Tackling the topology and geometry underlying big data

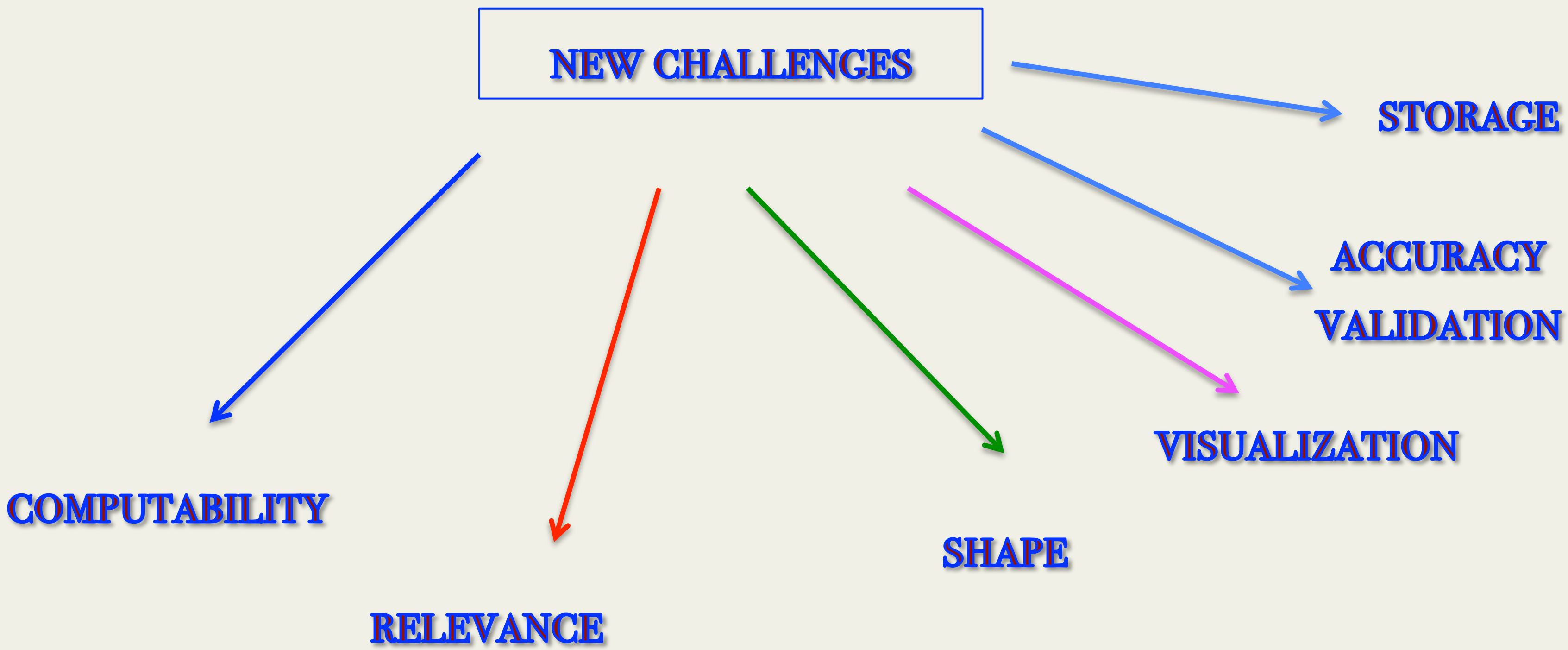
Monica Nicolau

Department of Mathematics
Department of Microbiology Immunology
&
Center for Cancer Systems Biology
Stanford University

Big data

Big data

Big data


NEW CHALLENGES
STORAGE
ACCURACY
VALIDATION
VISUALIZATION
COMPUTABILITY
RELEVANCE
SHAPE

Big data

NEW CHALLENGES

RELEVANCE

SHAPE

VISUALIZATION

Big data

NEW CHALLENGES

RELEVANCE

GEOMETRY
+
STATISTICS

SHAPE

TOPOLOGY
+
STATISTICS

VISUALIZATION

A little bit of history…

The Human Genome Project: *1990 – 2003* **DOE & NIH**

international effort to discover all the estimated **20,000-25,000** human genes.

determine the complete sequence of the **3 billion** *DNA* subunits

Data is large

Big data

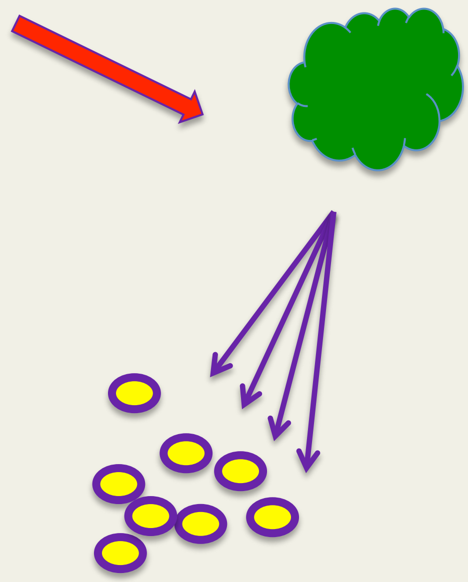
LOCALLY RICH

GLOBALLY MESSY

Nolan Lab
Monack Lab
Stanford

example:

Smoothing a hairball

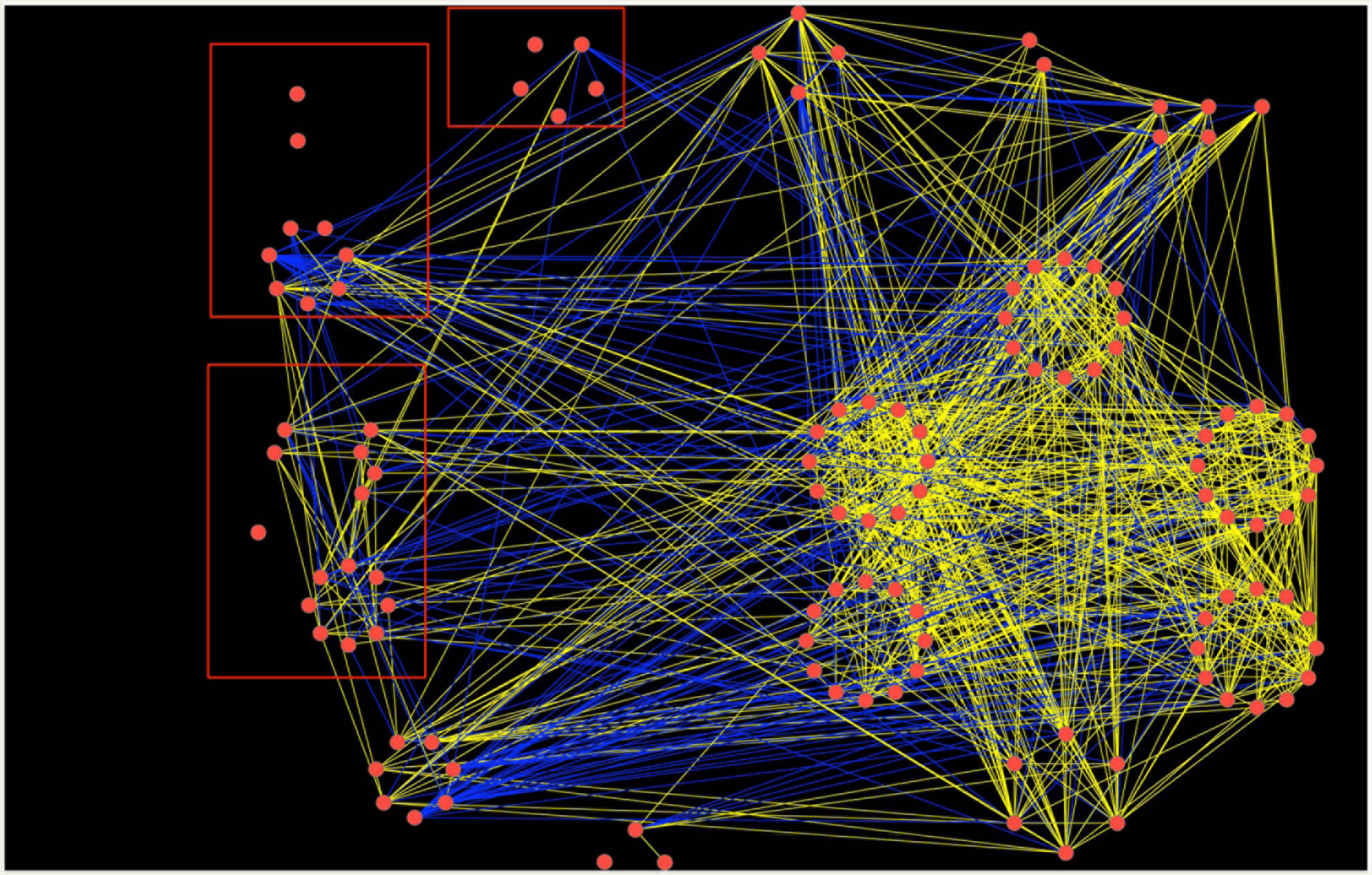


tissue sample

individual cells

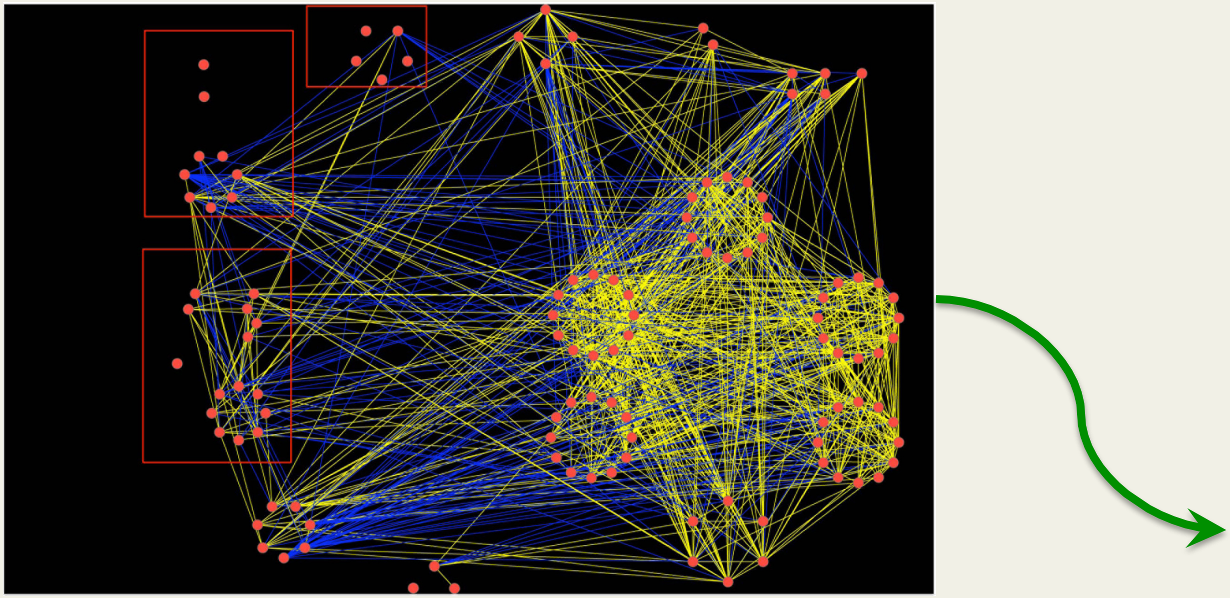
multiple measurements
for each cell

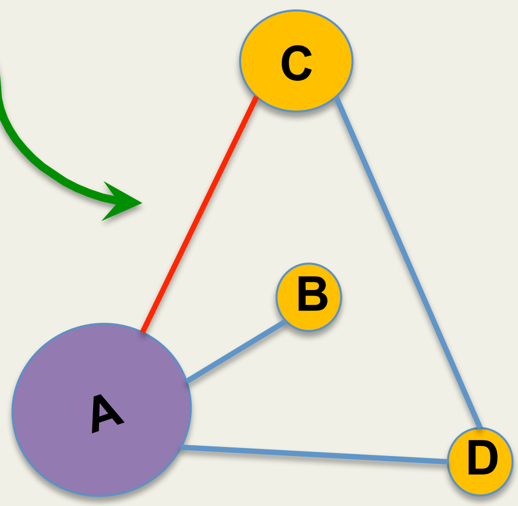
"Hairball"



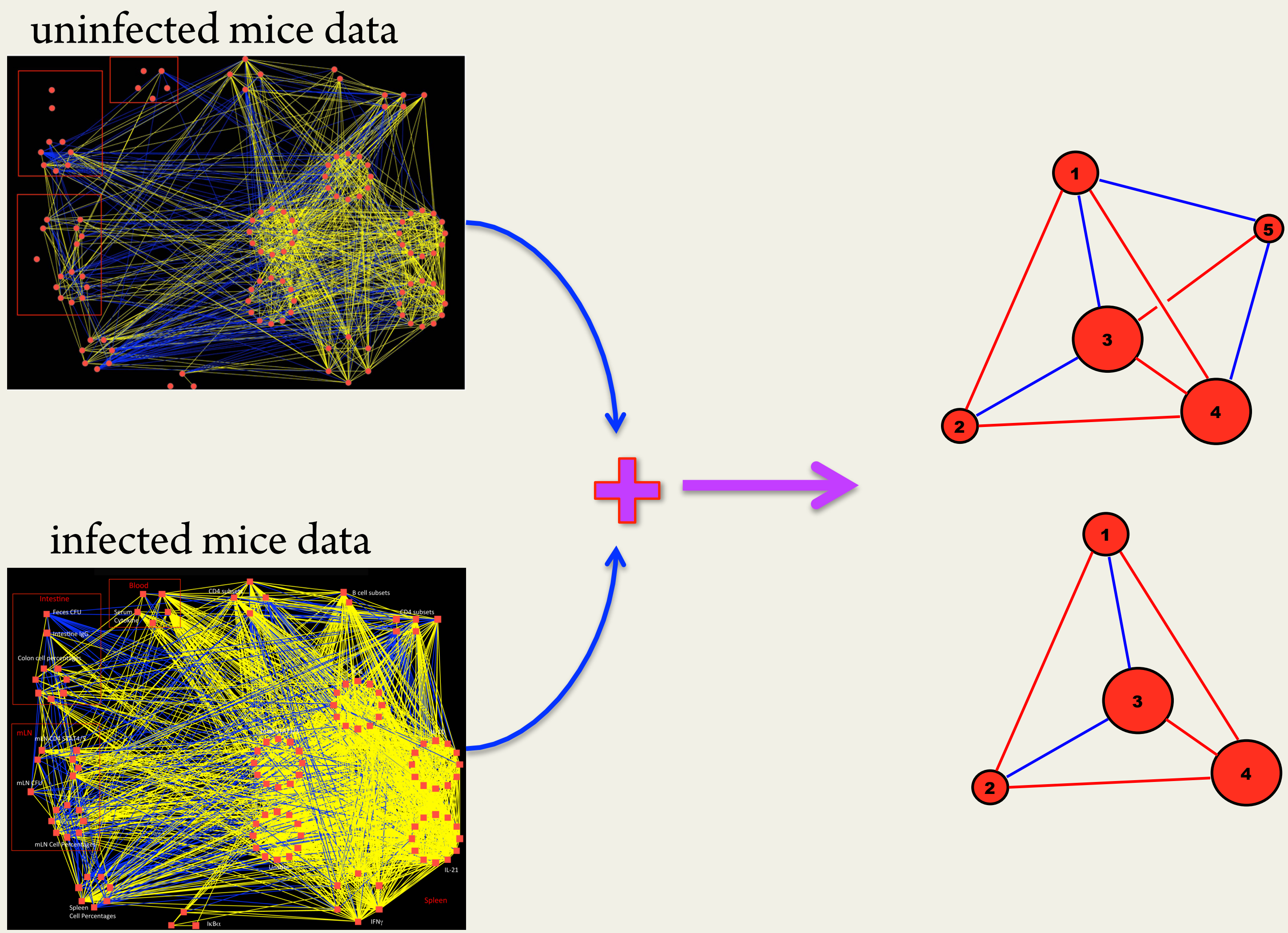
Nicolau, Hotson, Gopinath

Smoothing the "Hairball"





Nicolau, Hotson, Gopinath


uninfected mice data
infected mice data
1
2
3
4
5
1
2
3
4

A little bit of history…

The Human Genome Project: *1990 – 2003* **DOE & NIH**

international effort to discover all the estimated **20,000-25,000** human genes.

determine the complete sequence of the **3 billion** *DNA* subunits

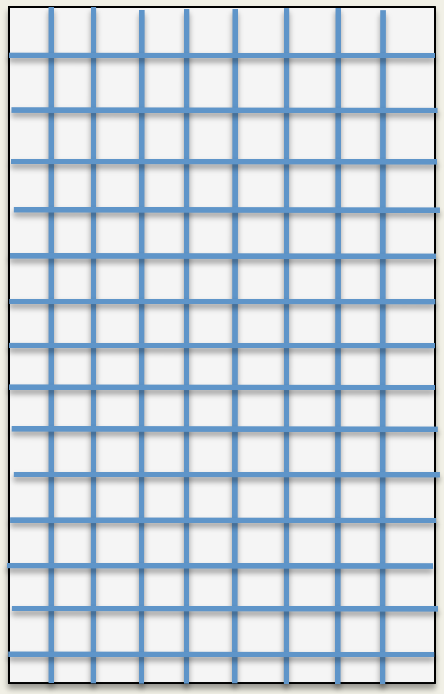
What do these genes do?

gene expression microarrays – 1995 (Science) & 1996 (Nature Biotechnology)

High throughput data

Data matrix:

several (100 – 500)
columns
TISSUE SAMPLES

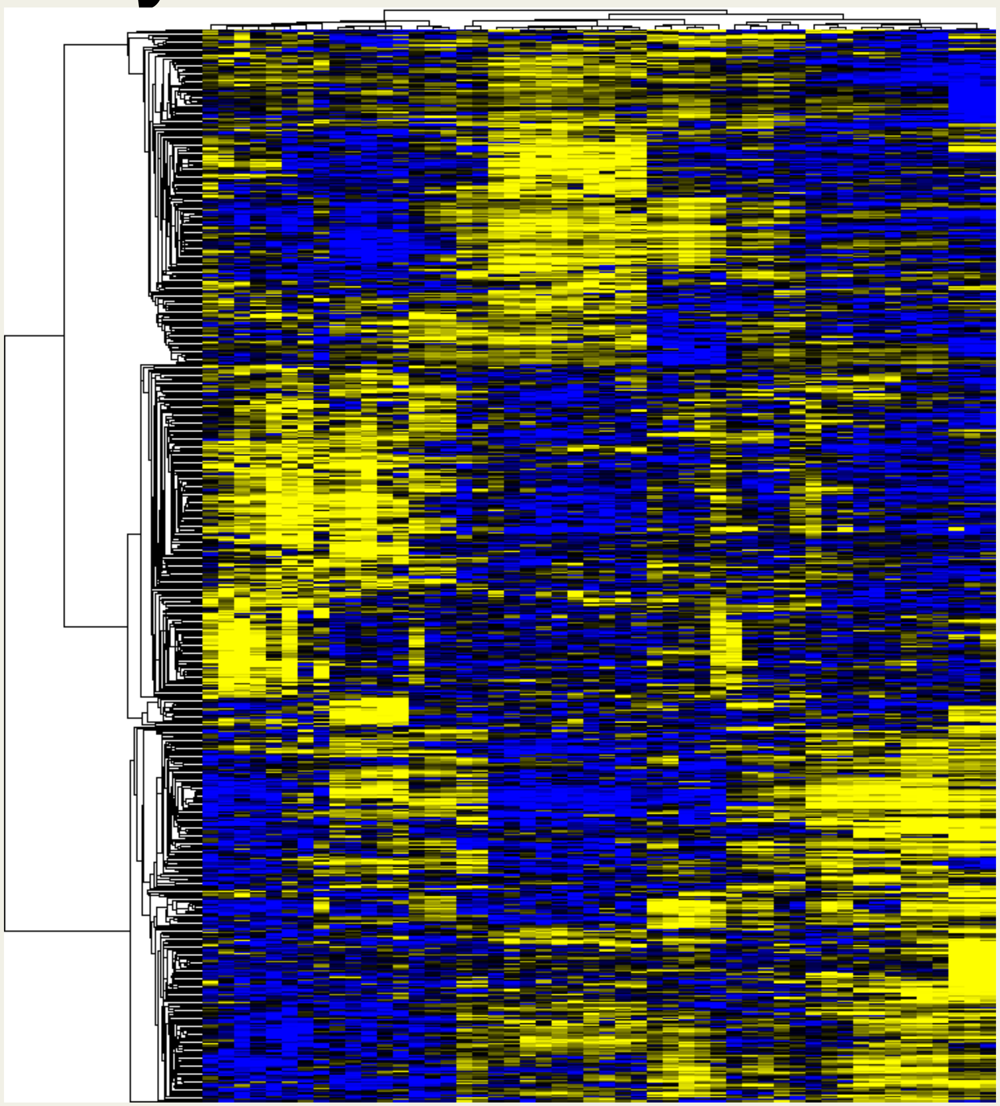


20,000 – 40,000
rows
GENES

function genes
distinctions diseases

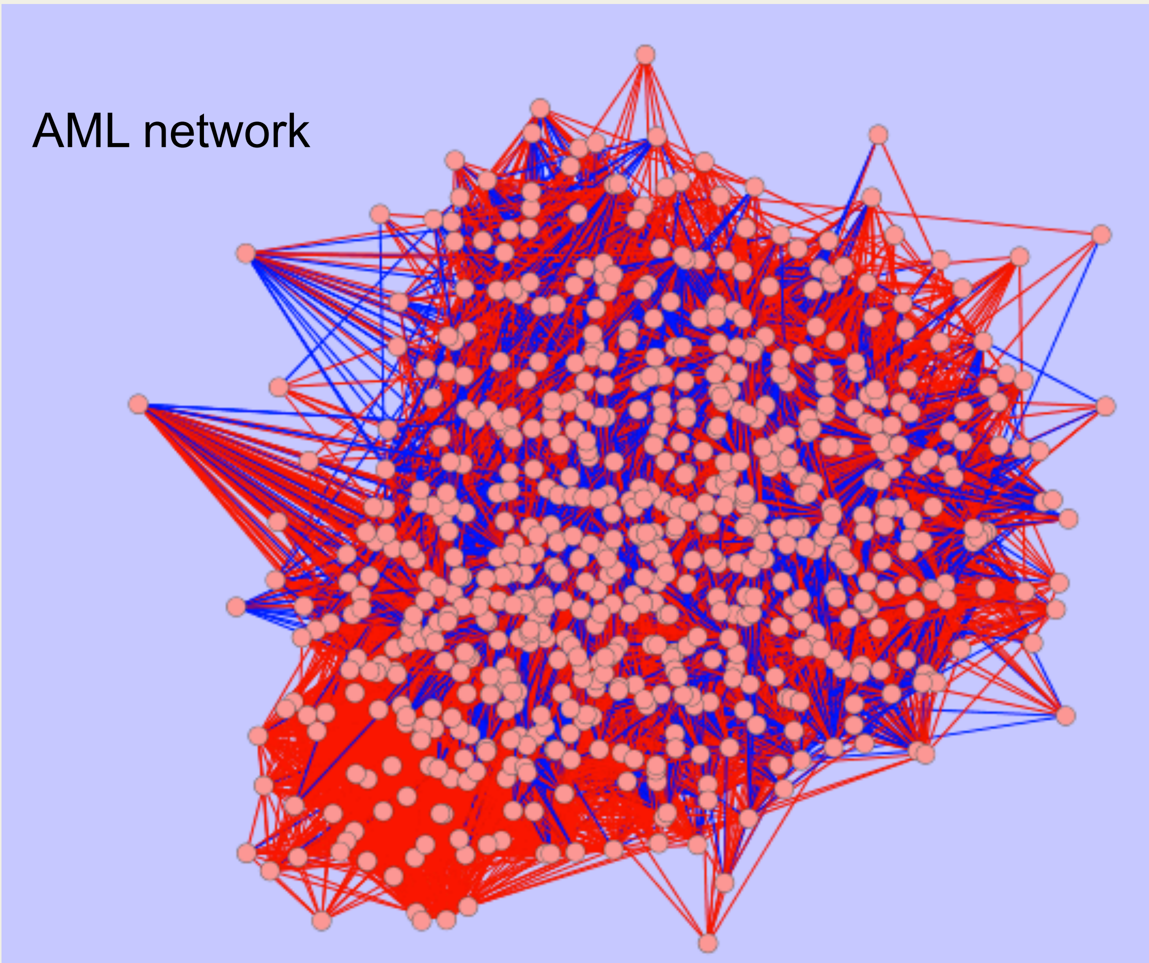
Gene expression microarrays

Acute Myeloid Leukemia (AML)



AML gene expression heatmap

Networks



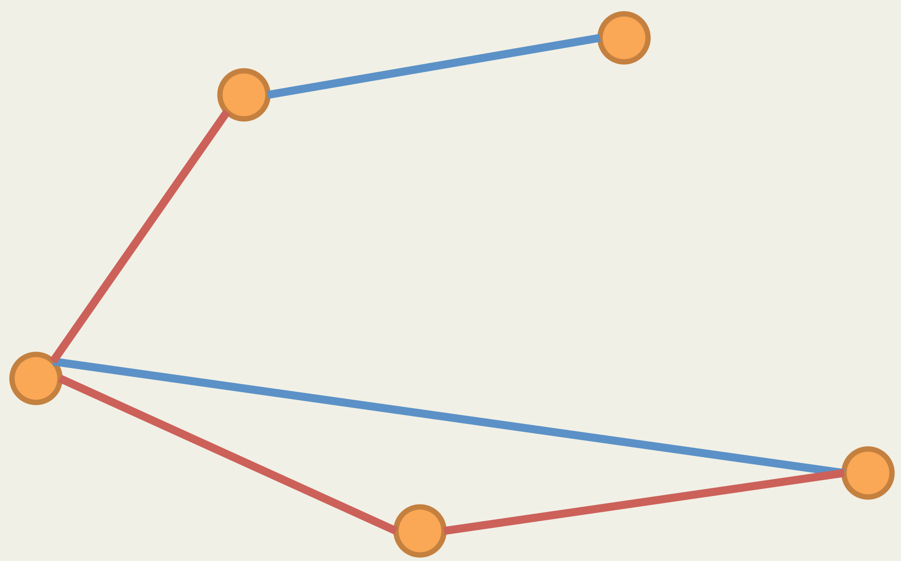
Legend:

node = gene

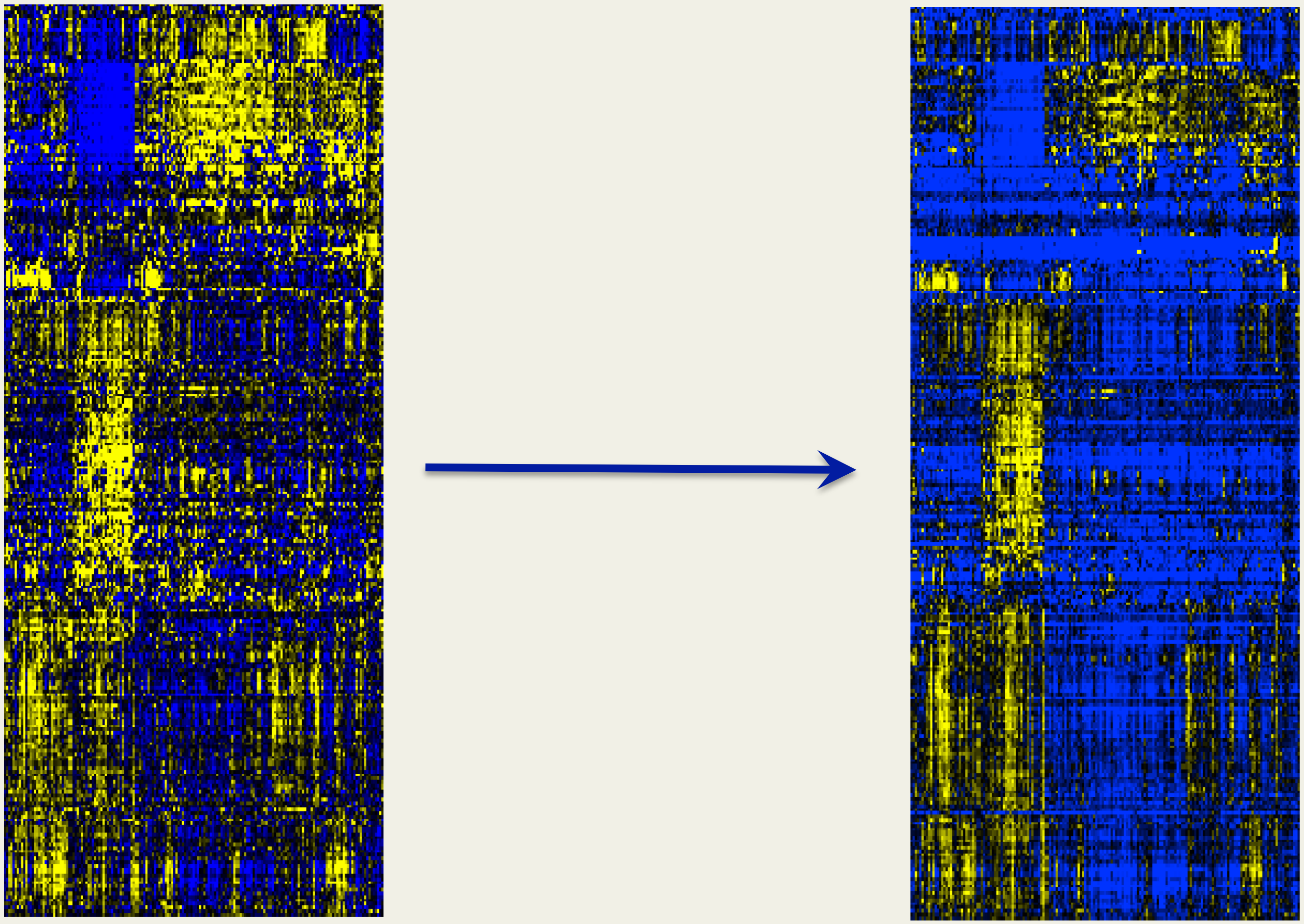
edge = positive correlation

edge = negative correlation

Simplified AML Hairball



Data smoothing: local smoothing of large data



DATA   FLAT DATA

Nicolau et al – Bioinformatics 2007

Data smoothing: local smoothing of large data

1. FLAT construction – data de-sparcing

$$N_1, N_2, \ \ldots N_k \longrightarrow \hat{N}_1, \hat{N}_2, \ \ldots \hat{N}_k$$

$\hat{N}_i$ fit to linear model in $N_1, N_2, \ldots, \ N_{i-1}, N_{i+1}, \ldots N_k$

Analysis of high throughput data

RELEVANCE

geometric transformations
hypothesis
definition & testing

SHAPE OF DATA

applied topology and
persistence - robustness

High throughput data & relevance

- What in the data is relevant to **my study**?

RELEVANCE

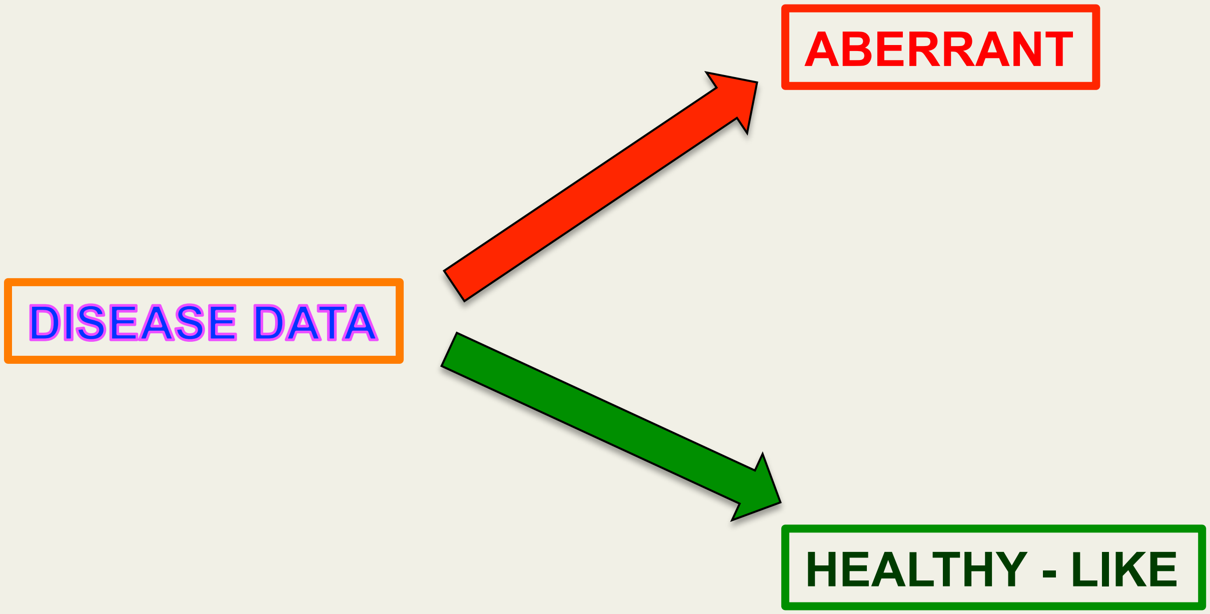
Understand disease processes from data

what is relevant to the disease?

transform data to emphasize aberrant patterns compared to healthy tissue data

Disease specific genomic analysis (*DSGA)*
Nicolau M, Tibshirani R, Børresen-Dale AL, Jeffrey SS Bioinformatics 2007

RELEVANCE –
Disease specific genomic analysis - DSGA



DSGA – Nicolau et al – Bioinformatics 2007

WHAT DOES DISEASE LOOK LIKE?

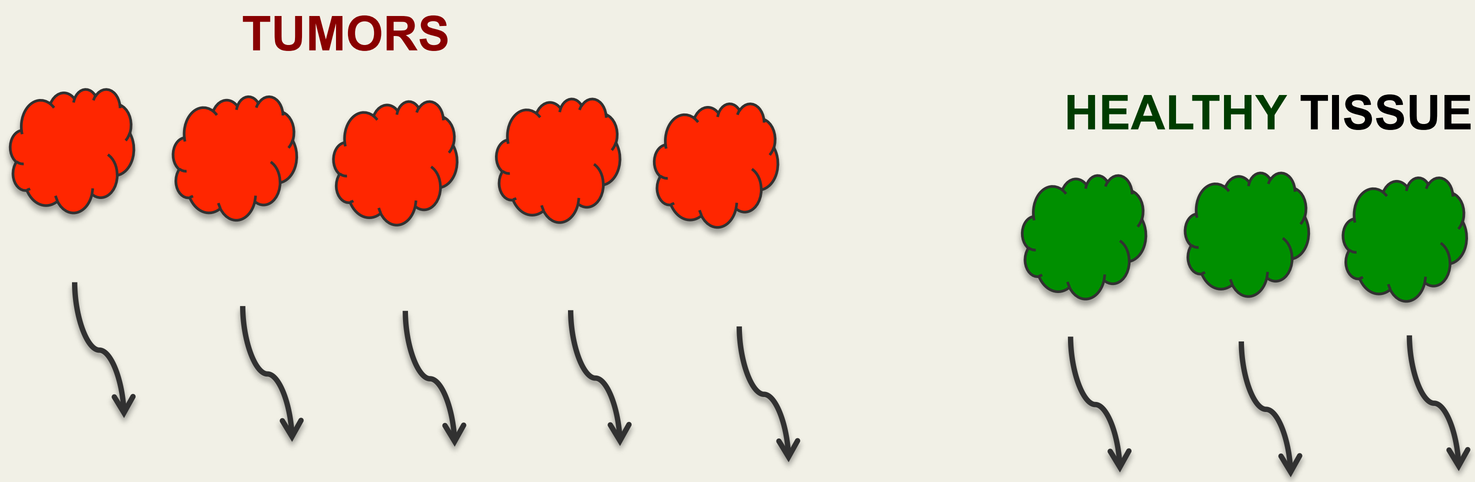
DSGA – Nicolau et al – Bioinformatics 2007

WHAT DOES DISEASE LOOK LIKE?

cancer cells retain memory of their (healthy) cell type signature

DSGA – Nicolau et al – Bioinformatics 2007

RELEVANCE –
Disease specific genomic analysis - DSGA

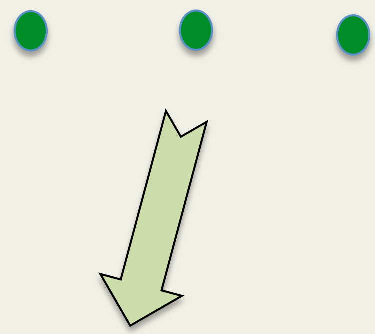


microarrays

DSGA – Nicolau et al – Bioinformatics 2007

RELEVANCE –
Disease specific genomic analysis - DSGA

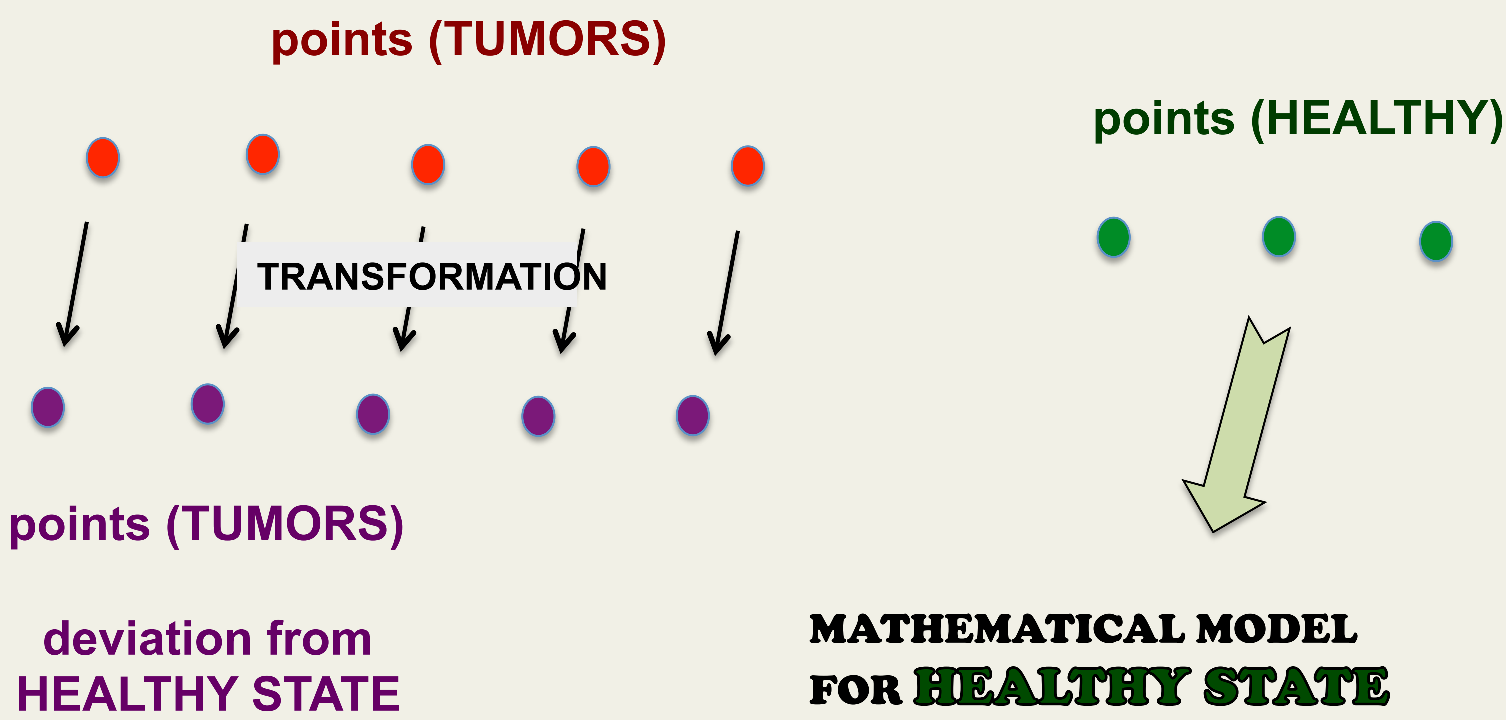
points (HEALTHY)



MATHEMATICAL MODEL
FOR HEALTHY STATE

DSGA – Nicolau et al – Bioinformatics 2007

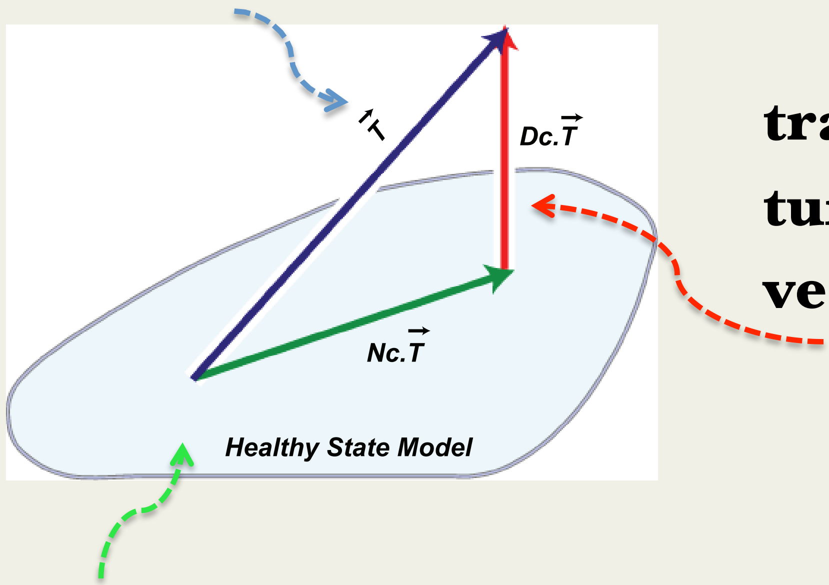
RELEVANCE –
Disease specific genomic analysis - DSGA



DSGA – Nicolau et al – Bioinformatics 2007

RELEVANCE –
Disease specific genomic analysis - DSGA

Tumor data



transformed
tumor data
vector of residuals

[Null Hypothesis Space]

⇧

Normal tissue data

DSGA – Nicolau et al – Bioinformatics 2007

RELEVANCE – *Disease specific genomic analysis - DSGA*

Benefits from *Disease component* of tumor data:

1. Highlight **extent of deviation** from normal – aberrant behavior

2. **Cleaner** identification of distinct classes

3. **Biology** you highlight is **different** from using original data.

DSGA – Nicolau et al – Bioinformatics 2007

Analysis of high throughput data

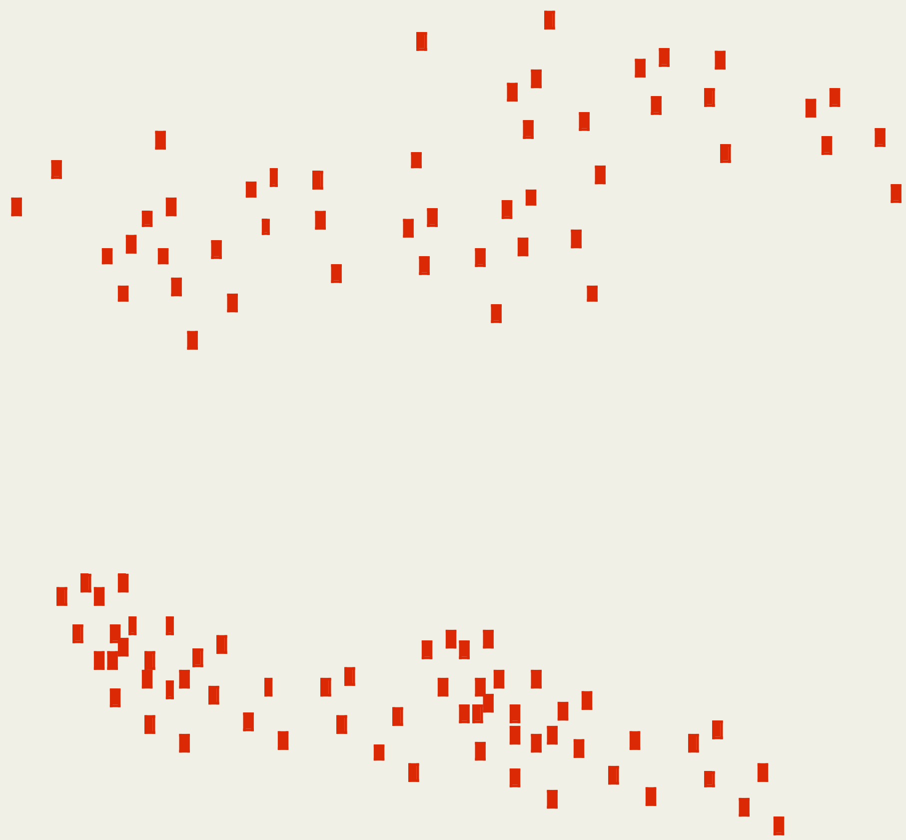
RELEVANCE

geometric transformations
hypothesis
definition & testing

SHAPE OF DATA

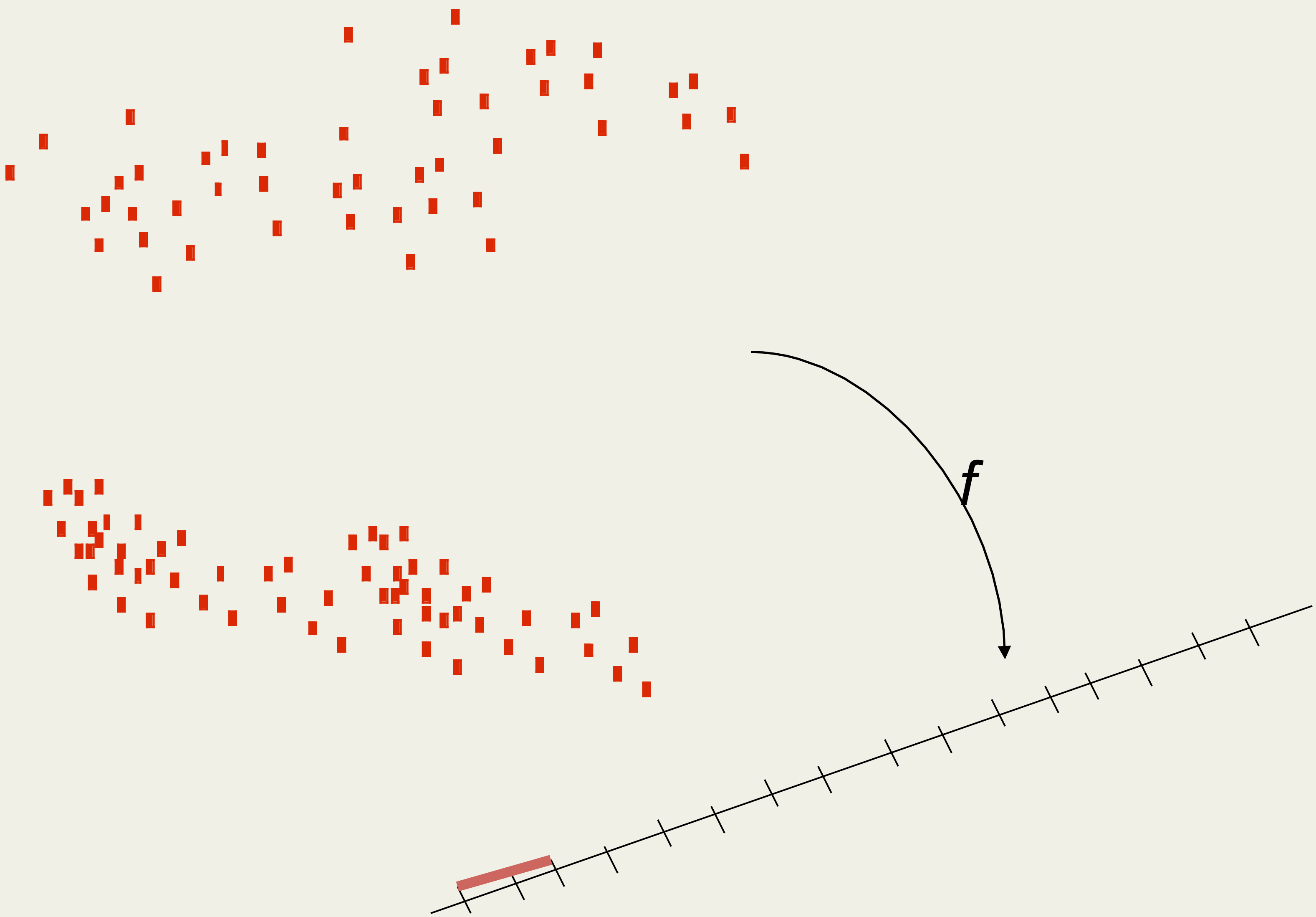
applied topology and
persistence - robustness

SHAPE OF DATA
TOPOLOGY & Mapper

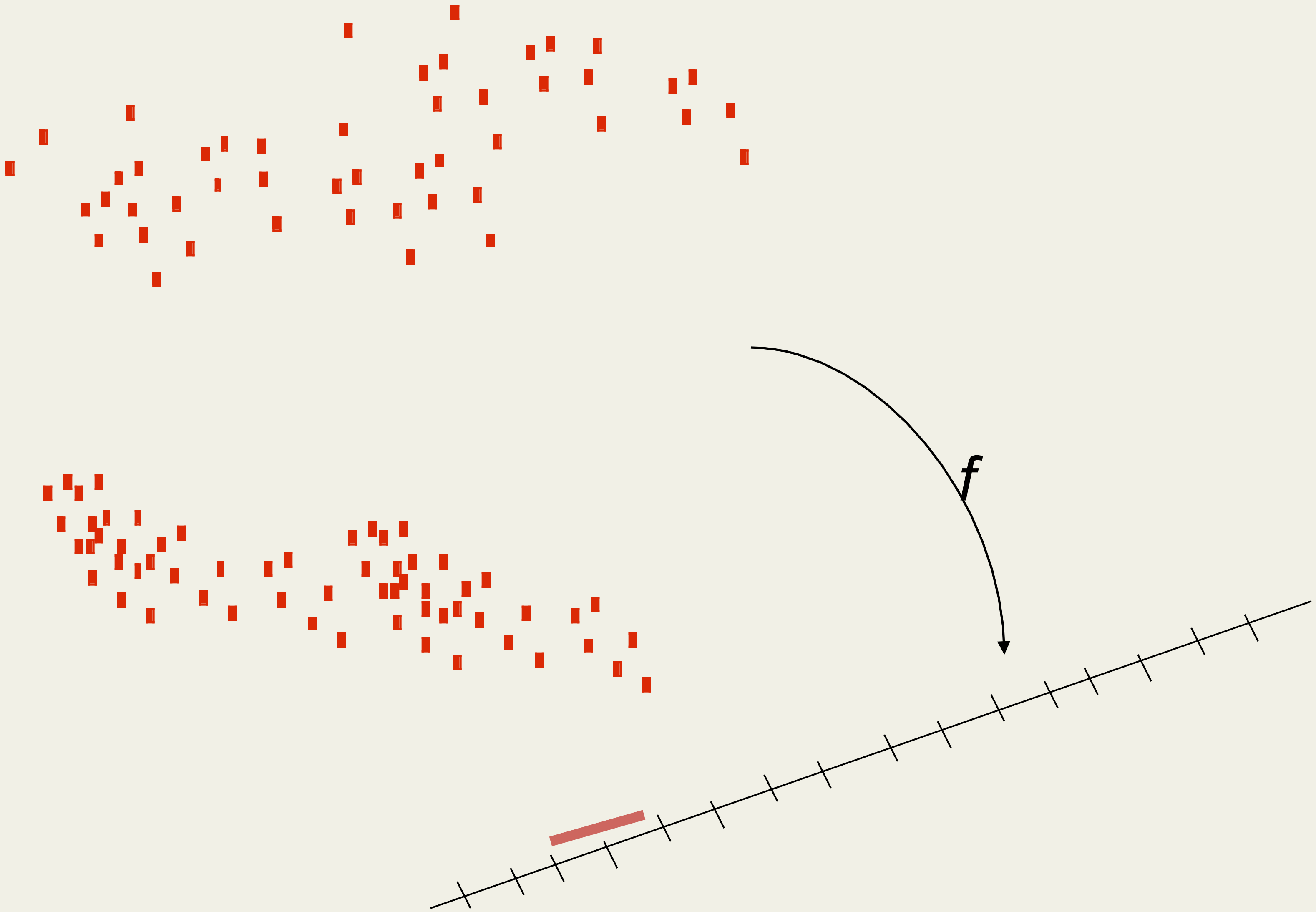


Singh, Memoli, Carlsson ***Point Based Graphics*** 2007

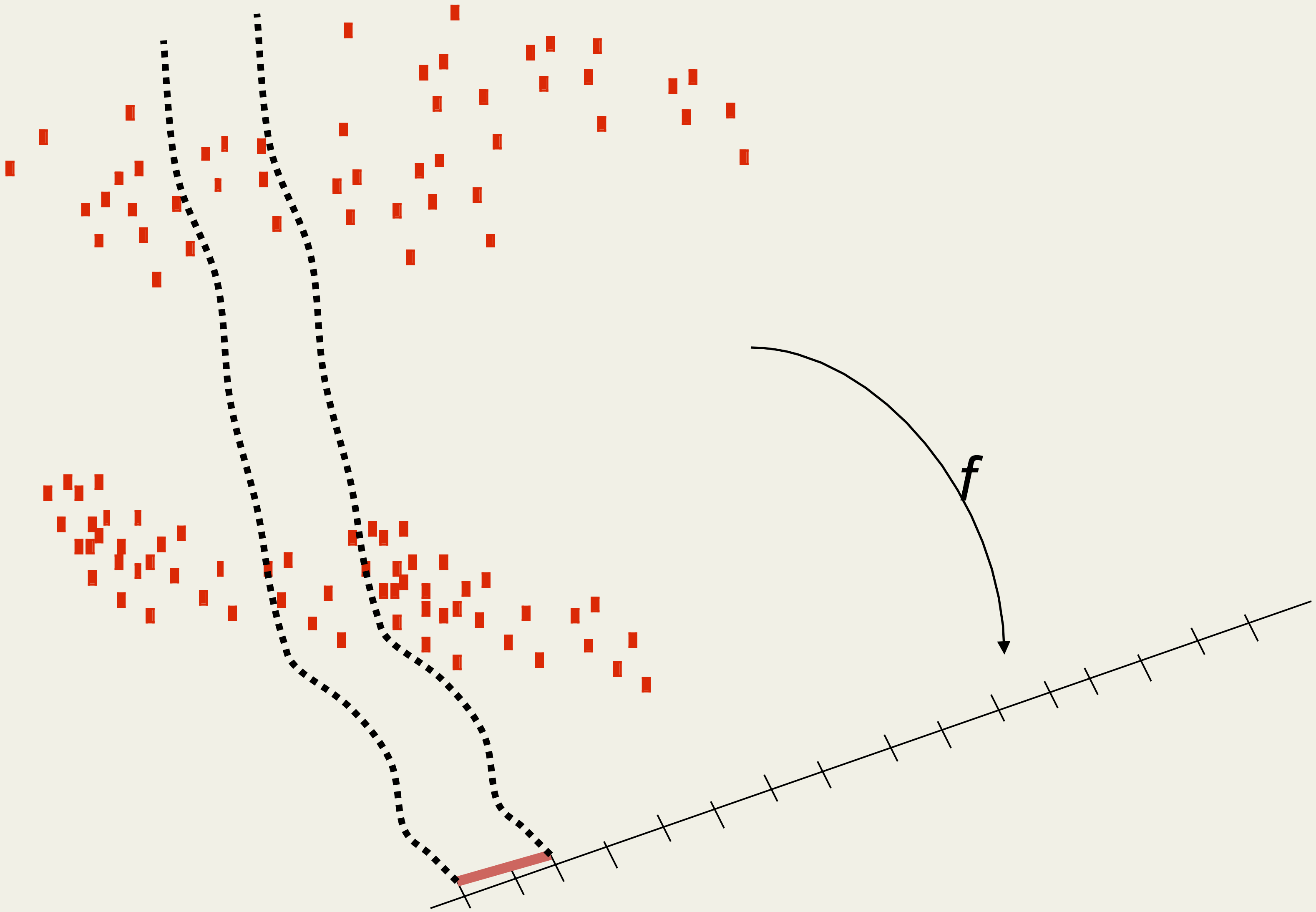
SHAPE OF DATA - mapper



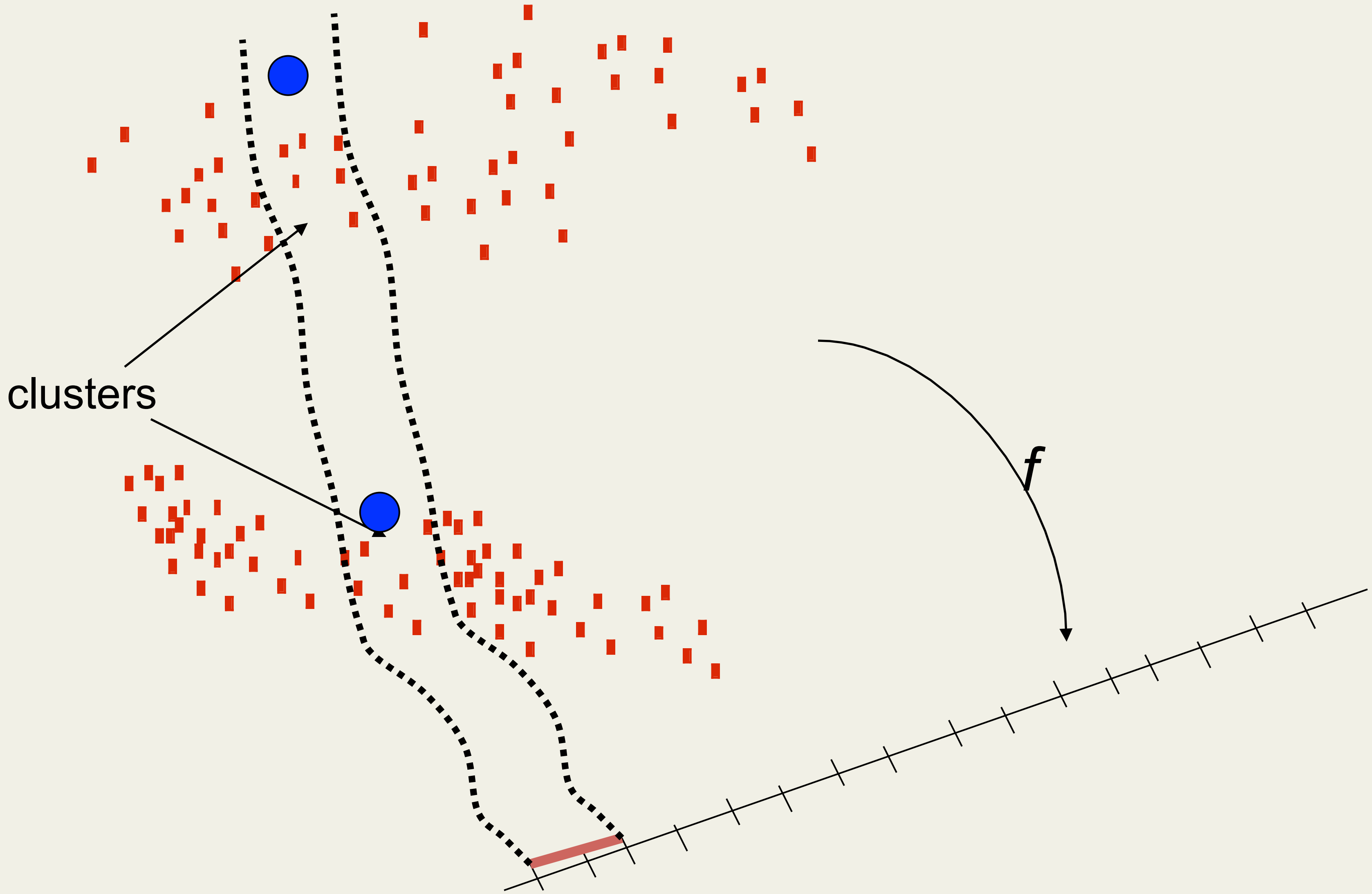
Mapper



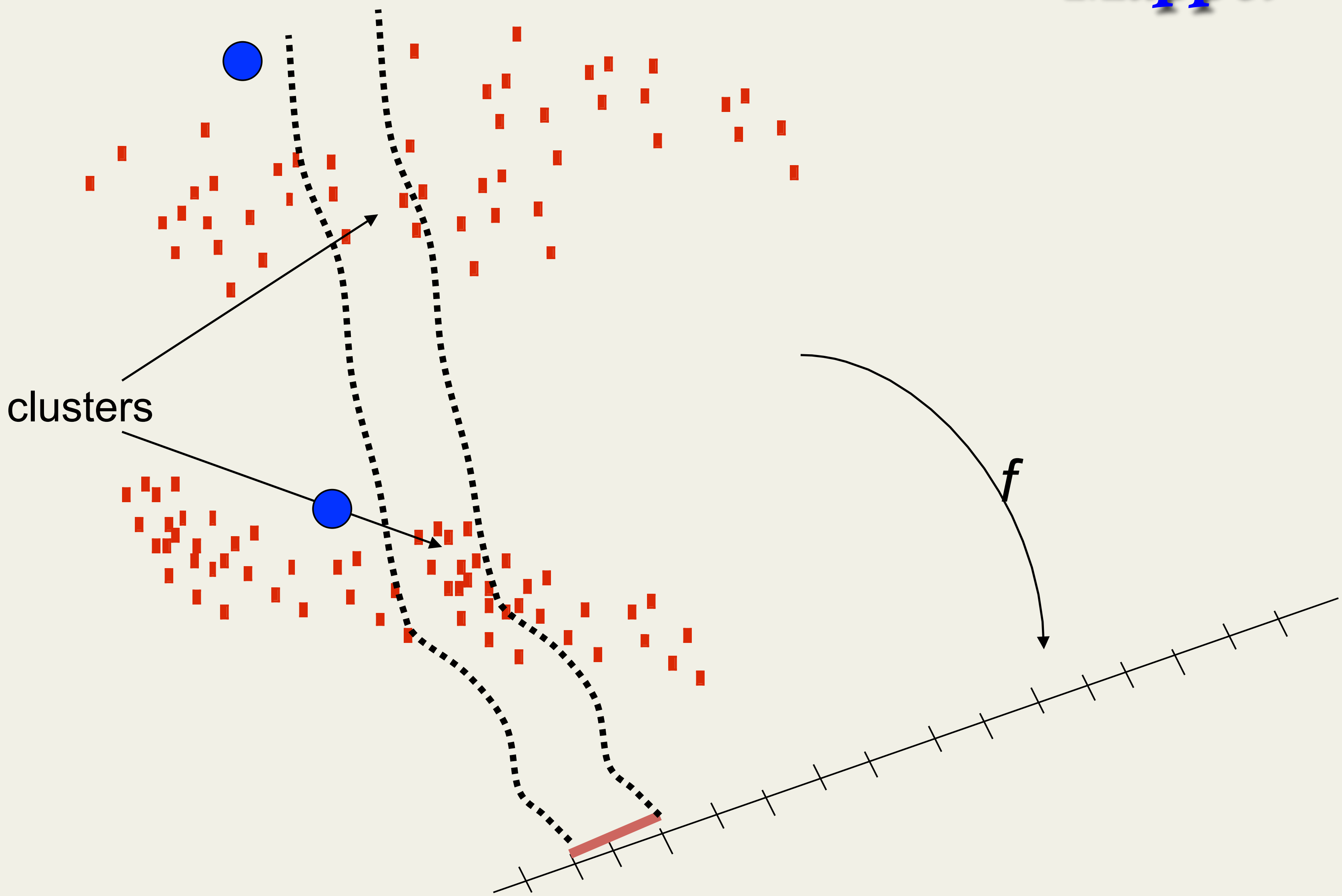
Mapper



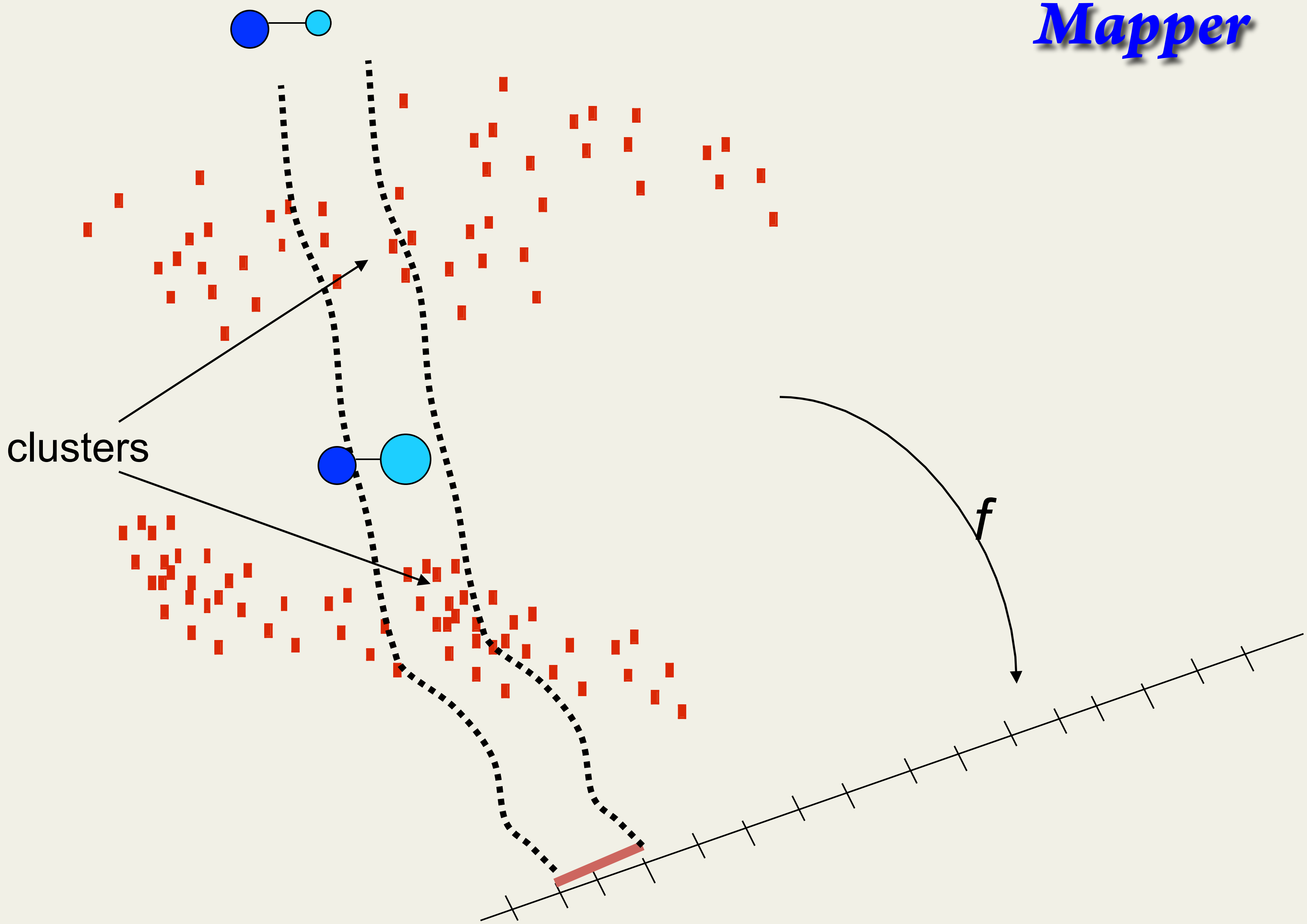
Mapper



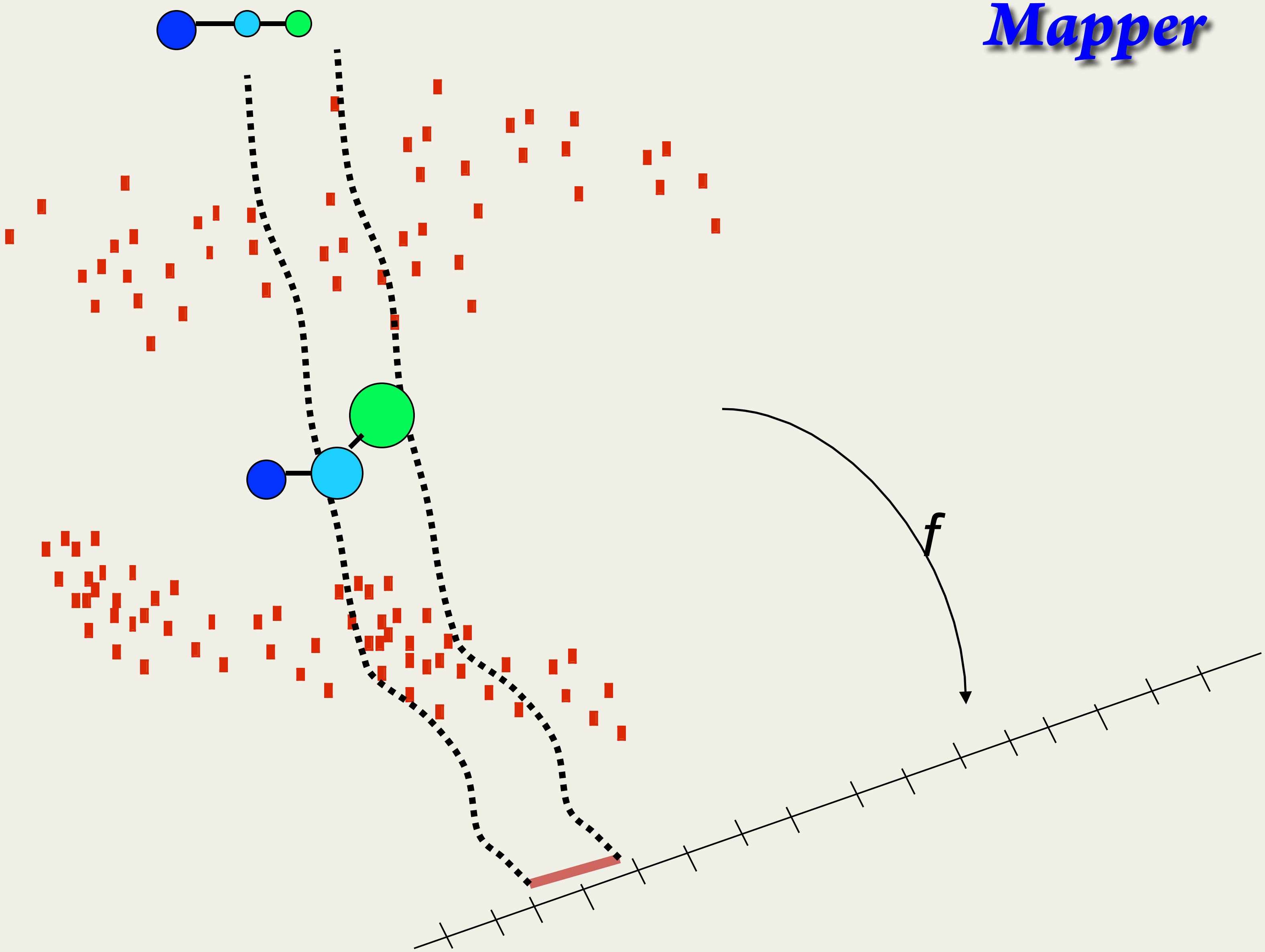
Mapper



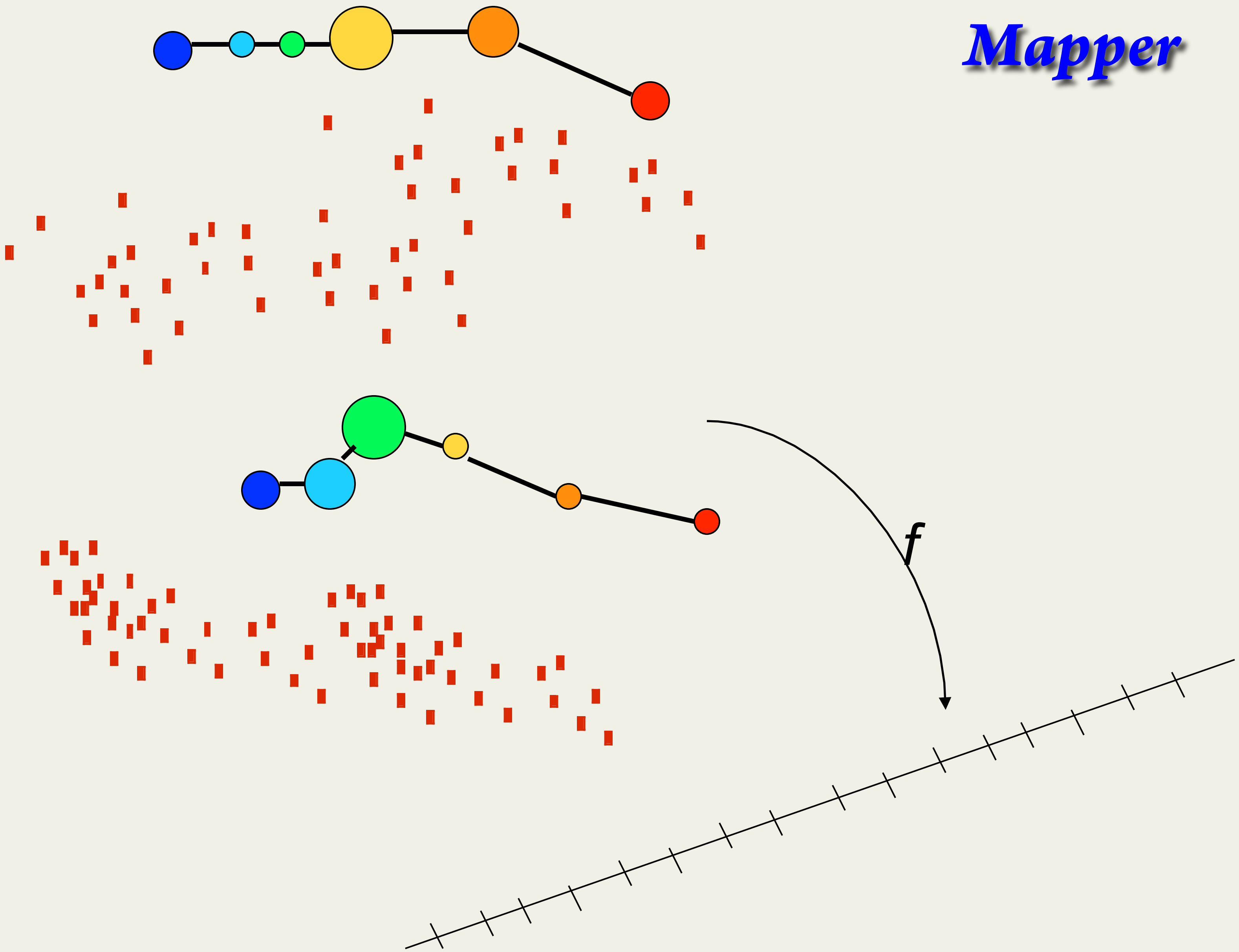
Mapper



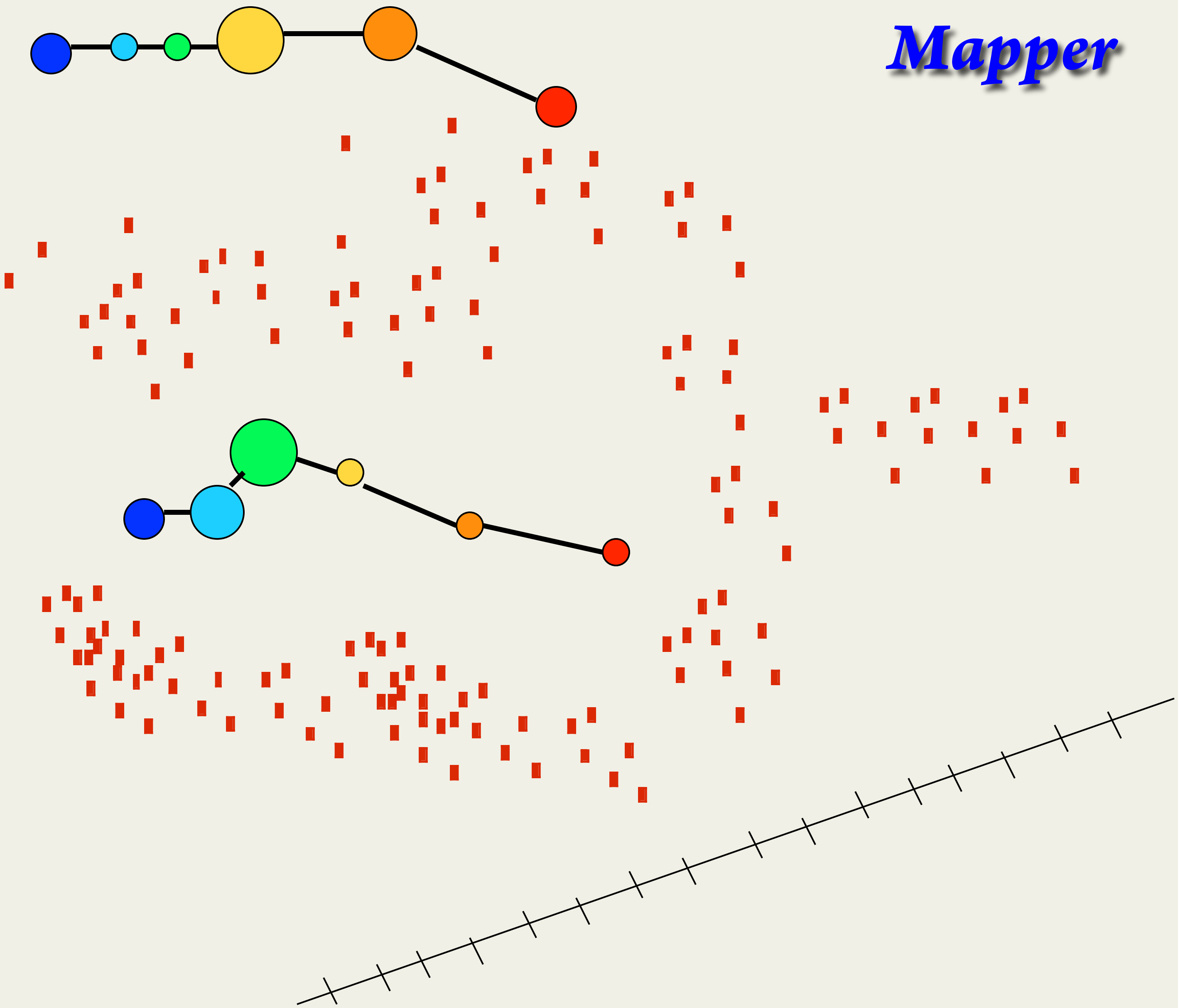
Mapper



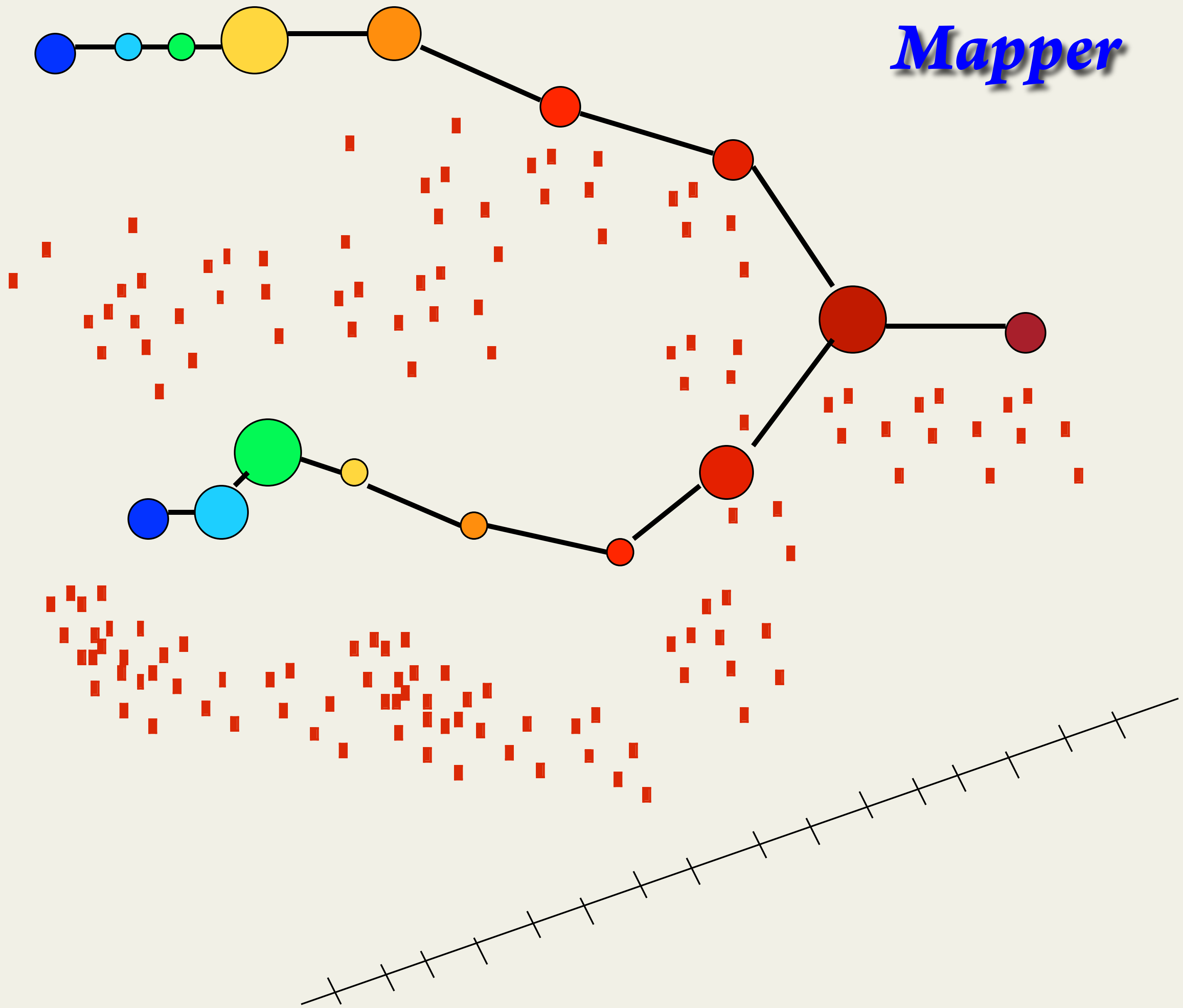
Mapper



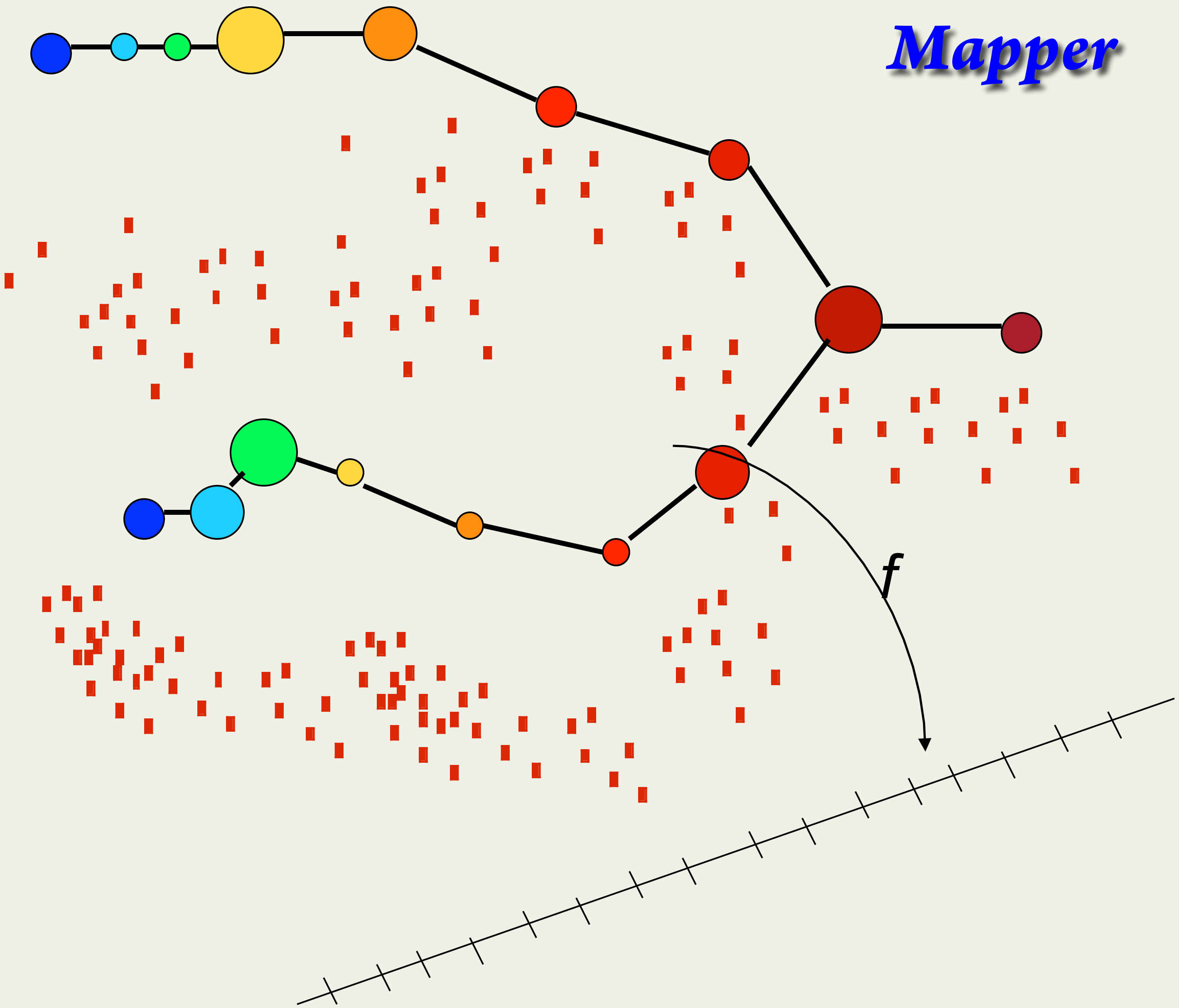
Mapper



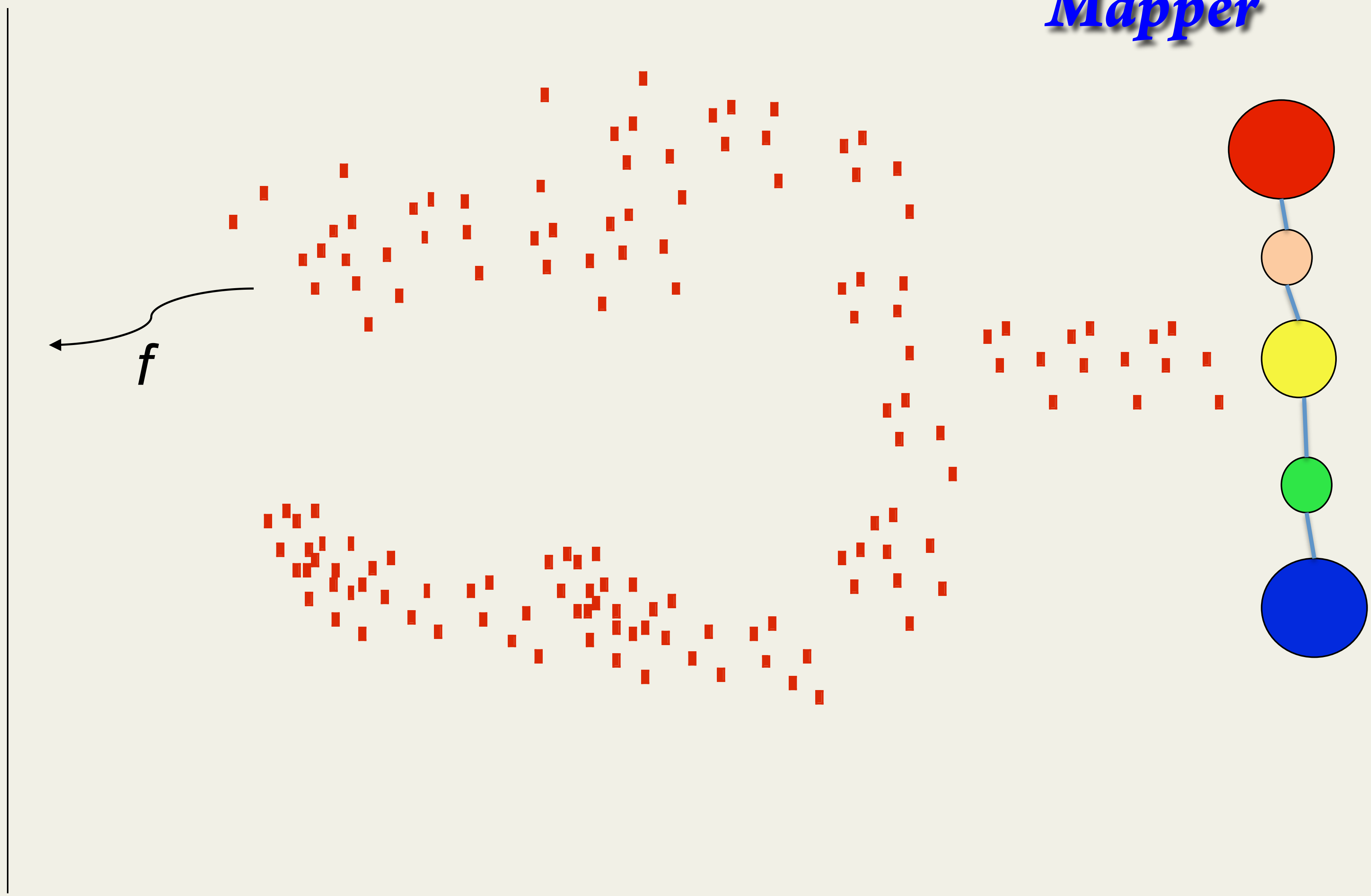
Mapper



Mapper

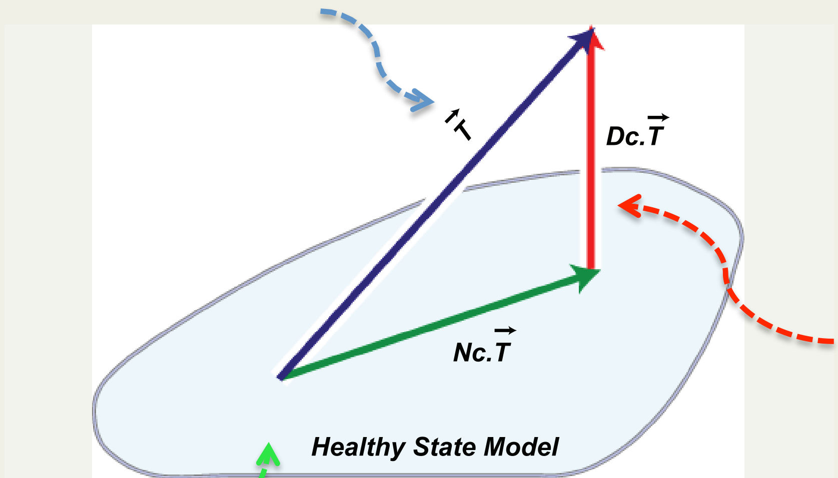


Mapper



Mapper filter function: overall deviation from Healthy State Model

Tumor data



transformed tumor data vector of residuals ***DcTumor***

vector magnitude of ***Disease Component***

[Null Hypothesis Space]

⇧

Normal tissue data

Progression Analysis of Disease – PAD

RELEVANCE

geometric transformations
DSGA

SHAPE OF DATA

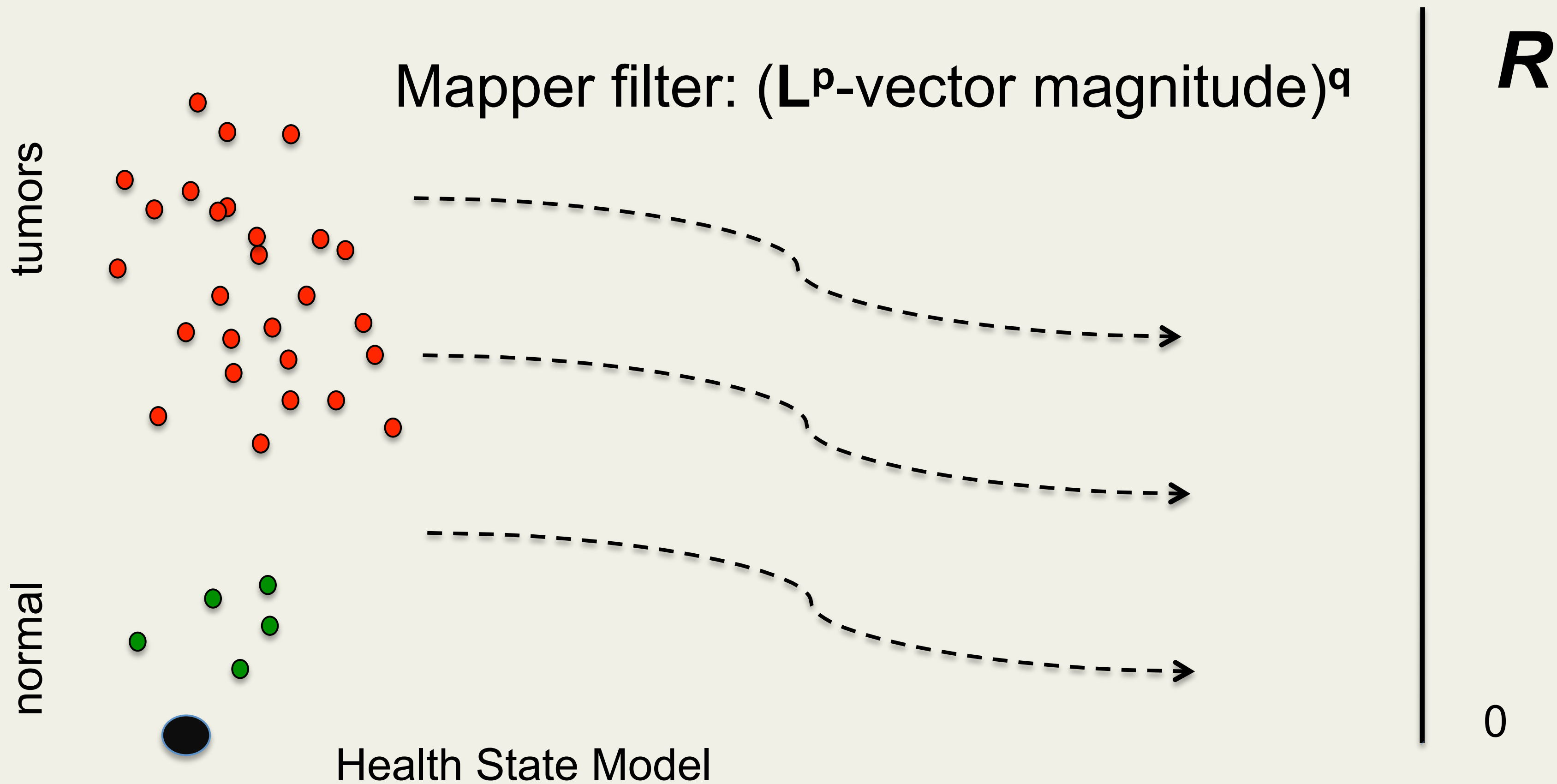
applied **topology**
Mapper

Topology based data analysis identifies subgroup of breast cancers with unique mutational profile and excellent survival
Nicolau M, Levine AJ, Carlsson G Proc Nat Acad Sci 2011

Progression Analysis of Disease: PAD

running Mapper on DSGA-transformed data

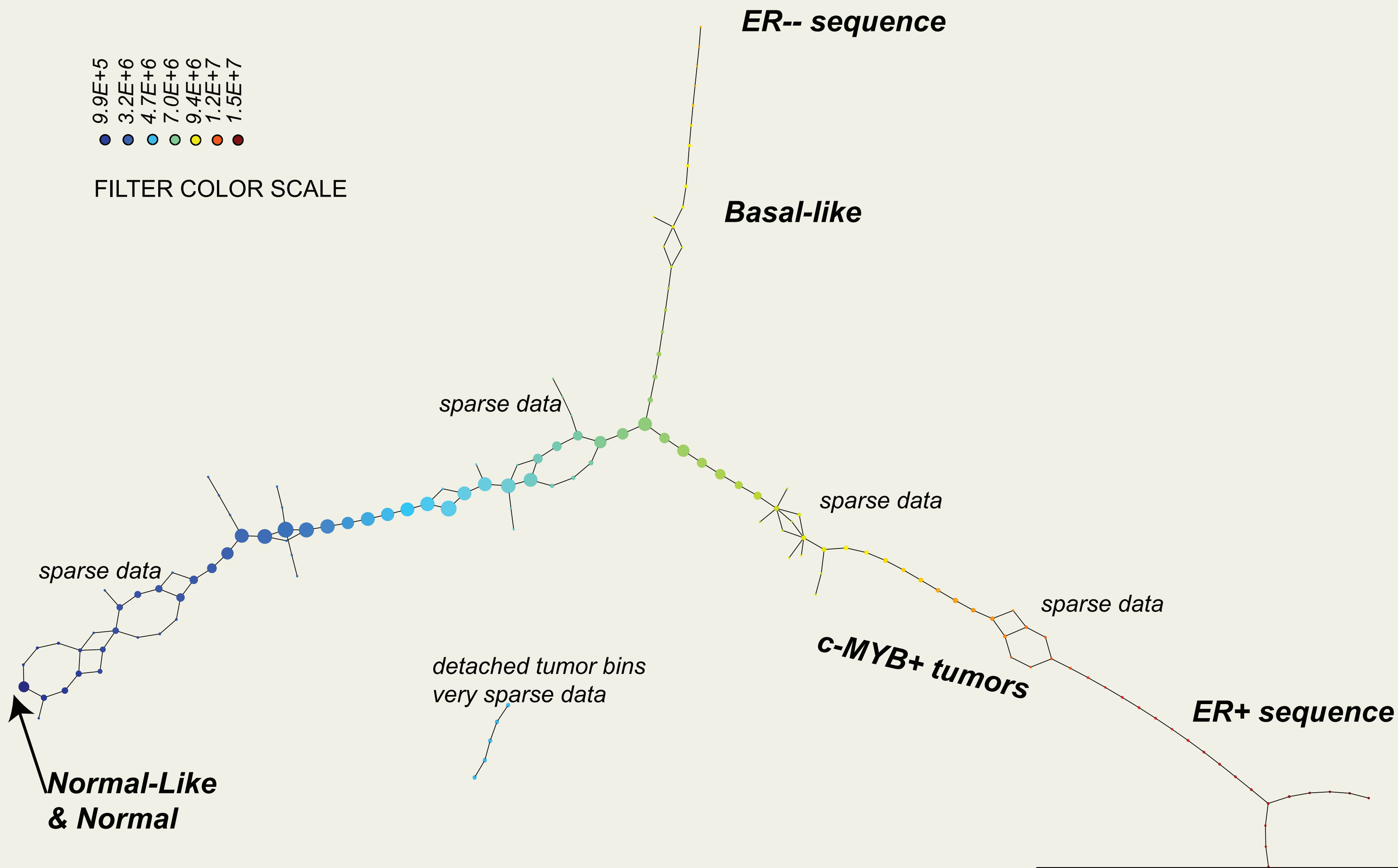
DSGA – transformed data from tumors & normals:
disease component



Nicolau et al PNAS 2011

Progression Analysis of Disease: PAD

running Mapper on DSGA-transformed data

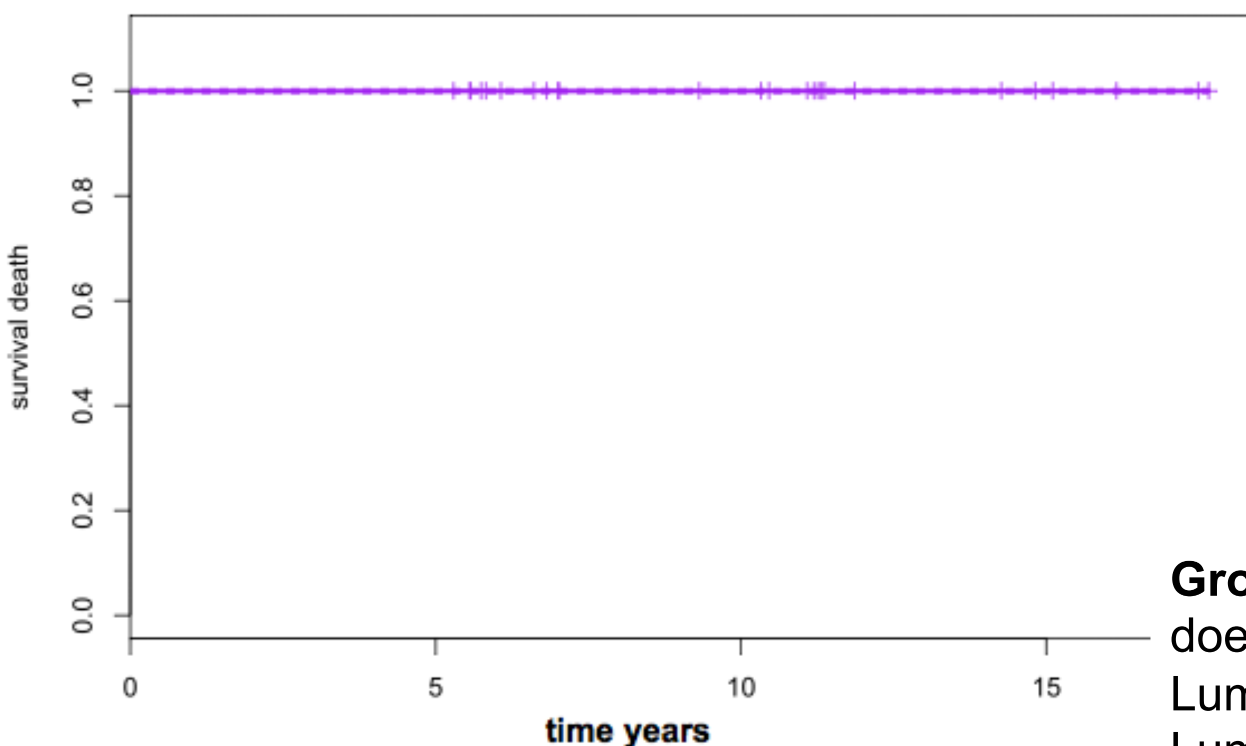


Nicolau *et al* PNAS 2011

c-MYB+ group

survival analysis

K-M survival
cMYB+group



Group is
homogeneous &
distinct
mathematically
biologically

predictor variables
few (1 or 2)

significant variables
biologically meaningfu

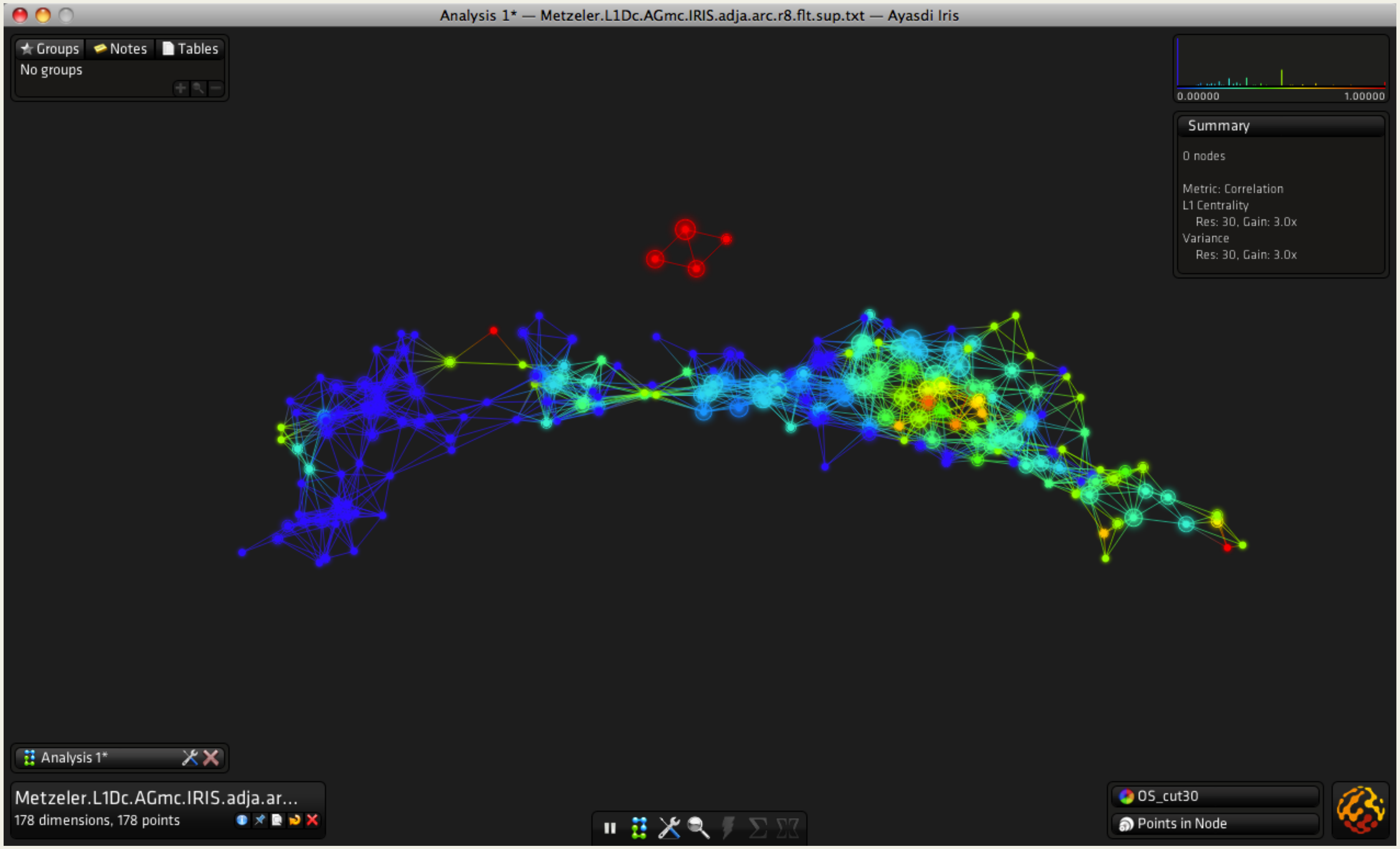
Group is new:
doesn't follow old classification
Luminal A
Luminal B
unclassified

Nicolau et al PNAS 2011

Acute Myeloid Leukemia

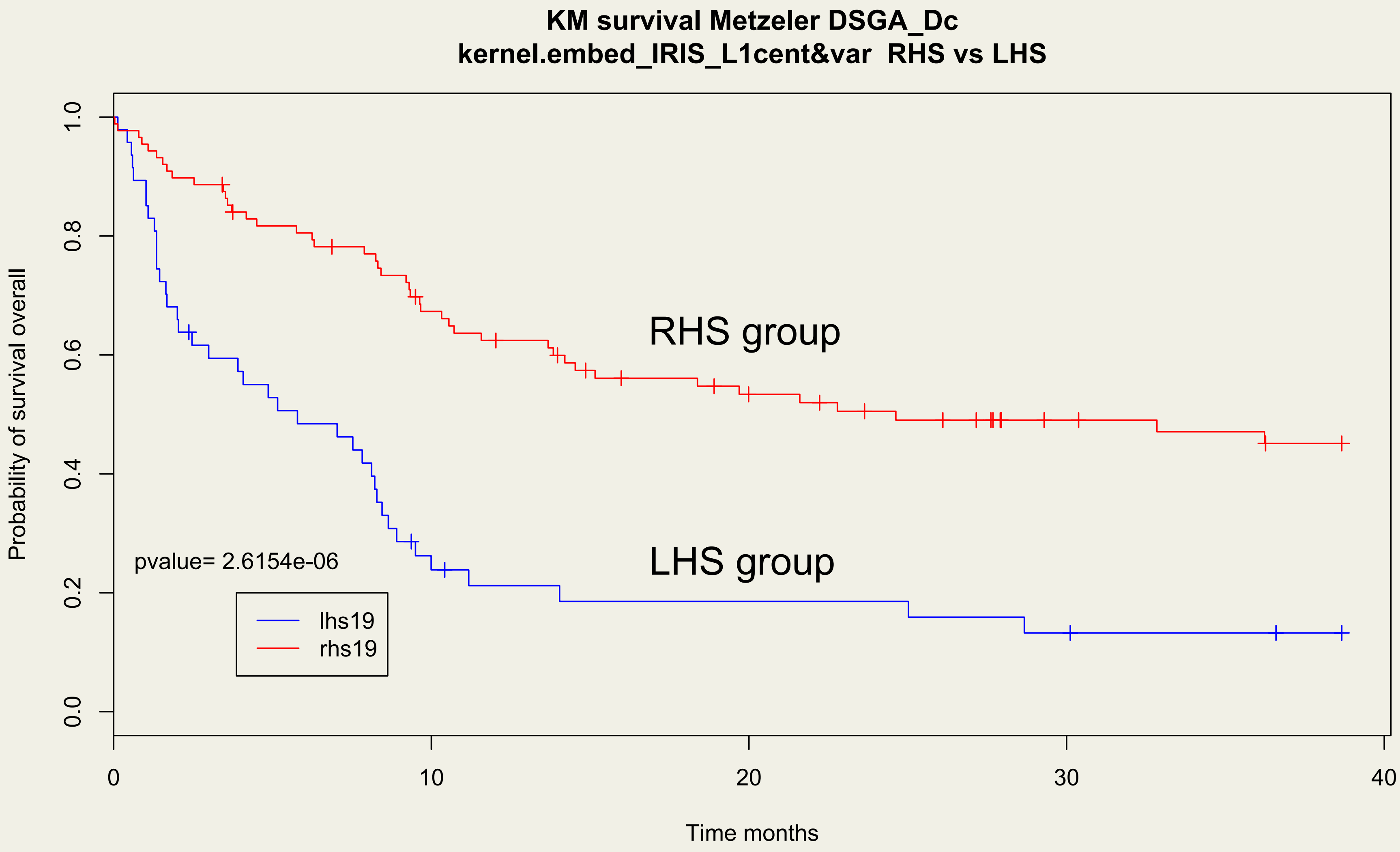
another example

Disease component IRIS: AML

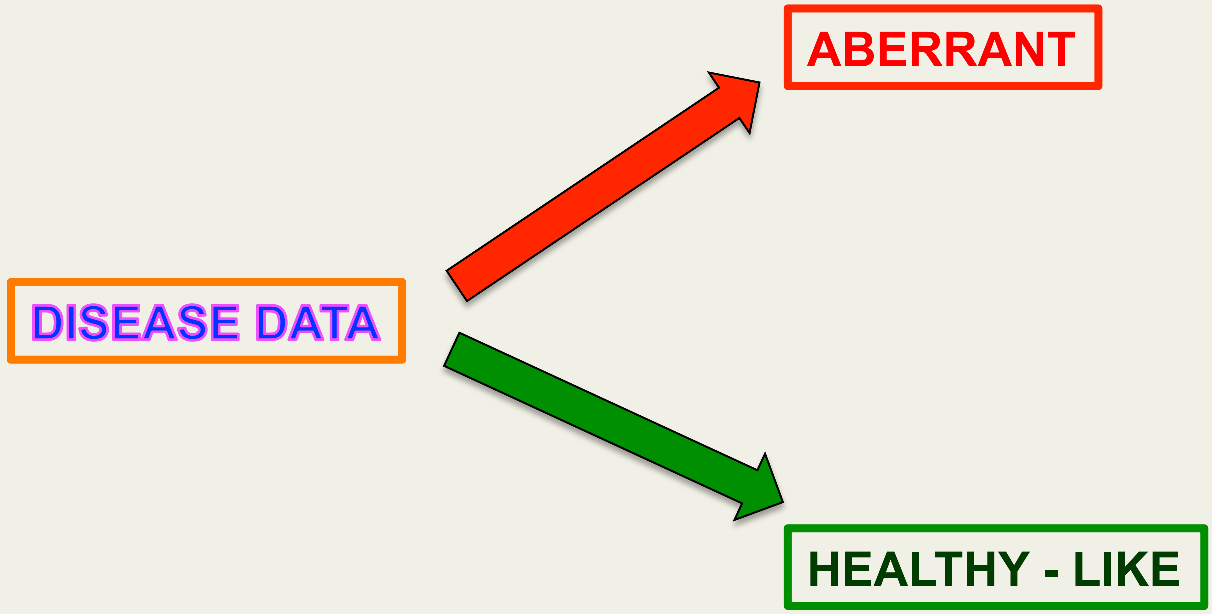


node color: survival

survival – RHS vs LHS

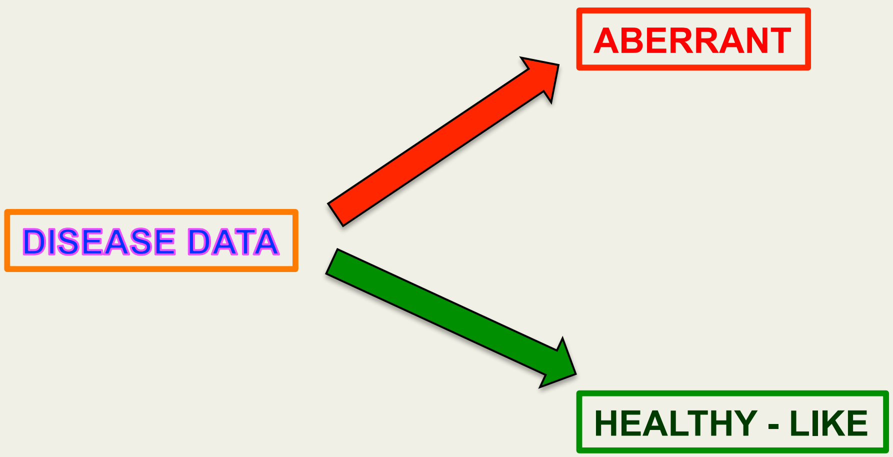


Another look at AML data



Disease Specific Genomic Analysis:
DSGA

Another look at AML data

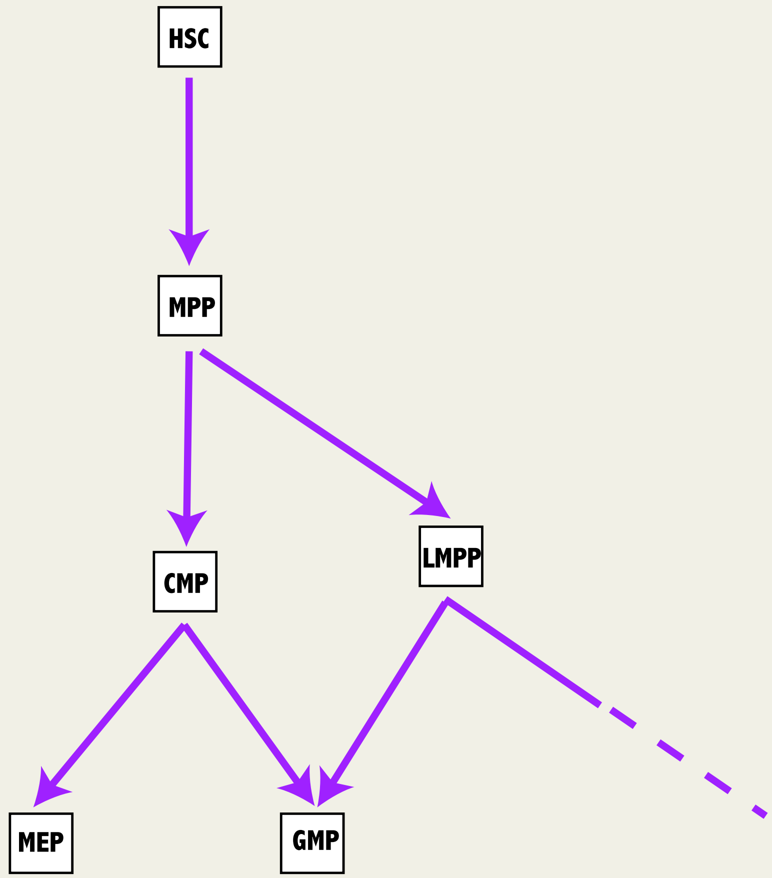


NORMAL COMPONENT OF AML CELLS

do AML tumor cells retain a memory of healthy signatures?

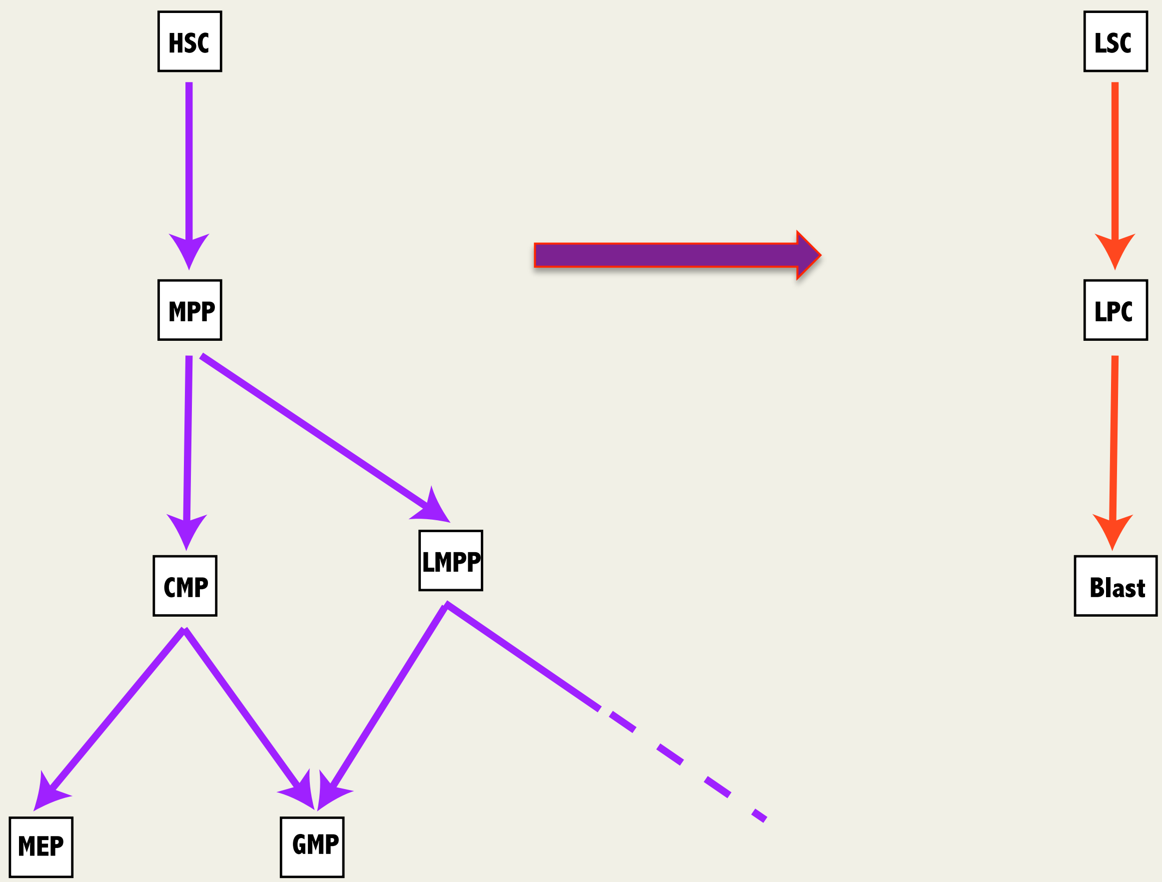
do differences in this memory have significance for disease?

Hematopoiesis



Hematopoiesis

AML developmental stages

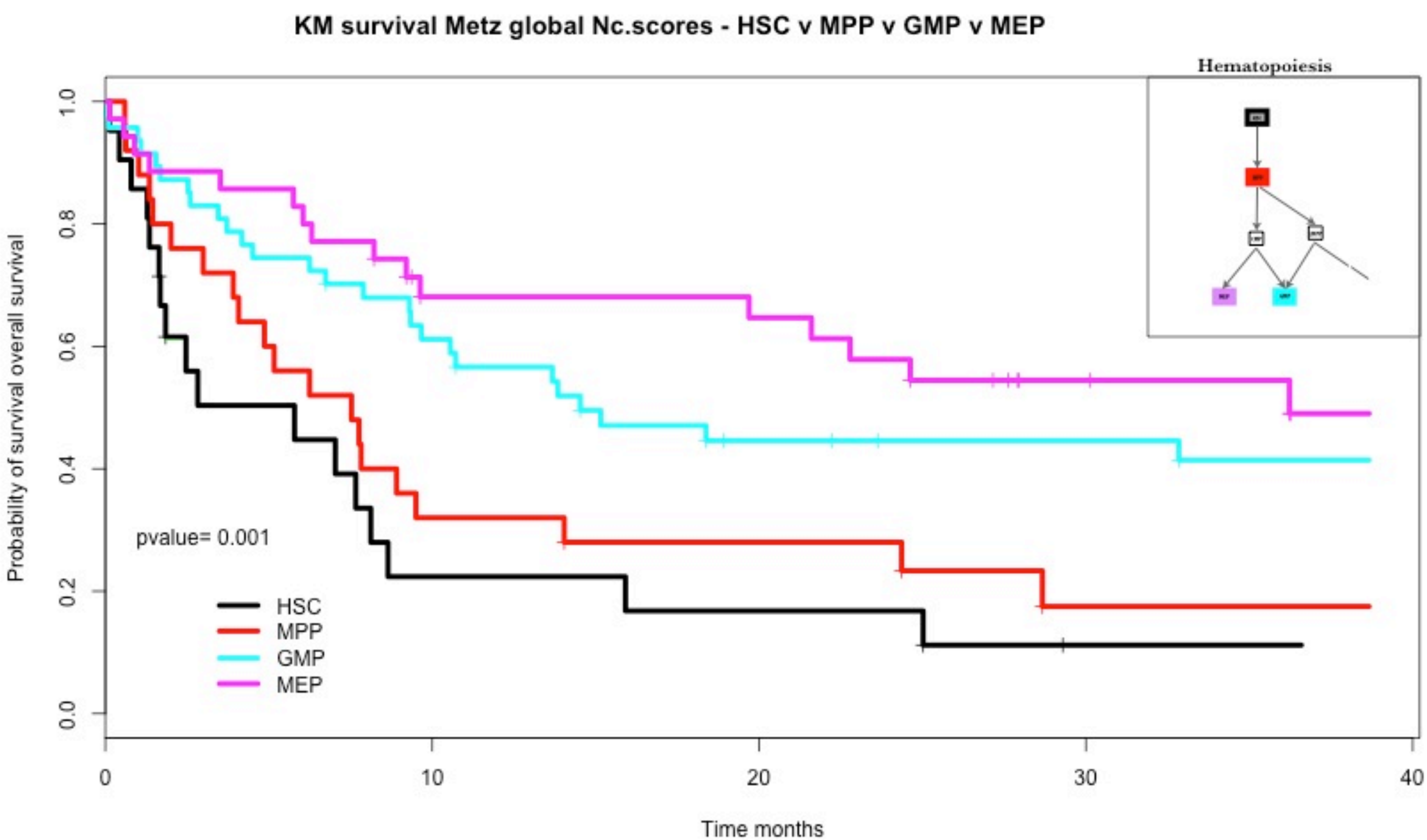


CD34+/CD38- *LSC*

CD34+/CD38+ *LPC*

CD34- *Blast*

Majeti markers


KM survival Metz global Nc.scores - HSC v MPP v GMP v MEP
Hematopoiesis
Probability of survival overall survival
1.0
0.8
0.6
0.4
0.2
0.0
pvalue= 0.001
HSC
MPP
GMP
MEP
0
10
20
30
40
Time months

conclusion

Disease component highlights aberrant behavior

association with clinical characteristics
cleaner groups of genes associated with distinct biology
together with Mapper found *novel group of* ***breast cancer***
& found strong *association with survival in* ***AML***

Normal component extracts memory of healthy types

association with clinical characteristics
identified *novel groups of* ***AML***

Networks/hairballs everywhere

Networks/hairballs everywhere

BREAST CANCER

overall deviation from healthy

DISEASE COMPONENT → **NETWORK**

AML

fit data to healthy template

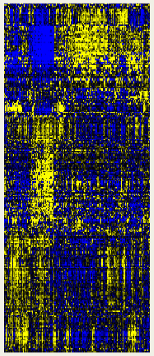
NORMAL COMPONENT → ***(directed)*** **NETWORK**

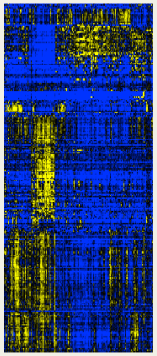
Large data

LOCALLY RICH

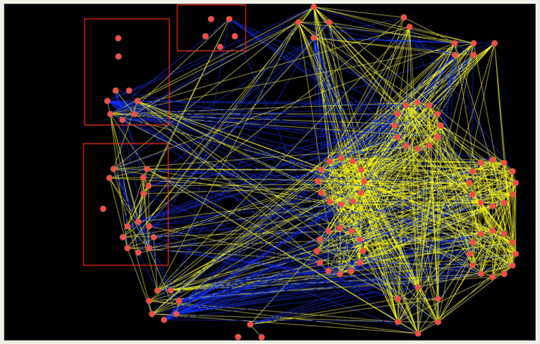
GLOBALLY MESSY

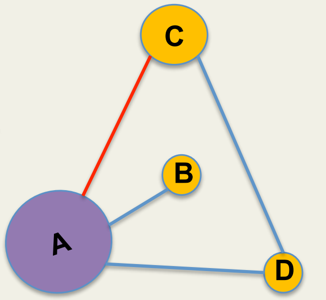
locally smooth data





locally smooth data similarities (hairball)





Thanks -

Computational:

Gunnar Carlsson (Stanford Mathematics)
Sylvia Plevritis (Stanford Cancer Center for Systems Biology)
Rob Tibshirani (Stanford Statistics)

Biology:

Arnold Levine (Princeton IAS – School of Natural Studies)
Anne-Lise Børresen-Dale (Genetics, University of Oslo, Norway)
Stefanie Jeffrey (Stanford Surgery)
Amato Giaccia (Stanford Radiation Oncology)
Janine Erler (Cell and Molecular Biology, Institute of Cancer Research, London, UK)
Ravindra Majeti (Stanford Hematology)
Garry Nolan (Stanford Immunology)

Thanks - funding

NIH – National Human Genome Research Institute (NHGRI)

California Breast Cancer Research Program

DARPA

Air Force Office of Scientific Research

National Institutes of Health