RLadies New York Data Visualization Workshop

Joyce Robbins

August 17, 2017

packages to install: dplyr, ggplot2, gridExtra, pgmm, scales, tibble, tidyr, vcdExtra, viridis

download this repo: www.github.com/jtr13/RLadies

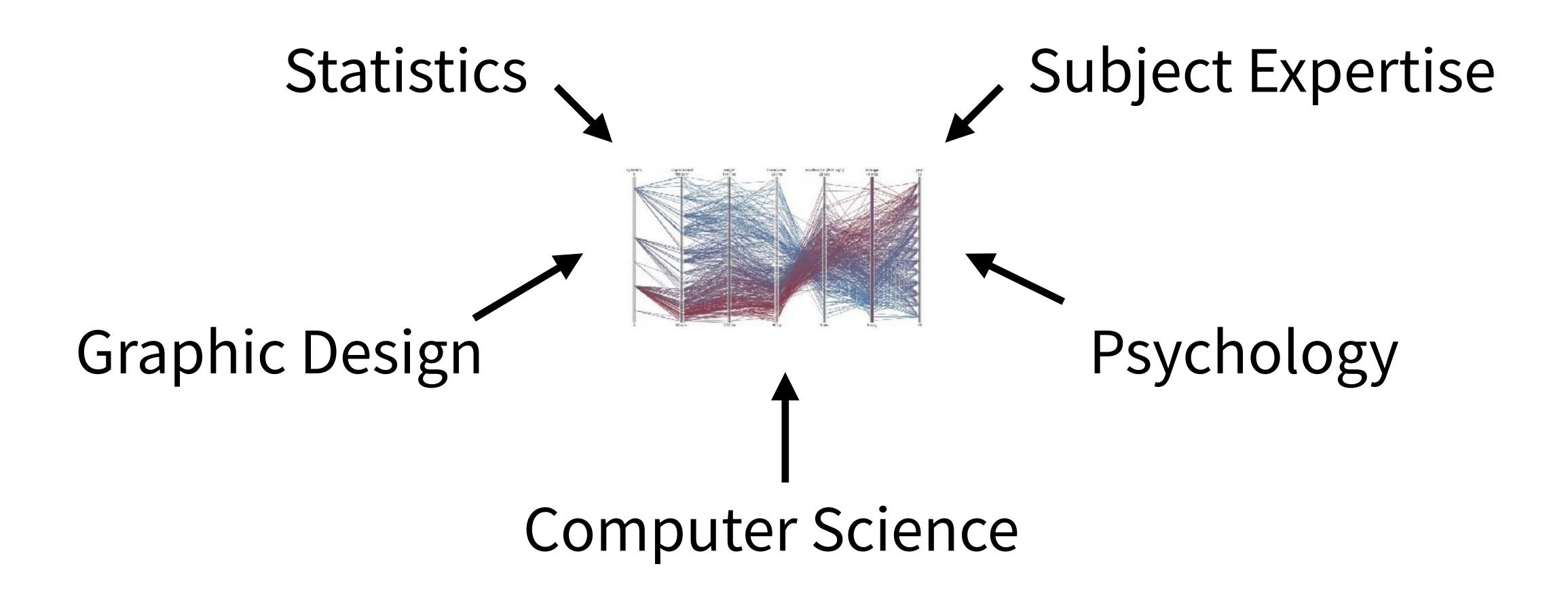
Agenda

- SLIDES: Datavis Intro & Heatmaps
- RSTUDIO: geom_tile, geom_rect, geom_raster
- PRACTICE: tiny data sets
- RSTUDIO: heatmaps
- PRACTICE: Yamaguchi87 dataset (vcdExtra)
- SLIDES: Color
- PRACTICE: Yamaguchi87 or wine dataset (pgmm)

What is data visualization?

- relatively new field (but long history)
- multidisciplinary
- lack of consensus

Interdisciplinary influences



Data visualization for data science

- detecting patterns
- finding outliers
- making comparisons
- identifying clusters

Why heatmaps?

- Employs color to add a third dimension
- Works well in cases in which there's an observation for every (x, y) pair, such as spatial data
- Works with categorical or numerical data, discrete or continuous data

Heat Map Examples

TEN TESTS OF EFFICIENCY

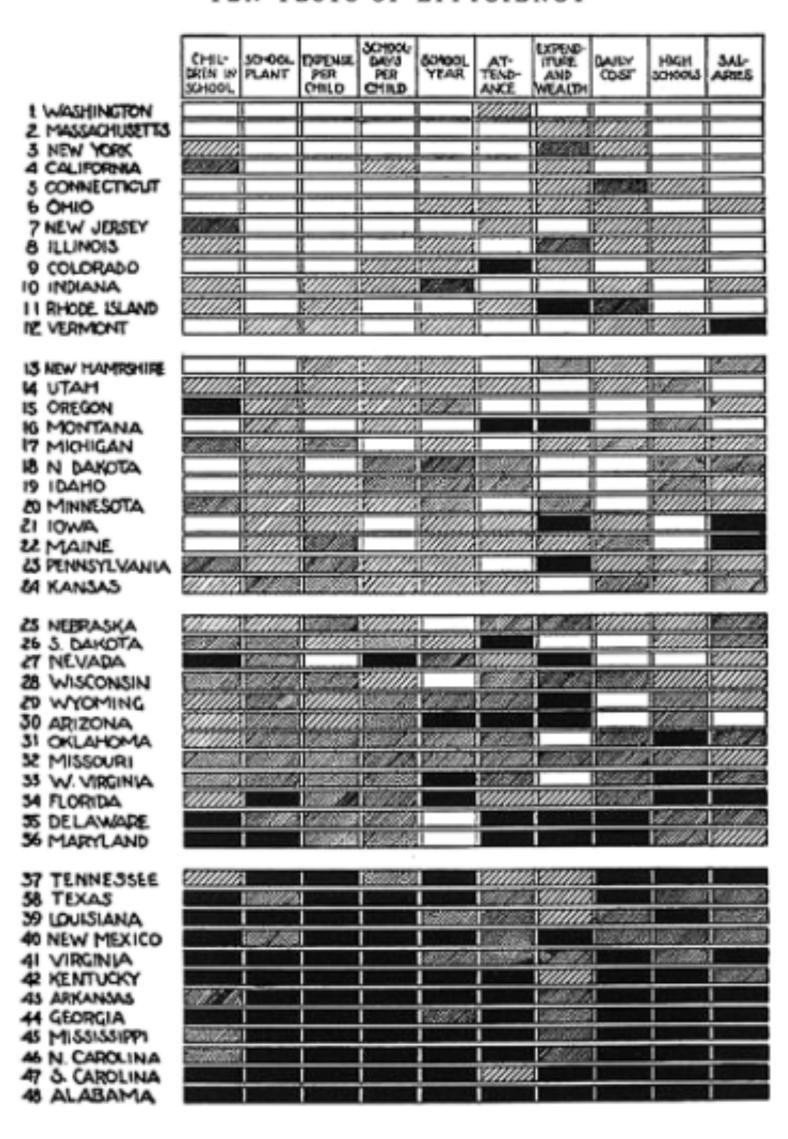


Figure 3. Sorted shaded display from Brinton (1914). The data are ranks of U.S. states on each of 10 educational features assessed in 1910. The matrix has been sorted by the row-marginal ranks.

TEN TESTS OF EFFICIENCY

	CHIL- DRIN IN SCHOOL	SO-OOL PLANT	DOENSE PER CHILD	SCHOOL: DAYS PER CHILD	60400L YEAR	AT- TEND- ANCE	EXPEND- ITURE AND WEALTH	COST	HIGH SCHOOLS	SAL- ARIES
! WASHINGTON						VIIIIII				
2. MASSACHUSETTS							ann.	401111		
3 NEW YORK	91111111						化的编	unn.		-
4 CALIFORNIA	93314						VIIIII			
5 CONNECTICUT							11111111	被视的	MIIIII.	
6 OHIO					11111111	VIIIIII.	71111111	Millin.		11111111
7 NEW JERSEY	160 066					VIIIIII		11/1/1/1/	<i>41111111</i> .	
8 ILLINOIS	91111111				anna.		1111	Mana.	Million .	
9 COLORADO			1	11/1/1/	Million.		911111		11111111	
IO INDIANA	444144		CHILLIAN .		77/2007			VIIIIIIII	1	41111111
I I RHODE ISLAND	94411111		4/8/1/10		1	90000		de 180	1	
TPOINTSY 30		9/11/14	MILLER		VIIIIIII			WIIIII.	3//////	
IS NEW HAMPSHIRE			11/2/10		William.		SCHOOL SELECTION OF THE PERSON	MANN.		1000
4 UTAH	umm.	1111111	Sentin.		MIIIIII	VIIIII.		11111111	10,000	
IS OREGON	1	11111/2	Manne.	3/1/////	14/20					
IG MONTANA		1111111	8	unnan					10000	
17 MICHIGAN	10000	CH11111	1524392	1	WIIIII.		19111111	100	SIMMI.	SHIII B
IR AL DAKOTA				CARDOON .	111111	100/97/2			100000	10000

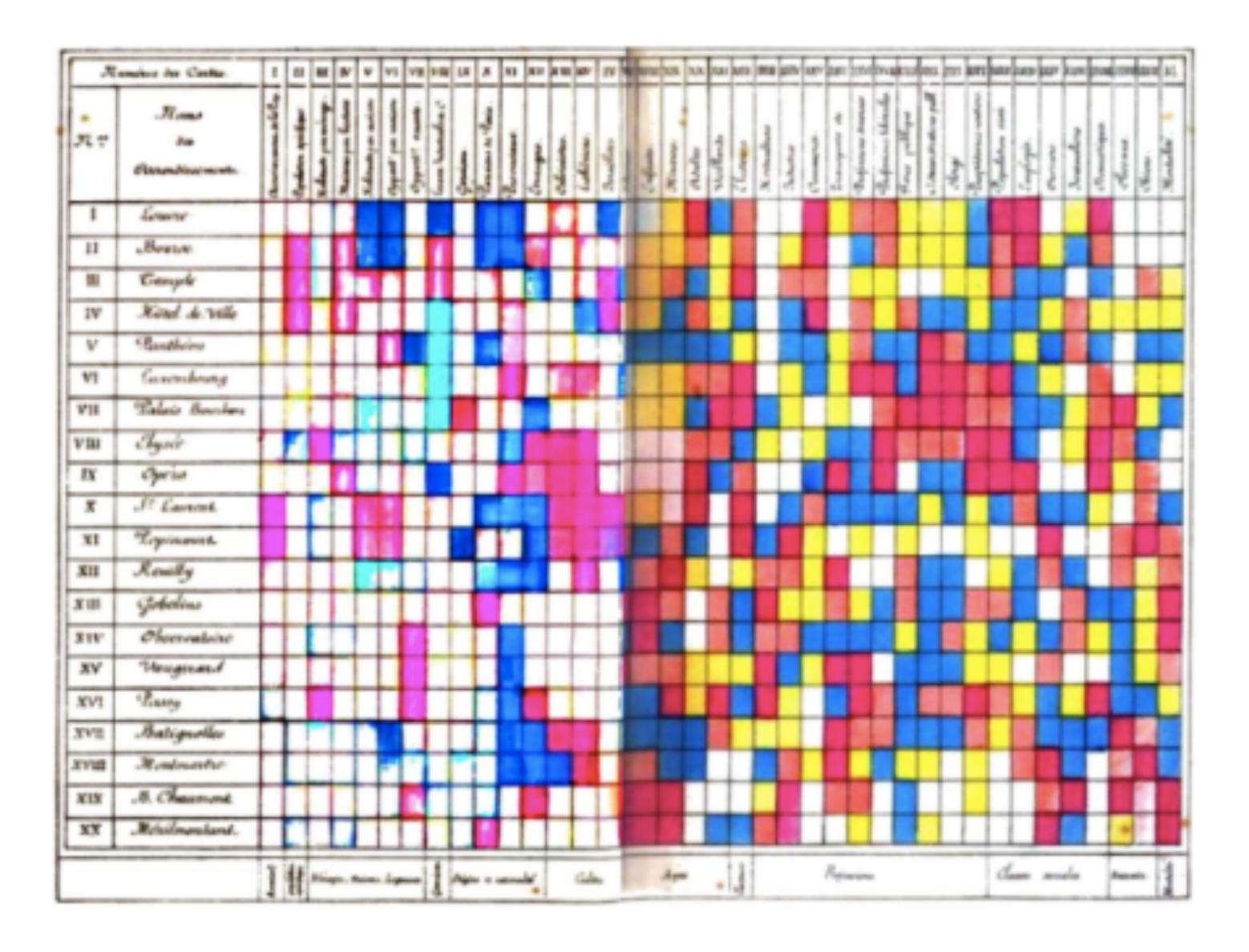
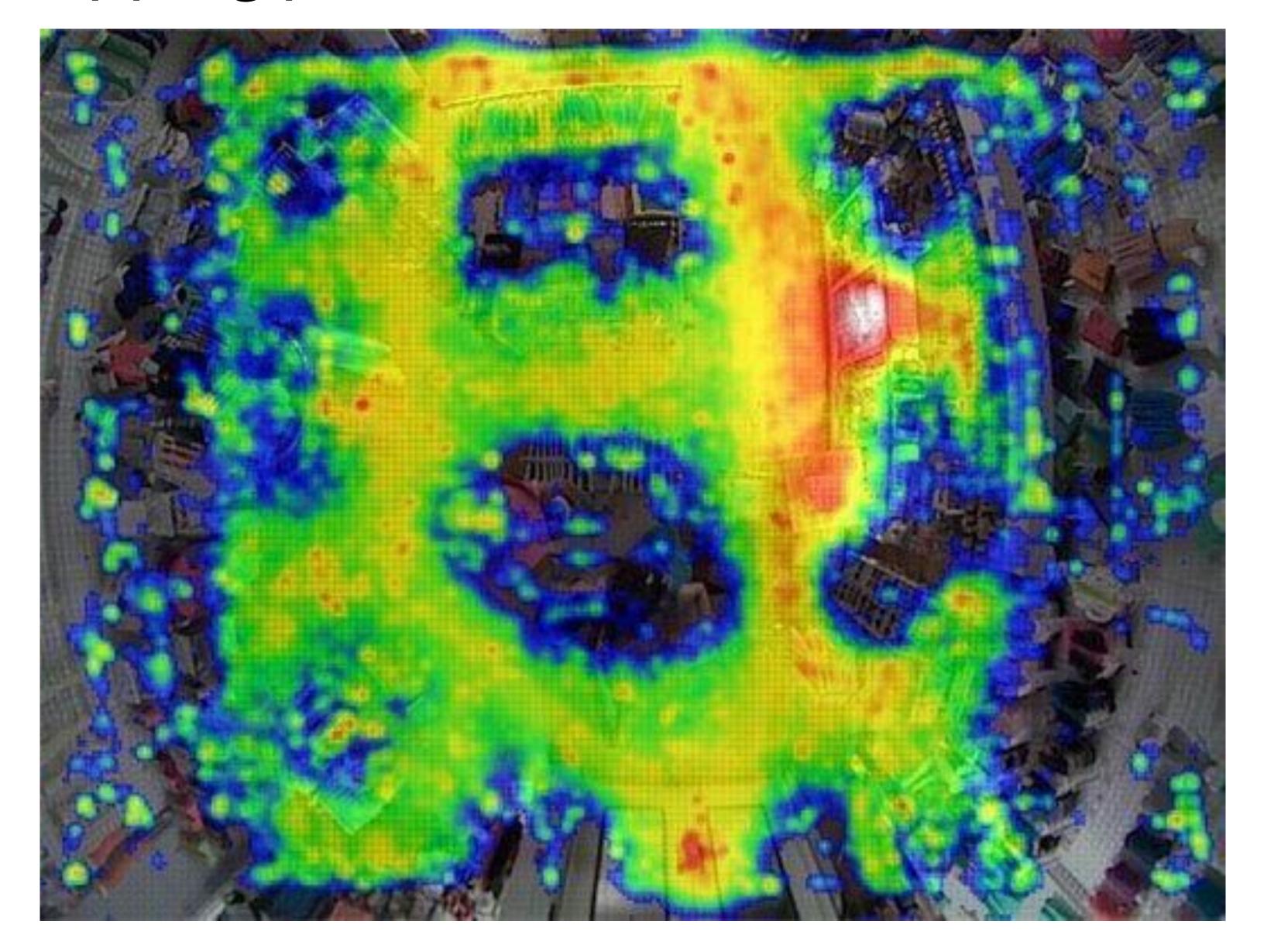
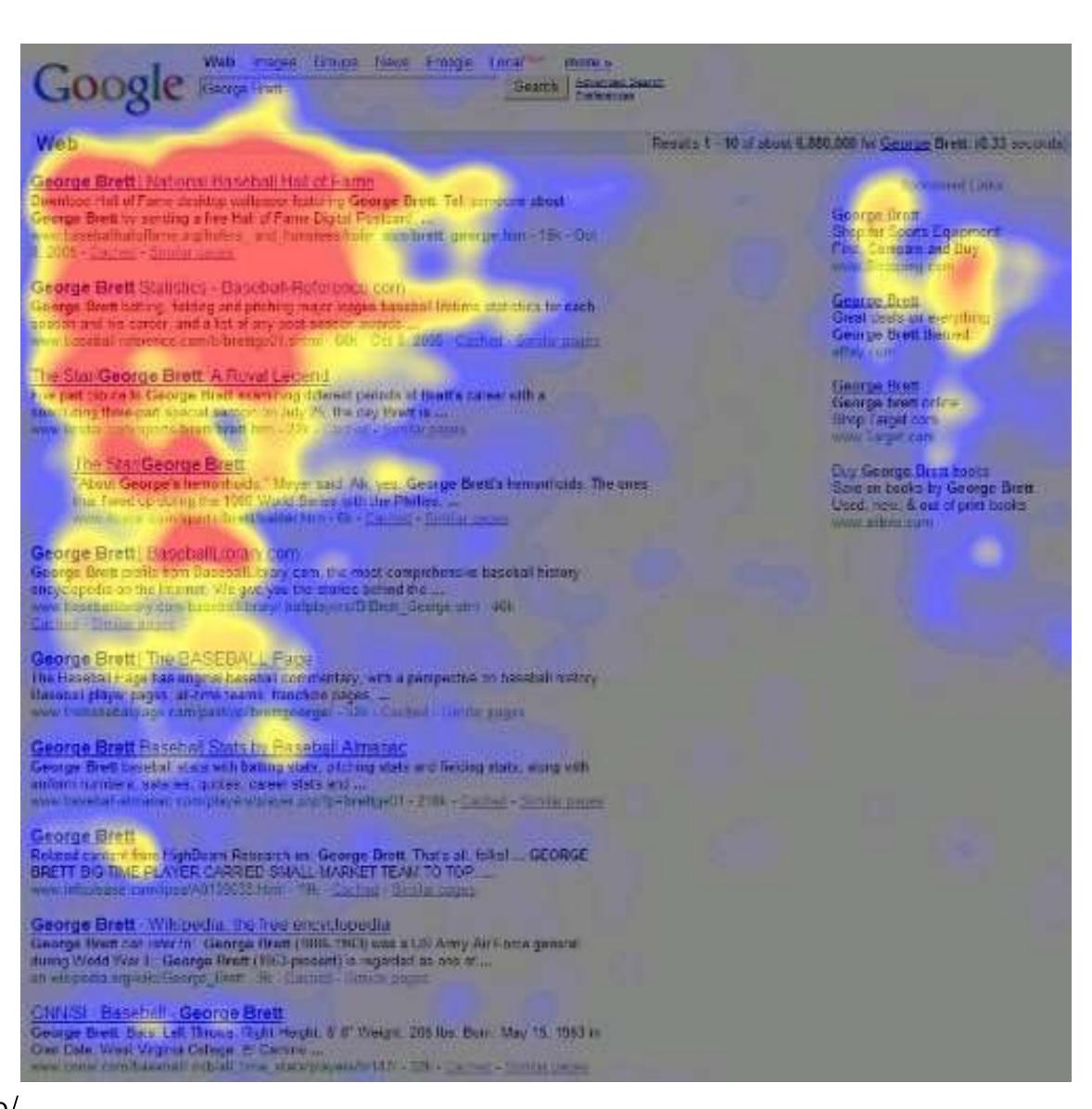


Figure 2. Shaded matrix display from Loua (1873), available online at http://books.google.com/books/. This was designed as a summary of 40 separate maps of Paris, showing the characteristics (e.g., national origin, professions, age, social classes) of 20 districts, using a color scale ranging from white (low) through yellow and blue to red (high).

In-store shopping pattern



physical location



screen location

21.75 secs



Ex ge le for the me sensitive skin.

S) mitive skin, add the chemicals and moisture to a lyou have diaper rash.

Bill or's unique high-absorbency natural-blend cotton odes cotton-soft, extra thick, gel-free protection you bully a sensitive skin. The chlorine-free materials and sorbent polymers is non-toxic and non-irritating. Clinically tested and pediatrician recommended for babies with allergies and sensitive skin.



Web site response

Six Weeks of Aggregate Average Response Time Data (in ms), By Day of Week and Hour o								
	SUN	MON	TUE	WED	THU	FRI	SAT	
Midnight								
01:00 AM PT								
02:00 AM PT								
03:00 AM PT								
04:00 AM PT								
05:00 AM PT								
06:00 AM PT								
07:00 AM PT								
08:00 AM PT								
09:00 AM PT								
10:00 AM PT								
11:00 AM PT								
Noon								
01:00 PM PT								
02:00 PM PT								
03:00 PM PT								
04:00 PM PT								
05:00 PM PT								
06:00 PM PT								
07:00 PM PT								
08:00 PM PT								
09:00 PM PT								
10:00 PM PT								
11:00 PM PT								

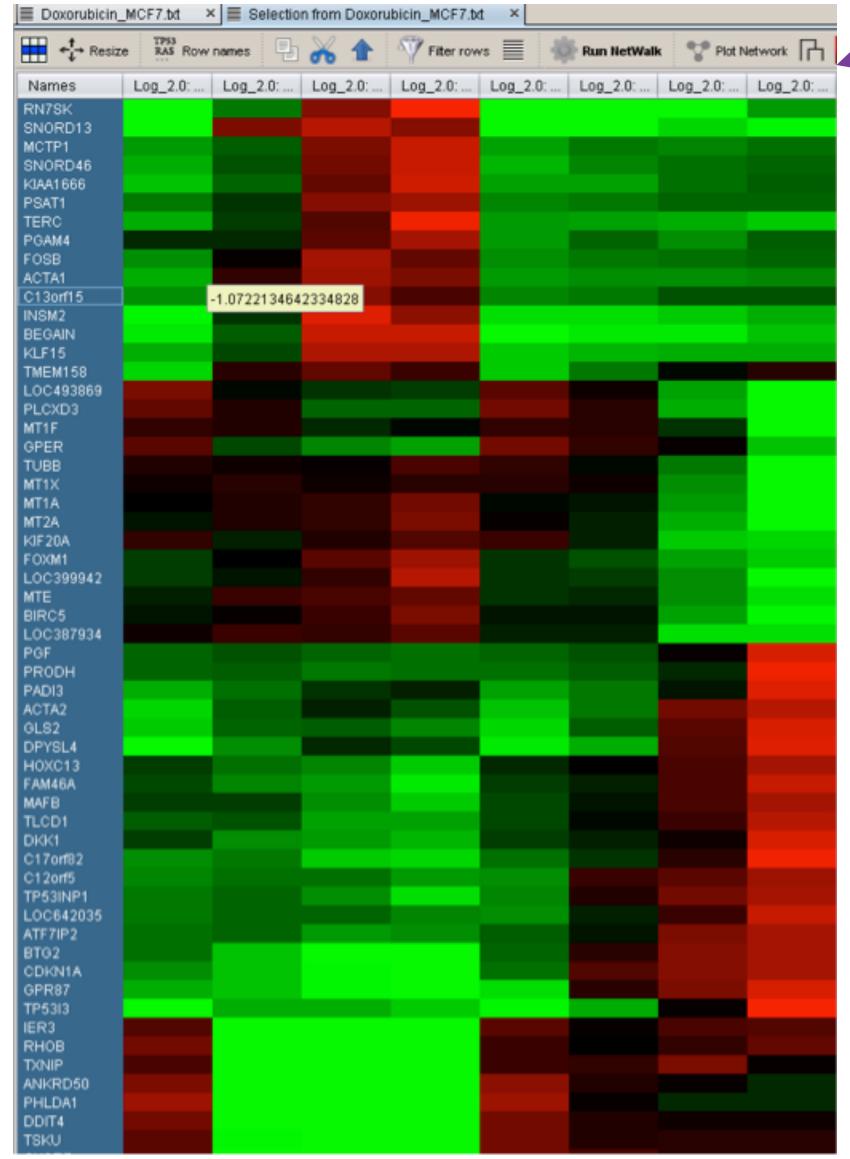
time

FASTEST SLOWEST

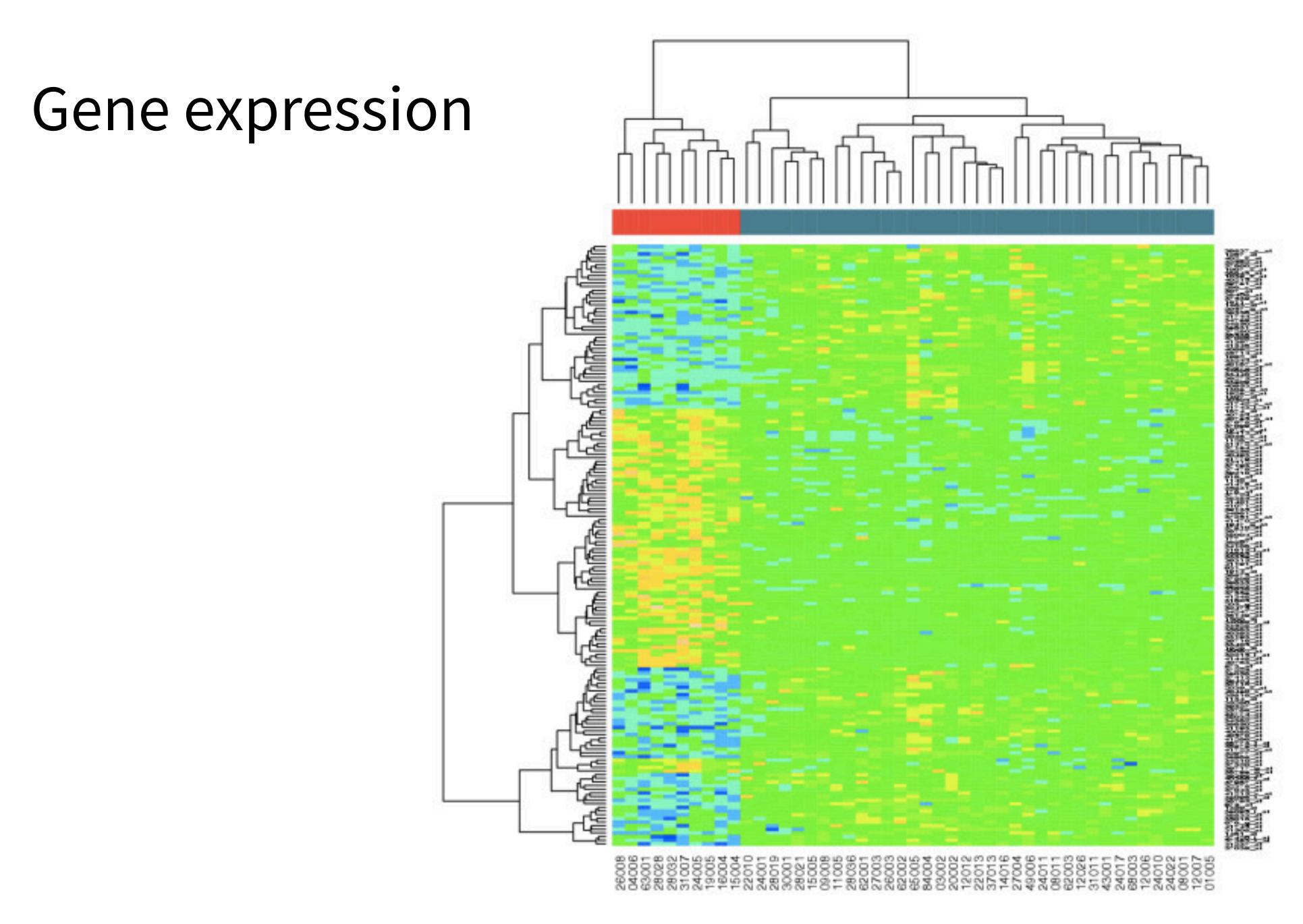
Gene expression

clustered samples





(normalized data)



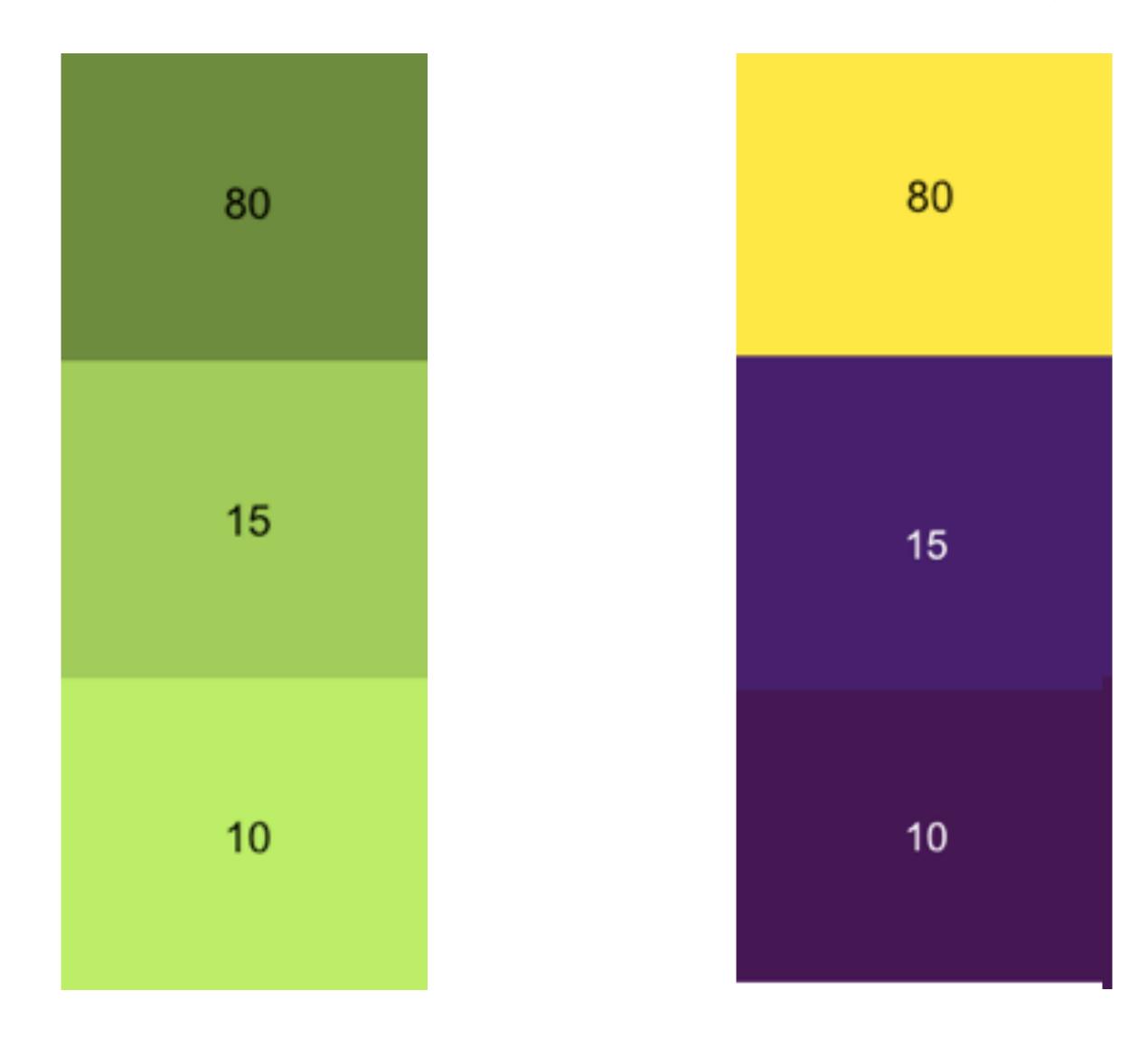
Treemap



Time to code...

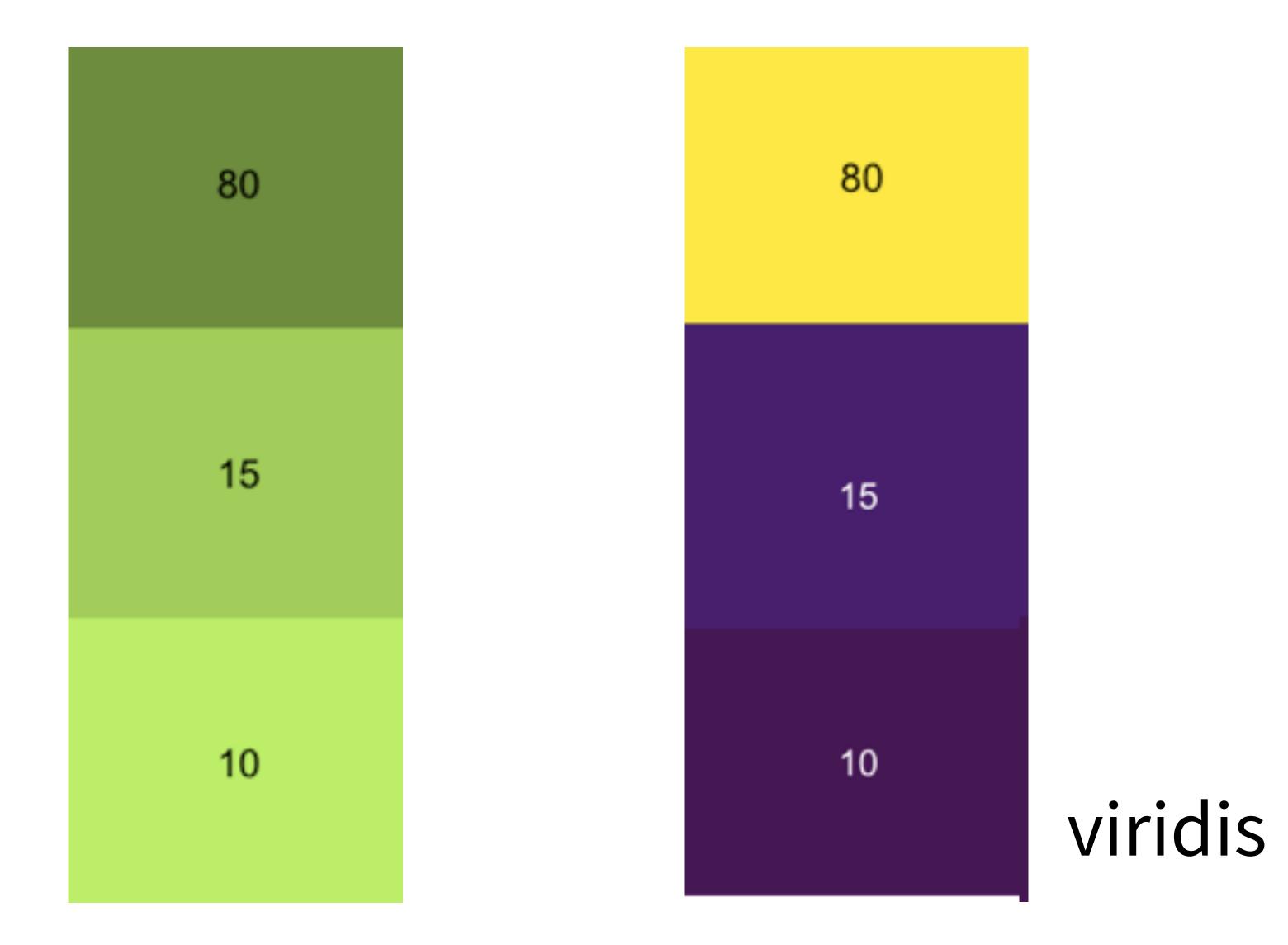
Perceptually uniform color space 1





Perceptually uniform color space 1





Perceptually uniform color spaces

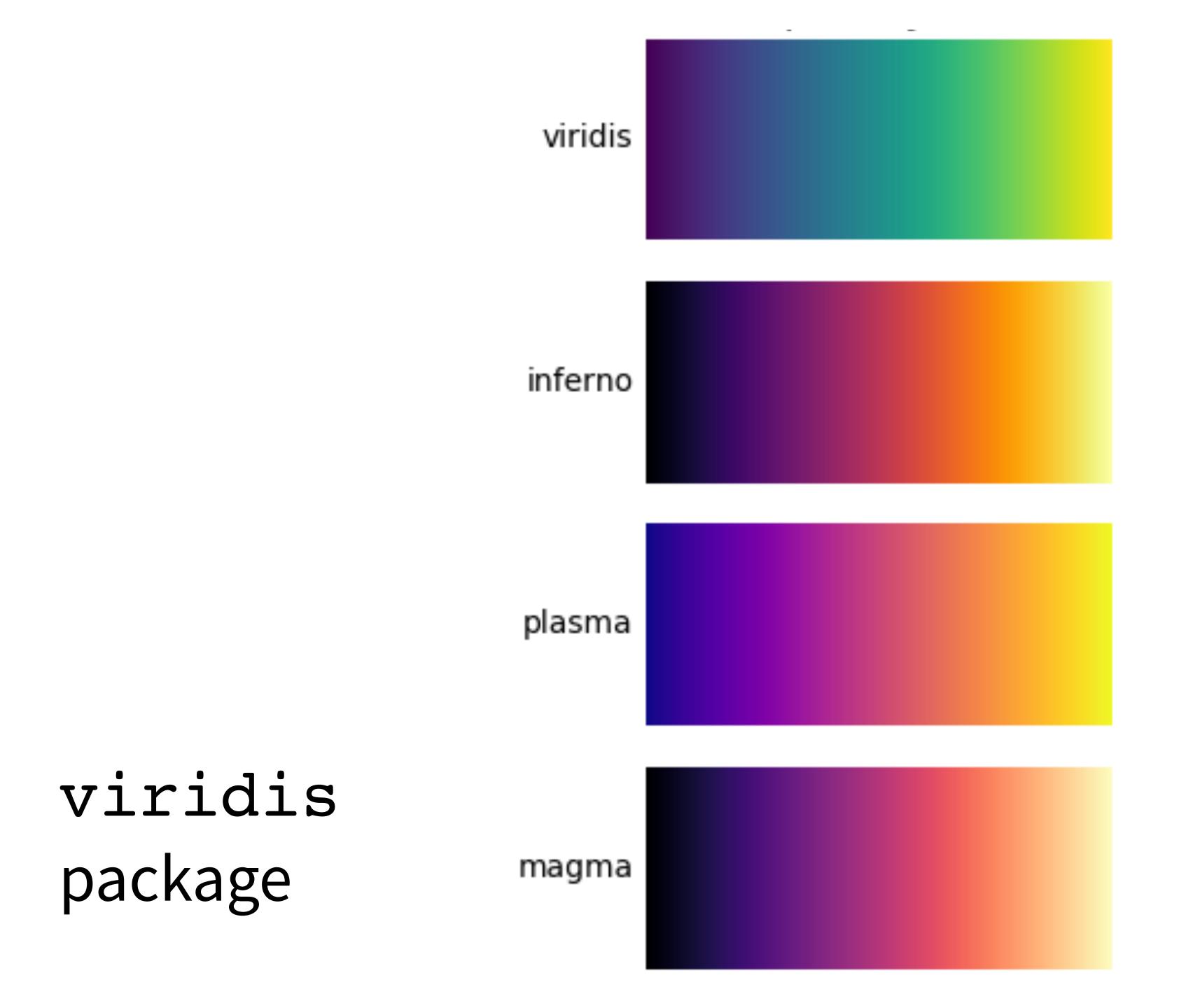
- don't blur important distinctions in the data
- don't add distinctions that don't exist in the data viridis package

tl;dr

Use the color scales in this package to make plots that are pretty, better represent your data, easier to read by those with colorblindness, and print well in grey scale.

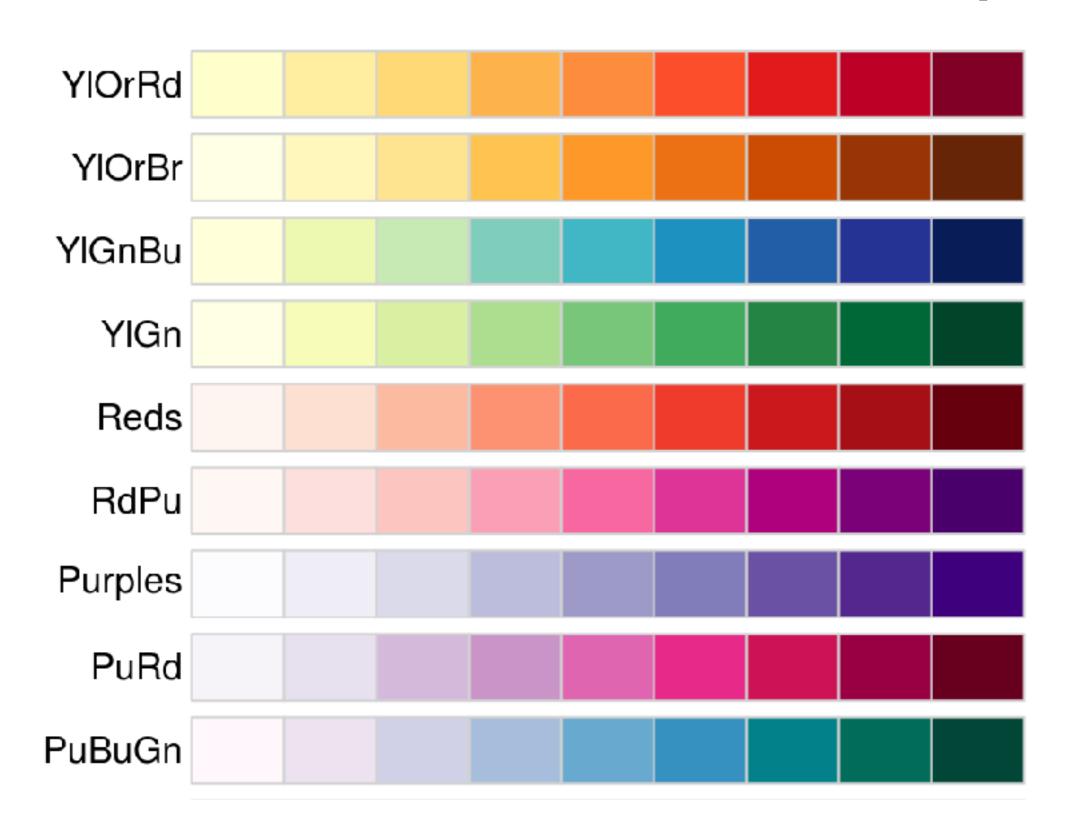
http://bids.github.io/colormap/

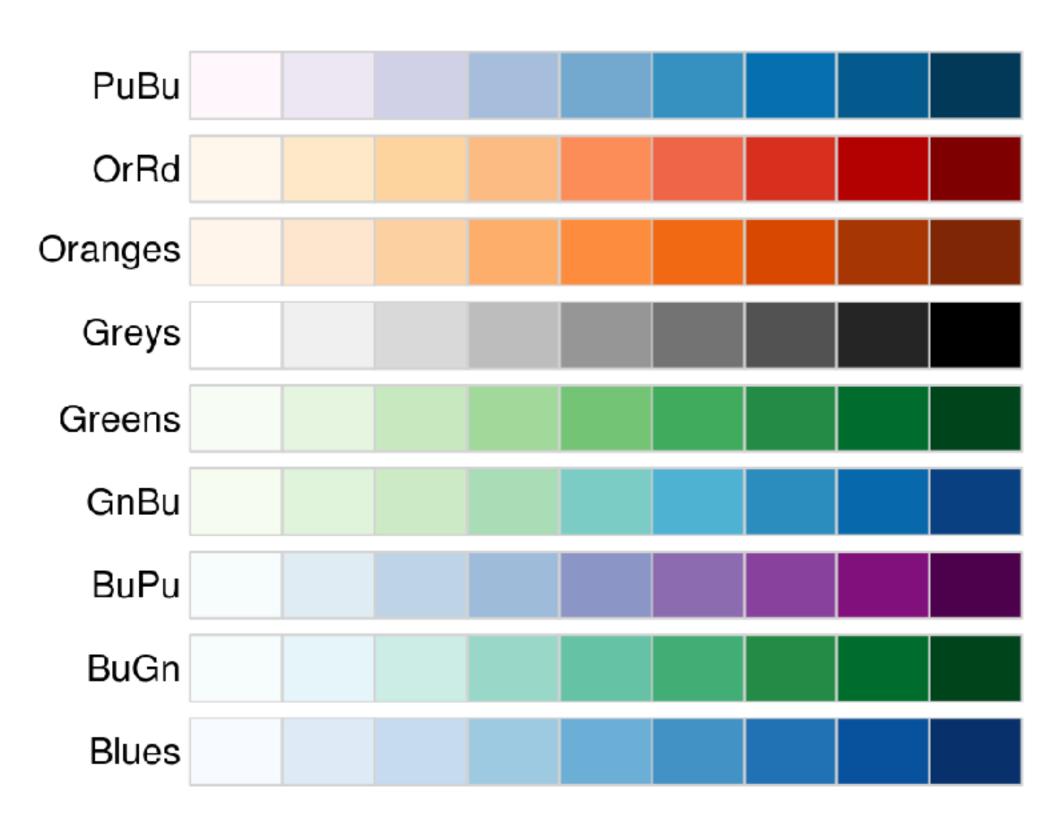
http://matplotlib.org/users/colormaps.html



RColorBrewer Color Schemes

sequential





RColorBrewer Color Schemes

diverging



RColorBrewer Color Schemes

qualitative (for categorical data)



Continuous data

VIRIDIS

+ scale_color_viridis()



RCOLORBREWER

```
+ scale_color_distiller(palette = "PuBu")
```

[+scale_color_brewer(palette = "PuBu"): Error: Continuous value supplied to discrete scale]

[+scale_color_continuous(palette = "PuBu") Error in f(..., self = self) : attempt to apply non-function]

CREATE YOUR OWN

```
+ scale_color_gradient(low = "white", high = "red")
```

- + scale_color_gradient2(low = "red", mid = "white", high = "blue", midpoint = 50)
- + scale_color_gradientn(colours = c("red", "pink", "lightblue", "blue"))

Discrete data

+ scale_color_manual(values = c("red", "yellow", "blue"))

```
VIRIDIS
+ scale_color_viridis() Error: Discrete value supplied to continuous scale
+ scale_color_viridis(discrete = TRUE)
RCOLORBREWER
+ scale_color_brewer(palette = "PuBu")
[+scale_color_discrete(palette = "PuBu") Error in f(..., self = self) : attempt to apply non-function]
+ scale_fill_grey()
CREATE YOUR OWN
```

Be careful...

```
> g + scale_fill_economist()
Error: Continuous value supplied to discrete scale
> g + scale_fill_viridis()
Error: Discrete value supplied to continuous scale
```

Color Vision Deficiency

approx. 8% of men, 0.5% of women have some form

```
missing or deficient cones:
protanopia (red)
deuteranopia (green)
tritanopia (blue)
```

How to make CVD friendly graphs

Use palettes that have already been tested (see viridis help)

Use a CVD simulator

www.vischeck.com

http://www.color-blindness.com/coblis-color-blindness-simulator/

Use high contrast

Back to RStudio...