Topological Data Analysis

Theory, Practice, Software, and Potential

Justin Skycak

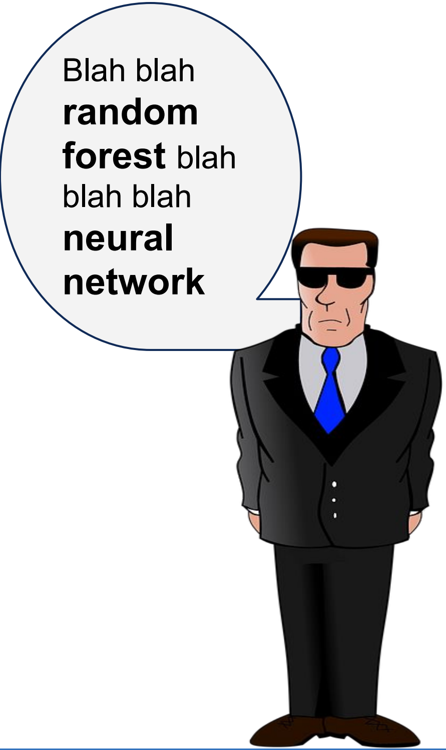
November 4, 2016

Contents

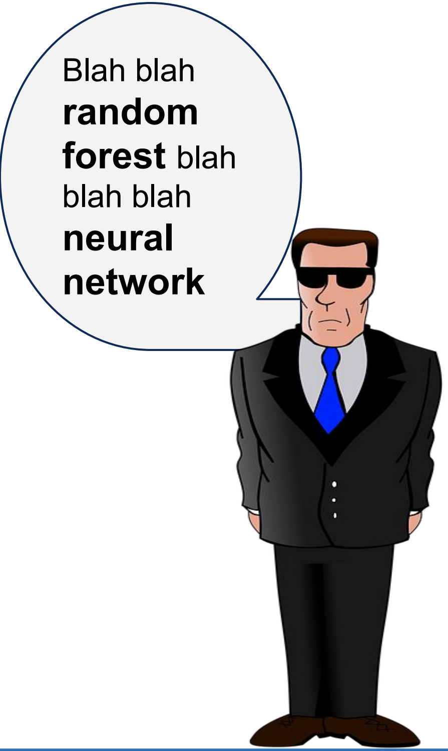
1. TDA and ML
2. TDA in Theory
3. TDA in Practice
4. TDA Software
5. TDA Potential

TDA and ML

Meet Matt, the Machine Learning Engineer

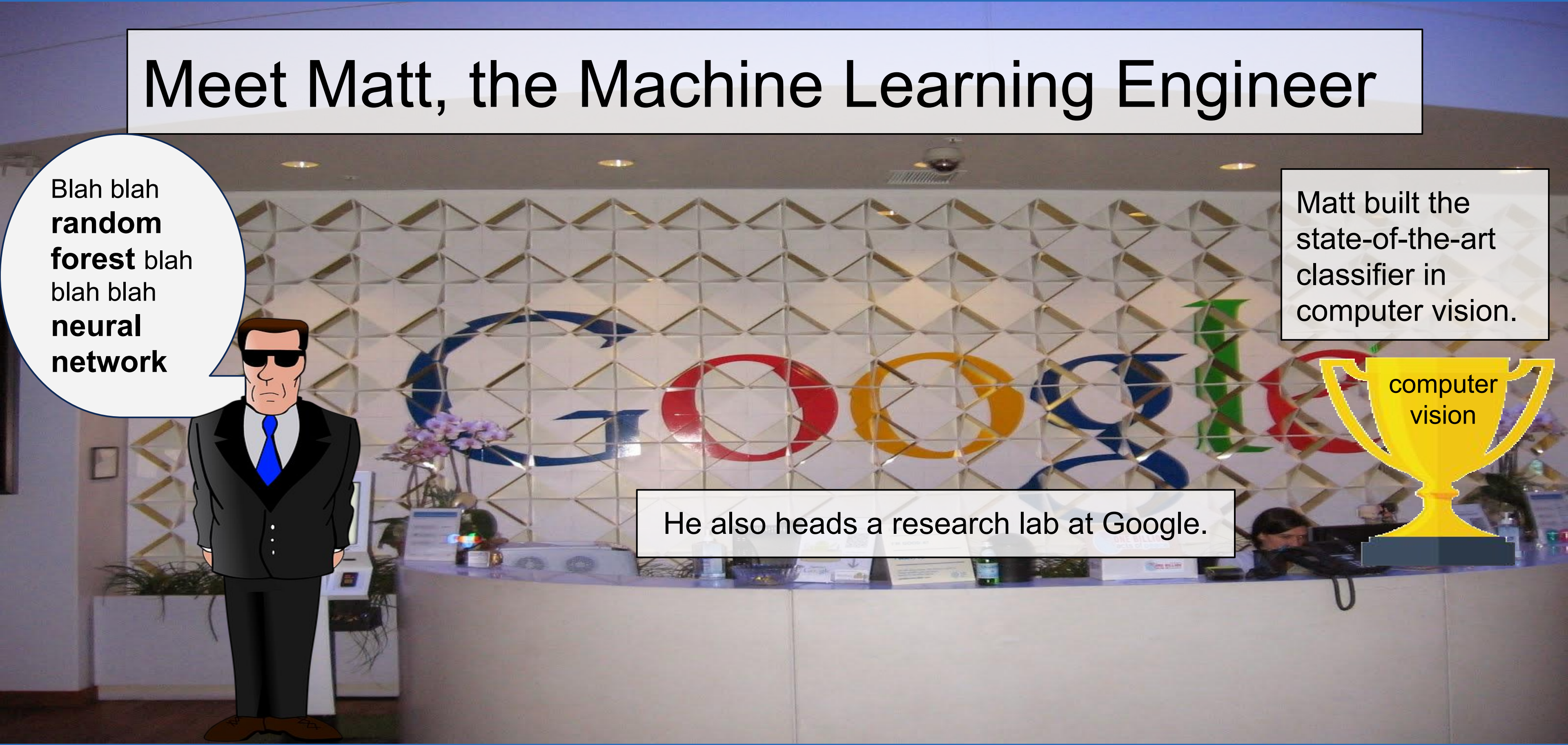


Meet Matt, the Machine Learning Engineer



Matt built the state-of-the-art classifier in computer vision.

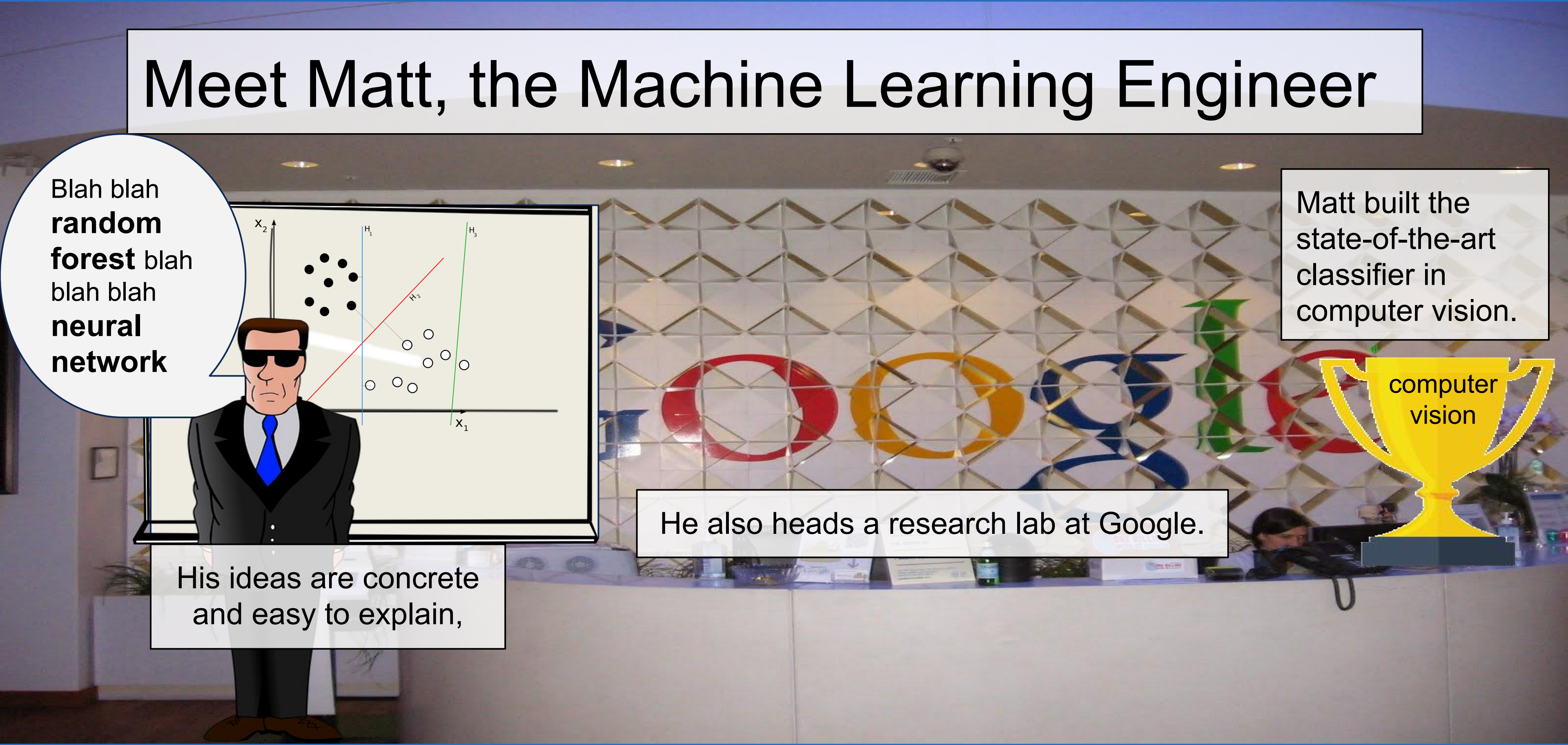
Meet Matt, the Machine Learning Engineer



Matt built the state-of-the-art classifier in computer vision.

He also heads a research lab at Google.

Meet Matt, the Machine Learning Engineer

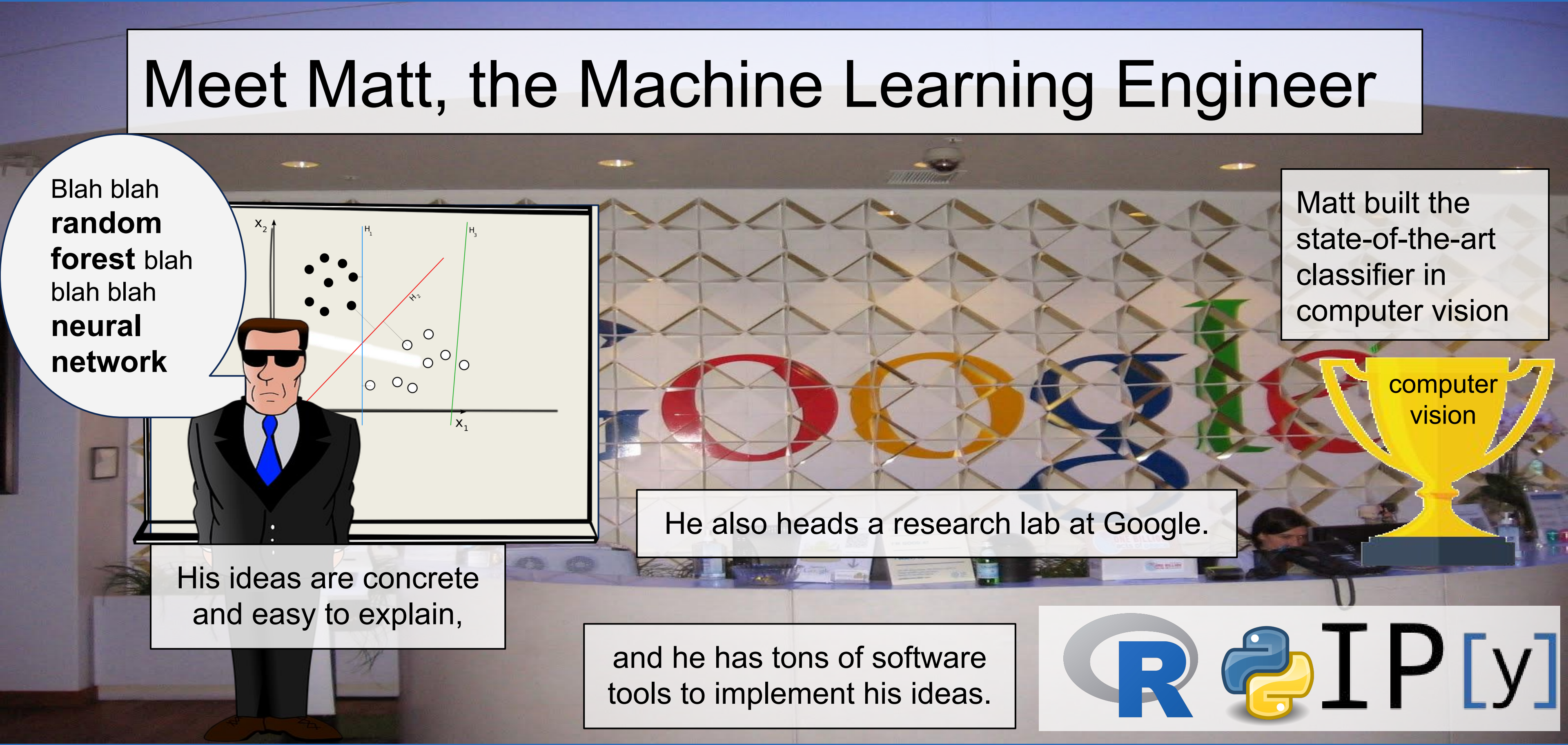


Matt built the state-of-the-art classifier in computer vision.

His ideas are concrete and easy to explain,

He also heads a research lab at Google.

Meet Matt, the Machine Learning Engineer



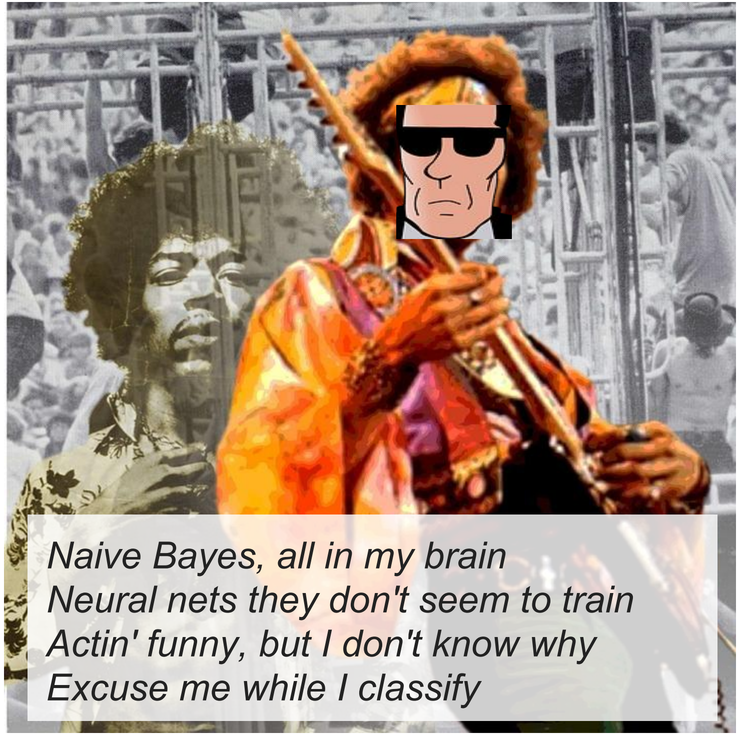
Matt built the state-of-the-art classifier in computer vision

He also heads a research lab at Google.

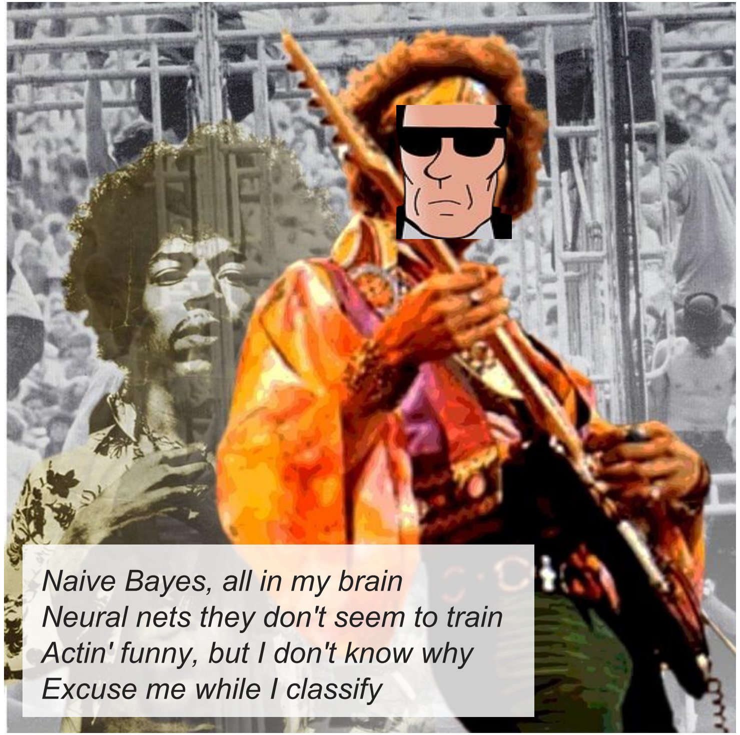
His ideas are concrete and easy to explain,

and he has tons of software tools to implement his ideas.

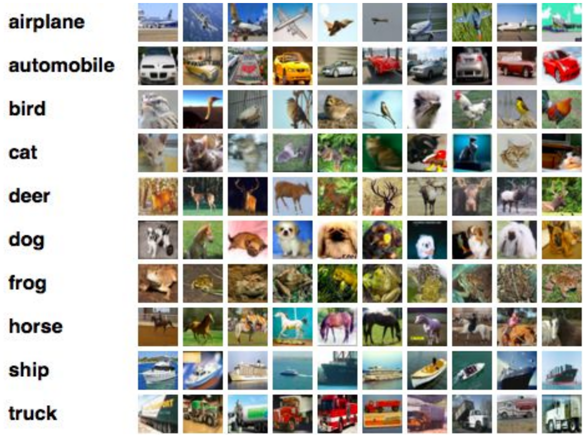
Matt may be a rock star...



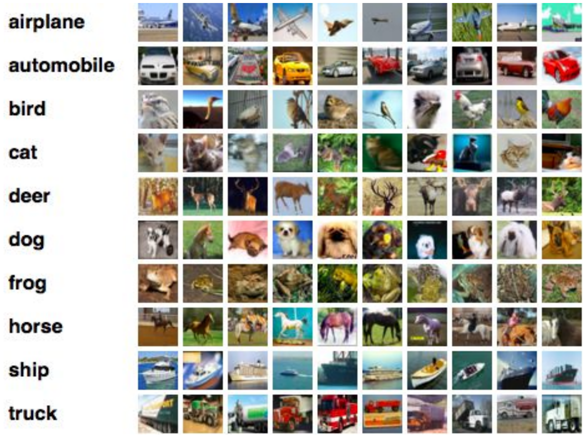
Matt may be a rock star...but there is a big skeleton in his closet.



Most of Matt’s tools only work when he has lots of labeled training data or a hypothesis he wishes to test

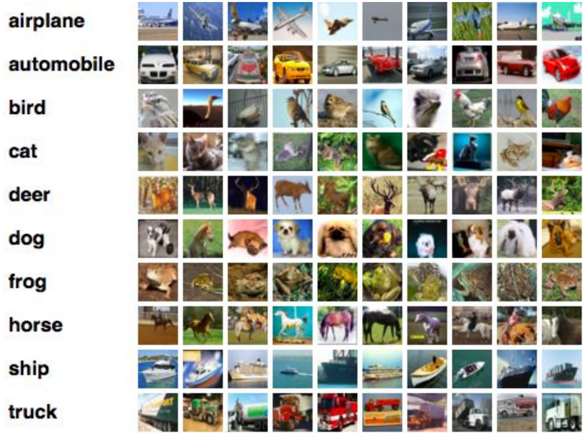


Most of Matt's tools only work when he has lots of labeled training data or a hypothesis he wishes to test



When Matt needs to extract some open-ended "insights" from unfamiliar data, he doesn't know where to begin.

Most of Matt's tools only work when he has lots of labeled training data or a hypothesis he wishes to test

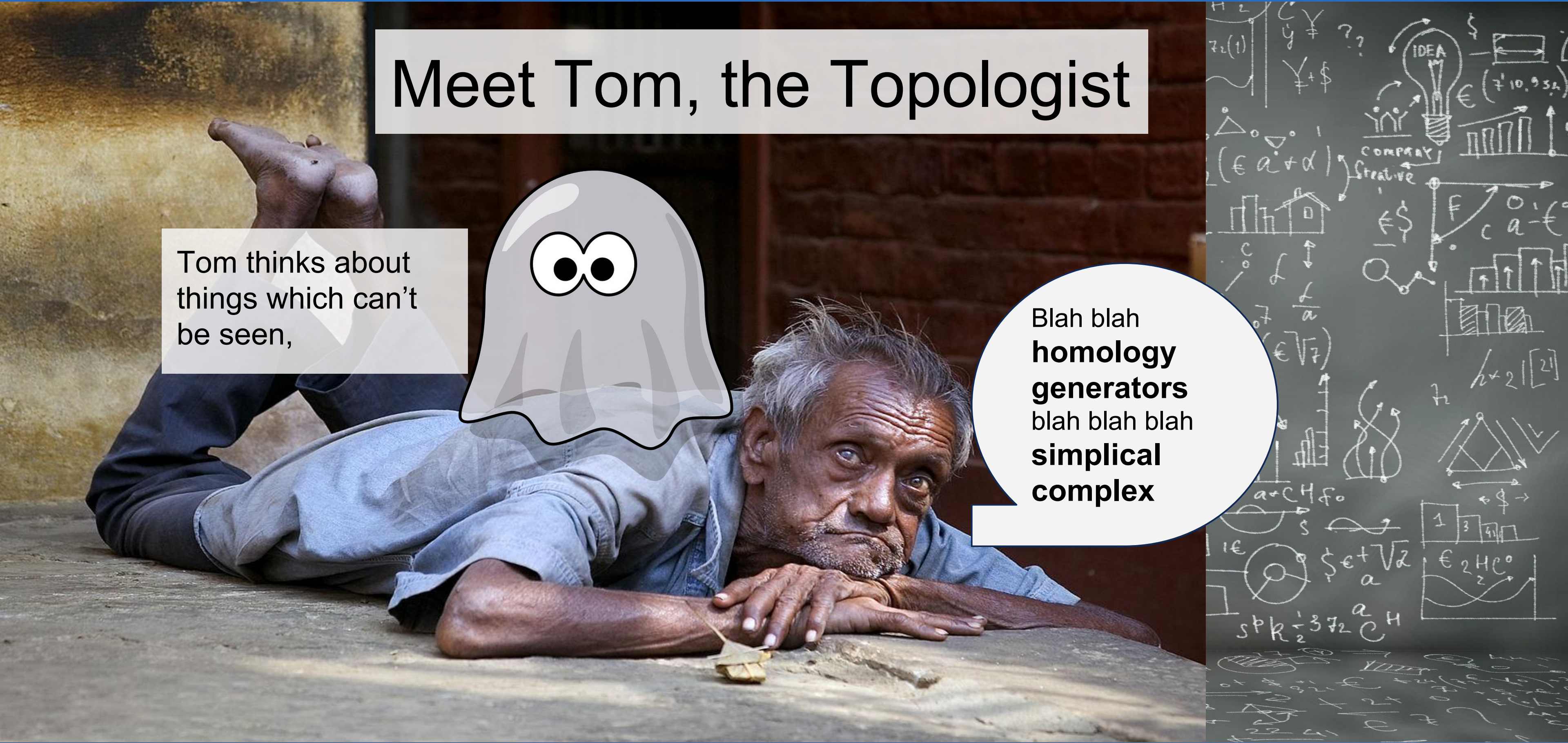


When Matt needs to extract some open-ended "insights" from unfamiliar data, he doesn't know where to begin.

So, Matt made a new friend to help him with open-ended data analysis.

Meet Tom, the Topologist

Meet Tom, the Topologist

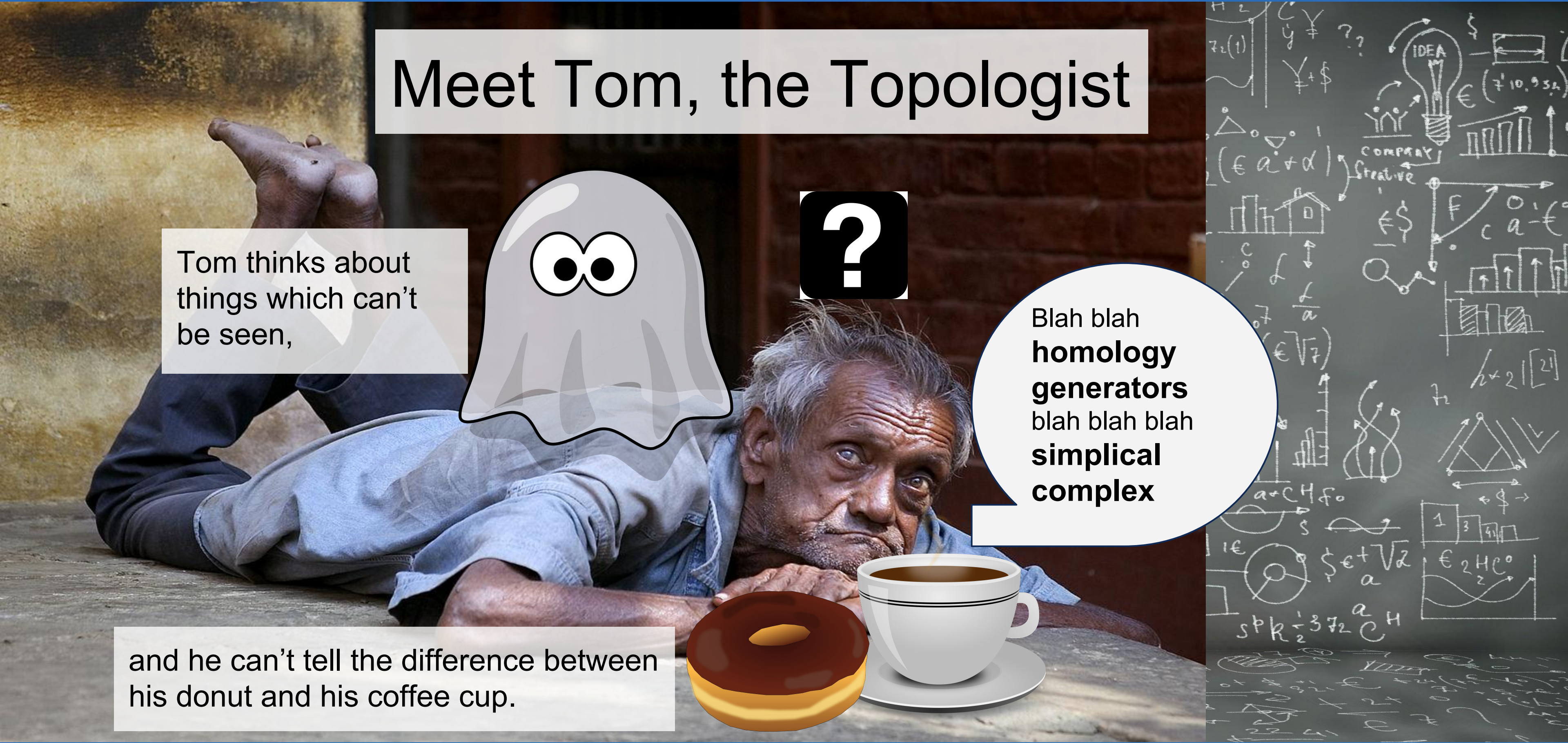


Tom thinks about things which can't be seen,

Blah blah **homology generators** blah blah blah **simplical complex**

Meet Tom, the Topologist

Tom thinks about things which can't be seen,

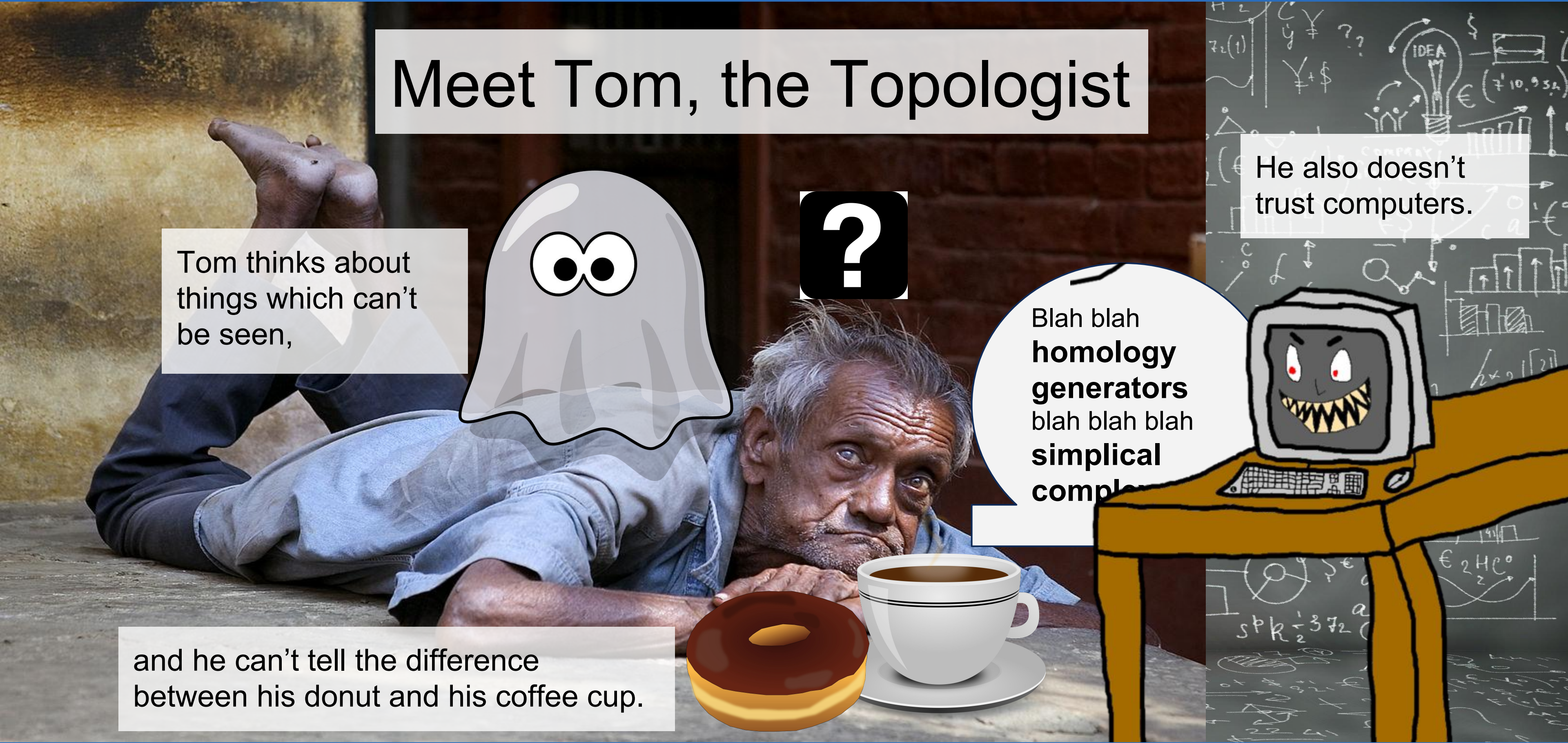


and he can't tell the difference between his donut and his coffee cup.

Meet Tom, the Topologist

Tom thinks about things which can’t be seen,

He also doesn’t trust computers.



and he can’t tell the difference between his donut and his coffee cup.

Why did Matt choose Tom?

Why did Matt choose Tom?

Tom understands shape of data.

Why did Matt choose Tom?

Tom understands shape of data.

His methods don’t require hypotheses, parameters, or even coordinates.

Why did Matt choose Tom?

Tom understands shape of data.

His methods don’t require hypotheses, parameters, or even coordinates.

To draw insights from data, all he needs is a measure of similarity between points.

Why did Matt choose Tom?

Tom understands shape of data.

His methods don't require hypotheses, parameters, or even coordinates.

To draw insights from data, all he needs is a measure of similarity between points.

How does Tom do it???

TDA in Theory

Ask yourself:

Is the shape of O more similar to that of P, or B?

Our intuition about shape is based on loops

O and P have similar shape: 1 loop

O and B have different shape: 1 loop vs 2 loops

Our intuition about shape is based on loops

O and P have similar shape: 1 loop
O and B have different shape: 1 loop vs 2 loops

The number of loops in a space turns out to depend precisely on the number of holes in the space

Our intuition about shape is based on loops

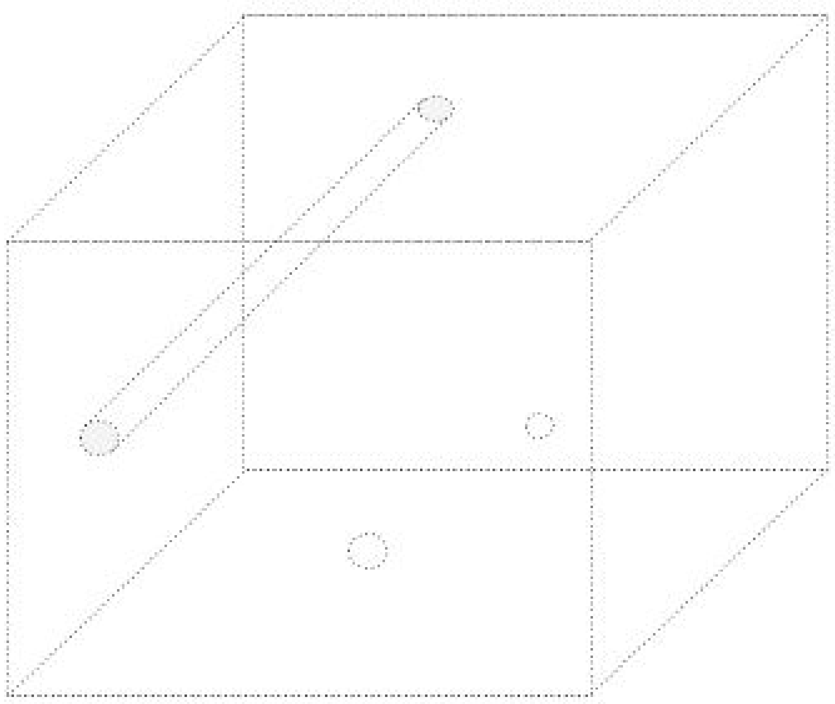
O and P have similar shape: 1 loop
O and B have different shape: 1 loop vs 2 loops

The number of loops in a space turns out to depend precisely on the number of holes in the space

We can classify a space by counting its holes in each dimension!

Example: Three-Dimensional Space

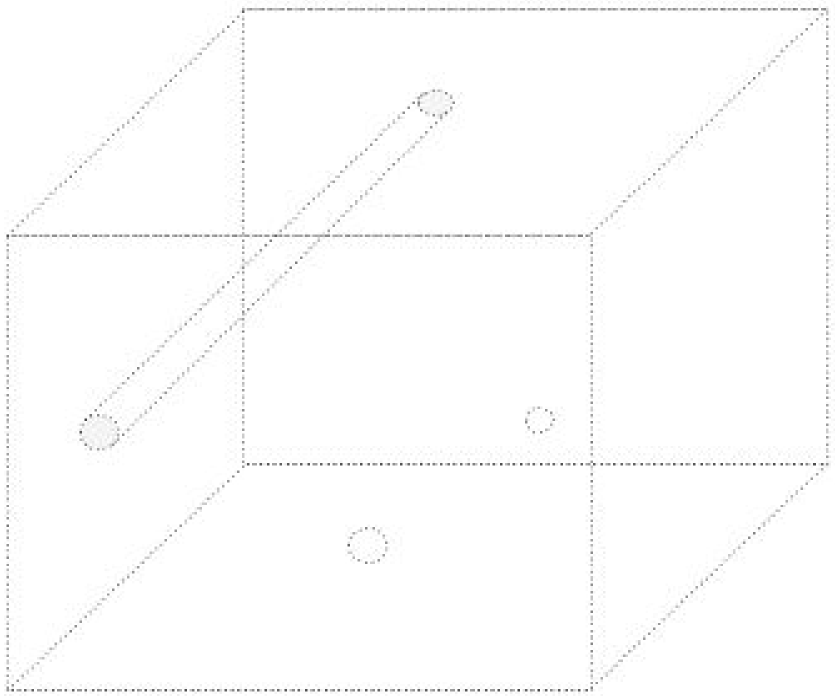
Missing two points → 2 one-dimensional holes



Example: Three-Dimensional Space

Missing two points → 2 one-dimensional holes

Missing one line → 1 two-dimensional hole

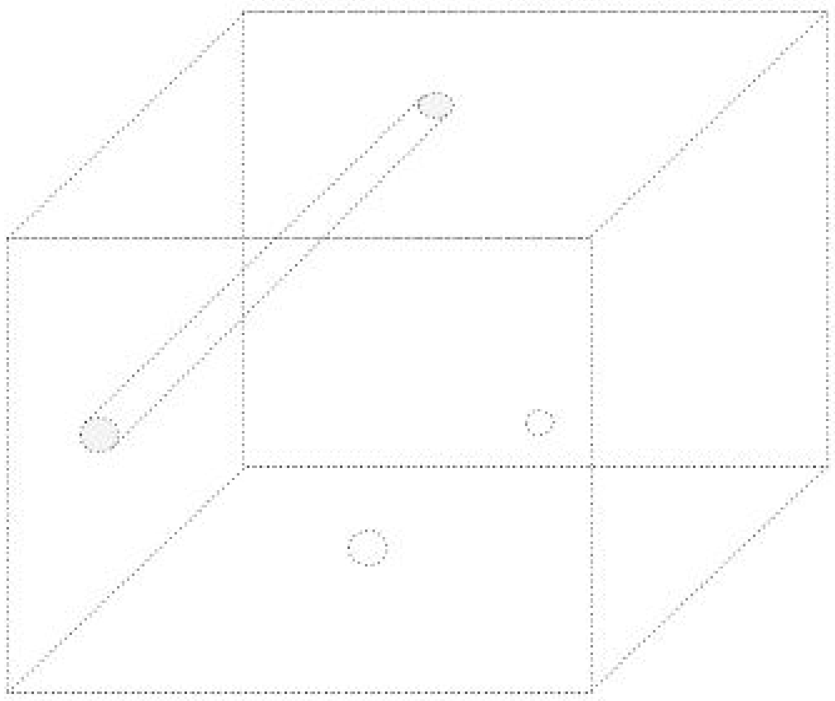


Example: Three-Dimensional Space

Missing two points → 2 one-dimensional holes

Missing one line → 1 two-dimensional hole

Classification: (2,1)



We can use these classifications to compare spaces!

A space can be represented as a tuple: (3, 1, 2) means

- 3 holes in dimension one
- 1 hole in dimension two
- 2 holes in dimension three

(3, 1, 2) is more similar to (4, 1, 2) than (2, 5, 5)

Philosophical Aside

We have lots of mathematical machinery to operate on transformations between points, e.g. probability and calculus.

Philosophical Aside

We have lots of mathematical machinery to operate on transformations between points, e.g. probability and calculus.

Up until topology, we were limited to using these tools within a particular space at a given time.

Philosophical Aside

We have lots of mathematical machinery to operate on transformations between points, e.g. probability and calculus.

Up until topology, we were limited to using these tools within a particular space at a given time.

Topology gives us a way to talk about entire spaces as points.

Philosophical Aside

We have lots of mathematical machinery to operate on transformations between points, e.g. probability and calculus.

Up until topology, we were limited to using these tools within a particular space at a given time.

Topology gives us a way to talk about entire spaces as points.

We can now use distance, probability, and calculus to study transformations between entire spaces! **(in theory)**

TDA in Practice

In theory...

TDA is about measuring similarity between spaces based on their topological features (holes)

In theory...

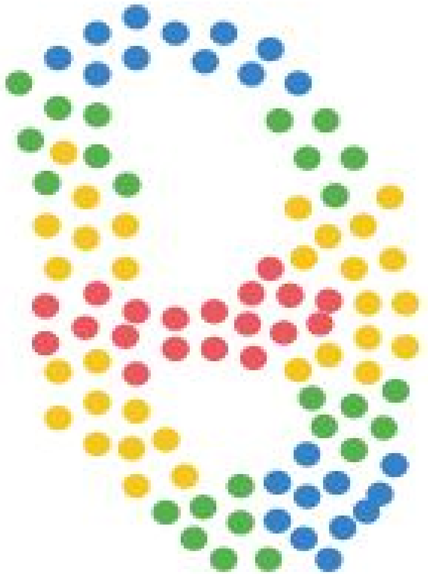
TDA is about measuring similarity between spaces based on their topological features (holes)

In practice...

TDA is about visualizing high-dimensional spaces as networks, without losing topological features

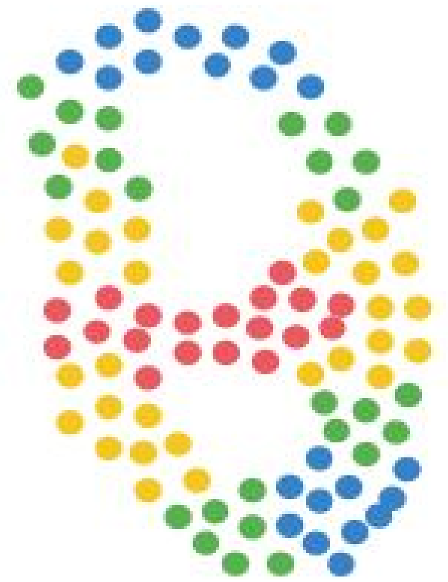
The Mapper Algorithm

Colored data points



The Mapper Algorithm

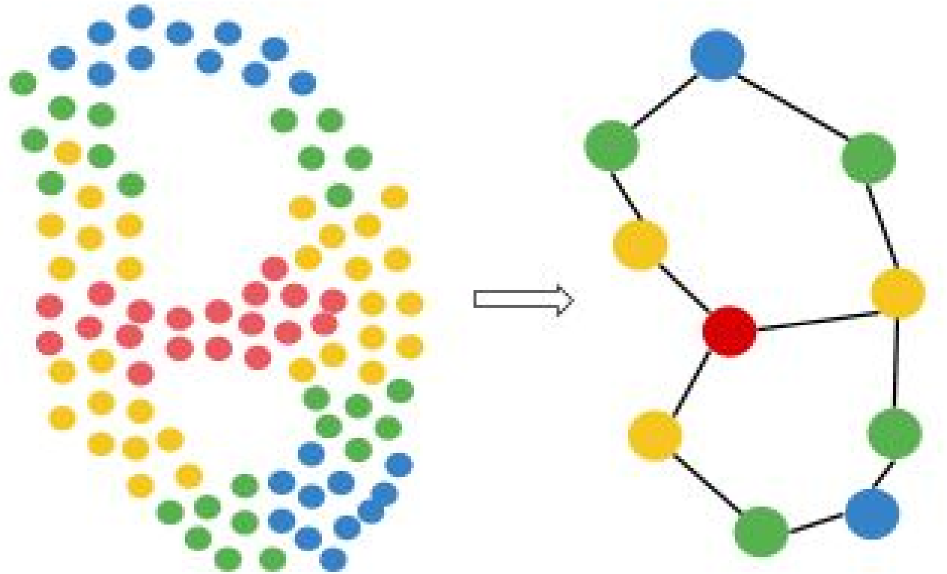
Colored data points



E.g. each point is of a correlation matrix for asset prices, colored according to returns %

The Mapper Algorithm

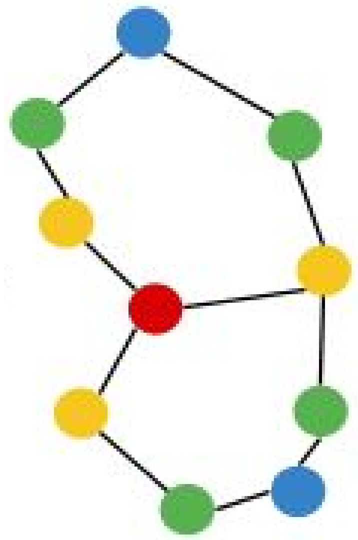
Colored data points



Colored network

The Mapper Algorithm

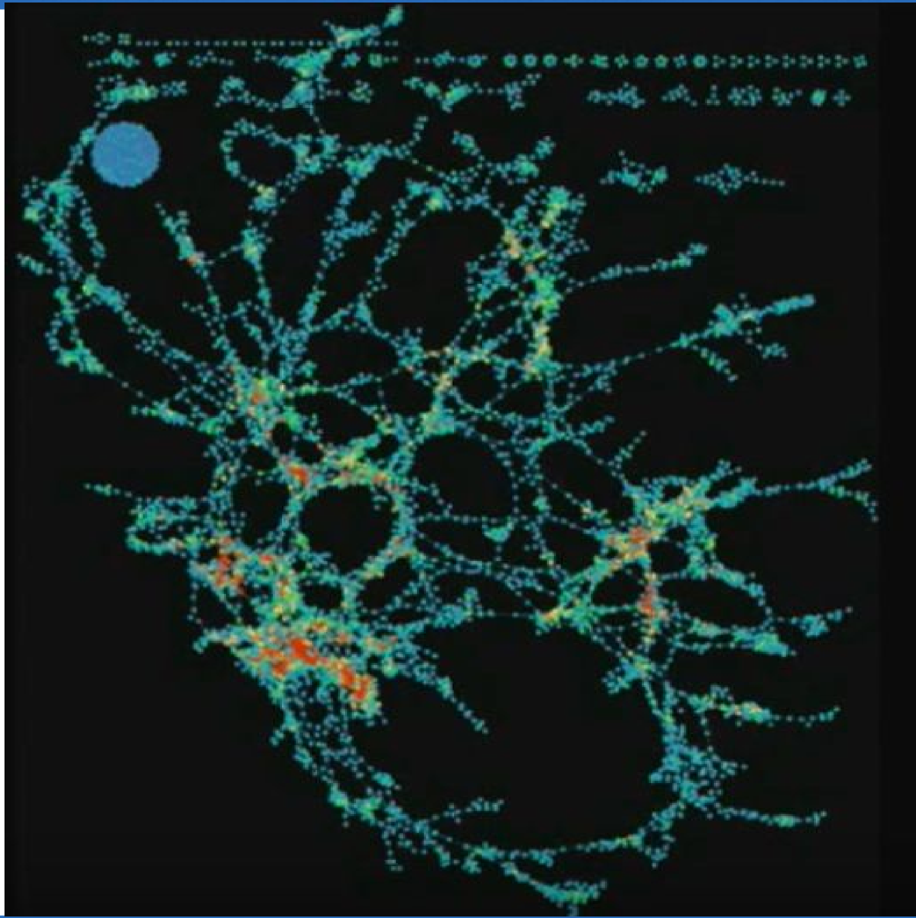
Nodes correspond to market regimes, colored by returns %



Colored network

Real Use Case: Forecasting Returns

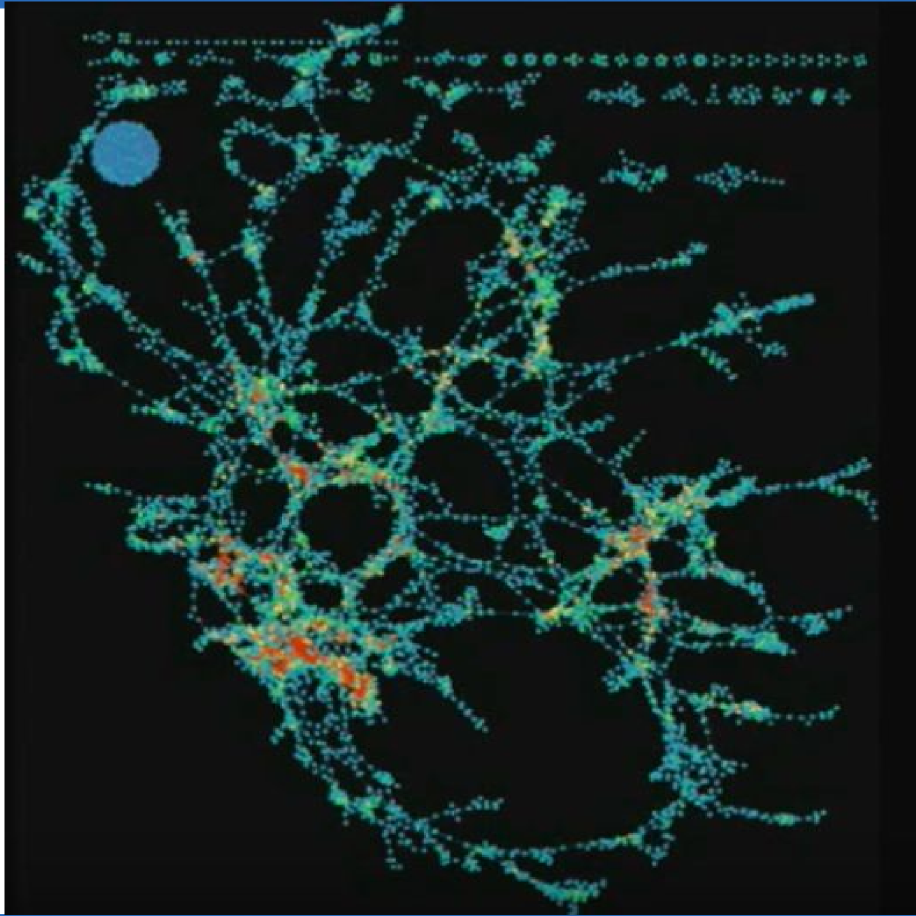
300+ market and economic variables, sampled over 25 years



Real Use Case: Forecasting Returns

300+ market and economic variables, sampled over 25 years

Nodes colored by year

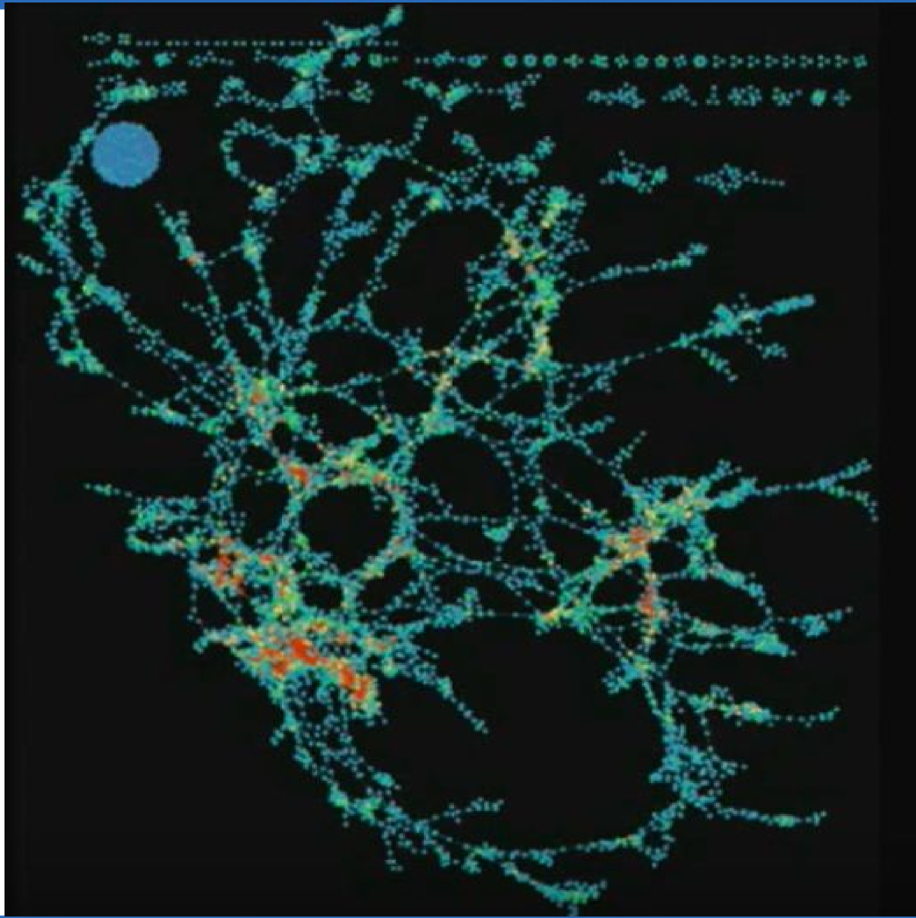


Real Use Case: Forecasting Returns

300+ market and economic variables, sampled over 25 years

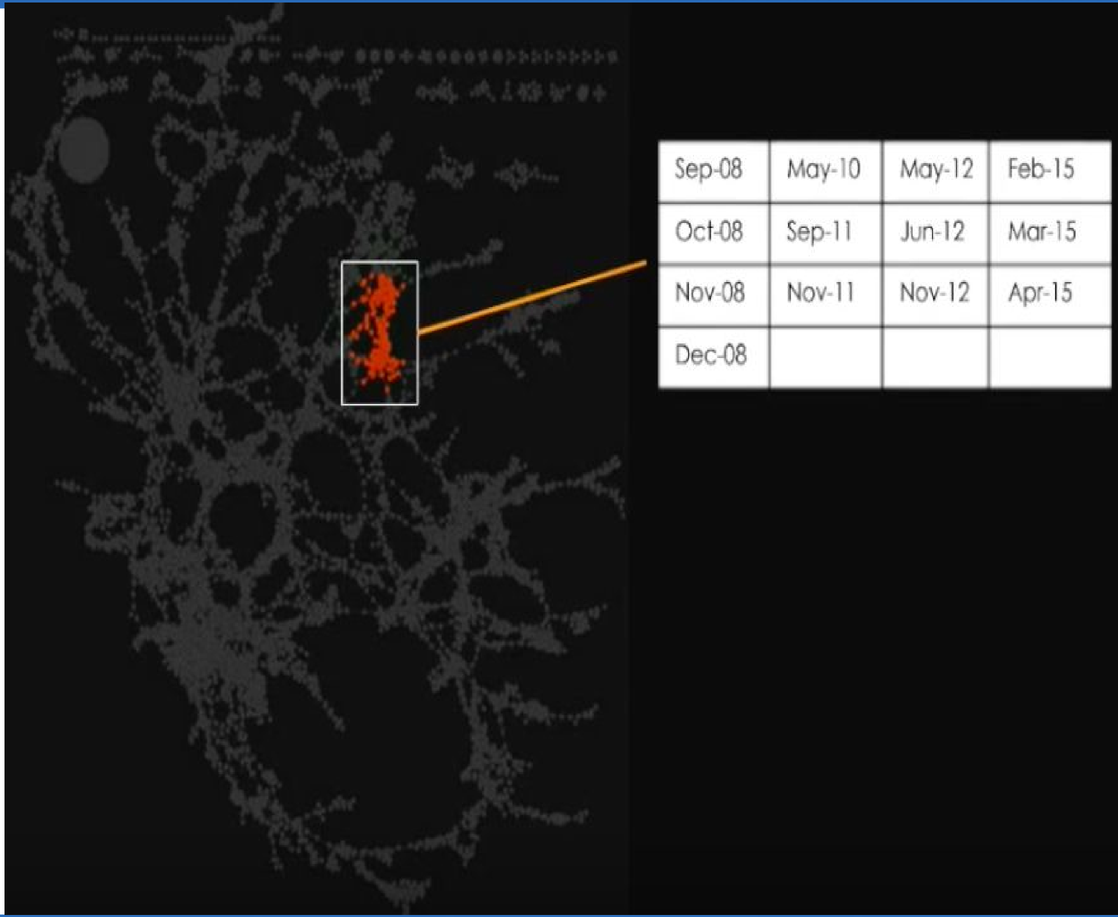
Nodes colored by year

Colors are spread out → indicates repeated patterns over time



Real Use Case: Forecasting Returns

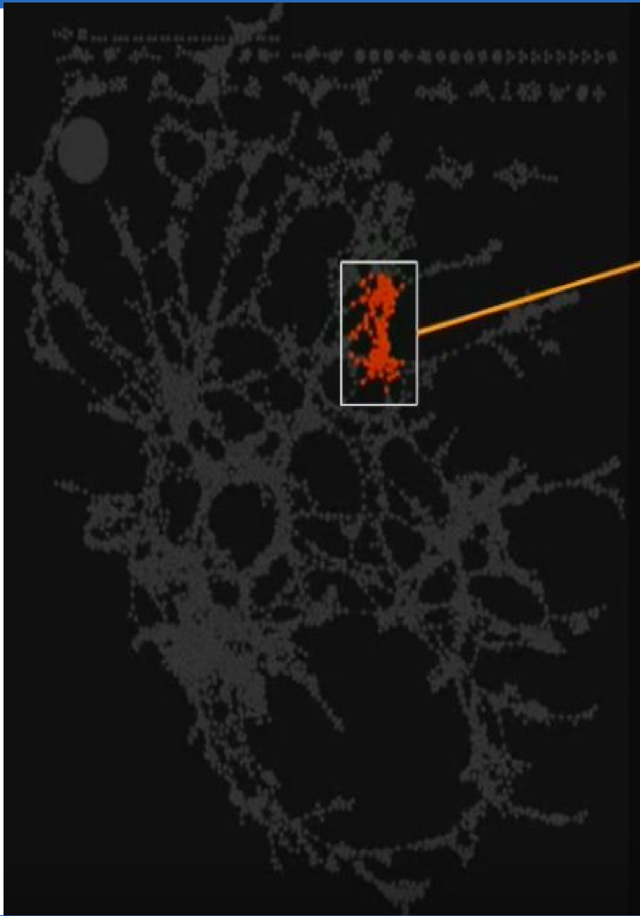
High-volatility and high-stress times are grouped together



Real Use Case: Forecasting Returns

High-volatility and high-stress times are grouped together

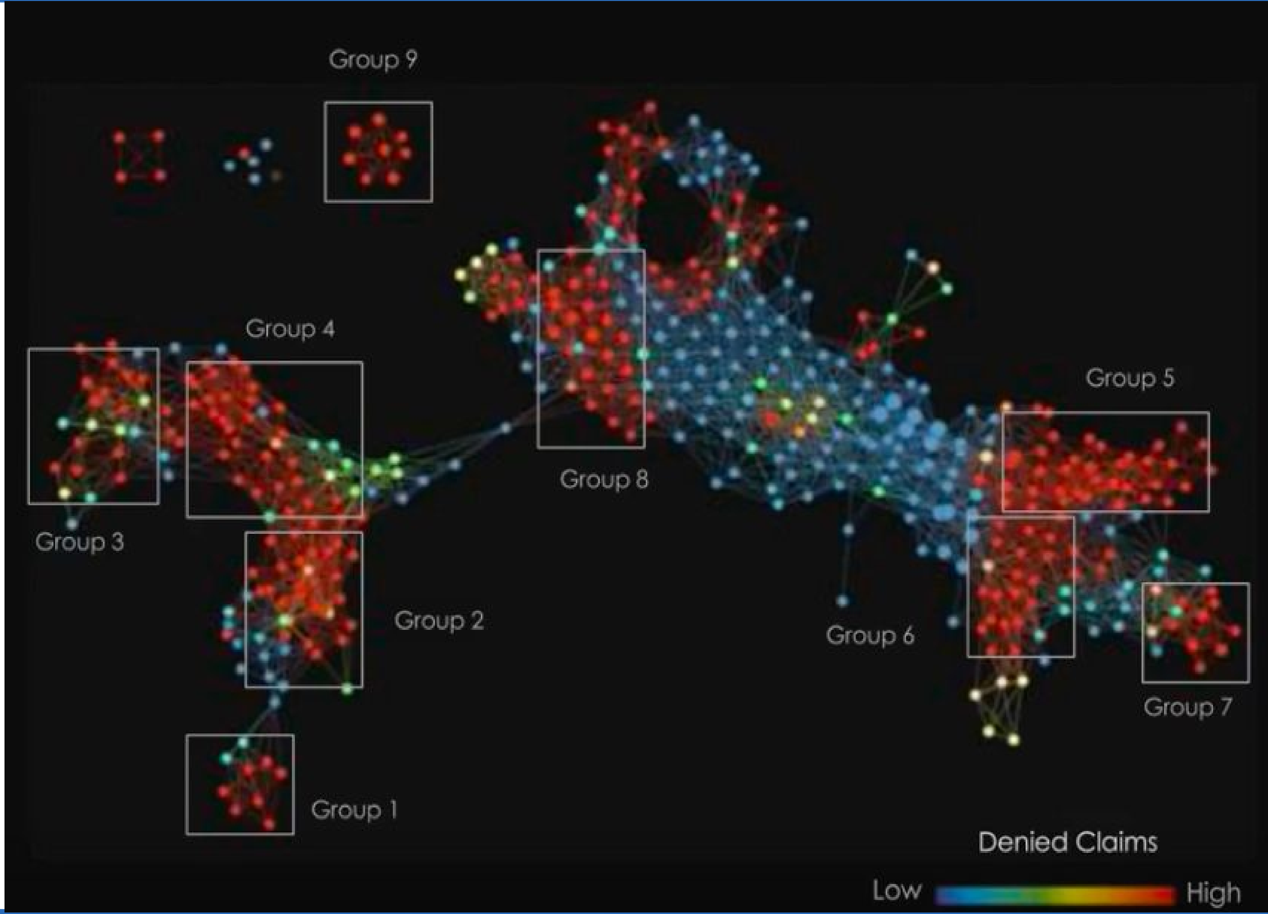
Implies similar market regimes



Sep-08	May-10	May-12	Feb-15
Oct-08	Sep-11	Jun-12	Mar-15
Nov-08	Nov-11	Nov-12	Apr-15
Dec-08			

Real Use Case: Diagnosing Denied Claims

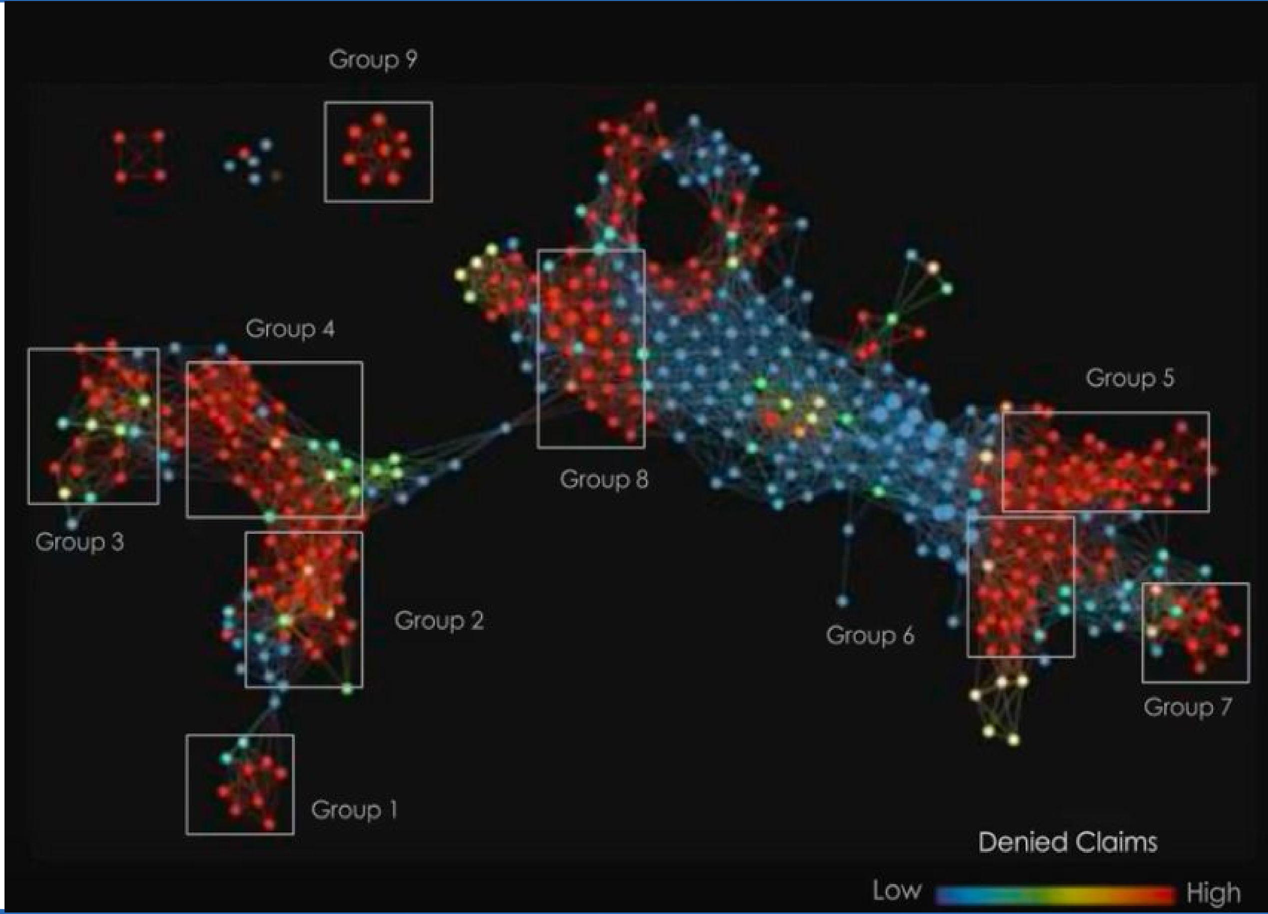
Structure: claim similarity (5 million medical claims)



Real Use Case: Diagnosing Denied Claims

Structure: claim similarity (5 million medical claims)

Color: claim denial frequency



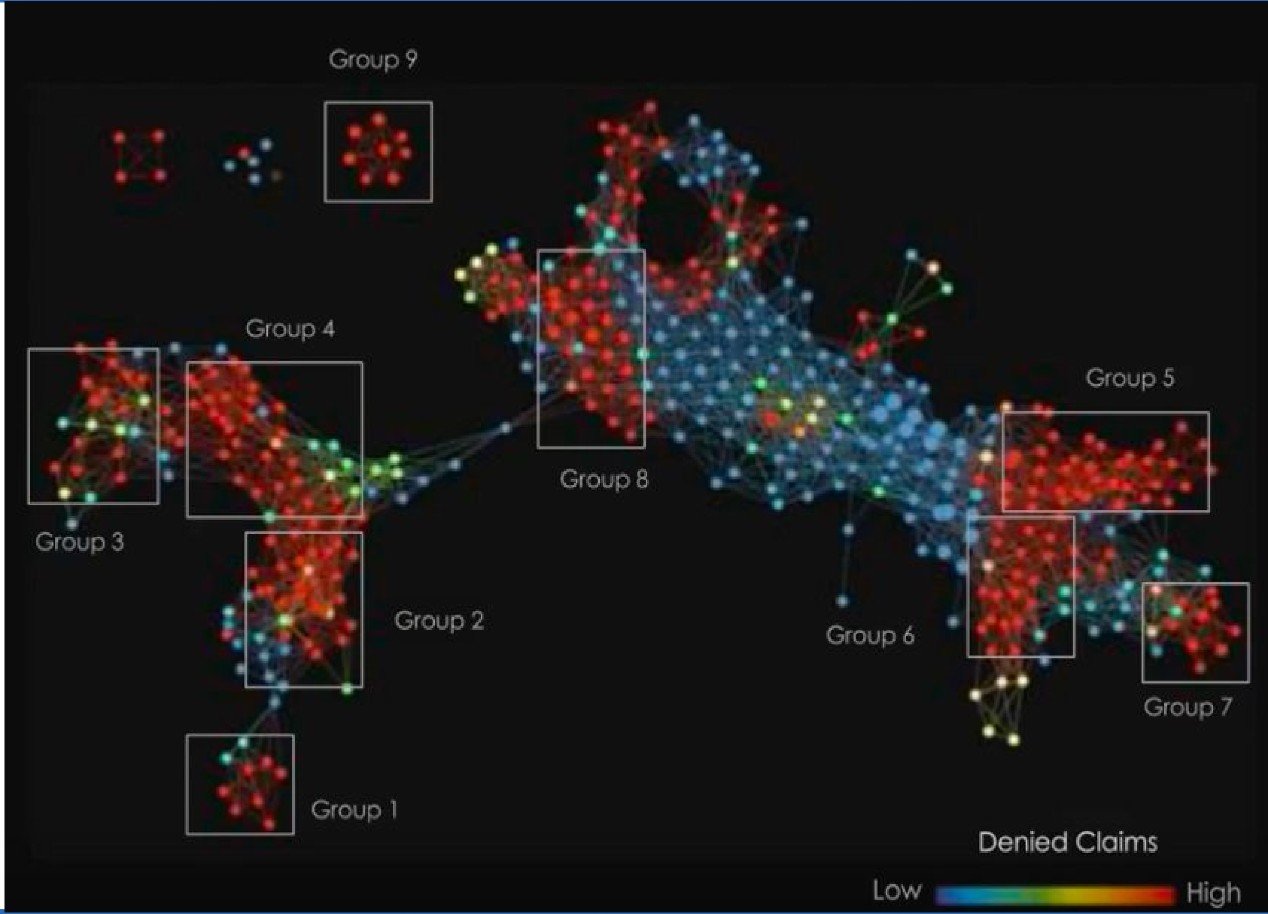
Real Use Case: Diagnosing Denied Claims

Structure: claim similarity (5 million medical claims)

Color: claim denial frequency

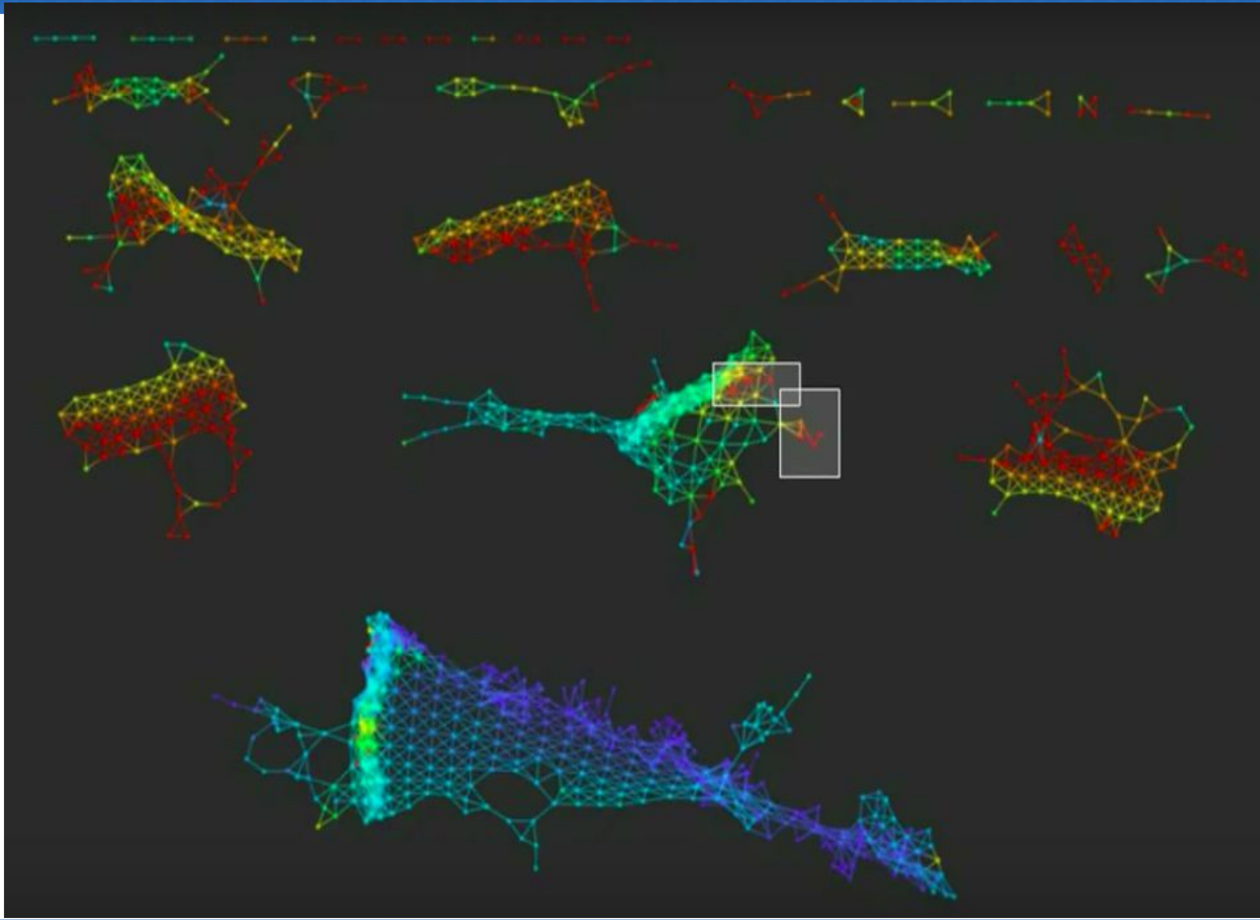
Result: advice for

- *pre-submission action*: modifying final code or supporting diagnosis
- *point of care*: seeking pre-authorization or reconsidering a procedure



Real Use Case: Identifying Fraud

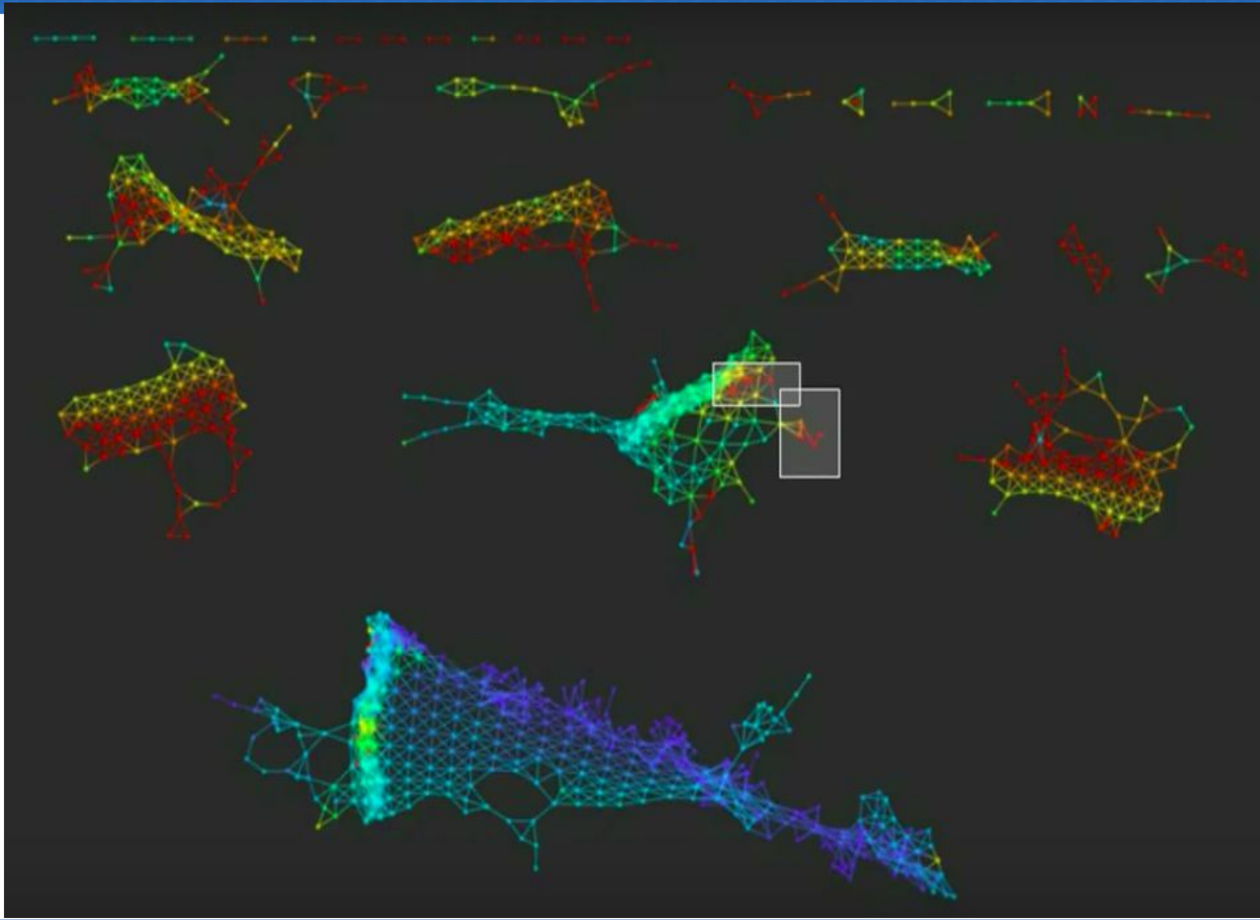
Structure: similarity in how providers practice
(CMS public health claims dataset)



Real Use Case: Identifying Fraud

Structure: similarity in how providers practice (CMS public health claims dataset)

Color: medicare payment amount

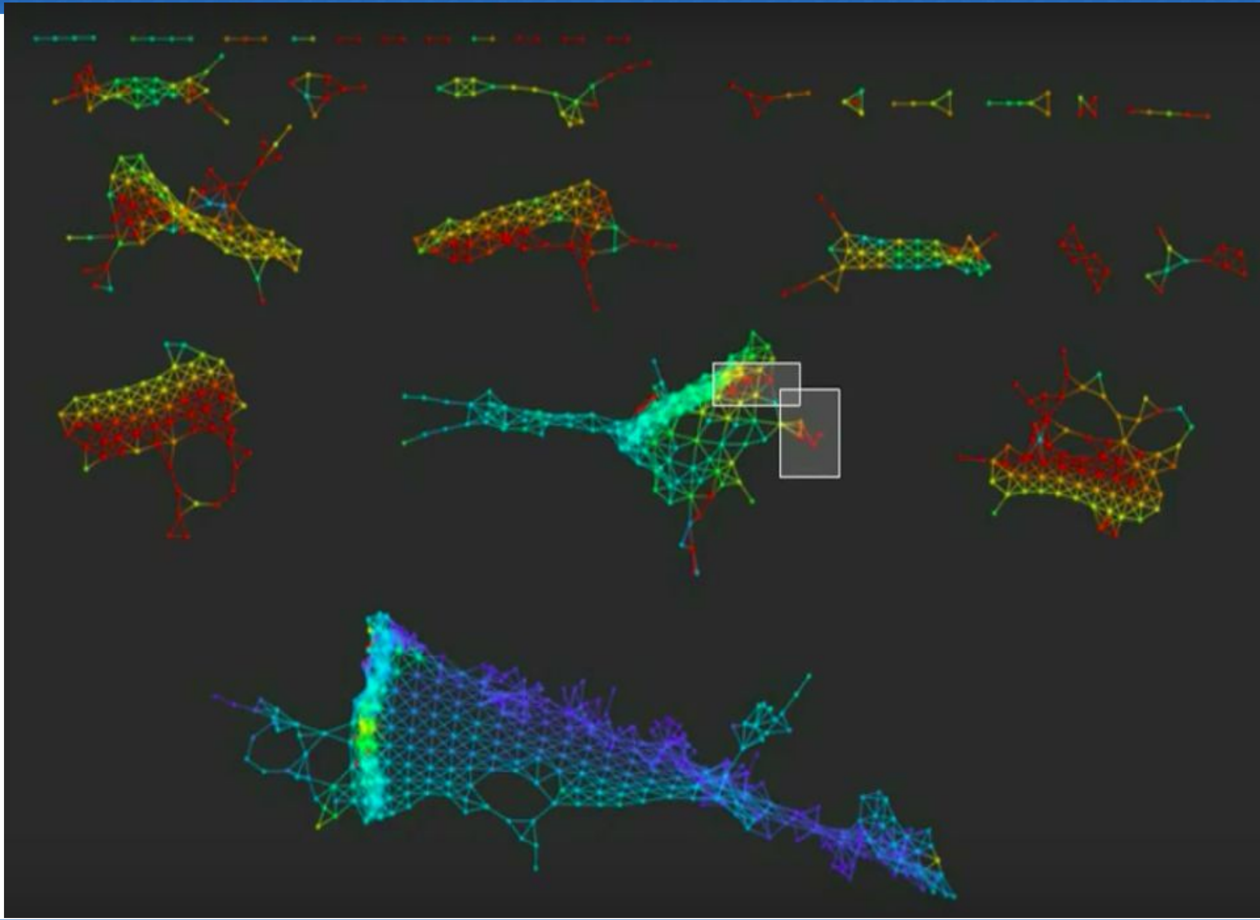


Real Use Case: Identifying Fraud

Structure: similarity in how providers practice (CMS public health claims dataset)

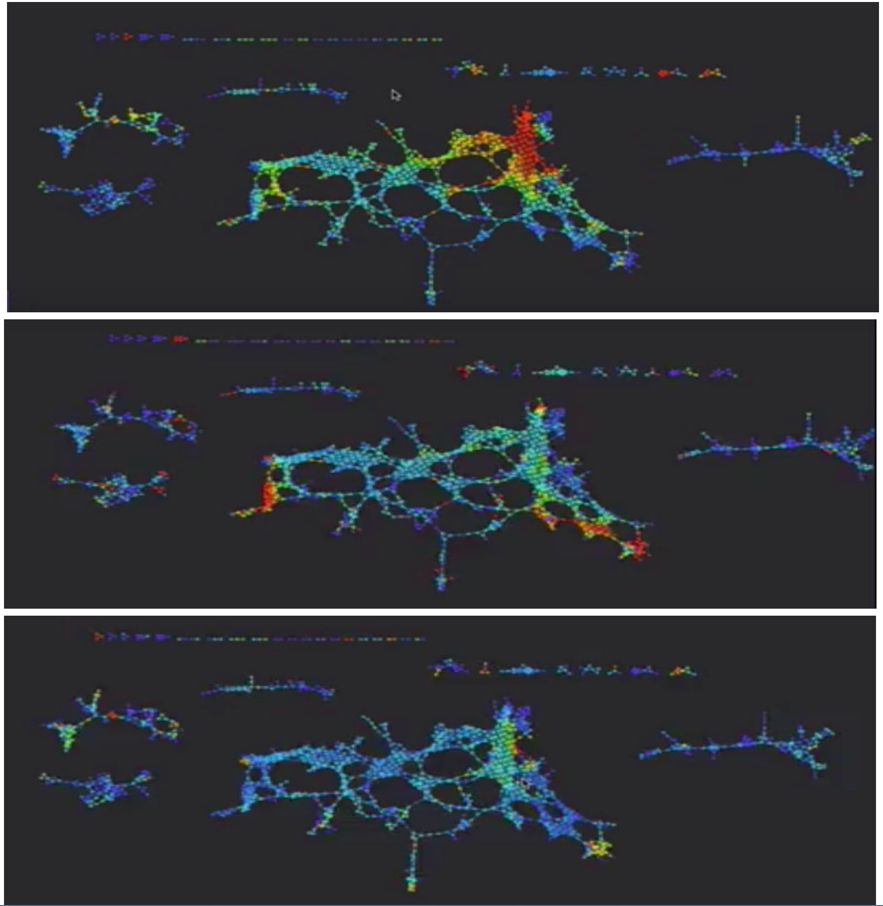
Color: medicare payment amount

Result: Identify leads for fraud investigation by looking for outlier providers who are getting paid abnormally much compared to similar providers



Real Use Case: Campaign Planning

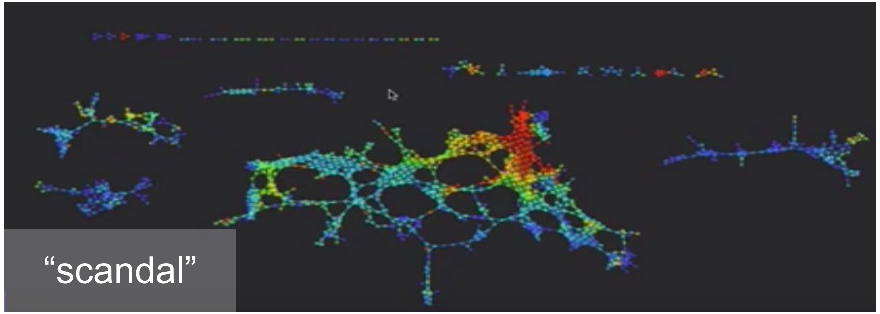
Structure: Twitter account similarity (36k users who tweeted about Chris Christie)

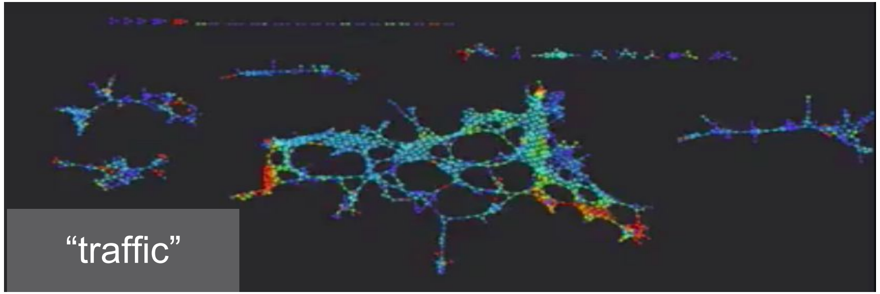


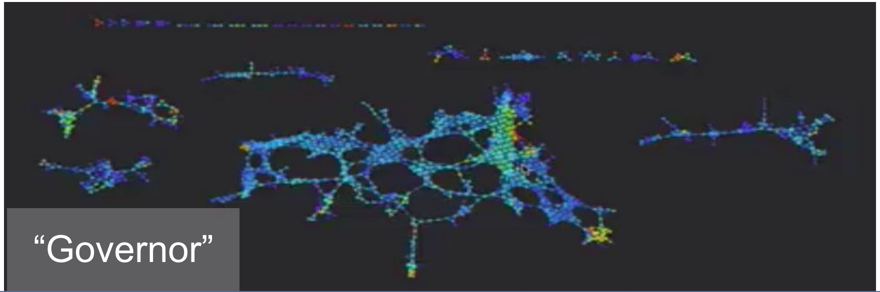
Real Use Case: Campaign Planning

Structure: Twitter account similarity (36k users who tweeted about Chris Christie)

Color: word frequency





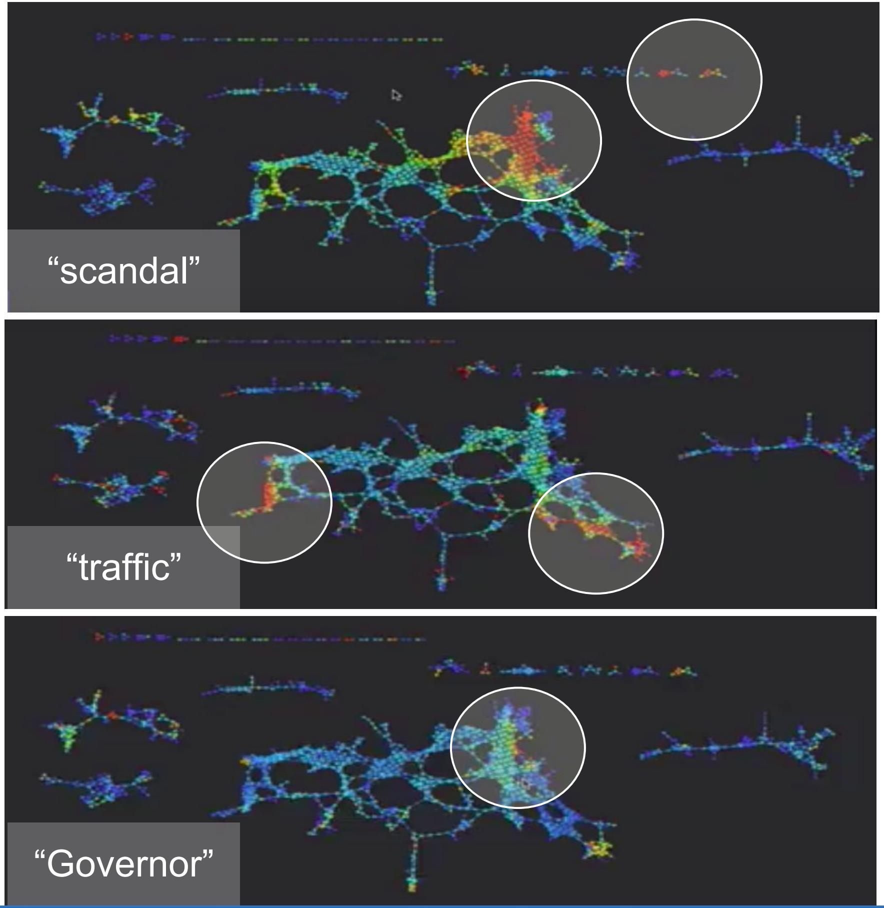


Real Use Case: Campaign Planning

Structure: Twitter account similarity (36k users who tweeted about Chris Christie)

Color: word frequency

Result: Identify niche conversations that are good targets for ads.



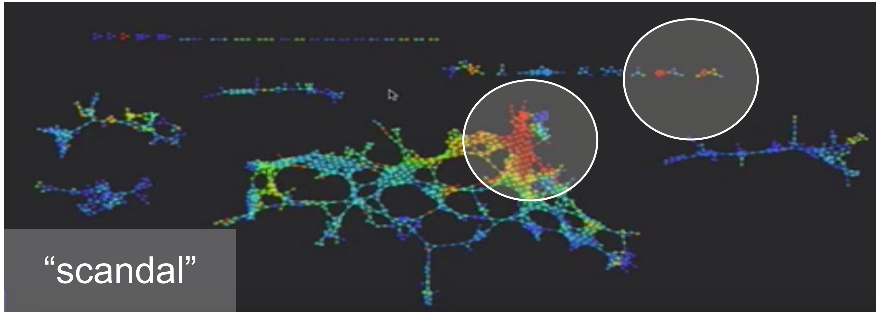
Real Use Case: Campaign Planning

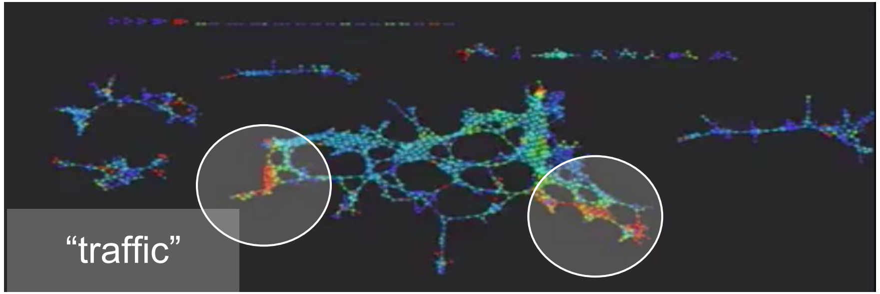
Structure: Twitter account similarity (36k users who tweeted about Chris Christie)

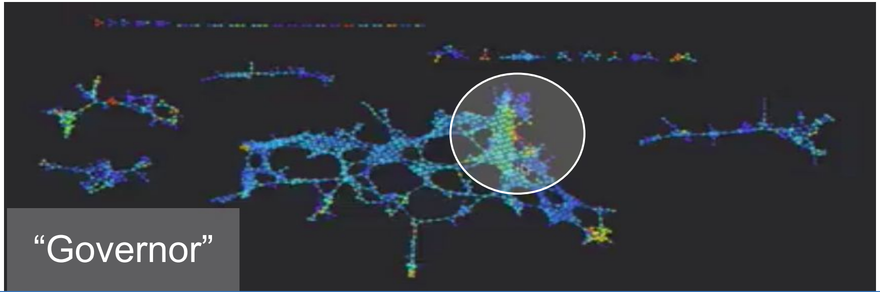
Color: word frequency

Result: Identify niche conversations that are good targets for ads.

(Can also investigate an individual group to see what other words differentiate the group from others.)







TDA Software

Commercial Software: Ayasdi

Ayasdi dominates the commercial TDA market.

Commercial Software: Ayasdi

Ayasdi dominates the commercial TDA market.

Every single use-case in the previous section is from an Ayasdi webinar.

Commercial Software: Ayasdi

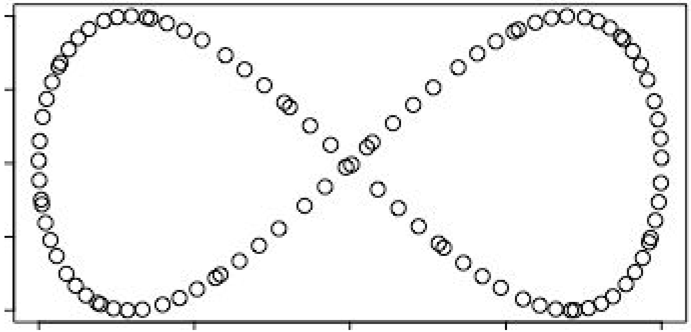
Ayasdi dominates the commercial TDA market.

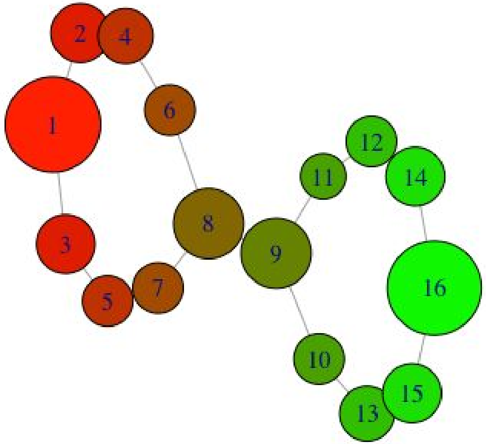
Every single use-case in the previous section is from an Ayasdi webinar.

Ayasdi not only implements the Mapper algorithm but also has an “explain” function which automatically differentiates clusters by running a barrage of statistical tests and ranking their most significant differences.

Open Source Software: TDAmapper

Not as pretty or easy as Ayasdi, but still not bad*:





*I don’t know how well (or poorly) it scales

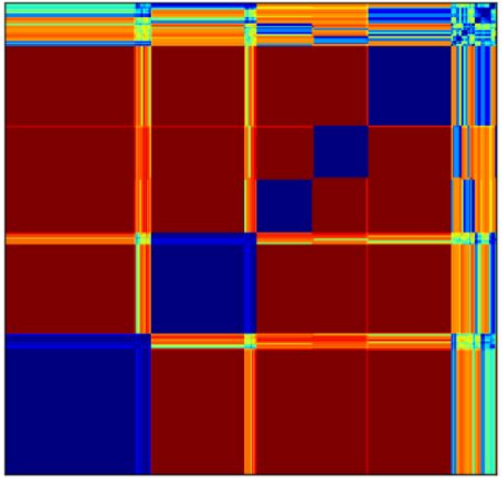
TDA Potential

Snoopy - Location Tracking

Dataset includes counts of visits to different location categories, for several thousand users

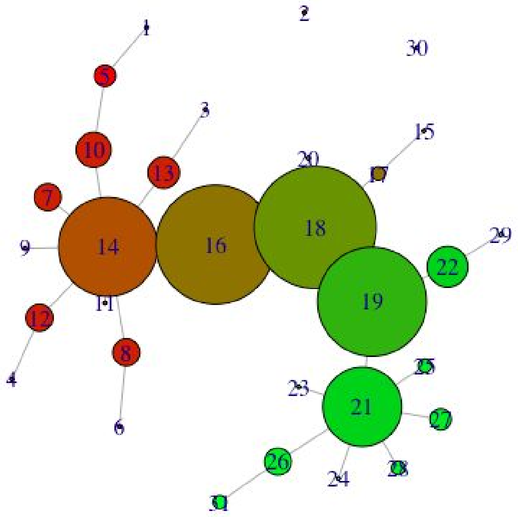
Clustering visit profiles within the "Recreation and Leisure" category revealed 5 niche clusters:

Recreation Category	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Athletic Fields	1%	1%	1%	99%	1%	49%
Golf	0%	95%	1%	1%	0%	7%
Gym and Fitness	1%	1%	1%	0%	97%	6%
Outdoors	1%	1%	96%	0%	1%	13%
Recreation Centers	0%	0%	0%	0%	0%	7%
Stadiums and Arenas	97%	1%	2%	0%	1%	17%
Swimming Pools	0%	0%	0%	0%	0%	7%
	High-end sports players/fans	Golfers	Hikers Campers	Recreational sports players	Gym rats	Everyone else



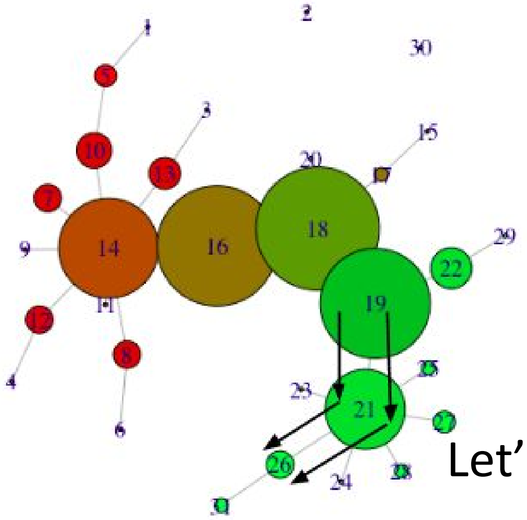
Snoopy - Location Tracking

Mapper algorithm reveals many more clusters. Moreover, we can see the paths by which they are connected.



Snoopy - Location Tracking

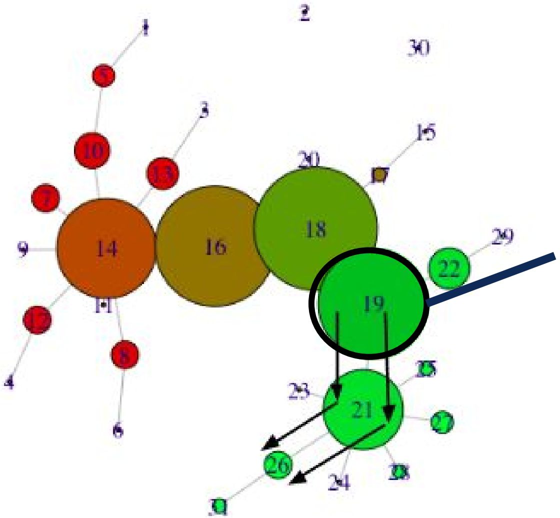
Mapper algorithm reveals many more clusters. Moreover, we can see the paths by which they are connected.

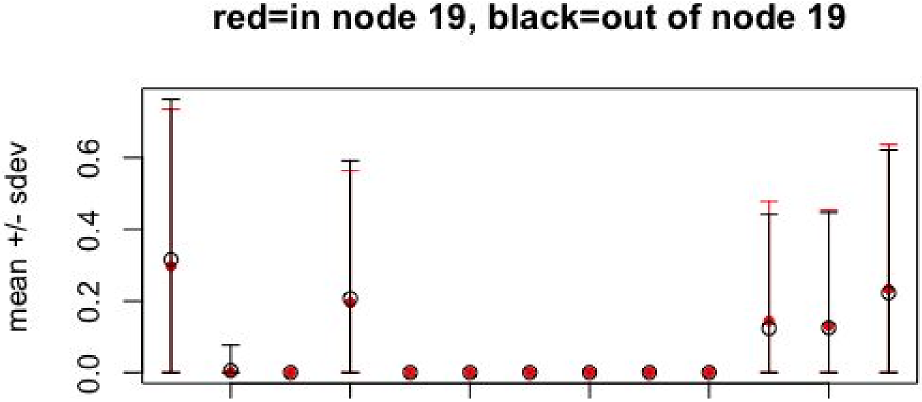


Let’s investigate this flare!

Snoopy - Location Tracking

Mapper algorithm reveals many more clusters. Moreover, we can see the paths by which they are connected.

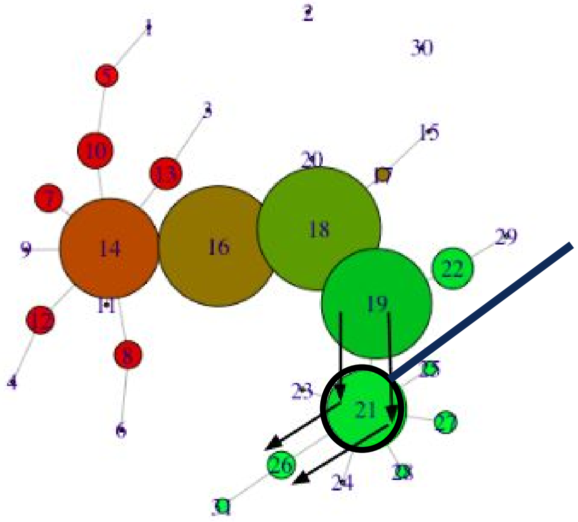


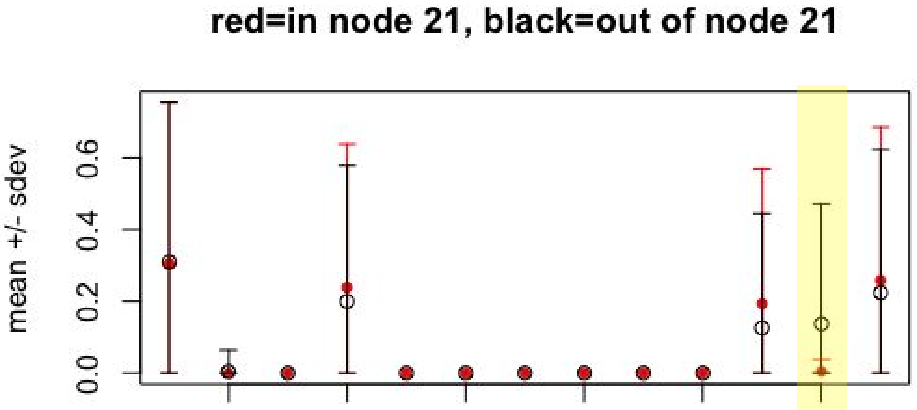


Node 19 has a normal profile

Snoopy - Location Tracking

Mapper algorithm reveals many more clusters. Moreover, we can see the paths by which they are connected.

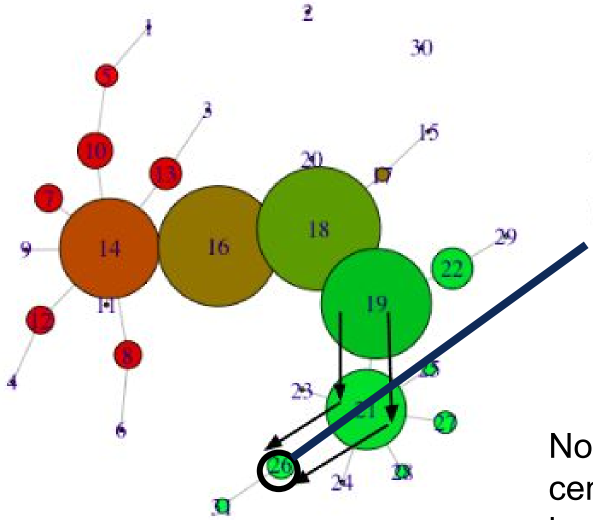




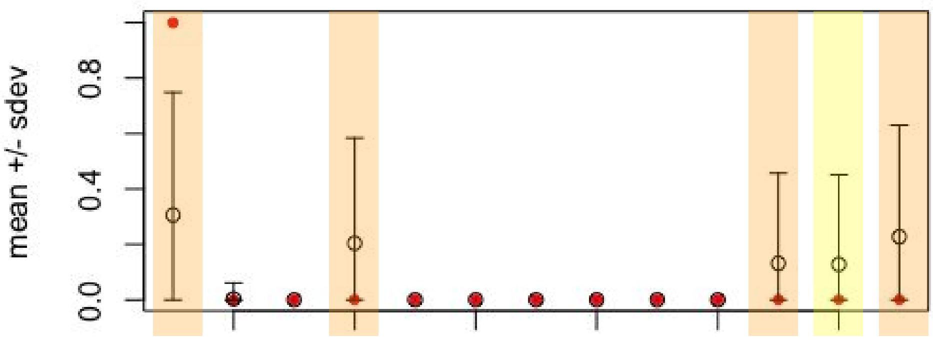
Node 21 has abnormally low visit frequency to gym/fitness centers

Snoopy - Location Tracking

Mapper algorithm reveals many more clusters. Moreover, we can see the paths by which they are connected.



red=in node 26, black=out of node 26



Node 21 has not only abnormally low visit frequency to gym/fitness centers, but also to athletic fields, golf, and outdoors. However, it has abnormally high visit frequency to stadiums/arenas.

Questions? :)