

ETC5521: Exploratory Data Analysis

Going beyond two variables, exploring high dimensions

Lecturer: *Di Cook*

✉ ETC5521.Clayton-x@monash.edu

CALENDAR Week 8 - Session 1



Read about the original book, and movie on [wikipedia](#)

Flatland: The Movie - Official Trailer



**More than two continuous
variables?**

Use a scatterplot matrix

synonyms: splom, draughtsman plot

Case study 1 Olive oils

data description R

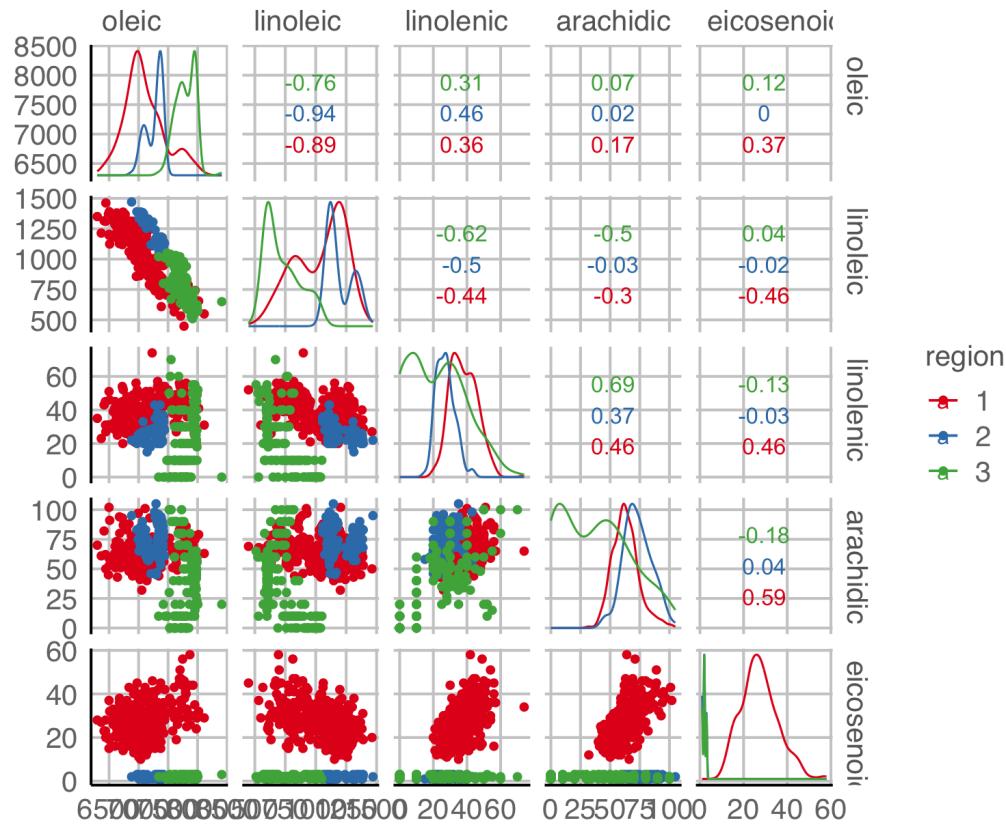
id	region	area	palmitic	palmitoleic	stearic	oleic	linoleic	linolenic	arachidic	eicosenoic
1.North-Apulia	1	1	1075	75	226	7823	672	36	60	29
2.North-Apulia	1	1	1088	73	224	7709	781	31	61	29
3.North-Apulia	1	1	911	54	246	8113	549	31	63	29
4.North-Apulia	1	1	966	57	240	7952	619	50	78	35
5.North-Apulia	1	1	1051	67	259	7771	672	50	80	46
6.North-Apulia	1	1	911	49	268	7924	678	51	70	44
7.North-Apulia	1	1	922	66	264	7990	618	49	56	29
8.North-Apulia	1	1	1100	61	235	7728	734	39	64	35
9.North-Apulia	1	1	1082	60	239	7745	709	46	83	33
10.North-Apulia	1	1	1037	55	213	7944	633	26	52	30
11.North-Apulia	1	1	1051	35	219	7978	605	21	65	24
12.North-Apulia	1	1	1036	59	235	7868	661	30	62	44
13.North-Apulia	1	1	1074	70	214	7728	747	50	79	33
14.North-Apulia	1	1	875	52	243	8018	655	41	79	32

Case study 1 Olive oils



learn

R



Case study 2 PISA

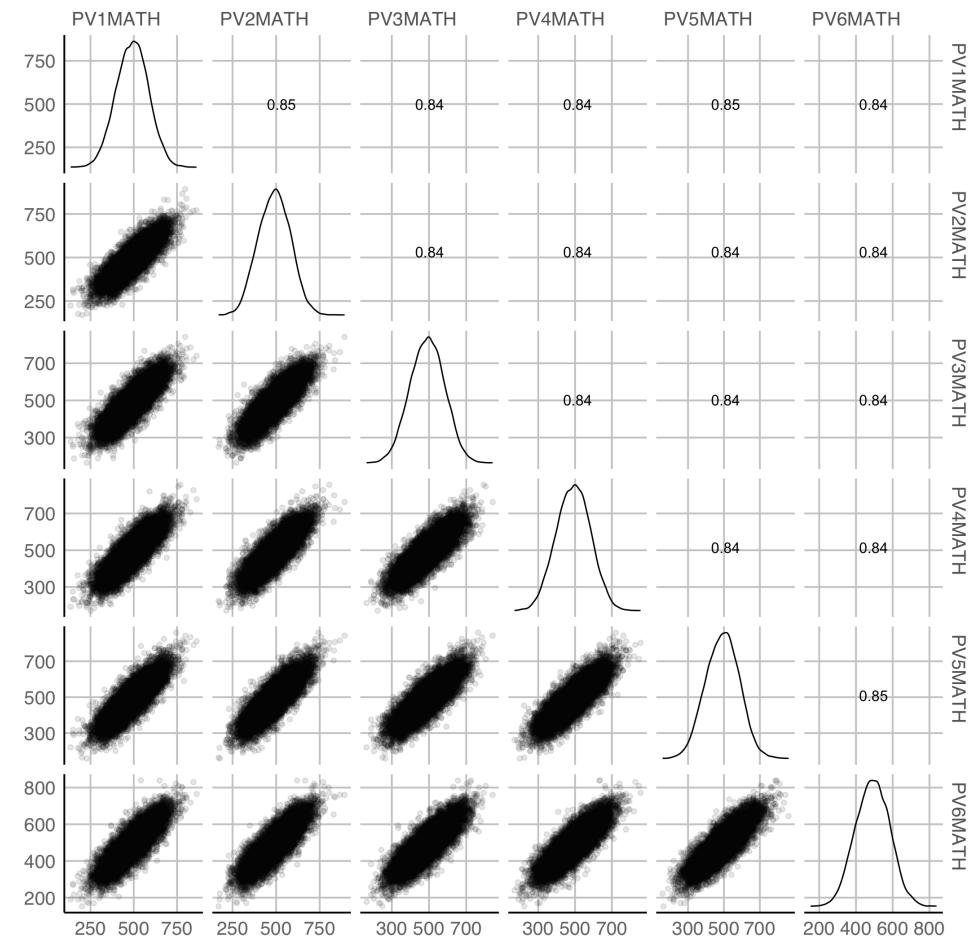


learn

R

The Programme for International Student Assessment (PISA) is a triennial survey conducted by the Organization for Economic Cooperation and Development (OECD) on assessment measuring 15-year-old student performances in reading, mathematics and science.

Math scores for Australia for 2018. (Only 6 of the 10 shown.)



Diversion

This is an example of fraudulent synthetic data, presented in a Lancet article in May 2020 claiming hydroxychloroquine increased risk of death.

RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19 (The Lancet, Volume 395, Issue 10234, 2020, pages 1073–1082)

Download PDF [1 MB]

Fi

[Summary](#)[Introduction](#)[Methods](#)[Results](#)[Discussion](#)[Supplementary Material](#)[References](#)[Article Info](#)[Figures](#)[Tables](#)

Background

Hydroxychloroquine or chloroquine, often in combination with a second-generation macrolide, are being widely used for treatment of COVID-19, despite no conclusive evidence of their benefit. Although generally safe when used for approved indications such as autoimmune disease or malaria, the safety and benefit of these treatment regimens are poorly evaluated in COVID-19.

Methods

We did a multinational registry analysis of the use of hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19. The registry comprised data from 671 hospitals in six continents. We included patients hospitalised between Dec 20, 2019, and April 14, 2020, with a positive laboratory finding for SARS-CoV-2. Patients who received one of the treatments of interest within 48 h of diagnosis were included in one of four treatment groups (chloroquine alone, chloroquine with a macrolide, hydroxychloroquine alone, or hydroxychloroquine with a macrolide), and patients who received none of these treatments formed the control group. Patients for whom one of the treatments of interest was initiated more than 48 h after diagnosis or while they were on mechanical ventilation, as well as patients who received remdesivir, were excluded. The main outcomes of interest were in-hospital mortality and the occurrence of de-novo ventricular arrhythmias (non-sustained or sustained ventricular tachycardia or ventricular fibrillation).

Table S3. Summary Data by Continent

Variable	North America	South America	Europe	Africa	Asia	Australia
N	63,315	3,577	16,574	4,402	7,555	609
Age (years)	54.4 +/- 17.8	53.6 +/- 17.1	52.7 +/- 17.0	53.9 +/- 16.9	51.9 +/- 17.2	55.8 +/- 17.7
BMI (Kg/m²)	28.1 +/- 5.3	26.4 +/- 5.4	28.1 +/- 5.3	23.8 +/- 5.4	24.8 +/- 5.3	28.1 +/- 5.4
Female sex	29,288 (46.3)	1,678 (46.9)	7,730 (46.6)	1,981 (45.0)	3,486 (46.1)	263 (43.2)
Coronary artery disease	7,850 (12.4)	485 (13.6)	2,169 (13.1)	614 (13.9)	980 (13.0)	39 (6.4)
Congestive heart failure	1,639 (2.6)	73 (2.0)	366 (2.2)	105 (2.4)	179 (2.4)	6 (1.0)
History of arrhythmia	2,293 (3.6)	118 (3.3)	543 (3.3)	146 (3.3)	256 (3.4)	25 (4.1)
Diabetes mellitus	8,654 (13.7)	521 (14.6)	2,360 (14.2)	570 (12.9)	1,069 (14.1)	86 (14.1)
Hypertension	17,159 (27.1)	954 (26.7)	4,368 (26.4)	1,140 (25.9)	2,010 (26.6)	179 (29.4)
Hyperlipidemia	20,032 (31.6)	1,088 (30.4)	5,131 (31.0)	1,380 (31.3)	2,374 (31.4)	193 (31.7)
COPD	2,069 (3.3)	97 (2.7)	590 (3.6)	132 (3.0)	254 (3.4)	35 (5.7)
Current smoker	6,316 (10.0)	347 (9.7)	1,604 (9.7)	453 (10.3)	707 (9.4)	61 (10.0)
Former smoker	10,707 (16.9)	670 (18.7)	2,936 (17.7)	830 (18.9)	1,301 (17.2)	109 (17.9)
Immunocompromised	1,997 (3.2)	52 (1.5)	463 (2.8)	127 (2.9)	208 (2.8)	21 (3.4)
ACE inhibitor	5,327 (8.4)	285 (8.0)	1,341 (8.1)	325 (7.4)	605 (8.0)	66 (10.8)
Statin	6,188 (9.8)	306 (8.6)	1,552 (9.4)	436 (9.9)	674 (8.9)	89 (14.6)
ARB	3,913 (6.2)	220 (6.2)	963 (5.8)	259 (5.9)	454 (6.0)	40 (6.6)
Antiviral Therapy use	25,646 (40.5)	1,444 (40.4)	6,747 (40.7)	1,771 (40.2)	3,085 (40.8)	234 (38.4)
Chloroquine alone	1,091 (1.7)	114 (3.2)	295 (1.8)	153 (3.5)	199 (2.6)	16 (2.6)
Hydroxychloroquine alone	2,127 (3.4)	72 (2.0)	540 (3.3)	83 (1.9)	184 (2.4)	10 (1.6)
CQ + macrolide	2,324 (3.7)	217 (6.1)	562 (3.4)	256 (5.8)	391 (5.2)	33 (5.4)
HCQ + macrolide	-	-	-	-	-	-

Table S3. Summary Data by Continent

Variable	North America	South America	Europe	Africa	Asia	Australia
N	25,200 (40.5)	1,070 (40.5)	7,750 (40.6)	1,581 (40.6)	3,480 (40.1)	205 (45.2)
Coronary artery disease	7,850 (12.4)	485 (13.6)	2,169 (13.1)	614 (13.9)	980 (13.0)	39 (6.4)
Congestive heart failure	1,620 (2.6)	72 (2.0)	366 (2.2)	105 (2.1)	172 (2.1)	6 (1.0)
Former smoker	6,316 (10.0)	347 (9.7)	1,604 (9.7)	453 (10.3)	707 (9.4)	61 (10.0)
Immunocompromised	10,707 (16.9)	670 (18.7)	2,936 (17.7)	830 (18.9)	1,301 (17.2)	109 (17.9)
ACE inhibitor	1,997 (3.2)	52 (1.5)	463 (2.8)	127 (2.9)	208 (2.8)	21 (3.4)
Statin	5,327 (8.4)	285 (8.0)	1,341 (8.1)	325 (7.4)	605 (8.0)	66 (10.8)
ARB	6,188 (9.8)	306 (8.6)	1,552 (9.4)	436 (9.9)	674 (8.9)	89 (14.6)
Antiviral Therapy use	3,913 (6.2)	220 (6.2)	963 (5.8)	259 (5.9)	454 (6.0)	40 (6.6)
Chloroquine alone	25,646 (40.5)	1,444 (40.4)	6,747 (40.7)	1,771 (40.2)	3,085 (40.8)	234 (38.4)
Hydroxychloroquine alone	1,091 (1.7)	114 (3.2)	295 (1.8)	153 (3.5)	199 (2.6)	16 (2.6)

Generalised pairs plot

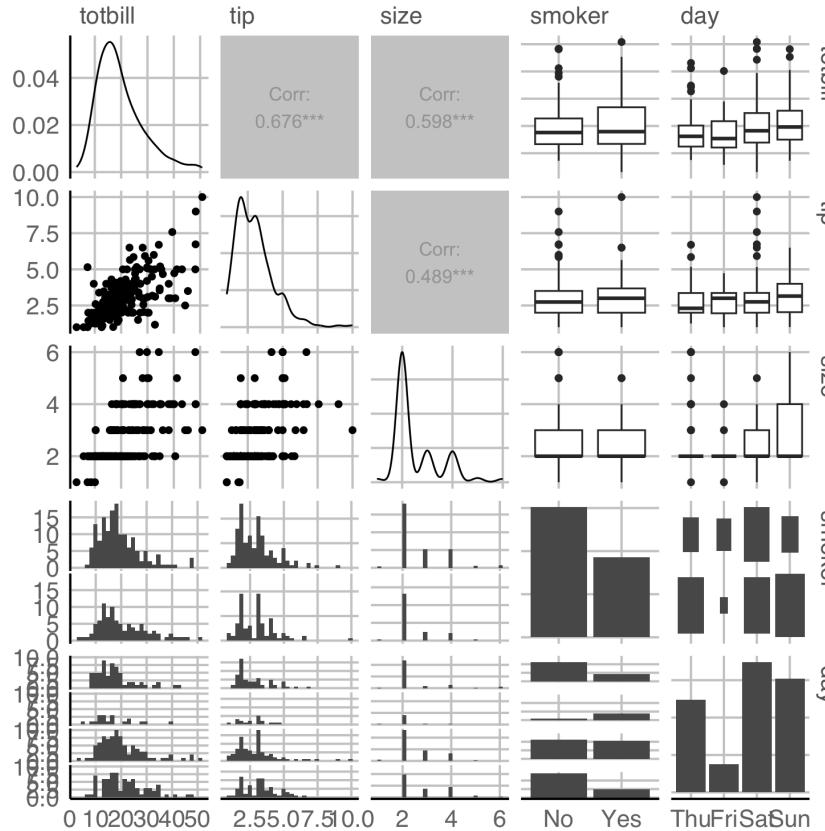
If the types of variables are not both quantitative, there are some other choices of mapping

Case study 3 Tips



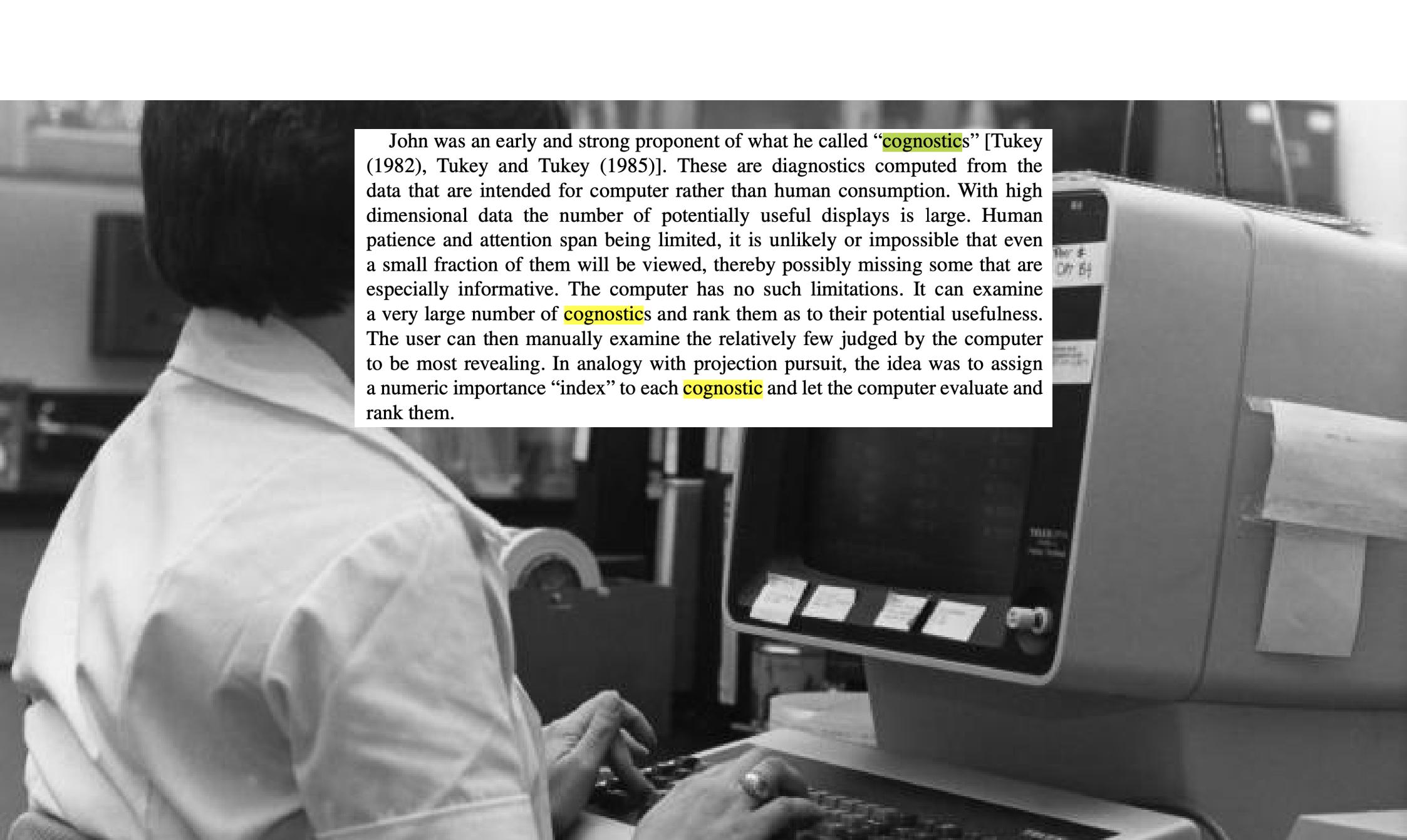
learn

R

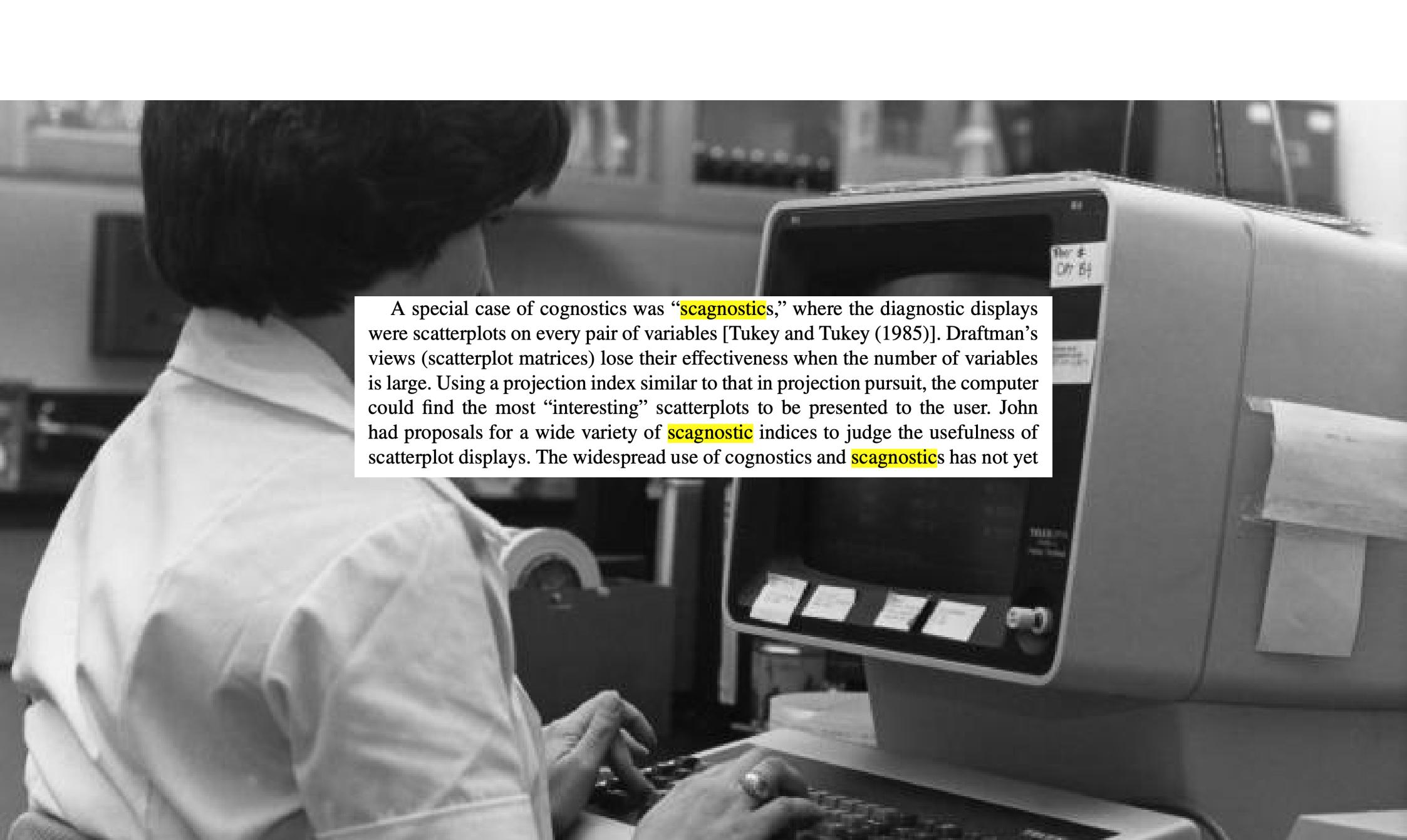


Scagnostics

Has your data got too many pairs of variables to scan easily?

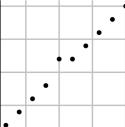
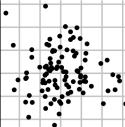
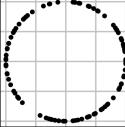
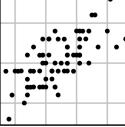
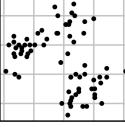


John was an early and strong proponent of what he called “**cognostics**” [Tukey (1982), Tukey and Tukey (1985)]. These are diagnostics computed from the data that are intended for computer rather than human consumption. With high dimensional data the number of potentially useful displays is large. Human patience and attention span being limited, it is unlikely or impossible that even a small fraction of them will be viewed, thereby possibly missing some that are especially informative. The computer has no such limitations. It can examine a very large number of **cognostics** and rank them as to their potential usefulness. The user can then manually examine the relatively few judged by the computer to be most revealing. In analogy with projection pursuit, the idea was to assign a numeric importance “index” to each **cognostic** and let the computer evaluate and rank them.



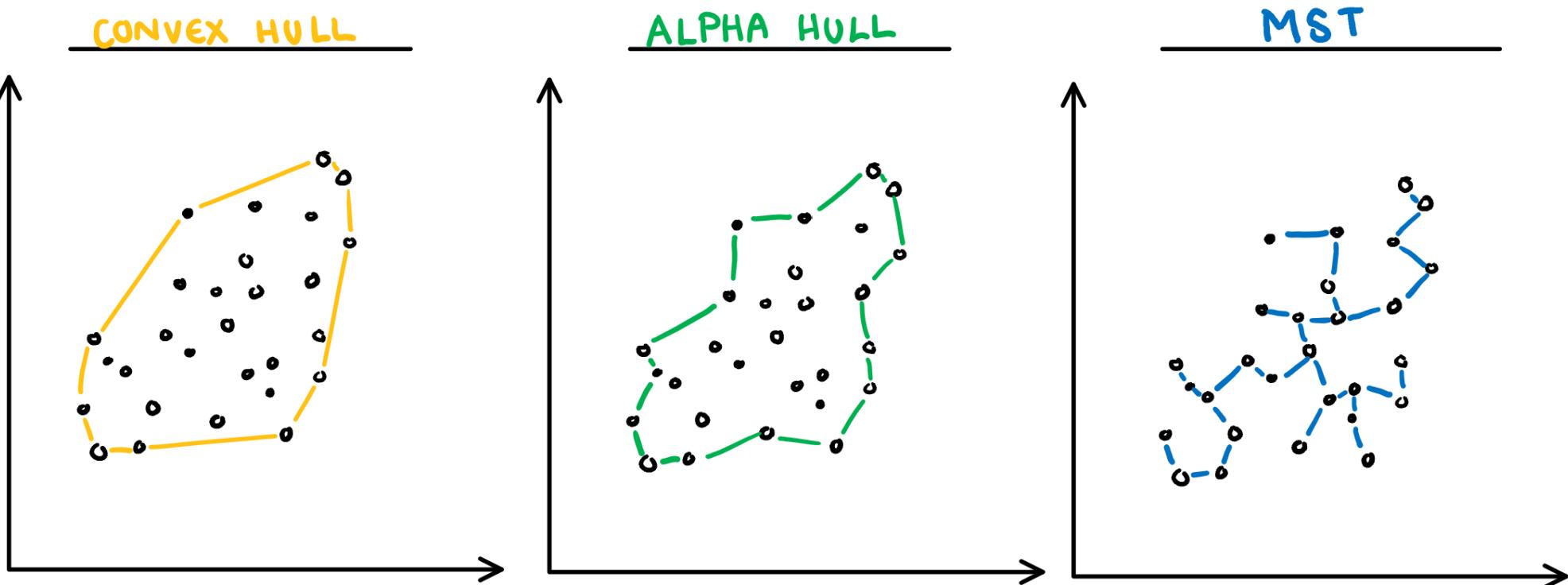
A special case of cagnostics was “scagnostics,” where the diagnostic displays were scatterplots on every pair of variables [Tukey and Tukey (1985)]. Draftman’s views (scatterplot matrices) lose their effectiveness when the number of variables is large. Using a projection index similar to that in projection pursuit, the computer could find the most “interesting” scatterplots to be presented to the user. John had proposals for a wide variety of scagnostic indices to judge the usefulness of scatterplot displays. The widespread use of cagnostics and scagnostics has not yet

Scagnostics

plot	set	outlying	stringy	striated	clumpy	sparse	monotonic	dcor
	line	0.000	1.000	0.600	0.368	0.157	0.997	0.991
	norm	0.190	0.789	0.330	0.603	0.095	0.013	0.160
	circle	0.000	1.000	0.980	0.966	0.065	0.009	0.248
	stripes	0.129	0.698	0.338	0.985	0.094	0.665	0.632
	clumps	0.038	0.608	0.233	0.992	0.107	0.375	0.502

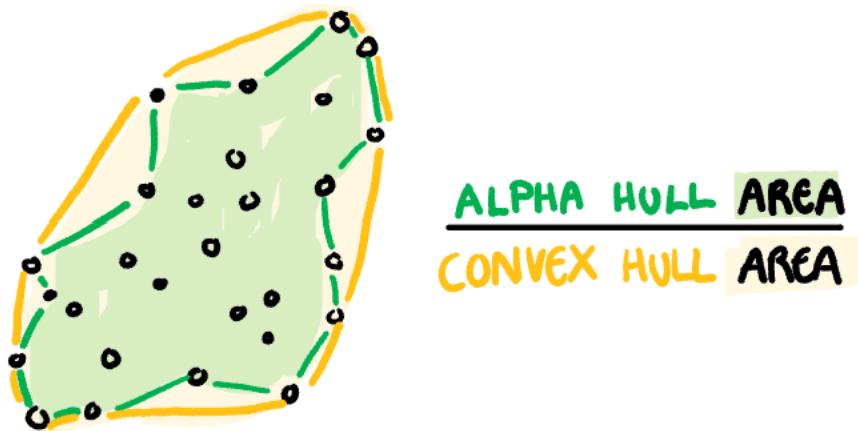
How are scagnostics calculated?

The building blocks are: convex hull, alpha hull, and minimal spanning tree



Convex: Measure of how convex the shape of the data is. Computed as the ratio between the area of the alpha hull (A) and convex hull (C).

$$S_{\text{convex}} = w \frac{\text{area}(A)}{\text{area}(C)}$$



Skinny: A measure of how "thin" the shape of the data is. It is calculated as the ratio between the area and perimeter of the alpha hull (A) with some normalisation such that 0 correspond to a perfect circle and values close to 1 indicate a skinny polygon.

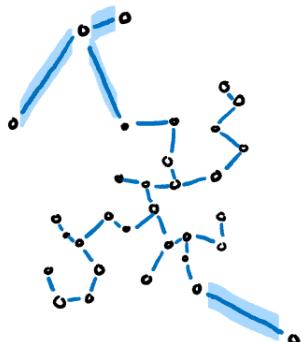
$$S_{\text{skinny}} = 1 - \frac{\sqrt{4\pi \text{area}(A)}}{\text{perimeter}(A)}$$



$$1 - \sqrt{\frac{4\pi \cdot \text{ALPHA HULL AREA}}{\text{ALPHA HULL PERIMETER}}}$$

Outlying: A measure of proportion and severity of outliers in dataset. Calculated by comparing the edge lengths of the outlying points in the MST with the length of the entire MST.

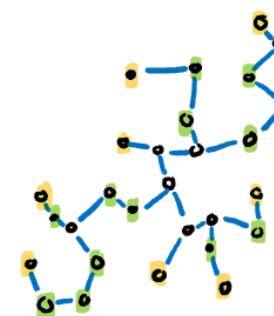
$$s_{\text{outlying}} = \frac{\text{length}(M_{\text{outliers}})}{\text{length}(M)}$$



EDGE LENGTHS OF OUTLYING POINTS
EDGE LENGTHS OF ORIGINAL MST

Stringy: This measure identifies a "stringy" shape with no branches, such as a thin line of data. It is calculated by comparing the number of vertices of degree two ($V^{(2)}$) with the total number of vertices (V), dropping those of degree one ($V^{(1)}$).

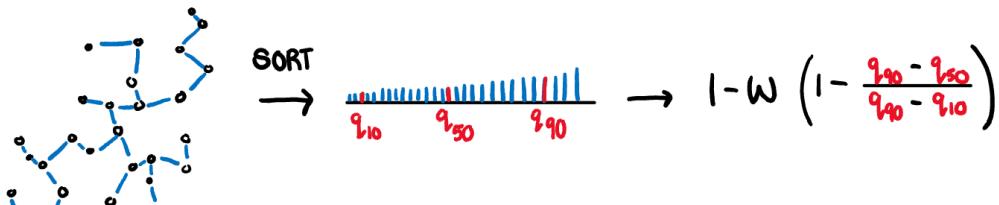
$$s_{\text{stringy}} = \frac{|V^{(2)}|}{|V| - |V^{(1)}|}$$



NUMBER OF $V^{(2)}$
TOTAL NUMBER OF V - NUMBER OF $V^{(1)}$

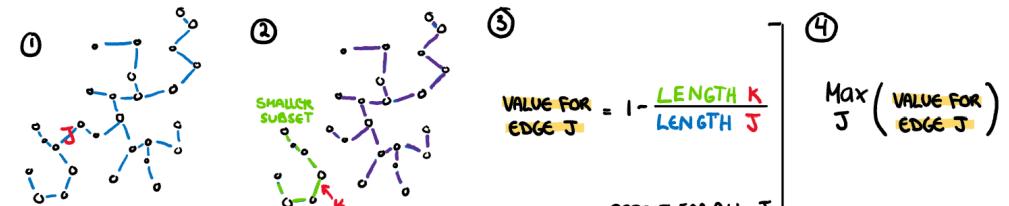
Skewed: A measure of skewness in the edge lengths of the MST (not in the distribution of the data). It is calculated as the ratio between the 40% IQR and the 80% IQR, adjusted for sample size dependence.

$$s_{\text{skewed}} = 1 - w \left(1 - \frac{q_{90} - q_{50}}{q_{90} - q_{10}} \right)$$



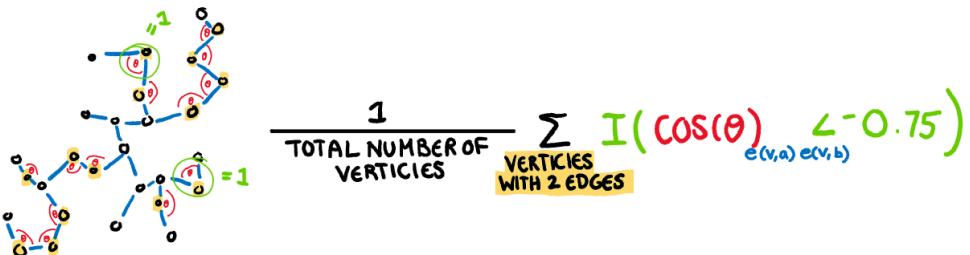
Clumpy: This measure is used to detect clustering and is calculated through an iterative process. First an edge J is selected and removed from the MST. From the two spanning trees that are created by this break, we select the largest edge from the smaller tree (K). The length of this edge (K) is compared to the removed edge (J) giving a clumpy measure for this edge. This process is repeated for every edge in the MST and the final clumpy measure is the maximum of this value over all edges.

$$\max_j \left(1 - \frac{\max_k(\text{length}(e_k))}{\text{length}(e_j)} \right)$$



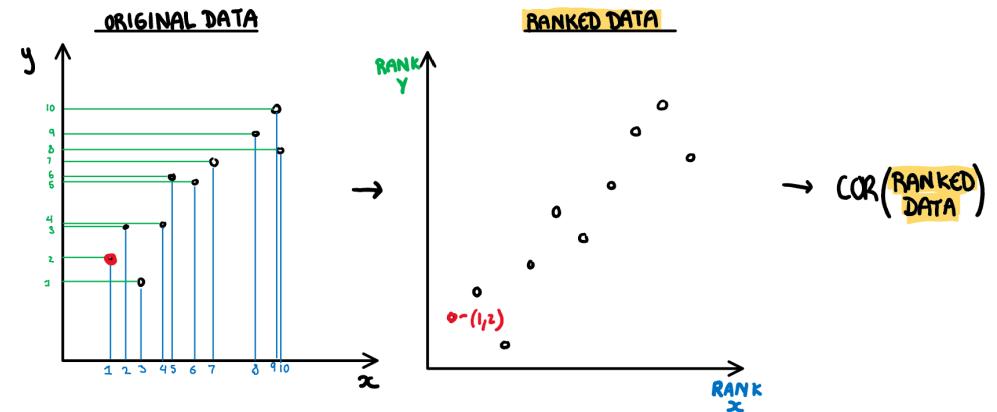
Striated: This measure identifies features such as discreteness by finding parallel lines, or smooth algebraic functions. Calculated by counting the proportion of acute (0 to 40 degree) angles between the adjacent edges of vertices with only two edges.

$$\frac{1}{|V|} \sum_{v \in V^2} I(\cos\theta_{e(v,a)e(v,b)} < -0.75)$$



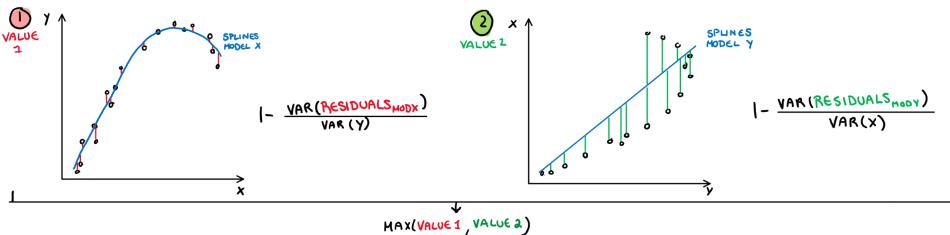
Monotonic: Checks if the data has an increasing or decreasing trend. Calculated as the Spearman correlation coefficient, i.e. the Pearson correlation between the ranks of x and y.

$$S_{\text{monotonic}} = r_{\text{spearman}}^2$$



Splines: Measures the functional non-linear dependence by fitting a penalised splines model on X using Y, and on Y using X. The variance of the residuals are scaled down by the axis so they are comparable, and finally the maximum is taken. Therefore the value will be closer to 1 if either relationship can be decently explained by a splines model.

$$S_{\text{splines}} = \max_{i \in x,y} \left[1 - \frac{\text{Var}(\text{Residuals}_{\text{model } i=..})}{\text{Var}(i)} \right]$$



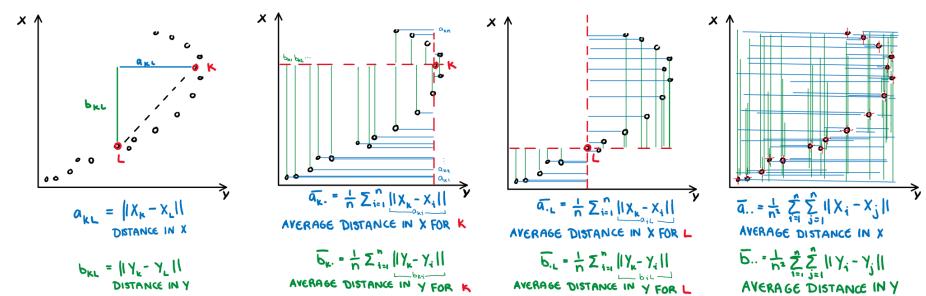
Dcor: A measure of non-linear dependence which is 0 if and only if the two variables are independent. Computed using an ANOVA like calculation on the pairwise distances between observations.

$$S_{\text{dcor}} = \sqrt{\frac{V(X,Y)}{V(X,X)V(Y,Y)}} \text{ where}$$

$$V(X,Y) = \frac{1}{n^2} \sum_{k=1}^n \sum_{l=1}^n A_{kl} B_{kl},$$

$$A_{kl} = a_{kl} - \bar{a}_k - \bar{a}_j - \bar{a}_{..}$$

$$B_{kl} = b_{kl} - \bar{b}_k - \bar{b}_j - \bar{b}_{..}$$



Scagnostics from familiar measures

There are many more ways to numerically characterise association that can be used as scagnostics too:

- ▶ We used those available in the `vaast` R package
- ▶ Slope, intercept, and error estimate from a simple linear model
- ▶ Correlation
- ▶ Principal component analysis: first eigenvalue
- ▶ Linear discriminant analysis: Between group SS to within group SS
- ▶ Cluster metrics
- ▶ Also see
 - ◀ Scagnostics for time series (`feasts` R package)
 - ◀ Longagnostics for longitudinal data (`brolgar` R package)

Resources

- Friendly and Denis "Milestones in History of Thematic Cartography, Statistical Graphics and Data Visualisation" available at <http://www.datavis.ca/milestones/>
- Schloerke et al (2020). GGally: Extension to 'ggplot2'. <https://ggobi.github.io/ggally>.
- Wilkinson, Anand, Grossmann (1994) Graph-Theoretic Scagnostics, <http://papers.rgrossman.com/proc-094.pdf>
- Grimm, K. (2016). Kennzahlenbasierte grafikauswahl (pp. III, 210) [Doctoral thesis]. Universität Augsburg.
- Hofmann et al (2020) binostics: Compute Scagnostics. R package version 0.1.2. <https://CRAN.R-project.org/package=binostics>
- O'Hara-Wild, Hyndman, Wang (2020). <https://CRAN.R-project.org/package=fabletools>
- Tierney, Cook, Prvan (2020) <https://github.com/njtierney/brolgar>



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](#).

Lecturer: *Di Cook*

✉ ETC5521.Clayton-x@monash.edu

CALENDAR Week 8 - Session 1

