SmarterAnalytics

IBM

Semantics of Visual Discrimination

John R. Smith
IBM T. J. Watson Research Center
jsmith@us.ibm.com

February 2015

Many Thanks To …

IBM T. J. Watson - Multimedia Research Team


Liangliang Cao


Michele Merler


Noel Codella


Matthew Hill


Quoc-Bao Nguyen


Wei Liu


Rosario Uceda-Sosa

IBM T. J. Watson - Exploratory Computer Vision Team


Sharath Pankanti


Rogerio Feris


Quanfu Fan


Lisa Brown


Nalini Ratha


Chiao-Fe Shu


Chung-Ching Lin

IBM Research Collaborators


Gang Hua


Shih-Fu Chang


John Kender


Daniel Ellis


Felix Yu

Big News Event in 2005

How many cameras?

Similar Big News Event Eight Years Later

How many cameras?

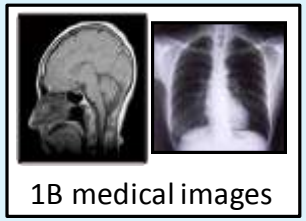
Cannot count them all … image and video is here!

Massive Multimedia is the Biggest Wave of All!

Safety / Security

10s millions cameras

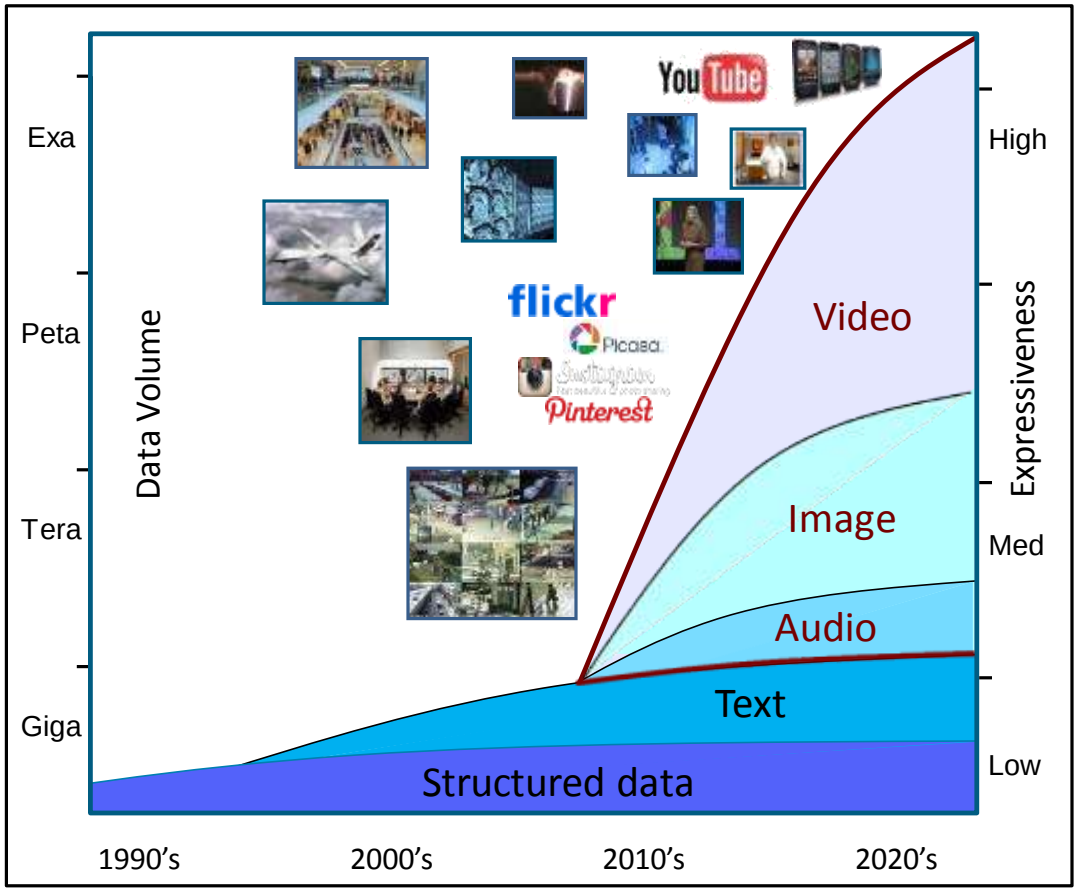
Medical



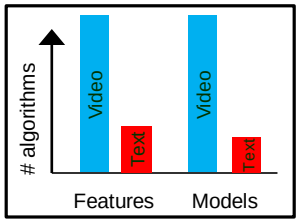
1B medical images

Customer

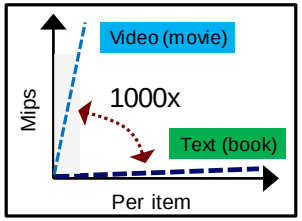
1B camera phones



>> Sophistication



>> Computation



Media

100 video hrs/minute

Wide Area Imagery

100's TB per day

Digital Marketing

12% of video views

Enterprise Video

Used by 1/3 of enterprises

Key Messages:

Image and video data is growing in volume and importance

Manual analysis is not scalable … cannot keep up

Vision system development has historically required deep expertise

Expert Analysis → Data-driven Visual Learning

Increasing availability of labeled training data

Emergence of increasingly sophisticated visual learning algorithms

Focus needed on Visual Semantic Modeling

How to best model visual world (concepts and relationships)

How to combine visual semantic modeling with learning systems

Use Cases

Multiple Industry Problems Require Large-Scale Visual Content Analysis

Safety and Security (IBM Intelligent Video Analytics)

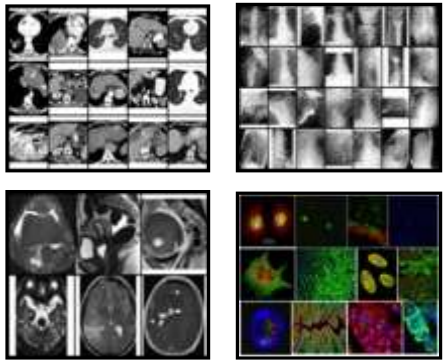
Volume: 10K's managed cameras per city

Velocity: real-time alerts, 20M video events/day

Variety: street scenes, rail stations, crowds, people, environmental conditions

Veracity: analysis of complex activities (trip wires, abandoned objects)

Healthcare Cognitive Systems (IBM Watson Research)

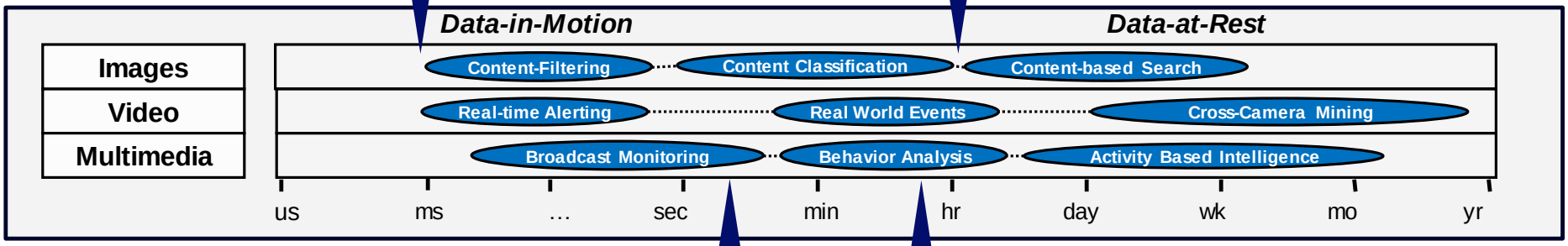


Volume: 1B medical images per year (growing 20-40% /yr)

Velocity: 50K radiology images per day per radiology dept.

Variety: images, video, text, patent records, cases, scientific literature, ontologies/semantics

Veracity: subjective interpretation across millions of categories (modalities, body views, organ systems, pathologies, anomalies)



Enterprise Content Management (IBM IMARS)

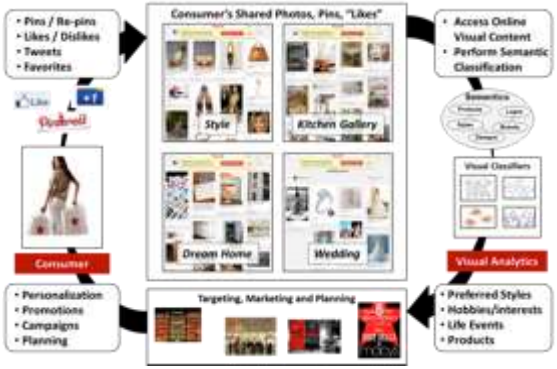
Volume: 70PB broadcast/yr, 40K hrs per news archive

Velocity: 100 video-hrs/min to YouTube

Variety: mobile, user generated, professional

Veracity: robust content extraction for objects, places, scenes, activities, people

Retail and Mobile Commerce (IBM System V)



Volume: 500B consumer photos/yr

Velocity: 100M customers per week for large retailers

Variety: transient and dynamic content

Veracity: predicting consumer attributes from diverse sources including visual data (images and video)

Safety and Security:

Urban Surveillance

Trip Wires

Object Detection

Alerts

Offline Training

Forensic Search

Object Tracking

- Expert developed algorithms
- Limited object detection and tracking
- Limited robustness to environment
- Manual tuning for new deployments

Semantic Extraction

Machine Learning

Visual Recognition

Attributes

Pattern Discovery

Activity Detection

- Large-scale data-driven visual learning
- Semantic extraction including attributes
- Integrated analysis across cameras
- Self-configuration, tuning, adaptation

IBM Intelligent Video Analytics (IVA) for Smarter Cities

Vehicle detection/classification

Person and face attributes

Activities and behaviors

Visual Learning

CAPTURE

ANALYZE

DECIDE

ACT

Traditional Computer Vision

Real world metrics (speed, size)

Trip wires and safety regions

Object detection and tracking

Media-in-the-Wild:

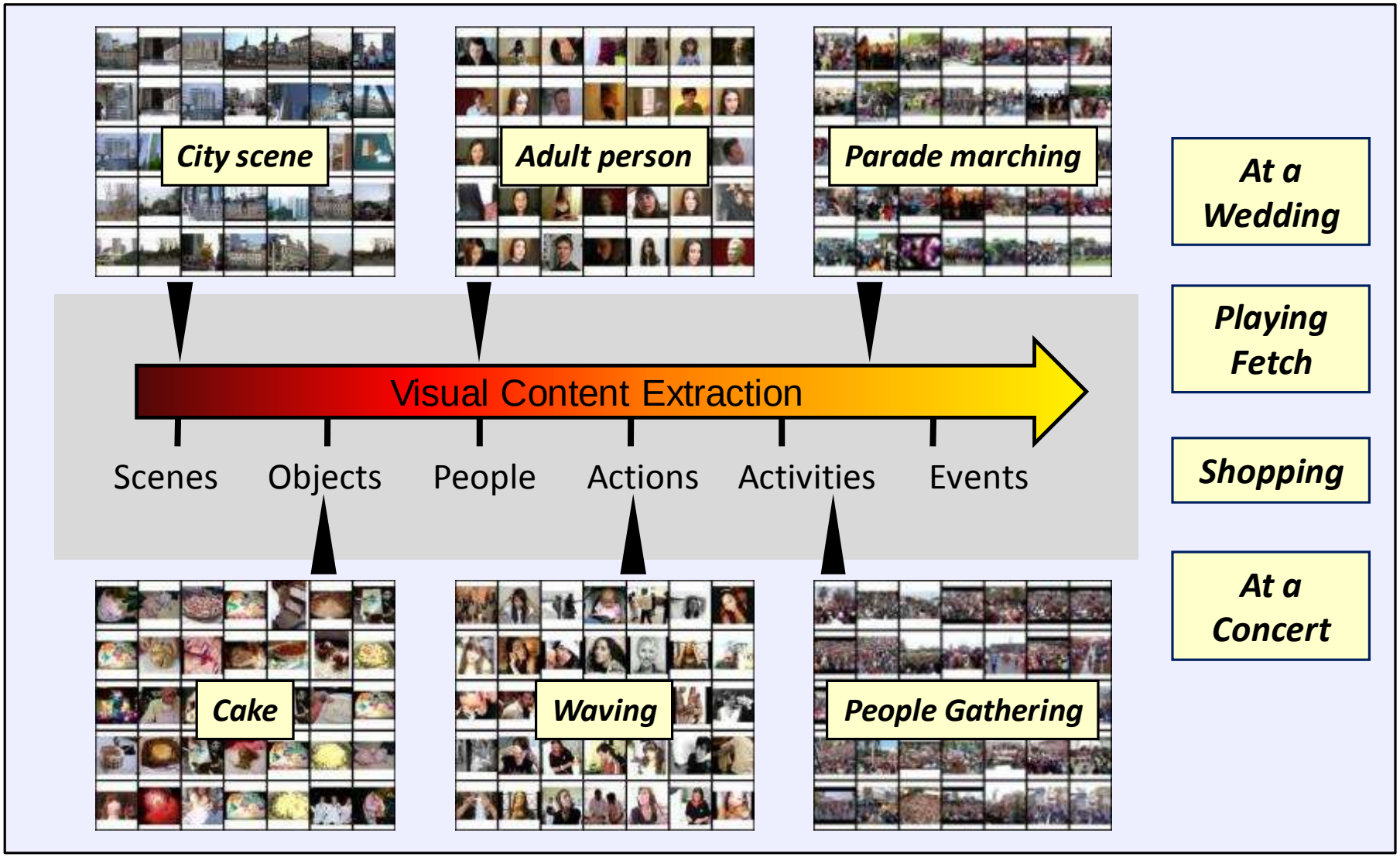
Traditional Media Archives

- Limited indexing and search of *news archives, TV programs, movies*
- Relies heavily on related information (e.g., *metadata, speech transcript, user tags)*

User Generated Content

- Want to understand diverse *user generated content* at a semantic level (e.g., *sports, activities, life events*)
- Extraction of visual insights
- Discovery of patterns across *users, segments, time, geographies*

Visual search uses thousands of Visual Classifiers across dozens of Facets *(Objects, Scenes, Locations, Activities, People, Events, etc.)*



Example: Visual Recognition for Image Semantic Indexing

Demo

Visual Semantic Label Categories

E.g., *Surfing*

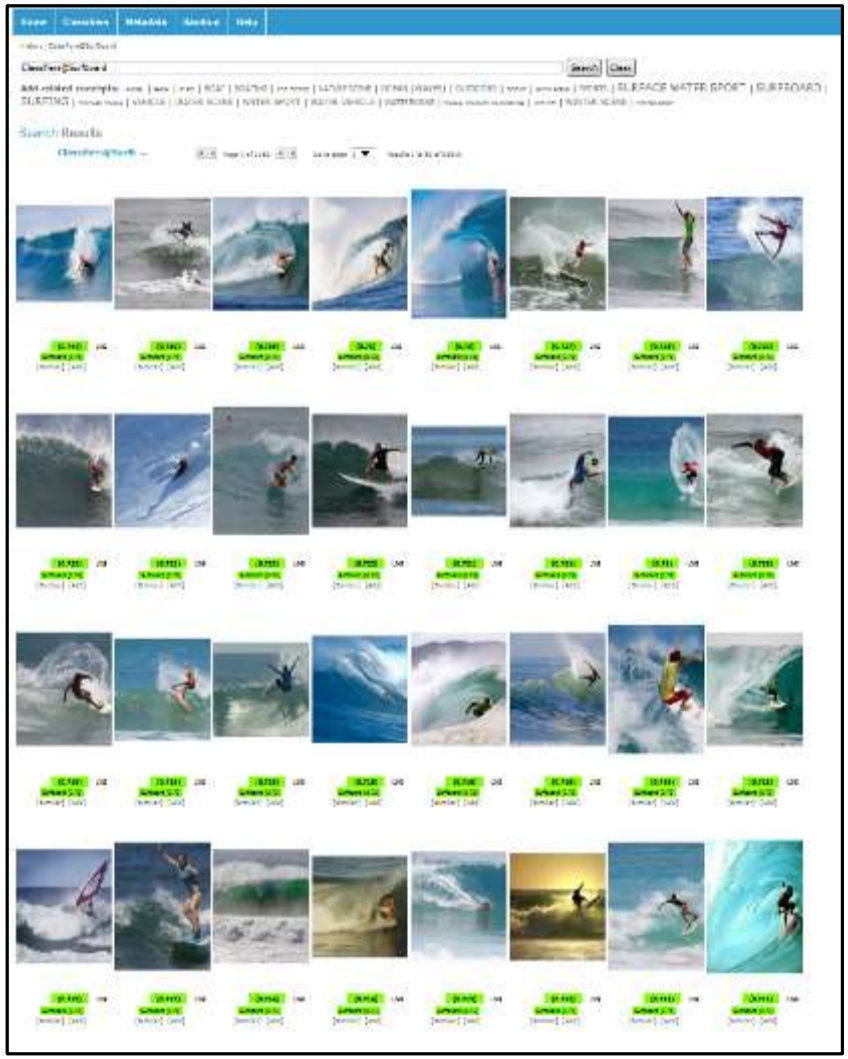
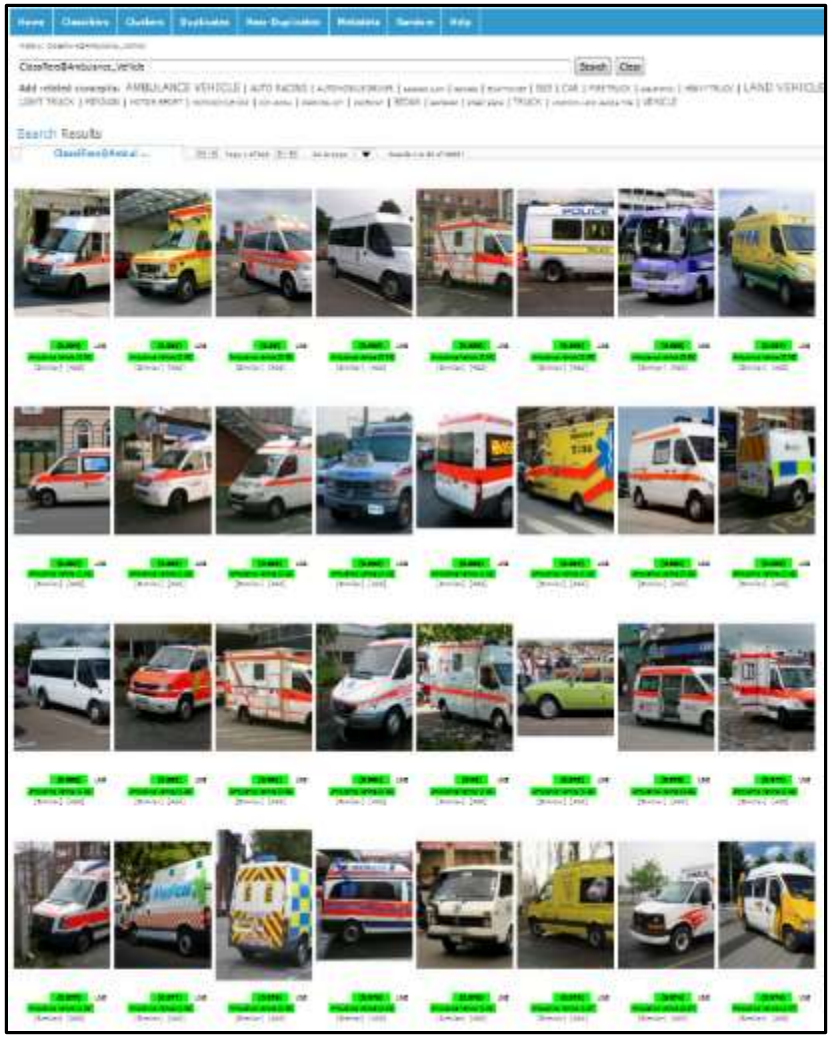
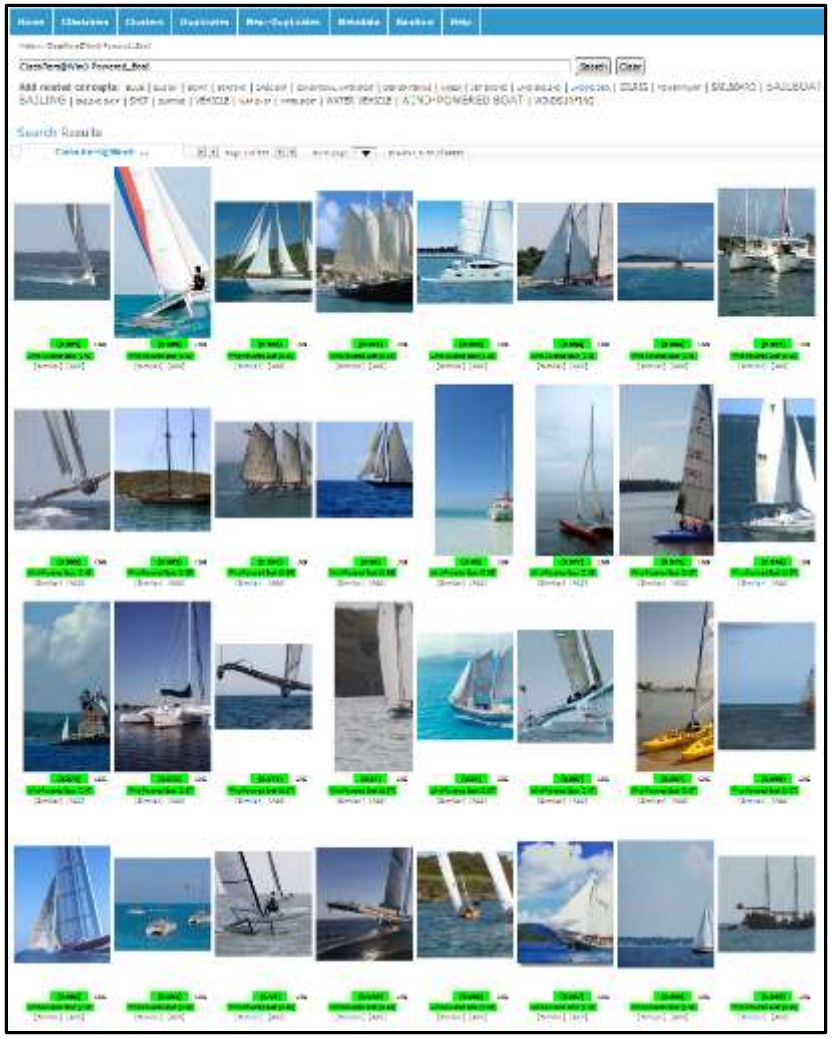


Image Classification Results using Visual Semantic Faceted Hierarchy

[OBJECT] *Ambulance Vehicle*

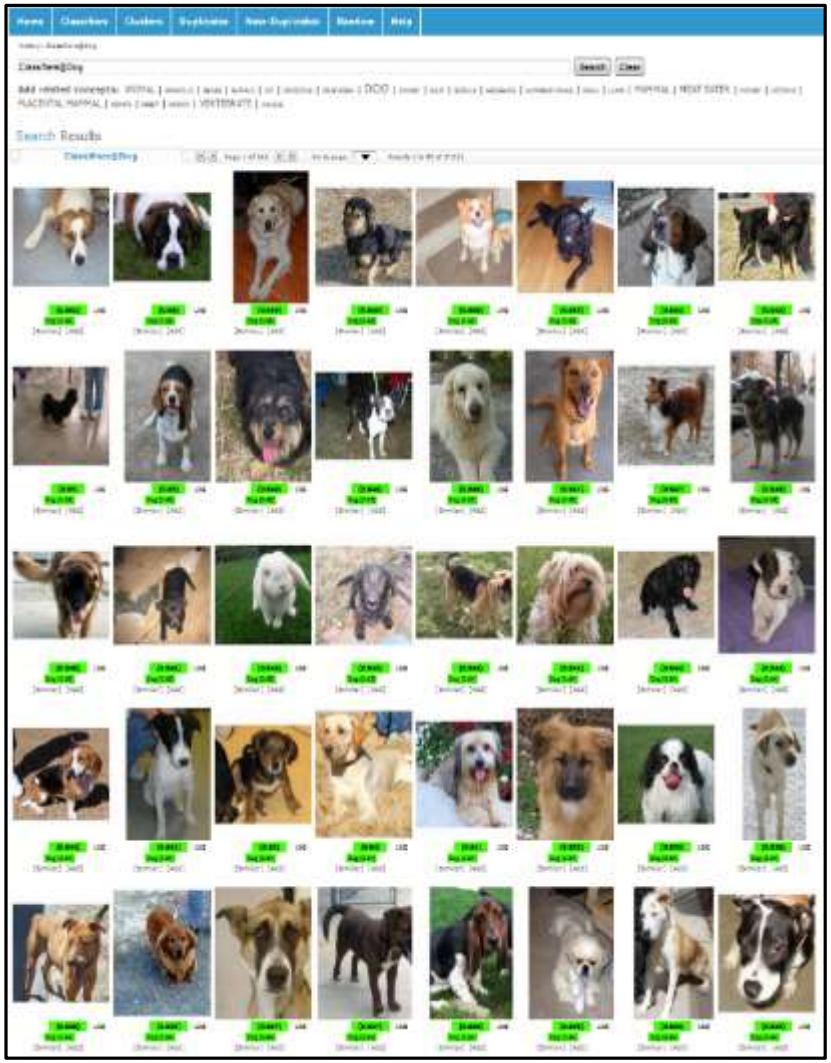


[OBJECT] *Wind-Powered Boat*

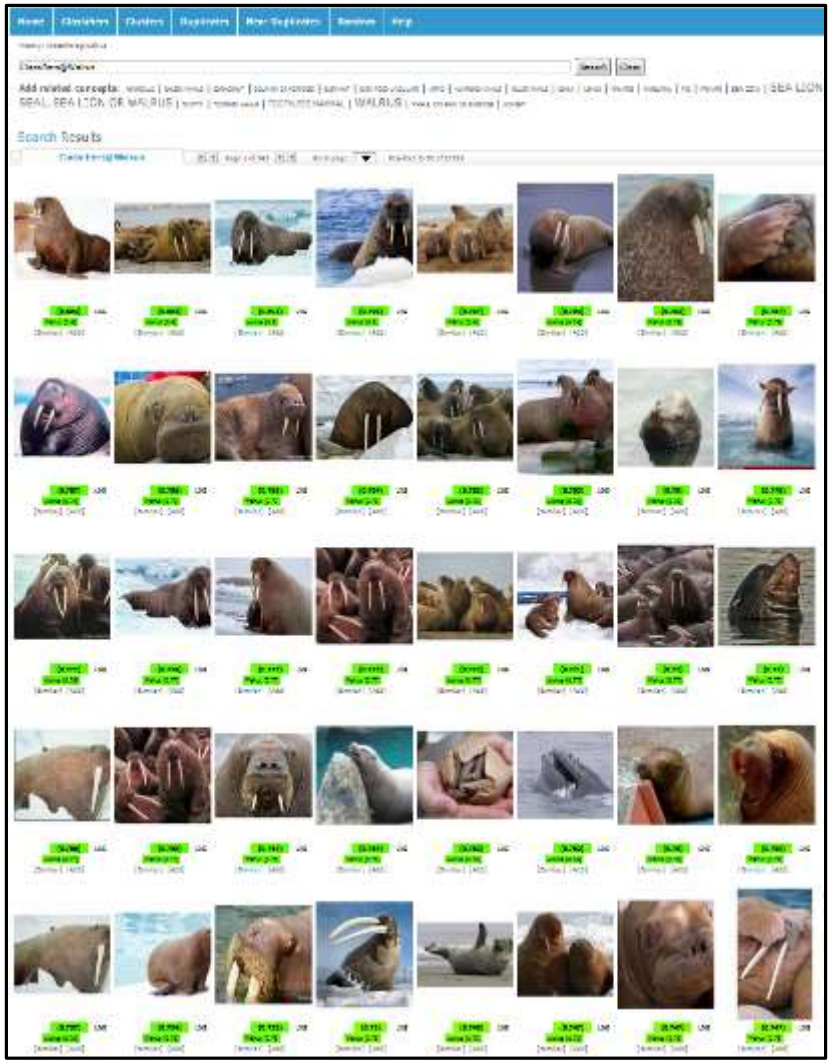


More Examples of Classification Results using Semantic Faceted Hierarchy

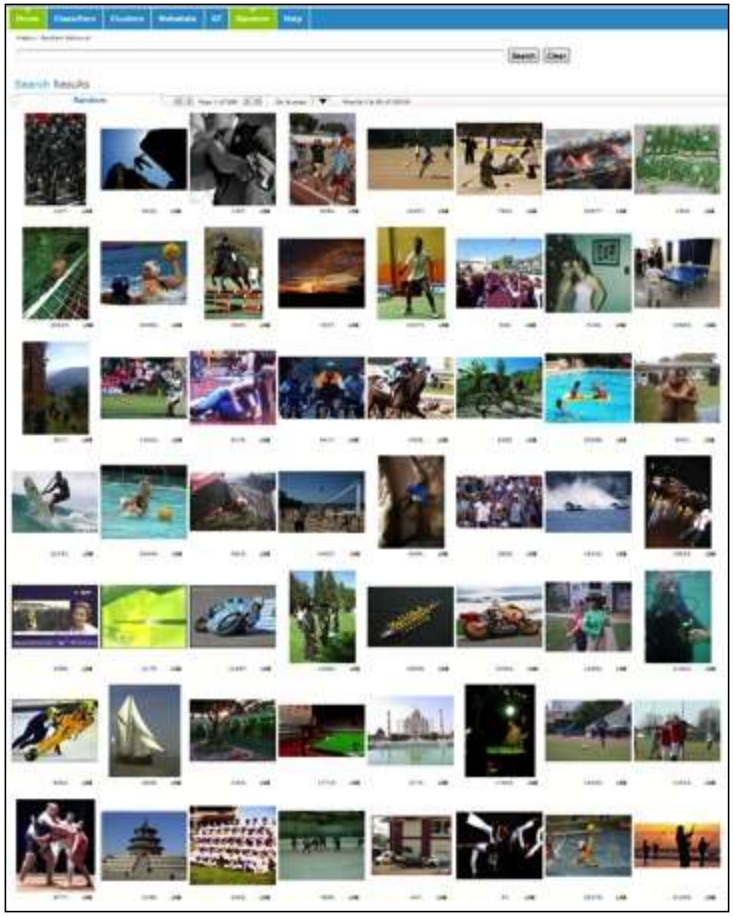
[ANIMAL] *Dog*



[ANIMAL] *Walrus*



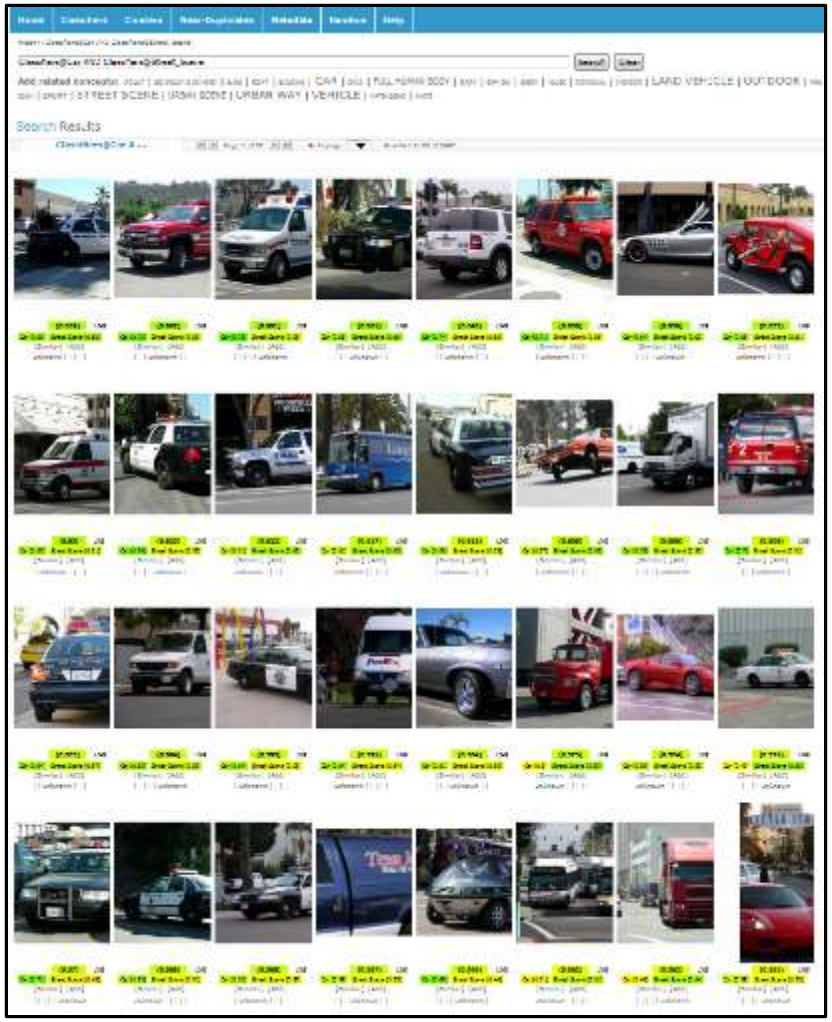
Machine Learning Shifts Effort to Organizing Image Data for Training



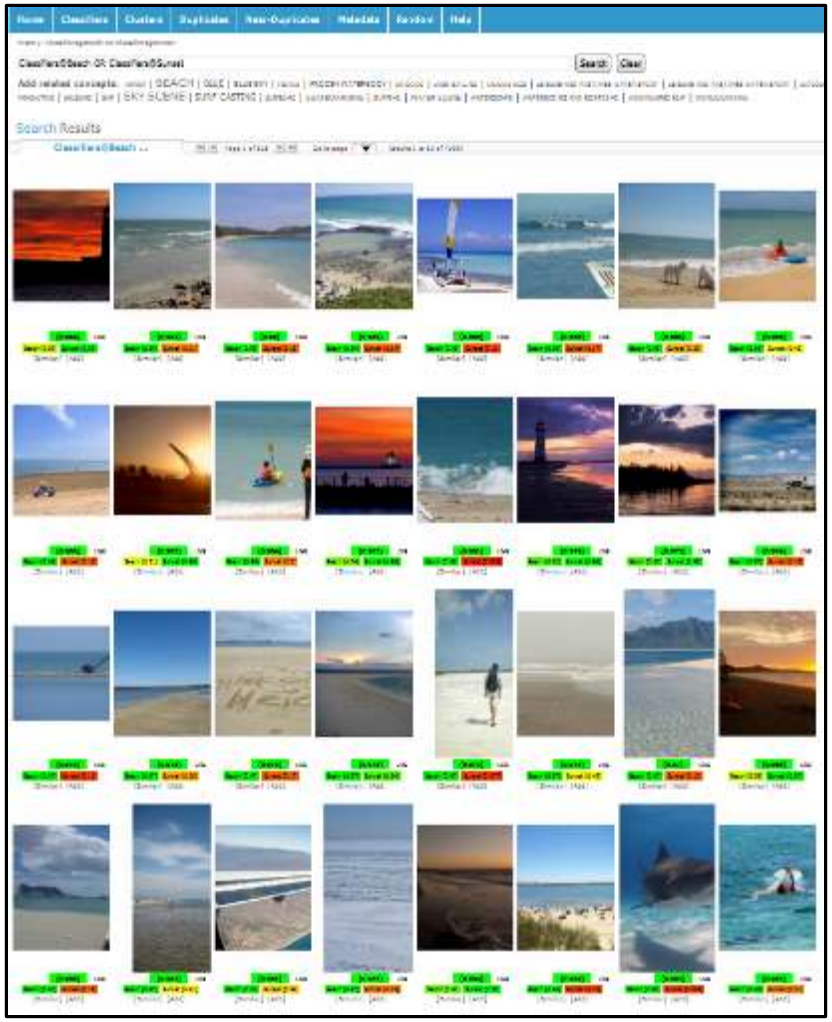
Accurate recognition of *sports* by training 150 categories

Semantic Searches is supported by Automatically Extracted Semantic Labels (Boolean = AND/OR)

Car* AND *Street Scene

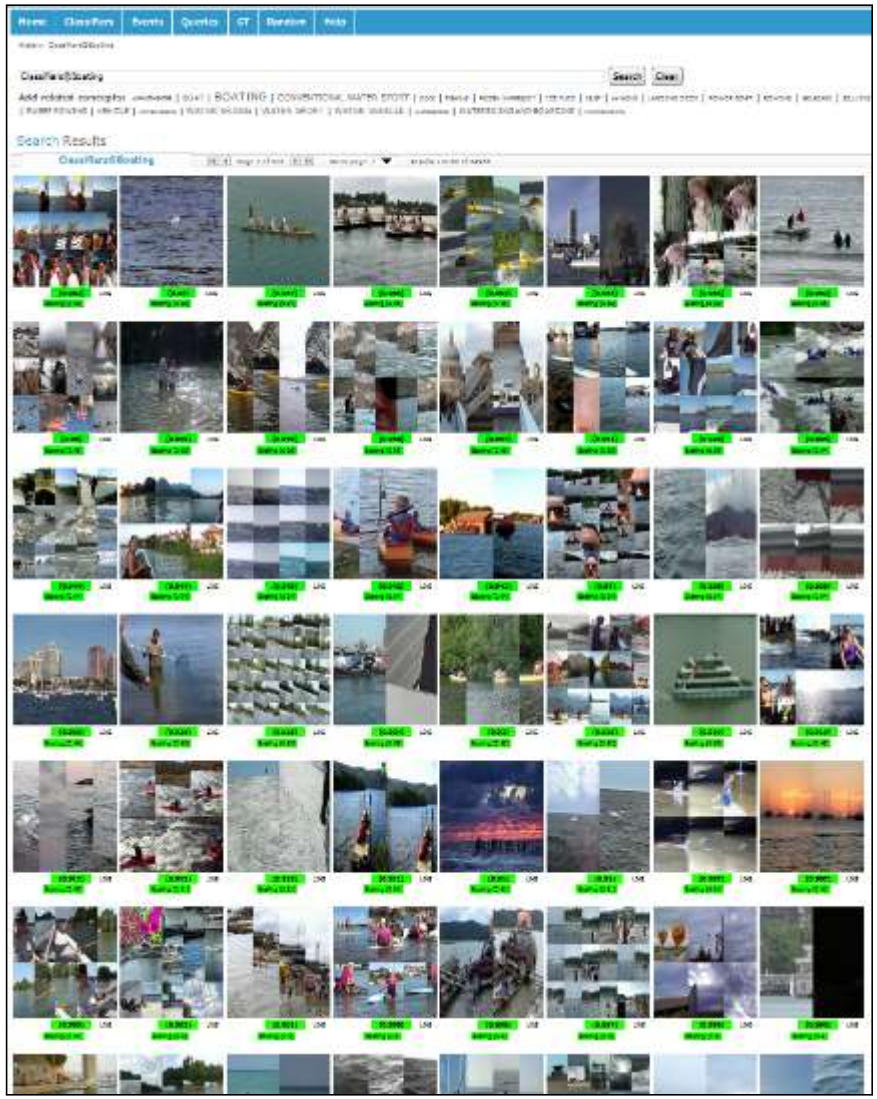


Beach* OR *Sunset



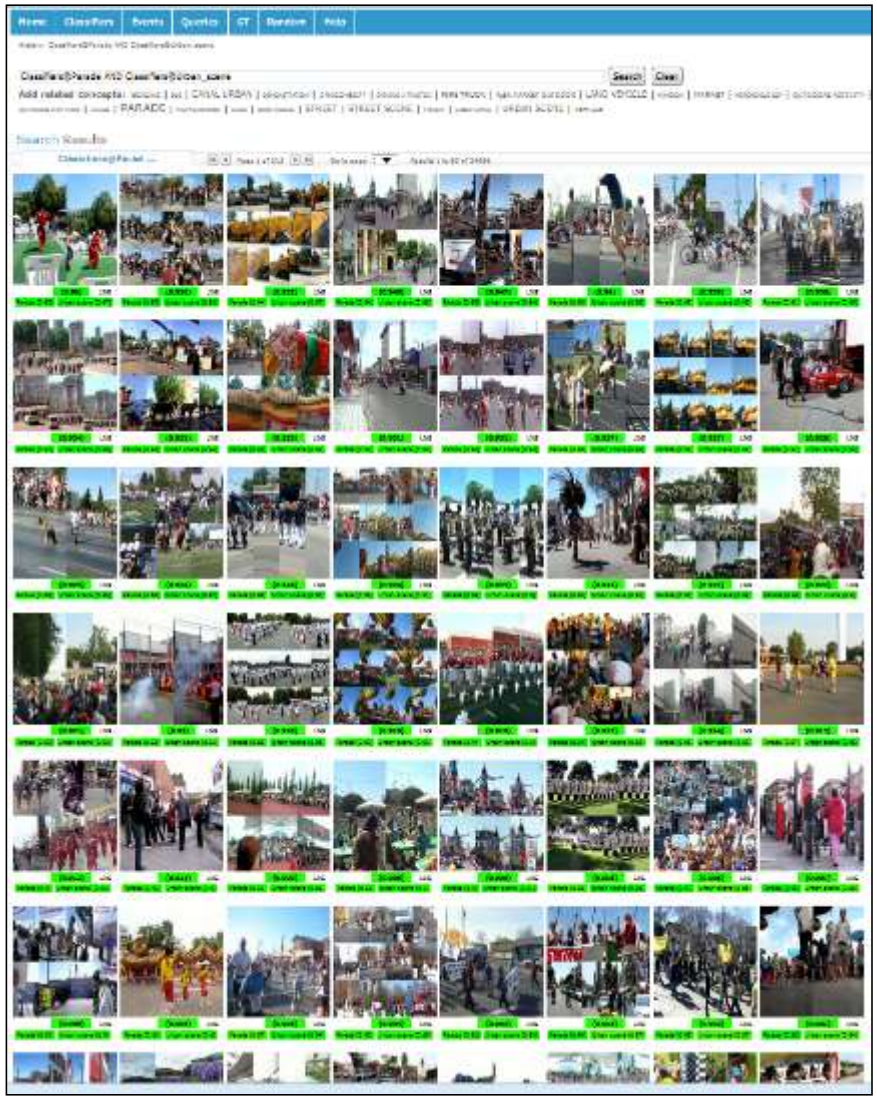
Semantic Indexing applies to Video Content Extraction and Search

Boating



Other activities: *Running*, *Skiing*, …

Parade* AND *Urban Scene



Other Combining Functions: *AND, OR, X, MIN*

On-line photos and videos have enormous potential as a rich source of information about consumers

¼ Trillion

photos hosted on Facebook from 1.3 Billion users

100 Hours

of video are uploaded to YouTube ***every minute***

51%

of the Class of 2015 (high school) use Instagram ***daily***

80%

of Pinterest users are ***female***

. . . and growing

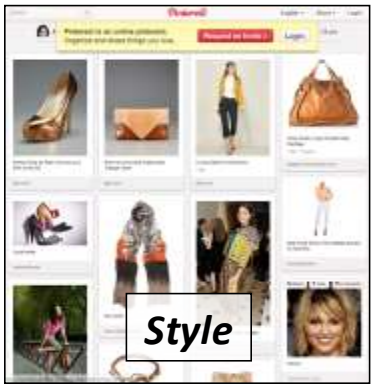
Growing Amount of Consumer Images and Video is Source for Insights

- Pins / Re-pins
- Likes / Dislikes
- Tweets
- Favorites

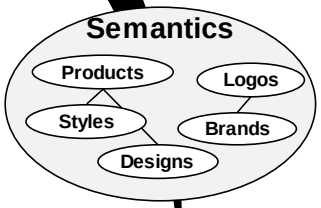
Like +1 Pinterest

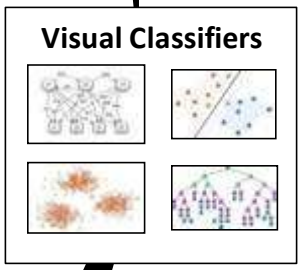
Consumer

Consumer Photos, Pins, Likes



- Access Public Visual Content
- Perform Semantic Classification





Visual Extraction

- *Preferred Styles*
- *Hobbies/interests*
- *Life Events*
- *Products*

Targeting, Marketing and Planning

macy's

- Personalization
- Promotions
- Campaigns
- Planning

Healthcare:

Radiology Image Analysis

- Specialized algorithms per medical *modality* and *disease*
- Limited scaling and coverage
- Bottleneck in algorithm development by computer vision experts

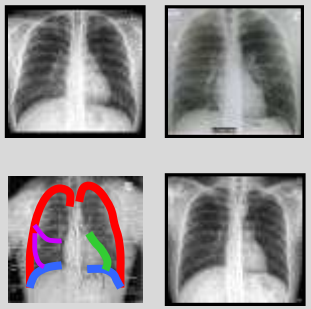
Medical Insights

- Diverse visual extraction across *modality, anatomy, pathology*
- Address wide spectrum across patient image data, images and figures in medical literature, visual knowledge repositories
- Beyond radiology to analyze images broadly in medicine

Multi-modal Analysis helps Build Medical Knowledge and Aid Diagnosis

Medical Semantics*

Multimedia Medical Data



Patient Records

Medical Image Taxonomy (RadLex, IRMA)

Anatomy/ Features (FMA, OBO)

Pathology/ Disease (UMLS, ICD, ICF)

Classification (modality, view, organ system, artifacts)

Anomaly Analysis (body/organ regions, local features)

Semantic Tagging (pathology, disease)

Cloudy (0.3)
Shadowed(0.3)
Darkened (0.1)

X-Ray (0.9)
Chest (0.8)
Lungs (0.8)

Categorical Examples

Annotated References

Multimodal Context Data

Massive Machine Learning*

Actionable Information

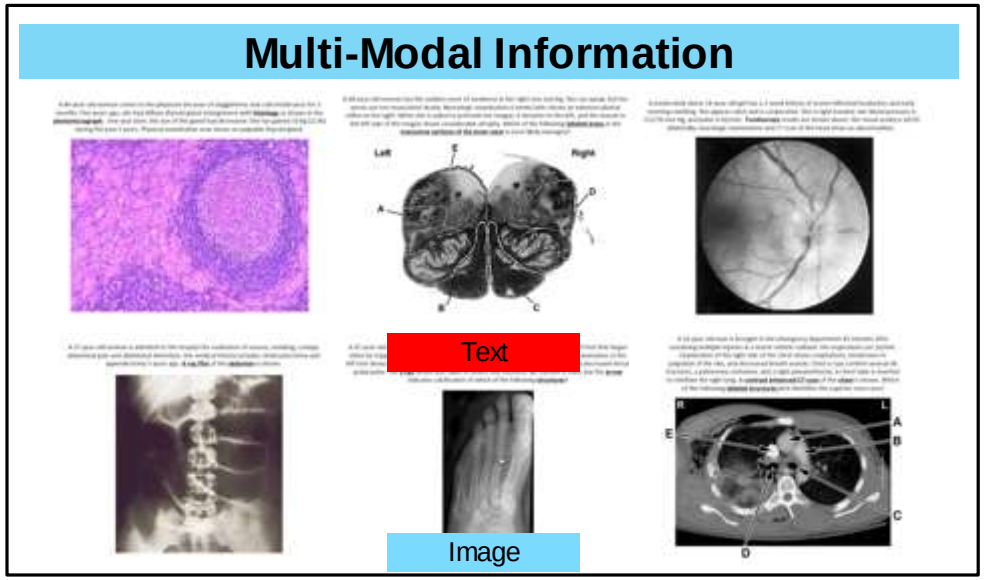
Pneumonia (0.3)
Tuberculosis (0.1)
Normal (0.4)

- Computer Aided Diagnosis
- Patient-Centric Medicine
- Automatic Triage and Retrieval
- Selected References
- Coincidental Diagnosis

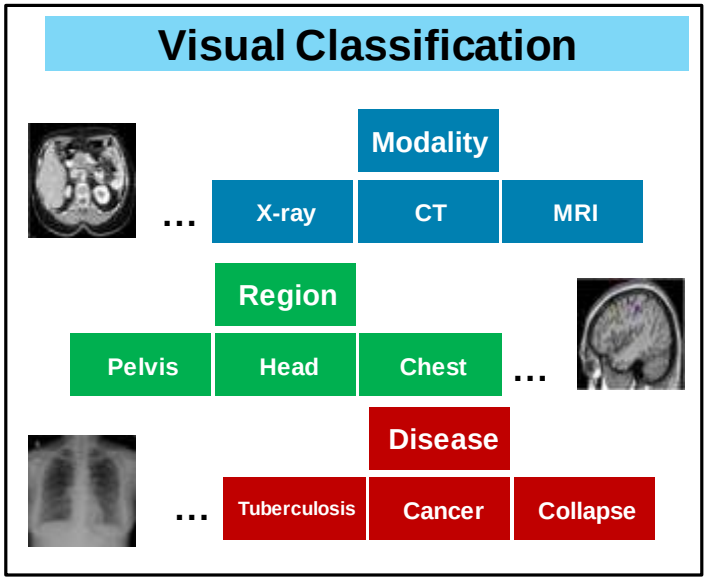
Medical Knowledge

Huge Diversity of Imagery in Medicine Spans Modality, Anatomy, Disease

Multi-Modal Information



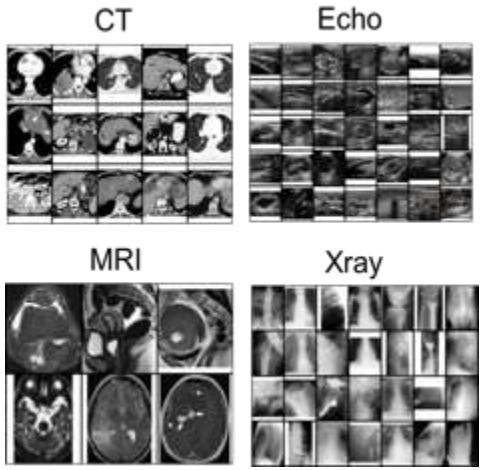
Visual Classification



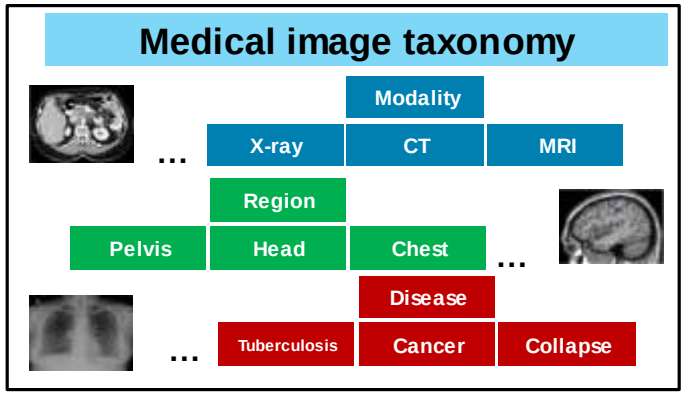
NIH PubMed Medical Modality Classification:

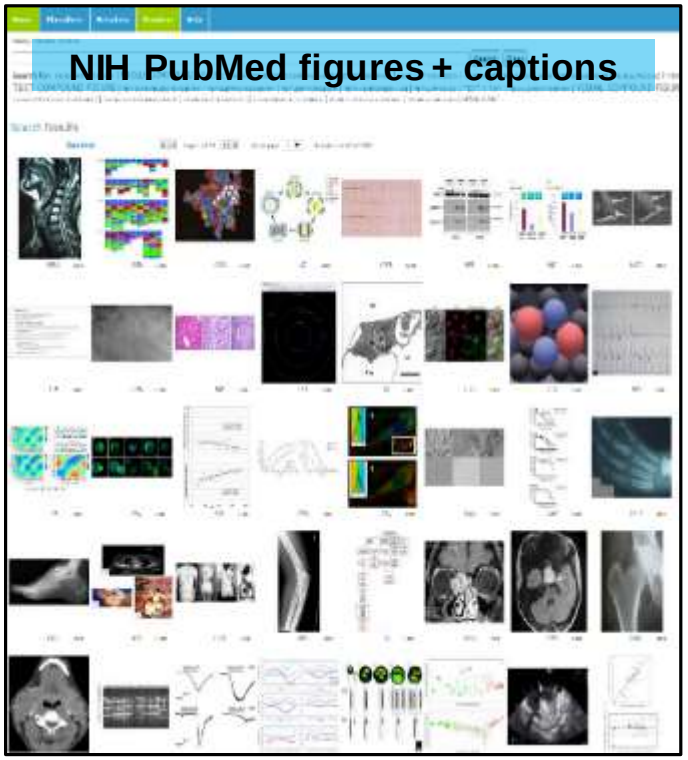
- **Extracting knowledge from NLM/NIH PubMed** 14.2 million articles (3.8 million *free* full-text articles)
- Rich **contextual information**: text + figures + captions
- Challenge: given an unknown PubMed image determine medical category automatically
- Automatically classifying millions of published medical images by ***modality, region, disease*** along with associated text builds multi-modal knowledge

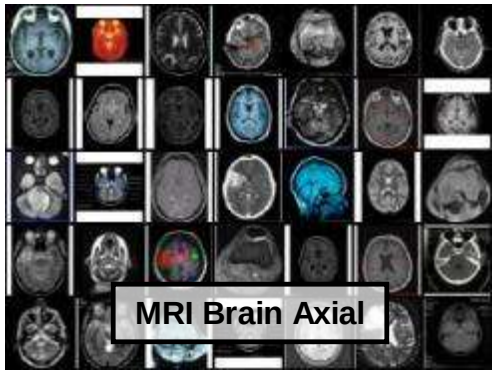
Visual Recognition

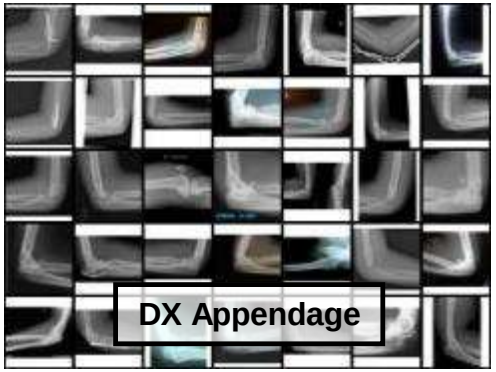


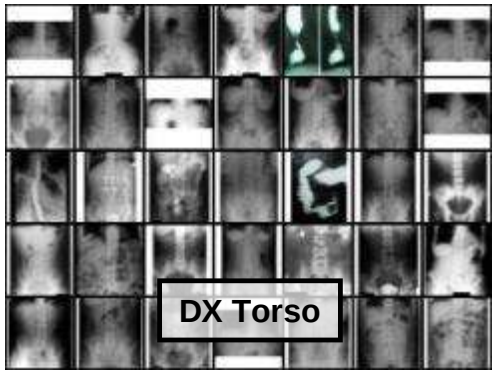
Medical Knowledge Management – Automatic Medical Image Classification

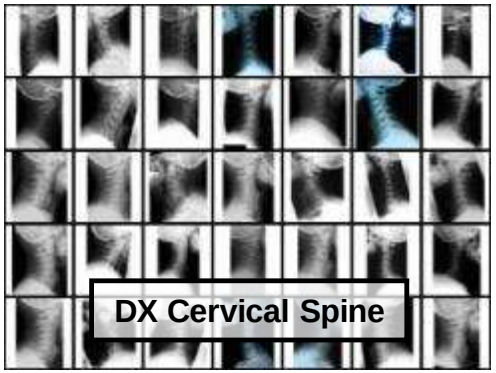








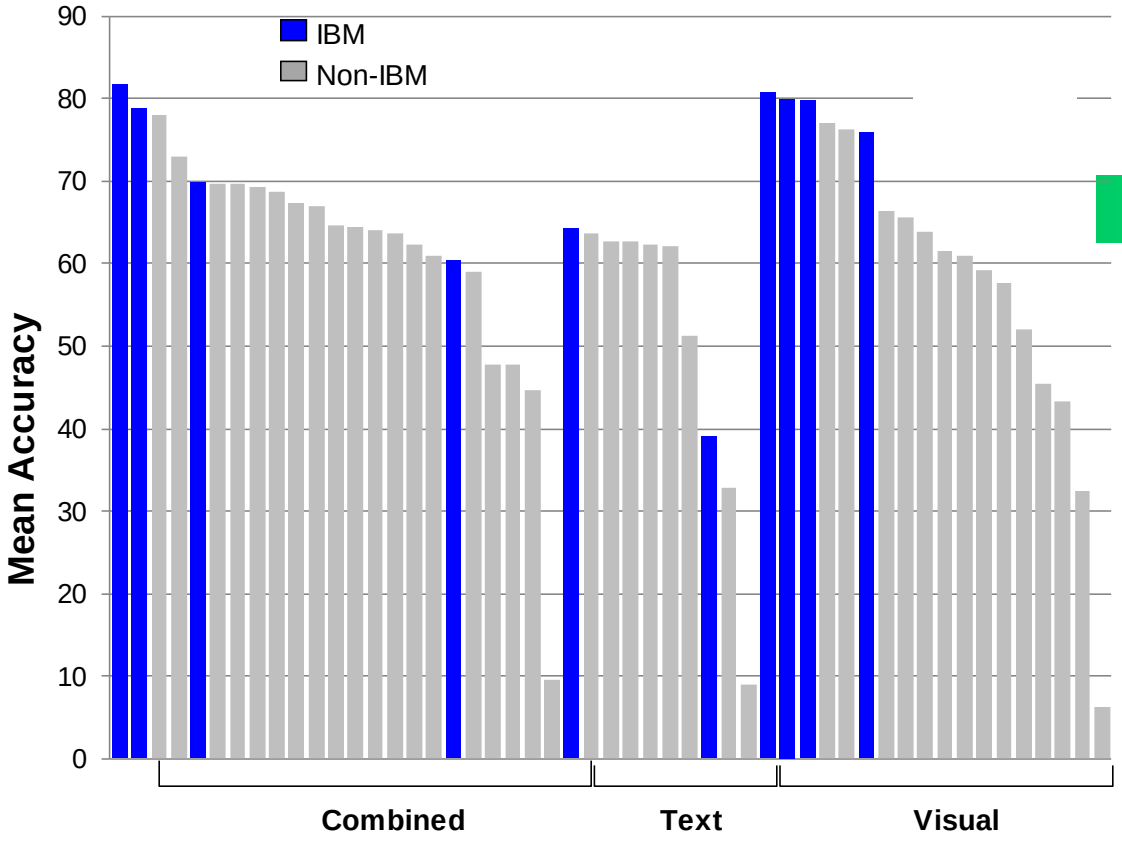


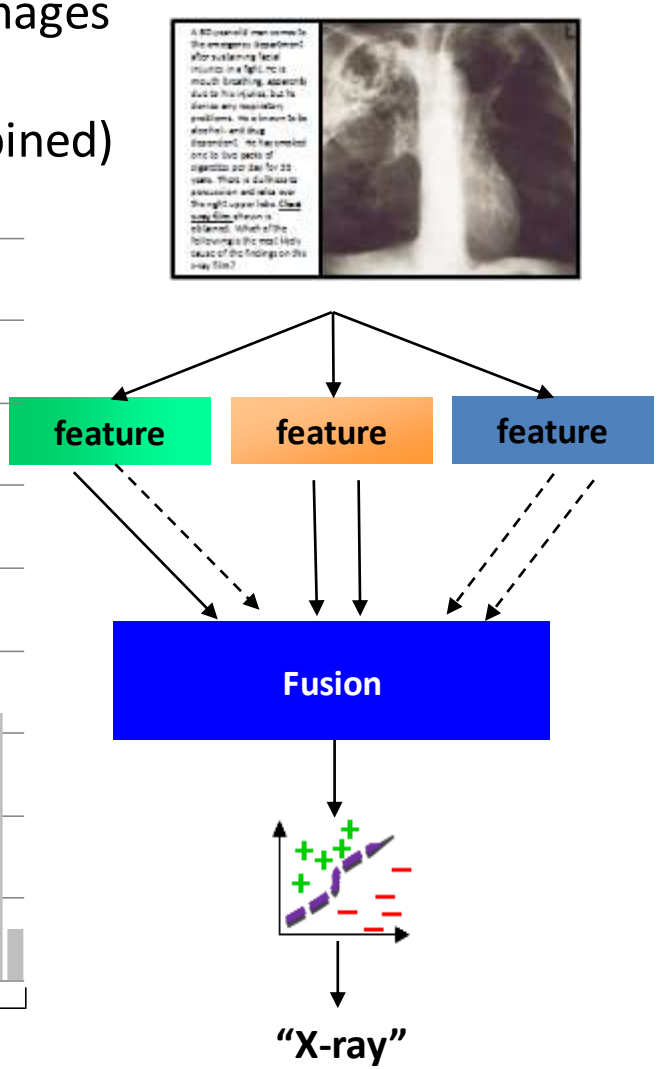


ImageCLEF Medical Image Classification*

Goal: automatically classify images into correct medical category

- Data-set of NIH PubMed articles with 305,000 images
- IBM achieved #1 performance in 2012 and 2013
- Top performance in each task (visual, text, combined)

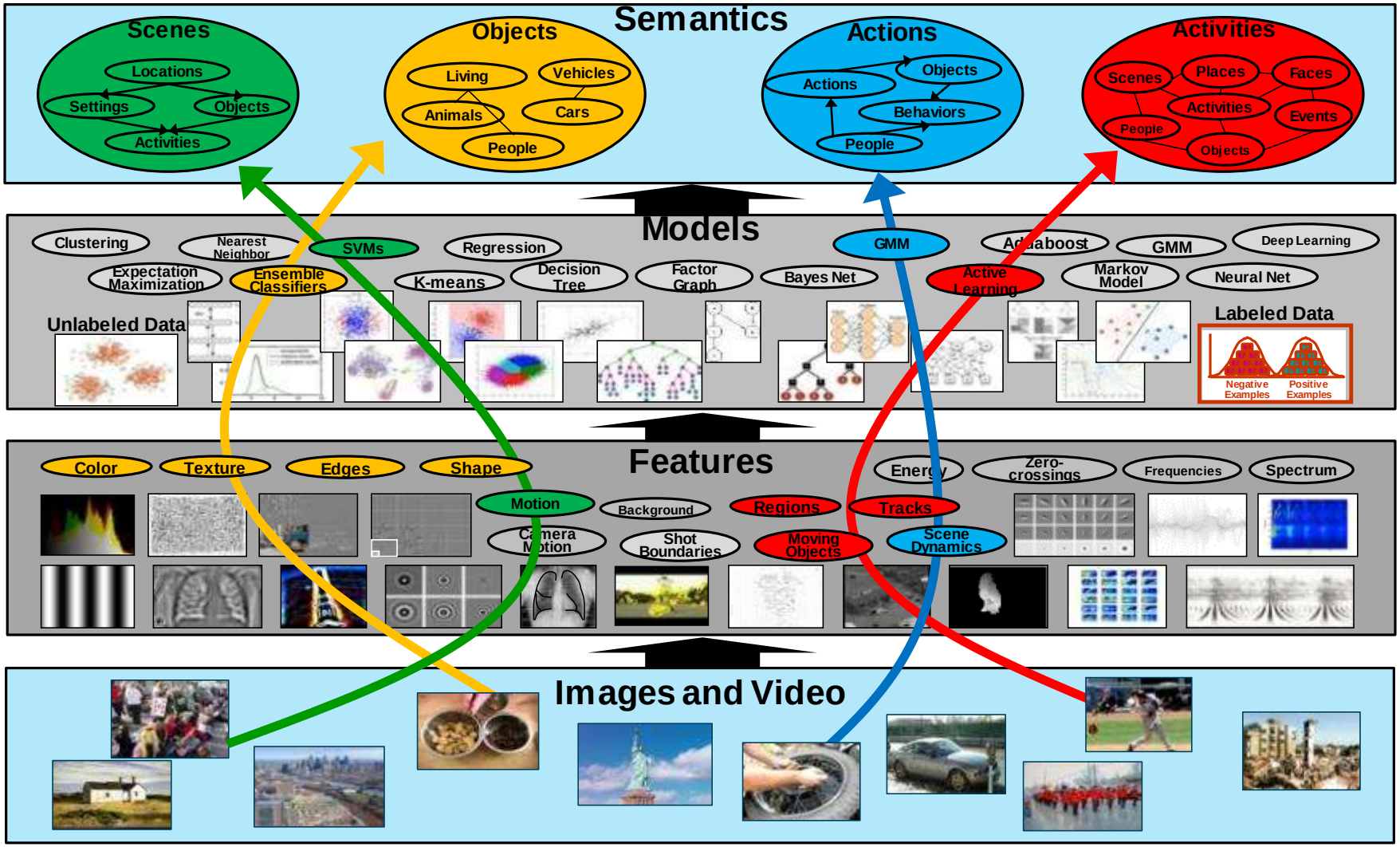




* http://imageclef.org/2013/medical

Visual Recognition
Technical Foundation

Multi-layer Learning Architecture for Image and Video Analysis



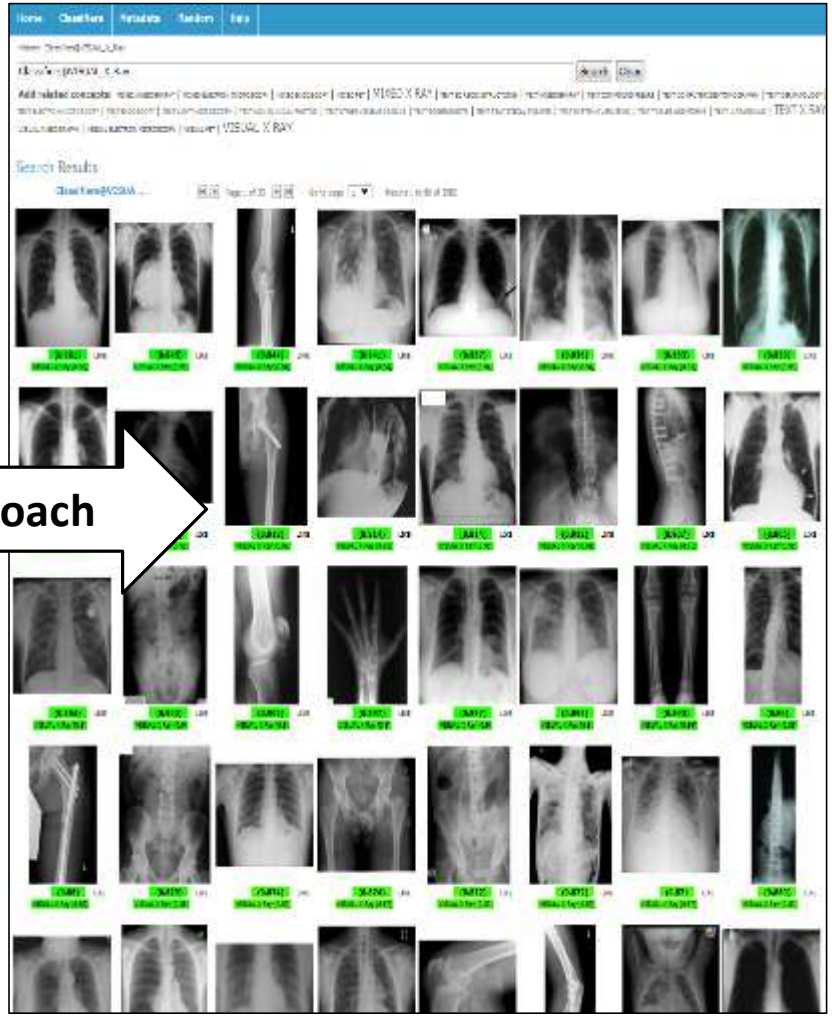
- **Visual Recognition** = predicting "semantic" labels for unknown images and video

Data-driven Machine Learning to Create Visual Semantic Classifiers

Natural Photos and Video

Medical Modalities & Viewpoints

Same Approach



Rich Set of Visual Features is Needed to Learn Semantic Discrimination

***Aardvark* [1]**
Spatial Scale Global (0.86)

***Even-Toed Ungulate* [14]**
Local Binary Pattern Global (0.68)

***Primate* [1]**
Local Binary Pattern Global (0.69)

***Seal, Sea Lion, Walrus* [3]**
Local Binary Pattern Global (0.83)

***Bat* [1]**
Thumbnail Vector Global (0.93)

***Meat-Eater* [12]**
Color Correlogram Global (0.68)

***Rabbit, Pika, Hare* [2]**
Color Moments Grid (0.79)

***Toothless Mammal* [2]**
Local Binary Pattern Global (0.85)

***Elephant* [1]**
Spatial Scale Global (0.81)

***Odd-Toed Ungulate* [4]**
Edge Histogram Layout (0.79)

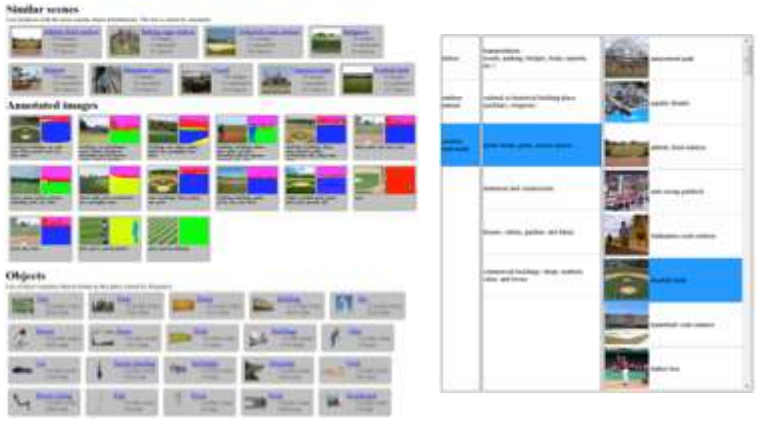
***Sea Cow* [1]**
Color Histogram Global (0.95)

***Whale, Dolphin, Porpoise* [5]**
Color Histogram Global (0.95)

Increasing Availability of Labeled Data is Accelerating Visual Recognition

Scenes and Objects:

(SUN Database: 131,072 images, 908 scenes, 3,819 objects, (http://groups.csail.mit.edu/vision/SUN/)



Objects and Actions:

(PASCAL VOC: 11,530 images with 27,450 ROI annotated objects and 6,929 segmentations (http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/)

20 objects

segmentations

10 actions

Scenes:

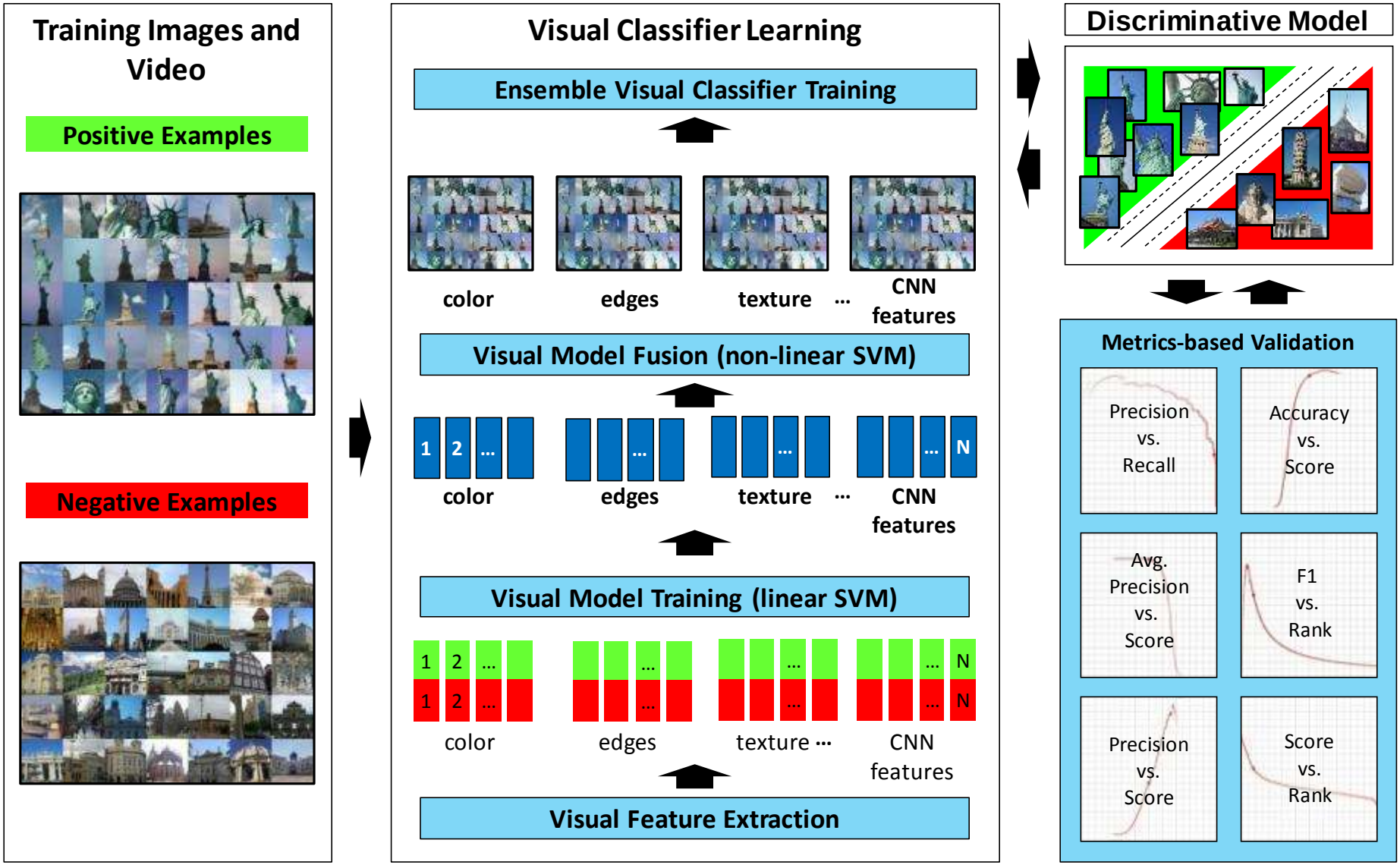
(INRIA Holidays Data Set): 1,491 images, 500 queries, 991 search images (http://lear.inrialpes.fr/~jegou/data.php)

Objects:

ImageNet linguistic visual objects (http://image-net.org/)

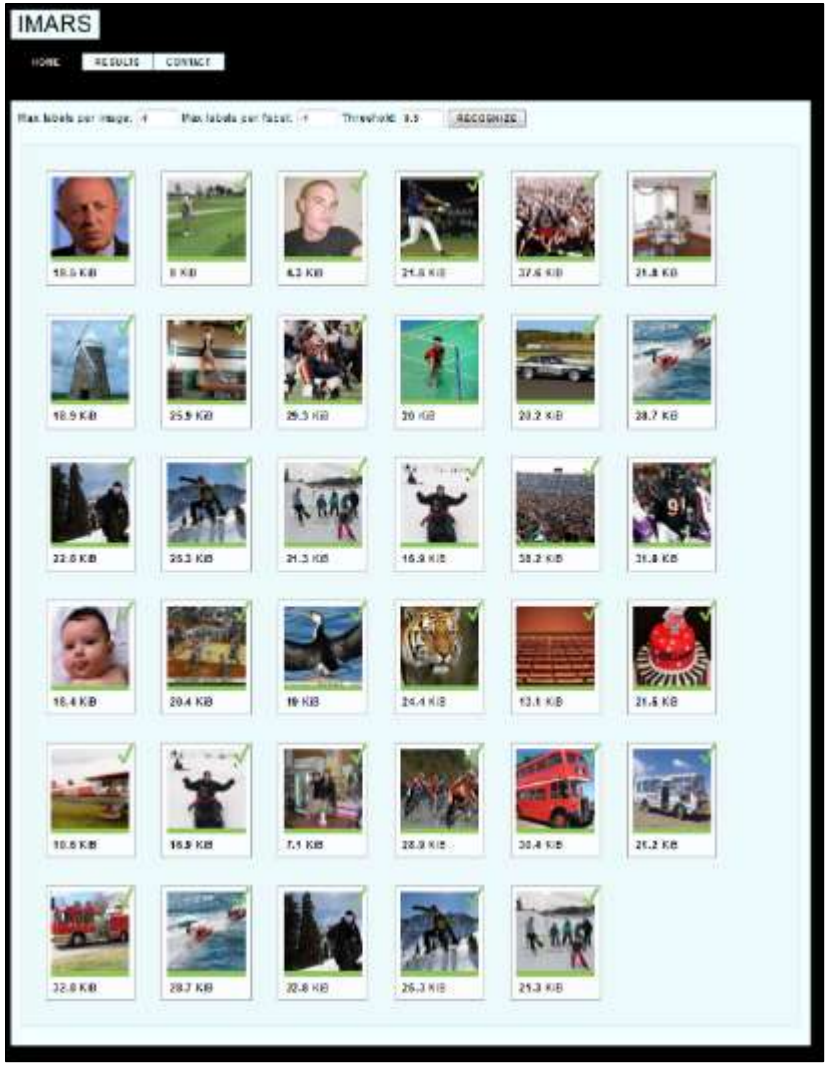
- **Full data**: 15,589 WordNet synsets
- **ILSVRC:** 1,000 object categories and 1.2 million images

Visual Classifier Learning Allows Metrics-based Optimization



Visual Recognition Example

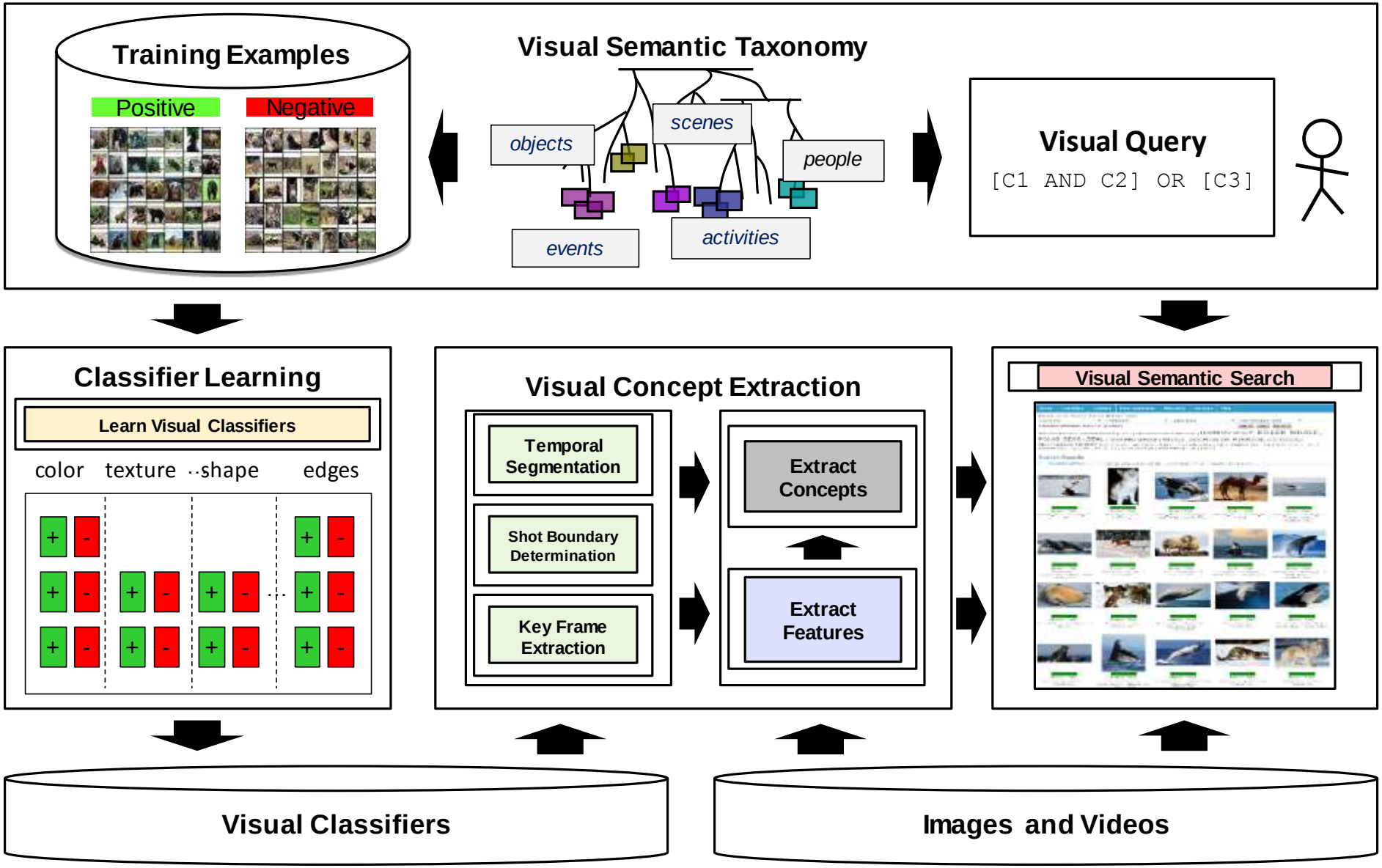
Upload Unknown Images



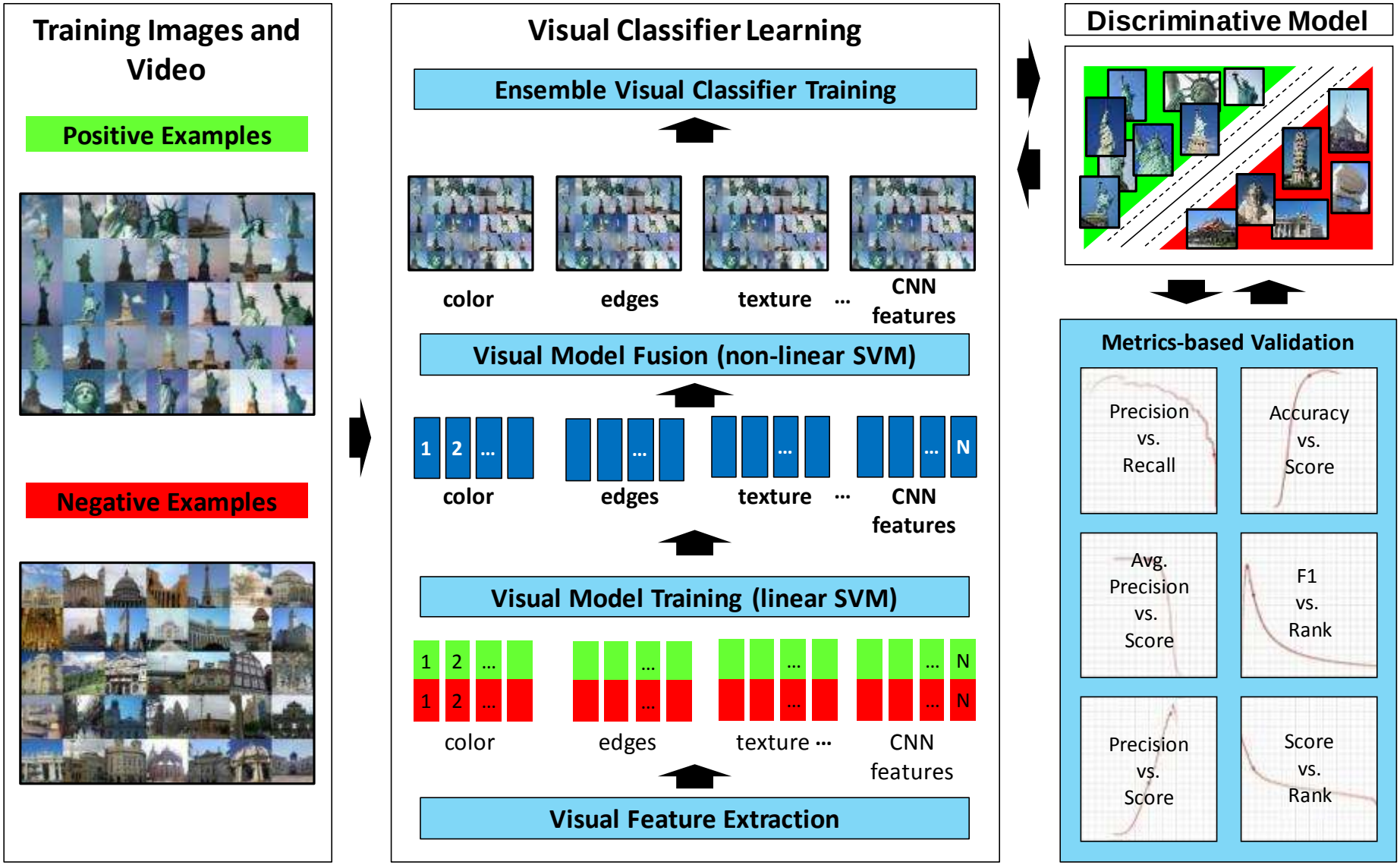
Visual Recognition Results

Demo

Visual Indexing and Search using Semantic Faceted Taxonomy



Visual Classifier Learning Allows Metrics-based Optimization



Different Approaches can be used for Managing and Selecting Negatives for Visual Discriminative Learning

Positive Examples		Negative Examples	
	*Whale*		*Person, Dog, Whale* **Common background** of all images (unlabeled positive images are included as negatives)
	*Whale*		*Zebra, Nature, Sunset* **Traditional taxonomy** mutually exclusivity (unlabeled positive images are included as negatives)
	*Whale*		*Zebra, Cat, Koala* **Faceted classification** ensures mutual exclusivity within facets (allows correct positives and negatives)

What label?

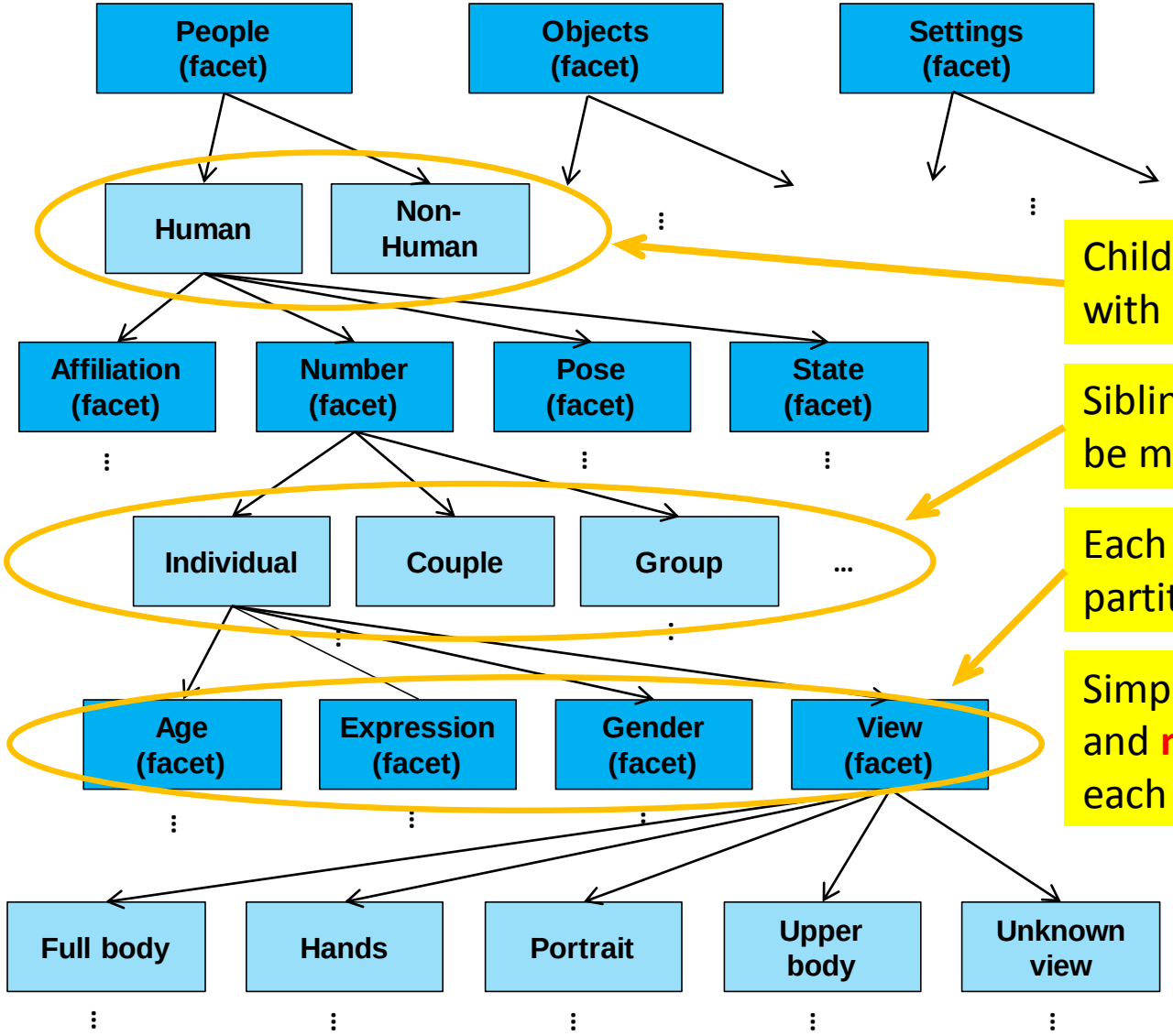
a) Man
b) Dog
c) Frisbee
d) Beach
e) Playing

FACETS:

[Person]
[Animal]
[Object]
[Setting]
[Activity]

Need to learn and assign labels from multiple facets!

Facets can be Nested in Hierarchical Faceted Classification Scheme



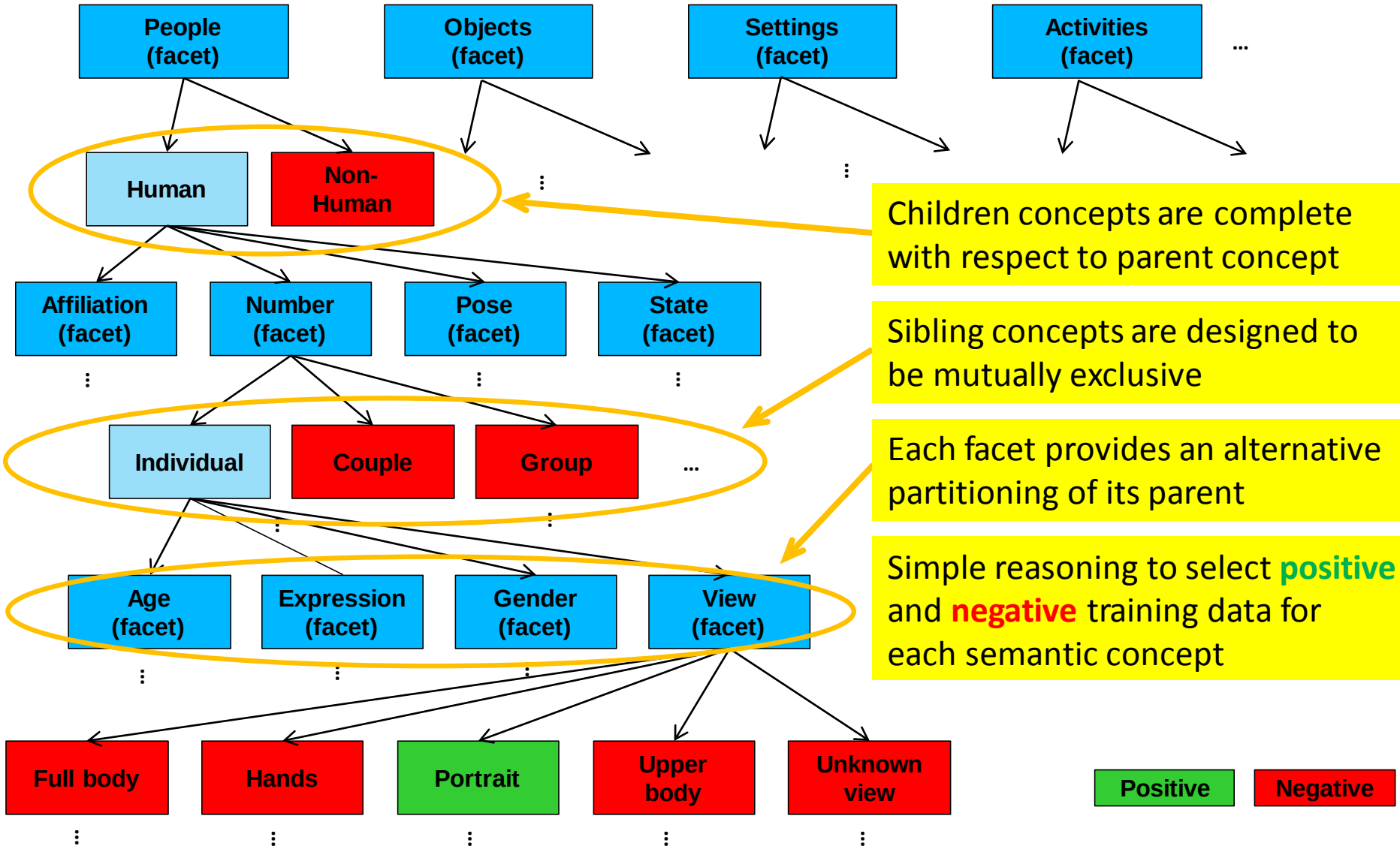
Children concepts are complete with respect to parent concept

Sibling concepts are designed to be mutually exclusive

Each facet provides an alternative partitioning of its parent

Simple reasoning to select **positive** and **negative** training data for each semantic concept

Facets can be Nested in Hierarchical Faceted Classification Scheme



Children concepts are complete with respect to parent concept

Sibling concepts are designed to be mutually exclusive

Each facet provides an alternative partitioning of its parent

Simple reasoning to select **positive** and **negative** training data for each semantic concept

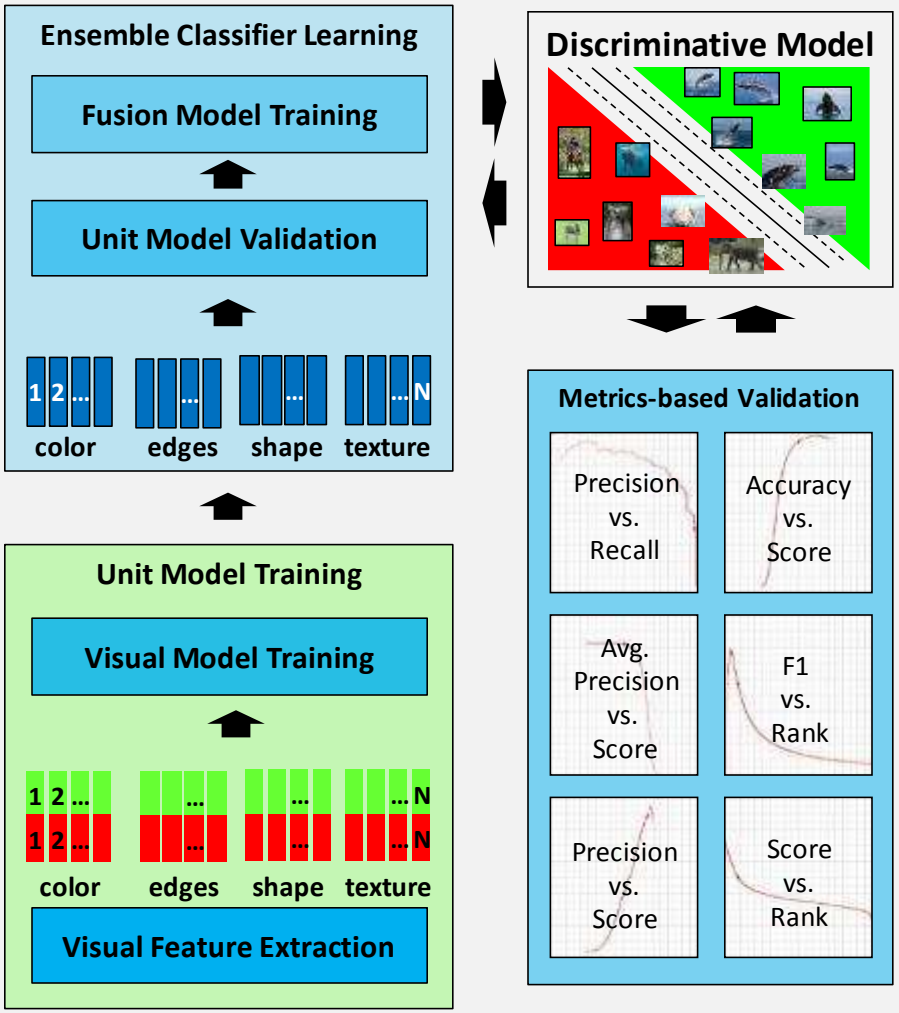
Key Challenge in Visual Discriminative Learning is Managing and Selecting of Negative Training Examples

Training Setup ?

Positive Examples

Negative Examples ???

Semantic Learning



Approaches for Labeling Training Images for Discriminative Learning

N Images

K Categories

	1	2	3	4	5	6	7	8	9	10	11	12
A	N	Y	N	N	N	N	Y	N	Y	N	N	N
B	N	N	Y	N	N	N	N	N	N	N	N	N
C	N	N	N	N	N	N	Y	N	N	Y	N	N
D	Y	N	N	N	N	N	N	N	N	N	N	Y
E	N	N	N	N	N	Y	N	Y	N	N	N	N
F	N	N	N	N	N	N	N	N	N	N	N	N
G	N	N	N	Y	Y	N	N	N	N	N	N	N
H	N	N	N	N	N	N	N	N	N	N	N	N
I	N	N	N	N	N	N	N	N	N	N	N	N
J	N	Y	N	N	N	N	N	N	N	N	Y	N
K	N	N	N	N	N	N	Y	N	N	N	N	N
L	N	N	N	N	N	N	N	N	Y	N	N	N
M	N	N	N	Y	N	N	N	N	N	N	N	N

Approaches for Labeling Training Images for Discriminative Learning

N Images

K Categories

	1	2	3	4	5	6	7	8	9	10	11	12
A	N	Y	N	N	N	N	Y	N	Y	N	N	N
B	N	N	Y	N	N	N	N	N	N	N	N	N
C	N	N	N	N	N	N	Y	N	N	Y	N	N
D	Y	N	N	N	N	N	N	N	N	N	N	Y
E	N	N	N	N	N	Y	N	Y	N	N	N	N
F	N	N	N	N	N	N	N	N	N	N	N	N
G	N	N	N	Y	Y	N	N	N	N	N	N	N
H	N	N	N	N	N	N	N	N	N	N	N	N
I	N	N	N	N	N	N	N	N	N	N	N	N
J	N	Y	N	N	N	N	N	N	N	N	Y	N
K	N	N	N	N	N	N	Y	N	N	N	N	N
L	N	N	N	N	N	N	N	N	Y	N	N	N
M	N	N	N	Y	N	N	N	N	N	N	N	N

- **Exhaustive**: label all *N* images by all *K* categories → *N* **K* is large (does not scale!)

- Assume a very "efficient" person can label 100K images per day per concept
- 1M image training set
- 10K concepts

→ 0.1 concept per day per person

→**Need 100K person-days**

Approaches for Labeling Training Images for Discriminative Learning

N Images

K Categories

	1	2	3	4	5	6	7	8	9	10	11	12
A	N	Y	N				Y	N	Y			N
B		N	Y	N						N		
C	N						Y	N		Y	N	N
D	Y	N			N							Y
E	N					Y	N	Y				
F		N	N									
G	N		N	Y	Y	N	N					
H					N	N						
I	N	N	N									N
J	N	Y	N						N	N	Y	N
K						N	Y	N			N	N
L	N							N	Y	N	N	N
M		N	N	Y					N			

- **Exhaustive**: label all *N* images by all *K* categories → *N* *K is large (does not scale!)
- **Binary**: label *p* positives (*p*<<*M*) and *n* negatives (*n*<<*N*) for each of *K* categories → (*p*+*n*)*K << *N**K (much better, still a lot)

Assume need 2K labeled images per concepts

"Efficient" person can label 50 concepts per day

→ **Need 200 person-days**

Approaches for Labeling Training Images for Discriminative Learning

N Images

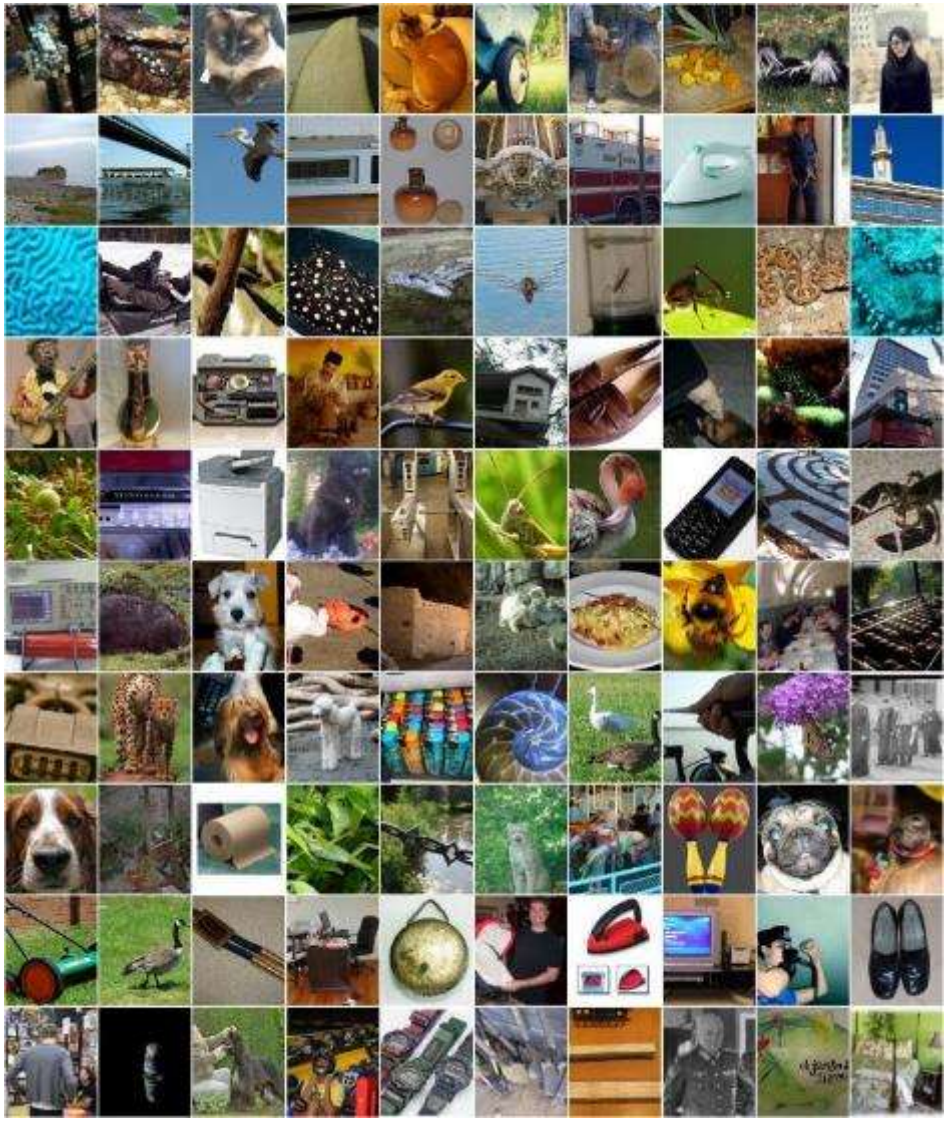
K Categories

	1	2	3	4	5	6	7	8	9	10	11	12
A		Y					Y		Y			
B			Y									
C							Y			Y		
D	Y											Y
E						Y		Y				
F												
G				Y	Y							
H												
I												
J		Y									Y	
K							Y					
L									Y			
M				Y								

- **Exhaustive**: label all *N* images by all *K* categories → *N* *K is large (does not scale!)
- **Binary**: label *p* positives (*p*<<*M*) and *n* negatives *(n<<N)* for each of *K* categories → *(p+n)*K* << *N*K* (much better, still a lot)
- **Multi-Class**: label *p* positives for each of *K* categories and infer negatives → *p*K*

→ **Need 100 person-days**

Problem: ImageNet ILSVRC Classification and Localization (CLS-LOC) Task



- 1,000 **object** categories that correspond to WordNet synsets
- No hierchical "overlap" between synsets
- For any pair of synsets *i* and *j*, *i* is not an ancestor of *j* in the WordNet hierarchy
- Every image has a **single image-level label** specifying the presence of **one object** category
- **No explicitly labeled negative** training images

Example Semantic Labeling Problems in ImageNet 2015

hay
→ **Not horse**

coffee mug
→ **Not lamp shade**
→ **Not table lamp**

ImageNet CLS-LOC Training Images for Discriminative Learning

N Images

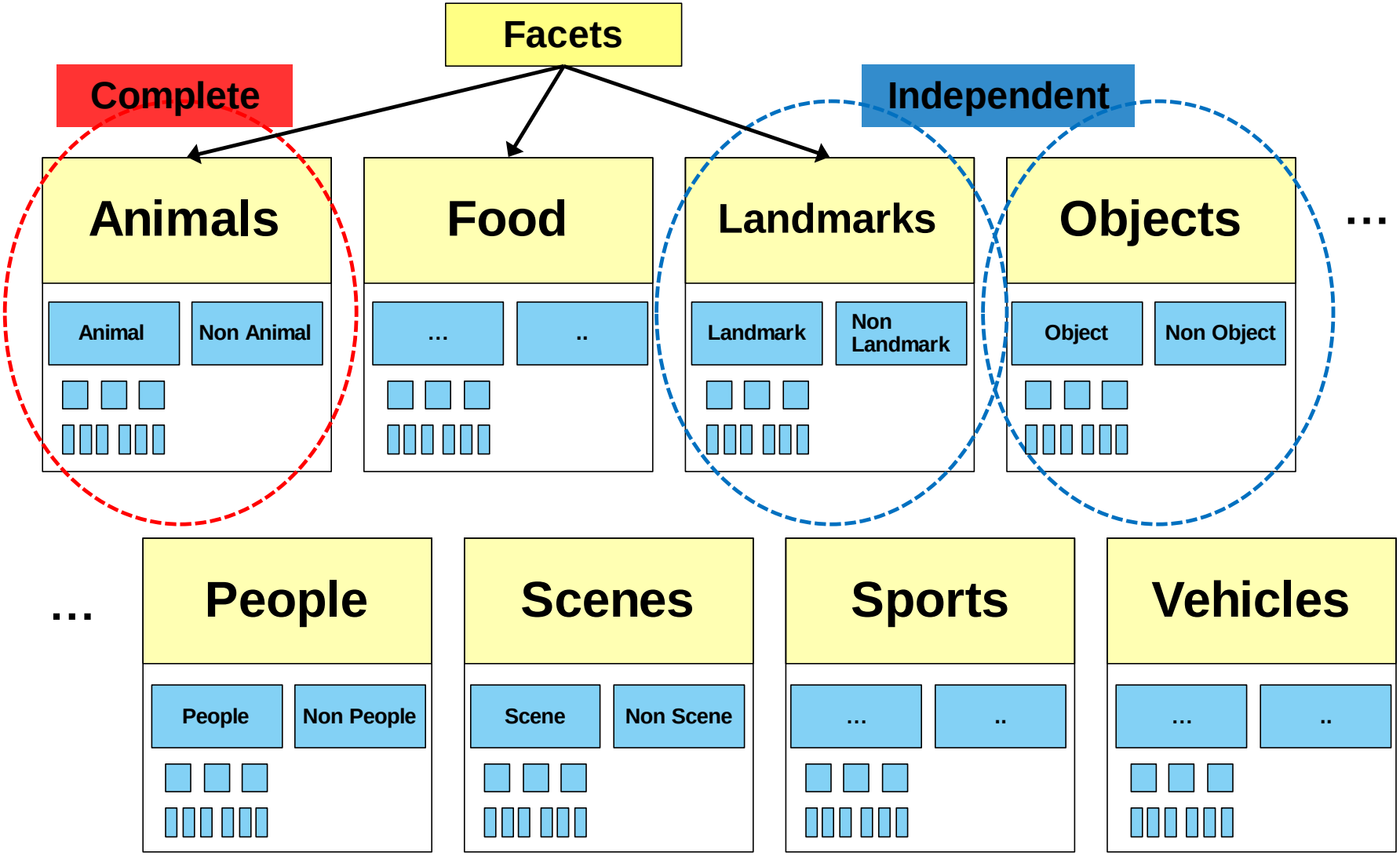
K Categories

	1	2	3	4	5	6	7	8	9	10	11	12
A	N	Y	N	N	N	N	Y	N	Y	N	N	N
B	N	N	Y	N	N	N	N	N	N	N	N	N
C	N	N	N	N	N	N	Y	N	N	Y	N	N
D	Y	N	N	N	N	N	N	N	N	N	N	Y
E	N	N	N	N	N	Y	N	Y	N	N	N	N
F	N	N	N	N	N	N	N	N	N	N	N	N
G	N	N	N	Y	Y	N	N	N	N	N	N	N
H	N	N	N	N	N	N	N	N	N	N	N	N
I	N	N	N	N	N	N	N	N	N	N	N	N
J	N	Y	N	N	N	N	N	N	N	N	Y	N
K	N	N	N	N	N	N	Y	N	N	N	N	N
L	N	N	N	N	N	N	N	N	Y	N	N	N
M	N	N	N	Y	N	N	N	N	N	N	N	N

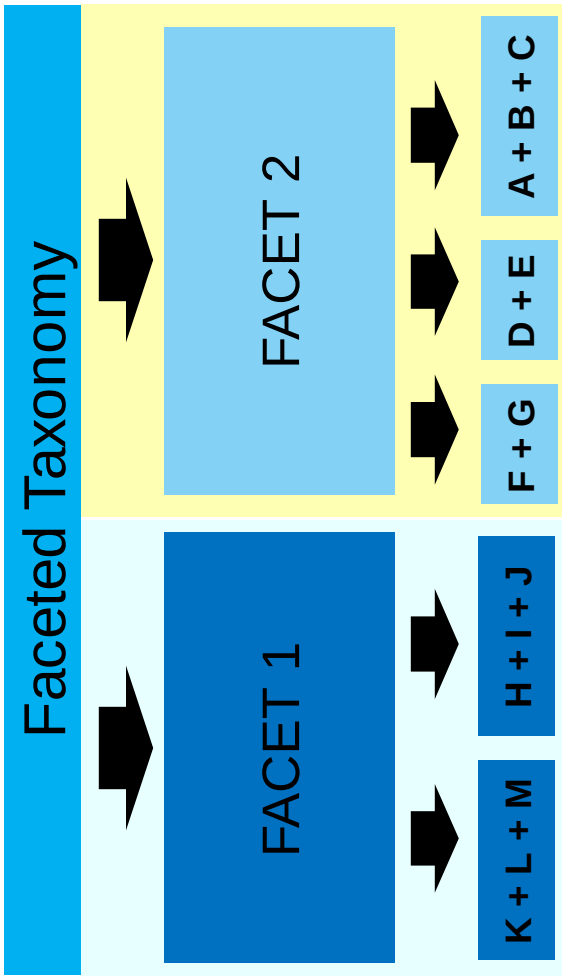
- **Exhaustive**: assign *K* labels to *N* images → *N* **K* is large (100K person-days for 10K labels and 1M images)
- **Binary**: label *p* positives (*p*<<*M*) and *n* negatives (*n*<<*N*) for each of *K* categories → (*p*+*n*)**K* << *N***K* (how to get negatives?)
- **Multi-Class**: label *p* positives for each of *K* categories and infer negatives → *p***K*

→Need 100 person-days

Better Managing Visual Semantics using a Faceted Taxonomy



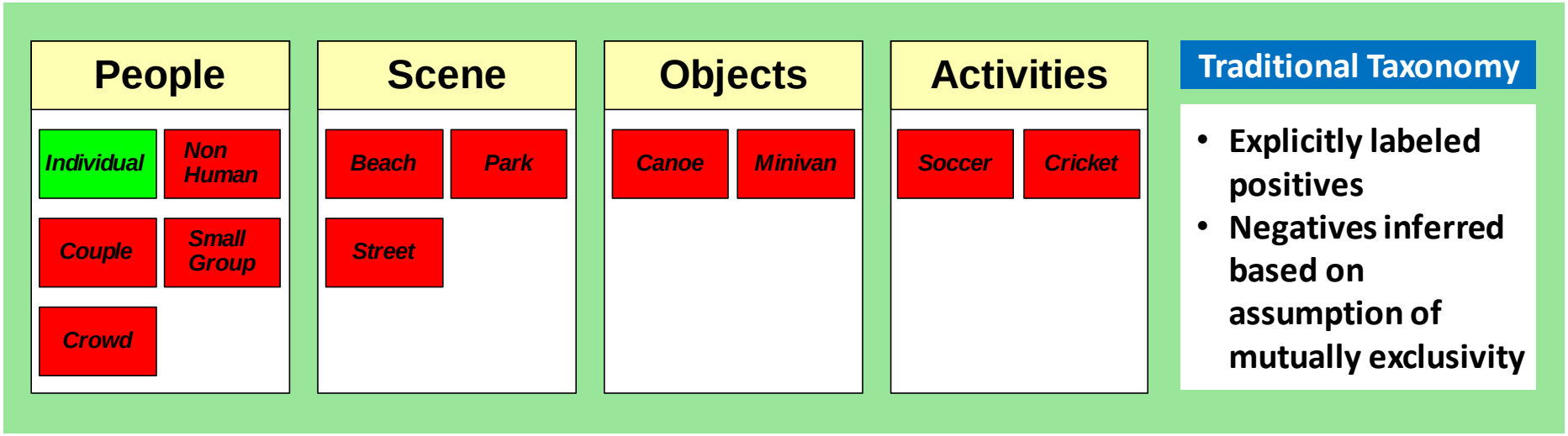
Faceted Taxonomy Authoring and Maintenance ($p*K_i$ labels per FACET i):

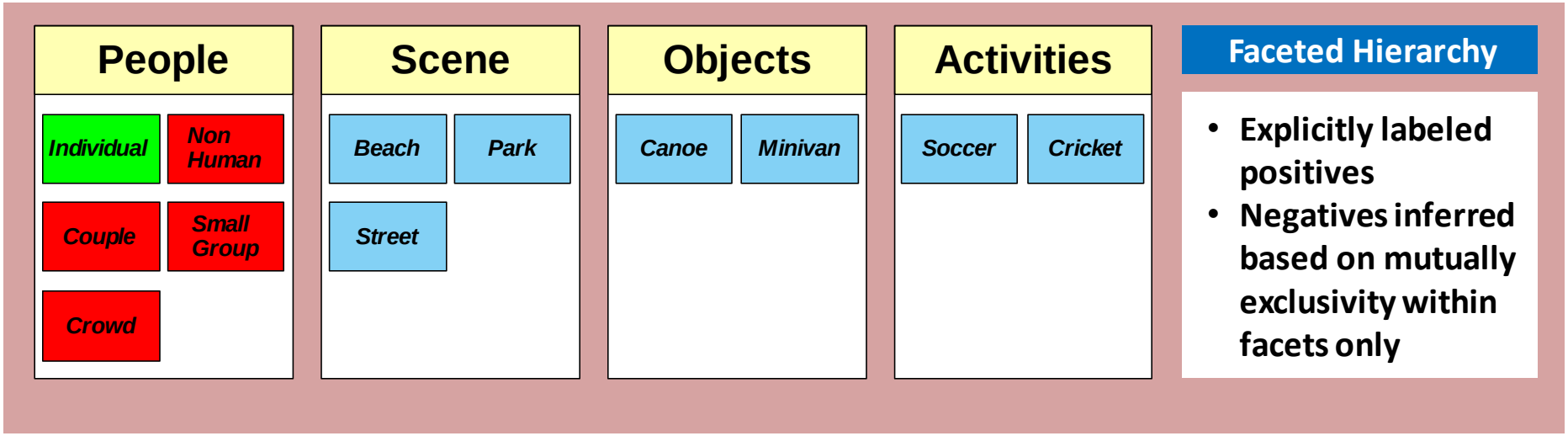


N Images

	1	2	3	4	5	6	7	8	9	10	11	12
A		Y							Y			
B			Y									
C										Y		
D	Y											Y
E						Y		Y				
F												
G				Y	Y							
H												
I												
J		Y									Y	
K							Y					
L									Y			
M				Y								

Experimental Setup: Faceted Hierarchy vs. Traditional Taxonomy





Results: Semantic Concepts are not Mutually Exclusive across Facets

[SCENE] *Beach*

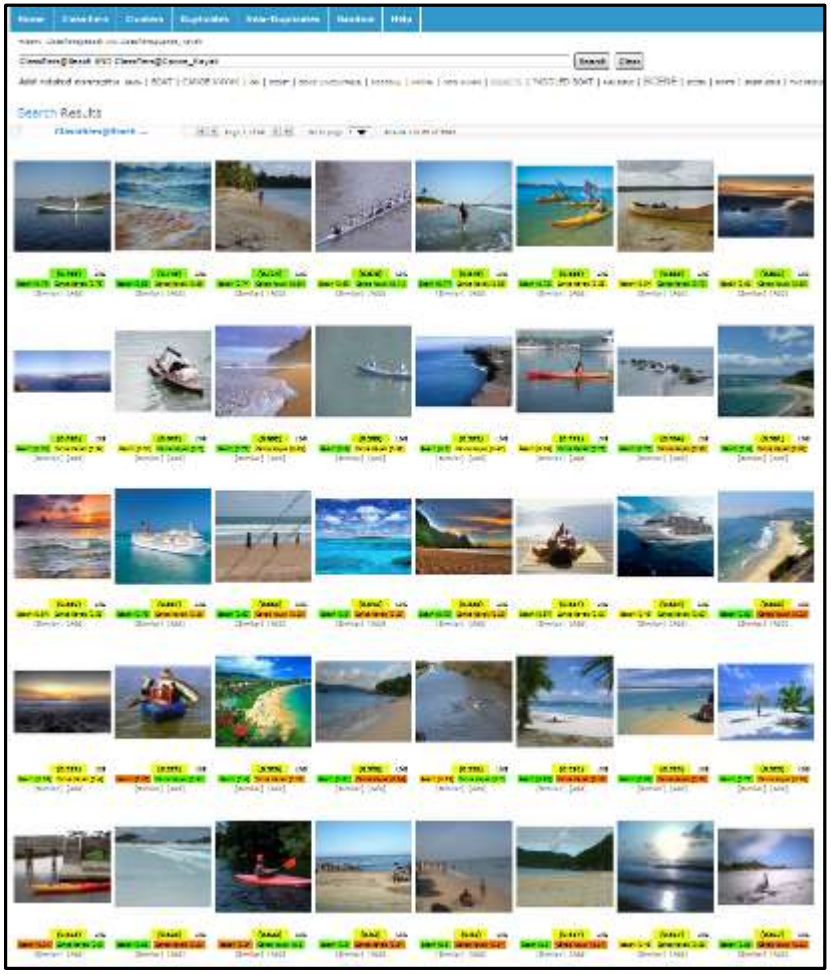
[OBJECT] *Canoe*

[SCENE] *Park*

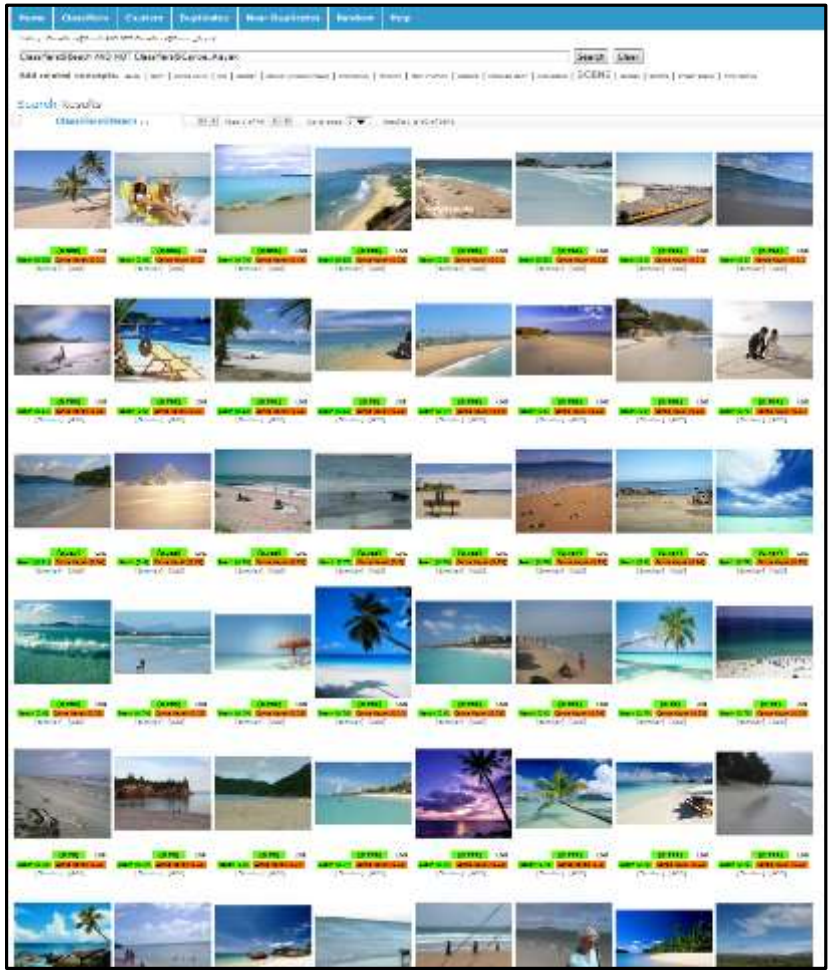
[PEOPLE] *Non Human*

Example Search Results using Visual Semantic Faceted Hierarchy

Beach* AND *Canoe



Beach* AND NOT *Canoe

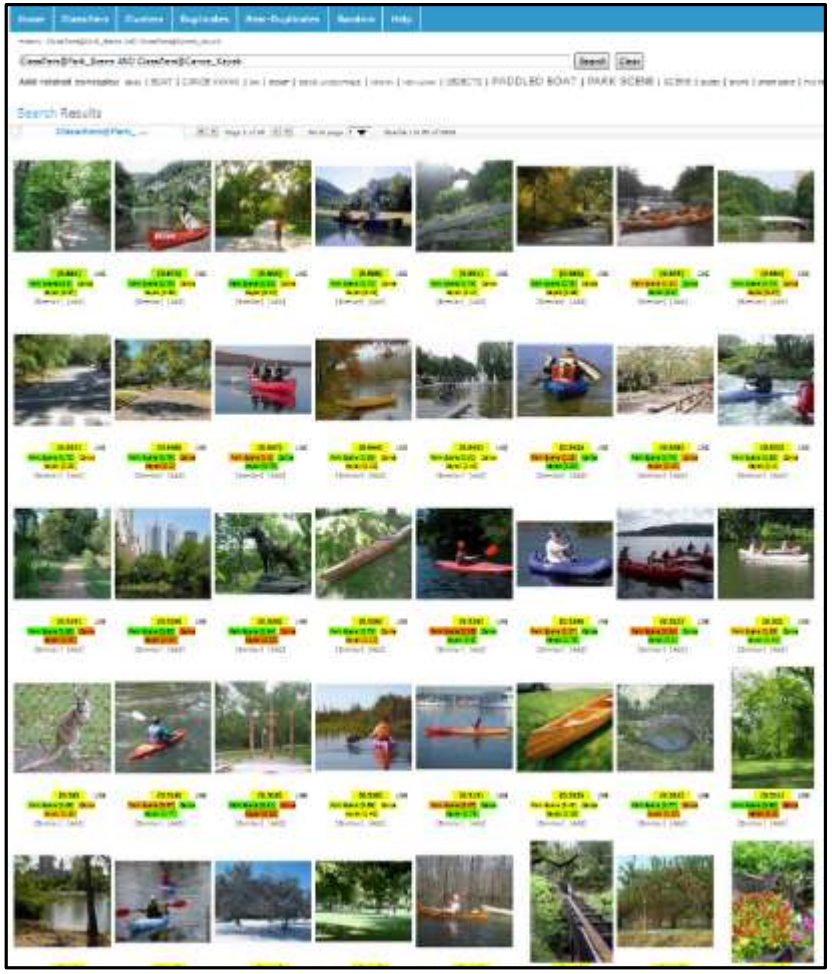


Beach and ***Canoe*** can be combined across facets → not mutually exclusive

Beach and **NOT *Canoe*** can also be combined across facets

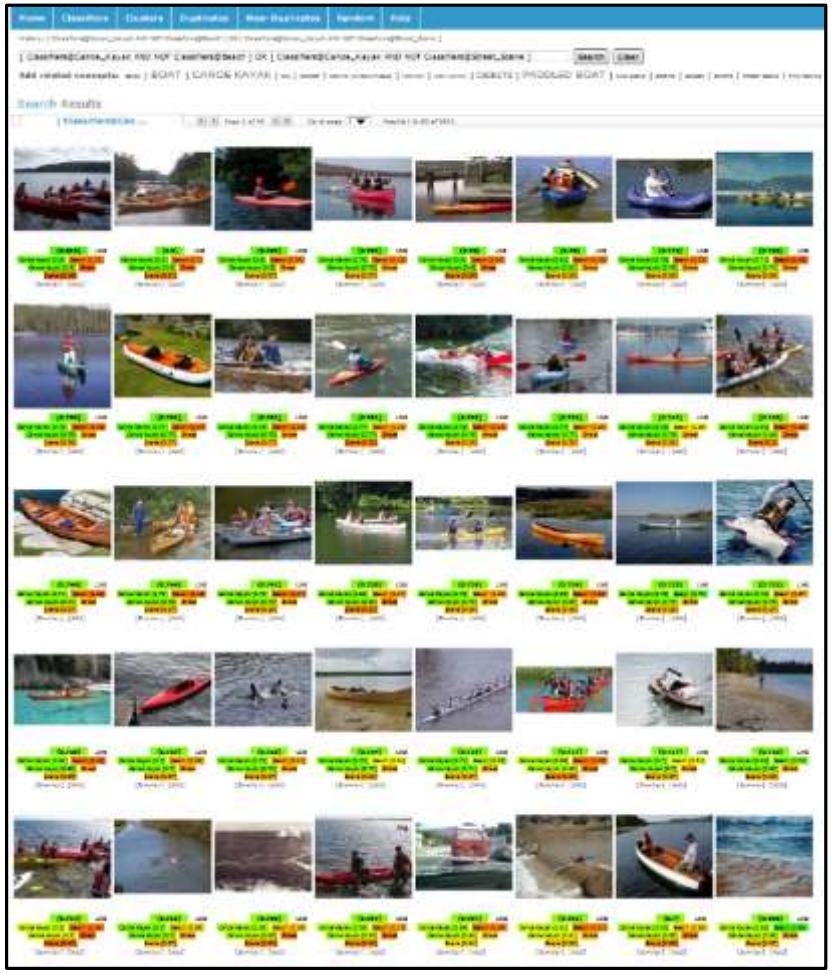
Example Search Results using Visual Semantic Faceted Hierarchy

Park* AND *Canoe



***Park* AND *Canoe* combines discriminative semantic concepts across facets**

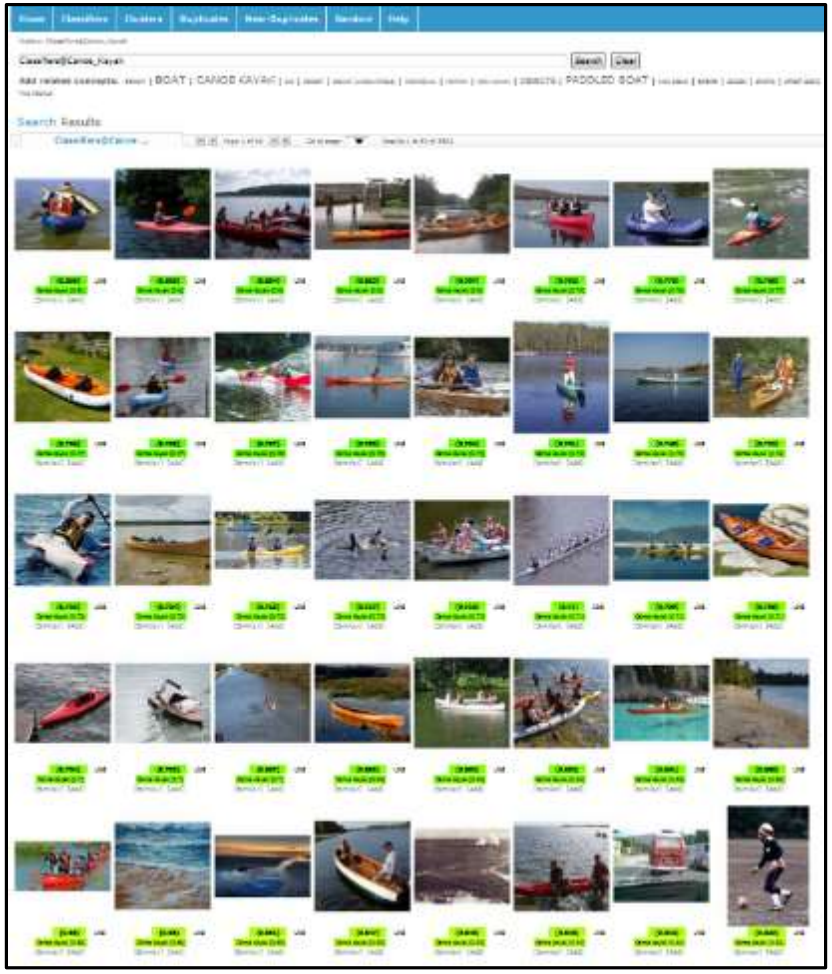
[*Canoe* AND NOT *Beach*]
OR [*Canoe* AND NOT *Street*]



[*Canoe* AND NOT *Beach*] OR [*Canoe* AND NOT *Street*] → approx. *Park* AND *Canoe*

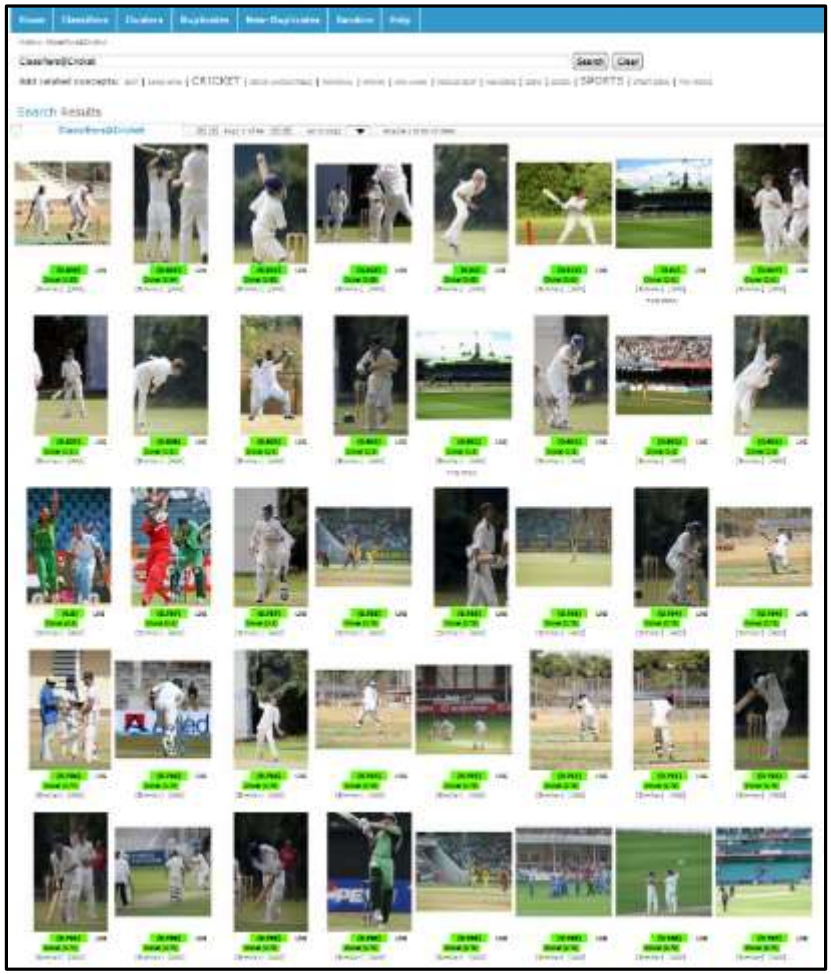
Example Classification Results using Visual Semantic Faceted Hierarchy

[OBJECT] Canoe



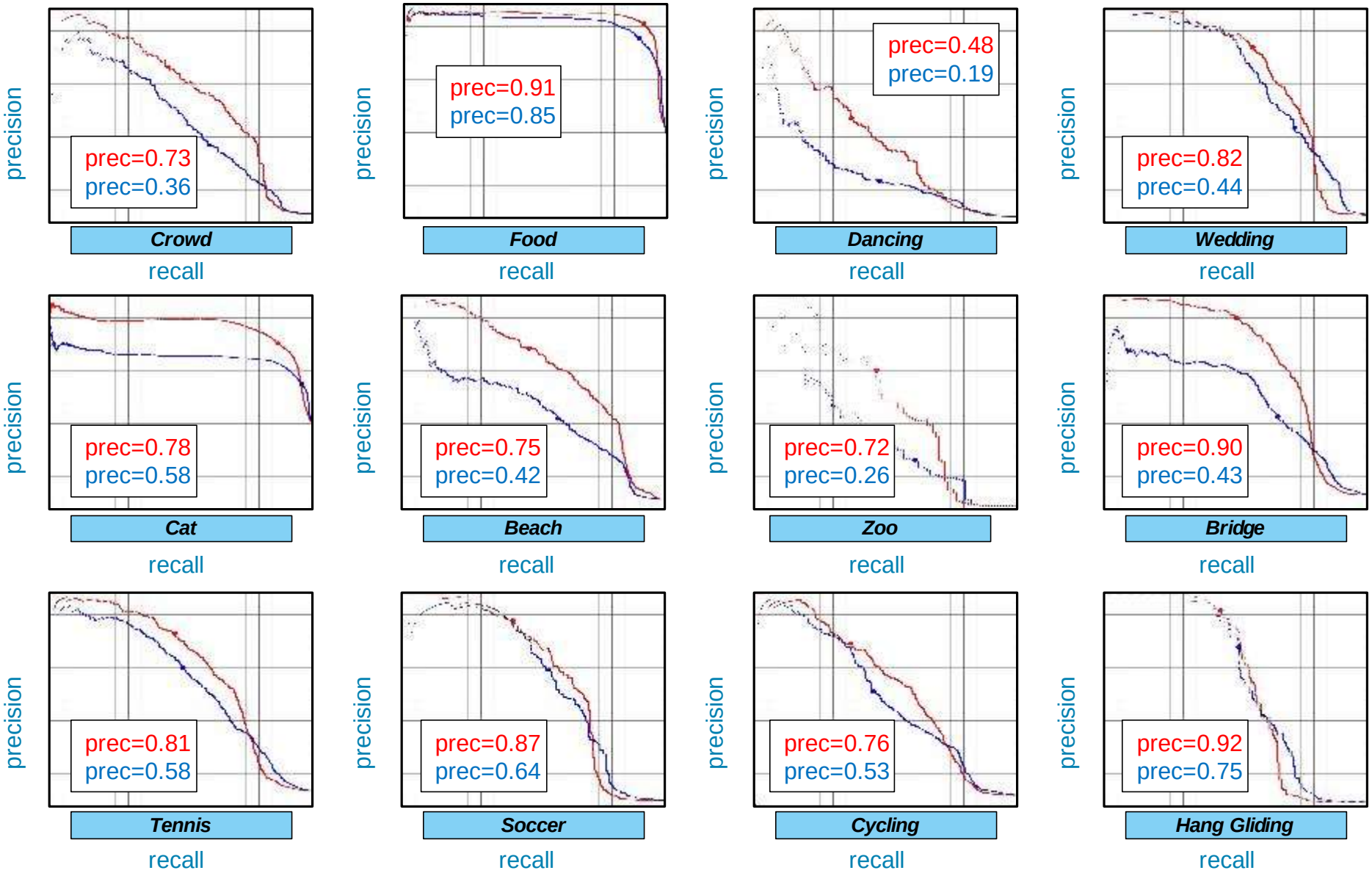
Scene for *canoe* can be a *beach* or *park,*
and *canoe* can have *people* or *no people*
→ **not mutually exclusive across facets**

[ACTIVITY] Cricket

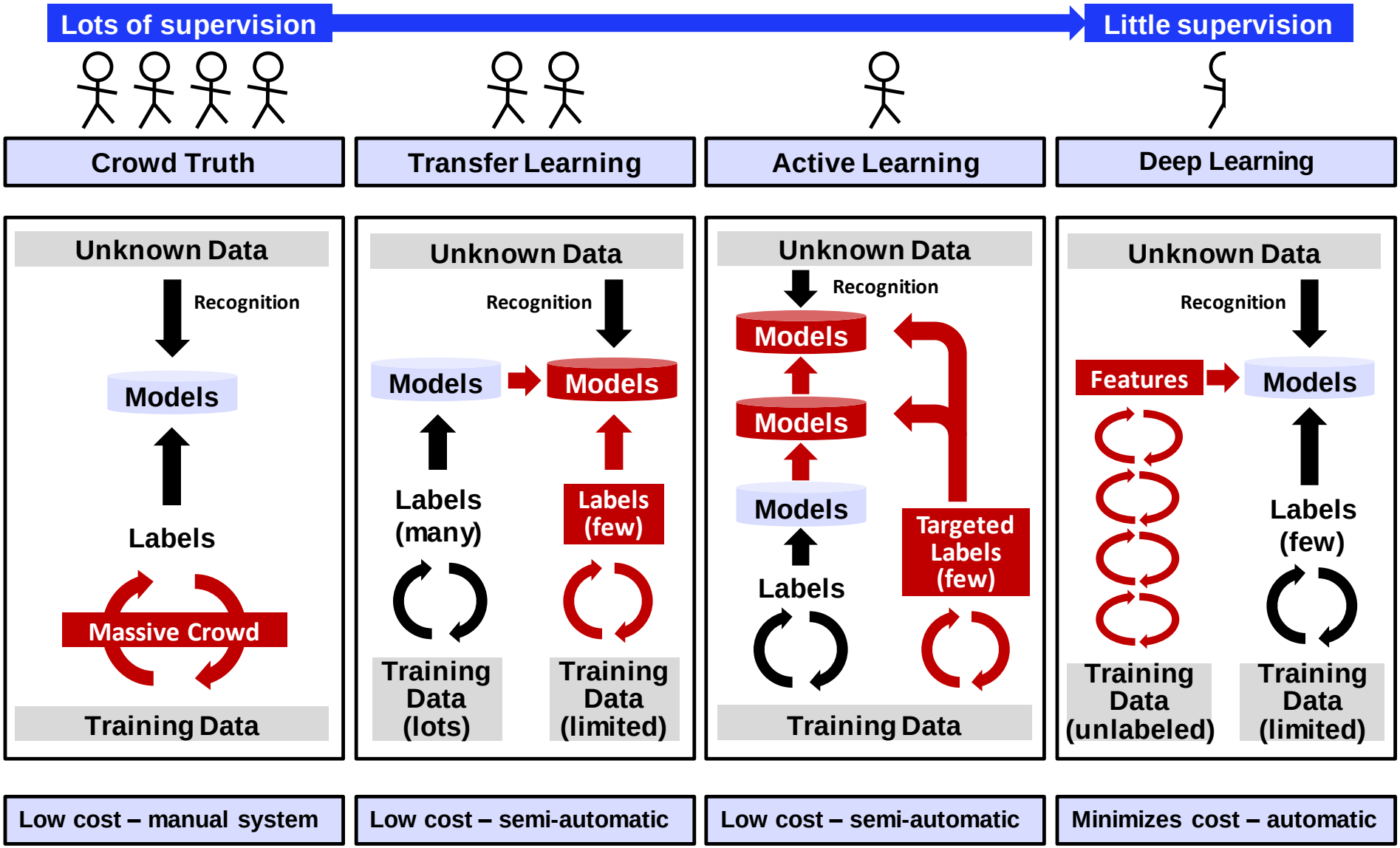


Scene for *cricket* can be a *park,*
and *cricket* can have *people* or *no people*
→ **not mutually exclusive across facets**

Results (582 labels): Faceted Hierarchy (acc=0.95) vs. Taxonomy (acc=0.89)

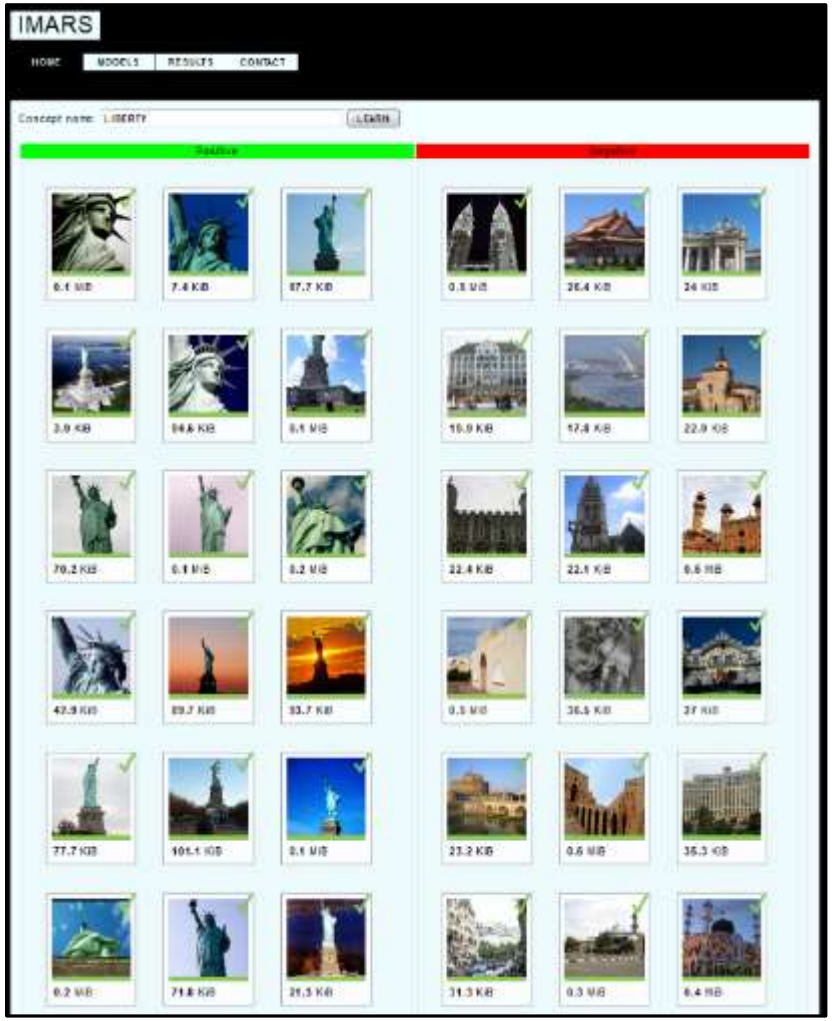


Putting the User in-the-Loop in Visual Semantic Learning

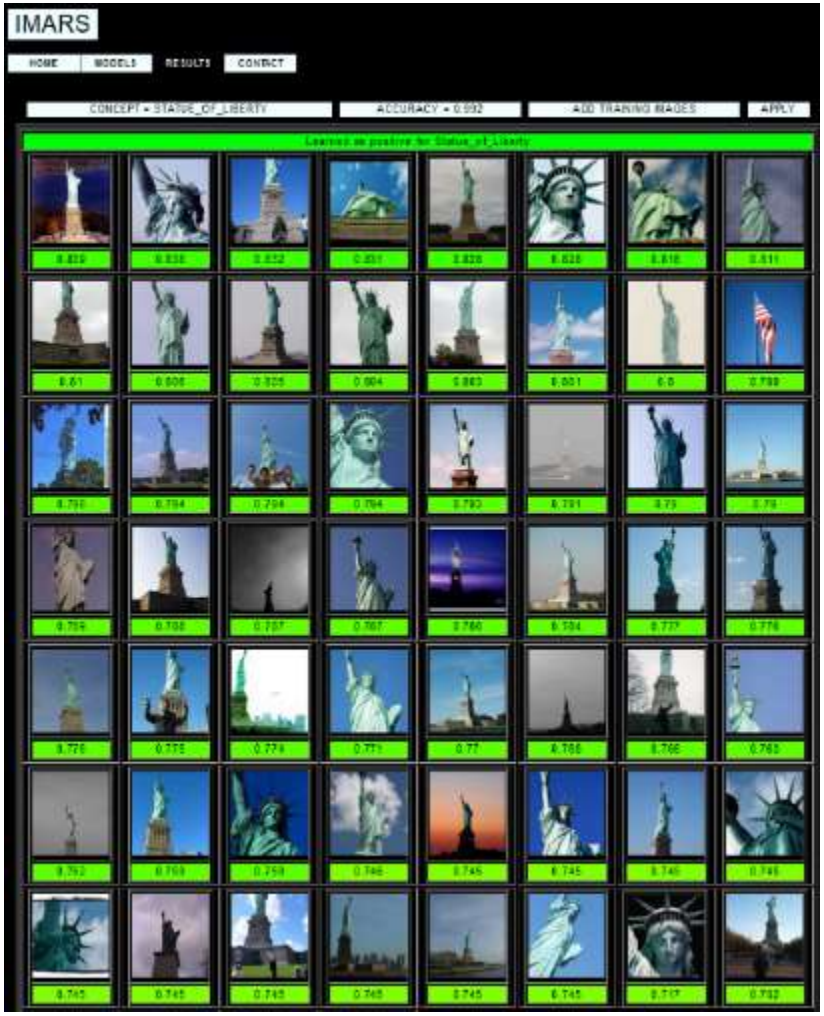


Visual Learning Service – Simple Drag-and-Drop Training Example

Upload Training Images



Learn Visual Classifier



Demo

Visual Learning can use Active Learning to Iteratively Train Models

Unknown Images

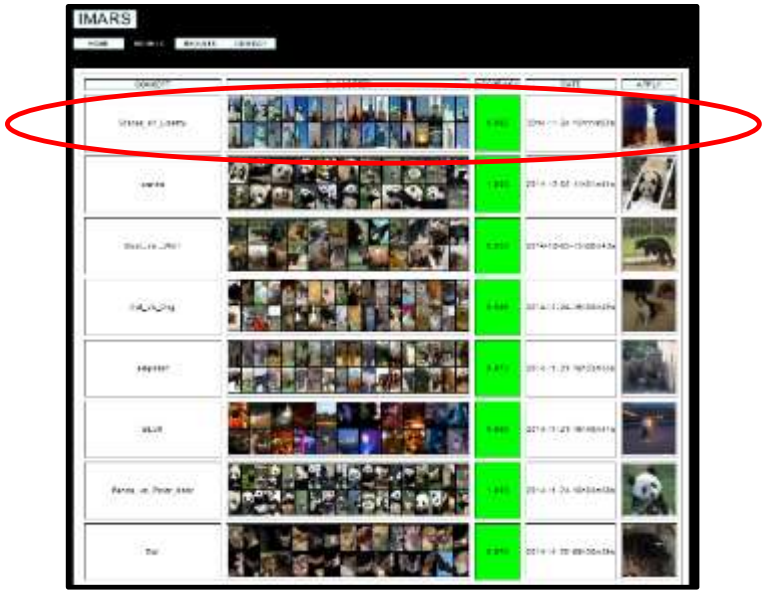
Labels and Scores

Refine

```
14 Statue_of_Liberty 0.754721
15 Statue_of_Liberty 0.793386
16 Statue_of_Liberty 0.774210
17 Statue_of_Liberty 0.775921
18 Statue_of_Liberty 0.688645
19 Statue_of_Liberty 0.756334
20 Statue_of_Liberty 0.611773
21 Statue_of_Liberty 0.800345
22 Statue_of_Liberty 0.745605
```

IN

OUT



Demo

Key Messages:

Image and video data is growing in volume and importance

Manual analysis is not scalable … cannot keep up

Vision system development has historically required deep expertise

Expert Analysis → Data-driven Visual Learning

Increasing availability of labeled training data

Emergence of increasingly sophisticated visual learning algorithms

Focus needed on Visual Semantic Modeling

How to best model visual world (concepts and relationships)

How to combine visual semantic modeling with learning systems