Statistical Thinking

Based on C. J. Wild and M. Pfannkuch (1999). Statistical thinking in Empirical Enquiry, *International Statistical Review*, **67**(3):223-265.

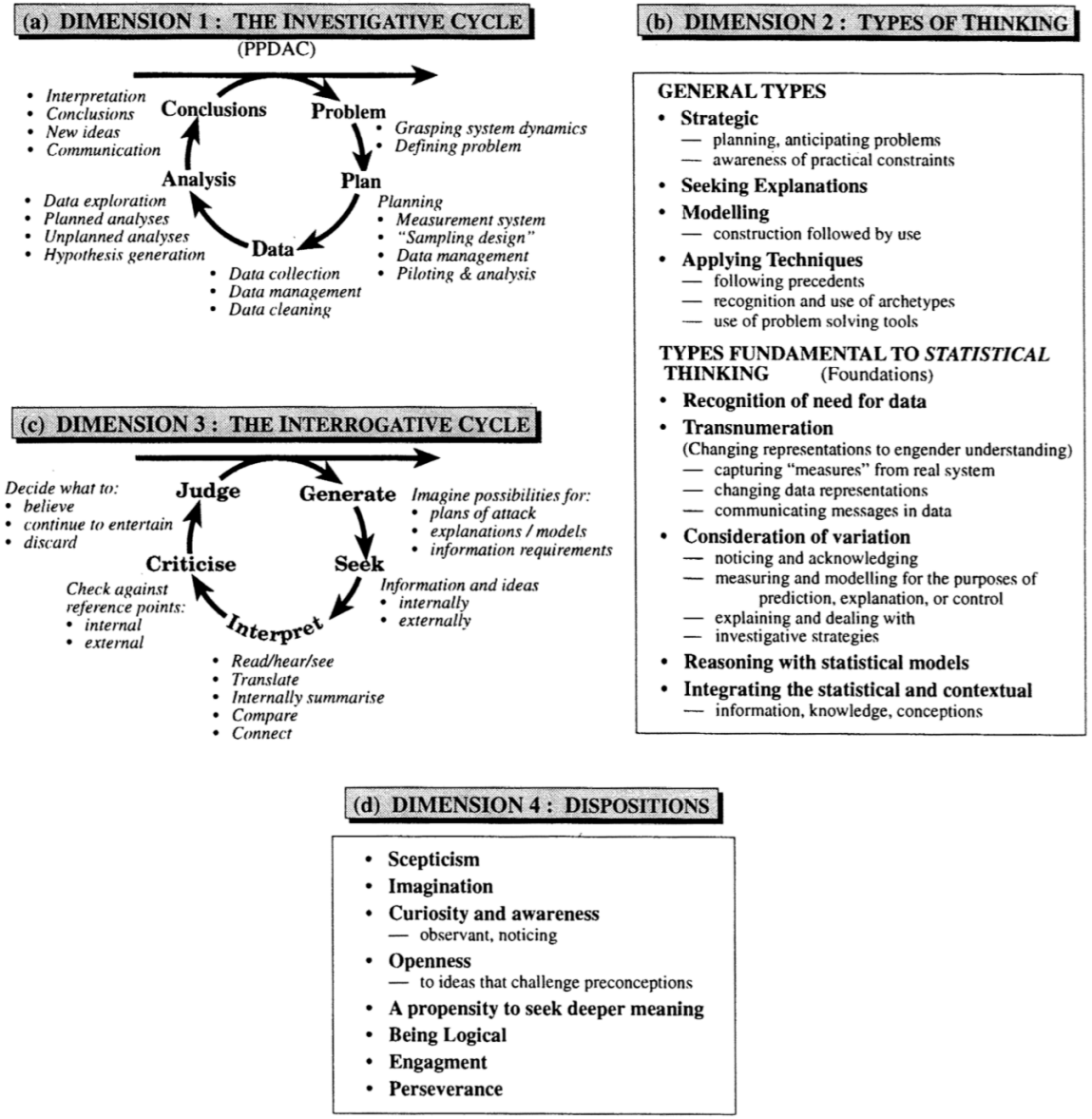
+

Professor Matt Waite's notes

Basic Ideas

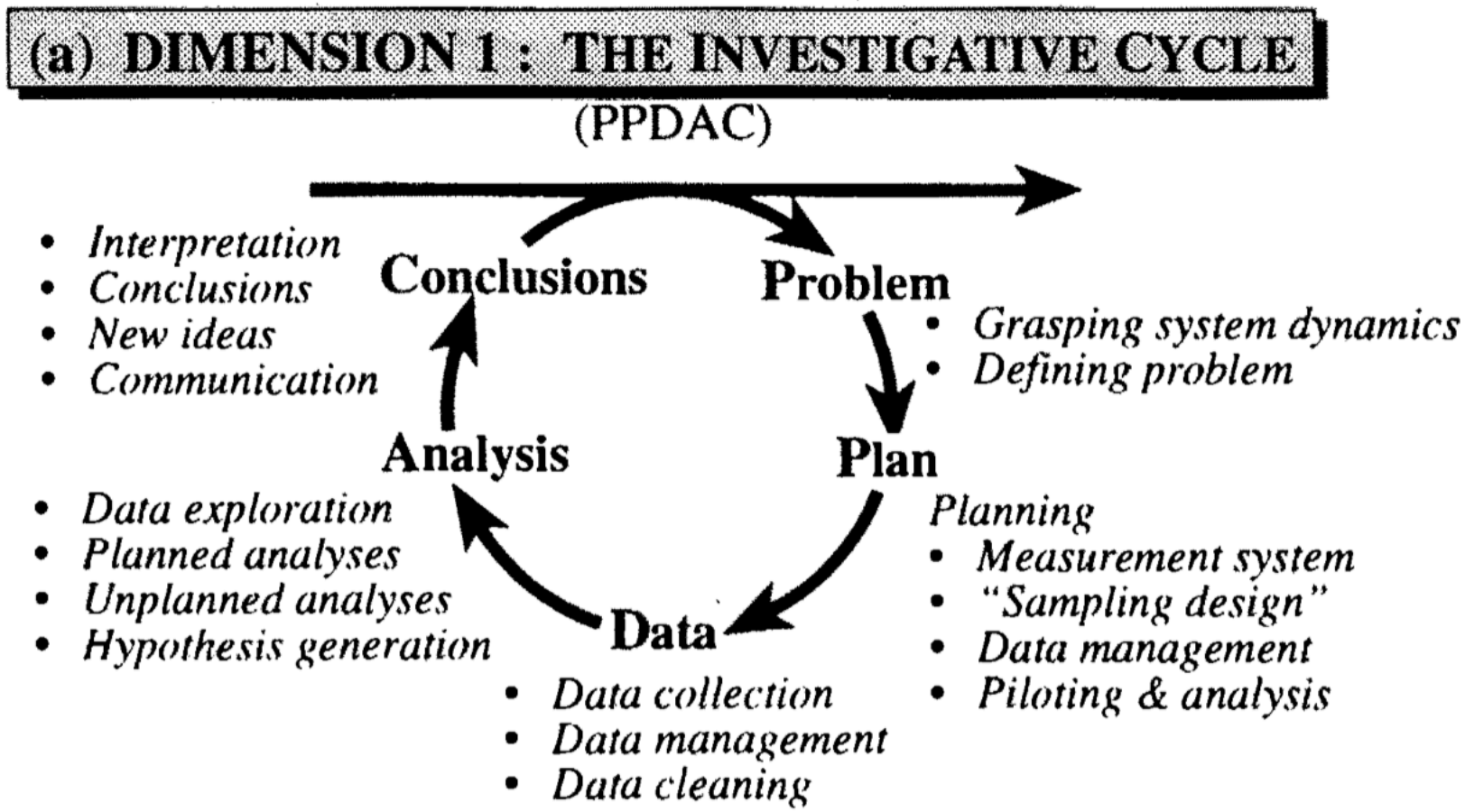
- Thought processes involved in statistical problem solving
  - From problem formulation to conclusions
- A four-dimensional framework for statistical thinking in empirical enquiry
  - Investigative cycle
  - Interrogative cycle
  - Types of thinking
  - Dispositions
- Central element: “**variation**”

Four-Dimensional Framework



Dimension 1: The Investigative Cycle

- Concerned with abstracting and solving a statistical problem grounded in a larger ”real” problem
- Based on the PPDAC model (Problem, Plan, Data, Analysis, Conclusions)



Dimension 2: Types of Thinking

- **Variation**
  - Thinking which is statistical is concerned with learning and decision making under uncertainty
  - for the purposes of explanation, prediction, or control

(b) DIMENSION 2 : TYPES OF THINKING

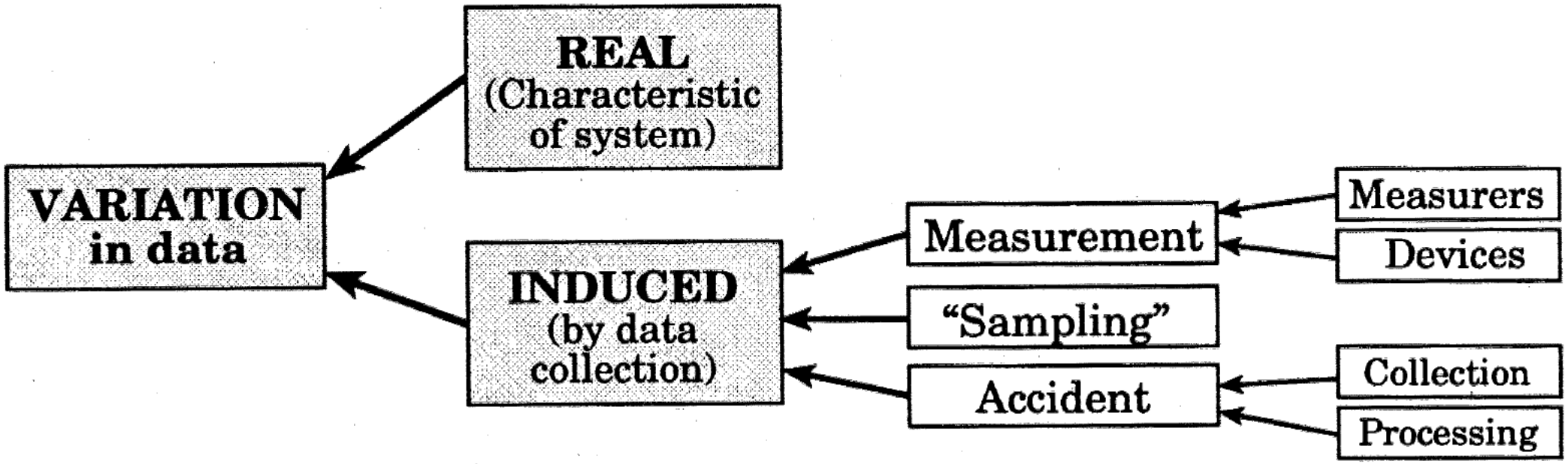
GENERAL TYPES

- **Strategic**
  - — planning, anticipating problems
  - — awareness of practical constraints
- **Seeking Explanations**
- **Modelling**
  - — construction followed by use
- **Applying Techniques**
  - — following precedents
  - — recognition and use of archetypes
  - — use of problem solving tools

TYPES FUNDAMENTAL TO *STATISTICAL* THINKING (Foundations)

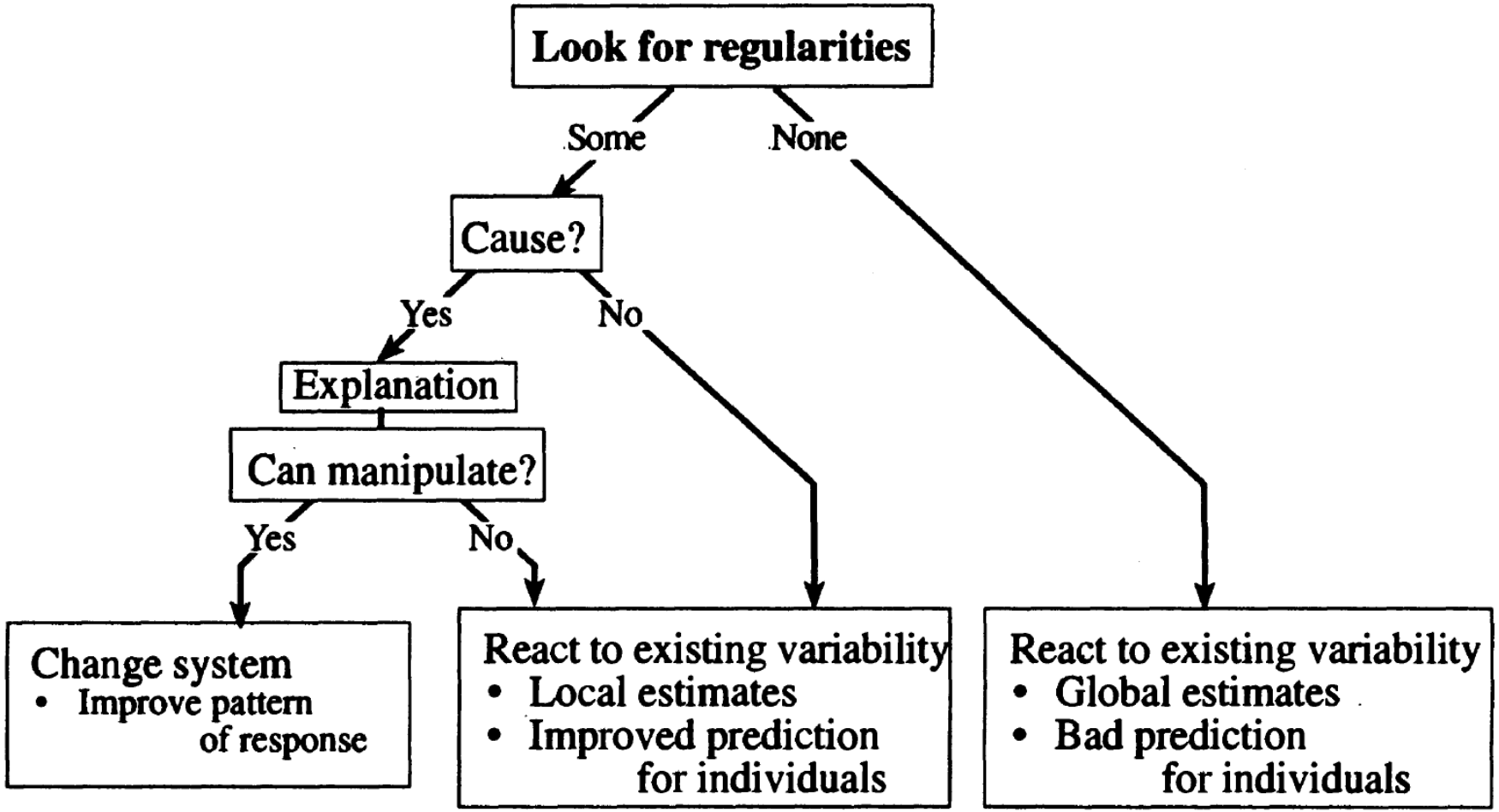
- **Recognition of need for data**
- **Transnumeration**
  (Changing representations to engender understanding)
  - — capturing "measures" from real system
  - — changing data representations
  - — communicating messages in data
- **Consideration of variation**
  - — noticing and acknowledging
  - — measuring and modelling for the purposes of prediction, explanation, or control
  - — explaining and dealing with
  - — investigative strategies
- **Reasoning with statistical models**
- **Integrating the statistical and contextual**
  - — information, knowledge, conceptions

Dimension 2: More on Variation | Sources



Dimension 2: More on Variation |

Prediction, Explain, Control



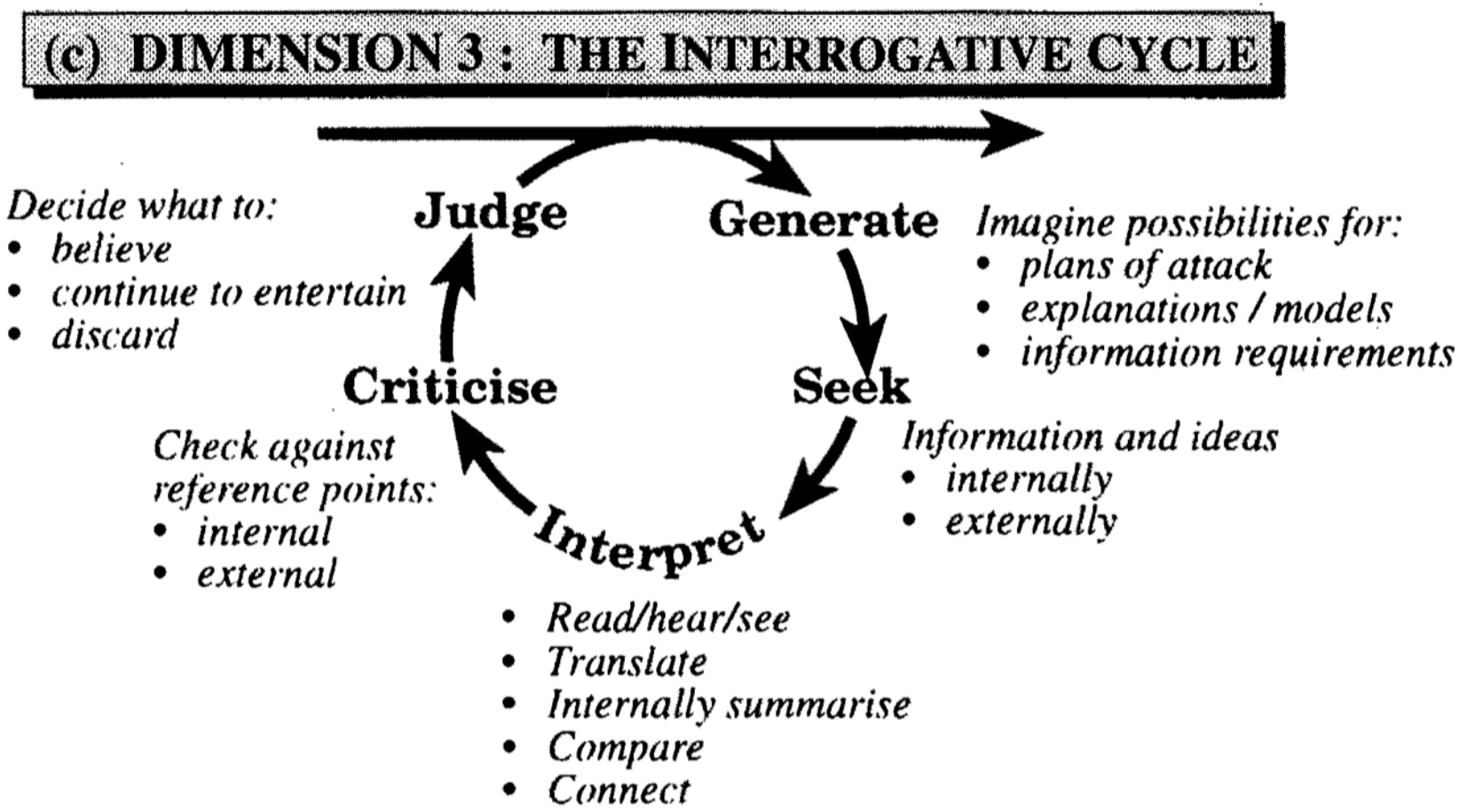
Dimension 2: Summary on Variation

- **Special-cause** vs. **common cause** variation
  - Useful when looking for causes
- **Explained** vs. **unexplained** variation
  - Useful when exploring data & building a model for them
- Suppositions
  - Variation is an **observable reality**
    - Some variation can be explained; other variation ***cannot*** be explained on current knowledge
  - **Random** variation is the way in which statisticians model unexplained variation
    - This unexplained variation may in part or in whole be produced by the process of observation through ***random sampling***
  - Randomness is a ***convenient*** human construct which is used to deal with variation in which patterns cannot be detected

Correlation is NOT causation

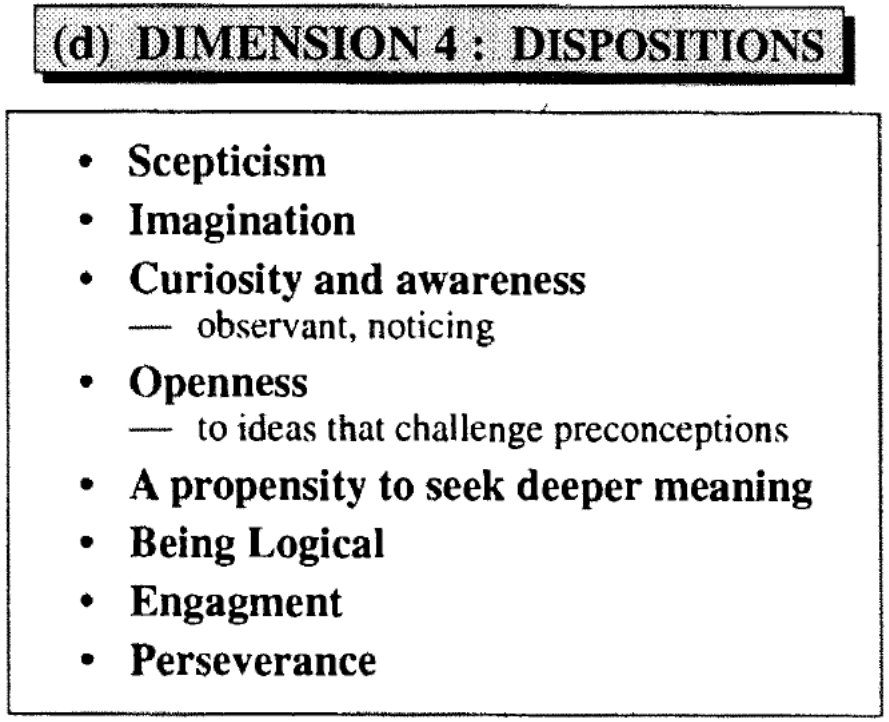
Dimension 3: The Interrogative Cycle

- Applies at macro levels
- Applies also at very detailed levels of thinking
  - **Recursive**
  - Subcycles are initiated within major cycles



Dimension 4: Dispositions

- When authors become intensely interested in a problem or are, ***a heightened sensitivity and awareness*** develops towards information on the peripheries of our experience that might be related to the problem
  - ***People are most observant in areas they find most interesting***
- **Engagement** intensities each dispositional element



Types of Analytics

- **Descriptive**
  - Describing characteristics or properties in the data
- **Predictive**
  - Predicting the types of outcomes given new sets of data, usually based on a classifier trained using labelled, existing datasets
- **Prescriptive**
  - Deciding on the best route or option or decision to make given data

Types of Data

- **Categorical** (cf. wikipedia)
  - Variable that can take on one of a limited, and usually fixed number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property
  - The blood type of a person: A, B, AB or O
  - The state that a person lives in
  - The political party that a voter might vote for
  - The type of a rock: igneous, sedimentary or metamorphic
  - ***Ordinal*** data?
- **Numerical**
  - Can be subdivided into discrete data (things that can be counted) and continuous data (all possible numbers).
  - # of children, age, scores, temperatures, etc.

Descriptive Statistics

- There are three main groups of descriptives
- The **distribution**
  - Works well with categorical data. How many of each thing is there?
- The **central tendency**
  - Only works with numerical data. What is the mean, median and mode?
- The **dispersion**
  - Only works with numerical data. How spread out is the data?

Descriptive Statistics: Distribution

- Grouping and counting by categorical data – group and count by town, or zip code or something like that
  - Often called a **frequency distribution**
  - **Histogram**
- With numerical data, **minimum** and **maximum** values are useful

Descriptive Statistics: Central Tendency

- **Mean**
  - ***Average or norm***: all up all values to find a total, and then divide the total by the number of values
- **Median**
  - ***Middle value***: Sort all values into order, and the median is the middle value; if there are 2 values in the middle, find the mean of these two
- **Mode**
  - ***Most frequent value***: Count how many each value appears, the mode is the value that appears the most
  - Can have more than one mode

Descriptive Statistics: Dispersion

- **Mean**
  - ***Average or norm***: all up all values to find a total, and then divide the total by the number of values
- **Median**
  - ***Middle value***: Sort all values into order, and the median is the middle value; if there are 2 values in the middle, find the mean of these two
- **Mode**
  - ***Most frequent value***: Count how many each value appears, the mode is the value that appears the most
  - Can have more than one mode

Descriptive Statistics: Dispersion

- **Range**
  - Difference between the lowest and highest values
  - Subject to extremes (e.g., ***outliers***)
- **Standard deviation**
  - It is the relation that a set of scores has to the mean
  - Subject to ***skewness*** in distribution
- For a Gaussian/normal distribution
  - 68% of all values will be within 1 standard deviation
  - 95% will be within 3 standard deviation

Dirty Data

- **Missing** data
  - Blanks in the database or spreadsheet.
  - Data missing from a period of time.
  - Missing states, counties, zip codes.
- **Wrong** data
  - Wrong type – numbers where they should be text and vice versa
  - Sharp curves – trends that continue normally that suddenly jump in one year
  - Conflicting data within a dataset or across datasets (race, percentages, etc)
- **Unusable** data
  - Non-standardized data
  - Inconsistent data
  - Abbreviations
  - Unit consistency

Correlation

- Pearson correlation coefficients (or Pearson product-moment correlation coefficient)
- It is a measure of how LINEARLY related two entities are.
- How often is a change in A related to a change in B? And is that positive or negative?

Correlation: For a population

$$\rho_{X,Y} = \frac{\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Standard deviation of *X*; standard deviation of *Y*

where:

- cov and σ_X are defined as above
- μ_X is the mean of X
- μ_Y is the mean of Y
- E is the expectation.

The formula for ρ can be expressed in terms of uncentered moments. Since

- $\mu_X = \mathrm{E}[X]$
- $\mu_Y = \mathrm{E}[Y]$
- $\sigma_X^2 = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - [\mathrm{E}[X]]^2$
- $\sigma_Y^2 = \mathrm{E}[(Y - \mathrm{E}[Y])^2] = \mathrm{E}[Y^2] - [\mathrm{E}[Y]]^2$
- $\mathrm{E}[(X - \mu_X)(Y - \mu_Y)] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y],$

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Correlation: For a sample

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is the sample size
- x_i , y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Correlation: What it means?

- It is based on a range from -1 to 1.
- 1 = **perfect positive correlation**
  - A goes up 1, B goes up 1
  - In the real world, almost never happens outside of a mistake
- 0 = **no correlation at all**
  - 0 rarely ever happens
  - NEAR zero happens all the time
- -1 = **perfect negative correlation**
  - A goes up 1, B goes down 1
  - It is just like 1: rare, probably a mistake

Significance: *t*-test

- The ***t*-test** is any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution under the ***null*** hypothesis.
- A *t*-test is most commonly applied when the test statistic would follow a ***normal*** distribution if the value of a scaling term in the test statistic were known
  - When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a Student's *t* distribution
  - ***The t-test can be used, for example, to determine if two sets of data are significantly different from each other***

https://en.wikipedia.org/wiki/Student%27s_t-test

Significance: *p*-value & null hypothesis

- In the context of null hypothesis testing: to quantify the idea of statistical significance of evidence
  - In essence, ***a claim is assumed valid if its counter-claim is improbable***
- The only hypothesis that needs to be specified in this test and which embodies the counter-claim is referred to as the ***null hypothesis***
  - i.e., the hypothesis to be nullified
- A result is said to be ***statistically significant*** if it allows us to ***reject*** the null hypothesis
  - ***The statistically significant result should be highly improbable if the null hypothesis is assumed to be true***
  - ***The rejection of the null hypothesis implies that the correct hypothesis lies in the logical complement of the null hypothesis***
- ***Caveat***: Unless there is a single alternative to the null hypothesis, the rejection of null hypothesis does *not* tell us which of the alternatives might be the correct one

https://en.wikipedia.org/wiki/Student%27s_t-test