LearnI— learning data integration

Gaël Varoquaux

Inria

From discrete entities to signals

Outline

1 Application context

- Social and health studies
- Dirty Data

Assembling health or social databases

- Prevalence of a pathology as a function of age?
- Prognosis: predict individual evolution
- Identify risk factors / causal links

Studies across sites / databases
bring more general conclusions

- Ingesting each dataset requires “alignment”

  finding correspondences in the information

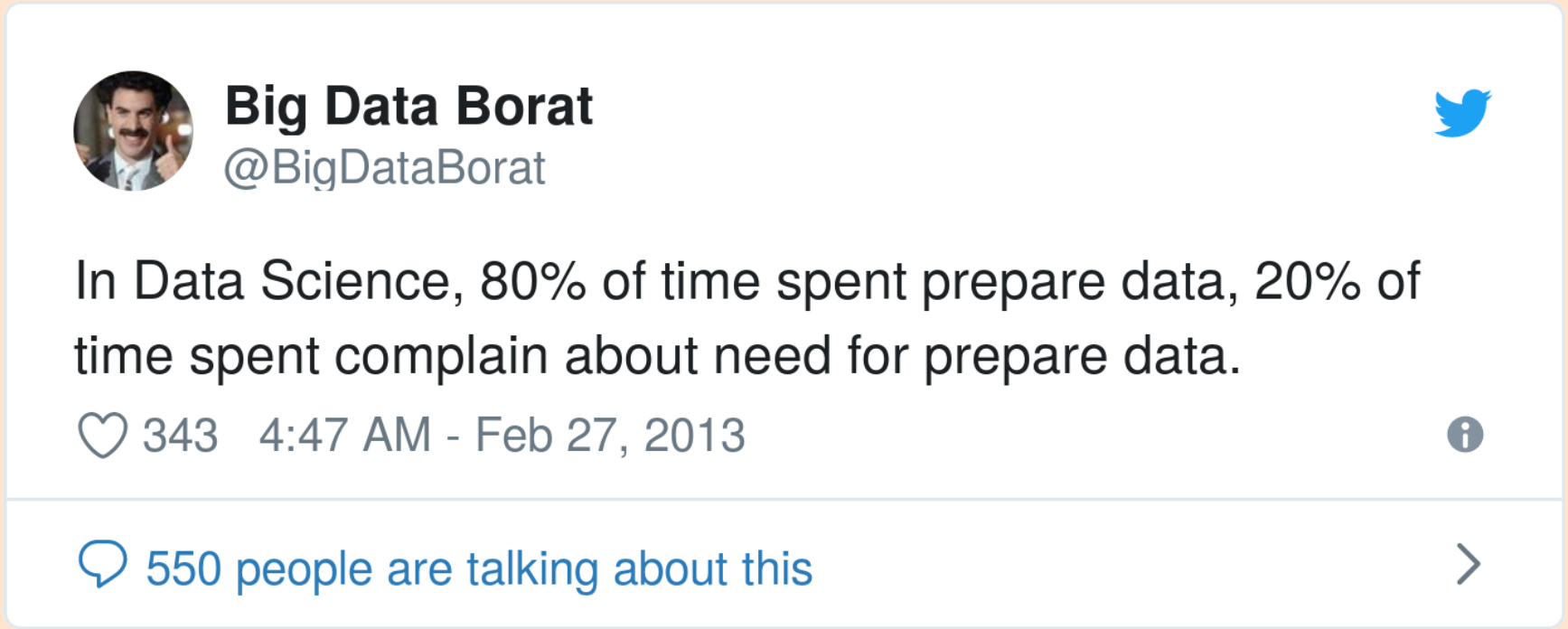
- Data must be “denormalized” for analysis

  joining / selecting / aggregating multiple tables into one

Data integration – a database-research topic

Data science and dirty data

We want to use statistics for meaningful answers



Statistics and dirty data

Statistical learning research: let $\mathbf{X} \in \mathbb{R}^{n \times p}$

Real-life data science:

Gender	Date Hired	Employee Position Title
M	09/12/1988	Master Police Officer
F	NA	Social Worker IV
M	07/16/2007	Police Officer III
F	02/05/2007	Police Aide
M	01/13/2014	Electrician I
M	04/28/2002	Bus Operator
M	NA	Bus Operator
F	06/26/2006	Social Worker III
F	01/26/2000	Library Assistant I
M	NA	Library Assistant I

Statistics and dirty data

Statistical learning research: let $\mathbf{X} \in \mathbb{R}^{n \times p}$

Real-life data science:

Missing values

Gender	Date Hired	Employee Position Title
	/1988	Master Police Officer
F	NA	Social Worker IV
M	07/16/2007	Police Officer III
F	02/05/2007	Police Aide
M	01/13/2014	Electrician I
M	04/28/2002	Bus Operator
M	NA	Bus Operator
F	06/26/2006	Social Worker III
F	01/26/2000	Library Assistant I
M	NA	Library Assistant I

Statistics and dirty data

Statistical learning research: let $\mathbf{X} \in \mathbb{R}^{n \times p}$

Real-life data science:

Missing values

Non normalized entries

Gender	Date Hired	Employee Position Title
	/1988	Master Police Officer
F	NA	Social Worker IV
	/2007	Police Officer III
F	02/05/2007	Police Aide
M	01/13/2014	Electrician I
M	04/28/2002	Bus Operator
M	NA	Bus Operator
F	06/26/2006	Social Worker III
F	01/26/2000	Library Assistant I
M	NA	Library Assistant I

Statistics and dirty data

Statistical learning research: let $\mathbf{X} \in \mathbb{R}^{n \times p}$

Real-life data science:

Missing values

Non normalized entries

Missing column names

Gender	Date Hired	Employee Position Title
	/1988	Master Police Officer
F	NA	Social Worker IV
	/2007	Police Officer III
	/2007	Police Aide
M	01/13/2014	Electrician I
M	04/28/2002	Bus Operator
M	NA	Bus Operator
F	06/26/2006	Social Worker III
F	01/26/2000	Library Assistant I
M	NA	Library Assistant I

Statistics and dirty data

Statistical learning research: let $\mathbf{X} \in \mathbb{R}^{n \times p}$

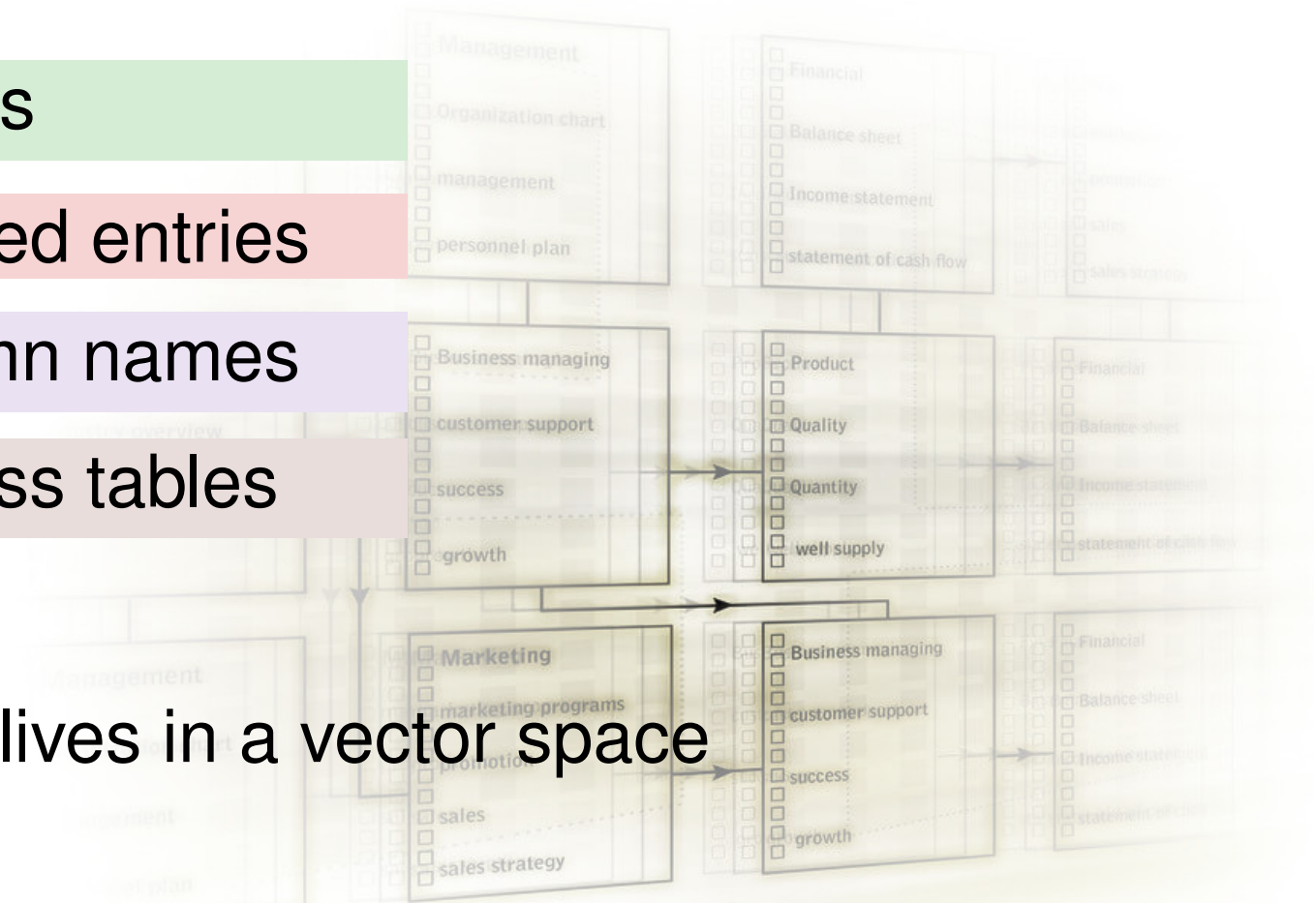
Real-life data science:

Missing values

Non normalized entries

Missing column names

Analysis across tables



None of this lives in a vector space

Databases and dirty data

Data of different nature are assembled

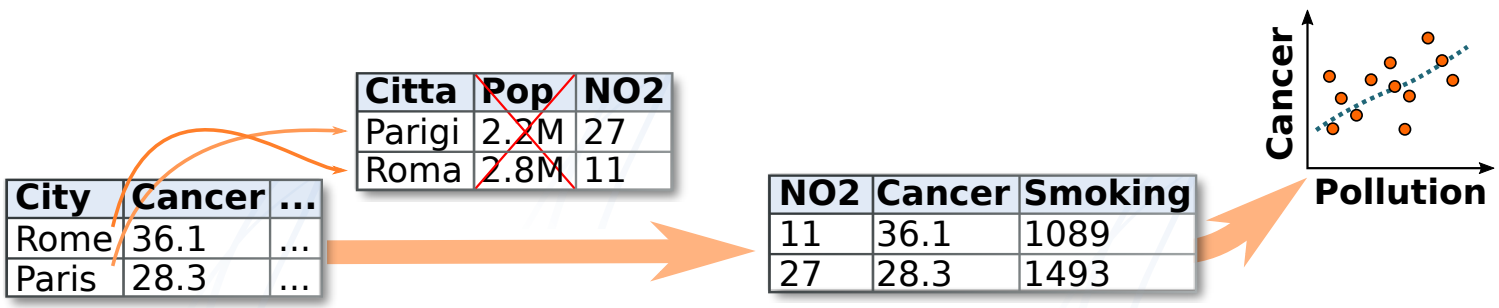
via common entities

eg patient characteristics and

pollution via location

Relational algebra

`union` `join` `aggregate...`



Breaks down with errors in symbols

Fundamental challenge

- Mix of numerical data & symbols with errors
- No topology / metrics on symbols

2 Project: Analysis on databases with embeddings

- Key intuitions
- Research axes

Overall strategy

≠ data cleaning

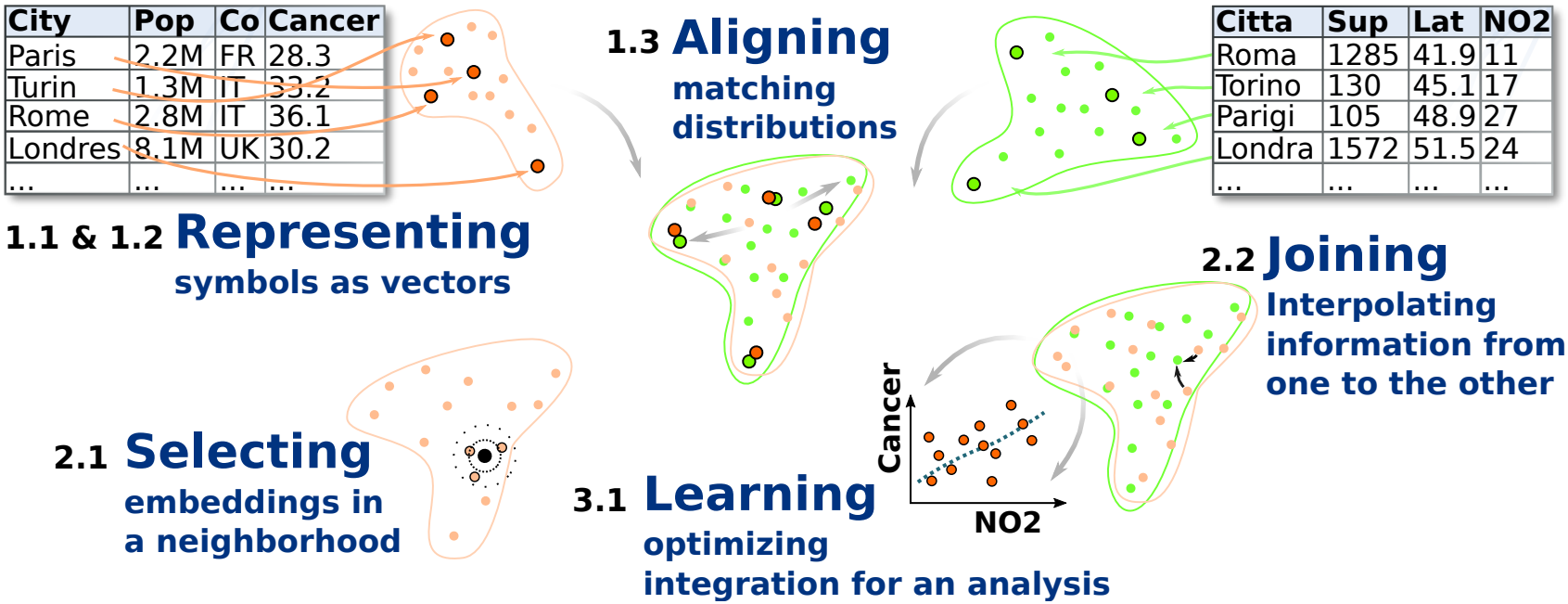
- Do not strive for a cleaned or assembled a data
- Rather, build a statistical model to answer analytical questions from the “dirty data”

♥ embeddings & continuous models

- Represent all discrete information in vector spaces
- Strive for statistical models fit with gradient descent

deep learning, differentiable programming

Axes



1 Representing:

from entity resolution to vector embedding

2 Transforming:

vector operations to assemble heterogeneous data

3 Learning:

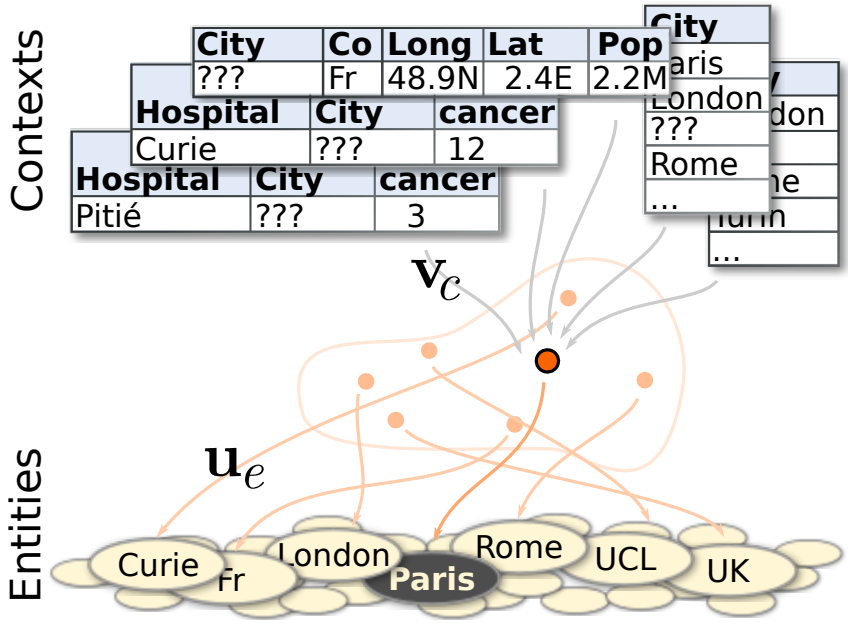
optimizing representations and operations

1 Representing

from entity resolution to vector embedding

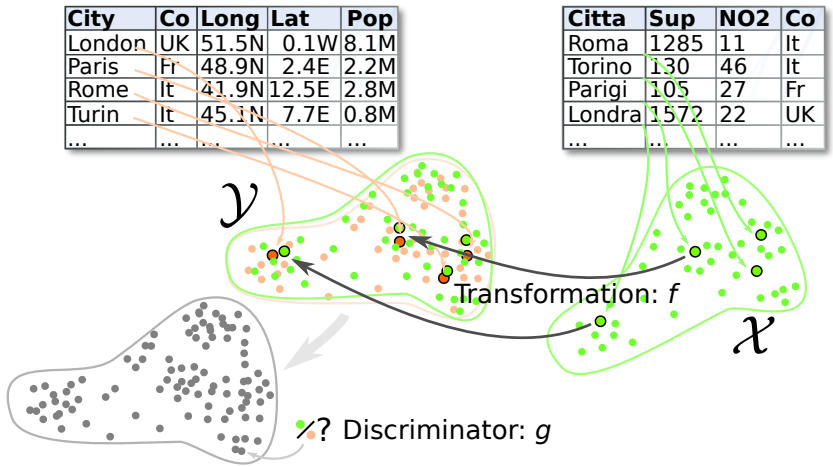
Embedding entities

from their context



Aligning entities

by matching distributions



2 Transforming

vector operations to assemble heterogeneous data

Select and aggregate

pattern matching

2.1 Selecting
embeddings in a neighborhood

Continuous joins

transfer of attributes

City	Pop	Co	Cancer
Paris	2.2M	FR	28.3
Turin	1.3M	IT	33.2
Rome	2.8M	IT	36.1
Londres	8.1M	UK	30.2
...	...	...	...

1.1 & 1.2 Representing
symbols as vectors

1.3 Aligning
matching distributions

Citta
Roma
Torino
Parigi
Londr
...

2.2 J
In

3 Learning

optimizing representations and operations

stochastic gradient descent & stochastic regularizations

Supervised learning

On a specific *type* of entity

eg patients/individuals

Transfer learning

Reuse representations and parts of architecture

4 Application tasks

enabling easy reuse of public and health data

Learning representations of public data

`data.gov`, `data.gouv.fr`...

many numerical dataset hard to reuse / integrate

Applications to opportunistic epidemiology

Understand population health with existing observational data

Electronic Health Records, AP-HP

3 Preliminary progress

String models of dirty categories

Vector representations capturing string regularities

String form of non-normalized entries often useful

Employee Position Title
Master Police Officer
Social Worker IV
Police Officer III
Police Aide
Electrician I
Bus Operator
Bus Operator
Social Worker III
Library Assistant I
Library Assistant I

[Cerda... 2018, Cerda and Varoquaux 2020]

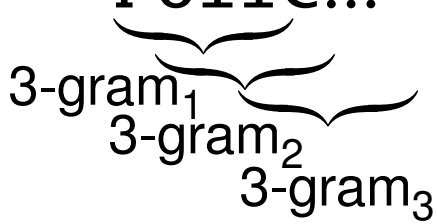
Vector representations capturing string regularities

String form of non-normalized entries often useful

Employee Position Title
Master Police Officer
Social Worker IV
Police Officer III
Police Aide
Electrician I
Bus Operator
Bus Operator
Social Worker III
Library Assistant I
Library Assistant I

Sub-string models

character-level n-grams

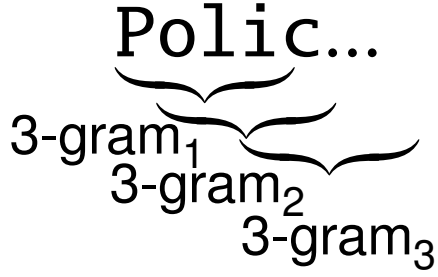


[Cerda... 2018, Cerda and Varoquaux 2020]

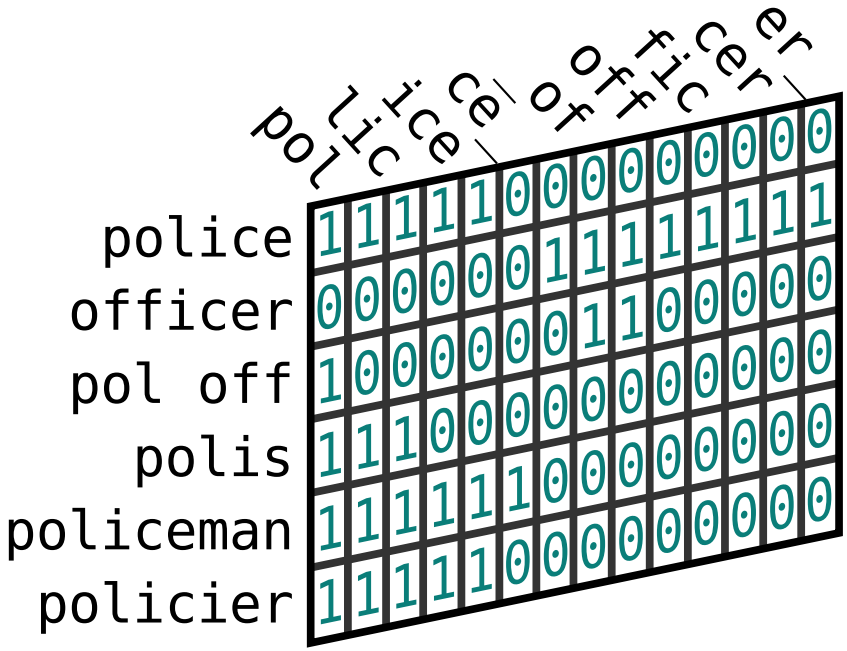
Vector representations capturing string regularities

Gamma-Poisson factorization

on sub-strings counts



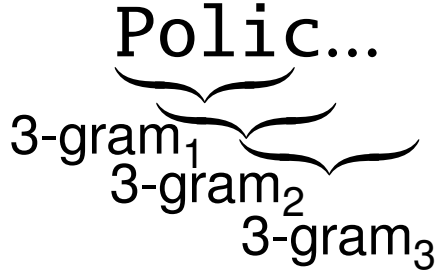
Models strings as a linear combination of substrings



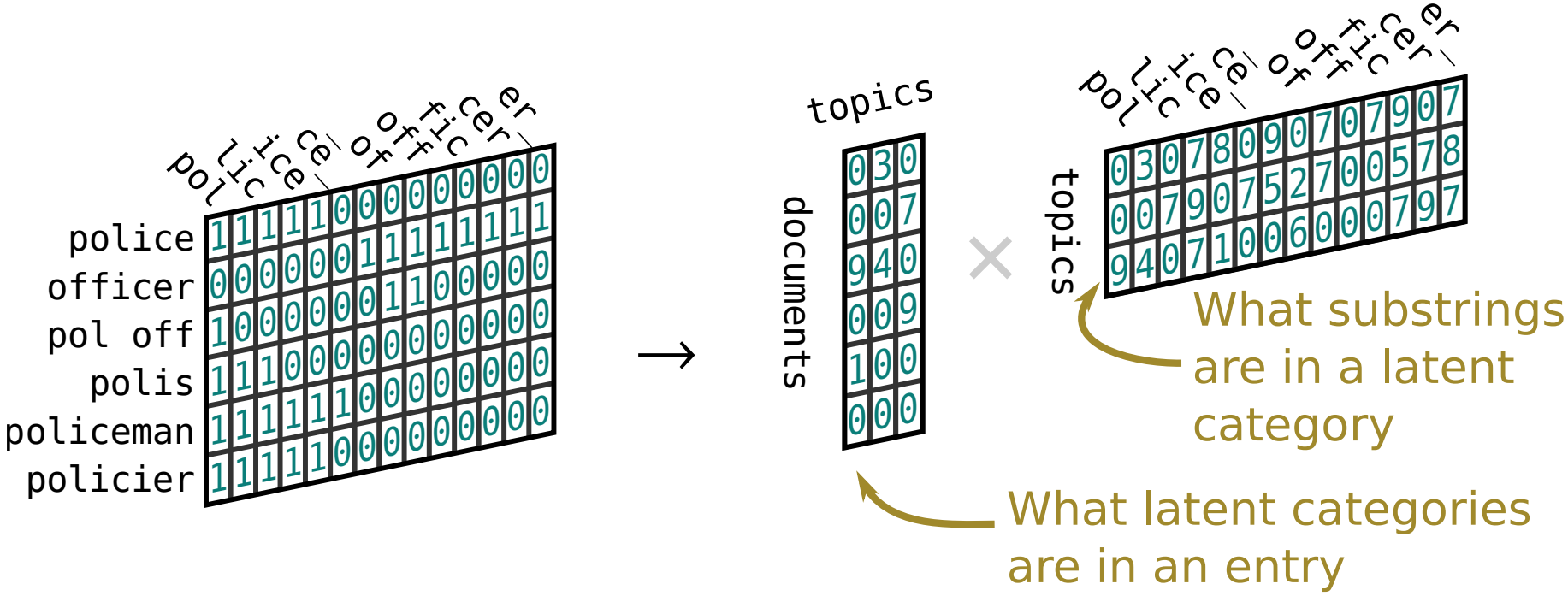
[Cerda and Varoquaux 2020]

Vector representations capturing string regularities

Gamma-Poisson factorization on sub-strings counts

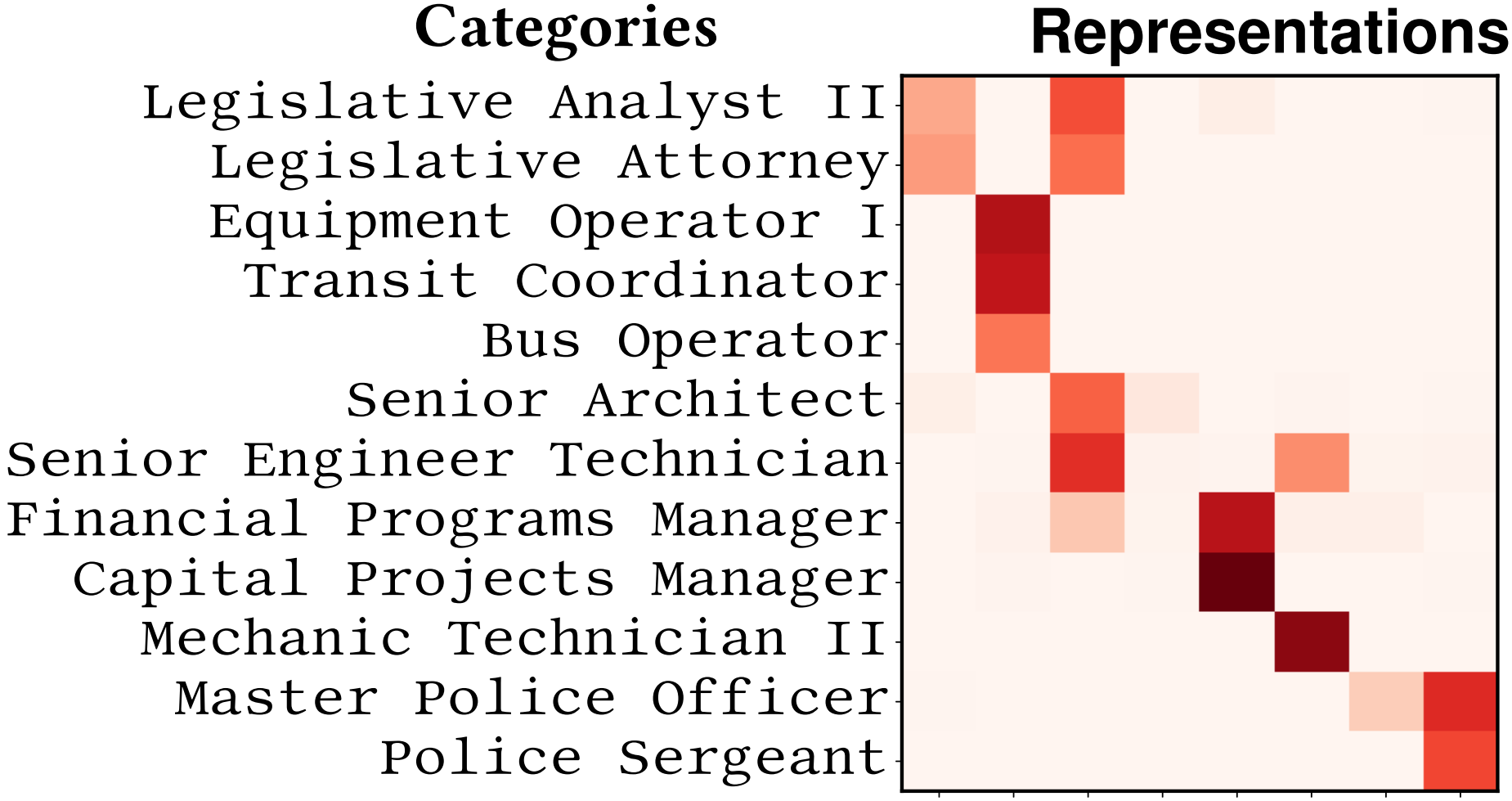


Models strings as a linear combination of substrings



[Cerda and Varoquaux 2020]

Vector representations capturing string regularities



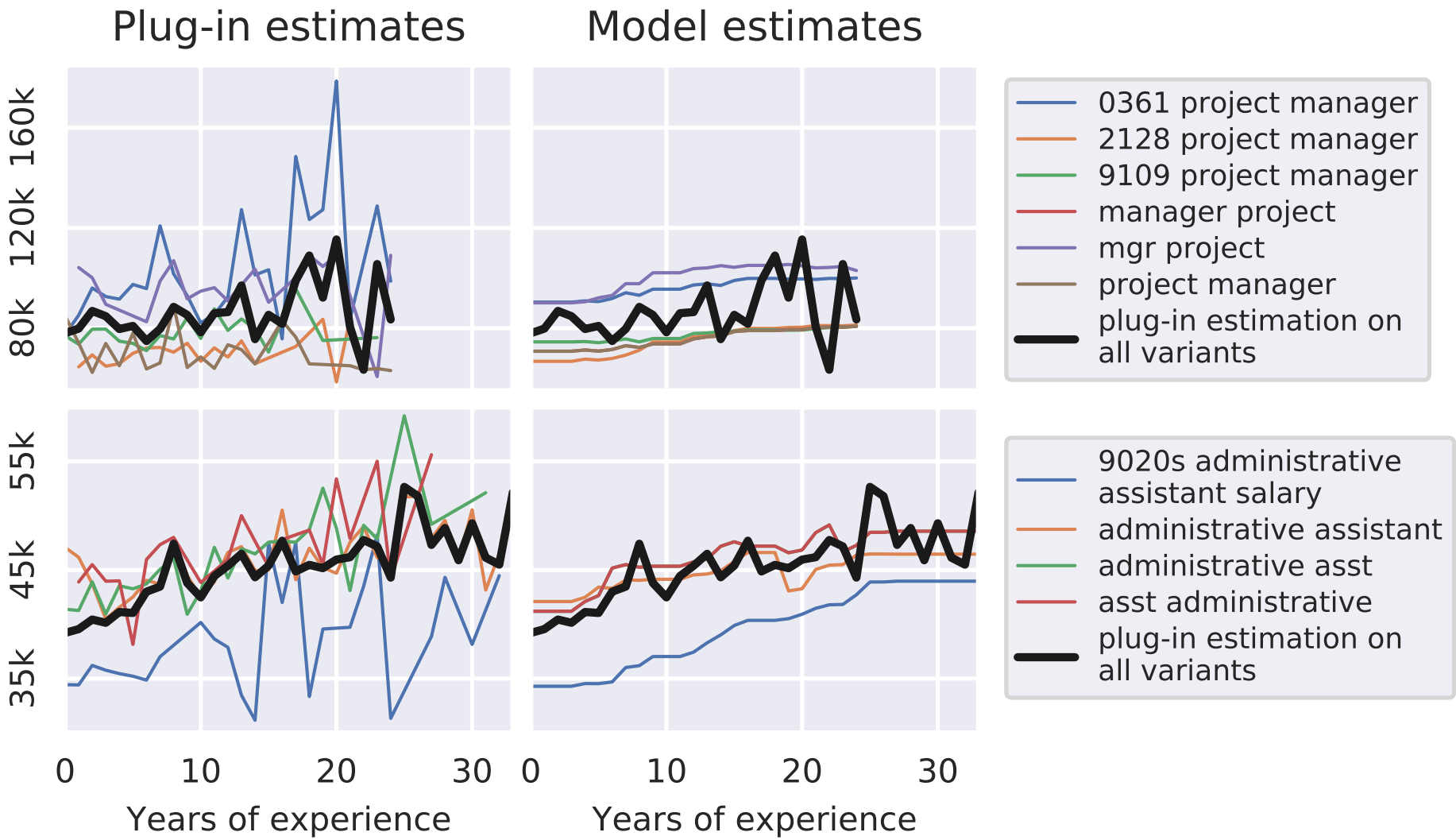
Good statistical analysis without deduplication

[Cerda and Varoquaux 2020]

Analysis with embedding rather than entity matching

Analyzing employee salaries across companies

Salary function of experience



[Cvetkov, in prep]

Analysis with embedding rather than entity matching

Analyzing employee salaries across companies

Relative error on salary

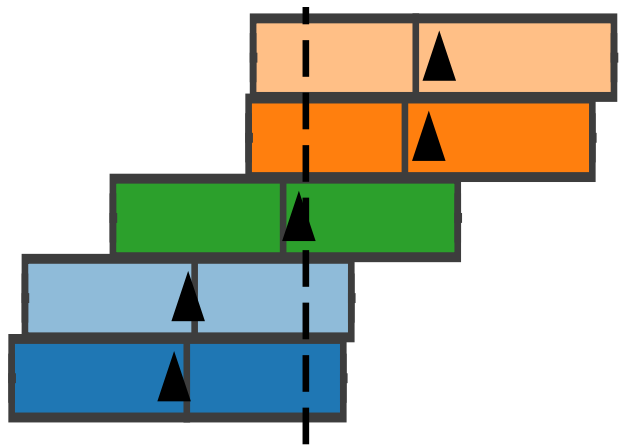
Matching & averaging

Matching & ...

Matching & weighted averaging

Embedding & machine learning

Matching, embedding & learning



Semantic embeddings + machine-learning model outperform manual entity matching

[Cvetkov, in prep]

Conclusion: LearnI – Learning data integration

- Enabling joint analysis of data of different nature / sources
- Difficulty: symbols with errors, lack of normalization
- Unconventional link between statistics and database research

Deep learning for database assembly


@GaelVaroquaux

References I

P. Cerda and G. Varoquaux. Encoding high-cardinality string categorical variables. *TKDE*, 2020.

P. Cerda, G. Varoquaux, and B. Kégl. Similarity encoding for learning with dirty categorical variables. 2018.