

Meaning of Data, Information, Knowledge, and Wisdom in Data Analytics

(source: by Gene Bellinger, Durval Castro, Anthony Mills at <http://www.systems-thinking.org/dikw/dikw.htm>)

- According to Russell Ackoff, a systems theorist and professor of organizational change, the content of the human mind can be classified into a hierarchy of categories:

1. **Data:** symbols (either numeric or iconic)

Data is raw. It simply exists and has no significance beyond its existence. It can exist in any form, usable or not.

2. **Information:** data that are processed to be useful; I.e., data that are placed into a context (relationship to similar data) providing answers to "who", "what", "where", "how much" and "when" questions

Information is data that has been given meaning by way of relational connection, such that $x_1 < x_2$ or x_1 preceeds x_2 .

3. **Knowledge:** application of data and information; answers "**how**" questions leading to some causal statements.

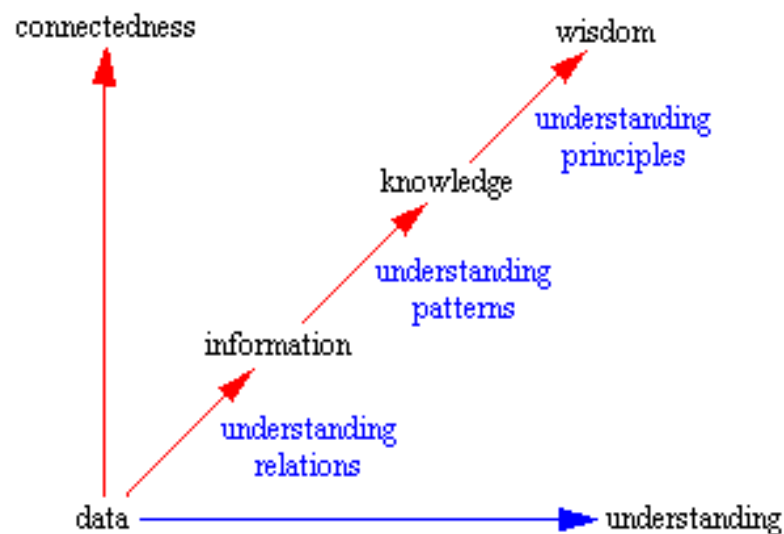
Knowledge is the appropriate connection of information, such that when X increases Y will also increase.

4. **Wisdom:** Construction of individual pieces of knowledge into a “*general*” theory including the explanation of apparent contradictions. A theory should be general enough to allow *predictions* based on new observations.

Developing an encompassing perspective of understanding real world events.

Note: “Big data” exploration and mining cannot improve our wisdom, it just provides a description of observed patterns in given data (see **BIGDATAANDSTATISTICS.PDF**).

- According to Gene Bellinger, Durval Castro, and Anthony Mills “understanding” connects all levels:



Science advances by

- [a] *uncovering meaningful patterns* within previously noisy information or
- [b] *explaining the mechanism and conditions* under which patterns emerge,

[b] *revising current knowledge* so it can accommodate in a unified framework (a theory) previously contradicting information.

Statistics provides tools that help us uncover patterns in observed data and test hypotheses about their *underlying data generating process*.

Philosophy of Scientific Inference (skipped)

(Based on Kenneth J. Rothman and Sander Greenland, 1998. “Chapter 2: Causation and Causal Inference” in *Modern Epidemiology*, 2nd edition, Lippincott Williams & Wilkins, Philadelphia. pp 7-28)

Objectives of Science:

- Develop a *logical structure* for the objects and events of the world surrounding us. This logical structure erects a *theoretical framework* that relates objects and events with each other and explains why these relationships exist.
- This logical structure should be *generally* applicable. The goal of the logical structure is to *understand* universal cause-effect relationships, *predict* and *control* objects and events.
- *Hypothesis* are derived from theory and establish *predictions* which can be tested against given empirical observations.

Experimental versus *Empirical* Research:

- One way to acquire knowledge is to conduct *controlled experiments* by systematically *varying a condition* that is a potential *cause* for a specific *effect*, while
 - [1] *holding constant* all other potential conditions or
 - [2] *randomly spreading* other potential conditions so that their influences onto an effect cancel each other out.

Theory in experimental studies allows us to *make predictions* about the experiment's outcome. If the observations match the prediction then the theory is *tentatively* supported, otherwise it needs to be *refuted*.

- We have only limited capabilities in *empirical research* to *control the variability* of potential causes as well as the initial conditions.

⇒ We have to use a small set of the empirical observations as they come available. There is no control over variability.

However, by carefully *designing an empirical study* and its associated data sampling scheme, we still can *control for biases* which are induced by unknown external circumstances.

Definition: Biases emerge from unknown external circumstances that systematically influence the observed data.

Bias can be controlled either by *randomization* of sample observations (the systematic influences are dispersed over the sample in the hope that they cancel each other out)

Alternatively it can be controlled in *case-control studies* by *stratified sampling* or *randomized controlled trials*.

- In addition, we may end up with only “*one*” *sample* (e.g., one specific census, one map pattern). However, by adopting the *process based perspective of a super-population* this one observation may also be conceived as a random sample from that hypothetical underlying process.

Importance of Variability in the Sciences

- Without variability we no uncertainty would exists. The larger the variability the higher the uncertainty.
- Suppression of variability changes perception of a causal mechanism:
For example: Yellow shanks (part of leg), a characteristics occurrence in a certain *genetic strain* of fowl, if it feds on *yellow corn*.

Two perceptions of causes due to *suppressed variability*:

- Genetic Explanation (no variation in feed): Fowl just fed only yellow corn → then the presence or absence of the gene determines yellow shanks.
- Environmental Explanation (no variation in genetics): Fowl with just the specific genetic strain is fed yellow corn or other corn → the corn determines yellow shanks.

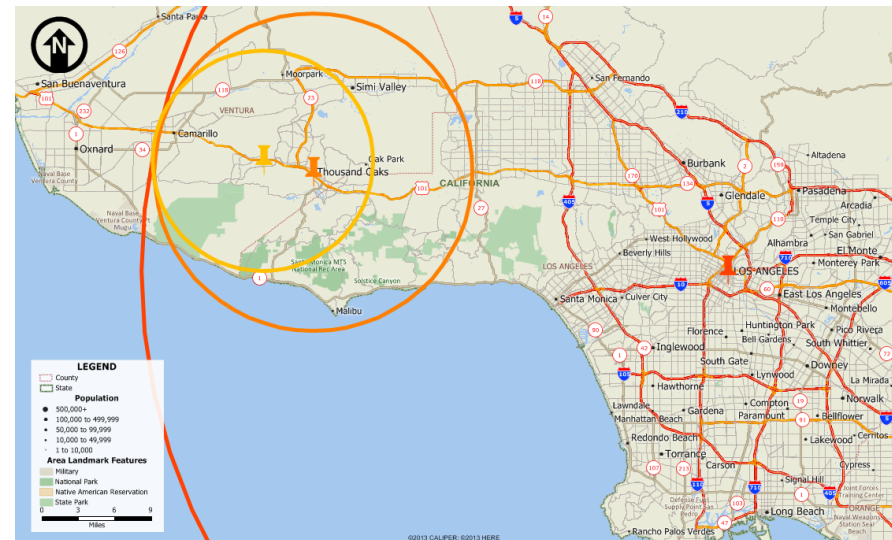
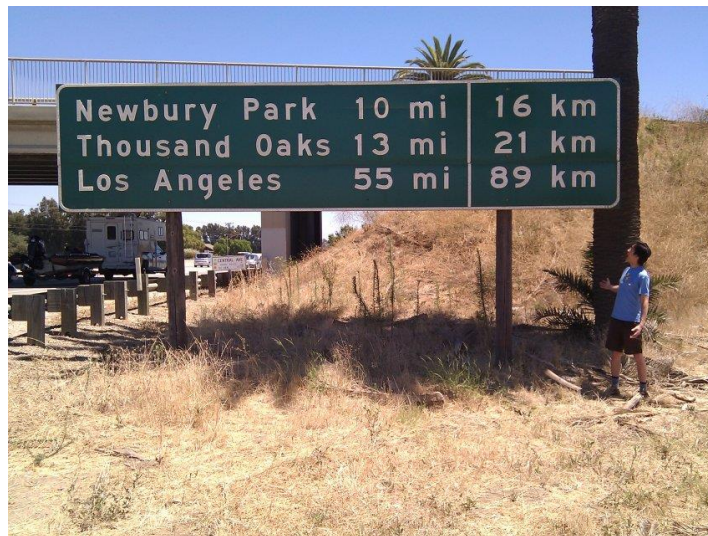
Measurement Theory

- A measurement *numerically* represents a property of a particular object/event and places it on a measurement *scale* (scale has a different meaning here than that of floating point numbers), so it can be *compared to* measurements taken from other objects (step from data → information). This implies that there must be *variability* among the objects.

What conditions need to be present in order to be able to perform measurements:

- Objects are *empirically related* to each other, e.g., person \mathcal{A} is taller than person \mathcal{B} , person \mathcal{A} has brown eyes but person \mathcal{B} does have green eyes.
- *Measurement theory* is concerned with the development of
 1. a *numerical scale* that reflects the empirically observed relationships among the objects and
 2. the assignment of a *numerical representation* to the objects' properties.
- Different assignment rules of numerical values to properties of objects, which *preserve the empirically observed relationships* among the objects, are classed together into a *measurement scale*.

E.g., kilometers and miles share a common measurement scaled




- *Physical properties* are easier to measure than social, economic or psychological *concepts*.

Measurement Scales:

- The measurement scales and their associated relationships are *hierarchically organized*:
 - Higher order scales are *inheriting* the properties of subordinate scales.
 - Progressing from a lower order scale to a higher order scale allows becoming *more specific* in expressing relationships among the measurements.
 - However, moving from a higher order scale to a lower order scale leads to a *loss of information*.
- **Overview over the system of scales:**
 - Nominal scale: Numbers are used only to *distinguish among objects* and label their attributes.

- Allows classifying similar observations
- Applicable mathematical relationships:

$$= \text{ and } \neq$$
- Notes: Categorical and discrete variable, which allows to count the frequency of equal values in a sample or a population.
-  codes categorical variables into **factors**, which group similar objects together.
- Ordinal scale: Numbers are used to place objects in an **order**.
 - Allows ranking observations.
 - Applicable mathematical relationships:

$$=, \neq, < \text{ and } >$$
 - Notes: Sample observations can be ranked. Due to the coarseness of the measurement instrument (lack of precision), some observations may have equal rank (also called *ties*).
- Interval scale: Scale on which **equal numerical intervals** between objects represent equal differences.
 - A fixed step length is defined, which is equal at each location on the scale.
 - Applicable mathematical relationship:

$$=, \neq, <, >, \text{ and transformation } y = b_0 + b_1 \cdot x$$
 with $b_1 > 0$ (the fixed distance unit/step length) and an arbitrary point of origin b_0 on the scale.
 - Notes: Equal difference allows performing summations and subtractions of values, but not size comparisons.

- Ratio scale: Scale with a **true zero** point allowing to size differences by ratios.

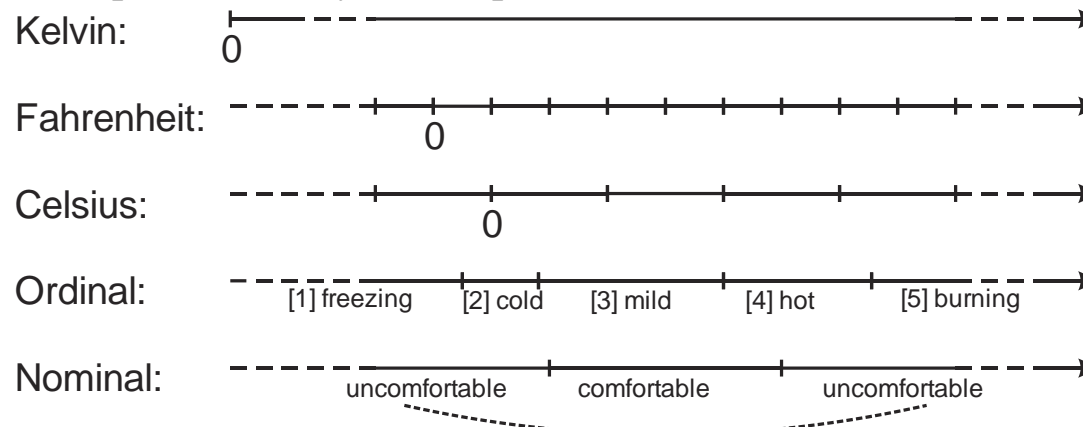
- Allow using size comparison phrases such as “twice as big”.
- Applicable mathematical relationships:

$$=, \neq, <, >, \text{ and transformation } y = \underbrace{b_0}_{=0} + b_1 \cdot x$$

with $b_1 > 0$ (the distance unit) the origin $b_0 = 0$ constraint to zero (fixed base level).

- Notes: Metric and continuous variable. Frequently used in the physical science.

- Example: Hierarchy of Temperature Scales



- Measurements can be **recoded** and **transformed**. Example: Transforming Celsius into Fahrenheit $[^{\circ}\text{F}] = [^{\circ}\text{C}] \times 9/5 + 32$ or Fahrenheit into Celsius $[^{\circ}\text{C}] = ([^{\circ}\text{F}] - 32) \times 5/9$ or *recoding* temperatures: mild \rightarrow comfortable, hot and cold \rightarrow uncomfortable.
- Metric measurement scales can have a fixed **upper and lower** bound within which the numerical representation of an attribute can vary. Example: The percentage scale.

- **Concatenation operations** allow combining and aggregating ratio-scaled observations.
Example: $[area\ of\ region\ A] + [area\ of\ region\ B]$ or number of people in both areas.
How many reference copies are necessary to give the observed aggregated value? The **unit size** of the reference copy determines the **resolution of the scale** and subsequent **measurement precision**.
The smaller the unit size the higher the precision (remember the machine epsilon).
- Further differentiation:
 - **Discrete** measurements comprise of a countable number of representations (usually a small number of integer numbers).
 - **Continuous** measurement allow for fractional representations. Continuous measurements are always quantitatively scaled (floating point numbers).
 - **Fundamental** measurements are directly observable, e.g., distances.
 - **Derived** measurement must be calculate from fundamental measurements,
 - the **ratio of different units** e.g., $kilo-watt-hours-per-day = power-consumption-during-billing-period / length-of-billing-period$, the *population density*
 - the **rate of equal units** where the numerator is part of the denominator: *murders among total number of crimes*, the *percentage of population living in urbanized areas*.

- The measurement scale and spatial objects:

	POINT	LINE	AREA
NOMINAL	<ul style="list-style-type: none"> • Town ✕ Mine ✕ Church BM ✕ Bench Mark 	<ul style="list-style-type: none"> — River — Road — Graticule - - - Boundary 	<ul style="list-style-type: none"> Swamp Desert Forest Census Regions
ORDINAL	<ul style="list-style-type: none"> Large Medium Small 	(Roads) <ul style="list-style-type: none"> Interstate U.S. numbered State County 	<ul style="list-style-type: none"> Major industrial region Minor industrial region
INTERVAL - RATIO	REPETITION Each dot represents 75 persons GRADUATED One-dimensional Bars Two-dimensional Circles, squares, triangles, etc.	REPETITION Isarithms GRADUATED Hachures Flowlines	Density Elevation

FIGURE 5.6 Some examples of the three classes of representation (point, line, area) and how they might be used to portray nominal, ordinal and interval-ratio data.

Data Problems

- Ultimately, the data **link back** to objects and events in the **real world** on which the data were measured. We want to make statements about the real world by using data generated by it.
We would like to know how our world is operating and perhaps manipulated it, that is, we would

like to understand the underlying data generating mechanism. (see **MATHMUSICSTATSLITERATURE.PDF**)

Data tell a story about what happened in the real world. Always be critical and check if the perceived story makes “sense”.

Based on the instructor’s experience advising graduate students, students have difficulties to ***relate data and their geo-statistical analyses to the real world situations***.

Geographic Information Systems just build ***models which are abstractions of the real world*** – we must understand the connection between the real world and the model.

- An analyst must know as much as possible about the ***measurement process*** that has been used to generate the observed data. She/he has to make sure that the data make “sense”:
 - What do we expect to see \Rightarrow do the data and analysis results ***match our expectations***? If they do not match our expectations, then we either have messed up or we are discovering something surprisingly and previously unknown to us.
 - One way to ***gain confidence*** is to use [a] two independent samples or [b] different analysis approaches. If the outcomes of both approaches match then we are more confident in the feasibility of the results. In statistics this is called ***meta-analysis***.

Problems to be expected with data are:

- instrument calibration, instrument precision and rounding
- Dealing with putative outliers and re-editing and inconsistent data.

Example: ozone layer hole over Antarctica has not been discovered before mid-1980 because the

software analyzing data from the Nimbus-7 satellite was programmed to delete extreme measurements. Extreme measurements were initially regarded to reflect a sensor problem.

- Omissions or selective over-sampling of specific groups of data.
- Software limitations and operation:
Example: Trying to import 120,000 records into an old 16-bit *dBase* file or incorrectly joining two different data sets.
- Aggregation into summary measures ***drops the variability*** among the objects by comprising their attributes into a simple summary measure.
- Errors in the data may be systematic (e.g., biased) or random variability:

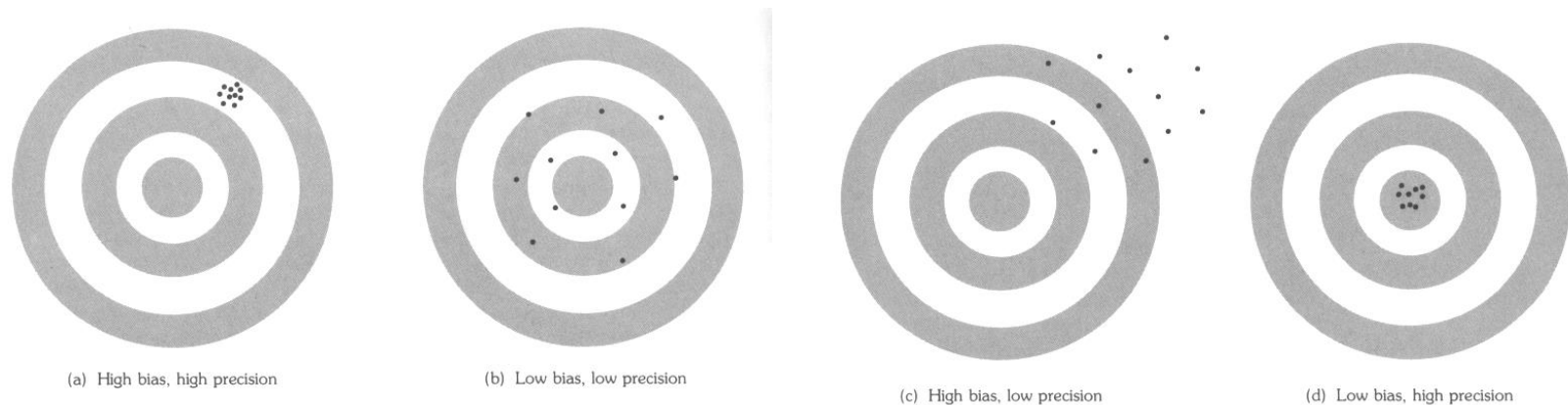
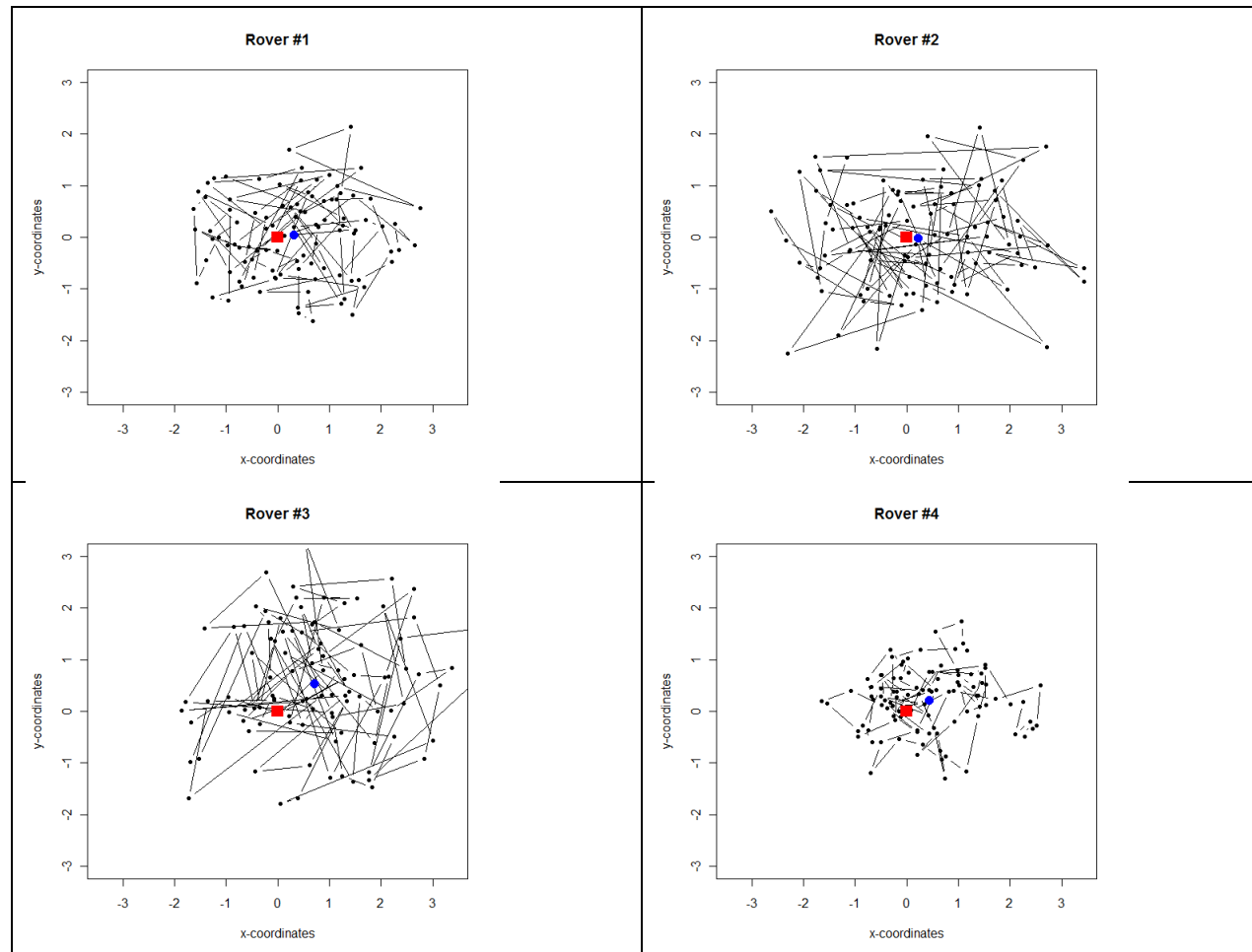


Figure 2. Bias and lack of precision in sample results.

Example: A temporal sequences of 100 measurements were taken by 4 GPS units at the same time and the same fixed location. The **red square** is the true location and the **blue dot** the average measured location.

- What errors and random variations influence the measured locational information?
- Is there some temporal persistence in the measurements?



Meaning of Statistics

- Statistics deals with *describing, summarizing, handling* and *analyzing* the **variability** within data
- Descriptive Statistics: Collection, organization, and presentation of data
Summary: Reducing the variability in a distribution into a small number of meaningful characteristics
=> minimize the effects of information loss. Learn how to meaningful interpret the summary statistics. Perform comparisons to understand the internal variability
- Specialized meaning: **Methodology** of the collection, presentation and analysis of data under **uncertainty**.
 - [a] Methods leading to summarization of vast amounts of data within a given **model framework**,
 - [b] theory validation and generalization, [c] forecasting (planning, selection of course of action)
- Inferential statistics: Draw conclusion about an **unknown population** (or data generating process) from a *random* or a *structured random* samples. **Bayesian statistics** incorporates prior believes about the underlying population. The sample has to be representative of its underlying population.

Other Key Terms

- | | |
|---|---|
| <ul style="list-style-type: none">○ Data organization in tables (matrices)○ Statistical population○ Population characteristics○ Variable○ Population census versus sample○ Random sample | <ul style="list-style-type: none">○ Sampling error versus non-sampling or data acquisition errors○ Statistical estimation versus hypothesis testing○ Validity, accuracy and precision of measurements |
|---|---|