



The CLEF 2011 Plant Images Identification Task

Alexis Joly, Hervé Goëau, Pierre Bonnet, Nozha Boujemaa, Daniel Barthelemy, Jean-François Molino, Philippe Birnbaum, Elise Mouysset and Marie Picard

Context

- **Global warming** affects **environment** as well as **agriculture** and food's **safety**



- Accurate **knowledge** about **plants distribution** and **evolution** is essential for **sustainable agriculture** and **biodiversity** preservation



Challenge

○ **Accessing basic information** about plants is **still challenging**:

1. Botanical data is highly **decentralized** and **heterogeneous**

○ Each actor has its own expertise domain, data and format

2. Botanical data is **complex**

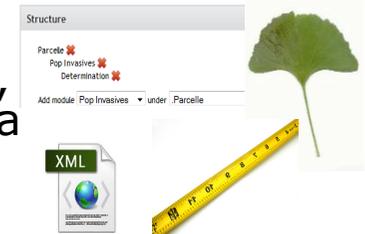
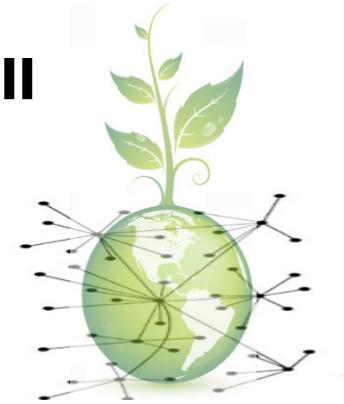
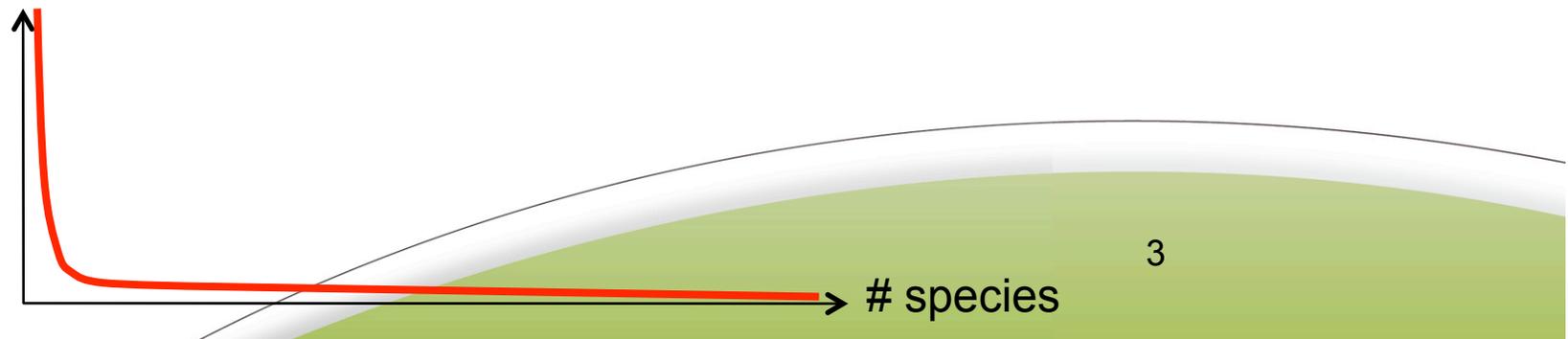
○ a single plant's observation includes (un)-structured tags, diverse images, Empirical measurements, Contextual data

3. Botanical data is **sparse** and **incomplete**

○ **Huge & Unknown** number of **species** (400K ?)

○ **Long tail** distribution (1 record per species !)

data



Taxonomic gap

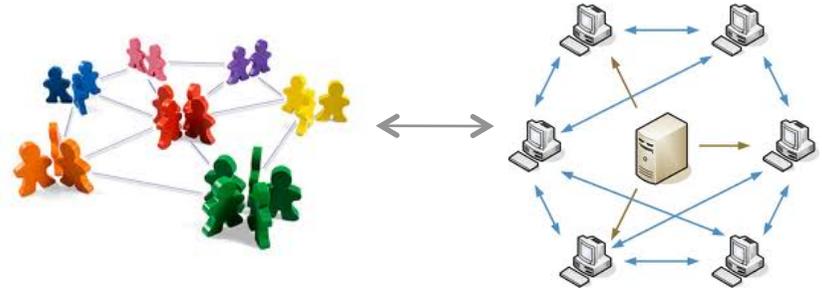
- As a consequence, **identifying plants is very difficult**
 - Large public & professionals: farmers, rangers, etc.
 - Even for the botanists themselves
- And managing ecosystems is problematic
 - How could a farmer select **adapted pesticides** ?
 - How to control **plant's distribution** and **evolution** ?



Towards bridging the gap

- **Collaborative Information Systems**

- Speeding up integration
- Sharing knowledge
- E.g. GBIF, EoL, PI@ntNet



- **Multimedia IR & Identification Tools**

- Notably **images** are: to acquire by anyone
- **Visual content** is very informative for characterisation

- **Image-based identification** (Task motivation)

- **CBIR** SoA not well studied **on plants**
- Few **datasets**, biased, narrow

ImageCLEF Plant Id Task

- A **trans-disciplinary** effort towards evaluating **multimedia IR** on **plants**

- Funded by  **Pl@ntNet** and  **CHORUS+**
AUDIO-VISUAL SEARCH

- Organized in collaboration with **botanists** and **environmental scientists**

- A focused & attractive **pilot task**

*as simple as possible **but** as realistic as possible*

- **Simple**

- 1 content type = images
- 1 organ = leaves (most studied)

- **Realistic**

- **Collaborative data** with contributors from 1 given climatic region

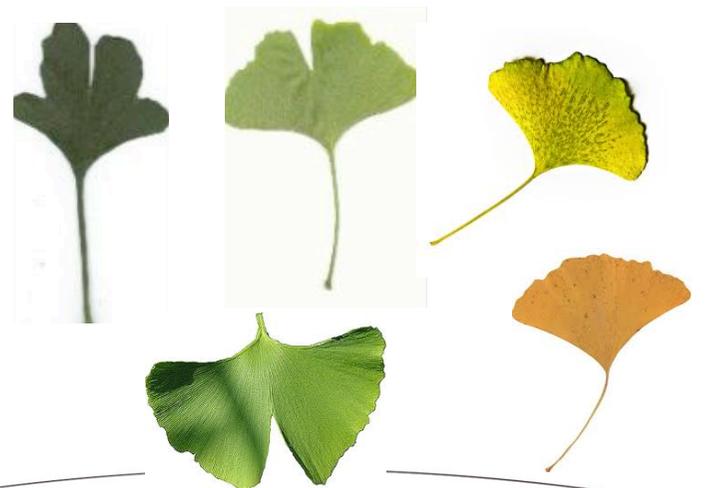
Data collection & creation

- Problem: How to reduce **bias** between **training** data and **real user's** data ?

Training data



User data



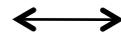
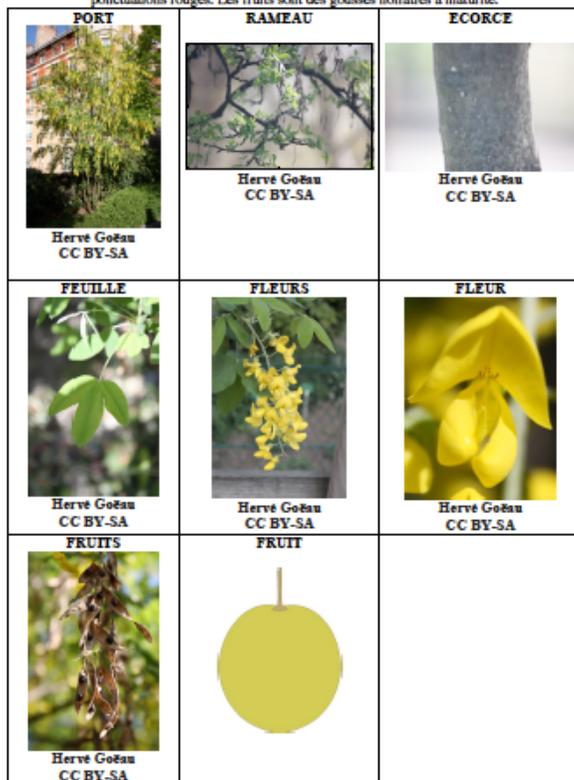
Data collection & creation



- o Solution 1: Let **real users collect** training data and botanists validate

Cytisse faux-ébénier : *Laburnum anagyroides* Medik.

<http://www.tela-botanica.org/s/flore/BDNFF/4.02/an/60511>
écorce lisse, brune et à rameaux duveteux pouvant mesurer jusqu'à 8 m de haut comme de large. Feuilles caduques de 4 à 8 cm de long à 3 folioles ovales et court pétiole. Fleurs en grappes pendantes de 10 à 20 cm de long, de couleur jaune vif avec des ponctuations rouges. Les fruits sont des gousses noirâtres à maturité.



Data collection & creation

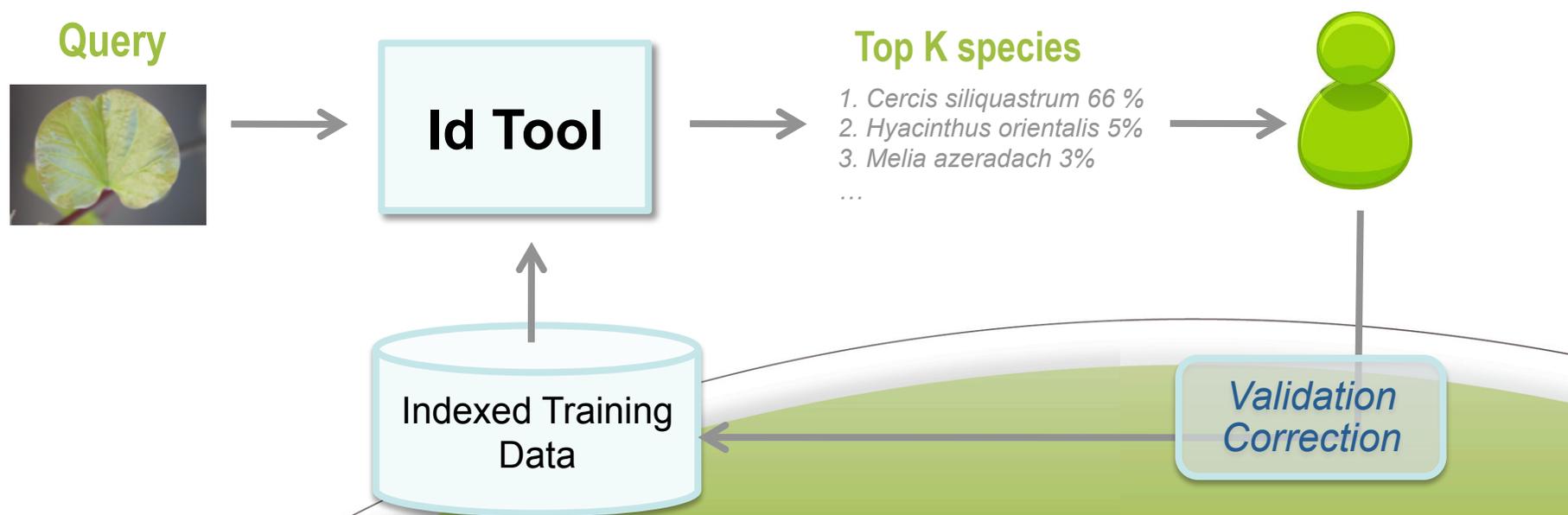
- Solution 1: Let **real users collect** training data
- Solution 2: Grow **training data** with **test data**

✓ **Online Identification Tool**

Play

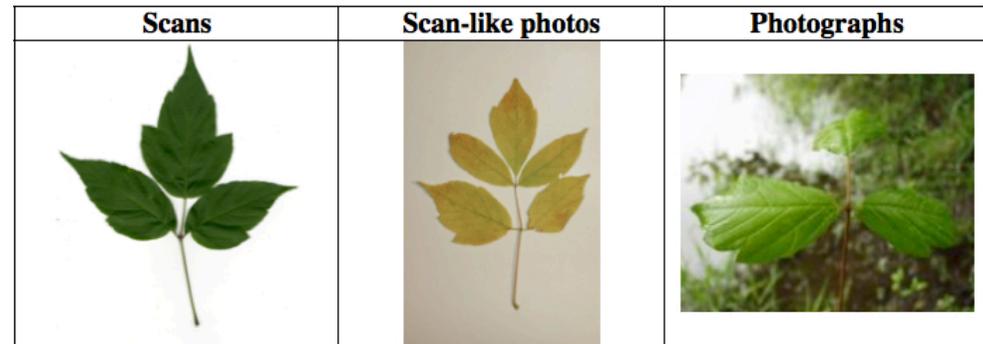
✓ **Online Validation & Correction**

Contribute



Pl@ntLeaves dataset

- **70 Mediterranean species**
- **5436 images of 3 types**
 - Scans
 - Photos with uniform background
 - Unconstrained photos

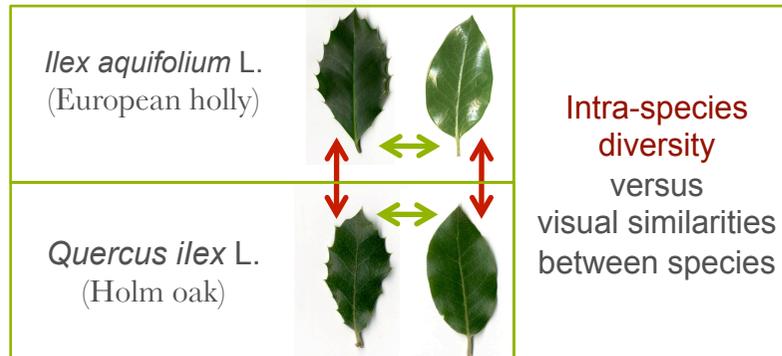
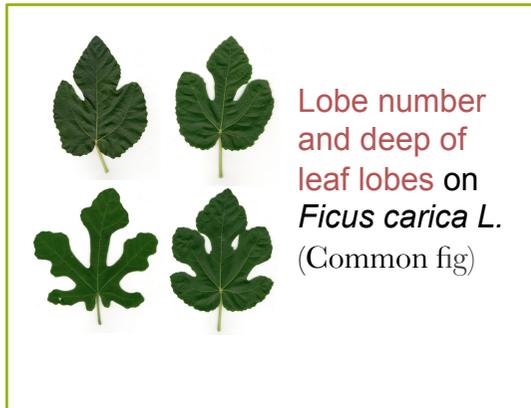


- **Metadata (XML)**
 - **Type** (scans, photos,...)
 - **GPS**
 - Author
 - Amount of leaves

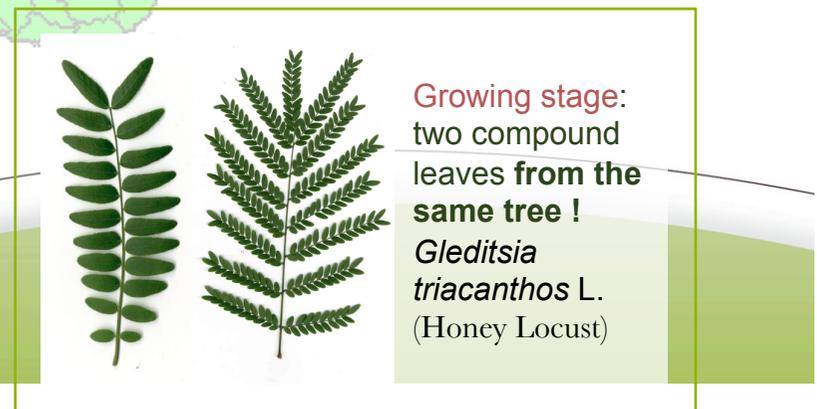


```
▼<Image>
  <FileName>997.jpg</FileName>
  <Date>07/07/10</Date>
  <Type>Scan</Type>
  <Organization>IRD</Organization>
  <Author>Molino Jean-francois</Author>
  <IndividualPlantId>262</IndividualPlantId>
  ▼<Taxon>
    <Sub-regnum>Tracheobionta</Sub-regnum>
    <Regnum>Plantae</Regnum>
    <Class>Magnoliopsida</Class>
    <Division>Magnoliophyta</Division>
    <Order>Lamiales</Order>
    <Family>Verbenaceae</Family>
    <Species>agnus-castus</Species>
    <Genus>Vitex</Genus>
  </Taxon>
  <VernacularNames>Vitex</VernacularNames>
  <Locality>France - Hérault</Locality>
  <Content>Leaf</Content>
  ▼<GPSLocality>
    <Longitude>3.258363889</Longitude>
    <Latitude>43.59123611</Latitude>
    <Altitude>0.0</Altitude>
  </GPSLocality>
</Image>
```

Pl@ntLeaves diversity



Localities,
seasons,
Users = # environments,
climate,
ecosystems



Task description

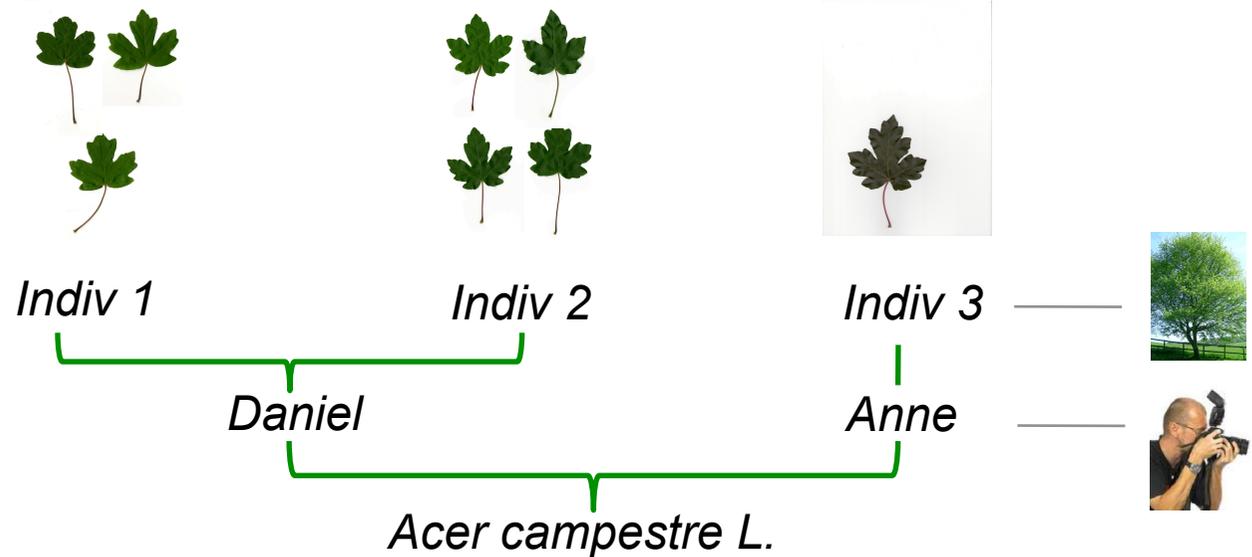
- A classification task (1 picture vs. 70 classes)
- A **plant-oriented** split strategy

		Nb of pictures	Nb of individual plants	Nb of contributors
Scan	Train	2349	151	17
	Test	721	55	13
Scan-like	Train	717	51	2
	Test	180	13	1
Photograph	Train	930	72	2
	Test	539	33	3
All	Train	3996	269	17
	Test	1440	99	14

- **Separate scores** for the **3** image **types**
 - Scans → plant's diversity study
 - Scan-like photos → real lighting conditions
 - Unconstrained photos → back to the future
- **Free training** strategy (with or without meta-data, photo vs. scans)

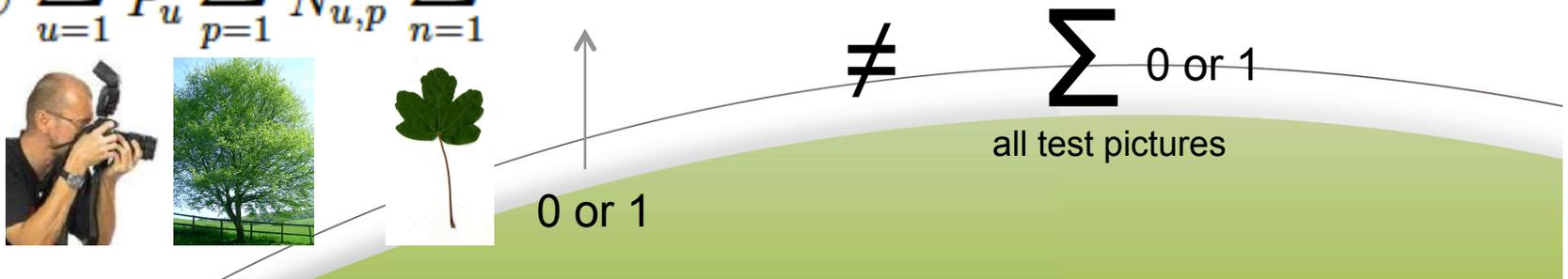
Task description

- **Unbalanced**
Real-world
Training data



- **Normalized** Average Classification Score (closer to **any** user's perception)

$$S = \frac{1}{U} \sum_{u=1}^U \frac{1}{P_u} \sum_{p=1}^{P_u} \frac{1}{N_{u,p}} \sum_{n=1}^{N_{u,p}} s_{u,p,n}$$



Participation

- About 40 groups registered
- 8 groups submitted a total of 20 runs (max was 4)

Group	Nb of runs	Methods/focus
DAEDALUS	1	SIFT visual features + NN classifier
IFSC	3	Boundary shape features
KMIMMIS	4	SIFT visual features + NN classifier
INRIA	2	Large scale matching, boundary shape
LIRIS	4	Model-driven boundary shape features
RMIT	2	GIFT visual features, 2 ML approaches
SABANCI-OKAN	1	Global visual features + SVM
UAIC	3	Metadata & visual features

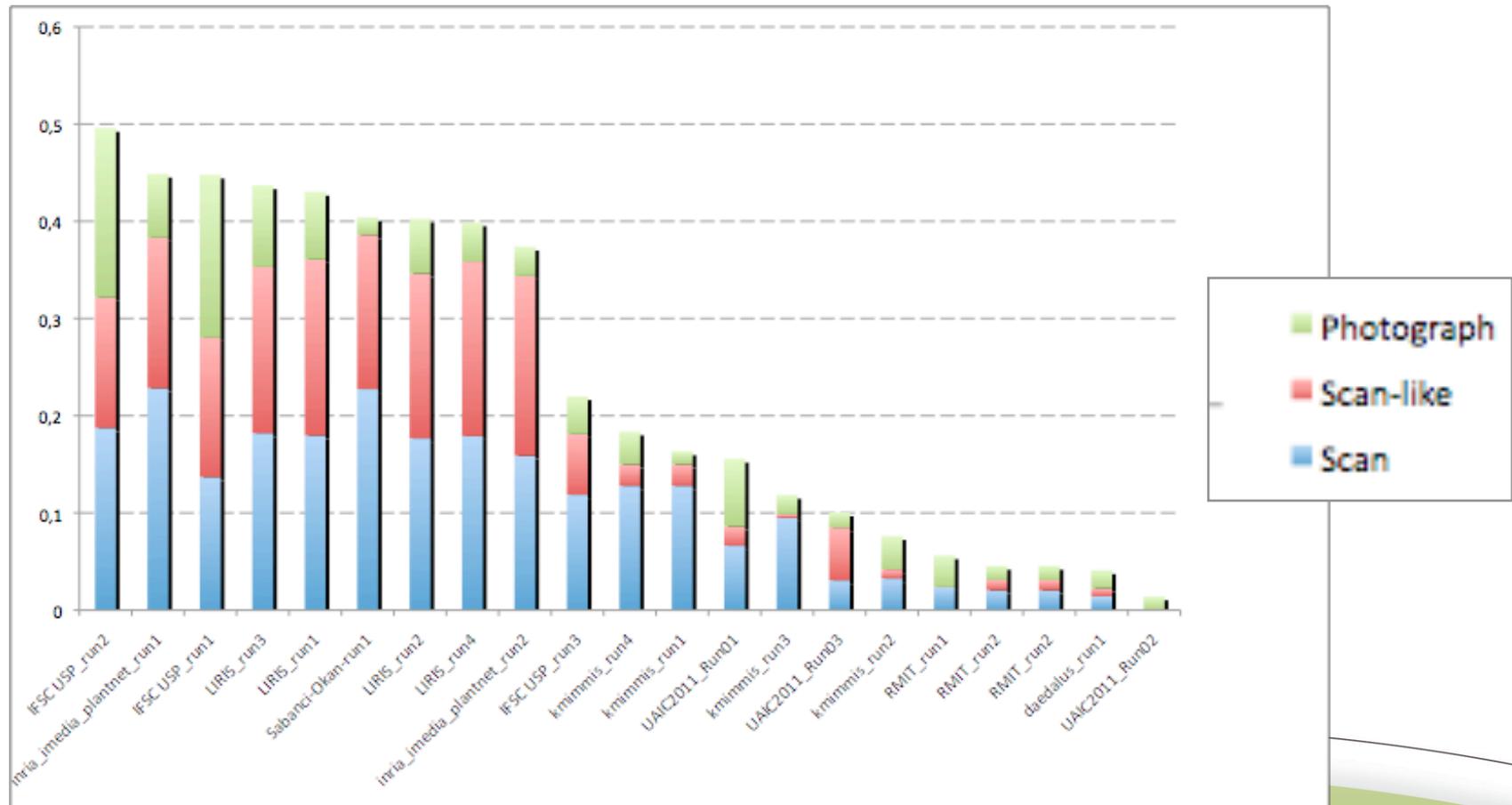


Participation

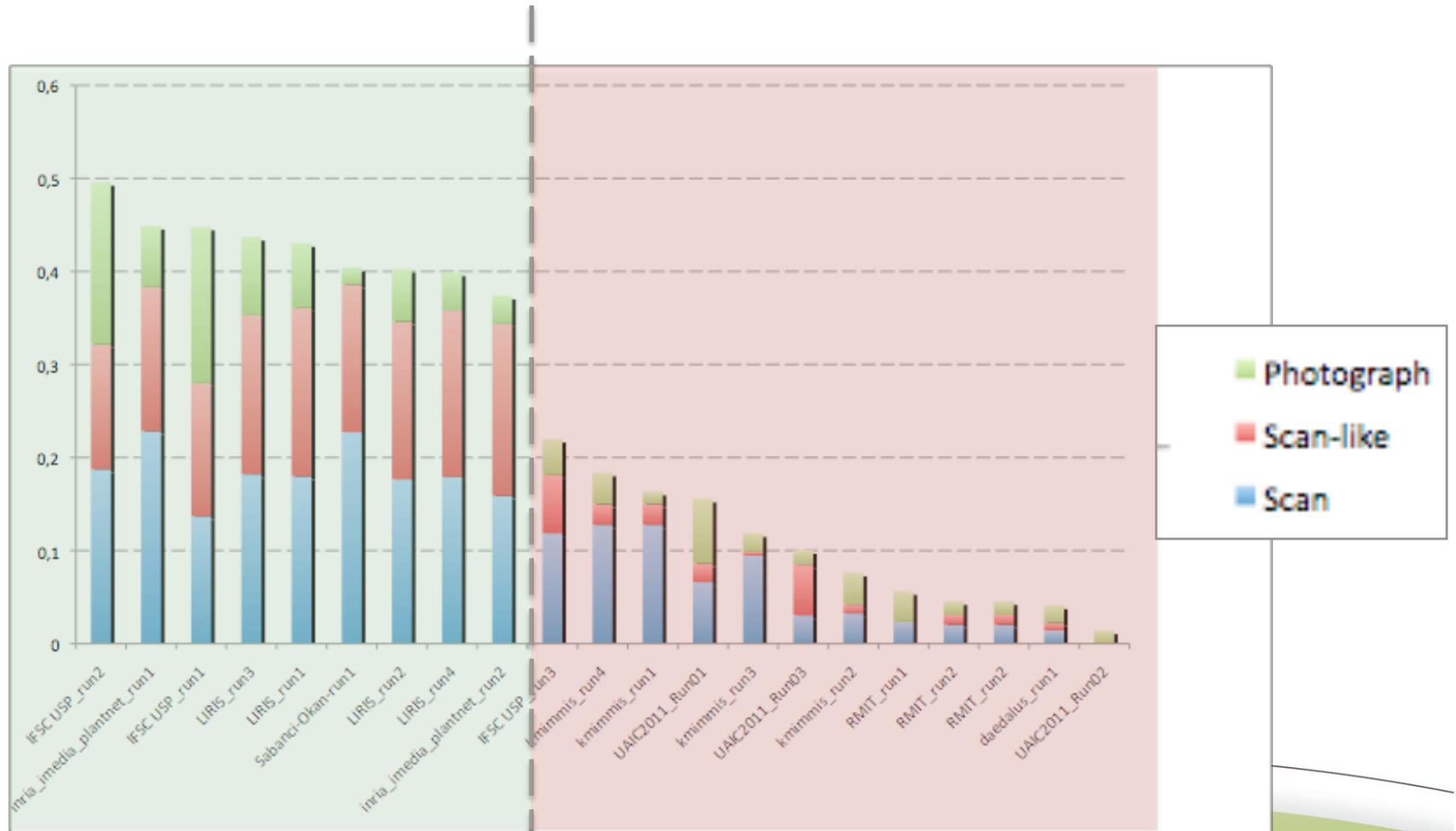
- *Used Media*
 - **Only visual content** 18/20 runs
 - Using metadata 2/20 runs
- *Used Visual features*
 - **Leaf boundary features** (SoA for leave's recognition) 8/20 runs
 - Global visual features 5/20 runs
 - SIFT features 5/20 runs
 - Other local features (harris, histograms) 1/20 runs
- *Classifiers*
 - **NN-classifiers** 15/20 runs
 - SVM 4/20 runs
 - Decision tree 1/20 runs



Results: overview



Results: overview



Results: overview

Using leaf
boundary
features

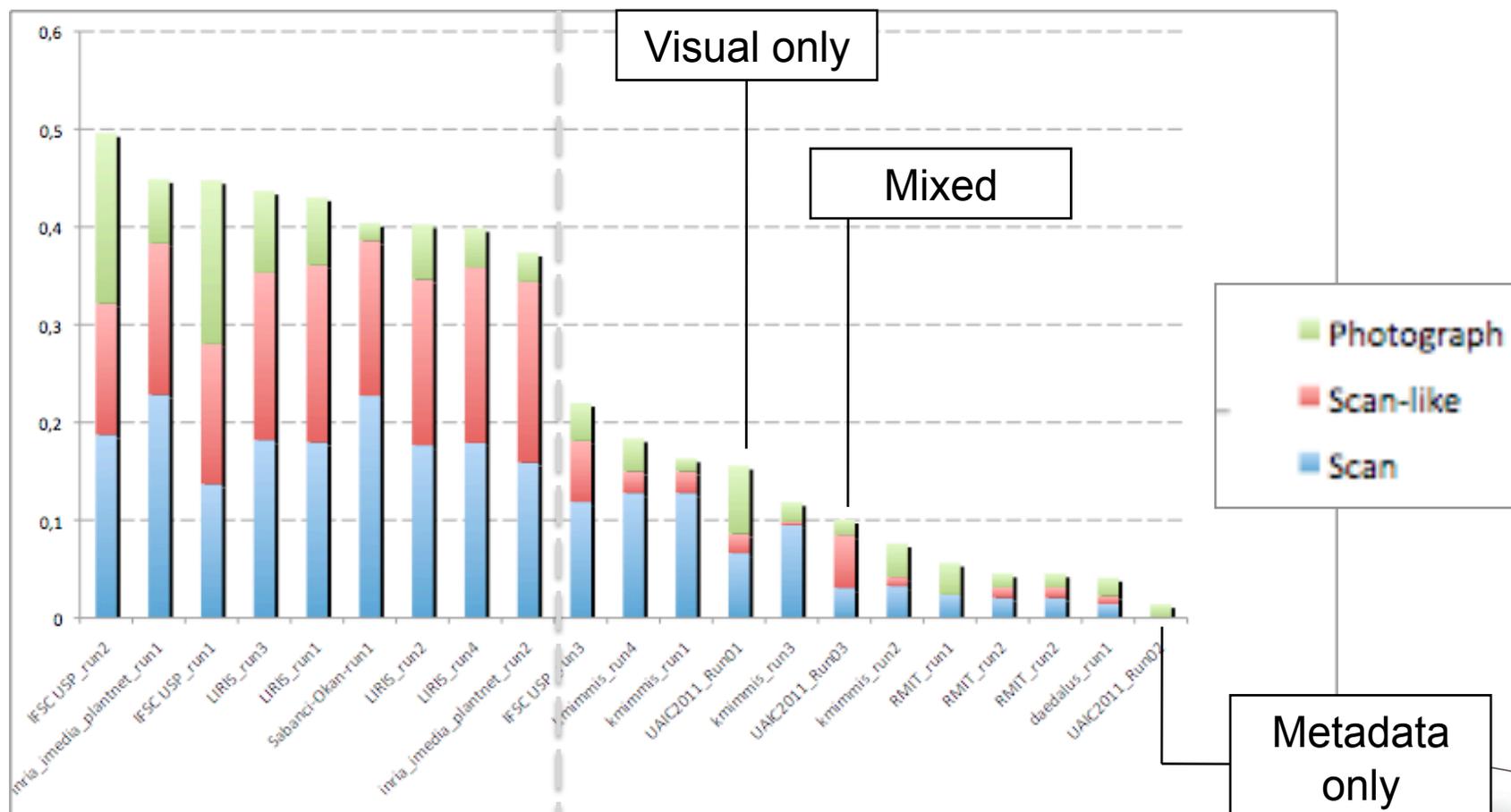


No leaf
boundary
features



Results: metadata

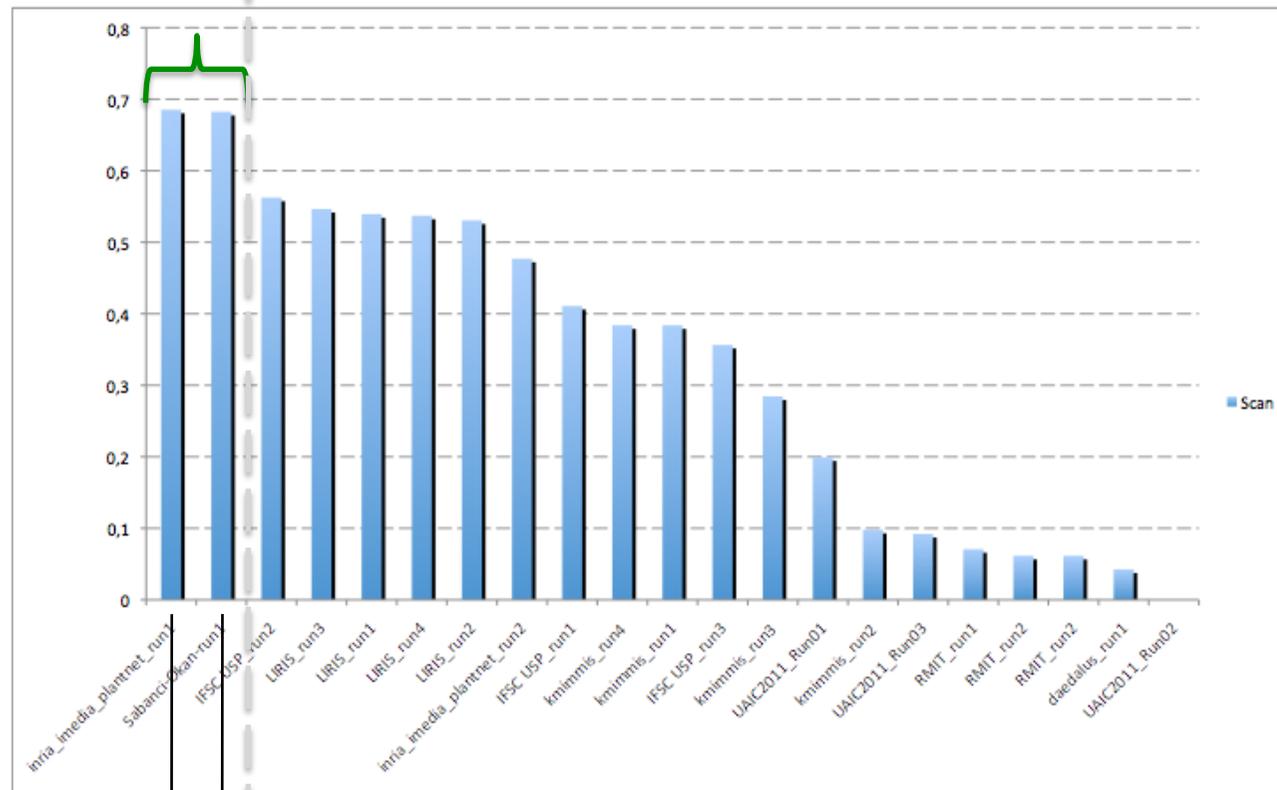
- Geo tags not so useful in a localized climate region ?
- Relations between metadata are complex



Results: focus on scans

Leaf boundary features alone not sufficient to disambiguate plant's morphologies

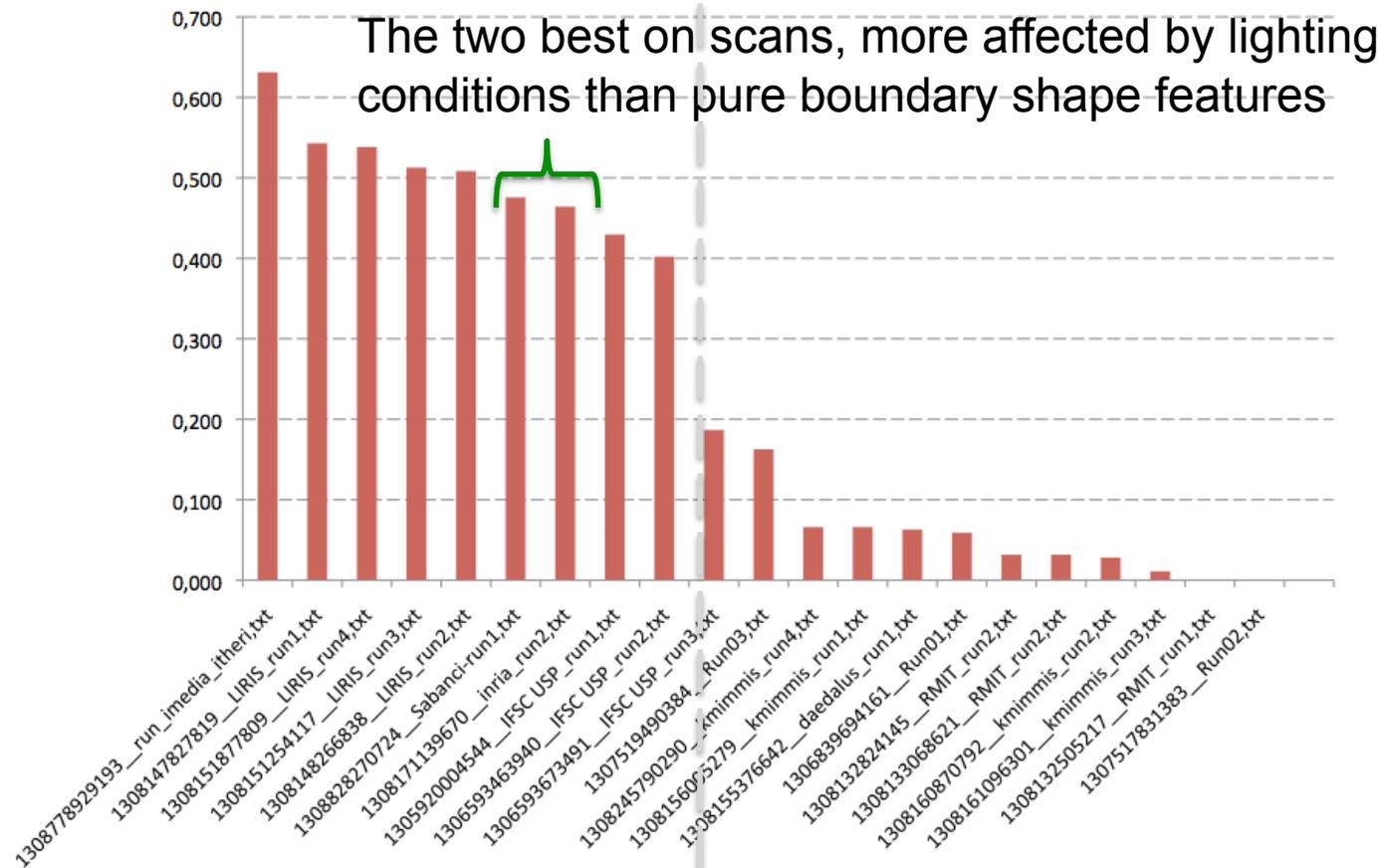
Not pure boundary shape features



Mixed global & shape boundary features
SVM

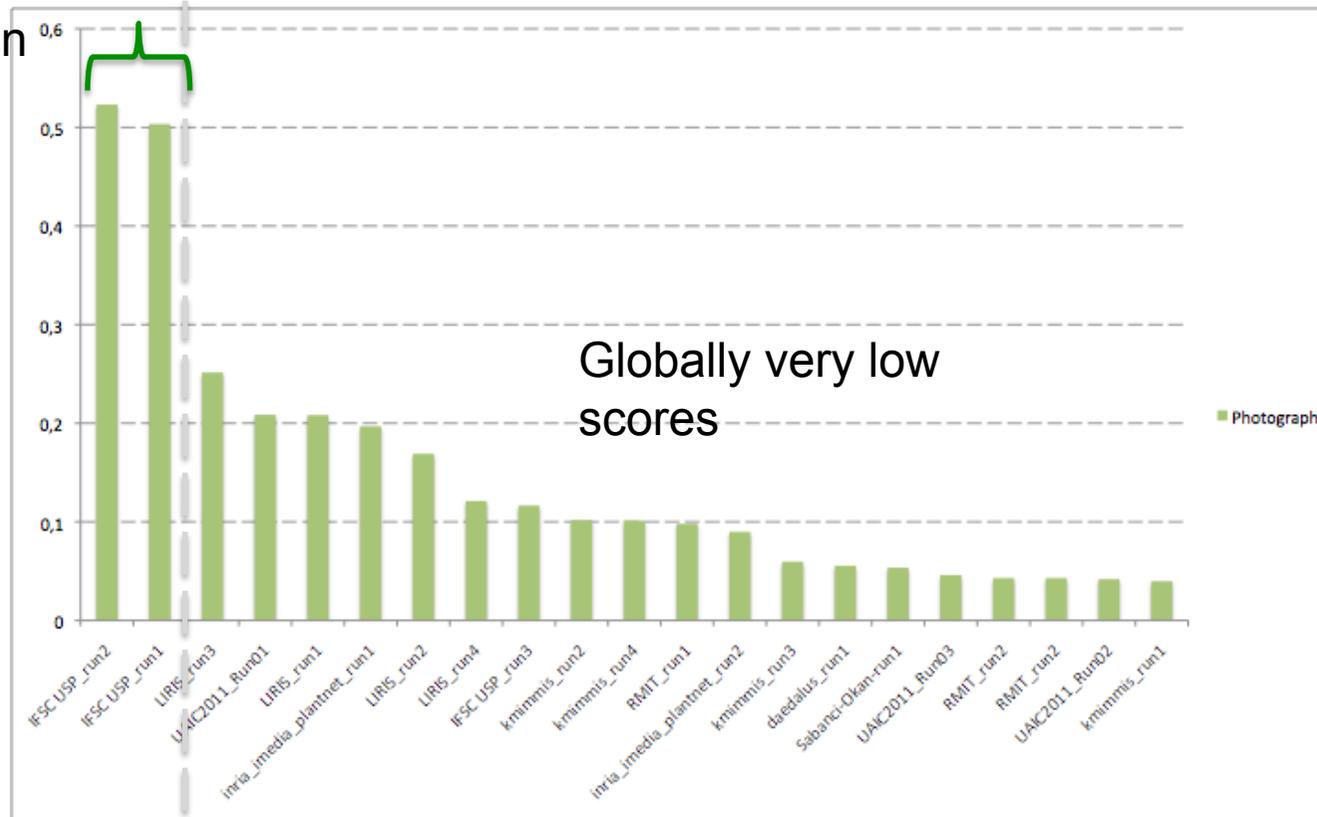
Mixed local features
Rigid object model

Results: focus on scan-like



Results: focus on unconstrained photos

1st, 2nd: with manual segmentation



Results: machine learning concerns

[Yanikoglu et al., Sabanci-Okan univ, ImageCLEF 2011]

- SVM vs. NN classifier: only slight improvements (Cross-validation scores)

Descriptors	1NN	SVM/SMO
1,3	79.59	80.27
1,2,3	85.02	86.93
1,2,3,4	85.45	87.11
1,2,3,4,5	87.36	89.15
1,2,3,4,5,6	88.59	90.20
1,2,3,4,5,6,7	88.41	90.57

- **Cross-validation** (random splits) vs. **Real test** (plant-based & author-based split)

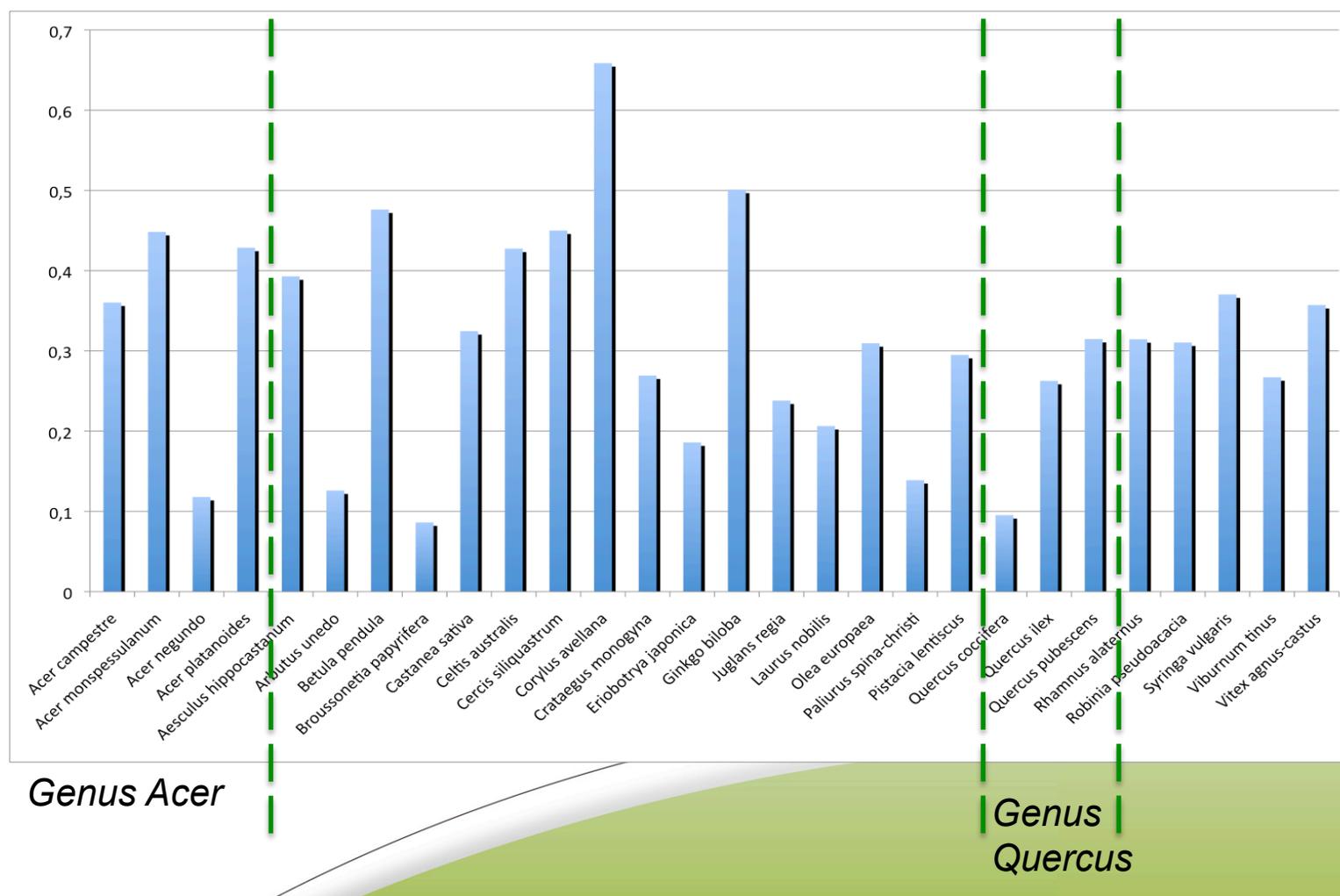
Cross-valid	Real
90.57	0.579



Back to botany

What is difficult across all runs ?
Results by species

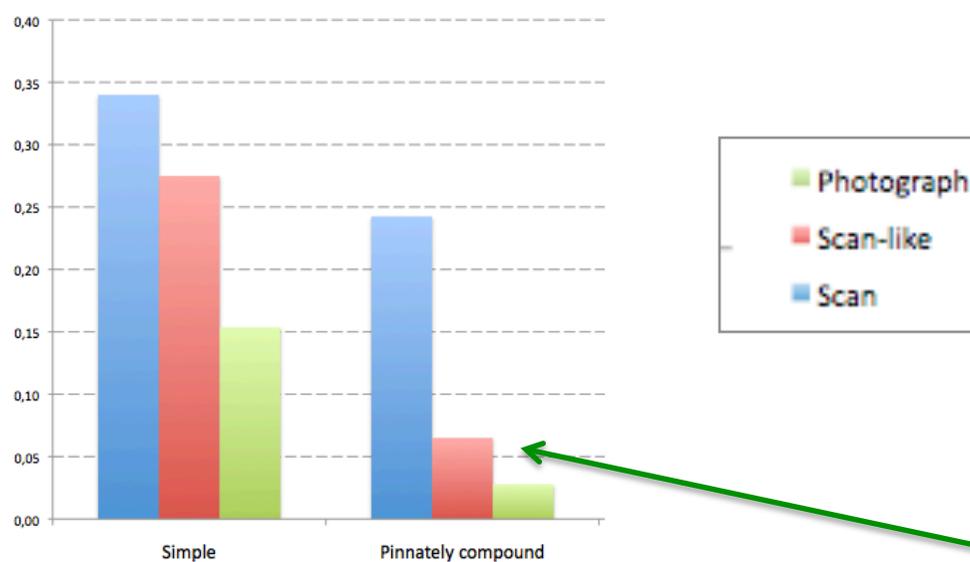
High variability
Not interpretable in terms of genus



Back to botany

What is difficult across all runs ?

Results by leaf morphology (based on botanist's manual classification)



Scan → scan-like :
Shadows ?
Overlaps ?



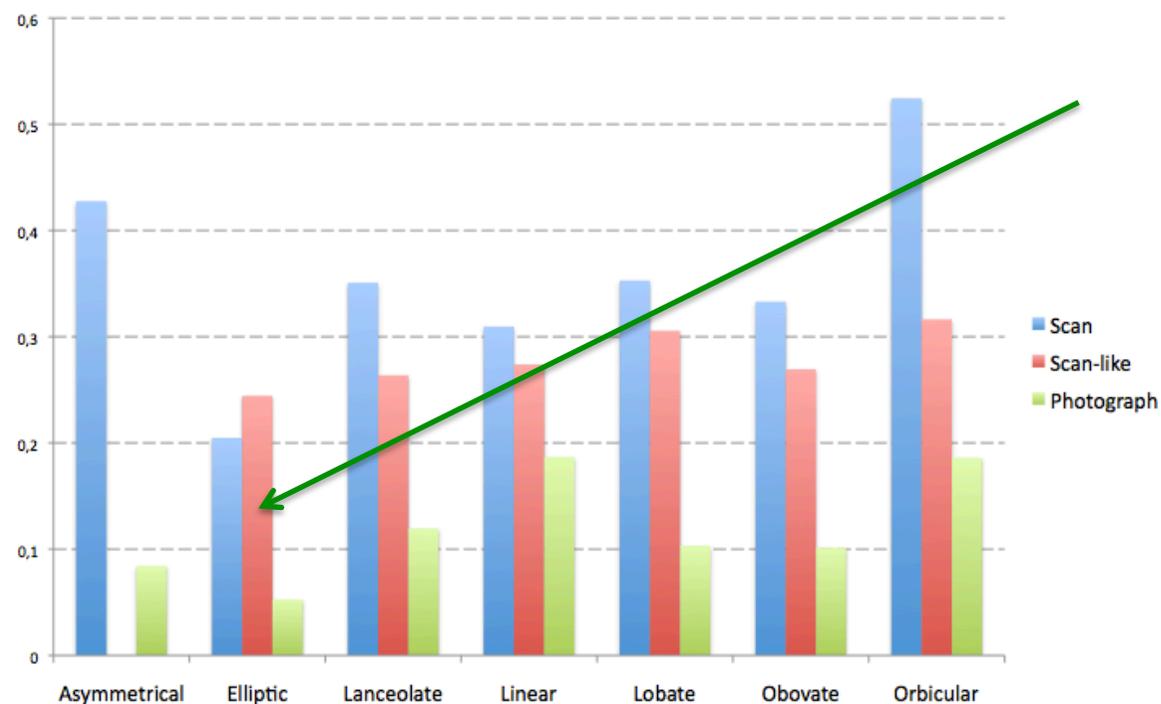
51 species

16 species

Back to botany

What is difficult across all runs ?

Results by leaf morphology: simple leaves only



Elliptic shapes:

Most ambiguous category

Most common leaf shape

The shape's information itself is poor

Species:

1

31

15

2

13

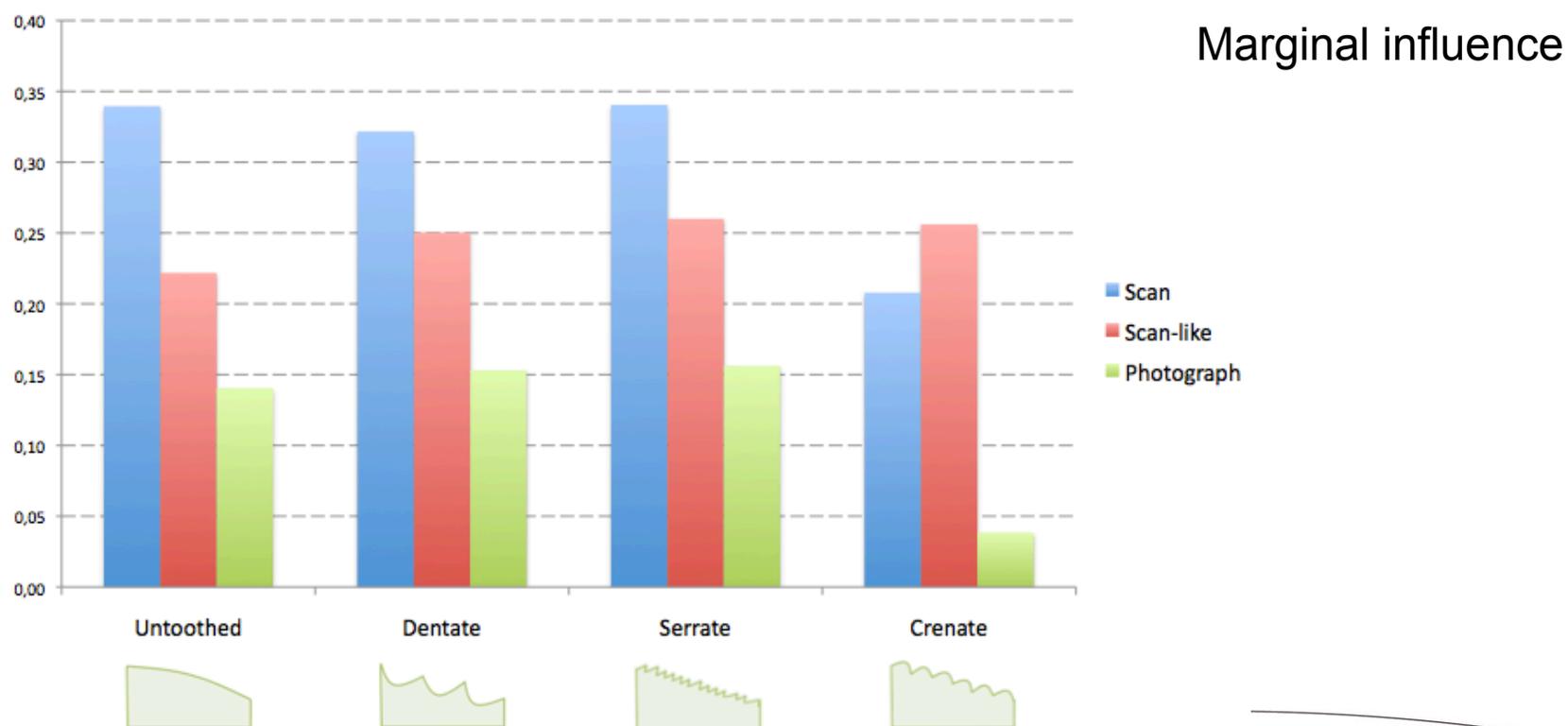
18

10

Back to botany

What is difficult across all runs ?

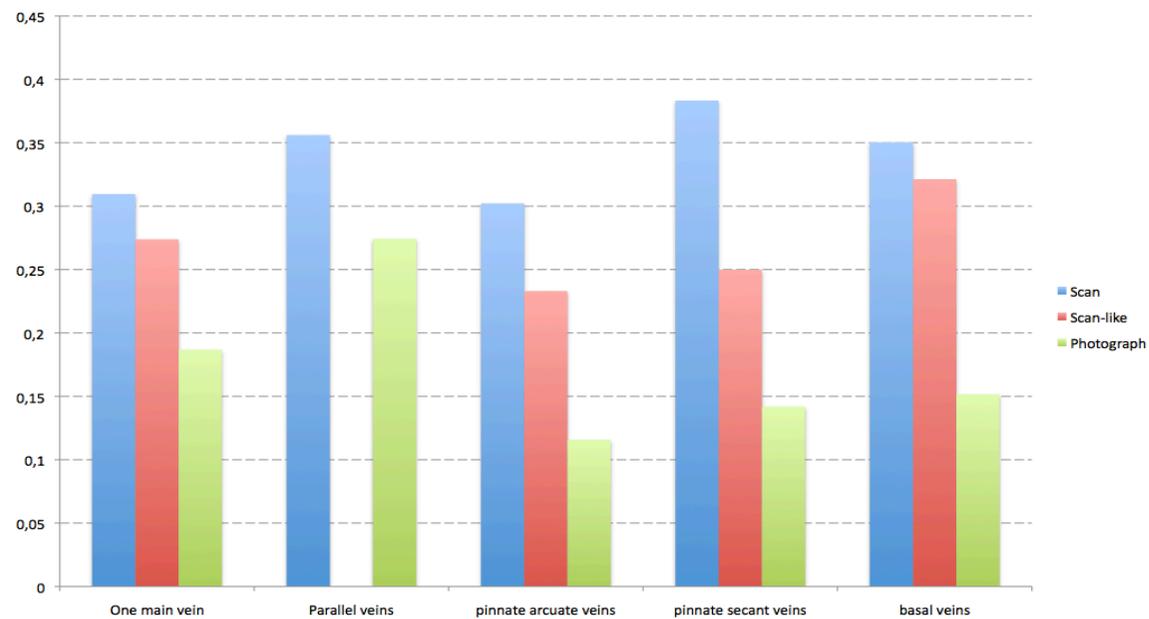
Results by leaf morphology: margin type (simple leaves)



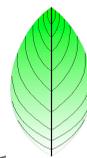
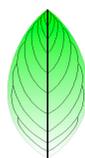
Back to botany

What is difficult across all runs ?

Results by leaf morphology: vein network type



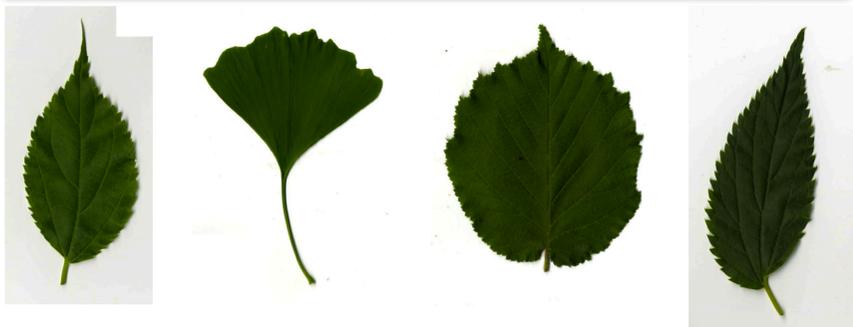
Marginal influence



Back to botany

What is difficult across all runs ?
Results by individual leaves

Easiest leaves (17/20)



Most difficult (0/20)



- On the long term, users would probably submit only **easy leaves**
- So that practical identification rate would increase (professionals, etc.)

Conclusions

- **Good identification scores on scans and scan-like**
 - Not far from mature
 - But high variability across species
 - Mainly due to leaf's morphology (compound leaves...)
 - **Unconstrained pictures much more challenging**
 - **SoA Shape boundary features are effective but**
 - Outperformed on pure scans when used alone
 - Combining them with other features may improve
 - **Local features + geometry effective as well**
 - Best results on scans
 - Most recent recognition methods not tested !!
 - **Benefits of advanced Machine Learning techniques is not clear**
- 

Perspectives & issues

- **More species ?**
 - European Trees \approx 200, European Plants \approx 2K, Tropical Plants \approx 200K
 - Pb = existing data too sparse (1 species = 1 image)
 - Depends on collaborative contributions...
 - **More organs ?**
 - Fruit, flower, trunk, etc.
 - Pb = existing data even more sparse
 - Depends on collaborative contributions...
 - **Beyond identification ?**
 - Towards a real botanical IR system
 - Rank-based task & metric
 - Multi-image & Multi-modal queries
 - Pb: task participation ...
- 

Thank you!!

And see you tomorrow morning

for  **CHORUS+** session
AUDIO-VISUAL SEARCH

8:30 – 10:00 Scientific Multimedia Data

- Earth observation
- BioSearch
- Collaborative botany
- Environmental informatics



Results: considering botany

Results by species: easiest species

Corylus avelana



Betula pendula



Ginkgo biloba



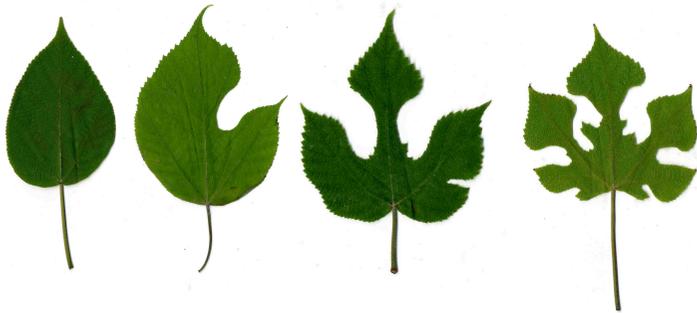
Acer monspessulanum



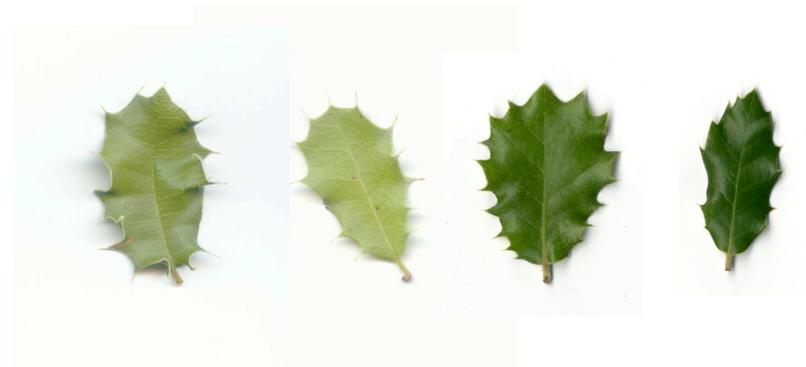
Results: considering botany

Results by species: hardest species

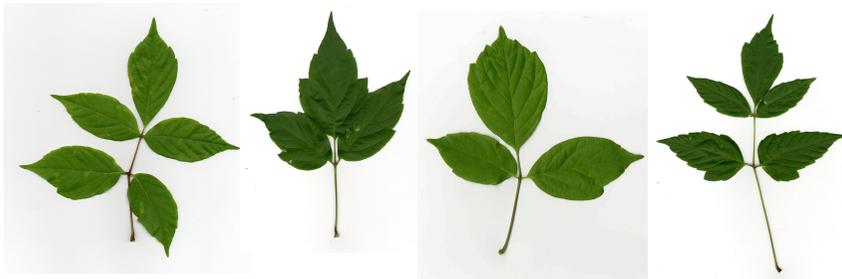
Broussonetia papyrifera



Quercus coccifera



Acer negundo



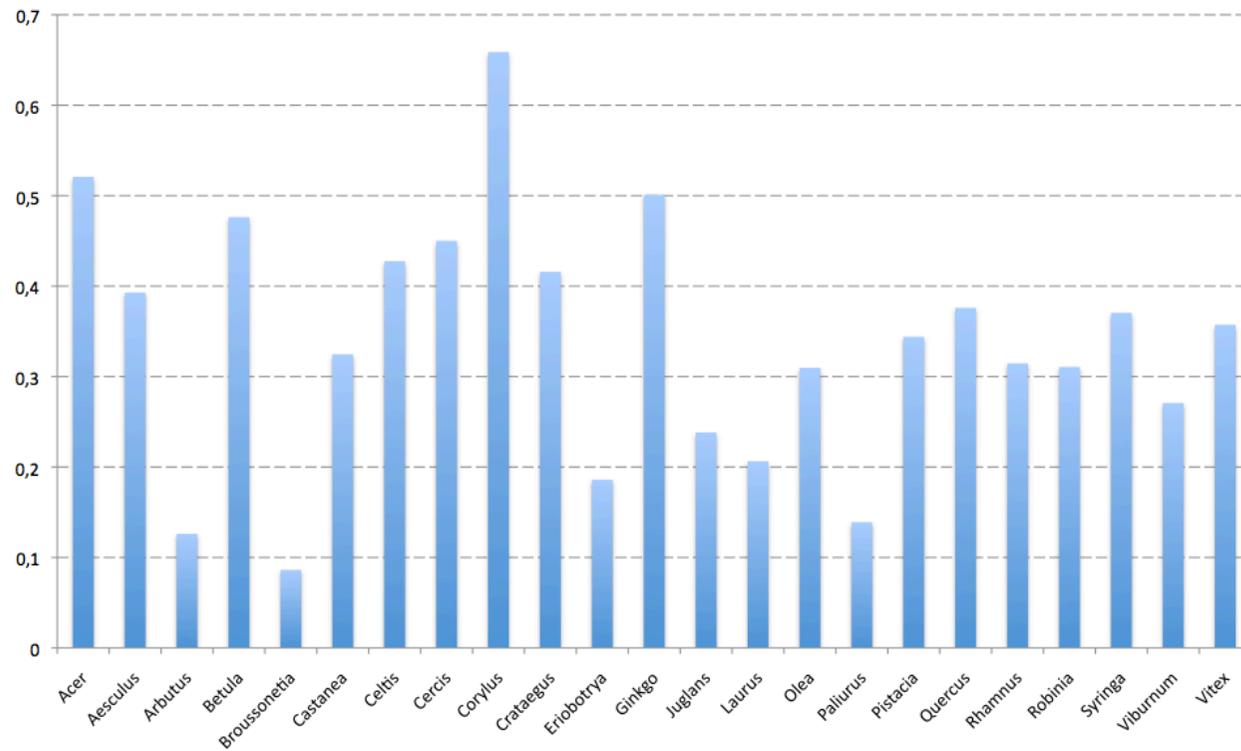
Arbutus unedo



Results: considering botany

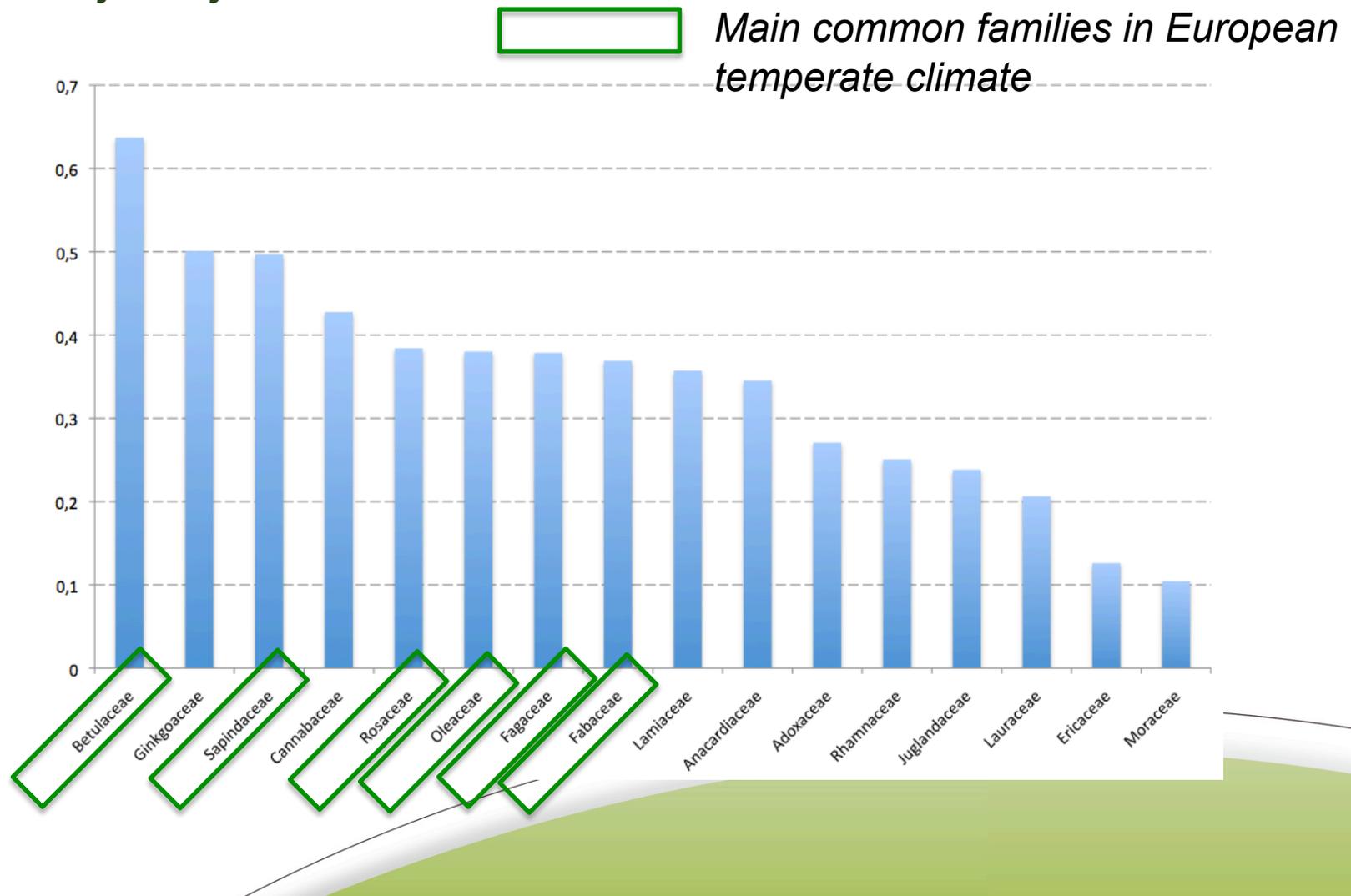
Results by genus

More or less the same for species



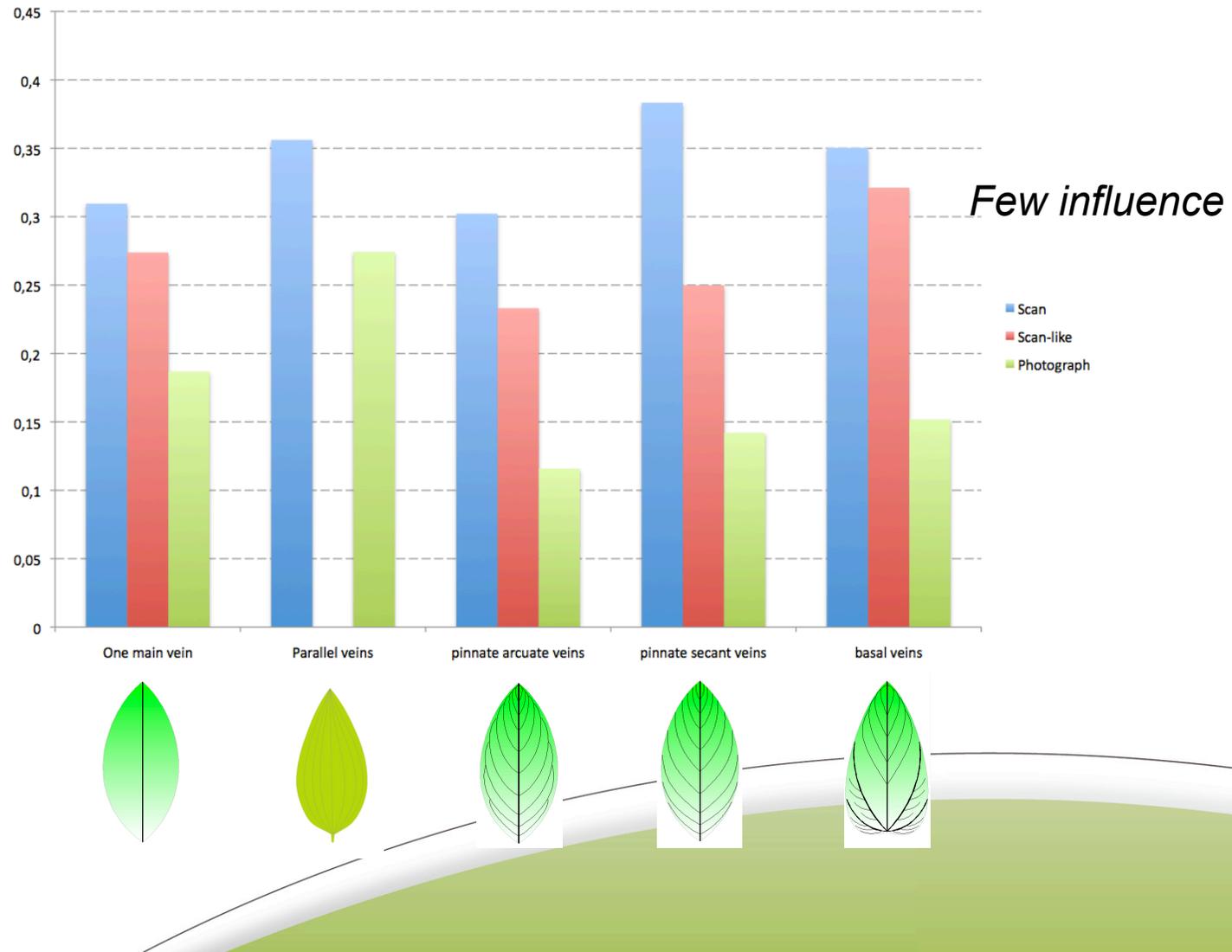
Results: considering botany

Results by family



Results: considering botany

Results by vein primary framework



Annexes





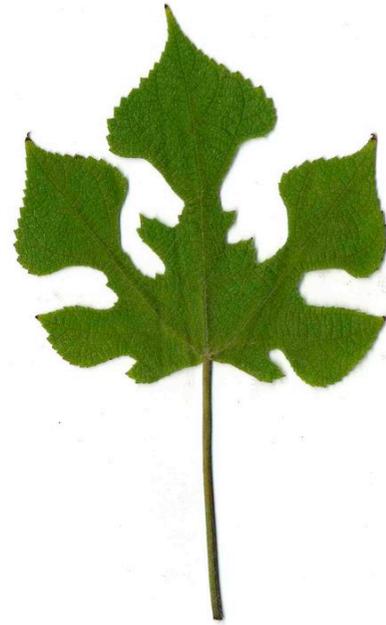
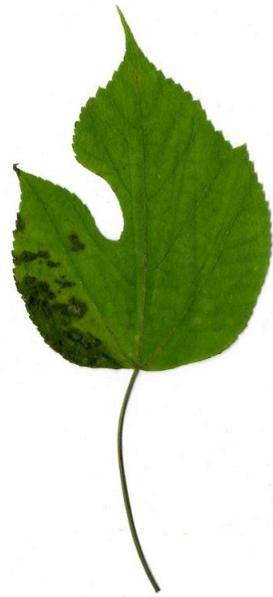
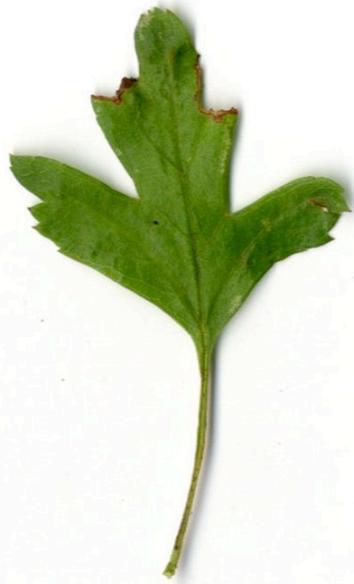
Voici les 3 images les mieux identifiées par l'ensemble de tous les participants (17 runs/20). De véritables mannequins !!





Et maintenant la galerie des monstres !!! les scans qui n'ont été identifiés par aucun run





Encore quelques « monstres » en photos cette fois-ci (« scan-like photos» identifiées par aucun système)



Quelques autres mal aimées, des un peu malade ou des marginales
(scans identifiés par aucun système)

