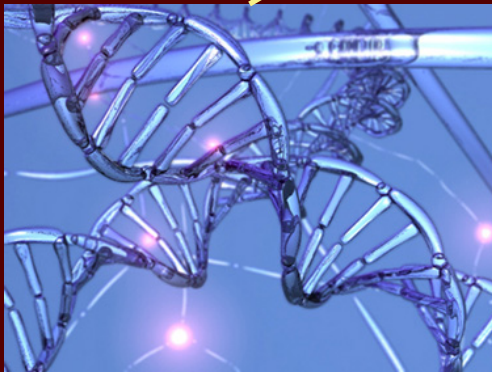**- Overview of concepts & methods**

- Hands-on data analyses

Gustavo de los Campos

(Michigan State University, March 19ht, 2019)

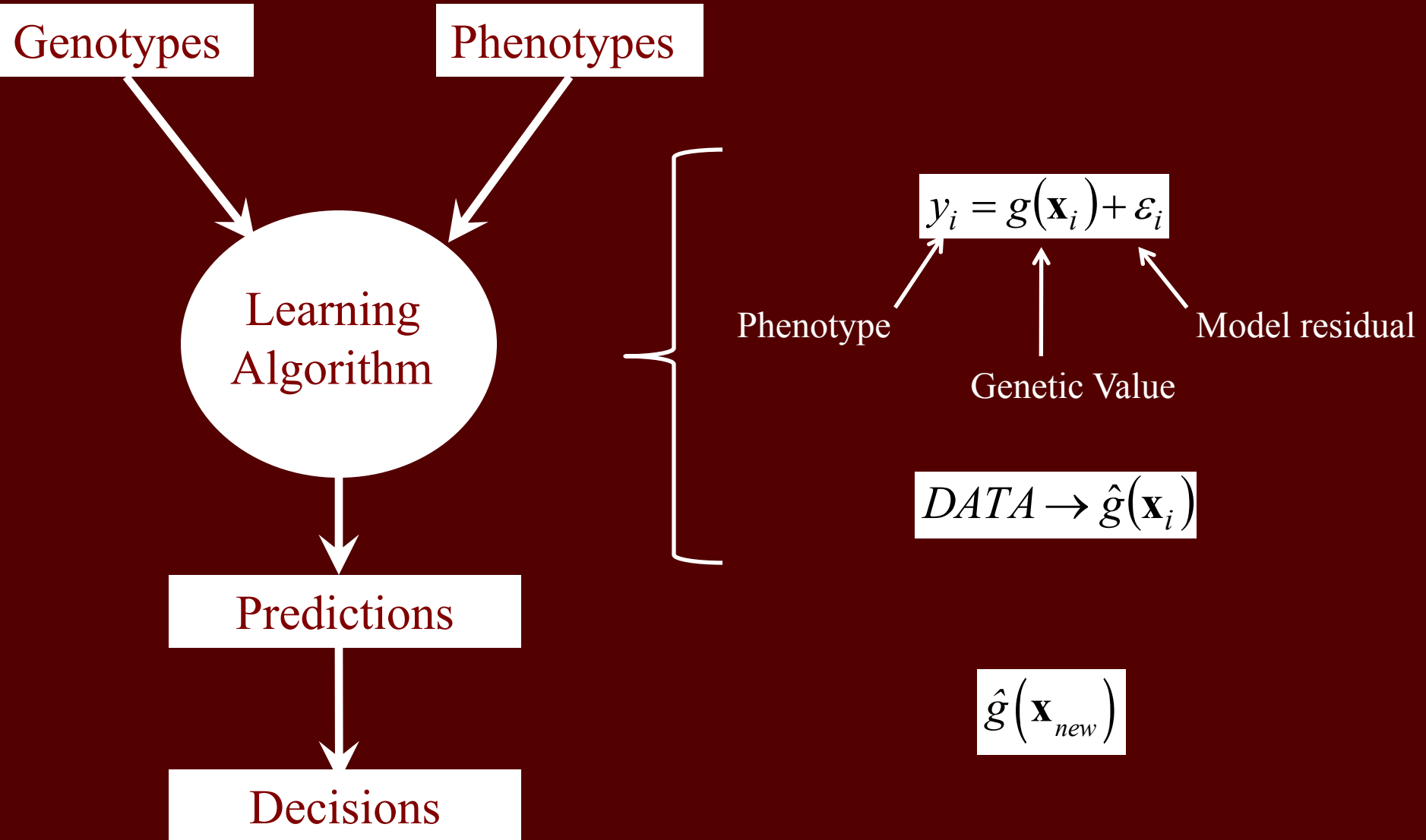# Genes, Environment & Phenotypes

Phenotype

Genome

Environment

Statistical Learning Task

Genotypes

Phenotypes

Learning Algorithm

Predictions

Decisions

$$y_i = g(\mathbf{x}_i) + \varepsilon_i$$

Phenotype

Genetic Value

Model residual

$$DATA \rightarrow \hat{g}(\mathbf{x}_i)$$

$$\hat{g}(\mathbf{x}_{new})$$

Confronting Complexity

$$x_{1i} \quad x_{2i} \quad x_{3i} \quad x_{4i} \ldots \quad x_{p-2,i} \quad x_{p-1,i} \quad x_{p,i}$$
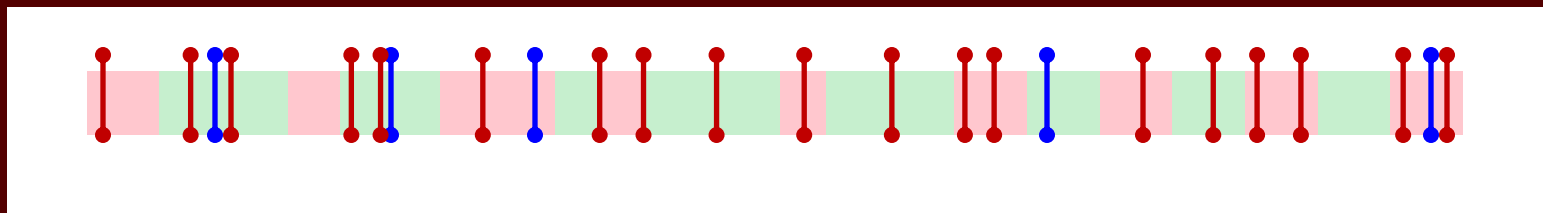
$\Rightarrow$ How many markers?

Two Different Approaches:

$\Rightarrow$ Dimension Reduction (subset selection)

$\Rightarrow$ High Dimensional Regressions (shrinkage & variable selection)

Genomic Data

The Case of Body Mass Index

"The genetic contribution to interindividual variation in common obesity has been estimated at 40–70%...
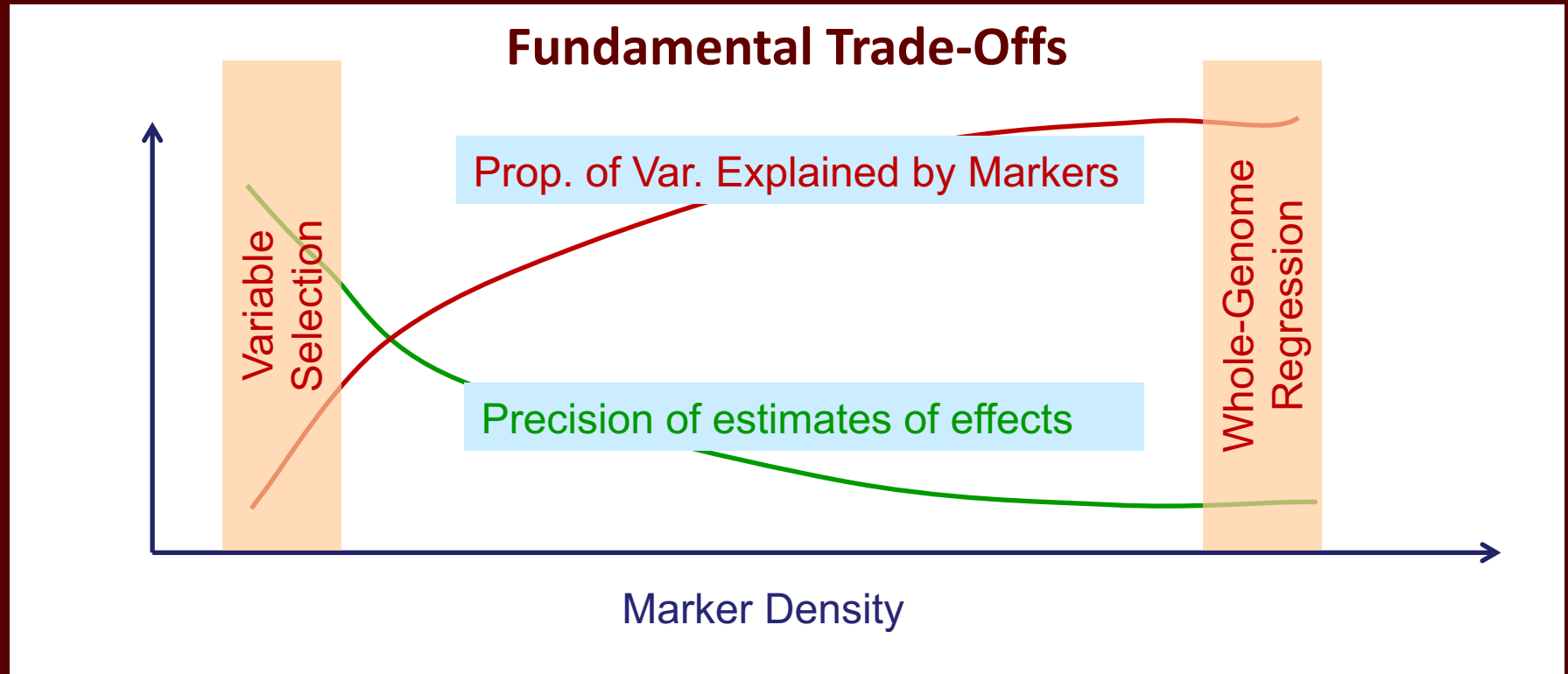
...three waves of GWA studies for obesity-related traits have been carried out...

...the currently established 15 loci explain only a small part of the inter-individual variation in BMI (<2%)."[1]

Chromosome

[1] Loss R., 2009 (PMCID: PMC2810793).

Genome Wide Association Studies

Example 1: Variable Selection Using SMR

Bayesian and Penalized Whole-Genome Regressions

Whole-Genome Regression

$\Rightarrow$ **First proposed by Meuwissen Hayes and Goddard (2001)**.

$\Rightarrow$ Idea: exploit multi-locus LD between markers and QTL.

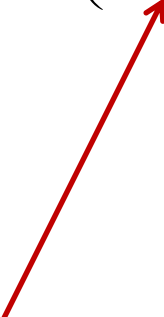$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \qquad\qquad p >> n$$

$\Rightarrow$ Methods: estimates marker effects using regularized (either penalized or Bayesian) regression methods.

Penalized Regression

$$\text{Penalized Regression}: \quad \hat{\boldsymbol{\beta}}_{\arg\min} = \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \times J(\boldsymbol{\beta}) \right\}$$

Lack of fit of the model to the training data set

Regularization Parameter
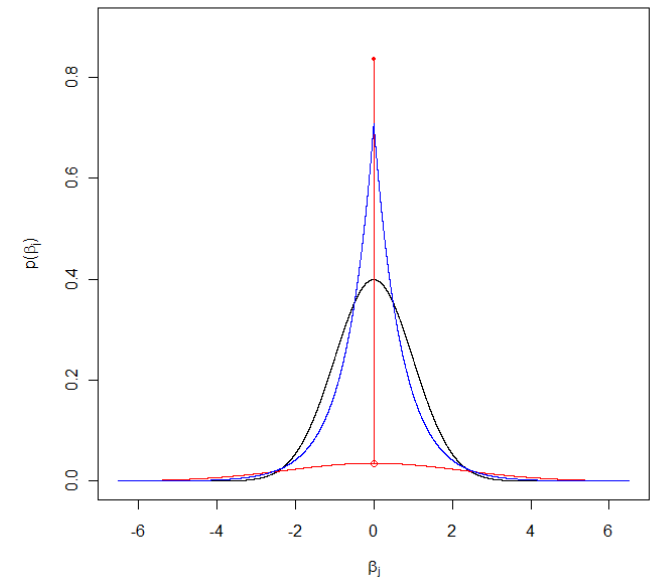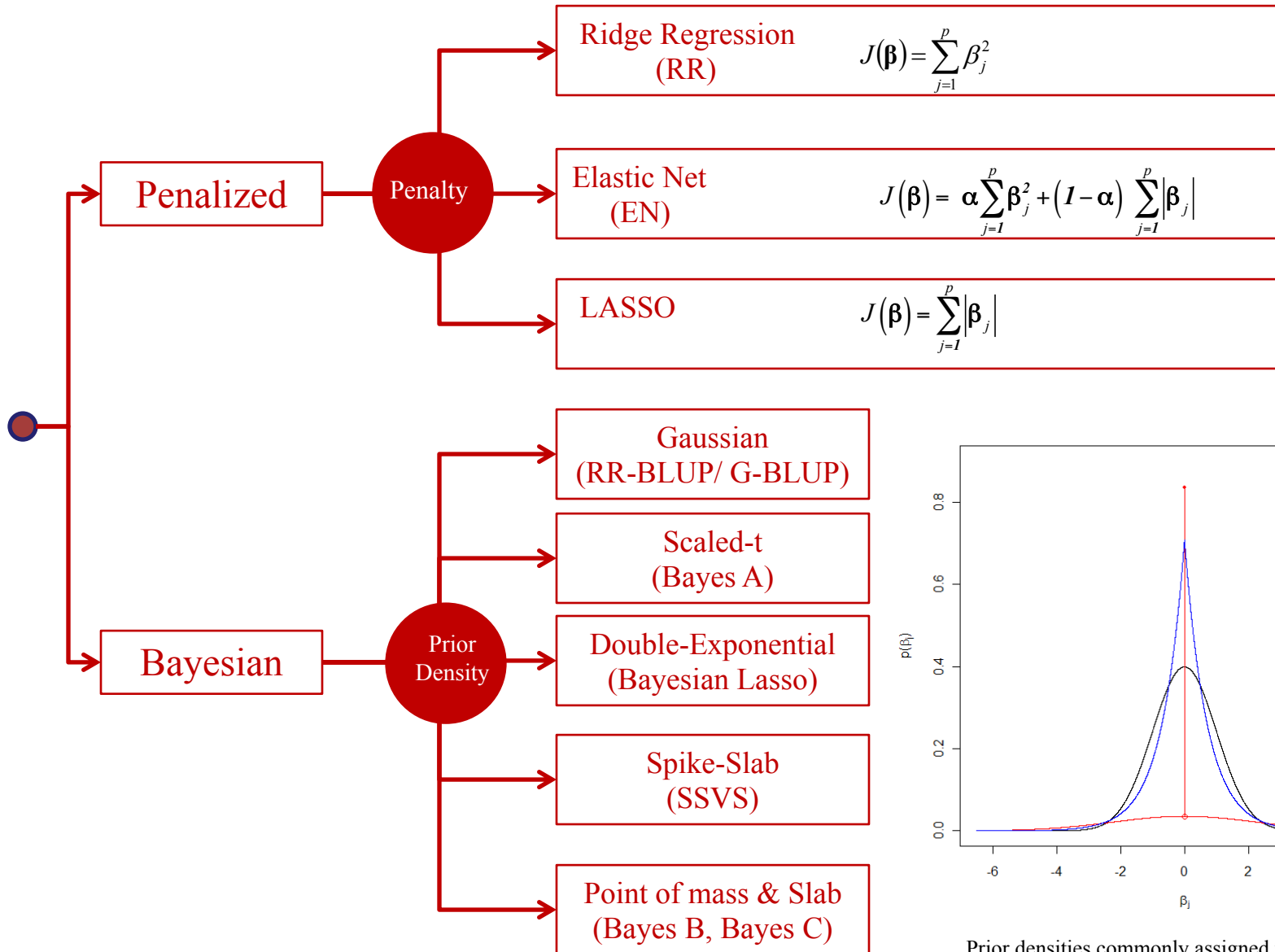
Penalty on Model Complexity

Example 2

$$\text{Penalized Regression}: \quad \hat{\boldsymbol{\beta}}_{\text{argmin}} = \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \times J(\boldsymbol{\beta}) \right\}$$

Elements of model specification:

- Type of penalty (L1, L2, Elastic-Net)
- Strength of penalty

- Example 2

Penalized

Penalty

Ridge Regression (RR)

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2$$

Elastic Net (EN)

$$J(\boldsymbol{\beta}) = \alpha \sum_{j=1}^{p} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p} |\beta_j|$$

LASSO

$$J(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$$

Bayesian

Prior Density

Gaussian (RR-BLUP/ G-BLUP)

Scaled-t (Bayes A)

Double-Exponential (Bayesian Lasso)

Spike-Slab (SSVS)

Point of mass & Slab (Bayes B, Bayes C)

Prior densities commonly assigned to marker effects

Example 3