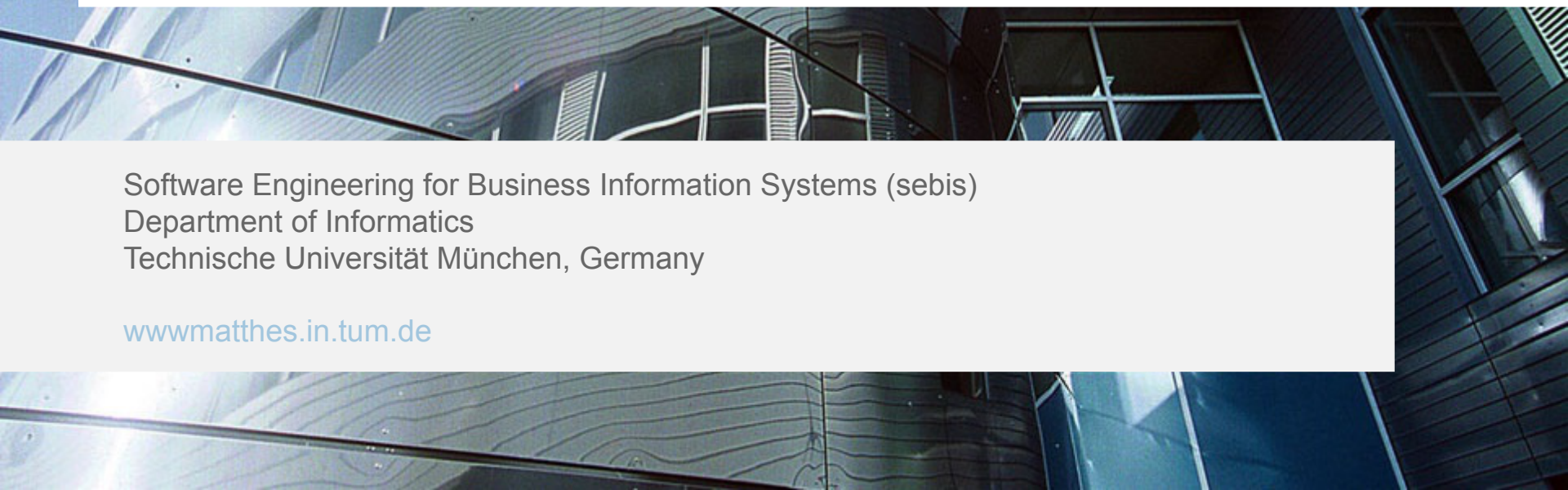
An Empirical Study in the Quality of Algorithms for Dimensionality Reduction and Visualization of High-Dimensional Data

Bachelor's Thesis Kick-Off Presentation

Lyubomir Stoykov 18.01.2015

Software Engineering for Business Information Systems (sebis)
Department of Informatics
Technische Universität München, Germany

wwwmatthes.in.tum.de



Introduction

Visualisation of High-Dimensional Data & Dimensionality Reduction

Most features in high-dimensional data are redundant

Curse of dimensionality

Machine learning, big data, artificial intelligence, biotechnology

Finding patterns in complexity

Visual analysis where analysis of raw data is beyond capabilities of human beings

Software analysis where analysis of raw data beyond capabilities of current technologies

Exponentially faster data processing with less resources

Research Questions

Which are the algorithms for dimensionality reduction and visualisation of high dimensional data?

What are their general properties and use cases?

In what extent do they reveal information about the underlying data?

Related Work

Big Data & Machine learning

No assesment, use „as it is“.

Overview of Data Visualisation Techniques

High-dimensional data often neglected

Overview of Dimensionality Reduction & Comparison

Not exhaustive

Outdated

Mainly comparison between the algorithms

Assesment of quality based on author's observation

Contribution

sebis

Fills a research gap for a very much needed current and comprehensive overview of algorithms for dimensionality reduction and visualisation of high-dimensional data, their general properties and use cases.

Fills a research gap for a more objective, tangible, author-detached evaluation of the quality of suchlike techniques.

Provides a not existing until now tool for the enduser without technical knowledge to apply state of the art dimensionality reduction and data visualisation algorithms on his own data sets.

Evaluation

sebis

Mathematical Approach

Train data sets on low dimensional, compressed representation. Measure error.

Calculate distances between related points in low dimensional representation.

Problems?

Ill-posed problem. Intrinsic dimensionality unknown. Mathematical approach doesn't quite suffice to adequately evaluate the techniques.

Heuristic, but tangible approach

Ask people concrete questions about properties of high dimensional data after dimensionality reduction & data visualisation tools have been applied.

Evaluation Tool

Choose between state of the art dimensionality reduction & data visualisation techniques

Apply dimensionality reduction & data visualisation techniques on default data sets

Apply dimensionality reduction & data visualisation techniques on own data sets

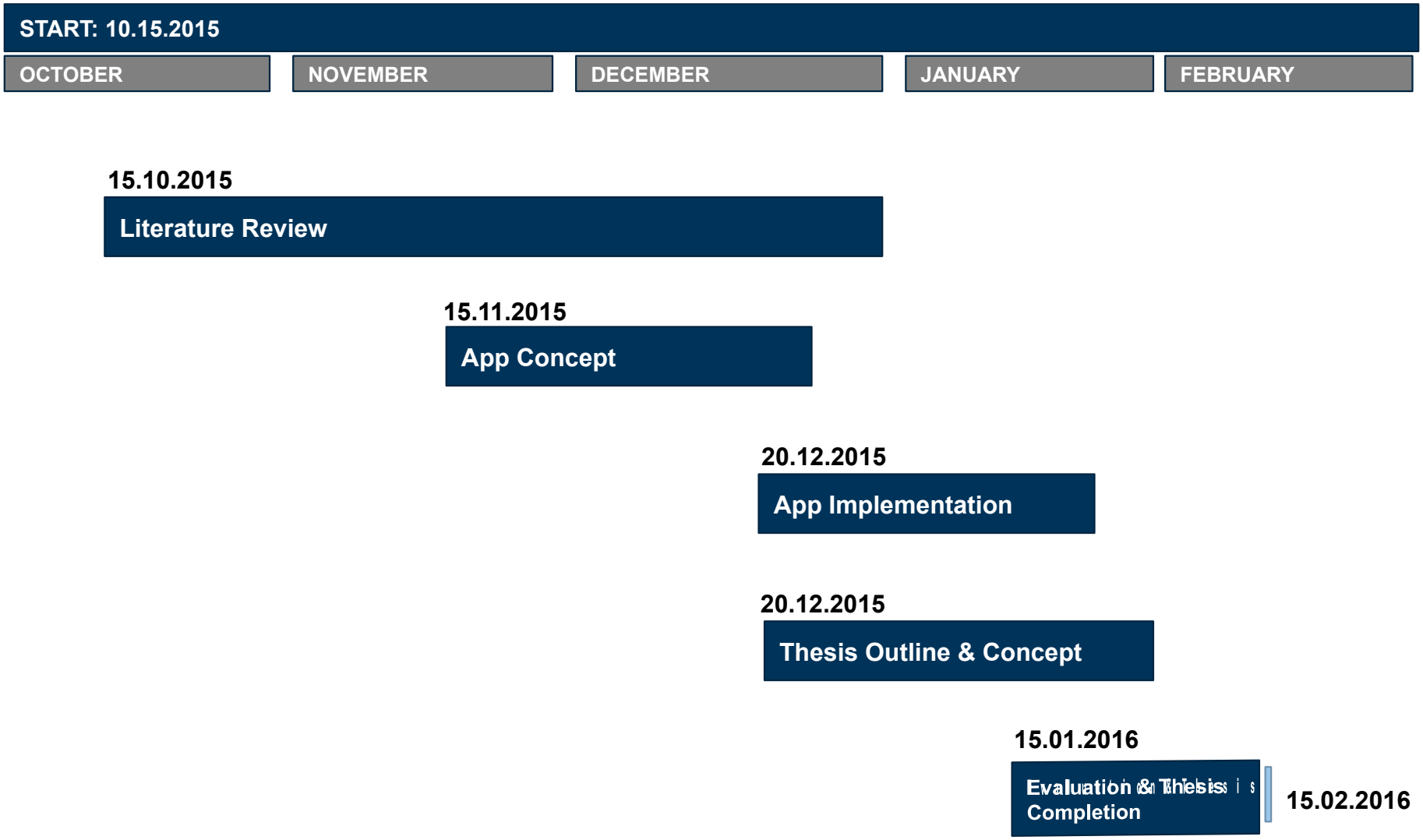
Change settings of dimensionality reduction & data visualisation techniques

Analyse non-metric data sets with user-defined mapping

Web Application for Browser & Desktop App for Mac, Linux & Windows

Demo

Timeline


START: 10.15.2015
OCTOBER
NOVEMBER
DECEMBER
JANUARY
FEBRUARY
15.10.2015
Literature Review
15.11.2015
App Concept
20.12.2015
App Implementation
20.12.2015
Thesis Outline & Concept
15.01.2016
Evaluation & Thesis Completion
15.02.2016

Thank you