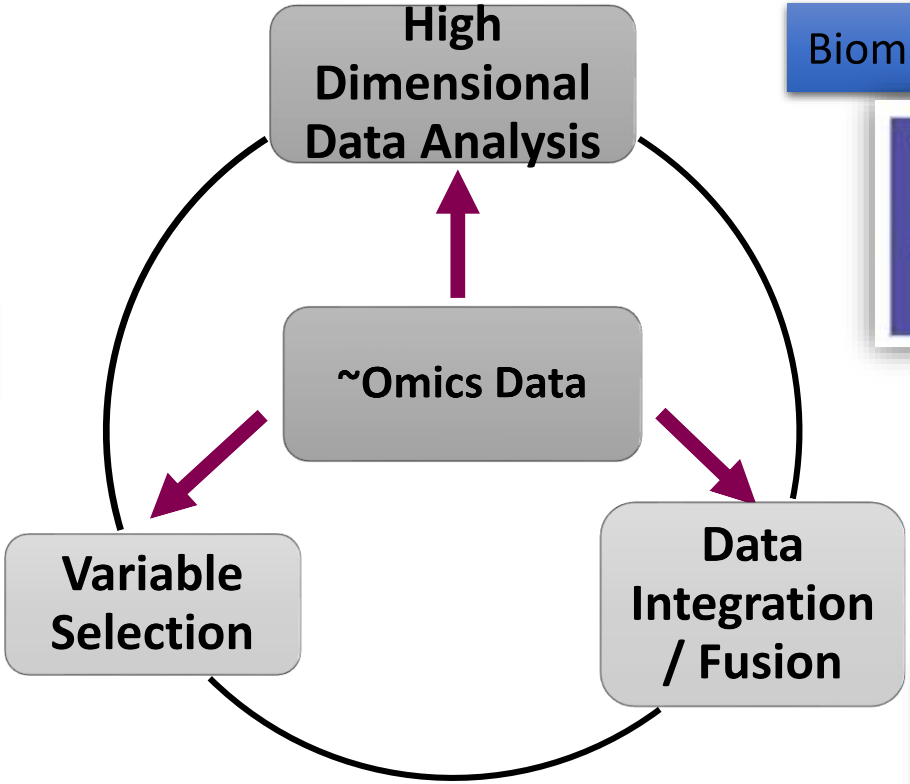
Metabolomics/lipidomics biomarker discovery

Dr. Animesh Acharjee

MRC Human Nutrition Research, Cambridge, UK

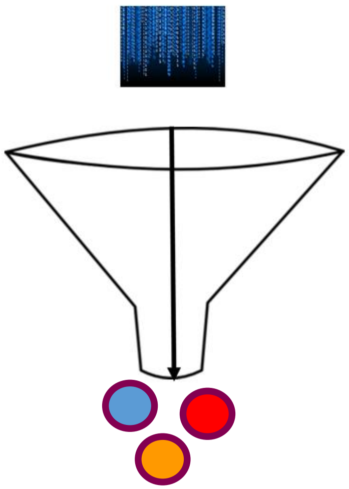
Department of Biochemistry, University of Cambridge, UK

Overview of core areas

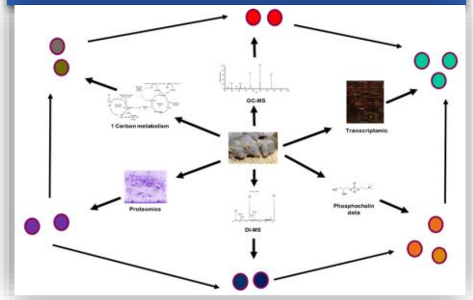


Biomarker discovery

Select best features



Linking multiple data



Agenda

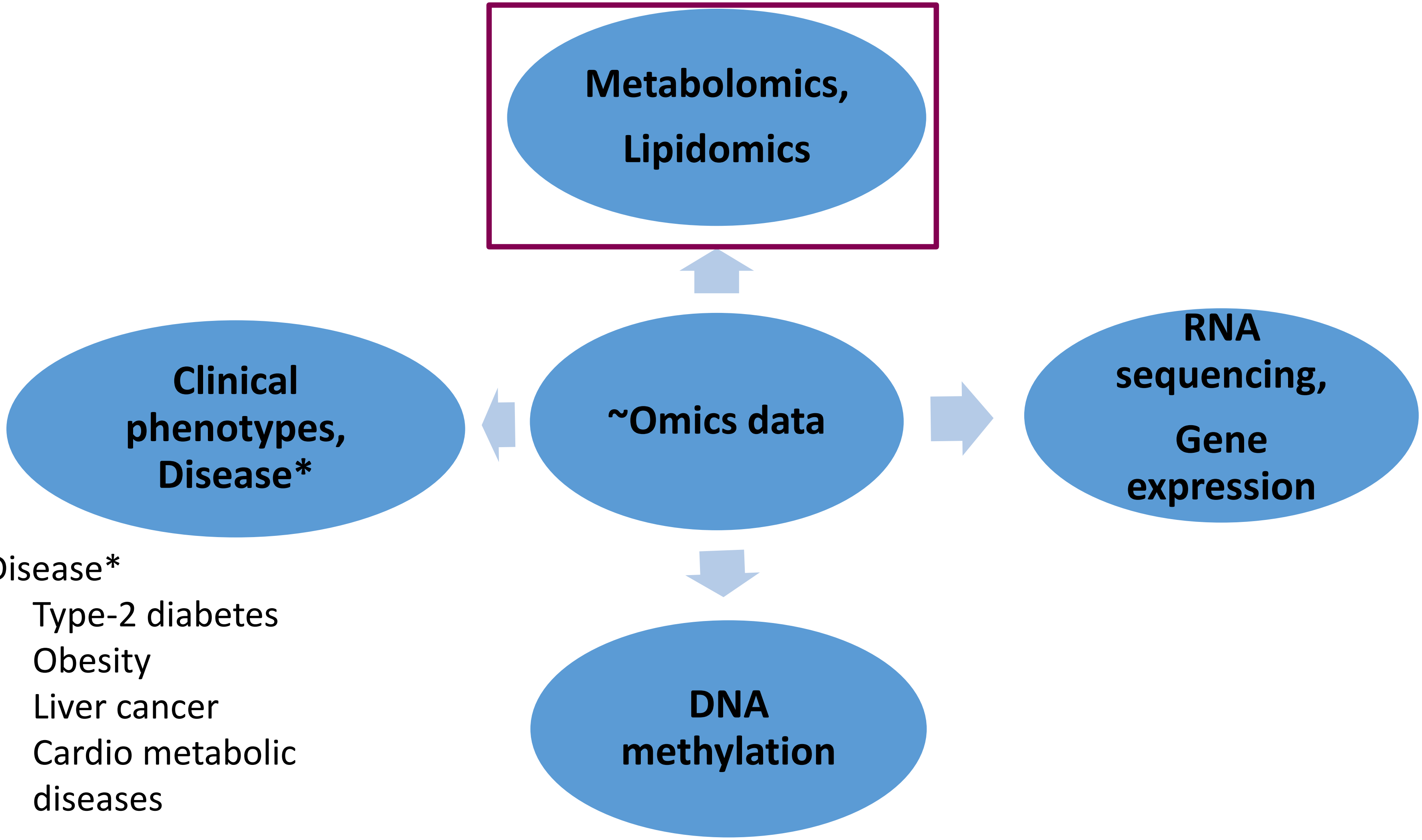
- Introduction : Biomarker Discovery
- Classification & regression
- Random Forest
- Hands on : Metabolomics data analysis

Background

What are biomarkers?

- A biomarker, or biological marker, is an indicator of a biological state or system (in the level of gene, metabolite and/or protein)
- Some of the important properties of a biomarker are
  - Robust
  - Predictive
  - Indicative

Overview of data sets



Metabolomics,
Lipidomics

~Omics data

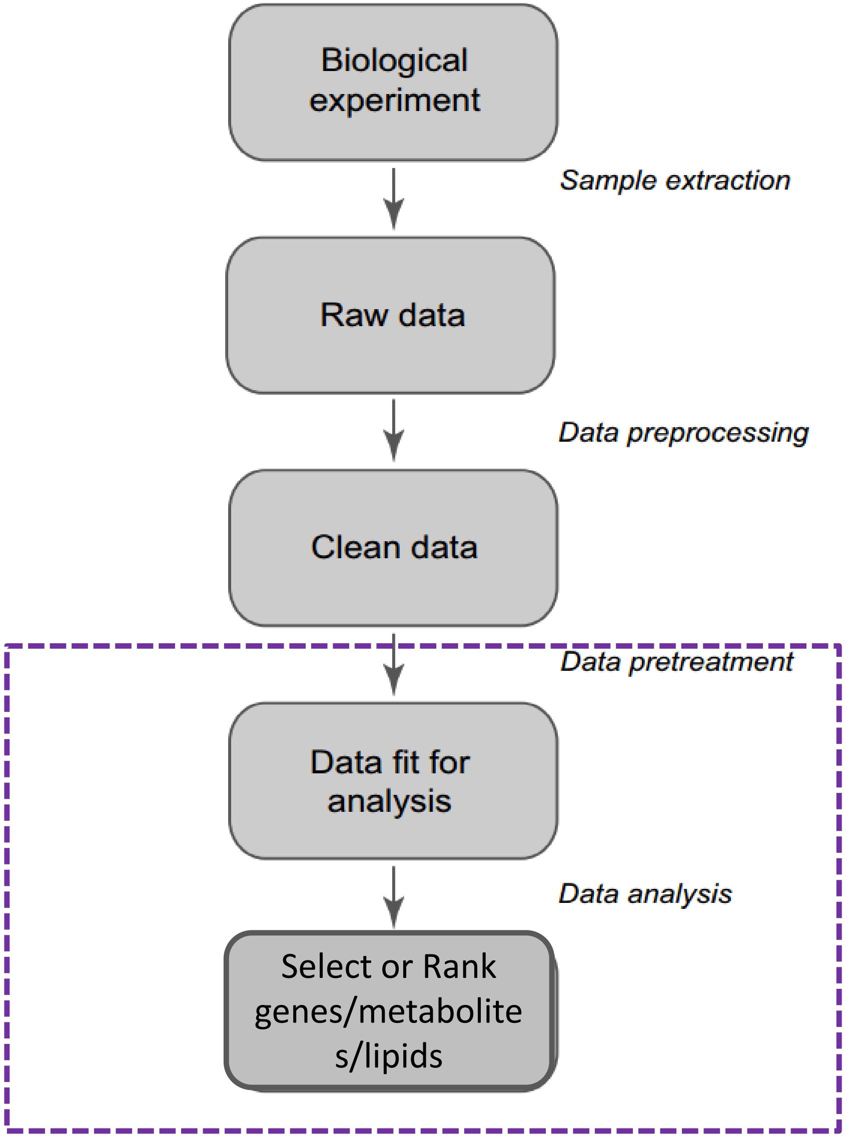
Clinical phenotypes, Disease*

RNA sequencing,
Gene expression

DNA methylation

Disease*
- Type-2 diabetes
- Obesity
- Liver cancer
- Cardio metabolic diseases

Which step I am talking about?



Biological experiment

Sample extraction

Raw data

Data preprocessing

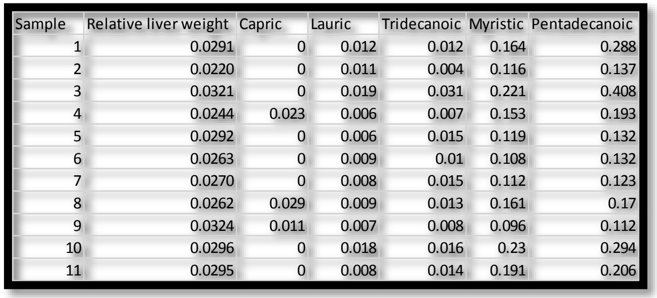
Clean data

Data pretreatment

Data fit for analysis

Data analysis

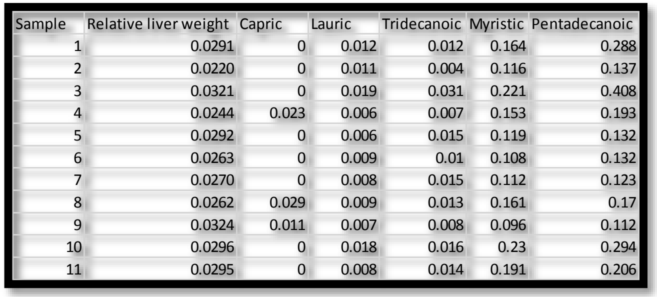
Select or Rank genes/metabolites/lipids



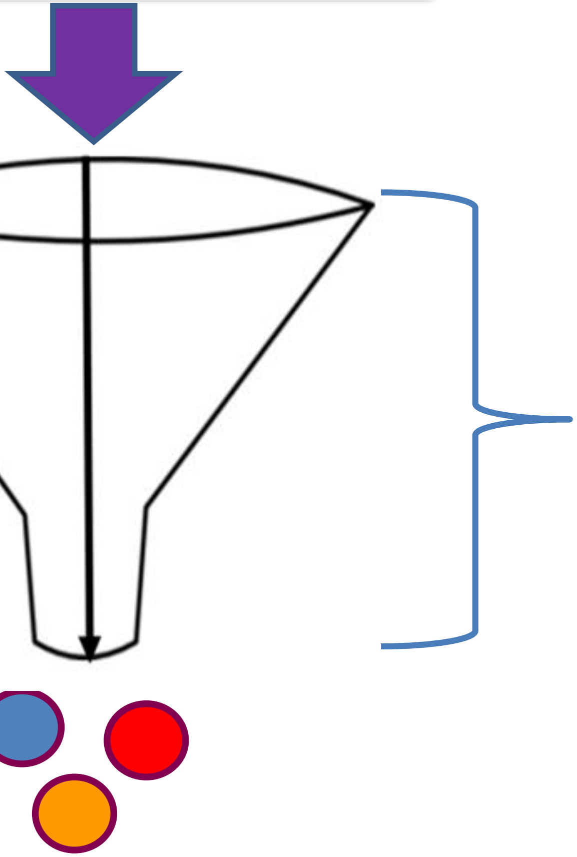
Sample	Relative liver weight	Capric	Lauric	Tridecanoic	Myristic	Pentadecanoic
1	0.0291	0	0.012	0.012	0.164	0.288
2	0.0220	0	0.011	0.004	0.116	0.137
3	0.0321	0	0.019	0.031	0.221	0.408
4	0.0244	0.023	0.006	0.007	0.153	0.193
5	0.0292	0	0.006	0.015	0.119	0.132
6	0.0263	0	0.009	0.01	0.108	0.132
7	0.0270	0	0.008	0.015	0.112	0.123
8	0.0262	0.029	0.009	0.013	0.161	0.17
9	0.0324	0.011	0.007	0.008	0.096	0.112
10	0.0296	0	0.018	0.016	0.23	0.294
11	0.0295	0	0.008	0.014	0.191	0.206

van den Berg et al., 2006

Candidate biomarkers

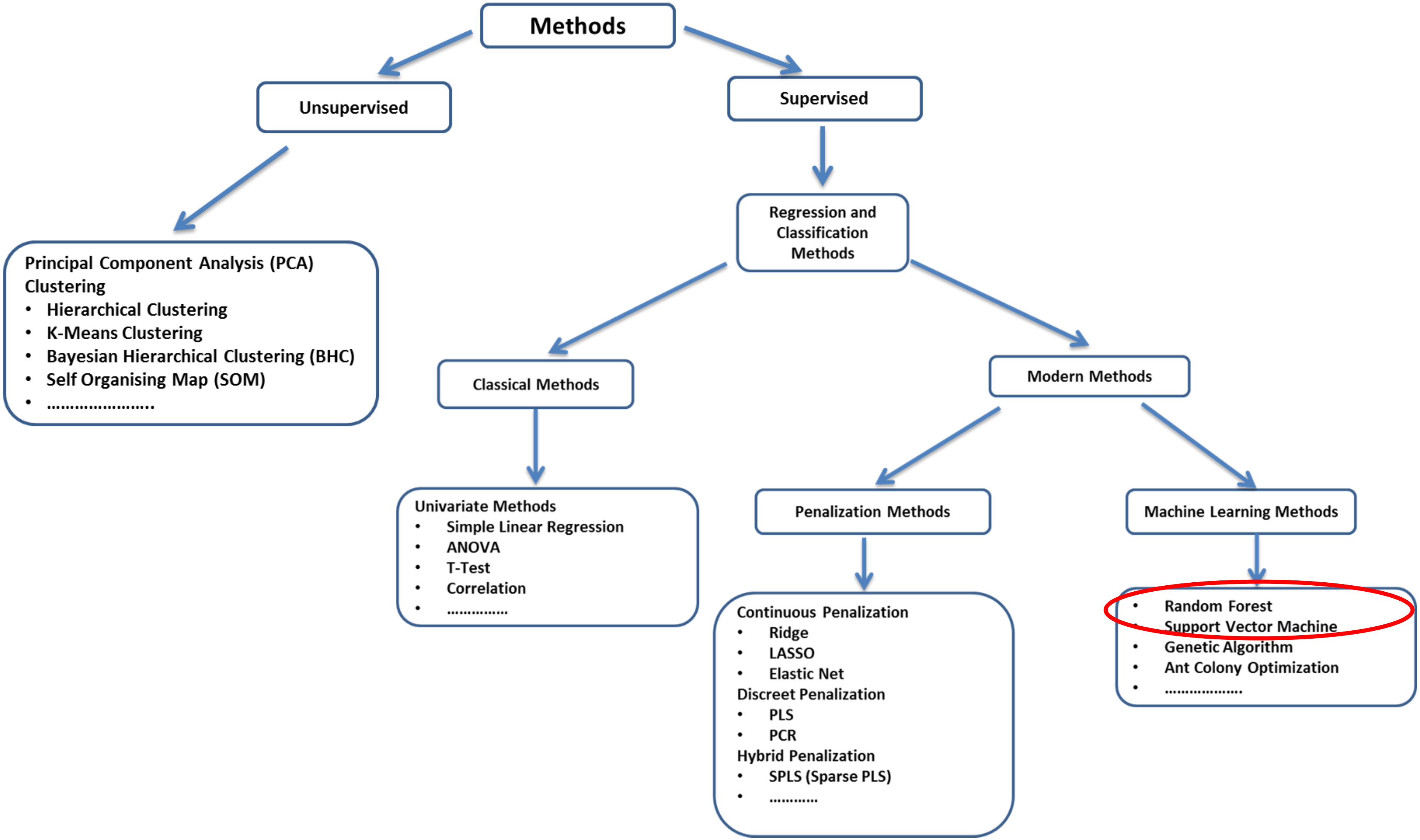


Sample	Relative liver weight	Capric	Lauric	Tridecanoic	Myristic	Pentadecanoic
1	0.0291	0	0.012	0.012	0.164	0.288
2	0.0220	0	0.011	0.004	0.116	0.137
3	0.0321	0	0.019	0.031	0.221	0.408
4	0.0244	0.023	0.006	0.007	0.153	0.193
5	0.0292	0	0.006	0.015	0.119	0.132
6	0.0263	0	0.009	0.01	0.108	0.132
7	0.0270	0	0.008	0.015	0.112	0.123
8	0.0262	0.029	0.009	0.013	0.161	0.17
9	0.0324	0.011	0.007	0.008	0.096	0.112
10	0.0296	0	0.018	0.016	0.23	0.294
11	0.0295	0	0.008	0.014	0.191	0.206



Statistical Methods

Selected variables (metabolites/gene/protein)



Methods

Unsupervised

Principal Component Analysis (PCA)

Clustering

- Hierarchical Clustering
- K-Means Clustering
- Bayesian Hierarchical Clustering (BHC)
- Self Organising Map (SOM)
- ……………………..

Supervised

Regression and Classification Methods

Classical Methods

Univariate Methods

- Simple Linear Regression
- ANOVA
- T-Test
- Correlation
- ……………

Modern Methods

Penalization Methods

Continuous Penalization

- Ridge
- LASSO
- Elastic Net

Discreet Penalization

- PLS
- PCR

Hybrid Penalization

- SPLS (Sparse PLS)
- …………

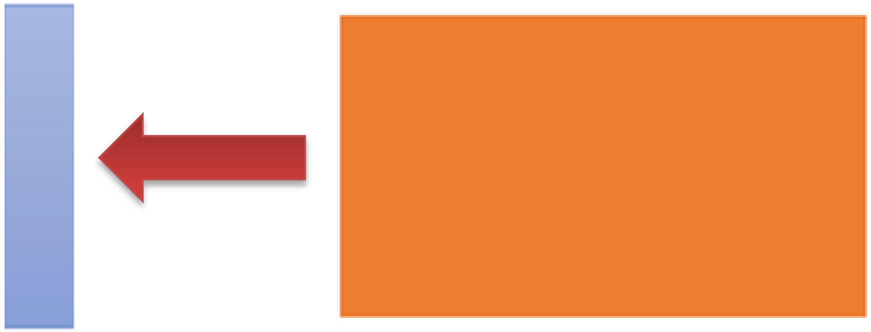
Machine Learning Methods

- Random Forest
- Support Vector Machine
- Genetic Algorithm
- Ant Colony Optimization
- ………………..

Agenda

- Introduction : Biomarker Discovery
- Classification & regression
- Random Forest
- Hands on : Metabolomics data analysis

Supervised Learning



y
(Continuous value
Or Class)

X
(-Omics data)

If “y” Continuous value => Regression approach
If “y” Class value => Classification approach

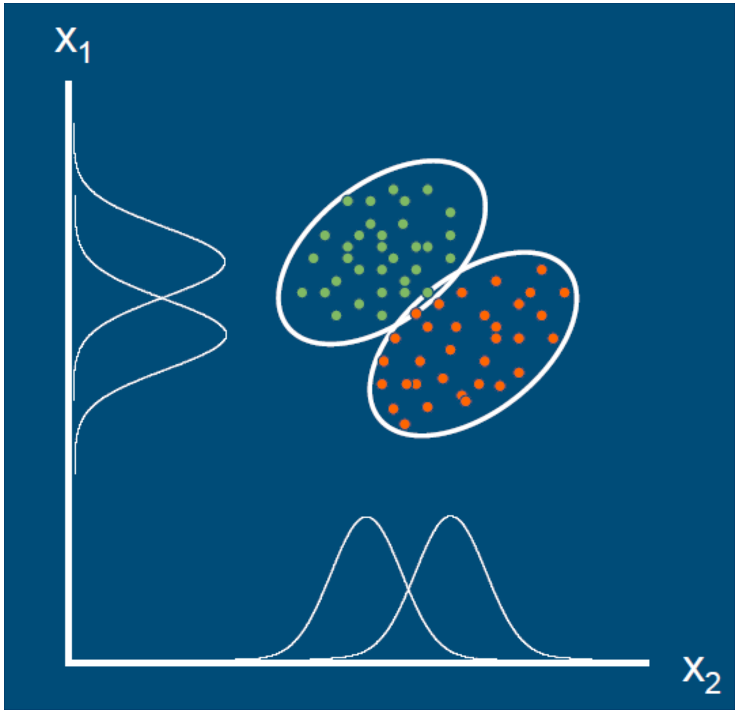
Classification : Basics

- Also known as discriminant analysis (DA)

- Find a ‘classifier’ distinguishing between two (or more) groups
- Groups (classes) – known
- Examples
  - Control vs. disease => Binary
  - Control vs. treatment 1, treatment 2, treatment 3 => Multi class

- Goal 1: predict class of a *new* sample/observation
  - Using its variable/feature values
  - With high precision
- Goal 2: selection of subset of variables with good classification

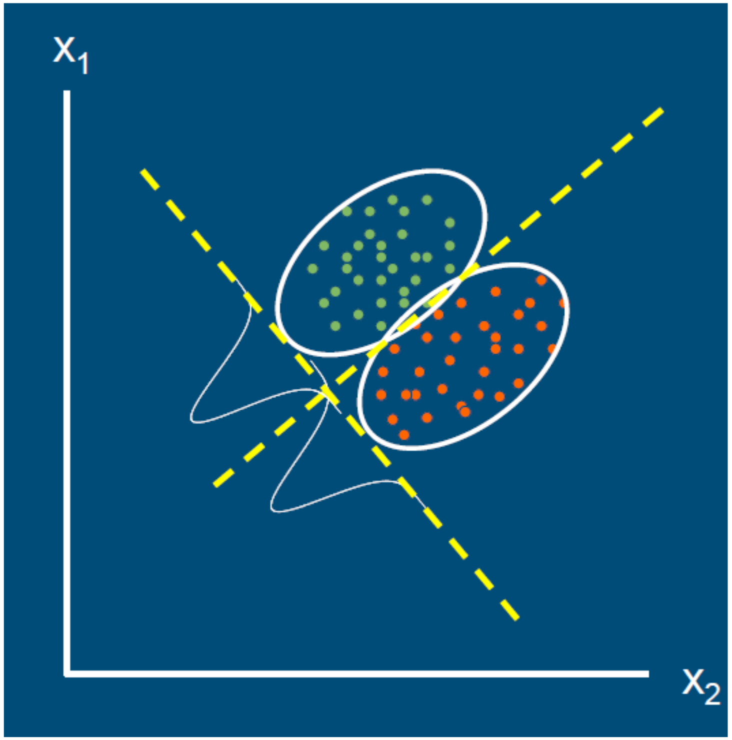
Classification / Discriminant analysis



- Objects (samples) not separated very well by either x_1 or x_2

- x_1 and x_2: Variables

Classification / Discriminant analysis



- Objects can be separated better using x_1 and x_2 simultaneously

- Criterion: maximize between-class difference as compared to within-class differences

Classification / Discriminant analysis

- Traditional methods for discriminant analysis
  - Rely on having more objects than variables !
  - (sometimes) assume equal ‘shape’
  - (sometimes) assume multivariate normality
  - (sometimes) use linear functions

- Often not possible in big data sets [variables (p)>>objects (n)]

“large p, small n” problem

- In reality, in big data:
  - Few samples (typically 10-100) (*n)*
  - Hundreds or variables *(p)*
  - ‘Wide data’:

- Few objects in a very high-dimensional space
  - Data space is almost ‘empty’
  - (Too) many possibilities for separating classes
  - Serious risk of ‘overfitting’
    - Perfect classification in current data
    - No or hardly any predictive value for *new* samples
    - Classifier uses also random differences and not just ‘true relationships’

- Need to evaluate quality of classifier

Regression: Basics

Why we use regression?

- Modeling relationship between variables
- Predict outcome of one variable as a function of others
- Investigate the relative importance of the predictors

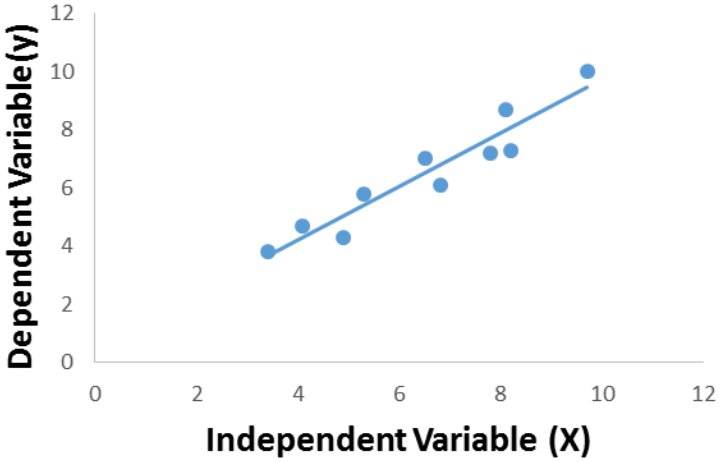
- Assumptions : Linearity, Homoscedasticity, Independence, Normality

Types of regression

- Linear
  - Simple linear
  - Multiple linear
- Nonlinear/Curvilinear

Regression: Basics

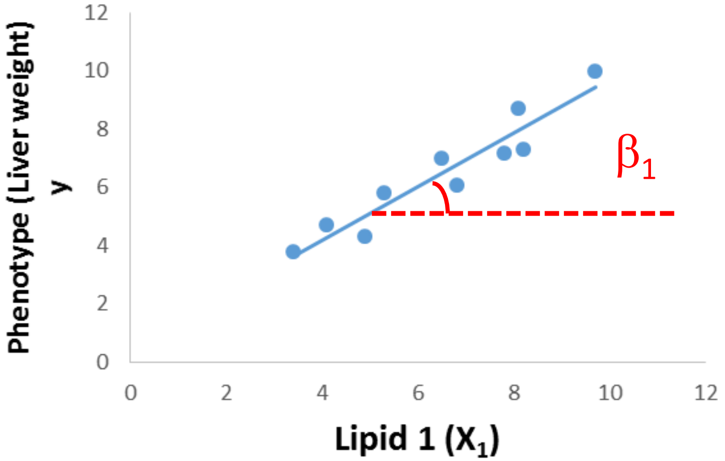
$$y = f(X) + \varepsilon$$



- y=Dependent / response / outcome
- X=Independent / regressor / predictor
- ε =Random variable representing the result of both errors in model specification and measurement.
- Number of observations (n)=10
- Number of response variables (y) =1
- Number of predictor variables (X) =1

- How one variable changes with another
- How "X" and "y" are behaving

Regression: One example



Here:

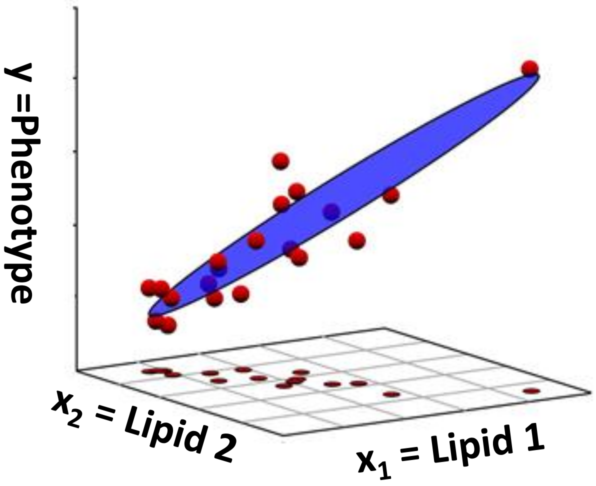
y = Phenotype (Liver weight)

x_1 = Lipid 1

$$y\ (\text{Phenotype}) = \beta_0 + \beta_1 {}^{*}\ X_1(\text{Lipid } 1) + \text{Error}$$

Slope / Coefficient/ Weights

Multiple linear regression



Same example but with x_2:

y = Phenotype (Liver weight)

x_1 = Lipid 1

x_2 = Lipid 2

Regression Equation : Surface

y (Phenotype) = $\beta_0 + \beta_1 * X_1$(Lipid 1) + $\beta_2 * X_2$(Lipid 2) + Error

If nr. of variable= "p", then

More general Equation: multidimensional surface

$$y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 \ldots + \beta_p * X_p + \text{Error}$$

Question:

Do you see any problem / Issue in this example?

Or in closed form

$$\sum_{i=1}^{N} \{y_i - \hat{y}_i\}^2 = \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j \, x_{ij} \right\}^2$$

Multicollinearity

- Lipid 1 (X_1)and Lipid 2 (X_2)might be correlated
- Exact or near linear relationships between the x variables
- Also called as : collinearity, near-collinearity or ill-conditioning

What are the consequences?

Regression coefficients

- Unstable (sensitive to small changes)
- Not uniquely defined
- Have high variance
- Coefficients can get the wrong sign
- Absolute values of regression coefficients can be absurdly high
- Impossible to interpret individually
- Relative importance of variables cannot be estimated reliably

Consequences of Multicollinearity

Bouncing betas

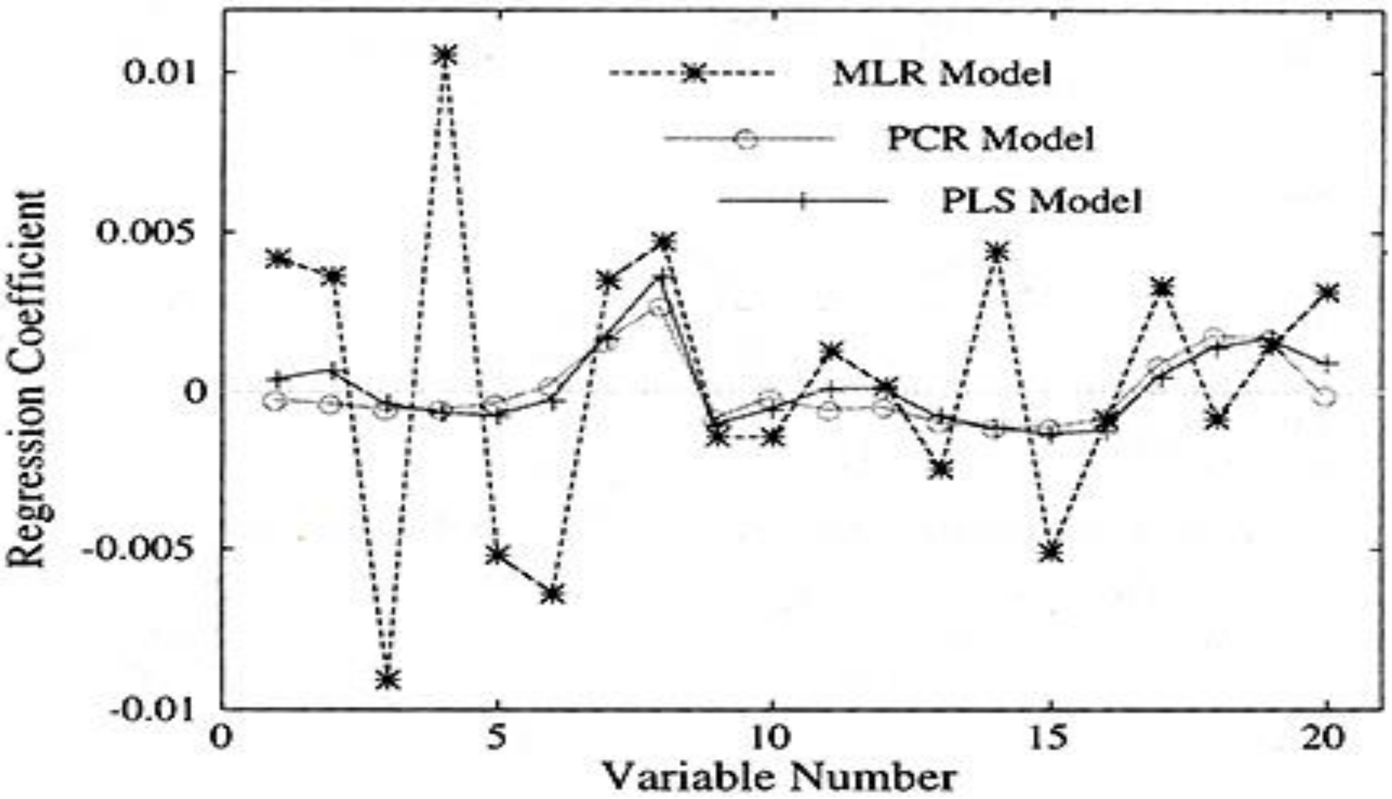


Figure 48. Regression Coefficients in MLR, PCR and PLS Models of SFCM Data.

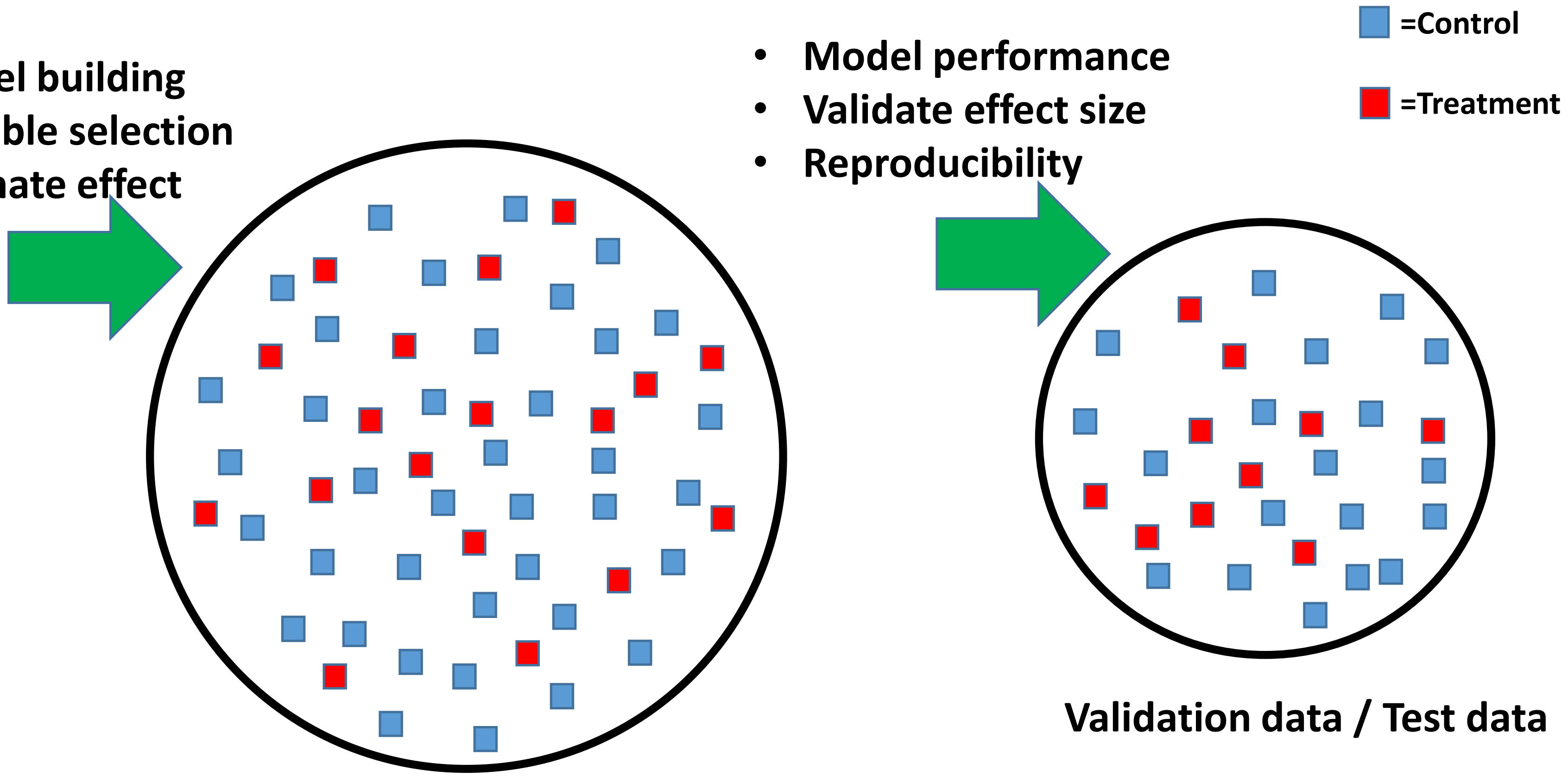
The Elements of Statistical Learning: Data Mining, Inference, and Prediction
Trevor Hastie, Robert Tibshirani, Jerome Friedman

Also possible solutions

- Use selected variables based on “some” criteria
- Filter first on univariate methods (t-test, correlation, ANOVA)
    - Problem
        - Assumption is variables are independent
        - Multiple testing
- Dimension reduction (PCA, clustering)
- Penalization methods
- Machine learning methods

Analysis Flow: Data

- Model building
- Variable selection
- Estimate effect size



- Model performance
- Validate effect size
- Reproducibility

🟦 =Control

🟥 =Treatment

Discovery data / Training data

Validation data / Test data

Agenda

- Introduction : Biomarker Discovery
- Classification & regression
- **Random Forest**
- Hands on : Metabolomics data analysis

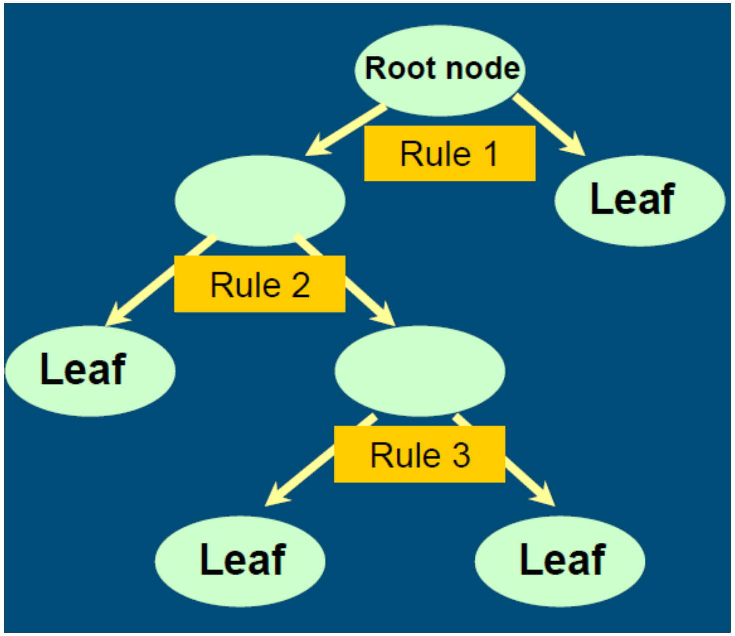
Random forests (Breiman 2001)

- Both classification and multiple regression
- Handles high numbers of variables (p >> n)
- Handles categorical and continuous predictors
- Two-class and multi-class
- Robust to large numbers of noise variables
- Incorporates interactions between variables
- Internal cross validation
- *Variable importance* is estimated
- Proximities between cases are computed
  - can be used to do clustering (unsupervised)
- Fast algorithm
- Extension of ‘Classification and regression trees’ (CART)

Random forest

- Ensemble method
  - not one tree, but many
- Each single tree unpruned
  - Low bias, high variance
- Introduce two forms of ‘randomness’
  - Random training sets (bootstrap samples)
  - Random variable selection at each node
- Effects
  - Individual trees are weak learners
    - Low bias, low correlation, high variance
  - Averaging over the trees retains low bias and reduces variance !
  - ‘Bagging’ = **b**ootstrap **ag**gregation

Example binary classification tree



- The root contains all samples

- Each subsequent node contains a subset of the samples

- Each decision rule splits up the samples into *two* subgroups

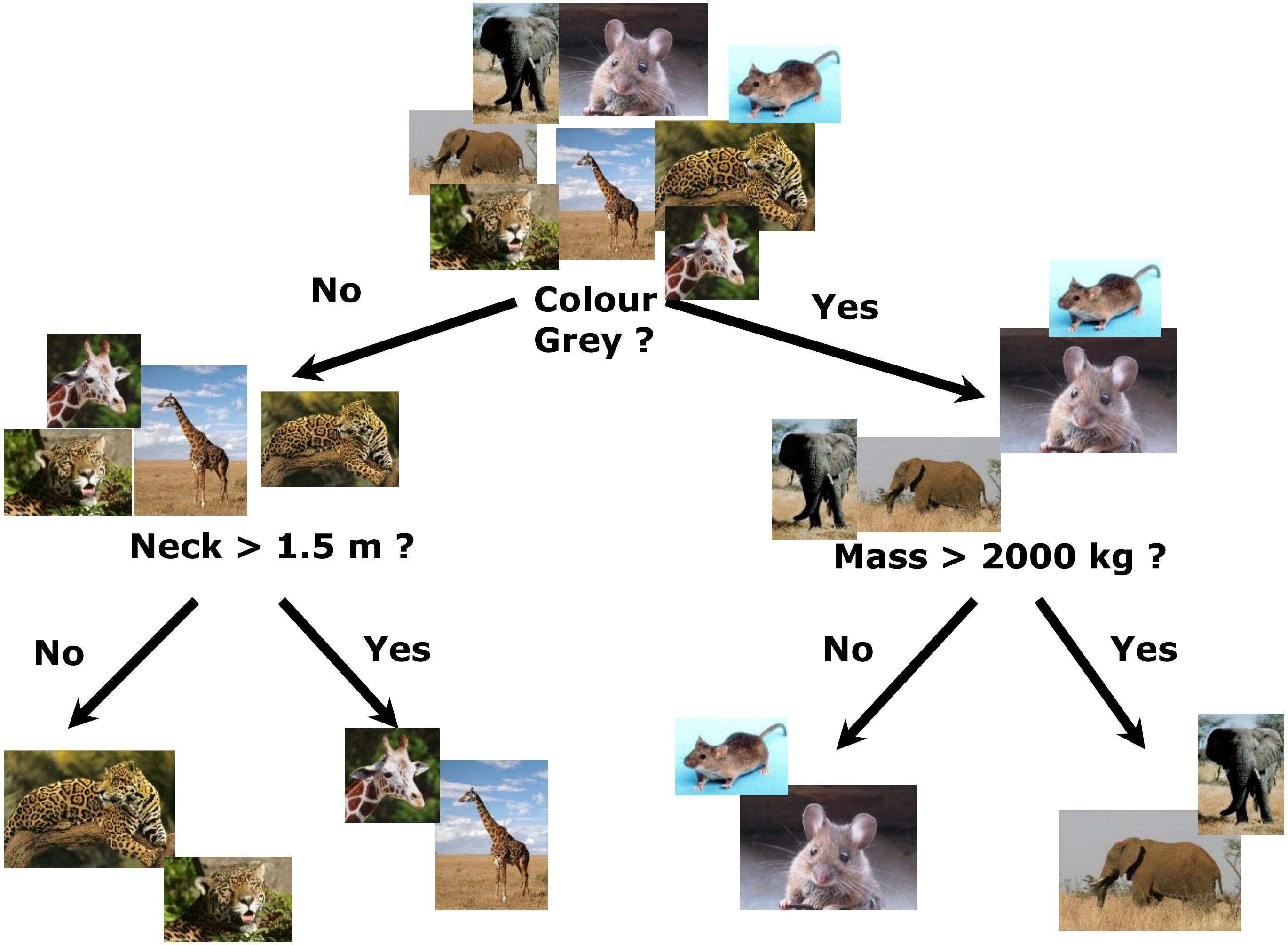
- Every rule is of the form
  - x > t for continuous x
  - x ∊ A for categorical x

- Only one variable per rule
- Same variable can be used again

- Each leaf more or less ‘pure’

- A new sample is run through the tree and one looks for the leaf it ends up

Four species



Colour Grey ?

No

Yes

Neck > 1.5 m ?

No

Yes

Mass > 2000 kg ?

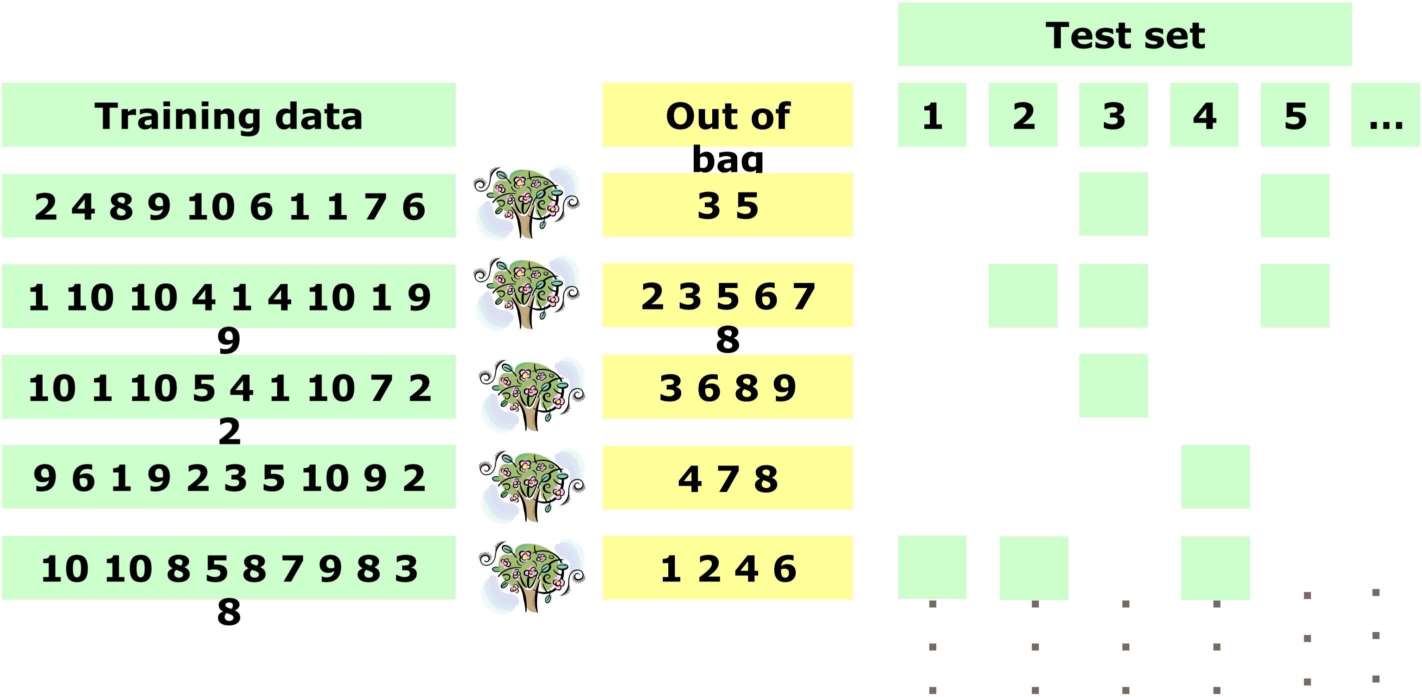
No

Yes

Bootstrap

Complete data	Training data		Out of bag
1 2 3 4 5 6 7 8 9 10	2 4 8 9 10 6 1 1 7 6	🌳	3 5
1 2 3 4 5 6 7 8 9 10	1 10 10 4 1 4 10 1 9 9	🌳	2 3 5 6 7 8
1 2 3 4 5 6 7 8 9 10	10 1 10 5 4 1 10 7 2 2	🌳	3 6 8 9
1 2 3 4 5 6 7 8 9 10	9 6 1 9 2 3 5 10 9 2	🌳	4 7 8
1 2 3 4 5 6 7 8 9 10	10 10 8 5 8 7 9 8 3 8	🌳	1 2 4 6
⋮	⋮		⋮

Internal cross validation using ‘out of bag’ samples



		Test set					
Training data	Out of bag	1	2	3	4	5	…
2 4 8 9 10 6 1 1 7 6	3 5			🟩		🟩	
1 10 10 4 1 4 10 1 9 9	2 3 5 6 7 8		🟩	🟩		🟩	
10 1 10 5 4 1 10 7 2 2	3 6 8 9			🟩			
9 6 1 9 2 3 5 10 9 2	4 7 8				🟩		
10 10 8 5 8 7 9 8 3 8	1 2 4 6	🟩	🟩		🟩		

- For each sample predict class: Use only trees for which it belongs to the OOB set
- Good estimate of test error: Information from i was not used for building these trees

Variable importance

- Idea: change values of a variable and check whether the OOB error changes dramatically

- For each variable x do the following:
    - For each tree of a forest permute the values of x for the ‘out of bag’ samples
    - Redo the classification for the OOB samples
    - Compare the OOB error with original OOB error
        - If unchanged, or just a bit: variable was not so important
        - If error increases a lot: variable *was* important
- Also: quantify the increase in impurity

Variable selection to obtain small sets

- Backward elimination procedure
  - Using variable importance from permutations
  - Using OOB classification error
- Procedure
  - Delete, iteratively, 20% variables with lowest importance
  - So: series of forests with decreasing numbers of vars
  - Compare their OOB errors
    - Choose smallest set within 1 st.error of the best (with min. error)

Díaz-Uriarte and De Andrés, 2006

Some criticism

- Breiman claims: no overfitting
  - Segal (2004): maybe due to the UCI data sets used
  - Overfitting when many highly correlated variables
- Strobl et al. (2006)
  - Random forests systematically prefer categorical variables with more categories over those with less
  - Use subsampling instead of bootstrap
  - Use smaller trees

References

- Acharjee et al. (2016), Integration of metabolomics, lipidomics and clinical data using a machine learning method. *BMC Bioinformatics*, 17(15):440

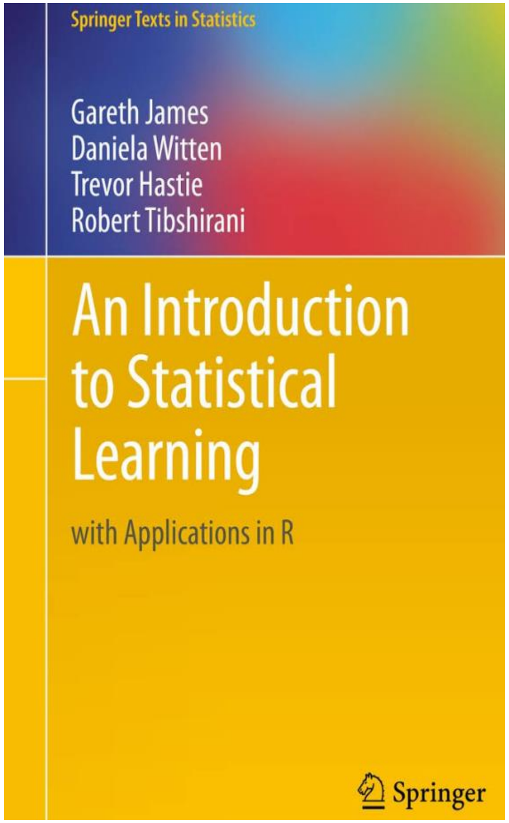
- Acharjee et al. (2016), Integration of multi-omics data for prediction of phenotypic traits using random forest. *BMC Bioinformatics*, 6;17 (5):180

- Acharjee et al. (2011), Data integration and network reconstruction with ~omics data using Random Forest regression in potato. *Analytica Chimica Acta,* 705(1-2):56-63.

- Breiman (2001), Random forests. *Machine learning,* 45: 5-32

- Segal (2004), Machine learning benchmarks and random forest regression. Techn. Paper.

Resources



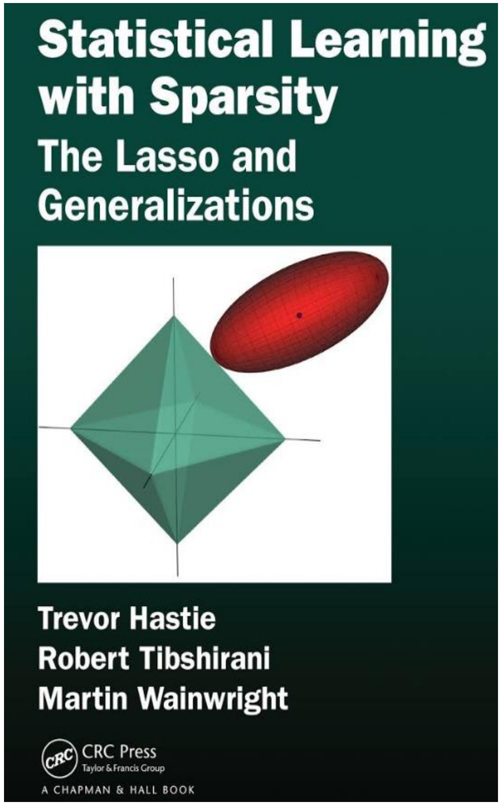
Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer

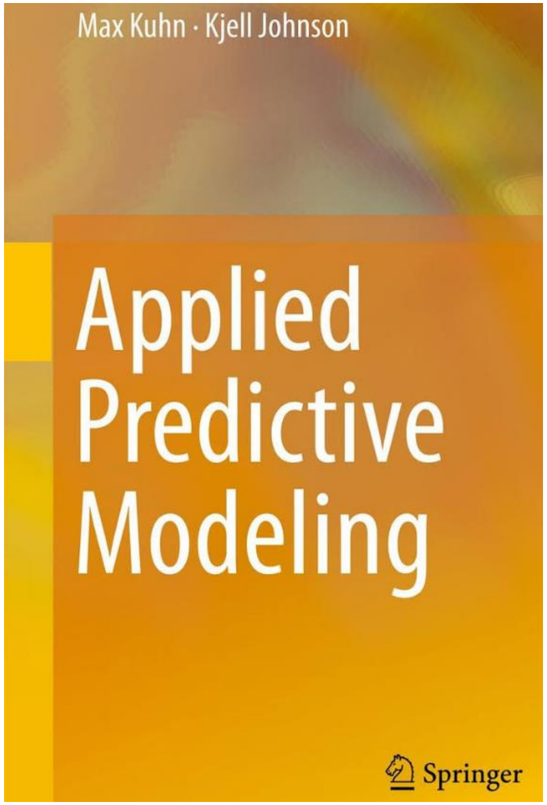


Statistical Learning with Sparsity

The Lasso and Generalizations

Trevor Hastie
Robert Tibshirani
Martin Wainwright

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK



Max Kuhn · Kjell Johnson

Applied Predictive Modeling

Springer

http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf

Agenda

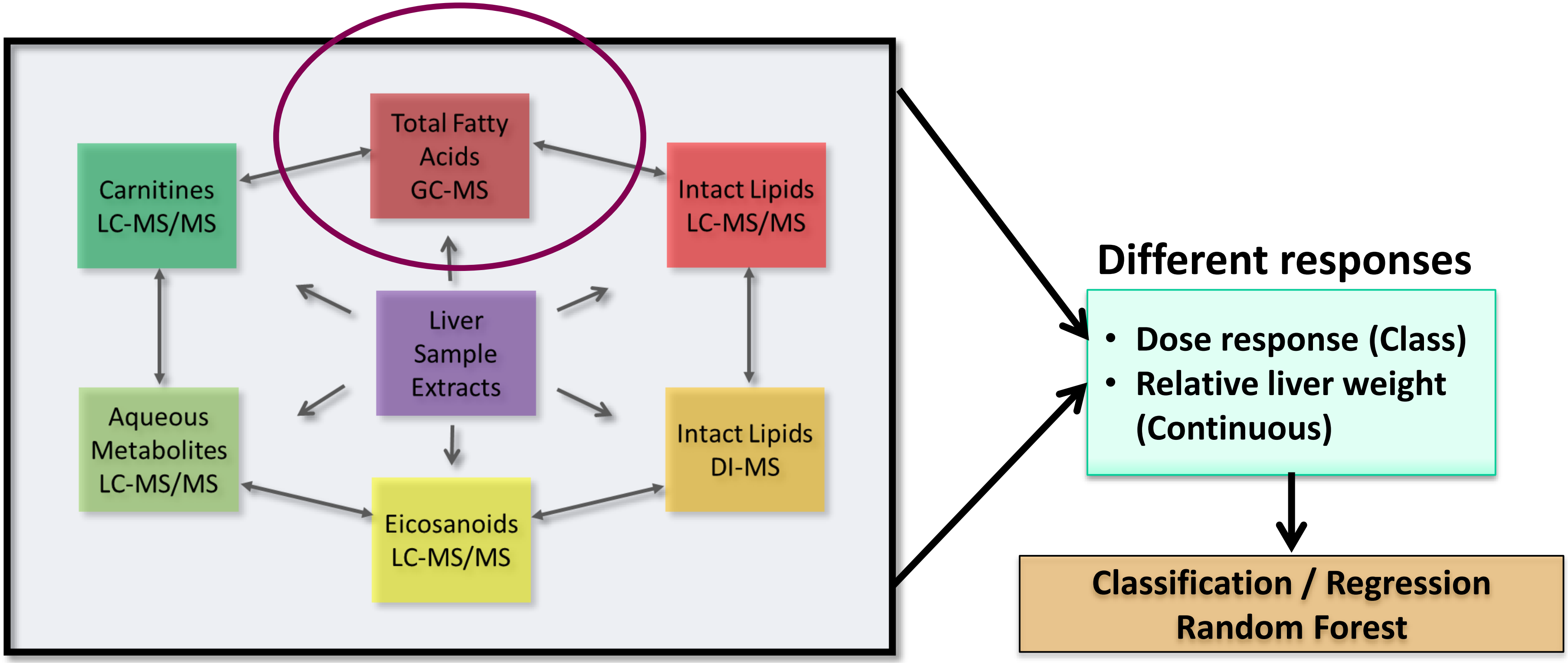
- Introduction : Biomarker Discovery
- Classification & regression
- Random Forest
- Hands on : Metabolomics data analysis

Data Set

Background

- Peroxisome proliferator-activated receptors (PPARs) play a central role in regulating metabolism.
- PPAR-pan agonist (a triple agonist of PPAR-α, -γ, and -δ) was investigated after dietary treatment of male rats (Sprague–Dawley) (Ament et al., 2015)
- Classical toxicological tests (clinical chemistry) & liver metabolomic and lipidomic changes were measured in order to understand metabolism and toxicity.

Data information



Total Fatty Acids GC-MS

Carnitines LC-MS/MS

Intact Lipids LC-MS/MS

Liver Sample Extracts

Aqueous Metabolites LC-MS/MS

Intact Lipids DI-MS

Eicosanoids LC-MS/MS

Different responses

- Dose response (Class)
- Relative liver weight (Continuous)

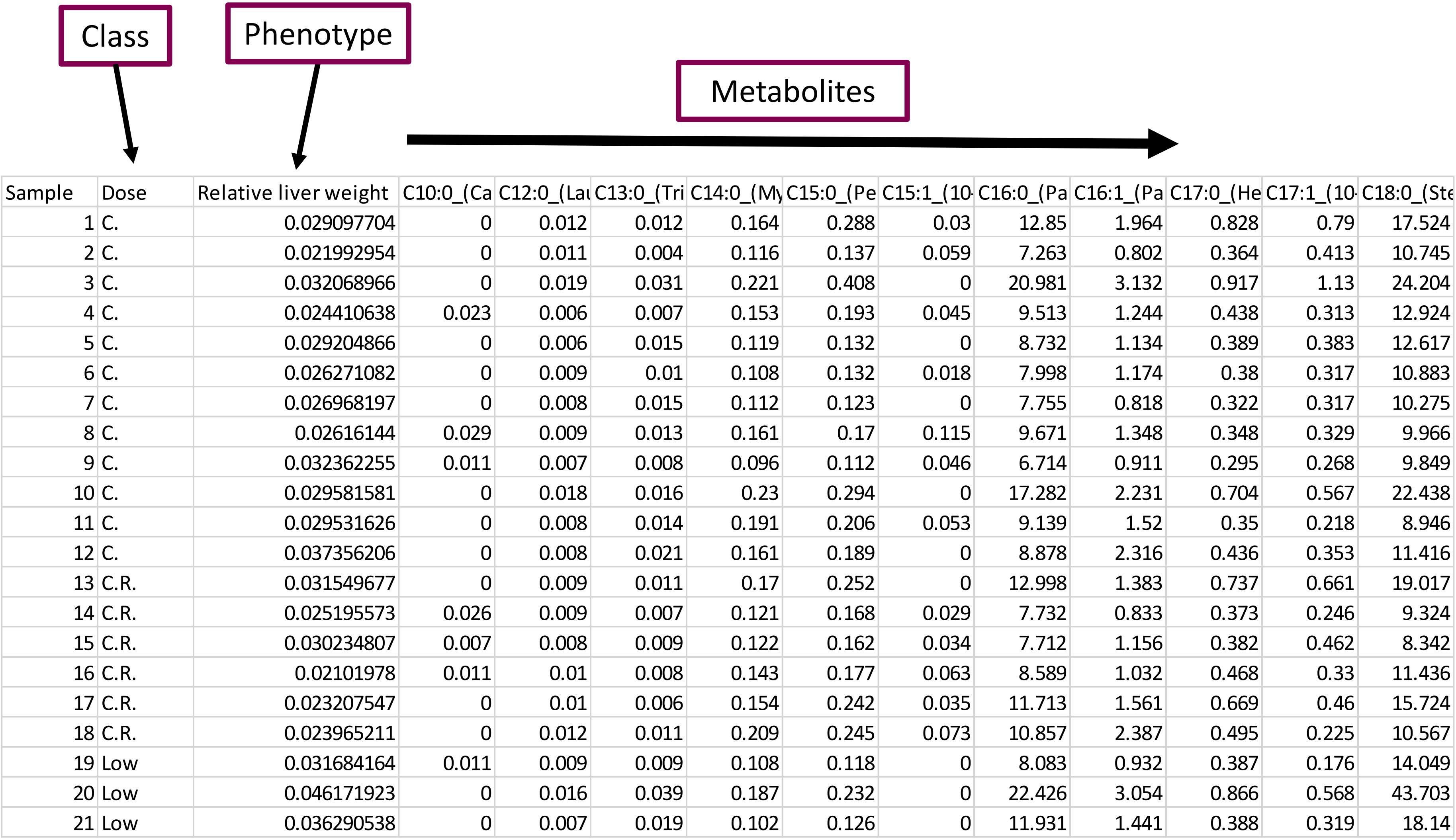
Classification / Regression Random Forest

Study design

Group	Dose (mg/kg/day)	Animal number	Recovery
Control	0	1-12	13-18
Low	30	19-30	-
Intermediate 1	100	31-42	-
Intermediate 2	300	43-54	55-60
High	1000	61-72	73-78

- Number of total groups/class : 8 (5 dose and 3 recovery doses)
- Relative liver weight as phenotype (Continuous)

Ament et al., 2015
Acharjee at al.,2016



Class

Phenotype

Metabolites

Sample	Dose	Relative liver weight	C10:0_(Ca	C12:0_(Lau	C13:0_(Tri	C14:0_(My	C15:0_(Pe	C15:1_(10-	C16:0_(Pa	C16:1_(Pa	C17:0_(He	C17:1_(10-	C18:0_(Ste
1	C.	0.029097704	0	0.012	0.012	0.164	0.288	0.03	12.85	1.964	0.828	0.79	17.524
2	C.	0.021992954	0	0.011	0.004	0.116	0.137	0.059	7.263	0.802	0.364	0.413	10.745
3	C.	0.032068966	0	0.019	0.031	0.221	0.408	0	20.981	3.132	0.917	1.13	24.204
4	C.	0.024410638	0.023	0.006	0.007	0.153	0.193	0.045	9.513	1.244	0.438	0.313	12.924
5	C.	0.029204866	0	0.006	0.015	0.119	0.132	0	8.732	1.134	0.389	0.383	12.617
6	C.	0.026271082	0	0.009	0.01	0.108	0.132	0.018	7.998	1.174	0.38	0.317	10.883
7	C.	0.026968197	0	0.008	0.015	0.112	0.123	0	7.755	0.818	0.322	0.317	10.275
8	C.	0.02616144	0.029	0.009	0.013	0.161	0.17	0.115	9.671	1.348	0.348	0.329	9.966
9	C.	0.032362255	0.011	0.007	0.008	0.096	0.112	0.046	6.714	0.911	0.295	0.268	9.849
10	C.	0.029581581	0	0.018	0.016	0.23	0.294	0	17.282	2.231	0.704	0.567	22.438
11	C.	0.029531626	0	0.008	0.014	0.191	0.206	0.053	9.139	1.52	0.35	0.218	8.946
12	C.	0.037356206	0	0.008	0.021	0.161	0.189	0	8.878	2.316	0.436	0.353	11.416
13	C.R.	0.031549677	0	0.009	0.011	0.17	0.252	0	12.998	1.383	0.737	0.661	19.017
14	C.R.	0.025195573	0.026	0.009	0.007	0.121	0.168	0.029	7.732	0.833	0.373	0.246	9.324
15	C.R.	0.030234807	0.007	0.008	0.009	0.122	0.162	0.034	7.712	1.156	0.382	0.462	8.342
16	C.R.	0.02101978	0.011	0.01	0.008	0.143	0.177	0.063	8.589	1.032	0.468	0.33	11.436
17	C.R.	0.023207547	0	0.01	0.006	0.154	0.242	0.035	11.713	1.561	0.669	0.46	15.724
18	C.R.	0.023965211	0	0.012	0.011	0.209	0.245	0.073	10.857	2.387	0.495	0.225	10.567
19	Low	0.031684164	0.011	0.009	0.009	0.108	0.118	0	8.083	0.932	0.387	0.176	14.049
20	Low	0.046171923	0	0.016	0.039	0.187	0.232	0	22.426	3.054	0.866	0.568	43.703
21	Low	0.036290538	0	0.007	0.019	0.102	0.126	0	11.931	1.441	0.388	0.319	18.14

Thank you

animesh.acharjee@gmail.com

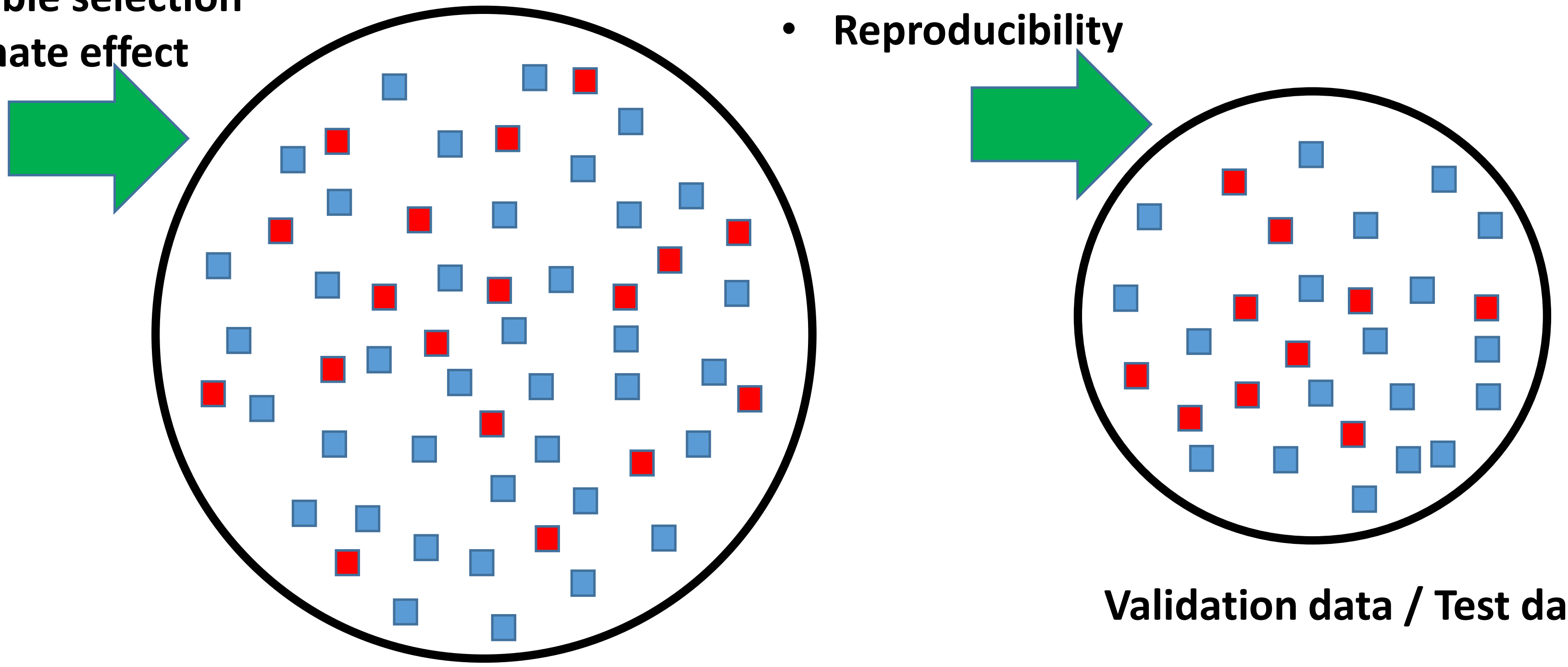
Analysis Flow: Data

- **Model building**
- **Variable selection**
- **Estimate effect size**

- **Model performance**
- **Validate effect size**
- **Reproducibility**

🟦 **=Control**

🟥 **=Treatment**



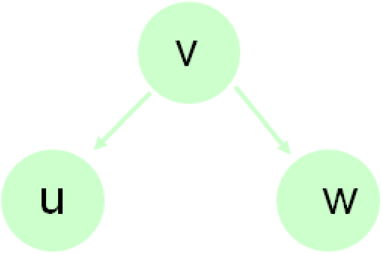
Discovery data / Training data

Validation data / Test data

Criteria for splitting

- Search, per step, ‘best’ variable and split point
  - Each step: splitting only one of the nodes into two
  - ‘best’: decreasing ‘impurity’ most
  - *E.g.* Gini index

$$G = i(v) - \left(p_u i(u) + p_w i(w)\right)$$



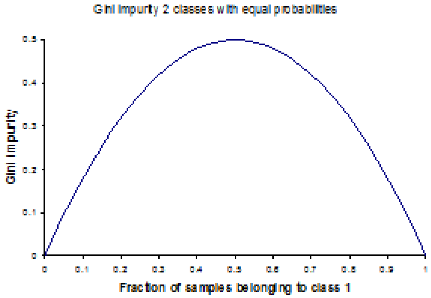
p_u = fraction of samples in node u

p_w = fraction of samples in node w

Impurity single node, 2 classes equal probabilities:

$$i(v) = \frac{2n_1(v)n_2(v)}{n(v)^2}$$

$$G = \frac{2}{n(v)} \cdot \left(\frac{n_1(v)n_2(v)}{n(v)} - \frac{n_1(u)n_2(u)}{n(u)} - \frac{n_1(w)n_2(w)}{n(w)} \right)$$

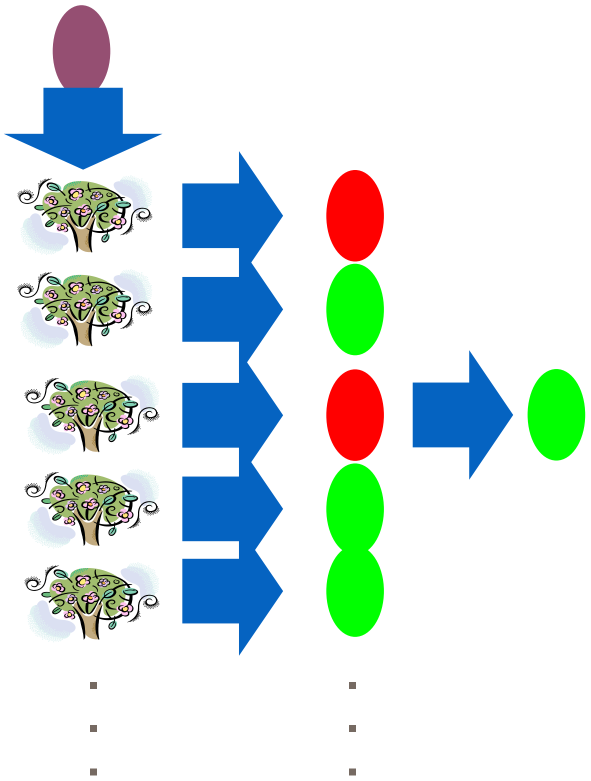


Stopping criteria

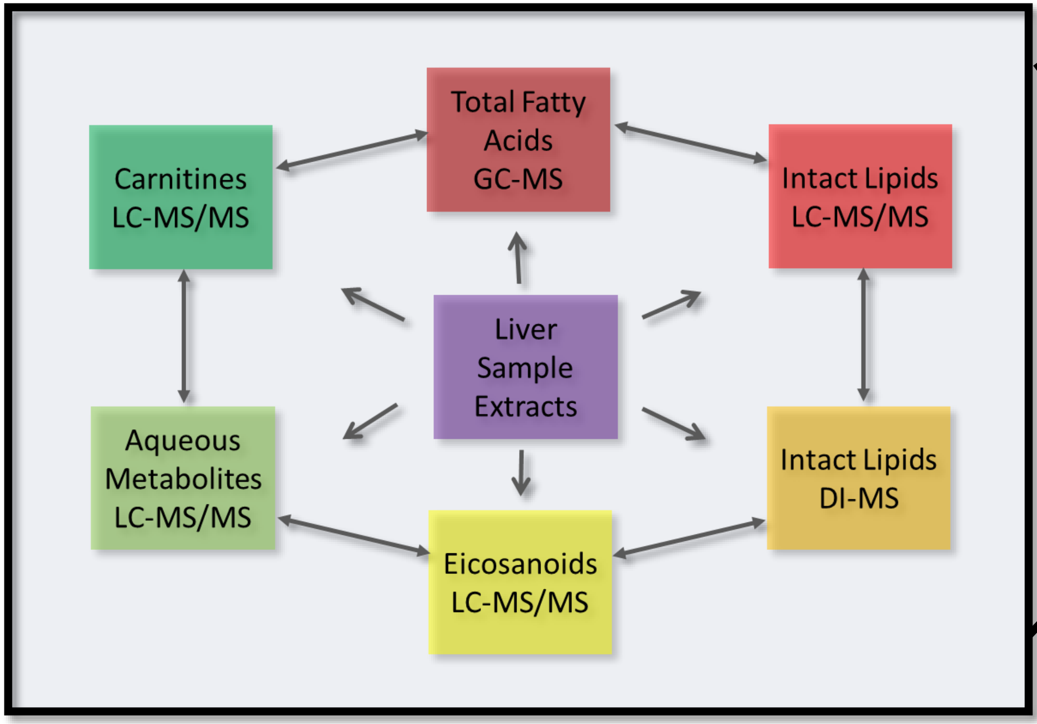
- All nodes are pure (single class) or have a ‘final’ number of elements, *e.g.* 5
- Prune the tree back somewhat
    - Remove splits with low decrease in impurity
    - To protect against overfitting
- Unpruned trees
    - Low bias (end nodes have maximum purity)
    - High variance
        - Widely different rules if the data or the samples change a little

New sample: each tree casts vote, then majority voting

Complete data	Training data
1 2 3 4 5 6 7 8 9 10	2 4 8 9 10 6 1 1 7 6
1 2 3 4 5 6 7 8 9 10	1 10 10 4 1 4 10 1 9 9
1 2 3 4 5 6 7 8 9 10	10 1 10 5 4 1 10 7 2 2
1 2 3 4 5 6 7 8 9 10	9 6 1 9 2 3 5 10 9 2
1 2 3 4 5 6 7 8 9 10	10 10 8 5 8 7 9 8 3 8
⋮	⋮



Data integration/fusion



Total Fatty Acids GC-MS

Carnitines LC-MS/MS

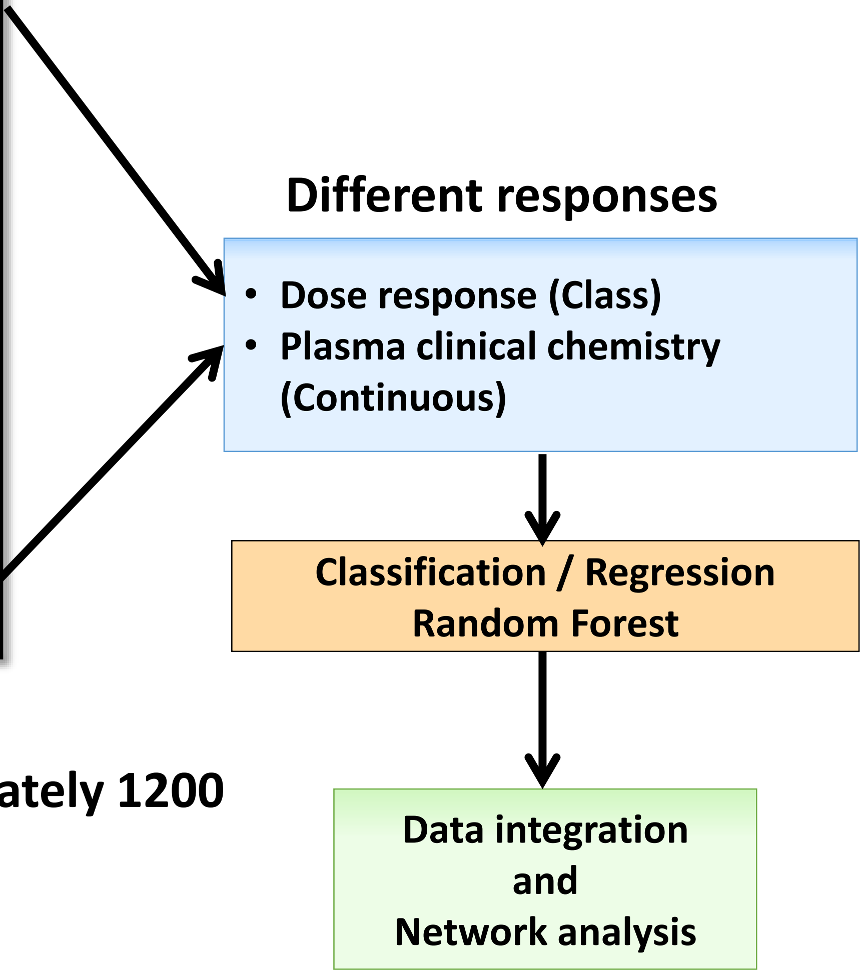
Intact Lipids LC-MS/MS

Liver Sample Extracts

Aqueous Metabolites LC-MS/MS

Intact Lipids DI-MS

Eicosanoids LC-MS/MS



Different responses

- Dose response (Class)
- Plasma clinical chemistry (Continuous)

Classification / Regression
Random Forest

Data integration
and
Network analysis

Total number of lipids/metabolites: Approximately 1200