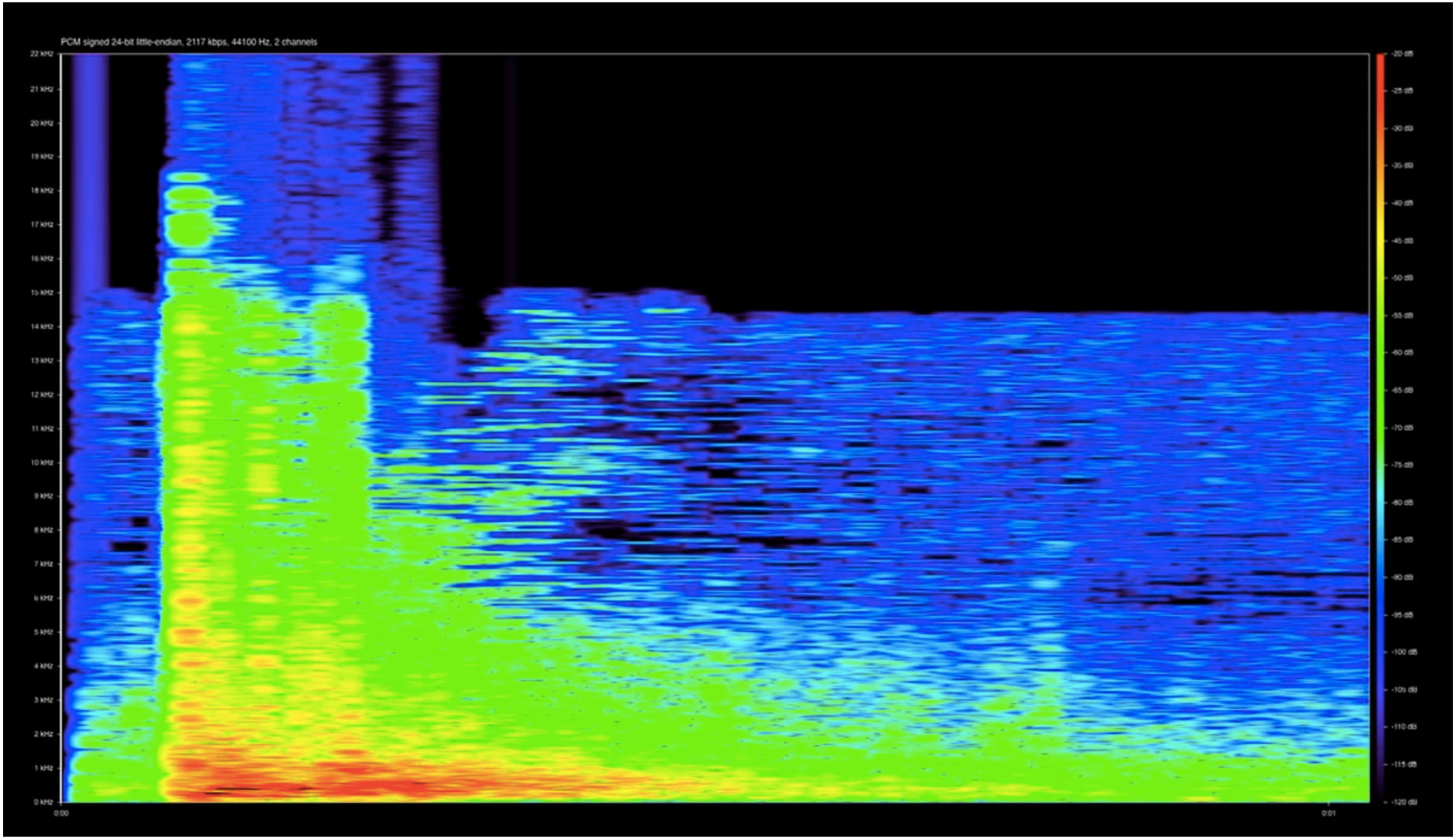
Data Visualization

Social Science Data Lab Series

MZES, December 16th, 2016

Richard Traunmüller, Goethe University Frankfurt

Making the Invisible Visible



http://beitunia.forensic-architecture.org/

On May 15th 2014, two teenagers were killed in a Nakba Day protest outside of Beitunia.

Nadeem Nawara (17) was shot in the chest, Mohammad Abu Daher (16) was shot in the back.

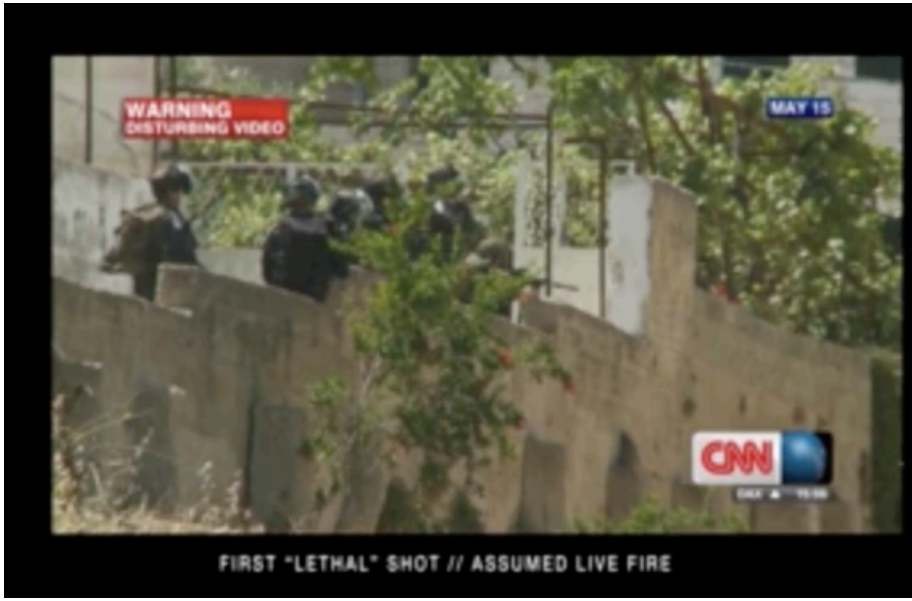
Security cameras of a local shop captured Nawara being mortally wounded.

An additional video shot by a local CNN crew shows Israeli soldiers discharging their weapons twice in the direction of protestors.

The video shows that the soldiers have rubber bullet extensions mounted on their M16 rifles.

No such visual material war available in the case of Abu Daher.

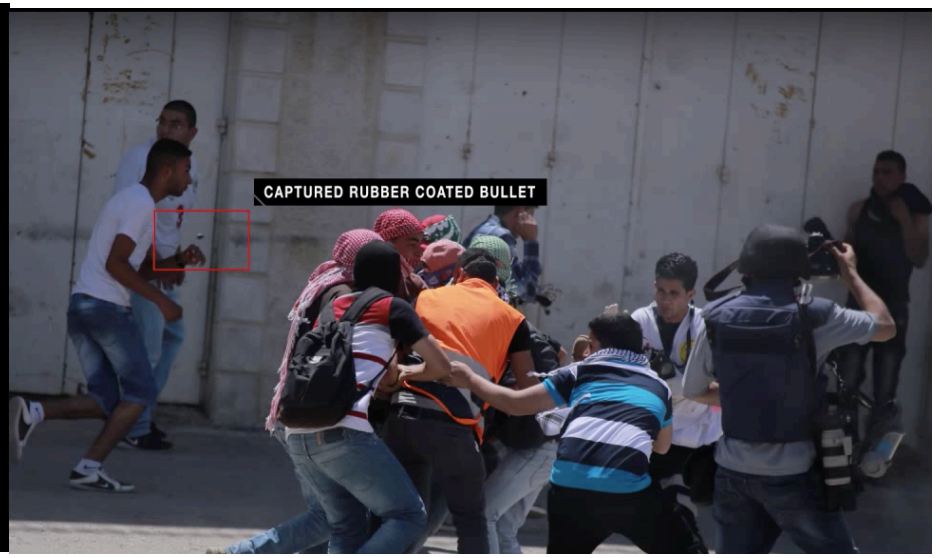
CNN live footage of first shot



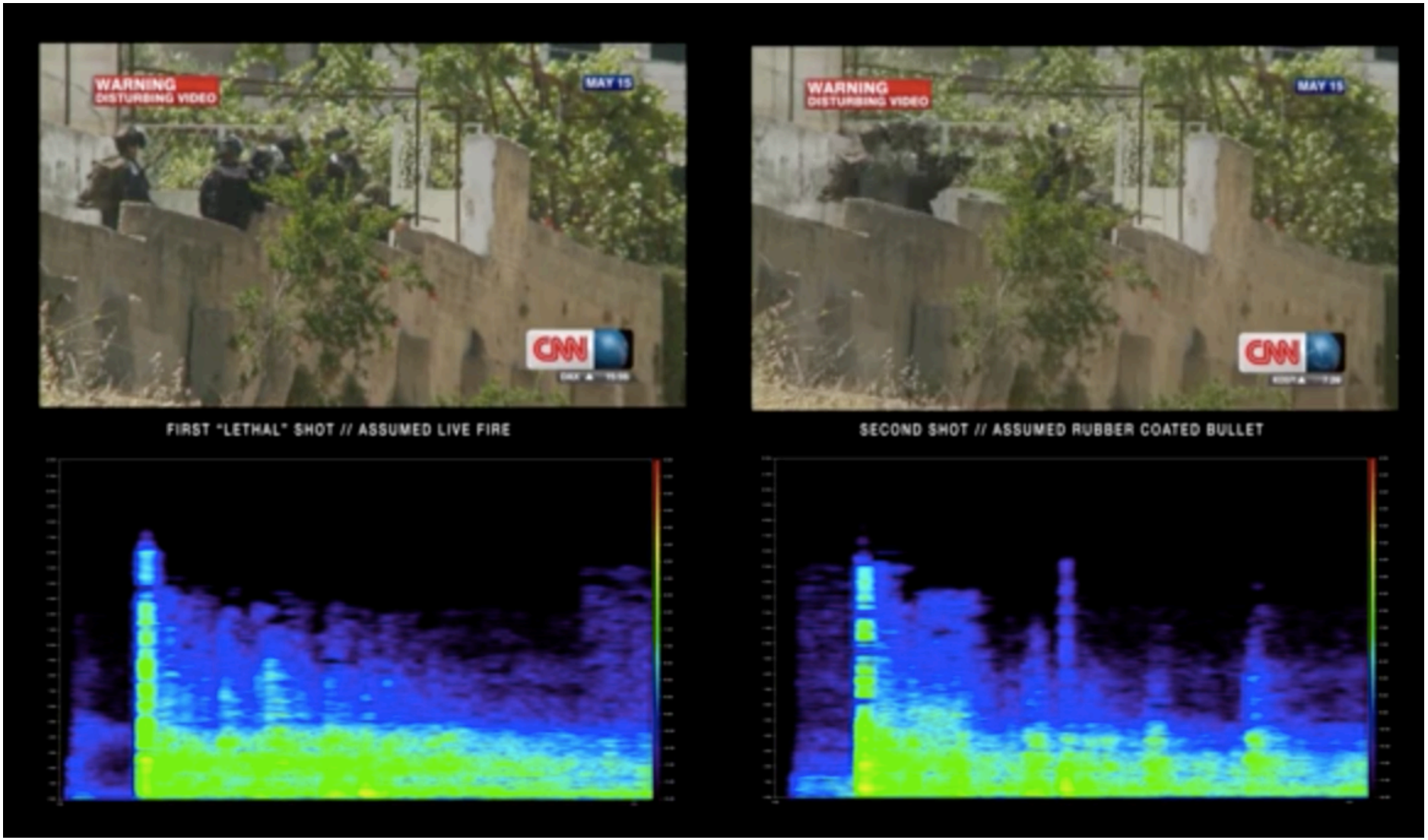
CNN live footage of second shot

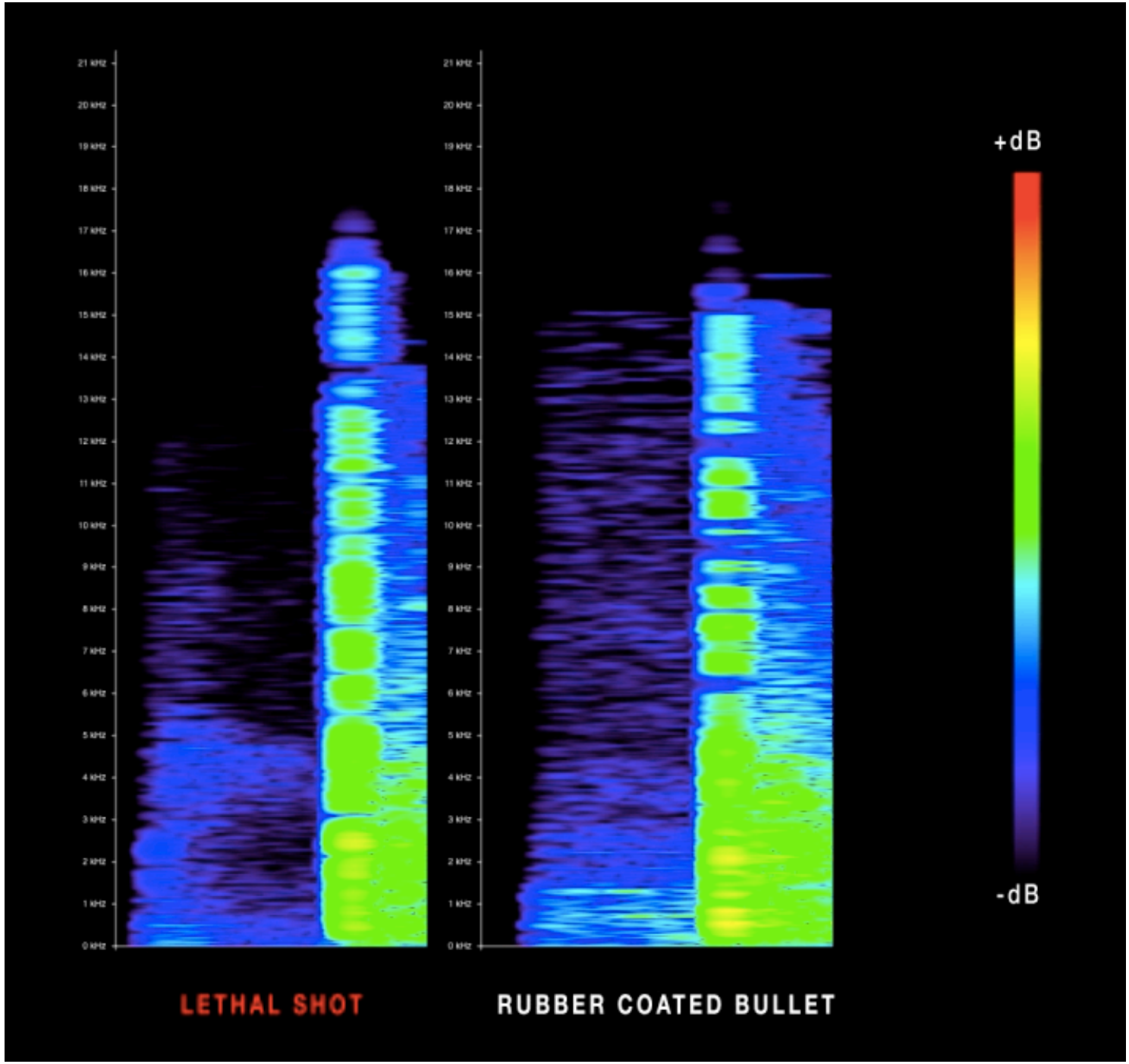
Nawara falling

Rubber bullet captured in video still



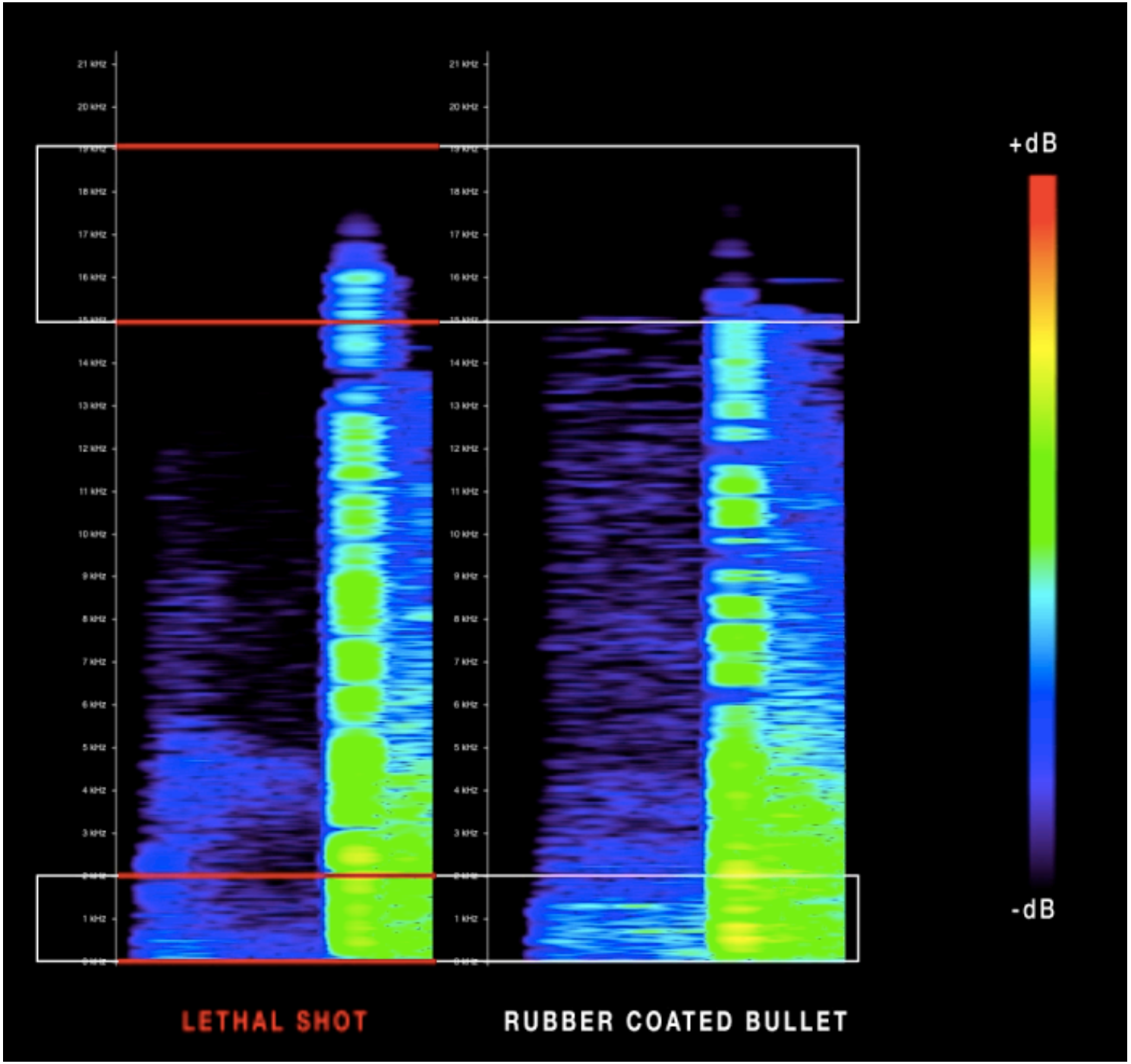
Sound signatures of the two shots (using spectrographs)




21 kHz
20 kHz
19 kHz
18 kHz
17 kHz
16 kHz
15 kHz
14 kHz
13 kHz
12 kHz
11 kHz
10 kHz
9 kHz
8 kHz
7 kHz
6 kHz
5 kHz
4 kHz
3 kHz
2 kHz
1 kHz
0 kHz
+dB
-dB
LETHAL SHOT
RUBBER COATED BULLET

The leathal shot is louder in the high frequencies.

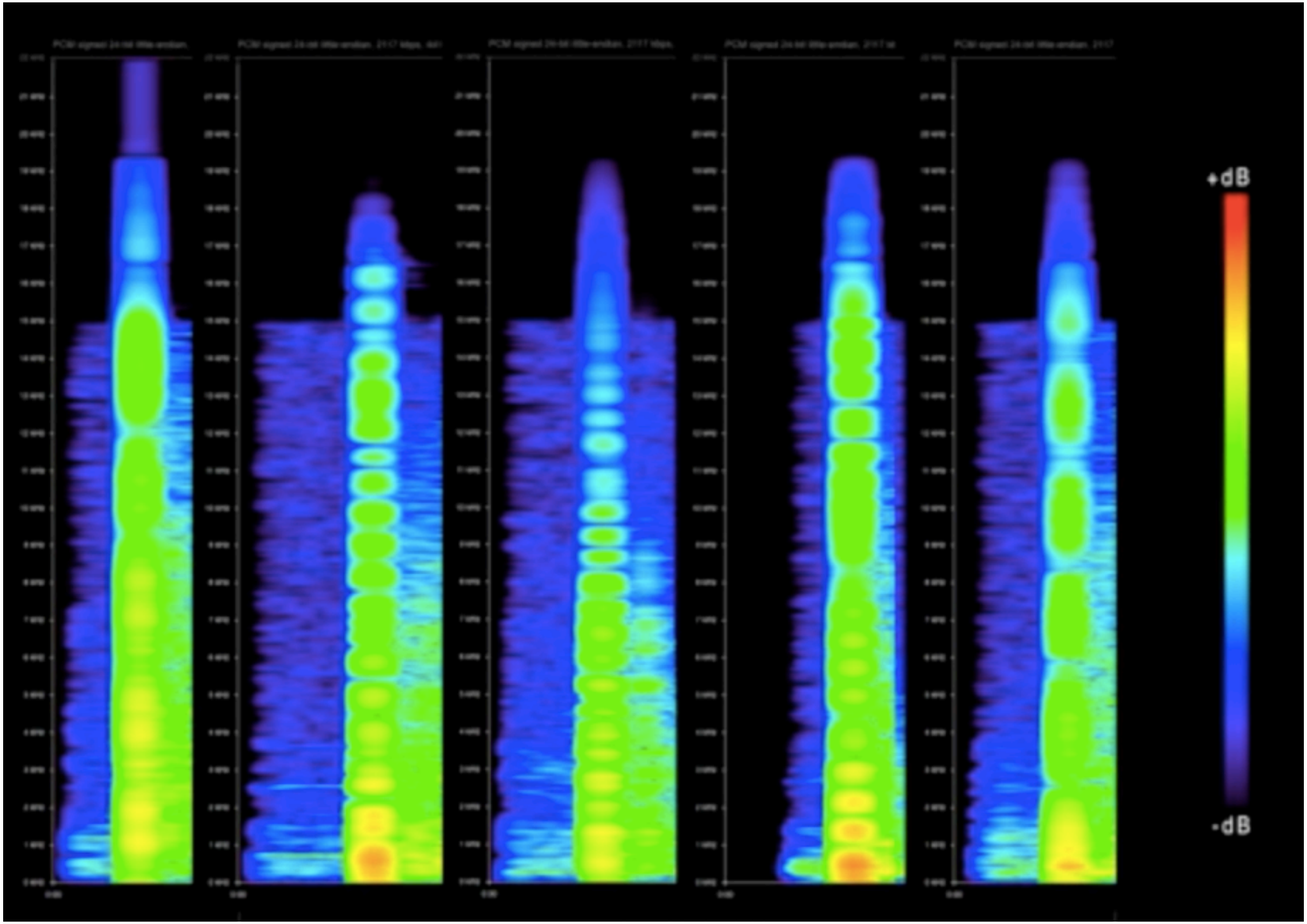
The rubber coated bullet is louder in the low frequencies.



Later that day, live footage by the news network PBC Palestine recorded the sound of five more gunshots.

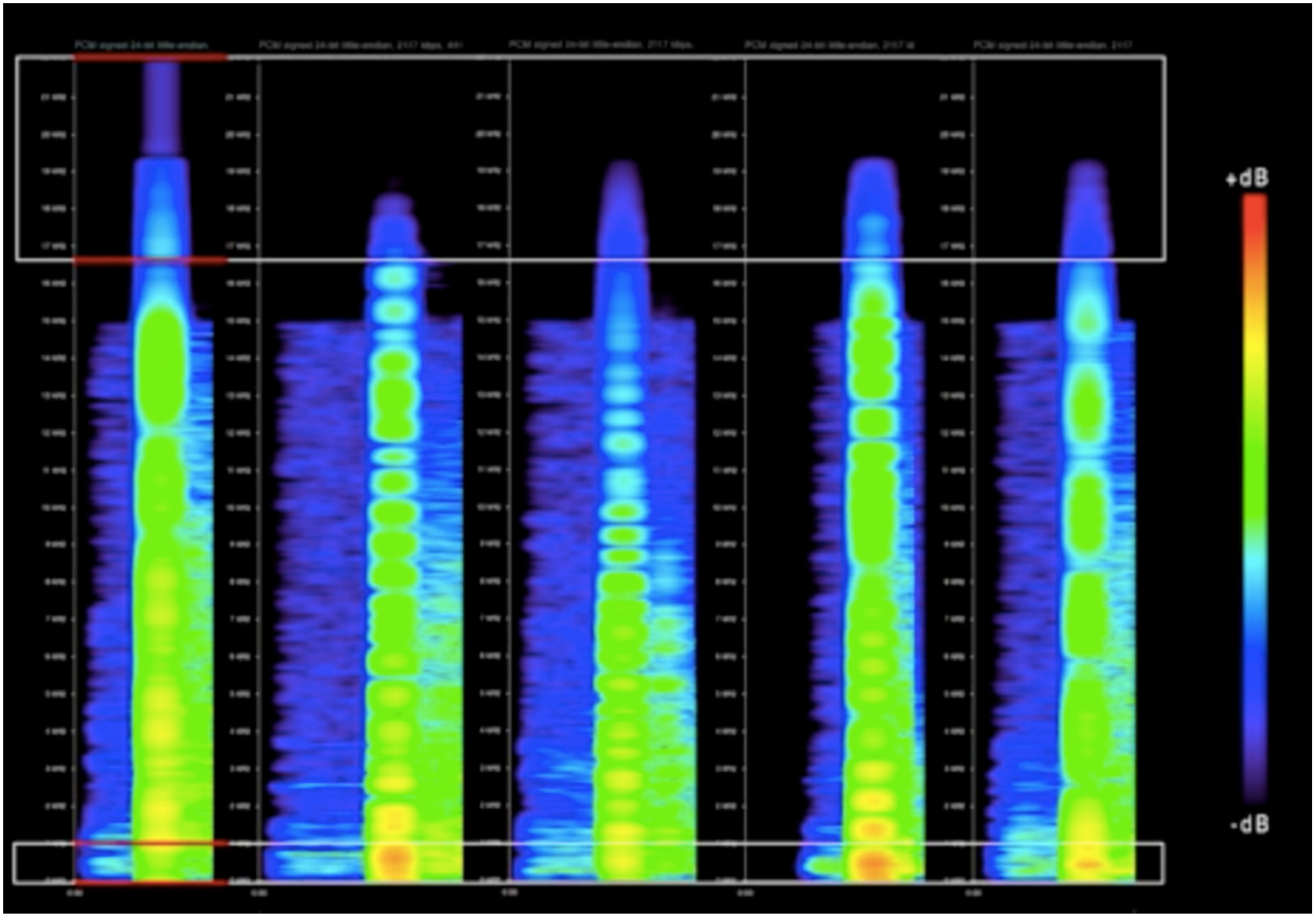
This included the gunshot that killed Mohammad Abu Daher.

Comparison of the sound signatures of the five shots captured on PBC Palestine footage



Comparison of the sound signatures of the five shots captured on PBC Palestine footage

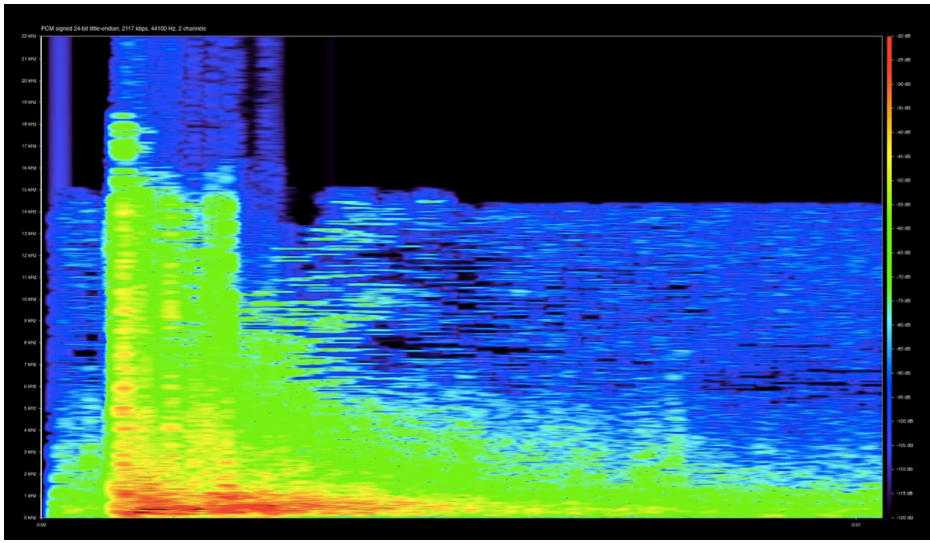
The leathal shot is louder in the high frequencies.



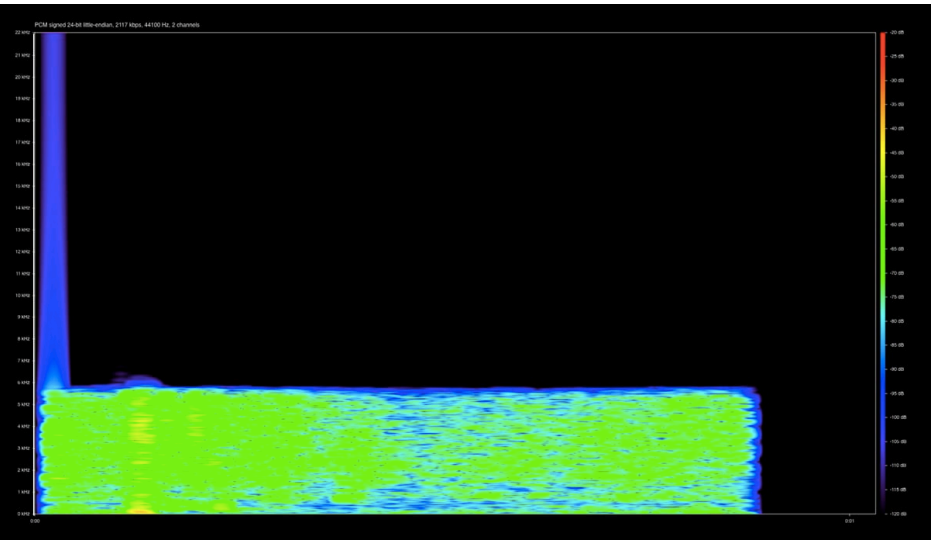
The leathal shot is not as loud in the low frequencies.

Further visual comparisons of distinct sound signatures of M16 rifle fire

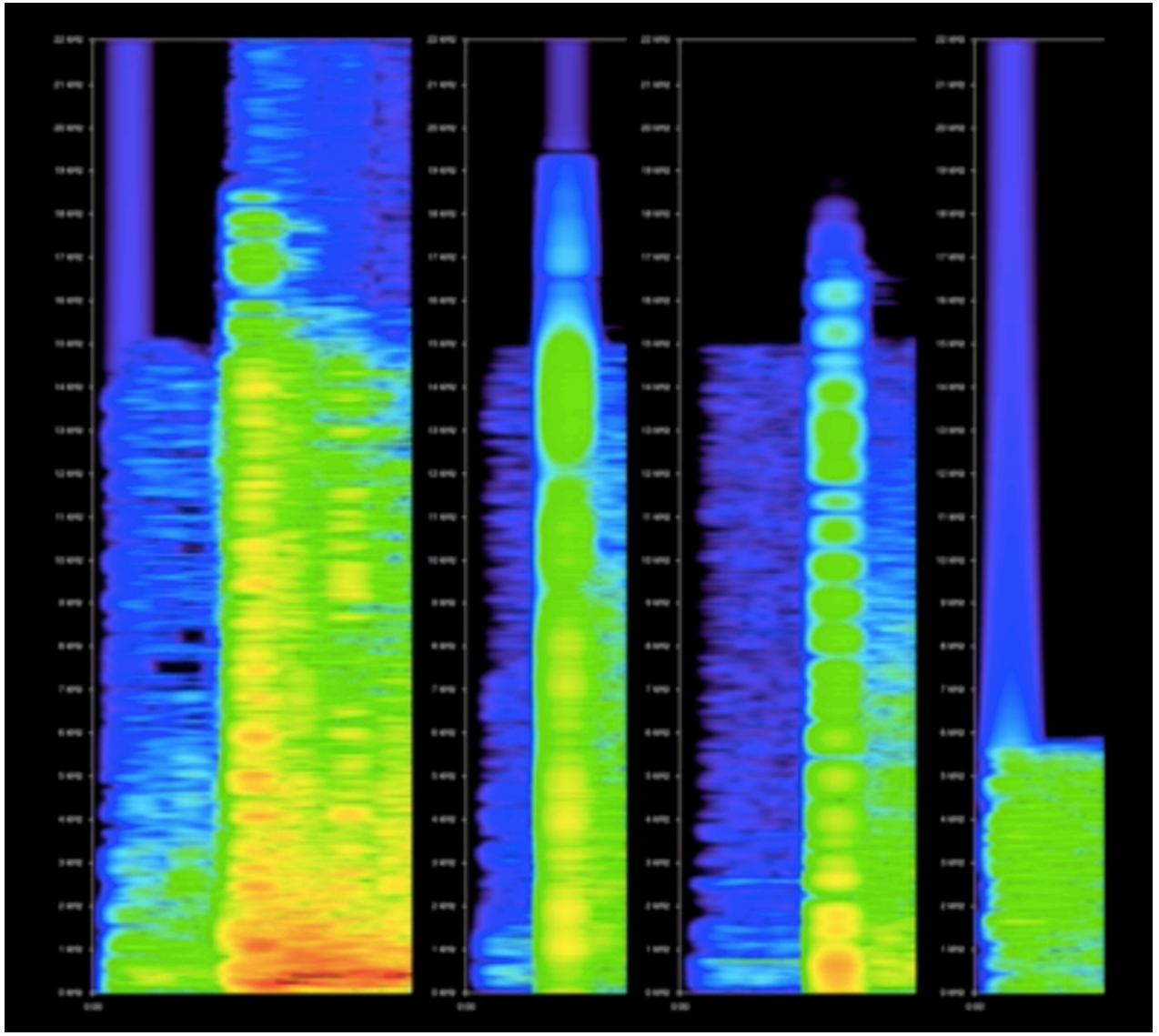
M16 rifle live fire without extension



M16 rifle live fire with silencer mounted



Further visual comparisons of distinct sound signatures of M16 rifle fire



Conclusion

Gunfire has a sound signature that can be made visible in a spectrograph.

Nawara was shot to death with live ammunition and not a rubber coated bullet.

We have identified a distinct sound signature that is a mix of live fire and rubber coated ammunition – live ammunition fired through a rubber bullet extension.

This particular sound signature is similar to the shot that killed Abu Daher.

This visual sound signature is important because its suppressed sound reveals the intention to disguise an act of murder.

What Can Learn From This Visualization Example?

1. Visualization is powerful tool, that let's us see things that would otherwise be invisible.
2. Visualization is very much like detective work.
3. Visualization is a great tool for persuasive communication.
4. Comparison is key.

Workshop Goals

1. Get to know some fundamental principles and tools of data visualization
2. How to do it in R

Course Outline

Principles of Graphical Data Analysis

Graphical Perception

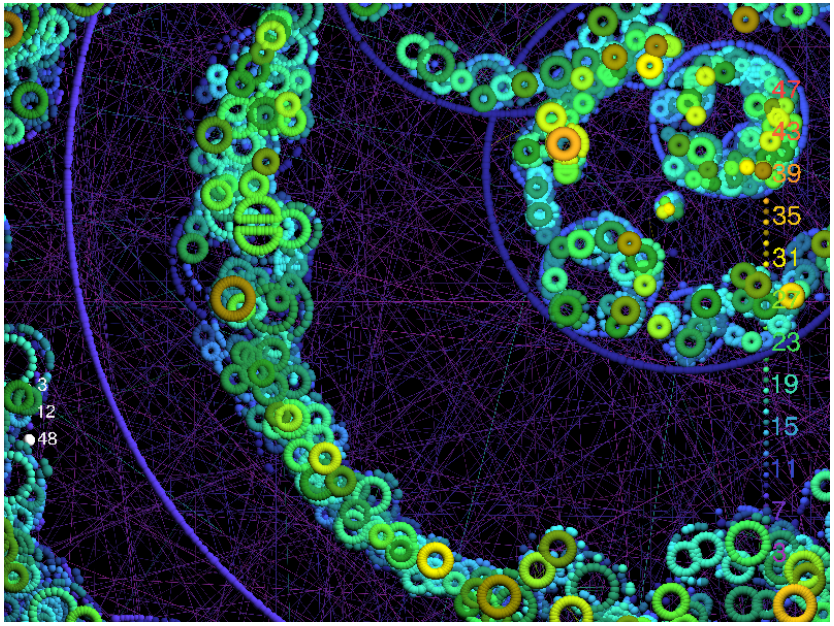
Visualization of Bivariate Data

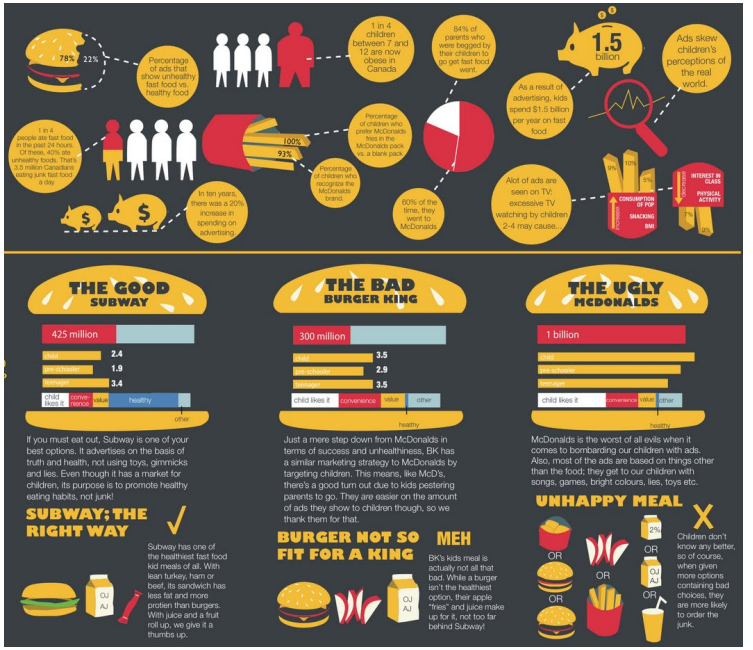
Visualization of Multivariate Data

Visual Inference

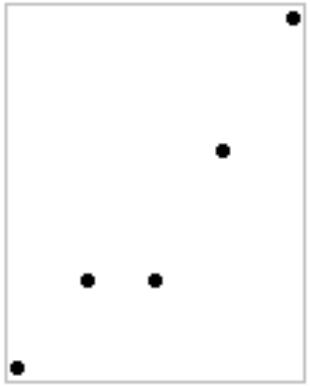
Visualization of Statistical Models

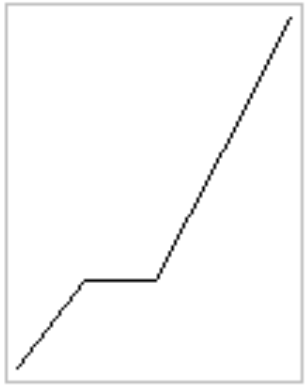
You will probably be disappointed if you expect to see things like this…

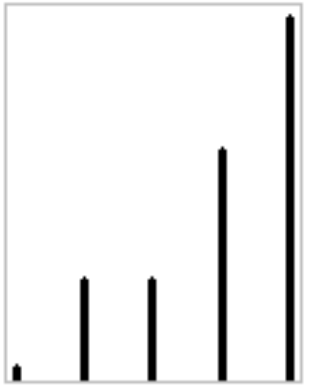




Instead I hope to convince you, that…







I know, I know – it's a bit like hoping for this...

And getting this...

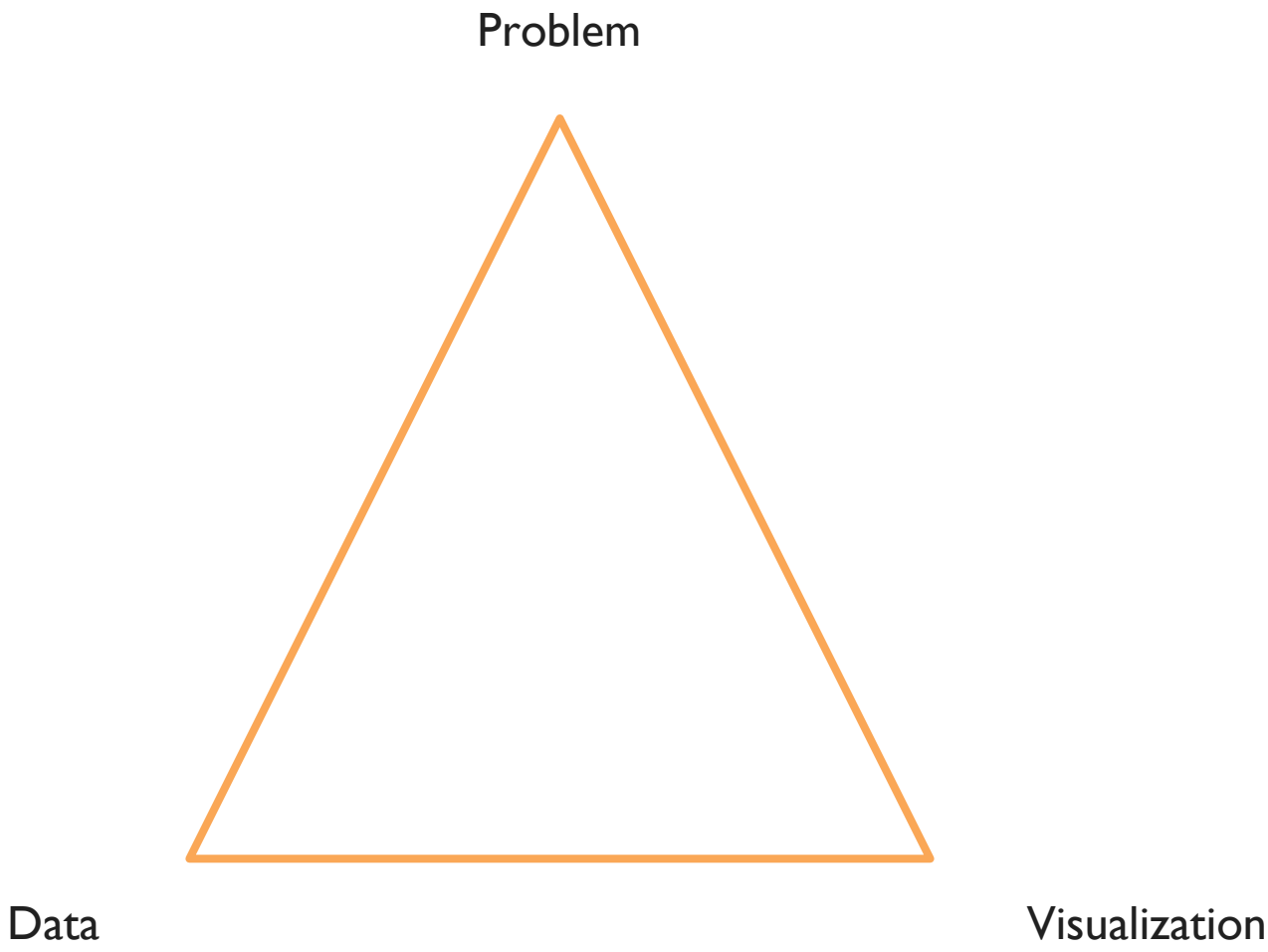
Data Visualization as a Methodology

"The critical question is how best to transform the data into something that people can understand for optimal decision making." (Ware 2013: 5)

Data visualization is a method for making sense of quantitative information – not to make pictures of data.

Note that this is more than data visualization in the narrow sense, i.e. the act of encoding quantitative information in visual objects.

Data Visualization as a Methodology



The Case for Visualization

Two ways to make sense of quantitative information:

Table Look-up

2	4	16

Pattern Perception

Social scientists are mostly interested in patterns, not in individual and exact values.

We are trying to make sense of relationships among data points and want to compare more than two values at a time.

The Case for Visualization

To see relationships among data – patterns, trends, and exceptions – we need a picture.

Visualization gives data form or shape, which allows us to see things that are otherwise difficult or impossible to see.

Advantages of data visualization

Useful summaries for large, complicated data sets – in fact, the utility of visualization increases with data size.

Little or no assumptions about the nature of the data.

Facilitates interaction between researcher and data – it's a hypothesis generating device.

I.		II.		III.		IV.	
y1	*x1*	*y2*	*x2*	*y3*	*x3*	*y4*	*x4*
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

I.		II.		III.		IV.	
y1	*x1*	*y2*	*x2*	*y3*	*x3*	*y4*	*x4*
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8
Mean(*y*)	7.50		7.5		7.50		7.5
Mean(*x*)	9.0		9.0		9.0		9.0

I.		II.		III.		IV.	
y1	*x1*	*y2*	*x2*	*y3*	*x3*	*y4*	*x4*
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

Mean(*y*)	7.50	7.5	7.50	7.5
Mean(*x*)	9.0	9.0	9.0	9.0
SD(*y*)	2.03	2.03	2.03	2.03
SD(*x*)	3.32	3.32	3.32	3.32

I.		II.		III.		IV.	
y1	*x1*	*y2*	*x2*	*y3*	*x3*	*y4*	*x4*
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

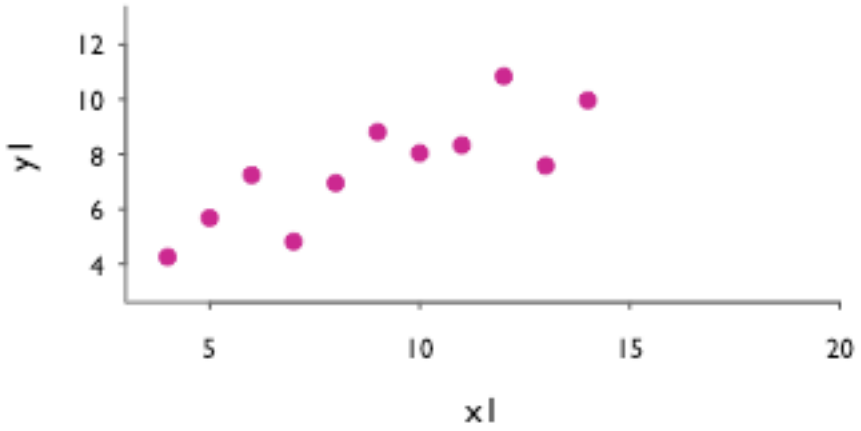
Mean(*y*)	7.50	7.5	7.50	7.5
Mean(*x*)	9.0	9.0	9.0	9.0
SD(*y*)	2.03	2.03	2.03	2.03
SD(*x*)	3.32	3.32	3.32	3.32
r(*y*, *x*)	.82	.82	.82	.82

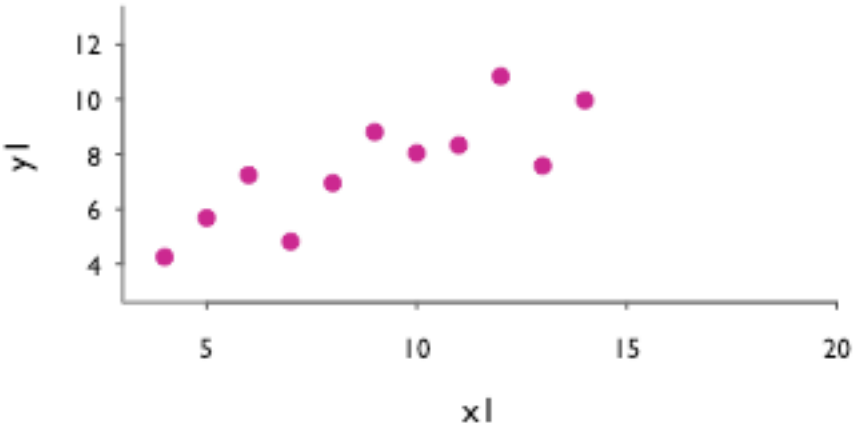
I.		II.		III.		IV.	
y1	*x1*	*y2*	*x2*	*y3*	*x3*	*y4*	*x4*
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

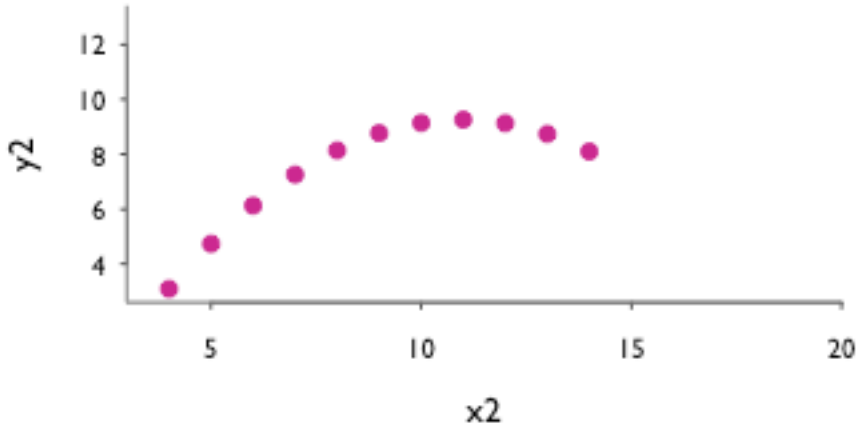
Mean(*y*)	7.50	7.5	7.50	7.5
Mean(*x*)	9.0	9.0	9.0	9.0
SD(*y*)	2.03	2.03	2.03	2.03
SD(*x*)	3.32	3.32	3.32	3.32
r(*y*, *x*)	.82	.82	.82	.82
y = a + b*x*	y = 3 + 0.5*x*	y = 3 + 0.5*x*	y = 3 + 0.5*x*	y = 3 + 0.5*x*

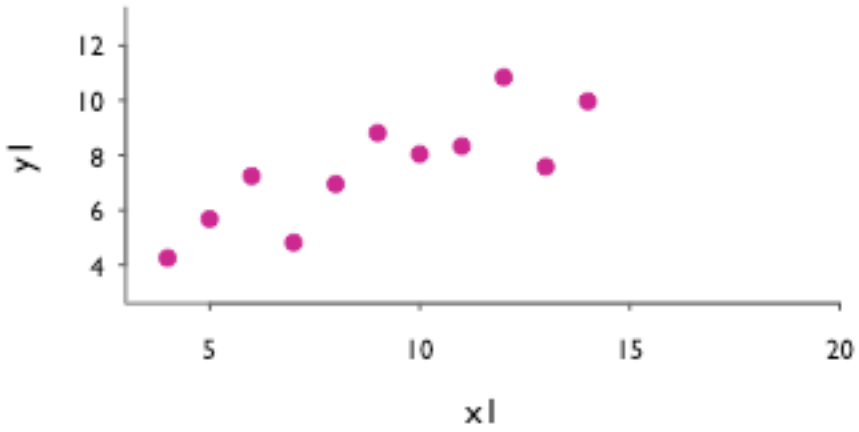
I.		II.		III.		IV.	
y1	*x1*	*y2*	*x2*	*y3*	*x3*	*y4*	*x4*
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

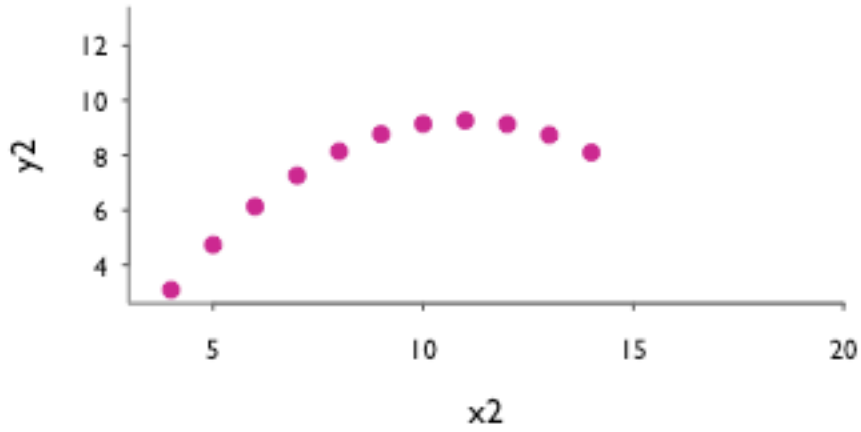
Mean(*y*)	7.50	7.5	7.50	7.5
Mean(*x*)	9.0	9.0	9.0	9.0
SD(*y*)	2.03	2.03	2.03	2.03
SD(*x*)	3.32	3.32	3.32	3.32
r(*y*, *x*)	.82	.82	.82	.82
y = a + b*x*	y = 3 + 0.5*x*	y = 3 + 0.5*x*	y = 3 + 0.5*x*	y = 3 + 0.5*x*
R^2	.67	.67	.67	.67

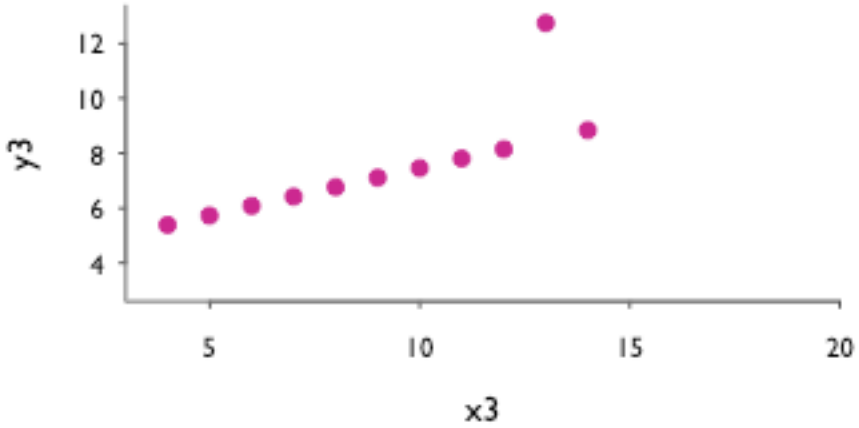

12
10
8
6
4
y1
5
10
15
20
x1

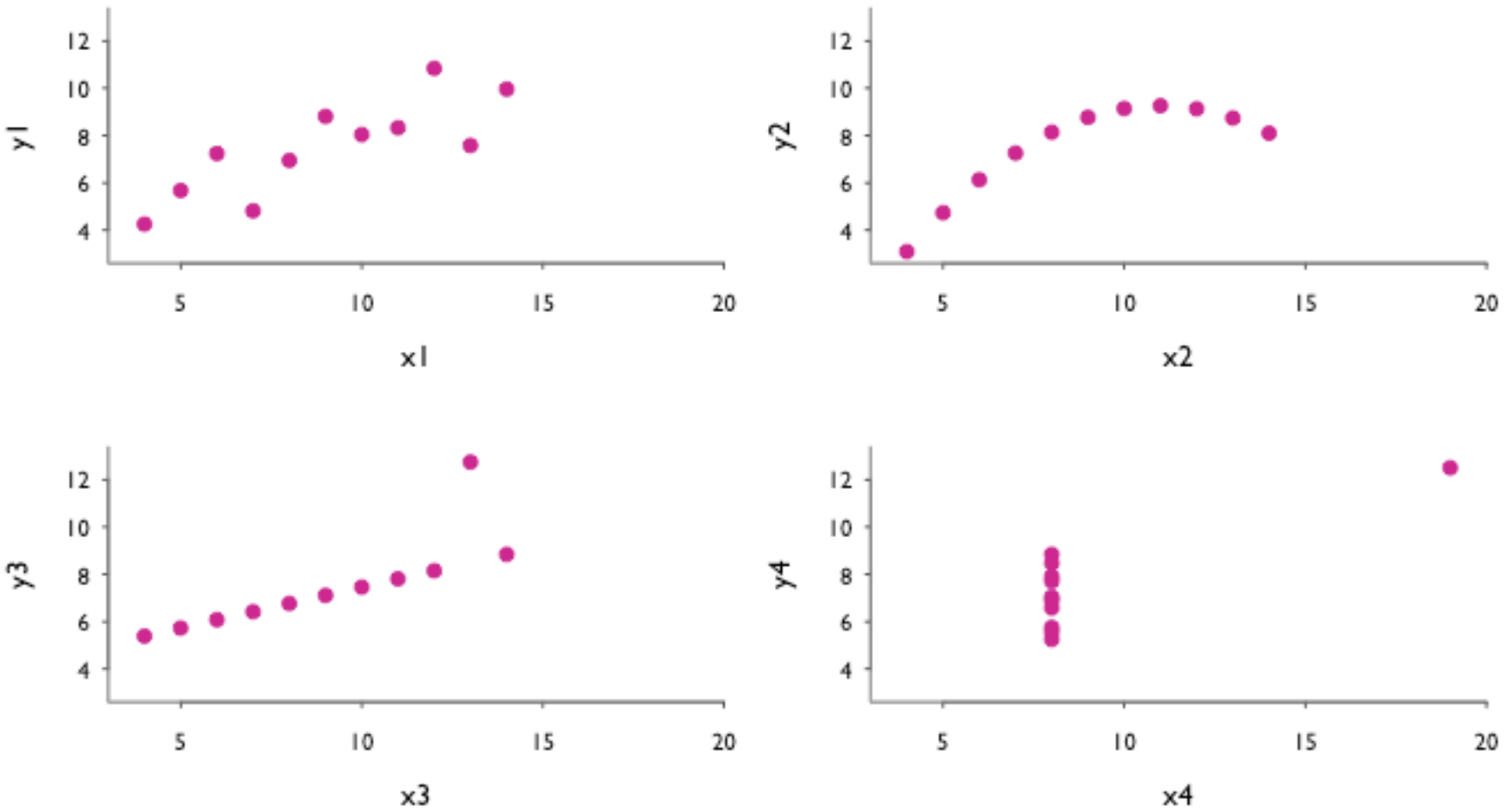

y1
12
10
8
6
4
5
10
15
20
x1

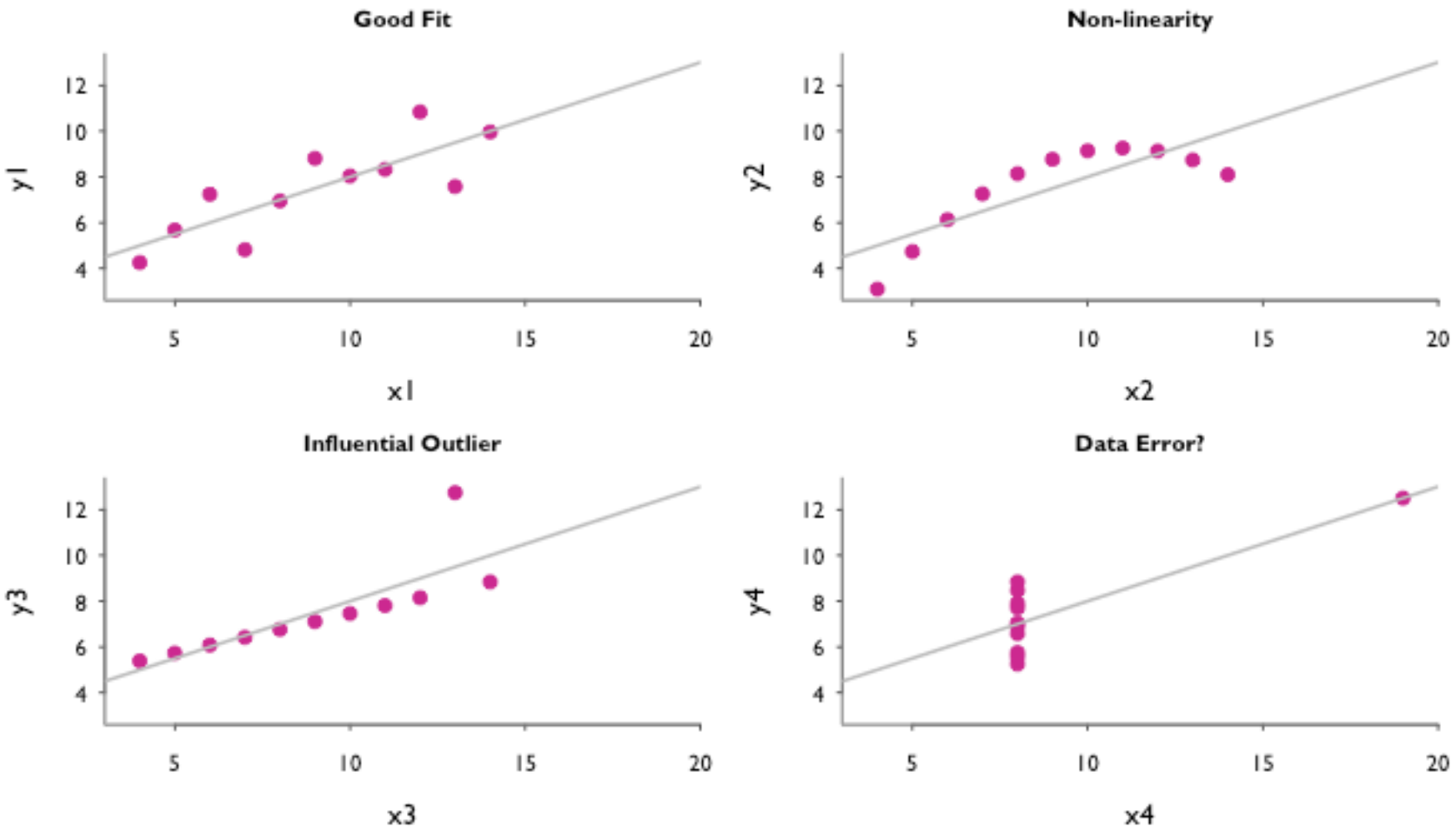

y2
12
10
8
6
4
5
10
15
20
x2


y1
12
10
8
6
4
5
10
15
20
x1


y2
12
10
8
6
4
5
10
15
20
x2


y3
12
10
8
6
4
5
10
15
20
x3

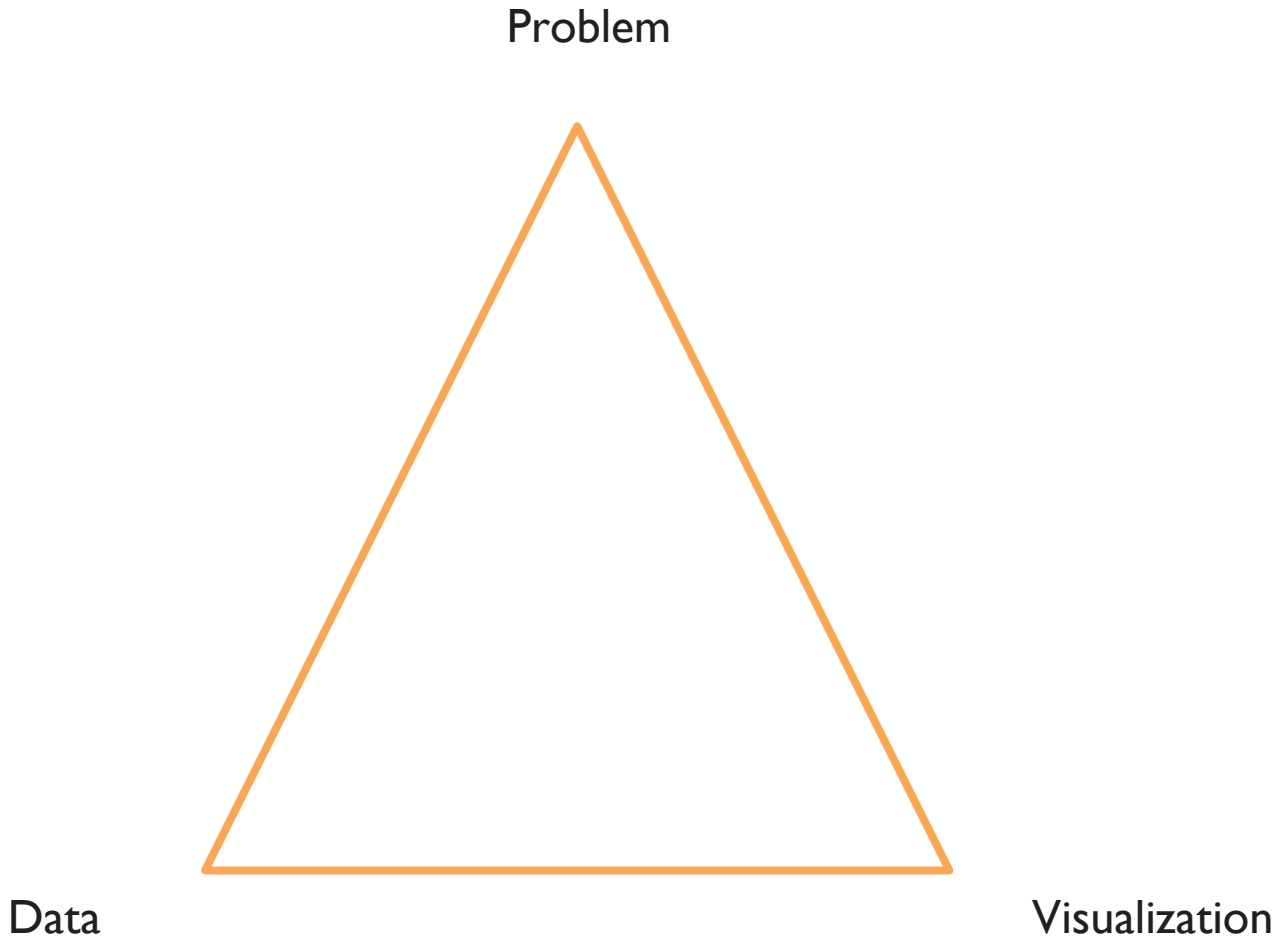

y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
5
10
15
20


Good Fit
y1
x1
Non-linearity
y2
x2
Influential Outlier
y3
x3
Data Error?
y4
x4
4
6
8
10
12
5
10
15
20

Conclusion

1. Data visualization allows us to see things we would otherwise miss
2. Data visualization reveals information beyond summary statistics and models
3. Data visualization complements statistical modelling

"Visualization is surprisingly difficult. Even the most simple matters can easily go wrong."
(W. Cleveland)



"Graphical excellence is […] a matter of *substance*, of *statistics*, and of *design*." (E. Tufte)

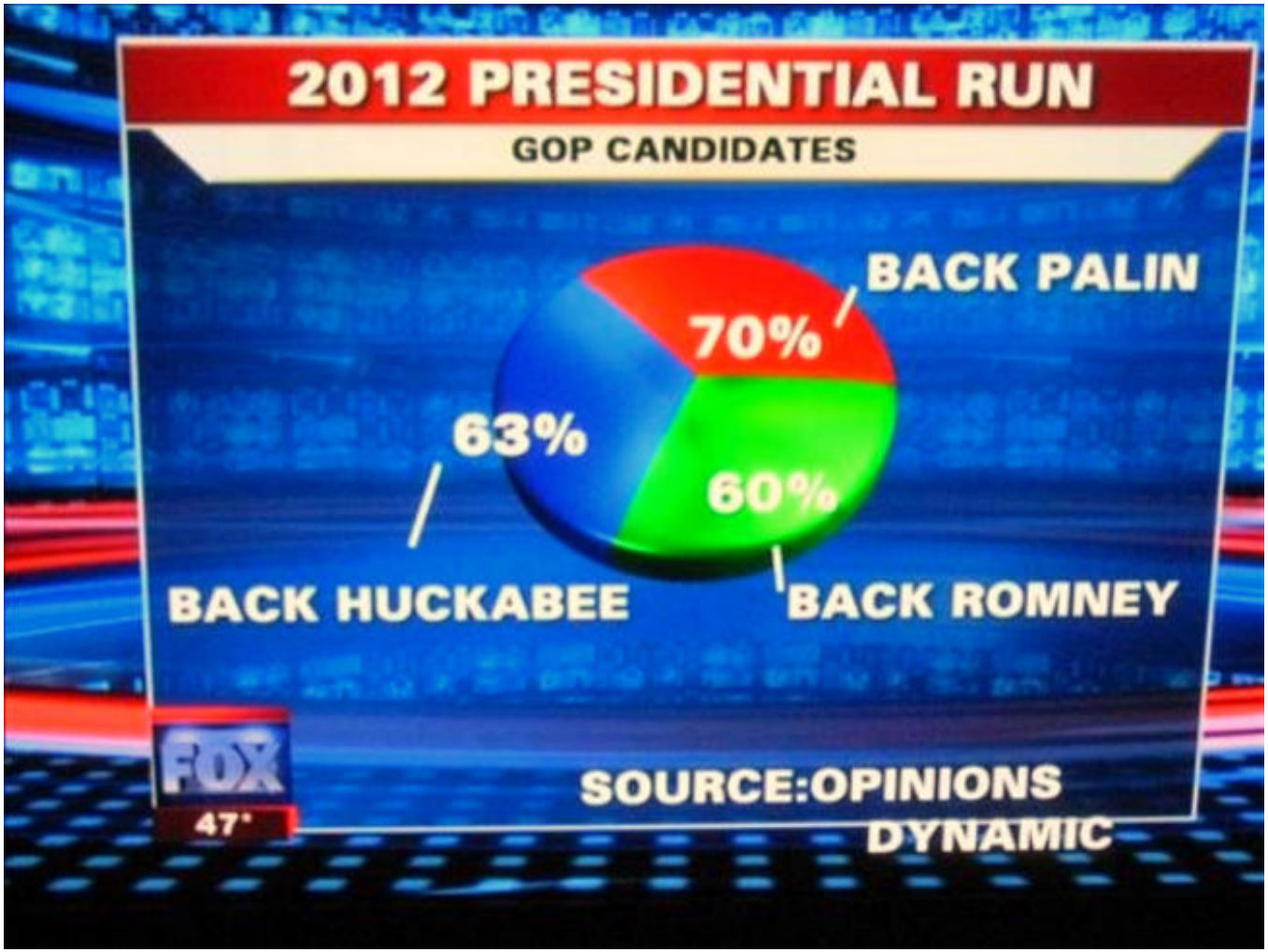
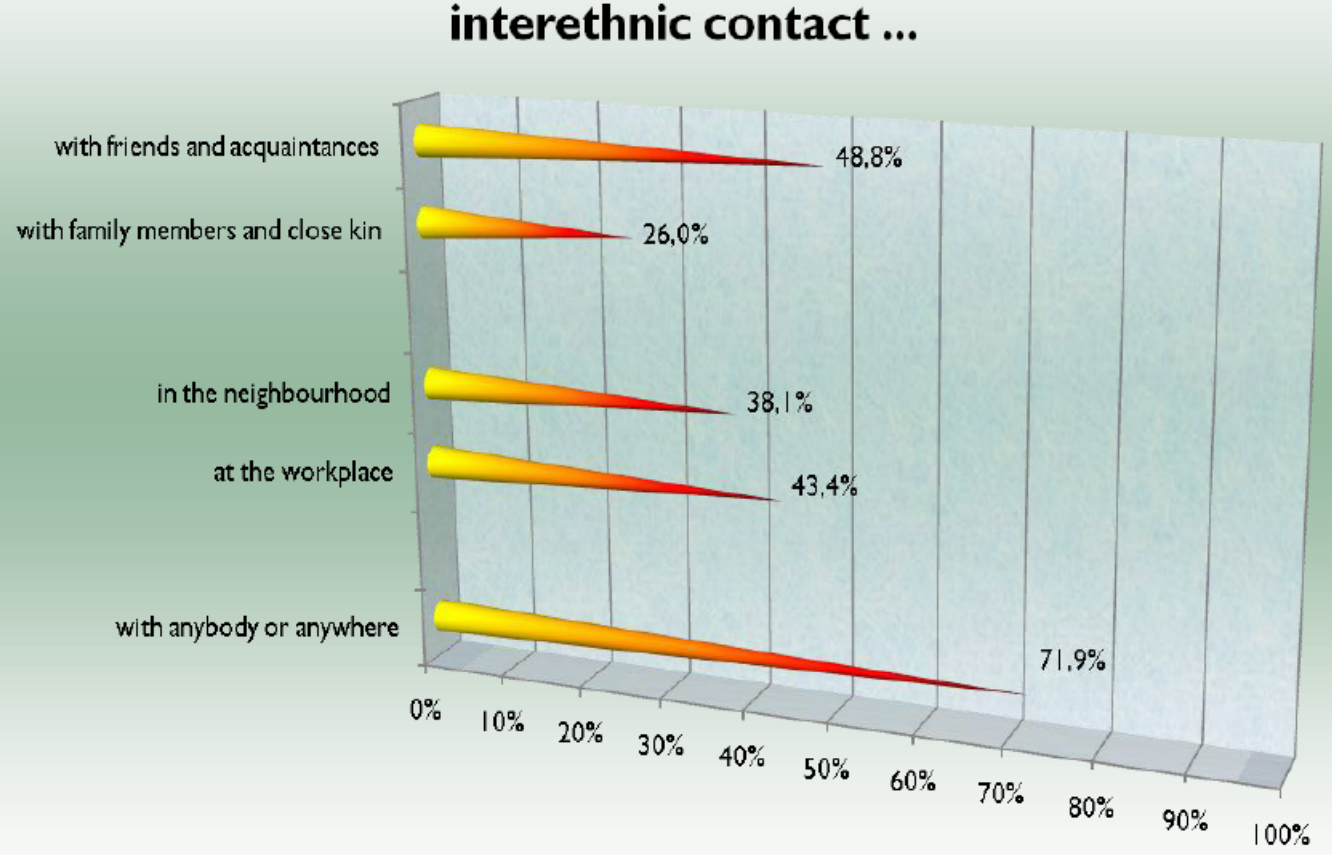

2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
47'
SOURCE:OPINIONS
DYNAMIC

Figure 1: Percentage interethnic contact of Germans

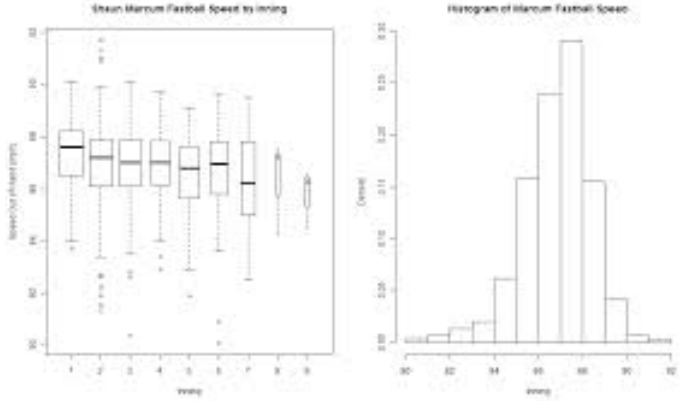


source: ALLBUS-survey 2006, n_{min}=3,063

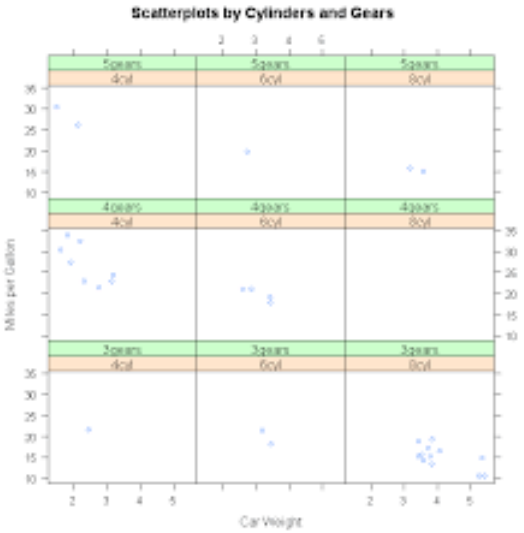
How to do it in R

R Graphic Packages

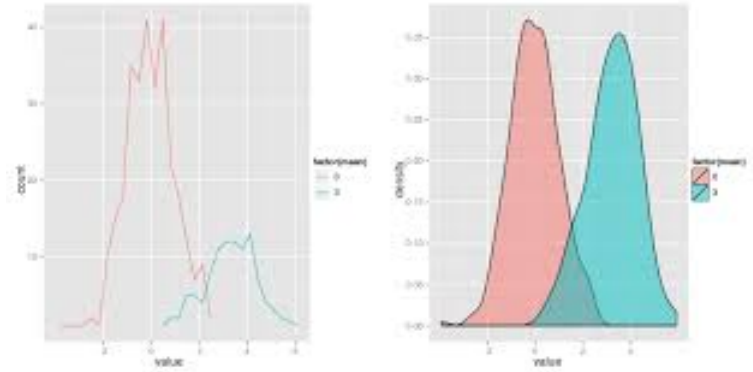
graphics



lattice



ggplot2



Graphics

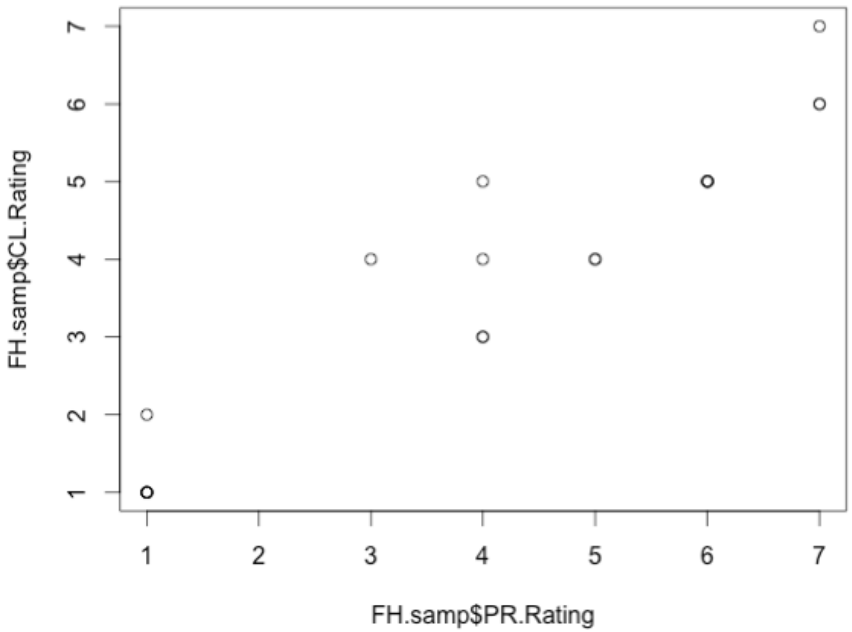
High-level commands: create a plot of a particular format

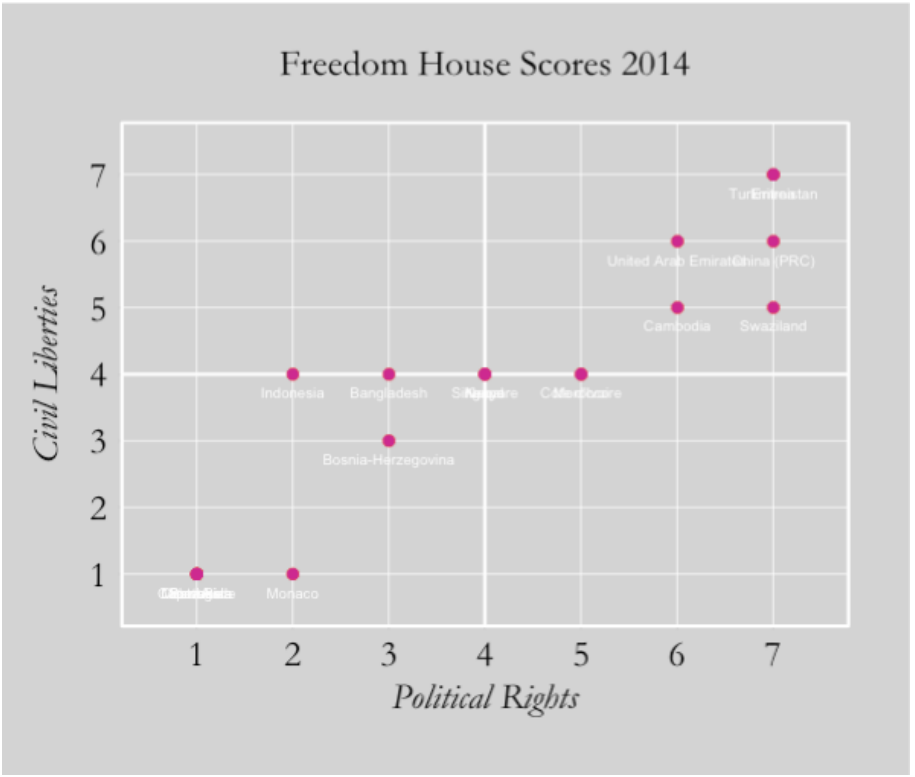
Low-level commands: add elements to a plot (points, lines, text, etc.)

Arguments and parameters: define and fine-tune various elements in a plot (color, size, etc.)

```
parameters(arguments)

highlevelcommand(data, arguments)

lowlevelcommand(data, arguments)

lowlevelcommand(data, arguments)

lowlevelcommand(data, arguments)
```

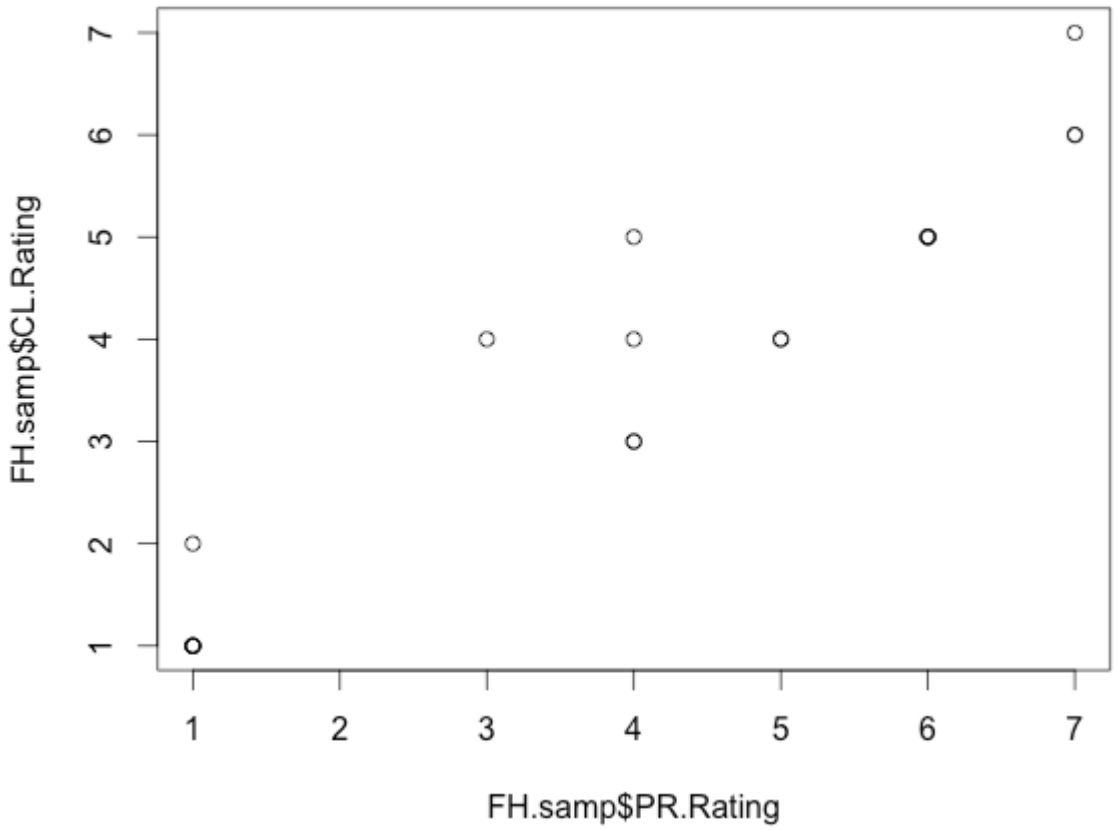
Before and After





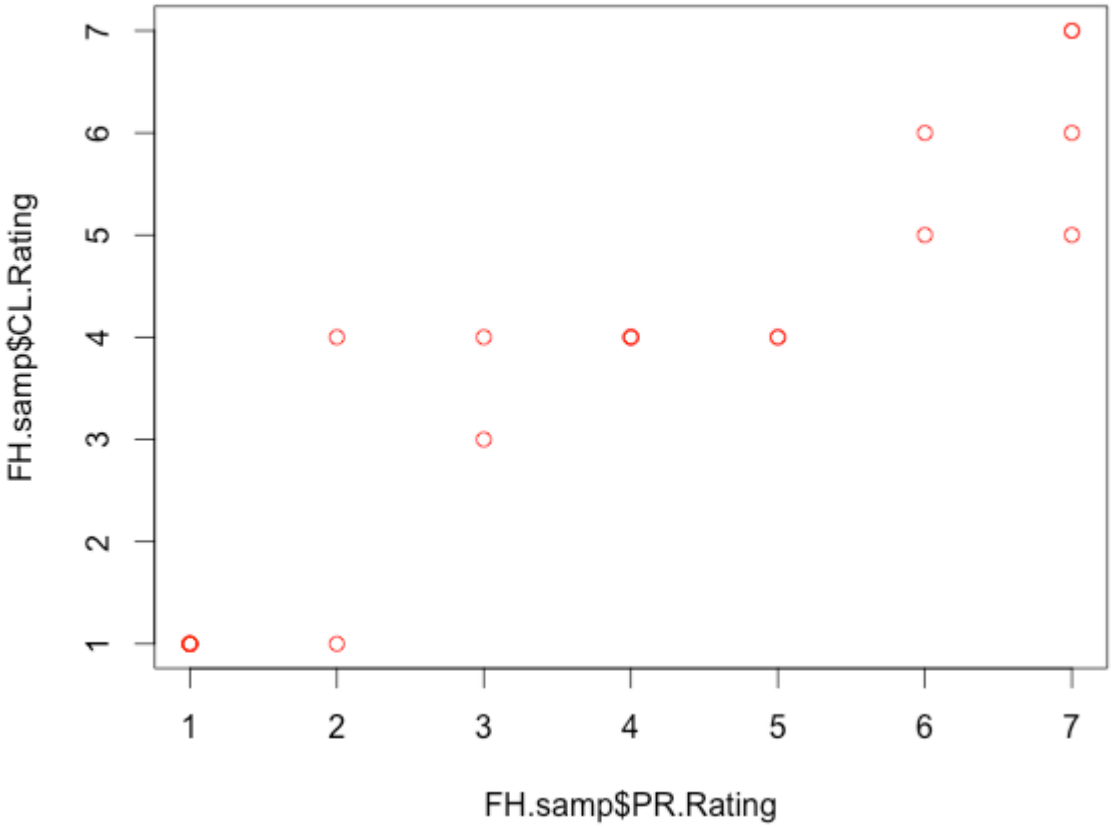
The `plot()` Function

```
plot(FH.samp$PR.Rating, FH.samp$CL.Rating)
```



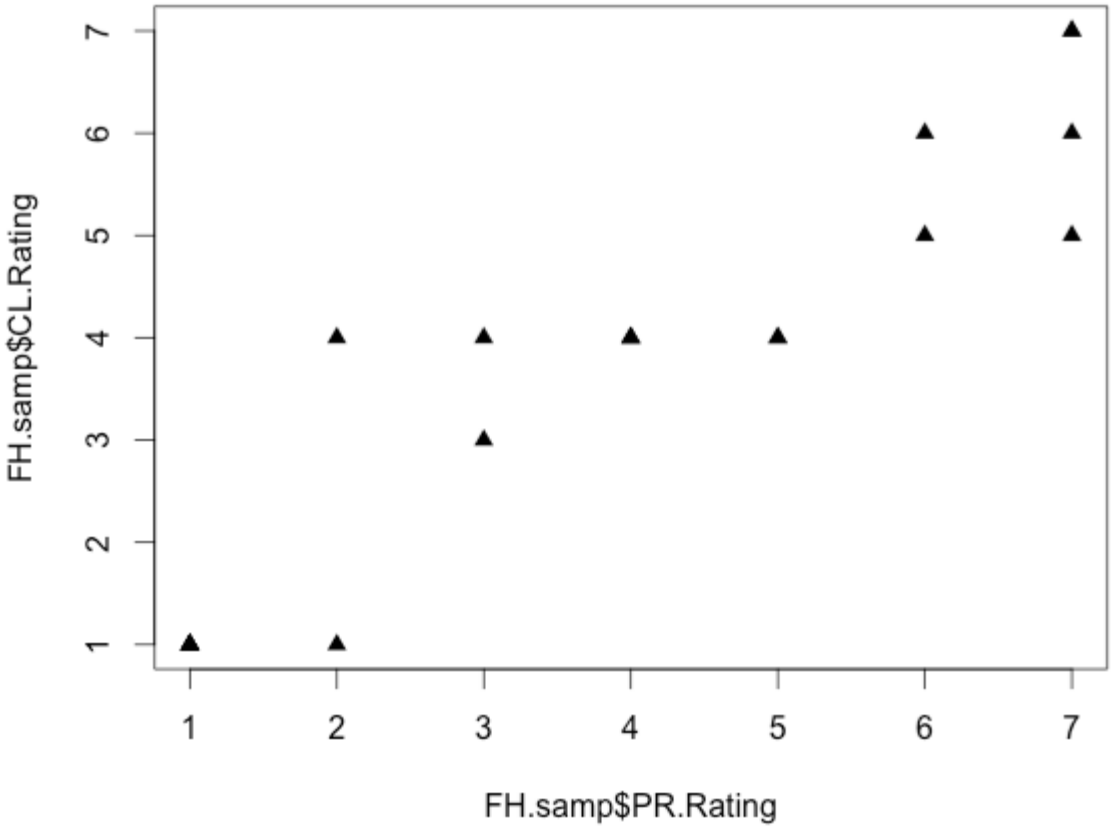
Some Arguments: Graphical Parameters

```
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, col="red")
```



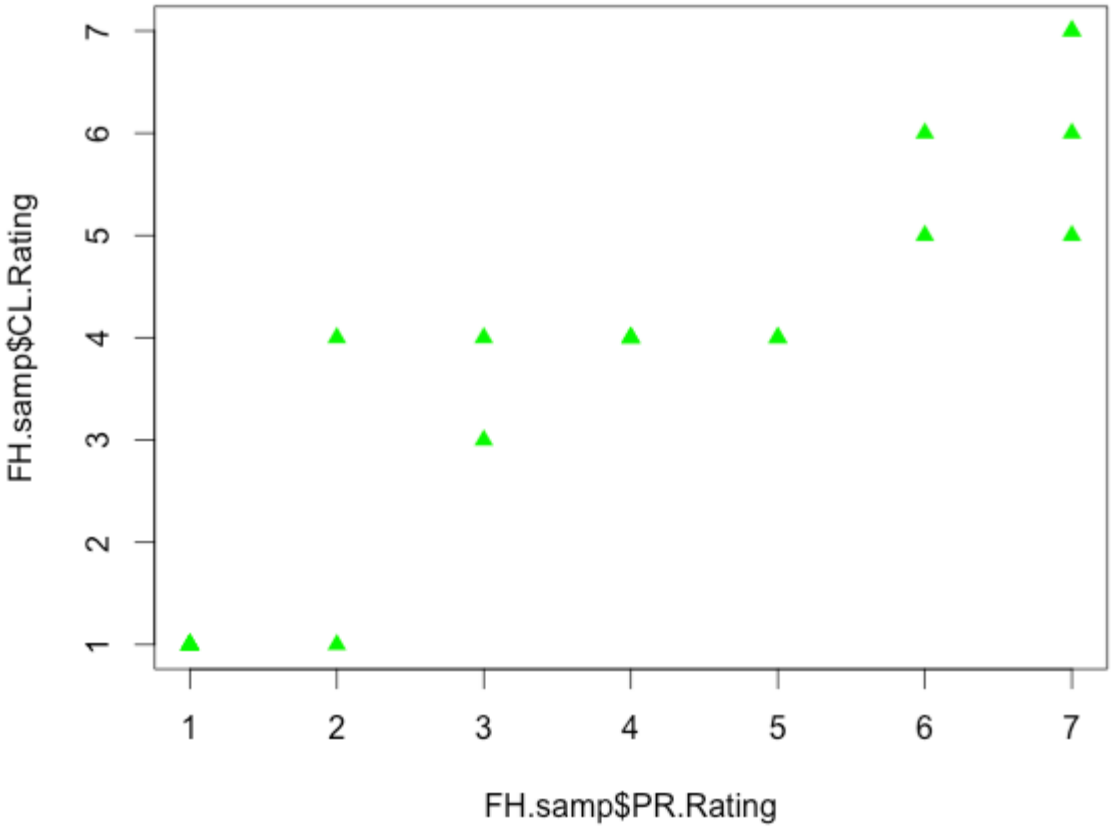
Some Arguments: Graphical Parameters

```
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, pch=17)
```

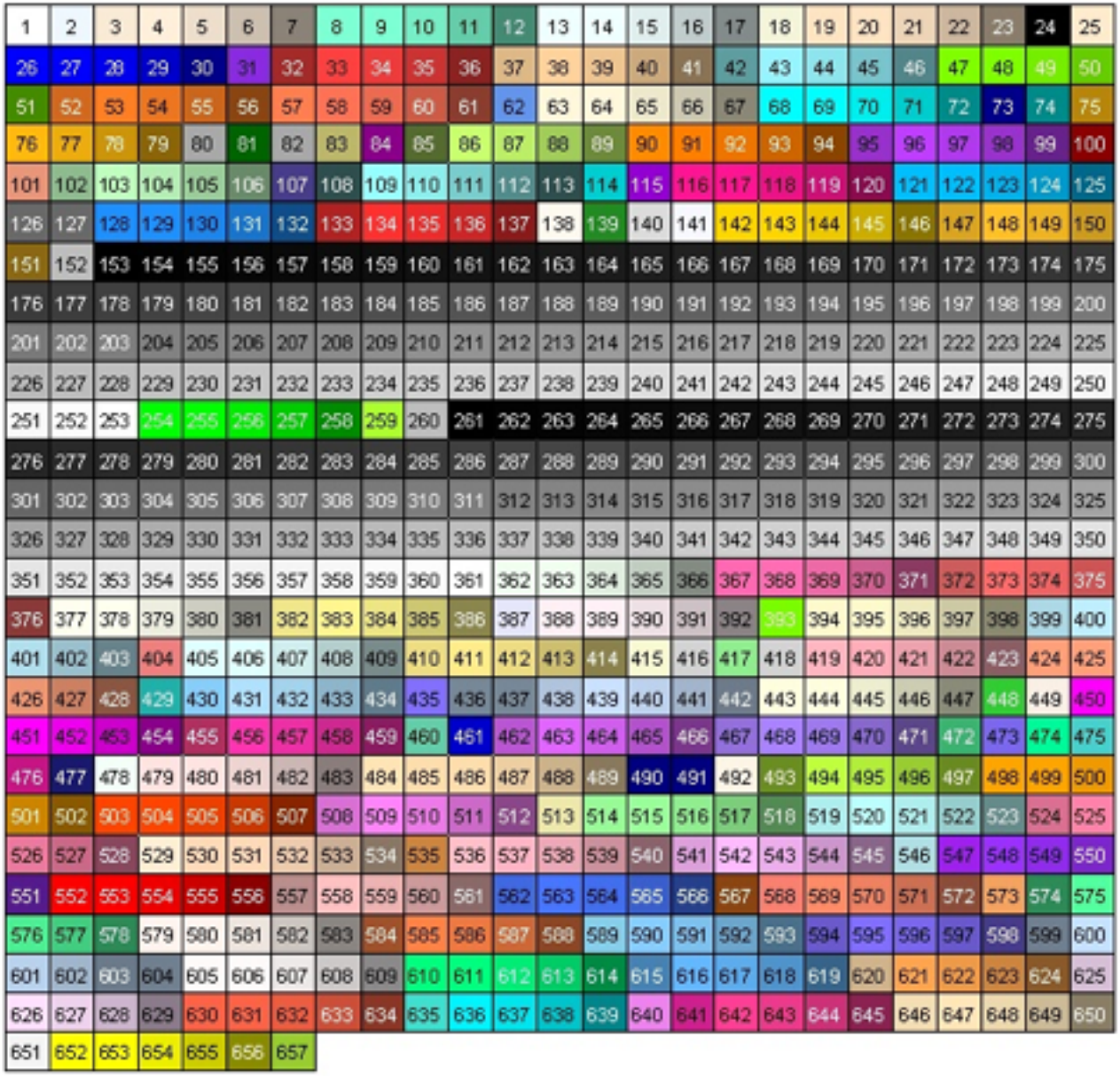


Some Arguments: Graphical Parameters

```
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, pch=17, col="green")
```



R colors

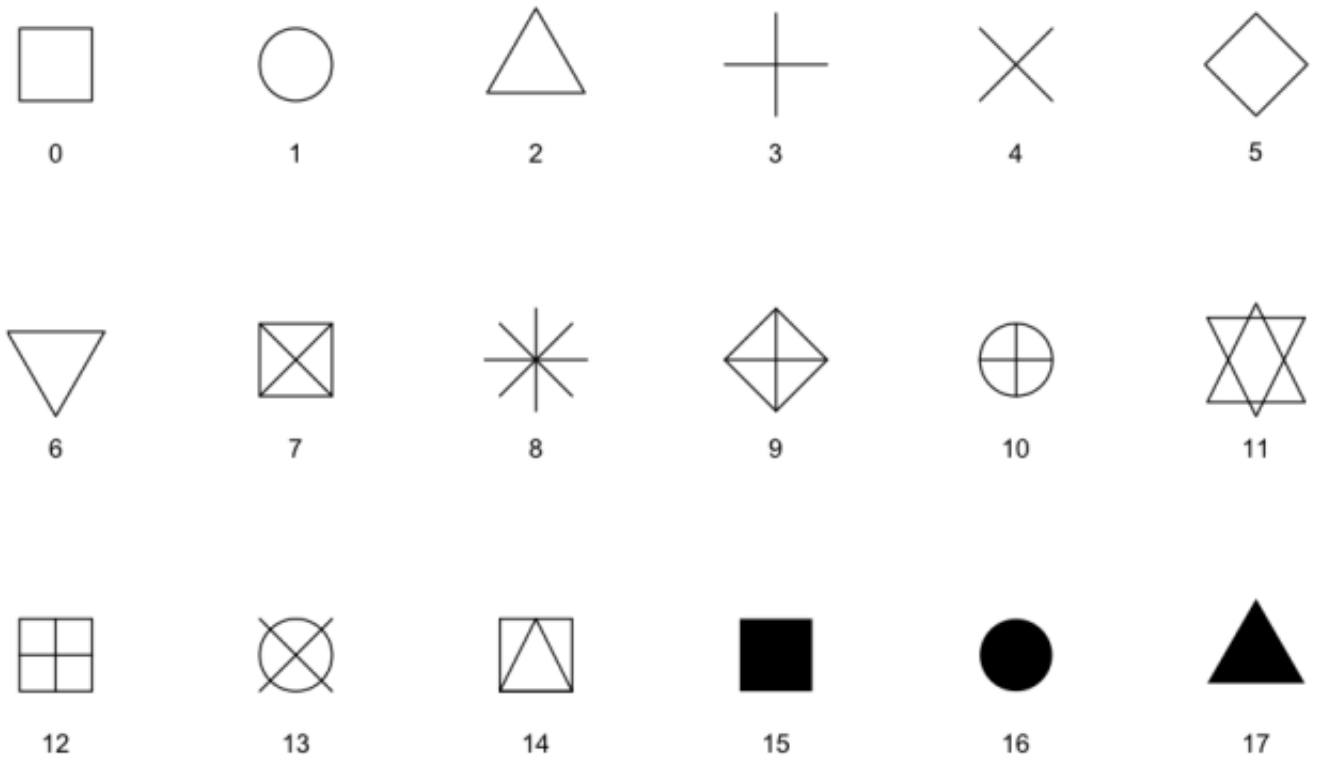

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125
126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150
151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175
176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200
201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225
226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250
251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275
276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300
301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325
326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350
351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375
376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400
401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425
426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450
451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475
476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500
501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525
526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550
551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575
576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600
601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625
626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650
651 652 653 654 655 656 657

How to find your favourite color

```
colors()[c(1, 280, 637)]
```

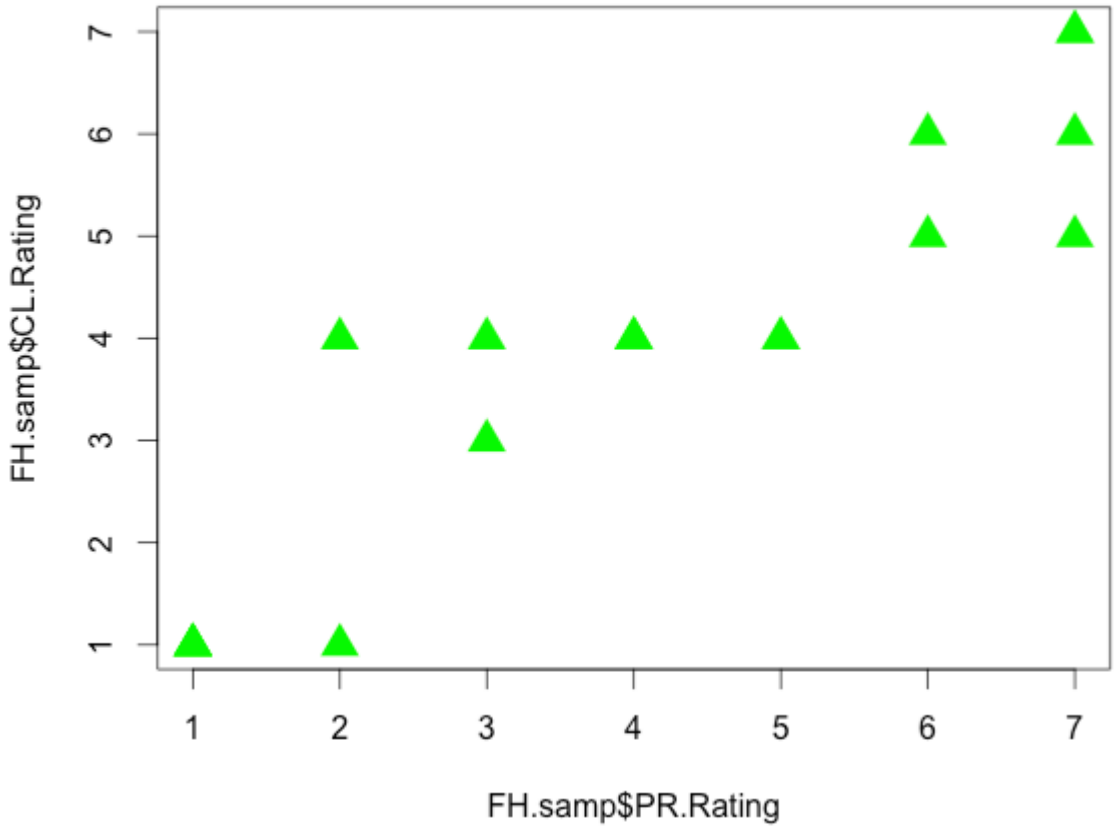
```
[1] "white" "grey19" "turquoise2"
```

Plotting Characters



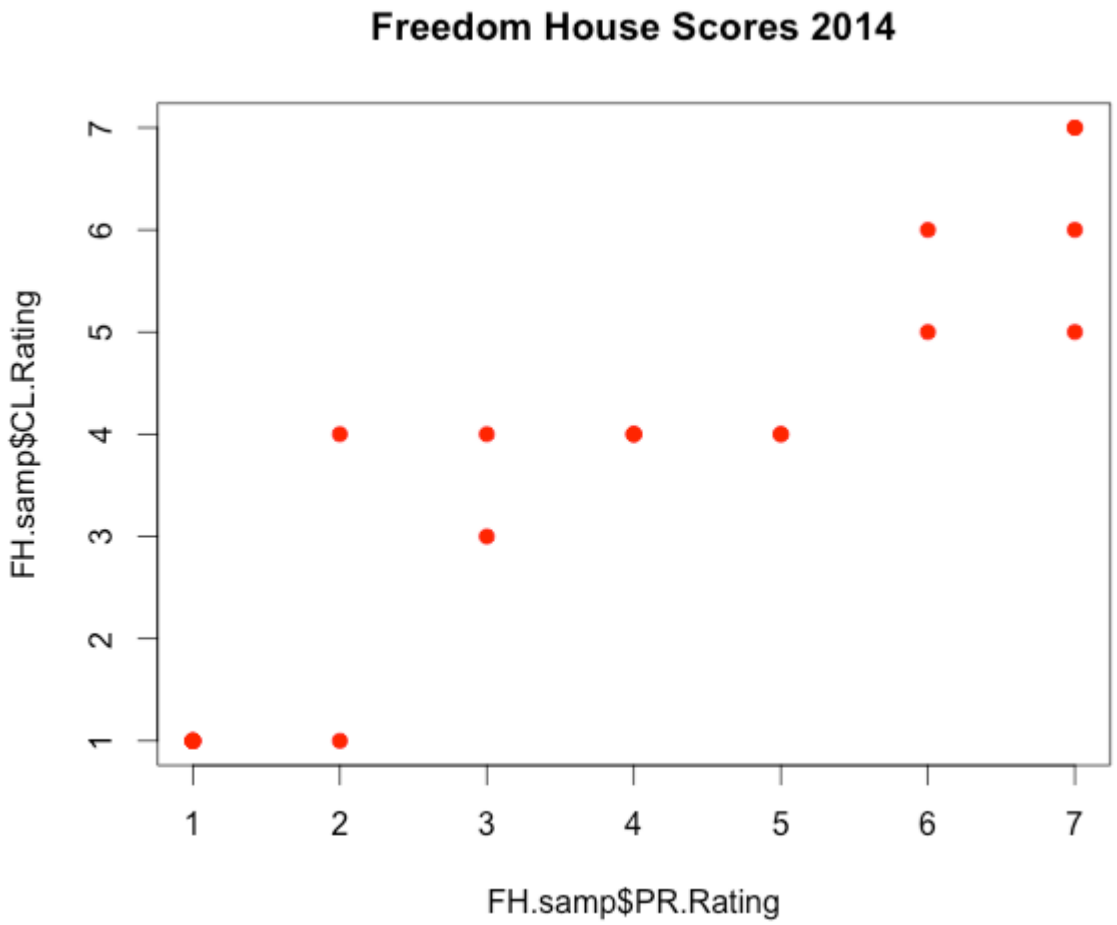
Some Arguments: Graphical Parameters

```
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, pch=17, col="green",
cex=2)
```



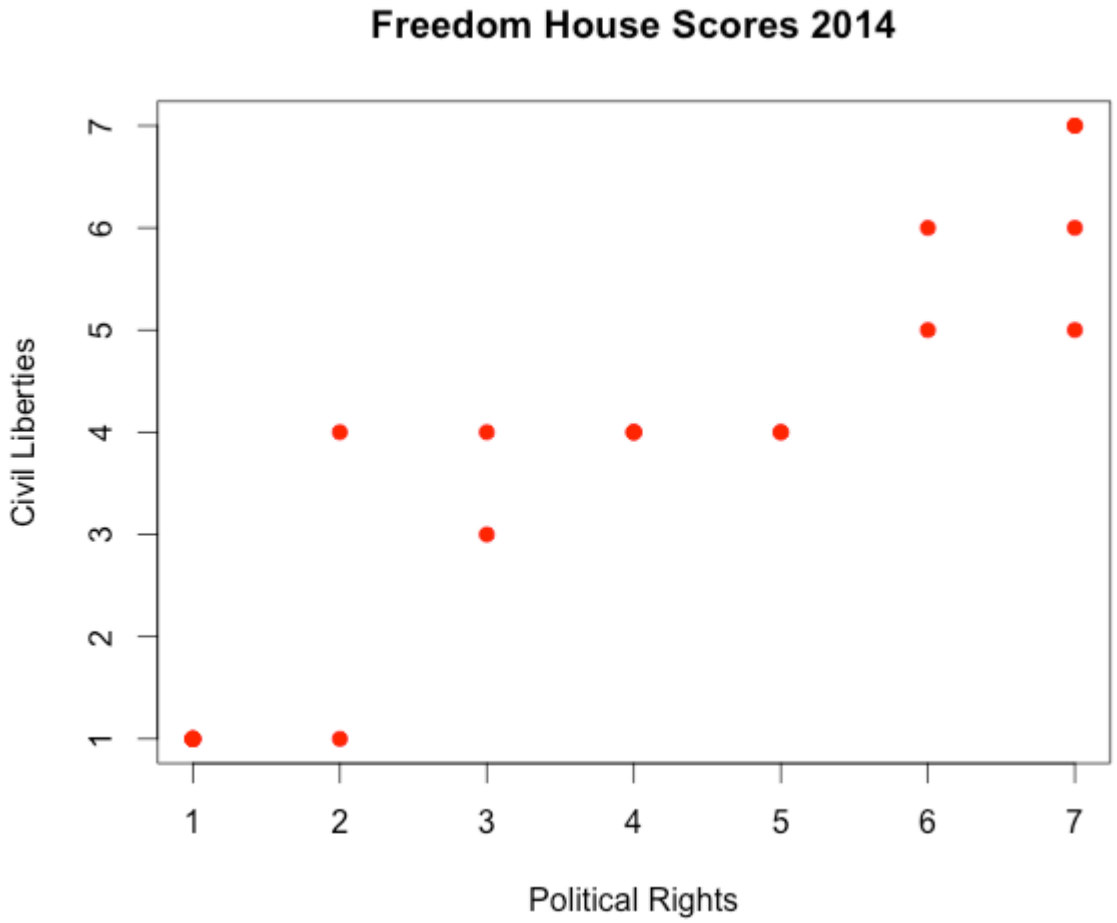
Some Arguments: Labels, Text and Axes

```
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, pch=19, col="red",
main="Freedom House Scores 2014")
```



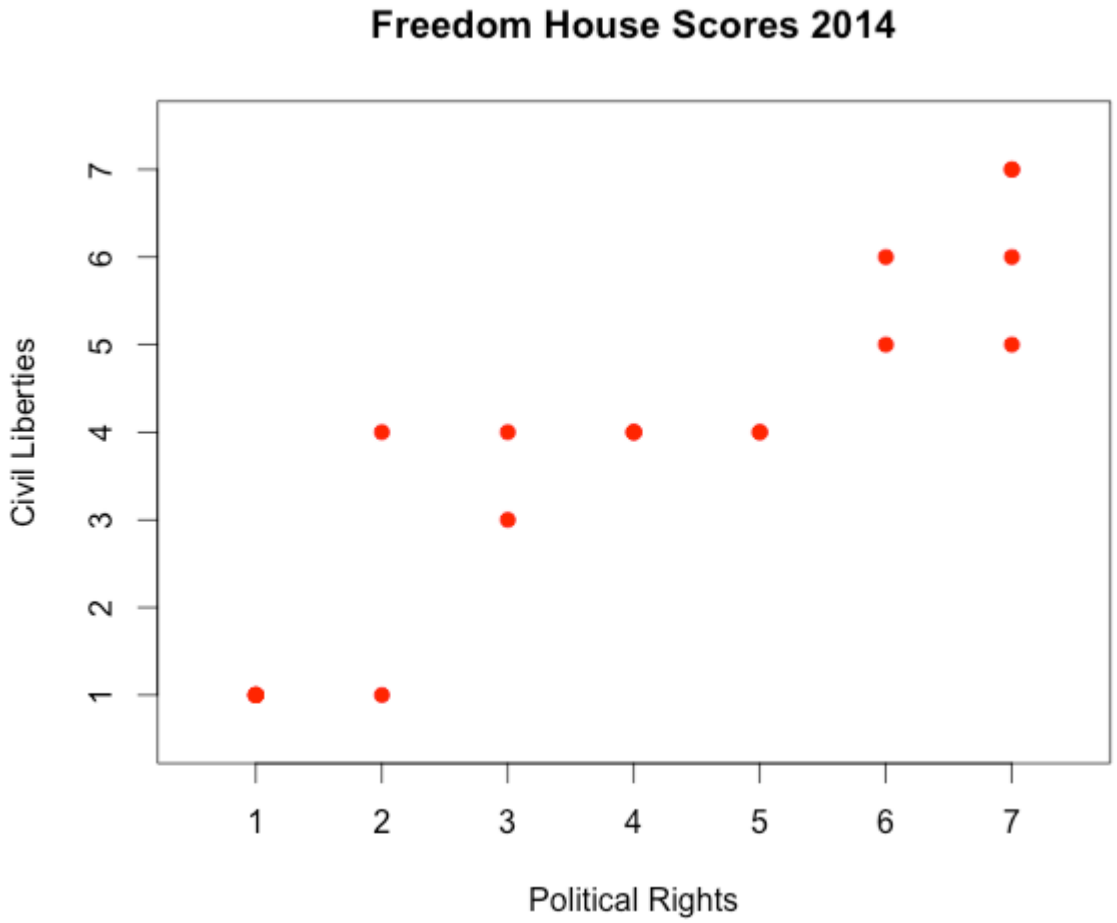
Some Arguments: Labels, Text and Axes

```
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, pch=19, col="red",
main="Freedom House Scores 2014", xlab="Political Rights",
ylab="Civil Liberties")
```



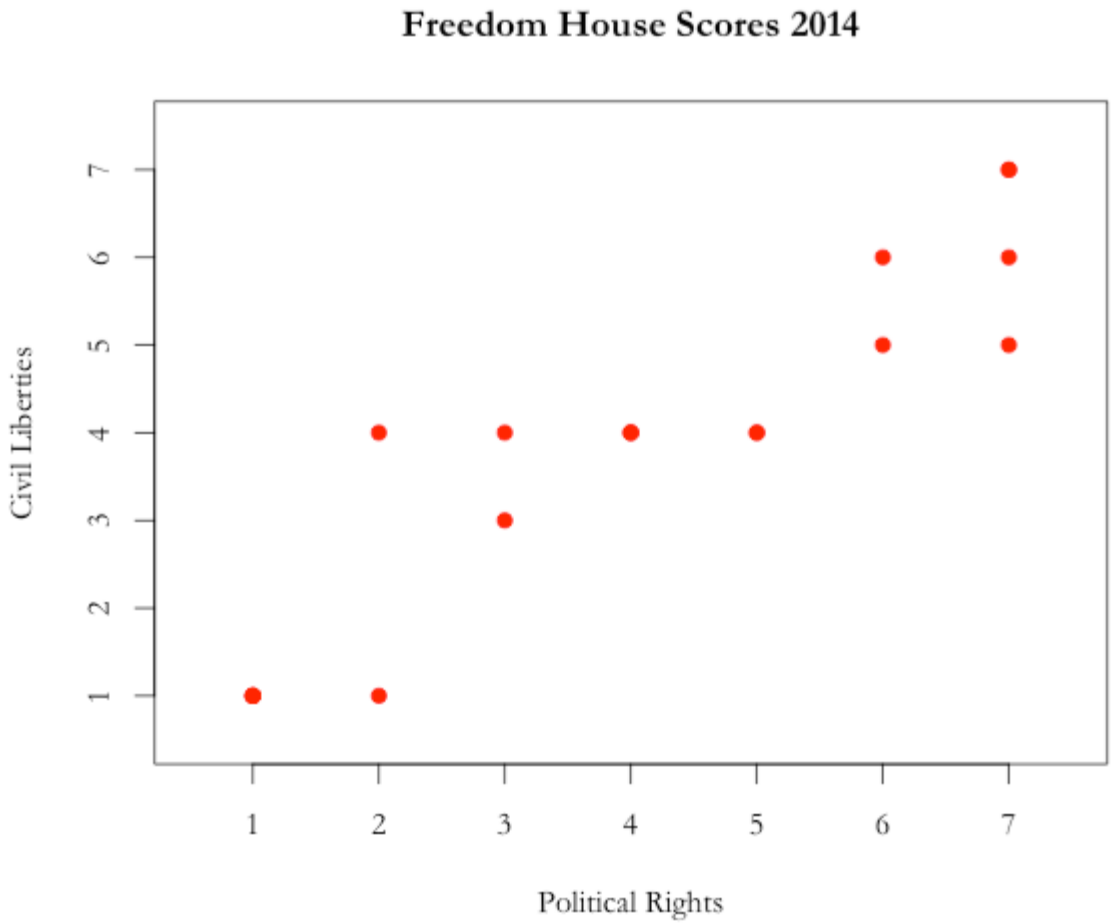
Some Arguments: Labels, Text and Axes

```
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, pch=19, col="red",
main="Freedom House Scores 2014", xlab="Political Rights",
ylab="Civil Liberties", ylim=c(.5, 7.5), xlim=c(.5, 7.5))
```



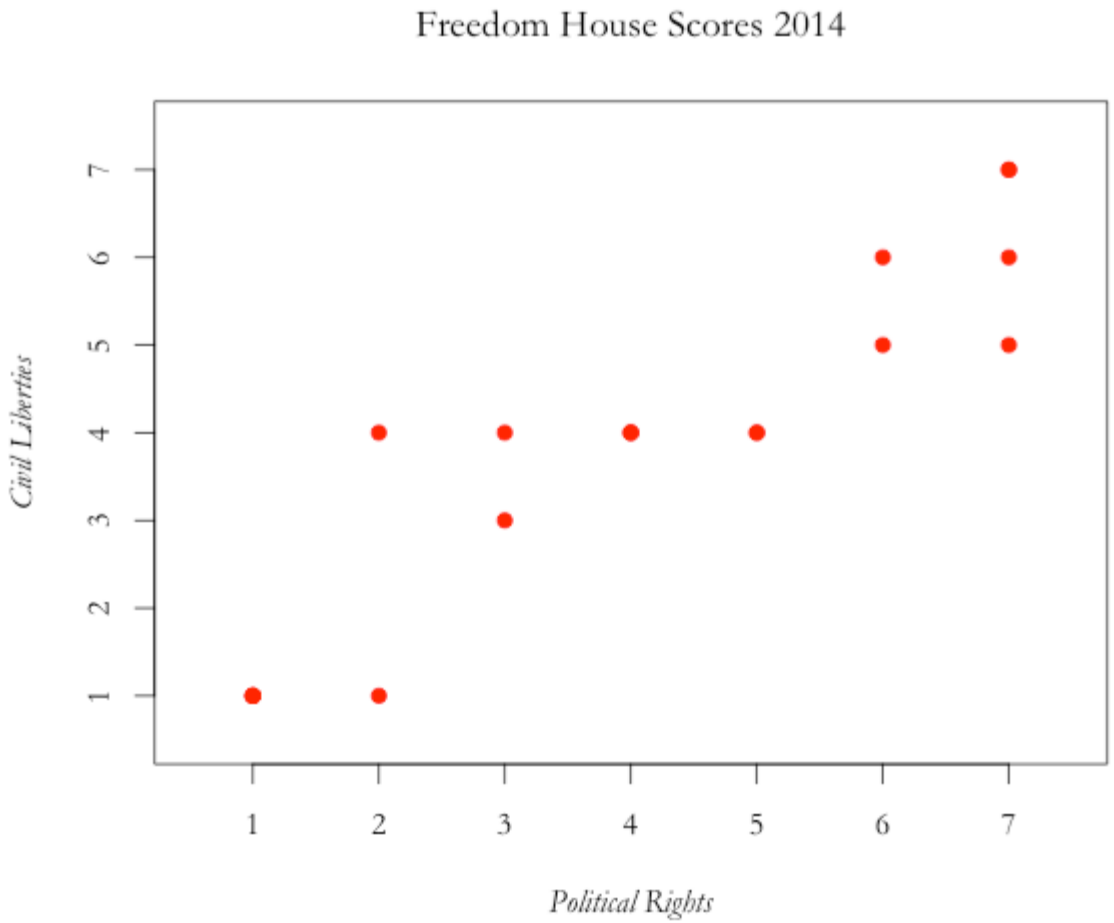
Some Arguments: Labels, Text and Axes

```
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, pch=19, col="red",
main="Freedom House Scores 2014", xlab="Political Rights",
ylab="Civil Liberties", ylim=c(.5, 7.5), xlim=c(.5, 7.5),
family="Garamond")
```

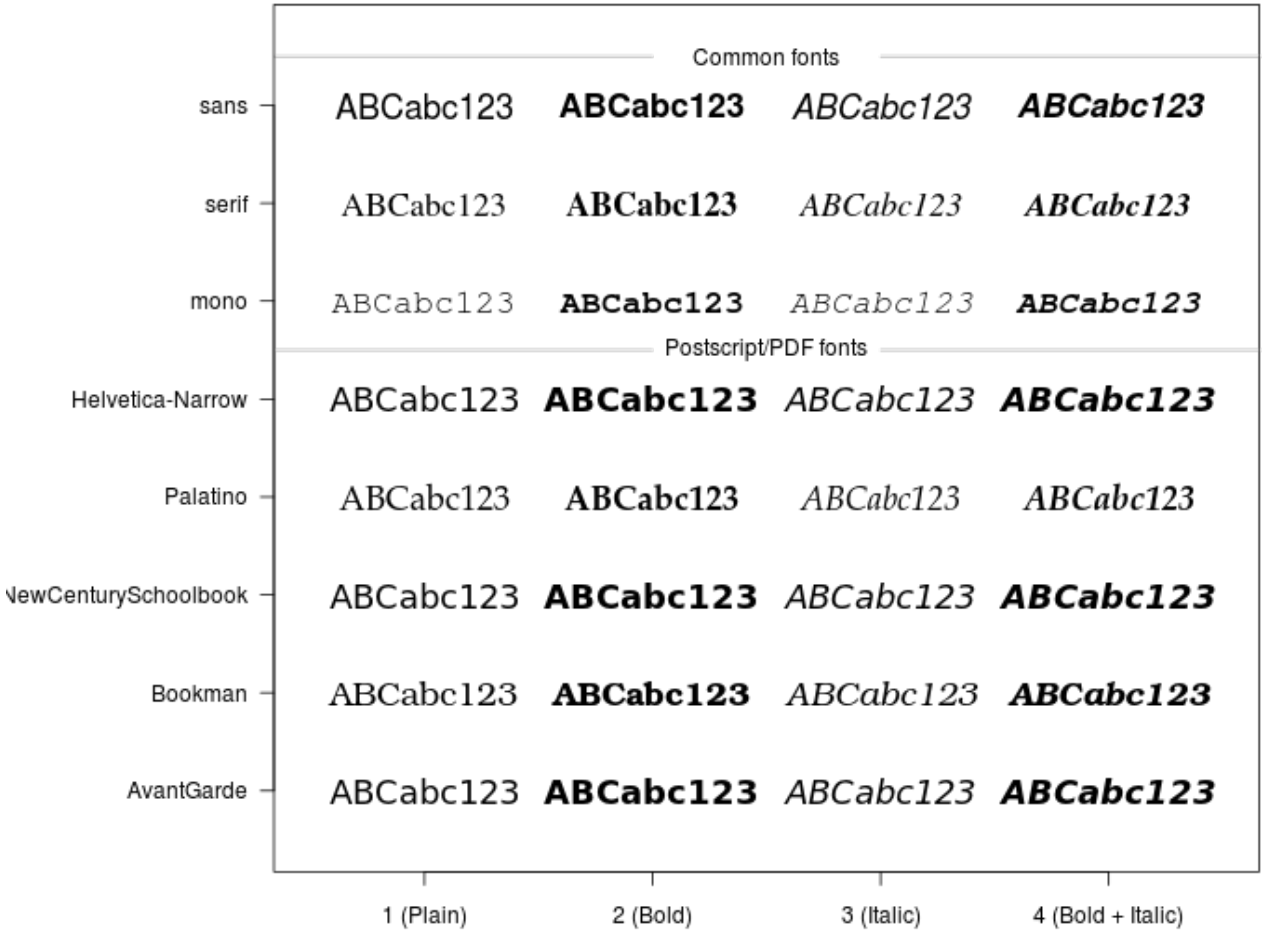


Some Arguments: Labels, Text and Axes

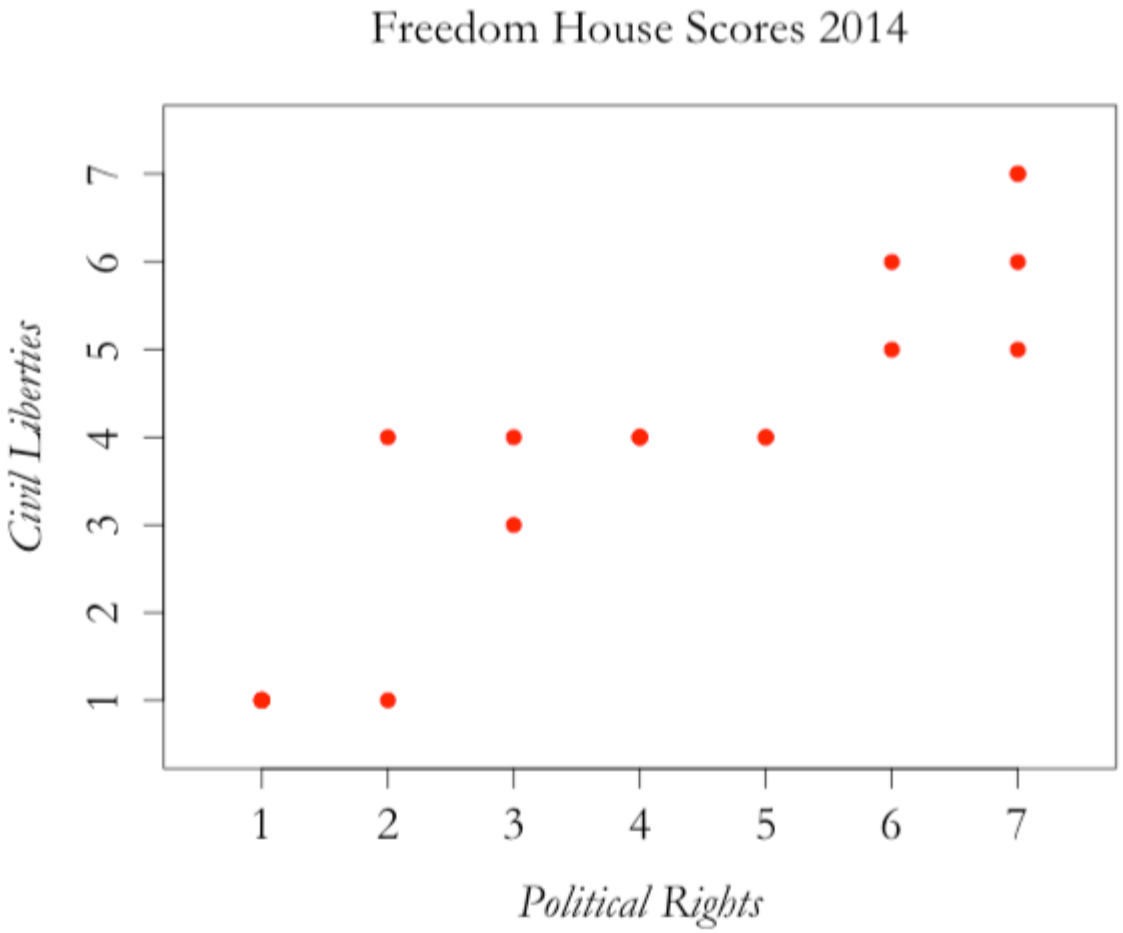
```
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, pch=19, col="red",
main="Freedom House Scores 2014", xlab="Political Rights",
ylab="Civil Liberties", ylim=c(.5, 7.5), xlim=c(.5, 7.5),
family="Garamond", font.main=1, font.lab=3)
```



Font Types and Families



```
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, pch=19, col="red",
main="Freedom House Scores 2014", xlab="Political Rights",
ylab="Civil Liberties", ylim=c(.5, 7.5), xlim=c(.5, 7.5),
family="Garamond", font.main=1, font.lab=3, , cex.main=1.5,
cex.lab=1.5, cex.axis=1.5)
```



The `par()` Function

Many arguments can be applied directly to high level functions such as `plot()`

See help for all possible arguments and how to define them: `?plot()`

Arguments can also be specified using the `par()` function

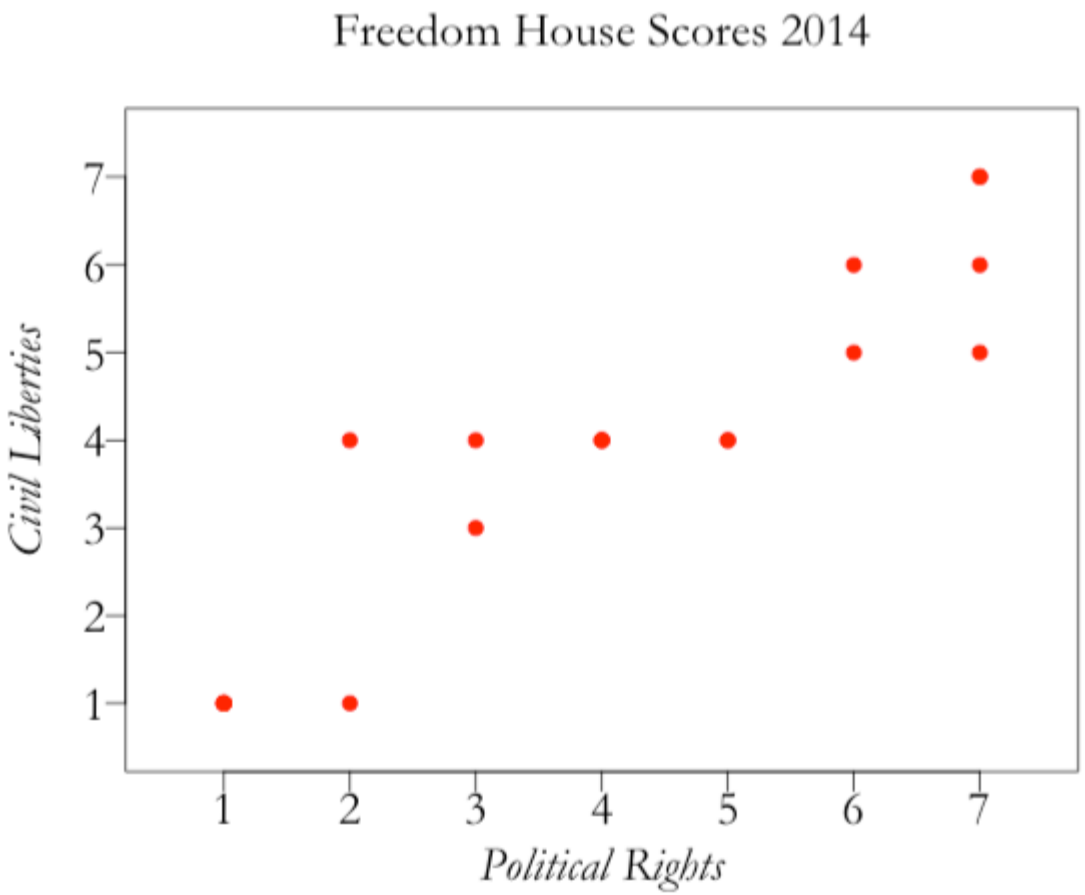
This sets parameters permanently for a session and applies them to all plots

Some more arguments with `par()`

```
par(mgp=c(2, .5, 0))
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, ...)
```

Some more arguments with `par()`

```
par(mgp=c(2, .5, 0), las=1)
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, ...)
```

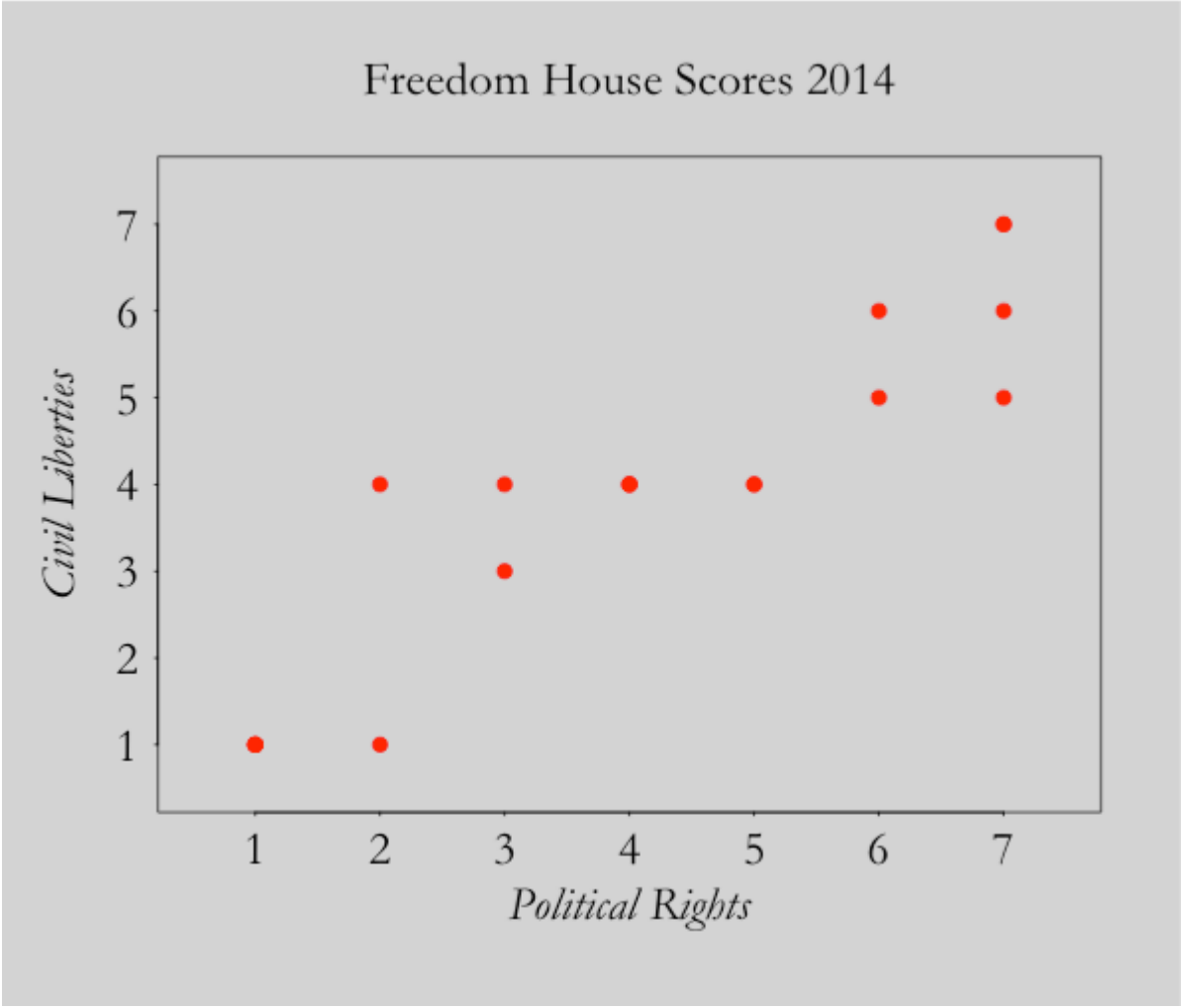


Some more arguments with `par()`

```
par(mgp=c(2, .5, 0), las=1, tick=-0.005)
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, ...)
```

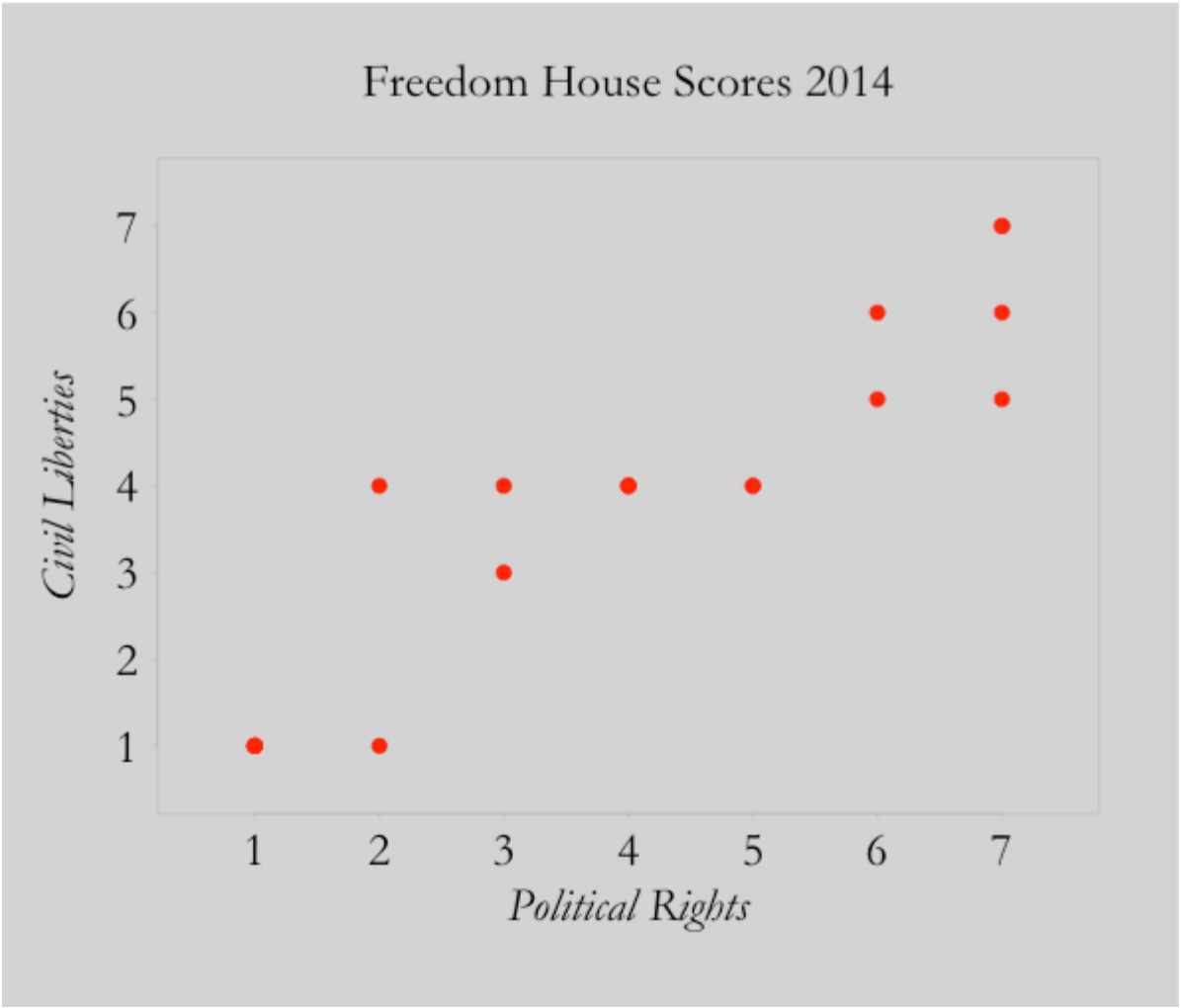
Some more arguments with `par()`

```
par(mgp=c(2, .5, 0), las=1, tck=-0.005, bg="lightgrey")
plot(FH.samp$PR.Rating, FH.samp$CL.Rating, ...)
```



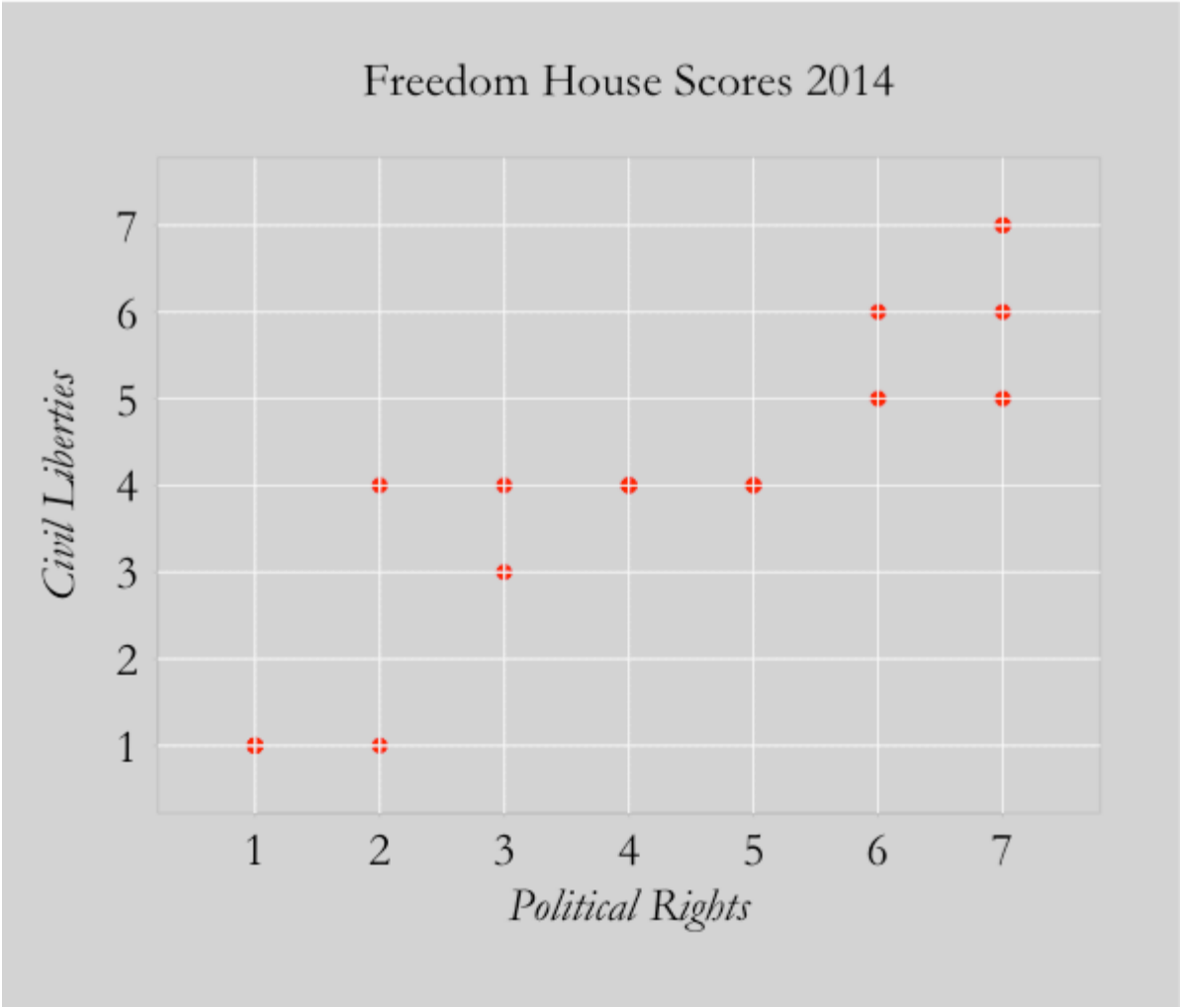
Some more arguments with `par()`

```
par(mgp=c(2, .5, 0), las=1, tck=-0.005, bg="lightgrey",
fg="grey")

plot(FH.samp$PR.Rating, FH.samp$CL.Rating, ...)
```



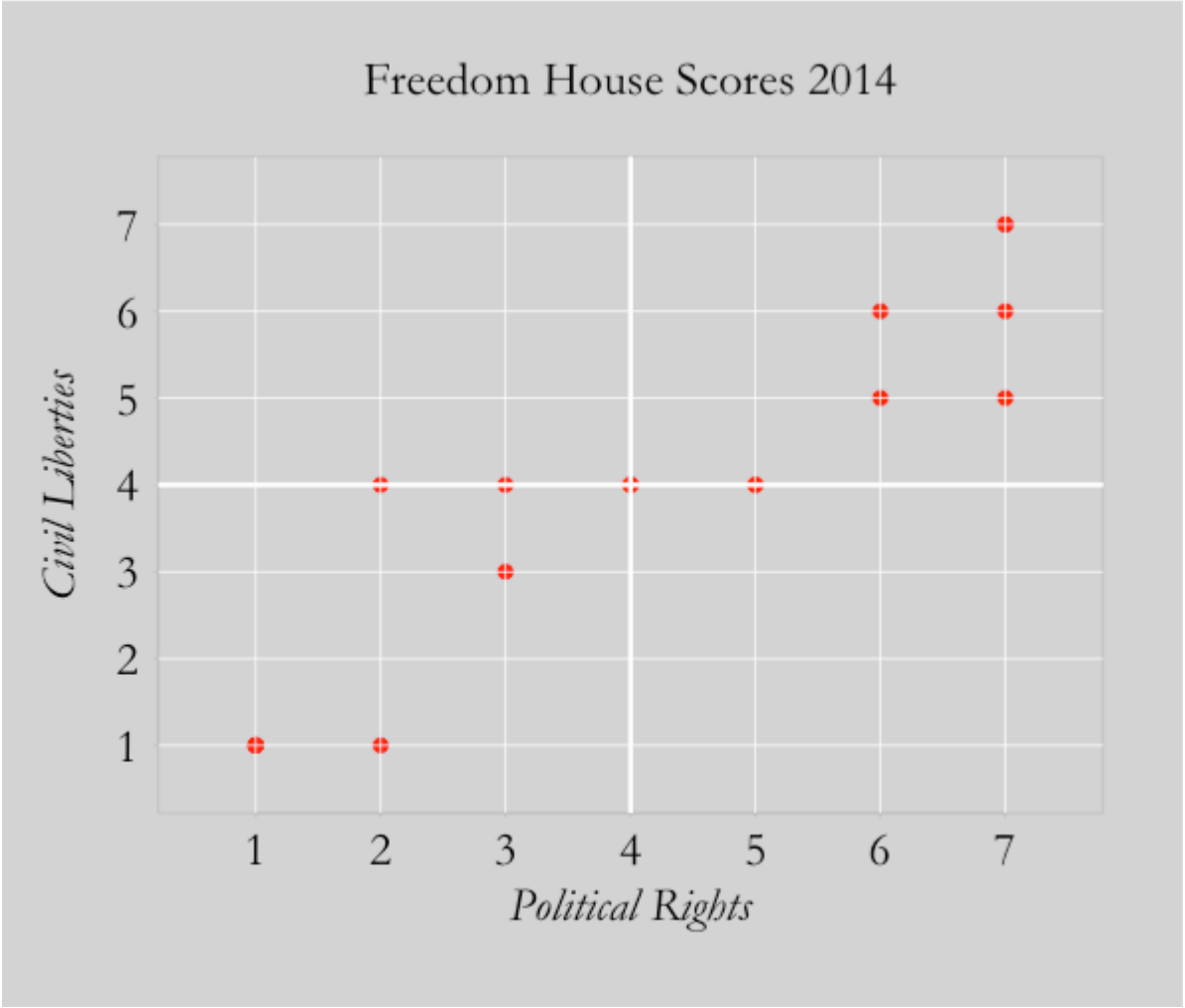
(Some) Low Level Functions

```
plot(...) grid(col="white", lty=1)
```



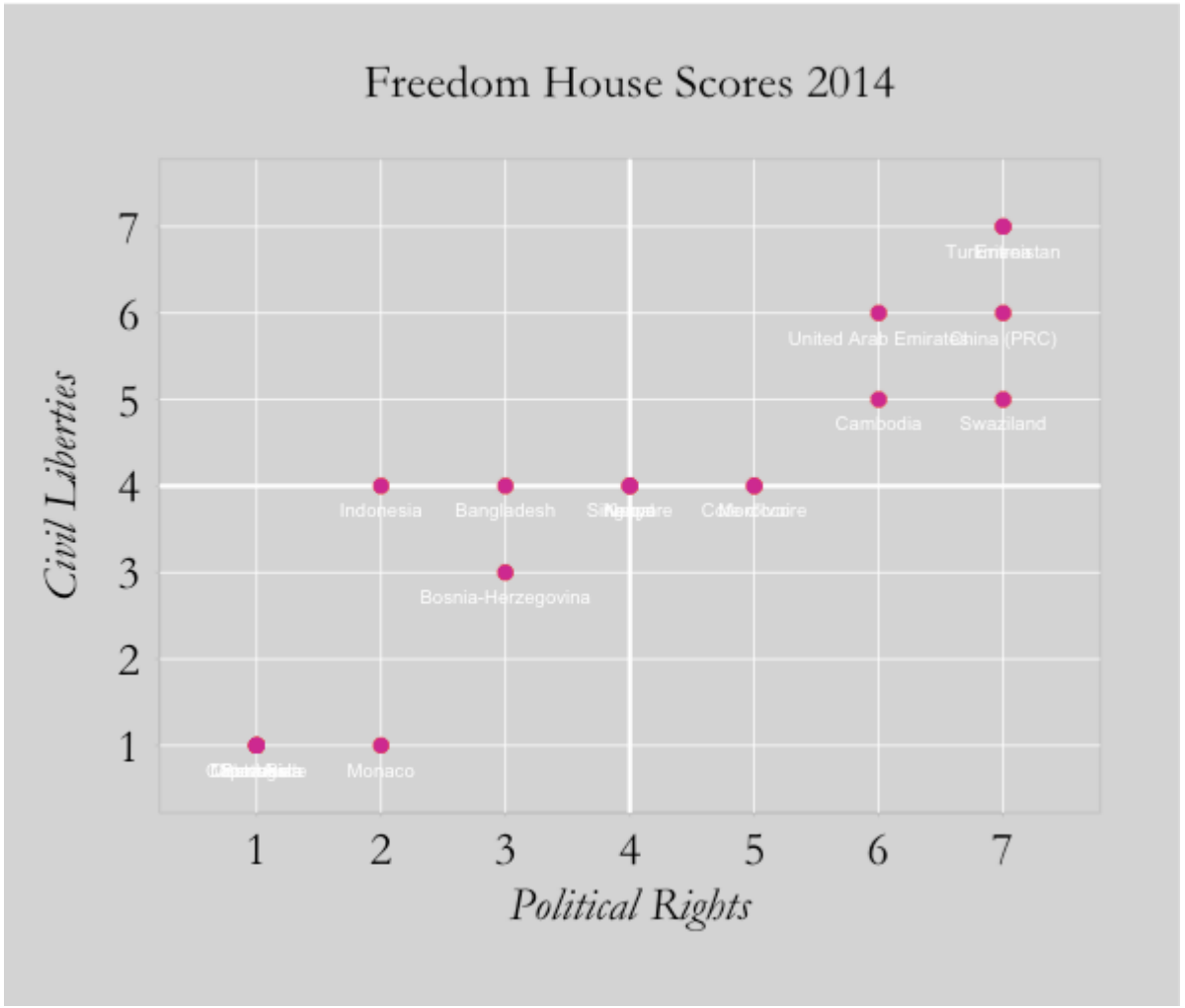
(Some) Low Level Functions

```
abline(v=4, col="white", lwd=3)
abline(h=4, col="white", lwd=3)
```



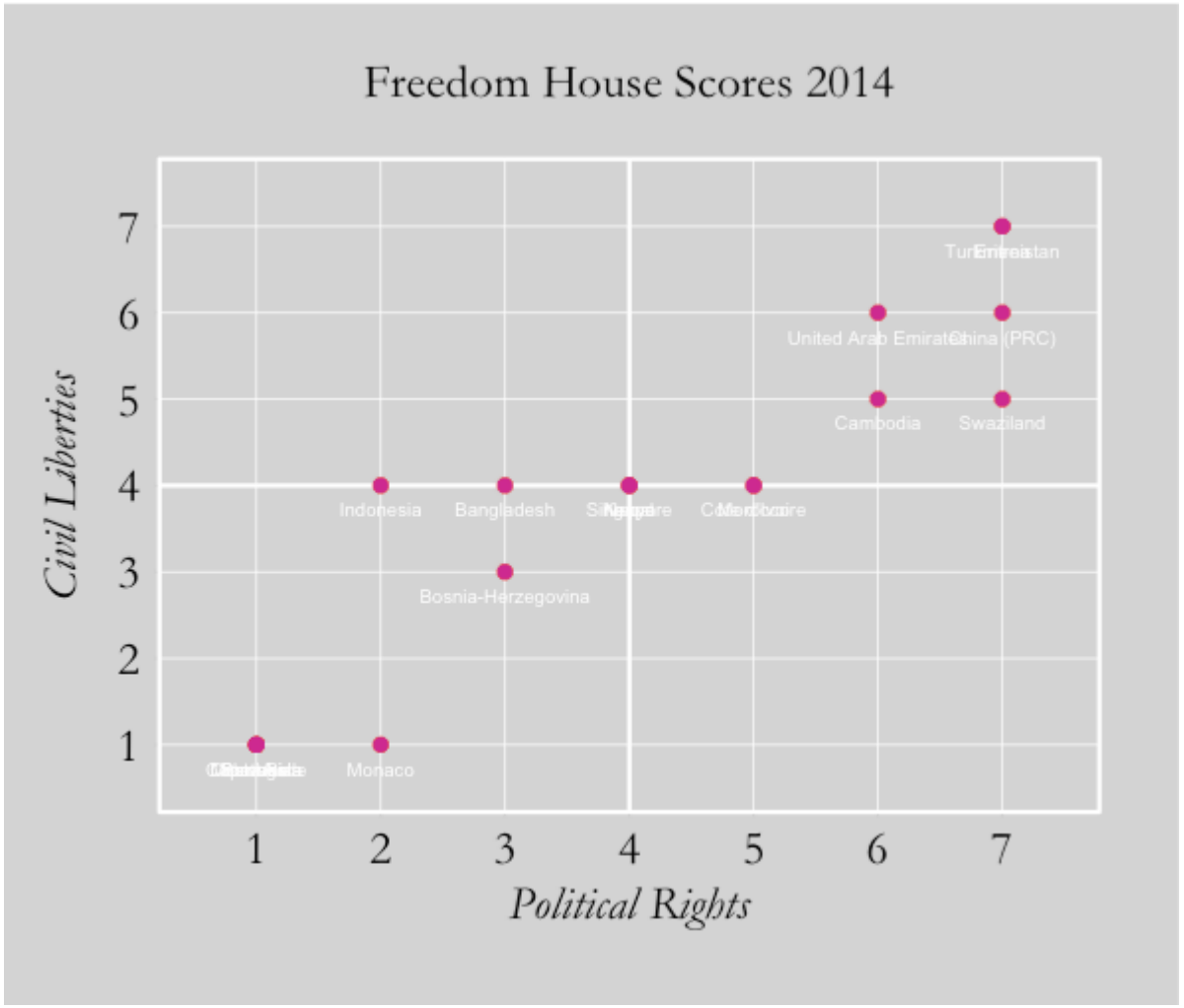
(Some) Low Level Functions

```
text(FH.samp$PR.Rating, FH.samp$CL.Rating, FH.samp$Country,
col="white", pos=1, cex=.6)
```



(Some) Low Level Functions

```
box(col="white", lwd=3)
```



Arranging Multiple Plots

```
par(mfrow=c(3,2))
```

Different Goals – Different Looks

Exploration Goals:

What's in the data?

Get a sense of size and complexity of data.

Explore and interact.

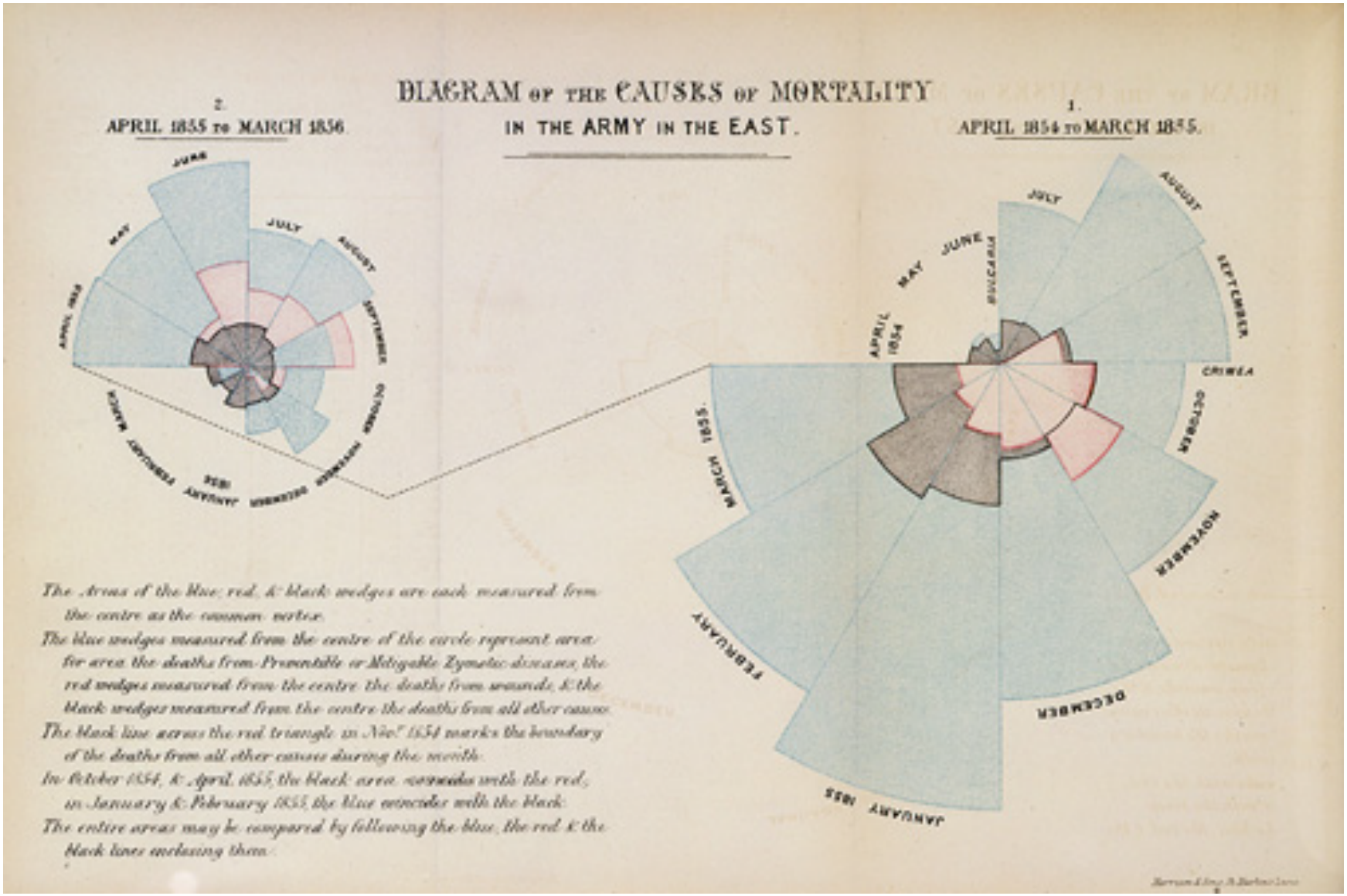
Communication Goals:

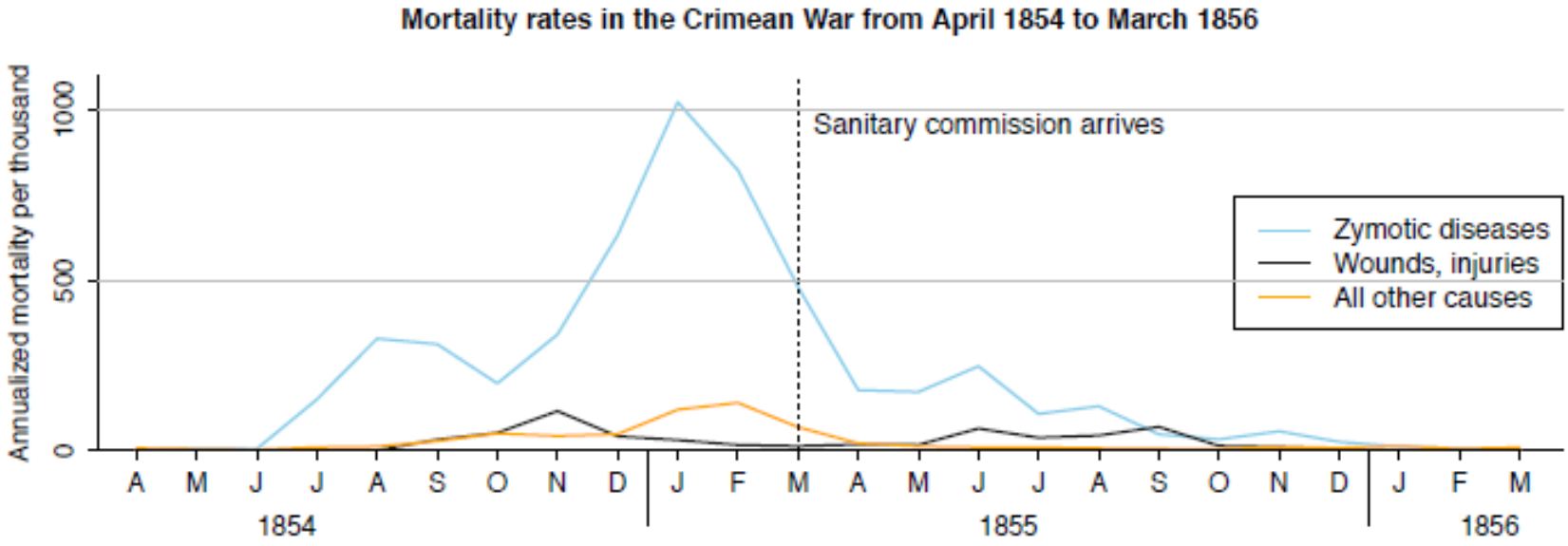
Communicate content of data.

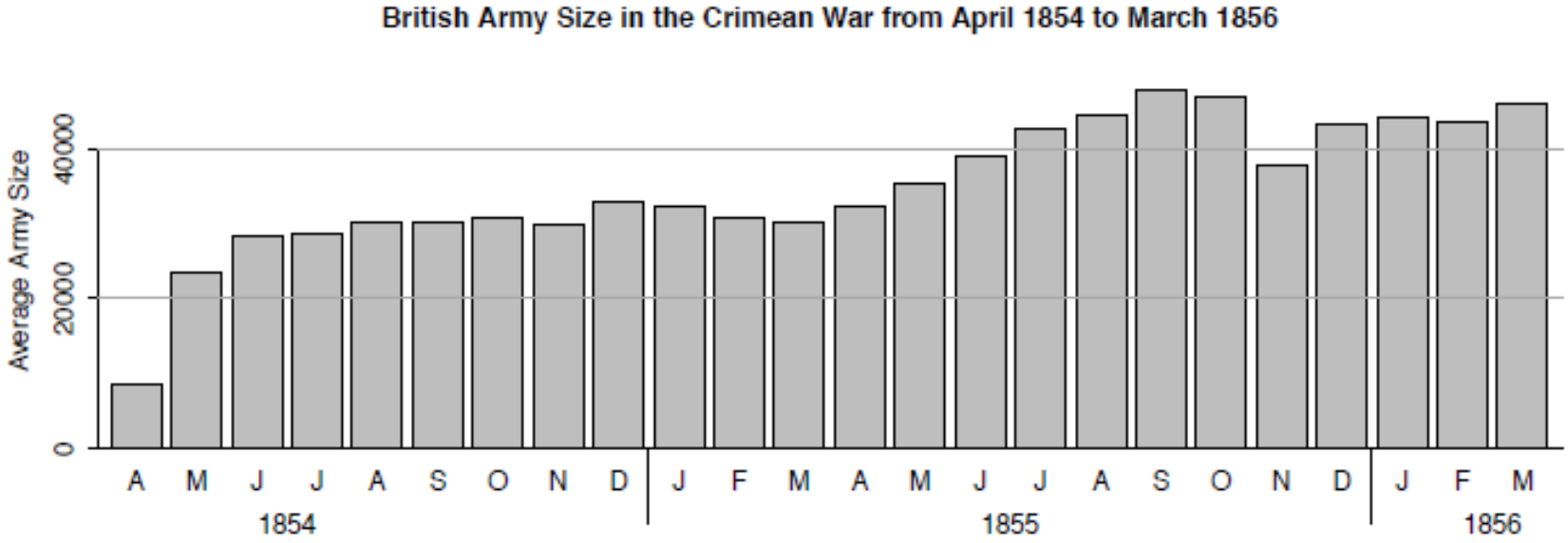
Tell a story with data.

Attract attention and interest.

Nightingale's Rose







Source: Gelman & Unwin 2012

Data Visualization as a Methodology

Data visualization is more than translating numbers into visual stimuli - it is a process where a problem is solved using data.

Visualization is the means by which the data is transformed into something that will help with the problem solving.

Focus is not so much on single graphical formats (although their proper choice remains essential!) but on how they are used in the larger context of a data analysis or presentation.

This almost immediately leads to questions of how to use and combine multiple graphs either of different subsets of the data or different formats of the same subset.

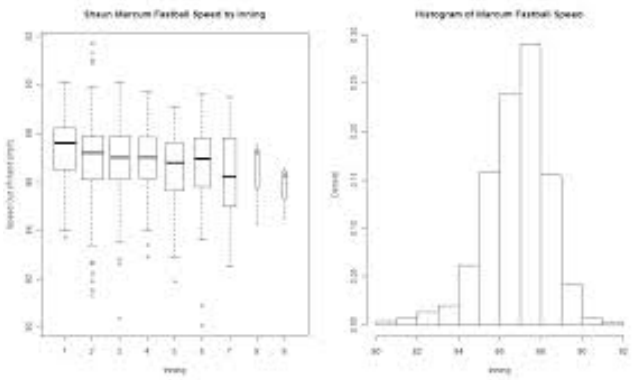
Put differently: single graphical formats are now just considered as building blocks in the more encompassing process that is data visualization.

Exploratory Visualization

„forces us to notice what we never expected to see“ (Tukey 1977: vi)

Mostly for ourselves in the course of the research process.

Many, quick and dirty, and rather unattractive graphs.



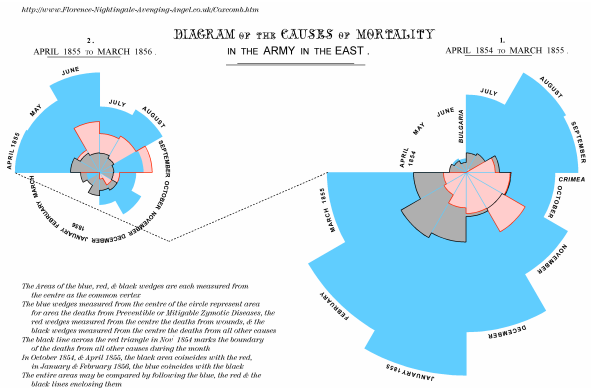
Analytical Graphs

Explanatory Visualization

„forces readers to see the information the designer wanted to convey“ (Kosslyn 1994: 271).

Mostly for others after the research is completed.

Few, carefully crafted, and attractive graphs.



Presentation Graphs

The Fundamental Principles of Analytic Design

1. Look at and Show: Comparisons, Contrasts, Differences

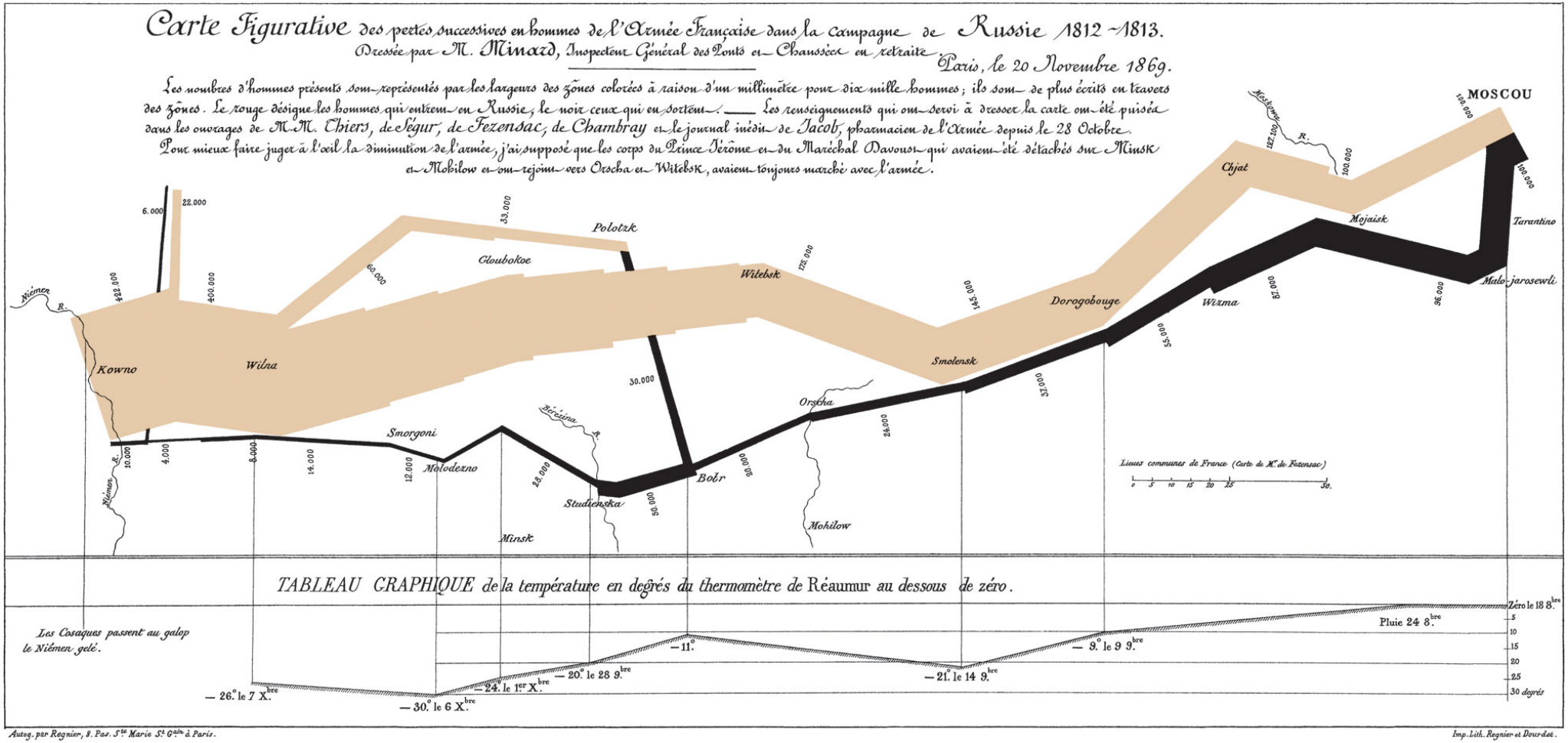
2. Look at and Show: Causality, Mechanism, Explanation, Systematic Structure

3. Look at and Show: Multivariate Data; that is, show more than 1 or 2 variables

4. Completely integrate words, numbers, images, diagrams

5. Thoroughly describe the evidence. Provide a detailed title, indicate the authors and sponsors, document the data sources, show complete measurement scales, point out relevant issues.

6. Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content. (What is the problem you want to solve?)


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
Niemen R.
Kowno
Wilna
Smorgoni
Minsk
Molodexno
Studienska
Polotzk
Gloubokoe
Witebsk
Bobr
Berezina R.
Orscha
Mohilow
Smolensk
Dorogobouge
Wixma
Chjat
Mojaisk
Moskwa R.
MOSCOU
Tarantino
Malo-jarosewli
Lieues communes de France (Carte de M. de Fezensac)
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26.° le 7 X.bre
— 30.° le 6 X.bre
— 24. le 1.er X.bre
— 20.° le 28 9.bre
— 11.°
— 21.° le 14 9.bre
— 9.° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. St.e Marie St. Gain à Paris.
Imp. Lith. Regnier et Dourdet.

High-level Goal: Exploratory Visualization

There is no correct way to visually explore data – but some general strategies are helpful.

Visual analysis can be divided into two approaches:

Directed: The analysis begins with a specific question and searches for a particular pattern to answer that question.

Exploratory: The analysis begins with simply looking at the data without any *a priori* idea – when we find something interesting, we ask a specific question and switch to directed analysis.

Of course, it is also possible to start with a directed analysis, find something unexpected and then switch to exploratory mode!

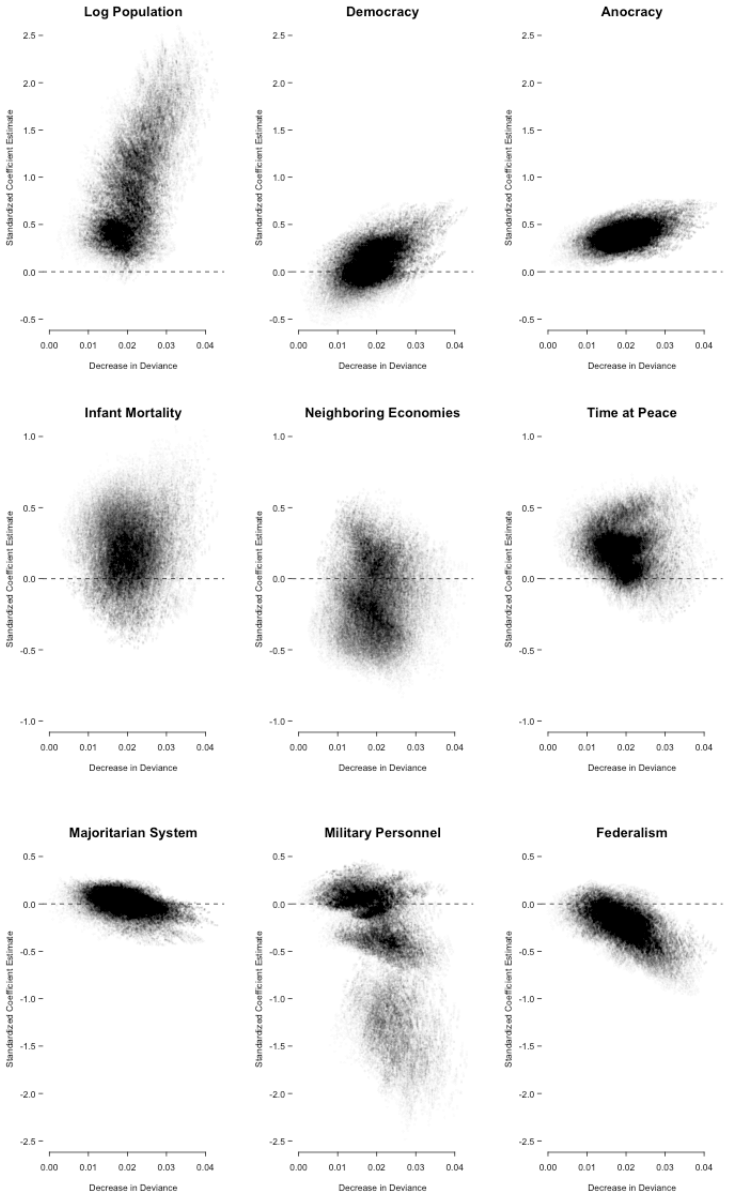
High-level Goal: Exploratory Visualization

How is “simply looking at data” done in practice?

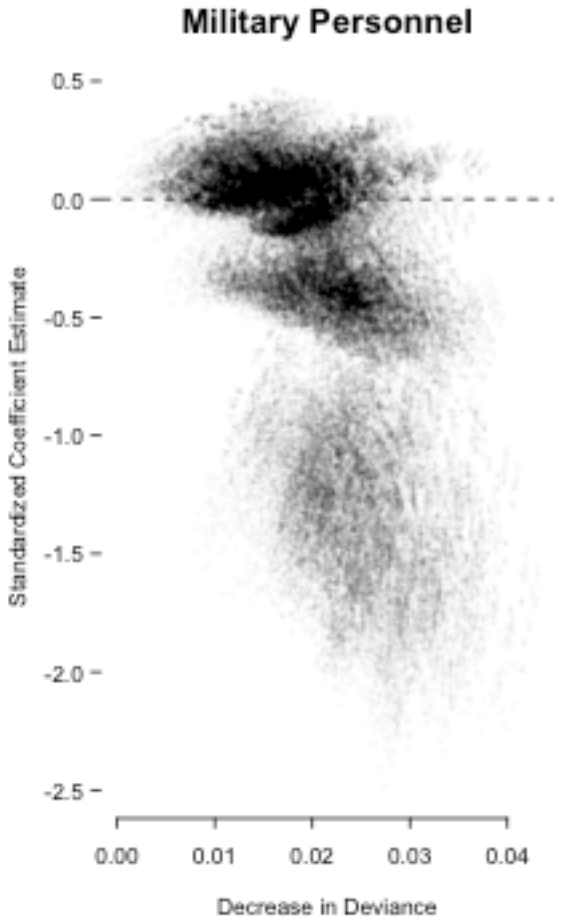
Shneiderman’s Information Seeking Mantra:

Overview first, zoom and filter, then details-on-demand.

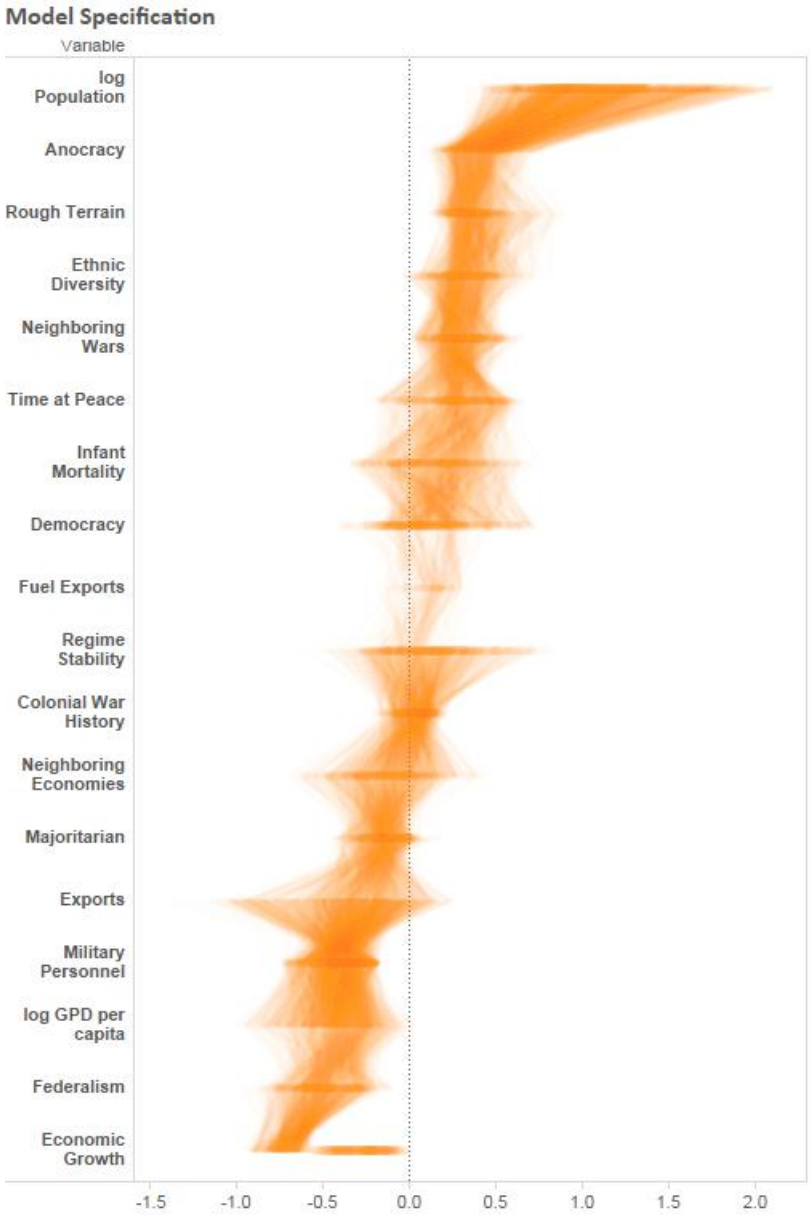
Overview first: Search for overall patterns and points of interest



Zoom and Filter: Zoom in on a point of interest, and filter out the rest



Details-on-demand: Bring in additional information as needed



And the exploration continues…

Low-level Goal: Making (Specific) Comparisons and Finding (Specific) Patterns

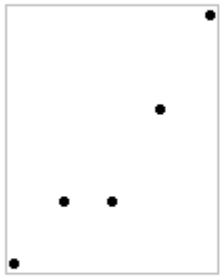
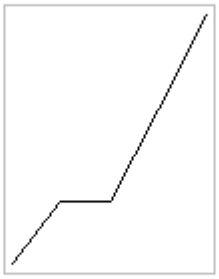
Graphs display relations and patterns in data by giving them a shape.

Graphical formats vary in terms of their usefulness in revealing specific patterns.

So having an idea about the comparisons you'd like to make and which patterns would like to find can guide your choice of a suitable graphical format.

Low-level Goal: Making (Specific) Comparisons and Finding (Specific) Patterns

Most relations can be displayed with some version of these very simple graphic formats

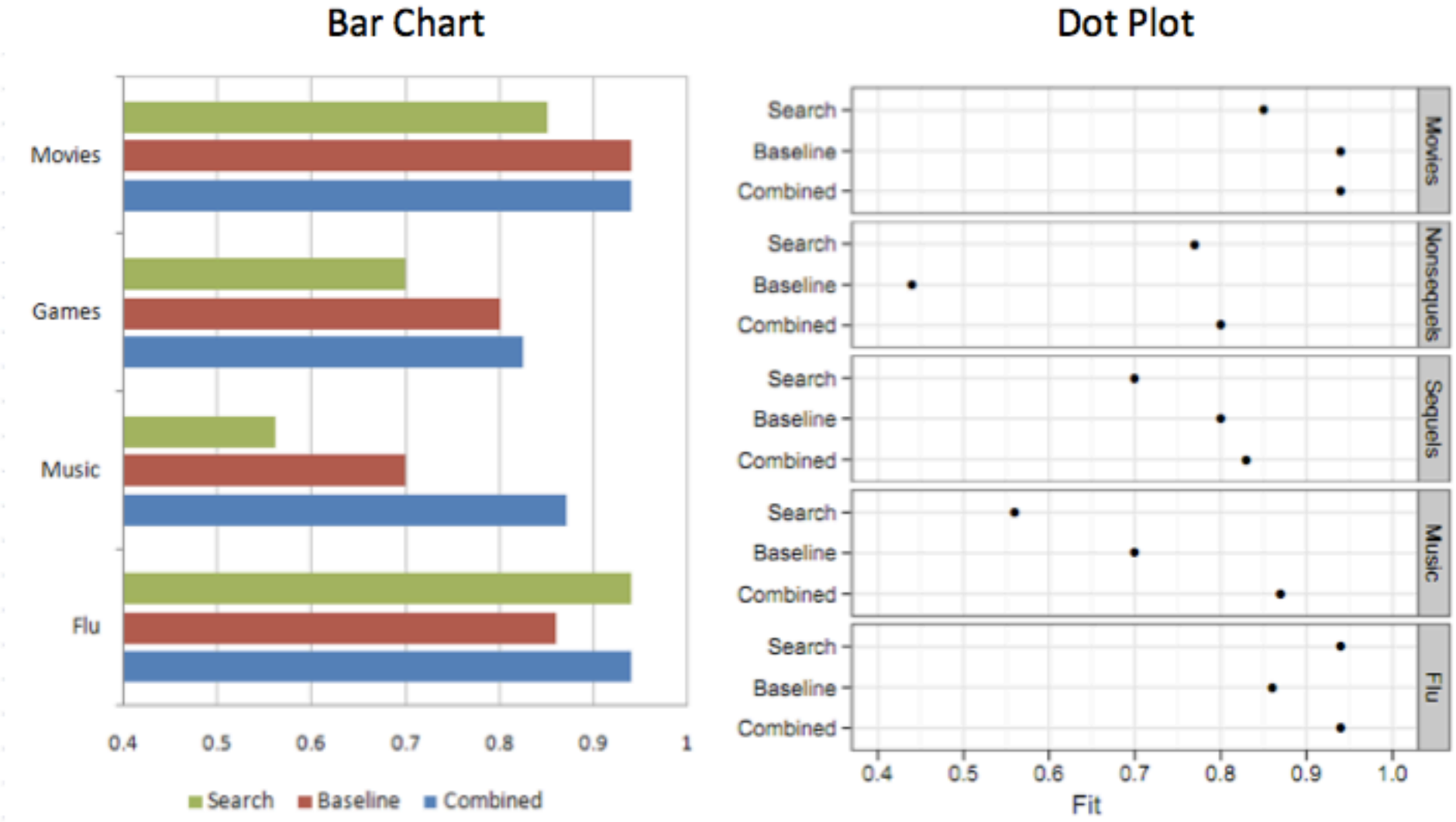




Don´t underestimate the power of common graphical formats - since most people are familiar with them, they are able to focus entirely on the content of the data.

„data analysts, both novices and experts, are, in general, incapable or unwilling to use the superb tools and techniques created for them“ (Andrienko & Andrienko 2006: 3).

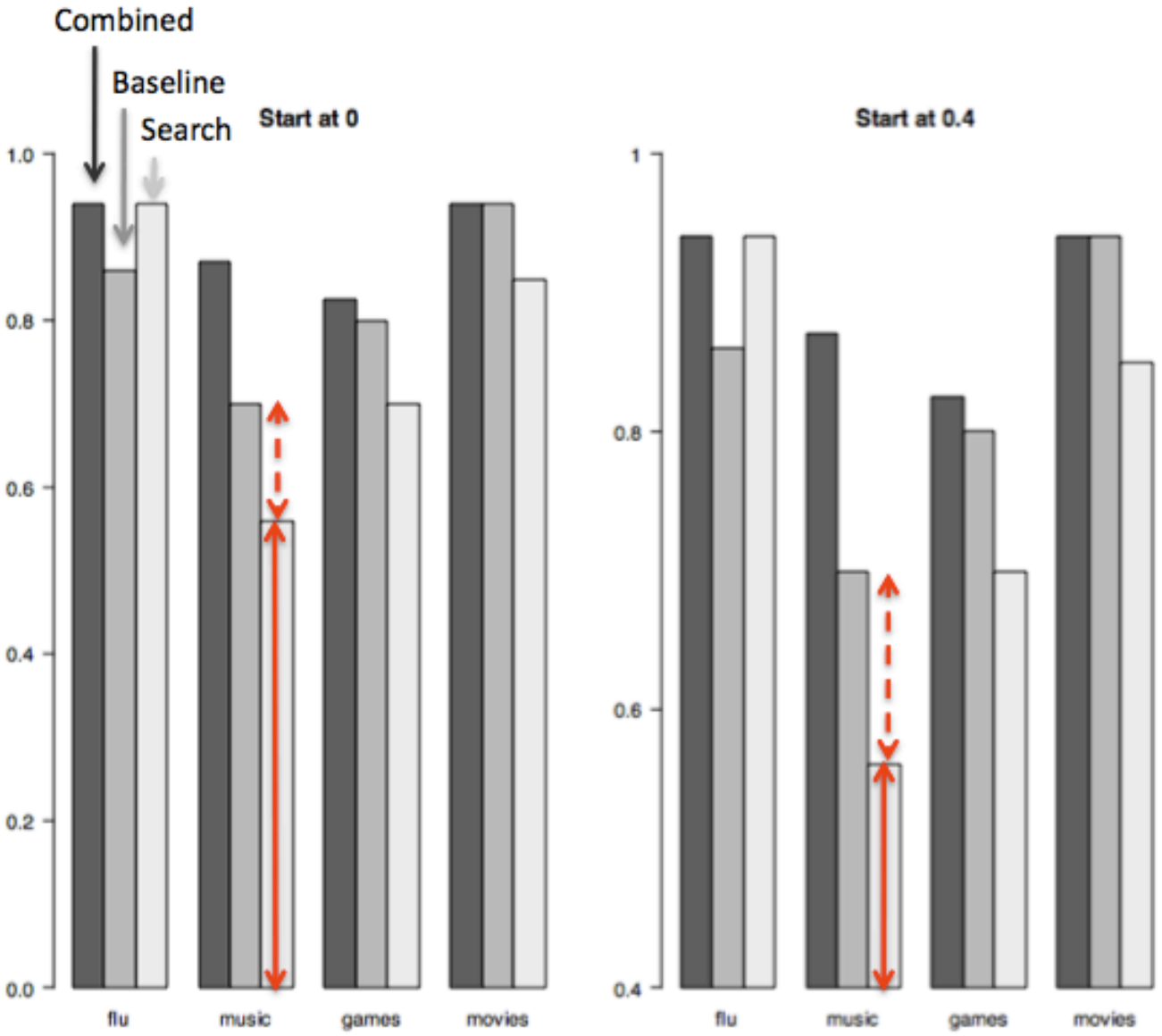
We will learn another important reason for using this formats later on (graphical perception!)

Trade-Offs and Comparative Advantages



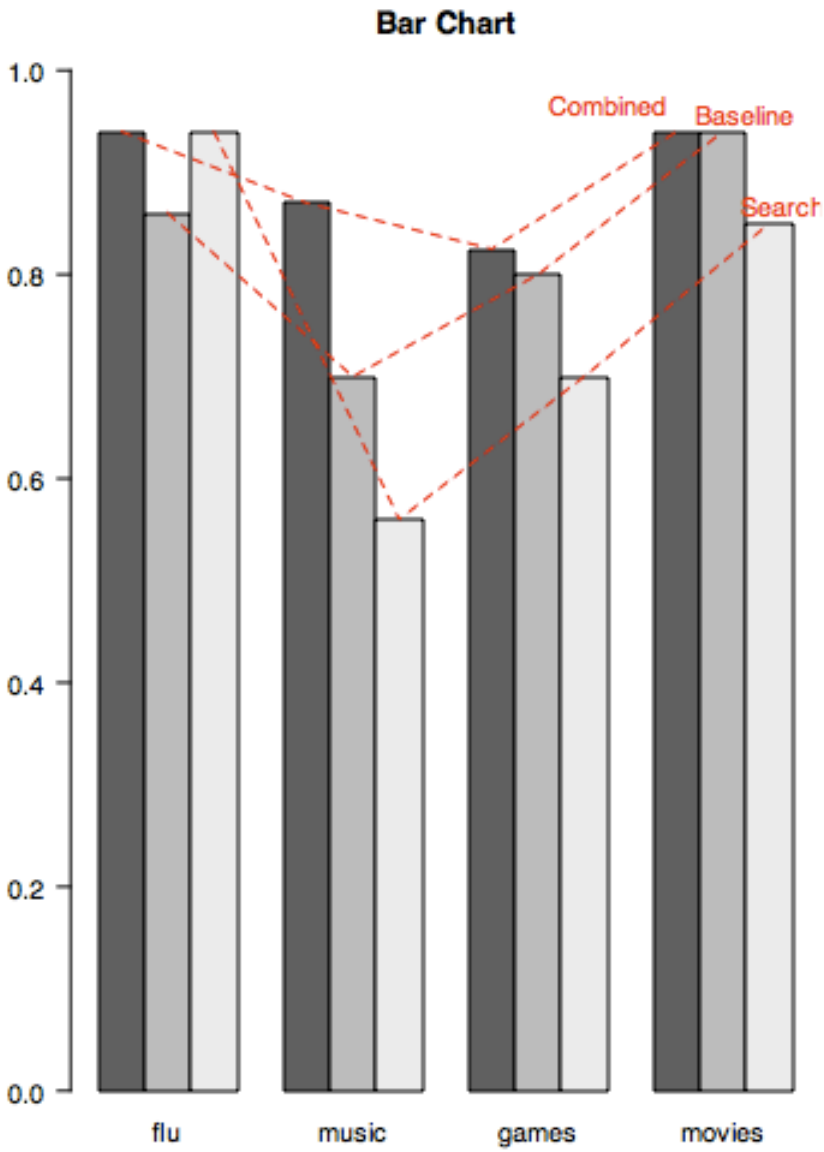
Source: Junkcharts.

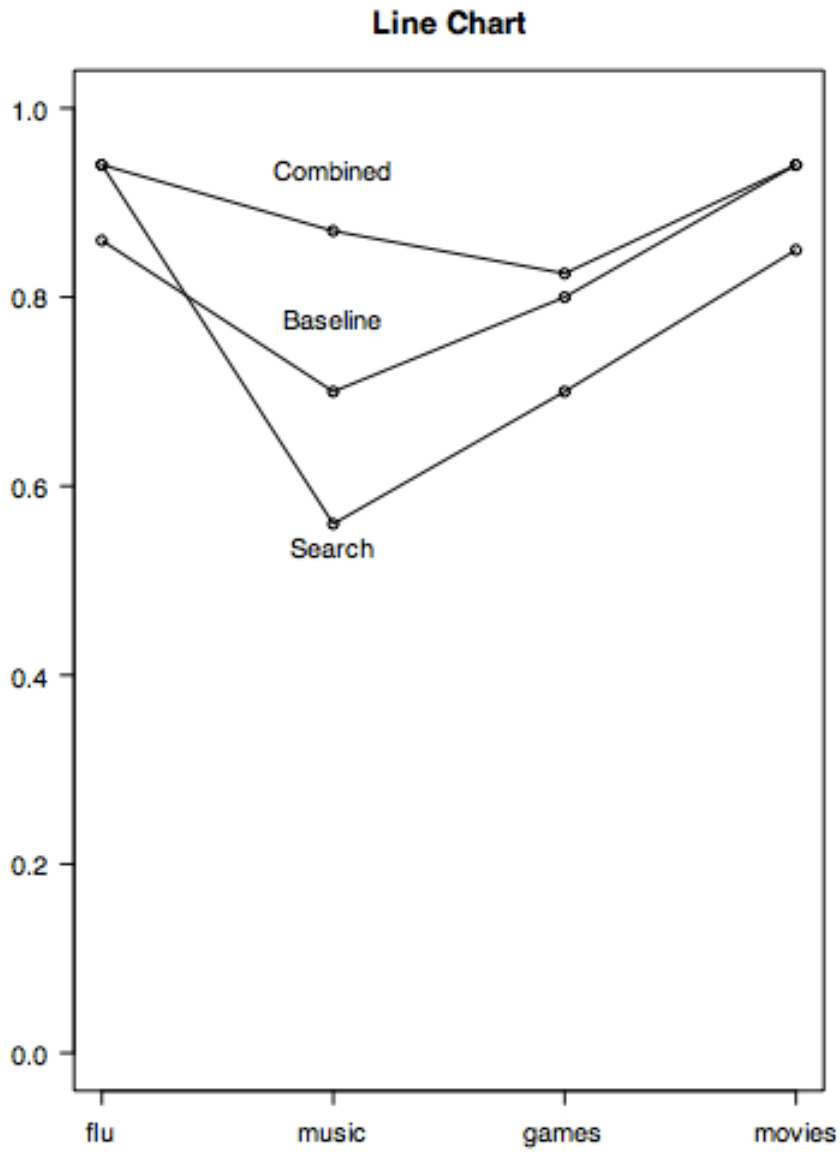
Trade-Offs and Comparative Advantages



Source: Junkcharts.

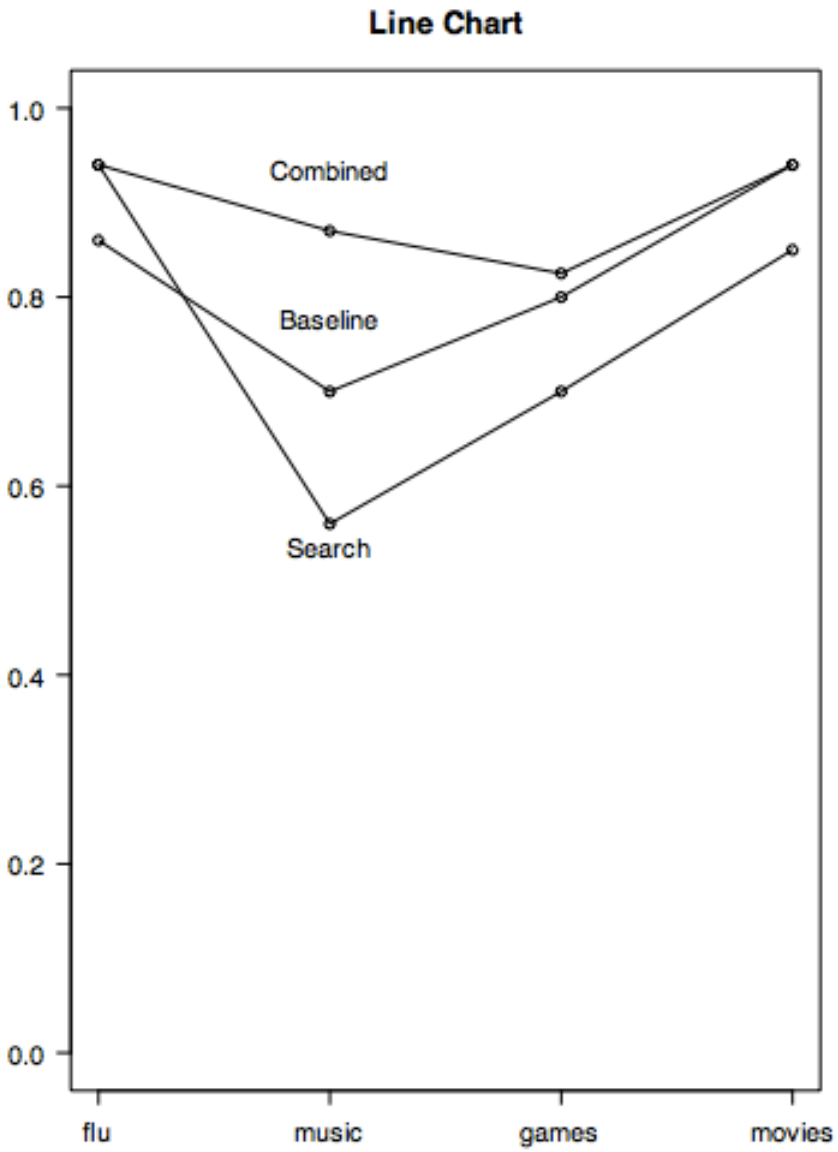
Trade-Offs and Comparative Advantages

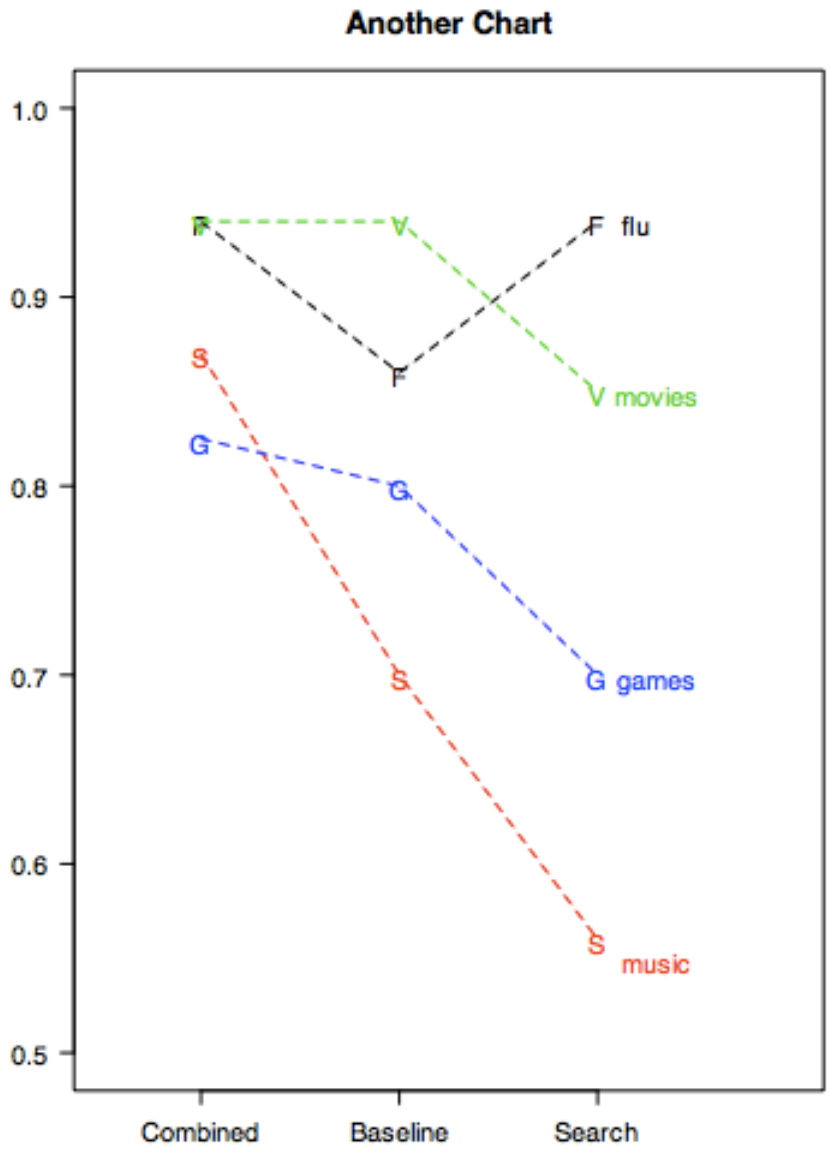




Source: Junkcharts.

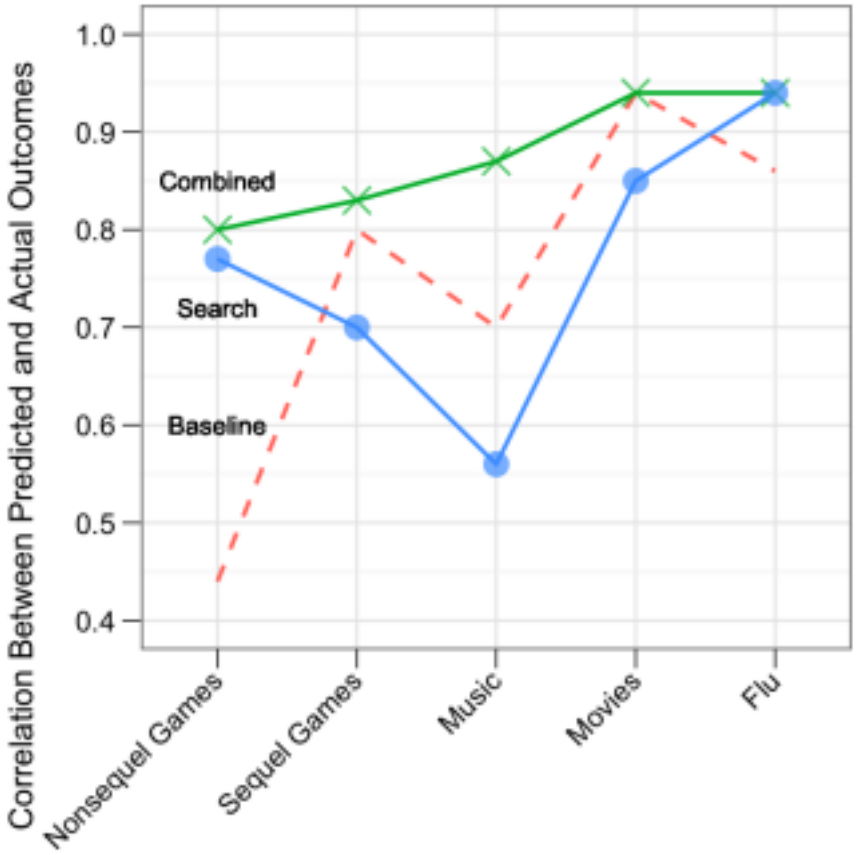
Trade-Offs and Comparative Advantages





Source: Junkcharts.

The Final Plot



Take Home Message

The choice of the right graphical format ultimately depends on the task or problem it is trying to solve.

Always try different graphical formats on the same data – they may reveal different aspects.

Constructing visualizations is almost always an iterative process – the first graph is rarely also the final one.

Comparison, Comparison, Comparison

"The fundamental analytical act in statistical reasoning is to answer the question ´compared to what?´

Whether we are evaluating changes over space or time, searching big data bases, adjusting and controlling for variables, designing experiments, specifying multiple regressions, or doing just about any kind of evidence-based reasoning, the essential point is to make intelligent and appropriate comparisons.

Thus visual displays, if they are to assist thinking, should show comparisons." (Tufte 2006: 127)

Some Visual Comparisons

Comparing before and after

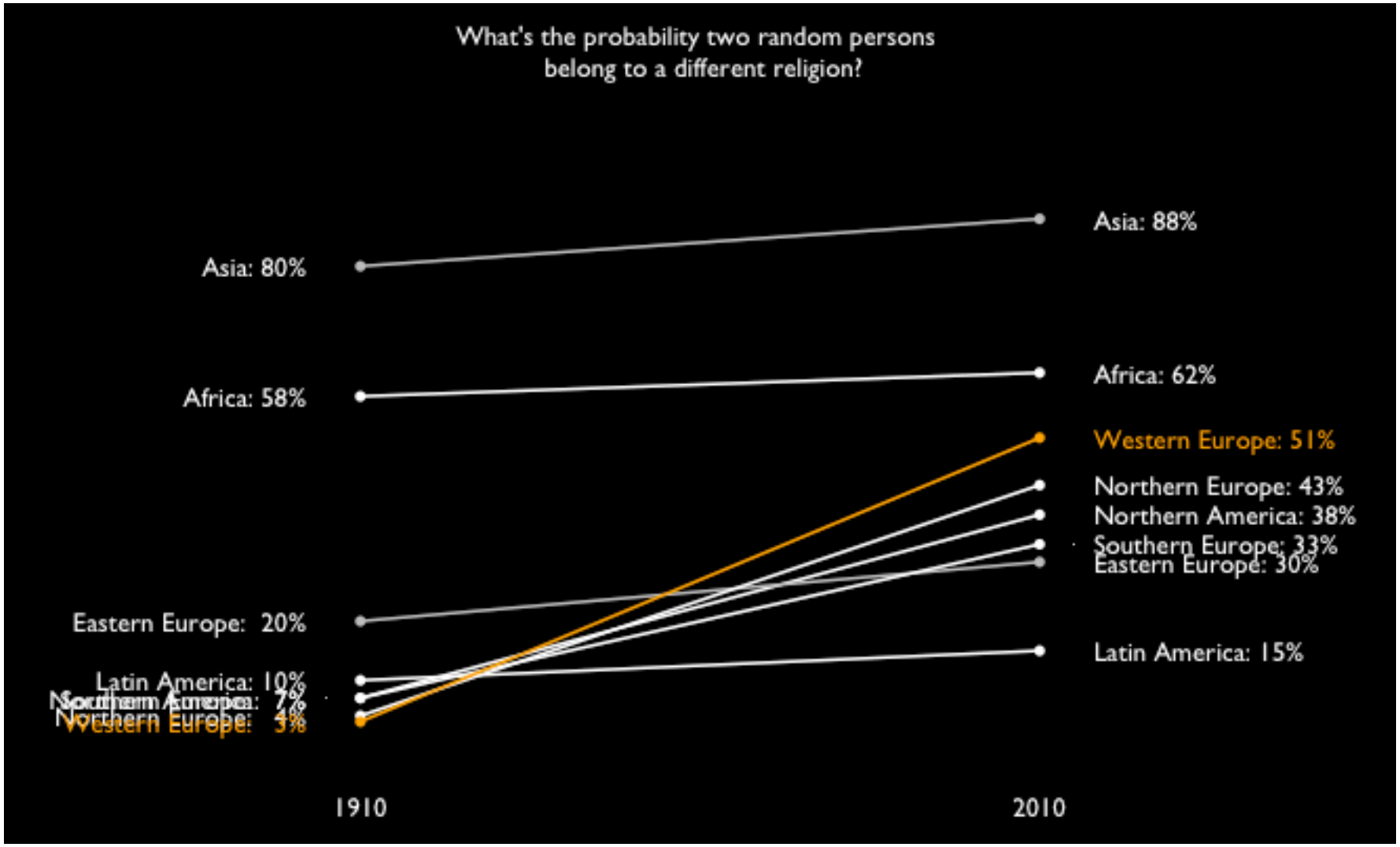
Comparing to a standard

Comparing by subgroups

Comparing to an implicit model ("what do we expect to see?")

A "visual discovery" is seeing a pattern, which *a priori* we did not expect to see.

Comparing before and after: Slope Graph



How many variables and data points? How many comparisons? How many insights?

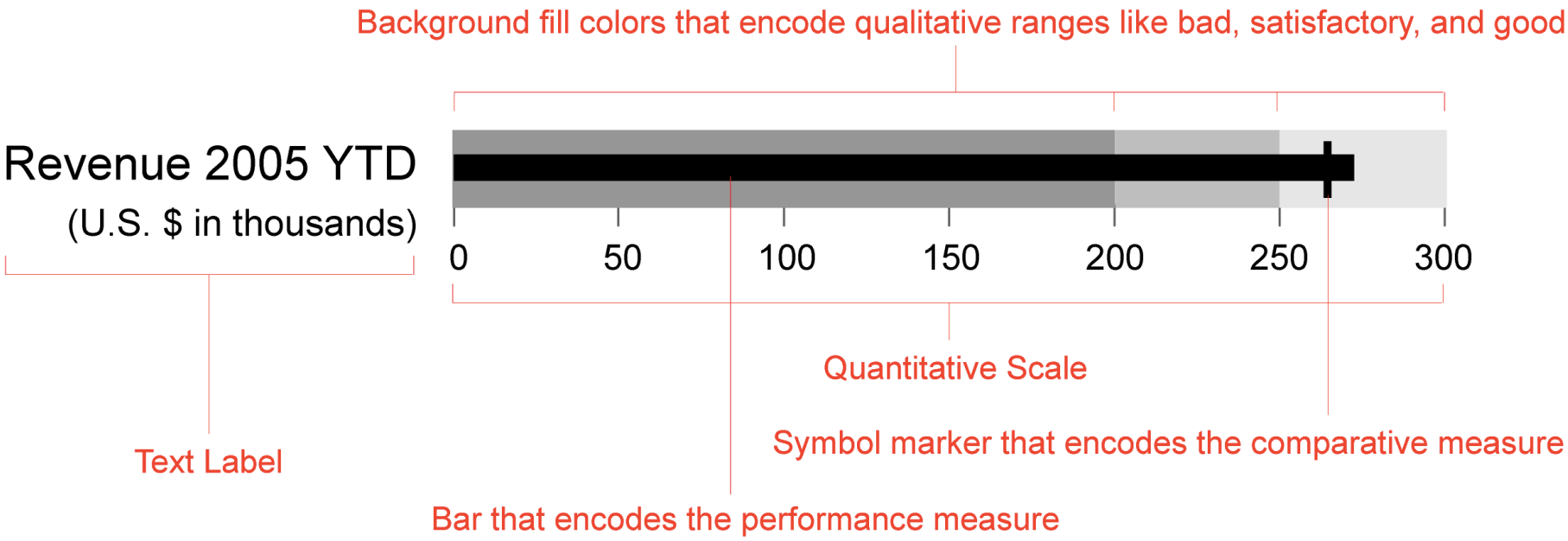
Current Receipts of Government as a Percentage of Gross Domestic Product, 1970 and 1979

Comparing to a Standard: Bullet Graph

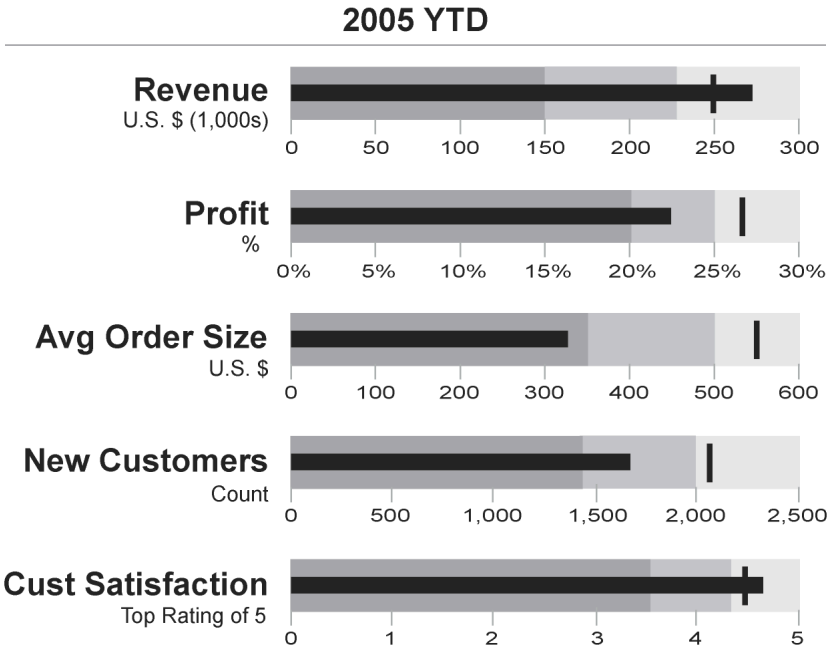
Bullet graphs compare a quantitative measure to

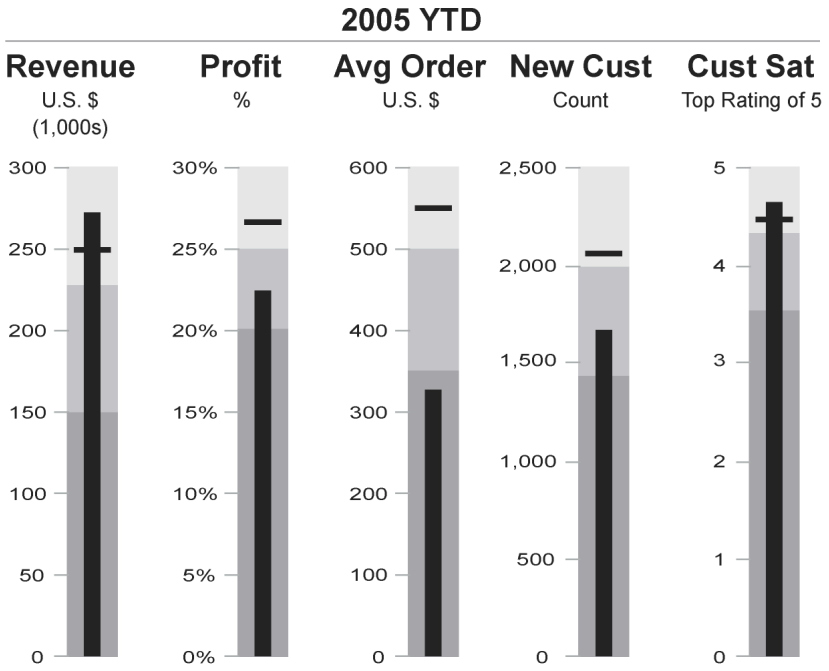
a) one or more related measures (e.g., a target or the same measure in the past) and

b) relate the measure to defined ranges that declare its qualitative state (e.g. good, satisfactory, and poor).



Comparing to a Standard: Bullet Graph





How many variables and data points?

How many comparisons?

How many insights?

Comparing by Subgroups: Small Multiples

Should the comparison be made within a single graph or across multiple graphs?

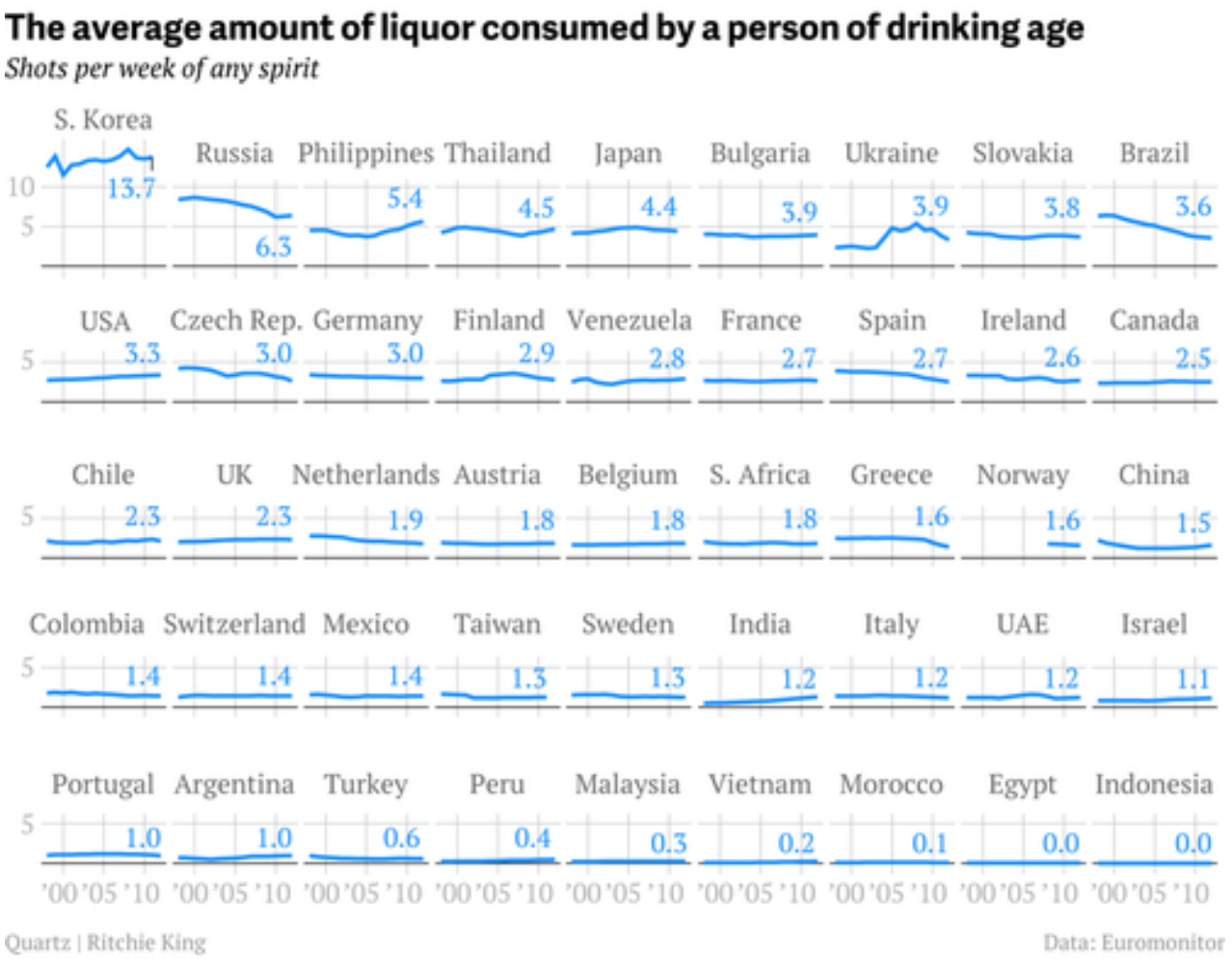
If multiple graphs, make sure to use:

Common graph size

Common scales

Helpful alignment

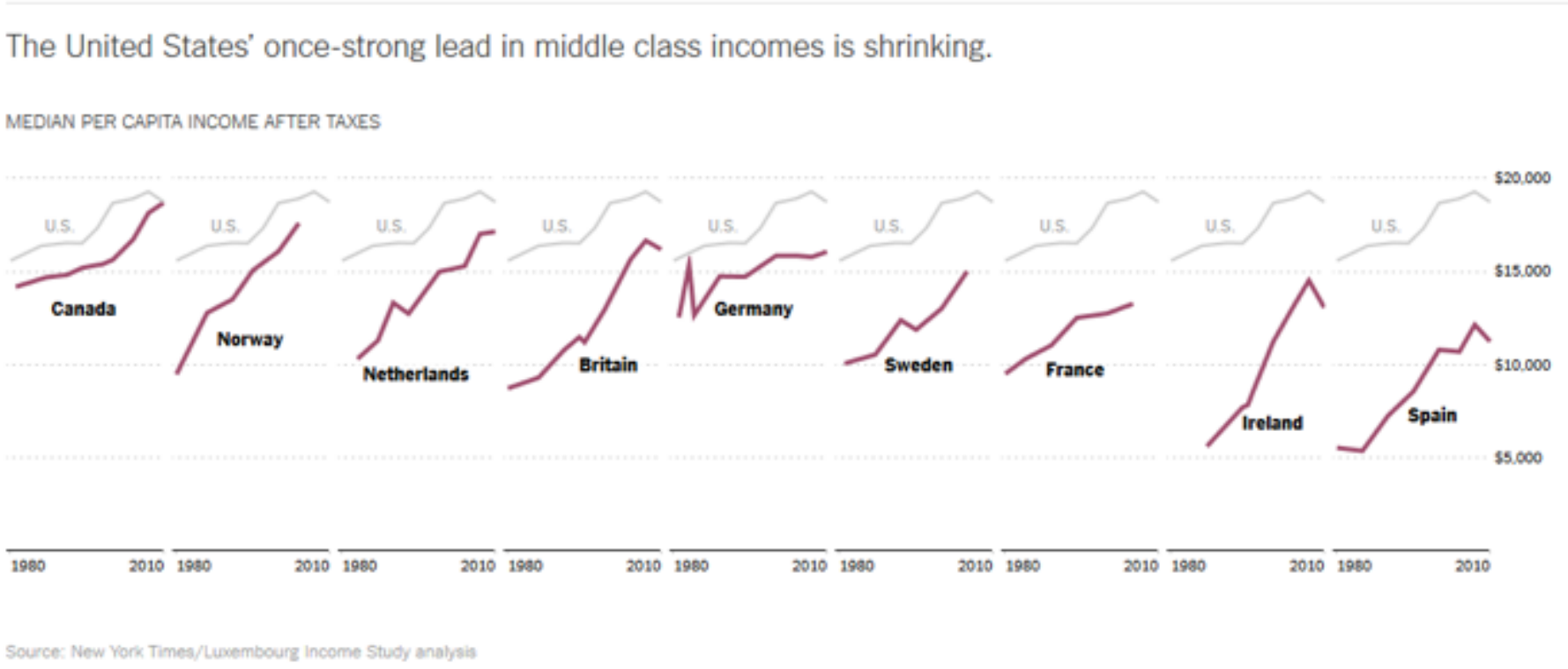
"High-density graphics help us to compare parts of the data by displaying much information within the view of the eye: we look at one page at a time and the more on the page, the more effective and comparative our eye can be." (Tufte 2001: 168).



How many variables and data points?

How many comparisons?

How many insights?



How many variables and data points?

How many comparisons?

How many insights?

2000: State-level support (orange) or opposition (green) on school vouchers, relative to the national average of 45% support

Orange and green colors correspond to states where support for vouchers was greater or less than the national average.
The seven ethnicreligious cagetories are mutually exclusive. "Evangelicals" includes Mormons as well as born-again Protestants.
Where a category represents less than 1% of the voters of a state, the state is left blank.

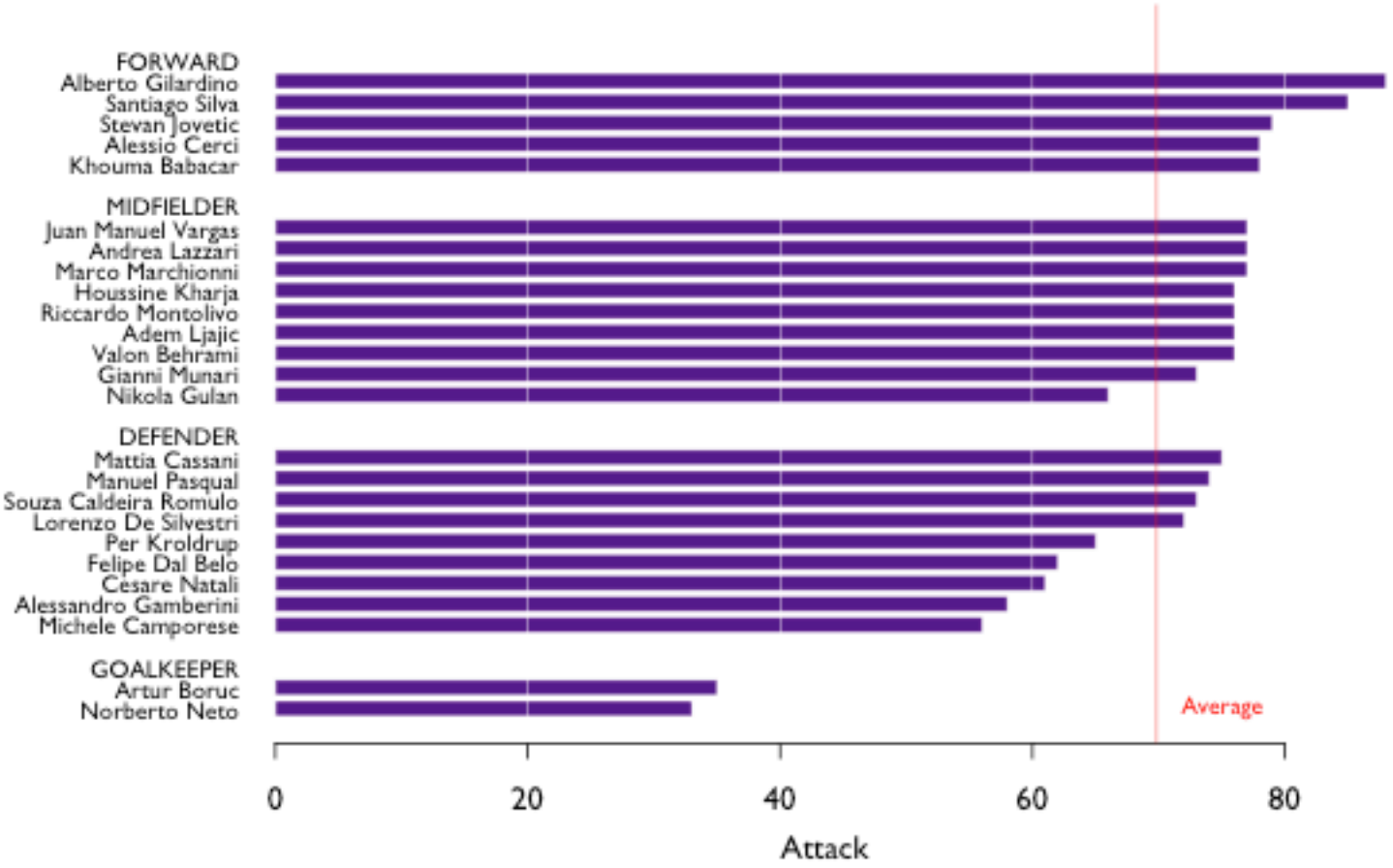
Comparing to a Standard & Comparing Subgroups: Lab Exercise

Football player skill ratings (at the time of September 2011)

The skill ratings are from the PES Stats Database (PSD), a community-based approach to create a database with accurate statistics and skill ratings for soccer players (originally for the video game "Pro Evolution Soccer" by Konami).

43 variables on 1851 players from 4 European Leagues (Bundesliga, La Liga, Premier League, Serie A)

I provided code using data from AC Fiorentina– you can use the code and data I provided, but be sure to pick another team!


FORWARD
Alberto Gilardino
Santiago Silva
Stevan Jovetic
Alessio Cerci
Khouma Babacar
MIDFIELDER
Juan Manuel Vargas
Andrea Lazzari
Marco Marchionni
Houssine Kharja
Riccardo Montolivo
Adem Ljajic
Valon Behrami
Gianni Munari
Nikola Gulan
DEFENDER
Mattia Cassani
Manuel Pasqual
Souza Caldeira Romulo
Lorenzo De Silvestri
Per Kroldrup
Felipe Dal Belo
Cesare Natali
Alessandro Gamberini
Michele Camporese
GOALKEEPER
Artur Boruc
Norberto Neto
Average
0
20
40
60
80
Attack

Lab Exercise with Football Data

Let's visually explore the offensive (`Attack`) and defensive (`Defence`) skills of football players.

What *a priori* expectations about the data and possible patterns do you have?

What general exploration strategy would be appropriate?

Which comparisons are possible and interesting?

Which graphical formats would be useful to make these comparisons?

Make sure you experiment with at least three graphical formats as well as with comparisons within a single plot and comparisons across multiple plots!

Graphical Perception

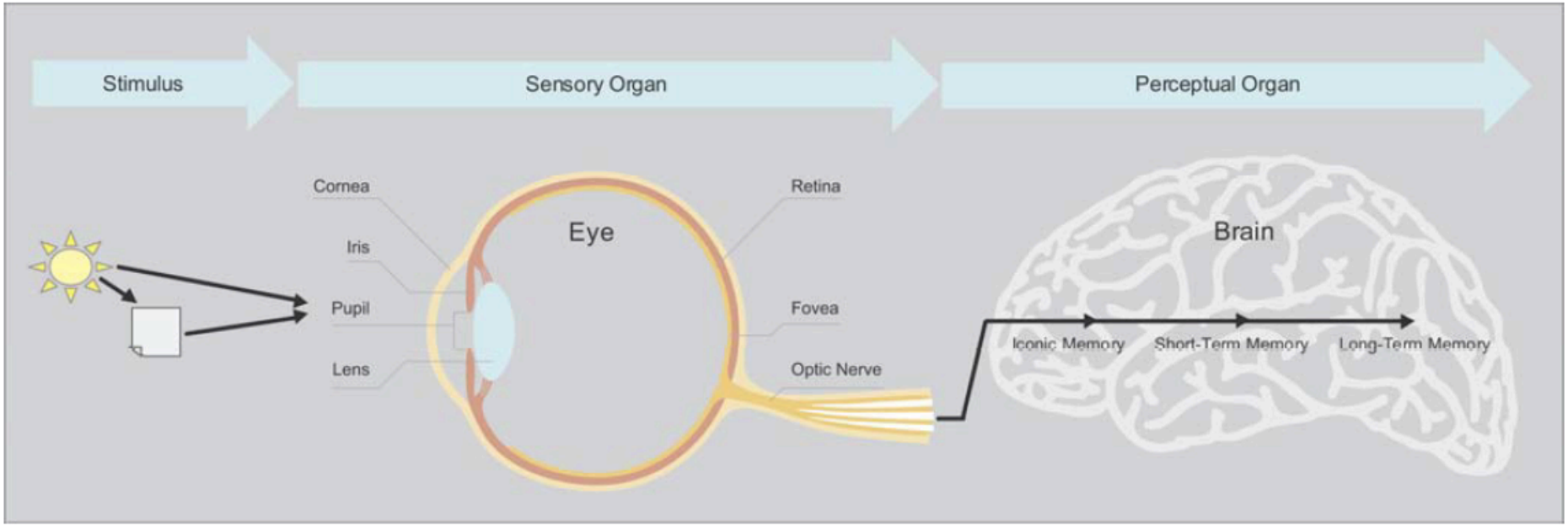
"If we can understand how perception works, our knowledge can be translated into rules for displaying information.

Following perception-based rules, we can present our data in such a way that the important and informative patterns stand out.

If we disobey the rules, our data will be incrompehensible or misleading."

(C. Ware 2004: xxi)

How Perception Works

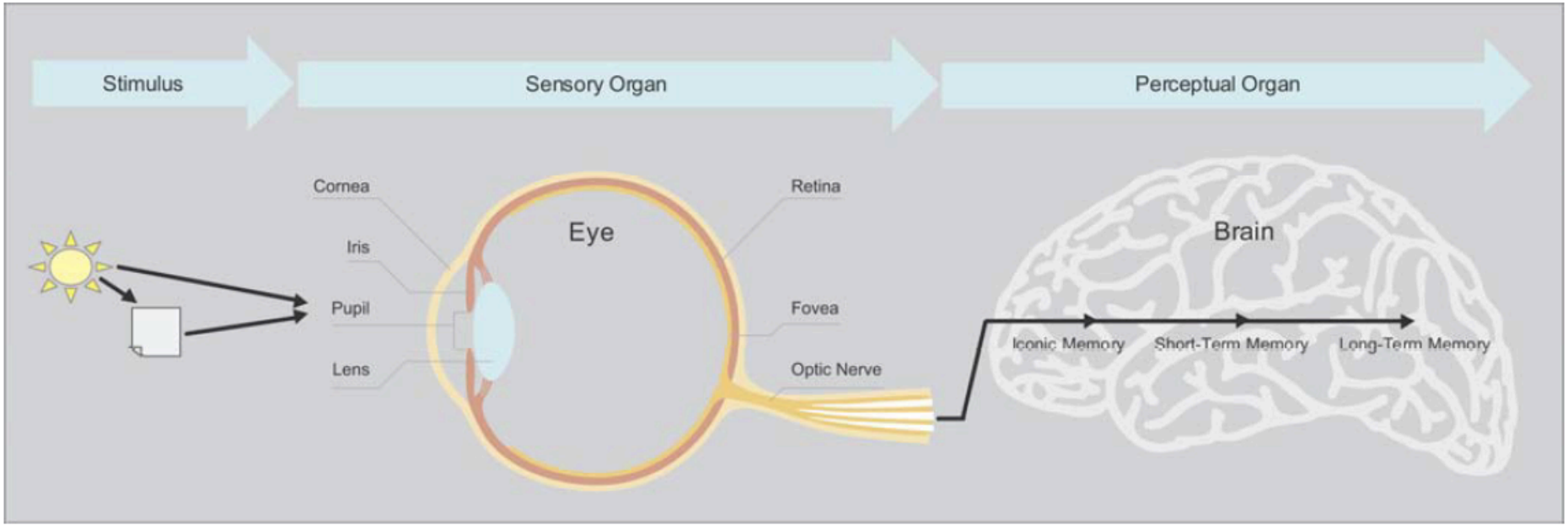


Light passes through the pupil and hits the retina – a thin surface coated with millions of nerves called rods and cones.

Rods and cones translate what they detect into neural signals and pass them on to the brain.

We don't see things with our eyes – we see them with our brains!

How Perception Works



Iconic Memory (Visual Sensory Register): Pre-attentive (unconscious and automatic) processing of visual attributes (colors, shapes, etc.)

Short Term or Working Memory: Attentive processing of "chunks" of information

Long Term Memory: Storing and recognizing familiar patterns

Attributes of Pre-attentive Processing

756395068473

658663037576

860372658602

846589107830

How many 3s are there?

Attributes of Pre-attentive Processing

7563950684 73

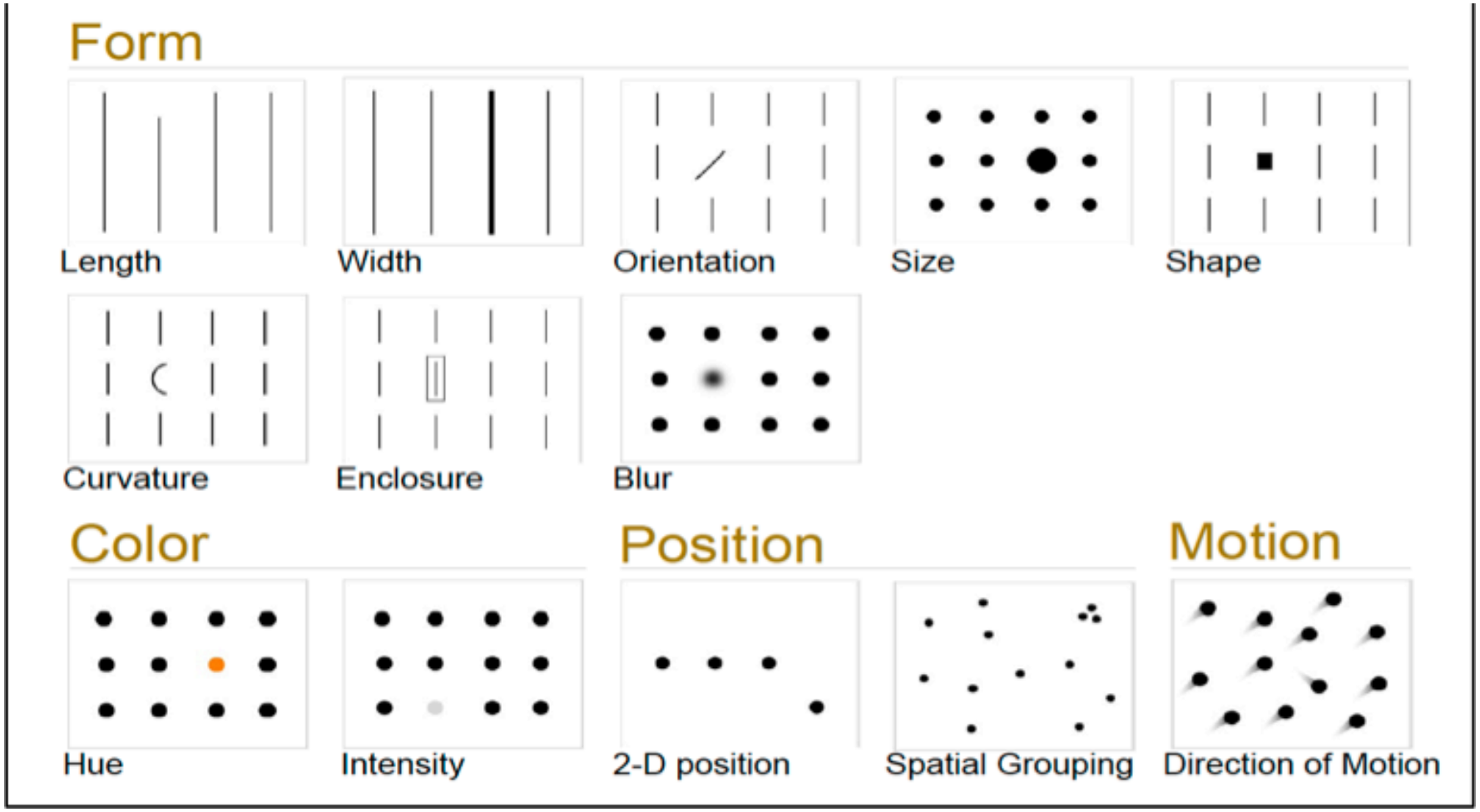
6586630375 76

8603726586 02

8465891078 30

How many 3s are there?

Attributes of Pre-attentive Processing



Attributes of Pre-attentive Processing

We can use attributes of pre-attentive processing to

1. Choose of visual encodings of abstract data

2. Direct audiences attention

3. Create a visual hierarchy

Attributes of Pre-attentive Processing

Attributes of Pre-attentive Processing

Pre-attentive attributes become less distinct as the variety of distractors increases.

Attributes of Pre-attentive Processing

How much bigger is the second circle?

Attributes of Pre-attentive Processing

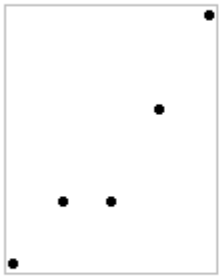
Bring the colors in an order from small to large!

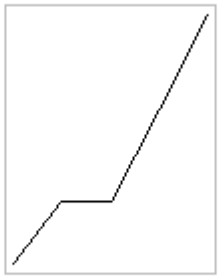
Attributes of Pre-attentive Processing

Visual attributes are not all created equal…

Attribute	*Quantitatively Perceived?*
2-D Position	Yes
Length	Yes
Width	Limited
Size	Limited
Color Intensity	Limited
Orientation	No
Shape	No
Enclosure	No
Color Hue	No

Now you know why…





Dot Chart: 2-D position of visual objects

Line Chart: 2-D position, connected to give shape to a series of values

Bar Chart: Length and 2-D position

Gestalt Principles of Visual Perception

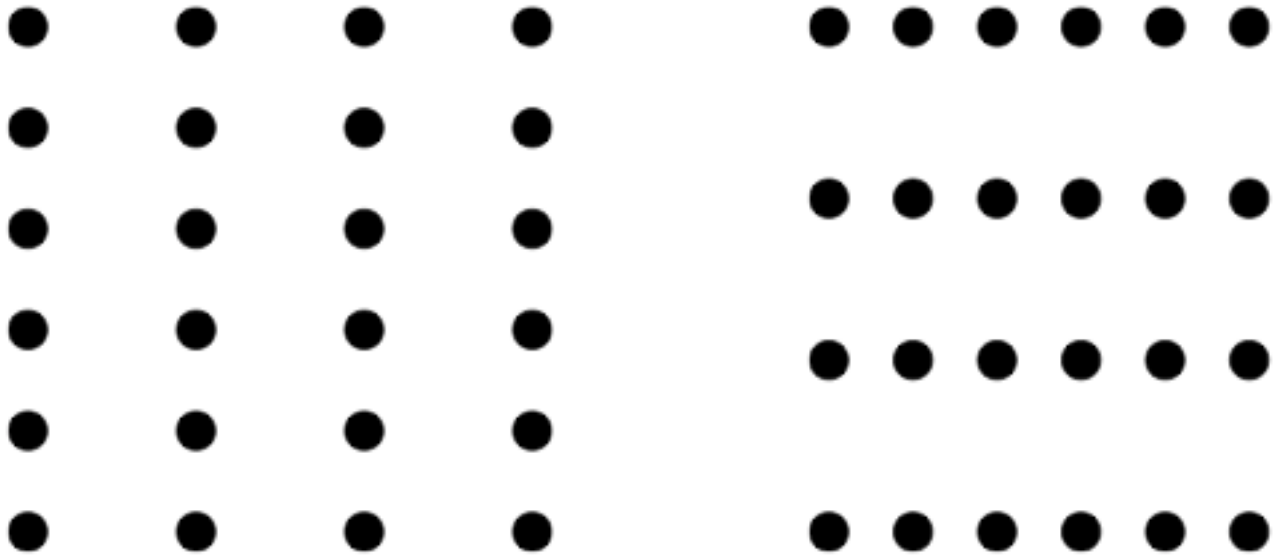
Visualization works because it abstracts from single data points and leverages the human brain´s ability to form these data points into emergent patterns.

This ability has been studied by the Gestalt School of Psychology in the early 1900s.

Understanding this ability allows us to use it strategically in the design of visualizations.

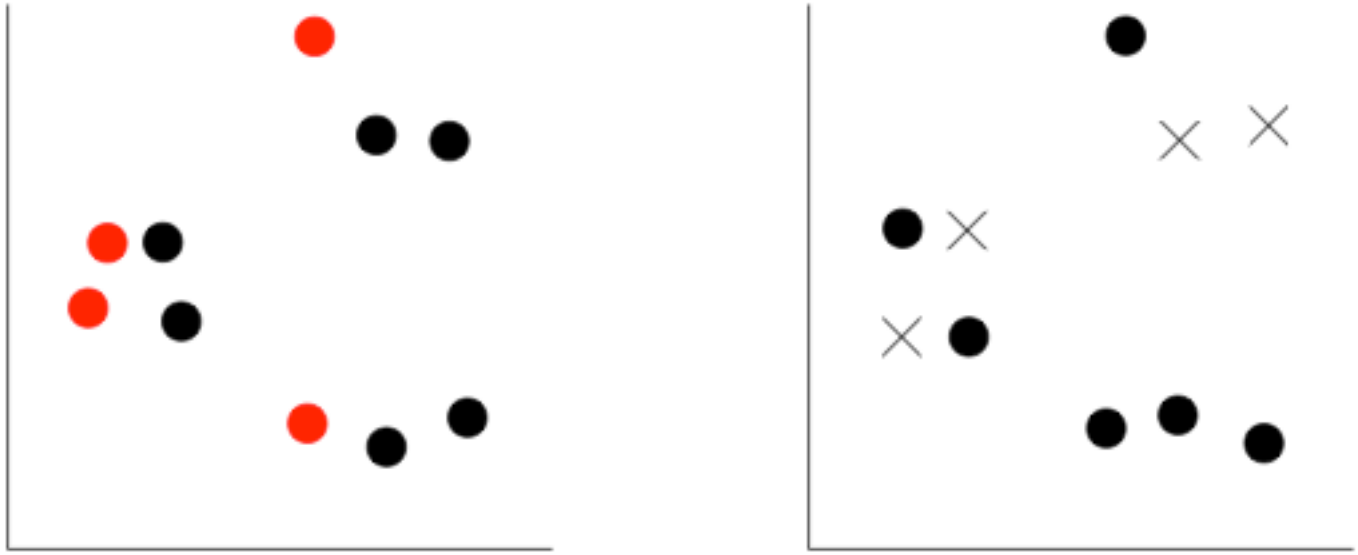
Gestalt Principles of Visual Perception

Proximity: Objects that are close together are perceived as a group.



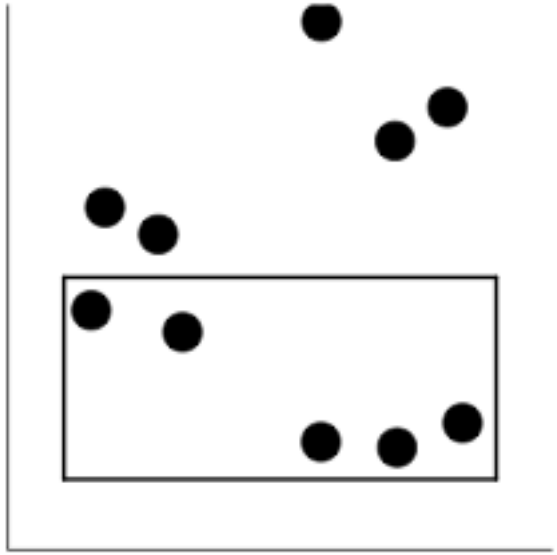
Gestalt Principles of Visual Perception

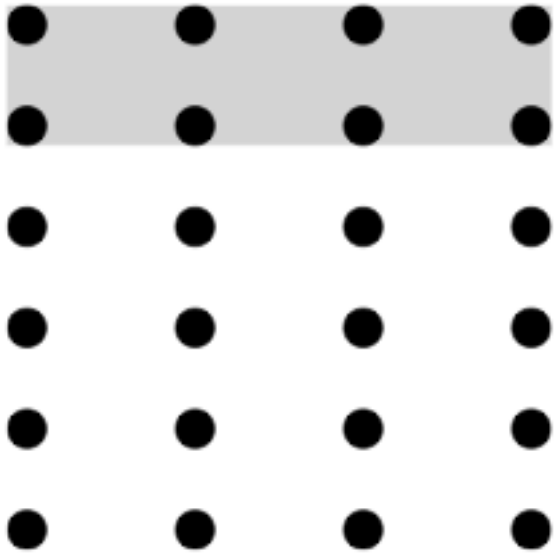
Similarity: Objects that share similar attributes (e.g. color or shape) are perceived as a group.



Gestalt Principles of Visual Perception

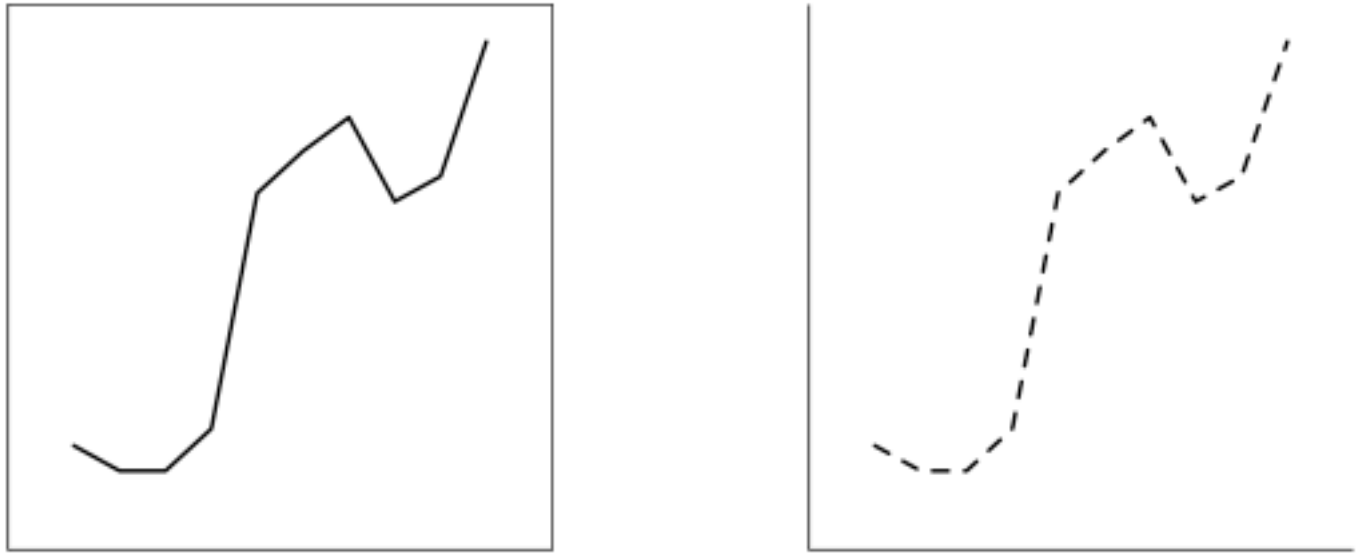
Enclosure: Objects that have boundary around them are perceived as a group.





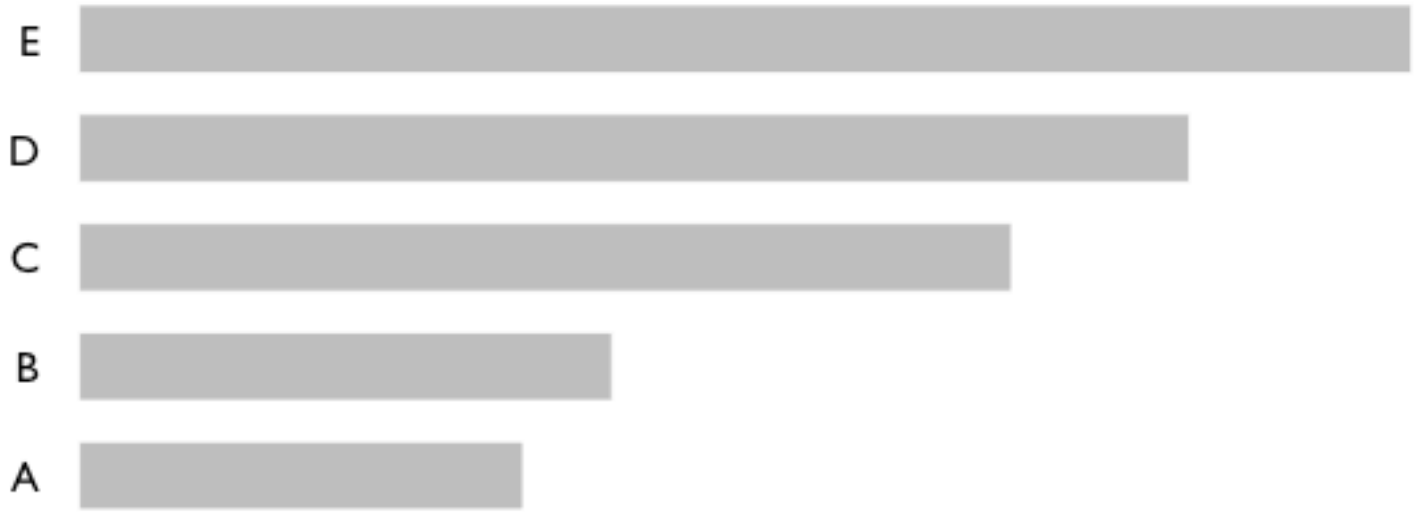
Gestalt Principles of Visual Perception

Closure: Open structures are perceived as closed, complete, and regular.



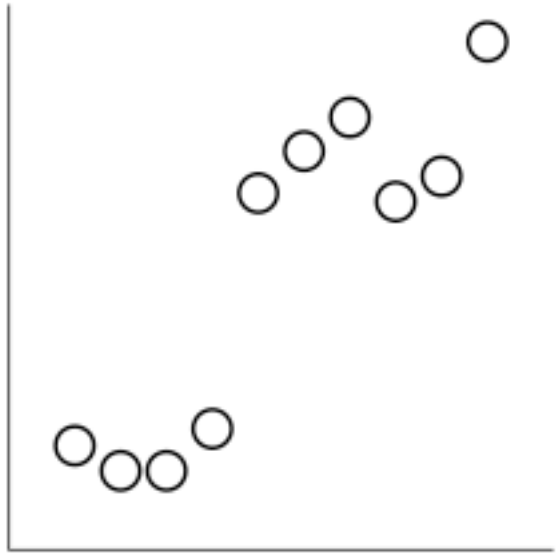
Gestalt Principles of Visual Perception

Continuity: Objects that are aligned together are perceived as a group.



Gestalt Principles of Visual Perception

Connection: Objects that are connected (e.g. by a line) are perceived as a group.



Reduce the Non-Data Pixels

1. Subtract unnecessary non-data pixels.

Ask yourself: "Would the data suffer any loss of meaning or impact if this were eliminated?"

If the answer is "no," then get rid of it.

2. De-emphasize and regularize the remaining non-data pixels.

e.g. use thin lines and light grey for supporting non-data components of the graph (axes, labels, etc.)

Enhance the Data Pixels

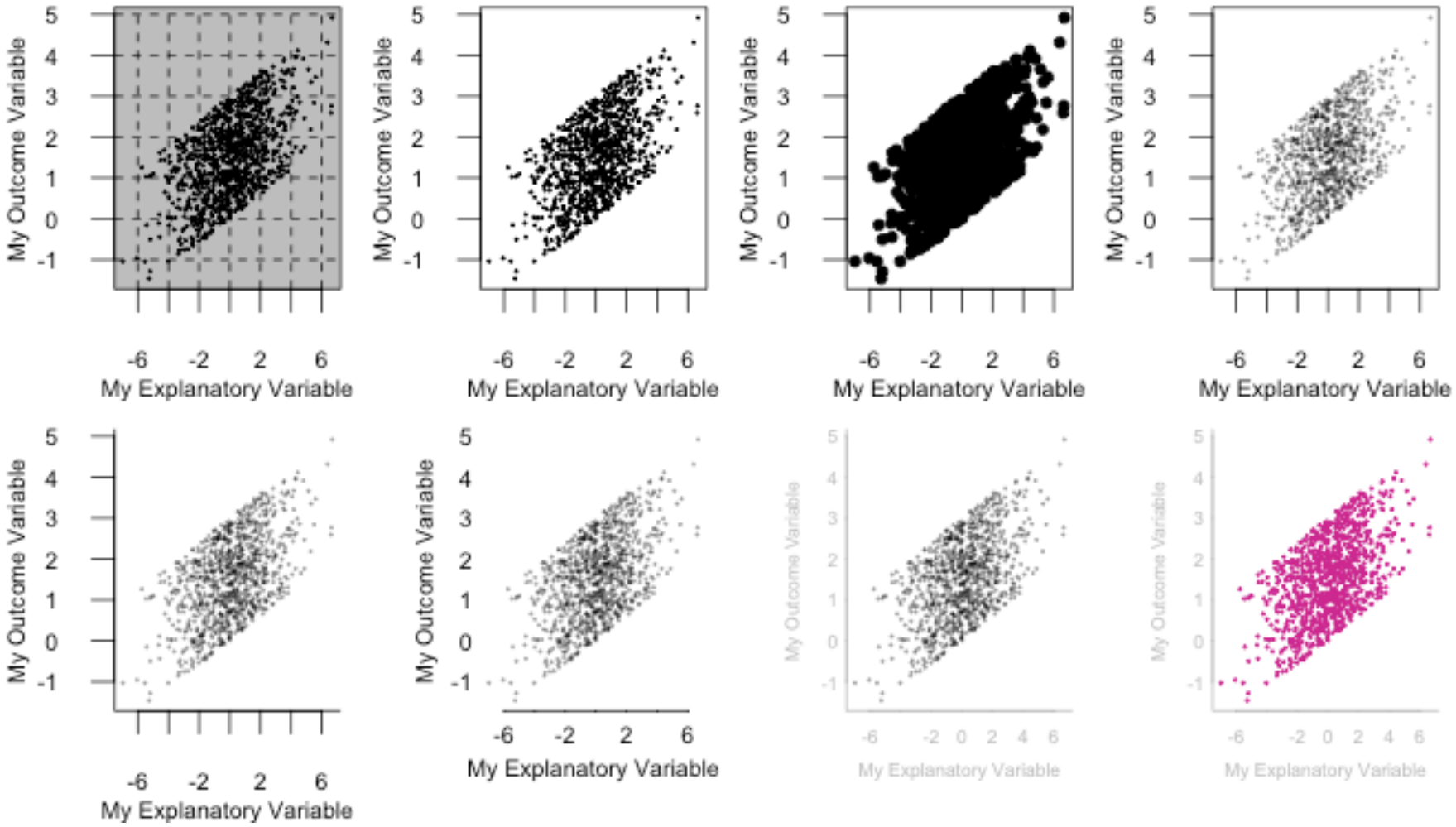
1. Subtract unnecessary data pixels

Not all information is equally important.

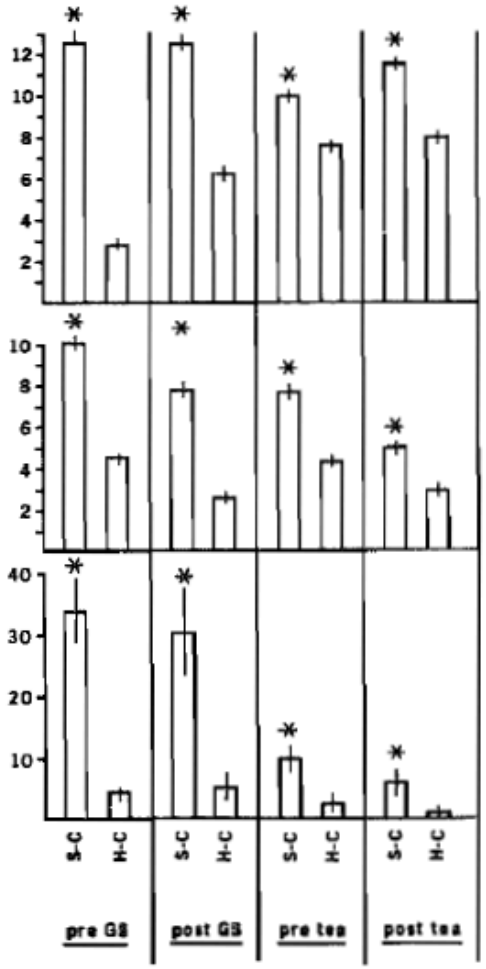
2. Emphasize the most important data pixels

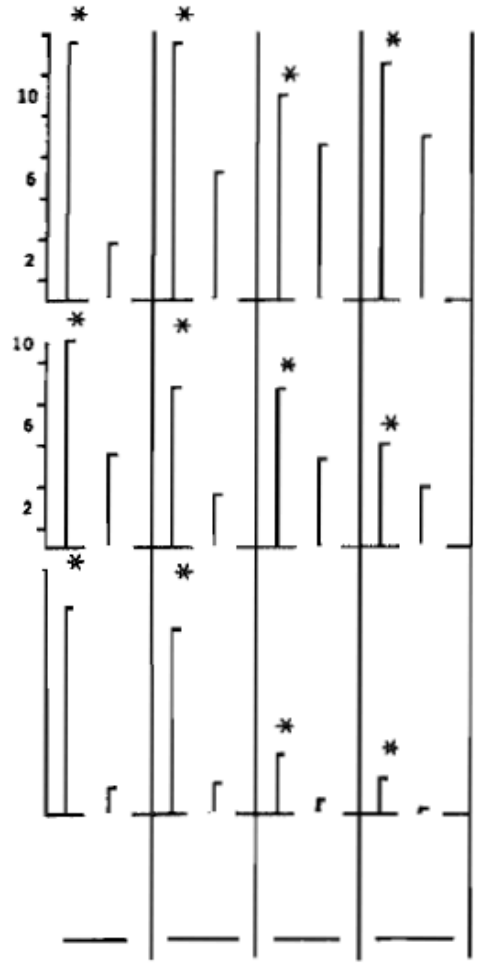
Use attributes of pre-attentive processing (e.g. color, size, width) to emphasize the most important data pixels.

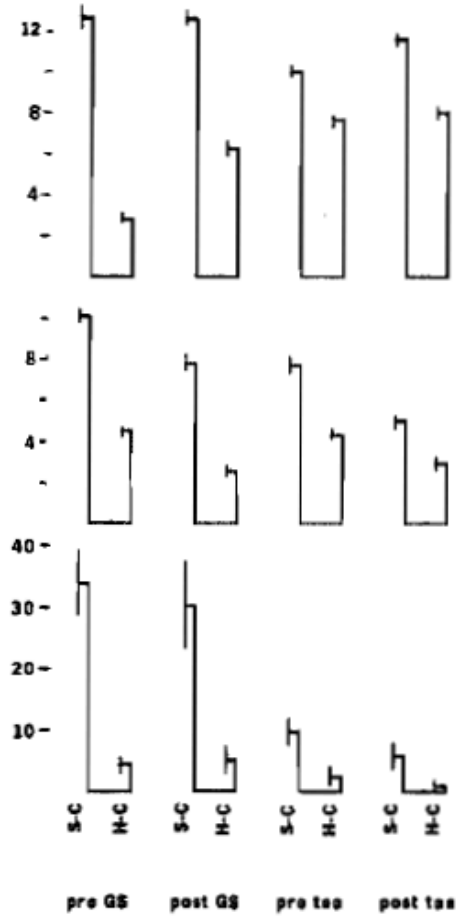
Example



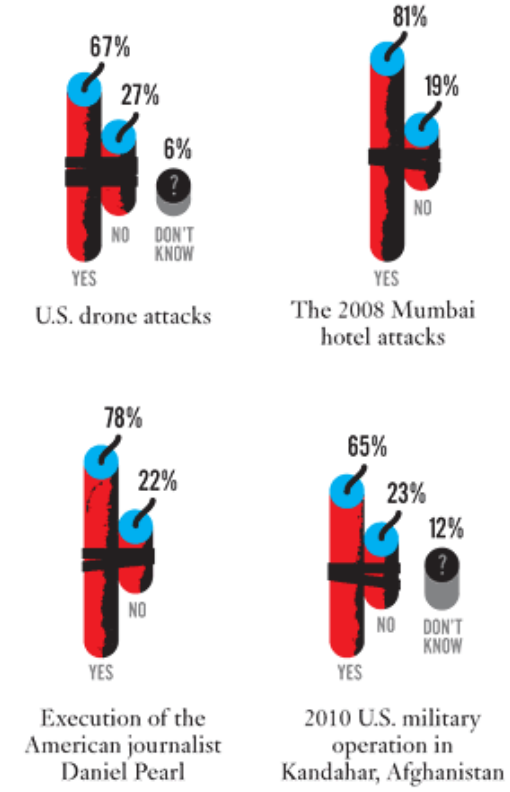
(Radical) Tufte Example



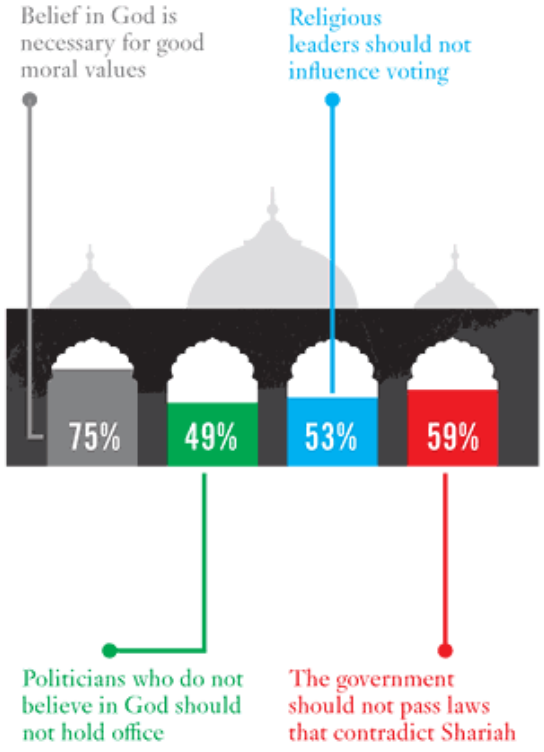




Which Is an Act of Terrorism?

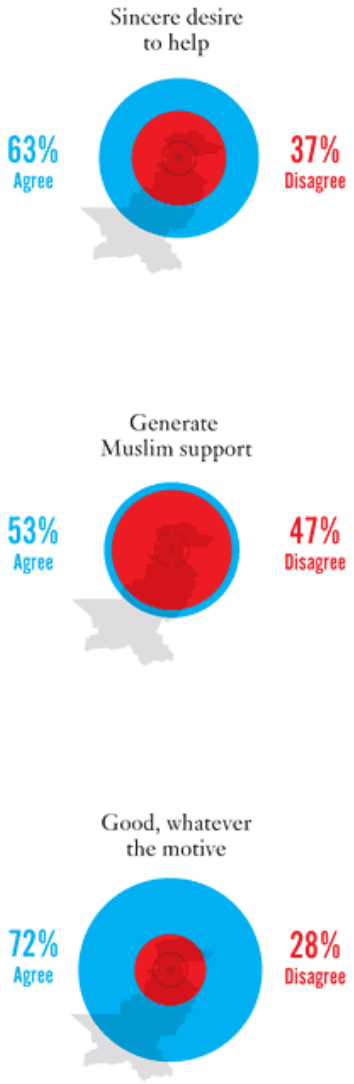


Religion



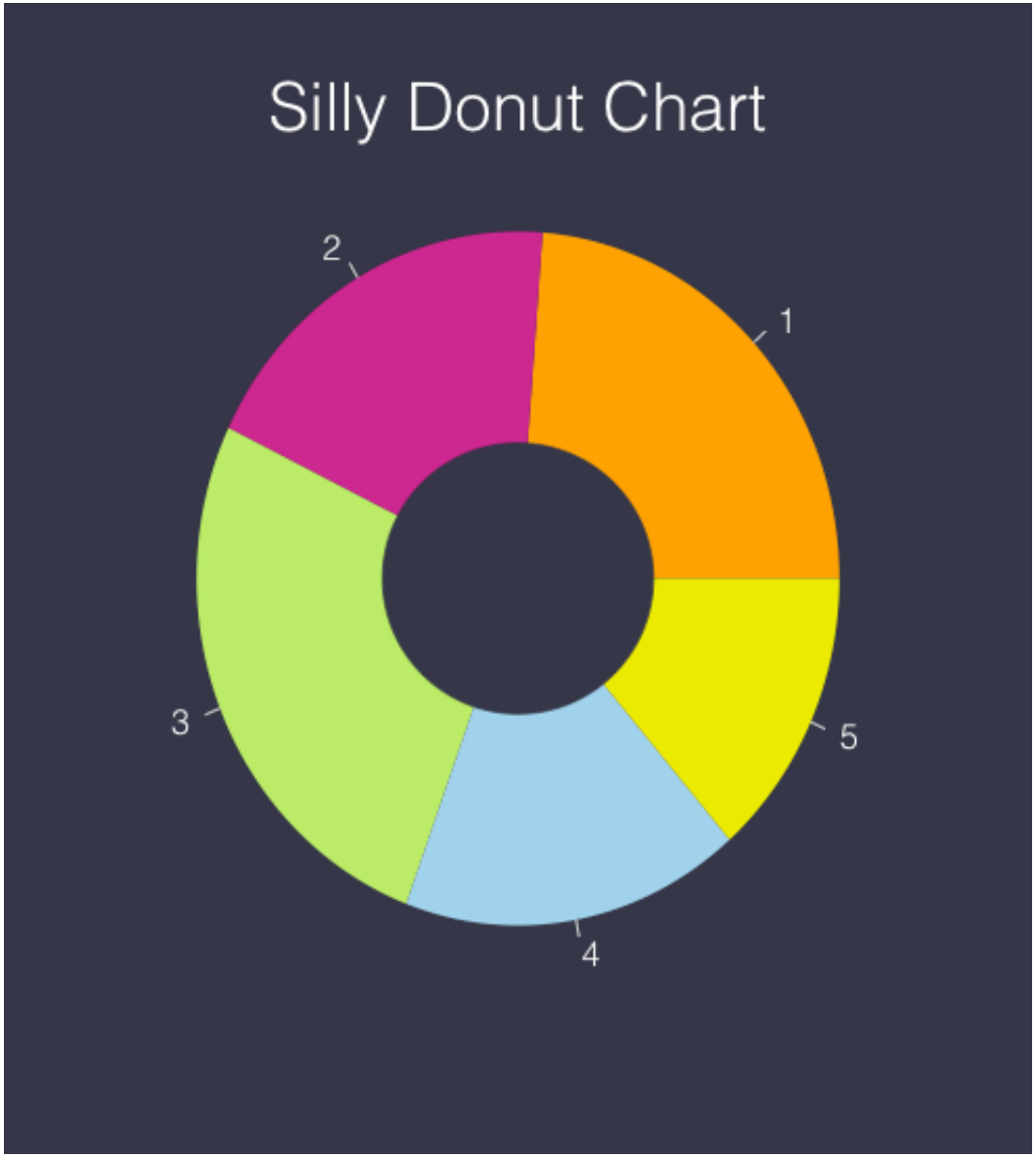
Earthquake Aid

WHY DID THE U.S. GIVE MILLIONS AFTER THE 2005 KASHMIR EARTHQUAKE?



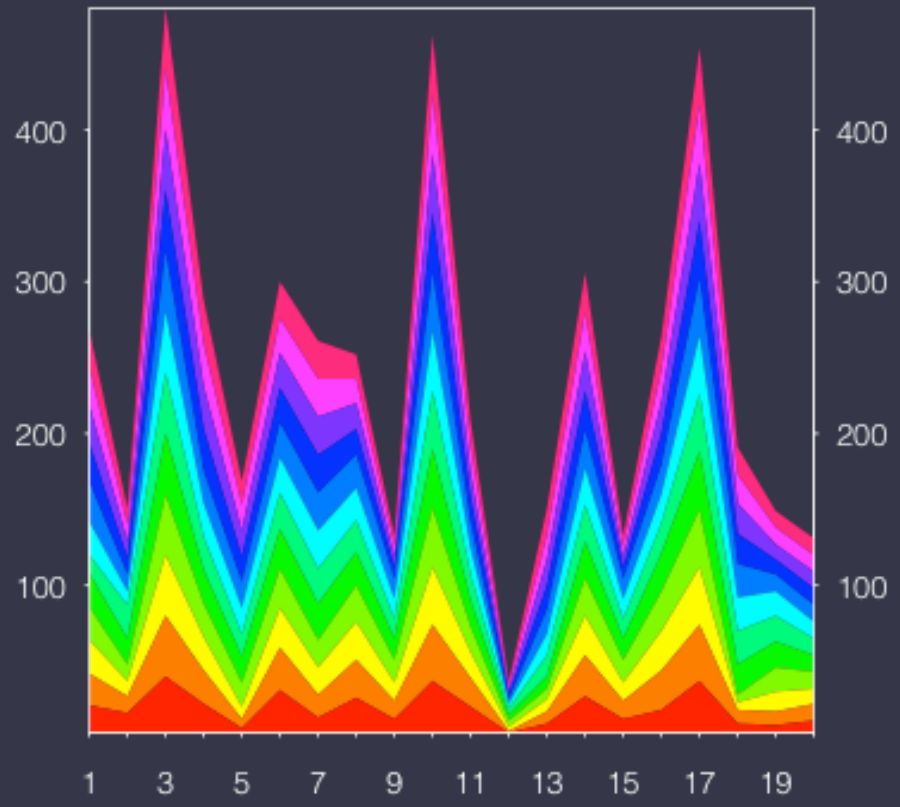
Intermission:

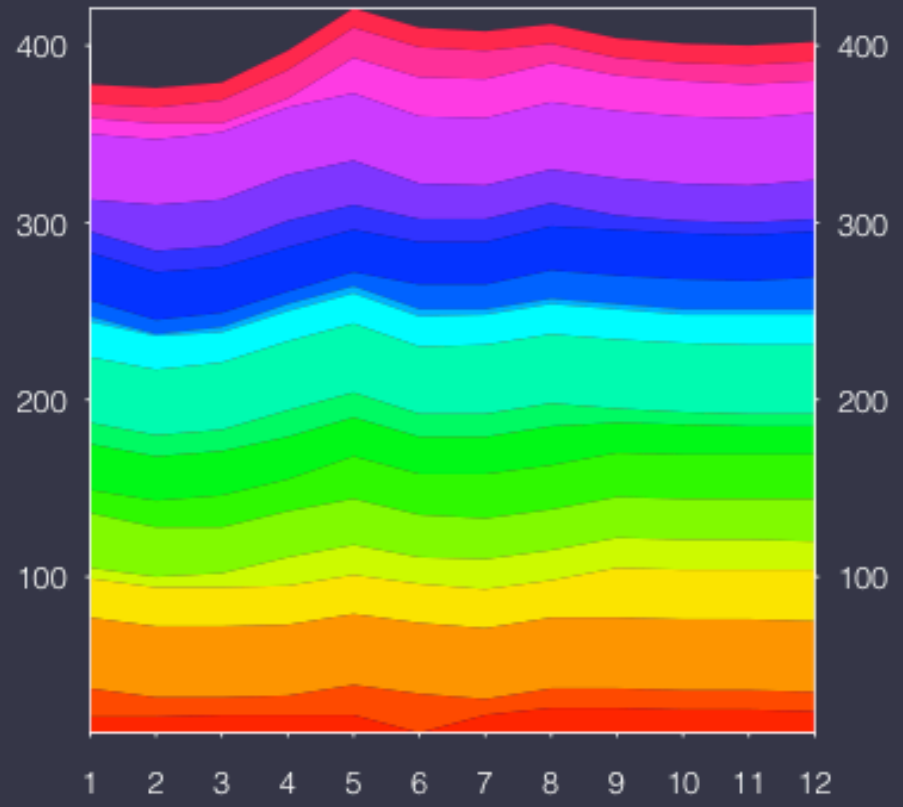
P-p-pop Charts!



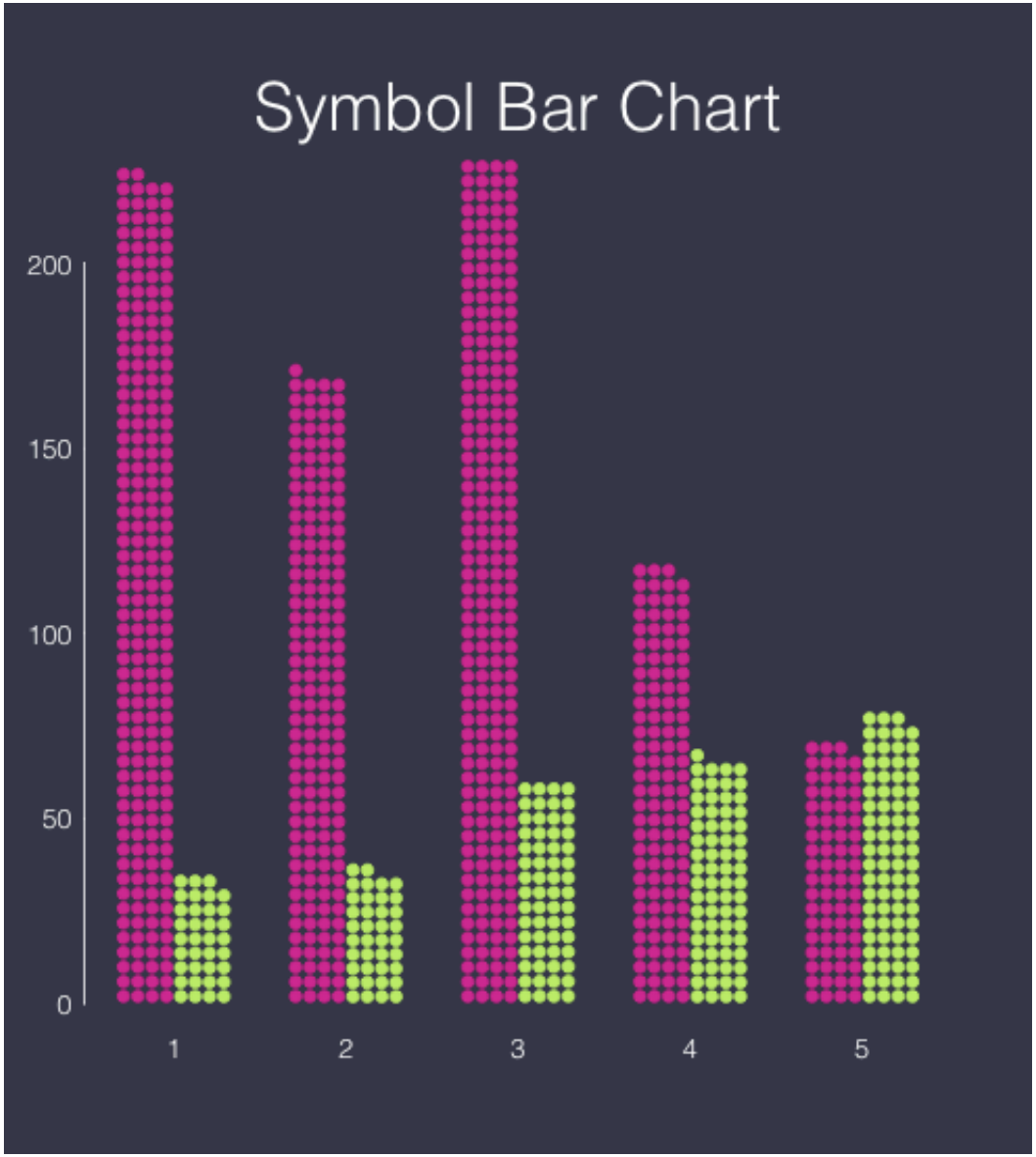

```
library(plotrix)

my.colors <- c("orange","
maroon3", "darkolivegreen2",
"lightskyblue2", "yellow2")

pie(table(data$x),
col=my.colors, border=F,
main="Silly Donut Chart",
cex.main=2)

points(0,0, col=rgb(54, 57,
74, 255, max=255), cex=18,
pch=19)
```

Stacked Line Chart



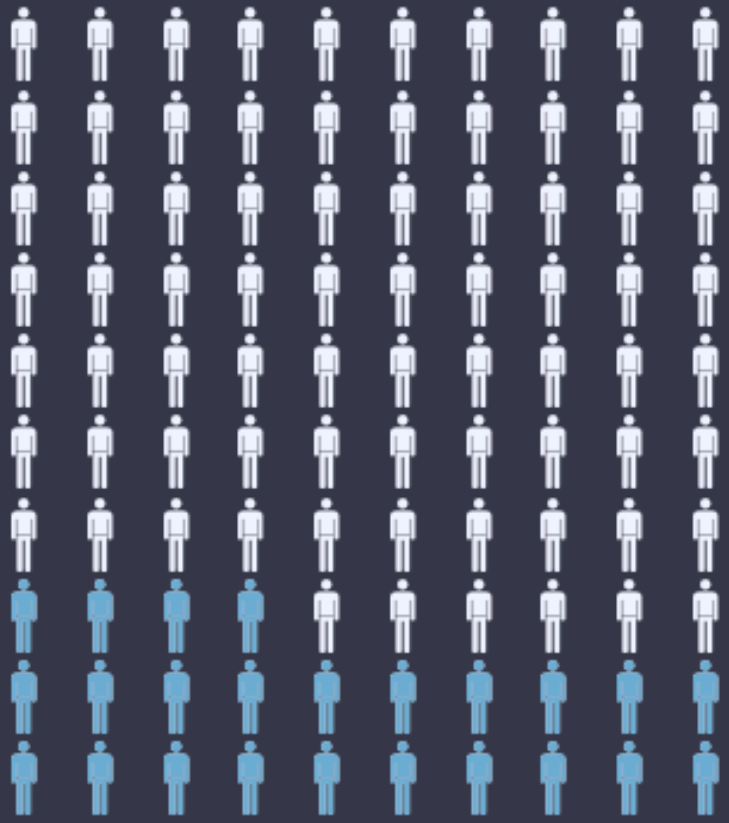


```
library(plotrix)

par(las=1, tck=-.005, mar=c(2,2,2,2), oma=c(0,1,0,1), family="Helvetica
Light", mfrow=c(1,2))

stackpoly(as.matrix(data), stack=T, ann=F)

mtext(side=3, line=3, "Stacked Line Chart", cex=2, adj=0)

stackpoly(t(as.matrix(data)), stack=T,  ann=F)
```




```
library(plotrix)

my.tab <- table(data$x,
data$y)

symbolbarplot(my.tab,
col=c("maroon3",
"darkolivegreen2"), beside=T,
symbbox=F, main="Symbol Bar
Chart", cex.main=2)
```

Unit Chart



24%

68%

```
par(family="Symbol Signs Basis set", mfrow=c(1,1), mar=c(1,1,1,1),
oma=c(1,1,1,1), mfrow=c(1,2))

perc1 <- 24
perc2 <- 68

s.row <- sort(rep(seq(.1, 1, by=.1), 10), decreasing=T)
s.col <- rep(seq(.1, 1, by=.1), 10)
s.mat <- cbind(s.row, s.col)

plot(s.mat[1:100,1], s.mat[1:100,2], pch="M", col=brewer.pal(5,
    "Blues")[1], cex=2.5, xlim=c(0,1), ylim=c(0,1), axes=F, ann=F)
points(s.mat[1:perc1, 2], s.mat[1:perc1, 1], pch="M", col=brewer.pal(5,
    "Blues")[3], cex=2.5)
mtext(side=1, line=3, paste(perc1, "%", sep=""), col=brewer.pal(5,
    "Blues")[3], cex=7, family="Helvetica Light")
mtext(side=3, line=3, "Unit Chart", col="white", cex=2,
family="Helvetica Light", adj=0)

plot(s.mat[1:100,1], s.mat[1:100,2], pch="F", col=brewer.pal(5,
    "PuRd")[1], cex=2.5, xlim=c(0,1), ylim=c(0,1), axes=F, ann=F)
points(s.mat[1:perc2, 2], s.mat[1:perc2, 1], pch="F", col=brewer.pal(5,
    "PuRd")[3], cex=2.5)
mtext(side=1, line=3, paste(perc2, "%", sep=""), col=brewer.pal(5,
    "PuRd")[3], cex=7, family="Helvetica Light")
```

Lab Exercise: Graph Re-Design

Use either the freedom house disaggregated dataset 2003-2014 (`FH_disaggregated_05_14.csv`) or your own data.

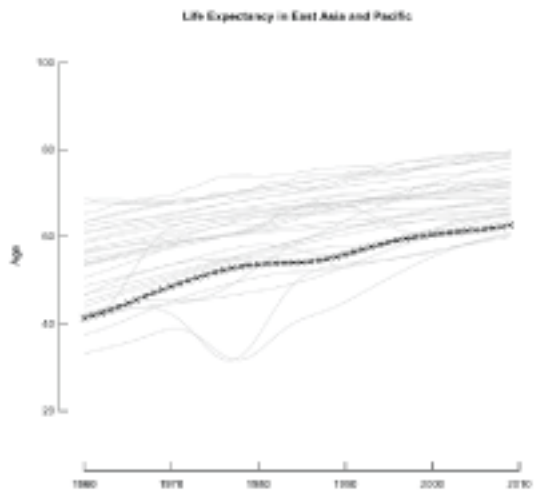
Mimic each of these six (more or less) famous graphic designs.

Then, re-design according to the principles we have discussed in the course.

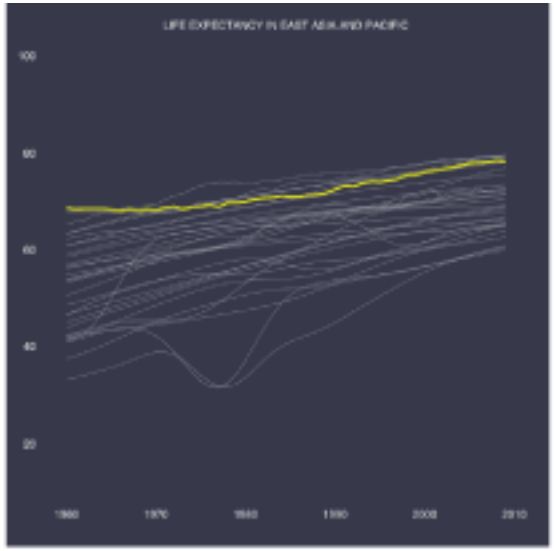
I am providing code to do this ("`Essex DataVis Code Redesign.R`") – but use it only when you are stuck yourself.

Finally, try to construct a "bullet graph" as invented by Stephen Few and a "slope graph" as invented by Edward Tufte.

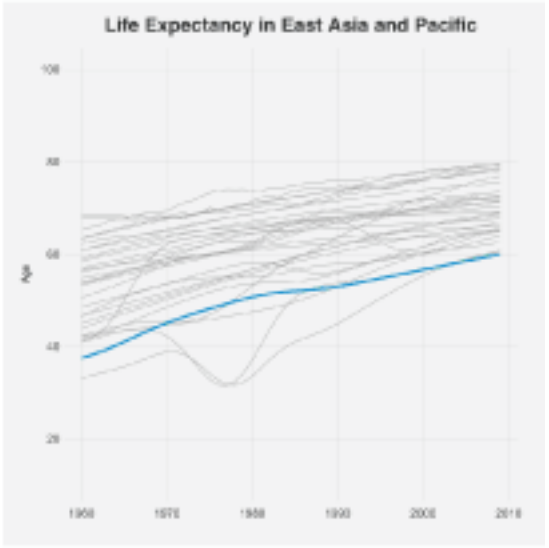
Tukey

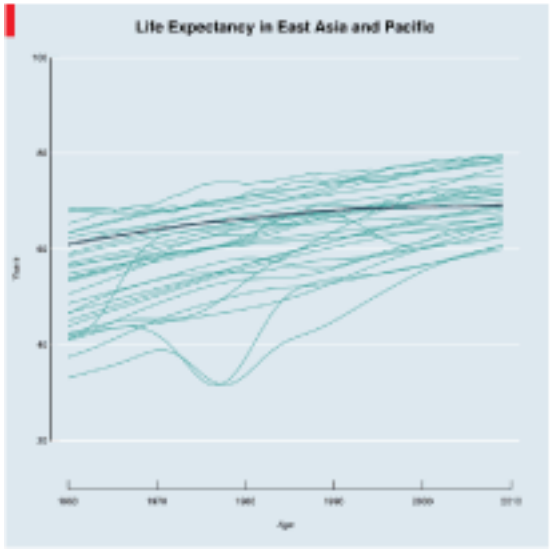


Feltron

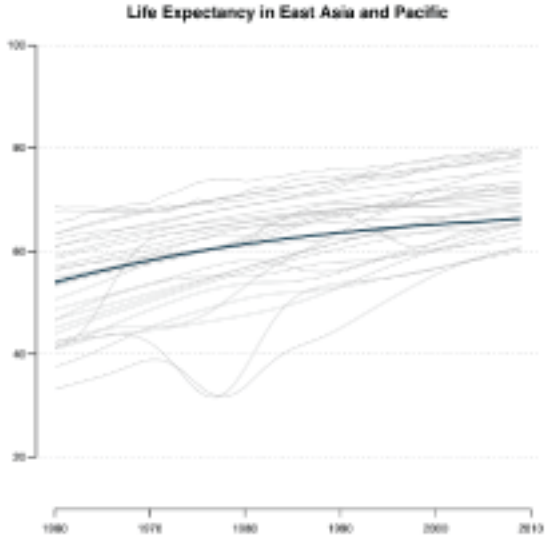


FiveThirtyEight

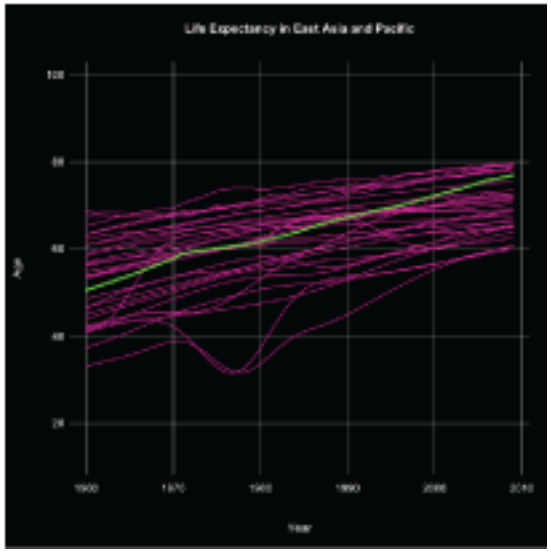




The Economist

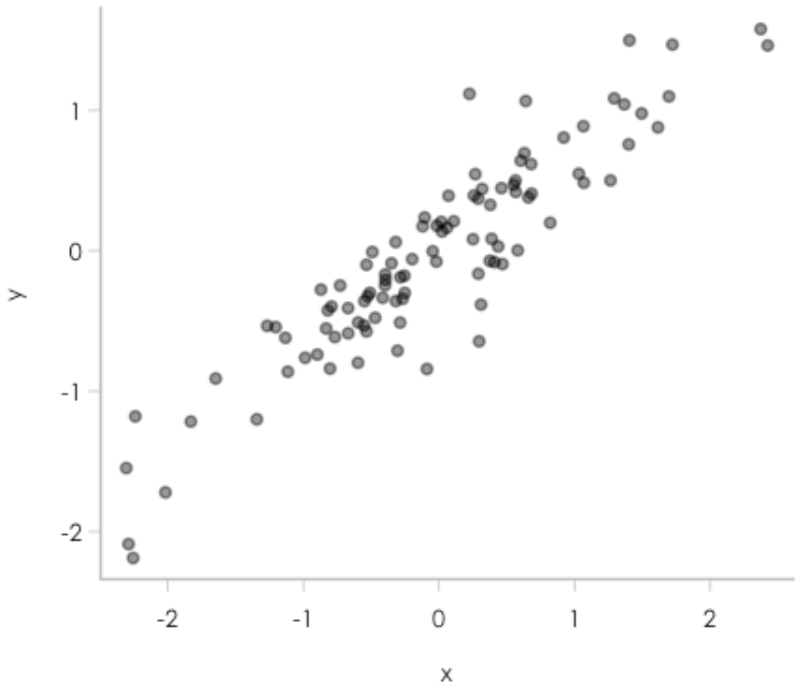


Newspaper

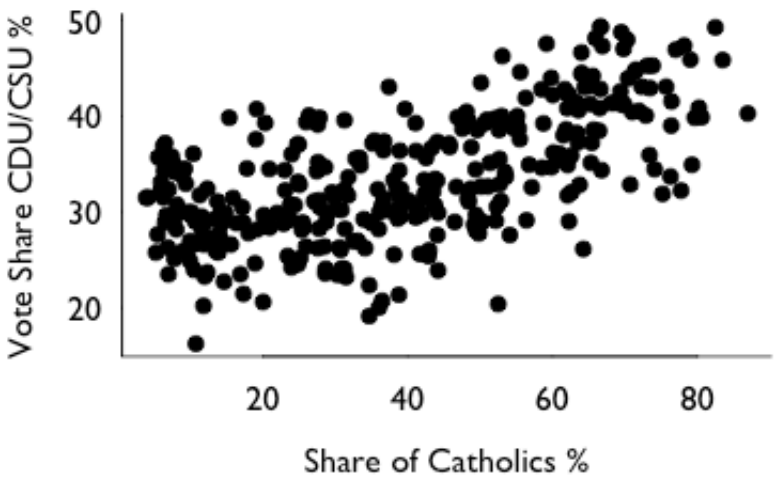


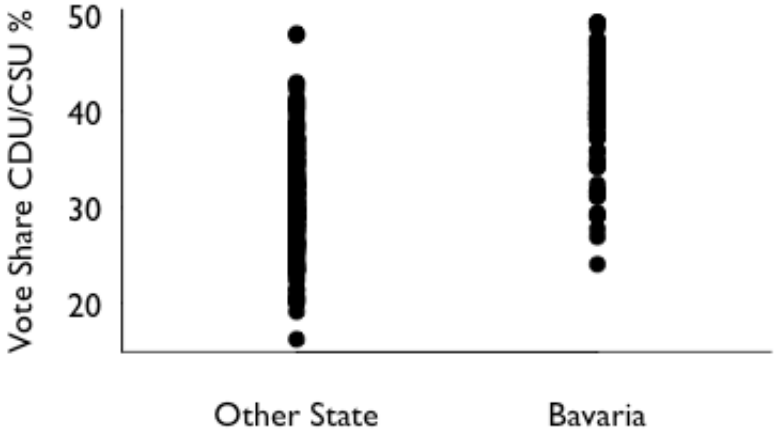
Old School Console

The Scatterplot



"Indeed, among all the forms of statistical graphics, the humble scatterplot may be considered the most versatile, polymorphic, and generally useful invention in the entire history of statistical graphics." (Friendly & Denis 2005)





Correlation?

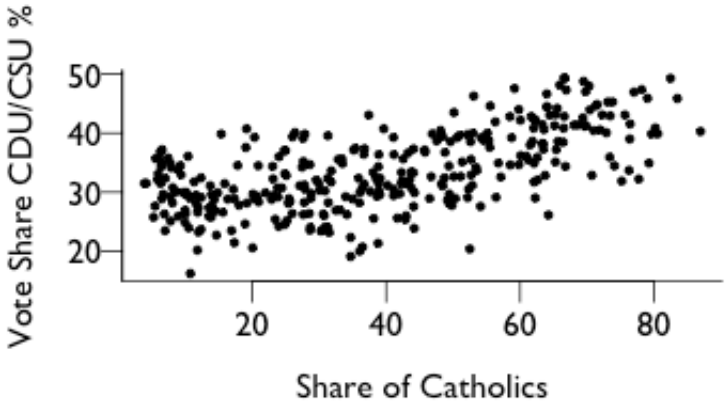
Functional Form?

Clusters?

Gaps?

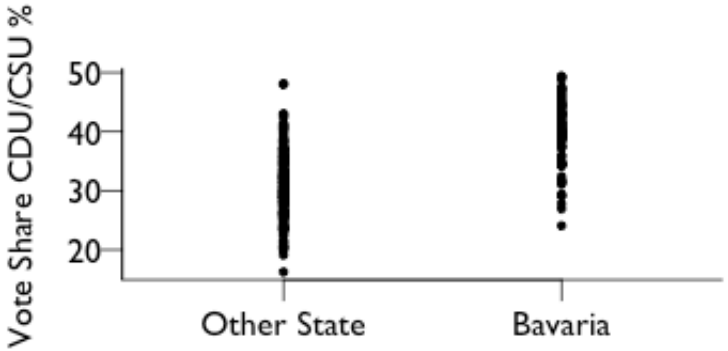
Outliers?

Reducing Overplotting

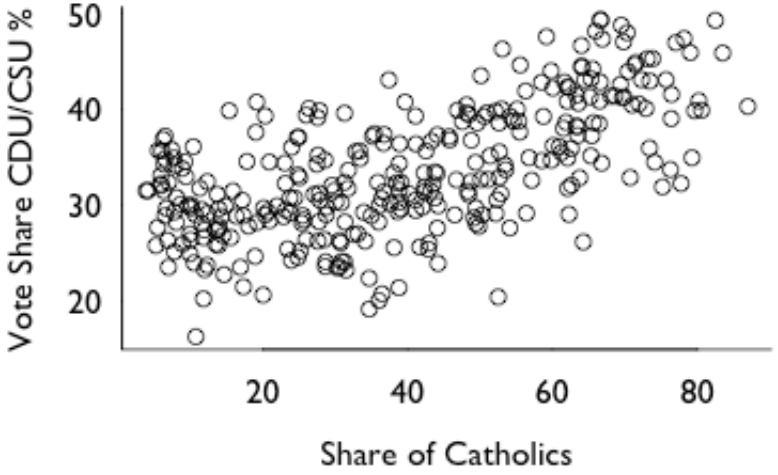


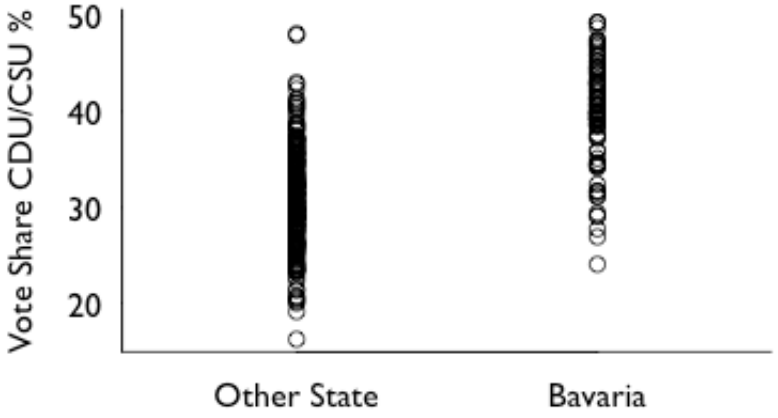
Decrease symbol size.

```
cex=.5
```



Reducing Overplotting

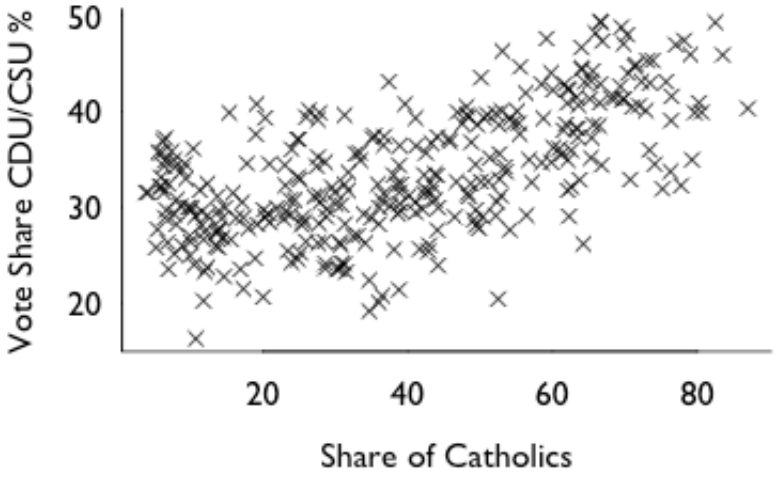


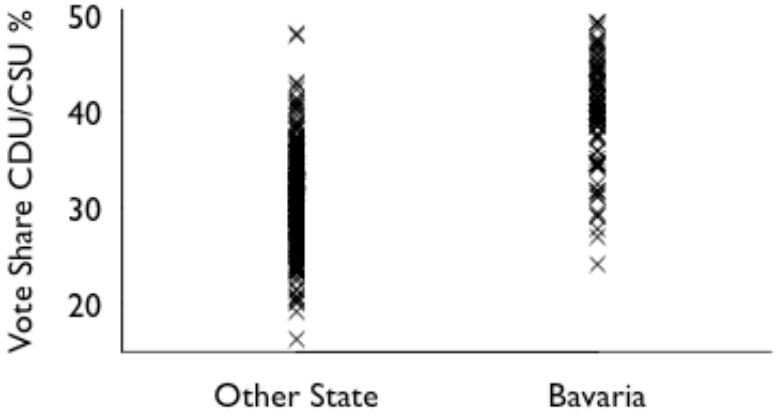


Remove fill color.

```
pch=21
```

Reducing Overplotting

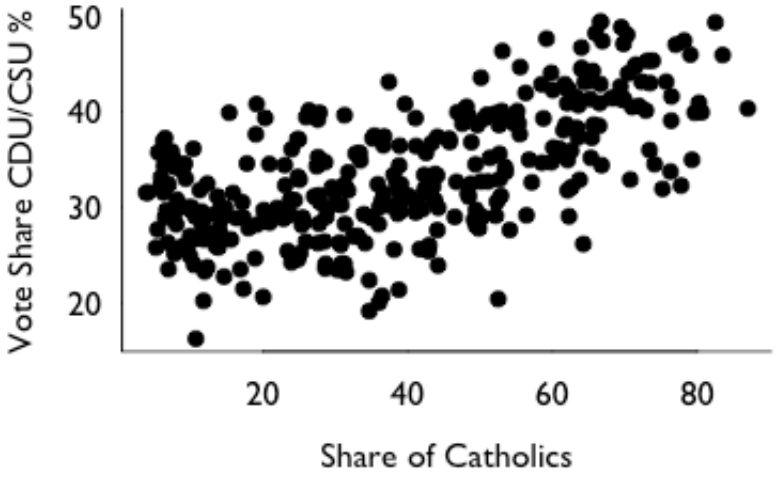




Choose other symbols.

```
pch=4
```

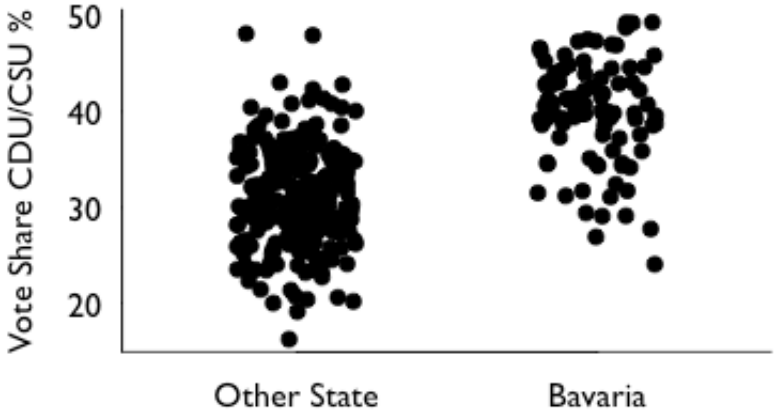
Reducing Overplotting



Jittering.

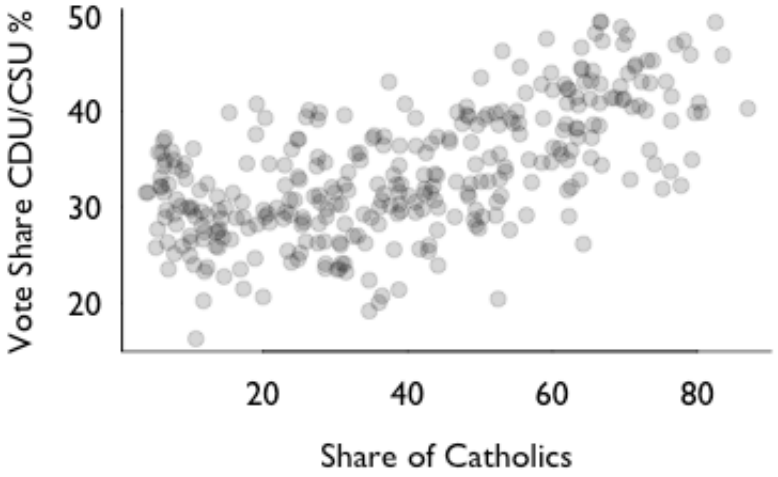
Adds random noise to data values.

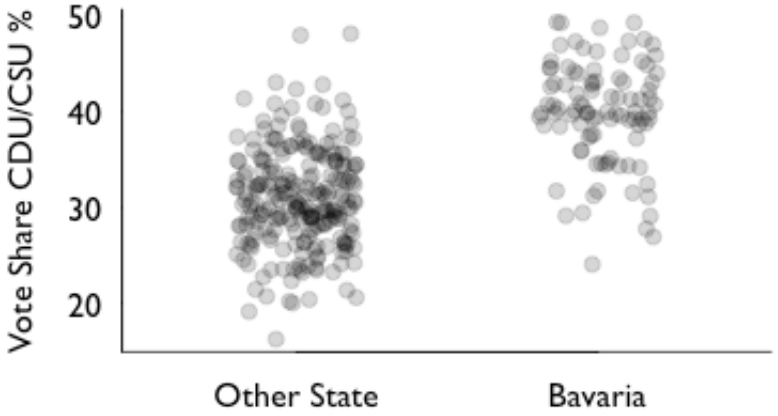
Doesn't make sense for continous variables!




```
points(jitter(daten05$bayern),
     daten05$cdu, pch=19)
```

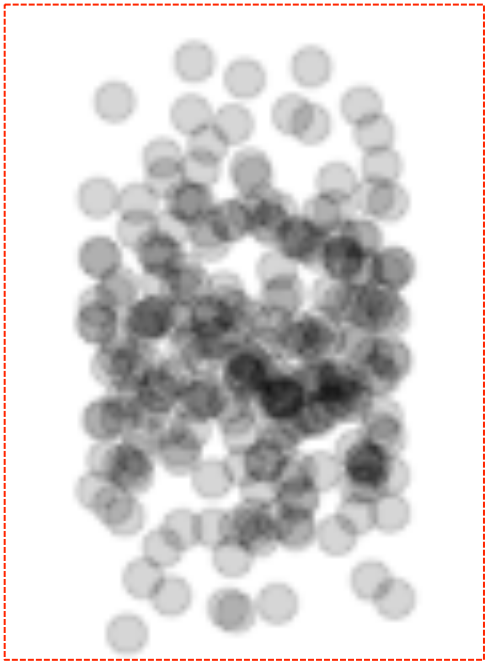
Reducing Overplotting





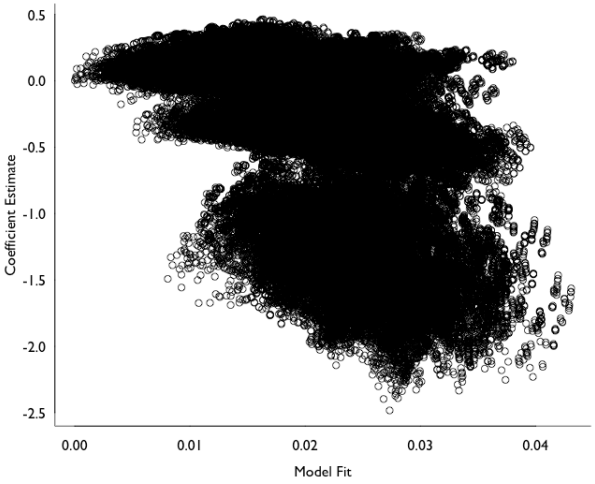
Alpha Blending.

Makes plotting symbols transparent.

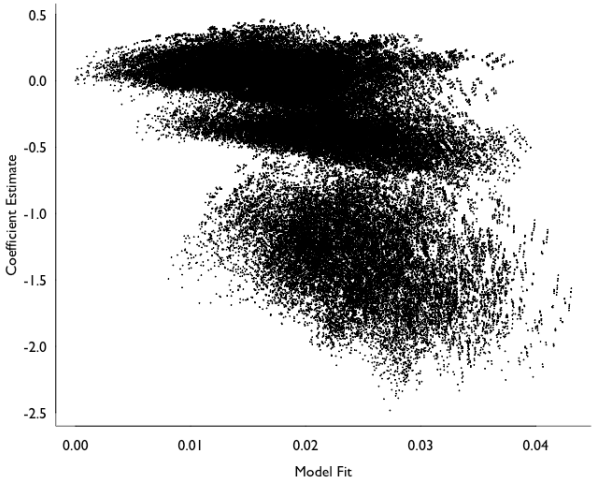


```
points(jitter(daten05$bayern),
daten05$cdu, pch=19, col=rgb(00,
00, 00, 50, max=255))
```

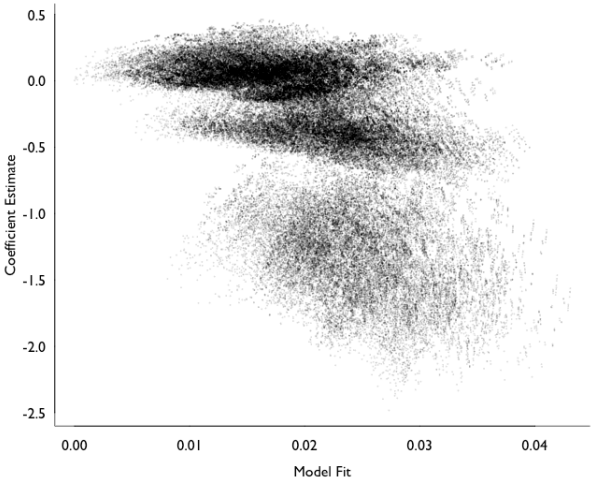
Same with „Big Data“ Example



No fill color...

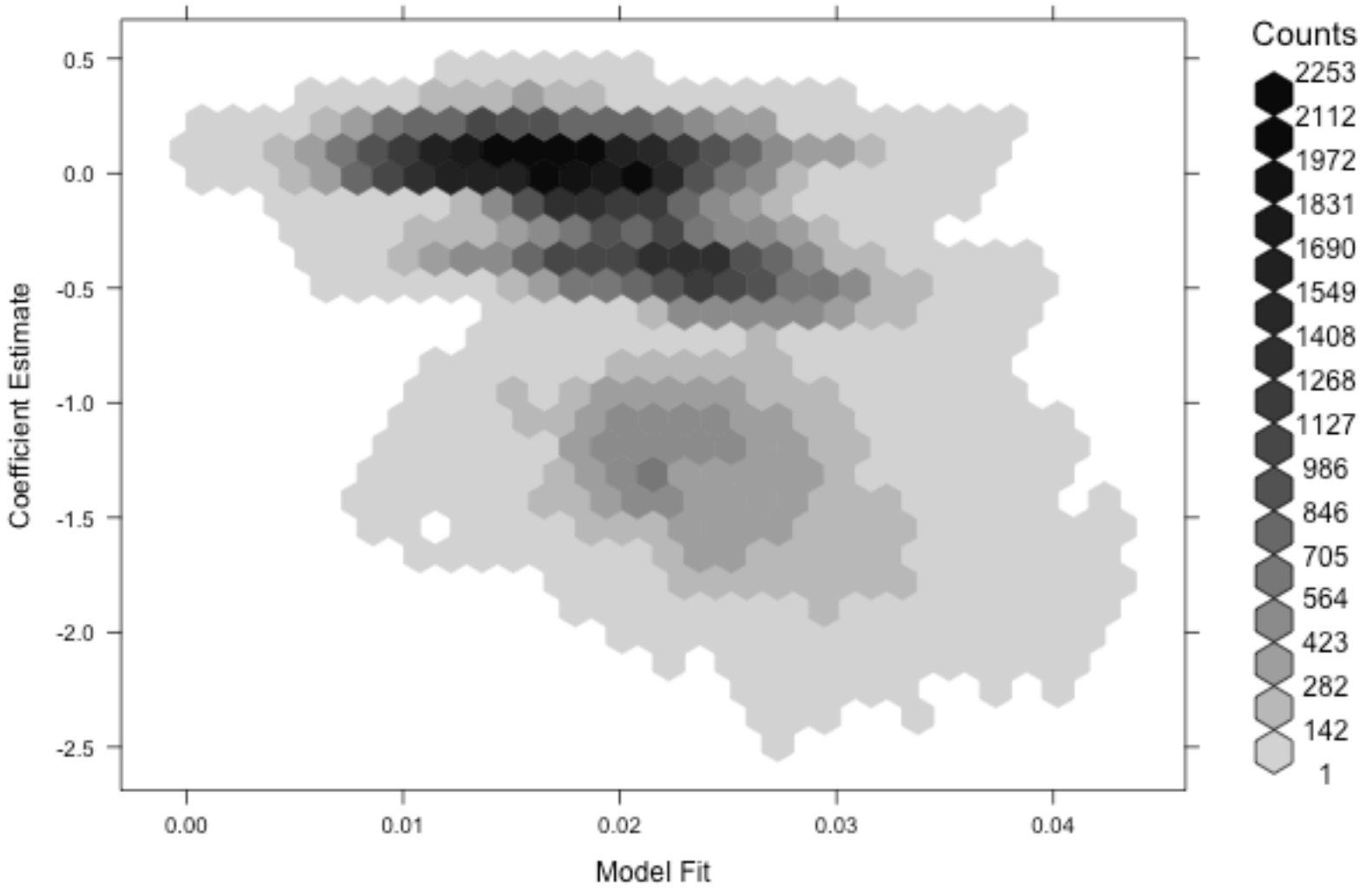


Smaller symbols...



Small symbols + Alpha blending

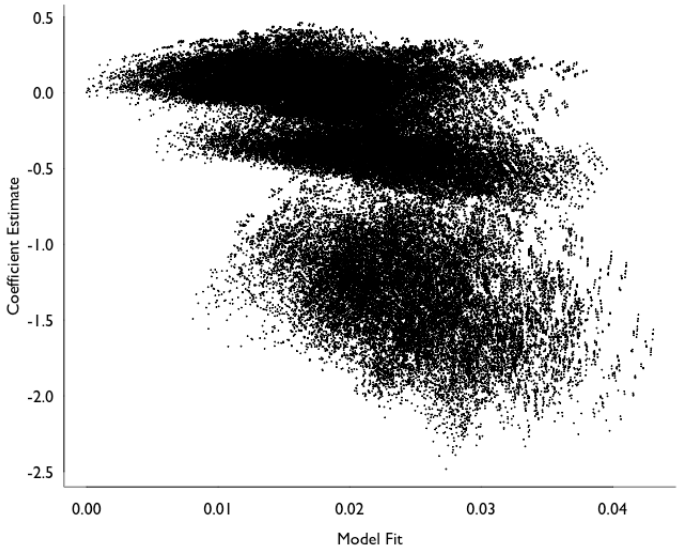
Density Encoding



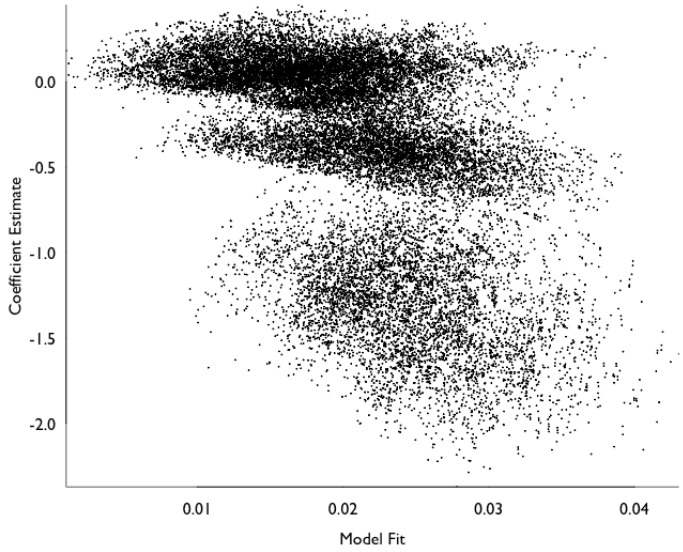

```
hexbinplot(DATA$Est.z[DATA$variable=="milper"] ~
DATA$dev_diff[DATA$variable=="milper"], xlab="Model Fit",
ylab="Coefficient Estimate")
```

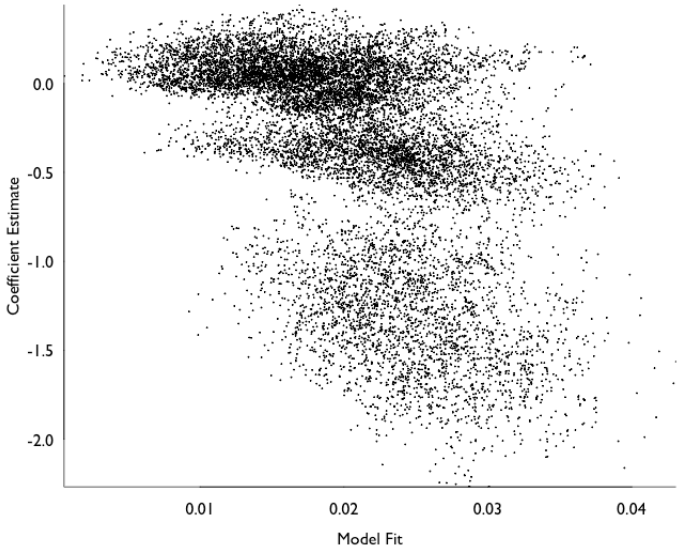
Sampling

262143 Points

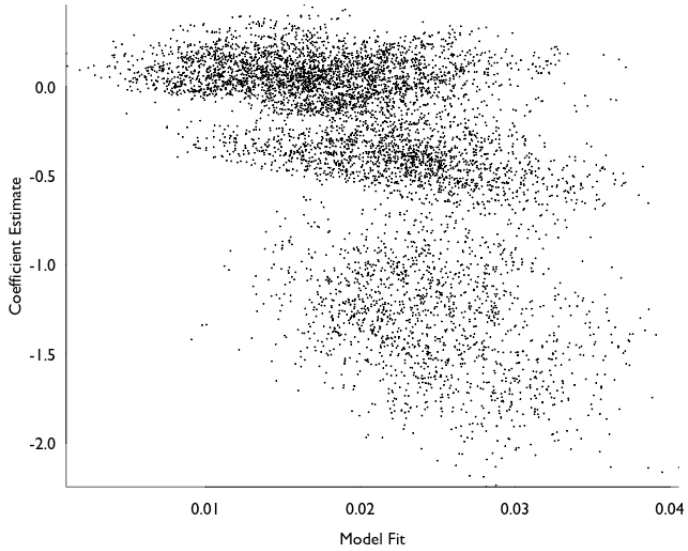


20% Sample





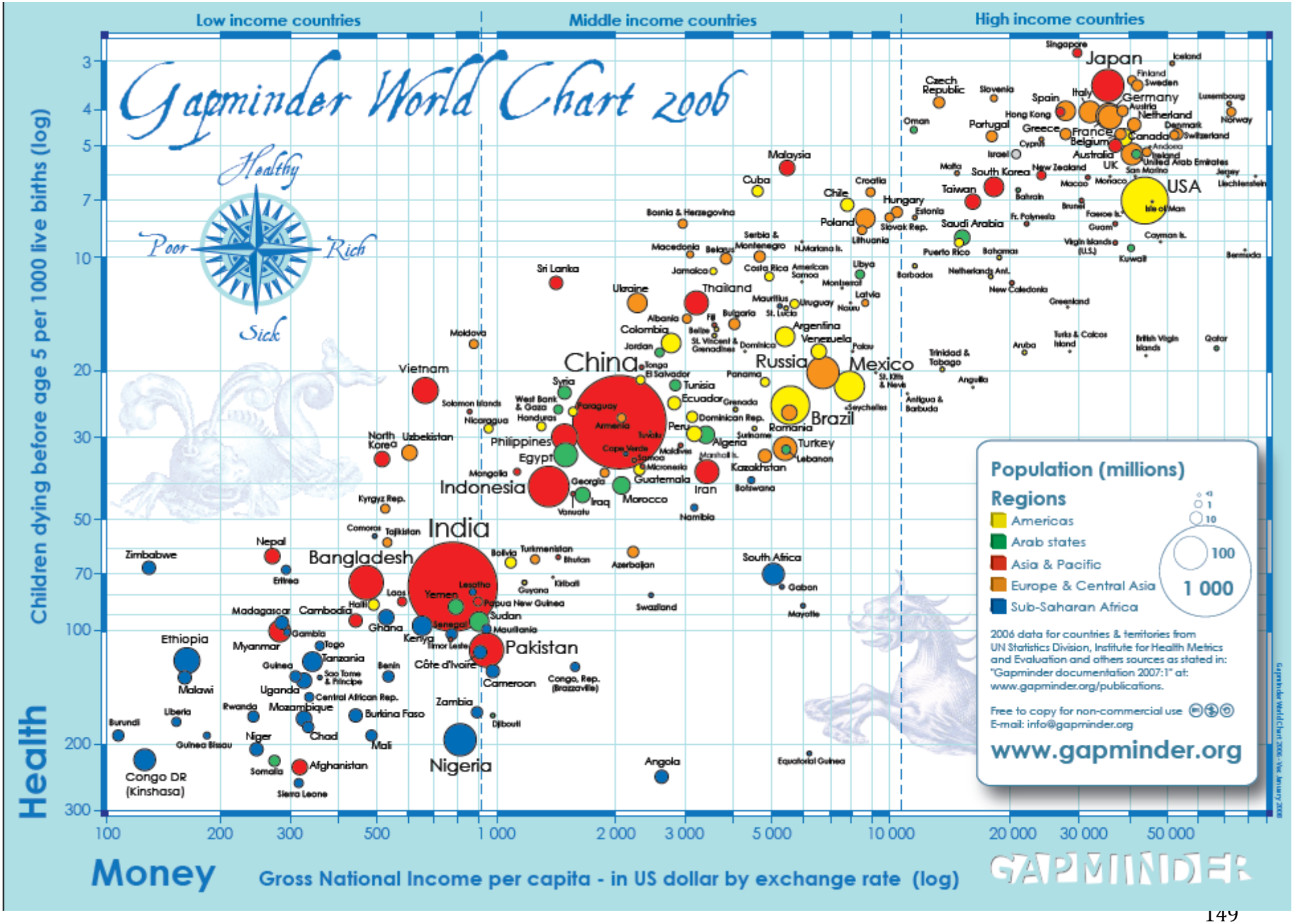
10% Sample



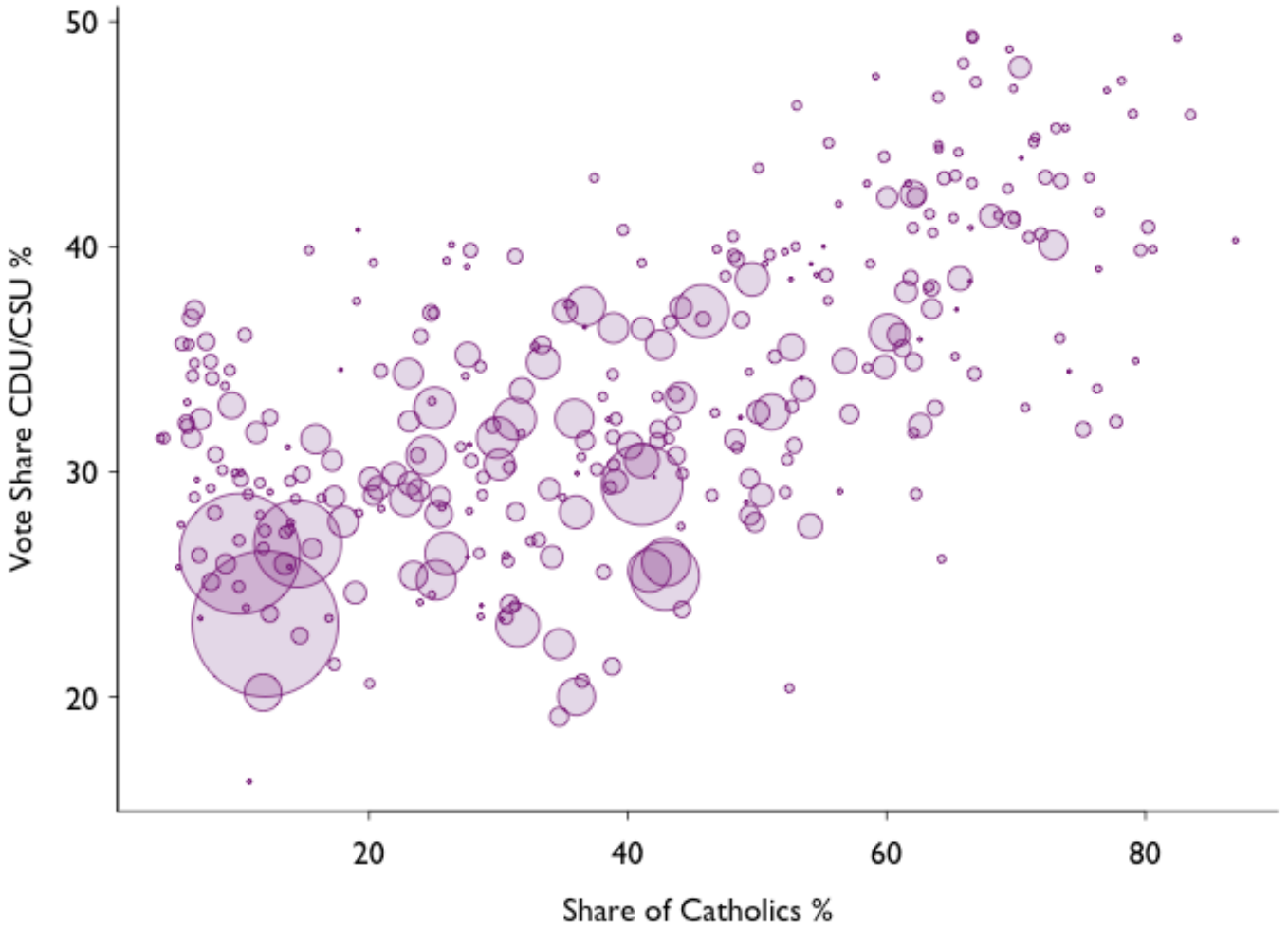
5% Sample

Sampling

```
samp.20 <- sample(length(DATA$Est.z[DATA$variable=="milper"]),
round(length(DATA$Est.z[DATA$variable=="milper"])*.2))


plot(DATA$dev_diff[DATA$variable=="milper"][samp.20],
    DATA$Est.z[DATA$variable=="milper"][samp.20], pch=19, cex=.1, axes=F,
    xlab="Model Fit", ylab="Coefficient Estimate")
axis(1)
axis(2, las=1)
box(bty="l")
```

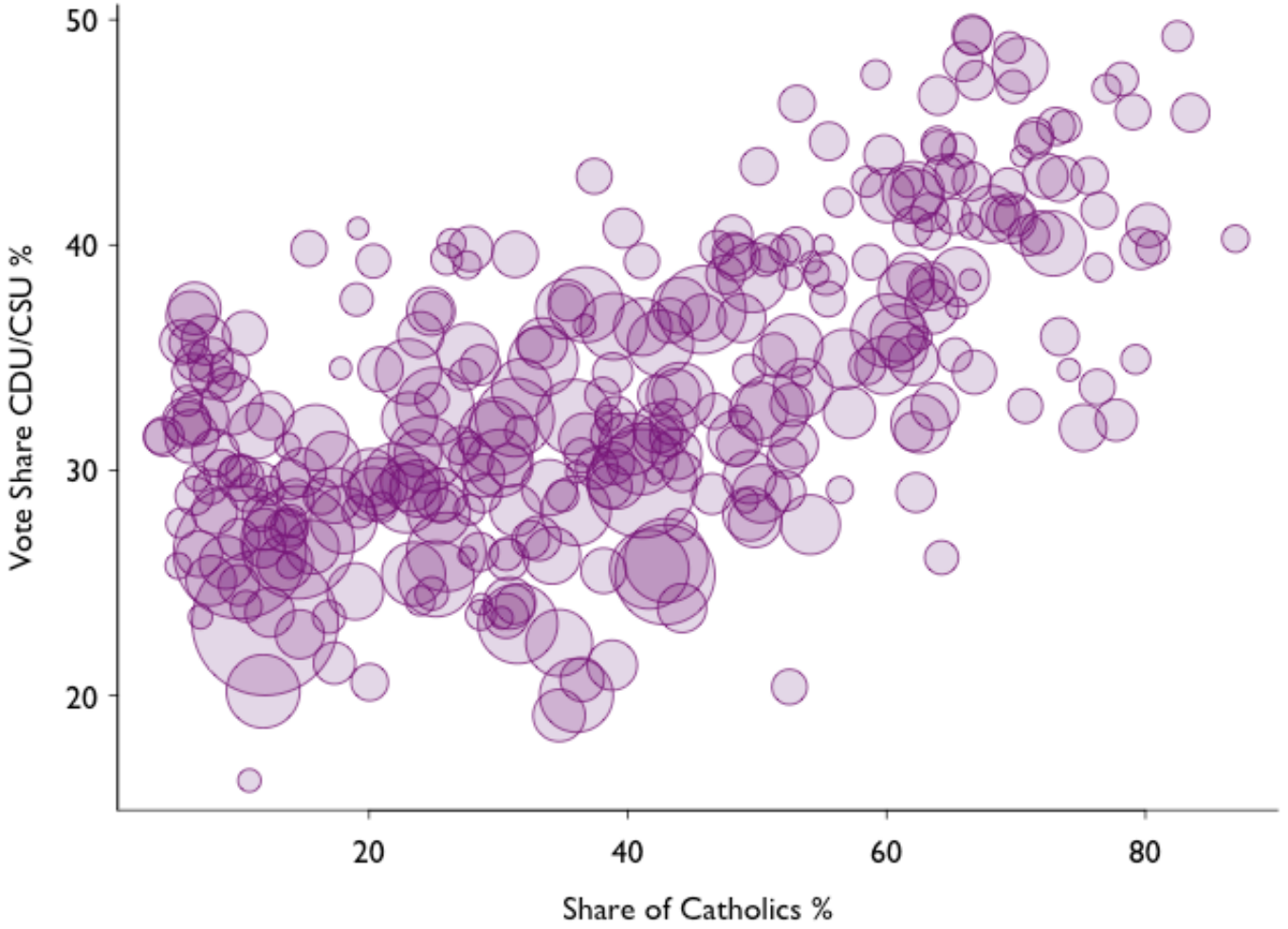

Low income countries
Middle income countries
High income countries
Gapminder World Chart 2006
Healthy
Poor
Rich
Sick
Children dying before age 5 per 1000 live births (log)
Health
Money
Gross National Income per capita - in US dollar by exchange rate (log)
Population (millions)
Regions
Americas
Arab states
Asia & Pacific
Europe & Central Asia
Sub-Saharan Africa
2006 data for countries & territories from UN Statistics Division, Institute for Health Metrics and Evaluation and others sources as stated in: "Gapminder documentation 2007:1" at: www.gapminder.org/publications.
Free to copy for non-commercial use
E-mail: info@gapminder.org
www.gapminder.org
GAPMINDER

The Bubble Plot




```
symbols(daten05$kath, daten05$cdu, circles=daten05$n, inches=0.5,
          bg=rgb(120, 00, 120, 50, max=255), fg=rgb(120, 00, 120,
           max=255), add=T)
```

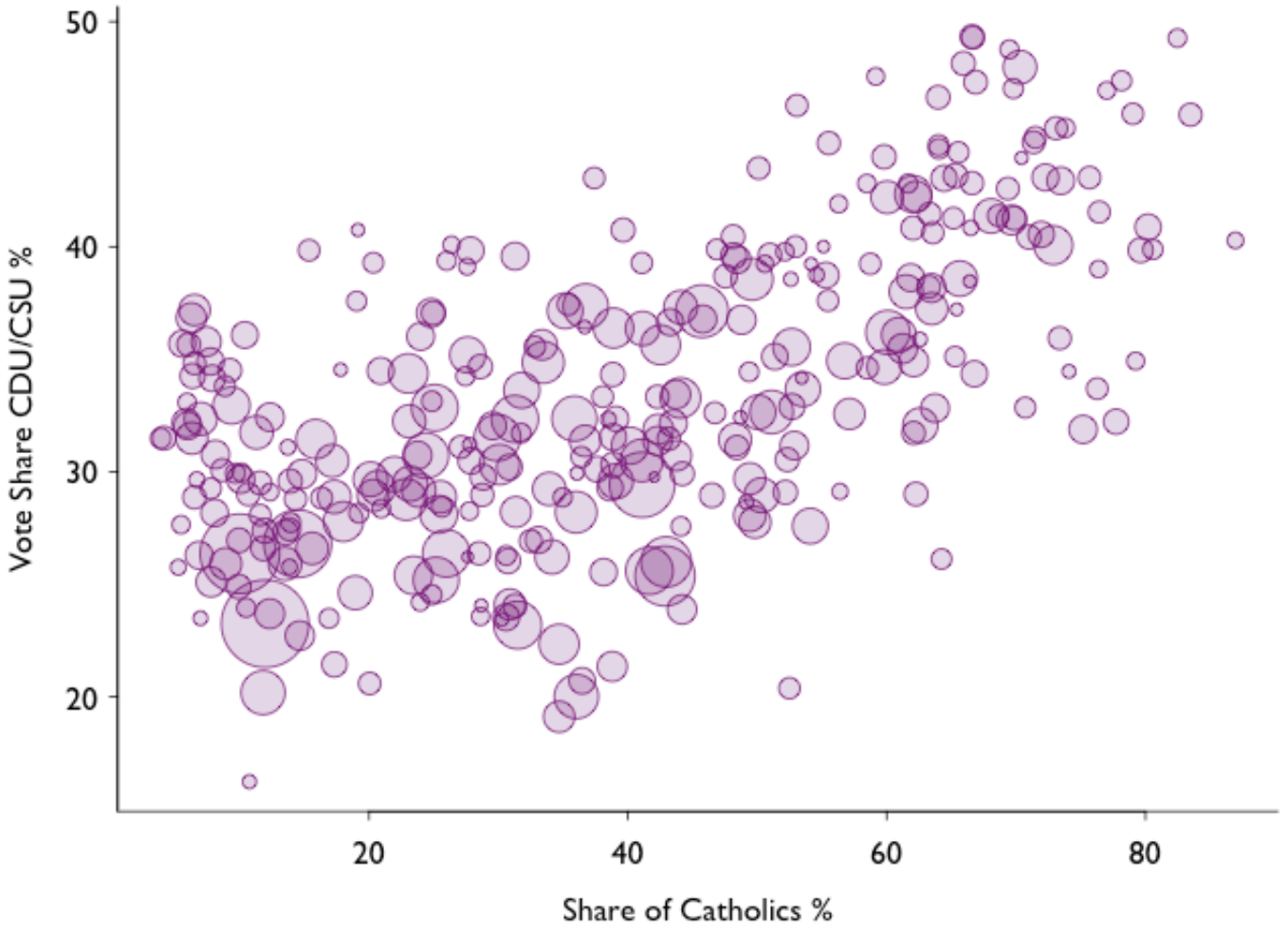
The Bubble Plot



ATTENTION: you need to size the bubbles proportional to area (not radius)!

$$Area = \pi r^2 \quad \text{and} \quad r = \sqrt{Area/\pi}$$

The Bubble Plot




```
symbols(daten05$kath, daten05$cdu, circles=sqrt(daten05$n/pi),
inches=0.3, bg=rgb(120, 00, 120, 50, max=255), fg=rgb(120, 00, 120,
max=255), add=T)
```

Adding Text Labels (This looks pretty Rosling-style to me!)


```
text(data$pr_2014, data$cl_2014, data$Country, cex=.6, col=rgb(0, 0, 0,
100, max=255), pos=2)
```

Visualizing Contingency Tables

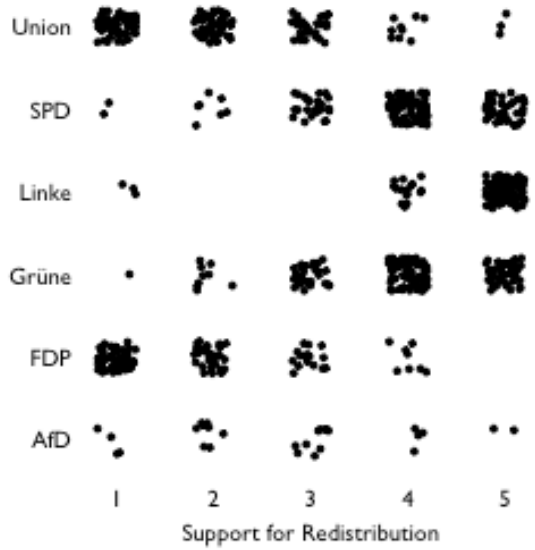
	1	2	3	4	5
Union	54	48	28	9	3
SPD	2	7	24	80	44
Linke	3	0	0	16	104
Grüne	1	9	31	76	44
FDP	57	32	16	8	0
AfD	4	8	10	4	2

Support for Redistribution

Visualizing Contingency Tables

	1	2	3	4	5
Union	54	48	28	9	3
SPD	2	7	24	80	44
Linke	3	0	0	16	104
Grüne	1	9	31	76	44
FDP	57	32	16	8	0
AfD	4	8	10	4	2

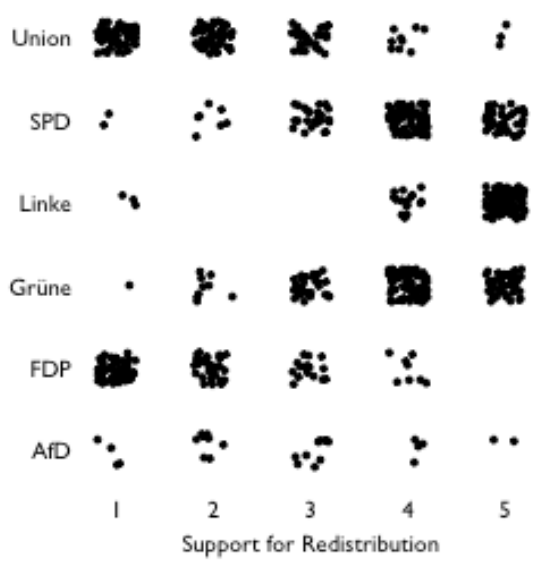
Support for Redistribution

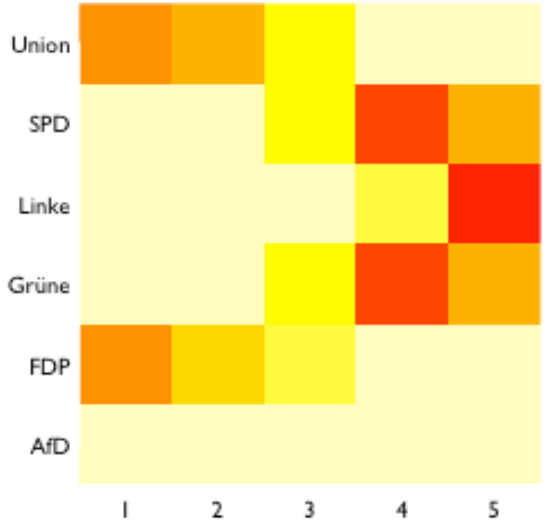


Visualizing Contingency Tables

	1	2	3	4	5
Union	54	48	28	9	3
SPD	2	7	24	80	44
Linke	3	0	0	16	104
Grüne	1	9	31	76	44
FDP	57	32	16	8	0
AfD	4	8	10	4	2

Support for Redistribution

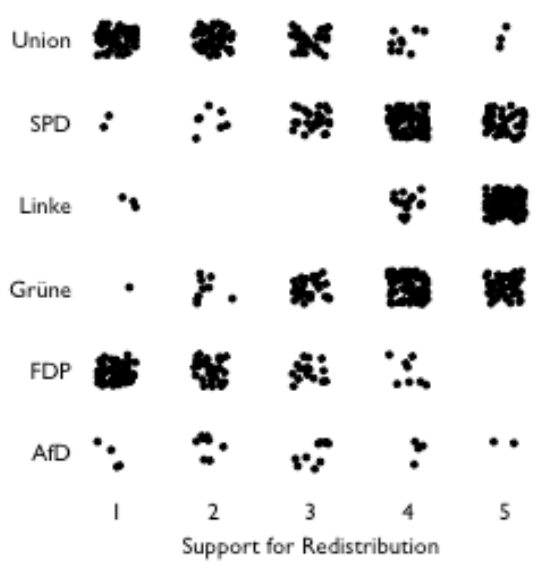


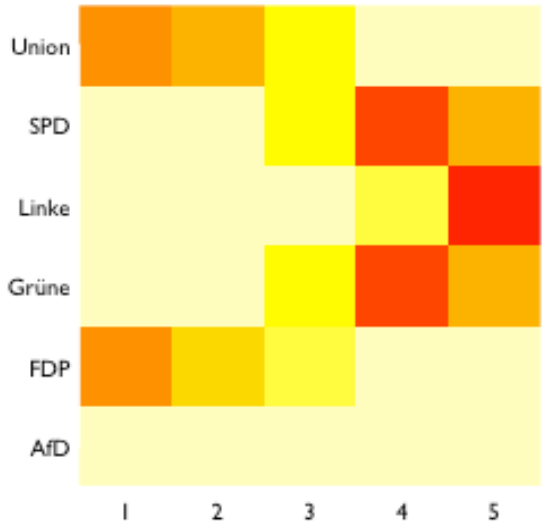


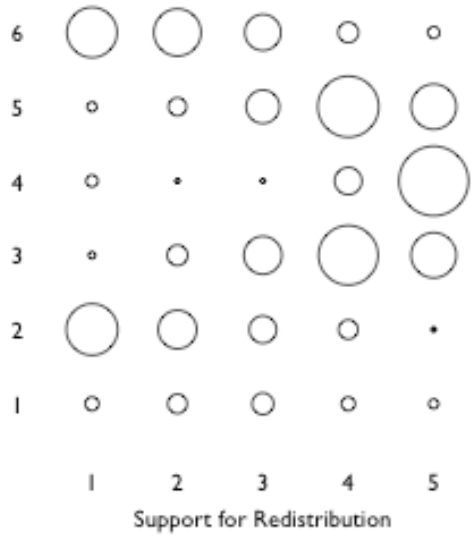
Visualizing Contingency Tables

	1	2	3	4	5
Union	54	48	28	9	3
SPD	2	7	24	80	44
Linke	3	0	0	16	104
Grüne	1	9	31	76	44
FDP	57	32	16	8	0
AfD	4	8	10	4	2

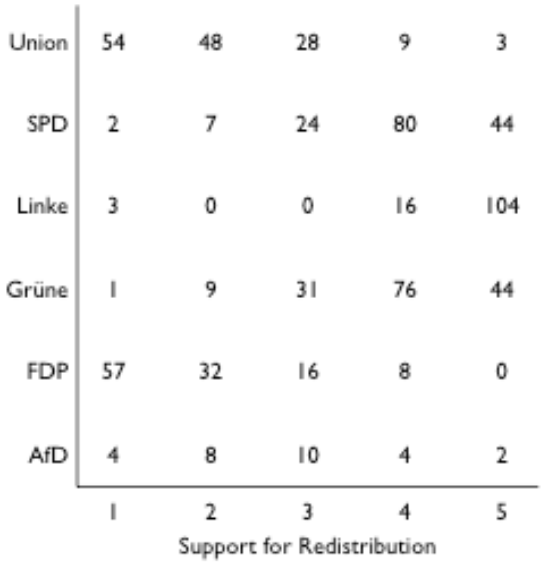
Support for Redistribution

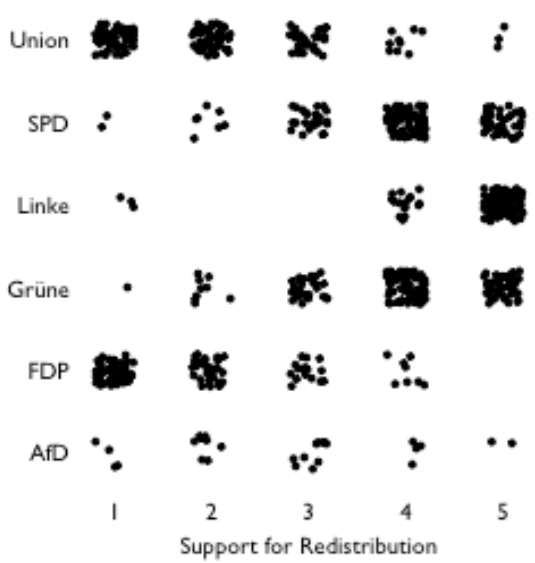


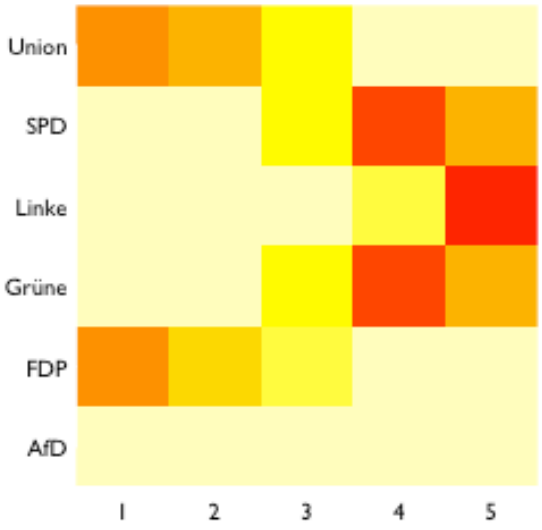


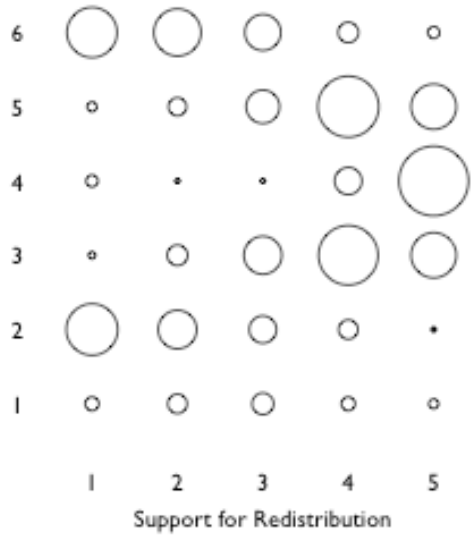


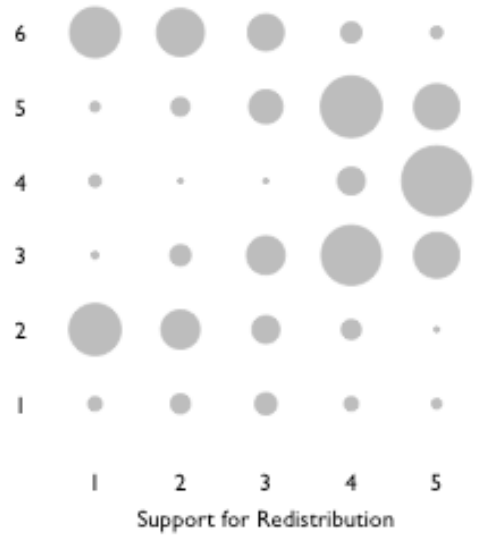
Visualizing Contingency Tables



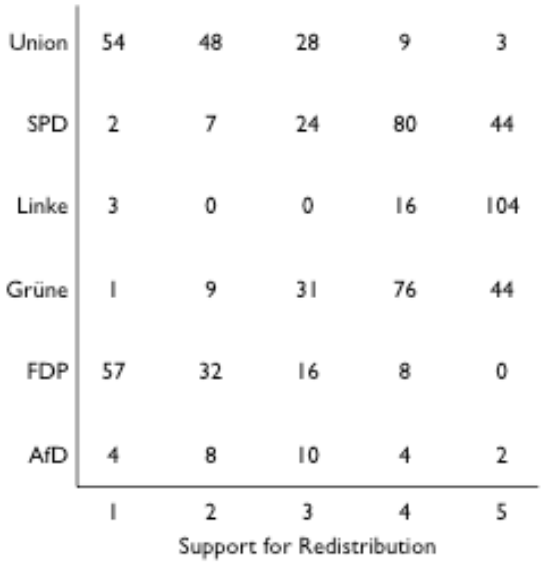


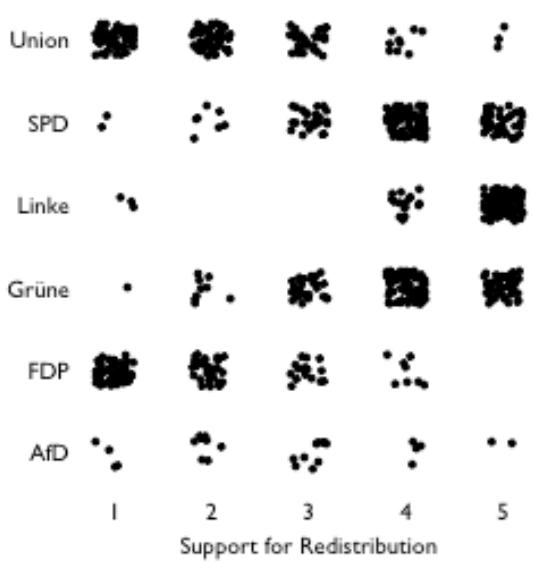


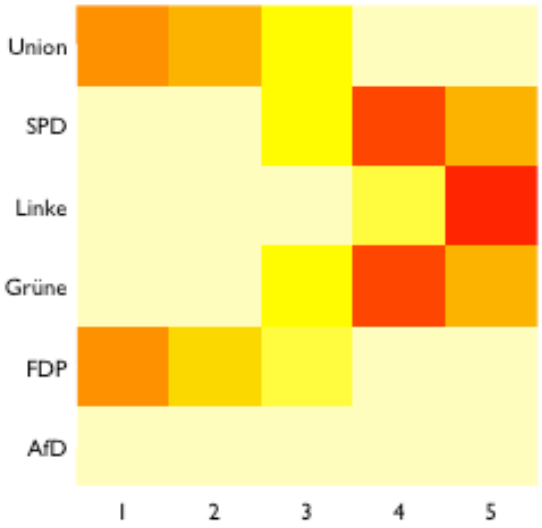


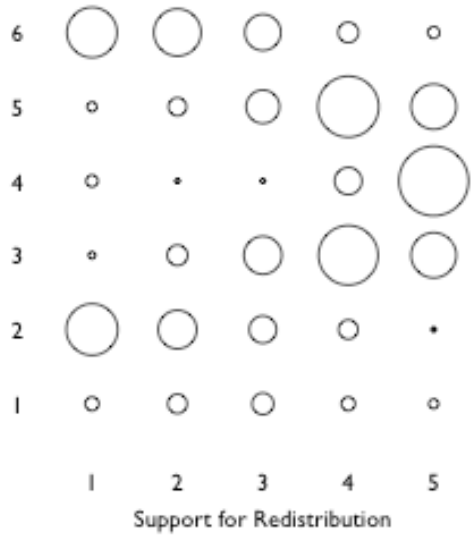


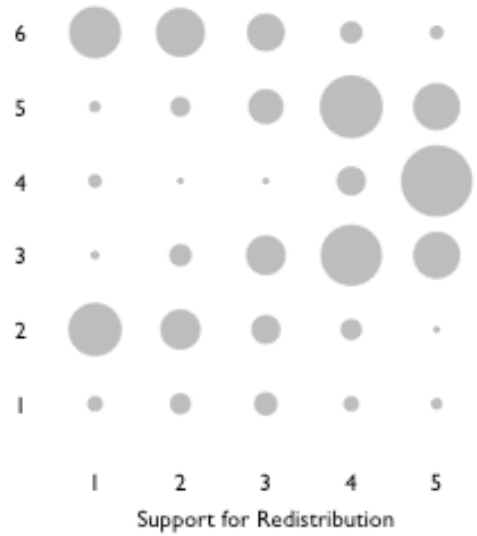
Visualizing Contingency Tables

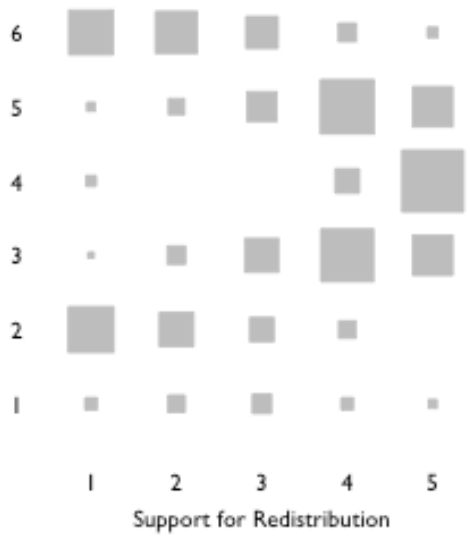












Visualizing Contingency Tables

Contingency Table

```
plot(jitter(data$redist), jitter(data$party.2), pch="", cex=.6,
xlab="Support for Redistribution", ylab="", axes=F)
axis(1, col="white", col.axis="black")
axis(2, at=c(1:6), labels=names(table(data$party)), col="white",
col.axis="black")
for(i in 1:6){
  text(c(1:5), rep(i, 5),  my.tab[i,])
}
box(bty="l")
```

Simple Scatter Plot of Categorical Variables

```
plot(jitter(data$redist), jitter(data$party.2), pch=19, cex=.6,
xlab="Support for Redistribution", ylab="", axes=F)
axis(1, col="white", col.axis="black")
axis(2, at=c(1:6), labels=names(table(data$party)), col="white",
col.axis="black")
```

Visualizing Contingency Tables

Simple Heat Map of Contingency Table

```
my.tab <- table(data$party.2, data$redist)
my.scale <- function(x, d){(x-1)/(d-1)}

image(t(my.tab), axes=F, col=rev(heat.colors(10)))
axis(2, at=my.scale(1:6, 6), label=names(table(data$party)),
col="white", col.axis="black", xlab="Support for Redistribution")
axis(1, at=my.scale(1:5, 5), label=c(1:5), las=1, col="white",
col.axis="black")
```

Visualizing Contingency Tables

Bubbles

```
par(mgp=c(1.5, .3, 0))
plot(0, 0, pch="", xlim=c(0.5, 5.5), ylim=c(0.5, 6.5), axes=F,
xlab="Support for Redistribution", ylab="")
for(i in 1:dim(my.tab)[1]){
  symbols(c(1:dim(my.tab)[2]), rep(i, dim(my.tab)[2]),
circle=sqrt(my.tab[i,]/200/pi), add=T, inches=F, fg="black")
}
```

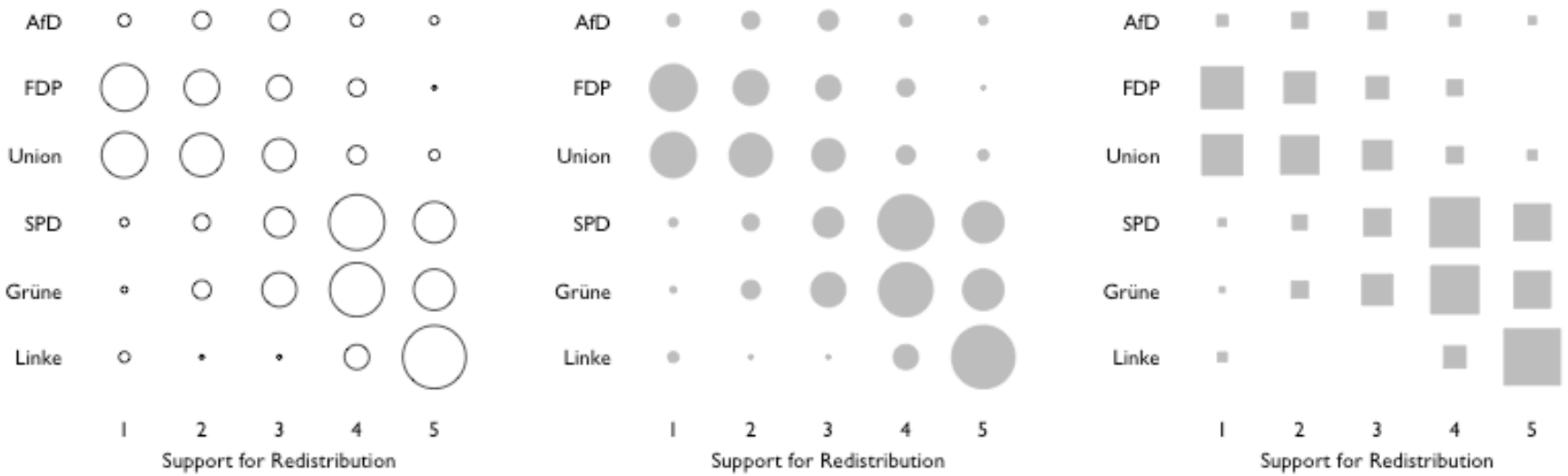
Filled Bubbles

```
symbols(c(1:dim(my.tab)[2]), rep(i, dim(my.tab)[2]),
circle=sqrt(my.tab[i,]/200/pi), add=T, inches=F, bg="grey", fg="grey")
```

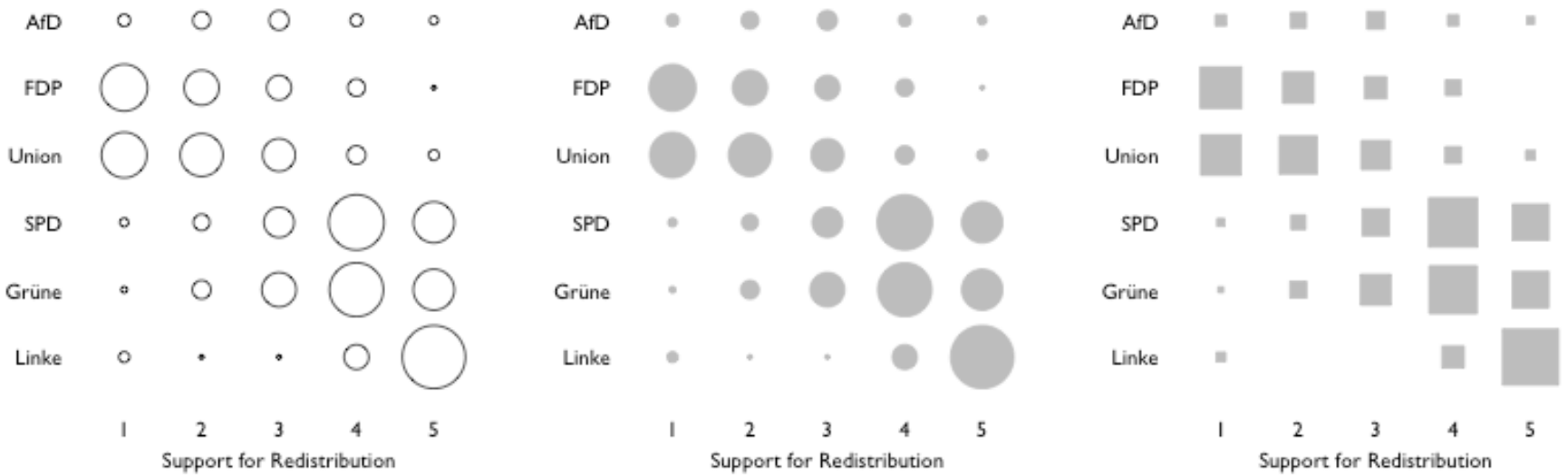
Squares

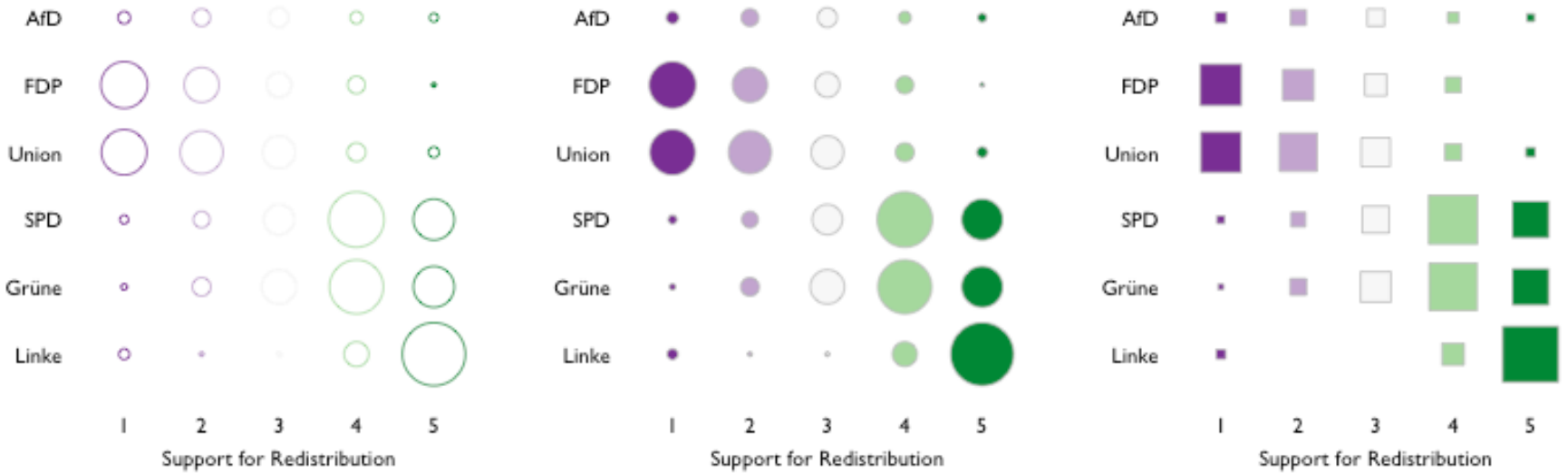
```
symbols(c(1:dim(my.tab)[2]), rep(i, dim(my.tab)[2]),
square=sqrt(my.tab[i,]/200), add=T, inches=F, bg="grey", fg="grey")
```

Visualizing Contingency Tables: Ordering



Visualizing Contingency Tables: Ordering + Color





Visualizing Contingency Tables: Ordering + Color

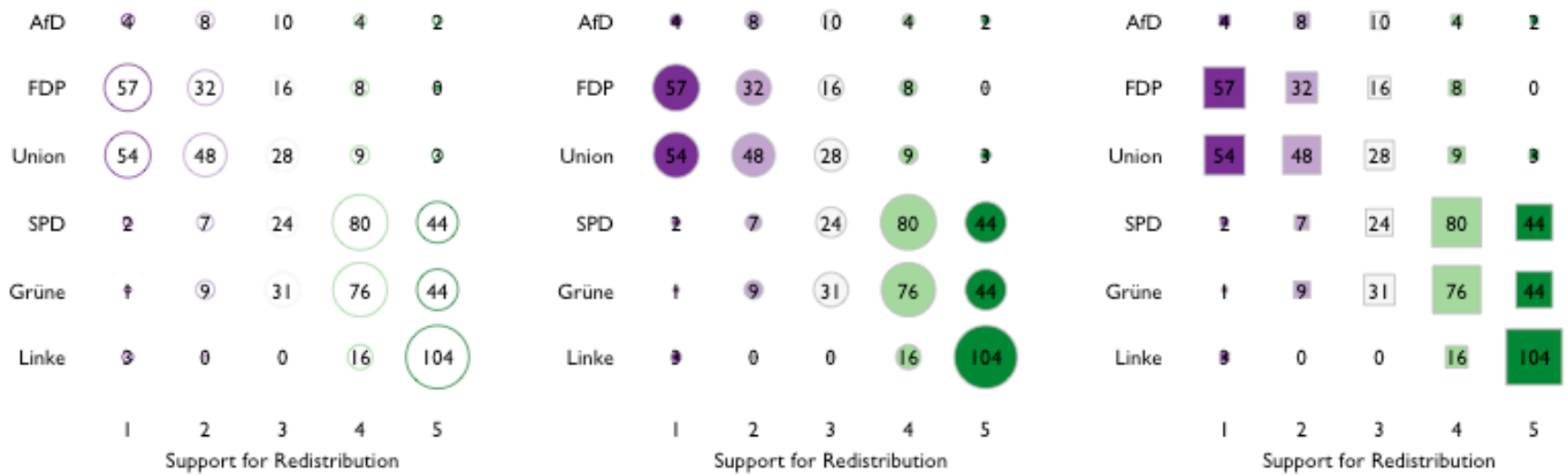
Re-order Levels of Categorical Variable (Logically)

```
data$party.ord <- factor(data$party, levels=c("Linke", "Grüne", "SPD",
"Union", "FDP", "AfD"))
my.tab <- table(data$party.ord, data$redist)
rownames(my.tab) <- names(table(data$party.ord))
```

Color Bubbles

```
plot(0, 0, pch="", xlim=c(0.5, 5.5), ylim=c(0.5, 6.5), axes=F,
xlab="Support for Redistribution", ylab="")
for(i in 1:dim(my.tab)[1]){
  symbols(c(1:dim(my.tab)[2]), rep(i,
dim(my.tab)[2]),circle=sqrt(my.tab[i,]/200/pi), add=T, inches=F,
fg=brewer.pal(5, "PRGn"))}
axis(1, col="white", col.axis="black")
axis(2, at=c(1:6), label=rownames(my.tab), las=1, col.axis="black",
col="white")
```

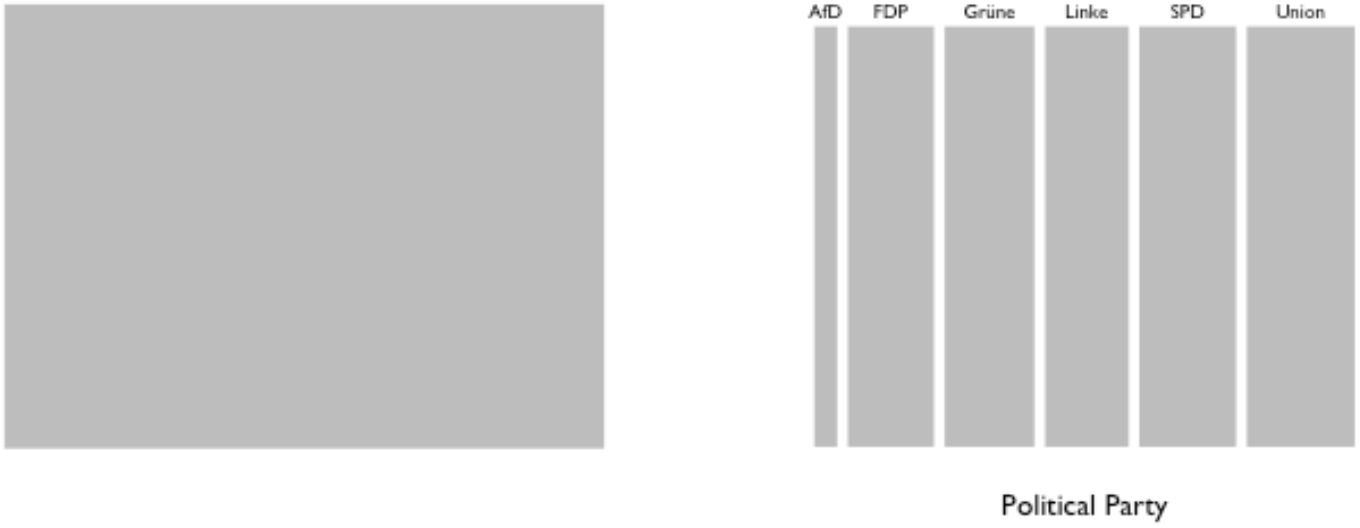
Visualizing Contingency Tables: Ordering + Color + Numbers



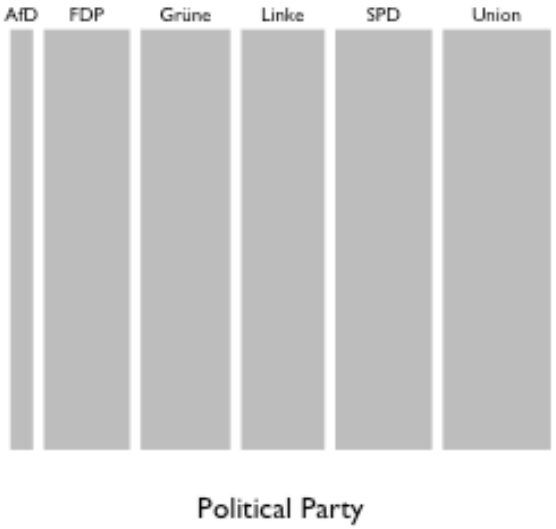
Add Numbers to Plot

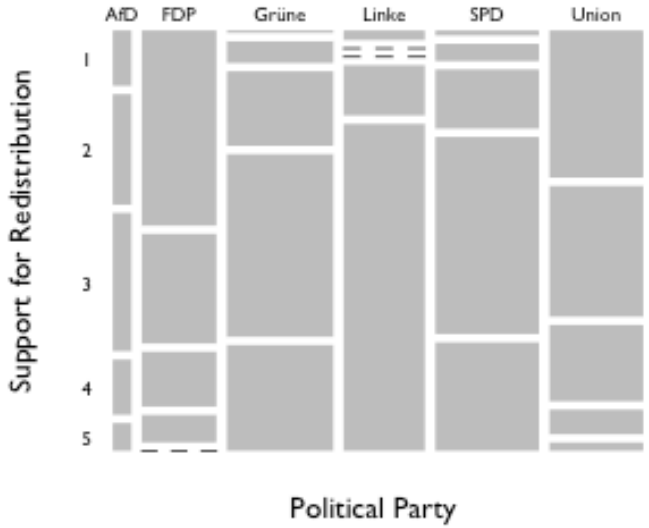
```
for(i in 1:6){
  text(c(1:5), rep(i, 5),  my.tab[i,])
}
```

The Mosaic Plot

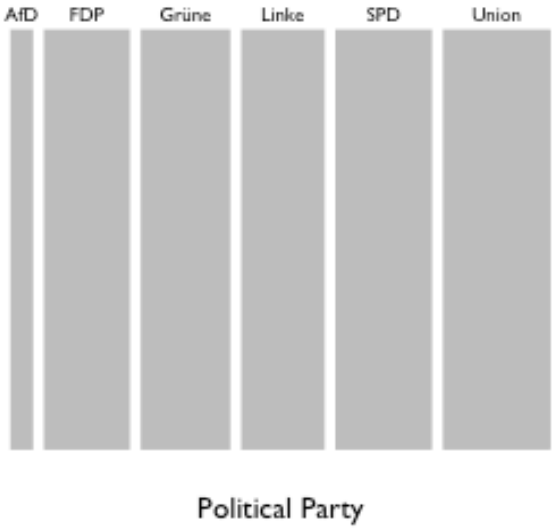


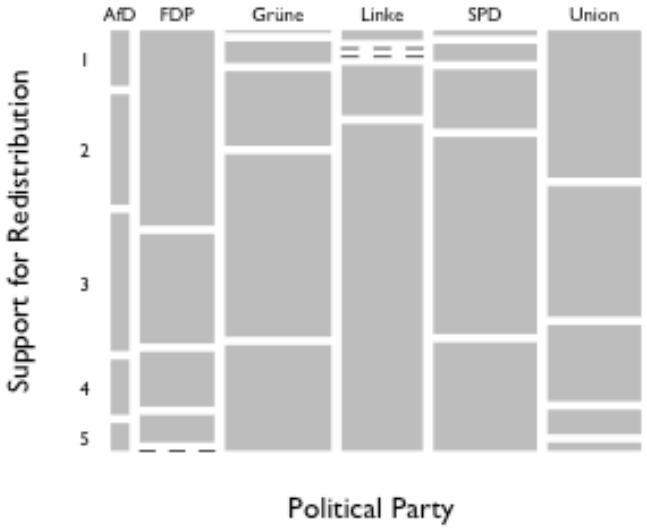
The Mosaic Plot

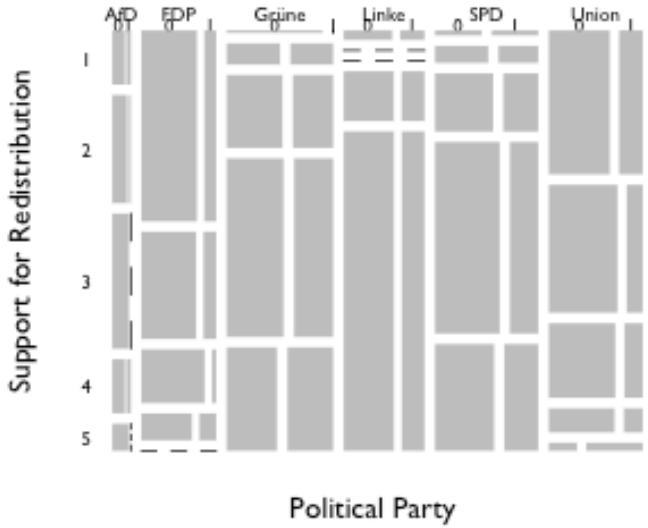




The Mosaic Plot





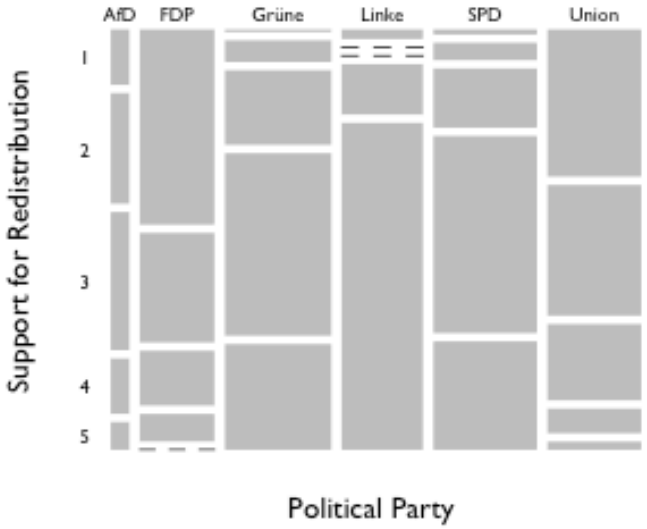


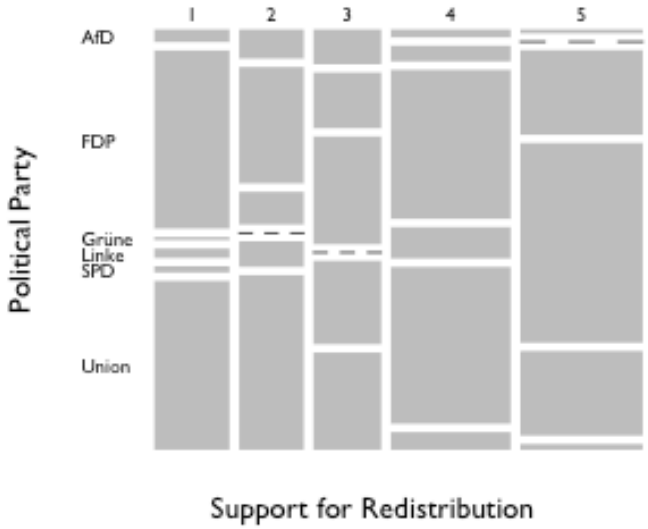
The Mosaic Plot

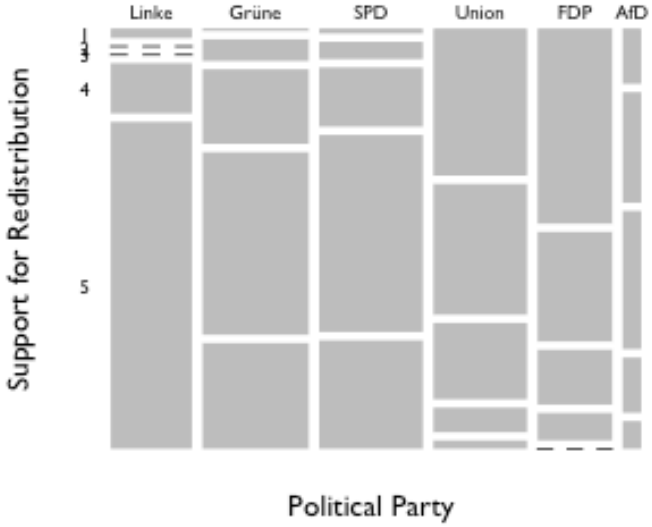
Highlevel Command

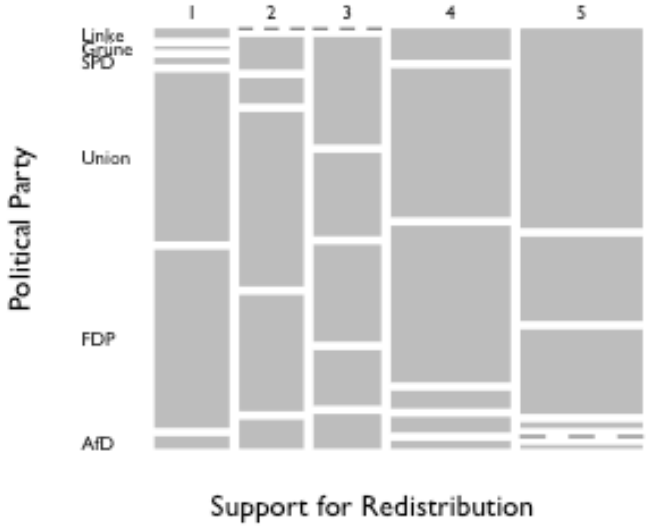
```
mosaicplot(table(data$party), xlab="Political Party", border=F, main="")

mosaicplot(table(data$party, data$redist), xlab="Political Party",
ylab="Support for Redistribution", las=1, border=F, main="")

mosaicplot(table(data$party, data$redist, data$female), xlab="Political
Party", ylab="Support for Redistribution", las=1, border=F, main="")
```

The Mosaic Plot: Sorting and Ordering







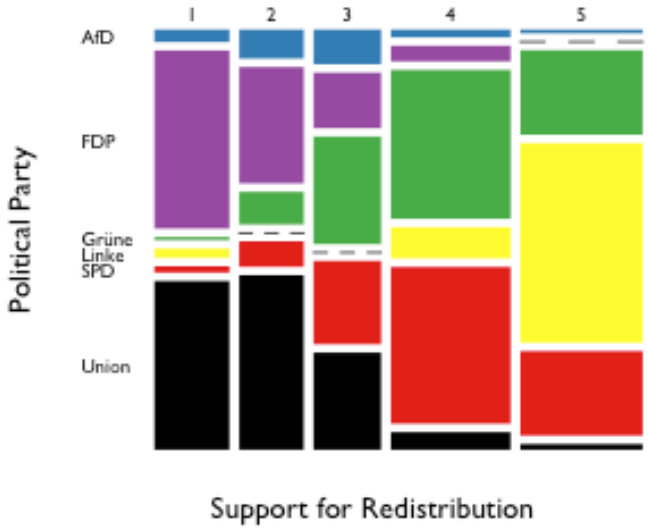


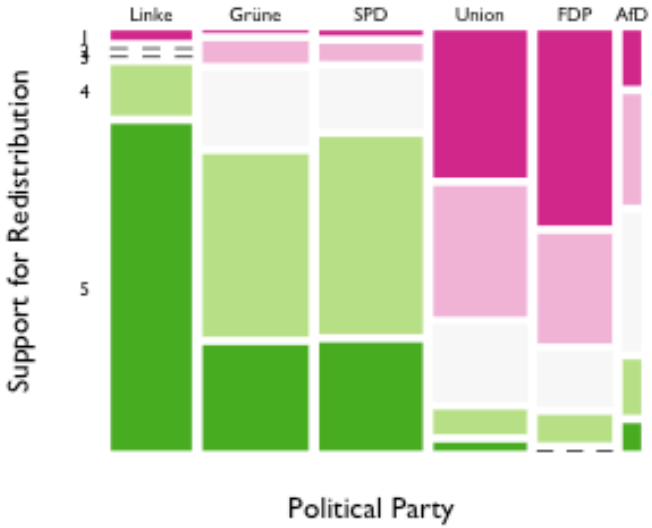
The Mosaic Plot: Sorting and Ordering

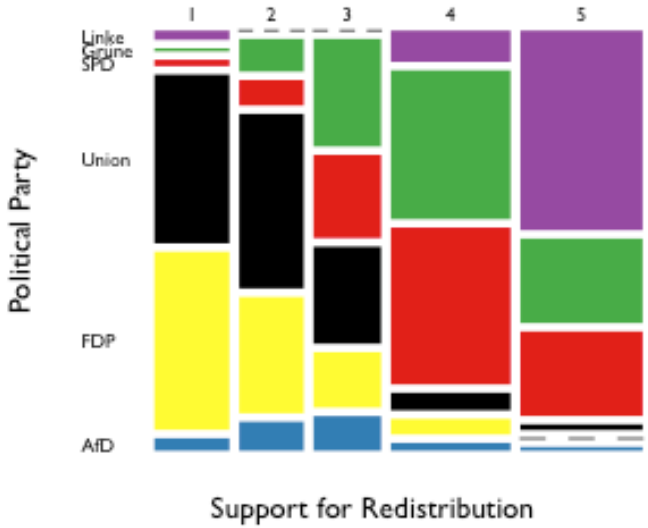
Different Sortings and Orderings

```
mosaicplot(table(data$party, data$redist), xlab="Political Party",
ylab="Support for Redistribution", las=1, border=F, main="")

mosaicplot(table(data$redist, data$party), xlab="Support for
Redistribution", ylab="Political Party", las=1, border=F, main="")

mosaicplot(table(data$party.ord, data$redist), xlab="Political Party",
ylab="Support for Redistribution", las=1, border=F, main="")

mosaicplot(table(data$redist, data$party.ord), xlab="Support for
Redistribution", ylab="Political Party", las=1, border=F, main="")
```

The Mosaic Plot: Adding Color





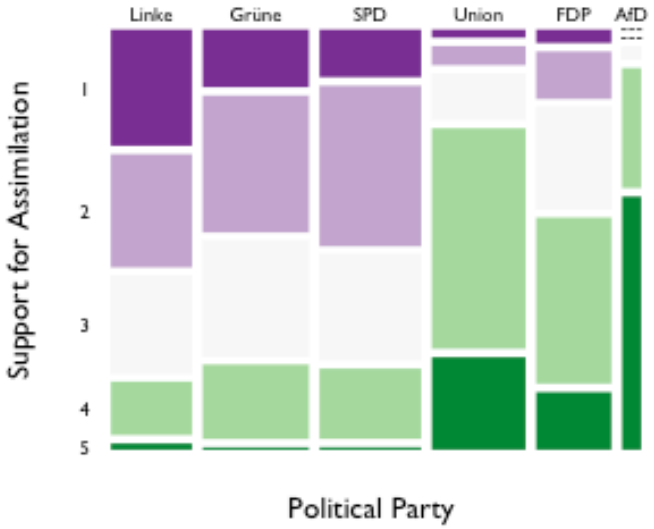


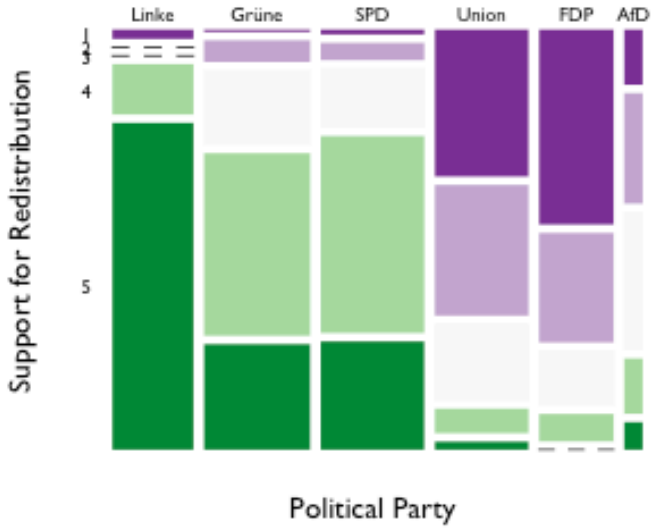
The Mosaic Plot: Adding Color

Add Color

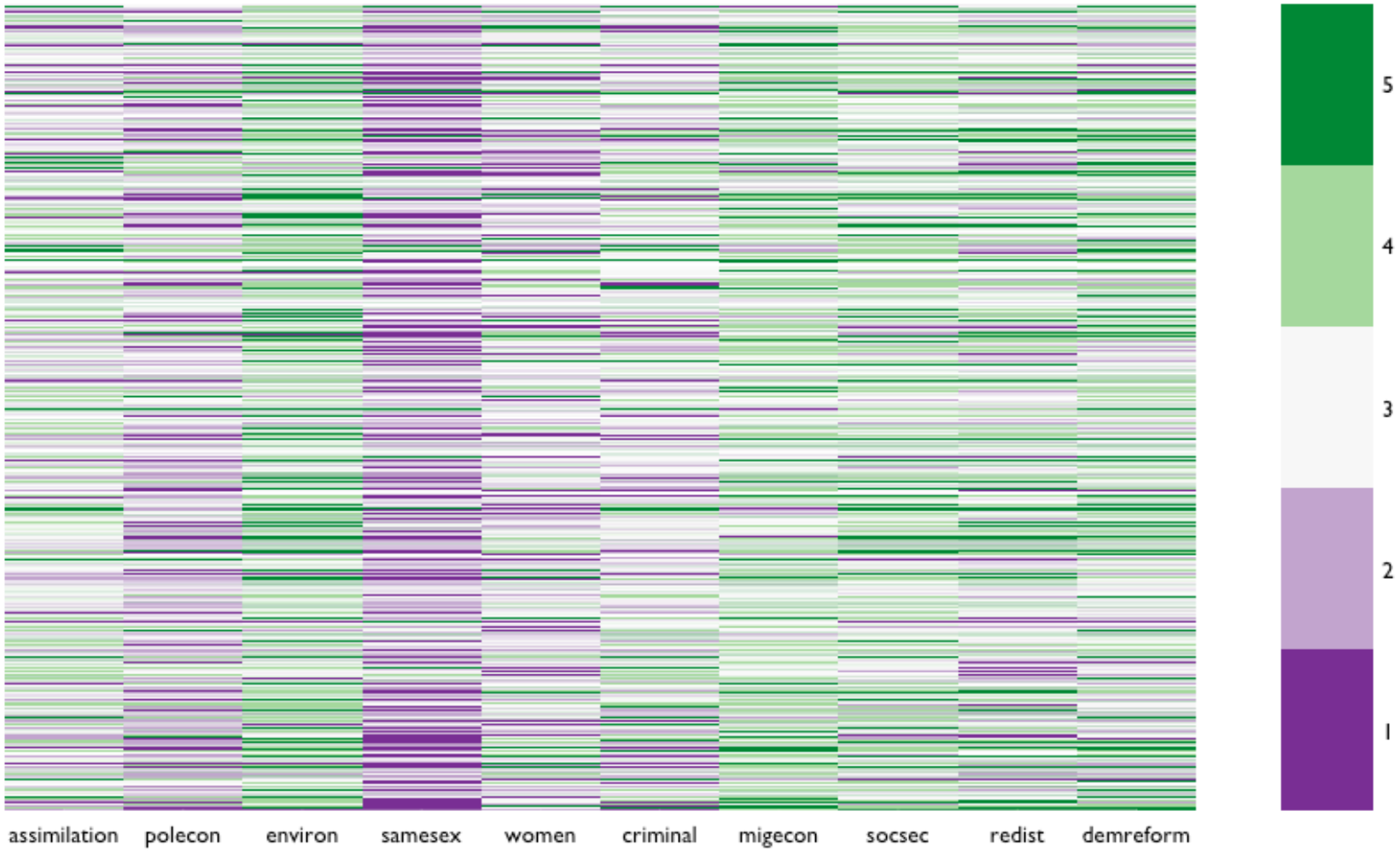
```
mosaicplot(table(data$party, data$redist), xlab="Political Party",
ylab="Support for Redistribution", las=1, border=F, main="",
col=brewer.pal(5, "PuOr"))

mosaicplot(table(data$redist, data$party), xlab="Support for
Redistribution", ylab="Political Party", las=1, border=F, main="",
col=rev(unique(data$part.col)[-6]))

mosaicplot(table(data$party.ord, data$redist), xlab="Political Party",
ylab="Support for Redistribution", las=1, border=F, main="",
col=brewer.pal(5, "PiYG"))

mosaicplot(table(data$redist, data$party.ord), xlab="Support for
Redistribution", ylab="Political Party", las=1, border=F, main="",
col=rev(unique(data$part.col)[-6][c(6, 3, 1, 2, 4, 5)]))
```

The Mosaic Plot: Comparing Policy Areas





The Heat Map


5
4
3
2
1
assimilation
polecon
environ
samesex
women
criminal
migecon
socsec
redist
demreform

The Heat Map

Heat Map of Full Data

Layout Grid

```
layout(rbind(c(1,2), c(1,2)), width=c(4,1))
```

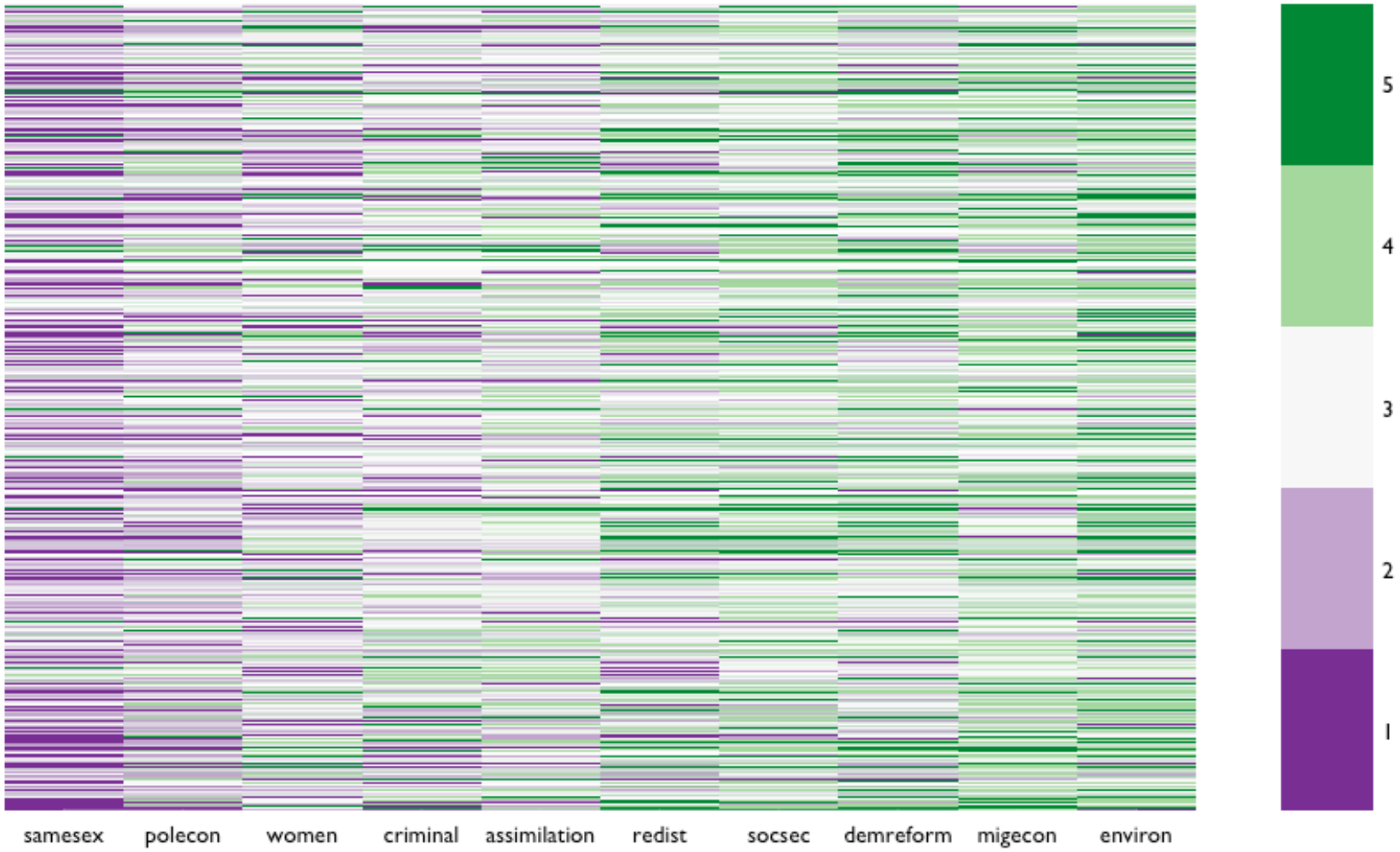
Plot Heat Map

```
par(mar=c(3,2,3,1))
image(t(as.matrix(data[,5:14])),  col=brewer.pal(5, "PRGn"), axes=F)
axis(1, at=my.scale(1:10, 10), label=colnames(data[,5:14]), las=1,
col="white", col.axis="black")
```

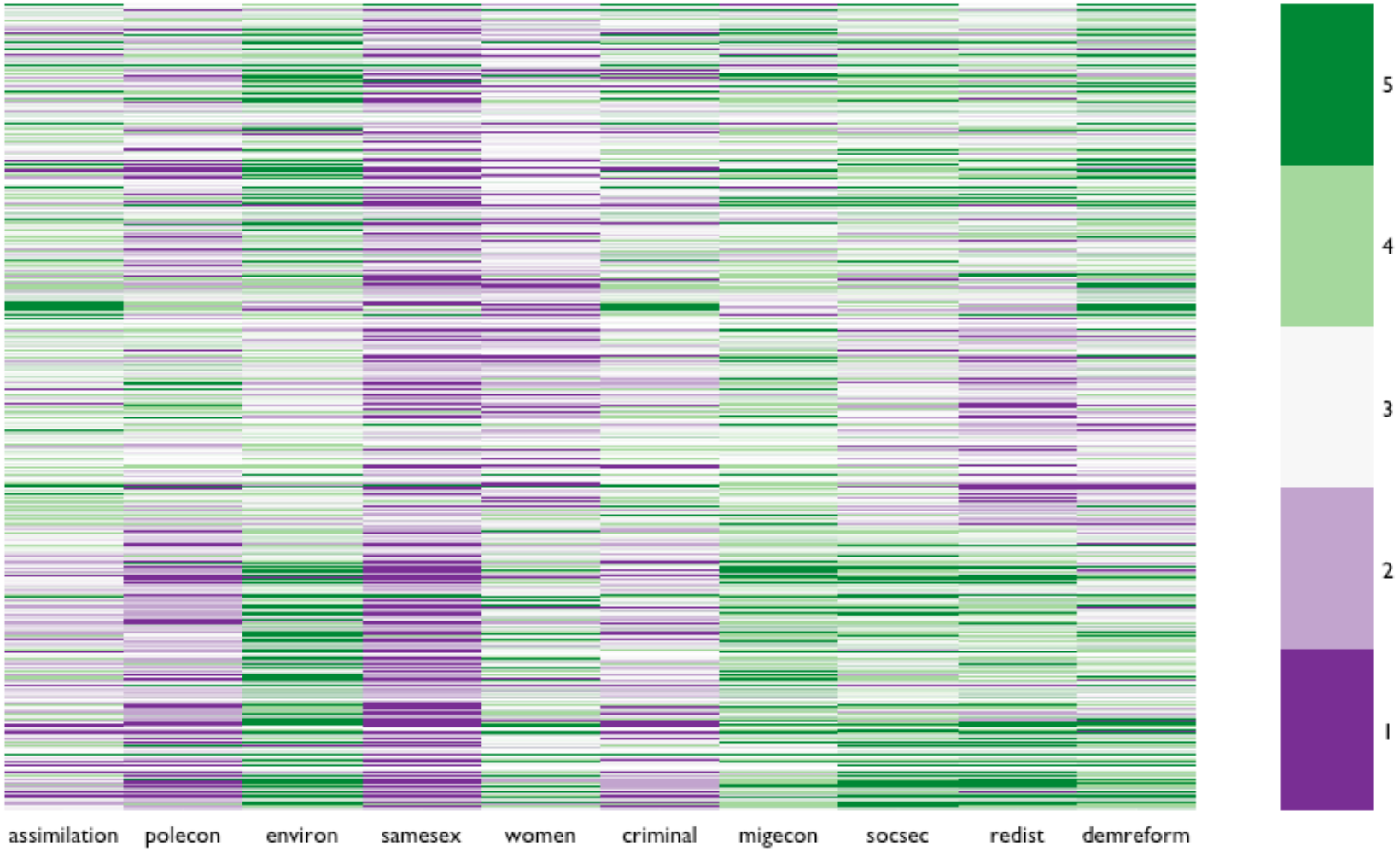
Add Color Legend

```
par(mar=c(3,2,3,6))
image(t(as.matrix(c(1:5))), col=brewer.pal(5, "PRGn"), axes=F)
axis(4,  at=my.scale(1:5, 5), label=c("1", "2", "3", "4", "5"),
col="white", col.axis="black")
```

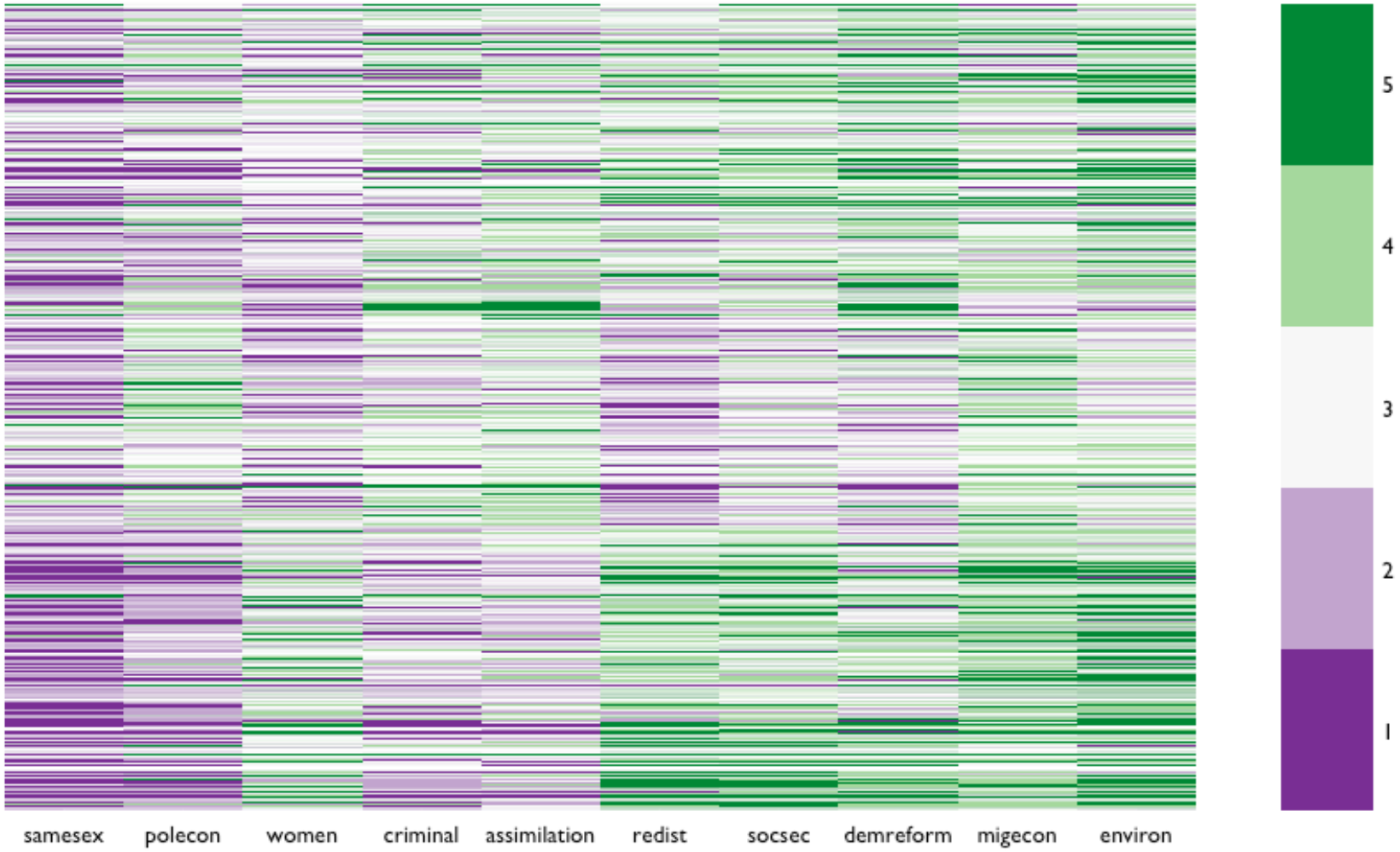
The Heat Map: Variables Ordered by Mean Support



The Heat Map: Oberservations Ordered by Party Affiliation



The Heat Map: Ordered by Both



The Heat Map

Order Observations by Party Affiliation

```
ord <- order(data$party.ord, na.last=T)
```

Order Variables by Mean Values

```
var.ord <- apply(data[,5:14], 2, mean, na.rm=T)
var.ord <- as.vector(order(var.ord))
```

Define Layout Grid

```
layout(rbind(c(1,2), c(1,2)), width=c(4,1))
```

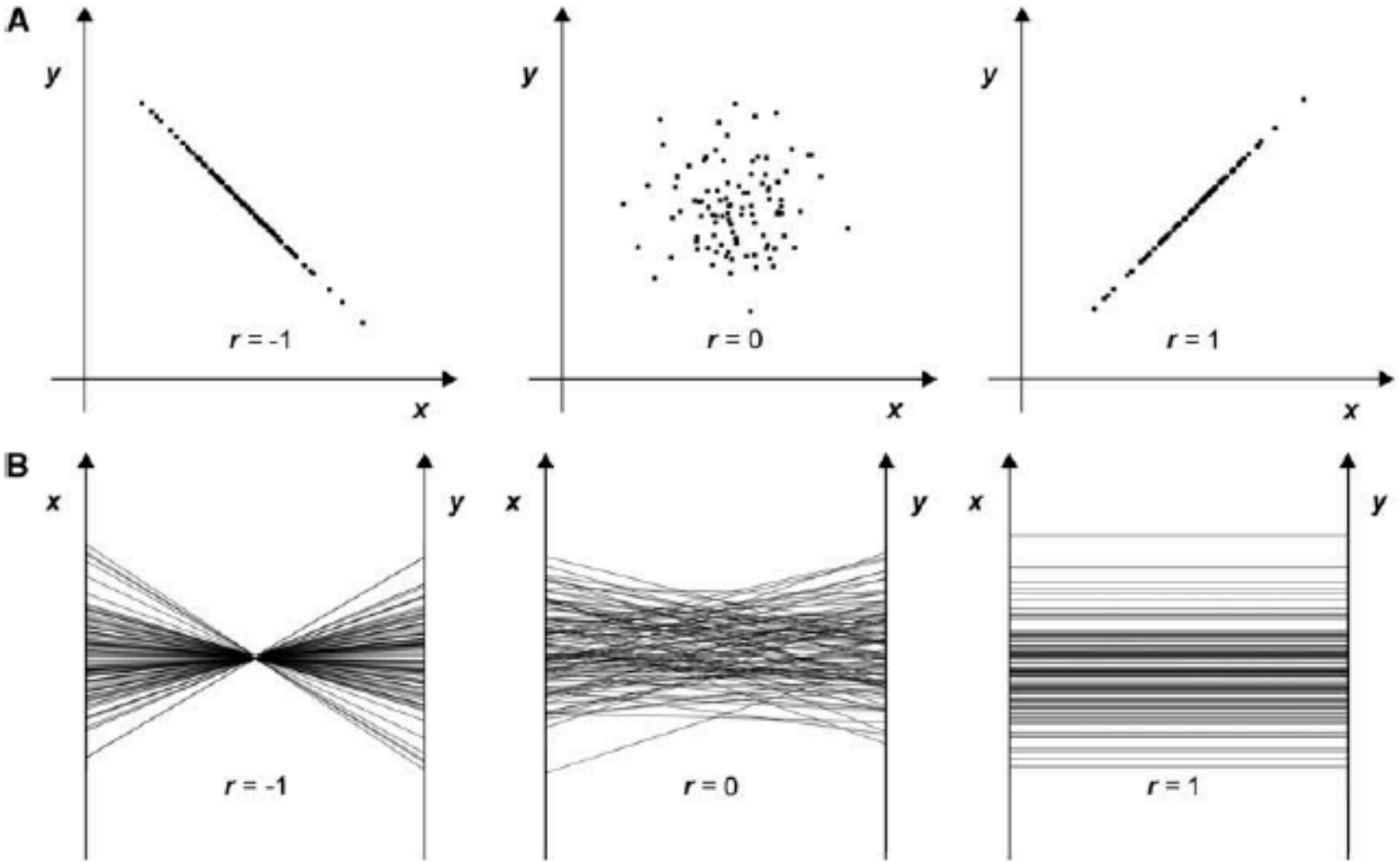
Plot Heat Map

```
par(mar=c(3,2,3,1))
image(t(as.matrix(data[ord, var.ord+4])),  col=brewer.pal(5, "PRGn"),
axes=F)
axis(1, at=my.scale(1:10, 10), label=colnames(data[,var.ord+4]), las=1,
col="white", col.axis="black")
```

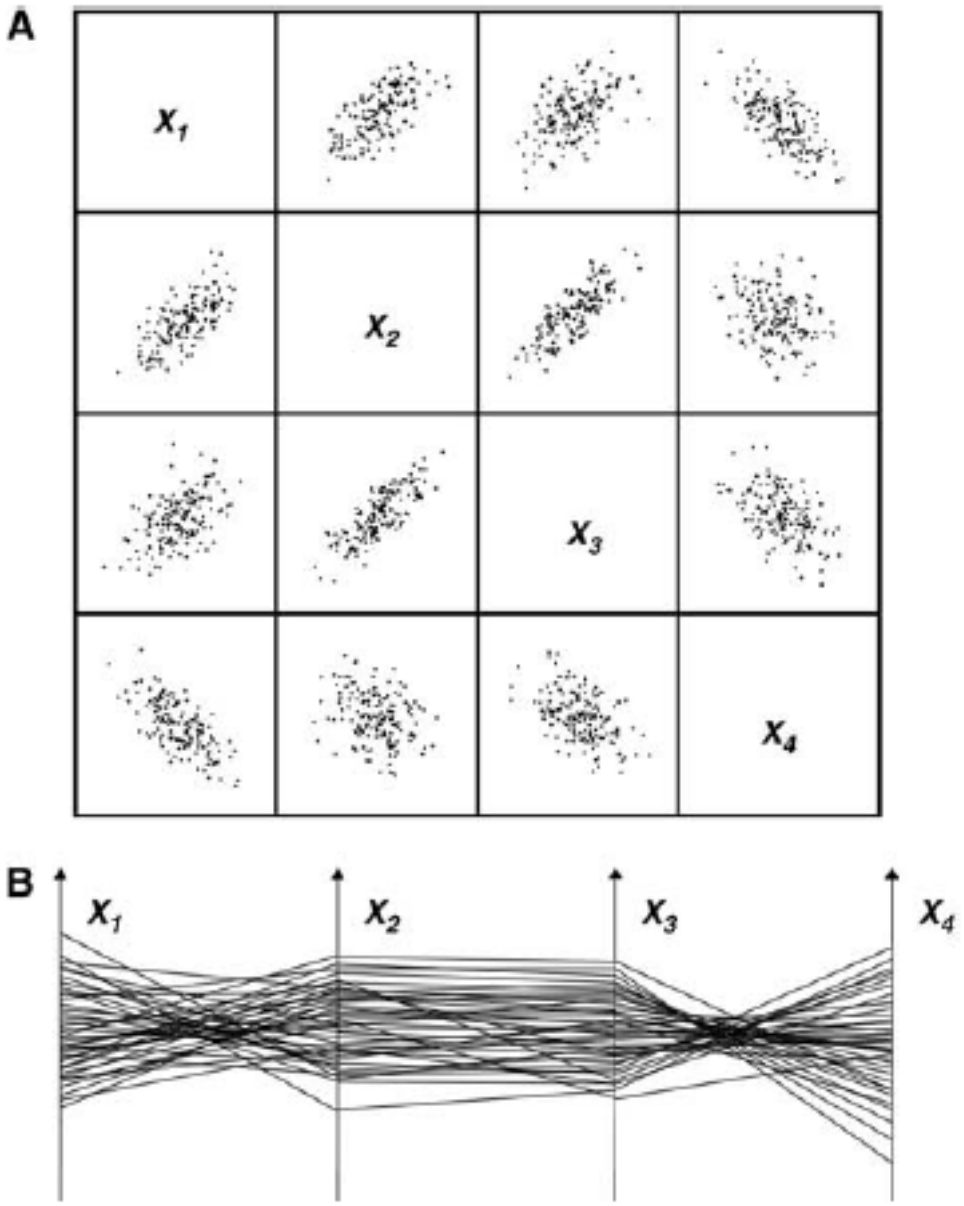
Add Color Legend

```
par(mar=c(3,2,3,6))
image(t(as.matrix(c(1:5))), col=brewer.pal(5, "PRGn"), axes=F)
axis(4,  at=my.scale(1:5, 5), label=c("1", "2", "3", "4", "5"),
col="white", col.axis="black")
```

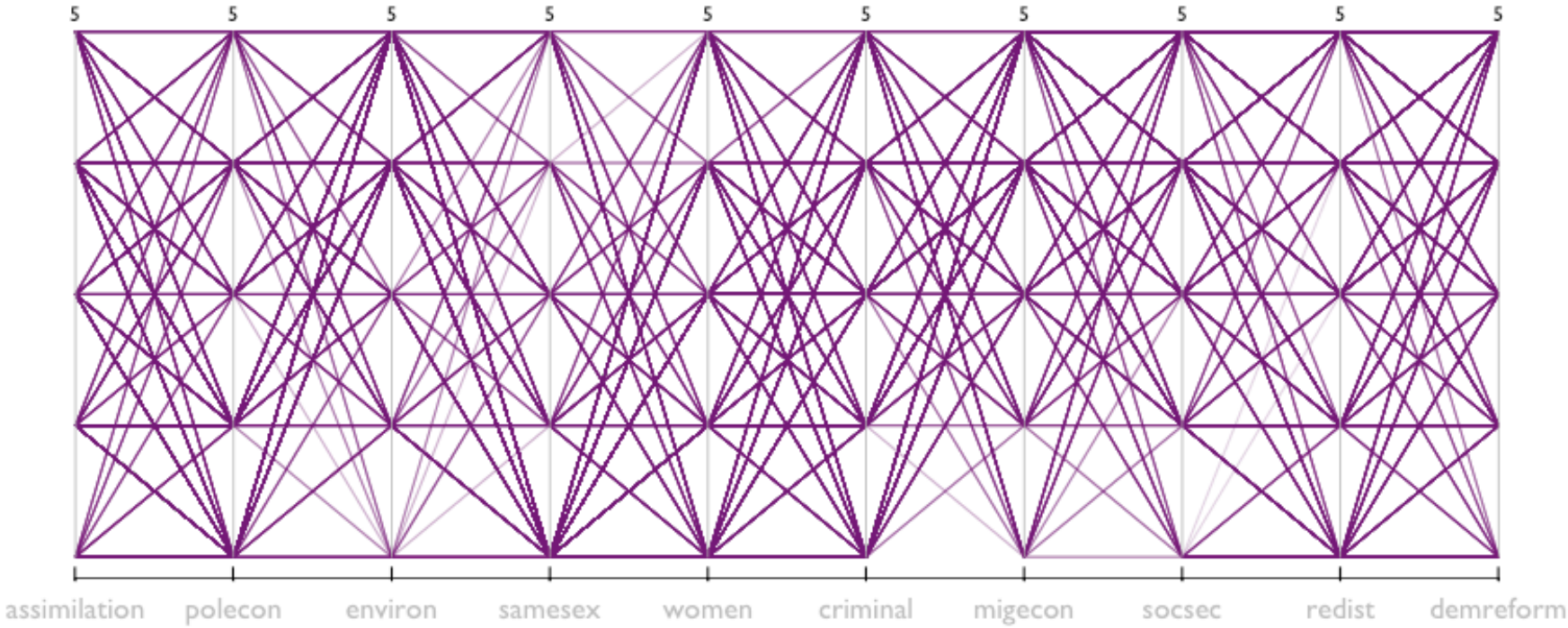
The Parallel Coordinates Plot



The Parallel Coordinates Plot

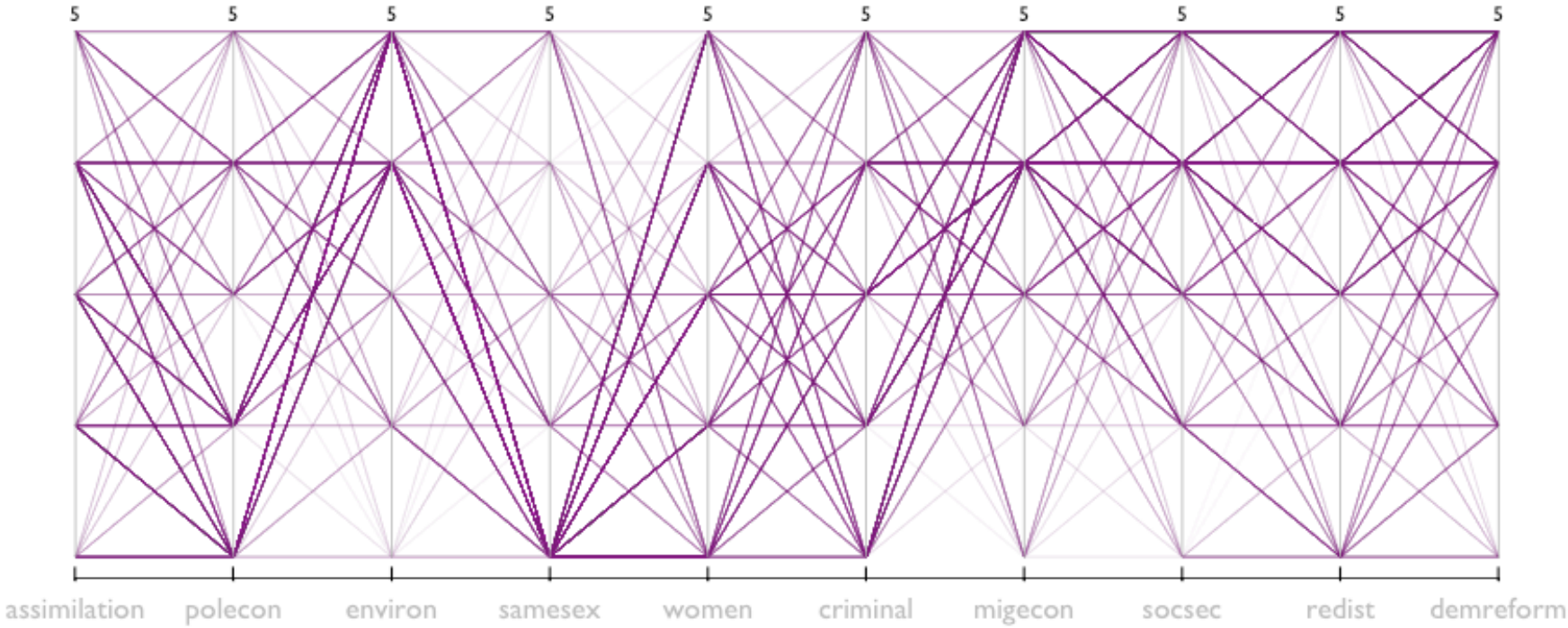


The Parallel Coordinates Plot



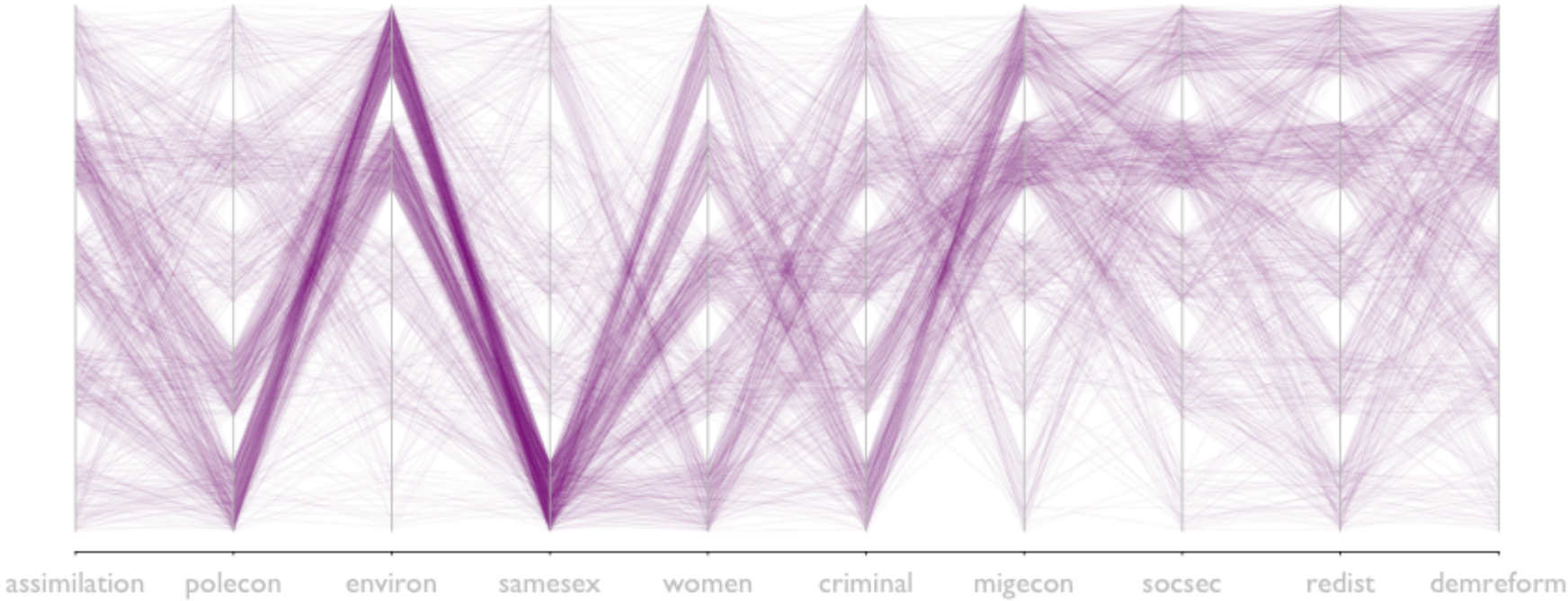

```
parcoord(data[, 5:14], lwd=.3, col=rgb(120, 00, 120, 255, max=255),
                                var.label=T)
```

The Parallel Coordinates Plot: Alpha Blending




```
parcoord(data[, 5:14], lwd=.3, col=rgb(120, 00, 120, 40, max=255),
                                                    var.label=T)
```

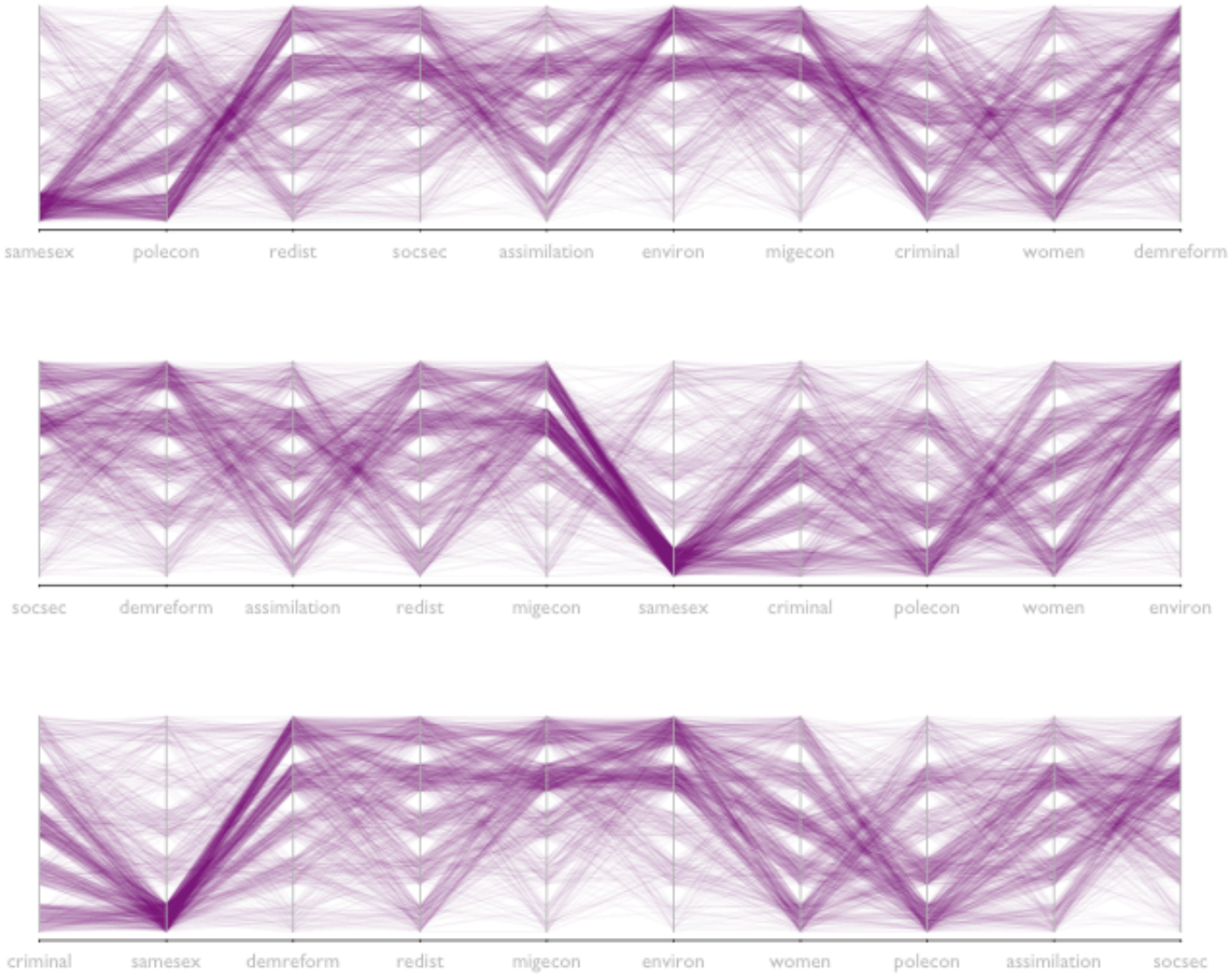
The Parallel Coordinates Plot: Alpha Blending and Jittering




```
parcoord(jitt.data, lwd=.3, col=rgb(120, 00, 120, 40, max=255),
                                var.label=T)
```

(I created jittered data just a in a couple of slides ago.)

The Parallel Coordinates Plot: Ordering of Axes



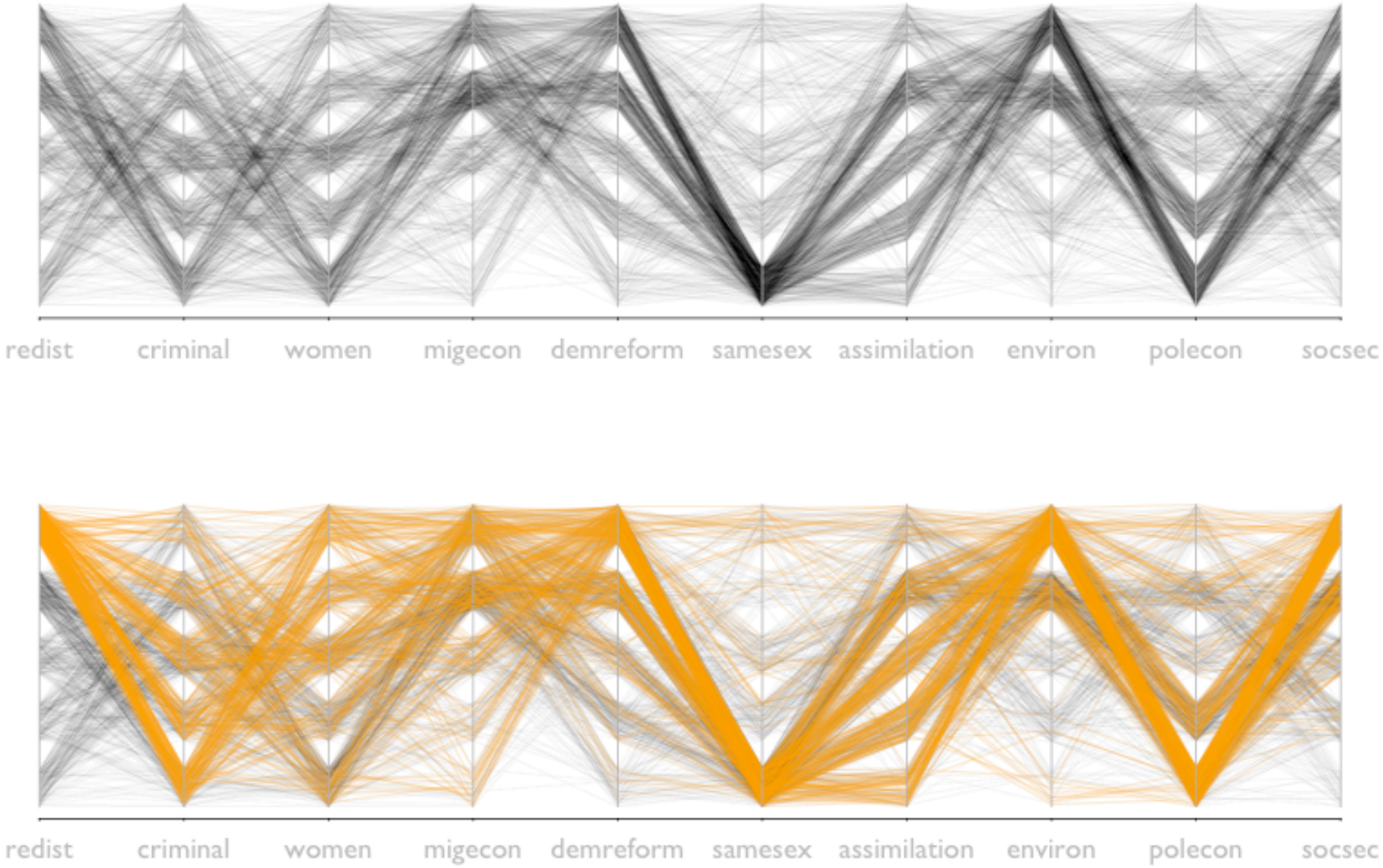
The Parallel Coordinates Plot: Ordering of Axes

```
par(mfrow=c(3,1))
```

Random Order of Axes

```
rand <- sample(1:10, 10, replace=F)

parcoord(jitt.data[,rand], lwd=.3, col=rgb(120,00,120, 40, max=255),
cex.axis=.6)
```

The Parallel Coordinates Plot: Brushing



The Parallel Coordinates Plot: Brushing

Assign Color to Variable Condition

```
data$col.code <- ifelse(data$redist>=5, "orange", rgb(00,00,00,
40, max=255))
```

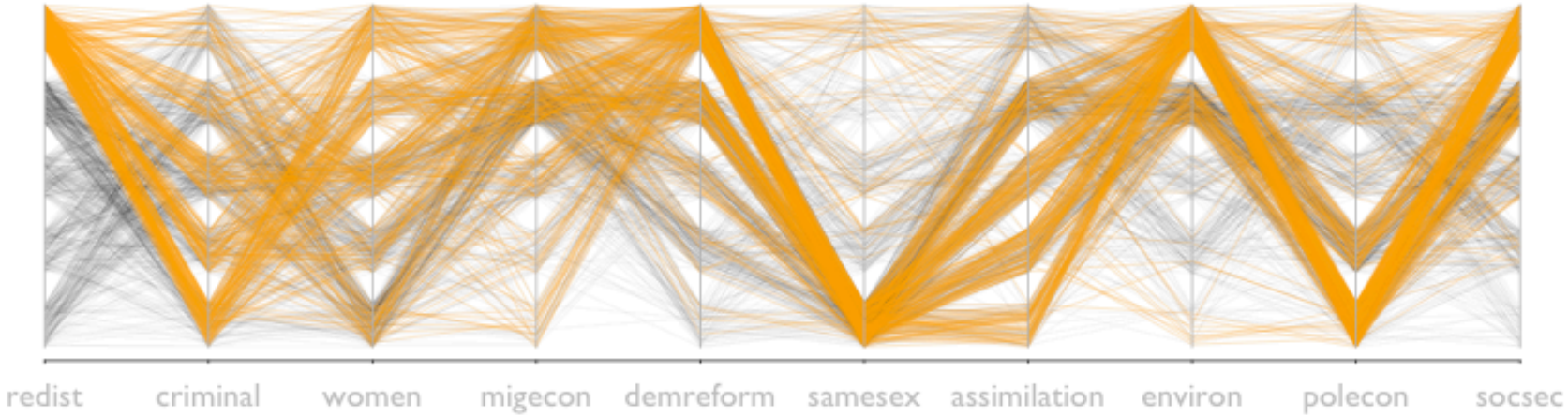
Order, so that colored observations are added last

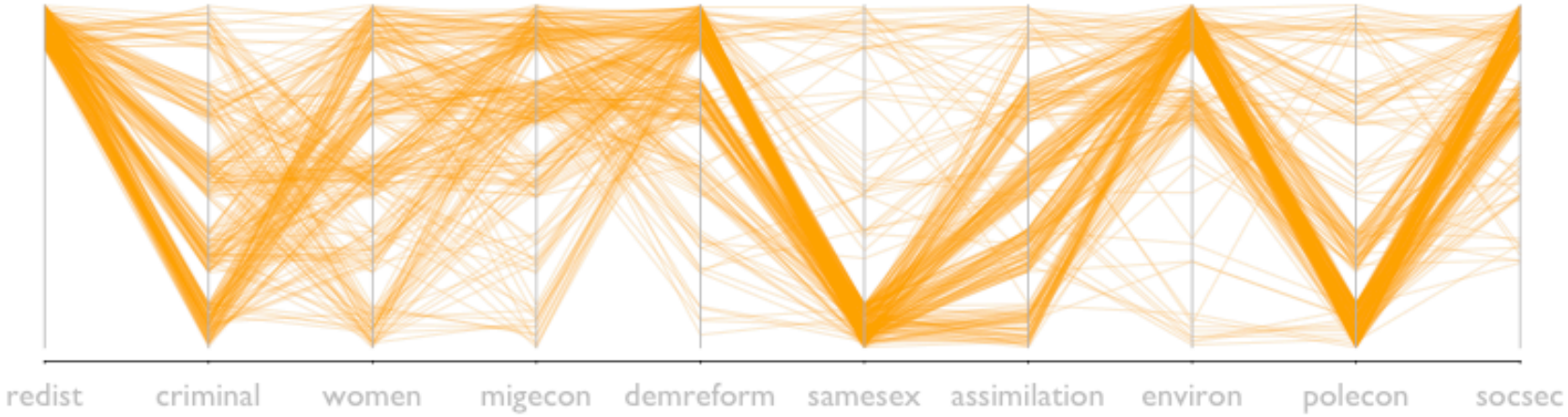
```
col.ord <- order(data$col.code, na.last=F)
```

Plot

```
parcoord(jitt.data[col.ord, rand], lwd=.3,
col=data$col.code[col.ord], cex.axis=.6)
```

The Parallel Coordinates Plot: Filtering





The Parallel Coordinates Plot: Filtering

Filtering = do not plot = NA

```
data$col.code <- ifelse(data$redist>=5, "orange", NA)

col.ord <- order(data$col.code, na.last=F)

parcoord(jitt.data[col.ord, rand], lwd=.3, col=data$col.code[col.ord],
cex.axis=.6)
```

Lab Exercise

Visually explore gender and age differences in political party ideology, by adding one of these or both variables to the graphic formats we have discussed.

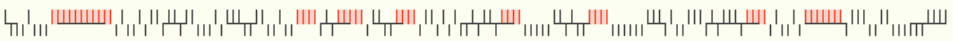
Be sure to experiment with different formats, orderings, and color schemes!

Use a parallel coordinates plot with brushing to explore party cluters of ideology.

Let's talk about baseball...

2009 Boston Red Sox win–loss record, 95–67, winning streaks of 4 or more games

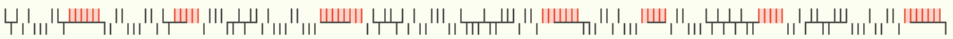
Longest winning streak



11 games

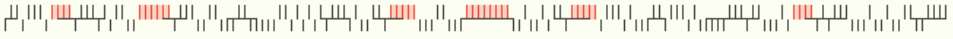
10 random seasons with the same win–loss record, both at home and on the road

Regenerate random seasons



8

7

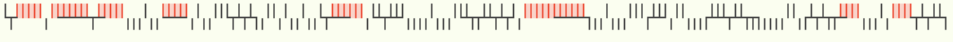


8

13

9

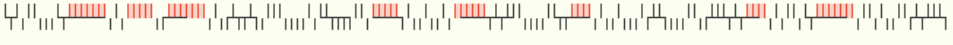
12



11

10

5



7

Let's talk about baseball...

2009 Boston Red Sox win–loss record, 95–67, winning streaks of 4 or more games

Longest winning streak

11 games

10 random seasons with the same win–loss record, both at home and on the road

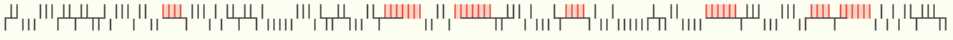
Regenerate random seasons

8

5

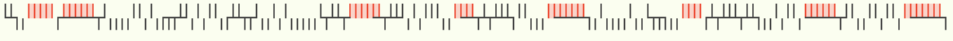
7

13



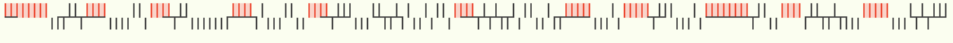
7

7



7

8

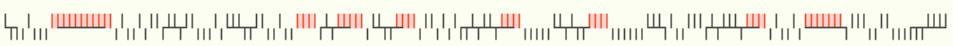


8

9

Let's talk about baseball...

2009 Boston Red Sox win–loss record, 95–67, winning streaks of 4 or more games

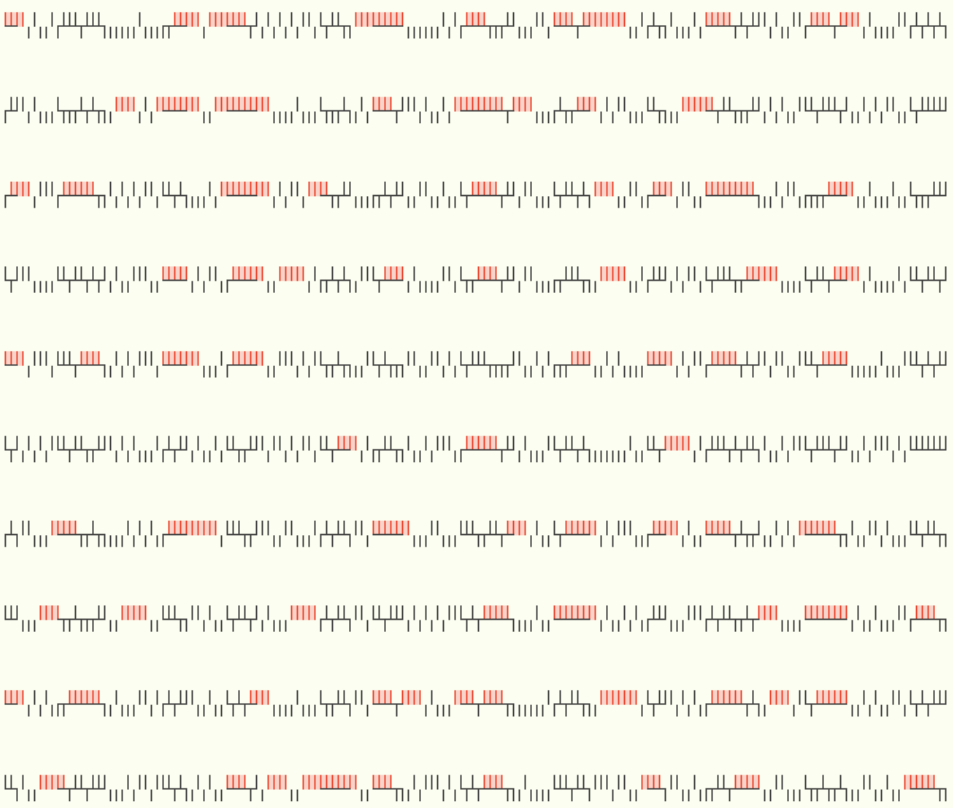


Longest winning streak

11 games

10 random seasons with the same win–loss record, both at home and on the road

Regenerate random seasons



9

10

9

6

7

6

9

8

7

10

Human Talent and Weakness

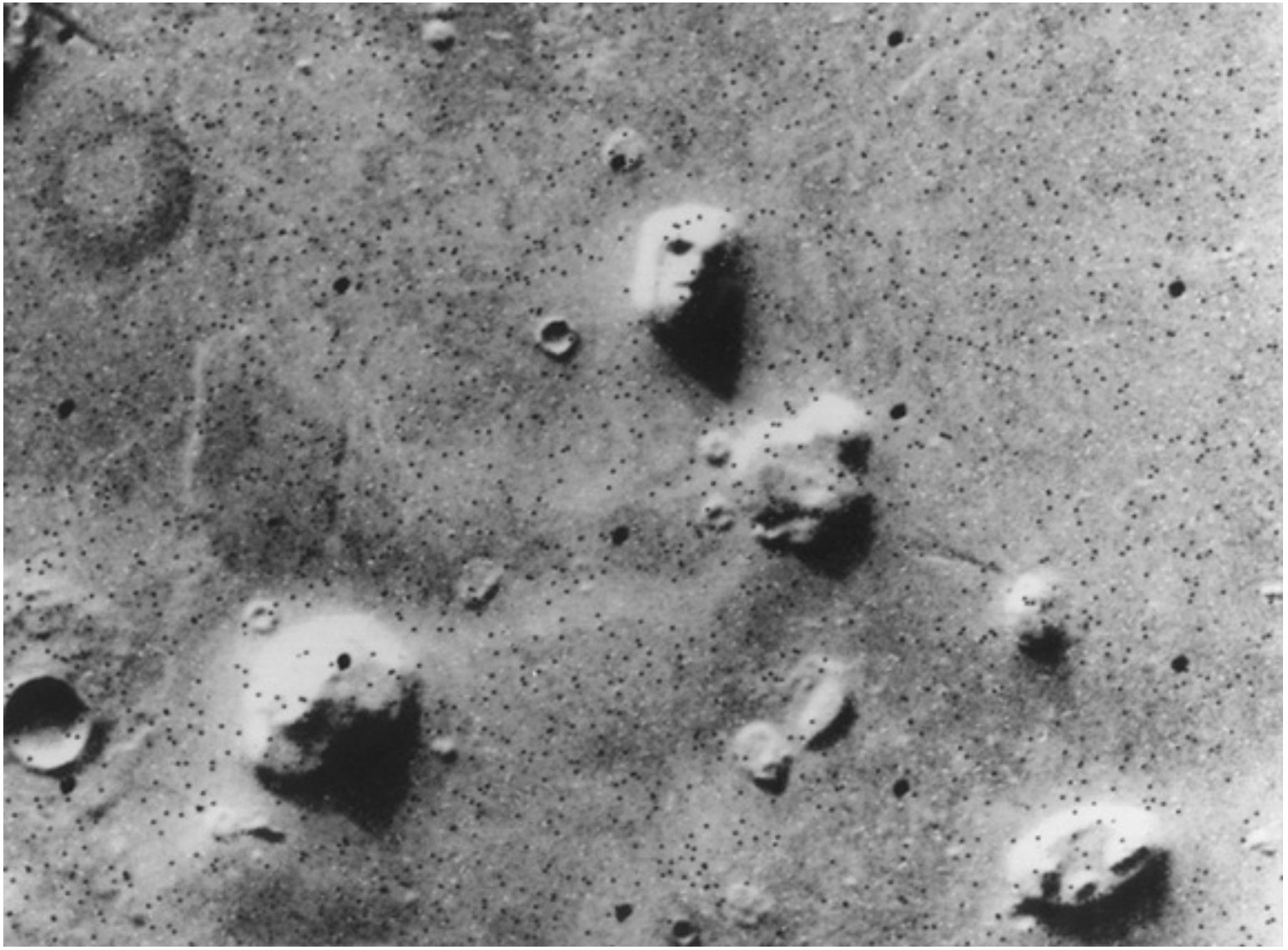
Humans are extremely good at seeing patterns.

At the same time, humans are also extremely bad in dealing with probability and randomness.

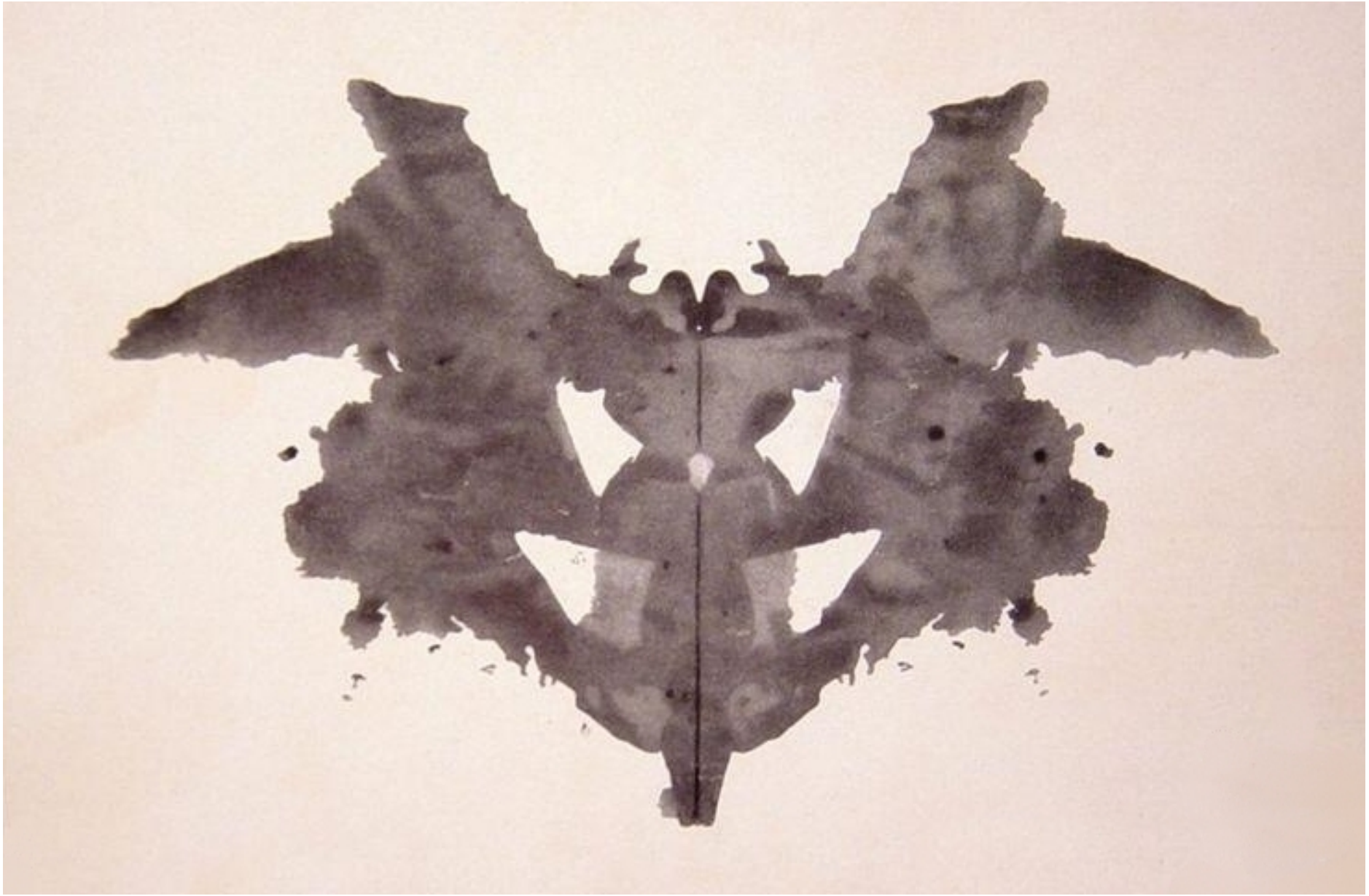
As a result of this talentsand weakness, they even see patterns, when there is nothing to see but pure randomness.

This tendency to see patterns in random data is called „apophenia“.

Image of Mars taken by NASA's Viking 1 orbiter, in grey scale, on July, 25 1976.



Rorschach Inkblot Test



Concerns with Exploratory Data Analysis

A concern that frequently arises with exploratory analysis is that it lacks the rigor of formal tests in confirmatory analysis or conventional statistical inference.

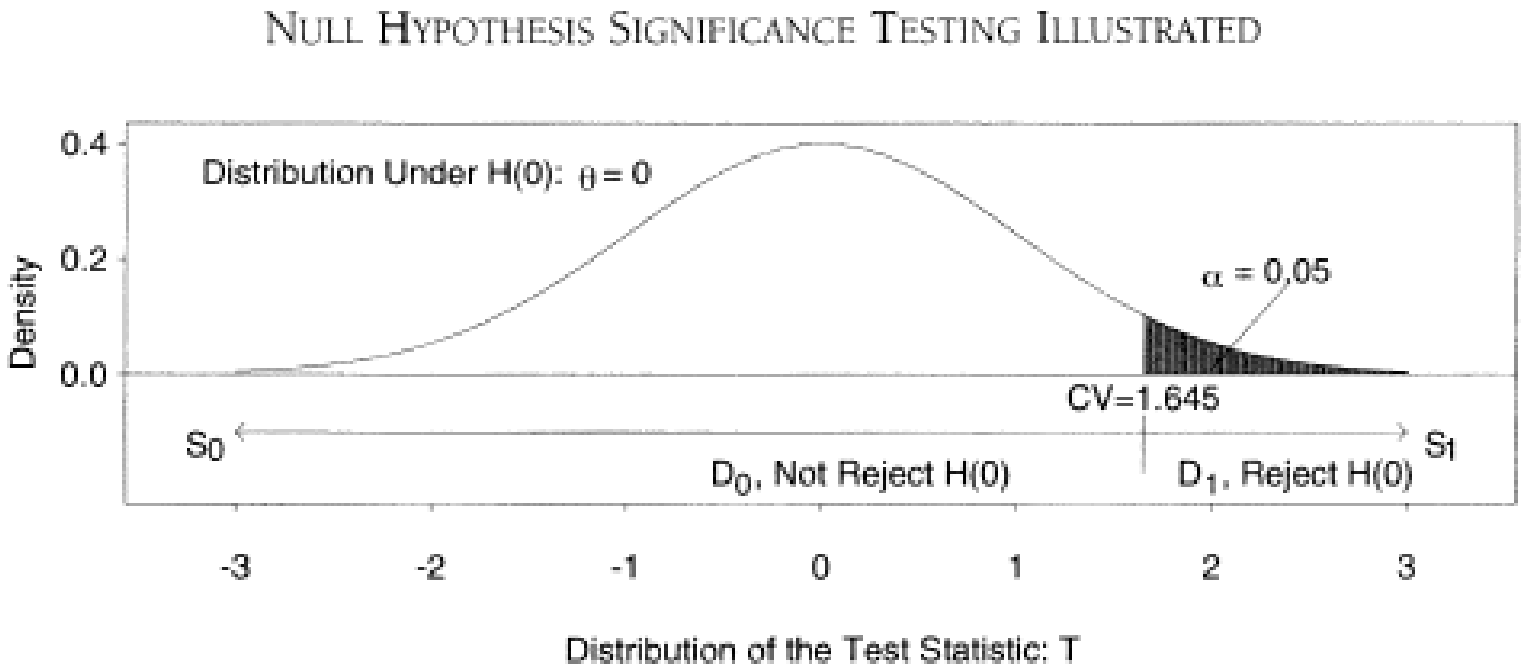
Exploratory data analysis and graphs give rise to an over-interpretation of patterns that may in fact be due to mere randomness:

"Humans' pattern recognition skills are amazing and the source of great insights, but sometimes they're too good. We are so adept at finding patterns that we sometimes detect ones that aren't really there" (Few 2009: 139).

Long-standing reservations against visualization as merely "informal" approach to data analysis and the fear that beautiful pictures may in fact not correspond to any meaningful patterns of substantive scientific interest.

The NHST

Formal testing involves the comparison of a test statistic to its reference distribution under the assumption of the null hypothesis.



If the test statistic is reasonably unlikely to have occurred under the null assumption, say $p < 0.05$, then the null hypothesis is rejected and one has a "statistically significant" result – unlikely to have occurred by chance.

Overcoming the Opposition of Exploratory vs. Confirmatory Analysis

Graphical displays are implicit or explicit comparisons to a reference distribution or baseline model (Gelman 2003, 2004).

If we discover an interesting pattern in data this usually means that it *looks different from what we expected*

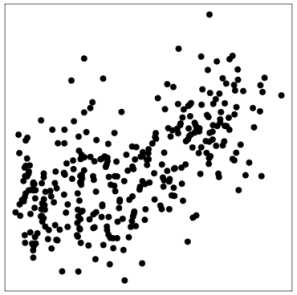
We usually we have implicit models in our mind to which we compare the data ("what do we expect to see?")

We can make these models *explicit* and use them to guard against "false discoveries"

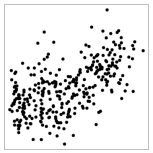
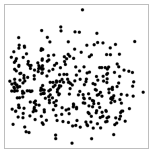
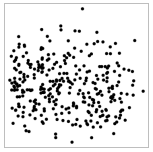
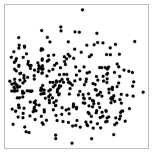
Visual discoveries correspond to the implicit or explicit rejection of null hypotheses (Buja et al. 2009).

Visual Inference as an analogue to NHST

The basic principle of formal testing remains the same in visual inference – with the exception that the test statistic is now a graphical display which is compared to a "reference distribution" of plots showing the null

Formal Test	Visual Inference
Null hypothesis H_0	Null hypothesis H_0
Test statistic $T = f(x)$	Visual feature in a plot 
Test: Reject? $T(x) > c$?	Human viewer: Discovery?

Simulation Based Testing

Formal Test	Visual Inference
Test statistic of true data $T = f(x)$	Plot of true data 
Test statistic simulated data $T = f(x_{s=1})$	Plot of simulated data 
Test statistic simulated data $T = f(x_{s=1})$	Plot of simulated data 
…	…
Test statistic simulated data $T = f(x_{s=S})$	Plot of simulated data 

Visual Inference: The Line-Up Protocol

Visual inference process that mimics conventional hypothesis tests (Buja et al. 2009, Wickham et al. 2010)

This method is called "after the 'police lineup' of criminal investigations […], because it asks the witness to identify the plot of the real data from among a set of decoys, the null plots, under the veil of ignorance" (Buja et al. 2009: 4369).

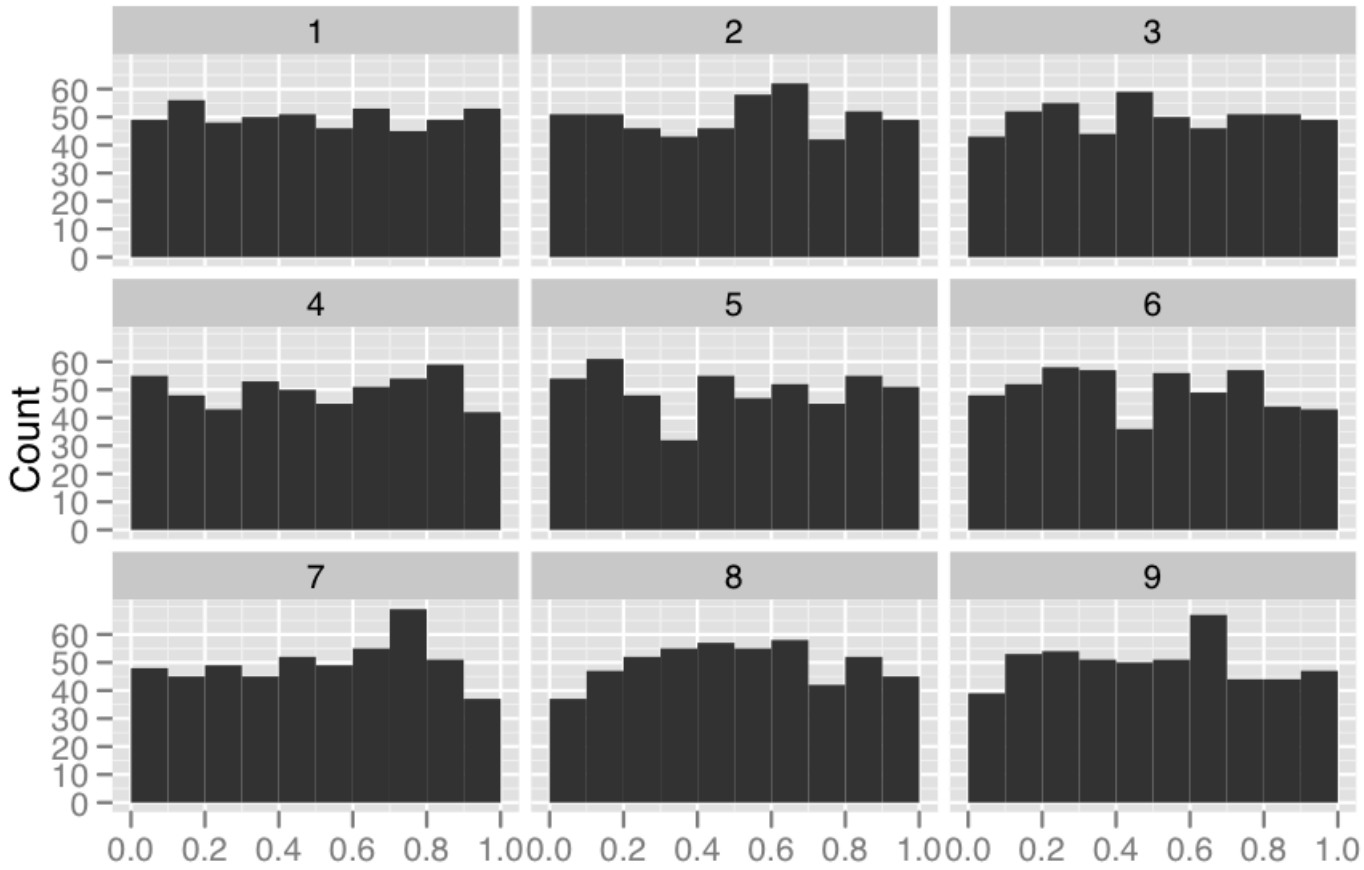
The visual hypothesis test involves the simulation of *m-1* null plots and randomly placing the plot of the real data among them, resulting in a total of *m* plots.

A human viewer is then asked to choose the plot that looks the most different from the rest.

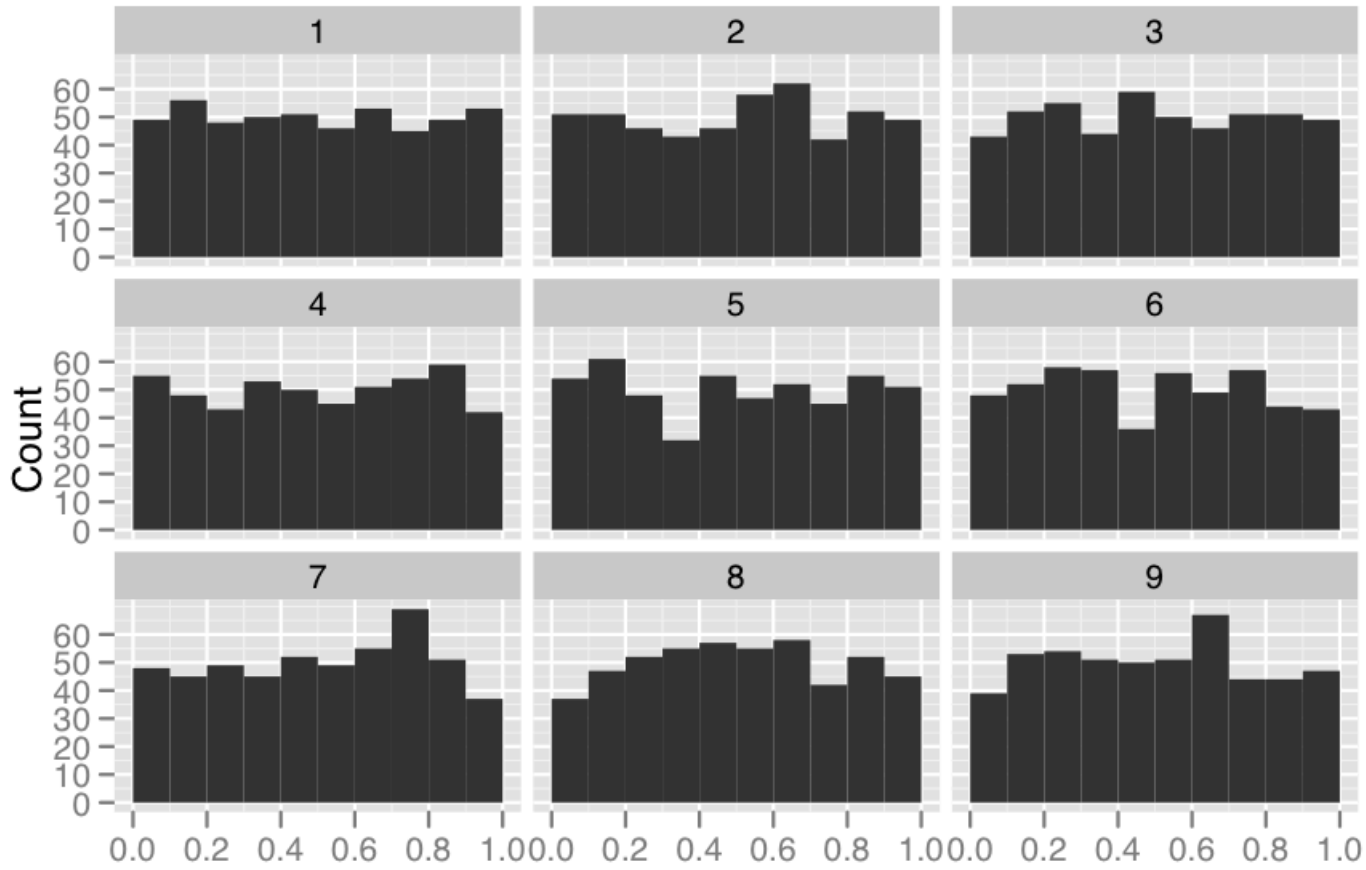
If the test person succeeds and picks the plot showing the actual data, then this visual discovery can be assigned a *p*-value of 1/m.

In other words, the probability of picking the true plot just by chance is 1/m.

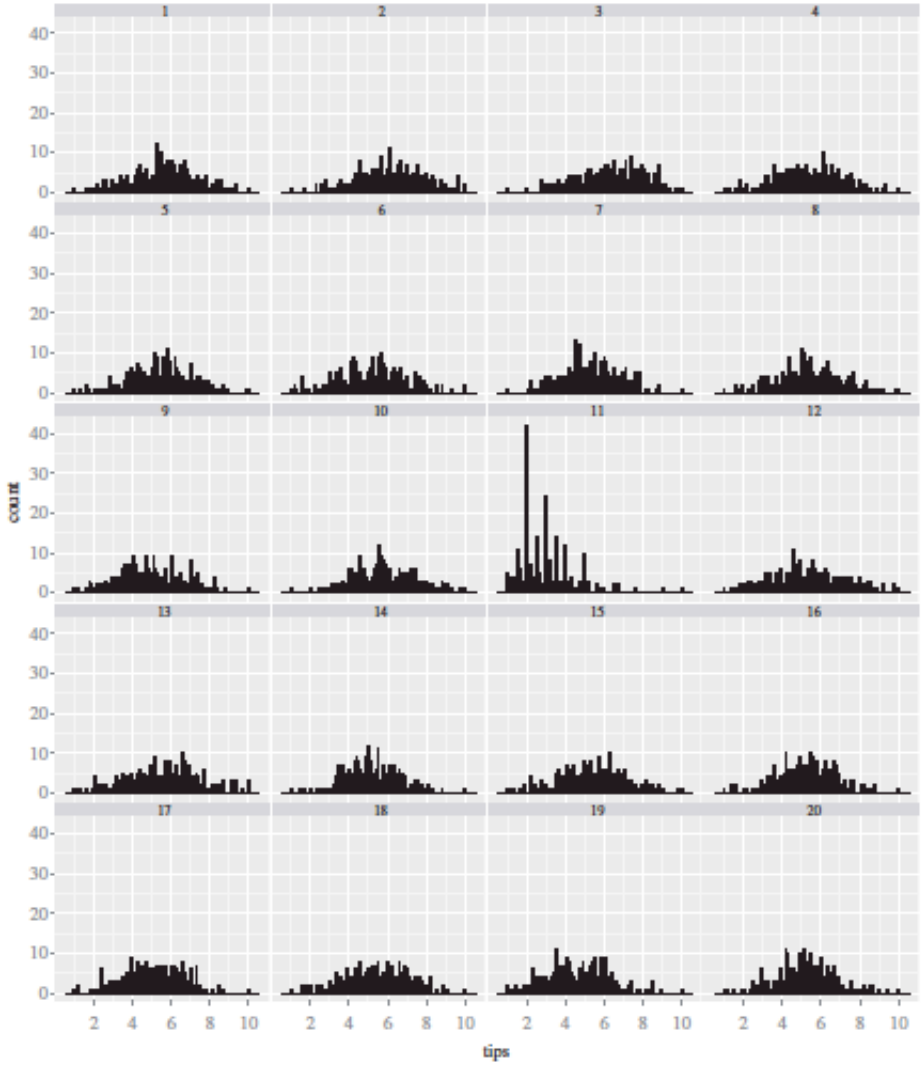
Which plot stands out from the rest?



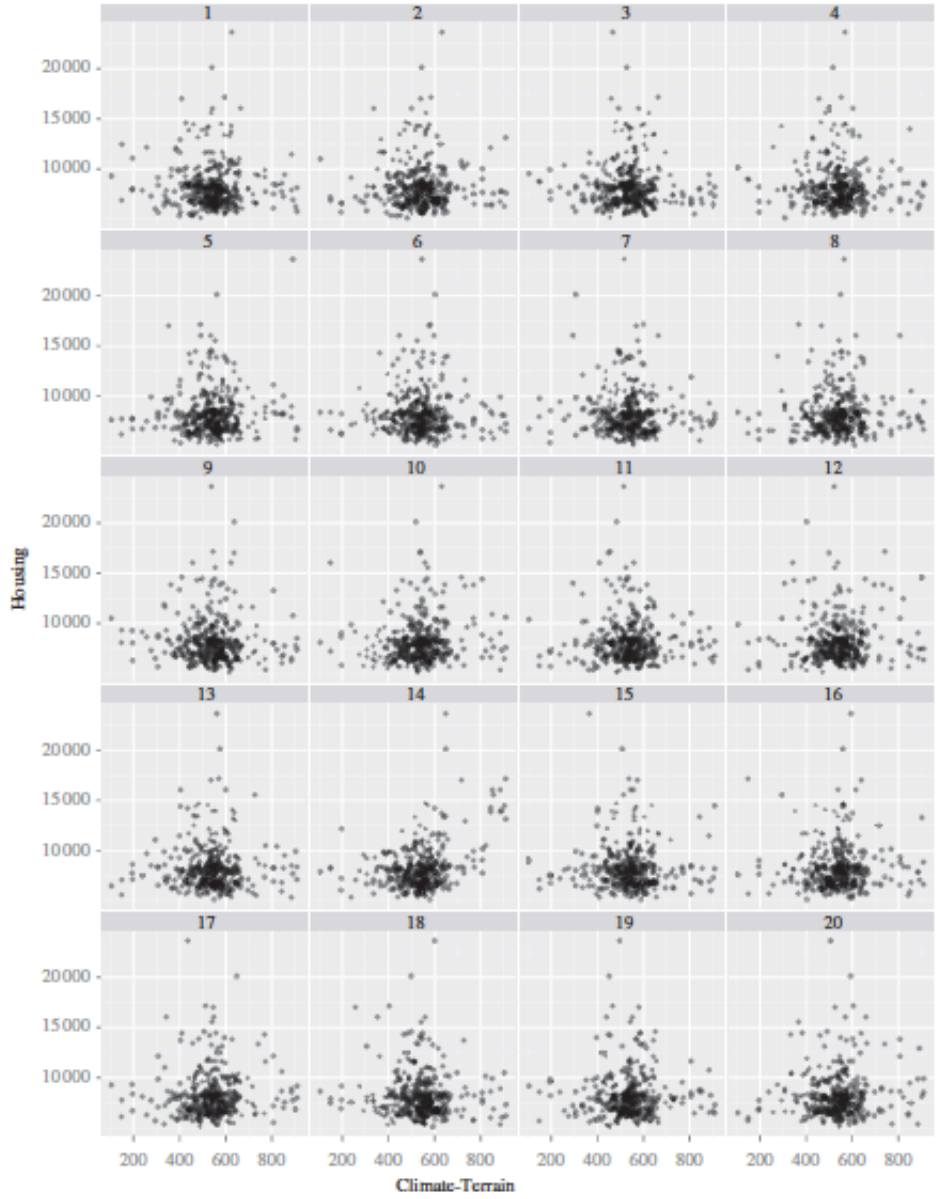
None! All just show 500 random draws from a uniform distribution U(0,1).



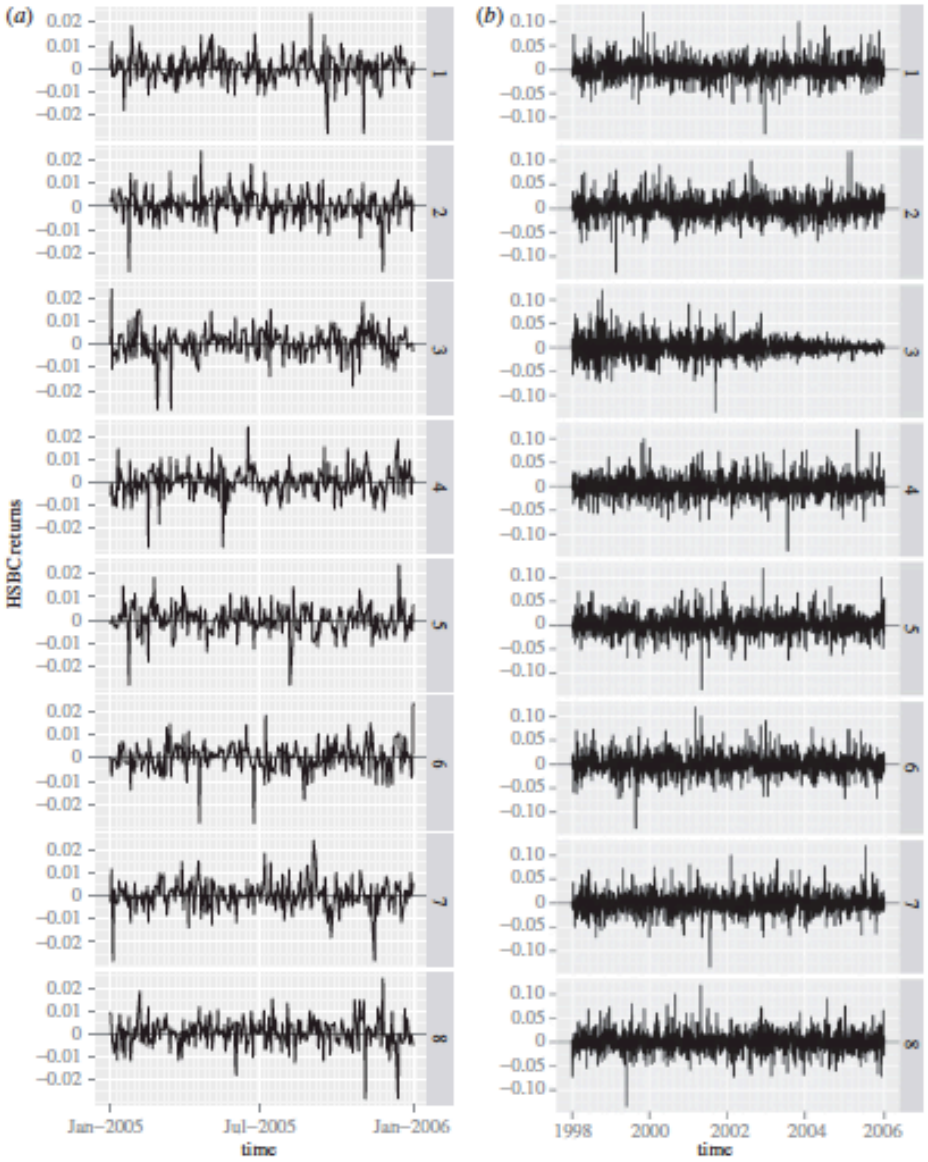
Ok, let's try again.



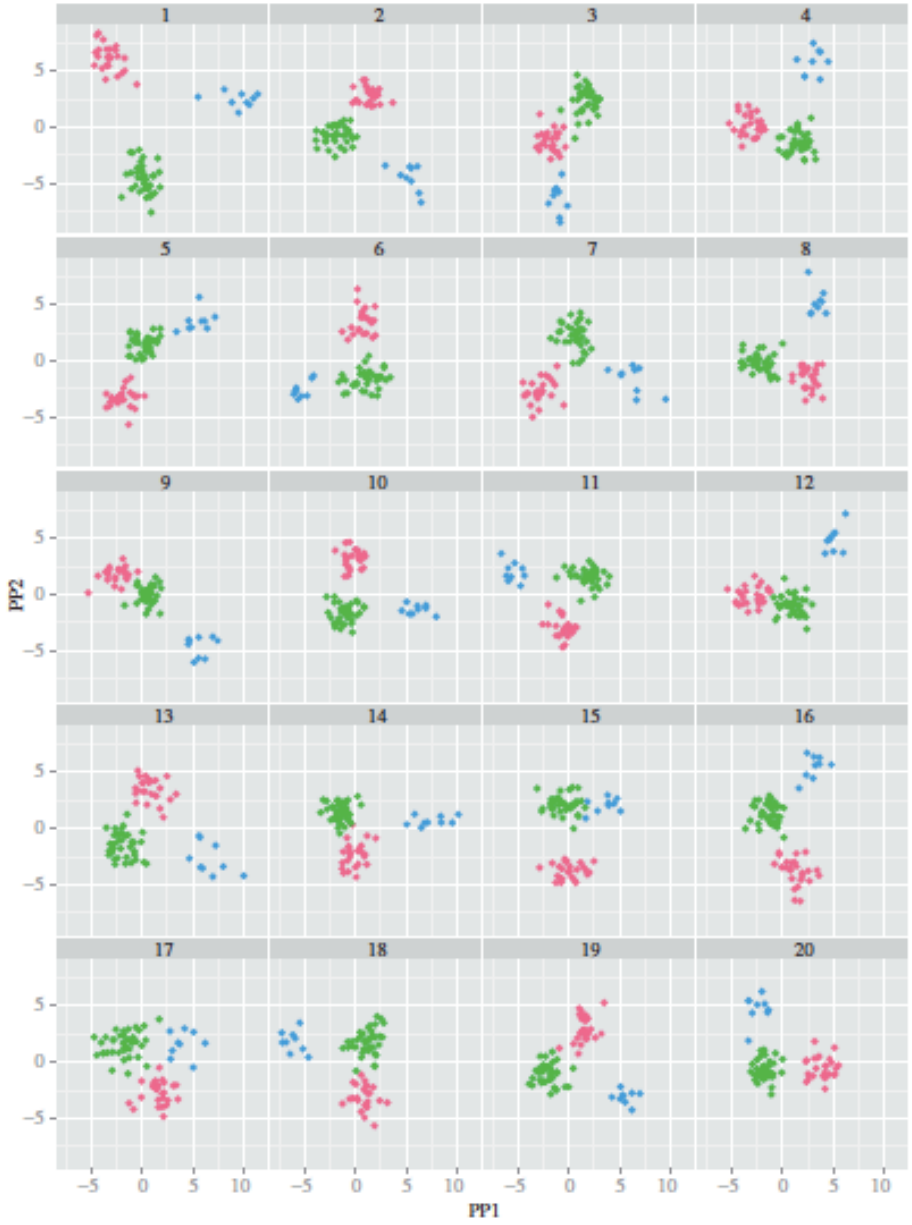
How about this?



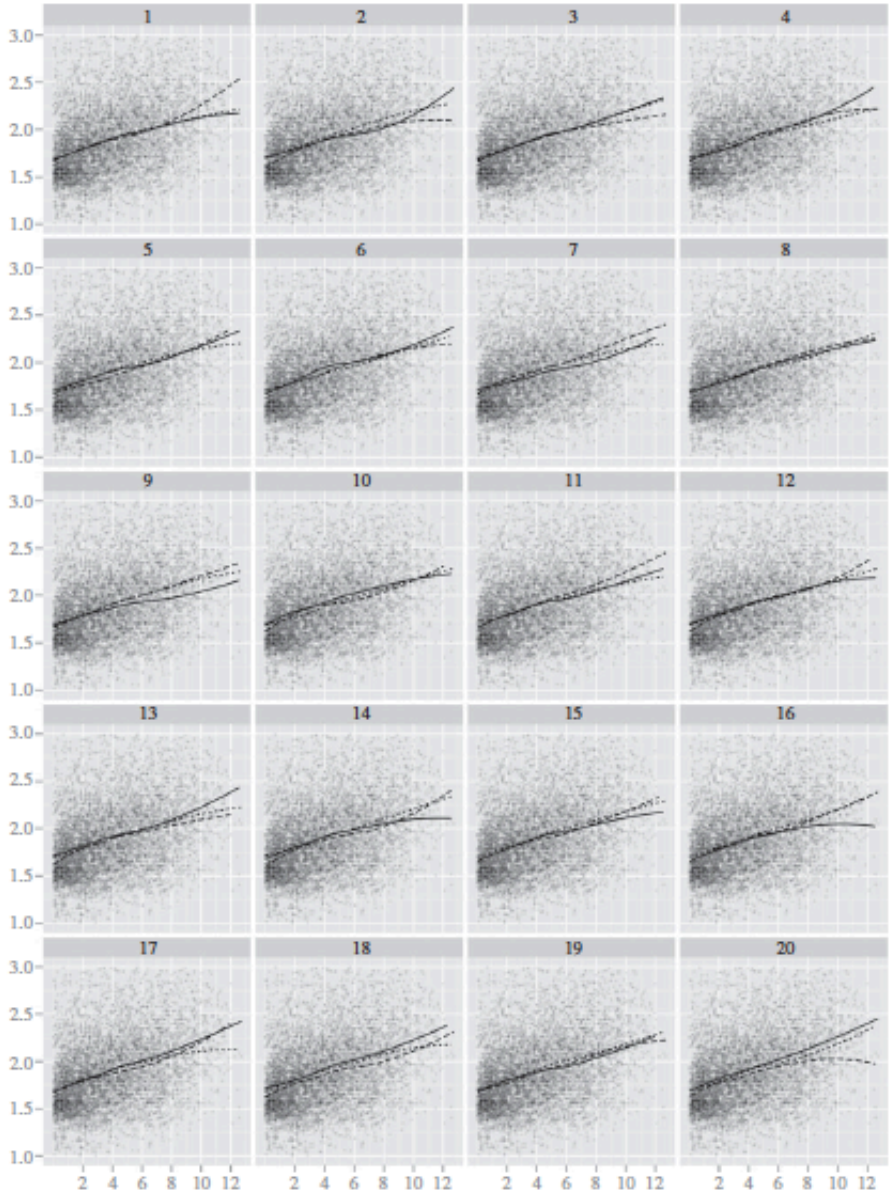
Which plots differ in the columns?



Which one is different?



Final one.



Solutions

Tipping data:	Plot 11
Housing price vs. Climate-terrain:	Plot 14
HSBC daily stock returns:	Plot 3 & Plot 3
Gene expression data:	Plot 1
Wage vs. Work experience:	Plot 9

Visual Inference: The Line-Up Protocol

To conduct visual inference with the line-up we have to:

1. Identify the question the plot is trying to answer or the pattern it is intended to show.

2. Formulate a null hypothesis (usually this will be H_0: "There is no pattern in the plot.")

3. Generate a null datasets to visualize (e.g. permutation of variable values, random simulation)

Some Graphical Displays and Patterns

Histogram: "Is the distribution normal (or uniform or …)?"

Scatterplot: "Is there a correlation between x and y?"

Scatterplot with Color: "Are there points clustered by color?"

Line Chart: "Is there a change over time?"

Choropleth Map: "Is there a spatial pattern?"

Null Hypothesis: "There is no normal distribution/relation/clustering/change/spatial pattern."

Example: Interviewer selection effects in the GLES

The GLES (German Longitudinal Election Study) is based on face-to-face interviews of the general population.

Researchers were worried that some of their interviewers might selectively contact households:

e.g. avoid low income areas and/or areas with high shares of foreigners.

The final methodological report did not give any information on this problematic interviewer behavior

What would we expect to see if interviewers indeed were to avoid certain areas?

Interviewer selection effects in the GLES

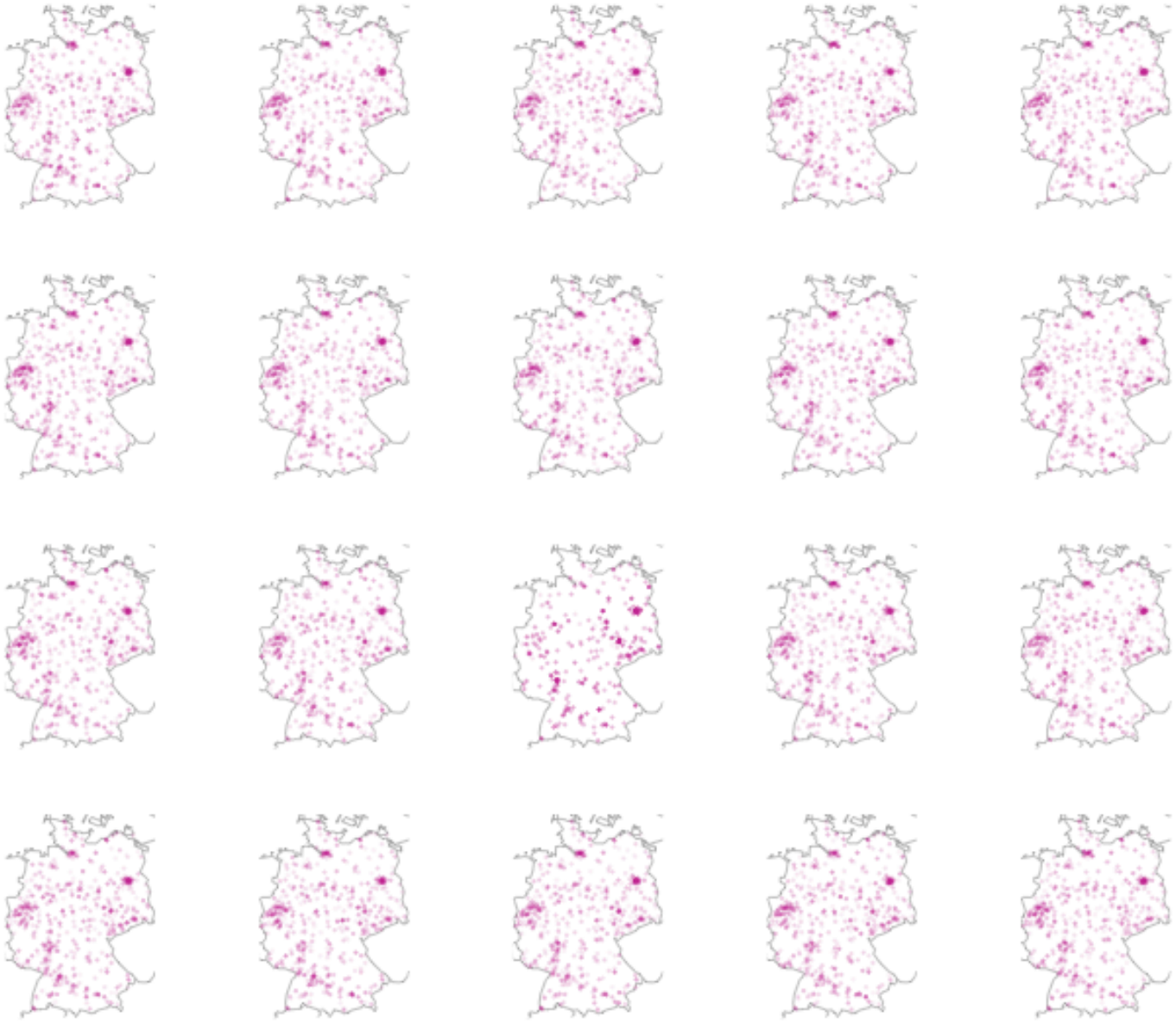
Question and graphical display: "Is there a spatial pattern in a map of interviewer contact behavior?"

Null hypothesis: "There is no spatial pattern in interviewer contact behavior: location and behavior are independent."

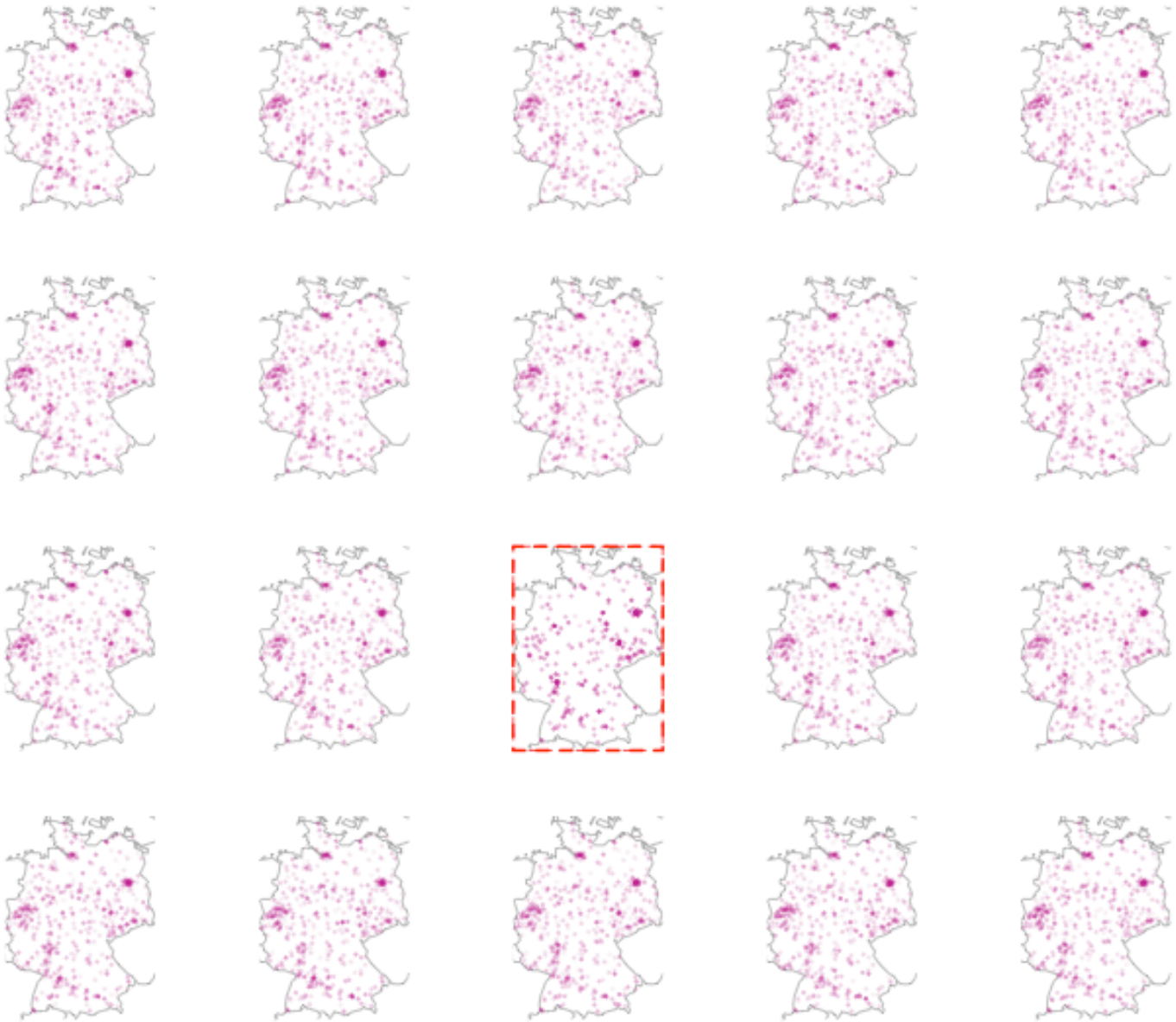
Generate null data sets: Just randomly permute the variable column for contact behavior

id4	status	g_lat	g_lon
2383	Interview completed	49.46432	11.08365
9633	Contact w/o Interview	53.48566	10.21064
9438	No Contact	51.35725	11.09774
14829	Contact w/o Interview	52.49620	13.51654
2671	Interview completed	51.96190	10.01162
1785	Interview completed	48.18359	10.82689

Visual Inference for “No Interviewer Contact”: Comparing the True Plot to 19 Null Plots



Visual Inference for “No Interviewer Contact”: Comparing the True Plot to 19 Null Plots



How to do it in R

Code interviewer behavior by color

```
data$col <- ifelse(data$status=="No Contact", "maroon3",
"darkolivegreen2")
```

Generate random plot placement

```
placement <- sample((1:20), 20)
layout(matrix(placement, 4, 5))
```

Generate 19 null plots

```
par(mar=c(.01, .01, .01, .01), oma=c(0, 0, 0, 0))
for(i in 1:19){
```

Randomize order

```
  random <- sample(c(1:15591), 15591)
```

Plot

```
  map(database="worldHires", fill=F, col="darkgrey", xlim=c(6, 15),
    ylim=c(47.3, 55))
  points(data$g_lon, data$g_lat, cex=.1, pch=19, col=data$col[random])}
```

How to do it in R

Add true plot

```
map(database="worldHires", fill=F, col="darkgrey", xlim=c(6, 15),
    ylim=c(47.3, 55))
points(data$g_lon, data$g_lat, cex=.1, pch=19, col=data$col)
```

Reveal true plot

```
box(col="red", lty=2, lwd=2)
which(placement==20)
```

Lab Exercise

Use the data set „`slop_2009_agg_example.dta`“ and produce some scatter plots to visualize relations between the variables.

Conduct visual inference to test whether the patterns you find „are actually there“.

Now, use any data set you like and construct line-ups to test whether visual patterns could have been produced by chance.

Make sure to try out different graphical formats.

Visualizing Statistical Models

Exploratory Data Analysis (EDA): Visualize and explore raw data

Exploratory Model Analysis (EMA): Visualize and explore cooked data: statistical models

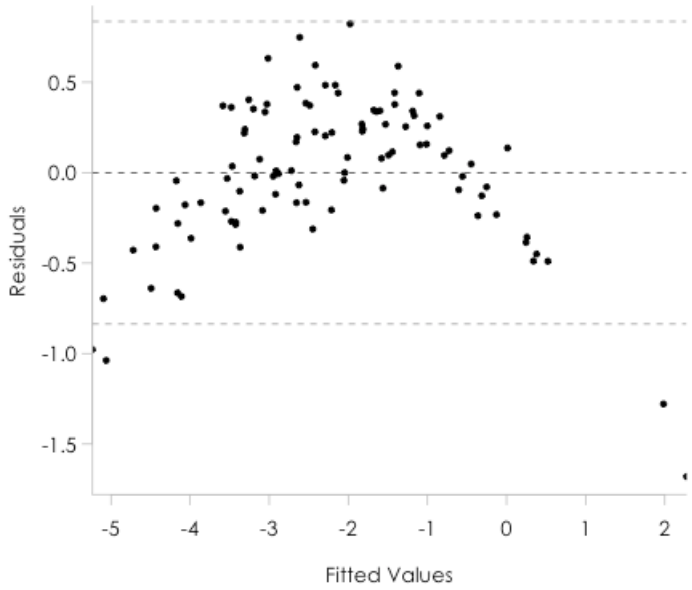
Several Uses:

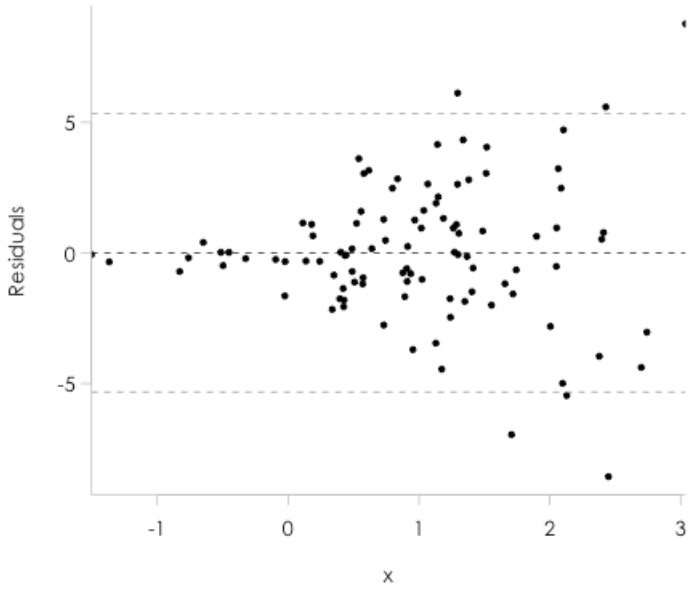
- Informal Precursor to More Complex Models
- Substitute to Complex Models if Assumptions Don´t Hold
- Model Checking and Diagnostics
- Sensitivity and Robustness Testing
- Presentation of Results

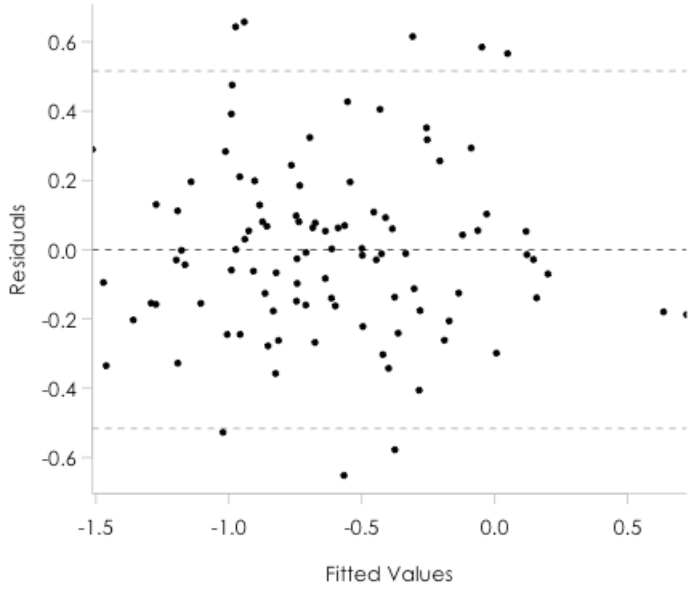
Quantities of Interest in EMA (adapted from Wickham et al. 2015)

Model Level:	M measures of model fit
Model-Estimate Level:	$M \times k$ coefficient estimates, standard errors, t-values
Estimate Level:	k estimate summaries over many models
Model-Observation Level:	$M \times N$ residuals and influence measures
Observation Level:	N original data and summaries of residual behavior

Classic Regression Diagnostics Using Residual Plots







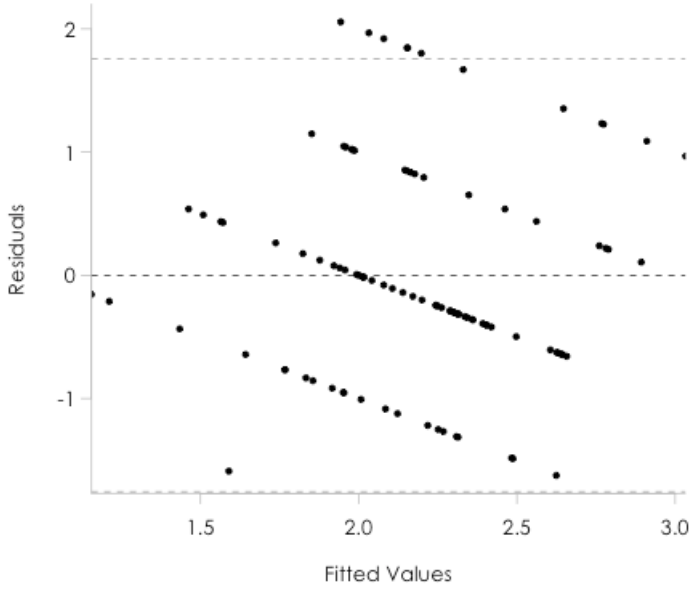


Table 4: Results from three Bayesian Hierachical Ordered Probit Models for Attitudes toward Muslims

	Too many Muslims (reversed)			Right to wear headscarf			Right to build minarets		
	Mean	*5%*	*95%*	*Mean*	*5%*	*95%*	*Mean*	*5%*	*95%*
Religious Support Index	-0.24	[-0.37	-0.10]	-0.09	[-0.21	0.03]	-0.15	[-0.28	-0.01]
% Muslims	-0.13	[-0.25	0.00]	0.26	[0.14	0.37]	0.06	[-0.06	0.19]
Integration Regime	0.08	[-0.04	0.21]	-0.13	[-0.25	-0.01]	-0.07	[-0.21	0.04]
Female	0.08	[-0.04	0.19]	0.09	[-0.02	0.20]	-0.09	[-0.20	0.02]
Age	-0.39	[-0.51	-0.28]	-0.10	[-0.20	0.02]	0.03	[-0.09	0.15]
Education	0.55	[0.44	0.66]	0.38	[0.27	0.50]	0.57	[0.46	0.68]
Left-Right-Ideology	-0.89	[-1.00	-0.76]	-0.75	[-0.87	-0.63]	-1.16	[-1.29	-1.01]
Close to SVP	-0.61	[-0.78	-0.45]	-0.38	[-0.54	-0.23]	-0.75	[-0.93	-0.56]
Urban Living Area	0.12	[0.00	0.25]	0.04	[-0.07	0.18]	0.19	[0.05	0.32]
Church Attendance	-0.07	[-0.20	0.06]	0.06	[-0.07	0.19]	-0.04	[-0.18	0.11]
Protestant	-0.06	[-0.23	0.12]	0.03	[-0.13	0.21]	0.02	[-0.17	0.19]
Catholic	-0.08	[-0.26	0.11]	-0.08	[-0.25	0.09]	0.03	[-0.16	0.21]
Other	-0.22	[-0.52	0.07]	0.04	[-0.25	0.34]	0.10	[-0.20	0.41]
Intercept	0.79	[0.62	1.00]	0.48	[0.30	0.66]	0.30	[0.11	0.48]
τ_1	0	0	0	0	0	0	0	0	0
τ_2	0.88	[0.80	0.96]	0.59	[0.53	0.65]	0.59	[0.53	0.65]
τ_3	1.53	[1.43	1.61]	0.86	[0.79	0.92]	0.88	[0.81	0.95]
τ_4	2.11	[2.00	2.22]	1.79	[1.70	1.89]	1.77	[1.66	1.87]
σ_α^2	0.01	[0.00	0.02]	0.00	[0.00	0.01]	0.00	[0.00	0.01]
σ_y^2	1	1	1	1	1	1	1	1	1

Tables of Coefficients are Usually Not Enough

Work ok...

...when models are linear and additive (even then graphs are better)

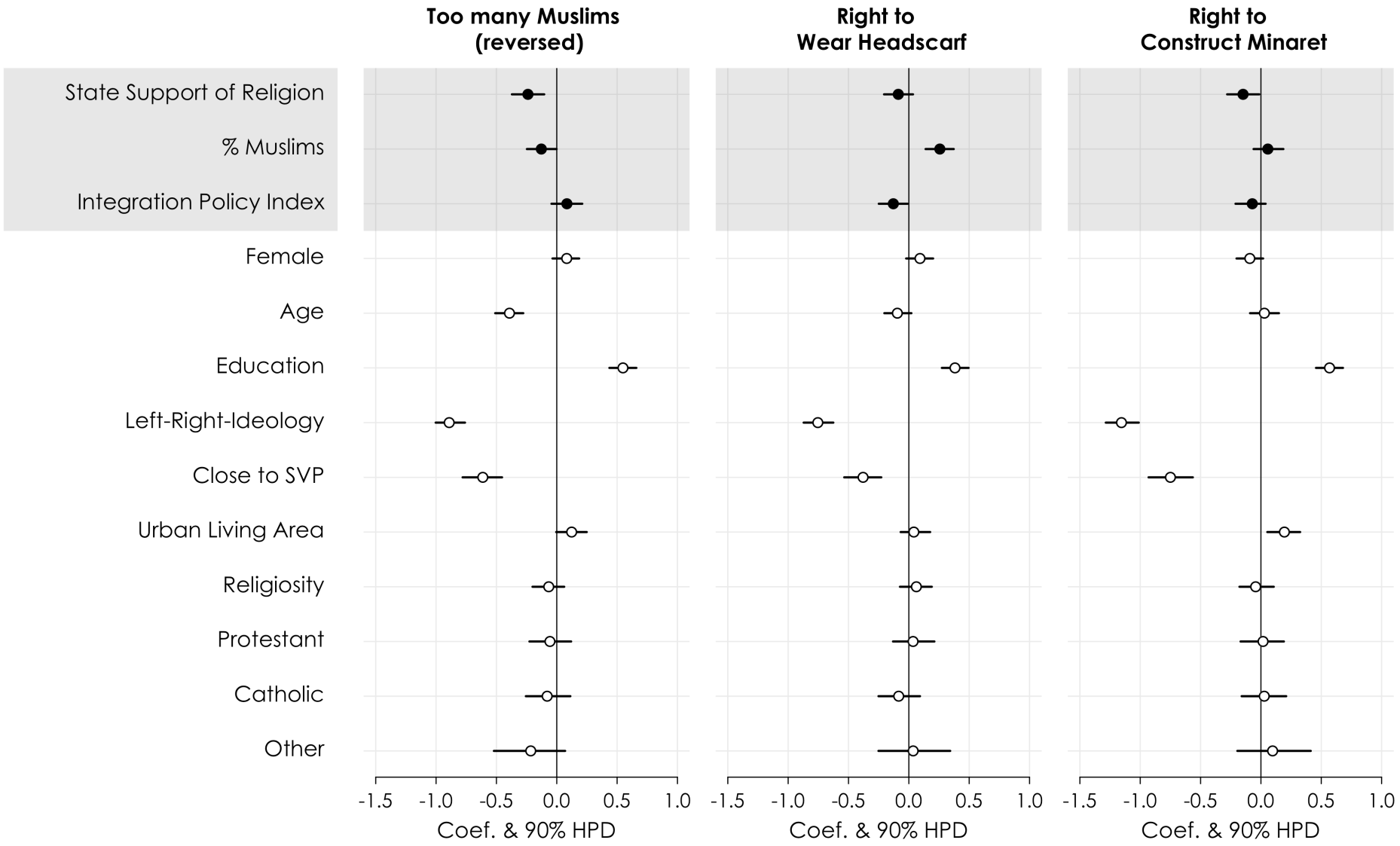
Less or not at all informative for...

...non-linear relationship between x and y (x^2, log(x), etc.)

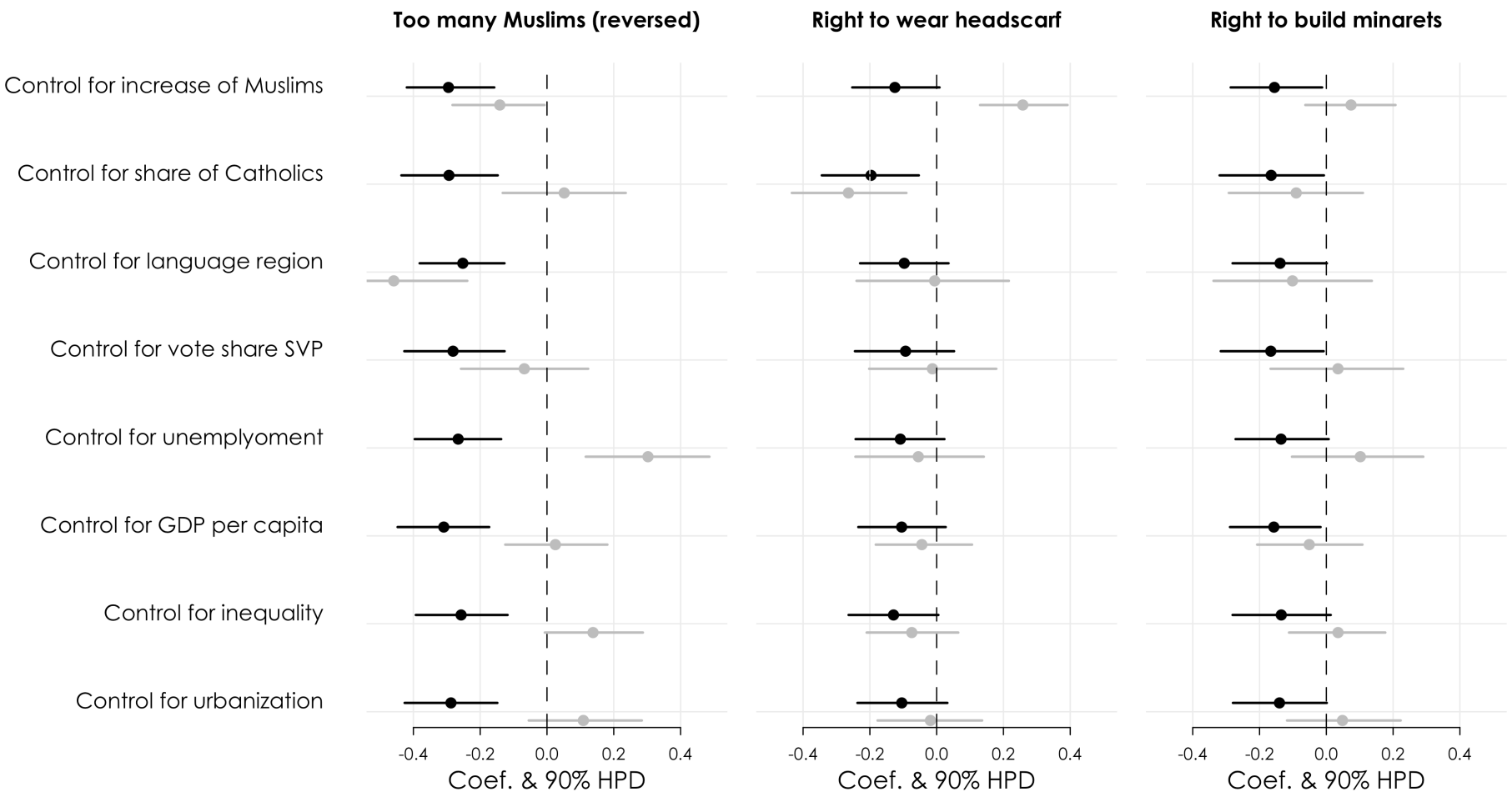
...interaction effects

...models for categorical data

Coefficient Plots



Coefficient Plots Comparing Two Groups of Coefficients



That's (Roughly) How I Produced the Plots

Create a plotting function

```
CoefPlot <- function(m.fit, title){

  par(mgp=c(1.5, 0.35, 0), tck=-0.02, mar=c(3,.5,3,.5))

  coefs <- coef(m.fit)

  plot(coefs, length(coefs):1, pch=19, main=title, axes=F, ylab="",
    xlab="Coef. & 95% CI", xlim=c(-1.5, 1), cex=.9, cex.main=1)

      abline(h=length(coefs):1, col="grey92", lwd=.5)
      abline(v=seq(-2, 2, by=.5), col="grey92", lwd=.5)
      axis(1, lwd=.5, cex.axis=.9, cex.lab=.9)

  segments(coefs+1.96*se.coef(m.fit), length(coefs):1,
      coefs-1.96*se.coef(m.fit), length(coefs):1)

     abline(v=0, lty=1, lwd=.5)
     rect(-2, 10.5, 2, 14, border=NA, col=rgb(0, 0, 0, .1))
     points(coefs, length(coefs):1, pch=19,col=c(rep("black",3),
     rep("white",10)), cex=.6)
}
```

That's (Roughly) How I Produced the Plots

Grid layout

```
par(mfrow=c(1,4))
```

Plot variable names

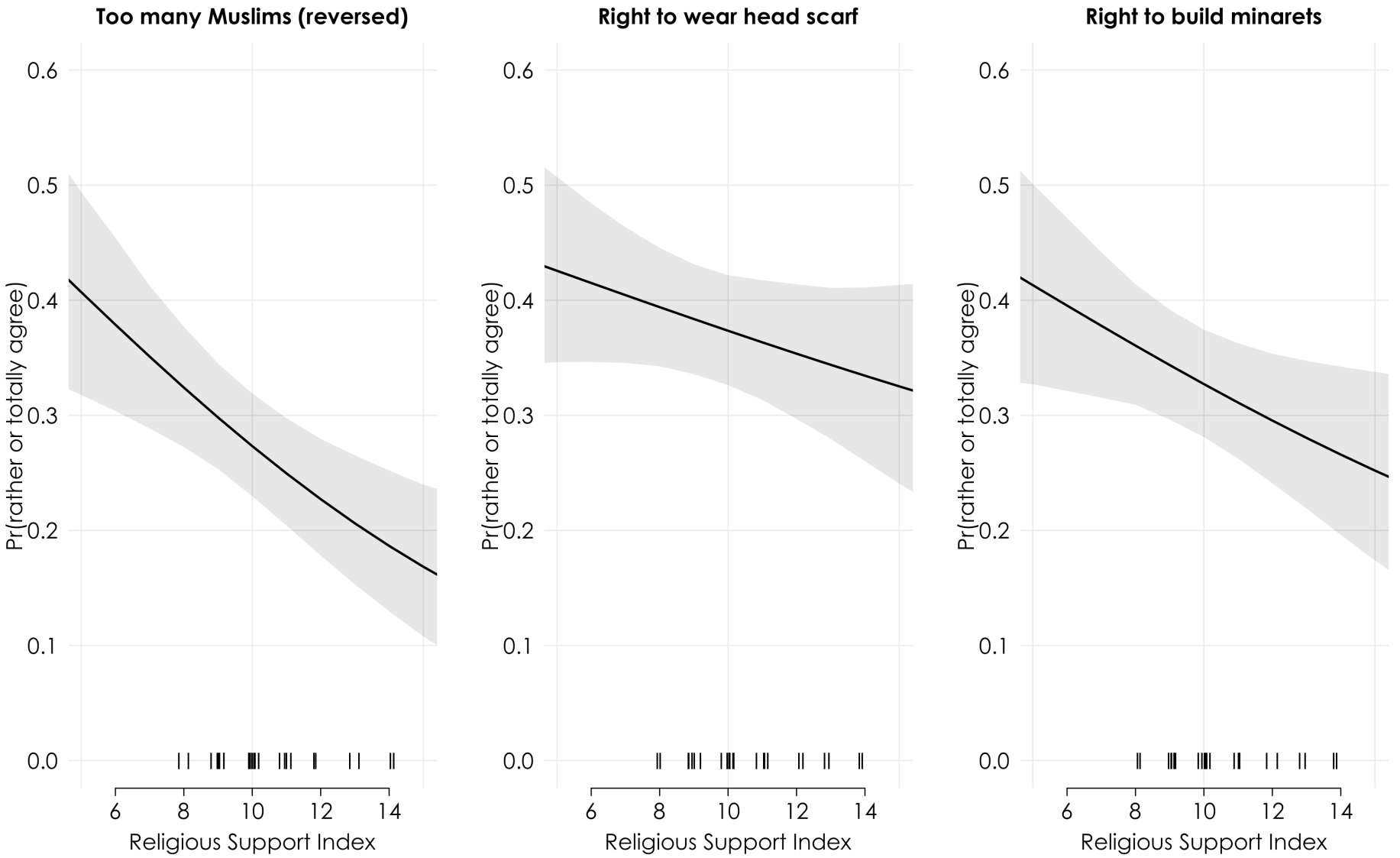
```
var.names <- c("State Support of Religion", "% Muslims", "Integration
Policy Index", "Female", "Age", "Education", "Left-Right-Ideology",
"Close to SVP", "Urban Living Area", "Religiosity", "Protestant",
"Catholic", "Other")

plot(rep(1, 13), c(13:1), ann=F, pch="", xlim=c(-18,1), axes=F)
  text(1, c(13:1), var.names, pos=2, cex=1)
  rect(-20, 10.5, 1, 14, border=NA, col=rgb(0, 0, 0, .1))
```

Now add the coefficient plots

```
CoefPlot(mcmc.1, "Too many Muslims \n (reversed)")
CoefPlot(mcmc.2, "Right to \n Wear Headscarf")
CoefPlot(mcmc.3, "Right to \n Construct Minaret")
```

A quick and dirty option is to just use the `coefplot()` function in the `arm` package.

Predicted Probability Plot



Using Graphs as Precursor or Substitute for More Complex Models

„Graphs of parameter estimates can be thought of as proto-multilevel models in that the graph suggests a relation between

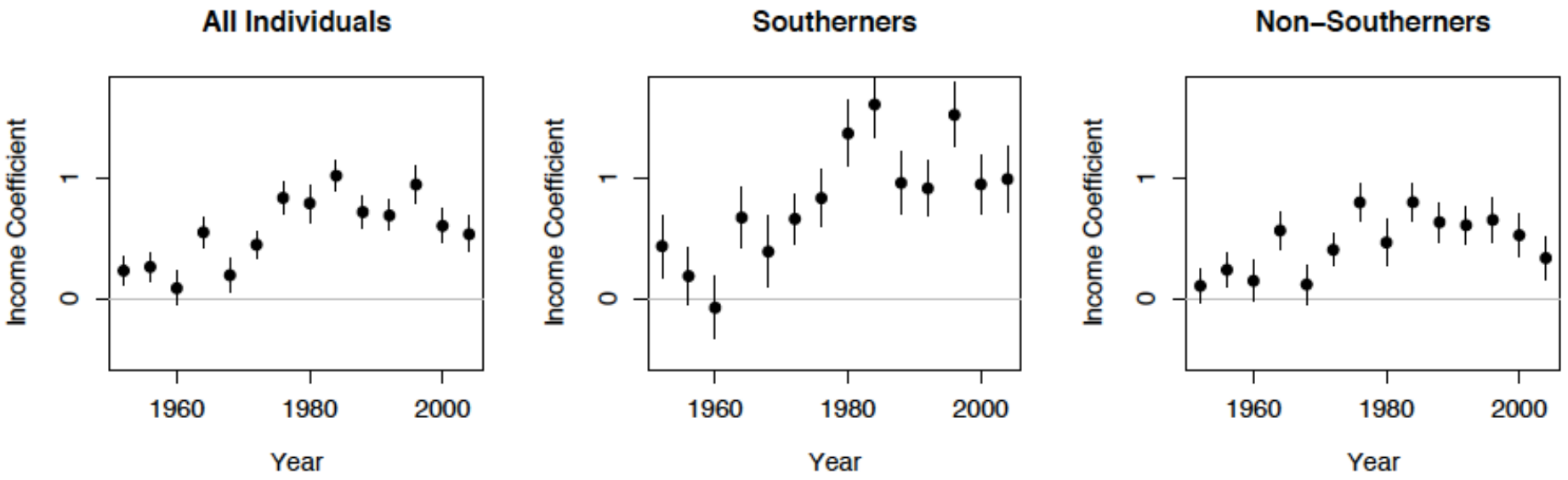
the y-axis (the parameter estimates being displayed) and

the x-axis (often time, or some other index of the different subsets being fit by a model).

These graphs contain an implicit model […] the same way that any scatterplot contains the seed of a regression or correlation model“ (Gelman & Hill: 553)

„The method of repeated modeling, followed by time-series plots of estimates, is sometimes called the ‚secret weapon' because it is so easy and powerful but yet is rarely used a as data-analytic tool“ (Gelman & Hill: 73).

The Secret Weapon



Three easy steps:

1. Run separate regression model for each subset j.

2. Extract coefficients (and standard errors) of interest.

3. Visualize coefficients (and standard errors) of interest.

Example

Does the effect of education on political participation depend on the educational context?

H1: Education is related to participation because it is related to relative social status.

→ Individual education should matter more in contexts with low average education.

H2: Education is related to participation because it is related to civic skills and resources.

→ Individual education should matter the same across all contexts.

Data

«Swiss Volunteers Monitor Communities»

N = 4399 respondents in J = 60 Swiss communities

Outcome:

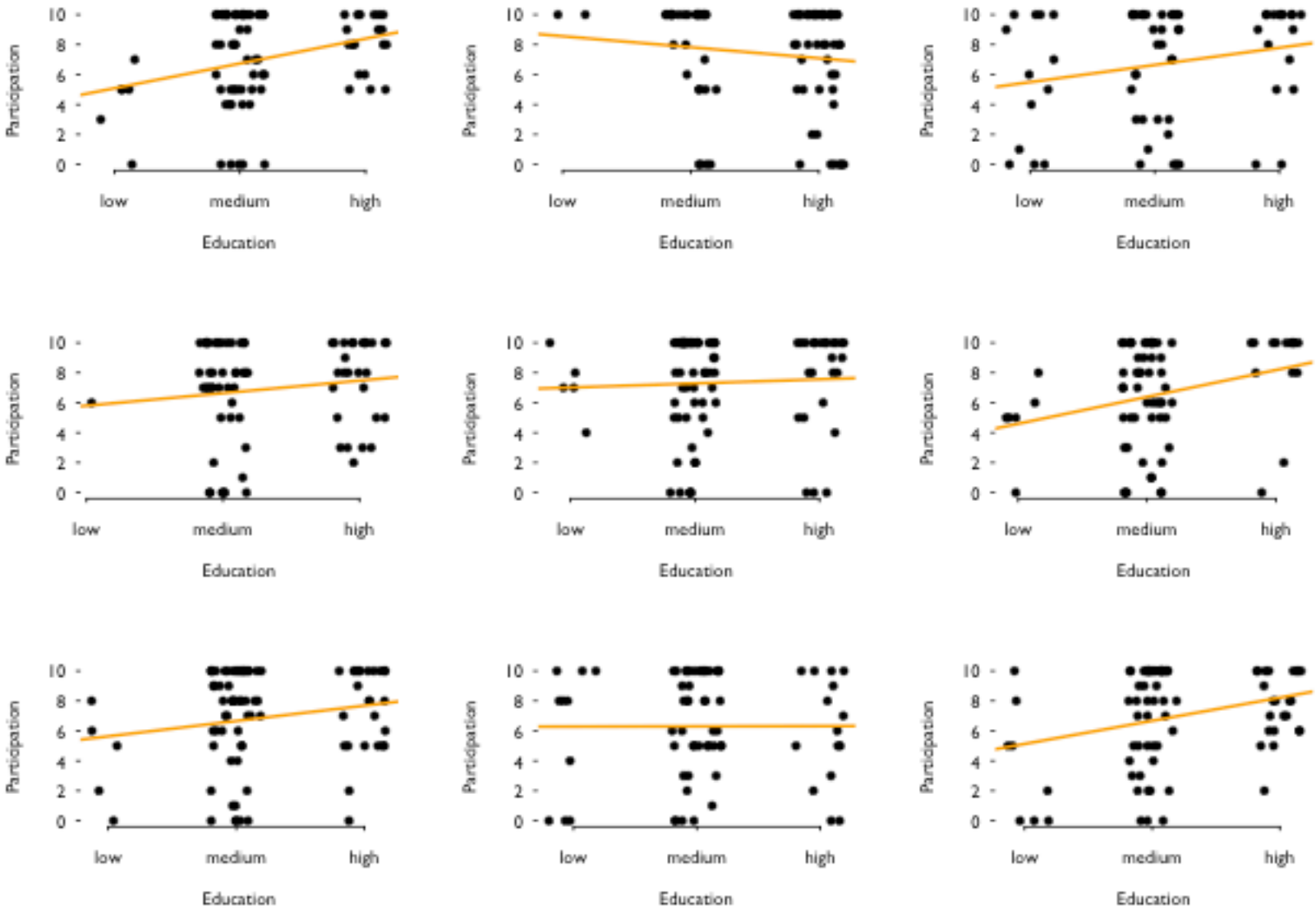
Out of ten times, number of times participated at local referenda (0-10)

Predictors:

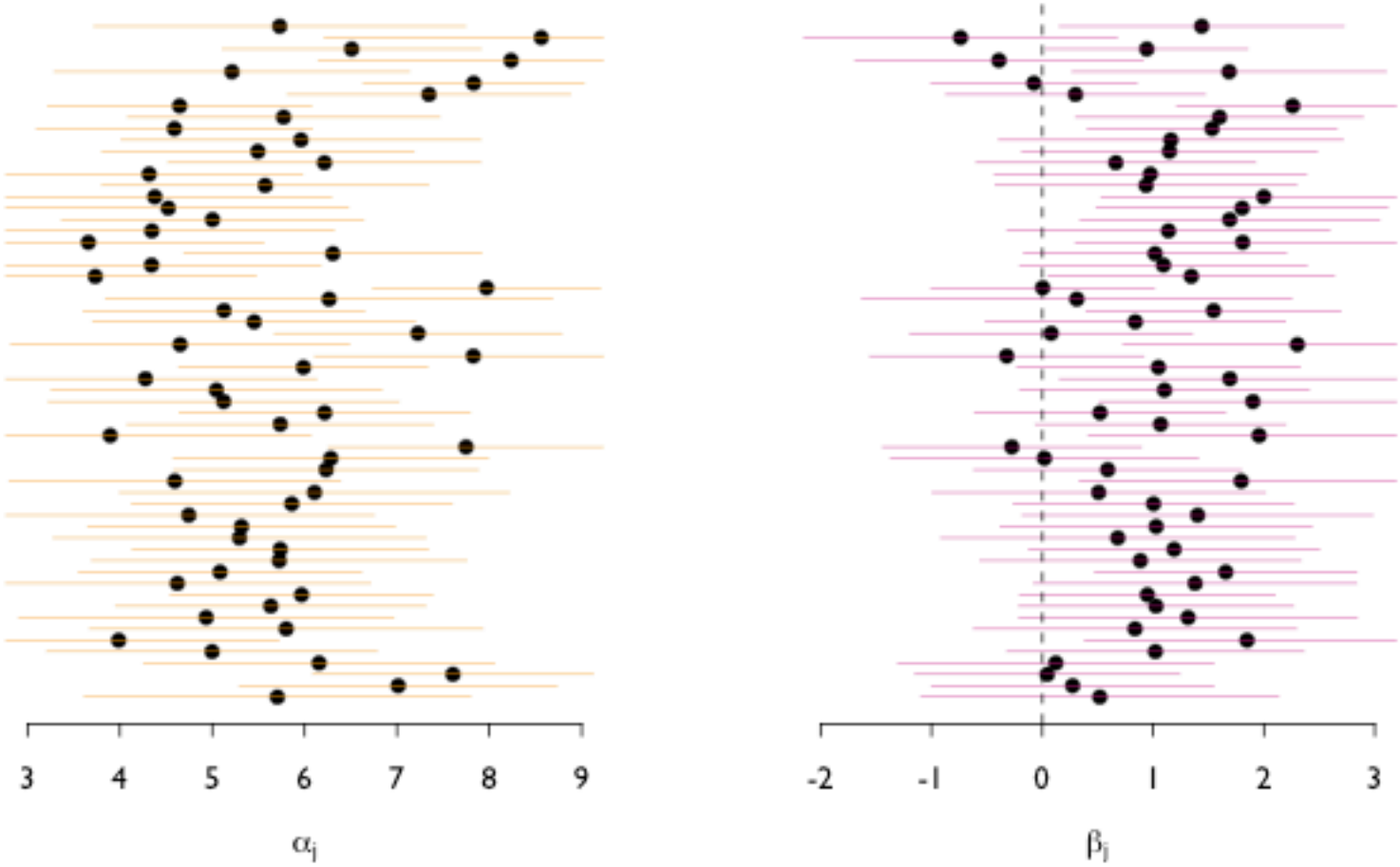
Individual level of education (1-3)

Share of highly educated in community context (in %)

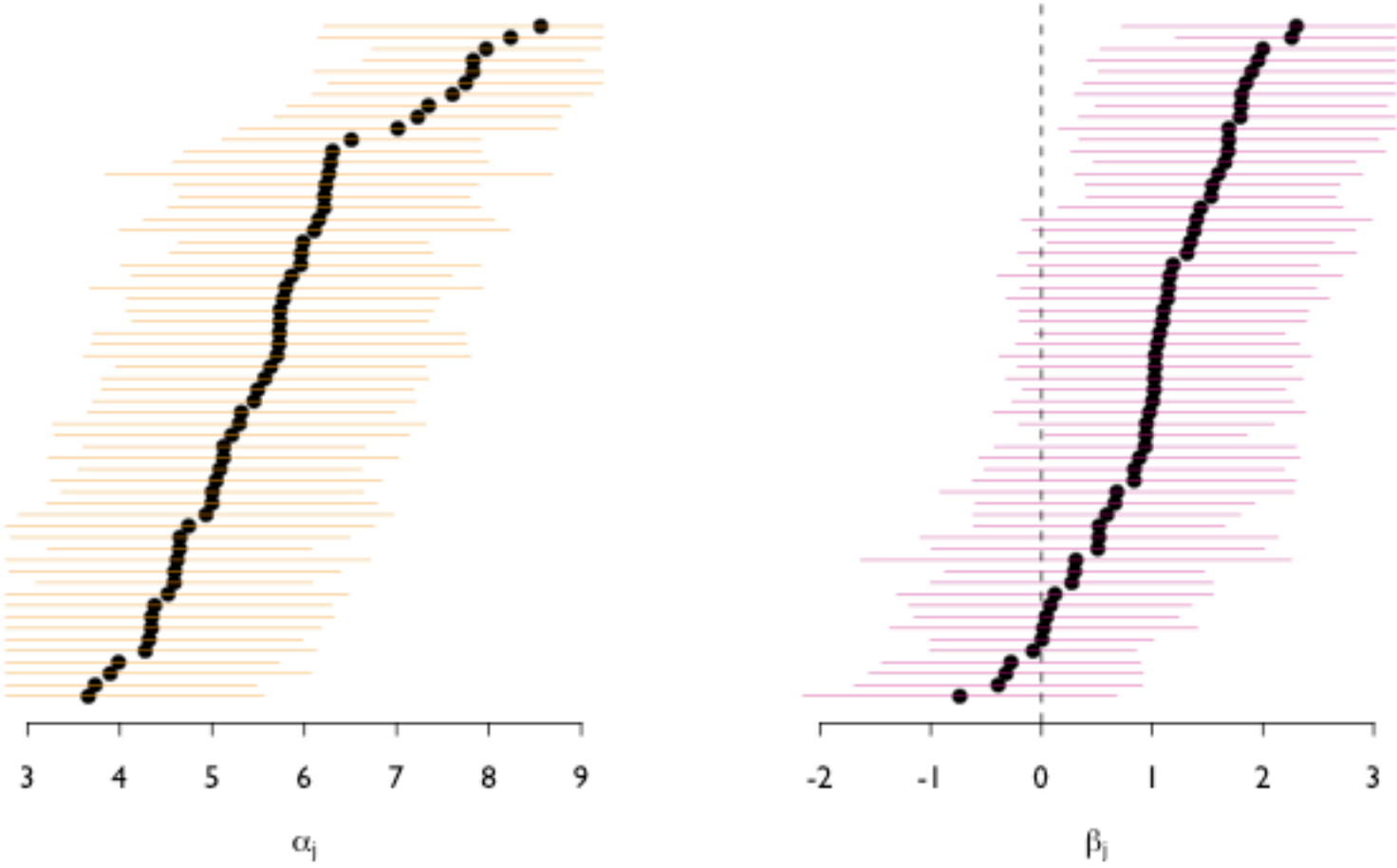
Scatterplots of 9 Randomly Sampled Communities



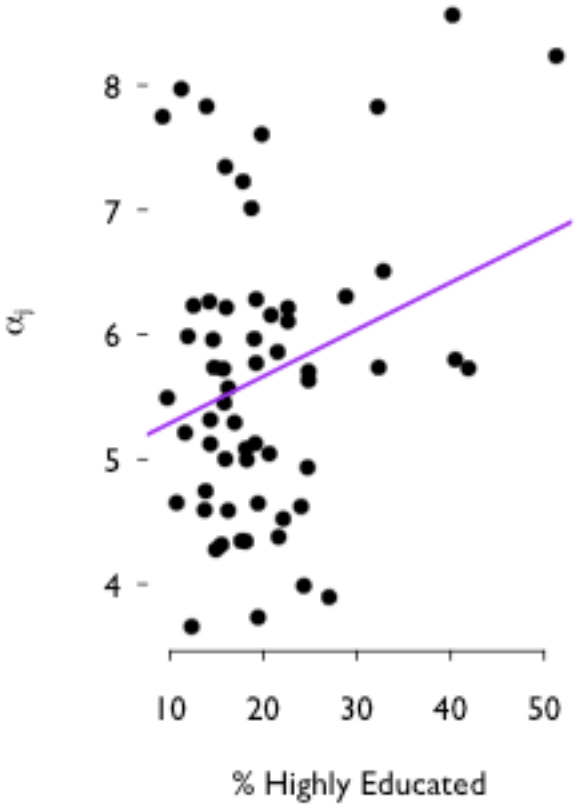
Plotting the Coefficients of 60 Regression Models

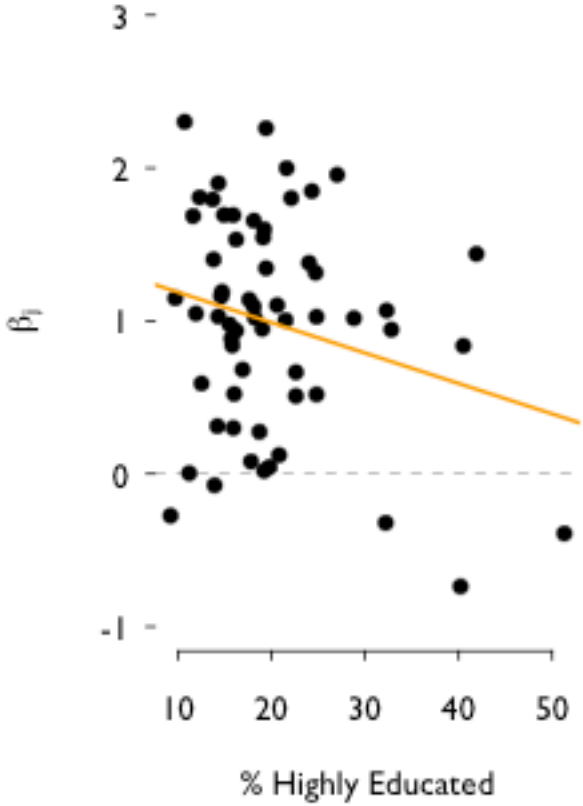


Plotting the Coefficients of 60 Regression Models (Sorted)

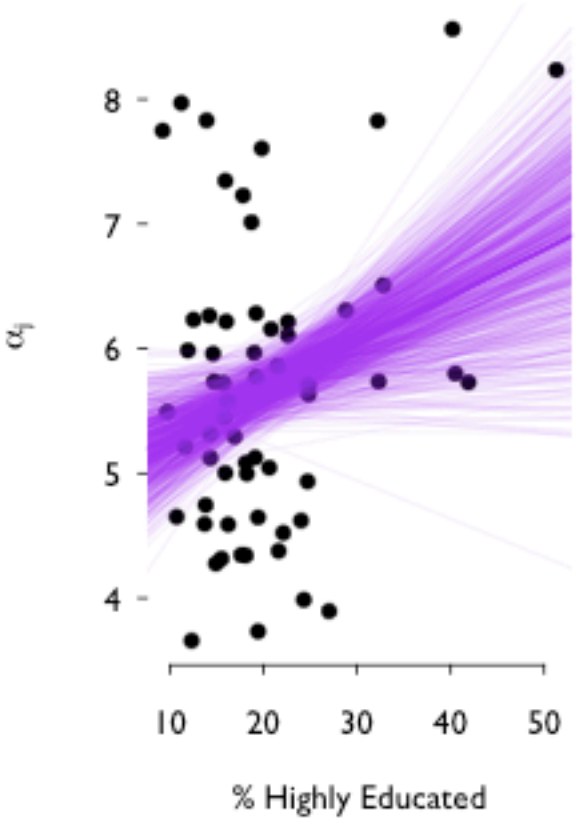


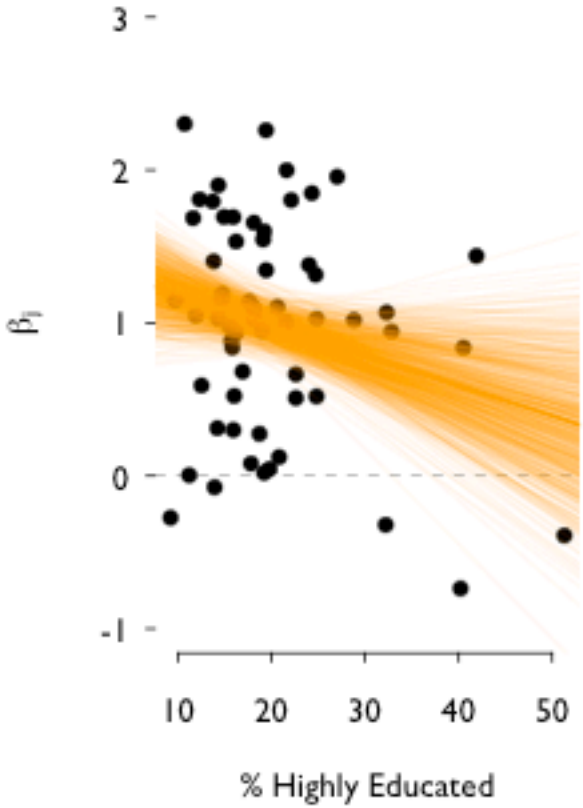
Relating Regressions Coefficients to Context Data





Relating Regressions Coefficients to Context Data





Sensitivity Analysis of the Determinants of Democracy

„[t]he lack of robust results is perhaps the most notable characteristic of this relatively large literature“ (Hegre et al. 2014:2)

Subset of Hegre at al.’s 2014 data

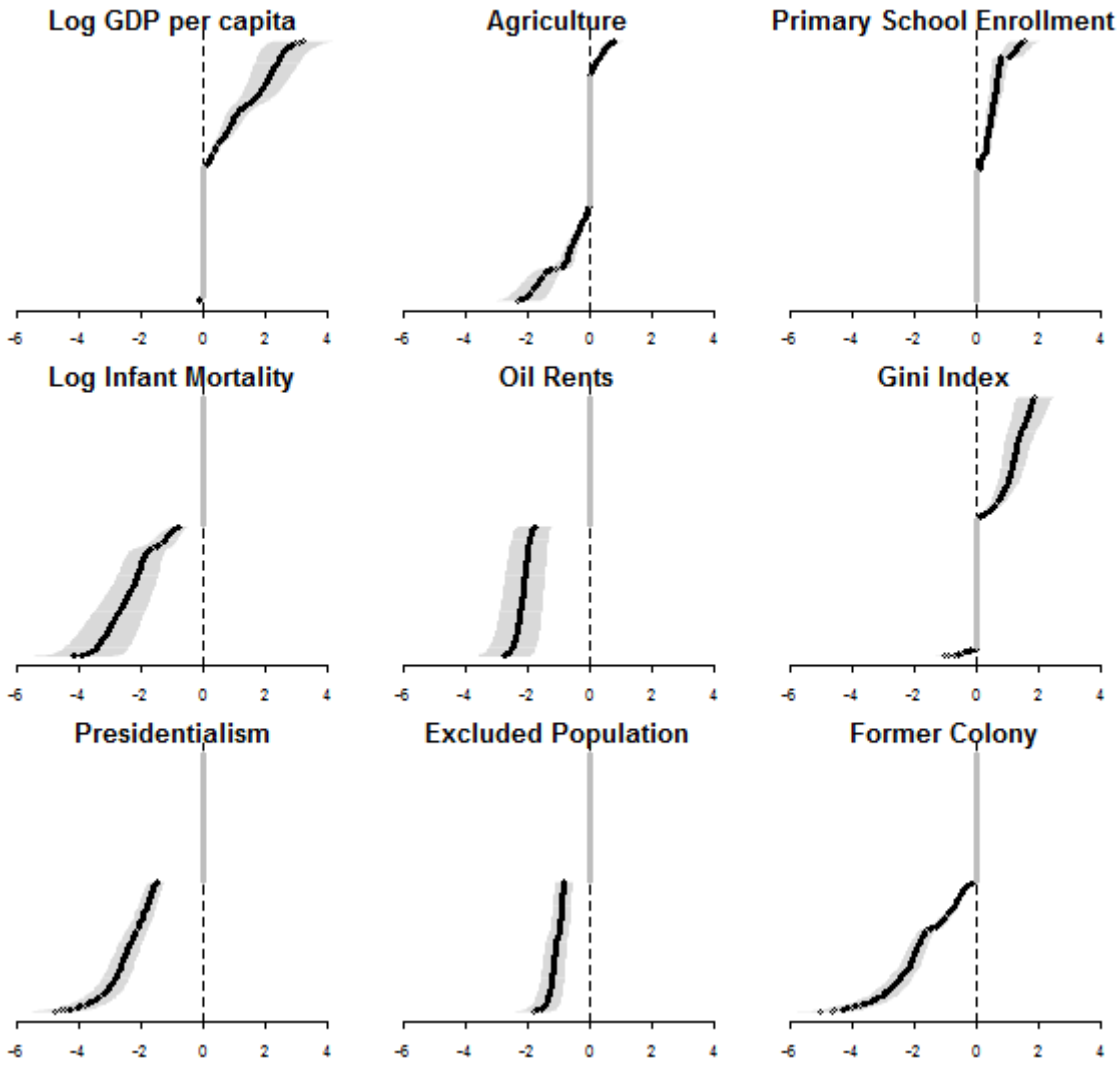
N = 168 countries, 1990-2010

Dependent Variable: Polity 2 scores (-10 to 10)

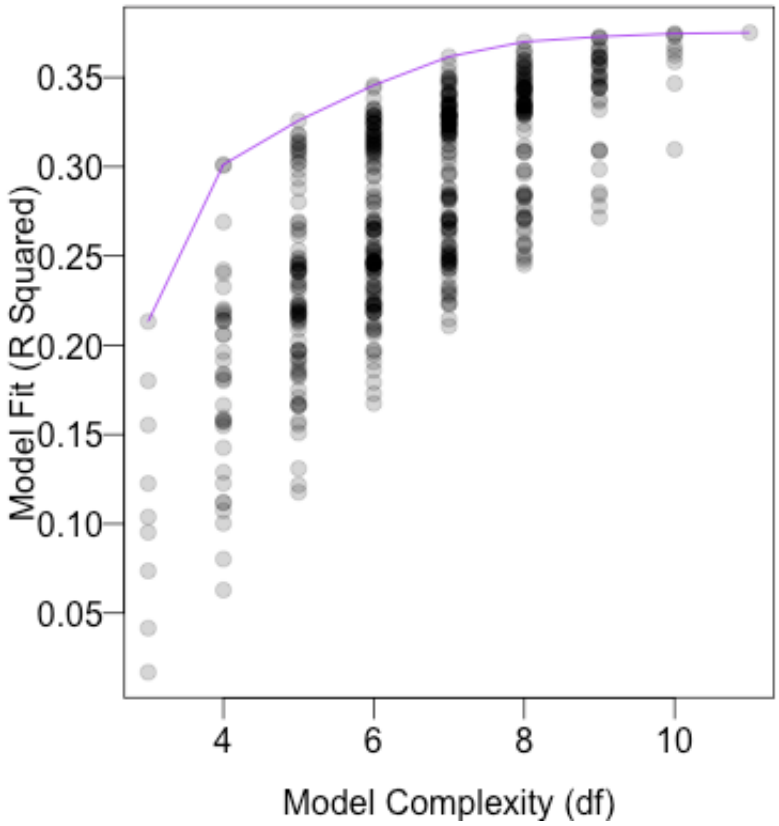
9 Determinants: log GDP per capita, agriculture, primary school enrollment, log infant mortality, oil rents, Gini index of income inequality, presidentialism, former colony, population excluded from power

Model Space: 2^9-1=511 simple linear regression models

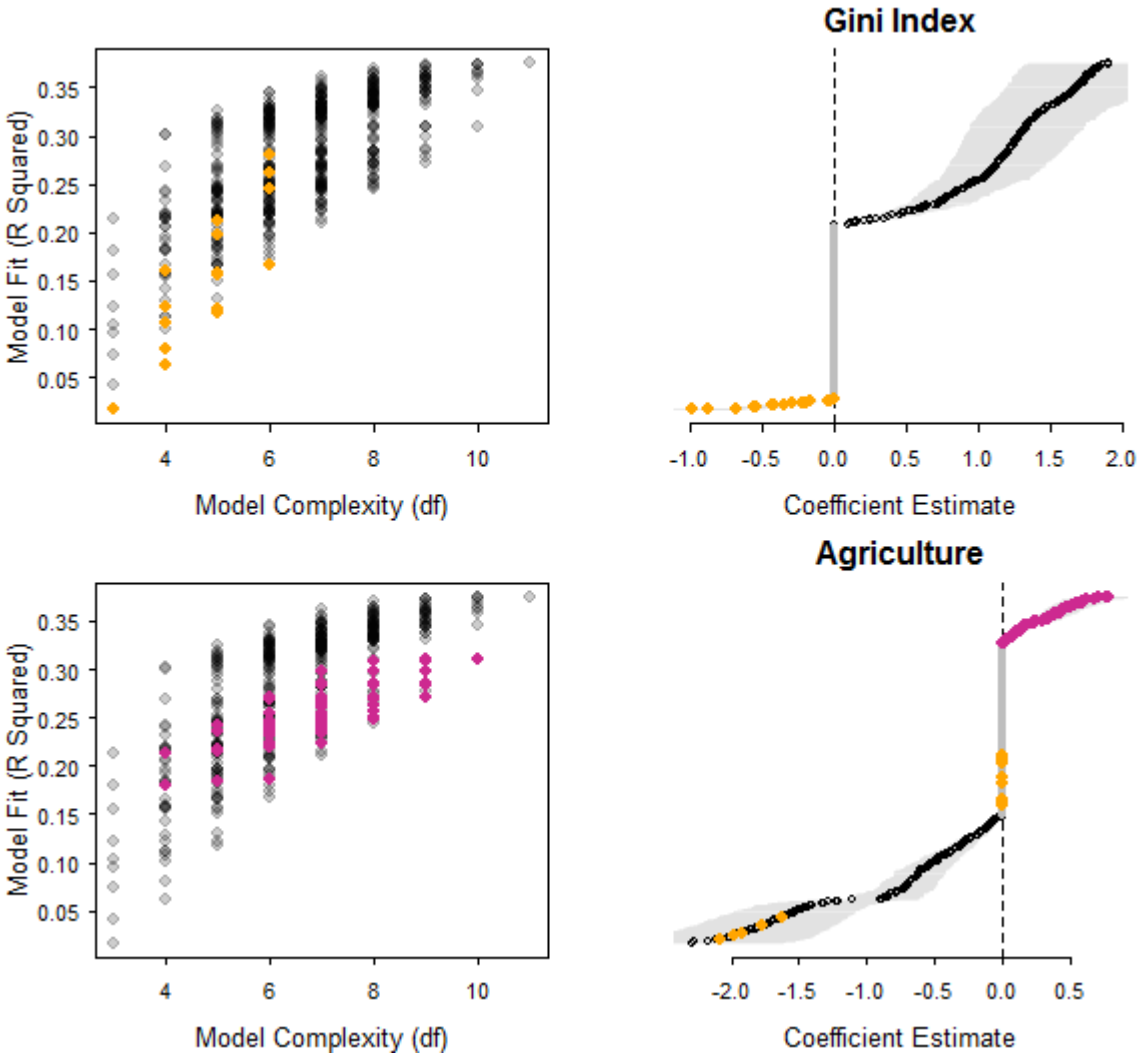
What Does Robustness „Look“ Like?



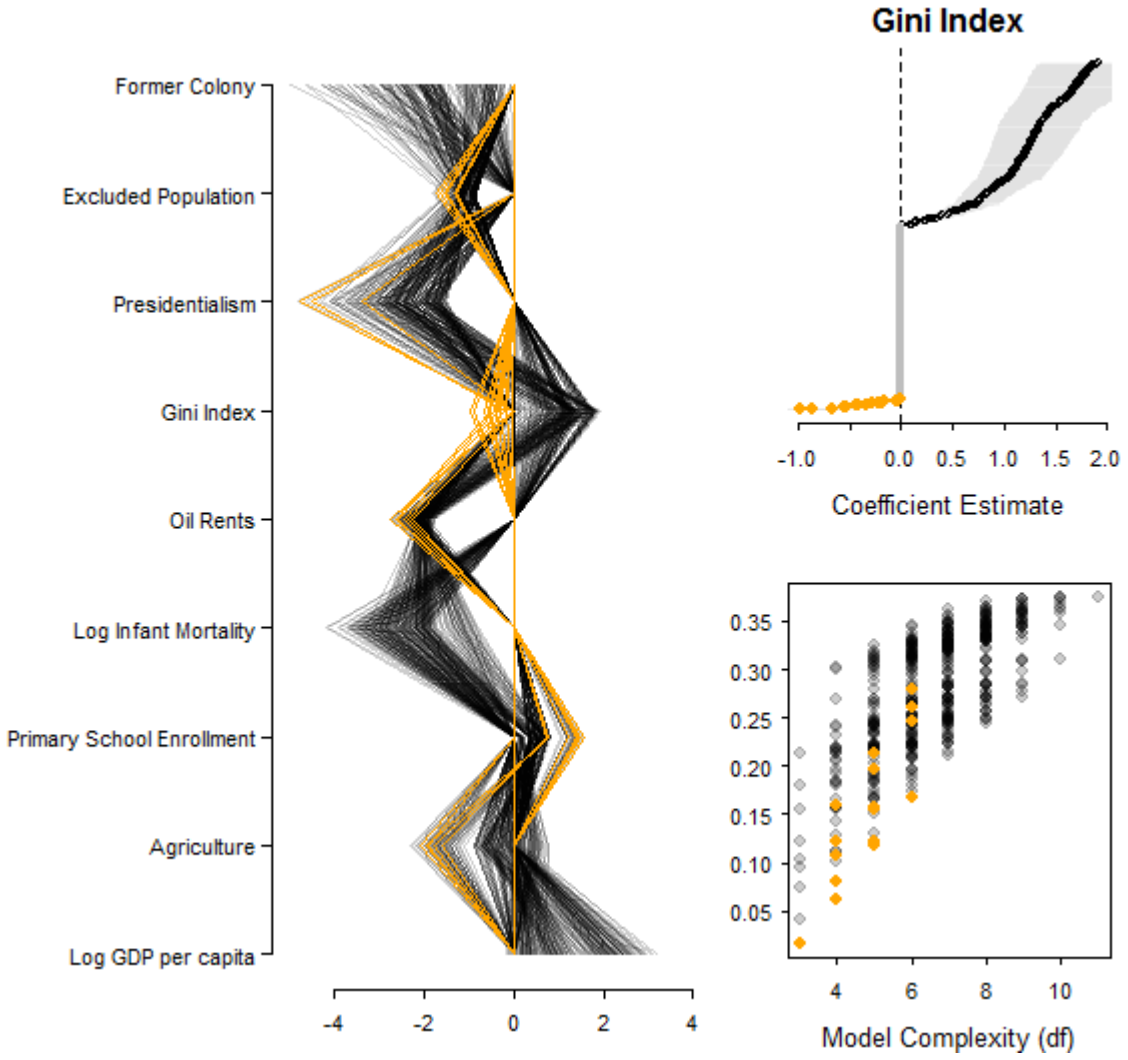
Which Specifications Produce Non-Robust Estimates?



Linking Model Level and Estimate Level



Linking Model-Estimate Level, Model Level and Estimate Level

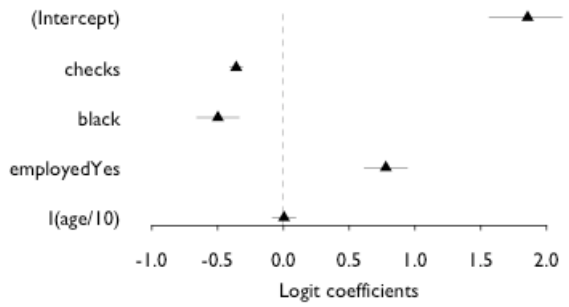


Lab Exercise: Logistic Regression Models of Marijuana Arrests

Data on police treatment of N = 5226 individuals arrested in Toronto for simple possession of small quantities of marijuana. The data are part of a larger data set featured in a series of articles in the Toronto Star newspaper.

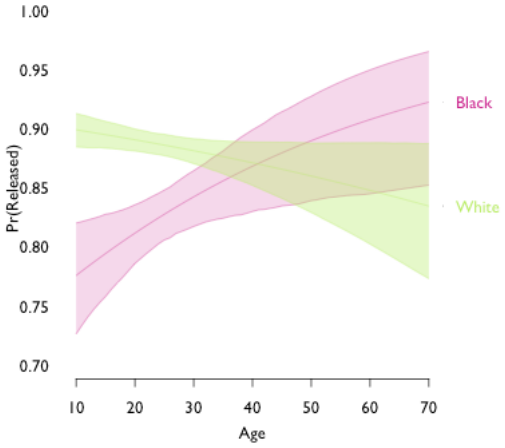
Run some models and then visualize the results.

Again, you can follow the code and data I provided, but be sure to experiment with different model specifications!



coefficients

predicted probabilities



interaction effects