Scikit-learn:

Machine learning in Python land

Gaël Varoquaux

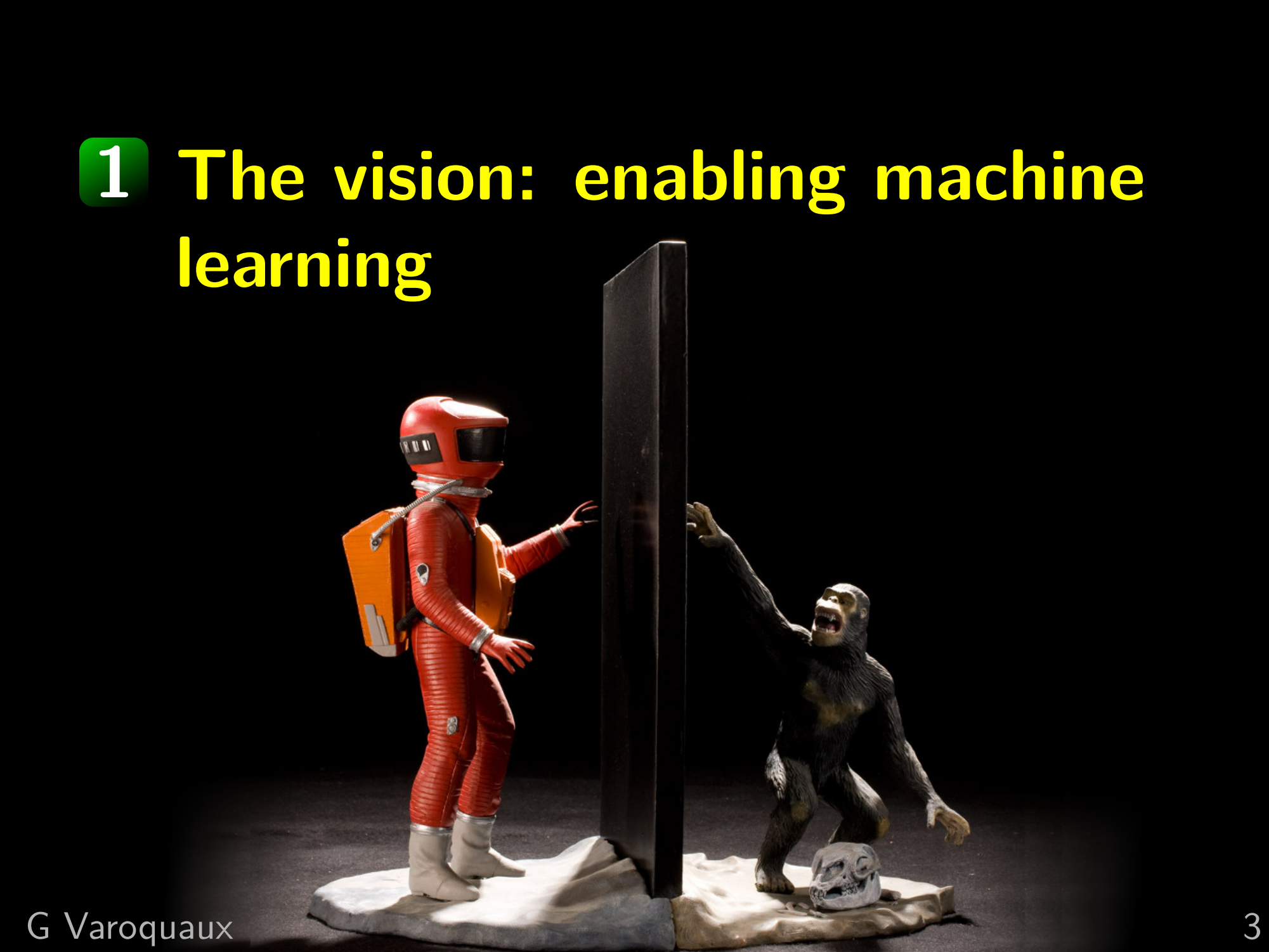
Inria

scikit learn

machine learning in Python

1. **The vision: enabling machine learning**
2. **The tool: a Python library**
3. **The project: a community**

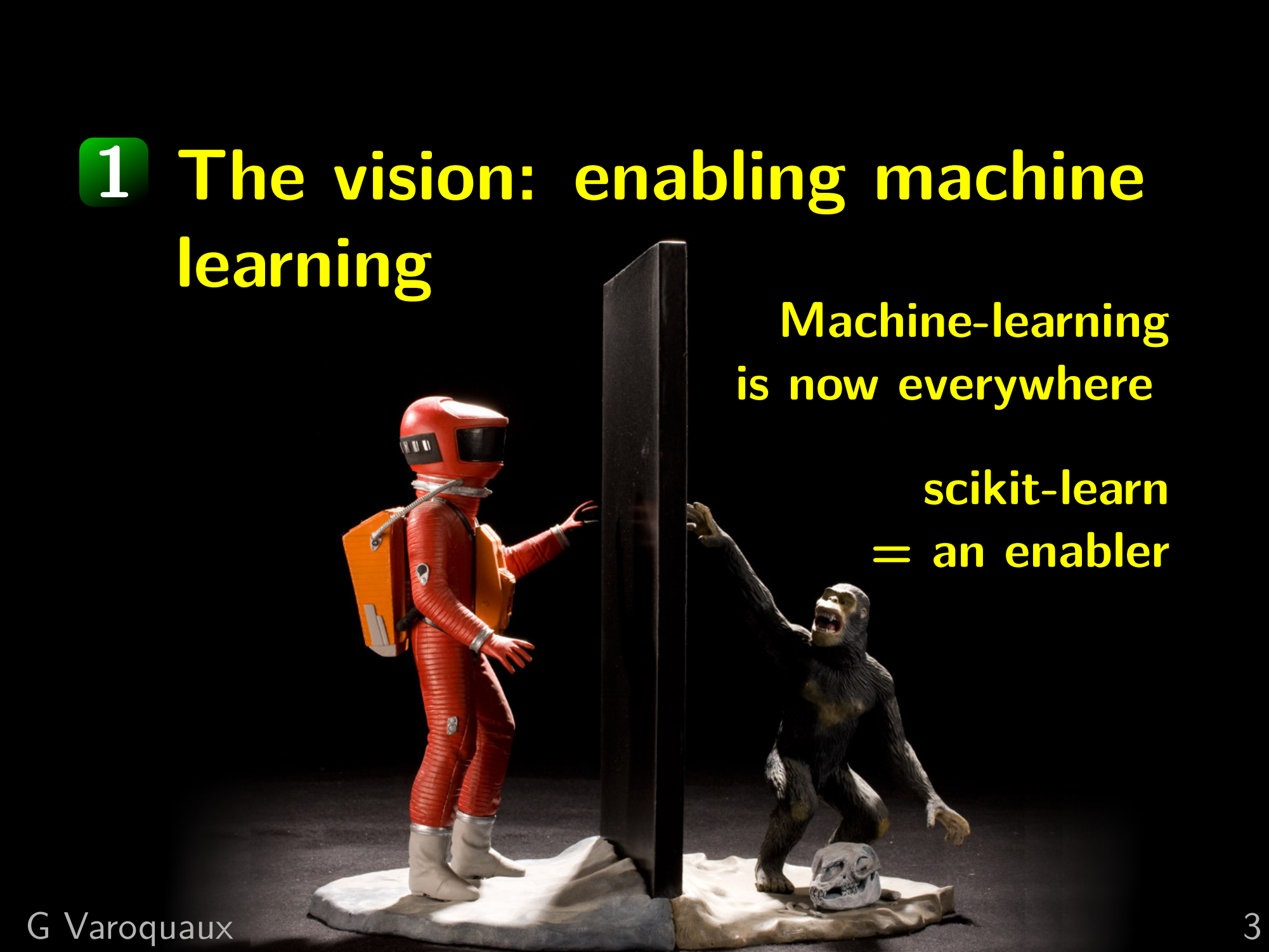
1 The vision: enabling machine learning



1 The vision: enabling machine learning

Machine-learning is now everywhere

1 The vision: enabling machine learning



Machine-learning is now everywhere

scikit-learn = an enabler

1 The vision: enabling machine learning

Machine-learning is now everywhere

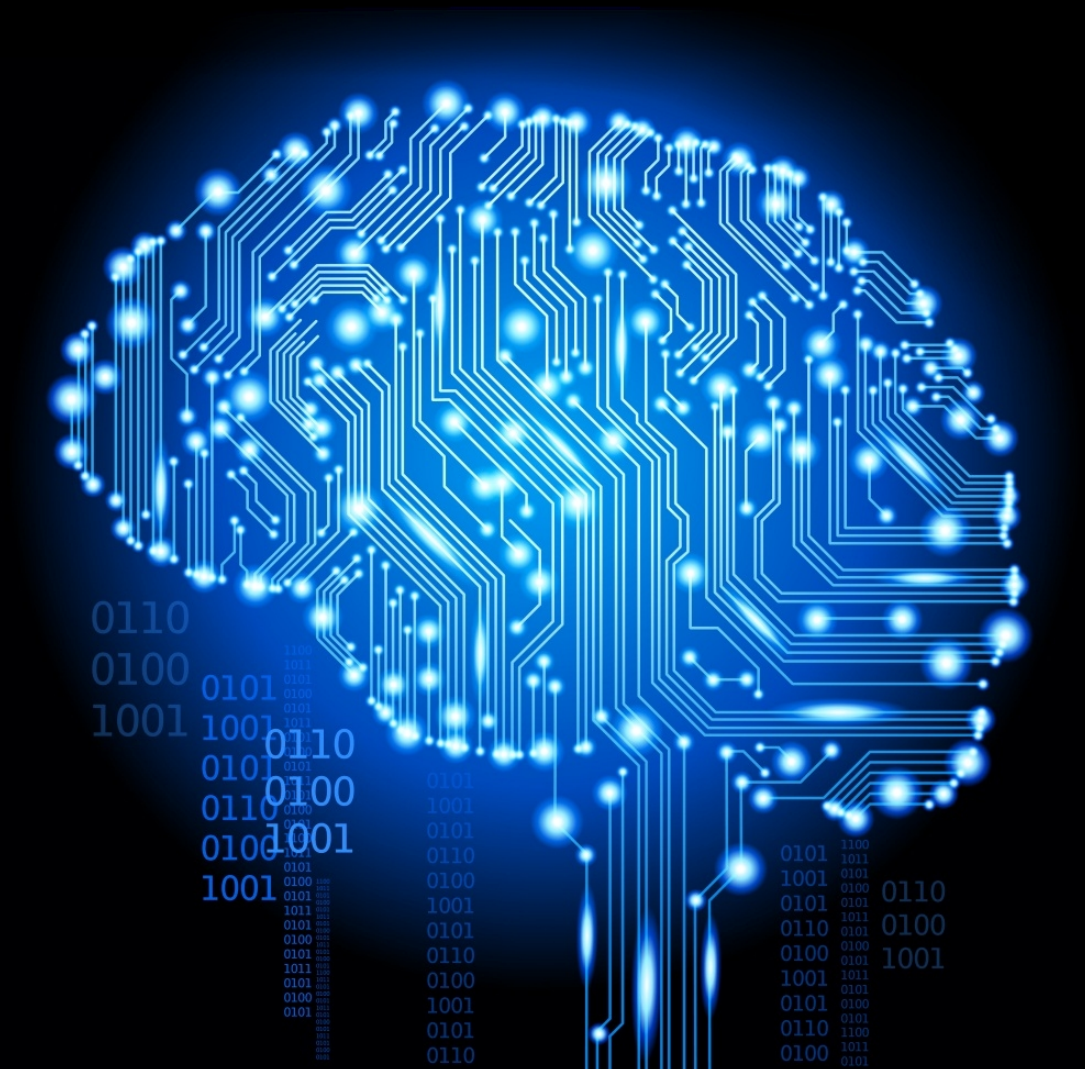
scikit-learn = an enabler

Machine learning for everybody and for everything

Machine learning without learning the machinery

Machine learning in a nutshell

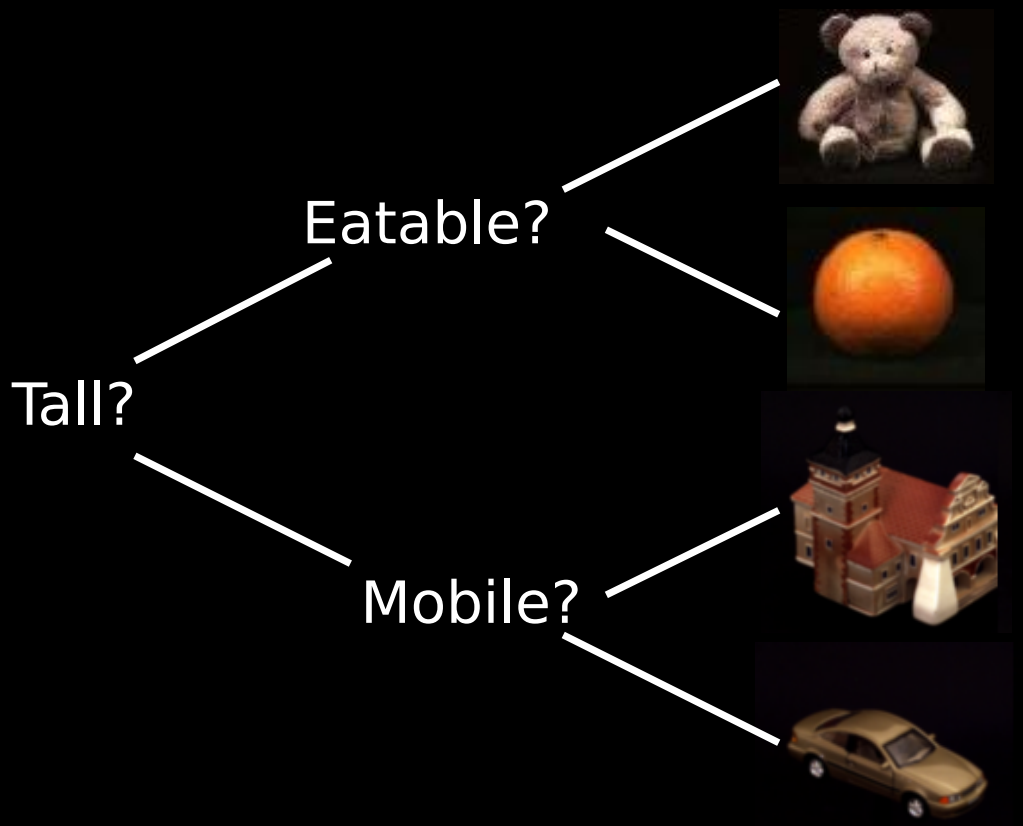
Machine learning is about making prediction from data



Expert systems

The 80s

- Building decision rules

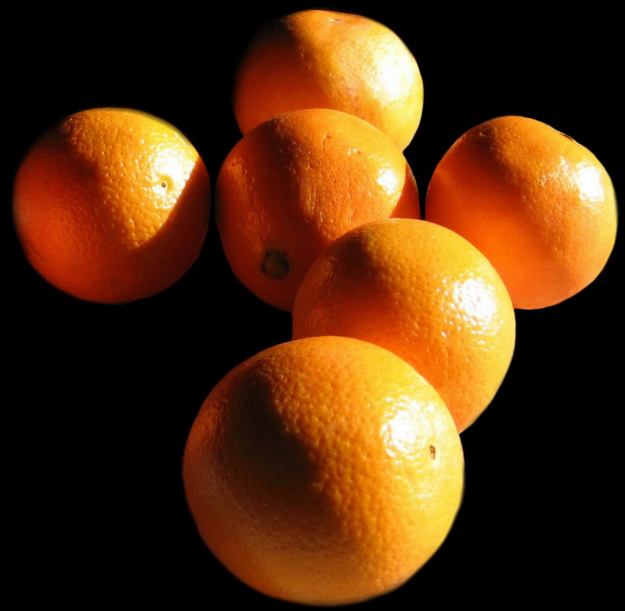


Expert systems The 80s

- Building decision rules

Machine learning The 90s

- Learn these from observations



1 Machine learning: a historical perspective

Expert systems The 80s

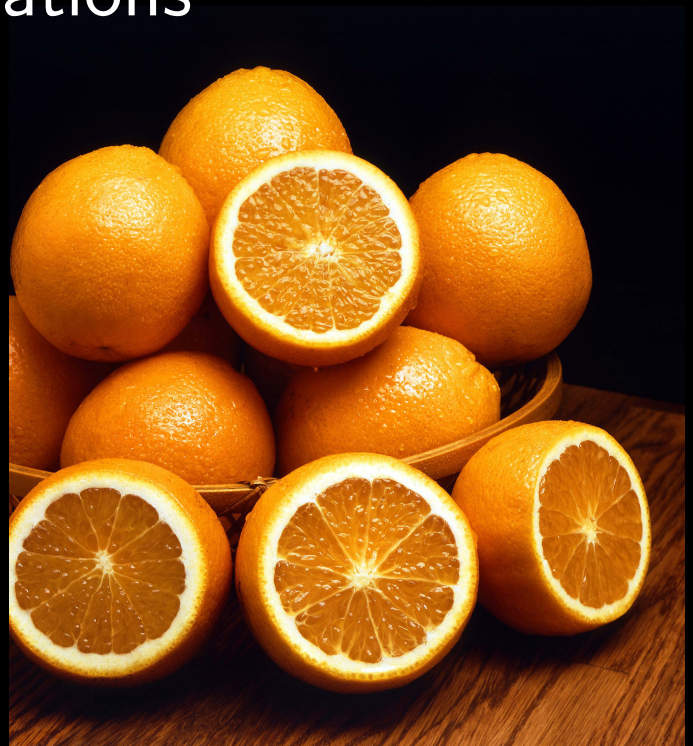
- Building decision rules

Machine learning The 90s

- Learn these from observations

Statistical learning 2000s

- Model the noise in the observations



1 Machine learning: a historical perspective

Expert systems The 80s

- Building decision rules

Machine learning The 90s

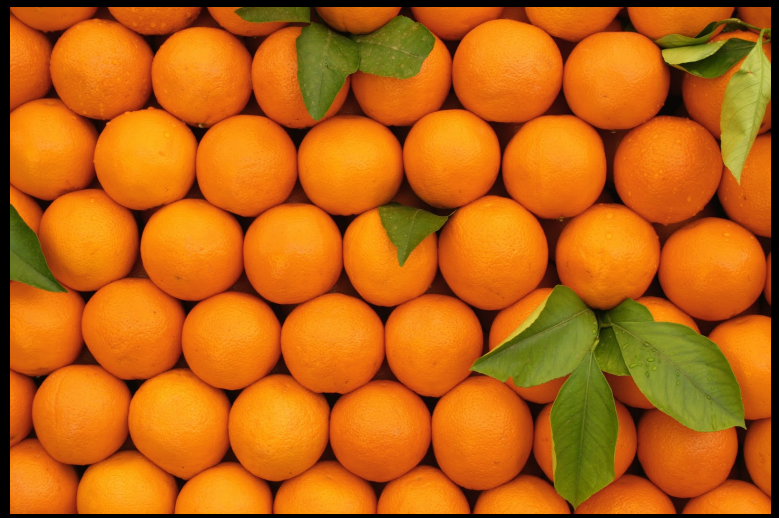
- Learn these from observations

Statistical learning 2000s

- Model the noise in the observations

Big data today

- Many observations



1 Machine learning: a historical perspective

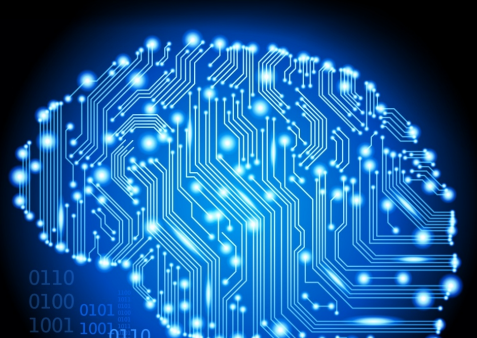
Expert systems — The 80s
- Building decision rules

Machine learning — The 90s
- Learn these from observations

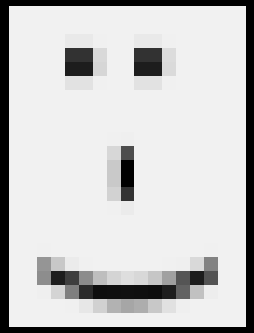
Statistical learning — 2000s
- Model the noise in the observations

Big data — today
- Many observations

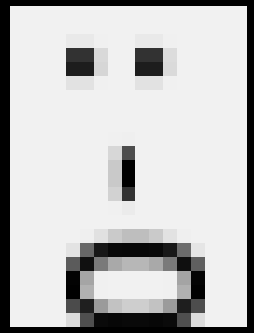
Artificial Intelligence — tomorrow
- Machine learning on hard problems



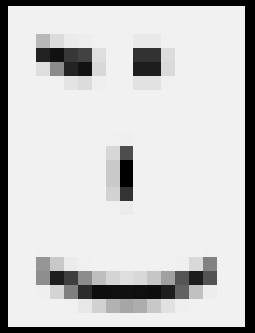
Face recognition



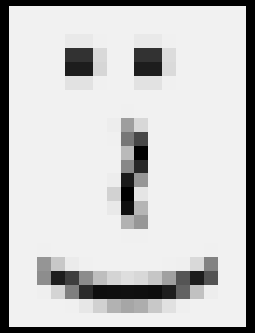
Andrew



Bill

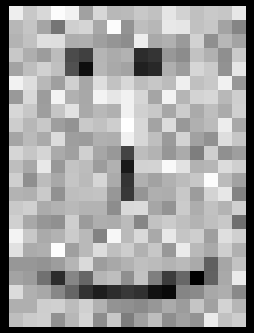


Charles

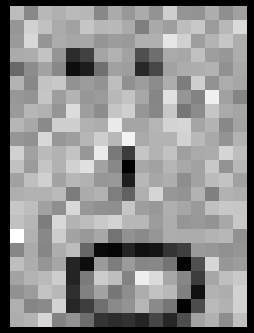


Dave

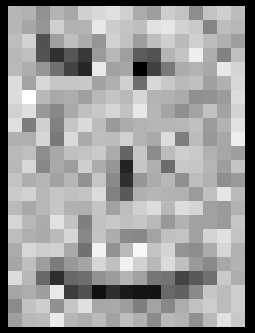
Face recognition



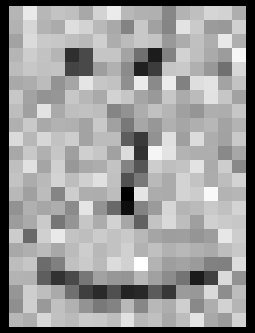
Andrew



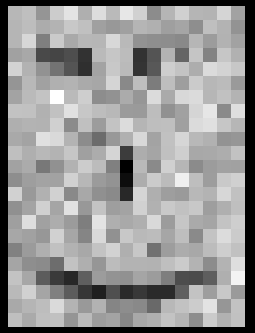
Bill



Charles



Dave



?

A simple method:

1. Store all the known (noisy) images and the names that go with them.
2. From a new (noisy) images, find the image that is most similar.

"Nearest neighbor" method

A simple method:

1. Store all the known (noisy) images and the names that go with them.
2. From a new (noisy) images, find the image that is most similar.

"Nearest neighbor" method

How many errors on already-known images?

... 0: no errors

A simple method:

1. Store all the known (noisy) images and the names that go with them.
2. From a new (noisy) images, find the image that is most similar.

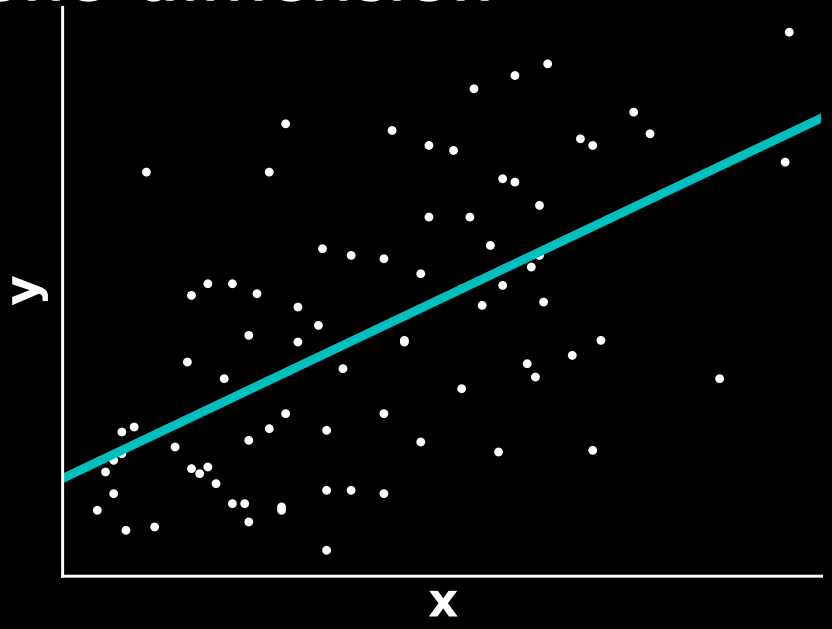
"Nearest neighbor" method

How many errors on already-known images?

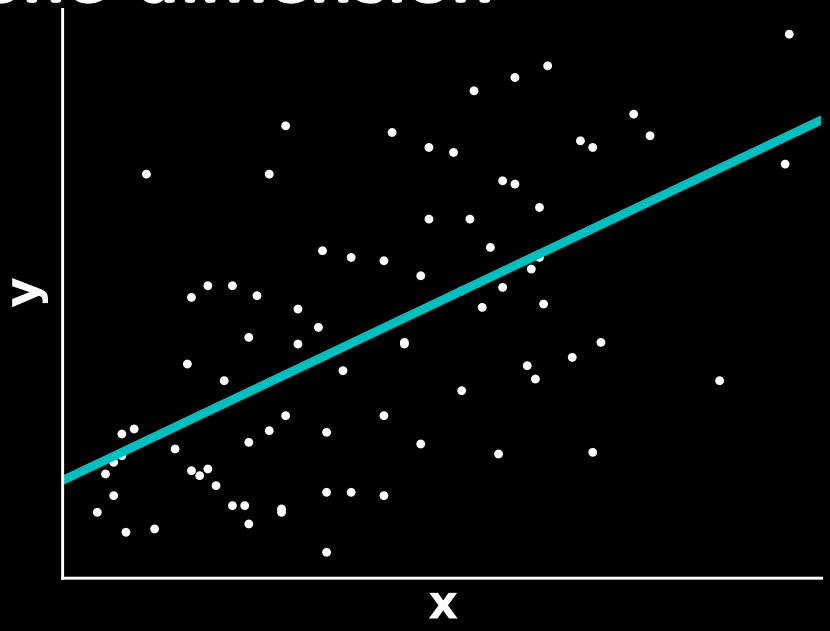
... 0: no errors

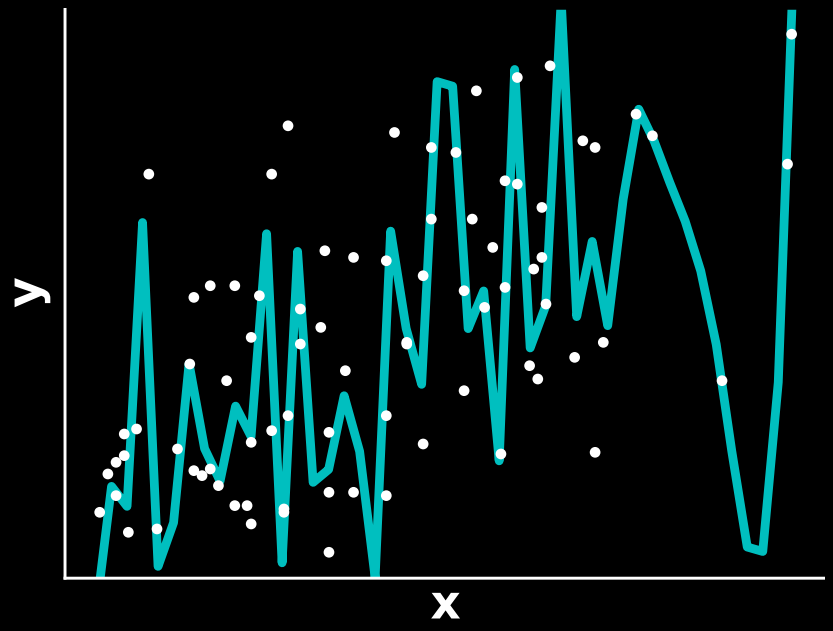
Test data $\neq$ Train data

A single descriptor: one dimension



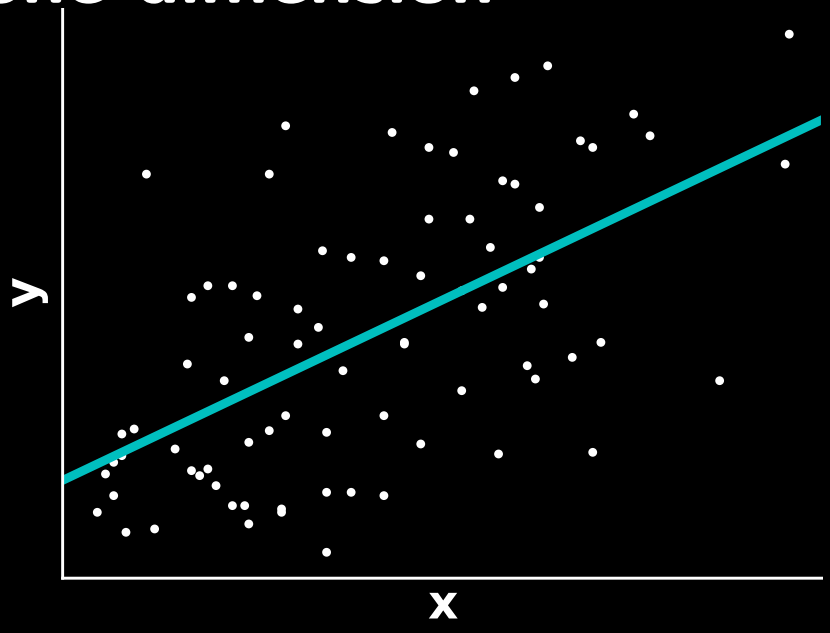
A single descriptor: one dimension

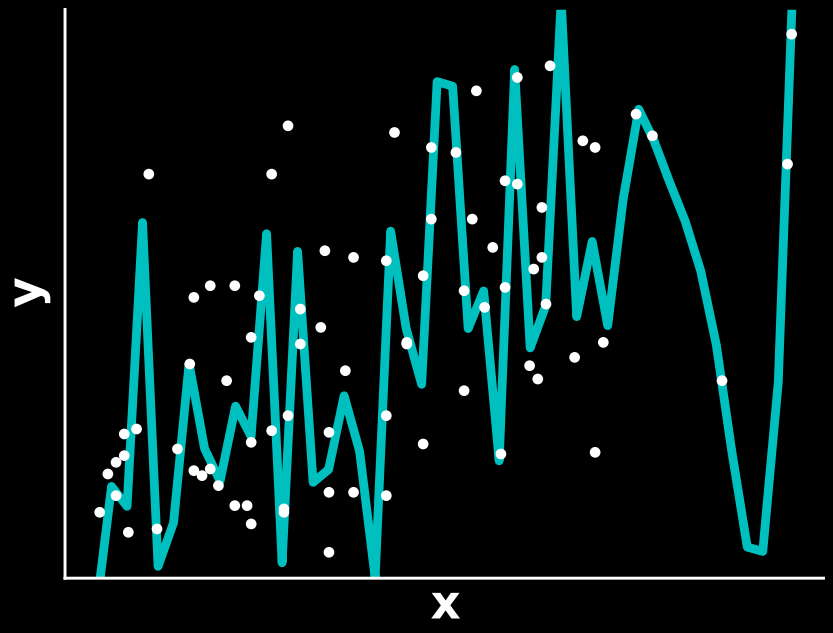




Which model to prefer?

A single descriptor: one dimension

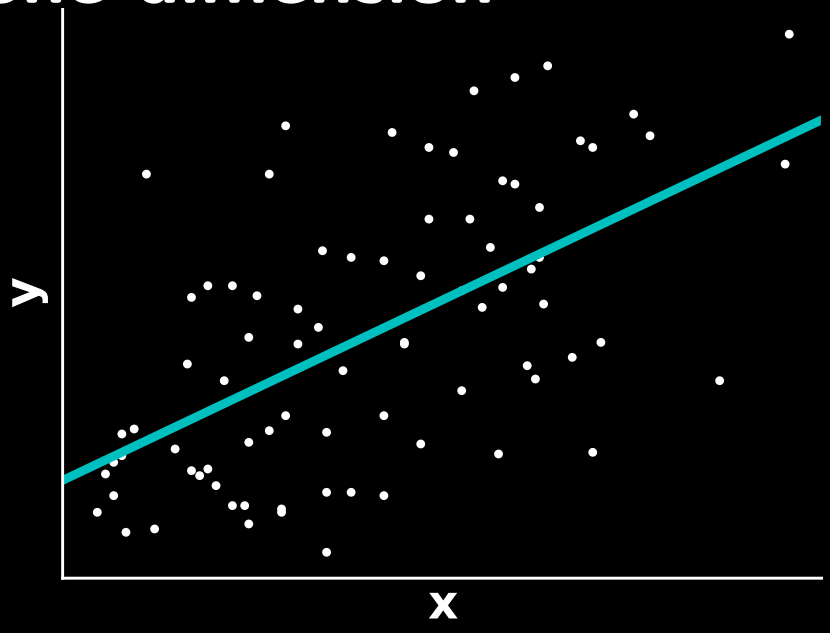


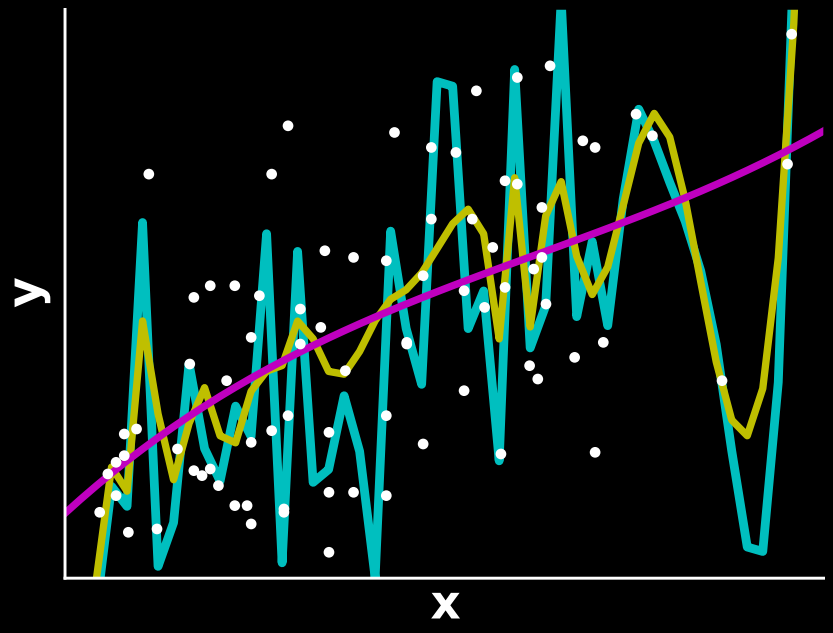


Problem of *"over-fitting"*

- Minimizing error is not always the best strategy (learning noise)
- Test data $\neq$ train data

A single descriptor: one dimension



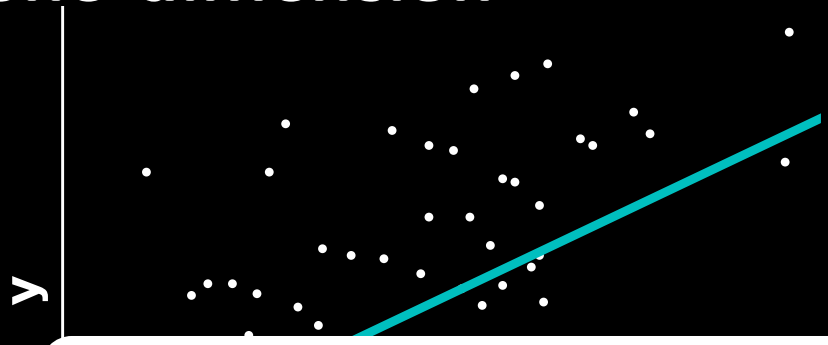


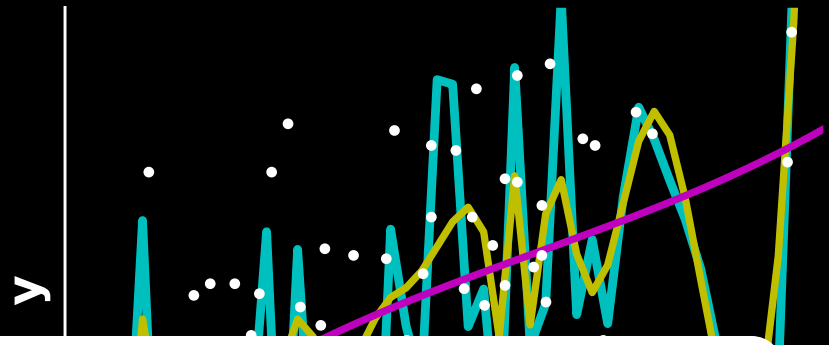
Prefer simple models

= concept of *"regularization"*

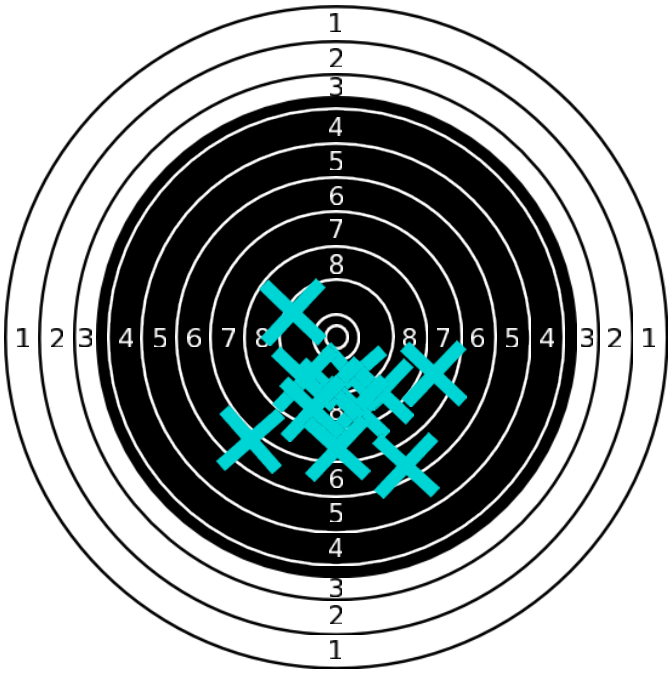
Balance the number of parameters to learn with the amount of data

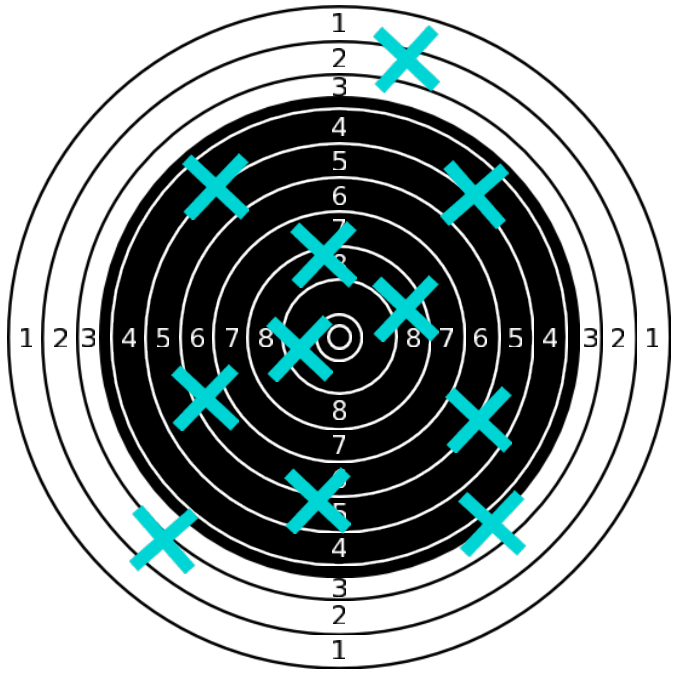
A single descriptor: one dimension





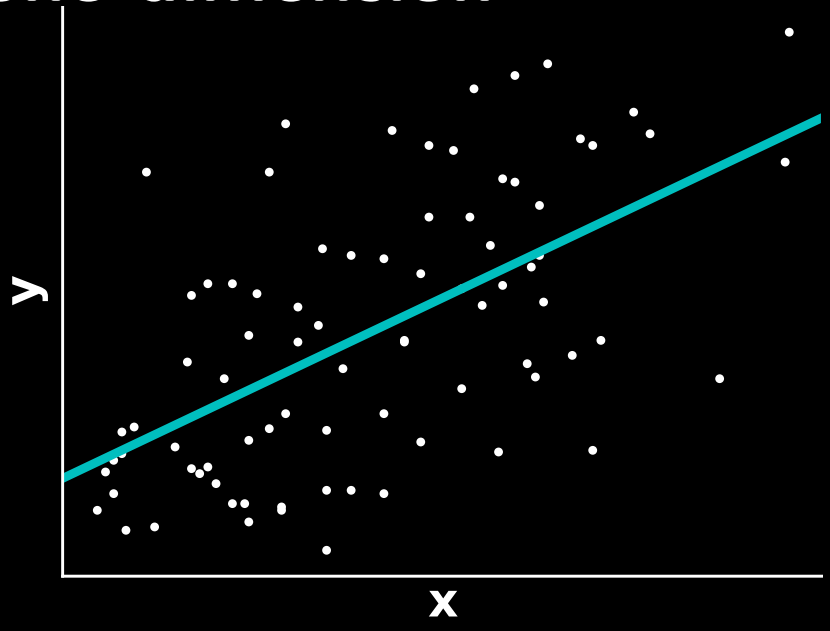
Bias **variance** tradeoff



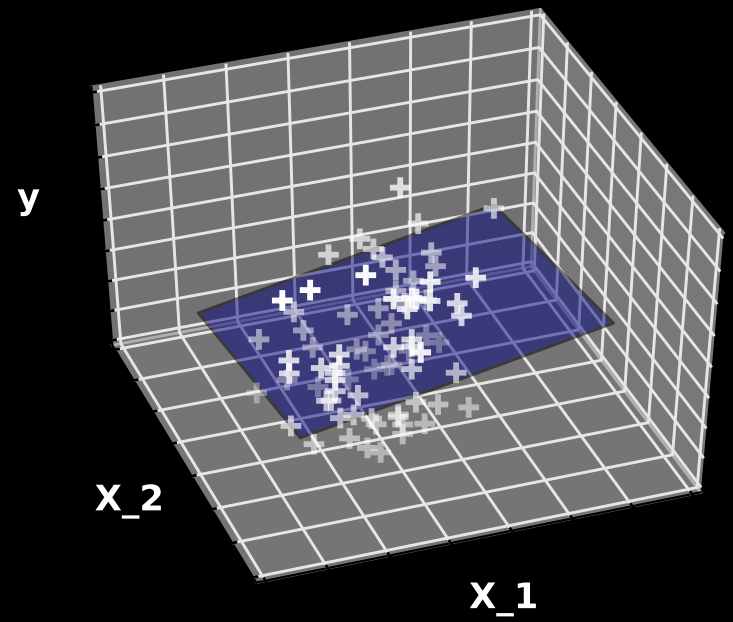


1 Machine learning in a nutshell: regression

A single descriptor: one dimension

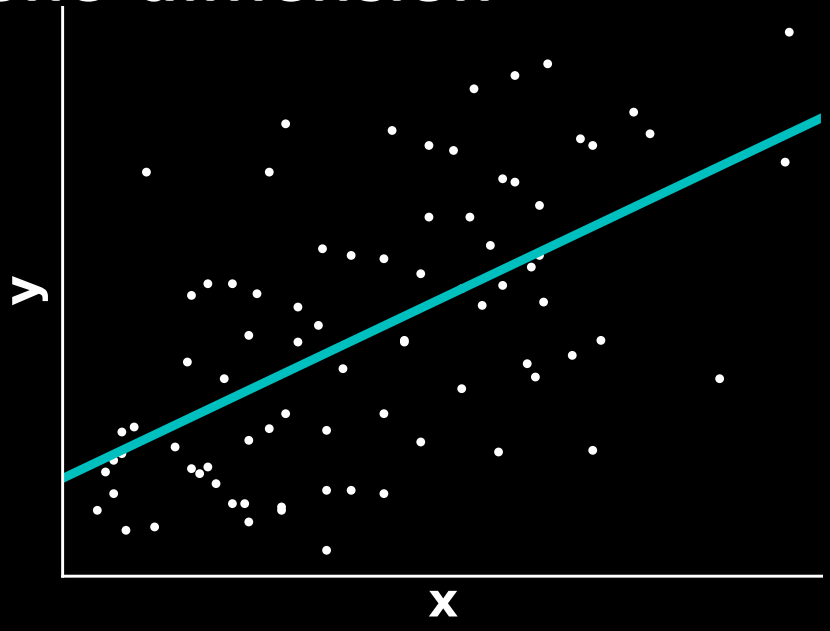


Two descriptors: 2 dimensions

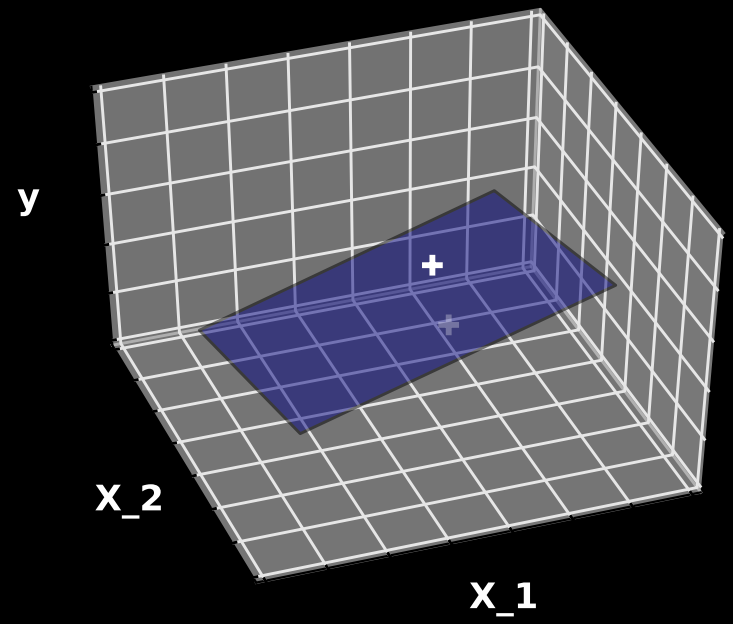


More parameters

A single descriptor: one dimension



Two descriptors: 2 dimensions

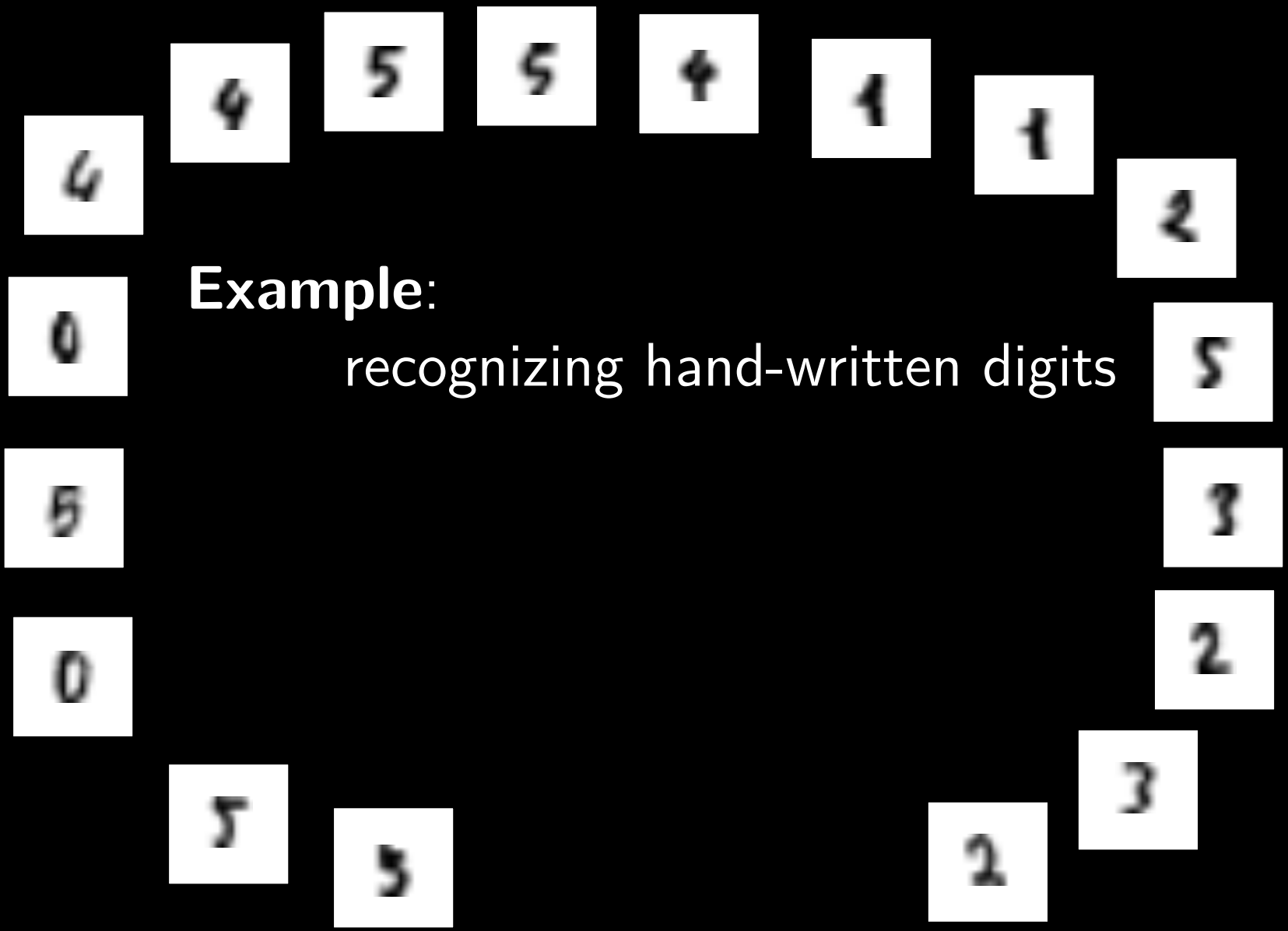


More parameters

$\Rightarrow$ need more data

"curse of dimensionality"

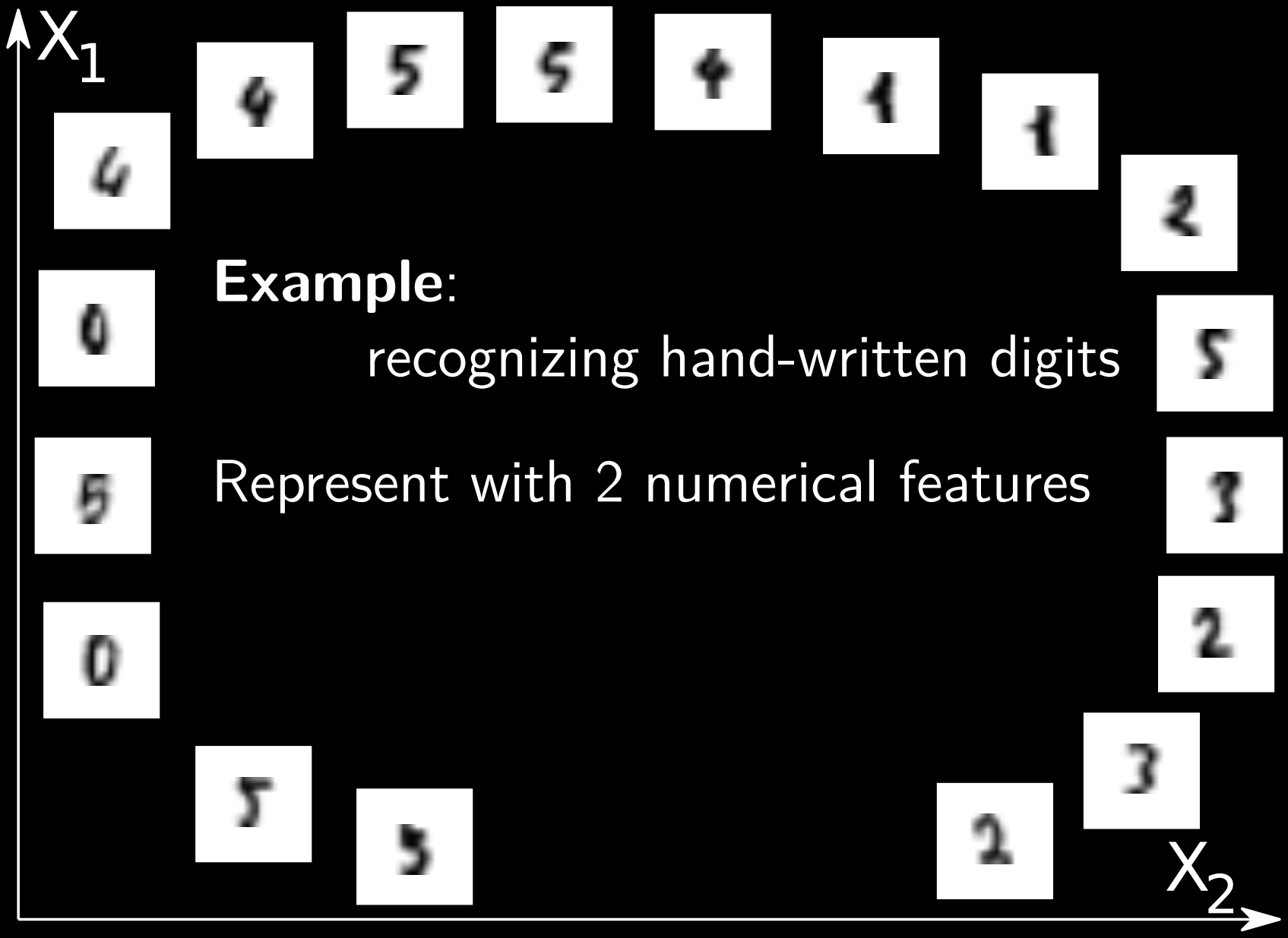
1 Machine learning in a nutshell: classification



Example:

recognizing hand-written digits

1 Machine learning in a nutshell: classification

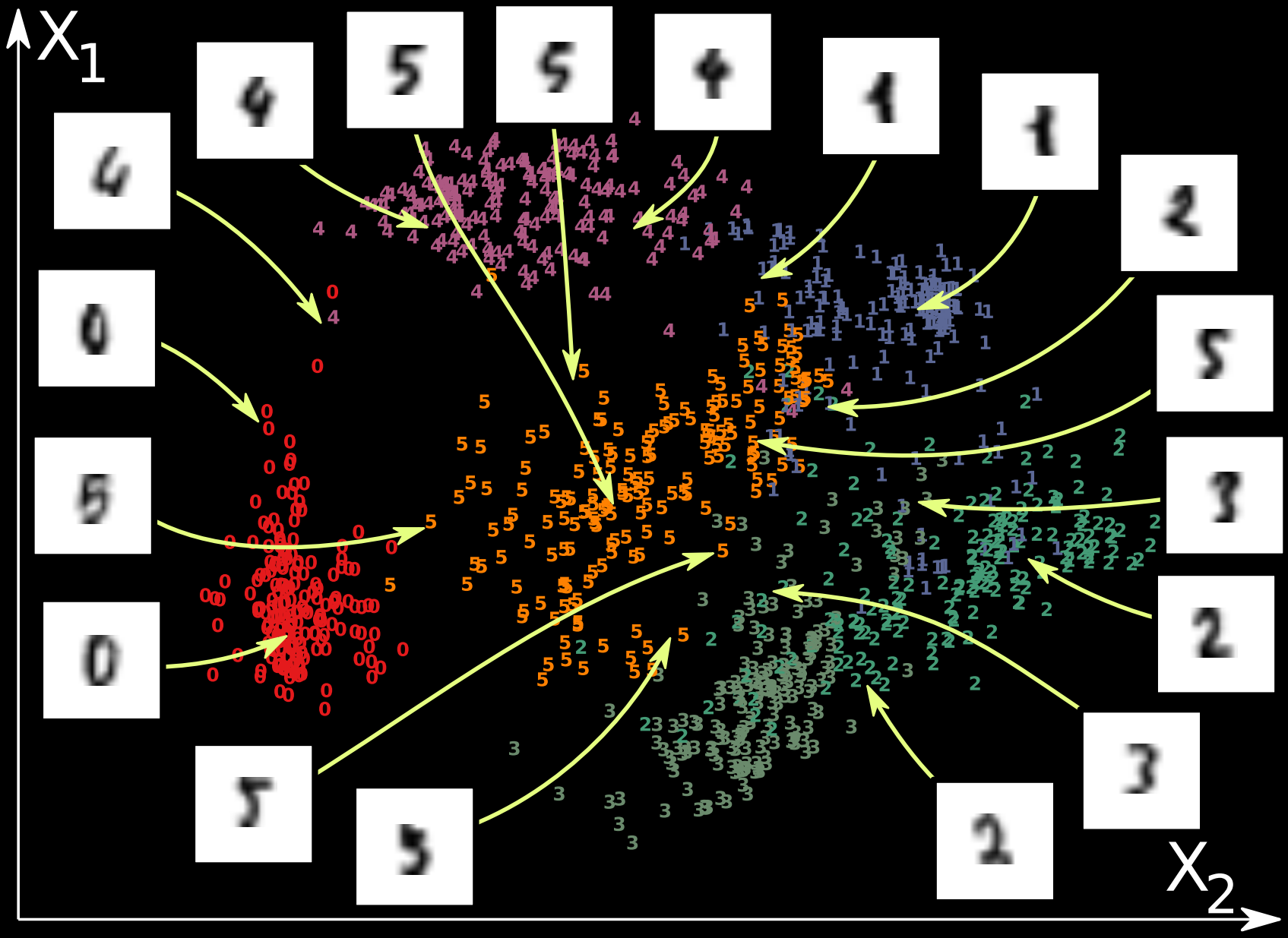


Example:

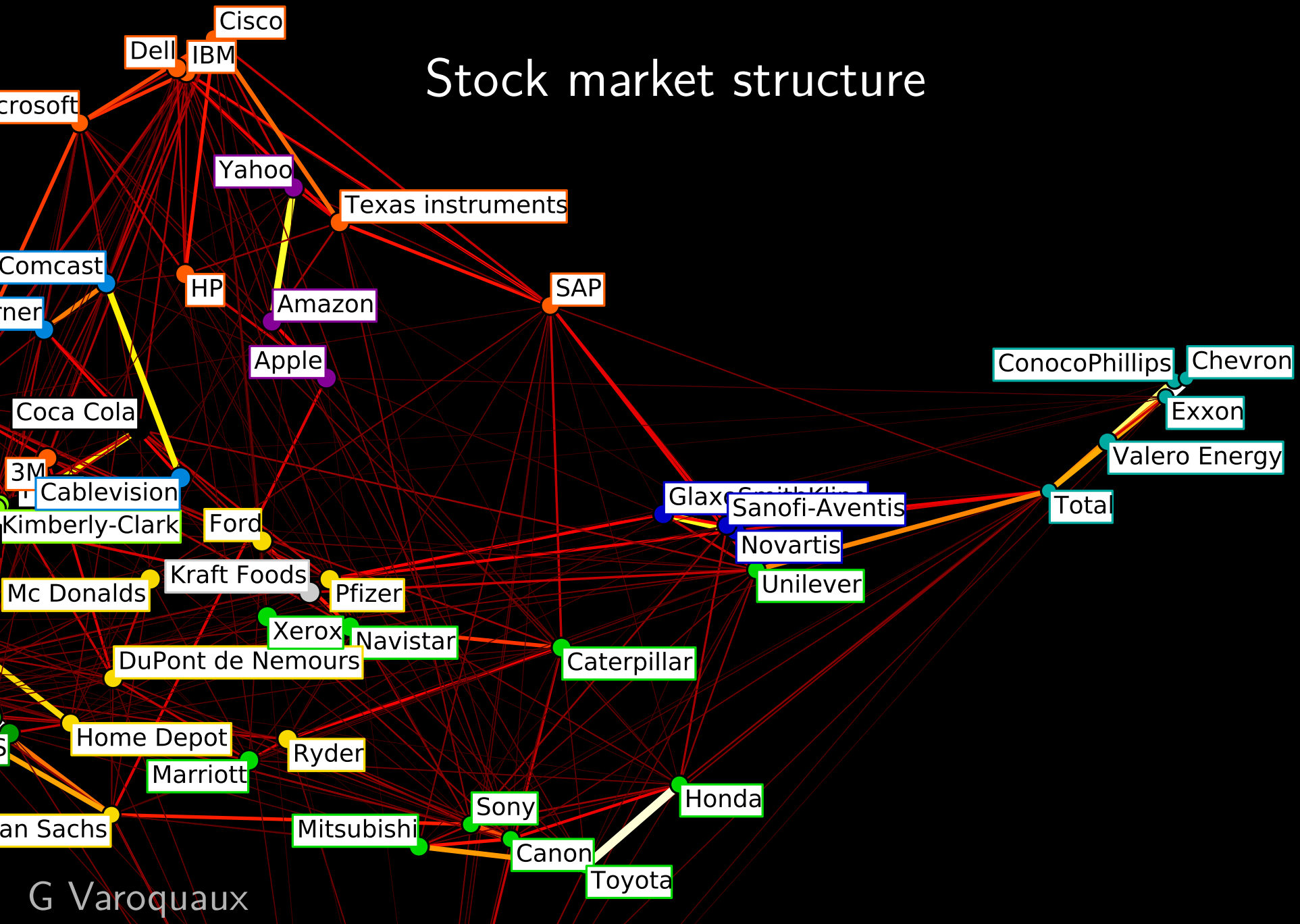
recognizing hand-written digits

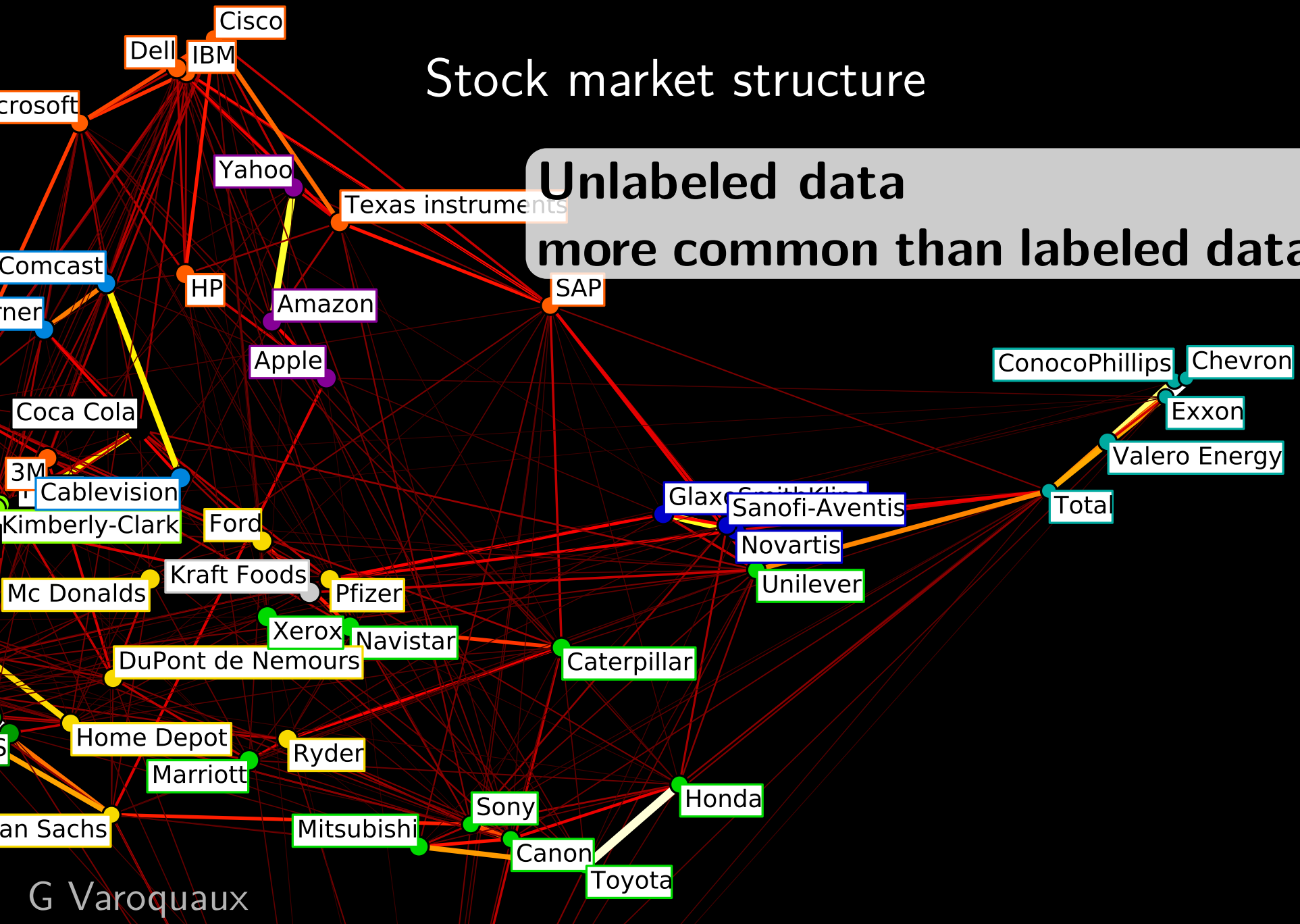
Represent with 2 numerical features

1 Machine learning in a nutshell: classification



1 Machine learning in a nutshell: unsupervised

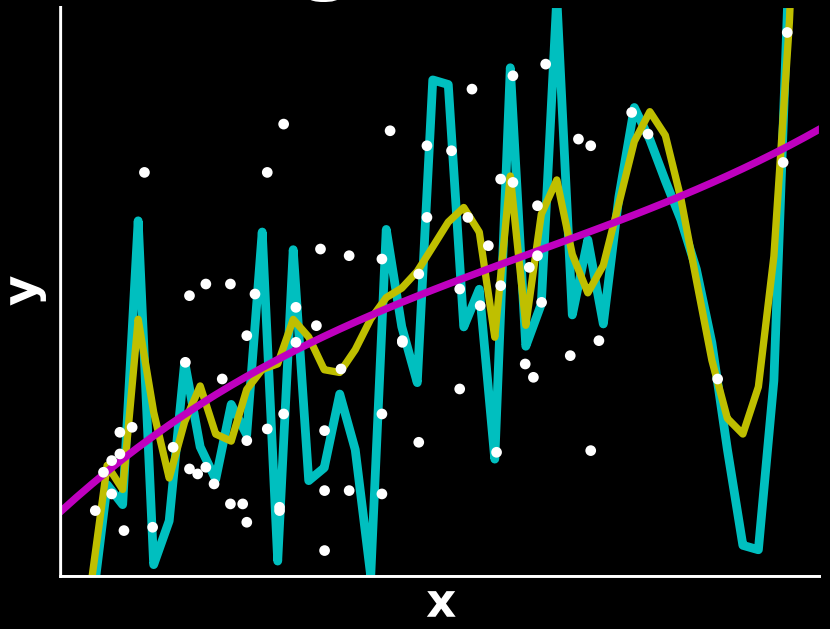



Stock market structure
Unlabeled data
more common than labeled data
Cisco
Dell
IBM
crosoft
Yahoo
Texas instruments
Comcast
HP
Amazon
SAP
rner
Apple
ConocoPhillips
Chevron
Exxon
Coca Cola
Valero Energy
3M
Cablevision
Glaxo
Sanofi-Aventis
Total
Kimberly-Clark
Ford
Novartis
Kraft Foods
Unilever
Mc Donalds
Pfizer
Xerox
Navistar
DuPont de Nemours
Caterpillar
Home Depot
Ryder
Marriott
Honda
Sony
an Sachs
Mitsubishi
Canon
Toyota

Machine learning

Mathematics and algorithms for fitting predictive models

Regression



Classification

Notions of overfit and test error

Machine learning is everywhere

- Image recognition
- Marketing (click-through rate)
- Movie / music recommendation
- Medical data
- Logistic chains (eg supermarkets)
- Language translation
- Detecting industrial failures

We built a machine learning library

In Python, in 2010

Real statisticians use R

- And real astronomers use IRAF
- Real economists use Gauss
- Real coders use ~~C~~ assembler
- Real experiments are controlled in Labview
- Real Bayesians use ~~BUGS~~ stan
- Real text processing is done in Perl
- Real Deep learner is best done with torch (Lua)
- And medical doctors only trust SPSS

Python, what else?

- General purpose
- Interactive language
- Easy to read / write

Fantastic scientific ecosystem

1 scikit-learn goals

Machine learning for all

No specific application domain

No requirements in machine learning

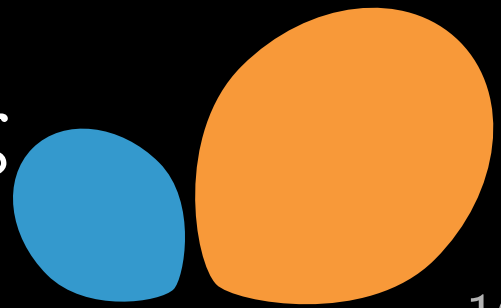
High-quality Pythonic software library

Interfaces designed for users

Community-driven development

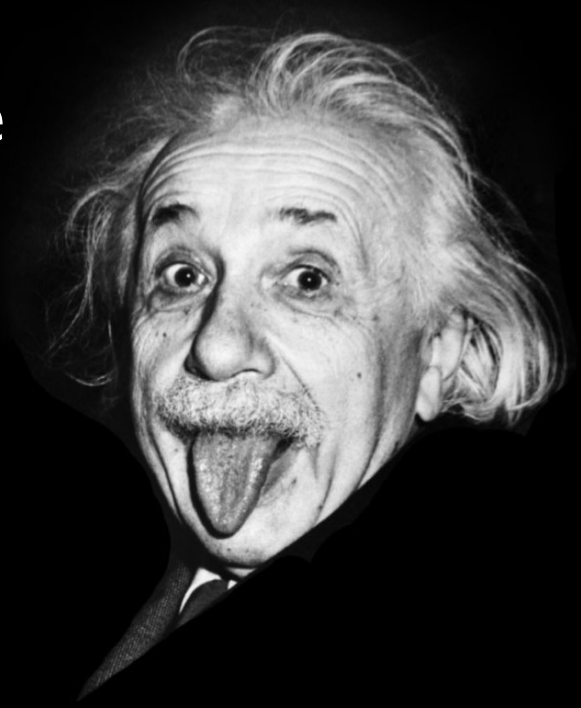
BSD licensed, very diverse contributors

`http://scikit-learn.org`



Machine learning research

- Conceptual complexity is not an issue
- New and bleeding edge is better
- Simple problems are old science



In the field

- Tried and tested (aka boring) is good
- Little sophistication from the user
- API is more important than maths

Solving simple problems matters
Solving them really well matters a lot