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 Recognitio*n, Singh, Memoli & Carlsson (2007).

how many are familiar with TDA? maybe say a few words about key ideas borrowed from mathematical topology.

it's a very active area of research, with multiple approaches. i'll discuss an approach called MAPPER, which gets it's name from it's algorithm. this is the approach behind Ayasdi.

i'm new to this. introduced by former colleague Anthony Bak….

What is `Mapper`?

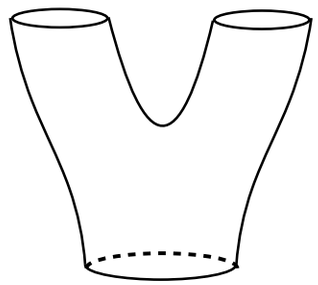


Illustration borrowed from Anthony Bak

… i’m going to borrow his pair-of-pants illustration.

Let's imagine our points cover the surface of this pair of pants. (real datasets don’t look like this, in fact we can’t see them)

lenses provide a geometrical summary and are a means of querying the dataset.

Walk through two examples. Guess the summary emitted by g(X).

What is `Mapper`?

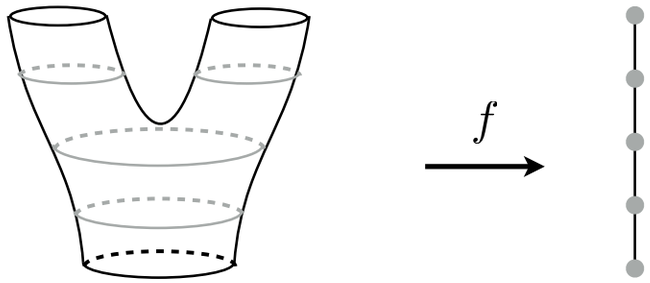


Illustration borrowed from Anthony Bak

… i’m going to borrow his pair-of-pants illustration.

Let's imagine our points cover the surface of this pair of pants. (real datasets don’t look like this, in fact we can’t see them)

lenses provide a geometrical summary and are a means of querying the dataset.

Walk through two examples. Guess the summary emitted by g(X).

What is `Mapper`?

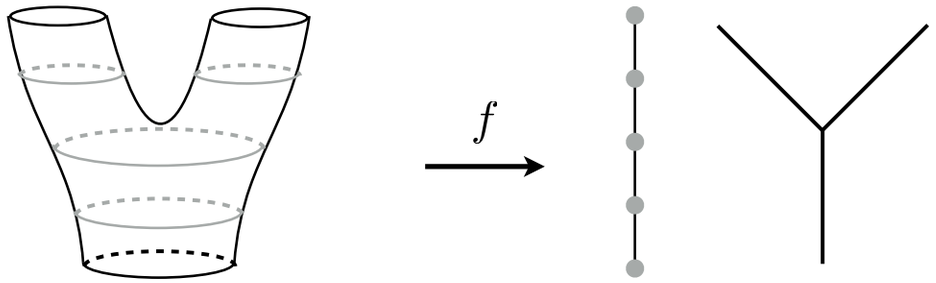


Illustration borrowed from Anthony Bak

… i’m going to borrow his pair-of-pants illustration.

Let's imagine our points cover the surface of this pair of pants. (real datasets don’t look like this, in fact we can’t see them)

lenses provide a geometrical summary and are a means of querying the dataset.

Walk through two examples. Guess the summary emitted by g(X).

What is `Mapper`?

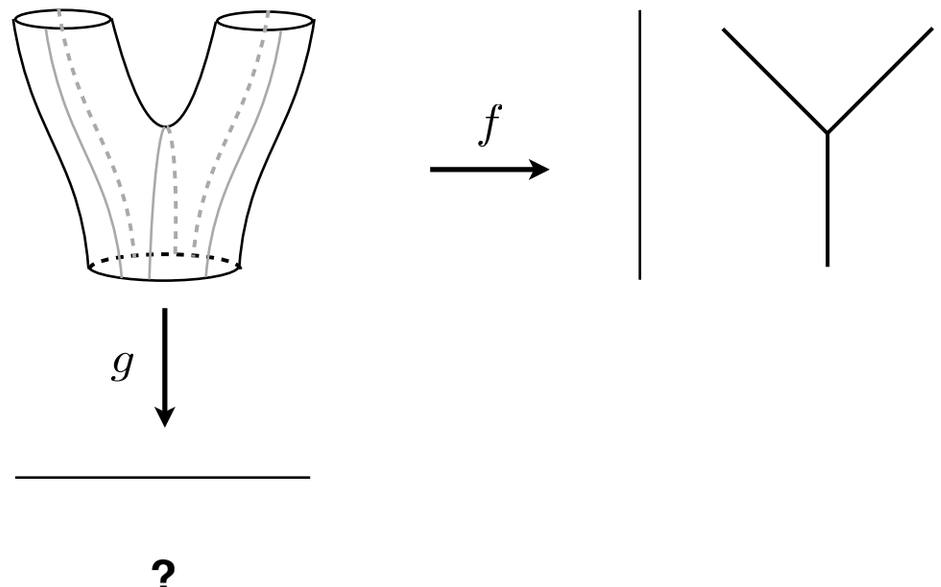


Illustration borrowed from Anthony Bak

… i’m going to borrow his pair-of-pants illustration.

Let's imagine our points cover the surface of this pair of pants. (real datasets don’t look like this, in fact we can’t see them)

lenses provide a geometrical summary and are a means of querying the dataset.

Walk through two examples. Guess the summary emitted by g(X).

What is Mapper?

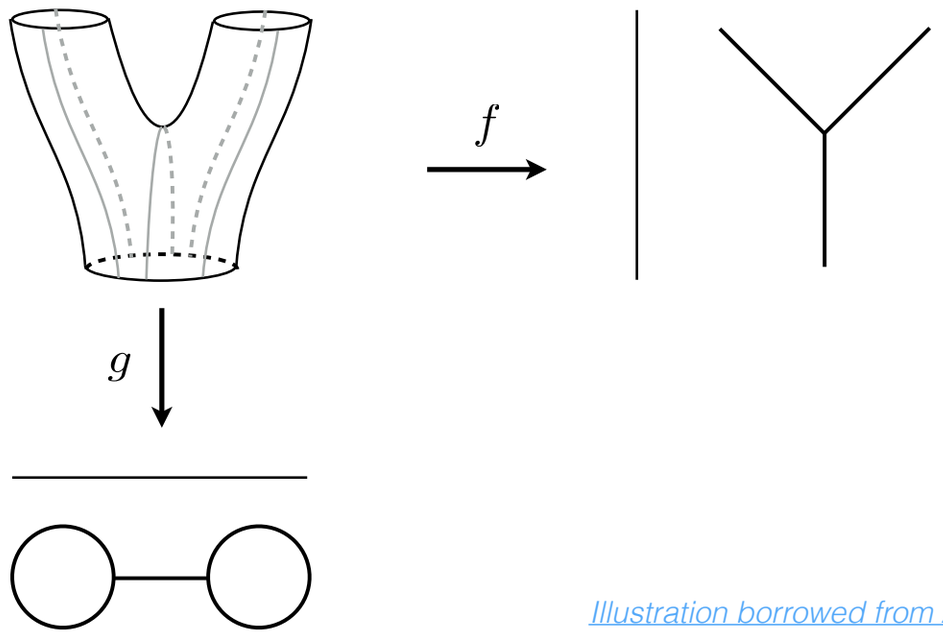


Illustration borrowed from Anthony Bak

… i’m going to borrow his pair-of-pants illustration.

Let's imagine our points cover the surface of this pair of pants. (real datasets don’t look like this, in fact we can’t see them)

lenses provide a geometrical summary and are a means of querying the dataset.

Walk through two examples. Guess the summary emitted by g(X).

What is Mapper?

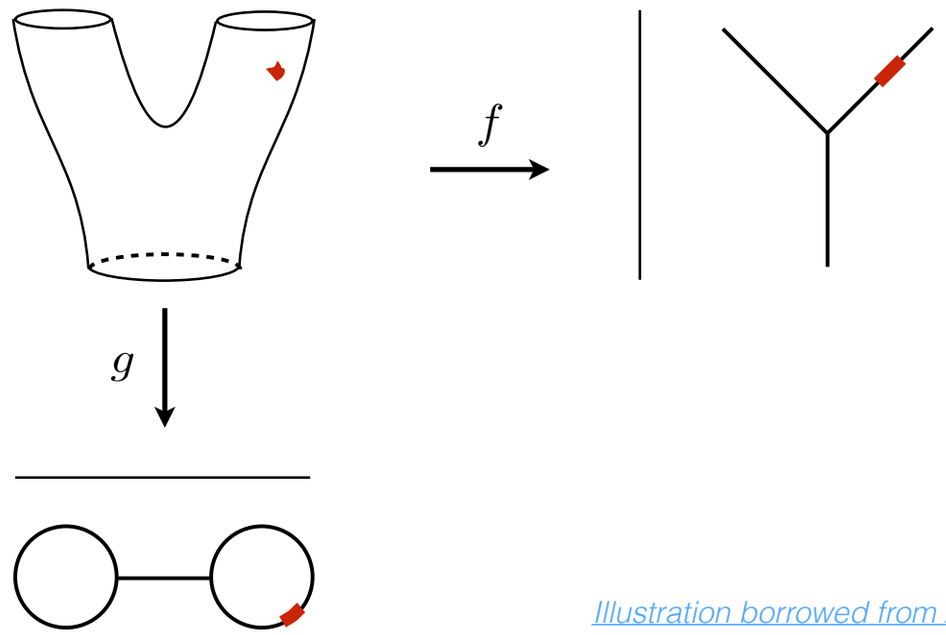


Illustration borrowed from Anthony Bak

A lens is just that — it provides a compressed and explorable view of the data.

what makes a good lens? Often (not always) we want to isolate some feature of the underlying data to be able to draw conclusions. Think fraud, disease, systematic error in ML models, etc.

Sometimes the geometric features themselves tell the story, e.g. closed loop corresponding to cycles in sequential data, like boom and bust cycles in economic time-series.

Whatever it is, they offer different views of the data, not all of them useful.

What is Mapper?

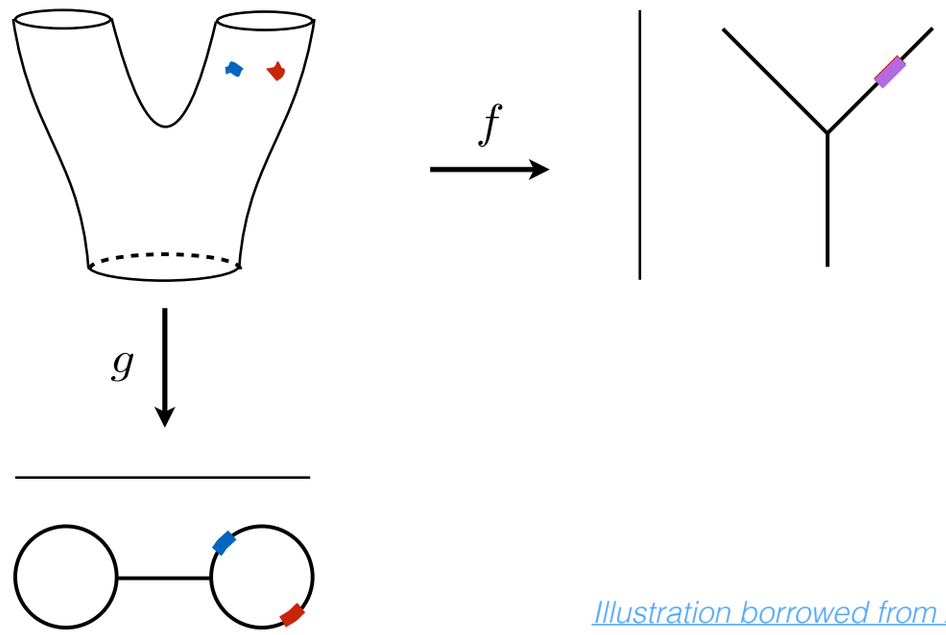


Illustration borrowed from Anthony Bak

A lens is just that — it provides a compressed and explorable view of the data.

what makes a good lens? Often (not always) we want to isolate some feature of the underlying data to be able to draw conclusions. Think fraud, disease, systematic error in ML models, etc.

Sometimes the geometric features themselves tell the story, e.g. closed loop corresponding to cycles in sequential data, like boom and bust cycles in economic time-series.

Whatever it is, they offer different views of the data, not all of them useful.

What is Mapper?

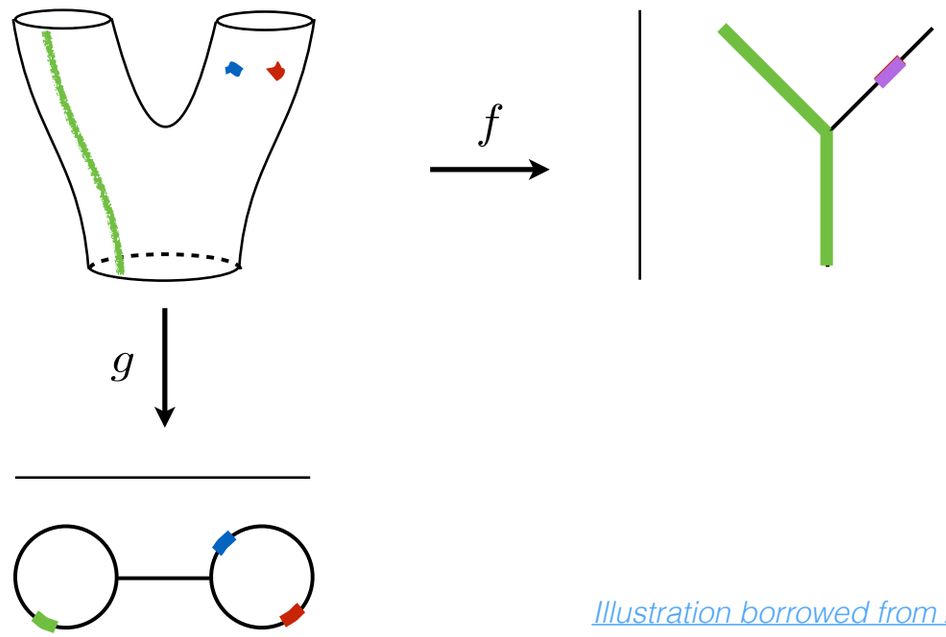


Illustration borrowed from Anthony Bak

A lens is just that — it provides a compressed and explorable view of the data.

what makes a good lens? Often (not always) we want to isolate some feature of the underlying data to be able to draw conclusions. Think fraud, disease, systematic error in ML models, etc.

Sometimes the geometric features themselves tell the story, e.g. closed loop corresponding to cycles in sequential data, like boom and bust cycles in economic time-series.

Whatever it is, they offer different views of the data, not all of them useful.

What is `Mapper`?

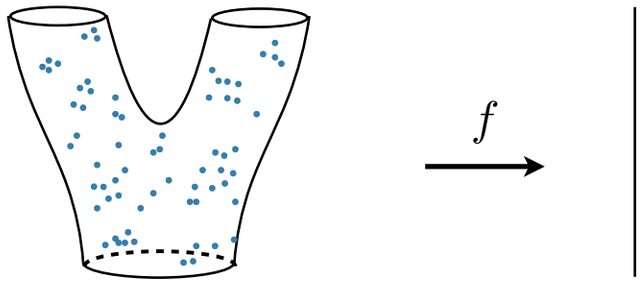


Illustration borrowed from Anthony Bak

Real data sets are discrete and noisy (and never look like a pair of pants). To deal with this, Mapper generalises on the previous summaries, producing similarity graphs using partial clustering on overlapping intervals.

To produce useful summaries, we need to make a lot of decisions — there's no free lunch! We need to me mindful of false positives.

Fortunately, intuitive chooses often bear fruit. Incremental knowledge can still be accrued with suboptimal choices. Remember these are just views/queries on the data.

What is `Mapper`?

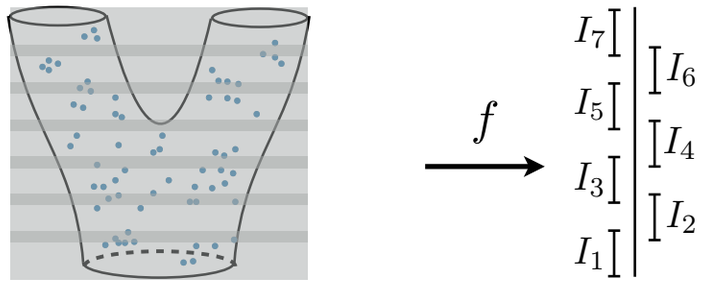


Illustration borrowed from Anthony Bak

Real data sets are discrete and noisy (and never look like a pair of pants). To deal with this, Mapper generalises on the previous summaries, producing similarity graphs using partial clustering on overlapping intervals.

To produce useful summaries, we need to make a lot of decisions — there's no free lunch! We need to me mindful of false positives.

Fortunately, intuitive chooses often bear fruit. Incremental knowledge can still be accrued with suboptimal choices. Remember these are just views/queries on the data.

What is Mapper?

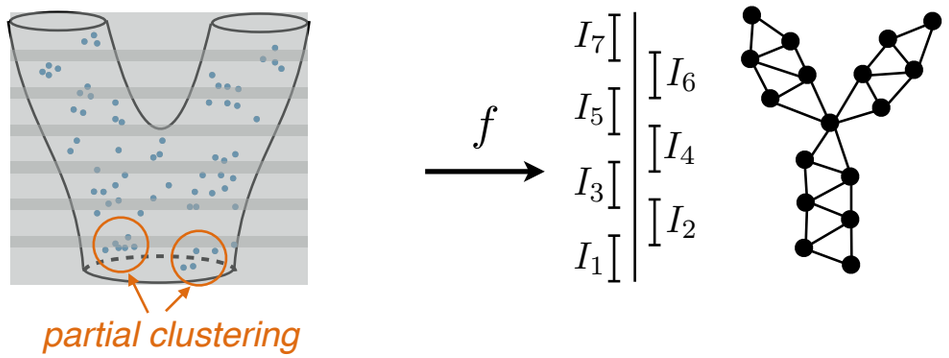


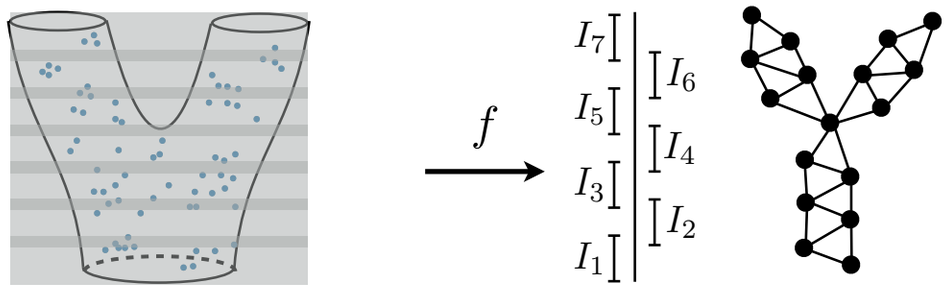
Illustration borrowed from Anthony Bak

Real data sets are discrete and noisy (and never look like a pair of pants). To deal with this, Mapper generalises on the previous summaries, producing similarity graphs using partial clustering on overlapping intervals.

To produce useful summaries, we need to make a lot of decisions — there's no free lunch! We need to me mindful of false positives.

Fortunately, intuitive chooses often bear fruit. Incremental knowledge can still be accrued with suboptimal choices. Remember these are just views/queries on the data.

What is Mapper?



Implementation decisions:

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

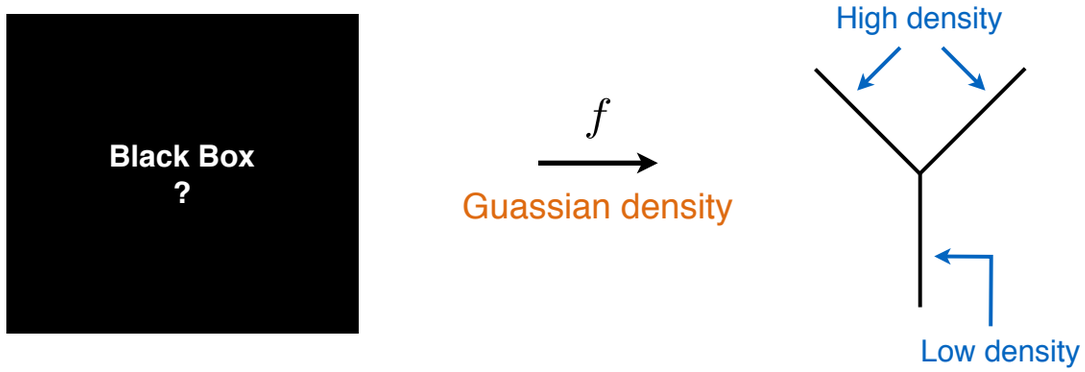
Illustration borrowed from Anthony Bak

Real data sets are discrete and noisy (and never look like a pair of pants). To deal with this, Mapper generalises on the previous summaries, producing similarity graphs using partial clustering on overlapping intervals.

To produce useful summaries, we need to make a lot of decisions — there's no free lunch! We need to me mindful of false positives.

Fortunately, intuitive chooses often bear fruit. Incremental knowledge can still be accrued with suboptimal choices. Remember these are just views/queries on the data.

What's in the box?

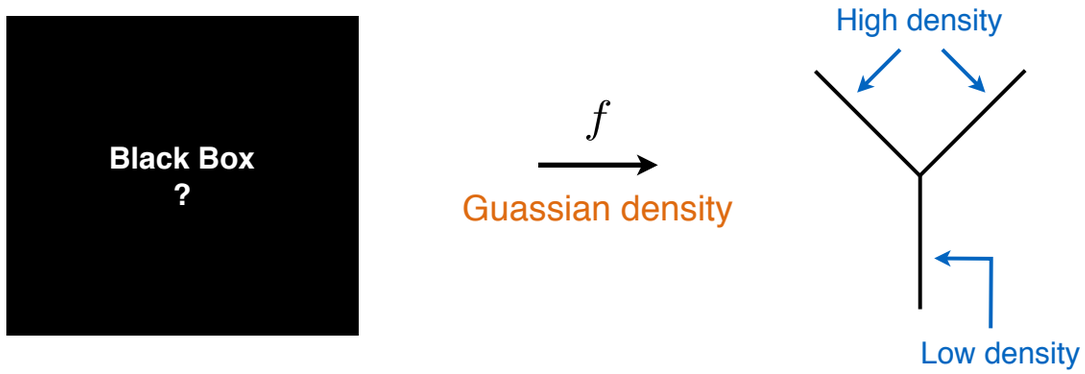


Real datasets are black boxes. Quiz: what's in the black box?

Let's test our understanding with some lenses that are useful in practice.
- Gaussian density
- Centrality

Other useful metrics:
- Statistics (mean, variance, etc.)
- Dimensionality reduction (PCA, MDS, Isomap, …)
- ML models (e.g. error)

What's in the box?



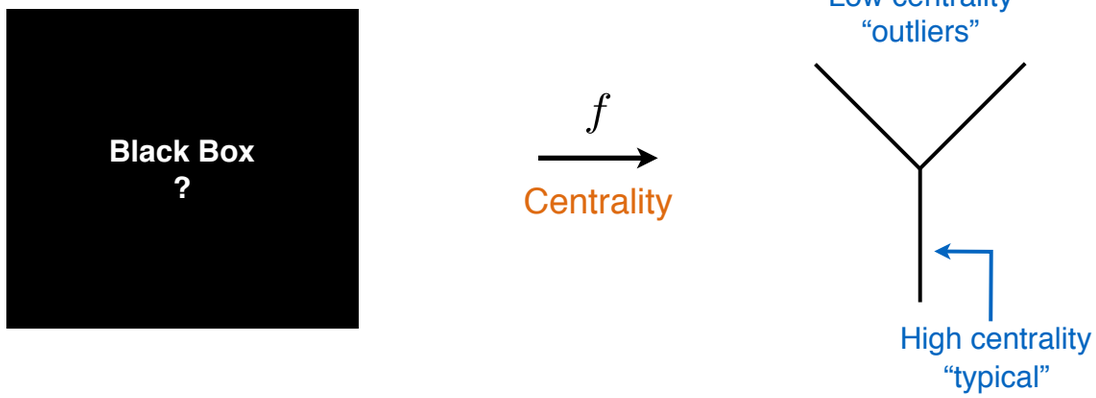
The data is drawn from a bimodal distribution

Real datasets are black boxes. Quiz: what's in the black box?

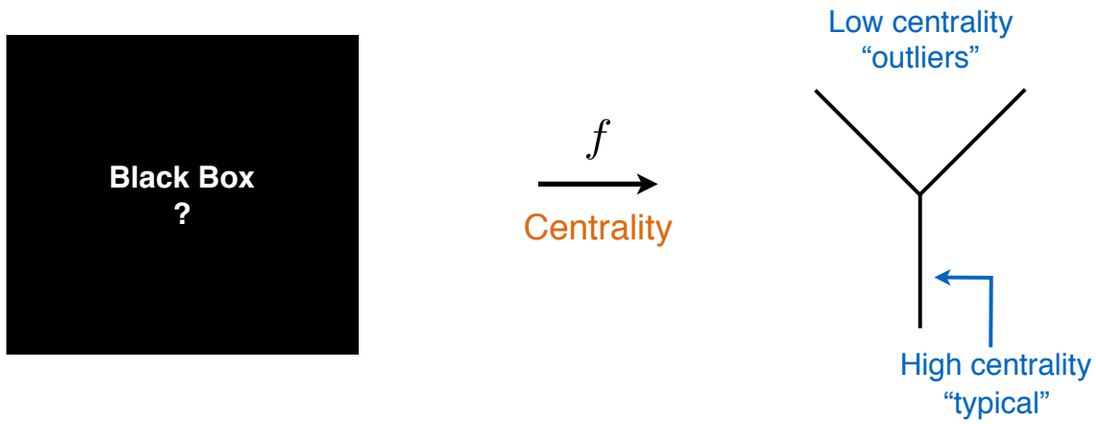
Let's test our understanding with some lenses that are useful in practice.
- Gaussian density
- Centrality

Other useful metrics:
- Statistics (mean, variance, etc.)
- Dimensionality reduction (PCA, MDS, Isomap, …)
- ML models (e.g. error)

What's in the box?



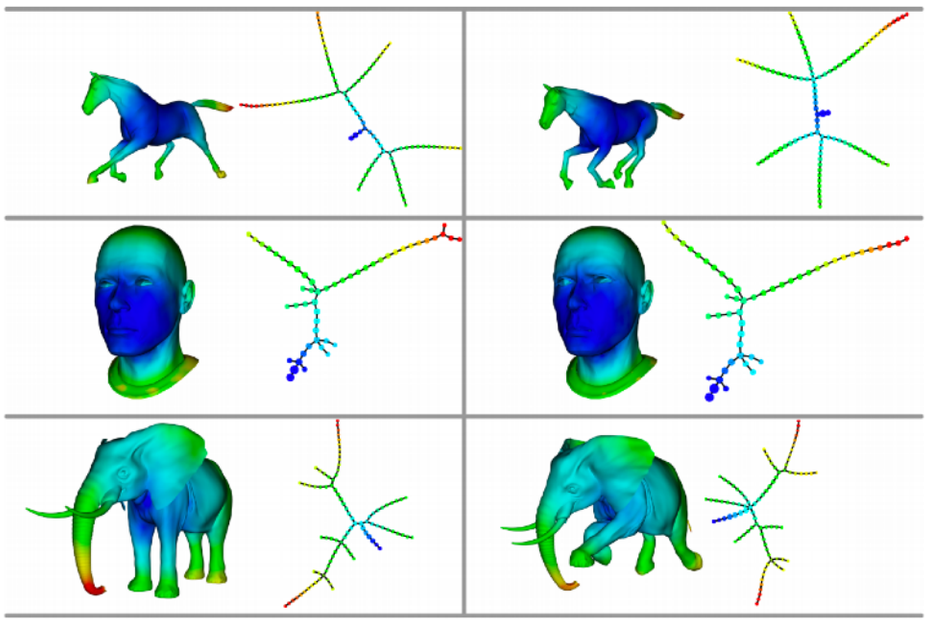
What's in the box?



The data has two qualitatively distinct outlier types

e.g. Type I & Type II diabetes

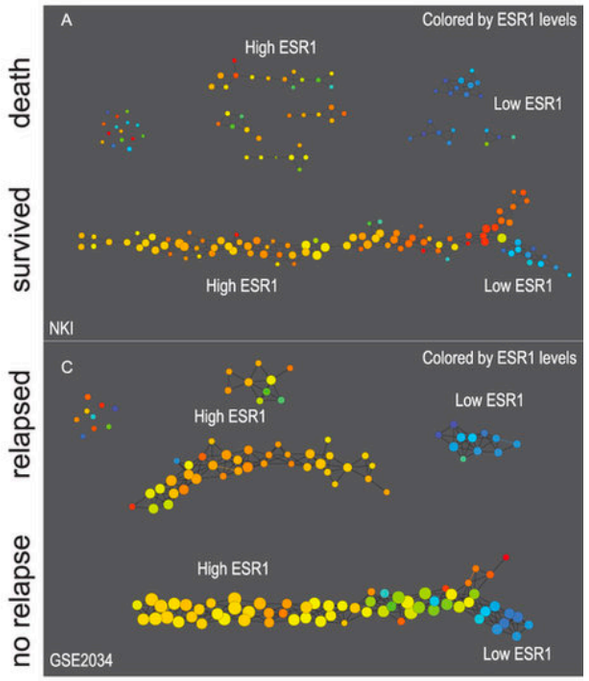
Example: Image Analysis



Singh, Memoli & Carlsson (2007)

Deformation invariance: comparing similarity metrics left and right shows equivalence.

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

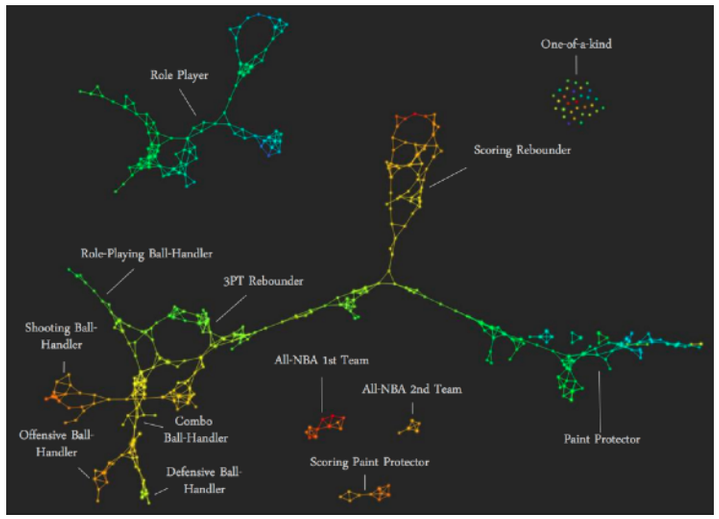
Clustering:
single-linkage clustering

expression levels of the oestrogen receptor ESR1 is correlated with improved prognosis; however, TDA reveals a is a patient subset that does not fair well.

** this is corroborated by independent datasets **

Through separate statistical tests, the authors identified a set of genes associated with a key pathway paired with survival.

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.