finding a story to tell: graphical research methods

Robert.Chung@berkeley.edu

October 2024

analytical graphics

- not presentation graphics, not exactly statistical graphics, not quite exploratory graphics

  presentation graphics tell your story; analytical graphics help you figure out what your story is

  analytical graphics are "work product"

- not always refined enough for publication or presentation

  Tufte? A lot of his techniques focus on effectively communicating quantitative findings. We'll focus on steps before that: uncovering interesting stories and questions in data

  Very valuable for grad students and early career researchers

analytical graphics are for analysis

- often analytical graphics are used not to prove hypotheses but to help generate them

  we don't always answer questions; we use graphical techniques to help us ask new ones

- often, the audience is YOU

  visualize differences and contrasts

  across time
  across places
  across treatments or policies
  across conditions

  there are tips, tricks, and techniques that help you in visualization

what if you already have a question?

- no problem. sometimes analytical graphics can help you focus on where to look, or to refine your question
- doesn’t replace theory, or your research question. You can do both: that’s allowed

basic approach

- maximize insight
- uncover underlying structure
- extract important variables
- detect outliers and anomalies
- develop (very) simple models
- not much testing; that's for later

what we'll do in these lectures

- some examples
- some principles
- some basic tricks
- some slightly more advanced tricks
- you'll have a chance to try out some of these tricks before next week

the three things we're looking for

- look for

    pattern

    unexpected pattern

    deviations from pattern

- these generate questions
- questions and how you address them are often the basis for papers or chapters or dissertations or careers
- "The data speak for themselves, but their voices are soft and sly" so we're looking for ways to amplify their voices

demography

- demography is the study of populations, their characteristics, relationships among characteristics, and how they change
- you probably already know how to examine characteristics; we'll look at ways to highlight relationships among the characteristics and how they change
- in particular, we'll often look for models to help us understand the relationships among characteristics

the purpose of models

- "The purpose of models is not to fit the data but to sharpen the questions" – Sam Karlin
- We'll use analytical graphics to help us sharpen questions

simple tools

- simple tools used intelligently (well, we can always dream) rather than complex tools used stupidly

  rules of thumb, not hard rules and regulations

- a handful of plots and a handful of tricks

  lots of specialty type graphs, but we try to avoid too many of them until we know our story

- xy plots are a hugely useful invention

  with one or two exceptions we'll focus mostly on ways to enhance xy plots

- decoding the language of graphs can be complicated, so we build on familiar beginnings

how graphing helps

- we can only make sense of a handful of numbers at a single time

  pages of dense tables are good for detail, evidence, and re-analysis but poor for understanding

- eye-brain is good at seeing patterns in large numbers of values

  though it can be fooled—we'll present some problems that can mislead the eye

- therefore

  use graphs when pattern is important

  use tables when exact details are important

  graphs and tables are complements, not replacements. (You can do both: that’s allowed)

apophenia

- apophenia is "the experience of seeing patterns of connections in random or meaningless data"
- we'll occasionally accept a little "type I error" when we're looking for interesting questions – as long as we back it up later with real confirmatory analysis

Anscombe's data

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

Anscombe's data

- same means, sd's, correlation, regression slope, fit

  mean(x1)=mean(x2)=mean(x3)=mean(x4) = 9

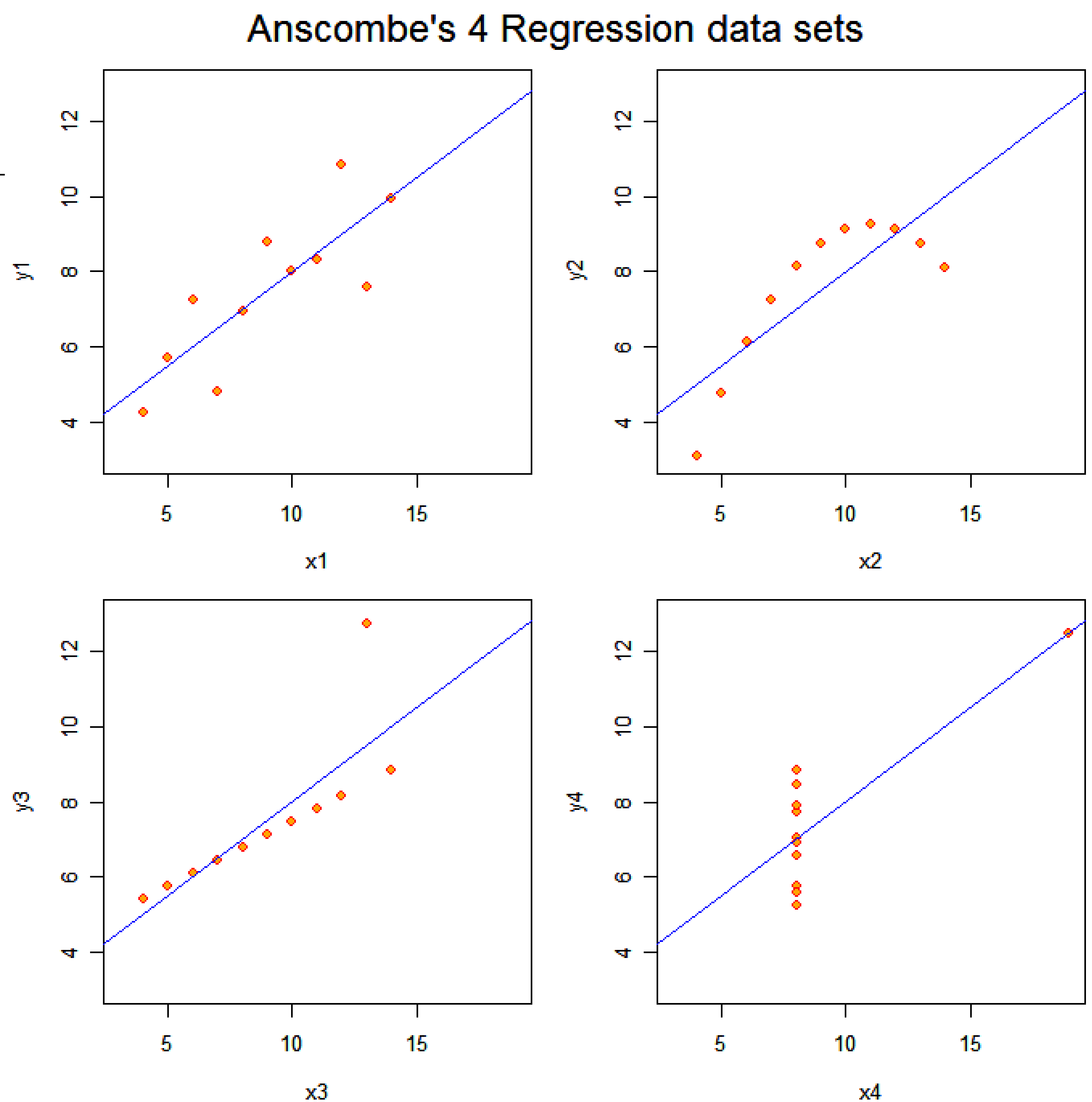
  mean(y1)=mean(y2)=mean(y3)=mean(y4) = 7.5

  sd(x1)=sd(x2)=sd(x3)=sd(x4) = 3.32

  sd(y1)=sd(y2)=sd(y3)=sd(y4) = 2.03

  r(x1,y1)=r(x2,y2)=r(x3,y3)=r(x4,y4) = 0.816

  y* = 3 + 0.5 x* with r^2 = 0.667
- so, conventional linear models make them look alike
- what will you see if you graph the data?


Anscombe's 4 Regression data sets
y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
5
10
15

the NJ Pick-It lottery

- each bettor selected a 3-digit number between 0 and 999
- each ticket cost 50 cents
- all bettors who held the winning number split the prize money. The size of the prize depended on selecting the winning number and on the number of players who chose that number
- what would you want to know?

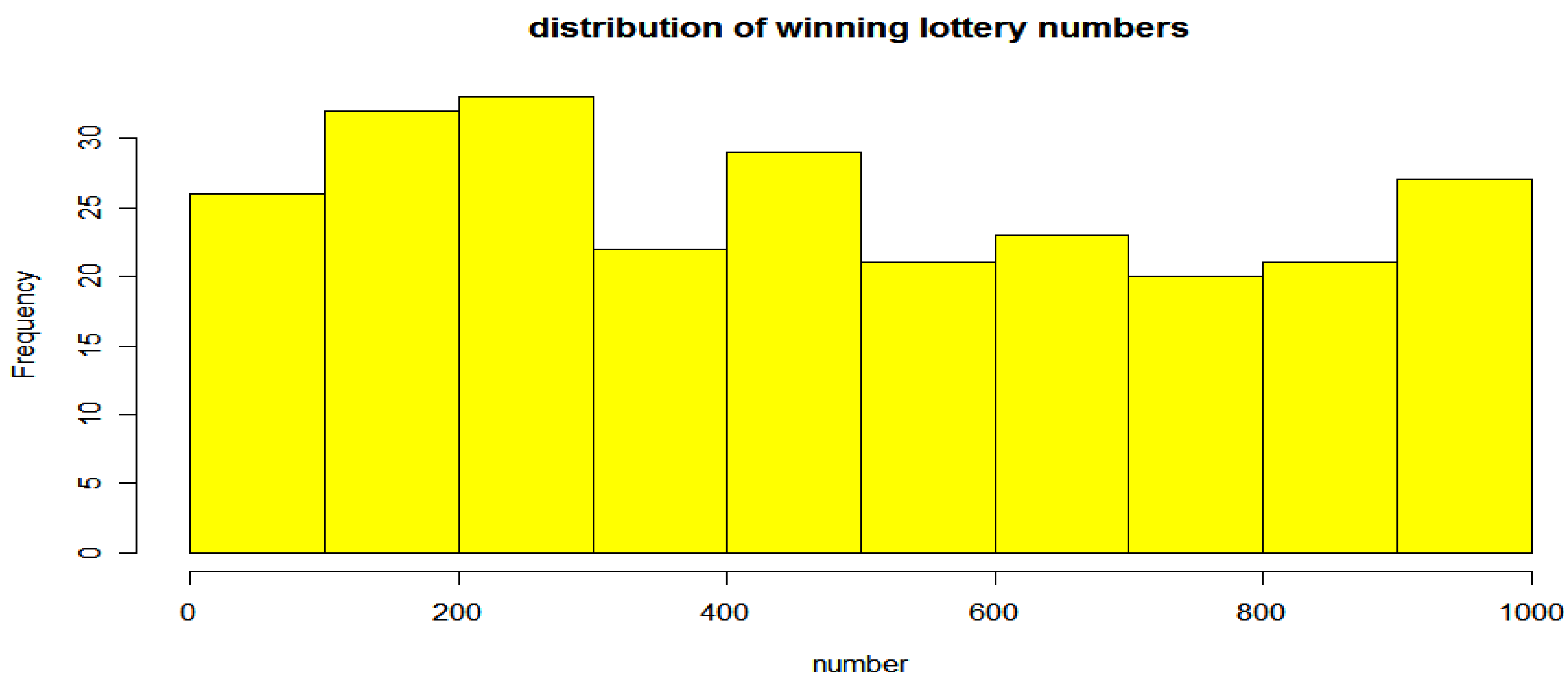
winning numbers and prize amounts

(810, \$190.0)
(156, \$120.5)
(140, \$285.5)
(542, \$184.0)

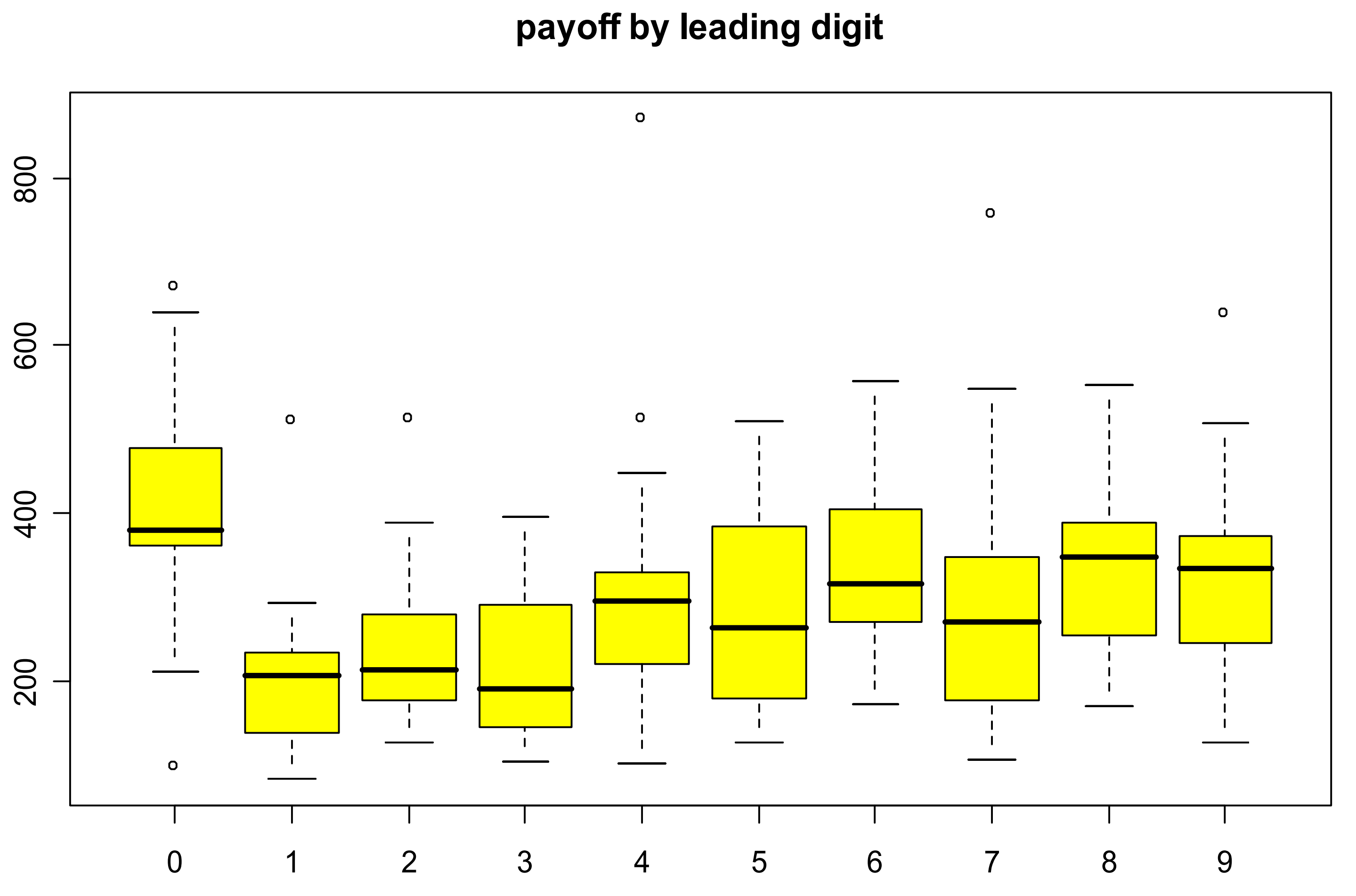
and so on for 254 consecutive days

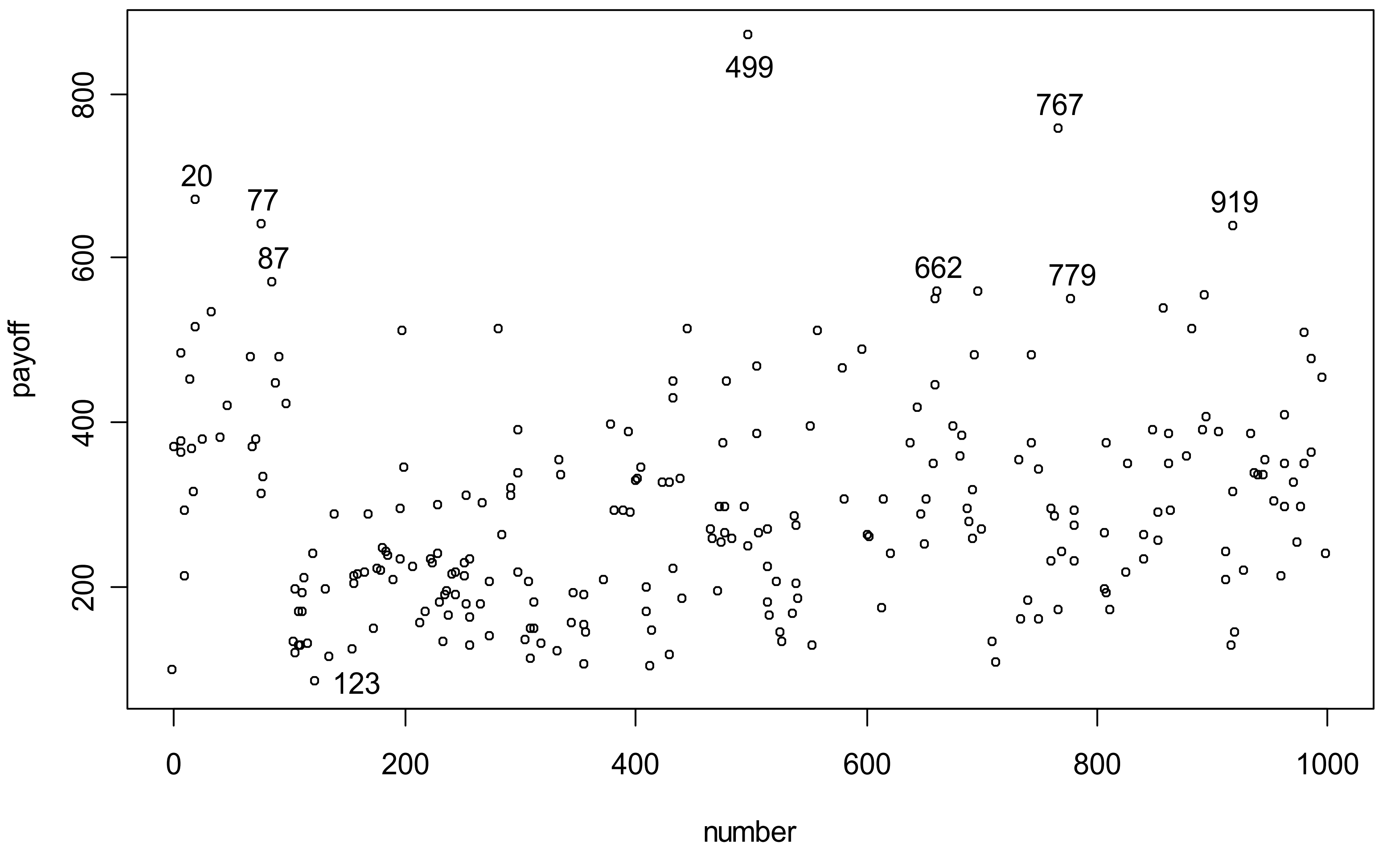
strategy 1: choose a winning number

- since we have data on the winning numbers, see if there's a pattern we can exploit to pick the winners
- examine the distribution of winning numbers using histograms or stem-and-leafs



strategy 2: choose a winning number that few others pick





499
767
20
77
87
919
662
779
123
800
600
400
200
payoff
0
200
400
600
800
1000
number

why little circles?

- easier to distinguish overlapping points
- especially with jittering

five rules

- rule 1: graph lots
- rule 2: use what the eye is good at (and avoid what the eye is bad at)
- rule 3: find the right contrast and show it
- rule 4: make it easy to spot pattern, and deviations from pattern
- rule 5: plot models, not just the data

rule 1: graph lots

- only one out of 50 graphs will “work” so to get a handful of workable graphs, graph lots
- good graphing principles help raise your yield of workable graphs
- better if you can generate lots of simple graphs quickly even if they’re not perfect
- for you, not for presentation (at this stage) so don’t obsess on look

  (though I’m showing you the survivors of hundreds of graphs, so they’re cleaner and not quite representative of the messy graphs I usually produce: my working graphs usually don't have titles, clear axis labels, etc.)

rule 2: use what the eye is good at (and avoid what it's bad at)

- we need to know something about how the eye-brain perceives graphics
  - what it's good and bad at, and an ordering or hierarchy
  - “optical illusions” and traps to avoid
  - techniques to exploit strengths and minimize weaknesses

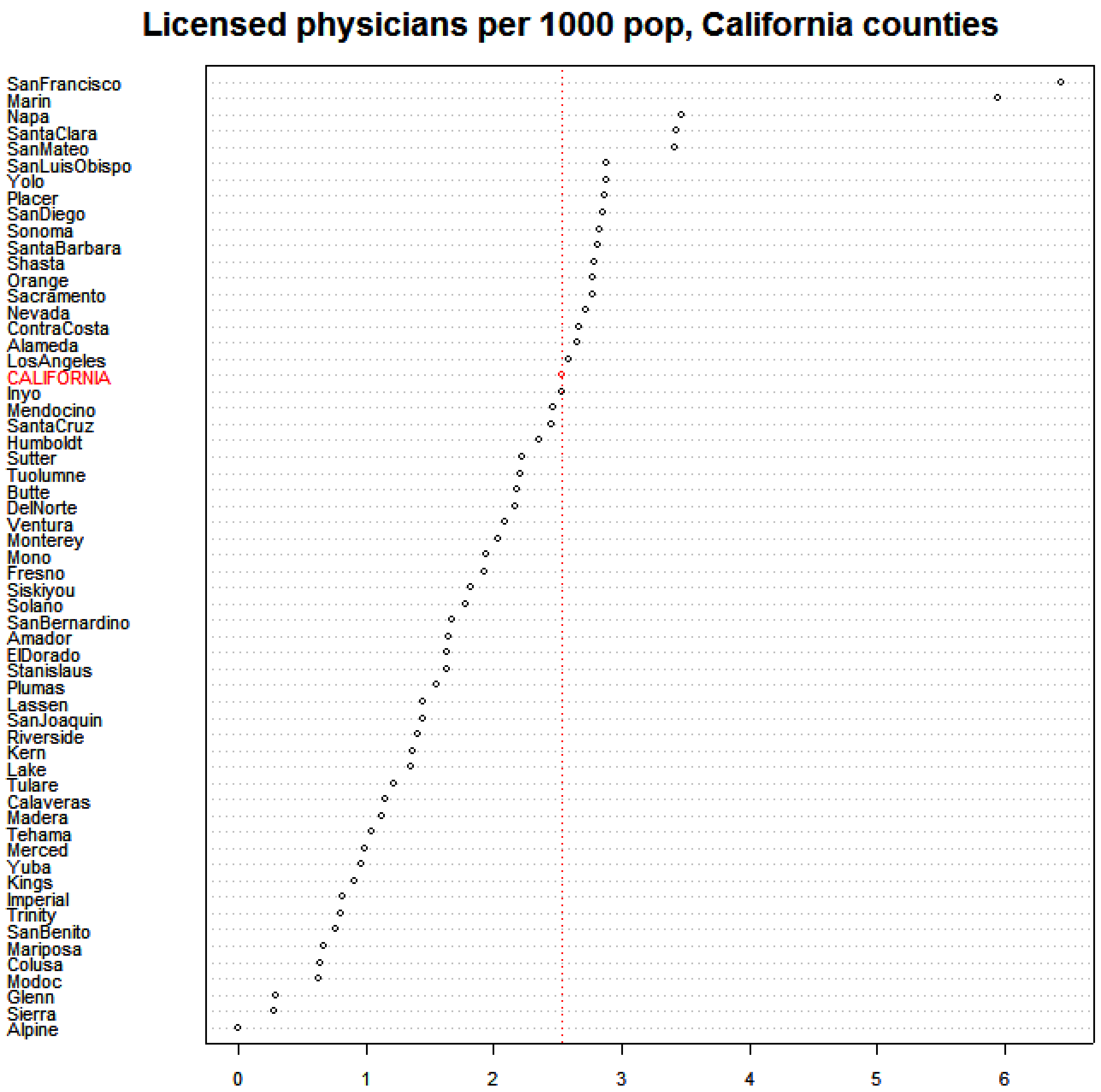
graphical perception

- quantitative pattern recognition by
  - detection: recognition of geometry
  - assembly: grouping of detected elements
  - estimation: assessment of relative magnitudes
- the human eye-brain can be fooled
  - optical illusions
- need to help it out
  - grouping, ordering, highlighting help to identify patterns

a hierarchy of graphical perception

- position along common scale
- position along identical non-aligned scales
- length
- angle, slope
- area
- volume
- shading, color (good discrimation but poor ordering)

dotplots


Licensed physicians per 1000 pop, California counties
SanFrancisco
Marin
Napa
SantaClara
SanMateo
SanLuisObispo
Yolo
Placer
SanDiego
Sonoma
SantaBarbara
Shasta
Orange
Sacramento
Nevada
ContraCosta
Alameda
LosAngeles
CALIFORNIA
Inyo
Mendocino
SantaCruz
Humboldt
Sutter
Tuolumne
Butte
DelNorte
Ventura
Monterey
Mono
Fresno
Siskiyou
Solano
SanBernardino
Amador
ElDorado
Stanislaus
Plumas
Lassen
SanJoaquin
Riverside
Kern
Lake
Tulare
Calaveras
Madera
Tehama
Merced
Yuba
Kings
Imperial
Trinity
SanBenito
Mariposa
Colusa
Modoc
Glenn
Sierra
Alpine
0
1
2
3
4
5
6

area

- pie charts require estimation of area
- human perception of relative areas is conservative, i.e., shrinkage toward 1.0
- shape affects estimation of area
  - concave shapes appear larger than convex
  - maps are good for context and clustering, not so good for comparisons of quantitative amounts
- color intensity affects estimation of area.
  - highly saturated colors appear larger

color

- human eye good at discrimination, poor at ordering
  - use for categories, not for quantitative coding
  - hues are not ordered
  - use for highlighting, patterning, especially in combination with small multiples
- more on color, later

rule 3: find the right contrast and show it

- don't rely on the eye to do differencing

  if you're interested in the difference between two lines, don't show the lines and rely on the eye to calculate the difference; calculate the difference itself and show it

- Tukey mean-difference (aka Bland-Altman) plots

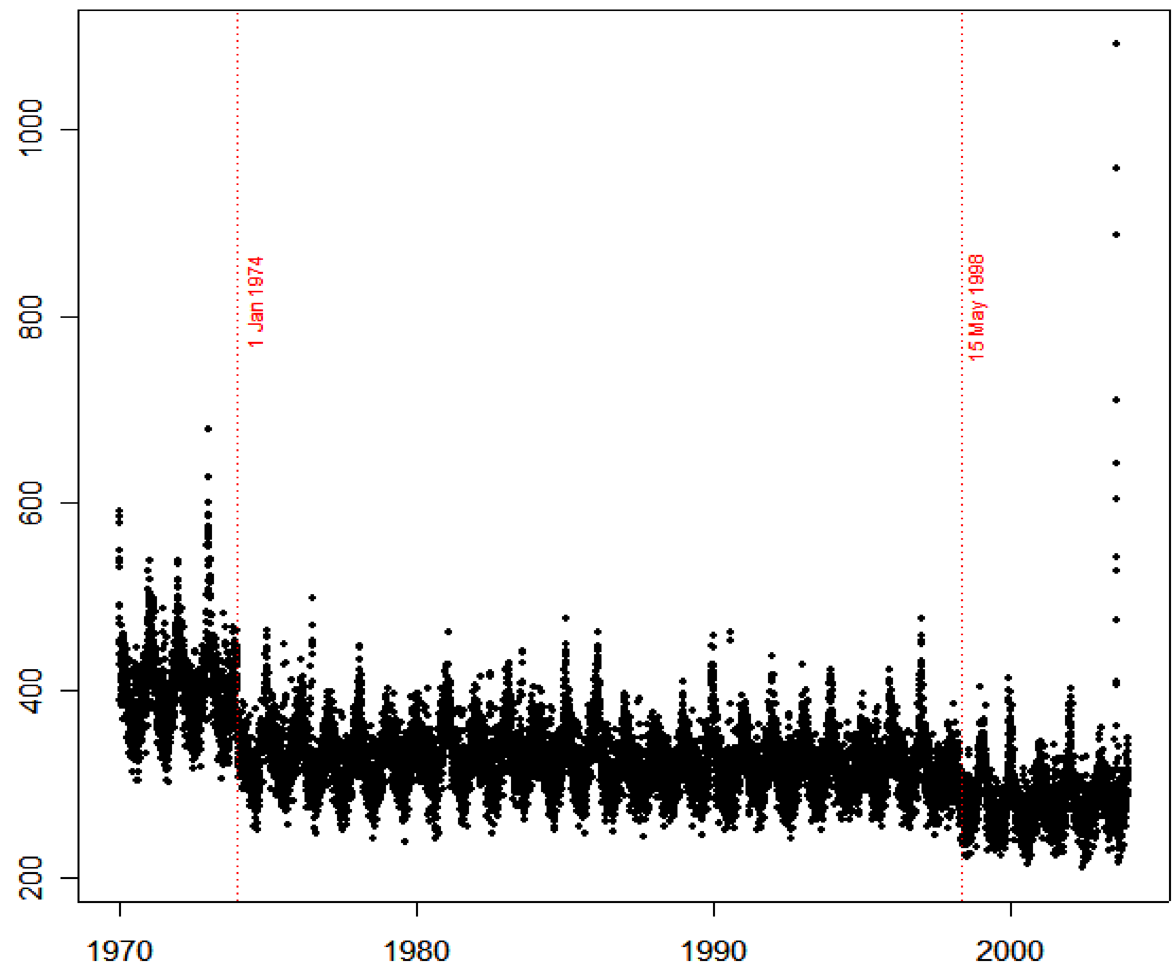
  levels vs. differences

- fits and residuals

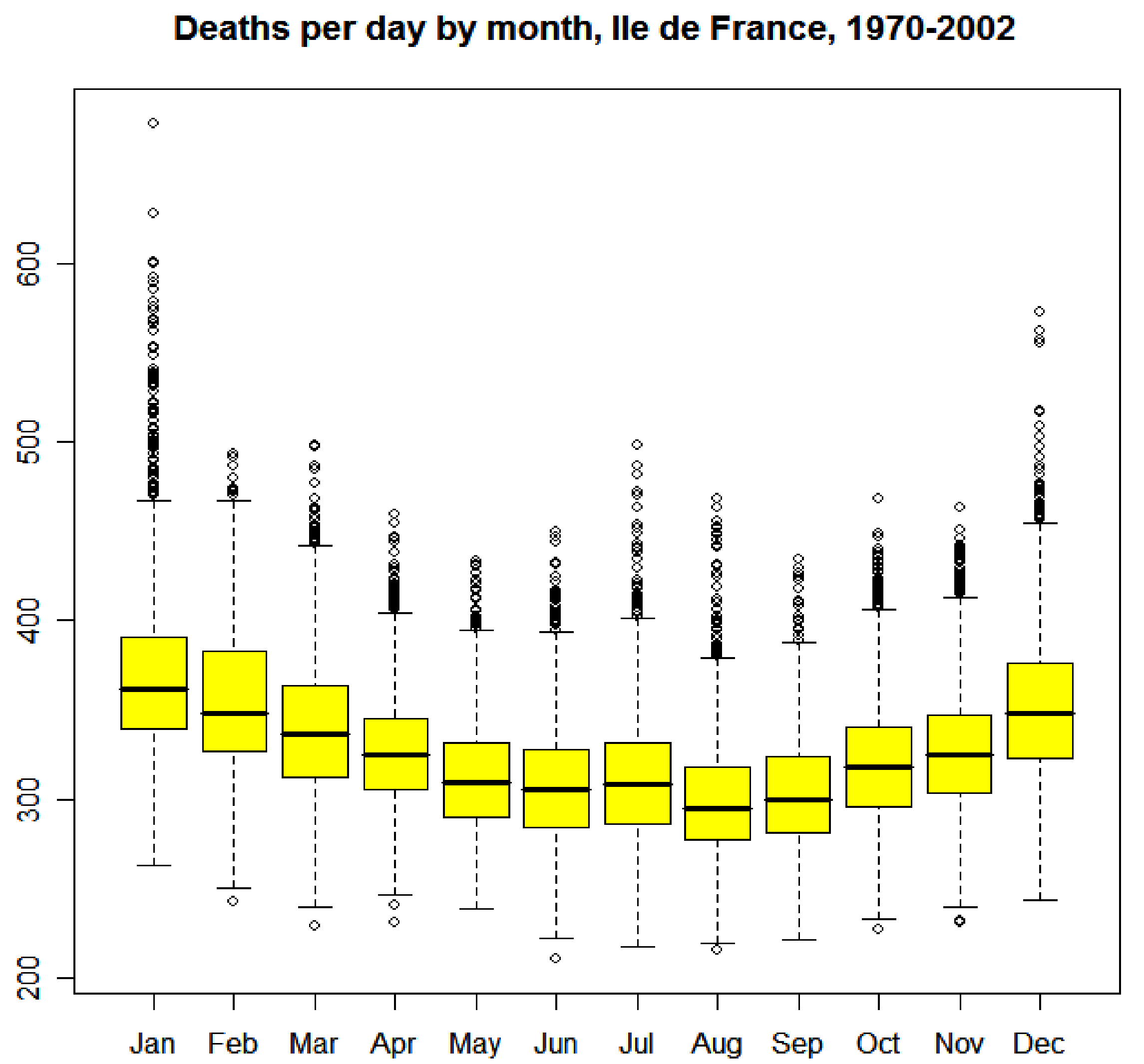
- different contrasts can give different insights

differences by time

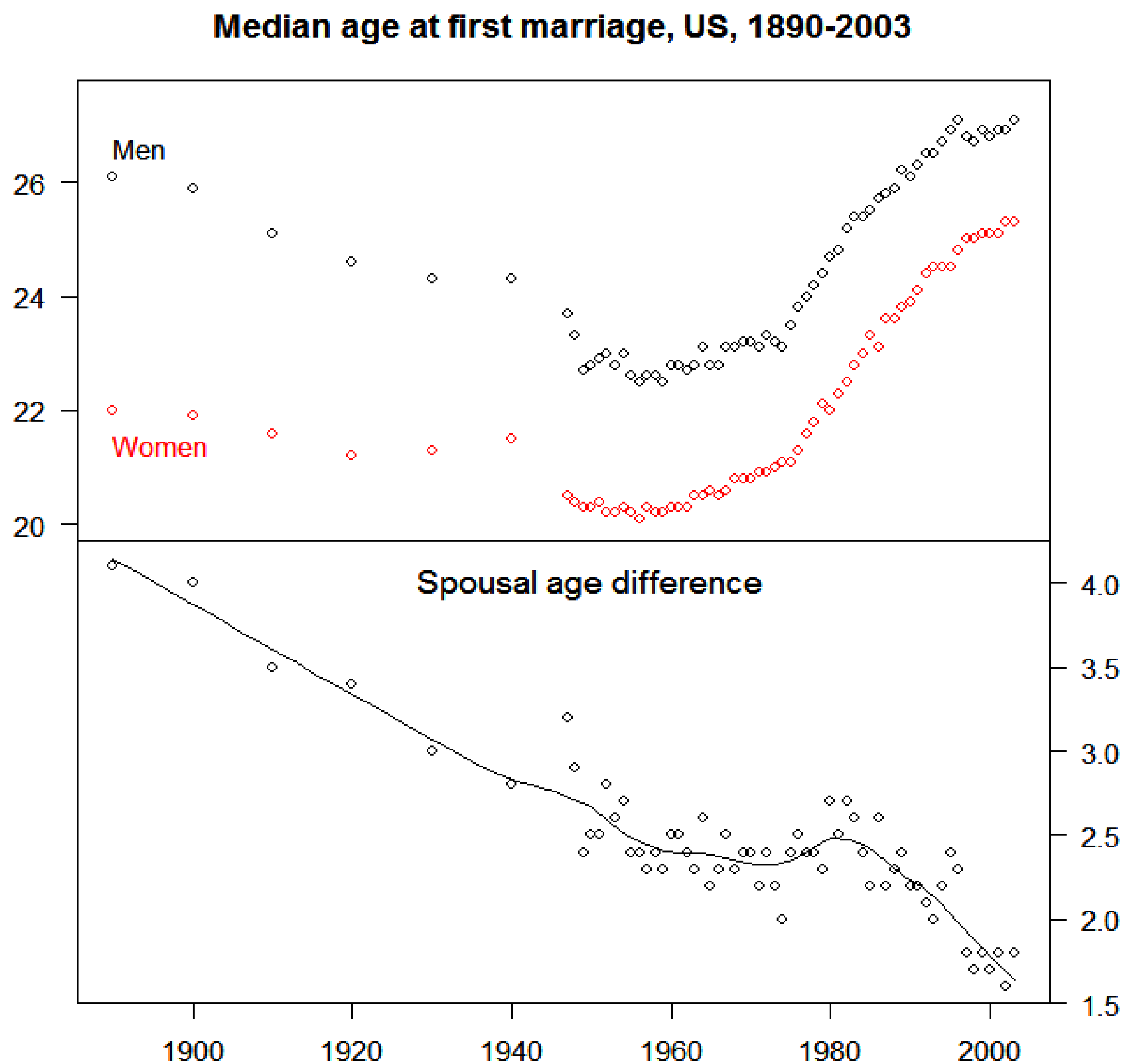
Daily deaths, Ile de France, 1970-2003



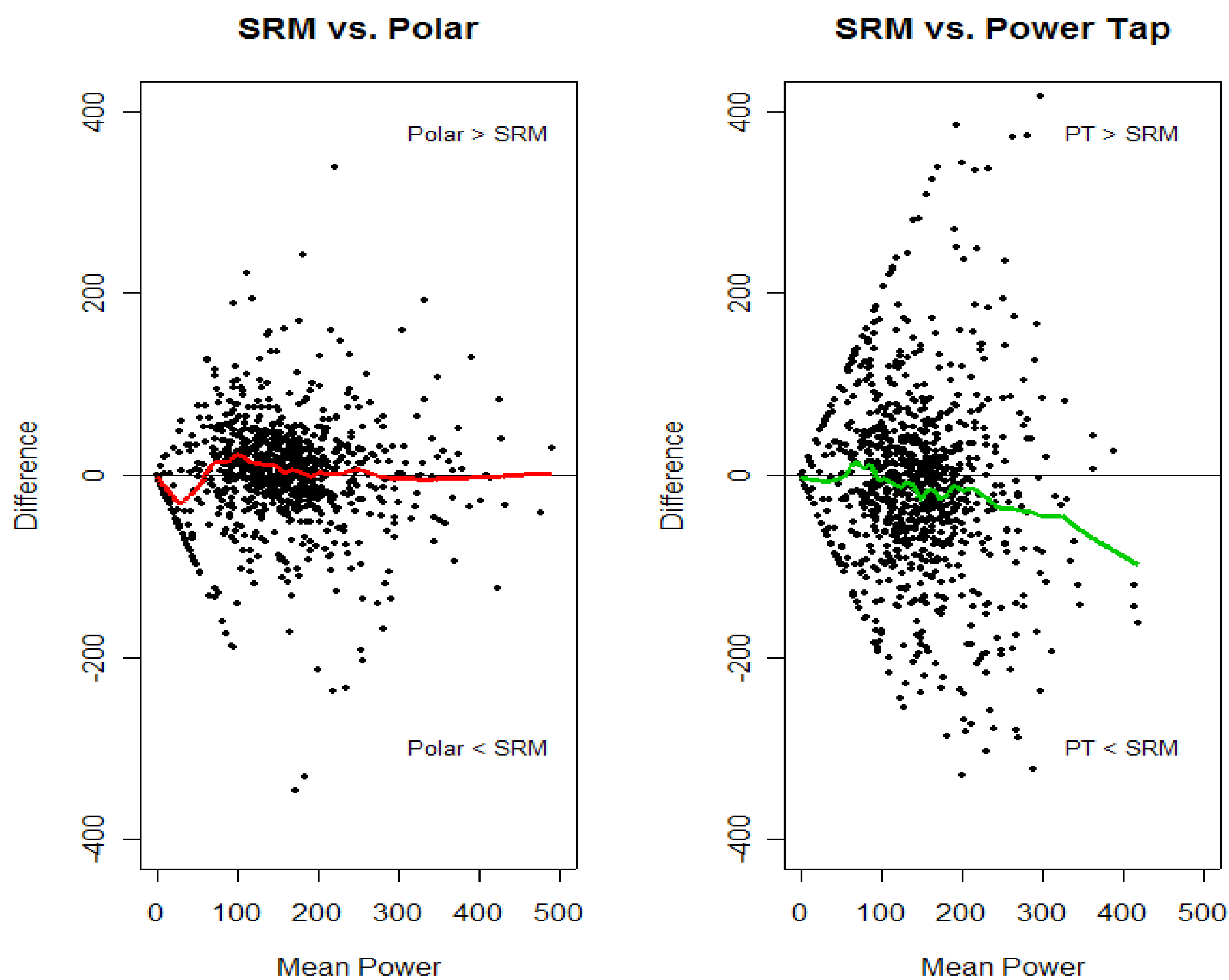
differences by category



differences between lines



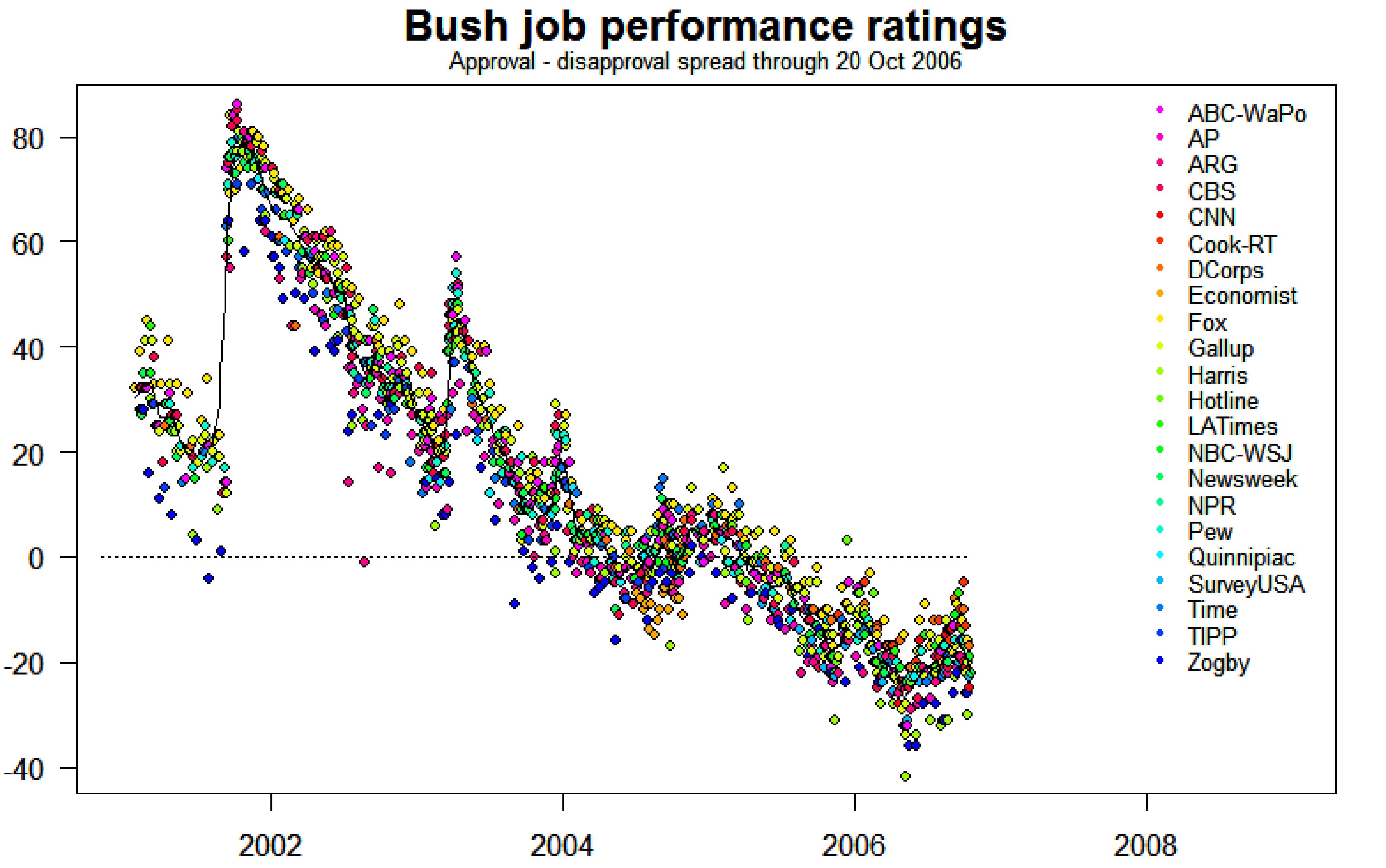
differences and means



differences from fits

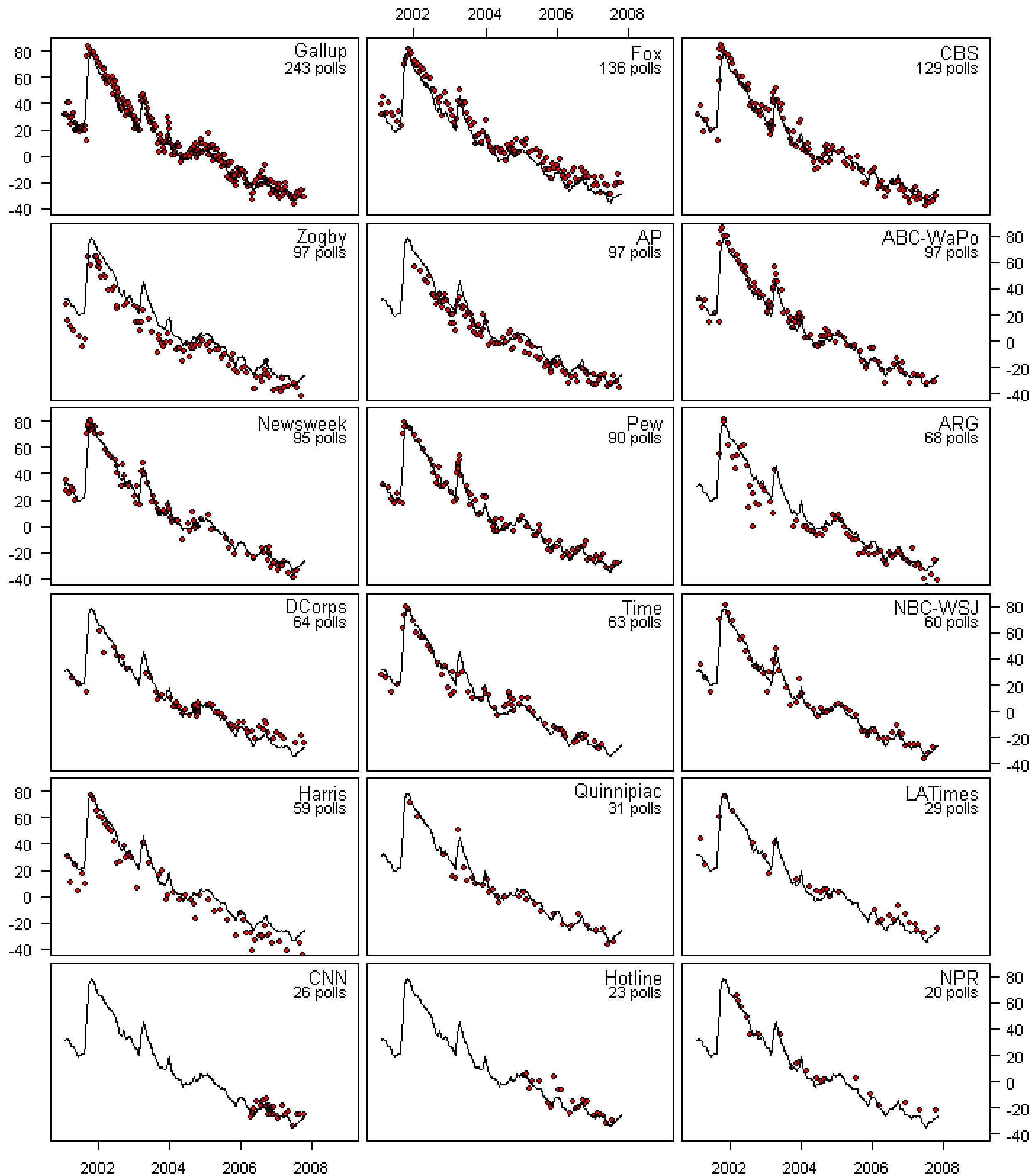
- data = fit + residual
- the classic residual plot

  “fit” can be broadly defined: data = smooth + residual

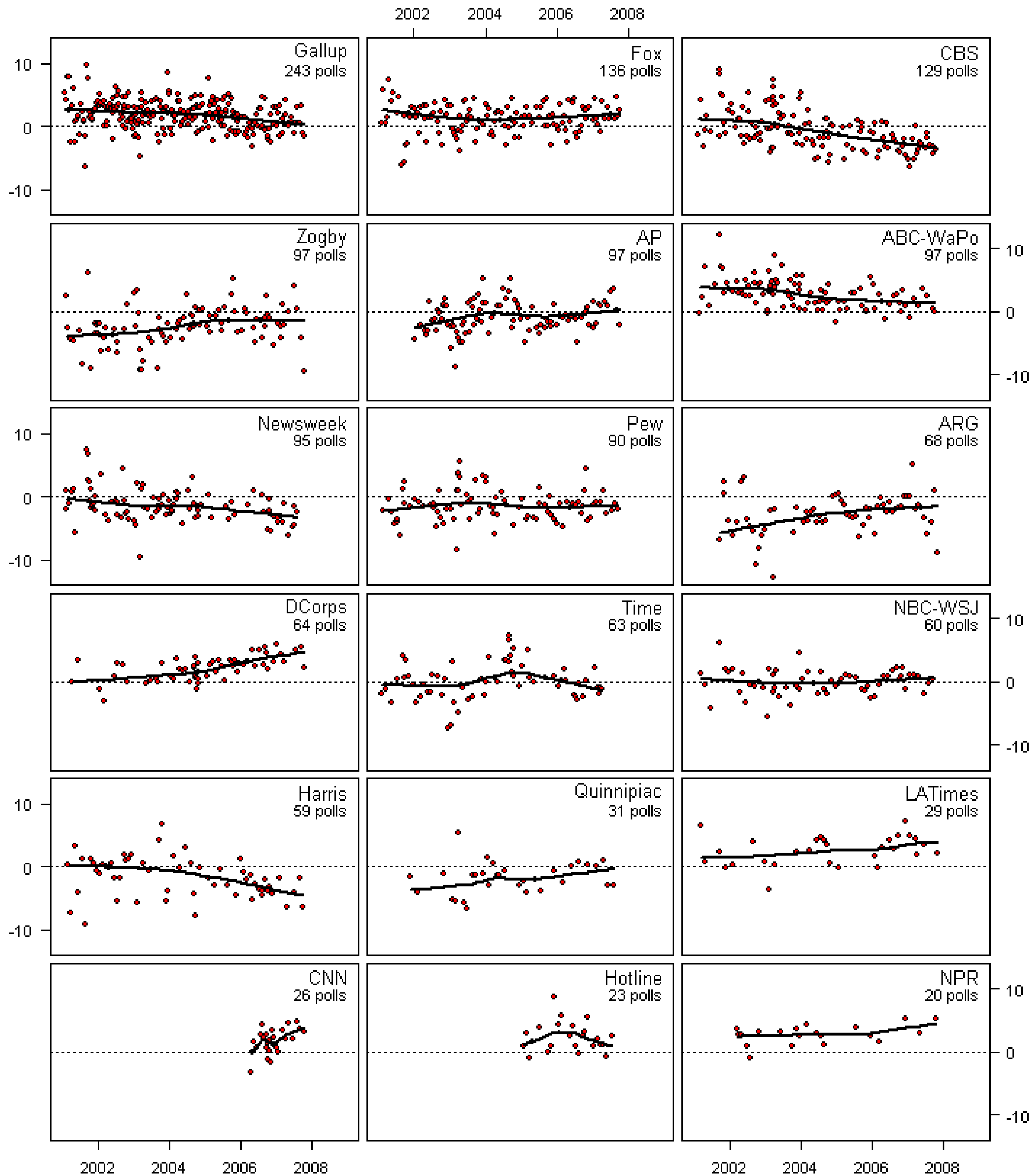


Bush job performance ratings
Approval - disapproval spread through 20 Oct 2006
80
60
40
20
0
-20
-40
2002
2004
2006
2008
ABC-WaPo
AP
ARG
CBS
CNN
Cook-RT
DCorps
Economist
Fox
Gallup
Harris
Hotline
LATimes
NBC-WSJ
Newsweek
NPR
Pew
Quinnipiac
SurveyUSA
Time
TIPP
Zogby

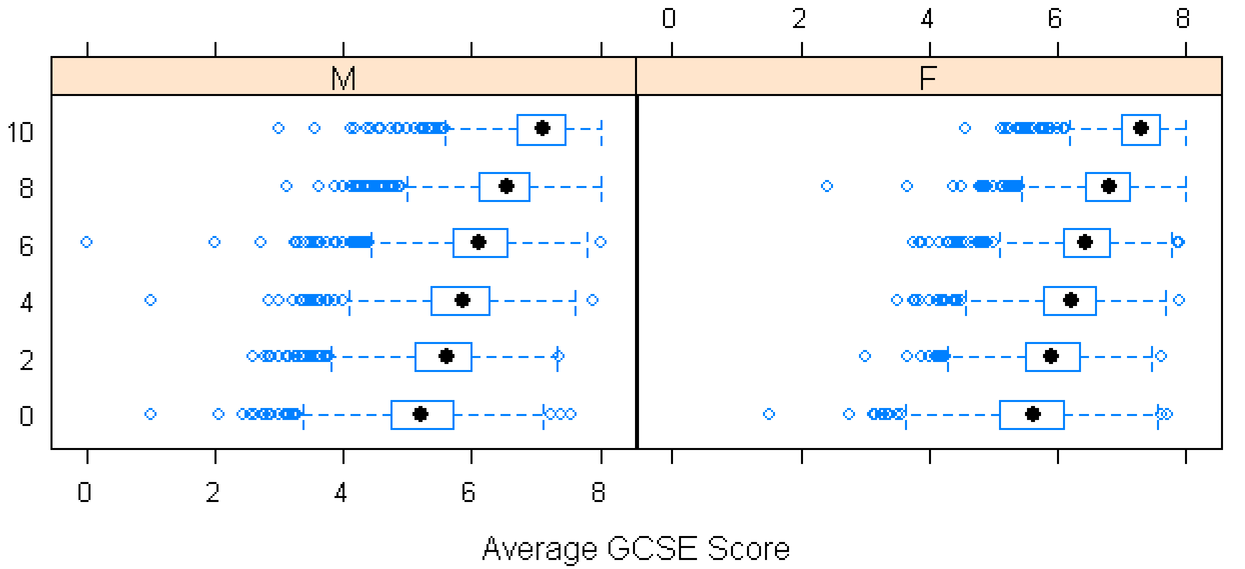
Net job approval for President Bush

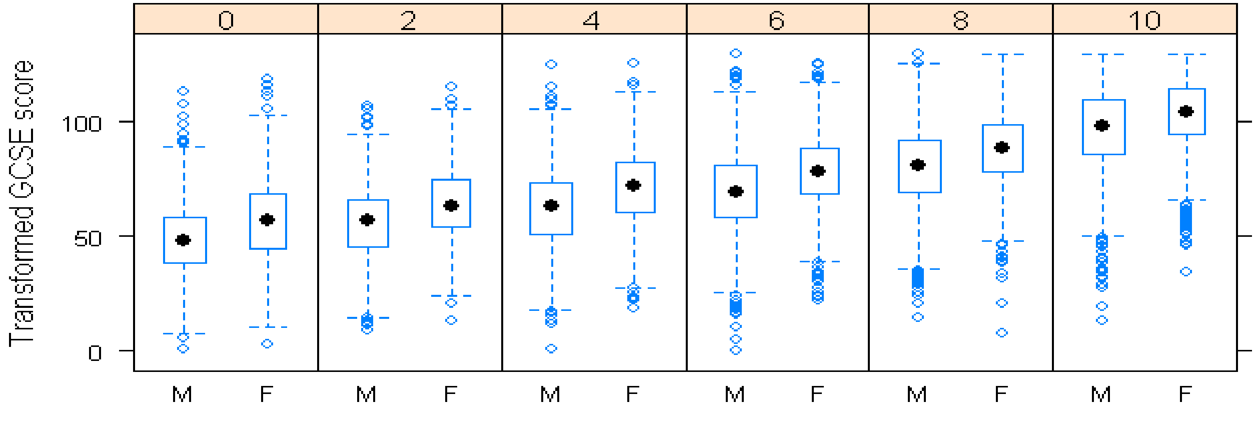


Some polls consistently below trend; others above



same data, different contrast

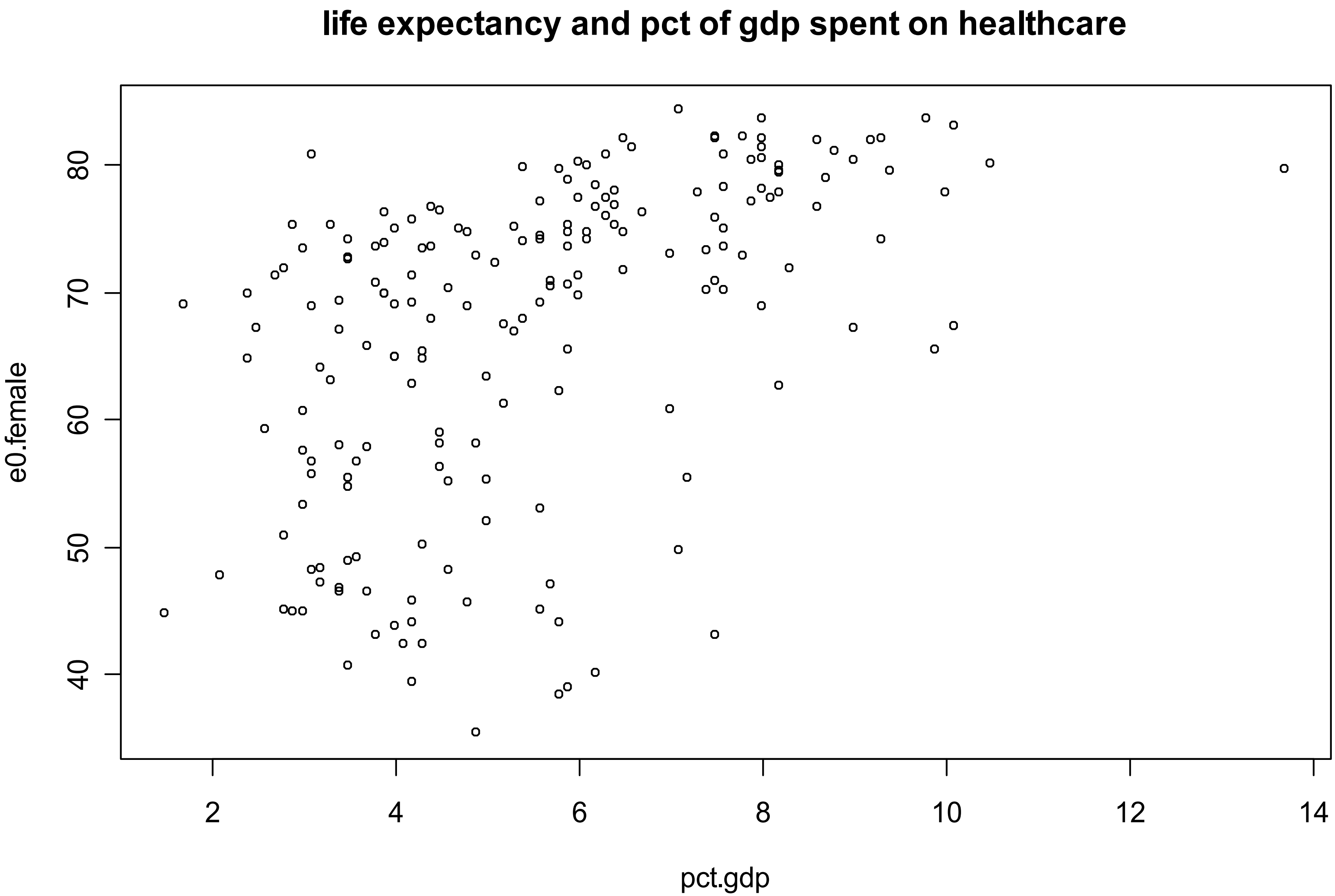




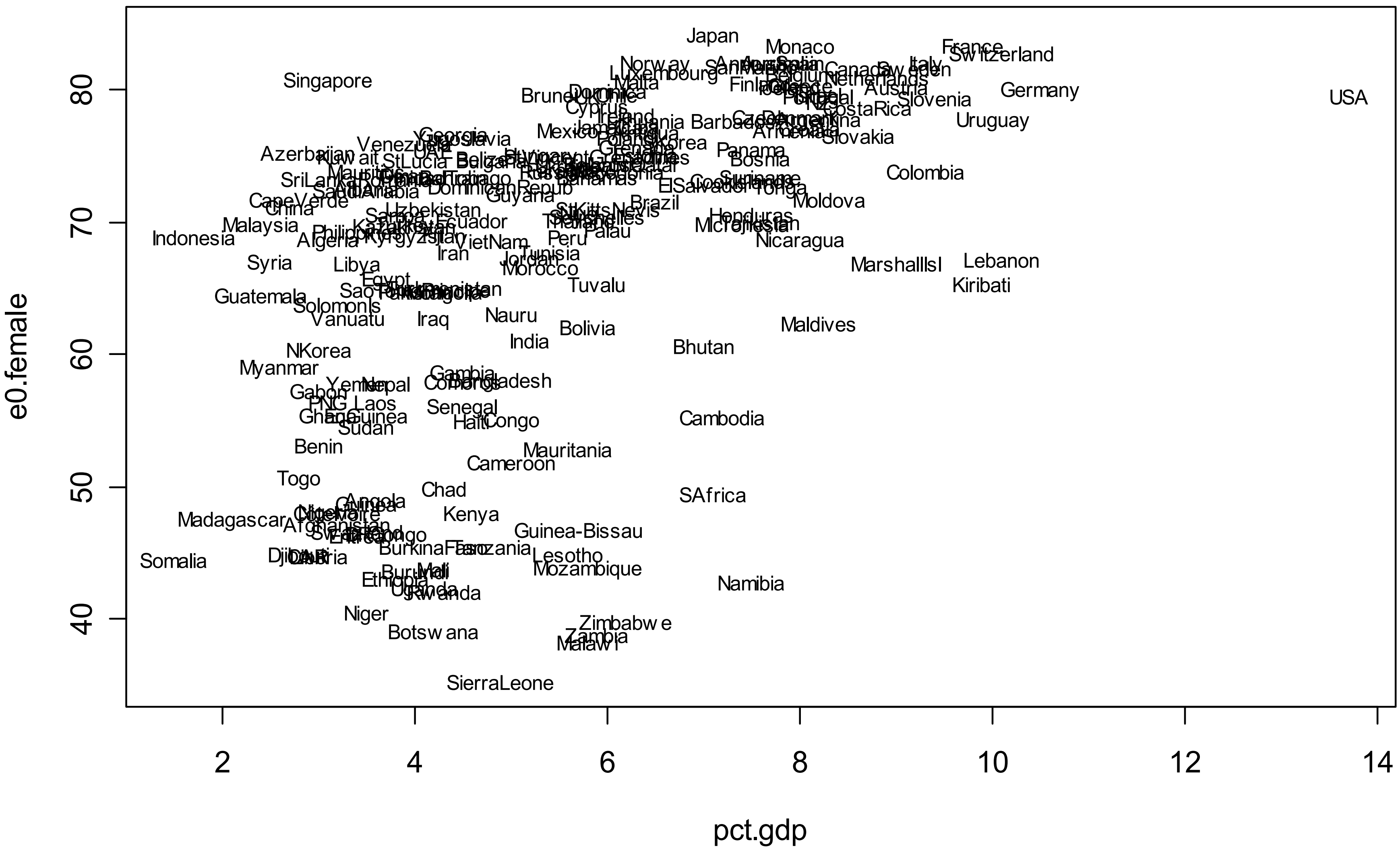
rule 4: make it easy to spot pattern

- add information depth, not (unnecessary) complexity
  - sometimes two plots are better than one complex plot (and sometimes it isn't)

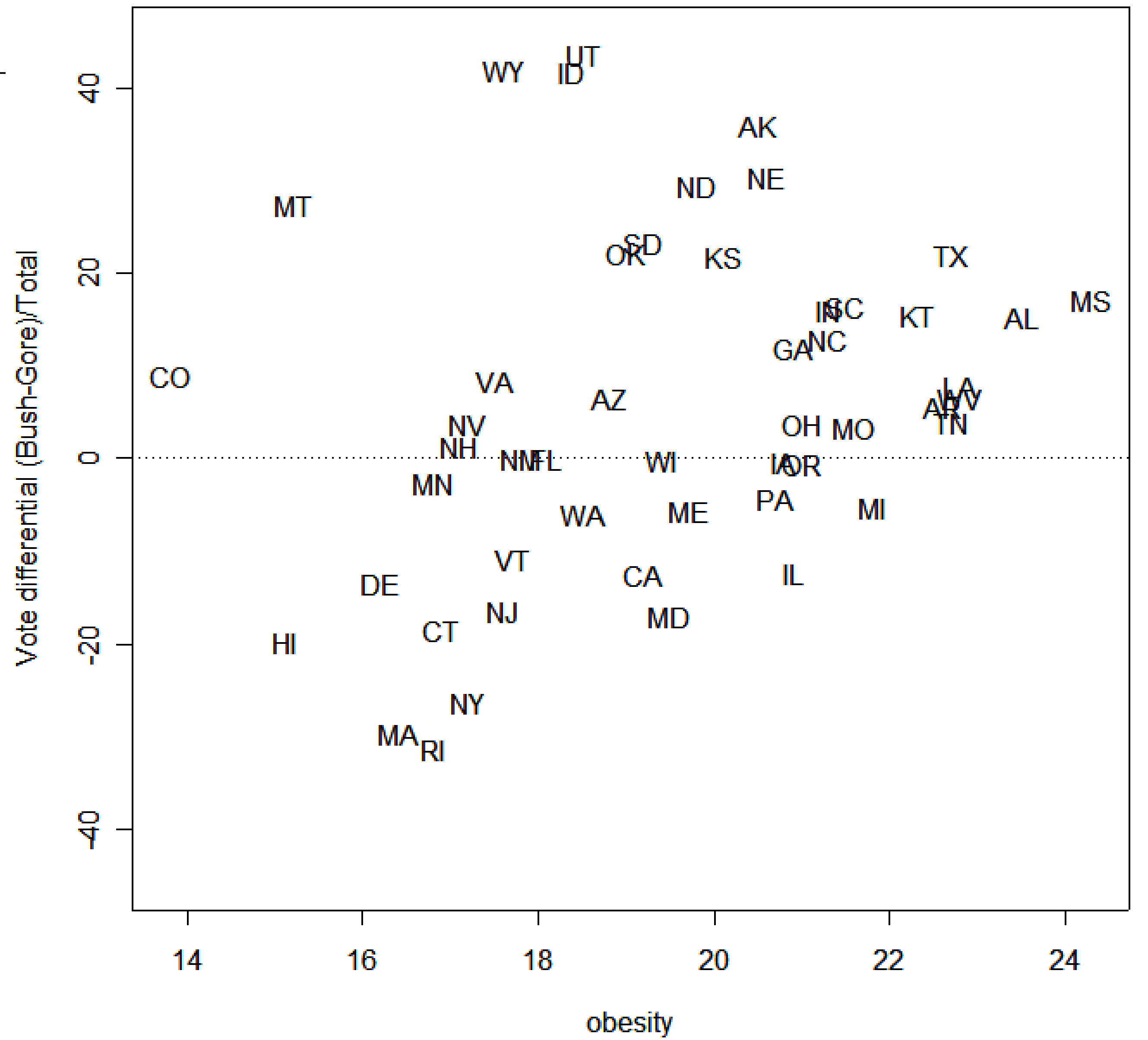
direct labeling



life expectancy and pct of gdp spent on healthcare



2000 Vote vs. self-reported obesity percentage



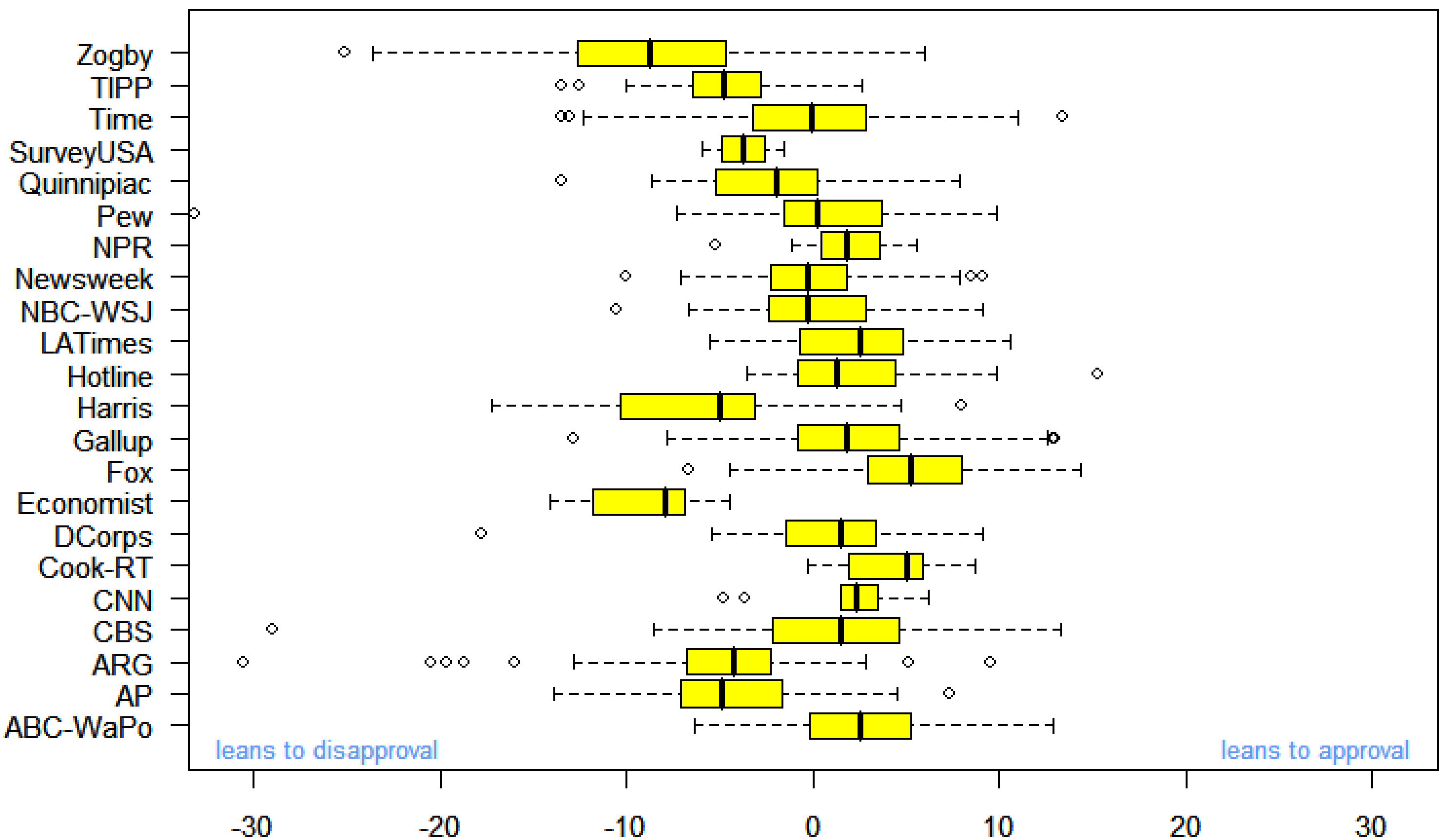
more on labeling

- direct labeling of lines often better than legends
    - particularly good when combined with line color
        - symbol plus line type often too busy to decode
        - looking back-and-forth at a legend is distracting

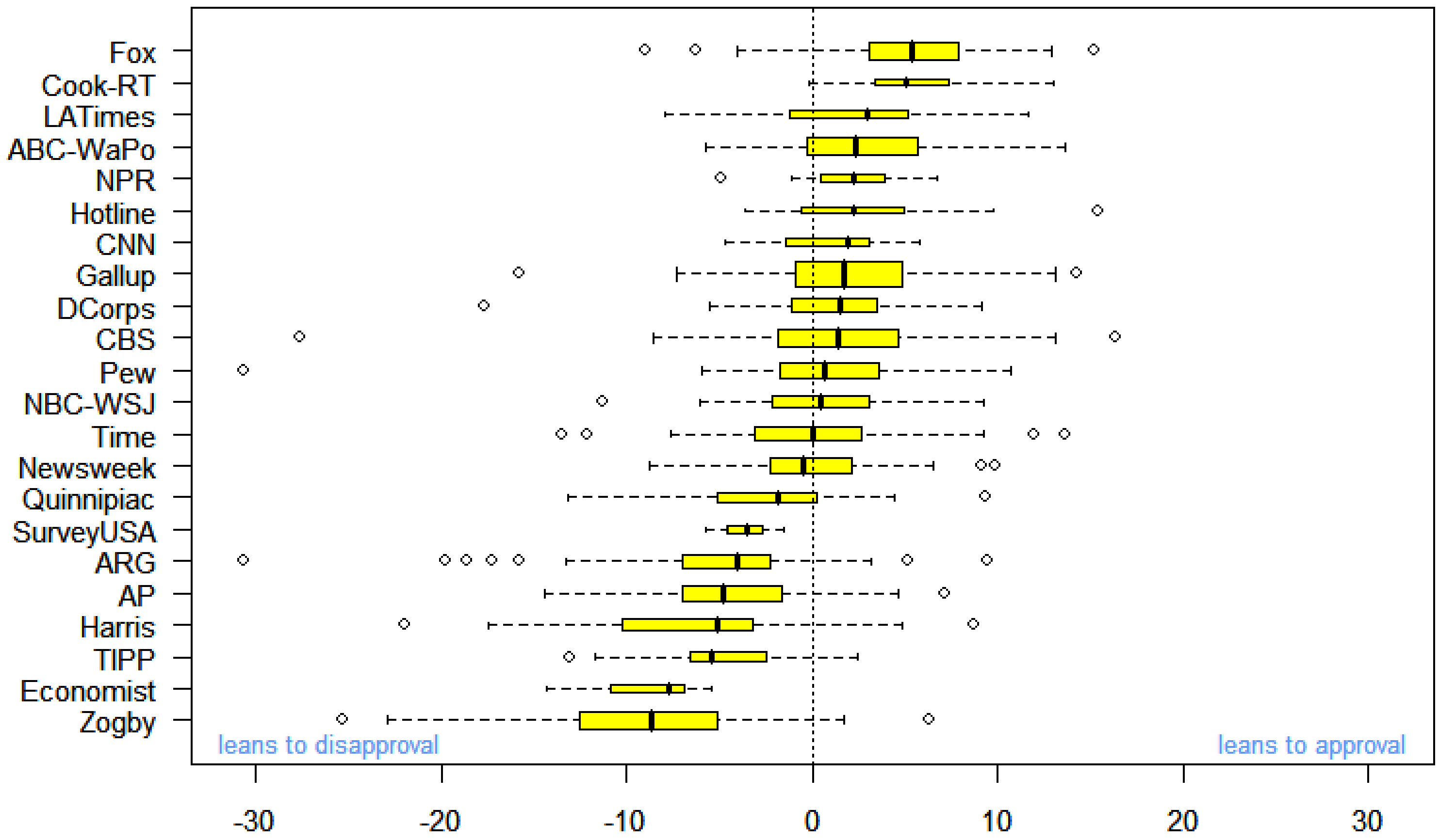
ordering

- default ordering for categorical variables is often alphabetical
  - that makes categoriess easy to find, but hard to compare
  - example: country data are often ordered by name of country rather than by the variable you're interested in
- find an ordering that makes sense and use it
  - if you are interested in mortality differences among countries, order by mortality not country name
  - this helps you spot and evaluate small differences between countries

Differences from trend, net job performance



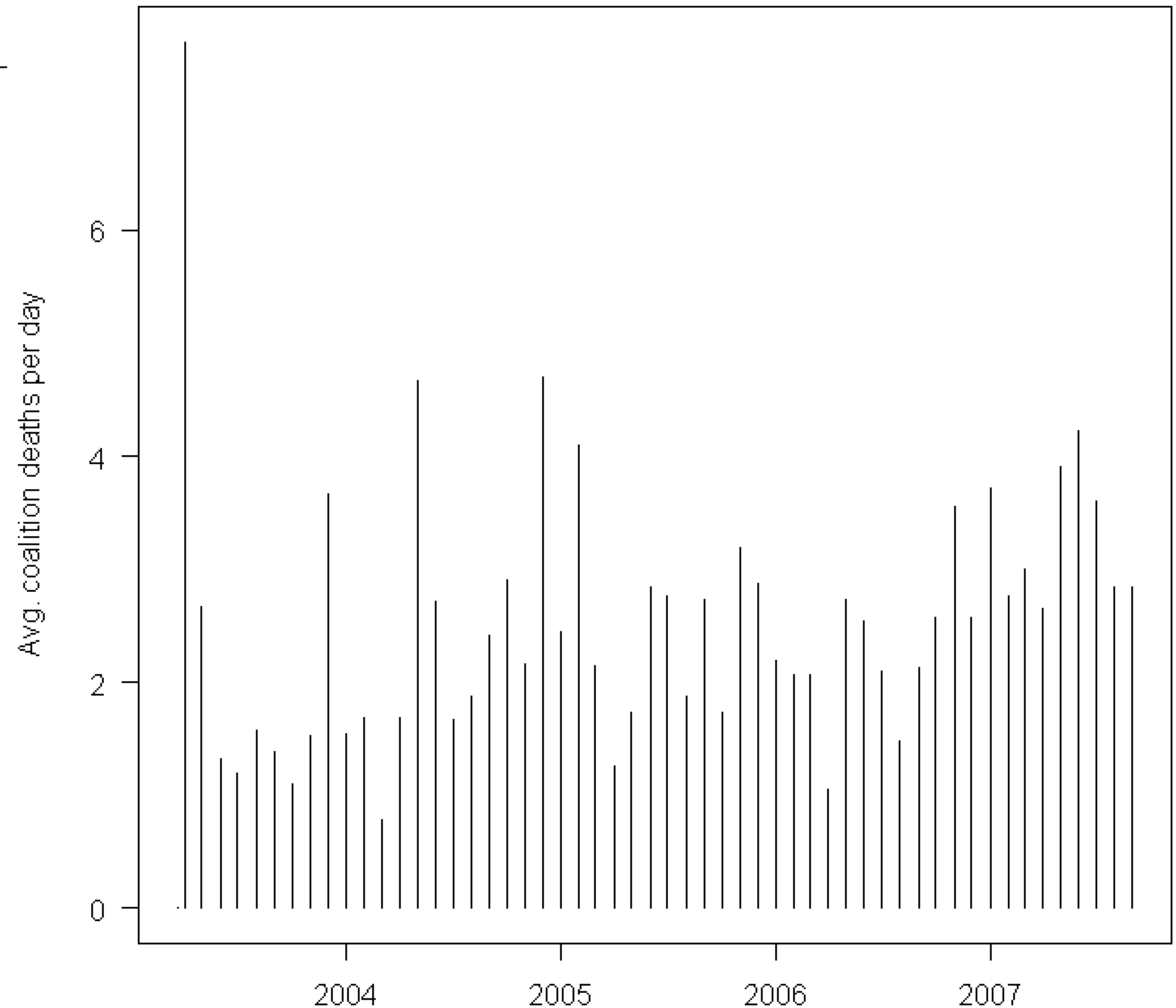
Differences from trend, net job performance

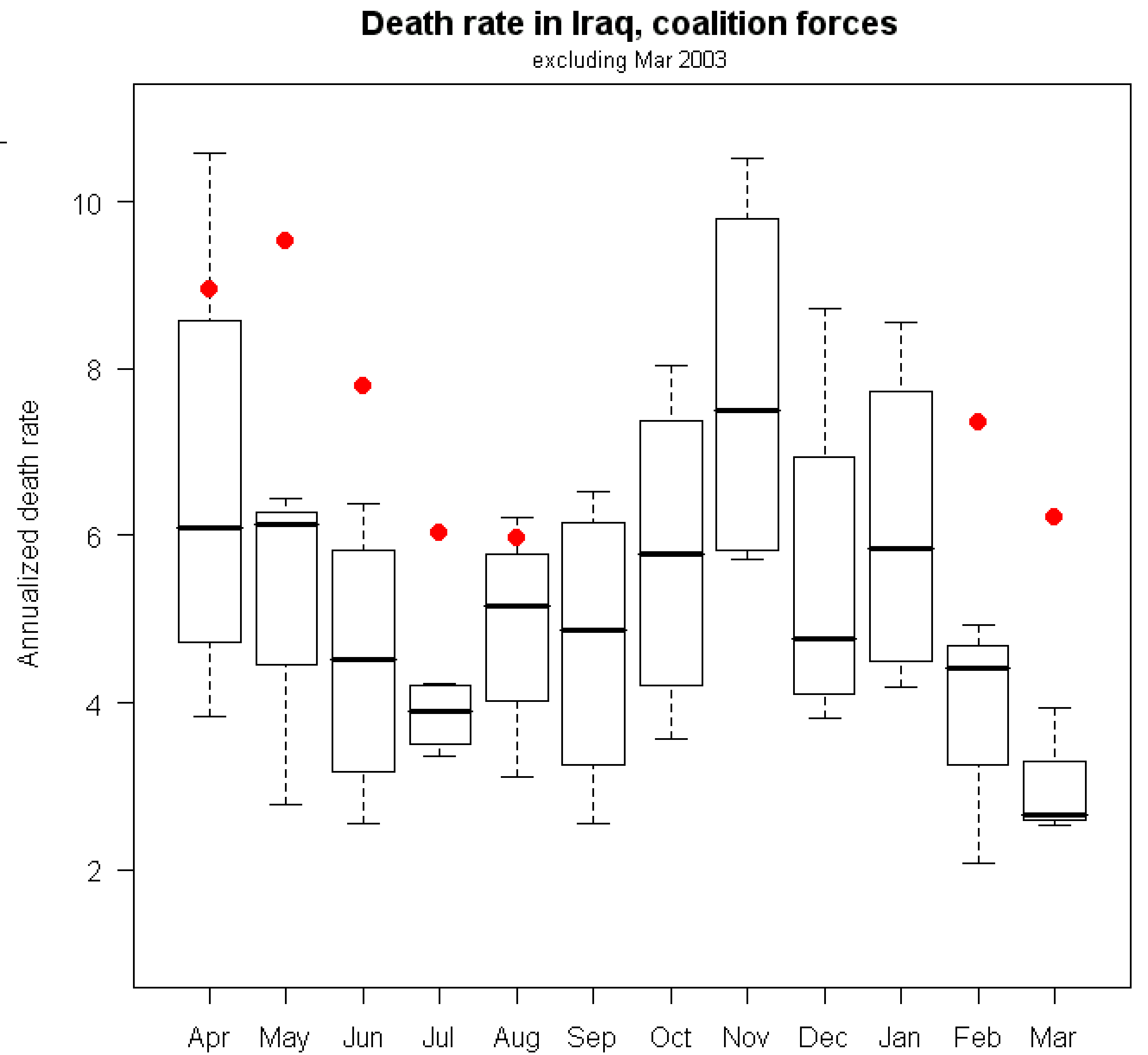


grouping

- grouping (done well) helps with pattern recognition
  - boxplots are a familiar way to group
- grouping (done poorly) obscures pattern
  - not all "obvious" groupings are informative
- next two slides show (almost) same data

Coalition deaths in Iraq, by month




Death rate in Iraq, coalition forces
excluding Mar 2003
Annualized death rate
2
4
6
8
10
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar

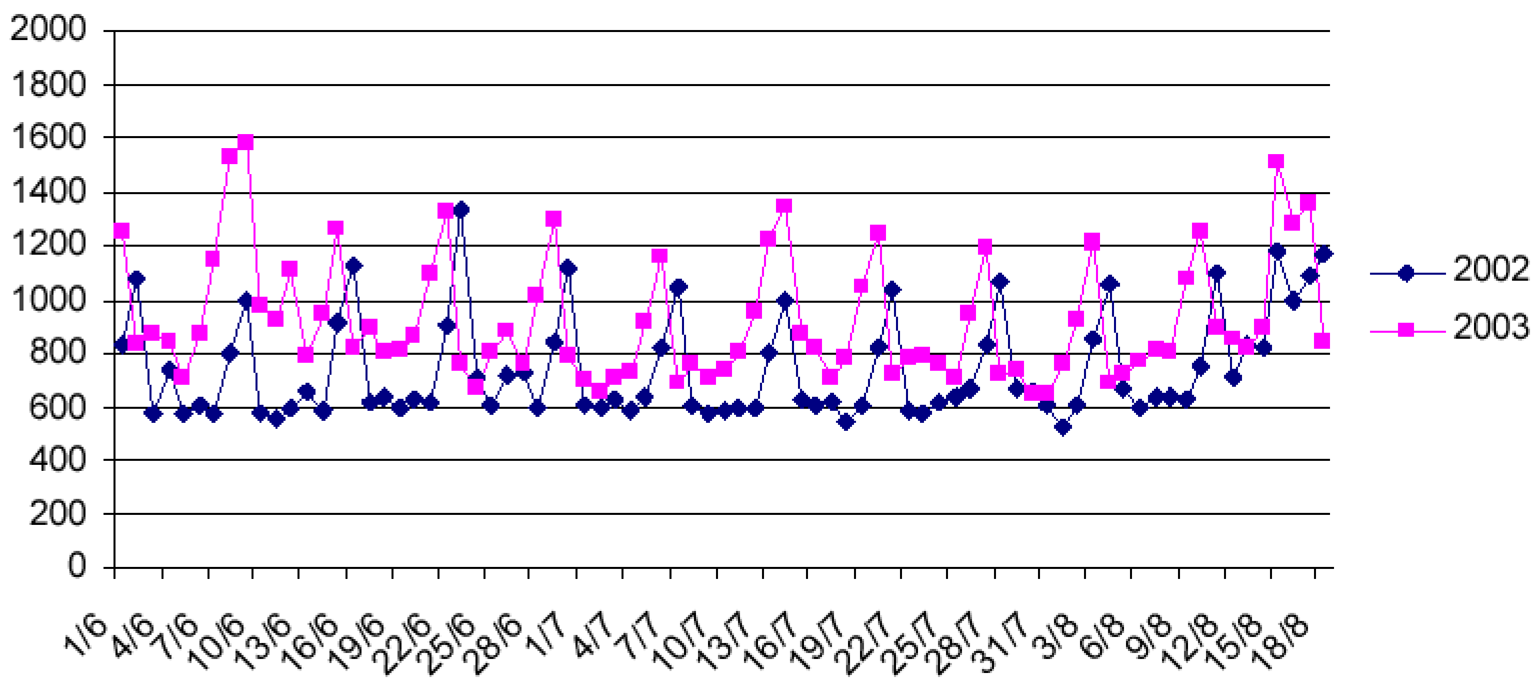
multivariate comparisons

- show relationships
  - more importantly, give you ideas
- time series plots show you what happened when, but rarely why they happened
  - we'll want to dig deeper into the data to generate new questions about the 'why?'

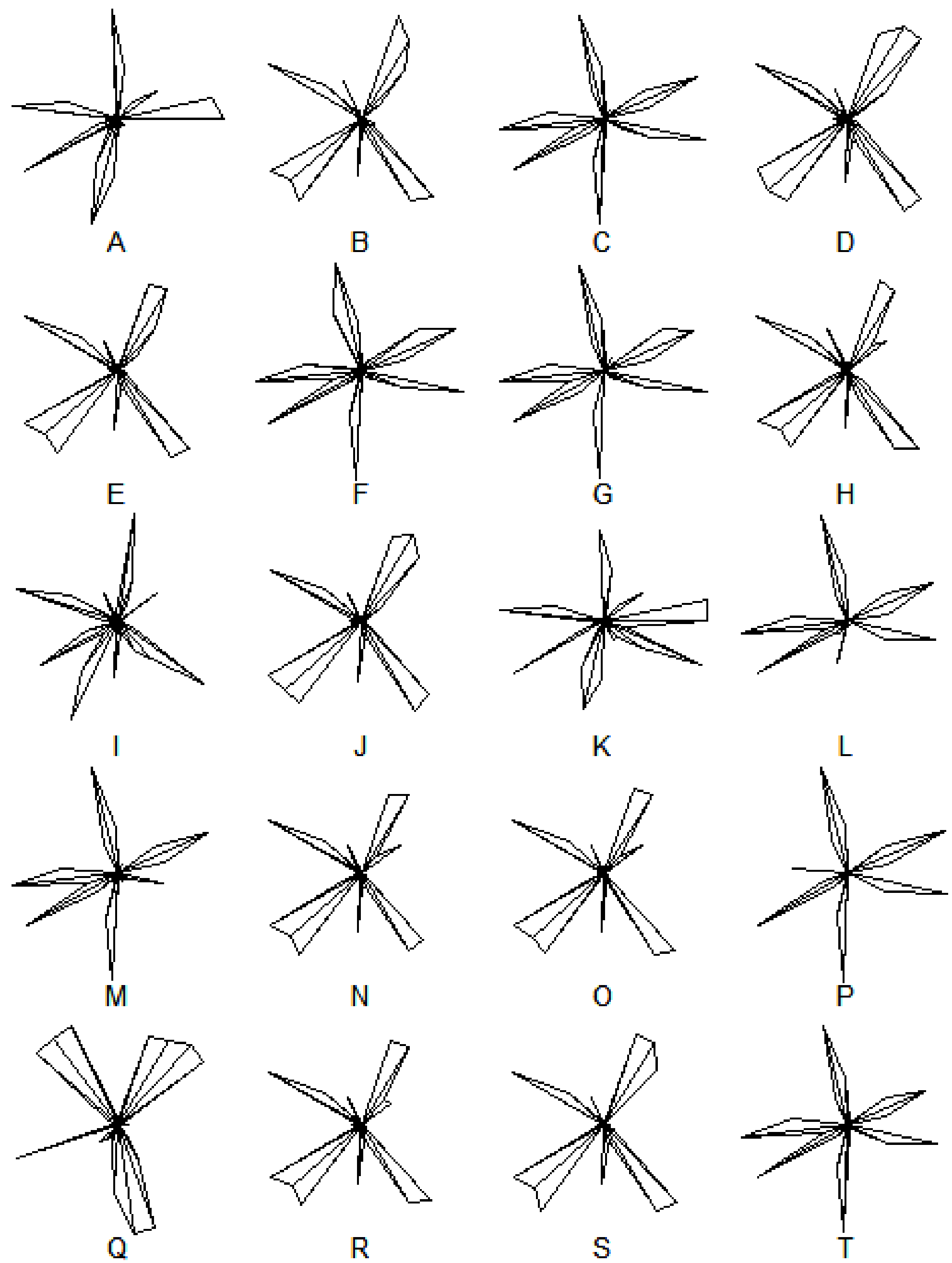
patterns in multivariate data

- twenty students read numbers of ambulance calls for July 2003 off a graph. How can we summarize the results?

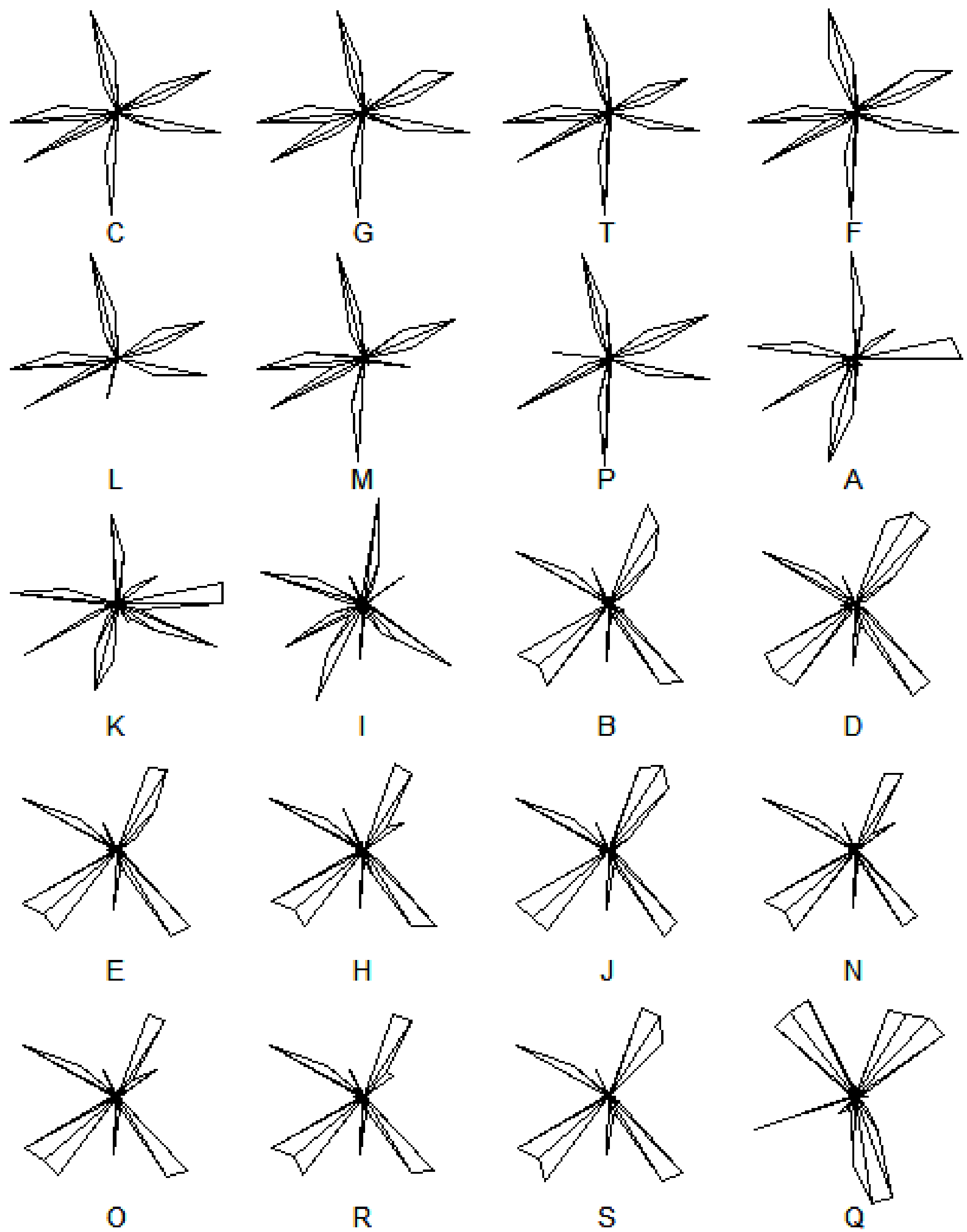
Graphique 5 : nombre d'interventions du SAMU 13 en 2003 par rapport à l'année précédente (2002)



Starplot for PS1



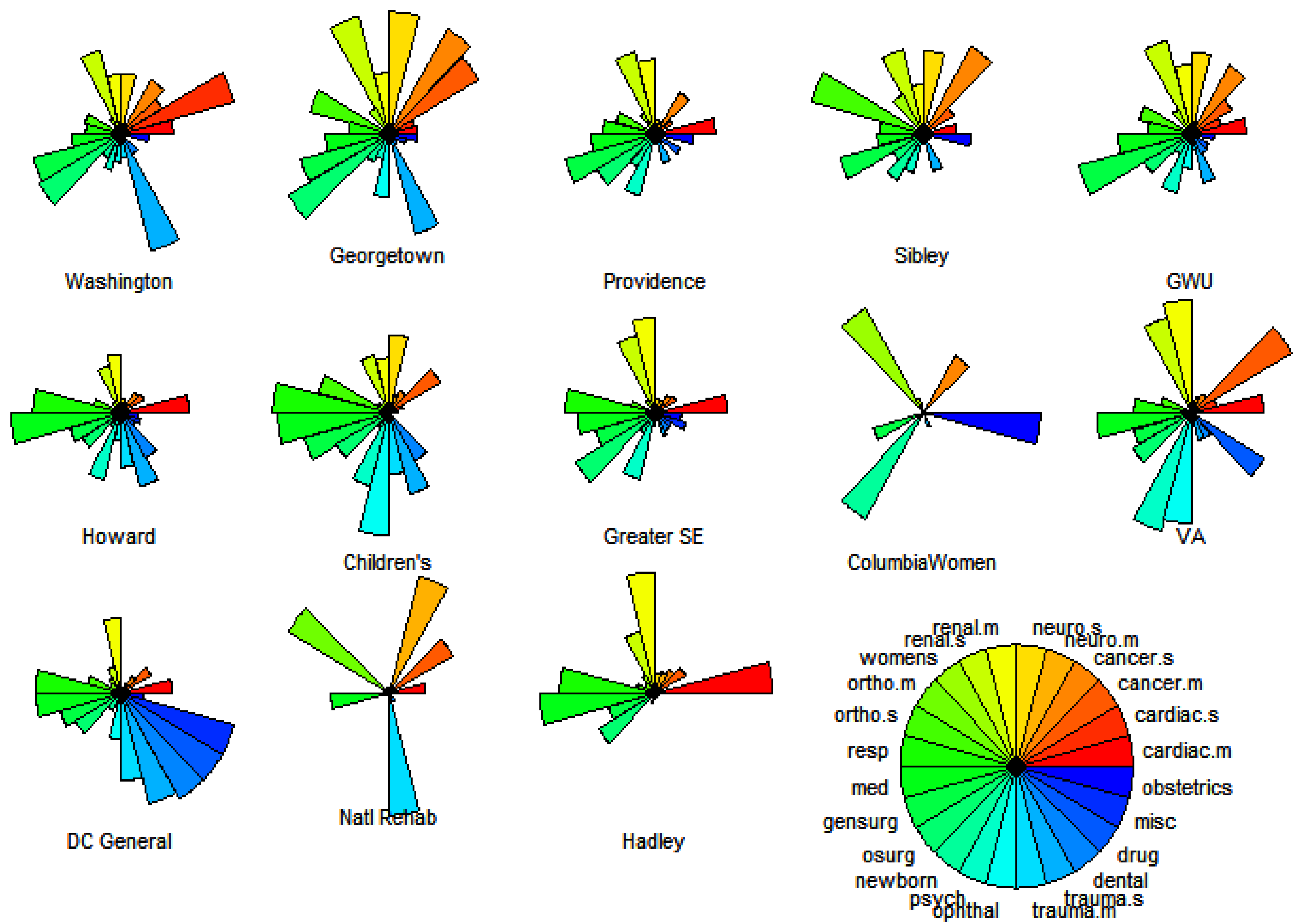
Rough grouping of PS1



dc hospitals

- thirteen hospitals
- twenty-four service lines
- do different hospitals specialize in different areas?

Hospitals and their service lines, 2000-1



stars are like pies

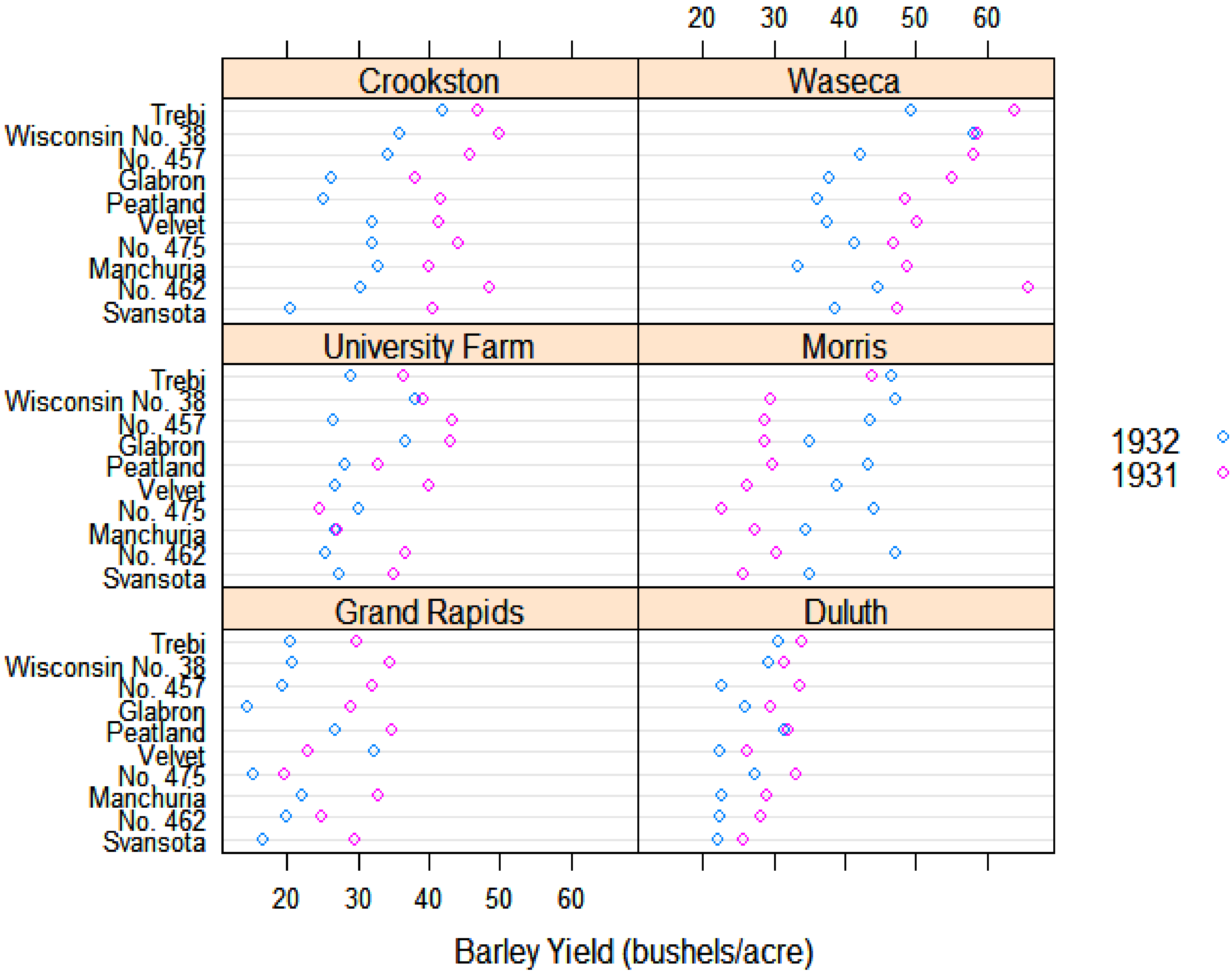
- except that angle is constant and radius varies
  - in pies, radius is constant and angle varies
  - that’s why pie segments need labels
- watch out for colors

scatterplots

- can (sometimes) show more than two variables
  - can code categorical variables with color
  - can code some interval variables with size
- small multiples can show varying conditions
  - lattice (i.e., trellis) plots
  - use same scale and ranges, if possible, to enhance comparison

dotplots (and trellis)

- conditioning plots
- barley yield
  - ten varieties
  - six plots
  - two years

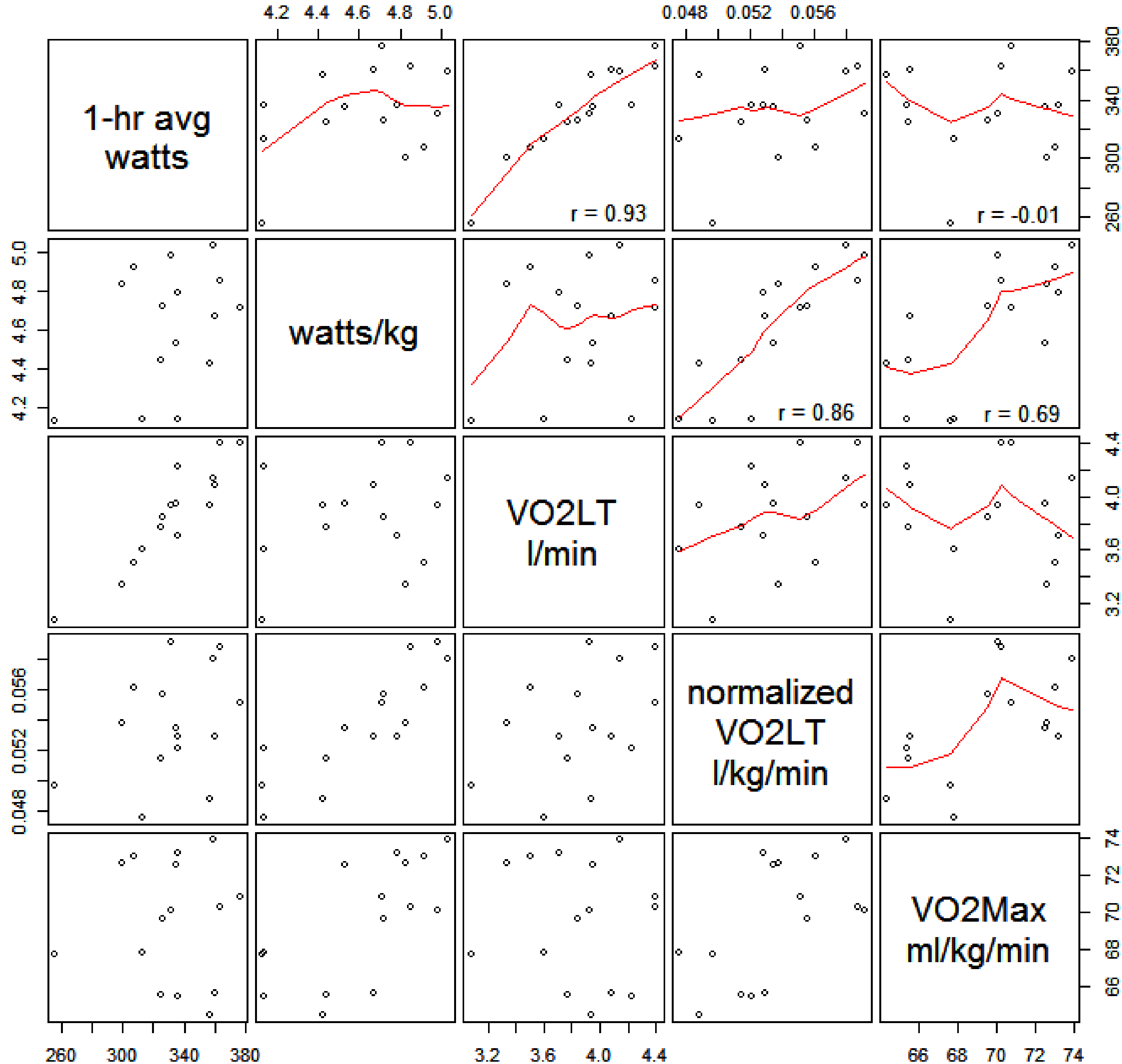


Crookston
Waseca
University Farm
Morris
Grand Rapids
Duluth
Trebi
Wisconsin No. 38
No. 457
Glabron
Peatland
Velvet
No. 475
Manchuria
No. 462
Svansota
20
30
40
50
60
1932
1931
Barley Yield (bushels/acre)

scatterplot matrices

- scatterplot matrices compress a lot of information on bivariate relationships into a small space
    - useful for winnowing out uninteresting variables and deciding which variables might be worth further investigation

Coyle, 1991

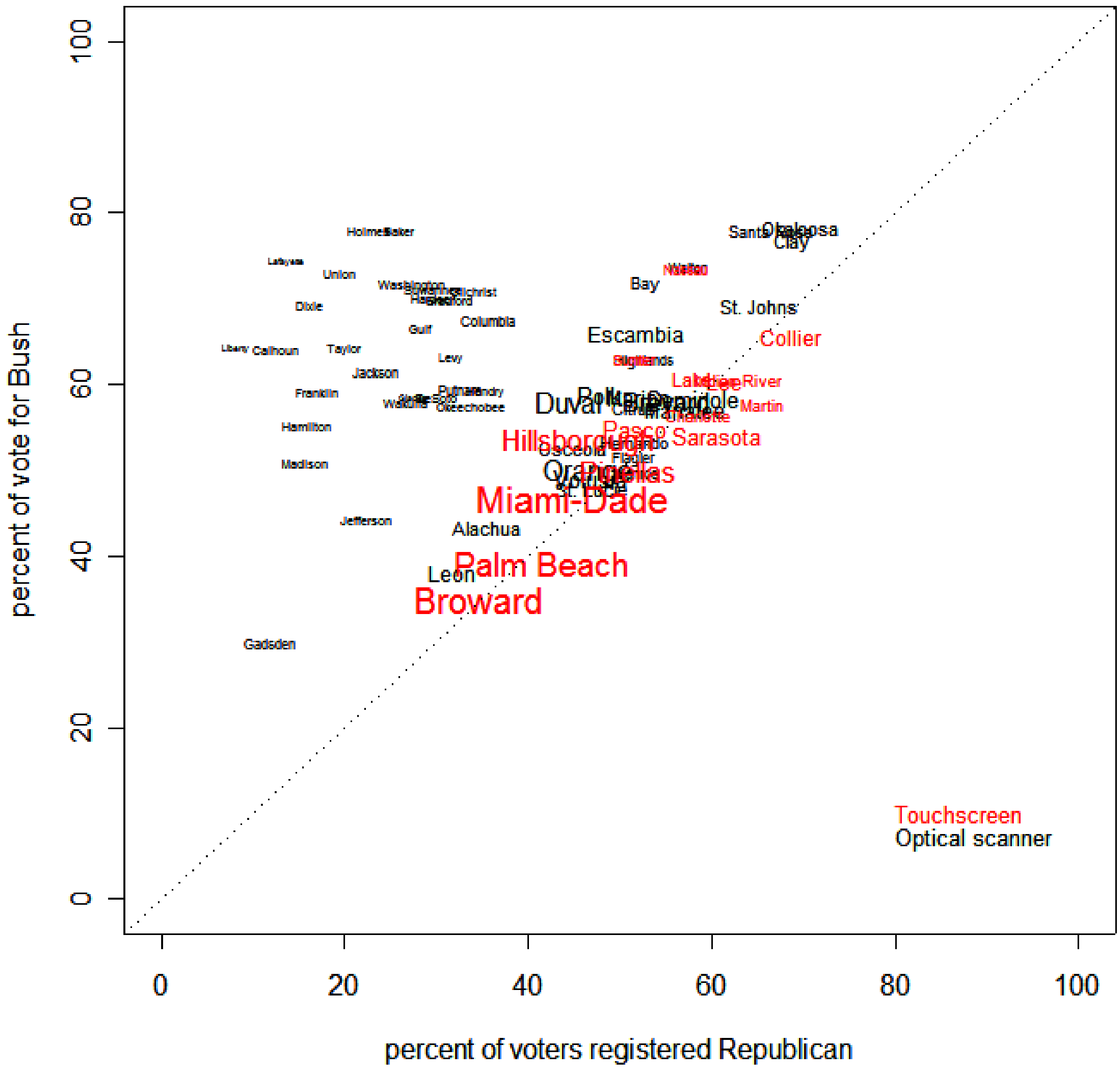


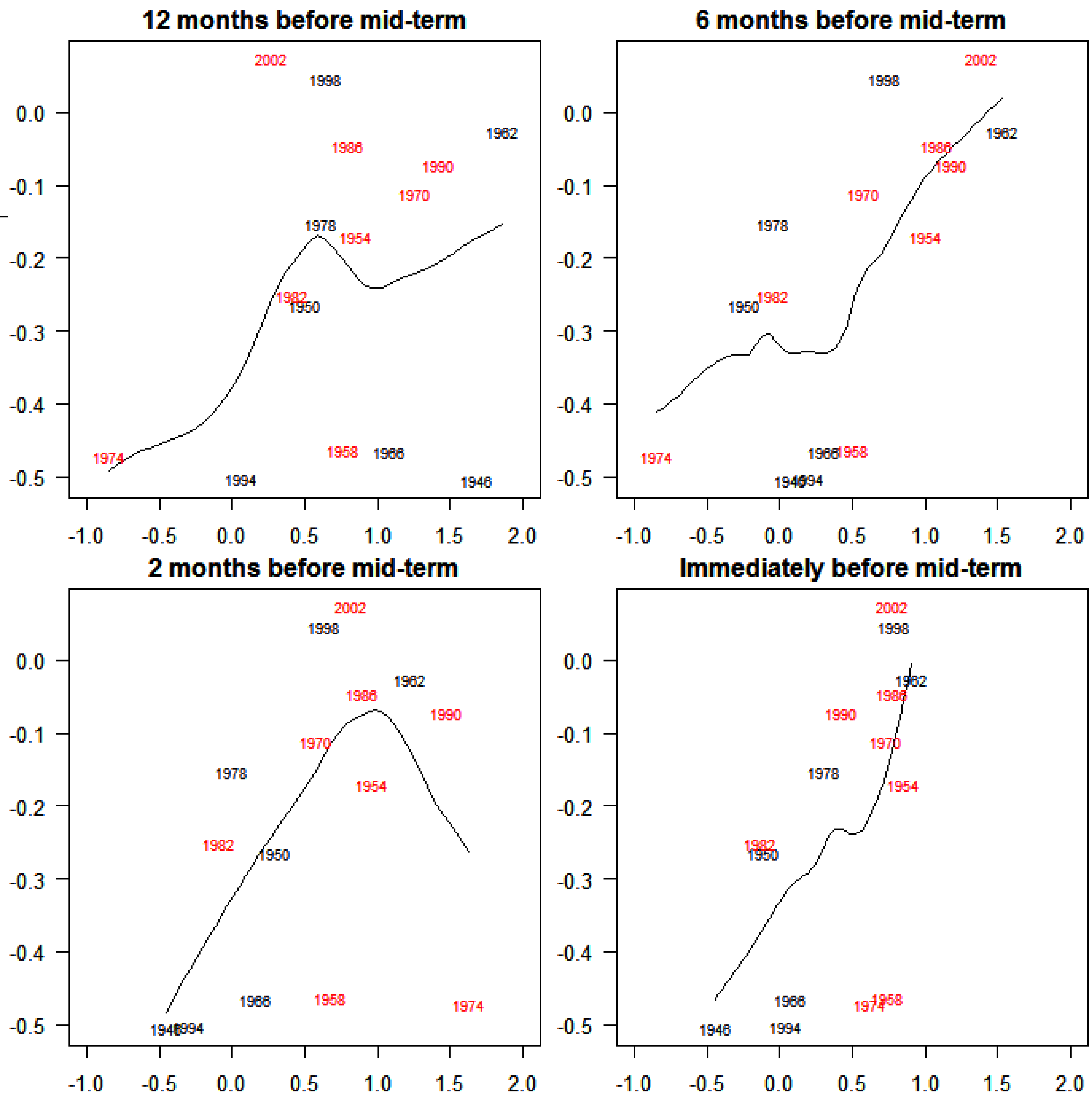
1-hr avg watts
watts/kg
VO2LT l/min
normalized VO2LT l/kg/min
VO2Max ml/kg/min
r = 0.93
r = -0.01
r = 0.86
r = 0.69
4.2 4.4 4.6 4.8 5.0
0.048 0.052 0.056
260 300 340 380
3.2 3.6 4.0 4.4
66 68 70 72 74

coding plotting symbols

- improves information density by tagging plotting symbols with attributes
    - you've seen this before using color or shape; can often combine with direct labeling

Florida vote by county



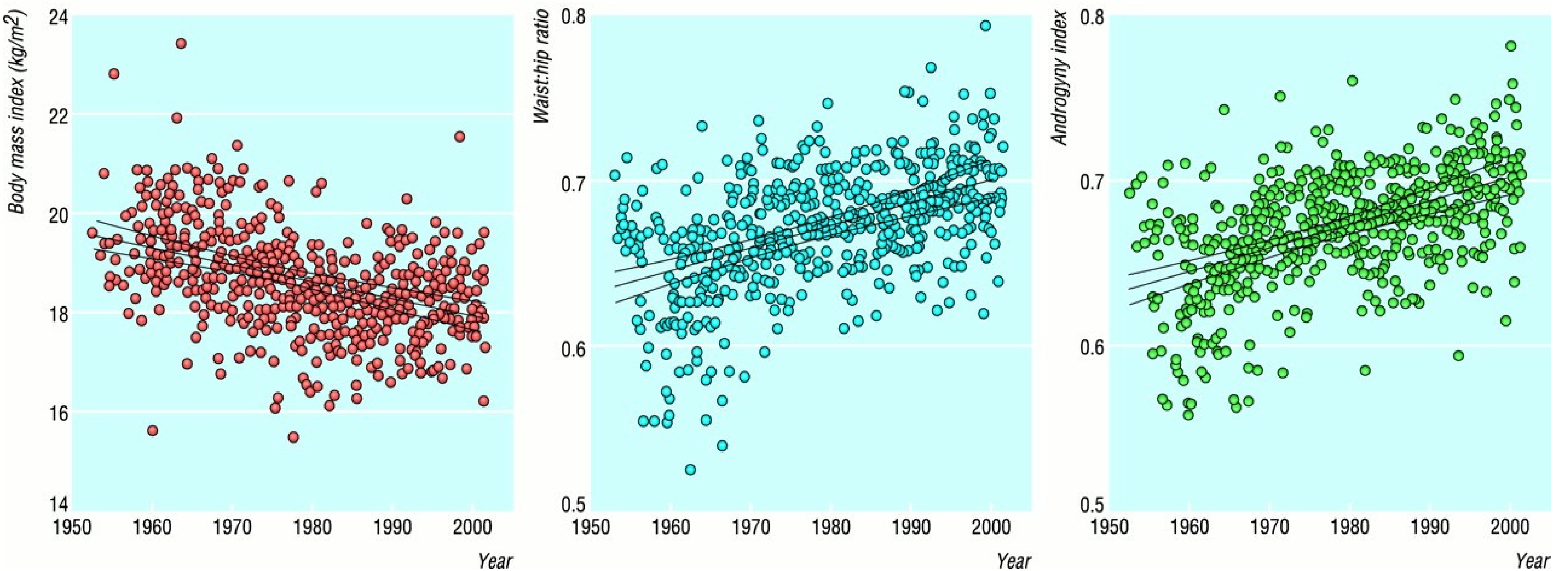


12 months before mid-term
6 months before mid-term
2 months before mid-term
Immediately before mid-term
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2002
1998
1962
1986
1990
1970
1978
1954
1982
1950
1974
1958
1966
1994
1946

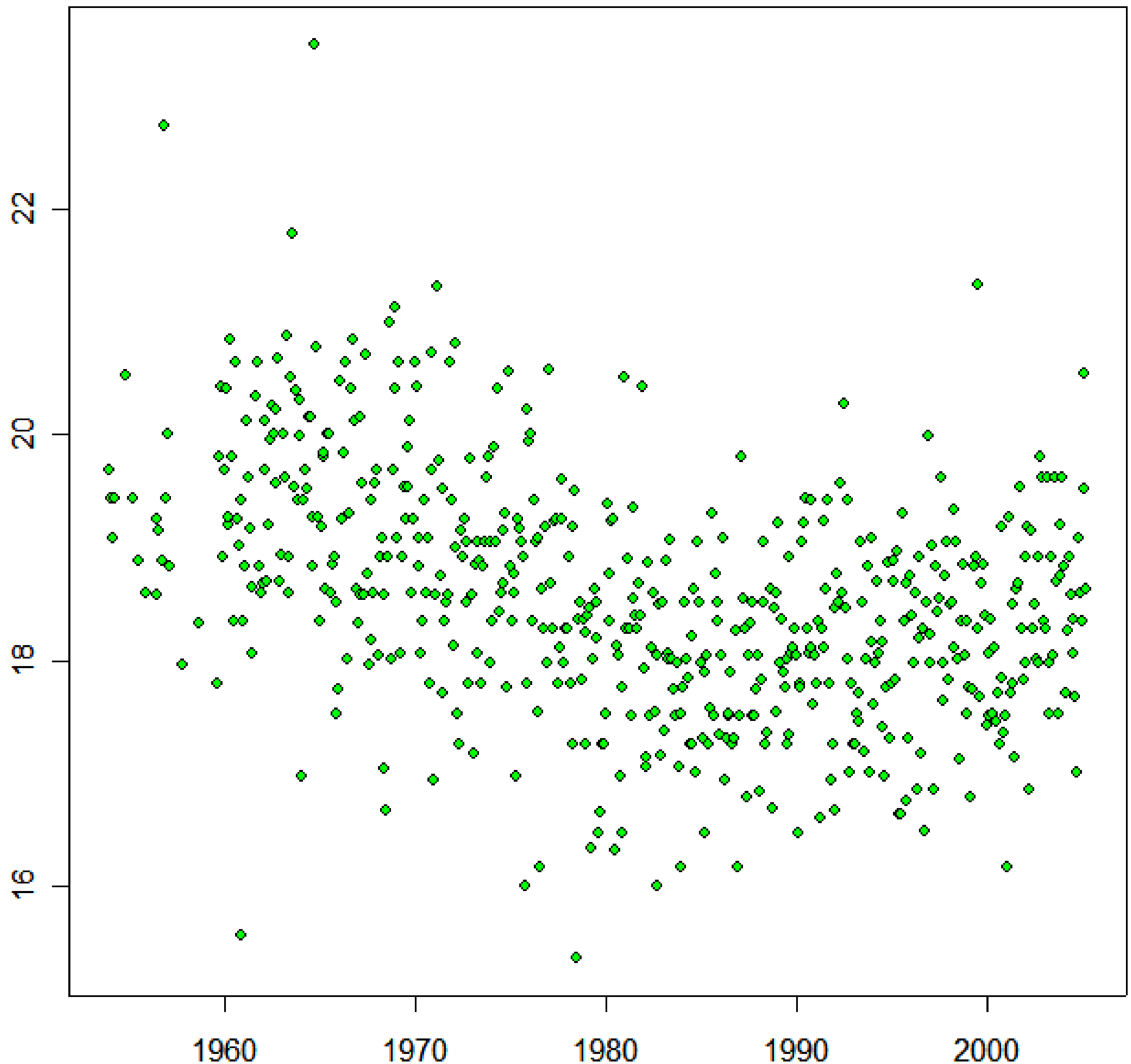
smoothing and straightening

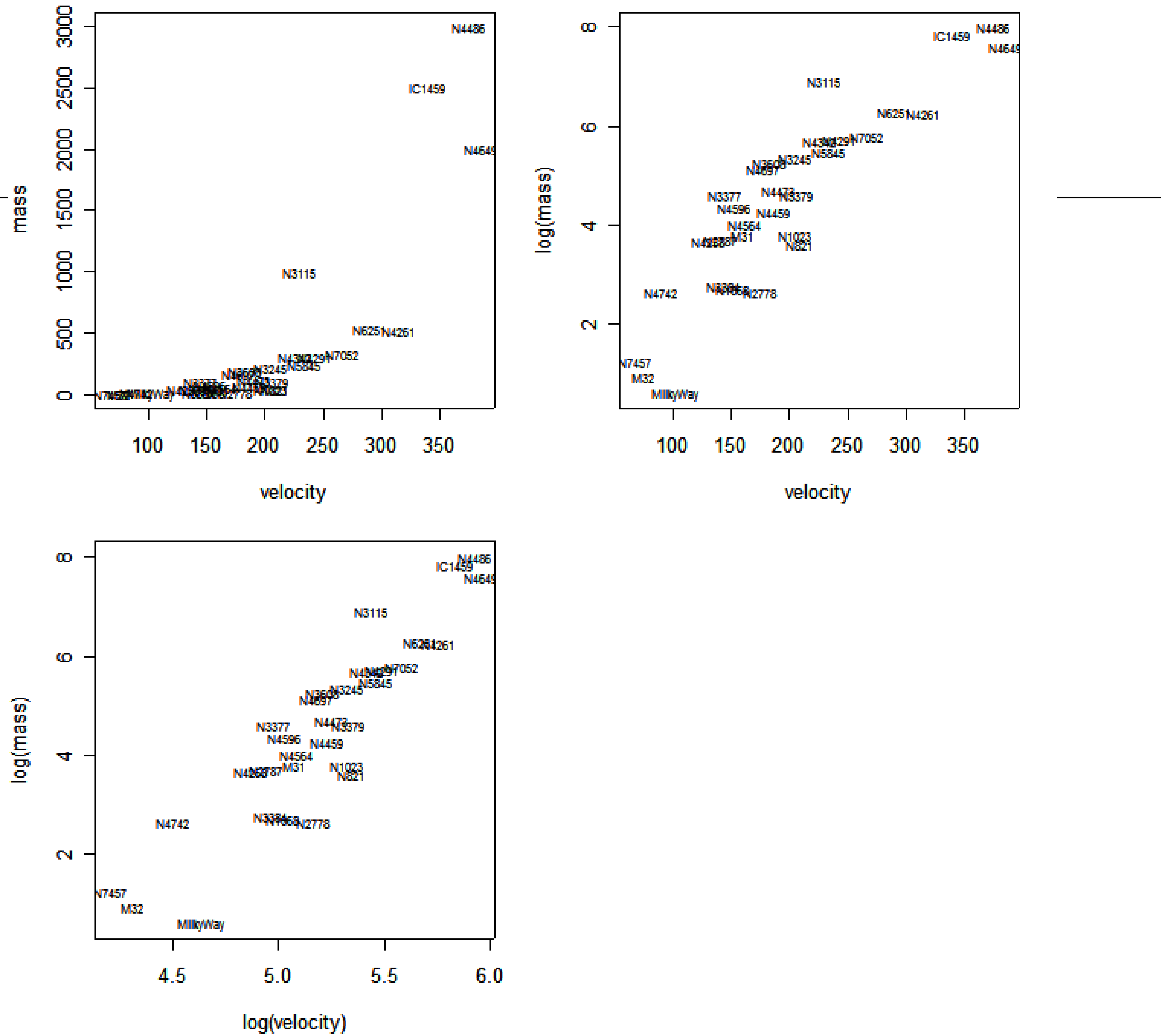
- smooth lines
  - piecewise linearity
  - splines and lo(w)ess
- a ladder of re-expression
- the re-expression rule

shapely centrefolds?



BMI for Playboy centerfolds, 1953-2005




mass
velocity
0
500
1000
1500
2000
2500
3000
100
150
200
250
300
350
N4486
IC1459
N4649
N3115
N6251
N4261
N7052
N3245
N5845
log(mass)
velocity
2
4
6
8
N3377
N4596
N4564
M31
N4473
N3379
N4459
N1023
N821
N4742
N2778
N7457
M32
MilkyWay
N4697
log(mass)
log(velocity)
4.5
5.0
5.5
6.0

a ladder of re-expressions...

3

2

1

½

#

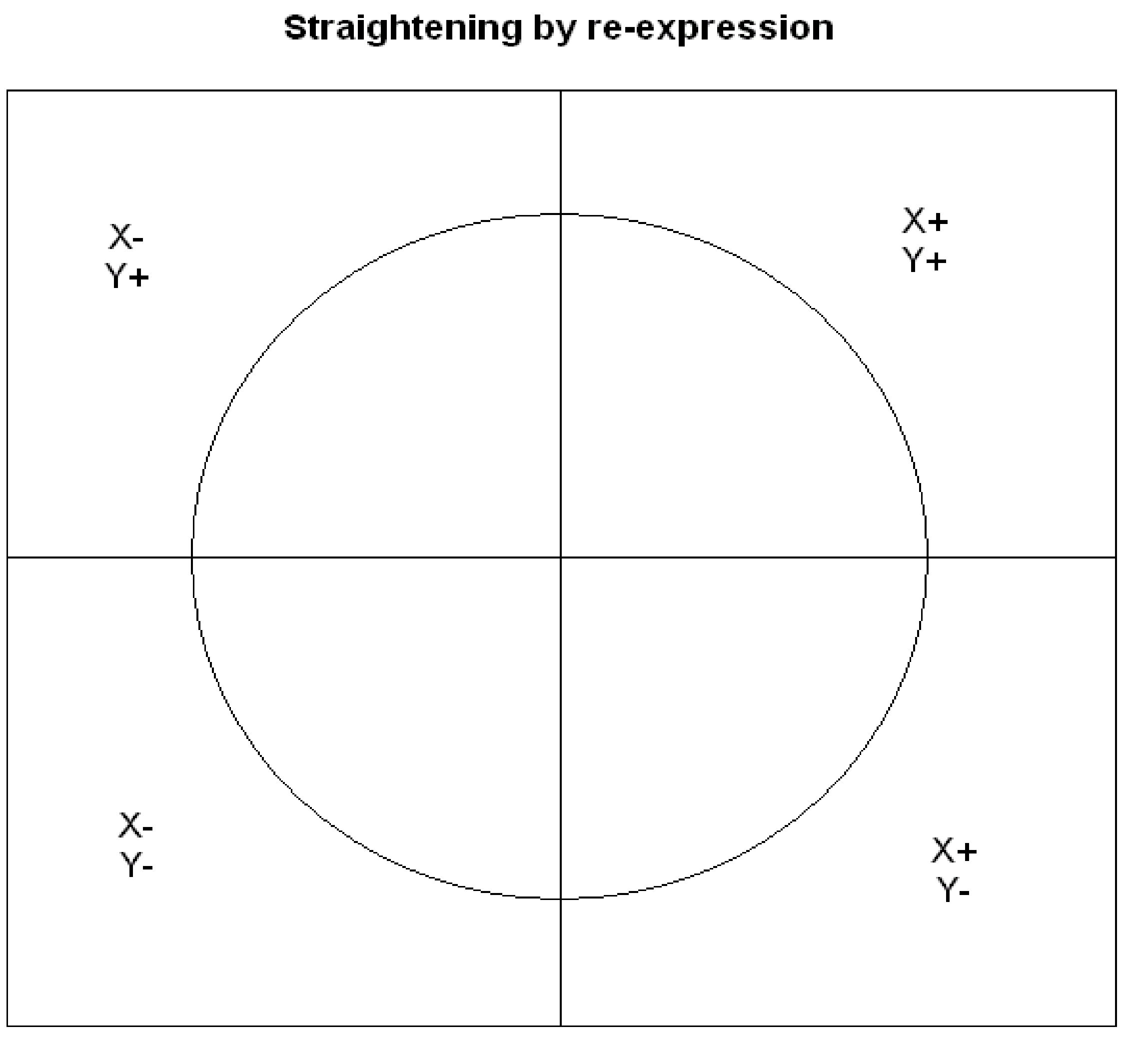
- ½

-1

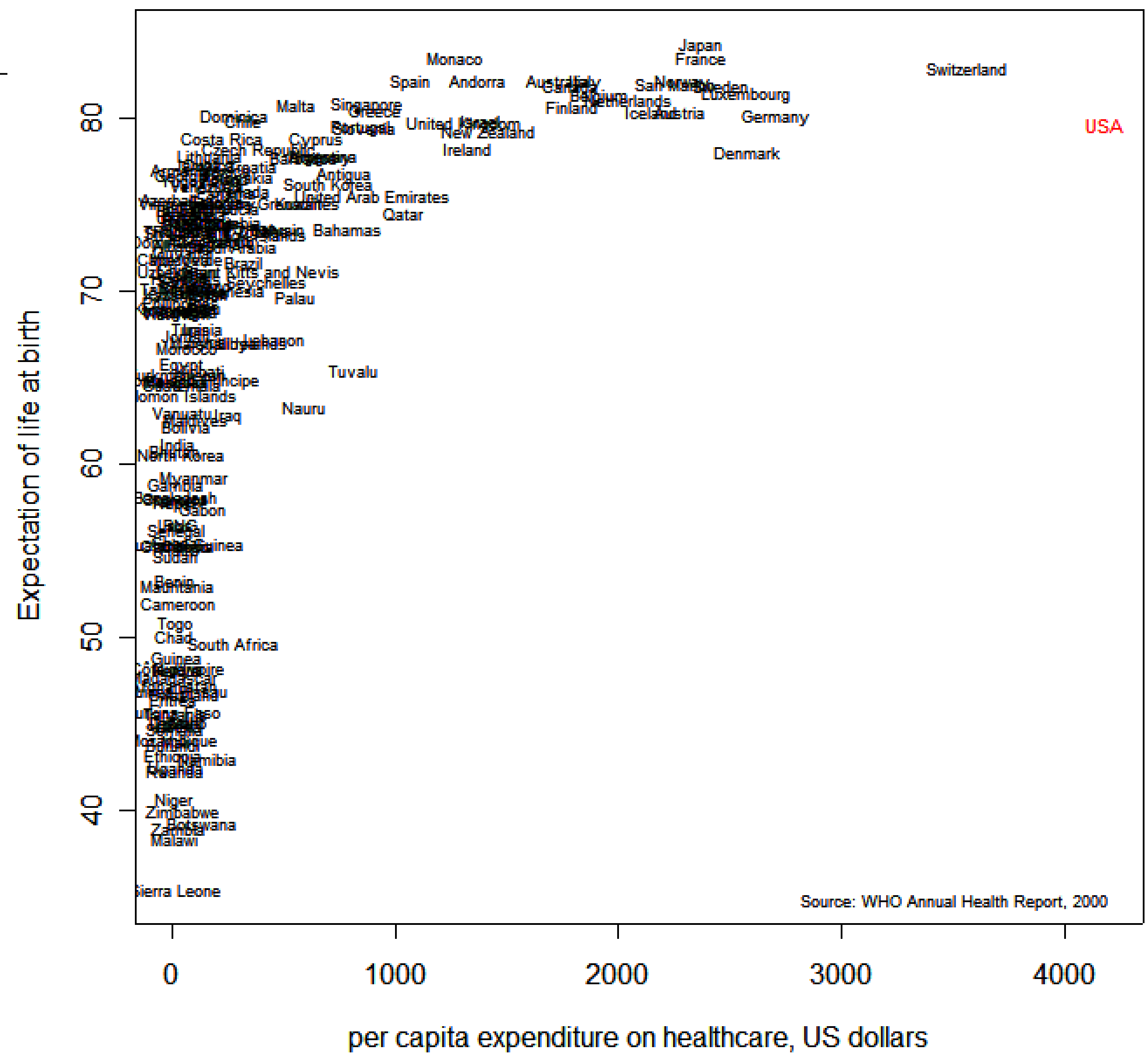
-2

-3

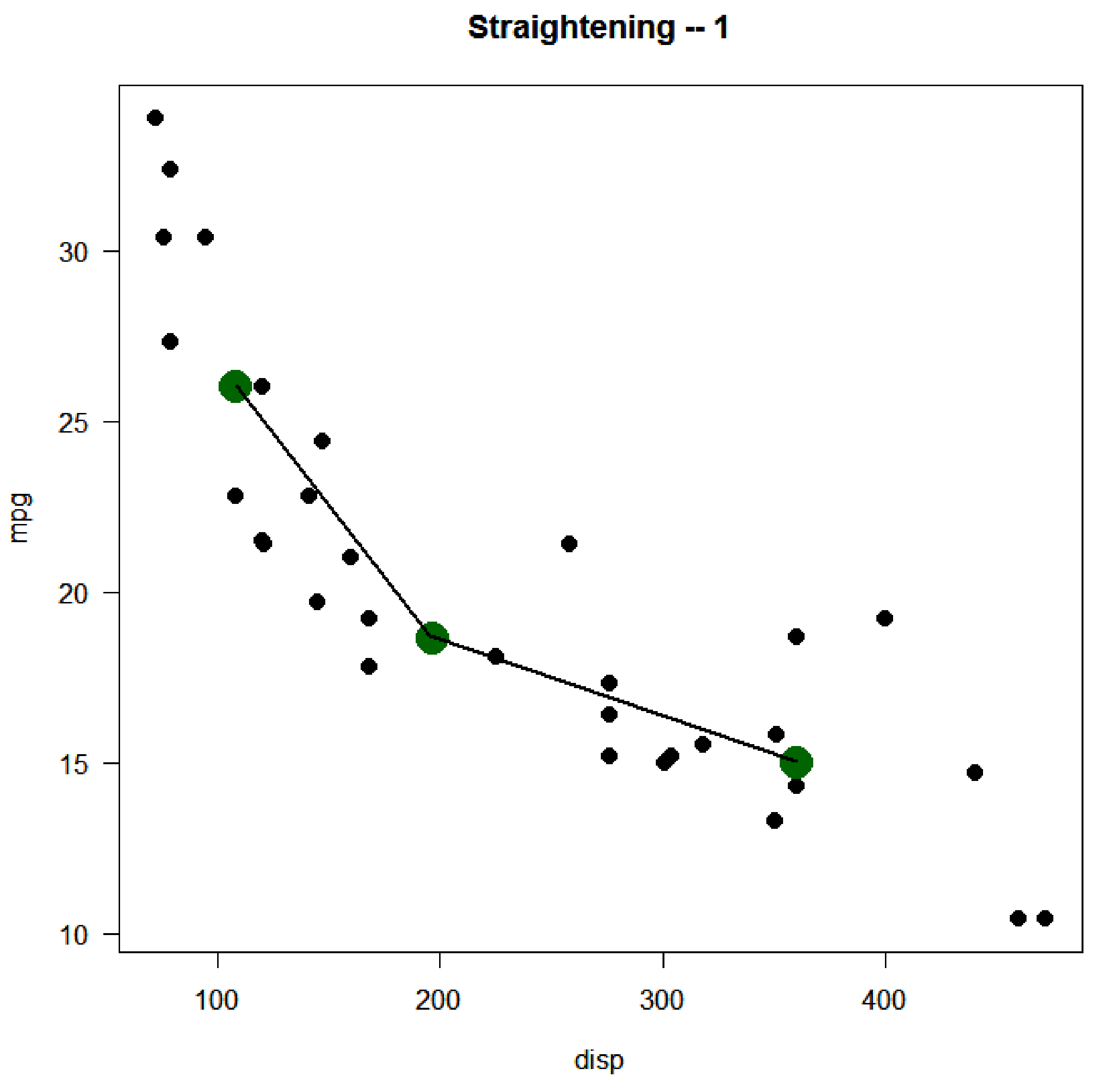
...and a rule for using them



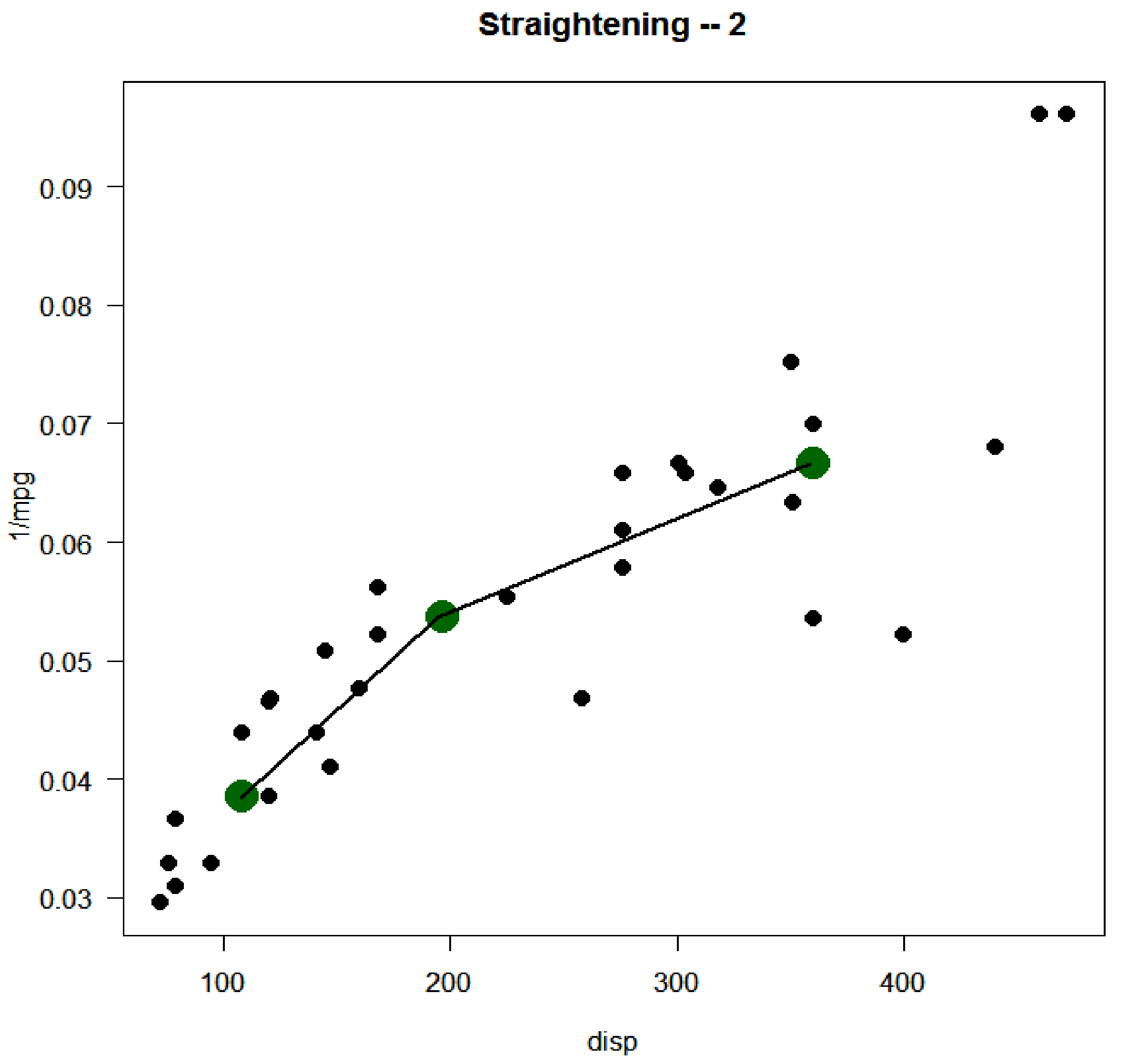
Life expectancy and national expenditures



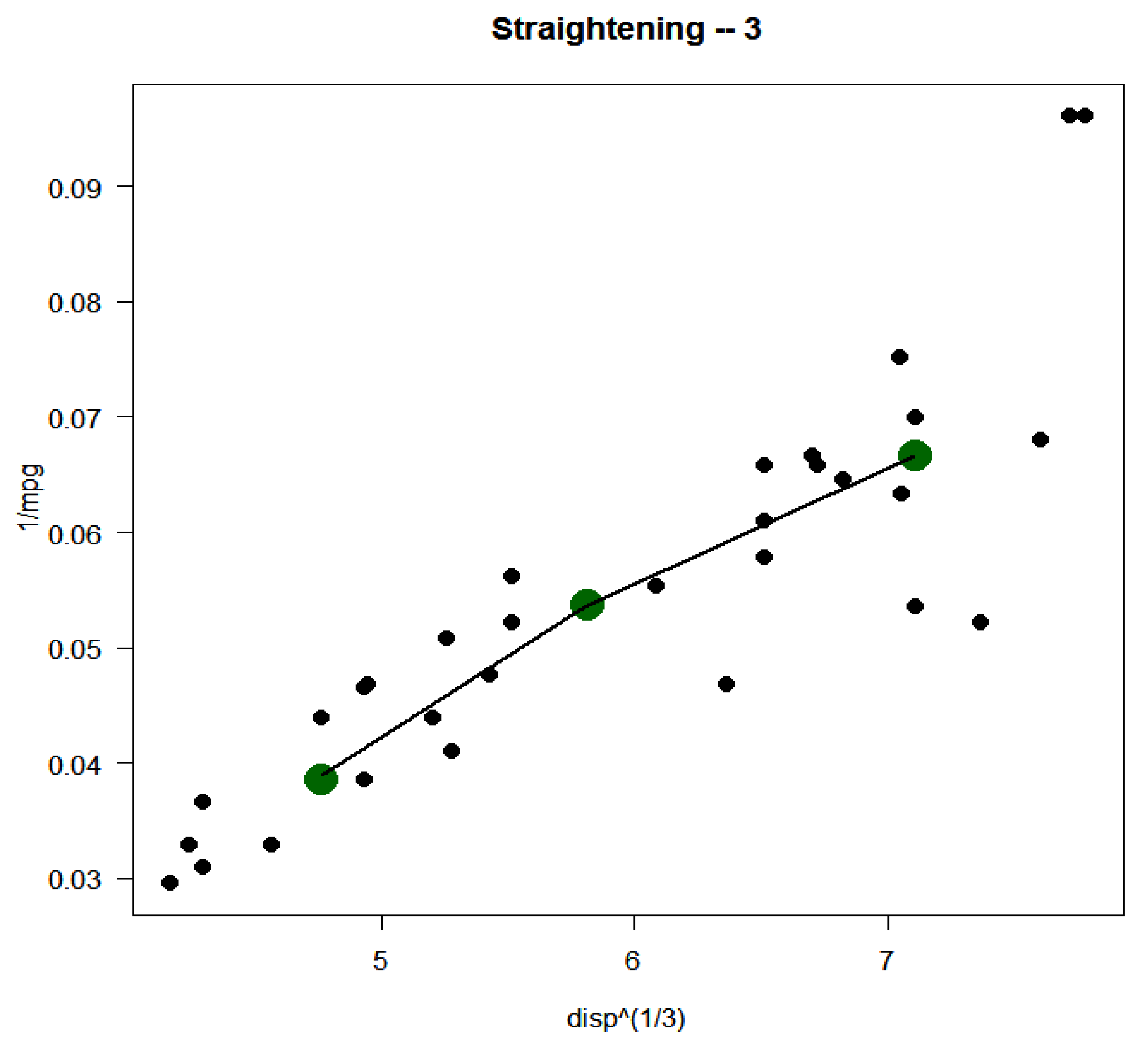
straightening



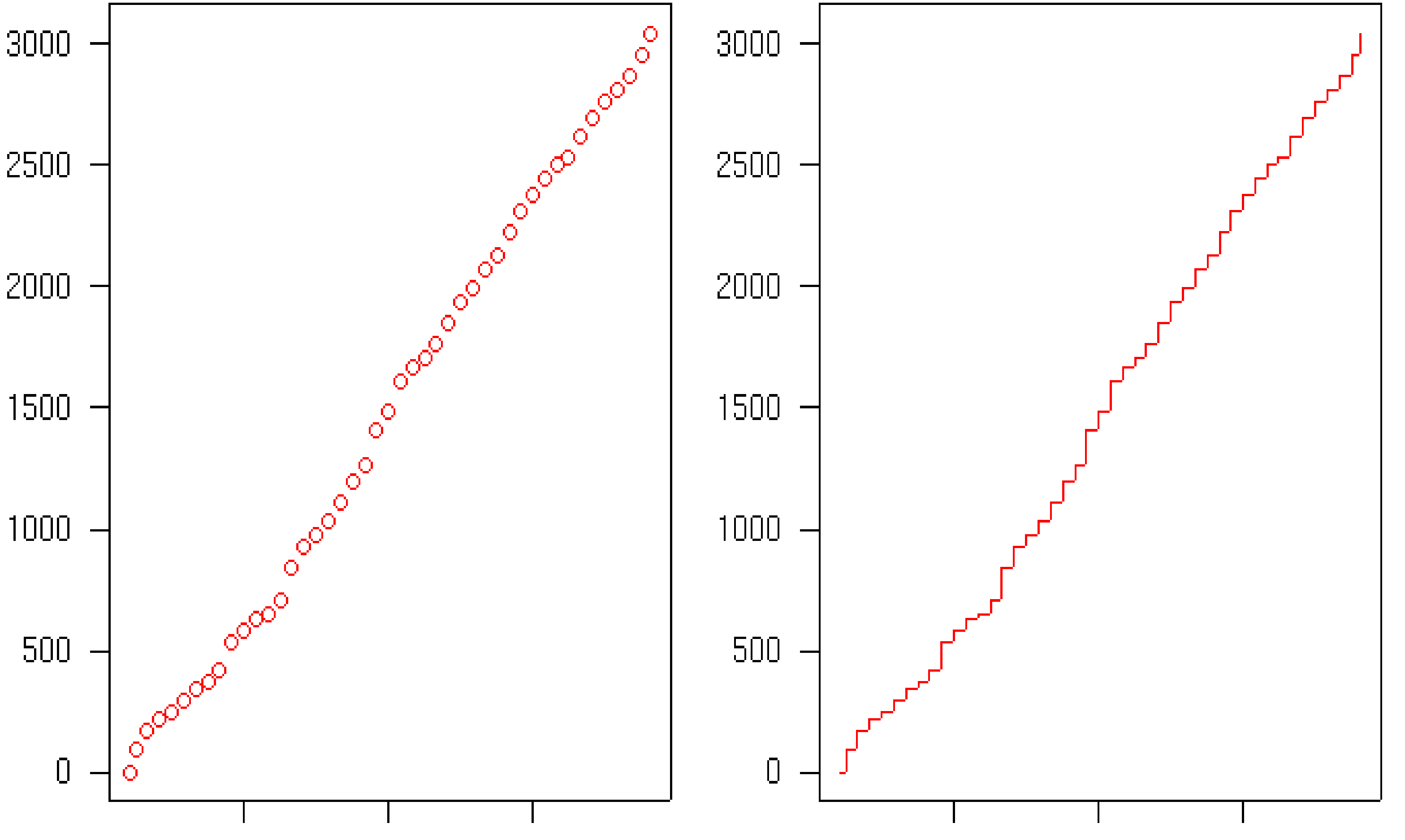
straightening 2



straightening 3

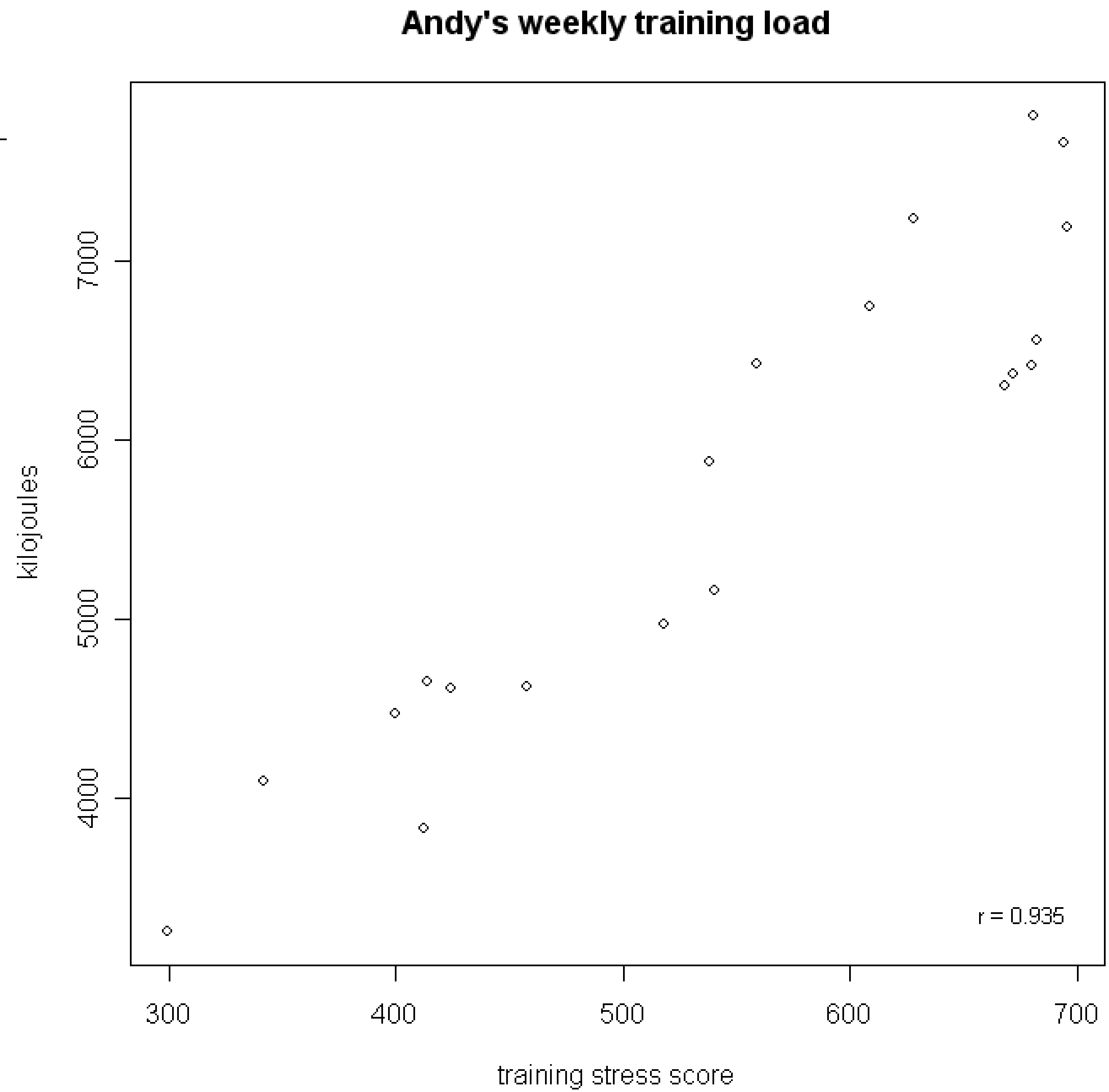


when is smooth too smooth?

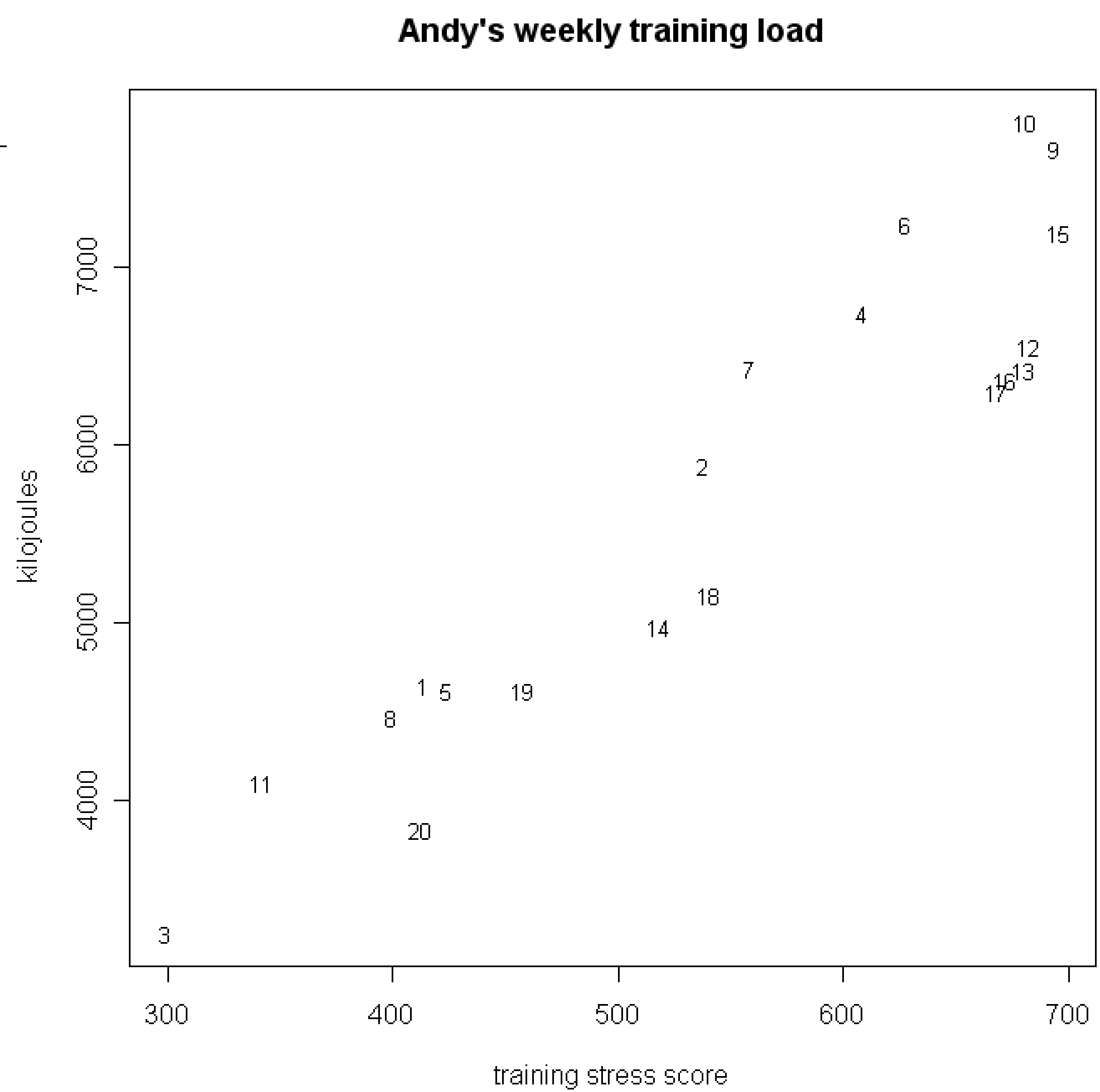


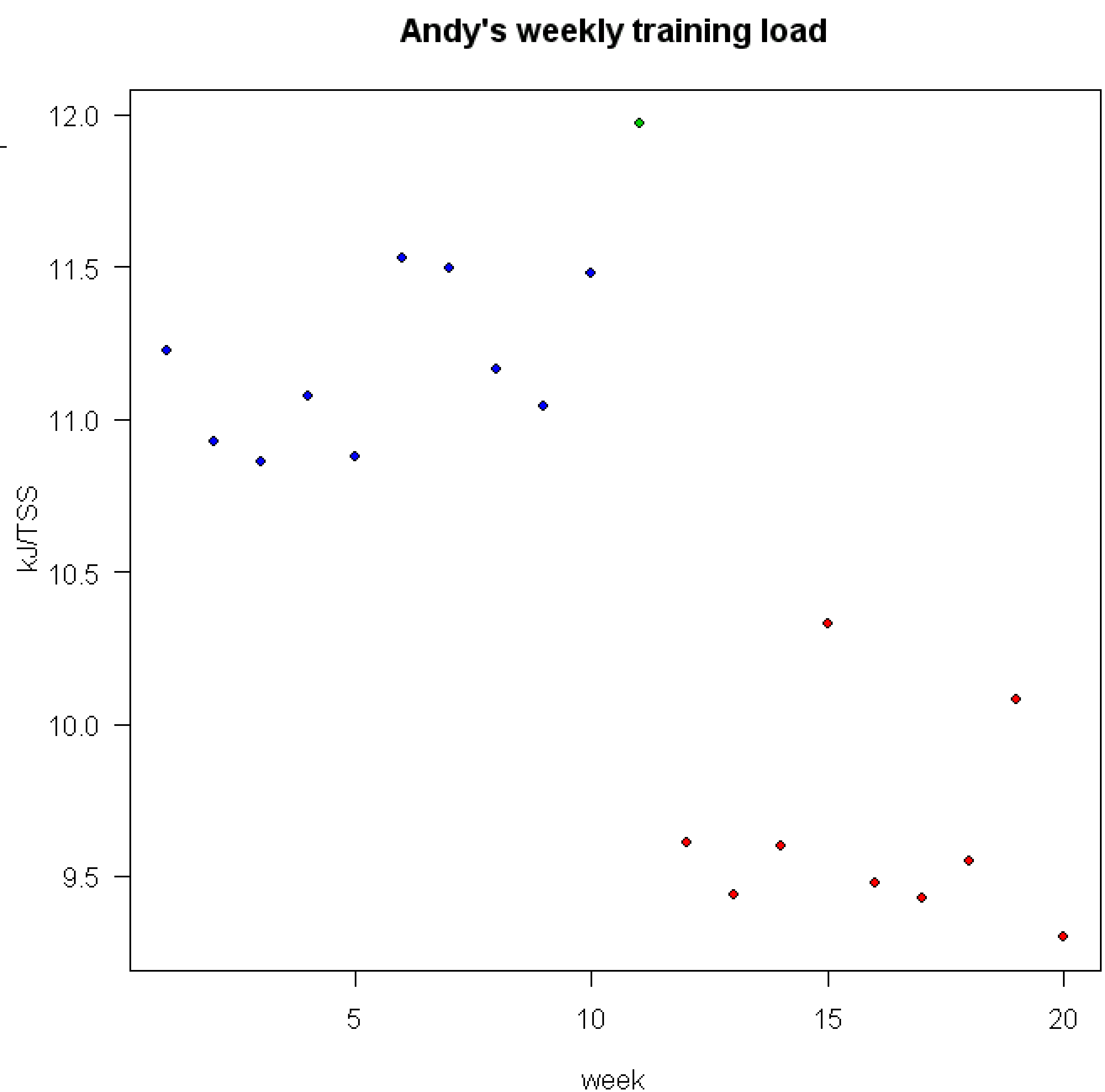
when is straight too straight?

- can straight be too straight?



Andy's weekly training load
kilojoules
7000
6000
5000
4000
r = 0.935
300
400
500
600
700
training stress score


Andy's weekly training load
kilojoules
training stress score
4000
5000
6000
7000
300
400
500
600
700
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

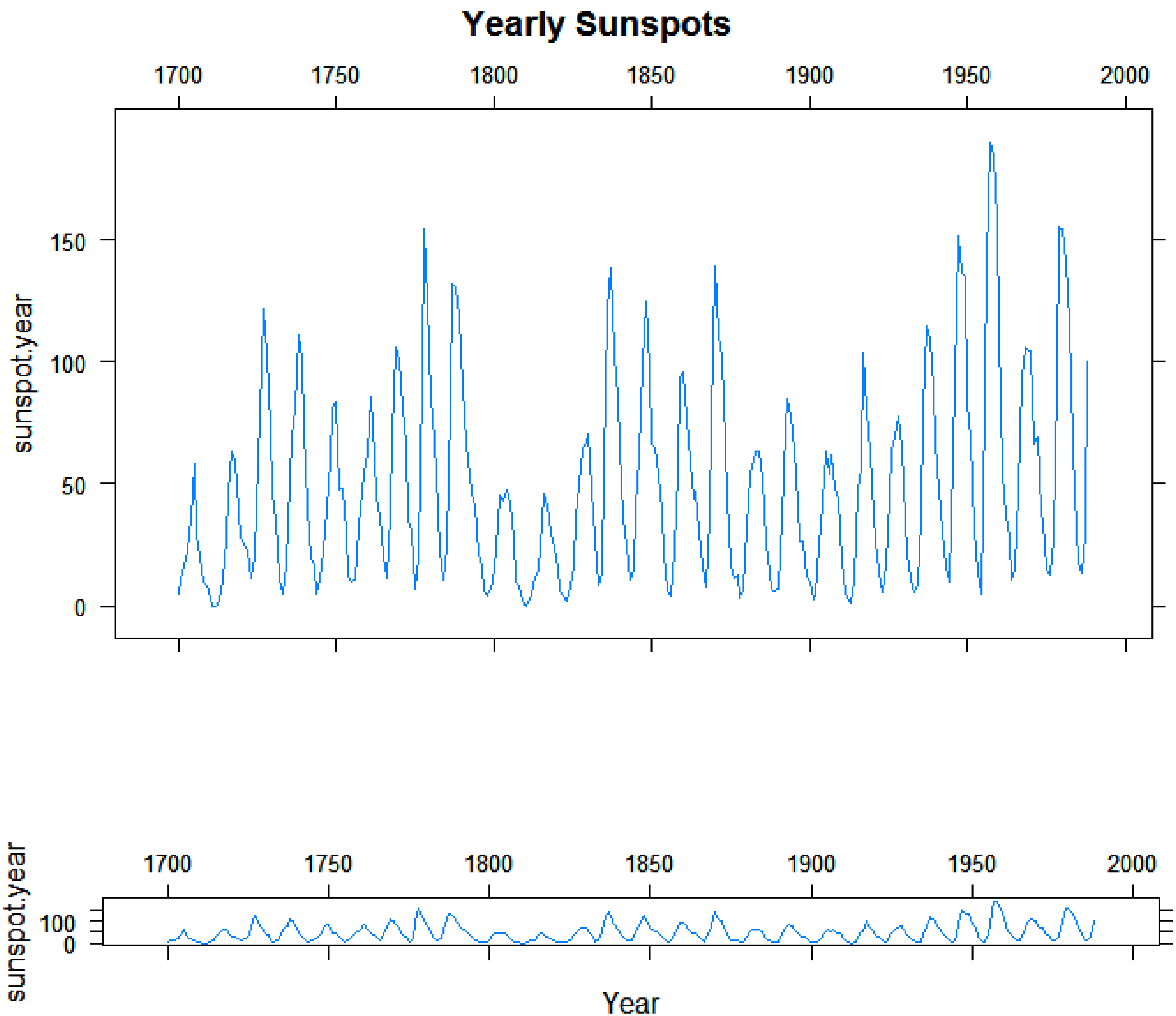


Andy's weekly training load
kJ/TSS
12.0
11.5
11.0
10.5
10.0
9.5
5
10
15
20
week

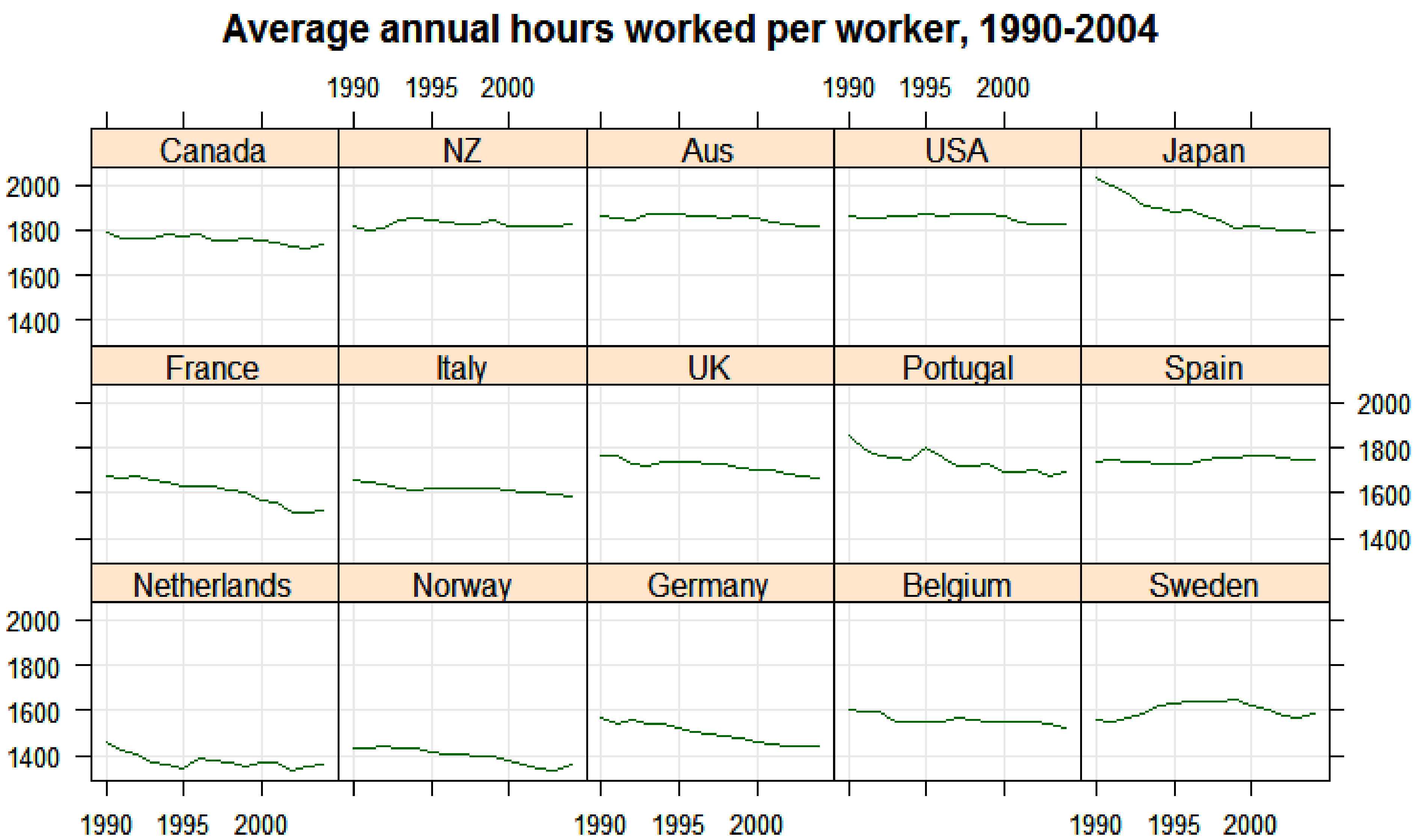
aspect ratios and banking

- human eye isn't great at decoding angles

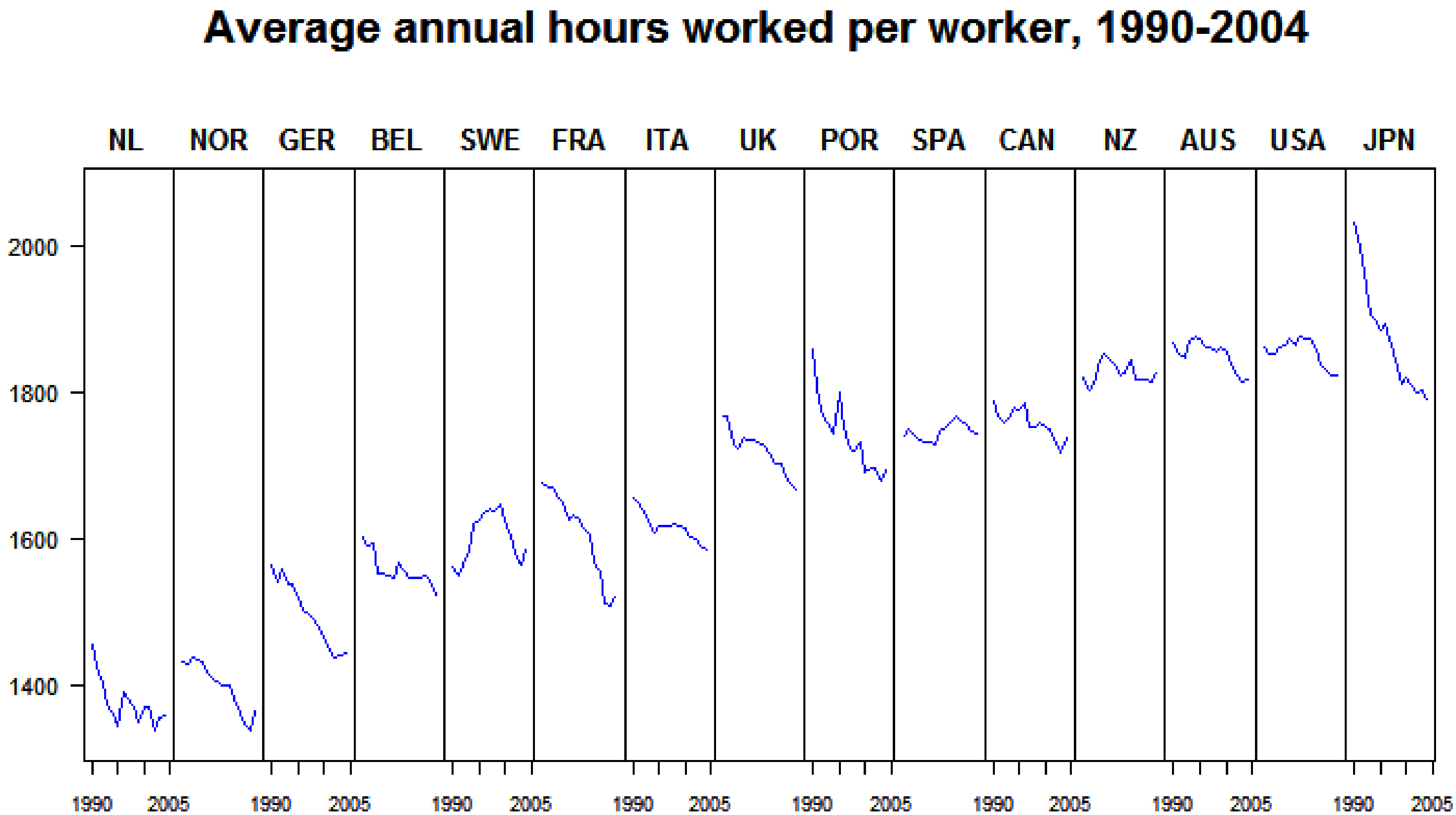
  banking helps the eye to decipher angles



Yearly Sunspots
1700
1750
1800
1850
1900
1950
2000
sunspot.year
150
100
50
0
100
0
Year



Average annual hours worked per worker, 1990-2004
Canada
NZ
Aus
USA
Japan
France
Italy
UK
Portugal
Spain
Netherlands
Norway
Germany
Belgium
Sweden
1990
1995
2000
1400
1600
1800
2000

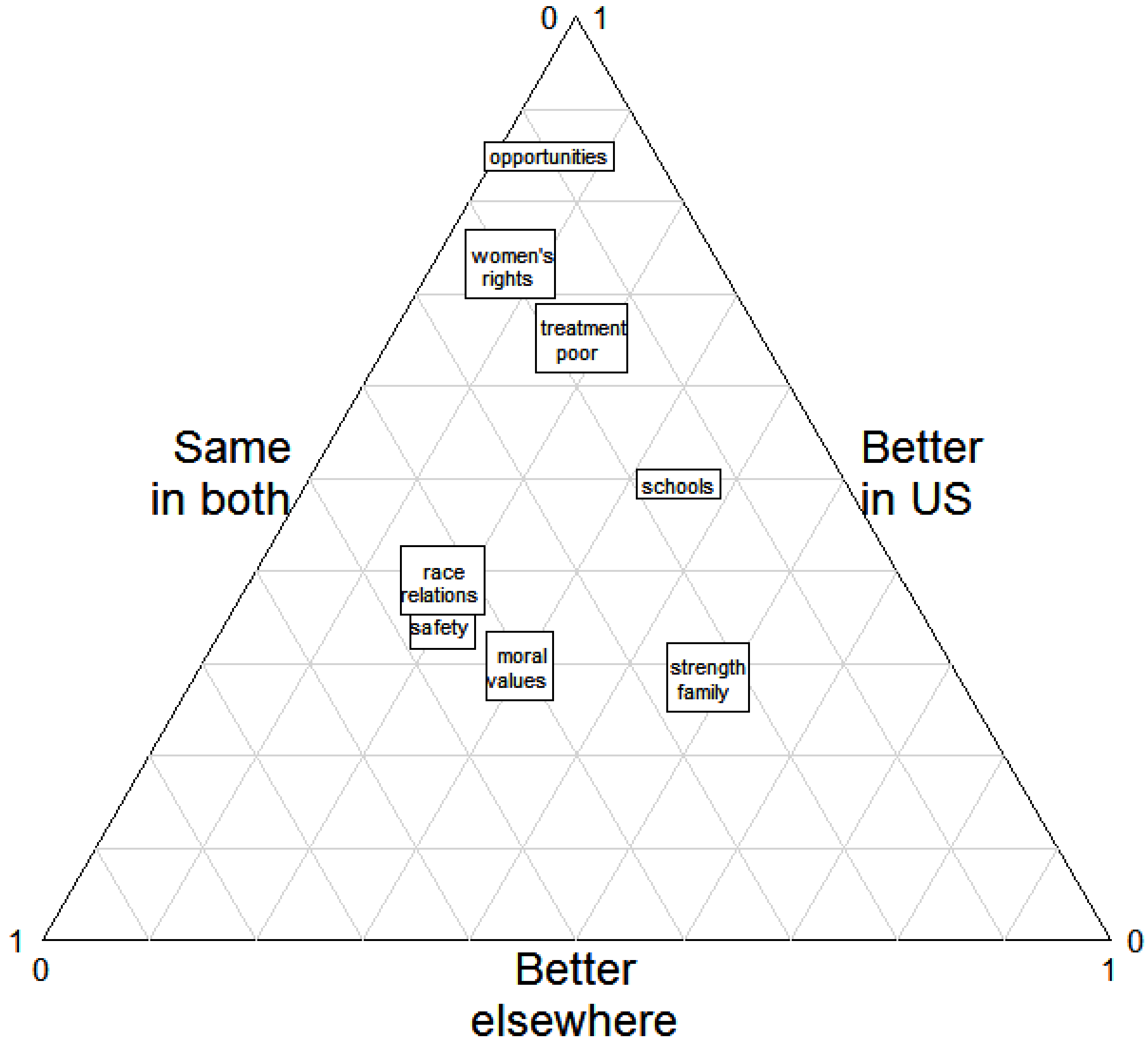


Average annual hours worked per worker, 1990-2004
NL
NOR
GER
BEL
SWE
FRA
ITA
UK
POR
SPA
CAN
NZ
AUS
USA
JPN
2000
1800
1600
1400
1990
2005

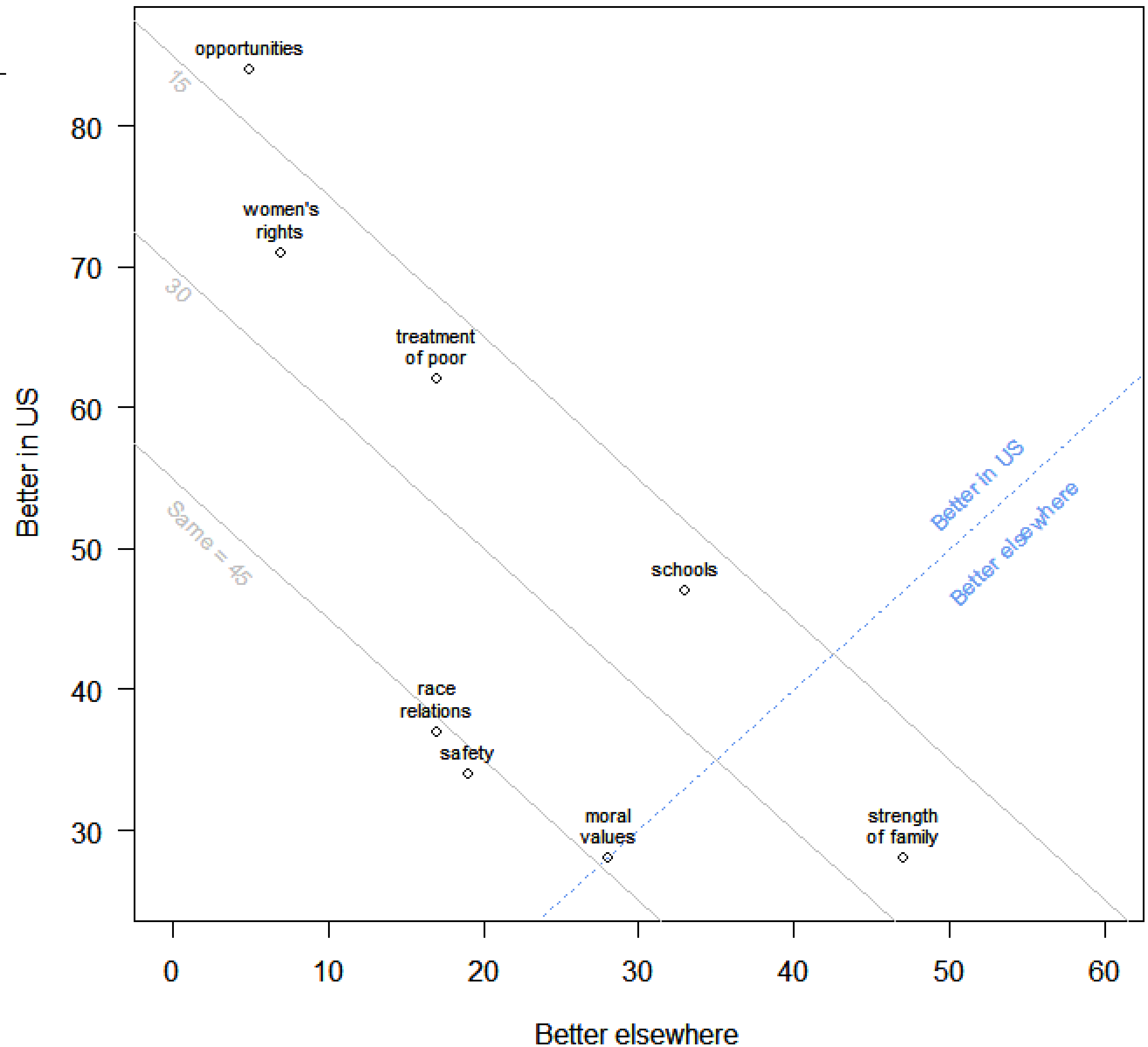
contours and bases

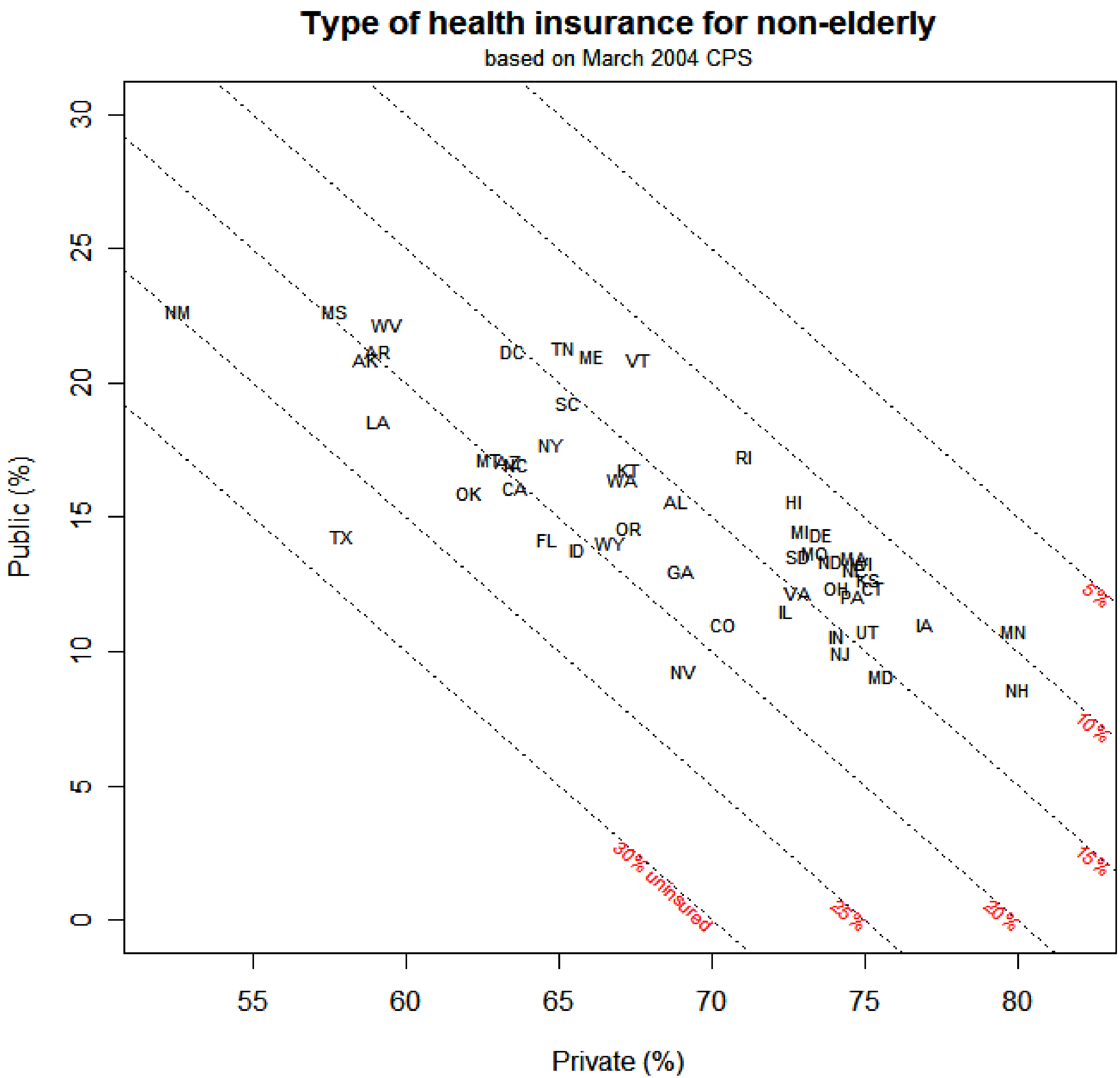
- triangle plots (= ternary plots)
  - soil texture plots
  - two degrees of freedom
- function contours can add context

Immigrant's assessments of US vs. country of origin

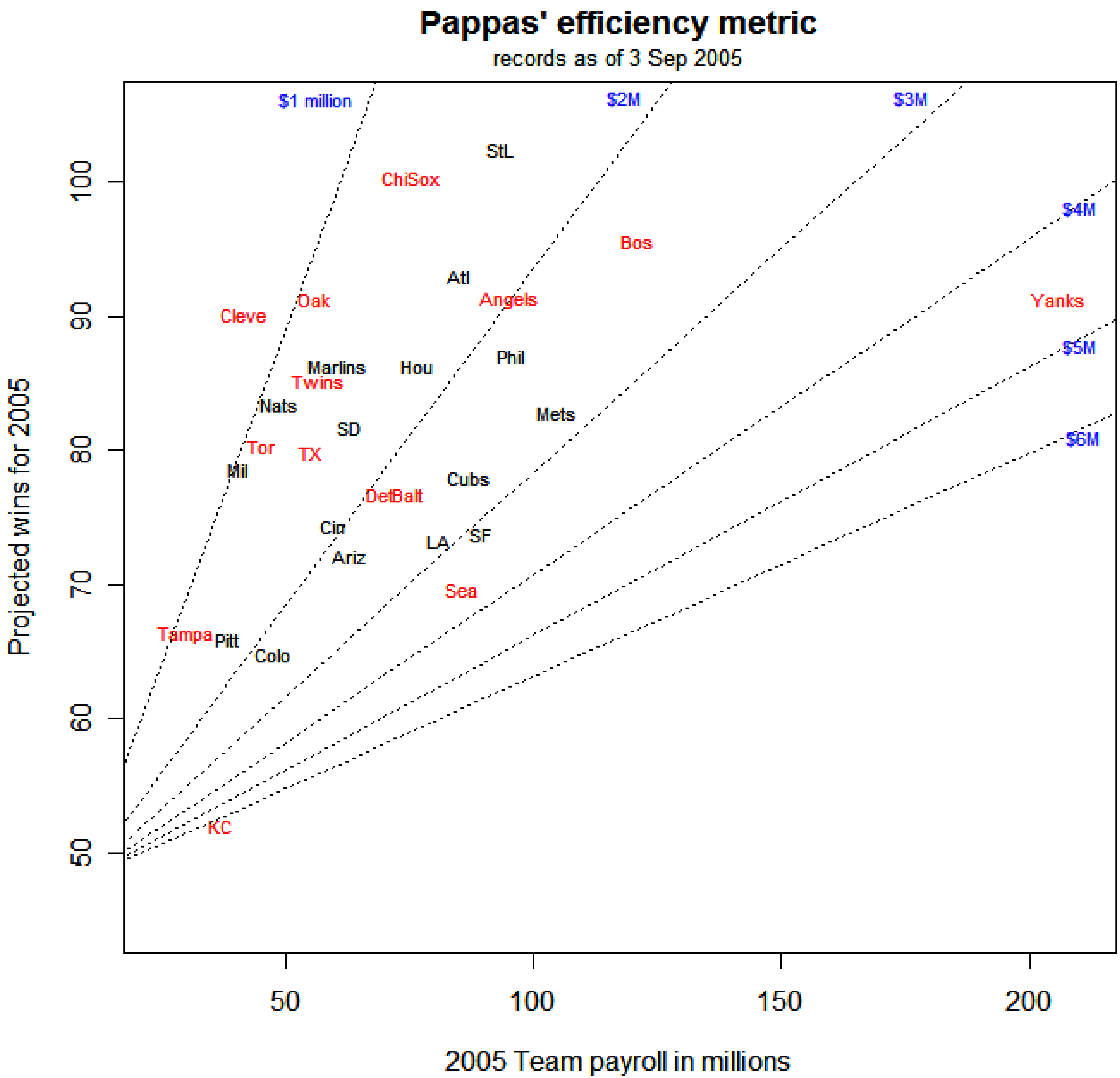


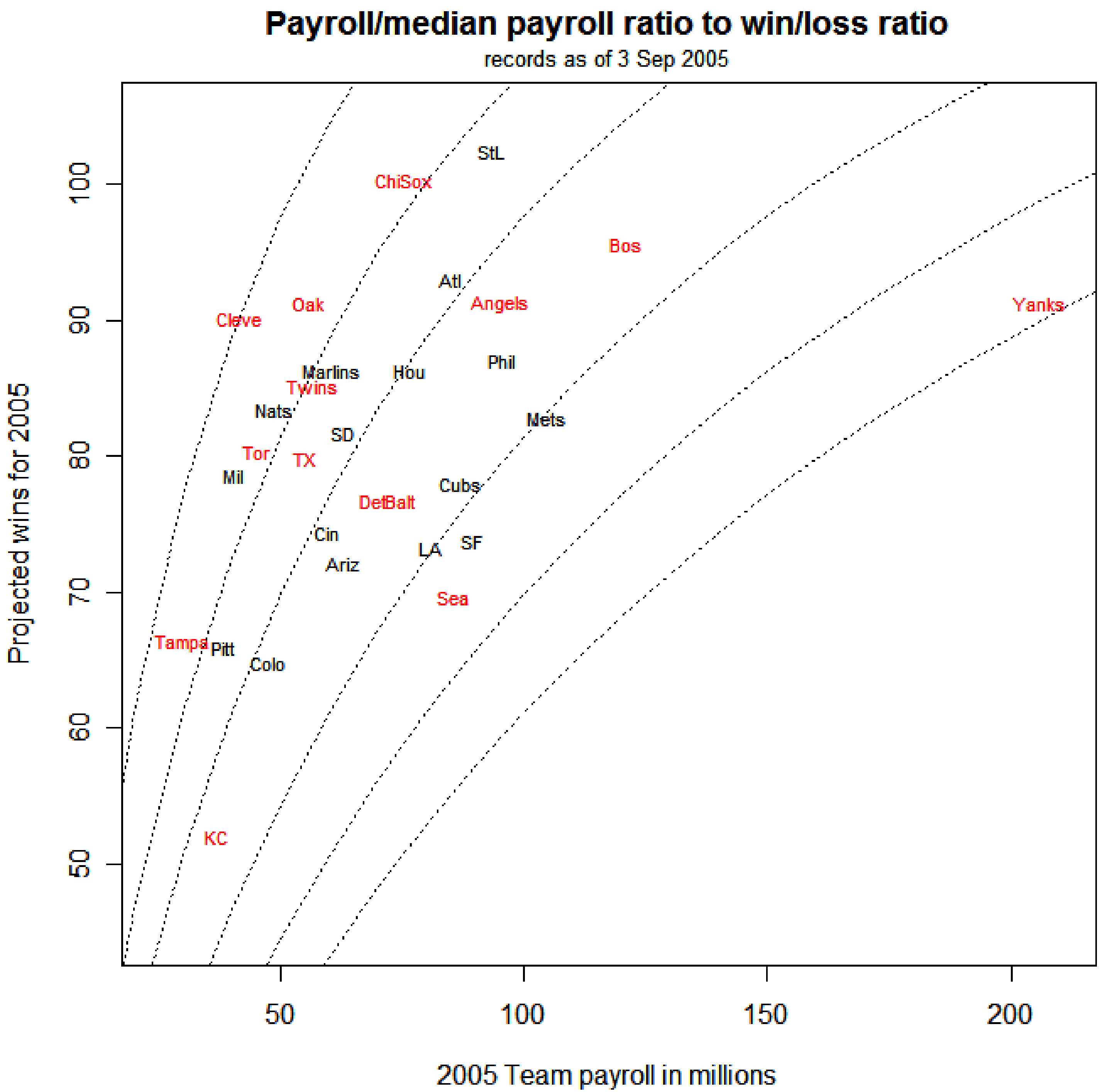
Immigrant's assessments of US vs. country of origin





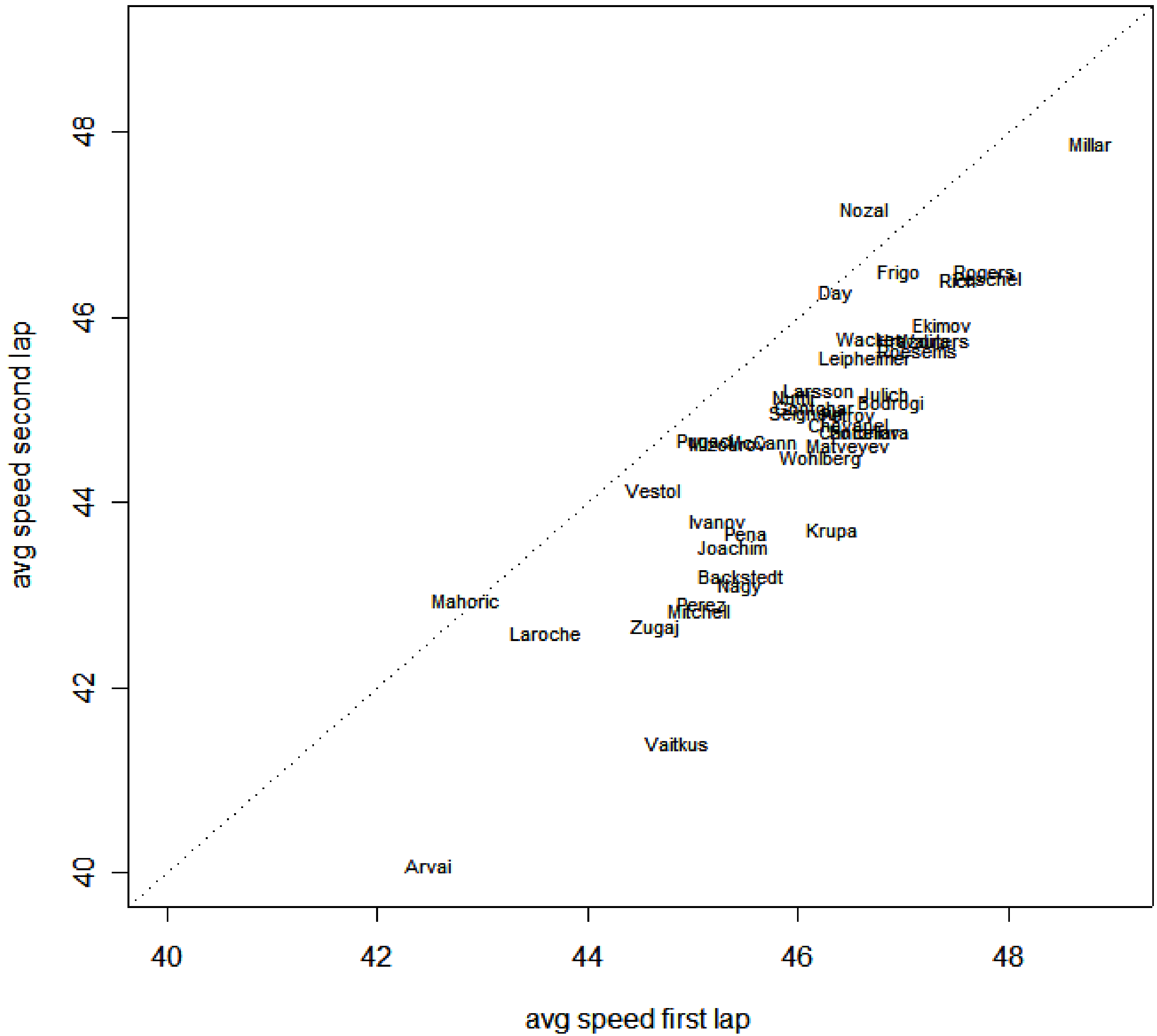
Type of health insurance for non-elderly
based on March 2004 CPS
Public (%)
Private (%)
30% uninsured
25%
20%
15%
10%
5%
NM
MS
WV
AR
AK
DC
TN
ME
VT
SC
LA
NY
RI
MT
AZ
NC
KY
WA
OK
CA
AL
HI
TX
FL
ID
WY
OR
MI
DE
SD
MO
ND
MA
NE
WI
KS
GA
OH
CT
PA
VA
IL
CO
IN
UT
IA
MN
NJ
NV
MD
NH


Pappas' efficiency metric
records as of 3 Sep 2005
$1 million
$2M
$3M
$4M
$5M
$6M
StL
ChiSox
Bos
Atl
Angels
Oak
Cleve
Yanks
Phil
Marlins
Hou
Twins
Nats
Mets
SD
Tor
TX
Mil
Cubs
DetBalt
Cin
SF
LA
Ariz
Sea
Tampa
Pitt
Colo
KC
100
90
80
70
60
50
Projected wins for 2005
50
100
150
200
2005 Team payroll in millions



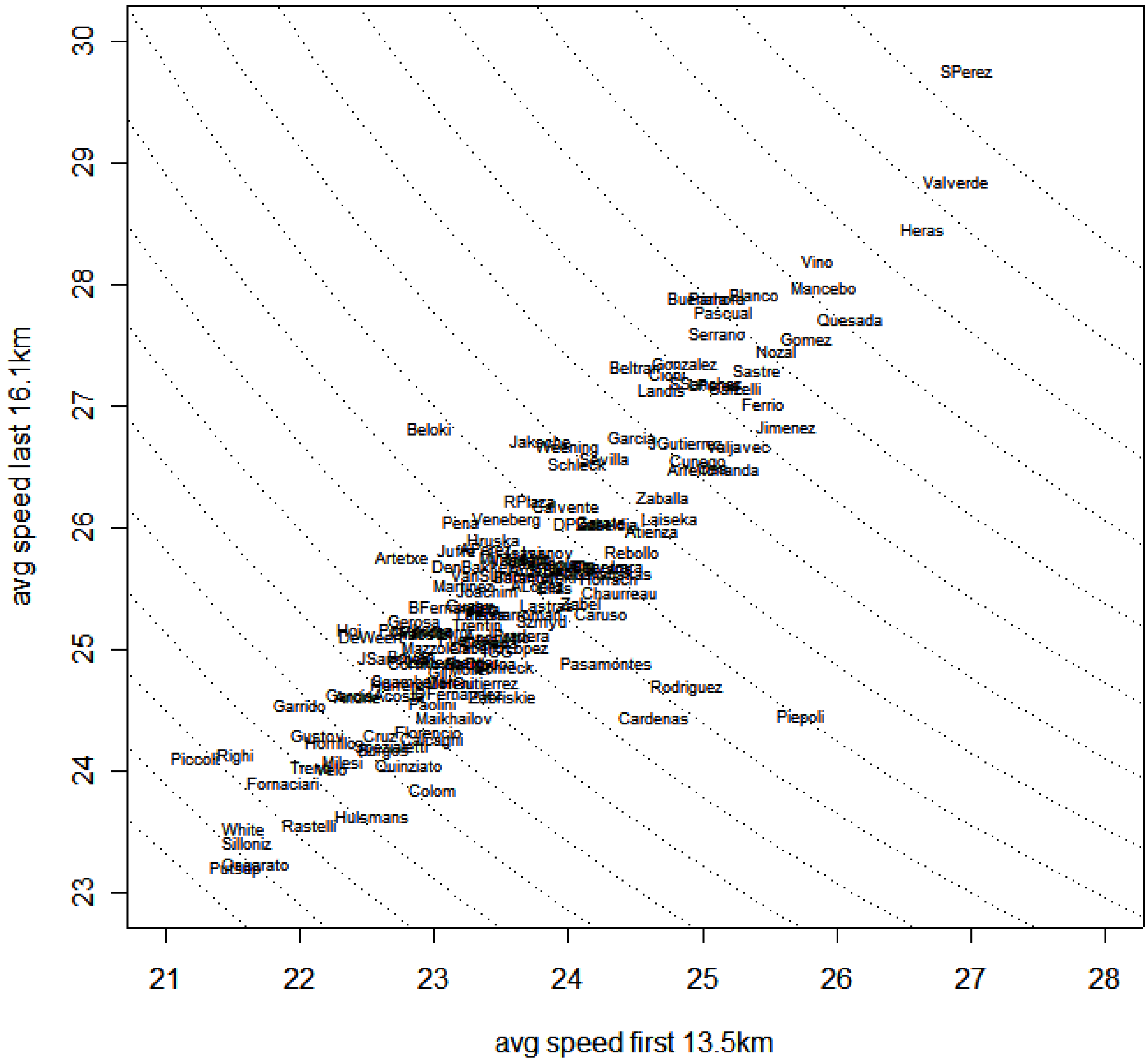
Payroll/median payroll ratio to win/loss ratio
records as of 3 Sep 2005
Projected wins for 2005
2005 Team payroll in millions
StL
ChiSox
Bos
Atl
Oak
Angels
Yanks
Cleve
Marlins
Hou
Phil
Twins
Nats
Mets
SD
Tor
TX
Mil
Cubs
DetBalt
Cin
LA
SF
Ariz
Sea
Tampa
Pitt
Colo
KC
50
100
150
200
60
70
80
90

World Championships, 2003, ITT

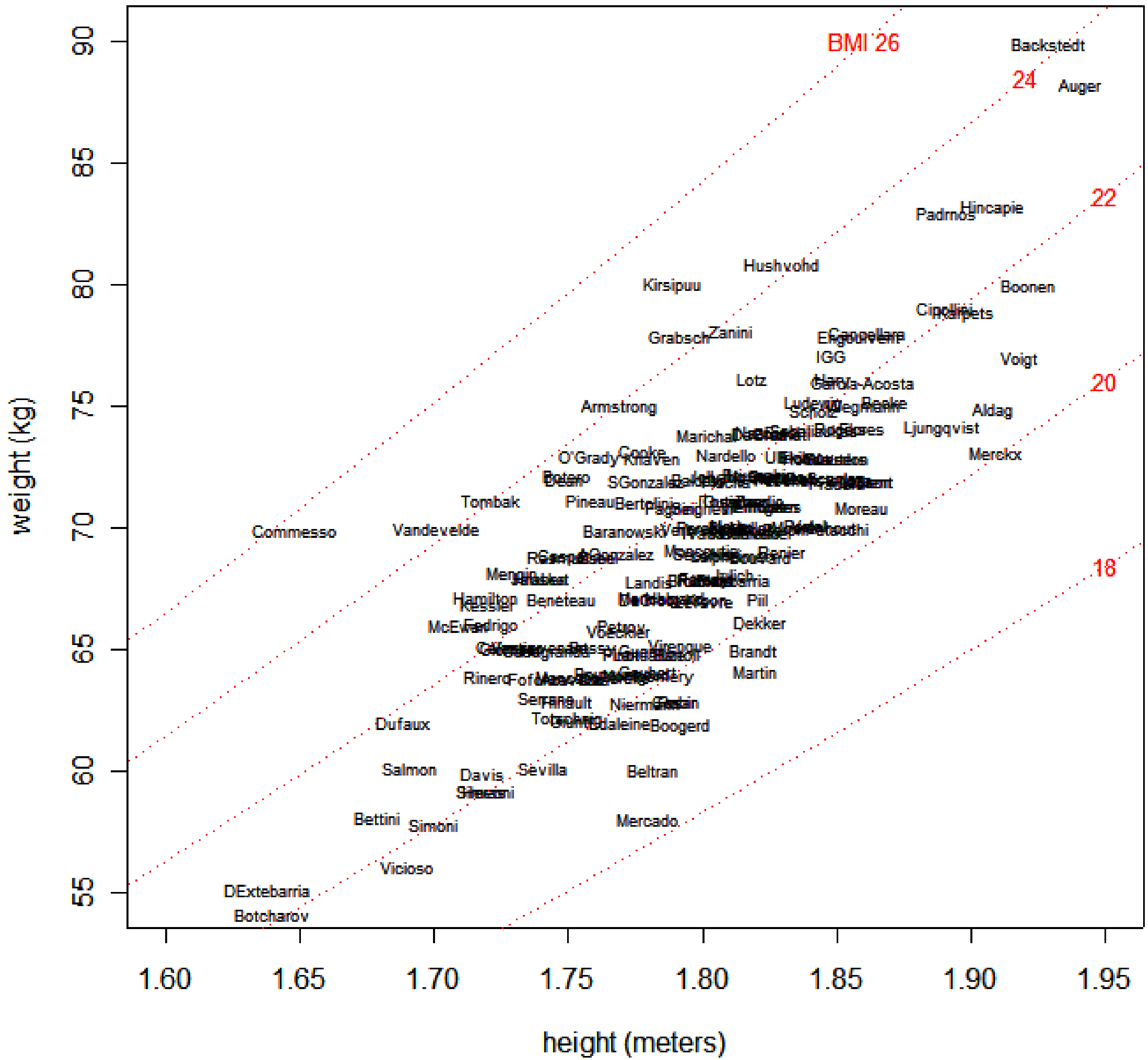


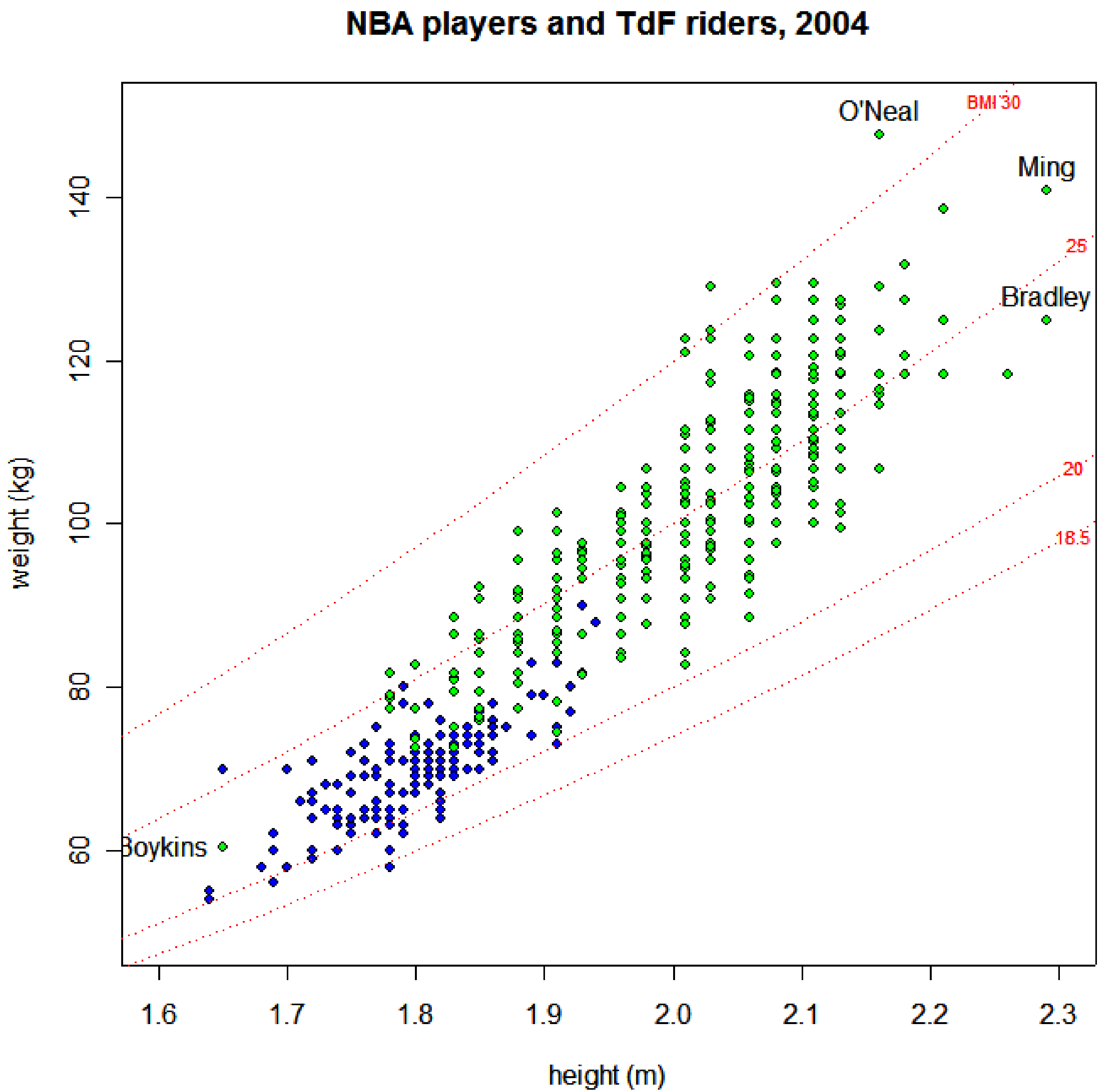
Granada - Sierra Nevada ITT

With one minute time contours



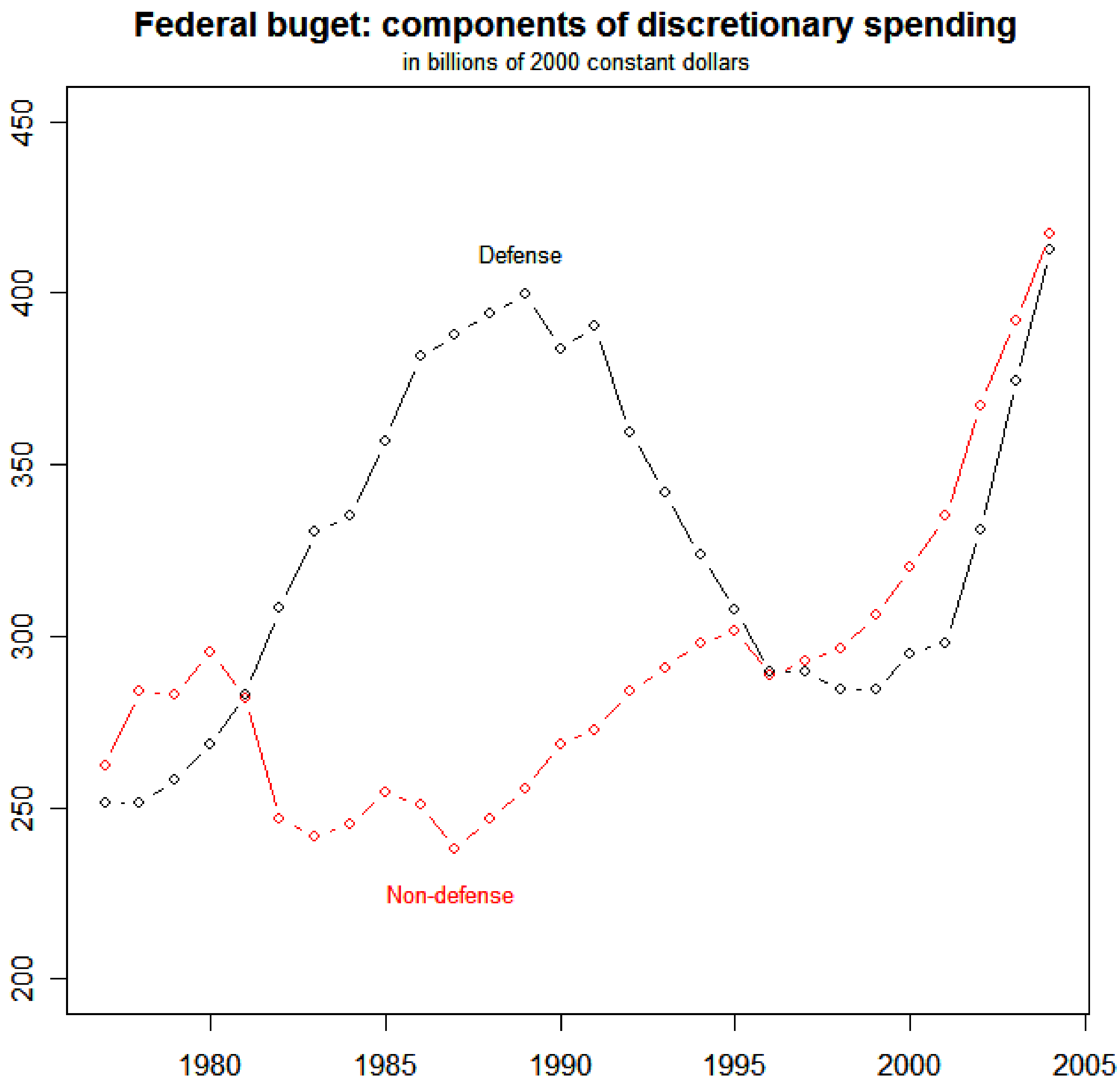
2004 TdF riders



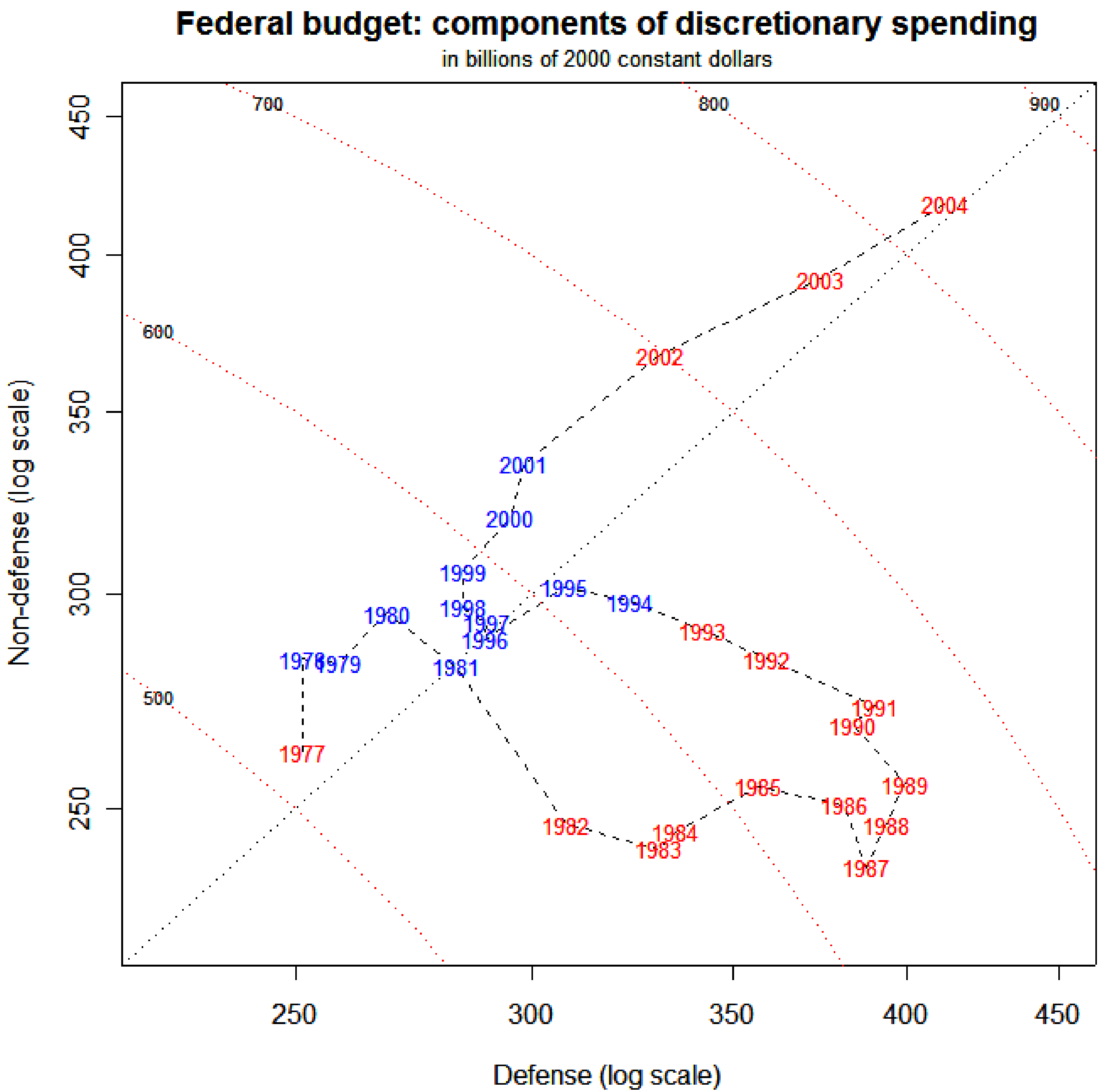

NBA players and TdF riders, 2004
O'Neal
Ming
Bradley
Boykins
BMI 30
25
20
18.5
weight (kg)
60
80
100
120
140
height (m)
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3

decomposing series into phase plots

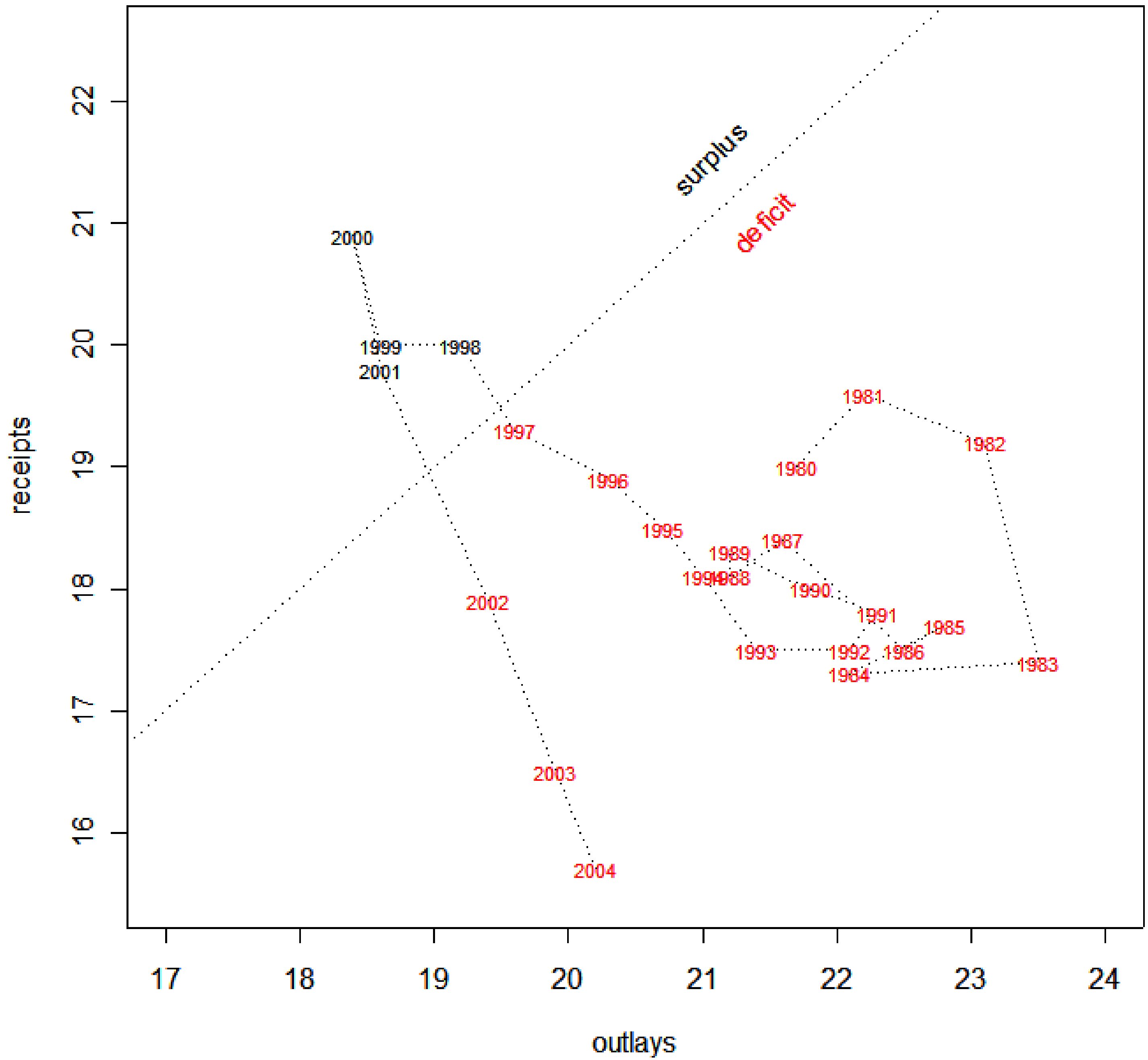
- another version of “show the difference”



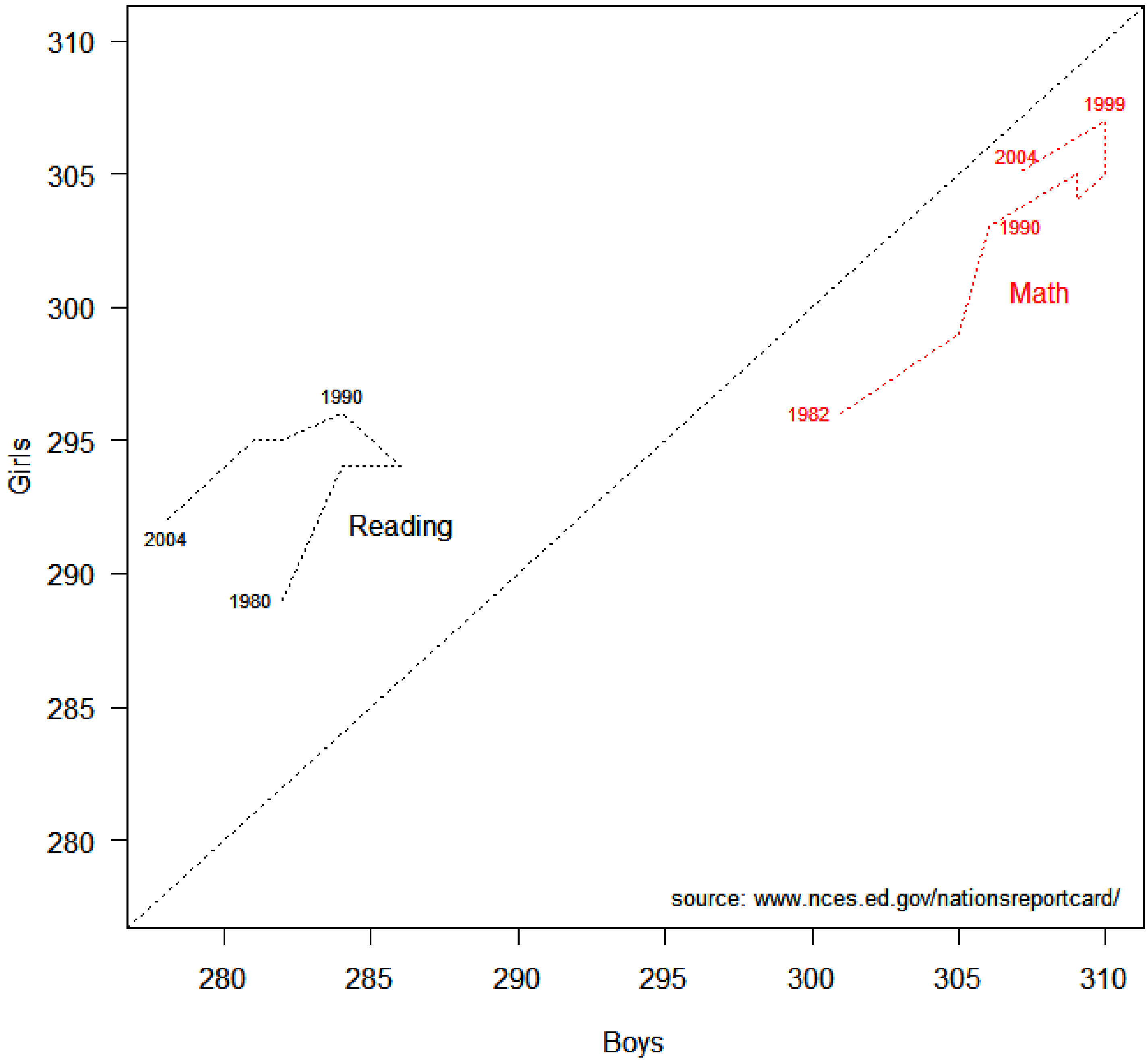
Federal buget: components of discretionary spending
in billions of 2000 constant dollars
450
400
350
300
250
200
1980
1985
1990
1995
2000
2005
Defense
Non-defense


Federal budget: components of discretionary spending
in billions of 2000 constant dollars
Non-defense (log scale)
Defense (log scale)
250
300
350
400
450
500
600
700
800
900
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004

Federal outlays and receipts as percentage of GDP

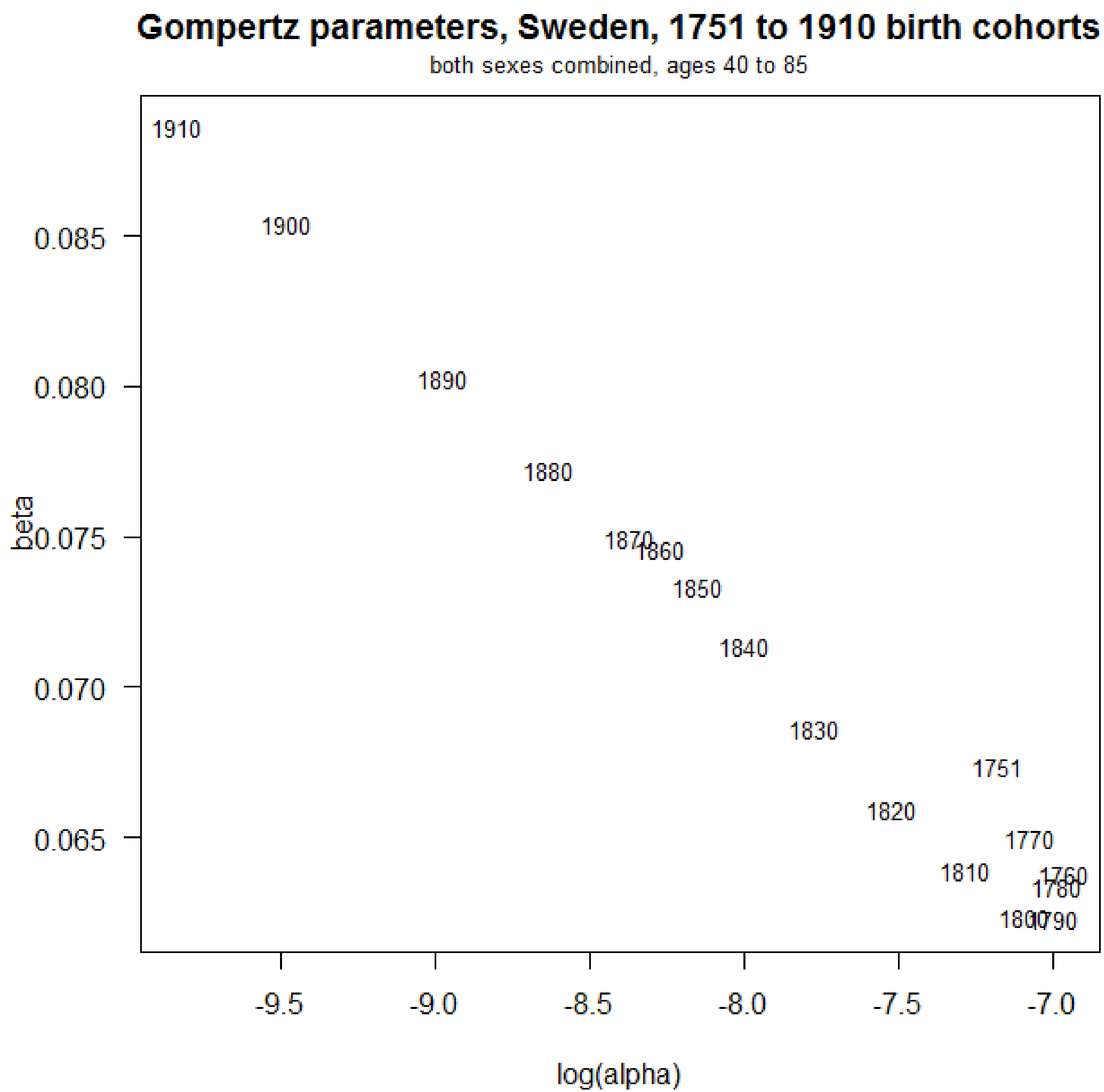


NAEP reading and math scores, 17 year olds



plot summaries for simplification

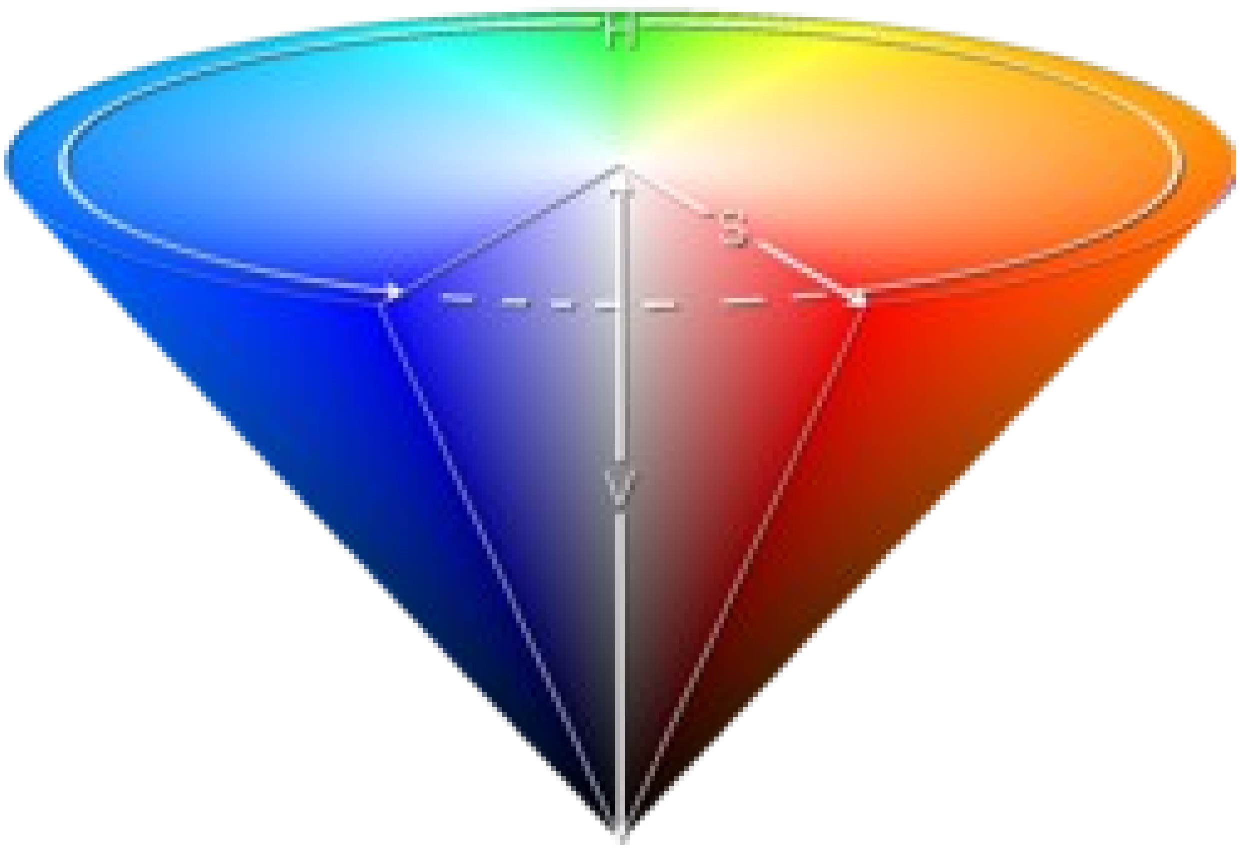
- when all subset have same contrasts, plot subset summaries
  - sometimes can get away with it even if not all subsets have all same contrasts—but then must be doubly careful
  - helps to identify patterns
  - plot and identify extremes, leave middle alone
    - this is the idea underlying “10 plus 10” plots
    - or, split into n groups (n small, like 3), and plot subsamples from each


Gompertz parameters, Sweden, 1751 to 1910 birth cohorts
both sexes combined, ages 40 to 85
1910
1900
1890
1880
1870
1860
1850
1840
1830
1751
1820
1770
1810
1760
1780
1800
1790
0.085
0.080
0.075
0.070
0.065
beta
-9.5
-9.0
-8.5
-8.0
-7.5
-7.0
log(alpha)

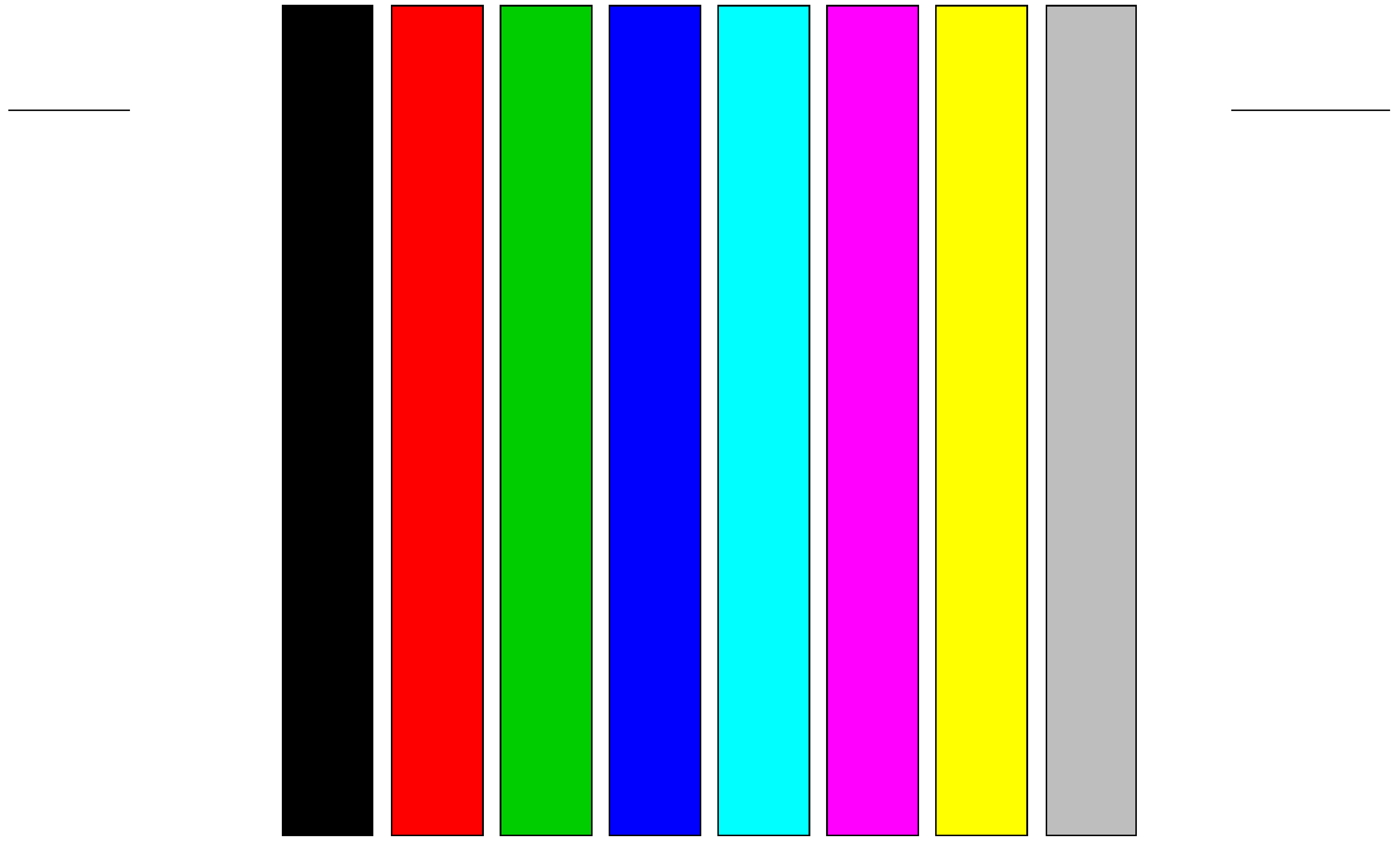
more on color

- HSV
  - h=hue, s=saturation, v=value
  - sometimes called HSL for hue, saturation, luminance
- equal impact colors
  - CIELUV and Munsell are systems of color perception
    - medium saturation, kind of pastel-like

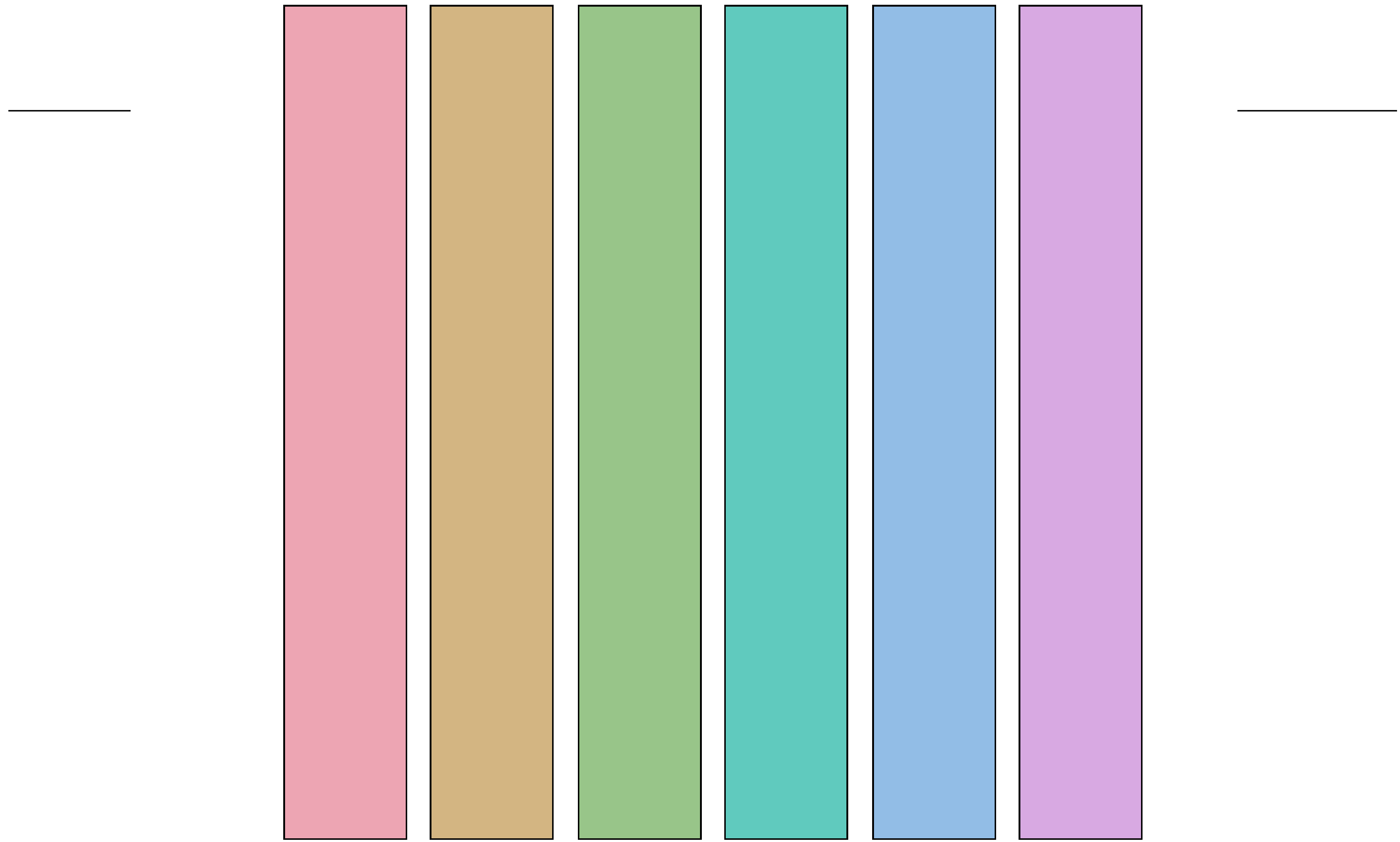
HSV cone



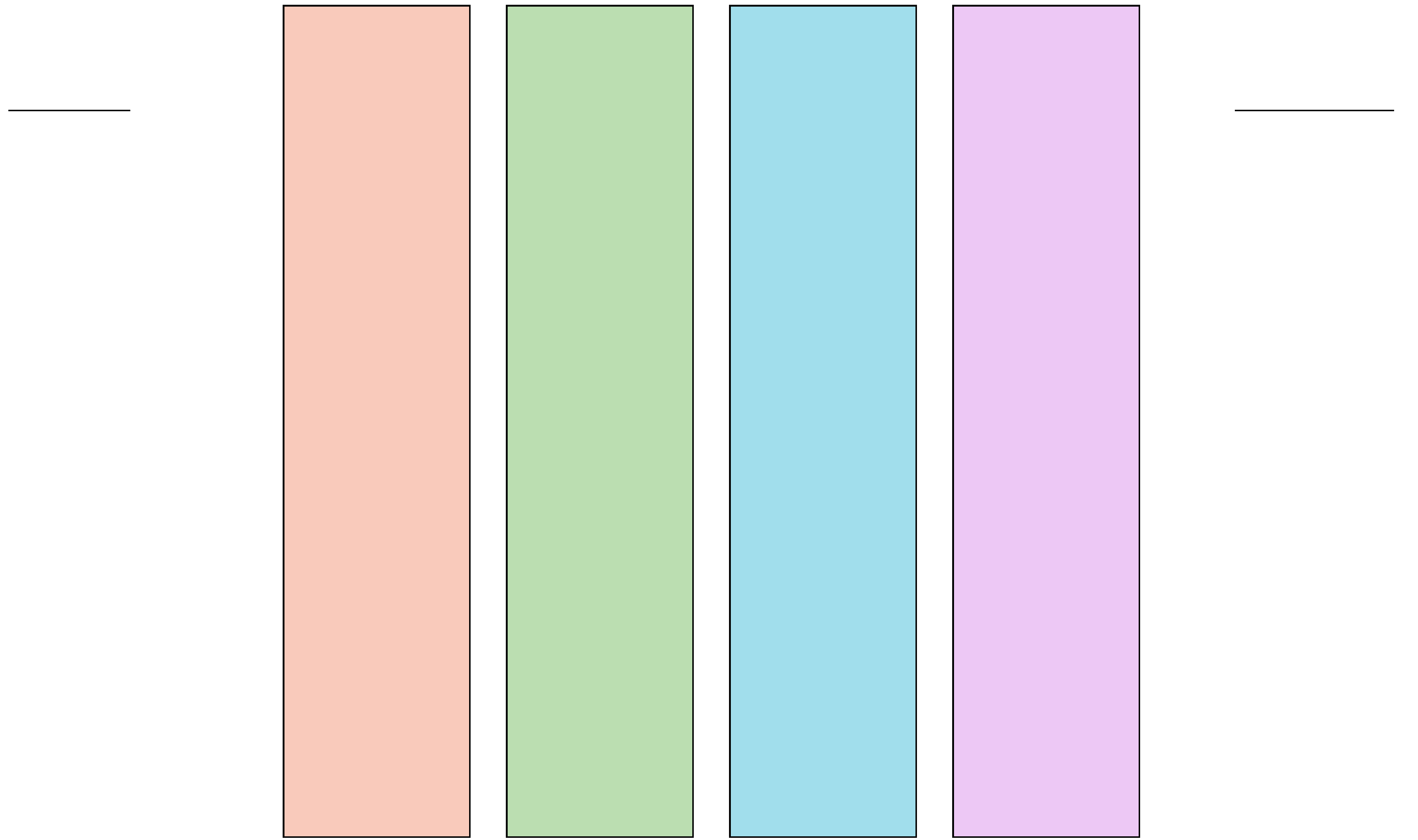
R's default colors



Example of equal impact colors



tetrad with maximal color differences



basic techniques

- show the difference
- identify outliers (or, label directly)
- group and order
- plot extremes
- multiple comparisons

slightly more advanced techniques

- smoothing
- straightening
- phase plots
- contours
- banking
- coloring

stuff I wanted to hide until the end

- friends don't let friends graph with Excel

  but let's be realistic: sometimes you have no choice

  dates in Excel are particularly a problem

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature > news > article

NEWS | 13 August 2021 | Correction 25 August 2021

Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

By Dyani Lewis

how a demographer changed
bicycle racing and design

finding a story to tell:
graphical research methods

Robert.Chung@berkeley.edu

October 2024

analytical graphics

- not presentation graphics, not exactly statistical graphics, not quite exploratory graphics
  - presentation graphics tell your story; analytical graphics help you figure out what your story is
  - analytical graphics are “work product”
- not always refined enough for publication or presentation
  - Tufte? A lot of his techniques focus on effectively communicating quantitative findings. We'll focus on steps before that: uncovering interesting stories and questions in data
  - Very valuable for grad students and early career researchers

You may be familiar with the work of Edward Tufte. Excellent stuff, but his focus is on how to display quantitative data in a way that tells a clear and compelling story. What we're focusing on is at an earlier step: how to figure out what your story is. Tufte focuses on presentation graphics that clearly render the information that they're supposed to convey. What we're doing is sort of mid-way between exploratory graphics and presentation graphics. We're still trying to figure out what that information might be, or if there is any interesting information there at all. These are *graphical research methods* and, like all research methods, they reveal the most if you approach it in a (semi-) organized way. In these lectures we'll present some principles, some rules of thumb, some shortcuts, and some tricks that can help you identify questions that are deep enough to hold your attention long enough to finish a dissertation.

analytical graphics are for analysis

- often analytical graphics are used not to prove hypotheses but to help generate them
  - we don't always answer questions; we use graphical techniques to help us ask new ones
- often, the audience is YOU
  - visualize differences and contrasts
    - across time
    - across places
    - across treatments or policies
    - across conditions
  - there are tips, tricks, and techniques that help you in visualization

Many of the (very beautiful) graphics in Tufte's books are “one-offs” that have been hand-tuned in Adobe Illustrator. We don't have that luxury – we're looking for techniques we can use over and over in a way that makes production relatively easy. We're optimizing for analysis, not for presentation.

what if you already have a question?

- no problem. sometimes analytical graphics can help you focus on where to look, or to refine your question
- doesn’t replace theory, or your research question. You can do both: that’s allowed

Many of the (very beautiful) graphics in Tufte's books are “one-offs” that have been hand-tuned in Adobe Illustrator. We don't have the luxury – we're looking for techniques we can use over and over in a way that makes production relatively easy. We're optimizing for analysis, not for presentation.

basic approach

- maximize insight
- uncover underlying structure
- extract important variables
- detect outliers and anomalies
- develop (very) simple models
- not much testing; that's for later

Testing hypotheses is important, but this isn’t (so much) about that. We’re looking for stories to tell, and just like writing, you try to separate the writing from the editing.

what we'll do in these lectures

- some examples
- some principles
- some basic tricks
- some slightly more advanced tricks
- you'll have a chance to try out some of these tricks before next week

In past years, we’ve used a prepared data set for this, but that requires more time commitment than I had this year.

Warning: a lot of these examples are pretty old, because I’ve been working on versions of this for about 20 years. So if the examples seem anachronistic, that’s cuz they are. The principles and tricks are (hopefully) still useful.

I sometimes wonder how AI will affect these lectures.

the three things we're looking for

- look for
  - pattern
  - unexpected pattern
  - deviations from pattern
- these generate questions
- questions and how you address them are often the basis for papers or chapters or dissertations or careers
- "The data speak for themselves, but their voices are soft and sly" so we're looking for ways to amplify their voices

demography

- demography is the study of populations, their characteristics, relationships among characteristics, and how they change
- you probably already know how to examine characteristics; we'll look at ways to highlight relationships among the characteristics and how they change
- in particular, we'll often look for models to help us understand the relationships among characteristics

We have, in demography, a pretty sweet situation: we have relatively strong models that give us clues about how variables might (ought?) to be related. So sometimes we'll want to look at potential relationships. That's an advantage that not all fields have.

the purpose of models

- “The purpose of models is not to fit the data but to sharpen the questions” – Sam Karlin
- We’ll use analytical graphics to help us sharpen questions

simple tools

- simple tools used intelligently (well, we can always dream) rather than complex tools used stupidly
  - rules of thumb, not hard rules and regulations
- a handful of plots and a handful of tricks
  - lots of specialty type graphs, but we try to avoid too many of them until we know our story
- xy plots are a hugely useful invention
  - with one or two exceptions we'll focus mostly on ways to enhance xy plots
- decoding the language of graphs can be complicated, so we build on familiar beginnings

Simple tools don't always mean that the doing is simple, or that the results are simple. Here we mean “simple” in the sense that these are basic foundations on which we'll build. We'll learn about strengths and weaknesses of different tools.

Excel is a simple (and widely available) graphics tool but, as we will see, in certain uses it's too simple and in others it's not complex enough. However, because it is ubiquitous, later on if you insist we can go over some things you can do to neutralize some of the bad things that Excel does.

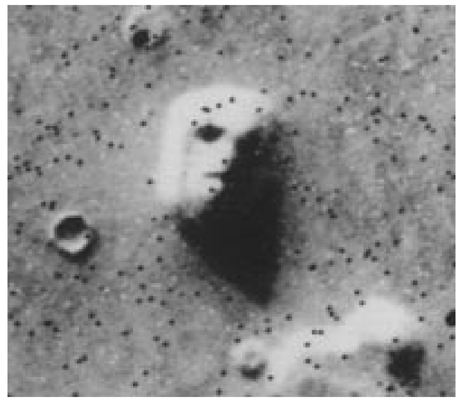
how graphing helps

- we can only make sense of a handful of numbers at a single time
  - pages of dense tables are good for detail, evidence, and re-analysis but poor for understanding
- eye-brain is good at seeing patterns in large numbers of values
  - though it can be fooled—we'll present some problems that can mislead the eye
- therefore
  - use graphs when pattern is important
  - use tables when exact details are important
  - graphs and tables are complements, not replacements. (You can do both: that's allowed)

People who focus on presentation (like, for example, Tufte) often appear to disdain tables in favor of graphics. However, because we tend to do both analysis and presentation, I think of them as complements. Use each (judiciously) when appropriate.

apophenia

- apophenia is “the experience of seeing patterns of connections in random or meaningless data”
- we'll occasionally accept a little “type I error” when we're looking for interesting questions – as long as we back it up later with real confirmatory analysis



This is the famous (?) “Martian face” as photographed by one of the Viking spacecraft missions to Mars.

Anscombe's data

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

built into R.

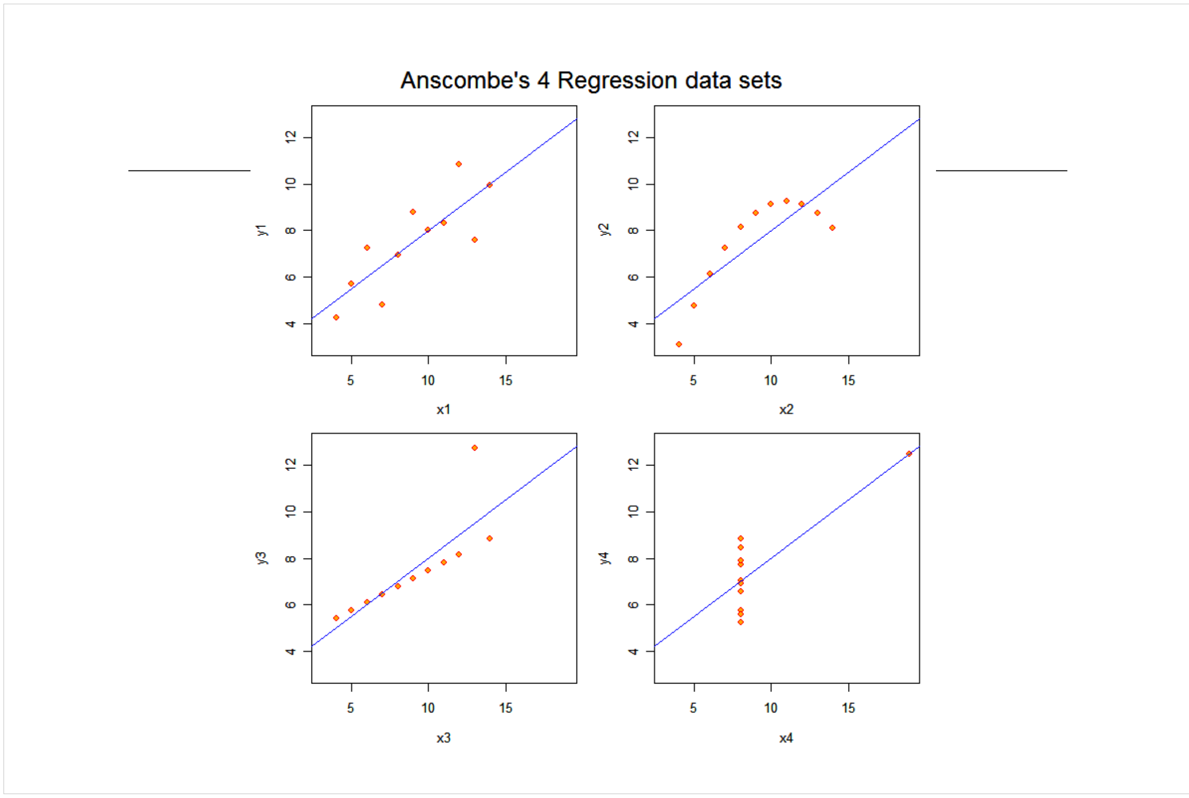
help(anscombe)

This is a pretty famous set of manufactured data. We'll see why in a moment.

Anscombe's data

- same means, sd's, correlation, regression slope, fit
  mean(x1)=mean(x2)=mean(x3)=mean(x4) = 9
  mean(y1)=mean(y2)=mean(y3)=mean(y4) = 7.5
  sd(x1)=sd(x2)=sd(x3)=sd(x4) = 3.32
  sd(y1)=sd(y2)=sd(y3)=sd(y4) = 2.03
  r(x1,y1)=r(x2,y2)=r(x3,y3)=r(x4,y4) = 0.816
  y* = 3 + 0.5 x* with r^2 = 0.667
- so, conventional linear models make them look alike
- what will you see if you graph the data?

What's worse, a depressingly large proportion of analysts will stop right there.



example(anscombe)

the NJ Pick-It lottery

- each bettor selected a 3-digit number between 0 and 999
- each ticket cost 50 cents
- all bettors who held the winning number split the prize money. The size of the prize depended on selecting the winning number and on the number of players who chose that number
- what would you want to know?

built into S-plus; in fact, I stole this example from “The Blue Book”

dat= read.table(“http://anonymous.coward.free.fr/misc/lottery.txt”,header=T)

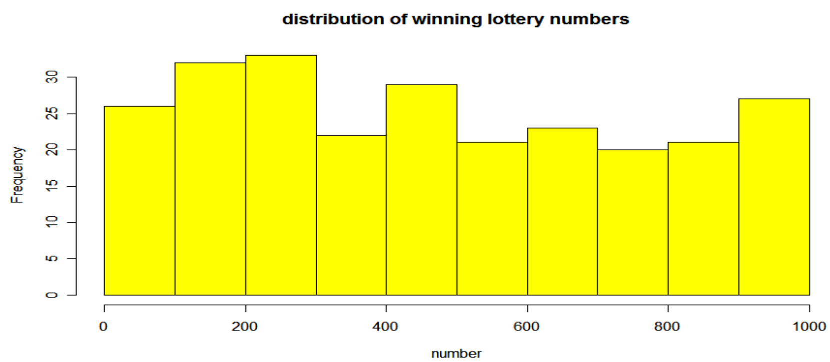
winning numbers and prize amounts

(810, $190.0)
(156, $120.5)
(140, $285.5)
(542, $184.0)

and so on for 254 consecutive days

strategy 1: choose a winning number

- since we have data on the winning numbers, see if there's a pattern we can exploit to pick the winners
- examine the distribution of winning numbers using histograms or stem-and-leafs




```
dat = read.table("lottery.txt",header=T)
attach(dat)
hist(number,col="yellow")
```

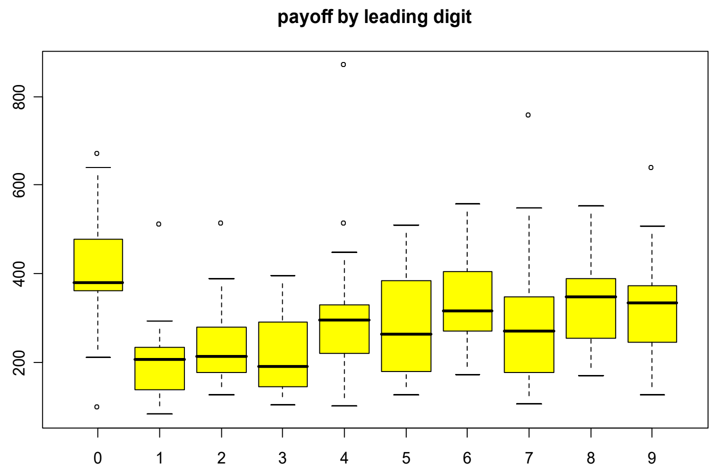
Are there more winning numbers between 100 and 299 than we'd expect by chance?

There are n=254 total "picks". We'd expect n*p in each of the 10 bars, or about 25.4, with sd=sqrt(n*p*q) = 4.8.

So we could draw horizontal lines at 25.4 +/- 2*4.8 to mark 2SD CI's. None of bars exceed CI.

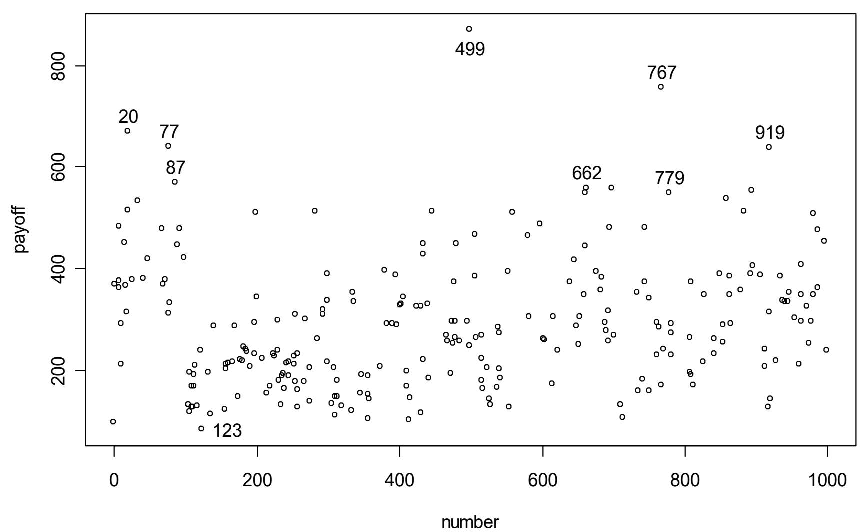
So since we can't pick a winning number, maybe we can see a pattern in how much each winning number won.

strategy 2: choose a winning number that few others pick



boxplot(payoff~number %/% 100, col=“yellow”)

Aha. It seems as if when asked to pick a number, not many people picked numbers below 100.



plot(number,payoff)

identify(number,payoff,number)

What do you see? Identifying the points helps you see that numbers with high returns either have a zero for the first digit, or double digits. Note that the lowest payoff for a winning number was “123”.

Identifying high and low payoffs tells you something interesting that you wouldn’t have known from more standard analysis. We’ll come back to this later when we talk about labeling.

why little circles?

- easier to distinguish overlapping points
- especially with jittering

little circles show up better when you have points close to another. This is part of the investigation into graphical perception that was done at Bell Labs.

five rules

- rule 1: graph lots
- rule 2: use what the eye is good at (and avoid what the eye is bad at)
- rule 3: find the right contrast and show it
- rule 4: make it easy to spot pattern, and deviations from pattern
- rule 5: plot models, not just the data

rule 1: graph lots

- only one out of 50 graphs will “work” so to get a handful of workable graphs, graph lots
- good graphing principles help raise your yield of workable graphs
- better if you can generate lots of simple graphs quickly even if they’re not perfect
- for you, not for presentation (at this stage) so don’t obsess on look

  (though I’m showing you the survivors of hundreds of graphs, so they’re cleaner and not quite representative of the messy graphs I usually produce: my working graphs usually don't have titles, clear axis labels, etc.)

Here's a tip that often helps me raise the yield: sketch what you think your graph will look like on a piece of paper. This sounds archaic but, for me at least, it seems to help. This may or may not work for you—you have to pay attention to what works for you.

rule 2: use what the eye is good at (and avoid what it's bad at)

- we need to know something about how the eye-brain perceives graphics
  - what it's good and bad at, and an ordering or hierarchy
  - “optical illusions” and traps to avoid
  - techniques to exploit strengths and minimize weaknesses

graphical perception

- quantitative pattern recognition by
    - detection: recognition of geometry
    - assembly: grouping of detected elements
    - estimation: assessment of relative magnitudes
- the human eye-brain can be fooled
    - optical illusions
- need to help it out
    - grouping, ordering, highlighting help to identify patterns

a hierarchy of graphical perception

- position along common scale
- position along identical non-aligned scales
- length
- angle, slope
- area
- volume
- shading, color (good discrimation but poor ordering)

Cleveland and McGill, 1984

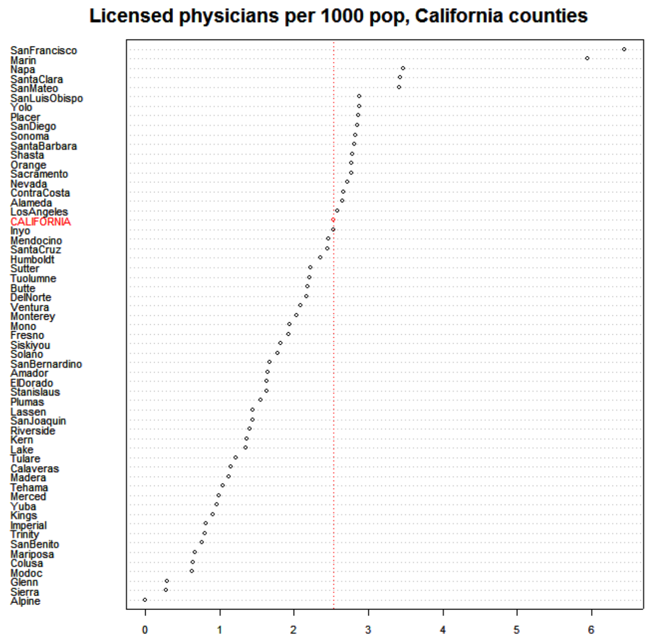
dotplots are plots where quantitative values are displayed as positions along a common scale.

Pie charts, which are ubiquitous and rely on decoding angles, are bad. (Stacked pie charts, aka “spie” charts, can be useful in certain circumscribed situations, but don't go hog wild).

Notice that the familiar age pyramid with males on one side and females on the other uses non-aligned scales, and lengths and areas when it could as easily use position on the same scale.

Color-coded maps are often good for detecting pattern *when the pattern is geographical*, but poor when the pattern is just about anything else. Because colors are poorly ordered by the human perception, it's often hard to decode quantitative variables by color – no matter how nice the final product looks. Use color to highlight, or to encode categorical variables, but avoid using it for quantities unless you know what you're doing. For example, if a pattern changes over time, sometimes maps just confuse things.

dotplots



dotplots are preferred to barcharts because bars have width and, thus, area and area is less reliably interpreted than position.

```
dat = read.csv("docs-percap.csv")
with(dat,dotchart(ratio,label=County))
```

area

- pie charts require estimation of area
- human perception of relative areas is conservative, i.e., shrinkage toward 1.0
- shape affects estimation of area
  - concave shapes appear larger than convex
  - maps are good for context and clustering, not so good for comparisons of quantitative amounts
- color intensity affects estimation of area.
  - highly saturated colors appear larger

shape affects perception of area: if you look at a map of the southeastern states, it appears that Florida is larger than Georgia – Florida is an odd shape, GA is more compact.

This is another potential problem with using maps to encode quantitative information: Denmark is about one-eighth the size of Norway in land area, but it's population is slightly larger, so encoding a mortality rate on top of a map adds one more layer of information for the eye-brain to decode (just so it can reject it as irrelevant).

One of the problems with “spie” aka “stacked pie” charts: are you comparing areas or lengths of radii?

color

- human eye good at discrimination, poor at ordering
  - use for categories, not for quantitative coding
  - hues are not ordered
  - use for highlighting, patterning, especially in combination with small multiples
- more on color, later

rule 3: find the right contrast and show it

- don't rely on the eye to do differencing

  if you're interested in the difference between two lines, don't show the lines and rely on the eye to calculate the difference; calculate the difference itself and show it

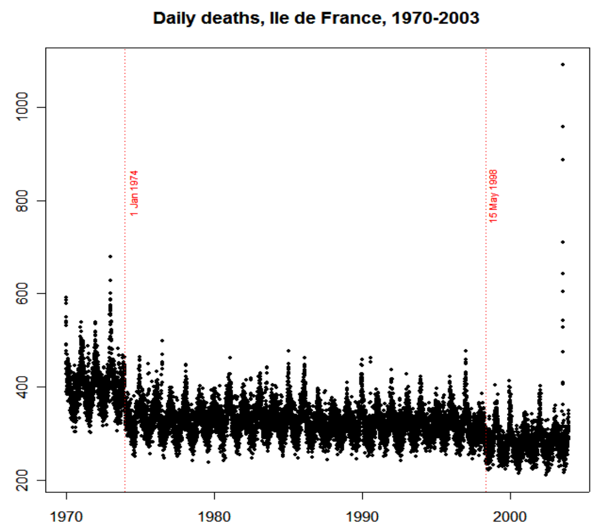
- Tukey mean-difference (aka Bland-Altman) plots

  levels vs. differences

- fits and residuals

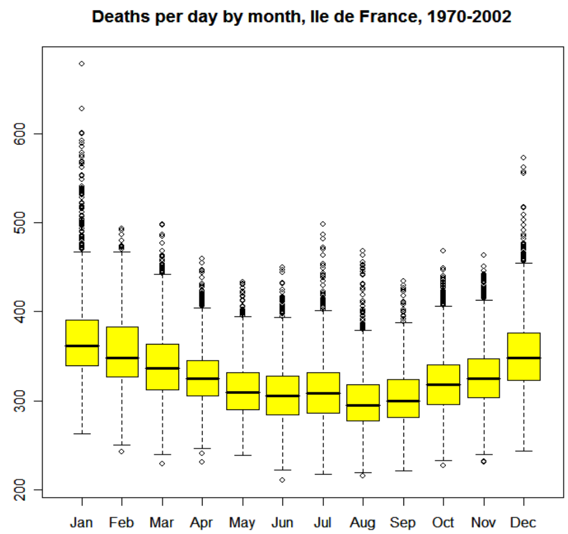
- different contrasts can give different insights

differences by time



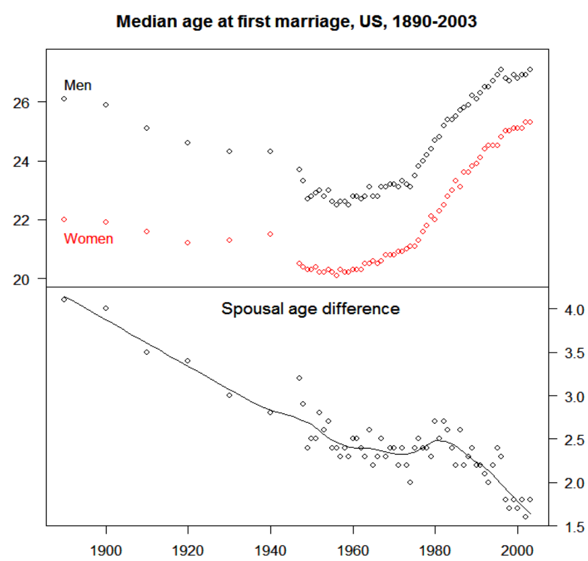
Time series plots are, next to pie charts, perhaps the most common of all charts. They do show how something changed over time; however, while they tell you when something happened, they don't tell you why it happened. You have to fill in the context. Most of what we're trying to do here is figure out ways to generate ideas and hyptheses, so showing when something happened isn't always very helpful.

differences by category



This gets a little more interesting. Simply by grouping by category, we see a pattern emerging that may be clearer than the previous slide.

differences between lines

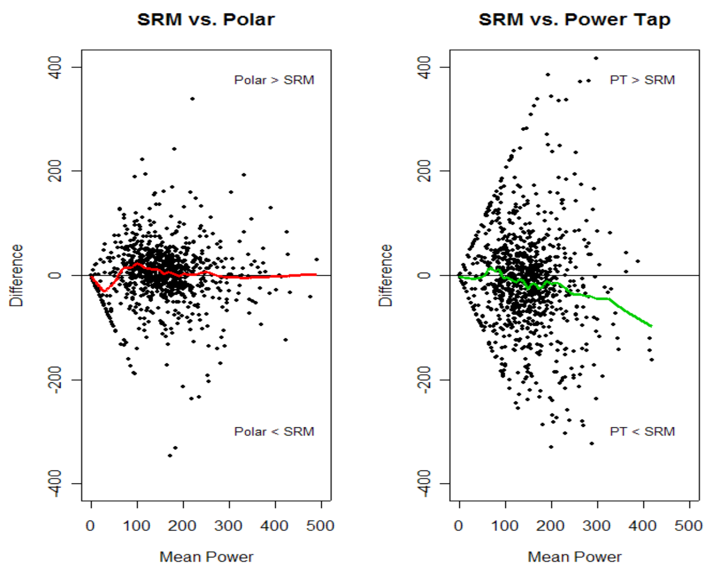


Just by looking at top panel, eye might have thought spousal age diff was constant after mid-1980's. Eye would never have picked up bump in late 1980's. Humans have trouble judging differences between two curves, *especially* when those two curves are steep.

If your story is about the difference, calculate and show the difference.

```
dat = read.csv("marriage-age.csv",comment="#")
head(dat)
attach(dat)
yl = range(Men,Women)
plot(Year,Men,ylim=yl)
points(Year,Women,col="red")

plot(Year,Men-Women)
lines(lowess(Year,Men-Women,f=0.5))
```

differences and means



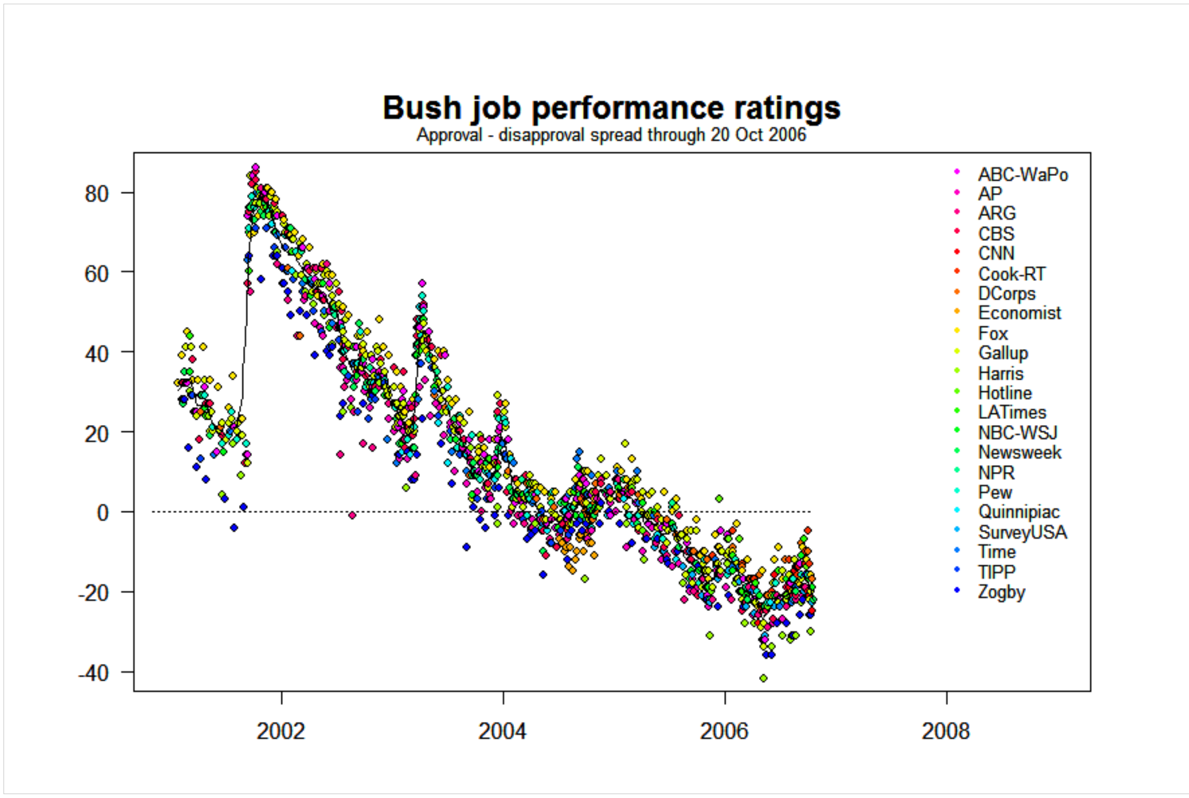
Tukey mean-difference plot. Shows whether difference grows with level.
x-axis is (SRM+PT)/2, y-axis is PT-SRM

plot((x+y)/2, (x-y))

differences from fits

- data = fit + residual
- the classic residual plot

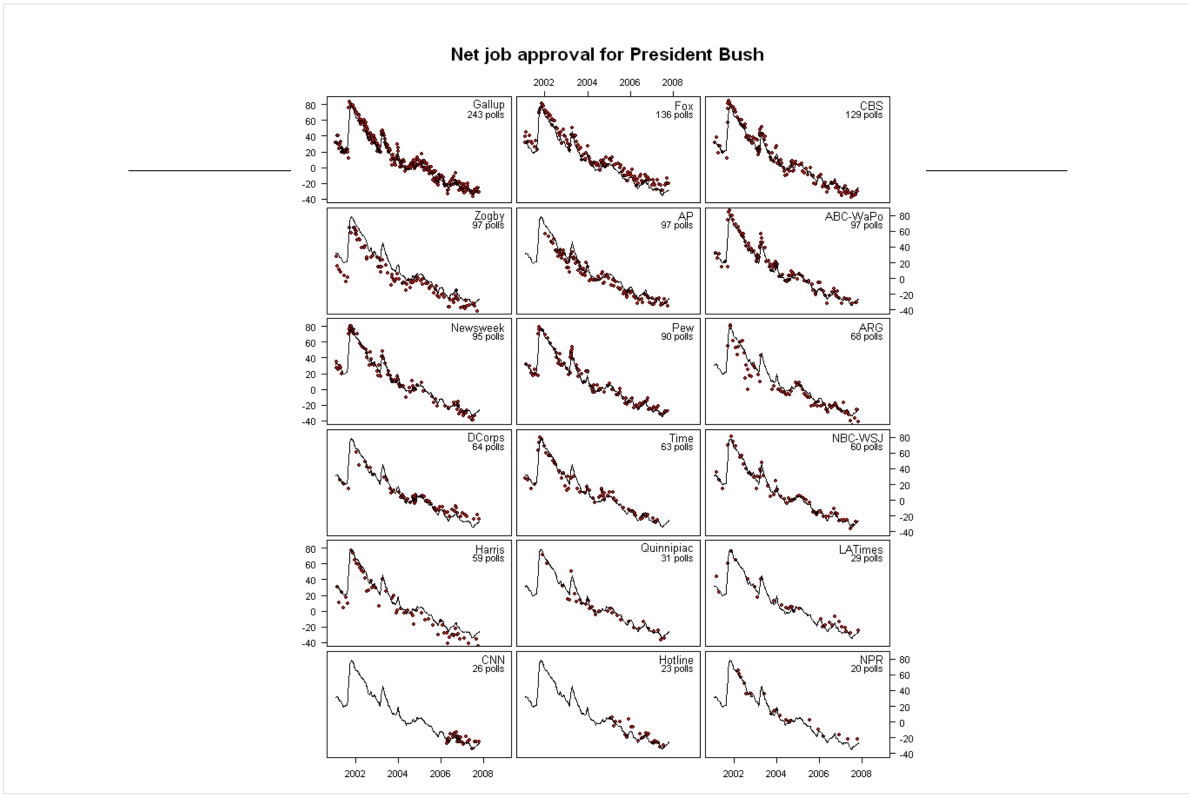
  “fit” can be broadly defined: data = smooth + residual



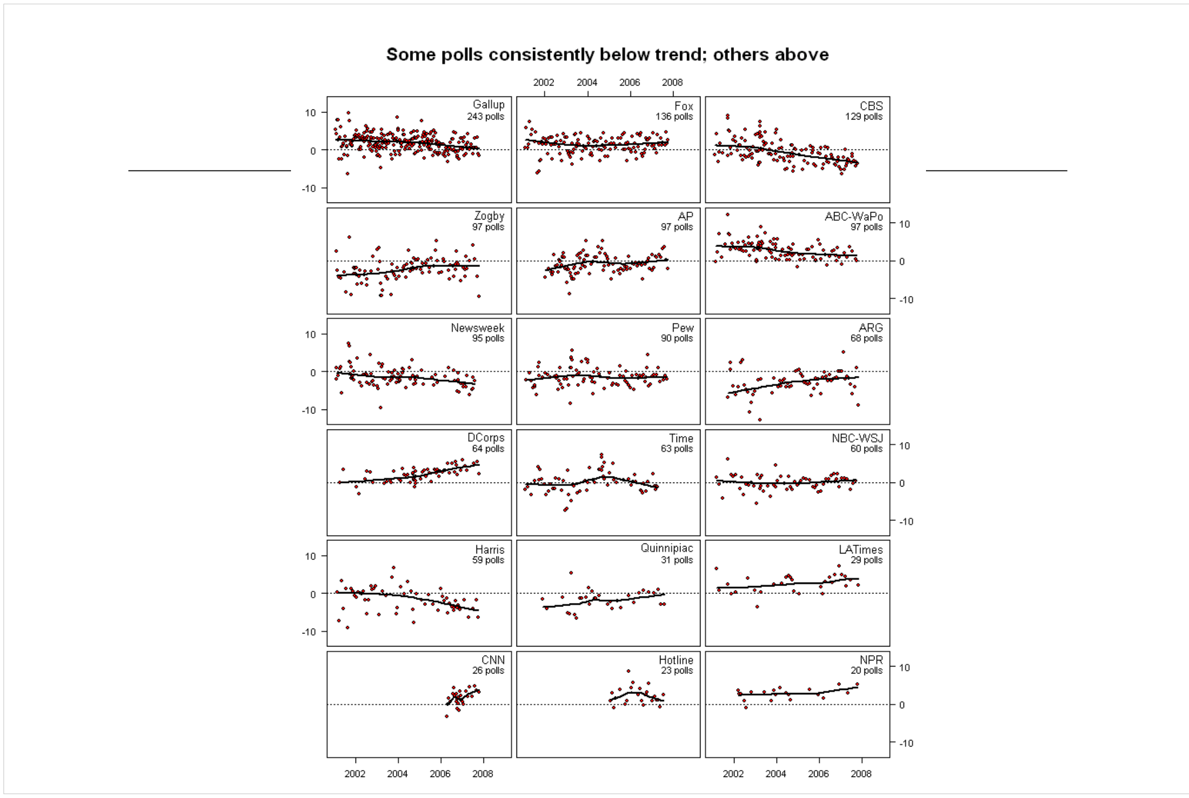
lowess

Too many colors
Legend hard to use

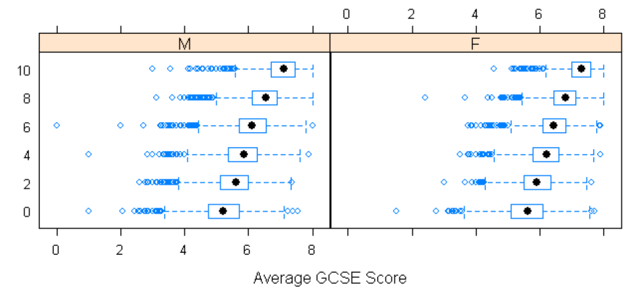
dat = read.csv("polldata.csv",comment="#")
head(dat)
dat$cdate = as.Date(dat$cdate,"%m/%d/%Y")
with(dat,plot(cdate,approval-disapproval))

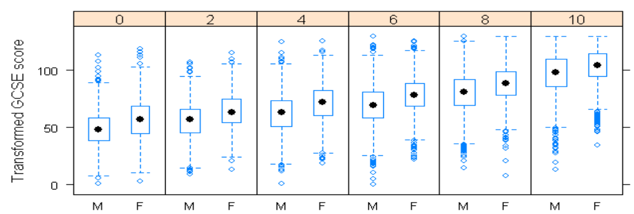


Net job approval for President Bush
Gallup
243 polls
Fox
136 polls
CBS
129 polls
Zogby
97 polls
AP
97 polls
ABC-WaPo
97 polls
Newsweek
95 polls
Pew
90 polls
ARG
68 polls
DCorps
64 polls
Time
63 polls
NBC-WSJ
60 polls
Harris
59 polls
Quinnipiac
31 polls
LATimes
29 polls
CNN
26 polls
Hotline
23 polls
NPR
20 polls
2002
2004
2006
2008


Some polls consistently below trend; others above
Gallup
243 polls
Fox
136 polls
CBS
129 polls
Zogby
97 polls
AP
97 polls
ABC-WaPo
97 polls
Newsweek
95 polls
Pew
90 polls
ARG
68 polls
DCorps
64 polls
Time
63 polls
NBC-WSJ
60 polls
Harris
59 polls
Quinnipiac
31 polls
LATimes
29 polls
CNN
26 polls
Hotline
23 polls
NPR
20 polls
2002
2004
2006
2008
10
0
-10

same data, different contrast





These are Figs. 3.11 and 3.12 from Sarkar's Lattice book. The bottom figure contrasts male and female scores on a test and makes it easier to see that transformed male scores seem to improve more than female scores.

To be clear, it's **still** hard to see the difference. If you were doing a presentation you might want to show this in a different way. (Q: how might you do it?) The point isn't that this contrast is the best way to see this effect – it's that if you hadn't done this contrast you might not have seen it at all.

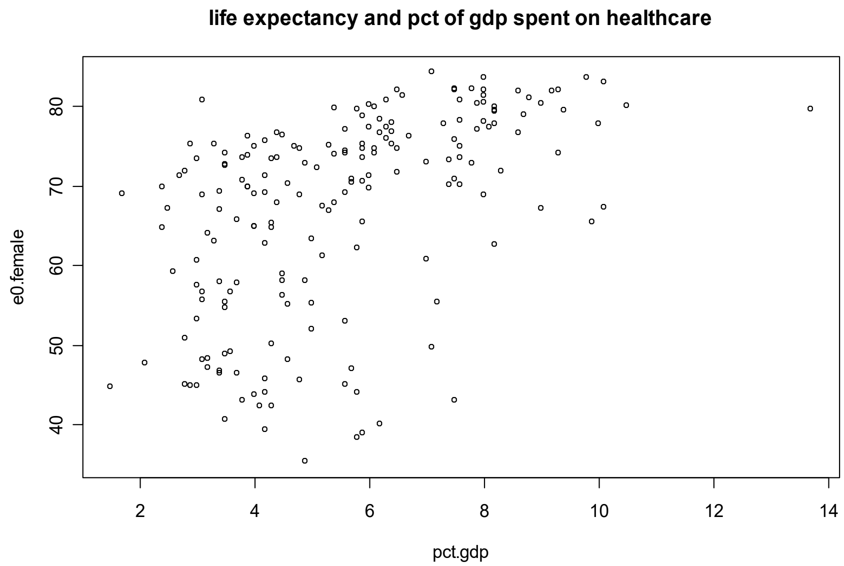
```
data(Chem97, package = "mlmRev")
## Figure 3.11
bwplot(factor(score) ~ gcsescore | gender, data = Chem97, xlab = "Average GCSE Score")
## Figure 3.12
bwplot(gcsescore^2.34 ~ gender | factor(score), Chem97, varwidth = TRUE, layout = c(6, 1), ylab = "Transformed GCSE score")
```

rule 4: make it easy to spot pattern

- add information depth, not (unnecessary) complexity

  sometimes two plots are better than one complex plot (and sometimes it isn't)

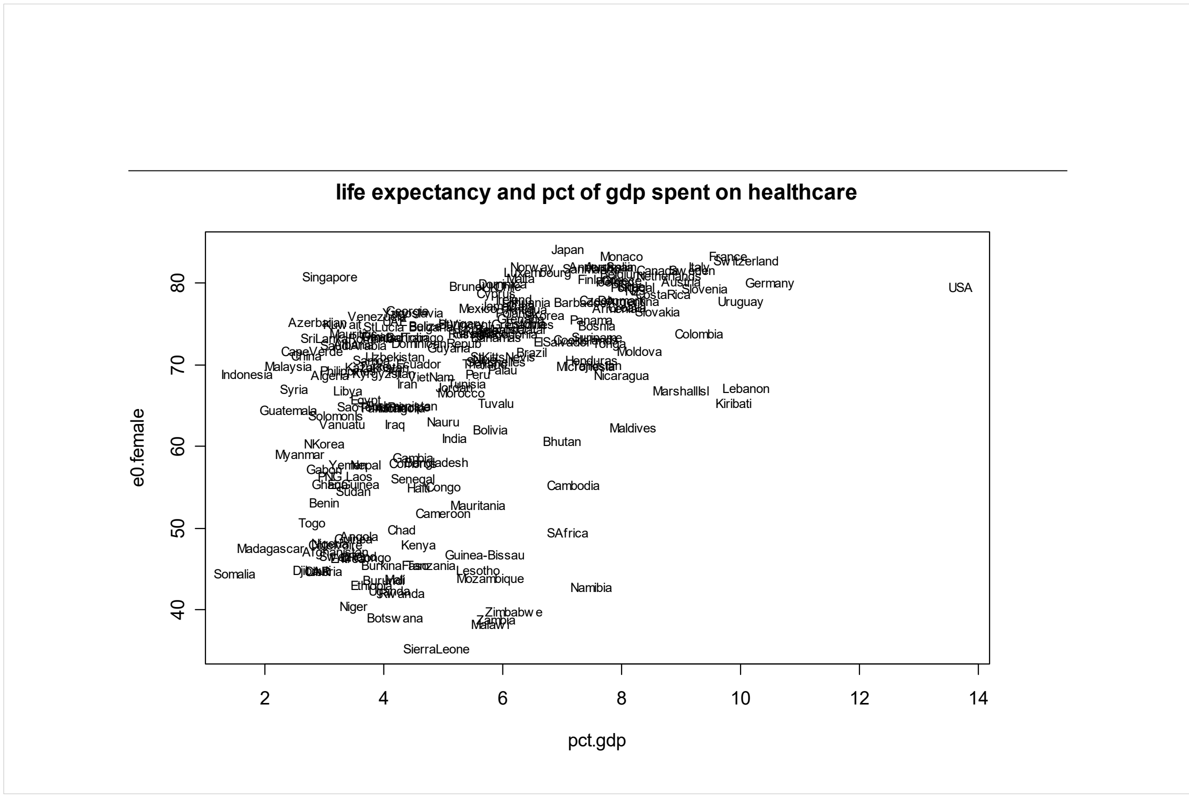
direct labeling



dat = read.csv(“http://anonymous.coward.free.fr/misc/who.csv”)

with(dat,plot(pct.gdp,e0.female))

This tells you something about the pattern (which is good) but it still kind of sits there on the screen. Can we do better?



plot(pct.gdp,e0.female,type="n")

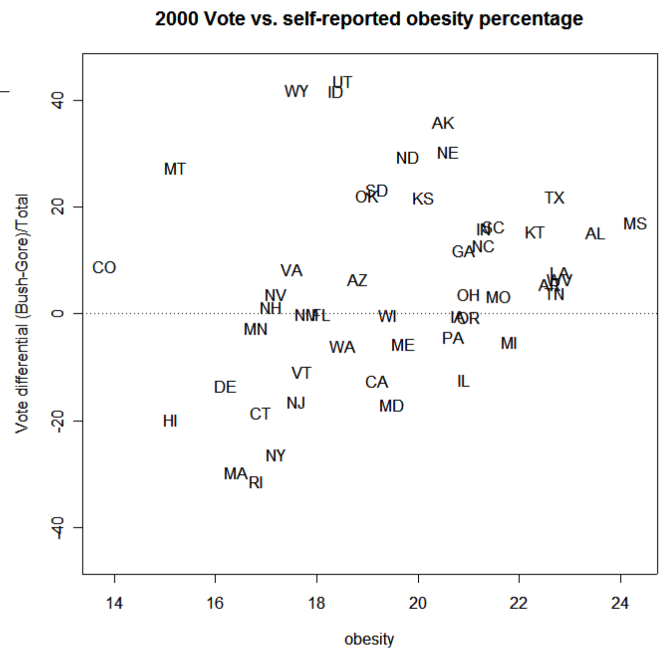
text(pct.gdp,e0.female,country,cex=.7)

Labeling helps a lot, even though you sacrifice exact placement of the point (since the names are of various lengths—in this example, the labels are centered on the data value).

You see the US, off the to right. You see Singapore way off to the left. You can spot sub-Saharan Africa. The labels have added context that helps you to think about new questions.

Note, your eye probably went to the outliers: we all tend to pick up on the outliers and to think about their story, and the mass in the middle looked squoze and muddled anyway. This is a good technique: label the outliers and leave the middle as points. This is, in spirit, related to Tukey's 10-plus-10 plots: he recommended plotting the top 10 and bottom 10 in order to see if he could pick out patterns. In our xy-plots, you might want to play with identifying the "envelope" and leaving the inside as dots.

If you have lots of rows of data, you can try splitting them into thirds by one of the variables and then plotting a subsample of each third.



1. The president has broad-based support.

2. Notice mountain states. We could have colored them separately if we wanted. Basically, the slope for the mountain west is about the same as for the rest of the states but offset with a different intercept. You could do a regression with a dummy variable with mountain west states.

3. This isn't causation. We're looking for interesting stories, not proving that Bush supporters are fat-assed.

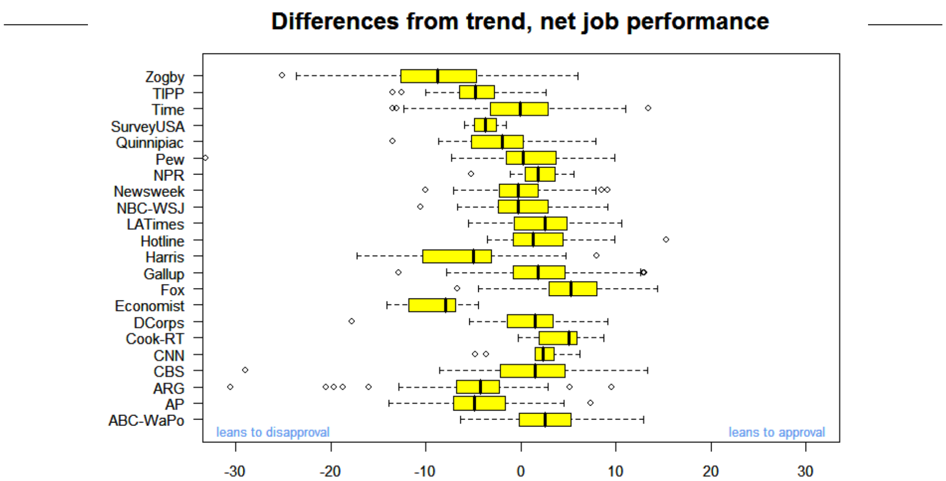
```
dat = read.csv("fat-vote.csv",comment="#")
head(dat)
with(dat,plot(obesity,(bush-gore)/total))

with(dat,plot(obesity,(bush-gore)/total,type="n"))
with(dat,text(obesity,(bush-gore)/total,as.character(state)))
```

more on labeling

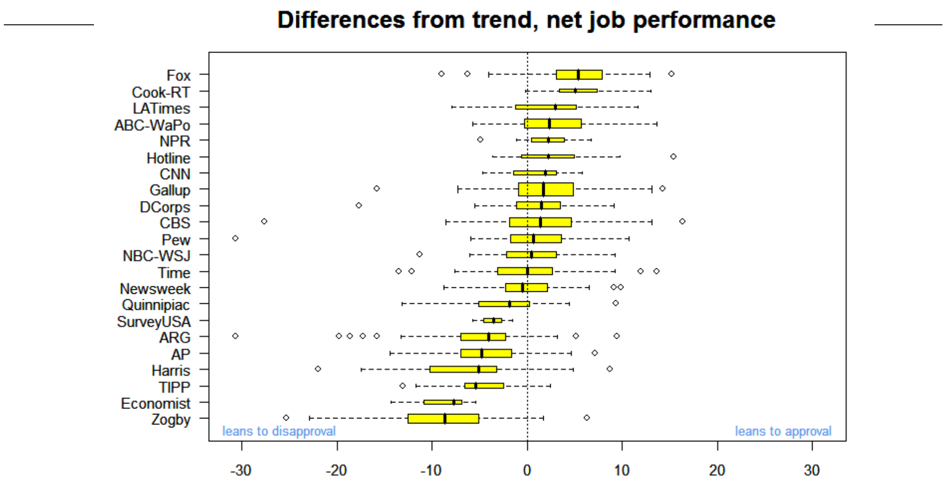
- direct labeling of lines often better than legends
    - particularly good when combined with line color
        - symbol plus line type often too busy to decode
        - looking back-and-forth at a legend is distracting

ordering

- default ordering for categorical variables is often alphabetical
  - that makes categoriess easy to find, but hard to compare
  - example: country data are often ordered by name of country rather than by the variable you're interested in
- find an ordering that makes sense and use it
  - if you are interested in mortality differences among countries, order by mortality not country name
  - this helps you spot and evaluate small differences between countries



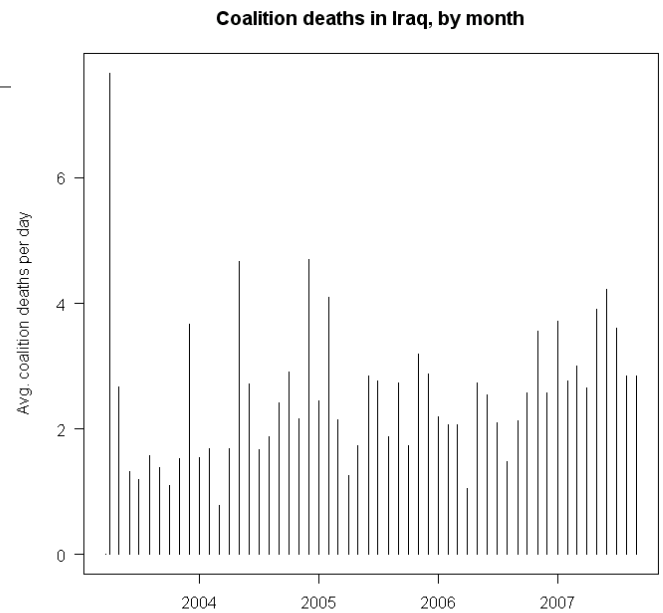
Different polling houses exhibit different amounts of "house effects" in their polls. These boxplots show, for each national pollling firm, the distribution of residuals from a model of presidential job approval. The ordering is alphabetical from bottom to top.



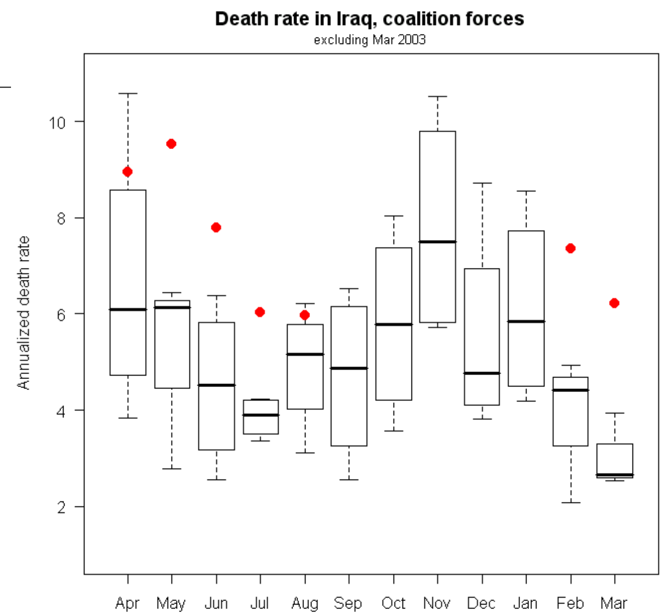
Similar plot to previous, but ordered by median residual, with boxwidth proportional to square root of sample size, and a reference line for zero average house effect.

grouping

- grouping (done well) helps with pattern recognition
  boxplots are a familiar way to group
- grouping (done poorly) obscures pattern
  not all “obvious” groupings are informative
- next two slides show (almost) same data



Coalition deaths in Iraq, by month
Avg. coalition deaths per day
0
2
4
6
2004
2005
2006
2007


Death rate in Iraq, coalition forces
excluding Mar 2003
Annualized death rate
10
8
6
4
2
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar

multivariate comparisons

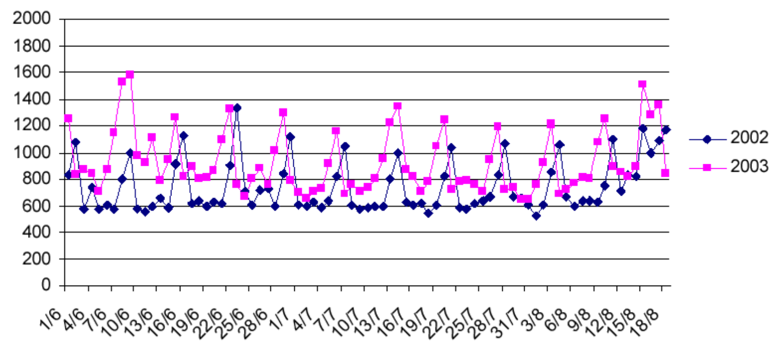
- show relationships
  - more importantly, give you ideas
- time series plots show you what happened when, but rarely why they happened
  - we'll want to dig deeper into the data to generate new questions about the 'why?'

Time-series plots are useful as descriptive summaries, but from an analytical point of view they don't usually provide enough motivation unless there are specific events that trigger sudden changes. They also tend to have relatively low information density, especially since one of the axes gets eaten up with time – time is a equi-interval variable, so it will often be more useful to use that axis for a variable whose observations are more complex. This is much more in the spirit of Descartes original insight.

patterns in multivariate data

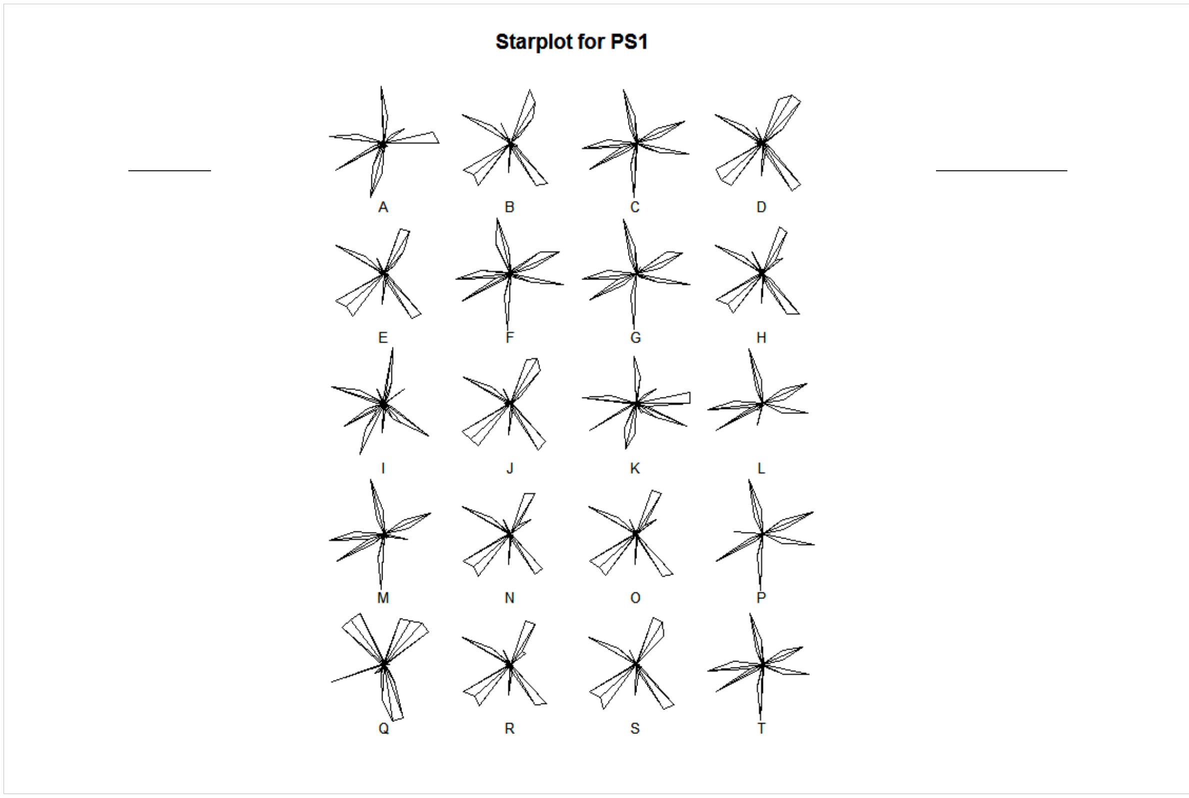
- twenty students read numbers of ambulance calls for July 2003 off a graph. How can we summarize the results?

Graphique 5 : nombre d'interventions du SAMU 13 en 2003 par rapport à l'année précédente (2002)

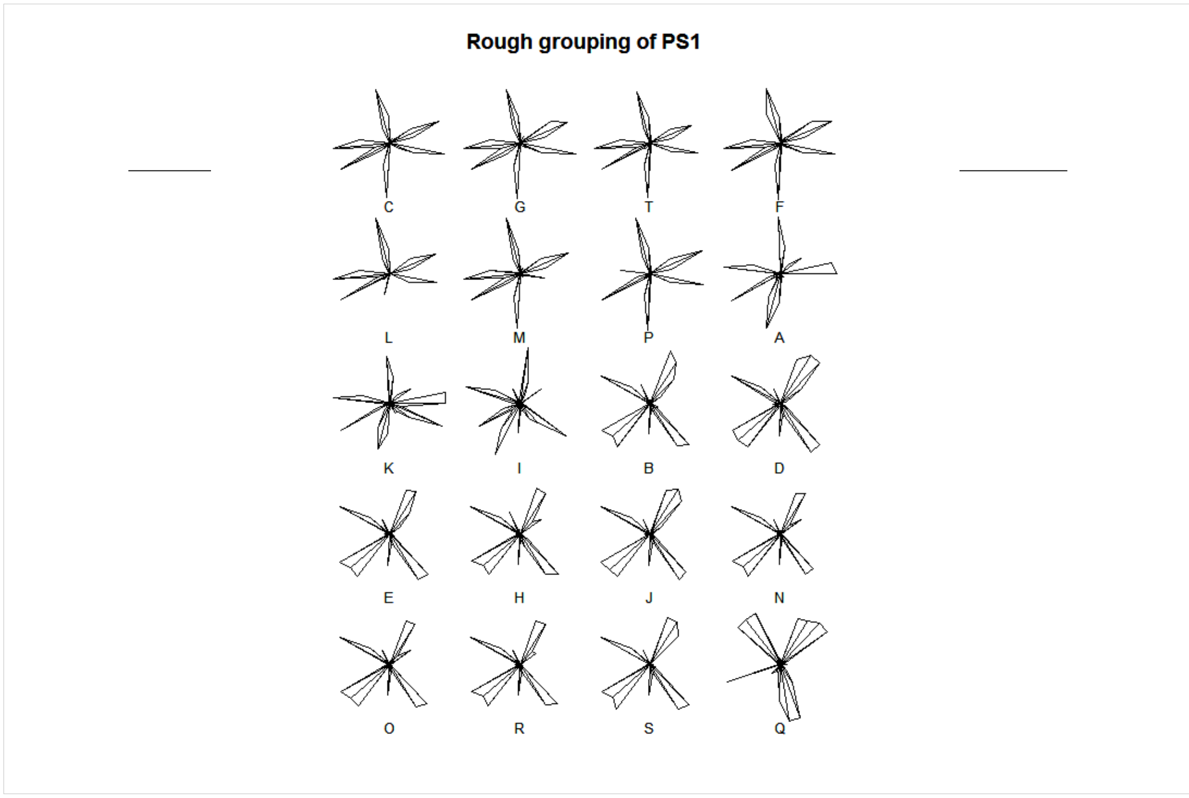



```
dat = read.csv("http://anonymous.coward.free.fr/mpa/ps2/ps1results.csv")

head(dat)

attach(dat)

stars(dat)

x <- as.matrix(dat)

stars(t(x))

stars(t(x),draw.seg=T)
```

Notice, by the way, that the original graph appears to have been done using Excel-like graphics. Grid lines are good, but dark grid lines compete for your attention. If you must use Excel, dim down the grid lines by making them a light gray.



stars(...)

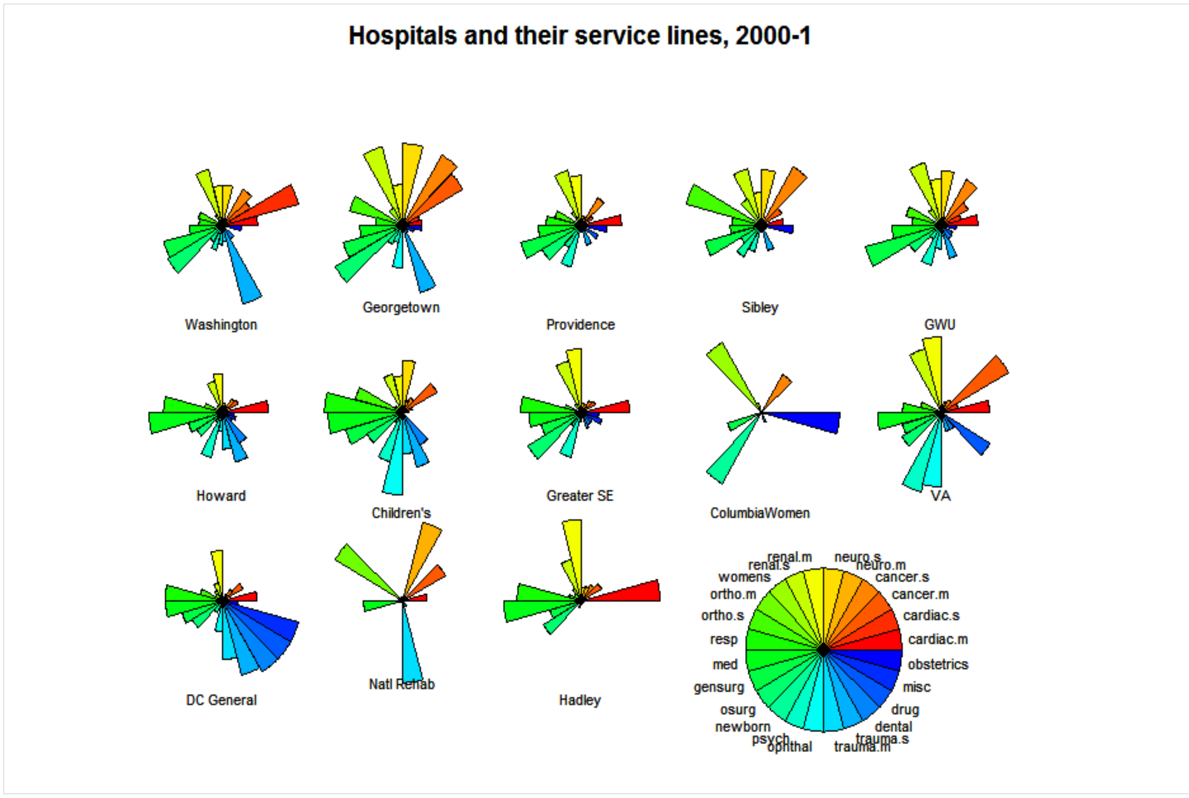


Stars can sometimes make it easy to spot patterns over many variables across individuals.

You can see another example using multivariate data on cars in R with example(stars)

dc hospitals

- thirteen hospitals
- twenty-four service lines
- do different hospitals specialize in different areas?



stars(..., draw.segments=T)

Notice how DC General looks different from all other hospitals. Georgetown and GWU look vaguely similar. Washington Hospital Medical Center has a large cardiac surgery program.

stars are like pies

- except that angle is constant and radius varies
    - in pies, radius is constant and angle varies
    - that’s why pie segments need labels
- watch out for colors

Though pie segments add to 100% and there's no such condition in stars.

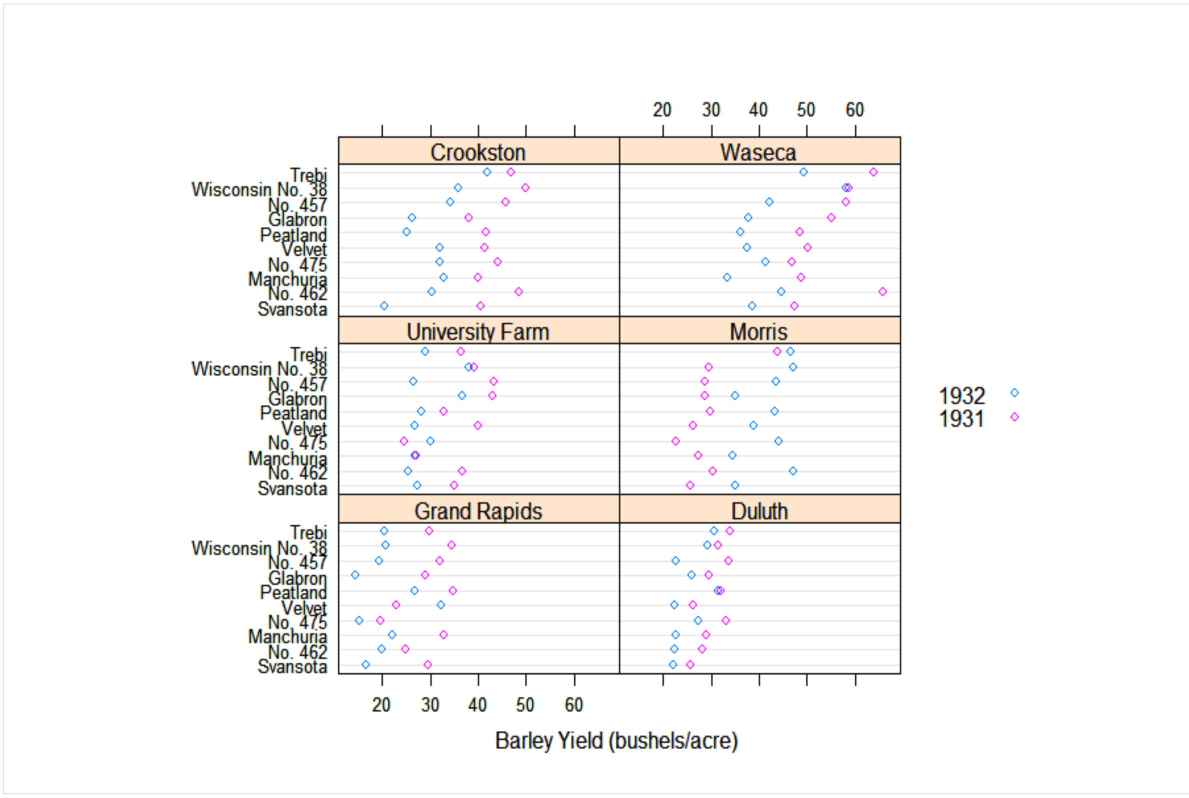
scatterplots

- can (sometimes) show more than two variables
  - can code categorical variables with color
  - can code some interval variables with size
- small multiples can show varying conditions
  - lattice (i.e., trellis) plots
  - use same scale and ranges, if possible, to enhance comparison

small multiples can be one of the best ways to exploit many comparisons at once. Try to use the same axis scales if at all possible.

dotplots (and trellis)

- conditioning plots
- barley yield
    - ten varieties
    - six plots
    - two years



library(lattice)

dotplot(variety ~ yield | site, data = barley, groups = year,

key = simpleKey(levels(barley$year), space = "right"),

xlab = "Barley Yield (bushels/acre) ",

aspect = 0.5, layout = c(2,3), ylab = NULL)

Though many statisticians worked on this data set, it took nearly half a century for someone to notice that the data for Morris appear to have been switched (Cleveland, 1984). That was realized from a plot like this.

You can see the “Morris” problem even more clearly if all six of the plots had been arranged in one long column: in R,

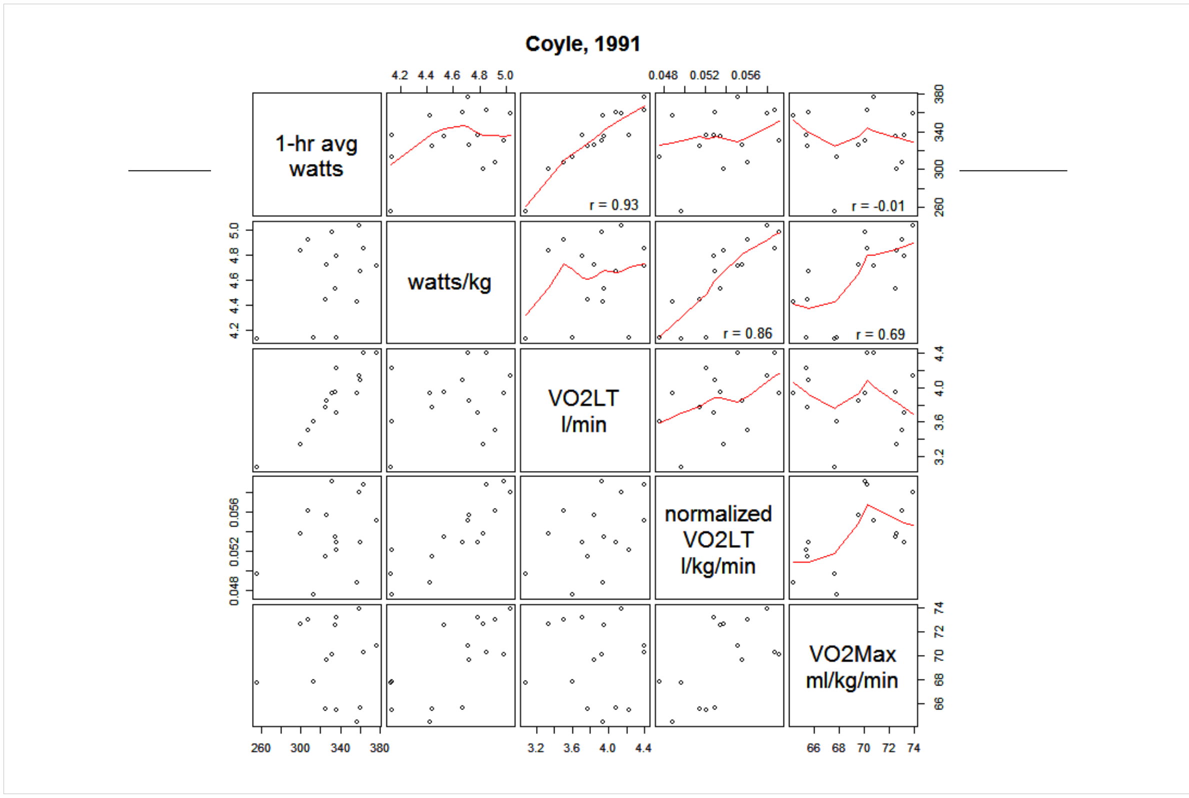
library(lattice)

example(barley)

scatterplot matrices

- scatterplot matrices compress a lot of information on bivariate relationships into a small space

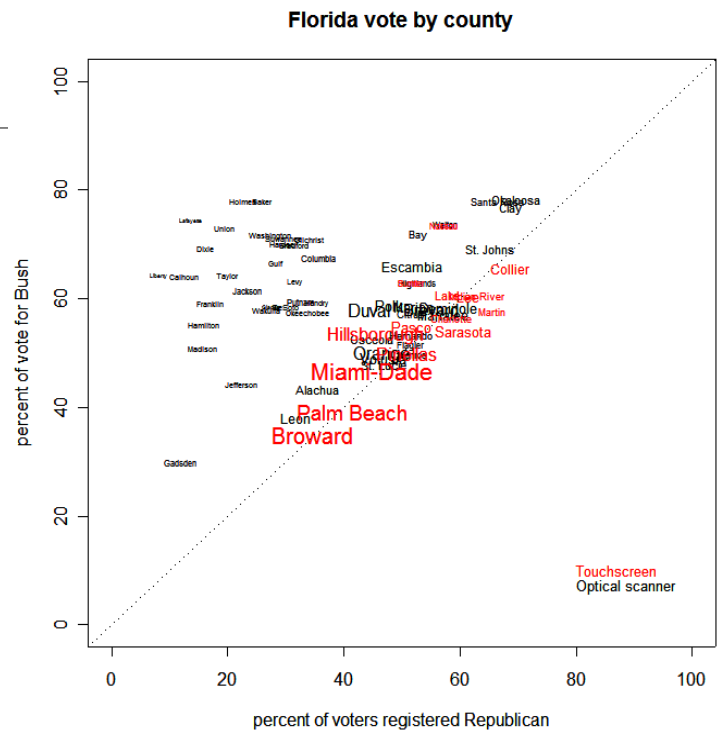
  useful for winnowing out uninteresting variables and deciding which variables might be worth further investigation



physiological measurements on 15 riders, elite and “state” class.

coding plotting symbols

- improves information density by tagging plotting symbols with attributes
  - you've seen this before using color or shape; can often combine with direct labeling



Here we show:

% repub

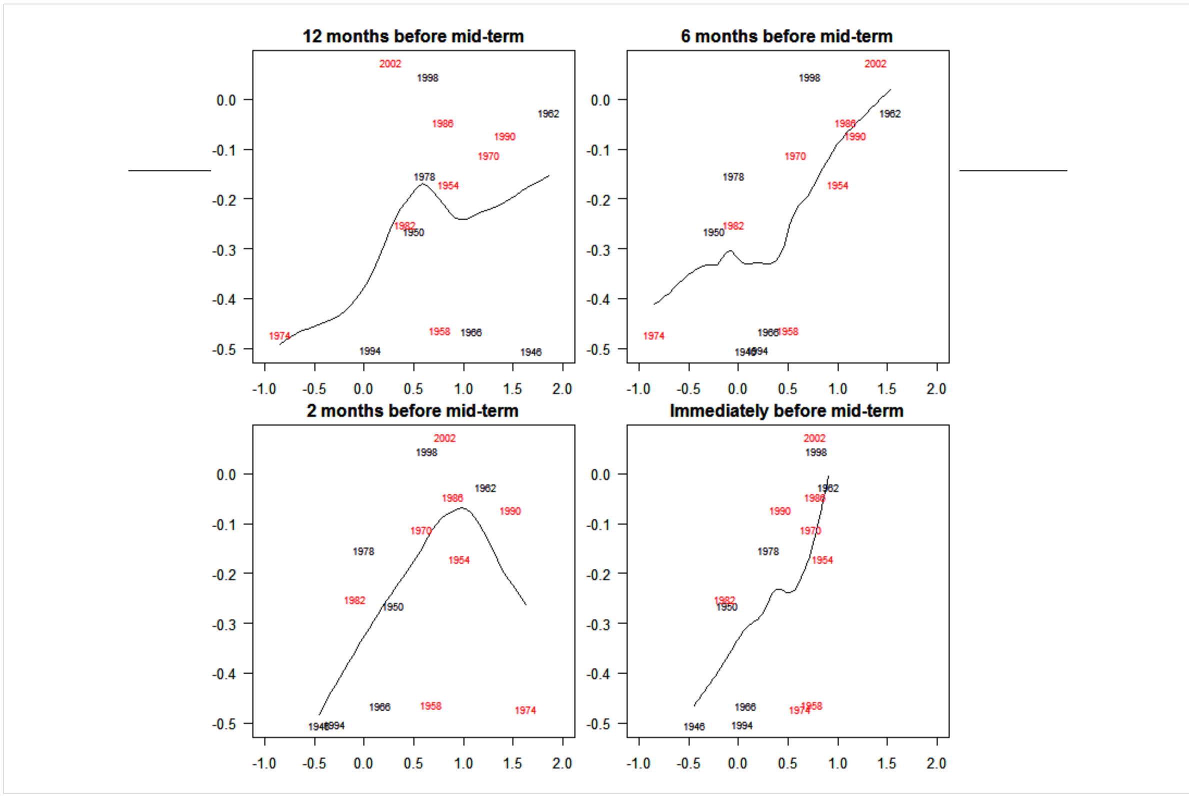
% Bush vote

name of county

type of voting machine

rough idea of county population

45-degree line to show that not many counties went more dem than their registration.



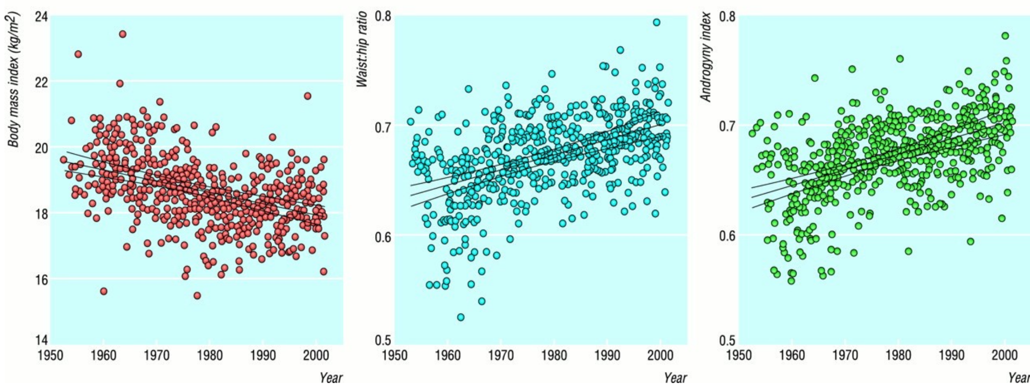
log(odds ratios) on y-axis, log(approval/disapproval) on x-axis
red means rep president, black is dem president
1974 and 1946 were strange years.

smoothing and straightening

- smooth lines
  - piecewise linearity
  - splines and lo(w)ess
- a ladder of re-expression
- the re-expression rule

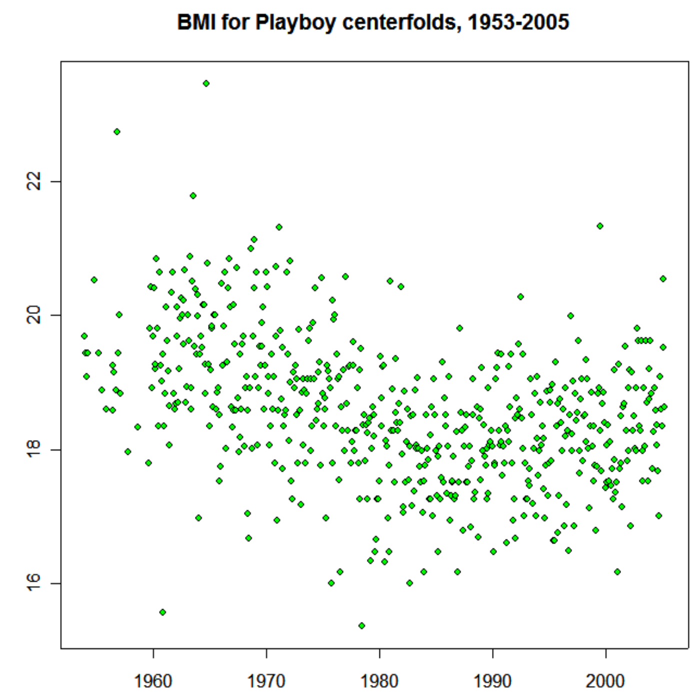
CABG EF relationship with in-hospital mortality: piecewise linear.

shapely centrefolds?



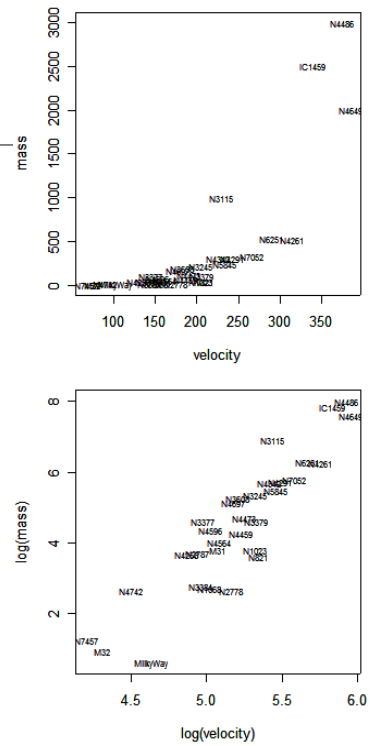
Voracek, M. et al. BMJ 2002;325:1447-1448
http://bmj.bmjjournals.com/cgi/content/full/325/7378/1447

Straight line or not?



Does that look straight to you?

I think the authors were looking for a straight line so they fit a straight line. We fit not-straight lines not only to help us guess relationships, but as a check to make sure our prejudices don't restrict our analyses.



transformation of axes in order to produce linear relationship

Black hole mass and dispersion of stellar velocities near galactic centers
Data from:
http://www.physics.ucsb.edu/~jatila/astro/astro2/b_hole_dispersion.html

a ladder of re-expressions...

3

2

1

½

#

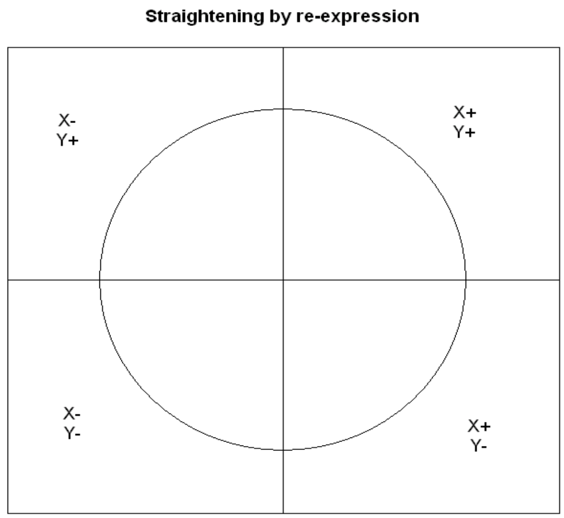
- ½

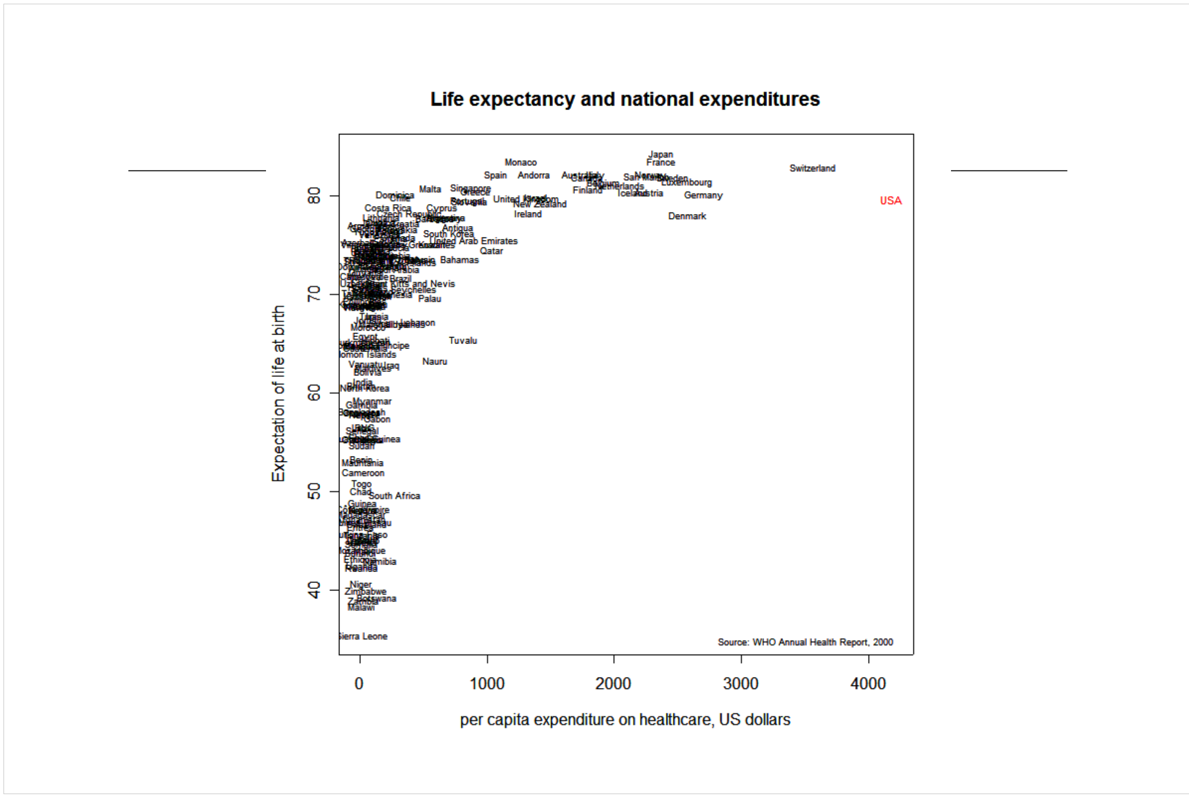
-1

-2

-3

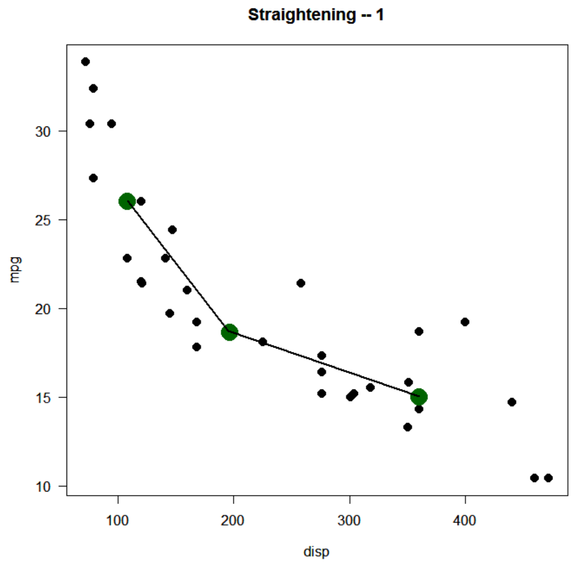
...and a rule for using them



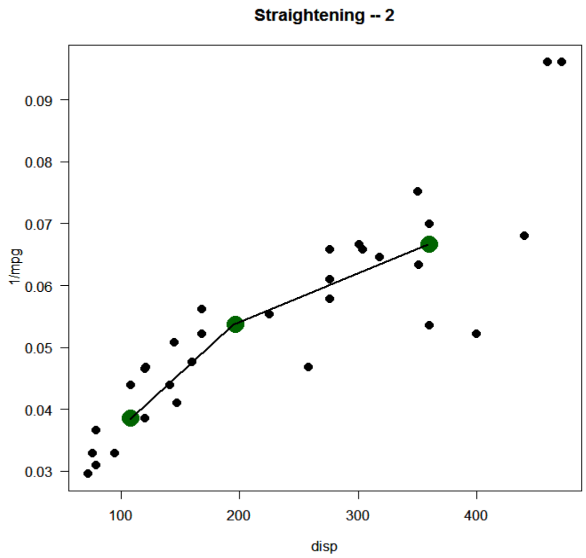



```
dat = read.csv(“who.csv”,comment=”#”)
head(dat)
summary(dat)
with(dat,plot(exp.percap,e0.female))
with(dat,plot(exp.percap,e0.female,log=”x”))
with(dat,identify(exp.percap,e0.female,country))
```

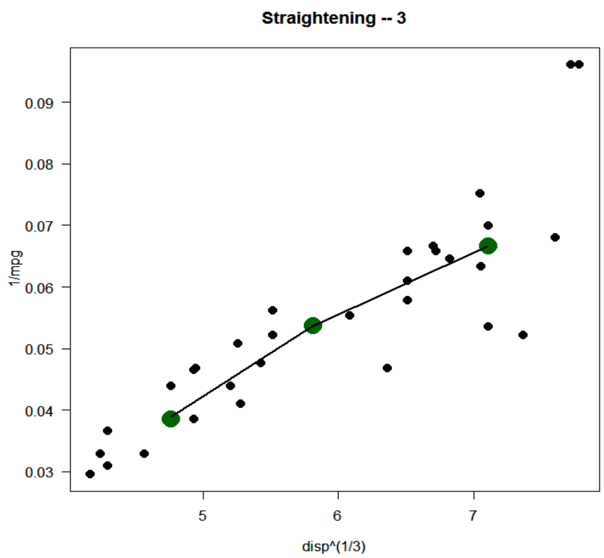
straightening



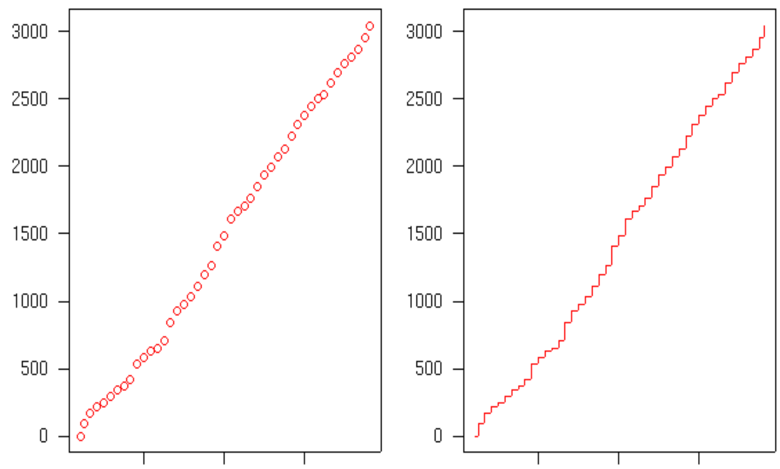
straightening 2



straightening 3



when is smooth too smooth?



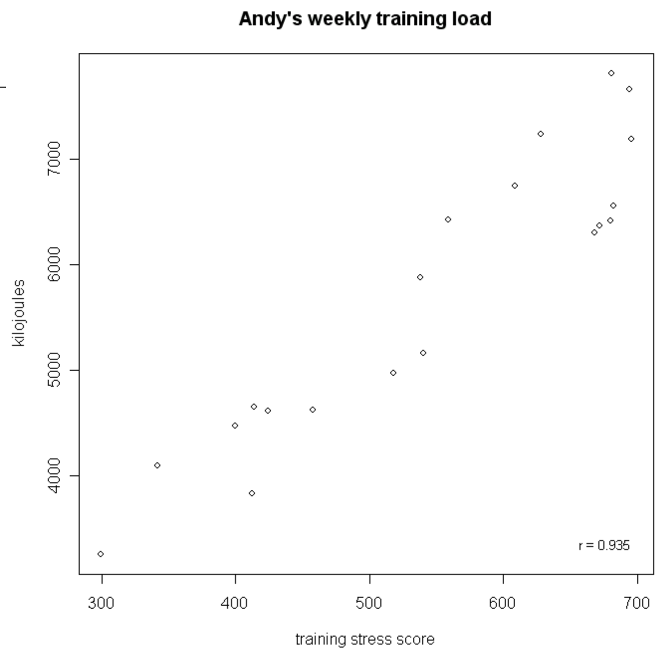
Coalition deaths in Iraq, 30 Mar 2003 - 23 Oct 2006

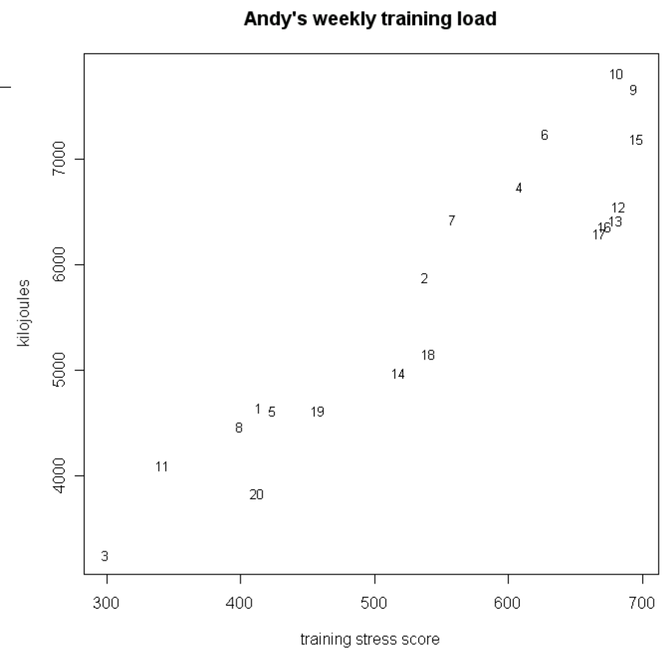
Data from http://icasualties.org/oif/

The eye wants to see “lines” connecting the dots, particularly when the dots are almost straight. In the right hand panel, it’s easier to force the eye not to connect the dots by using horizontal and vertical segments. It’s easier to see that a handful of months had large numbers of deaths.

when is straight too straight?

- can straight be too straight?


Andy's weekly training load
kilojoules
4000
5000
6000
7000
300
400
500
600
700
training stress score
r = 0.935



Andy's weekly training load
kilojoules
training stress score
300
400
500
600
700
4000
5000
6000
7000

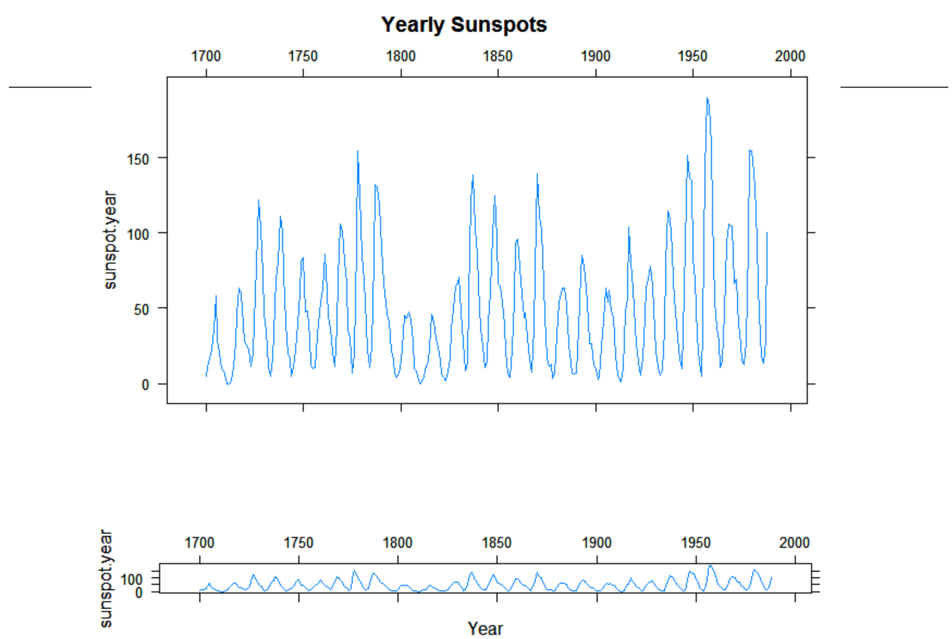
Andy's weekly training load
kJ/TSS
12.0
11.5
11.0
10.5
10.0
9.5
5
10
15
20
week

aspect ratios and banking

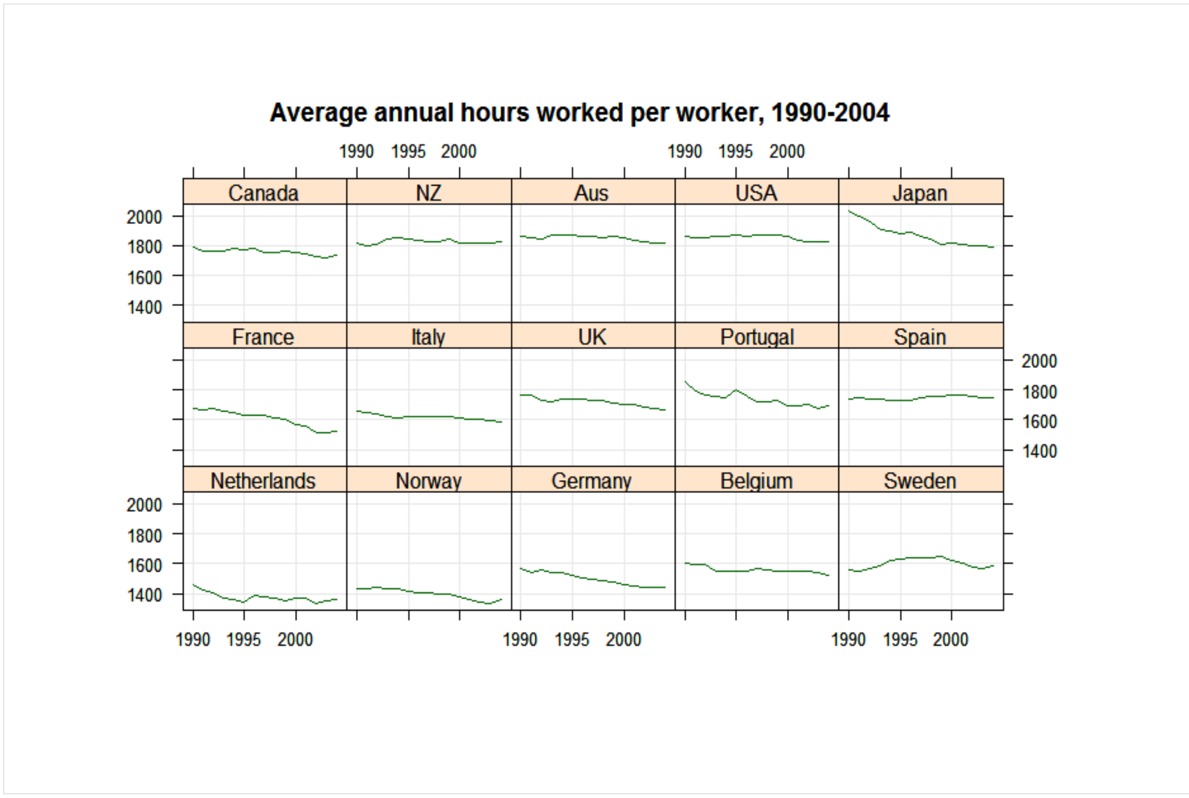
- human eye isn't great at decoding angles
  - banking helps the eye to decipher angles

If you use base R, and you have separate graph window you can just grab the corner to change the apparent aspect ratio.

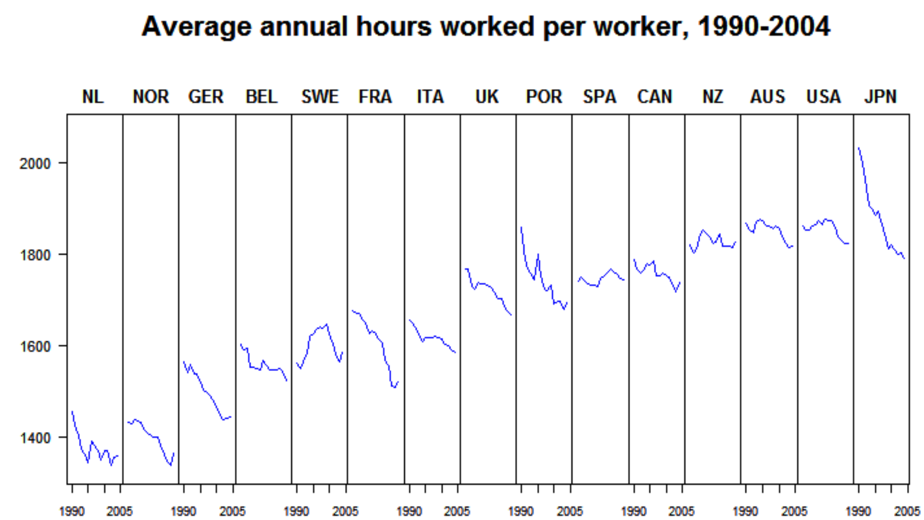
If you use Rstudio, you can write graphs to dev.new() and grab the corner to change the aspect ratio. When you want ot close the window, use dev.off()



Lower panel is low-aspect ratio. Note that for the 18th and 19th C. that sunspots ramped up faster than they declined, but that in the 20th they were more symmetric. Can't see that in the top panel.



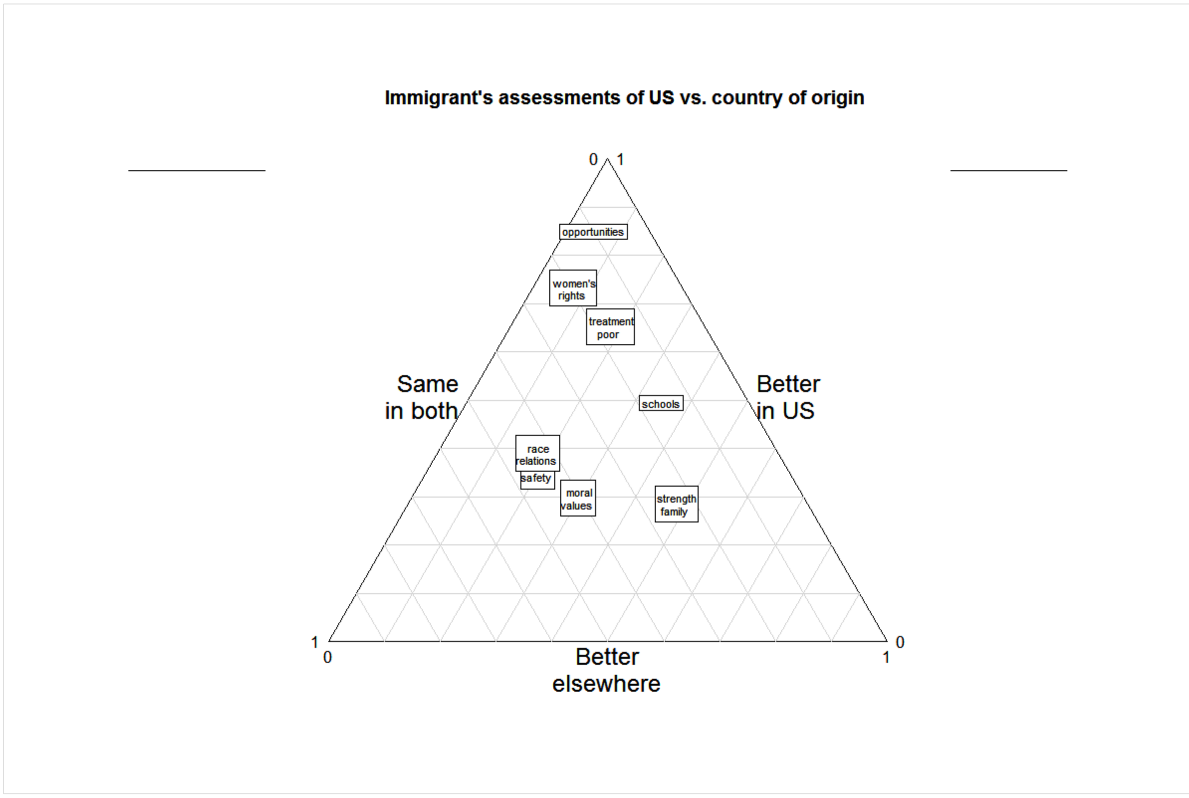
Because of the previous graph, Tufte generally recommends low-aspect ratios for time series. Here and on the next page are a counter-example.

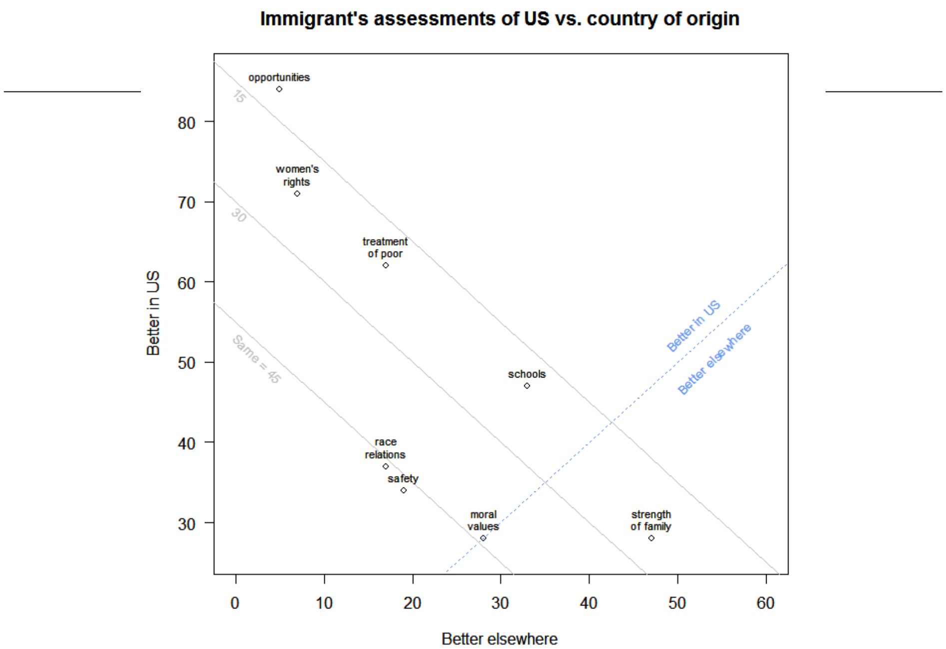


The aspect ratio is so high here that I had to abbreviate the country names, but it shows both the levels among the countries and also that Sweden's work hours went up and then came back down (sort of the same with the US).

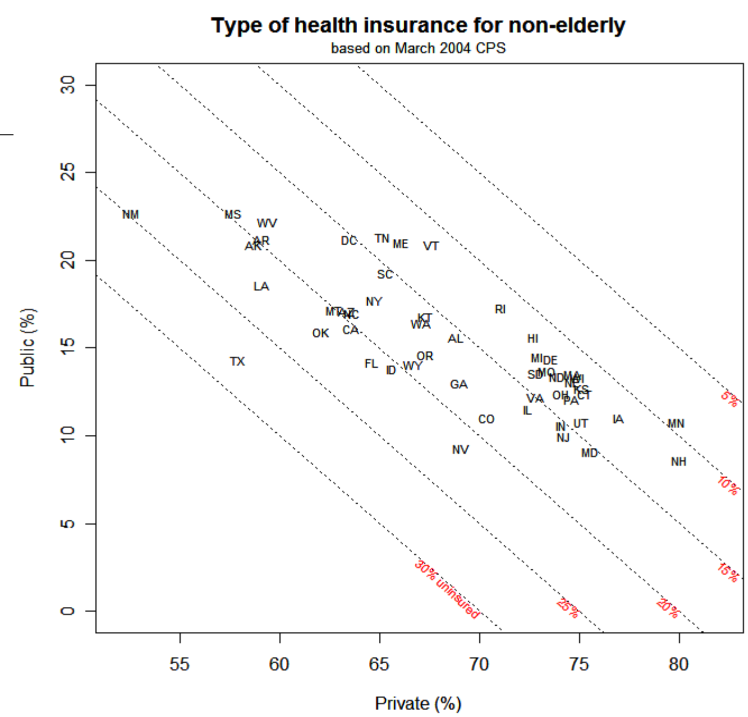
contours and bases

- triangle plots (= ternary plots)
  - soil texture plots
  - two degrees of freedom
- function contours can add context


Immigrant's assessments of US vs. country of origin
Same in both
Better in US
Better elsewhere
opportunities
women's rights
treatment poor
schools
race relations
safety
moral values
strength family
0
1
1
0
0
1



Immigrant's assessments of US vs. country of origin
opportunities
women's rights
treatment of poor
schools
race relations
safety
moral values
strength of family
15
30
Same = 45
Better in US
Better elsewhere
Better in US
Better elsewhere
80
70
60
50
40
30
0
10
20
30
40
50
60



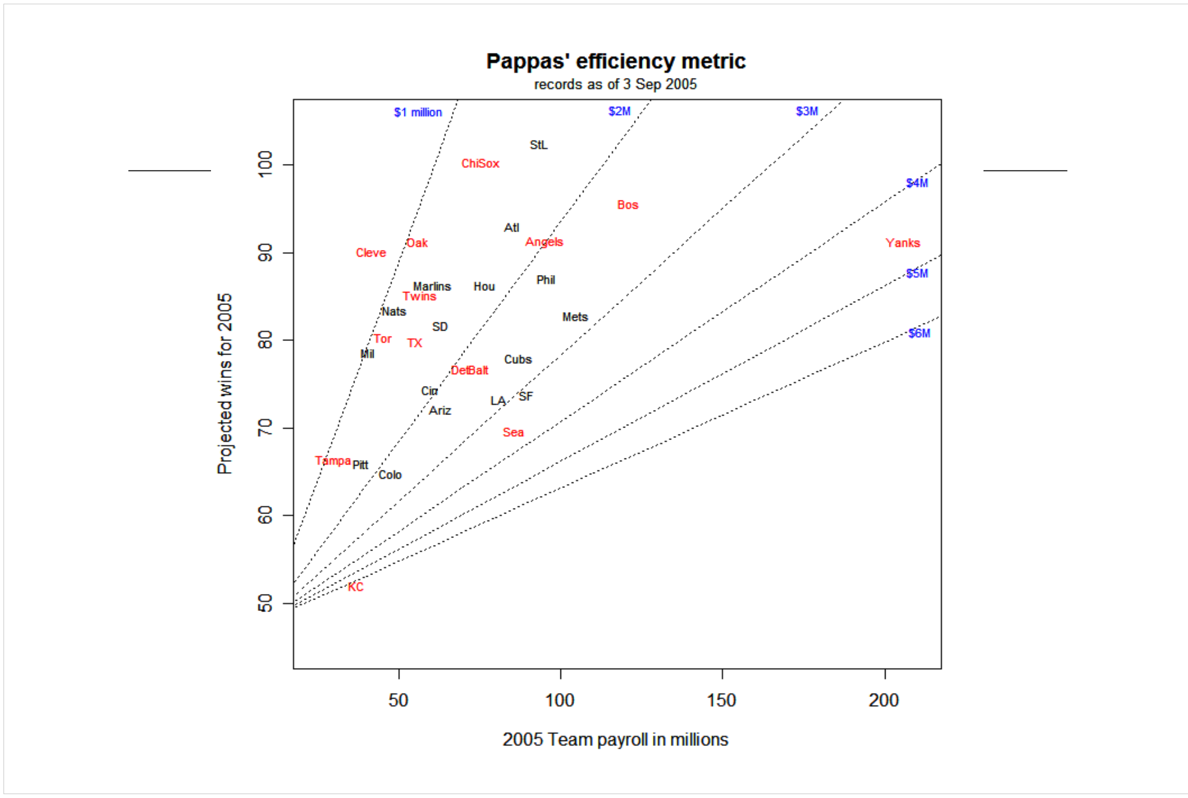
curve(…)

This is a ternary plot on normal x-y axes. Some people have problems "visualizing" a ternary plot.

Private insurance + public insurance + uninsured add to 100%, so plot any two and the third can be shown by contours counting down from upper right.

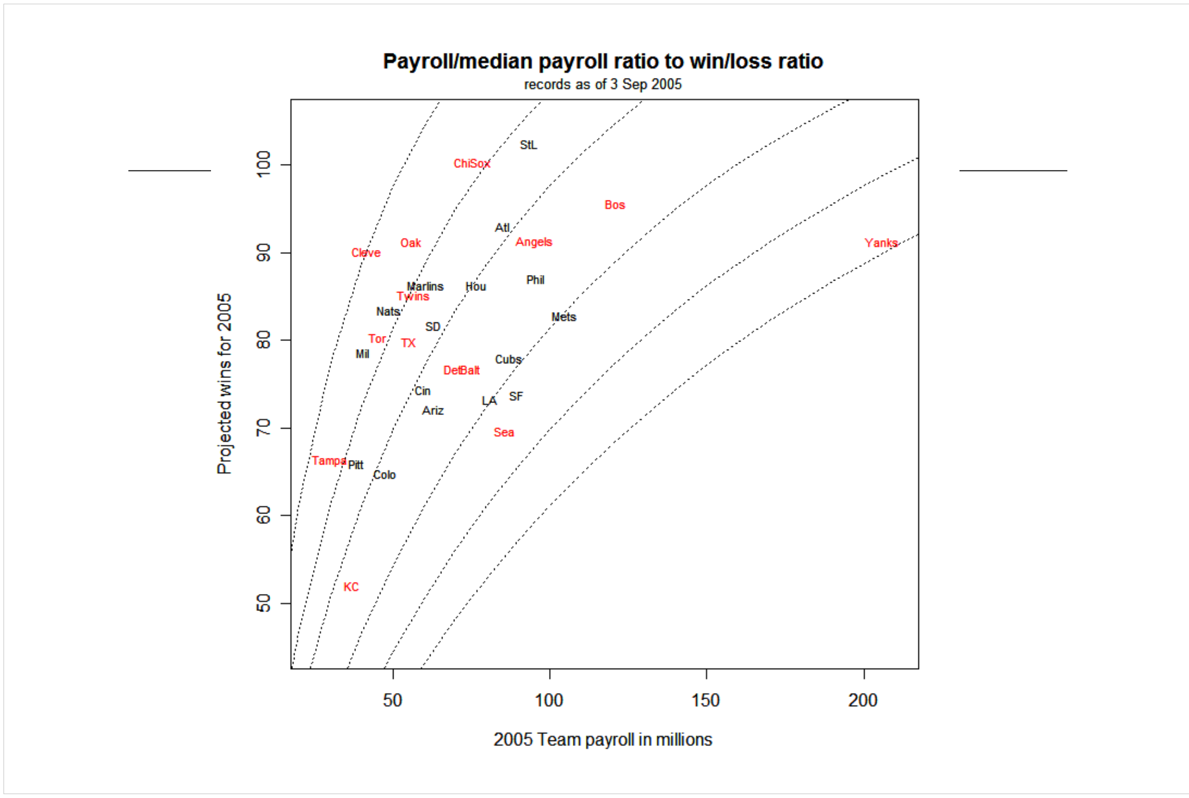
Texas has highest percentage of uninsured.

Note that across states, as private% decreases, public% increases – but not at the same rate. Cross-state, safety net isn't uniform.

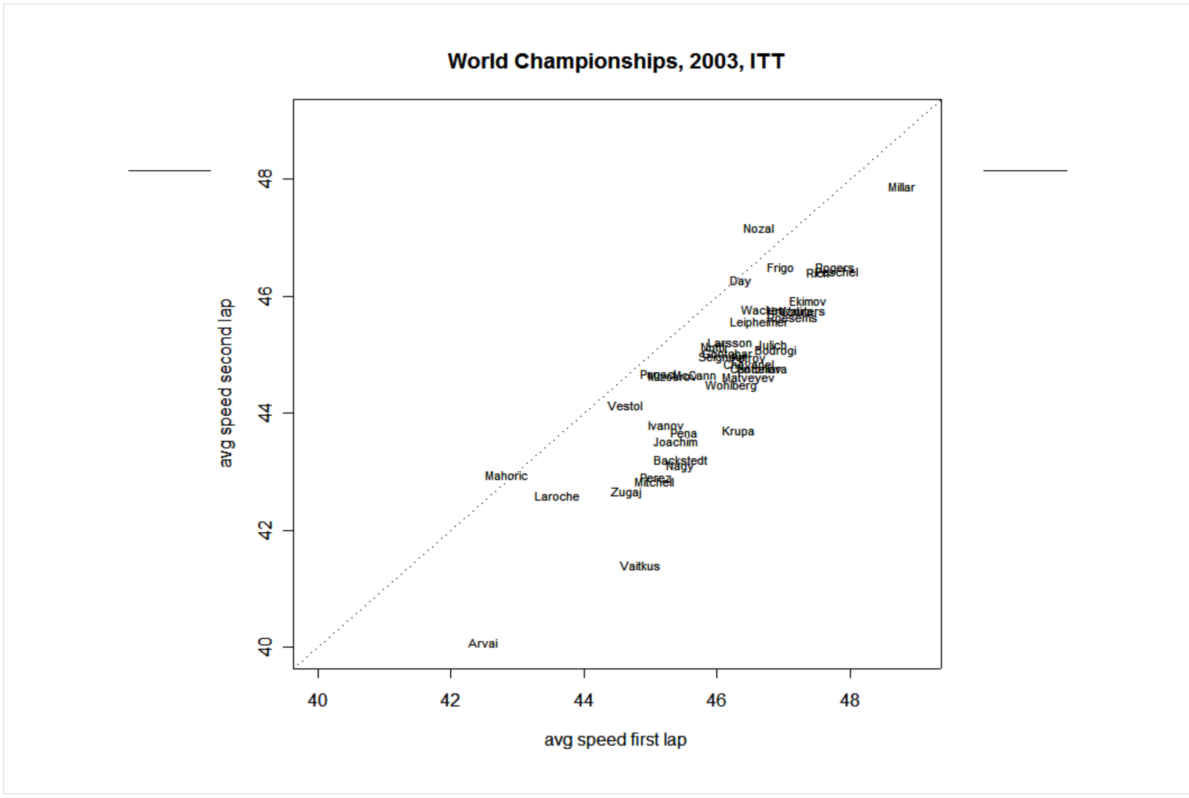


For context, read:

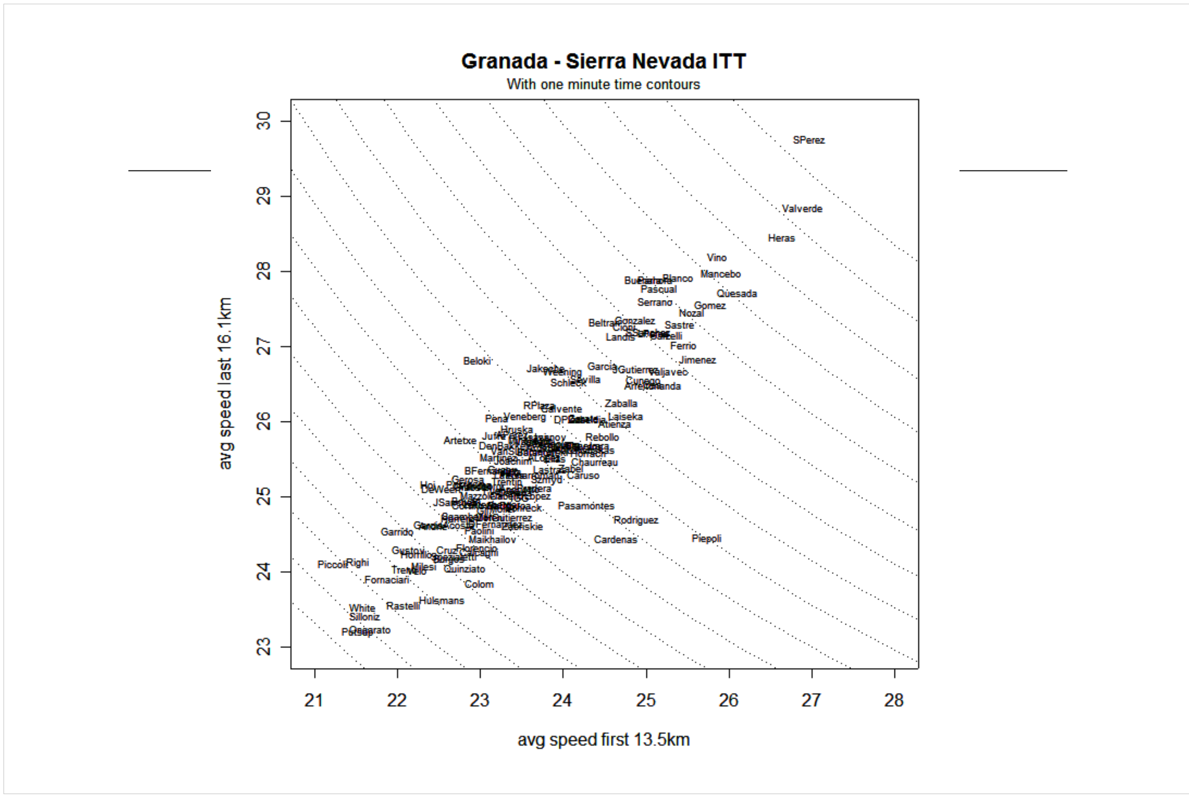
http://junkcharts.typepad.com/junk_charts/2005/08/baseball_roi_3_.html



Payroll/median payroll ratio to win/loss ratio
records as of 3 Sep 2005
Projected wins for 2005
2005 Team payroll in millions
50
60
70
80
90
100
150
200
StL
ChiSox
Bos
Atl
Angels
Yanks
Oak
Cleve
Marlins
Hou
Phil
Twins
Nats
SD
Mets
Tor
TX
Mil
Cubs
DetBalt
Cin
SF
LA
Ariz
Sea
Tampa
Pitt
Colo
KC



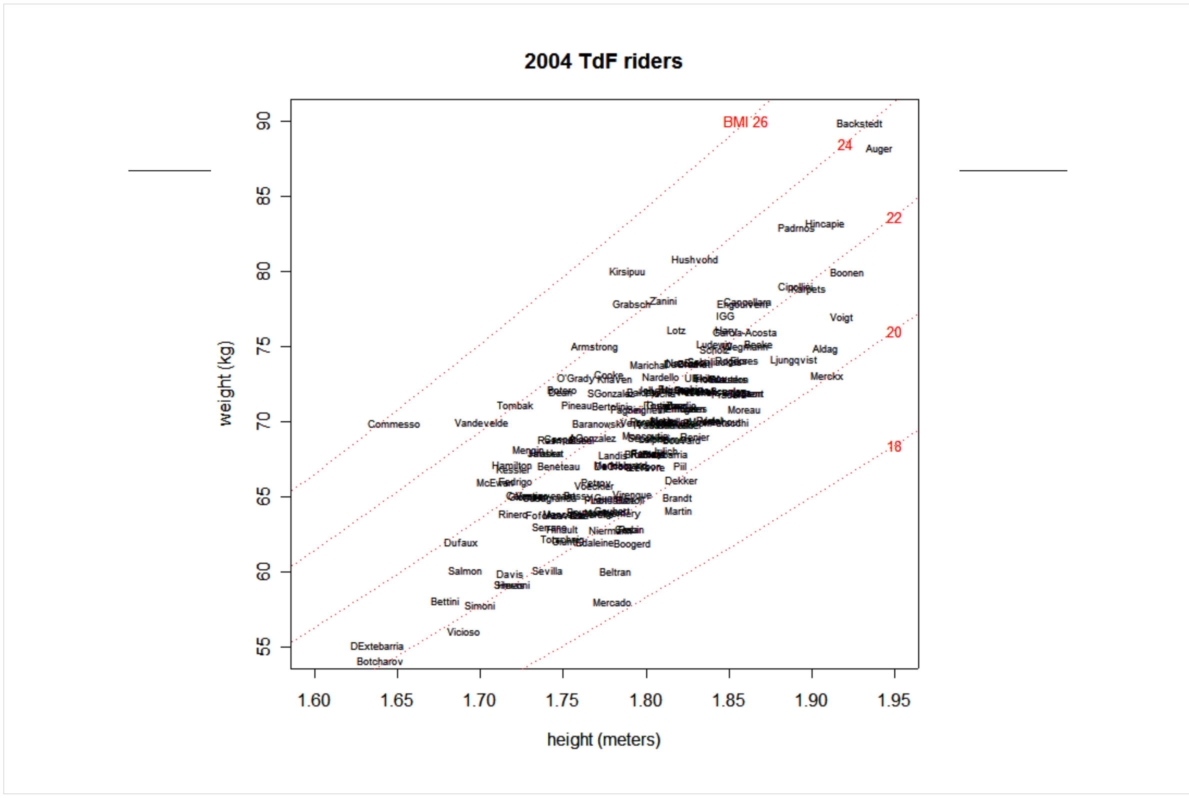
Millar was way faster than anyone else. A few months later he forced to renounce world championship after admitting using EPO prior to the race.



Similar to previous plot, but with one-minute contours added (I think the contours are too much info -- if I were to do this again, I’d restrict them to a smallish area around the data.)

2004 Vuelta a Espana. Stage 15
Perez was tossed out of Vuelta for using EPO

The contours extending across the page are distracting. If I re-made this thing I'd cut down on size or number.



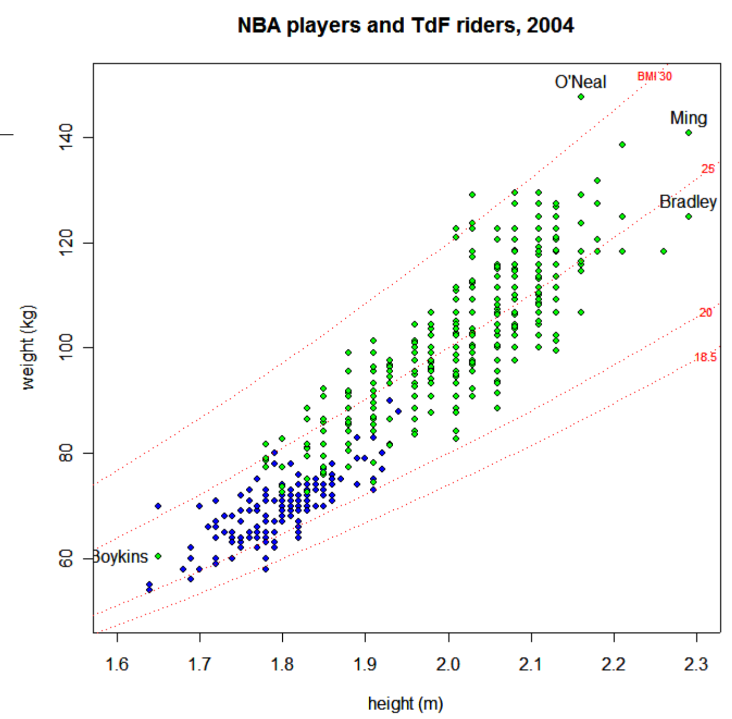
BMI = kg/m^2. In the US, the current standard is

BMI<18 = “underweight”

18<BMI<25 = “normal”

25 < BMI < 30 = “overweight”

BMI > 30 = “obese”

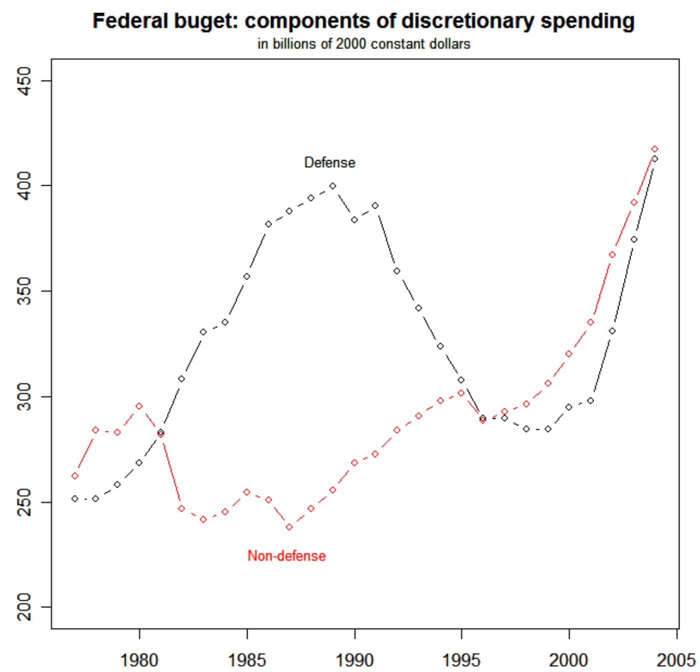


BMI = kg/(m^2)

This graph shows that BMI doesn't scale well with height -- Yao Ming is very, very tall but he doesn't appear overweight. Shaquille O'Neal isn't obese.

decomposing series into phase plots

- another version of “show the difference”

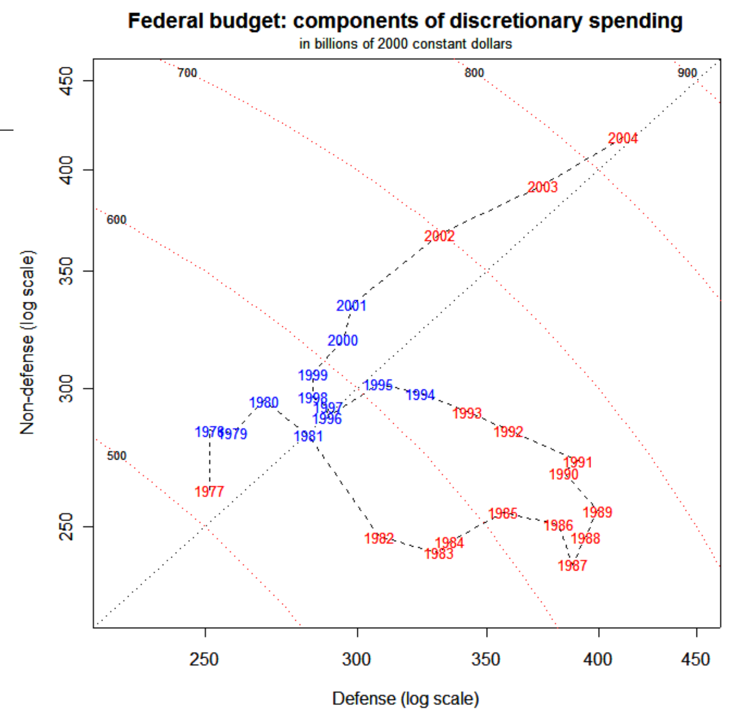


This is the conventional way to show the data – plotted against time. However, as we've already seen, it's hard for the eye to estimate distances between lines.

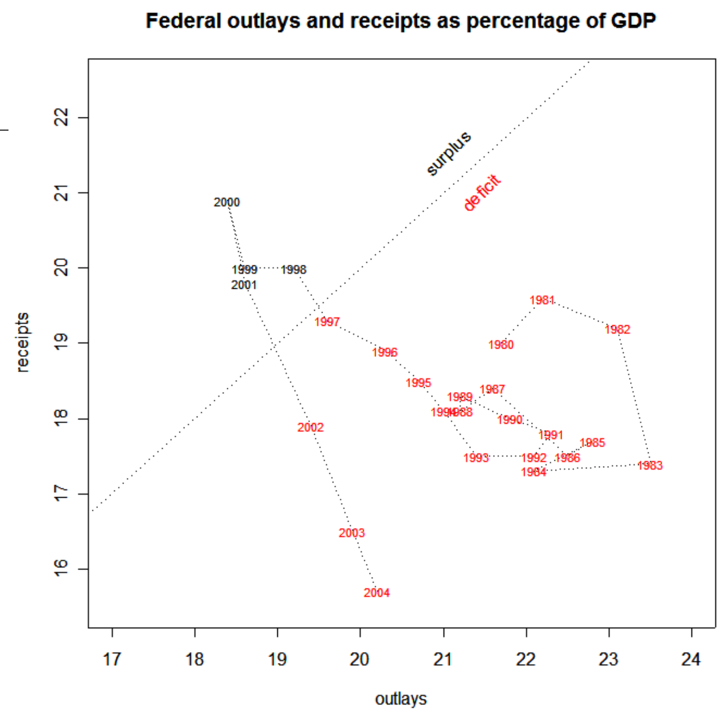
note: lines labeled, so no need for legend

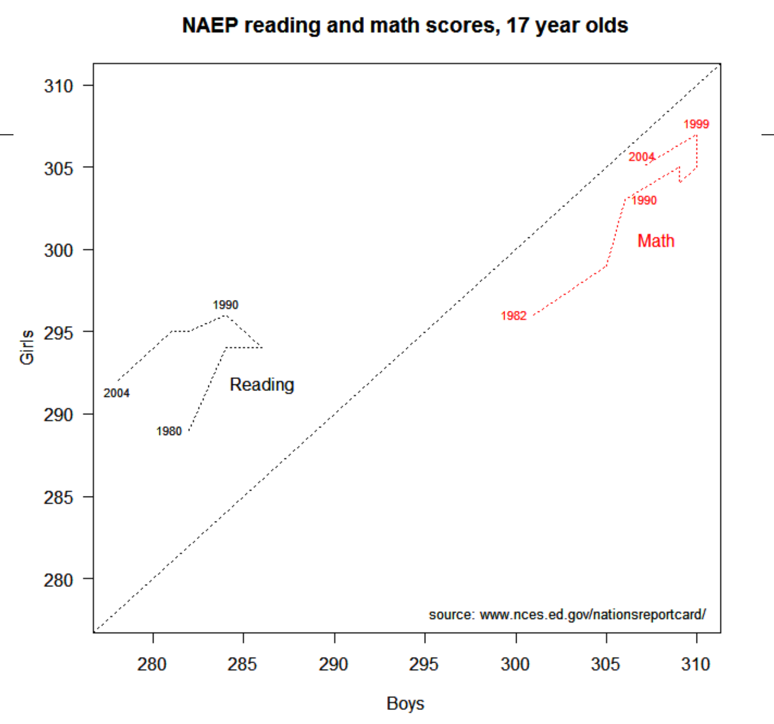
Alternatively, you could look at percent of GDP

```
dat = read.table(“discretionary.txt”,comment=”#”,header=T)
head(dat)
with(dat,plot(year,defense))
with(dat,lines(year,domestic,col=2))

with(dat,plot(defense,domestic,type=”n”)
with(dat,text(defense,domestic,year,cex=.7))
```



Of discretionary spending, this graph shows amount for defense, amount for non-defense, total, growth rate, year, presidential party (Red=budget proposed by a Republican president, blue=budget proposed by a Democratic president)


Federal outlays and receipts as percentage of GDP
receipts
outlays
surplus
deficit
2000
1999
1998
2001
1997
1996
1995
1981
1982
1980
1987
1989
1988
1990
1991
1985
1986
1992
1993
1984
1983
2002
2003
2004
16
17
18
19
20
21
22
23
24

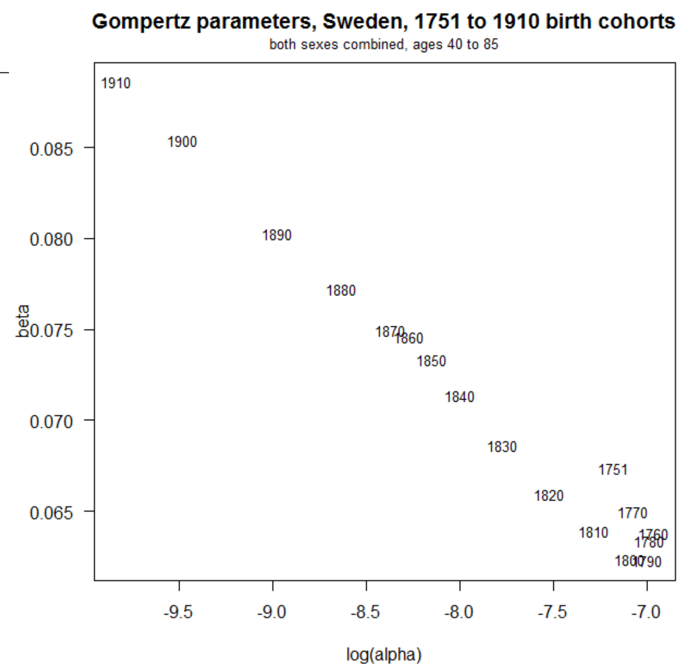


national assessment of educational progress.

You can see that math scores for boys and girls seems to move up and down together, but not so for reading scores. Also, it's easy to pick out and compare differences between boys and girls.

plot summaries for simplification

- when all subset have same contrasts, plot subset summaries
  - sometimes can get away with it even if not all subsets have all same contrasts—but then must be doubly careful
  - helps to identify patterns
  - plot and identify extremes, leave middle alone
    - this is the idea underlying “10 plus 10” plots
    - or, split into n groups (n small, like 3), and plot subsamples from each

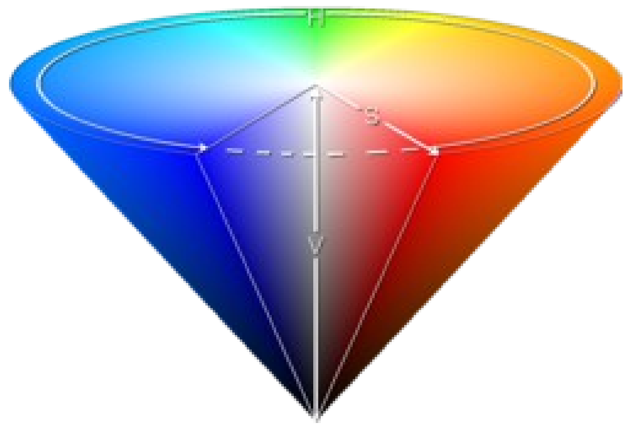


The Gompertz alpha and beta parameters summarize a particular kind of "fit" to the hazard rates. Here's a tip: if you're estimating model parameters over time, or over space, or for some other kind of contrast, plot the parameters.

more on color

- HSV
  - h=hue, s=saturation, v=value
  - sometimes called HSL for hue, saturation, luminance
- equal impact colors
  - CIELUV and Munsell are systems of color perception
    - medium saturation, kind of pastel-like

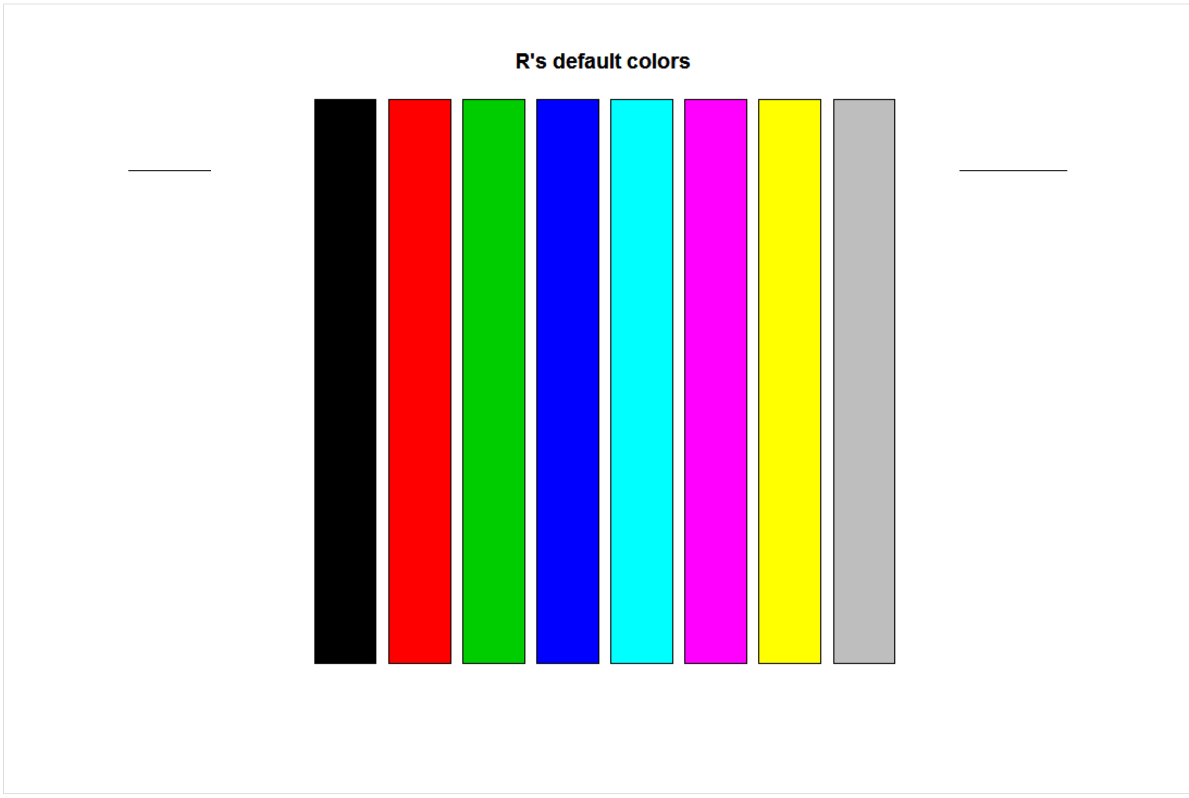
HSV cone



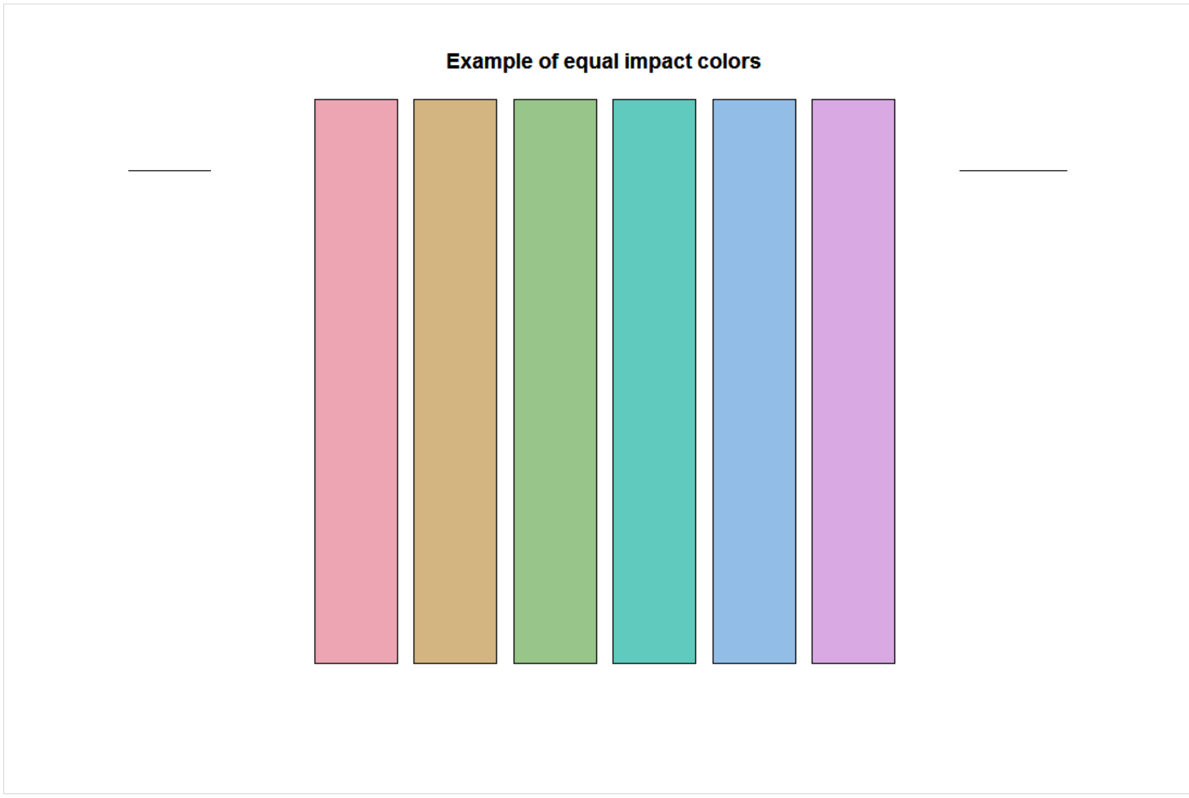
Think of the outside edge of the top of the cone as an old-fashioned color wheel. That's why hues aren't ordered – they form a circle.

As saturation goes up, you go from center to edge, so highly saturated colors appear brighter.

The human eye's ability to distinguish color depends on how much light there is. As value goes up, the luminance goes up, so the bottom of the cone is black. The top center is white because it has high value and low saturation for any particular hue.

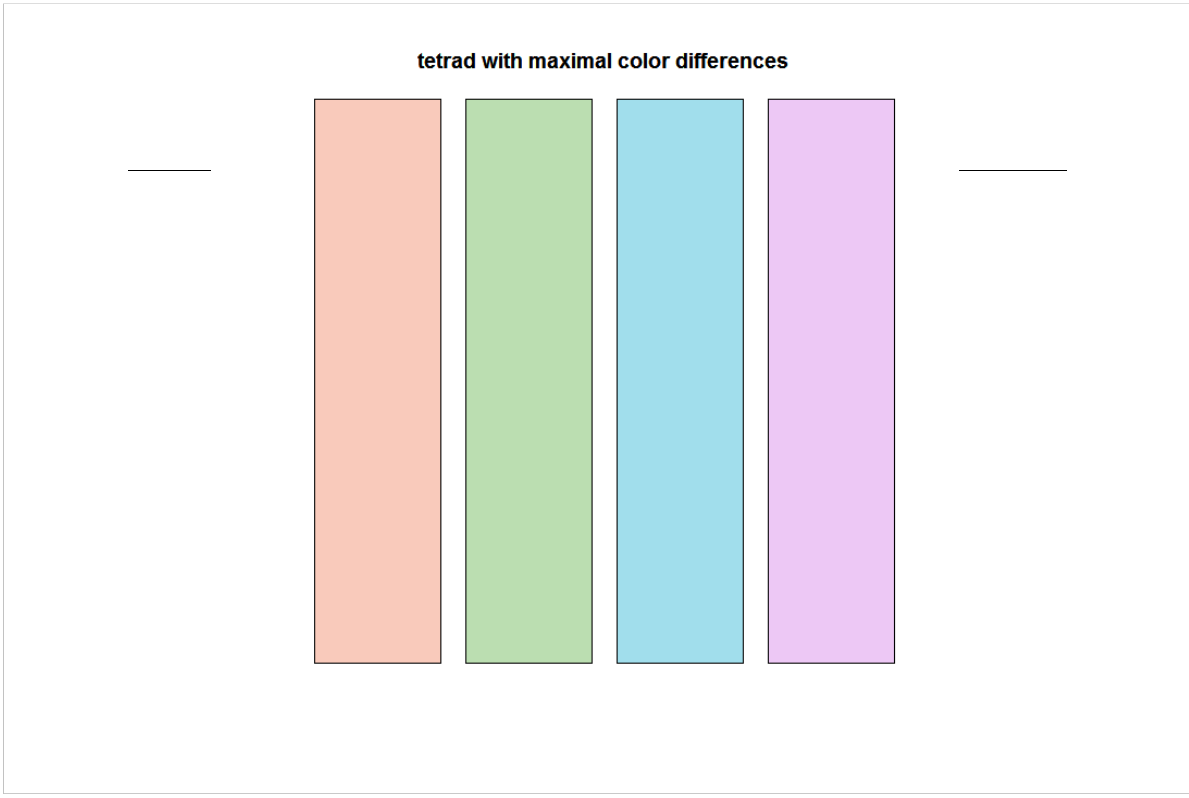


Yikes.



hcl() is one way to get equal impact colors. These bars were produced with:

barplot(rep(1,6),col=hcl(h=seq(0,300,len=6),l=75,c=45),axes=F,main="Example of equal impact colors")




```
barplot(rep(1,4),col=hcl(h=seq(30,300,len=4)),axes=F,main="tetrad with maximal color differences")
```

basic techniques

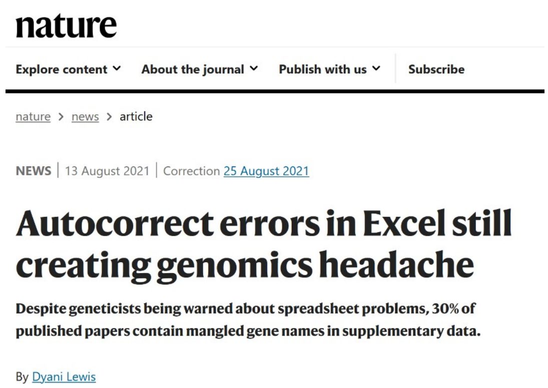
- show the difference
- identify outliers (or, label directly)
- group and order
- plot extremes
- multiple comparisons

slightly more advanced techniques

- smoothing
- straightening
- phase plots
- contours
- banking
- coloring

stuff I wanted to hide until the end

- friends don't let friends graph with Excel
  - but let's be realistic: sometimes you have no choice
  - dates in Excel are particularly a problem



nature

Explore content About the journal Publish with us Subscribe

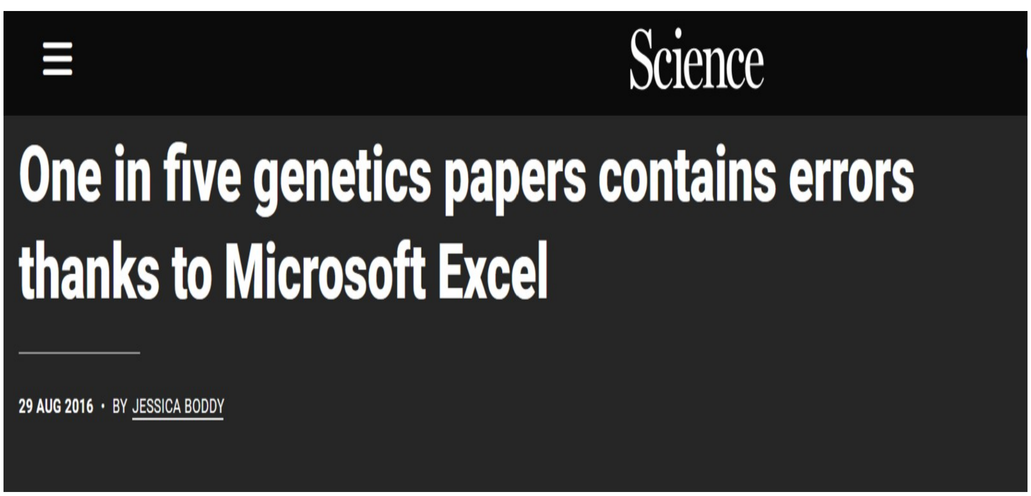
nature > news > article

NEWS | 13 August 2021 | Correction 25 August 2021

Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

By Dyani Lewis



Science

One in five genetics papers contains errors thanks to Microsoft Excel

29 AUG 2016 • BY JESSICA BODDY

how a demographer changed
bicycle racing and design