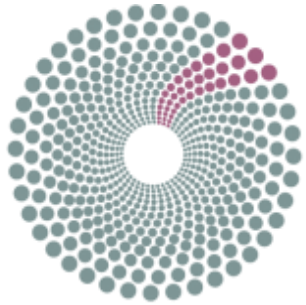
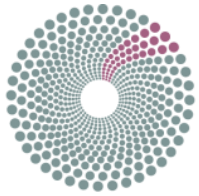


BIG DATA EUROPE

Empowering Communities
with Data Technologies

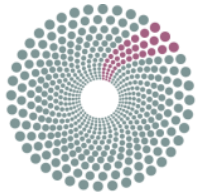
Methodologies for Developing Big Data Applications

Ivan Ermilov @ ICTCS, Amman, Jordan



Outline

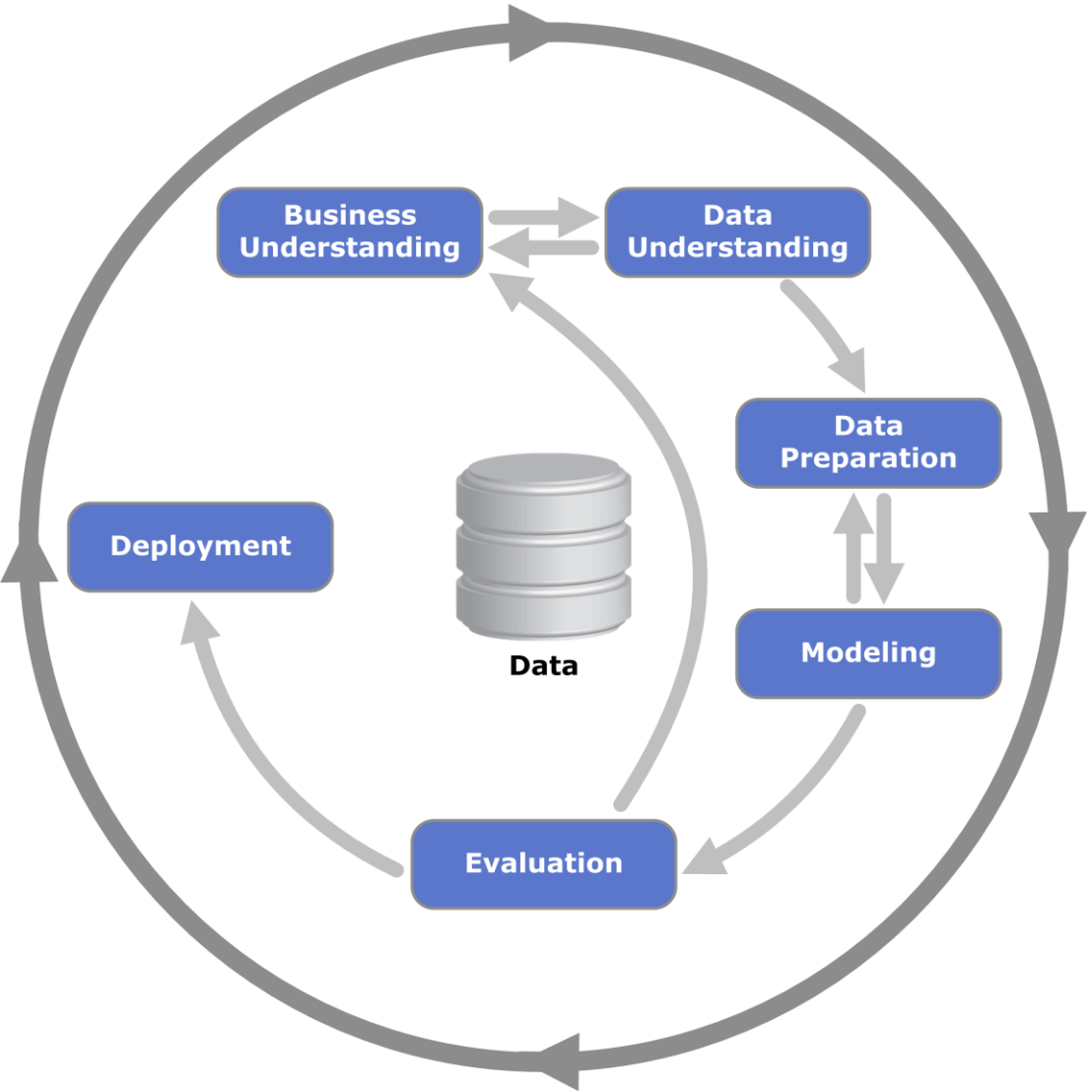
- CRISP DM
- Adaptations of CRISP DM for Big Data
- Examples of Big Data Applications
- BDI SL Methodology



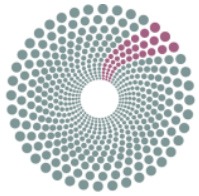
CRISP DM

- CRoss-Industry Standard Process for Data Mining
- Why Data Mining?
- Developed by industry leaders
- Industry-, tool- and application-neutral model

CRISP DM

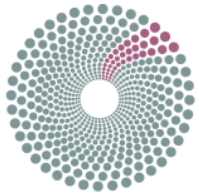


- Six Phases
- Data centric
- No team communications
- No development methodologies (e.g. agile)



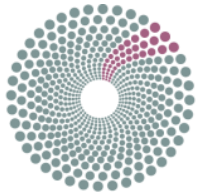
Business Understanding

- Determine business objectives
  - Background
  - Business objectives
  - Business success criteria
- Assess situation
  - Inventory of resources
  - Requirements, assumptions, and constraints
- Determine data mining goals
  - Data mining goals & success criteria
- Produce Project Plan



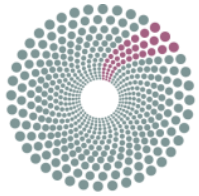
Data Understanding

- Collect initial data
  - Initial data collection report
- Describe data
  - Data description report
- Explore data
  - Data exploration report
- Verify data quality
  - Data quality report



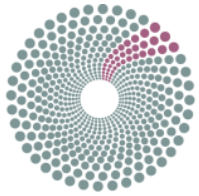
Data Preparation

- Dataset
  - Dataset description
- Select data
  - Rationale for inclusion/exclusion
- Clean data
  - Data cleaning report
- Construct data
  - Derived attributes & generated records
- Integrate data
- Format data



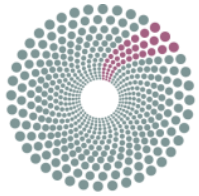
Modelling

- Select modelling technique
  - Modelling technique & assumptions
- Generate test design
- Build model
  - Parameter settings
  - Models
  - Model description
- Assess Model
  - Model Assessment
  - Revised parameter settings



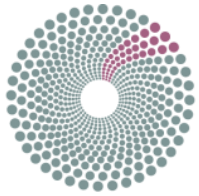
Evaluation

- Evaluate results
  - Assessment of data mining results wrt business success criteria
  - Approved models
- Review process
  - Review of process
- Determine next steps
  - List of possible actions
  - Decision



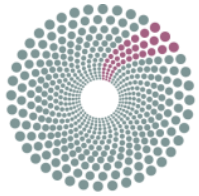
Deployment

- Plan deployment
  - Deployment plan
- Plan monitoring and maintenance
  - Monitoring and maintenance plan
- Produce final report
  - Final report
  - Final presentation
- Review project
  - Experience documentation



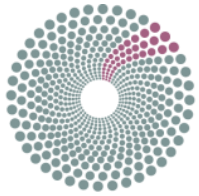
CRISP DM for SNA

- Data acquisition
  - Initial keywords
- Data cleaning
  - Compound or nested filtering
- Data formatting
  - Unstructured → structured data (e.g. Hive)
- Data validation
  - e.g. checking for a power-law distribution



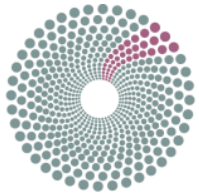
CRISP DM for SNA

- Data Analysis
  - Based on identified questions
- Deployment
  - Documentation and results



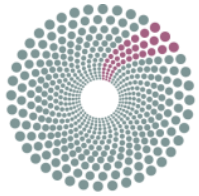
CRISP DM: Gaps

- Project management perspective
- Software development methodology
  - Waterfall
  - Agile
- Team communication



Big Data Apps: Examples

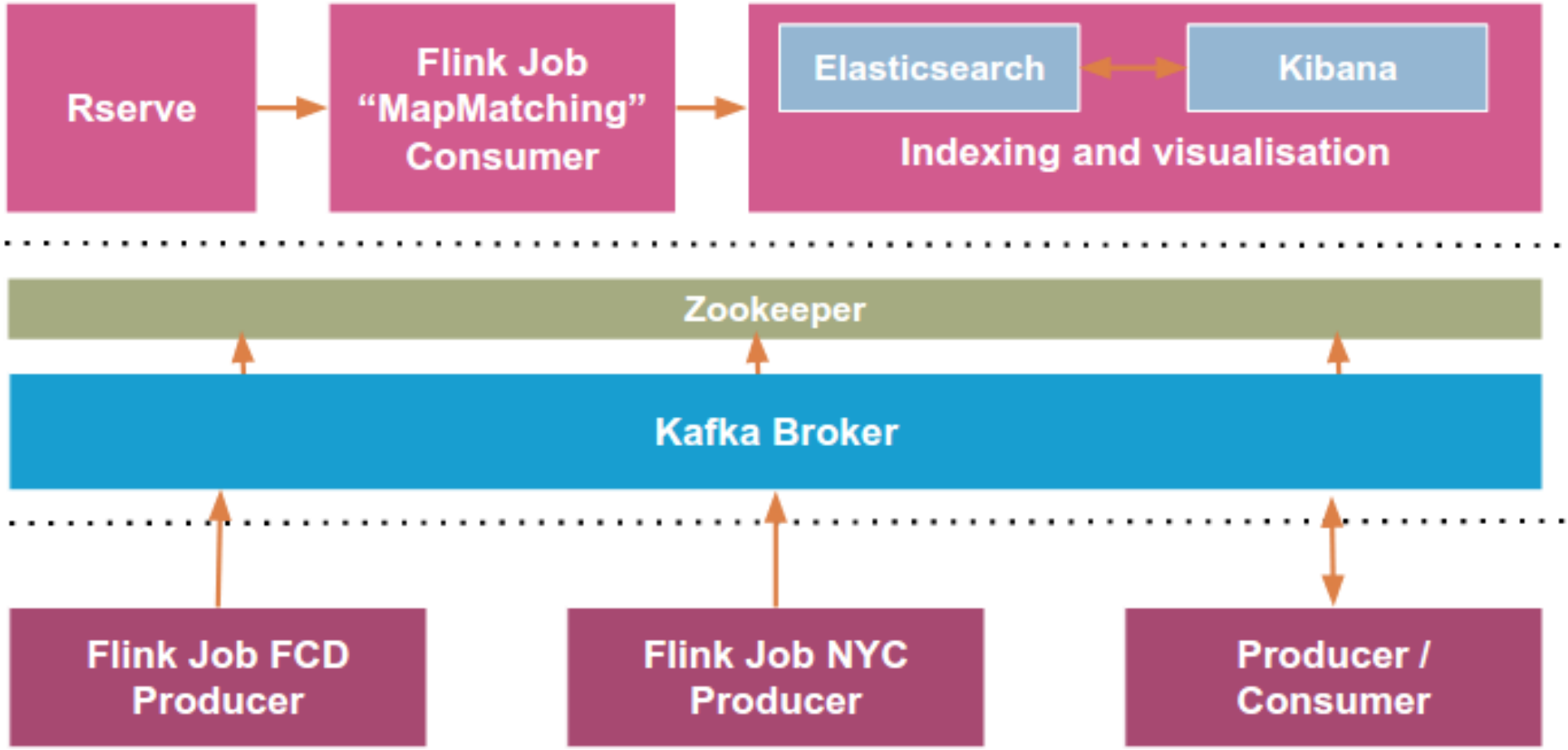
- SC4: transport domain
- SC6: social sciences domain
- SC7: security domain



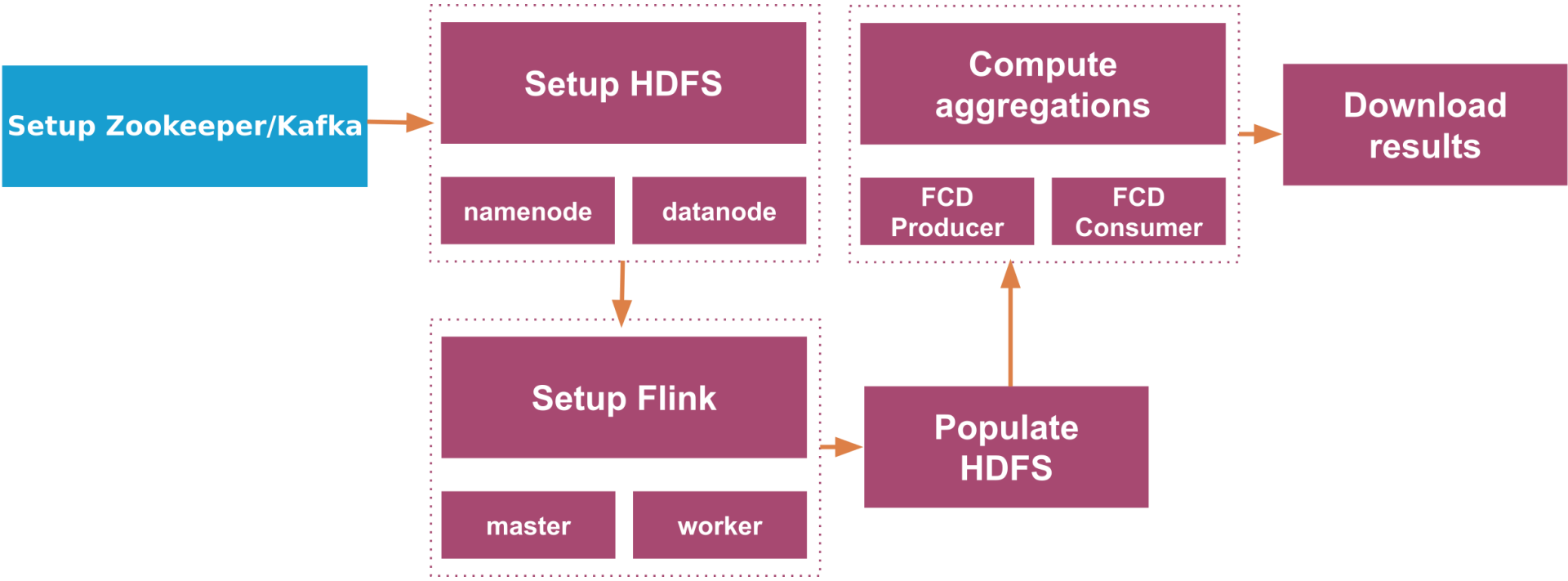
SC4: Transport

- Monitors traffic flow in Thessaloniki, Greece
- Relational database, stored procedures and R scripts for map matching
- Traffic monitoring and forecasting
- How to scale?
  - Migrate to BDF

SC4: Architecture



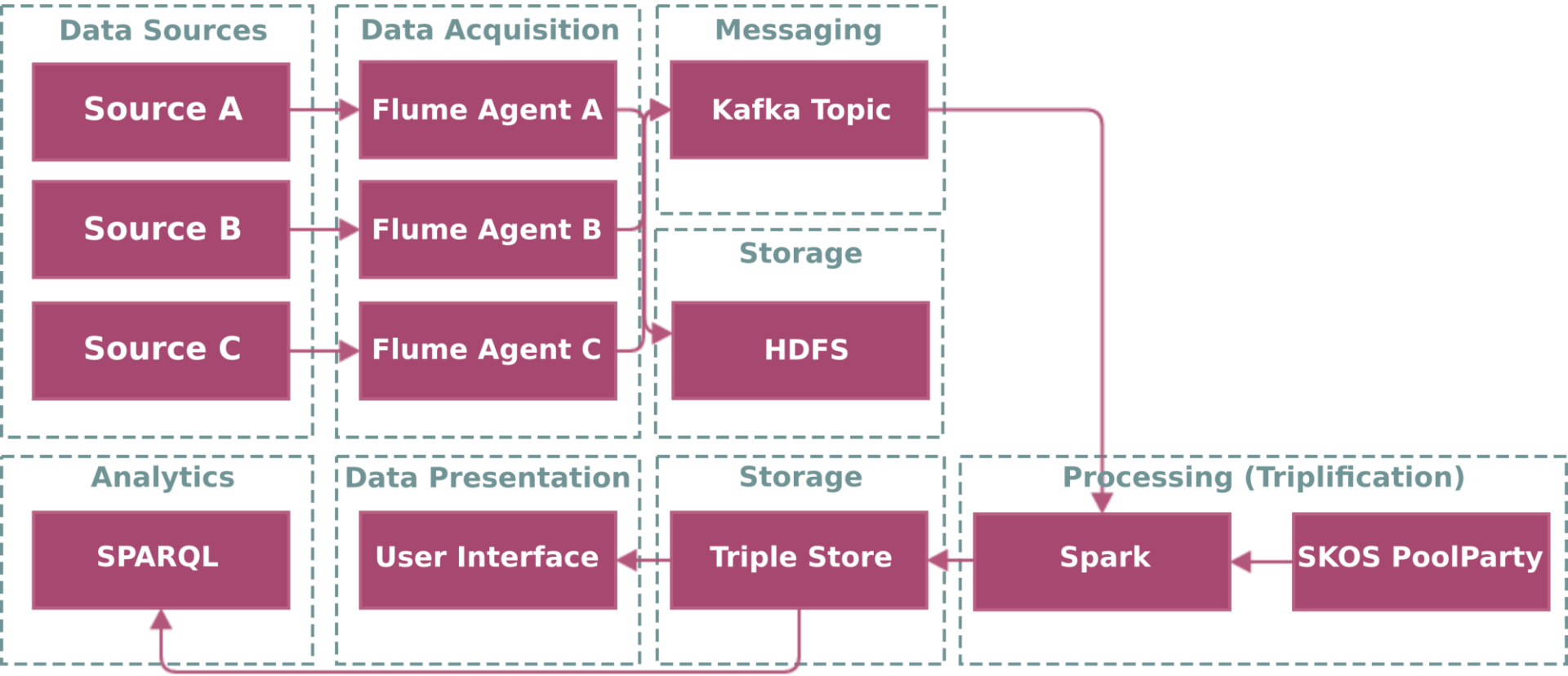
SC4: Initialization Pipeline



SC6: Social Sciences

- Making budget data comparable across EU municipalities
- On example of: Athens, Thessaloniki, Kalamaria
- Uses proprietary docker components
  - PoolParty Semantic Suite
  - PoolParty Graph Search

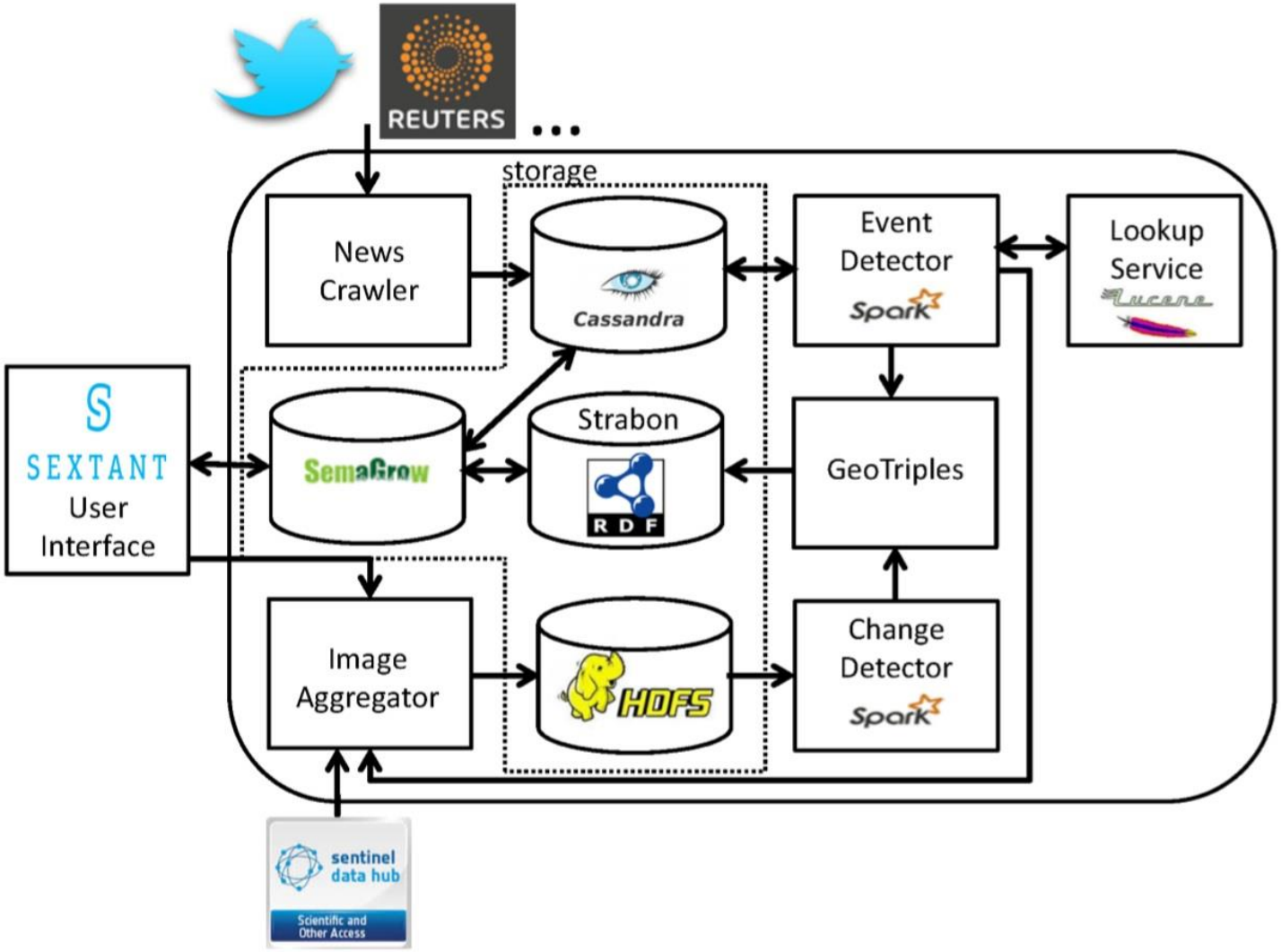
SC6: Architecture



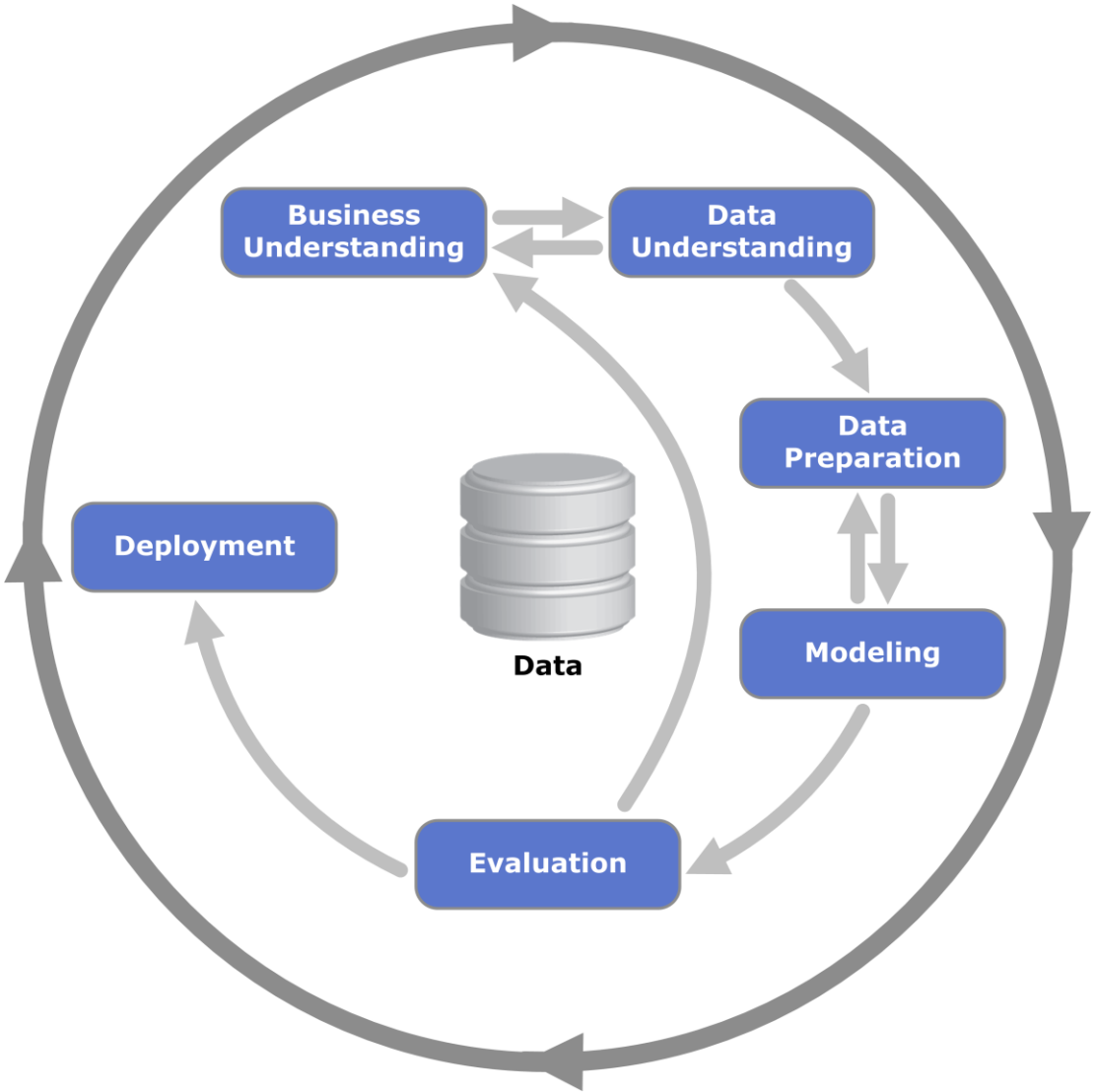
SC7: Security

- Detection of land cover and land use change
- Three workflows
  - Change detection
  - Activation
  - Event detection

SC7: Architecture



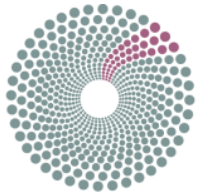
CRISP DM: Where are we?



Developing with BDI

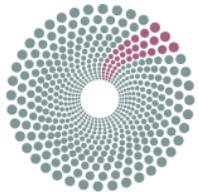
- Docker Hadoop Spark Workbench
  - HDFS
  - YARN
  - HIVE
  - Spark
  - Hue
  - Spark-notebook (or Zeppelin)

https://github.com/big-data-europe

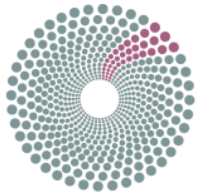


Hadoop Spark Workbench

- Docker based
- Expose ports to dockerhost for development
- Deploy on server for production

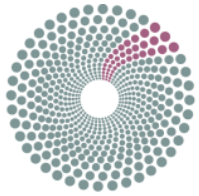


Hadoop Spark Workbench: Demo



Spark Templates

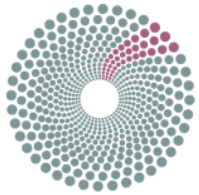
- Create Dockerfile for your application
- Build docker image
- Deploy along the BDI stack
- Make sure that template and Spark versions match



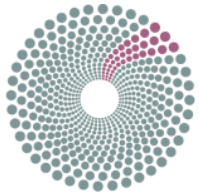
Spark Templates


```
FROM bde2020/spark-java-template:2.2.0-hadoop2.7

ENV SPARK_APPLICATION_JAR_NAME my-app-1.0-SNAPSHOT-with-dependencies
ENV SPARK_APPLICATION_MAIN_CLASS eu.bde.my.Application
ENV SPARK_APPLICATION_ARGS "foo bar baz"
```



Spark Template: Demo



Thank you

Questions?

Github: https://github.com/earthquakesan
@AKSW: http://aksw.org/IvanErmilov.html
Email: iermilov@informatik.uni-leipzig.de
Twitter: @earthquakesan
LinkedIn: https://www.linkedin.com/in/iermilov/