Applying Machine Learning to Plant Literature: Augmenting Human Curation

Tanya Berardini, Ph.D

TAIR/Phoenix Bioinformatics

Problem: Data isolation

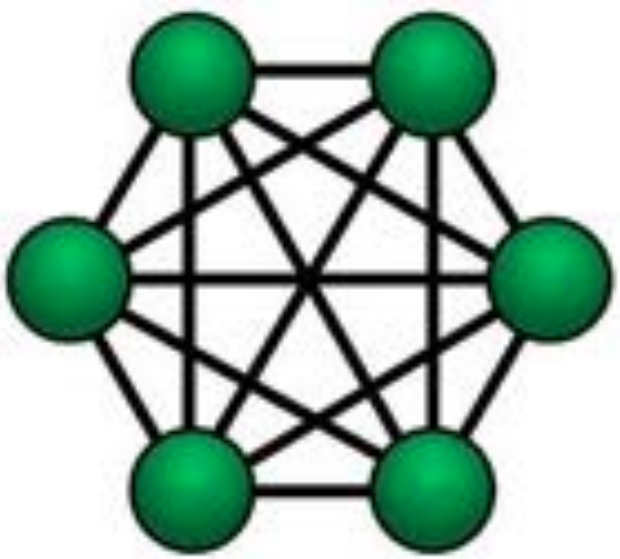
What do we want to know?

- What is known about my gene or gene set of interest?
- Which genes lack information?

Solution: Capture data in a structured way and interconnect them

Function

Phenotype



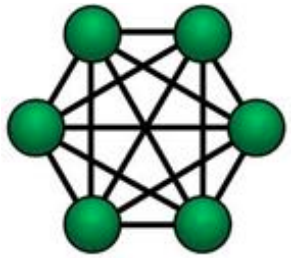
Sequence

Gene

Allele

Paper

Benefits of structured and interconnected data



- Answer questions more easily.
  - What is known about my gene or gene set of interest?
  - Which genes lack information?
  - What are the functions of the genes in my newly sequenced genome?
- structured data, completeness of the answers.

TAIR's manual literature curation structures data

What is a Gene Ontology (GO) annotation?

gene

method

publication

Gene Ontology term

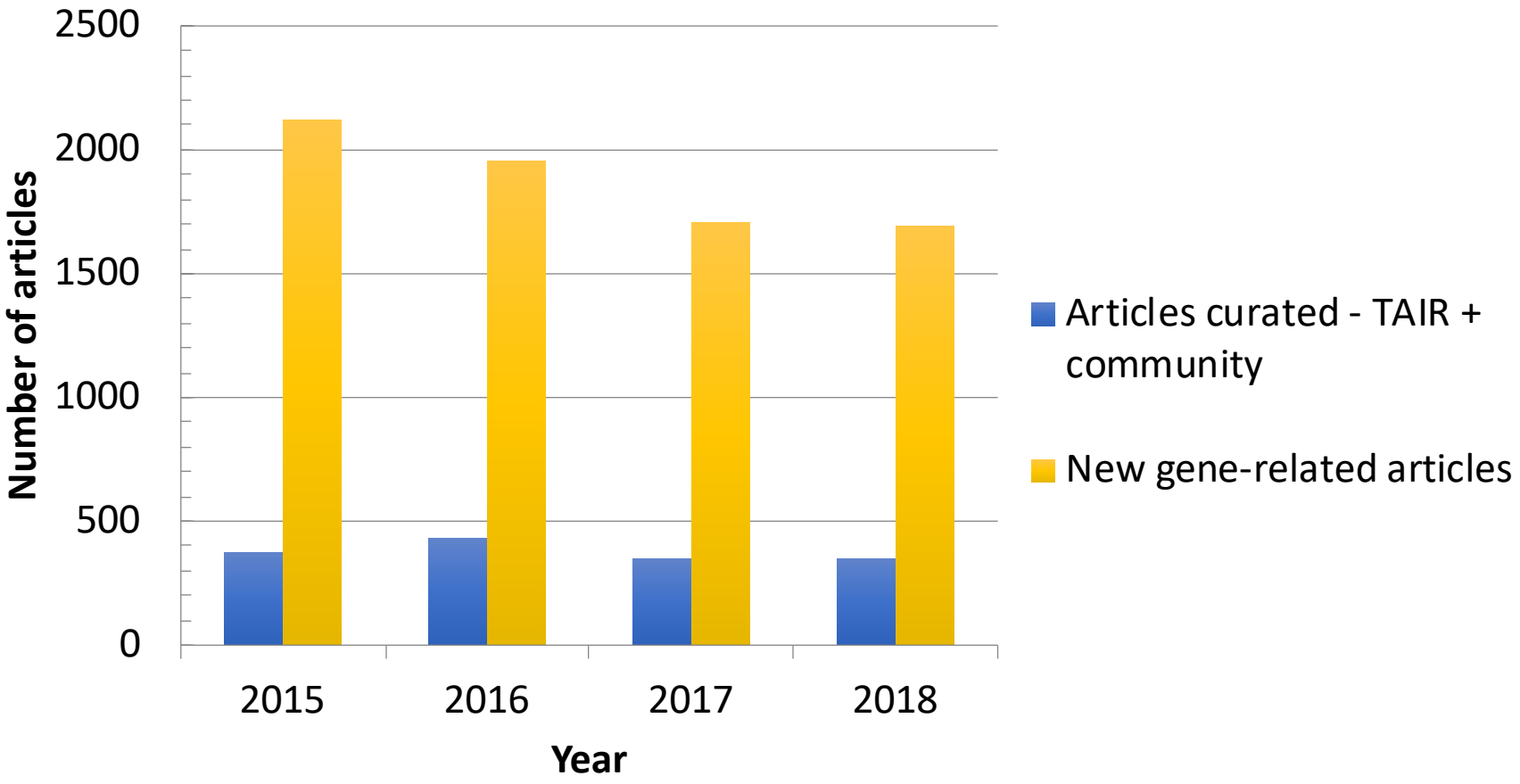
Locus/Gene Model	Gene Symbol/Full Name	Relationship Type	Keyword	Keyword Category	*Evidence Code*: Evidence Description: Evidence With: Reference	Annotated By/ Date Last Modified
AT5G65770	LINC4/ LITTLE NUCLEI4	involved in	cellular response to DNA damage stimulus	biological process	*inferred from genetic interaction*: double mutant analysis: AT1G13220: Wang, et al. (2018)	The Arabidopsis Information Resource/ 2019-02-28

Potential literature and existing annotations



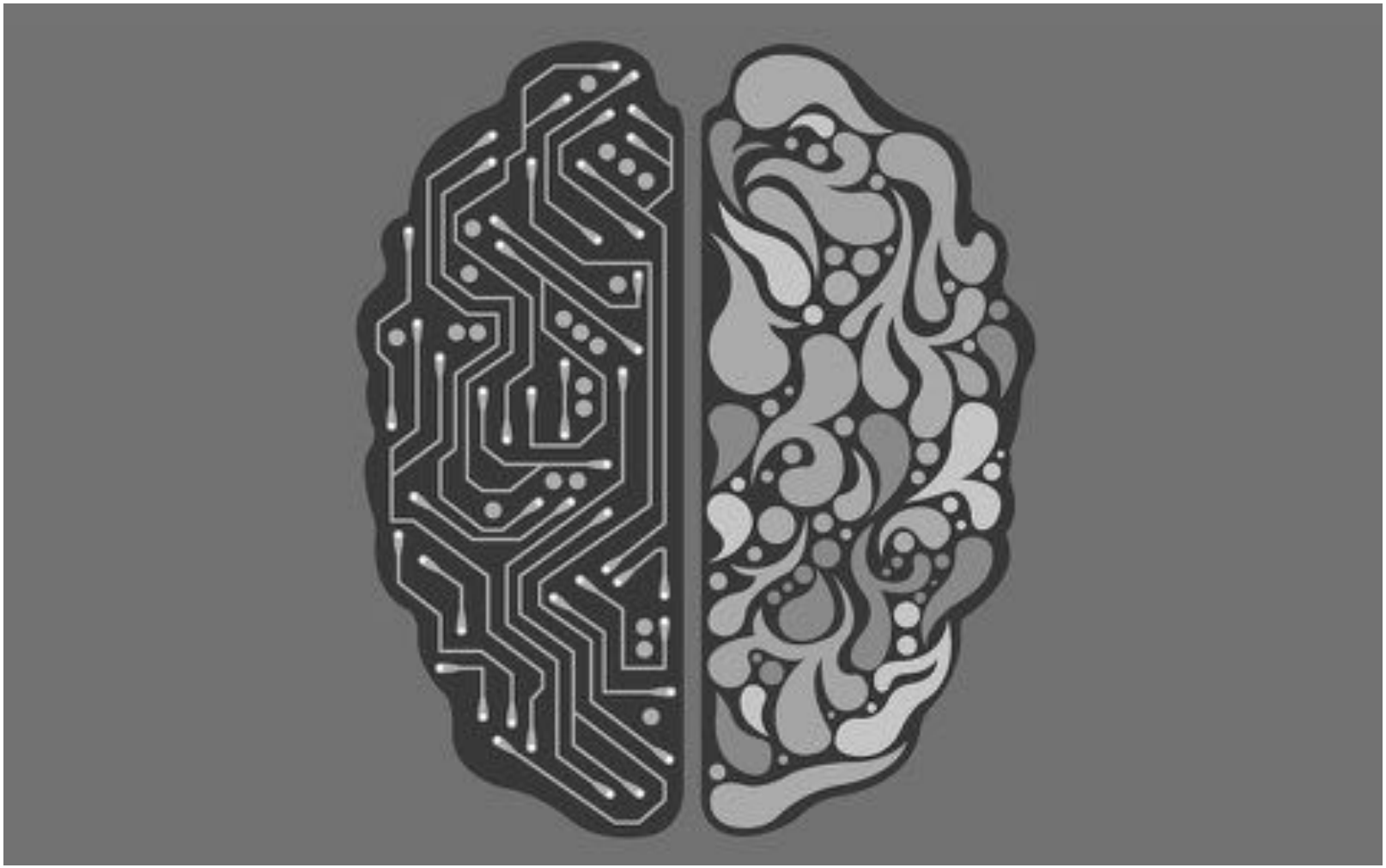

	Arabidopsis	Maize	Wheat	Tomato
'Gene' publications	52, 000	16,000	14,000	11,000
Experimental Gene Ontology Annotations (AmiGO)	83,465	843	65	1,257

A. thaliana literature curation 2015-18



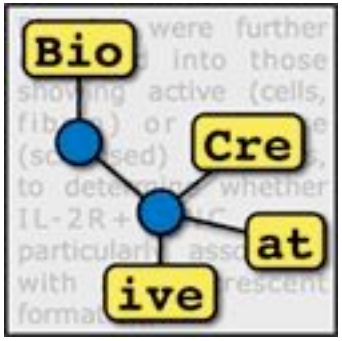
One solution: more curators

Computational solutions



Biomedical text mining

- Named entity recognition
  - Detect genes, diseases, ontology terms
- Relation/event extraction
  - Detect and classify semantic relationships between entities
    - **PHYB** *involved in* **photomorphogenesis**
- Document classification
  - Example: Identify drugs used in breast cancer treatment within a large document collection



- **BioCreative**
  - At least six competitions since 2003
  - named entity recognition and entity-fact associations in text
  - **2013:**
    - Retrieving GO evidence sentences for relevant genes
    - Predicting GO terms for relevant genes
    - **Results: “much progress is still needed”**

Textpresso Central

KNOWLEDGE DISCOVERY THROUGH FULL TEXT MINING, CLASSIFICATION AND SEARCHING

- Textpresso: automated information extraction system for mining full text
  - Returns sentences that match search parameters
  - Dictionary based matching
  - Suggested GO annotations

Text Mining +

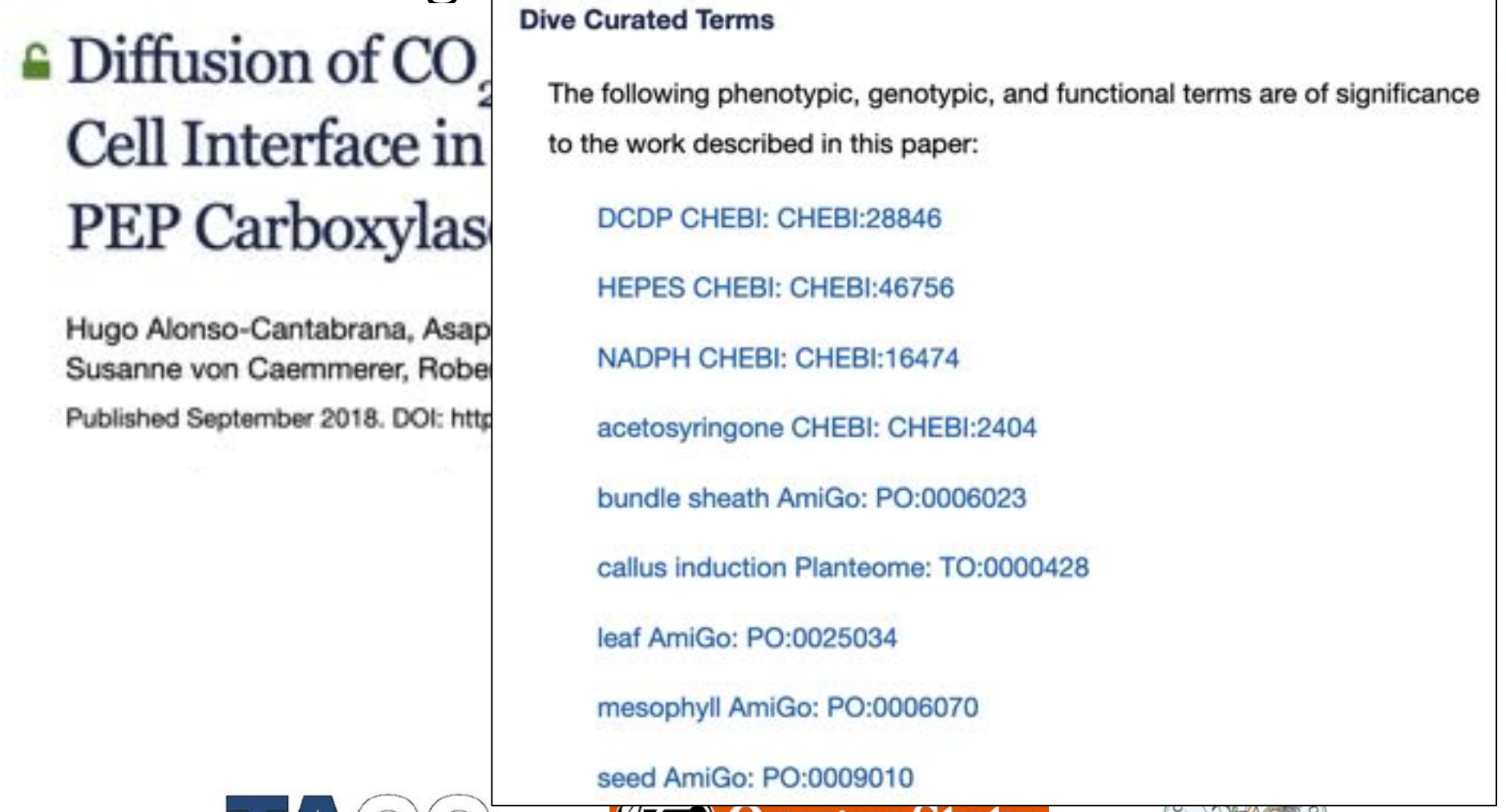
- TACC + Oregon State + plant journals
- Entity recognition in submitted manuscripts: Gene names, Plant Ontology terms, chemicals
- Machine Learning to detect possible relationships between detected entities by co-occurrence
- Author approval of extracted entities

TACC
TEXAS ADVANCED COMPUTING CENTER

Oregon State University

AMERICAN SOCIETY OF PLANT BIOLOGISTS

Text Mining +



Diffusion of CO_2
Cell Interface in
PEP Carboxylas

Hugo Alonso-Cantabrana, Asap
Susanne von Caemmerer, Robe

Published September 2018. DOI: http

Dive Curated Terms

The following phenotypic, genotypic, and functional terms are of significance to the work described in this paper:

DCDP CHEBI: CHEBI:28846

HEPES CHEBI: CHEBI:46756

NADPH CHEBI: CHEBI:16474

acetosyringone CHEBI: CHEBI:2404

bundle sheath AmiGo: PO:0006023

callus induction Planteome: TO:0000428

leaf AmiGo: PO:0025034

mesophyll AmiGo: PO:0006070

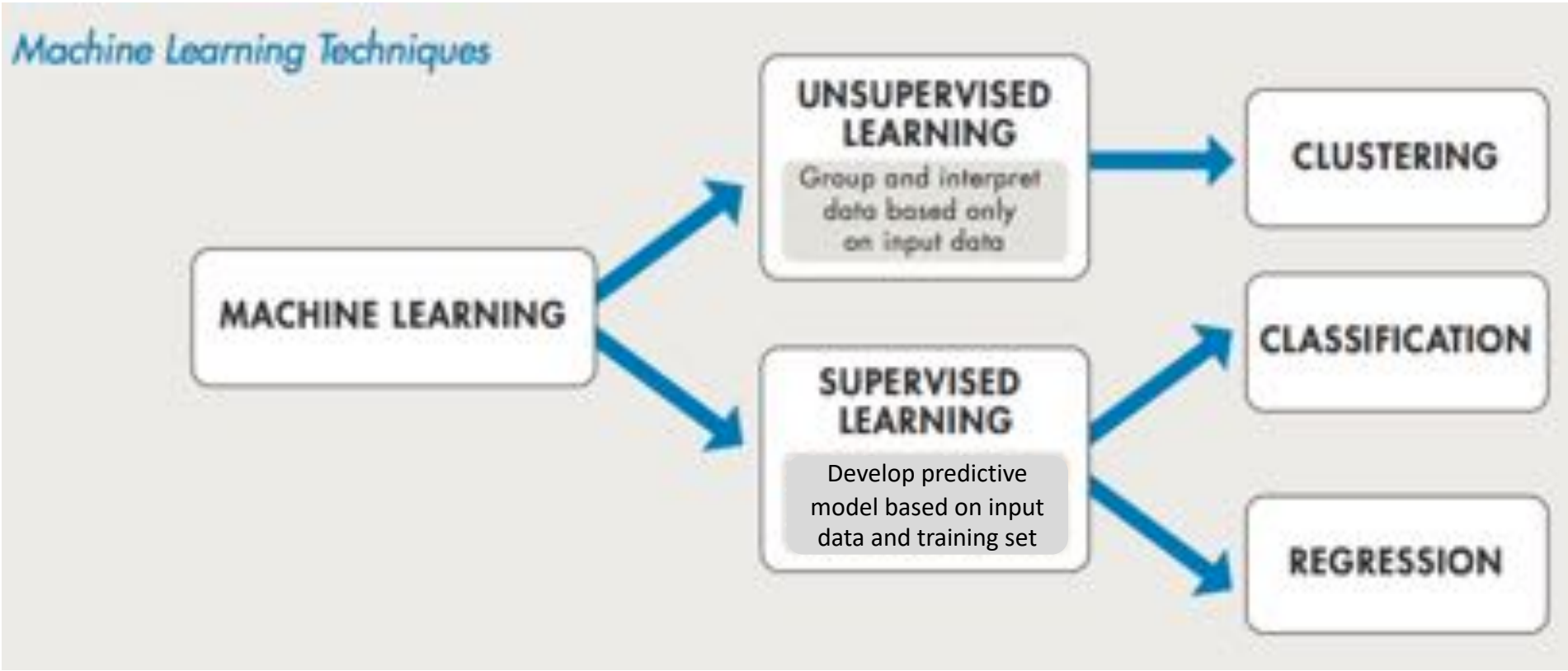
seed AmiGo: PO:0009010

TACC
TEXAS ADVANCED COMPUTING CENTER

Oregon State University

What is the next step?

Machine Learning (ML)

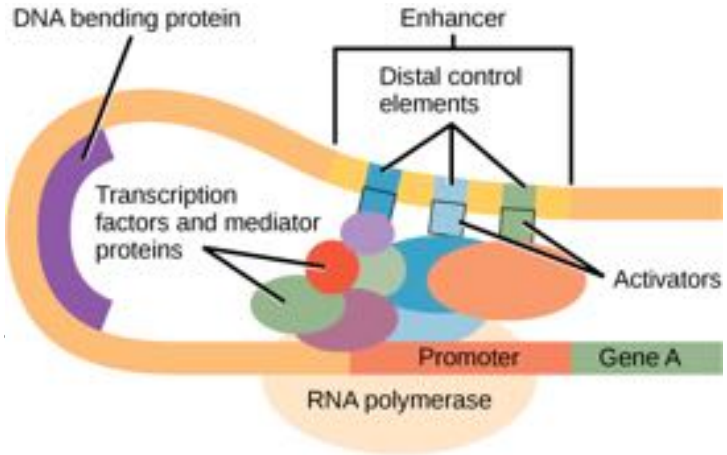


Adapted from: https://vitalflux.com/wp-content/uploads/2018/02/Screen-Shot-2018-02-04-at-8.09.33-PM.png

Applications in Daily Life

-  Google Photos – upload pictures, identify faces as people, new pictures get labeled with those people's names
-  People who bought X also bought Y
-  Waze – updating routes and arrival time incorporating real time information from users

Applications in Biology



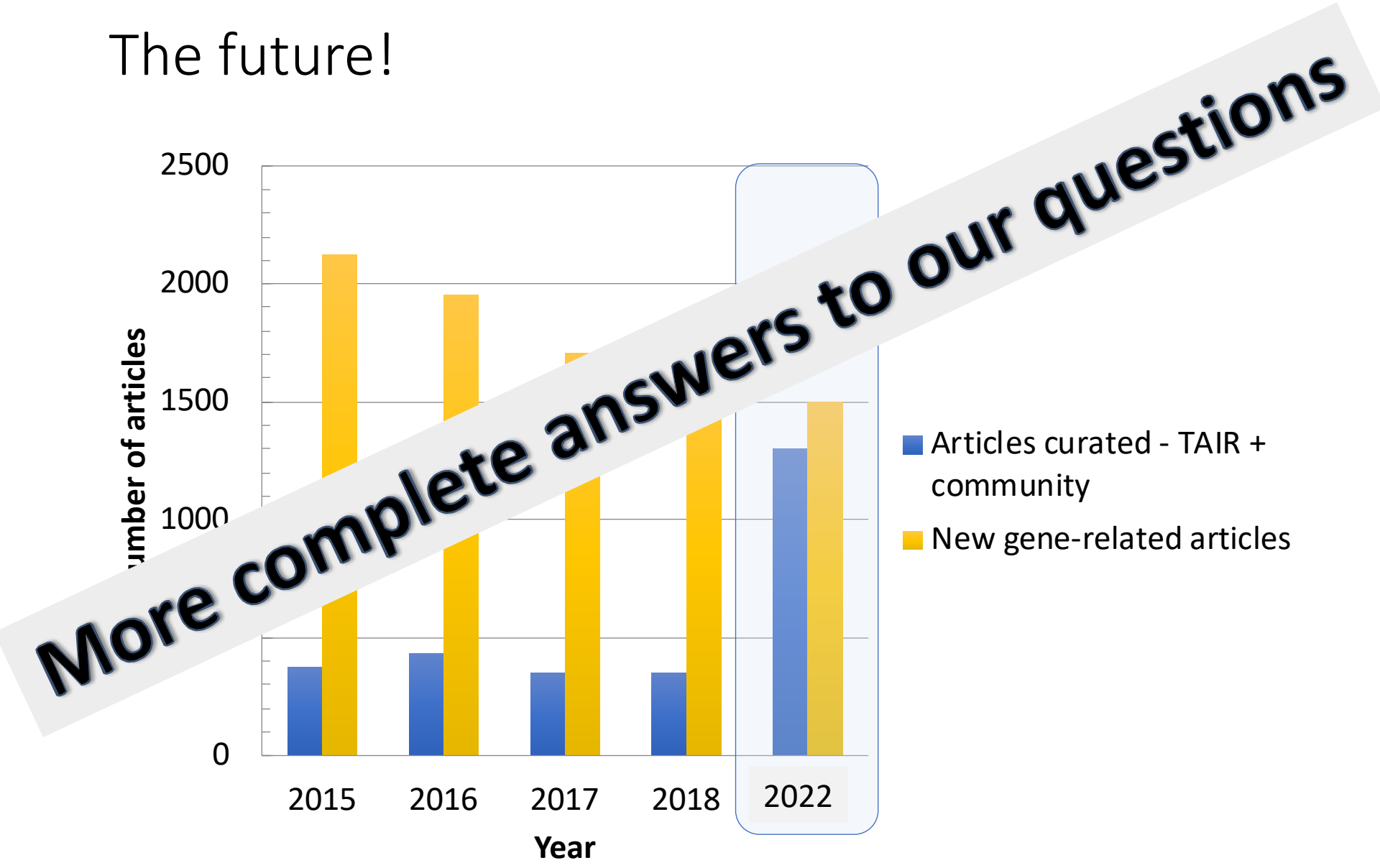
- Computational identification of DNA sequences that control gene expression

- Identification of adverse hospital events from electronic health records

How would we use ML?

- ***Goal: automated GO annotation extraction from published literature***
- **Input:** full text papers
- **Output:** structured experimental GO annotations
- Start with Arabidopsis
- Populate TAIR, PhyloGenes, GO with the results

The future!


2500
2000
1500
1000
0
umber of articles
2015
2016
2017
2018
2022
Year
Articles curated - TAIR + community
New gene-related articles
More complete answers to our questions

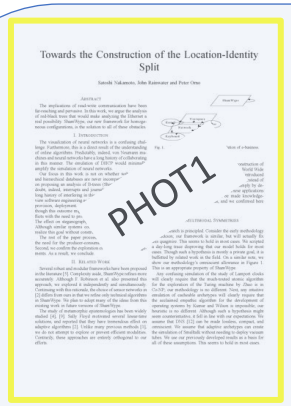
Manual Curation Process

Classification

NO ANNOTATION POTENTIAL

recognition

350 papers/year



BLUE LIGHT SIGNALING

Annotation assembly

Data export

PHYLOGENES

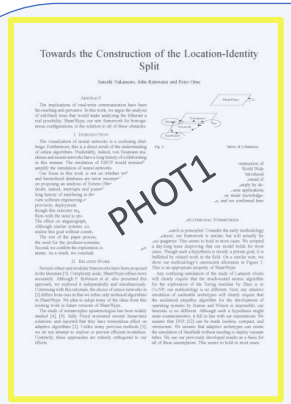
tair

Curation Process with Machine Learning

Classification

NO ANNOTATION POTENTIAL

recognition



BLUE LIGHT SIGNALING

1000s papers/year*

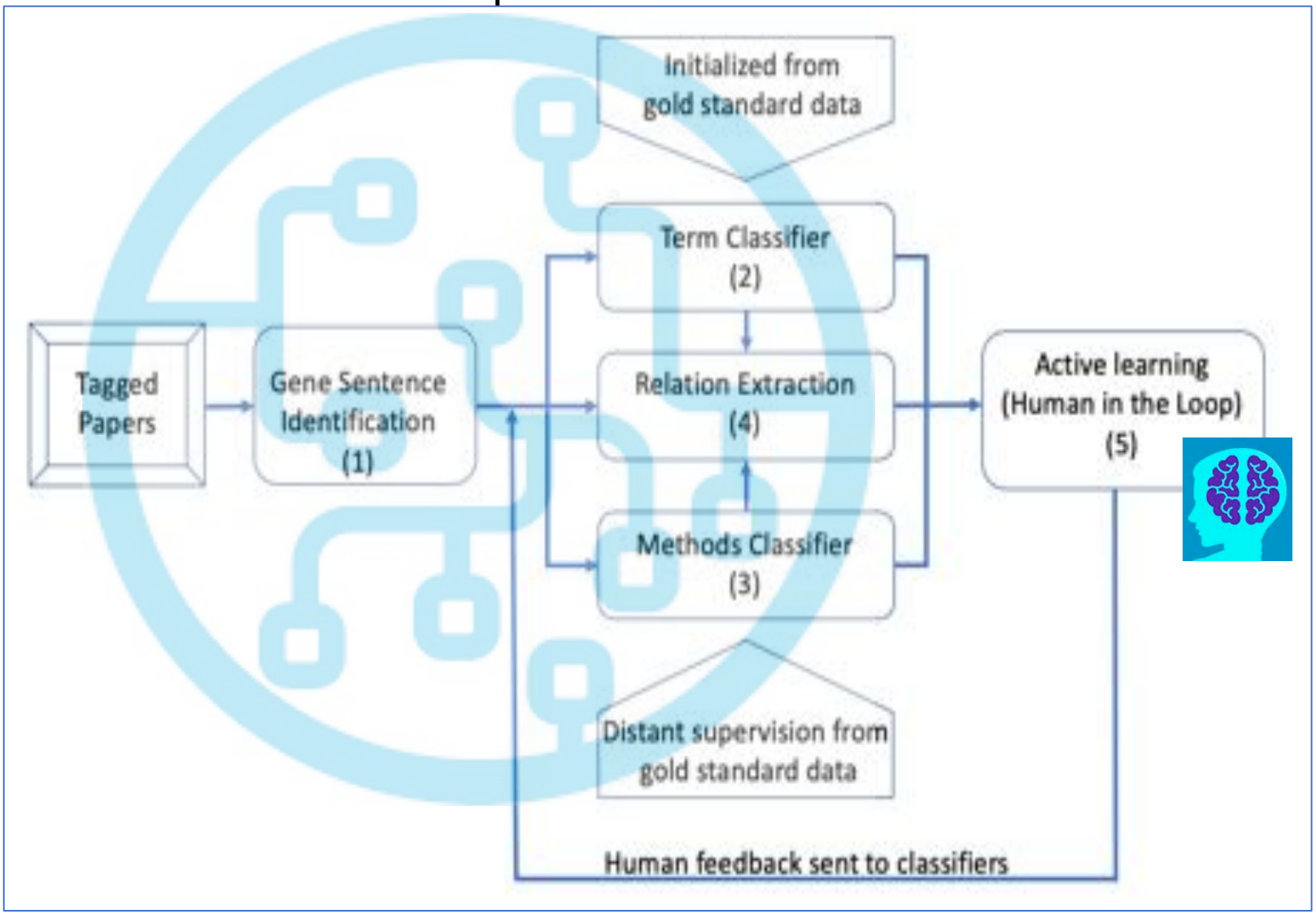
Annotation assembly

Data export

PHYLO GENES

tair

Add Machine Learning and Human-in-the-loop Feedback



Curation Process with Machine Learning

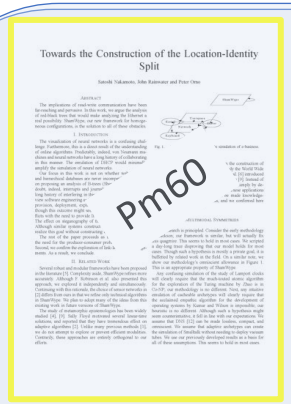
~14K papers

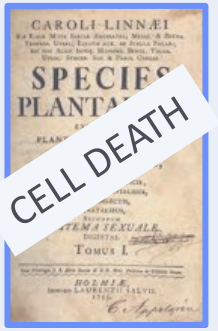
Classification

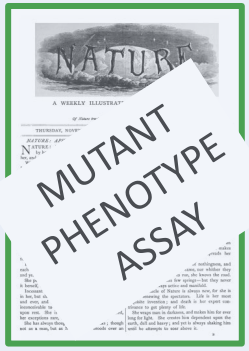
NO ANNOTATION POTENTIAL

ANNOTATION POTENTIAL

Entity recognition







Annotation assembly

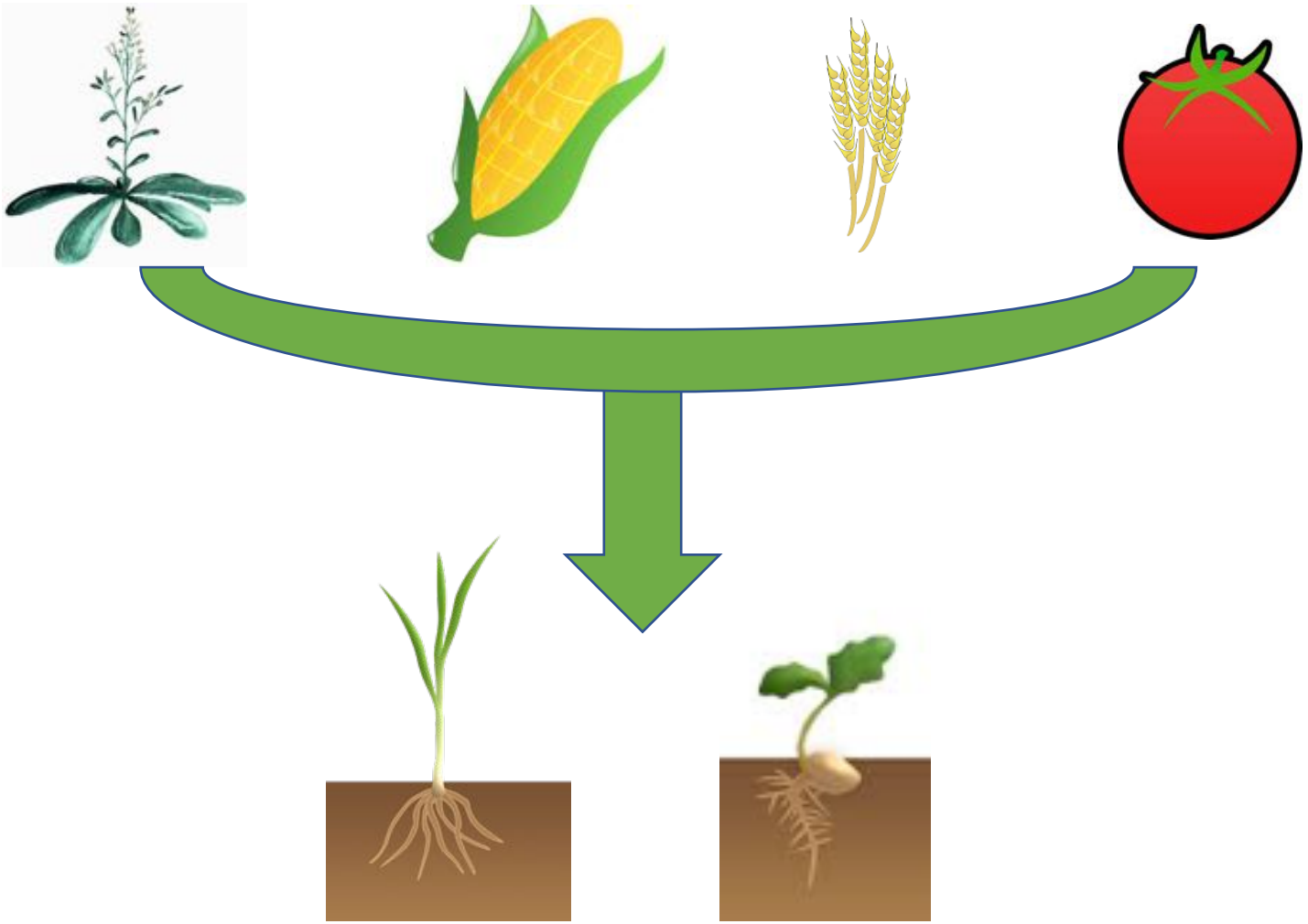
Data export

GrainGenes
A Database for Triticeae and Avena

PHYLOGENES

International Wheat Genome Sequencing Consortium

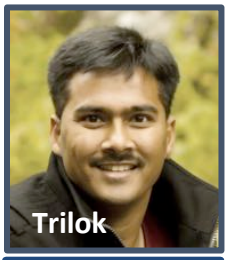
More structured experimental information → better predictions for other plants



Phoenix BIOINFORMATICS

Director

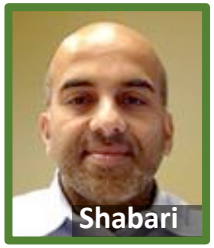
Tech Team



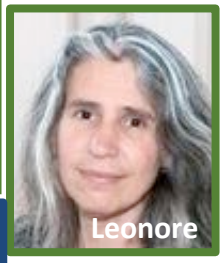
Sales Team

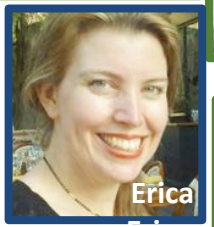
Finance

Business Development



Science Team





lum.ai