Lecture 2: Introduction To Data Data Visualization

Instructor: Saravanan Thirumuruganathan

Outline

1. Data Mining Terminology
2. Basics of Visualization
   - Graph integrity
   - 2D visualization
   - Basics of higher dimensional visualization

Piazza

- Enrollment done for registered students
- Auditing students - send me your email id
- "Search for Teammates" enabled

Socrative

- Instant Student Feedback
- Accessible via Smart Phone, Tablet, Laptop
- No login needed from Student's end
- Use a consistent name throughout the semester

In-Class Quizzes

- **URL:** `http://m.socrative.com/`
- **Room Name:** **4f2bb99e**

Misc Announcements

- Slides for Lecture 1 updated
- Change Office hour timings?
- Installation of Scientific Python

Other Relevant Online Classes

- Machine Learning, Stanford: https://www.coursera.org/course/ml
- Mining of Massive Datasets, Stanford: https://www.coursera.org/course/mmds
- Statistical Learning, Stanford: https://class.stanford.edu/courses/HumanitiesandScience/StatLearning/Winter2015/about

Data Mining Terminology

Data Matrix

Table 1.1. Extract from the Iris dataset

	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

Data Matrix

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array}\right)$$

Data Matrix

- n rows and d columns
- Row $\Rightarrow$ Tuple/Entities
- Column $\Rightarrow$ attribute/feature
- Special column called **Class**
- x_i: i-th row, X_j: j-th column
- Row $\Rightarrow$ entities, instances, examples, records, transactions, objects, points, feature-vectors, tuples
- Column $\Rightarrow$ attributes, properties, features, dimensions, variables, fields
- $n \Rightarrow$ size, $d \Rightarrow$ dimensionality of data

Geometric View

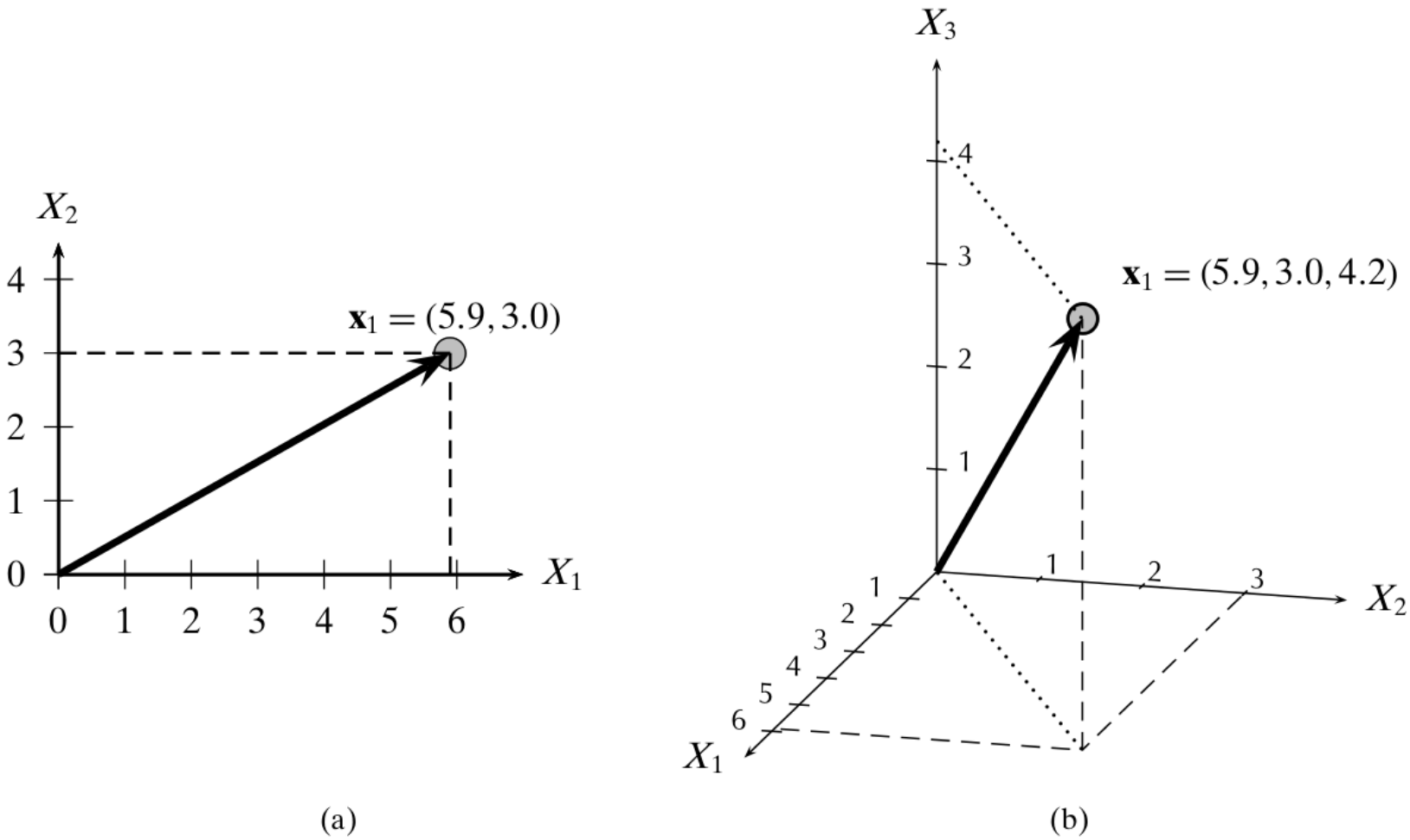
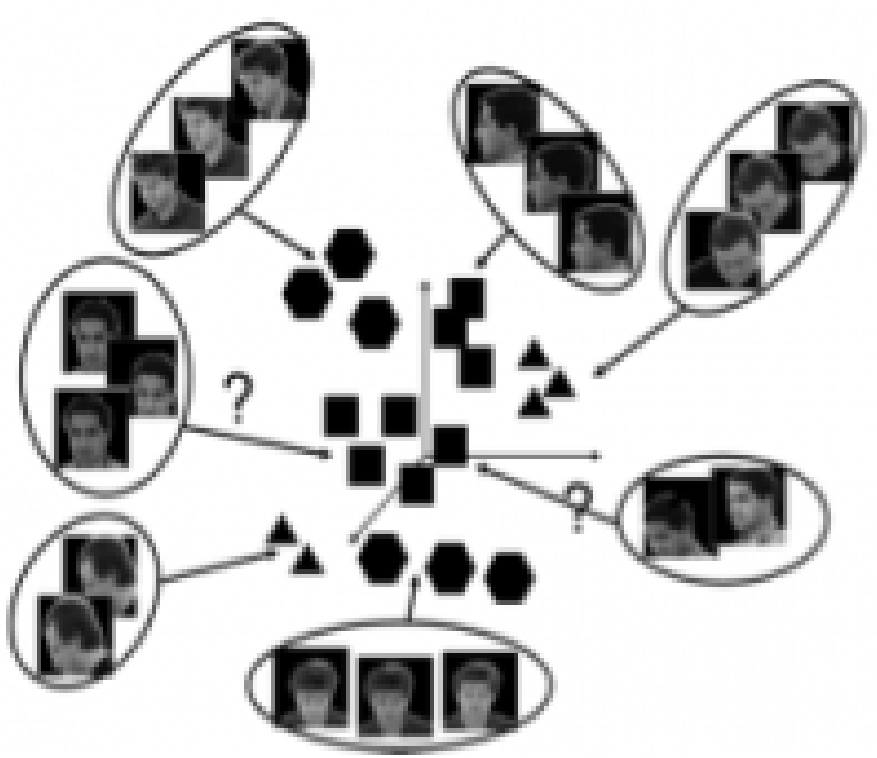


Figure 1.1. Row $\mathbf{x}_1$ as a point and vector in (a) $\mathbb{R}^2$ and (b) $\mathbb{R}^3$.

Implications

- Each photo in the universe is some point in high dimension
- Each book (written or in future) are some point in high dimension



Data Types

Ben Shneiderman, 1996:[1]

- 1D (sequences)
- Temporal
- 2D (maps)
- 3D (shaped)
- nD (relational)
- Trees (hierarchical)
- Networks (graphs)
- Others (text)

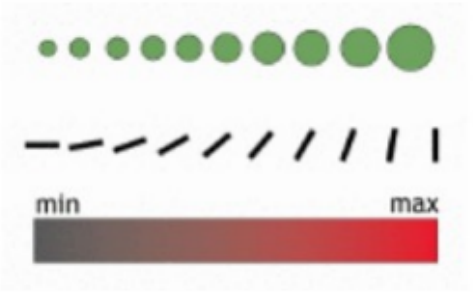
[1] The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization [Shneiderman, 96]

Semantics vs. Types

- Data Semantics: real-world meaning
  - e.g., company name, day of the month, person height, etc.
- Data Type: Interpretation in terms of scales of measurements
  - e.g., quantity or category, sensible mathematical operations etc.

Data Types

- Nominal (Categorical) (N)
  - Are = or ≠ to other values
  - Apples, Oranges, Bananas,...
- Ordinal (O)
  - Obey a < relationship
  - Small, medium, large
- Quantitative (Q)
  - Can do arithmetic on them
  - 10 inches, 23 inches, etc.



On the theory of scales and measurements [S. Stevens, 46]

Data Types

- Q - Interval (location of zero arbitrary)

  Dates: Jan 19; Location: (Lat, Long)

  Like a geometric point. Cannot compare directly.

  Only differences (i.e., intervals) can be compared

- Q - Ratio (zero fixed)

  Measurements: Length, Mass, Temp, ...

  Origin is meaningful, can measure ratios & proportions

  Like a geometric vector, origin is meaningful

Data Types

- N - Nominal (labels)
  - Operations: $=, \neq$
- O - Ordinal (ordered)
  - Operations: $=, \neq, >, <$
- Q - Interval (location of zero arbitrary)
  - Operations: $=, \neq, >, <, +, -$
- Q - Ratio (zero fixed)
  - Operations: $=, \neq, >, <, +, -, \times, \div$

Quiz!

What is the data type of:

- Gender:

Quiz!

What is the data type of:

- Gender: Categorical/Nominal
- Age:

Quiz!

What is the data type of:

- Gender: Categorical/Nominal
- Age: Ordinal
- Height:

Quiz!

What is the data type of:

- Gender: Categorical/Nominal
- Age: Ordinal
- Height: Quantitative - Ratio
- Date:

Quiz!

What is the data type of:

- Gender: Categorical/Nominal
- Age: Ordinal
- Height: Quantitative - Ratio
- Date: Quantitative - Interval

Data Dimensions

- Univariate (1D)
- Bivariate (2D)
- Trivariate (3D)
- Multivariate (nD)

Introduction To Data Visualization

Visualization Goals

- **Presentation**
  - Known facts about data
  - Task: Communicate results
- **Exploration**
  - Data without hypothesis
  - Task: Generate hypothesis
- **Confirmation**
  - Hypothesis is given
  - Task: Verify / falsify hypothesis

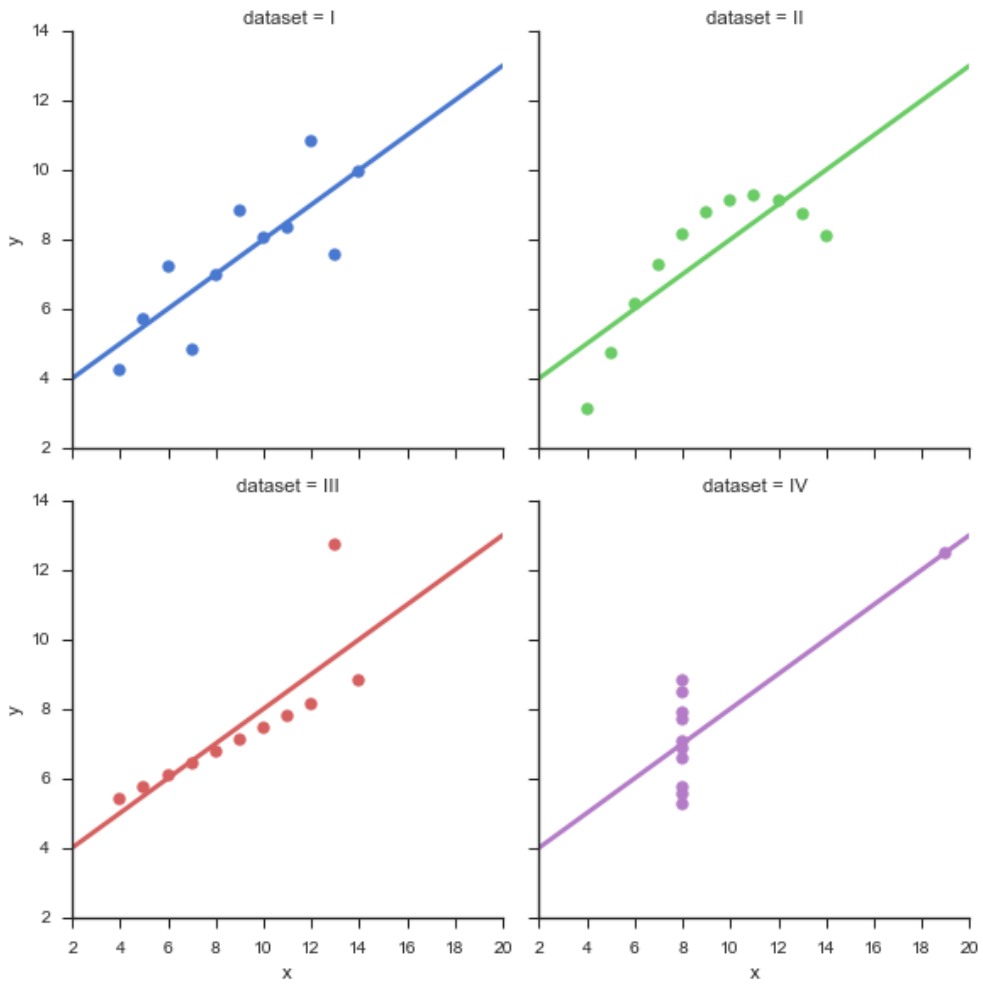
Visualization Goals

"The greatest value of a picture is when it forces us to notice what we never expected to see."

-John Tukey (1915 - 2000)

Anscombe's Quartet

Same mean, variance, correlation, and linear regression line

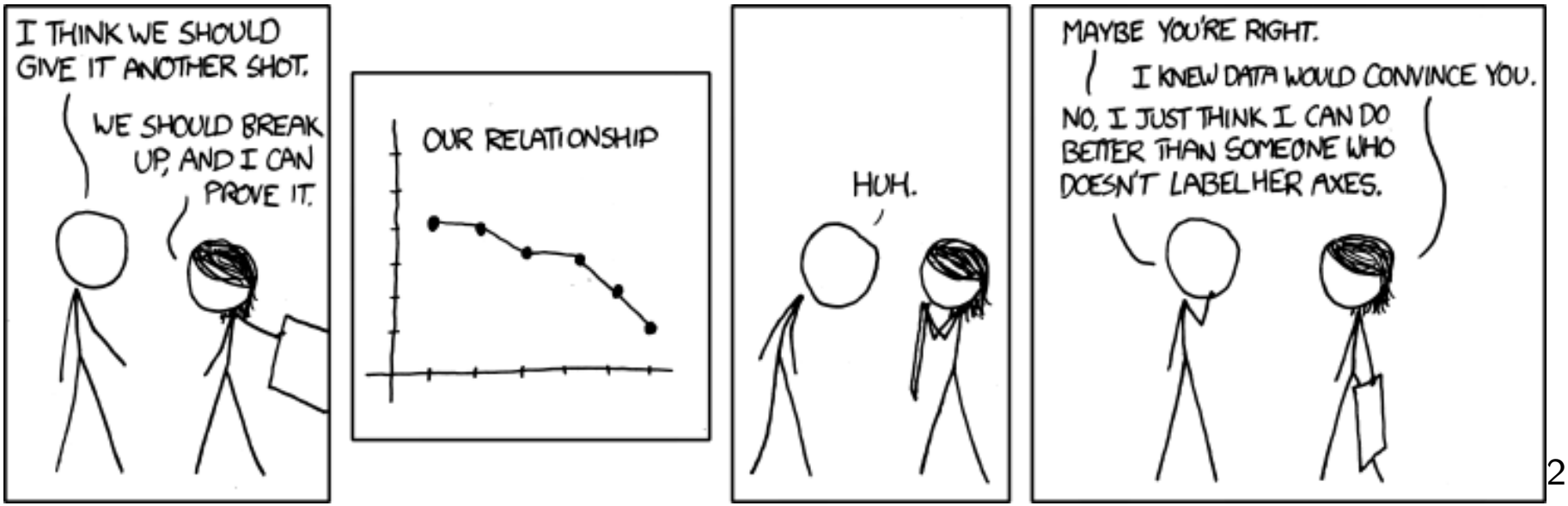


Graphical Integrity

"There are three kinds of lies: lies, damned lies, and statistics."

- attributed to Benjamin Disraeli in 19th Century

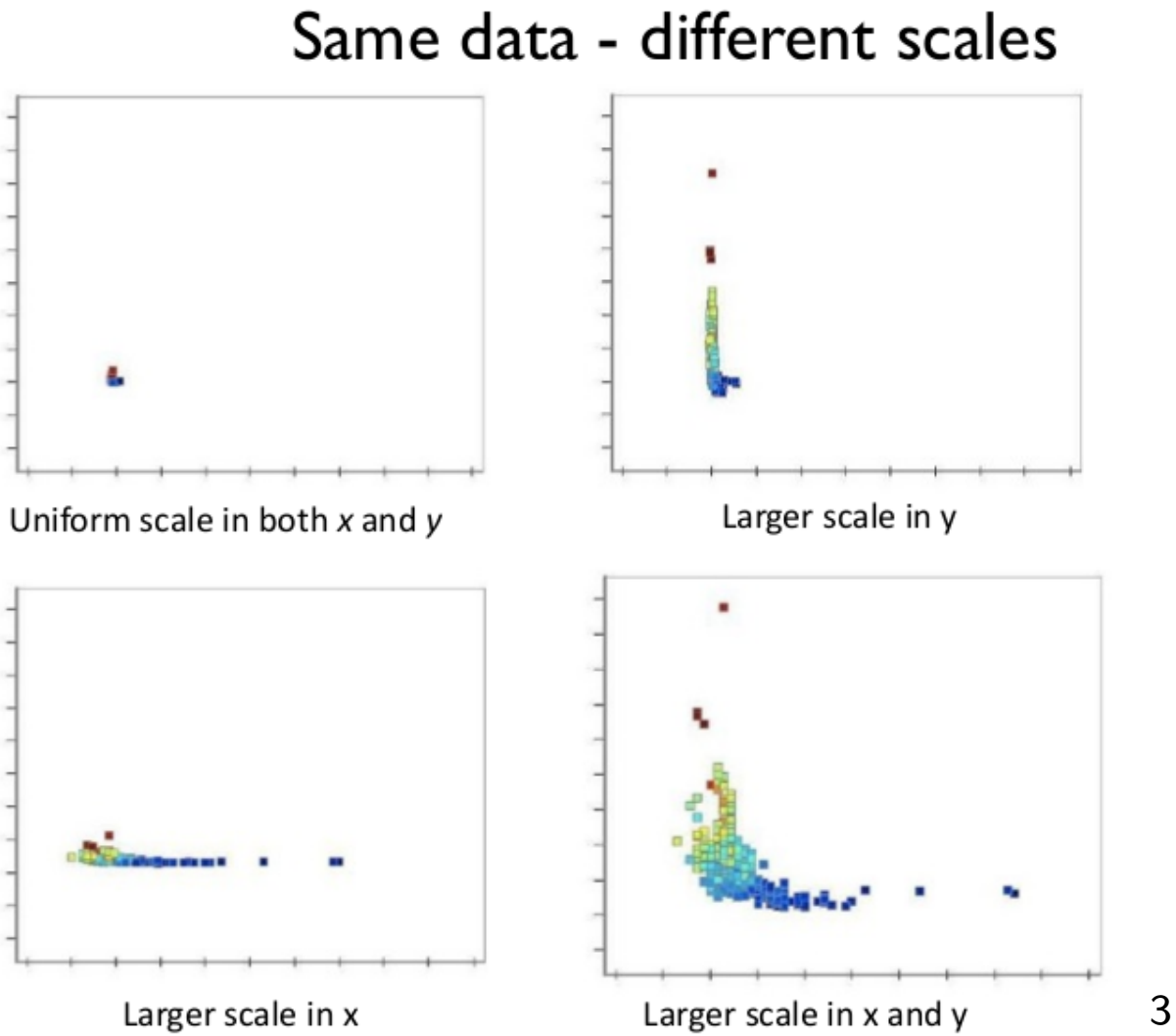
Labelling Chart Axes



[2]

[2] http://xkcd.com/833/

Lying with Scales



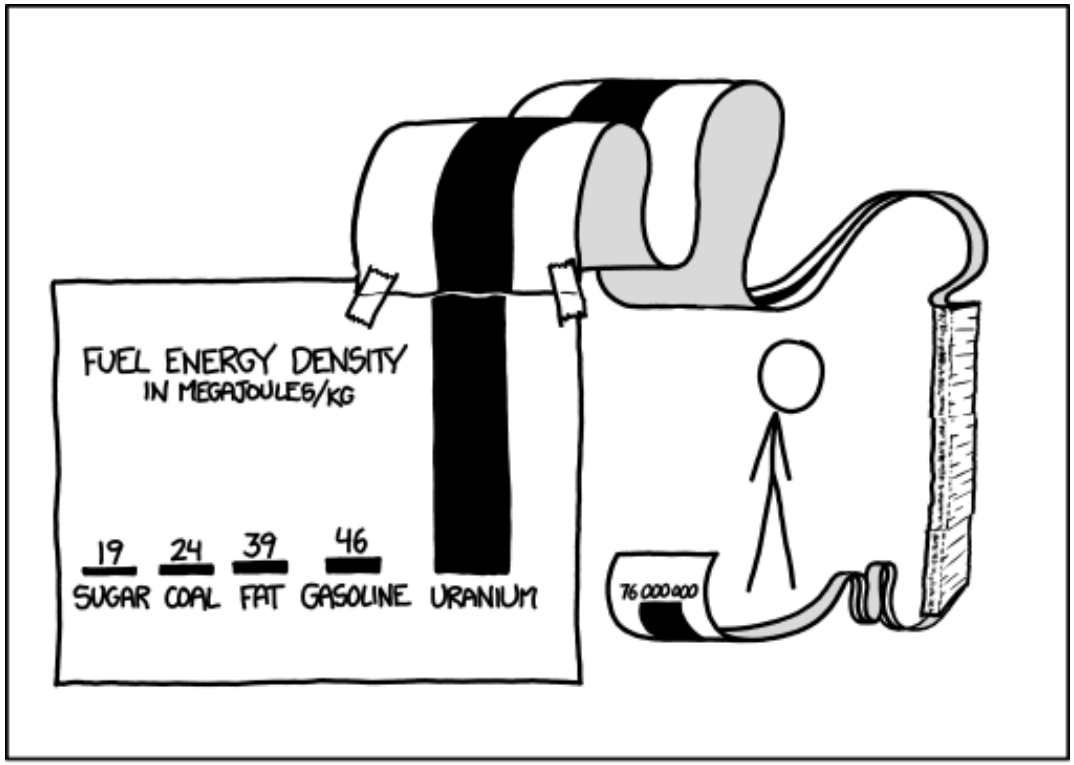
[3]

[3] Ward, Grinstein, Keim, 2011

Scales are Critical!

- What are your bounds upper and lower?
- What scale works? - Linear? Log? Clipping? Breaks?
- Relative or absolute values?
- How can you make things comparable?

Log Scale



SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY.*

4

[4] http://xkcd.com/1162/

Graph Types (2D and nD)

Summary

Major Concepts:

- Data mining Terminology
- Visualization basics
- Graphical Integrity
- Graph Types for 2D and nD

Slide Material References

- Slides from Harvard CS 109 (2013 and 2014)