Topological Data Analysis

"Data has shape and shape has meaning"

– *Gunnar Carlsson*

Anthony Gillan-Anderson
Data Visualization-ers
28.03.2017

github.com/amanderson/tda

What is TDA?

- Active area of research, with multiple approaches.
- One approach in particular - `Mapper`

  Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, Singh, Memoli & Carlsson (2007).

What is Mapper?

point cloud: a pair of pants

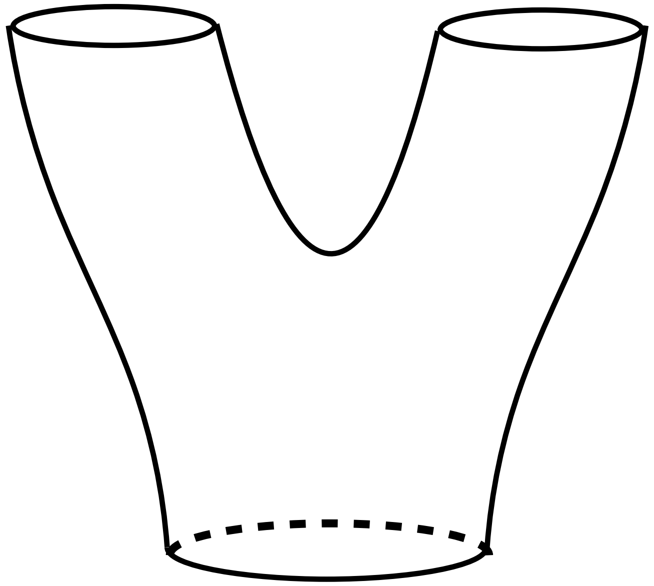
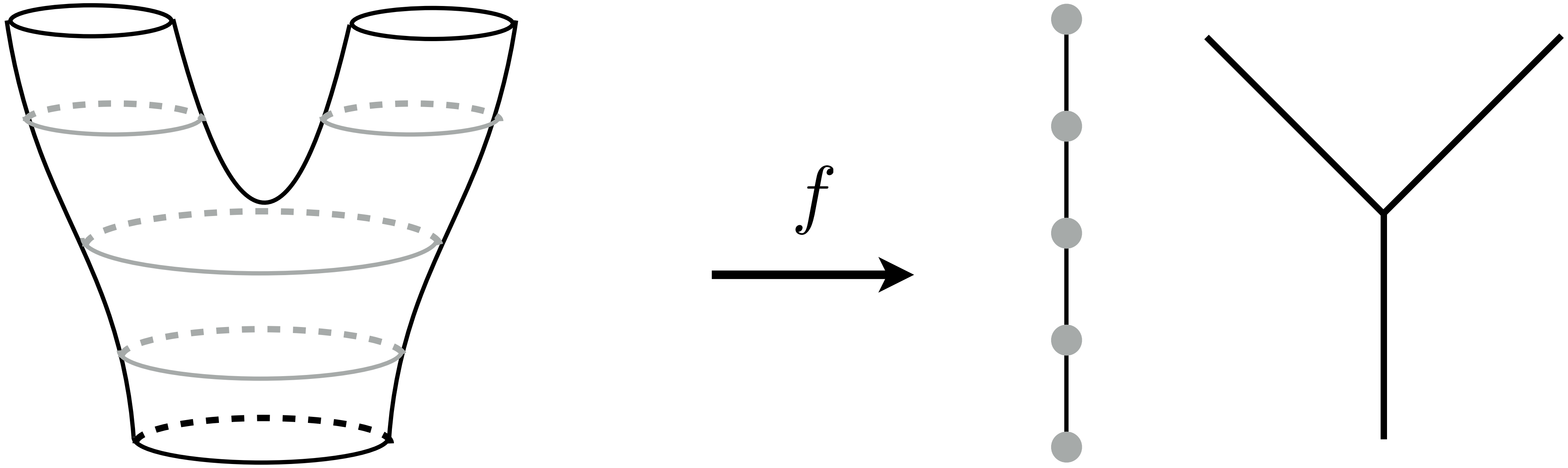


Illustration borrowed from Anthony Bak

What is Mapper?

point cloud: a pair of pants



Choose a “lens”:

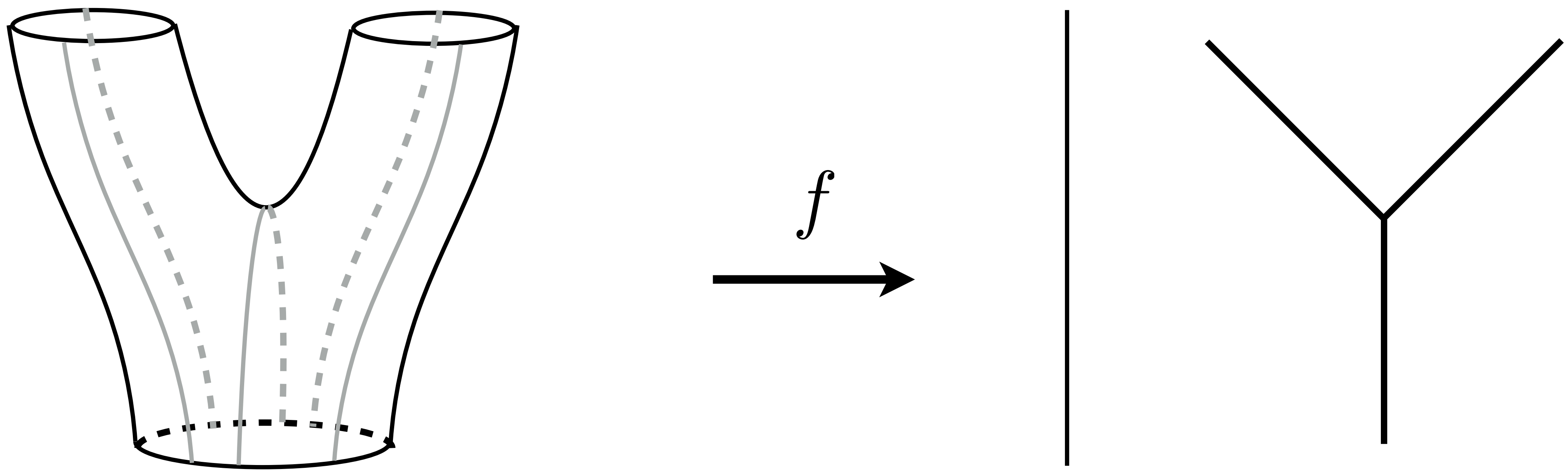
Here ***f*** is the vertical height of a data point on the pair of pants.

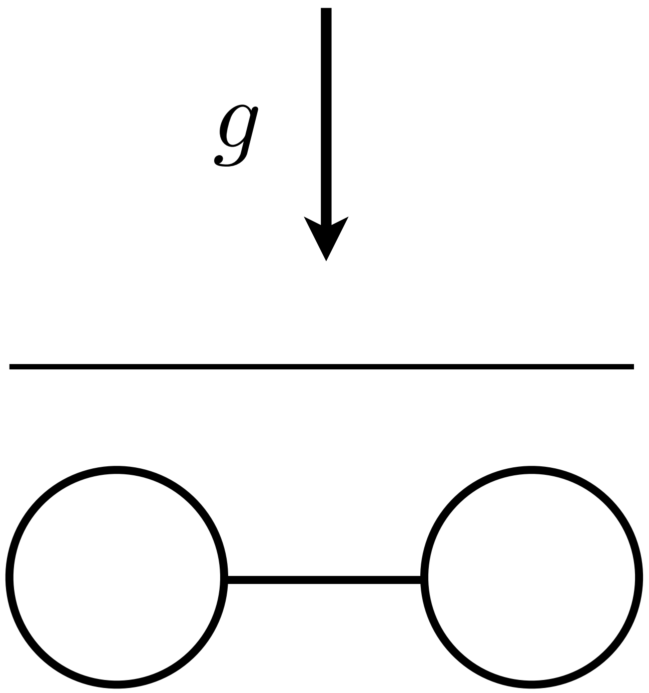
Through lens ***f***, shape is summarised by a **Y**: the inverse image of ***f*** has a single isoline split in two.

Illustration borrowed from Anthony Bak

What is Mapper?

point cloud: a pair of pants





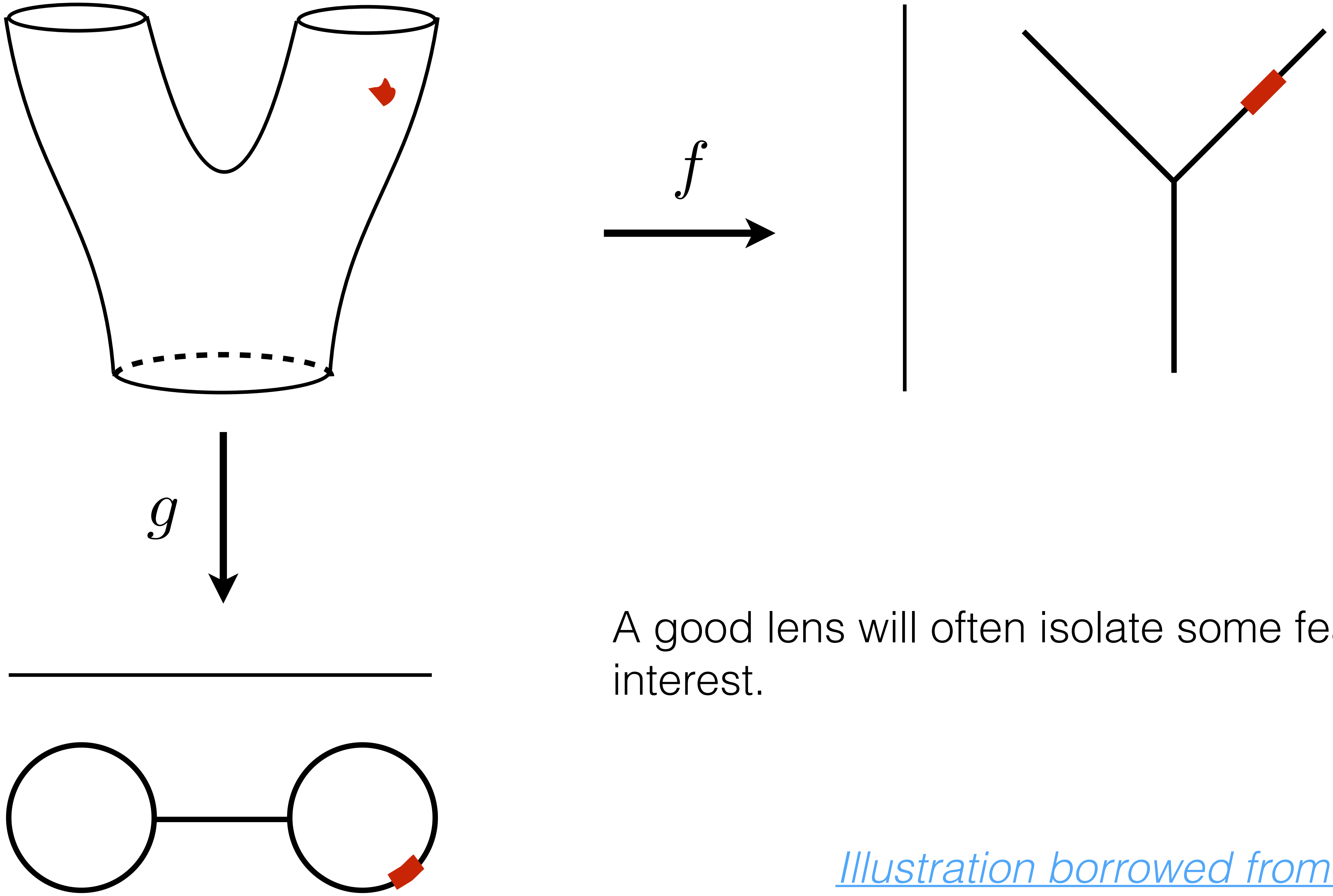
Choose a “lens”:

Here ***g*** is the horizontal position of a data point on the pair of pants.

Through lens ***g***, shape is summarised by a **O-O**: the inverse image of ***g*** has contours splitting through the legs.

Illustration borrowed from Anthony Bak

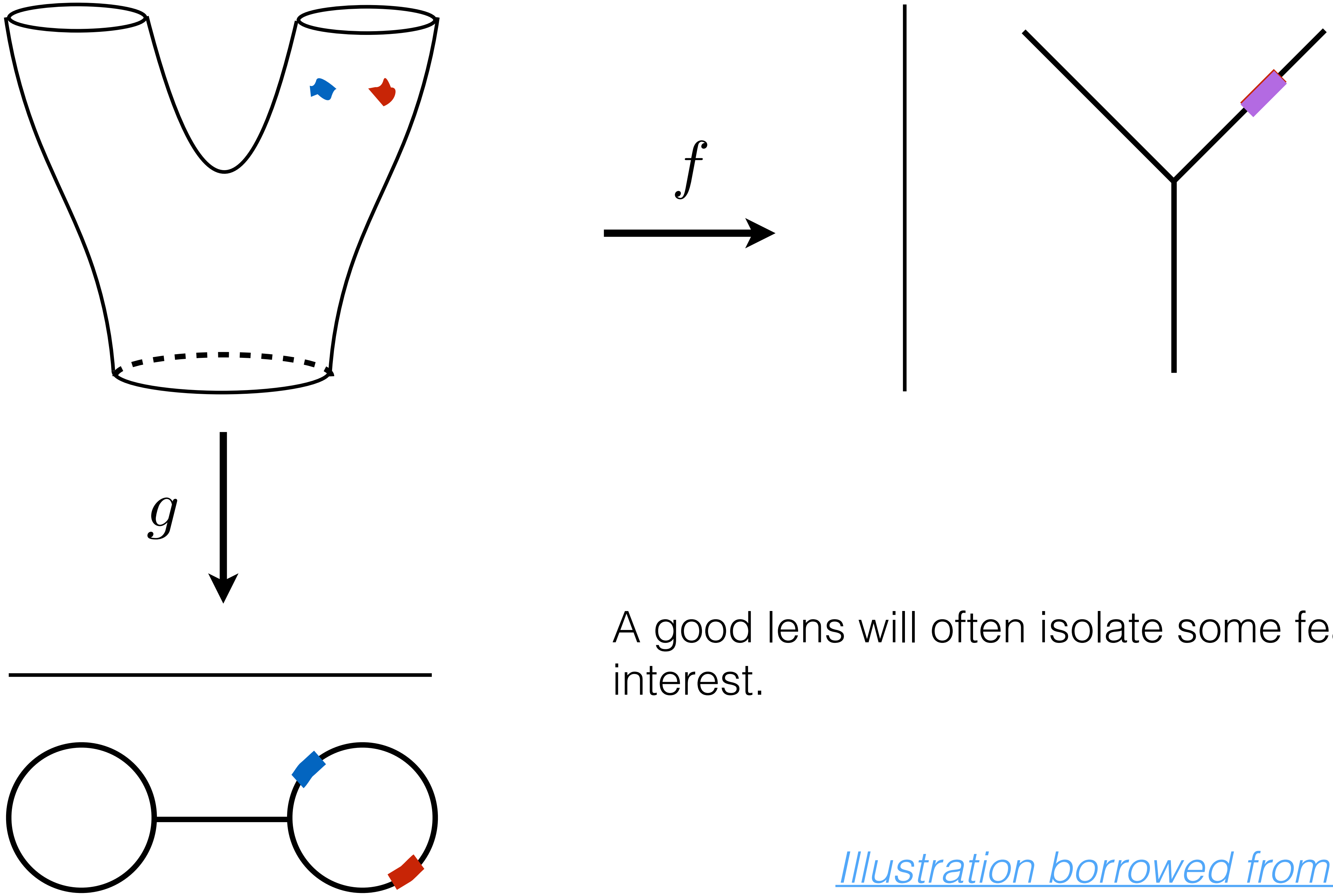
What is Mapper?



A good lens will often isolate some feature of interest.

Illustration borrowed from Anthony Bak

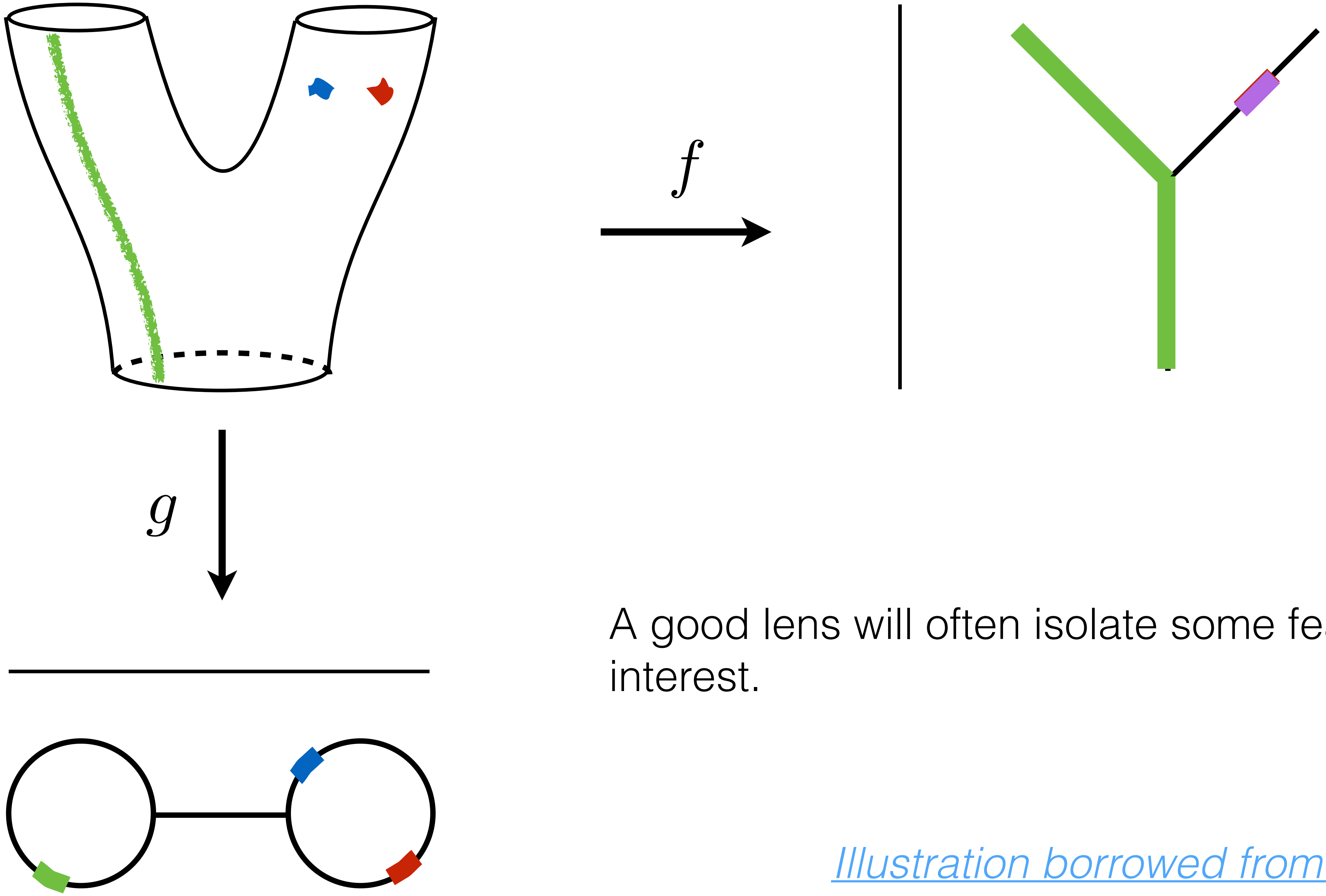
What is Mapper?



A good lens will often isolate some feature of interest.

Illustration borrowed from Anthony Bak

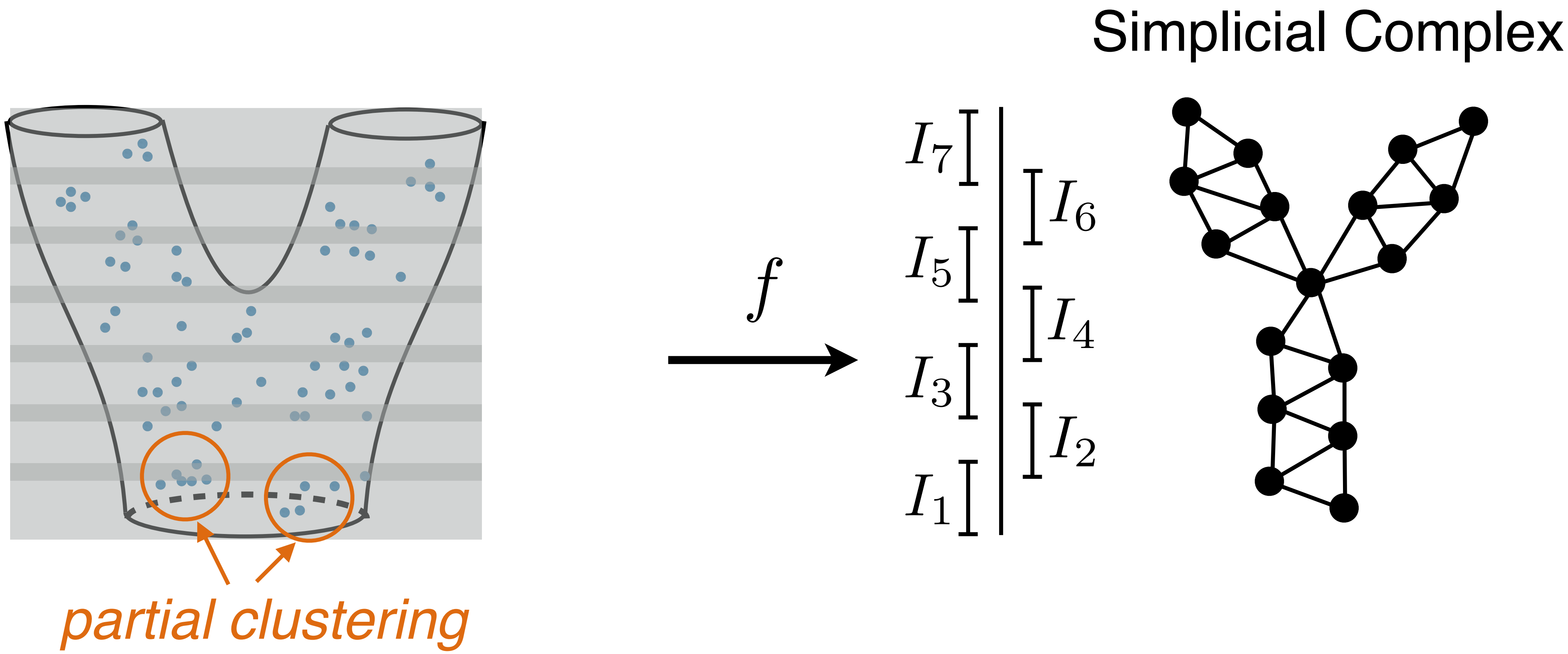
What is Mapper?



A good lens will often isolate some feature of interest.

Illustration borrowed from Anthony Bak

What is Mapper?

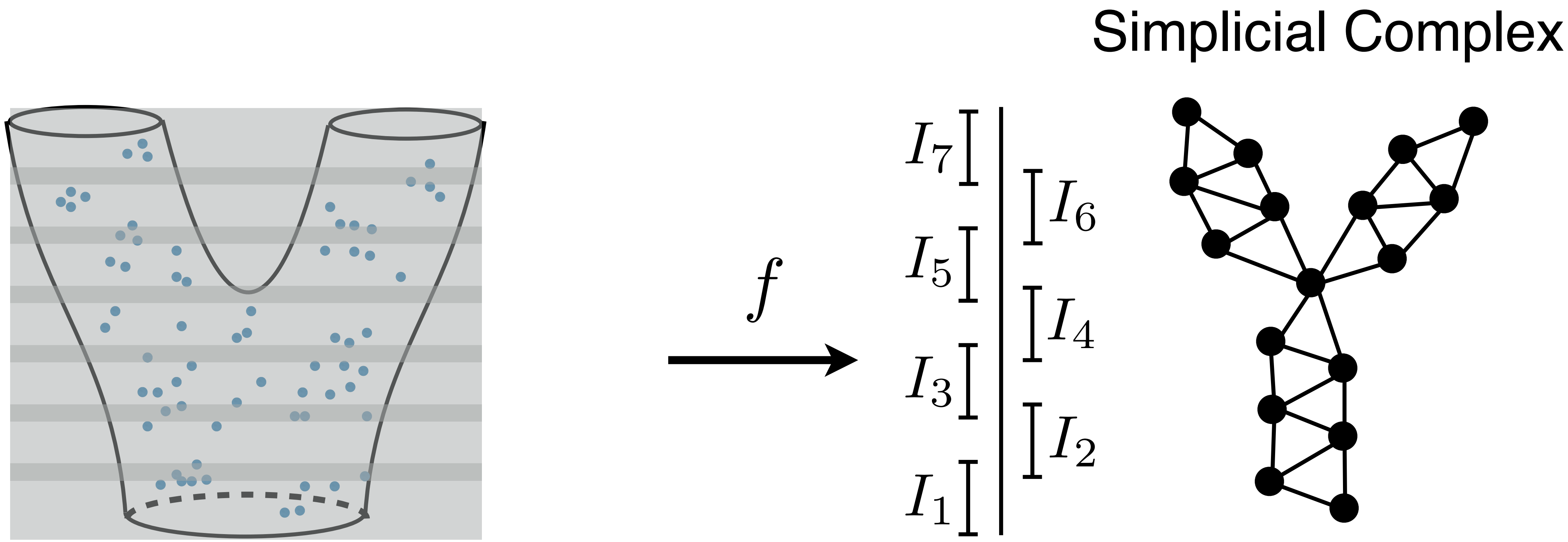


Partial clusters on each interval become nodes in a simplicial complex or similarity graph.

Nodes are connected by edge if their clusters share common data points. This is made possible be over sampling with overlapping intervals.

Illustration borrowed from Anthony Bak

What is Mapper?



Implementation decisions:

- Distance/dissimilarity metric
- Filter function(s)
- Partition of *f*
- Clustering algorithm

Illustration borrowed from Anthony Bak

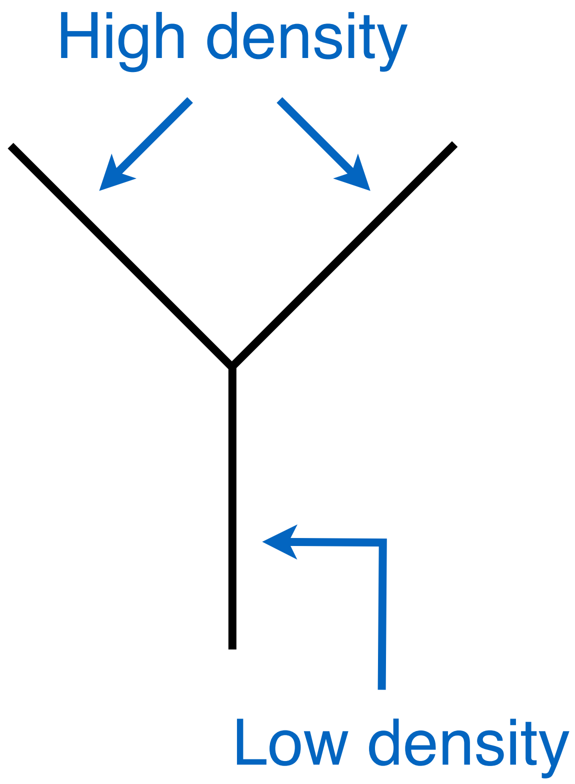
What’s in the box?

Black Box

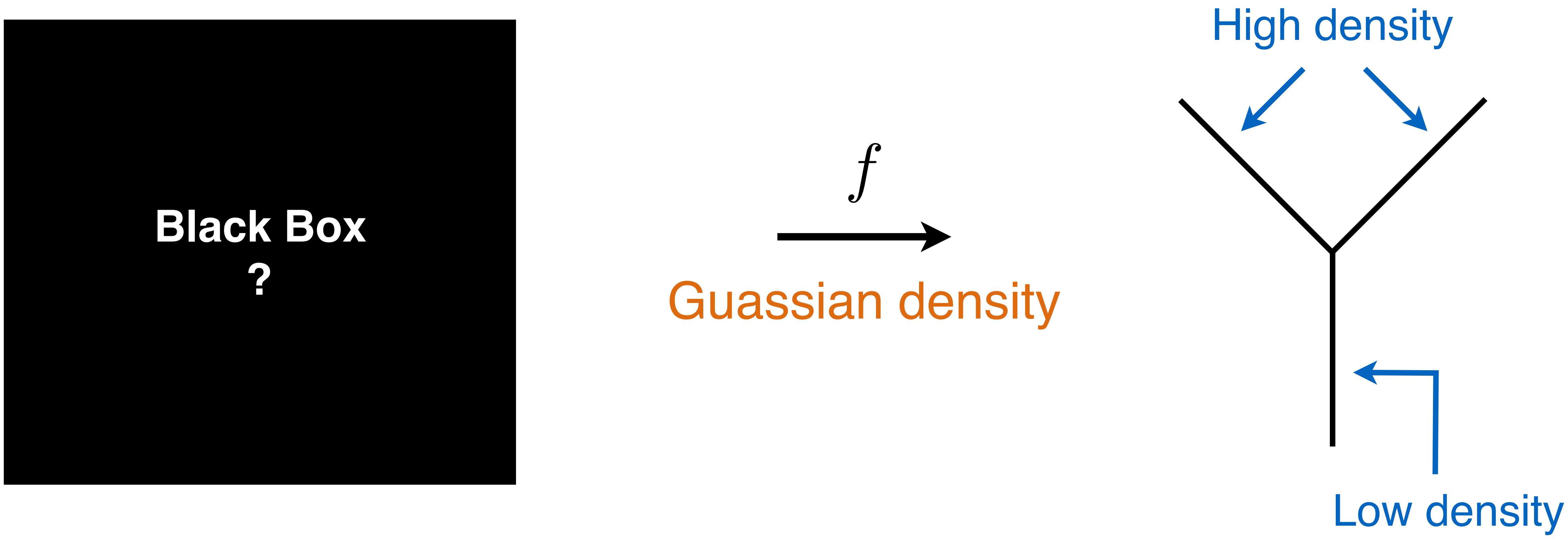
?

$$\xrightarrow{f}$$

Guassian density



What’s in the box?



Black Box
?

f

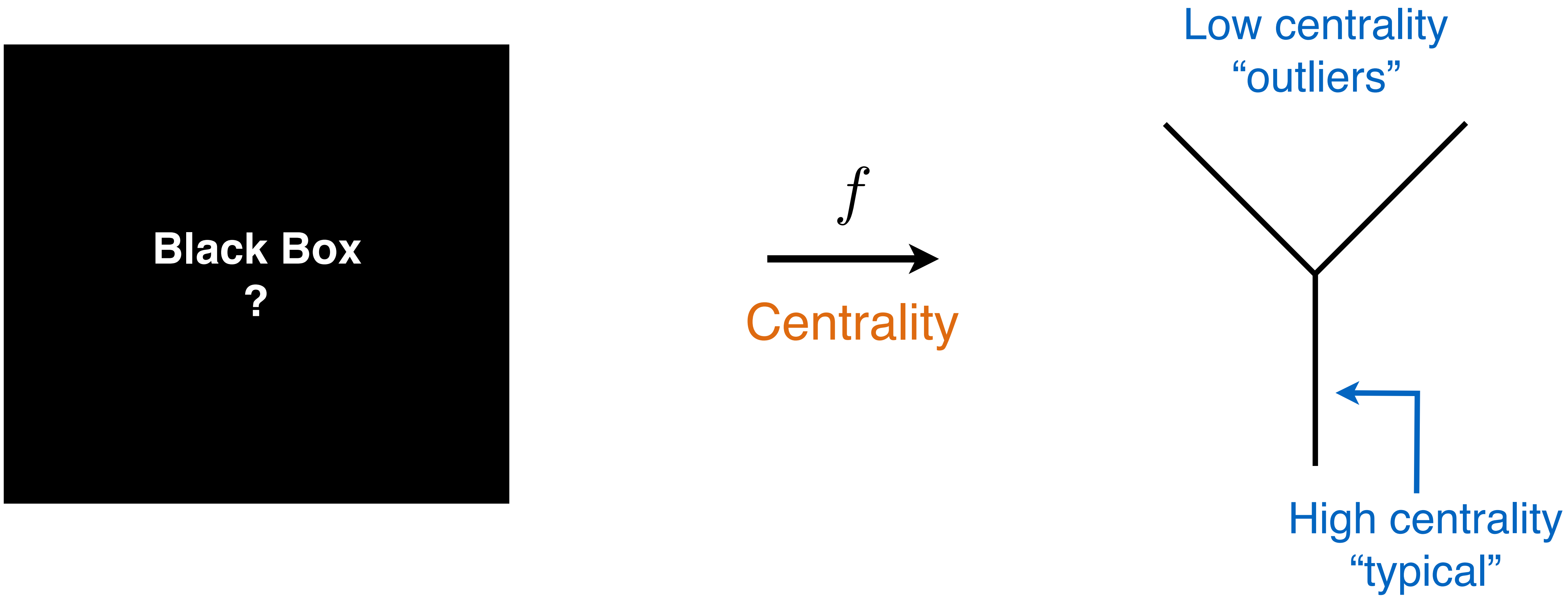
Guassian density

High density

Low density

The data is drawn from a bimodal distribution

What’s in the box?



Black Box

?

$\xrightarrow{f}$

Centrality

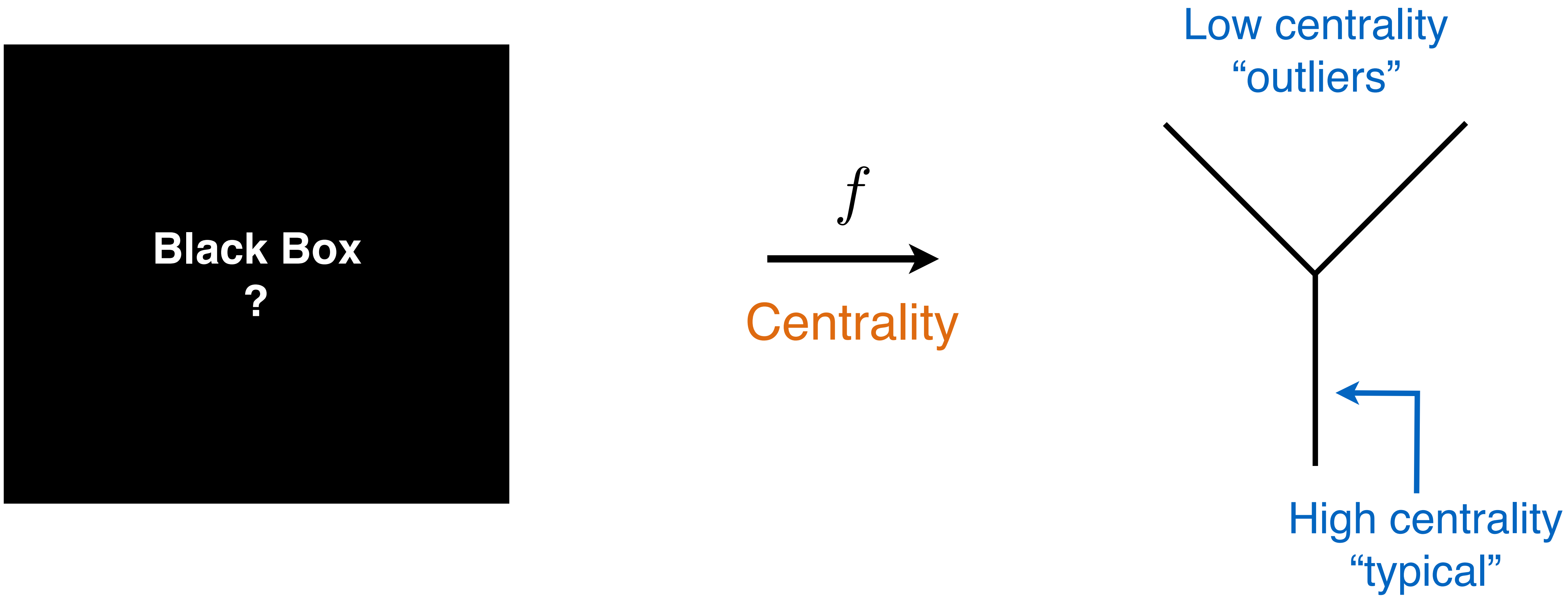
Low centrality

“outliers”

High centrality

“typical”

What’s in the box?



Black Box

?

f

Centrality

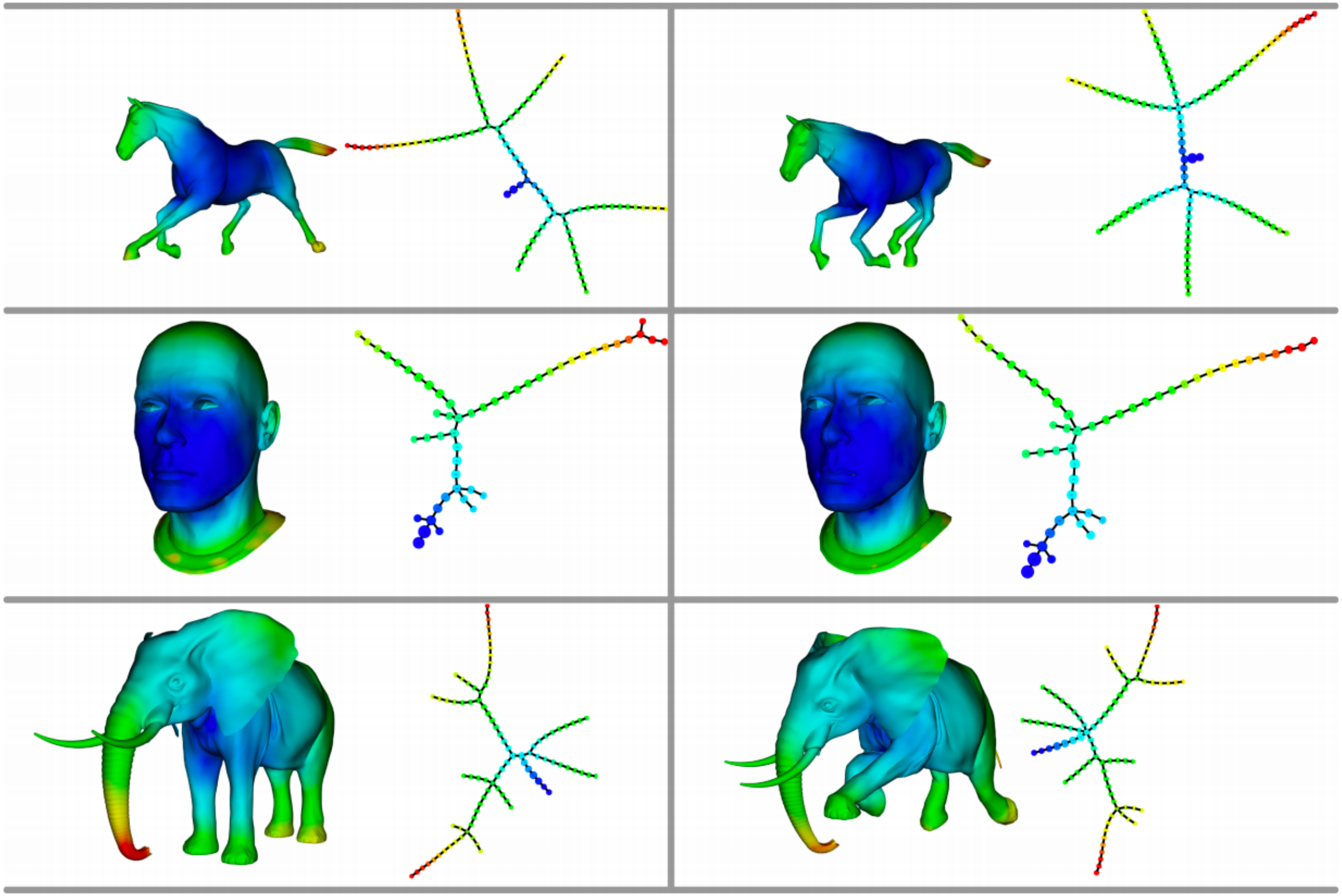
Low centrality
“outliers”

High centrality
“typical”

The data has two qualitatively distinct outlier types

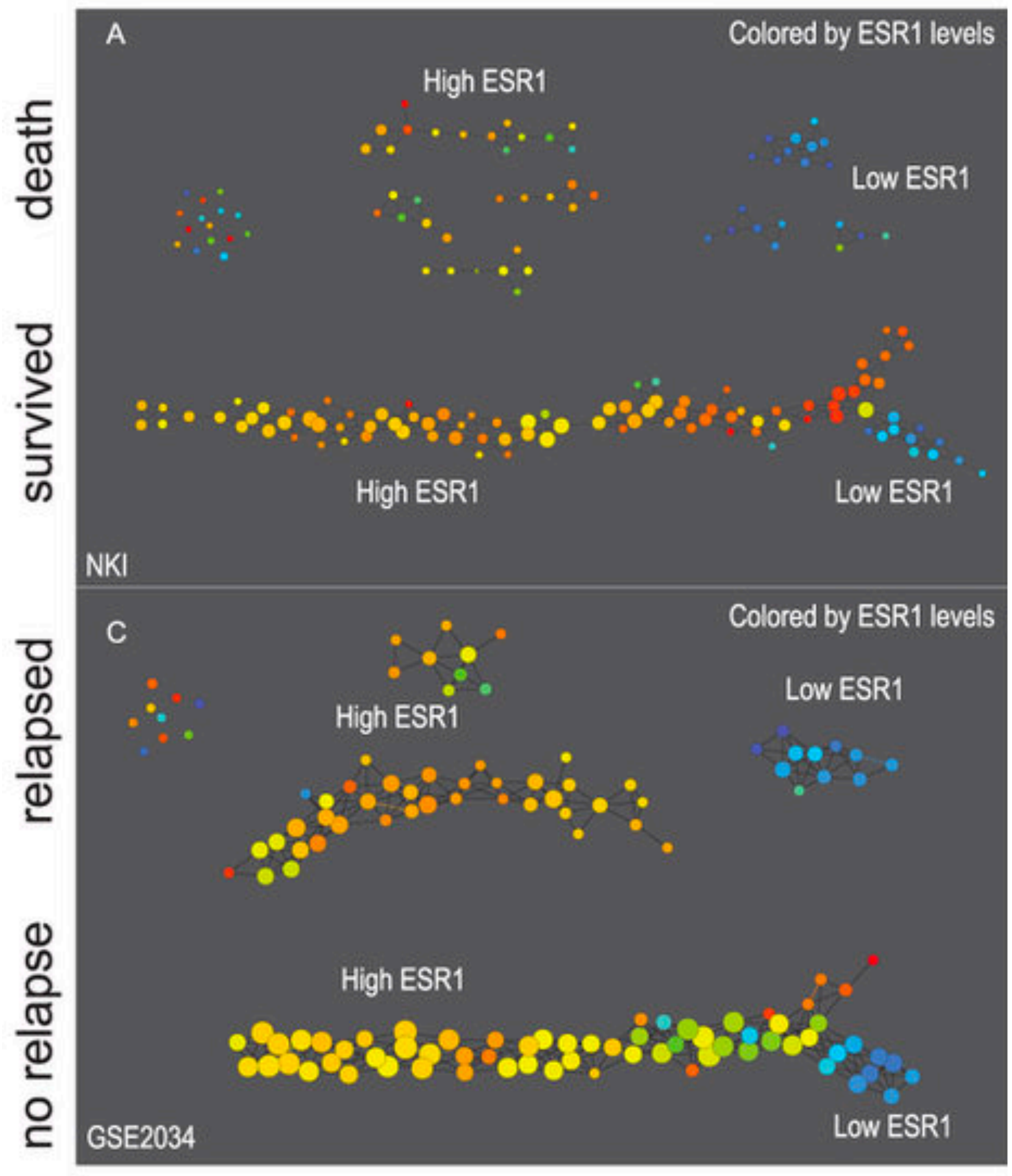
e.g. Type I & Type II diabetes

Example: Image Analysis



Singh, Memoli & Carlsson (2007)

Example: Breast Cancer



Lum *et al*, Nature (2013)

Data sets:

(a) NKI - patient survival based on 1.5k gene expression levels.

(b) GSE2034 - patient relapse time on 1.5k genes with highest variance.

Dissimilarity metric:

Correlation distance

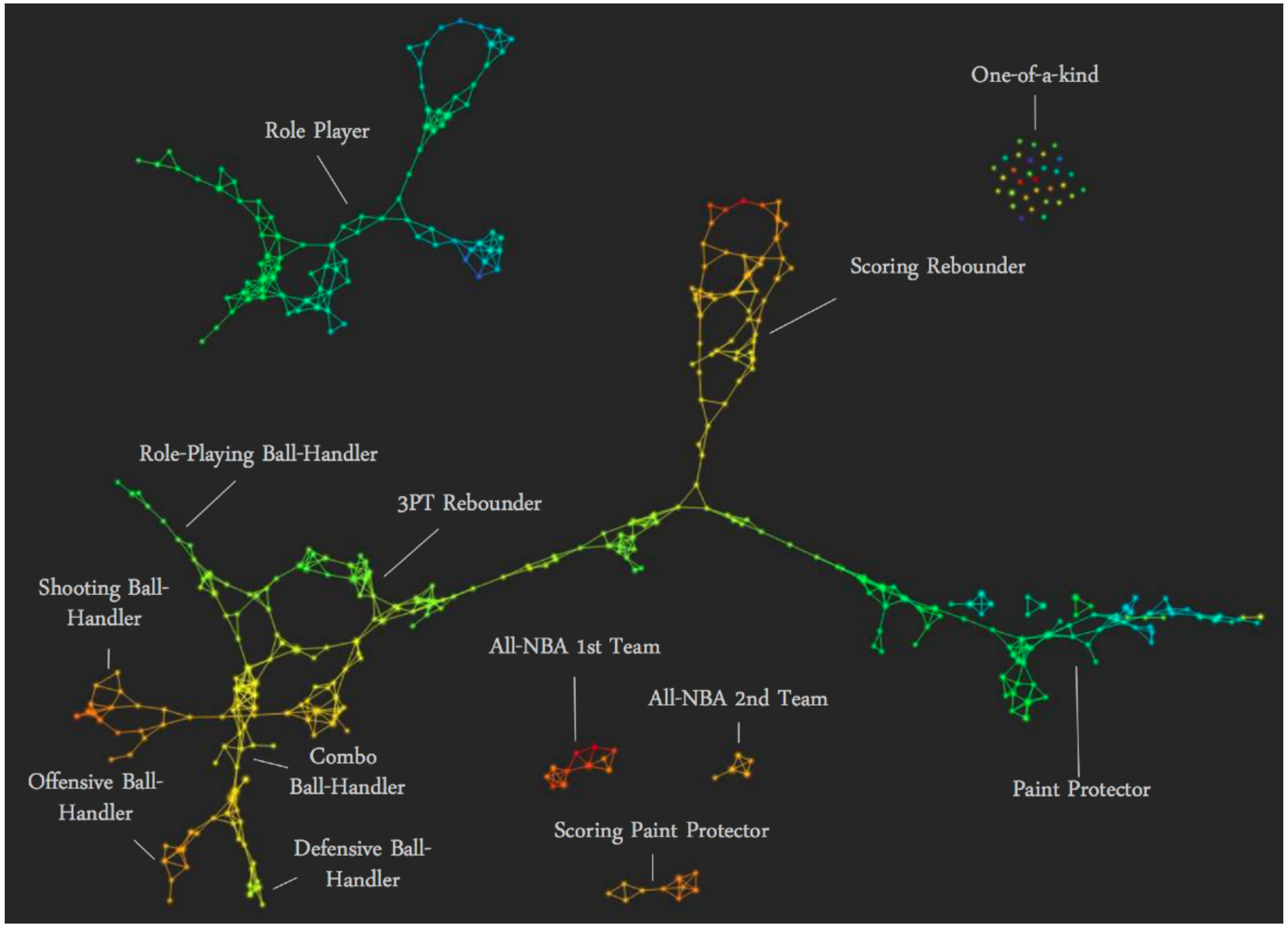
Filter functions:

Survival outcome, L-infinity centrality

Clustering:

single-linkage clustering

Example: NBA



Alagappan, Ayasdi (2012)
MIT Slone Sports Analytics Conference

Data set:
452 players, 7 stats categories (pts, rebs, blk, ast, stl, tov, pf)

Dissimilarity metric:
variance-normalised Euclidean

Filter functions:
1st & 2nd SVD components

Clustering:
single-linkage clustering

Open-source Libraries

- Python
  - **Mapper** (http://danifold.net/mapper/)
  - KeplerMapper (https://github.com/MLWave/kepler-mapper)
- R
  - TDAMapper (https://cran.r-project.org/web/packages/TDA)
- Matlab
  - Original Mapper paper

demos: NBA, hand-written digits

More Resources

- Anthony Bak is an actual expert in TDA and speaks very well on this topic from the viewpoint of a practitioner:
  - How Ayasdi used TDA to Solve Complex Problems
  - TDA for the Working Data Scientist
- The Ayasdi website has an archive of blog postings and white papers describing their platform and applications for TDA.
- Technical articles:
  - Original Mapper article
  - TDA for breast cancer outcomes (including statistical analysis of shape).
  - If you're interested in the maths, see Carlsson's seminal article.
- My GitHub page (@amanderson) has a TDA repo with a (hopefully growing) set of notebooks.