Last Time: Design & Re-Design
TRIPPING POINT

Half the power plants planned over the last five years have still not come on steam due to slackness of the policy machinery.

Source: India Today
Music: Super Cuts (page 92)
Gene Expression Time-Series [Meyer et al. 11]

Color Encoding

Position Encoding
Artery Visualization [Borkin et al 11]

Rainbow Palette

2D

Diverging Palette

3D
Artery Visualization [Borkin et al 11]

Rainbow Palette

Accuracy: 62%

Shear Stress (Pa)

2D

Diverging Palette

Accuracy: 92%

Shear Stress (Pa)

3D

39%

71%
What was the first data visualization?
~6200 BC Town Map of Catal Hyük, Konya Plain, Turkey
~950 AD Position of Sun, Moon and Planets
MACVLAE IN SOLE APPARENTES OBSERVATAE
anno 1611. ad latitudinem grad. 48. min. 40.

Sunspots over time, Scheiner 1626
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
Exports and Imports to and from Denmark & Norway from 1700 to 1780.


Balance against.

The Commercial and Political Atlas, William Playfair 1786
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”
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 millièmes pour un million de tonnes.

Données admises pour former le Tableau ci-contre.

<table>
<thead>
<tr>
<th>Consommations</th>
<th>Sources des Renseignements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exportations</td>
<td>Mineral statistics 1865 page 214 et Renseignements Parlementaires</td>
</tr>
<tr>
<td>District de Londres</td>
<td>id. id. page 215</td>
</tr>
<tr>
<td>Produits de la Fonte</td>
<td>id. id. page 215 et pour les années avant 1855 calculée à raison de 3/4 de houille pour 1 tonne de fonte, en admettant les quantités annuelles de fonte du Coal question page 192.</td>
</tr>
<tr>
<td>Production du fer</td>
<td>Mineral statistics page 215 et pour les années avant 1865 calculée à raison de 2/5 de houille pour 1 tonne de fonte convertie en fer, et admettant 24/10 de la fonte produite convertie en fer.</td>
</tr>
</tbody>
</table>

Foyers domestiques: En y comprenant les petites manufactures.
On l'estimait en 1848 à 19 millions de tonnes, (A) qu'on peut réduire à 18 millions 1/2 pour les foyers seuls, mais qu'on peut porter à 20 millions pour la population de 1852.


Exploitation des Chemins de Fer. En supposant pour consommation totale 10.5 par Kilomètre parcouru par les trains d'après les renseignements parlementaires.


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.
The Rise of Statistics
Rise of **formal methods** in statistics and social science – Fisher, Pearson, …

**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**

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

[Anscombe 1973]
Topics

Exploratory Data Analysis
Data Diagnostics
Graphical Methods
Data Transformation

Incorporating Statistical Models
Data Space & Model Space
Graphical Inference
Data Diagnostics
<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>4523755</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
- Writing custom code (dplyr in R, Pandas in Python)
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
Graph Viewer

Roll-up by:
- All

Visualization:
- Matrix

Sort by:
- Linkage

Edge centrality filters:
Visualize Friends 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

Missing Data
no measurements, redacted, …?

Erroneous Values
misspelling, outliers, …?

Type Conversion
e.g., zip code to lat-lon

Entity Resolution
diff. values for the same thing?

Data Integration
effort/errors when combining data

**LESSON**: Anticipate problems with your data. Many research problems around these issues!
Analysis Example: Motion Pictures Data
Motion Pictures Data

Title | String (N)
IMDB Rating | Number (Q)
Rotten Tomatoes Rating | Number (Q)
MPAA Rating | String (O)
Release Date | Date (T)
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>
<thead>
<tr>
<th>Bacteria</th>
<th>Penicillin</th>
<th>Streptomycin</th>
<th>Neomycin</th>
<th>Gram Staining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aerobacter aerogenes</td>
<td>870</td>
<td>1</td>
<td>1.6</td>
<td>negative</td>
</tr>
<tr>
<td>Brucella abortus</td>
<td>1</td>
<td>2</td>
<td>0.02</td>
<td>negative</td>
</tr>
<tr>
<td>Brucella anthracis</td>
<td>0.001</td>
<td>0.01</td>
<td>0.007</td>
<td>positive</td>
</tr>
<tr>
<td>Diplococcus pneumoniae</td>
<td>0.005</td>
<td>11</td>
<td>10</td>
<td>positive</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>100</td>
<td>0.4</td>
<td>0.1</td>
<td>negative</td>
</tr>
<tr>
<td>Klebsiella pneumoniae</td>
<td>850</td>
<td>1.2</td>
<td>1</td>
<td>negative</td>
</tr>
<tr>
<td>Mycobacterium tuberculosis</td>
<td>800</td>
<td>5</td>
<td>2</td>
<td>negative</td>
</tr>
<tr>
<td>Proteus vulgaris</td>
<td>3</td>
<td>0.1</td>
<td>0.1</td>
<td>negative</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>850</td>
<td>2</td>
<td>0.4</td>
<td>negative</td>
</tr>
<tr>
<td>Salmonella (Eberthella) typhosa</td>
<td>1</td>
<td>0.4</td>
<td>0.008</td>
<td>negative</td>
</tr>
<tr>
<td>Salmonella schottmuelleri</td>
<td>10</td>
<td>0.8</td>
<td>0.09</td>
<td>negative</td>
</tr>
<tr>
<td>Staphylococcus albus</td>
<td>0.007</td>
<td>0.1</td>
<td>0.001</td>
<td>positive</td>
</tr>
<tr>
<td>Staphylococcus aureus</td>
<td>0.03</td>
<td>0.03</td>
<td>0.001</td>
<td>positive</td>
</tr>
<tr>
<td>Streptococcus fecalis</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>positive</td>
</tr>
<tr>
<td>Streptococcus hemolyticus</td>
<td>0.001</td>
<td>14</td>
<td>10</td>
<td>positive</td>
</tr>
<tr>
<td>Streptococcus viridans</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 faecalis</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: \( \frac{1}{\log(MIC)} \)
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?
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?

Not a streptococcus! (realized ~30 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]
## Common Data Transformations

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalize</td>
<td>$\frac{y_i}{\sum_i y_i}$</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.
Analysis Example: MTurk Participation
# Data Set: Turker Participation

<table>
<thead>
<tr>
<th>Field</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turker ID</td>
<td>String (N)</td>
</tr>
<tr>
<td>Avg. Completion Rate</td>
<td>Number [0,1] (Q)</td>
</tr>
</tbody>
</table>

Collected in 2009 by Heer & Bostock.

What questions might we ask of the data? What charts might provide insight?
Box Plot

Turker Completion Percentage

Min

Median

Max

Lower Quartile

Upper Quartile

Box Plot
Dot Plot (with transparency for overlap)
Dot Plot (with Reference Lines)
Histogram (binned counts)
Used to compare two distributions; in this case, one actual and one theoretical.

Plots the quantiles (here, the percentile values) against each other.

Similar distributions lie along the diagonal. If linearly related, values will lie along a line, but with potentially varying slope and intercept.

Quantile-Quantile Plot
Quantile-Quantile Plots

Turker Task Group Completion %

Uniform Distribution
Gaussian Distribution
Fitted Mixture of 3 Gaussians
Histogram (+ Fitted Mixture of 3 Gaussians)
Lessons

Even for “simple” data, a variety of graphics might provide insight. Tailor the choice of graphic to the questions being asked, but be open to surprises.

Graphics can be used to guide and help assess the quality of statistical models.

Premature commitment to a model and lack of verification can lead an analysis astray.
Administrivia
A2: Exploratory Data Analysis

Use visualization software to form & answer questions

First steps:
Step 1: Pick domain & data
Step 2: Pose questions
Step 3: Profile the data
Iterate as needed

Create visualizations
Interact with data
Refine your questions

Make a notebook
Keep record of your analysis
Prepare a final graphic and caption

Due by 5:00pm
Friday, April 15
Tutorials!

Visualization Tools
Tue 4/12, 3:00-4:20pm PAA 114A
Introduction to Tableau, plus a few others.

d3.js: Data-Driven Documents
Tue 4/19, 3:00-4:20pm PAA 114A
Focus on D3, touches on HTML/CSS/JS
Visualization + Statistics
[The Elements of Graphing Data. Cleveland 94]
By Eye ...
Linear regression w/out outlier ...

[The Elements of Graphing Data. Cleveland 94]
Transforming Data

How well does the curve fit the data?

[Cleveland 85]
Plot the Residuals

Plot vertical distance from best fit curve
Residual graph shows accuracy of fit

[Cleveland 85]
Multiple Plotting Options

Plot model in data space

Plot data in model space

[Cleveland 85]
Graphical Inference
[Buja, Cook, Hofmann, Wickham, et al.]
Choropleth maps of cancer deaths in Texas.

One plot shows a real data set. The others are simulated under the null hypothesis of spatial independence.

Can you spot the real data? If so, you have some evidence of spatial dependence in the data.
Distance vs. angle for 3 point shots by the LA Lakers.

One plot is the real data. The others are generated according to a null hypothesis of quadratic relationship.
Distance vs. angle for 3 point shots by the LA Lakers.

One plot is the real data. The others are generated according to a null hypothesis of quadratic relationship.
Exploratory analysis may combine graphical methods, data transformations, and statistics. Use questions to uncover more questions. Formal methods may be used to confirm, sometimes on held-out or new data. Visualization can further aid assessment of fitted statistical models.
Data Quality
A Detective Story

You have accounting records for two firms that are in dispute. One is lying. *How to tell?*

<table>
<thead>
<tr>
<th>Firm A</th>
<th></th>
<th>Firm B</th>
<th></th>
<th>LIARS!</th>
</tr>
</thead>
<tbody>
<tr>
<td>283.08</td>
<td>25.23</td>
<td>283.08</td>
<td>75.23</td>
<td></td>
</tr>
<tr>
<td>153.86</td>
<td>385.62</td>
<td>353.86</td>
<td>185.25</td>
<td></td>
</tr>
<tr>
<td>1448.97</td>
<td>12371.32</td>
<td>5322.79</td>
<td>9971.42</td>
<td></td>
</tr>
<tr>
<td>18595.91</td>
<td>1280.76</td>
<td>8795.64</td>
<td>4802.43</td>
<td></td>
</tr>
<tr>
<td>21.33</td>
<td>257.64</td>
<td>61.33</td>
<td>57.64</td>
<td></td>
</tr>
<tr>
<td>Amt. Paid: $34823.72</td>
<td></td>
<td>Amt. Rec’d: $29908.67</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Benford’s Law (Benford 1938, Newcomb 1881)

The *logarithms* of the values (not the values themselves) are uniformly randomly distributed.

Hence the leading digit *1* has a ~30% likelihood. Larger digits are increasingly less likely.
Benford’s Law (Benford 1938, Newcomb 1881)

The logarithms of the values (not the values themselves) are uniformly randomly distributed. Holds for many (but certainly not all) real-life datasets: Addresses, Bank accounts, Building heights, … Data must span multiple orders of magnitude. Evidence that records do not follow Benford’s Law is admissible in a court of law!
Model-Driven Data Validation

Deviations from the model may represent errors

Find Statistical Outliers

# std dev, Mahalanobis dist, nearest-neighbor, non-parametric methods, time-series models

Robust statistics to combat noise, masking

Data Entry Errors

Product codes: PZV, PZV, PZR, PZC, PZV

Which of the above is most likely in error?

Opportunity: combine with visualization methods
Confirmatory Analysis
Incorporating Models

**Hypothesis testing**: What is the probability that the pattern might have arisen by chance?

**Prediction**: How well do one (or more) data variables predict values of interest?

**Summarization**: With what parameters does data fit a given function? What is the goodness of fit?

**Scientific theory**: Which model explains reality?
Example: Heights by Gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Male / Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height (in)</td>
<td>Number</td>
</tr>
</tbody>
</table>

\[ \mu_m = 69.4 \quad \sigma_m = 4.69 \quad N_m = 1000 \]
\[ \mu_f = 63.8 \quad \sigma_f = 4.18 \quad N_f = 1000 \]

Is the difference in heights significant? In other words: assuming no true difference, what is the probability that our data is due to chance?
Histograms
Formulating a Hypothesis

Null Hypothesis (H₀): \( \mu_m = \mu_f \) (population)
Alternate Hypothesis (Hₐ): \( \mu_m \neq \mu_f \) (population)

A statistical hypothesis test assesses the likelihood of the null hypothesis.

What is the probability of sampling the observed data assuming the population means are equal?

This is called the p-value.
Testing Procedure

Compute a **test statistic**. This is a number that in essence summarizes the difference.
Calculate the Test Statistic

\[ Z = \frac{\mu_m - \mu_f}{\sqrt{\sigma^2_m/N_m + \sigma^2_f/N_f}} \]

\[ \mu_m - \mu_f = 5.6 \]
Testing Procedure

Compute a test statistic. This is a number that in essence summarizes the difference.

The possible values of this statistic come from a known probability distribution.

According to this distribution, determine the probability of seeing a value meeting or exceeding the test statistic. This is the p-value.
Lookup Probability of Test Statistic

Normal Distribution
\( \mu = 0, \ \sigma = 1 \)
\( Z \sim \mathcal{N}(0, 1) \)

95% of Probability Mass

\( Z = .2 \)
\( Z > +1.96 \)

\( p > 0.05 \)
\( p < 0.05 \)
Statistical Significance

The threshold at which we consider it safe (or reasonable?) to reject the null hypothesis.

If \( p < 0.05 \), we typically say that the observed effect or difference is statistically significant.

This means that there is a less than 5% chance that the observed data is due to chance.

Note that the choice of 0.05 is a somewhat arbitrary threshold (chosen by R. A. Fisher).
## Common Statistical Methods

<table>
<thead>
<tr>
<th>Question</th>
<th>Data Type</th>
<th>Parametric</th>
<th>Non-Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Assumes a particular distribution for the data, e.g., normal (Gaussian).</td>
<td>Does not assume a distribution. Typically works on rank orders.</td>
</tr>
</tbody>
</table>
# Common Statistical Methods

<table>
<thead>
<tr>
<th>Question</th>
<th>Data Type</th>
<th>Parametric</th>
<th>Non-Parametric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do data distributions have different “centers”? (aka “location” tests)</td>
<td>2 uni. dists</td>
<td>t-Test</td>
<td>Mann-Whitney U</td>
</tr>
<tr>
<td></td>
<td>&gt; 2 uni. dists</td>
<td>ANOVA</td>
<td>Kruskal-Wallis</td>
</tr>
<tr>
<td></td>
<td>&gt; 2 multi. dists</td>
<td>MANOVA</td>
<td>Median Test</td>
</tr>
<tr>
<td>Are observed counts significantly different?</td>
<td>Counts in categories</td>
<td></td>
<td>χ² (chi-squared)</td>
</tr>
<tr>
<td>Are two vars related?</td>
<td>2 variables</td>
<td>Pearson coeff.</td>
<td>Rank correl.</td>
</tr>
<tr>
<td>Do 1 (or more) variables predict another?</td>
<td>Continuous</td>
<td>Linear regression</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Binary</td>
<td>Logistic regression</td>
<td></td>
</tr>
</tbody>
</table>