Rensselaer

why not change the world?®

Data Analysis – Part II

Ahmed Eleish

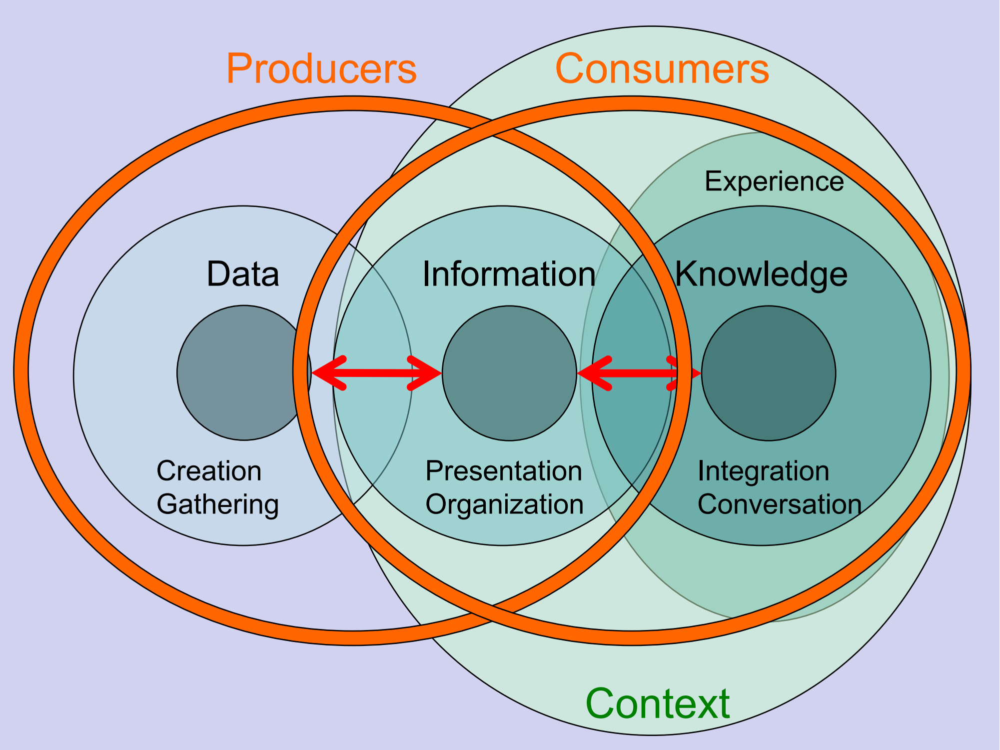
Data Science – ITWS/CSCI/ERTH-4350/6350 Module 4, October 10th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

Contents

- Data Analysis I review
- Errors and uncertainty…
- Visualization as an information tool and analysis tool
- New visualization methods (new types of data)
- Use, citation, attribution and reproducibility

Data-Information-Knowledge Ecosystem

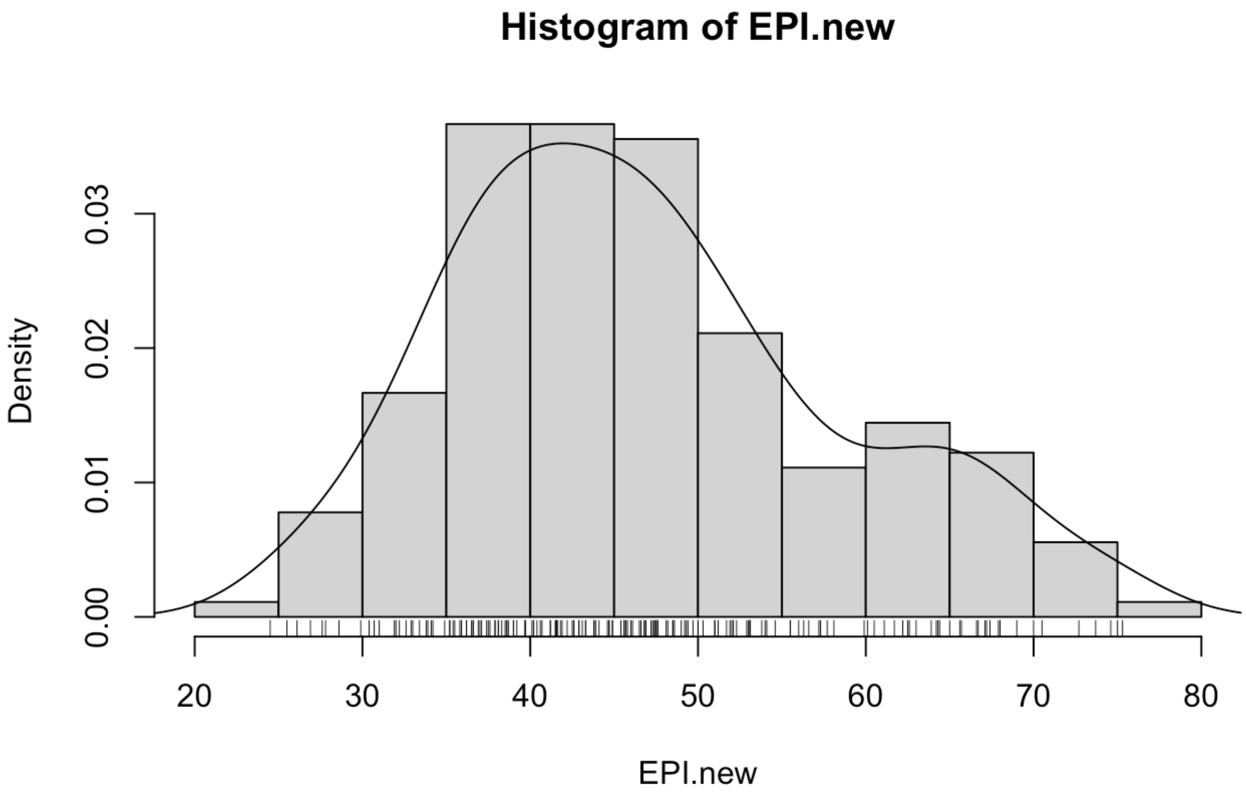


Review: Data Analysis I

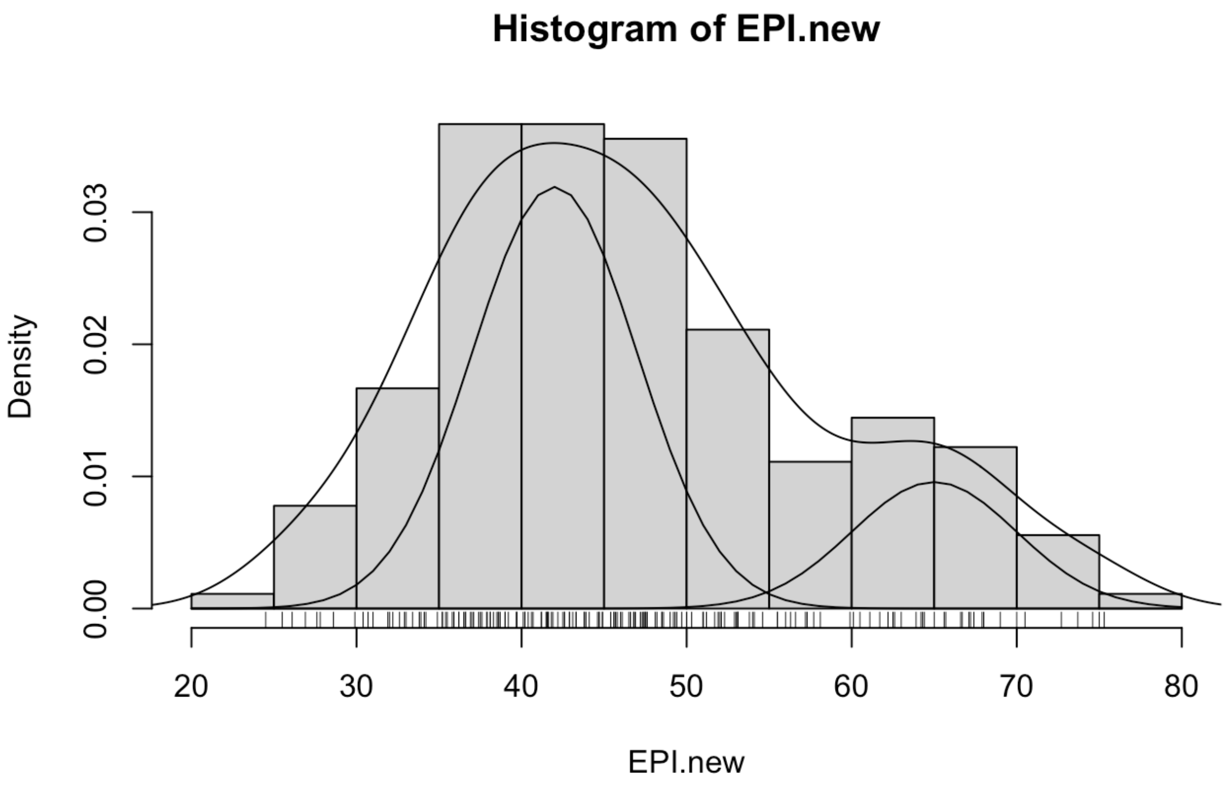
Types of Data

Type of data	Level of measurement	Examples
Categorical	**Nominal** (no inherent order in categories)	Eye colour, ethnicity, diagnosis
	Ordinal (categories have inherent order)	Job grade, age groups
	Binary (2 categories – special case of above)	Results of some tests, e.g. positive/negative
Quantitative (Interval/Ratio) (NB units of measurement used)	Discrete (usually whole numbers)	Size of household **(ratio)**
	Continuous (can, in theory, take any value in a range, although necessarily recorded to a predetermined degree of precision)	Temperature °C/°F (no absolute zero) **(interval)** Height, age **(ratio)**

Histogram – type of bar plot



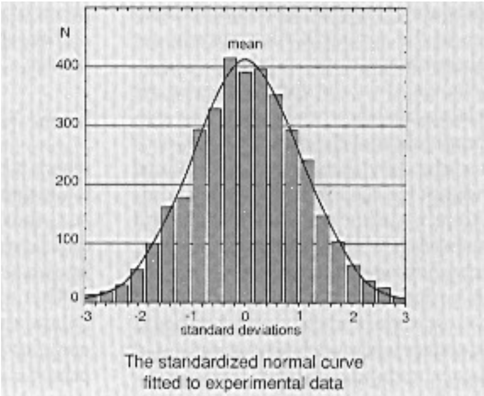
Histogram – type of bar plot



Statistics

• We will most often use a Gaussian distribution (*aka* normal distribution, or bell-curve) to describe the statistical properties of a group of measurements.

• The variation in the measurements taken over a finite spatial region may be caused by intrinsic spatial variation in the measurement, by uncertainties in the measuring method or equipment, by operator error, ...



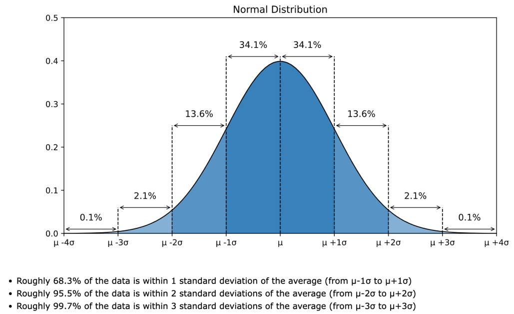


Image Credit: W3C school:
https://www.w3schools.com/statistics/statistics_normal_distribution.php

Mean and standard deviation

•The mean, *m*, of *n* values of the measurement of a property *z* (the average).

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

•The standard deviation *s* of the measurements is an indication of the amount of spread in the measurements with respect to the mean.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}, \text{ where } \mu = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

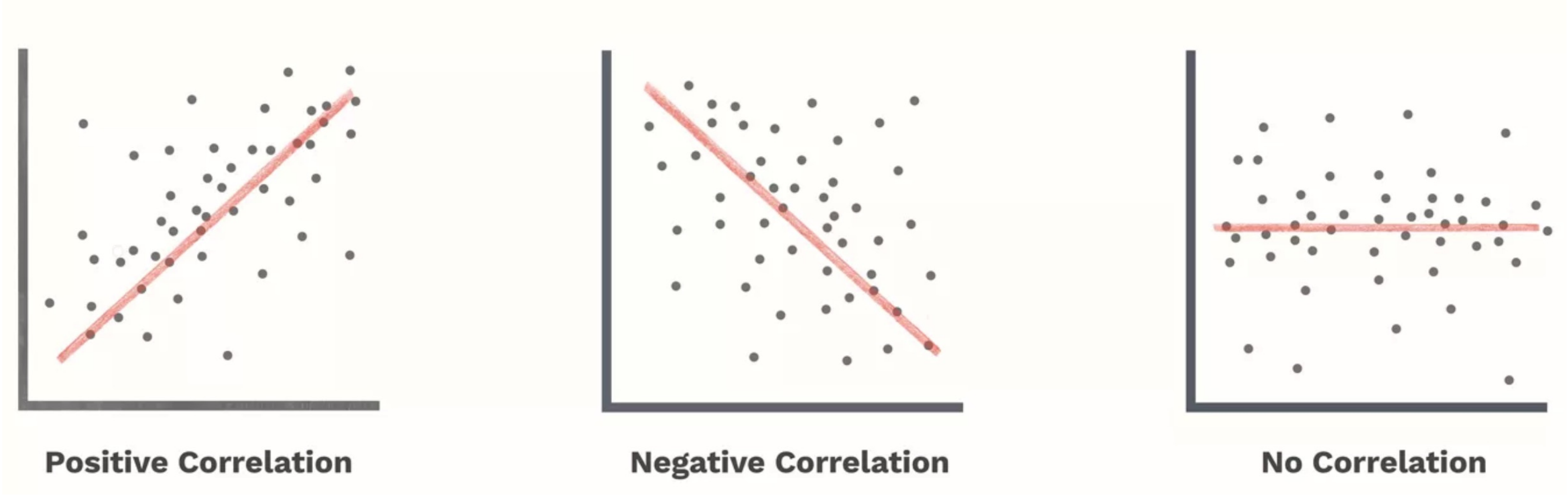
• The quantity σ^2 is known as the variance of the measurements.

Correlation

- One measure of the strength of the association between two numerical variables is correlation.
- Correlation describes the strength of the linear association between two variables.
- Correlation coefficient is between -1 and +1
- -1 indicates a perfect negative linear association and +1 indicates a perfect positive linear association. The correlation coefficient of 0, indicates that there is no linear relationship in the two variables. -0.1 and +0.1, indicates no linear relationship *or a very weak* linear relationship
- Correlation coefficient is sensitive to outliers.
- Correlation coefficient is unitless.

Reference(s): https://www.investopedia.com/terms/c/correlationcoefficient.asp
https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp

Correlation...



Image/Photo Credit: https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp

Input/Output

- Input: input go by different names,

input: *features, predictors*, *independent variables*, sometimes just variables

$$X = (x_1, x_2, \ldots, x_n)$$

- **Output**: The output variable called *response* or *dependent variable*, typically denoted by Y

- Suppose that we observe quantitative response Y with *p* different predictor variables,

$x_1, x_2, \ldots x_p$.

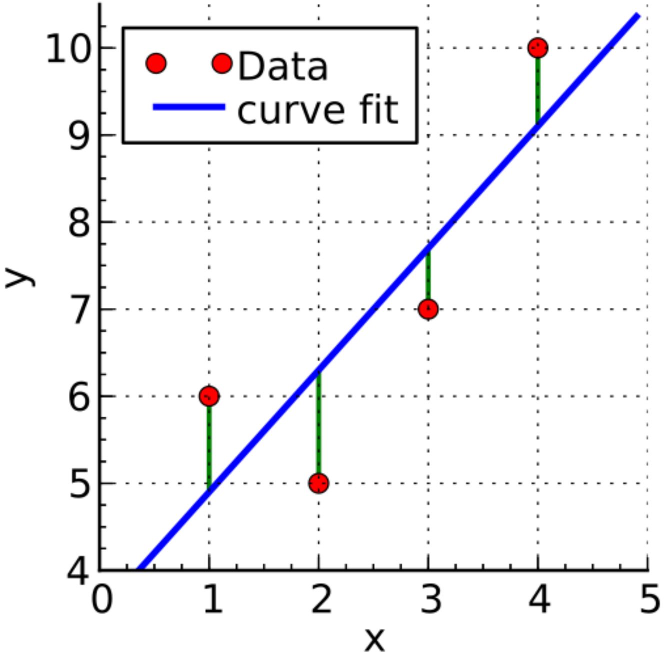
- We assume some relationship between Y and $X = (x_1, x_2, \ldots x_p)$, which can be written as:

$$Y = f(x) + \varepsilon$$

f is an unknown function of x

random error term, which is independent of x

Regression



Simple Linear Regression

- The most commonly used approach is the *Least Squares*
- Least Squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Predicted response

Intercept

Slope

Explanatory variable

- $\hat{y}$ = Predicted value of the response variable
- x = Explanatory variable (x)
- $\hat{\beta}_0$ = Intercept
- $\hat{\beta}_1$ = Slope

Residuals ...

• The residual is defined as the difference between the observed value and the predicted value.(Difference between the observed value and the predicted value of the response variable for a given data point).

$$e_i = y_i - \hat{y}_i$$

represents the *ith* residual,

this is the difference between the *ith* observed response value and the *ith* response value that is predicted by the linear model.

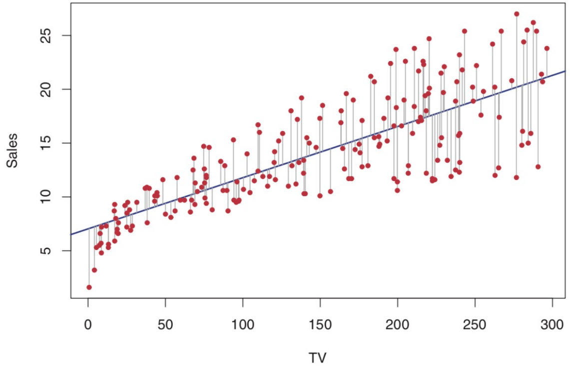
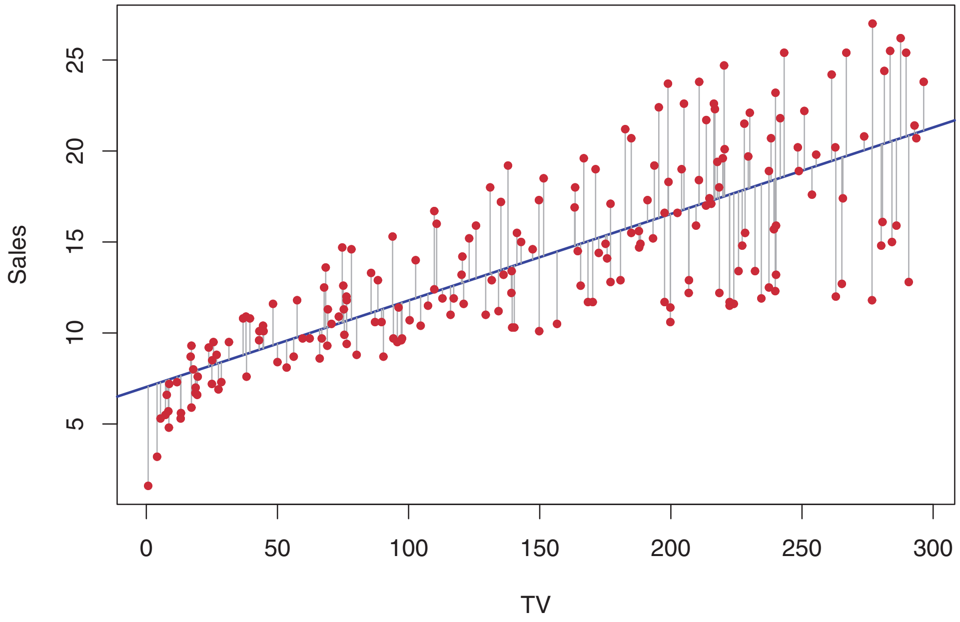


Image Credit: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 3 – Linear Regression

Reference: Introduction to Statistical Learning with Applications in R, 7th Edition, Chapter 3 - Linear Regression

Linear Model

- Sales vs. TV ad spending
- Sales in 1000s of units
- TV ad spending in 1000s of $



Evaluating the Linear Model

Values of coefficients >> their Std. errors

High t-statistic

Very low p-value

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Hypothesis (more TV ads → more sales)

H0 : There is no relationship between X and Y

Ha : There is some relationship between X and Y

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{\mathrm{s.e.}(\hat{\beta})}$$

Reject the null hypothesis!

Data Analysis II

Errors & Uncertainty

Errors

- Personal errors are mistakes on the part of the experimenter. It is your responsibility to make sure that there are no errors in recording data or performing calculations
- Systematic errors tend to decrease or increase all measurements of a quantity, (for instance all of the measurements are too large). E.g. calibration
- Random errors are also known as statistical uncertainties, and are a series of small, unknown, and uncontrollable events

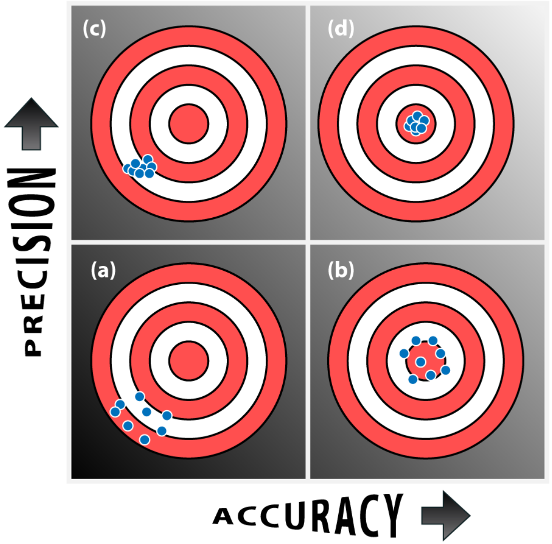
Errors

- Statistical uncertainties are much easier to assign, because there are rules for estimating the size

e.g. If you are reading a ruler, the statistical uncertainty is half of the smallest division on the ruler. Even if you are recording a digital readout, the uncertainty is half of the smallest place given. This type of error should always be recorded for any measurement

Standard measures of error

- Absolute deviation
  - is simply the difference between an experimentally determined value and the true value
- Relative deviation
  - is a more meaningful value than the absolute deviation because it accounts for the relative size of the error. The relative percentage deviation is given by the absolute deviation divided by the true value and multiplied by 100%
- Standard deviation

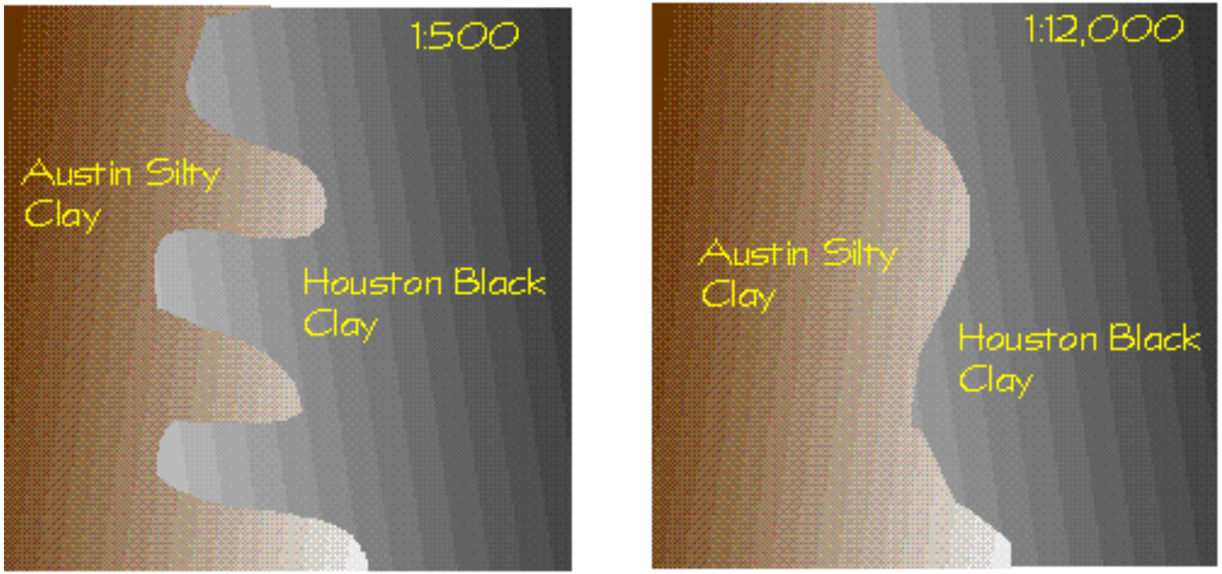


Some considerations

- Possibly more important than our answer is our confidence in the answer.
- Our confidence is quantified by uncertainties.
- Once we combine numbers, we need to be able to assess how the uncertainties change for the combination.
- This is called **propagation of errors** or more correctly the propagation of our understanding/ estimate of errors in the result we are looking at…

Resolution

Accuracy and Generalization

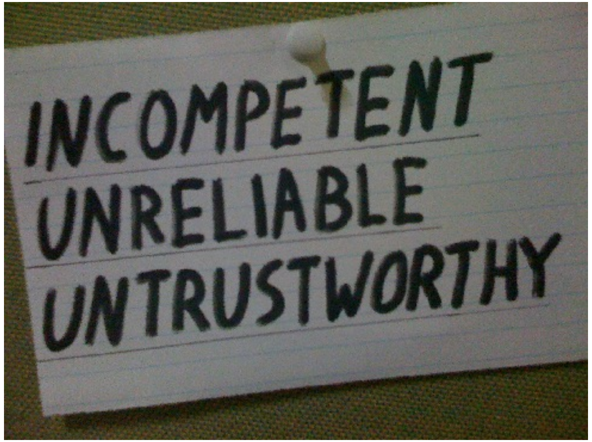


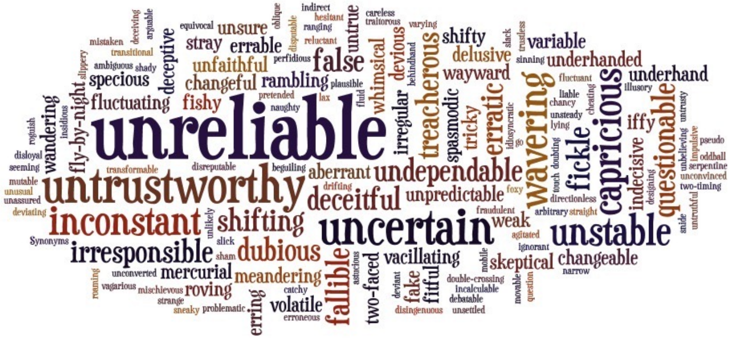
Actual soil interdigitation | Generalization on map

Different soil type boundaries are generalized when mapping an area, but are actually vague and graduated. Differences in scale allow finer resolution, but only if the original data was collected at a finer resolution.

Reliability

- Changes in data over time
- Non-uniform coverage
- Map scales
- Observation density
- Sampling theorem (aliasing)
- Surrogate data and their relevance
- Round-off errors in computers





Propagating errors

- This is an unfortunate term – it means making sure that the result of the analysis carries with it a calculation (rather than an estimate) of the error

- E.g. if C=A+B (your analysis), then ∂C=∂A+∂B
- E.g. if C=A-B (your analysis), then ∂C=∂A+∂B!

- It's not as simple for other calcs.

- When the function is not merely addition, subtraction, multiplication, or division, the error propagation must be defined by the total derivative of the function.

Error propagation

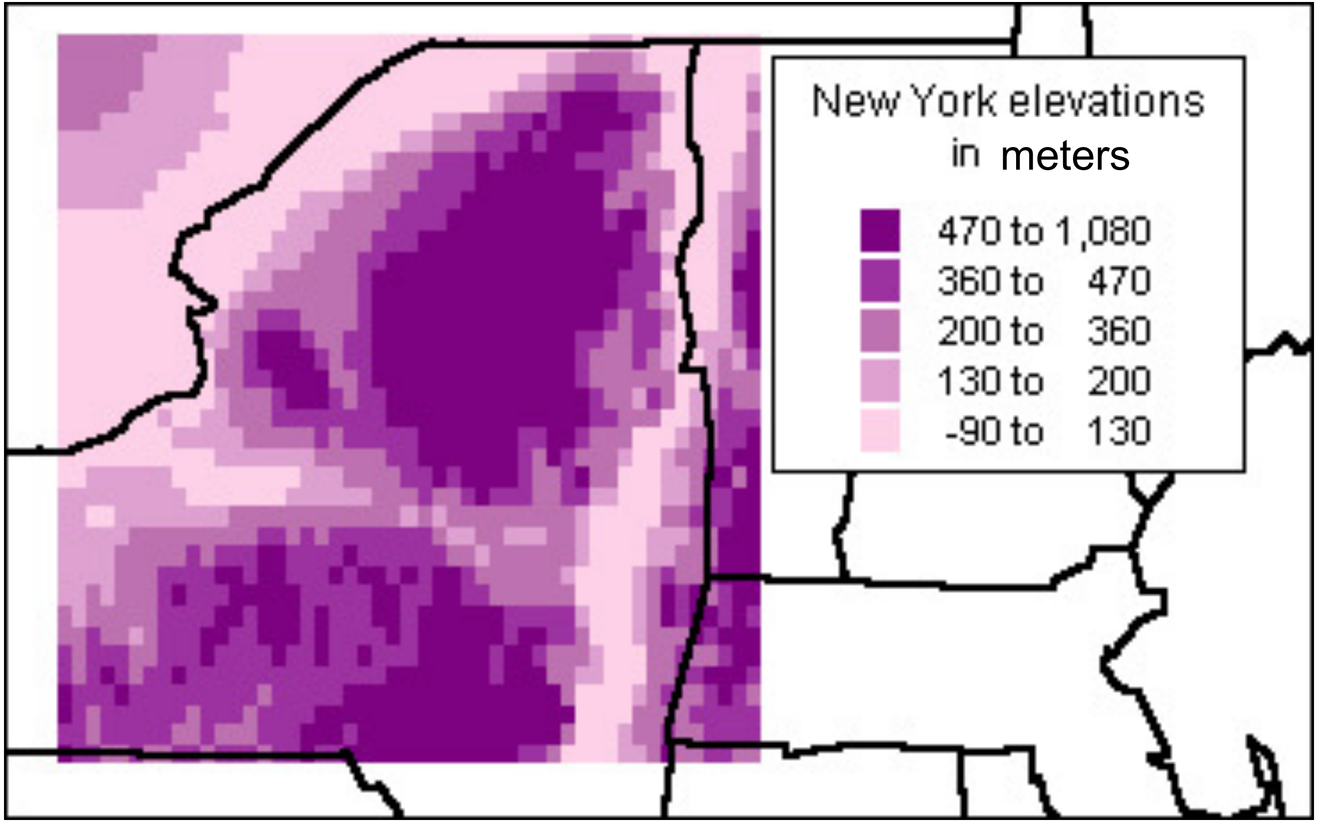
- Errors arise from data quality, model quality and data/model interaction.
- We need to know the sources of the errors and how they propagate through our model.
- Simplest representation of errors is to treat observations/attributes as statistical data – use mean and standard deviation.

Dealing with errors

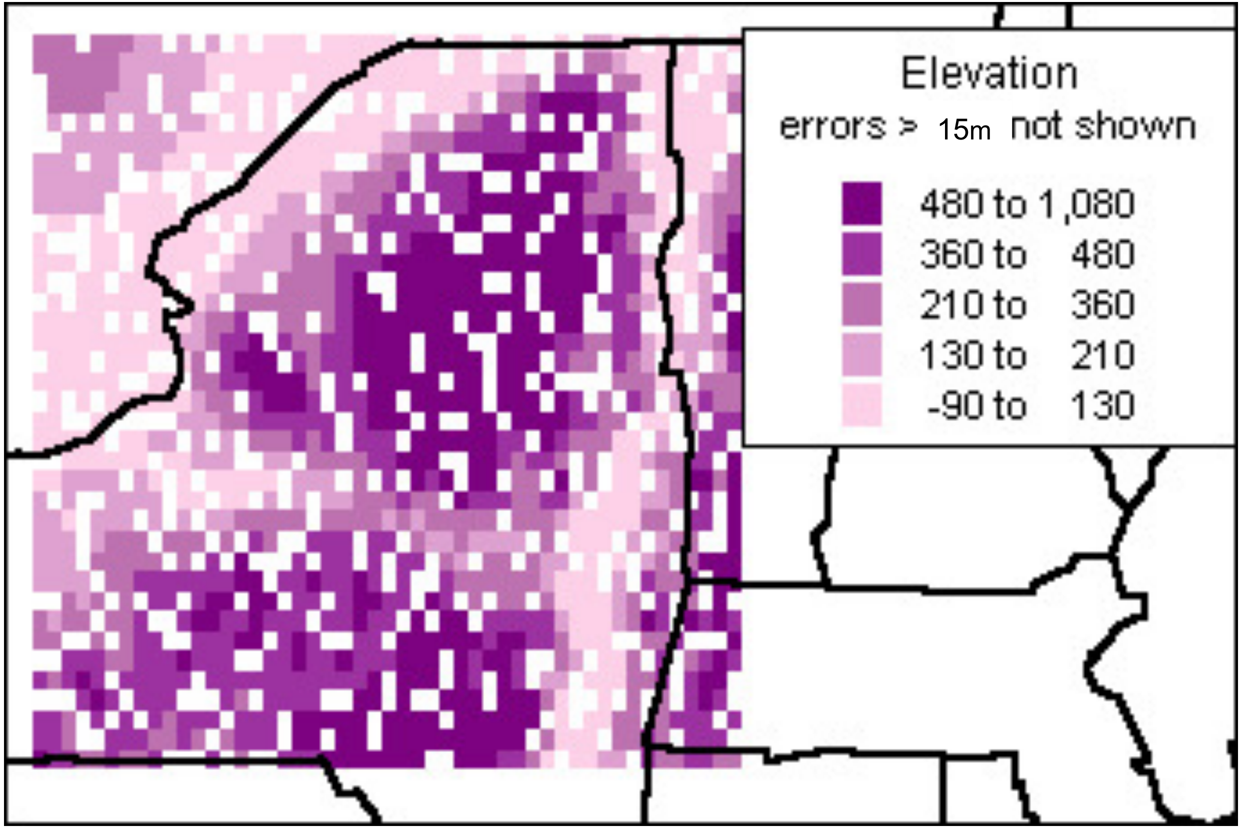
- In analyses:
  - report on the statistical properties
  - does it pass tests at some confidence level?

- On maps:
  - exclude data that are not reliable (map only subset of data)
  - show additional map of some measure of confidence

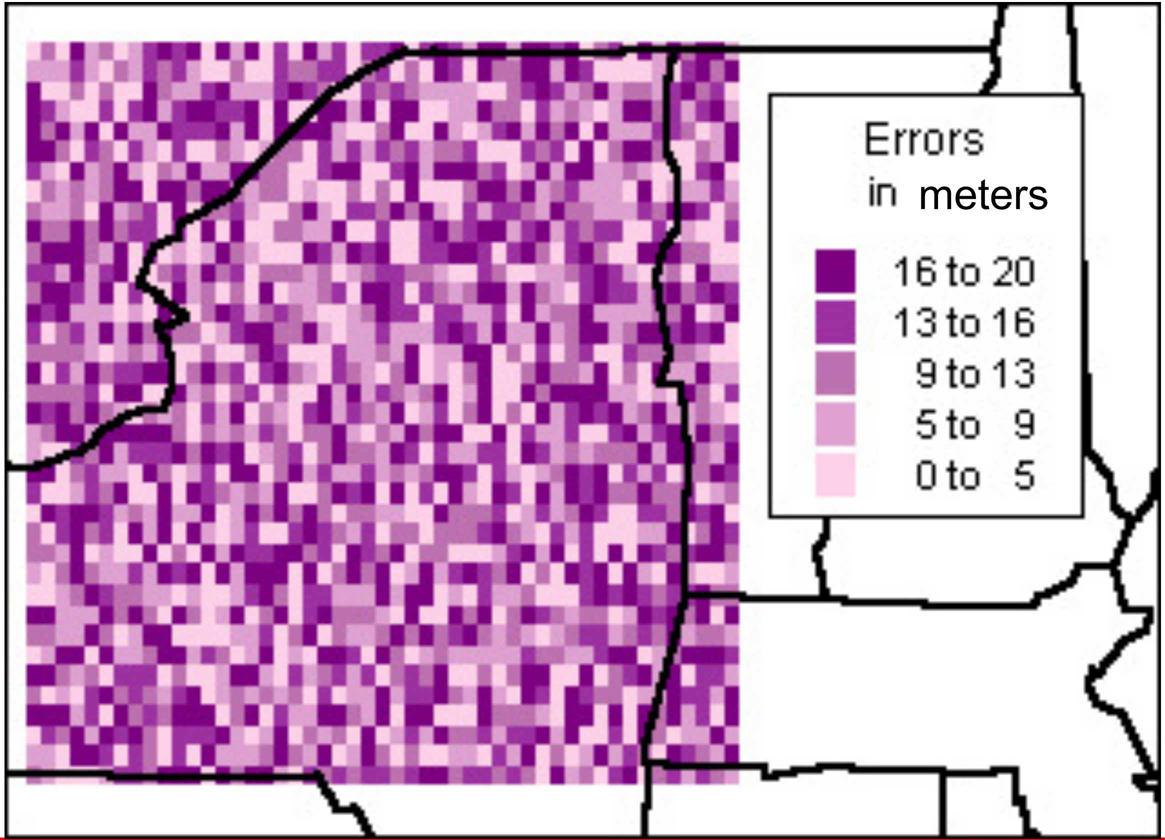
Elevation map



Larger errors ‘whited out’



Elevation errors



Reporting results/ uncertainty

- Consider the number of significant digits in the result which is indicative of the certainty of the result
- The number of significant digits depends on the measuring equipment you use and the precision of the measuring process - do not report digits beyond what was recorded
- The number of significant digits in a value defines the precision of that value

Reporting results…

- In calculations, it is important to keep enough digits to avoid round off error.
- In general, keep at least one more digit than is significant in calculations to avoid round off error
- It is not necessary to round every intermediate result in a series of calculations, but it is very important to round your final result to the correct number of significant digits.

Uncertainty

- Results are usually reported as result ± uncertainty (or error)
- The uncertainty is given to one significant digit, and the result is rounded to that place
- For example, a result might be reported as 12.7 ± 0.4 m/s^2. A more precise result would be reported as 12.745 ± 0.004 m/s^2. A result should not be reported as 12.70361 ± 0.2 m/s^2
- Units are very important to any result

Secondary analysis

- Depending on where you are in the data analysis
- Having a clear enough awareness of what has been done to the data (either by you or others) prior to the next analysis step is very important – it is very similar to sampling bias
- Read the metadata (or create it) and documentation

Visualizations

Considerations for visualizations as analysis

- What is the improvement in the understanding of the data as compared to the situation without visualization?

- Which visualization techniques are suitable for one's data?
  - e.g. Are direct volume rendering techniques to be preferred over surface rendering techniques?

Why visualization?

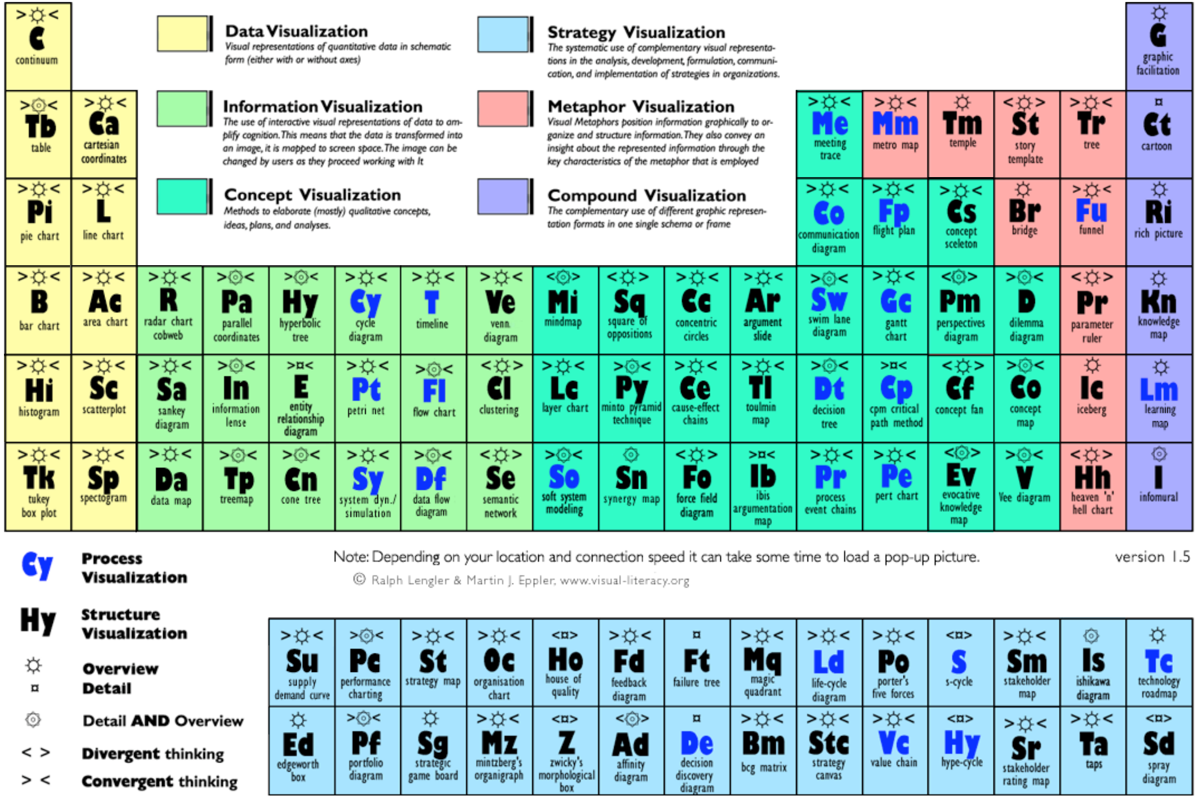
- Reducing amount of data, quantization
- Patterns
- Features
- Events
- Trends
- Irregularities
- Leading to presentation of data, i.e. information products
- *Exit points for analysis*

Types of visualization

- Color coding (including false color)
- Classification of techniques is based on
    - Dimensionality
    - Information being sought, i.e. purpose
- Line/scatter/bar plots
- Networks
- Contours
- Volume rendering techniques
- Animation techniques
- Non-realistic, including ‘cartoon/ artist’ style

Remember – metadata!

A PERIODIC TABLE OF VISUALIZATION METHODS



https://www.visual-literacy.org/periodic_table/periodic_table.html

Managing visualization products

- The importance of a ‘self-describing’ product
- Visualization products are not *just* consumed by people
- How many images, graphics files do you have on your computer for which the origin, purpose, use is still known?
- How are these logically organized?

Use, citation, attribution

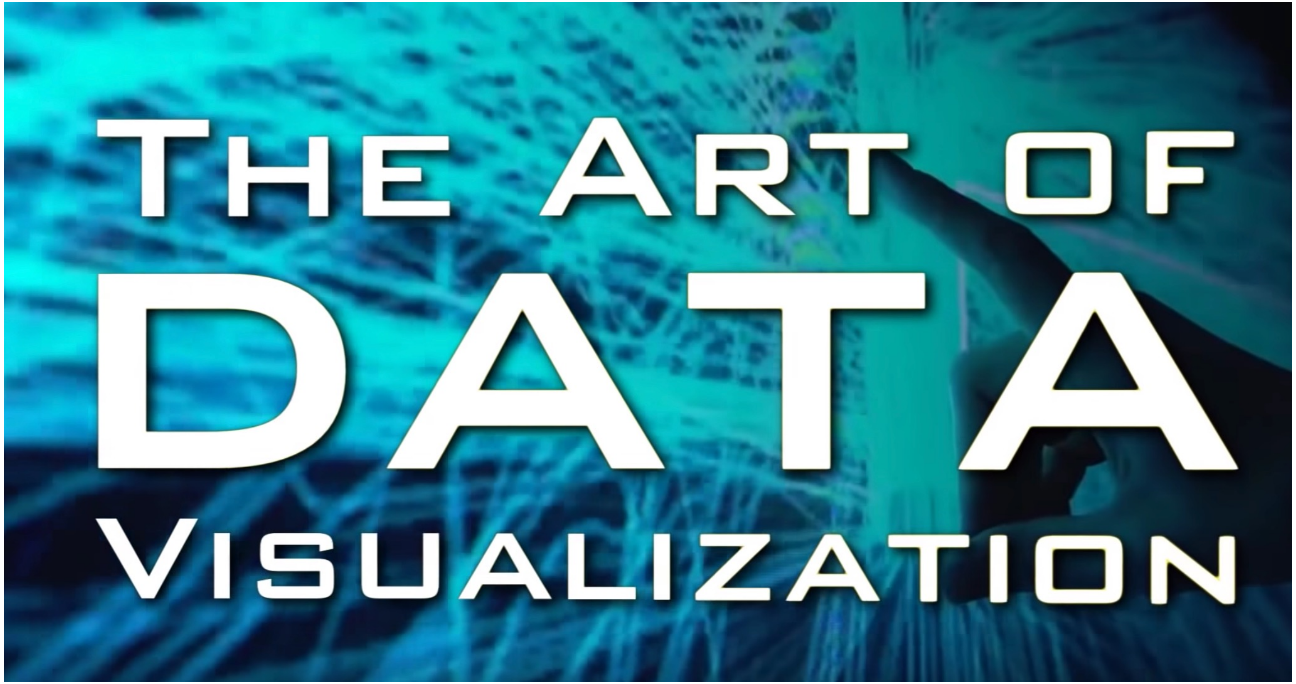
- Think about and implement a way for others (including you) to easily use, cite, attribute any analysis or visualization you develop
- This *must* include suitable connections to the underlying (aka backbone) data – and note this may not just be the full data set!
- Naming, logical organization, etc. are key
- Make them a resource, e.g. URI / URL

See http://commons.esipfed.org/node/308

Reproducibility

- The documentation around procedures used in the analysis and visualization are very often neglected – DO NOT make this mistake
- Treat this *just* like a data collection (or generation) exercise
- Follow your management plan
- Despite the lack or minimal metadata/ metainformation standards, capture and record it
- Get someone else to verify that it works

Motivation: Art of Data Visualization



https://www.youtube.com/watch?v=AdSZJzb-aX8

Thanks!

Form your teams!