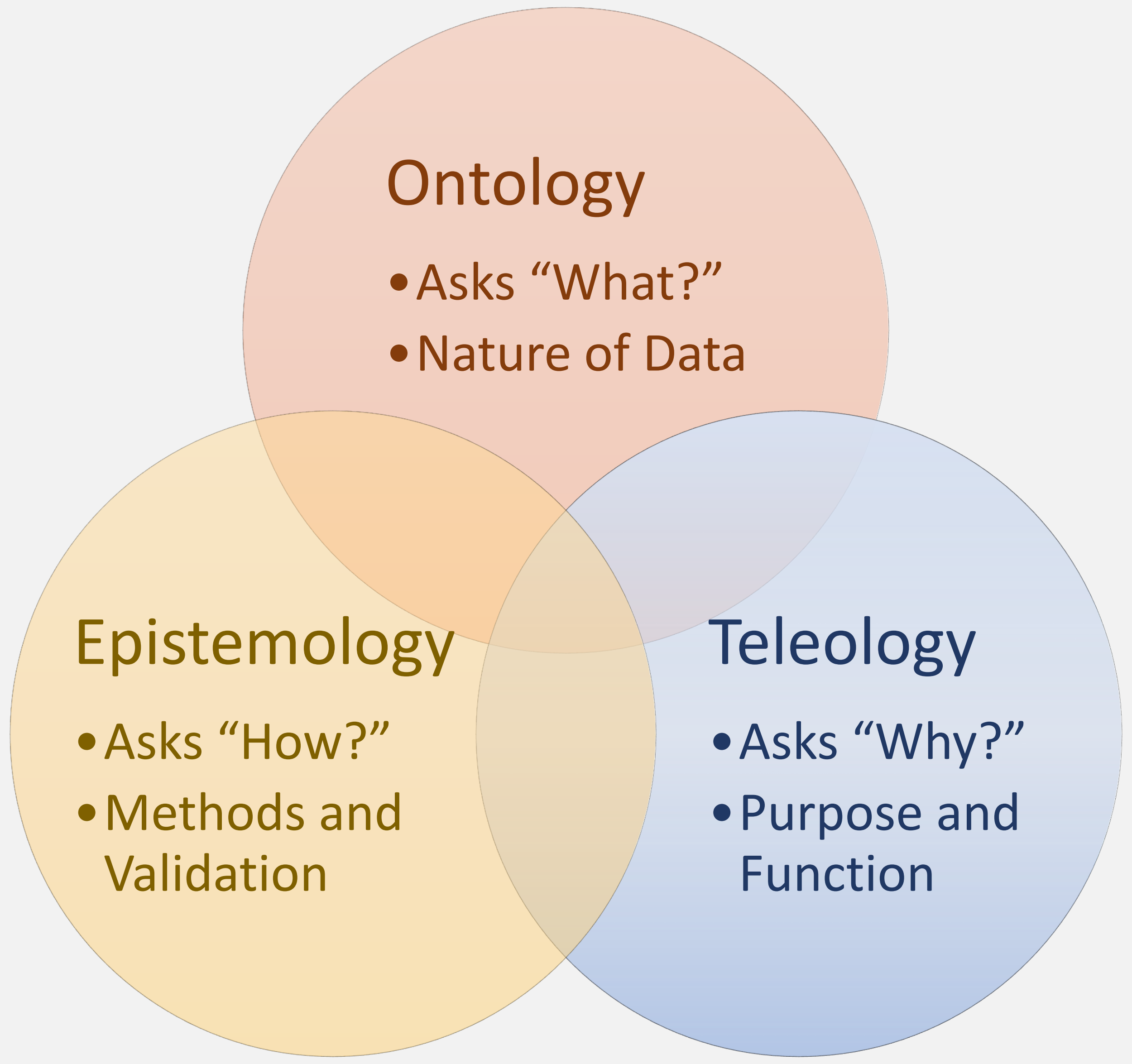
The Archaeology of Data

Ancient Problems, New Science, and a Philosophy of Analysis

J. Scott Cardinal
Georgia Institute of Technology

Email: cardinal.js@gatech.edu
LinkedIn: https://www.linkedin.com/in/cardinal-jscott

When one thinks of archaeology, it may suggest ancient civilizations, exotic art and objects in museums, or portrayals in popular media such as "Indiana Jones" or "Lara Croft". It might be difficult to imagine a world any *more* different from the advanced technologies and mathematics of data science. The actual practice of archaeology, however, has many surprising parallels with modern data science – stochastic and inhomogeneous spatial processes, classification and clustering problems, social network and graph analysis, and managing unstructured data. The true commonality, though, is that both are interested in developing methodologies for converting *data* into *information*, with the goal of transforming that information into *knowledge*.



Ontology
- Asks "What?"
- Nature of Data

Epistemology
- Asks "How?"
- Methods and Validation

Teleology
- Asks "Why?"
- Purpose and Function

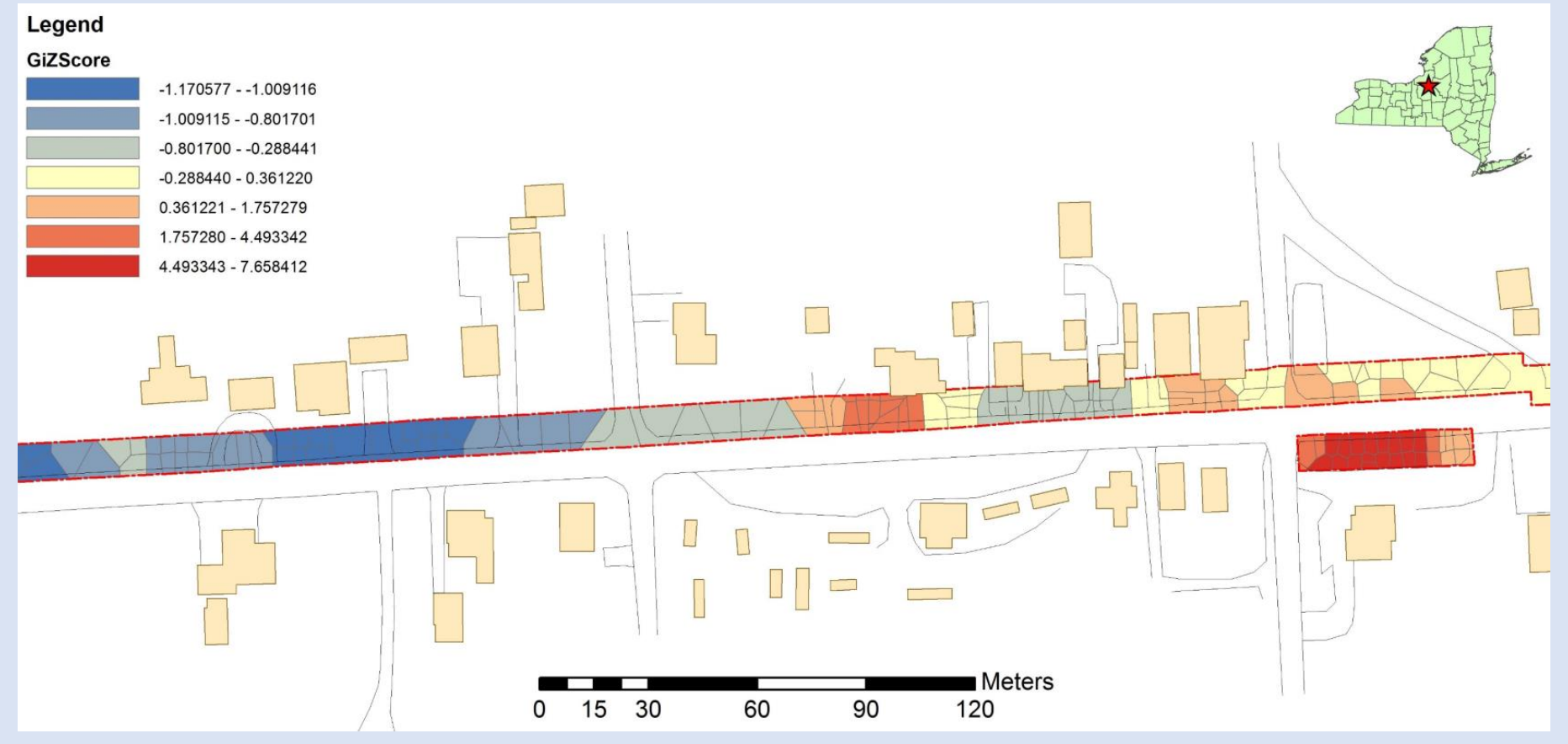
The Philosophy of Data

Philosophy could be considered the *original* data science, and there is a long history of thought concerning the nature and discovery of knowledge. Deriving knowledge from data is at the intersection of three specific areas of philosophical study:

- **Ontology:** the study of the nature of the things
- **Epistemology:** the study of methods and validation
- **Teleology:** the study of purpose and function

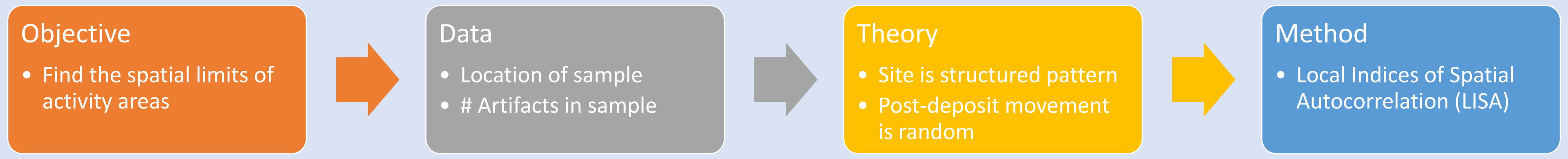
These constitute the "What", "How", and "Why" of any substantial research question. *What* is the empirical nature and limitations of the data (ontology)? *How* do the methods reflect the underlying processes (epistemology) ? *Why* do those processes result in the observable outcomes (teleology)? Each of these must be properly addressed and specified in order to produce meaningful analyses.

Detect Activity Areas as Local Indices of Spatial Autocorrelation



A common research question in archaeology is identifying a specific area of habitation or activity within a larger area of occupation. One way to do this is to look for areas with a greater spatial density of artifacts related to that activity. Over time, those artifacts tend to get spread out due to various natural or behavioral processes. The solution was to filter the random spatial dispersion by looking for the clustered patterns of autocorrelation.

J. S. Cardinal, "Site Identification, Delineation, and Evaluation through Quantitative Spatial Analysis: Geostatistical and GIS Methods to Facilitate Archaeological Resource Assessment," MA Thesis, University at Albany (SUNY), 2011.



Objective
- Find the spatial limits of activity areas

Data
- Location of sample
- # Artifacts in sample

Theory
- Site is structured pattern
- Post-deposit movement is random

Method
- Local Indices of Spatial Autocorrelation (LISA)

Mixed Assemblages as Weighted Gene Co-expression Networks Analysis

When the same location is occupied repeatedly or continuously over time (whether decades or centuries), each subsequent occupation tends to churn up artifacts out of the ground or dig down through the layers of the previous occupations. These mixtures of older and newer artifacts then get re-deposited and buried as a mixed assemblage. The problem was to find their original sequence of deposit. The solution was in evaluating the topological overlap of a network analogous to the way gene coexpressions are clustered.

J. S. Cardinal, "Analysis of Data Recovery Assemblages." In: Davis, NL (ed.), Fort Edward Village Site, Fort Edward Feeder Canal Bridge Site, and Hilfinger Pottery Site, 38–86. Cultural Resource Survey Program 8. New York State Education Department. Chap. 3.

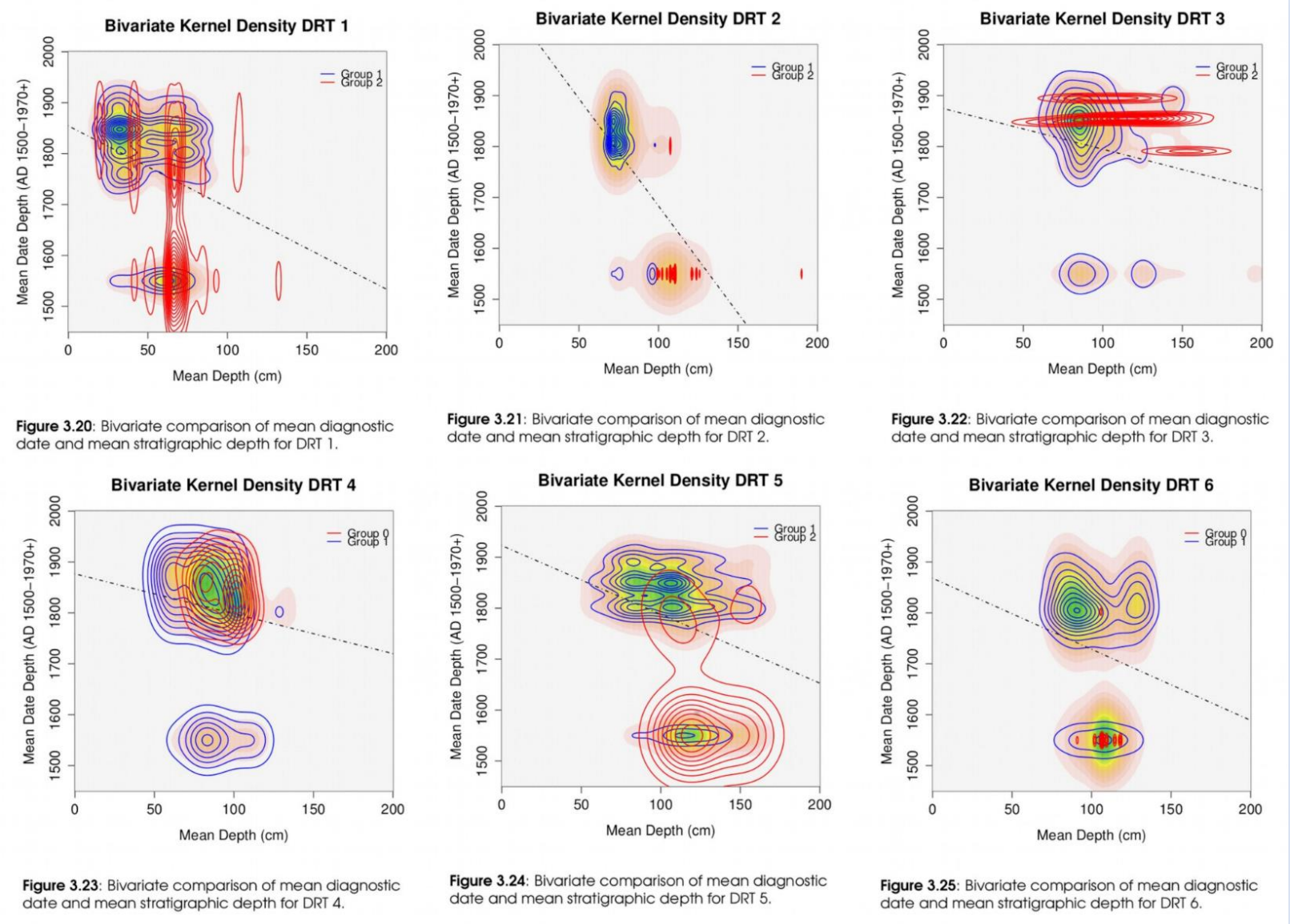


Figure 3.20: Bivariate comparison of mean diagnostic date and mean stratigraphic depth for DRT 1.

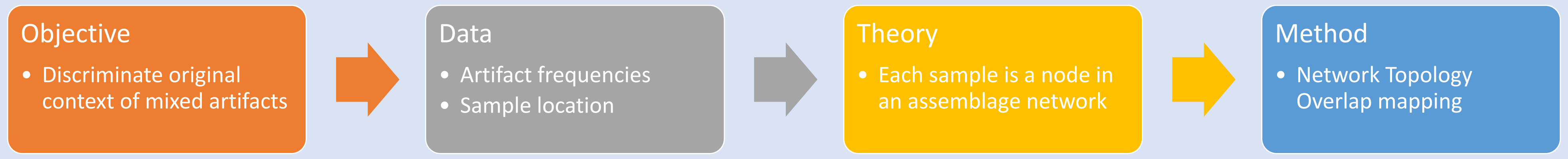
Figure 3.21: Bivariate comparison of mean diagnostic date and mean stratigraphic depth for DRT 2.

Figure 3.22: Bivariate comparison of mean diagnostic date and mean stratigraphic depth for DRT 3.

Figure 3.23: Bivariate comparison of mean diagnostic date and mean stratigraphic depth for DRT 4.

Figure 3.24: Bivariate comparison of mean diagnostic date and mean stratigraphic depth for DRT 5.

Figure 3.25: Bivariate comparison of mean diagnostic date and mean stratigraphic depth for DRT 6.



Objective
- Discriminate original context of mixed artifacts

Data
- Artifact frequencies
- Sample location

Theory
- Each sample is a node in an assemblage network

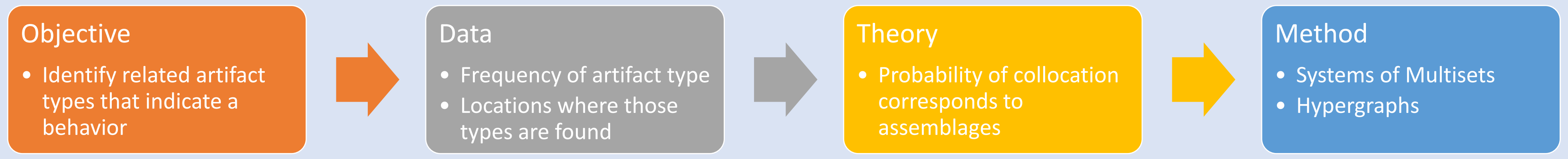
Method
- Network Topology Overlap mapping

Partitioning Assemblages and Contexts as Systems of Multisets

$$
\begin{aligned}
\text{Provenience } \mathcal{V}_i &= \left\{a_1^{m_{V_i}(a_1)}, \ldots, a_n^{m_{V_i}(a_n)}\right\} \\
&= \left\{\forall\, a_i \in^{m_{V_i}(a_i)} \Omega_{\mathcal{S}} \,\middle|\, m_{V_i}(a_i) \in \mathbb{N} \geq 1\right\} \\
\text{Site } \mathcal{S} &= \left\{a_1^{m_S(a_1)}, \ldots, a_n^{m_S(a_n)}\right\} = \biguplus_{i=1}^{n} \mathcal{V}_i \\
\text{All proveniences } \mathbb{V} &= \{\mathcal{V}_1, \ldots, \mathcal{V}_n\} \\
&= \left\{\exists \mathcal{V}_i \subseteq \wp(\mathcal{S}) - \emptyset \,\middle|\, \biguplus_{i=1}^{n} \mathcal{V}_i = \mathcal{S} \wedge \bigcap_{i=1}^{n} \mathcal{V}_i = \emptyset\right\} \\
\text{Assemblage } (\mathcal{A}_i, \mathcal{C}_i) &= \left\{\exists \mathcal{X}_i \in^{m(x)} \wp(\mathcal{S}) - \emptyset, \exists \mathcal{Y}_i \in^{m(y)} \wp(\mathbb{V}) - \emptyset \right. \\
&\qquad \left. \middle|\, \mathcal{X}_i \approx \biguplus_{j=1}^{n} \mathcal{V}_j \in^{m(y)} \mathcal{Y}_i \right\}
\end{aligned}
$$

When different types of artifacts are related in some way, whether by function or as part of a "toolkit" of objects used in some activity, archaeologists refer to them as an "assemblage". The problem is that we don't necessarily know beforehand what artifact types are part of which assemblage… we only know what types are found together and how often. The solution was in formalizing the probabilities by defining that relationship in combinatorial terms.

J. S. Cardinal, "Sets, Graphs, and Things We Can See: A Formal Combinatorial Ontology for Empirical Intra-Site Analysis," J. Comput. Appl. Archaeol., vol. 2, no. 1, pp. 56–78, Apr. 2019.



Objective
- Identify related artifact types that indicate a behavior

Data
- Frequency of artifact type
- Locations where those types are found

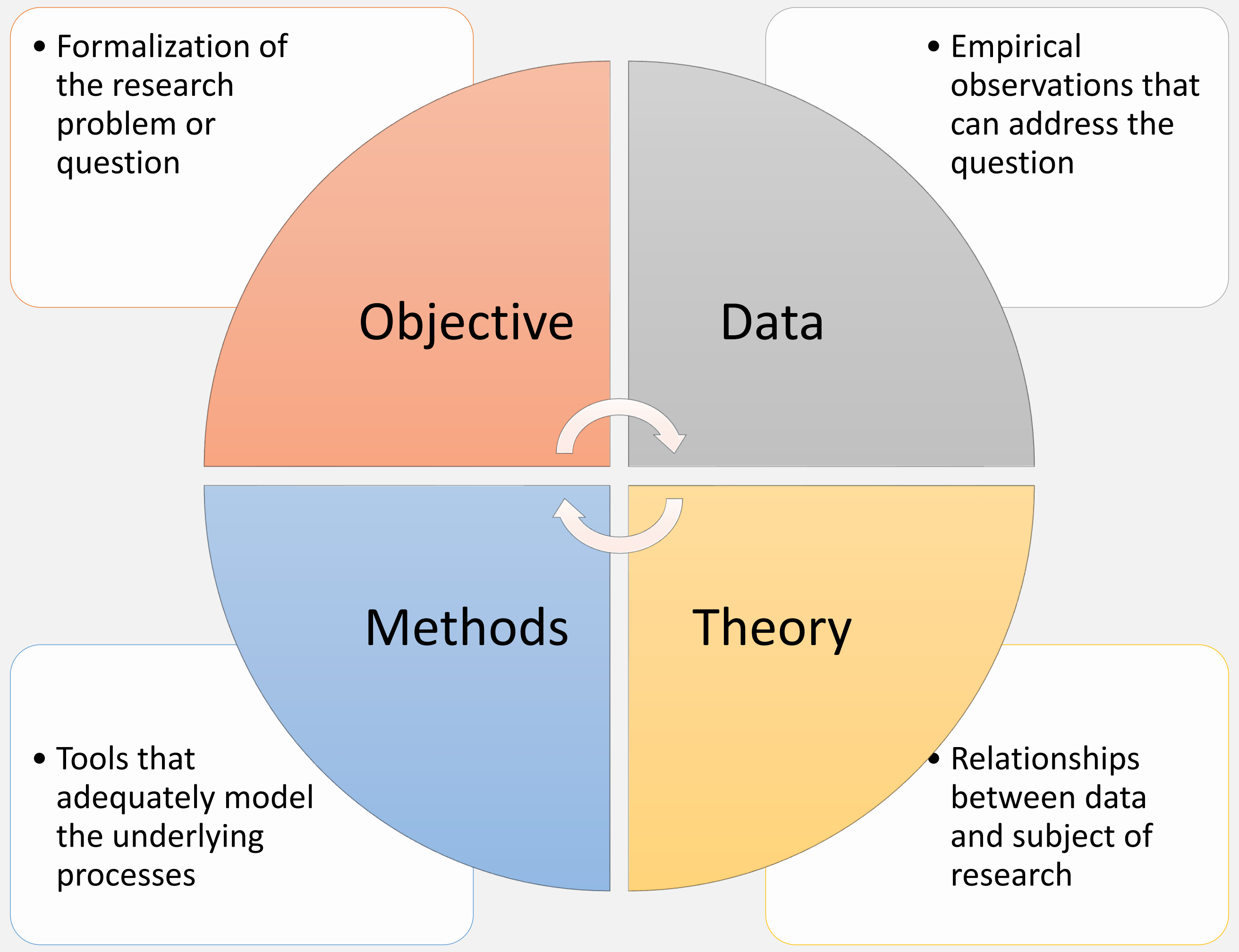
Theory
- Probability of collocation corresponds to assemblages

Method
- Systems of Multisets
- Hypergraphs

Balancing Objective, Data, Theory, and Method

Technical sophistication won't help if either the research question is poorly framed, the data are underspecified, or the model is a poor fit. Each aspect of analyses must be considered in terms its what, how, and why. It's easy to think of the what and how only in terms of the data and models, respectively. It can be less obvious that both data and model are intrinsically linked to both objectives and theory.

The *selection* of data and the *selection* of model are driven by assumptions regarding the underlying relationships and processes of the phenomena being studied. These assumptions are, however, inherently theoretical in nature. Well-conceived research questions balance these four facets of analysis



Objective
- Formalization of the research problem or question

Data
- Empirical observations that can address the question

Methods
- Tools that adequately model the underlying processes

Theory
- Relationships between data and subject of research

Discussion

Unfortunately, the path from data to knowledge can be convoluted, regardless of how rigorous the quantitative methods. Data scientists are trained to evaluate various diagnostics such as bias-variance tradeoffs, goodness of fit, or Type I/II error rates. Unless research questions or objectives are properly framed and formalized, the sophistication of the models and assessments are rendered moot. Essentially, these are specialized cases of more generalizable questions of problem formalization and epistemology. To insure properly framed questions requires an optimal balance between Objectives, Data, Theory, and Method that can be rendered into clear and formal statements of research parameters.