Machine Learning for Multi Environmental Trials

Danilo A. Sarti*, Alessandra Lemos, Rafael Moral and Andrew Parnell
danilo.sarti@mu.ie
http://danilosarti.github.io/

INNOVAR

EUROPEAN UNION

Maynooth University
National University of Ireland Maynooth

AGENDA

- InnoVar project
- Machine Learning
- Machine Learning in InnoVar
- Achievements
- Next steps

InnoVar

- 21 Partners
- 15 Locations - 5 Agroclimatic zones
- Integration of new science into DUS and VCU testing processes
- Combine VCU and DUS characters
- Incorporate variety information into decision making on farm.
- *Triticum durum* and *Triticum aestivum*

ADAS

afbi AGRI-FOOD & BIOSCIENCES INSTITUTE

AHDB AGRICULTURE & HORTICULTURE DEVELOPMENT BOARD

CONSULAI www.consulai.com

crea Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria

CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

An Roinn Talmhaíochta, Bia agus Mara Department of Agriculture, Food and the Marine

Department for Environment Food & Rural Affairs

HORT@ From research to field

ICARDA Science for resilient livelihoods in dry areas

IP PRAGMATICS

ISRIC World Soil Information

LESPROJEKT služby

Maynooth University National University of Ireland Maynooth

Origin ENTERPRISES PLC

TystofteFonden

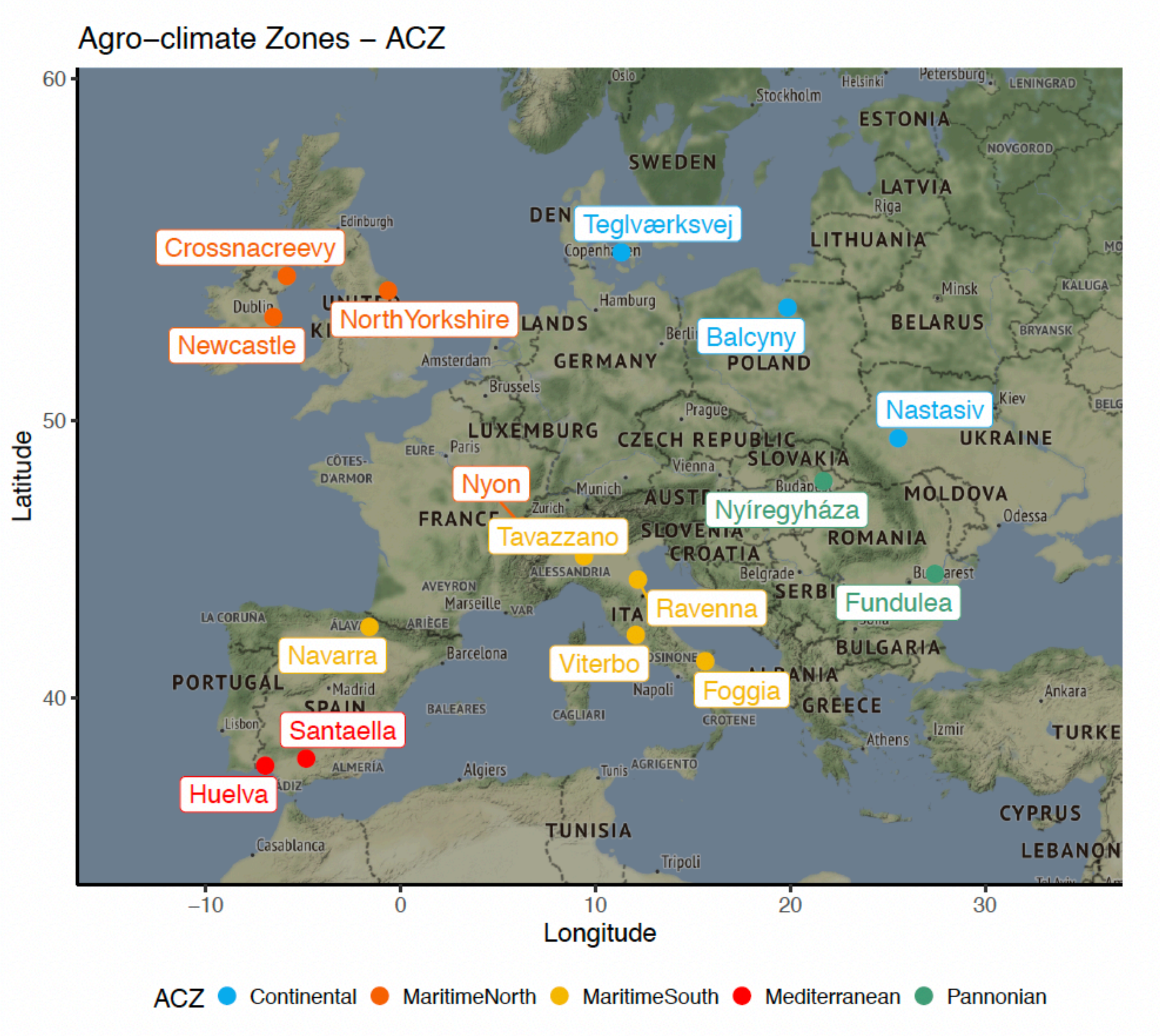
UCD DUBLIN

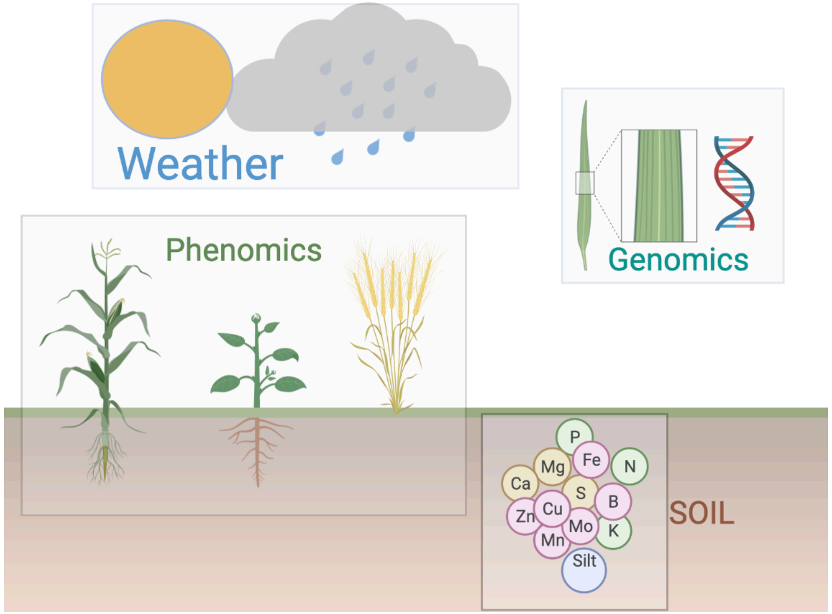
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA DIPARTIMENTO DI SCIENZE E TECNOLOGIE AGRO-ALIMENTARI

DEBRECENI EGYETEM

UNIVERSITÀ DEGLI STUDI DELLA Tuscia

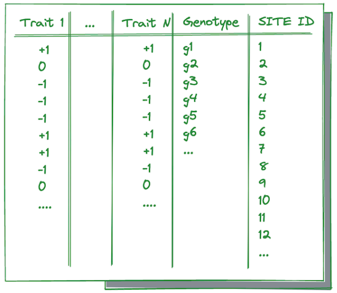
POLITÉCNICA UNIVERSIDAD POLITÉCNICA DE MADRID


Agro−climate Zones − ACZ
Latitude
60
50
40
Longitude
−10
0
10
20
30
Crossnacreevy
Newcastle
NorthYorkshire
Teglværksvej
Balcyny
Nastasiv
Nyon
Tavazzano
Nyíregyháza
Ravenna
Fundulea
Navarra
Viterbo
Foggia
Santaella
Huelva
ACZ
● Continental
● MaritimeNorth
● MaritimeSouth
● Mediterranean
● Pannonian

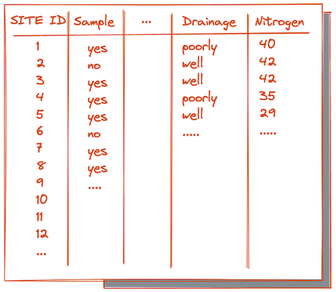


- Multi Environmental Trials
- Phenomics, DUS and VCU
- Genomics
- Weather
- Soil
- Drone images

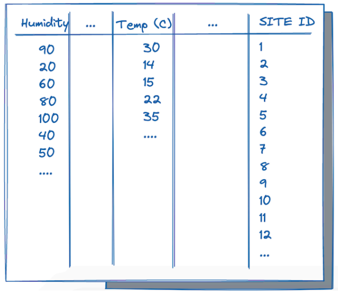
- Data collected will be stored in a database.
- Data will be used to create models and systems.
- **COVID** delays



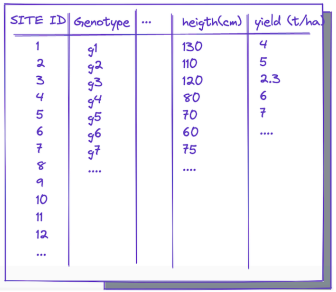
Trait 1	...	Trait N	Genotype	SITE ID
+1		+1	g1	1
0		0	g2	2
-1		-1	g3	3
-1		-1	g4	4
-1		-1	g5	5
+1		+1	g6	6
+1		+1	...	7
-1		-1		8
0		0		9
....		....		10
				11
				12
				...



SITE ID	Sample	...	Drainage	Nitrogen
1	yes		poorly	40
2	no		well	42
3	yes		well	42
4	yes		poorly	35
5	yes		well	29
6	no		.....	.....
7	yes			
8	yes			
9	....			
10				
11				
12				
...				

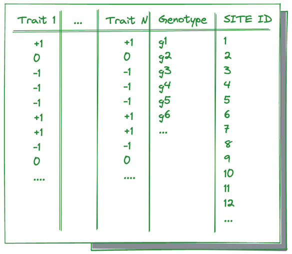


Humidity	...	Temp (C)	...	SITE ID
90		30		1
20		14		2
60		15		3
80		22		4
100		35		5
40		....		6
50				7
....				8
				9
				10
				11
				12
				...

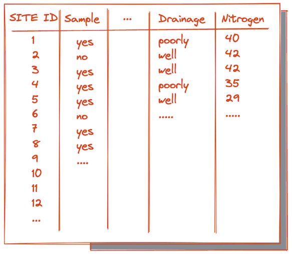


SITE ID	Genotype	...	heigth(cm)	yield (t/ha)
1	g1		130	4
2	g2		110	5
3	g3		120	2.3
4	g4		80	6
5	g5		70	7
6	g6		60	....
7	g7		75	
8	....		....	
9				
10				
11				
12				
...				

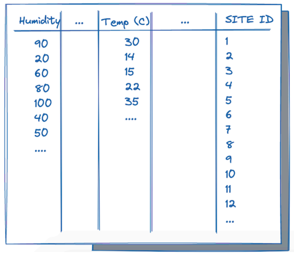
Machine Learning



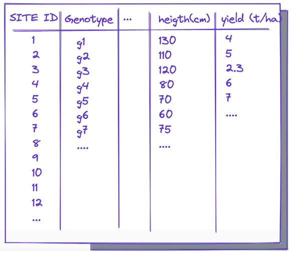
Trait 1	...	Trait N	Genotype	SITE ID
+1		+1	g1	1
0		0	g2	2
-1		-1	g3	3
-1		-1	g4	4
-1		-1	g5	5
+1		+1	g6	6
+1		+1	...	7
-1		-1		8
0		0		9
....		....		10
				11
				12
				...



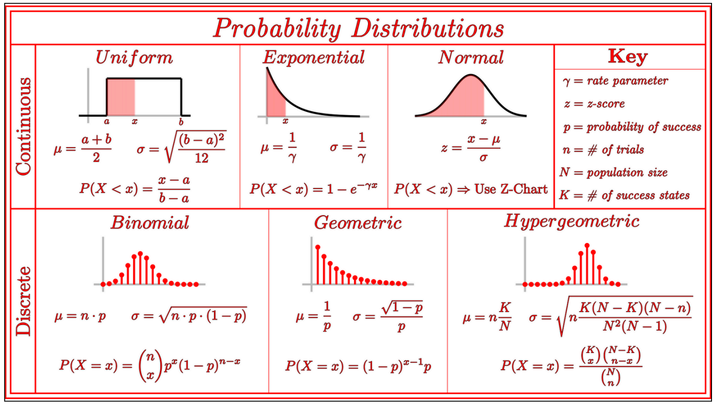
SITE ID	Sample	...	Drainage	Nitrogen
1	yes		poorly	40
2	no		well	42
3	yes		well	42
4	yes		poorly	35
5	yes		well	29
6	no		.....	.....
7	yes			
8	yes			
9	....			
10				
11				
12				
...				



Humidity	...	Temp (C)	...	SITE ID
90		30		1
20		14		2
60		15		3
80		22		4
100		35		5
40		....		6
50				7
....				8
				9
				10
				11
				12
				...



SITE ID	Genotype	...	heigth(cm)	yield (t/ha)
1	g1		130	4
2	g2		110	5
3	g3		120	2.3
4	g4		80	6
5	g5		70	7
6	g6		60	....
7	g7		75	
8	....		....	
9				
10				
11				
12				
...				



Probability Distributions

Continuous

Uniform

$\mu = \frac{a+b}{2} \qquad \sigma = \sqrt{\frac{(b-a)^2}{12}}$

$P(X < x) = \frac{x-a}{b-a}$

Exponential

$\mu = \frac{1}{\gamma} \qquad \sigma = \frac{1}{\gamma}$

$P(X < x) = 1 - e^{-\gamma x}$

Normal

$z = \frac{x-\mu}{\sigma}$

$P(X < x) \Rightarrow$ Use Z-Chart

Key

γ = rate parameter
z = z-score
p = probability of success
n = # of trials
N = population size
K = # of success states

Discrete

Binomial

$\mu = n \cdot p \qquad \sigma = \sqrt{n \cdot p \cdot (1-p)}$

$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

Geometric

$\mu = \frac{1}{p} \qquad \sigma = \frac{\sqrt{1-p}}{p}$

$P(X = x) = (1-p)^{x-1} p$

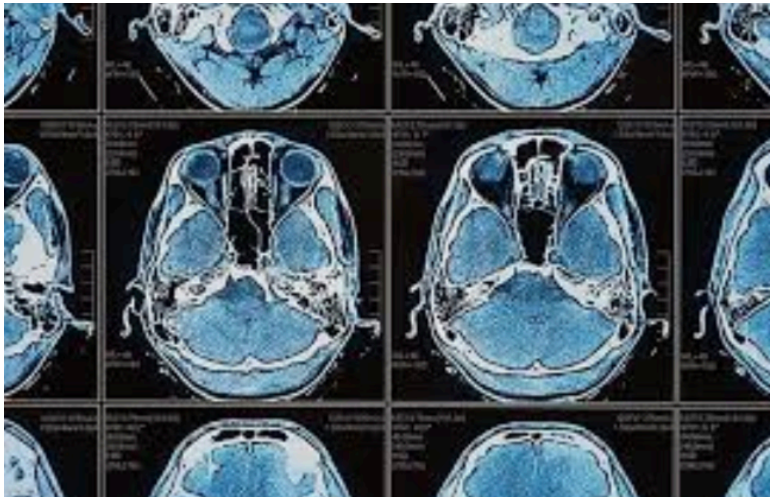
Hypergeometric

$\mu = n\frac{K}{N} \qquad \sigma = \sqrt{n\frac{K(N-K)(N-n)}{N^2(N-1)}}$

$P(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$

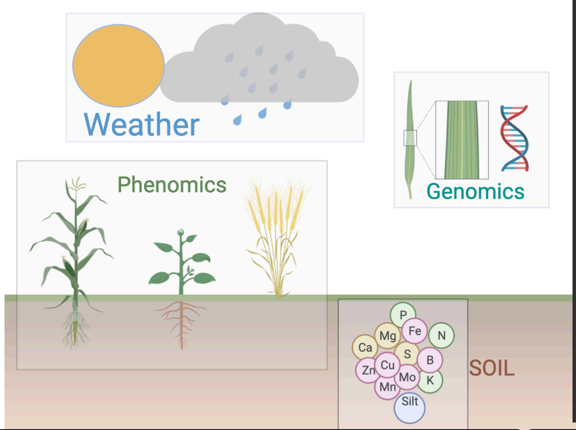
Study of computer algorithms that can improve automatically through experience and use of data.

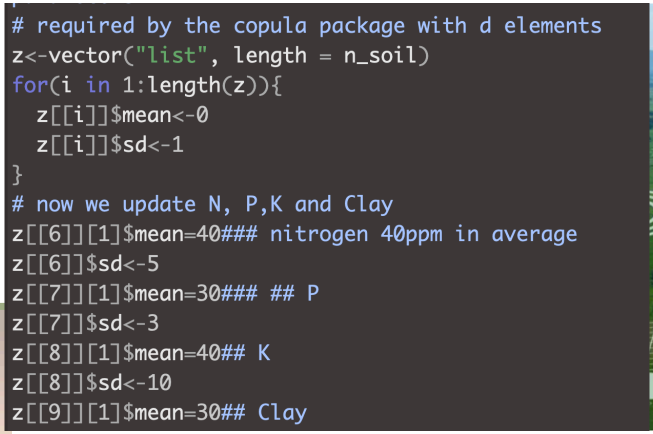
Machine Learning Applications



Social Media
Streaming platforms
Precision medicine
Precision agriculture
Now: MET, DUS and VCU

Machine Learning in InnoVar



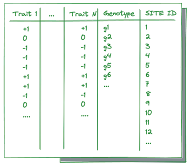


```
# required by the copula package with d elements
z<-vector("list", length = n_soil)
for(i in 1:length(z)){
  z[[i]]$mean<-0
  z[[i]]$sd<-1
}
# now we update N, P,K and Clay
z[[6]][1]$mean=40### nitrogen 40ppm in average
z[[6]]$sd<-5
z[[7]][1]$mean=30### ## P
z[[7]]$sd<-3
z[[8]][1]$mean=40## K
z[[8]]$sd<-10
z[[9]][1]$mean=30## Clay
```

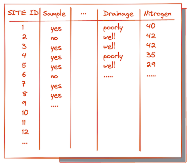
Data Collection

Computational Solutions

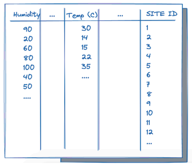
Decision Support Systems



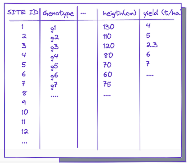
Trait 1	...	Trait N	Genotype	SITE ID
+1		+1	g1	1
0		0	g2	2
-1		-1	g3	3
-1		-1	g4	4
-1		-1	g5	5
+1		+1	g6	6
+1		+1	...	7
-1		-1		8
0		0		9
....		....		10
				11
				12
				...



SITE ID	Sample	...	Drainage	Nitrogen
1	yes		poorly	40
2	no		well	42
3	yes		well	42
4	yes		poorly	35
5	yes		well	29
6	no		.....	.....
7	yes			
8	yes			
9	....			
10				
11				
12				
...				

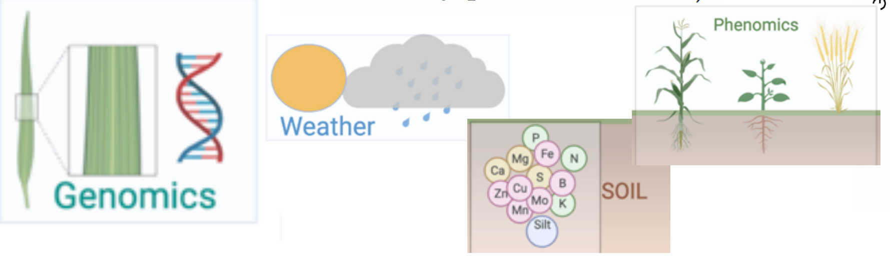


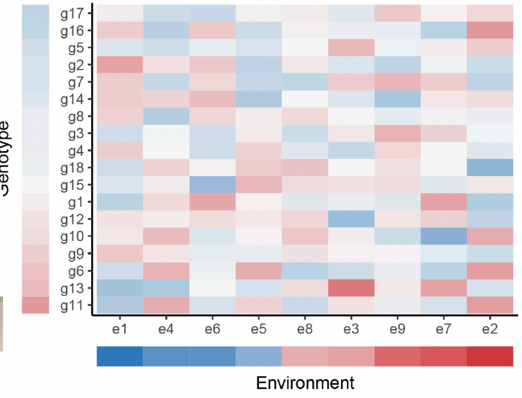
Humidity	...	Temp (C)	...	SITE ID
90		30		1
20		14		2
60		15		3
80		22		4
100		35		5
40		....		6
50				7
....				8
				9
				10
				11
				12
				...



SITE ID	Genotype	...	heigth(cm)	yield (t/ha)
1	g1		130	4
2	g2		110	5
3	g3		120	2.3
4	g4		80	6
5	g5		70	7
6	g6		60	....
7	g7		75	
8	....		....	
9				
10				
11				
12				
...				

$$y_{ij}|\mathbf{x}_{ij},\Theta \sim \mathrm{N}\left(\mu + g_i + e_j + \sum_{t=1}^{T} h(\mathbf{x}_{ij},\mathcal{M}_t,\mathcal{T}_t),\sigma^2\right)$$



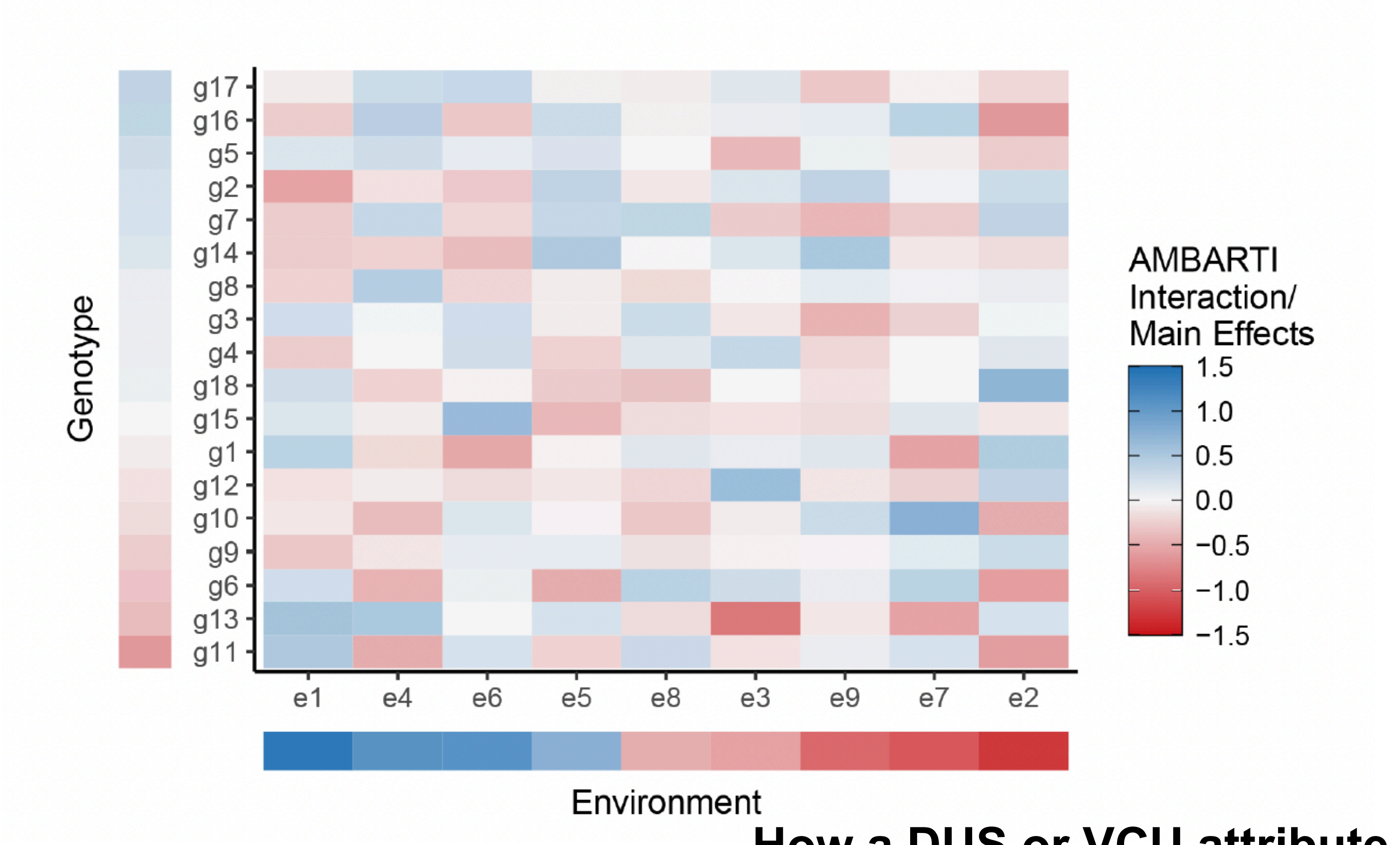


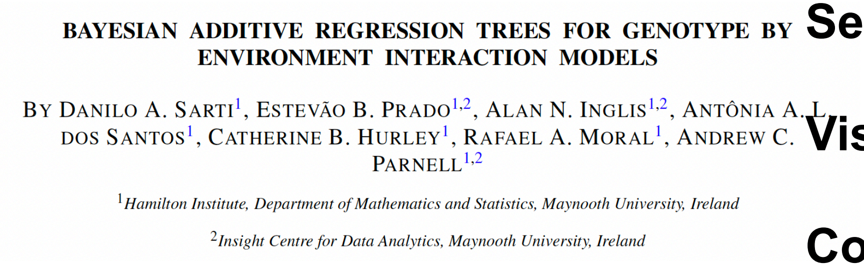
Comprehend and predict the behaviour of DUS and VCU attributes considering interactions (G by E and G by E by T)

Achievements: innovaR

- R programming language package
- Developed to simulate a complete multi environmental trial (soil, genomics, and phenomics), considering dependence between attributes of VCU and DUS.
- Can be used to validate statistical models based on VCU and DUS data.
- Free and open source.
- It will be available at: https://github.com/danilosarti/innovaR

Achievements: AMBARTI





BAYESIAN ADDITIVE REGRESSION TREES FOR GENOTYPE BY ENVIRONMENT INTERACTION MODELS

By Danilo A. Sarti[1], Estevão B. Prado[1,2], Alan N. Inglis[1,2], Antônia A. L. dos Santos[1], Catherine B. Hurley[1], Rafael A. Moral[1], Andrew C. Parnell[1,2]

[1] *Hamilton Institute, Department of Mathematics and Statistics, Maynooth University, Ireland*

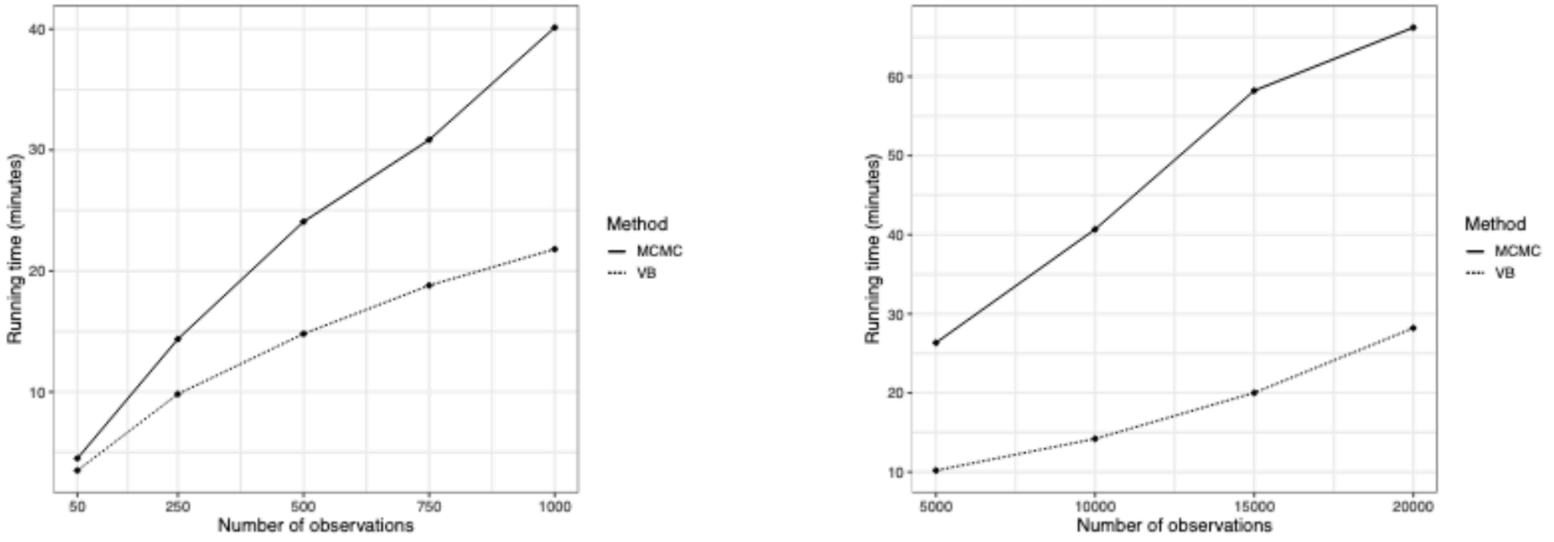
[2] *Insight Centre for Data Analytics, Maynooth University, Ireland*

How a DUS or VCU attribute behaves in Several Environments.

Visual appeal. Could be an app for farmers.

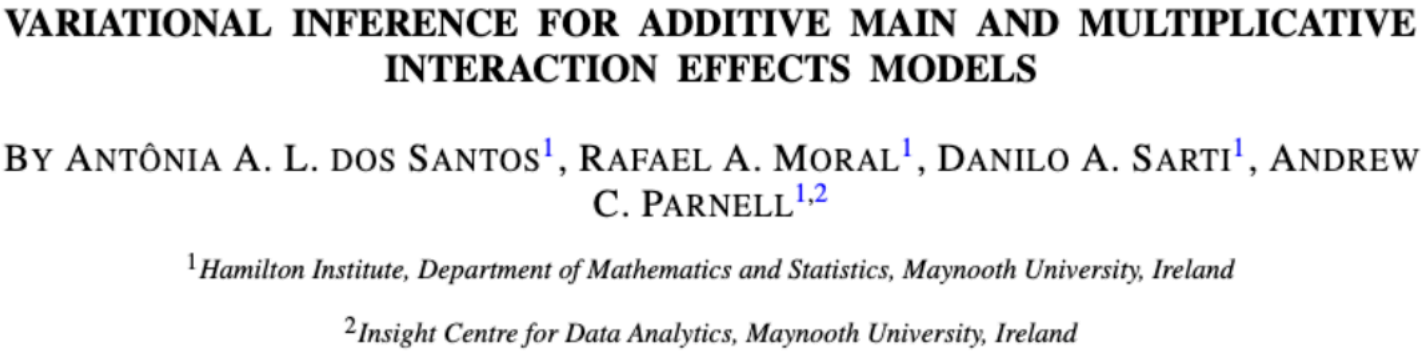
Could we predict the performance of a genotype Considering some genomic attribute?

Achievements: Variational Inference



(a) Comparison of simulation times for y = $\{100, 250, 500, 1000\}$.

(b) Comparison of simulation times for y = $\{5000, 10000, 15000, 20000\}$.



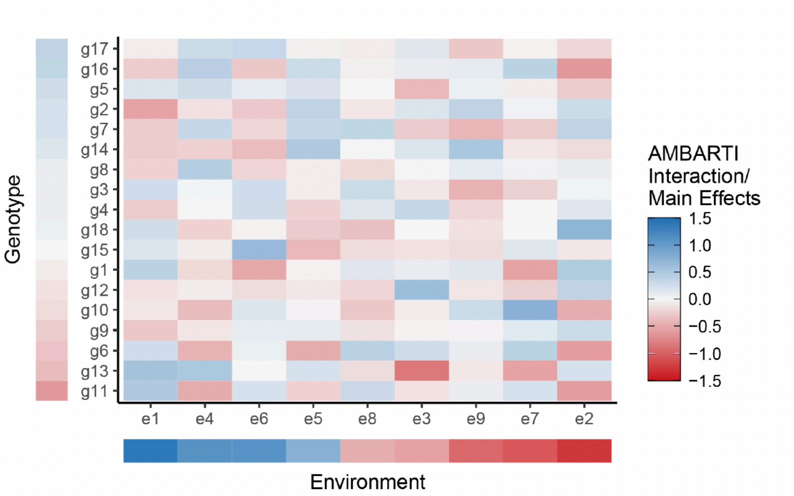
VARIATIONAL INFERENCE FOR ADDITIVE MAIN AND MULTIPLICATIVE INTERACTION EFFECTS MODELS

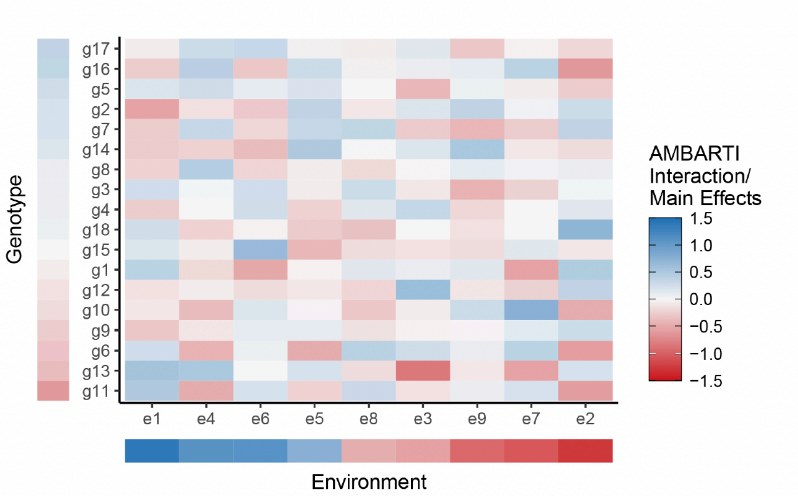
BY ANTÔNIA A. L. DOS SANTOS[1], RAFAEL A. MORAL[1], DANILO A. SARTI[1], ANDREW C. PARNELL[1,2]

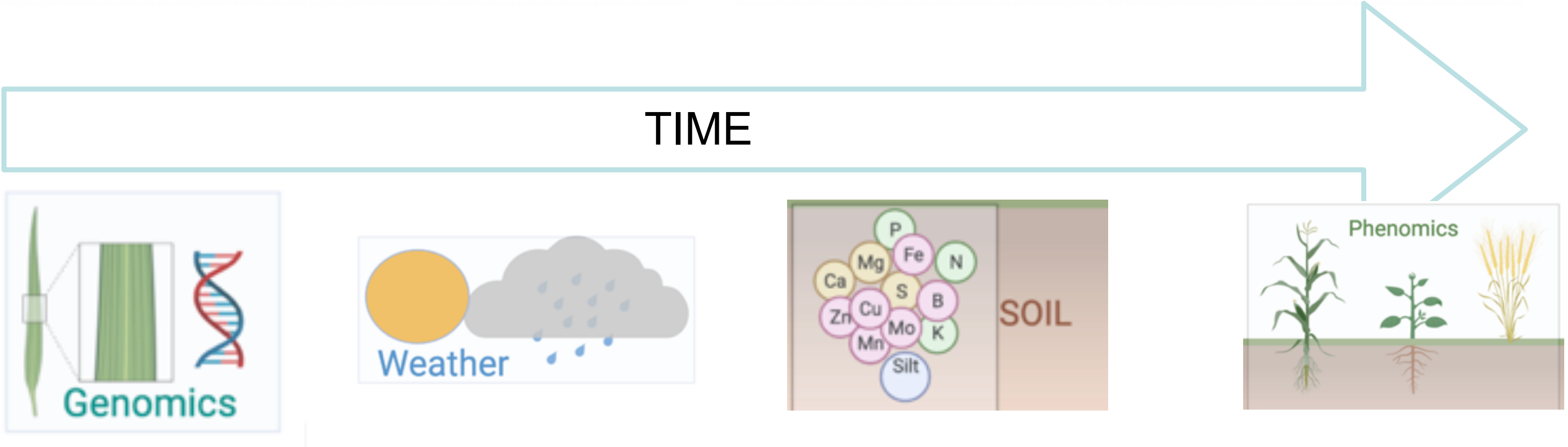
[1] *Hamilton Institute, Department of Mathematics and Statistics, Maynooth University, Ireland*

[2] *Insight Centre for Data Analytics, Maynooth University, Ireland*

Other interactions: G by E by T by Soil
How could such models incorporate genomic information … epigenetic markers…







Short Summary

- InnoVar has been collecting massive data from the soil, weather, genomics, and phenomics.
- The data will be used to develop machine learning algorithms and to create decision support systems.
- There is an opportunity for such models to be used to optimize DUS and VCU by allowing the prediction of the genotypes' attributes across environments and time, considering information on genomics and other sources. This could eventually speed up the process of regulation of new varieties.