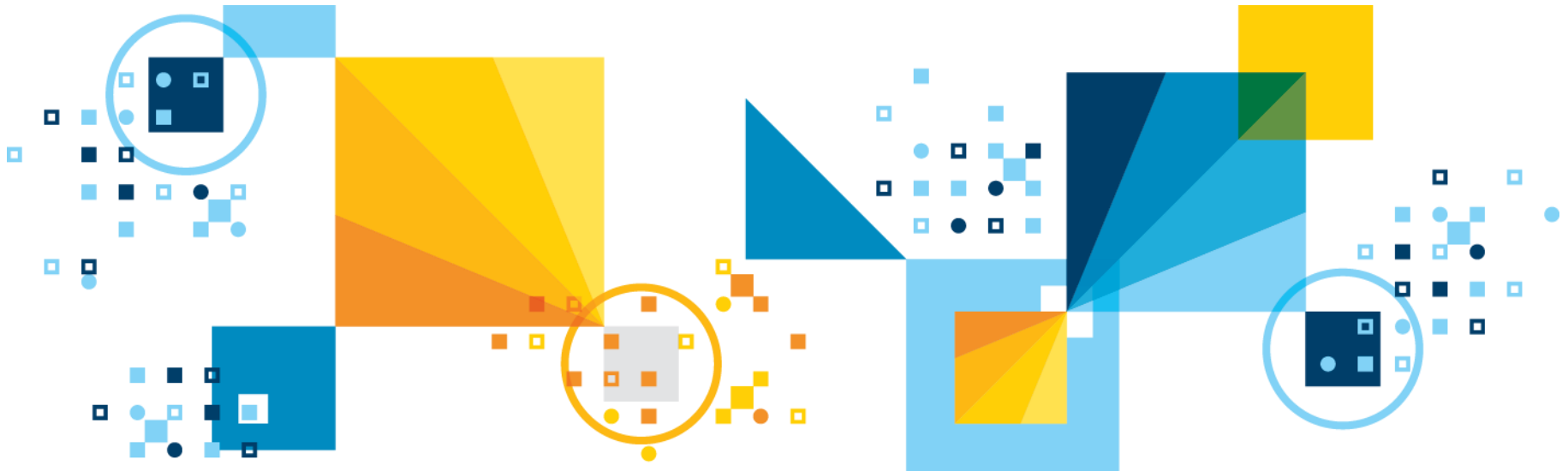


ResearchMap for Data Science

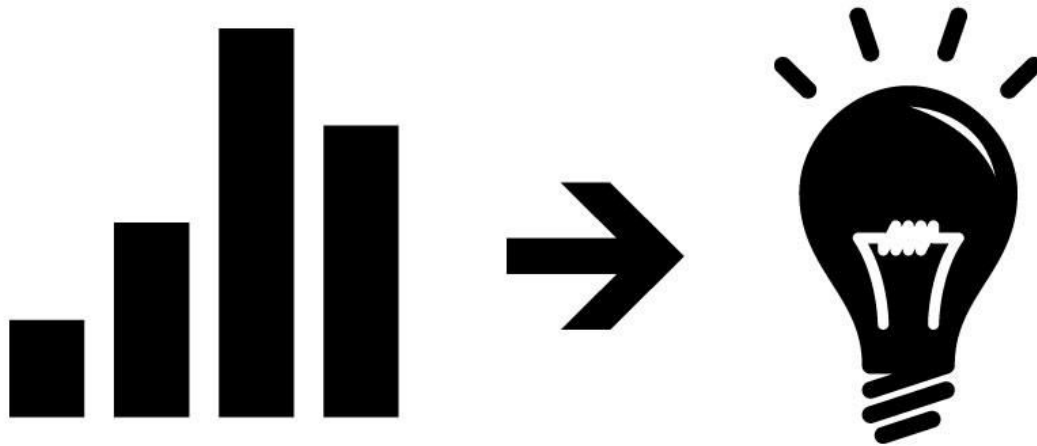
Dr. Alex Liu, Principal Data Scientist

September 21, 2015



Data Science Definition

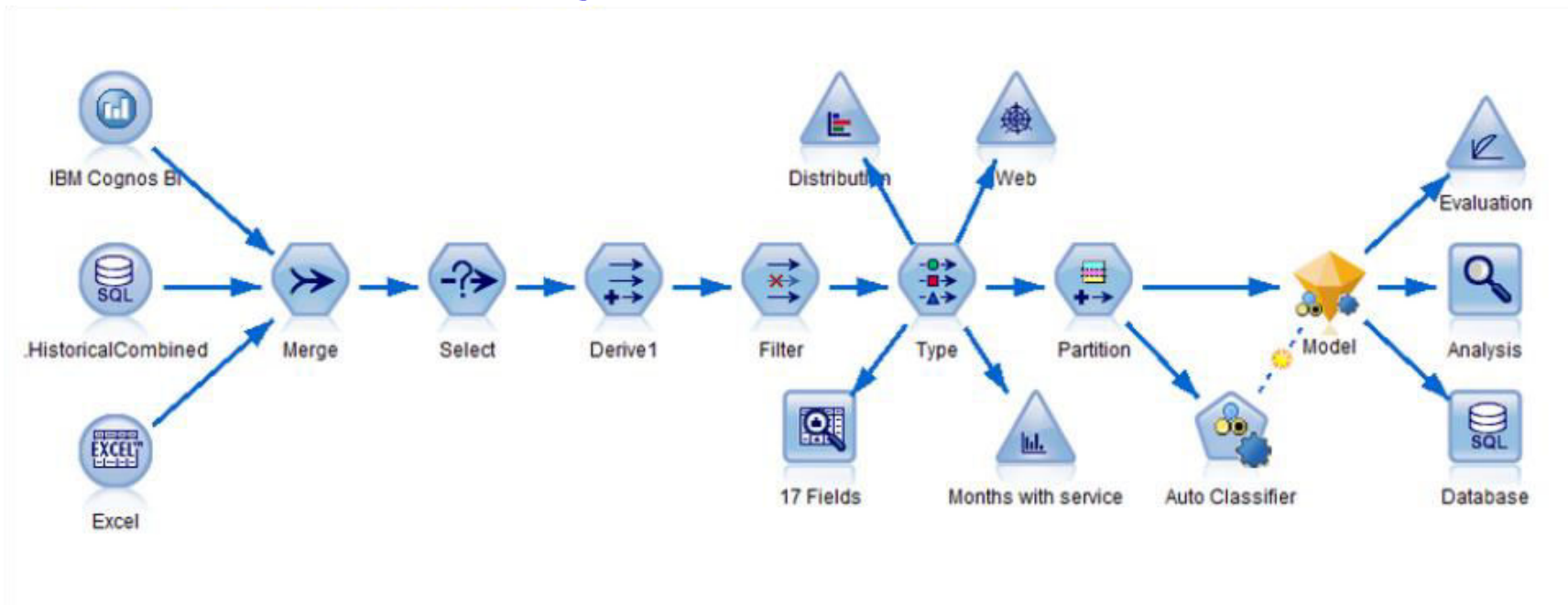
- **Data Science** is an interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of **data** in various forms either structured or unstructured, which is a continuation of some of the data analysis fields such as **data** mining and predictive analytics, as well as knowledge discovery and **data** mining (KDD).
- **Data Science** is about turning data into insights.



Data Science is a process

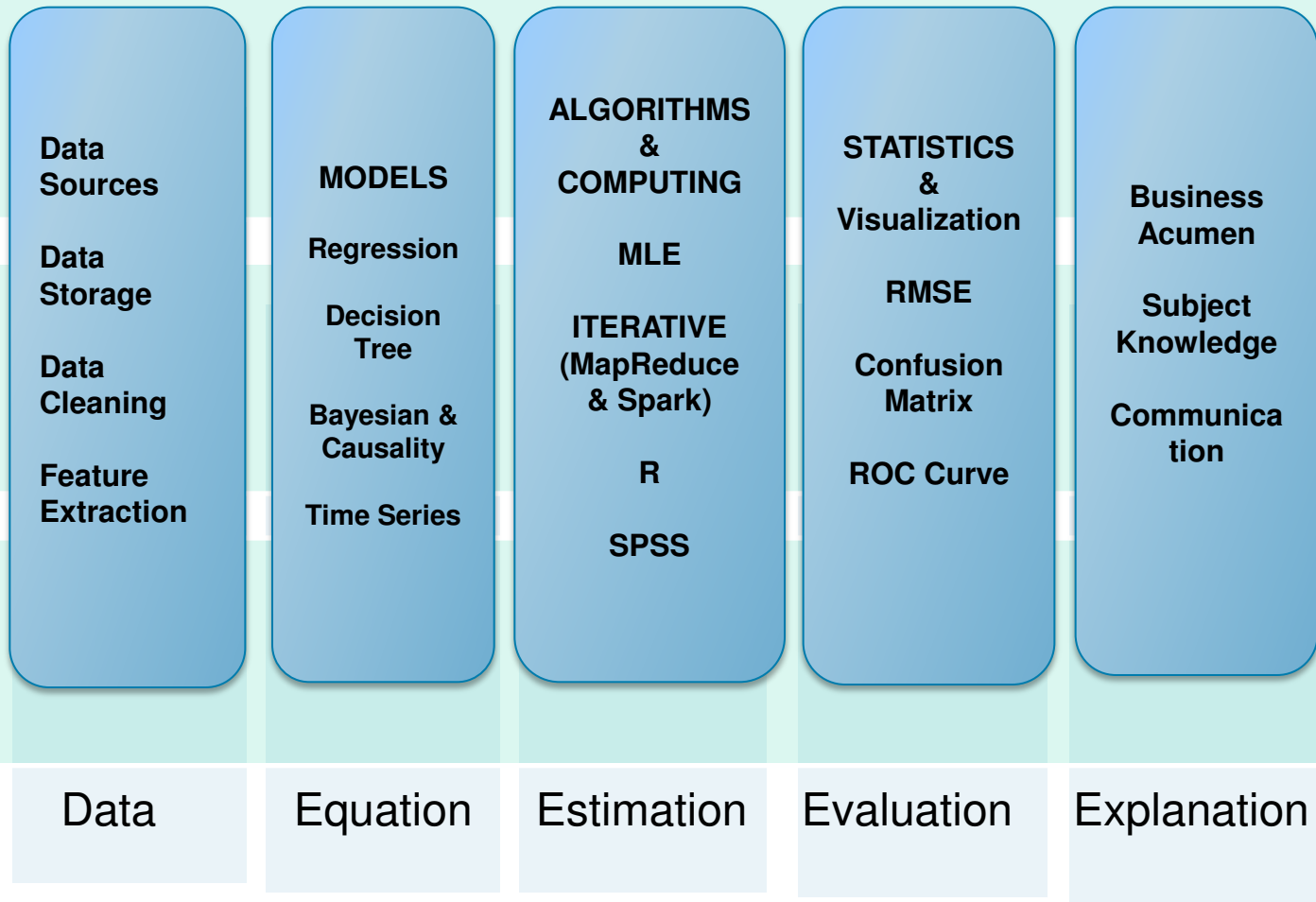


SPSS on Hadoop

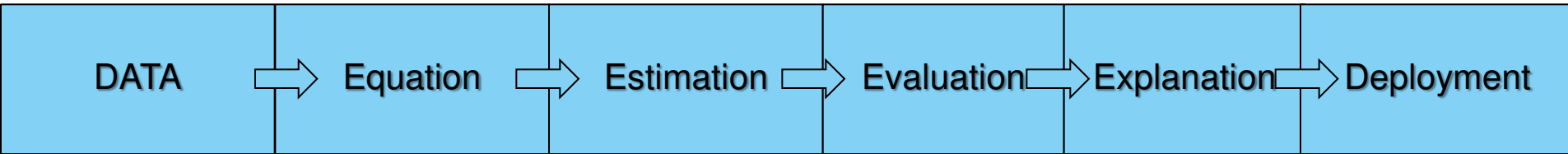


4Es – Equation – Estimation – Evaluation – Explanation – as foundation

Data Science Framework – RM4Es Based Workflow

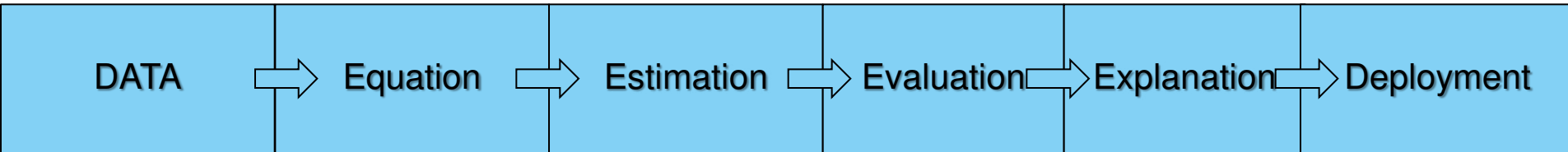


ResearchMap for Data Science



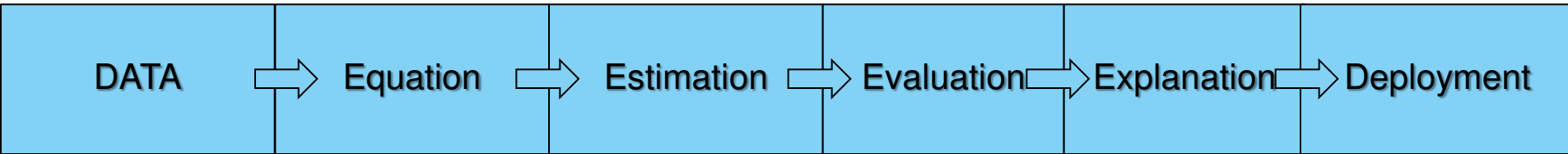
- Hadoop HDFS
- Data Sources
- Data Loading
- Data Cleaning
- Data re-organizing
- Feature Development

ResearchMap for Data Science



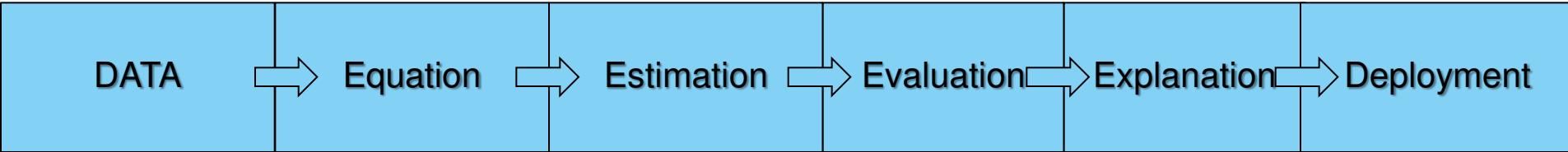
- Business Case Development
- Concept Operationalization
- Feature Selection
- Dimension Reduction
- Model Selection


ResearchMap for Data Science



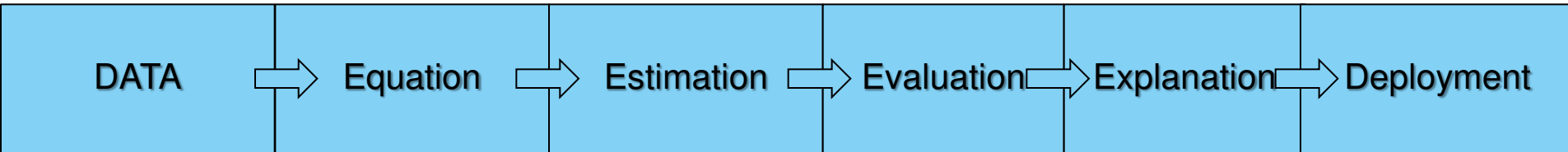
- Algorithms
- Iterative Computing
- Parallel Computing
- Map Reduce
- Spark

ResearchMap for Data Science



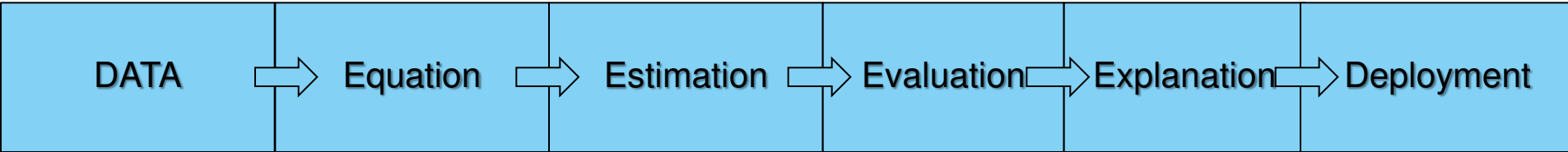
- 
- Statistics
 - RMSE
 - Confusion Matrix
 - ROC Curve

ResearchMap for Data Science



- 
- Subject Knowledge
 - Visualization
 - Communication
 - ROI

ResearchMap for Data Science



- 
- Experiments
 - Tests
 - Implementation
 - Monitoring

Past Work

- Apache Taverna at <https://taverna.incubator.apache.org/>
- Kepler at <https://kepler-project.org/>
- KNIME at <https://www.knime.org/>
- RM4Es System at <http://www.researchmethods.org/userguide.htm>
- ...