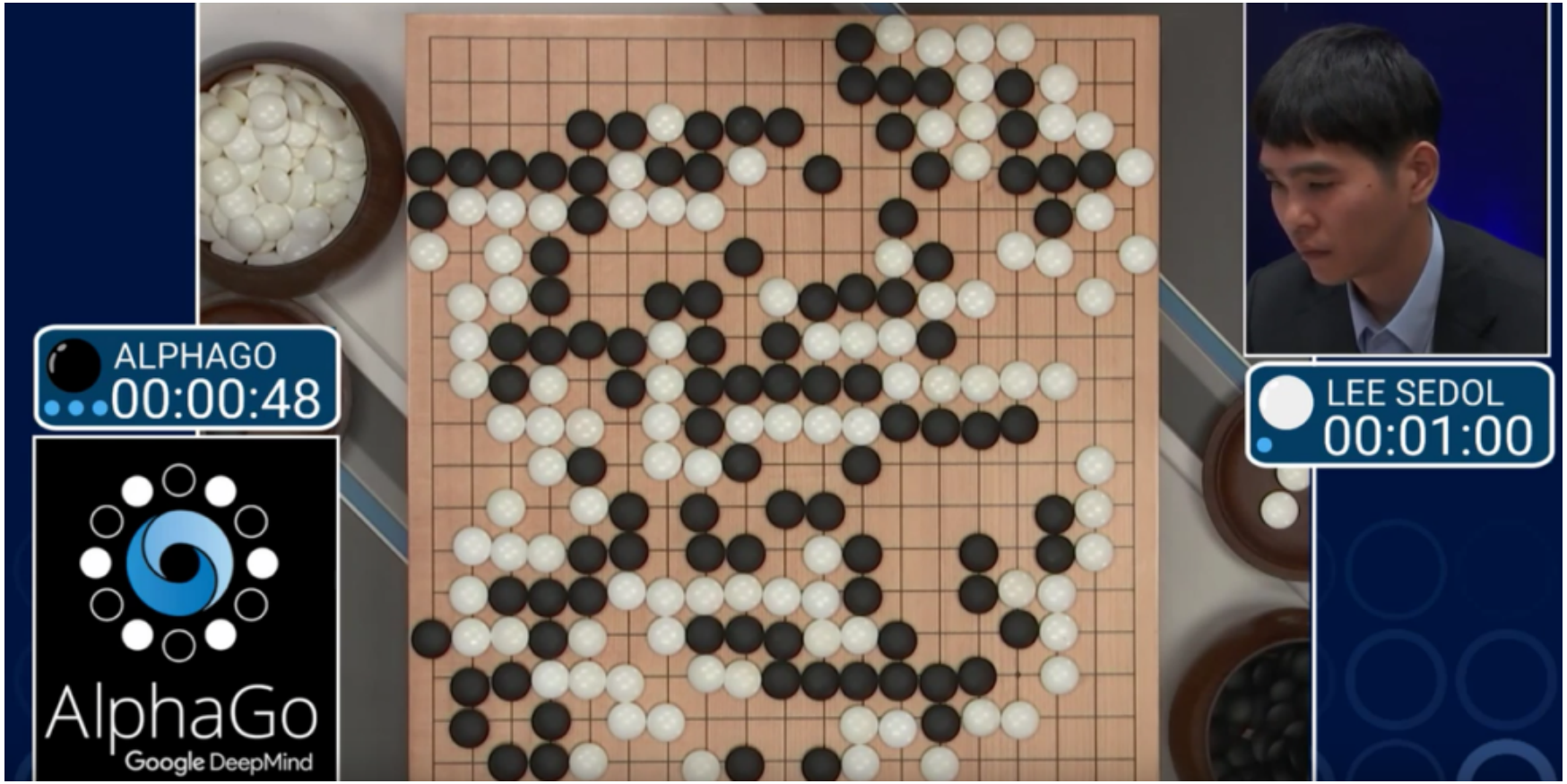
- Apprentissage profond (*Deep Learning*)
  est un ensemble de méthodes d'apprentissage automatique tentant de modéliser avec un haut niveau d'abstraction des données grâce à des architectures articulées de différentes transformations non linéaires

- Intelligence artificielle (*AI*)
  est l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence

AI in the news

AlphaGo

- 2016: computer beats a top-ranked professional player at the go game, known as one of the most difficult for computers
  - You can check the documentary movie, e.g. on Netflix

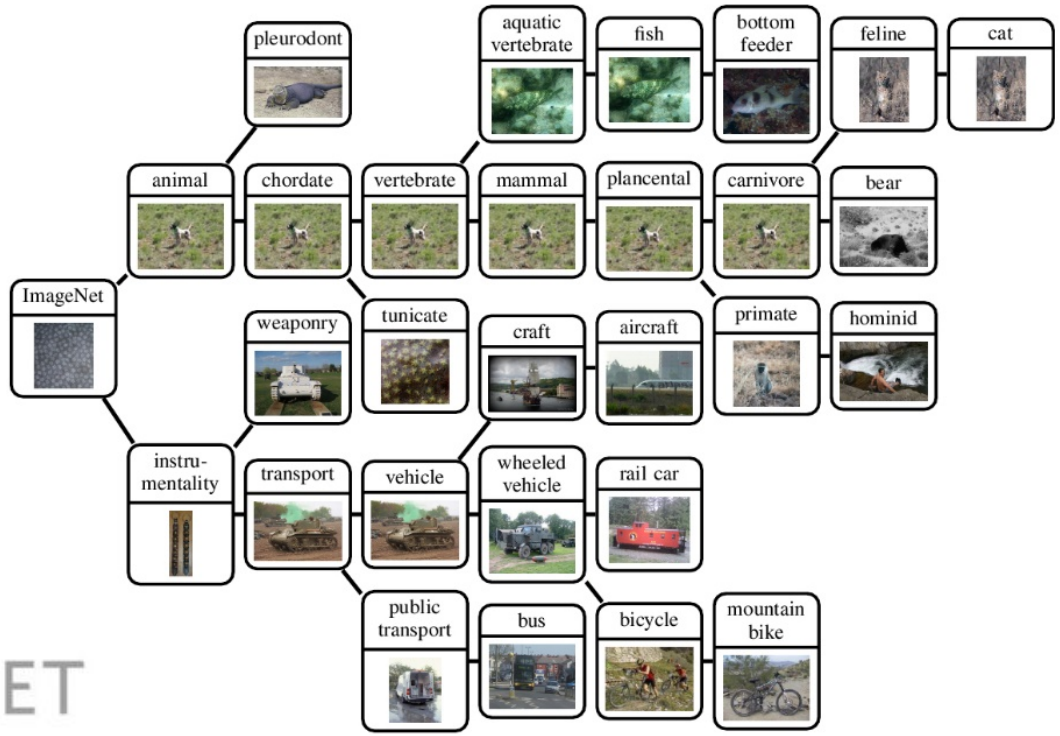


AI in the news

2019: Deep fake – e.g. https://www.youtube.com/watch?v=3vHvOyZ0GbY

Pushing the state of the art

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

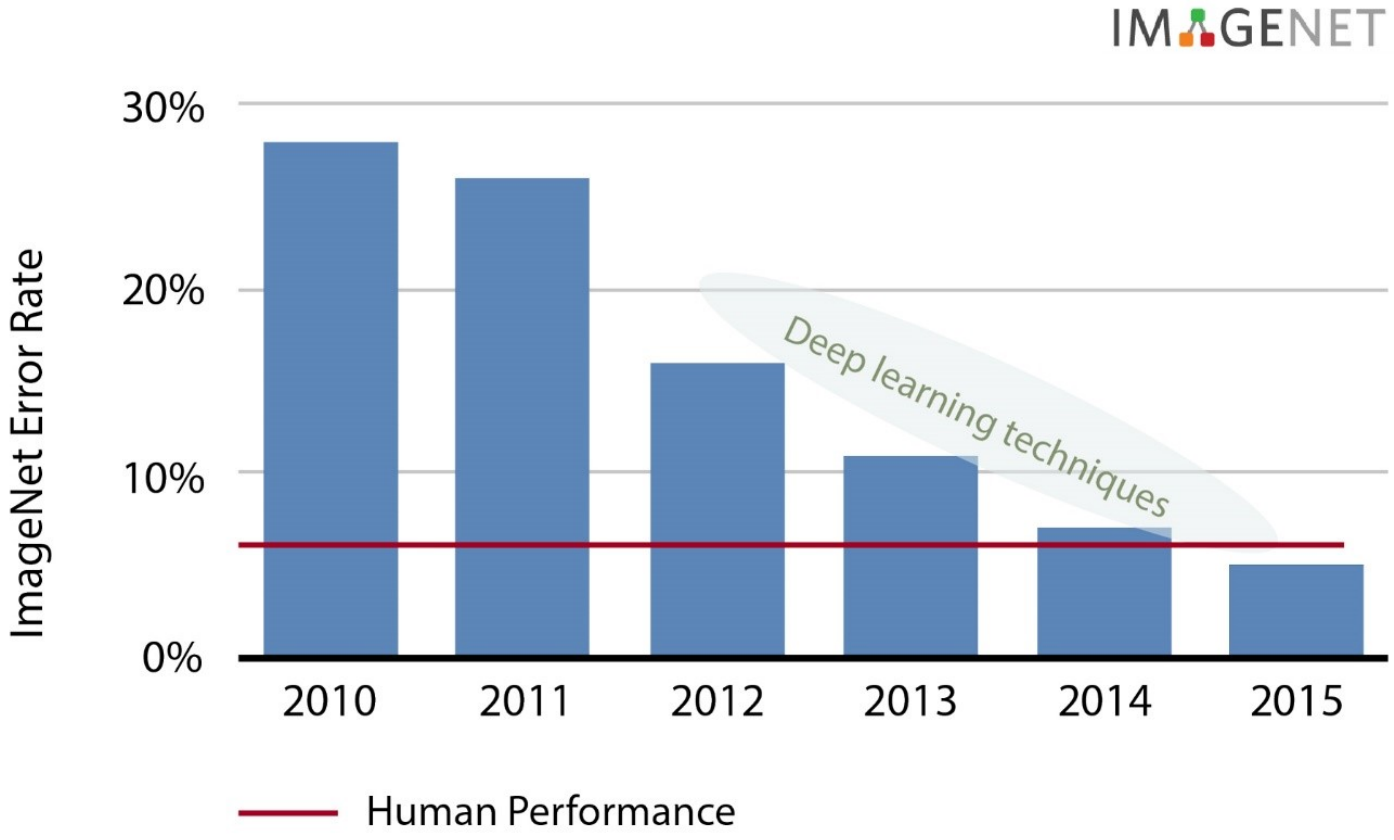


IMAGENET

NAVER LABS EUROPE

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)



Deep Learning techniques have led to significant performance improvements in recent years (Source: Nervana)

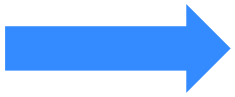
Recherche visuelle d’images – introduction

Comprendre les données visuelles à grande échelle

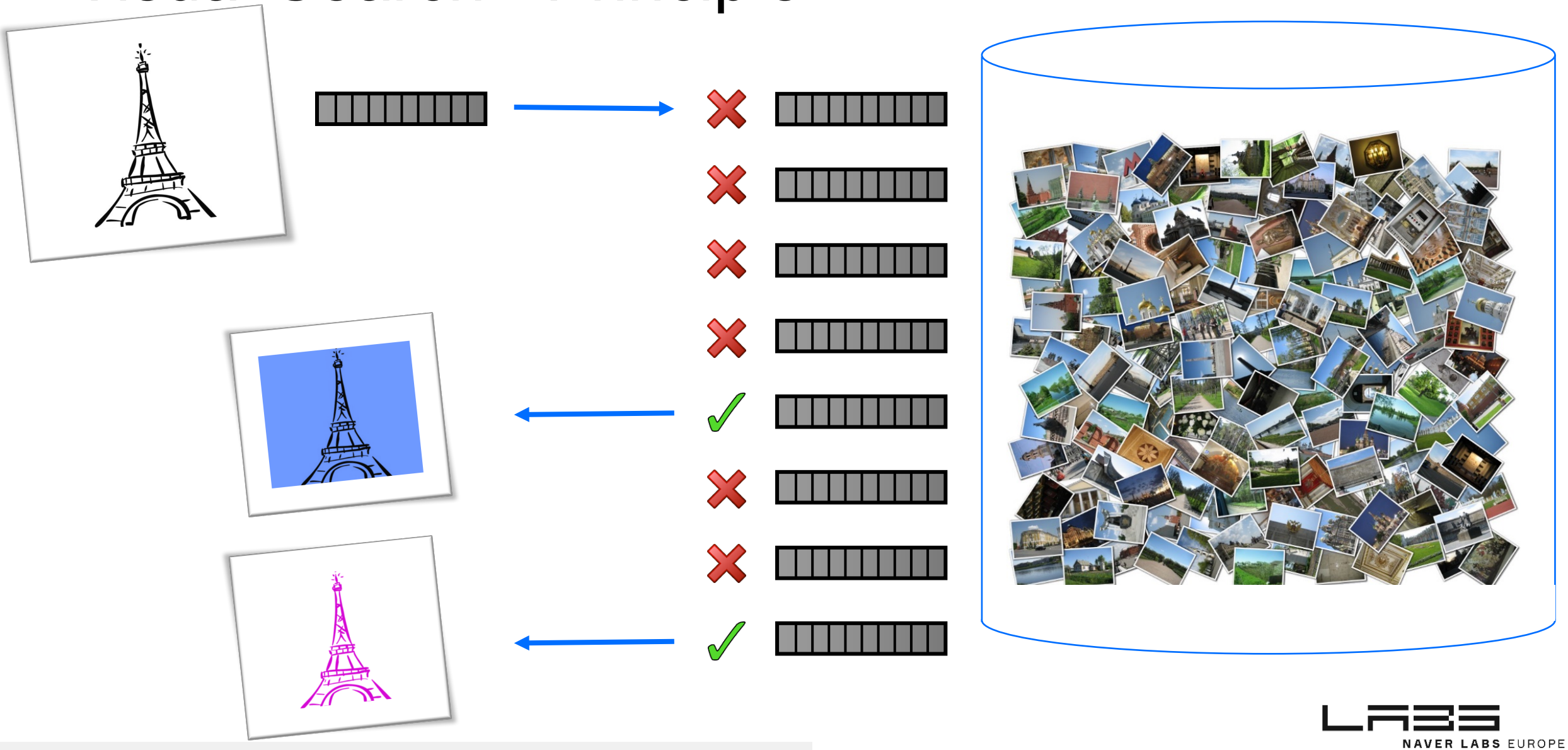
Cours 1: Introduction, 17 octobre 2019

NAVER LABS EUROPE

Visual Search - Principle



Visual Search - Principle

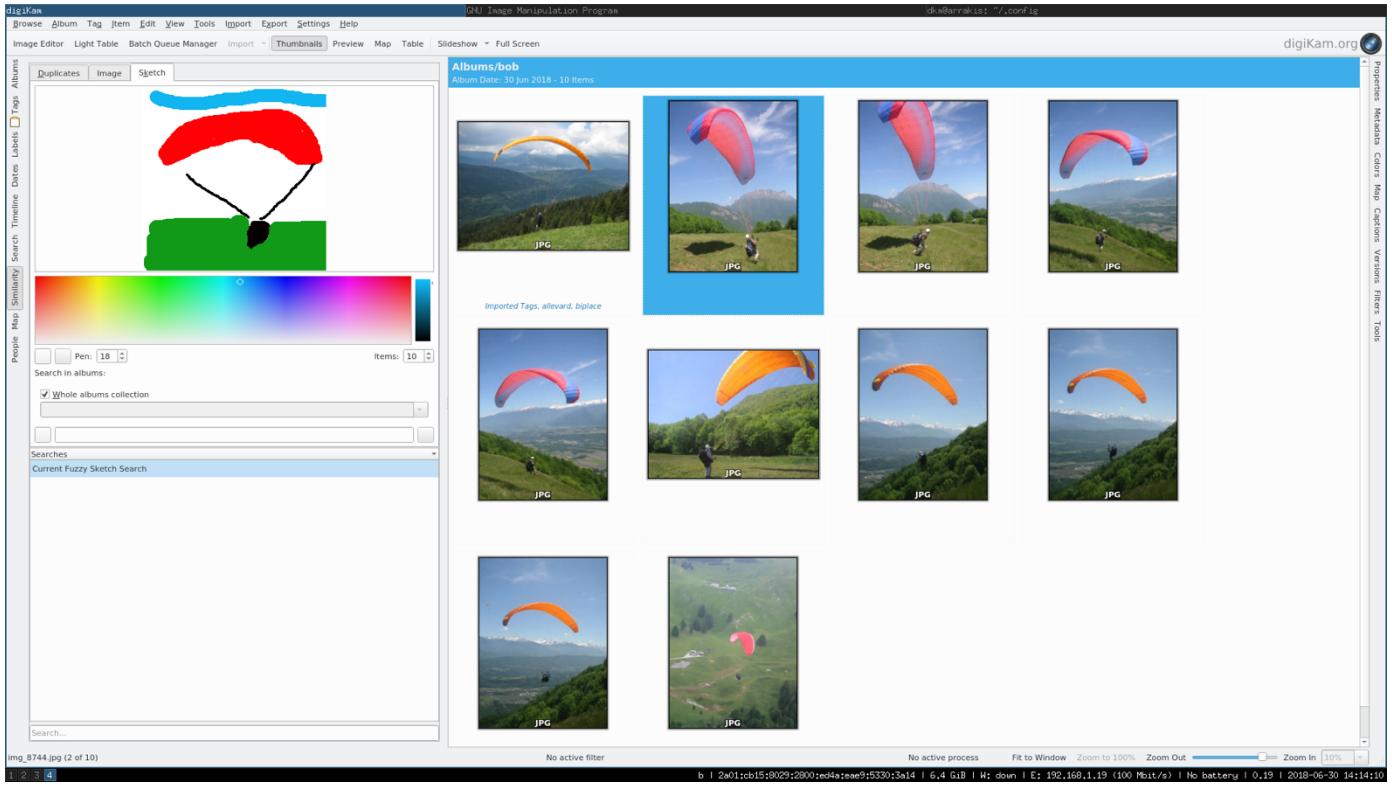


NAVER LABS EUROPE

Visual Search - Applications

Many applications

- Reverse Image search
  - Web search engine
  - Personal photo collection

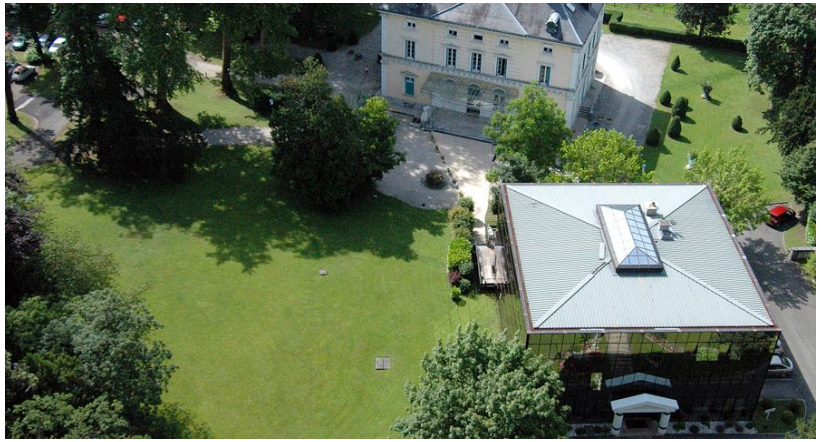


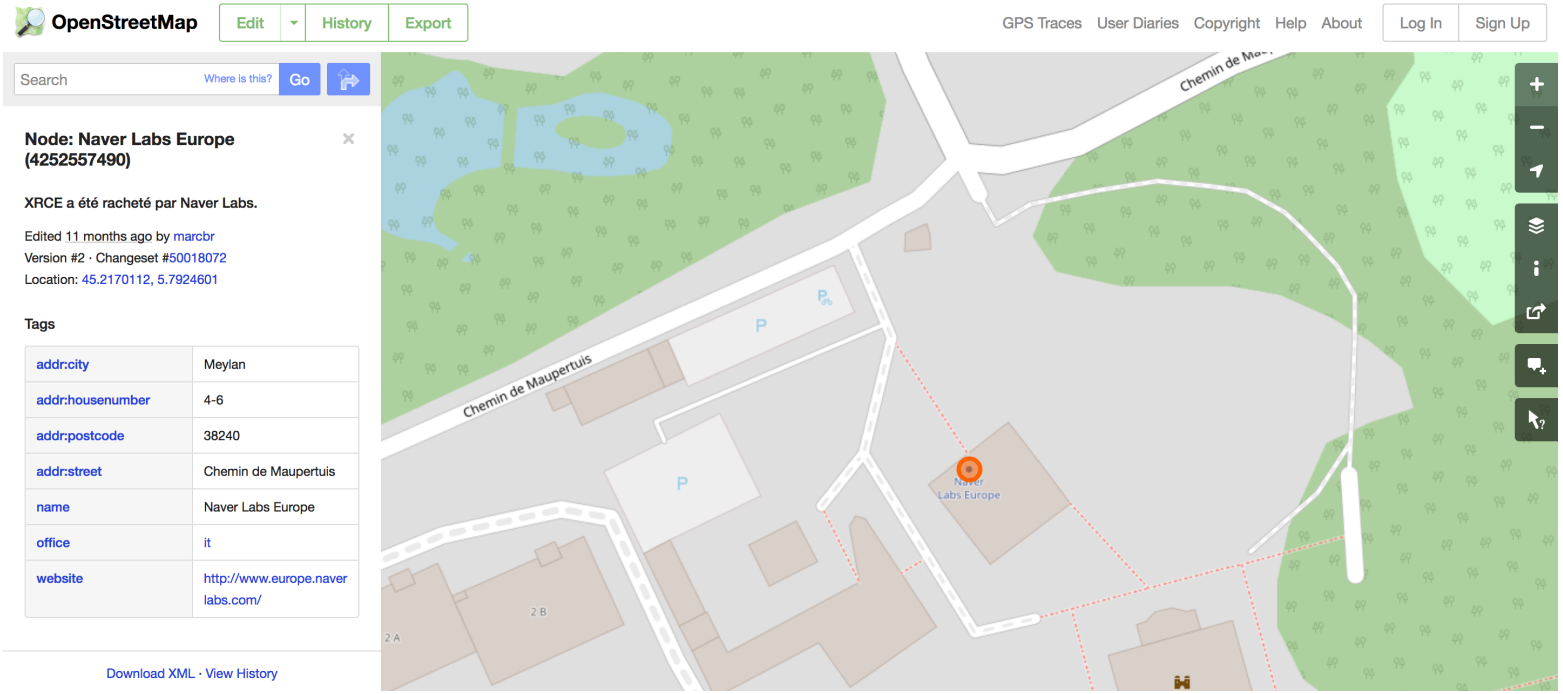
NAVER LABS EUROPE

Visual Search - Applications

Many applications

- Geolocalization



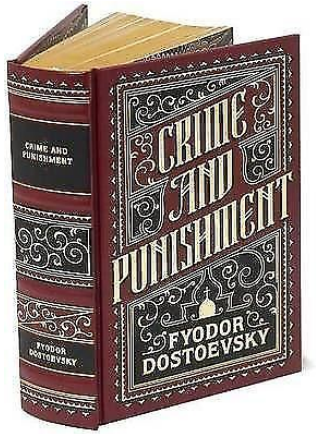


NAVER LABS EUROPE

Visual Search - Applications

Many applications

- Query for more information
  - Landmarks
  - Paintings
  - Movies
  - Book covers
  - Game covers
  - Packaged food



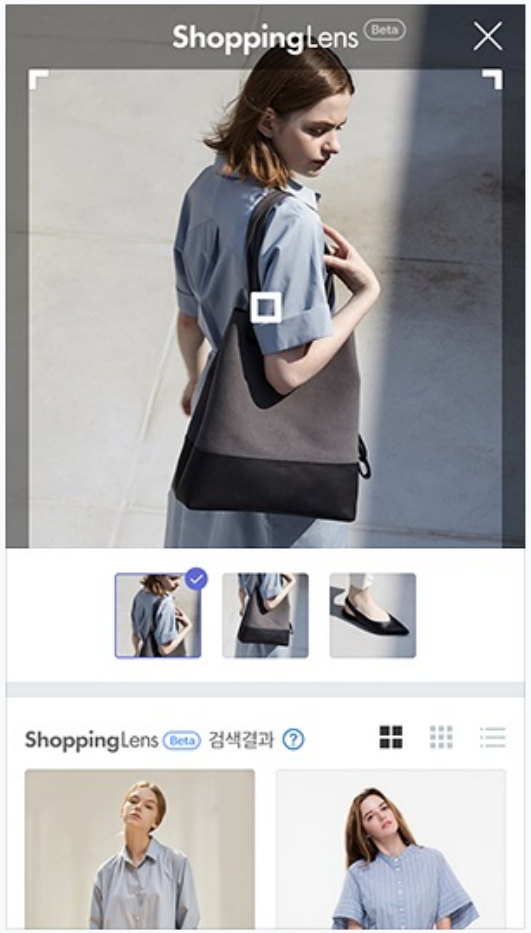
NAVER LABS EUROPE

Visual Search - Applications

NAVER

Many applications

- Shopping interfaces



NAVER LABS EUROPE

Inherent ambiguity

What can the user mean with such a single query?

Inherent ambiguity

What can the user mean with such a single query?

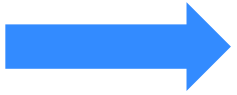
NAVER LABS EUROPE

Inherent ambiguity

What can the user mean with such a single query?

Application dependent!

- Inject prior information
- Leverage training



Inherent ambiguity

What can the user mean with such a single query?

Application dependent!

- Inject prior information
- Leverage training

Information a priori:

- *Design à la main de descripteurs pertinents ou de méthodes pertinentes pour la vérification en fonction de l'application visée*

***Apprentissage**:*

- *On laisse le système apprendre tout ou partie du design, en particulier le système « choisit » les caractéristiques pertinentes pour la tâche*