

# Algorithms for Automatic Tagging of Medieval Manuscripts

Swati Chandna

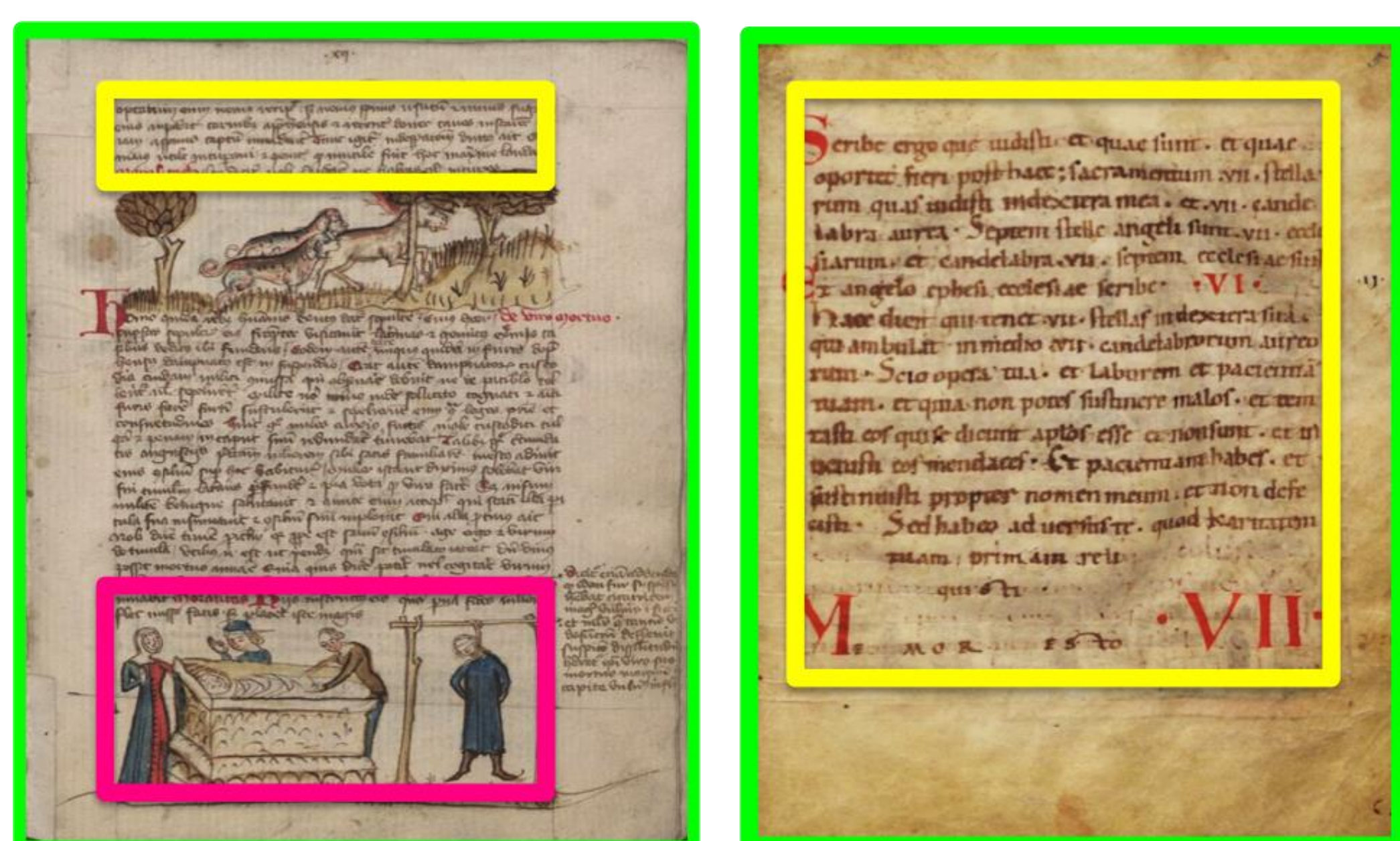


- Joint research project of Darmstadt, Trier and Karlsruhe
- Funded by the Federal Ministry of Research and Education (BMBF)
- In cooperation with DARIAH-DE (Digital Research Infrastructure for the Arts and Humanities)



## Aims

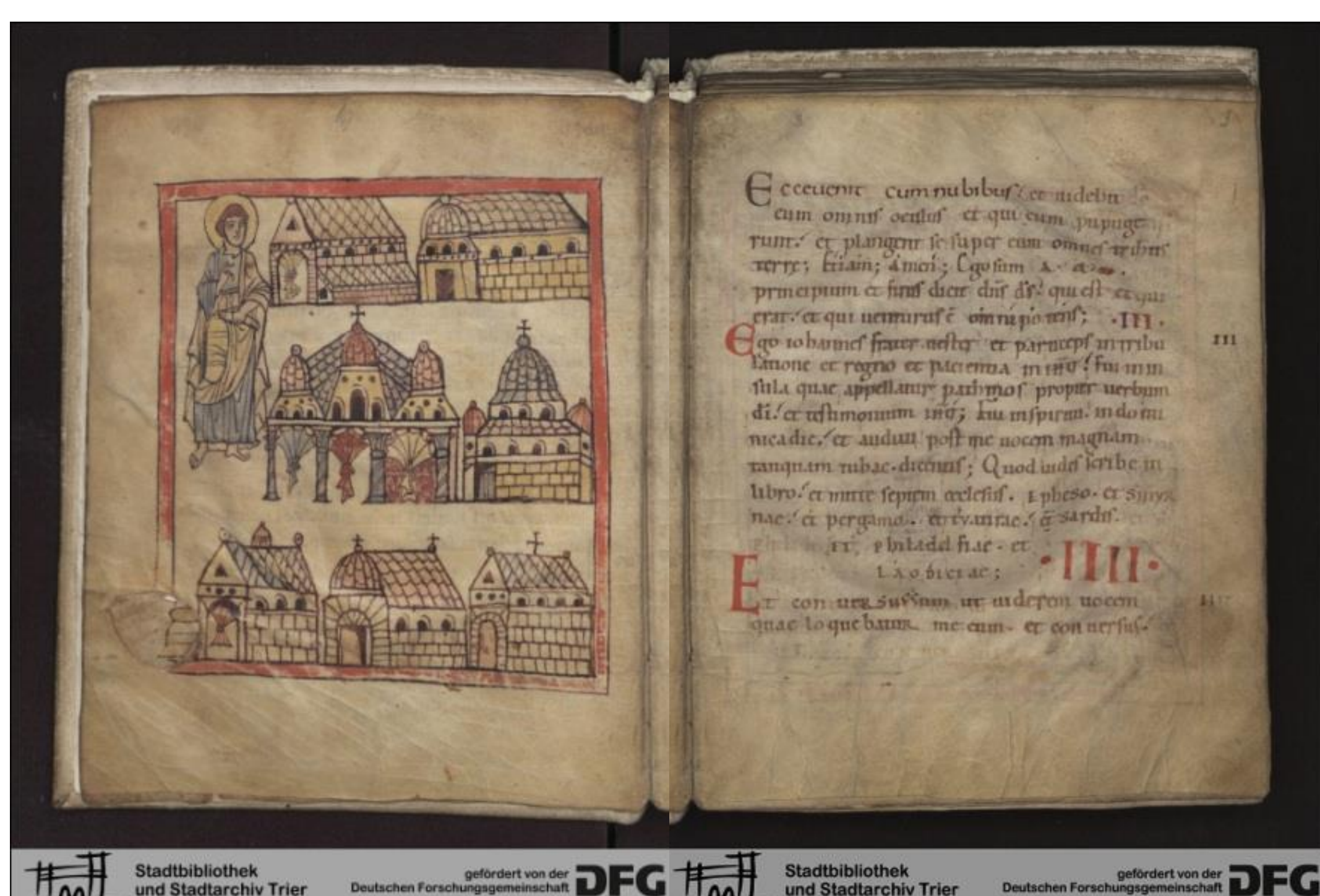
- **Automatic identification** of macro and micro structural *layout elements*, e.g. number of lines
- **Statistical analysis** of objectified, reproducible and at micro level differentiated features
- **Visual analysis** of hidden relationships between many groups of codices



□ page size, □ written space, □ pictorial space

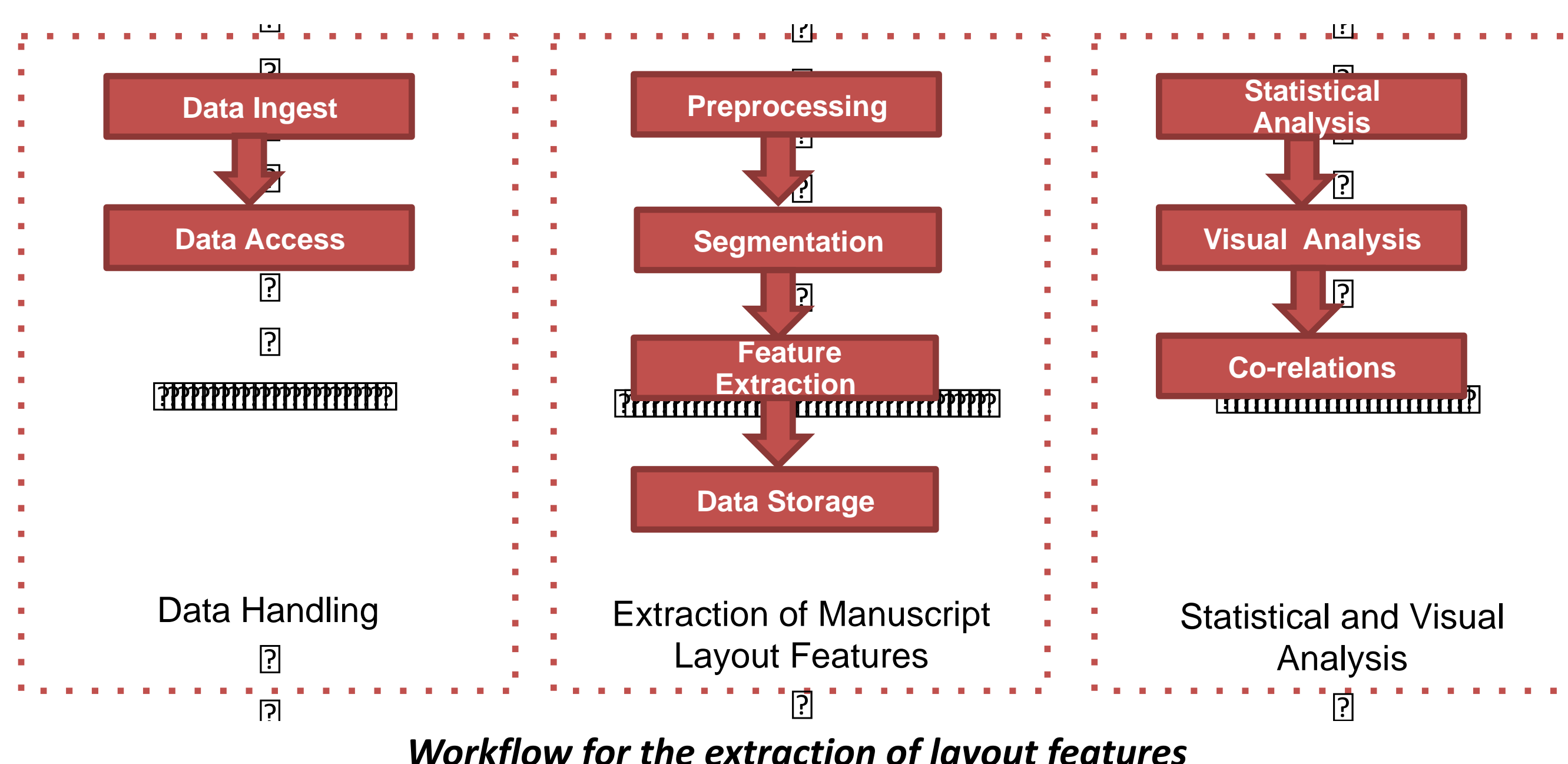
## Virtual Scriptorium St. Matthias

- Digital reconstruction of the library of the Benedictine abbey St. Matthias, Trier
- 440 codices (8<sup>th</sup> – 16<sup>th</sup> century)
- 170,000 digitized codex pages
- 1,000,000 files in various formats (~5 TB)
- Online available: <http://stmatthias.uni-trier.de>

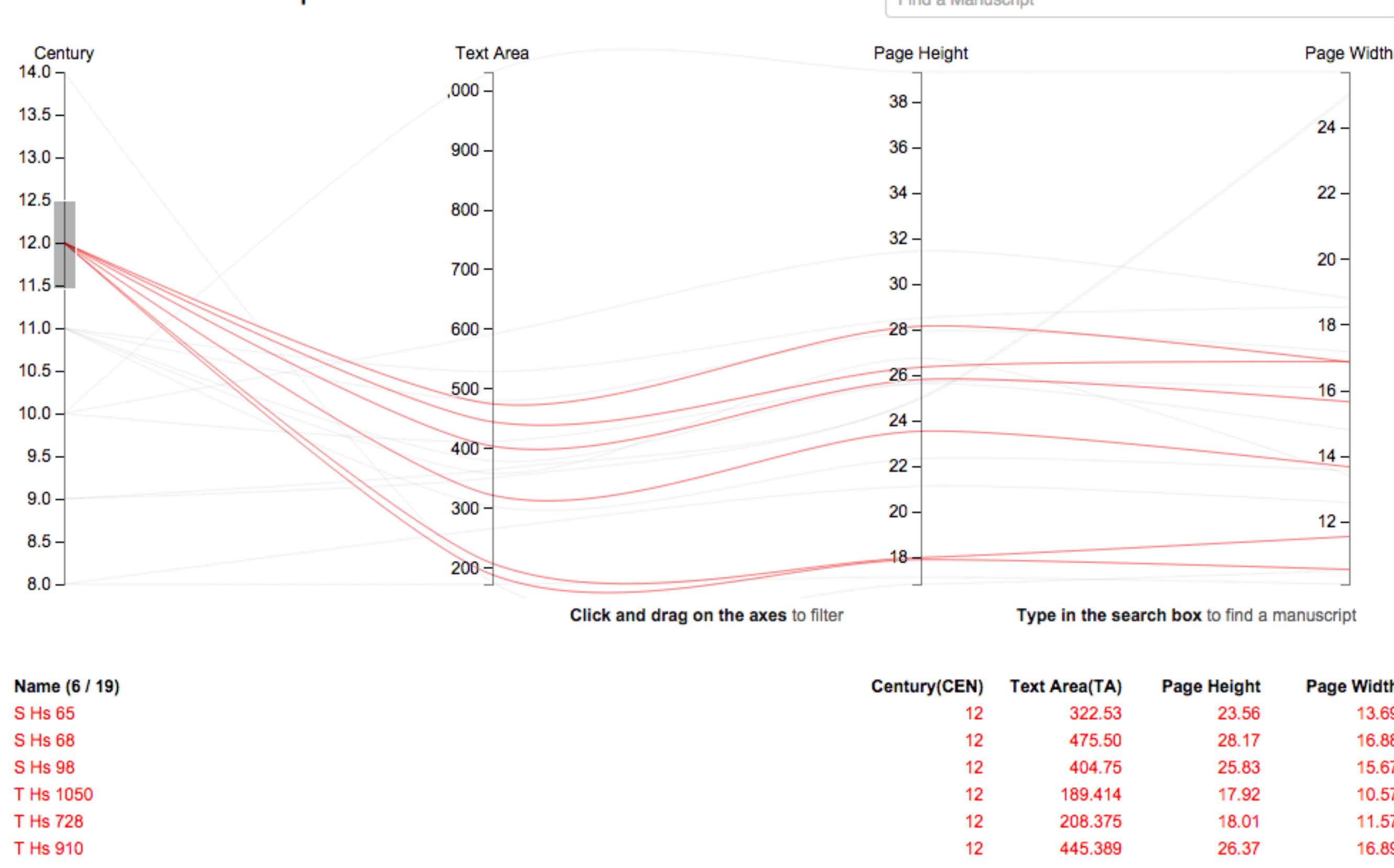


## Data Processing and Visualization

- **Data Handling:** data is managed by repository using KIT Data Manager services
- **Preprocessing** of the manuscript images: color calibration, spatial calibration and noise removal
- **Object Segmentation:** pages, images, text areas
- **Feature Extraction:** e.g. page size, pictorial space, written space
- **Metadata:** results are extended in XML schema according to TEI P5 in DARIAH Annotation Infrastructure
- **Visual Analysis:** gain better insights into multi dimensional datasets of digitized medieval manuscripts



## Medieval Manuscripts Visualization



An example of interactive visual analysis with highlighted dataset