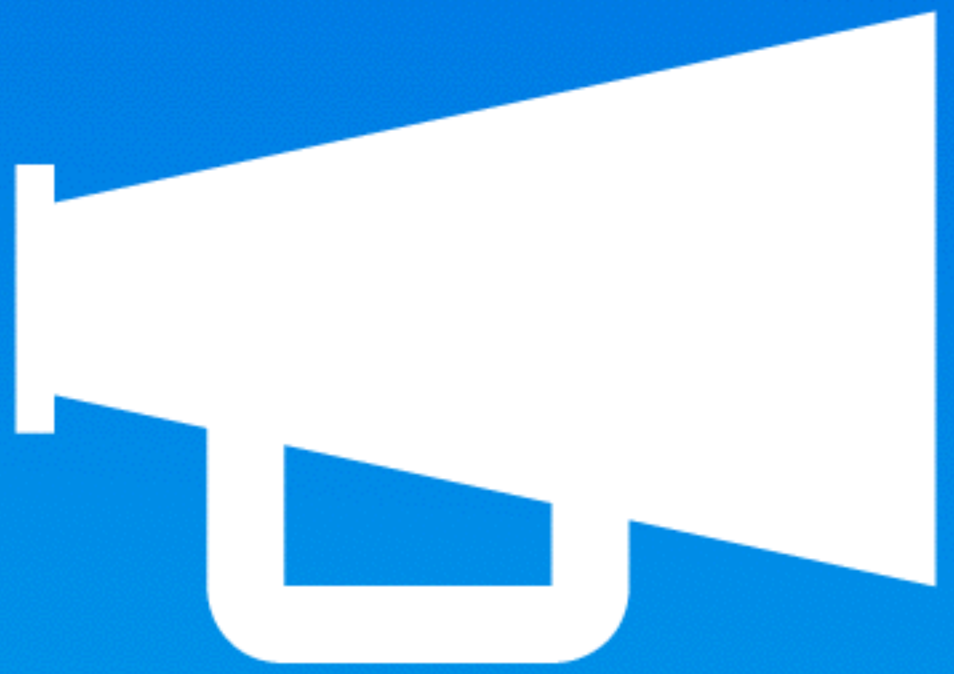
Modern Fraud Prevention using Deep Learning

Phil Winder

1430 CET Scandic Grandball
6th October 2015

goto; conference



Click ‘engage’
to rate sessions
and ask questions

Introduction

Phil Winder

Engineer at Trifork Leeds

Current project:
  Elasticsearch framework for Apache Mesos

pnw@trifork.com
@DrPhilWinder

Line Christa Amanda Sørensen

- Group COO
- las@trifork.com

Tom Benedictus

- Trifork Leeds CEO
- tob@trifork.com

Trifork

We

	make	teach	advise
	apps	agile	NoSQL

- 6,000+ attended our conferences in 2014
- 30+ companies worldwide
- 400+ employees
- 30,000,000+ revenue

Trifork in finance and beyond

- CMS
- Custom Solutions
- Internet of Things
- Mobile
- NoSQL and Search
- Academy

Outline

Background	Machine learning	Demos	Architectures
1	2	3	4

https://github.com/philwinder/MortgageMachineLearning

Introduction

Background	Machine learning	Demos	Architectures
1	2	3	4

Introduction: Financial crime

Serious Fraud Office

“Put simply, fraud is an act of deception intended for personal gain or to cause a loss to another party.”

UK *Current account* fraud

“151 in every 10,000” [2]

“69% due to identity theft” [2]

UK *Mortgage* Fraud

1.2 Million residential properties sold in 2014 [1]

“83 in every 10,000 mortgage applications were found to be fraudulent” [2]

Approximately **£1B** in fraudulent applications. [3]

UK *Retail* fraud

“SMBs are losing £18bn every year to fraudulent transactions” [4]

[1] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/461354/UK_Tables_Sep_2015__cir_.pdf
[2] http://www.experian.co.uk/blogs/latest-thinking/dramatic-increase-current-account-fraud/
[3] http://www.moneywise.co.uk/news/2013-05-16/average-outstanding-uk-mortgage-100000
[4] http://www.retailfraud.com/fraud-costs-uk-smbs-18bn-a-year/

Introduction: Legislation

2017 AML legislation

- Businesses: credit, finance, legal and financial services, gambling, anyone facilitating transactions over 10,000 EUR
- Major changes:
  - Maximum “out of scope” limit dropped to 1,000 EUR
  - Must prove “due diligence”
  - Public central registry of business information

[1] DIRECTIVE (EU) 2015/849 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 20 May 2015 on the prevention of the use of the financial system for the purposes of money laundering or terrorist financing, amending Regulation (EU) No 648/2012 of the European Parliament and of the Council, and repealing Directive 2005/60/EC of the European Parliament and of the Council and Commission Directive 2006/70/EC

Introduction: Common technologies

Origination based

Verifies identity. Some practices are very poor, e.g. services verifying identity using DOB.

Rules based

Static set of rules searching for very specific patterns. Very poor accuracy.

Credit checks

Expensive services that aim to provide risk profile. Fraudsters are easily able to overcome credit checks.

Aggregation and monitoring

A reactive, but worthwhile solution. E.g. many payments from same account, large transactions, etc.

Machine Learning

Background	Machine learning	Demos	Architectures
1	2	3	4

ML: How humans learn

How do we learn?

Time

Many diverse tasks
But it takes time

Practise

Requires practise
Repetition of tasks
New examples

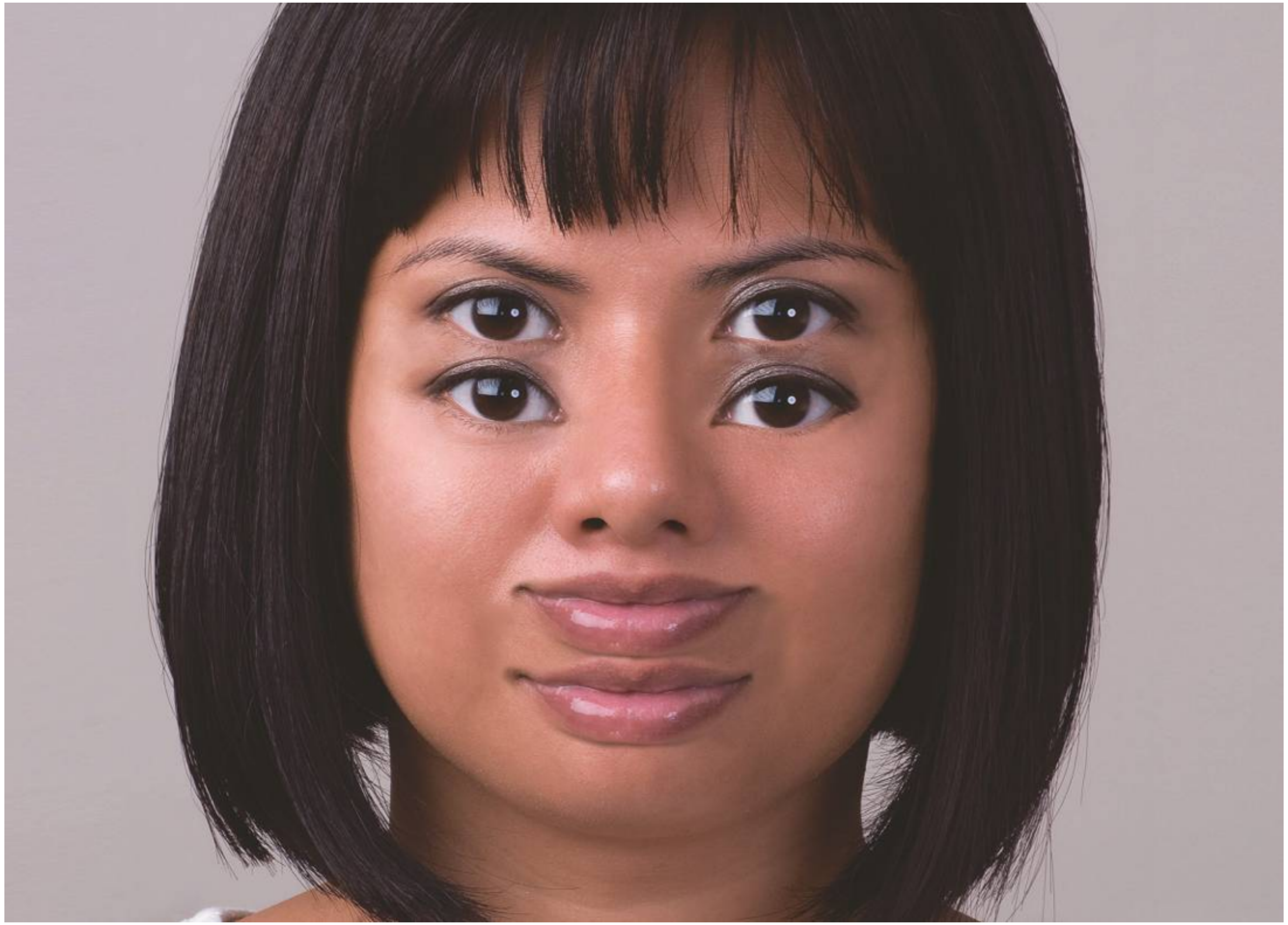
ML: How humans get it wrong

Misuse of features

Misclassification

Bad data

ML: How humans get it wrong

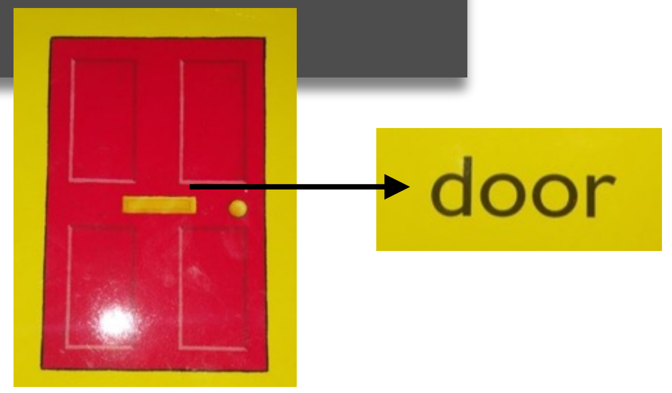


http://visitcanberra.com.au/events/9005967/perception-deception

ML: Main categories of algorithms

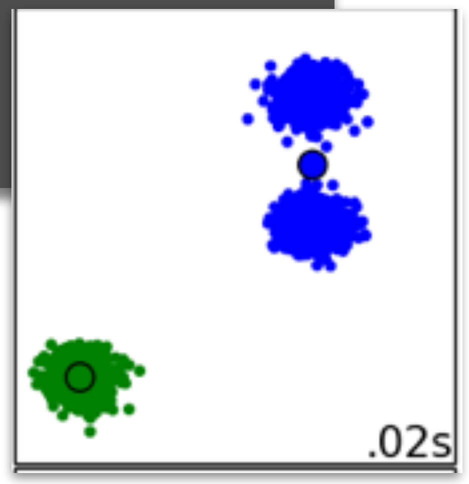
Dimensionality reduction

Curse of dimensionality
Reduce number of inputs



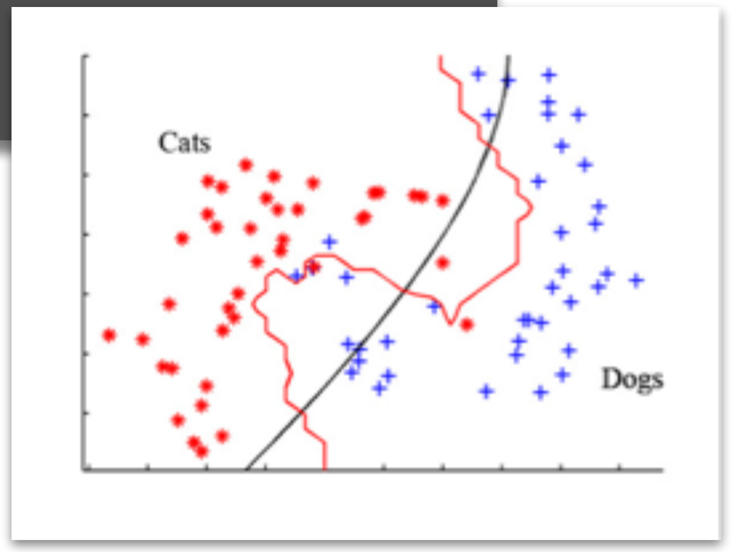
Clustering

Assign output to a class



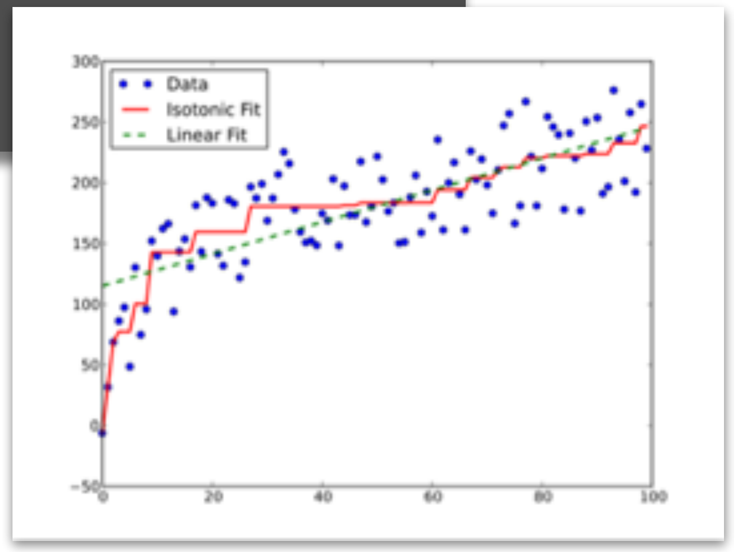
Classification

Decide to which class an input belongs



Regression

Predict value given input



ML: Supervised vs. Unsupervised

Training

Supervised

Expected result is provided

Algorithm is trained to produce the correct result

New data is classified according to the training

Unsupervised

No result is expected

Algorithm is trained so that:

- Similar data are "close"
- Dissimilar data is "far"

Generally, new data is specified as belonging to a group

Semi-Supervised

Some results are provided

Users interact with unsupervised data to find new results

ML: Decision trees

What are they?

Classifier & Regression

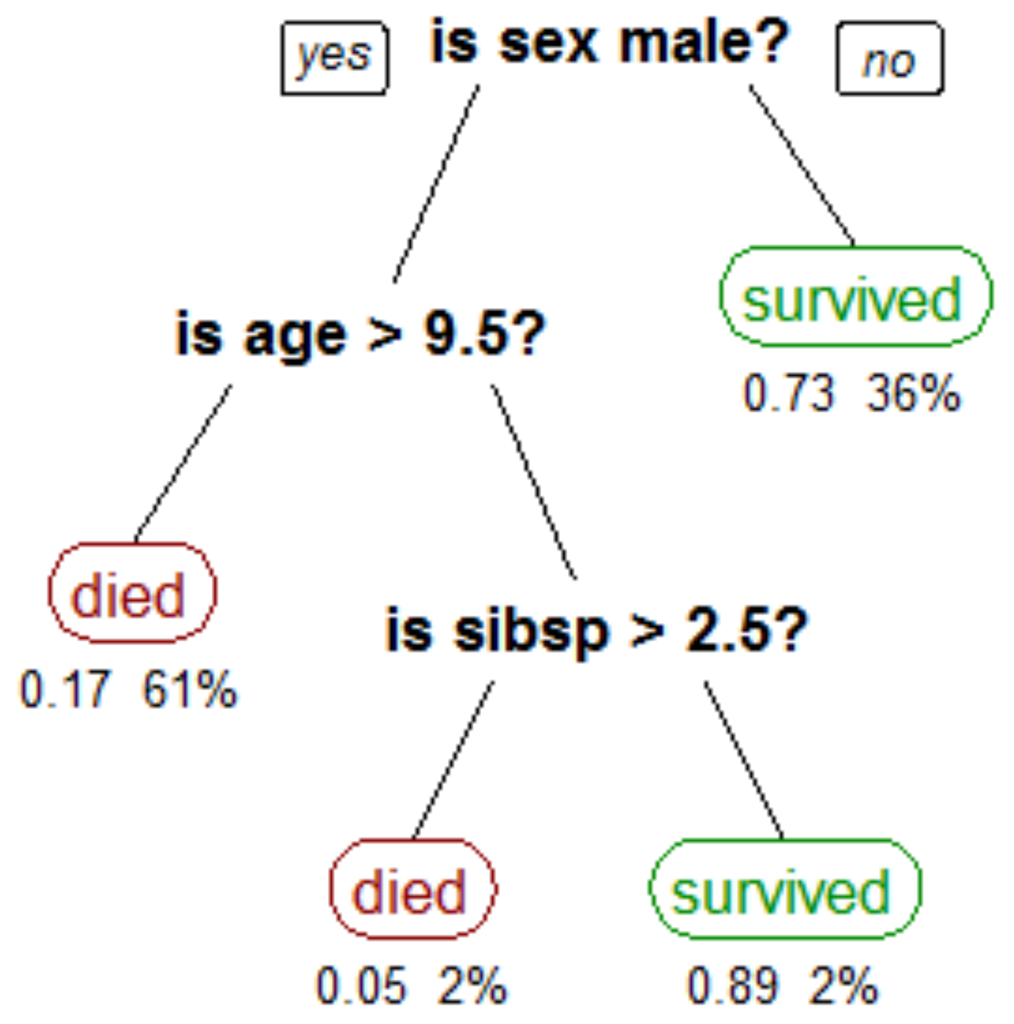
Predict value of target by learning simple decision rules

Pros & Cons

Conceptually simple

Handle categorical data

Overfitting



https://en.wikipedia.org/wiki/Decision_tree_learning

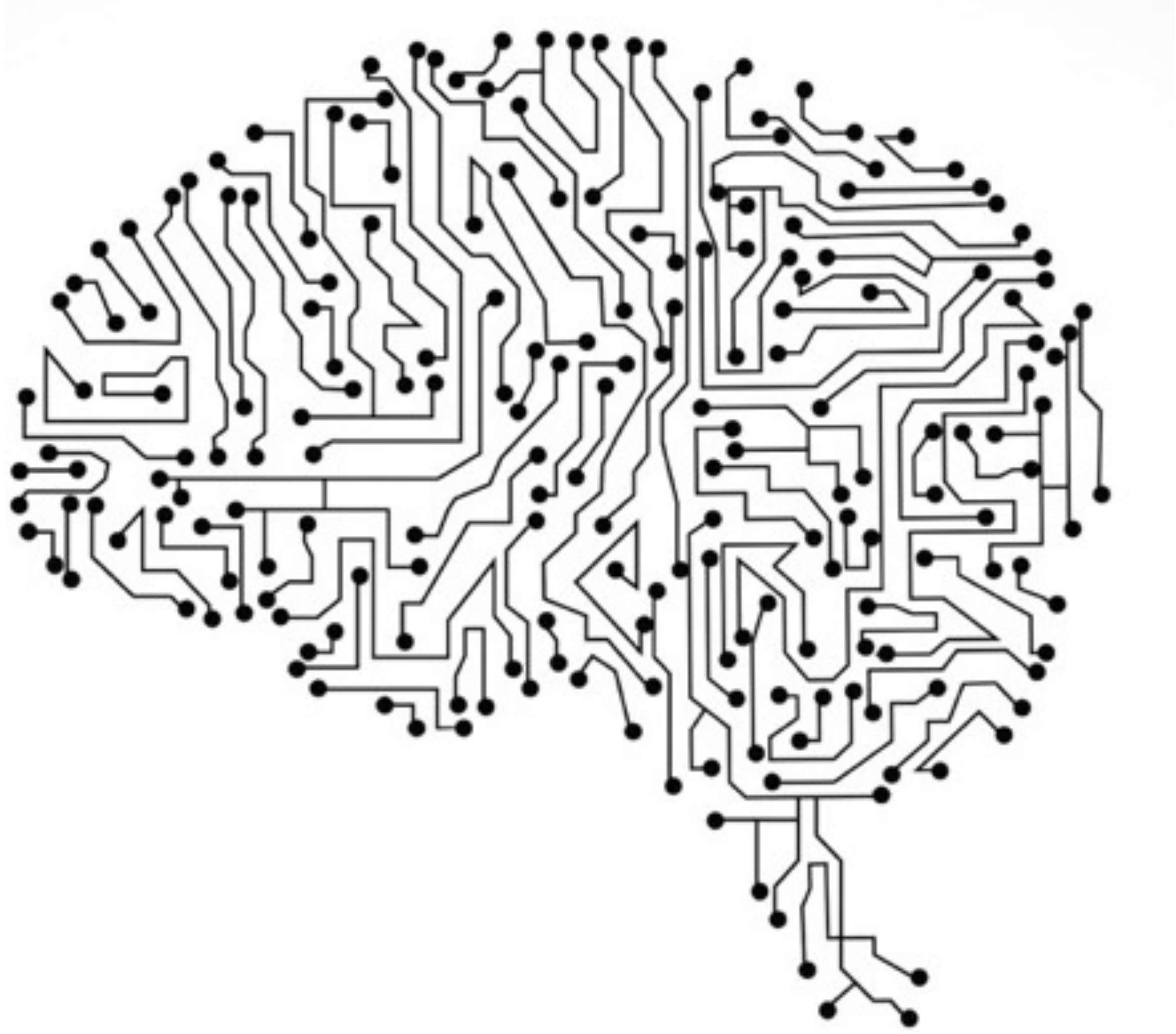
ML: Deep learning

What is deep learning?

What is it?

Dimensionality reduction, classifier, regression & clustering.

Attempts to mimic human brain. Modelled by neurons and weights.

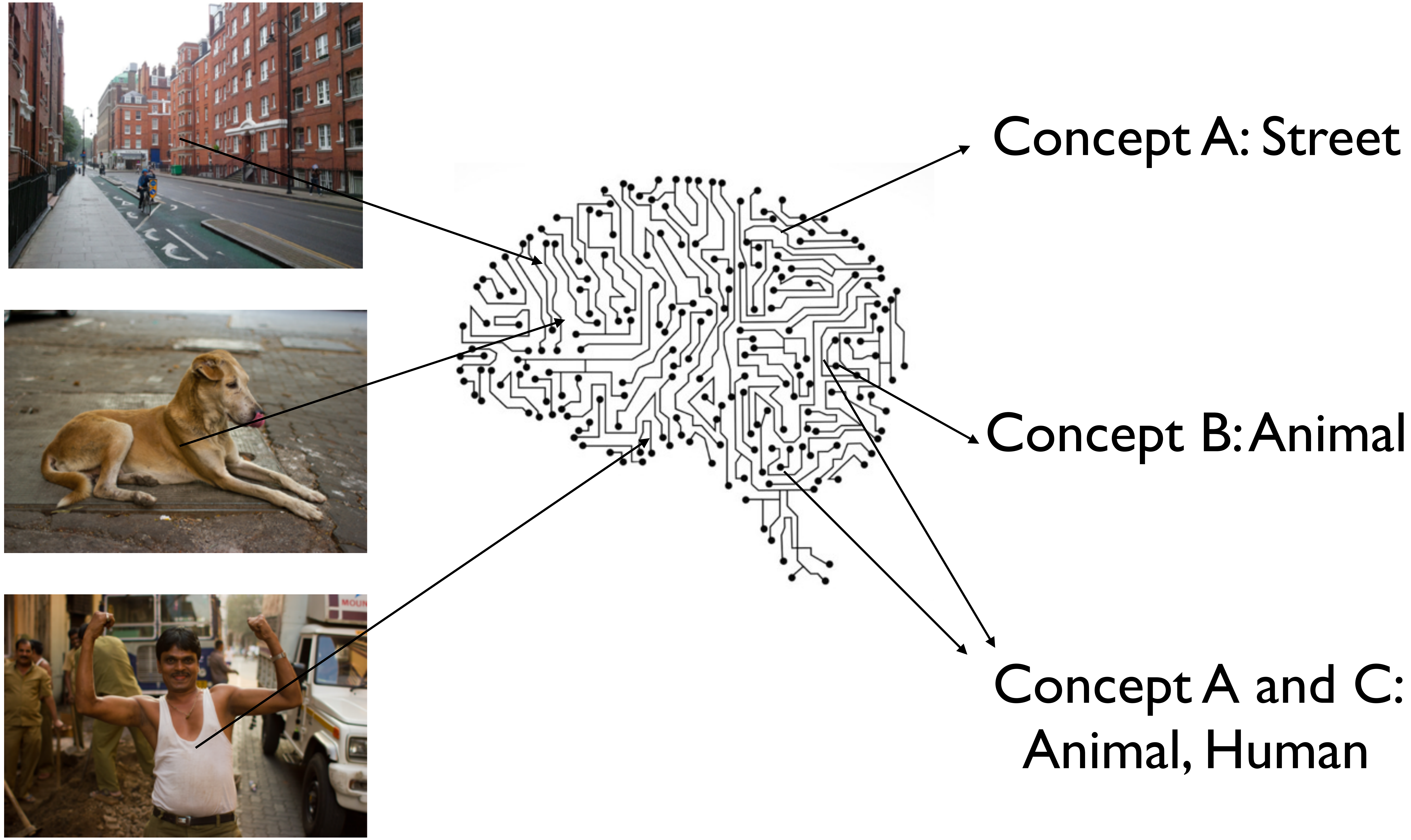


Pros & Cons

- Versatile
- Automated feature engineering
- Hard to visualise

ML: Deep learning

What is deep learning?

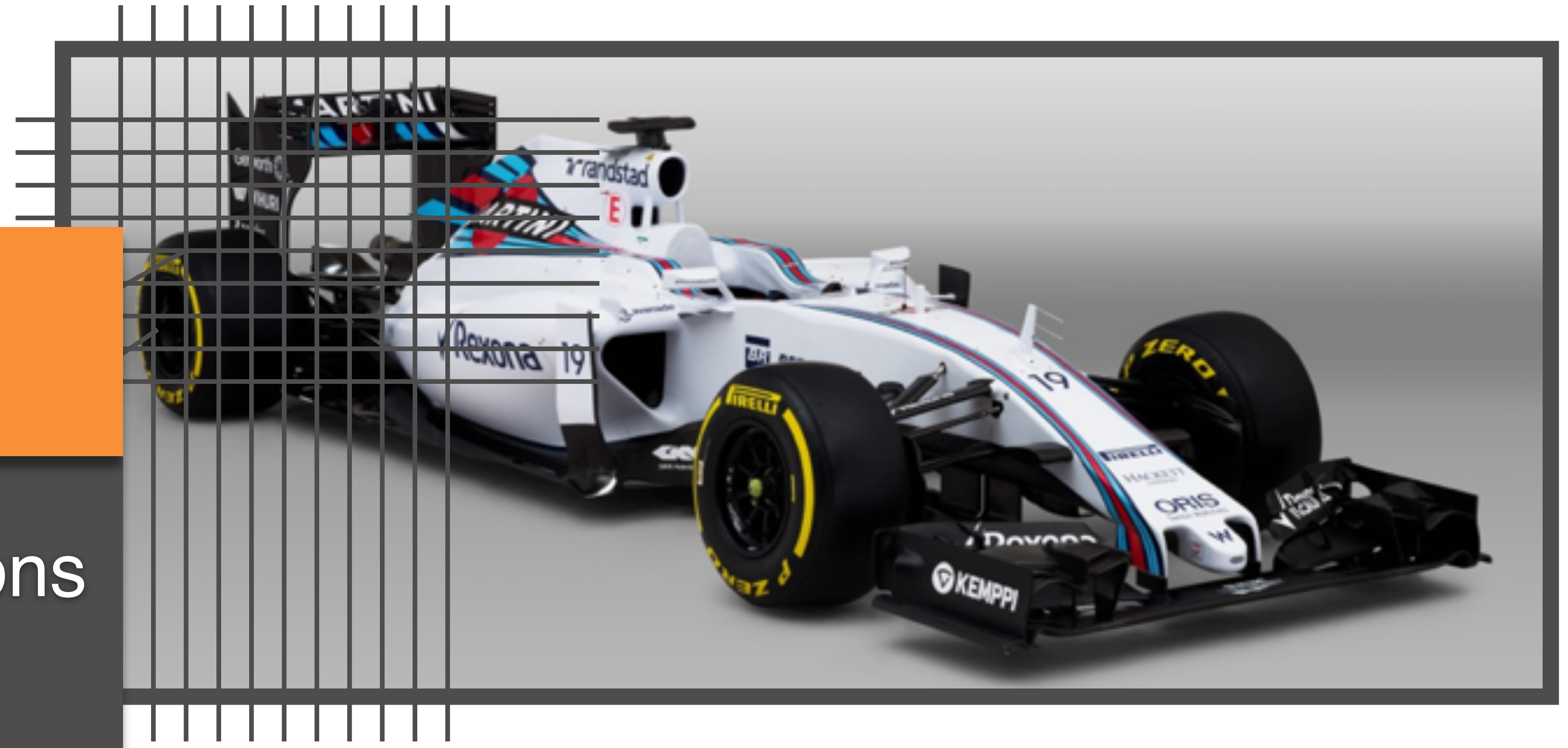


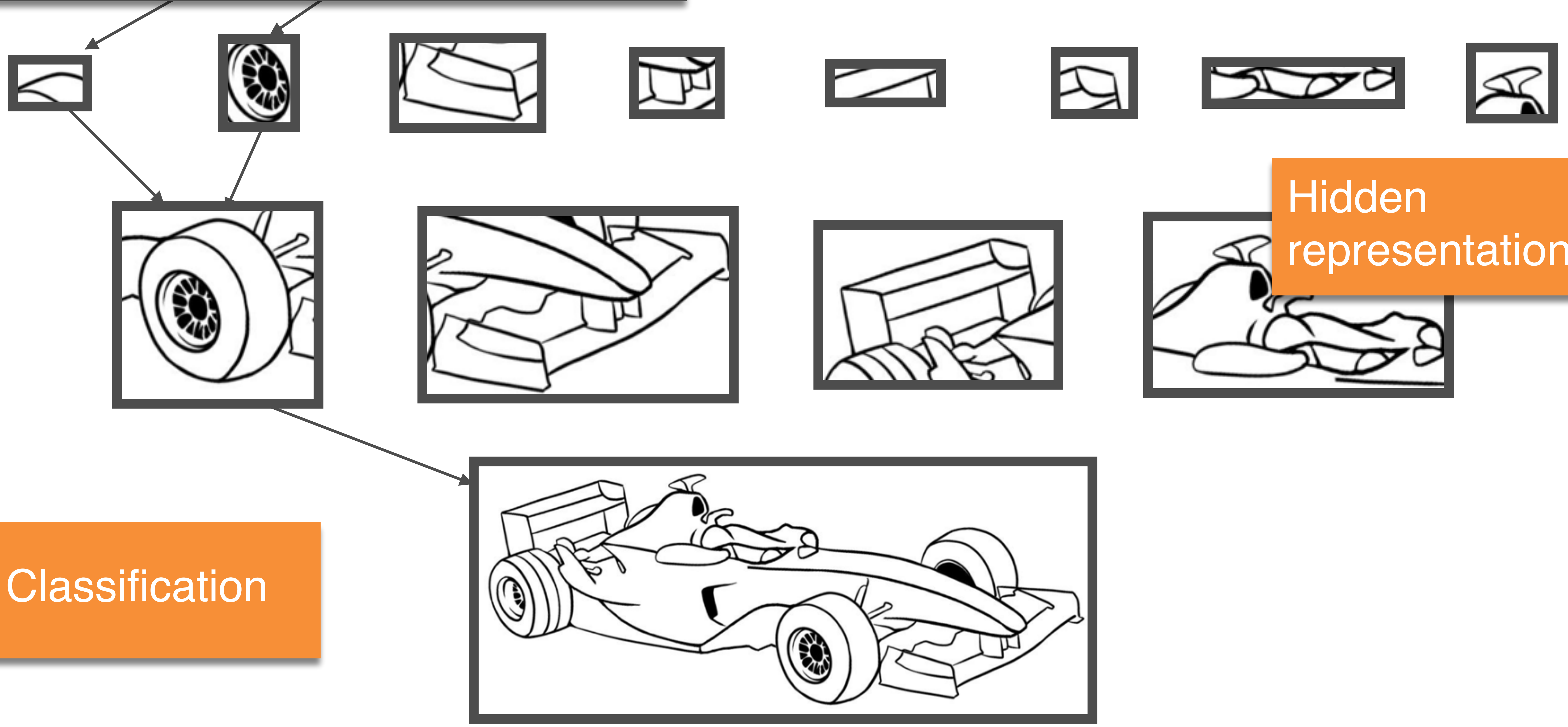
ML: Deep learning

A simple graphical example

How does it work?

- Attempts to model high level abstractions using a cascade of transformations





Hidden representation

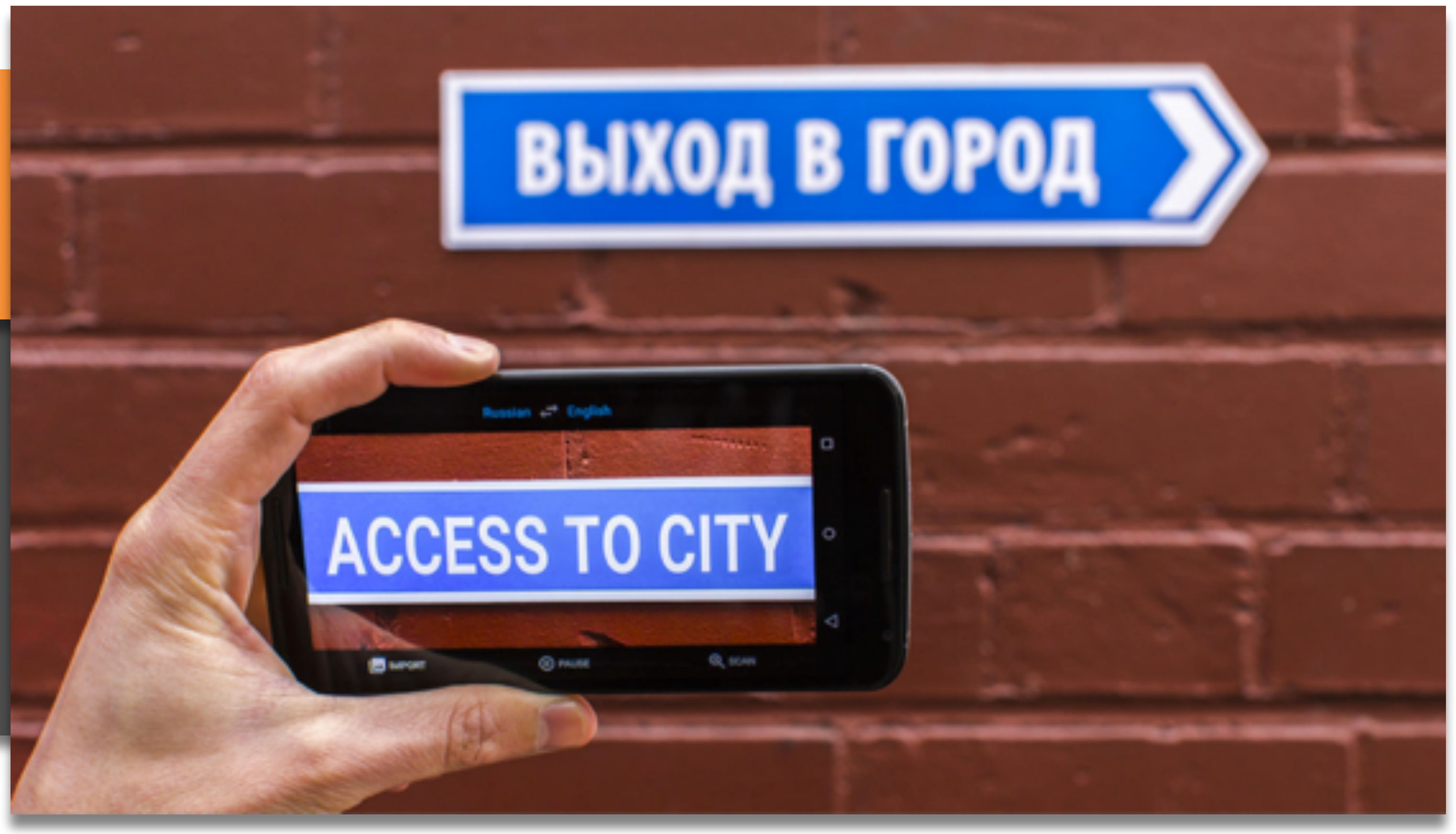
Classification

Machine Learning (ML)

"Machine learning explores the study and construction of algorithms that can learn from and make predictions on data." [1]

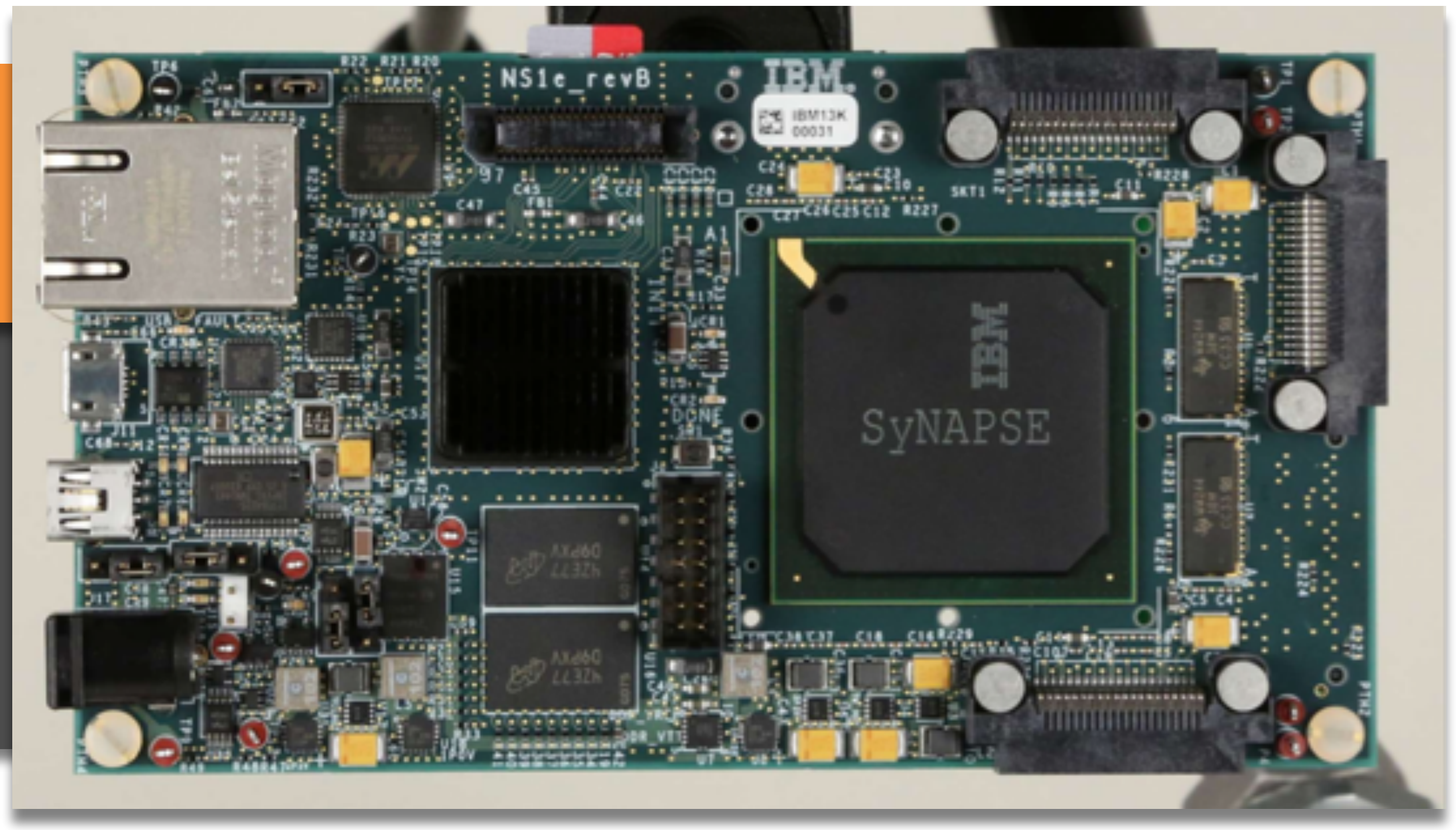
Google

- Google uses deep learning in phones for translation
- http://googleresearch.blogspot.co.uk/2015/07/how-google-translate-squeezes-deep.html?m=1



IBM

- IBM creates deep learning chip
- http://www.wired.com/2015/08/ibms-rodent-brain-chip-make-phones-hyper-smart/



[1] Ron Kohavi; Foster Provost (1998). "Glossary of terms". Machine Learning 30: 271–274.

ML: Deep learning demo

A simple graphical example

http://keras.io/

ML: Deep learning demo

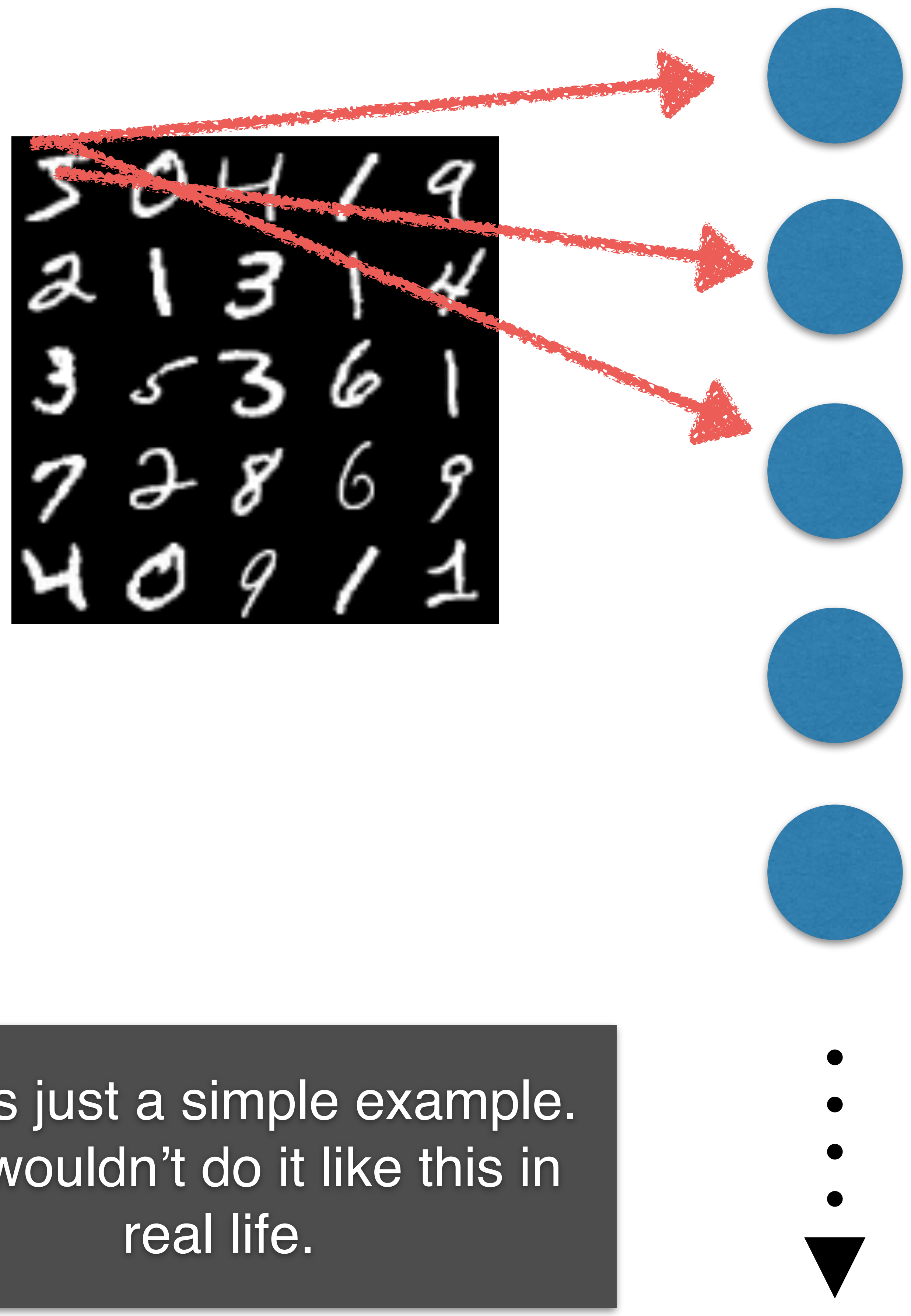
A simple graphical example

Is it a 3 or a 5?

ML: Deep learning demo

A simple graphical example

Input layer

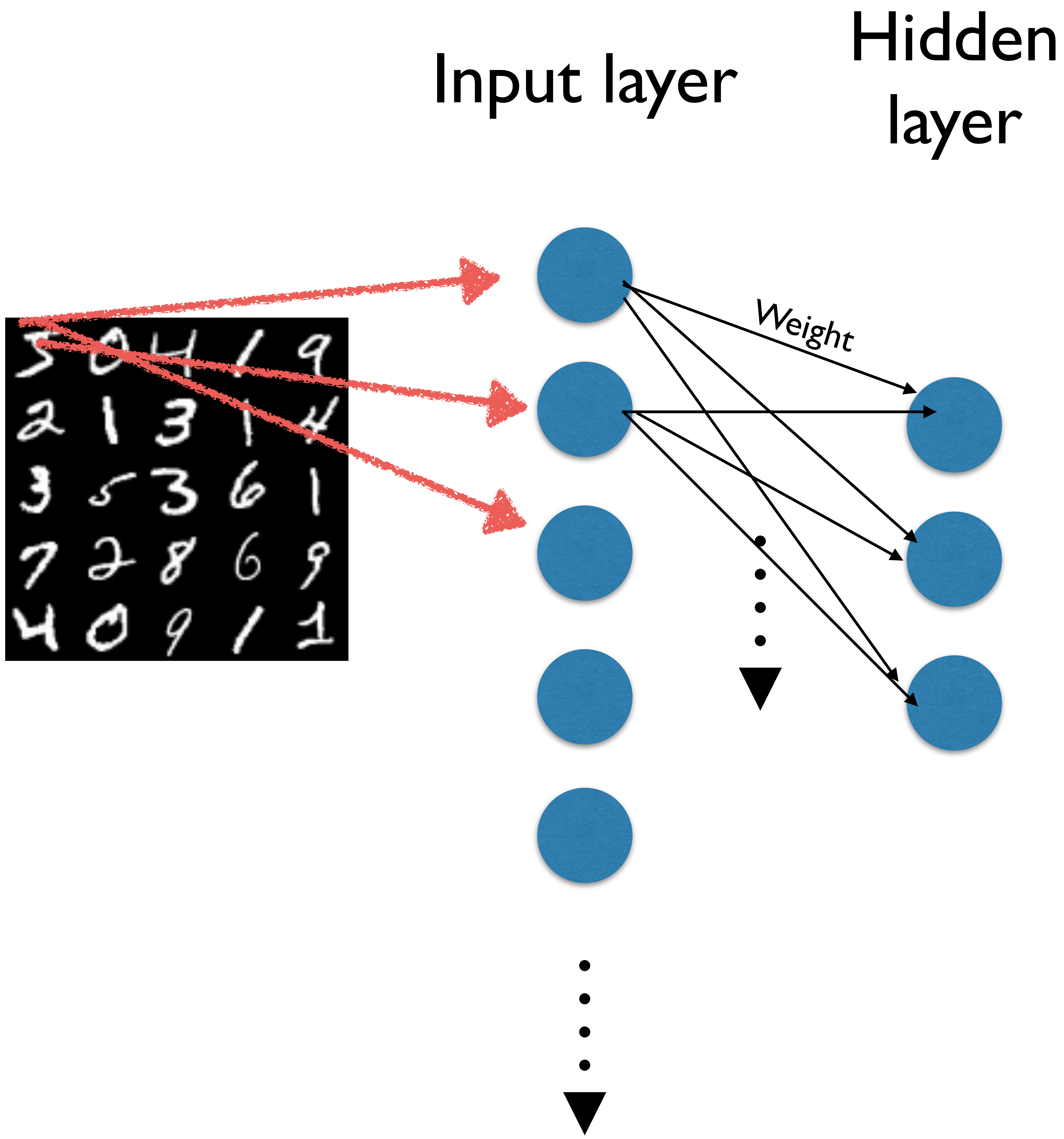


Each pixel is mapped to an input neuron

This is just a simple example. You wouldn't do it like this in real life.

ML: Deep learning demo

A simple graphical example

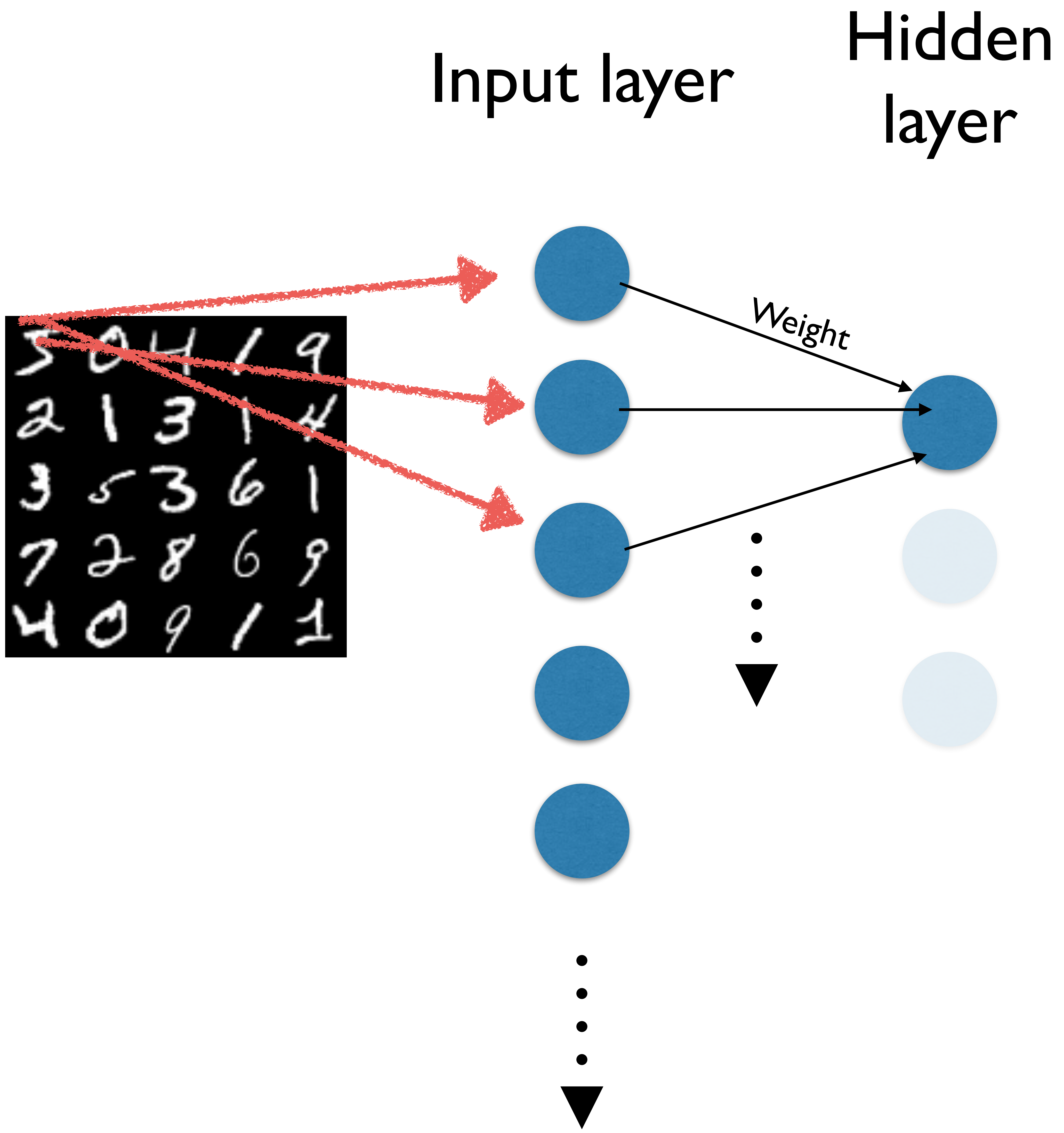


@DrPhilWinder

TRIFORK.
...think software

ML: Deep learning demo

A simple graphical example

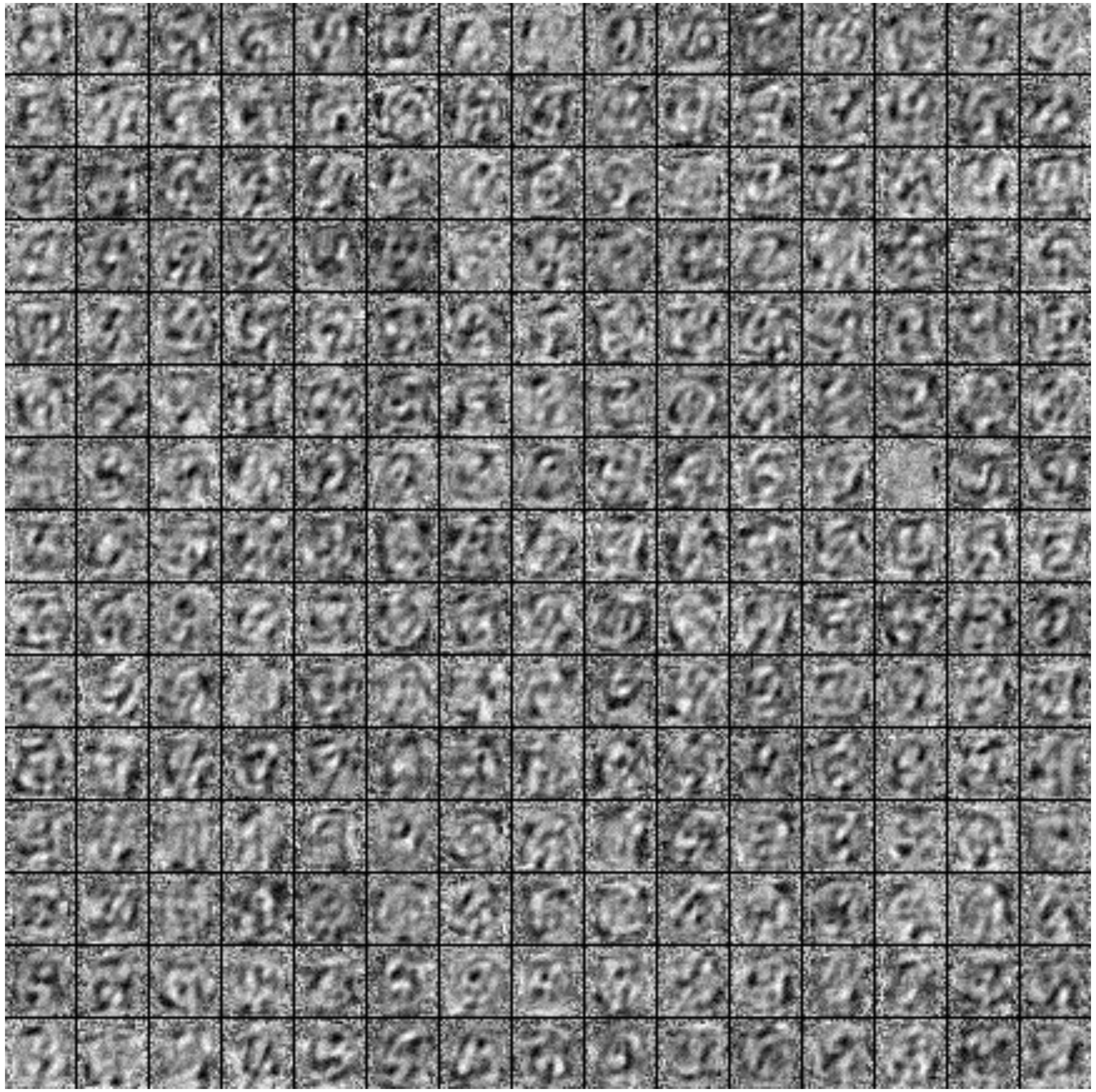


Features are learned

ML: Deep learning demo

A simple graphical example

Visualise the features



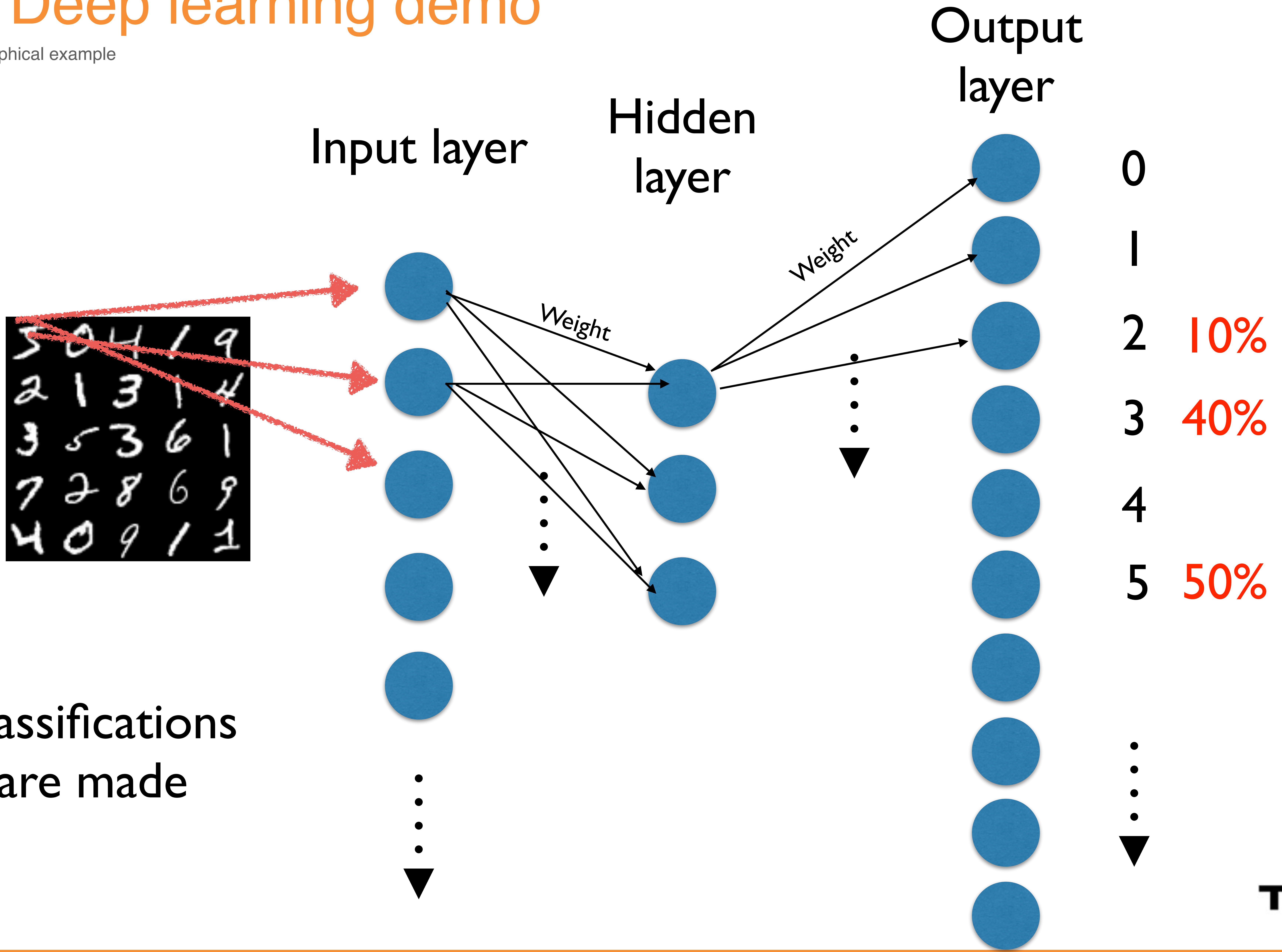
@DrPhilWinder

TRIFORK.

...think software

ML: Deep learning demo

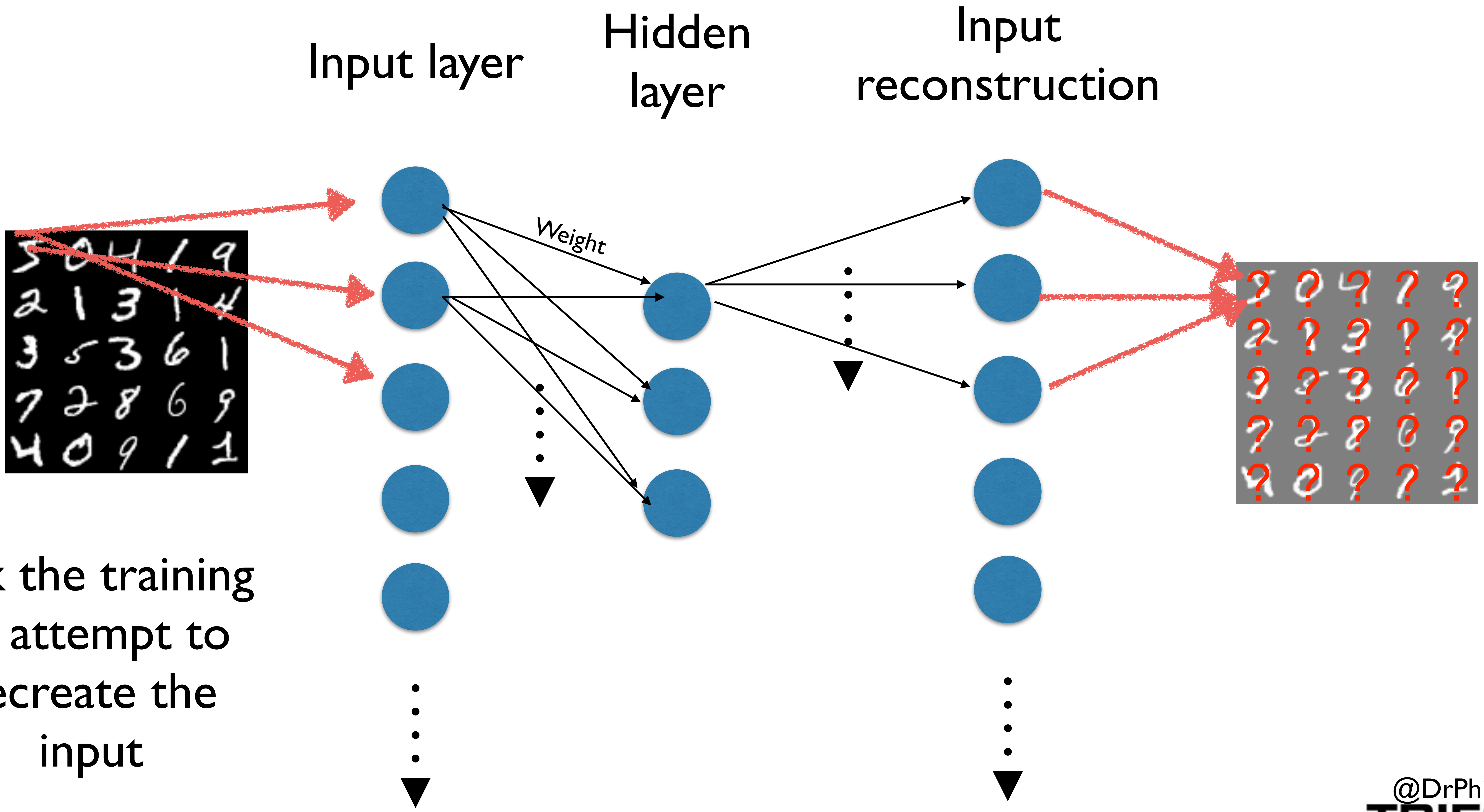
A simple graphical example



Classifications are made

ML: Deep learning demo

A simple graphical example



Ask the training to attempt to recreate the input

ML: Deep learning demo

A simple graphical example

ML: Deep learning demo

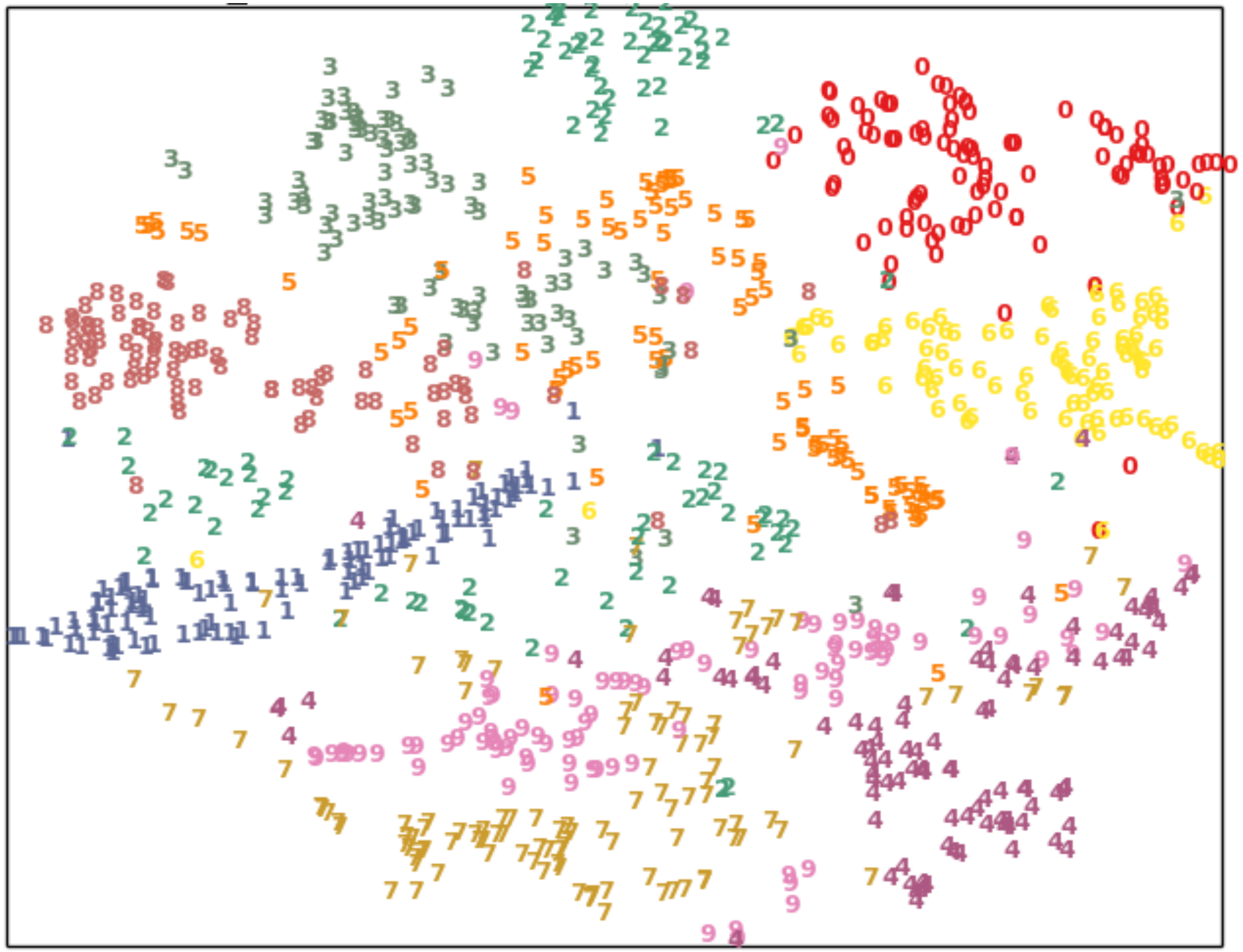
A simple graphical example

ML: Deep learning demo

A simple graphical example

Flatten the output into 2D, for plotting

(Imagine flattening a 3D cube to a 2D square)



Precision

0.84

0.98-0.99 is possible on this dataset

Financial Crime Demos

Background	Machine learning	Demos	Architectures
1	2	3	4

Rules based:
Graph databases

What is a graph database?

1

It's a database

NoSQL

2

It's a graph

Terminology:

- Node

  An object, a thing, a noun

- Relationship

  A link, a relationship, a verb

3

A natural representation of your data

A graph structure may be a more natural fit of your data. Use the right tool for the job.

What is a graph?

Terminology and examples

Node	Relationship	Node
Bob	Is friends with	Jane
A chair	Is contained within	The meeting room
Jane	Bought	Catch 22
Jane	Placed a transaction of £20	At WH Smiths

The power of graphs

The motivation

Better represents problem domain

Performance

Agility

Flexibility

Neo4j

A (very) quick look

Cypher makes queries intuitive: (nodes), [relationships], -[]-> direction


```
MERGE (:PhoneNumber {number:"01234524312"})<-[:HAS_PHONENUMBER]
  -(:AccountHolder {first:"John",last:"Smith",id:"JohnSmithID"})-[:HAS_NI]->(:NI {id:" JW123294D"})
```

`MATCH (n)-[r]-() RETURN n,r;`
- Match all nodes with a relationship.

`MATCH (ni:NI) RETURN ni;`
- Match any node of type NI

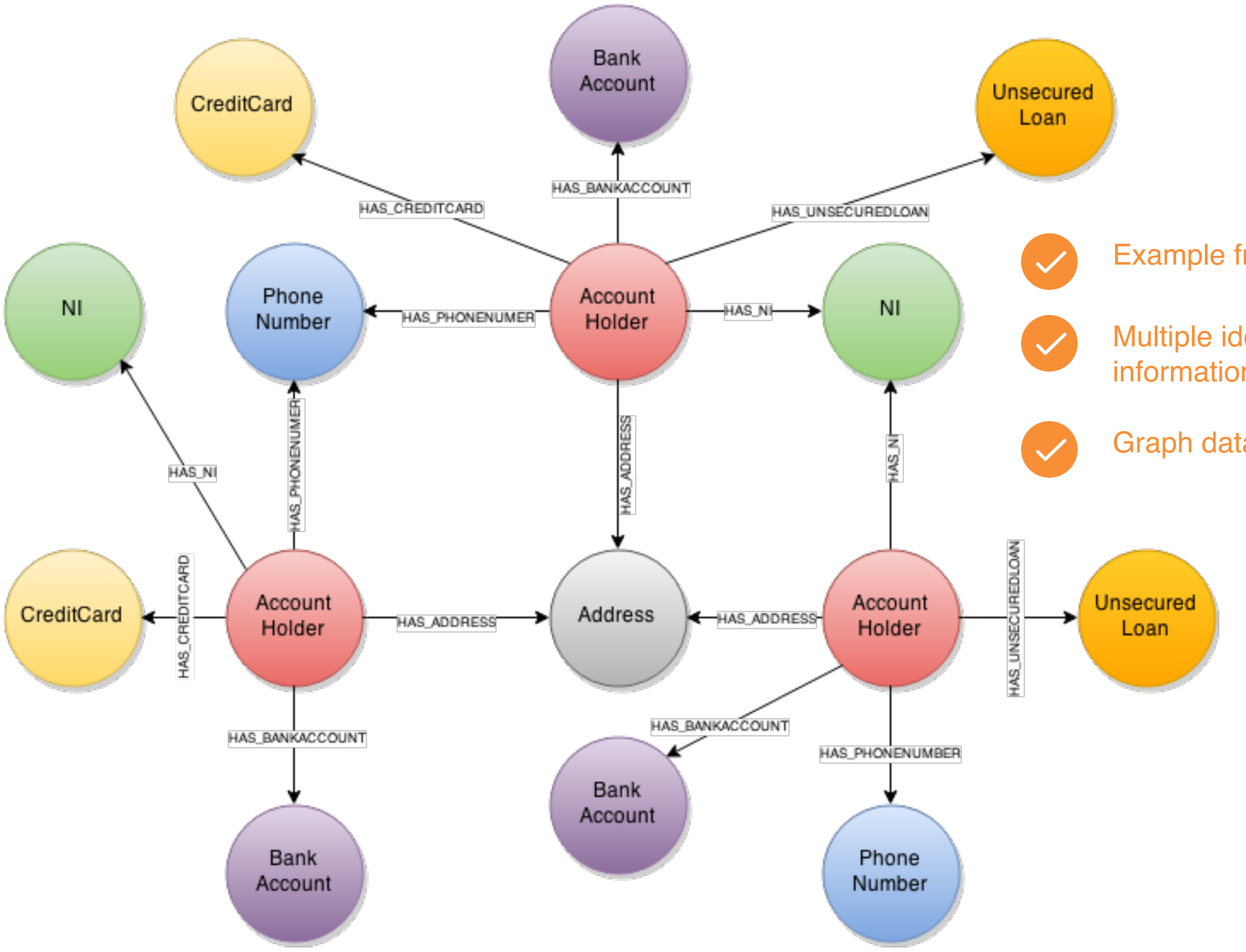
`MATCH (n)-[:HAS_NI]-() return n;`
- Match any node that has a HAS_NI relationship

@DrPhilWinder

TRIFORK. ...think software

Neo4j

A (very) quick look



- Example fraud ring
- Multiple identities sharing legitimate information
- Graph databases can help

@DrPhilWinder

TRIFORK. ...think software

Deep Learning: Voice “fingerprinting” for origination

Goal

Origination
Prove the identity of the customer

Offline

Record customer's voice
Record

Pre-process data to generate features.
Process

Train deep learning model
Training

Store "fingerprint" for verification
Save

Online

Record customer's voice
Record

Pre-process data to generate features.
Process

Compare result to "fingerprint"
Test

Verified

Overview

Three people, eight phrases

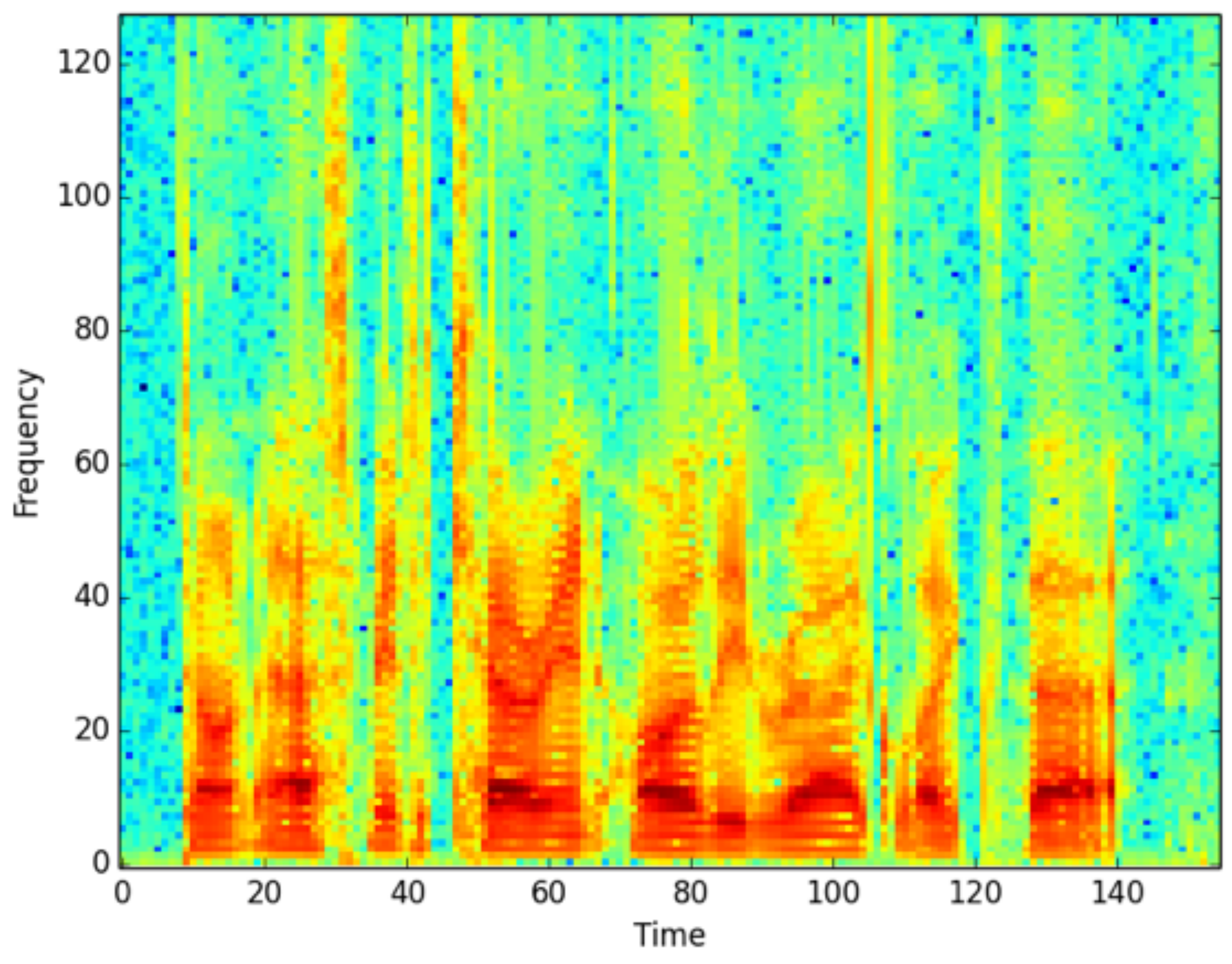
Record

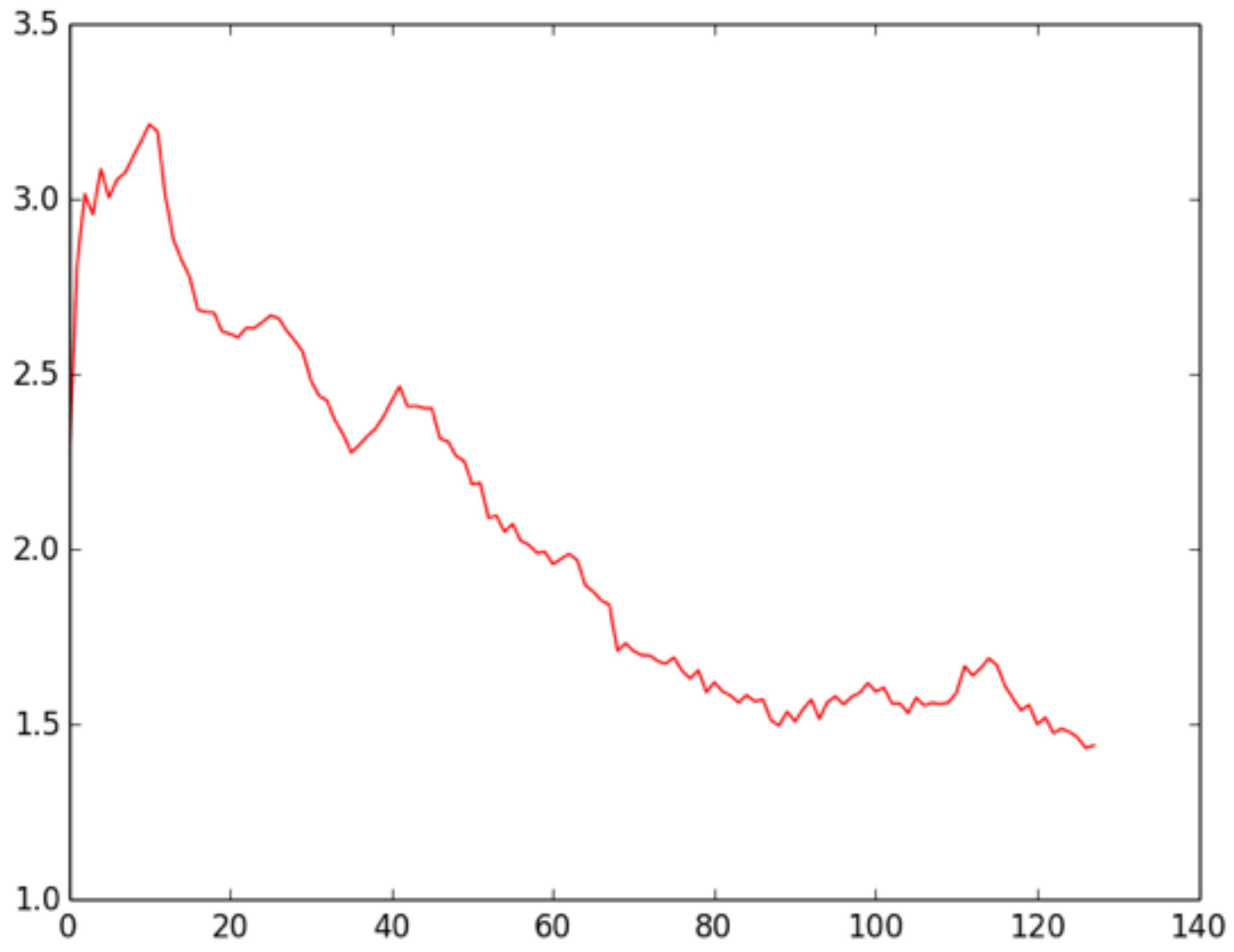
FFT and average

Process

Deep learning based Classification

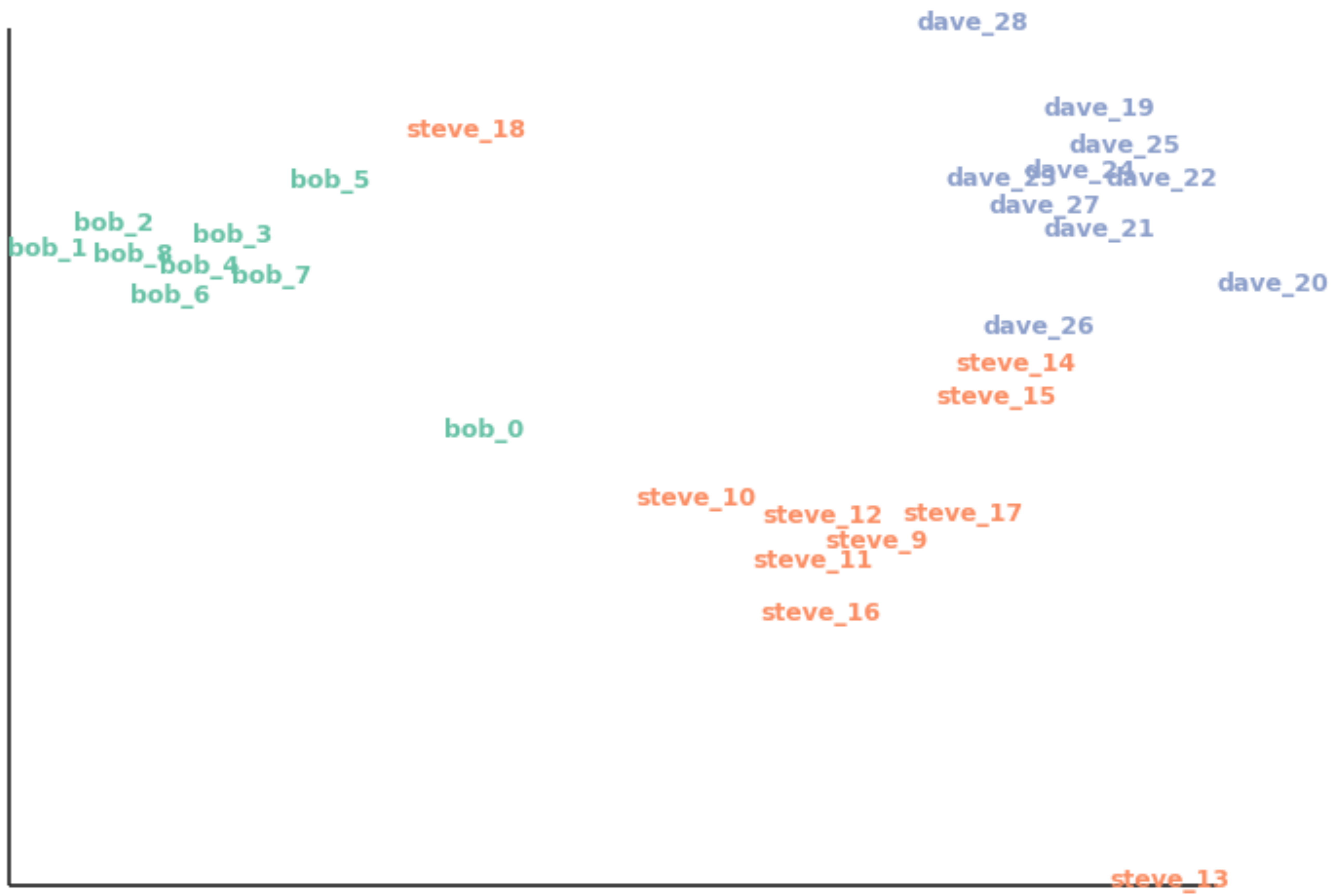
Training





Deep learning

Each colour/name represents a person.
Each example is a phrase.



Classification

Probability
Bob Steve Dave

[0.98 0.01 0.01]

[0.01 0.97 0.01]

[0.02 0.03 0.96]

Voice data: http://web.mit.edu/6.863/share/nltk_lite/timit/

Python + Keras + SkLearn

Decision trees:
Predicting Mortgage Defaults

Demo: Mortgage default prediction

Can we predict defaults?

- Given labelled mortgage applications, is it possible to predict defaults?
- What data have we got access to?
- Is it enough?

Freddie Mac / Fannie Mae

Huge datasets released by publicly owned US lenders.

Provides default label

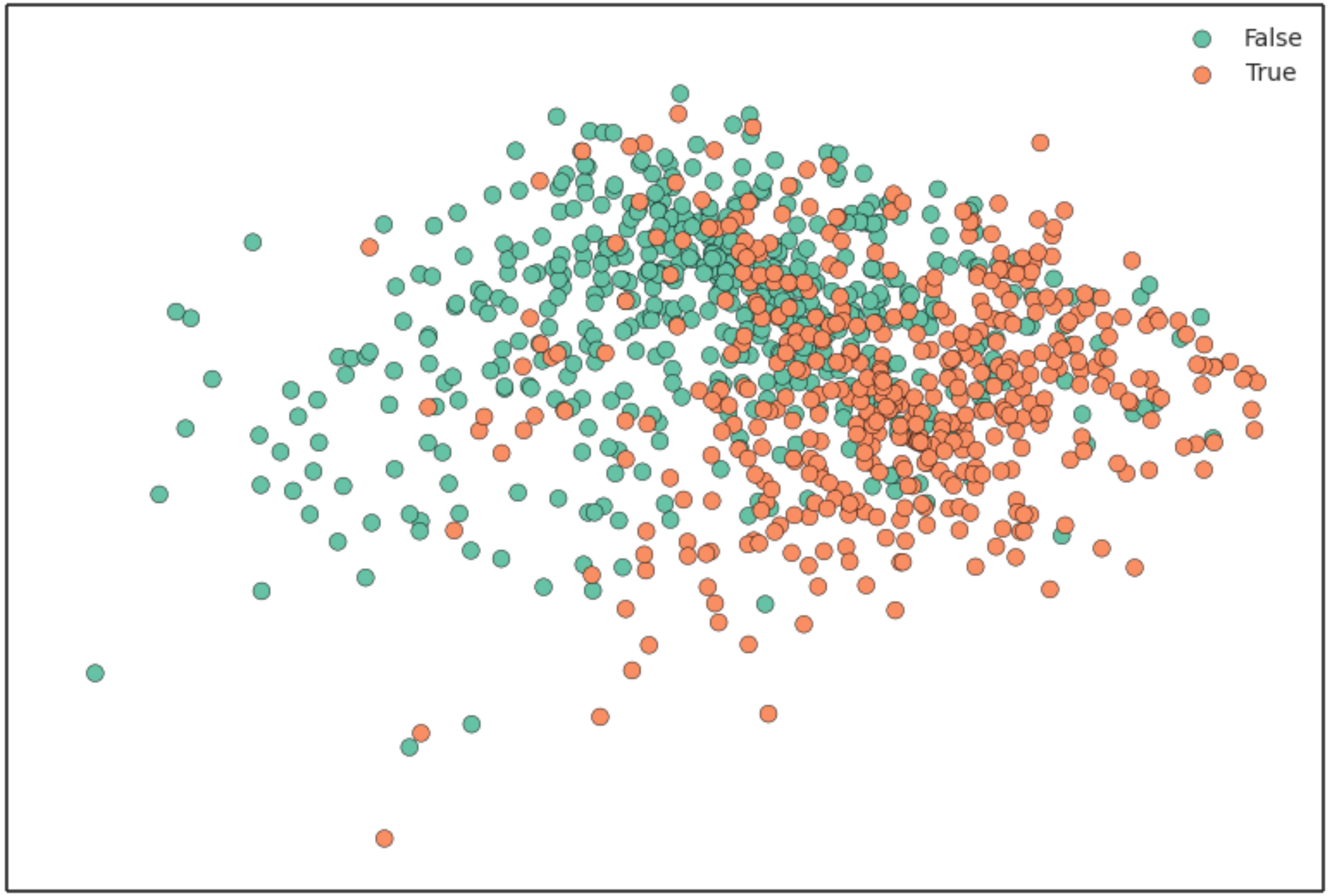
Let’s take a look at the data

Big cleaning effort

Remove as much as feasible

```
CREATE TABLE loans_learning (
  id integer NOT NULL,
  first_payment_date date,
  credit_score integer NOT NULL,
  first_time_homebuyer_flag integer NOT NULL,
  mip integer,
  number_of_units integer,
  occupancy_status integer NOT NULL,
  ocltv numeric,
  dti integer NOT NULL,
  original_upb numeric,
  oltv numeric,
  original_interest_rate numeric,
  channel integer NOT NULL,
  prepayment_penalty_flag integer NOT NULL,
  property_type integer NOT NULL,
  loan_sequence_number char(12),
  loan_purpose integer NOT NULL,
  original_loan_term integer,
  number_of_borrowers integer NOT NULL,
  hpi_at_origination numeric,
  default_flag boolean
);
```

Let’s take a look at the data

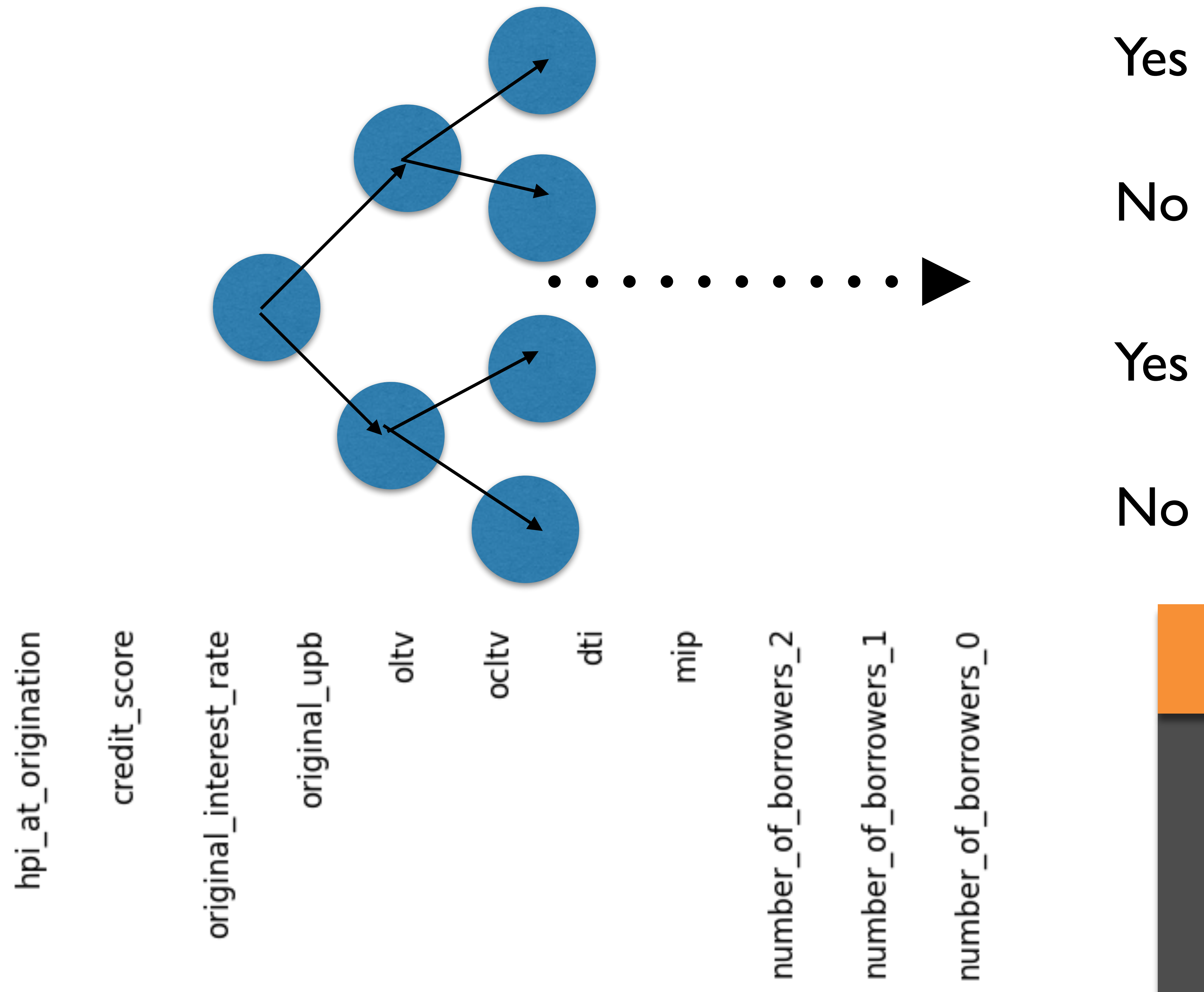


Flatten the output into 2D, for plotting

Method

Decision tree

Classification

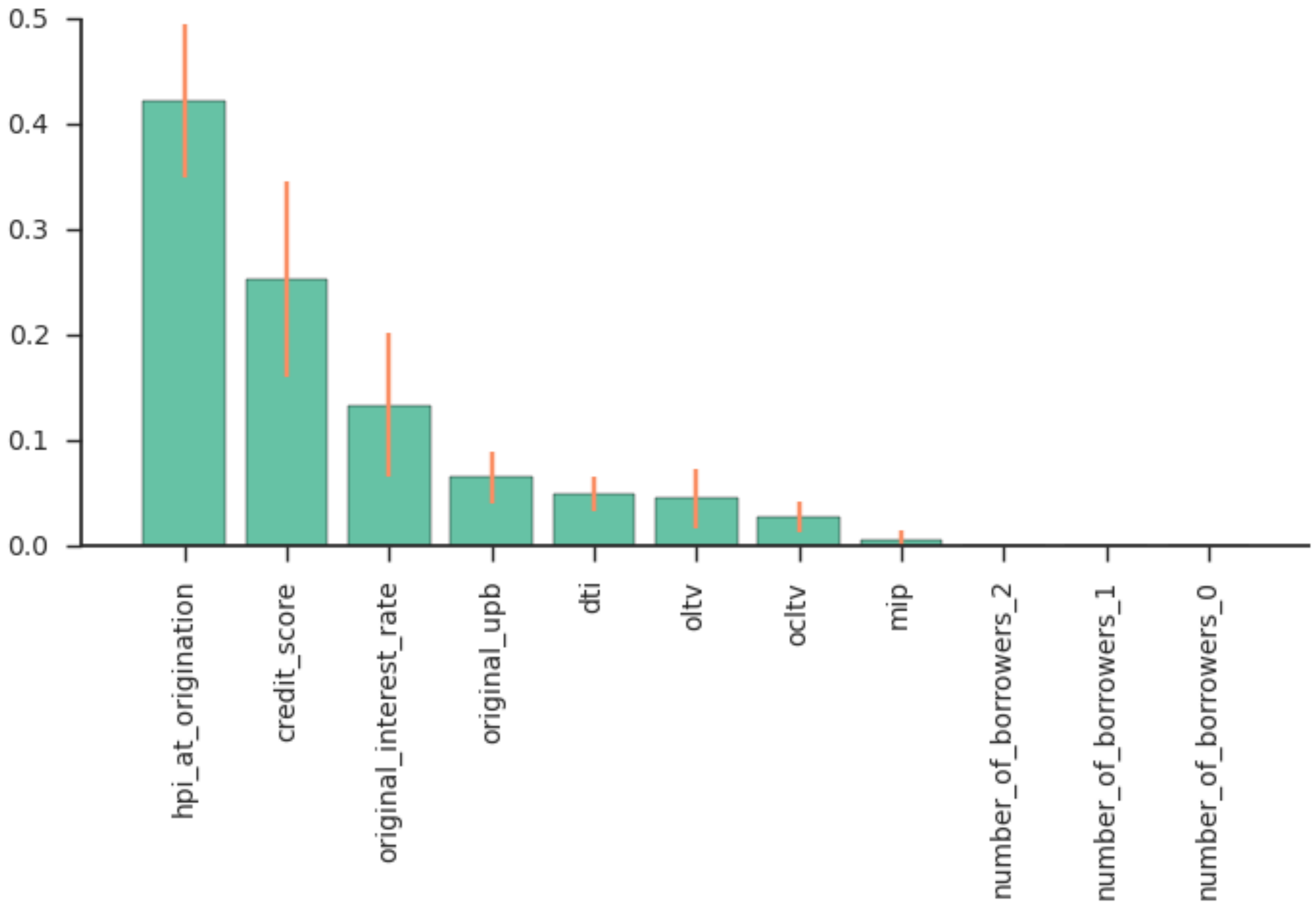


Data

- Approx. 10,000 default examples (20,000 total)
- Random Forest classifier
- 11 input features (very small)

Results 1: Feature importance

Mortgage



Results 2: Classification

	Precision	Recall	F1-score	Support
FALSE	0.84	0.83	0.84	995
TRUE	0.84	0.84	0.84	1005

Deep learning: Detecting unknown crime

Demo: Detecting unknown fraud

You're always one step behind

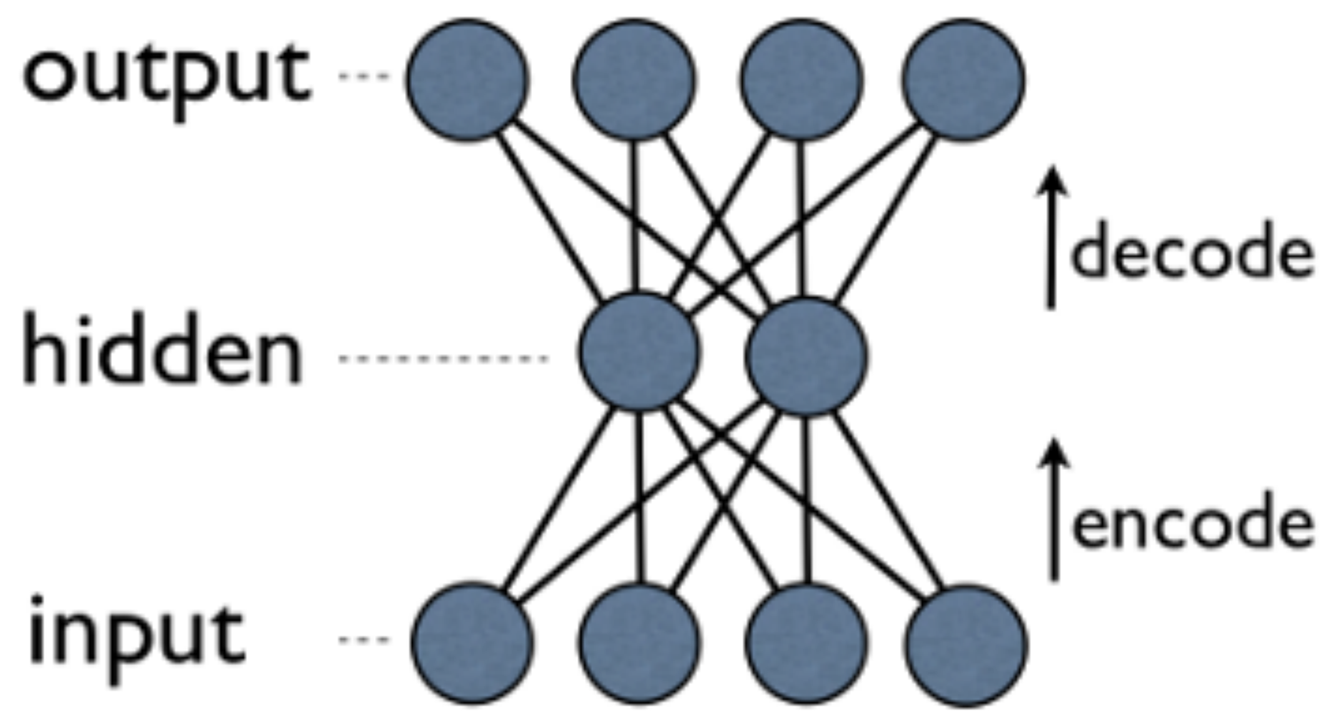
- What about when the rules don't catch the fraudster?
- What should we look for?

Deep learning

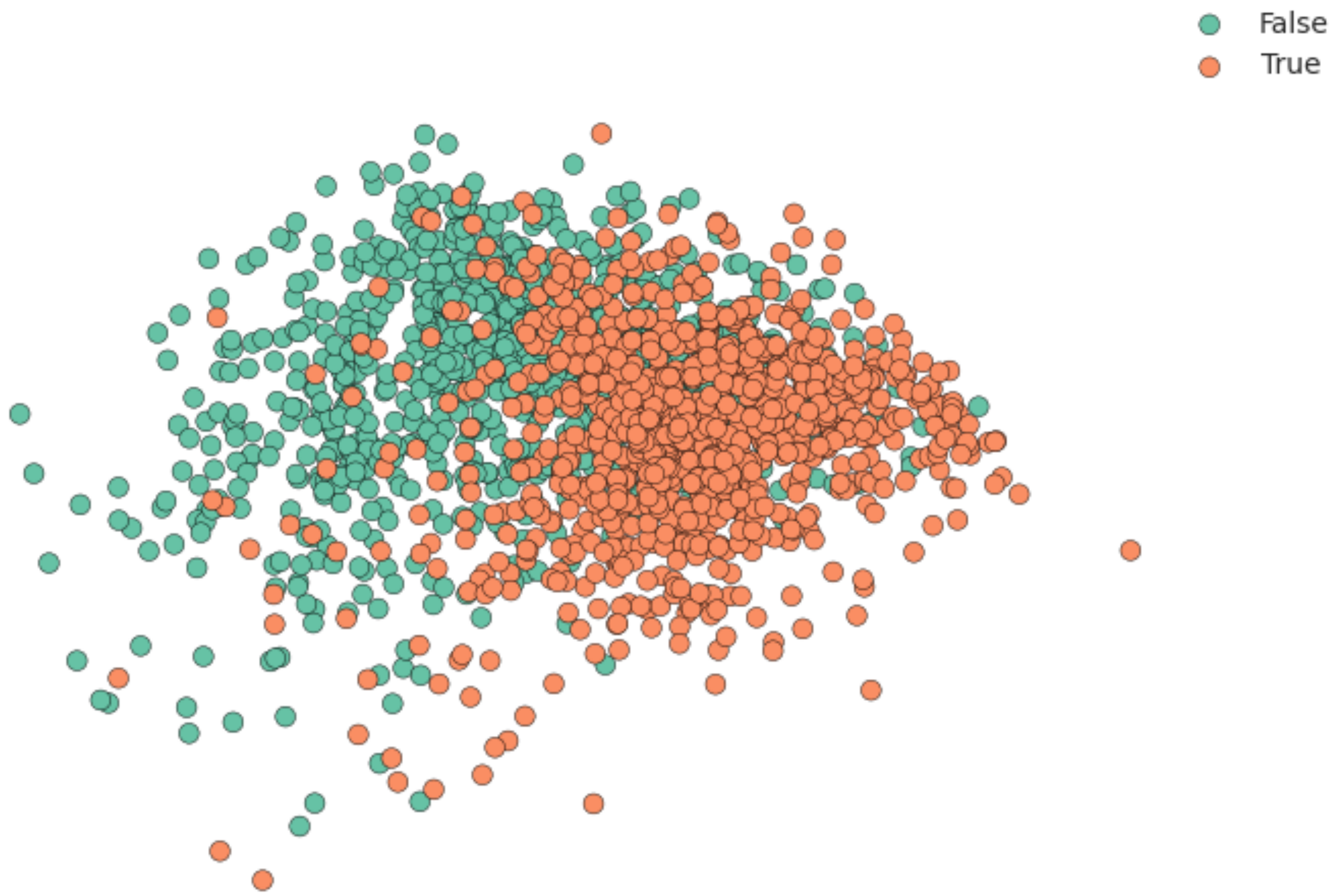
Lets ask deep learning to investigate the data.

Completely unsupervised, I have no data on fraudulent mortgages.

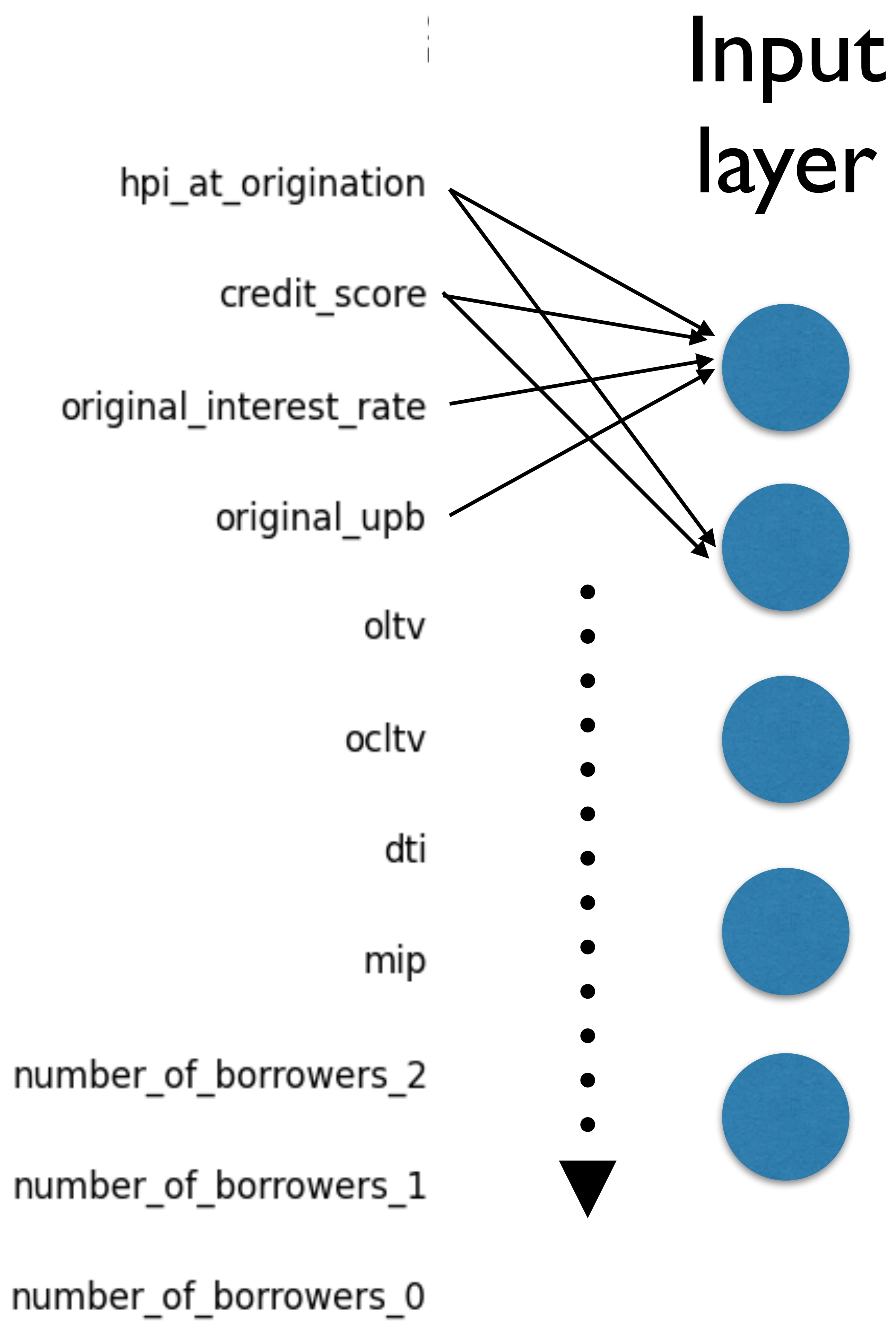
How? An Auto-Encoder



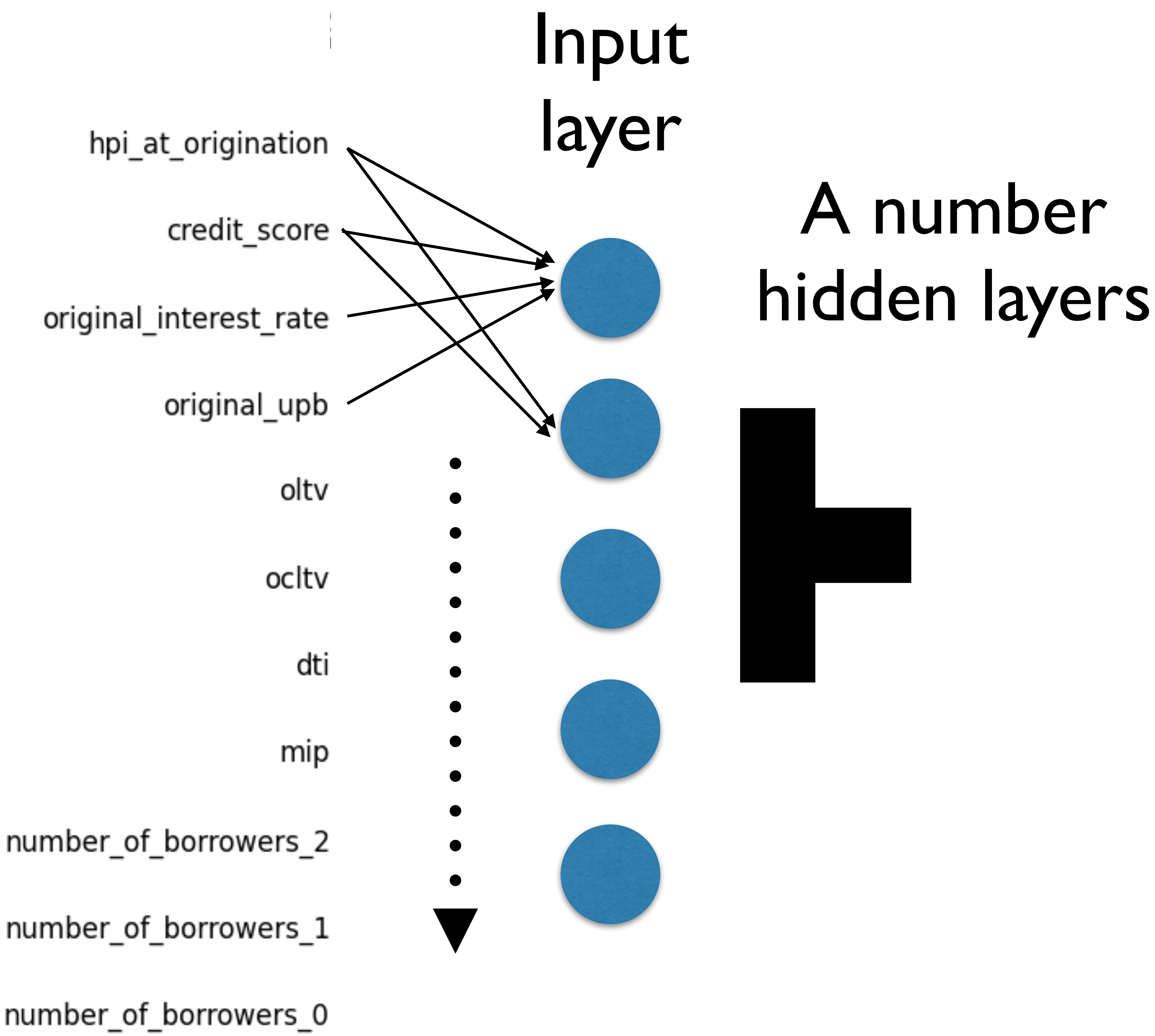
Before



Method

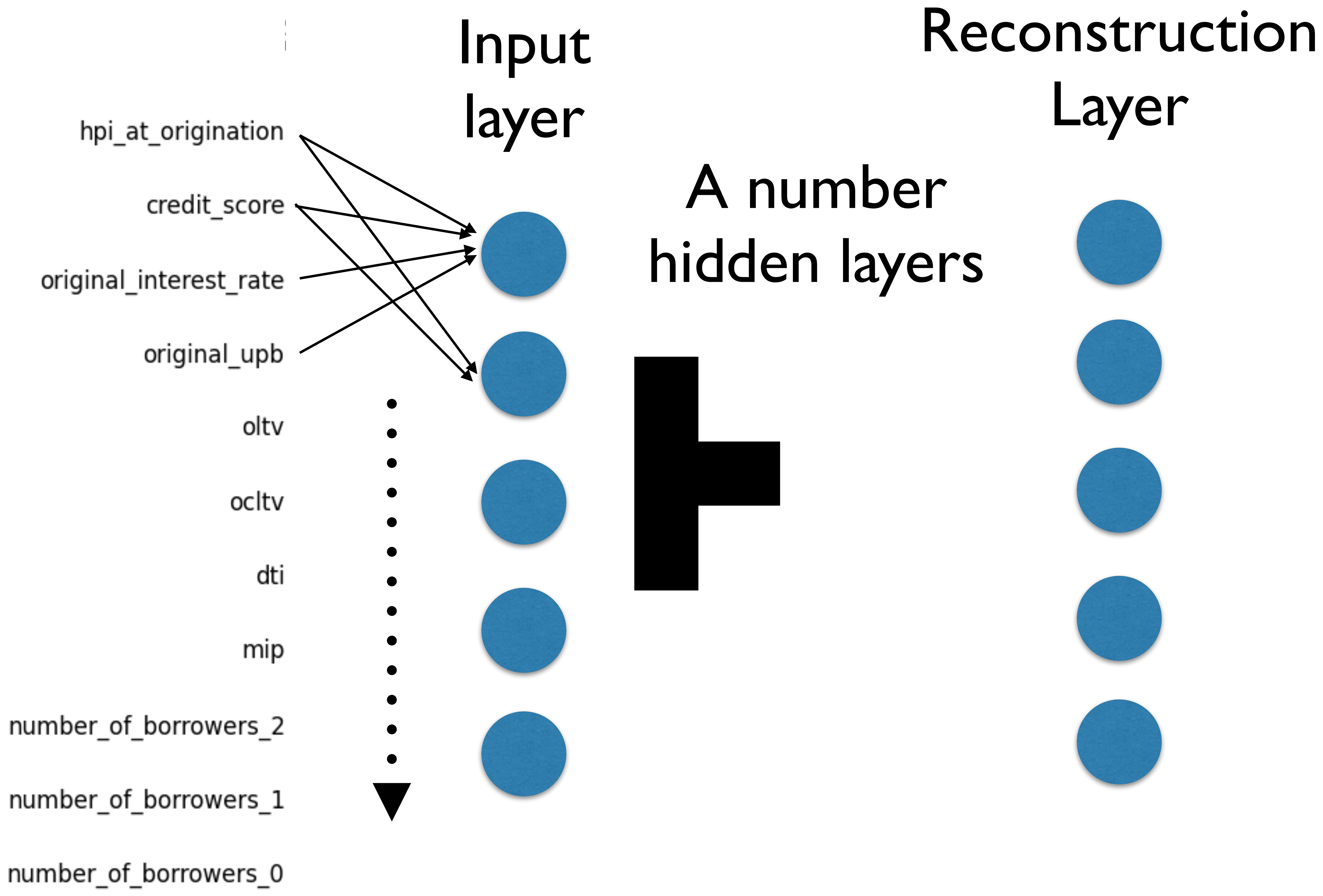

Input layer
hpi_at_origination
credit_score
original_interest_rate
original_upb
oltv
ocltv
dti
mip
number_of_borrowers_2
number_of_borrowers_1
number_of_borrowers_0

Method

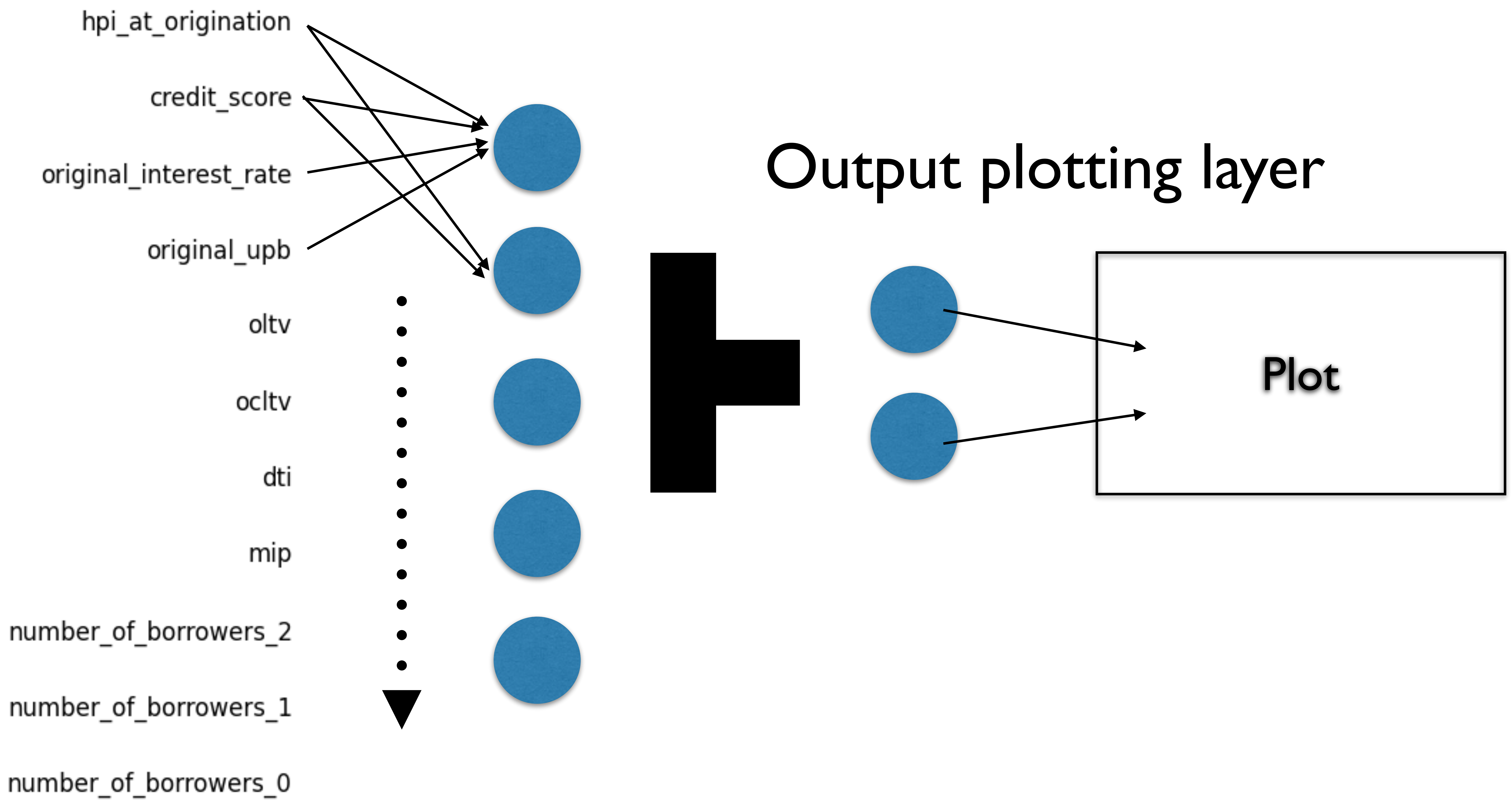

Input layer
A number hidden layers
hpi_at_origination
credit_score
original_interest_rate
original_upb
oltv
ocltv
dti
mip
number_of_borrowers_2
number_of_borrowers_1
number_of_borrowers_0

Method

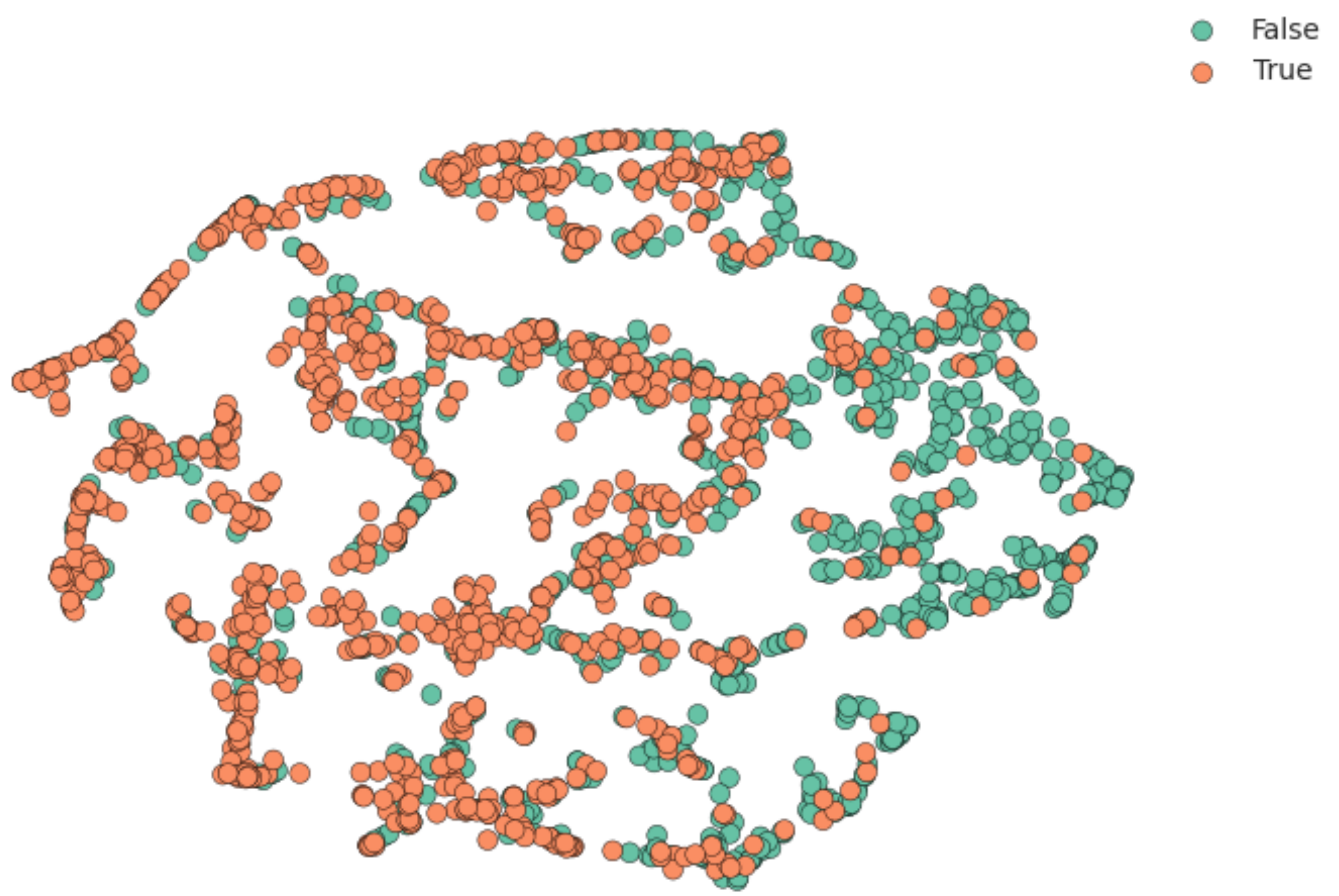
During training…

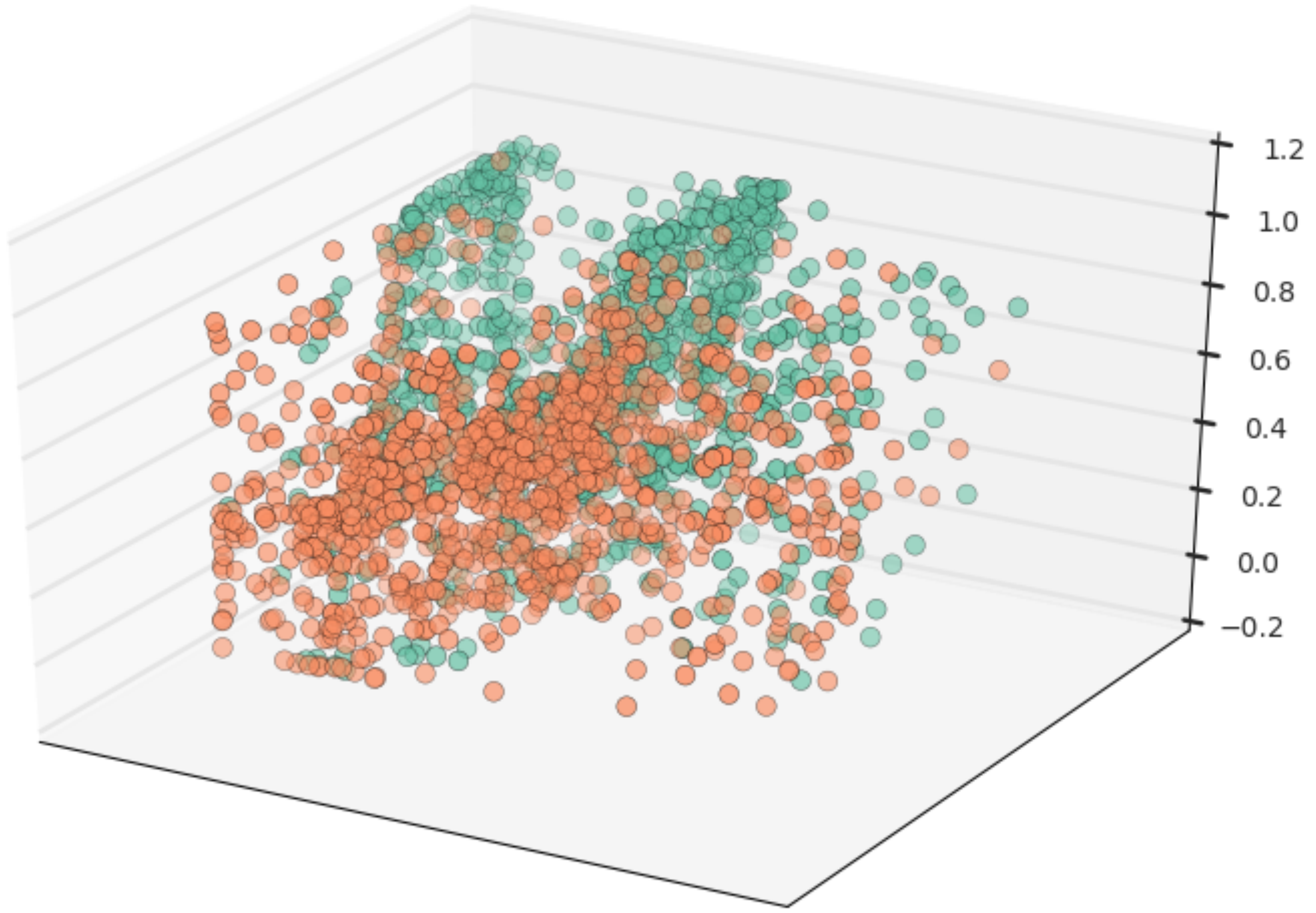


Method


hpi_at_origination
credit_score
original_interest_rate
original_upb
oltv
ocltv
dti
mip
number_of_borrowers_2
number_of_borrowers_1
number_of_borrowers_0
Output plotting layer
Plot

After





One of many possible visualisations

Tools and techniques

Background	Machine learning	Demos	Architectures
1	2	3	4

Tech: Proof of concepts (R&D)

Python (R/Matlab)

- sklearn
- Keras, Theano
- A database of some kind (Elasticsearch + elasticsearch-py)
- Laptop

Tech: Production

Computing

- Apache spark

Databases

- Riak
- Elasticsearch
- Neo4j

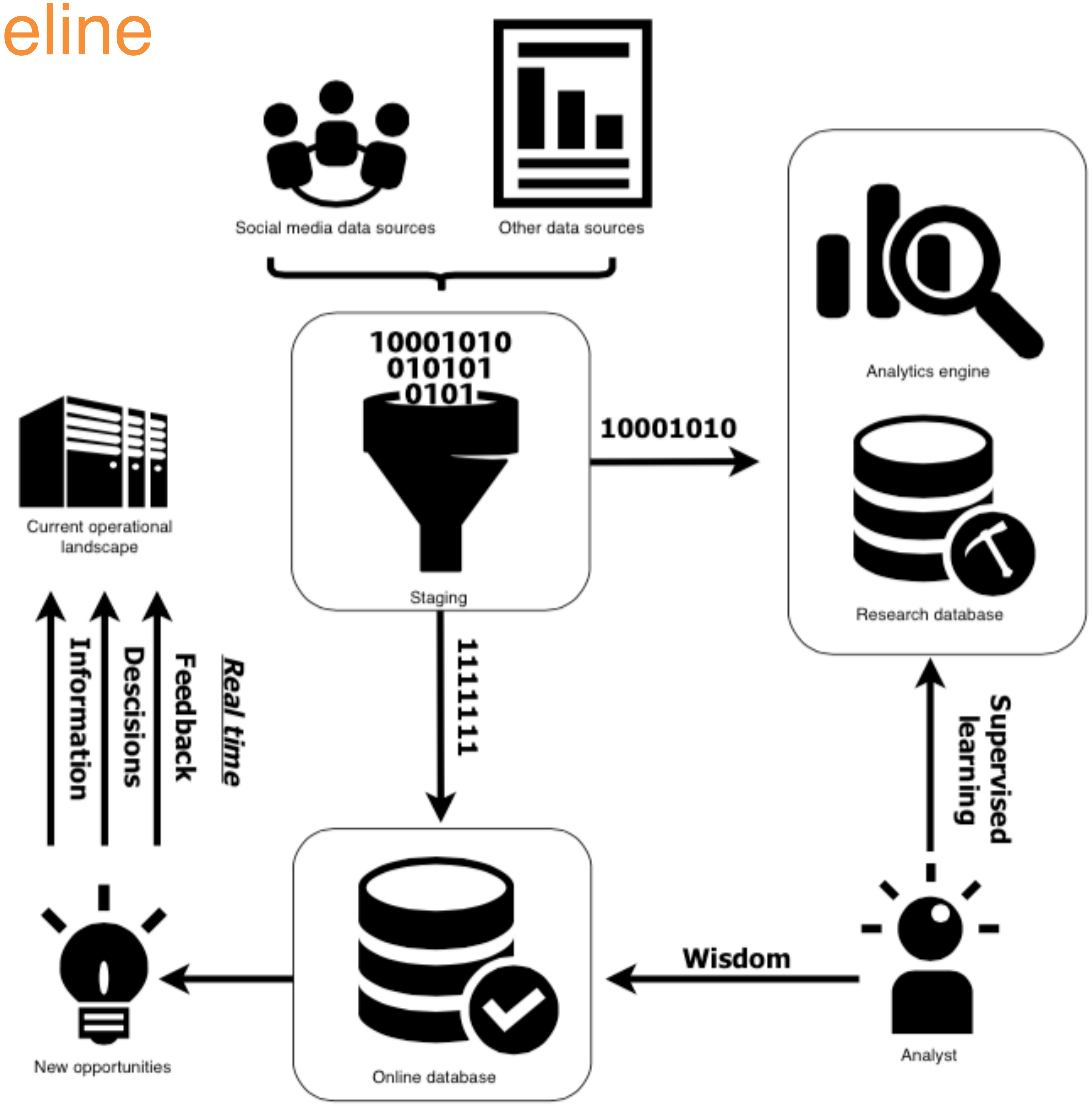
Infrastructure/Comms

- Apache Mesos
- Docker
- Akka
- Consul/Terraform
- etc.

And many more…

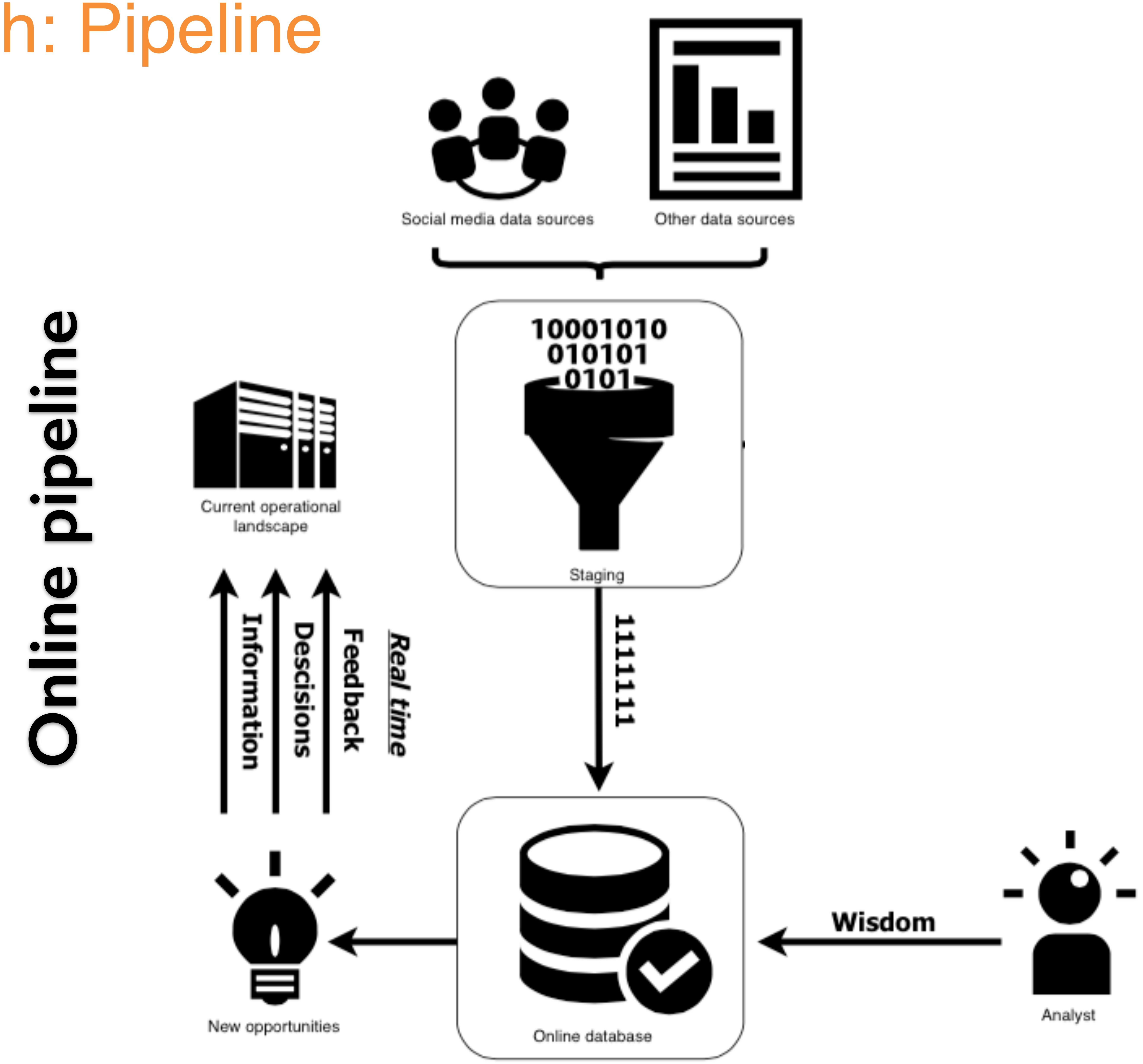
- Legacy integration
- APIs
- Data management
- User management
- Reporting
- Front end
- Etc. etc.

Tech: Pipeline

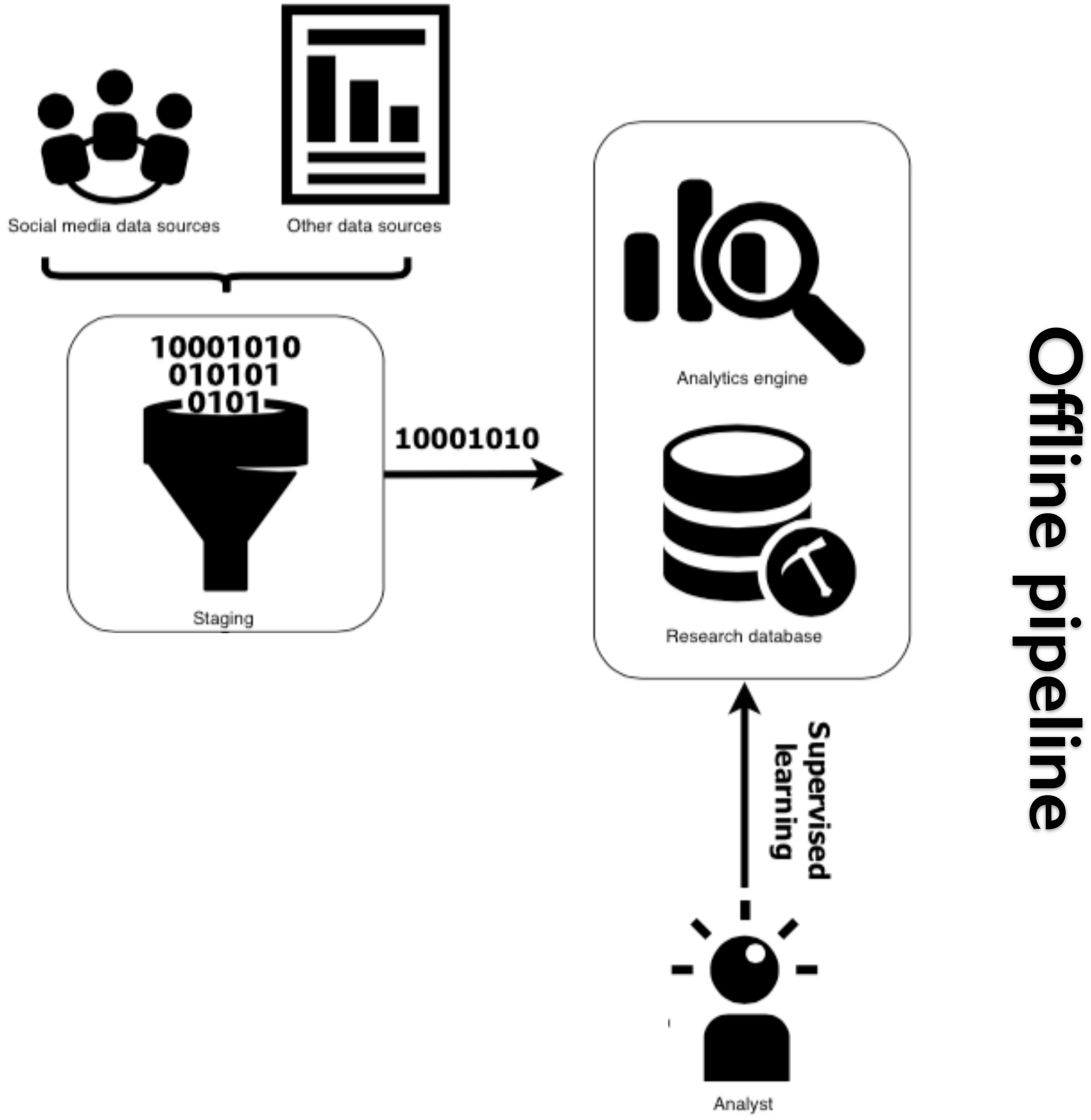


@DrPhilWinder
TRIFORK.
...think software

Tech: Pipeline



Tech: Pipeline



Summary

- Fraud evolves rapidly, legislation evolves even faster!
- Need for a disruptive approach
- Deep learning reveals new methods of analysis and sophisticated automation
- Profit drivers:
  - Being able to trust valid applications through analysis and verification
  - Automation improves efficiency

goto; conference

Please

Remember to rate this session

Thank you!

github.com/philwinder

@DrPhilWinder

pnw@trifork.com