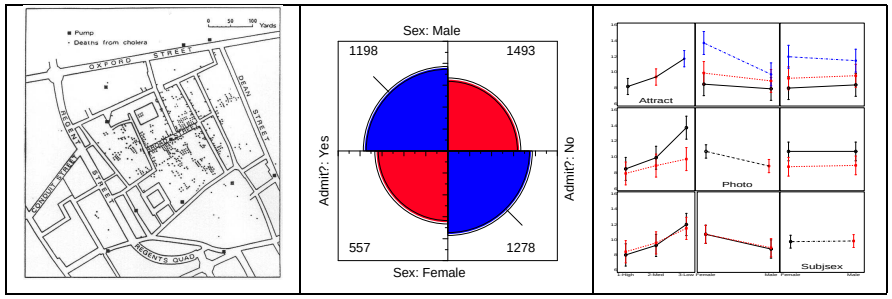
Visions of The Past, Present and Future of Statistical Graphics

(An Ideo-Graphic and Idiosyncratic View)



Michael Friendly

York University

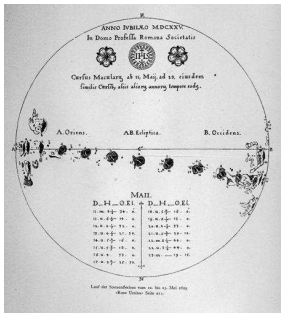
American Psychological Association

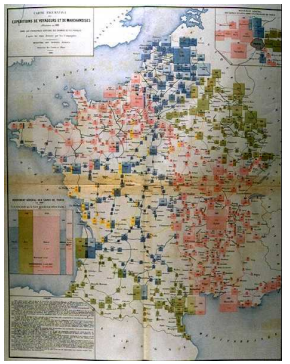
August, 2003

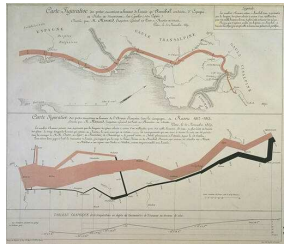
Visions of the Past

The only new thing in the world is the history you don't know. Harry S. Truman

- The *Milestones* Project
- The Golden Age of Statistical Graphics
- Re-Visions of Minard







Milestones Project: Roots of Data Visualization

- **Cartography**
  - early map-making ⟶ geo-measurement ⟶ thematic cartography
  - GIS, geo-visualization
- **Statistics, statistical thinking**
  - probability theory ⟶ distributions ⟶ estimation
  - statistical models ⟶ diagnostic plots ⟶ interactive graphics
- **Data collection**
  - early recording devices
  - "statistics" (numbers of the state): population, mortality ⟶ census, surveys
  - economic, social, moral, medical, . . . statistics
- **Visual thinking**
  - geometry, functions, mechanical diagrams, EDA
- **Technology**
  - paper, printing, lithography, computing, displays, . . .

Milestones Project: Goals

- Comprehensive catalog of historical developments in *all* fields related to data visualization
- ⟶ collect detailed bibliography, images, cross-references, web links, etc.
  - 220 milestone items (6200 BC – present)
  - 240 images, portraits
  - 140 web links (biographies, commentary)
  - 250 references
- ⟶ enable researchers to study themes, antecedants, influences, trends, etc.
- Web version: `http://www.math.yorku.ca/SCS/Gallery/milestone/`
  - Present form: hyperlinked, chronological listing (HTML, PDF)
  - Next: searchable by subject, content, author, country, etc. (LaTeX⟶ XML)

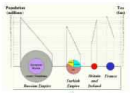
Milestones: Content Overview

Every picture has a story – Rod Stewart

c. 550 BC: The first world map? (Anaximander of Miletus)

1669: First graph of a continuous distribution function (Gaunt's life table)– Christiaan Huygens.

1782: First topographical map- M. du Carla-Boniface

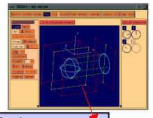


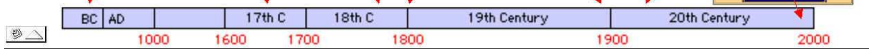
1801: Pie chart, circle graph - William Playfair

1896: Area rectangles to display two variables and their product- Jacques Bertillon

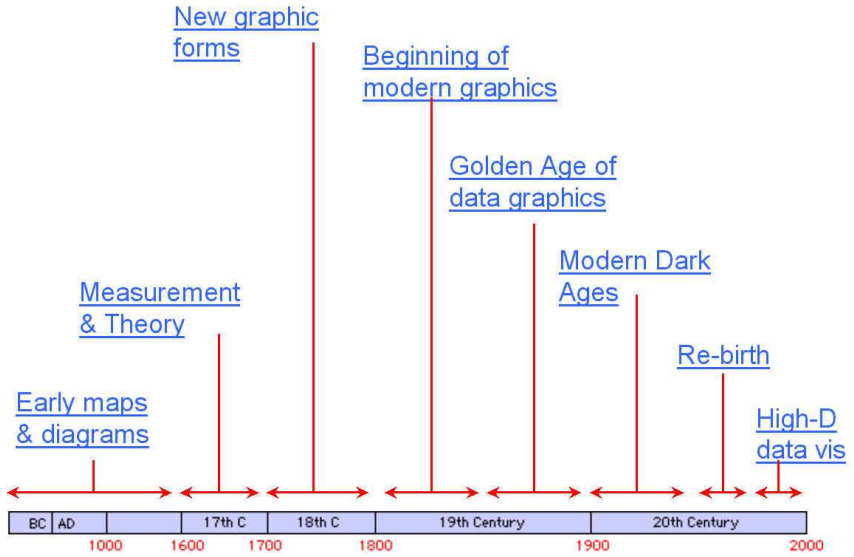
1924: Museum of Social Statistical Graphics- Otto Neurath

1991-1996: Interactive data visualization systems (Xgobi, ViSta)



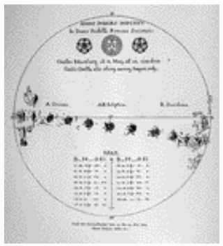


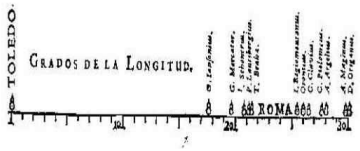
Milestones Tour



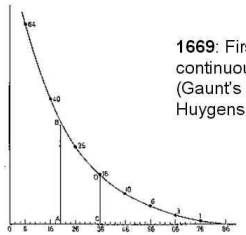
1600-1699: Measurement and Theory

1626: Visual representations used to chart the changes in sunspots over time- Christopher Scheiner

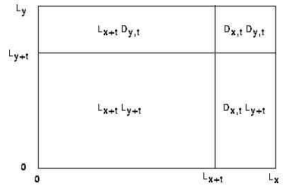




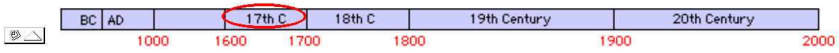
1644: First visual representation of statistical data- M.F. van Langren, Spain



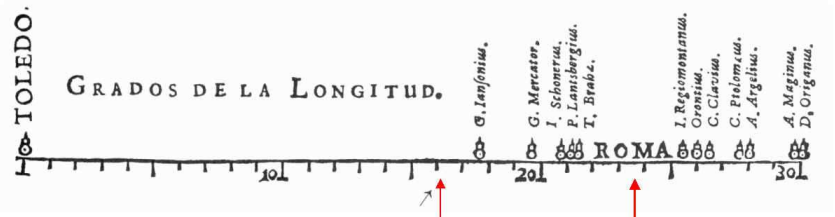
1669: First graph of a continuous distribution function (Gaunt's life table)– Christiaan Huygens.

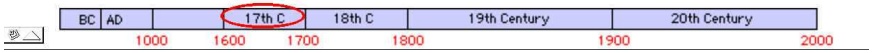


1693: First use of areas of rectangles to display probabilities of independent binary events- Edmund Halley, England



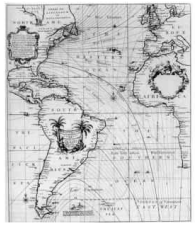
1644: First visual representation of statistical data: determination of longitude between Toledo and Rome- M. F. van Langren, Spain



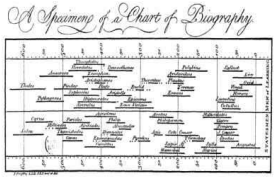


1700-1799: New graphic forms

1701: Isobar map, lines of equal magnetic declination – Edmund Halley

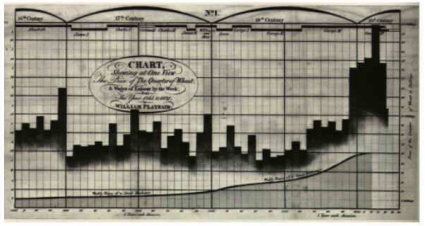


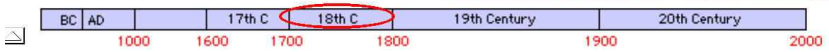
1765: Historical time line (life spans of famous people) Joseph Priestley



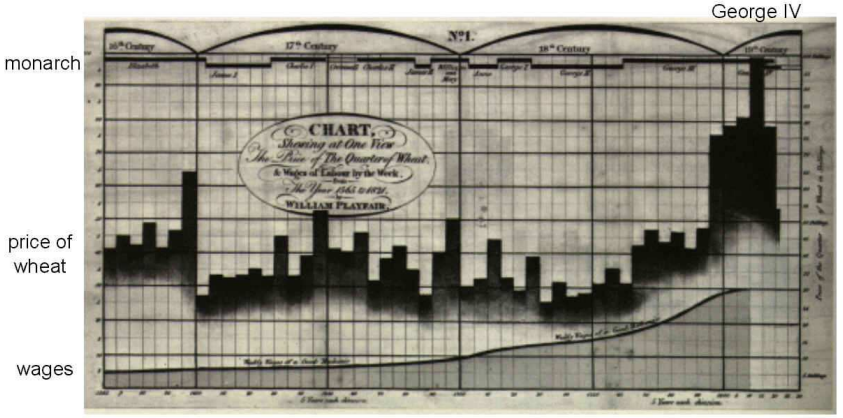
1782: First topographical map- Marcellin du Carla-Boniface

1786: Bar chart, line graphs of economic data- William Playfair



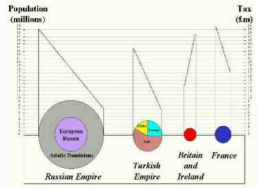


1786: Bar chart and line graph showing three time series: Price of wheat, weekly wages and reigning monarch over a 250+ year span- William Playfair

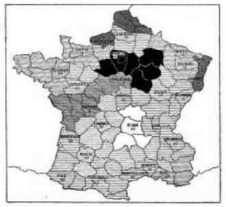


1800-1849: Beginning of modern data graphics

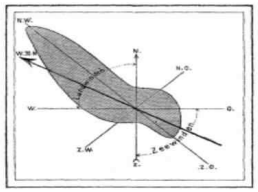
1801: Pie chart, circle graph invented- William Playfair

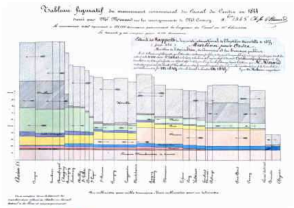


1819: First modern statistical map (illiteracy in France)- Charles Dupin

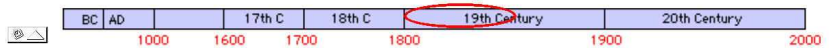


1843: Wind-rose (polar coordinates)- L. Lalanne





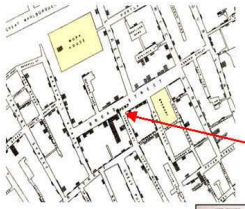
1844: variable-width, divided bars, area ~ cost of transport- C. J. Minard



Beginning of Modern Data Graphics: 1800–1849

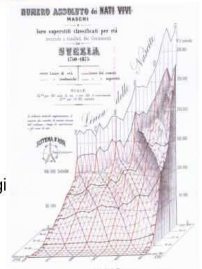
- Playfair's linear arithmetic (1780–1800): line plot, pie chart, etc.
- Adolphe Quetelet (1835) ,"average man" as central tendency in a normal curve.
- Moral, social and medical statistics collected systematically (1820–)
  - Dupin: distributions of years of schooling; prostitutes in Paris.

1850-1899: Golden Age

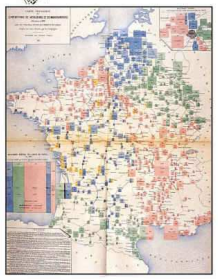


1855: Dot map of disease data (cholera)- John Snow

Broad St. pump

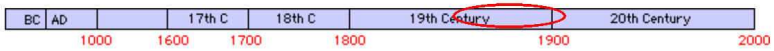


1879: Stereogram (3D population pyramid)- Luigi Perozzo



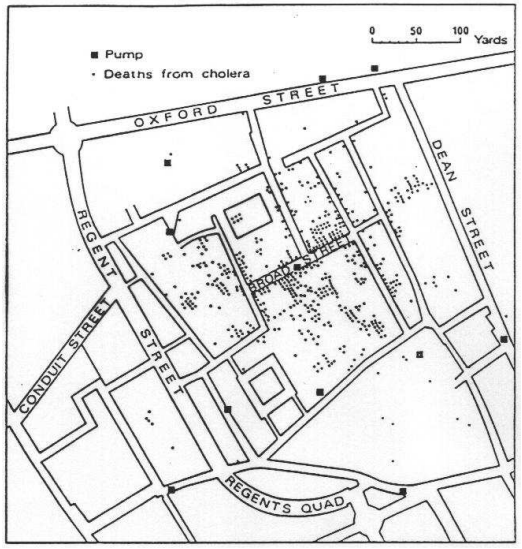
1884: Recursive multi-mosaic on a map- Emile Cheysson

1896: Area rectangles on a map to display two variables and their product- Jacques Bertillon



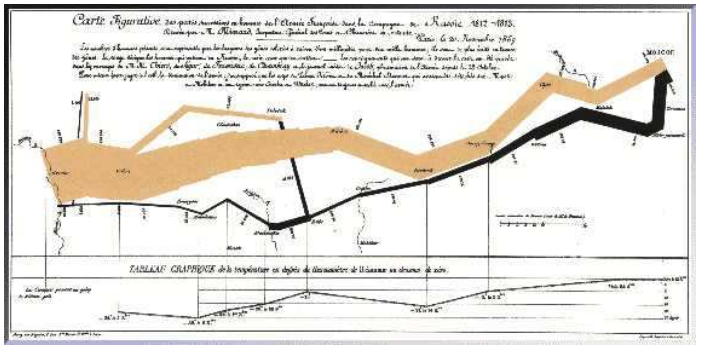
The Golden Age of Statistical Graphics

- Snow: map of cholera cases (Aug 31–Sep 8, 1854) → Broad Street pump.



- cf. Water in Walkerton: Outbreak of E. coli contamination (May 16–22, 2000) → 6 died, > 2000 ill.
  - Source: undetermined until Jan. 2001
  - No one thought to make a map!

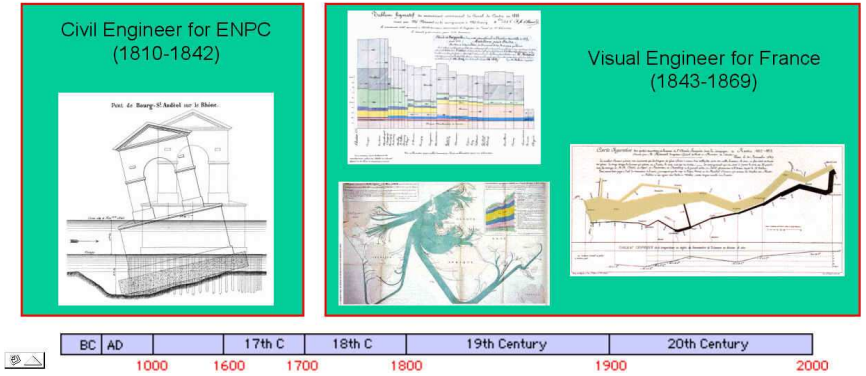
"The Best Statistical Graphic Ever Produced"



- E-J Marey (1878): "defies the pen of the historian by its brutal eloquence".
- Funkhouser (1937): Minard, the Playfair of France.
- Tufte (1983): "multivariate complexity integrated so gently that viewers are hardly aware that they are looking into a world of six dimensions ... the best statistical graphic ever produced."

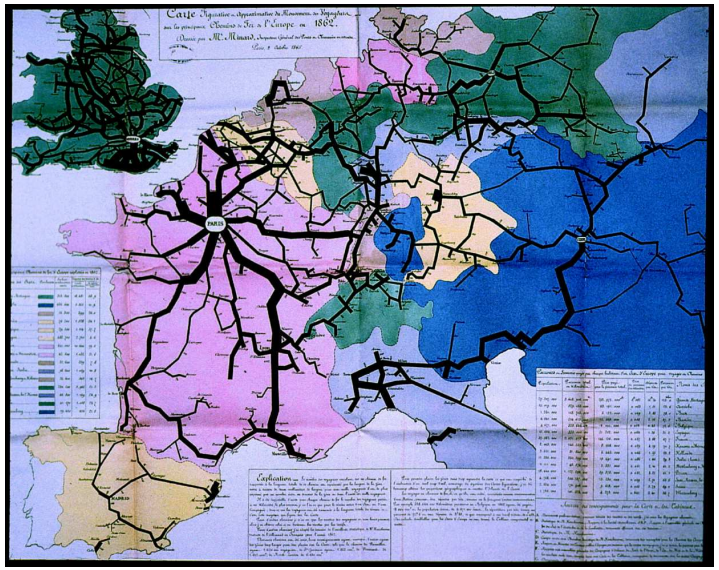
Why Minard?

- Study breadth and depth of his work
  - How related to work in his time?
  - How related to modern statistical graphics?
  - How related to his personal history?



Flow maps as visual tools

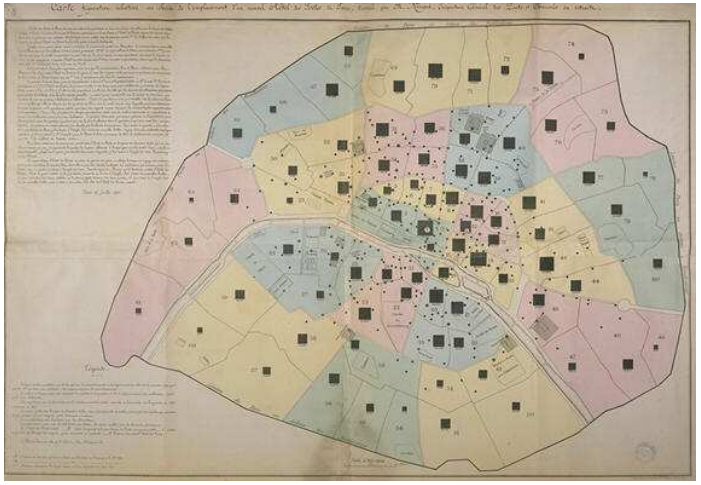
- Movement of people and goods was a consistent theme of most of Minard's work
- Data represented *both* visually and numerically
- Extensive legends, describing how the information should be understood and interpreted
- Visual engineer for France: the dawn of globalization, emergence of the modern French state.



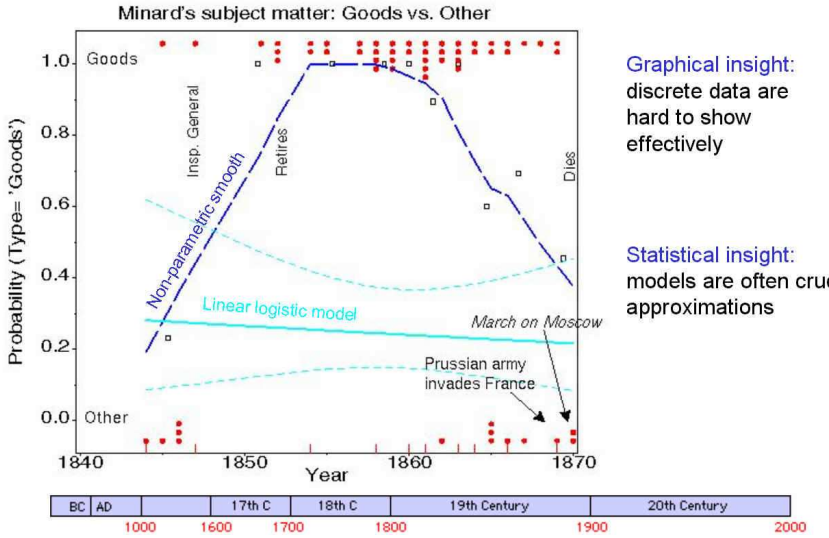
Carte figurative et approximative du mouvement des voyageurs sur les principal chemin de fer de l'Europe en 1862 (1865) [ENPC: 5862/C351]

Minard's graphic inventions

- Population represented by squares, area $\sim$ population
- Visual center of gravity used to choose location for new post office



Minard's themes: Goods vs. Other

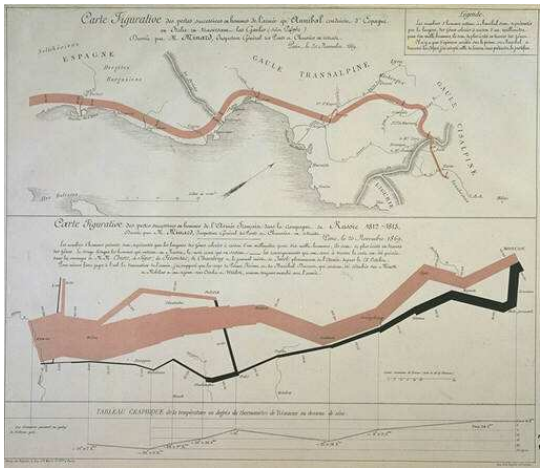


Graphical insight: discrete data are hard to show effectively

Statistical insight: models are often crude approximations

The March Re-visited

- March on Moscow was part of a pair, along with Hannibal's campaign



- Aug. 1869: Prussian army invades, Minard flees to Bordeau
- Personal meaning: horrors of war, the human cost of thirst for military glory.

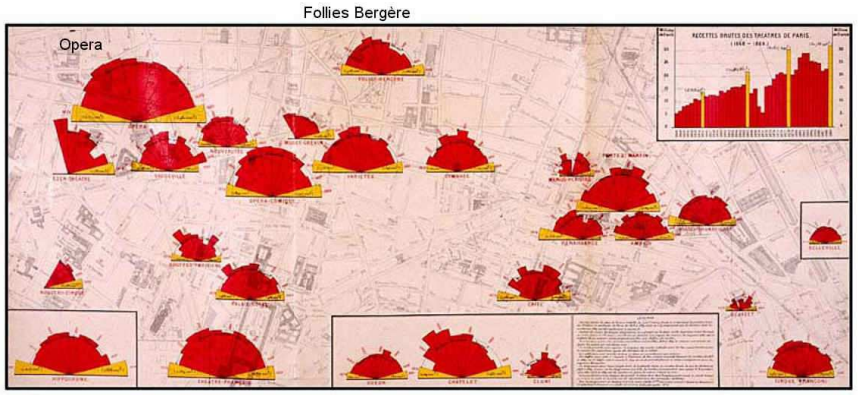
Why the Golden Age?

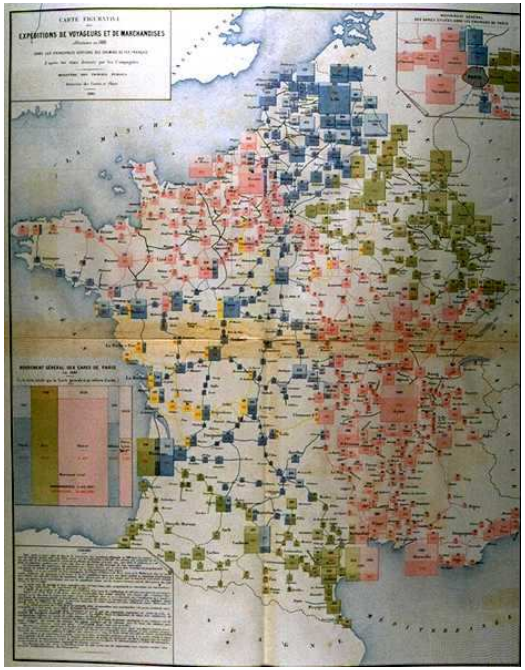
- Statistics as a discipline:
  - 1st International Statistics Congress (1853) [Quetelet]
  - 3rd ISC: Expo. & standardization of graphical methods (Vienna, 1857)
  - la Société de statistique de Paris (1860)
  - Royal Statistical Society (1860)
- Expansion of industrialization, trade, transport ⟶ government initiatives in data collection and analysis.
- Statistics: Numbers of the State
  - Ministry of Public Works (France): Statistical Bureau (Émile Chasson)
  - Similar efforts in Germany, Switzerland, etc.
  - U.S. Census Bureau (Edward Walker)— first US census (1860)

L'Album de Statistique Graphique

- The pinnacle of the Golden Age of Graphics
- 18 volumes published 1879–1899
- Les Chevaliers des Album

1889: Gross receipts in theaters in Paris, 1848-1889

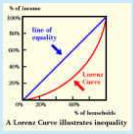




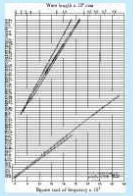
1900-1949: Modern Dark Ages of Statistical Graphics

Few innovations; by the mid-1930s, enthusiasm for visualization of the late 1800s had been supplanted by the rise of quantification and statistical models. But graphical methods entered the mainstream, were applied, and popularized.

1905: Lorenz curve (cumulative distribution by rank order)- M. O. Lorenz



1913: Discovery of atomic number, based on graphical analysis- H. Moseley



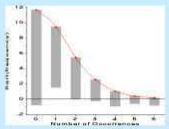
1924: Museum of Social Statistical Graphics, and the Isotype method- Otto Neurath

1944: Harvard's Mark I, the first digital computer- Howard Aiken, Grace Hopper

1950-1974: Re-birth of Modern Statistical Graphics

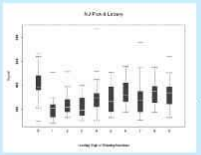
Visualization began to rise from dormancy in the mid 1960s, spurred largely by: 1) Tukey's *Exploratory Data Analysis*, 2) Bertin's *Semiologie Graphique*, and 3) the advent of graphics software.

1965: Beginnings of EDA: improvements on histogram (hanging rootogram)- John W. Tukey

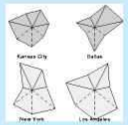


1967: Comprehensive theory of graphical symbols and modes of graphics representation- Jacques Bertin

1967: Reorderable semi-graphic matrix- Jacques Bertin

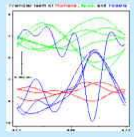


1969: Graphical innovations for EDA (stem-and-leaf, box-plots, etc.)- Tukey



1971: Star plot to represent multivariate data- J. H. Siegel etal.

1972: Fourier series plots of multivariate data- David F. Andrews



1973: Cartoons of human face to represent multivariate data- Herman Chernoff,

Visions of the Present

Look not mournfully into the past. It comes not back again. Wisely improve the present. It is thine. Henry Wadsworth Longfellow

- Graphical methods for categorical data
  - Fourfold displays
  - Mosaic displays
  - Diagnostic plots for GLIMs
- Graphical principles: Rendering and effect ordering
  - Corrgrams
  - Effect ordering for data display
- Other innovations
  - JMP— Graphs as first-place objects; graphic scripting
  - VISTA— dynamic graphics (spreadplots), workmaps
  - `ggobi`→R— interconnectivity
  - Graphical excellence: e.g., linked micromaps (Dan Carr)
    - God is in the details
  - NVIZN— *Grammar of Graphics*→JAVA

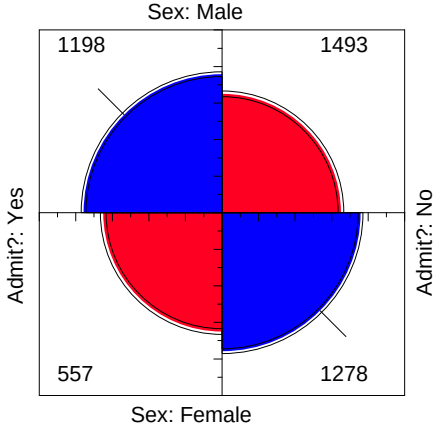
Graphical methods for categorical data

- *Visualizing Categorical Data* (Friendly, 2000)
  - Goals:
    - Develop graphical methods comparable to those used for quantitative data
    - Make them *available* and *accessible* in SAS Software
  - Visualizing odds ratios— Fourfold displays
  - Visual fitting for loglinear models— Mosaic displays
  - Visualizing model diagnostics for GLIMs— Influence plots
  - Multi-variable overviews— Mosaic matrices

See: `http://www.math.yorku.ca/SCS/vcd/`

Fourfold displays for 2 × 2 tables

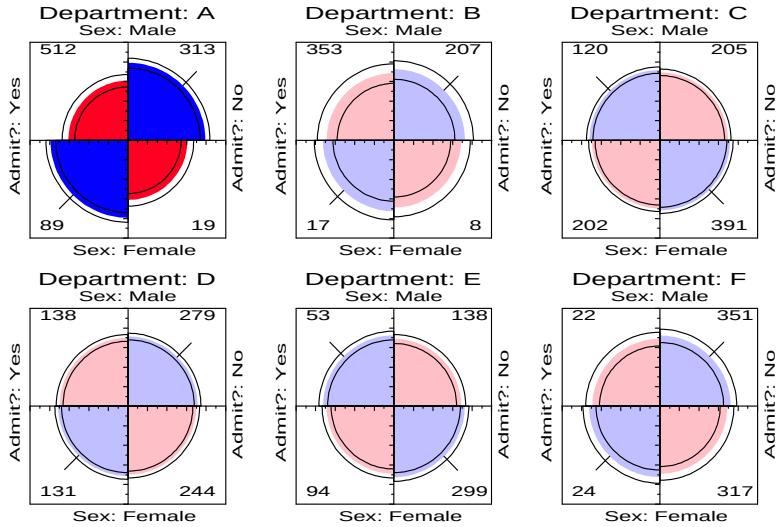
- ***Quarter circles***: radius $\sim \sqrt{n_{ij}} \Rightarrow$ **area** $\sim$ **frequency**
- ***Independence***: Adjoining quadrants $\approx$ align
- ***Odds ratio***: ratio of areas of diagonally opposite cells
- ***Confidence rings***: Visual test of $H_0 : \theta = 1 \leftrightarrow$ adjoining rings overlap



- Confidence rings do not overlap: $\theta \neq 1$

Fourfold displays for 2 × 2 × *k* tables

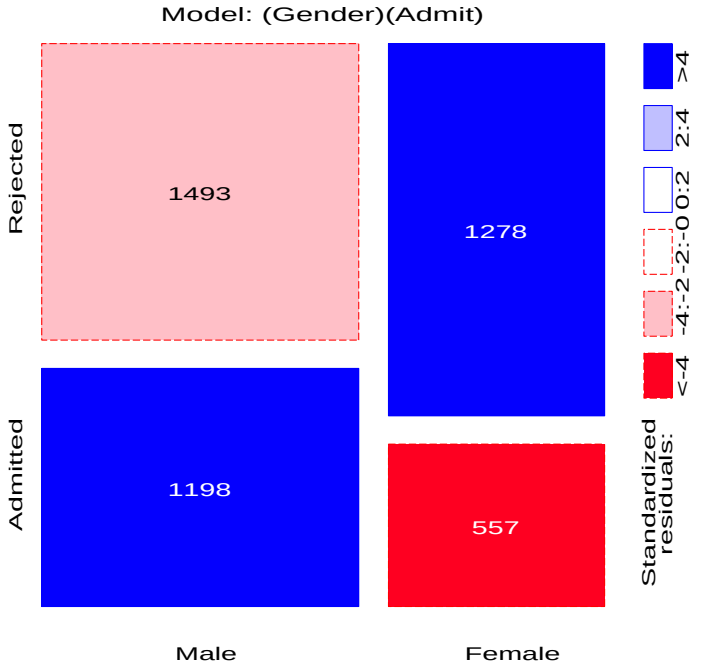
- Data had been pooled over departments
- Stratified analysis: one fourfold display for each department
- Each 2×2 table standardized to equate marginal frequencies
- Shading: highlight departments for which $H_a : \theta_i \neq 1$



- Only one department (A) shows association; $\theta_A = 0.349 \rightarrow$ women $(0.349)^{-1} = 2.86$ times as likely as men to be admitted.

Mosaic displays

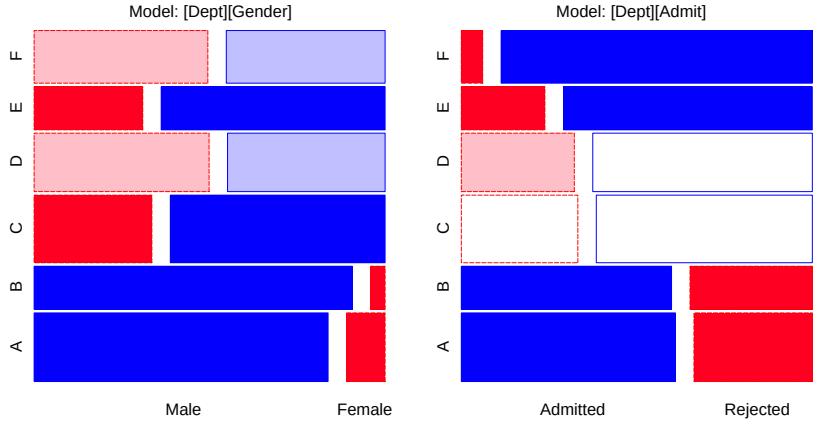
- ***Width*** $\sim$ one set of marginals
- ***Height*** $\sim$ relative proportions of other variable
- $\Rightarrow$ **area** $\sim$ **frequency**
- ***Shading***: Sign and magnitude of Pearson χ^2 residual, $d_{ij} = (n_{ij} - \hat{m}_{ij})/\sqrt{\hat{m}_{ij}}$ (or L.R. G^2)
  - Sign: $-$ negative in red; $+$ positive in blue
  - Magnitude: intensity of shading: $|d_{ij}| > 0, 2, 4, \ldots$
- ***Independence***: Rows $\approx$ align, *or* cells are empty!
- E.g., aggregate data:



Mosaic displays— Other two-way views

Department $\times$ Gender, Department $\times$ Admit

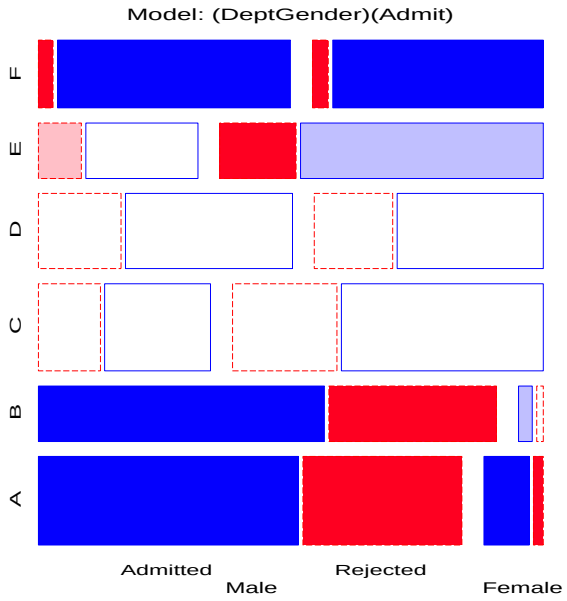
- Did men and women apply differentially to departments?
- Did departments differ in overall rate of admission?



Mosaic displays for multiway tables

- Generalizes to n-way tables: divide cells recursively
- Can fit *any* log-linear model, e.g. (3-way),
  - Mutual independence, $[A][B][C] \leftrightarrow A \perp B \perp C$
  - Joint independence, e.g., $[AB][C] \leftrightarrow (A,\ B) \perp C$
  - Conditional independence, e.g., $[AC][BC] \leftrightarrow (A \perp B)\,|\,C$
- Shows:
  - **DATA** (size of tiles)
  - (some) **marginal** frequencies (spacing $\rightarrow$ visual grouping)
  - **RESIDUALS** (shading)

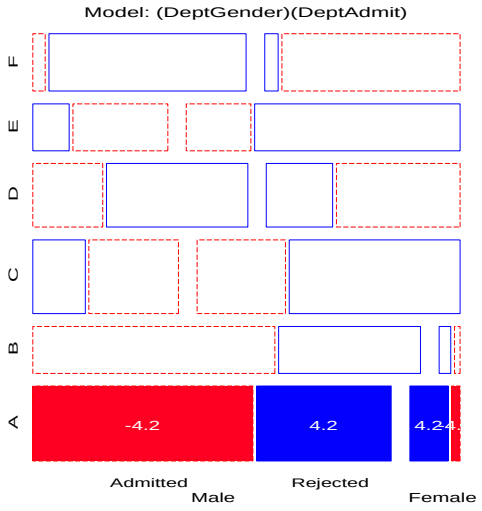
- E.g., Joint independence (null model, Admit as response) [$G^2_{(11)}$ = 877.1]:



Mosaic displays for multiway tables

- Typical numerical methods for loglinear models:
  - Fit model → remove NS terms → "better" model— NS increase in G^2
  - Backward elimination: let the computer do your thinking!
- Mosaics → Visual fitting:
  - Pattern of lack-of-fit (residuals) → "better" model— smaller residuals
  - "cleaning the mosaic" → "better" model— empty cells
  - best done interactively

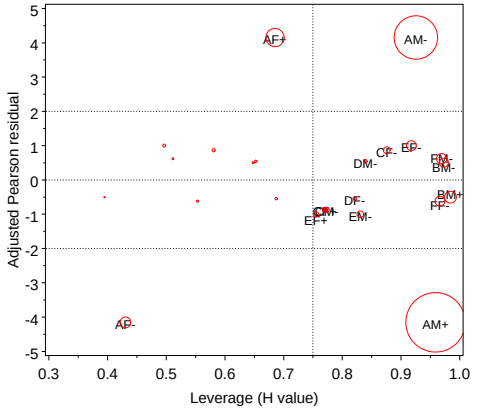
- E.g., Add [Dept Admit] association → Conditional independence:



- Fits poorly overall ($G^2_{(6)} = 21.74$)
- But, only in Department A!

Diagnostic plots for GLIMs

- Diagnostic displays for categorical data ≈ those for regression, GLMs.
- `INFLGLIM` macro: GENMOD→ Influence plots bubble plot of residual vs. Hat value, area ∼ Cook's D.
  - Model $[AD][GD] \longleftrightarrow$ logit model $L_{ij} = \alpha + \beta_i^{\text{Dept}}$

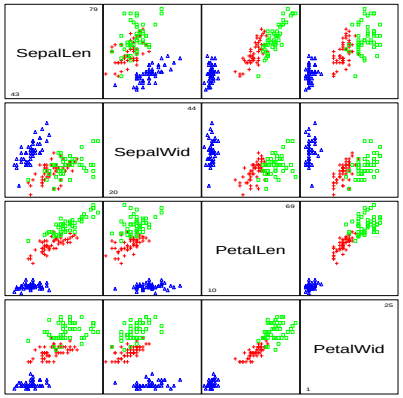


- All cells which do not fit ($|r_i| > 2$) are for department A.

Mosaic matrices

Quantitative data: scatterplot matrix shows $p \times (p-1)$ marginal views in a coherent display;

- Each scatterplot a projection of data
- Detect patterns not easily seen in separate graphs.
- Only shows bivariate relations.



Mosaic matrices

Categorical data: Mosaic matrix shows all $p \times (p-1)$ marginal views

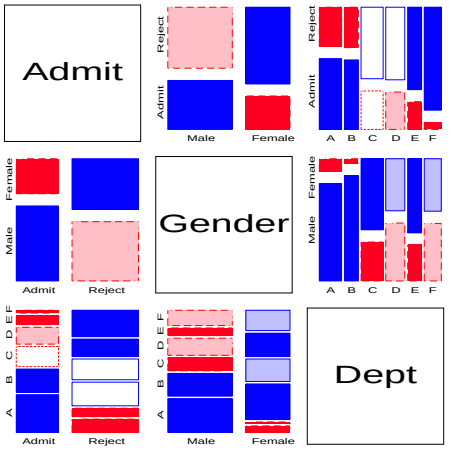
- Each mosaic shows bivariate relation
- Fit: bivariate independence
- Direct visualization of the "Burt" matrix analyzed in MCA to account for all pairwise associations among p variables

$$\boldsymbol{B} = \boldsymbol{Z}^{\mathsf{T}} \mathrm{diag}(\boldsymbol{n}) \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{N}_{[1]} & \boldsymbol{N}_{[12]} & \cdots \\ \boldsymbol{N}_{[21]} & \boldsymbol{N}_{[2]} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

where $\boldsymbol{N}_{[i]}$ = diagonal matrix of one-way margin; $\boldsymbol{N}_{[ij]}$ = two-way margin for variables i and j,

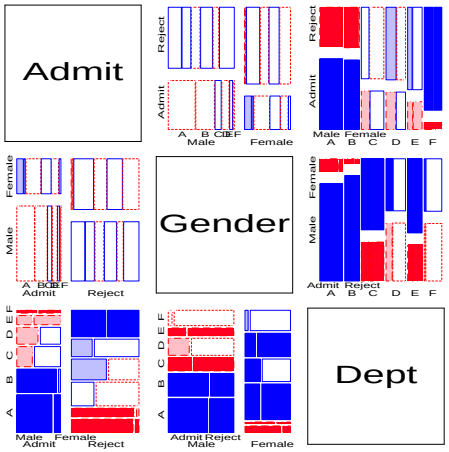
Mosaic matrices: Berkeley admissions

- Admission, Gender: overall, more males admitted
- Dept A, B: highest admission rate; E, F lowest
- Males apply most to A, B, women more to C–F.



Conditional mosaic matrices

- Show 3-way conditional relations, fitting conditional independence, $[AC][BC]$ for each A, B.
- $\Rightarrow$ Admission $\perp$ Gender $\mid$ Dept. (except for Dept. A)



"Mixed" models: Categorical and Continuous Data

- **Marginal views**
  - X, Y pairs: scatterplot
  - A, B pairs: mosaic
  - X, A pairs: boxplot
- **Conditional views**
  - Fit graphical mixed model: $AB \,//\, XY$ (Edwards, 1995)
  - Fit GLMs:

$$g(\mu_i) = x_{\text{others}}^{\mathsf{T}}\beta$$
$$g(\mu_j) = x_{\text{others}}^{\mathsf{T}}\beta$$

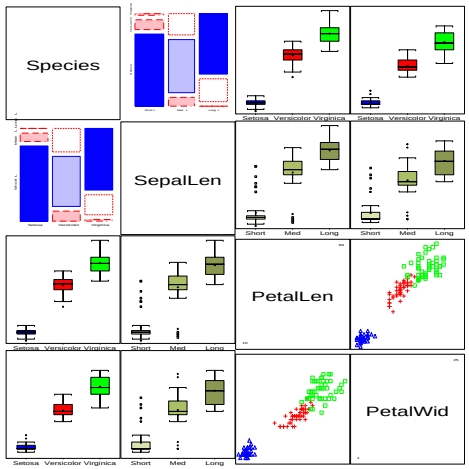
with identity link for X, Y, log link for A, B

  - Plot residuals as in marginal views

"Mixed" models: Categorical and Continuous Data

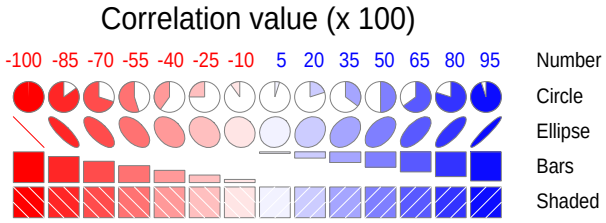
Iris data — Mixed scatterplot matrix

- Discrete: Species, SepalLen (divided into thirds)
- Continuous: PetalLen, PetalWid



Corrgrams— Correlation matrix displays

- Render a correlation to depict sign and magnitude (tasks: lookup, comparison, detection)

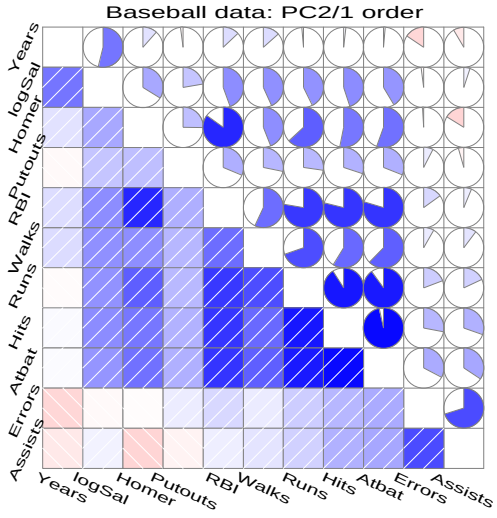


Task-specific renderings:

Task	Lookup	Comparison	Detection
Rendering	Number	Circle	Shading

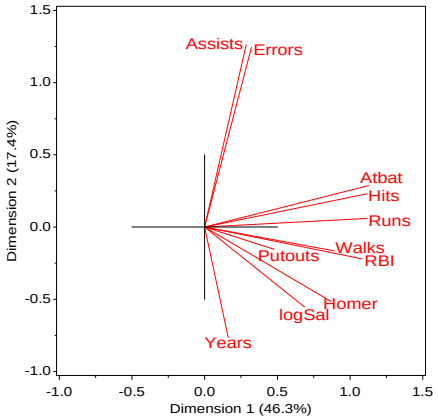
Corrgrams— Rendering

Baseball data: (lower) Patterns vs. (upper) comparison



Corrgrams— Variable ordering

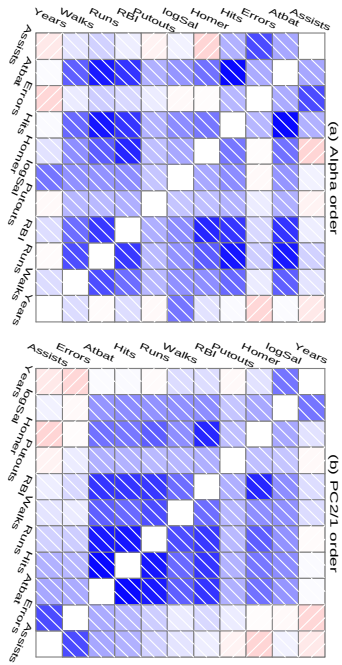
- Reorder variables to show similarities: PC1 or angles (PC2/PC1)



- Generalizations to partial ($R(\boldsymbol{Y} \mid \boldsymbol{X})$), conditional correlations ($r_{ij \mid \text{rest}} \sim \boldsymbol{R}^{-1}$)

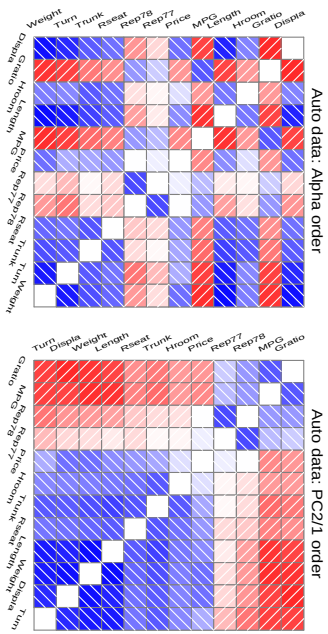
Corrgrams— Correlation matrix displays

Baseball data: (a) alpha vs. (b) correlation ordering



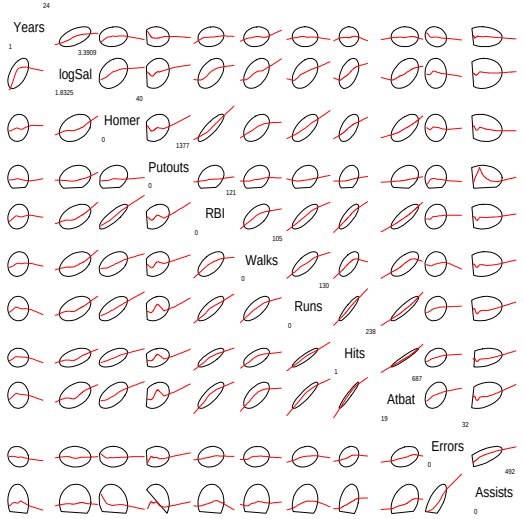
Corrgrams— Auto data

- Correlation ordering shows a coherent pattern
- Size variables positively correlated
- Gratio, MPG, repair record positively correlated
- Negative correlations between the two sets



Corrgrams— Other renderings

Baseball data: schematic scatterplot matrix: 68% data ellipse + loess smooth



- Different renderings for look-up, comparison, detection of patterns, anomalies!

Effect ordering for data displays

- **Information presentation is *always* ordered**—
  - in ***time, or sequence*** (a talk, a written paper),
  - in ***space*** (a table, or graph)
  - Constraints of time and space are dominant— can conceal or reveal the important message.
- **Effect ordering for data display** (Friendly and Kwan, 2003)

  Sort the data by the effects to be seen
- Applies to:
  - unordered factors for quantitative data
  - categories of variables in frequency tables
  - arrangement of observations and variables in multivariate displays

Effect ordering for data displays

- Multiway quantitative data
  - Main effects ordering— sort unordered factors by means/medians
- Multiway frequency data
  - Association ordering— sort by CA Dim 1 (SVD of residuals from independence)
- Multivariate displays
  - Correlation ordering for variables
  - Clustering/sorting for observations

Effect ordering for frequency tables

Table 1: Hair color - Eye color data: Alpha ordered

Eye color	Hair color			
	Blond	Black	Brown	Red
Blue	94	20	17	84
Brown	7	68	26	119
Green	10	15	14	54
Hazel	16	5	14	29

Table 2: Hair color - Eye color data: Effect ordered

Eye color	Hair color			
	Black	Brown	Red	Blond
Brown	68	119	26	7
Hazel	15	54	14	10
Green	5	29	14	16
Blue	20	84	17	94

Model:	*Independence*: [Hair][Eye] χ^2 (9)= 138.29						
Color coding:	<-4	<-2	<-1	0	>1	>2	>4
n in each cell:	$n <$ expected				$n >$ expected		

Visions of the Future

Prediction is very difficult, especially about the future Niels Bohr

The best way to predict the future is to invent it Alan Kay

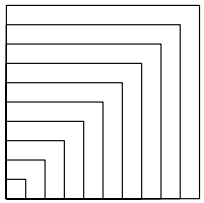
- Statistical graphics: Models for growth?
  - Different strokes: graphics *user* vs. *developer*
  - Minard's lessons for statistical graphics
  - JMP— Model summary = Graphs + Numbers
  - ViSta— Dynamic, interactive graphics (spreadplots, workmaps)
  - Innovation and Graphical excellence
- Wider visions
  - Visions from the Forrest
  - Visions for graphic users and developers
- Conclusions

Minard's lessons for statistical graphics

- What can we learn from the process of *programming* to duplicate Minard's March?
- Elegance factors: Power and expressiveness—
  - Simplicity, transparency of data representation
  - Simplicity, transparency of procedural representation
- Turtle graphics: *Logo: A Language for Learning* (Friendly, 1988)
  - Concise and transparent
  - Specification (program statements) tightly linked to display
  - Thinking ⟷ doing ⟷ seeing

```
TO SQUARE :SIZE
   REPEAT 4 [FORWARD :SIZE
             RIGHT 90]
   END
```

```
TO GROW.SQUARE :SIZE
   IF :SIZE > 100 [STOP]
   SQUARE :SIZE
   GROW.SQUARE (:SIZE+10)
   END
```

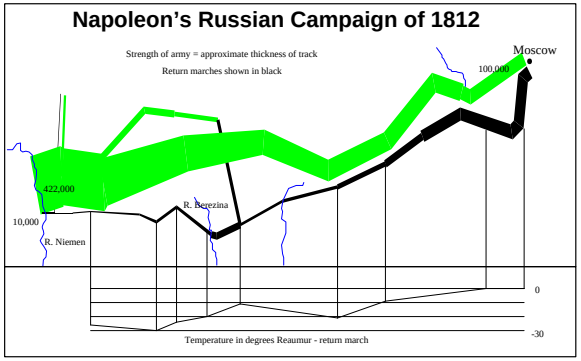


Mathematica

Mathematica:

- list processing,
- recursion,
- modularity,
- function mapping

Shaw and Tigg (1994): `NapoleonicMarchOnMoscowAndBackAgainPlot[]`



- Data structure: Nested lists– (x, y) coordinates, troop strength, temperature, rivers, etc.

```
StrengthData = {
  { {0.142, 0.238, 50000}, {0.257, 0.331, 50000},
    {0.312, 0.326, 50000}, {0.312, 0.326, 33000},
    {0.392, 0.318, 33000} },
  { {0.056, 0.230, 422000}, {0.105, 0.242, 422000},
    {0.105, 0.242, 400000}, {0.181, 0.234, 400000},
    {0.181, 0.234, 340000}, {0.333, 0.273, 257000}, ...}, ...
   };

TempData = {
  {955, 306, 0},   {885, 304, 0},   {700, 259, -9},
  {612, 228, -21}, {433, 177, -11}, {372, 170, -20},
  {316, 201, -24}, {279, 181, -30}, {158, 195, -26}};
```


Mathematica

- Procedural structure:
  - Nested functions:

```
NapoleonicMarchOnMoscowAndBackAgainPlot[] :=
  Show[Graphics[
   {ProcessStrength[StrengthData],
    ProcessTemp[TempData],
    ProcessRivers[RiverData],
    ProcessBoxes[BoxData],
    ProcessTitle[TitleData],
    ProcessPoints[PointData],
    ProcessText[TextData]}
  ]]
```

  ⇒ *[list of Graphics instructions]* → `Graphics[]` → `Show[]`
  - Function mapping: Applying a function pattern over a list

```
ProcessRivers[riverdata_] :=
  Map[(RGBColor[0, 0, 1], Thickness[0.001], Line[#]&),
      riverdata]
```

A Grammar for Graphics

Wilkinson (1999) - grammar for representing:

- data (variables, attributes, transformations)
- graph elements (coordinates, frames, scales, guides)
- specification: declarative, not procedural (Java: GPL)

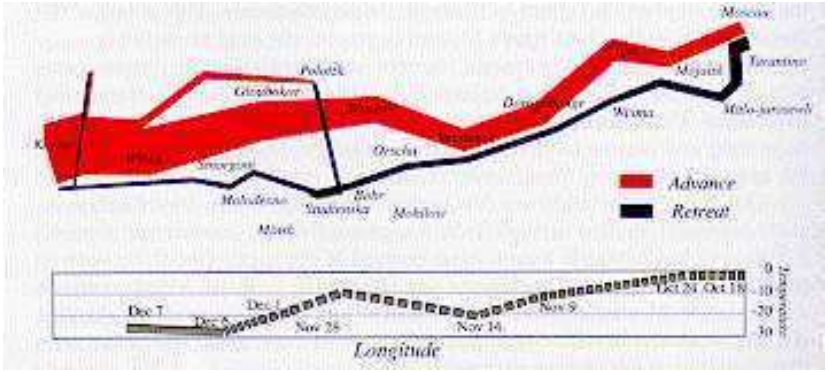
⇒ Two sub-graphics (march and temperature), linked by common horizontal scale of longitude.

- The March graphic

```
FRAME: lonc*latc
GRAPH: point(label(city), size(0))
GRAPH: path(position(lonp*latp), size(survivors),
       color(direction), split(group))
GUIDE: legend(color(direction))
```

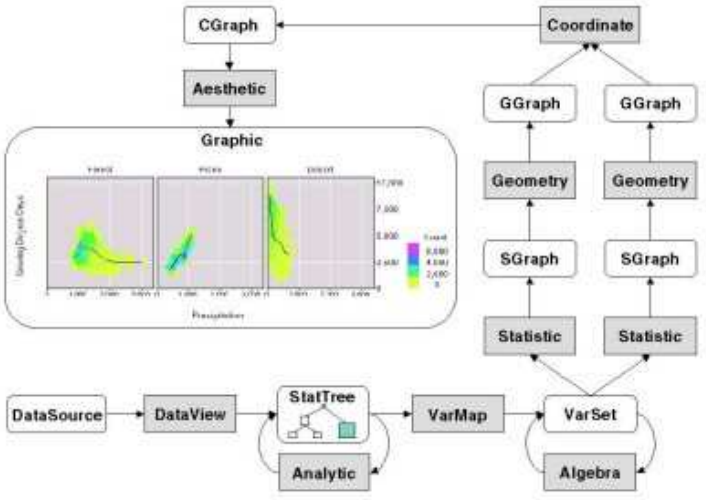
- Temperature graphic

```
TRANS: ldate = lag(date,-1)
TRANS: days = diff( date, ldate )
FRAME: lonc*latc
GRAPH: point(label(city), size(0))
GRAPH: path(position(lonp*temp), label(date),
       texture.granualarity(days))
GUIDE: legend(color(direction))
```

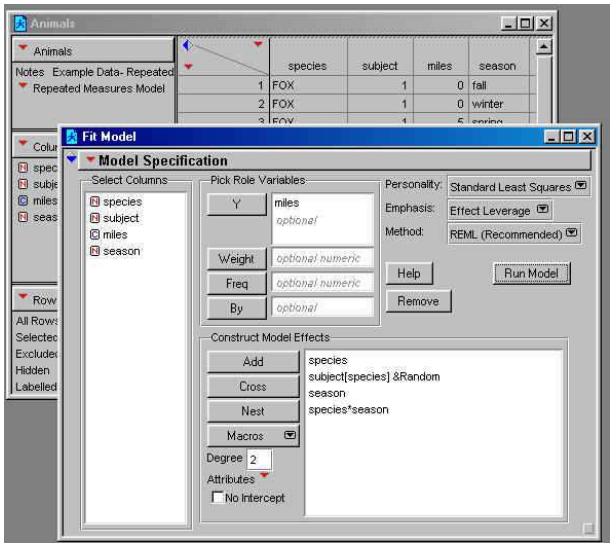
nViZn— Grammar of Graphics

- Java implementation of *Grammar of Graphics* framework
  - DataView: Abstraction of dataset; multiple input sources.
  - StatTree: Data objects + analysis nodes (filter, recode, summarize, etc.)
  - Graph algebra: Frame + operators (*cross, nest, blend*) on variable subset (*VarSet*) ⟶ statistical graph (*Sgraph*)
  - Coordinate transforms (log, polar, etc.) + rendering methods (Aesthetics) ⟶ Graphic



See: `http://www.illumitek.com`,
`http://www.spss.com/research/wilkinson/nViZn/nvizn.html`

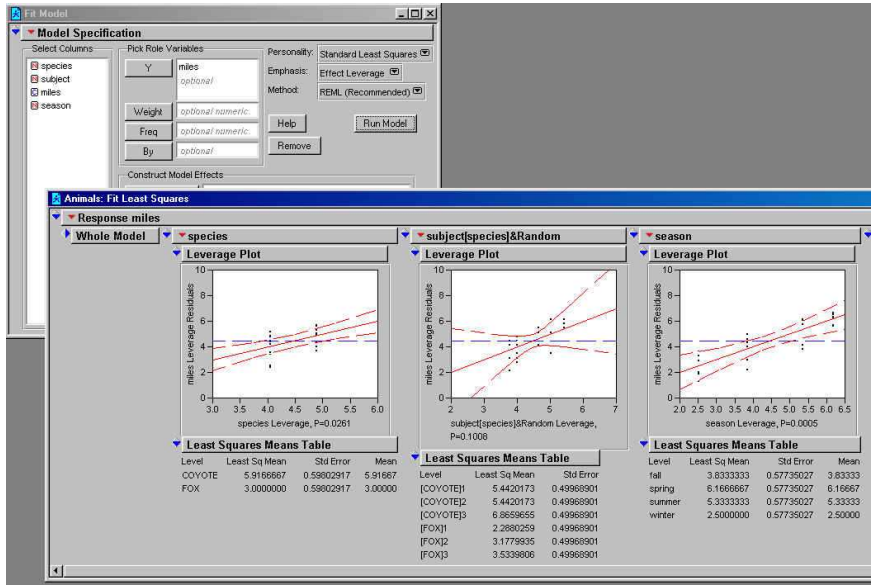
JMP— Model summary = graphs + numbers

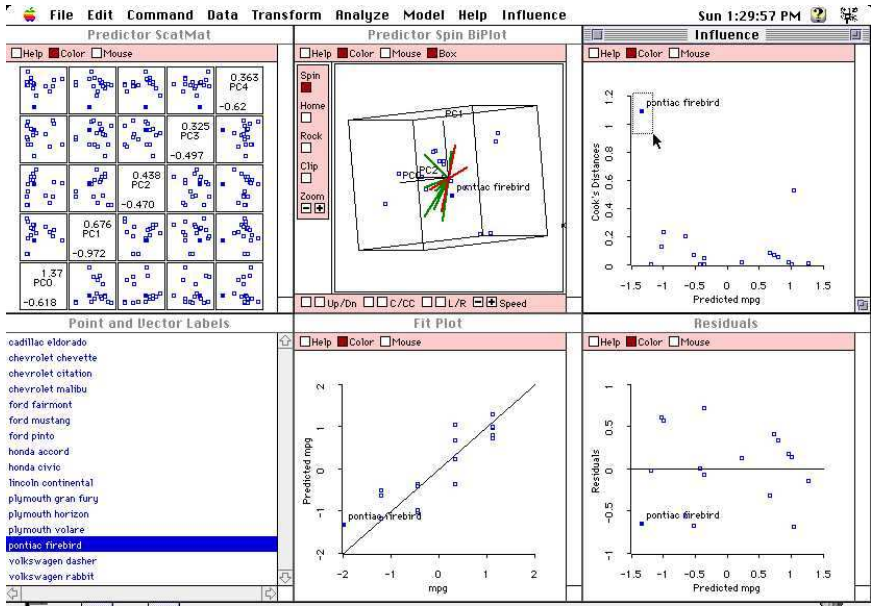


ViSta— spreadplots, work maps

- Spreadplots
  - Graphic equivalent of a spreadsheet
  - Dynamically linked views of *data* and *model* objects
  - Highly interactive: every action → data, model, plots
  - (Message passing architecture)
- e.g., Spreadplot for multiple regression
  - Scatterplot matrix— overview
  - 3D spin predictor biplot— leverage, collinearity
  - Influence plot, fit plot, residual plot— influential cases
  - Observation, variable labels, interactive brushing, etc.

See: `http://forrest.psych.unc.edu/research/`

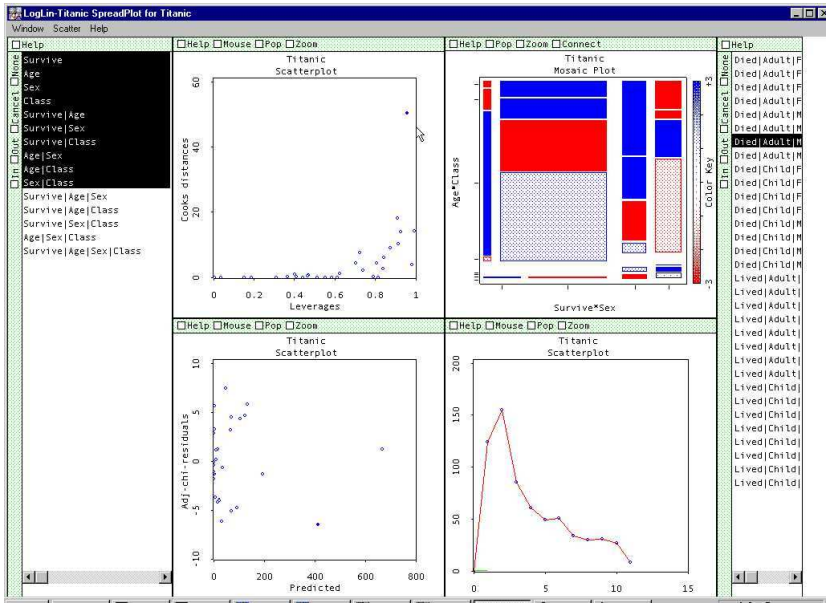




ViSta— Categorical data

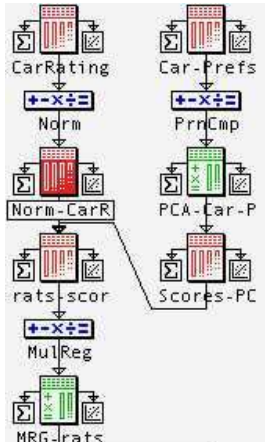
- Visual model fitting— select terms
- Mosaic display for current model
- Influence plot: Cook's D vs. Leverage (Hat values)
- Model summary graph: Deviance vs. df
- All dynamically linked, manipulable!

See: Valero et al. (2003),
`http://www.math.yorku.ca/SCS/Papers/viscat.pdf`



ViSta— Workmaps

- Workmap— visual GUI for path(s) of analysis
- Each item: dynamic links to table-view, numerical summary, spreadplot visualization



ViSta— Expandability

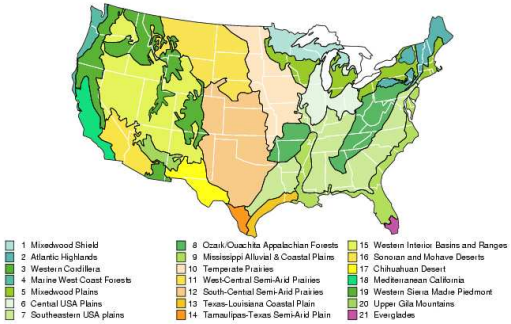
- Other features:
  - Plugins — add new analysis and visualizations
  - Web Applets, Scripts
  - Data analysis language

See: `http://forrest.psych.unc.edu/research/`

Innovation and Graphical Excellence

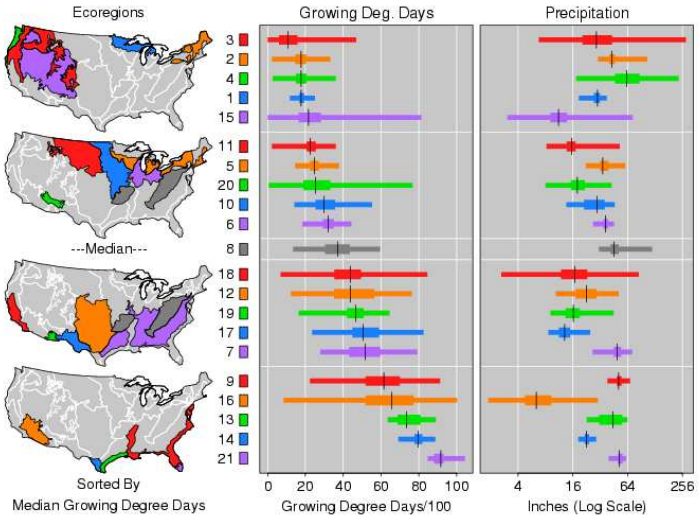
e.g., Dan Carr (Carr et al., 1998)

- Omernick ecoregions - ecological distinctive areas



- Linking regions with labels is difficult
- Hard to use distinct colors
- How to show spatial variation of analysis variables?

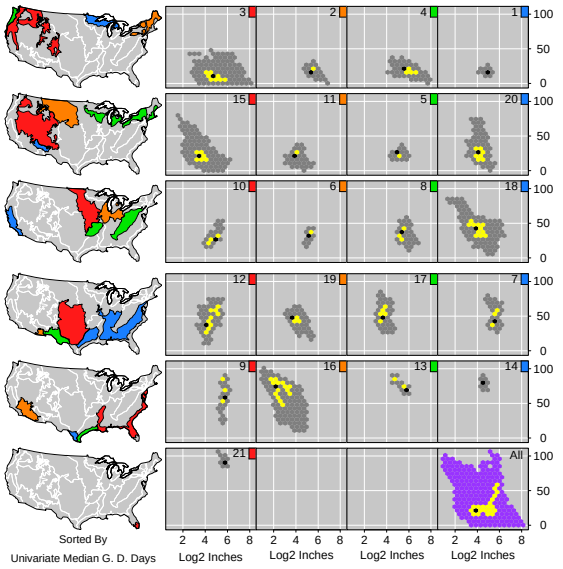
- ⟶ Linked micromaps
  - Boxplots of growing degree days & precipitation
  - Effect ordering: sorted by median growing degree days
  - Color linking is clear; attention to detail exemplary



Innovation and Graphical Excellence

- Relationship of growing days and precipitation hard to see in univariate views.
- Bivariate density estimation (481K grid cells)
- Bivariate boxplots (50% high-density region, bivariate median)
- Sorted by median growing degree days

Figure 2: LM Bivariate Boxplots
1961-1990 Precipitation (x) versus Growing Degree Days/100 (y)



Visions from the Forrest

The Statistician's 3D Virtual-Reality Workroom

- A 3D, VR statistical analysis environment:
  - Data sources, data streams, data views
  - Tools (and a glove?) for manipulating data
  - Analysis and visualization devices
  - An amenuensis— virtual assistant
- Data sources, data streams, data views
  - Visual, manipulable building blocks (lego?)
  - Snap together to form statistical objects (tables, datasets)
  - Spigots for incoming streams, trapdoors to the data mine, hoses, valves, connectors...
  - Lassos and windows for data views
- Tools for manipulating data
  - transformations,
  - subset, merge, join, ...
  - $\longrightarrow$ new data objects, views, ...

Visions from the Forrest

The Statistician's 3D Virtual-Reality Workroom

- Analysis and visualization devices
  - Data toasters: data $\longrightarrow$ toast (model summary) $+$ crumbs (residuals)— all plug 'n play
  - Data/Model/Residual VCR's, with controls: pop in the data, out comes a visualization.
  - Recepticles for making new connections, plugging in new appliances
  - Hand-held devices— controls to interact with transformations, models, summaries, residuals, ...
  - Workmaps to show you where you've been, Guidemaps to show you where you might want to go
- An amenuensis— virtual assistant
  - take notes,
  - offer guidance,
  - suggest visualizations,
  - summarize results,
  - write results section,
  - serve virtual coffee, ...

The Future for Graphics Users

- Statistical procedures extensively developed— will continue
  - regression $\longrightarrow$ GLM $\longrightarrow$ GLIM $\longrightarrow$ HLM, GAM
  - PCA $\longrightarrow$ FA $\longrightarrow$ Lisrel, SEM
- Need to simplify the environment— for most users
- 80–20 rule: 80% of a graph takes 20% of effort. The last 20% is hard work.
- Statistical graphics is on the right track when ...
  - it allows you to picture what your data have to say
  - the picture is faithful to some (possibly complex) model
  - the picture leverages the perceptual and cognitve capabilities of the viewer.

The Future for Graphics Developers

- Statistical graphics now well-developed, but many different systems— mostly incompatible, different capabilities
  - SAS $\longrightarrow$ macros, SAS/INSIGHT, ...
  - R/S-Plus $\longrightarrow$ general `plot()` methods, packages, connections to interactive graphics (`ggobi`)
- Need to provide paths of growth for new visualizations, methods of interaction, ...
- 80–20 rule: 80% of software development takes 20% of effort. The last 20% is hard work.
- Statistical graphics is on the right track when ...
  - it allows one to develop a new method of visualization or interaction with ease
  - it provides elegant connections between statistical analysis (summarization) and visualization (exposure)
  - it leverages the capabilities of different software systems

Conclusions

- The past history of statistical graphics teaches us that:
  - Statistical graphics can have both *beauty* and *truth*
  - Statistical graphics had a purpose— tell a story, inform a decision, ...
  - Statistical graphics was hard work.
- The present history of statistical graphics teaches us that:
  - We need graphical methods for categorical data on a par with those for quantitative data.
  - Languages for graphics development differ in *power* and *simplicity of expression*: Thinking → doing → seeing.
  - Users— Different strokes for different folks:
    - Most want *graphical toasters*: data in, picture out (but, what picture?)
    - Some want/need complete control of graphic styles, rendering details
    - Graphic developers want it all: freedom to invent!

... Conclusions

- The future of statistical graphics?
  - Statistical graphics is on the right track when ...
    - it allows one to construct a pretty picture of data,
    - the picture is faithful to some (possibly complex) model,
    - the picture leverages the perceptual and cognitve capabilities of the viewer.
  - Statistical graphics is on the right track when ...
    - it moves the 80–20 rule in favor of the user/developer,
    - it nurtures future growth of tools, techniques → insight,
    - it allows for *beauty* as well as *truth*.

References

Carr, D., Olsen, A. R., Pierson, S. M., and Courbois, J.-Y. Boxplot variations in a spatial context: An Omernik ecoregion and weather example. *Statistical Computing & Statistical Graphics Newsletter*, 9(2):4–13, 1998.

Edwards, D. *Introduction to Graphical Modelling*. Springer-Verlag, New York, NY, 1995.

Friendly, M. *Advanced Logo: A Language for Learning*. L. Erlbaum Associates, Hillsdale, NJ, 1988.

Friendly, M. *Visualizing Categorical Data*. SAS Institute, Cary, NC, 2000.

Friendly, M. and Kwan, E. Effect ordering for data displays. *Computational Statistics and Data Analysis*, 43(4):509–539, 2003.

Shaw, W. T. and Tigg, J. *Applied Mathematica: Getting Started, Getting It Done*. Addison-Wesley, Reading, MA, 1994.

Valero, P., Young, F., and Friendly, M. Visual categorical analysis in ViSta. *Computational Statistics and Data Analysis*, 43(4):495–508, 2003.

Wilkinson, L. *The Grammar of Graphics*. Springer, New York, 1999.