IDAKS 2018

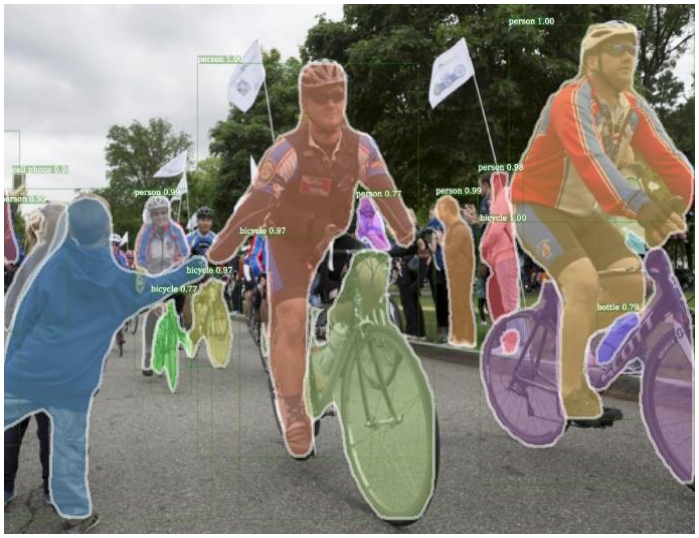
Semantic & interaction: the meeting points between Document Image Analysis and Computer Vision

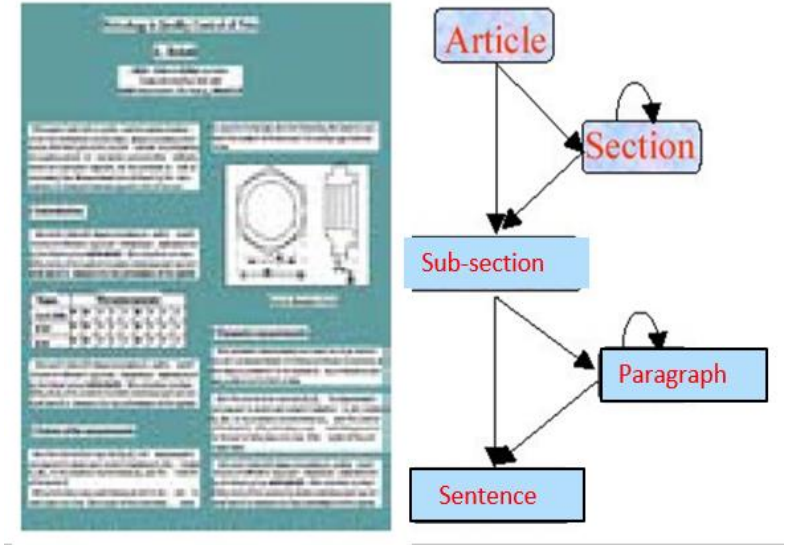
Jean-Yves RAMEL

October 2018

Starting point…

Differences and similarities between CV and DIA problems?



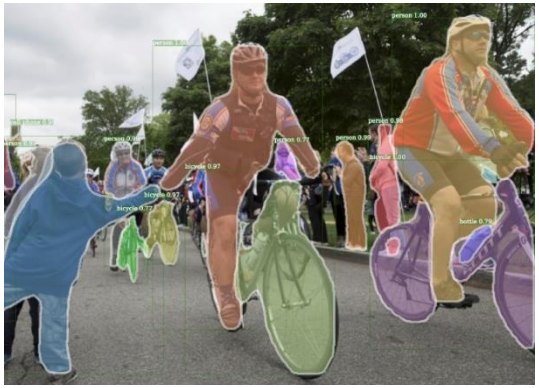


Starting point…

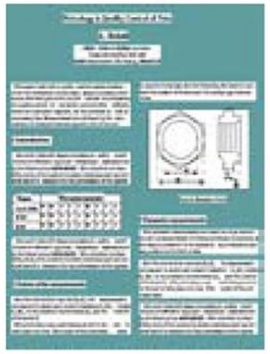
More and more similarity between CV and DIA problems

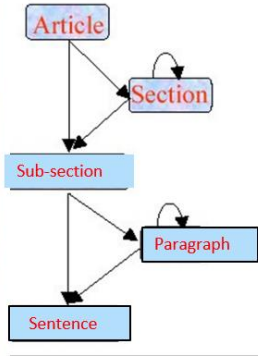
- "New" goals in CV
  - Scene and Image understanding (VQA, VRD, …)
  - More genericity by using machine learning & interaction

- Reformulation: associating semantical labels to images (semantical meta-data)
  - Objects (face, people, cat, car, …) detection (segmentation) and recognition (label)
  - Analysis of Spatial and Temporal relations between objects or subparts of objects ➔ sematic description of the content, behavior, pose and emotion recognition, object tracking, …
  - Using the numerous toolboxes (tensorflow, Detectron, …)



- This goal is targeted since many years in DIA
  - Analysis of spatial and temporal relations between elements is mandatory in OCR, layout analysis, line drawing analysis, …
  - Extraction of elements of contents (EoC) at different levels: lexical, syntactical, semantical
  - Knowledge representations for the analysis of relation between them (dictionaries, models of language, …)





Starting point…

In this new context…

Is DIA expertise useless or not?
Is it CNN compatible ?

What are the good directions?

- CNN ➔ A low level vision of real world (The data are considered as a set of pixels)
- The learning algorithms only consider annotated data to fix the parameters
- The human ➔ a higher level of vision of the real world (looks for a semantical segmentation of the data)
- Contextual information should be integrated such as recently recommended by Yann Lecun in a French conference (RFIAP2018)
- Future systems have to work at a higher level (semantic)
- Future systems have to be more transparent (interaction) and adaptable (plasticity)

Shall deep learning be the mandatory future of document analysis problems? Nicole Vincent, Jean-Marc Ogier, Pattern Recognition, 2019, Vol 86 p281-289

Systems and methods taxonomy

Could we position the DIA and CV methods or systems into categories ?

Is it CNN compatible ?

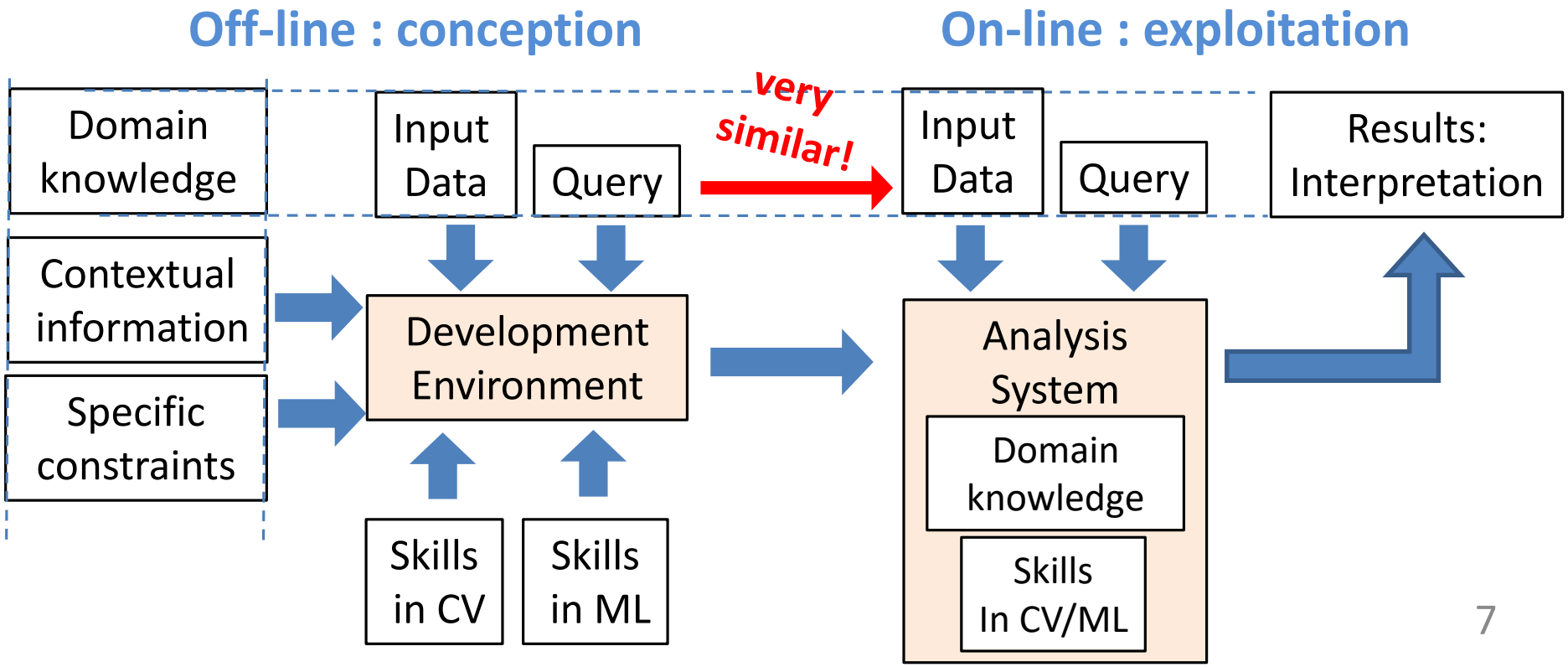
Systems and methods taxonomy

- Categories of DIA and CV methods and systems
  - Static systems (no learning / no interaction) ← Part 1
    - Handcrafted bottom-up or top/down or hybrid approaches (CV & DIA)
  - Adaptable methods (off-line data driven and interaction) ← Part 2
    - Toolboxes for IP, statistical PR and Machine Learning (CV)
    - Syntactical and structural pattern recognition (DIA)
  - Adaptive methods (on-line data driven and interaction) ← Part 3
    - Robustness ➔ plasticity ➔ User interaction, user feedbacks
    - Robustness ➔ plasticity ➔ Incrementality, active and on-line learning,
    - *New constraints (real-time, understandability of parameters and decisions, …)*

- Different goals / deadlocks inside different fields
  - Computer Vision and Image Analysis (matrix, vectors, datasets)
  - Pattern Recognition and Machine Learning (matrix, vectors, datasets)
  - Data and Knowledge Representation (models, architectures, graphs, …)
  - Understanding Visualization, CHI, …

Static handcrafted systems

- Inside the system, the designer encodes:
  - All the algorithms for signature extraction and EoC recognition
  - Using the a priori knowledge about the data
  - Regarding the known future inputs (query, images)
  - Without separations between algorithms, levels, models, …

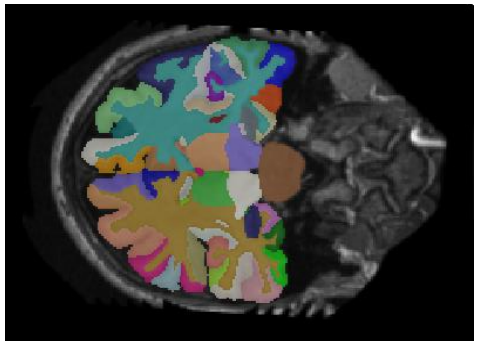


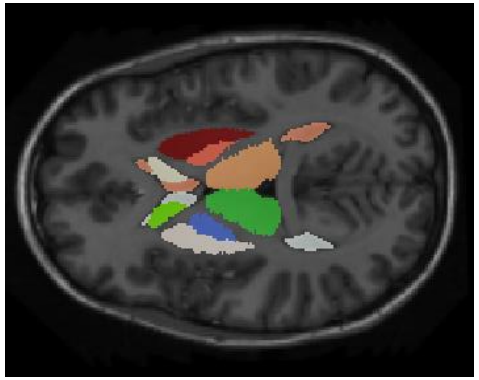
Static systems

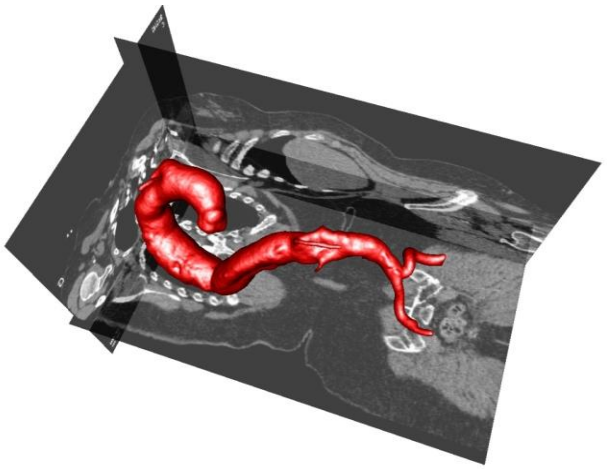
In CV, lot's of methods for segmentation and object detection

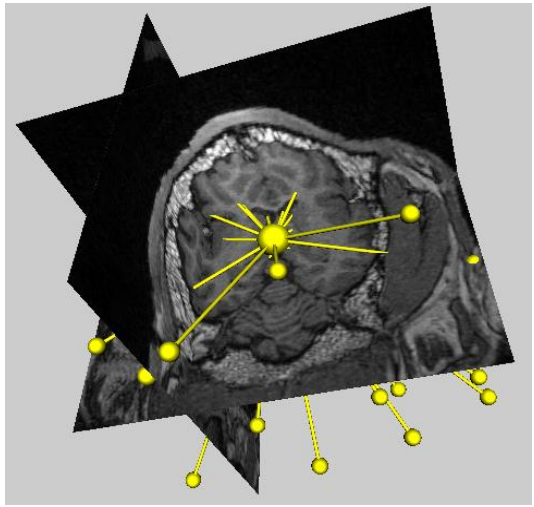
- Global approaches (atlas and scene models)
- Local approaches (active contour and shape model)

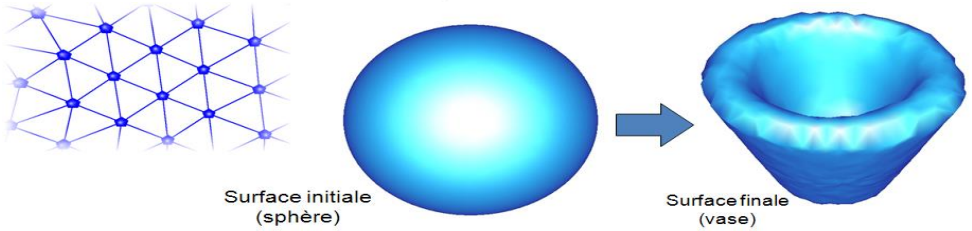
➔ lexical/syntactical level









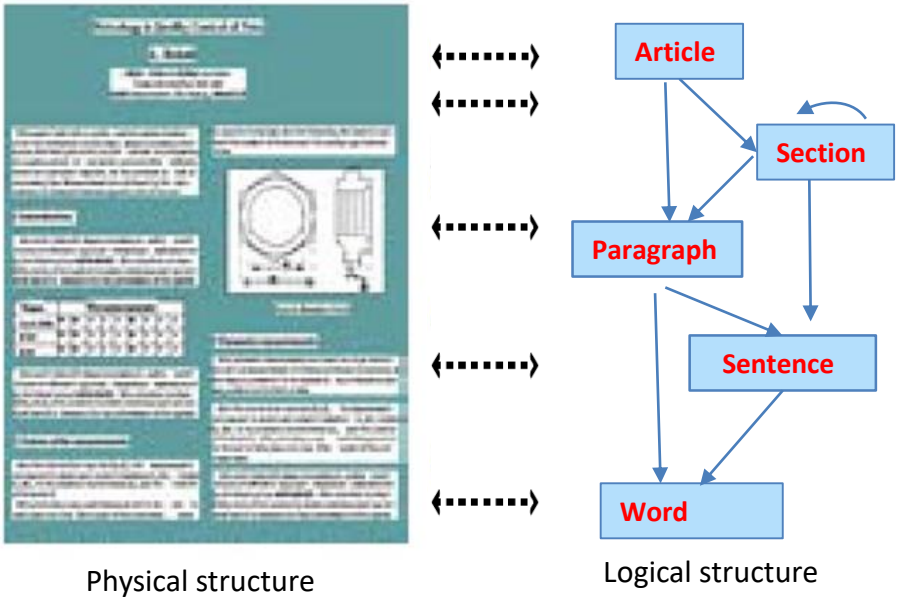


Static systems

More separation between levels in DIA systems (layout analysis) ?

Two kind of structures have been identified by researchers in DIA:

- The logical structure ➔ the generic one corresponding to a priori knowledge about the content of the document (scene model)
- The physical structure ➔ the analysed instance corresponding to the extracted EoC inside the image, each one associated to descriptive features (size, position, number of sub-patterns, …)
- Layout analysis tries to recognize these 2 structures (EoC + relationships identification)
- The analysis of the EoC is usually achieved based on a **rule based system** defined through a grammar (static one).



Physical structure

Logical structure

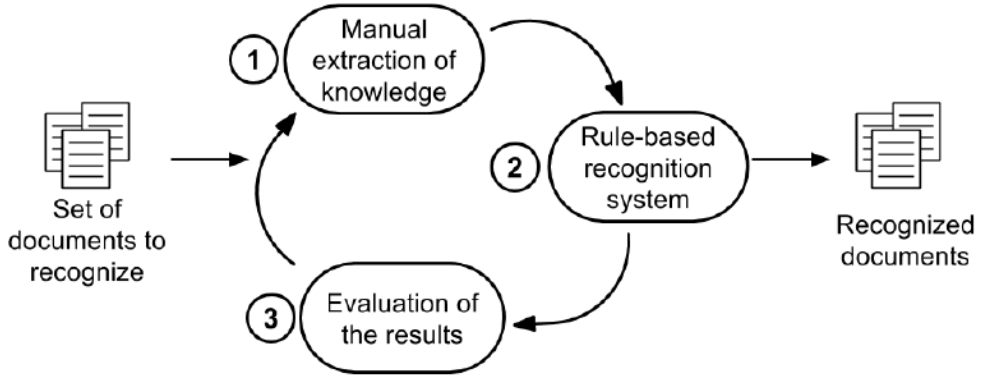


Fig. 1. Current design of a rule-based recognition system, based on manual extraction of knowledge combined with a trial and error approach

Eyes Wide Open: an interactive learning method for the design of rule-based systems. Cérès Carton, Aurélie Lemaitre, Bertrand Coüasnon.

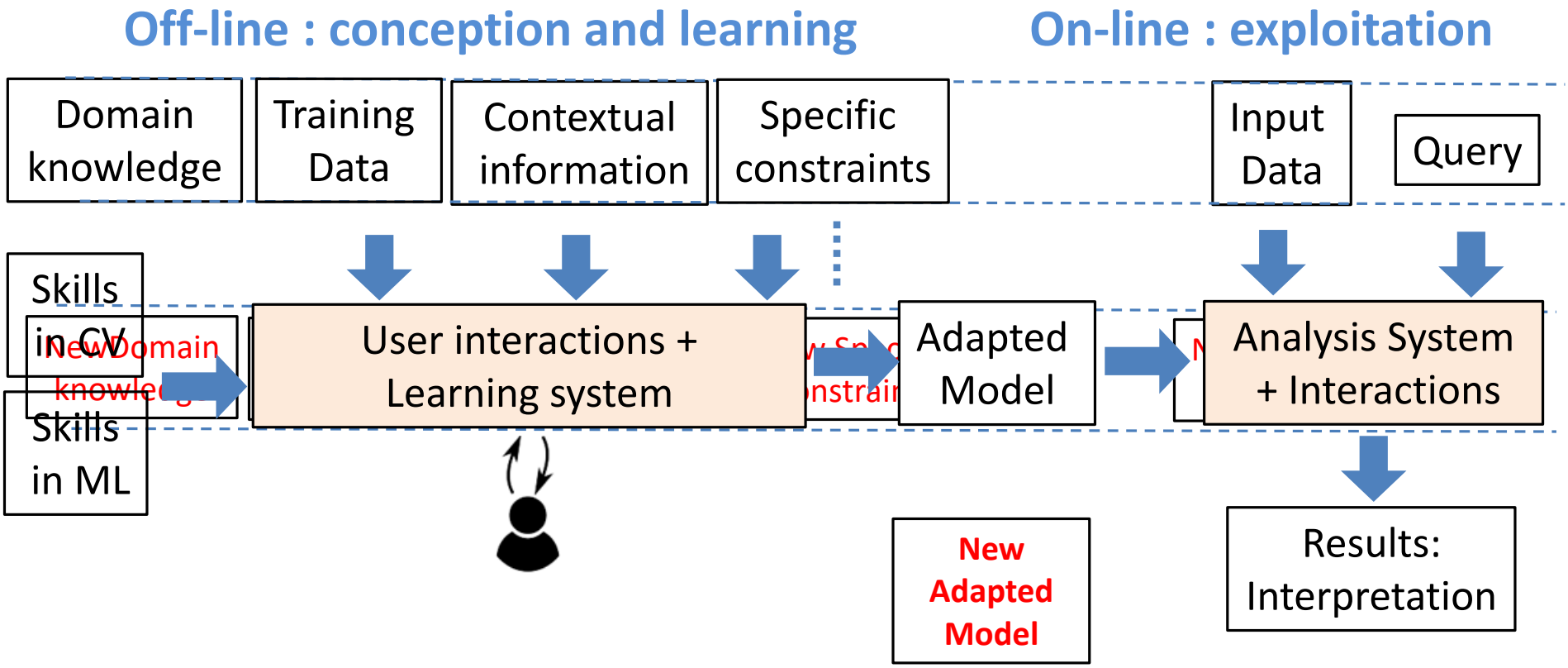
Adaptable & Interactive systems

- Inside the system, the designer/user tune what ?

  - at which level (lexical, syntactical, semantical)
  - in which part of the system (off-line or on-line) ?

Adaptable & Interactive systems

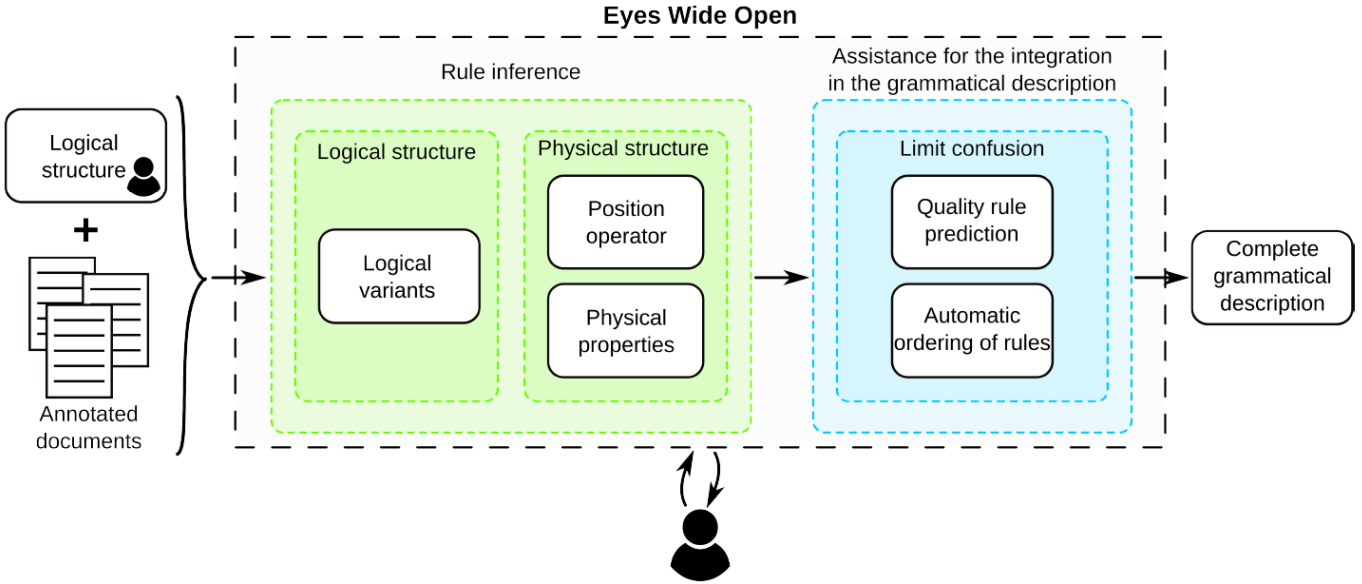
- Inside the system,
  - Adaptable models that can be learned or user-defined



Adaptable & Interactive systems

Interactive learning for the design of rule-based systems (off-line, syntactical level)

- Interactive building / learning of a complete grammatical description of a set of documents
- Main steps:
  - Automatic and exhaustive analysis of an annotated data set (logical structure)
  - The rules are built progressively using a clustering algorithm
  - The interaction with the **grammar writer** brings semantic in the automatically inferred structures.
  - Evaluation of the pertinence of the built grammar

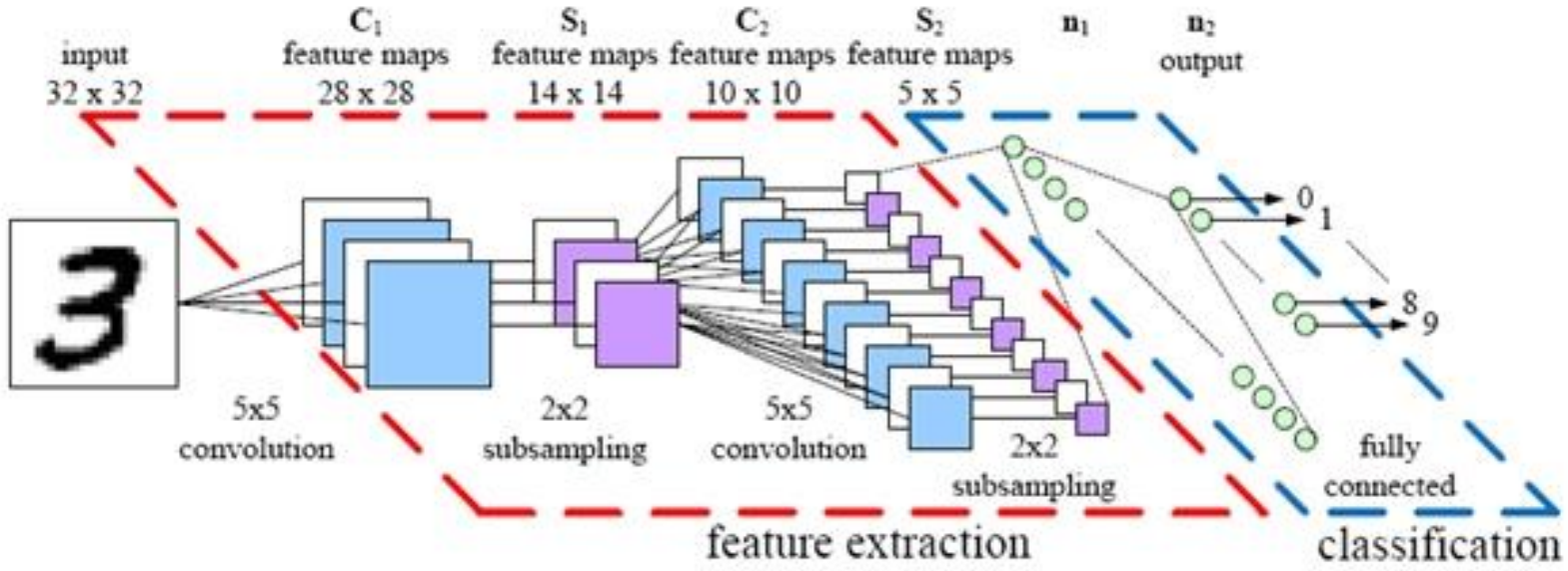


- Advantages of the syntactical methods → expressiveness, understandable, introduction of user knowledge
- Without their main drawbacks→ time needed to adapt the system to a new type of document

Eyes Wide Open: an interactive learning method for the design of rule-based systems. Cérès Carton, Aurélie Lemaitre, Bertrand Coüasnon.

Adaptable & Interactive systems

- Inside Deep Learning system,
  - Adaptable semantic models that can be learned or user-defined ?

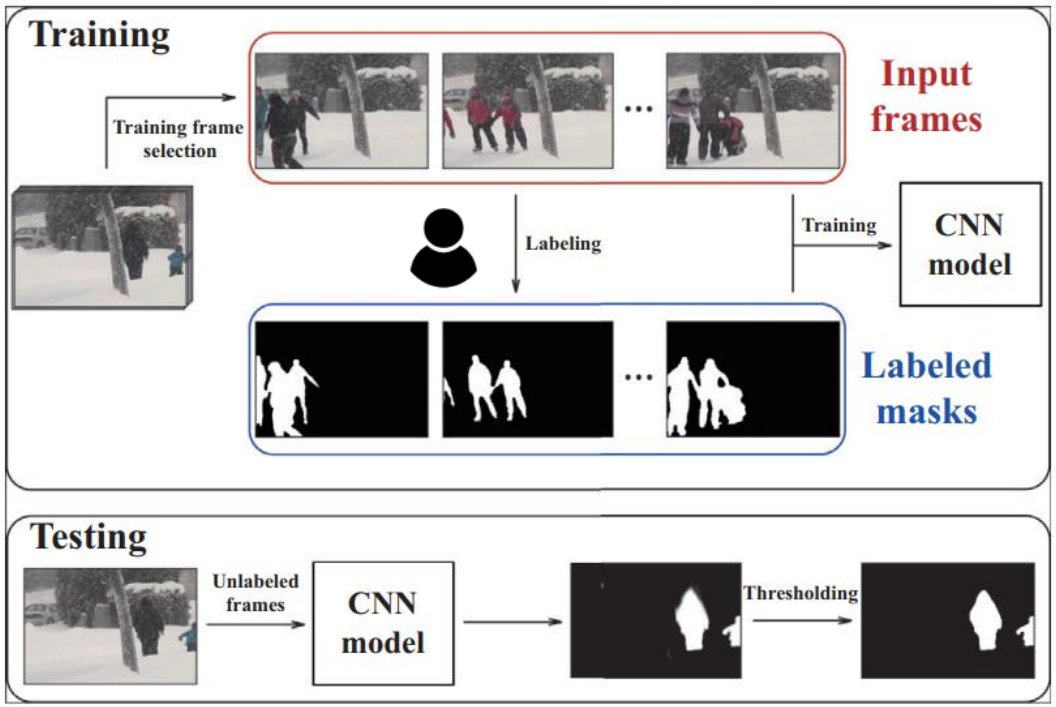


As I am not specialist of CNN, I wonder
Can we do more than automatic features selection
(lexical level, off-line)?

Adaptable & Interactive systems

Interactive (deep) learning ➔ Only off-line and at the lexical level ?

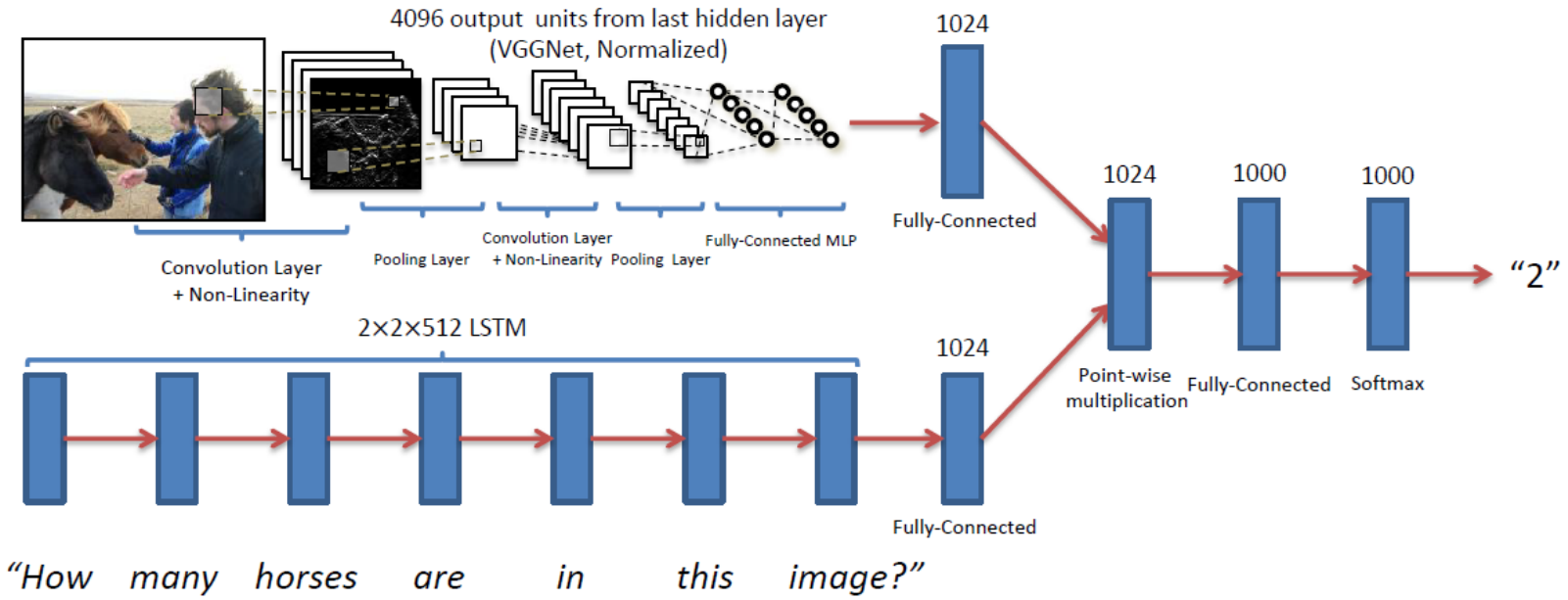
- The users can also interact with the training data (off-line)
- Transfer learning (off-line): multi-task learning, featuriser, …
- Curriculum learning (off-line)
- …



Dealing with semantic models

Semantic Models for Visual Question Answering

- VGGNet to encode the image content
- LSTM to encode the question
- Question and images features are transformed into a common space and pass through a FCL to select the best answers
- Is it really a semantic model?

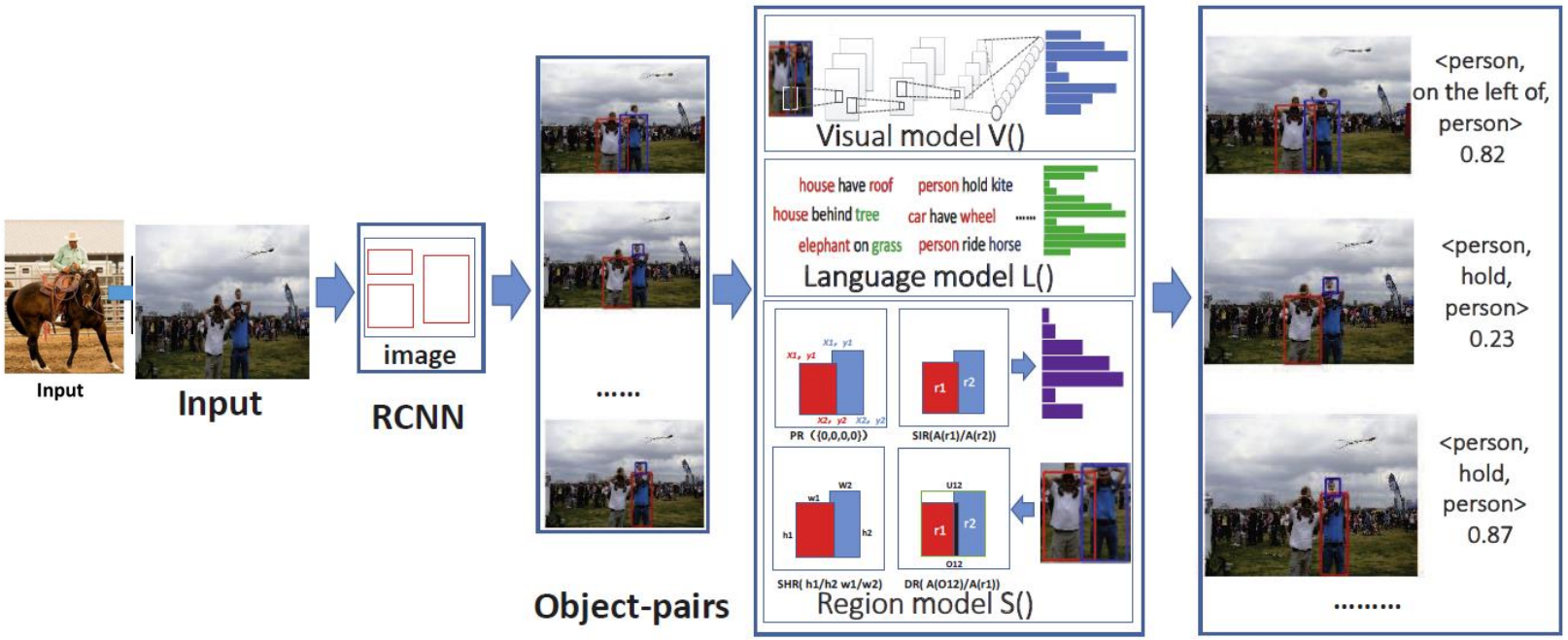


VQA: Visual Question Answering Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

Dealing with semantic models

Semantic Models for Visual Relationship Detection

- Introduction of a more clear separation between different models/levels
  - Visual model (CNN features)
  - Language model (dictionaries of n-grams)
  - Region model (spatial relation: distance, size, position, …)

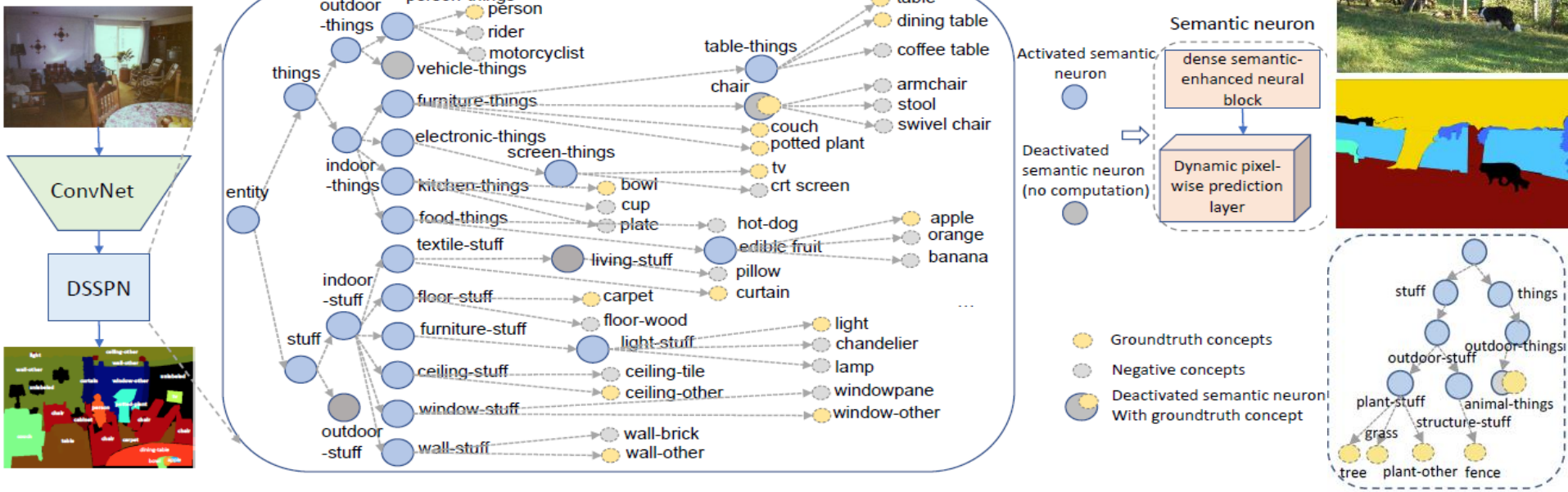


VISUAL RELATIONSHIP DETECTION WITH OBJECT SPATIAL DISTRIBUTION Yaohui Zhu, Shuqiang Jiang, Xiangyang Li ICME 2017

Dealing with semantic models

More structural ML models

- Proposition of a **Dynamic-structured Semantic Propagation Network**
- A semantic hierarchy (neuron graph network) ➔ Model of the world (manually built?)
- CNN features are propagated into a graph for hierarchical pixel-wise recognition
- Sub-graphs activation during training/testing (feed-forward and back propagation)
- Use of a Hierarchical description (document structures)

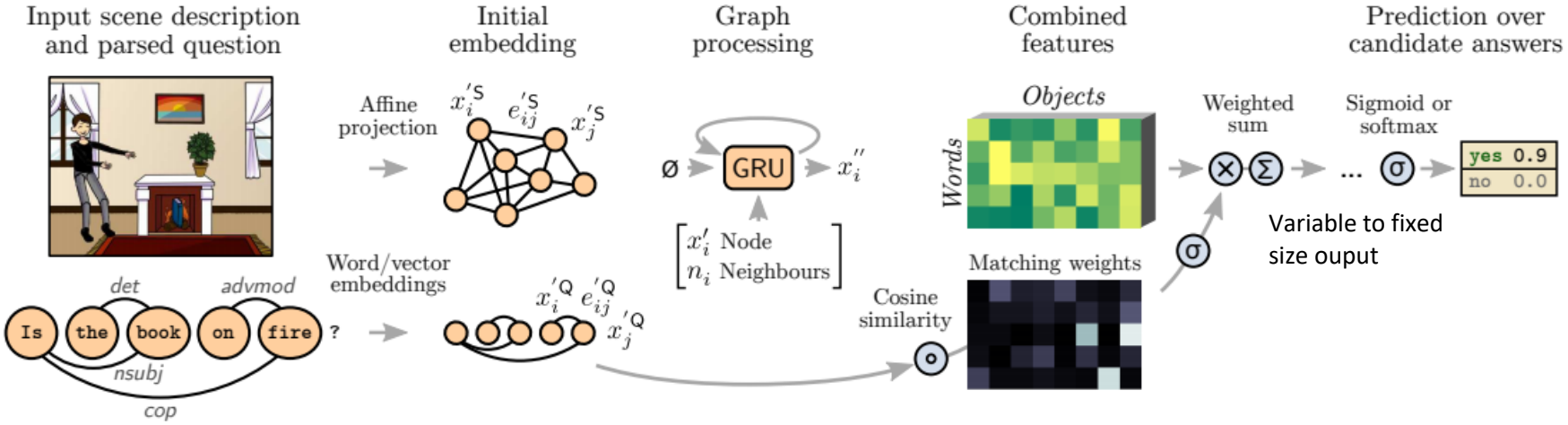


Dynamic-structured Semantic Propagation Network Xiaodan Liang Hongfei Zhou Eric Xing arXiv:1803.06067v1 [cs.CV] 16 Mar 2018

Dealing with semantic models

More structural representations (graphs)

- A scene graph with attributed nodes (objects) and edges (spatial relationships)
- A question graph with node (words) and edges (type of syntactic)
- A recurrent unit (GRU) transform the 2 graphs into word and object features
- Both features are concatenated pairwise (inside a matrix)
- A final classifier predicts scores over a fixed set of candidate answers
- One step toward sub-graph matching ?



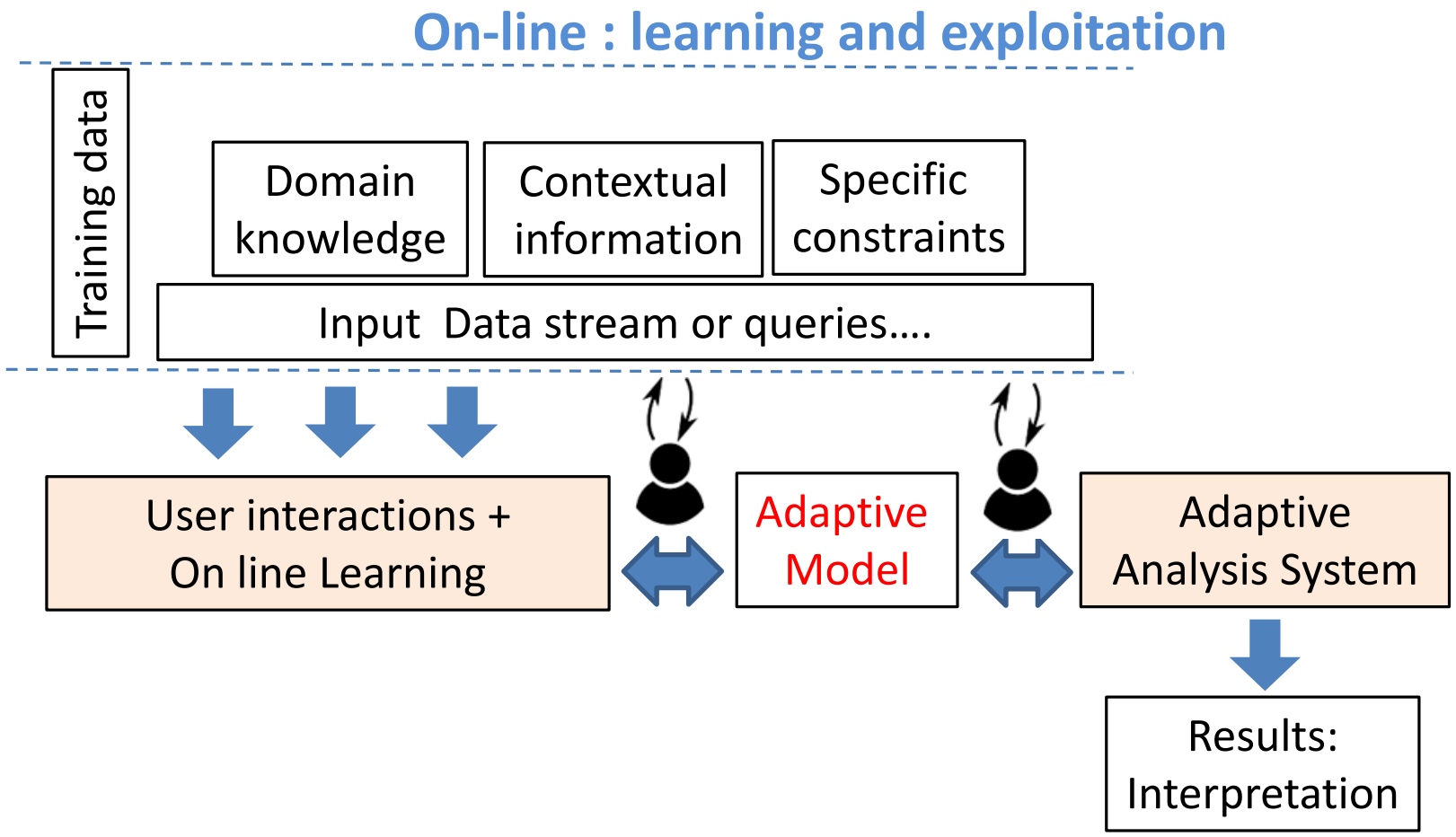
Graph-structured representations for visual question answering. Damien Teney, Lingqiao Liu, and Anton van den Hengel. CVPR, July 2017.

Adaptive & Interactive systems

- Inside the system,
  - At which level is the adaptation?
  - At what time (on-line or off-line)?
  - Names of techniques?

Adaptive & Interactive systems

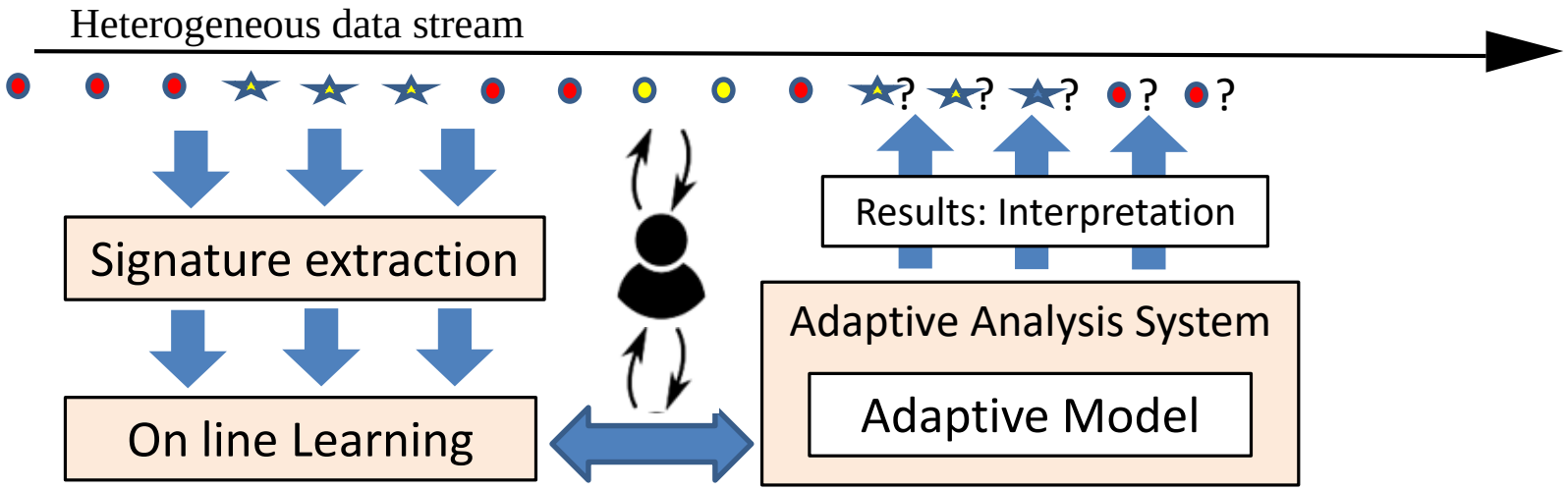
- Inside the system,
  - Adaptive models are updated on-line
  - Adaptation are supervised by the system or by the user



Adaptive & Interactive systems

1. Online learning

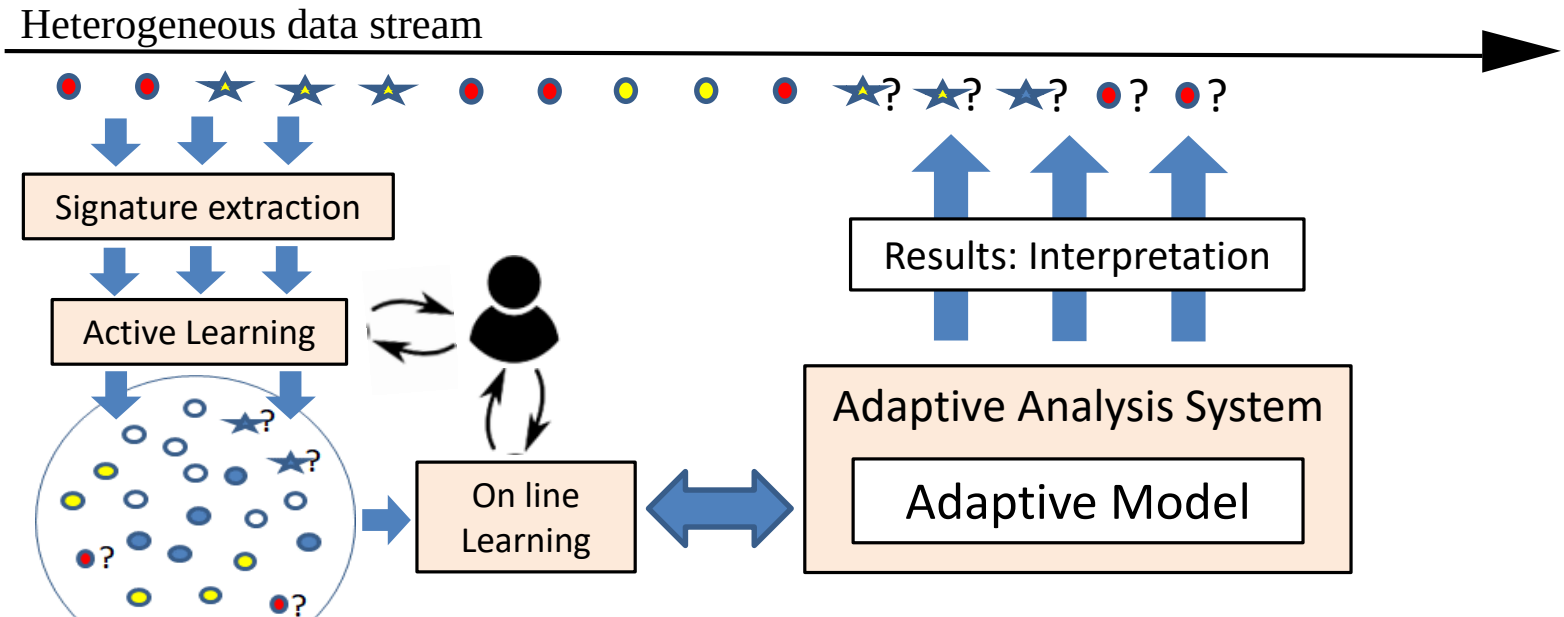
- Requirements for online evolving systems are:
  - Incremental learning from few initial learning data
  - Each data sample must be processed only once
  - Adapt models according to new data without requiring all the original data
  - Preserve previously acquired knowledge (no catastrophic forgetting)
  - Memory and computing time must be limited
  - System learning can be interrupted and its quality shouldn't be altered



Adaptive & Interactive systems

2. On-line active learning

- A classifier can achieve equivalent performance with only part of the learning data, if those data have been correctly chosen.
- The learning system itself will choose which data samples will be used
- Need method to evaluate the classifier confidence during recognition (Sampling decision)
  - **Ask the users** to decide when to query the label of the sample
- Decide the label of the new samples (Semi supervised learning)
  - **Ask the users** to label data samples for which the system is likely to make a recognition error and which will be very interesting for the evolving classifier learning



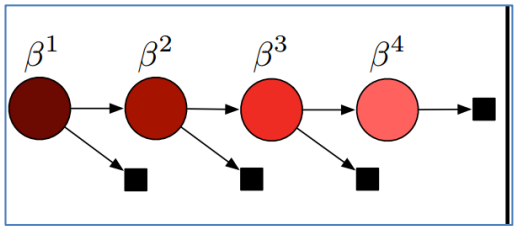
Adaptive & Interactive systems

3. Budgeted Learning & incremental classification

- New systems need problem resolution under time and memory constraints
- One possible solution to explore is **budgeted learning & classification**

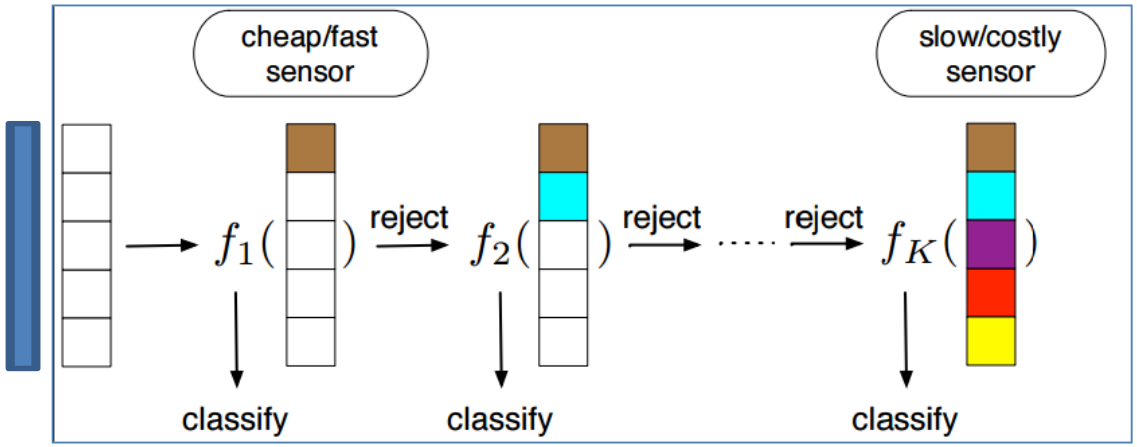
Main ideas

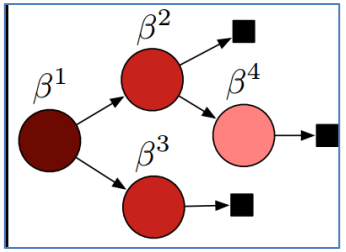
- At test time, compute & use costly features only if necessary (utility scores)
- **Definition of new learning and classification strategies / architectures** cost sensitive ones →



Example of architecture

A cost vector is associated to the feature vector





Supervised sequential classification under budget constraints K Trapeznikov, V Saligrama - Artificial Intelligence and Statistics, 2013
Supervised Machine Learning Under Test-Time Resource Constraints: A Trade-off Between Accuracy and Cost. Zhixiang Xu Washington University in St. ETDs Thesis. Louis 2014

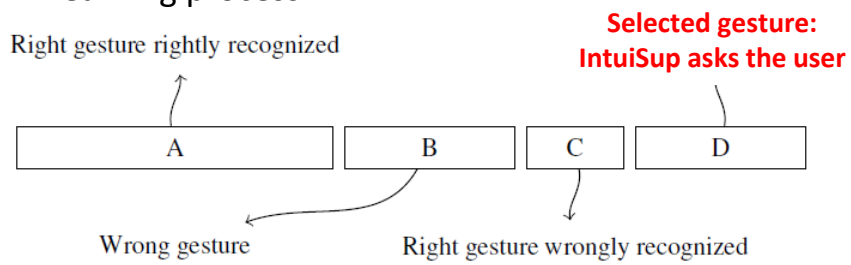
Adaptive & Interactive systems

On line active learning in DIA

Online and Active supervision of a recognition system

- Context = customized gesture command
- Goal = Optimizing user-system interaction in this cross-learning context ➔ stream sampling
- Method = Evolving fuzzy classifier + IntuiSup

- **Combining implicit and explicit supervision**
- The implicit supervision mechanism takes advantage of user next action to implicitly label the majority of correctly classified data
- The **explicit supervision** mechanism makes it possible to learn from complex data samples that are hard to recognize, and from which it is very beneficial to learn
- The Evolving Sampling by Uncertainty (ESU) algorithm triggers user interactions using the classifier confidence measure as input.
- Possibility to increase the **interaction rate** at the beginning of system use and during concept drifts to fasten system learning process.



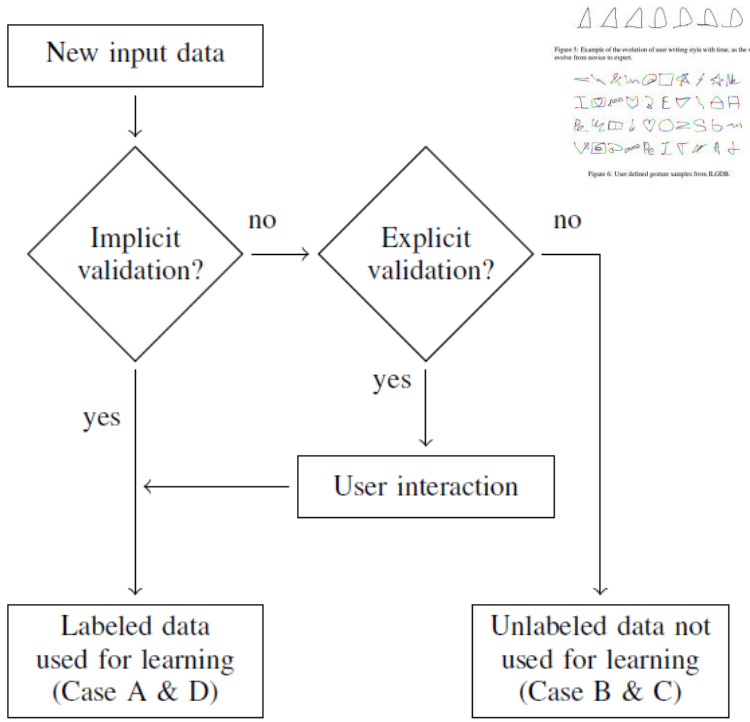


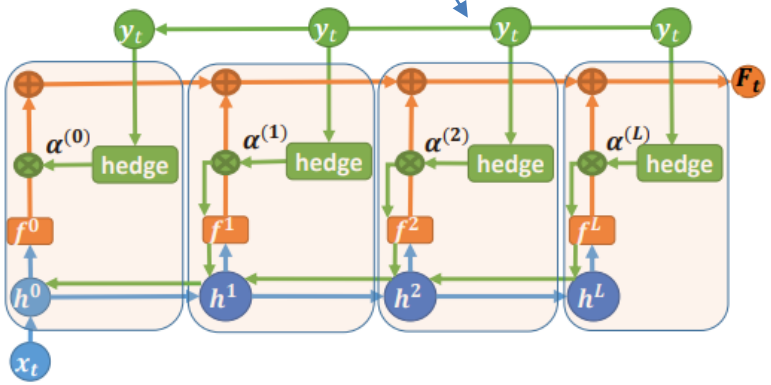
Figure 4: Online active learning supervision process.

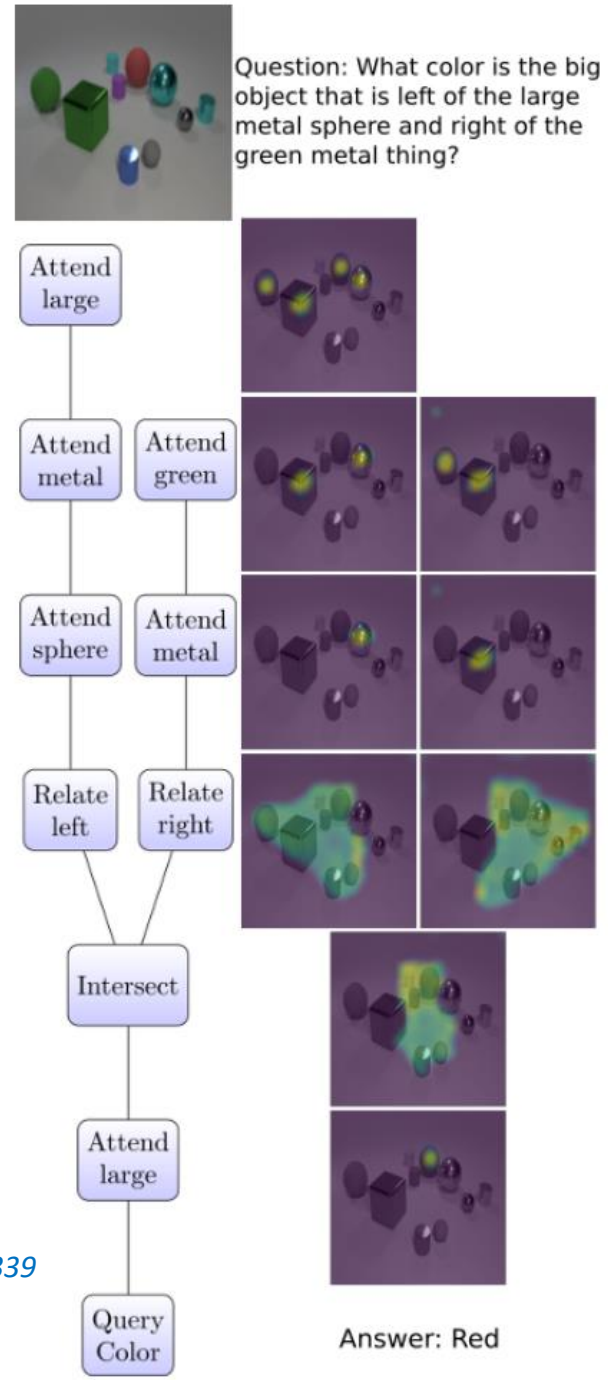
Online Active Supervision of an Evolving Classifier for Customized-Gesture-Command Learning. Manuel Bouillon, Eric Anquetil. Neurocomputing, 2017

Adaptive & Interactive systems

Toward transparent, interactive, incremental deep learning systems ?

- **Design of (transparent) DL systems** TbD-net (with resNet)
- **On line supervision of the learning**
- **On-line deep learning** Hedge backpropagation
- **Active deep learning**





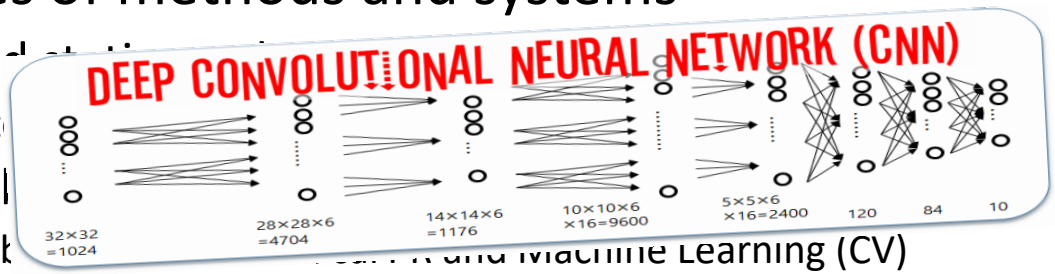
Online Deep Learning: Learning Deep Neural Networks on the Fly. Doyen Sahoo, Quang Phan, Jing Lu, Steven C.H. Hoi. arXiv:1711.0370

Sequential Labeling with online Deep Learning. Gang Chen, Ran Xu, Sargur Srihari. arXiv:1412.339

Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning David Mascharka∗1 Philip Tran2 Ryan Soklaski1 Arjun Majumdar. arXiv:1803.0526

Conclusion (Version 1)

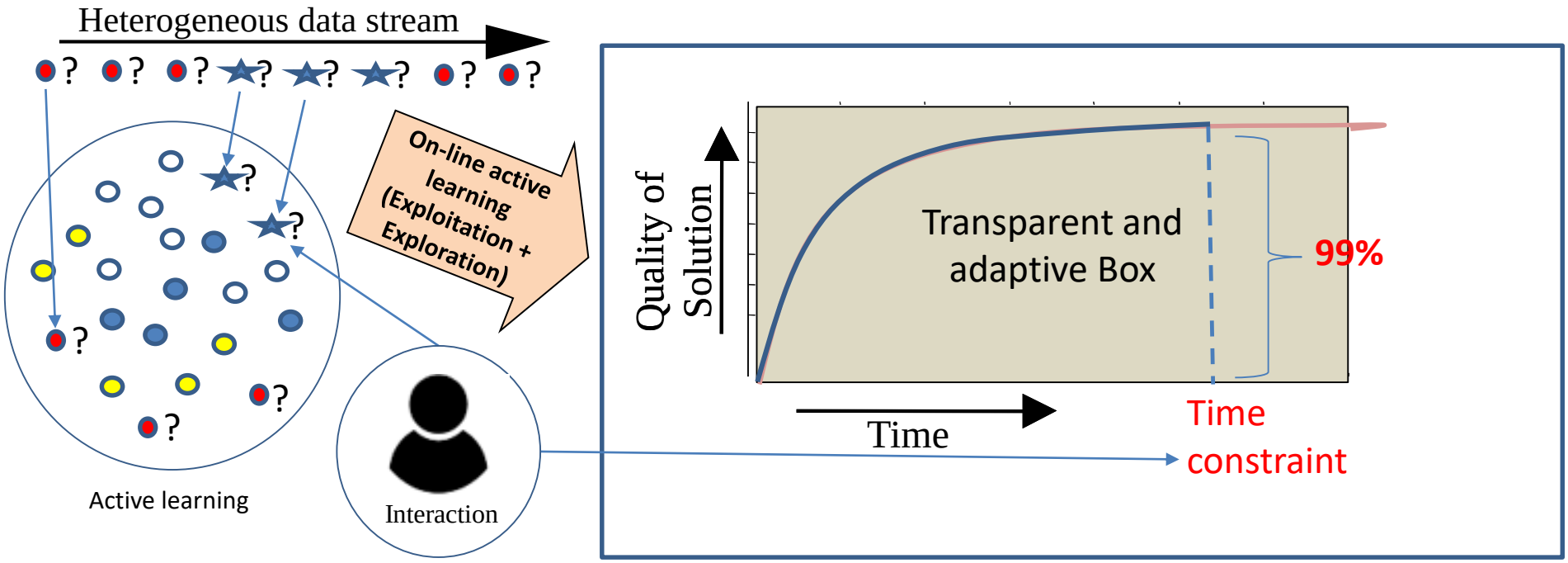
- Categories of methods and systems
  - Adapte
    - Bott
  - Adaptal)
    - Toolb and Machine Learning (CV)
    - Syntactical and structural pattern recognition (DIA)



  - Adaptive methods (on-line data driven and interaction)
    - Robustness ➔ plasticity ➔ User interaction, user feedbacks
    - Robustness ➔ plasticity ➔ Incrementality, active and on-line learning,
    - *New constraints (real-time, understandability of parameters and decisions, …)*

- Different goals / deadlocks inside different fields
  - Computer Vision and Image Analysis (matrix, vectors, datasets)
  - Pattern Recognition and Machine Learning (matrix, vectors, datasets)
  - Data and Knowledge Representation (models, architectures, graphs, …)
  - Perception ⇔ Understanding, Visualization, CHI, …

Conclusion (version 2)



- What should we remember?
  - Toward robust adaptive system design instead of mono-dataset accurate system
  - Adapted & static methods → a lot of operational toolboxes in CV, PR, ML, …
  - Adaptable methods → Off-line learning (from datasets) and from human interaction
  - Adaptive, incremental, interactive systems ➔ Human supervision, active learning
  - Time and memory constraints ➔ Anytime, budgeted & distributed systems
  - My keywords → Active, Budgeted, Interactive, Incremental **but less sequential more dynamic** (perceptive cycles, saccades ?)

Thanks

The Workshop > Important dates

Dates:

- GBR 2019: 19-21th June 2019
- Regular paper submission: 12th December 2018
- Notification of acceptance: 1st February 2019
- Camera ready due: 15th March 2019
- Early Registration: 15th March 2019