

The Key Concepts and Steps in Data Science

Engin A. Sungur

Statistics Discipline

University of Minnesota, Morris

OUTLINE



PRESENTATION



SURVEY



TEST

- **INTRODUCTIONS & BACKGROUND INFORMATION**
- **STEPS/STAGES OF DATA SCIENCE/STATISTICS**
 - **QUESTION/PROBLEM**
 - **DATA COLLECTION**
 - **DATA MANIPULATIONS**
 - **EXPLORATORY DATA ANALYSIS**
 - **CONFIRMATORY DATA ANALYSIS**
 - **COMMUNICATING THE FINDINGS**
 - **FORMULATING NEW QUESTIONS/PROBLEMS**
- **GENERAL REMARKS**

LEARNING OBJECTIVES



PRESENTATION

- IDENTIFY THE SEVEN STAGES OF THE DATA SCIENCE (you)
- LEARN COMMON CONCEPTS IN EACH (you)
- GET FAMILIAR WITH SOME STATISTICAL TECHNIQUES/METHODS/TOOLS THAT ARE AVAILABLE (you)



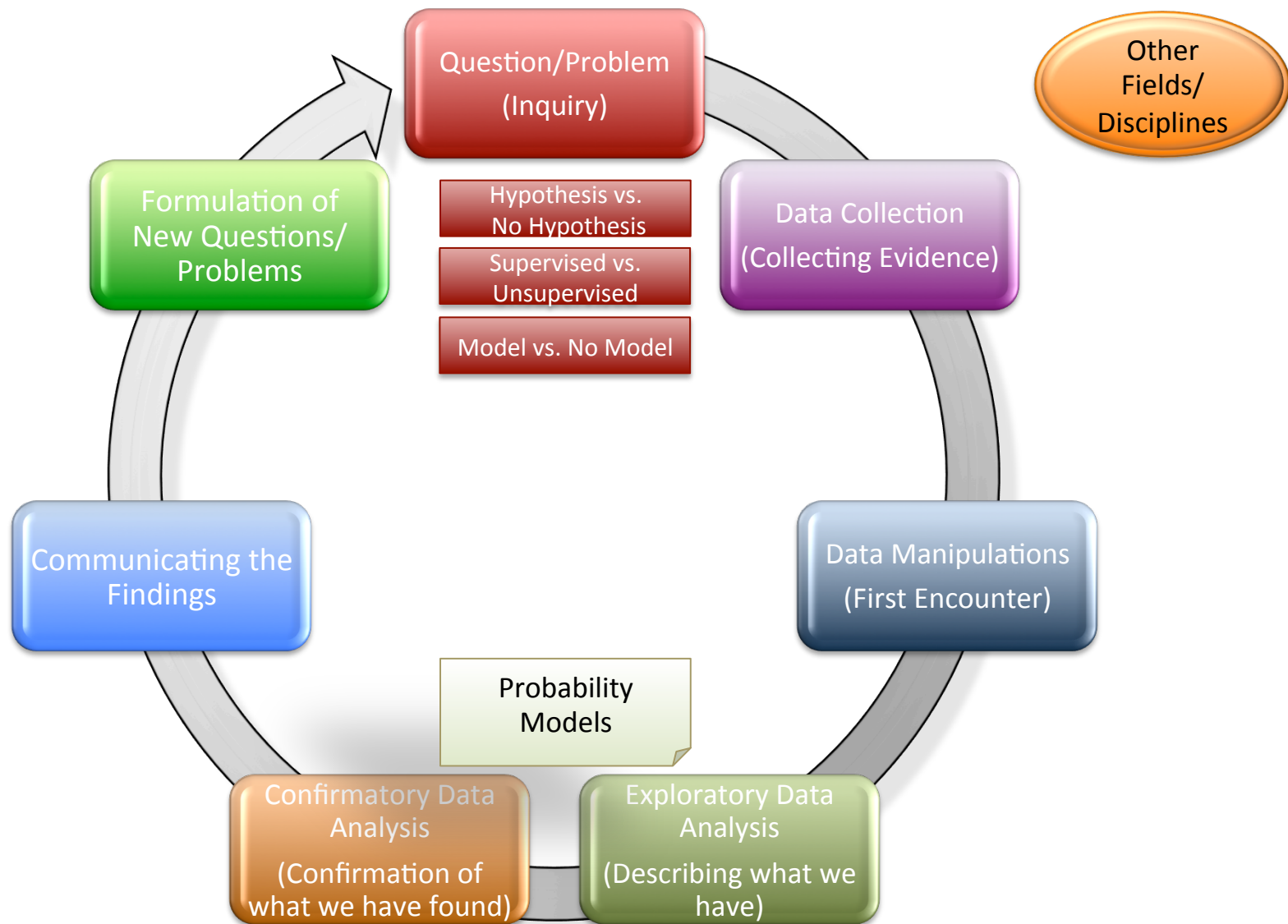
SURVEY

- LEARN ABOUT THE STATISTICAL ASPECT OF THE RESEARCH PROJECTS (me)
 - FORMALIZE THE TYPE OF STATISTICAL NEED OF YOUR PROJECTS (me)
- GET READY FOR YOUR QUESTIONS (me)

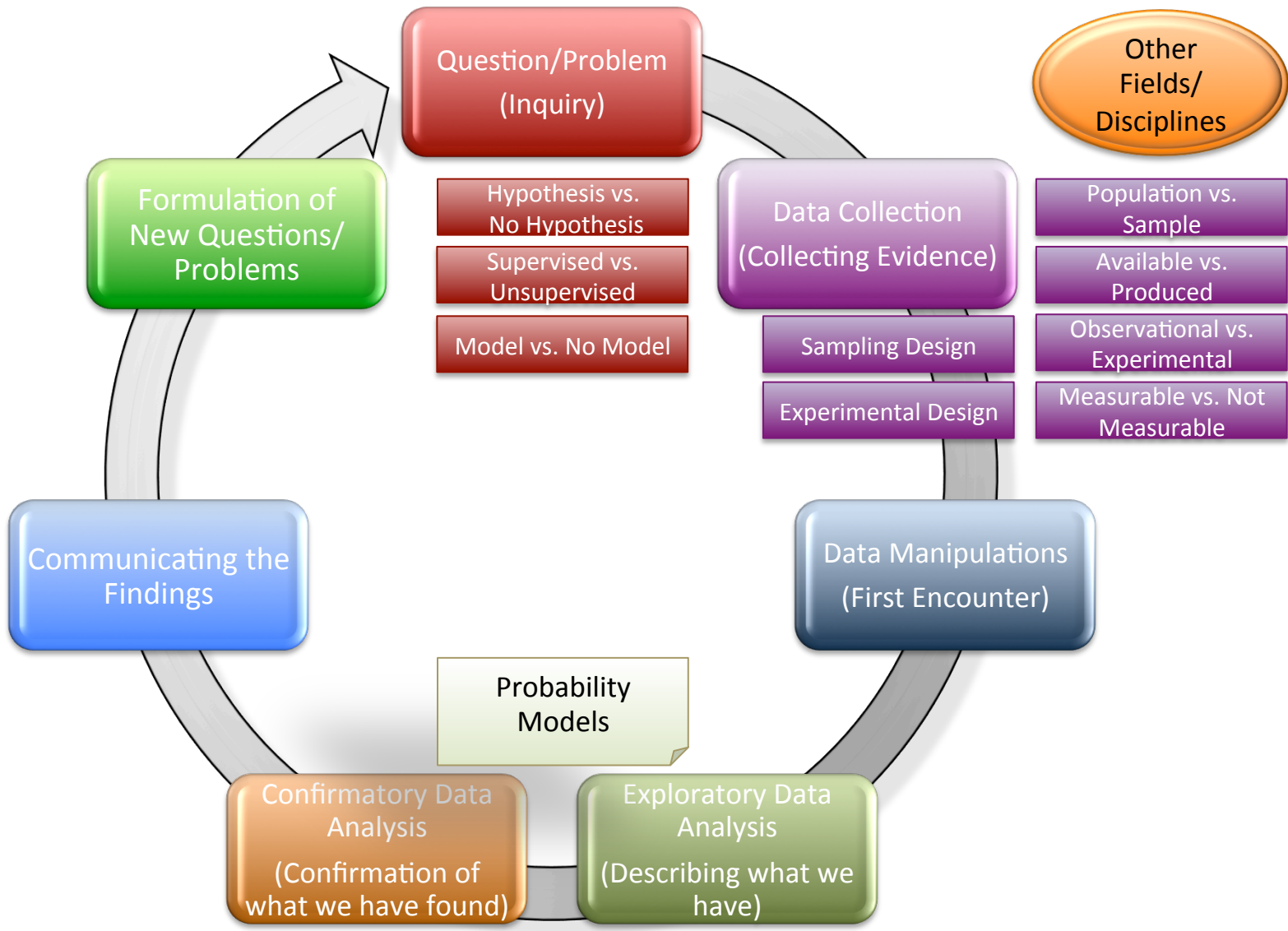
STAGES OF DATA SCIENCE



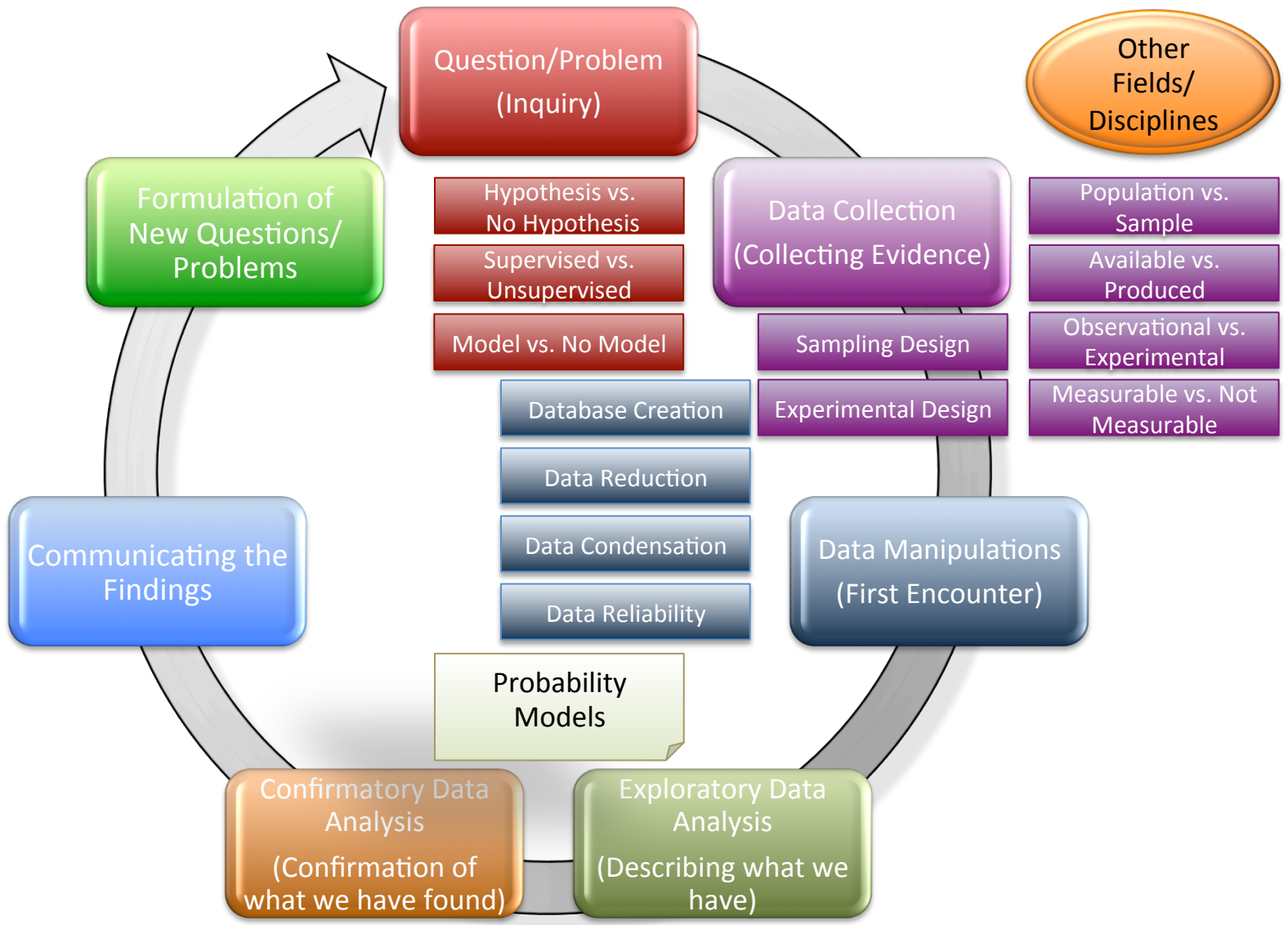
STAGES OF DATA SCIENCE (Contd.)



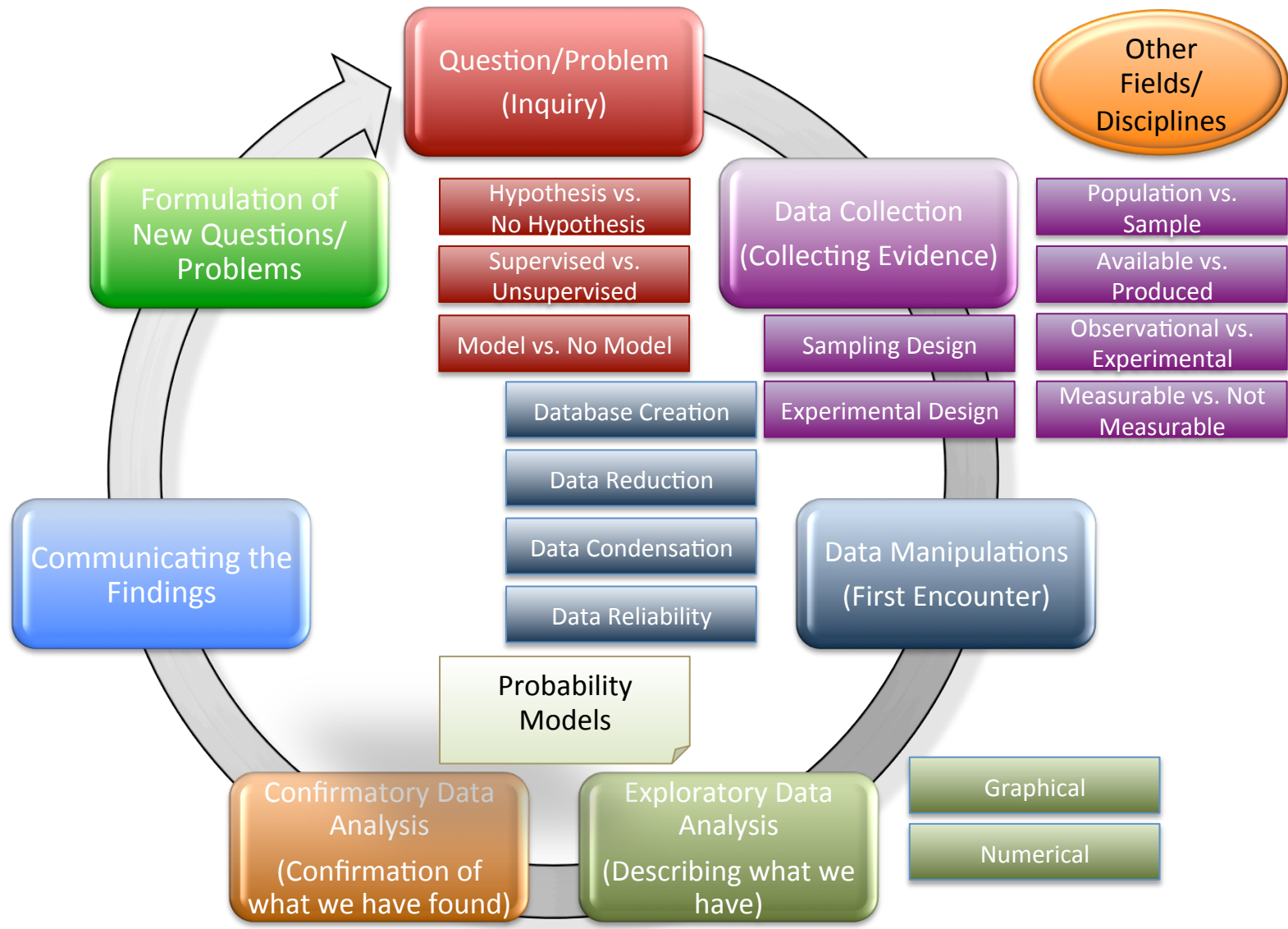
STAGES OF DATA SCIENCE (Contd.)



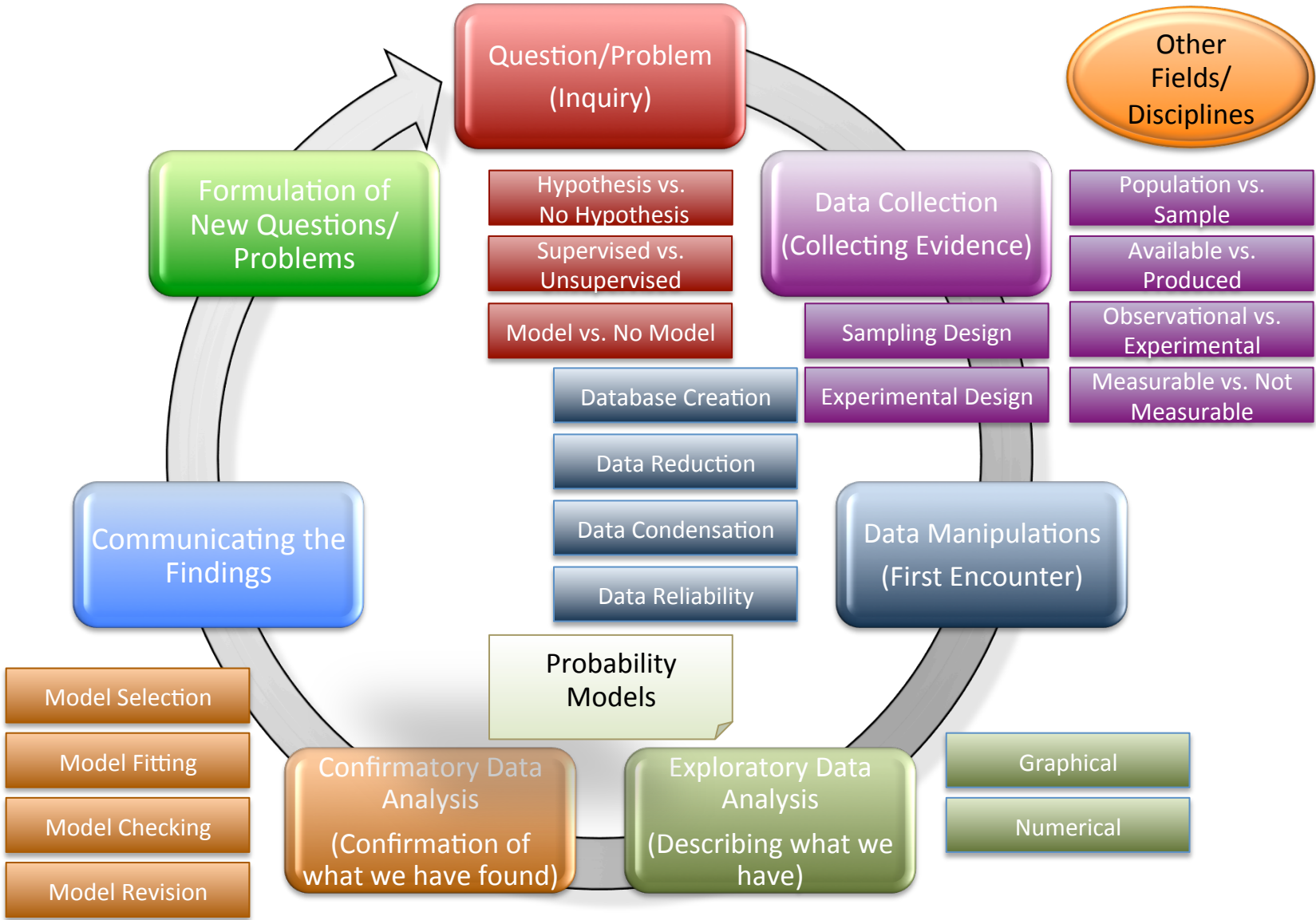
STAGES OF DATA SCIENCE (Contd.)



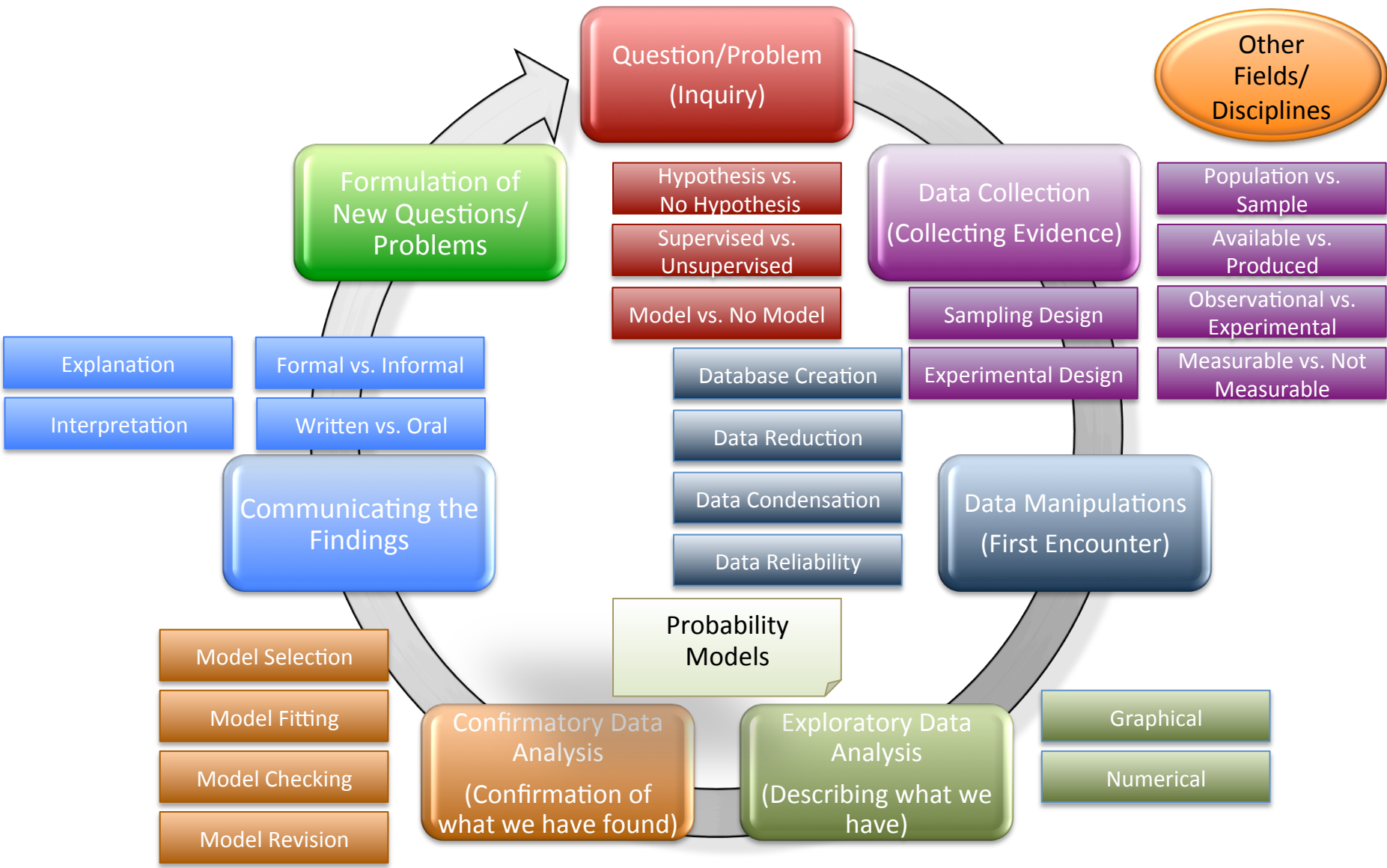
STAGES OF DATA SCIENCE (Contd.)



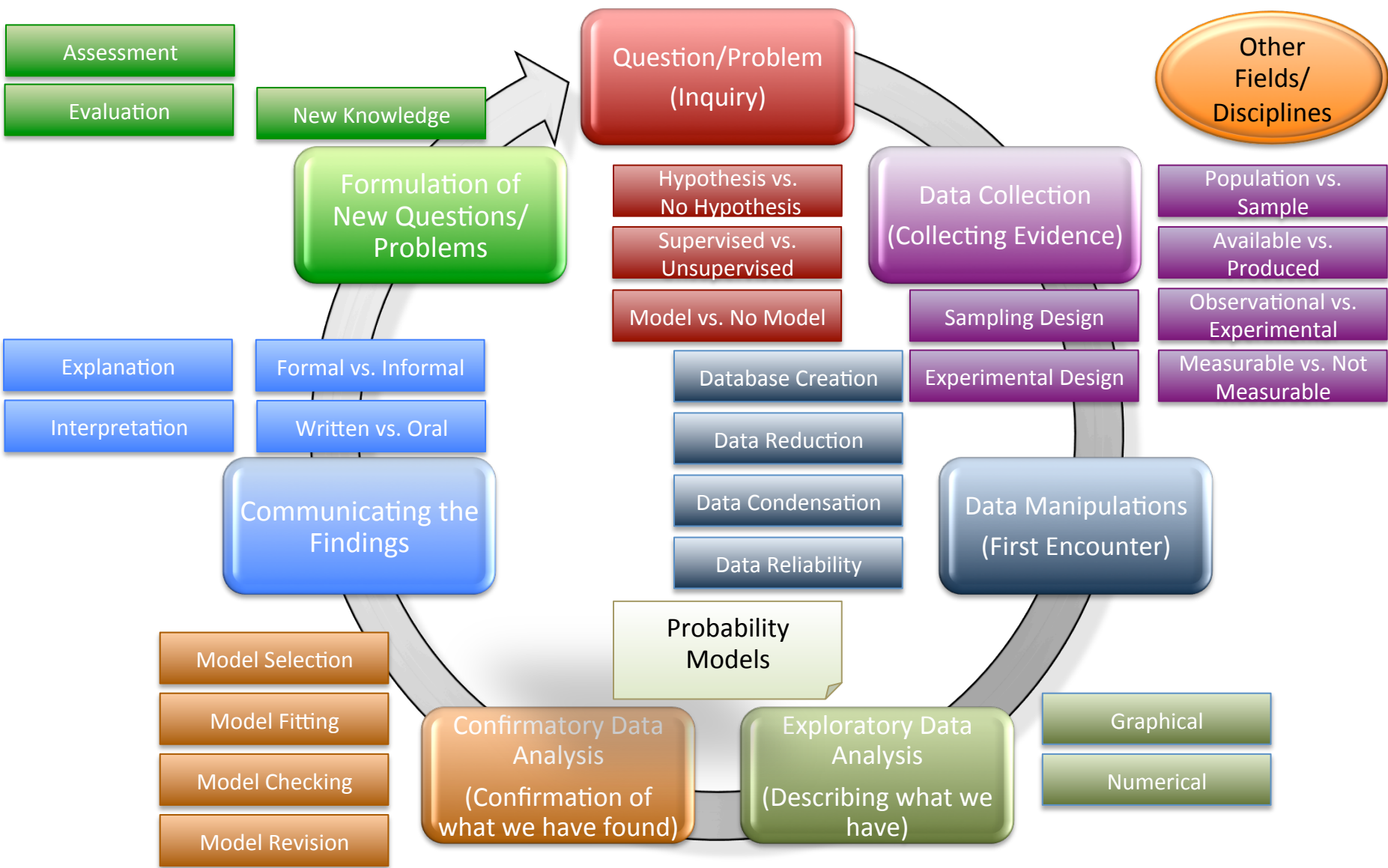
STAGES OF DATA SCIENCE (Contd.)



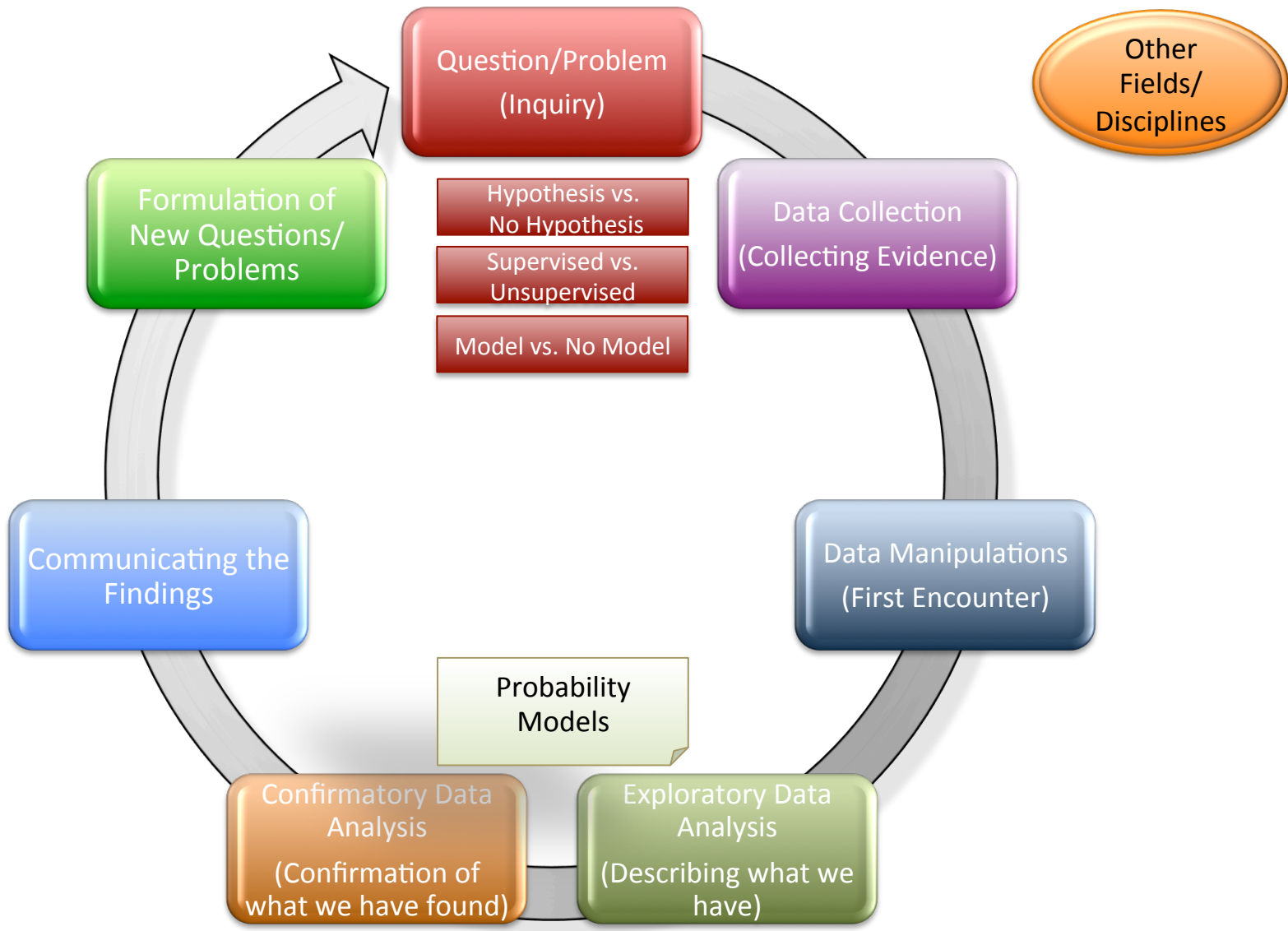
STAGES OF DATA SCIENCE (Contd.)



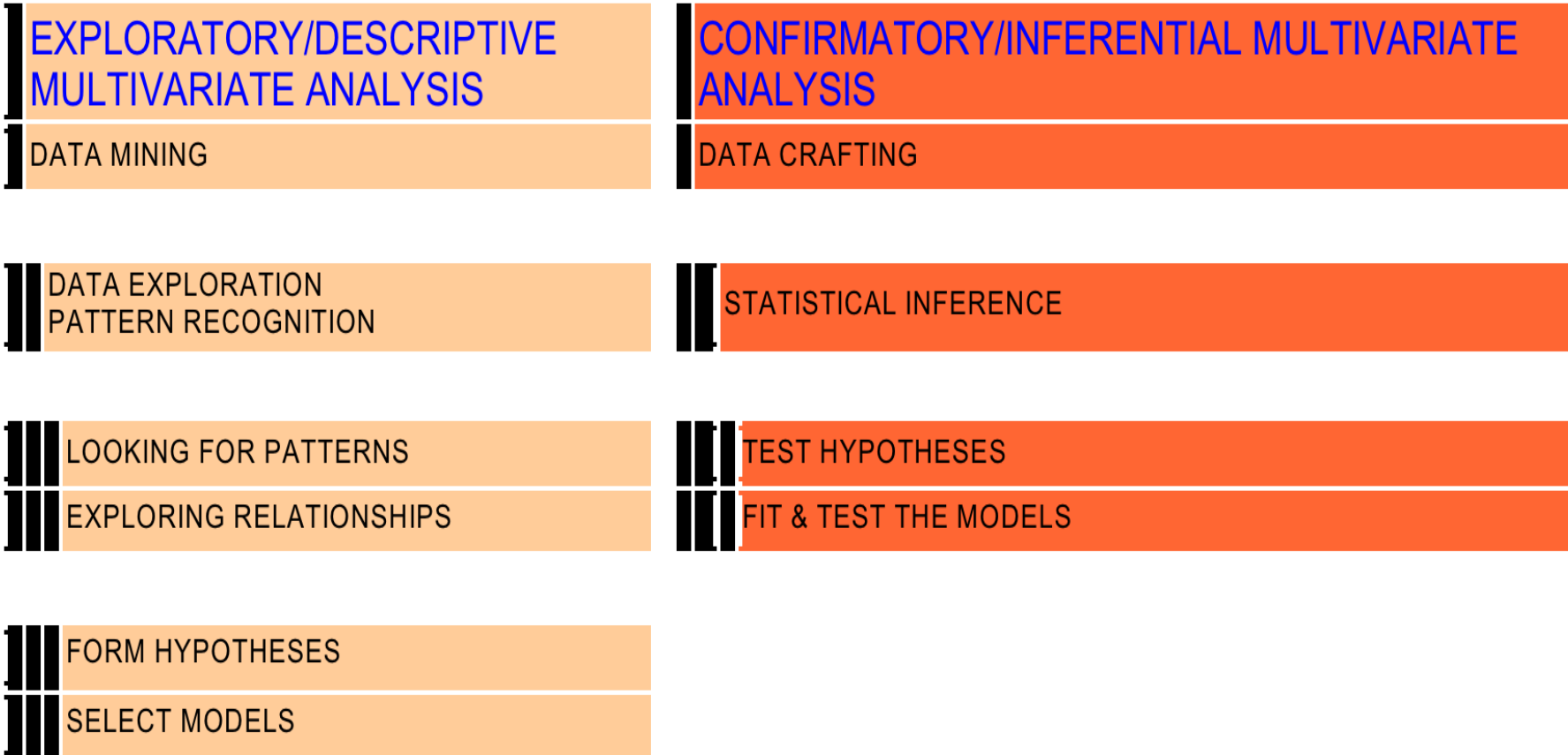
STAGES OF DATA SCIENCE (Contd.)



QUESTION/PROBLEM



QUESTION/PROBLEM



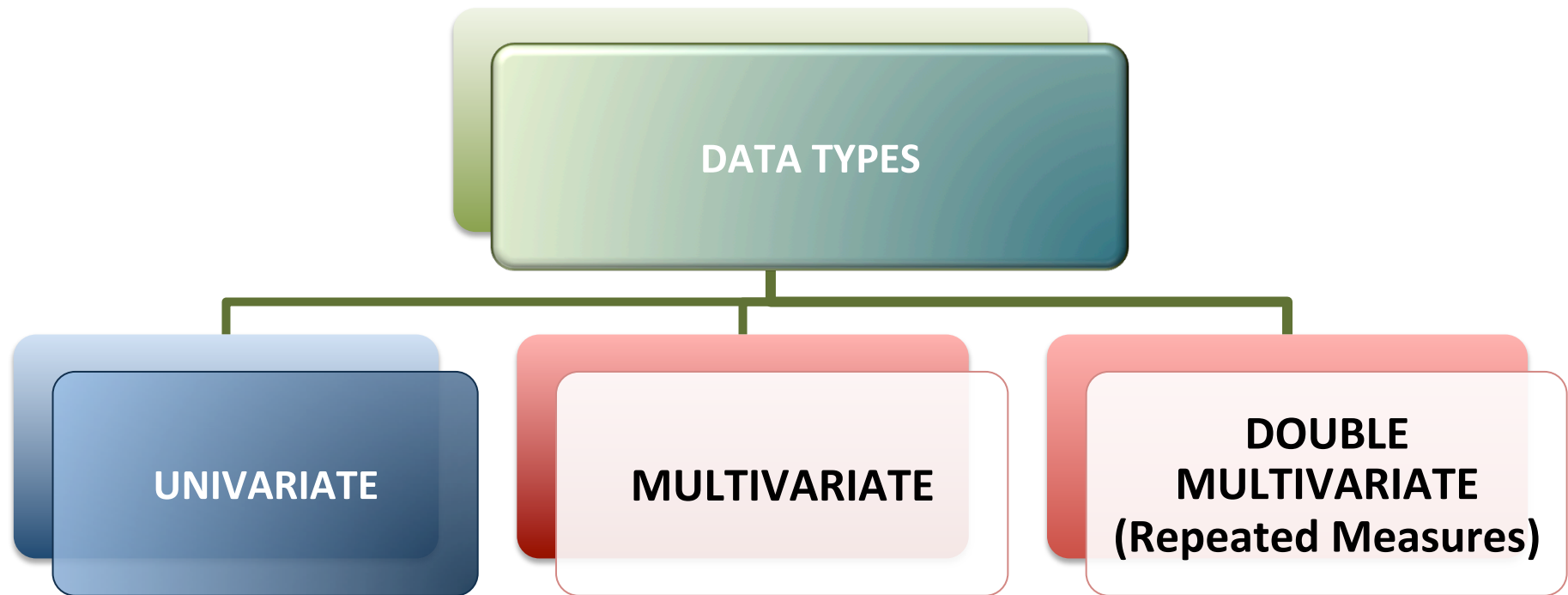
QUESTION/PROBLEM

PATTERN RECOGNITION		
	UNSUPERVISED (NO PRIOR KNOWLEDGE)	SUPERVISED (PRIOR KNOWLEDGE)
PATTERNS OF "SIMILARITY" BETWEEN VARIABLES	ORDINATION PCA FACTOR ANALYSIS	DISCRIMINANT ANALYSIS GLM REGRESSION PATH ANALYSIS
PATTERNS OF "SIMILARITY" BETWEEN INDIVIDUALS	ORDINATION PRINCIPAL COMPONENT ANALYSIS MDS CLUSTER ANALYSIS	

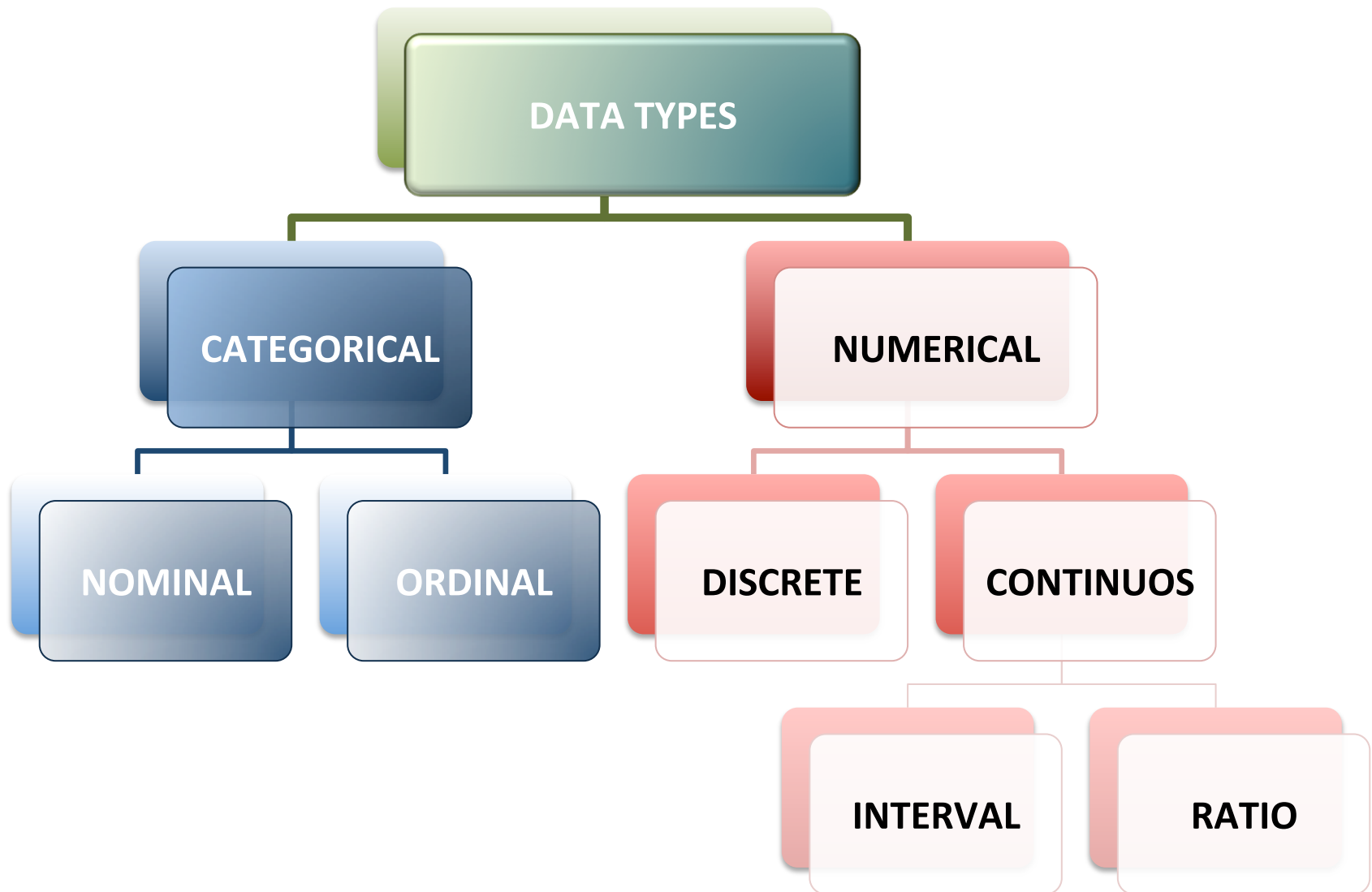
DATA COLLECTION



DATA COLLECTION: DATA TYPES



DATA COLLECTION: DATA TYPES



DATA COLLECTION: DATA TYPES

		Response Variable	
Explanatory Variable	Categorical (Nominal or Ordinal)	Categorical (Nominal or Ordinal)	Numerical (Interval or Ratio)
	Numerical (Interval or Ratio)		

DATA COLLECTION: DATA TYPES

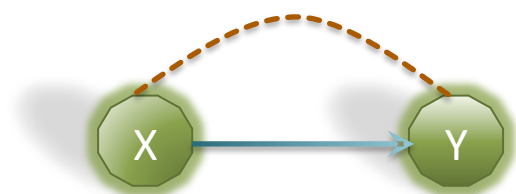
		Response Variable	
Explanatory Variable	Categorical (Nominal or Ordinal)	Categorical (Nominal or Ordinal)	Numerical (Interval or Ratio)
	Numerical (Interval or Ratio)	Chi-square Analysis Through Crosstabulation Logistic Regression Log-linear Models	Independent/Dependent t-test ANOVA Regression Correlation

DATA COLLECTION: TYPES OF RELATIONSHIPS

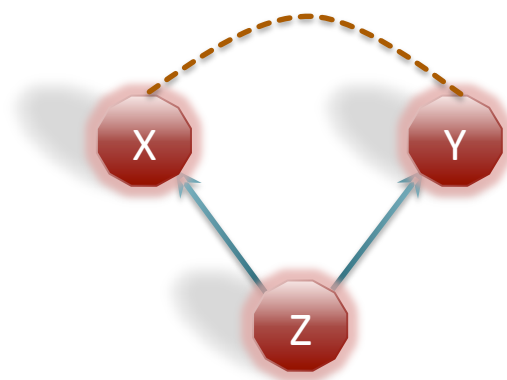
Association/Correlation does not imply Causation

Dependence does not imply Causation

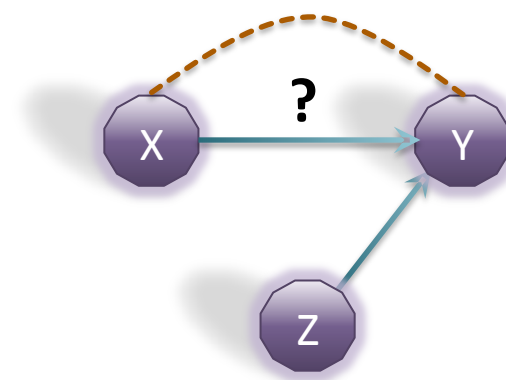
(but it sure is a hint Lynd & Stevenson (2007), Tufte (2006), von Eye & DeShon (2011)).



CAUSAL



COMMON RESPONSE



CONFOUNDING



Association/Correlation

Cause-and-effect Relationship

DATA COLLECTION: DESIGN OF EXPERIMENTS

Causal relationships can only be set through experiments.

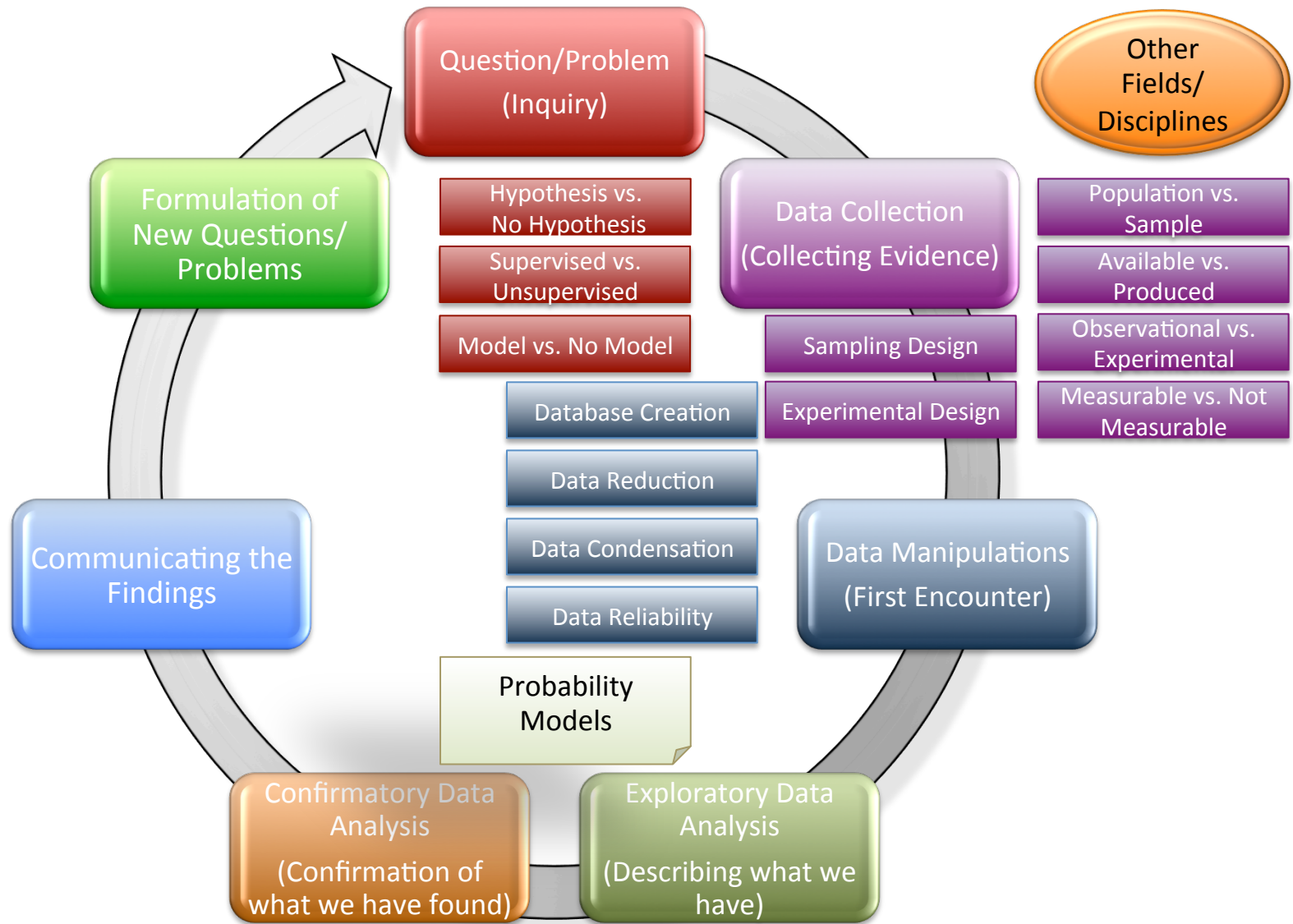
PRINCIPALS OF DESIGN OF EXPERIMENTS

CONTROL

RANDOMIZE

REPLICATE

DATA MANIPULATIONS



DATA MANIPULATIONS: DATA RELIABILITY

Data reliability is a state that exists when data is sufficiently complete and error free to be convincing for its purpose and context.

- **COMPLETE:** Includes all of the data elements (variables/fields) needed for the analysis
- **ACCURATE:**
 - **CONSISTENT:** The data was obtained and used in a manner that is clear and well-defined enough to yield similar results in similar analysis
 - **CORRECT:** The data set reflects the data entered at the source and/or properly represents the intended results.
- **UNALTERED:** The data reflects source and has not been tampered with.

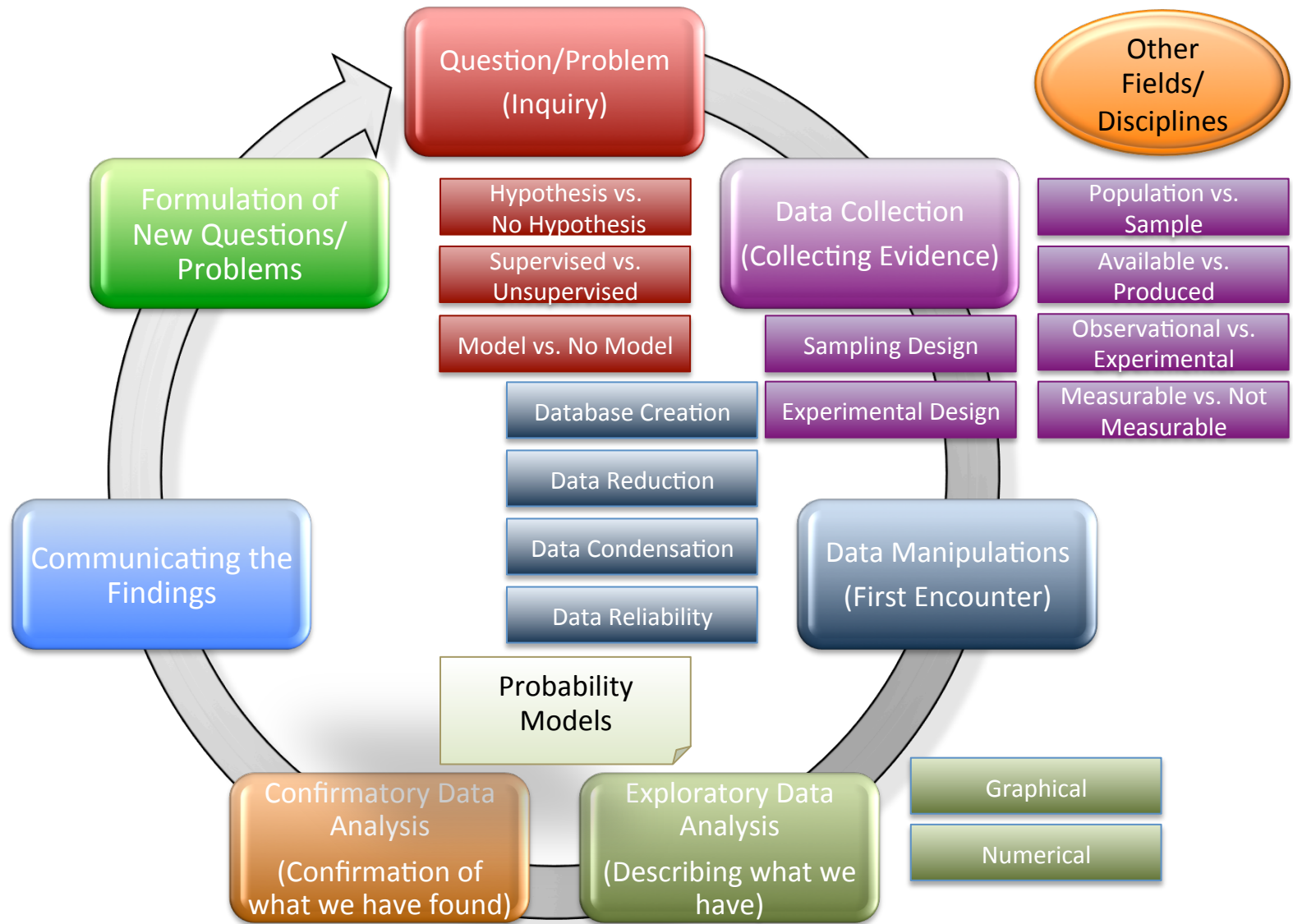
DATA MANIPULATIONS: DATABASE

Database is an organized collection of data

- Easy to use (data entry and data manipulations)
 - Dynamic
 - Interactive
 - Open to collaboration
 - Integrated
-
- Piece of paper
 - Word processor (Microsoft Word)
 - Microsoft Excel
 - Microsoft Access
 - Statistical software package (R, StatCrunch, SPSS, SAS etc.)
 - Any program that uses SQL (Structured Query Language)
 - Google Docs
 - [Google Fusion Tables](#)

(UMM Data Services Center: <http://mnstats.morris.umn.edu/UMMDataServicesCenter.html>)

EXPLORATORY DATA ANALYSIS

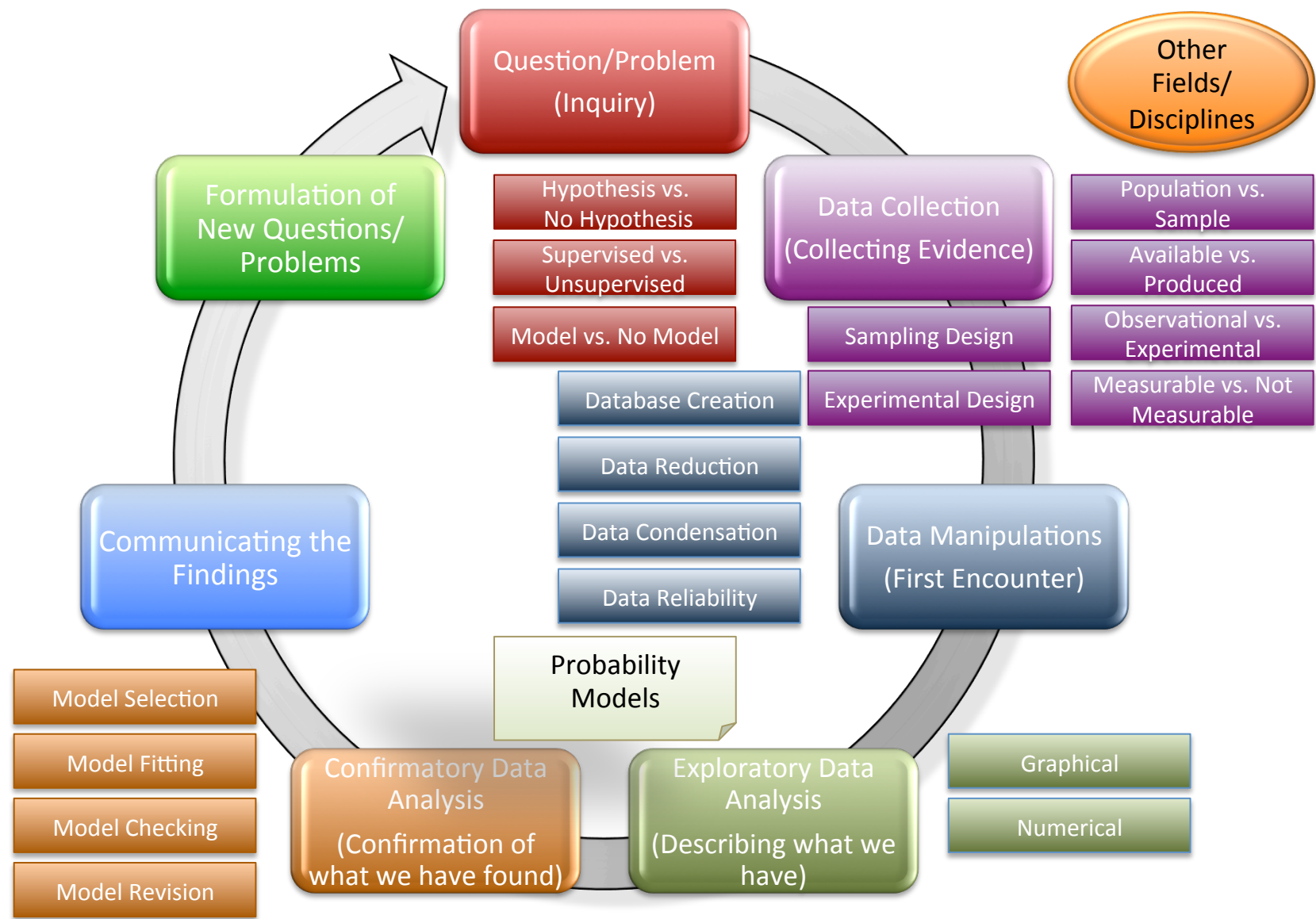


EXPLORATORY DATA ANALYSIS

- ☐ [Dynamic](#)
- ☐ [Interactive](#)
- ☐ Database integrated graphical displays

Correct selection of numerical and graphical summary techniques and methods

CONFIRMATORY DATA ANALYSIS



CONFIRMATORY DATA ANALYSIS

$$Y_1, Y_2, \dots, Y_n \text{ are i.i.d. from } N(\mu, \sigma^2)$$

$$Y_1, Y_2, \dots, Y_n \Rightarrow DATA$$

i.i.d. \Rightarrow independent and identically distributed

\Rightarrow Simple random sample from the same population

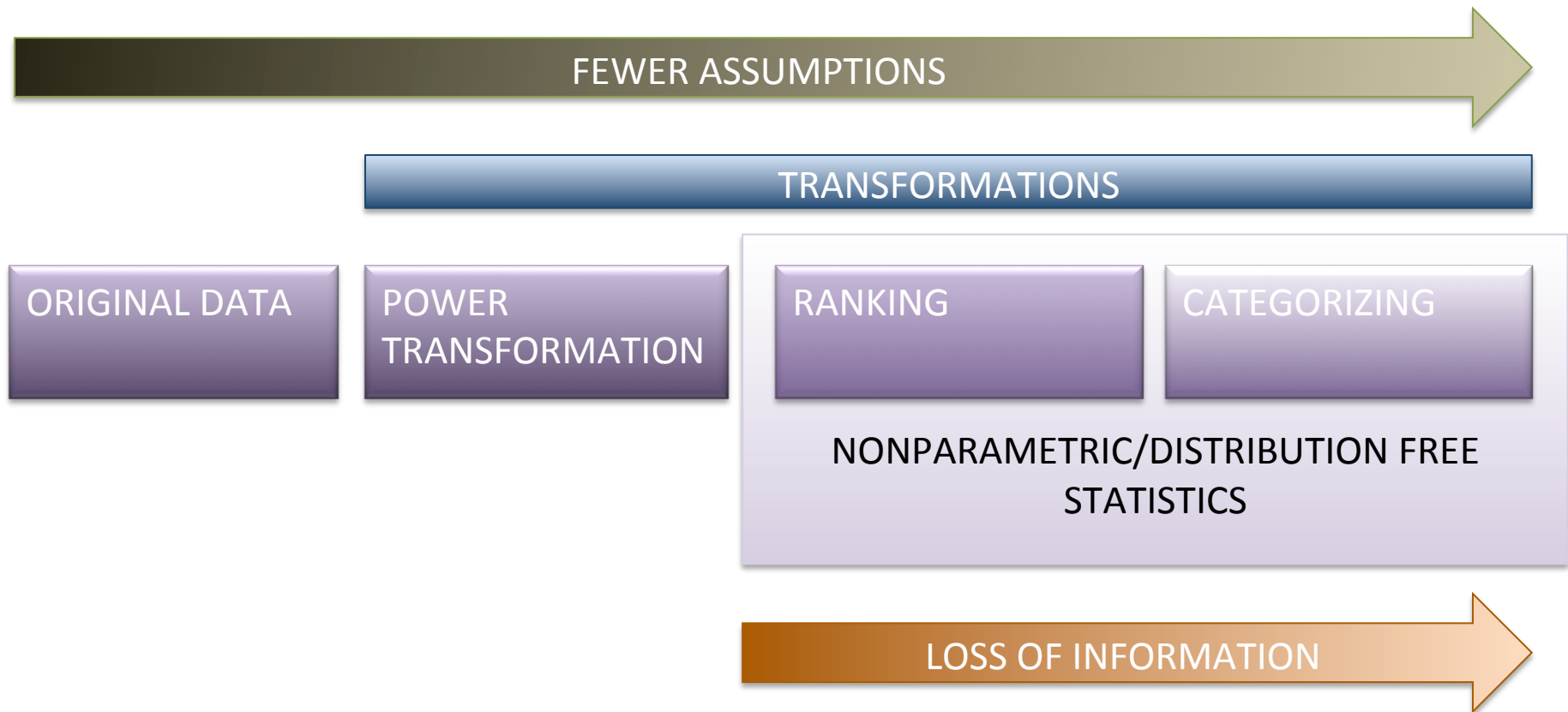
$$N(\mu, \sigma^2) \Rightarrow \text{Normal Distribution}$$

$$Y_1, Y_2, \dots, Y_n \text{ are i.i.d. from } N\left(\beta_0 + \sum_{i=1}^p \beta_i \mu_{X_i}, \sigma^2\right)$$

$$\beta_0 + \sum_{i=1}^p \beta_i \mu_{X_i} \Rightarrow \text{Linear Model } N(\dots, \sigma^2) \Rightarrow \text{Constant Variance}$$

CONFIRMATORY DATA ANALYSIS: TRANSFORMATIONS

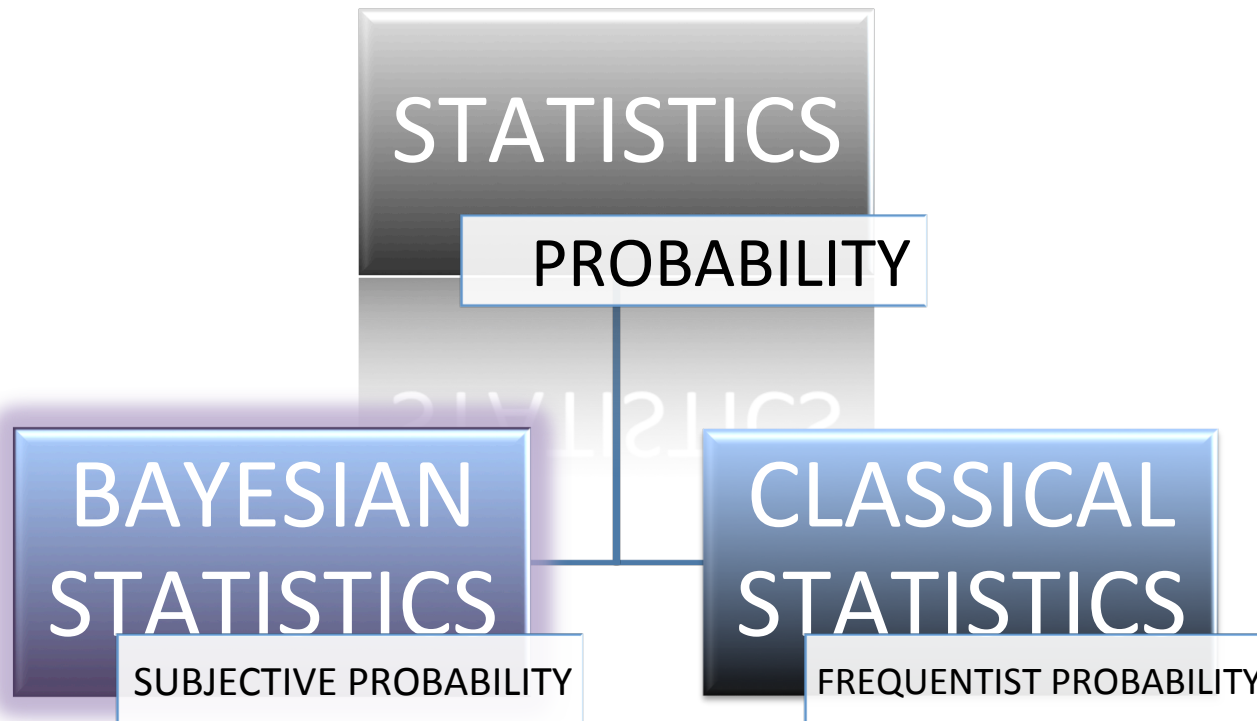
WHAT TO DO WHEN THE MODEL ASSUMPTIONS ARE VIOLATED?



CONFIRMATORY DATA ANALYSIS: NONPARAMETRIC STATISTICS

- RANK-BASED METHODS
- PERMUTATION TESTS
R.A. FISHER (1935)
- BOOTSRAP METHODS
TAKE A SAMPLE OF SAME SIZE FROM THE SAMPLE WITH REPLACEMENT
- CURVE SMOOTHING
NO LINEAR OR NONLINEAR MODEL

CONFIRMATORY DATA ANALYSIS



CONFIRMATORY DATA ANALYSIS

How would you describe in plain English the characteristics that distinguish Bayesian from Frequentist reasoning?
(<http://stats.stackexchange.com>)

Here is how I would explain the basic difference to my grandma:

I have misplaced my phone somewhere in the home. I can use the phone locator on the base of the instrument to locate the phone and when I press the phone locator the phone starts beeping.

Problem: Which area of my home should I search?

Frequentist Reasoning:

I can hear the phone beeping. I also have a **mental model** which helps me identify the area from which the sound is coming from. Therefore, upon hearing the beep, I infer the area of my home I must search to locate the phone.

Bayesian Reasoning:

I can hear the phone beeping. Now, apart from a **mental model** which helps me identify the area from which the sound is coming from, I also know the locations where I have misplaced the phone in the past. So, I combine my inferences using the beeps and my **prior information** about the locations I have misplaced the phone in the past to identify an area I must search to locate the phone.

CONFIRMATORY DATA ANALYSIS

Tongue firmly in cheek:

A **Bayesian** defines a "probability" in **exactly the same way that most non-statisticians do** - namely an indication of the plausibility of a proposition or a situation. If you ask him a question, he will give you **a direct answer** assigning probabilities describing the plausibilities of the possible outcomes for the particular situation (and state his prior assumptions).

A **Frequentist** is someone that believes probabilities represent long run frequencies with which events occur; if needs be, he will **invent a fictitious population** from which your particular situation could be considered a random sample so that he can meaningfully talk about long run frequencies. If you ask him a question about a particular situation, he will **not give a direct answer**, but instead make a statement about this (possibly imaginary) population. Many non-frequentist statisticians will be easily **confused** by the answer and interpret it as Bayesian probability about the particular situation.

P-VALUE? <https://www.youtube.com/watch?feature=endscreen&NR=1&v=ax0tDcFkPic>
<https://www.youtube.com/watch?v=ez4DgdurRPg>

CONFIRMATORY DATA ANALYSIS

Very crudely I would say that:

Frequentist: Sampling is infinite and decision rules can be sharp. Data are a **repeatable random sample** - there is a frequency. **Underlying parameters are fixed i.e. they remain constant during this repeatable sampling process.**

Bayesian: Unknown quantities are treated probabilistically and the **state of the world can always be updated.** Data are observed from the realised sample. Parameters are unknown and described probabilistically. **It is the data which are fixed.**

CONFIRMATORY DATA ANALYSIS: SOME TECHNIQUES

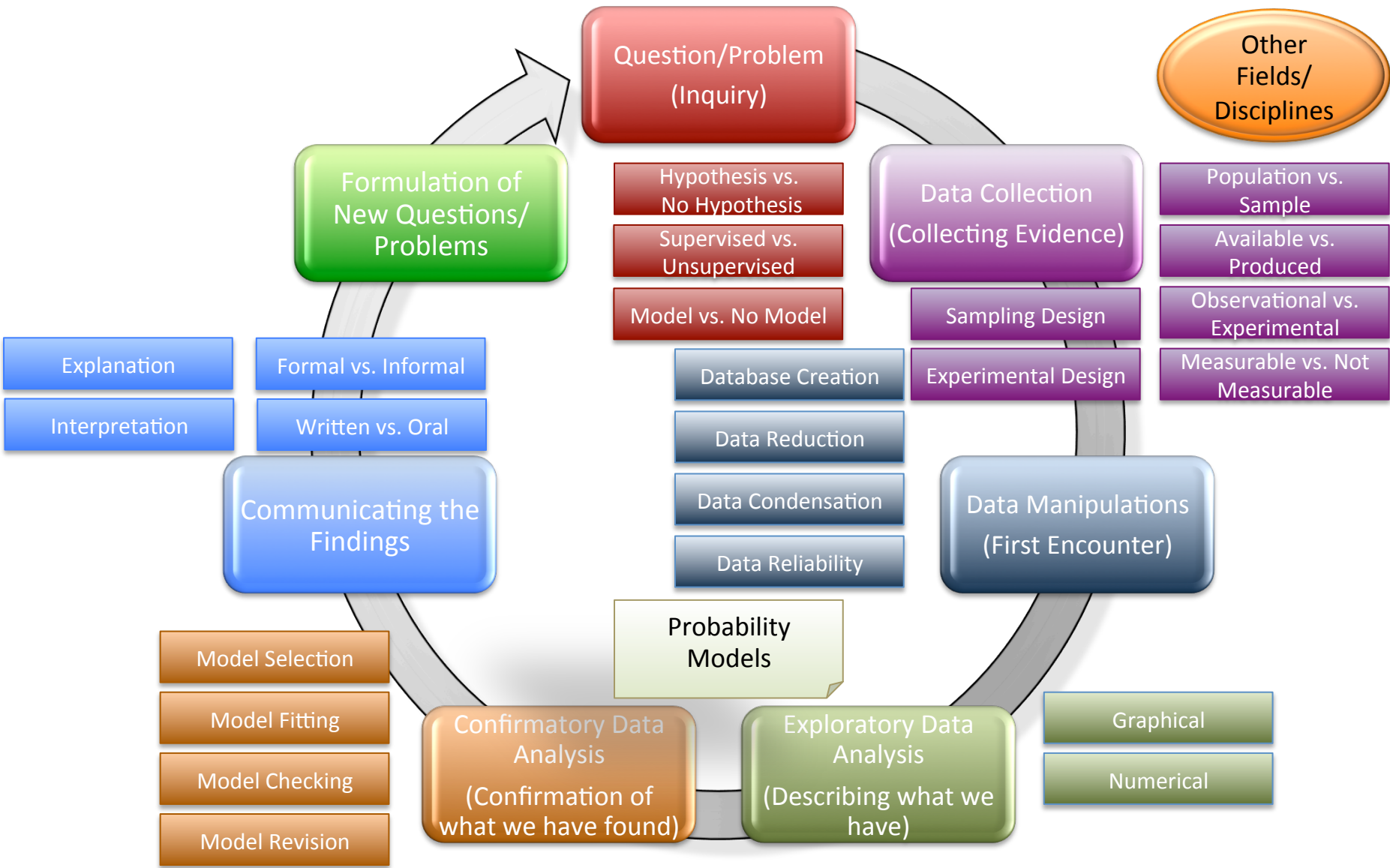
◆ MULTIVARIATE TECHNIQUES

<http://mnstats.morris.umn.edu/multivariatestatistics/overview.html>

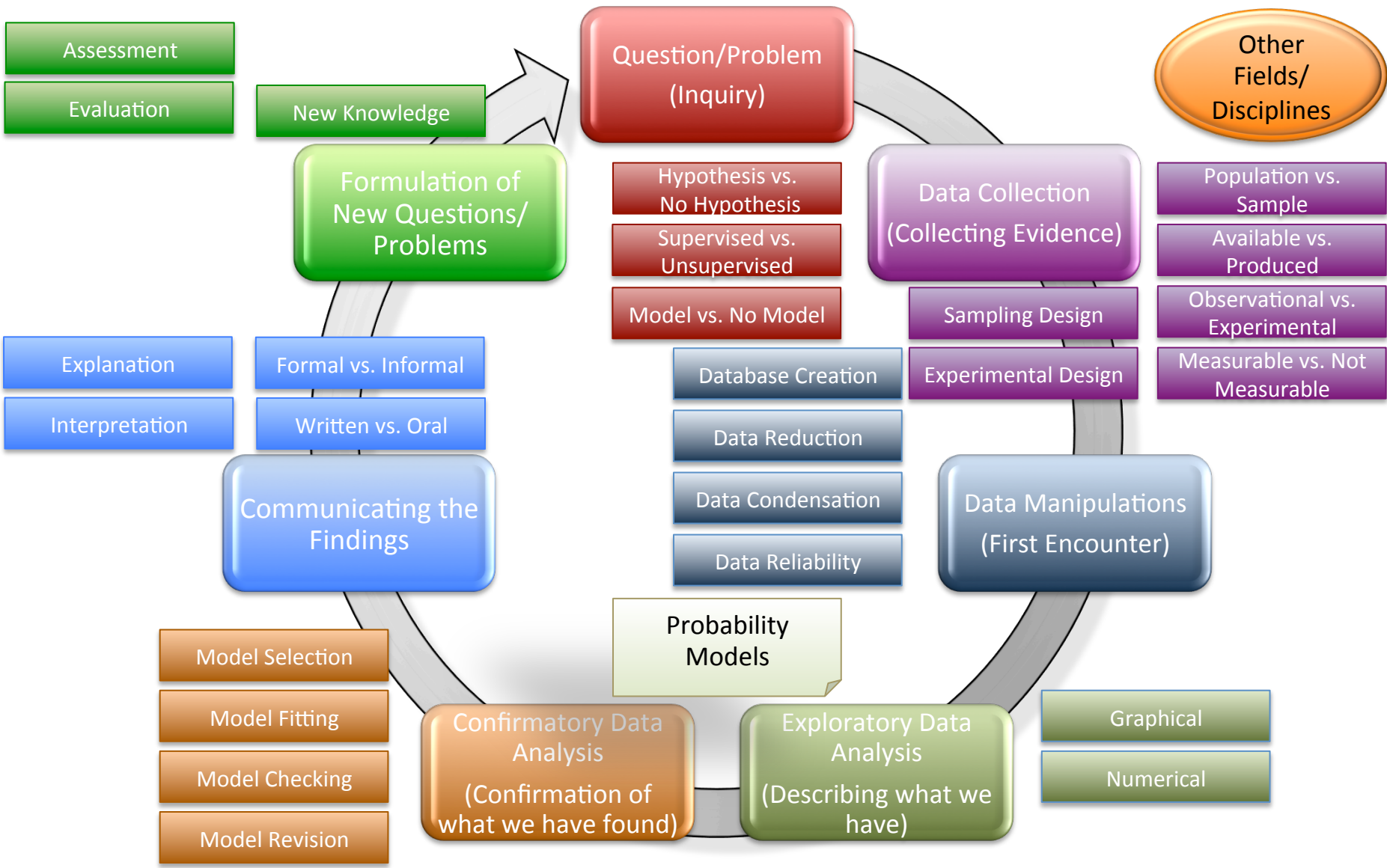
◆ NONPARAMETRIC/DISTRIBUTION FREE TECHNIQUES

<http://mnstats.morris.umn.edu/introstat/nonparametric/learningtools.html>

COMMUNICATING THE FINDINGS



FORMULATING NEW QUESTIONS/PROBLEMS



CONCLUDING REMARKS

- SEE ME FOR A HELP
- QUESTIONS?
- IF NOT, I HAVE SOME FOR YOU. PLEASE
TAKE THE TEST BEFORE YOU LEAVE