Plotting as a valuable data analysis tool

Kevin Lin
@linnylin92

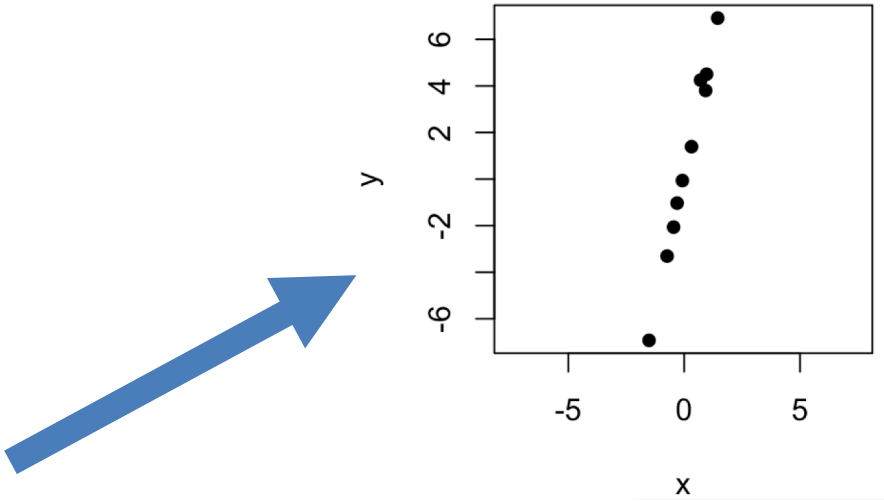
A large portion of statistics involves understanding the relationship between variables.

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

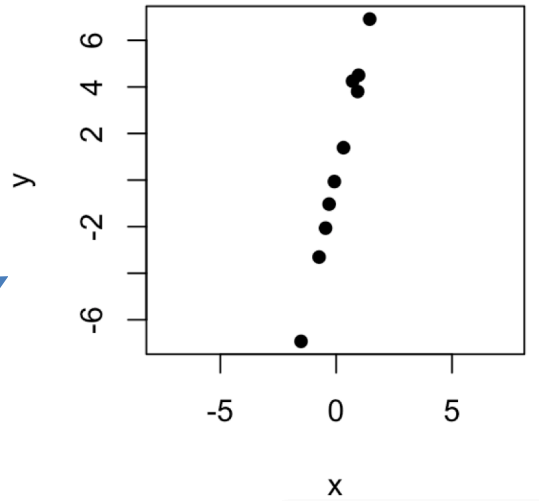
A large portion of statistics involves understanding the relationship between variables.

	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.

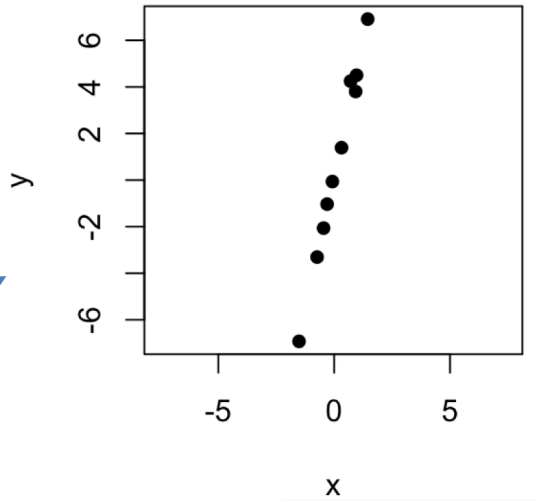
	x	y
1	0.92833370	3.8042839
2	-0.30295657	-1.0313626
3	-0.45507998	-2.0610312
4	1.44913302	6.9135100
5	0.96801747	4.5033052
6	0.31809710	1.3910063
7	-0.07636837	-0.0629351
8	-1.51404029	-6.9245794
9	0.70839776	4.2499442
10	-0.73403137	-3.3051033




```
> 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

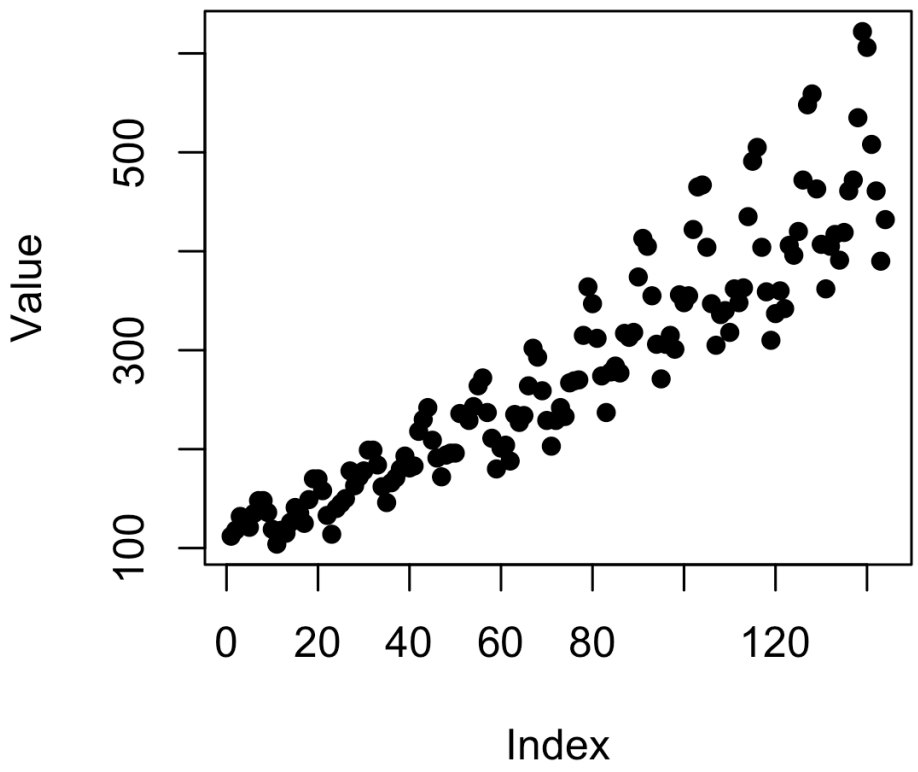


Visual

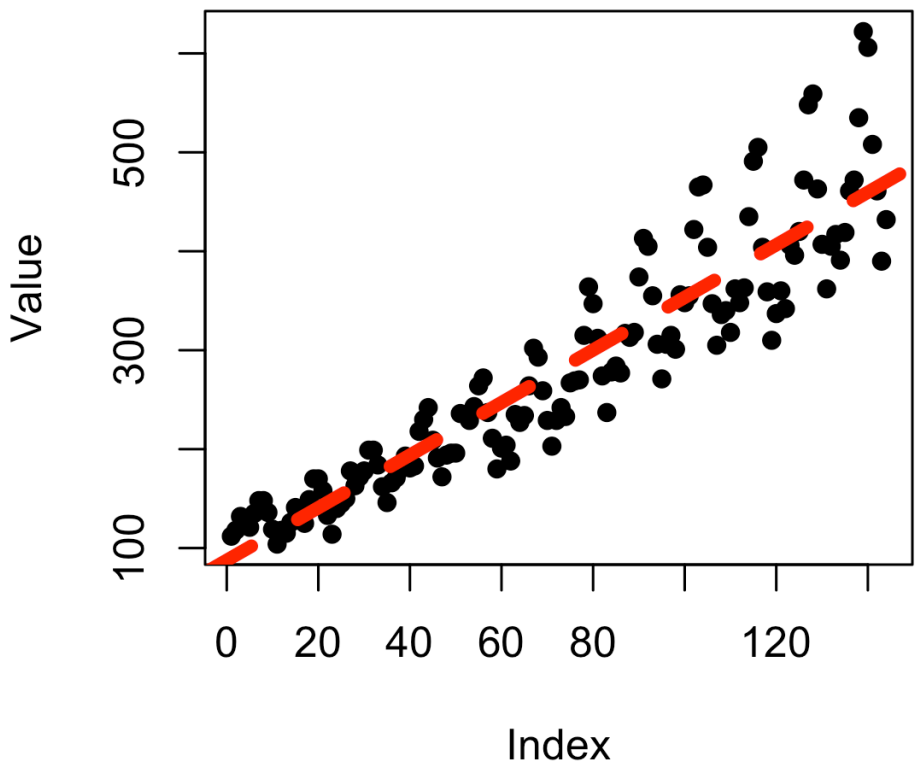
```
> cor.test(x,y)

        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

If all our data were only two-dimensional, then we would not need a lot of statistics.



If all our data were only two-dimensional, then we would not need a lot of statistics.



If all our data were only two-dimensional, then we would not need a lot of statistics.

- Visual methods are inherently two-dimensional

If all our data were only two-dimensional, then we would not need a lot of statistics.

- 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

If all our data were only two-dimensional, then we would not need a lot of statistics.

- 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

If all our data were only two-dimensional, then we would not need a lot of statistics.

- 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
  - Can seem quite fancy/impressive

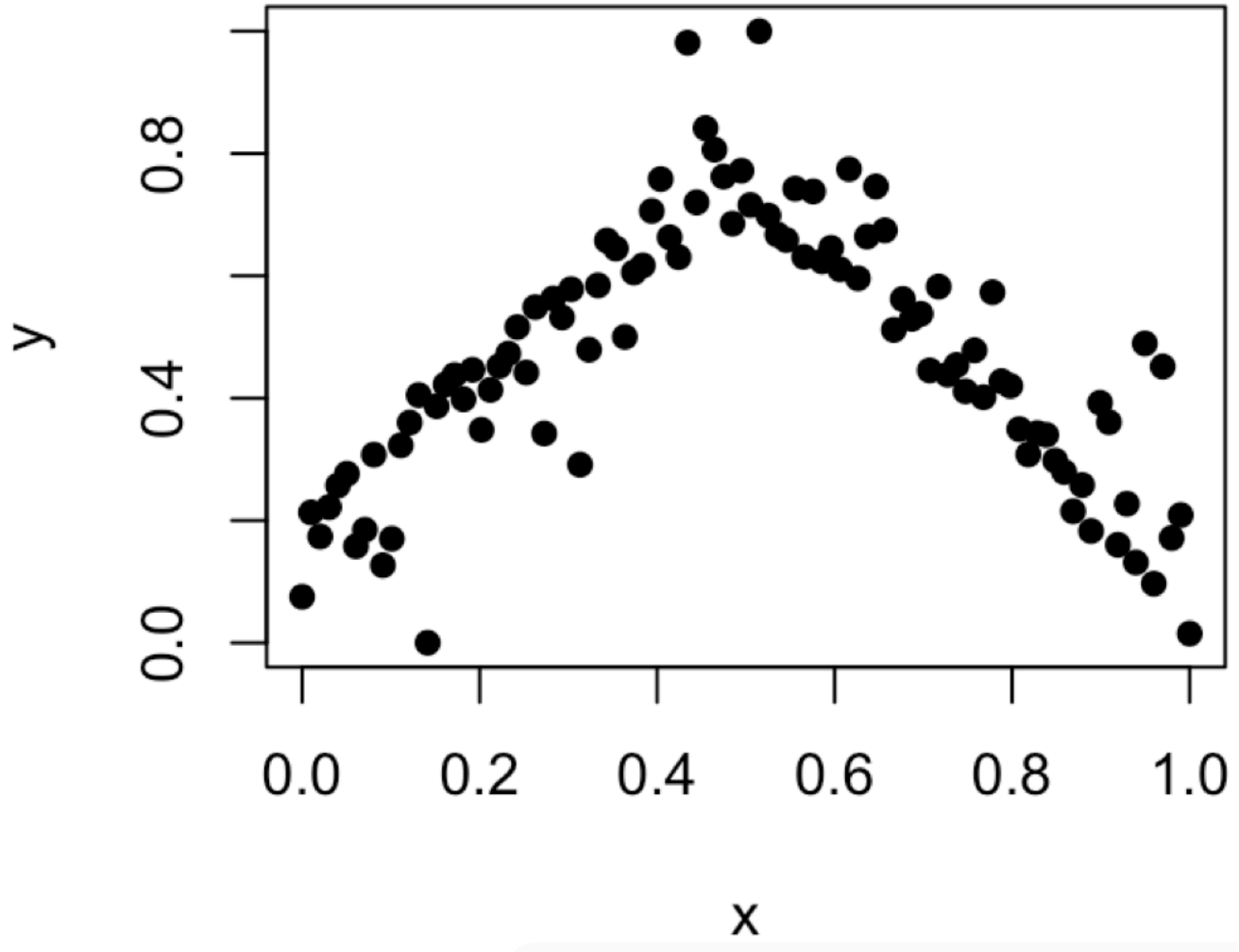
However, numerical methods can be misleading for “most real data.”

```
> summary(fitlm)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46205 -0.15543 -0.01442  0.16566  0.53487

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.460878   0.042622  10.813   <2e-16 ***
x           0.008261   0.073637   0.112    0.911
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2147 on 98 degrees of freedom
Multiple R-squared:  0.0001284, Adjusted R-squared:  -0.01007
F-statistic: 0.01258 on 1 and 98 DF,  p-value: 0.9109
```

However, numerical methods can be misleading for “most real data.”

```
> shapiro.test(fitlm$residuals)

        Shapiro-Wilk normality test

data:  fitlm$residuals
W = 0.98894, p-value = 0.5795
```

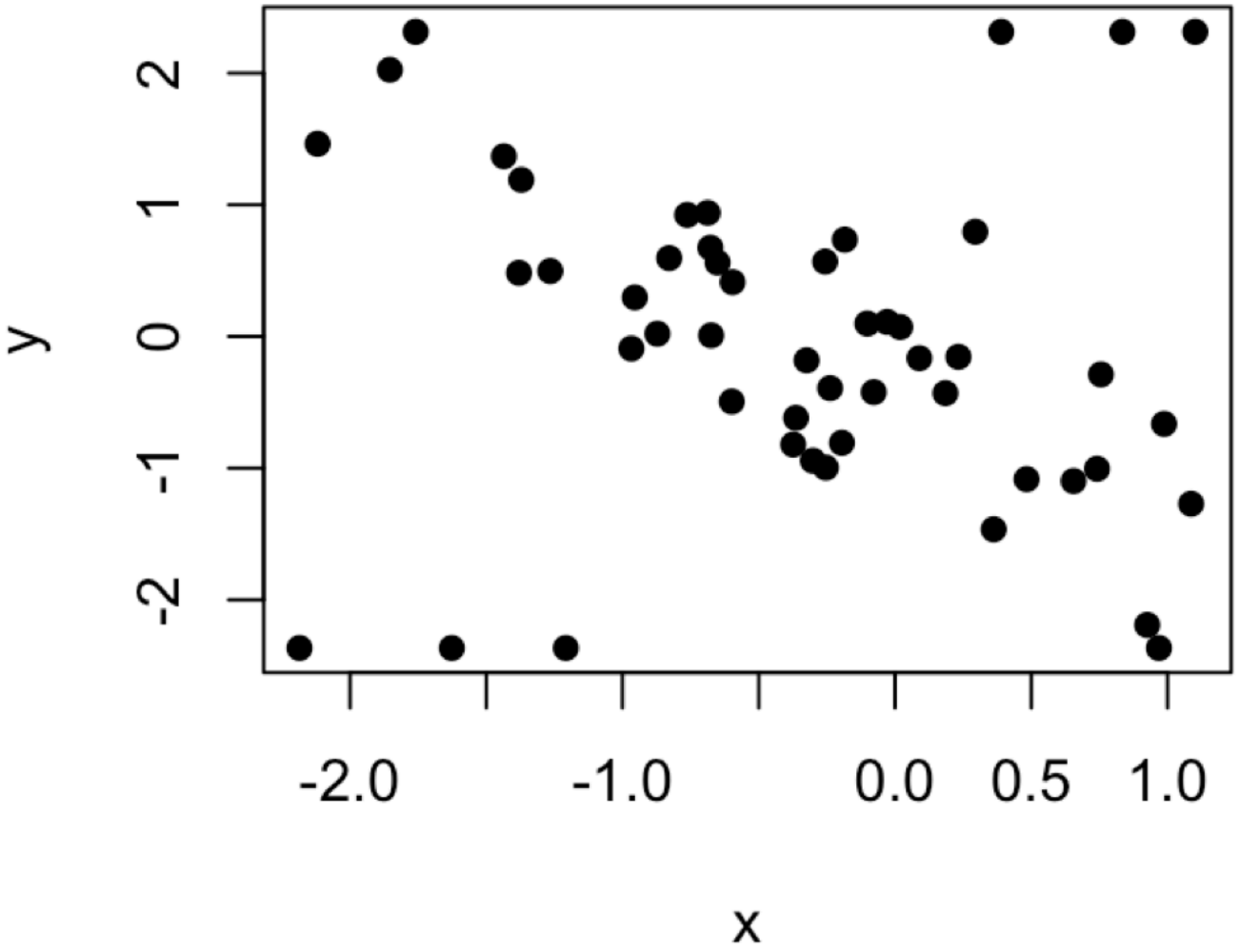
However, numerical methods can be misleading for “most real data.”



However, numerical methods can be misleading for “most real data.”

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

        Pearson's product-moment correlation

data:  x and y
t = -1.3305, df = 48, p-value = 0.1896
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.44365768  0.09471973
sample estimates:
       cor
-0.1886004
```

However, numerical methods can be misleading for “most real data.”



However, numerical methods can be misleading for “most real data.”

- Of course, you can argue that “had you known these particular numerical methods were not enough, you could’ve used better numerical methods.” But how would’ve have known this without plotting?

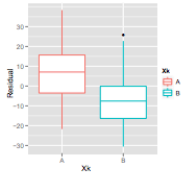
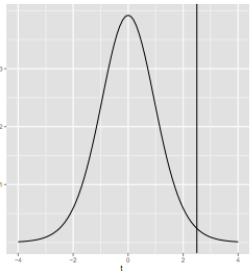
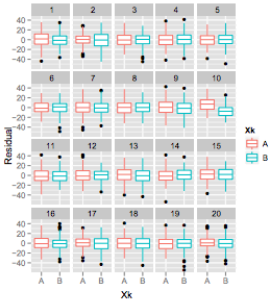
However, numerical methods can be misleading for “most real data.”

- Of course, you can argue that “had you known these particular numerical methods were not enough, you could’ve used better numerical methods.” But how would’ve have known this without plotting?

- In multivariate analyses, numerical methods become more fragile. Hence, it is even *more* important to think of how to plot data.

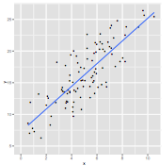
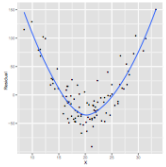
In fact, some research is done on showing people can assess plots better than hypothesis tests.

Table 1. Comparison of visual inference with existing inference technique

	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$	$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$
Test statistic	$T(y) = \frac{\hat{\beta}}{se(\hat{\beta})}$	$T(y) =$ 
Null Distribution	$f_{T(y)}(t);$ 	$f_{T(y)}(t);$ 
Reject H_0 if	observed T is extreme	observed plot is identifiable

From: “Validation of Visual Statistical Inference, with Application to Linear Models”, 2013

In fact, some research is done on showing people can assess plots better than hypothesis tests.

Case	Null Hypothesis	Statistic	Test Statistic	Description
1	$H_0 : \beta_0 = 0$	Scatter plot		Scatter plot with least square line overlaid. For lineup plot, we simulate data from fitted null model.
5	$H_0 : X$ Linear	Residual Plot		Residual vs predictor plots with loess smoother overlaid. For lineup plot, we simulate residual data from normal with mean 0 variance $\hat{\sigma}^2$.

From: “Validation of Visual Statistical Inference, with Application to Linear Models”, 2013

In fact, some research is done on showing people can assess plots better than hypothesis tests.

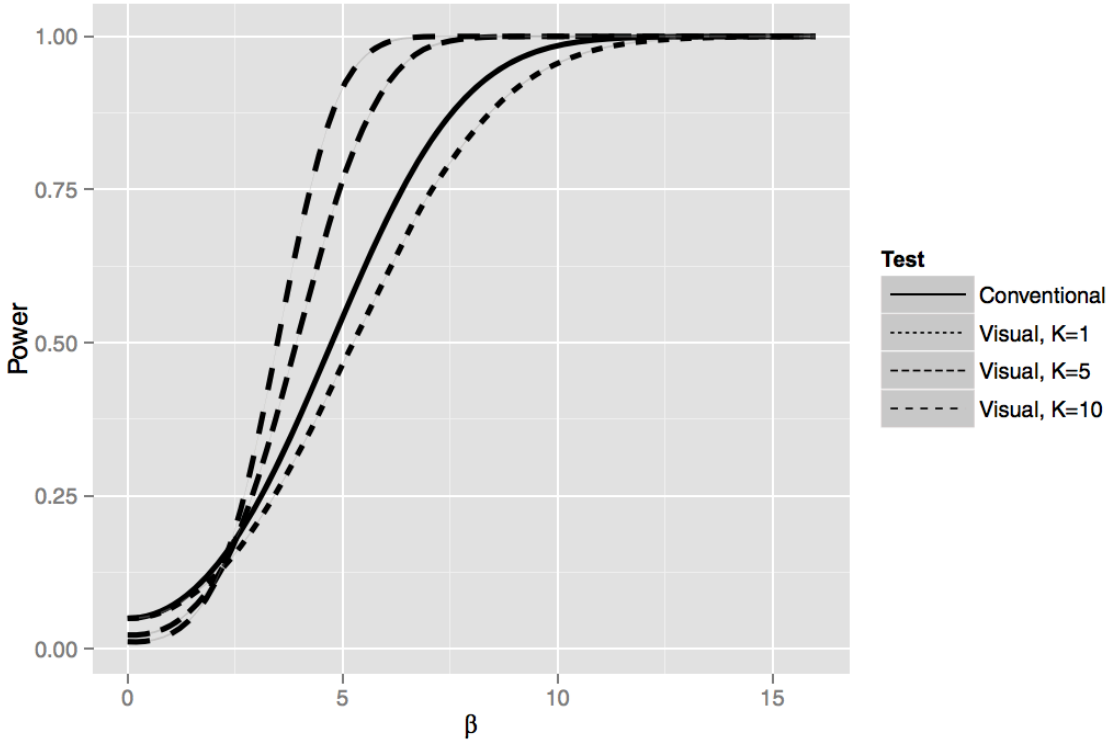


Figure 3. Comparison of the expected power of a visual test of size $m = 20$ for different K (number of observers) with the power of the conventional test, for $n = 100$ and $\sigma = 12$.

From: “Validation of Visual Statistical Inference, with Application to Linear Models”, 2013

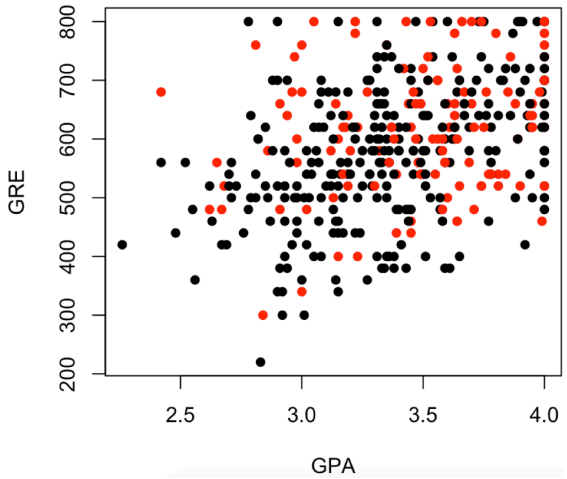
In fact, some research is done on showing people can assess plots better than hypothesis tests.

Statistical graphics play a crucial role in exploratory data analysis, model checking, and diagnosis. The lineup protocol enables statistical significance testing of visual findings, bridging the gulf between exploratory and inferential statistics. In this article, inferential methods for statistical graphics are developed further by refining the terminology of visual inference and framing the lineup protocol in a context that allows direct comparison with conventional tests in scenarios when a conventional test exists. This framework is used to compare the performance of the lineup protocol against conventional statistical testing in the scenario of fitting linear models. A human subjects experiment is conducted using simulated data to provide controlled conditions. Results suggest that the lineup protocol performs comparably with the conventional tests, and expectedly outperforms them when data are contaminated, a scenario where assumptions required for performing a conventional test are violated. Surprisingly, visual tests have higher power than the conventional tests when the effect size is large. And, interestingly, there may be some super-visual individuals who yield better performance and power than the conventional test even in the most difficult tasks. Supplementary materials for this article are available online.

From: “Validation of Visual Statistical Inference, with Application to Linear Models”, 2013

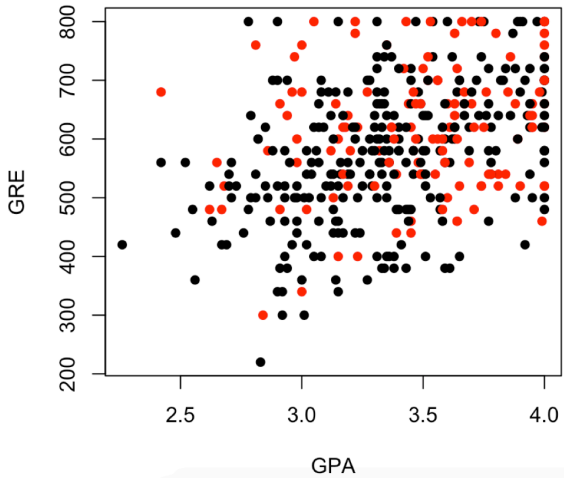
Plotting occurs at different “proximities” to the data.

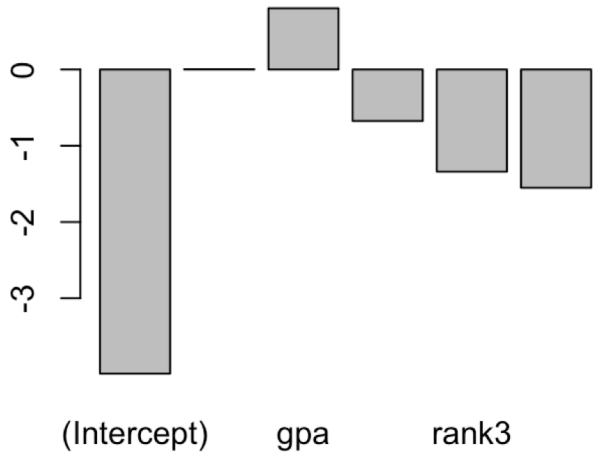
Data



Plotting occurs at different “proximities” to the data.

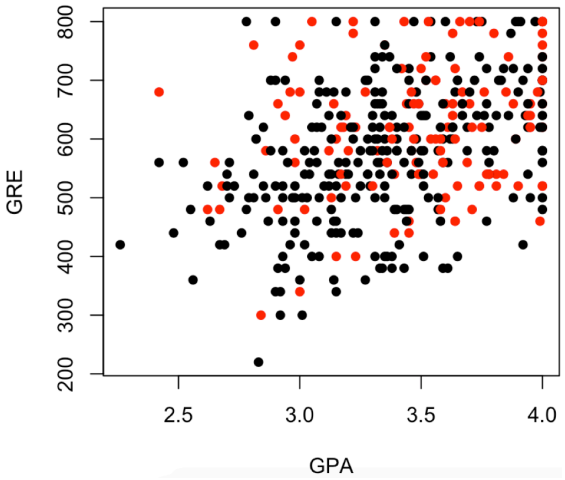
Data → Estimator

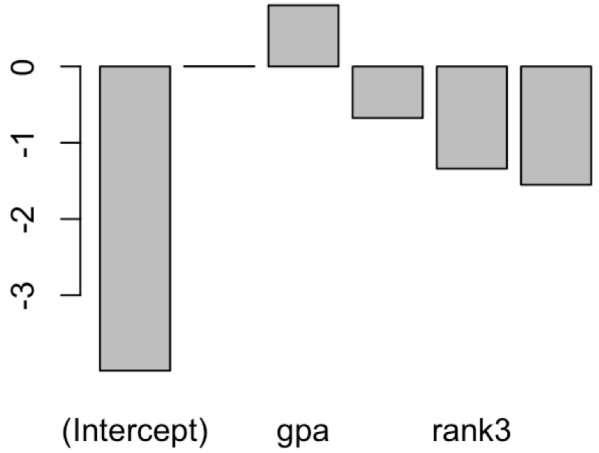


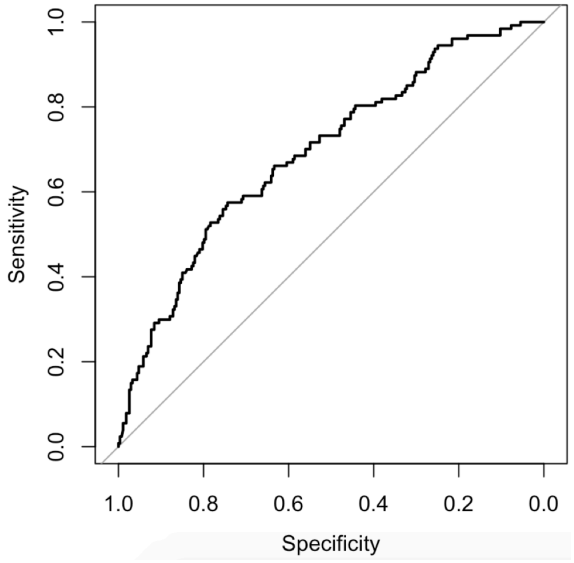


Plotting occurs at different “proximities” to the data.

Data → Estimator → Calibration → …

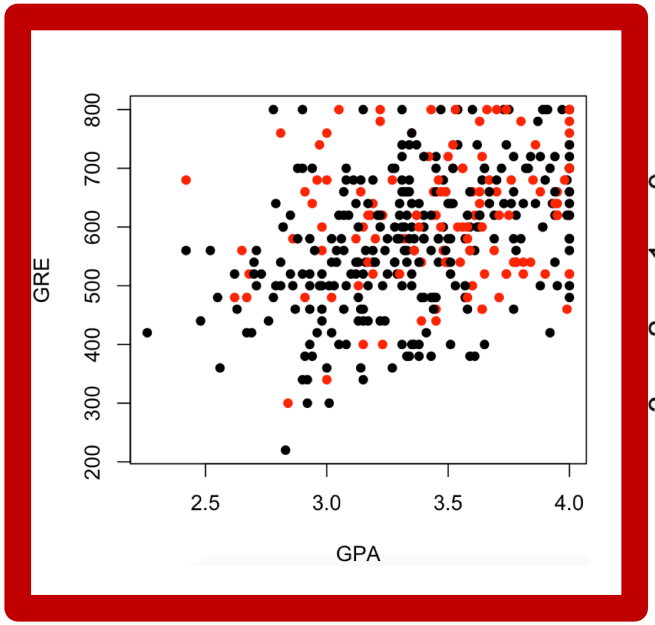


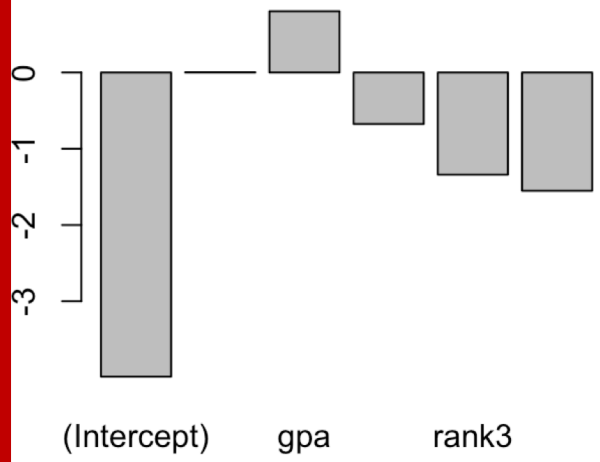


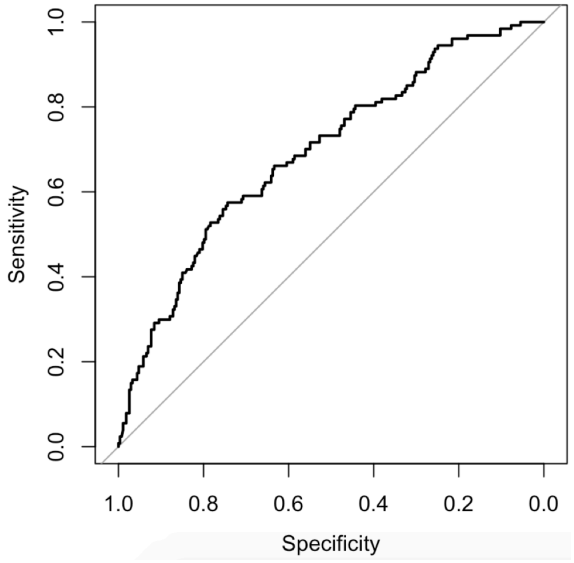


Plotting occurs at different “proximities” to the data.

Data → Estimator → Calibration → …







- Many systems have been developed to overcome high dimensions.

Avenue 1: Dimension reductions offer ways to embed multivariate data into 2 dimensions.

- Goal: Represent (embed) a multivariate dataset (i.e., many variables) on a two-dimensional plane

Avenue 1: Dimension reductions offer ways to embed multivariate data into 2 dimensions.

- Goal: Represent (embed) a multivariate dataset (i.e., many variables) on a two-dimensional plane

- Although there is a loss of information, we want to "preserve" the most "important" aspects of our dataset to the best of our abilities

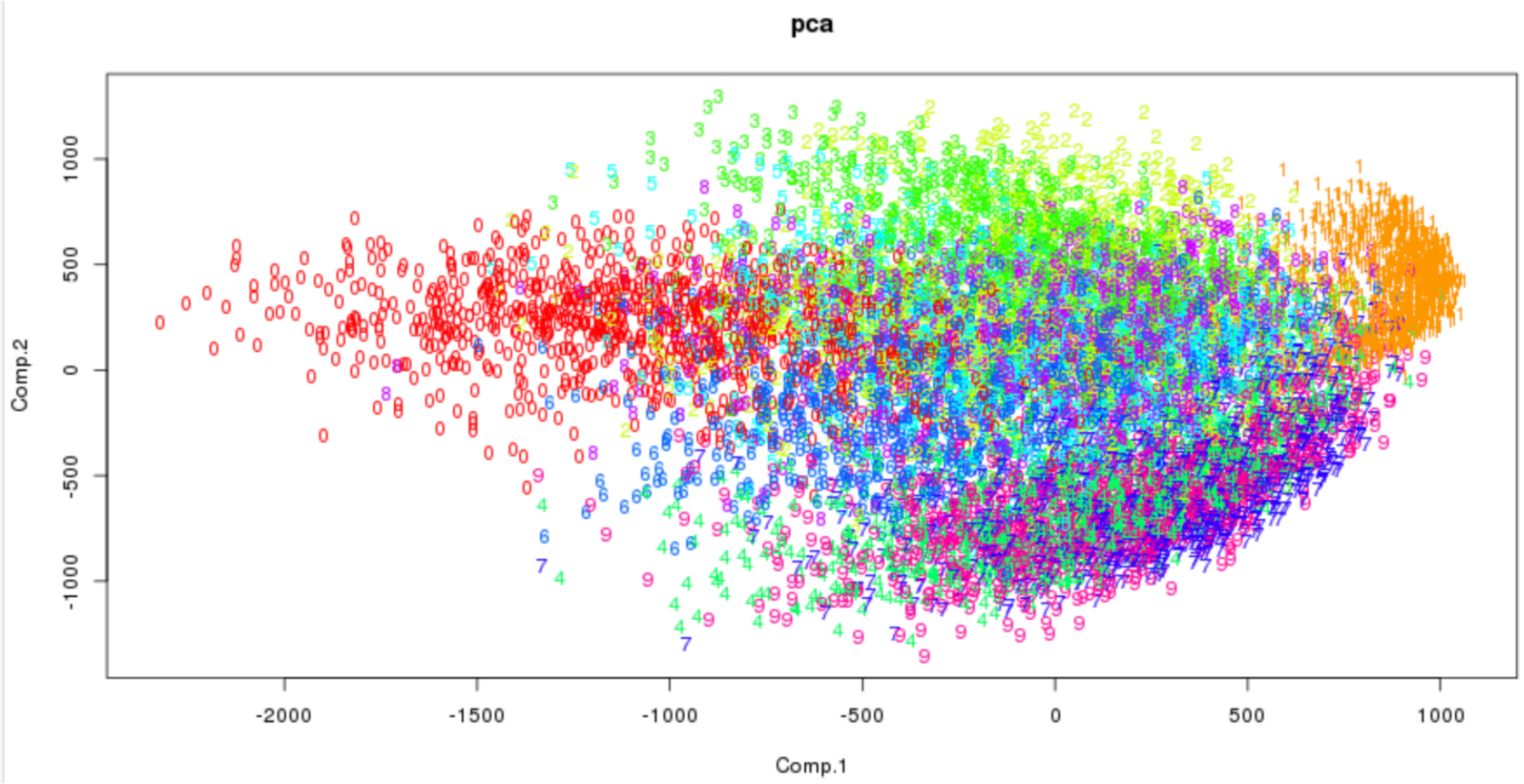
Avenue 1: Dimension reductions offer ways to embed multivariate data into 2 dimensions.

- Goal: Represent (embed) a multivariate dataset (i.e., many variables) on a two-dimensional plane

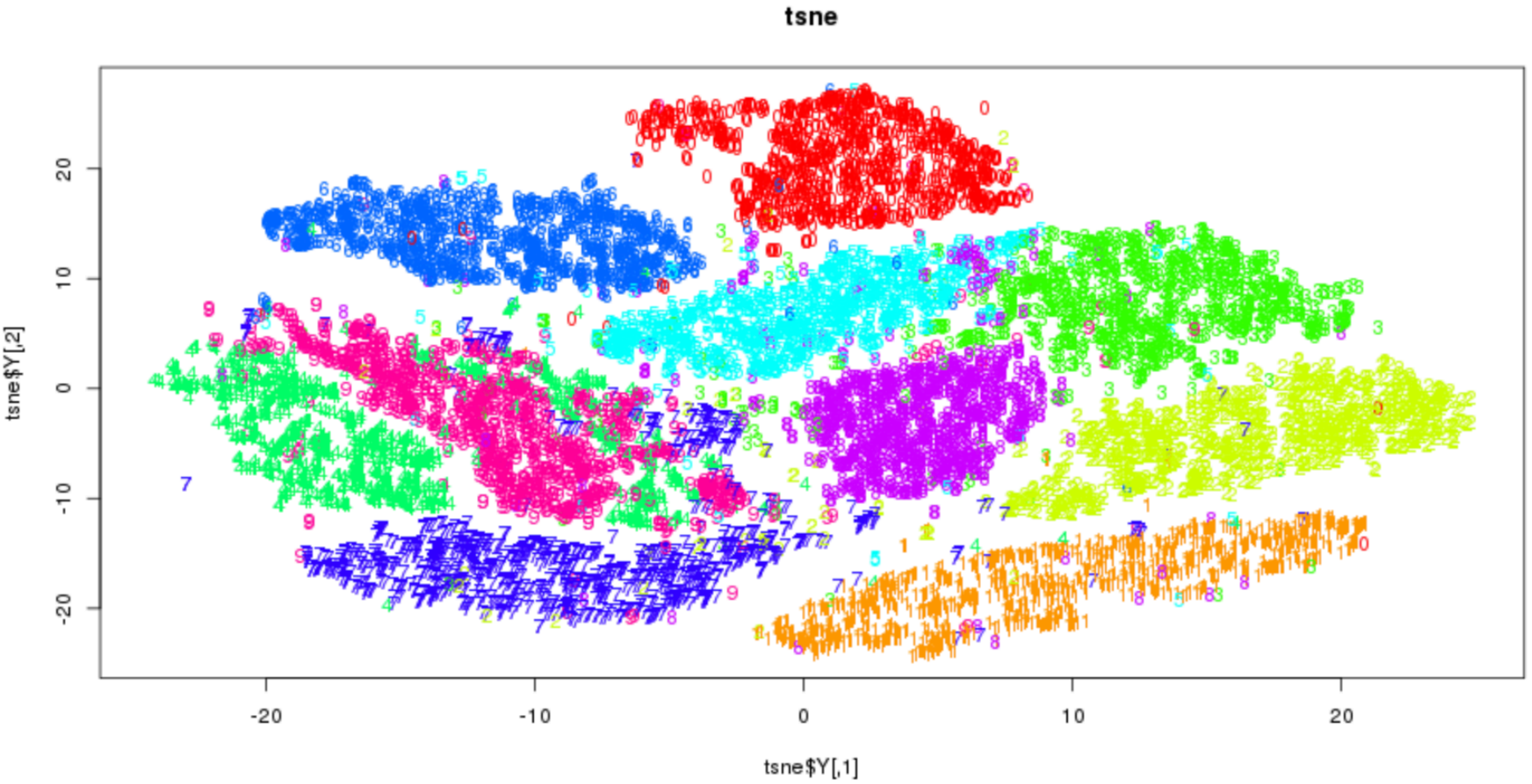
- Although there is a loss of information, we want to "preserve" the most "important" aspects of our dataset to the best of our abilities
  - Principal component analysis (PCA): Linear method to preserve as much variability
  - t-distributed stochastic neighbor embedding (t-SNE): Nonlinear method to minimize distortion in distribution

Avenue 1: Dimension reductions offer ways to embed multivariate data into 2 dimensions.

Avenue 1: Dimension reductions offer ways to embed multivariate data into 2 dimensions.



Avenue 1: Dimension reductions offer ways to embed multivariate data into 2 dimensions.



See: https://ai.googleblog.com/2016/12/open-sourcing-embedding-projector-tool.html

Avenue 2: A series of projections can offer a tour of the dataset via animation.

- What if we're not happy with just one projection?

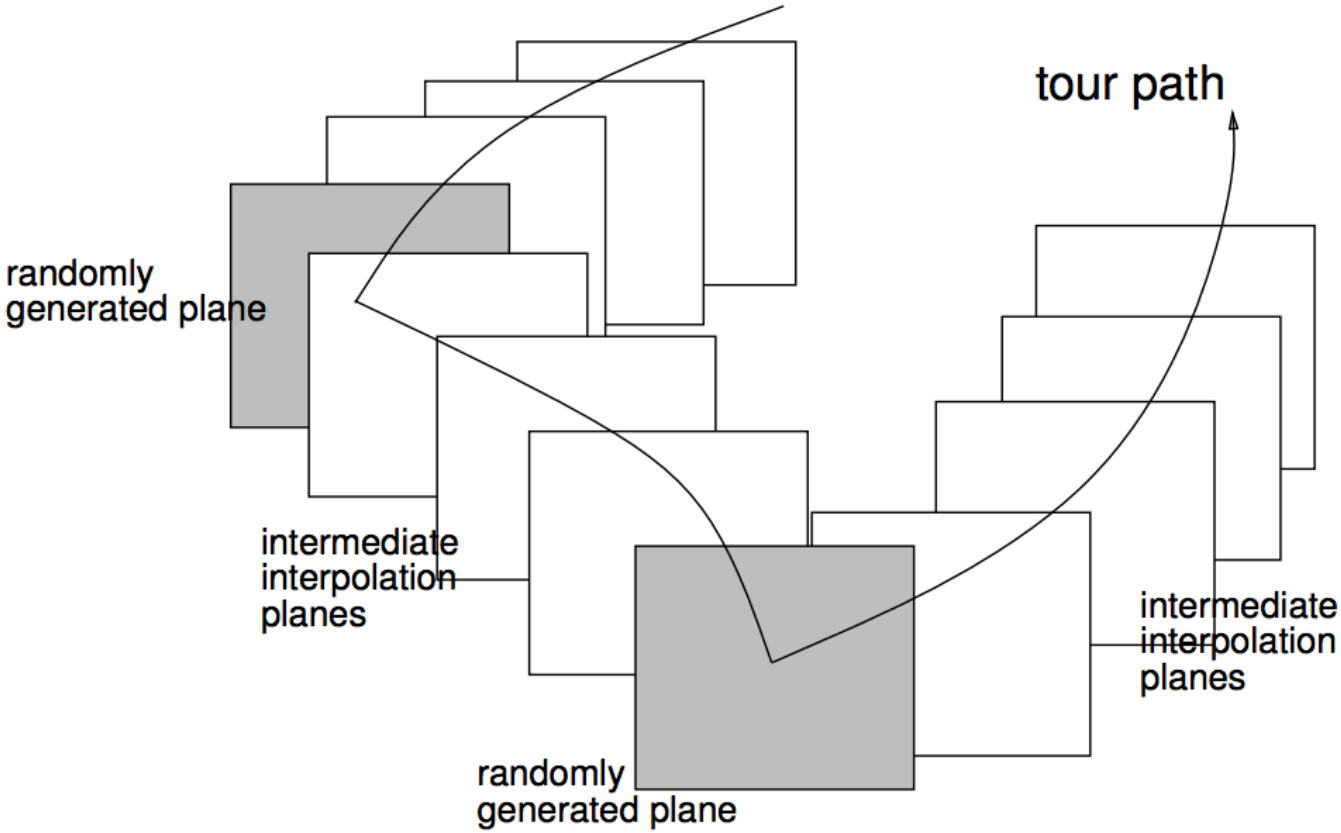
Avenue 2: A series of projections can offer a tour of the dataset via animation.

- What if we're not happy with just one projection?
- Tour: a "continuous" sequence of projections of the dataset into 2-dimensions, so we can animate these plots like a movie.

Avenue 2: A series of projections can offer a tour of the dataset via animation.

- What if we're not happy with just one projection?
- Tour: a "continuous" sequence of projections of the dataset into 2-dimensions, so we can animate these plots like a movie.
- The plots "look like they're rotating in 3D", but this works for any dimension.

Avenue 2: A series of projections can offer a tour of the dataset via animation.



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

- If we're interested in the variables themselves, not embeddings, we can use scatterplots.

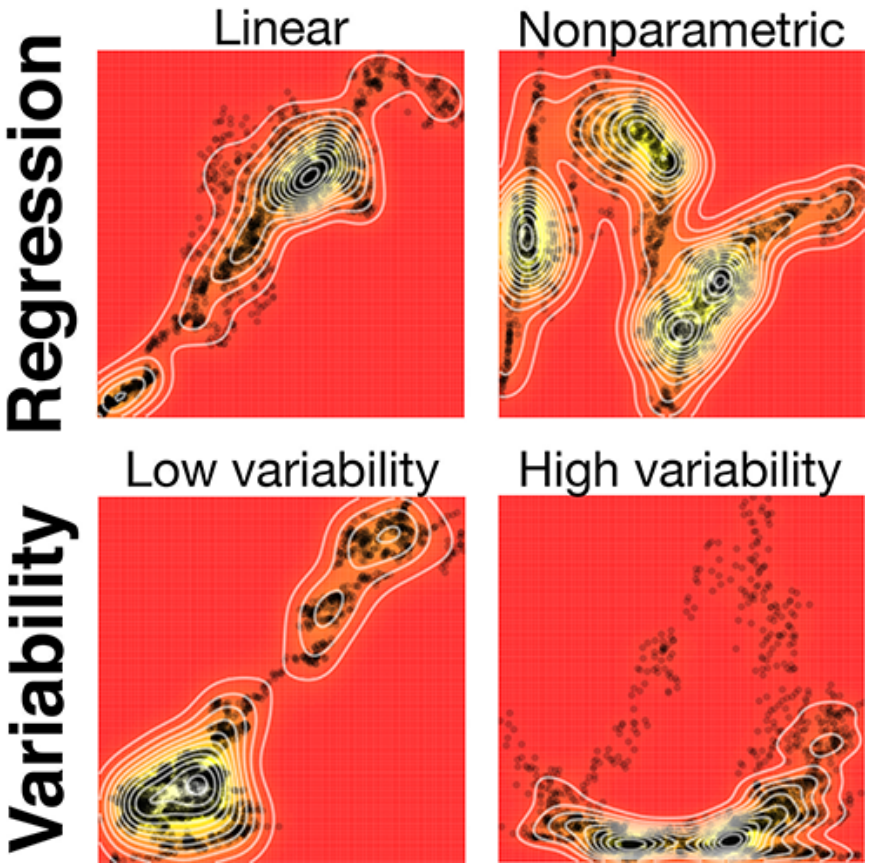
Avenue 3: Image recognition software to learn what patterns the user is interested in.

- If we're interested in the variables themselves, not embeddings, we can use scatterplots.
- But if there are too many variables, there are too many scatterplots to visualize.

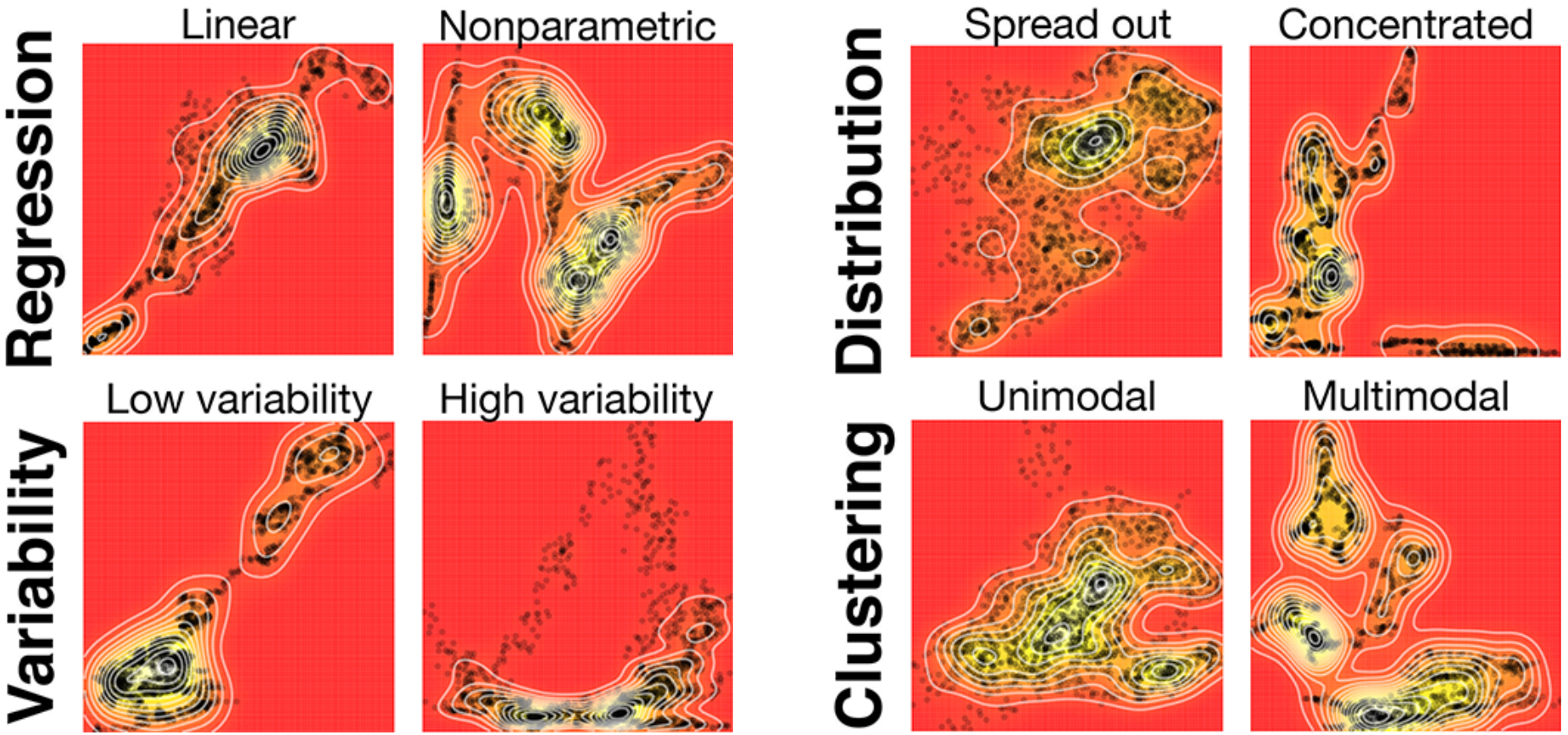
Avenue 3: Image recognition software to learn what patterns the user is interested in.

- If we're interested in the variables themselves, not embeddings, we can use scatterplots.
- 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.

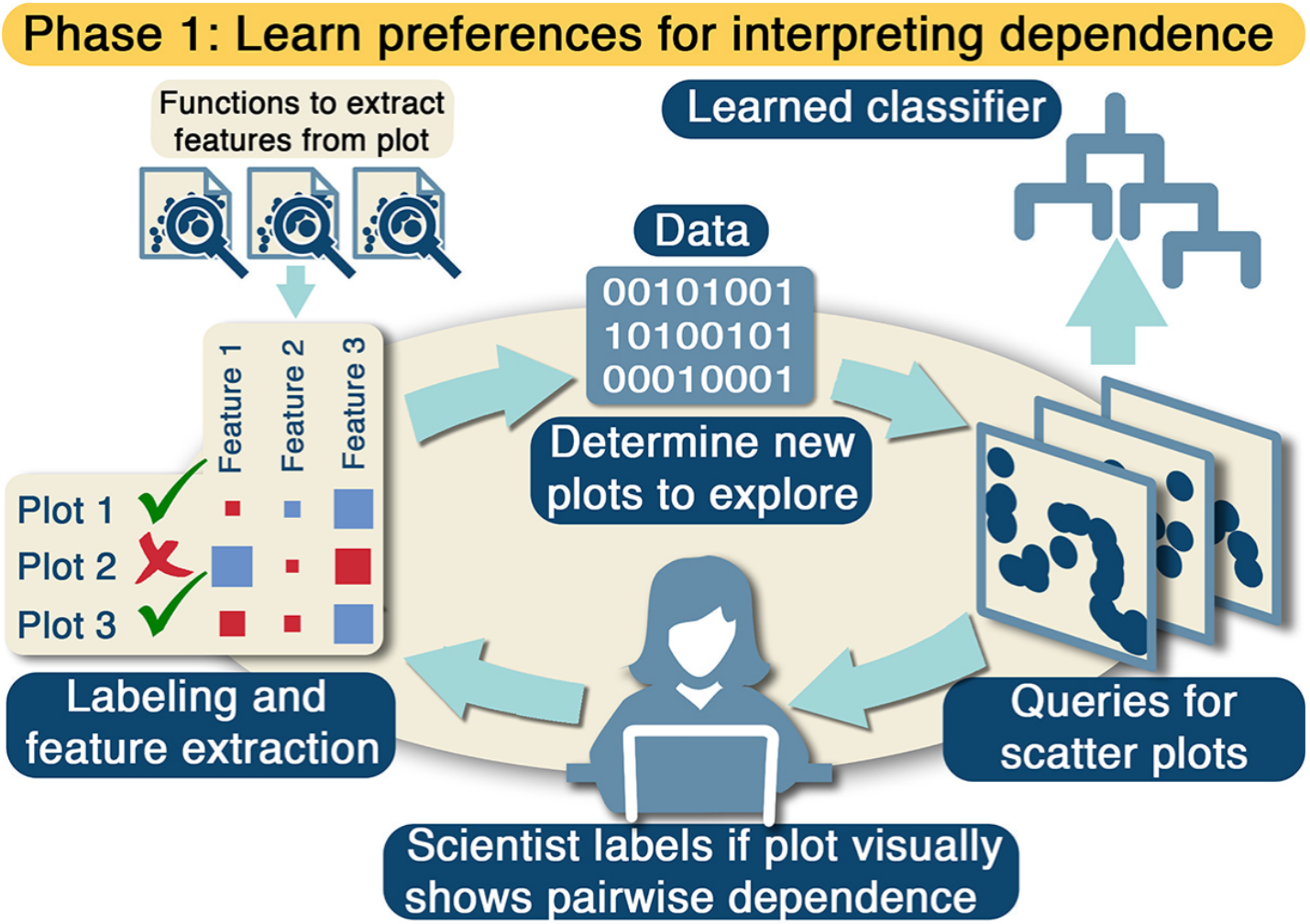
Avenue 3: Image recognition software to learn what patterns the user is interested in.



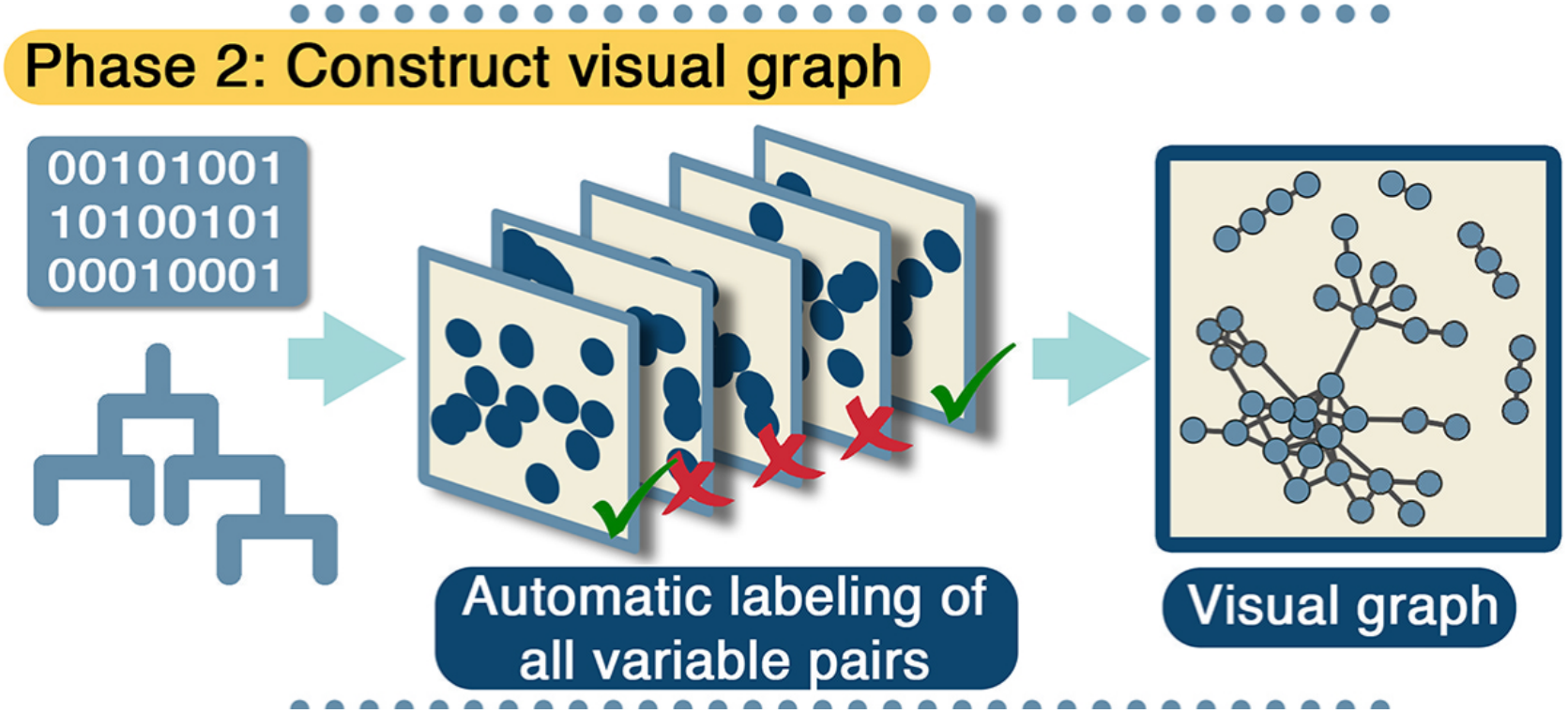
Avenue 3: Image recognition software to learn what patterns the user is interested in.



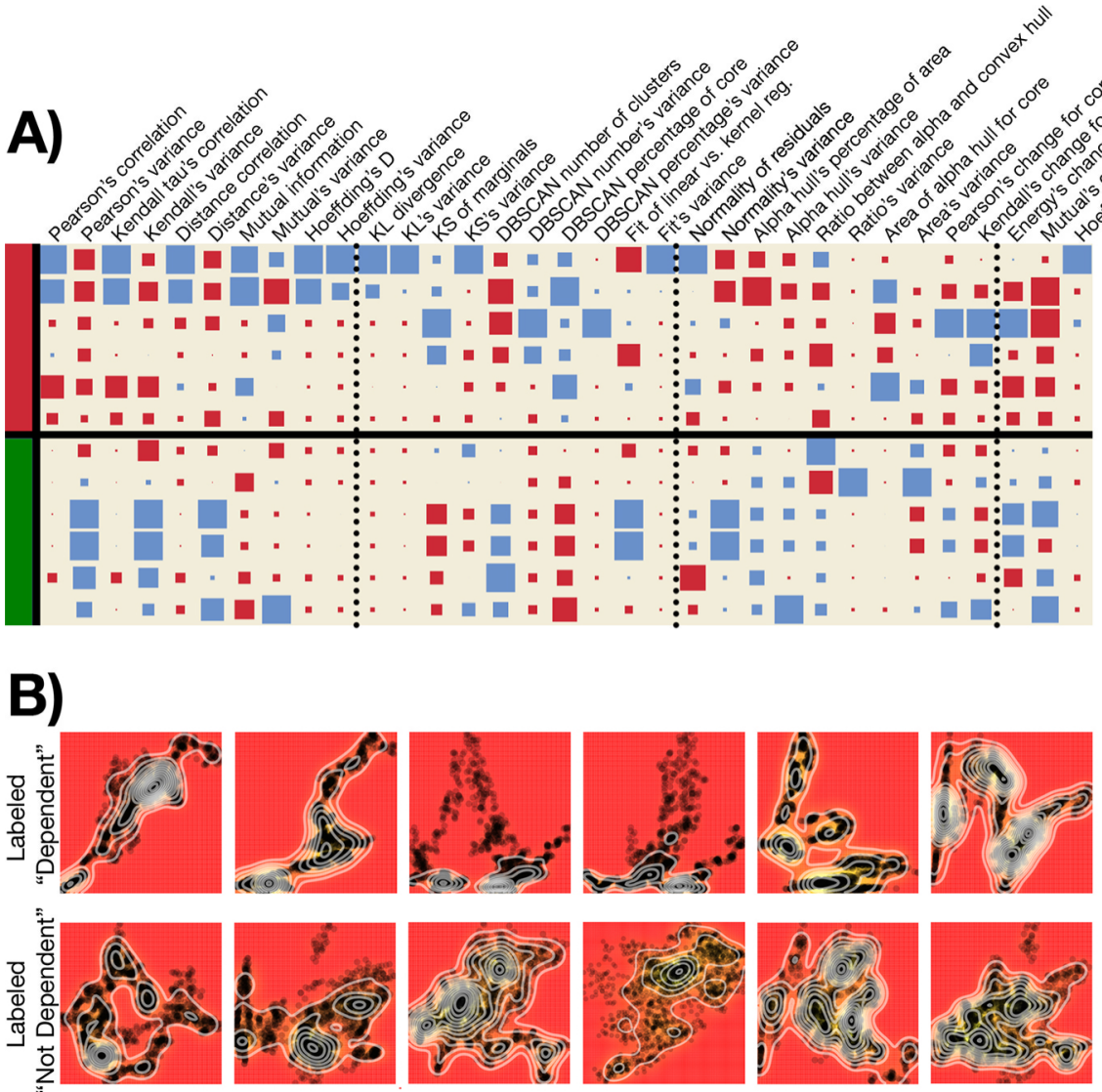
Avenue 3: Image recognition software to learn what patterns the user is interested in.



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



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



C)

Dependency measures:

Measures dependency between two variables, such as Pearson correlation, Kendall's tau correlation, distance correlation, mutual information, and Hoeffding's D.

Distributional measures:

Measures how ideal the empirical distribution is, such as the KL divergence between the bivariate Gaussian and product of univariate Gaussians, and sum of the KS statistic of each marginal distribution to a univariate Gaussian.

Clustering measures:

Measures clustering properties such as the number of clusters/modes or the number of core points found by DBSCAN.

Regression measures:

Measures how well one variable can predict the other, such as comparing the MSE between linear and kernel regression or the KS statistic between the residuals from kernel regression to a Gaussian.

Shape measures:

Measures the shape of the empirical density, such as the ratio between the area of the alpha hull and of the bound box or of the convex hull, and the change in area when only the alpha hull of the core points (found by DBSCAN) are considered.

Outlier measures:

Measures the change in dependency measures when only core points (found by DBSCAN) are considered.

So what?

- As you learn more about statistics, be sure to ask yourself how you would convince yourself how valid the results are.

So what?

- As you learn more about statistics, be sure to ask yourself how you would convince yourself how valid the results are.
- Be aggressive and unyielding in trying to find plots that can back up your numerical results.

So what?

- As you learn more about statistics, be sure to ask yourself how you would convince yourself how valid the results are.
- Be aggressive and unyielding in trying to find plots that can back up your numerical results.
- Neither numerical and visual methods are foolproof, so we have to always use both concurrently.