

Discovering Data Insights Through Neural Networks

DataEngConf
October 31, 2017



Sonia Sen

DATALOGUE

Head of Product

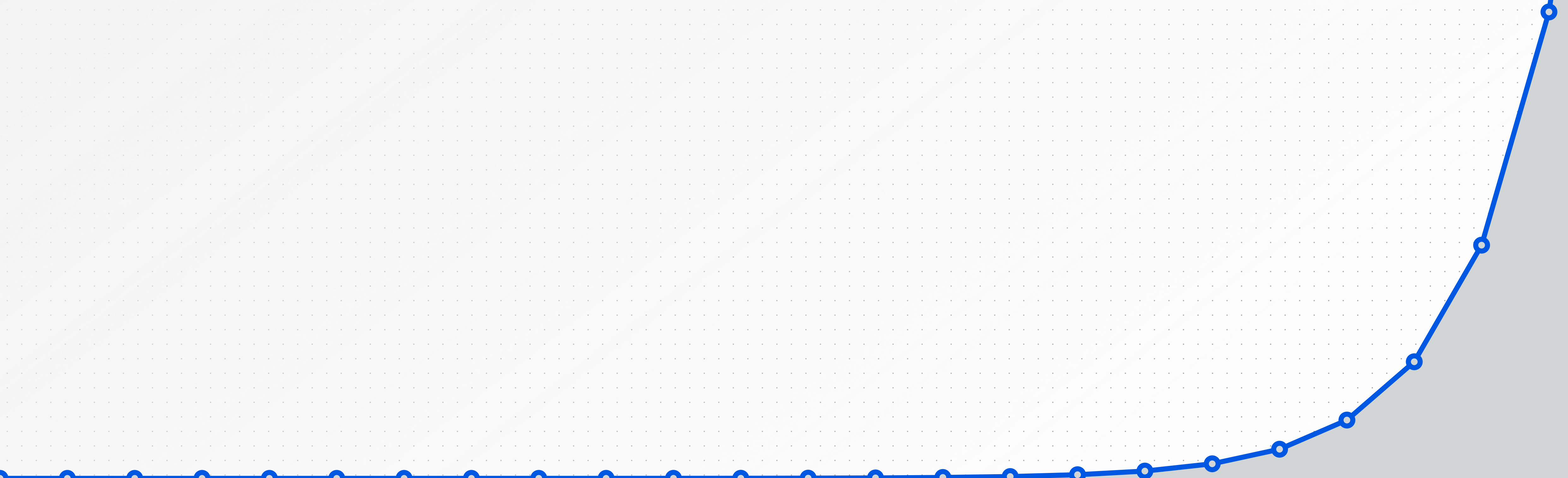


**CORNELL
TECH**

M.S. in Information Science - Health Tech



That data diagram

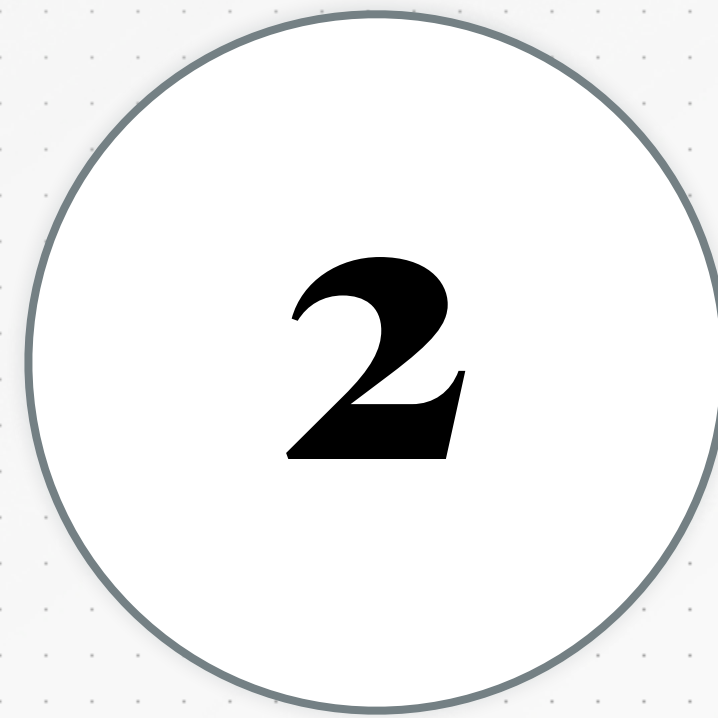


ETL 1.0



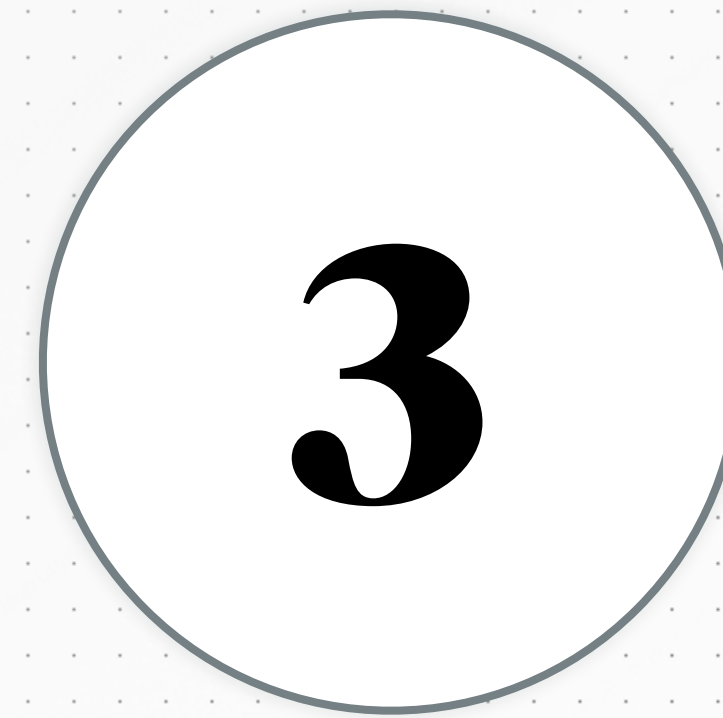
GET

Ad hoc connectors,
scrapers and queries



KNOW

Manual data inspection
and tagging (sampled)
and regular expressions



TRANSFORM

Reliant scripting to manage
schemas, little to no control
over semantic data formats

ETL 1.0: regular expressions

REGEX TO IDENTIFY

```
([A-Z]{3,}|( |,))(\D*)*\W*| (\D*)|  
(\D*)\V(\D*)
```

SCRIPTING

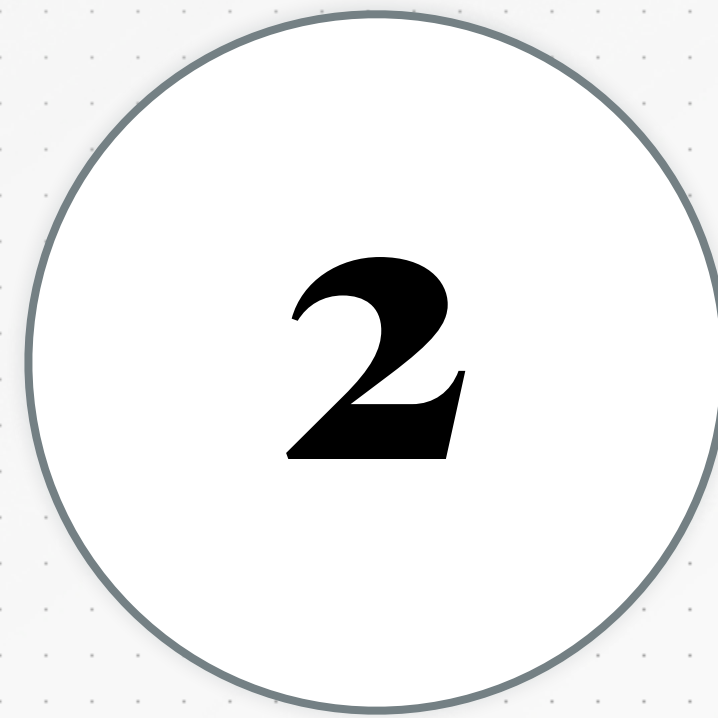
```
parse(string, regex):  
    a = string.find(regex)  
    return a
```

ETL 2.0



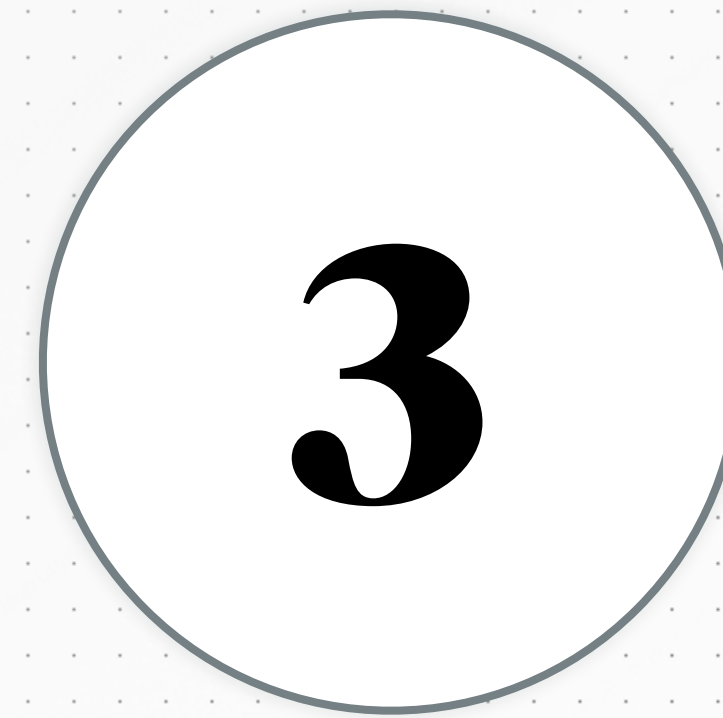
GET

Shared, largely open
source connectors



KNOW

Regular expression and
simple rules based
data mapping



TRANSFORM

Visual, shared, component
based transformations

ETL 2.0: GUI based scripting

TRIFACTA

extract

Ex

The Extract transform extracts data that follows a specified pattern from a given column and creates a new column containing that data. The original column remains unchanged.

Syntax:

```
extract col: column on: 'pattern'  
limit: integer quote: 'string'  
at: position after: 'pattern'  
before: 'pattern' from: 'pattern'  
to: 'pattern' urlparam: 'string'
```

Example:

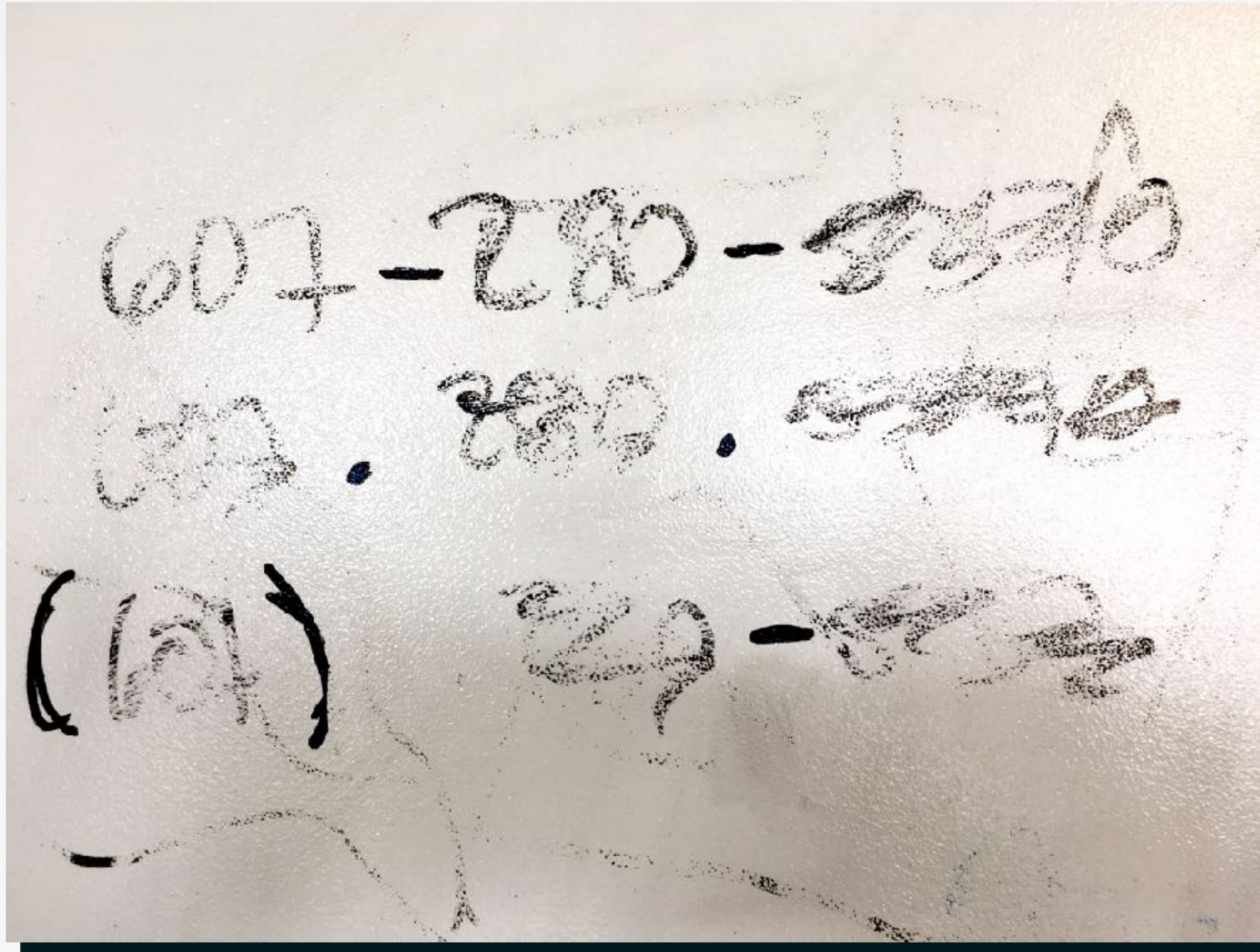
```
extract col: text on: /[0-9]+/ limit: 10
```

PAXATA

The screenshot shows the Paxata interface for a project titled "Prep for customer product segmentation". A "Compute Values" tool is configured to calculate "Total Revenue" by summing "Interest Revenue" and "Fees Revenue". Below the tool configuration, a data table is displayed with the following columns: Address, City, State, ZipCode, Country, MainPhoneNum..., Interest Revenue, Fees Revenue, and Total Revenue. The table contains five rows of data:

Address	City	State	ZipCode	Country	MainPhoneNum...	Interest Revenue	Fees Revenue	Total Revenue
	CULVER CITY	CA	90232-2425	USA	3102020011	9383.1	2841.19	
	ROSEMEAD	CA	91770-2328	USA	6263073200	498	9332.1	
	LA HABRA	CA	90631-3933	USA	5626946551	9271.1	90377.01	
	LOS ANGELES	CA	90025-4260	USA	3104994150	5557.25	200403	
	LOS ANGELES	CA	90025-0000	USA	3102316100	3532.08	7851.15	
	LOS ANGELES	CA	90025-0000	USA	3102316100	9458.73	7669.15	

The aha! moment



(568) 487-7267
464.211.3998
465-111-1061
173 296 1782
309 612-8634
(754) 488-1728
917.181.8932
(168) 785-8583 ext. 2343
+1 946.456.7514
+1 (281) 229-6261

Regex + machine learning

Dates

Mardi 22 Mars, 1991

Tuesday March 22nd 1991

03/22/1991

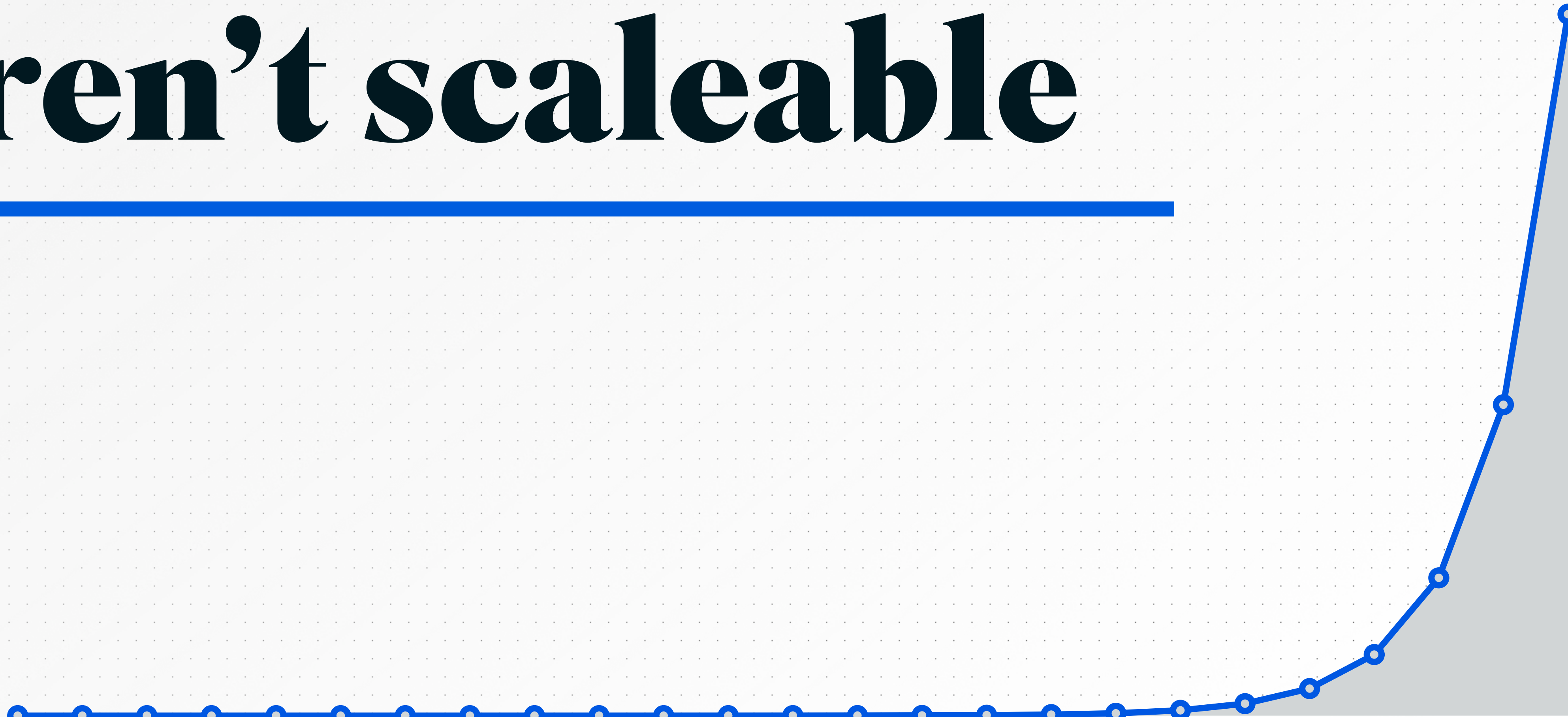
Length	4
# Letters	0
# Digits	4
# Special chars	0
Index special char	-1
...	...

 Text

 Numbers

 Special chars

**Existing systems
aren't scaleable**



ETL 3.0: deep learning

1

VDCNN

Semantic + structural
understanding

2

LSTM + CRF

Parsing of
unstructured data

3

SEQ₂SEQ

Translation of data from one
format to another

Deep learning

1

VDCNN

Semantic + structural
understanding

2

LSTM + CRF

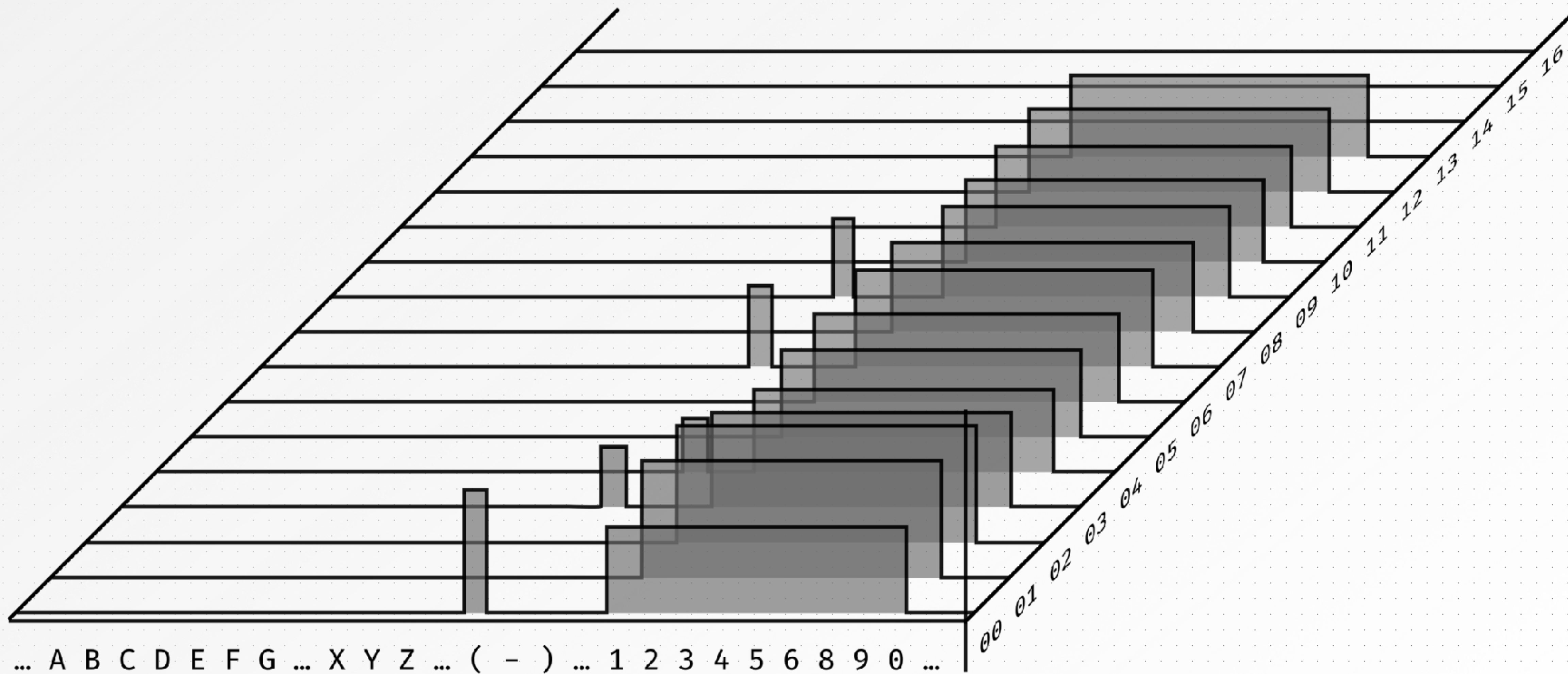
Parsing of
unstructured data

3

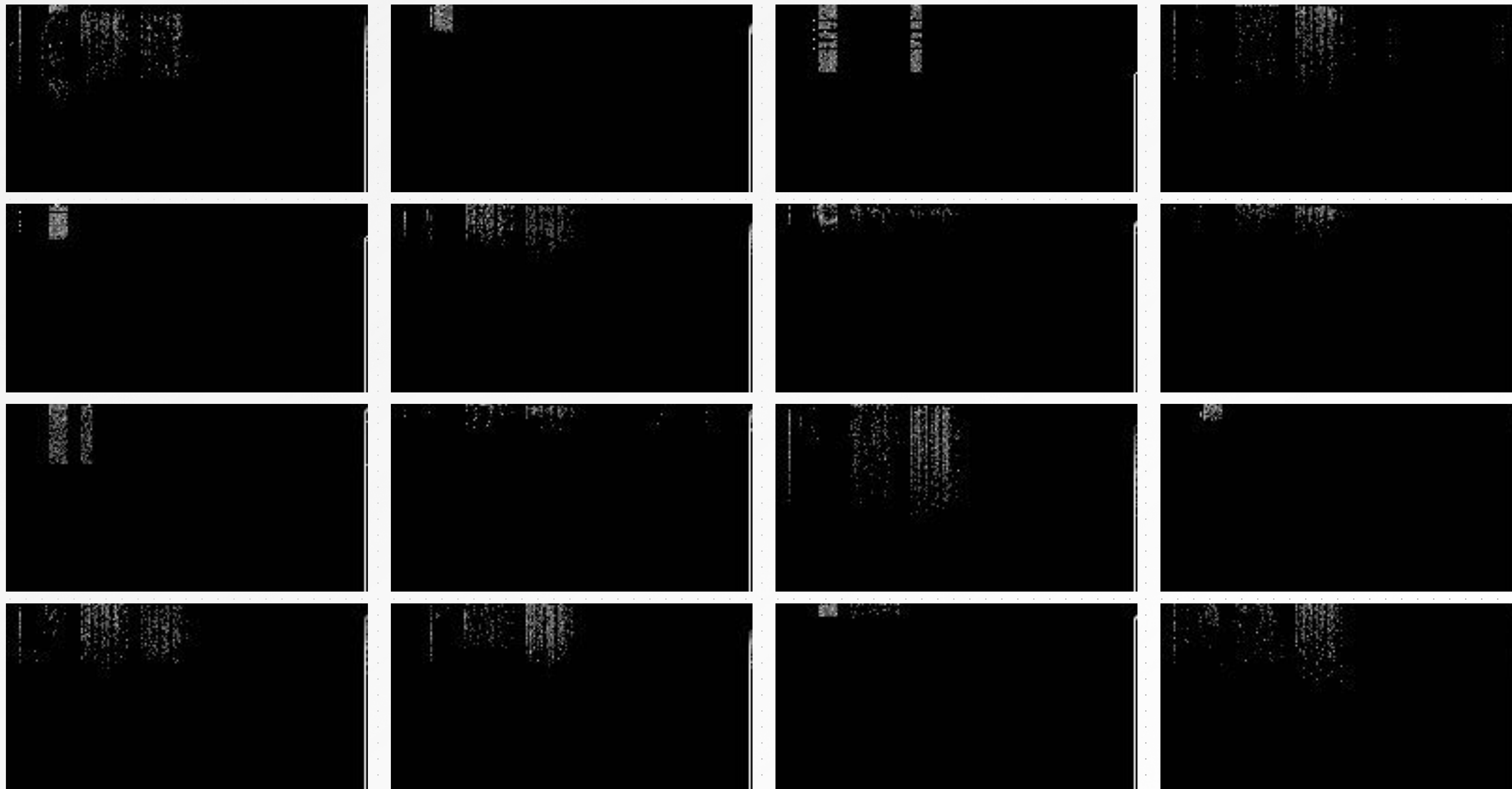
SEQ₂SEQ

Translation of data from one
format to another

VDCNN inputs



VDCNN inputs



VDCNN outputs

"Oliver Wyman"



```
{  
  "score": "0.936947",  
  "tag": "organization"  
}
```

Organization

```
{  
  "score": "0.152489",  
  "tag": "address"  
}
```

Address

```
{  
  "score": "0.908231",  
  "tag": "full_name"  
}
```

Full name

Deep learning

1

VDCNN

Semantic + structural
understanding

2

LSTM + CRF

Parsing of
unstructured data

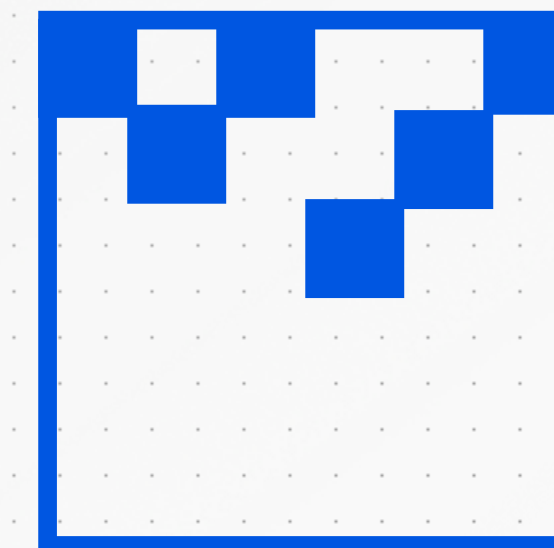
3

SEQ₂SEQ

Translation of data from one
format to another

LSTM + CRF

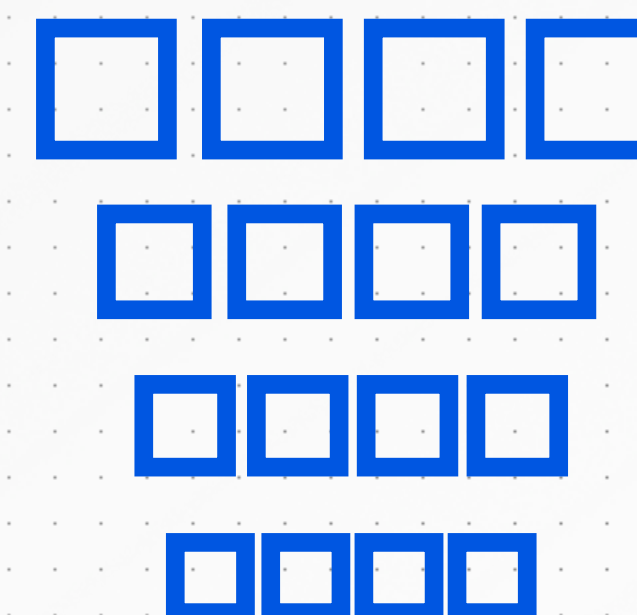
Char Embedding



Bidirectional LSTM

+

CRF



Parsed String

10 Airport Road SE,Salem,NY,97301
AAAAAAAAAAAAAAAAAUCUUUUSSUZZZZZ

Deep learning

1

VDCNN

Semantic + structural
understanding

2

LSTM + CRF

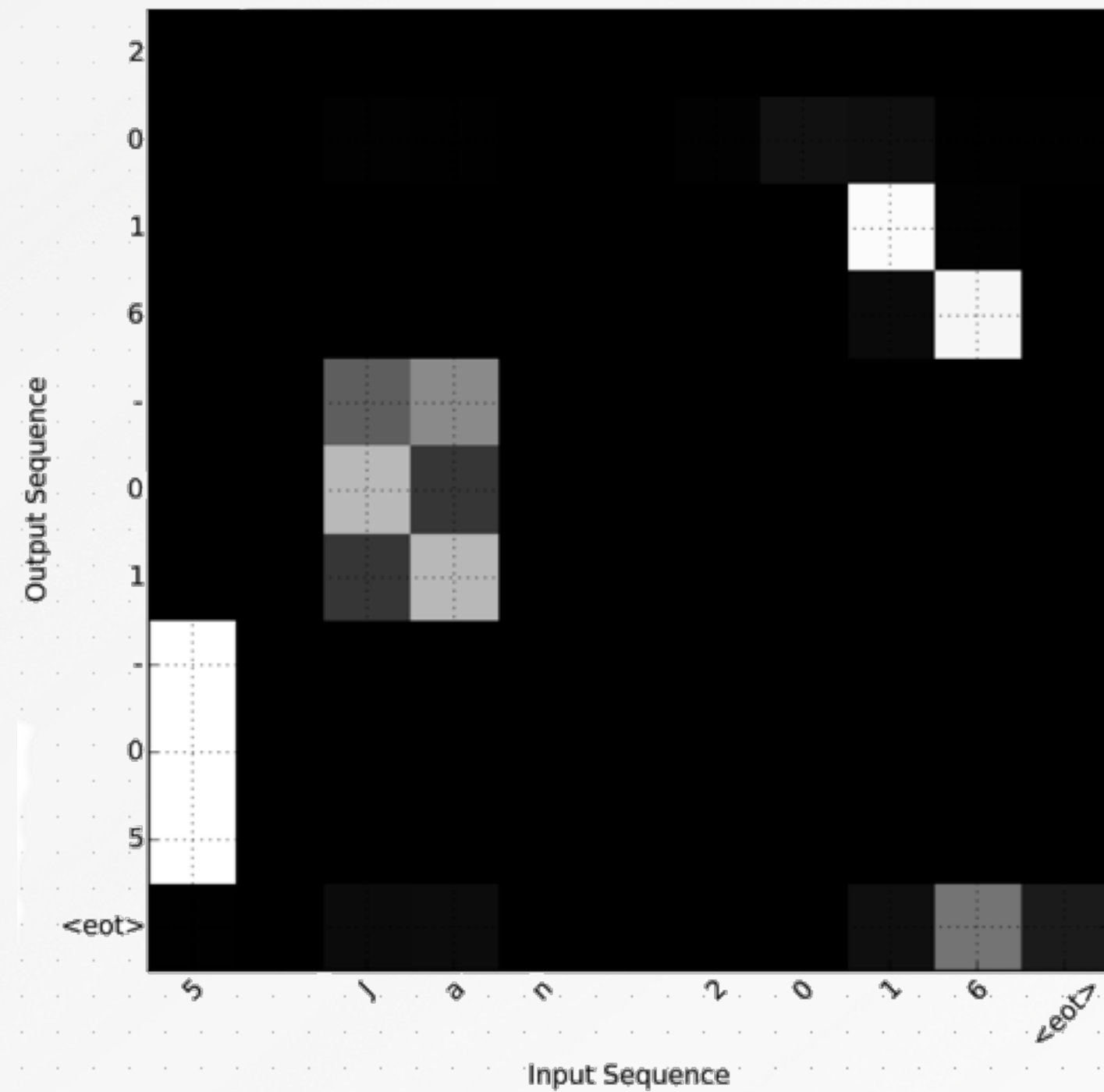
Parsing of
unstructured data

3

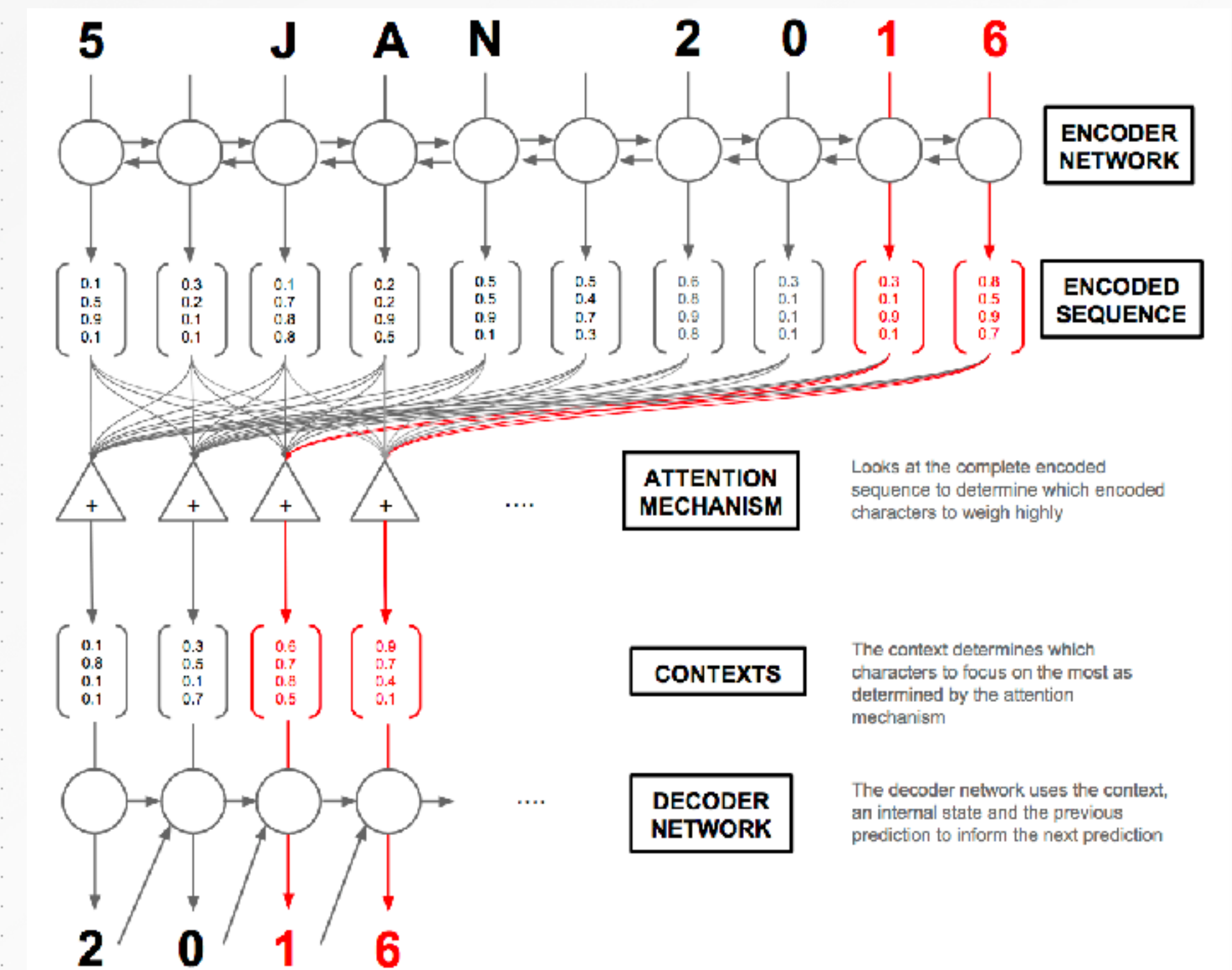
SEQ₂SEQ

Translation of data from one
format to another

Seq2seq



SEQUENCE TO SEQUENCE MODELS TRANSFORM DATAPOINTS TO THE FORMAT OF YOUR CHOOSING



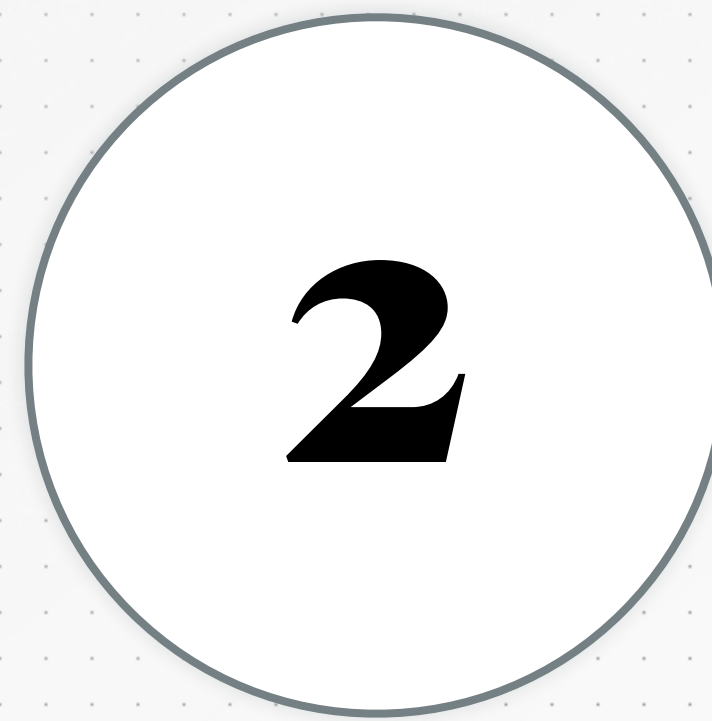
ATTENTION MECHANISMS CREATE AUDITABLE NEURAL NETWORKS FOR TRANSLATION

Deep learning-powered pipelines



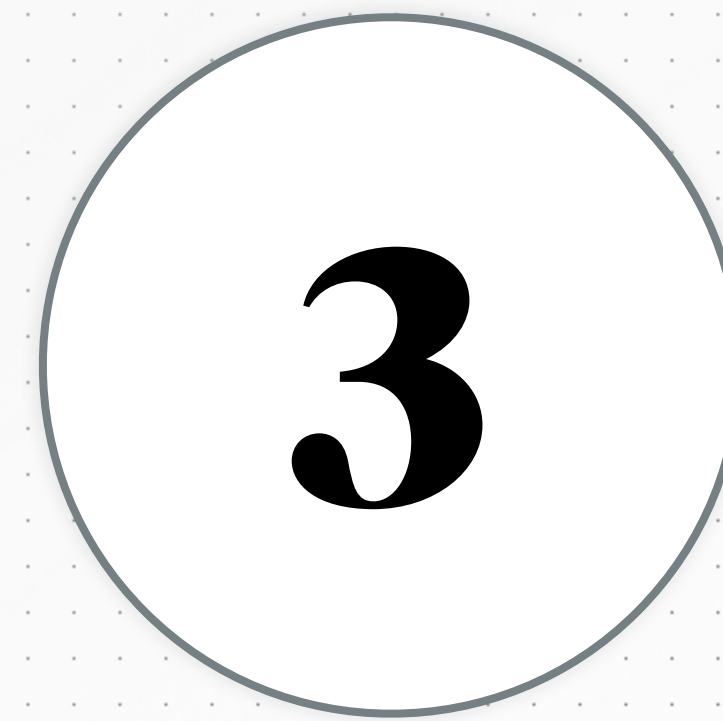
GET

Stream data from
any source



KNOW

Annotate data with
automatically detected
ontologies



TRANSFORM

Transform data to any
isoform and format

Thank you!

sonia@datalogue.io
@your_datalogue