Intro to Topological Data Analysis

Statistics and Dimension Reduction

Chunyin Siu, Feb 14, 2025

Intro to Topological Data Analysis

Statistics and Dimension Reduction

Chunyin Siu, Feb 14, 2025

Intro to Topological Data Analysis

Statistics and Dimension Reduction

Chunyin Siu, Feb 14, 2025

What is Data Analysis

What is Data Analysis

Get Data

Sérgio Valle Duarte https://commons.wikimedia.org/wiki/File:BinaryData.jpg

What is Data Analysis

```
const LOCALE = globalThis.navigator.language

const div = document.body.appendChild(document.createElement('div'))
const list = div.appendChild(document.createElement('ol'))

const dayNames = new Map()

for (let i = 0; i < 7; ++i) {
    const d = Temporal.PlainDate.from({
        year: Temporal.Now.plainDateISO().year,
        month: 1,
        day: i + 1,
    })

    dayNames.set(d.dayOfWeek, d.toLocaleString(LOCALE, { weekday: 'long' }))
}

for (const num of [...dayNames.keys()].sort((a, b) => a - b)) {
    list.appendChild(Object.assign(
        document.createElement('li'),
        { textContent: dayNames.get(num) },
    ))
}
```

Get Data

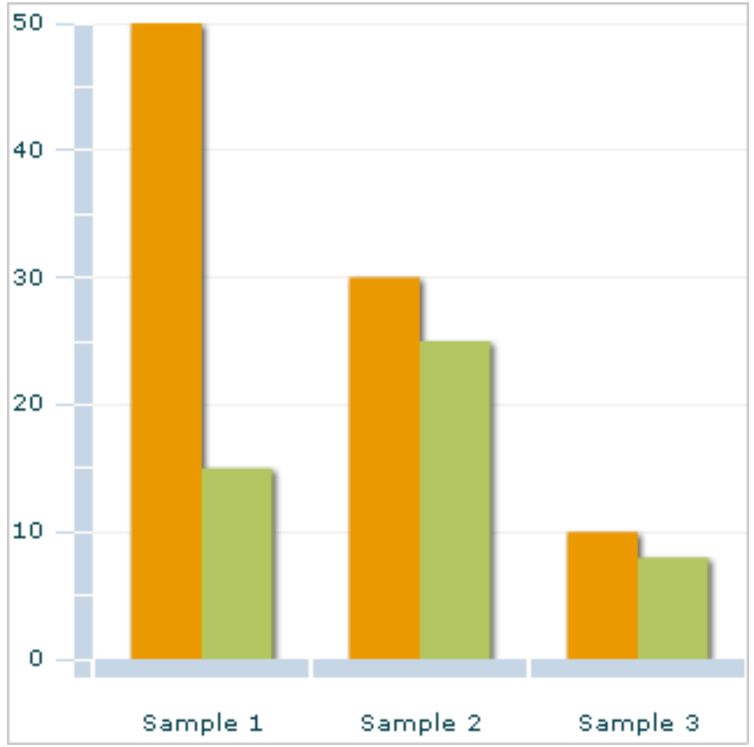
Sérgio Valle Duarte https://commons.wikimedia.org/wiki/File:BinaryData.jpg

Analyze Data

Lionel Rowe https://commons.wikimedia.org/wiki/File:JavaScript_code.png

What is Data Analysis

```
const LOCALE = globalThis.navigator.language

const div = document.body.appendChild(document.createElement('div'))
const list = div.appendChild(document.createElement('ol'))

const dayNames = new Map()

for (let i = 0; i < 7; ++i) {
    const d = Temporal.PlainDate.from({
        year: Temporal.Now.plainDateISO().year,
        month: 1,
        day: i + 1,
    })

    dayNames.set(d.dayOfWeek, d.toLocaleString(LOCALE, { weekday: 'long' }))
}

for (const num of [...dayNames.keys()].sort((a, b) => a - b)) {
    list.appendChild(Object.assign(
        document.createElement('li'),
        { textContent: dayNames.get(num) },
    ))
}
```



Get Data

Sérgio Valle Duarte https://commons.wikimedia.org/wiki/File:BinaryData.jpg

Analyze Data

Lionel Rowe https://commons.wikimedia.org/wiki/File:JavaScript_code.png

Get Nice Plots

https://commons.wikimedia.org/wiki/File:Adobe_Flex_ColumnChart.png

What is Data Analysis

Get Data

https://banknxt.com/2021/12/08/tarot-cards-meanings/

What is Data Analysis

Get Data

https://banknxt.com/2021/12/08/tarot-cards-meanings/

Analyze Data

Lionel Rowe https://commons.wikimedia.org/wiki/File:JavaScript_code.png

What is Data Analysis

Get Data

https://banknxt.com/2021/12/08/tarot-cards-meanings/

Analyze Data

Lionel Rowe https://commons.wikimedia.org/wiki/File:JavaScript_code.png

Get Nice Plots

https://slashandscroll.com/when-to-use-the-celtic-cross-spread/

Explaining data variability while acknowledging the effect of random fluctuations

Explaining data variability
while acknowledging the effect of random
fluctuations

Explaining data variability
while acknowledging the effect of random
fluctuations

BELIEF

Four Levels of Statistical Belief

Four Levels of Statistical Belief

Descriptive	mean, variance, etc

Four Levels of Statistical Belief

Frequentist	parameter estimation, p-value
Descriptive	mean, variance, etc

Four Levels of Statistical Belief

Bayesian	posterior distribution
Frequentist	parameter estimation, p-value
Descriptive	mean, variance, etc

Four Levels of Statistical Belief

Causal Inference	treatment effect size
Bayesian	posterior distribution
Frequentist	parameter estimation, p-value
Descriptive	mean, variance, etc

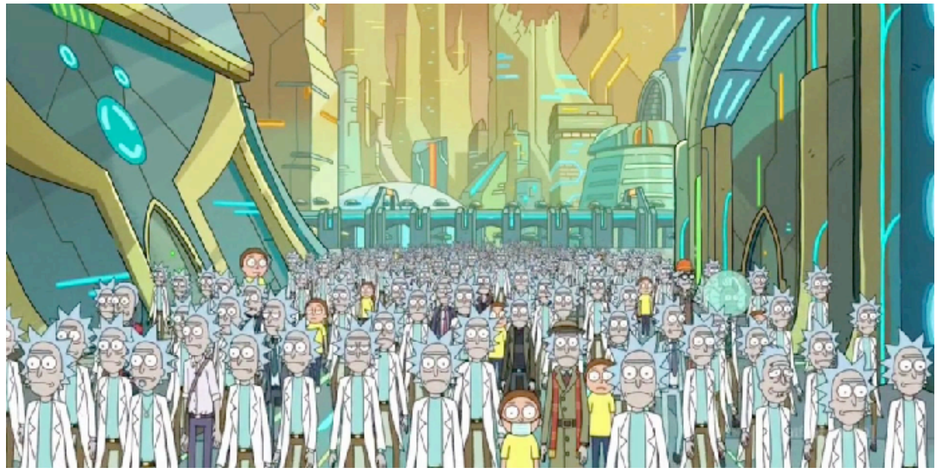
Four Levels of Statistical Belief

Causal Inference

Bayesian

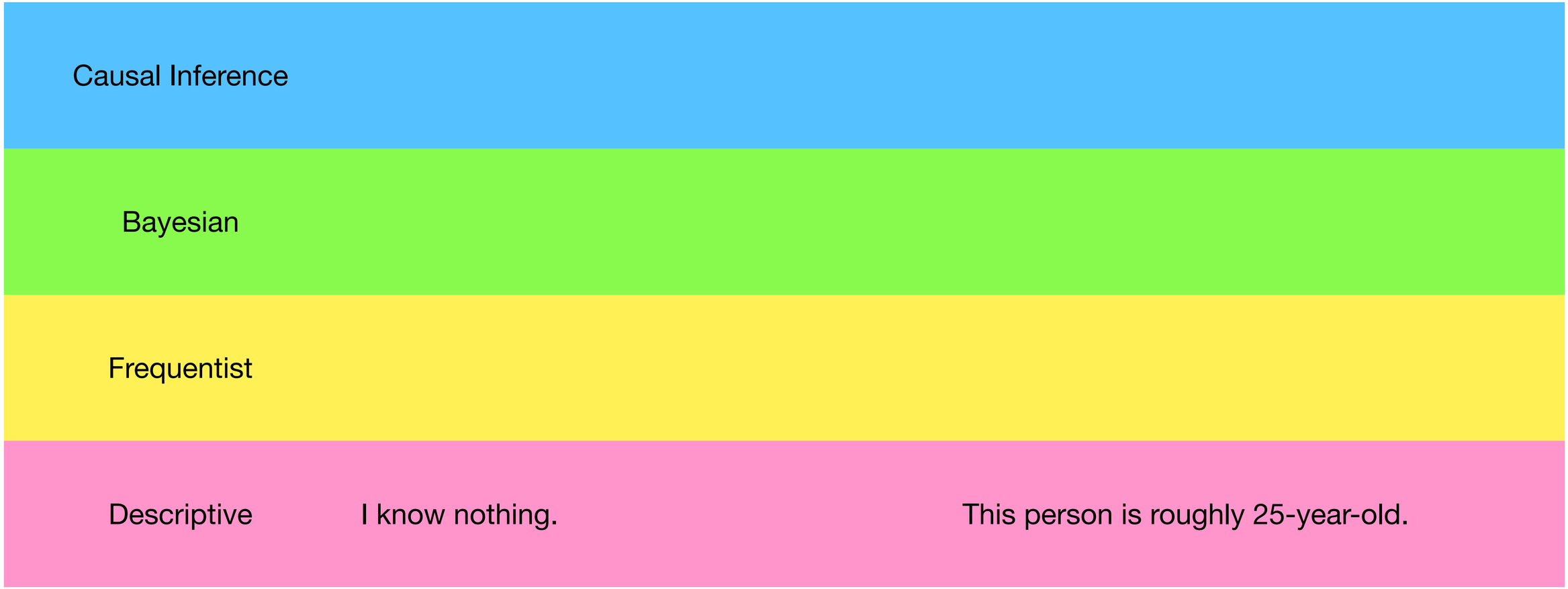
Frequentist

Descriptive



Rick? Or Morty?

Rick? Or Morty?



Rick? Or Morty?

Causal Inference		
Bayesian		
Frequentist	I know all the possible worlds.	This person is either Rick, or Morty.
Descriptive	I know nothing.	This person is roughly 25-year-old.

Rick? Or Morty?

Causal Inference		
Bayesian	I know all possible worlds, and I know the probability of each world.	This person was found in Mortytown.
Frequentist	I know all the possible worlds.	This person is either Rick, or Morty.
Descriptive	I know nothing.	This person is roughly 25-year-old.

Rick? Or Morty?

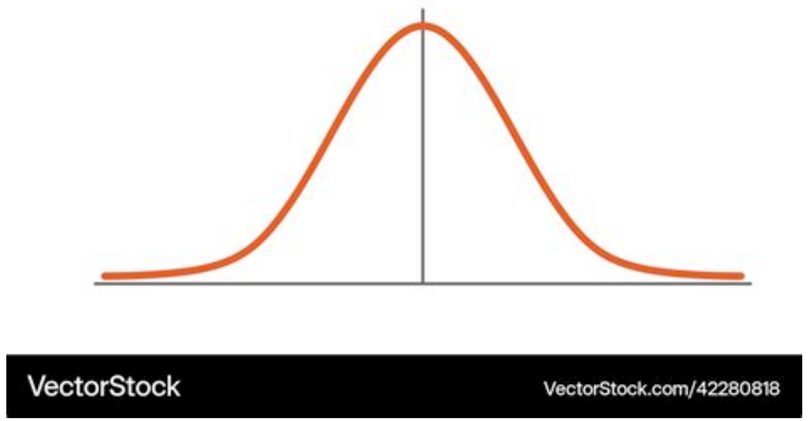
Causal Inference	I know all parallel universes, even though only one can materialize.	Will Jessica make Morty happy?
Bayesian	I know all possible worlds, and I know the probability of each world.	This person was found in Mortytown.
Frequentist	I know all the possible worlds.	This person is either Rick, or Morty.
Descriptive	I know nothing.	This person is roughly 25-year-old.

In the Citadel

- mean age = 42
- sd = 28

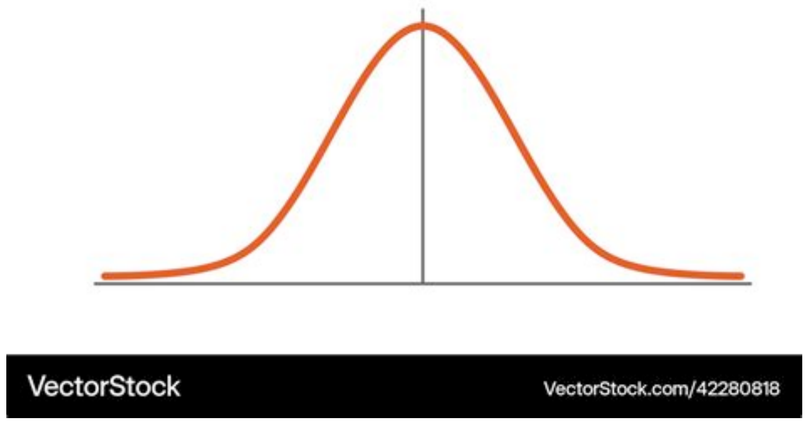
In the Citadel

- mean age = 42
- sd = 28



In the Citadel

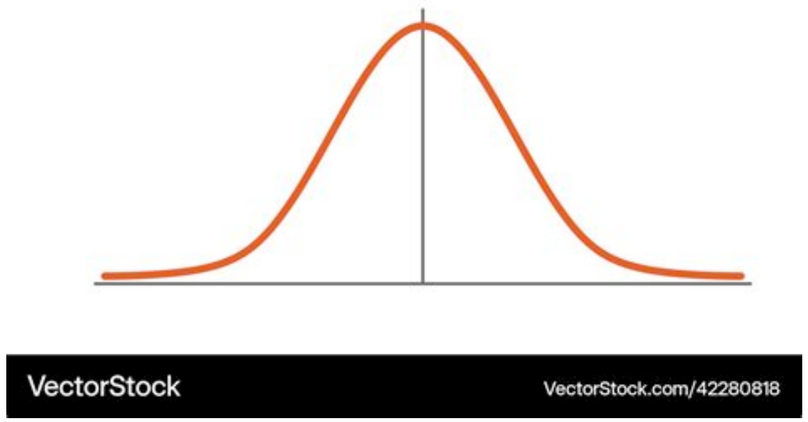
- mean age = 42
- sd = 28

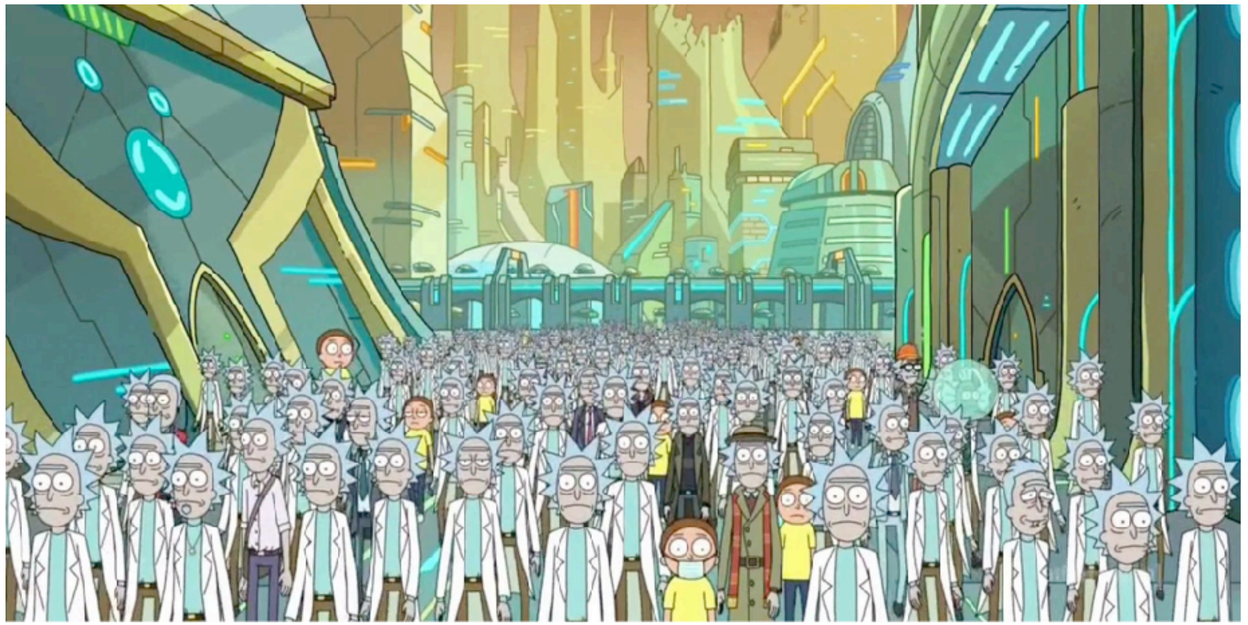


- age of 38% of the population in [28, 56]

In the Citadel

- mean age = 42
- sd = 28

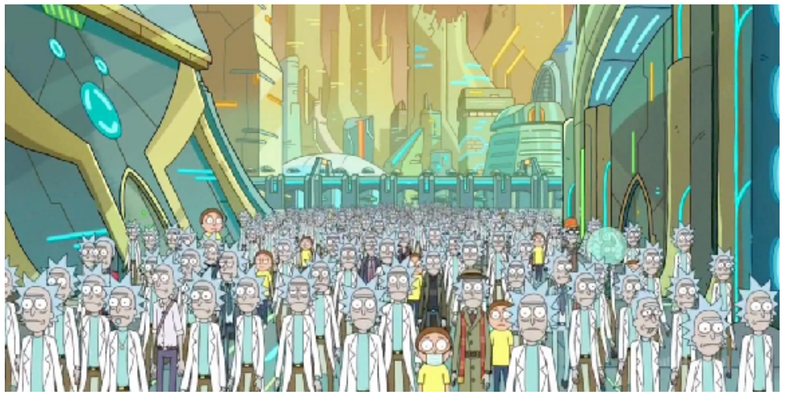




- age of 38% of the population in [28, 56]

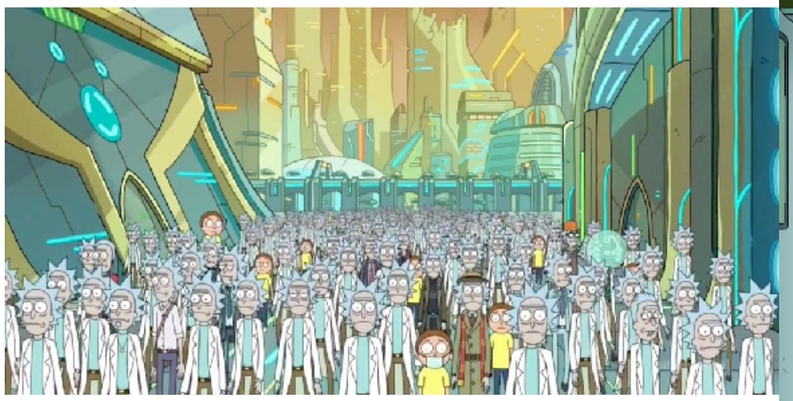
Only Descriptive Statistics Works Here

The cult of mean



mean ≠ mode

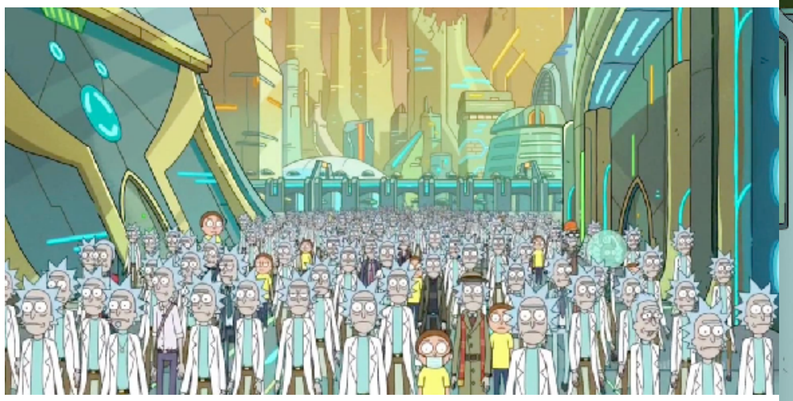
The cult of mean



mean ≠ mode

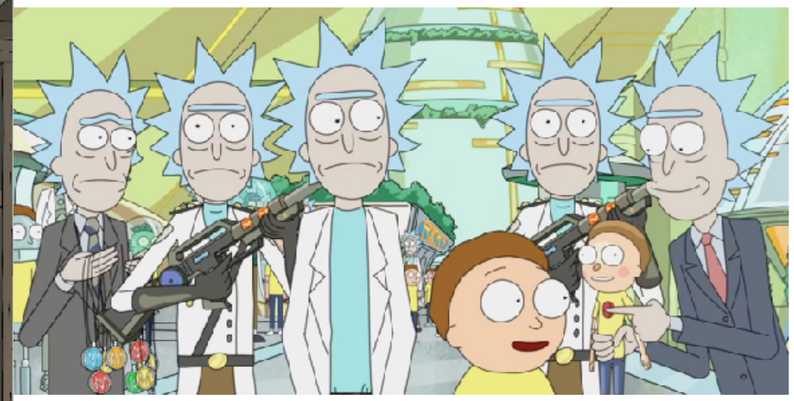
(positive + negative)/2 = 0

The cult of mean



mean ≠ mode

(positive + negative)/2 = 0



(signal + noise)/2 = weak signal

The cult of $p < 0.05$

The cult of $p < 0.05$

p-value = probability of positive result if it is actually wrong

The cult of p < 0.05

p-value = probability of positive result if it is actually wrong

Assumptions

The cult of p < 0.05

p-value = probability of positive result if it is actually wrong

Assumptions

A non-Rick is not necessarily a Morty

The cult of p < 0.05

p-value = probability of positive result if it is actually wrong

Assumptions	Number of Tests
A non-Rick is not necessarily a Morty	

The cult of p < 0.05

p-value = probability of positive result if it is actually wrong

Assumptions	Number of Tests
A non-Rick is not necessarily a Morty	For 100 wrong claims, you get 5 positive results.

The cult of p < 0.05

p-value = probability of positive result if it is actually wrong

Assumptions	Number of Tests
A non-Rick is not necessarily a Morty	For 100 wrong claims, you get 5 positive results. Multiple hypothesis testing / FDR control comes at a price.

The cult of p < 0.05

p-value = probability of positive result if it is actually wrong

Assumptions	Number of Tests	Effect Size
A non-Rick is not necessarily a Morty	For 100 wrong claims, you get 5 positive results. Multiple hypothesis testing / FDR control comes at a price.	

The cult of $p < 0.05$

p-value = probability of positive result if it is actually wrong

Assumptions	Number of Tests	Effect Size
A non-Rick is not necessarily a Morty	For 100 wrong claims, you get 5 positive results. Multiple hypothesis testing / FDR control comes at a price.	0.0000000001 > 0, so?

Explaining data variability
while acknowledging the effect of random
fluctuation

BELIEF

ATOM

[Wasserstein, Schirm, Lazar, 2019]

ATOM

[Wasserstein, Schirm, Lazar, 2019]

- Be **Thoughtful** about your belief.

ATOM

[Wasserstein, Schirm, Lazar, 2019]

- Be **Thoughtful** about your belief.
- Be **Open** about your belief.

ATOM

[Wasserstein, Schirm, Lazar, 2019]

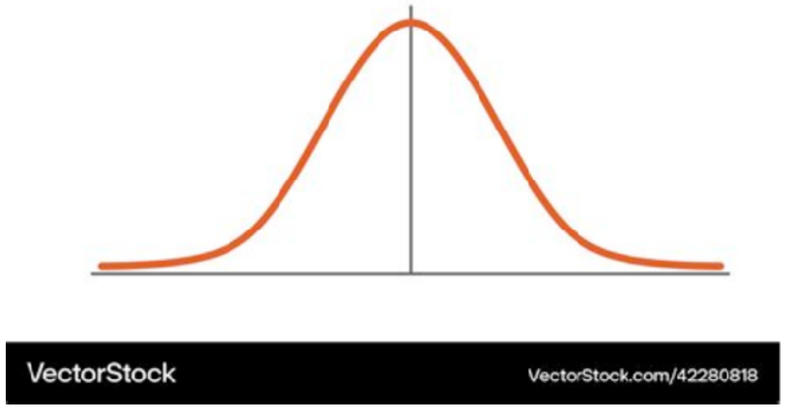
- Be **Thoughtful** about your belief.
- Be **Open** about your belief.
- Be **Modest** about your results.

ATOM

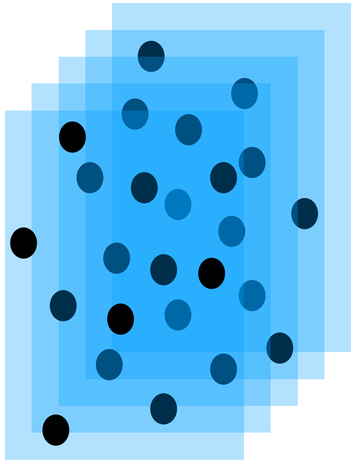
[Wasserstein, Schirm, Lazar, 2019]

- **Accept** randomness.
- Be **Thoughtful** about your belief.
- Be **Open** about your belief.
- Be **Modest** about your results.

How to Gauge Randomness

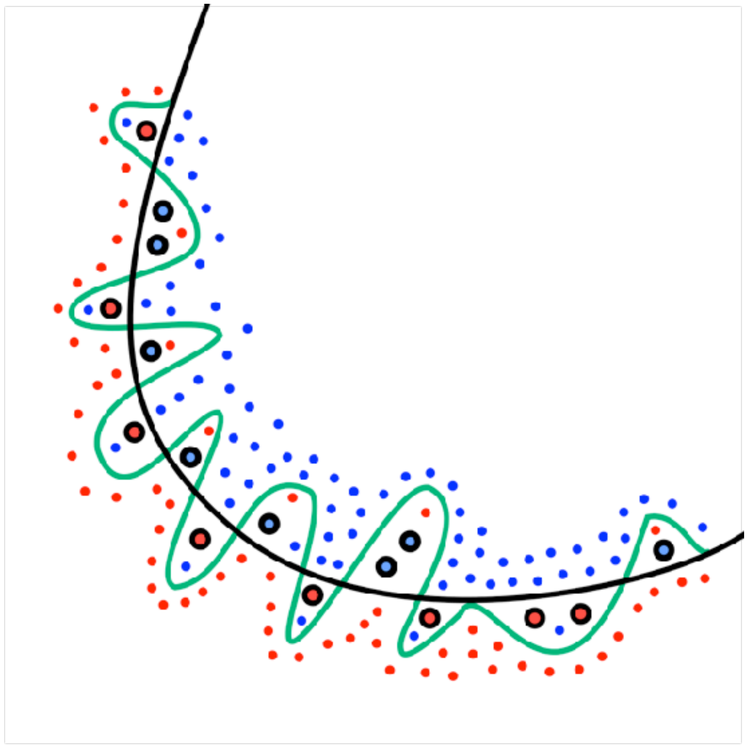


Normal Approximation



Bootstrap

Lionel Rowe https://commons.wikimedia.org/wiki/File:JavaScript_code.png

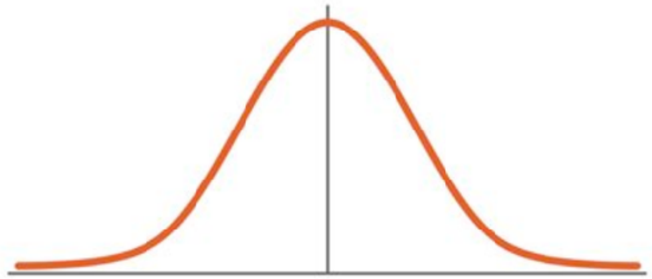


Cross Validation

Chabacano https://commons.wikimedia.org/wiki/File:Overfitting.svg

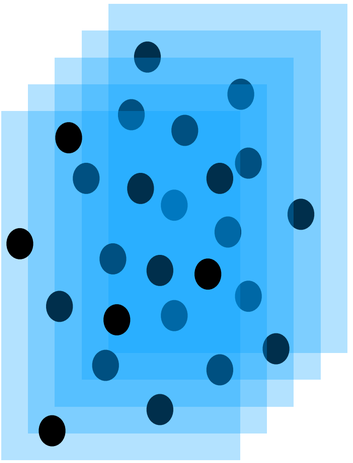
Normal Approximation

- when you compute mean and standard deviation and nothing else



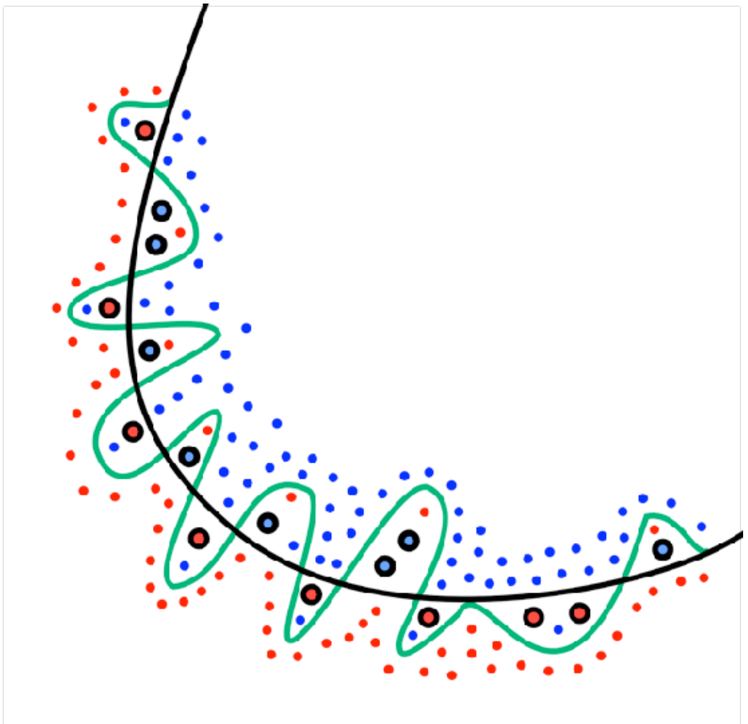
Bootstrap / Subsampling

- Bootstrap
  - Generate many **same-sized** samples by **re**sampling from the data set
- Subsampling
  - Generate many small **sub**samples
- Look at the resultant distribution of the sample means



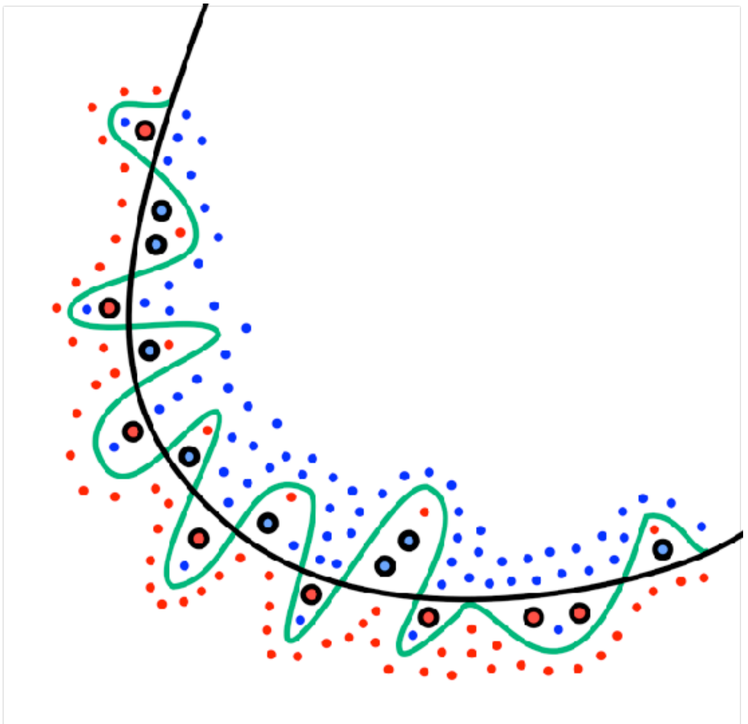
Cross-Validation

- Maybe this is neither Rick nor Morty

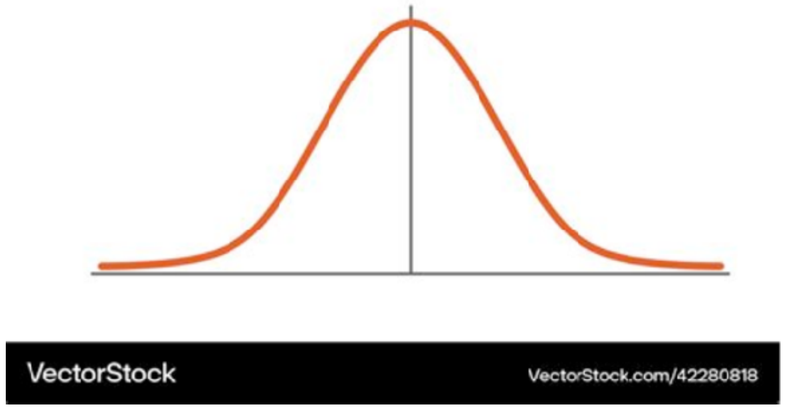


Cross-Validation

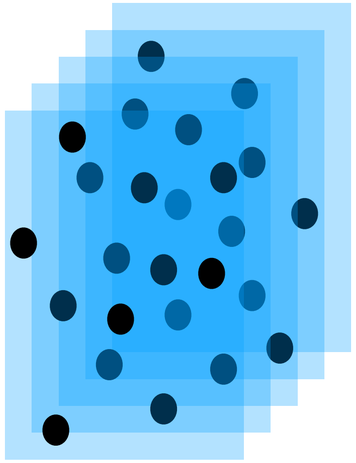
- Maybe this is neither Rick nor Morty
- But all models are wrong…



How to Gauge Randomness

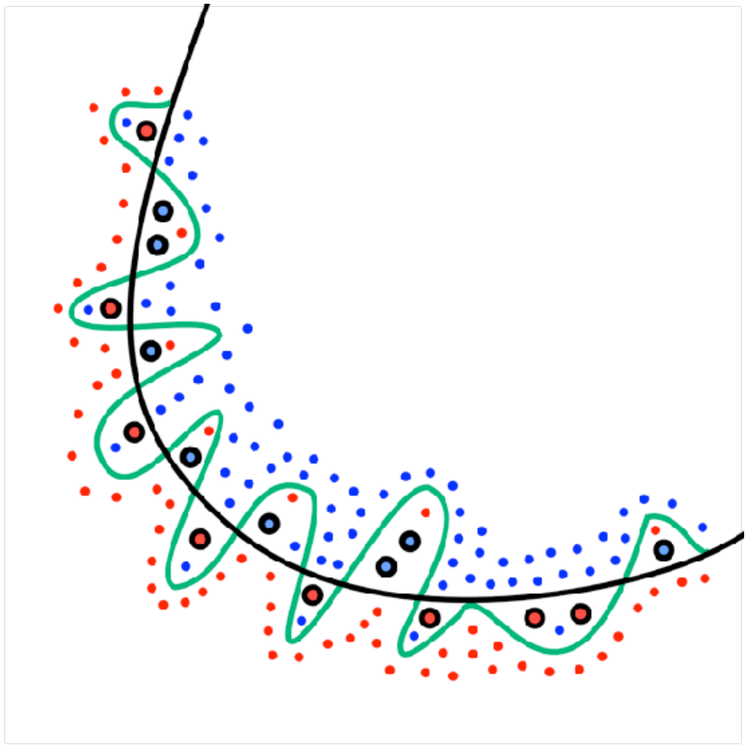


Normal Approximation



Bootstrap

Lionel Rowe https://commons.wikimedia.org/wiki/File:JavaScript_code.png



Cross Validation

Chabacano https://commons.wikimedia.org/wiki/File:Overfitting.svg

Dimension Reduction

Dimension Reduction

Dimension Reduction

Enchanted, 2007

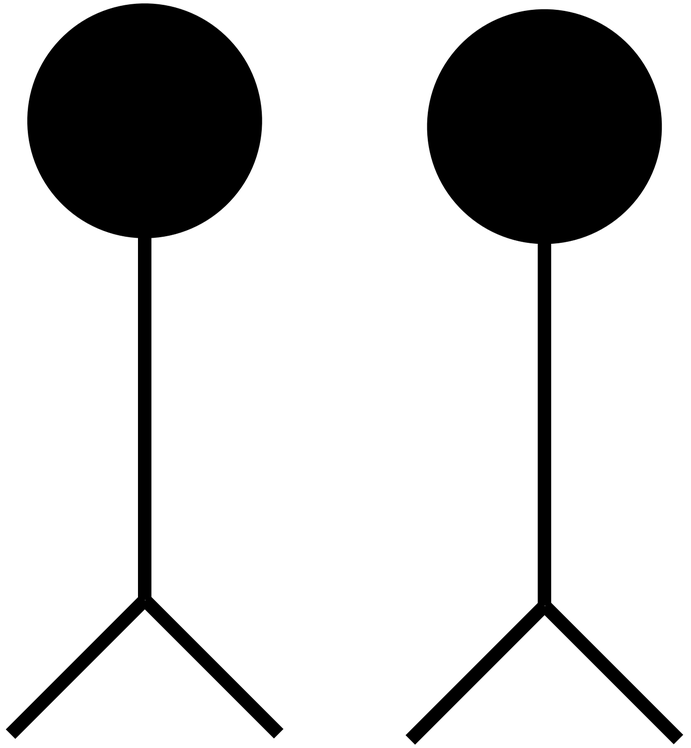
Dimension Reduction

Walt Disney Pictures / Barry Wetcher

Enchanted, 2007

Input, Output

Enchanted, 2007



Different Methods

Different Methods

Principal Component Analysis	Preserve linear information
Multi-Dimensional Scaling	
UMAP	
Mapper	

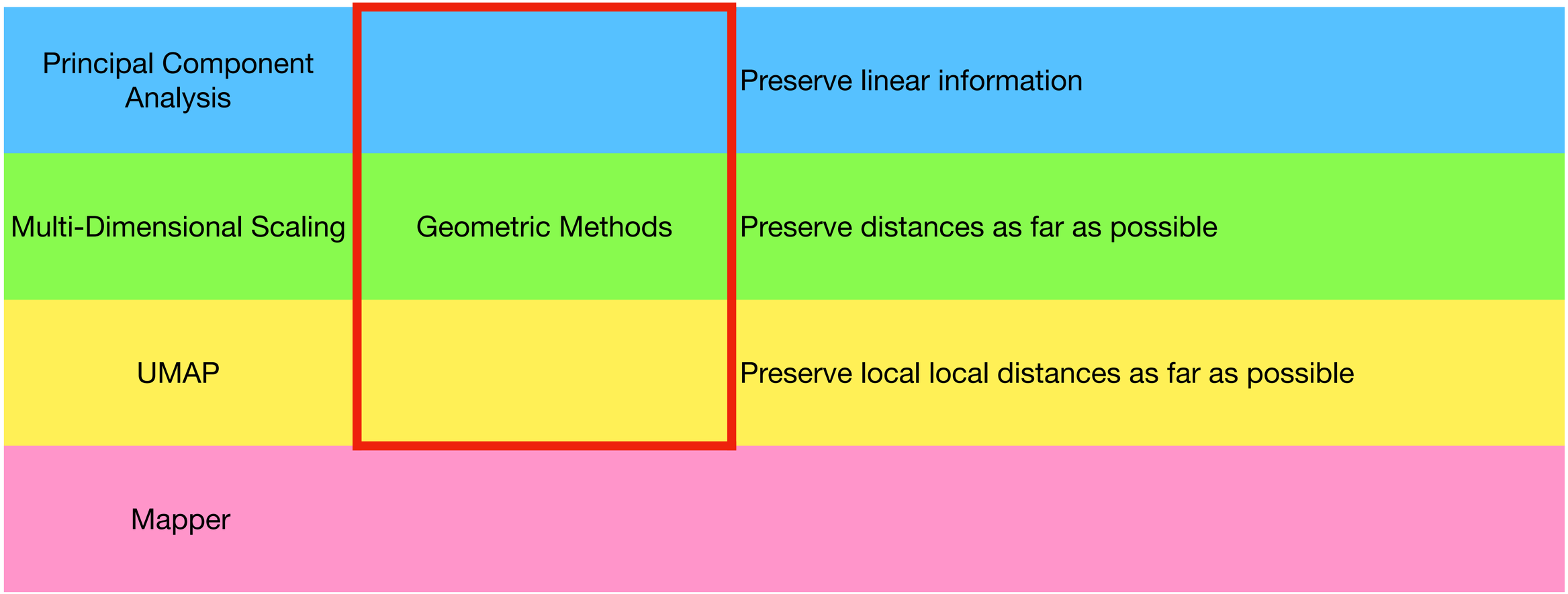
Different Methods

Principal Component Analysis	Preserve linear information
Multi-Dimensional Scaling	Preserve distances as far as possible
UMAP	
Mapper	

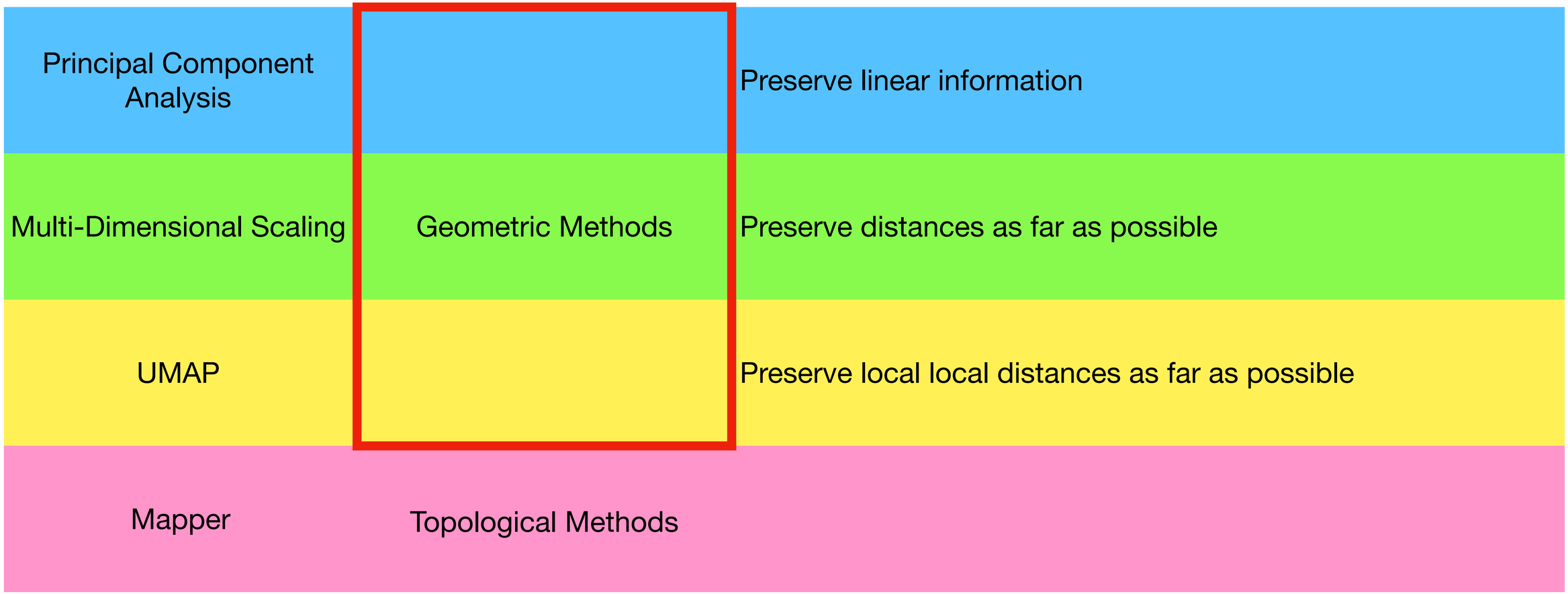
Different Methods

Principal Component Analysis	Preserve linear information
Multi-Dimensional Scaling	Preserve distances as far as possible
UMAP	Preserve local local distances as far as possible
Mapper	

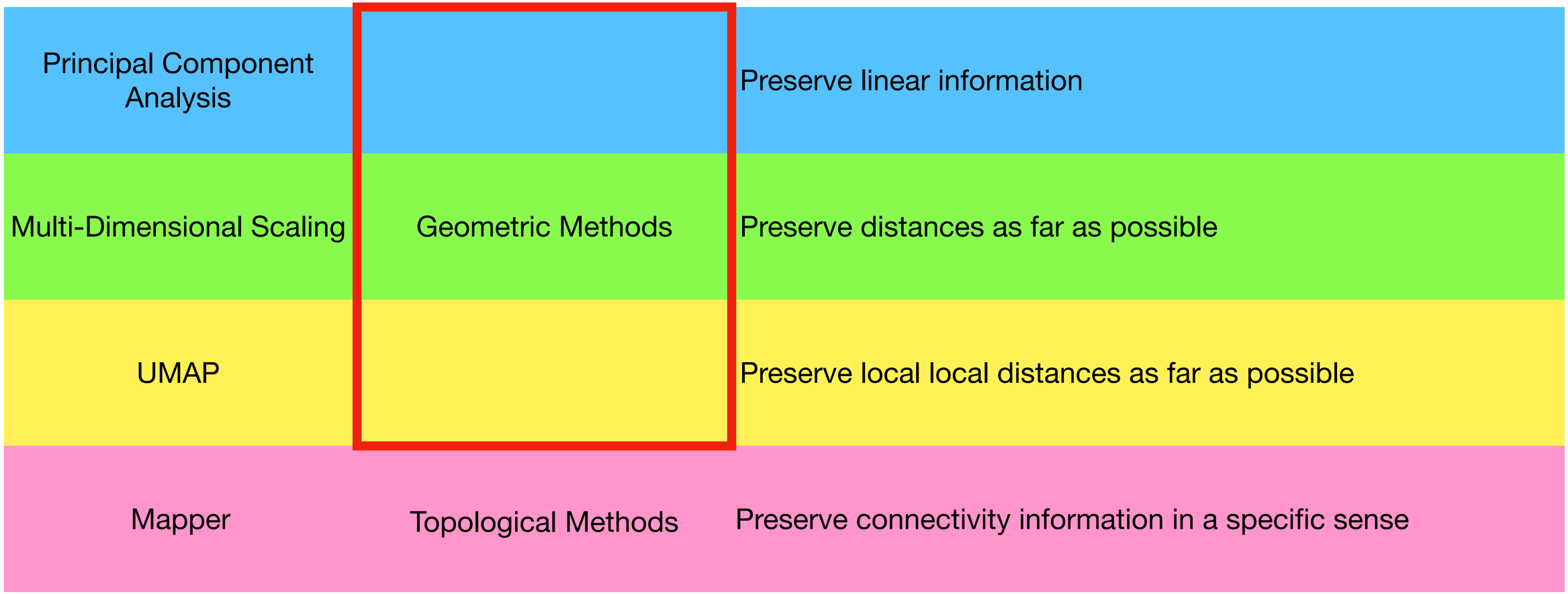
Different Methods



Different Methods



Different Methods



As good as possible…

As good as possible…

- Nearby points in high-dim remains nearby in low-dim.
- Reverse?

As good as possible…

- Nearby points in high-dim remains nearby in low-dim.
- Reverse?

Cinderella III: A Twist in Time (2017)

As good as possible…

- Nearby points in high-dim remains nearby in low-dim.
- Reverse needs cheating

Alice in the Wonderland (1951)

Mapper

- Nearby points in high-dim remains nearby in low-dim.
- Reverse guaranteed by using graph topology to transcend Euclidean geometry

Alice in the Wonderland (1951)

Mapper

- Nearby points in high-dim remains nearby in low-dim.
- Reverse guaranteed by using graph topology to transcend Euclidean geometry
- Price: loss of metric control

Alice in the Wonderland (1951)

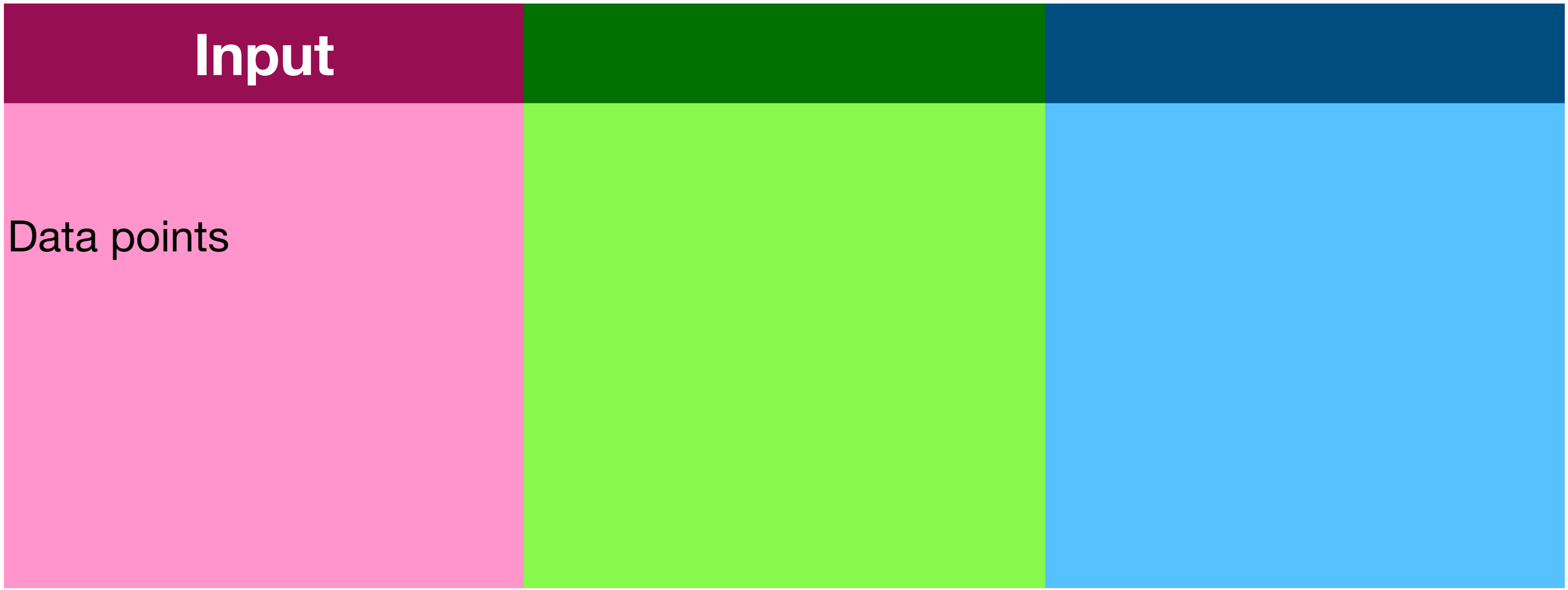
Mapper

- Nearby points in high-dim remains nearby in low-dim.
- Reverse guaranteed by using graph topology to transcend Euclidean geometry
- Price: loss of metric control
- Get to keep: connectivity

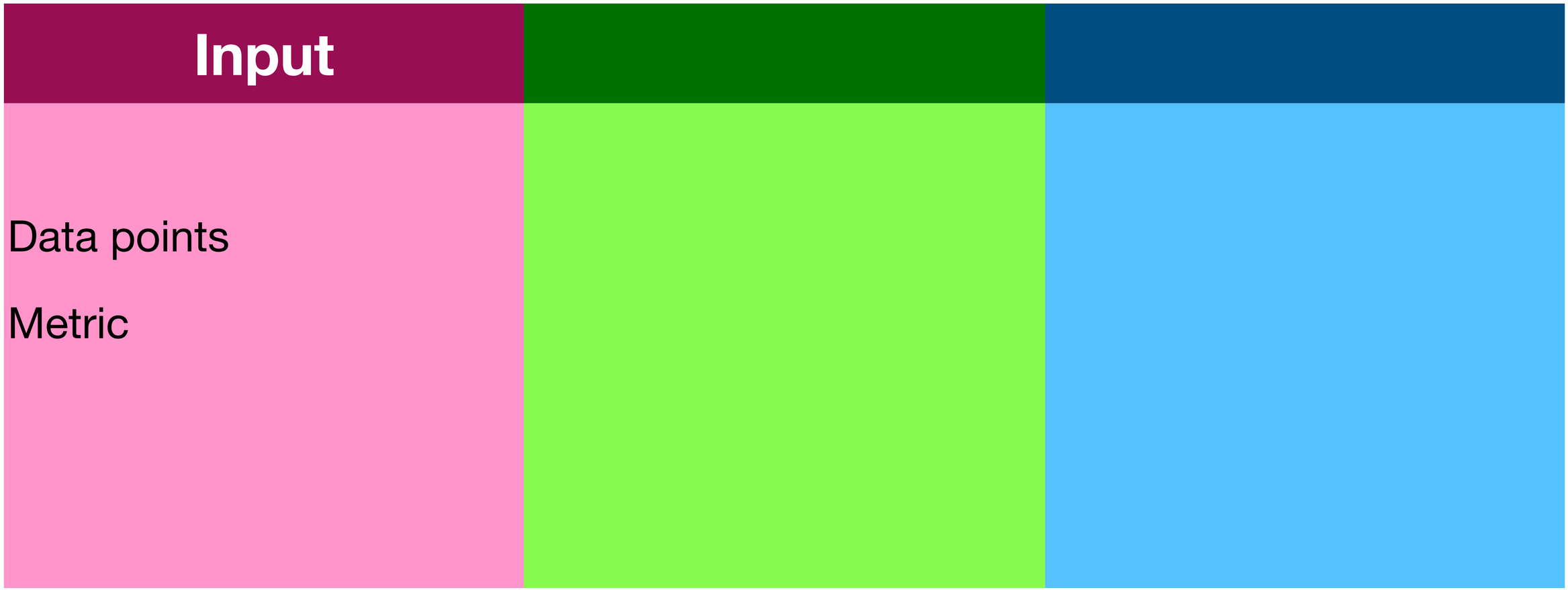
Alice in the Wonderland (1951)

What is Mapper?

What is Mapper?



What is Mapper?



What is Mapper?

Input		
Data points Metric Coloring scheme (optional)		

What is Mapper?

Input		Output
Data points Metric Coloring scheme (optional)		A (colored) graph, where

What is Mapper?

Input		Output
Data points Metric Coloring scheme (optional)		A (colored) graph, where nodes are clusters of data points

What is Mapper?

Input		Output
Data points Metric Coloring scheme (optional)		A (colored) graph, where nodes are clusters of data points edges are nonempty intersection of clusters

What is Mapper?

Input	Parameters	Output
Data points Metric Coloring scheme (optional)	Discretization parameters	A (colored) graph, where nodes are clusters of data points edges are nonempty intersection of clusters

What is Mapper?

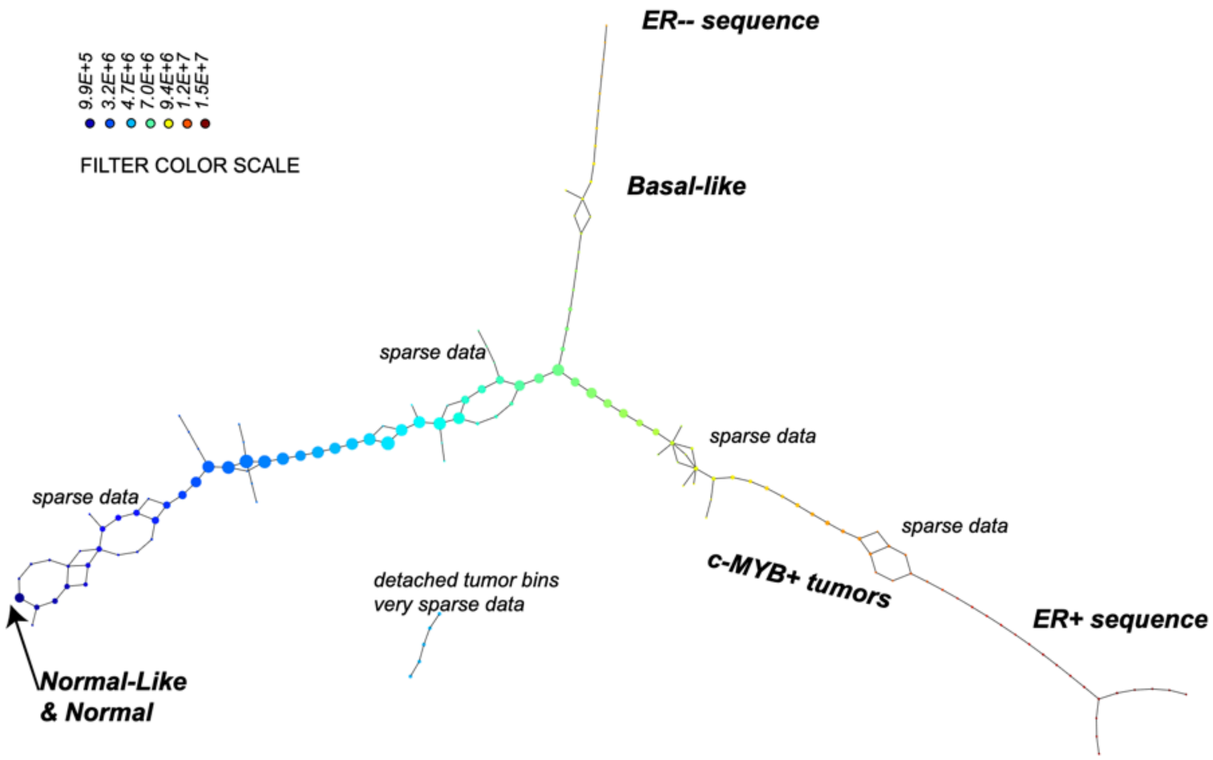
Input	Parameters	Output
Data points Metric Coloring scheme (optional)	Discretization parameters Lens (if lens-based)	A (colored) graph, where nodes are clusters of data points edges are nonempty intersection of clusters

What is Mapper?

Input	Parameters	Output
Data points Metric Coloring scheme (optional)	Discretization parameters Lens (if lens-based) Subroutine parameters	A (colored) graph, where nodes are clusters of data points edges are nonempty intersection of clusters

Nicolau et al, 2011

- data: Transcriptional microarray data
- sample: 295 tumors
- features: 262 genes
- lens: Normal component to the linear subspace of healthy tissues



What is Mapper Good For?

proximity guarantee, and…?

What is Mapper Good For?

proximity guarantee, and…?

- Distinguish points that are indistinguishable through the lens

What is Mapper Good For?

proximity guarantee, and…?

- Distinguish points that are indistinguishable through the lens
- Good for sub-typing

What is Mapper Good For?

proximity guarantee, and…?

- Distinguish points that are indistinguishable through the lens
- Good for sub-typing
- Finding out relationship between them

Limitations of Mapper

Limitations of Mapper

- cannot retain metric information

Limitations of Mapper

- cannot retain metric information
- choices of lens

Limitations of Mapper

- cannot retain metric information
- choices of lens
- parameter tuning

Limitations of Mapper

- cannot retain metric information
- choices of lens
- parameter tuning
- difficulty of clustering

Intro to Topological Data Analysis

Statistics and Dimension Reduction

Chunyin Siu, Feb 14, 2025