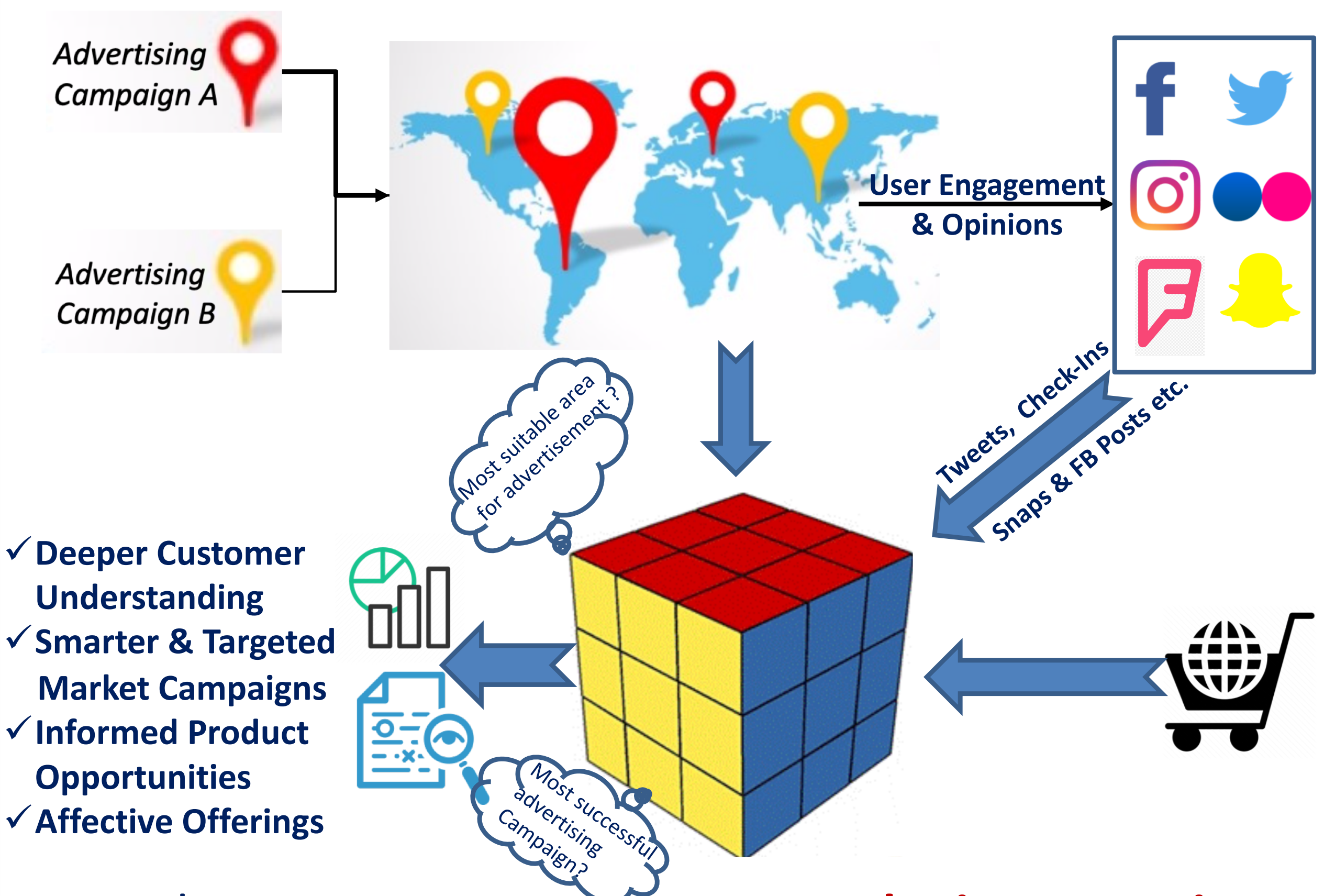
Ninth European Big Data Management & Analytics Summer School (eBISS 2019)
June 30 - July 5, 2019 Berlin, Germany

Spatio-Textual Analytics

Mohsin Iqbal[1,2], Torben Bach Pedersen[1], Esteban Zimányi [2], Matteo Lissandrini[1]
[1]Aalborg University, Aalborg
[2] Université Libre de Bruxelles, Brussels

Background & Motivation


Advertising Campaign A
Advertising Campaign B
User Engagement & Opinions
Tweets, Check-Ins
Snaps & FB Posts etc.
Most suitable area for advertisement ?
Most successful advertising Campaign?
✓Deeper Customer Understanding
✓Smarter & Targeted Market Campaigns
✓Informed Product Opportunities
✓Affective Offerings

Analyze **customer response** to **marketing campaign** and compare **effect on sales** in different regions

Limitations of existing **Spatio-Textual Analytics:**

- Lacks **formalization** and **definition** of **spatio-textual cube**
- No support for **OLAP** over spatio-textual data
- No **framework** for **exploring regions interactively** for major **activities**, **events** and discussion **topic**

Hypothesis

Analyzing structured and **unstructured** data **together** produces **better**, **elaborated** and **more powerful insights.**

Objectives

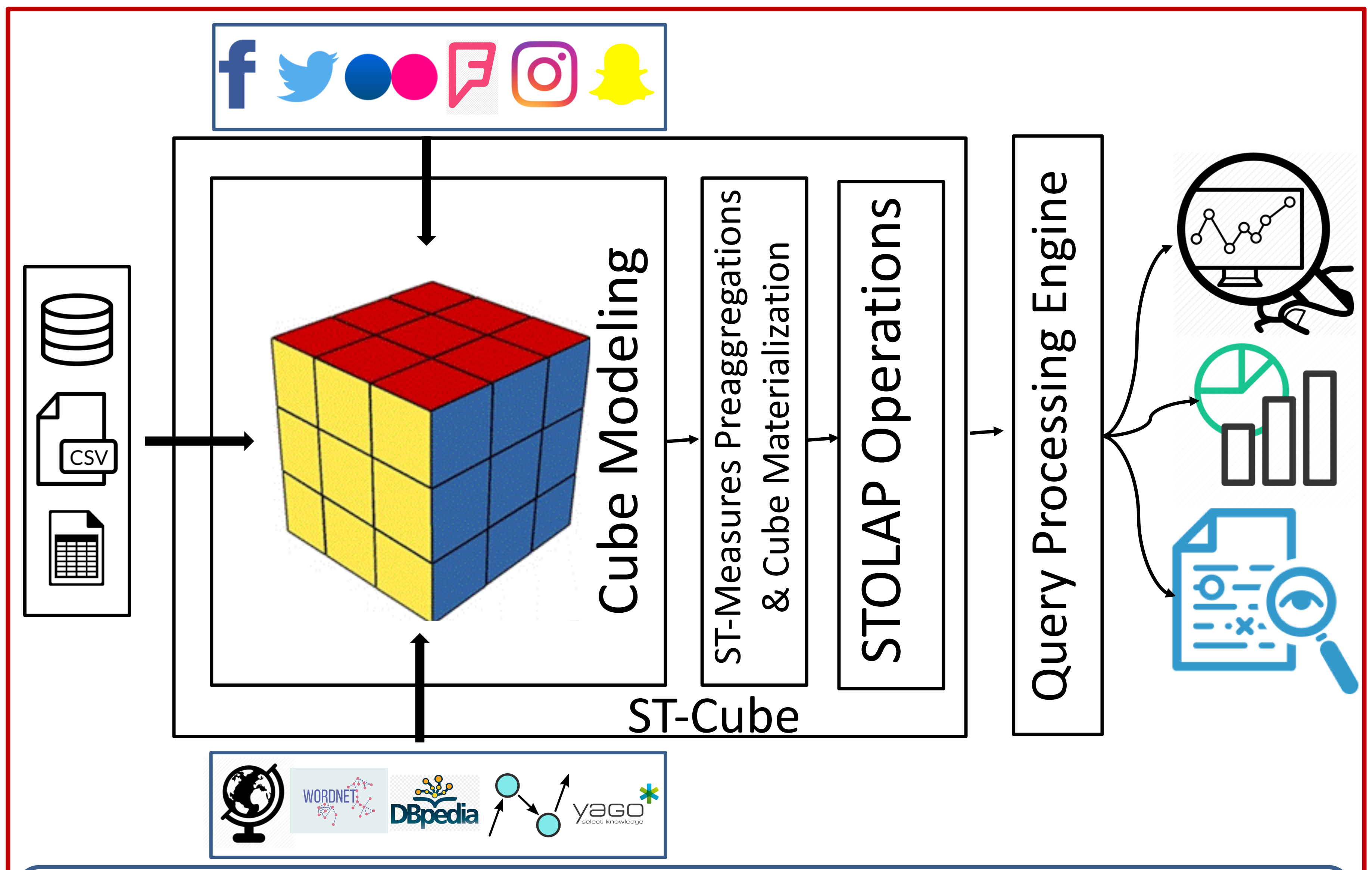
Defining and Formalizing a **Spatio-Textual Cube**

- ✓ Defining dimensions and hierarchies for **spatial** and **textual** data
- ✓ Spatio-Textual **Measures**
- ✓ Spatial-Textual **OLAP** Operations

Building a **Region Exploration Framework**

- ✓ Supports analysis of spatio-textual data **combined** with traditional data
- ✓ Links **external resources**
- ✓ Allows to **compare** & find **similar regions**
- ✓ Mechanism for **ranking** similar regions
- ✓ Functionality like **magnifying glass**

Spatio-Textual Cube



Spatio-Textual Object **Obj$_{st}$** is formally defined as **Obj$_{st}$ = <τ, λ, ϕ>** where **τ**, **λ** and **ϕ** represents the **timestamp**, **location** and **textual** components respectively**.**

An ***n-dimensional*** **ST-Cube schema** CS$_{stc}$ is a tuple **CS$_{stc}$ = (D, M, F)**, with a set of dimensions D = {**d$_{time}$**, **d$_{location}$**, **d$_{text}$**, **d$_4$**, . . . , **d$_n$**}, a set of measures **M = {m$_1$, m$_2$, m$_3$, . . . , m$_k$}**, and a **fact type F.**

Spatio-Textual **Dimensions**

- ✓ **Spatial** Dimension
  - ✓ **Grid** Based Hierarchy
  - ✓ **Semantic** Based Hierarchy
- ✓ **Textual** Dimension
  - ✓ **Individual** Based Hierarchy
  - ✓ **Importance** Based Hierarchy
  - ✓ **Majority** Based Hierarchy
  - ✓ **Custom** Hierarchy

Spatio-Textual **Dimension Hierarchies**

- ✓ **τ → Day → Month → Quarter → Year**
- ✓ **τ → Second → Minute → Hour**
- ✓ **λ → City → Region → Country**
- ✓ **ϕ → Term → Theme → Topic → Concept**

→ **T** ***(Top)***

Spatio-Textual **Measures**

- ✓ Top *k* **keywords** discussed within a **geographical region**
- ✓ **Keywords density** in a region
- ✓ **Most frequent keywords** in a area defined by a **polygon**

Spatio-Textual Operations & Experiments

- ✓ **Aggregation** Operations **– combines** more than one spatio-textual objects →**O$_{agg}$(obj$^{st}_1$, obj$^{st}_2$, ..., obj$^{st}_n$) → obj$^{st\prime}$**, e.g., Top-*K* **Topics, Union** of regions (e.g., polygons) and Most **frequent KW** in a **region**
- ✓ **Comparison** Operations – compares two or more spatio-textual objects for **relevance → O$_{com}$(obj$^{st}_1$, obj$^{st}_2$, ..., obj$^{st}_n$) → *true/false***
- ✓ **Numeric** Operations – takes more than one spatio-textual objects and return a **real value O$_n$(obj$^{st}_1$,..., obj$^{st}_n$) → $\mathbb{R}$**
- ✓ **Extension** of OnLine Analytical Processing (OLAP) operation to **spatio-textual OLAP (STOLAP)** using spatio-textual operations

Performance Optimization

- ✓ **Preaggregation** and **materialization** of spatio-textual **measure** and **cube** (space-time trade off) for efficient analysis
- ✓ **Comparison** of proposed preaggregation & **partial**-materialization technique with baseline **fully**-materialized and **no**-materialization
- ✓ **Experimental Evaluation** using real-world **twitter** dataset **(8.5M)**
- ✓ **ST-Cube modeling** using **snowflake** schema in **MSSQL** Server
- ✓ **Spatial** dimension implementation using the **Geo Names dataset**[1]
- ✓ **Textual** dimension implementation using **WordNet**[2] knowledge source

[2]https://wordnet.princeton.edu [1] http://download.geonames.org/export/dump/

Future Work - Region Exploration

- ✓ **Exploring** a **geographical region** and finding **similar** regions
- ✓ A **magnifying glass** like system to figure out the general **trends**, **events** and popular discussion **topic** of an interested **region**
- ✓ **Parameters selection** to capture any regions characteristics
- ✓ **Comparison** of different regions (e.g., future business opportunities)
- ✓ **Ranking** function for similar regions queries
- ✓ **Evaluation** of proposed techniques using **real-world** data and use cases