Learning Goals

What is exploratory data analysis and why is it important?

What factors should we consider when exploring a dataset?

How do visualization researchers design tools to support exploratory data analysis? (one example)
Topics

Exploratory Data Analysis
Historical Context
Visualizations vs Statistical Models
Data Wrangling
Exploratory Analysis Examples
Tableau / Polaris
What was the first data visualization?
~6200 BC Town Map of Catal Hyük, Konya Plain, Turkey 0 BC
~950 AD Position of Sun, Moon and Planets
MACVLAE IN SOLE APPARENTES, OBSERVATAE
anno 1611. ad latitudinem in grad. +8. min. +0.


a c horizontis arcus solis diurnus. Solerius ex parte a, maculas exhibet quas vides, occasus et c, eadem ratione primi motus, nonnulla incertit. Et hane matutinam vesperinamque mutationem, omnes macula quotidian subeunt quad semel exhibe et nonusse, sufficit.

Macula fuit valde conspicua, propter notabilis longa novem magnitudinem.

Macula M, es hactenus usus eum maior nullius praeritio magnitudinem.

Figura quae habet antennam signum X, est O'letter.
Longitudinal distance between Toledo and Rome, van Langren 1644
The Rate of Water Evaporation, Lambert 1765
The Rate of Water Evaporation, Lambert 1765
The **Golden Age of Data Visualization**

1786 - 1900
1786 1826(?) Illiteracy in France, Pierre Charles Dupin
“to affect thro’ the Eyes what we fail to convey to the public through their word-proof ears”
1864 British Coal Exports, Charles Minard
Consommations approximatives de la Houille dans la Grande Bretagne de 1850 à 1864.

Les abscisses représentent les années et les ordonnées les quantités annuelles de houille consommée. Les couleurs indiquent les espèces de consommations. Les longueurs d'ordonnées comprises dans une couleur sont les quantités de houille consommées à raison de deux millimètres pour un million de tonnes.

Données admises pour former le Tableau ci-dessus.

Consommations — Sources des Renseignements.

Exportations — Mineral statistics 1865 page 91 et renseignements parlementaires.

District de Londres — id. — page 93.

Produits de la Fonte — id. — page 215 et pour les années avant 1855 calculée à raison de 3° de houille pour 1° de fonte, en admettant la quantité annuelle de fonte du Coal question page 192.


Foyers domestiques — En y compréhens les petites manufactures. On l'estimait en 1848 à 19 millions de tonnes, (A) qu'on peut réduire à 18 millions. Pour les foyers seuls, mais qu'on peut porter à 20 millions pour la population de 1864.

Éclairage au Gaz — Consommation estimée généralement du 3° au 8° de la production totale.

Exploitation des Chemins de Fer — En supposant pour consommation totale 10° pour kilomètre parcouru par les trains d'après les renseignements parlementaires.

Navigation à Vapeur — Calculée à raison de 5° houille par cheval-vapeur et par heure, le nombre de chevaux étant celui des Steam Vessels pour 1864, et les steamers étant supposés marcher la moitié de l'année.

Avant 1864, j'ai supposé les consommations proportionnelles aux tonnages annuels des steamers du statistical abstract et du Board of Trade.

(A) Voir l'excellent article houille de M. Lamé Fleury, Dictionnaire du Commerce Page III.
1786

1884 Rail Passengers and Freight from Paris
66. Interstate Migration—Number of Native Immigrants and Native Emigrants, by States and Territories: 1890.

Native immigrants. [Hundreds of thousands.] Native emigrants.

NEW YORK
OHIO
PENNSYLVANIA
ILLINOIS
VIRGINIA
INDIANA
TENNESSEE
KENTUCKY
MISSOURI
IOWA
GEORGIA
ALABAMA
WISCONSIN
NORTH CAROLINA
MASSACHUSETTS
MISSISSIPPI
MARYLAND
MAINE
SOUTH CAROLINA
MICHIGAN
NEW JERSEY
VERMONT
KANSAS
CONNECTICUT
NEW HAMPSHIRE
ARKANSAS
LOUISIANA
MINNESOTA
TEXAS
WEST VIRGINIA
KANSAS
RHODE ISLAND
CALIFORNIA
DELWARE
OREGON
DISTRICT OF COLUMBIA
FLORIDA
UTAH
COLORADO
NEW MEXICO
NORTH DAKOTA
SOUTH DAKOTA
NEVADA
WASHINGTON
MONTANA
IDaho
Wyoming
ARIZONA

1786
1890 Statistical Atlas of the Eleventh U.S. Census
Negro business men in the United States.
Nègres Americans dans les affaires.
Done by Atlanta University.

Estimated capital
Capital évalué

General merchandise stores
Magasins de provisions et d'objets divers

Grocers
Epiciers

Bankers
Banquiers

 Undertakers
Entrepreneurs de pompes funèbres

Building contractors
Entrepreneurs de batiments

Drugists
Pharmacists

Publishers
Éditeurs

Building and loan associations
Institutions financières coopératives

1900 Visualizing Black America, W. E. B. DuBois et al.
The Rise of Statistics
Rise of **formal statistical methods** in the physical and social sciences

**Little innovation** in graphical methods

A period of **application and popularization**

Graphical methods enter textbooks, curricula, and **mainstream use**
Four major influences act on data analysis today:
1. The formal theories of statistics.
2. Accelerating developments in computers and display devices.
3. The challenge, in many fields, of more and larger bodies of data.
4. The emphasis on quantification in a wider variety of disciplines.
The last few decades have seen the rise of formal theories of statistics, "legitimizing" variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions, and restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with "known" probabilities of error.
While some of the influences of statistical theory on data analysis have been helpful, others have not.
Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.
Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the flexibility of the informed human mind.

Accordingly, both approaches and techniques need to be structured so as to facilitate human involvement and intervention.
<table>
<thead>
<tr>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Set D</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.1</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.1</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.11</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
<td>7</td>
<td>7.26</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
</tr>
</tbody>
</table>

**Summary Statistics**

- $u_X = 9.0$
- $\sigma_X = 3.317$
- $u_Y = 7.5$
- $\sigma_Y = 2.03$

**Linear Regression**

- $Y = 3 + 0.5X$
- $R^2 = 0.67$

[Anscombe 1973]
Set A

Set B

Set C

Set D

[Anscombe 1973]
Data Wrangling
I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any “analysis” at all.

Anonymous Data Scientist

[Kandel et al. ’12]
In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.
**Reported crime in Alabama**

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>4525375</td>
<td>4029.3</td>
<td>987</td>
<td>2732.4</td>
<td>309.9</td>
</tr>
<tr>
<td>2005</td>
<td>4548327</td>
<td>3900</td>
<td>955.8</td>
<td>2656</td>
<td>289</td>
</tr>
<tr>
<td>2006</td>
<td>4599030</td>
<td>3937</td>
<td>968.9</td>
<td>2645.1</td>
<td>322.9</td>
</tr>
<tr>
<td>2007</td>
<td>4627851</td>
<td>3974.9</td>
<td>980.2</td>
<td>2687</td>
<td>307.7</td>
</tr>
<tr>
<td>2008</td>
<td>4661900</td>
<td>4081.9</td>
<td>1080.7</td>
<td>2712.6</td>
<td>288.6</td>
</tr>
</tbody>
</table>

**Reported crime in Alaska**

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>657755</td>
<td>3370.9</td>
<td>573.6</td>
<td>2456.7</td>
<td>340.6</td>
</tr>
<tr>
<td>2005</td>
<td>663253</td>
<td>3615</td>
<td>622.8</td>
<td>2601</td>
<td>391</td>
</tr>
<tr>
<td>2006</td>
<td>670053</td>
<td>3582</td>
<td>615.2</td>
<td>2588.5</td>
<td>378.3</td>
</tr>
<tr>
<td>2007</td>
<td>683478</td>
<td>3373.9</td>
<td>538.9</td>
<td>2480</td>
<td>355.1</td>
</tr>
<tr>
<td>2008</td>
<td>686293</td>
<td>2928.3</td>
<td>470.9</td>
<td>2219.9</td>
<td>237.5</td>
</tr>
</tbody>
</table>

**Reported crime in Arizona**

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>5739879</td>
<td>5073.3</td>
<td>991</td>
<td>3118.7</td>
<td>963.5</td>
</tr>
<tr>
<td>2005</td>
<td>5953007</td>
<td>4827</td>
<td>946.2</td>
<td>2958</td>
<td>922</td>
</tr>
<tr>
<td>2006</td>
<td>6166318</td>
<td>4741.6</td>
<td>953</td>
<td>2874.1</td>
<td>914.4</td>
</tr>
<tr>
<td>2007</td>
<td>6338755</td>
<td>4502.6</td>
<td>935.4</td>
<td>2780.5</td>
<td>786.7</td>
</tr>
<tr>
<td>2008</td>
<td>6500180</td>
<td>4087.3</td>
<td>894.2</td>
<td>2605.3</td>
<td>587.8</td>
</tr>
</tbody>
</table>

**Reported crime in Arkansas**

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>2750000</td>
<td>4033.1</td>
<td>1096.4</td>
<td>2699.7</td>
<td>237</td>
</tr>
<tr>
<td>2005</td>
<td>2775708</td>
<td>4068</td>
<td>1085.1</td>
<td>2720</td>
<td>262</td>
</tr>
<tr>
<td>2006</td>
<td>2810872</td>
<td>4021.6</td>
<td>1154.4</td>
<td>2596.7</td>
<td>270.4</td>
</tr>
<tr>
<td>2007</td>
<td>2834797</td>
<td>3945.5</td>
<td>1124.4</td>
<td>2574.6</td>
<td>246.5</td>
</tr>
<tr>
<td>2008</td>
<td>2855390</td>
<td>3843.7</td>
<td>1182.7</td>
<td>2433.4</td>
<td>227.6</td>
</tr>
</tbody>
</table>

**Reported crime in California**

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>35842038</td>
<td>3423.9</td>
<td>686.1</td>
<td>2033.1</td>
<td>704.8</td>
</tr>
<tr>
<td>2005</td>
<td>36154147</td>
<td>3321</td>
<td>692.9</td>
<td>1915</td>
<td>712</td>
</tr>
<tr>
<td>2006</td>
<td>36457549</td>
<td>3175.2</td>
<td>676.9</td>
<td>1831.5</td>
<td>666.8</td>
</tr>
<tr>
<td>2007</td>
<td>36553215</td>
<td>3032.6</td>
<td>648.4</td>
<td>1784.1</td>
<td>600.2</td>
</tr>
<tr>
<td>2008</td>
<td>36756666</td>
<td>2940.3</td>
<td>646.8</td>
<td>1769.8</td>
<td>523.8</td>
</tr>
</tbody>
</table>

**Reported crime in Colorado**

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
<th>Property crime rate</th>
<th>Burglary rate</th>
<th>Larceny-theft rate</th>
<th>Motor vehicle theft rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>4601821</td>
<td>3918.5</td>
<td>717.3</td>
<td>2679.5</td>
<td>521.6</td>
</tr>
</tbody>
</table>
Data Wrangling

One often needs to manipulate data prior to analysis. Tasks include reformatting, cleaning, quality assessment, and integration.

Approaches include:
Manual manipulation in spreadsheets
Code: arquero (JS), dplyr (R), pandas (Python)
Trifacta Wrangler  http://www.trifacta.com/products/wrangler/
Open Refine  http://openrefine.org/
How do rows, columns, and tables match up with observations, variables, and types? In “tidy” data:
1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

The advantage is that this provides a flexible starting point for analysis, transformation, and visualization.

Our pivoted table variant was not “tidy”!

(This is a variant of normalized forms in DB theory)
Data Quality

“The first sign that a visualization is good is that it shows you a problem in your data...

...every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.”

Martin Wattenberg
Violent Infants!

Marauding Centenarians!

Query Result: 4792 out of 4792 (100%)
Visualize Degrees by School?

- Berkeley
- Cornell
- Harvard
- Harvard University
- Stanford
- Stanford University
- UC Berkeley
- UC Davis
- University of California at Berkeley
- University of California, Berkeley
- University of California, Davis
Data Quality Hurdles

Erroneous Values  misspelling, outliers, …?
Entity Resolution  diff. values for the same thing?
Missing Data  no measurements, redacted, …?
Type Conversion  e.g., zip code to lat-lon
Data Integration  effort/errors when combining data

LESSON: Anticipate problems with your data. Many research problems around these issues!
Administrivia
A1: Visualization Design

Pick a **guiding question**, use it to title your vis. Design a **static visualization** for that question. You are free to **use any tools** (inc. pen & paper).

**Deliverables** (upload via Canvas; see A1 page)
- Image of your visualization (PNG or JPG format)
- Short description + design rationale (≤ 4 paragraphs)

Due by **11:59 pm, Wednesday April 6**.
Tableau Tutorial (Optional)

Friday April 8, 1-2pm
Zoom link available on Canvas
Session will be recorded.
Break Time!
Analysis Example: Motion Pictures Data
<table>
<thead>
<tr>
<th>Title</th>
<th>IMDB Rating</th>
<th>Rotten Tomatoes Rating</th>
<th>MPAA Rating</th>
<th>Release Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>String (N)</td>
<td>Number (Q)</td>
<td>Number (Q)</td>
<td>String (O)</td>
<td>Date (T)</td>
</tr>
</tbody>
</table>
Lesson: Exercise Skepticism

Check **data quality** and your **assumptions**.
Start with **univariate summaries**, then start to consider **relationships among variables**. **Avoid premature fixation!**
Analysis Example: Antibiotic Effectiveness
Data Set: Antibiotic Effectiveness

Genus of Bacteria  String (N)
Species of Bacteria String (N)
Antibiotic Applied  String (N)
Gram-Staining? Pos / Neg (N)
Min. Inhibitory Concent. (g) Number (Q)

Collected prior to 1951.
What questions might we ask?

**Table 1: Burtin’s data.**

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Penicillin</th>
<th>Streptomycin</th>
<th>Neomycin</th>
<th>Gram Staining</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Aerobacter aerogenes</em></td>
<td>870</td>
<td>1</td>
<td>1.6</td>
<td>negative</td>
</tr>
<tr>
<td><em>Brucella abortus</em></td>
<td>1</td>
<td>2</td>
<td>0.02</td>
<td>negative</td>
</tr>
<tr>
<td><em>Brucella anthracis</em></td>
<td>0.001</td>
<td>0.01</td>
<td>0.007</td>
<td>positive</td>
</tr>
<tr>
<td><em>Diplococcus pneumoniae</em></td>
<td>0.005</td>
<td>11</td>
<td>10</td>
<td>positive</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>100</td>
<td>0.4</td>
<td>0.1</td>
<td>negative</td>
</tr>
<tr>
<td><em>Klebsiella pneumoniae</em></td>
<td>850</td>
<td>1.2</td>
<td>1</td>
<td>negative</td>
</tr>
<tr>
<td><em>Mycobacterium tuberculosis</em></td>
<td>800</td>
<td>5</td>
<td>2</td>
<td>negative</td>
</tr>
<tr>
<td><em>Proteus vulgaris</em></td>
<td>3</td>
<td>0.1</td>
<td>0.1</td>
<td>negative</td>
</tr>
<tr>
<td><em>Pseudomonas aeruginosa</em></td>
<td>850</td>
<td>2</td>
<td>0.4</td>
<td>negative</td>
</tr>
<tr>
<td><em>Salmonella (Eberthella) typhosa</em></td>
<td>1</td>
<td>0.4</td>
<td>0.008</td>
<td>negative</td>
</tr>
<tr>
<td><em>Salmonella schottmuelleri</em></td>
<td>10</td>
<td>0.8</td>
<td>0.09</td>
<td>negative</td>
</tr>
<tr>
<td><em>Staphylococcus albus</em></td>
<td>0.007</td>
<td>0.1</td>
<td>0.001</td>
<td>positive</td>
</tr>
<tr>
<td><em>Staphylococcus aureus</em></td>
<td>0.03</td>
<td>0.03</td>
<td>0.001</td>
<td>positive</td>
</tr>
<tr>
<td><em>Streptococcus fecalis</em></td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>positive</td>
</tr>
<tr>
<td><em>Streptococcus hemolyticus</em></td>
<td>0.001</td>
<td>14</td>
<td>10</td>
<td>positive</td>
</tr>
<tr>
<td><em>Streptococcus viridans</em></td>
<td>0.005</td>
<td>10</td>
<td>40</td>
<td>positive</td>
</tr>
</tbody>
</table>
How do the drugs compare?

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>Penicillin</th>
<th>Antibiotic Streptomycin</th>
<th>Neomycin</th>
<th>Gram stain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerobacter aerogenes</td>
<td>870</td>
<td>1</td>
<td>1.6</td>
<td>–</td>
</tr>
<tr>
<td>Brucella abortus</td>
<td>1</td>
<td>2</td>
<td>0.02</td>
<td>–</td>
</tr>
<tr>
<td>Bacillus anthracis</td>
<td>0.001</td>
<td>0.01</td>
<td>0.007</td>
<td>+</td>
</tr>
<tr>
<td>Diplococcus pneumoniae</td>
<td>0.005</td>
<td>11</td>
<td>10</td>
<td>+</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>100</td>
<td>0.4</td>
<td>0.1</td>
<td>–</td>
</tr>
<tr>
<td>Klebsiella pneumoniae</td>
<td>850</td>
<td>1.2</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>800</td>
<td>5</td>
<td>2</td>
<td>–</td>
</tr>
<tr>
<td>Proteus vulgaris</td>
<td>3</td>
<td>0.1</td>
<td>0.1</td>
<td>–</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>850</td>
<td>2</td>
<td>0.4</td>
<td>–</td>
</tr>
<tr>
<td>Salmonella (Eberthella) typhosa</td>
<td>1</td>
<td>0.4</td>
<td>0.008</td>
<td>–</td>
</tr>
<tr>
<td>Salmonella schottmuelleri</td>
<td>10</td>
<td>0.8</td>
<td>0.09</td>
<td>–</td>
</tr>
<tr>
<td>Staphylococcus albus</td>
<td>0.007</td>
<td>0.1</td>
<td>0.001</td>
<td>+</td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>0.03</td>
<td>0.03</td>
<td>0.001</td>
<td>+</td>
</tr>
<tr>
<td>Streptococcus fecalis</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>+</td>
</tr>
<tr>
<td>Streptococcus hemolyticus</td>
<td>0.001</td>
<td>14</td>
<td>10</td>
<td>+</td>
</tr>
<tr>
<td>Streptococcus viridans</td>
<td>0.005</td>
<td>10</td>
<td>40</td>
<td>+</td>
</tr>
</tbody>
</table>

Original graphic by Will Burtin, 1951
How do the drugs compare?

Radius: 1 / log(MIC)
Bar Color: Antibiotic
Background Color: Gram Staining
How do the drugs compare?
How do the drugs compare?

X-axis: Antibiotic | log(MIC)
Y-axis: Gram-Staining | Species
Color: Most-Effective?
**All bacteria**

Streptomycin and Neomycin are more efficient against broad-spectrum antibiotics than Penicillin.

**Gram-negative bacteria only**

Neomycin and Streptomycin are more efficient against gram-negative bacteria, so can be used at a lower dosage here than above.

**Gram-positive bacteria only**

Penicillin is more efficient than either Streptomycin or Neomycin if the bacteria is known to be gram-positive.

---

**Effectiveness of Antibiotics**

- A. aerogenes
- B. abortus
- E. coli
- K. pneumoniae
- M. tuberculosis
- P. vulgaris
- P. aeruginosa
- S. typhosa
- S. schottmuelleri
- B. anthracis
- D. pneumoniae
- S. albus
- S. aureus
- S. fecalis
- S. hemolyticus
- S. viridans

Penicillin

<table>
<thead>
<tr>
<th>MIC (μg/mL)</th>
<th>No Effect</th>
<th>Low Effect</th>
<th>Moderate Effect</th>
<th>High Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>0.01</td>
<td>0.1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>0.002</td>
<td>0.02</td>
<td>0.2</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>0.005</td>
<td>0.05</td>
<td>0.5</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>0.01</td>
<td>0.1</td>
<td>1</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>10</td>
<td>100</td>
<td>1000</td>
</tr>
</tbody>
</table>

Streptomycin

<table>
<thead>
<tr>
<th>MIC (μg/mL)</th>
<th>No Effect</th>
<th>Low Effect</th>
<th>Moderate Effect</th>
<th>High Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>0.01</td>
<td>0.1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>0.002</td>
<td>0.02</td>
<td>0.2</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>0.005</td>
<td>0.05</td>
<td>0.5</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>0.01</td>
<td>0.1</td>
<td>1</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>10</td>
<td>100</td>
<td>1000</td>
</tr>
</tbody>
</table>

Neomycin

<table>
<thead>
<tr>
<th>MIC (μg/mL)</th>
<th>No Effect</th>
<th>Low Effect</th>
<th>Moderate Effect</th>
<th>High Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>0.01</td>
<td>0.1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>0.002</td>
<td>0.02</td>
<td>0.2</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>0.005</td>
<td>0.05</td>
<td>0.5</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>0.01</td>
<td>0.1</td>
<td>1</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>10</td>
<td>100</td>
<td>1000</td>
</tr>
</tbody>
</table>

---

**Minimum Inhibitory Concentration (MIC)**

- Penicillin
- Streptomycin
- Neomycin

---

**Log10 Minimum Inhibitory Concentration (μg/mL)**

- Penicillin
- Streptomycin
- Neomycin

---

**Antibiotic Dose Comparison**

- Penicillin
- Streptomycin
- Neomycin

---

**Antibiotic Effectiveness Comparison**

- Penicillin
- Streptomycin
- Neomycin

**Legend**

- **Darker colors**: more effective
Which antibiotic should one use?
Do the bacteria group by antibiotic resistance?
Do the bacteria group by antibiotic resistance?
Do the bacteria group by antibiotic resistance?

Wainer & Lysen
American Scientist, 2009
Do the bacteria group by antibiotic resistance?

Really a streptococcus! (realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009
Do the bacteria group by antibiotic resistance?

Not a streptococcus! (realized ~30 yrs later)

Really a streptococcus! (realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009
Do the bacteria group by resistance?
Do different drugs correlate?
Do the bacteria group by resistance?
Do different drugs correlate?

Wainer & Lysen
American Scientist, 2009
Lesson: Iterative Exploration

Exploratory Process
1. Construct graphics to address questions
2. Inspect “answer” and assess new questions
3. Repeat...

Transform data appropriately (e.g., invert, log)

Show data variation, not design variation [Tufte]
Tableau / Polaris
Polaris [Stolte et al.]
Tableau

Data Display

Encodings

Data Model
Tableau / Polaris Approach

Insight: can simultaneously specify both database queries and visualization
Choose data, then visualization, not vice versa
Use smart defaults for visual encodings
Can also suggest encodings upon request
Tableau Demo
Specifying Table Configurations

Operands are the database fields
Each operand interpreted as a set \{...\}
Quantitative and Ordinal fields treated differently

Three operators:
concatenation (+)
cross product (x)
nest (/)
GROUP BY Category, Region, Segment
The operators (+, x, /) and operands (O, Q) provide an algebra for tabular visualization. Algebraic statements are then mapped to:

**Visualizations** - trellis plot partitions, visual encodings

**Queries** - selection, projection, group-by aggregation

In Tableau, users make statements via drag-and-drop.

Note that this specifies operands *NOT* operators! Operators are inferred by data type (O, Q).
Table Algebra: Operands

**Ordinal fields**: interpret domain as a set that partitions table into rows and columns.
Quarter = \{(Qtr1),(Qtr2),(Qtr3),(Qtr4)\} ->

<table>
<thead>
<tr>
<th>Qtr1</th>
<th>Qtr2</th>
<th>Qtr3</th>
<th>Qtr4</th>
</tr>
</thead>
<tbody>
<tr>
<td>95892</td>
<td>101760</td>
<td>105282</td>
<td>98225</td>
</tr>
</tbody>
</table>

**Quantitative fields**: treat domain as single element set and encode spatially as axes.
Profit = \{(Profit[-410,650])\} ->
Concatenation (+) Operator

Ordered union of set interpretations
Quarter + Product Type
= {((Qtr1),(Qtr2),(Qtr3),(Qtr4)) + {(Coffee), (Espresso)}}
= {((Qtr1),(Qtr2),(Qtr3),(Qtr4),(Coffee),(Espresso))}

<table>
<thead>
<tr>
<th></th>
<th>Qtr1</th>
<th>Qtr2</th>
<th>Qtr3</th>
<th>Qtr4</th>
<th>Coffee</th>
<th>Espresso</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>48</td>
<td>59</td>
<td>57</td>
<td>53</td>
<td>151</td>
<td>21</td>
</tr>
</tbody>
</table>

Profit + Sales = {((Profit[-310,620]),(Sales[0,1000]]))
Cross (x) Operator

Cross-product of set interpretations
Quarter x Product Type =
{(Qtr1,Coffee), (Qtr1, Espresso), (Qtr2, Coffee), (Qtr2, Espresso), (Qtr3, Coffee), (Qtr3, Espresso), (Qtr4, Coffee), (Qtr4, Espresso)}

Product Type x Profit =

<table>
<thead>
<tr>
<th>Qtr1</th>
<th>Qtr2</th>
<th>Qtr3</th>
<th>Qtr4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coffee</td>
<td>Espresso</td>
<td>Coffee</td>
<td>Espresso</td>
</tr>
<tr>
<td>131</td>
<td>19</td>
<td>160</td>
<td>20</td>
</tr>
</tbody>
</table>

Coffee

profit

Espresso

profit
Nest (/) Operator

Cross-product filtered by existing records

Quarter x Month ->

creates twelve entries for each quarter. i.e., (Qtr1, December)

Quarter / Month ->

creates three entries per quarter based on tuples in database (not semantics)
<table>
<thead>
<tr>
<th>State</th>
<th>Coffee</th>
<th>Espresso</th>
<th>Herbal Tea</th>
<th>Tea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colorado</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Connecticut</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Florida</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Illinois</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Iowa</td>
<td>⬤</td>
<td></td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Louisiana</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Missouri</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Nevada</td>
<td>⬤</td>
<td>⬤</td>
<td></td>
<td>⬤</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>New Mexico</td>
<td>⬤</td>
<td>⬤</td>
<td></td>
<td>⬤</td>
</tr>
<tr>
<td>New York</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Ohio</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Oregon</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Texas</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Utah</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Washington</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
<td>⬤</td>
</tr>
</tbody>
</table>
Quantitative-Quantitative
Ordinal-Quantitative
Querying the Database

(1) Select records from the database, filtering by user-defined criteria.

(2) Partition the records into layers and panes. The same record may appear in multiple partitions.

(3) Group, sort, and aggregate the relations within each pane.

(4) Render and compose layers.
Summary: Connecting Queries and Visualizations in Tableau

Tableau maintains a joint representation of analysis operations as both data queries and visualizations using a table algebra.

This allows Tableau to support a graphical user interface for expressing data queries.

This also enables Tableau to automatically map queries to visualizations and vice versa.
## Common Data Transformations

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalize</td>
<td>$y_i / \sum_i y_i$</td>
</tr>
<tr>
<td>Log</td>
<td>$\log y$</td>
</tr>
<tr>
<td>Power</td>
<td>$y^{1/k}$</td>
</tr>
</tbody>
</table>
| Box-Cox Transform            | $(y^\lambda - 1) / \lambda$ if $\lambda \neq 0$  
|                              | $\log y$ if $\lambda = 0$                   |
| Binning                      | e.g., histograms                             |
| Grouping                     | e.g., merge categories                       |

Often performed to aid comparison (% or scale difference) or better approx. normal distribution.