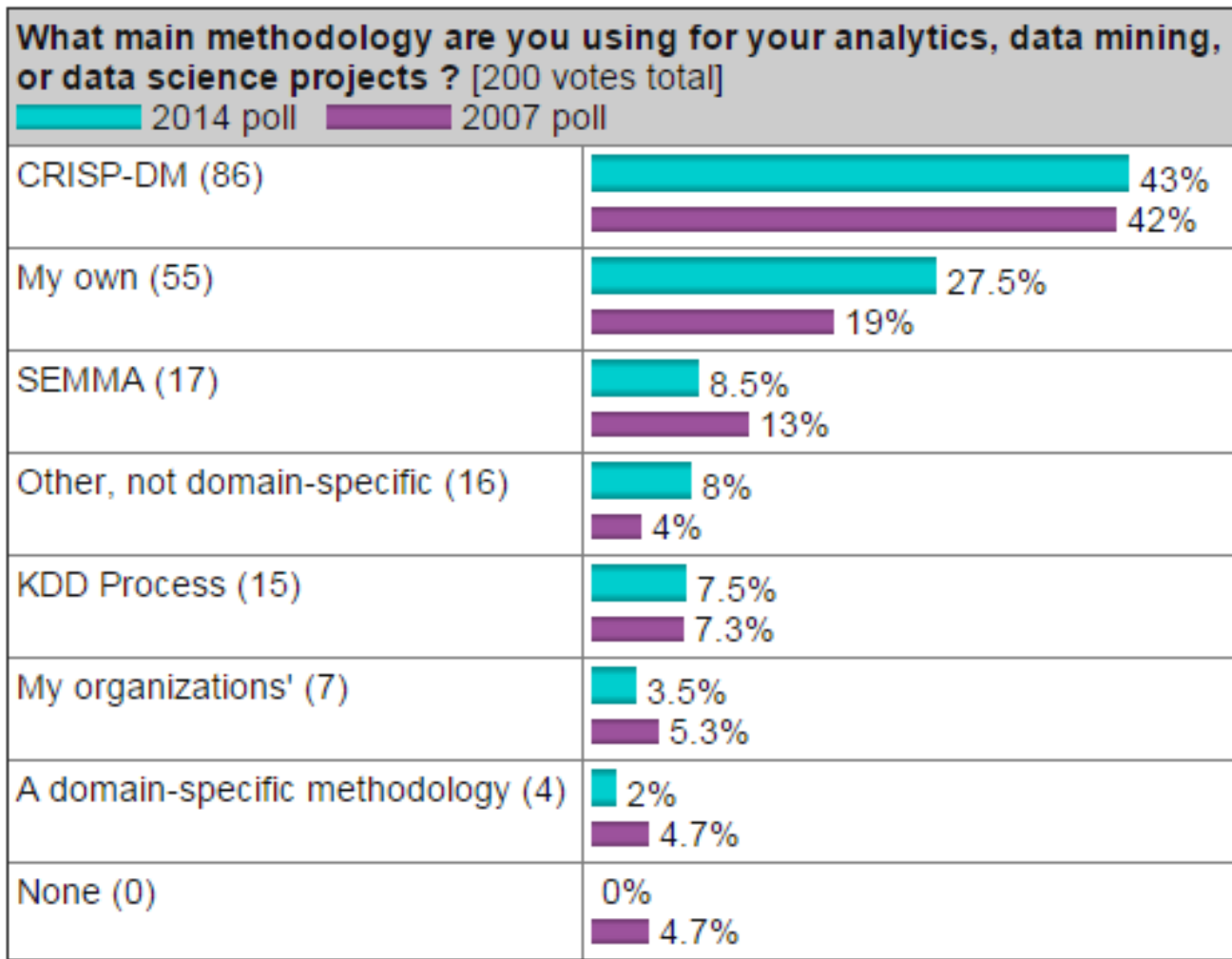


The background is a light gray gradient. It features several realistic water droplets of various sizes, some with highlights and shadows, scattered across the frame. In the upper center, there is a faint, circular, textured pattern that resembles a lens flare or a subtle watermark.

BUSINESS ANALYTICS

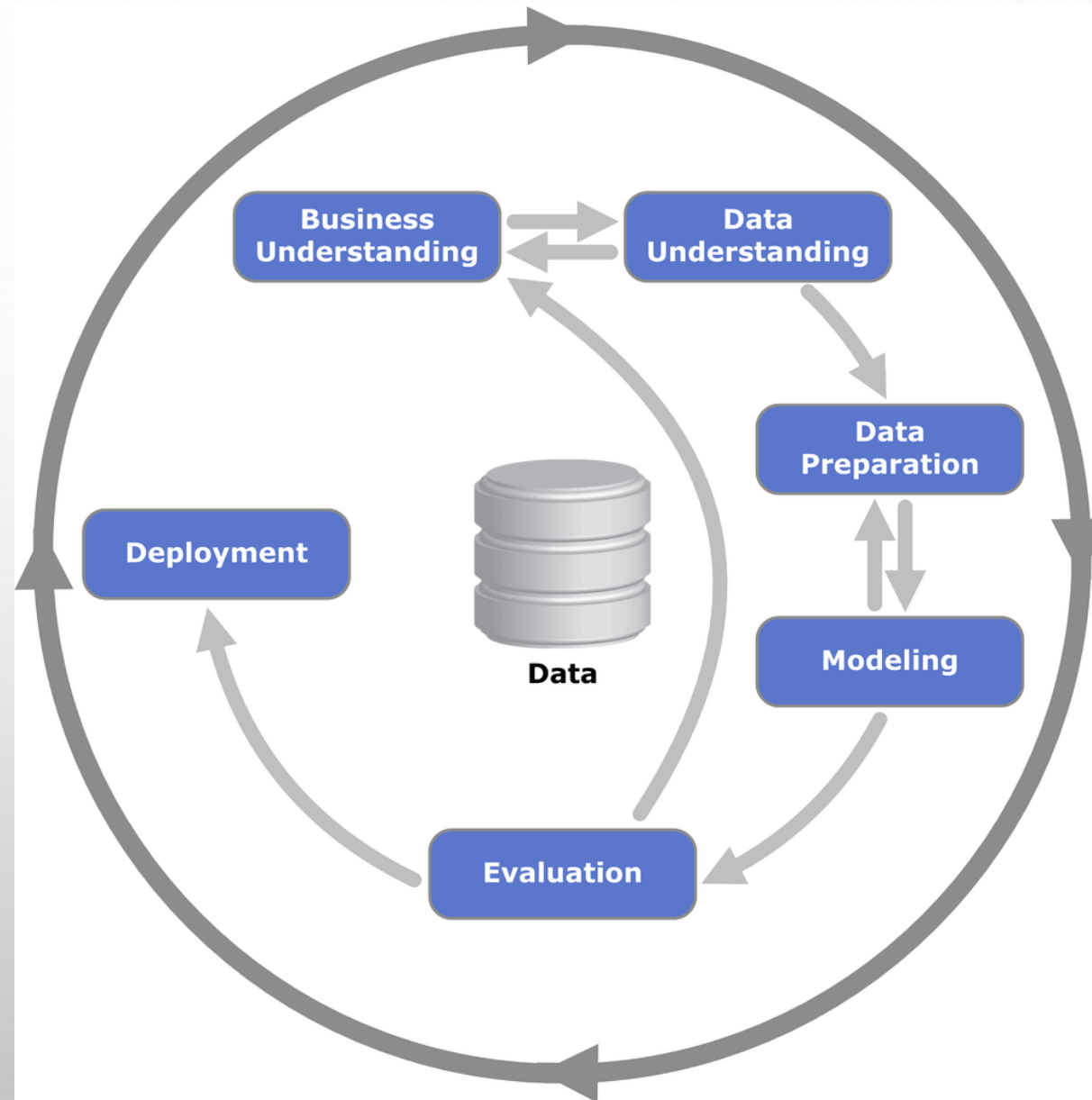
DATA MINING PROCESS

- A MANIFESTATION OF BEST PRACTICES
- A SYSTEMATIC WAY TO CONDUCT DM PROJECTS
- MOST COMMON STANDARD PROCESSES:
 - CRISP-DM (CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING)
 - SEMMA (SAMPLE, EXPLORE, MODIFY, MODEL, AND ASSESS)
 - KDD (KNOWLEDGE DISCOVERY IN DATABASES)



Source: KDNuggets.com, Oct 2014

DATA MINING PROCESS: CRISP-DM



DATA MINING PROCESS: CRISP-DM

STEP 1: BUSINESS UNDERSTANDING

STEP 2: DATA UNDERSTANDING

STEP 3: DATA PREPARATION (!)

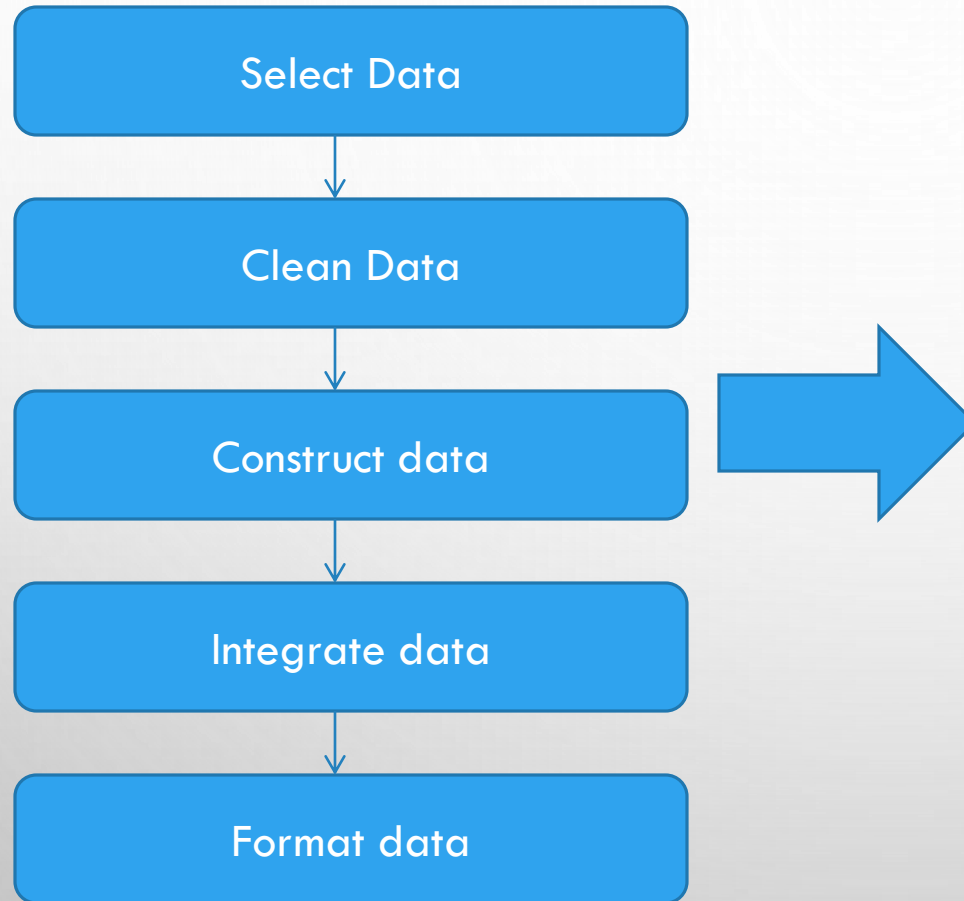
STEP 4: MODEL BUILDING

STEP 5: TESTING AND EVALUATION

STEP 6: DEPLOYMENT

- THE PROCESS IS HIGHLY REPETITIVE AND EXPERIMENTAL

DATA PREPARATION – A CRITICAL DM TASK



Output:

Dataset and a dataset description,
both used for modeling or for the
major analysis work of the project.

DATA PREPARATION – A CRITICAL TASK

Select Data

- ACTIVITIES:
 - SOURCE DATA COLLECTED.
 - SIGNIFICANCE AND CORRELATION TESTS TO DETERMINE INCLUSION
 - DATA SELECTION CRITERIA REVISITED (STEP2)
 - DATA SUBSETS
 - SAMPLING TECHNIQUES
 - DOCUMENT RATIONALE FOR INCLUSION/EXCLUSION
- OUTPUT:
 - LIST OF DATA TO BE USED/EXCLUDED AND REASONS

DATA PREPARATION – A CRITICAL TASK

Clean Data

- ACTIVITIES:
 - NOISE
 - SPECIAL VALUES.
 - DATA SELECTION CRITERIA
- OUTPUT:
 - DATA CLEANING REPORT

DATA PREPARATION – A CRITICAL TASK

Construct Data

- ACTIVITIES:
 - CONSTRUCTION MECHANISMS AND TOOLS
 - PRODUCTION OF DERIVED ATTRIBUTES, NEW RECORDS, TRANSFORMED VALUES
 - SELECTION CRITERIA
- OUTPUT:
 - DERIVED ATTRIBUTES

DATA PREPARATION – A CRITICAL TASK

Integrate Data

- ACTIVITIES:
 - INTEGRATE SOURCES
 - DATA SELECTION CRITERIA
- OUTPUT:
 - MERGED DATA

DATA PREPARATION – A CRITICAL TASK

Format Data

- ACTIVITIES:
 - REARRANGING ATTRIBUTES
 - REORDERING RECORDS
 - REFORMATTING VALUES
- OUTPUT:
 - REFORMATTED DATA