



POLITECNICO
MILANO 1863



USC University of
Southern California

Information Sciences Institute

Exploiting the Semantic Web for the Automatic Extraction of Los Angeles City Data

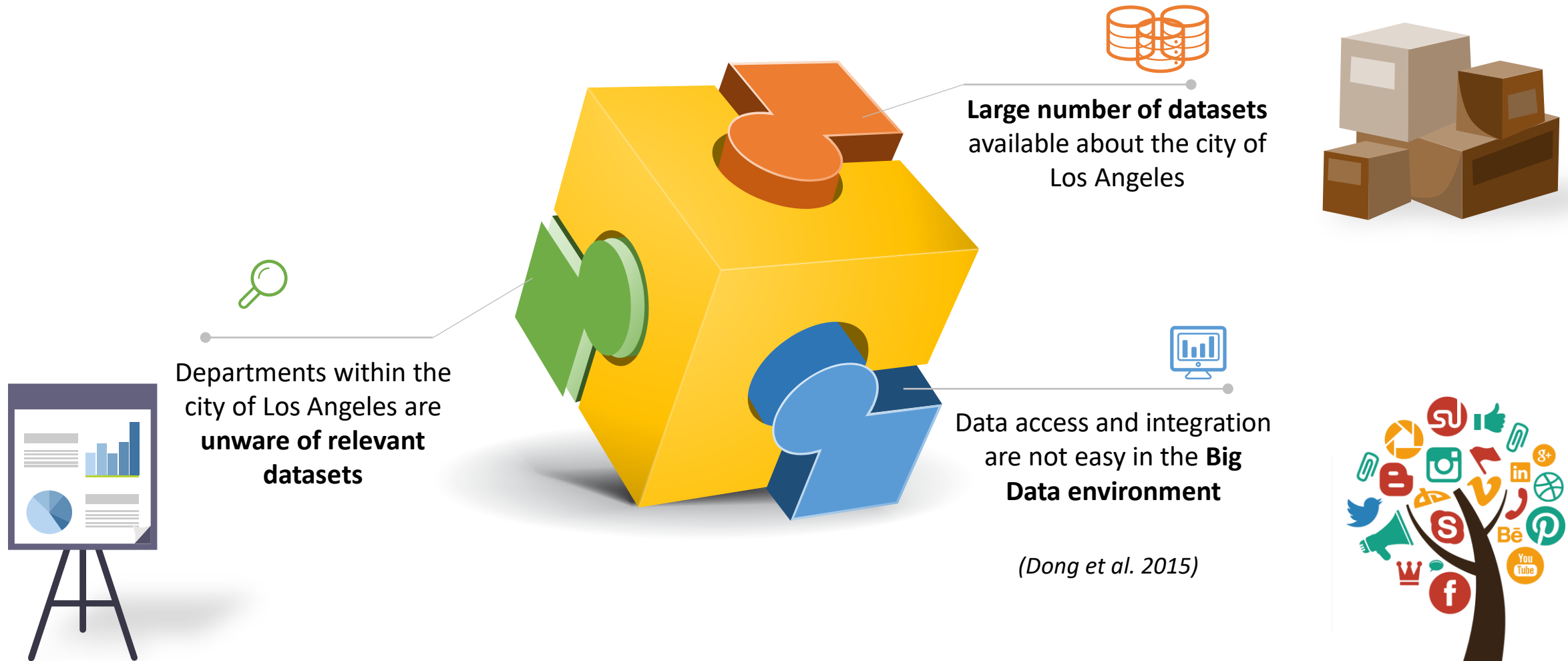
Advisor: Letizia Tanca

Co-Advisor: Craig Knoblock

Author: Marianna Bucchi

Student ID: 898422

Introduction: Problem Identification

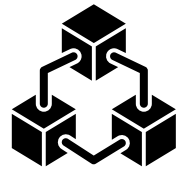


The Solution

Provide a **better access** to the data the city already has



How?



Automatically extracting the **content** from each dataset in a form of a class

Preliminary Notions

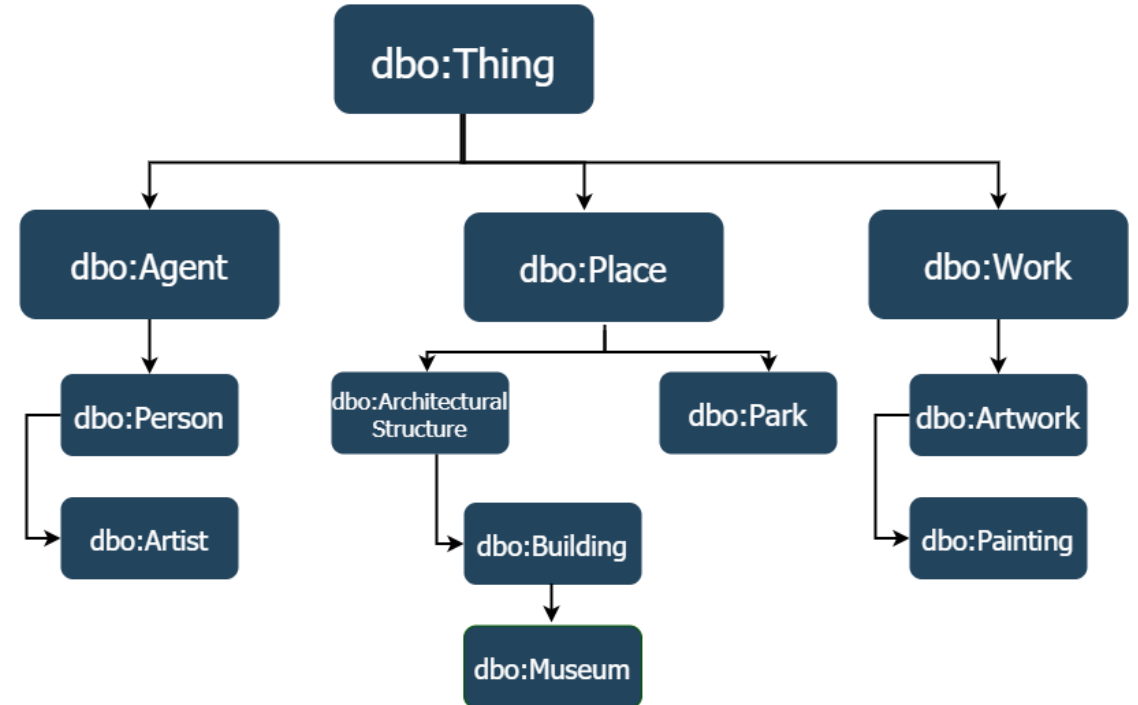
The Web was designed as an information space with the goal to be useful for human-to-human communication, while the **Semantic Web** approach develops methods and languages for expressing information in a machine-processable form.

(Berners-Lee et al. 2001)

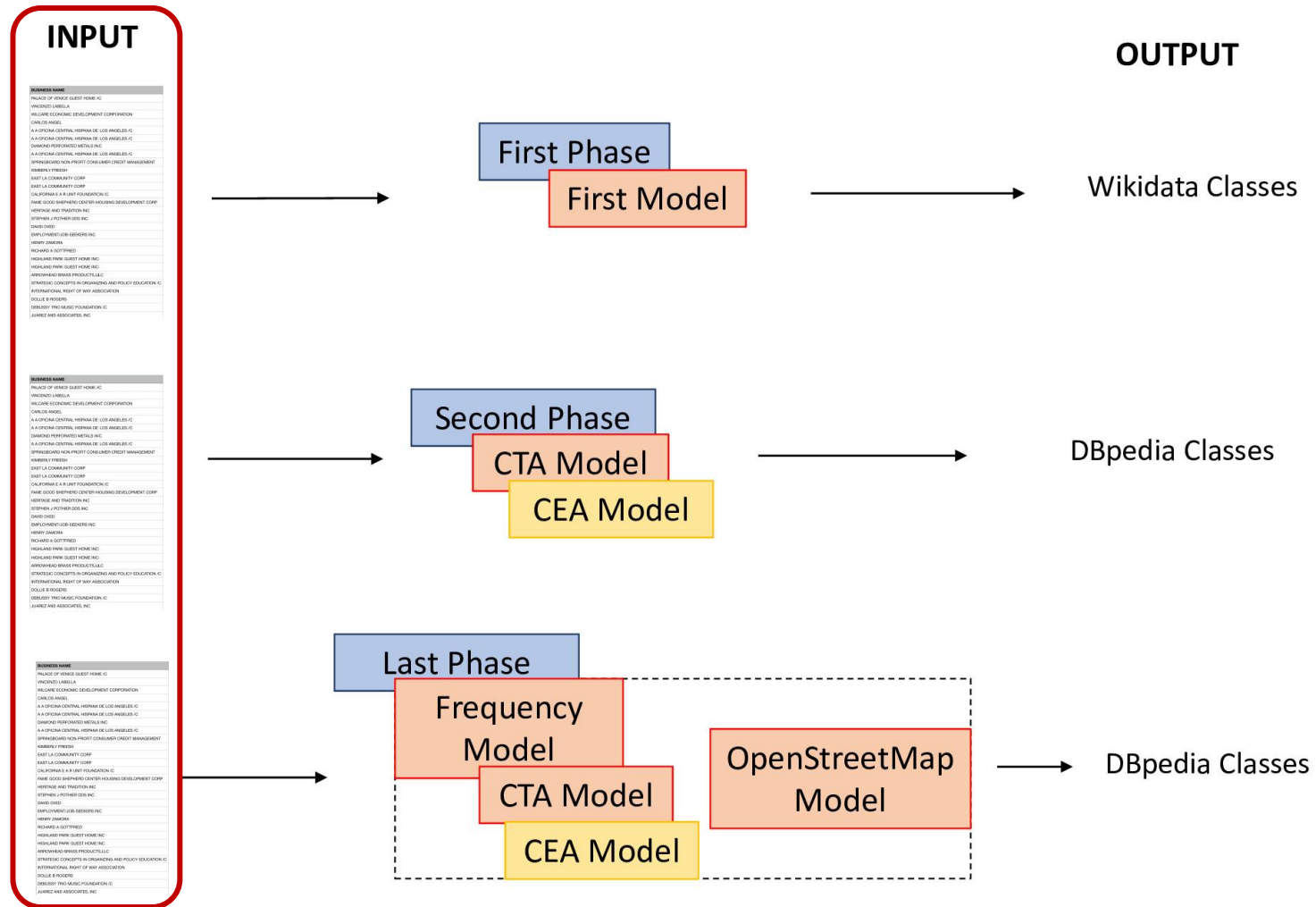


Ontologies are stores of information accessible through queries on the Web, which describe the contextual relations between concepts and specify logical rules for reasoning about them.

(Ismayilov et al. 2018)

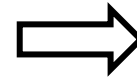


The Model Development



Input: Column Selection Phase

From the observation of the datasets, we collected **some exclusion criteria** to isolate a single column from a dataset



INPUT

BUSINESS NAME	DBA NAME	STREET ADDRESS	CITY
PALACE OF VENICE GUEST HOME /C		1727 CRENSHAW BLVD	LOS ANGELES
VINCENZO LABELLA		921 SWARTHMORE AVENUE	PACIFIC PALISADES
WILCARE ECONOMIC DEVELOPMENT CORPORATION		9911 AVALON BLVD	LOS ANGELES
CARLOS ANGEL		1221 W 7TH STREET SUITE 4H-407	LOS ANGELES
A A OFICINA CENTRAL HISPANA DE LOS ANGELES /C		4917 S BROADWAY	LOS ANGELES
A A OFICINA CENTRAL HISPANA DE LOS ANGELES /C		1330 WILSHIRE BLVD #208	LOS ANGELES
DIAMOND PERFORATED METALS INC		11090 BEECH AVENUE	FONTANA
A A OFICINA CENTRAL HISPANA DE LOS ANGELES /C		2015 W TEMPLE STREET	LOS ANGELES
SPRINGBOARD NON-PROFIT CONSUMER CREDIT MANAGEMENT	MONEY MANAGEMENT INTERNATIONAL	1605 W OLYMPIC BLVD #9023	LOS ANGELES
KIMBERLY FREESH		10926 OWENSMOUTH AVENUE	CHATSWORTH
EAST LA COMMUNITY CORP		121 N CHICAGO STREET	LOS ANGELES
EAST LA COMMUNITY CORP		115 N SOTO STREET	LOS ANGELES
CALIFORNIA E A R UNIT FOUNDATION /C		29654 DRIVER AVENUE	CASTAIC
FAME GOOD SHEPHERD CENTER HOUSING DEVELOPMENT CORP		2420 S WESTERN AVENUE	LOS ANGELES
HERITAGE AND TRADITION INC		3756 ALOHA STREET	LOS ANGELES
STEPHEN J POTHER DDS INC		9720 RESEDA BLVD #2	NORTHRIDGE
DAVID OVED		12304 SANTA MONICA BLVD #209	LOS ANGELES
EMPLOYMENT/JOB-SEEKERS INC		4005 10TH AVENUE #4	LOS ANGELES
HENRY ZAMORA		6730 TULUNGA AVENUE	NORTH HOLLYWOOD
RICHARD A GOTTFRIED	RICHARD A GOTTFRIED JD MBA MFT	12304 SANTA MONICA BLVD #215	LOS ANGELES
HIGHLAND PARK GUEST HOME INC	HIGHLAND PARK GST. HM INC HIGHLAND PARK GUEST HOME	345 N AVENUE 57	LOS ANGELES
HIGHLAND PARK GUEST HOME INC		346 N AVENUE 57	LOS ANGELES
ARROWHEAD BRASS PRODUCTS,LLC		5147 ALHAMBRA AVENUE	LOS ANGELES
STRATEGIC CONCEPTS IN ORGANIZING AND POLICY EDUCATION /C	SCOPE	1715 W FLORENCE AVENUE	LOS ANGELES
INTERNATIONAL RIGHT OF WAY ASSOCIATION		19750 S VERMONT AVENUE #220	TORRANCE
DOLLIE B ROGERS		5420 COMPTON AVENUE	LOS ANGELES
DEBUSSY TRIO MUSIC FOUNDATION /C		223 S BLUNDY DRIVE	LOS ANGELES
JUAREZ AND ASSOCIATES, INC		12139 NATIONAL BLVD	LOS ANGELES

Columns' Selection

OUTPUT

BUSINESS NAME
PALACE OF VENICE GUEST HOME /C
VINCENZO LABELLA
WILCARE ECONOMIC DEVELOPMENT CORPORATION
CARLOS ANGEL
A A OFICINA CENTRAL HISPANA DE LOS ANGELES /C
A A OFICINA CENTRAL HISPANA DE LOS ANGELES /C
DIAMOND PERFORATED METALS INC
A A OFICINA CENTRAL HISPANA DE LOS ANGELES /C
SPRINGBOARD NON-PROFIT CONSUMER CREDIT MANAGEMENT
KIMBERLY FREESH
EAST LA COMMUNITY CORP
EAST LA COMMUNITY CORP
CALIFORNIA E A R UNIT FOUNDATION /C
FAME GOOD SHEPHERD CENTER HOUSING DEVELOPMENT CORP
HERITAGE AND TRADITION INC
STEPHEN J POTHER DDS INC
DAVID OVED
EMPLOYMENT/JOB-SEEKERS INC
HENRY ZAMORA
RICHARD A GOTTFRIED
HIGHLAND PARK GUEST HOME INC
HIGHLAND PARK GUEST HOME INC
ARROWHEAD BRASS PRODUCTS,LLC
STRATEGIC CONCEPTS IN ORGANIZING AND POLICY EDUCATION /C
INTERNATIONAL RIGHT OF WAY ASSOCIATION
DOLLIE B ROGERS
DEBUSSY TRIO MUSIC FOUNDATION /C
JUAREZ AND ASSOCIATES, INC

- Containing all **equal values**
- Including **numbers** for the majority of the rows (i.e. more than 50% of rows are of integer or float type)
- Containing specific **symbols** for the majority of the rows (i.e. more than 50% of rows have symbols)
- Containing **underlined text**
- Including any information related to **time and space** (i.e. dates)
- Containing **e-mail** addresses

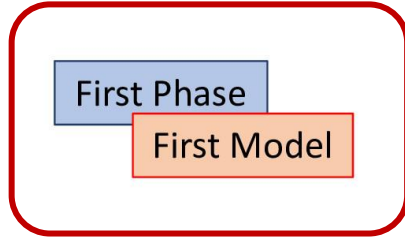
The Model Development

INPUT

BUSINESS NAME
PALACE OF VENICE GUEST HOME-IC
UNICREDIT LABELLA
MILANO ECONOMIC DEVELOPMENT CORPORATION
CARLOS ANGEL
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
SPRINGFIELD NON-PROFIT CONSUMER CREDIT MANAGEMENT
ARMANDO FREZZO
SAIT LA COMMUNITY CORP
SAIT LA COMMUNITY CORP
CALIFORNIA A R FIRST FOUNDATION-IC
FRANK GOOD BROTHERS CENTER HOUSING DEVELOPMENT CORP
HERNANDEZ AND TRANSTON INC
STEPHEN J. POTTS INC INC
SHAW GROUP
EMPLOYMENT-LOB-BENEFERS INC
HEPBY JAVANA
RICHARD A. SOTTI INC
HIGHLAND PARK GUEST HOME INC
HIGHLAND PARK GUEST HOME INC
APPROXIMAD BINDER PRODUCTS LLC
STRATEGIC CONCEPTS IN ORGANIZING AND POLICY EDUCATION-IC
INTERNATIONAL RIGHT OF WAY ASSOCIATION
DOUGLE BROTHERS
DELBERT TRIMMUS FOUNDATION-IC
JAUZES AND ASSOCIATES INC

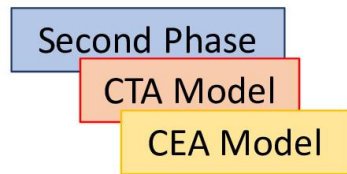
BUSINESS NAME
PALACE OF VENICE GUEST HOME-IC
UNICREDIT LABELLA
MILANO ECONOMIC DEVELOPMENT CORPORATION
CARLOS ANGEL
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
SPRINGFIELD NON-PROFIT CONSUMER CREDIT MANAGEMENT
ARMANDO FREZZO
SAIT LA COMMUNITY CORP
SAIT LA COMMUNITY CORP
CALIFORNIA A R FIRST FOUNDATION-IC
FRANK GOOD BROTHERS CENTER HOUSING DEVELOPMENT CORP
HERNANDEZ AND TRANSTON INC
STEPHEN J. POTTS INC INC
SHAW GROUP
EMPLOYMENT-LOB-BENEFERS INC
HEPBY JAVANA
RICHARD A. SOTTI INC
HIGHLAND PARK GUEST HOME INC
HIGHLAND PARK GUEST HOME INC
APPROXIMAD BINDER PRODUCTS LLC
STRATEGIC CONCEPTS IN ORGANIZING AND POLICY EDUCATION-IC
INTERNATIONAL RIGHT OF WAY ASSOCIATION
DOUGLE BROTHERS
DELBERT TRIMMUS FOUNDATION-IC
JAUZES AND ASSOCIATES INC

BUSINESS NAME
PALACE OF VENICE GUEST HOME-IC
UNICREDIT LABELLA
MILANO ECONOMIC DEVELOPMENT CORPORATION
CARLOS ANGEL
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
LA OFICINA CENTRAL, HERMANA DE LOS ANGELES-IC
SPRINGFIELD NON-PROFIT CONSUMER CREDIT MANAGEMENT
ARMANDO FREZZO
SAIT LA COMMUNITY CORP
SAIT LA COMMUNITY CORP
CALIFORNIA A R FIRST FOUNDATION-IC
FRANK GOOD BROTHERS CENTER HOUSING DEVELOPMENT CORP
HERNANDEZ AND TRANSTON INC
STEPHEN J. POTTS INC INC
SHAW GROUP
EMPLOYMENT-LOB-BENEFERS INC
HEPBY JAVANA
RICHARD A. SOTTI INC
HIGHLAND PARK GUEST HOME INC
HIGHLAND PARK GUEST HOME INC
APPROXIMAD BINDER PRODUCTS LLC
STRATEGIC CONCEPTS IN ORGANIZING AND POLICY EDUCATION-IC
INTERNATIONAL RIGHT OF WAY ASSOCIATION
DOUGLE BROTHERS
DELBERT TRIMMUS FOUNDATION-IC
JAUZES AND ASSOCIATES INC

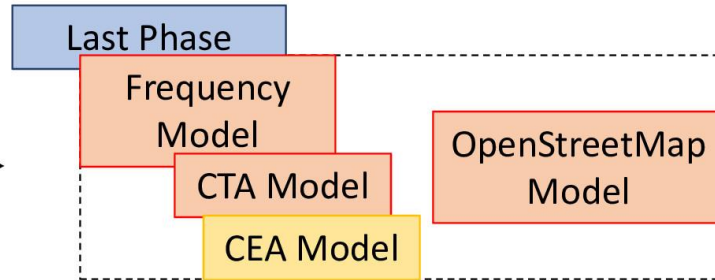


OUTPUT

Wikidata Classes

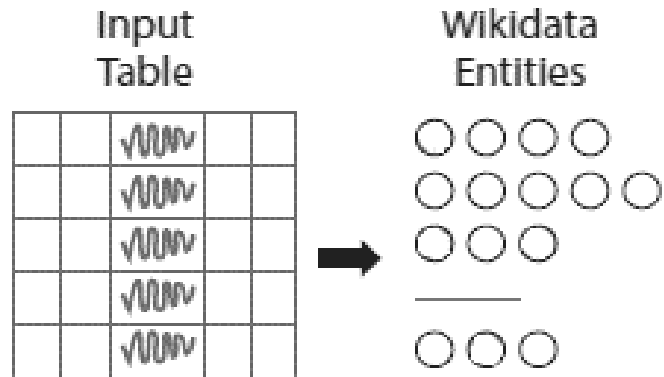


DBpedia Classes



DBpedia Classes

First Phase: Wikidata Model



items	Q13220204	Q13360155	Q13410400	Q1496967
Autauga	1.0	1.0	1.0	5.0
Baldwin	2.0	2.0	2.0	3.0
Barbour	2.0	2.0	1.0	2.0
Bibb	2.0	2.0	1.0	2.0
Blount	2.0	2.0	1.0	3.0
Bullock	1.0	1.0	1.0	1.0
Butler	0.0	0.0	0.0	3.0
Calhoun	2.0	2.0	1.0	4.0

RESULTS

- Meaningless classes in Wikidata
- Correctness in the generation of the candidates

DATASET	OUTPUT	CLASS EXPECTED
Cultural Event	Interaction	Event
Department of Recreation	Publication	Park
Education Facilities	Subject	Educational Institution

The Model Development

INPUT

BUSINESS NAME
PHILADELPHIA SELECT HOME CO
UNIONED LABELLA
WELLS ECONOMIC DEVELOPMENT CORPORATION
ORCA DE ANGEL
A L'OPERA CENTRAL HERMAN DE LOS ANGELES CO
A L'OPERA CENTRAL HERMAN DE LOS ANGELES CO
UNIONED REFINANCED METALL INC
A L'OPERA CENTRAL HERMAN DE LOS ANGELES CO
SPRINGFIELD NON PROFIT CONSUMER CREDIT MANAGEMENT
AMERICA PRESS
FAST LA COMMUNITY CORP
FAST LA COMMUNITY CORP
CALIFORNIA A L'ABRI FOUNDATION CO
PAINE GOOD SHEPHERD CENTER HOUSING DEVELOPMENT CORP
HERSCHE AND TRAXTER INC
STERNEN J POTHER GOO INC
DAVID DASH
EMPLOYMENT JOB SEEKERS INC
HEERY ZIGMON
HOWARD A GOTTFRID
HIGHLAND PARK QUEST HOME INC
HIGHLAND PARK QUEST HOME INC
WELLS ECONOMIC DEVELOPMENT CORPORATION
SPRINGFIELD CONCEPTS IN FOUNDATION AND POLICY EDUCATION CO
INTERNATIONAL RIGHT OF WAY ASSOCIATION
DOUGL E RICHIE
REINDEY TRUST MUSIC FOUNDATION CO
JARRIEZ AND ASSOCIATES INC

BUSINESS NAME
PHILADELPHIA SELECT HOME CO
UNIONED LABELLA
WELLS ECONOMIC DEVELOPMENT CORPORATION
ORCA DE ANGEL
A L'OPERA CENTRAL HERMAN DE LOS ANGELES CO
A L'OPERA CENTRAL HERMAN DE LOS ANGELES CO
UNIONED REFINANCED METALL INC
A L'OPERA CENTRAL HERMAN DE LOS ANGELES CO
SPRINGFIELD NON PROFIT CONSUMER CREDIT MANAGEMENT
AMERICA PRESS
FAST LA COMMUNITY CORP
FAST LA COMMUNITY CORP
CALIFORNIA A L'ABRI FOUNDATION CO
PAINE GOOD SHEPHERD CENTER HOUSING DEVELOPMENT CORP
HERSCHE AND TRAXTER INC
STERNEN J POTHER GOO INC
DAVID DASH
EMPLOYMENT JOB SEEKERS INC
HEERY ZIGMON
HOWARD A GOTTFRID
HIGHLAND PARK QUEST HOME INC
HIGHLAND PARK QUEST HOME INC
WELLS ECONOMIC DEVELOPMENT CORPORATION
SPRINGFIELD CONCEPTS IN FOUNDATION AND POLICY EDUCATION CO
INTERNATIONAL RIGHT OF WAY ASSOCIATION
DOUGL E RICHIE
REINDEY TRUST MUSIC FOUNDATION CO
JARRIEZ AND ASSOCIATES INC

BUSINESS NAME
PHILADELPHIA SELECT HOME CO
UNIONED LABELLA
WELLS ECONOMIC DEVELOPMENT CORPORATION
ORCA DE ANGEL
A L'OPERA CENTRAL HERMAN DE LOS ANGELES CO
A L'OPERA CENTRAL HERMAN DE LOS ANGELES CO
UNIONED REFINANCED METALL INC
A L'OPERA CENTRAL HERMAN DE LOS ANGELES CO
SPRINGFIELD NON PROFIT CONSUMER CREDIT MANAGEMENT
AMERICA PRESS
FAST LA COMMUNITY CORP
FAST LA COMMUNITY CORP
CALIFORNIA A L'ABRI FOUNDATION CO
PAINE GOOD SHEPHERD CENTER HOUSING DEVELOPMENT CORP
HERSCHE AND TRAXTER INC
STERNEN J POTHER GOO INC
DAVID DASH
EMPLOYMENT JOB SEEKERS INC
HEERY ZIGMON
HOWARD A GOTTFRID
HIGHLAND PARK QUEST HOME INC
HIGHLAND PARK QUEST HOME INC
WELLS ECONOMIC DEVELOPMENT CORPORATION
SPRINGFIELD CONCEPTS IN FOUNDATION AND POLICY EDUCATION CO
INTERNATIONAL RIGHT OF WAY ASSOCIATION
DOUGL E RICHIE
REINDEY TRUST MUSIC FOUNDATION CO
JARRIEZ AND ASSOCIATES INC

First Phase

First Model

OUTPUT

Wikidata Classes

Second Phase

CTA Model

CEA Model

DBpedia Classes



(Thawani et al. 2019)

Last Phase

Frequency Model

CTA Model

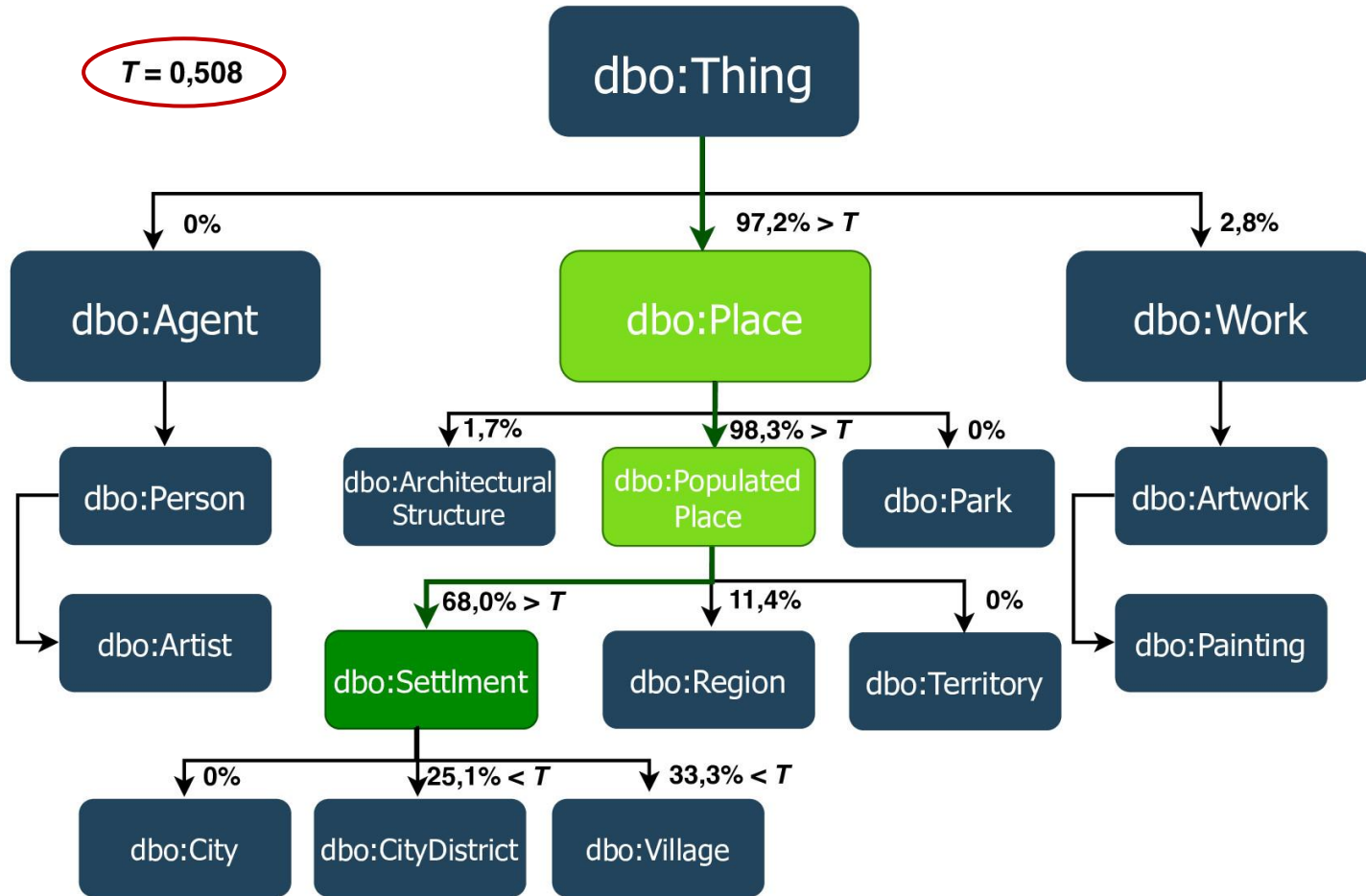
CEA Model

OpenStreetMap Model

DBpedia Classes



Second Phase: the Column Type Annotation Model



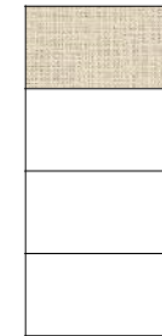
(Vembunarayanan, 2013)

Semantic Feature



Semantic Coherence among cells in a column

Lexical Feature



Lexical Similarity

Knowledge Graph

Four Configurations

- Levenshtein Multiplied by TFIDF
- Weighted Average between Levenshtein and TFIDF
- Jaro-Winkler Multiplied by TFIDF
- Weighted Average between Jaro-Winkler and TFIDF

Results Evaluation

$$Accuracy_i = \frac{n_i}{D}$$

- n is the number of correct classes detected by the methodology i applied
- D is the total number of datasets evaluated (41)

DATASET	OUTPUT	CLASS EXPECTED
Cultural Centers	Venue	Venue
Cultural Events	Museum	Event
Education Facilities	Organisation	Educational Institution

MEASURE	Levenshtein multiplied by the TFIDF
Permissive Accuracy	0,36585
Restrictive Accuracy	0,31707

The Model Development

INPUT

BUSINESS NAME
PRINCE OF WILHELM SELECT HOME CO
UNICREDIT LABELLA
WELDON ECONOMIC DEVELOPMENT CORPORATION
CAROL ANSEL
A-L-SPRINK CENTRAL HERMAN DE LOS ANGELES CO
A-L-SPRINK CENTRAL HERMAN DE LOS ANGELES CO
UNION PACIFIC DEVELOPMENT METALL INC
A-L-SPRINK CENTRAL HERMAN DE LOS ANGELES CO
SPRINGFIELD NON PROFIT CONSUMER CREDIT MANAGEMENT
AMBERLY PRESS
FAST LA COMMUNITY CORP
FAST LA COMMUNITY CORP
CALIFORNIA E-A-STATE FOUNDATION CO
PARK GOOD IMPROVEMENT CENTER HOUSING DEVELOPMENT CORP
HERSCHE AND TRAXTER INC
STERNEN J-POWER-USA INC
DAVID PERE
EMPLOYMENT JOB SEEKERS INC
HEIDI ZIGMUND
RICHARD A GOTTFRIED
HIGHLAND PARK QUEST HOME INC
HIGHLAND PARK QUEST HOME INC
WELDON ECONOMIC DEVELOPMENT CORPORATION
SPRINGFIELD CONCEPTS IN HOUSING AND POLICY EDUCATION CO
INTERNATIONAL RIGHT OF WAY ASSOCIATION
DOUGLE E RICHARDS
REINOLDY TRIMM MUSIC FOUNDATION CO
JARRINE AND ASSOCIATES INC

BUSINESS NAME
PRINCE OF WILHELM SELECT HOME CO
UNICREDIT LABELLA
WELDON ECONOMIC DEVELOPMENT CORPORATION
CAROL ANSEL
A-L-SPRINK CENTRAL HERMAN DE LOS ANGELES CO
A-L-SPRINK CENTRAL HERMAN DE LOS ANGELES CO
UNION PACIFIC DEVELOPMENT METALL INC
A-L-SPRINK CENTRAL HERMAN DE LOS ANGELES CO
SPRINGFIELD NON PROFIT CONSUMER CREDIT MANAGEMENT
AMBERLY PRESS
FAST LA COMMUNITY CORP
FAST LA COMMUNITY CORP
CALIFORNIA E-A-STATE FOUNDATION CO
PARK GOOD IMPROVEMENT CENTER HOUSING DEVELOPMENT CORP
HERSCHE AND TRAXTER INC
STERNEN J-POWER-USA INC
DAVID PERE
EMPLOYMENT JOB SEEKERS INC
HEIDI ZIGMUND
RICHARD A GOTTFRIED
HIGHLAND PARK QUEST HOME INC
HIGHLAND PARK QUEST HOME INC
WELDON ECONOMIC DEVELOPMENT CORPORATION
SPRINGFIELD CONCEPTS IN HOUSING AND POLICY EDUCATION CO
INTERNATIONAL RIGHT OF WAY ASSOCIATION
DOUGLE E RICHARDS
REINOLDY TRIMM MUSIC FOUNDATION CO
JARRINE AND ASSOCIATES INC

BUSINESS NAME
PRINCE OF WILHELM SELECT HOME CO
UNICREDIT LABELLA
WELDON ECONOMIC DEVELOPMENT CORPORATION
CAROL ANSEL
A-L-SPRINK CENTRAL HERMAN DE LOS ANGELES CO
A-L-SPRINK CENTRAL HERMAN DE LOS ANGELES CO
UNION PACIFIC DEVELOPMENT METALL INC
A-L-SPRINK CENTRAL HERMAN DE LOS ANGELES CO
SPRINGFIELD NON PROFIT CONSUMER CREDIT MANAGEMENT
AMBERLY PRESS
FAST LA COMMUNITY CORP
FAST LA COMMUNITY CORP
CALIFORNIA E-A-STATE FOUNDATION CO
PARK GOOD IMPROVEMENT CENTER HOUSING DEVELOPMENT CORP
HERSCHE AND TRAXTER INC
STERNEN J-POWER-USA INC
DAVID PERE
EMPLOYMENT JOB SEEKERS INC
HEIDI ZIGMUND
RICHARD A GOTTFRIED
HIGHLAND PARK QUEST HOME INC
HIGHLAND PARK QUEST HOME INC
WELDON ECONOMIC DEVELOPMENT CORPORATION
SPRINGFIELD CONCEPTS IN HOUSING AND POLICY EDUCATION CO
INTERNATIONAL RIGHT OF WAY ASSOCIATION
DOUGLE E RICHARDS
REINOLDY TRIMM MUSIC FOUNDATION CO
JARRINE AND ASSOCIATES INC

OUTPUT

Wikidata Classes

DBpedia Classes

DBpedia Classes

First Phase

First Model

Second Phase

CTA Model

CEA Model

Last Phase

Frequency Model

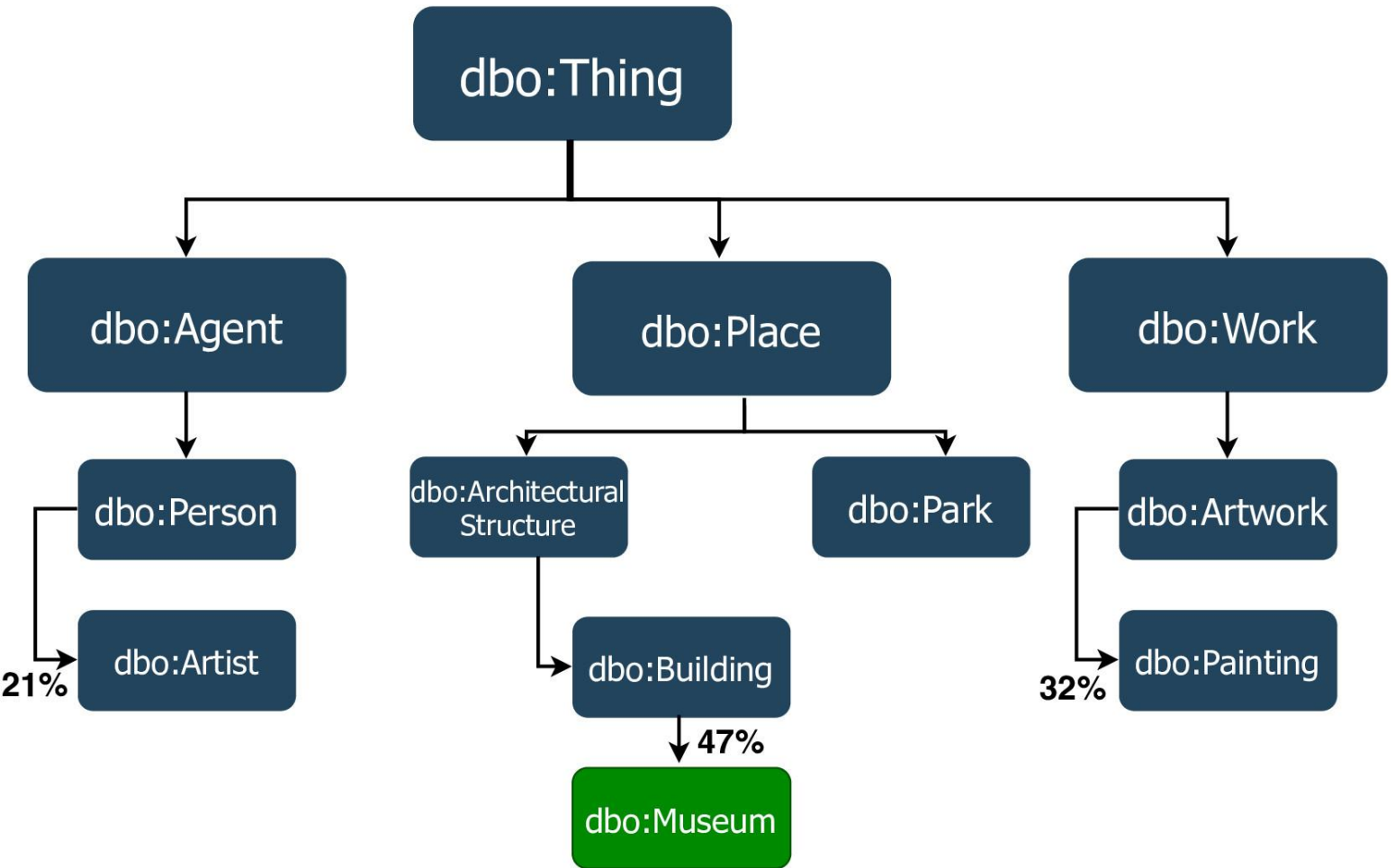
CTA Model

CEA Model

OpenStreetMap Model

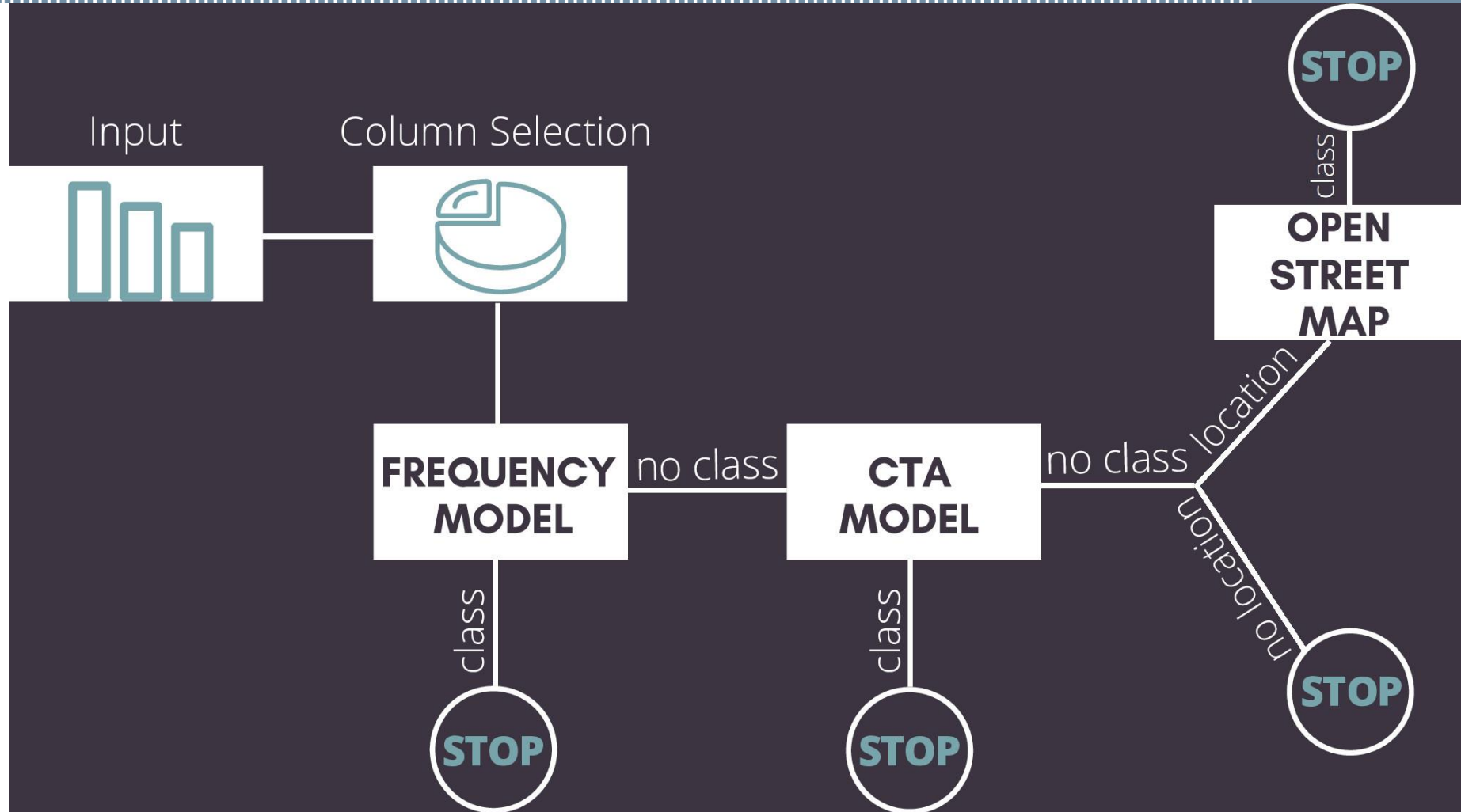


Last Phase: Final Model Development



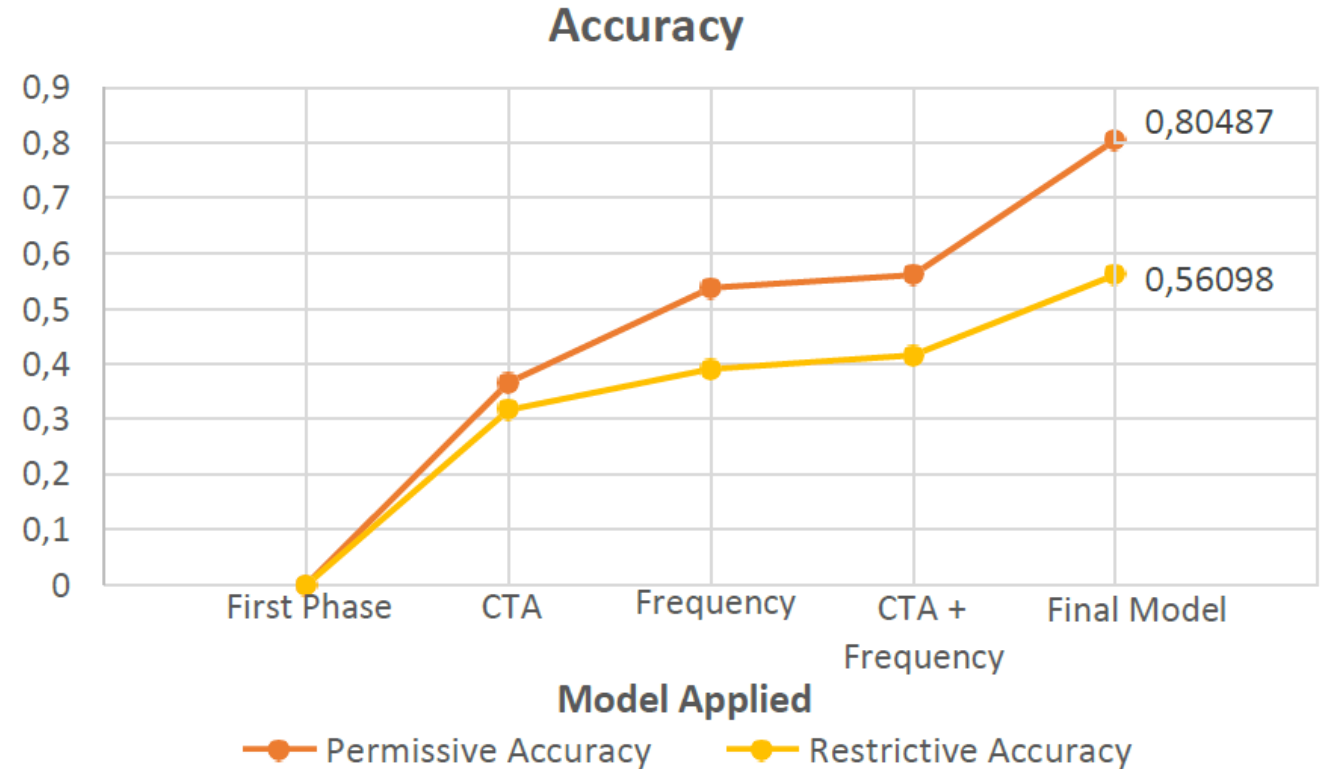
MEASURE	CTA	Frequency
Permissive Accuracy	0,36585	0,53658
Restrictive Accuracy	0,31707	0,39024

Final Model Design



Performance Evaluations

MEASURE	CTA + Frequency	Final Model
Permissive Accuracy	0,56098	0,80487
Restrictive Accuracy	0,41463	0,56098



Conclusions

Exploitation of Web ontologies



A good reasoning capacity by integrating different approaches



Seeking for efficiency



Improve candidates' generation phase

(Manning et. al, 2008)



Apply more advanced methodologies for column selection

(Pham et. al, 2016)



Introduce more sophisticated approaches for Frequency model



POLITECNICO
MILANO 1863

Thank you for the attention!

Any Questions?