DATA ANALYTICS USING DEEP LEARNING

GT 8803 // VENKATA KISHORE PATCHA

LECTURE #06:

SMELLY RELATIONS: MEASURING AND UNDERSTANDING DATABASE SCHEMA QUALITY

CREATING THE NEXT®

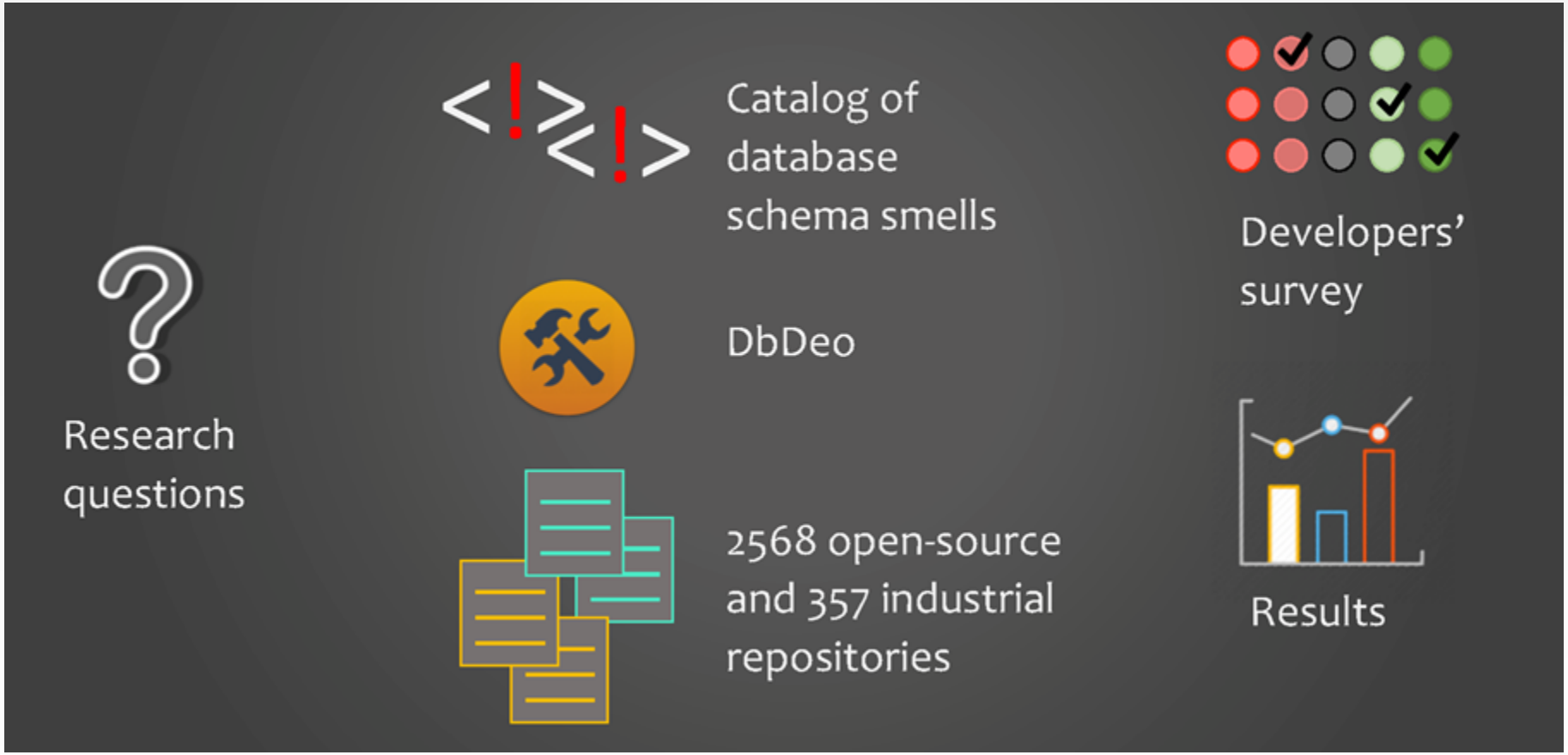
TODAY'S PAPER

- **Smelly Relations: Measuring and Understanding Database Schema Quality**
  - **Authors:**
    - Tushar Sharma, Marios Fragkoulis , Diomidis Spinellis
      - affiliated with Athens University of Economics and Business, Athens, Greece
    - Stamatia Rizou
      - Affiliated with Singular Logic Athens, Greece
    - Magiel Bruntink
      - Affiliated with Software Improvement Group Amsterdam, The Netherlands
  - **Areas of focus:**
    - Data Base Schema; Software Development and quality.
  - Slides based on a presentation by Tushar Sharma @ ICSE 2018 * SEIP

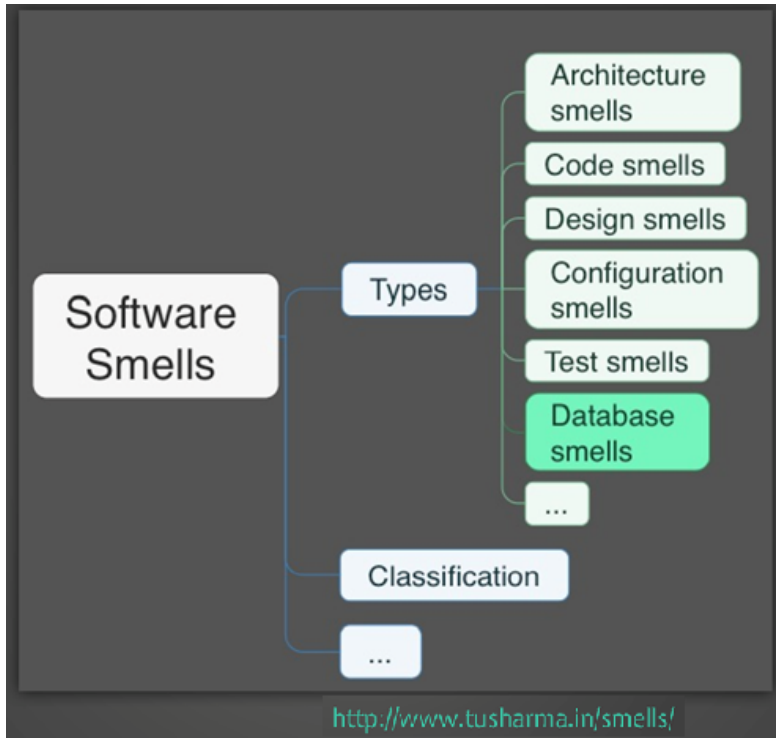
TODAY'S AGENDA

- Study Overview
- Context: Background Info on Relevant Concepts
- Key Idea
- Technical Details
- Experiments
- Discussion Questions

STUDY OVERVIEW



CONTEXT: SOFTWARE SMELLS



- **certain structures** in the **code** that **suggest(sometimes they scream for) the possibility of refactoring**. - Kent Beck

CONTEXT: DATA BASE SMELLS

- Not following the recommended best practices and potentially affecting the quality of the software system in a negative way.

CONTEXT: CLASSIFICATION OF DB SMELLS

- Schema smells – The paper is about this.
- Query smells - Smells arising from poorly written sql queries are specified as database query smells.
- Data smells – Poor data. Example: typos

CONTEXT: CATALOG

1. Compound attribute – Comma separated list
2. Adjacency list - recursive relation in a table.
3. Superfluous key – Unwanted Surrogate key. Dup validation
4. Missing constraints - foreign keys are missing
5. Metadata as data – Key value pairs

CONTEXT: CATALOG

6. Polymorphic association – SQL don't allow two fk. Don't force

Person

CustID	Name	--
4	Dave	
9	Tom	

Business

CustID	Company Name	--
4	Coco	
5	Times	

OrderID	CustType	CustID
4	Person	4
5	Business	9

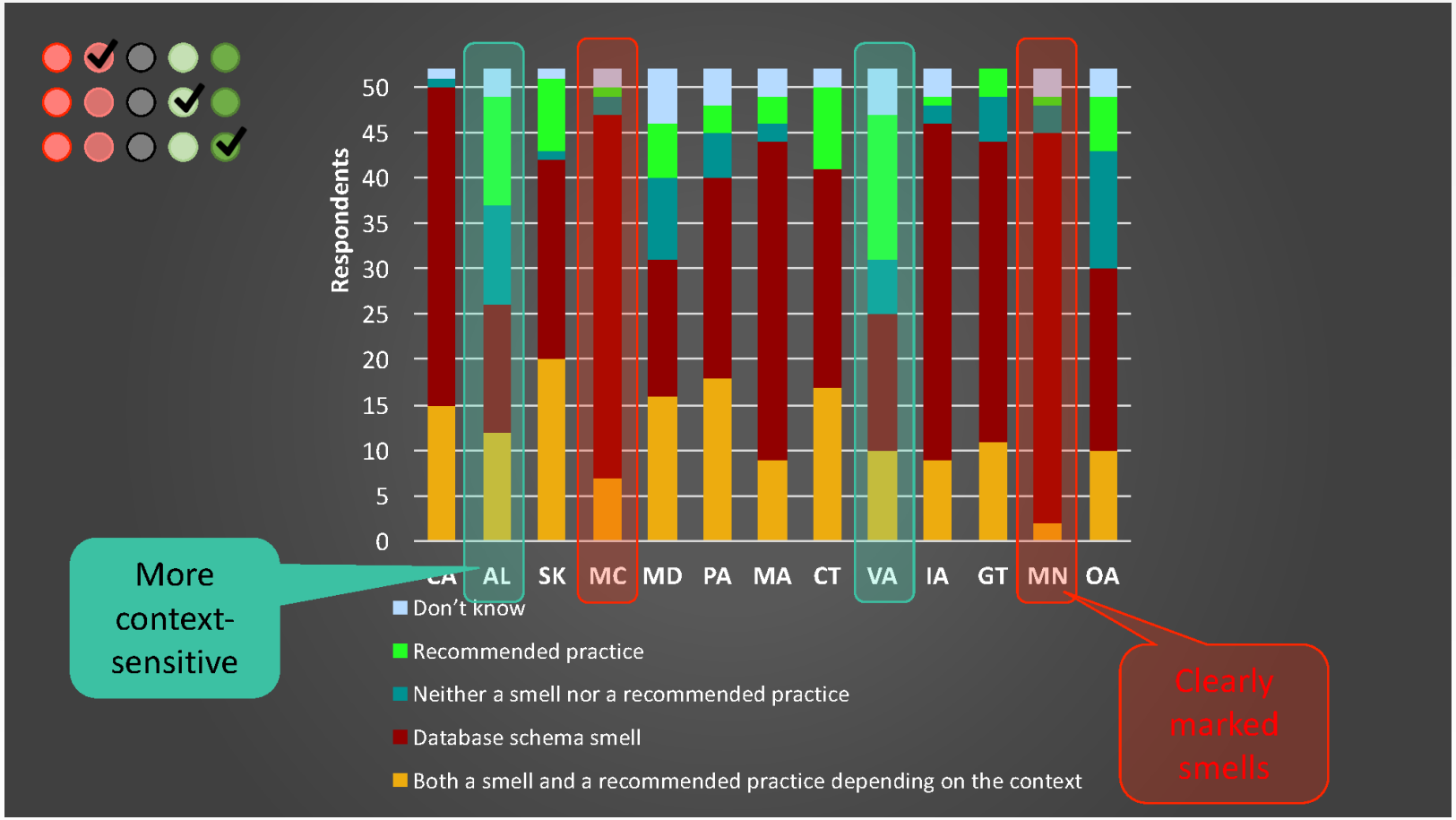
CONTEXT: CATALOG

7. Multicolumn attribute – Tag1, Tag2 and so on
8. Clone table – Orders2017, Orders2010
9. Values in attribute definition – Choice/check list in schema
10. Index abuse – Over or under use
11. God table – Anti-Normalization
12. Meaningless name
13. Overload attribute names – Attributes have similar names but different type in different tables. Example ID.

KEY IDEA

- **Objective: Developers opinion on DB Schema smells. Collect code from industry & OSS and answer RQs.**
  - What are the occurrence patterns of database smells?
  - Does the size of the project or the database play a role in smell density?
  - Does the nature of code (type of the application, or usage of ORM frameworks) affect the smell density?
  - What is the degree of co-occurrence among database smells?
- DbDeo – An open-source tool to
    - extract embedded SQL statements and
    - detect database schema smells

TECHNICAL DETAILS - SURVEY



TECHNICAL DETAILS - DBDEO

- 9 smells are automated.
- Compound attribute: Look for pattern-matching expressions in an sql query
- Adjacency list: We look for a foreign key constraint referring to an attribute in the same table.
- Metadata as data: look for a schema definition containing only three attributes. We detect the smell if we find two of the attributes, among three, of type varchar

TECHNICAL DETAILS - DBDEO

- Multicolumn attribute: Check the schema for a pattern "N where N is a number
- Clone tables: Check all the schema definitions within a database
- Values in attribute definition: check the schema for "enum" or "check"

TECHNICAL DETAILS - DBDEO

Index abuse:

- Missing indexes: 0 indexes in schema
- Insufficient indexes: Missing index for FK
- Unused indexes: Indexed column is not present in where clause

.

TECHNICAL DETAILS - DBDEO

- God table: More than 10 columns in a table.
- Overloaded attribute names: Same column name found in different tables but with different datatype.

RQ1. OCCURRENCE PATTERNS OF DATABASE SMELLS

Smells	Occurrences		Avg. smell density	
	I	OSS	I	OSS
CA	5,517	7,966	0.04	0.04
AL	733	297	0.15	0.02
GT	4,428	5,507	0.44	0.24
VA	85	326	0.00	0.02
MD	944	1,003	0.16	0.09
MA	1,624	3,137	0.10	0.07
CT	101	3,704	0.00	0.05
OA	1,814	7,300	0.20	0.21
IA	12,643	9,475	1.25	1.76

Most frequently occurring smell

Index abuse

RQ1. OCCURRENCE PATTERNS OF DATABASE SMELLS

Smells	Occurrences		Avg. smell density	
	I	OSS	I	OSS
CA	5,517	7,966	0.04	0.04
AL	733	297	0.15	0.02
GT	4,428	5,507	0.44	0.[illegible]
VA	85	326	0.00	0.0[illegible]
MD	944	1,003	0.16	0.09
[illegible]	1,624	3,137	0.10	0.07
CT	101	3,704	0.00	0.05
OA	1,814	7,300	0.20	0.21
IA	12,643	9,475	1.25	1.76

Adjacency list prone to occur more in industrial projects

OSS projects report more *Clone table*

RQ2. DOES THE SIZE OF THE PROJECT OR THE DATABASE PLAY A ROLE IN SMELL DENSITY?

Smell density

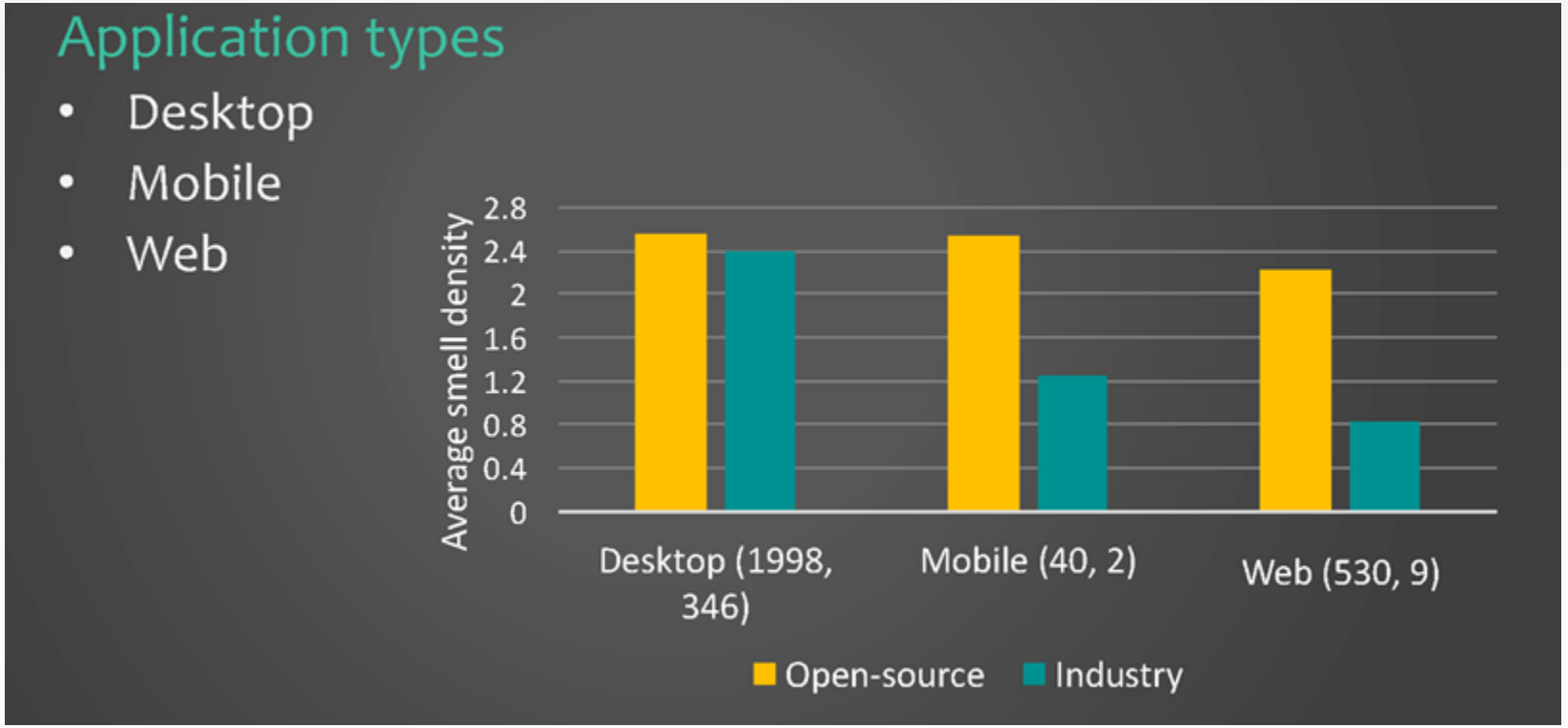
- Number of database smells per ten SQL statements

- LOC vs smell density
  - $\rho = 0.2420$ (p-value = 3.724×10^{-6}) for Industry
  - $\rho = 0.0006$ (p-value = 0.9731) fo

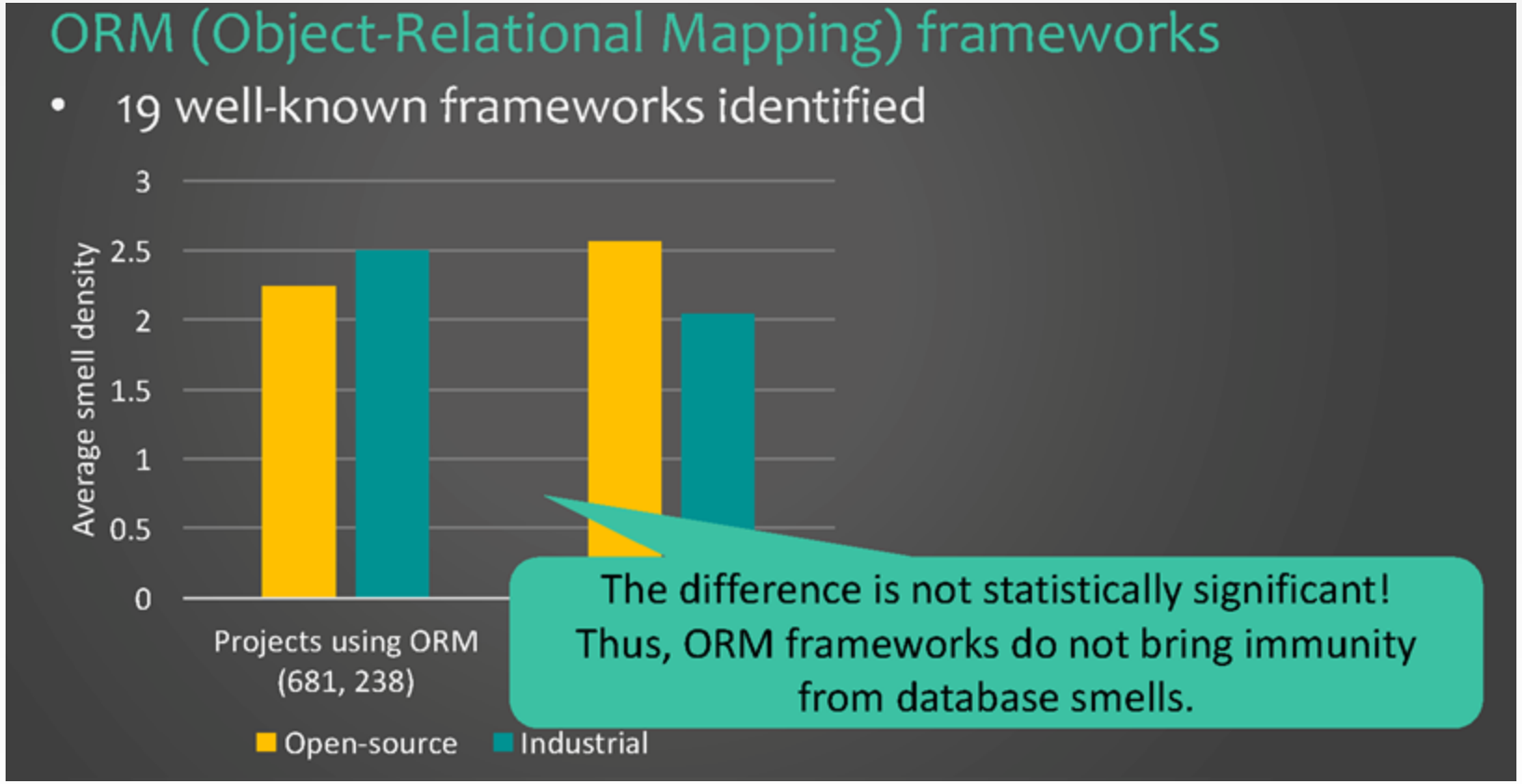
Strong correlation between database size and smell density.

- Database size vs smell density
  - $\rho = 0.7338$ (p-value < 2.2×10^{-16}) for Industry
  - $\rho = 0.6174$ (p-value < 2.2×10^{-16}) for OSS

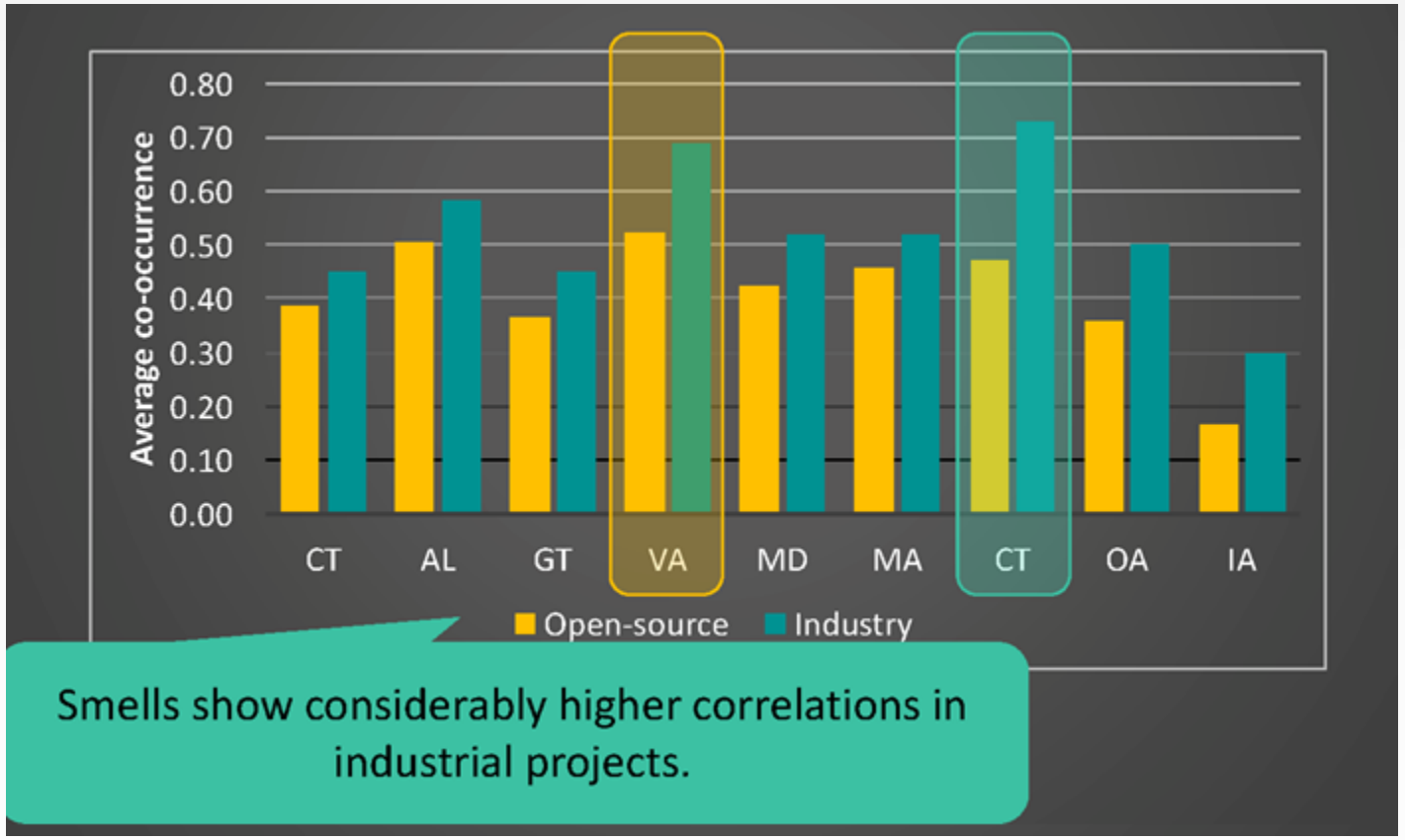
RQ3.
DOES THE NATURE OF CODE (TYPE OF THE APPLICATION, OR USAGE OF ORM FRAMEWORKS) AFFECT THE SMELL DENSITY?



RQ3.
DOES THE NATURE OF CODE (TYPE OF THE APPLICATION, OR USAGE OF ORM FRAMEWORKS) AFFECT THE SMELL DENSITY?



RQ4. WHAT IS THE DEGREE OF CO-OCCURRENCE AMONG DATABASE SMELLS?



DISCUSSION QUESTIONS

- What are key strengths of this approach?
- What are key weaknesses/limitations?
- How could this DbDeo be modified to capture more smells and/or with better accuracy?
- Can Schema be fixed automatically?

BIBLIOGRAPHY

- Bill Karwin. 2010. SQL Antipatterns: Avoiding the Pitfalls of Database Programming (1st ed.). Pragmatic Bookshelf