Plant genetics meets text mining at the SeeDev challenge

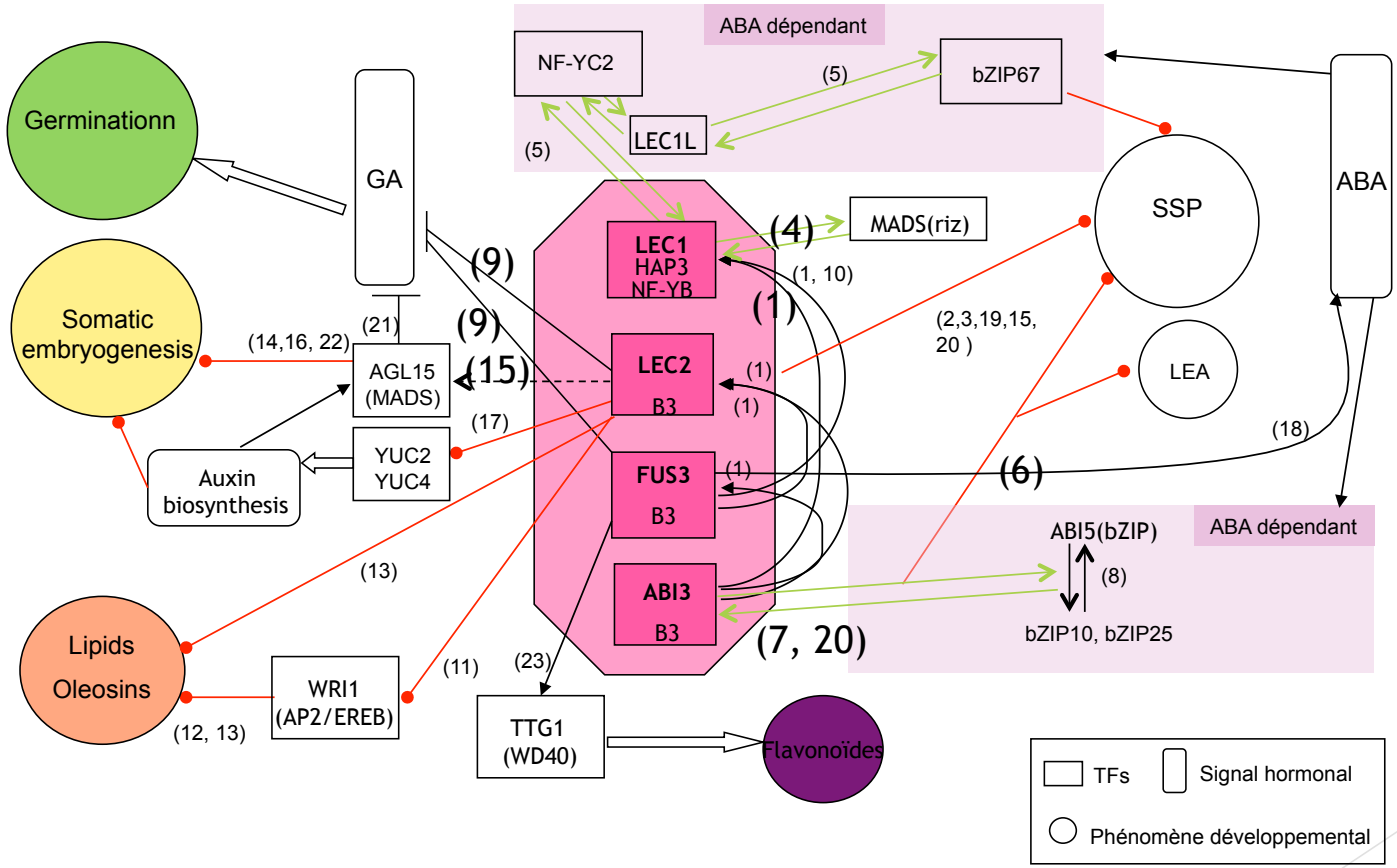
Bertrand Dubreucq & Robert Bossy

Seed maturation and reserves accumulation

a major challenge to improve the components of yield, nutritional value and industrial value of the seed.

Understand the limiting factors of accumulation and quality of reserves.

Seed Development involves many regulatory element:



Objectifs:

Reconstruction of regulatory networks including

- - the genetic and molecular levels,
- - environmental factors
- - the associated phenotypes.

In a systems biology approach, that is, in an explanatory and predictive model.

Using the data existing in the litterature and Text Data Mining (TDM).

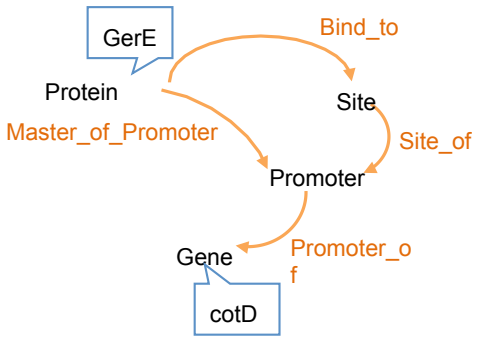
Objectif : from text to network

*[..] We show that **GerE** binds to two **sites** that span the -35 region of the **cotD** **promoter** [..]*

Objectif : from text to network

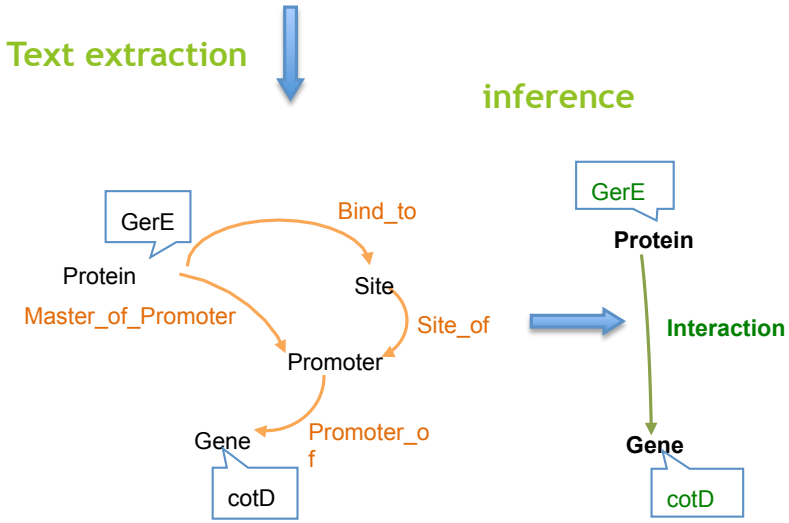
*[..] We show that **GerE** binds to two **sites** that span the -35 region of the **cotD** **promoter** [..]*

Text extraction



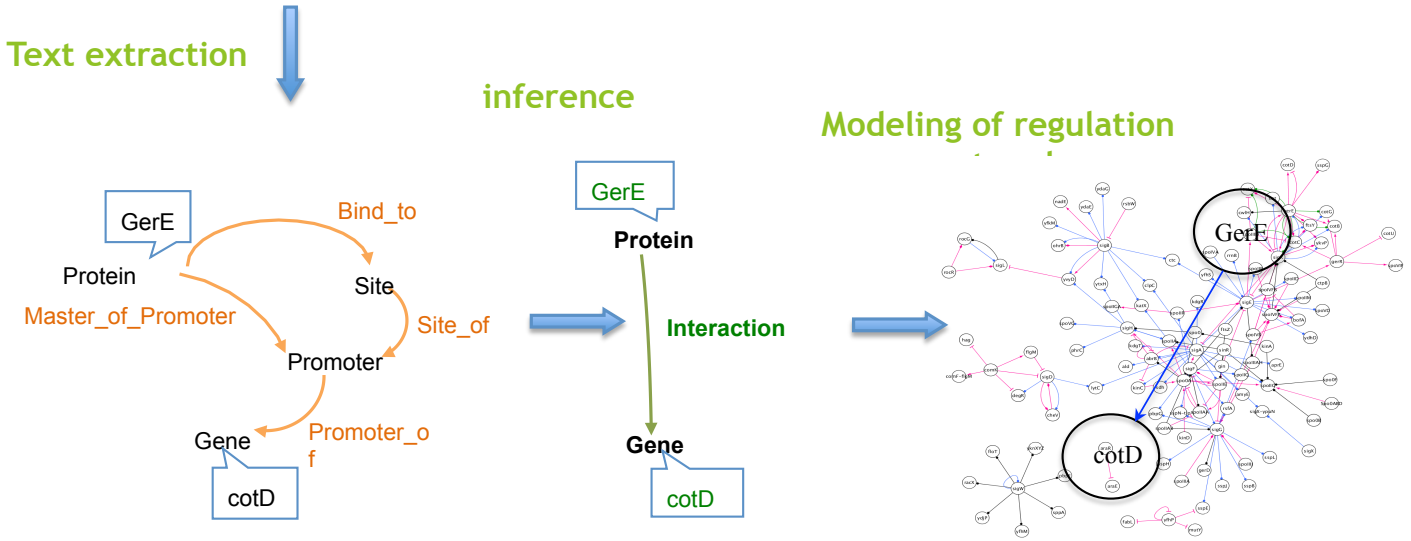
Objectif : from text to network

*[..] We show that **GerE** binds to two **sites** that span the -35 region of the **cotD** **promoter** [..]*

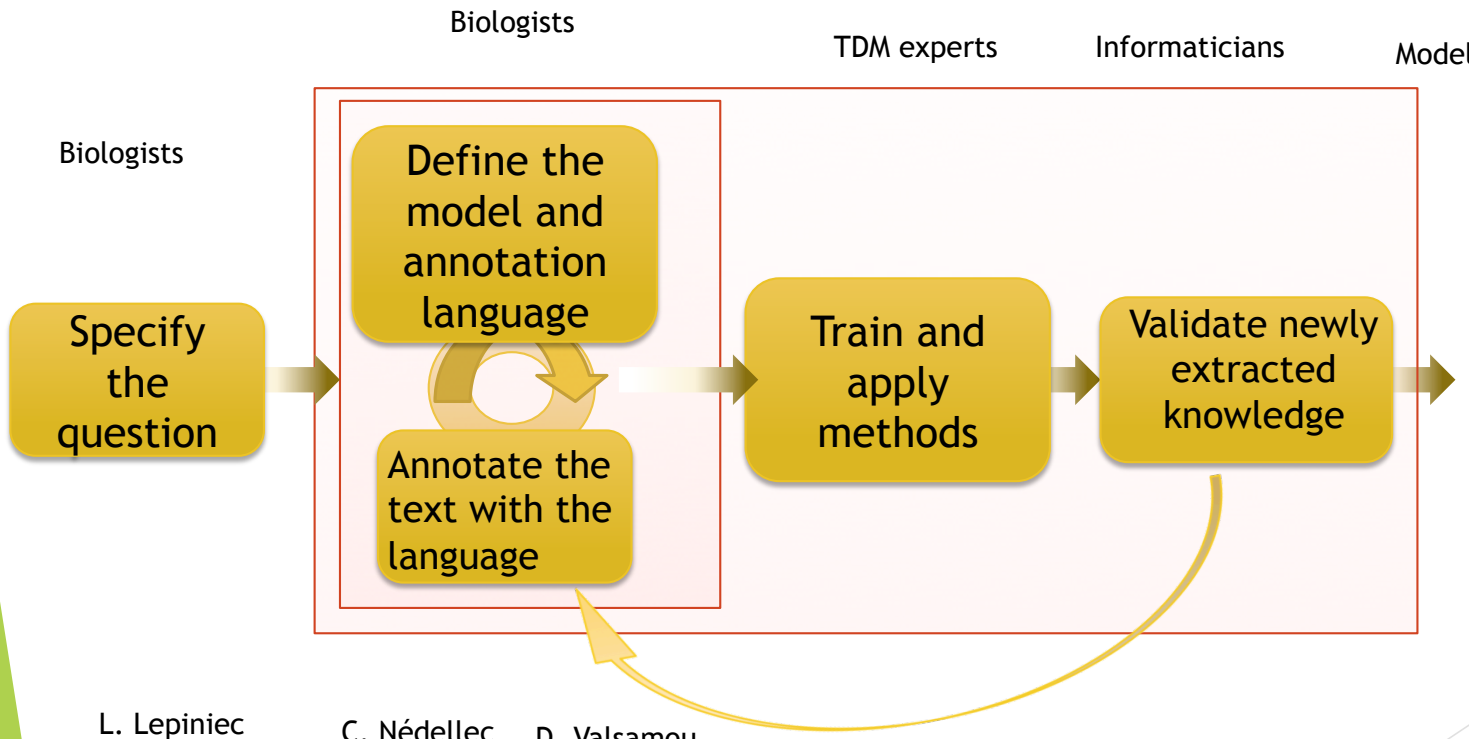


Objectif : from text to network

*[..] We show that **GerE** binds to two **sites** that span the -35 region of the **cotD** **promoter** [..]*



Mutlidisciplinary approach:



L. Lepiniec
A. Fatihi
B. Dubreucq

C. Nédellec
R. Bossy
E. Chaix

D. Valsamou
P. Zweigenbaum
P. Bessierres

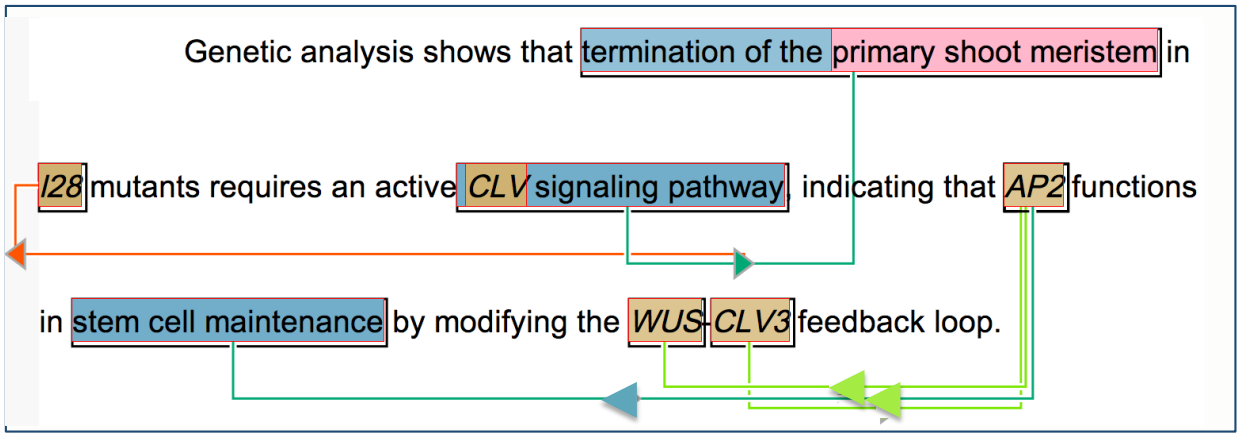
Annotation language: using the Alvisae editor

Genetic analysis shows that termination of the primary shoot meristem in

I28 mutants requires an active CLV signaling patway, indicating that AP2 functions

In stem cell maintenance by modifying the WUS-CLV3 feedback loop.

Annotation language: using the Alvisae editor



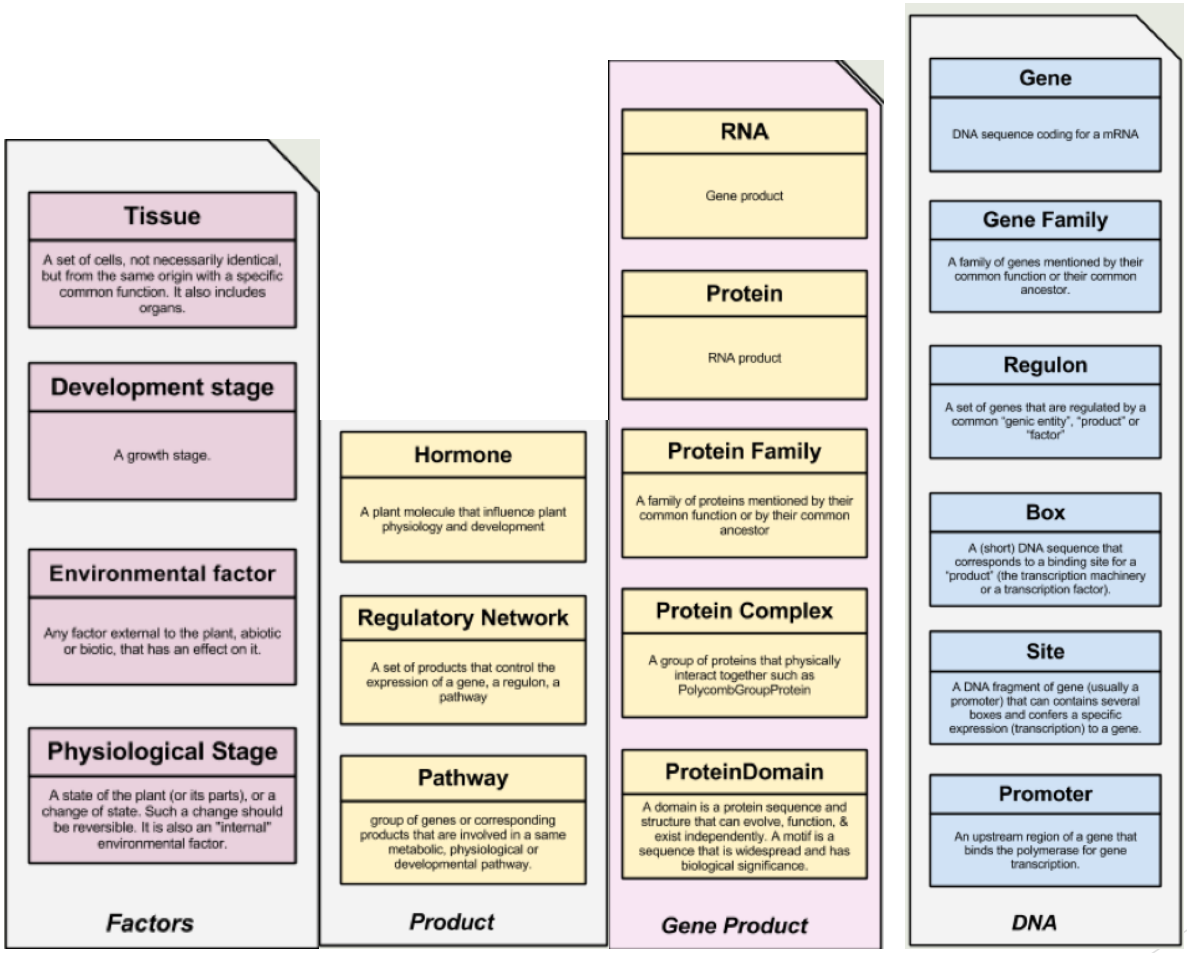
Entities

Genes.
Tissues.
Metabolic pathway
Regulatory network

Relations

Agent **(Gene)** Regulates activity of Target **(metabolic pathway).**
Agent **(Gene)** Regulates expression of Target **(Gene).**
…

Ath annotation model: definition of the entities



Ath annotation model: definition of the relations.

Types of relations

10 relation types were defined :

→Regulation :

- *RegulatesActivityOf*
- *RegulatesAccumulationOf*
- *RegulatesExpressionOf*

→Interaction :

- *InteractWith*
- *BindTo*

→Localisation :

- *IsFoundIn*
- *IsFoundDuring*

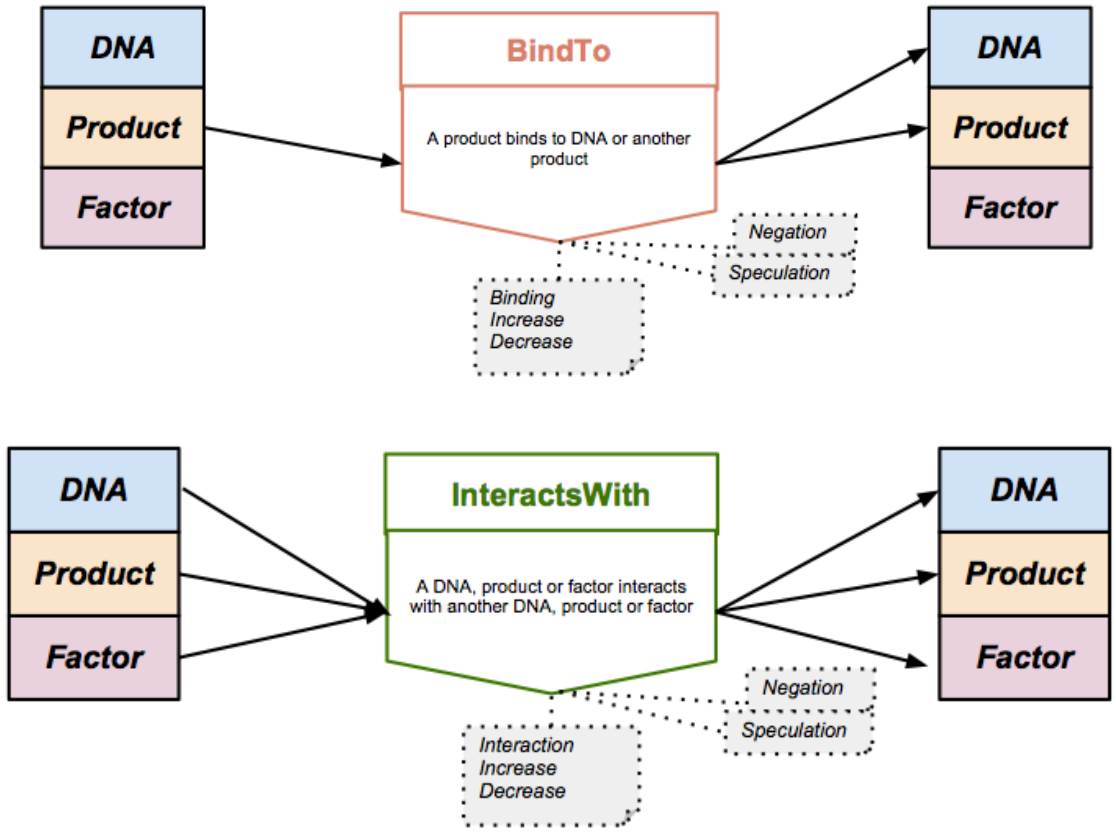
→Similarity :

- *Comparison*
- *Belongs to*
- *Encodes*

1 relation to define n-ary events

- *Condition*

Integration of entities and relations:



Ath annotation model: definition of the relations.

Event type	Description	Arguments
AccumulatesIn	*A product* accumulates in a given *factor (environment)*	Agent : [*Product*] Target : [*Factor*]
BindTo	*A product* binds to DNA or another *product*	Agent : [*product*] Target : [*DNA* | *product*]
BelongsTo	A DNA, product or factor belongs to another DNA, product or factor	Agent : [*DNA* | *product* | *factor*] Target [*DNA* | *product* | *factor*]
InteractWith	A DNA, product or factor interacts with another DNA, product or factor	Agent : [*DNA* | *product* | *factor*] Target: [*DNA* | *product* | *factor*]
RegExpression (activates, inhibits, not)	A DNA, product or factor regulates the activity of a DNA entity	Agent : [*DNA* | *product* | *factor*] Target : [*DNA*]
RegAccumulation (activates, inhibits, not)	A DNA, product or factor regulates the accumulation of a product or a factor	Agent : [*DNA* | *product* | *factor*] Target : [*product* | *factor*]
RegActivity (activates, inhibits, not)	A product or a factor regulates the activity of a product or a factor	Agent : [*product* | *factor*] Target : [*product* | *factor*]

Instructions, *exemples*

Event types	**A few examples illustrating all the possible relationships**
AccumulatesIn	*FUS3* mRNA accumulates in seed FUS3 (protein) accumulates in torpedo embryo ABA (hormone) accumulates in seed during maturation The AFL (protein complex) accumulates in embryo
BindTo	FUS3 (protein) binds p*At2S3* (promoter) FUS3 (protein) binds RY element (box) WDR (protein complex) binds *pBANYULS* (promoter) FUS3 (protein) binds LEC2 (protein) TT2 (protein) binds TT8-TTG1 (protein complex) ABP (protein) Binds Auxin (hormones) PCG (protein complex) bind pLEC2 (promoter) ABA (hormone) bind to ABP (protein)
BelongsTo	*FUSCA3* (gene) belongs to the B3 (gene family) *RY* element (box) belongs to *pAt2S3* (promoteur) FUSCA3 (protein) belongs to AFL (regulatory network) AKIN10 (protein) belongs to Kinases (protein family) The endothelium (tissue) belongs to the seed coat (tissue)

Guidelines

A.1.2 Gene product entities

RNA : gene product

Protein : RNA product

ProteinFamily : a family of proteins mentioned by their common function or by their common ancestor (ex. **MADS**-box, MYB, bHLH, bZIP transcription factor families).

ProteinComplex: a group of proteins that physically interact together such as PolycombGroupProtein (PCG), MYB-bHLH-WD40 (MBW),…

Hormone : is a plant molecule that influence plant physiology and development, They include auxin, ethylene, abscisic acid (ABA), GA Gibberellic ac (GA), Cytokinin (CK), (jasmonate/jasmonic acid / JA), strigolactone, Brassinosteroid, Salicylic Acid (SA), Nitric oxide (NO)

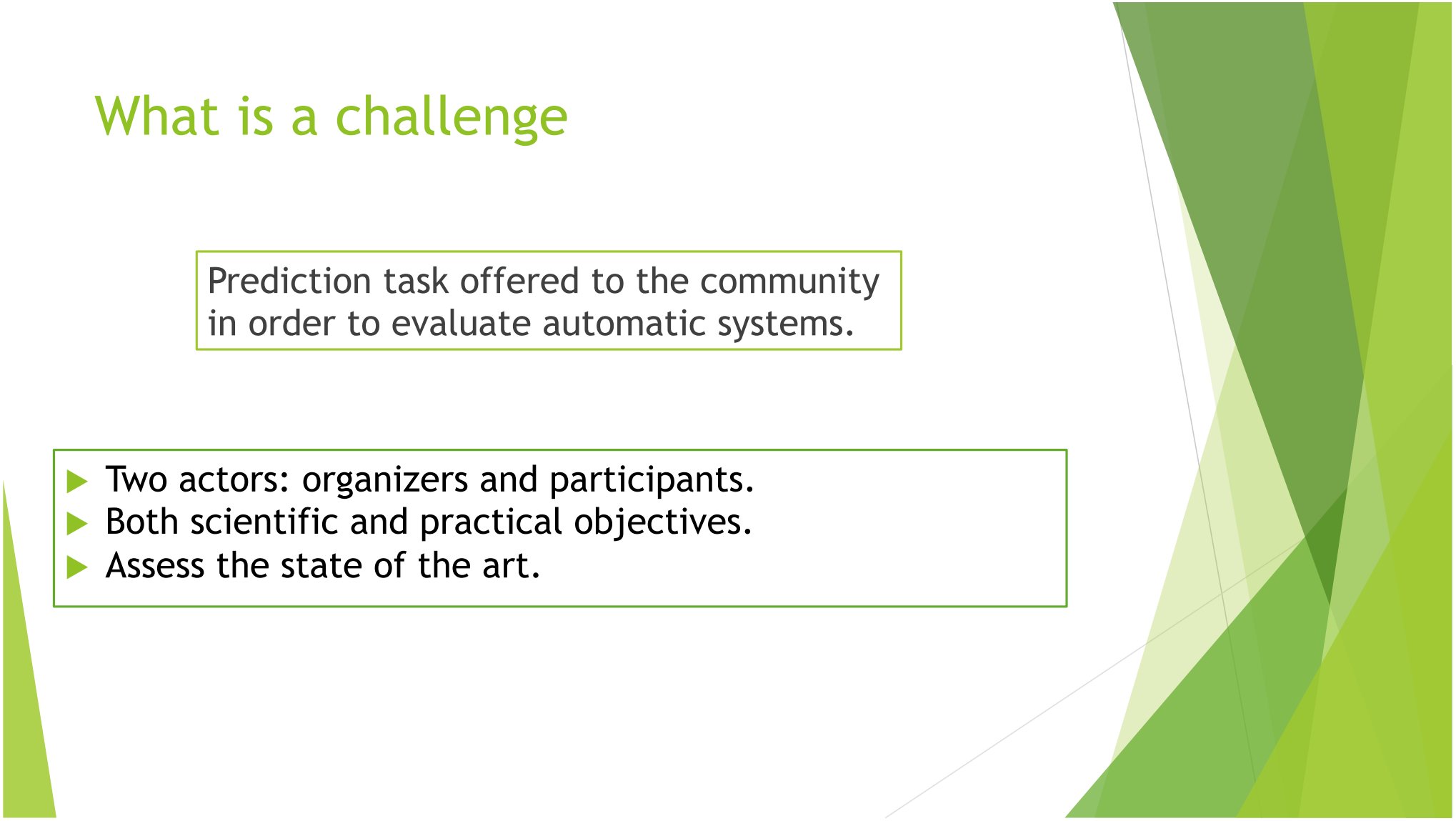
RegulatoryNetwork; a set of products that control the expression of a gene, a regulon, a pathway

Pathway: group of genes or corresponding products that are involved in a same metabolic, physiological or developmental pathway. Examples of metabolic pathways, lipid/triacylglycerol/oil, fatty acids, storage proteins, starch, carbohydrate, secondary metabolites, flavonoid, tannin. Une nomenclature complète existe sur le serveur TAIR à l'adresse ftp://ftp.arabidopsis.org/home/tair/Pathways/

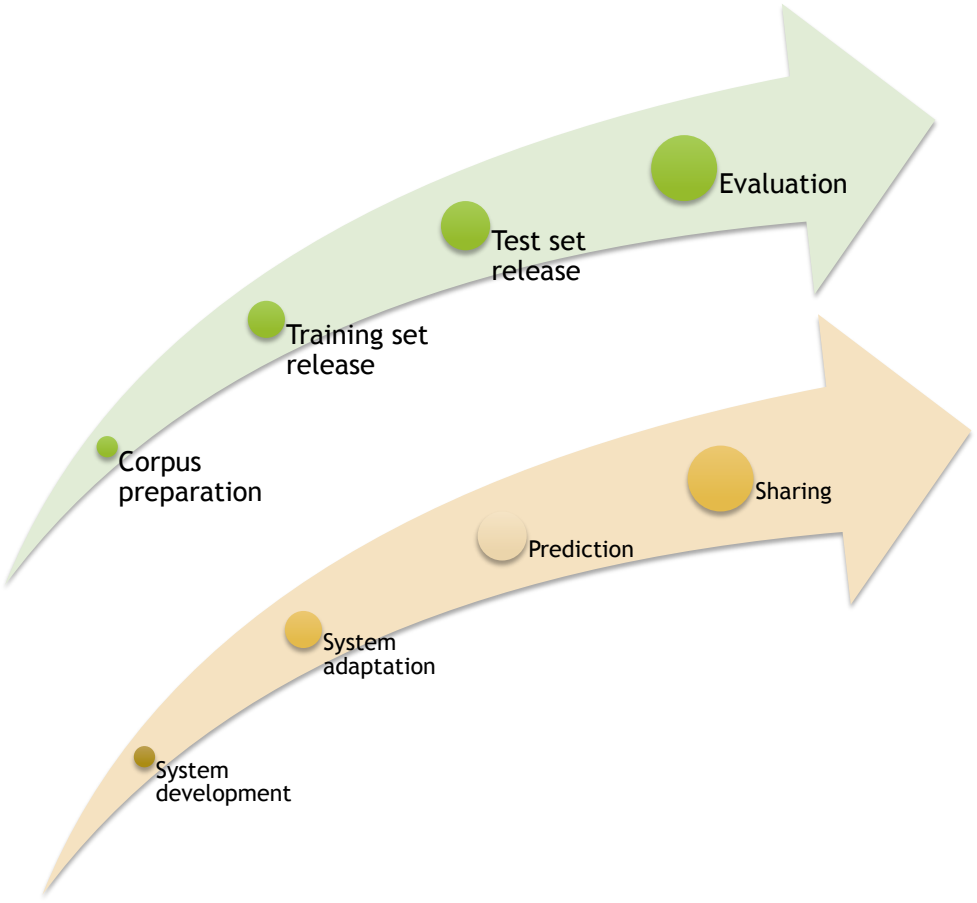
What is a challenge

Prediction task offered to the community in order to evaluate automatic systems.

- Two actors: organizers and participants.
- Both scientific and practical objectives.
- Assess the state of the art.



The schedule of a challenge



Workshop & Papers

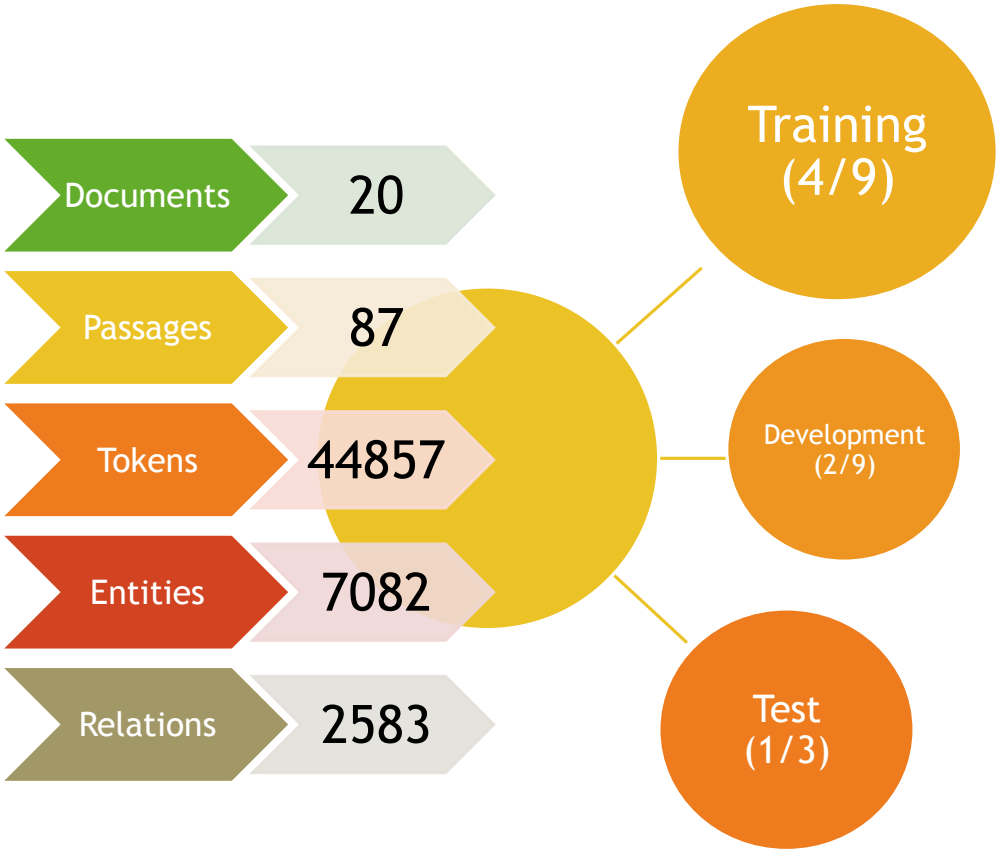
BioNLP-ST 2016
(ACL, Berlin, 2016)

BioNLP-OST 2019
(EMNLP, Hong-Kong, 2019)

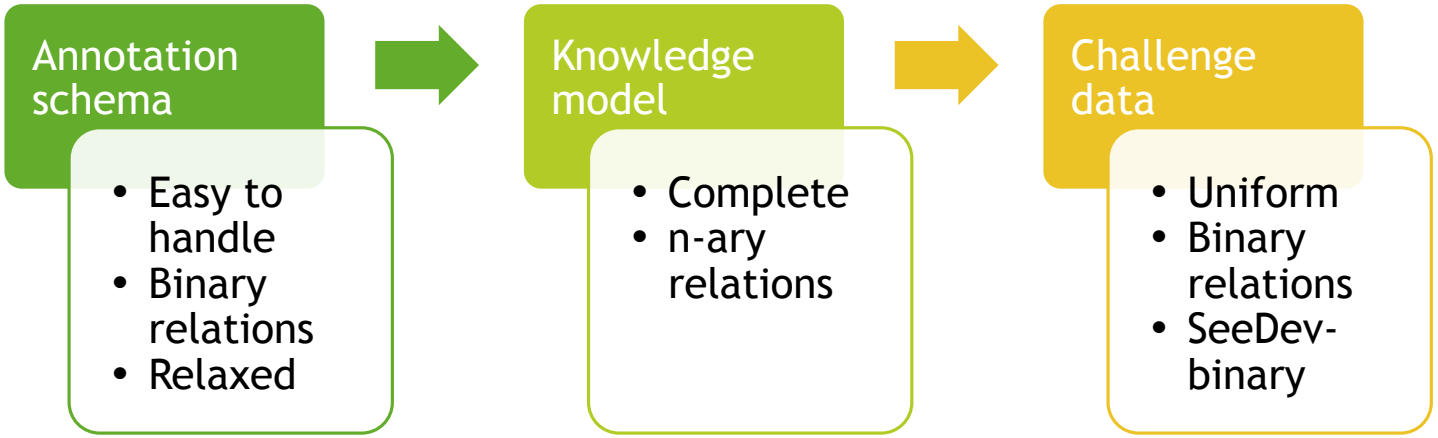
Our strategy

- Medical topics dominate the challenge landscape.
- Increase awareness on INRA's topics.
- Try out automatic systems genericity.

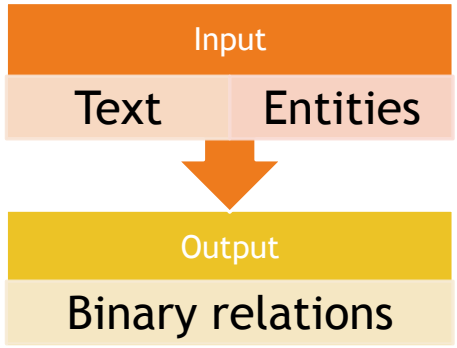
Corpus statistics & Split

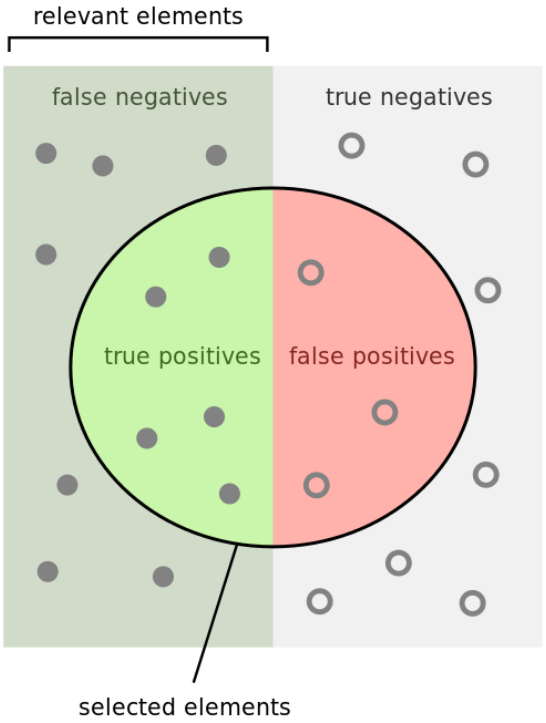


Three data models

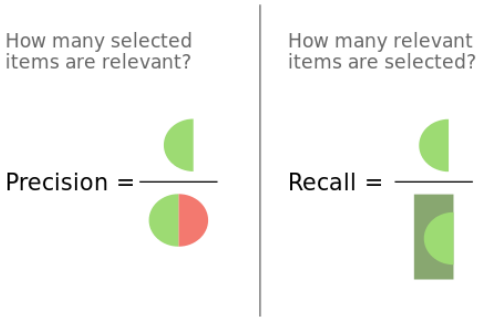


Evaluation

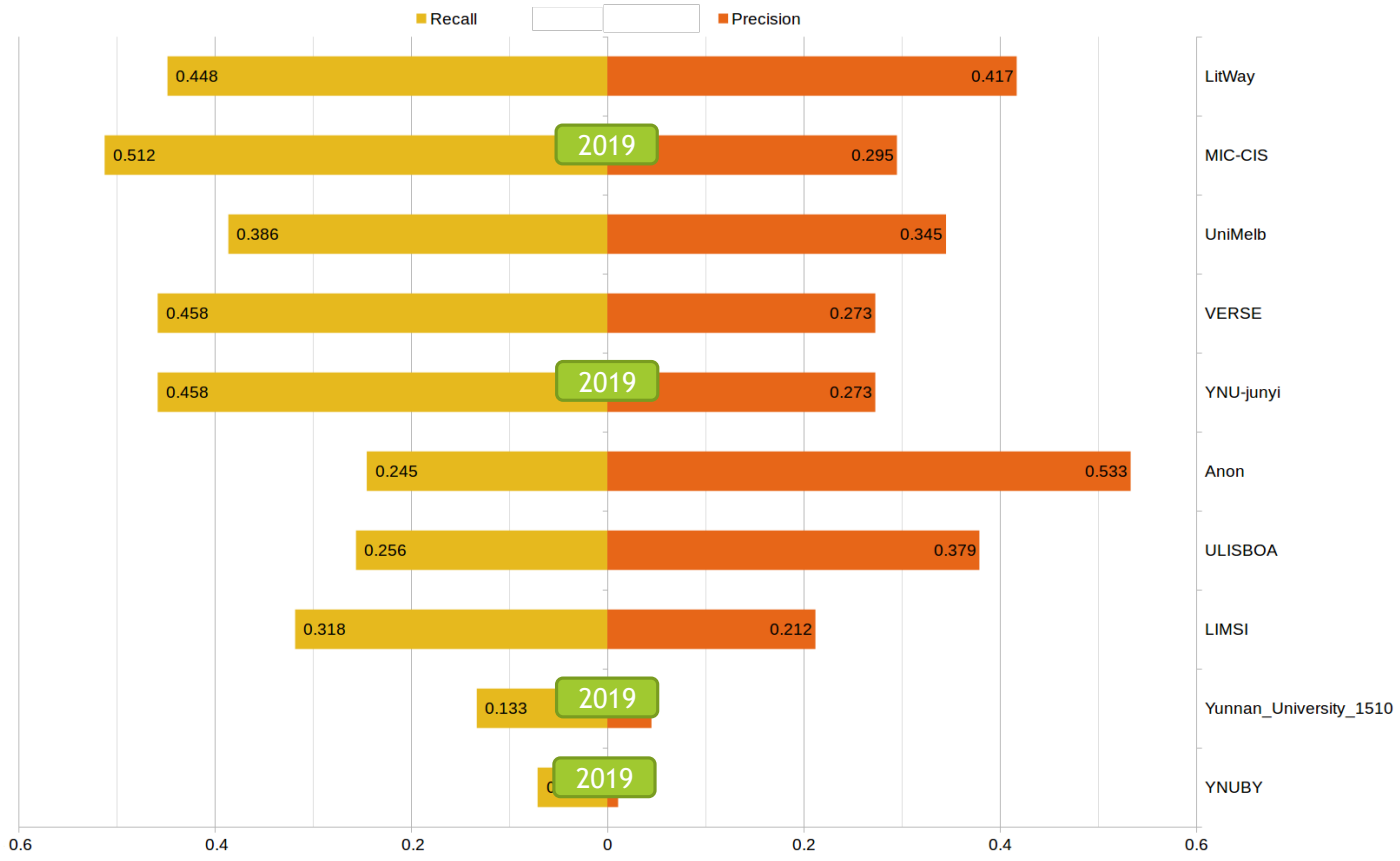




$$F_1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$



Results


Recall
Precision
0.448
0.417
LitWay
2019
0.512
0.295
MIC-CIS
0.386
0.345
UniMelb
0.458
0.273
VERSE
2019
0.458
0.273
YNU-junyi
0.245
0.533
Anon
0.256
0.379
ULISBOA
0.318
0.212
LIMSI
2019
0.133
Yunnan_University_1510
2019
YNUBY
0.6
0.4
0.2
0
0.2
0.4
0.6

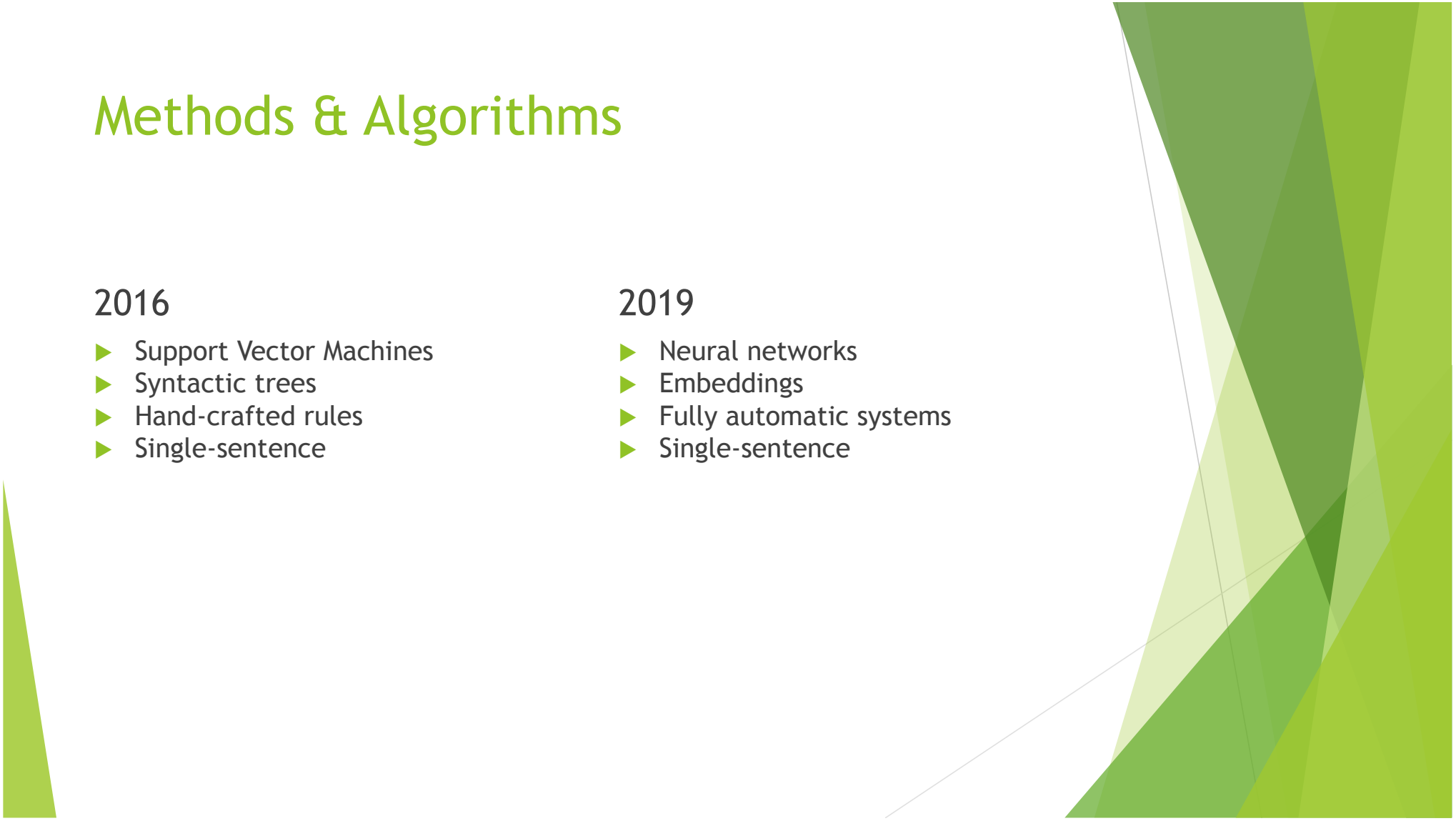
Methods & Algorithms

2016

- Support Vector Machines
- Syntactic trees
- Hand-crafted rules
- Single-sentence

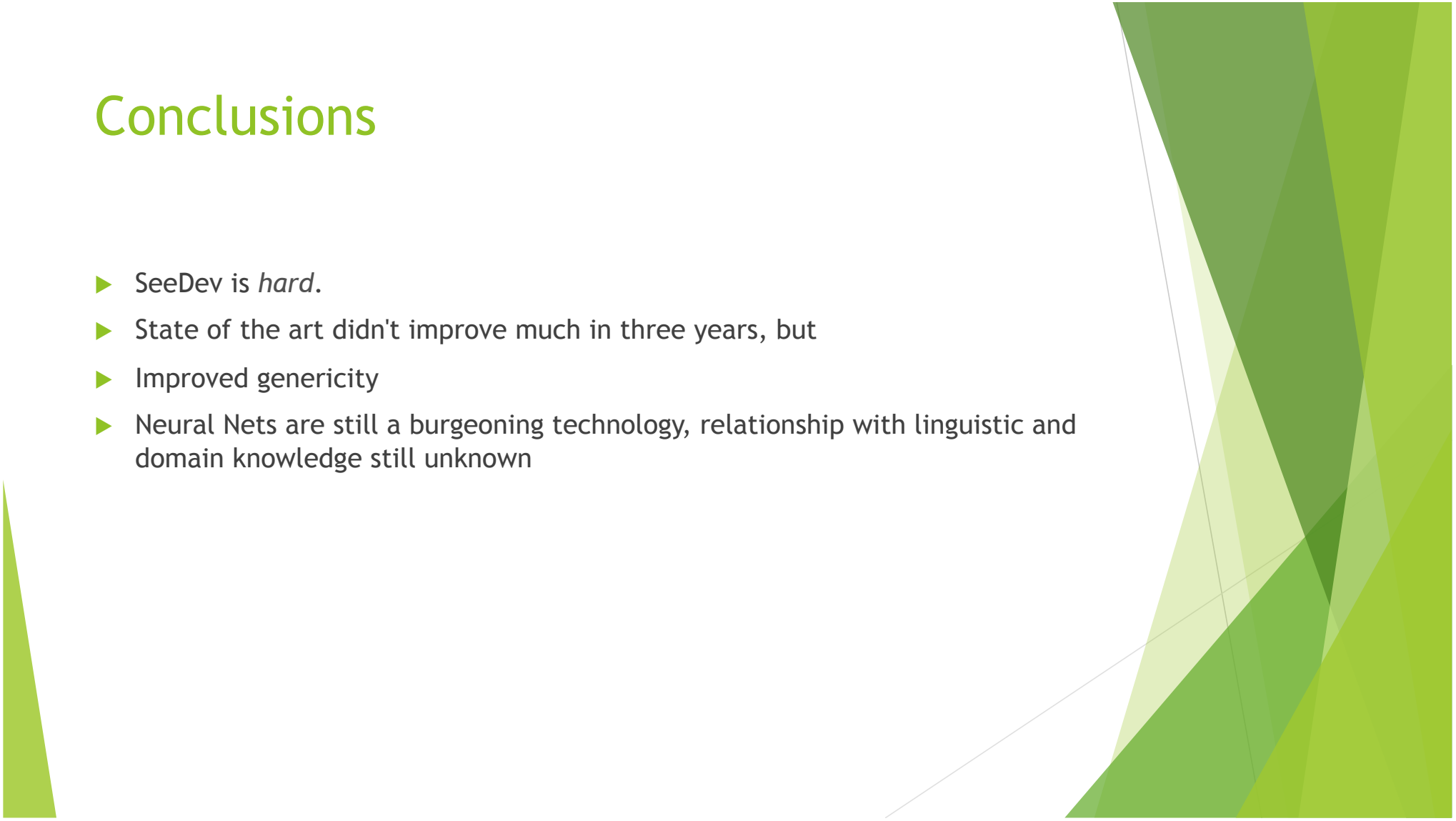
2019

- Neural networks
- Embeddings
- Fully automatic systems
- Single-sentence



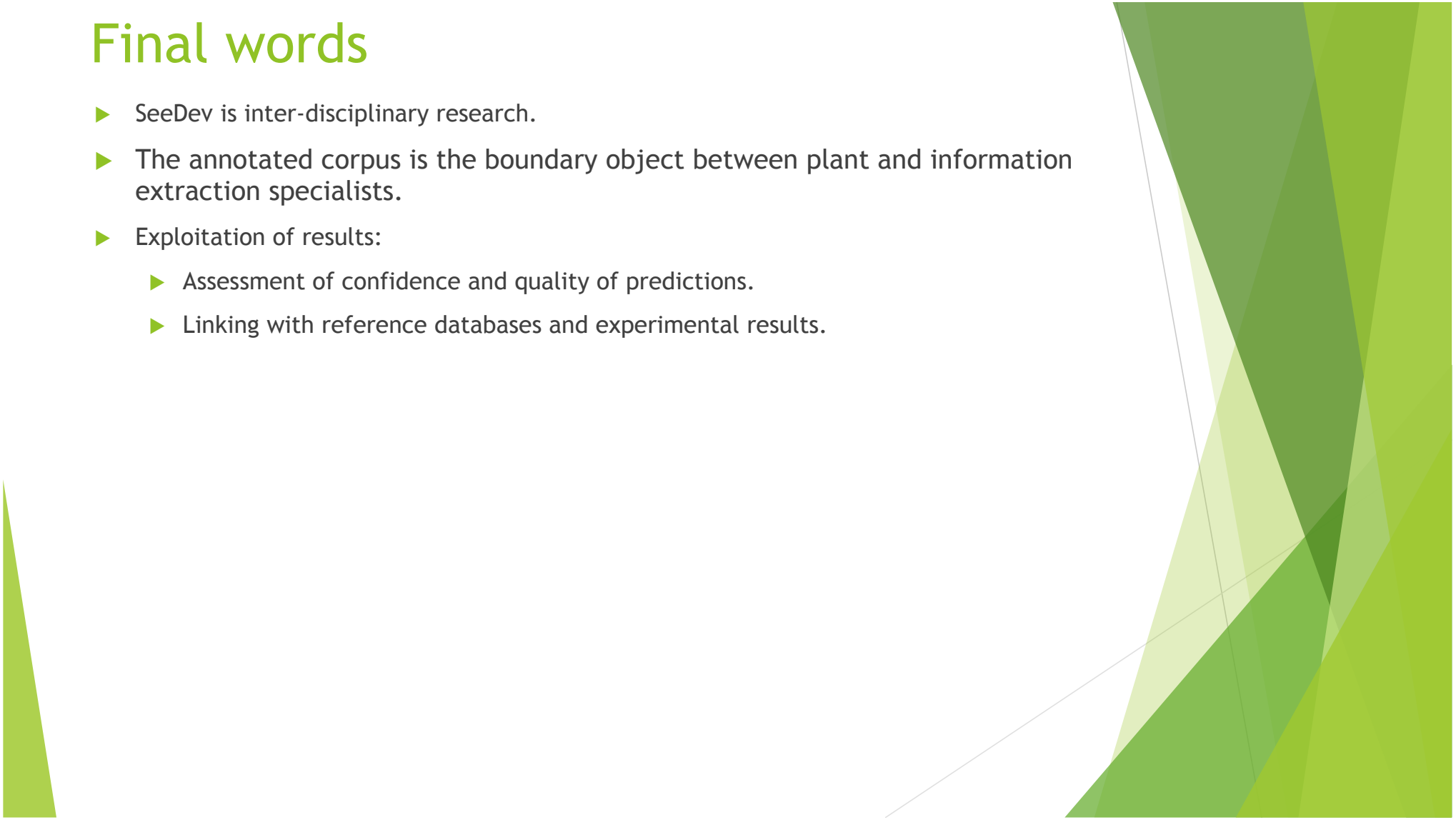
Conclusions

- SeeDev is *hard*.
- State of the art didn't improve much in three years, but
- Improved genericity
- Neural Nets are still a burgeoning technology, relationship with linguistic and domain knowledge still unknown



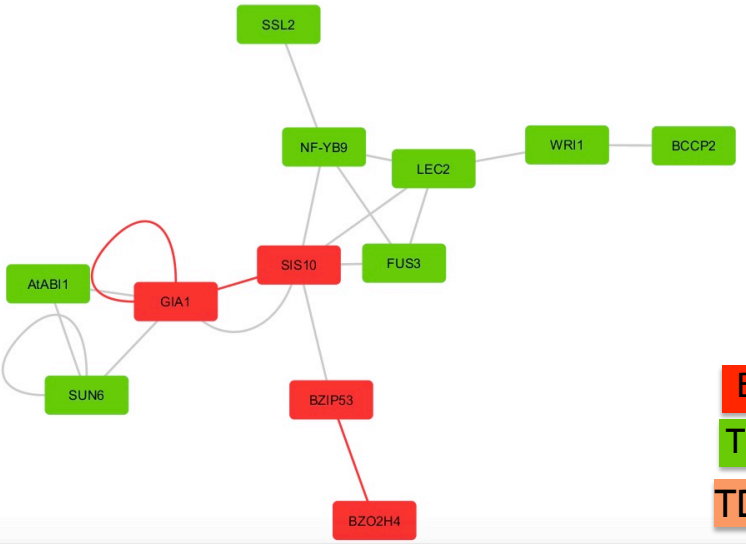
Final words

- SeeDev is inter-disciplinary research.
- The annotated corpus is the boundary object between plant and information extraction specialists.
- Exploitation of results:
  - Assessment of confidence and quality of predictions.
  - Linking with reference databases and experimental results.



Final words

- SeeDev is inter-disciplinary research.
- The annotated corpus is the boundary object between plant and information extraction specialists.
- Exploitation of results:
  - Assessment of confidence and quality of predictions.
  - Linking with reference databases and experimental results.



Biogrid Database

TDM, manual annotatio

TDM, predicted annotation