Healthcare Decision-Making over a Geographic, Socioeconomic, and Image Data Warehouse

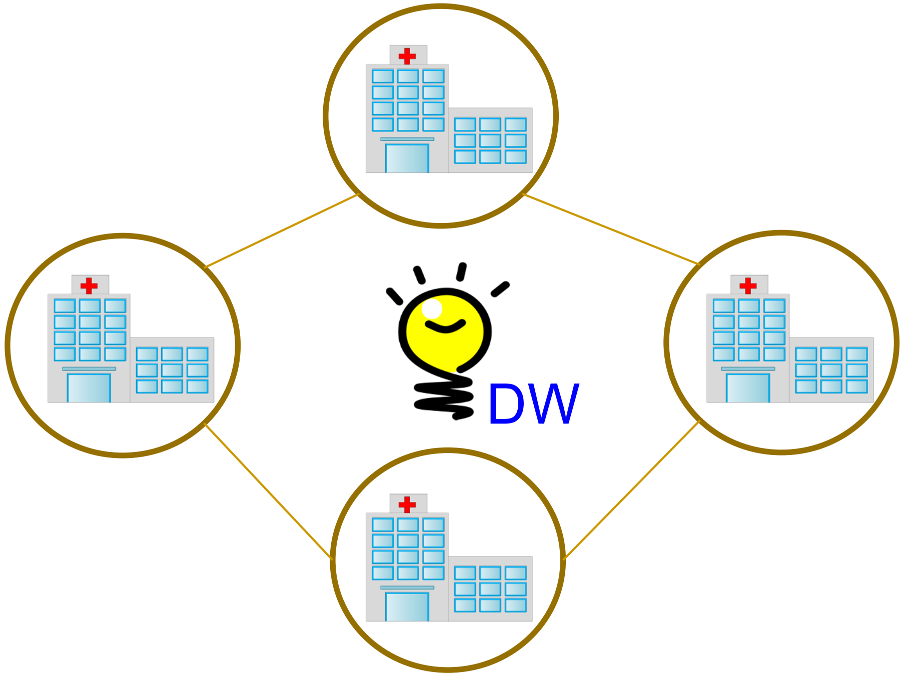
Guilherme M. Rocha
Piero L. Capelo
Cristina D. A. Ciferri

Outline

- Motivation
- Contributions
  - Designs of Star Schema
  - The SimSparkOLAP Method
- Experimental Evaluation
- Conclusions and Future Work

Motivation

- Huge volume of healthcare data generated by different sources
- Types of data
  - conventional
  - image
  - geographic
  - socioeconomic



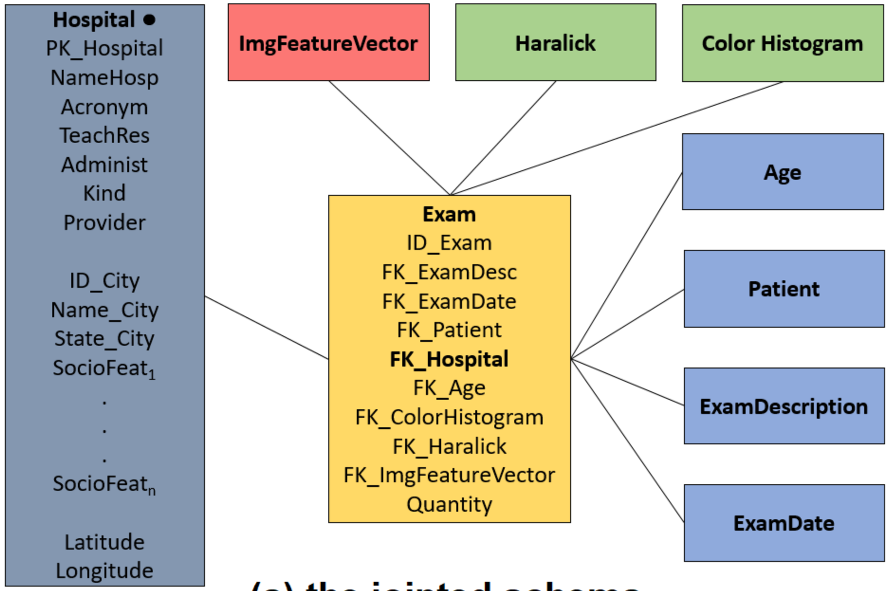
- Data Warehouse (DW)
  - can provide support for the healthcare decision-making

Our Work

- Execution of analytical queries over geographic, socioeconomic, and image DWs
- Contributions
  - three designs of star schema for the extended DW
  - a Spark method, called SimSparkOLAP, to efficiently process extended analytical queries
  - investigation of the method's performance over the proposed star schemas
  - study of semantic analytical queries and their importance to the healthcare decision-making

Outline

- Motivation
- Contributions
  - Designs of Star Schema
  - The SimSparkOLAP Method
- Experimental Evaluation
- Conclusions and Future Work



Hospital ●
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider

ID_City
Name_City
State_City
SocioFeat$_1$
.
.
.
SocioFeat$_n$

Latitude
Longitude

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVector
Quantity

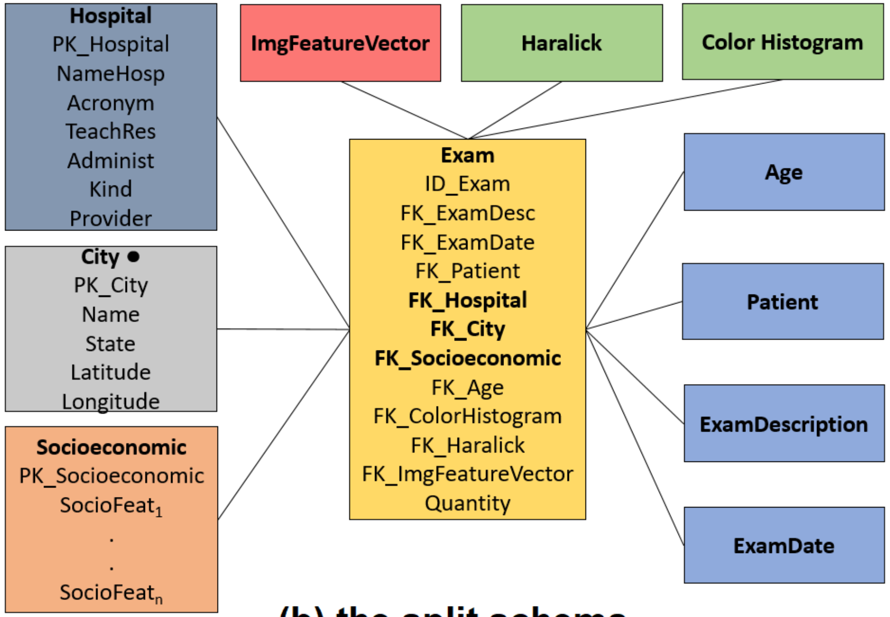
Age

Patient

ExamDescription

ExamDate

(a) the jointed schema



Hospital
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider

City ●
PK_City
Name
State
Latitude
Longitude

Socioeconomic
PK_Socioeconomic
SocioFeat$_1$
.
.
SocioFeat$_n$

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_City
FK_Socioeconomic
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVector
Quantity

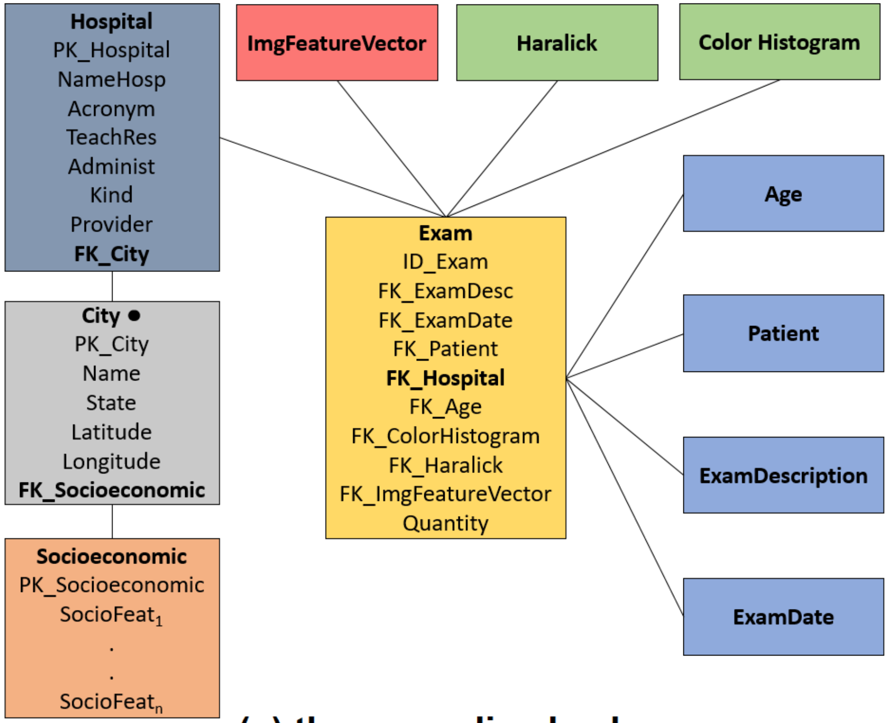
Age

Patient

ExamDescription

ExamDate

(b) the split schema



Hospital
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider
FK_City

City ●
PK_City
Name
State
Latitude
Longitude
FK_Socioeconomic

Socioeconomic
PK_Socioeconomic
SocioFeat$_1$
.
.
SocioFeat$_n$

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVector
Quantity

Age

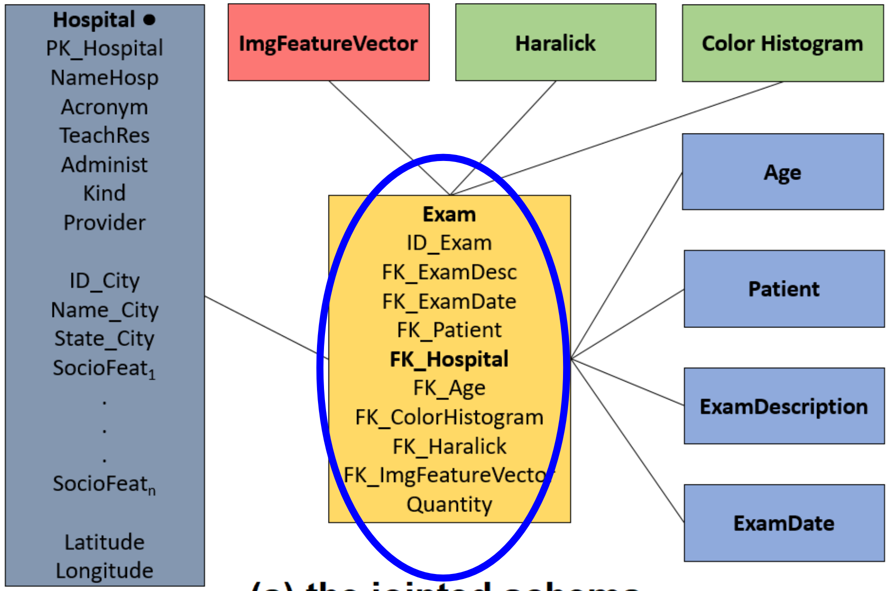
Patient

ExamDescription

ExamDate

(c) the normalized schema

fact table



Hospital ●
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider

ID_City
Name_City
State_City
SocioFeat$_1$
.
.
.
SocioFeat$_n$

Latitude
Longitude

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVecto
Quantity

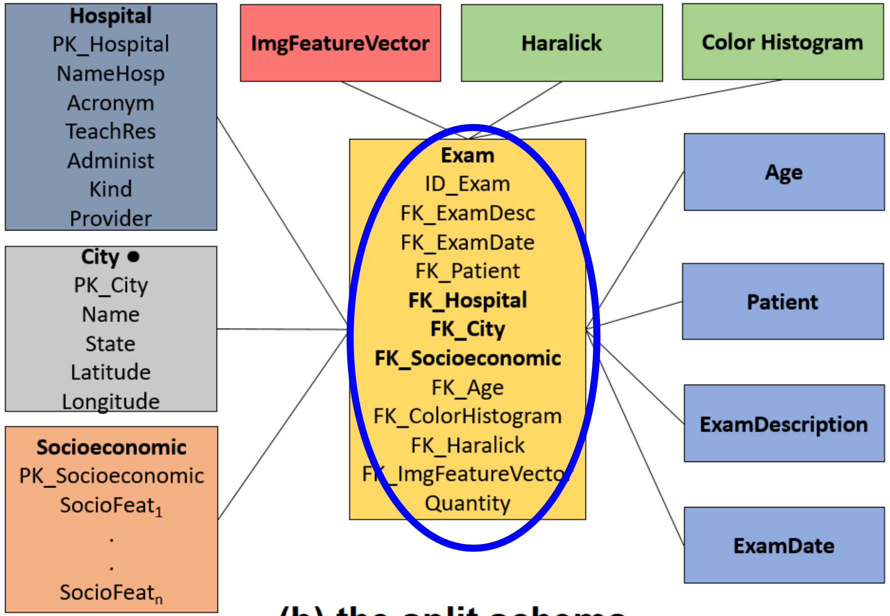
Age

Patient

ExamDescription

ExamDate

(a) the jointed schema



Hospital
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider

City ●
PK_City
Name
State
Latitude
Longitude

Socioeconomic
PK_Socioeconomic
SocioFeat$_1$
.
.
SocioFeat$_n$

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_City
FK_Socioeconomic
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVecto
Quantity

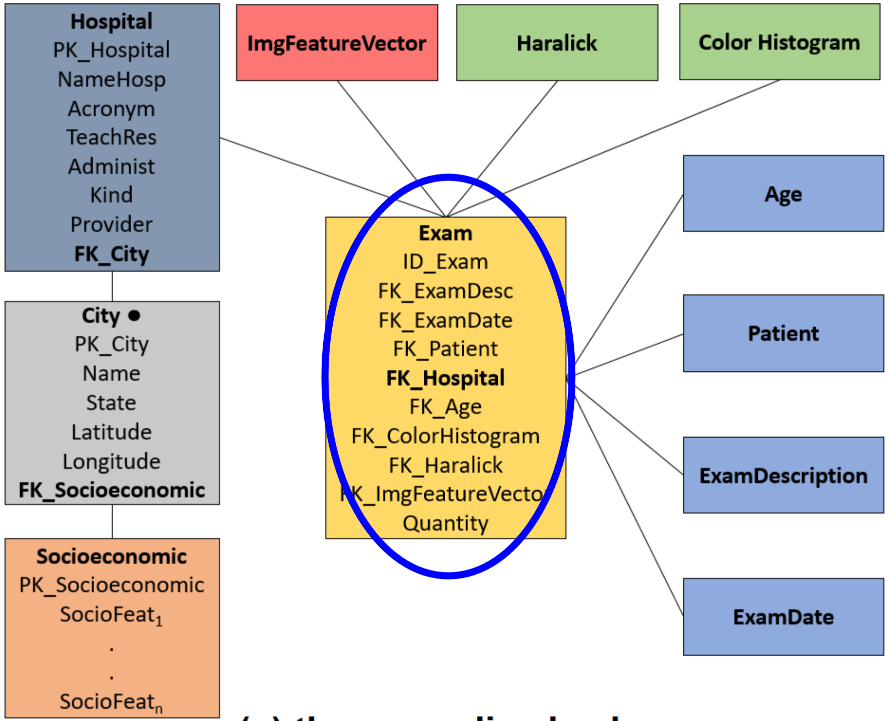
Age

Patient

ExamDescription

ExamDate

(b) the split schema



Hospital
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider
FK_City

City ●
PK_City
Name
State
Latitude
Longitude
FK_Socioeconomic

Socioeconomic
PK_Socioeconomic
SocioFeat$_1$
.
.
SocioFeat$_n$

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVecto
Quantity

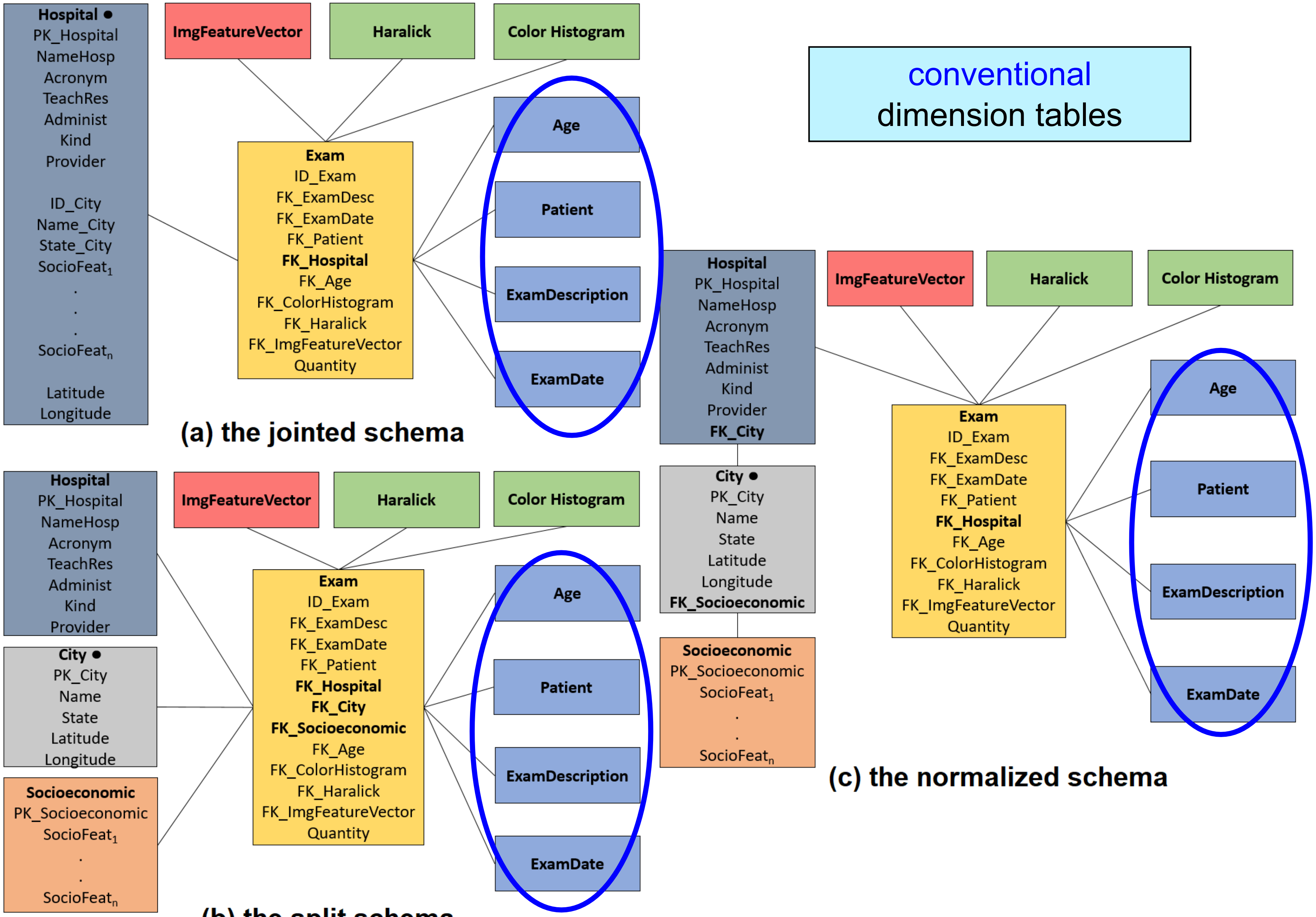
Age

Patient

ExamDescription

ExamDate

(c) the normalized schema



conventional
dimension tables

Hospital ●
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider

ID_City
Name_City
State_City
SocioFeat$_1$
.
.
.
SocioFeat$_n$

Latitude
Longitude

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVector
Quantity

Age

Patient

ExamDescription

ExamDate

(a) the jointed schema

Hospital
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider

City ●
PK_City
Name
State
Latitude
Longitude

Socioeconomic
PK_Socioeconomic
SocioFeat$_1$
.
.
SocioFeat$_n$

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_City
FK_Socioeconomic
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVector
Quantity

Age

Patient

ExamDescription

ExamDate

(b) the split schema

Hospital
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider
FK_City

City ●
PK_City
Name
State
Latitude
Longitude
FK_Socioeconomic

Socioeconomic
PK_Socioeconomic
SocioFeat$_1$
.
.
SocioFeat$_n$

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVector
Quantity

Age

Patient

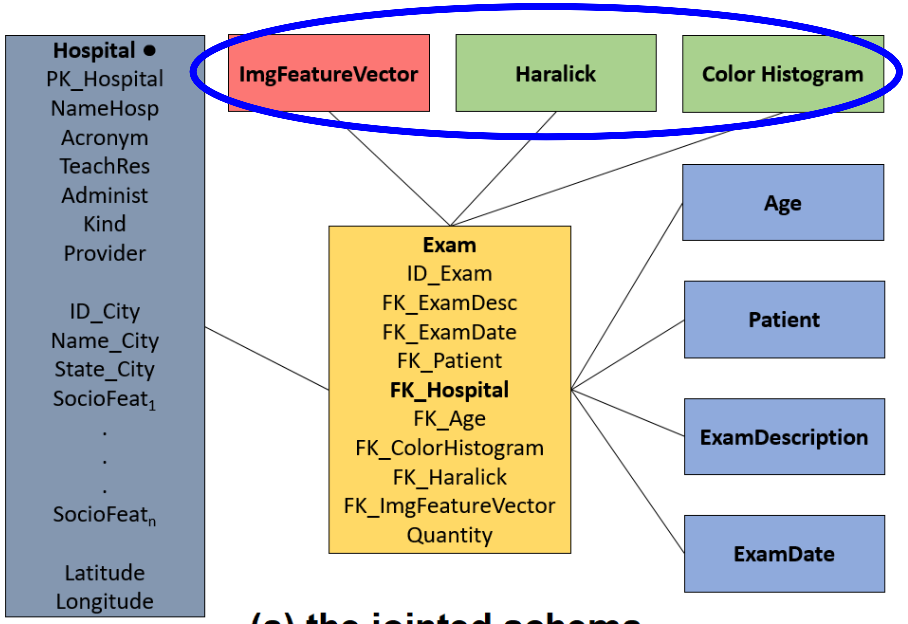
ExamDescription

ExamDate

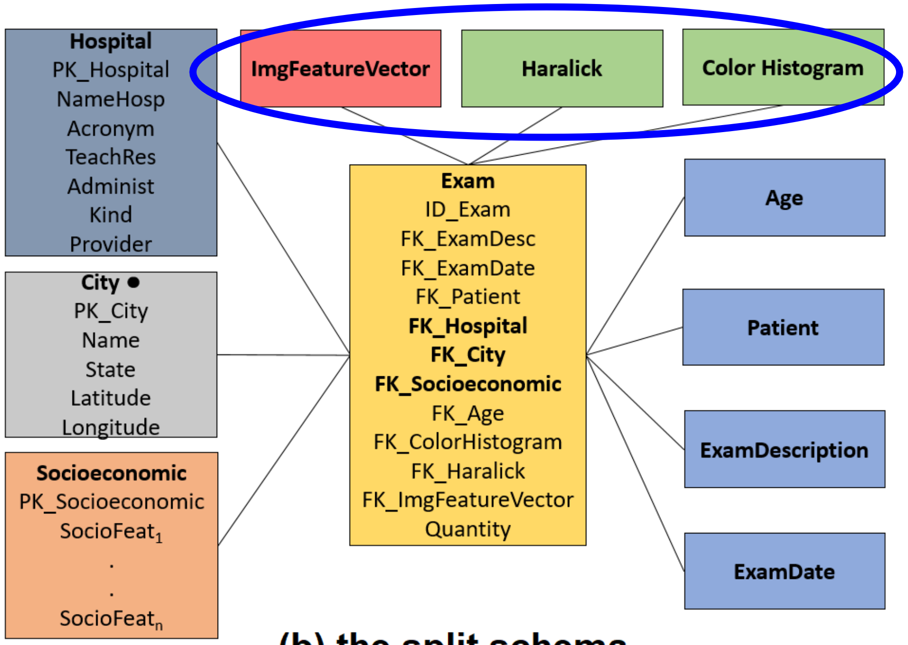
(c) the normalized schema

image similarity factor

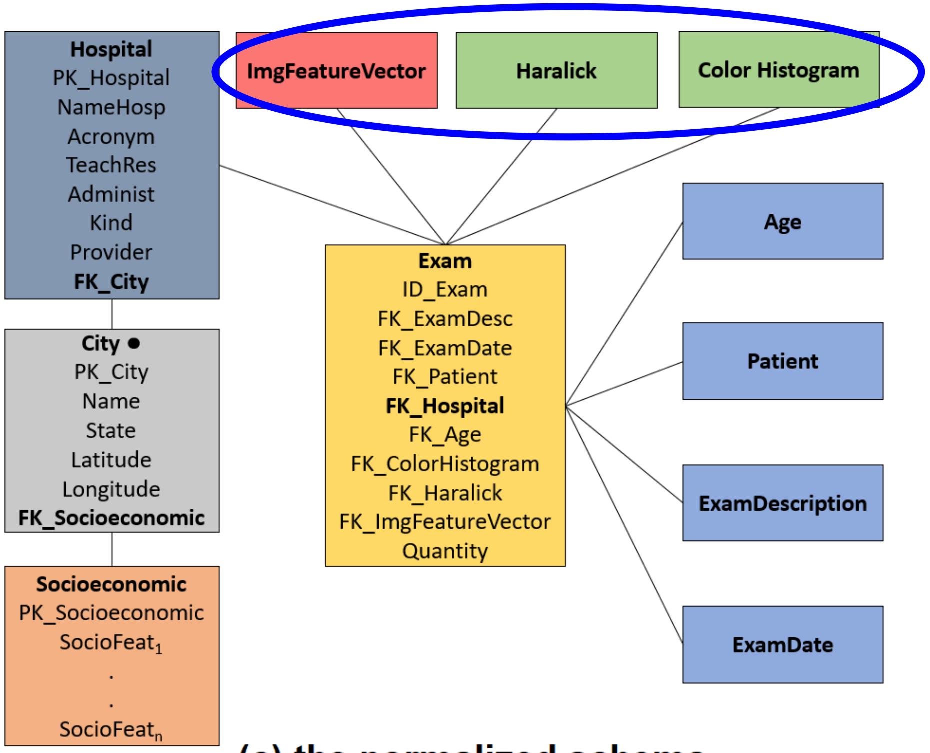
feature vector table
perceptual layer tables



(a) the jointed schema



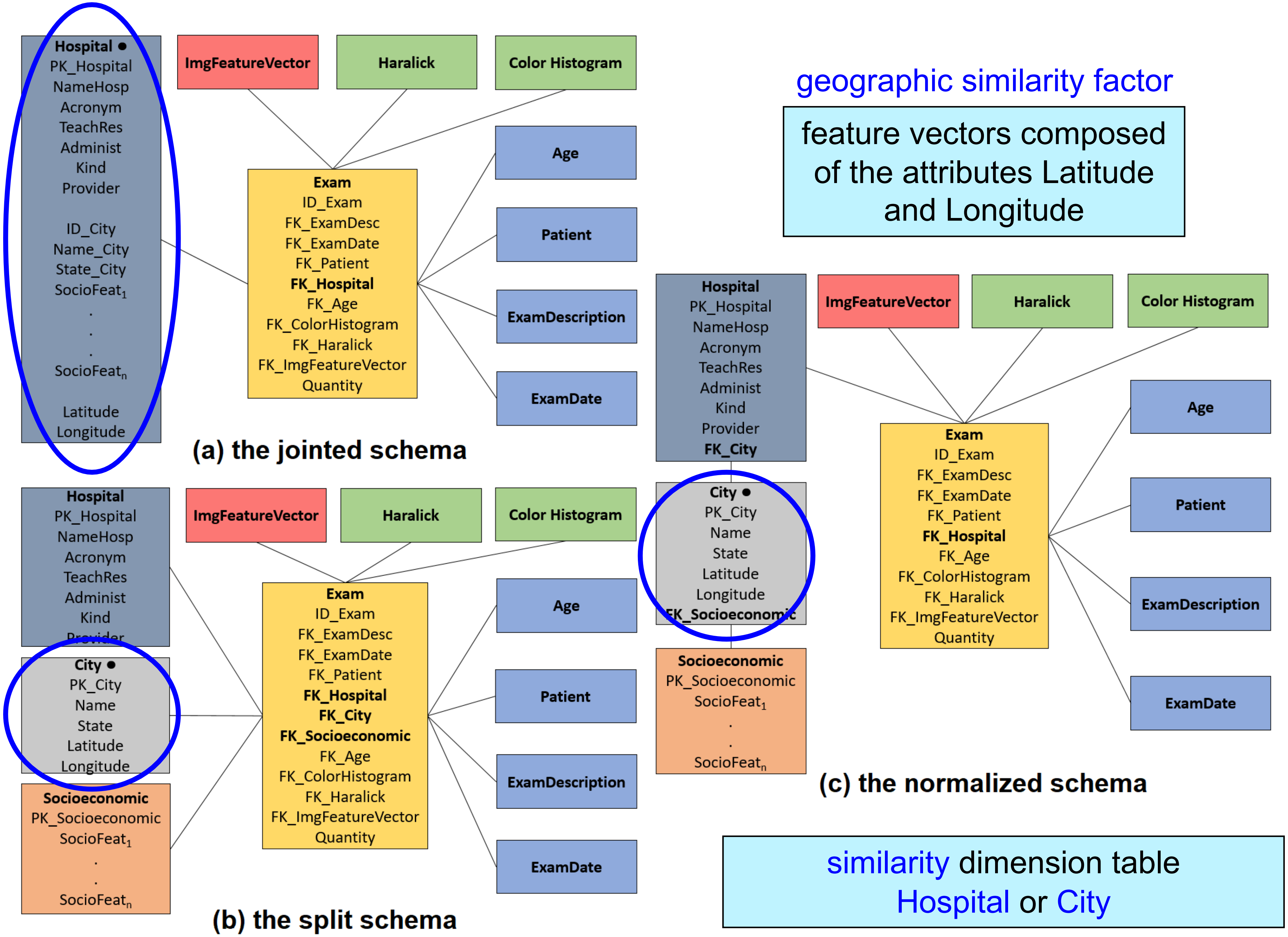
(b) the split schema



(c) the normalized schema

geographic similarity factor

feature vectors composed of the attributes Latitude and Longitude

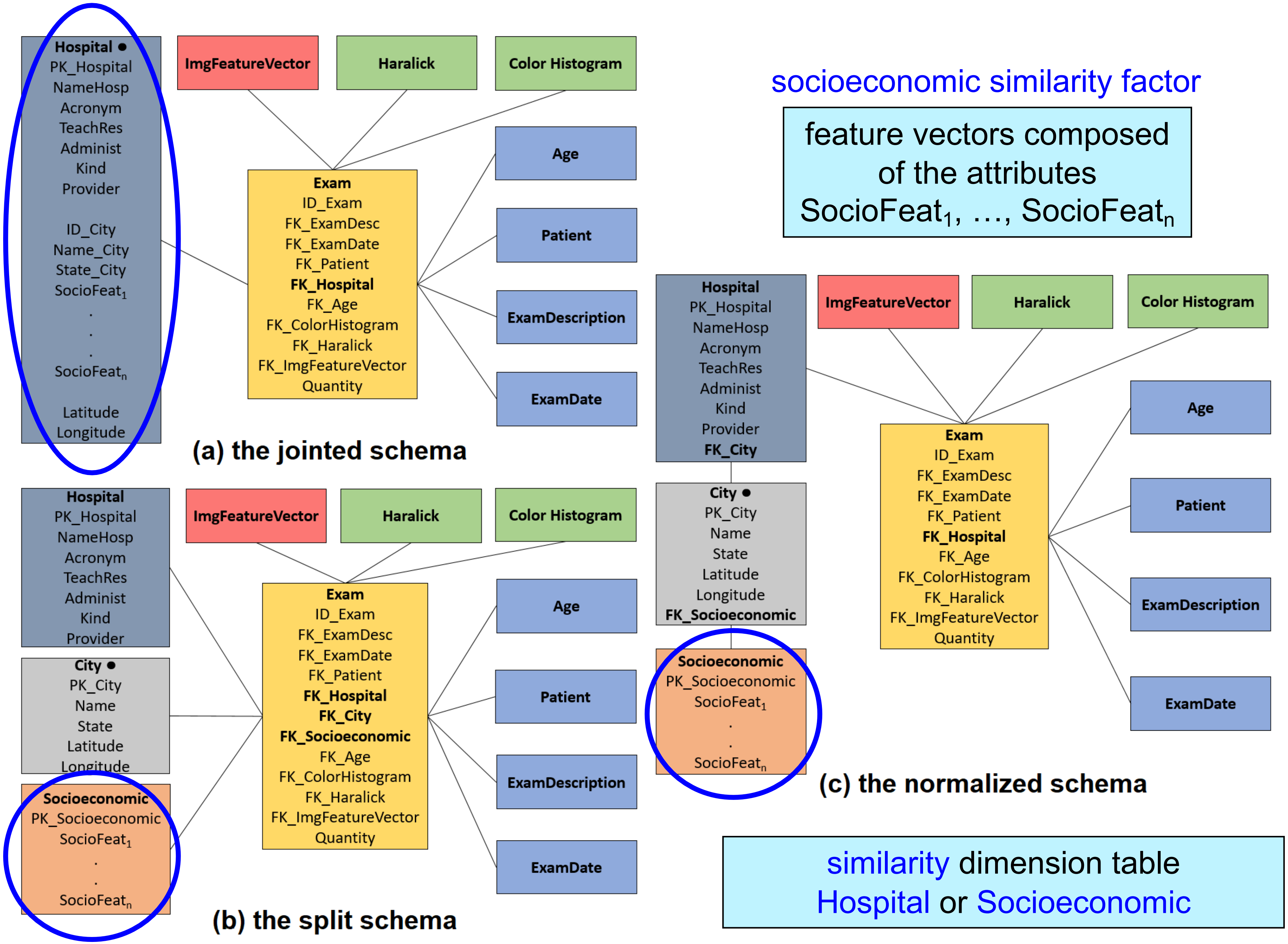


(a) the jointed schema

(b) the split schema

(c) the normalized schema

similarity dimension table
Hospital or City



(a) the jointed schema

(b) the split schema

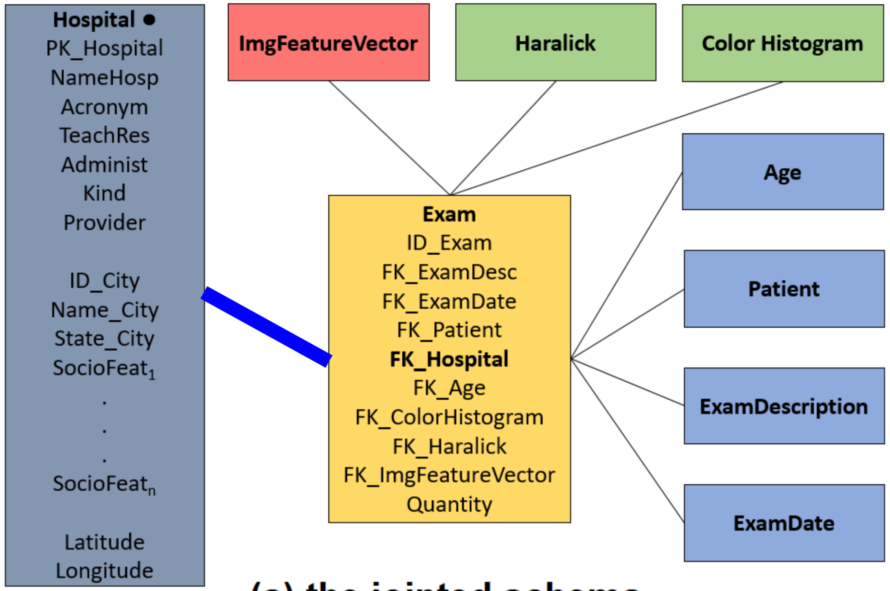
(c) the normalized schema

socioeconomic similarity factor

feature vectors composed of the attributes SocioFeat$_1$, …, SocioFeat$_n$

similarity dimension table Hospital or Socioeconomic

relationship with Exam



Hospital ●
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider

ID_City
Name_City
State_City
SocioFeat$_1$
.
.
.
SocioFeat$_n$

Latitude
Longitude

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVector
Quantity

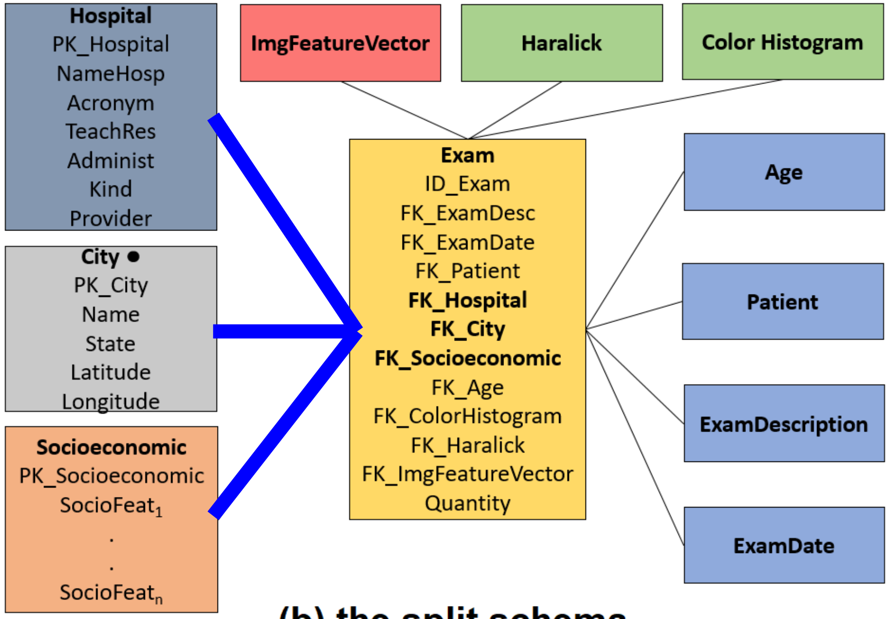
Age

Patient

ExamDescription

ExamDate

(a) the jointed schema



Hospital
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider

City ●
PK_City
Name
State
Latitude
Longitude

Socioeconomic
PK_Socioeconomic
SocioFeat$_1$
.
.
SocioFeat$_n$

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_City
FK_Socioeconomic
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVector
Quantity

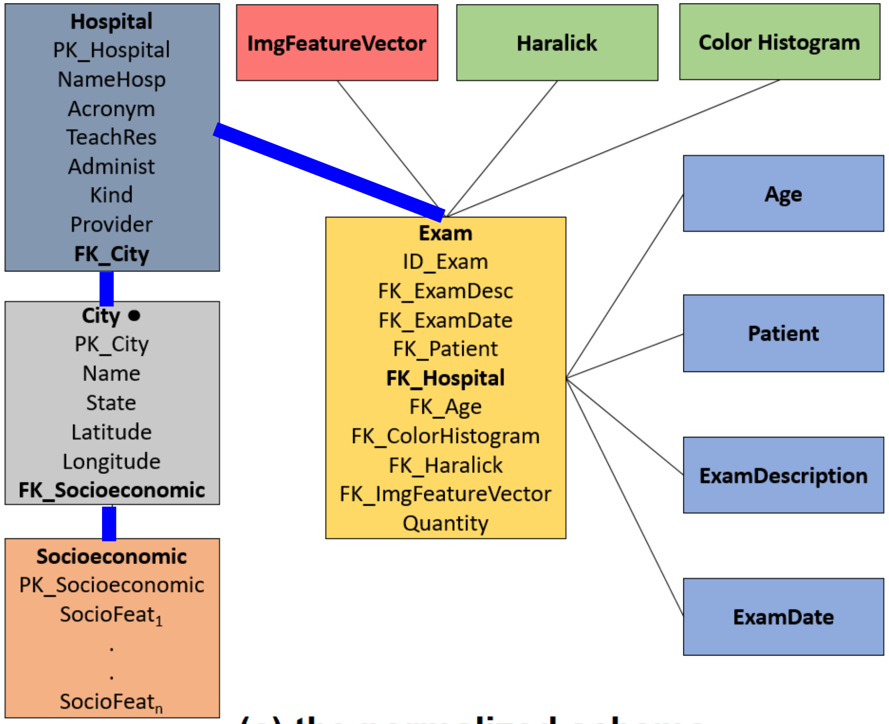
Age

Patient

ExamDescription

ExamDate

(b) the split schema



Hospital
PK_Hospital
NameHosp
Acronym
TeachRes
Administ
Kind
Provider
FK_City

City ●
PK_City
Name
State
Latitude
Longitude
FK_Socioeconomic

Socioeconomic
PK_Socioeconomic
SocioFeat$_1$
.
.
SocioFeat$_n$

ImgFeatureVector

Haralick

Color Histogram

Exam
ID_Exam
FK_ExamDesc
FK_ExamDate
FK_Patient
FK_Hospital
FK_Age
FK_ColorHistogram
FK_Haralick
FK_ImgFeatureVector
Quantity

Age

Patient

ExamDescription

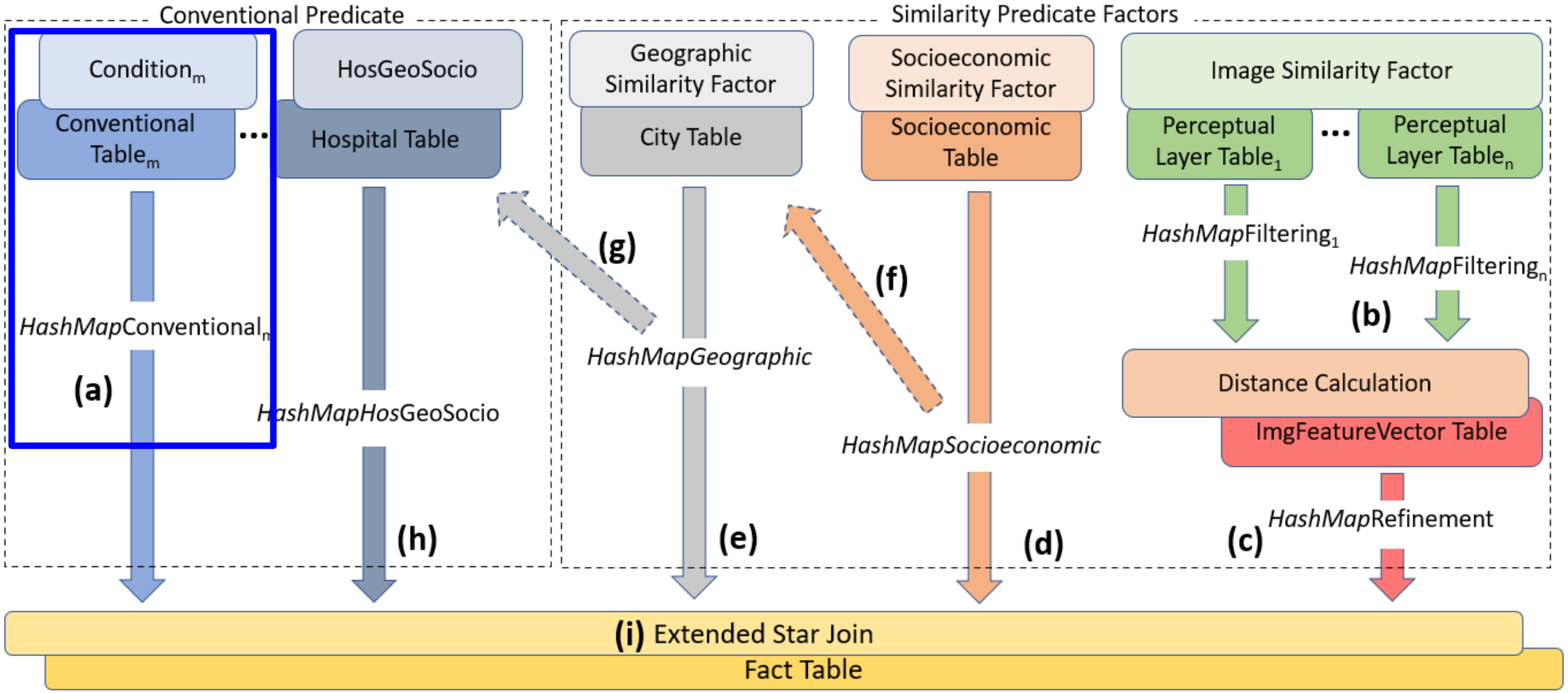
ExamDate

(c) the normalized schema

Outline

- Motivation
- Contributions
  - Designs of Star Schema
  - The SimSparkOLAP Method
- Experimental Evaluation
- Conclusions and Future Work

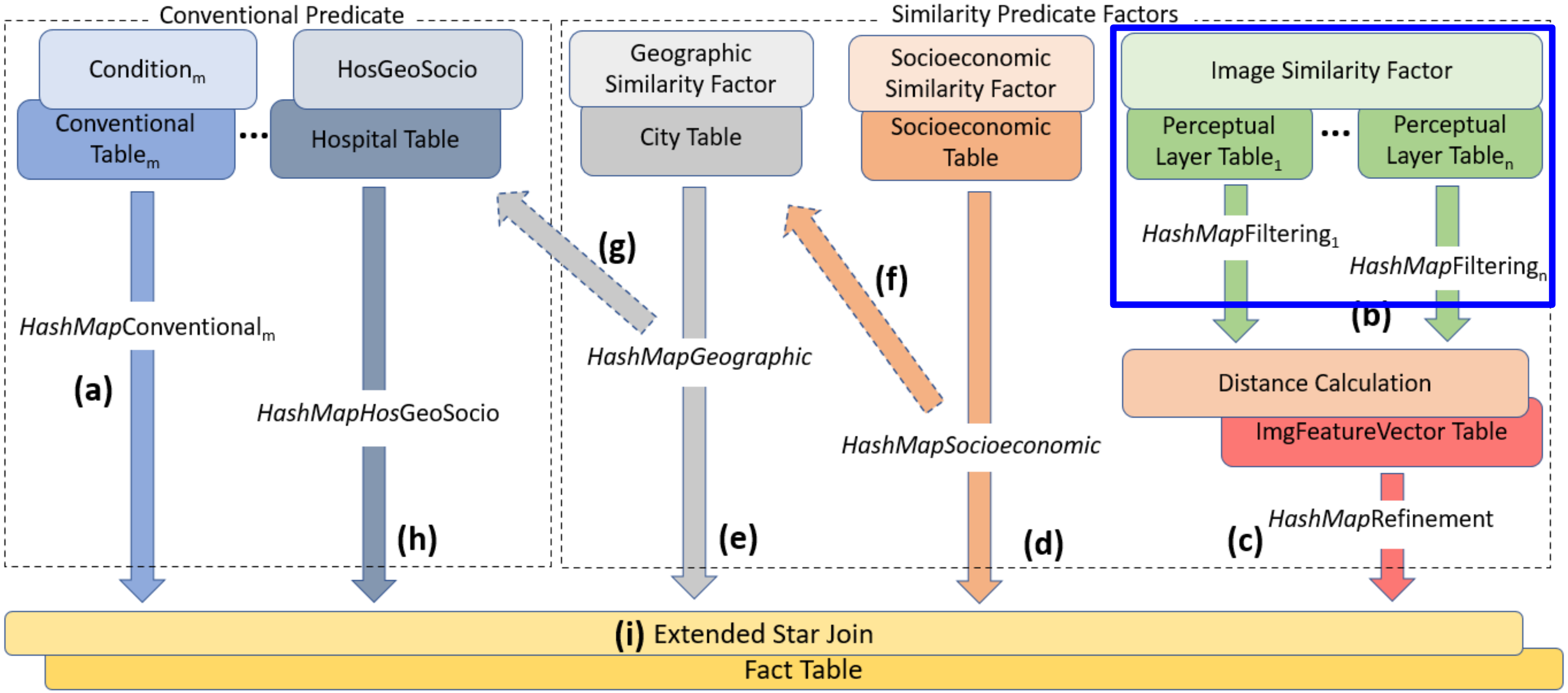
General View of SimSparkOLAP



a) Processing the conventional predicate (common task)

- each conventional dimension table is accessed to process the selection conditions
- the results are stored in the structures *HashMapConventional*

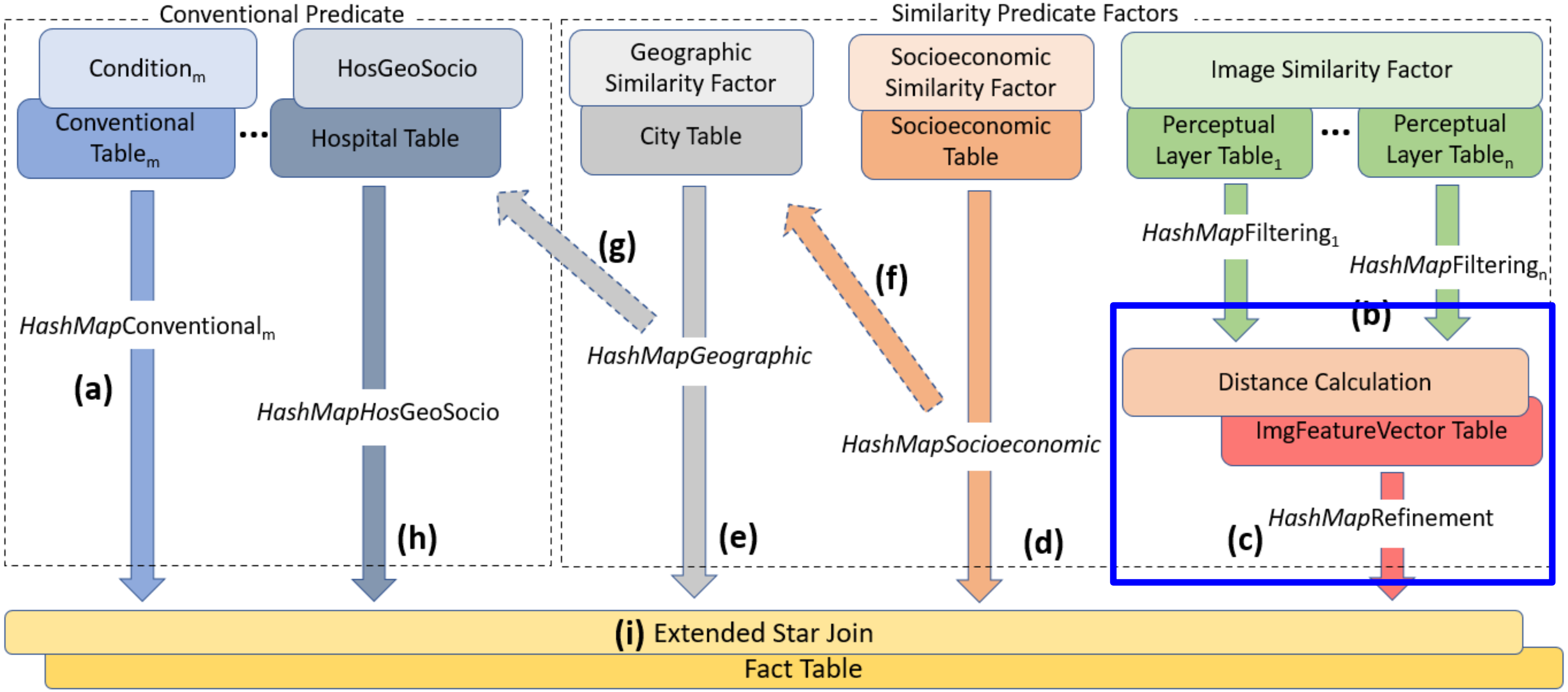
General View of SimSparkOLAP



b) Processing the image similarity predicate (common task)

- Each perceptual layer table is accessed to filter the image data
- The results are stored in the structures *HashMapFiltering*

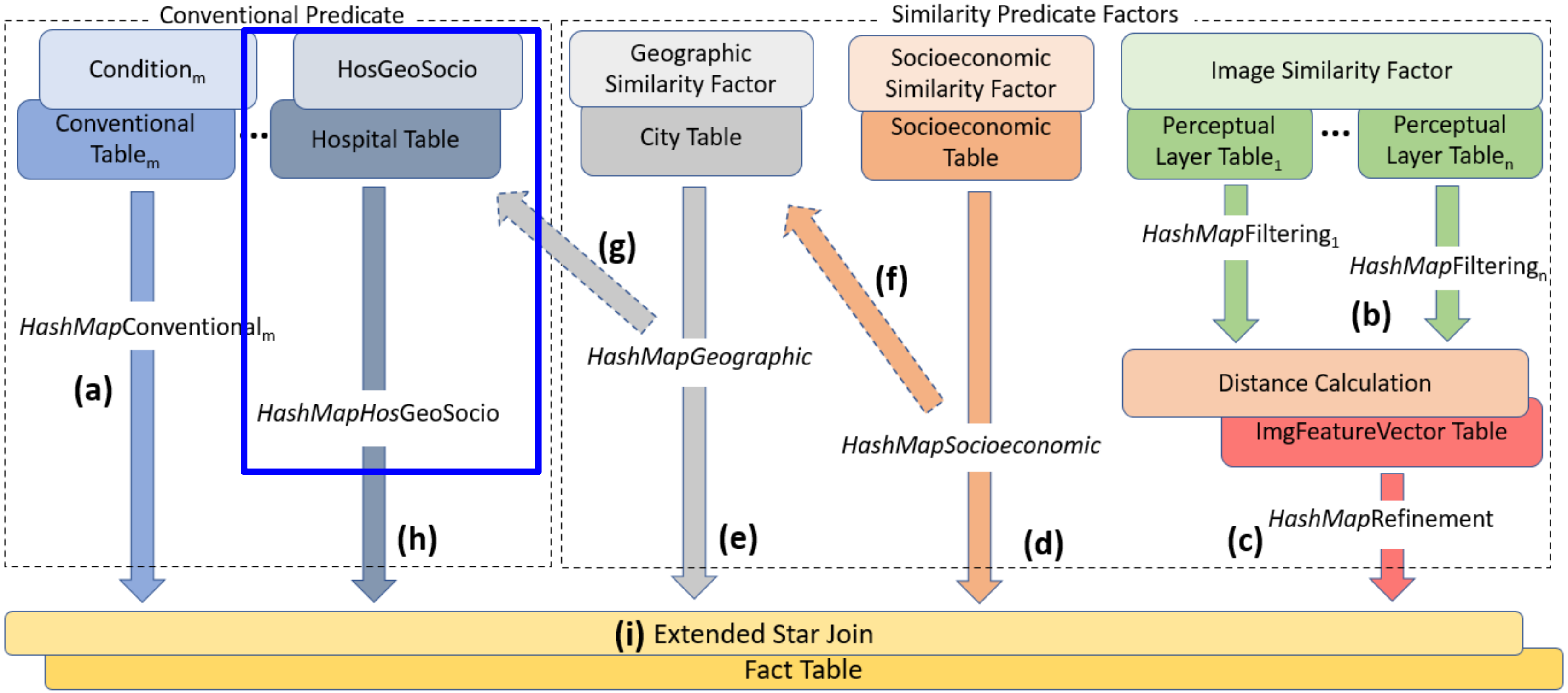
General View of SimSparkOLAP



c) Processing the image similarity predicate (common task)

- the feature vector table is accessed to eliminate false positives
- the results are stored in the structure *HashMapRefinement*

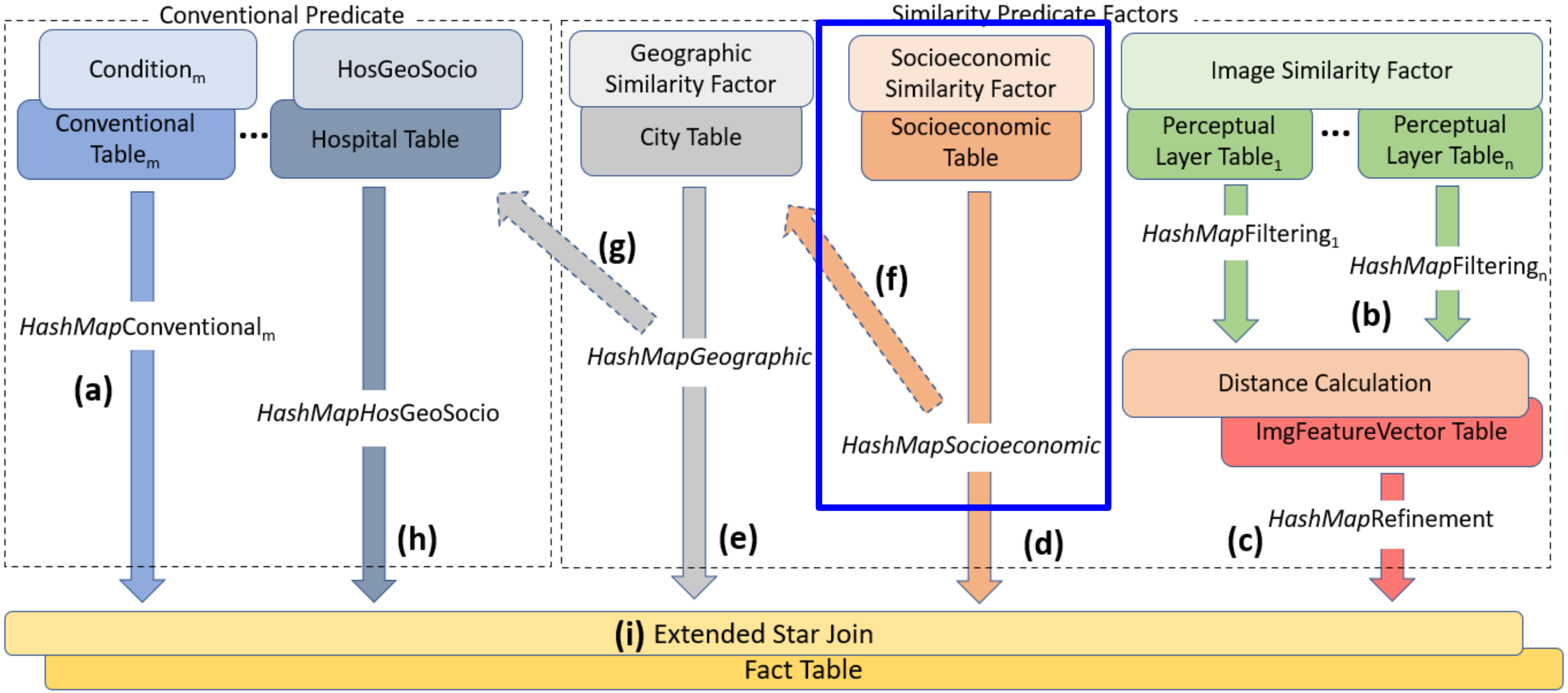
General View of SimSparkOLAP



h) The jointed schema (a → b → c → h)

- The geographic, socioeconomic, and conventional predicates are processed against the table Hospital
- The results are stored in the structure *HashMapHosGeoSocio*

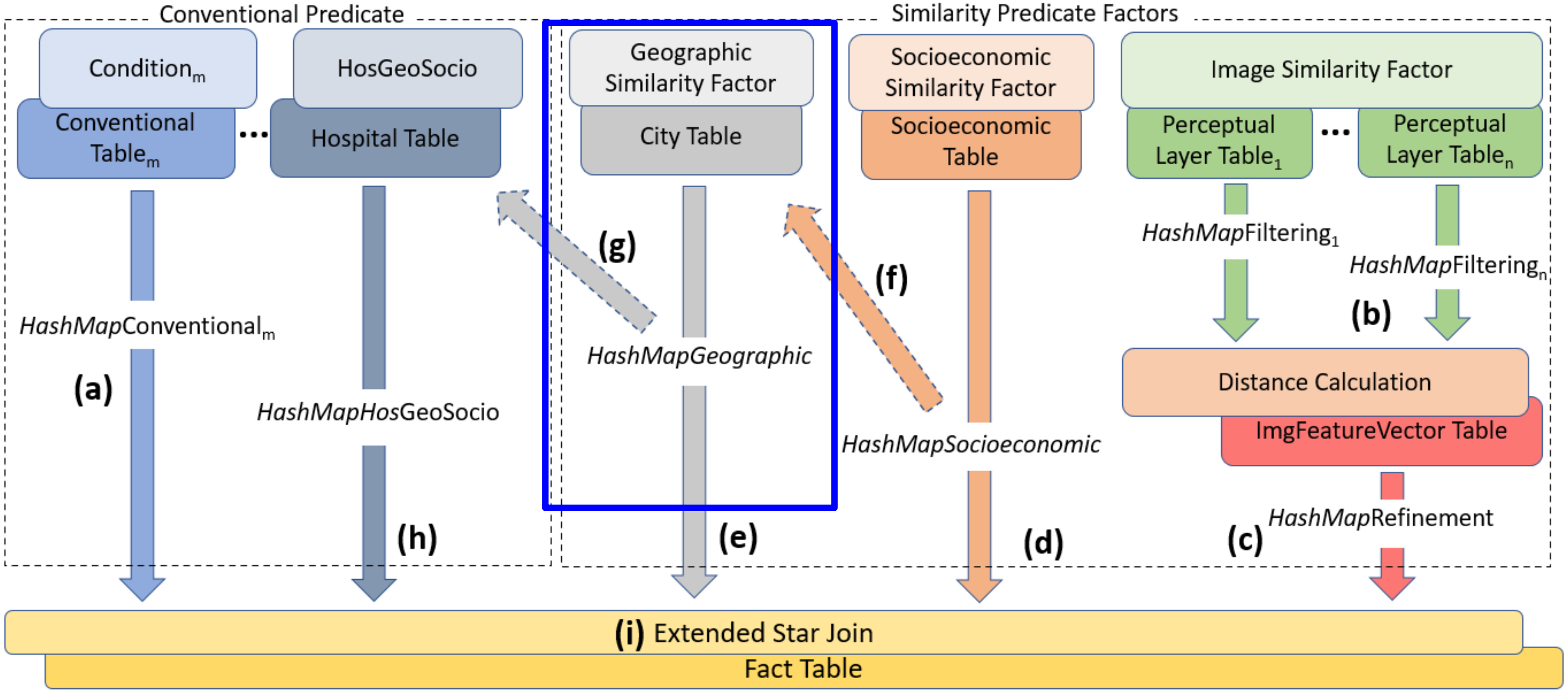
General View of SimSparkOLAP



d) The split schema (1/3) (a → b → c → d)

- The socioeconomic similarity predicate is processed against the table Socioeconomic
- The results are stored in the structure *HashMapSocioeconomic*

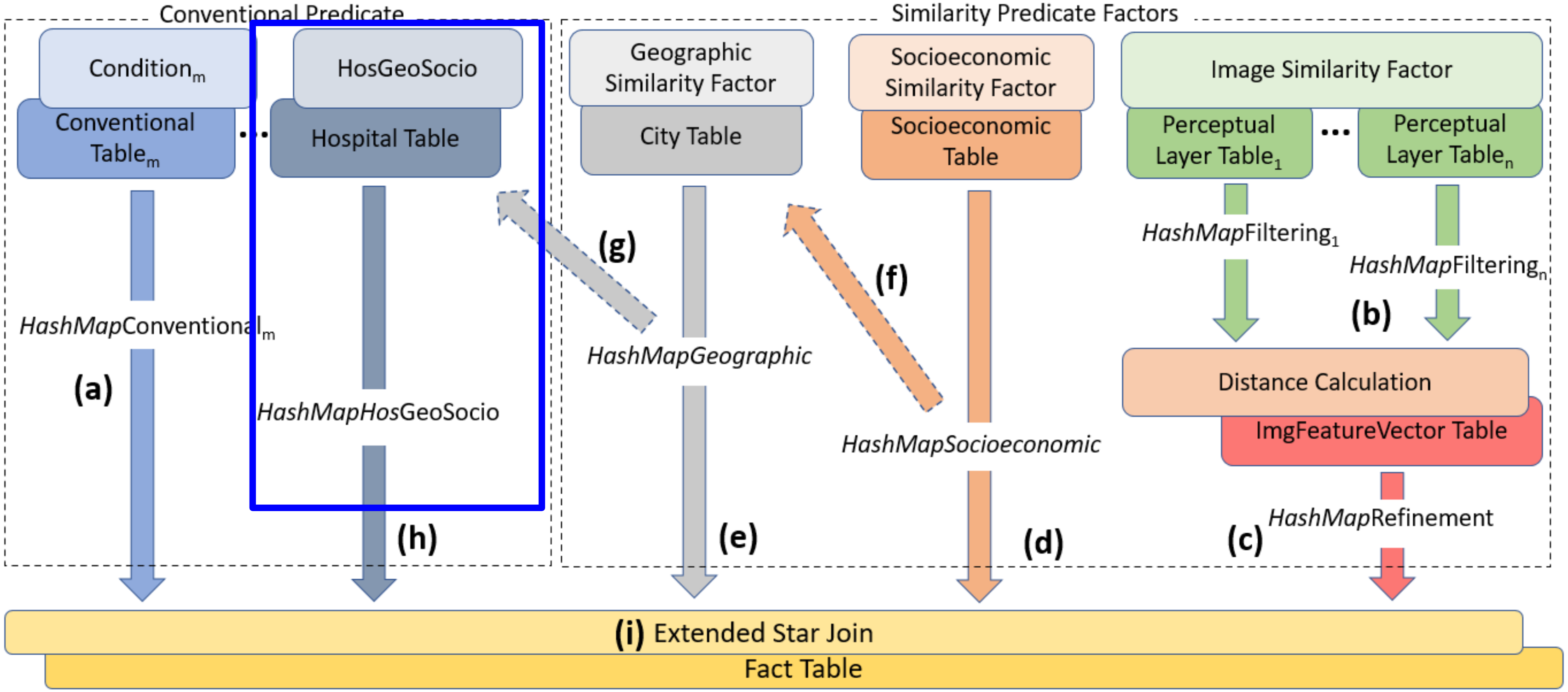
General View of SimSparkOLAP



e) The split schema (2/3) (a → b → c → d → e)

- The geographic similarity predicate is processed against the table City
- The results are stored in the structure *HashMapGeographic*

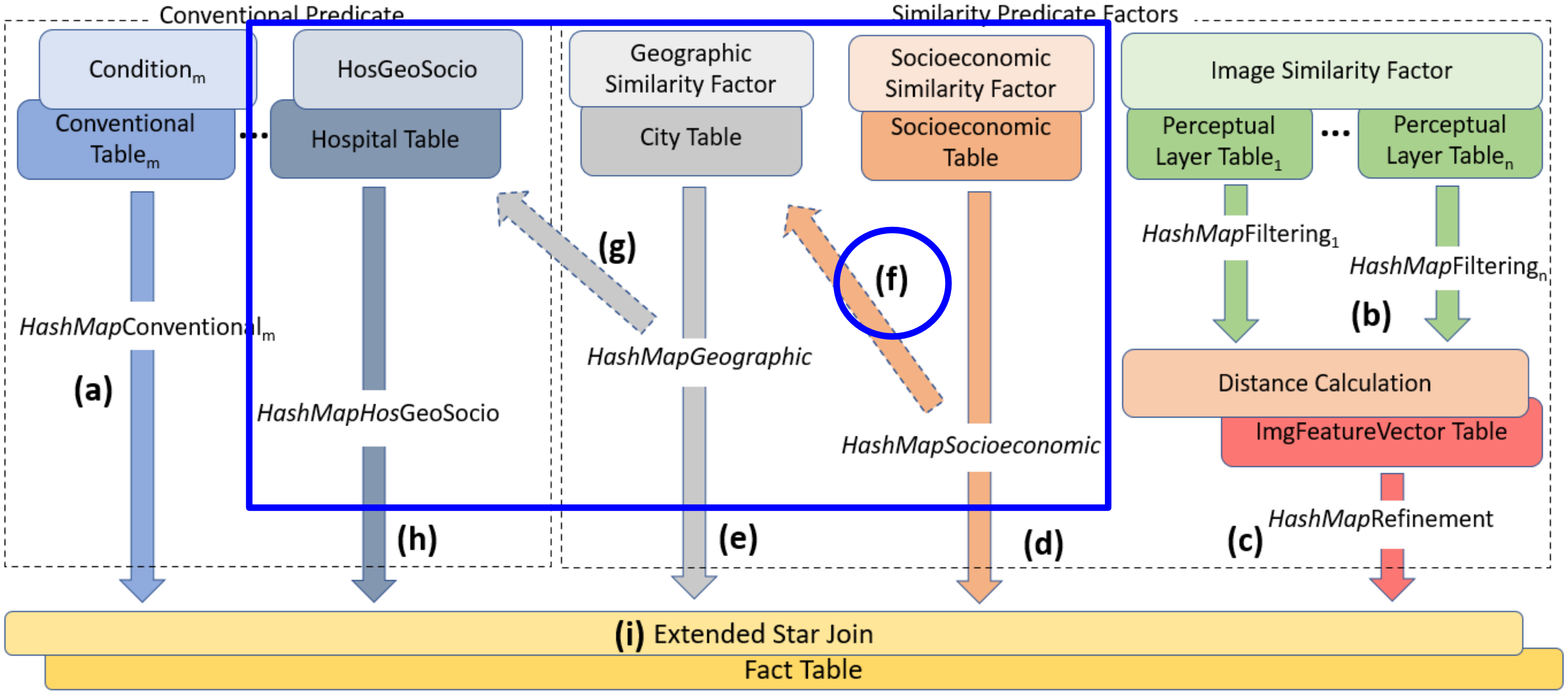
General View of SimSparkOLAP



h) The split schema (3/3) (a → b → c → d → e → h)

- The conventional predicate defined on the attributes of Hospital is processed against this table
- The results are stored in the structure *HashMapHosGeoSocio*

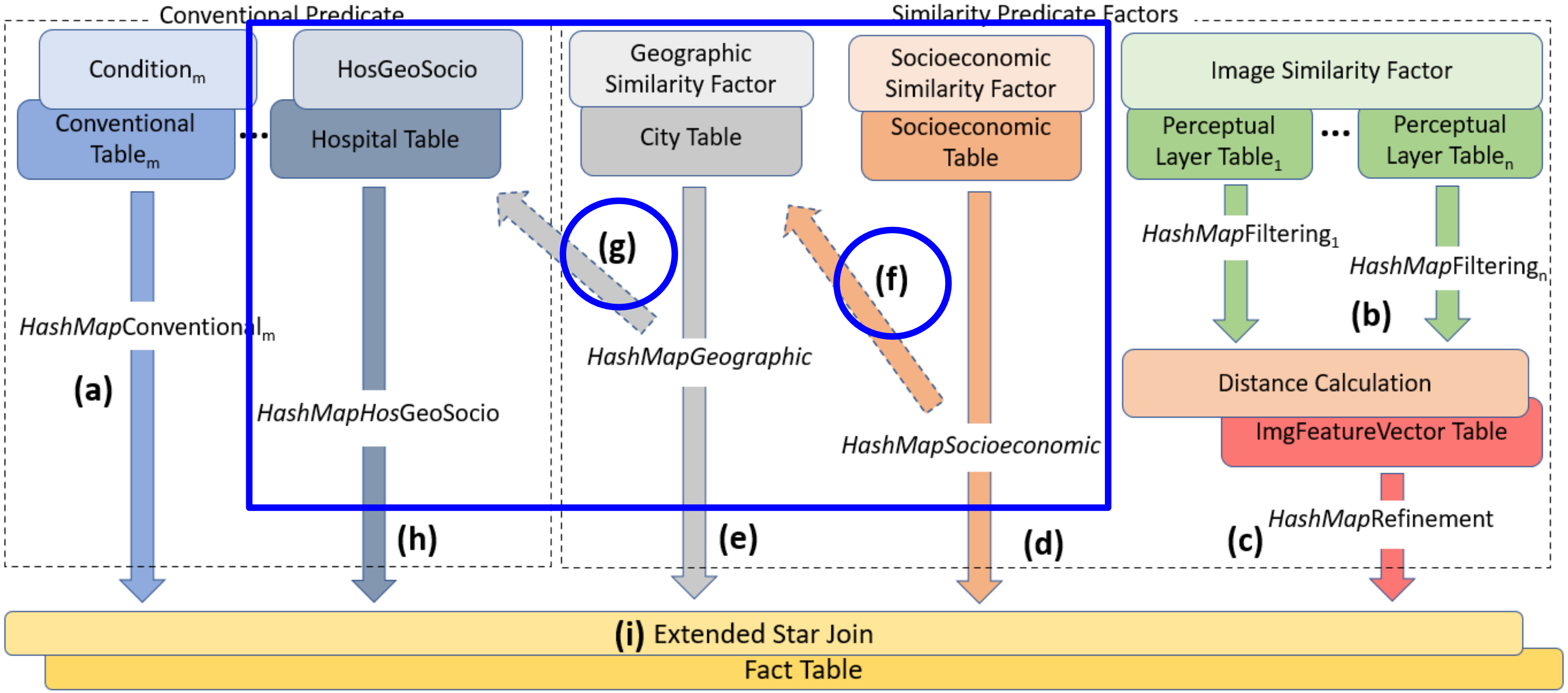
General View of SimSparkOLAP



f) The normalized schema (1/3) (a → b → c → f)

- The socioeconomic similarity predicate is processed against the table Socioeconomic
- The results are stored in the structure *HashMapSocioeconomic*

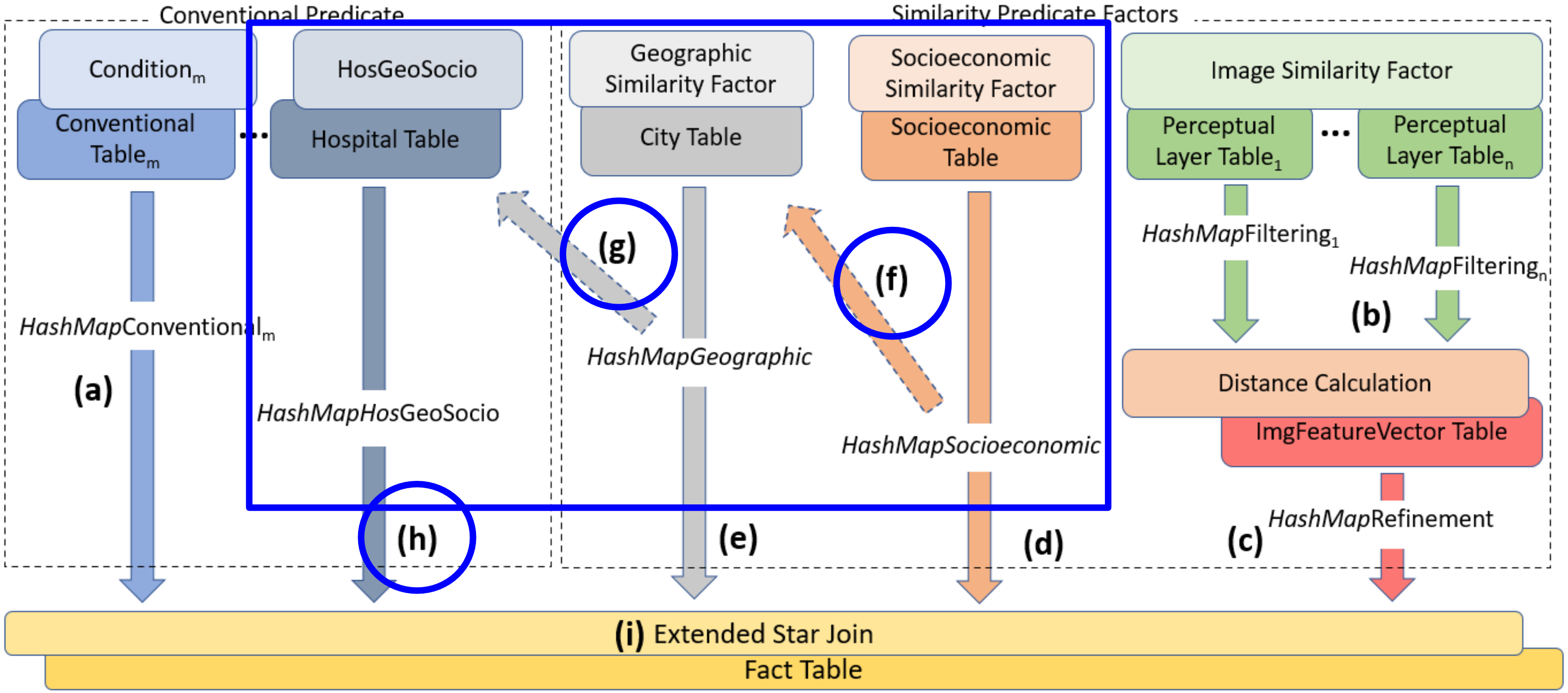
General View of SimSparkOLAP



g) The normalized schema (2/3) (a → b → c → f → g)

- The structure *HashMapSocioeconomic* is associated to the table City to process the geographic similarity predicate
- The results are stored in the structure *HashMapGeographic*

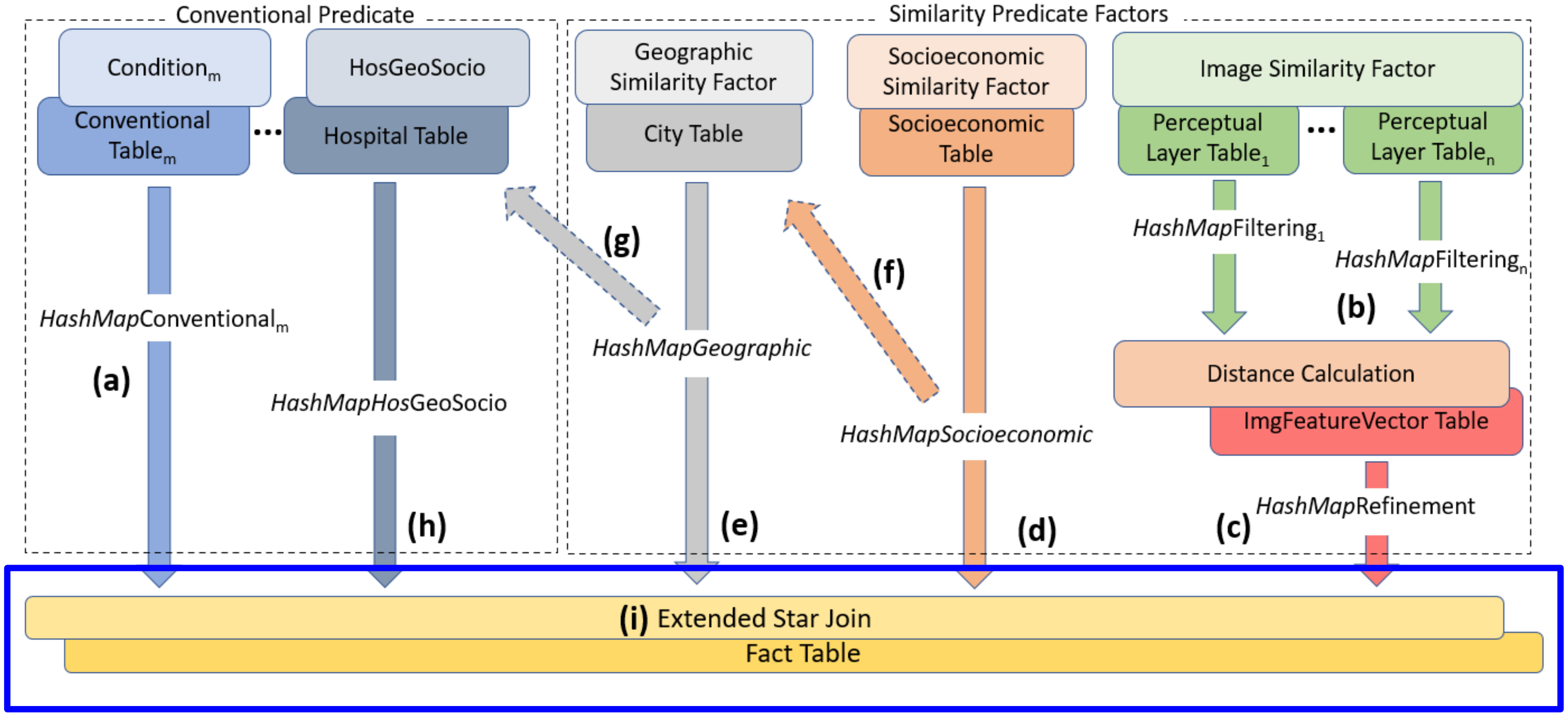
General View of SimSparkOLAP



h) The normalized schema (3/3) (a → b → c → f → g → h)

- The structure *HashMapGegraphic* is associated to the table Hospital to process the conventional predicate
- The results are stored in the structure *HashMapHosGeoSocio*

General View of SimSparkOLAP



i) Broadcasting the structures (common task)

- All the structures are broadcasted to all nodes of the cluster
- The extended star join is performed against the fact table Exam

Outline

- Motivation
- Contributions
  - Designs of Star Schema
  - The SimSparkOLAP Method
- Experimental Evaluation
- Conclusions and Future Work

Experimental Setup

- Real and synthetic data
  - medical images and conventional data from the ImgDW Generator tool
  - geographic and socioeconomic data from US cities from the Census dataset from year 2000

Tables	# Tuples
Exam	30 millions
ExamDate	18,268
ExamDescription	3 millions
Patient	300,000
Age	121

Tables	# Tuples
Color Histogram	3 millions
Haralick	3 millions
ImgFeatureVector	3 millions
Hospital	100,000
City	25,000 pairs of (Lat, Long)
Socioeconomic	25,000 sets of 95 features

Experimental Setup

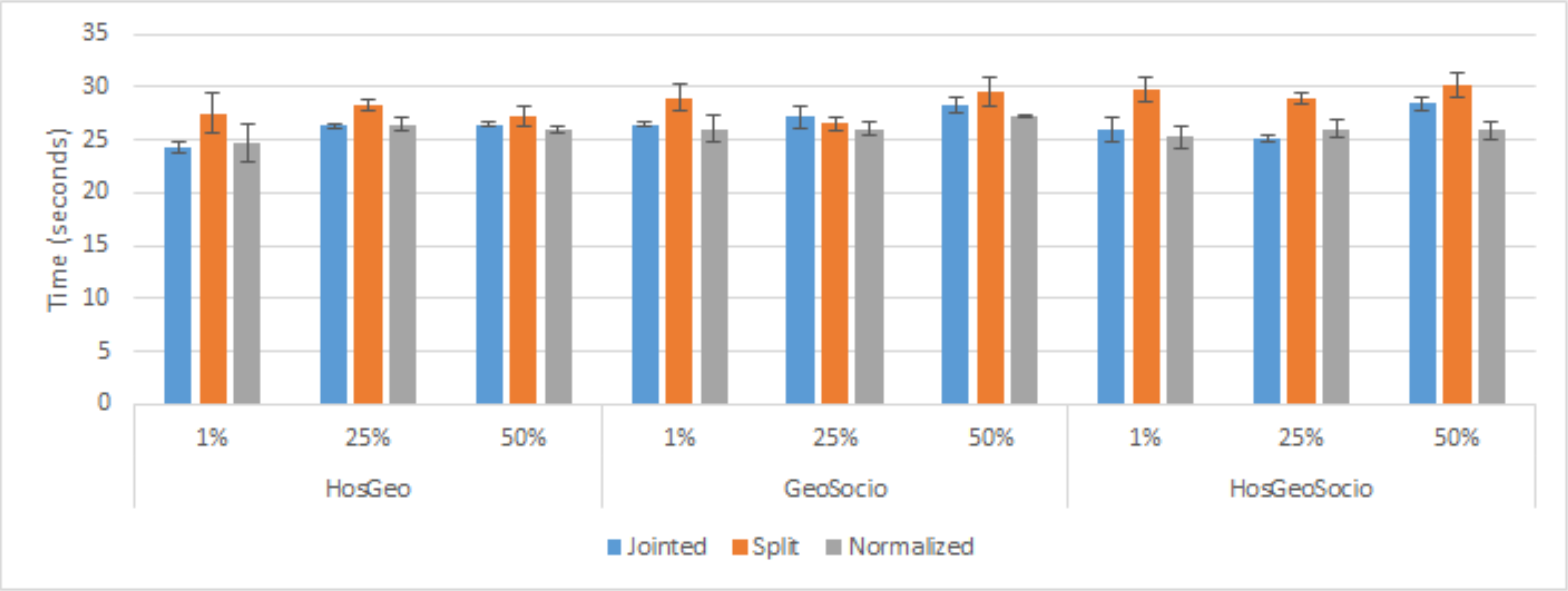
- Machine
  - cluster with 5 nodes
  - each node had, at least, 3GB of RAM
- Execution
  - Each query was executed 10 times
  - All cache and buffers were flushed after finishing each query
  - Outliers were removed

Experimental Setup

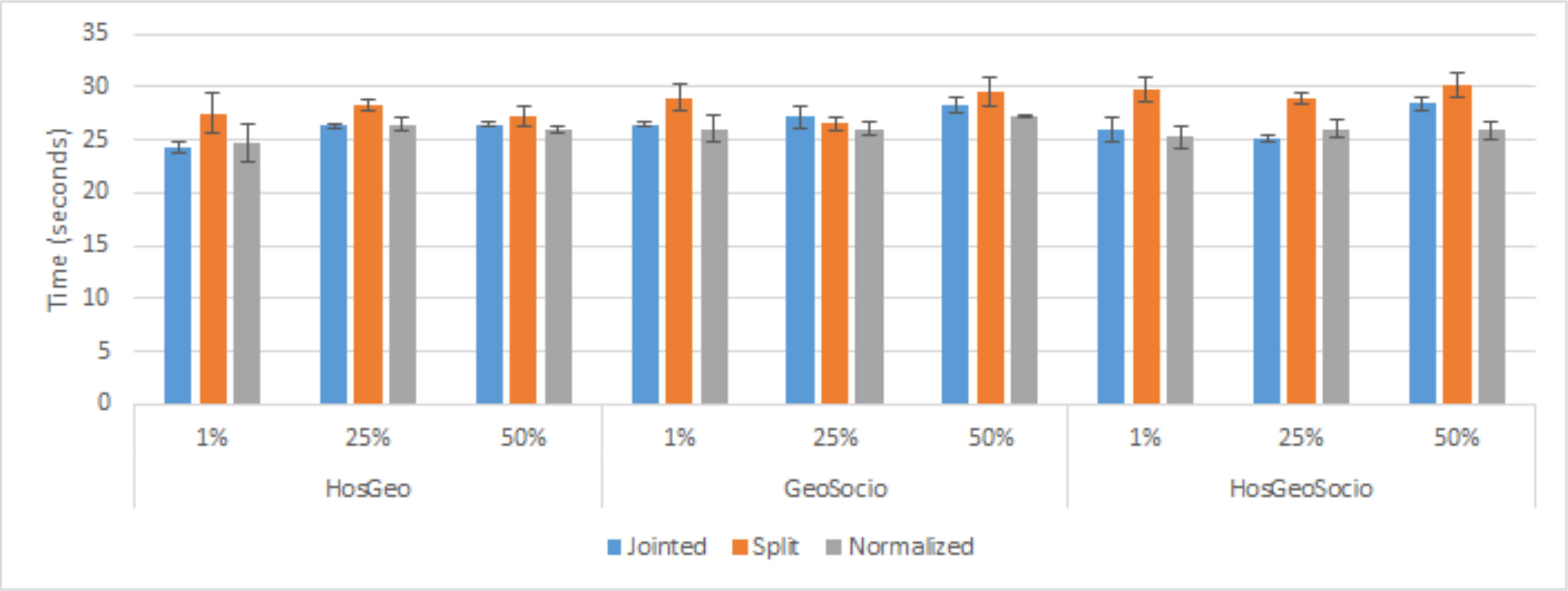
Configurations	**Predicates**		
	Conventional	Geographic	Socioeconomic
HosGeo	X	X	
GeoSocio		X	X
HosGeoSocio	X	X	X

- Other parameters
  - values of selectivity: 1%, 25%, and 50%
  - values of radius for the geographic range query: 35 km, 900 km, and 1,500 km from New York
  - distance functions: Euclidean (image and socioeconomic) and DGDist (geographic)

Effect of the Star Schemas



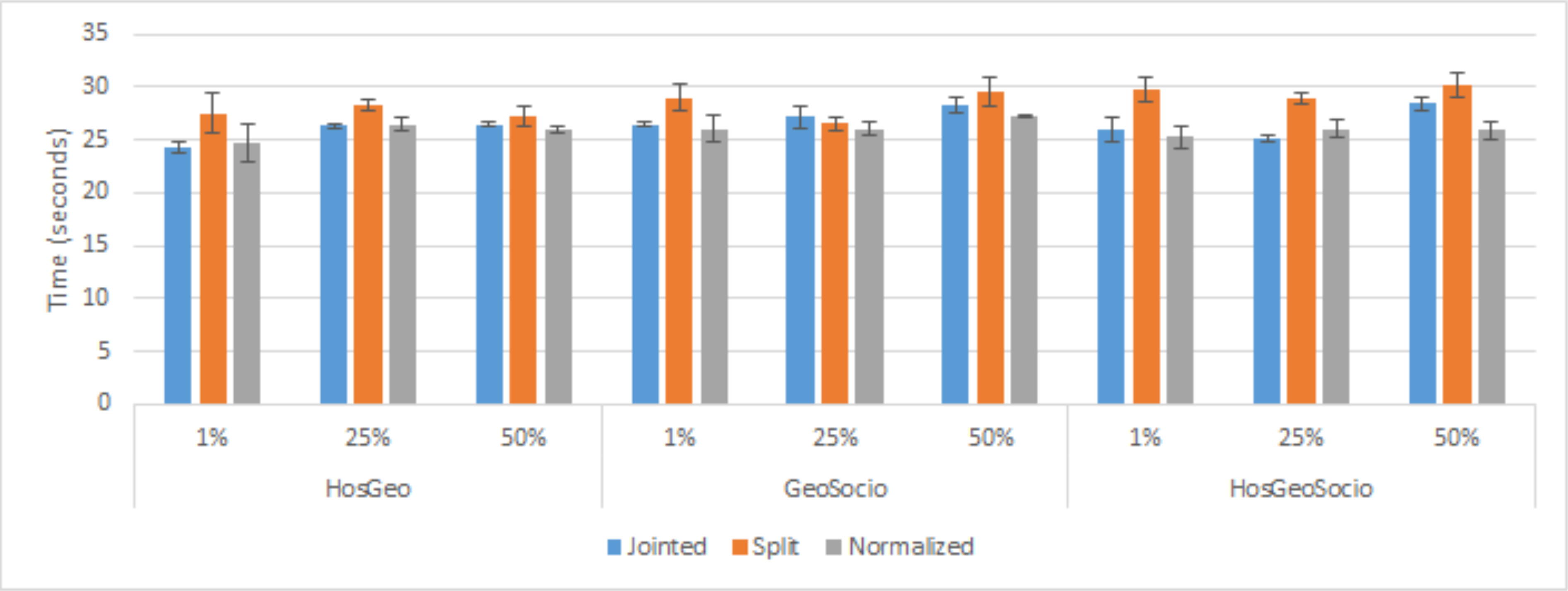
Effect of the Star Schemas



the normalized schema produced the best performance results, followed by the jointed schema, followed by the split schema

normalized requires joins between significantly smaller similarity tables

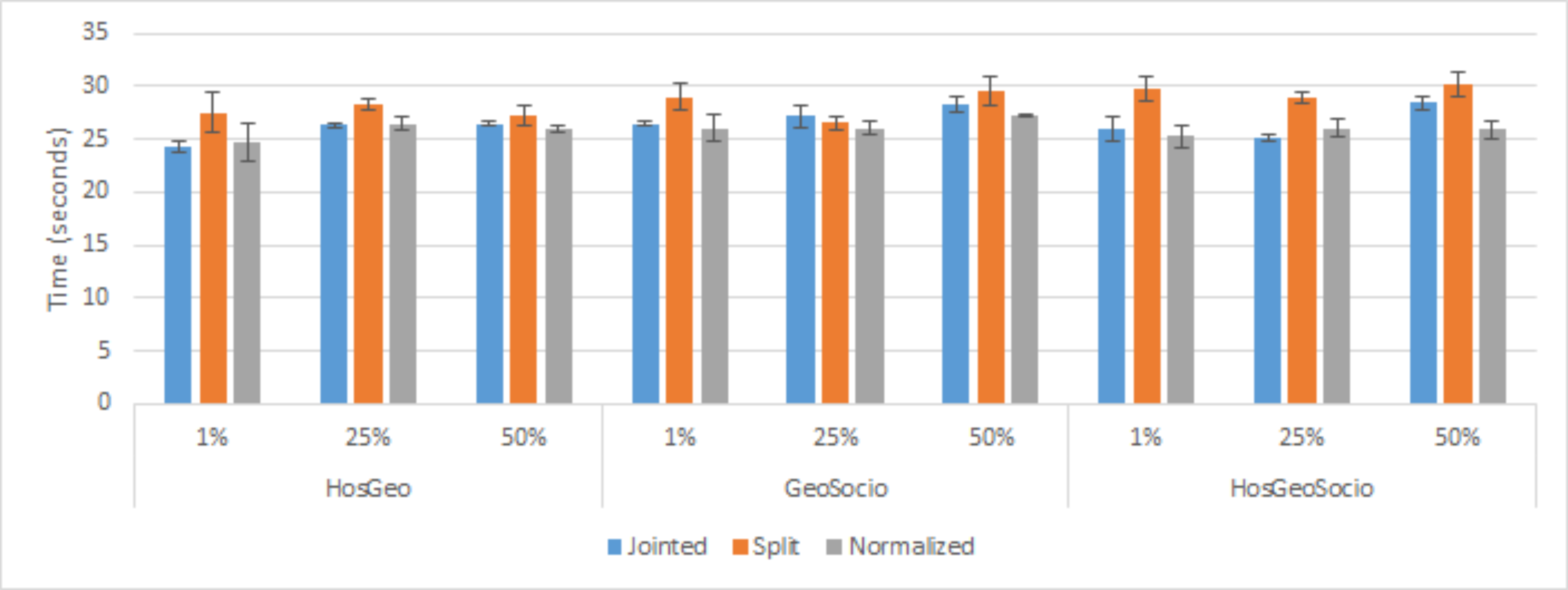
Effect of the Star Schemas



the normalized schema produced the best performance results, followed by the jointed schema, followed by the split schema

normalized x jointed
up to 15.89%

Effect of the Star Schemas



the normalized schema produced the best performance results, followed by the jointed schema, followed by the split schema

jointed x split
up to 13.68%

Semantic Queries

- Importance to the healthcare decision-making
  - analyzing conventional and image data may show the evolution curve of a given disease over time
  - investigating geographic areas around a point of interest may reveal an epicenter
  - studying socioeconomic data may demonstrate how a given disease affect people from different age ranges, salary ranges, and education levels

Outline

- Motivation
- Contributions
  - Three Designs of Star Schema
  - The SimSparkOLAP Method
- Experimental Evaluation
- Conclusions and Future Work

Conclusions

- Management of geographic, socioeconomic, and image similarity factors
  - how to store these similarity factors in DWs
    - three designs of star schema
  - how to process analytical queries extended with these similarity factors in Spark
    - the SimSparkOLAP method
  - how to use semantic analytical queries extended with these similarity factors to improve the healthcare decision-making

Future Work

- Execution of new performance tests, considering different
  - data volumes
  - healthcare datasets
- Analysis of other types of extended analytical queries

Thank you!

Guilherme M. Rocha, Piero L. Capelo, Cristina D. A. Ciferri

{guilherme.muzzi.rocha,piero.capelo}@usp.br

cdac@icmc.usp.br

USP ICMC

FAPESP

CNPq

CAPES