

Big Data and Big Questions: Vignettes and lessons from the history of data visualization

Michael Friendly York University Workshop on Visualization for Big Data Fields Institute, Feb. 2015

Big meta Q:Visualization-based discoveries ??

- When have graphics led to discoveries that might not have been achieved otherwise?
 - Snow (1854): cholera as a water-borne disease
 - Galton (1883): anti-cyclonic weather patterns
 - E.W. Maunder (1904): 11-year sunspot cycle
 - Hertzsprung/Russell (1911): spectral classes of stars



Big meta Q:Visualization-based discoveries ??

- In the history of graphs, what features and data led to such discoveries?
 - What visual ideas/representations were available?
 - What was needed to see/understand something new?
- As we go forward, are there any lessons?
 - What are the Big Questions for today?
 - How can data visualization help?

Context: Milestones Project





Web site: http://datavis.ca/milestones

Vignettes: 4 heros in the history of data visualization

- The 1st statistical graph: M.F. van Langren and the "secret" of longitude
- "Moral statistics": A.M. Guerry and the rise of modern social science
- Visual tools for state planning: C.J. Minard and the Albums de Statistique Graphique in the "Golden Age"
- · Mapping data: Galton's discoveries & visual insight



Underlying themes

- Escaping flatland: $1D \rightarrow 2D \rightarrow nD$
- The rise of visual thinking and explanation
- Mapping the invisible
- Data \rightarrow Theory \rightarrow Practice
- Graphical excellence
- Appreciating the rich history of DataVis in what we do today

1. Big questions of the 17th century

- Geophysical measurement: distance, time and space
- Astronomy
 - Shape of the earth



- Orbits of planets, comets
- → analytic geometry, errors of measurement, least squares
- Navigation & territorial expansion
 - Map-making and surveying
 - Navigation at sea: latitude (easy) and longitude (hard)

7

Big data of the 17th century

- Astronomical and geodetic tables
 - Ptolemy's Geography (~150 AD): lat/long catalog of known earth
 - Positions of planets, moon, etc. observed from given locations
 - Alfonsine Tables (~1260), Rudolphine tables (Kepler, 1627)
 - Tycho Brahe's star/planet tables

Cabula equanomic Surger ann modely fine	Trabula Tenterie Dinerus Ster Congress water > -
WAY OF A FAR ANALY A WAY A FOOL S AN	Talula in standard & second for water and second the standard and and a second a se
	at a state the state of the state
	a law and a set of the
The consess sector sector sparts so	an and menute ways much i is a first
	a bran guing a bran a gara
二二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二、二	
ም ሮ ቤ ጅ ስ ጅ ስ ጅ ስ ጅ ስ ጅ ስ ጅ ስ ጅ ስ ጅ ስ ጅ ስ ጅ	
1 0 10 10 0 11 16 11 10 10 10 10 10 11 10 11 40 10 11 1 11 1	
1 0 40 10 11 11 11 LI 35 20 12 24 24 20 12 40 1 0 8	
\$ 10 0 10 15 5 5 10 55 16 15 40 13 12 1 58 12	
10 10 10 10 50 5 5 10 4 10 48 14 15 42 12 14 1 32 14	and the second se
	0 1 20 0 0 11 30 1 46 11 20 10 3 41 40 24 24
A VI & LO & LI GO LA TA LA TA TA SO LI SA I LA I A	
S 11 46 10 50 46 30 10 8 30 X8 30 16 11 10 1 0 2 9	
1 11 85 20 21 31 26 10 11 18 21 0 40 18 45 10 42 0 42 2 2	4 0 44 10 0 45 10 0 10 A 60 3 10 0
	4 30 1 46 10 40 5 41 58 10 10 10 50 00 0
10 12 12 10 40 10 12 10 14 SKLAN 31 14 24 13 14 50 1- 33	
11 11 12 10 41 5 14 e 10 15 4 14 15 18 14 45 10 9 31 2 1	
	4 20 - 24 20 30 29 39 30 15 9 10 20 20 20 3
	T 11 1 11 11 14 A GO 10 10 10 10 10 AG
ie sa 30 ao	· · · · · · · · · · · · · · · · · · ·
it is a so is is in the source is it is in the so is it is in the so	
	A 10 35 11 20 10 10 10 10 10 10 10 10

Alfonsine tables, Toledo, ca.1260

Early "firsts" in data visualization

- Map projections, latitude, longitude Claudius Ptolemy, c. 150 AD
- First modern atlas *Teatrum Orbis Terrarum*, Abraham Ortelius, 1570
- First world map showing isogons Guillaume Le Nautonier, 1640
- First visual representation of statistical data M. F. van Langren, 1644



Early "firsts" in data visualization

- Map projections, latitude, longitude Claudius Ptolemy, c. 150 AD
- First modern atlas *Teatrum Orbis Terrarum*, Abraham Ortelius, 1570
- · First world map showing isogons Guillaume Le Nautonier, 1640
- First visual representation of statistical data M. F. van Langren, 1644



M. F. van Langren (1644). La Verdadera Longitud por Mar y Tierra. Antwerp (n.p.)

LA VERDADERA LONGITVD

MAR Y TIERRA.



Mathematico y Cofmographo de fu Mag⁴ reprefenta los puntos figuientes, de la Longitud por Mary Tierra; y dize que fu Padre y Abuelo hizieron profetion de las artes, como Afronomia y Geographia, y en particilar ed dicho fu Padre affitti o en las obferaciones Celefres del famolo Aftronomo TIGHO BRAME, de quien receiblo fus primeras obfervaciones, como confla por las obras del dicho TICHO, estimitino fervió fu Pa-

Las obras del dicho T10H0, still milmo ferviò fu Padre 26, años à fu Magé - en calidad de Cofinographo, en los Eftados de Flandee 26, Y el dicho VAN LANGREN, à imitacion de fus Antepafiados, fe ha exercitado en cflas artes, y difcubierto cofas que hafta agora no fe fabian, incinandofe mas à lo effencial de la dicha ficiencia, que à lo effeculativo, por conocer que todo el mundo necefitava de la Verdadera Longitud por Mar y por Tierra. Y haviendo hallado cofa confiderable en dicha materia, la propufo als Serenifiima Infanta Doña ISAS EL, laqual por fer muy afficionada à las dichas artes, encomendò à fu Magé al dicho van Laworan de fu mano propia en ciano 1620, pidendole le encargafie la correccion general de la Geographia: Lo que confentió fu Magé, por fu Real Cedula, por fer losertine de las diffancias, que los mas graves Aftronomos y Geographos ponen nure Roma y Tolecdo, por laqual le puede conjecturar lo que fera de lagrars mas diflantes.



The Truth about Longitude for Sea and Land (1644) Sent to King Phillip of Spain

- The best astronomers have made their measurements
- You can plainly see the large errors in estimates of distance from Toledo to Rome
- I know the secret of longitude, and will tell you if you grant me a Patent.

Early history of statistical graphs

Timeline of Invention of Basic Forms for Statistical Graphs



The problem of longitude

- Latitude- easy to determine by inclination of sun, moon, stars above/below equator
- Longitude:
 - No fixed 0 reference; distance varies with latitude
 - Only known fact: 360° = 24hrs \rightarrow 15° / hr
 - $\blacksquare \to$ Need to know $\Delta time \mbox{ precisely, for given location}$
- Solution classes:
 - I 2 clocks: ∆time = time_{Here} time_{There}
 - Astrononomic: Δtime = time_{Here} time_{There}

13

Longitude: Lunar distance method

Measure the angle between moon and some other body

Lookup in almanac time at which that distance would be observed in Greenwhich: time_{There}

Determine time_{Here} (sextant)

Longitude = $15^{\circ} \times \Delta$ time





14

Why the first graph got it right

"Every picture tells a story" - Rod Stewart

- · What was van Langren's communication goal?
- What else could he have done?
- Why did the idea of a graph occur to him?



What else could he have done?

- What would occur to men of his time to convey a message to the King?
- ... he could have used a table have sorted by year to establish priority (or show change).

	Sorted	by	Priority	
--	--------	----	----------	--

Year	Name	Longitude	Where
150	Ptolomeus, C.	27.7	Egypt
1463	Regiomontanus,	25.4	Germany
1530	Lantsbergius, P.	21.1	Belgium
1536	Schonerus, I.	20.8	Germany
1542	Ortonius	26.0	France
1567	Mercator, G.	19.6	Flanders
1567	Clavius, C.	26.5	Germany
1578	Brahe, T.	21.5	Denmark
1582	Maginus, A.	29.8	Italy
1601	Organus, D.	30.1	Germany
1605	lansonius, G.	17.7	Flanders
1610	Argelius, A.	28.0	Italy

Answers: Who did it when?

•... he could have sorted by *longitude*, to show the *range*.

Answers: How much did they vary?

Longitude	Name	Year	Where
17.7	G. lansonius	1605	Flanders
19.6	G. Mercator	1567	Flanders
20.8	I. Schonerus	1536	Germany
21.1	P. Lantsbergius	1530	Belgium
21.5	T. Brahe	1578	Denmark
25.4	I. Regiomontanus	1463	Germany
26.0	Orontius	1542	France
26.5	C. Clavius	1567	Germany
27.7	C. Ptolomeus	150	Egypt
28.0	A. Argelius	1610	Italy
29.8	A. Maginus	1582	Italy
30.1	D. Organus	1601	Germany

Sorted by Longitude

Sorted by Authority

•... he could have sorted by *name*, to show *authority*.

Answers: What did XXX say?

Name	Longitude	Year	Where
Argelius, A.	28.0	1610	Italy
Brahe, T.	21.5	1578	Denmark
Clavius, C.	26.5	1567	Germany
lansonius, G.	17.7	1605	Flanders
Lantsbergius, P.	21.1	1530	Belgium
Maginus, A.	29.8	1582	Italy
Mercator, G.	19.6	1567	Flanders
Organus, D.	30.1	1601	Germany
Orontius	26.0	1542	France
Ptolomeus, C.	27.7	150	Egypt
Regiomontanus, I.	25.4	1463	Germany
Schonerus, I.	20.8	1536	Germany

Only a graph shows...



Only a graph shows...

We can see what van Langren was trying to show in relation to a modern map

- Scales aligned to position Rome at the correct location (unknown to him)
- Main goal: show wide variability
- Added benefit: we can see that all estimates were positively biased
- What's not to like?



Further details: Friendly etal. (2010), The First (Known) Statistical Graph: Michael Florent van Langren and the "Secret" of Longitude. The American Statistician, 64, 185-191

The "secret:" Longitude & selenography

- Van Langren jealously guarded his secret an improvement in the lunar method
- By 1628, he had the idea to use rotation of the moon– rather than mere position in the sky
 - Sunrise/sunset on peaks/craters: continuous set of reference events
- Required:
 - Accurate lunar maps with named features
 - Ephemeris tables, giving onset of sunrise and sunset for those features in the lunar cycle



By 1645, he produced the 1st comprehensive lunar map

 325 named locations (illustrious men, saints, his patrons– Phillip, Isabella)

• Never completed the manual and tables showing exactly how the method could be used

• Today, he is better remembered for his lunar map than for the 1st statistical graph

• Nearly all of his nomenclature was later replaced, except for the self-named *Langrenus Crater*

2. Big questions of the early 1800s

Issues for European states

- Demography: taxes, raising an army (Süssmilch, 1741)
- "Statistik": Numbers of the state (Achenwall, 1748)
- Social problems: crime, suicide, literacy, etc.
- Anthropometry: the measure of Man
 - Distributions of human characteristics
 - Birth, mortality, lifespan
- Beginnings of statistical theory and application
 - Normal distⁿ (de Moivre, 1733)
 - L'homme moyen (Quetelet, 1835)

22

Big data of the early 1800s: "An avalanche of social numbers"

- J.-B.J. Fourier: *Recherches statistique sur la ville de Paris* (1821-1829)
 - Massive tabulations: births, deaths (by cause), admission to insane asylums (age, sex, affliction)
- Ministry of Justice: Compte generale (1825--)
 - First national compilation of criminal justice data
 - All charges & dispositions, quarterly, 86 departments
- Other sources:
 - Bureau de Longitudes (illegitimate births)
 - Parent-Duchatelet (prostitution); Min. of War (desertions)
 - Suicide notes in Paris collected and analyzed for motives
- Social issues could now be addressed with DATA

A. M. Guerry and the rise of social science

Essai sur la statistique moral de la France The launching pad of modern social science

- Presented to Academie des Sciences Français July 2, 1832
- First systematic analysis of comprehensive data on crime, suicide, and other social variables.
- ► Along with Quetelet (1831, 1835), established the study of "moral statistics" → modern social science, criminology, sociology



Social context of crime in 1820s France

- Crime a serious concern:
 - Explosive growth in Paris
 - Widespread unemployment,
 - Emergence of "dangerous classes"
- Liberal ("philanthrope") view
 - Increase education
 - Better prison conditions, diet (bread and soup)
 - Religious instruction
- Conservative view
 - Build more prisons
 - Harsher treatment of recidivists
- Now, there was finally some DATA!

25

The discovery of "social facts" Stability and Variation

Guerry's results were both compelling and startling:

- Rates of crime and suicide remained remarkably invariant over time, yet varied sytematically by region, sex of accused, type of crime, etc.
- In any given French city or department, almost the same number committed suicide, stole, gave birth out of wedlock, etc.

Year	1826	1827	1828	1829	1830	Avg			
Sex	All accused (%)								
Male	79	79	78	77	78	78			
Female	21	21	22	23	22	22			
Age	Accused of Theft (%)								
16-25	37	35	38	37	37	37			
25–25	31	32	30	31	32	31			
Crime	Committed in summer (%)								
Indecent assault		36	36	35	38	36			
Assault & battery		28	27	27	27	28			

The discovery of "social facts" Social laws á la physical laws

Do crime and other moral variables represent:

- structural, lawful characteristics of society, or are they
- simply indicants of individual behaviour?

Guerry argued:

Each year sees the same number of crimes of the same degree reproduced in the same regions.(Guerry, 1833, p.10)

... We are forced to recognize that the facts of the moral order are subject, like those of the physical order, to invariable laws (Guerry, 1833, p14)

1829: Statistique comparée de l'état de l'instruction...

- First shaded thematic maps of crime data
- First comparative maps of social data
- inversely related to crime against property!
- ► Instruction: → France obscure and France éclairée (Dupin, 1826)
- North of France highest in education, but also in property crime!



1833: Essai sur la statistique morale de la France

- Divided the 86 departments into 5 regions
- Supplemented data from the Compte général with:
 - Suicides in Paris, 1794–1832
 - Prostitutes in Paris (Parent-Duchâtelet)
 - Wealth (taxes per inhabitant)
 - Distribution of clergy
 - ٠...
- First study to use crime data to 'test' hypotheses
- Attracted widespread interest in Europe



Guerry's 1833 map of literacy in France

29

1833: Semi-graphic tables

How does type of crime vary with age?

- ▶ → Used ranked tables of crime/1000 connected by colored lines
- First instance of modern parallel coordinates plot



Figure: Relative ranking of crimes at different ages

1833: Semi-graphic tables

Crimes against persons

- Indecent assault on adults (viol sur des adultes) decreases with age
- Indecent assault on children increases with age (top for 70+)
- Paricide rises to max at age 60–70



Figure: Ranking of crimes against persons at different ages

1864: Statistique morale de l'Angleterre comparée...

- Proposes to replace simple "moral statistics" (tables) with "analytical statistics"
 - calculation, graphic display
 - ► → general, abstract results
- ► 17 large color plates (56 × 39 cm):
 - data for France (1825–1855), England (1834–1855)
 - crimes against persons and property decomposed in various ways
 - first attempt to delineate multivariate relations among moral variables
- Voluminous data:
 - 85,564 suicide records (1836–1860), classified by motive
 - 226,224 accused of personal crime
 - $\blacktriangleright\,$ numbers, in a line $\rightarrow\,$ 1170 meters!

STATISTIQUE MORALE L'ANGLETERRE LA STATISTIQUE NOLALE LA STATISTIQUE NOLALE LA FRANCE

· PAN A.-N. GUERRY

АТЦАВ СООТВАТИИ СТАТИТИСТИ СТАТИТИ СТАТИТИ СТАТИТИСТИ СТАТИТИСТИ СТАТИТИ ПО ПО ТОТИТИ СТАТИТИ СТ

Antipolitika da aprilizio consulta par futura a successo antipolitika entrepetata outerane consuma a claracettor an manesa at entre e senare l'h est, deglale area (1) plantes parely a comparison e contexe, Phi e 1800 db.





Further details: Friendly, M. (2007). A.-M. Guerry's Moral Statistics of France: Challenges for Multivariable Spatial Analysis, Statistical Science, 22, 368-399

1864: Statistique morale de l'Angleterre comparée... Comparing France and England



33

Statistique analytique: General causes of crime

Plate XVII: M. Guerry's magnum opus



3. The graphic vision of C. J. Minard



- Marey (1878): "defies the pen of the historian in its brutal eloquence"
- Tufte (1983): "the best statistical graphic ever produced"

Visual thinking, visual explanation

1840: Why did the bridge at Bourg-St. Andèol collapse?

Minard's report consisted essentially of this self-explaining diagram.



Big questions of the mid 1800s

- 1830—1860: emergence of modern French state, dawn of globalization
- Trade, commerce, transportation:
 - Where to build railroads, canals?
 - How to compete with imports/exports?
 - Visualizing changes over time, differences over space
 - ${\ }{\ }{\ }\rightarrow$ Flow maps and other graphical innovations
- These questions led to the "Golden Age" of statistical graphics.

See: Friendly, M. (2008). The Golden Age of Statistical Graphics, Statistical Science, 23, 502-535

37

Flow maps as visual tools

Transport of passengers on the principal railroads in Europe in 1862



Effect of US civil war on cotton trade





Statistical atlases: Data → practice, national identity & graphical excellence

- Statistical albums ~ 1870—1910
 - France: Album de Statistique Graphique: 1879-1899
 - USA: Census atlases: 1870/80/90
 - Gemany: local albums (Berlin, Frankfurt, etc.)
 - Switzerland: Atlas graphique de la Suisse:1897, 1914
 - Others: Latvia, Romania, Bulgaria, etc.
- Goals
 - Visualize progress in commerce, industry, transport
 - Provide a graphic portrait of a nation
 - Consolidate a national identity

Album de statistique graphique

- Published by the Statistical Graphics Bureau, Ministry of Public Works, Émile Cheysson, director
- 18 volumes: 1879-1899, 12-34 plates each, ~ 11"x15" pages
- Graphic forms:
 - Flow maps (simple, double, multi)
 - Pie maps, star, radial, polar time-series, proportional circles
 - Mosaic maps, anamorphic maps, planetary diagrams
 - Choropleth, bi-polar scales
 - Charts: line, bar, time-series
- Themes:
 - Recurrent: railroads, navigation, transport--- bread & butter topics
 - Occasional: agriculture, Paris, expositions, ...
- Pinnacle of the *Golden Age*: exquisite sampler of all known graphic forms!

41

4. Galton's visual discoveries-Bivariate normal correlation surface (1886)

Table 9.1	One o	of Galt	on's co	orrelat	ion ta	bles								
Height of the mid-	mid-Height of the adult child											-2014-002-00-0		
parent in inches	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7
>73.0					_		_	_	—		_	1	3	-
72.5		<u>, 1997</u>	10			-		1	2	1	2	7	2	4
71.5			-		1	3	4	3	5	10	4	9	2	2
70.5	1		1		1	1	3	12	18	14	7	4	3	3
69.5			1	16	4	17	27	20	33	25	20	11	4	5
68.5	1		7	11	16	25	31	34	48	21	18	4	3	· .
67.5	· <u> </u>	3	5	14	15	36	38	28	38	19	11	4	-	
66.5		3	3	5	2	17	17	14	13	4		10000 1000	_	
65.5	1	-	9	5	7	11	11	7	7	5	2	1	—	—
64.5	1	1	4	4	1	5	5		2					<u> </u>
<64.0	1		2	4	1	2	2	1	1		-			—
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14
Medians			66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	

Source: Galton (1886), p. 68.

43

Recursive multi-mosaic map

Distribution of passengers and goods from the Paris railways to the rest of France [*Album*, 1884, pl. 11]





Visual smoothing \rightarrow Insight

Table 9.1 One of Galton's correlation tables



Source: Galton (1886), p. 68.

Visual insight → Theory





... that Galton should have evolved all this ... is to my mind one of the most noteworthy scientific discoveries arising from analysis of pure observation (Pearson 1920, p37)



Galton's big data : Isochronic chart (1881)



Galton's discovery of weather patterns-Perhaps the most notable *purely graphic* discovery ever!

METEOROGRAPHICA,

METHODS OF MAPPING THE WEATHER;

ILLUSTRATED BY UPWARDS OF 600 PRINTED AND LITHOGRAPHED DIAGRAMS RETERING TO

THE WEATHER OF A LARGE PART OF EUROPE,

During the Month of December 1861.

By FRANCIS GALTON, F.R.S.

(Galton, 1863)

Method: All weather stations across Europe asked to record data 3x/day for all of Dec., 1861

Data: recordings of barometric pressure, wind dir/speed, rain, temp., cloud: 3x/day, 50 weather stations in Europe.

Graphic analysis: 3x31=93 maps, each with multivariate glyphs showing all variables

Visual ideas:

Iconic symbols

• Multivariate glyphs (stamps!)





Visual abstraction \rightarrow Patterns

How to see patterns of geographical variation over time?

.

- · Iconic symbols on a geographical grid
- "Small multiples:" separate graphs laid out for direct comparison





Symbols in Barometrical Charts. Inches. 29,20 and below. Iuches. Inches. 29,70 to 29,46 Inches. Inches. 29.45 to 29.21 Inches. Inches. - 29,95 to 29,71 Black 0 \odot ¥. 0 30.71 and above. 30.21 to 30.45 30.46 to 30.70 - 29.96 to 30.20

49

Visual abstraction → Patterns

What varies with what, over time and space?

• mini, abstract maps: vars x TOD

· iso-contours, shading to show equivalence

arrows to show wind direction





Data for Dec 5, 1861

50

The large picture \rightarrow Insight

Pattern:

Low pressure (black) in early Dec. \rightarrow CCW wind High pressure (red) in late Dec. \rightarrow CW wind

Graphic: 3x3x31 grid, mapping {pressure, wind/ rain, temperature} x {AM, 12, PM} x day {1:31}

(try this with your software!)



Visual insight \rightarrow Theory

Visual insight from 93 (3x31) high-D graphs:

- · Changes in wind dir w/ pressure over time
- \rightarrow Winds revolve inwardly
- (CCW) in low pressure areas-
- as in a cvclone:
- → revolve outwardly (CW) in high pressure areas- "anticyclone"

Theory:

· Explained by Dove's 'Law of Gyration'

 Prediction: reversed pattern (CW/CCW) in southern hemisphere - confirmed!



Theory → Practice

The first modern weather map, *London Times*, Apr. 1, 1875

Galton did for weathermen what Kepler did for Tycho Brahe. This is no small accomplishment. (Wainer 2005)



Looking forward: Big data and big questions

- What are the Big Questions for today?
- Global warming & climate change
 - Monitoring and predicting artic ice, sea level rise
 - Extreme weather events
- Disease outbreaks
 - Ebola, avian flu
- Threat assessment, terrorism
- Genomics

Conclusions

- In the history of science, visualization often:
 - ...proved crucial in discovery
 - ... required simplification and summary of "big" data
 - ... provided simple explanations for complex phenomena
- Notable examples in this history illustrate ...
 - Importance of visual thinking
 - Interoccularity: message hits you between the eyes
 - Role of *smoothing* in seeing patterns, gaining insight
 - Necessity to escape flatland:
 - Progress in display of increasingly rich and complex data
 - Data \rightarrow Visual abstraction \rightarrow Theory \rightarrow Practice

54

How can graphics help?

• Plot the data and smooth



How can graphics help?

• Note departures from general pattern



57

Interoccularity

Make the message hit you between the eyes



Source: http://www.politically-incorrect-humor.com/2010/03/positive-proof-of-global-warming/

Validation?

• Test predictions from a fitted model



59

Causal & visual explanation

Lack of pirates causes global warming!



Theory \rightarrow practice:

To stop global warming, become a pirate!