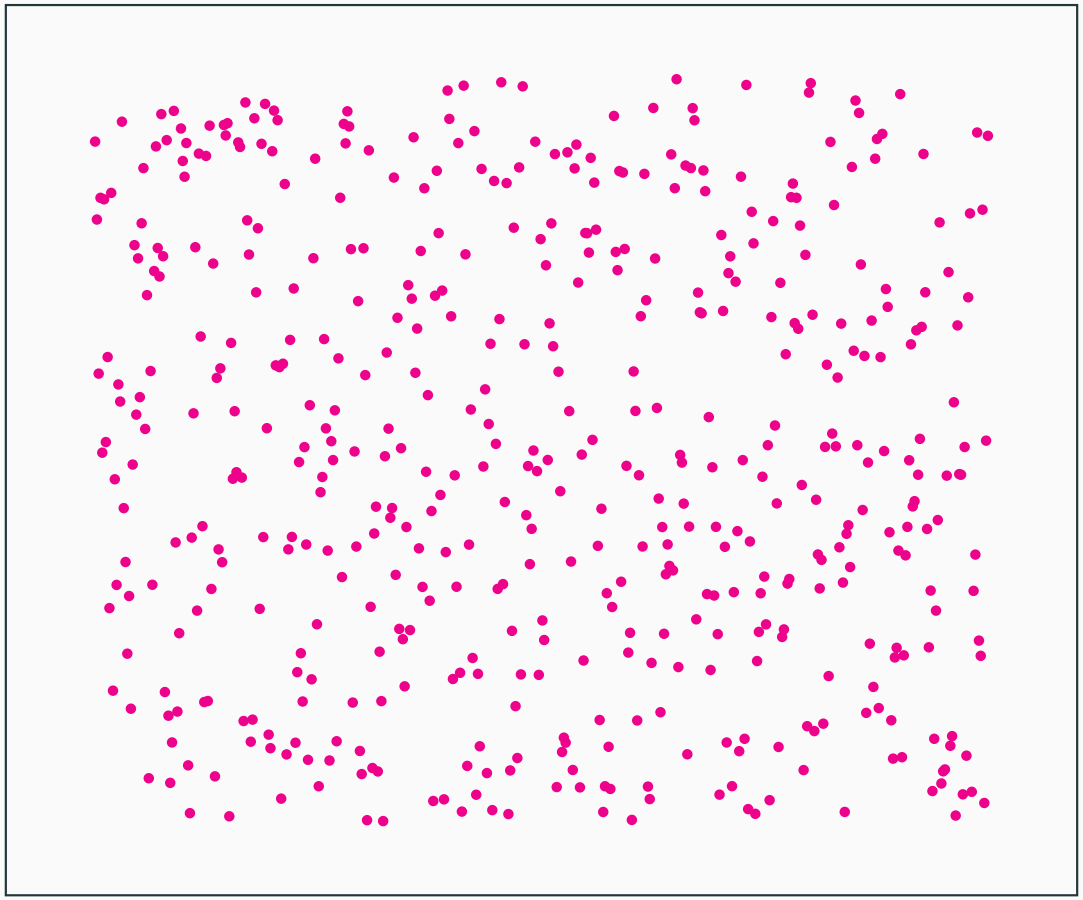
The Shape of Data

Roderic Guigó Corominas

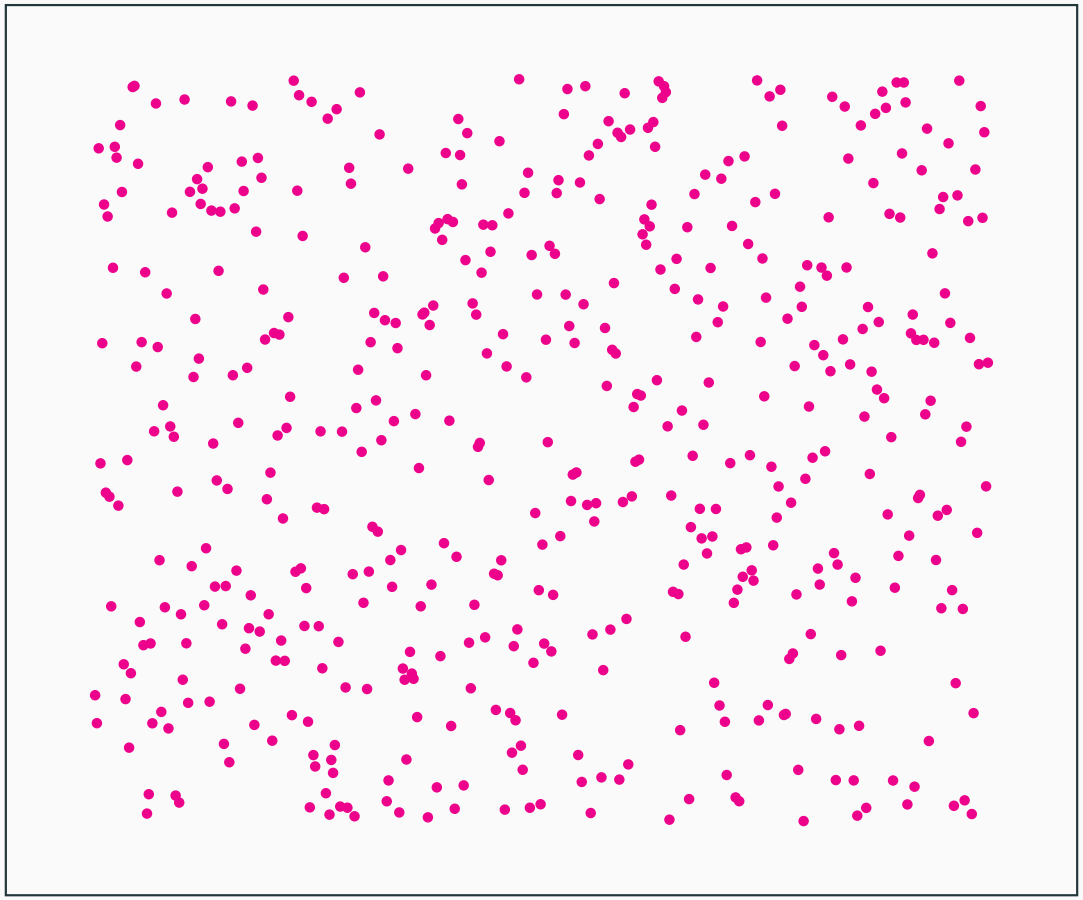
August 12th 2019

SMTB 2019

Are there any patterns in this dataset?

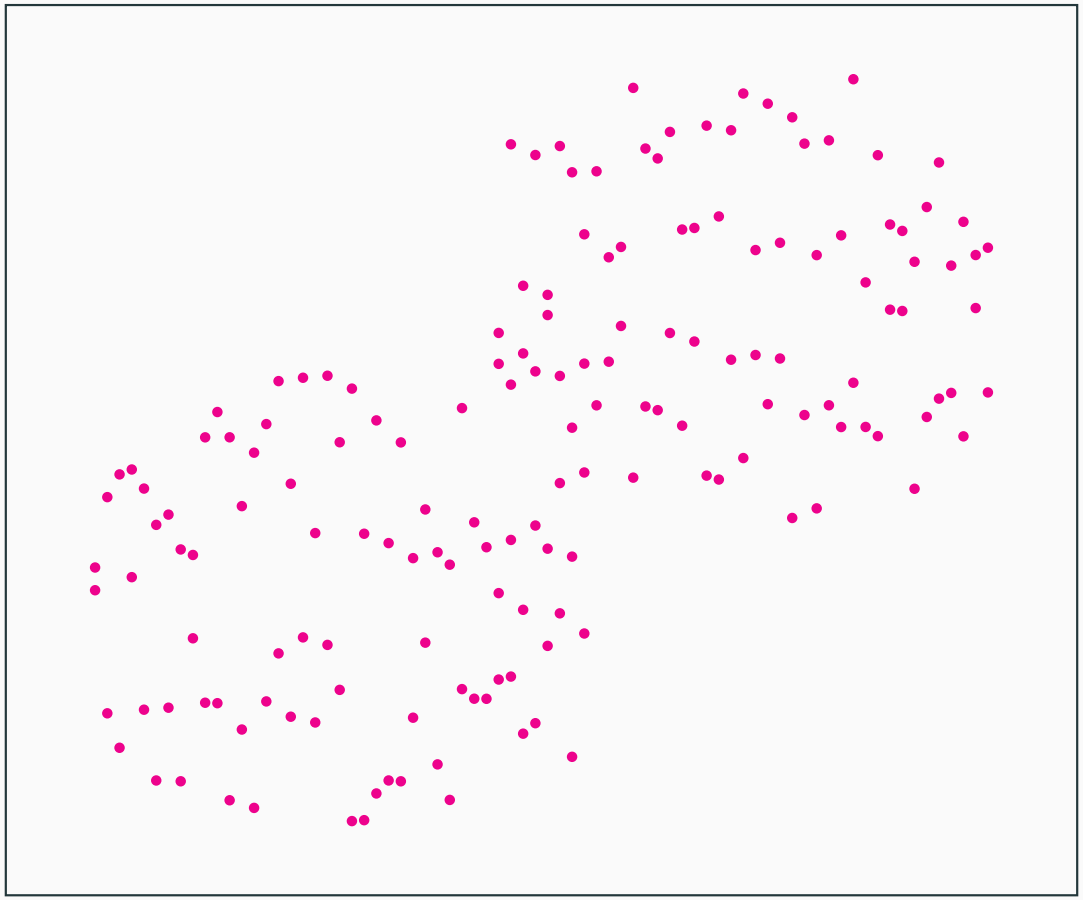


Are there any patterns in this dataset?

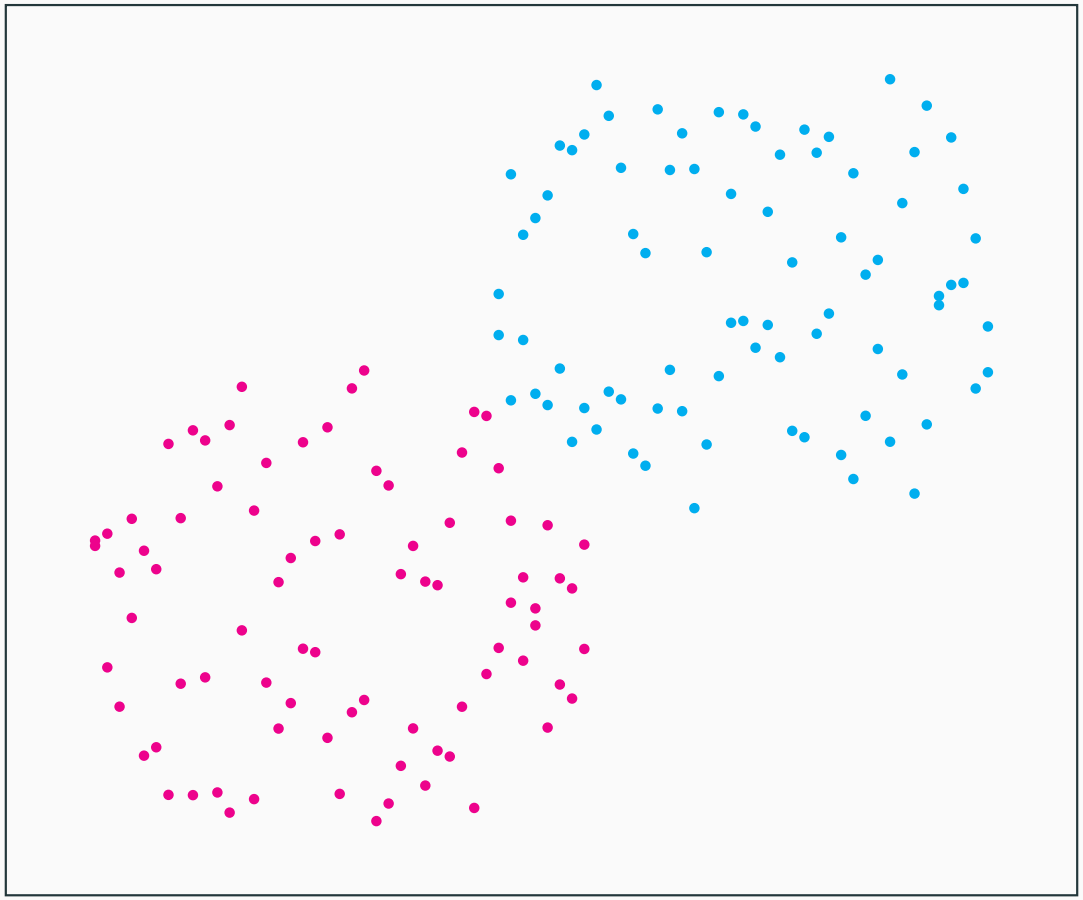


No apparent pattern.

How about this one?

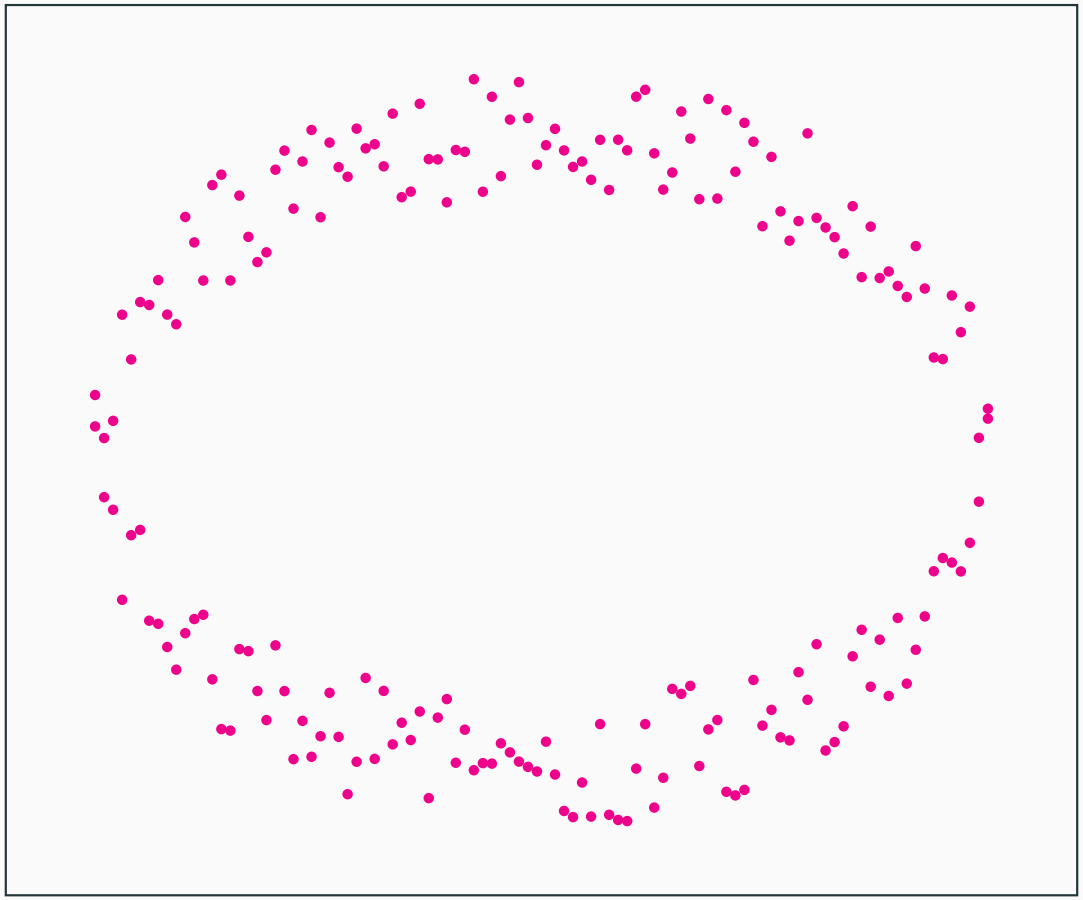


How about this one?

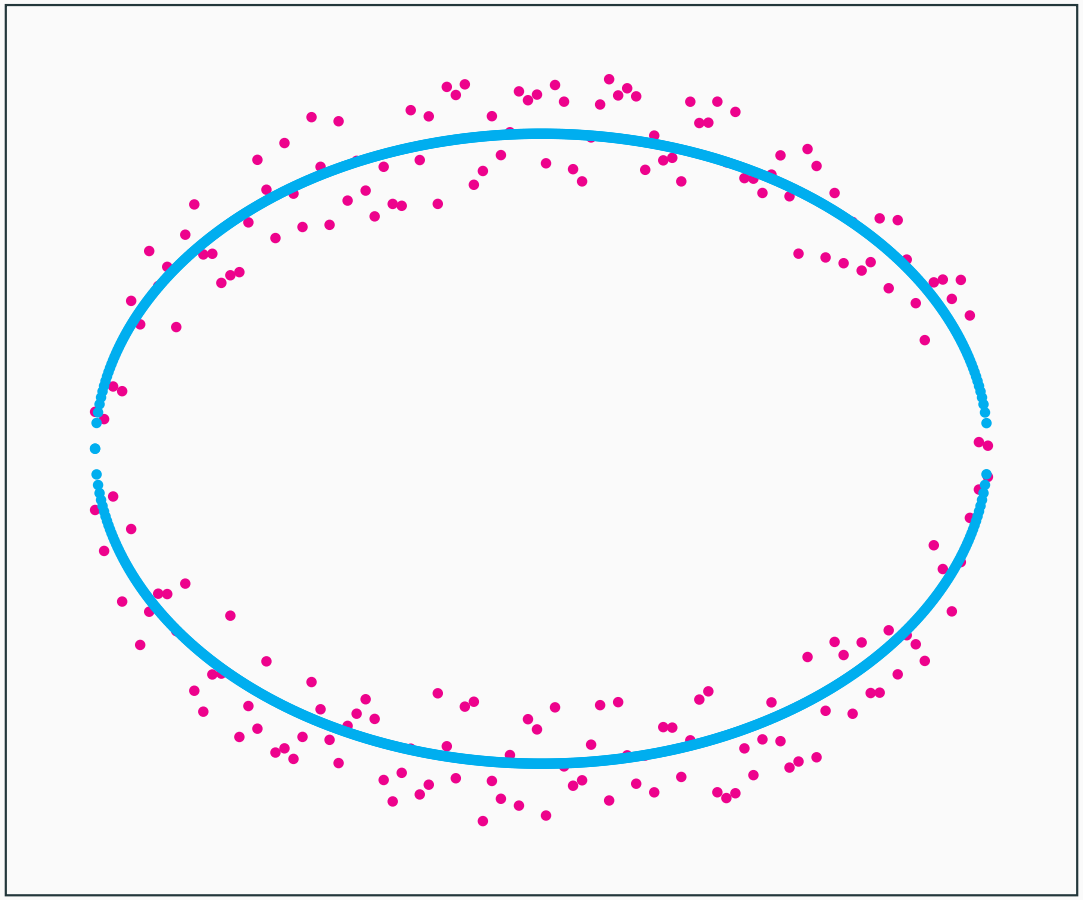


Two distinguished groups of data.

And this one?



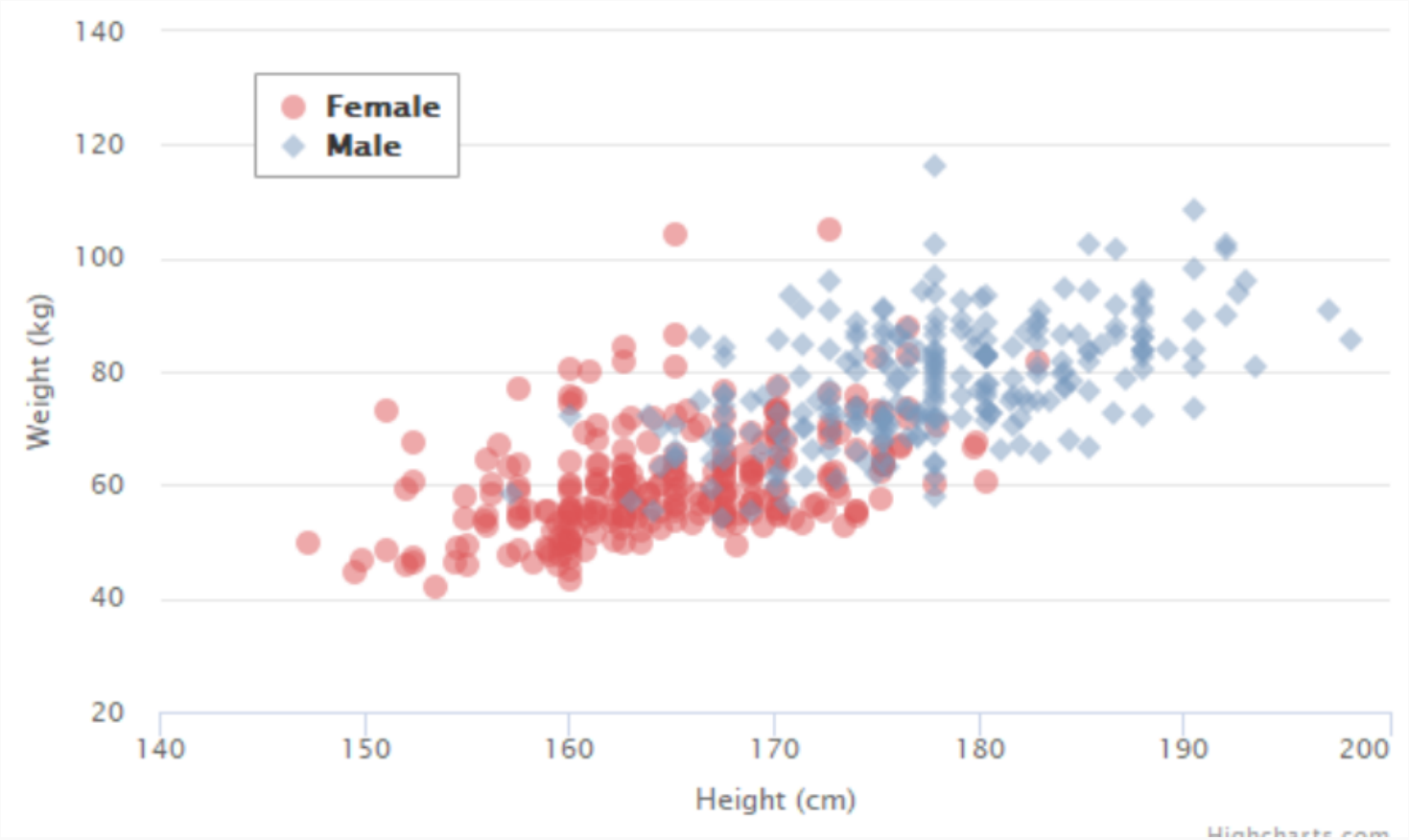
And this one?



The data is approximately distributed in a circle.

The shape of data

Example: height and weight of a group of people. Every person is represented by two numbers: $(185\text{cm}, 80\text{kg})$. We can plot these numbers!



The shape of data

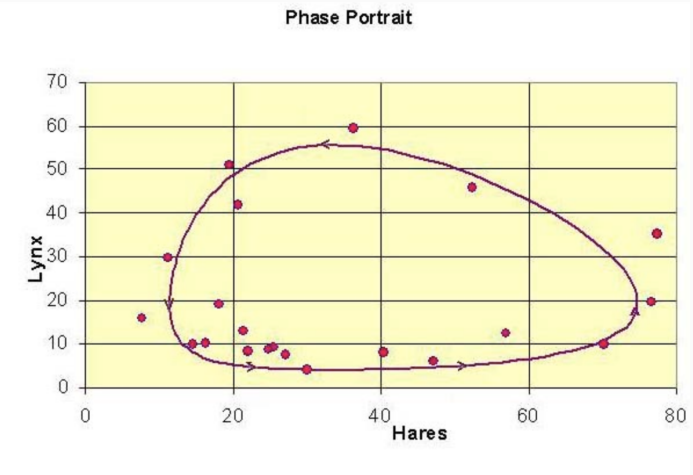
Example: the Lotka-Voterra equations describe the dynamics of an environment in which two species, a predator and a prey, interact. Data of Lynx and Hare pelts from 1900 to 1920 collected by the Hudson Bay company.

Year	Hares (x1000)	Lynx (x1000)	Year	Hares (x1000)	Lynx (x1000)
1900	30	4	1911	40.3	8
1901	47.2	6.1	1912	57	12.3
1902	70.2	9.8	1913	76.6	19.5
1903	77.4	35.2	1914	52.3	45.7
1904	36.3	59.4	1915	19.5	51.1
1905	20.6	41.7	1916	11.2	29.7
1906	18.1	19	1917	7.6	15.8
1907	21.4	13	1918	14.6	9.7
1908	22	8.3	1919	16.2	10.1
1909	25.4	9.1	1920	24.7	8.6
1910	27.1	7.4			

The shape of data

Example: the Lotka-Voterra equations describe the dynamics of an environment in which two species, a predator and a prey, interact. Data of Lynx and Hare pelts from 1900 to 1920 collected by the Hudson Bay company.

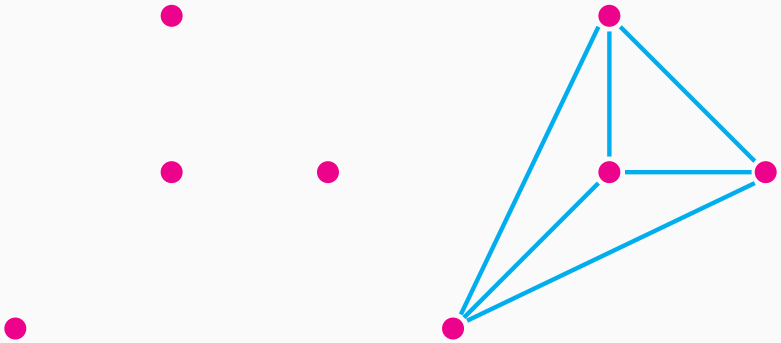
Year	Hares (x1000)	Lynx (x1000)	Year	Hares (x1000)	Lynx (x1000)
1900	30	4	1911	40.3	8
1901	47.2	6.1	1912	57	12.3
1902	70.2	9.8	1913	76.6	19.5
1903	77.4	35.2	1914	52.3	45.7
1904	36.3	59.4	1915	19.5	51.1
1905	20.6	41.7	1916	11.2	29.7
1906	18.1	19	1917	7.6	15.8
1907	21.4	13	1918	14.6	9.7
1908	22	8.3	1919	16.2	10.1
1909	25.4	9.1	1920	24.7	8.6
1910	27.1	7.4			



Shape matters!!!

Well, sometimes, at least...

Topological Data Analysis (TDA)



Data $\Rightarrow$ Topology $\Rightarrow$ Properties $\Rightarrow$ Meaning

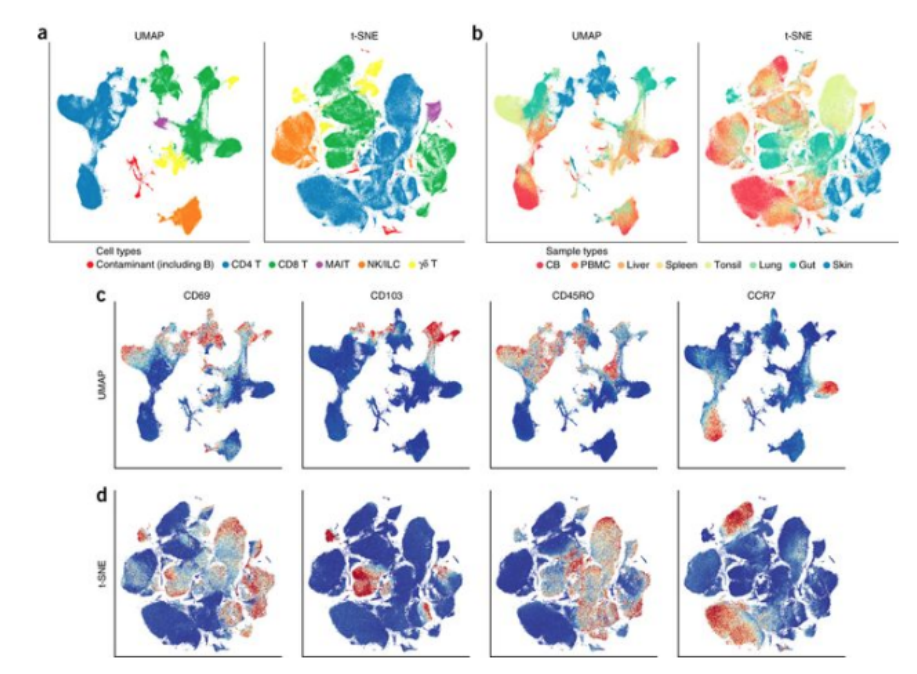
Existing Applications

There are 2.5 quintillions of bytes of data created each day (in 2018).

- Hockey Analytics
- Breast Cancer
- Biomolecular Data and Protein Flexibility
- Image Processing
- ...

Buzzwords

- Image processing
- Persitent homology
- Clustering
- Dimension reduction



Ideal world: combine powerful statistical methods (machine learning) with problem specific knowledge (TDA).

Aims & Interests

In this course you will learn:

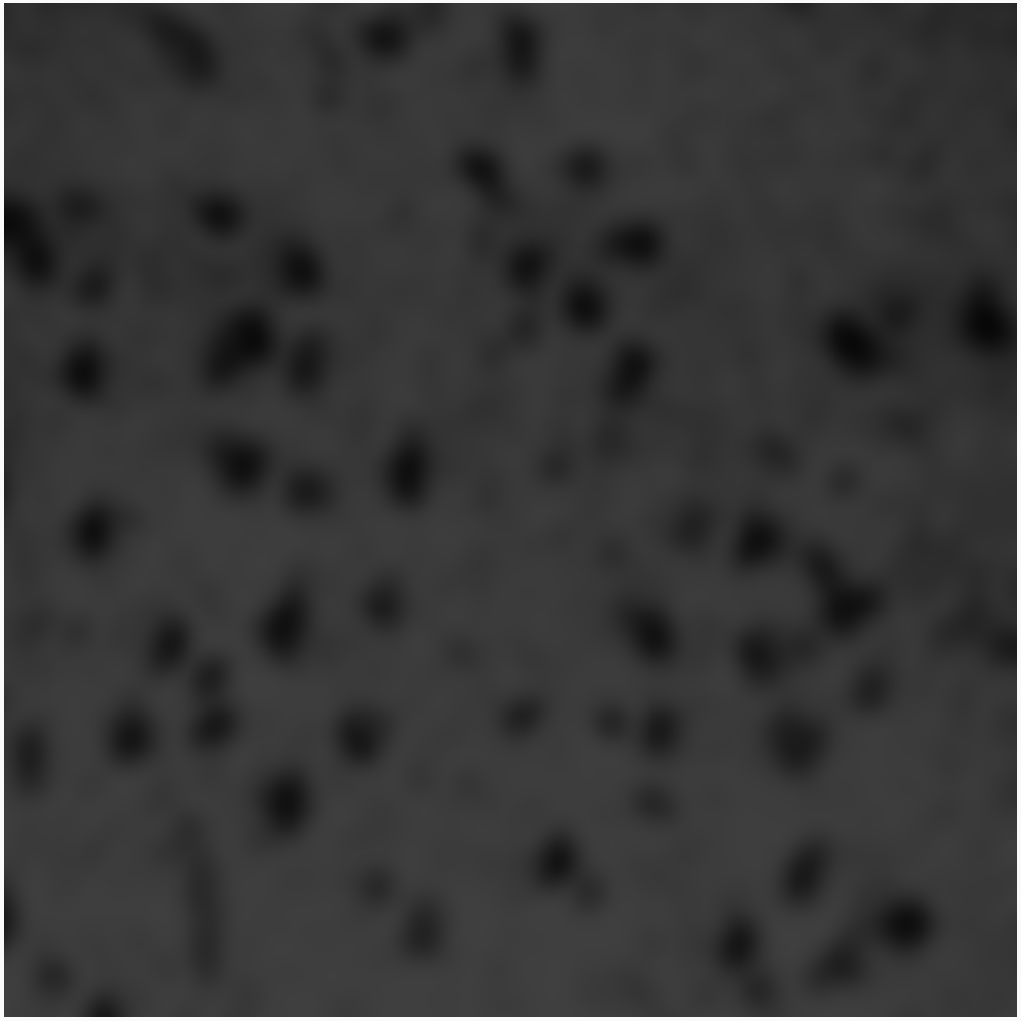
- Mathematics of shape: topology.
- Code in Python.
- Use TDA specific external Python libraries.
- Maybe some statistics.
- Applications to the field of biology.

Sign up for this course if: you enjoy mathematics and programming, with a biological flavour.

Course Plan

- Day 1: Math Theory, Basics of Topology
- Day 2: Math Theory, Simplices
- Day 3: Math Theory, Persistent Homology
- Day 4: Intro to Python & Numpy
- Day 5: Python applications, image processing
- Day 6: Python applications, Bio
- Day 7: Python applications, Bio

How many cells do you see?



How many cells do you see?

How many cells do you see?

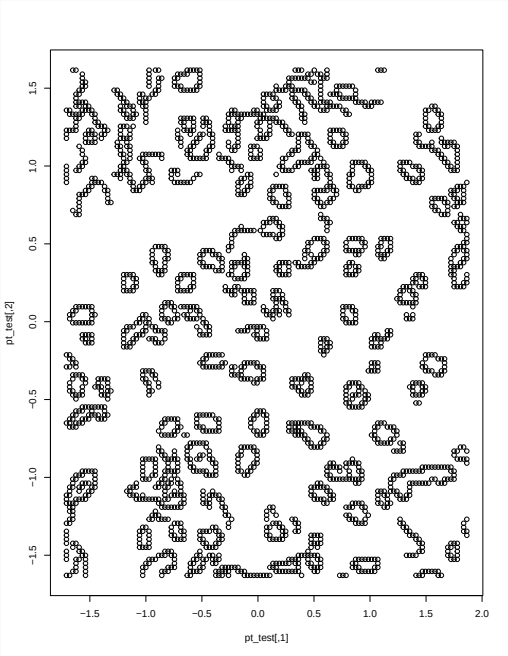
- Rough count: 100 cells, 5 seconds.

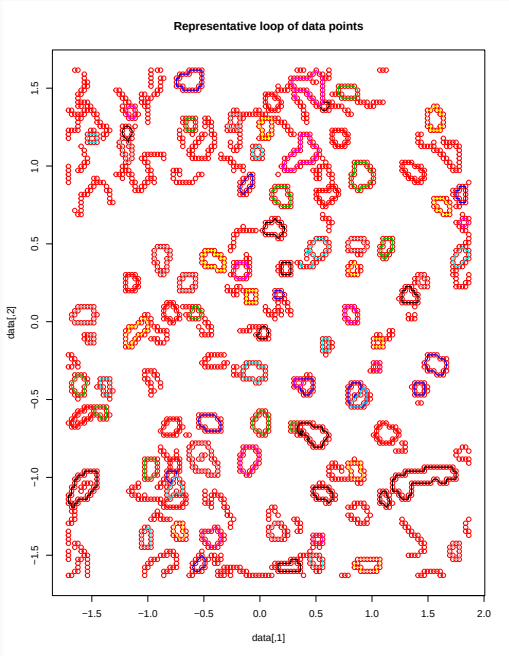
How many cells do you see?

- Rough count: 100 cells, 5 seconds.
- Careful count: 70 cells, 60 seconds.

How many cells do you see?

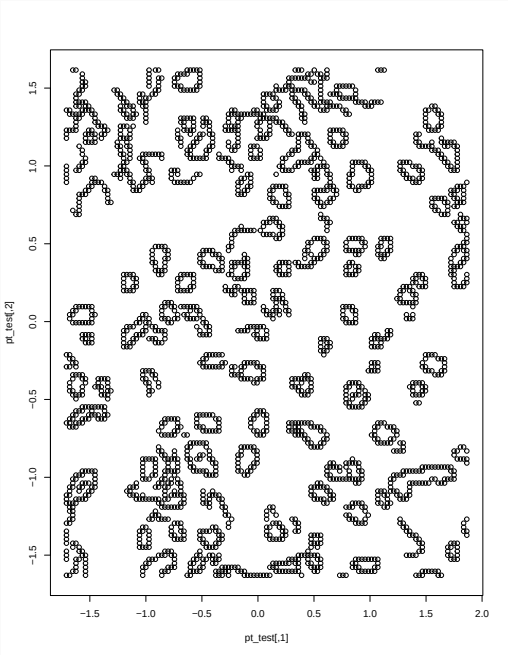
- Rough count: 100 cells, 5 seconds.
- Careful count: 70 cells, 60 seconds.
- Computer count: 68 cells, $\ll 1$ second.

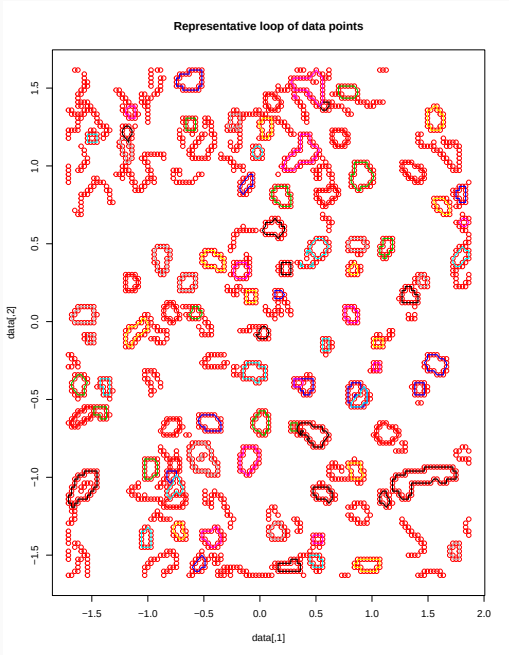




How many cells do you see?

- Rough count: 100 cells, 5 seconds.
- Careful count: 70 cells, 60 seconds.
- Computer count: 68 cells, $\ll 1$ second.





However: are all black dots cells? does it depend on the size? the computer performs faster, but also better? accuracy? parameters?

Thanks for coming!