Data Visualization.

1. Purposes of visualization

- Organize and display data
  - doesn't take much data to be baffling in numerical form
    * show Anscombe datasets
  - making pictures of the data helps the scientist / engineer / whomever understand it.
  - designing the visualizations forces you to organize the data.
  - → visualization as part of the intellectual, creative process.
    e.g. introducing 3D molecular viz
    (can discover things otherwise impossible to find)
    * show Snow cholera example.

- Present data
  - data is used to support arguments (scientific, policy, ...)
  - data is used to make decisions (funding, operational, ...)
  - presenting data well is key to having any kind of influence on others' understanding or decisions.
  - → visualization as part of informative / persuasive communication
  * show Challenger launch decision example.

2. Mapping data into a visual display.

- Data types (programming language)

- Datatypes (semantic)
  - Nominal – unorganized set of names.
    apples, oranges, tomatoes, ...
    Toyota, Ford, Subaru, ...
  - Ordinal – ordered set of values.
    Jan, Feb, Mar, ...
    Trial 1, Trial 2, Trial 3, ...
    123 Oak St., 125 Oak St., 129 Oak St., ...

- Quantitative - values themselves are meaningful.
  - interval (arbitrary zero) vs. ratio (relevant zero)
  - discrete (noticeable steps) vs. continuous.

  e.g. interval discrete: # of revolutions.
  ratio discrete: counts (e.g. # of patients with symptom)
  interval continuous: °C, position, potential energy
  ratio continuous: °K, length, mass

- Graphical information channels.
  - spatial —
    - length
    - position
    - size (area, volume?)
  - color —
    - value: lightness (black-to-white)
    - hue: color
    - texture: fill pattern
  - details —
    - shape
    - orientation

- Matching data types to available channels
  - want to pay attention to semantics
  - choose visual channel that carries the semantics well.
  - a rough chart:

		N	O	Q	int	nt
spatial	len			✓		✓
	pos	✓	✓	✓	✓	
	sz		✓	~		~
color	val	✓	✓	~		
	hue	✓				
	texture	✓				
	shape	✓				
	orient.			✓	✓	

- Examples of some common types of vizs:
  - time series
    - horizontal axis: position → Q, interval (time)
    - vertical axis: often $, but any.
    - oldest form of data visualization
    - readily interpreted with little training / effort.
  - relational plots
    - horizontal axis: alleged "cause" } but sometimes only correlation is indicated.
    - vertical axis: alleged "effect" }
    - very powerful way to investigate relationships.
    - scatter plot vs. connected line.
      - scatter plot → unordered data
      - connected line → ordered data. functional "law"
  - bar charts
    - bars are length, so ought to be ratio datatype (and keep zero at the right place!)
    - bars are separate, so no implied connection (best for Nominal, Ordinal).
  - polar plots.
    - independent var. maps to orientation, or direction (no zero point; naturally cyclic) (ideal for directional / angular data)
    - dependent var. maps to radius (natural zero point)
    - distortion of sizes
  - histograms: counts, or frequencies, over categories.
    - categories v. often Q → O (aggregated).
    - dangers of aggregation
    - usefulness of also plotting raw data.

- color maps : encoding quantitative data as color.
  value is the best attribute to map to a quantitative variable.
  hue is best used nominally.
  * map example

  Note Matlab has atrocious default colormaps.

3. Graphical Integrity (Tufte's phrase)

- Avoid misleading the viewer
  - use appropriate mappings rather than designing for exaggeration
  - don't add decoration designed to change interpretation
  - report money in real terms.

- Easy to mislead by mistake as well as on purpose

- Be very cautious with size ↔ quantity mappings.
  (perceived size = area? length?)

4. Graphical Excellence (or, Graphical Makeovers)

- Key principle : data/ink ratio.

- erase unnecessary marks!
  - no extra frames
  - shorten frames
  - not too much grid
  - not extra decoration w/o purpose

- shrink to increase density
  - small multiples.