DATALOGUE

Discovering Data Insights Through Neural Networks

DataEngConf

October 31, 2017

Sonia Sen

DATALOGUE

Head of Product

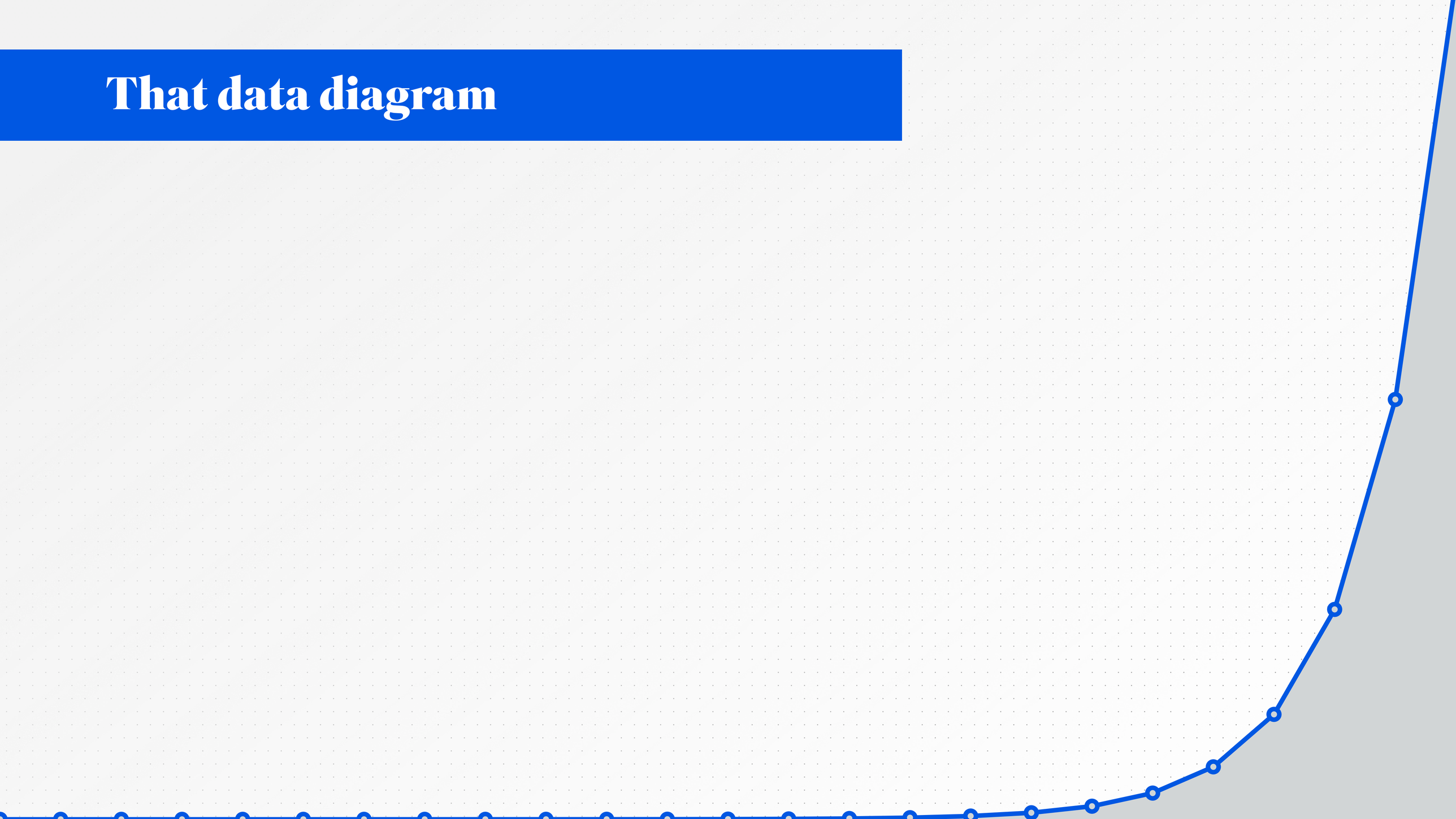
CORNELL TECH

M.S. in Information Science - Health Tech

BOEING

NASA

That data diagram



ETL 1.0

1

GET

Ad hoc connecters, scrapers and queries

2

KNOW

Manual data inspection and tagging (sampled) and regular expressions

3

TRANSFORM

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

ETL 1.0: regular expressions

REGEX TO IDENTIFY

```
([A-ZA-Z]{3,}( I,))(\D*)*\W*I (\D*)I
(\D*)\/(\D*)
```

SCRIPTING

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

ETL 2.0

1

GET

Shared, largely open source connectors

2

KNOW

Regular expression and simple rules based data mapping

3

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.

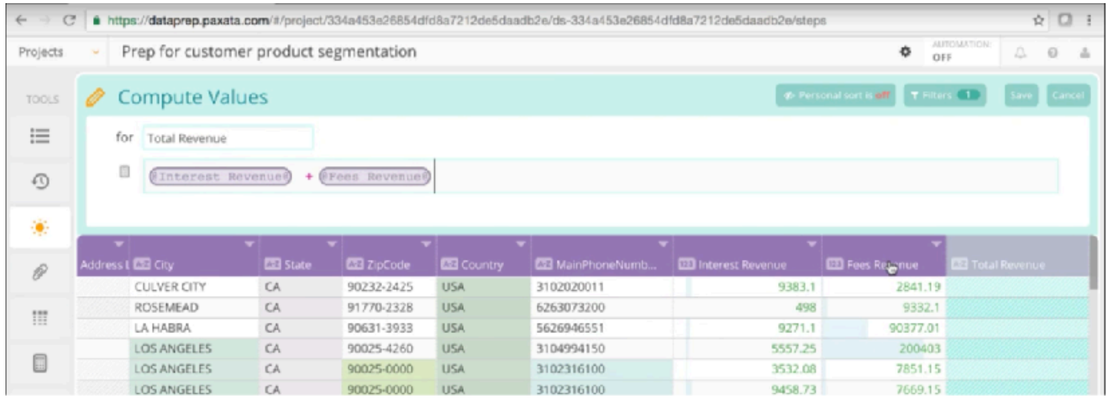
Syntax:

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

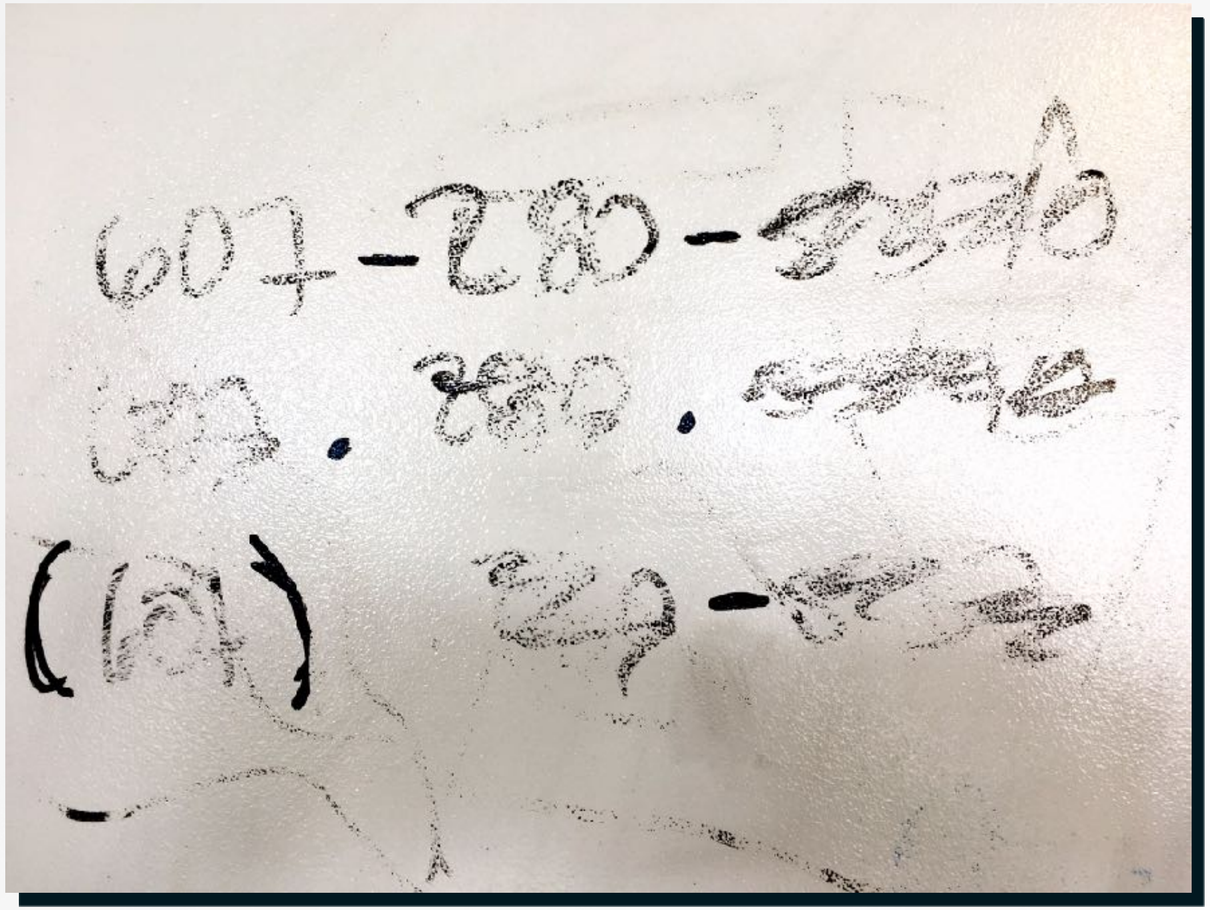
Example:

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

PAXATA



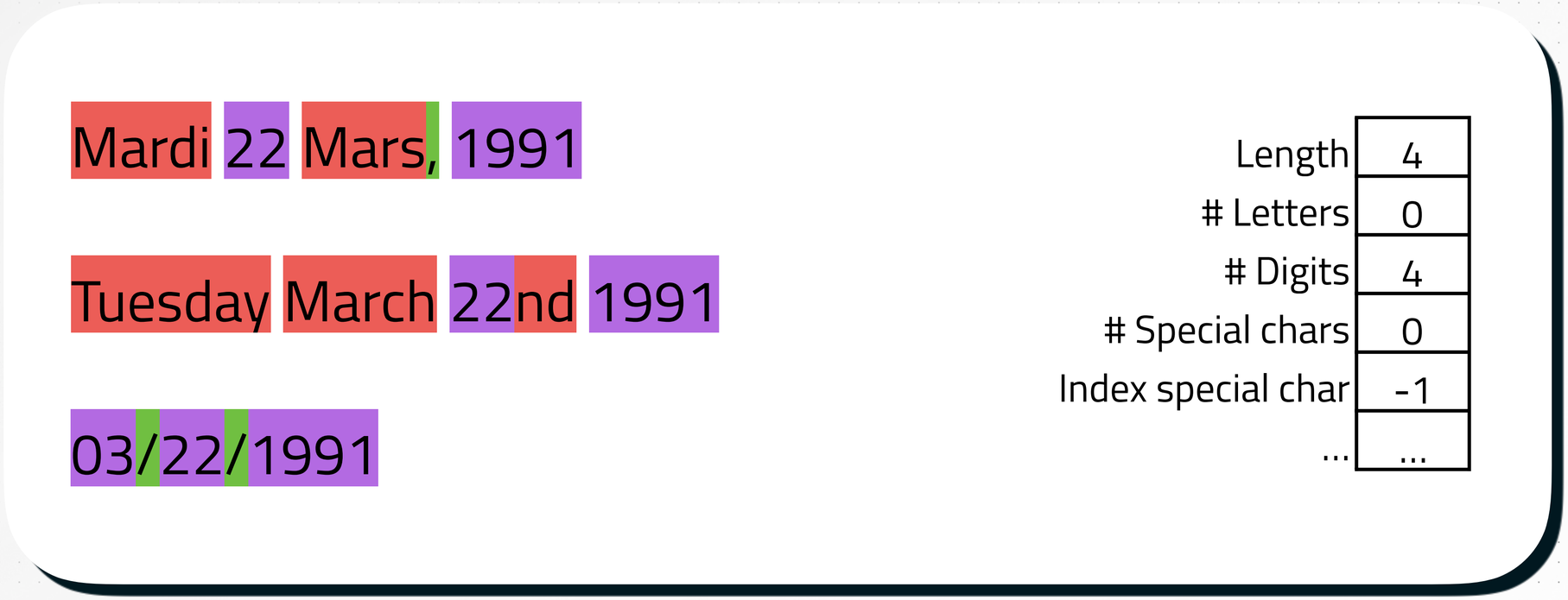
The aha! moment



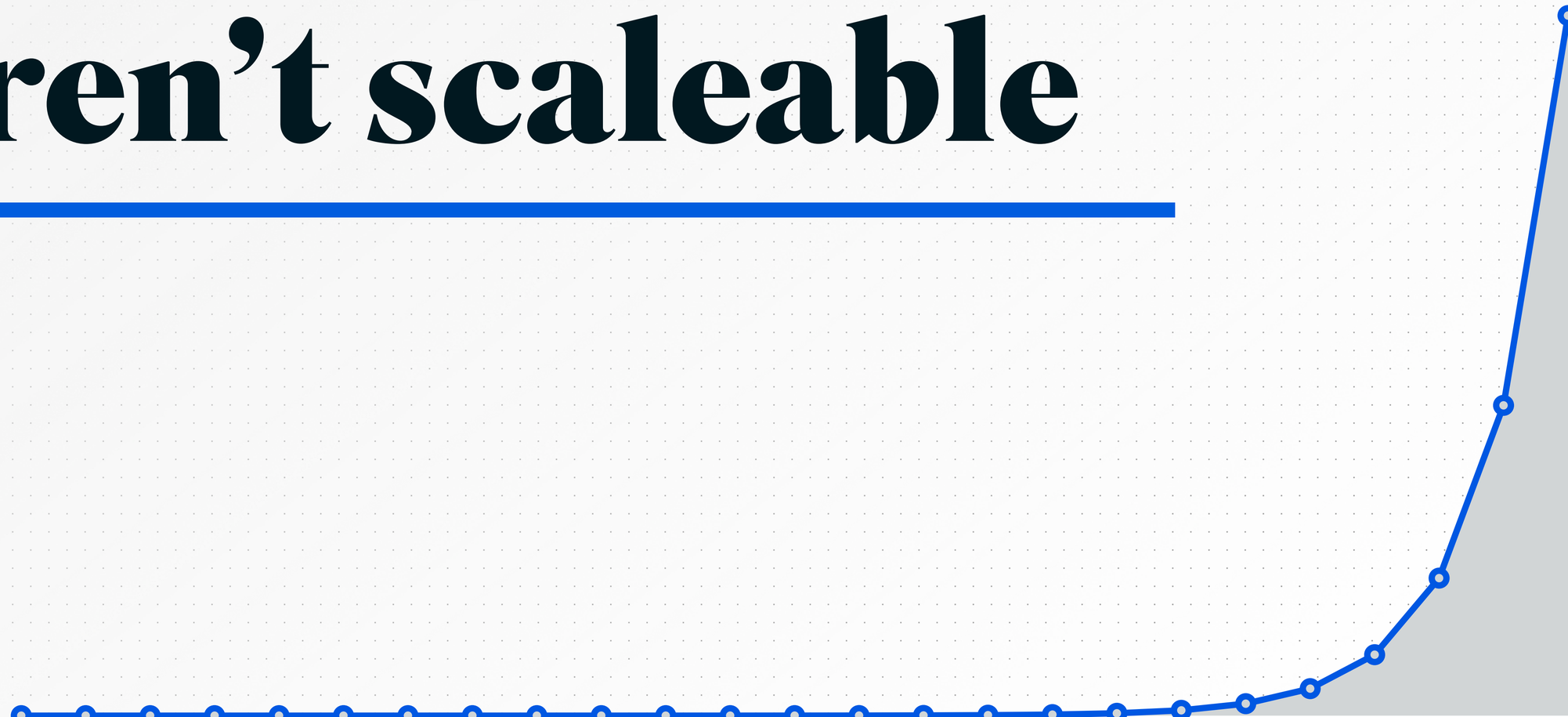
(568) 487-7267
464.211.3990
465-133-3861
173 296 1702
309 612-8634
(754) 480-1720
917.181.8932
(169) 705-8583 ext. 2343
+1 946.456.7514
+1 (281) 229-6261

Regex + machine learning

Dates



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

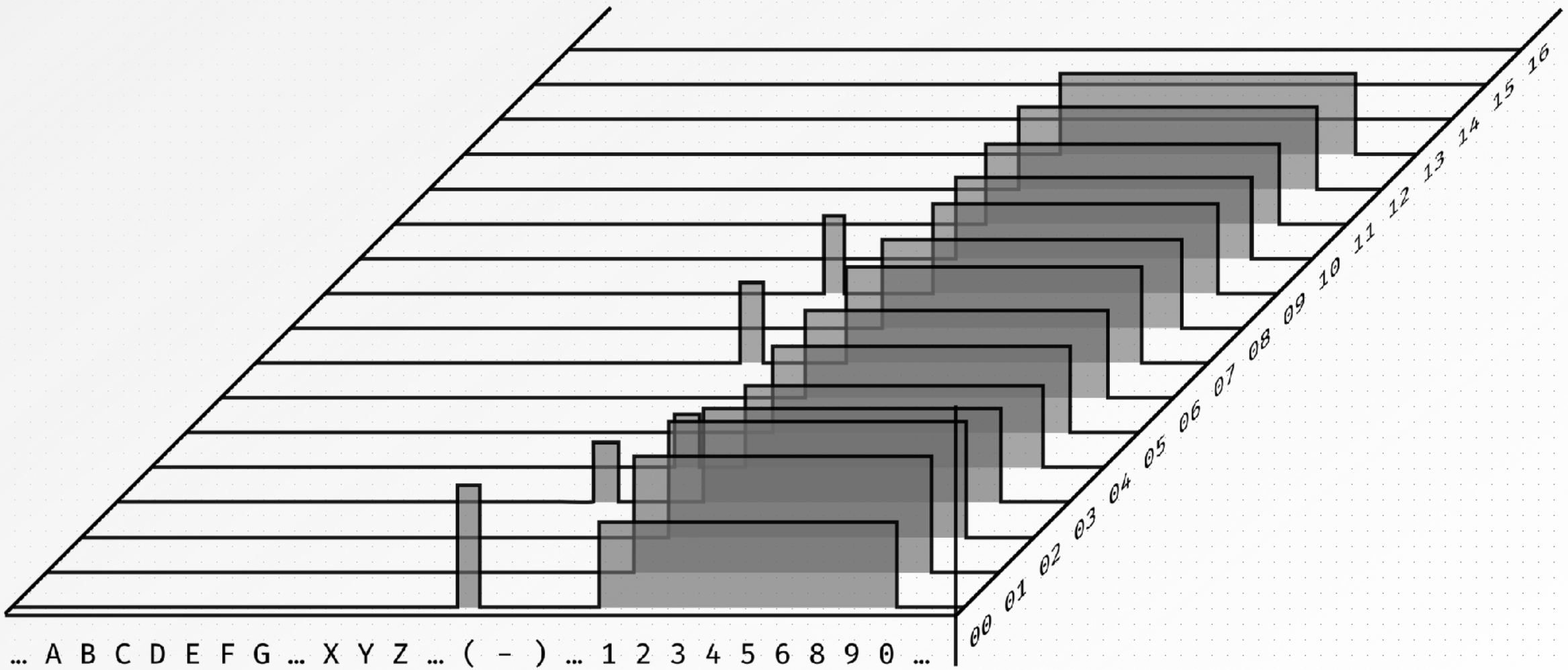
Parsing of unstructured data

3

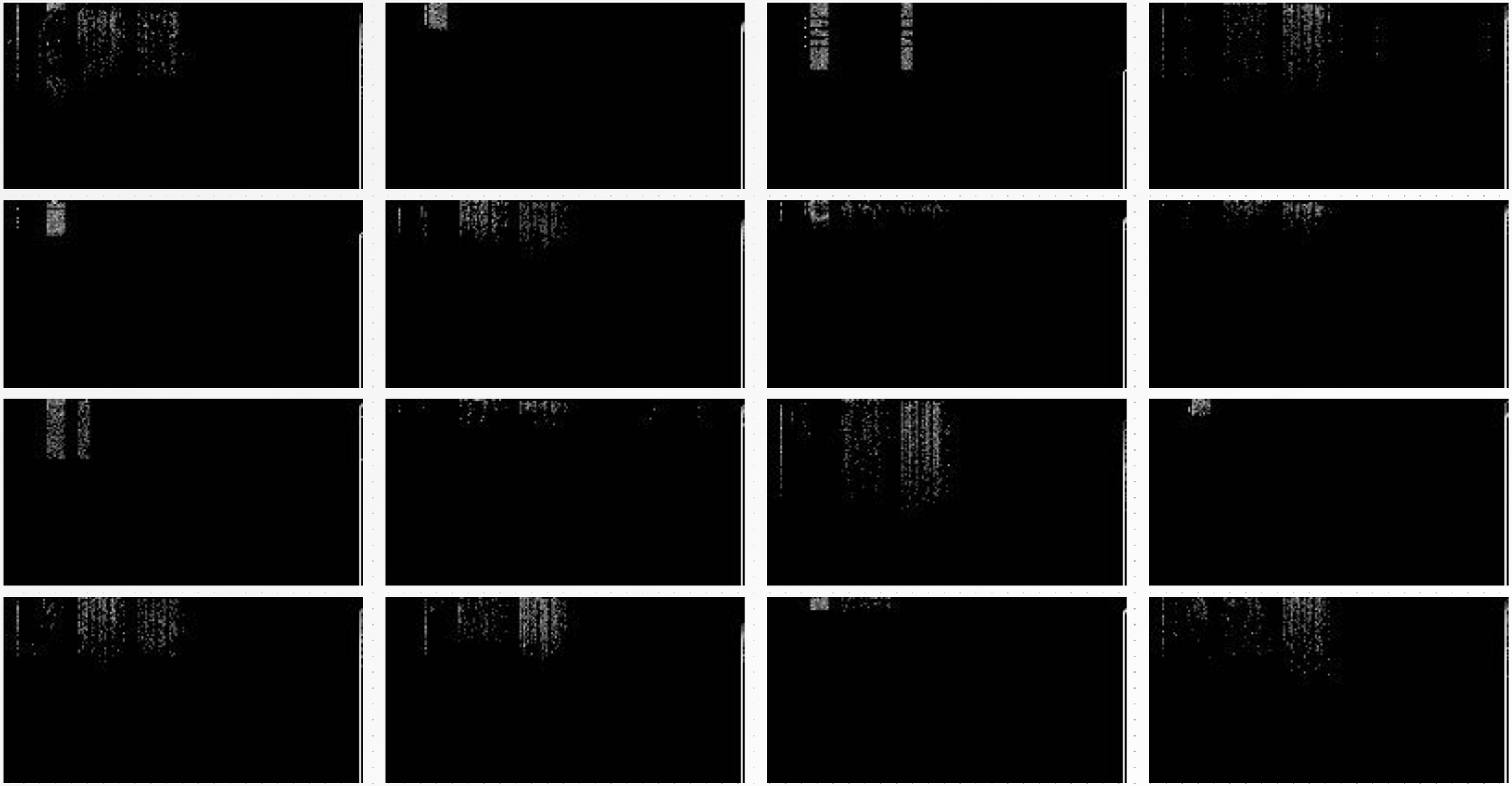
SEQ2SEQ

Translation of data from one format to another

VDCNN inputs

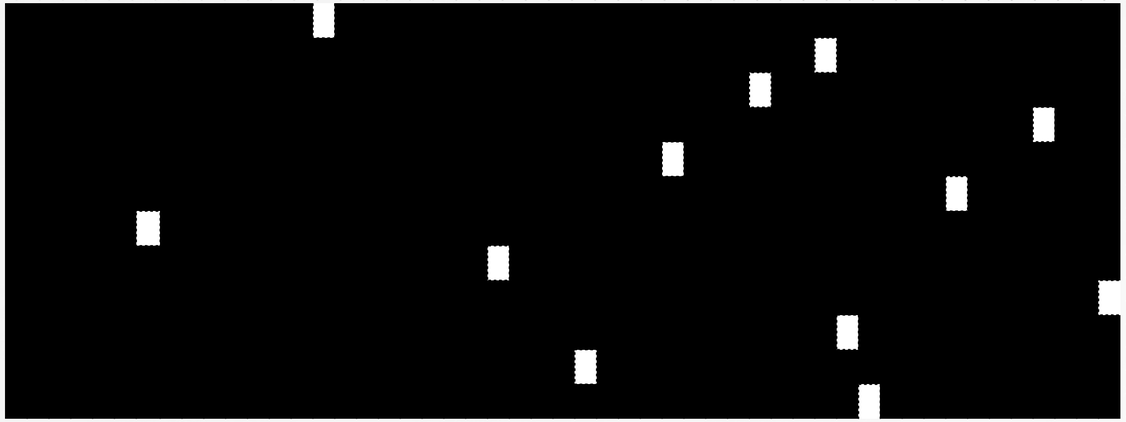


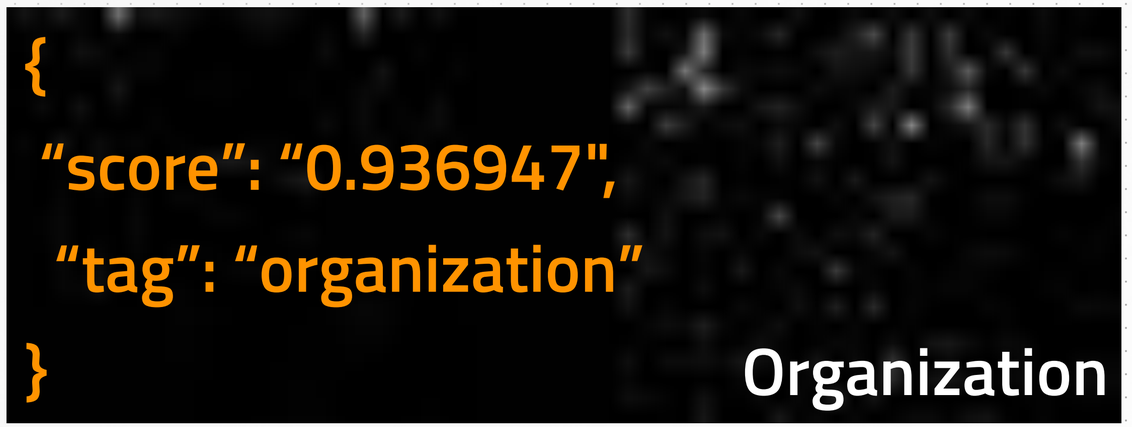
VDCNN inputs

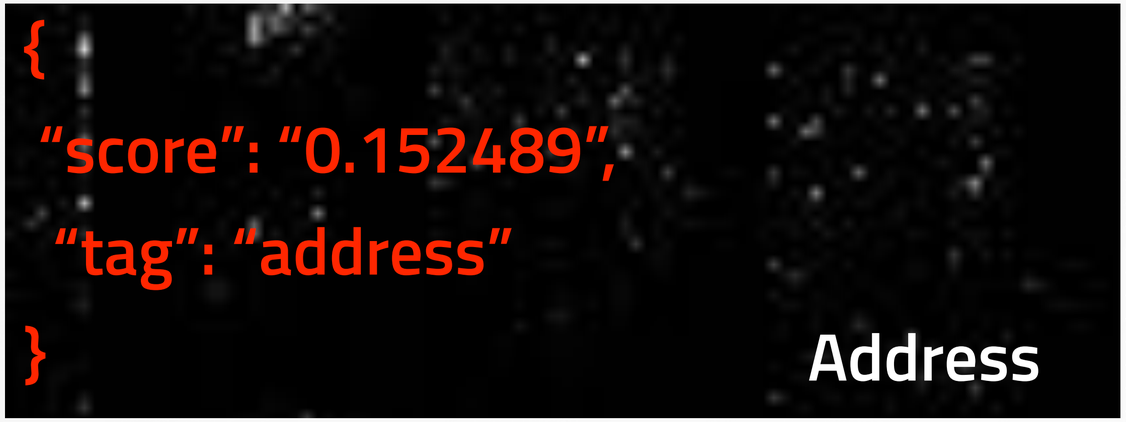


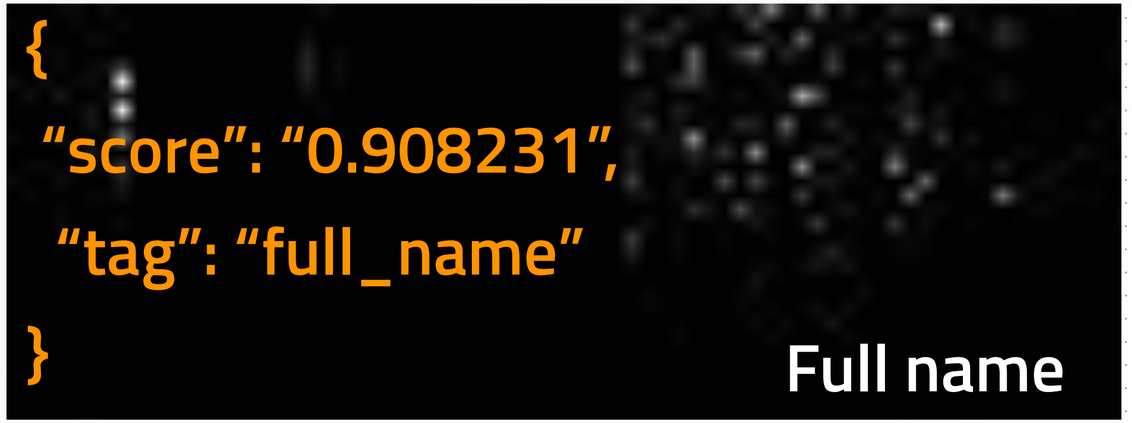
VDCNN outputs

“Oliver Wyman”









Deep learning

1

VDCNN

Semantic + structural understanding

2

LSTM + CRF

Parsing of unstructured data

3

SEQ2SEQ

Translation of data from one format to another

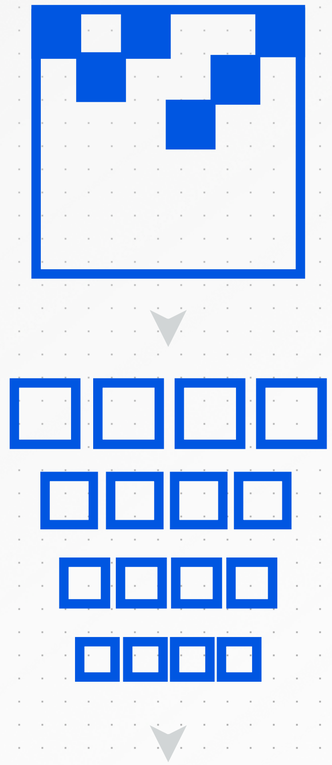
LSTM + CRF

Char Embedding

Bidirectional LSTM

+

CRF



Parsed String

10 Airport Road SE,Salem,NY,97301

AAAAAAAAAAAAAAAAAAUCCCCCUSSUZZZZZ

Deep learning

1

VDCNN

Semantic + structural understanding

2

LSTM + CRF

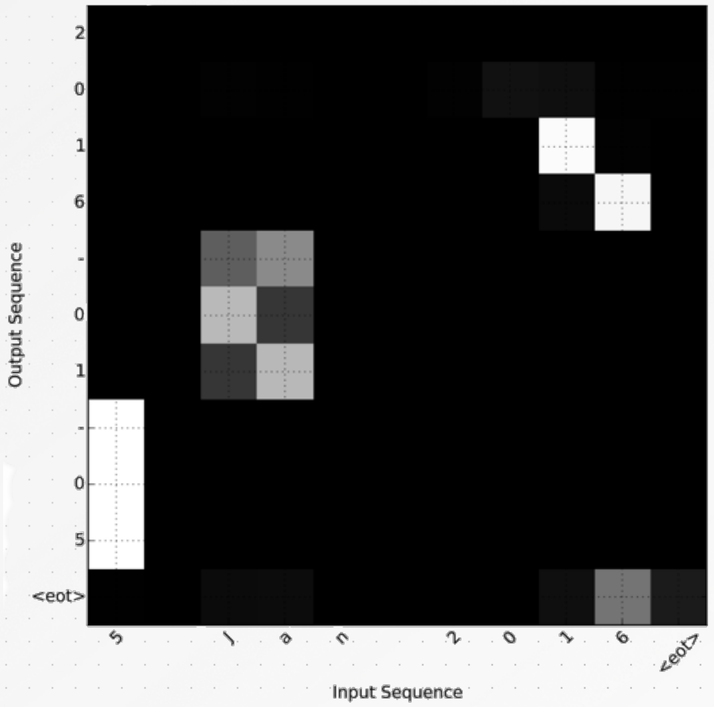
Parsing of unstructured data

3

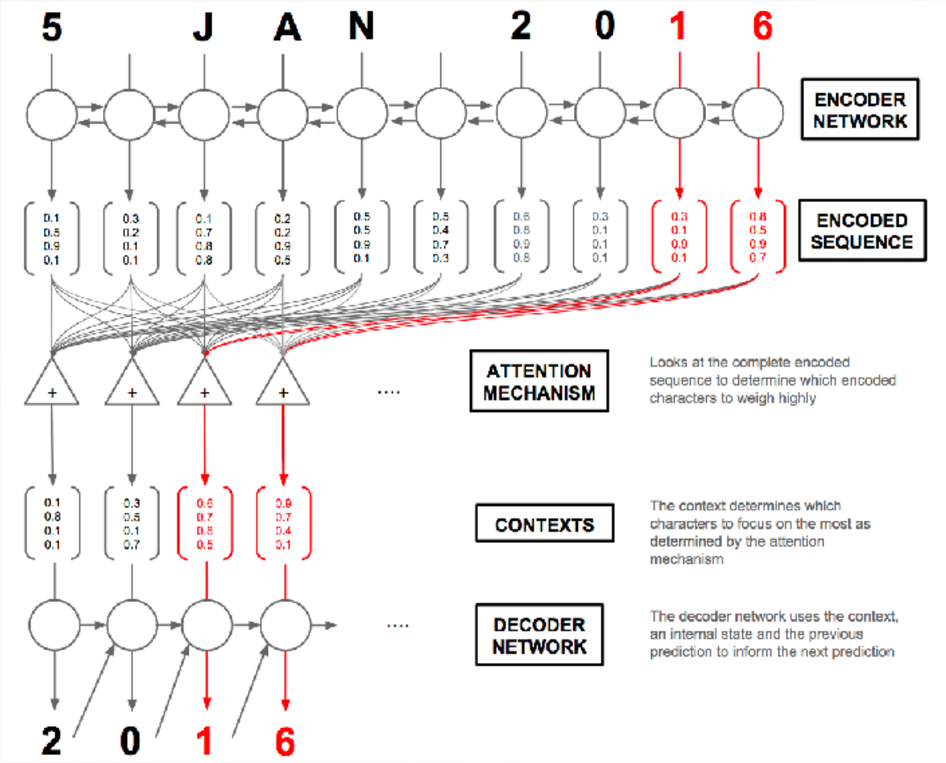
SEQ2SEQ

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

1

GET

Stream data from any source

2

KNOW

Annotate data with automatically detected ontologies

3

TRANSFORM

Transform data to any isoform and format

DATALOGUE

Thank you!

sonia@datalogue.io

@your_datalogue