Mathematical Foundations of Plant Semantics

Simon Castellan[1], Jos Käfer[2], Eric Tannier[3]

[1] Inria Rennes
[2] CNRS
[3] Inria Lyon

February 16th, 2023

Another research

Can we do research that:

- solve a real need arising from society,
- empowers users without enslaving them (Illich's *conviviality*),
- is somewhat aligned with the ecological transition.

Another research

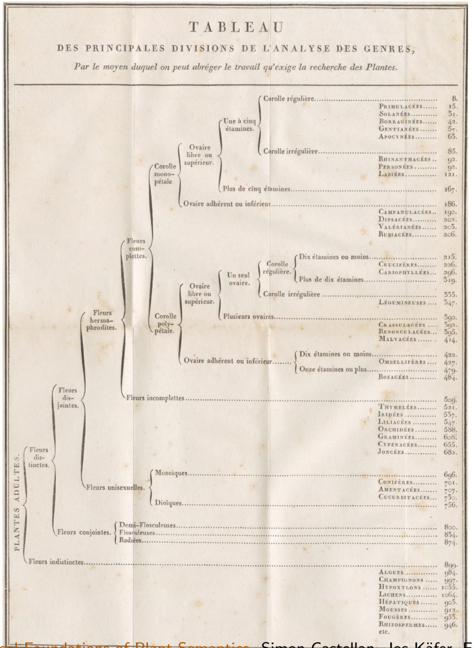
Can we do research that:

- solve a real need arising from society,
- empowers users without enslaving them (Illich's *conviviality*),
- is somewhat aligned with the ecological transition.

A case study, Pl@ntNet: can we improve on the points above?

- Hard to trust
- Very knowledgeable but a bad teacher
- What if it goes away?

Botanist technology for Plant ID [Lamarck'1805]



TABLEAU

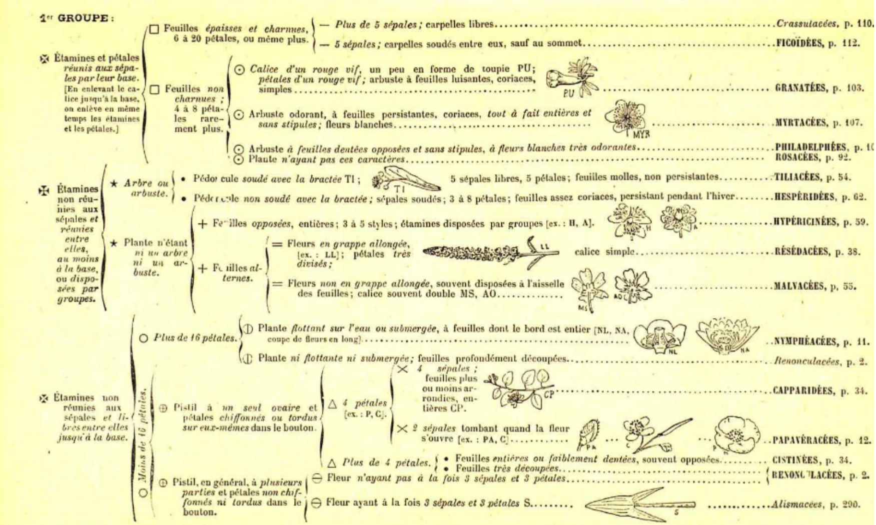
DES PRINCIPALES DIVISIONS DE L'ANALYSE DES GENRES,

Par le moyen duquel on peut abréger le travail qu'exige la recherche des Plantes.

PLANTES ADULTES.

- Fleurs distinctes.
  - Fleurs disjointes.
    - Fleurs hermaphrodites.
      - Fleurs complettes.
        - Corolle monopétale.
          - Ovaire libre ou supérieur.
            - Une à cinq étamines.
              - Corolle régulière .... 8.
                - Primulacées...... 15.
                - Solanées.......... 31.
                - Borraginées...... 42.
                - Gentianées....... 57.
                - Apocynées......... 63.
              - Corolle irrégulière .... 85.
                - Rhinanthacées .. 92.
                - Personées......... 92.
                - Labiées............ 121.
            - Plus de cinq étamines .... 167.
          - Ovaire adhérent ou inférieur .... 186.
            - Campanulacées.. 190.
            - Dipsacées.......... 202.
            - Valérianées...... 203.
            - Rubiacées.......... 206.
        - Corolle polypétale.
          - Ovaire libre ou supérieur.
            - Un seul ovaire.
              - Corolle régulière.
                - Dix étamines ou moins .... 215.
                  - Crucifères........ 226.
                  - Cariophyllées... 296.
                - Plus de dix étamines .... 319.
              - Corolle irrégulière .... 335.
                - Légumineuses.... 347.
            - Plusieurs ovaires .... 390.
              - Crassulacées .... 392.
              - Renonculacées... 395.
              - Malvacées .... 414.
          - Ovaire adhérent ou inférieur........
            - Dix étamines ou moins .... 422.
              - Ombellifères .... 427.
            - Onze étamines ou plus .... 479.
              - Rosacées.......... 484.
      - Fleurs incomplettes .... 509.
        - Thymelées........ 521.
        - Iridées .... 537.
        - Liliacées .... 547.
        - Orchidées .... 588.
        - Graminées......... 608.
        - Cyperacées........ 655.
        - Joncées............ 682.
    - Fleurs unisexuelles.
      - Monoïques .... 696.
        - Conifères.......... 701.
        - Amentacées....... 707.
        - Cucurbitacées.... 730.
      - Dioïques .... 756.
  - Fleurs conjointes.
    - Demi-Flosculeuses .... 800.
    - Flosculeuses .... 834.
    - Radiées .... 874.
- Fleurs indistinctes .... 899.
  - Algues............. 984.
  - Champignons ..... 997.
  - Hypoxylons ...... 1055.
  - Lichens............ 1064.
  - Hépatiques .... 905.
  - Mousses........... 912.
  - Fougères.......... 933.
  - Rhizospermes..... 946.
  - etc.

Modern botanist technology [Bonnier'1904]



1er GROUPE :

✠ Étamines et pétales *réunis aux sépales par leur base.* [En enlevant le calice jusqu'à la base, on enlève en même temps les étamines et les pétales.]

- □ Feuilles *épaisses et charnues,* 6 à 20 pétales, ou même plus.
  - — *Plus de 5 sépales ;* carpelles libres .... *Crassulacées,* p. 110.
  - — *5 sépales ;* carpelles soudés entre eux, sauf au sommet .... FICOÏDÉES, p. 112.
- □ Feuilles *non charnues ;* 4 à 8 pétales rarement plus.
  - ⊙ *Calice d'un rouge vif,* un peu en forme de toupie PU; *pétales d'un rouge vif ;* arbuste à feuilles luisantes, coriaces, simples .... GRANATÉES, p. 103.
  - ⊙ Arbuste odorant, à feuilles persistantes, coriaces, *tout à fait entières et sans stipules ;* fleurs blanches .... MYRTACÉES, p. 107.
  - ⊙ Arbuste *à feuilles dentées opposées et sans stipules, à fleurs blanches très odorantes* .... PHILADELPHÉES, p. 10
  - ⊙ Plante *n'ayant pas ces caractères* .... ROSACÉES, p. 92.

✠ Étamines non réunies aux sépales *et réunies entre elles, au moins à la base, ou disposées par groupes.*

- ★ *Arbre ou arbuste.*
  - • Pédoncule *soudé avec la bractée* TI ; 5 sépales libres, 5 pétales; feuilles molles, non persistantes .... TILIACÉES, p. 54.
  - • Pédoncule *non soudé avec la bractée ;* sépales soudés; 3 à 8 pétales; feuilles assez coriaces, persistant pendant l'hiver .... HESPÉRIDÉES, p. 62.
- ★ Plante *n'étant ni un arbre ni un arbuste.*
  - + Feuilles *opposées,* entières; 3 à 5 styles; étamines disposées par groupes [ex. : H, A] .... HYPÉRICINÉES, p. 59.
  - + Feuilles *alternes.*
    - = Fleurs *en grappe allongée,* [ex. : LL]; pétales *très divisés ;* calice simple .... RÉSÉDACÉES, p. 38.
    - = Fleurs *non en grappe allongée,* souvent disposées à l'aisselle des feuilles; calice souvent double MS, AO .... MALVACÉES, p. 55.

✠ Étamines non réunies aux sépales *et libres entre elles jusqu'à la base.*

- ○ *Plus de 16 pétales.*
  - ⦶ Plante *flottant sur l'eau ou submergée,* à feuilles dont le bord est entier [NL, NA, coupe de fleurs en long] .... NYMPHÉACÉES, p. 11.
  - ⦶ Plante *ni flottante ni submergée ;* feuilles profondément découpées .... *Renonculacées,* p. 2.
- ○ *Moins de 16 pétales.*
  - ⊕ Pistil *à un seul ovaire* et pétales *chiffonnés ou tordus sur eux-mêmes* dans le bouton.
    - △ *4 pétales* [ex. : P, C].
      - ✕ *4 sépales ;* feuilles plus ou moins arrondies, entières CP .... CAPPARIDÉES, p. 34.
      - ✕ *2 sépales* tombant quand la fleur s'ouvre [ex. : PA, C] .... PAPAVÉRACÉES, p. 12.
    - △ *Plus de 4 pétales.*
      - • Feuilles *entières ou faiblement dentées,* souvent opposées .... CISTINÉES, p. 34.
      - • Feuilles *très découpées* .... RENONCULACÉES, p. 2.
  - ⊕ Pistil, en général, à *plusieurs parties* et pétales *non chiffonnés ni tordus* dans le bouton.
    - ⊖ Fleur *n'ayant pas à la fois 3 sépales et 3 pétales* .... RENONCULACÉES, p. 2.
    - ⊖ Fleur ayant *à la fois 3 sépales et 3 pétales* S .... *Alismacées,* p. 290.

Flore complète de la France et de la Suisse, pour trouver facilement les noms de plantes, SANS MOTS TECHNIQUES

The rabbithole

Can we automate the work of building determination keys and alleviate some of their limits?

Output: offline app or even paper version of the key.

Challenges:

- No open source morphological database.
- No formal description of plants.

The rabbithole

Can we automate the work of building determination keys and alleviate some of their limits?

Output: offline app or even paper version of the key.

Challenges:

- No open source morphological database.
- No formal description of plants.
- Avoid over-engineering !
- Participative research to build data.

The ID3 algorithm [Quinlan'86]

In what order should we ask the questions?

The ID3 algorithm [Quinlan'86]

In what order should we ask the questions?

Bayesian algorithm using **information theory**:

1. Start with an **initial probability distribution** d.

The ID3 algorithm [Quinlan'86]

In what order should we ask the questions?

Bayesian algorithm using **information theory**:

1. Start with an **initial probability distribution** d.
2. For every question q:
   - compute the average **information** *after* q.
3. Ask the question with the **largest information**.

The ID3 algorithm [Quinlan'86]

In what order should we ask the questions?

Bayesian algorithm using **information theory**:

1. Start with an **initial probability distribution** d.
2. For every question q:
   - compute the average **information** *after* q.
3. Ask the question with the **largest information**.
4. **Update** d with the user answer and go back to (1).

The ID3 algorithm [Quinlan'86]

In what order should we ask the questions?

Bayesian algorithm using **information theory**:

1. Start with an **initial probability distribution** d.
2. For every question q:
   - compute the average **information** *after* q.
3. Ask the question with the **largest information**.
4. **Update** d with the user answer and go back to (1).

Greedy and non-optimal, but good enough for now.

An aside about information theory [Shannon'48]

Information is usually defined as the opposite of **entropy**:

$$\begin{array}{rrcl} \text{entropy}: & \mathscr{D}(X) & \to & \mathbb{R} \\ & d & \mapsto & -\sum_{x \in X} d(x) \times \log(d(x)) \end{array}$$

An aside about information theory [Shannon'48]

Information is usually defined as the opposite of **entropy**:

$$\begin{array}{rrcl} \text{entropy}: & \mathscr{D}(X) & \to & \mathbb{R} \\ & d & \mapsto & -\sum_{x \in X} d(x) \times \log(d(x)) \end{array}$$

Extremal values:

$$\begin{array}{rcl} \text{entropy}(\text{uniform}(\{1, ..., n\})) & = & \log(n) \\ \text{entropy}(\text{dirac}(x)) & = & 0 \end{array}$$

An aside about information theory [Shannon'48]

Information is usually defined as the opposite of **entropy**:

$$\begin{array}{rrcl} \text{entropy}: & \mathscr{D}(X) & \to & \mathbb{R} \\ & d & \mapsto & -\sum_{x \in X} d(x) \times \log(d(x)) \end{array}$$

Extremal values:

$$\begin{array}{rcl} \text{entropy}(\text{uniform}(\{1, ..., n\})) & = & \log(n) \\ \text{entropy}(\text{dirac}(x)) & = & 0 \end{array}$$

Maximizing information $\leftrightarrow$ **minimizing entropy**

Bayesian Update

Input:

- $d \in \mathscr{D}(\text{Species})$: prior knowledge of what species it may be.
- $o \in \mathrm{Obs}$: user answer ("red flower")

Output:

- $d' \in \mathscr{D}(\text{Species})$: posterior knowledge

Bayesian Update

Input:

- $d \in \mathscr{D}(\text{Species})$: prior knowledge of what species it may be.
- $o \in \mathrm{Obs}$: user answer ("red flower")

Output:

- $d' \in \mathscr{D}(\text{Species})$: posterior knowledge

For $s \in \mathrm{Species}$: (Bayes' Law)

$$\underbrace{d'(s)}_{P(\text{we see } s \mid \text{observing } o)} \propto \underbrace{d(s)}_{P(\text{we see } s)} \times \underbrace{\text{score}(s, o)}_{P(\text{observing } o \mid \text{we see } s)}$$

Input to the ID3 algorithm

```
(* Representation of the description of a species *)
(* e.g. ``white flowers, simple lanceolate leaves'' *)
type species

(* Representation of observation (user answers). *)
type observation

(* Given an observation, how likely is
   it to be a particular species ? *)
val score : species -> observation -> float
```

How to describe species and observation ?

$\rightsquigarrow$ `score` tangles species and observation.

General shape of model

We distinguish two things:

- **Plant description**, which are certain and exhaustive
  “Tree with white flowers, large leaves ...“
- **Observation**, which are uncertain
  “Tree with whiteish flowers (maybe rose), not sure about leaves (its winter)”

General shape of model

We distinguish two things:

- `Plant`: **Plant description**, which are certain and exhaustive
  "Tree with white flowers, large leaves ..."
- `Obs`: **Observation**, which are uncertain
  "Tree with whiteish flowers (maybe rose), not sure about leaves (its winter)"

We let $\texttt{Obs} = \mathscr{D}(\texttt{Plant})$.

General shape of model

We distinguish two things:

- `Plant`: **Plant description**, which are certain and exhaustive
  "Tree with white flowers, large leaves ..."
- `Obs`: **Observation**, which are uncertain
  "Tree with whiteish flowers (maybe rose), not sure about leaves (its winter)"

We let $\mathtt{Obs} = \mathscr{D}(\mathtt{Plant})$.

$$P(\text{observing } o \mid \text{we see } s)$$

General shape of model

We distinguish two things:

- `Plant`: **Plant description**, which are certain and exhaustive
  "Tree with white flowers, large leaves ..."
- `Obs`: **Observation**, which are uncertain
  "Tree with whiteish flowers (maybe rose), not sure about leaves (its winter)"

We let $\texttt{Obs} = \mathscr{D}(\texttt{Plant})$.

$$\begin{aligned} & P(\text{observing } o \mid \text{we see } s) \\ = \quad & \textstyle\sum_{p \in \text{Plant}} o(p) \times P(\text{observed plant is } p \mid \text{we see } s) \end{aligned}$$

General shape of model

We distinguish two things:

- `Plant`: **Plant description**, which are certain and exhaustive
  "Tree with white flowers, large leaves ..."
- `Obs`: **Observation**, which are uncertain
  "Tree with whiteish flowers (maybe rose), not sure about leaves (its winter)"

We let $\texttt{Obs} = \mathscr{D}(\texttt{Plant})$.

$$
\begin{aligned}
& P(\text{observing } o \mid \text{we see } s) \\
= \; & \textstyle\sum_{p \in \text{Plant}} o(p) \times \underbrace{P(\text{observed plant is } p \mid \text{we see } s)}_{s(p)}
\end{aligned}
$$

Thus we can also let $\text{Species} = \mathscr{D}(\text{Plant})$!

General shape of model

We distinguish two things:

- `Plant`: **Plant description**, which are certain and exhaustive
  "Tree with white flowers, large leaves ..."
- `Obs`: **Observation**, which are uncertain
  "Tree with whiteish flowers (maybe rose), not sure about leaves (its winter)"

We let $\texttt{Obs} = \mathscr{D}(\texttt{Plant})$.

$$\begin{aligned} & P(\text{observing } o \mid \text{we see } s) \\ = \; & \textstyle\sum_{p \in \text{Plant}} o(p) \times \underbrace{P(\text{observed plant is } p \mid \text{we see } s)}_{s(p)} \end{aligned}$$

Thus we can also let $\text{Species} = \mathscr{D}(\text{Plant})$!

An example

```
type color = Red | Blue | White | Rose
type plant = { flower_color: color }
```

An example

```
type color = Red | Blue | White | Rose
type plant = { flower_color: color }

type observation = plant distribution
let whiteish : observation = fun p ->
  match p.flower_color with
  | White -> 0.8
  | Rose -> 0.2
  | _ -> 0.
```

An example

```
type color = Red | Blue | White | Rose
type plant = { flower_color: color }

type observation = plant distribution
let whiteish : observation = fun p ->
  match p.flower_color with
  | White -> 0.8
  | Rose -> 0.2
  | _ -> 0.

type species = plant distribution
let laurier_rose : species = fun p ->
  match p.flower_color with
  | Rose -> 0.8
  | White -> 0.2 (* cultivars *)
  | _ -> 0.0
```

Where are we ?

We have all the conceptual ingredients to make it work:

- The decision tree **algorithm**
- Species and observation with **uncertainty**
- How to make species and observation **interact** (`score`)

Where are we ?

We have all the conceptual ingredients to make it work:

- The decision tree **algorithm**
- Species and observation with **uncertainty**
- How to make species and observation **interact** (`score`)

However:

- What is the real type for `plant` ?
- Where do we get the data?
- If `plant` is big, isn't `score` intractable ?

The project in a nutshell

1. How to have expert botanists write the type `plant` ?

2. How to describe the distribution probabilities for species ?
   - User-friendlyness: we need a lot of data, hence of a lot of people!
   - Link formal/informal: bibliographical info linked to data

The Flat Model – `Plant`

The naive approach:

$$\mathsf{Plant} = \prod_{i \in I} C_i$$

where:

- I is the set of **characters**
- Each C_i is a simple sum $1 + \ldots + 1$

The Flat Model – `Plant`

The naive approach:

$$\mathsf{Plant} = \prod_{i \in I} C_i$$

where:

- I is the set of **characters**
- Each C_i is a simple sum $1 + \ldots + 1$

Can be described by a textual format:

```
flower-color = [ red blue white rose ... ];
flower-petal-number = [ 1 2 3 ... ];
leaf-structure = [ simple divided ];
```

The Flat Model – `Species`

Distributions over Plant are inconvenient.
However, we have a correspondance:

$$\mathscr{D}(S \times T) \rightleftarrows \mathscr{D}(S) \times \mathscr{D}(T)$$

The Flat Model – `Species`

Distributions over Plant are inconvenient.
However, we have a correspondance:

$$\mathscr{D}(S \times T) \rightleftarrows \mathscr{D}(S) \times \mathscr{D}(T)$$

Thus we use

$$\mathscr{D}(\mathsf{Plant}) \rightleftarrows \prod_{i \in I} \mathscr{D}(C_i)$$

The Flat Model – Species

Distributions over Plant are inconvenient.
However, we have a correspondance:

$$\mathscr{D}(S \times T) \rightleftarrows \mathscr{D}(S) \times \mathscr{D}(T)$$

Thus we use

$$\mathscr{D}(\mathsf{Plant}) \rightleftarrows \prod_{i \in I} \mathscr{D}(C_i)$$

```
# laurier-rose.species
flower-color = [ white: 0.8 rose: 0.2 ];
# we can omit the rest and use uniform distribution
```

The Flat Model – Species

Distributions over Plant are inconvenient.
However, we have a correspondance:

$$\mathscr{D}(S \times T) \rightleftarrows \mathscr{D}(S) \times \mathscr{D}(T)$$

Thus we use

$$\mathscr{D}(\text{Plant}) \rightleftarrows \prod_{i \in I} \mathscr{D}(C_i)$$

```
# laurier-rose.species
flower-color = [ white: 0.8 rose: 0.2 ];
# we can omit the rest and use uniform distribution
```

With this representation, `score` can be computed in time $O(C_1 + \ldots + C_n)$ instead of $O(C_1 \times \ldots \times C_n)$!

Limits of the flat model

1. `Plant` is too **rigid**: what about plants without flowers?

2. `Species` cannot represent all probability distributions.
   No **correlation** between trait distributions.

3. Compositional structure on species?
   Merge different descriptions ...

Going full algebraic type

What if we allow `Plant` to be an algebraic type?

```
plant := {
  leaf;
  flower;
}.
leaf := {
     position: [ base | stem {disposition} ];
     venation;
     attachment;
}.
flower := {
  inflorescence;
  sex: [ unisexual | hermaphrodism ];
  color: [ red | blue | white | rose ];
}.
```

$\rightsquigarrow$ Similar to ontologies (e.g. RDFS).

What about Species?

We need to extend our **abstraction**:

$$\mathscr{D}(S \times T) \rightleftarrows \mathscr{D}(S) \times D(T)$$

to:

$$\mathscr{D}(S \oplus T) \rightleftarrows$$

What about Species?

We need to extend our **abstraction**:

$$\mathscr{D}(S \times T) \rightleftarrows \mathscr{D}(S) \times D(T)$$

to:

$$\mathscr{D}(S \oplus T) \rightleftarrows [0,1] \times \mathscr{D}(S) \times \mathscr{D}(T)$$

Example of a distribution:

```
{ leaf = { position = stem };
  flower = { color = [ 0.8: rose | 0.2: white ] }
}
```

What about Species?

We need to extend our **abstraction**:

$$\mathscr{D}(S \times T) \rightleftarrows \mathscr{D}(S) \times D(T)$$

to:

$$\mathscr{D}(S \oplus T) \rightleftarrows [0,1] \times \mathscr{D}(S) \times \mathscr{D}(T)$$

Example of a distribution:

```
{ leaf = { position = stem };
  flower = { color = [ 0.8: rose | 0.2: white ] }
}
```

Abstract type interpretation:

$$[\![S \times T]\!] = [\![S]\!] \times [\![T]\!] \qquad [\![S + T]\!] = [0,1] \times [\![S]\!] \times [\![T]\!].$$

Biological species model of Linear Logic

We cannot still express **polymorphism**:

simple basal leaves, compound stem leaves

Biological species model of Linear Logic

We cannot still express **polymorphism**:

simple basal leaves, compound stem leaves

$\rightsquigarrow$ Can only say: *simple or compound*, *basal or compound*.

Biological species model of Linear Logic

We cannot still express **polymorphism**:

simple basal leaves, compound stem leaves

$\rightsquigarrow$ Can only say: *simple or compound*, *basal or compound*.We add an exponential $!S$ to the type language:

$$[\![!S]\!] = \mathscr{D}(S)$$

We can thus write:

```
[ 0.5 { leaf = { simple ; basal } }
| 0.5 { leaf = { compound; stem } } ]
```

In the plant description, we can add modalities:

```
plant = { leaf!; flower!; stem }
```

Conclusion

We have a working prototype for the flat model:

- An editor for entering species distribution
- Greedy algorithm implementation

Conclusion

We have a working prototype for the flat model:

- An editor for entering species distribution
- Greedy algorithm implementation

Inria exploratory project *Back to the trees*:

- Extend the model
- Work with local associations to build a database
- Implement non-greedy algorithms ?


WE NEED
YOU