Lessons learned: Job Titles Classification & Deep-Learning

2017-09-24

AIUkraine

ALEKSEI PUPYSHEV

R&D Team Lead at Wrike Inc.

Main idea

The deep learning can be a great tool for Data Engineering and Business Intelligence

- It’s easy to start without having PhD in math
- It’s bringing us to development of flexible and scalable solution for data processing
- It’s bringing us to IR for complex data and big data democratization

Structure: business case of Job Titles Classification

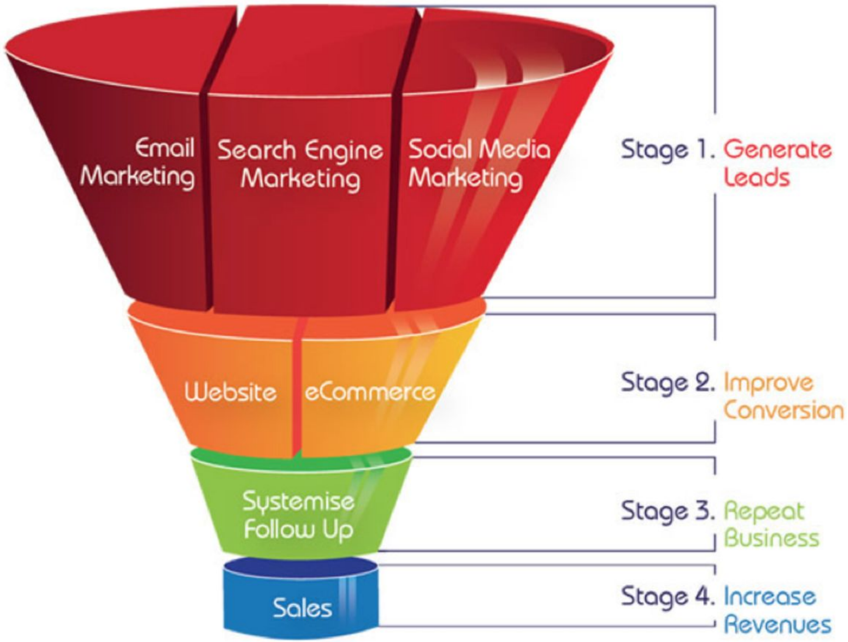
- Why we need a segmentation?
- What data do we have? Why it isn’t a good data format for BI?
- Clusterization *vs* Topic Distribution models
- Why heuristic approaches are good for start only? Maintenance issues
- Deep learning approach
- Alternatives

Why we need a segmentation?

- Wrike: Leading Work Management & Project Management Solution
- Wrike Customer: isn't a single person & not an entire organization
- Wrike Customer can be a single person (lead), or team, or department, or even a group of departments within one organization for different time stage in Wrike account

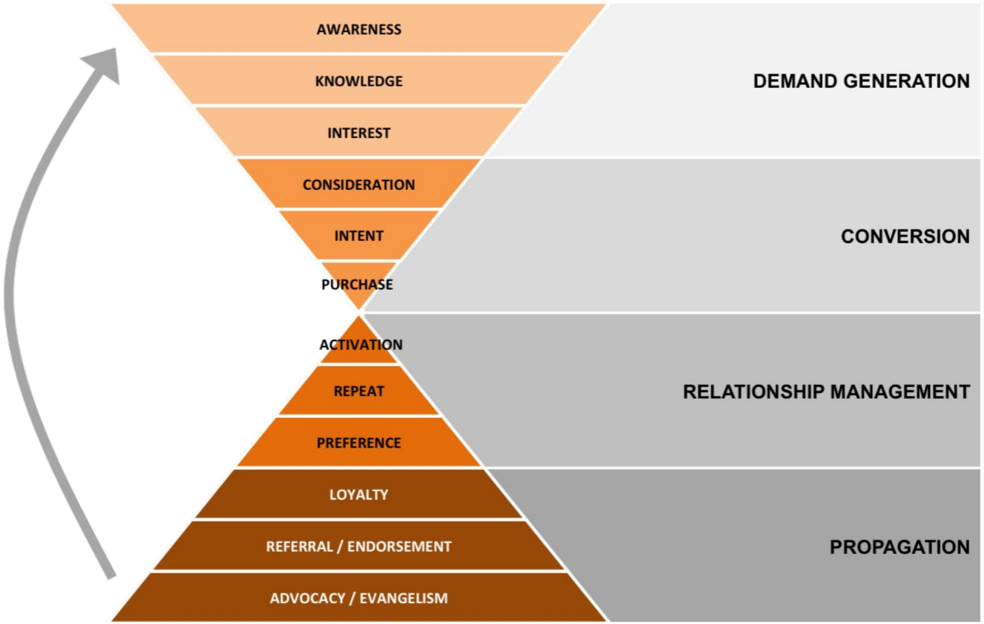
Why we need a segmentation?

Classical funnel



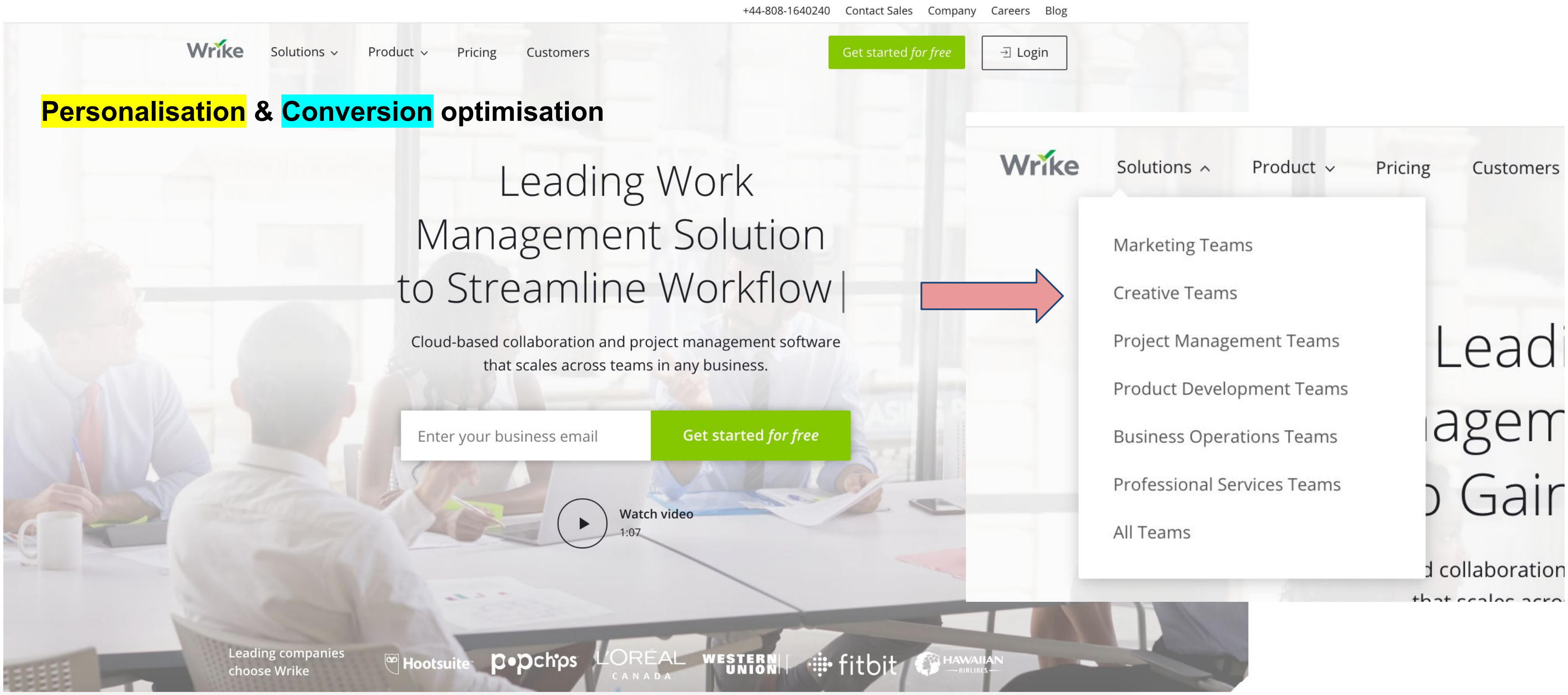
Why we need a segmentation?

B2B SaaS funnel

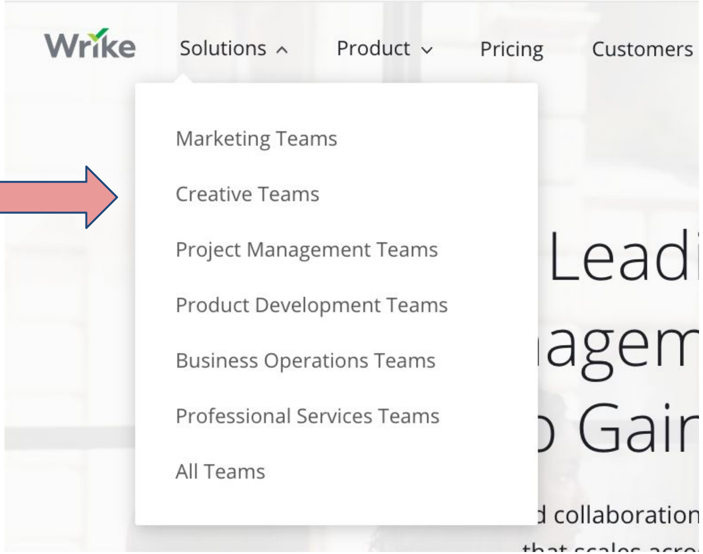


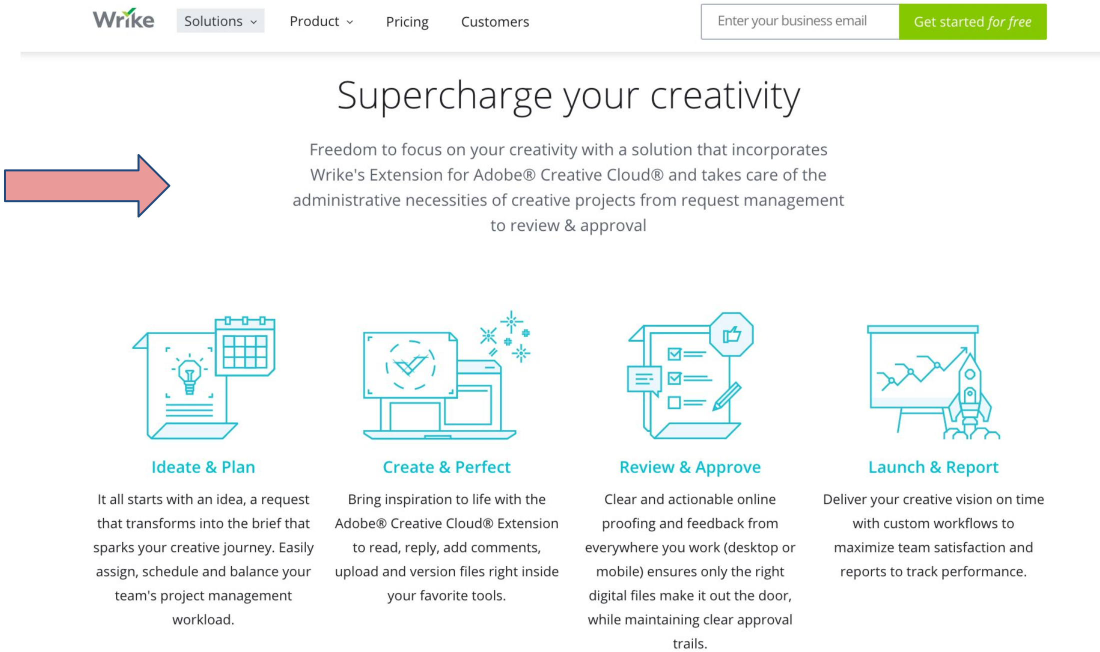
Why do we need a segmentation?

Personalisation & **Conversion** optimisation



Why do we need a segmentation?





https://www.wrike.com/creative-project-management/

What data do we have? Why is it not a good data format for BI?

Segmentation of: visitors, trials, teams, departments, companies, users etc

Segmentation based on factors

Company Size	Industry*	Country*	Channel

Segmentation based on behaviour

Retention	Sessions	Target Actions	Metrics per Account

What data do we have? Why is it not a good data format for BI?

Segmentation of: visitors, trials, teams, departments, companies, users etc

Segmentation based on factors

Company Size	Industry*	Country*	Channel

Segmentation based on behaviour

Retention	Sessions	Target Actions	Metrics per Account

Segmentation based on text data

Free-Text forms	CRM data calls / emails etc	Web Mined data	Bought datasets

Which data do we have? Why is it not a good data format for BI?

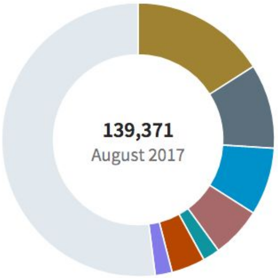
PREMIUM

Employee distribution and headcount growth by function

Based on LinkedIn data. Excludes subsidiaries.

Functions

Functional distribution



Apple Inc.

Headcount growth	6m	1y
Engineering	▲ 4%	▲ 5%
Information Technology	▲ 3%	▲ 4%
Sales	0%	▼ 1%
Operations	▲ 1%	▲ 1%
Administrative	▲ 1%	0%
Arts and Design	▼ 1%	▼ 5%
Business Development	▲ 2%	▲ 2%

Want!

Which data do we have? Why is it not a good data format for BI?

PREMIUM

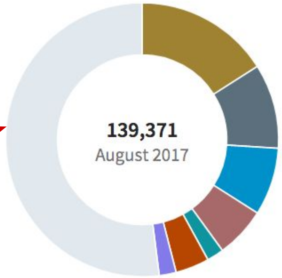
Employee distribution and headcount growth by function

Based on LinkedIn data. Excludes subsidiaries.

Functions

Functional distribution

Other - ?



Apple Inc.

Headcount growth	6m	1y
Engineering	▲ 4%	▲ 5%
Information Technology	▲ 3%	▲ 4%
Sales	0%	▼ 1%
Operations	▲ 1%	▲ 1%
Administrative	▲ 1%	0%
Arts and Design	▼ 1%	▼ 5%
Business Development	▲ 2%	▲ 2%

Want!

Which data do we have? Why is it not a good data format for BI?

SEO, UI/UX, Mktg Ops	Event Planner
President	Administrative Assistant to John MacMillan
Marketing Assistant	**Designer**
VP of Design	Vice President, Marketing
Traffic Coordinator	Communication and Change Manager, West Gulf Coast Region
Lead Graphic Designer	Senior Manager, Marketing Promotions
Sales & Marketing Director	Customer Marketing Manager
VP Mktg	Queen of the World
Department Manager	Communications Specialist
VP Software & Infrastructure Engineering	**Designer**
Associate Marketing Manager	Marketing Operations Manager
Graphic Designer / Jr. Project Manager	Chief Police Officer
Marketing Manager	Assistant Events Manager
Application Analyst/Designer	SR Underwriter
President	**President**
VP of Op	
Marketing Projects Manager	

Have!

Which data do we have? Why is it not a good data format for BI?

- **Stemming**:
  - removing: Jr, Sr. etc.
  - replacing: &,/ etc.

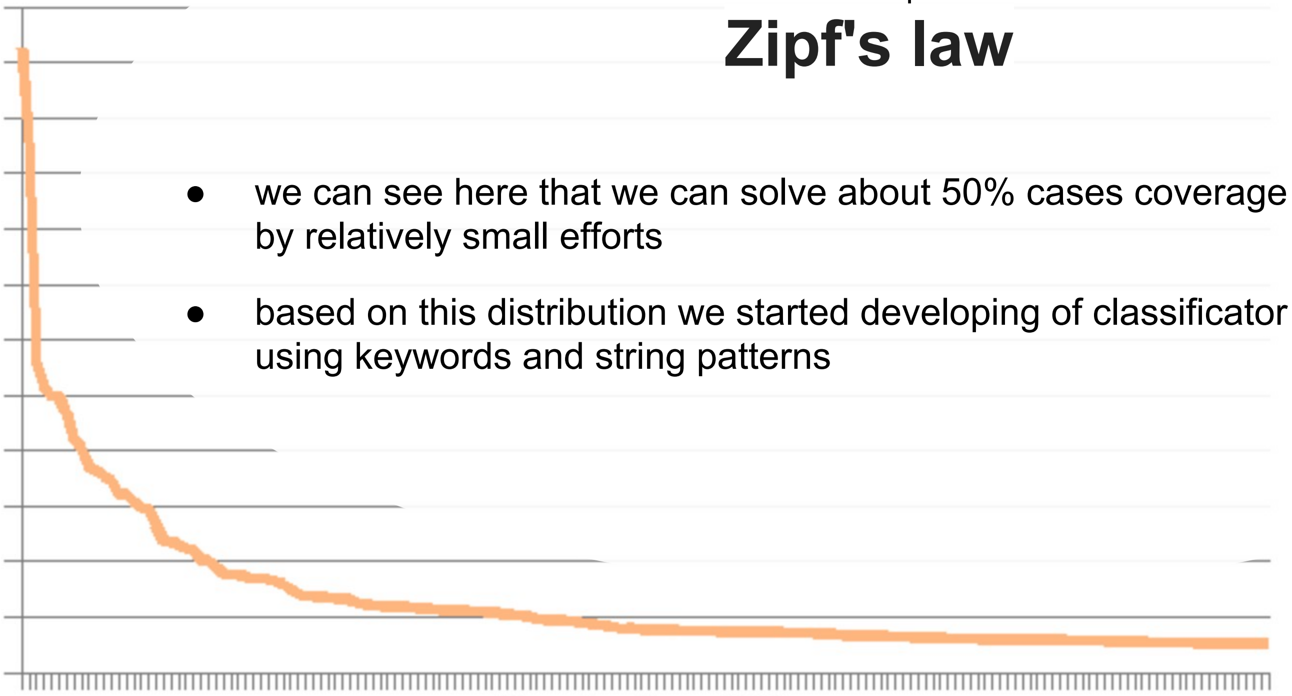
- **Tokenization** and **di-gramms** generating:

“Account Manager” -> [‘account’, ‘manager’, ‘account_manager’]

“Sales Team Lead” -> [sales, ‘manager’, …, ‘team’, ‘team_lead’]

Which data do we have? Why is it not a good data format for BI?

- Tokens distribution picture



- we can see here that we can solve about 50% cases coverage by relatively small efforts
- based on this distribution we started developing of classificator using keywords and string patterns

Which data do we have? Why is it not a good data format for BI?

- There is new question: **which classes should we use?**

- **Standard Occupational Classification** (SOC) and other lists isn’t really helpful for our clients representation

Clusterization *vs* Topic Distribution models

- Clusterization doesn’t work - clusters is really hard for interpretation

- User in general can have a different roles at the same time

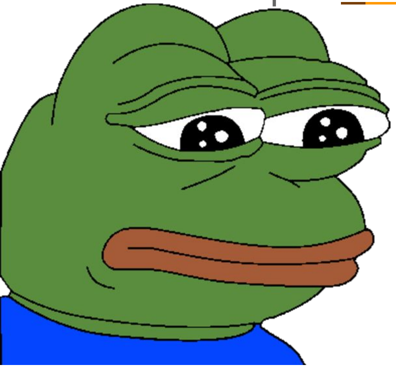
example: Sales & Marketing professional

Clusterization *vs* Topic Distribution models

- Clusterization doesn't work - clusters is really hard for interpretation

- User in general can have a different roles at the same time

example: Sales & Marketing professional



Clusterization *vs* Topic Distribution models

- Clusterization doesn't work - clusters is really hard for interpretation

- User in general can have a different roles at the same time

example: Sales & Marketing professional

Clusterization *vs* Topic Distribution models

- So idea is: let’s define user as a ‘document’ and tokens related to this users as ‘words’ in this document

example: *Aleksei* have

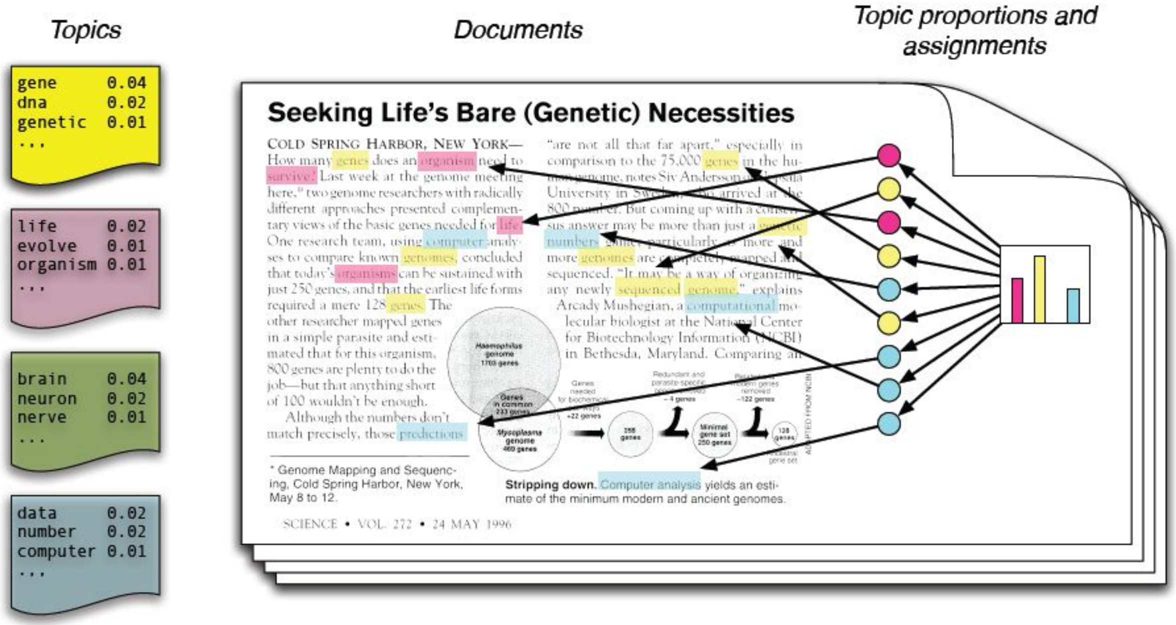
- groups [“data engineers”, “analytics”, “R&D”],
- titles in profile [“software engineer”, “data scientist”, “data analyst”]
- ...

- Now we can use Topic Distribution tools like pLSA or LDA - **Latent Dirichlet Allocation**

Clusterization vs Topic Distribution models

$$P(w,d) = \sum_c P(c)P(d|c)P(w|c) = P(d)\sum_c P(c|d)P(w|c)$$

- Now we can use Topic Distribution tools like **pLSA** or **LDA** - **Latent Dirichlet Allocation**



Clusterization vs Topic Distribution models

'0.066*"marketing_strategy" + 0.055*"marketing" + 0.034*"sales" + 0.033*"management" + 0.032*"social_media_marketing" + 0.031*"strategic_planning"')	**Social Media, PR & Brand Marketing**
'0.077*"training" + 0.054*"leadership_development" + 0.048*"recruiting" + 0.044*"human_resources" + 0.044*"coaching" + 0.040*"organizational_development"')	HR
'0.079*"css" + 0.068*"html" + 0.063*"web_development" + 0.058*"javascript" + 0.042*"php" + 0.040*"jquery"')	Production & QA
'0.058*"nonprofits" + 0.056*"community_outreach" + 0.052*"public_speaking" + 0.043*"fundraising" + 0.030*"volunteer_management" + 0.030*"event_planning"')	Campaign & Event Management
'0.043*"sales" + 0.040*"sales_management" + 0.039*"customer_service" + 0.039*"account_management" + 0.038*"team_building" + 0.035*"management"')	Sales & Account Management
'0.057*"saas" + 0.046*"enterprise_software" + 0.046*"cloud_computing" + 0.044*"software_development" + 0.041*"sql" + 0.038*"java"')	Production & QA

$$P(w,d) = \sum_c P(c)P(d|c)P(w|c) = P(d)\sum_c P(c|d)P(w|c)$$

- **LDA is awesome!**
- **We found interpretable topics and right describing keywords**

Classes

Topics also should be splitted for at least 3 groups - Category Type

- Functional Roles
- Level
- Professional Industry

example: **VP of Marketing**

- High level manager
- Marketing Role

Professional Industry is also important to separate Job Titles like:

- Professor of **Marketing** Management - isn’t really marketing role
- Chief Police **Officer** - isn’t really C-Level like COO/CTO etc.

Classes: Functional Role

Executives
Marketing
Creatives
Customer Services & Customer Support
Product Development
IT Ops & Technology
Engineering
Sales & Account Mgmt
Other Operations
Product mgmt.
Project/Program mgmt.
Consulting & Professional Services
Accounting, Finance & Audit
Administrative
HR
Education
Analytics, Research & Data Science
Other Mgmt Role
Other

Classes: Level

CEO/(Co)Founder/Owner
C-Level & Partner
VP/Head of
Director, Sr. Director, Associate Director
Group Manager, Lead, Coordinator
Consultant, Advisor
Manager, Sr. Manager, Supervisor, Strategist
Specialist, Professional, Analyst
Entry-level
Other

Classes: Professional Industry

Marketing & Advertising
Churches & Charity
Financial & Law Services
Administrative & Government
Manufacturing
Health & Fitness
Science & Education
Information Technology
Media & Entertainment
Design, Photography & Publishing
Environment
Transportation & Supply Chain, Retail
Consumer Services & Leisure
Other

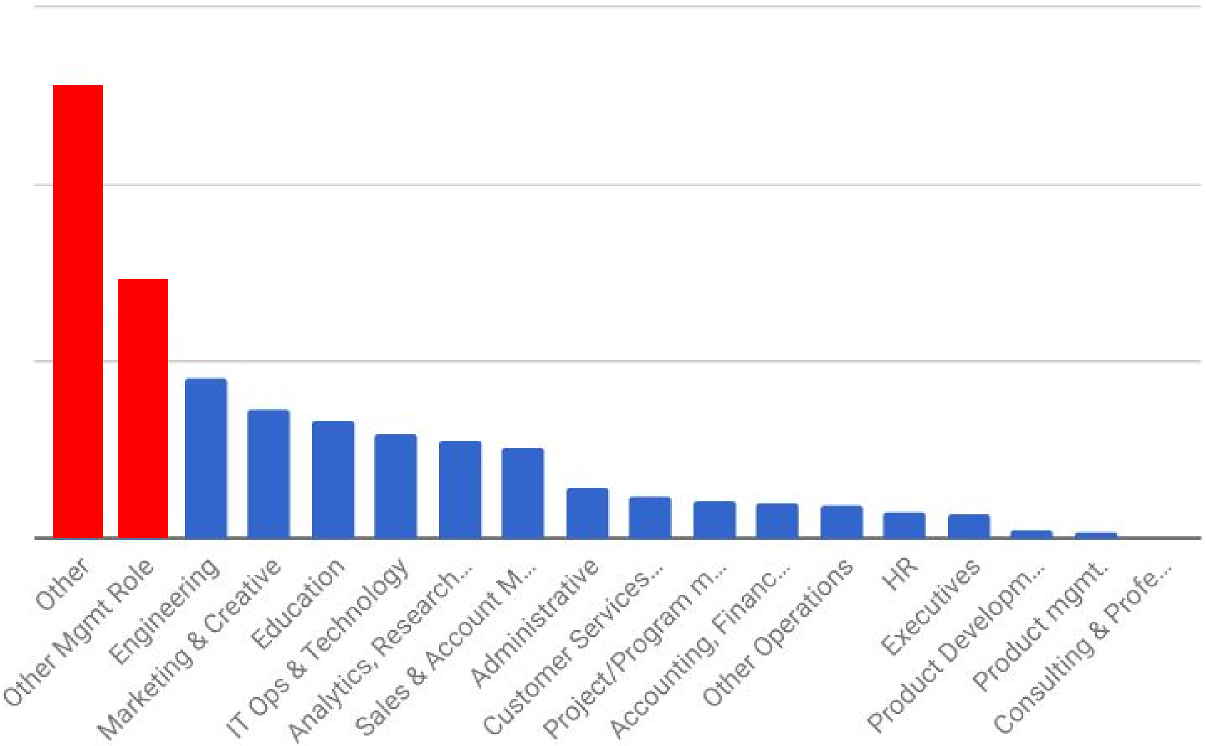
Problems: Misclassification

Sales Engineer -> Engineering

Business Developer -> Engineering

Programmer Marketing Team -> Marketing

Problems: Other



Problems: Other

PREMIUM

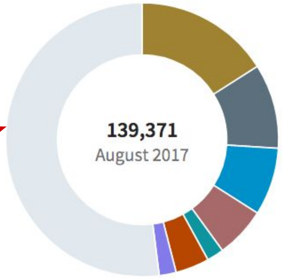
Employee distribution and headcount growth by function

Based on LinkedIn data. Excludes subsidiaries.

Functions

Functional distribution

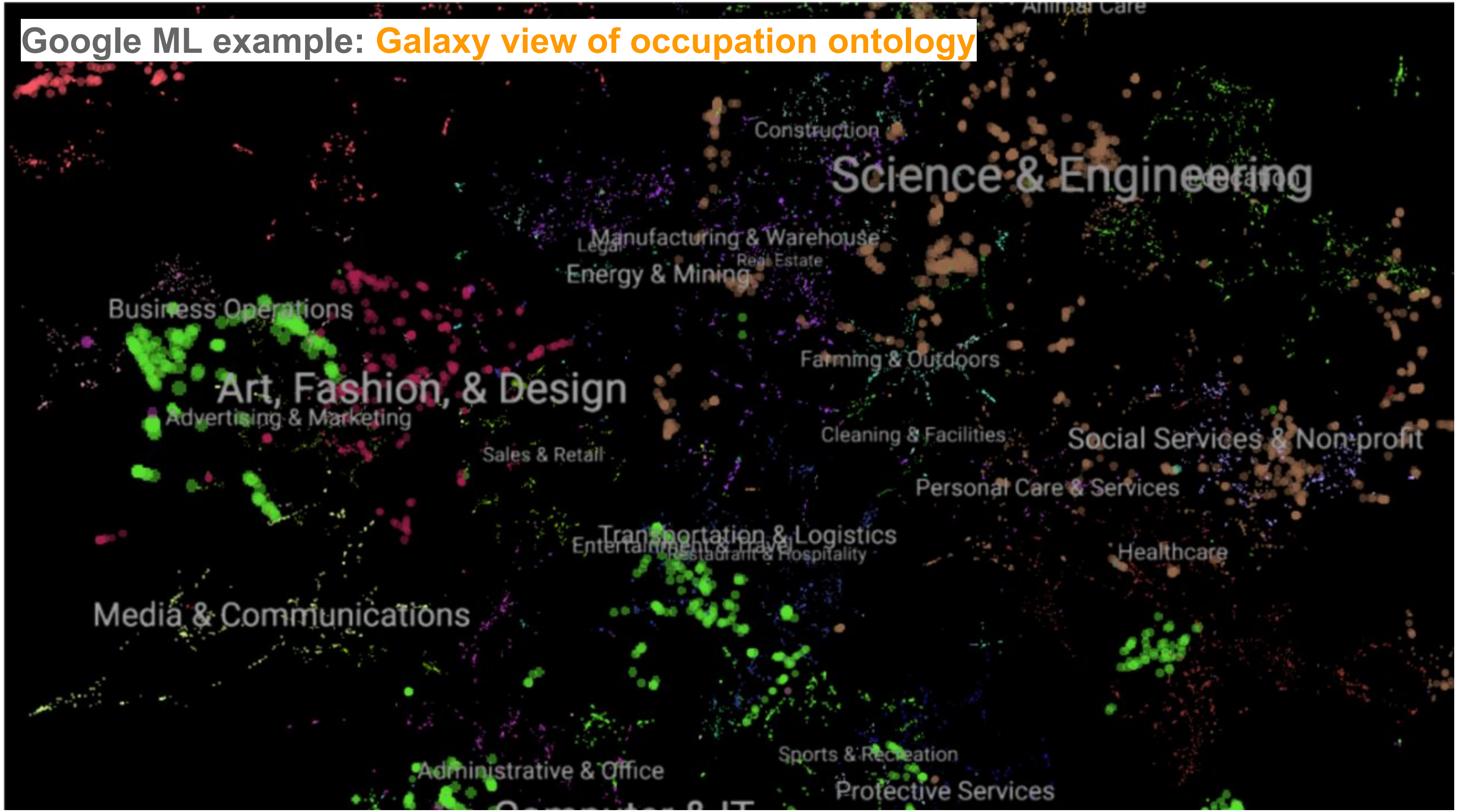
Other - ?



Apple Inc.

Headcount growth	6m	1y
Engineering	▲ 4%	▲ 5%
Information Technology	▲ 3%	▲ 4%
Sales	0%	▼ 1%
Operations	▲ 1%	▲ 1%
Administrative	▲ 1%	0%
Arts and Design	▼ 1%	▼ 5%
Business Development	▲ 2%	▲ 2%

Industry Cases: Google Cloud ML Job Search API



Space of Job Titles can be really helpful

similarity(Sales, Account Manager) -> **0.9**

similarity(Sales, Marketing) -> 0.7

similarity(Sales, Creative Director) -> **0.2**

Deep learning approach

- **Textkernel** team developed a solution special for Job Titles classification

example:

similarity(Sales, Account Manager) -> **0.9**

similarity(Sales, Marketing) -> 0.7

similarity(Sales, Creative Director) -> **0.2**

- **Siamese Neural Network** for ‘similar / dissimilar task’ based on RNN

Deep learning approach

Job title 1	Job title 2	Similarity
Developer	Code Ninja	Similar
Service Desk Agent	Agile Java Tester	Dissimilar
Recruiter	Recriuter	Similar
Data Scientist	Buzz Saw Operator	Dissimilar
Talent Sourcer	Recruiter	Similar

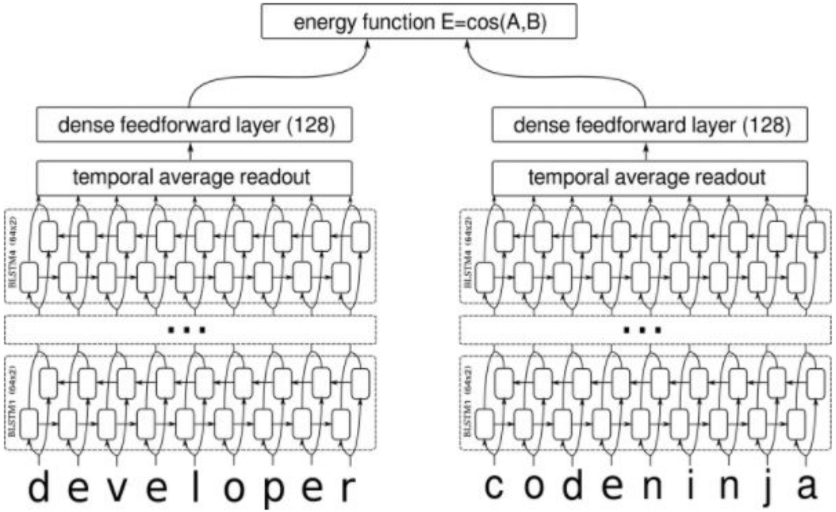
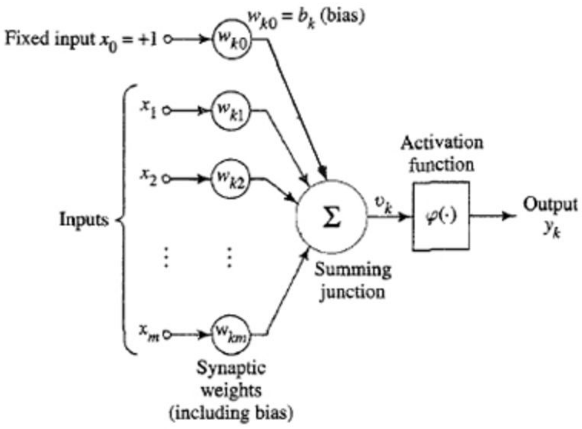
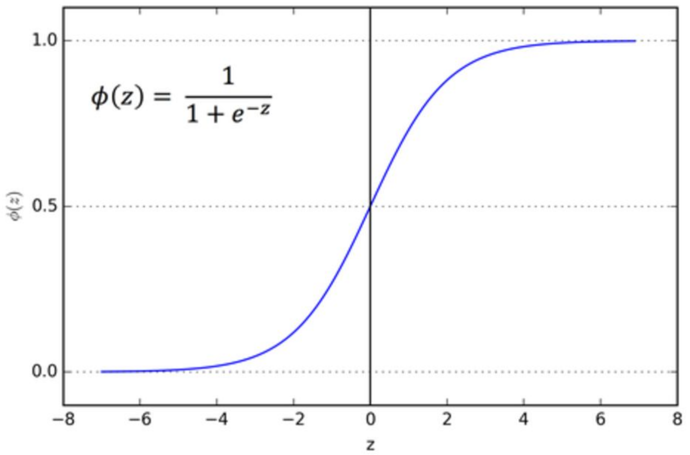


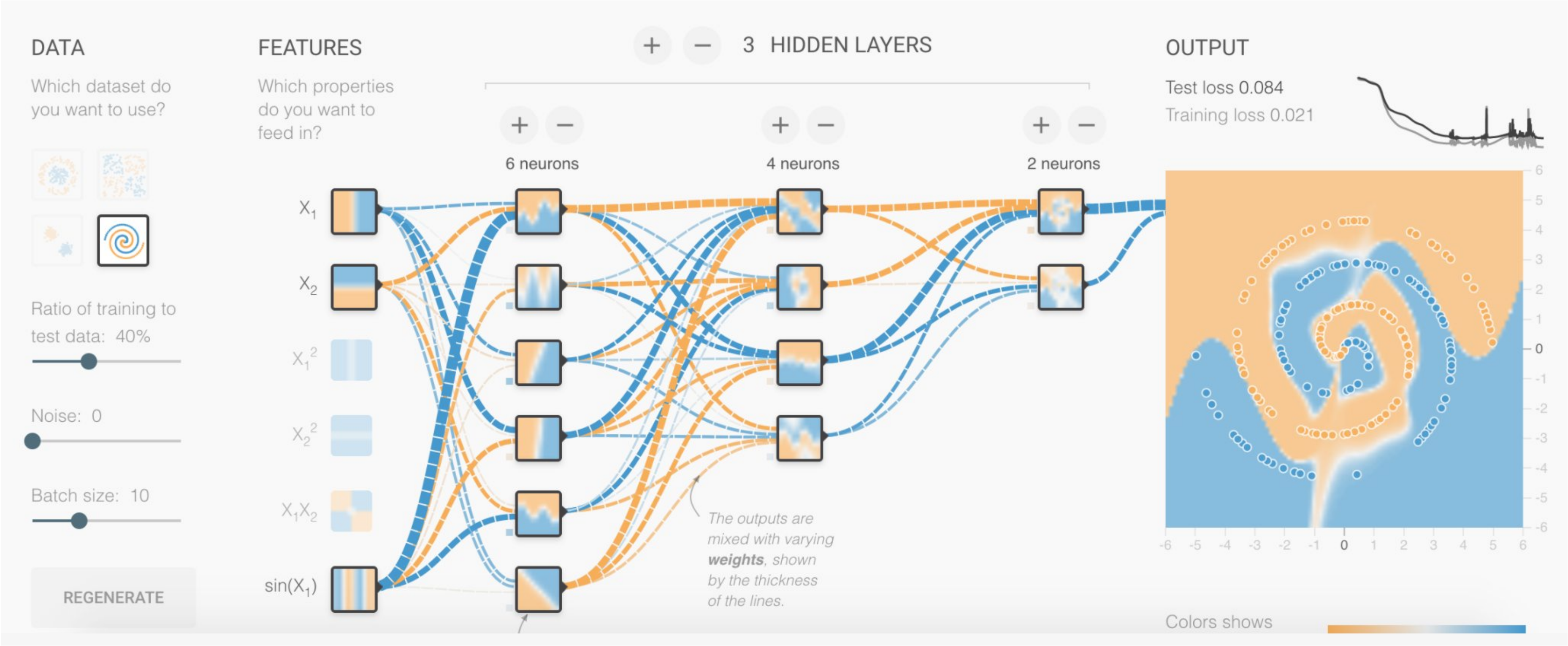
Figure 2: Network overview.

How it works: Neural Network Principles





How it works: Neural Network Principles



Deep learning approach

Job title 1	Job title 2	Similarity
Developer	Code Ninja	Similar
Service Desk Agent	Agile Java Tester	Dissimilar
Recruiter	Recriuter	Similar
Data Scientist	Buzz Saw Operator	Dissimilar
Talent Sourcer	Recruiter	Similar

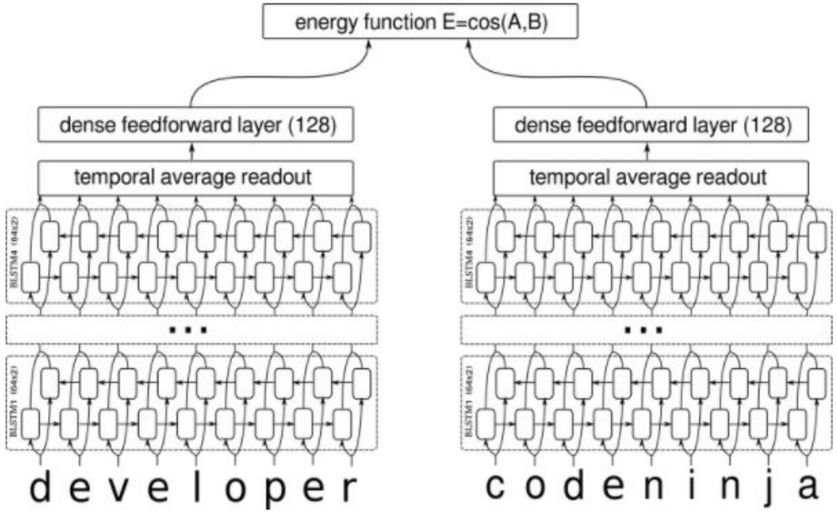
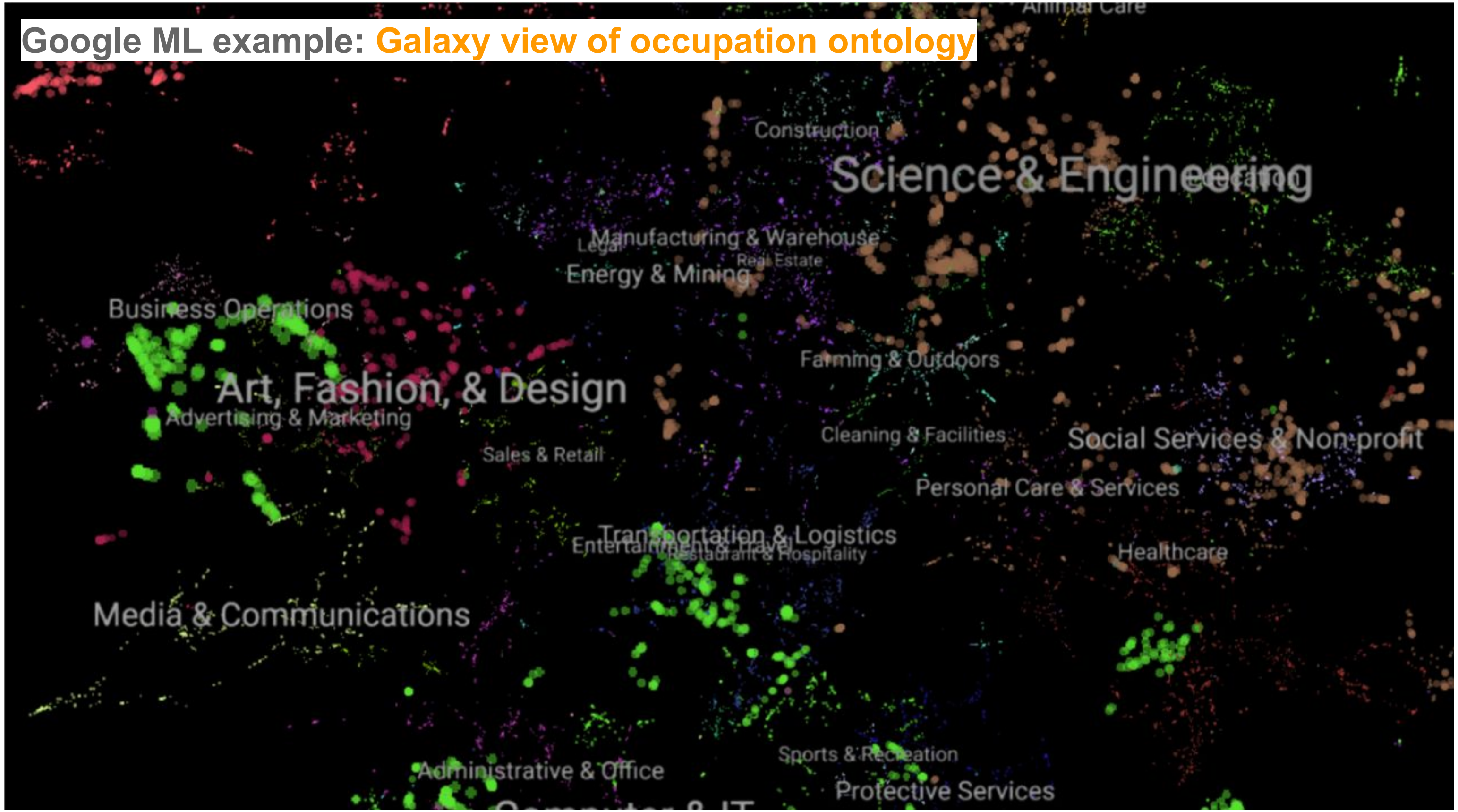


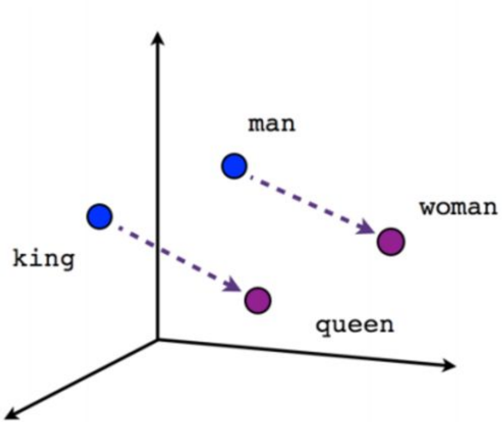
Figure 2: Network overview.

Industry Cases: Google Cloud ML Job Search API

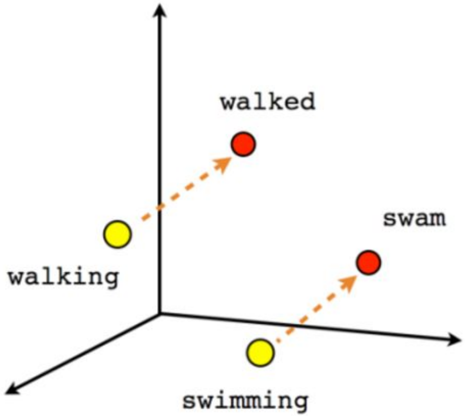
Google ML example: Galaxy view of occupation ontology



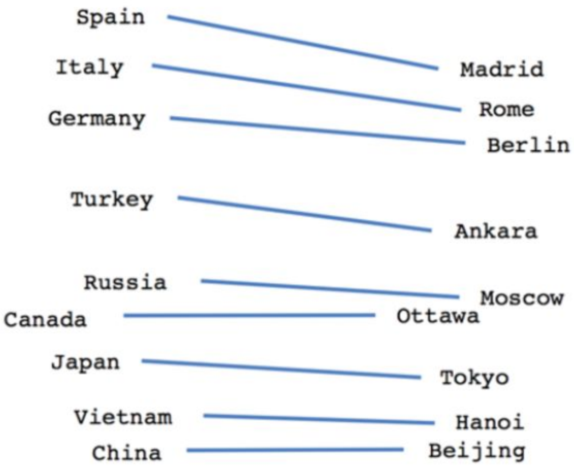
Theory: Word2Vec



Male-Female



Verb tense



Country-Capital

How it works: Dataset Creation Example

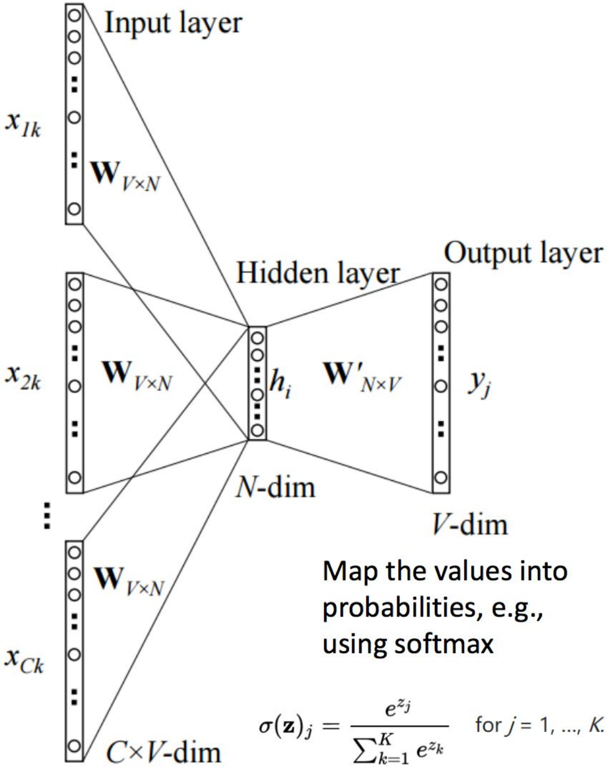
The solution: find a <u>common factor</u>
which combines
different job titles
with adequate sense

Theory: Word2Vec

CBOW

- **Representations**
  - The W and W' are shared for all words
  - W and W' are the things we need
  - Each row in W and W' is the representation of a word in a new N-dimensional space
  - Input/output vectors

$$\mathbf{h} = \frac{1}{C}\mathbf{W}^T(\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_C)$$
$$= \frac{1}{C}(\mathbf{v}_{w_1} + \mathbf{v}_{w_2} + \cdots + \mathbf{v}_{w_C})^T$$



Map the values into probabilities, e.g., using softmax

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } j = 1, \ldots, K.$$

Clusterization vs Topic Distribution models

- So idea is: let's define user as a 'document' and tokens related to this users as 'words' in this document

example: *Aleksei* have

- groups ["data engineers", "analytics", "R&D"],
- titles in profile ["software engineer", "data scientist", "data analyst"]
- ...

- Now we can use Topic Distribution tools like LDA - **Latent Dirichlet Allocation**

Space of Job Titles can be really helpful

similarity(Sales, Account Manager) -> **0.9**

similarity(Sales, Marketing) -> 0.7

similarity(Sales, Creative Director) -> **0.2**

How it works: Tech Stack

How it works: Dataset Creation Example

The solution: find a <u>common factor</u>
which combines
different job titles
with adequate sense

Clusterization vs Topic Distribution models

- So idea is: let's define user as a 'document' and tokens related to this users as 'words' in this document

example: ***Aleksei*** have

- groups ["data engineers", "analytics", "R&D"],
- titles in profile ["software engineer", "data scientist", "data analyst"]
- ...

- Now we can use Topic Distribution tools like LDA - **Latent Dirichlet Allocation**

How it works: Dataset Creation Example

Aleksei Pupyshev
Research and Development Team Lead / Data Scientist at Wrike

Experience:

Research and Development Team Lead at Wrike Inc.

Data Engineer, Data Analyst at Wrike Inc.

Software Engineer at Wrike Inc.

Data Scientist, Quantitative Researcher at QuantumBrains Hedge Fund

Data Scientist, Research Scientist at HBImed AG

How it works: Dataset Creation Example

Aleksei's Experience	Aleksei's Experience
...	...
Data Engineer	Data Engineer
Data Analyst	Data Analyst
Software Engineer	Software Engineer
Data Scientist	Data Scientist
...	...

Job Title 1	Job Title 2	Target
...	...	1
Data Engineer	Software Engineer	1
Data Analyst	Data Scientist	1
...	...	1

within one person
job titles pairing
as POSITIVE

How it works: Dataset Creation Example

random permutations of job titles pairing as NEGATIVE

Aleksei's Experience	Bob's Experience
...	...
Data Engineer	Sales Engineer
Data Analyst	Account Executive
Software Engineer	Business Developer
Data Scientist	Marketing Manager
...	...

Job Title 1	Job Title 2	Target
…	...	0
Data Engineer	Account Executive	0
Data Analyst	Sales Engineer	0
…	...	0

How it works: Dataset Creation Example

Aleksei's Experience	Aleksei's Experience
...	...
Data Engineer	Data Engineer
Data Analyst	Data Analyst
Software Engineer	Software Engineer
Data Scientist	Data Scientist
...	...

Job Title 1	Job Title 2	Target
…	...	1
Data Engineer	Software Engineer	1
Data Analyst	Data Scientist	1
…	...	1

Aleksei's Experience	Bob's Experience
...	...
Data Engineer	Sales Engineer
Data Analyst	Account Executive
Software Engineer	Business Developer
Data Scientist	Marketing Manager
...	...

Job Title 1	Job Title 2	Target
…	...	0
Data Engineer	Account Executive	0
Data Analyst	Sales Engineer	0
…	...	0

How it works: Dataset Creation Example

```
In [47]: df[(df.target == 1) & (df.job_from_headline.str.contains('HR ', case = False))].tail(100)
```

Out[47]:

	job_from_headline	job_from_experience	target	seq_expe	seq_head
50418	HR Technology and Operations Executive	Sr. Manager, Cloud HR Consulting & Transformation	1	[14, 4, 45, 2, 20, 5, 3, 5, 12, 1, 4, 31, 2, 1...	[43, 29, 2, 30, 1, 10, 21, 3, 8, 11, 8, 12, 25...
50882	HR Business Partner, Energy	Recruiter	1	[29, 1, 10, 4, 13, 7, 6, 1, 4]	[43, 29, 2, 35, 13, 9, 7, 3, 1, 9, 9, 2, 22, 5...
50957	EHR Product Manager	Technical Engineer - ETL Data Integrator	1	[30, 1, 10, 21, 3, 7, 10, 5, 11, 2, 27, 3, 12,...	[27, 43, 29, 2, 22, 4, 8, 15, 13, 10, 6, 2, 20...
51624	Manager HR Data Reporting and Analytics	Forecast Manager	1	[34, 8, 4, 1, 10, 5, 9, 6, 2, 20, 5, 3, 5, 12,...	[20, 5, 3, 5, 12, 1, 4, 2, 43, 29, 2, 26, 5, 6...
51692	Senior HR Generalist	Labor Rep	1	[33, 5, 37, 8, 4, 2, 29, 1, 17]	[14, 1, 3, 7, 8, 4, 2, 43, 29, 2, 39, 1, 3, 1,...
53742	HR Business Partner	Branch Manager	1	[35, 4, 5, 3, 10, 21, 2, 20, 5, 3, 5, 12, 1, 4]	[43, 29, 2, 35, 13, 9, 7, 3, 1, 9, 9, 2, 22, 5...
54771	HR Program Manager	Assistant Head Instructor	1	[16, 9, 9, 7, 9, 6, 5, 3, 6, 2, 43, 1, 5, 15, ...	[43, 29, 2, 22, 4, 8, 12, 4, 5, 18, 2, 20, 5, ...
55002	HR & Organizational Effectiveness Leader succe...	Human Resource Manager - Acquisition Integration	1	[43, 13, 18, 5, 3, 2, 29, 1, 9, 8, 13, 4, 10, ...	[43, 29, 2, 49, 2, 32, 4, 12, 5, 3, 7, 56, 5, ...

How it works: Dataset Creation Example

```
In [48]: df[(df.target == 0) & (df.job_from_headline.str.contains('HR ', case = False))].tail(100)
```

Out[48]:

	job_from_headline	job_from_experience	target	seq_expe	seq_head
38419	Sr. HR Information Systems Specialist	Colorado State University Student, Resident As...	0	[19, 8, 11, 8, 4, 5, 15, 8, 2, 14, 6, 5, 6, 1,...	[14, 4, 45, 2, 43, 29, 2, 28, 3, 23, 8, 4, 18,...
38736	Manager, Human Resources/HR Business Partner	Graduate Student	0	[39, 4, 5, 15, 13, 5, 6, 1, 2, 14, 6, 13, 15, ...	[20, 5, 3, 5, 12, 1, 4, 31, 2, 43, 13, 18, 5, ...
40785	HR Staffing	President & CEO	0	[22, 4, 1, 9, 7, 15, 1, 3, 6, 2, 49, 2, 19, 27...	[43, 29, 2, 14, 6, 5, 23, 23, 7, 3, 12]
40803	Retail HR Manager	staff rn	0	[9, 6, 5, 23, 23, 2, 4, 3]	[29, 1, 6, 5, 7, 11, 2, 43, 29, 2, 20, 5, 3, 5...
41505	EHR Applications Specialist	Hilton San Diego Bayfront Finance MDP	0	[43, 7, 11, 6, 8, 3, 2, 14, 5, 3, 2, 26, 7, 1,...	[27, 43, 29, 2, 16, 17, 17, 11, 7, 10, 5, 6, 7...
41556	Talent Acquisition/HR Shared Services Leader	Content Analysis	0	[19, 8, 3, 6, 1, 3, 6, 2, 16, 3, 5, 11, 25, 9,...	[30, 5, 11, 1, 3, 6, 2, 16, 10, 60, 13, 7, 9, ...
41731	HR Generalist	Airline/Aviation Professional	0	[16, 7, 4, 11, 7, 3, 1, 41, 16, 24, 7, 5, 6, 7...	[43, 29, 2, 39, 1, 3, 1, 4, 5, 11, 7, 9, 6]
42022	HR Associate/Compliance Coordinator	Manager of Engineering Services	0	[20, 5, 3, 5, 12, 1, 4, 2, 8, 23, 2, 27, 3, 12...	[43, 29, 2, 16, 9, 9, 8, 10, 7, 5, 6, 1, 41, 1...

How it works: Neural Net. Architecture

Step 0: **Stemming**

.lower

.replace '[^a-z]' to '_'

How it works: Neural Net. Architecture
Step 1: Embeddings Matrix preparation

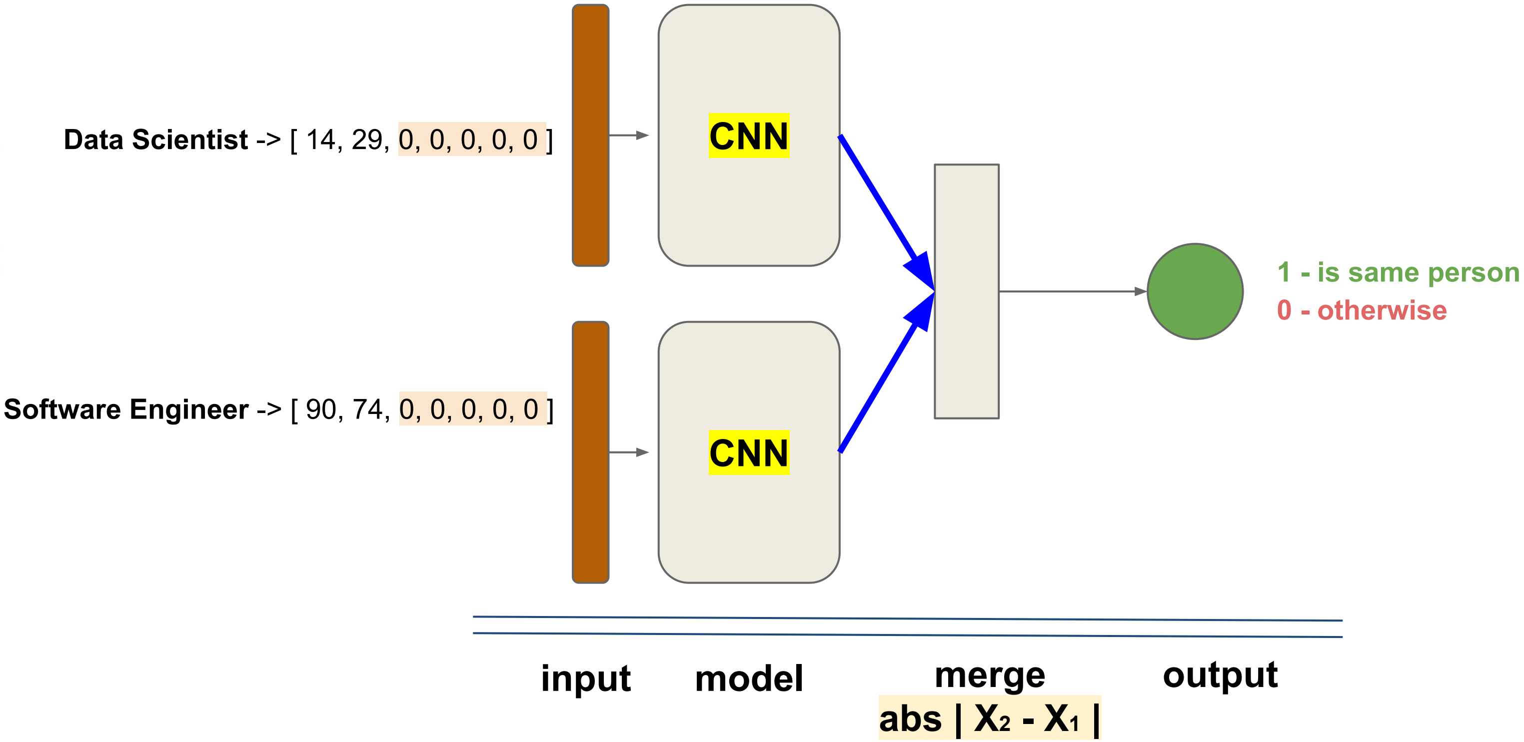
GloVe

'glove.6B.50d.txt'

VOCAB_SIZE = 1000

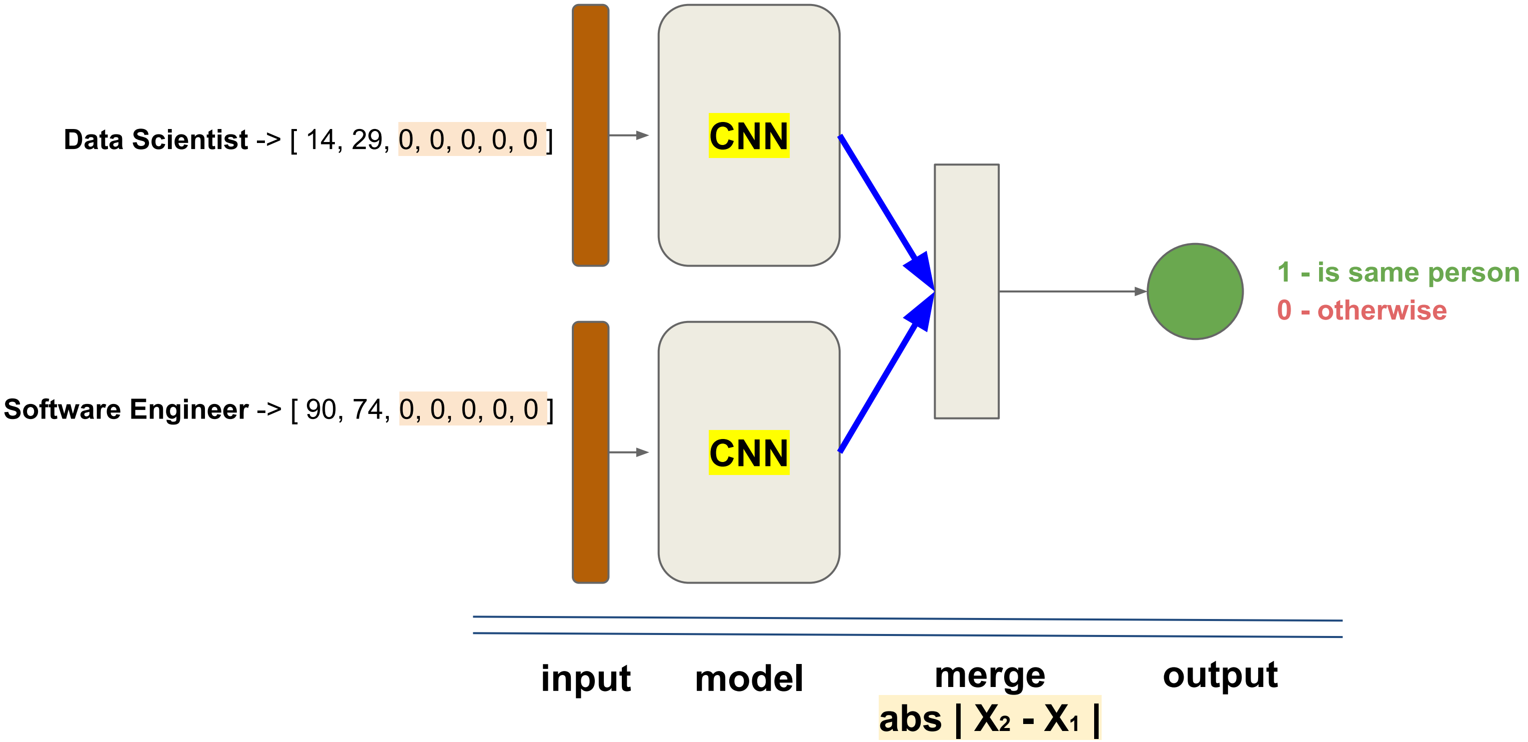
How it works: Neural Net. Architecture

Step 2: **Model Construction [Siamese neural network]**

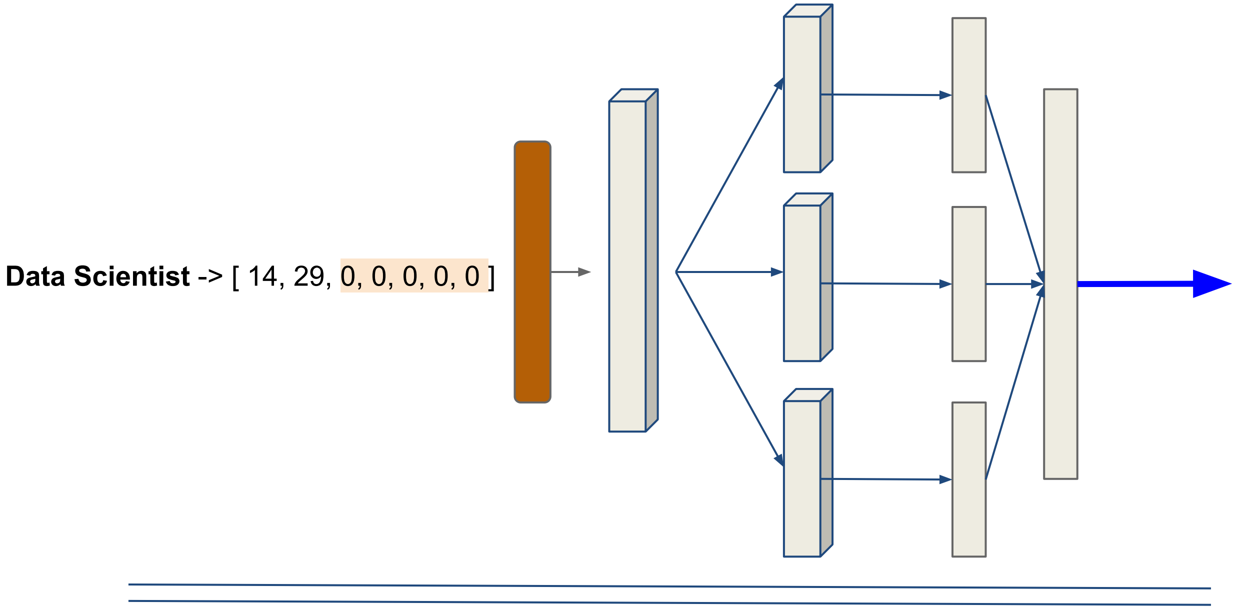


How it works: Neural Net. Architecture

Step 2: Model Construction [Siamese neural network]

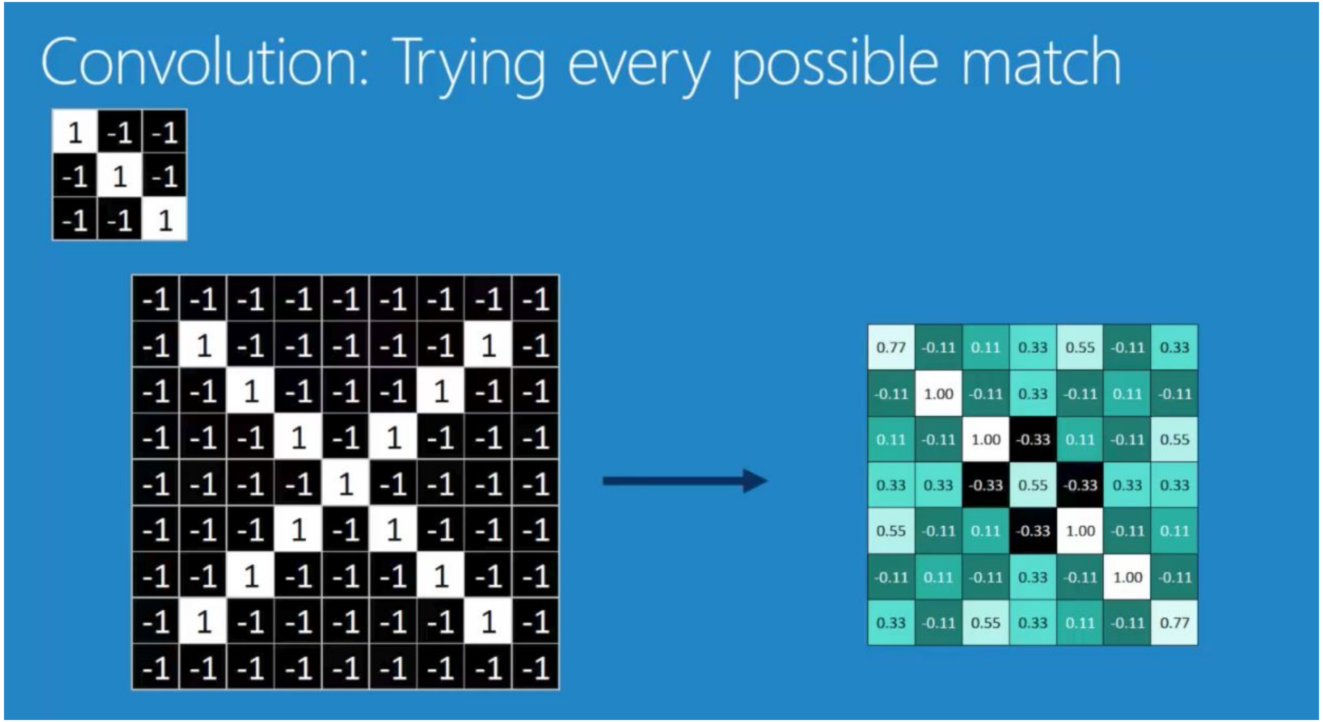


How it works: Neural Net. Architecture
Step 2: Model Construction [Convolution Model]



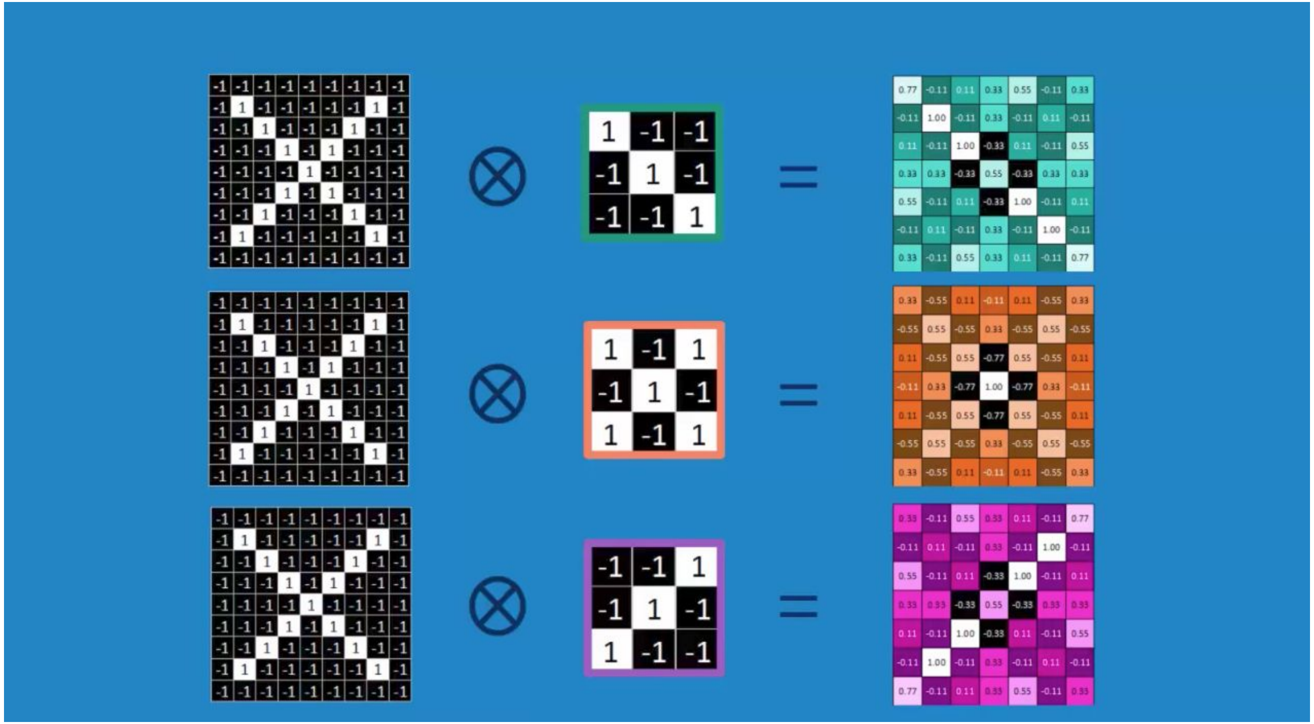
Input -> embeddings -> 3-convolution -> 3-global max pooling -> concatenation -> output

How it works: Neural Net. Architecture
Step 2: Model Construction [Convolution Model]

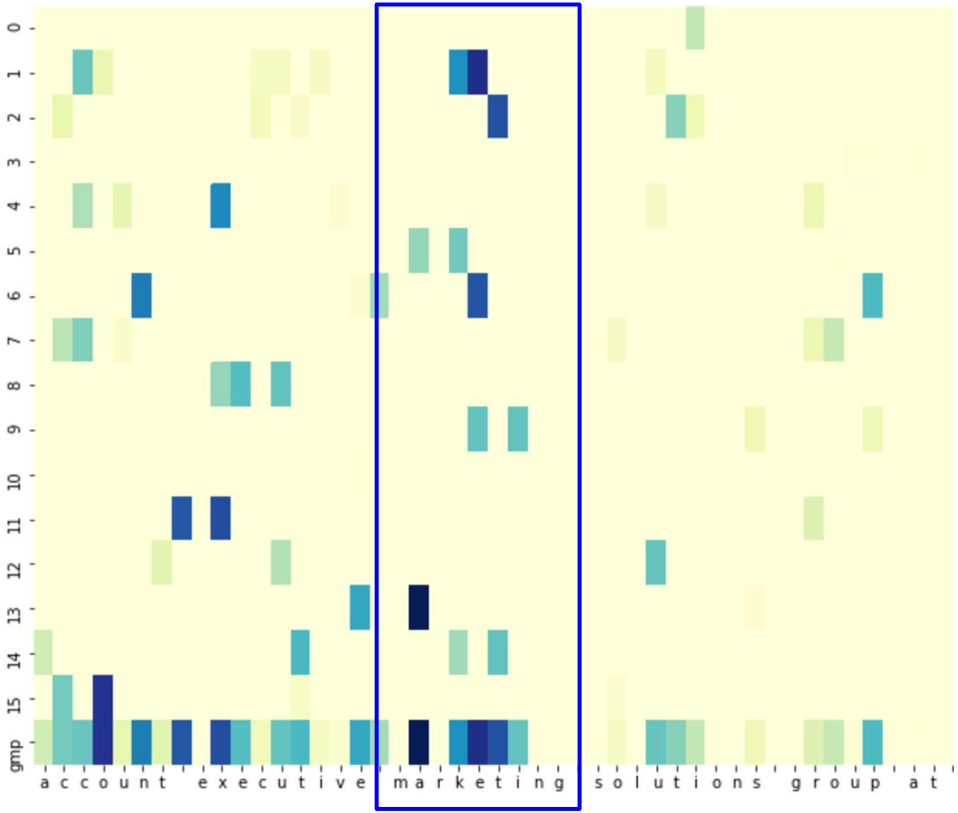


How it works: Neural Net. Architecture

Step 2: Model Construction [Convolution Model]

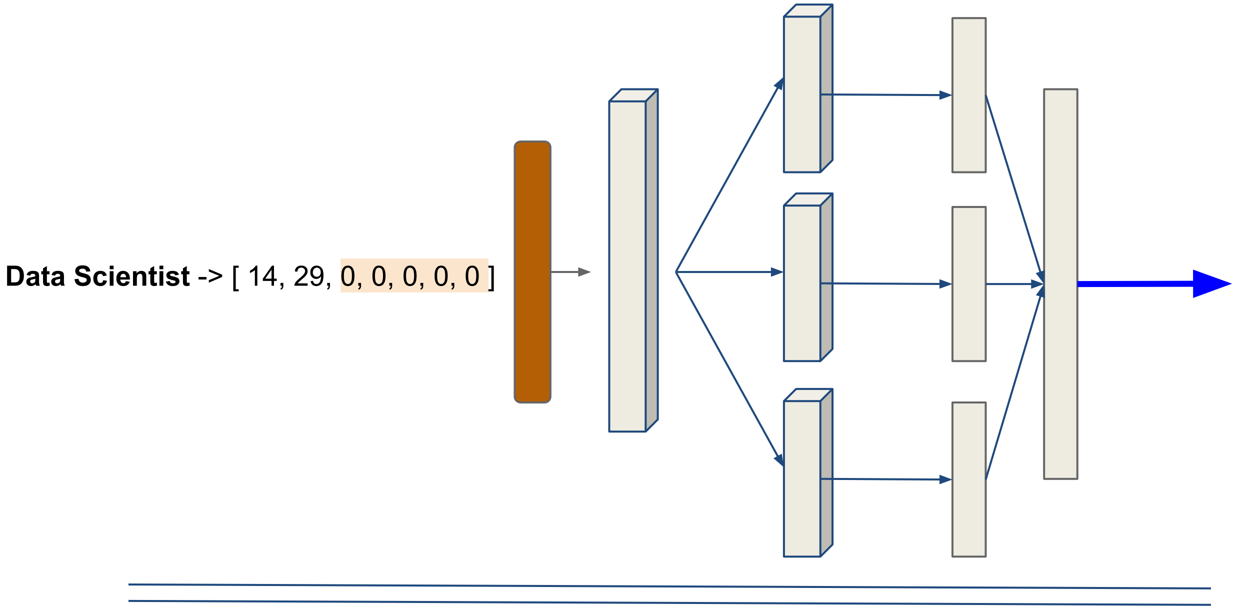


How it works: Neural Net. Architecture [based on characters]



Example: Convolution layer visualization

How it works: Neural Net. Architecture
Step 2: Model Construction [Convolution Model]



Input -> embeddings -> 3-convolution -> 3-global max pooling -> concatenation -> output

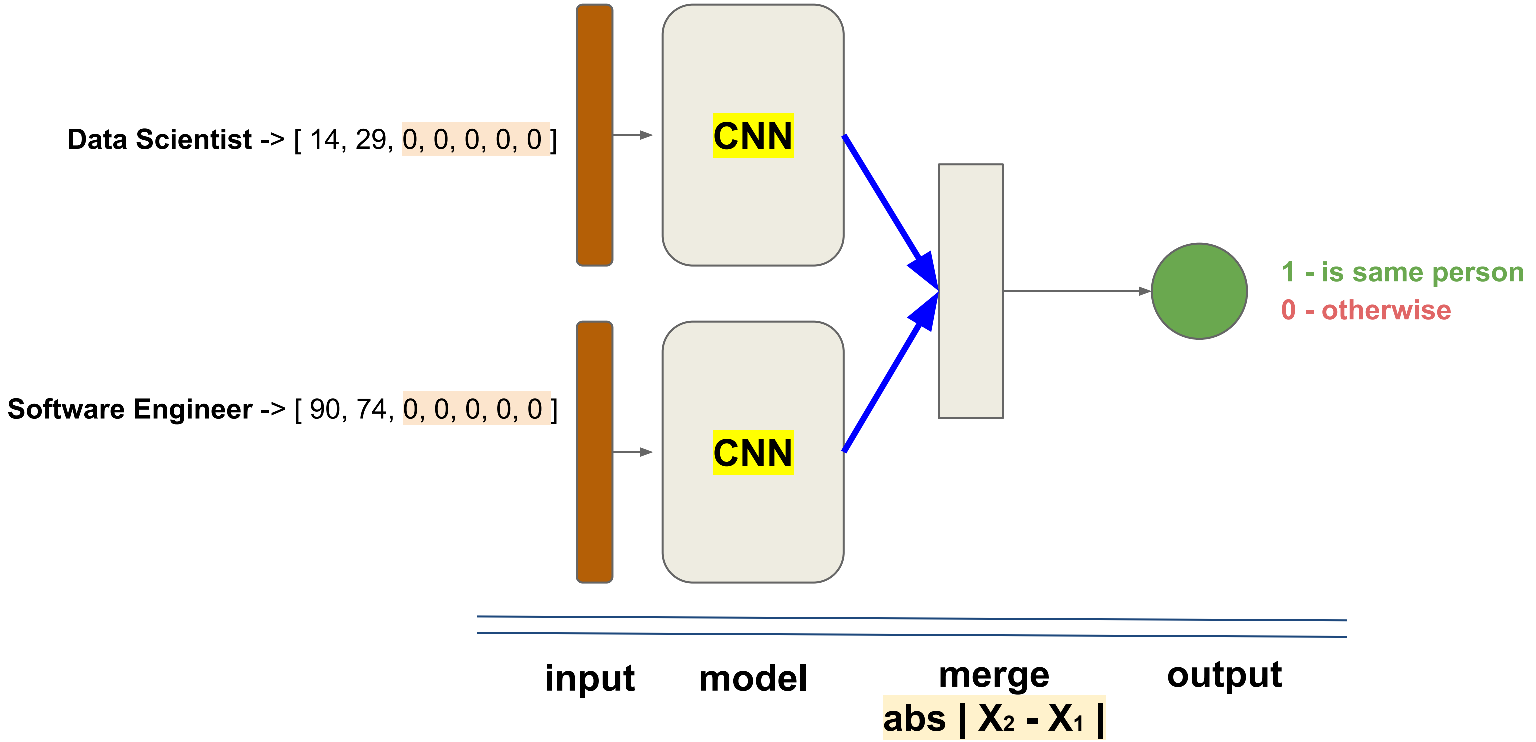
How it works: Neural Net. Architecture

Layer (type)	Output Shape	Param #	Connected to
input_22 (InputLayer)	(None, 7)	0	
embedding_4 (Embedding)	(None, 7, 50)	50050	input_22[0][0]
conv1d_10 (Conv1D)	(None, 7, 20)	1020	embedding_4[0][0]
conv1d_11 (Conv1D)	(None, 7, 20)	2020	embedding_4[0][0]
conv1d_12 (Conv1D)	(None, 7, 20)	3020	embedding_4[0][0]
global_max_pooling1d_10 (GlobalM	(None, 20)	0	conv1d_10[0][0]
global_max_pooling1d_11 (GlobalM	(None, 20)	0	conv1d_11[0][0]
global_max_pooling1d_12 (GlobalM	(None, 20)	0	conv1d_12[0][0]
concatenate_4 (Concatenate)	(None, 60)	0	global_max_pooling1d_10[0][0] global_max_pooling1d_11[0][0] global_max_pooling1d_12[0][0]

Total params: 56,110
Trainable params: 56,110
Non-trainable params: 0

How it works: Neural Net. Architecture

Step 2: Model Construction [Siamese neural network]

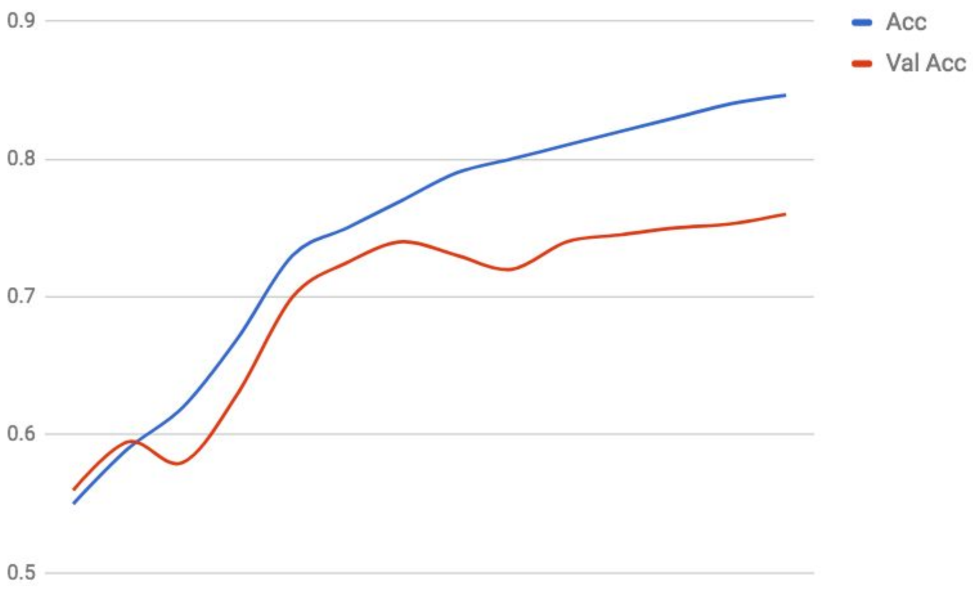


How it works: Neural Net. Architecture

```
Layer (type)                Output Shape       Param #      Connected to
==================================================================================
input_23 (InputLayer)       (None, 7)          0

input_24 (InputLayer)       (None, 7)          0

model_13 (Model)            (None, 60)         56110        input_23[0][0]
                                                            input_24[0][0]

merge_10 (Merge)            (None, 60)         0            model_13[1][0]
                                                            model_13[2][0]

dense_10 (Dense)            (None, 1)          61           merge_10[0][0]
==================================================================================
Total params: 56,171
Trainable params: 56,171
Non-trainable params: 0
```

How it works: Neural Net. Architecture
Step 3: Training

75% Accuracy on test set
84% on train
AUC = 0.85 on test



Flexibility

Example: Senior Java Developer

Functional Role	Most Representative Examples	Similarity	Threshold	Result
Engineering	*Software Engineer*	**0.8**	0.5	**Engineering**
Accounting, Finance & Audit	*Accountant*	0.5		
HR	*HR Manager*	0.3		
Marketing	*Marketing Specialist*	0.6		

Flexibility

Example: Founder & CEO

Functional Role	Most Representative Examples	Similarity	Threshold		Result
Engineering	*Software Engineer*	0.1	0.5	≠	Other
Accounting, Finance & Audit	*Accountant*	0.1			
HR	*HR Manager*	**0.3**			
Marketing	*Marketing Specialist*	0.2			

Maintenance

Before

Set of patterns
Keywords
Priority Rules
Rude Mismatches

No quality metrics
No change management
Un-flexible for new categories
Low coverage

After

Dataset
NN-Model & Tokenizer
Threshold
Relation: Most Representative Examples to Categories

Accuracy
Flexible for fixes
Flexible for new categories
Coverage depends on threshold
Almost same code for other tasks

Flexibility

Example: Senior Java Developer

Functional Role	Most Representative Examples	Similarity	Threshold	Result
Engineering	*Software Engineer*	**0.8**	0.5	**Engineering**
Accounting, Finance & Audit	*Accountant*	0.5		
HR	*HR Manager*	0.3		
Marketing	*Marketing Specialist*	0.6		

Flexibility

Example: Founder & CEO

Functional Role	Most Representative Examples	Similarity	Threshold	Result
Engineering	*Software Engineer*	0.1	0.5	Other
Accounting, Finance & Audit	*Accountant*	0.1		
HR	*HR Manager*	**0.3**		
Marketing	*Marketing Specialist*	0.2		

Flexibility

Example: Founder & CEO

Functional Role	Most Representative Examples	Similarity	Threshold	Result
Executives & Top Management	***Vice President***	<u>**0.9**</u>	0.5	Executives & Top Management
Accounting, Finance & Audit	*Accountant*	0.3		
HR	*HR Manager*	0.4		
Marketing	*Marketing Specialist*	0.6		

Restrictions: New Types of Categories

Example: Director of IT

Authority Level	Most Representative Examples	Similarity	Threshold	Result
CEO/(Co)Founder/Owner	*max { CEO, …, Founder }*	0.6	0.5	C-Level & Partner
C-Level & Partner	*max { COO, CFO, … , CTO }*	0.5		
Director, Associate Director	*max { Head of IT, …, Director of IT }*	0.3		
Specialist, Professional	*max { Data Scientist, …, Recruiter }*	0.5		

Doesn't work!

Restrictions: New Types of Categories

Example of Category: Authority Level

… until you'll find

a <u>common factor</u>

and build

a good accuracy ml-model …

Possible solution: Job Responsibilities

Scalability: Semantic Vectorization & IR

User ID	Job Title	Engineering Score	HR Score	...	Marketing Score
876786876	Founder & CEO	0.3	0.6	...	0.6
509801942	Java Developer	**0.8**	0.1	...	0.2
912300427	Recruiter	0.1	**0.9**	...	0.4
....	....	....	....	....	....

where: **Engineering Score = similarity(Job Title, 'Software Engineer')**

Scalability: Semantic Vectorization & IR

User ID	Job Title	Engineering Score	HR Score	...	Marketing Score
876786876	Founder & CEO	0.3	0.6	...	0.6
509801942	Java Developer	**0.8**	0.1	...	0.2
912300427	Recruiter	0.1	**0.9**	...	0.4
....	....	....	....	....	....

… or: Engineering Score = Agg {

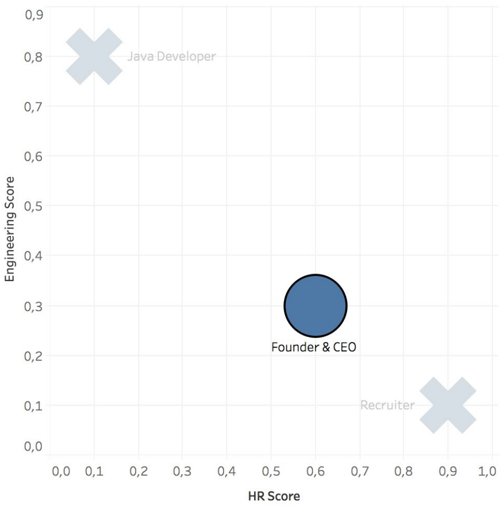
similarity(Job Title, ‘Software Engineer’),
…,
similarity(Job Title, ‘Data Scientist’)

}

where: Agg can be [Max, Min, Avg etc.]

Scalability: Semantic Vectorization & IR

User ID	Job Title	Engineering Score	HR Score	...	Marketing Score
876786876	Founder & CEO	0.3	0.6	...	0.6
509801942	Java Developer	**0.8**	0.1	...	0.2
912300427	Recruiter	0.1	**0.9**	...	0.4
....	....	....	....	....	....



Main idea again

The solution: find a common factor which combines different job titles with adequate sense

Alternatives: Doc2Vec

Distributed Bag of **Words version of Paragraph Vector** (PV-DBOW)

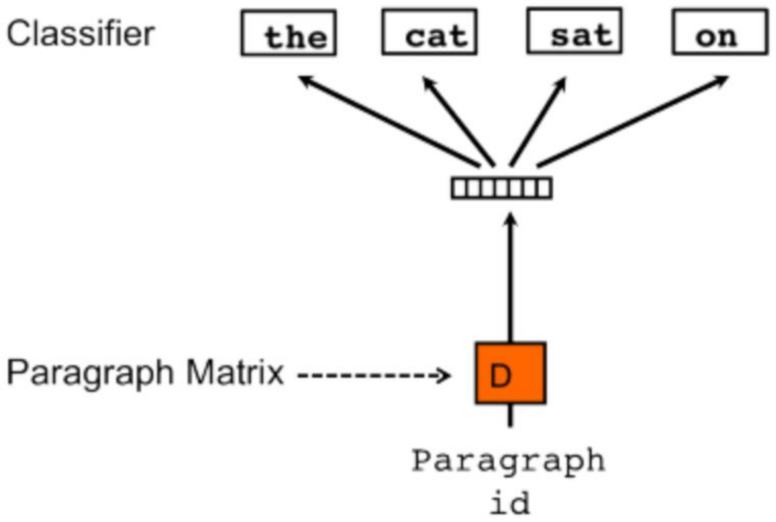


fig 4: PV-DBOW model

Main idea again

The solution: **find a** **<u>common factor</u>**
which combines
different job titles
with adequate sense

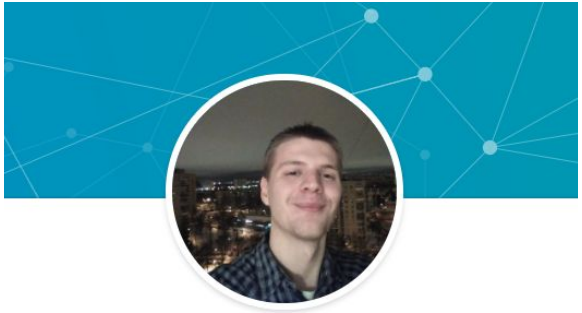
Main idea

The deep learning can be a great tool for Data Engineering and Business Intelligence

- **It’s easy to start without having PhD in math**
- **It’s bringing us to development of flexible and scalable solution for data processing**
- **It’s bringing us to IR for complex and big data democratization**

Special thanks!

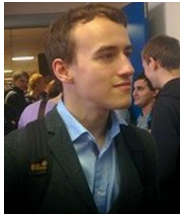
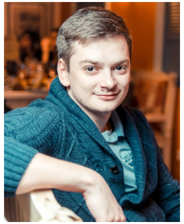
Alexandr Ozerov •
Data Scientist



Pavel Plotnikov • 1st
Data Engineer at Wrike

Revekka Viktorova
Data Analyst

Special thanks!








Q&A