Urszula Czerwinska; Systems Biology Lab of Benno Schwikowski

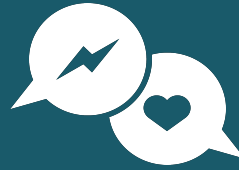# Pair biomarkers screening with isotonic regression classifier

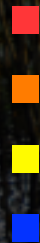Context   Results   Methods   Discussion   Perspectives   Conclusions

Context

WORLD POPULATION

DEN GUE

40 AT% RISK

Aedes aegypti

no specific medication or cure available for dengue
no vaccine against dengue
5 vaccines in clinical trials

Dengue Shock Syndrome (DSS)

Dengue Hemmoragic Fever (DHF)

Dengue Fever (DF)

Asymptomatic (AS)

Problematic diganostics

Dataset

📅 Sept 2011-Oct 2012

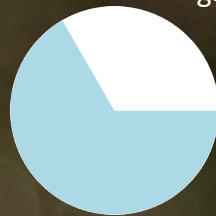📍 Krong Kampong Cham, Cambodia

🟥 13 (DSS)

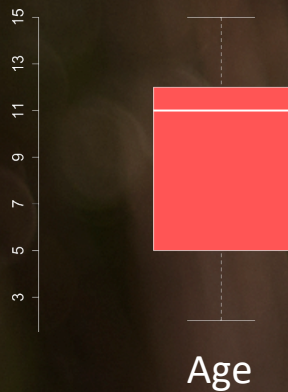🟧 13 (DHF)

🟨 13 (DF)

🟦 10 (AS)

Total 49

girls

boys

- platelet count
- DENV serotype
- hematocrit
- symptoms
- cholesterol
- triglycerides
- HDL
- VLDL
- bleeding...

Age

mRNA expression data Affymetrics probes of PBMC
70524 probes
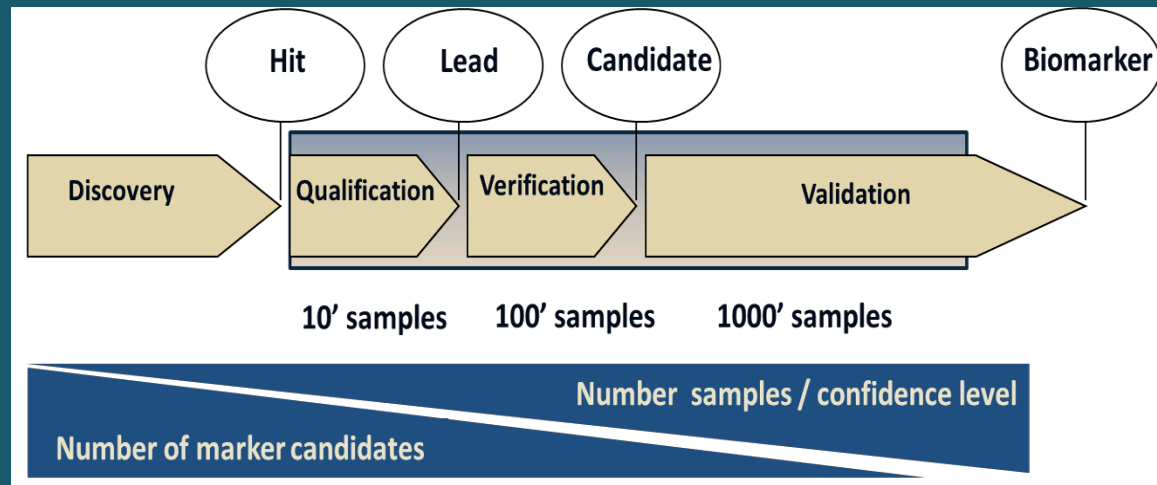70524x49 matrix

Pair biomarkers screening

What is a biomarker?

a measurable <u>indicator</u> of some biological state or condition[1]

==In vitro diagnostic test?

Types of biomarkers[3]:
- Diagnostic
- Prognostic
- Predictive
- Pharmacodynamic
- Recurrence

1. Strimbu, Kyle; Jorge, Tavel (2010). "What are Biomarkers?". *Current Opinion in HIV and AIDS* **5** (6): 463–466.
2. (Schemat) Fournier-Majoie Foundation for Innovation (2014)
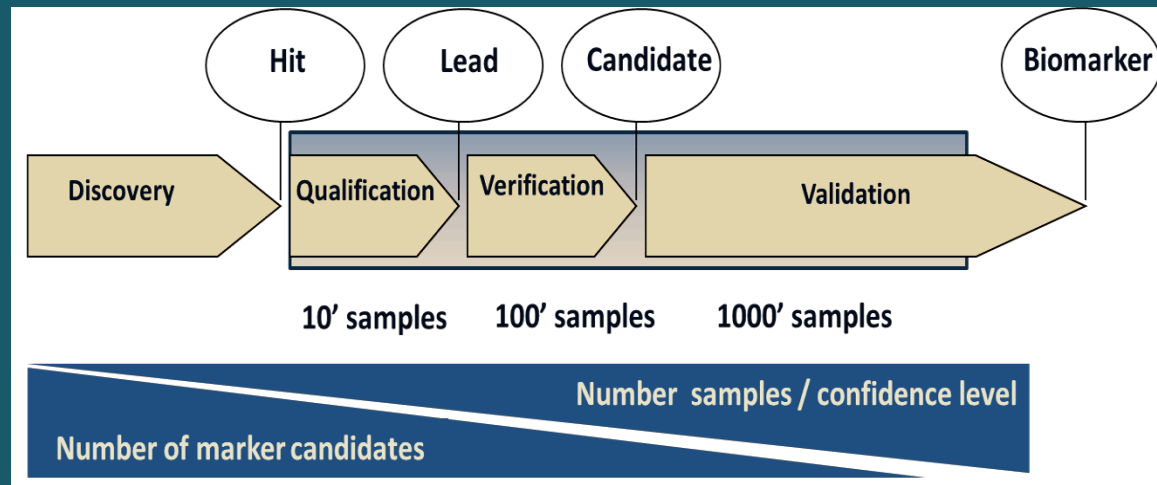3. Courtesy of Mariela Skendi and Antonio **V. Larrauri**

What is a biomarker?

a measurable <u>indicator</u> of some biological state or condition[1]

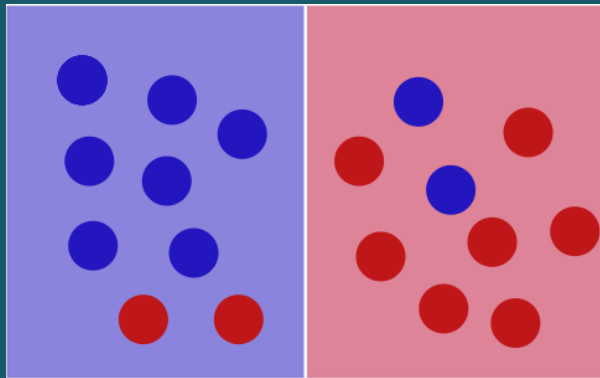*==In vitro* diagnostic test?

Types of biomarkers[3]:
- Diagnostic
- **<u>Prognostic</u>**
- Predictive
- Pharmacodynamic
- Recurrence

1. Strimbu, Kyle; Jorge, Tavel (2010). "What are Biomarkers?". *Current Opinion in HIV and AIDS* **5** (6): 463–466.
2. (Schemat) Fournier-Majoie Foundation for Innovation (2014)
3. Courtesy of Mariela Skendi and Antonio **V. Larrauri**

What is a *good* biomarker?

sensitivity > 0.9[1]
specificy > 0.9[1]
positive predictive value
reproductabilty

PPV= TP/ TP+FN

1. Brower V (March 2011). "Biomarkers: Portents of malignancy". *Nature* **471** (7339): S19–21.
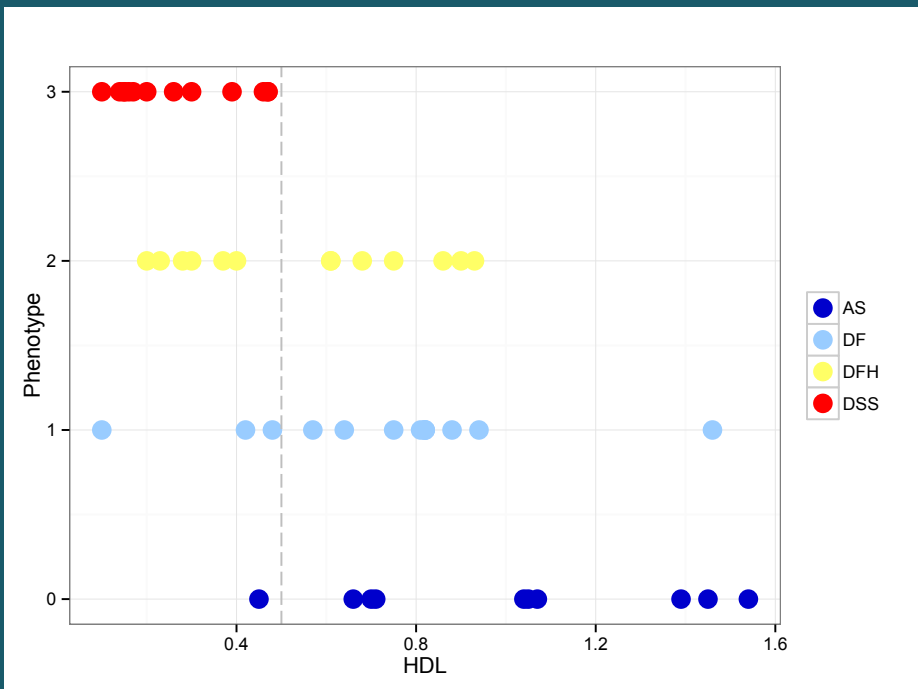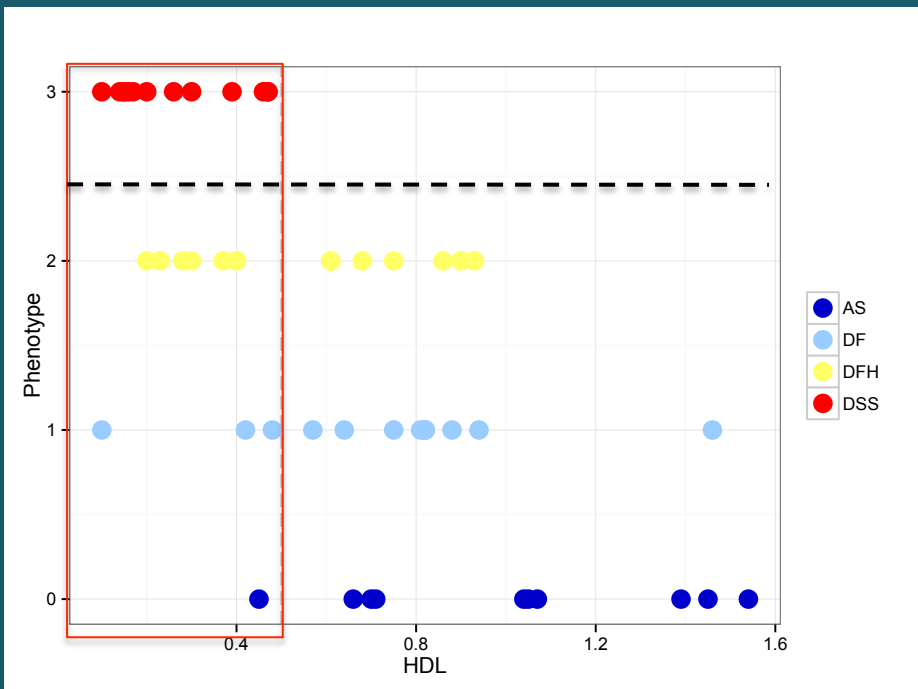
HDL as Dengue biomarker

GENETICS OF HUMAN RESPONSE TO INFECTION

Anavaj Sakuntabhai

HDL as Dengue biomarker

GENETICS OF HUMAN RESPONSE TO INFECTION
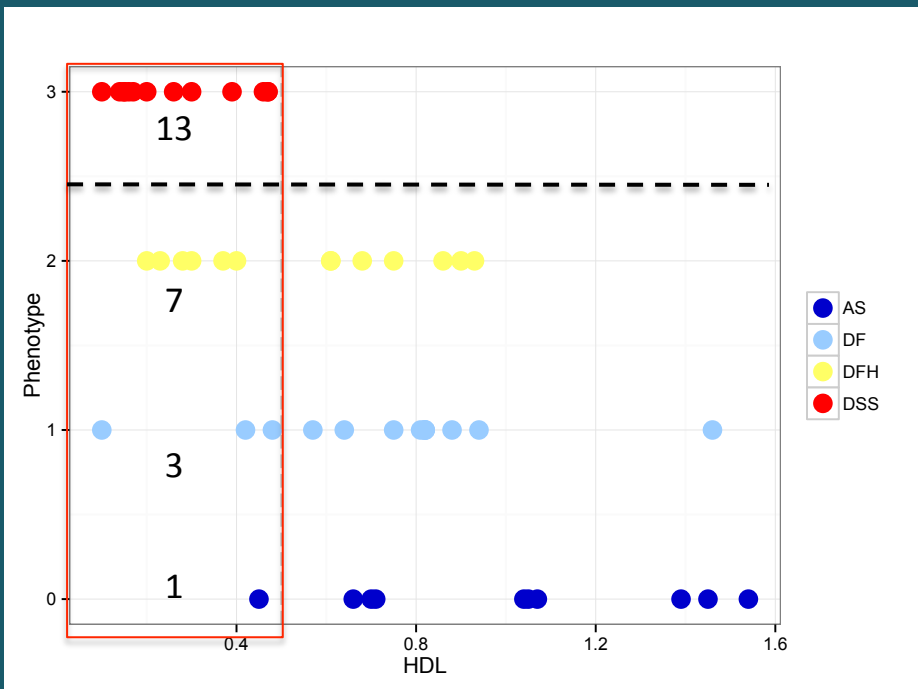
Anavaj Sakuntabhai

HDL as Dengue biomarker

GENETICS OF HUMAN RESPONSE TO INFECTION

Anavaj Sakuntabhai

TP = 13
FP = 0
Sensitivity = 1
Specificity = 0.69
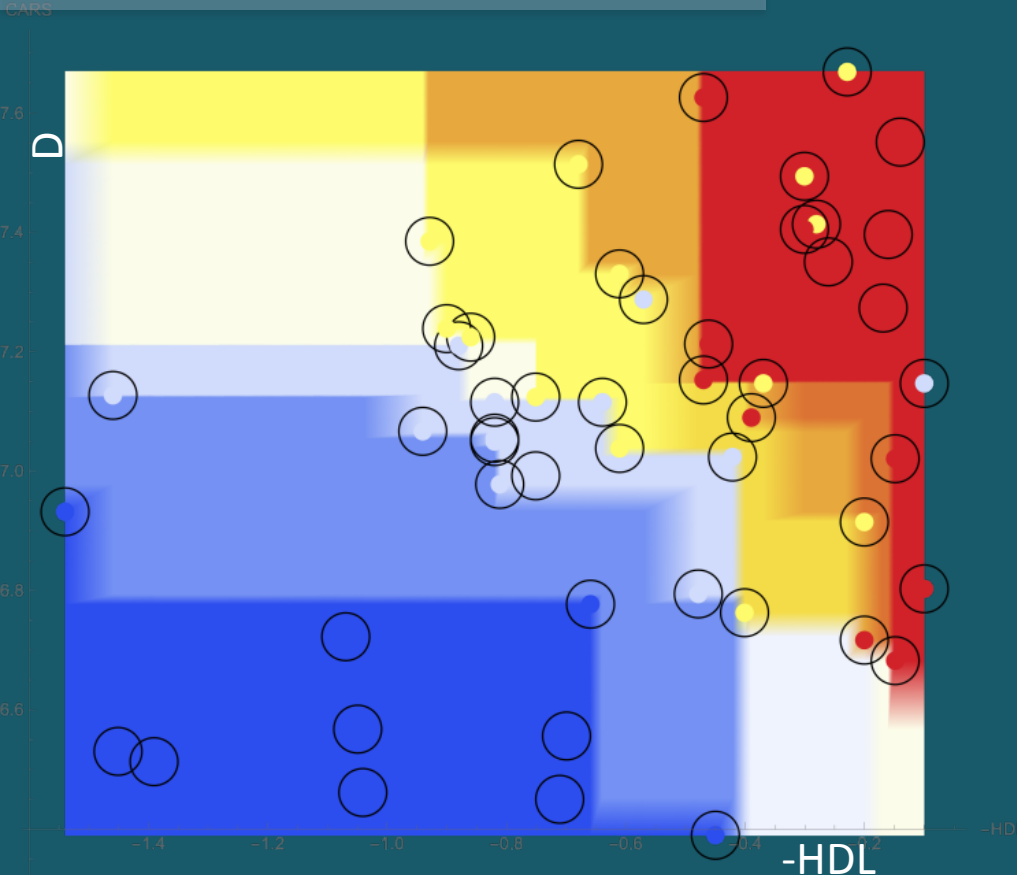PPV = 0.5

Pair biomarker screening
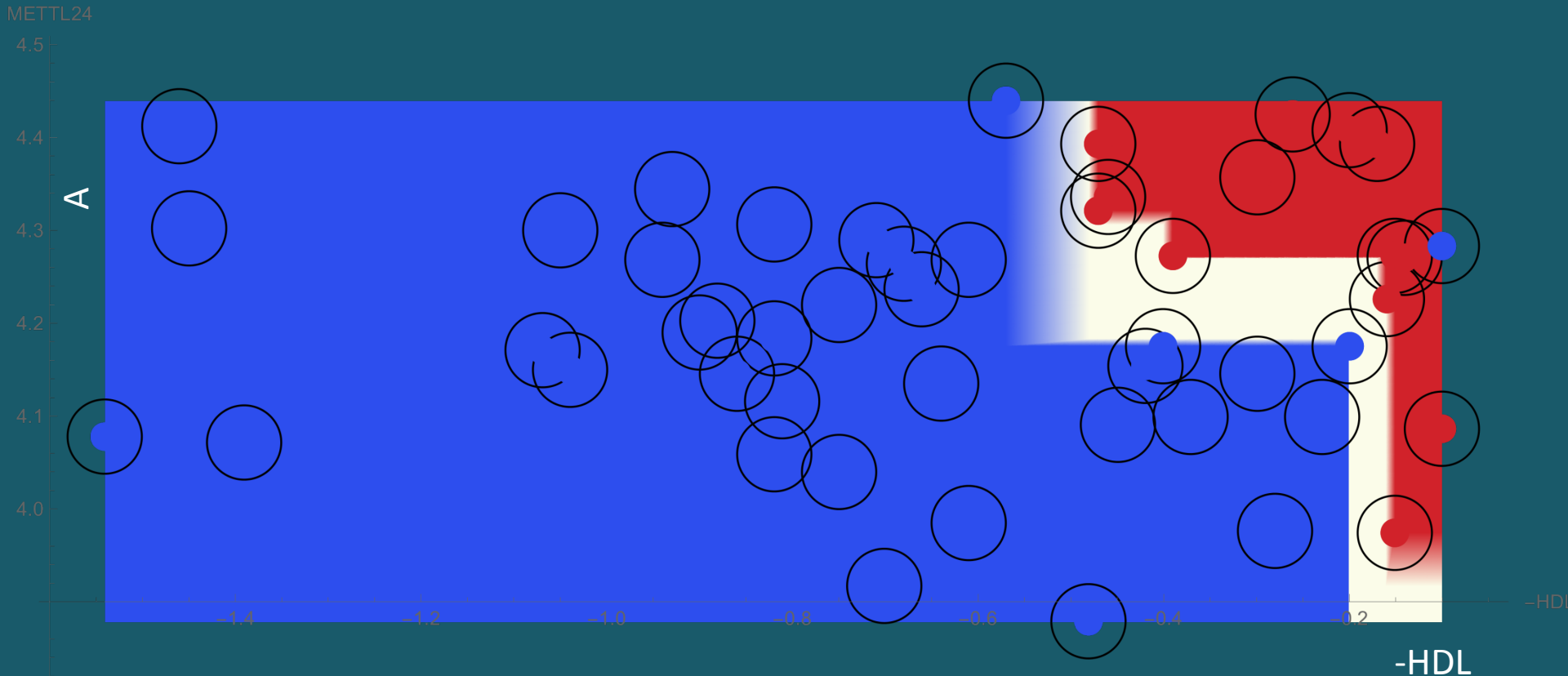

SYSTEMS BIOLOGY

Benno Schwikowski

Feature: D *top secret*

Mean cross validation error over 100 cross validations (leave 5 out) is 5.54.

Mean global fitting error is 23.05.

- 🔴 DSS
- 🟡 DHF
- 🔵 DH
- 🔵 AS

DSS vs DHF biomarker pairs *with HDL*

Feature: A *top secret*

Mean cross validation error over 100 cross validations (leave 5 out) is **0.33.**
Mean global fitting error is **1.22**.

DSS vs DHF biomarker pairs *with HDL*

A
E   *top secret*

F[2]

G[2]   *top secret*
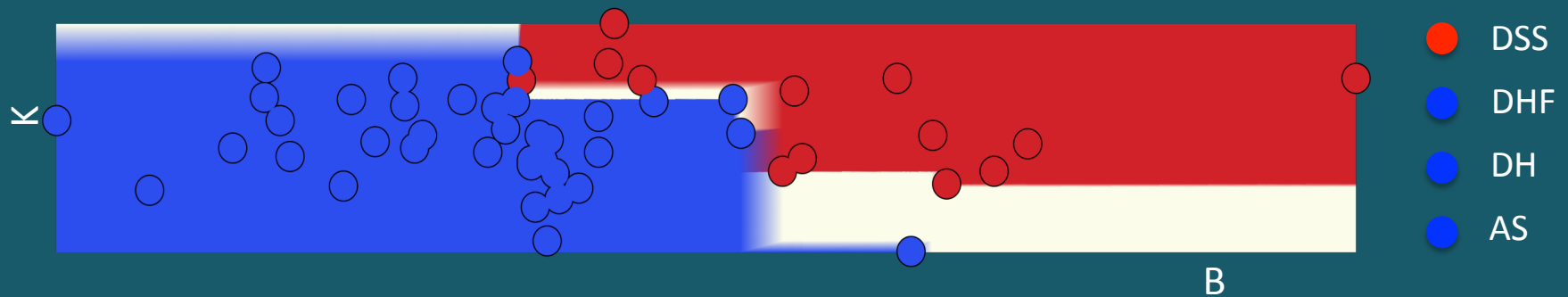
H[2]

I[1]
J[1]   *top secret*

GO terms

o   helix loop helix DNAbinding
o   mitochndrial membrane
o   liver
o   translational activity factor
o   *top secret*
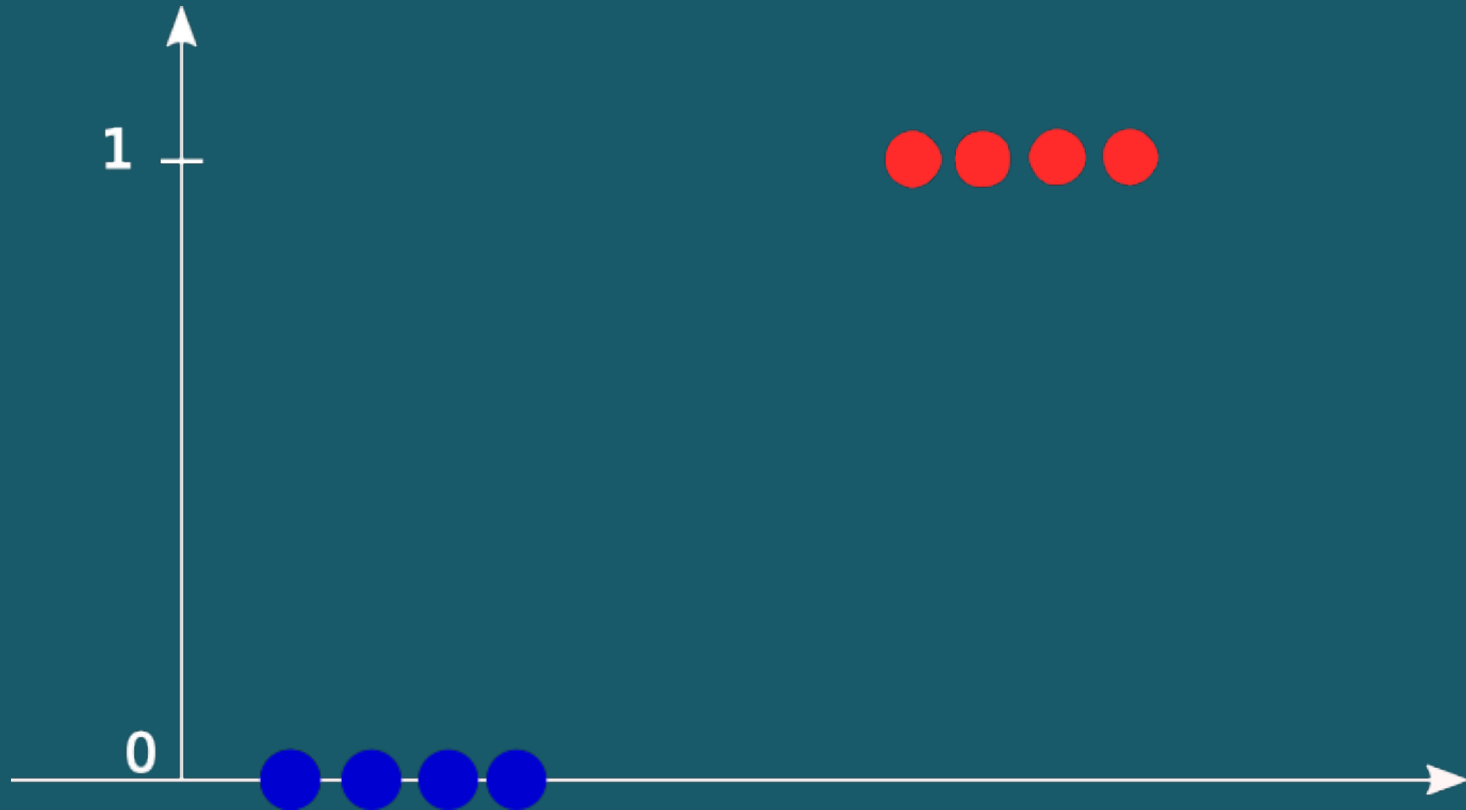o   cholesterol biosynthetis
o   NADPH activity

1.   Kwissa M et al. Dengue virus infection induces expansion of a CD14(+)CD16(+) monocyte population that stimulates plasmablast differentiation. Cell Host Microbe 2014 Jul 9;16(1):115-27.
2.   Tantibhedhyangkul Wet al. Orientia tsutsugamushi stimulates an original gene expression program in monocytes: relationship with gene expression in patients with scrub typhus. PLoS Negl Trop Dis 2011 May;5(5):e1028

DSS vs DHF biomarker pairs

what if we forget about HDL?

Feature 1: B *top secret*

Feature 2: K *top secret*

Mean cross validation error over 100 cross validations (leave 5 out)
is **0.22.**
Mean global fitting error is **0.32.**

DSS vs DHF biomarker pairs

what if we forget about HDL?

L[2]

M[1]   *top secret*

B[3,4]   *top secret*

K

1. Kwissa M et al. Dengue virus infection induces expansion of a CD14(+)CD16(+) monocyte population that stimulates plasmablast differentiation. Cell Host Microbe 2014 Jul 9;16(1):115-27.
2. Tantibhedhyangkul Wet al. Orientia tsutsugamushi stimulates an original gene expression program in monocytes: relationship with gene expression in patients with scrub typhus. PLoS Negl Trop Dis 2011 May;5(5):e1028
3. Tolfvenstam T et al. Characterization of early host responses in adults with dengue disease. BMC Infect Dis 2011 Aug 2;11:209
4. Lin, Ren-Jye, et al. "Distinct antiviral roles for human 2', 5'-oligoadenylate synthetase family members against dengue virus infection." The Journal of Immunology 183.12 (2009): 8035-8043.

GO terms
o transmembrane
o mitochondria
o B
o Zinc fingers
o RNA binding immunoglobulin
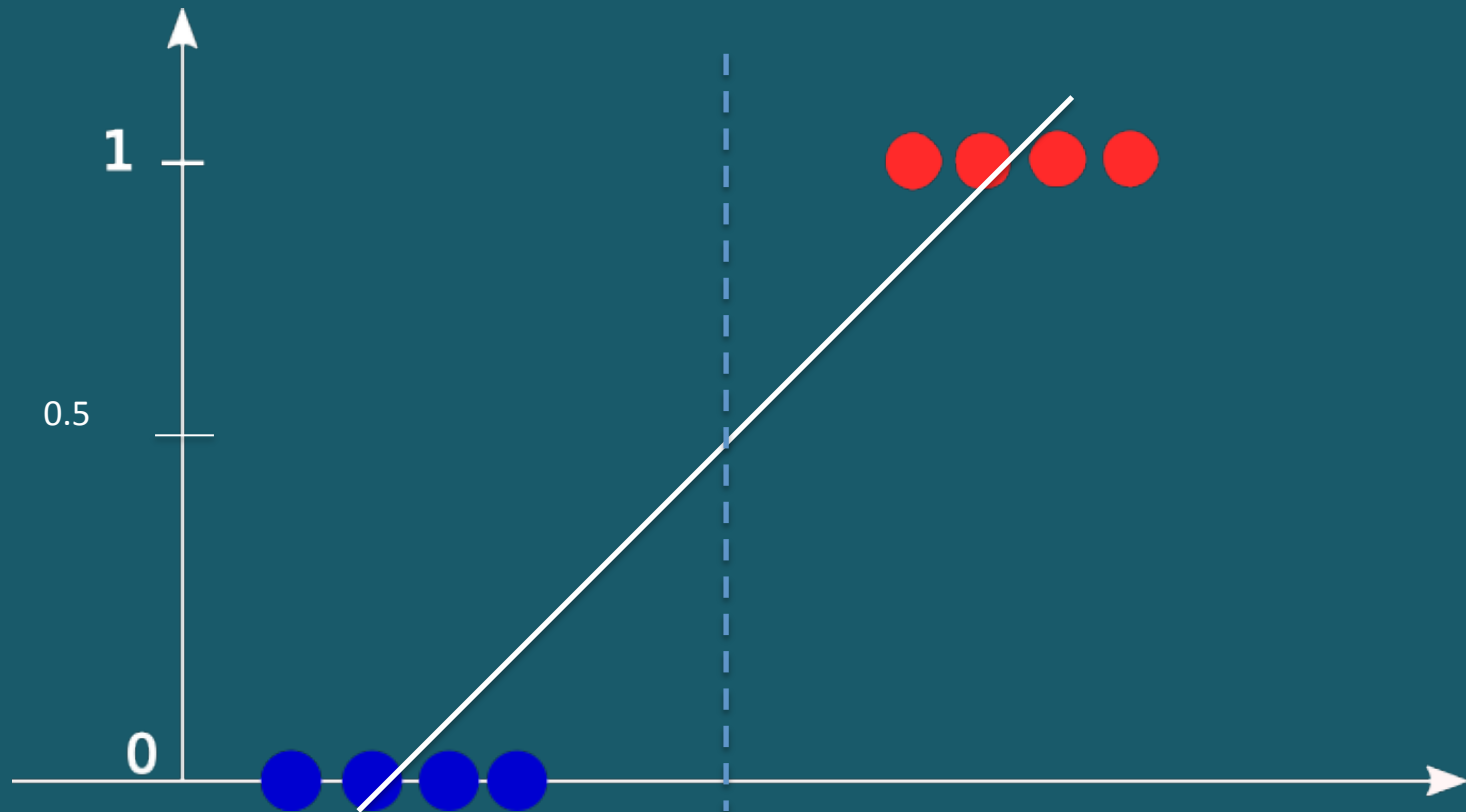o G-protein coupled signaling pathway
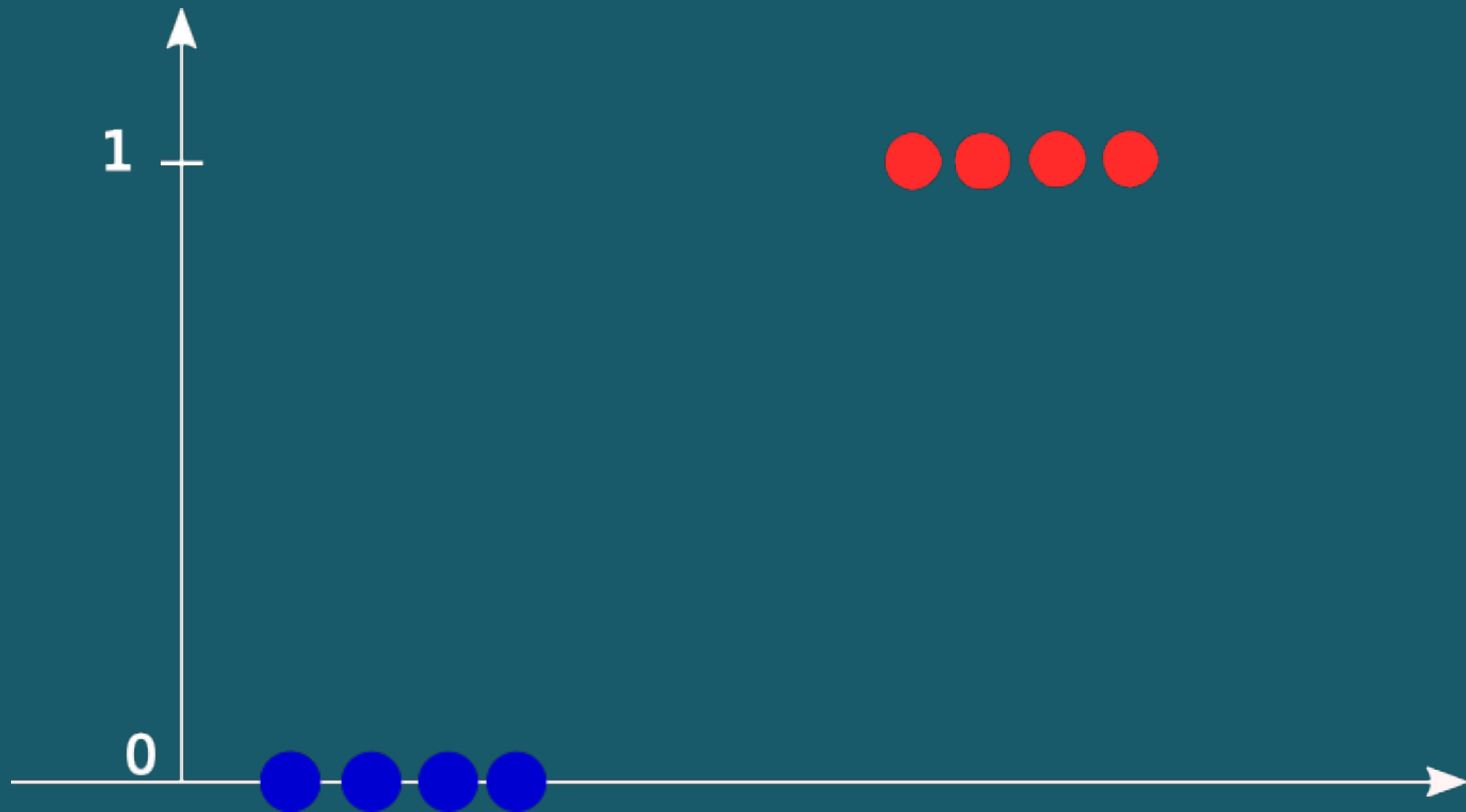
Methods

Linear regression
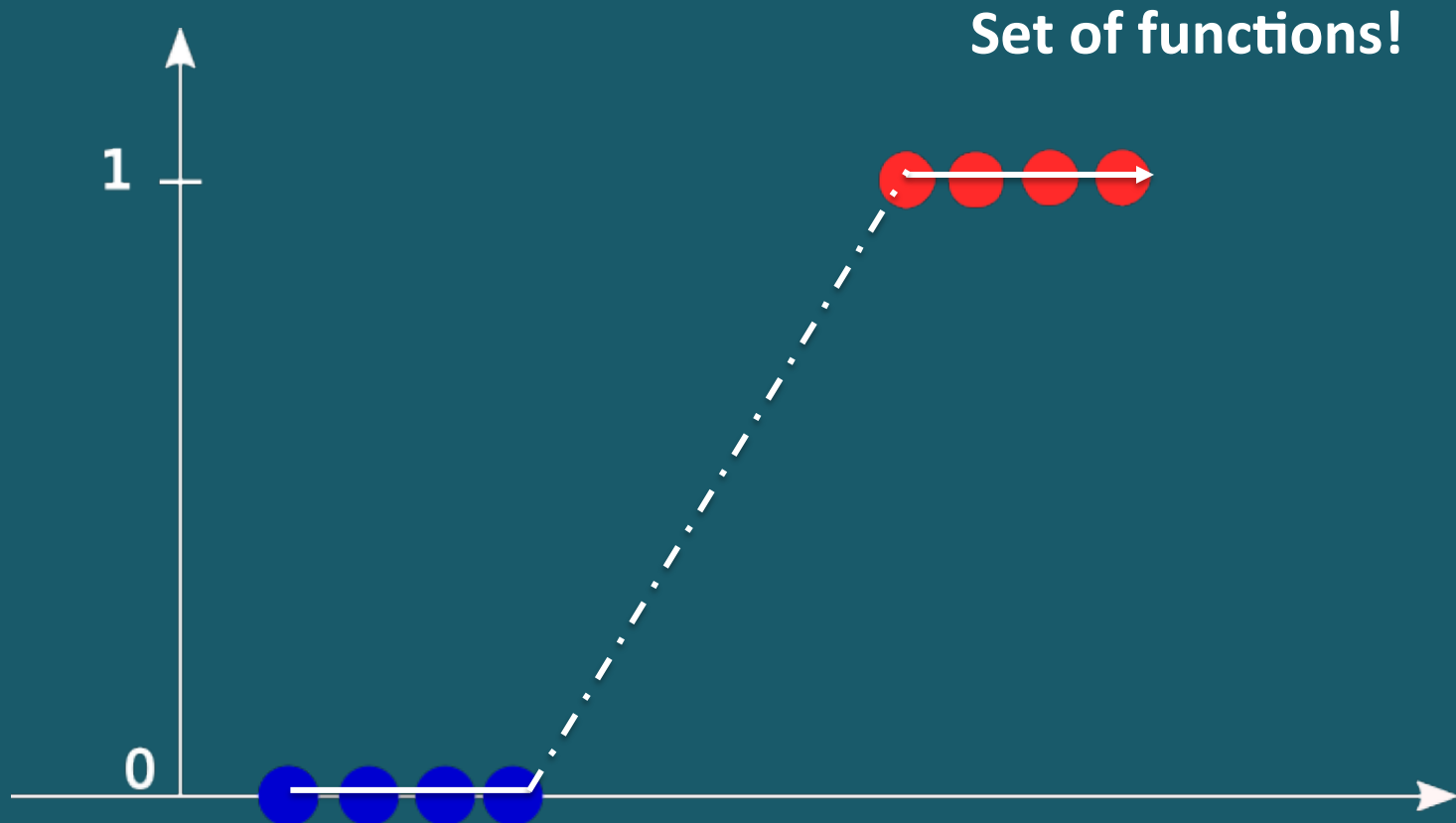
Linear regression

$$y = \alpha + \beta x$$

Isotonic regression

Isotonic regression

Isotonic regression
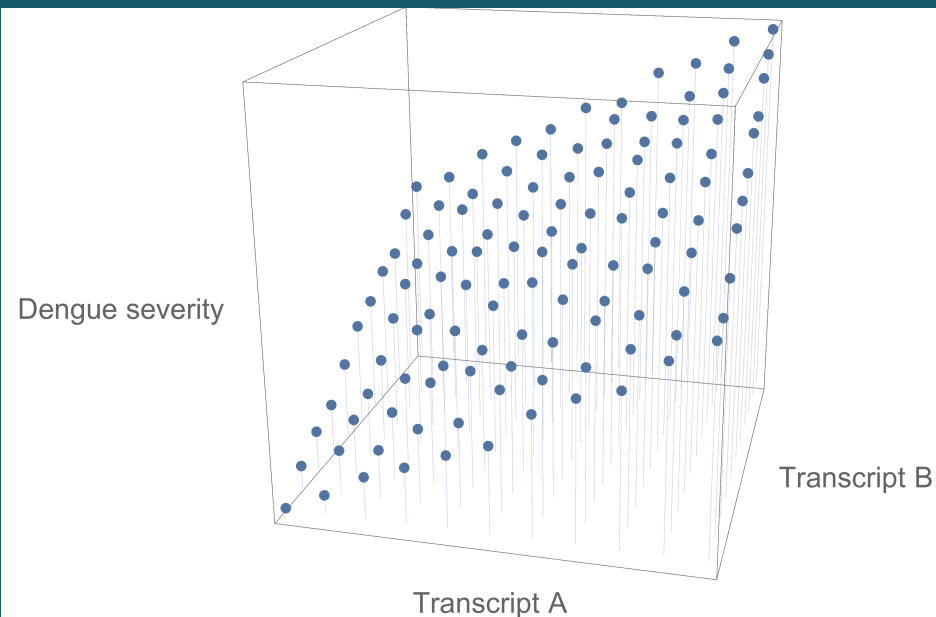
Set of functions!

$$x_1 \leq x_2 \leq \ldots \leq x_n \qquad f(x_1) \leq f(x_2) \leq \ldots \leq f(x_n)$$
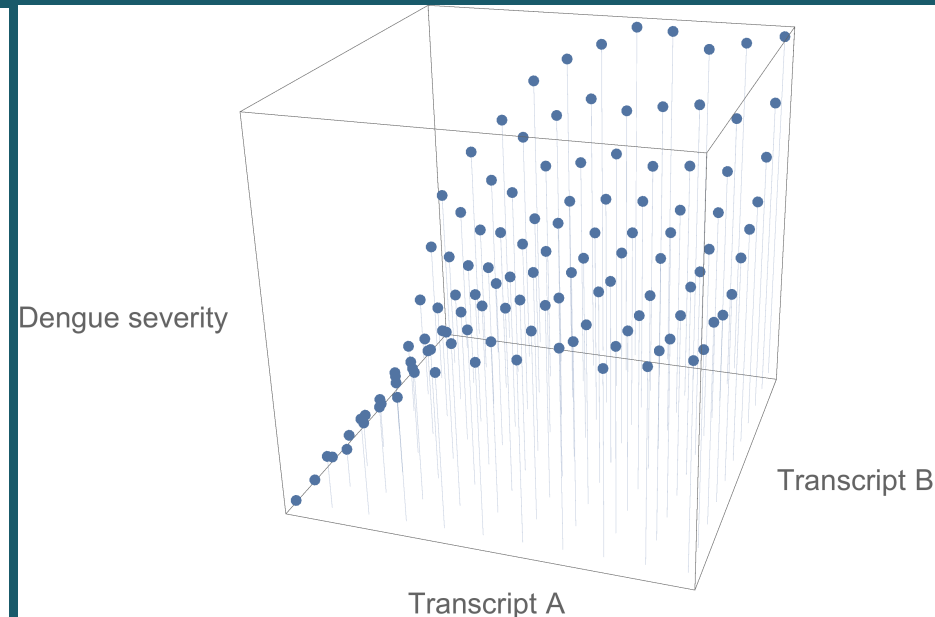
$$\min_g \sum_{i=1}^{n} w_i(g(x_i) - f(x_i))^2$$

Linear vs. isotonic regression

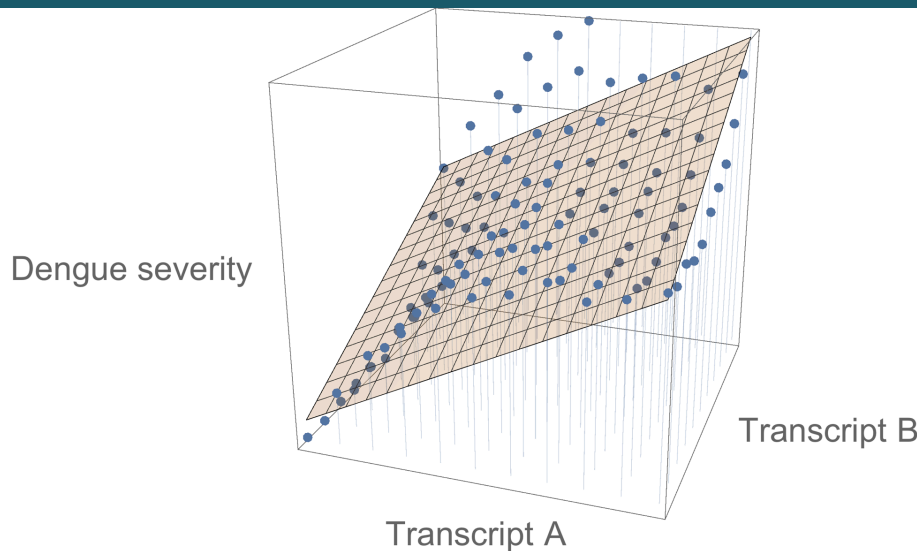**Interacton between transcripts**
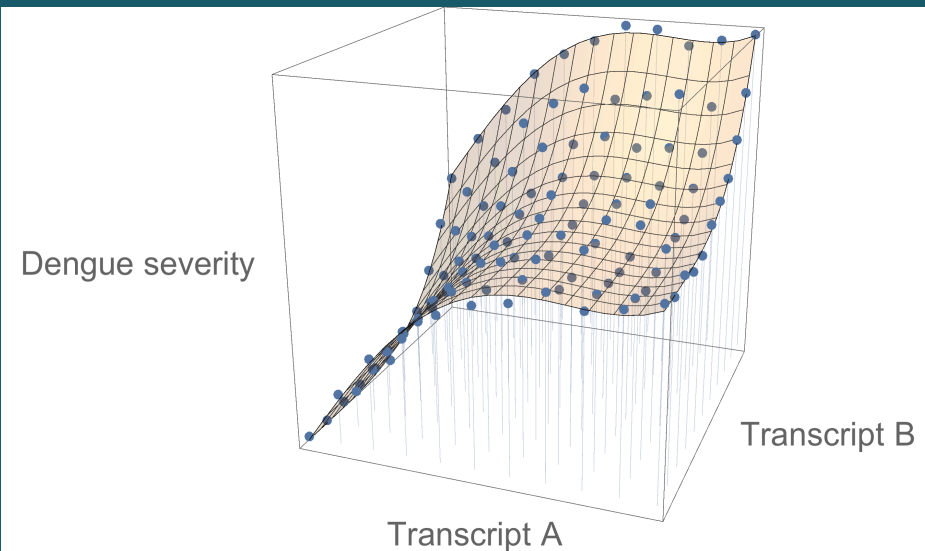
Linear interaction

Nonlinear, monotonic interaction

Linear vs. isotonic regression

Monotonic regression
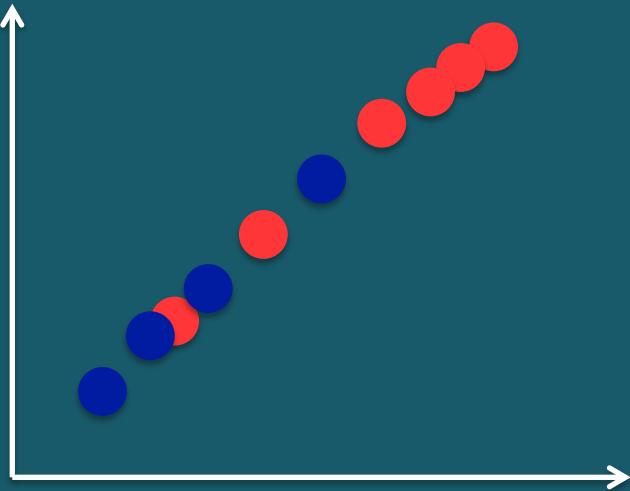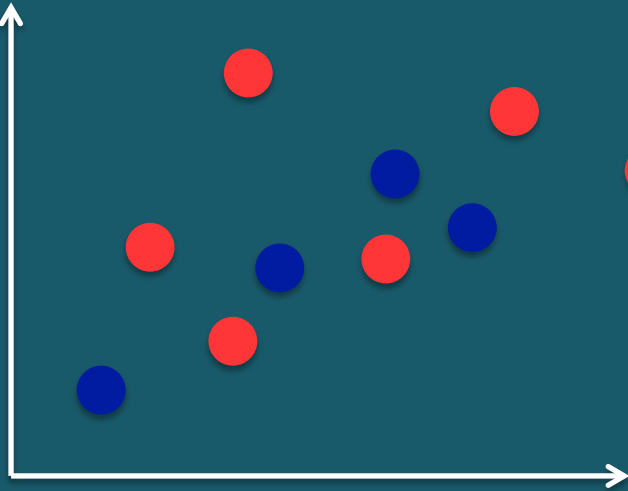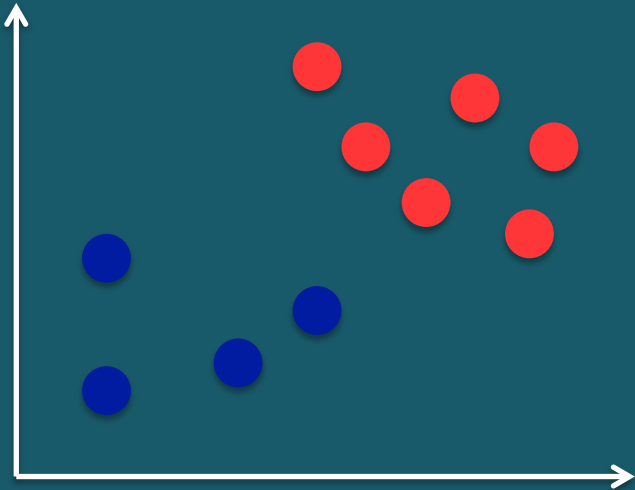
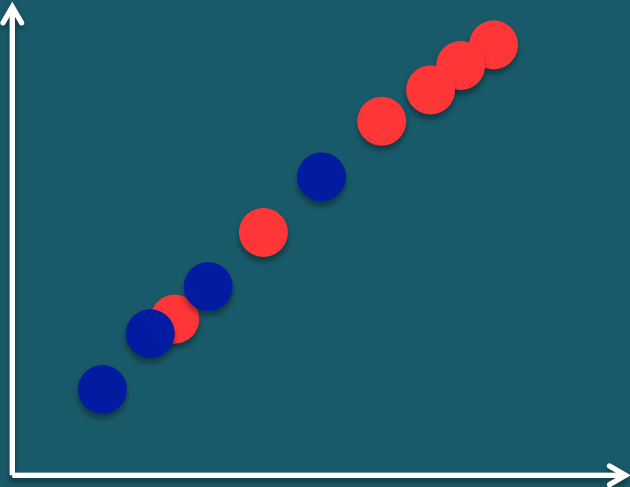Linear regression:
Inaccurate

Monotonic regression:
Accurate

Isotonic or not?

Isotonic or not?

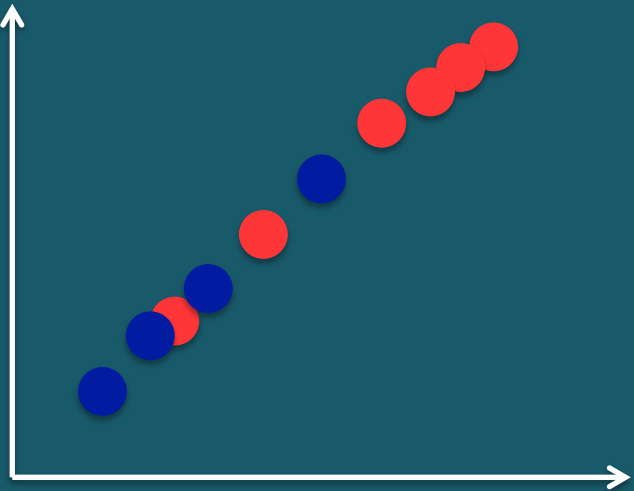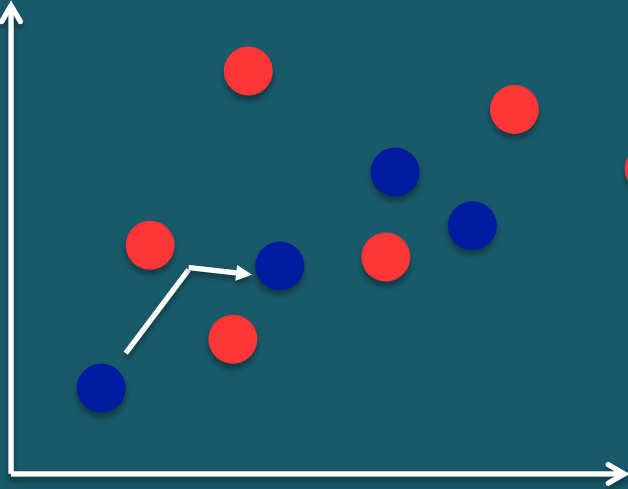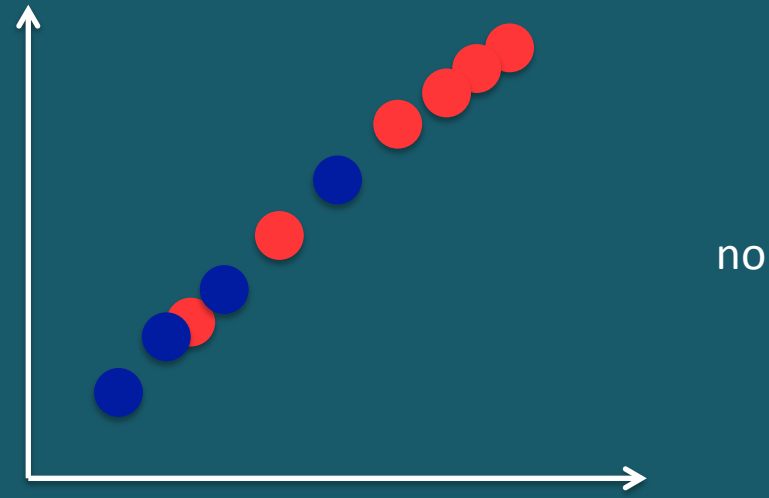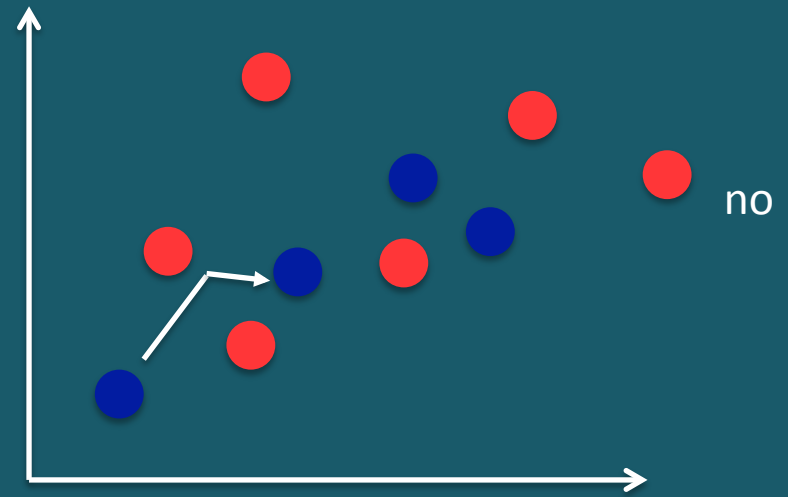# Isotonic or not?

Isotonic or not?

From a bike to Formula1

This appears in *Algorithmica* 66 (2013), pp. 93–112.

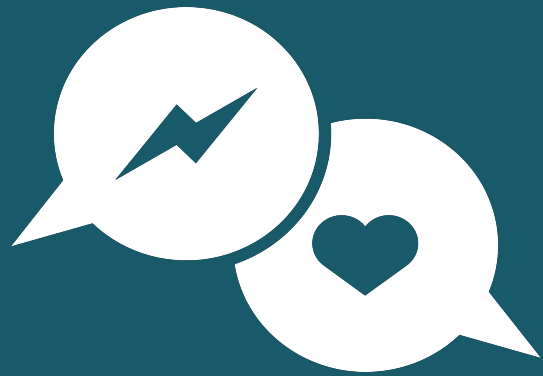## Isotonic Regression via Partitioning

Quentin F. Stout

Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109–2121   USA

$$n^3 \longrightarrow n\log^2 n$$

Discussion

Classifiers everywhere

SVN

Naive base

PCA

Logit Boost

Logistic regression pairs

Logistic regression lasso/elmap

Classifiers everywhere

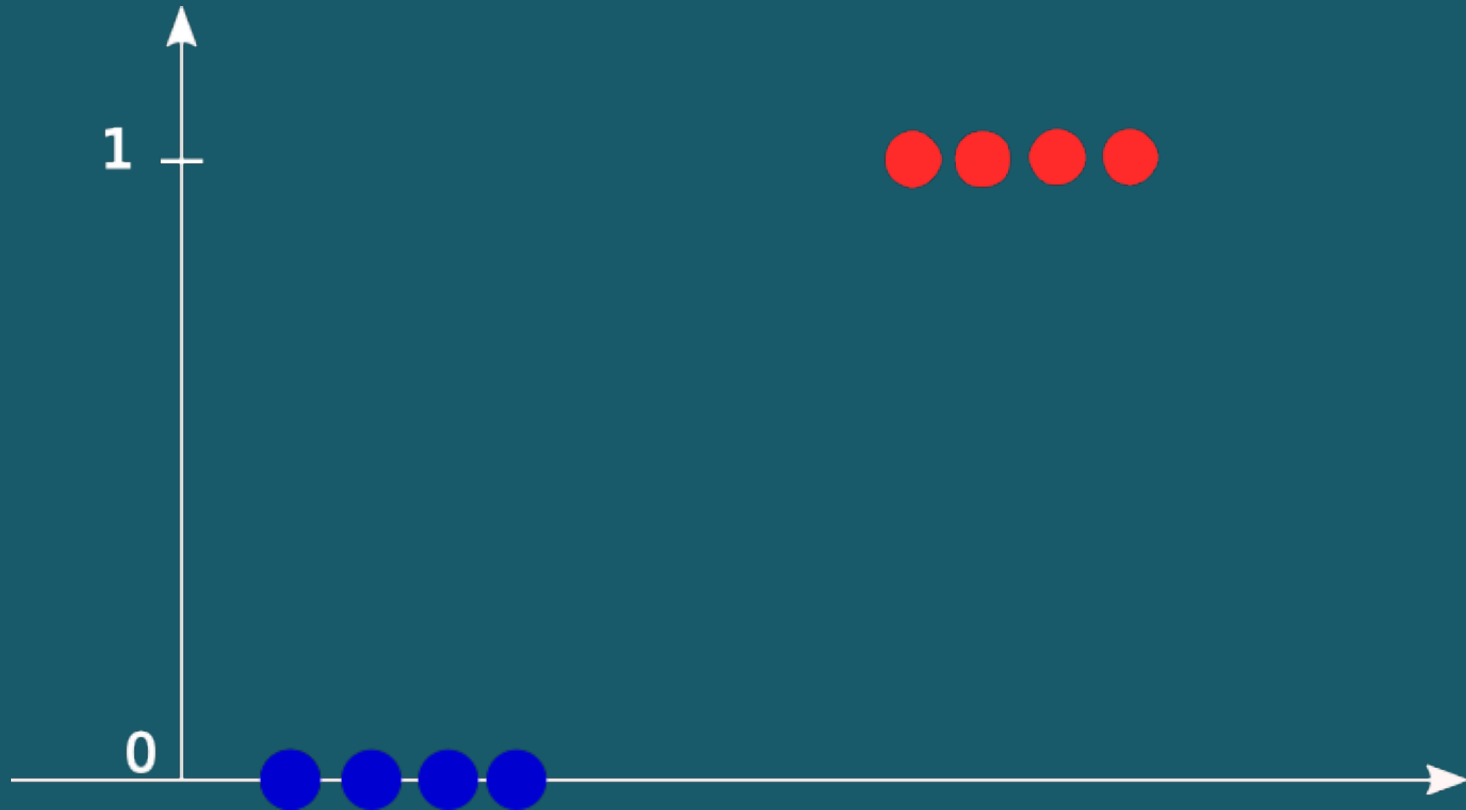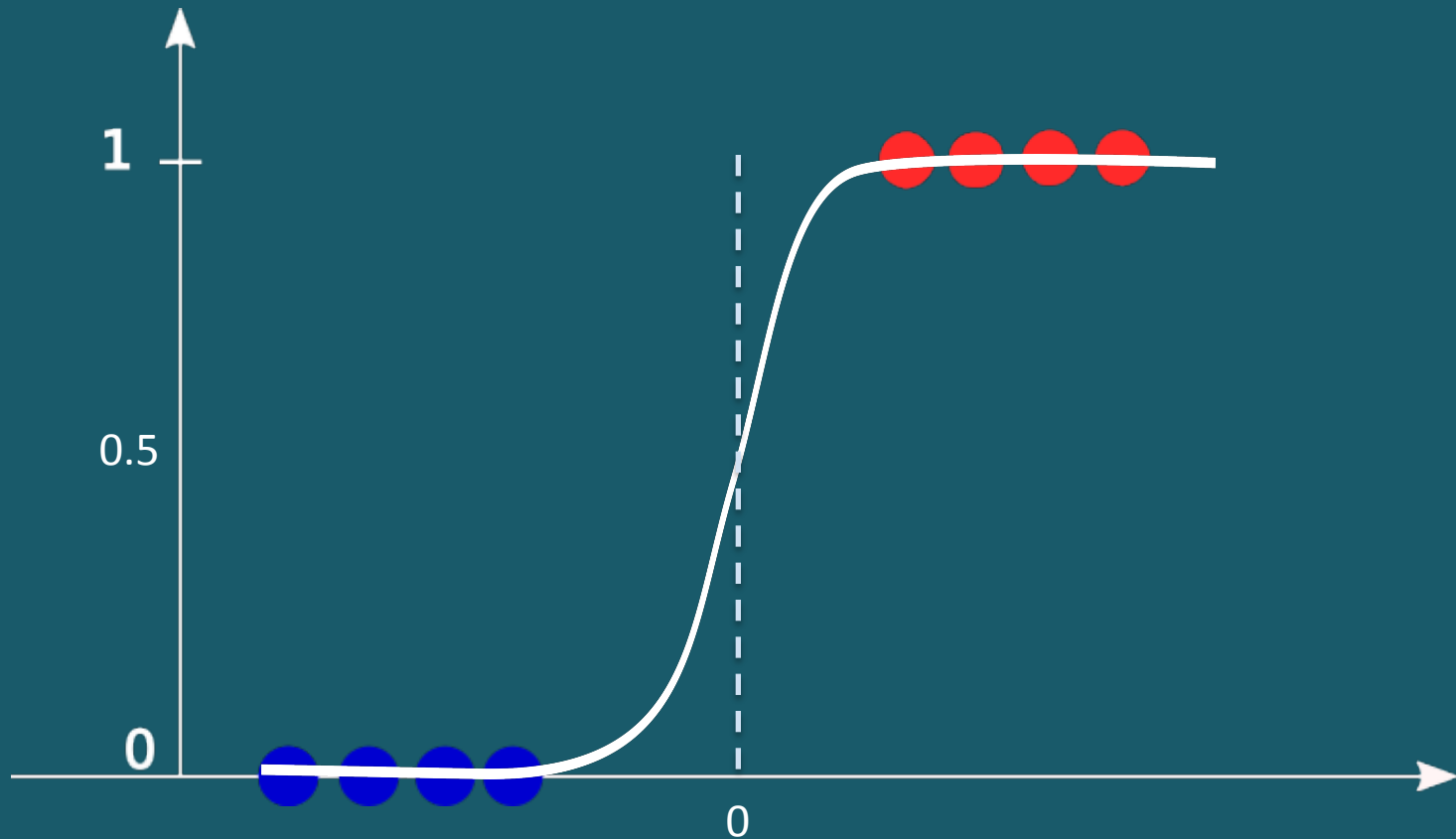|	
SVN	too big subsets (70 genes), too slow
Naive Bayes	too big subsets (76 genes), poor performance AUC=0.65
PCA	very poor performance
Logit Boost	too big subsets (76 genes) poor performance AUC=0.65
Logistic regression pairs	too slow (229 weeks)
Logistic regression lasso/elmap	intersting candidate 2 good estimators

Logistic regression

Logistic regression
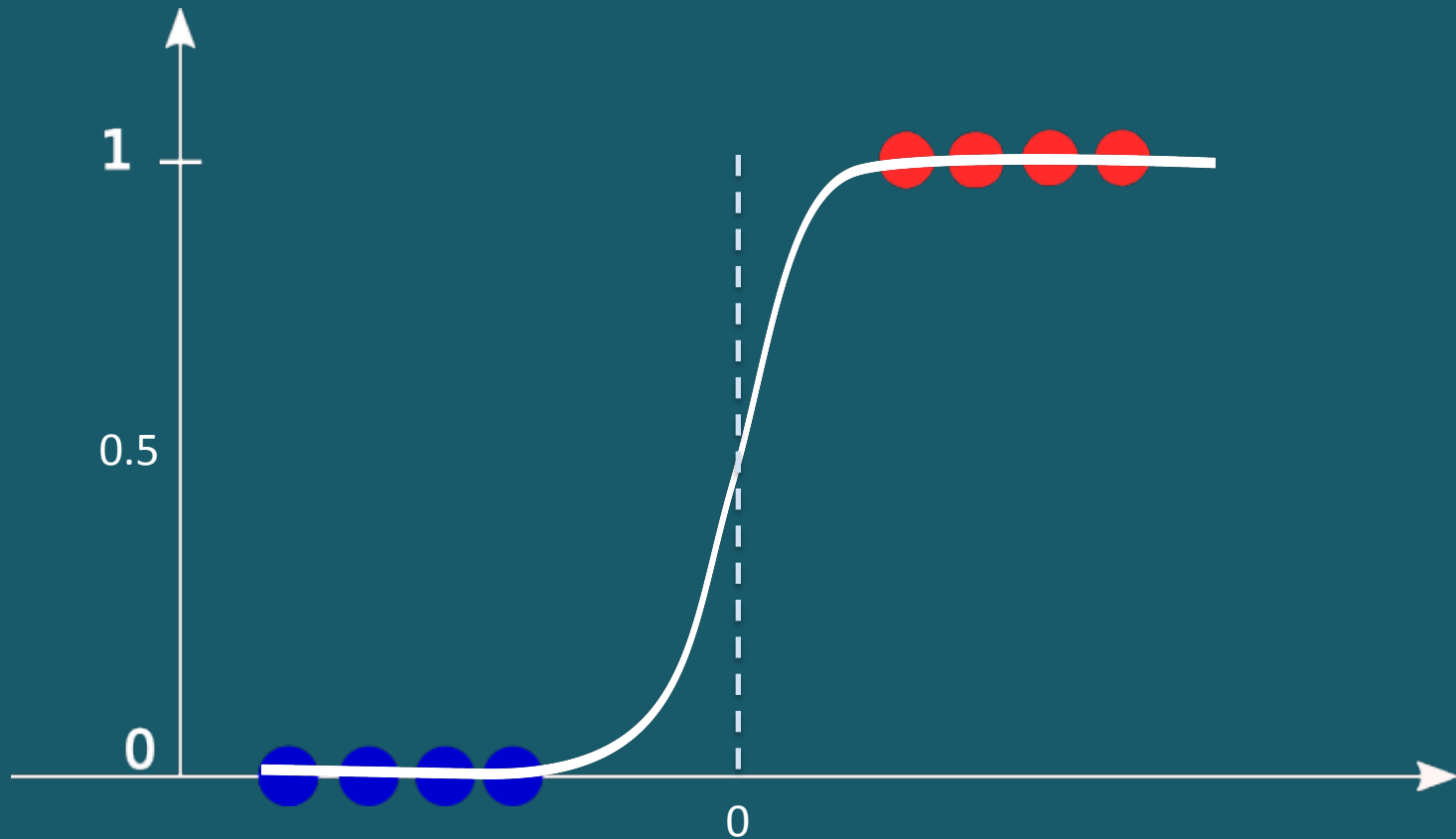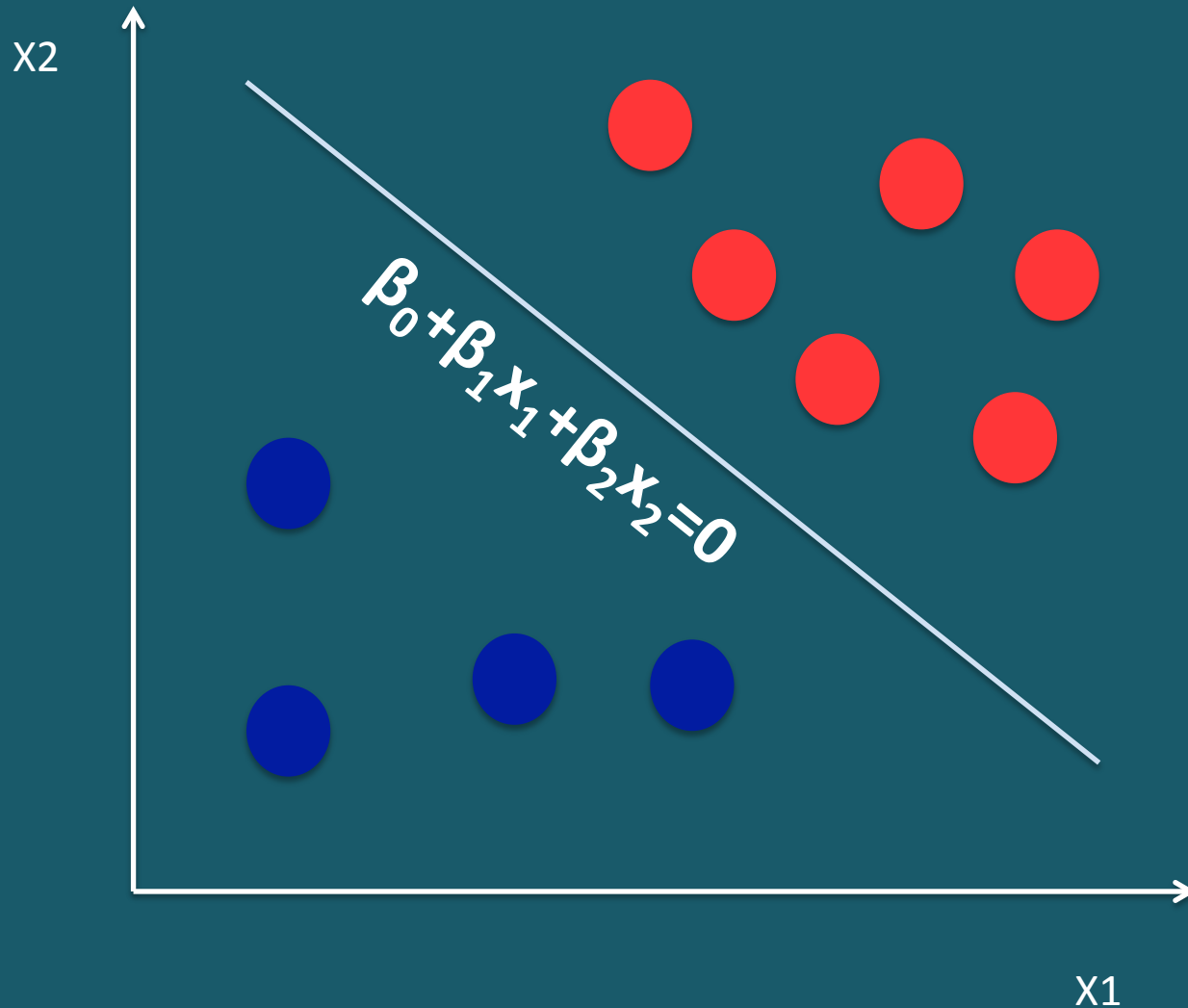

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Logistic regression

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Logistic regression

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

X2

X1

Logistic regression

Logistic regression

DSS

DHF

DH

AS

Mean cross validation error = 1.3

TP = 11
FP = 1
FN = 1
TN = 36
PPV = 0.91
Sensitivity = 0.91
Specificity = 0.97

Isotonic regression

Feature 1: C *top secet*

Feature 2: Q *top secret*

Mean cross validation error over 100 cross validations (leave 5 out) is 0.83.`

Mean global fitting error is 1.72.

What about p-value?

CDF of the best score in 1000 phenotype pemutaions
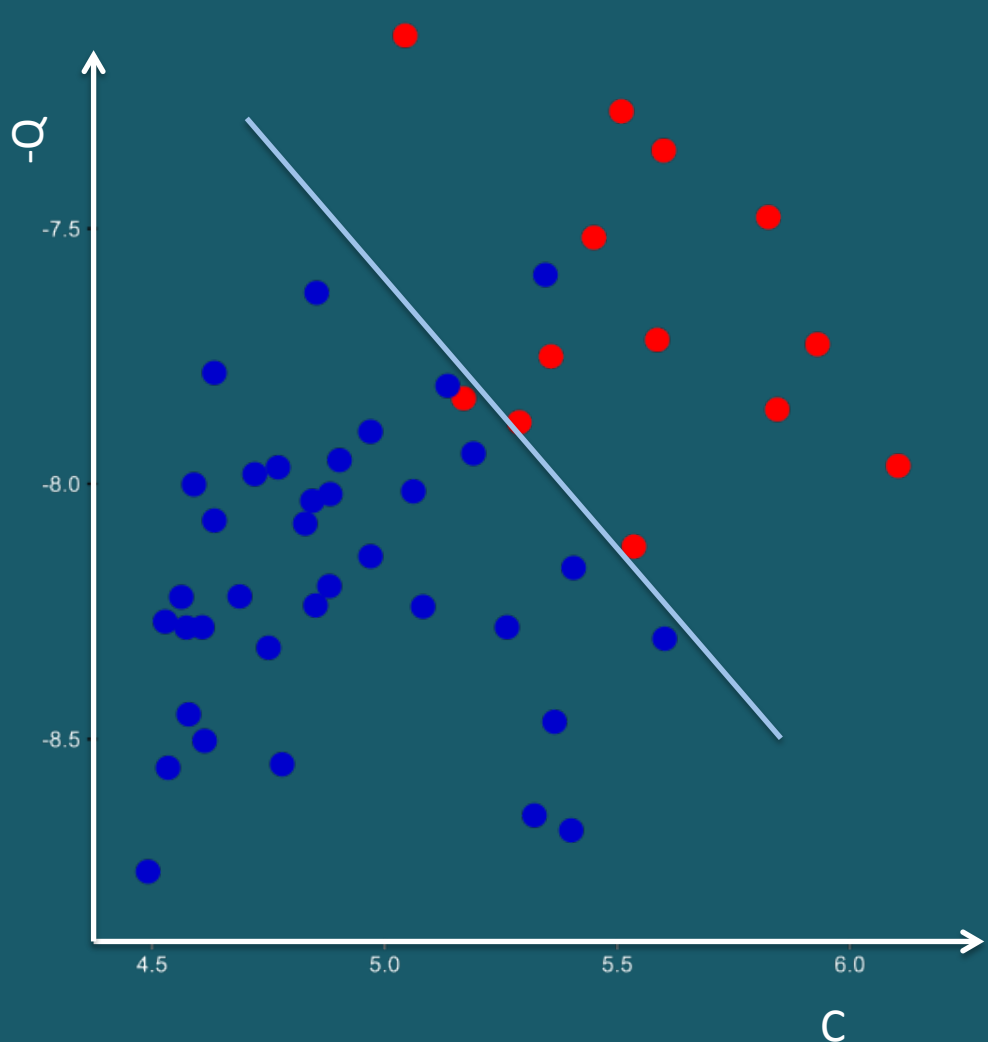

our best score

p-value estimation

$$\frac{\text{number of scores} \leq \text{our score}}{\text{number of trials}}$$

Perspectives

Back to the lab

qRT PCR

Interpret the hypothesis of A

Mesure the platelet volume

Study non coding RNA

Study the role of B family in Dengue severity groups

Theoretical extentions

P.values distribution for transcript pairs

Multiple testing correction

AUC

Mutliple biomarkers testing: 3, 4...

Test the method with other data

- more individuals
- continuous phenotype

Take home message

Dengue is a dangerous disease and there is no working cure so far

We discovered pairs of transcripts classifying Dengue severity groups with no error

Computing p-values is not always an easy task

Isotonic regression is nonlinear generalisation of regression

Higher dimensionality of isotonic regression are to be tested

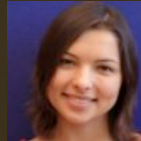Isotonic regression can be very fast now

Thank you

SYSTEMS BIOLOGY
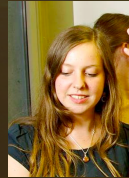
GENETICS OF HUMAN
RESPONSE TO INFECTION

Benno Schwikowski

Iryna Nikolayeva
FdV

Anavaj Sakuntabhai

Oriol Guitart Plat

Urszula Czerwinska
AIV

Etienne Simon-Loriere

urszula.czerwinska@cri-paris.org

AIV

Institut Pasteur

Context   Results   Methods   Discussion   Perspectives   Conclusions

a)

classe	n
DSS	13
DHF	13
DH	13
AS	10

Color Key

high   low

Expression level

p =70 000

Transcripts
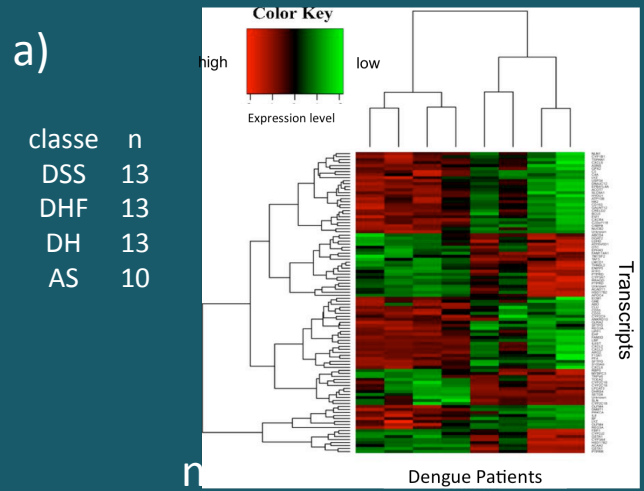
Dengue Patients
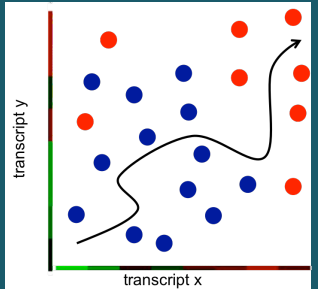
matrix of transcripts
«class» file as weights

b)   C code

Isotonic regression model built
for all pairs of transcripts (70 000^2)
Cross-validation (leave-five-out) x 100
Training: 44, **Test**: 5

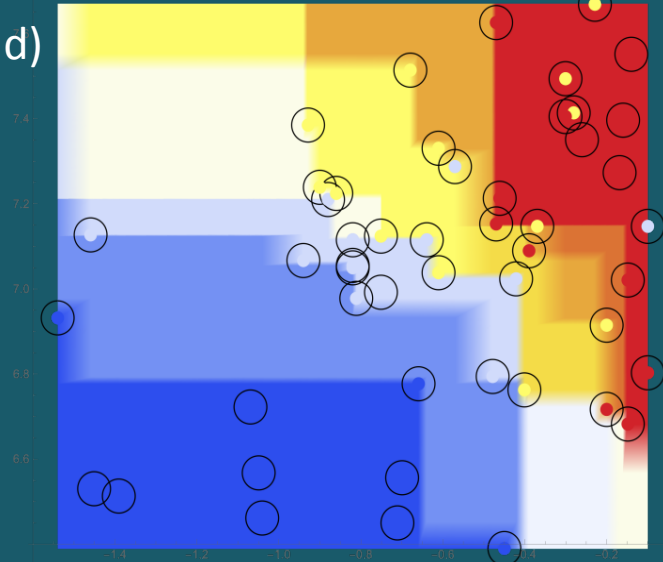$$\min_g \sum_{i=1}^n w_i (g(x_i) - f(x_i))^2$$

g – estimated function
f – known function
w – weights

expression data
classes
imported to *Mathematica*

transcript y

transcript x

text file

d)

Ranking
imported to
Mathematica

DSS
DHF
DH
AS

Wolfram
Mathematica

c)

	Output			
rank	cross-validation	fitting	transcript x n°	transcript y n°
1	58	58	8	7
2	77	188	0	1667
3	90	100	56	9876
4	120	123	9865	23