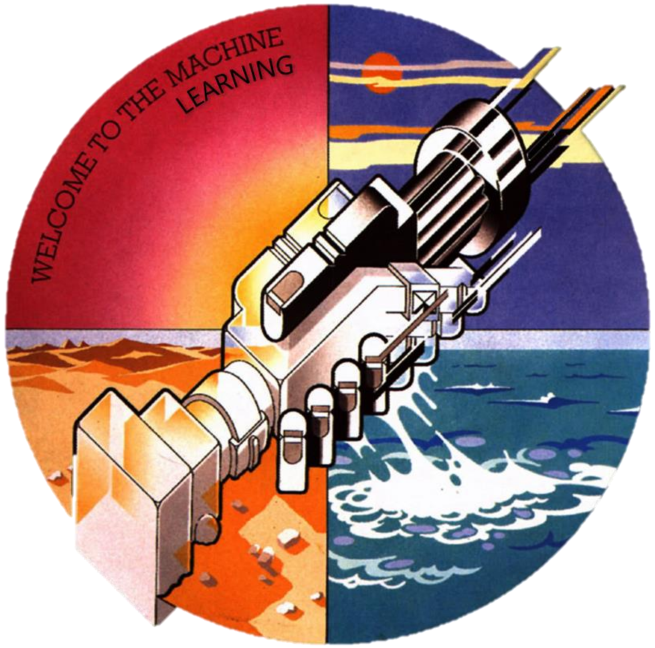


Alencar Xavier

Learning from Data: Machine Learning in plant breeding

CORTEVA™ agriscience

PURDUE UNIVERSITY

Key messages

1. **ML in hands-on breeding: HTPs, BLUPs and GEBVs**
2. **Usage of the data relies on the nature of the signal**
3. **Breeding applications mostly on Gaussian process**
4. **Machines are important when signal is scarce**

Outline

Overview of ML

Topic 1 - Data processing

Topic 2 - Signal detection

Topic 3 - Complex workflow

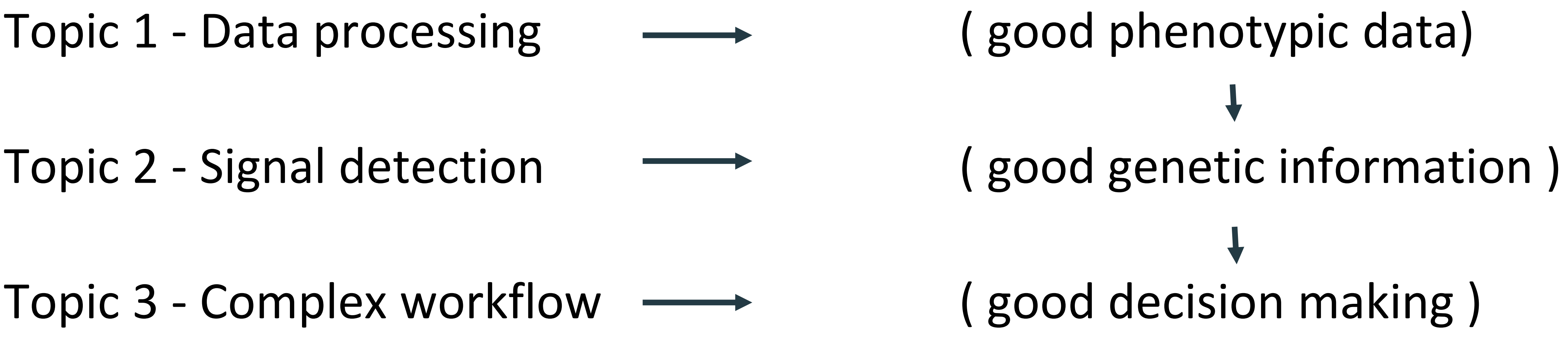
Concluding remarks

"Begin at the beginning," "the King said, *"and go on till you come to the end: then stop."*

Lewis Carroll, Alice in Wonderland

Outline

Overview of ML



Topic 1 - Data processing → (good phenotypic data)

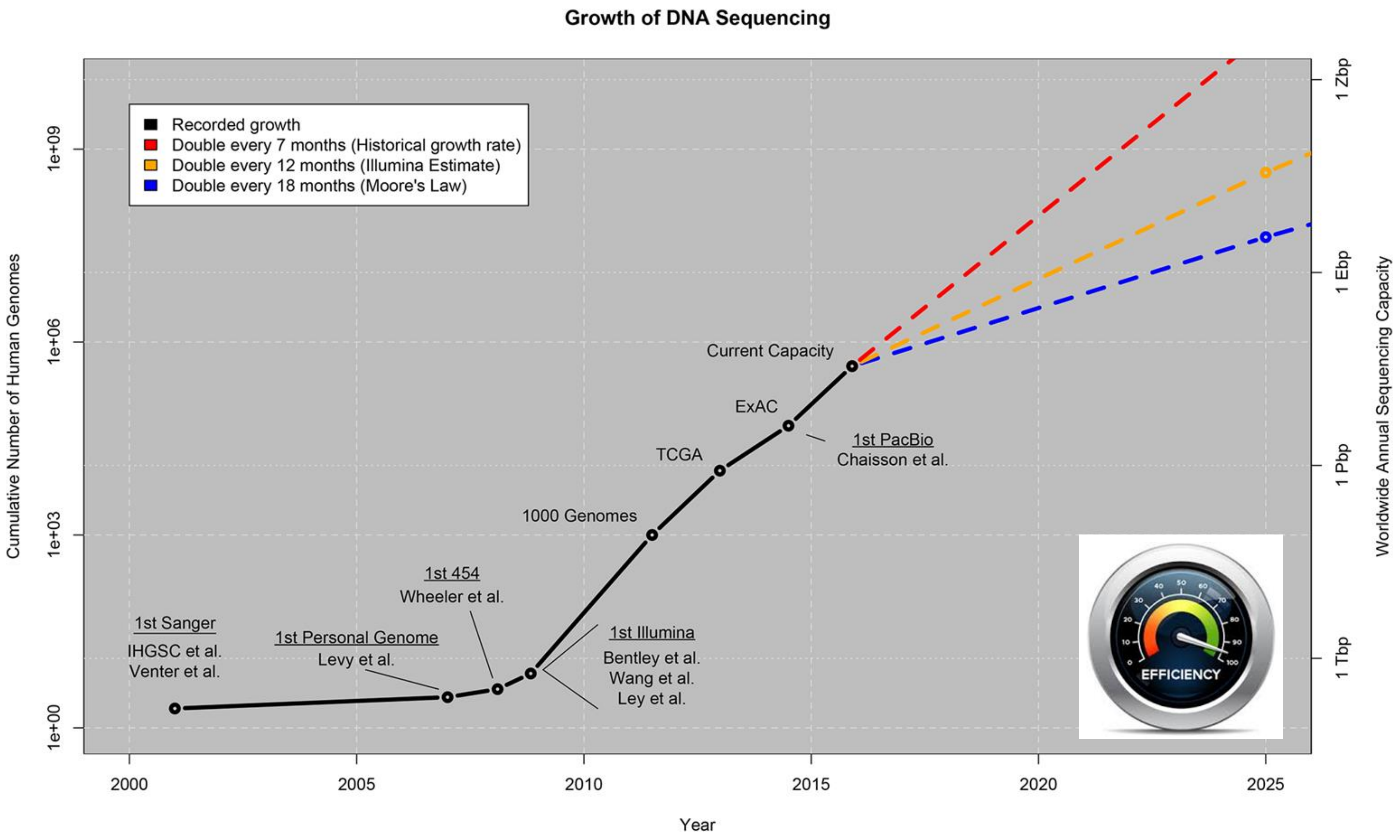
Topic 2 - Signal detection → (good genetic information)

Topic 3 - Complex workflow → (good decision making)

Concluding remarks

"Begin at the beginning," the King said, *"and go on till you come to the end: then stop."*

Lewis Carroll, Alice in Wonderland



Stephens, Z. D.et al. (2015). Big data: astronomical or genomical?. *PLoS biology*, *13*(7), e1002195.

- **Machine learning is** a major component of artificial intelligence (AI). ML is concerning with capturing specific patterns from data, often with the purpose of de-noising, classification and predictions for decision making.

No machine can be optimally efficient in more than one task

The example below postulates that one machine that climb stairs and make pancakes will be less efficient than two machines exclusively focus on climbing start and making pancakes, respectively.

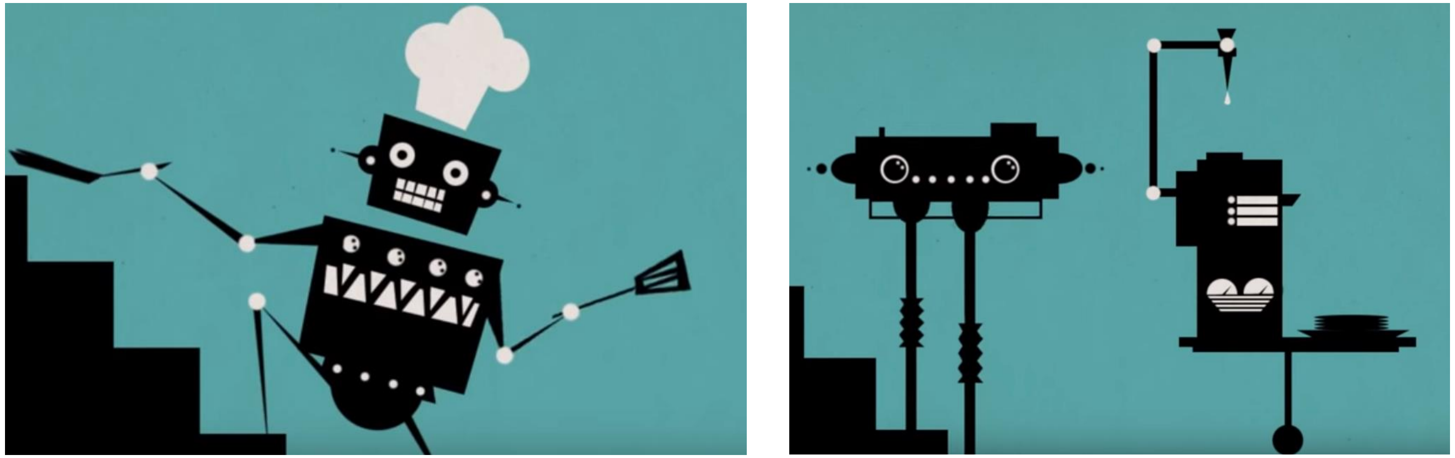


Figure source: https://www.youtube.com/watch?v=MPR3o6Hnf2g

Where ML is being deployed on plant breeding

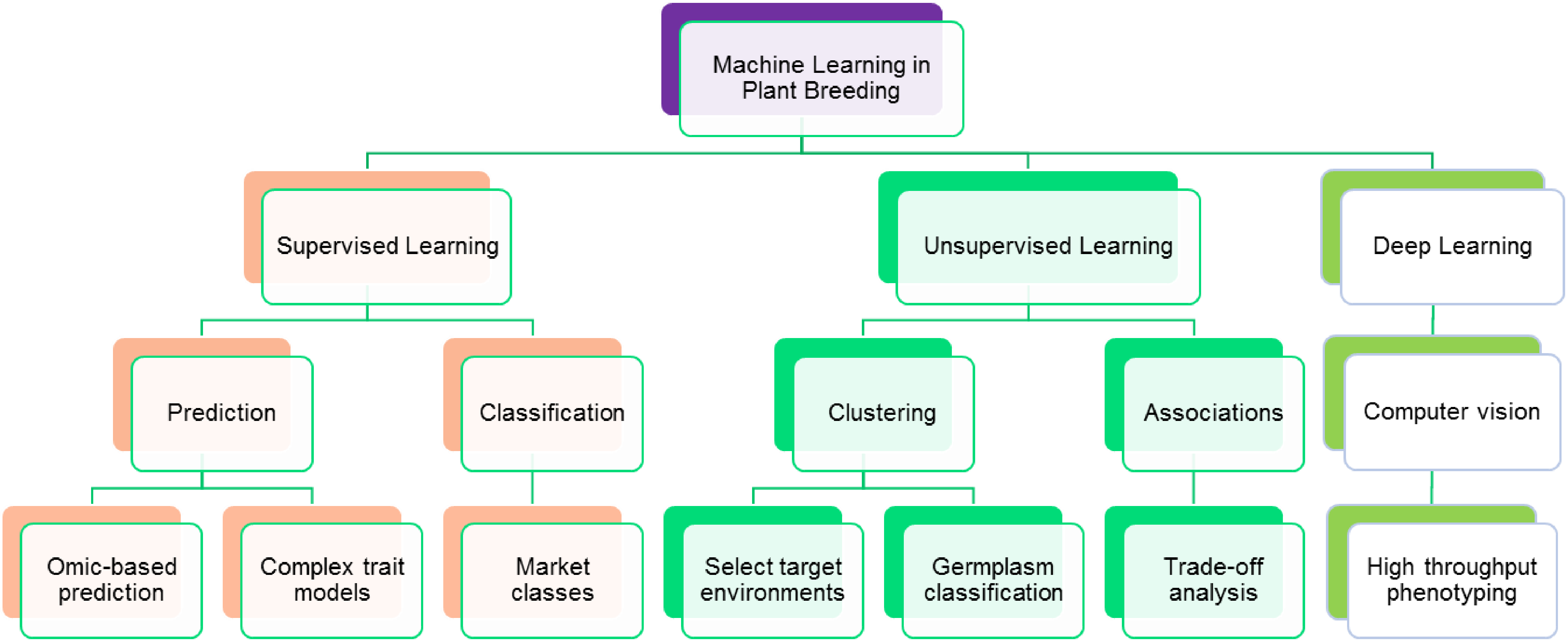
On phenotyping and phenotypic analysis

- Automated and precise scoring using drones and computer vision
- Discovery of (index and longitudinal) traits correlated to yield
- Phenotypic denoising: competition, environmental and spatial noise

On genotyping and genotypic analysis

- Unsupervised analysis: Imputation and germplasm classifications
- Early generation selection and recycling: Additive genetic signal
- Advanced generations selection and stability: G and GxE signal

Where is ML in plant breeding today?



Machine Learning in Plant Breeding

Supervised Learning

Prediction

Omic-based prediction

Complex trait models

Classification

Market classes

Unsupervised Learning

Clustering

Select target environments

Germplasm classification

Associations

Trade-off analysis

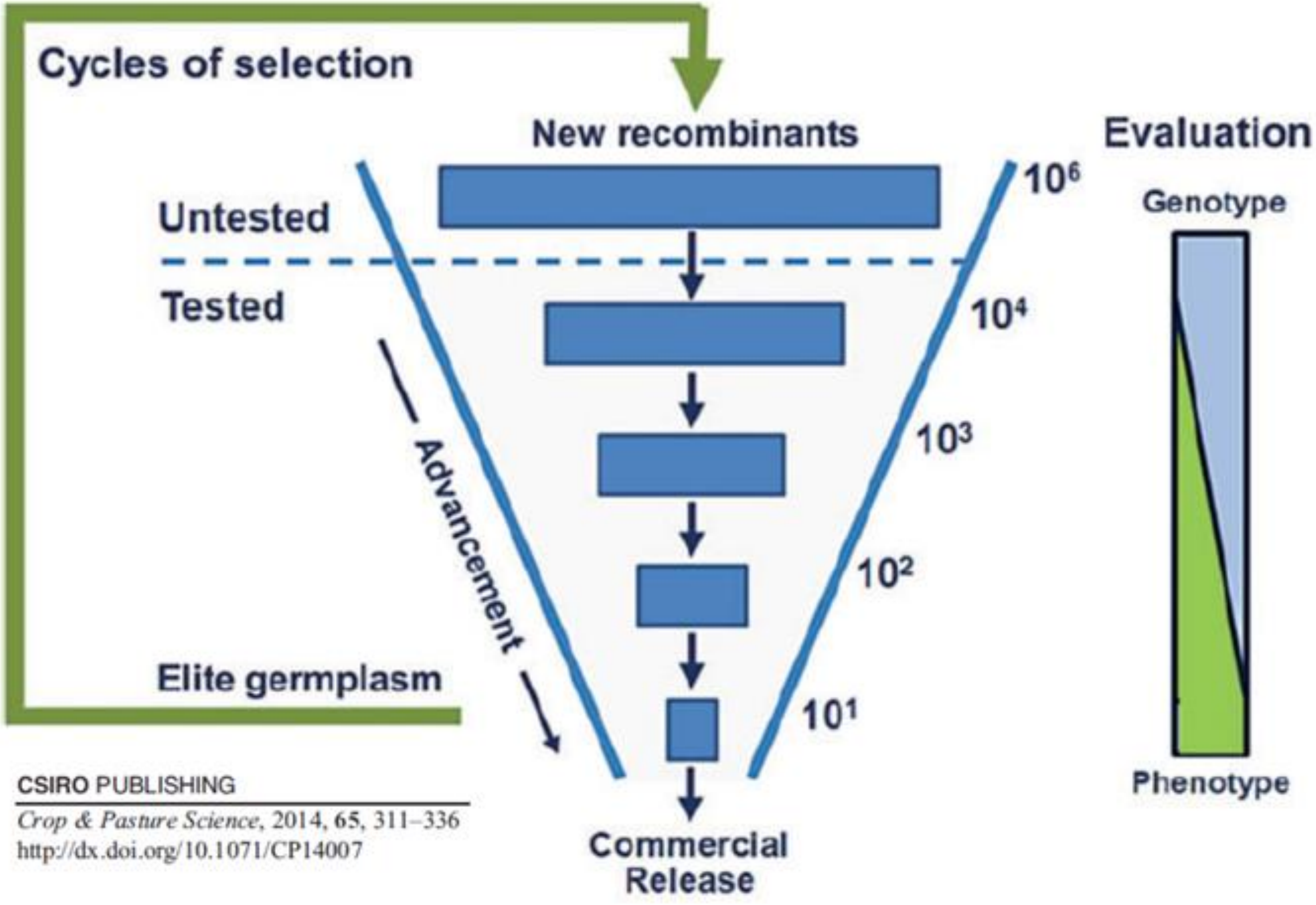
Deep Learning

Computer vision

High throughput phenotyping

1. Data processing

Where does ML fit in the breeding pipeline?



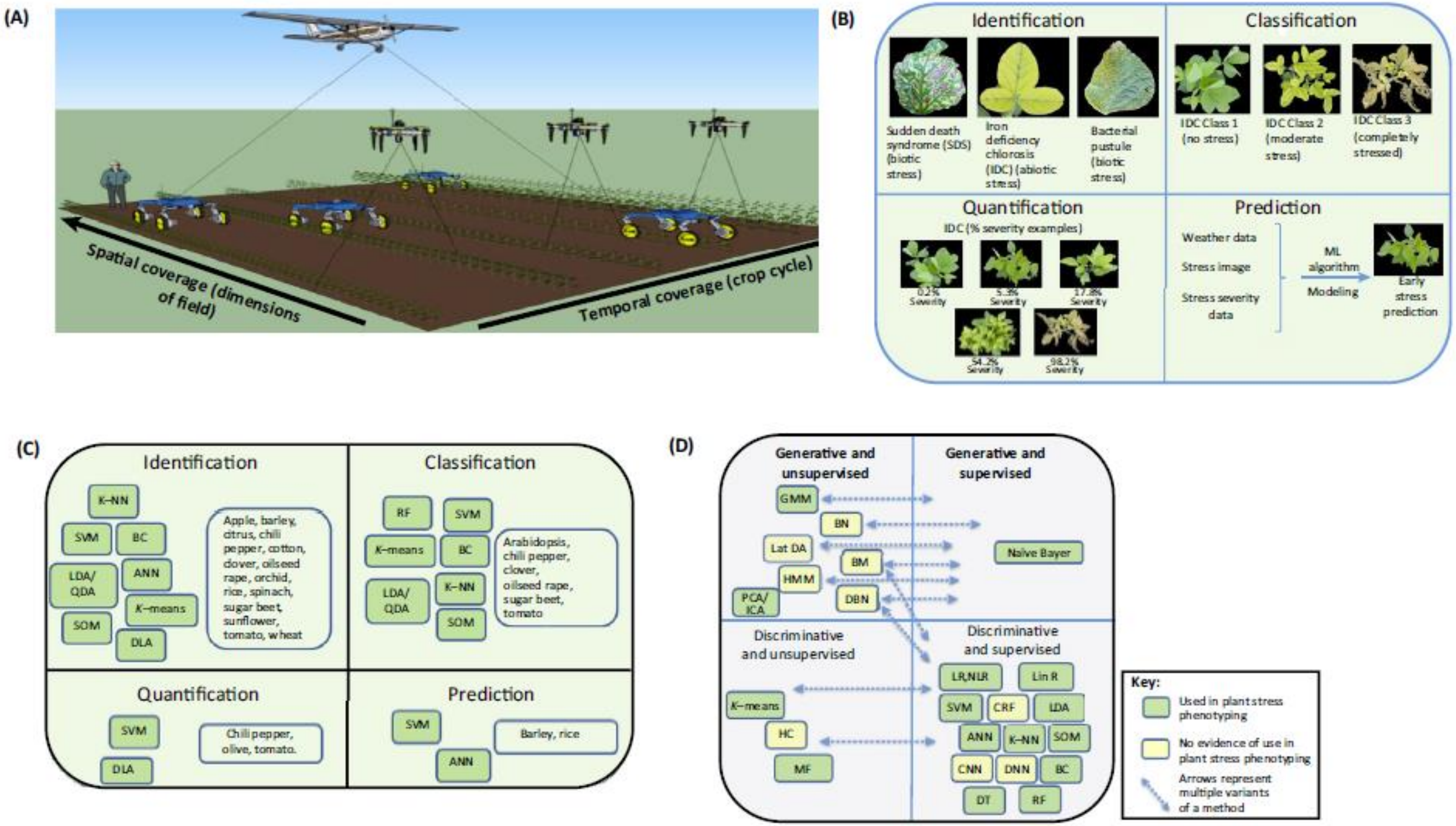
CSIRO PUBLISHING

Crop & Pasture Science, 2014, **65**, 311–336
http://dx.doi.org/10.1071/CP14007

Pipeline complexity (highly computational)

Key Figure Trends in Plant Science, February 2016, Vol. 21, No. 2 111

Machine Learning (ML) Tools for High-Throughput Stress Phenotyping


(A)
Spatial coverage (dimensions of field)
Temporal coverage (crop cycle)
(B)
Identification
Sudden death syndrome (SDS) (biotic stress)
Iron deficiency chlorosis (IDC) (abiotic stress)
Bacterial pustule (biotic stress)
Classification
IDC Class 1 (no stress)
IDC Class 2 (moderate stress)
IDC Class 3 (completely stressed)
Quantification
IDC (% severity examples)
0.2% Severity
5.3% Severity
17.8% Severity
54.2% Severity
98.2% Severity
Prediction
Weather data
Stress image
Stress severity data
ML algorithm
Modeling
Early stress prediction
(C)
Identification
K–NN
SVM
BC
LDA/ QDA
ANN
K–means
SOM
DLA
Apple, barley, citrus, chili pepper, cotton, clover, oilseed rape, orchid, rice, spinach, sugar beet, sunflower, tomato, wheat
Classification
RF
SVM
K–means
BC
LDA/ QDA
K–NN
SOM
Arabidopsis, chili pepper, clover, oilseed rape, sugar beet, tomato
Quantification
SVM
DLA
Chili pepper, olive, tomato.
Prediction
SVM
ANN
Barley, rice
(D)
Generative and unsupervised
GMM
BN
Lat DA
BM
HMM
PCA/ ICA
DBN
Generative and supervised
Naive Bayer
Discriminative and unsupervised
K–means
HC
MF
Discriminative and supervised
LR,NLR
Lin R
SVM
CRF
LDA
ANN
K–NN
SOM
CNN
DNN
BC
DT
RF
Key:
Used in plant stress phenotyping
No evidence of use in plant stress phenotyping
Arrows represent multiple variants of a method

Pipeline complexity (highly computational)

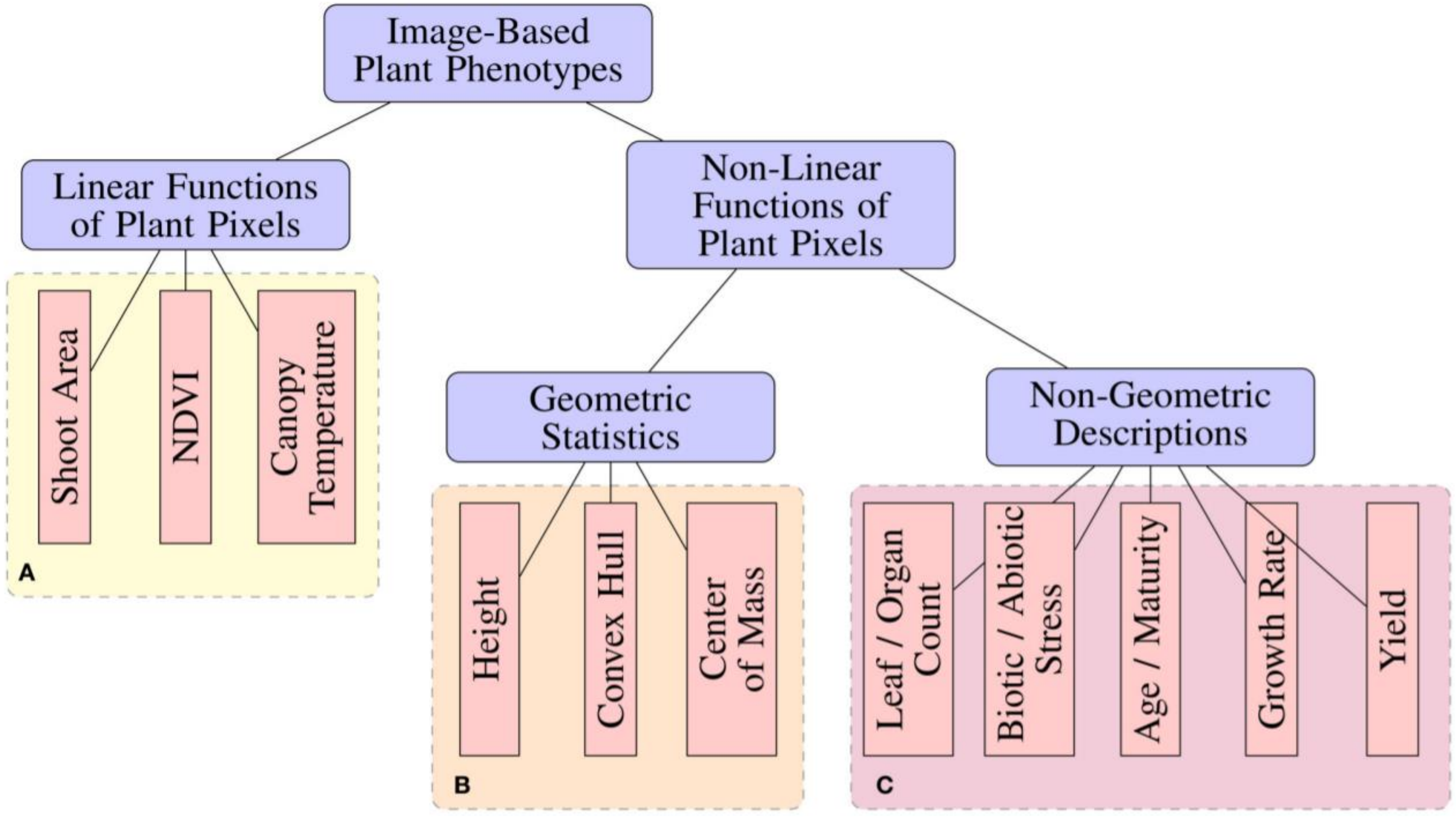
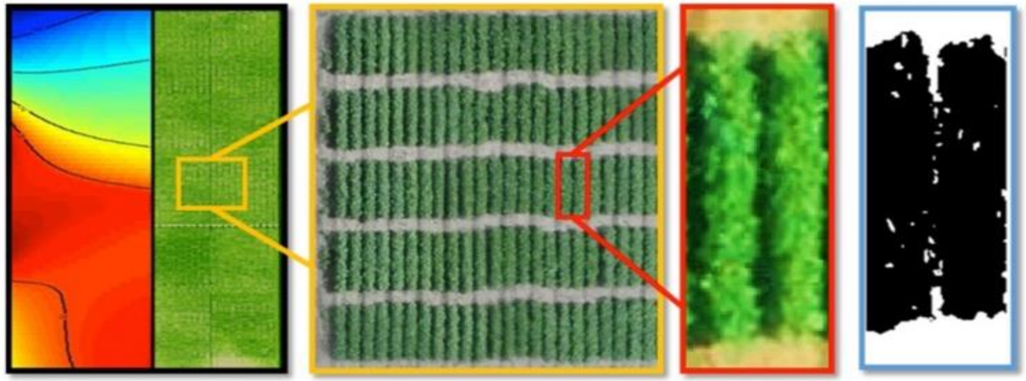
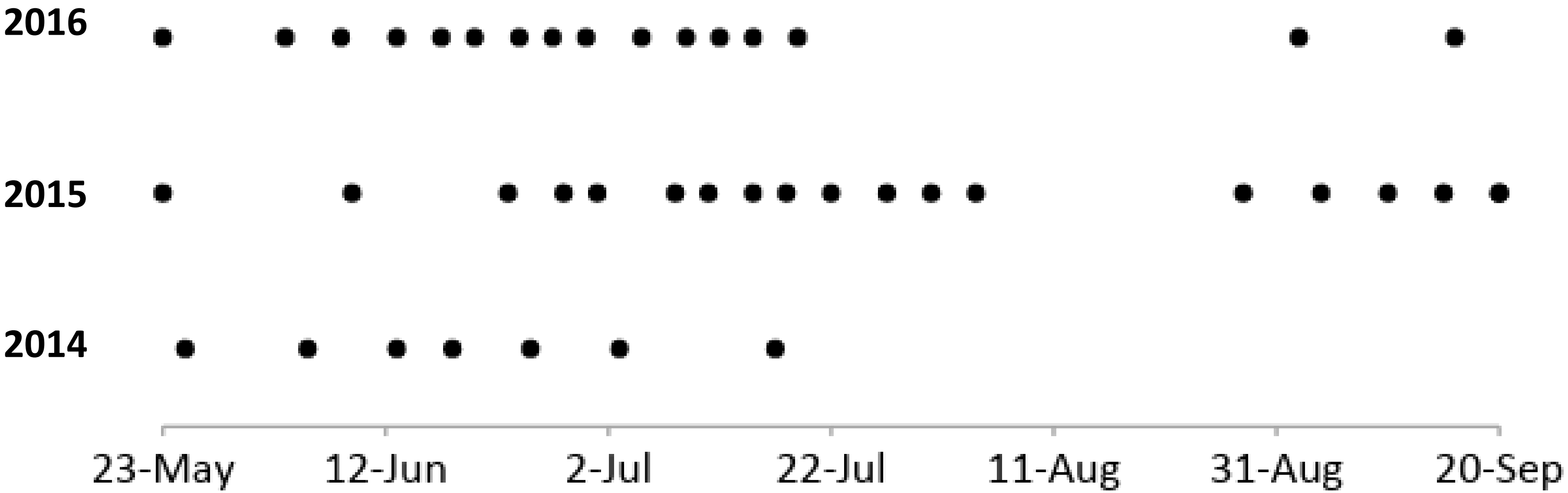


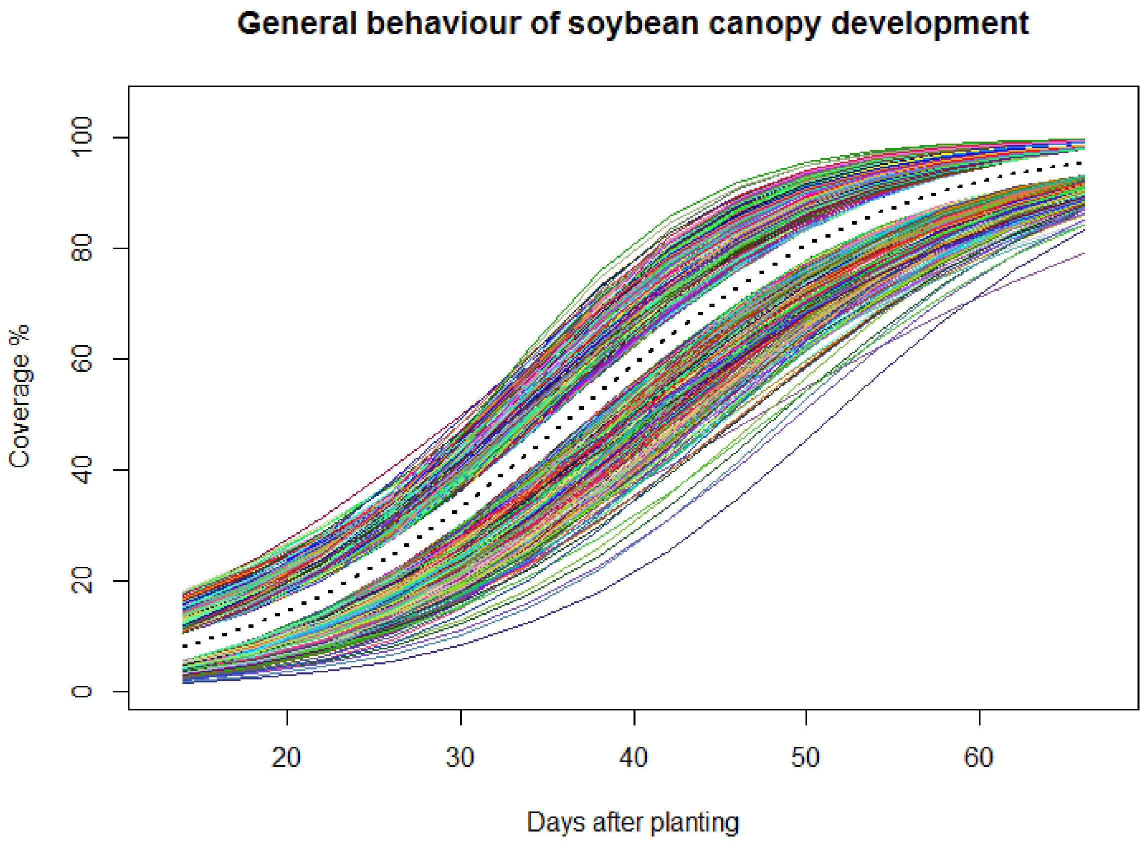
FIGURE 1 | A taxonomy of image-based plant phenotyping tasks. Phenotypes in **(A,B)** can be accurately measured using classical image processing techniques (Fahlgren et al., 2015b), while those in **(C)** are complex phenotyping tasks requiring more sophisticated analyses.

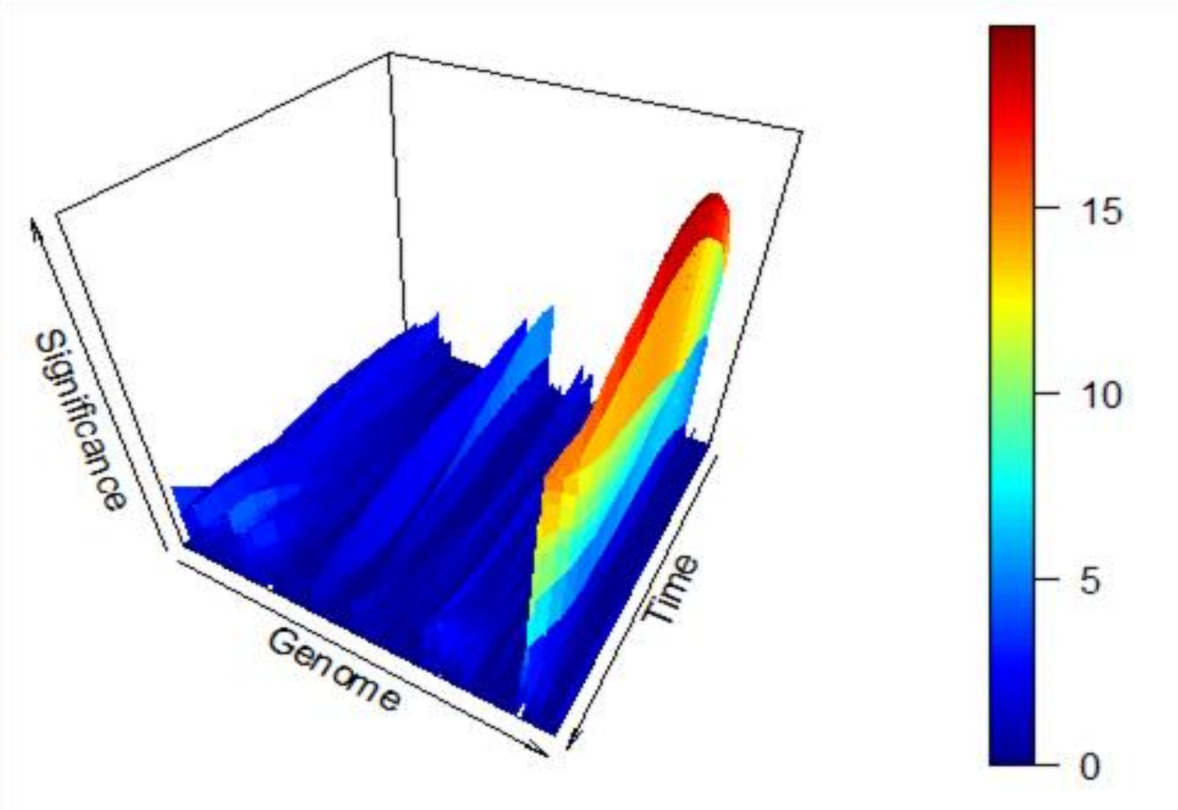
Ubbens, J. R., & Stavness, I. (2017). Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. Frontiers in plant science, 8, 1190.



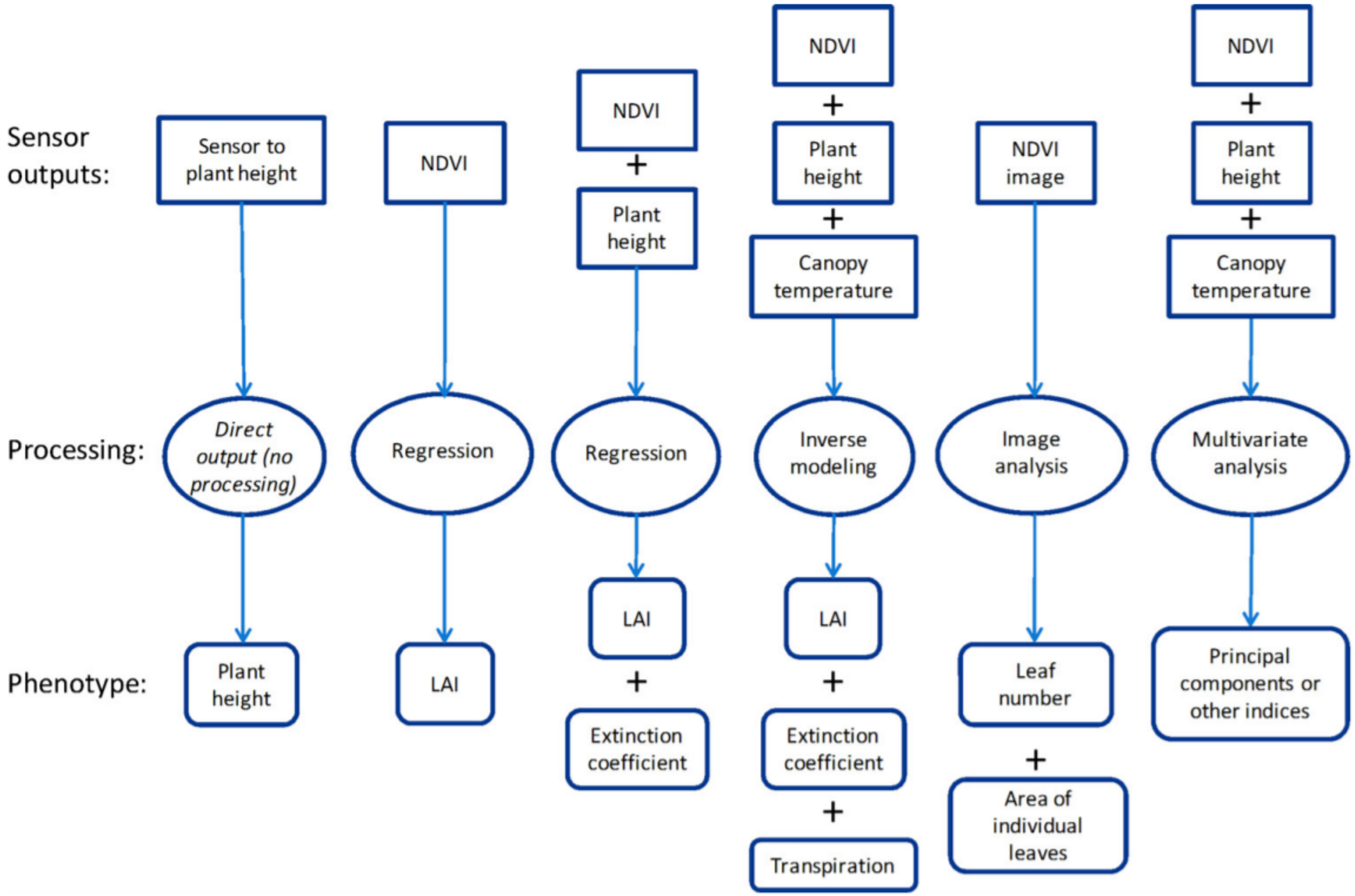

2016
2015
2014
23-May
12-Jun
2-Jul
22-Jul
11-Aug
31-Aug
20-Sep

Different time-points, different genetic architecture





Number and quality of phenotypes increase



J.W. White et al. / Field Crops Research 133 (2012) 101–112

Fig. 6. Examples of possible paths of data analysis whereby field measurements are processed to provide more biologically meaningful data. Field data usually would be recorded as time series, allowing estimation of growth or developmental rates.

2. Signal detection

2.1 Robust machines

2.2 Capturing signals

Robust machines

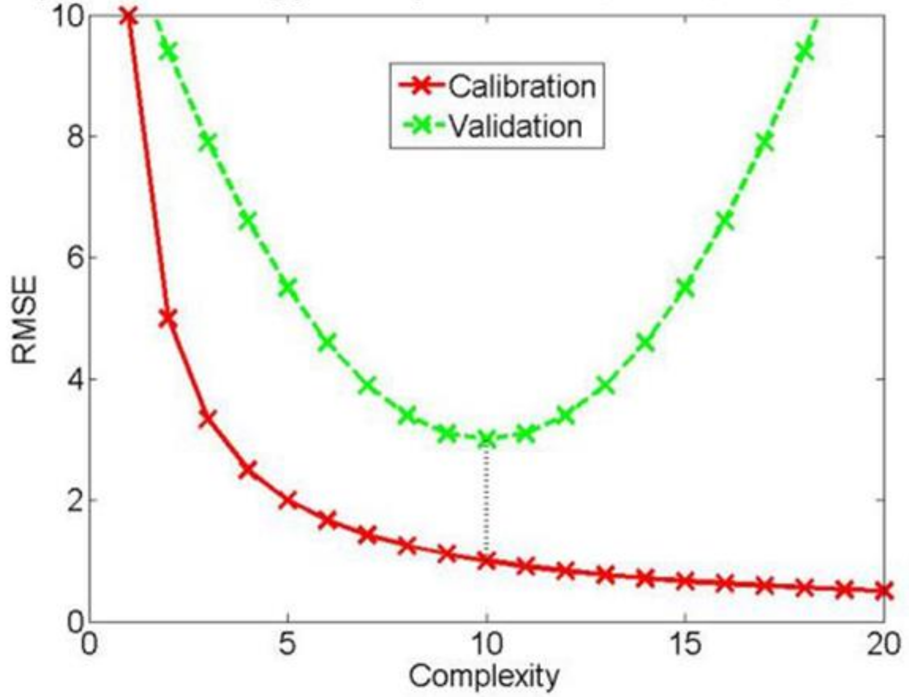
1. **Regularization**

2. **Parsimony**

3. **Interactions**

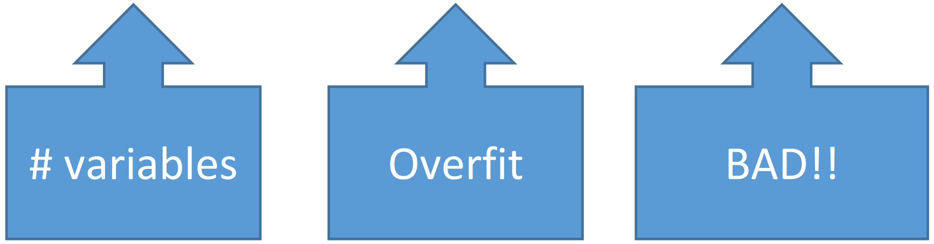
1) Overfitting: Complexity-Variance tradeoff

dynamicecology.wordpress.com/2013/08/19/

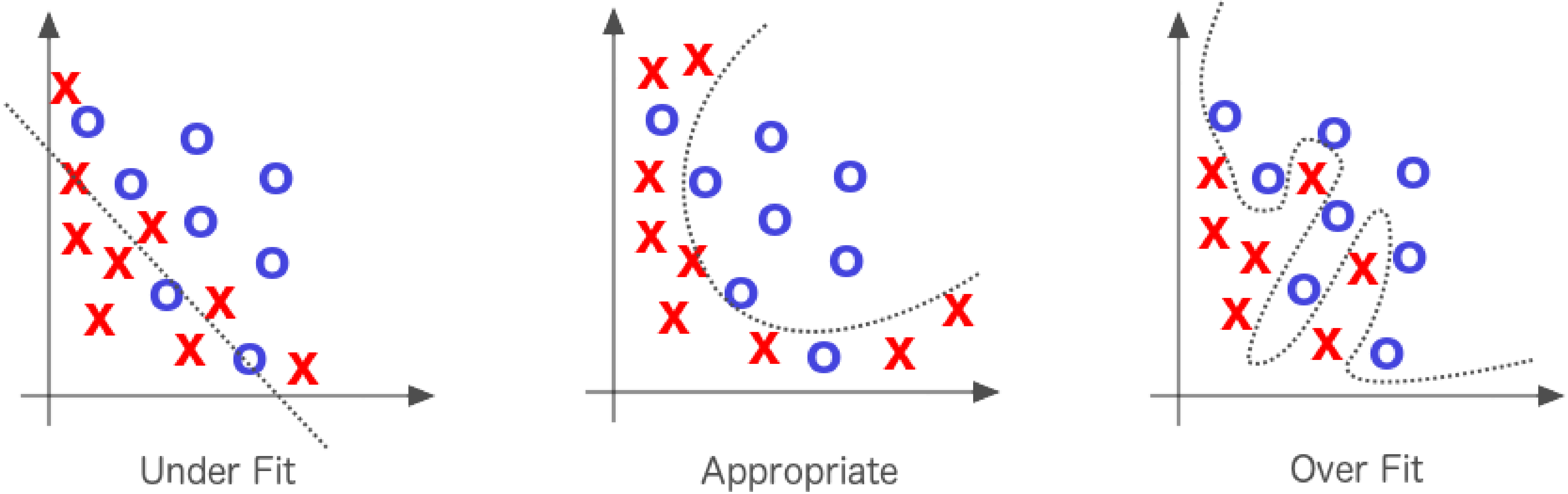


$$\lambda \uparrow = Fitness\ (R^2)\ \downarrow = Number\ of\ SNPs \downarrow = Shrinkage \uparrow$$

$$\uparrow Complexity = \downarrow Variance = \downarrow Prediction$$



Overfitting (kNN example)



Under Fit

Appropriate

Over Fit

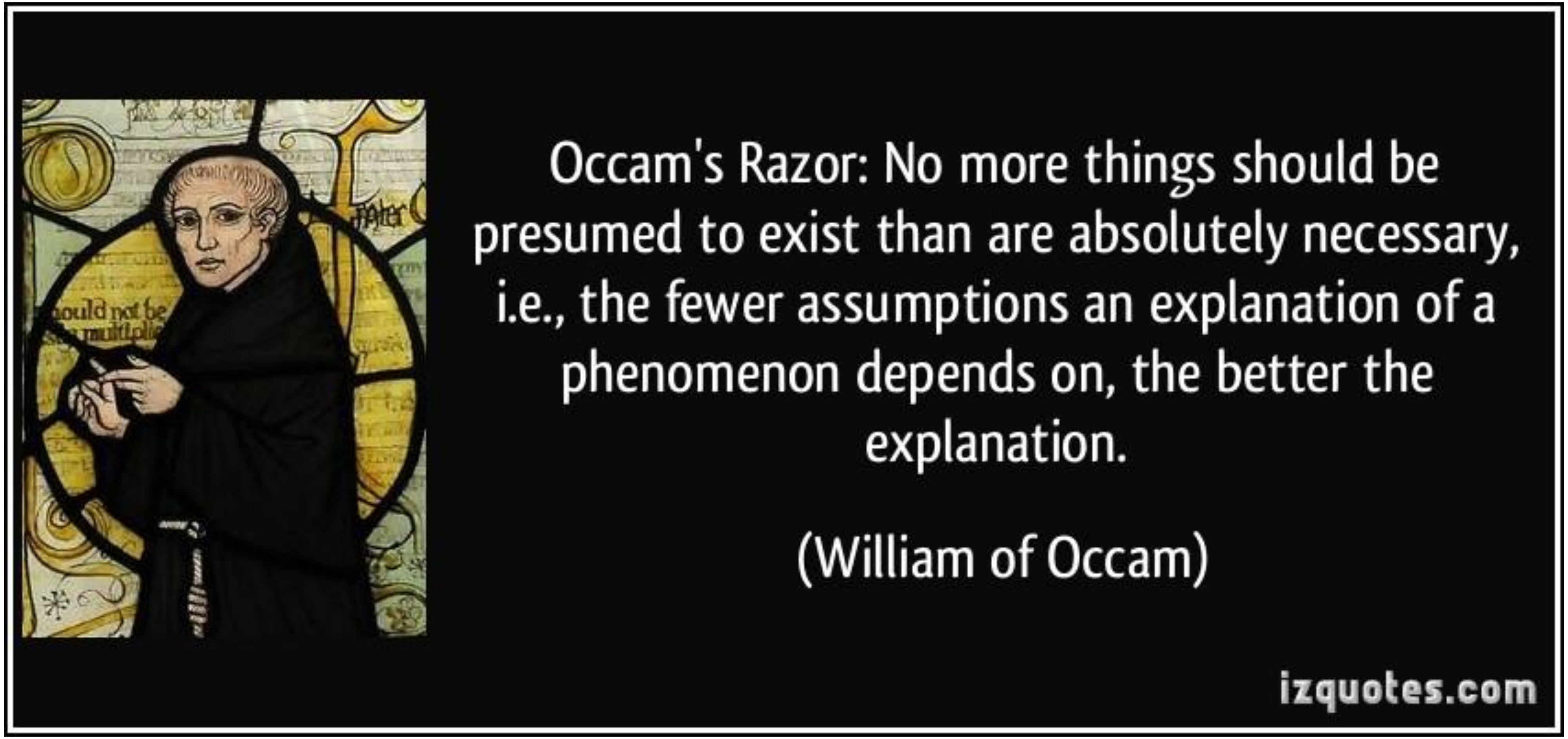
https://www.safaribooksonline.com/library/view/deep-learning/9781491924570/ch01.html

2) Parsimony

- **Occam's razor** – A simpler explanation is better than a complex one
  - The less terms you have in your model, the better
  - An attempt to comprise most information with the least amount of factors

- **Example (model for genetic values). Check the two models:**
  1) **Yield = Block + Location + Year + Genotype + (Genotype x Year x Location) + …**
  2) **Yield = Block + Genotype**

- **Genomics: Models with too many parameters (p>>n)**
  - Most high-dimensional machines automatically perform variable selection



Occam's Razor: No more things should be presumed to exist than are absolutely necessary, i.e., the fewer assumptions an explanation of a phenomenon depends on, the better the explanation.

(William of Occam)

izquotes.com

3) Hierarchical principle

- **CLAIM**: Lower order effects more important than higher order effects
- Effects of same order equally important

POWER IS AN ISSUE TO DETECT SIGNAL OF INTERACTIONS!!!!

- Consider the type of the variable (continuous or categorical)
- This principle makes one wonder about the relevance of **Epistasis** and **GxE**
- Higher order terms are good to **run out of degrees of freedom**

Capturing signals

1. **Signal extraction**

2. **Supervised machines**

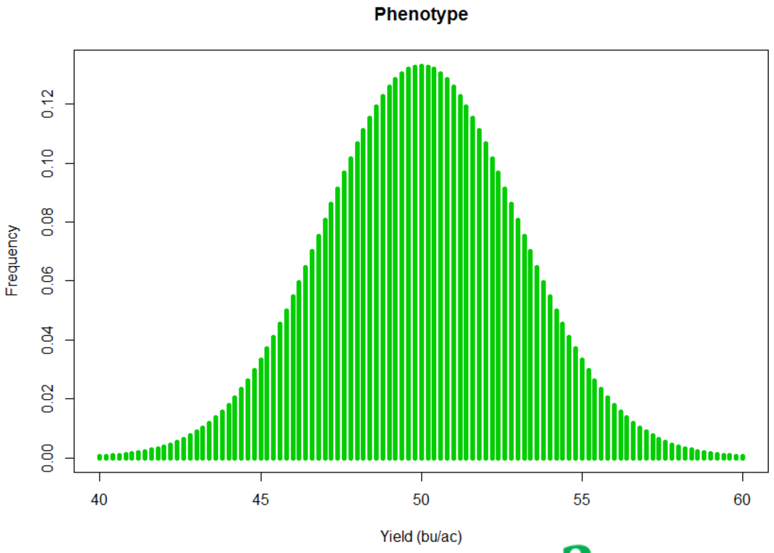
3. **Multiple signals**

Gaussian process: Signal and noise

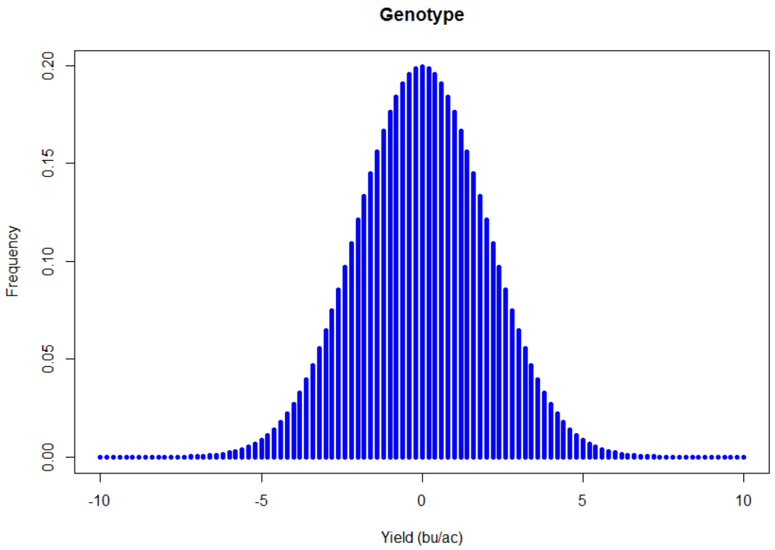
$$y = 1\mu + g + e$$

$$\sigma_y^2 = \sigma_g^2 + \sigma_e^2$$

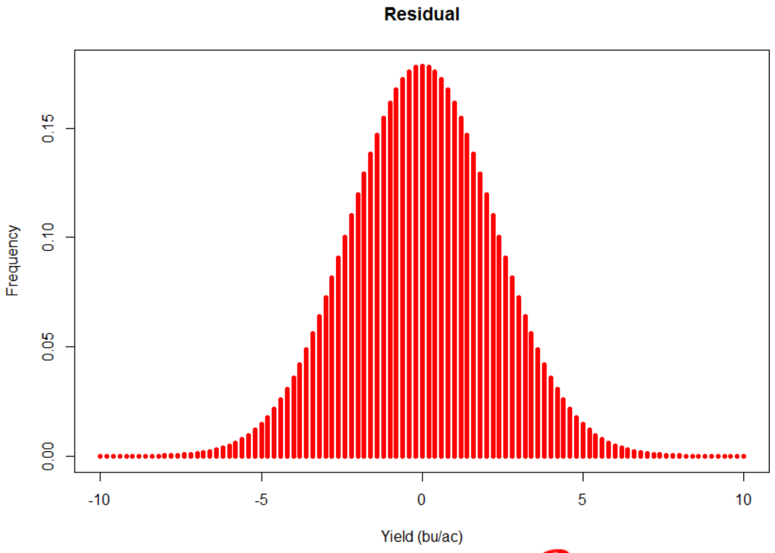
$$\mu = 50$$
$$\sigma_y^2 = 9$$
$$\sigma_g^2 = 4$$
$$\sigma_e^2 = 5$$
$$cov(g, e) = 0$$



$$y \sim N(\mu, I\sigma_y^2)$$



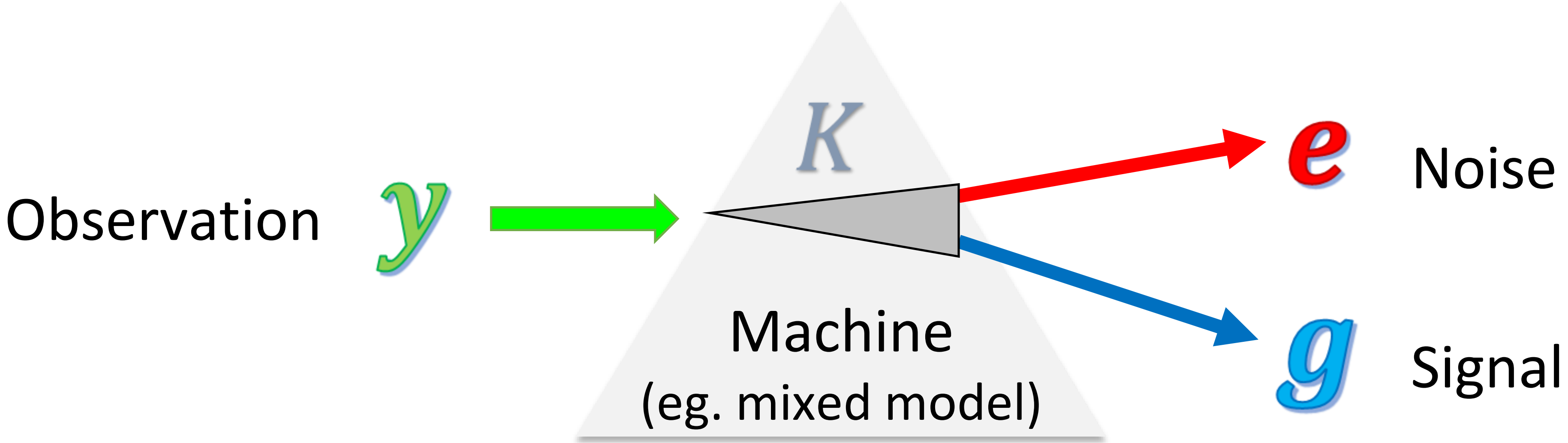
$$u \sim N(0, I\sigma_g^2)$$



$$e \sim N(0, I\sigma_e^2)$$

Distinction of signal and noise

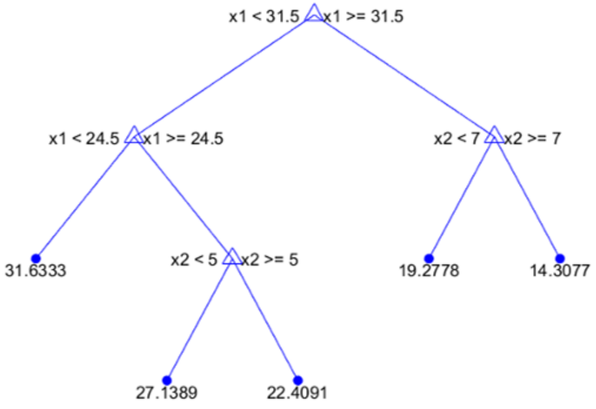
$$y = \mu + g + e$$



Most common supervised machines

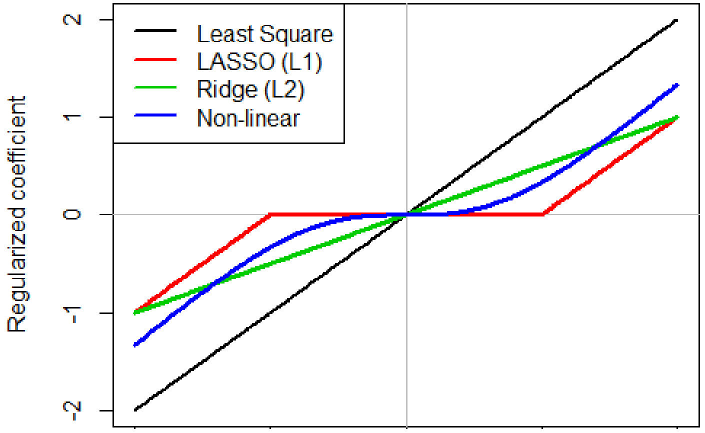
Regression trees

(eg. simplified decision model)



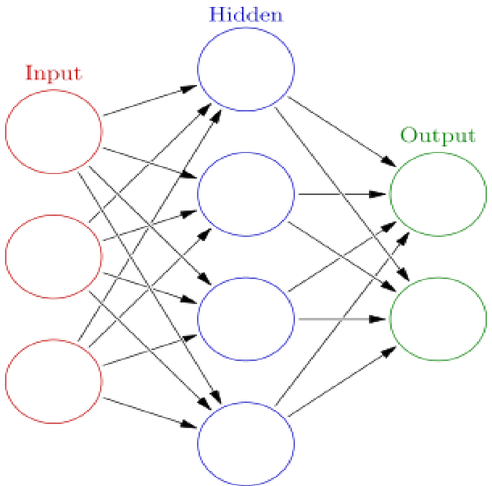
Penalized regression

(eg. random effects in mixed models)



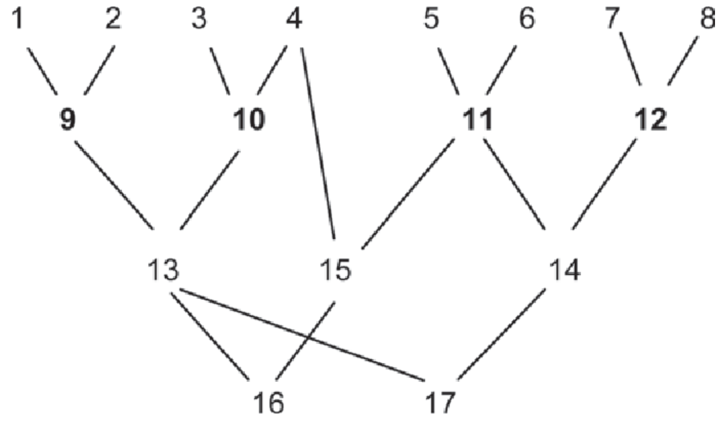
Neural networks

(eg. Google search engine)

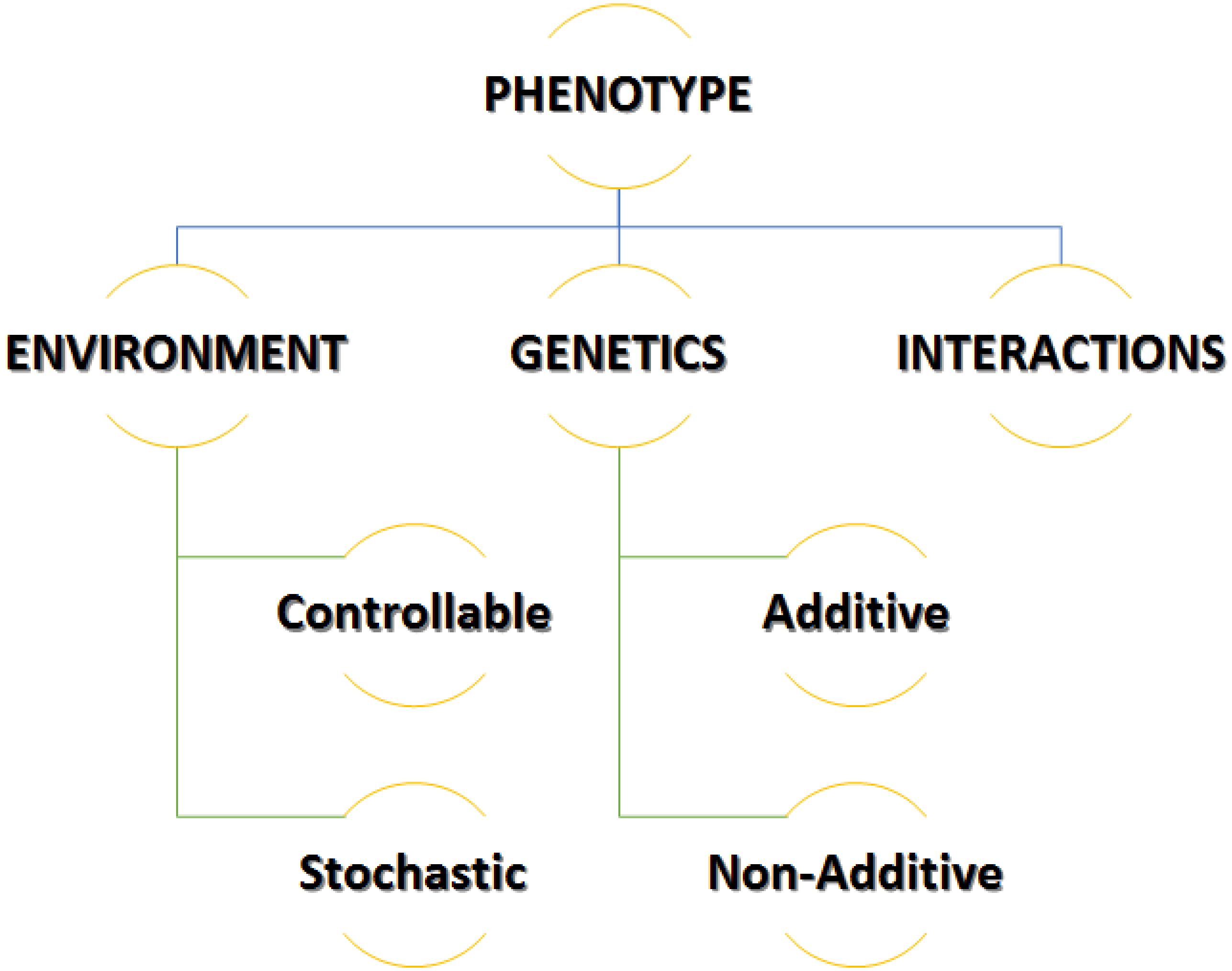


Kernel regression

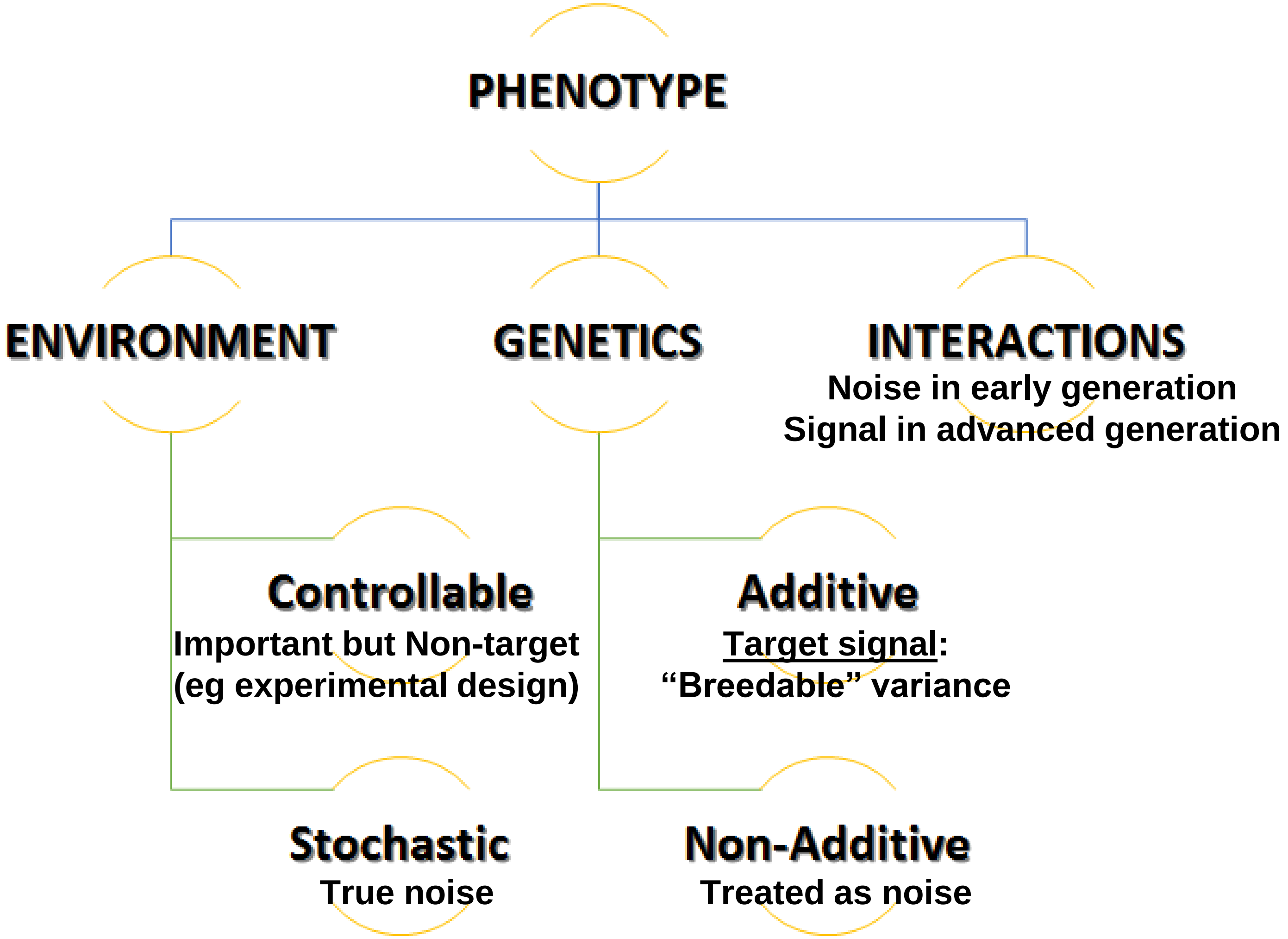
(eg. GBLUP and GPS interpolation)



Multiple signals



Multiple signals



PHENOTYPE

ENVIRONMENT

GENETICS

INTERACTIONS

Noise in early generation
Signal in advanced generation

Controllable

Important but Non-target
(eg experimental design)

Additive

Target signal:
"Breedable" variance

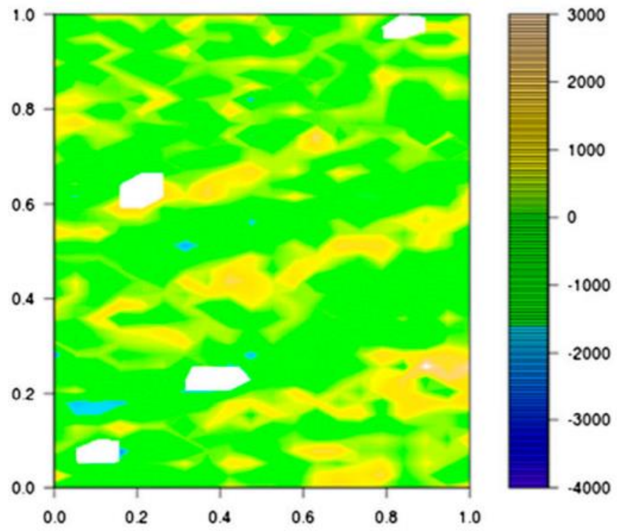
Stochastic

True noise

Non-Additive

Treated as noise

Case of non-target signal (Lado et al. 2013)



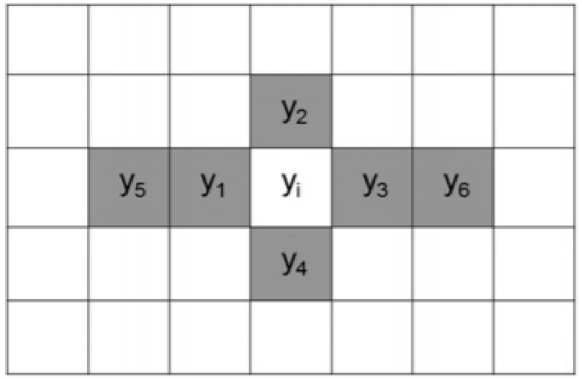


Figure 1 Diagram to calculate the covariable x_i. Y_i is the phenotypic value in the plot. The neighboring plots are indicated with gray color.

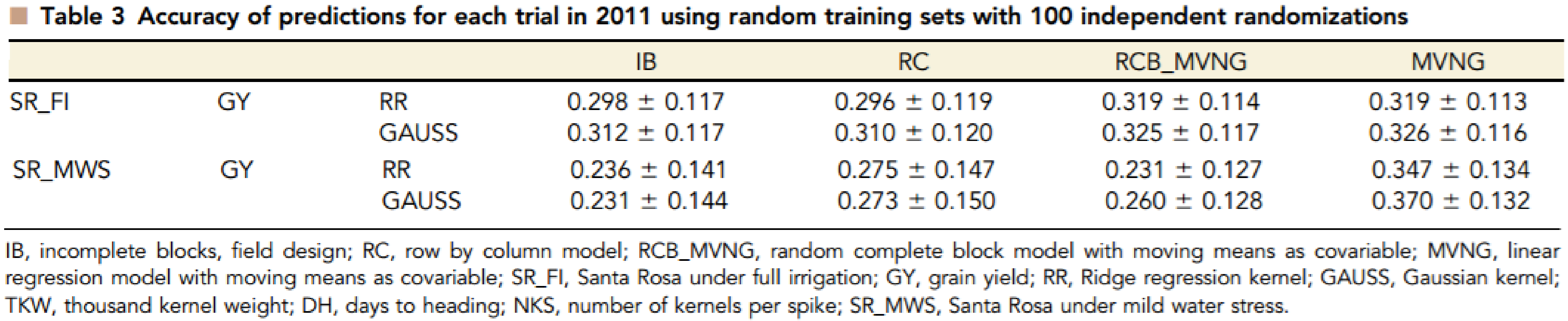


Table 3 Accuracy of predictions for each trial in 2011 using random training sets with 100 independent randomizations

			IB	RC	RCB_MVNG	MVNG
SR_FI	GY	RR	0.298 ± 0.117	0.296 ± 0.119	0.319 ± 0.114	0.319 ± 0.113
		GAUSS	0.312 ± 0.117	0.310 ± 0.120	0.325 ± 0.117	0.326 ± 0.116
SR_MWS	GY	RR	0.236 ± 0.141	0.275 ± 0.147	0.231 ± 0.127	0.347 ± 0.134
		GAUSS	0.231 ± 0.144	0.273 ± 0.150	0.260 ± 0.128	0.370 ± 0.132

IB, incomplete blocks, field design; RC, row by column model; RCB_MVNG, random complete block model with moving means as covariable; MVNG, linear regression model with moving means as covariable; SR_FI, Santa Rosa under full irrigation; GY, grain yield; RR, Ridge regression kemel; GAUSS, Gaussian kemel; TKW, thousand kernel weight; DH, days to heading; NKS, number of kemels per spike; SR_MWS, Santa Rosa under mild water stress.

3. Complex workflow

Phenomics & genomics are usually tight together

Review

Trends in Plant Science December 2011, Vol. 16, No. 12

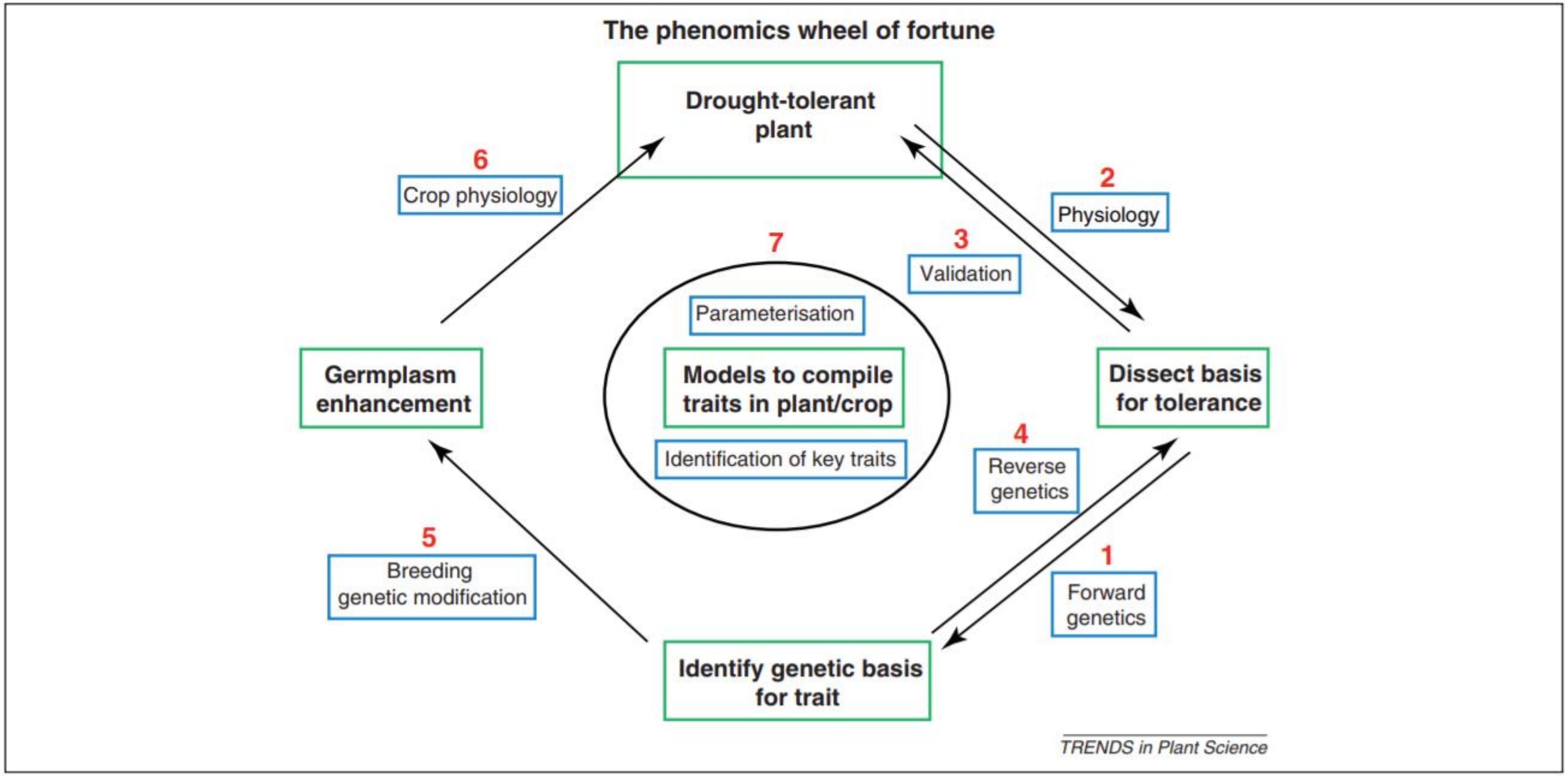

The phenomics wheel of fortune
Drought-tolerant plant
2
Physiology
3
Validation
Dissect basis for tolerance
4
Reverse genetics
1
Forward genetics
Identify genetic basis for trait
5
Breeding genetic modification
Germplasm enhancement
6
Crop physiology
7
Parameterisation
Models to compile traits in plant/crop
Identification of key traits
TRENDS in Plant Science

Figure 2. Closing the gene to genotype loop with phenomics.

Analysis of multiple traits

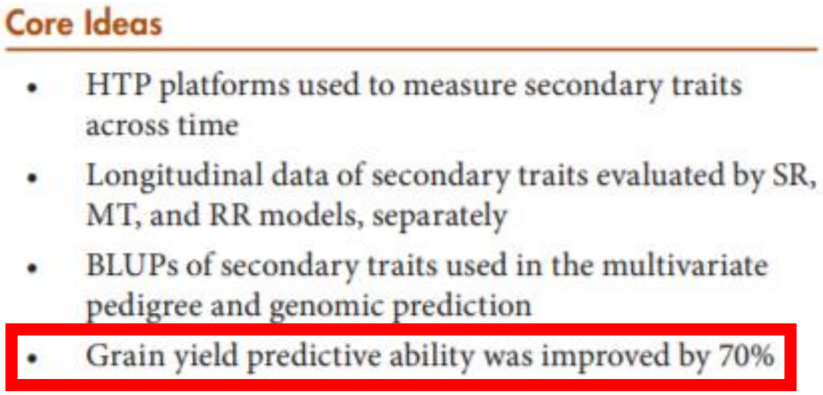
- In linear models: Covariate vs Multivariate
  - Covariate – on trait is used to predict the other. No strings attached to genetics.
  - Multivariate – modeling 2+ traits simultaneously. Genetics connected across traits.

- Computational burden increases exponentially $O(k)^7$ with the number of traits

- Multicollinearity
  - Many traits are nearly identical (eg. neighbor bandwidths from hyperspectral image)
  - OLS will not work due to singularities

- Modeling is often improved by accounting for time-space domains

Multivariate models



Core Ideas

- HTP platforms used to measure secondary traits across time
- Longitudinal data of secondary traits evaluated by SR, MT, and RR models, separately
- BLUPs of secondary traits used in the multivariate pedigree and genomic prediction
- Grain yield predictive ability was improved by 70%

Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J. L., & Sorrells, M. E. (2017). Multitrait, Random Regression, or Simple Repeatability Model in High-Throughput Phenotyping Data Improve Genomic Prediction for Wheat Grain Yield. *The Plant Genome*.

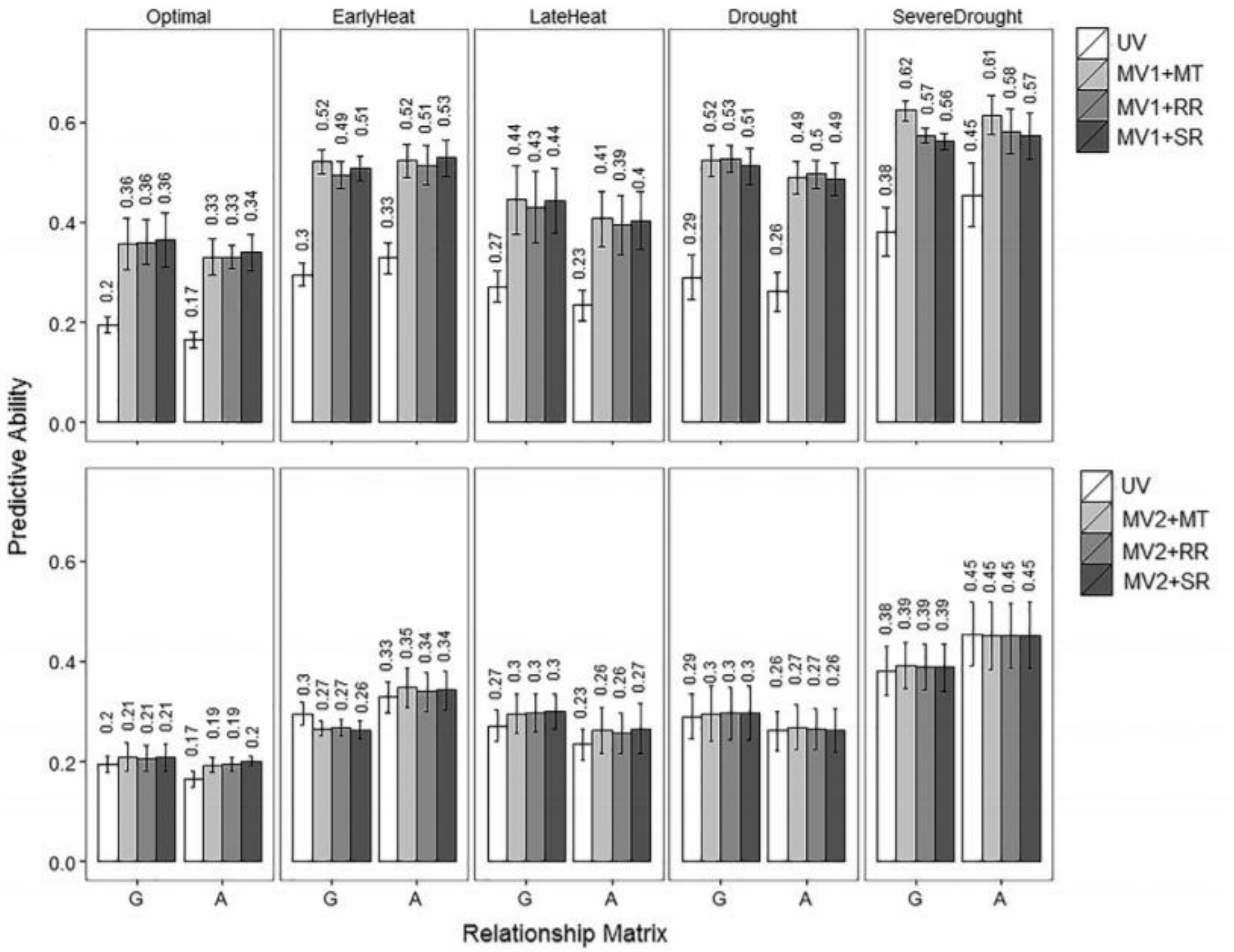
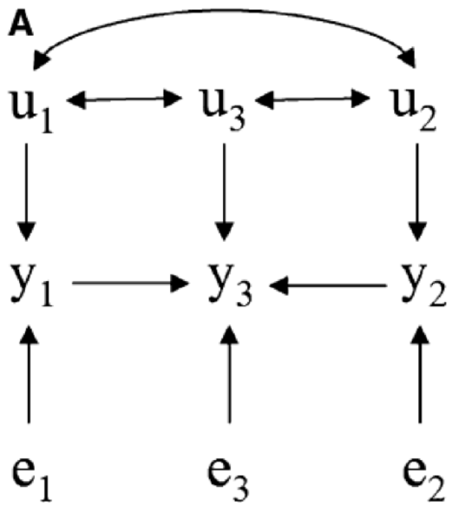


Fig. 1. Predictive ability comparison for grain yield between prediction models with secondary traits (MV1 and MV2) and without secondary trait (UV). MV1, multivariate prediction model with secondary traits in both training and testing populations; MV2, multivariate prediction model with secondary traits in training population only; UV, univariate prediction model with grain yield only; MT/RR/SR, multivariate prediction model MV1 or MV2 using best linear unbiased predictions (BLUPs) of secondary traits from multitrait (MT), random regression (RR), or simple repeatability (SR) model; *G/A*, genomic/pedigree relationship matrix.

Alternatives to dense multivariate models:

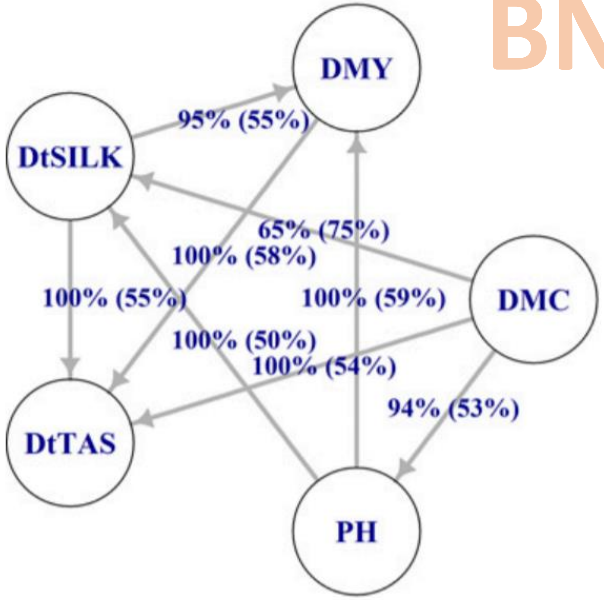
Structural Equation Models (SEM), Bayesian Networks (BN), and Markov Random Fields (MRF)

SEM



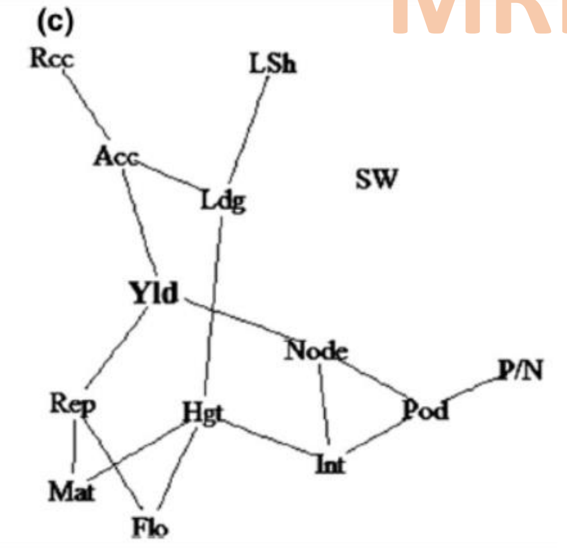
Valente, B. D., Rosa, G. J., de los Campos, G., Gianola, D., & Silva, M. A. (2010). Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*, *185*(2), 633-644.

BN



Töpner, K., Rosa, G. J., Gianola, D., & Schön, C. C. (2017). Bayesian Networks Illustrate Genomic and Residual Trait Connections in Maize (Zea mays L.). *G3: Genes, Genomes, Genetics*, *7*(8), 2779-2789.

MRF



Xavier, A., Hall, B., Casteel, S., Muir, W., & Rainey, K. M. (2017). Using unsupervised learning techniques to assess interactions among complex traits in soybeans. *Euphytica*, *213*(8), 200.

Better the signal, less important is the machine

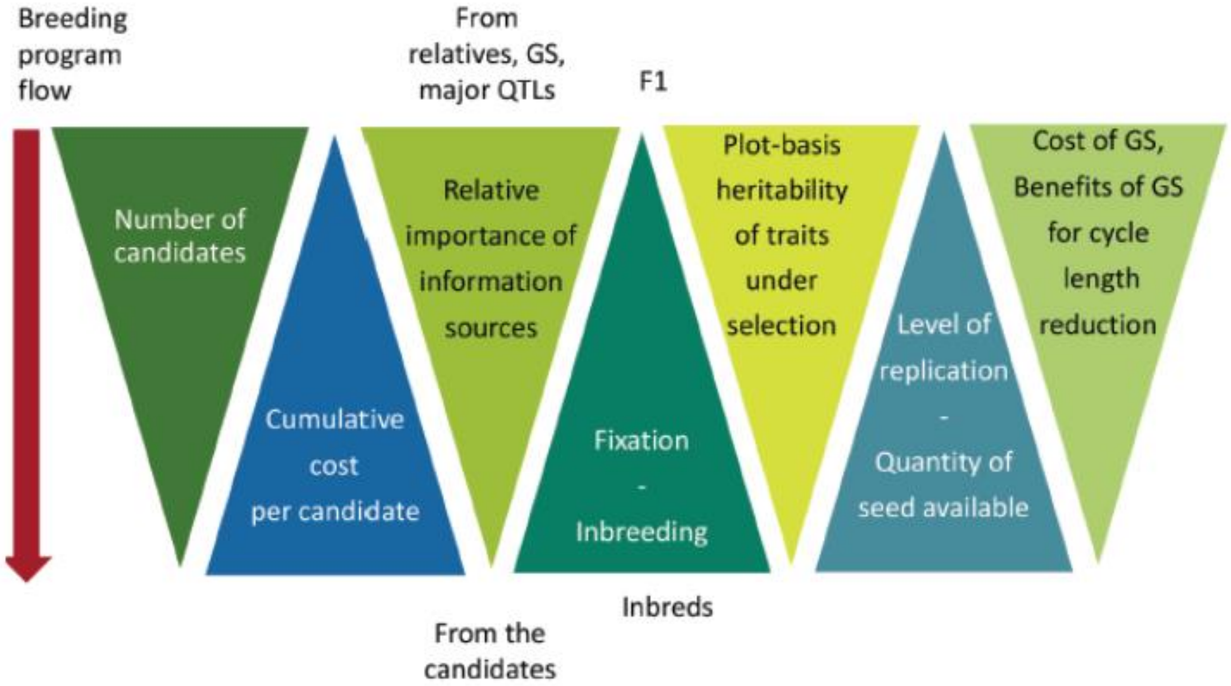
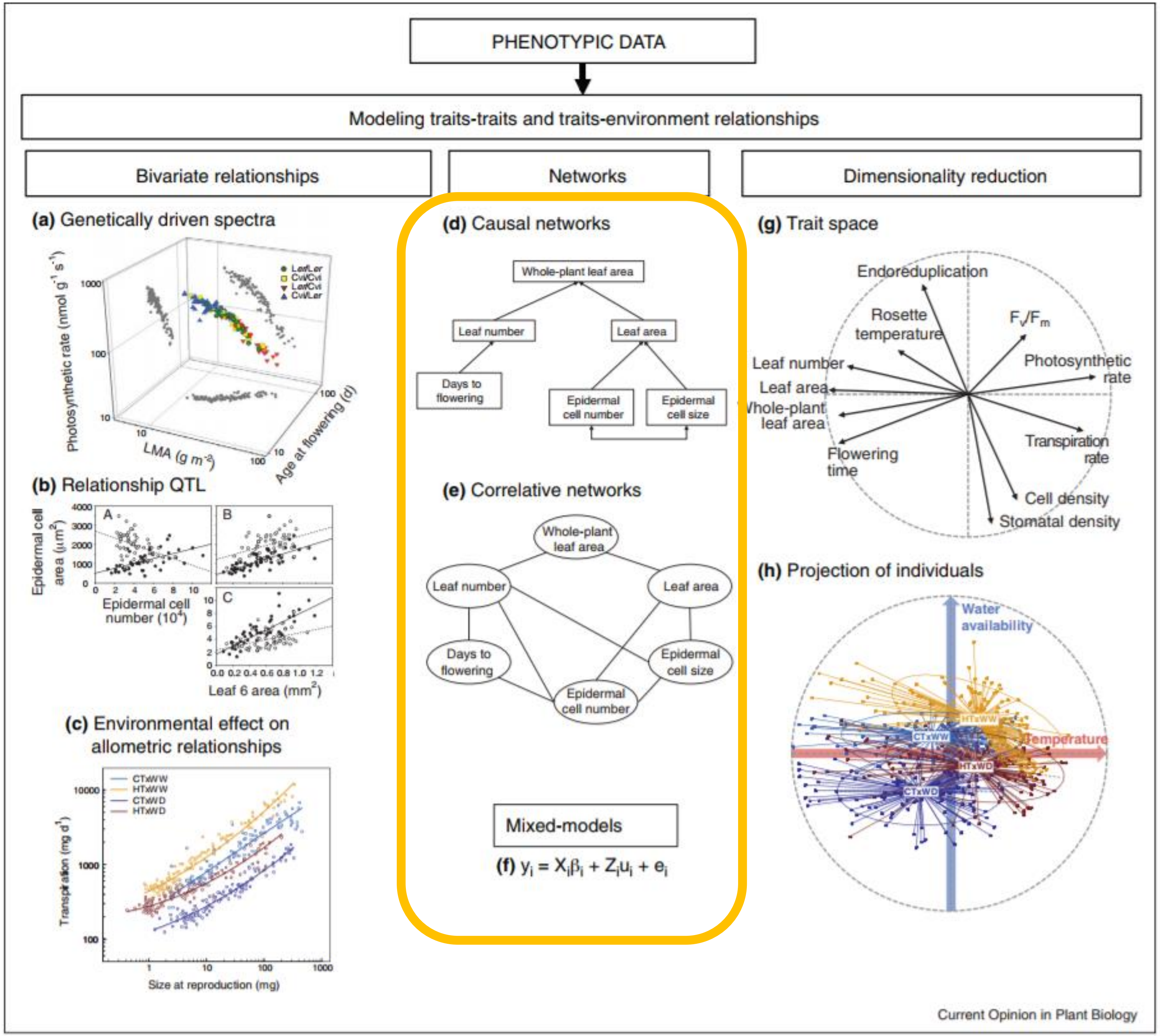

Breeding program flow
From relatives, GS, major QTLs
F1
Number of candidates
Cumulative cost per candidate
Relative importance of information sources
Fixation - Inbreeding
Plot-basis heritability of traits under selection
Level of replication - Quantity of seed available
Cost of GS, Benefits of GS for cycle length reduction
From the candidates
Inbreds

Figure 1. Key parameters and changes during a breeding cycle, to consider in implementing genomic selection (GS). The triangles indicate increase or decrease of the quantity considered. QTL, quantitative trait loci.

Heslot, N., Jannink, J. L., & Sorrells, M. E. (2015). Perspectives for genomic selection applications and research in plants. *Crop Science*, *55*(1), 1-12.



Granier, C., & Vile, D. (2014). Phenotyping and beyond: modelling the relationships between traits. *Current opinion in plant biology*, *18*, 96-102.

Concluding remarks

Revisiting key messages

- ★ **ML in hands-on breeding: HTPs, BLUPs and GEBVs**
  - ○ *Obtain better phenotypes and perform more accurate selections*
- ★ **Usage of the data relies on the nature of the signal**
  - ○ *Best results come from simple and mindful model*
- ★ **Breeding applications mostly on Gaussian process**
  - ○ *Most breeding problems can be tackled with mixed models*
- ★ **Machines are important when signal is scarce**
  - ○ *Genomic breeding suits best modeling early generations*

Acknowledgements

Funding
- USB (Genotyping, experiments, phenotyping), Corteva (experiment and phenotyping), USDA and NSF (students)

Genetic resource and genotyping
- USDA (Perry Cregan, Qijian Song), UNL (James Specht), UofIL (Brian Diers, Randy Nelson)

Drone data
- Purdue ABE (Keith Cherkauer, Anthony Hearst)

Modeling
- Purdue (William Muir), UC Riverside (Shizhong Xu), ISU (William Beavis, Vishnu Ramasubramanian), UNL (Diego Jarquin, Reka Howard), UMN (Aaron Lorenz)

Phenotypes
- Purdue (Katy Rainey, Ben Hall), UofIL (Randy Nelson, Brian Diers), UNL (James Specht, George Graeg), ISU (William Beavis), OSU (Leah McHale), KSU (William Schapaugh), MSU (Dechun Wang), UofMO (Grover Shannon), NCSU (Rouf Mian)

Support
- Corteva (Tabare Abadie, Radu Totir, Mak Geha, David Habier)

That’s all!

Thanks!

Questions?

Purdue Soybean Breeding

AX102418

http://alenxav.wix.com/home

Corteva Biostatistics Whole Genome Analytics

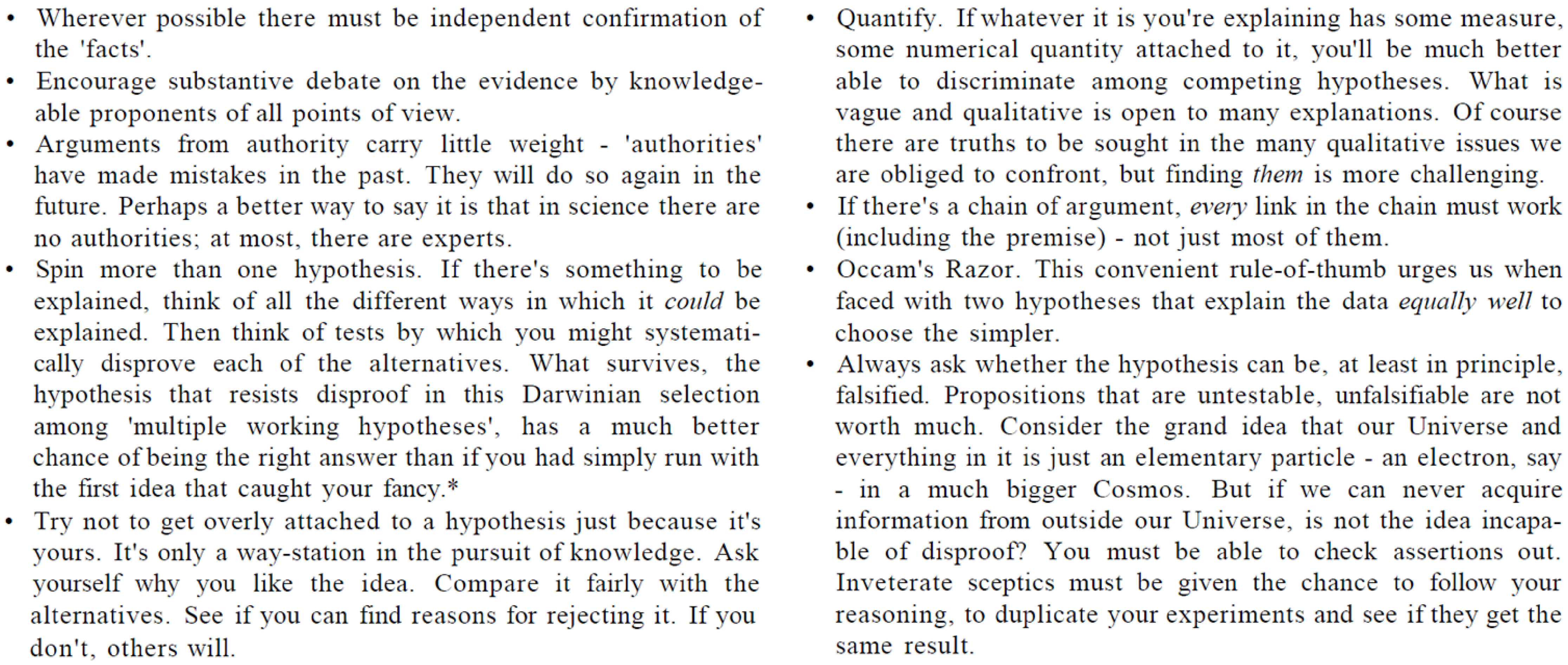
Implementations in R

- Genetic signal based on linear (mixed) models: lme4, pedigreemm, SpATS
- Genomic signal based on supervised machine learning: glmnet, kernlab, ranger, pls, keras
- Genomic signal based on linear (mixed) models: BGLR, bWGR, rrBLUP, EMMREML

Follow up readings

- Xu (2013). Mapping quantitative trait loci by controlling polygenic background effects. Genetics, genetics-113
- Morota and Gianola (2014). Kernel-based whole-genome prediction of complex traits: a review. Frontiers in genetics, 5, 363.
- Henryon et al. (2014). Animal-breeding schemes using genomic information need breeding plans designed to maximise long-term genetic gains. Livestock Science, 166, 38-47.
- Heslot, N., Jannink, J. L., & Sorrells, M. E. (2015). Perspectives for genomic selection applications and research in plants. Crop Science, 55(1), 1-12.
- Xavier et al. (2016). Walking through the statistical black boxes of plant breeding. Theoretical and applied genetics, 129(10), 1933-1949.
- Hickey et al. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nature genetics, 49(9), 1297.

Tools for sceptical thinking



- Wherever possible there must be independent confirmation of the 'facts'.
- Encourage substantive debate on the evidence by knowledgeable proponents of all points of view.
- Arguments from authority carry little weight - 'authorities' have made mistakes in the past. They will do so again in the future. Perhaps a better way to say it is that in science there are no authorities; at most, there are experts.
- Spin more than one hypothesis. If there's something to be explained, think of all the different ways in which it *could* be explained. Then think of tests by which you might systematically disprove each of the alternatives. What survives, the hypothesis that resists disproof in this Darwinian selection among 'multiple working hypotheses', has a much better chance of being the right answer than if you had simply run with the first idea that caught your fancy.*
- Try not to get overly attached to a hypothesis just because it's yours. It's only a way-station in the pursuit of knowledge. Ask yourself why you like the idea. Compare it fairly with the alternatives. See if you can find reasons for rejecting it. If you don't, others will.
- Quantify. If whatever it is you're explaining has some measure, some numerical quantity attached to it, you'll be much better able to discriminate among competing hypotheses. What is vague and qualitative is open to many explanations. Of course there are truths to be sought in the many qualitative issues we are obliged to confront, but finding *them* is more challenging.
- If there's a chain of argument, *every* link in the chain must work (including the premise) - not just most of them.
- Occam's Razor. This convenient rule-of-thumb urges us when faced with two hypotheses that explain the data *equally well* to choose the simpler.
- Always ask whether the hypothesis can be, at least in principle, falsified. Propositions that are untestable, unfalsifiable are not worth much. Consider the grand idea that our Universe and everything in it is just an elementary particle - an electron, say - in a much bigger Cosmos. But if we can never acquire information from outside our Universe, is not the idea incapable of disproof? You must be able to check assertions out. Inveterate sceptics must be given the chance to follow your reasoning, to duplicate your experiments and see if they get the same result.

Carl Sagan - The Demon Haunted World (p.197)