Dependency diagnostic: Visually understanding pairwise variable relationships

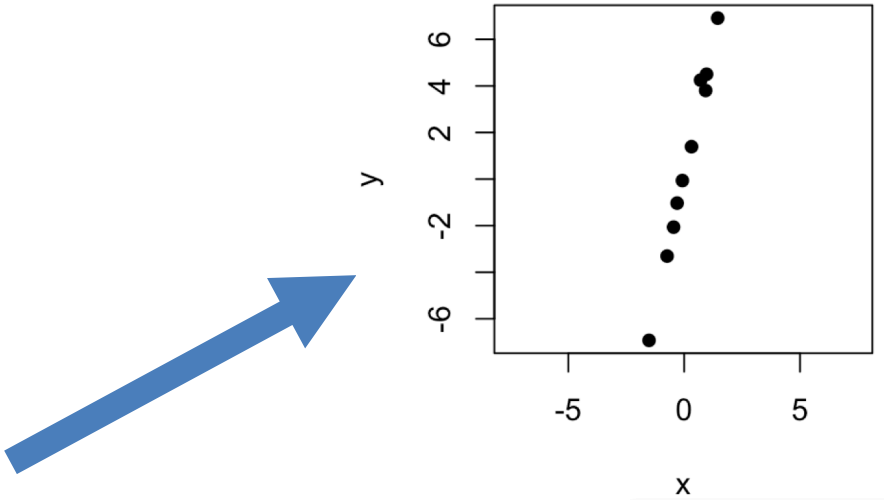
Kevin Lin
Carnegie Mellon University,
Department of Statistics and Data Science
@linnylin92

A large portion of statistics involves understanding the relationship between variables.

	x	y
1	0.92833370	3.8042839
2	-0.30295657	-1.0313626
3	-0.45507998	-2.0610312
4	1.44913302	6.9135100
5	0.96801747	4.5033052
6	0.31809710	1.3910063
7	-0.07636837	-0.0629351
8	-1.51404029	-6.9245794
9	0.70839776	4.2499442
10	-0.73403137	-3.3051033

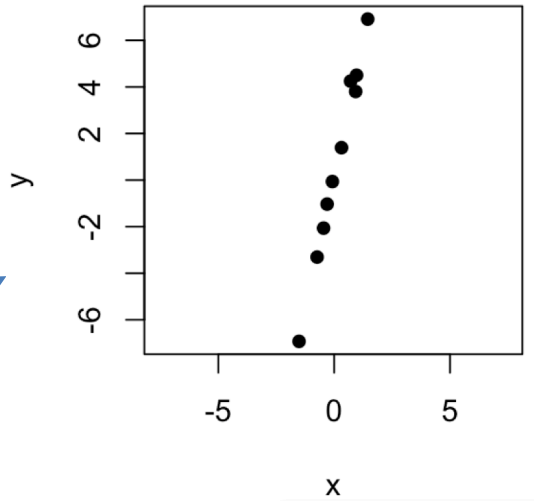
A large portion of statistics involves understanding the relationship between variables.

	x	y
1	0.92833370	3.8042839
2	-0.30295657	-1.0313626
3	-0.45507998	-2.0610312
4	1.44913302	6.9135100
5	0.96801747	4.5033052
6	0.31809710	1.3910063
7	-0.07636837	-0.0629351
8	-1.51404029	-6.9245794
9	0.70839776	4.2499442
10	-0.73403137	-3.3051033



A large portion of statistics involves understanding the relationship between variables.

	x	y
1	0.92833370	3.8042839
2	-0.30295657	-1.0313626
3	-0.45507998	-2.0610312
4	1.44913302	6.9135100
5	0.96801747	4.5033052
6	0.31809710	1.3910063
7	-0.07636837	-0.0629351
8	-1.51404029	-6.9245794
9	0.70839776	4.2499442
10	-0.73403137	-3.3051033



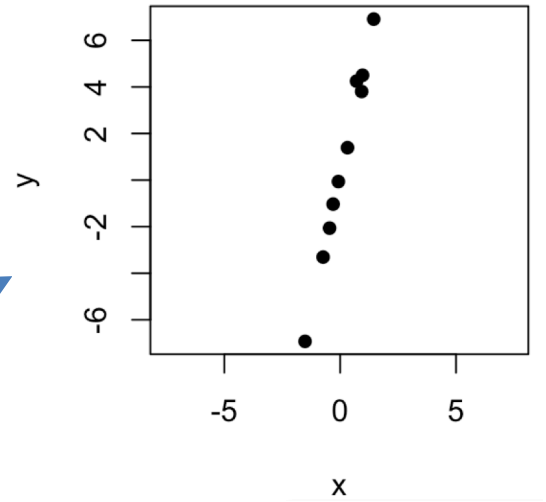

```
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 32.026, df = 8, p-value = 9.84e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9830525 0.9991175
sample estimates:
      cor
0.9961229
```

A large portion of statistics involves understanding the relationship between variables.

	x	y
1	0.92833370	3.8042839
2	-0.30295657	-1.0313626
3	-0.45507998	-2.0610312
4	1.44913302	6.9135100
5	0.96801747	4.5033052
6	0.31809710	1.3910063
7	-0.07636837	-0.0629351
8	-1.51404029	-6.9245794
9	0.70839776	4.2499442
10	-0.73403137	-3.3051033



Visual

```
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = 32.026, df = 8, p-value = 9.84e-10
alternative hypothesis: true correlation is not equa
95 percent confidence interval:
 0.9830525 0.9991175
sample estimates:
      cor
0.9961229
```

Numerical

There's quite a lot of appeal of using numerical methods over visual methods…

- Visual methods are inherently two-dimensional

There's quite a lot of appeal of using numerical methods over visual methods…

- Visual methods are inherently two-dimensional
- Numerical methods seem more appealing for large data analyses
  - Give a principled way to do multivariate analyses

There's quite a lot of appeal of using numerical methods over visual methods…

- Visual methods are inherently two-dimensional

- Numerical methods seem more appealing for large data analyses
  - Give a principled way to do multivariate analyses
  - More concrete; not based on opinions
  - Not affected by the plotting illusions

… but no numerical method can serve as a general-purpose tool for any situation.

> "Traditional statistical tests are well studied, well-formulated and work best when data is well-behaved, following a known distribution in relatively simple scenarios. But ... traditional statistical tests do not cover all of the complexities that arise when exploring data." – Wickham et al. (2010)

… but no numerical method can serve as a general-purpose tool for any situation.

> "Traditional statistical tests are well studied, well-formulated and work best when data is well-behaved, following a known distribution in relatively simple scenarios. But ... traditional statistical tests do not cover all of the complexities that arise when exploring data." – Wickham et al. (2010)

- However, there are often too many things to plot naively

We build a system to help determine how to construct a pairwise dependency graph.

- We can use various numerical estimates of dependency to estimate a pairwise dependency graph.

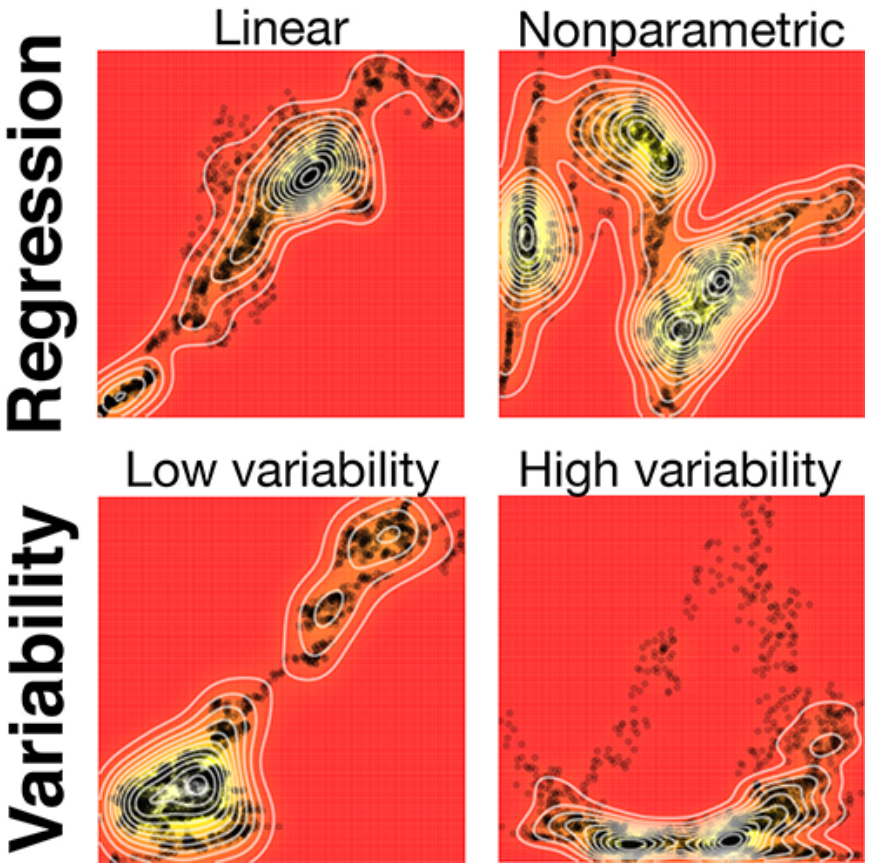
We build a system to help determine how to construct a pairwise dependency graph.

- We can use various numerical estimates of dependency to estimate a pairwise dependency graph.
- However, we want to use visualizations to diagnose this numerical method.
- But if there are too many variables, there are too many scatterplots to visualize.

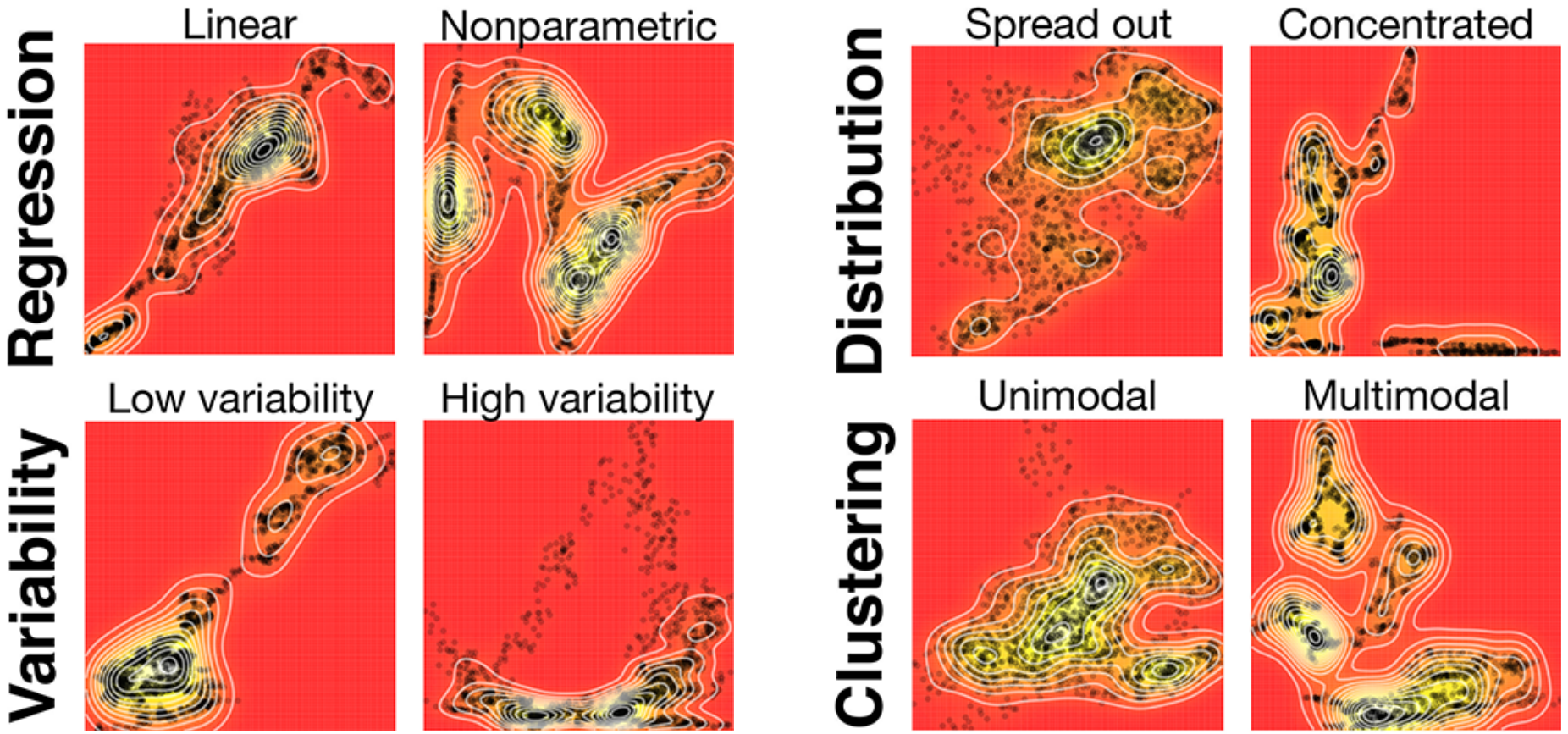
We build a system to help determine how to construct a pairwise dependency graph.

- We can use various numerical estimates of dependency to estimate a pairwise dependency graph.
- However, we want to use visualizations to diagnose this numerical method.
- But if there are too many variables, there are too many scatterplots to visualize.
- Solution: A system to learn what kind of patterns we're looking for.

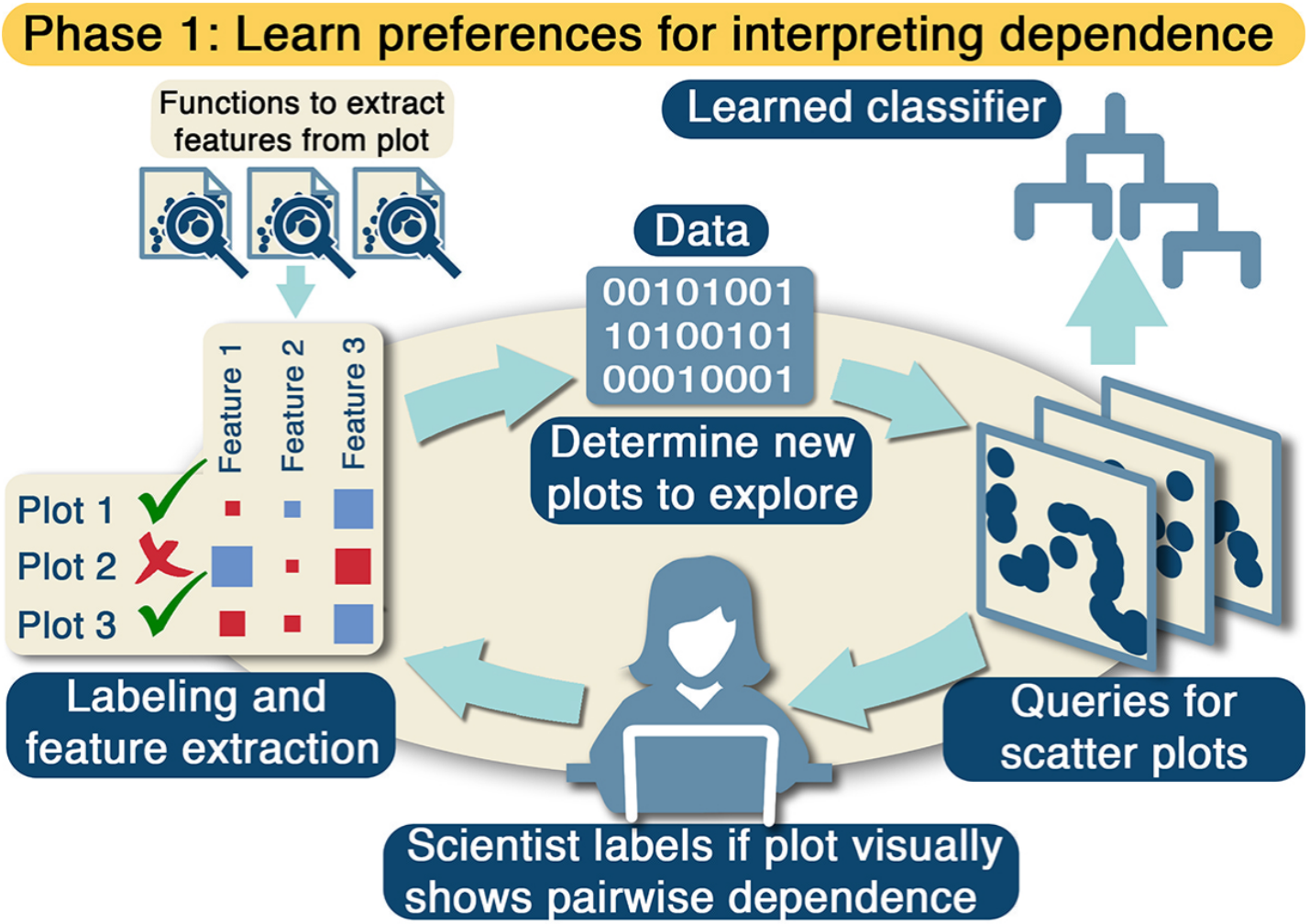
There are many types of patterns that can describe a scatterplot.



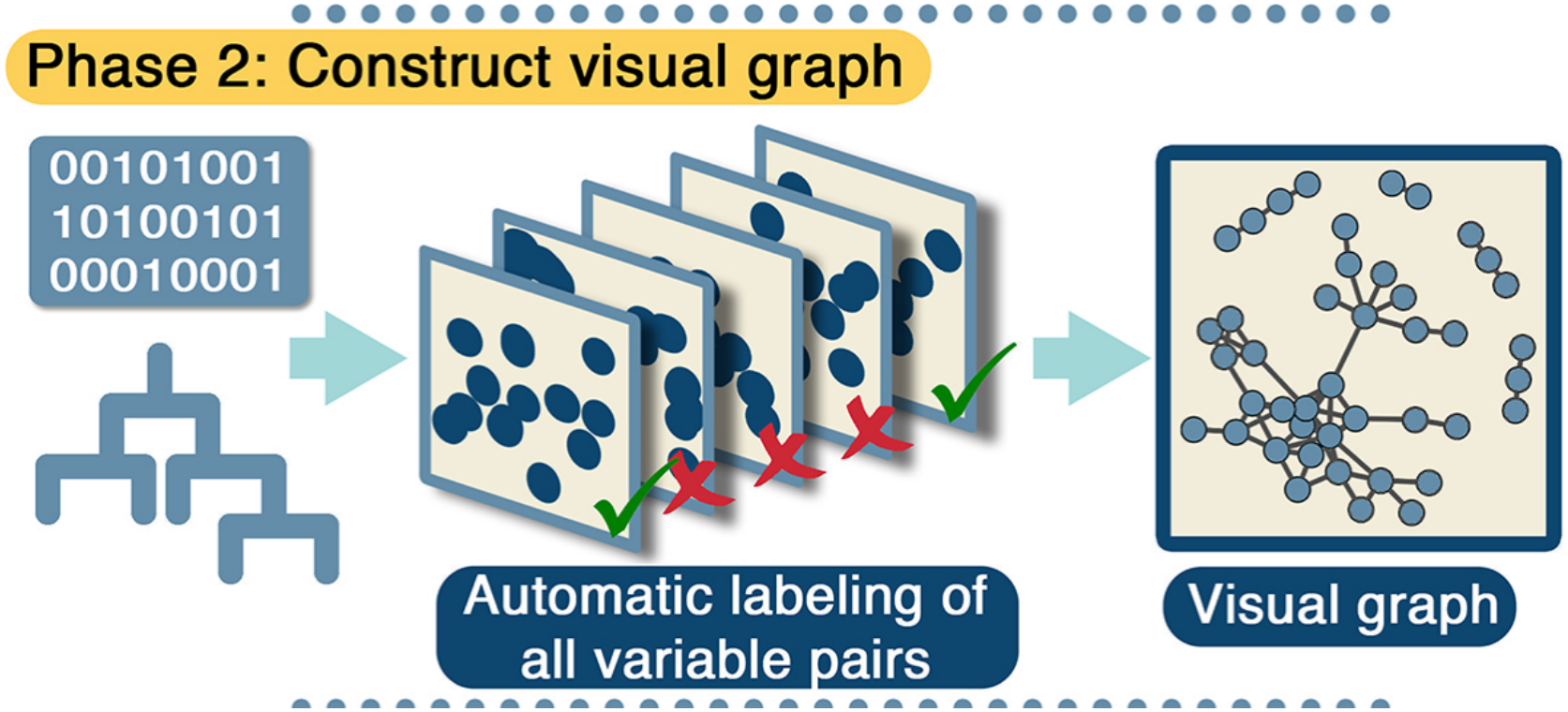
There are many types of patterns that can describe a scatterplot.



Flowchart of system

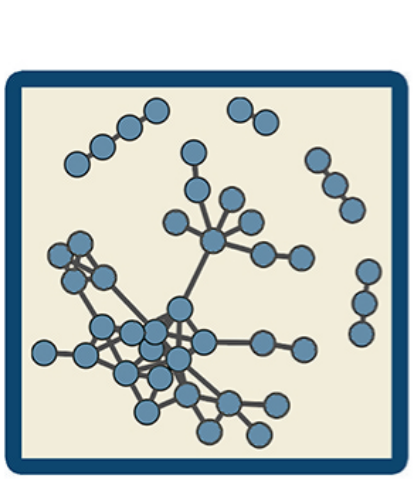


Flowchart of system

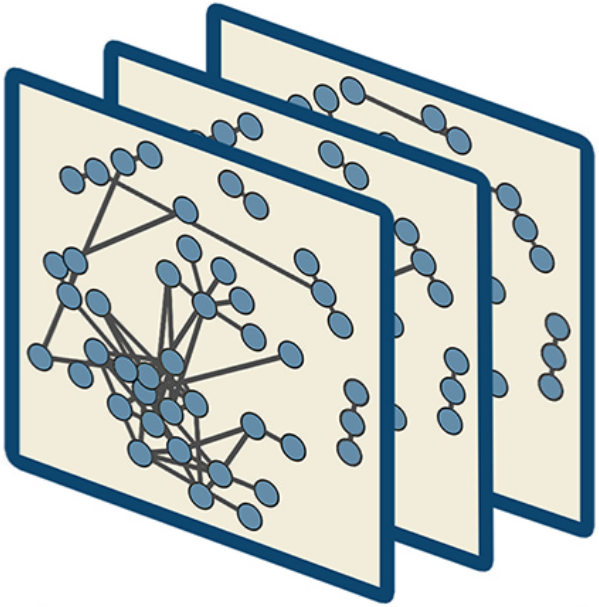


Flowchart of system

Phase 3: Compare visual and numerical graphs



vs.



Graphs estimated via numerical methods

The system advocates using Method 3 as the appropriate method for this dataset.

The engineering novelties are in designing a flexible system to handle numerical features and to do active learning.

- Design the system to be as interpretable as possible

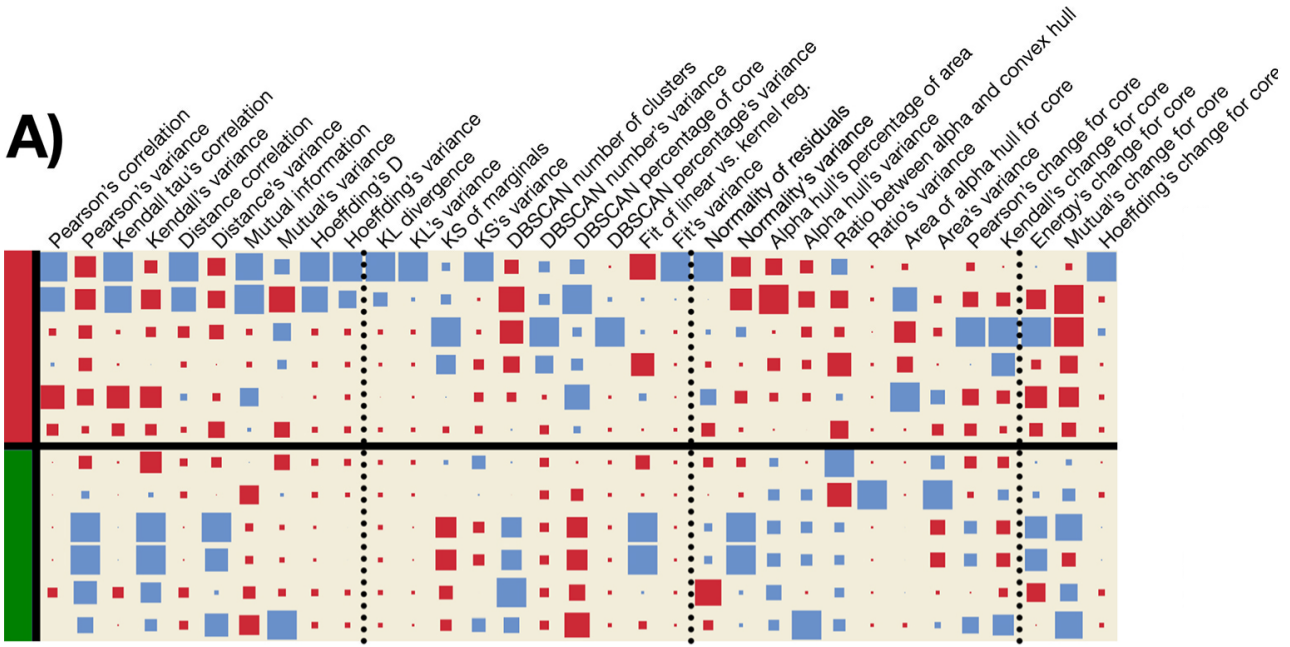
The engineering novelties are in designing a flexible system to handle numerical features and to do active learning.

- Design the system to be as interpretable as possible
- Numerical features: Many numerical ways to quantify different aspects of a scatter plot

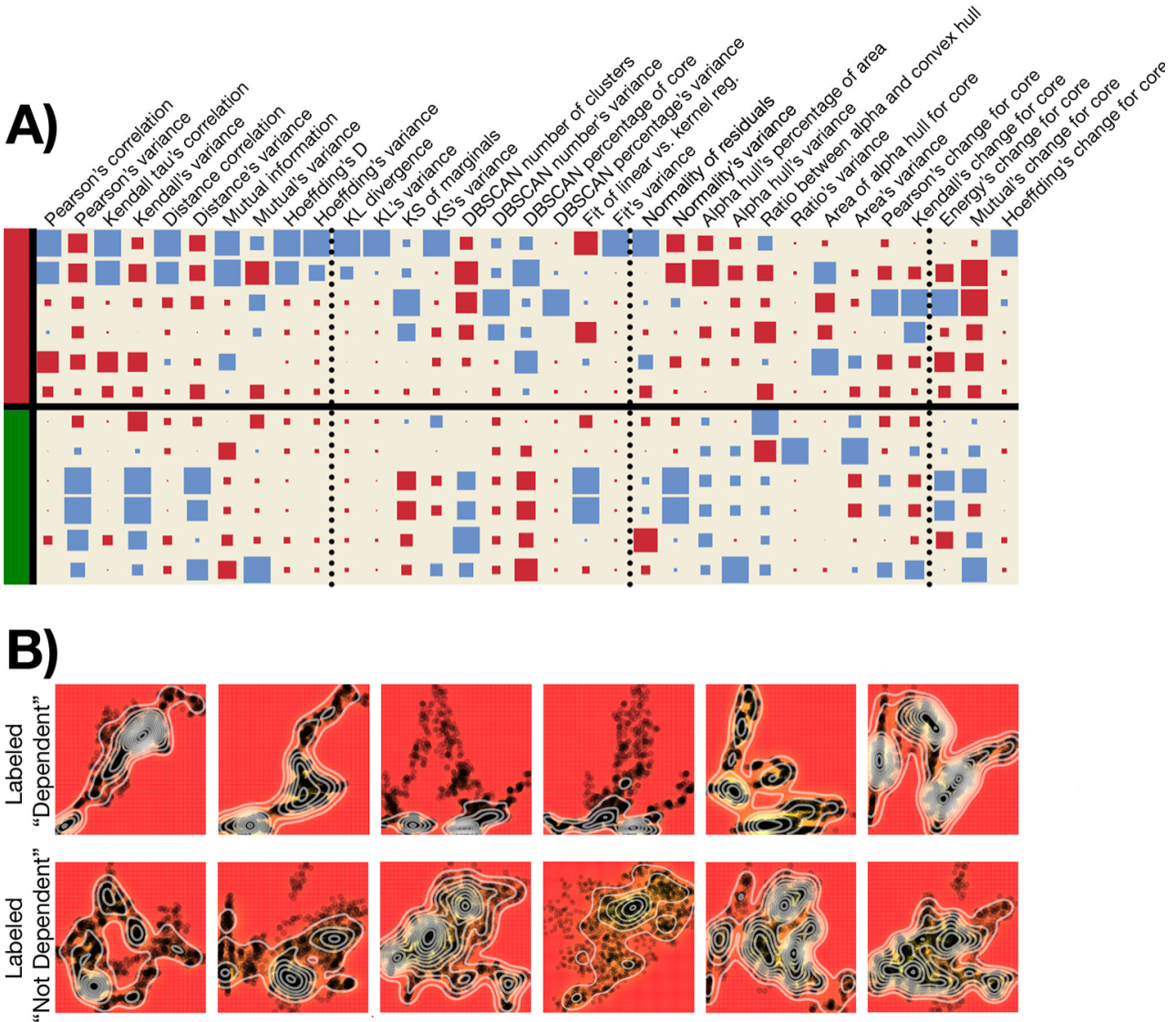
The engineering novelties are in designing a flexible system to handle numerical features and to do active learning.

- Design the system to be as interpretable as possible
- Numerical features: Many numerical ways to quantify different aspects of a scatter plot
- Active learning: System to ensure that as few scatter plot as possible need to be manually labeled

We can make other plots to assess the quality of the learned classifier.



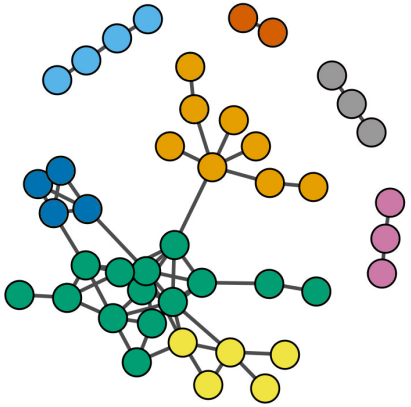
Avenue 3: Image recognition software to learn what patterns the user is interested in.



We apply our method for building a stock portfolio from 43 different healthcare companies.

- Data from 1998 to 2006

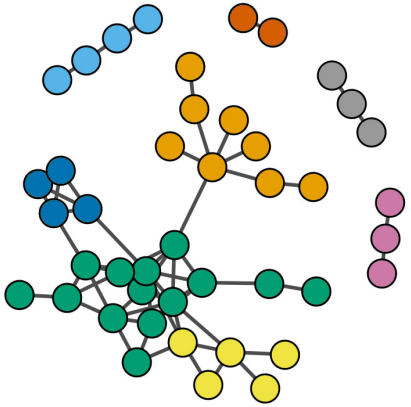
Visual graph



We apply our method for building a stock portfolio from 43 different healthcare companies.

- Data from 1998 to 2006
- Build a portfolio using stocks that are as independent of each other as possible, according to our pairwise dependency graph

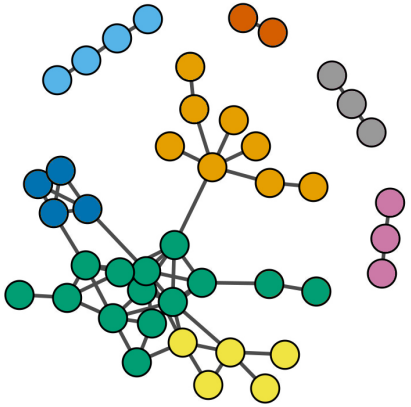
Visual graph



We apply our method for building a stock portfolio from 43 different healthcare companies.

- Data from 1998 to 2006
- Build a portfolio using stocks that are as independent of each other as possible, according to our pairwise dependency graph
- Construct metrics to compare visual and numeric graphs

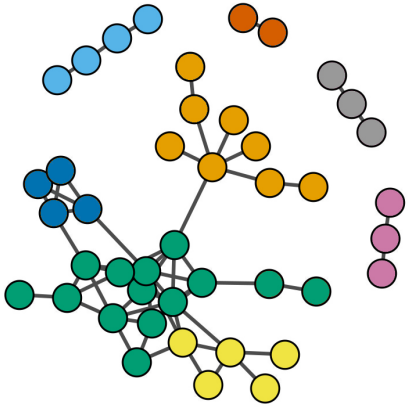
Visual graph



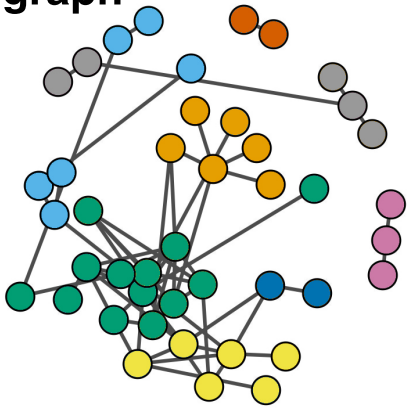
We apply our method for building a stock portfolio from 43 different healthcare companies.

- Data from 1998 to 2006
- Build a portfolio using stocks that are as independent of each other as possible, according to our pairwise dependency graph
- Construct metrics to compare visual and numeric graphs

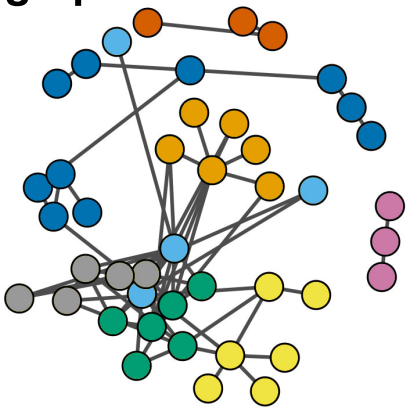
Visual graph



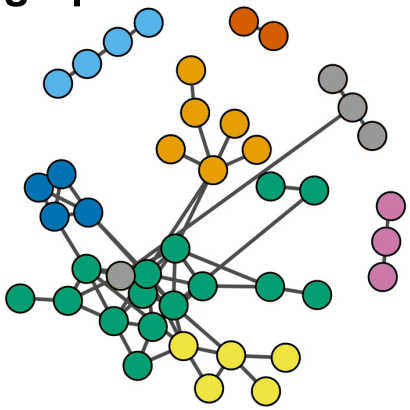
Pearson correlation graph



Distance correlation graph



Mutual information graph



Takeaway: We build a system to help users explore data for constructing more accurate pairwise dependency graphs.

- Strive for fully interpretable system

Takeaway: We build a system to help users explore data for constructing more accurate pairwise dependency graphs.

- Strive for fully interpretable system
- Users classify plots, system quantifies the plots behind-the-scenes and uses active learning to determine which new plots to show, then system labels the remaining plots

Takeaway: We build a system to help users explore data for constructing more accurate pairwise dependency graphs.

- Strive for fully interpretable system
- Users classify plots, system quantifies the plots behind-the-scenes and uses active learning to determine which new plots to show, then system labels the remaining plots

Kevin Lin
@linnylin92