MACHINE LEARNING

INTRODUCTION FOR SOFTWARE DEVELOPERS

NIKLAS ANTONČIĆ

CADEC 2018.03.08 | CALLISTAENTERPRISE.SE

CALLISTA

— ENTERPRISE —

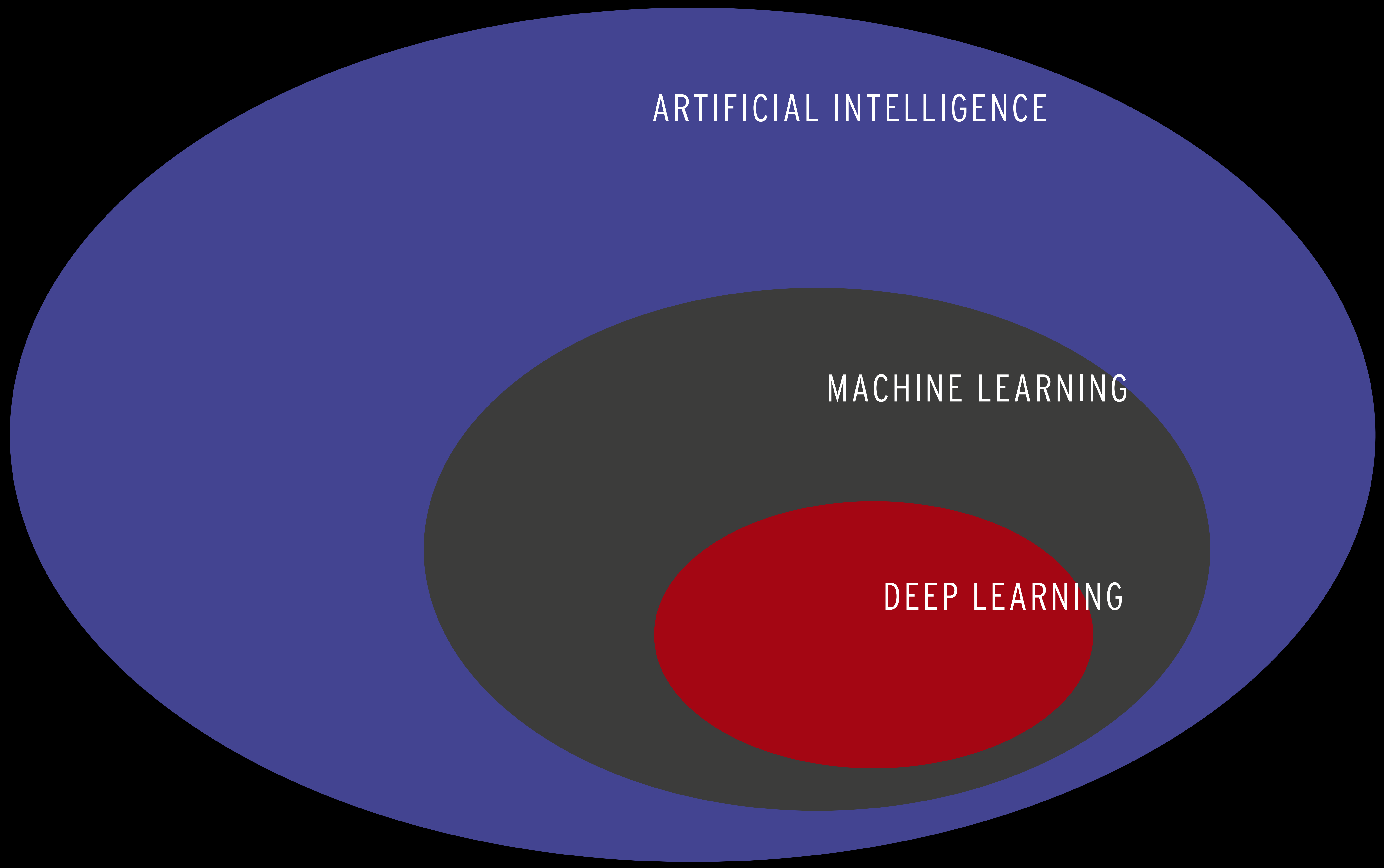
AGENDA

- Introduction and context
- The work process
- The learning problem
- Validation and overfitting
- Tools
- Risks and ethics
- Demo

AGENDA

- **Introduction and context**
- The work process
- The learning problem
- Validation and overfitting
- Tools
- Risks and ethics
- Demo

AI VS ML VS DL



WHAT CLASS OF PROBLEMS DOES MACHINE LEARNING SOLVE?

WHAT CLASS OF PROBLEMS DOES MACHINE LEARNING SOLVE?

Complex problems where the human brain cannot find an analytical solution.

PREREQUISITES

PREREQUISITES

- No analytical solution known

PREREQUISITES

- No analytical solution known
- A pattern, a hunch of the problem domain

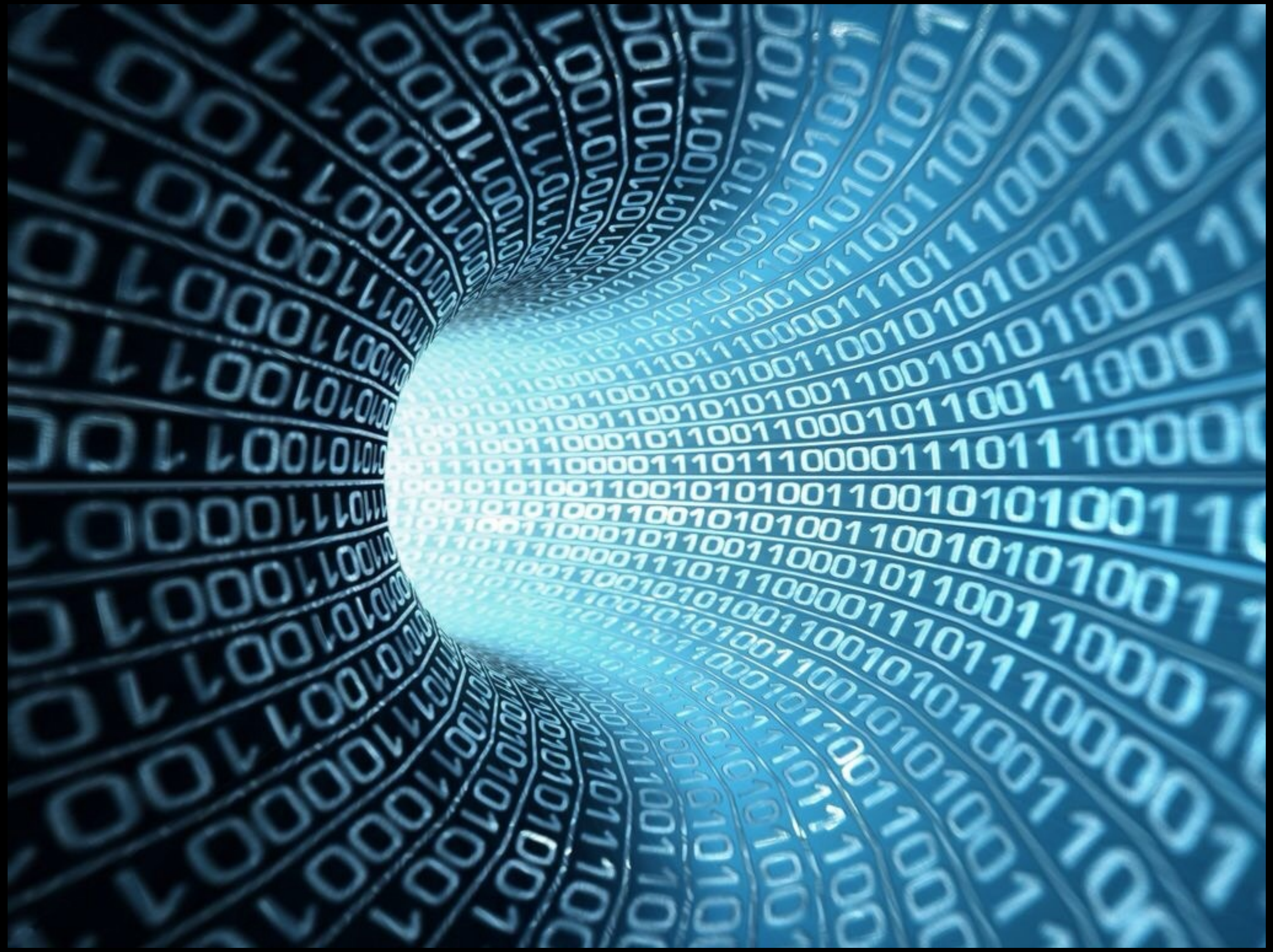
PREREQUISITES

- No analytical solution known
- A pattern, a hunch of the problem domain
- Lots of data

THE HYPE - WHY NOW?

THE HYPE - WHY NOW?

- IoT, Web-scale, Big Data



THE HYPE - WHY NOW?

- IoT, Web-scale, Big Data
- CPU perfomance vs GPU performance

THE HYPE - WHY NOW?

- IoT, Web-scale, Big Data
- CPU perfomance vs GPU performance
- Deep Learning (Google Brain, 2012)

Andrew Y. Ng, a Stanford computer scientist, is cautiously optimistic about neural networks.
Jim Wilson/The New York Times

DIFFERENT ML PARADIGMS

DIFFERENT ML PARADIGMS

- Supervised learning

DIFFERENT ML PARADIGMS

- Supervised learning
- Unsupervised learning

DIFFERENT ML PARADIGMS

- Supervised learning
- Unsupervised learning
- Reinforced learning

REAL WORLD EXAMPLES

REAL WORLD EXAMPLES

REAL WORLD EXAMPLES

REAL WORLD EXAMPLES

REAL WORLD EXAMPLES

AGENDA

- Introduction and context
- **The work process**
- The learning problem
- Validation and overfitting
- Tools
- Risks and ethics
- Demo

THE PROCESS

%

BUSINESS TARGET

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA

PRE PROCESS

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA

PRE PROCESS

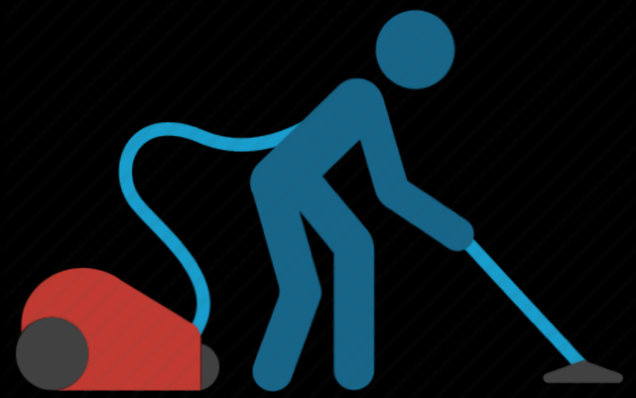
SELECT MODEL

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA



PRE PROCESS

SELECT MODEL

TRAIN

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA

PRE PROCESS

SELECT MODEL

TRAIN

FINAL HYPOTHESIS

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA

PRE PROCESS

SELECT MODEL

TRAIN

VALIDATE RESULT

FINAL HYPOTHESIS

THE PROCESS

%

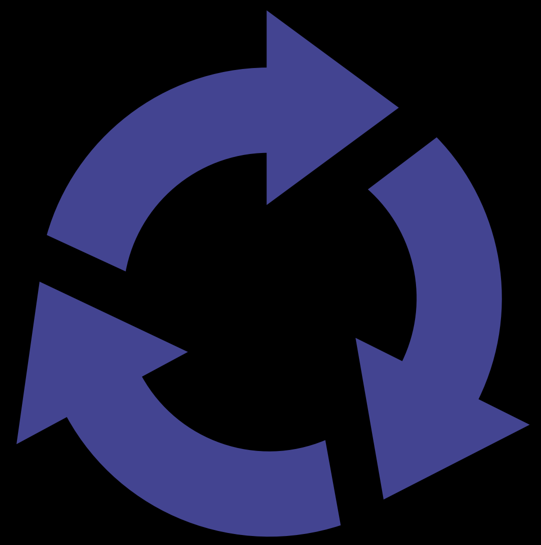
BUSINESS TARGET

AQUIRE RAW DATA

PRE PROCESS

SELECT MODEL

TRAIN



TRIM OR CHANGE MODEL

VALIDATE RESULT

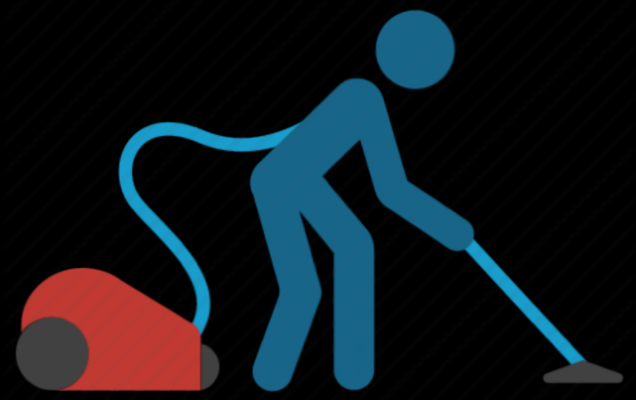
FINAL HYPOTHESIS

THE PROCESS

%

BUSINESS TARGET

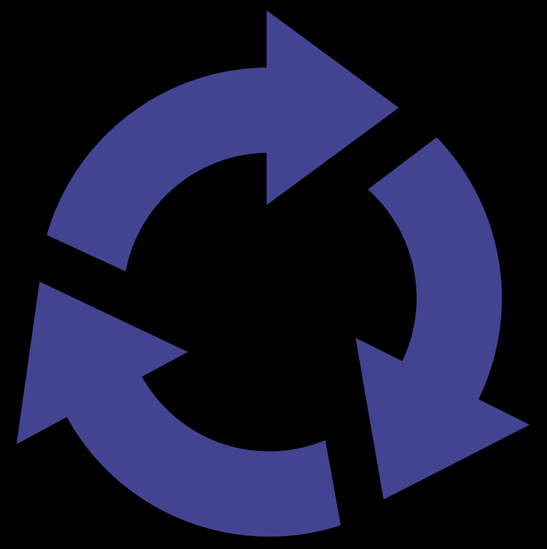
AQUIRE RAW DATA



PRE PROCESS

SELECT MODEL

TRIM OR CHANGE MODEL



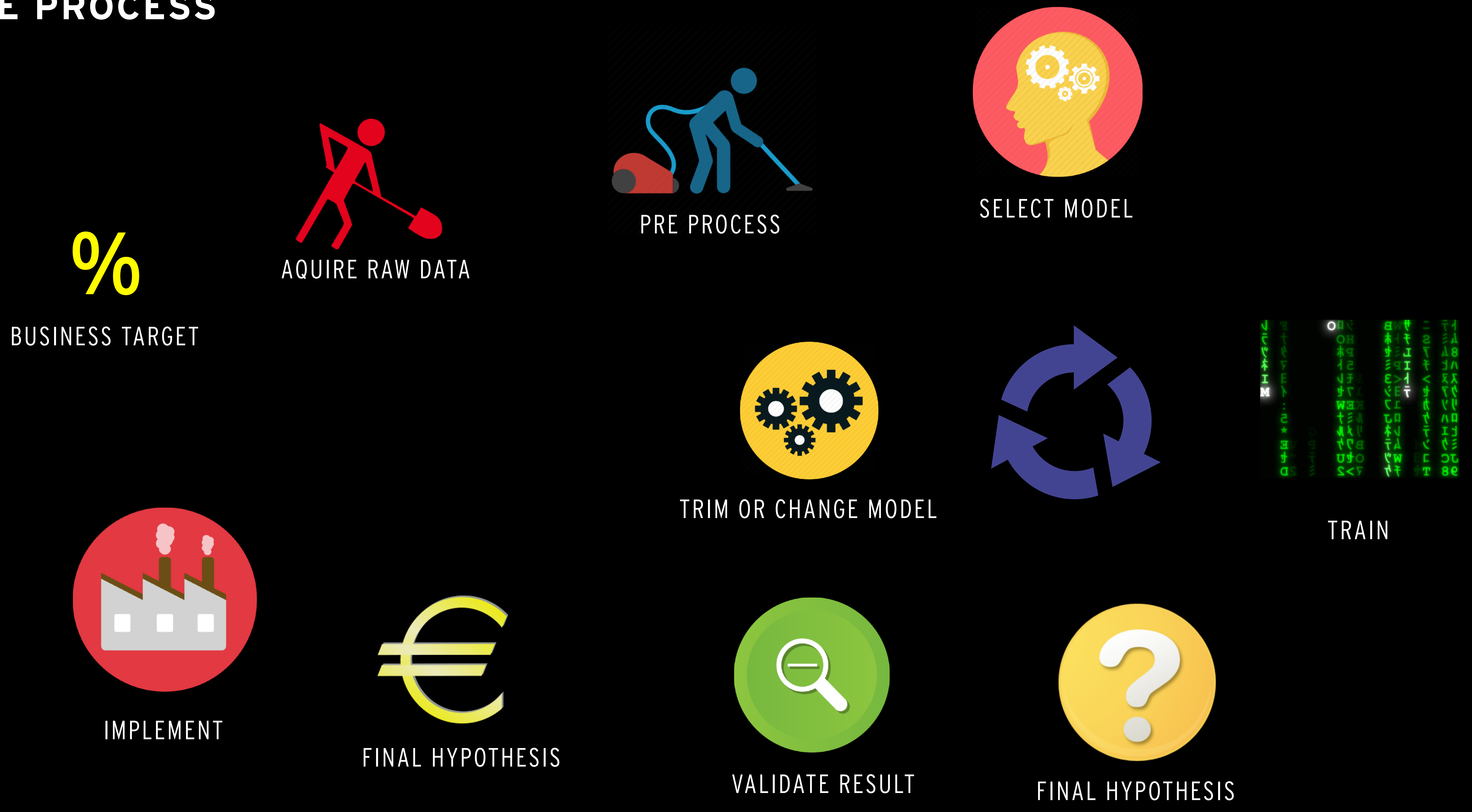
TRAIN

FINAL HYPOTHESIS

VALIDATE RESULT

FINAL HYPOTHESIS

THE PROCESS



SELECT MODEL
PRE PROCESS
AQUIRE RAW DATA
%
BUSINESS TARGET
TRIM OR CHANGE MODEL
TRAIN
IMPLEMENT
FINAL HYPOTHESIS
VALIDATE RESULT
FINAL HYPOTHESIS

AGENDA

- Introduction and context
- The work process
- **The learning problem**
- Validation and overfitting
- Tools
- Risks and ethics
- Demo

EXAMPLE: CREDIT APPROVAL

APPLICATION (INPUT):

Age	34
Yearly Income	400 000
Years in residence	6
Loans	2 000 000

CORRECT CREDIT DECISION (OUTPUT) :

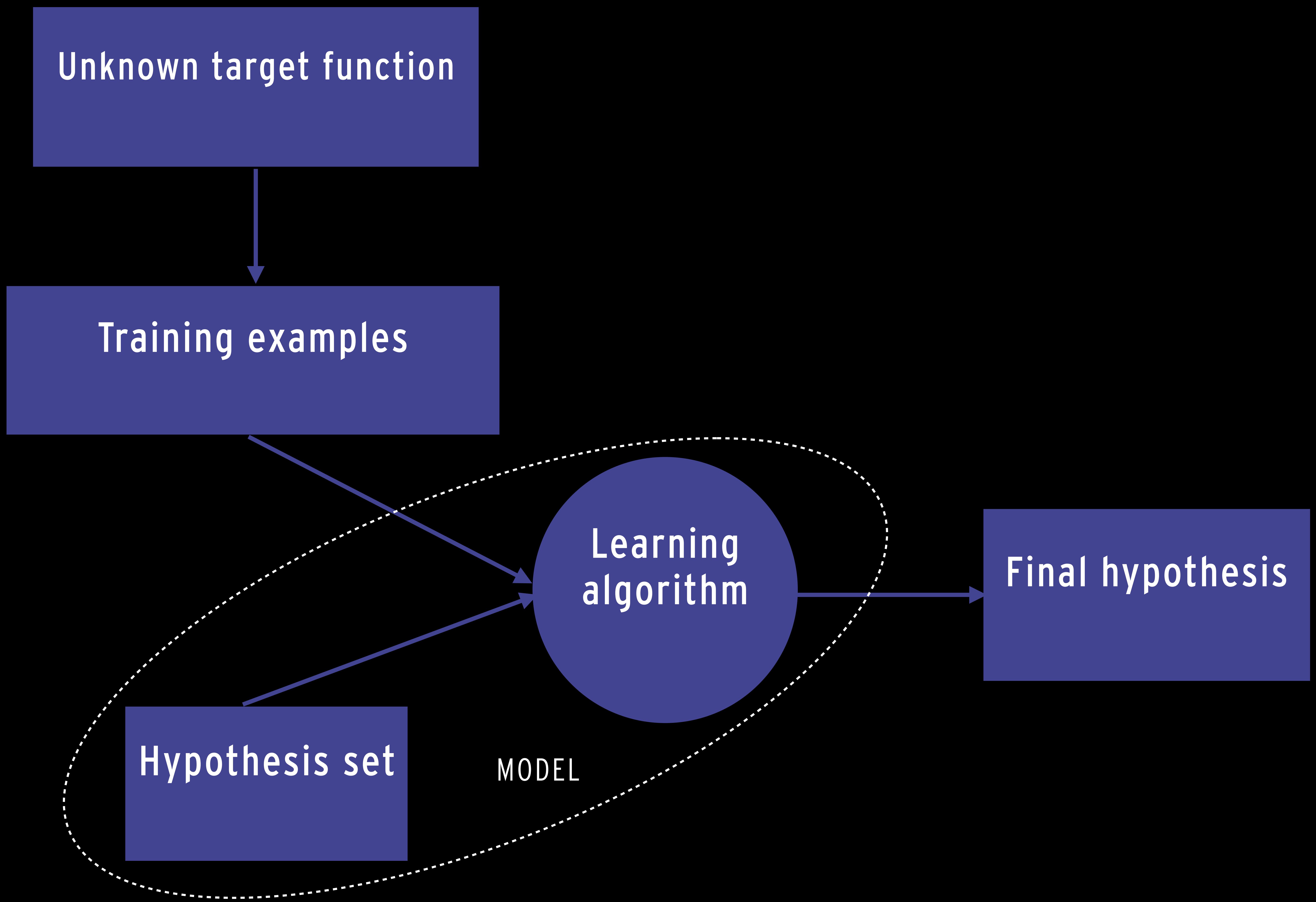
Good customer	yes/no

EXAMPLE: CREDIT APPROVAL

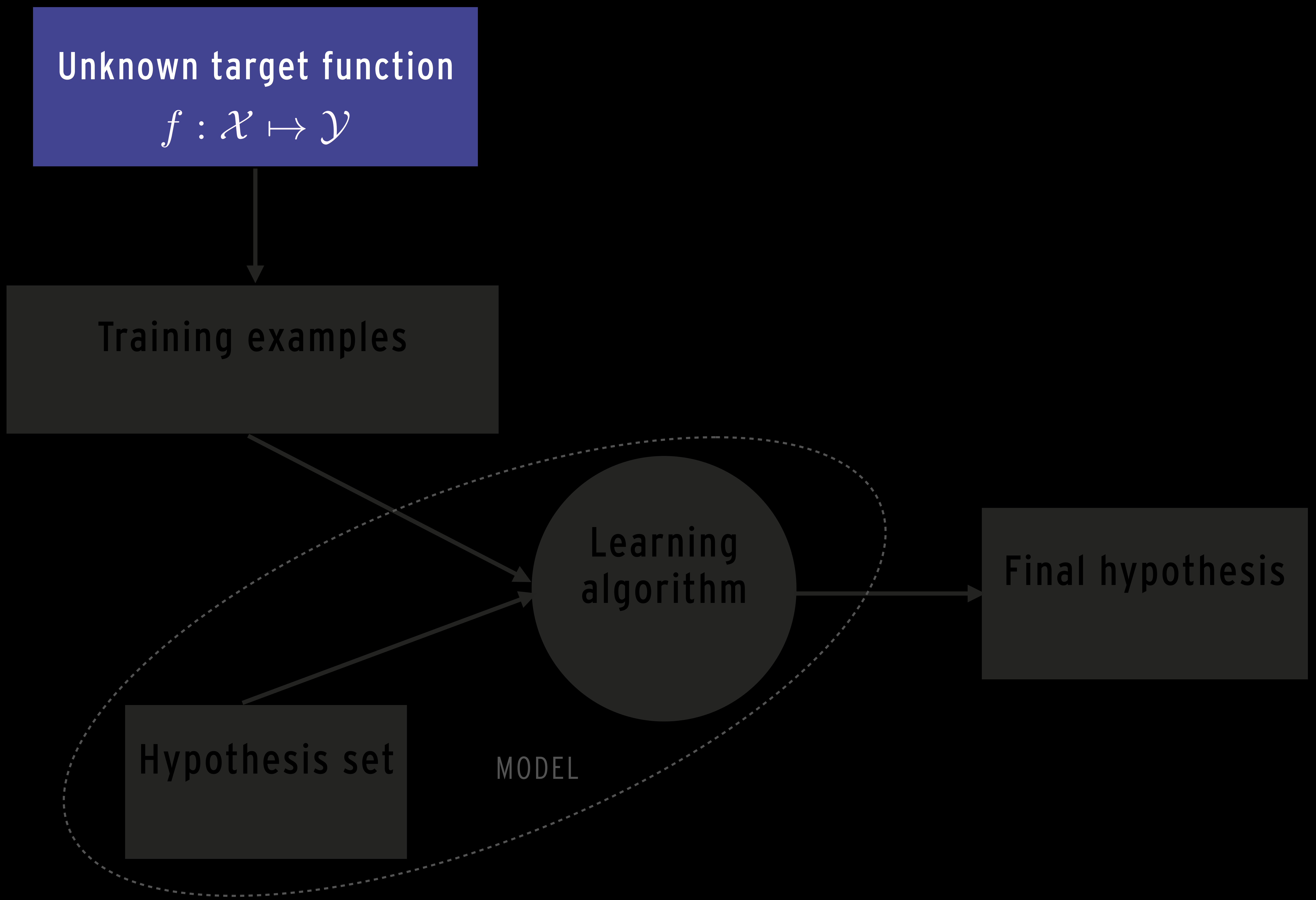
AVAILABLE DATA

	Age	Years in residence	Yearly income	Loans	Good Customer
1	36	4	400000	3000000	Yes
2	54	17	700000	1000000	Yes
…	…	…	…	…	…
N	18	1	80000	0	No

THE LEARNING PROBLEM - MAIN COMPONENTS



THE UNKNOWN TARGET FUNCTION



THE UNKNOWN TARGET FUNCTION - INPUT AND OUTPUT DATA

APPLICATION (INPUT):

Age	34
Yearly Income	400 000
Years in residence	6
Loans	2 000 000

CORRECT CREDIT DECISION (OUTPUT) :

Good customer	yes/no

THE UNKNOWN TARGET FUNCTION - INPUT AND OUTPUT DATA

APPLICATION (INPUT):

Age	34
Yearly Income	400 000
Years in residence	6
Loans	2 000 000

$x_1, x_2, ..., x_d \in \mathcal{X}$ $\mathcal{X} = \mathbb{R}^d$

CORRECT CREDIT DECISION (OUTPUT) :

Good customer yes/no

THE UNKNOWN TARGET FUNCTION - INPUT AND OUTPUT DATA

APPLICATION (INPUT):

Age	34
Yearly Income	400 000
Years in residence	6
Loans	2 000 000

$x_1, x_2, ..., x_d \in \mathcal{X}$ $\mathcal{X} = \mathbb{R}^d$

CORRECT CREDIT DECISION (OUTPUT) :

Good customer yes/no

THE UNKNOWN TARGET FUNCTION - INPUT AND OUTPUT DATA

APPLICATION (INPUT):

Age	34
Yearly Income	400 000
Years in residence	6
Loans	2 000 000

$$x_1, x_2, ..., x_d \in \mathcal{X} \qquad \mathcal{X} = \mathbb{R}^d$$

$$\mathbf{x} = [x_1, x_2, ..., x_d]$$

CORRECT CREDIT DECISION (OUTPUT) :

Good customer yes/no

THE UNKNOWN TARGET FUNCTION - INPUT AND OUTPUT DATA

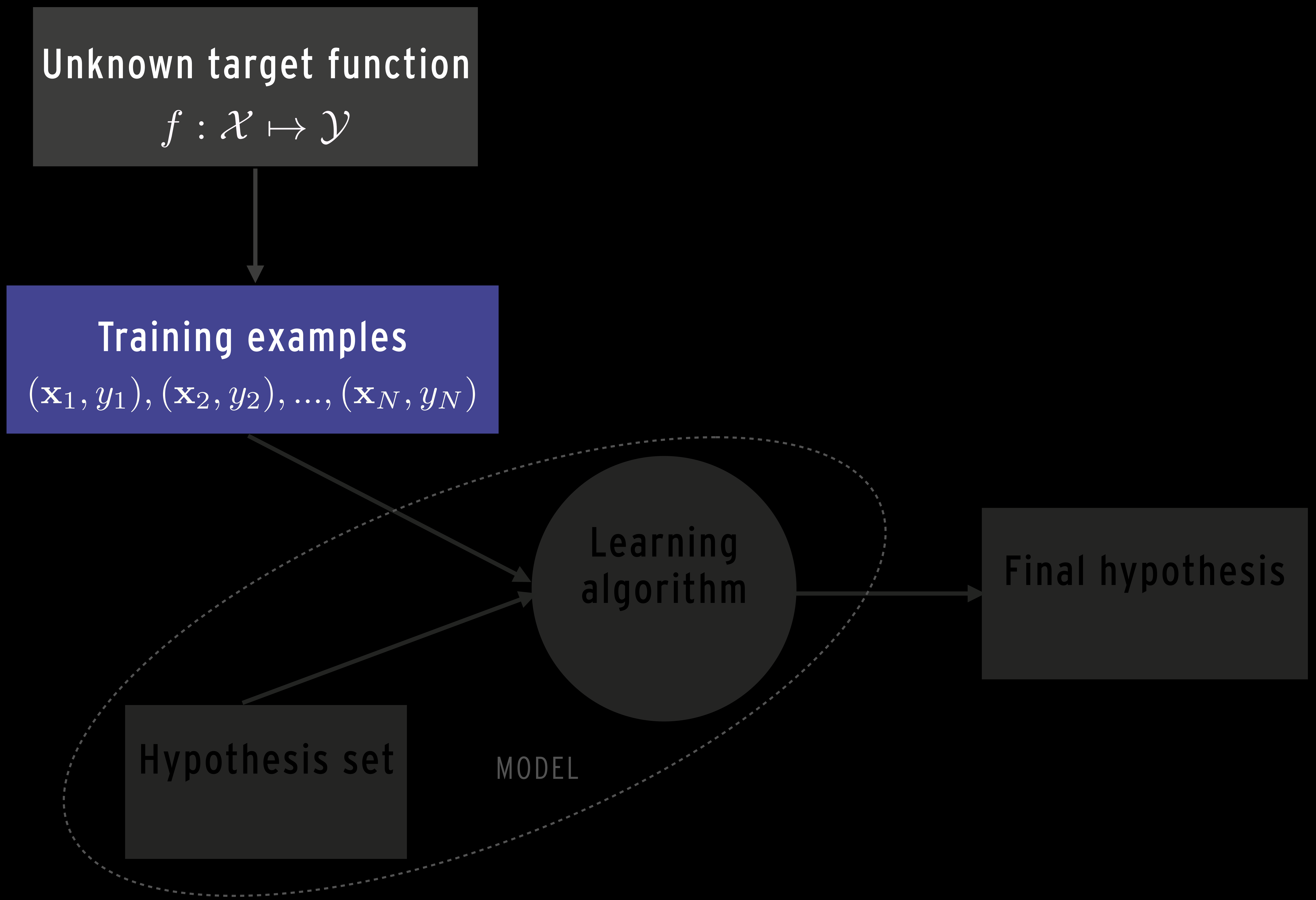
APPLICATION (INPUT):

Age	34	$x_1, x_2, ..., x_d \in \mathcal{X}$	$\mathcal{X} = \mathbb{R}^d$
Yearly Income	400 000		
Years in residence	6	$\mathbf{x} = [x_1, x_2, ..., x_d]$	
Loans	2 000 000		

CORRECT CREDIT DECISION (OUTPUT) :

Good customer	yes/no	$y \in \mathcal{Y}$	$\mathcal{Y} = \{-1, 1\}$

TRAINING EXAMPLES



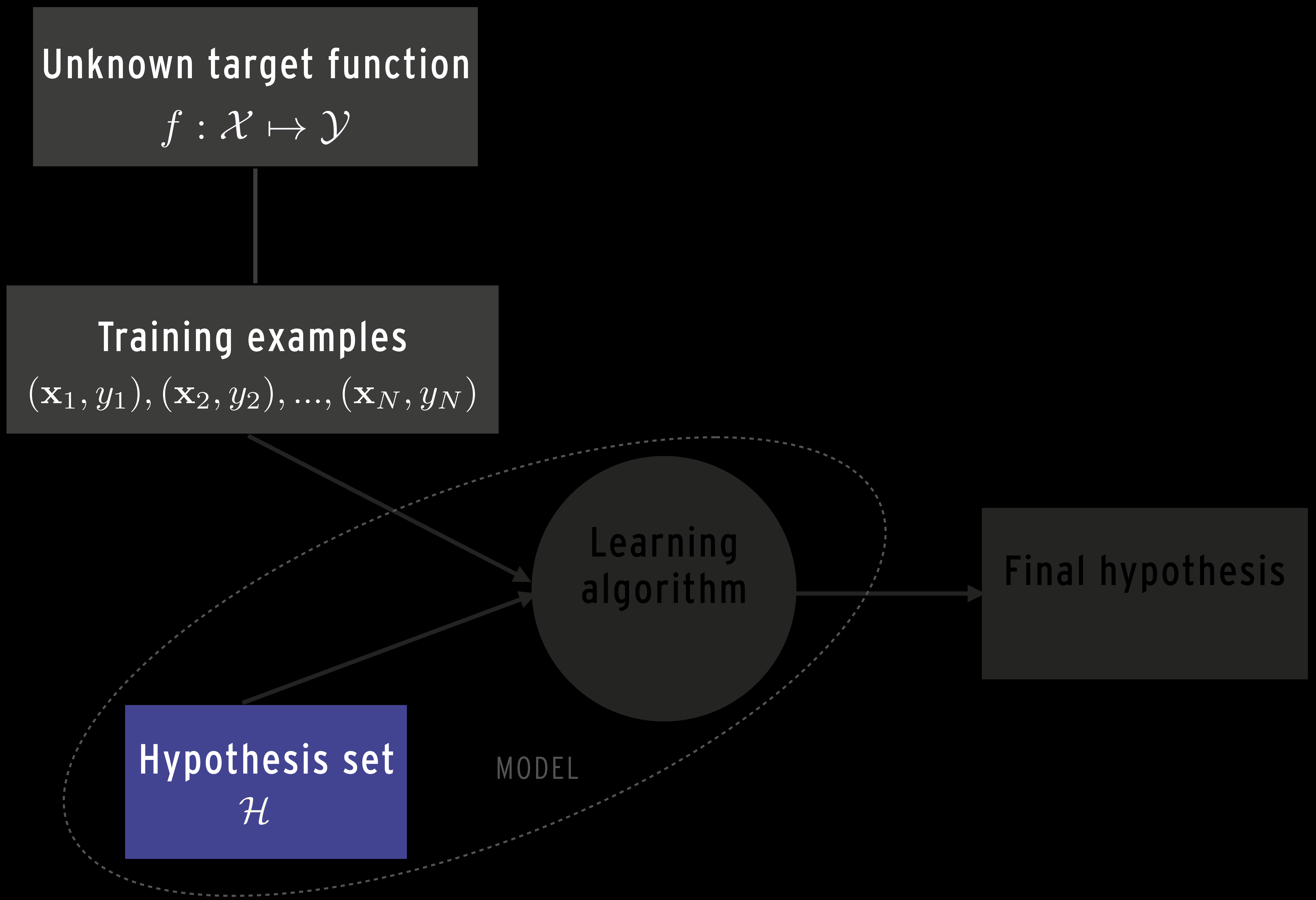
TRAINING EXAMPLES

TRAINING DATA

	Age x_1	Years in residence x_2	Yearly income x_3	Loans x_4	Good Customer y_1
1	36	4	400000	3000000	1
2	54	17	700000	1000000	1
...	...	...	...	...	...
N	20	1	80000	0	-1

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$$

TRAINING EXAMPLES



EXAMPLE: THE HYPOTHESIS SET

$x_1, x_2, ..., x_d$ Has something to do with it …

EXAMPLE: THE HYPOTHESIS SET

$x_1, x_2, ..., x_d$ Has something to do with it …

Lets combine them into a credit score with weights since the attributes has different importance

EXAMPLE: THE HYPOTHESIS SET

$x_1, x_2, ..., x_d$ Has something to do with it …

Lets combine them into a credit score with weights since the attributes has different importance

$$Approve \ \ if \qquad w_1x_1 + w_2x_2 + ... + w_dx_d > threshold$$

EXAMPLE: THE HYPOTHESIS SET

$x_1, x_2, ..., x_d$ Has something to do with it …

Lets combine them into a credit score with weights since the attributes has different importance

$$Approve \ \ if \qquad w_1x_1 + w_2x_2 + ... + w_dx_d > threshold$$

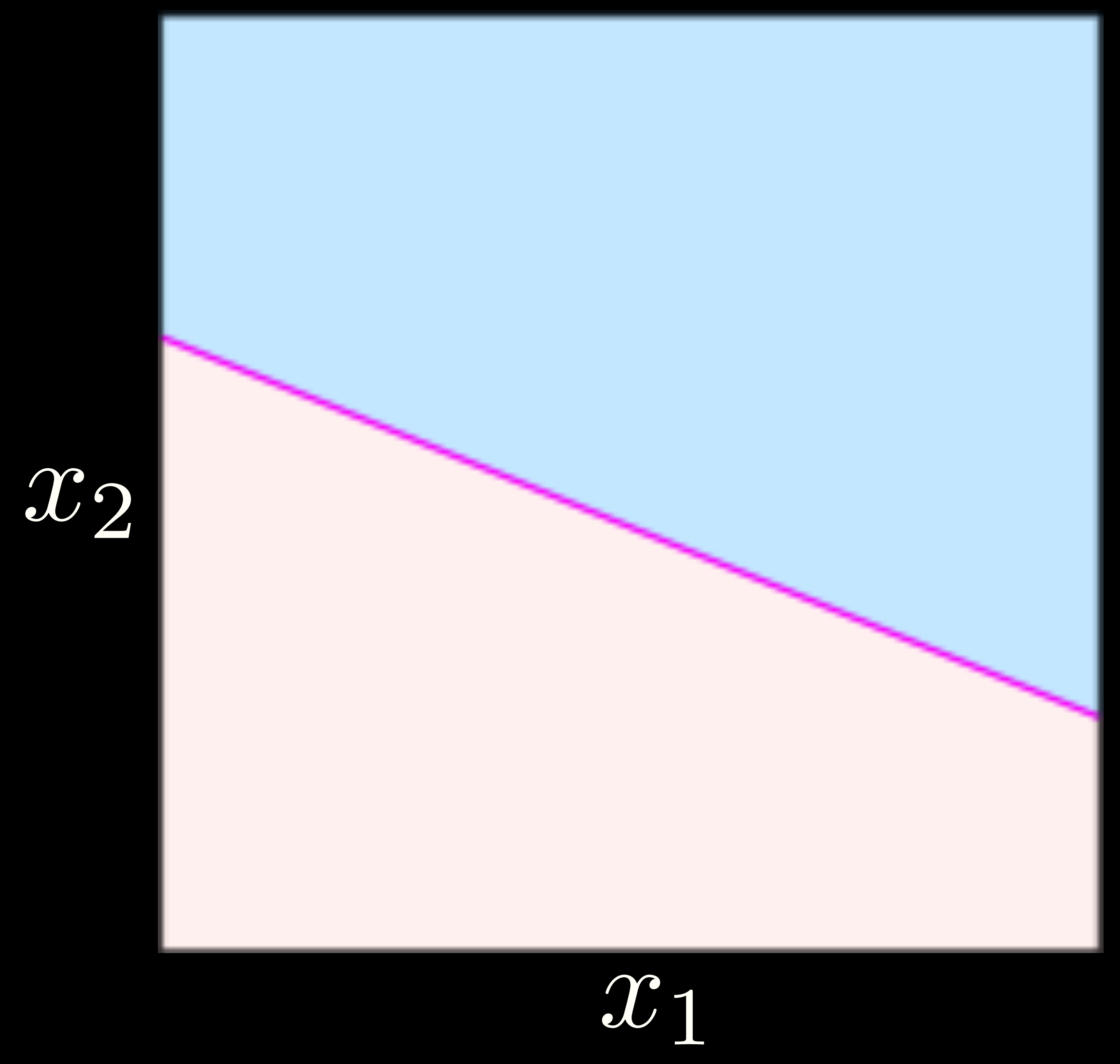
$$Deny \ \ if \qquad w_1x_1 + w_2x_2 + ... + w_dx_d < threshold$$

$$w_1x_1 + w_2x_2 + threshold = 0$$

EXAMPLE: A 2D PERCEPTRON

$$w_1 x_1 + w_2 x_2 + threshold = 0$$

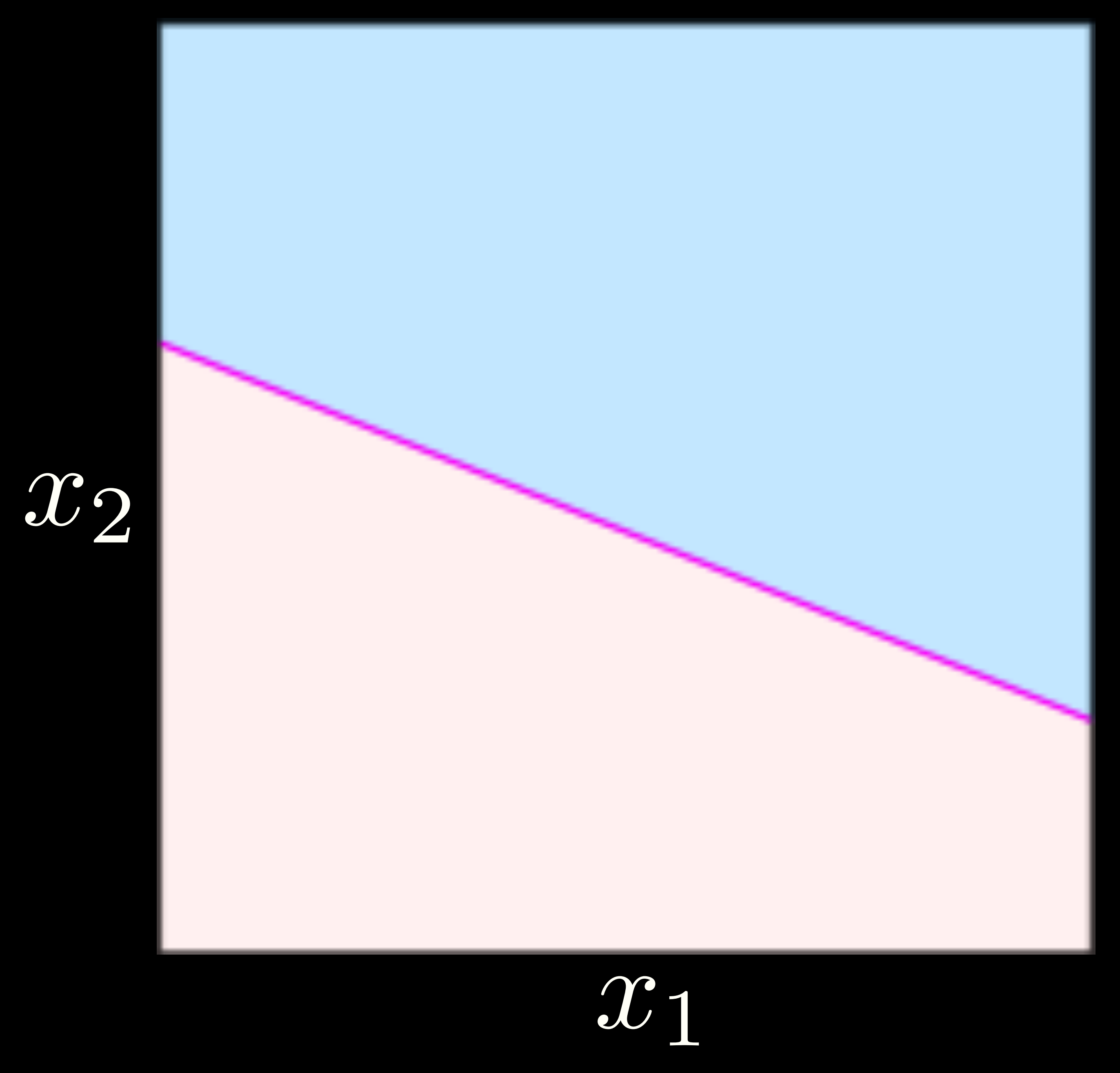
$$w_0 x_0 + w_1 x_1 + w_2 x_2 = 0$$



EXAMPLE: A 2D PERCEPTRON

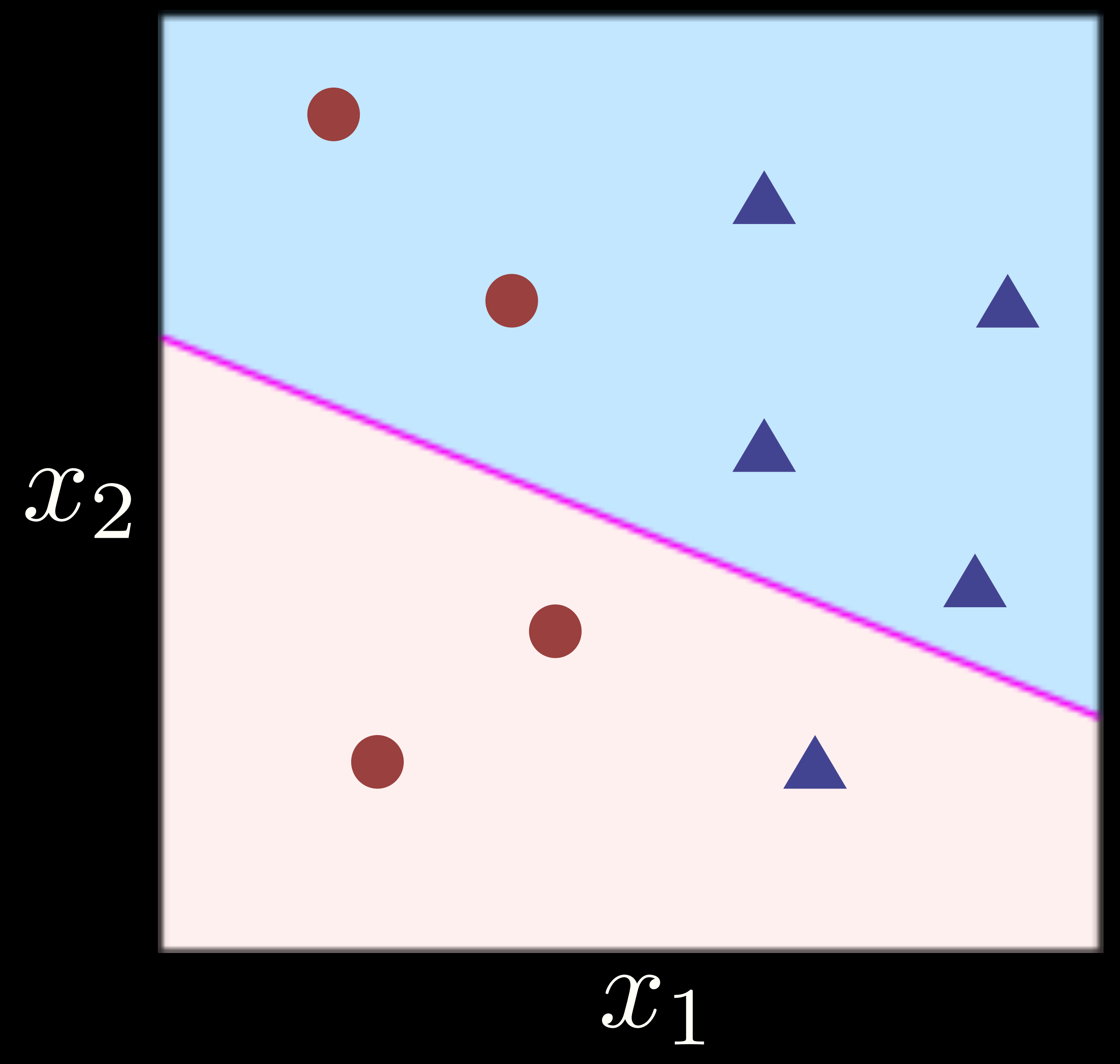
$$w_1 x_1 + w_2 x_2 + threshold = 0$$

$$w_0 x_0 + w_1 x_1 + w_2 x_2 = 0$$



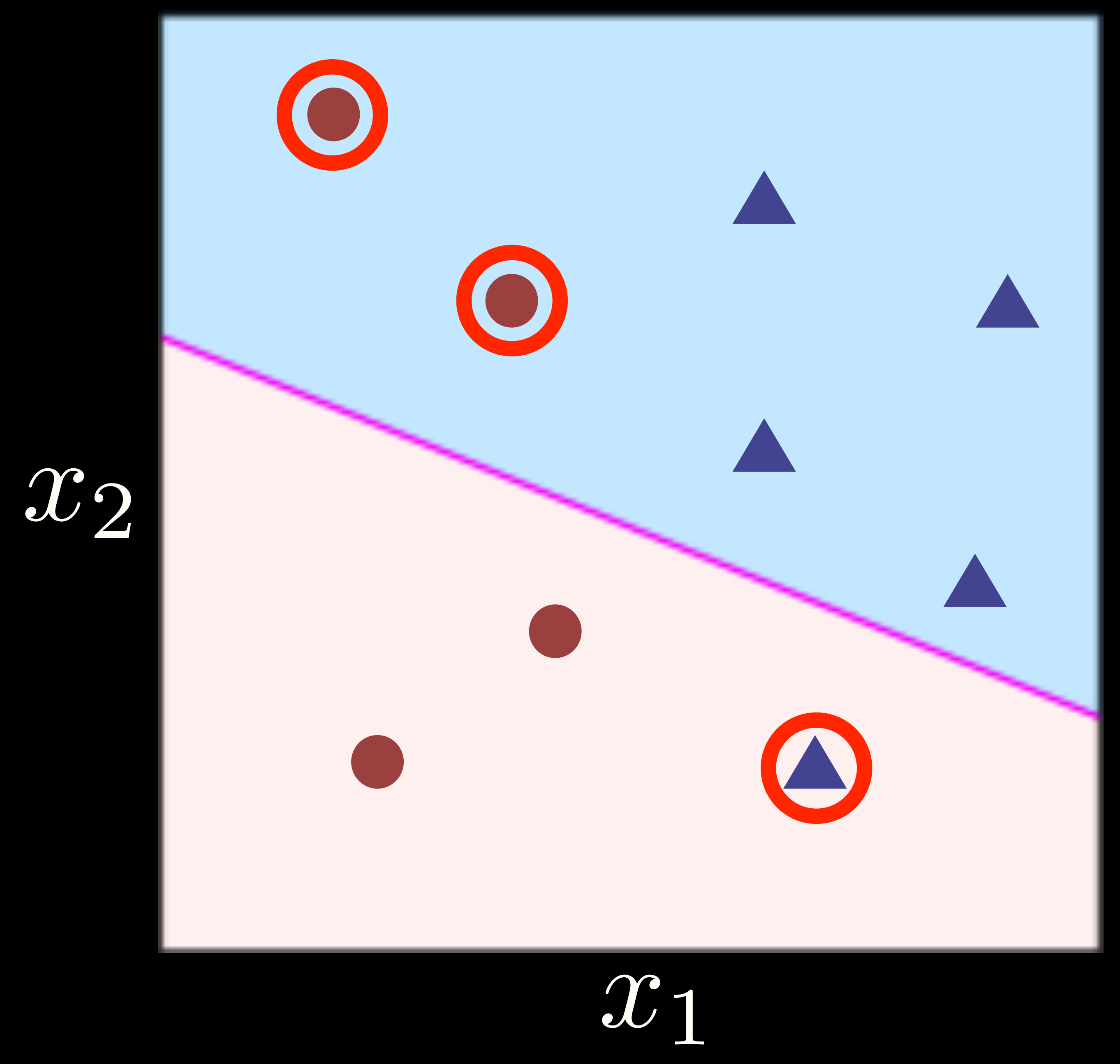
EXAMPLE: A 2D PERCEPTRON

$$w_1 x_1 + w_2 x_2 + threshold = 0$$



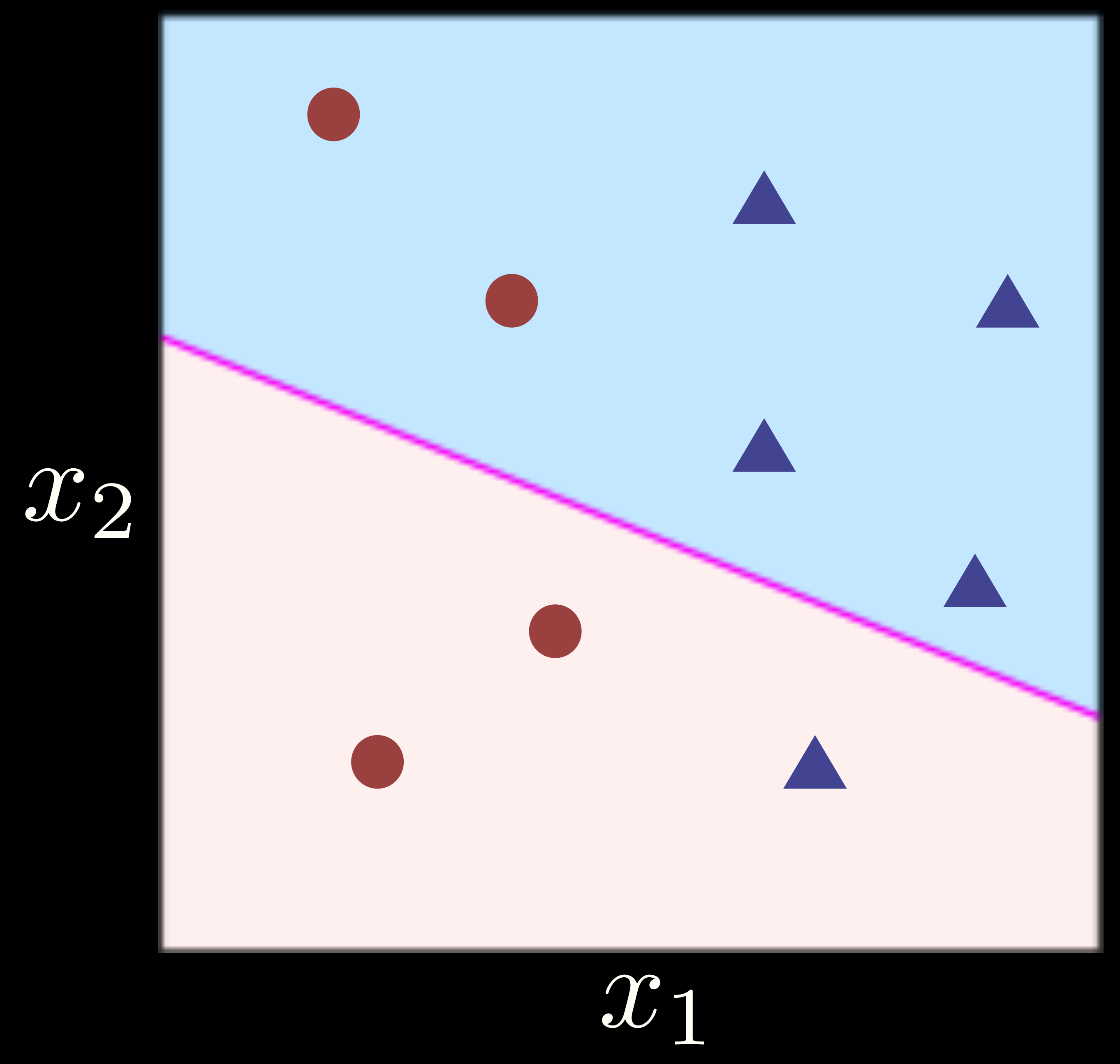
EXAMPLE: A 2D PERCEPTRON

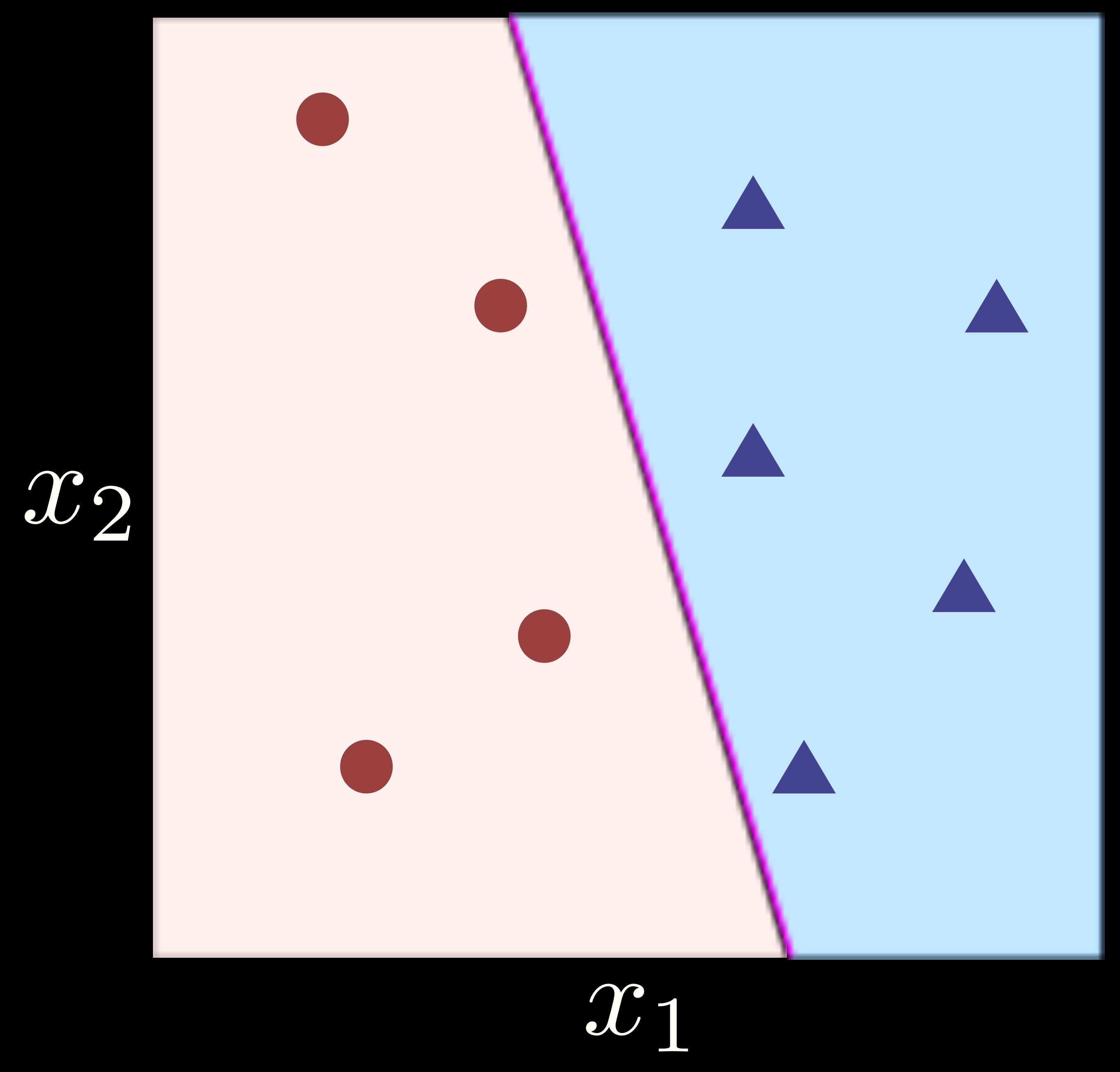
$$w_1 x_1 + w_2 x_2 + threshold = 0$$



EXAMPLE: A 2D PERCEPTRON

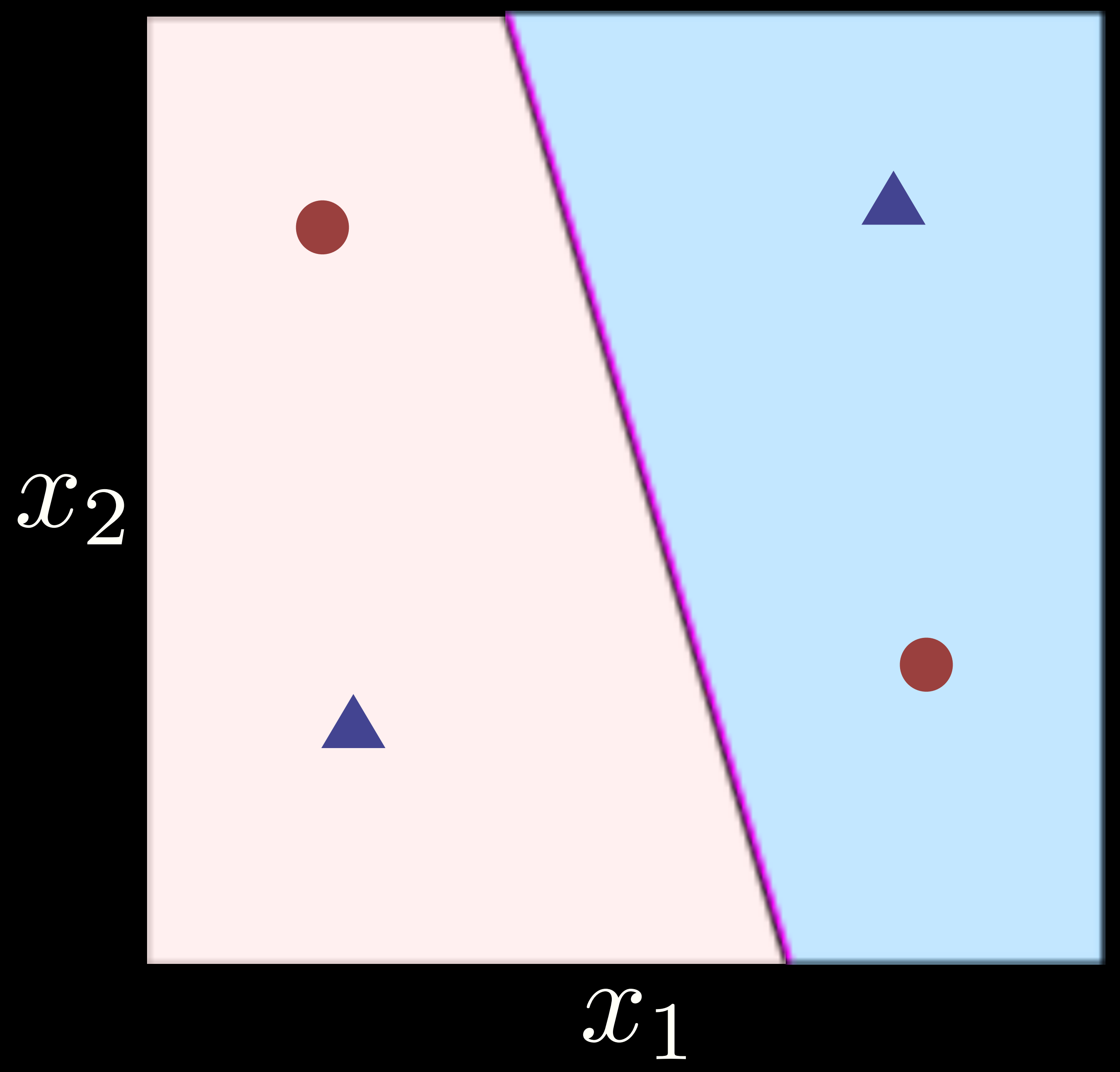
$$w_1 x_1 + w_2 x_2 + threshold = 0$$





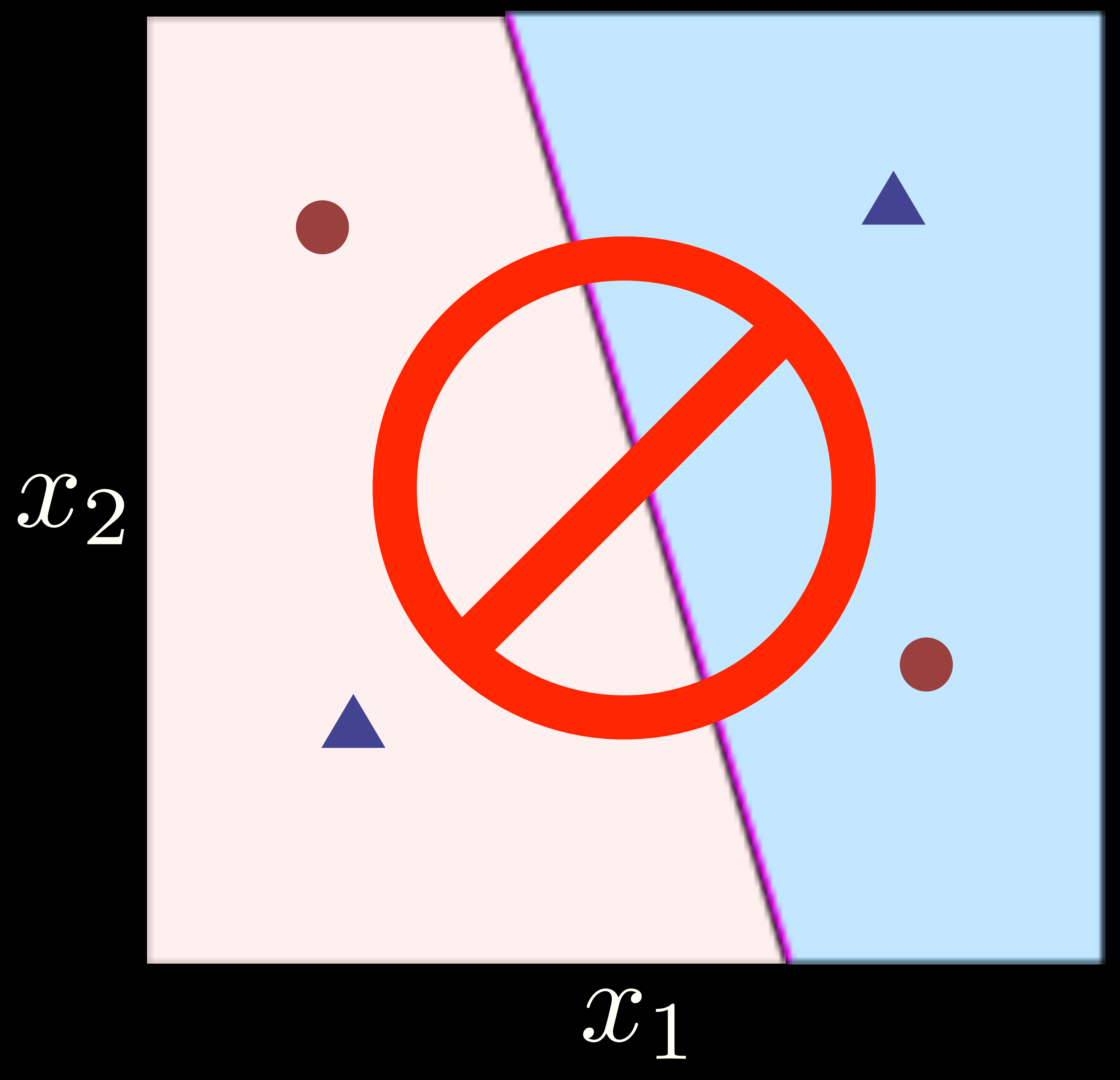
EXAMPLE: A 2D PERCEPTRON

$$w_1 x_1 + w_2 x_2 + threshold = 0$$



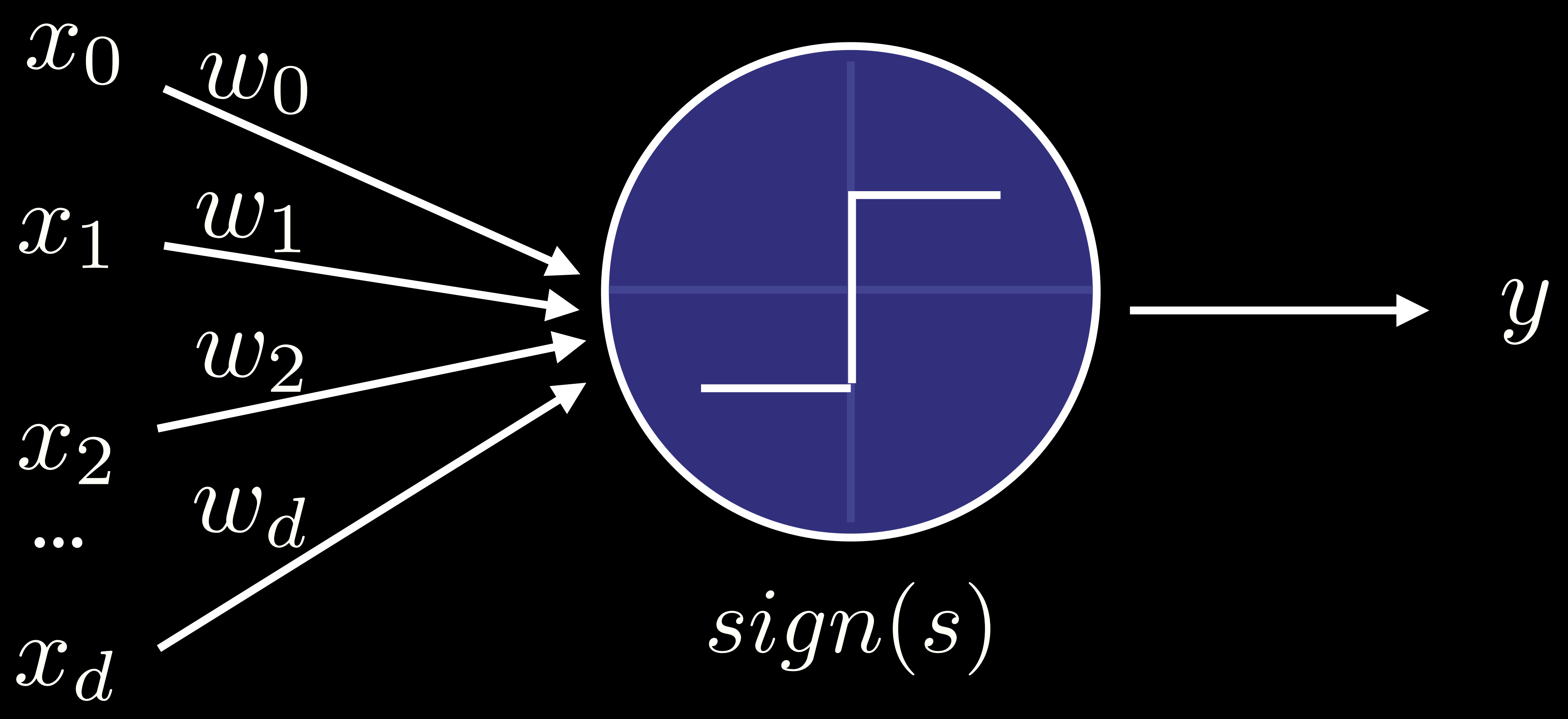
EXAMPLE: A 2D PERCEPTRON

$$w_1 x_1 + w_2 x_2 + threshold = 0$$



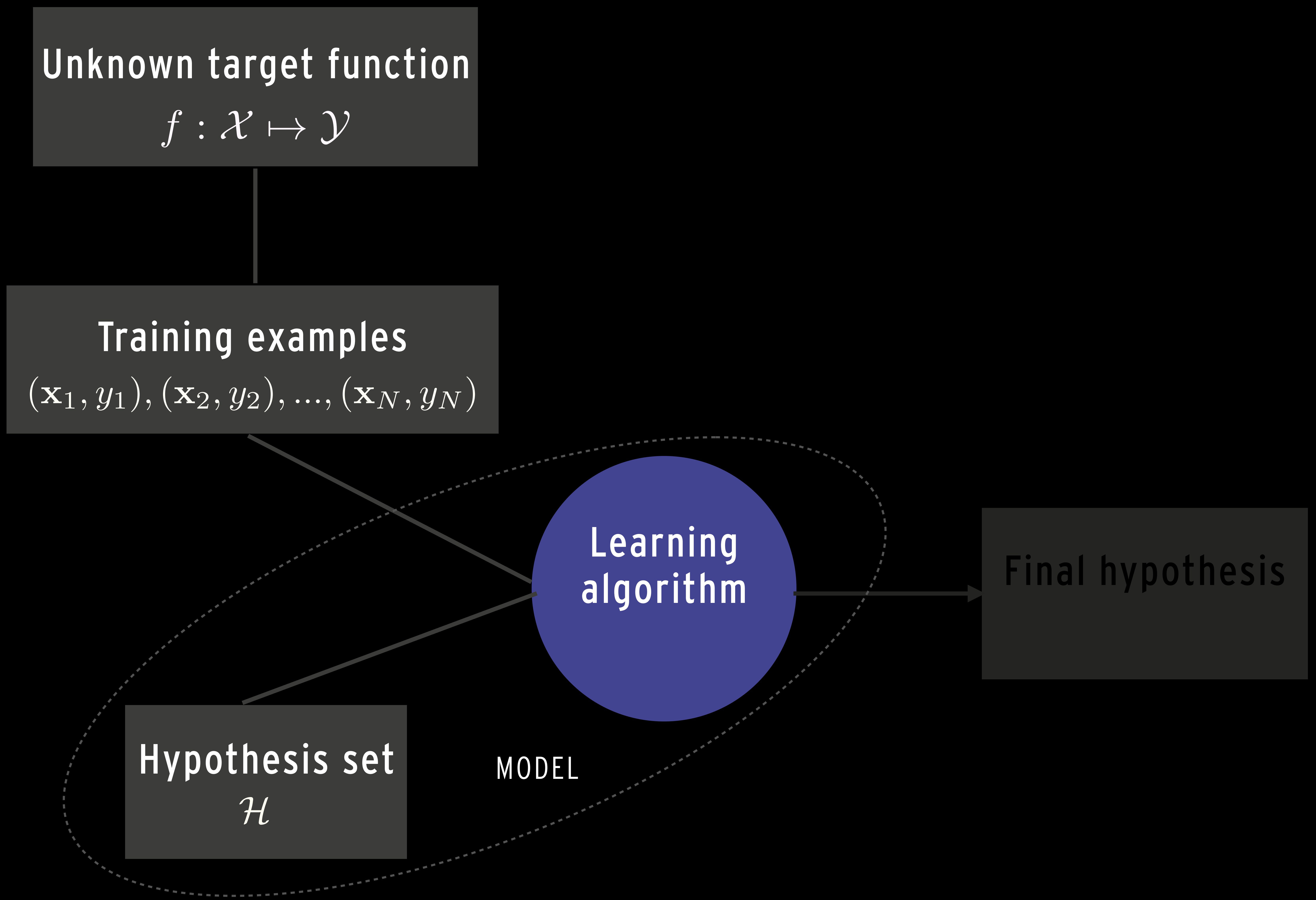
PERCEPTRON

- Weighted input, activation function and output



$$h(\mathbf{x}) = sign\left(\sum_{i=0}^{d} w_i x_i\right)$$

TRAINING EXAMPLES



EXAMPLE: THE PLA ALGORITHM

Perceptron Learning Algorithm

EXAMPLE: THE PLA ALGORITHM

Perceptron Learning Algorithm

1. Pick a specific hypothesis combination of weights, a weight vector $\mathbf{w}(i)$

EXAMPLE: THE PLA ALGORITHM

Perceptron Learning Algorithm

1. Pick a specific hypothesis combination of weights, a weight vector $\mathbf{w}(i)$
2. Take the first test data vector and run it in the perceptron

EXAMPLE: THE PLA ALGORITHM

Perceptron Learning Algorithm

1. Pick a specific hypothesis combination of weights, a weight vector **w**(i)
2. Take the first test data vector and run it in the perceptron
   A. If the perceptrons result is the same as the test data output then take next testdata vector.

EXAMPLE: THE PLA ALGORITHM

Perceptron Learning Algorithm

1. Pick a specific hypothesis combination of weights, a weight vector $\mathbf{w}(i)$
2. Take the first test data vector and run it in the perceptron
   A. If the perceptrons result is the same as the test data output then take next testdata vector.
   B. Else correct the weights according to $\mathbf{w}(i+1) = \mathbf{w}(i) + y(i)\mathbf{x}(i)$

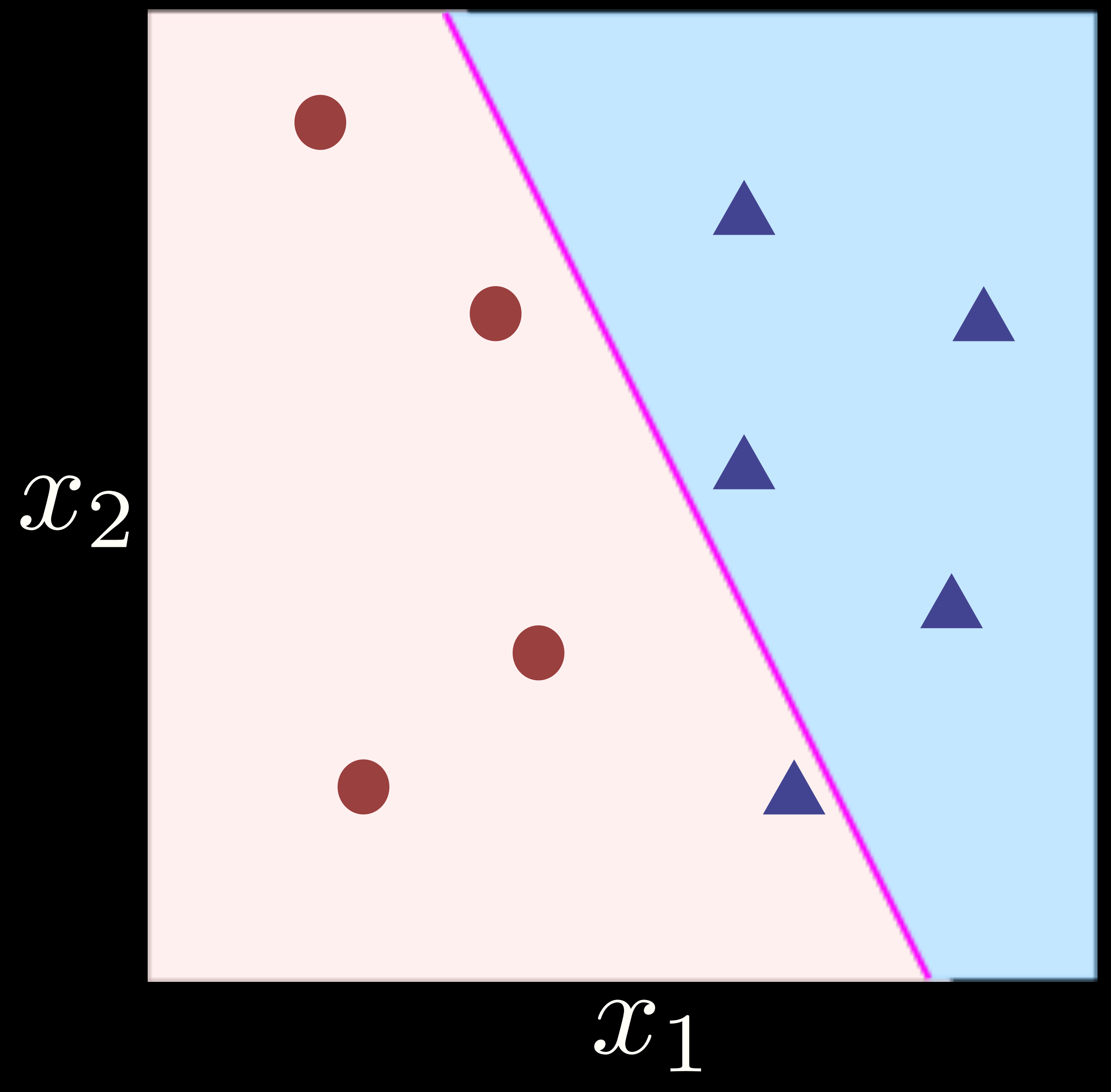
EXAMPLE: THE PLA ALGORITHM

Perceptron Learning Algorithm

1. Pick a specific hypothesis combination of weights, a weight vector **w**(i)
2. Take the first test data vector and run it in the perceptron
   A. If the perceptrons result is the same as the test data output then take next testdata vector.
   B. Else correct the weights according to $\mathbf{w}(i+1) = \mathbf{w}(i) + y(i)\mathbf{x}(i)$
3. Continue with new testdata points until there are no misclassified left.

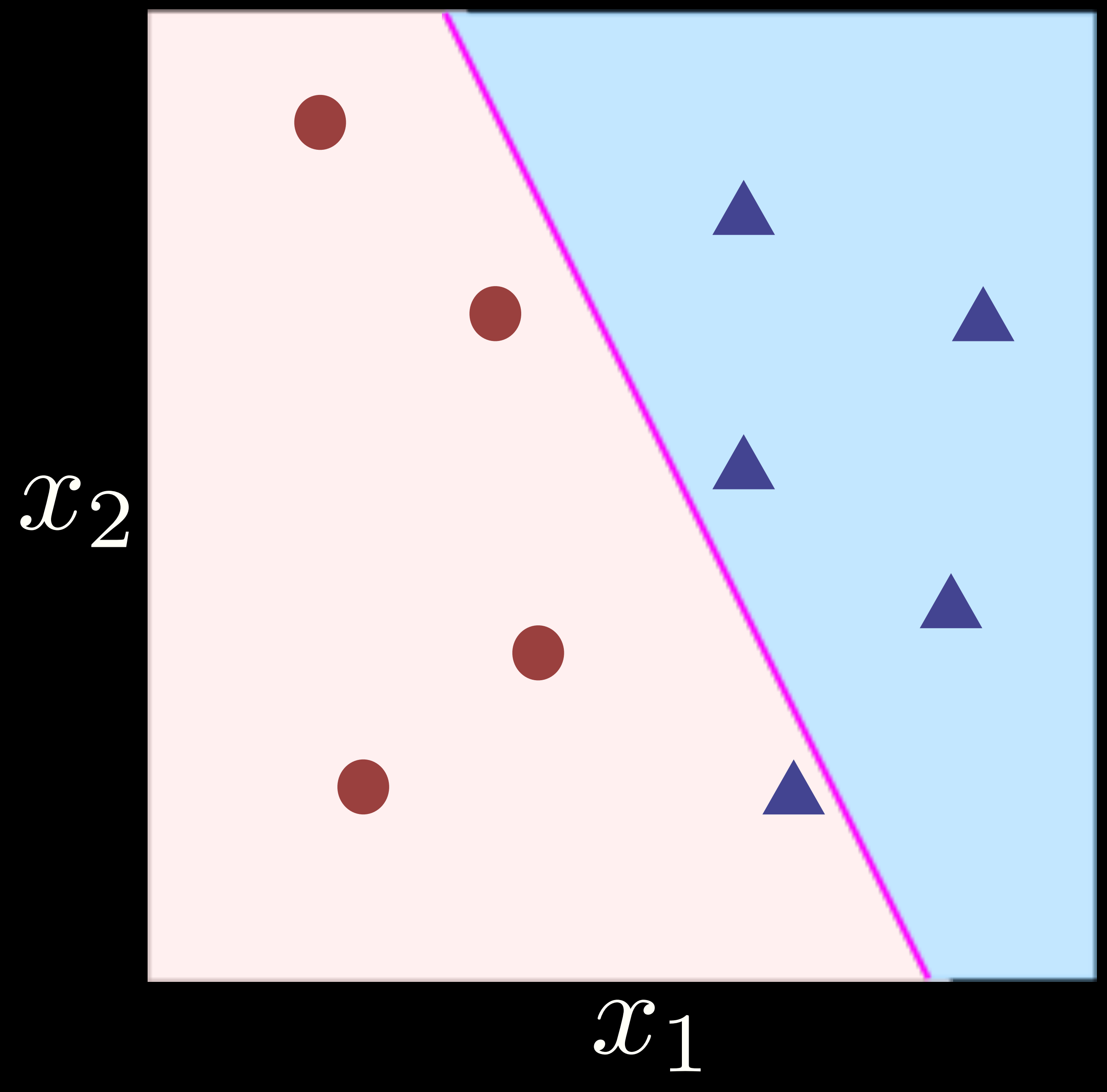
EXAMPLE: THE PLA ALGORITHM ON A 2D PERCEPTRON

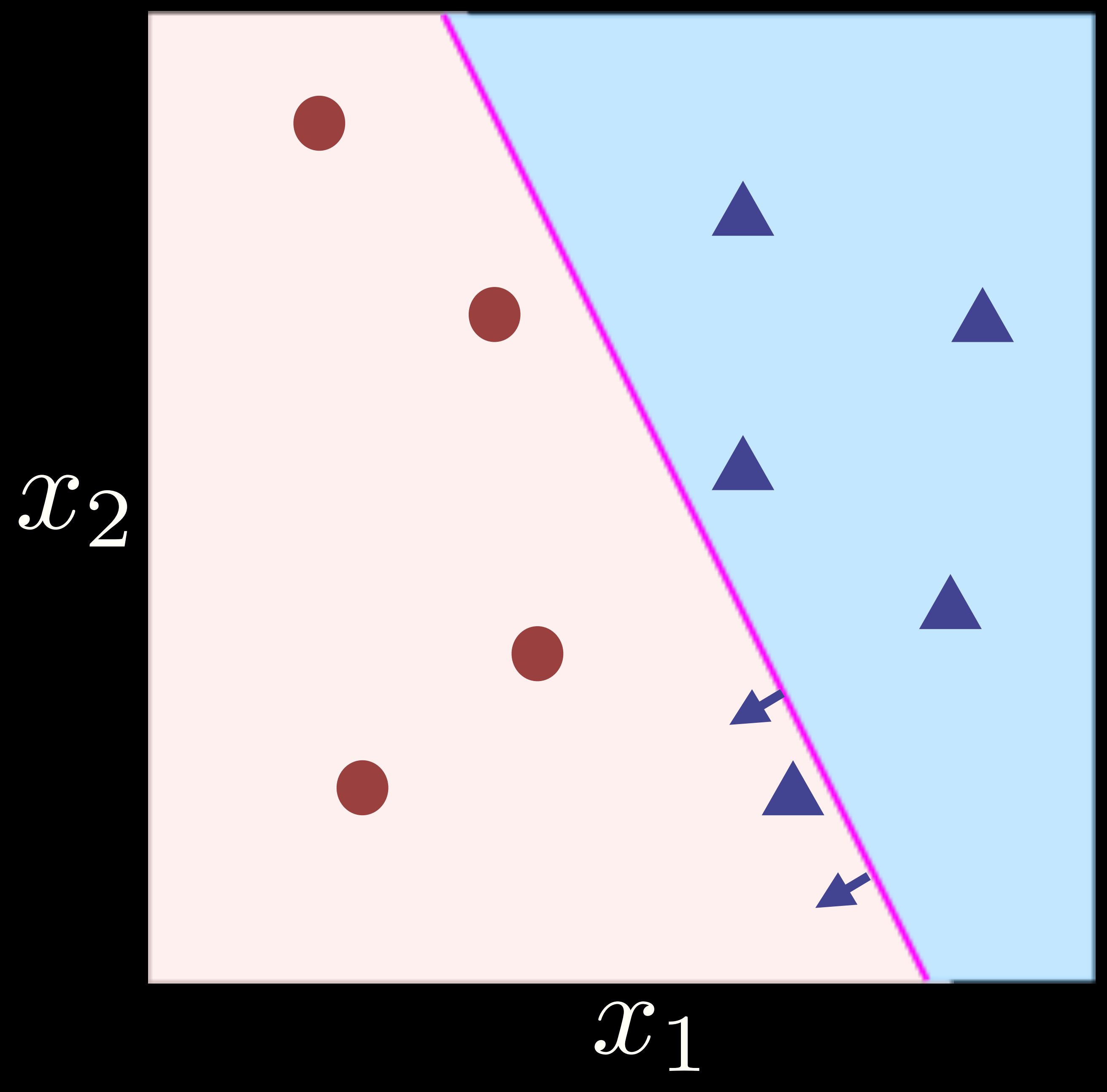
$$w_1 x_1 + w_2 x_2 + threshold = 0$$



EXAMPLE: THE PLA ALGORITHM ON A 2D PERCEPTRON

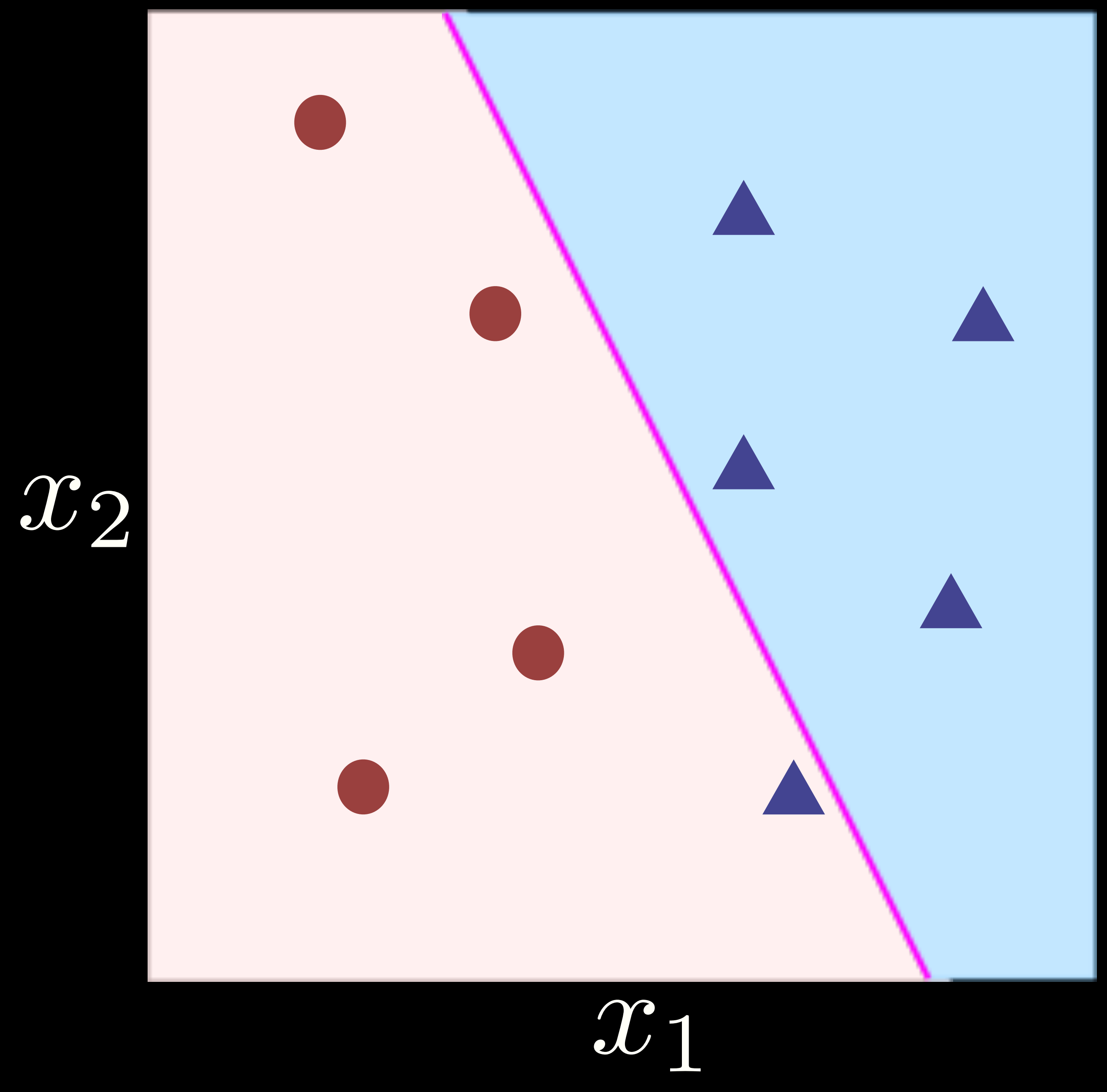
$$w_1 x_1 + w_2 x_2 + threshold = 0$$

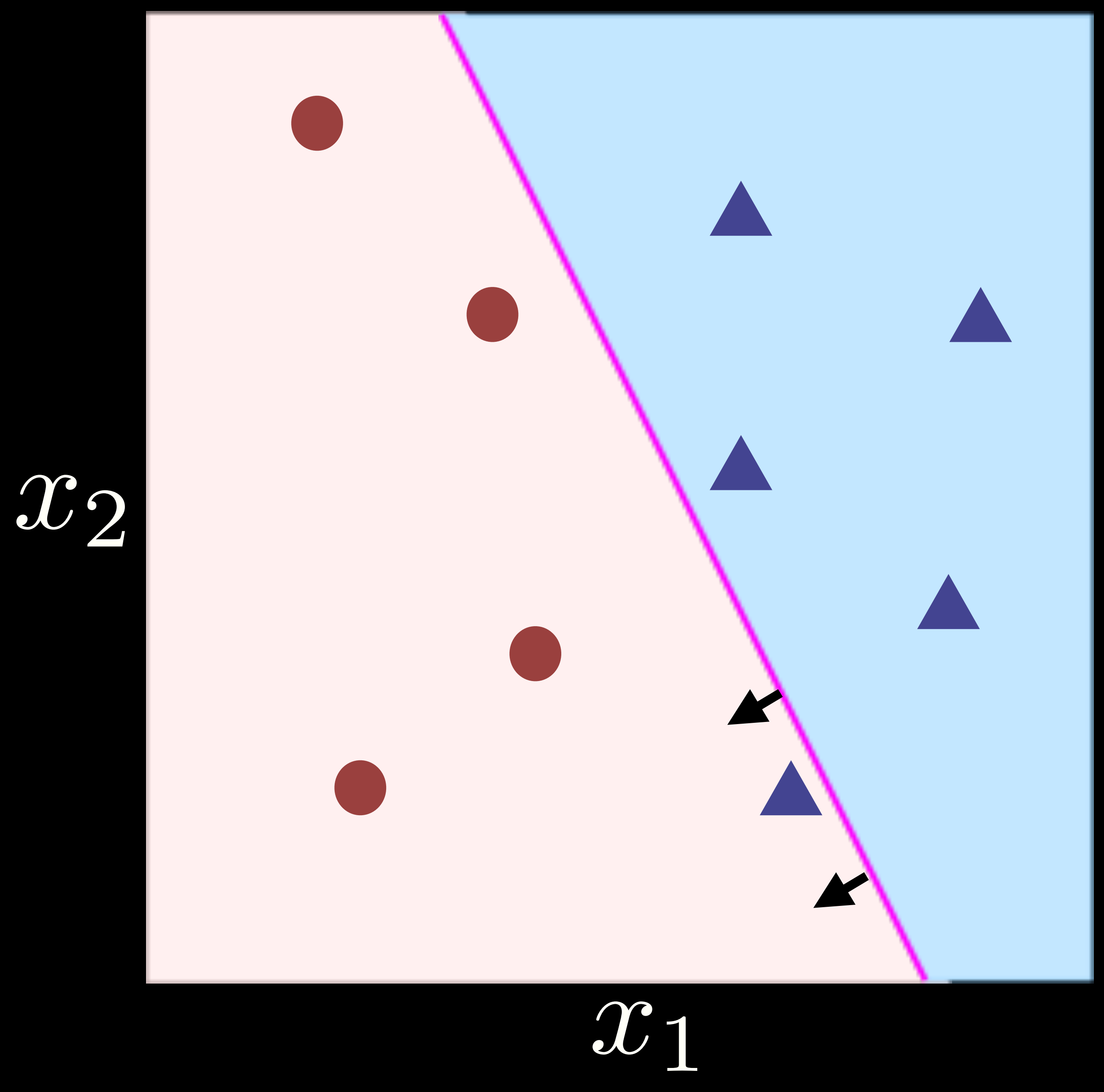


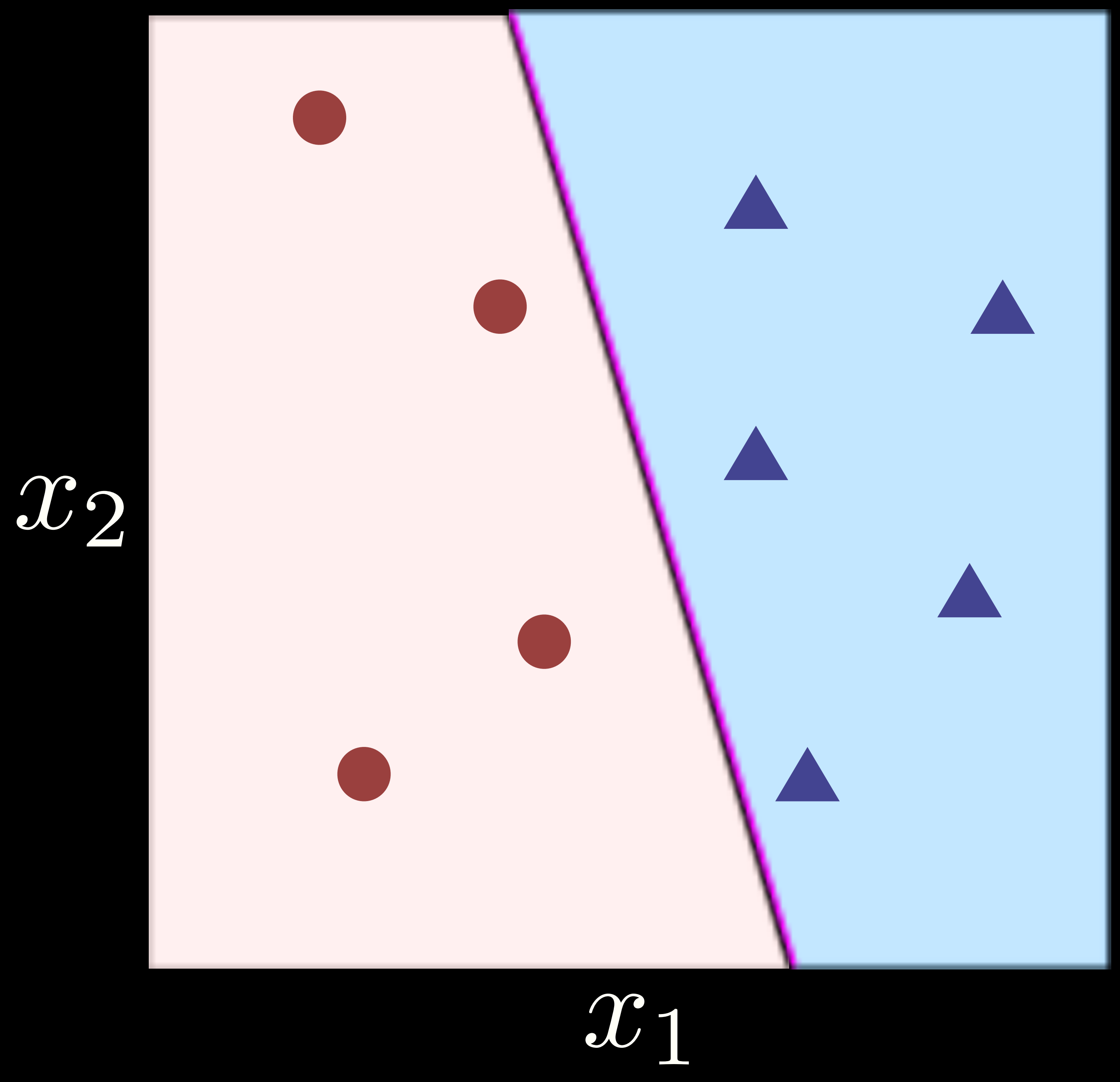


EXAMPLE: THE PLA ALGORITHM ON A 2D PERCEPTRON

$$w_1 x_1 + w_2 x_2 + threshold = 0$$





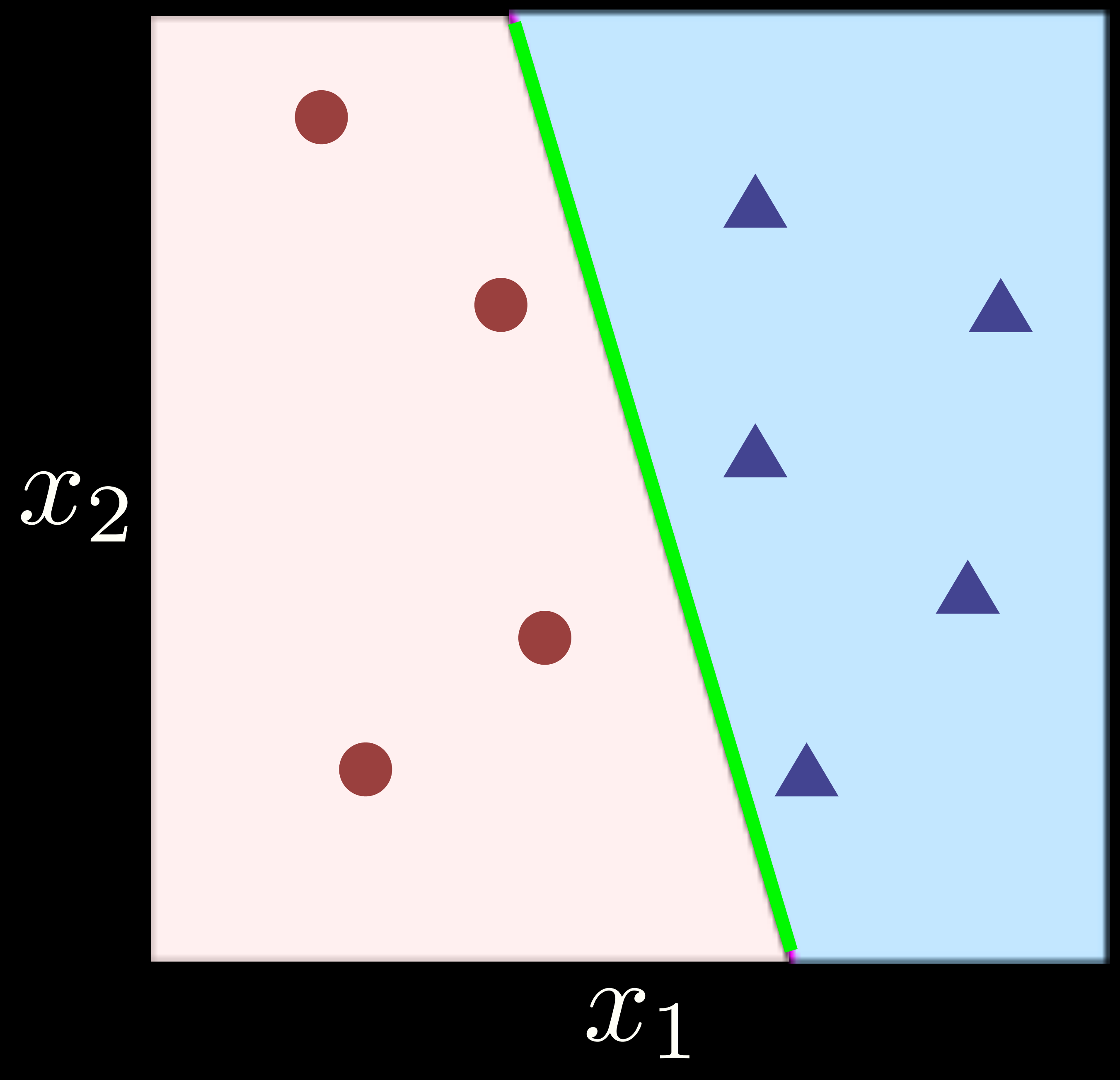


FINAL HYPOTHESIS

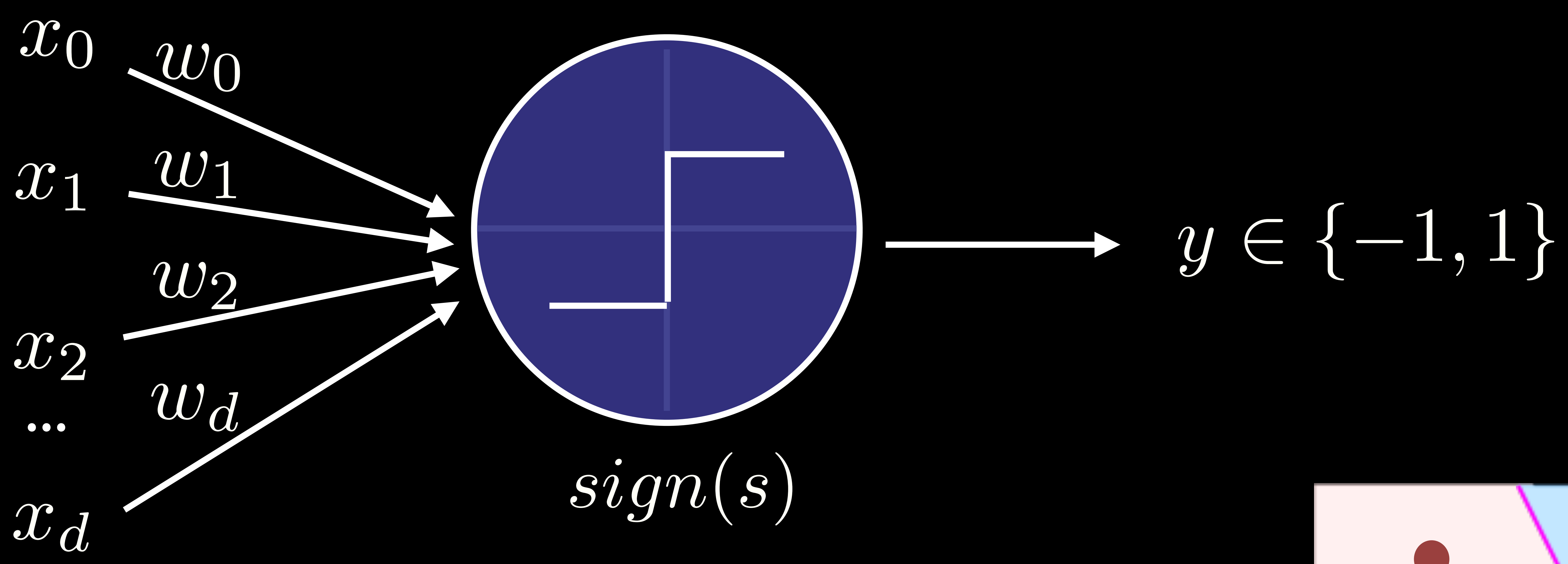
- We have a result:

$$g = sign(w_1 x_1 + w_2 x_2 + threshold)$$

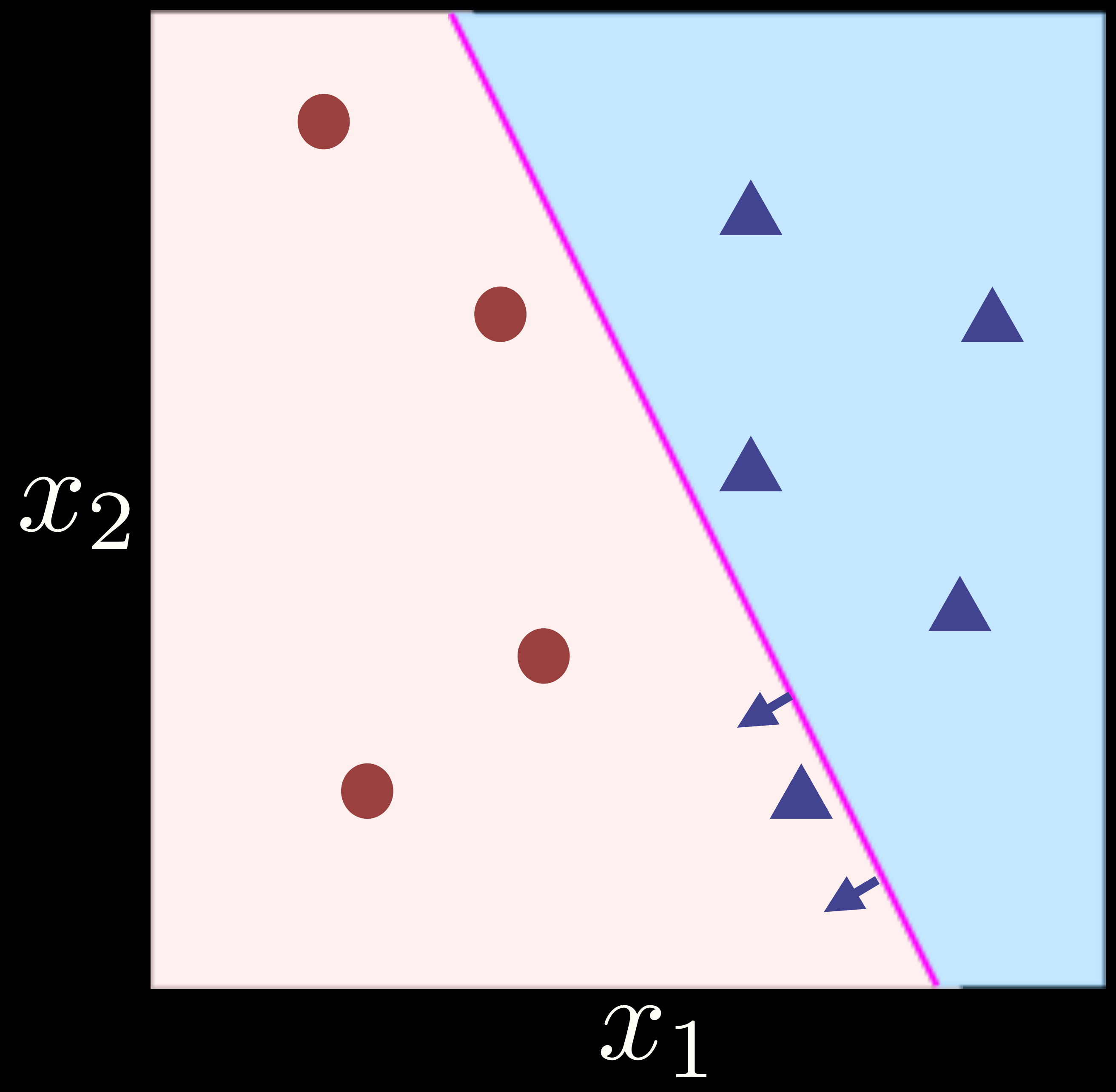
$$g \approx f$$



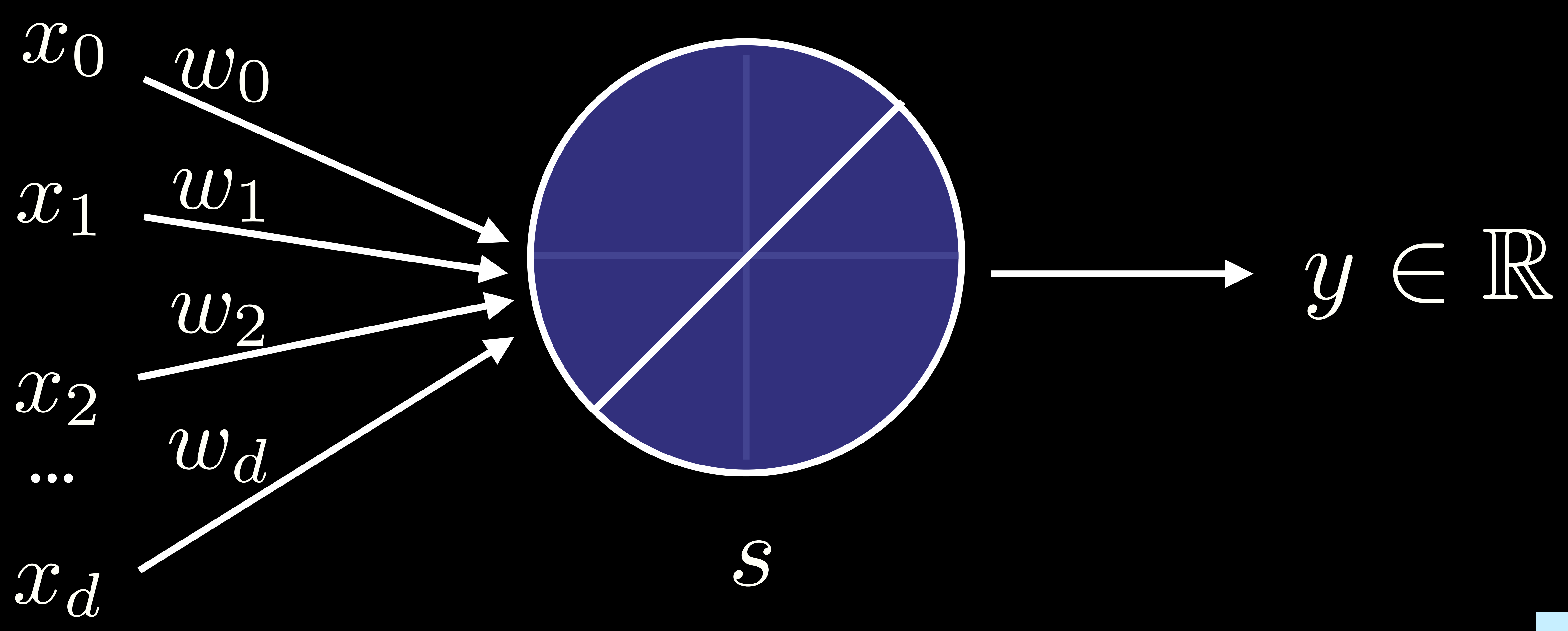
OTHER LINEAR MODELS



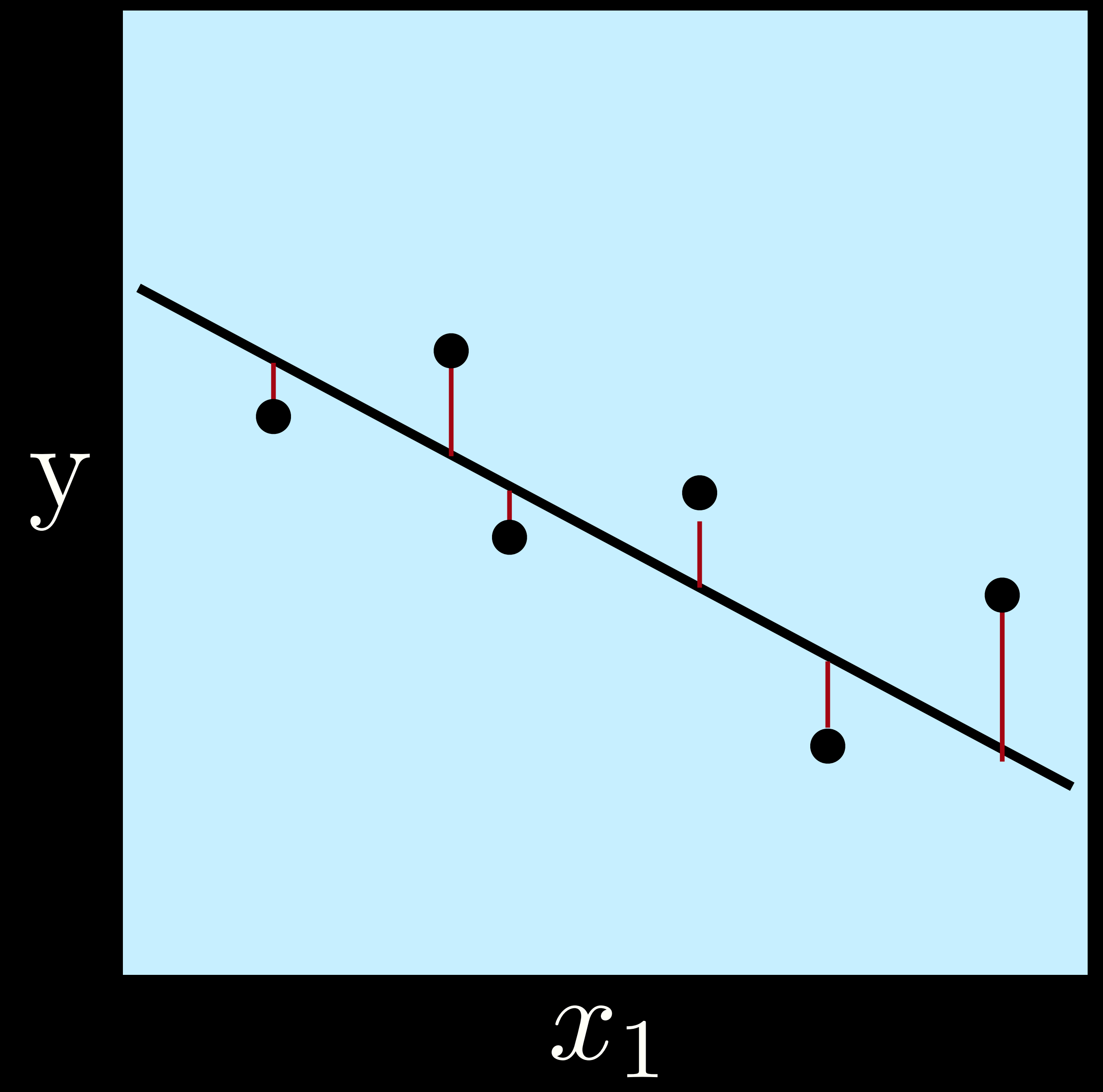
Perceptron



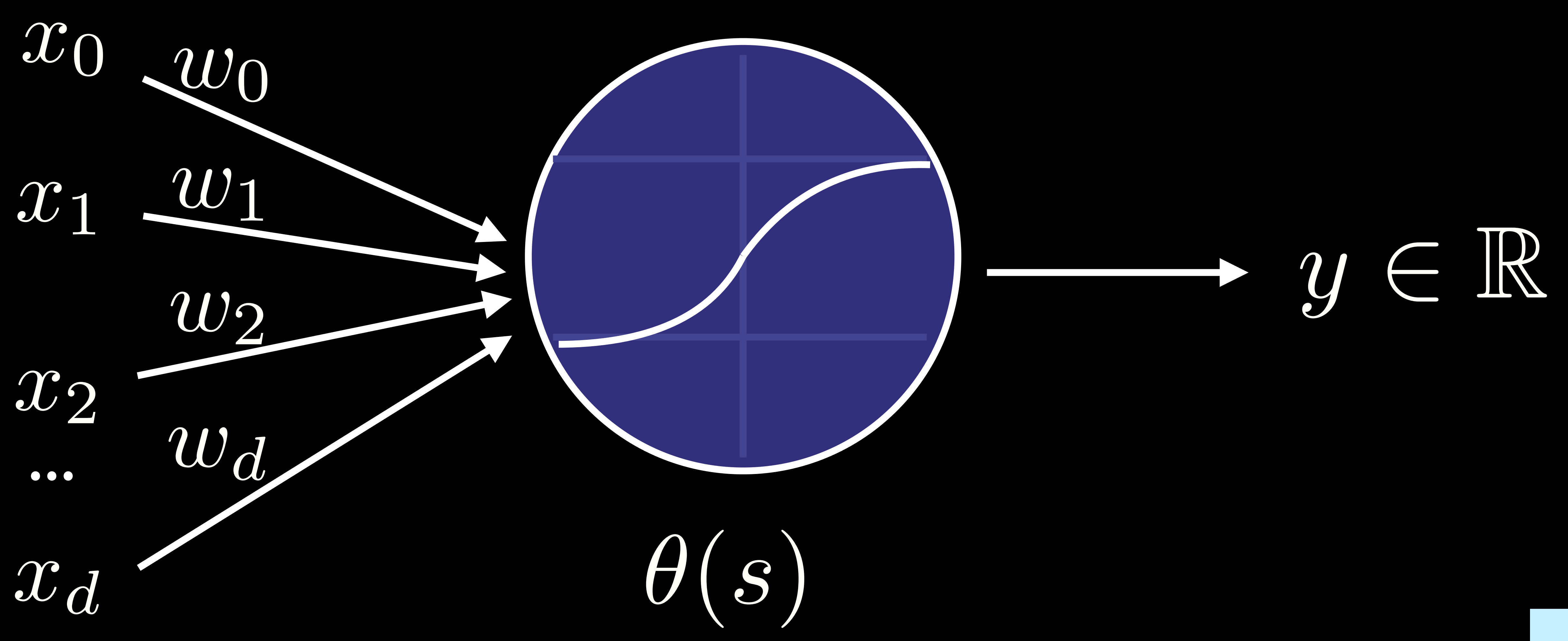
OTHER LINEAR MODELS



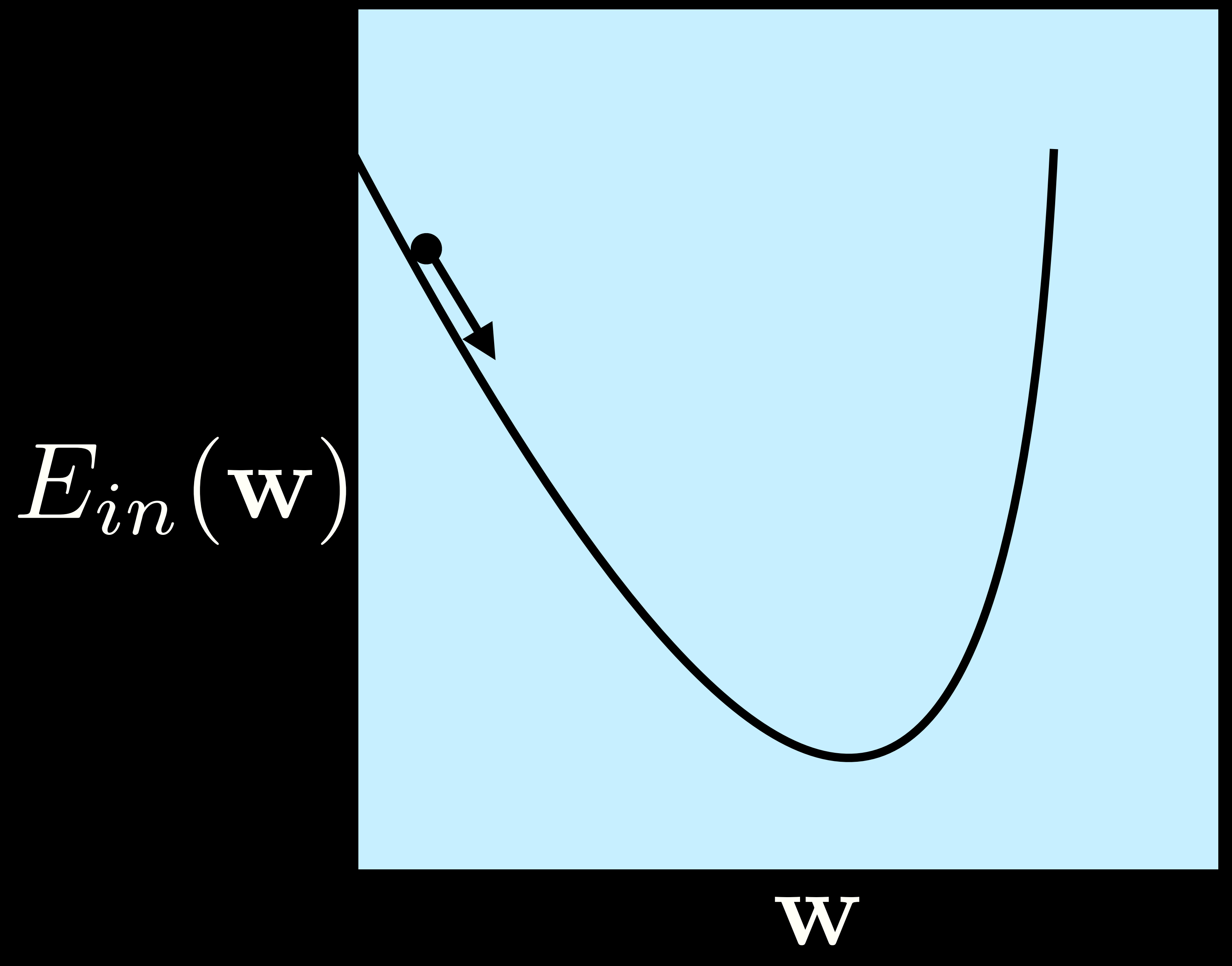
Linear regression



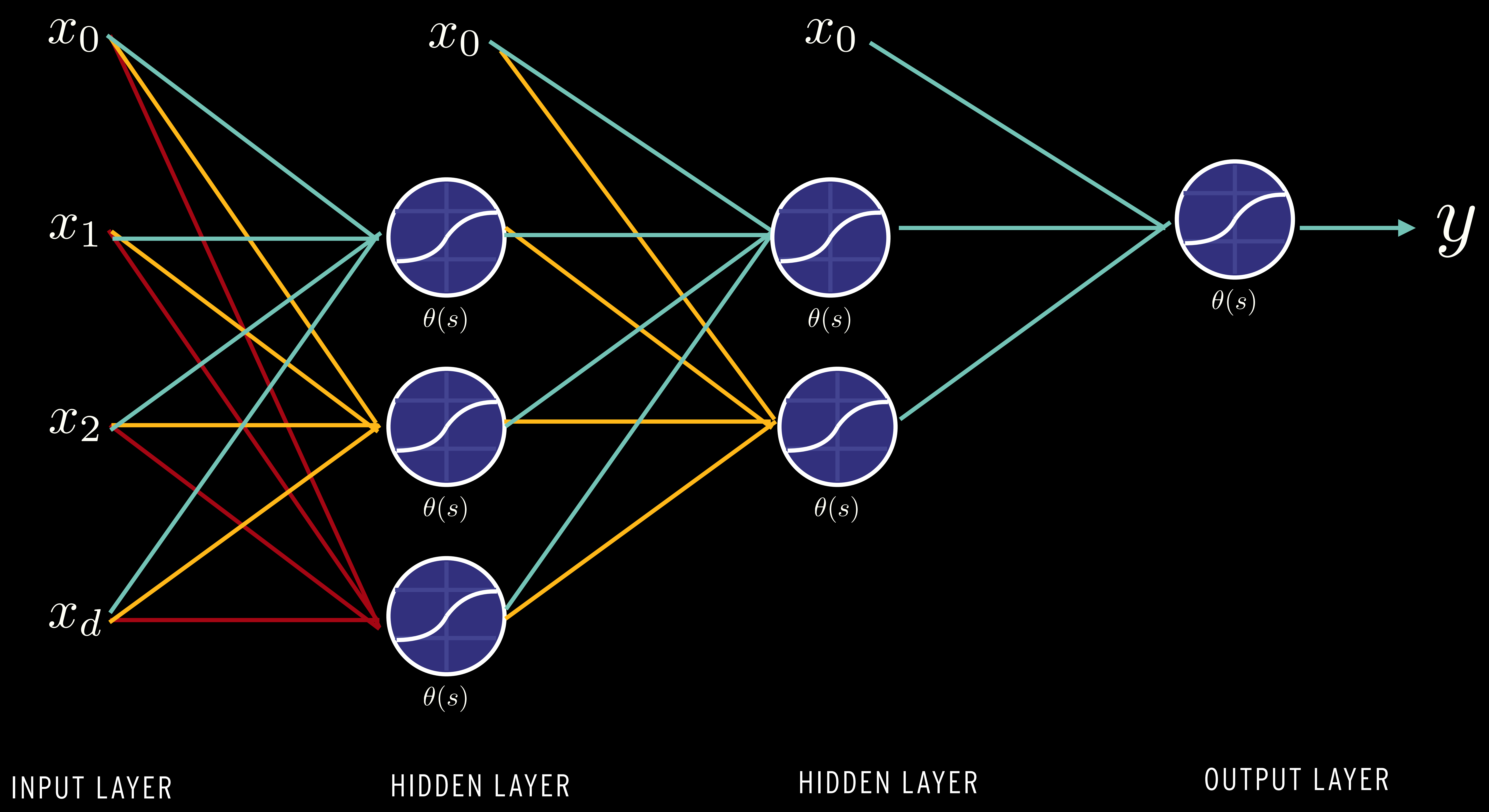
OTHER LINEAR MODELS



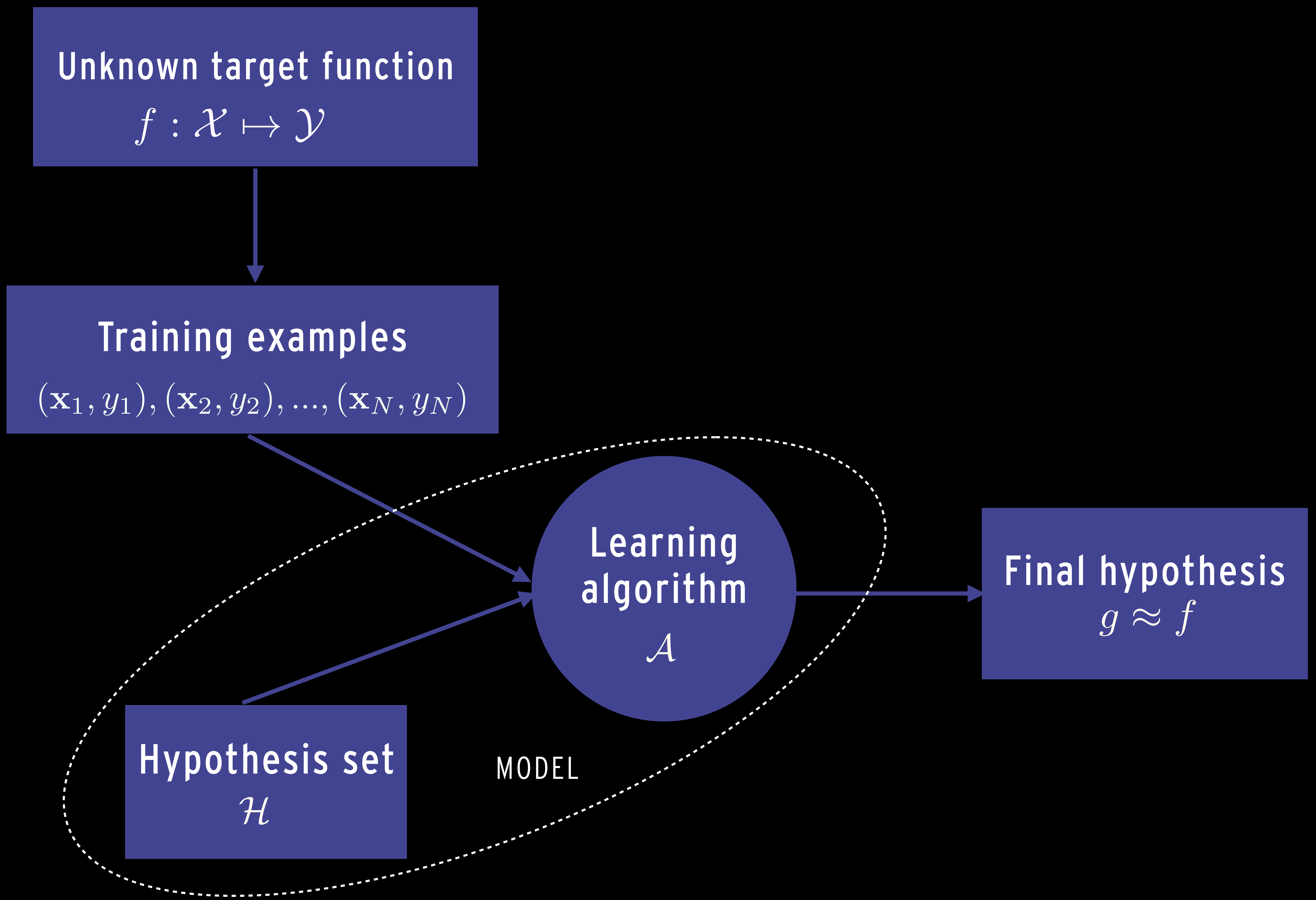
Logistic regression



NEUARAL NETWORKS



THE LEARNING PROBLEM



WE HAVE A RESULT!

$$g \approx f$$

How do we know that it works outside of the training data?

AGENDA

- Introduction and context
- The work process
- The learning problem
- **Validation and overfitting**
- Tools
- Risks and ethics
- Demo

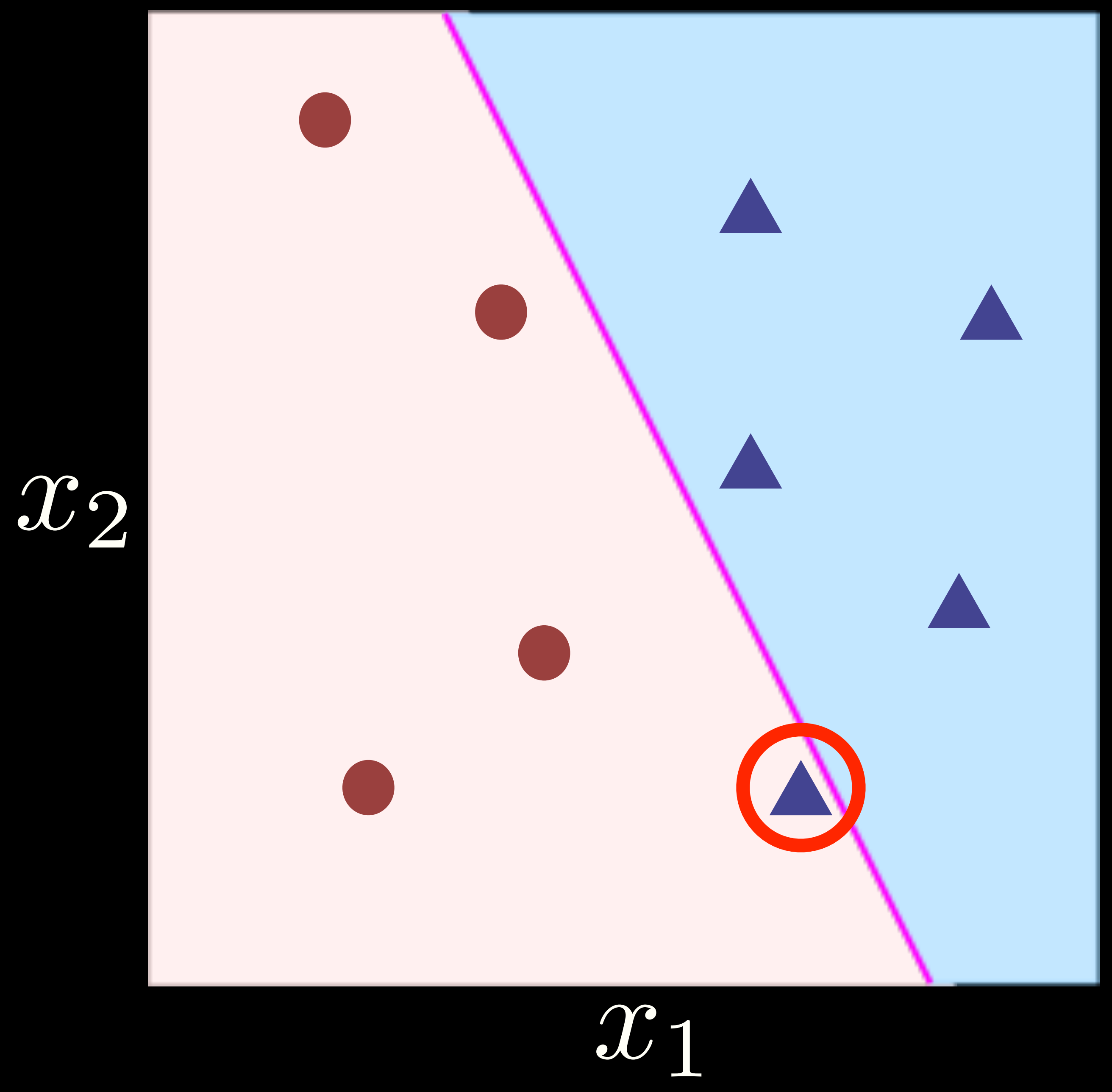
HOW DO WE VALIDATE THE RESULT?

- Error
- Validation
- Noise
- Overfitting

IN-SAMPLE ERROR

E_{in} (in-sample error), how unsuccessful one hypothesis is on the training data set.

The fraction of misclassified points in the training data set.



$$E_{in} = \frac{1}{N} \sum_{n=1}^{N} [\![h(x_n) \neq f(x_n)]\!]$$

OUT-OF-SAMPLE ERROR

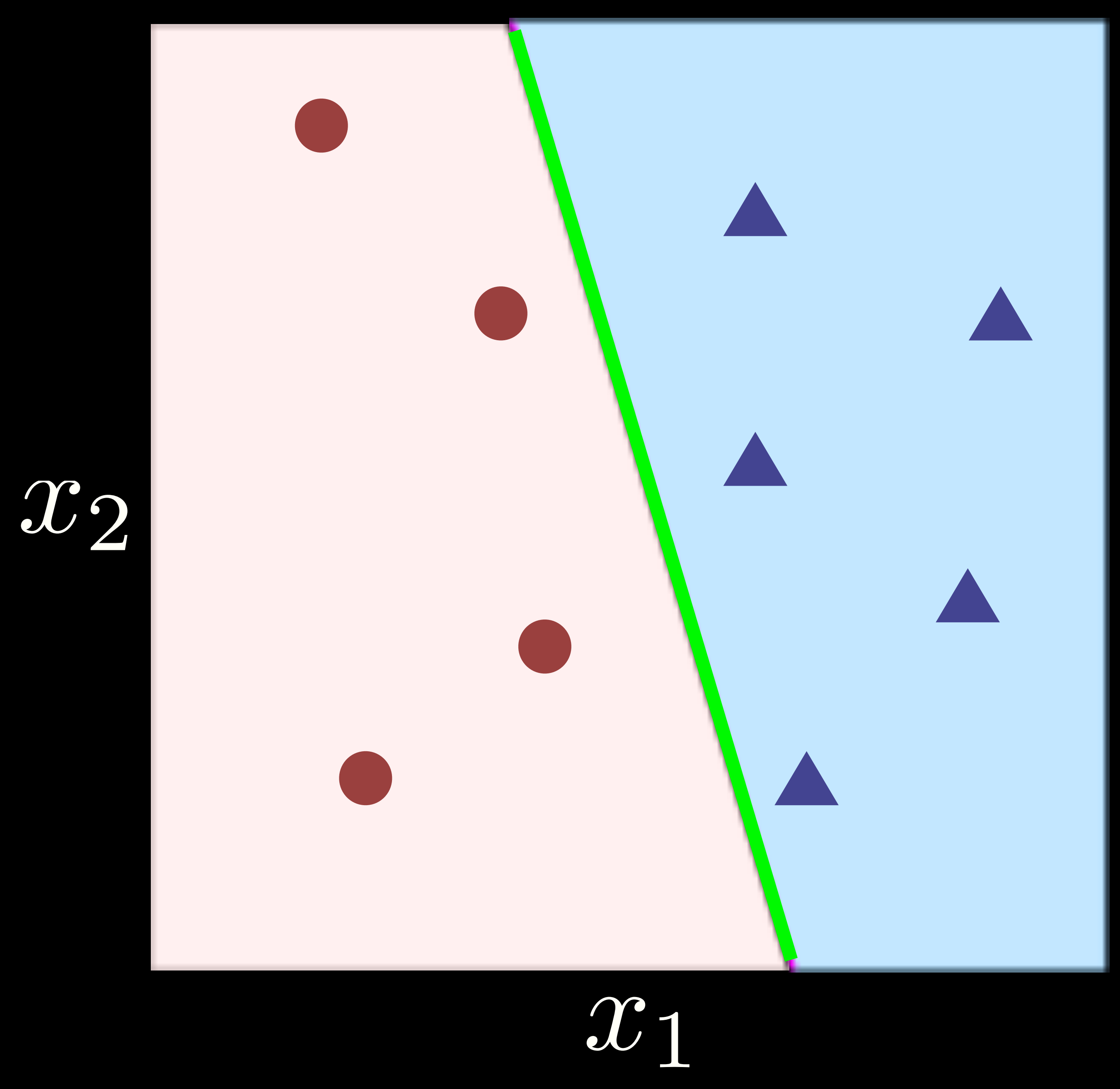
E_{out} imperfectness of the final hypothesis outside of the training data

$E_{out} = f(x_{out}) - g(x_{out})$ which is unknown

OUT-OF-SAMPLE ERROR

E_{out} imperfectness of the final hypothesis outside of the training data

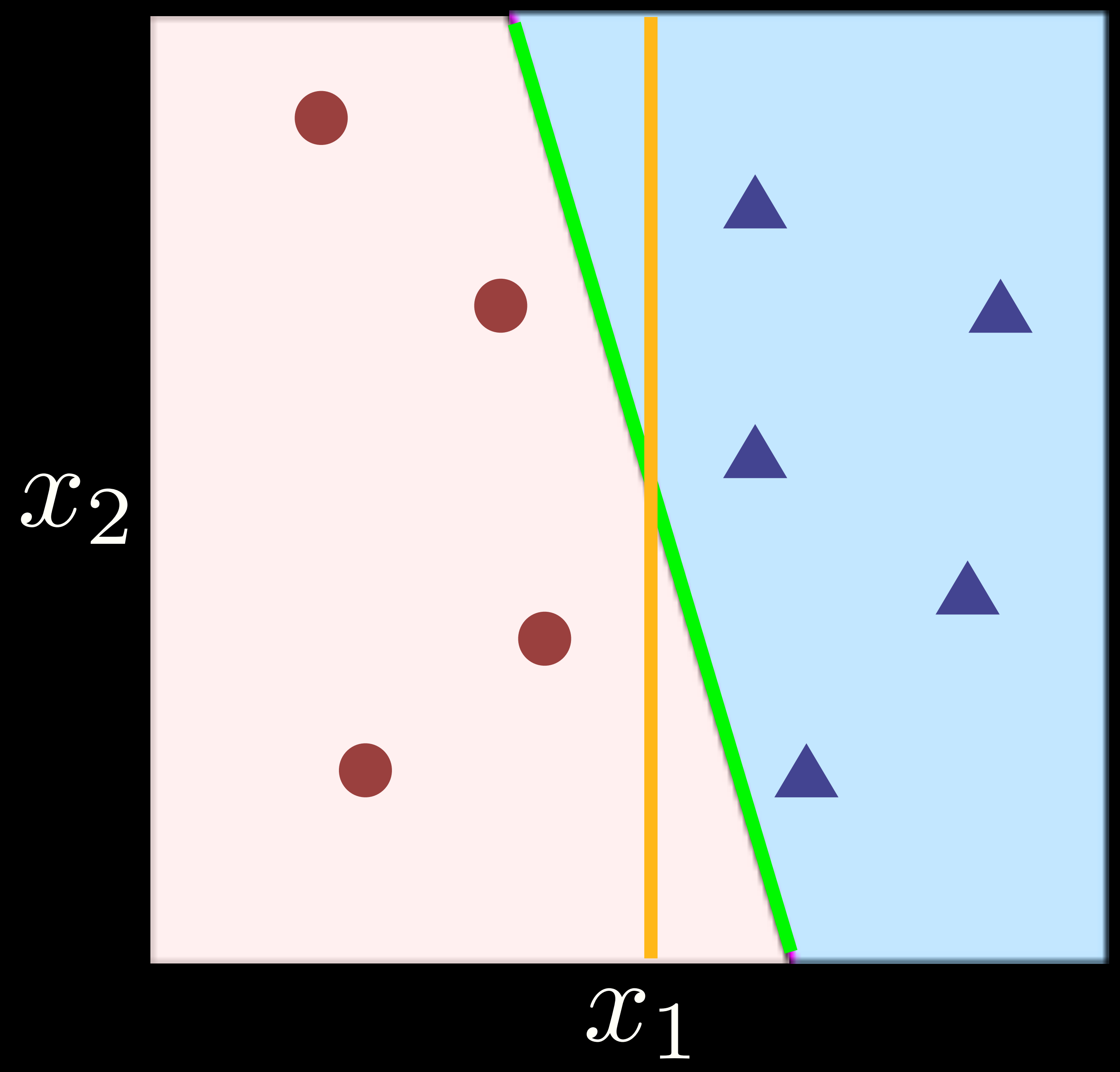
$E_{out} = f(x_{out}) - g(x_{out})$ which is unknown



OUT-OF-SAMPLE ERROR

E_{out} imperfectness of the final hypothesis outside of the training data

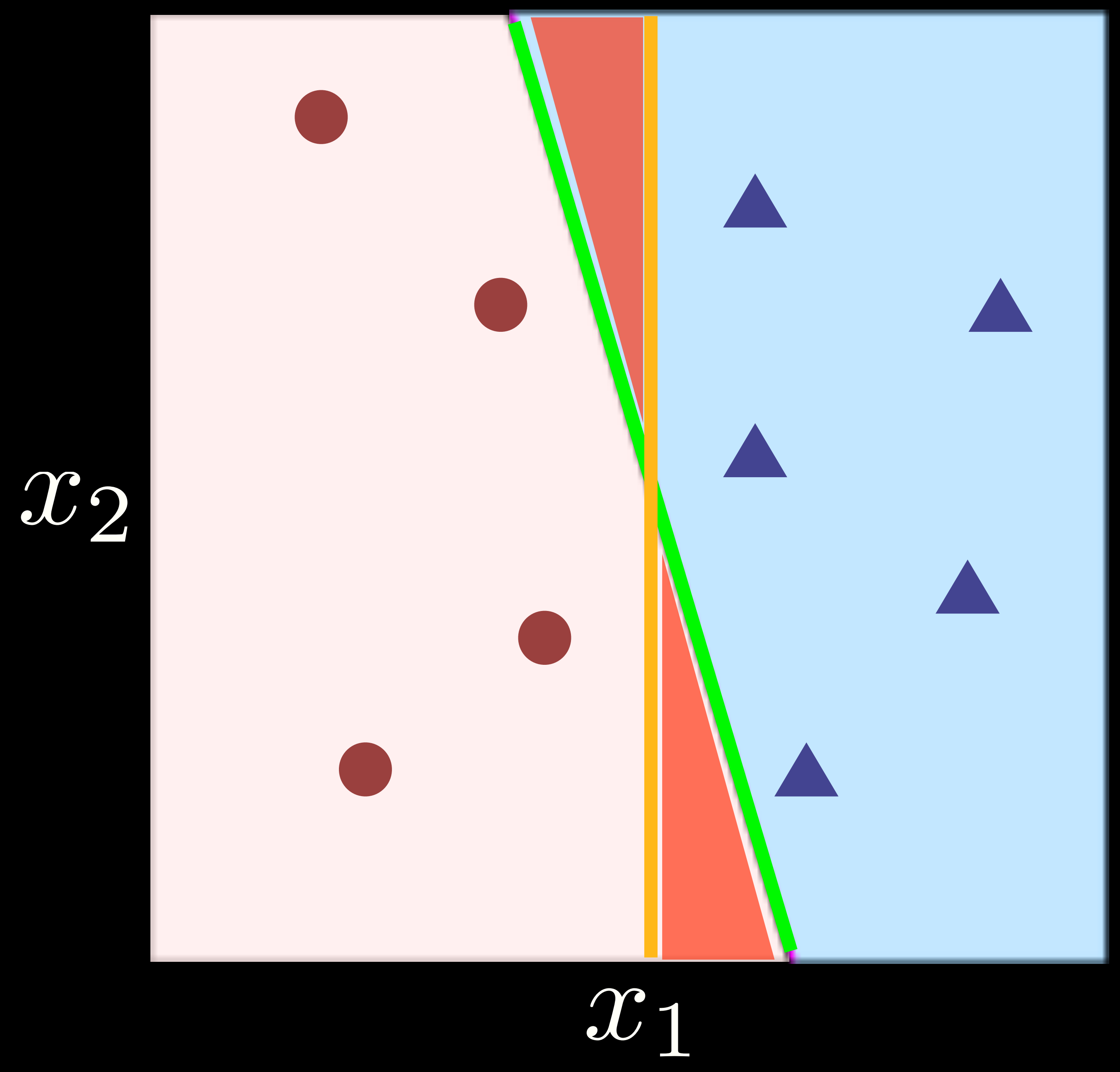
$E_{out} = f(x_{out}) - g(x_{out})$ which is unknown



OUT-OF-SAMPLE ERROR

E_{out} imperfectness of the final hypothesis outside of the training data

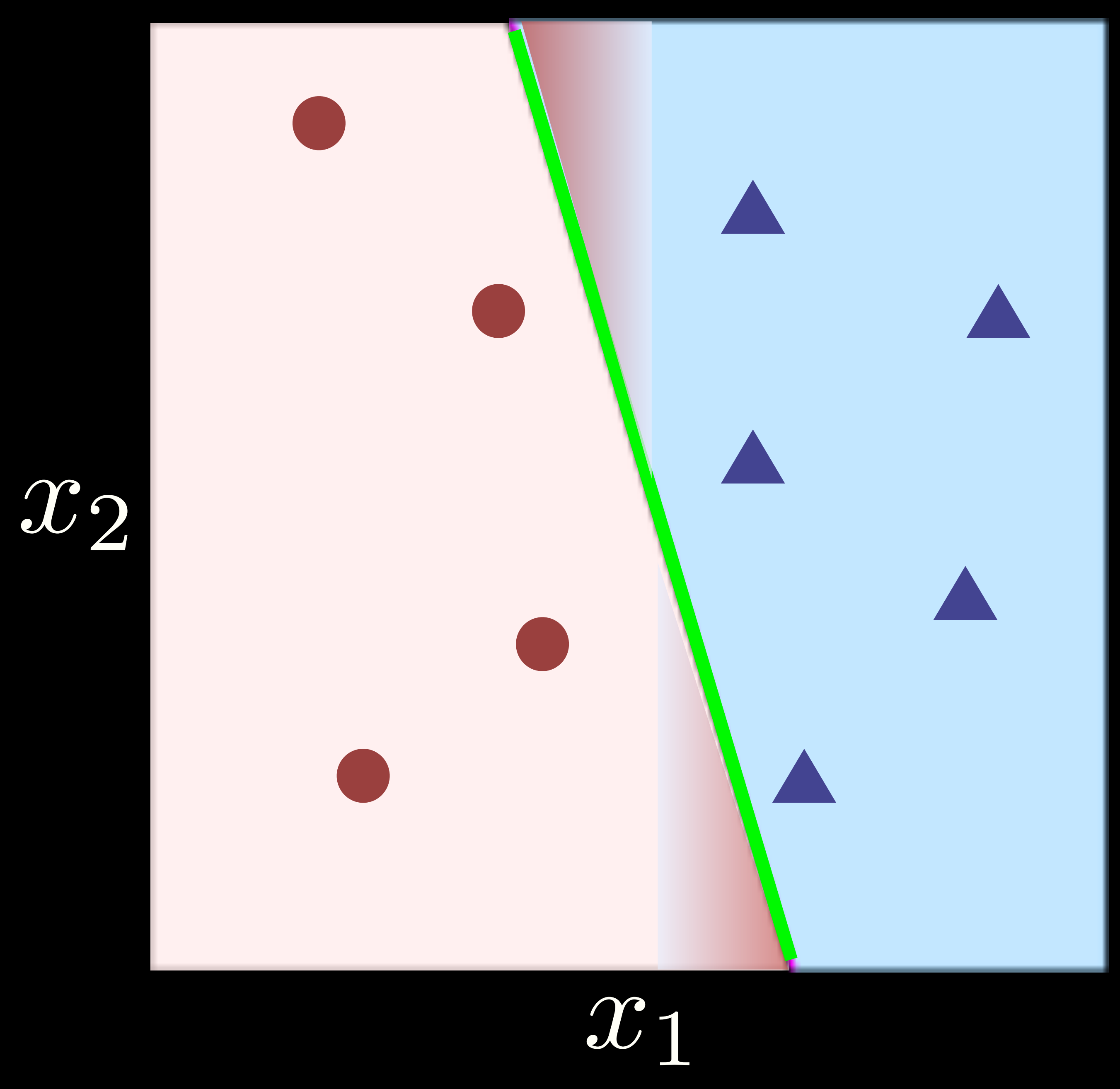
$E_{out} = f(x_{out}) - g(x_{out})$ which is unknown



OUT-OF-SAMPLE ERROR

E_{out} imperfectness of the final hypothesis outside of the training data

$E_{out} = f(x_{out}) - g(x_{out})$ which is unknown



SO WHAT SHOULD WE USE?

Virtual Reality!

SO WHAT SHOULD WE USE?

Virtual Reality!

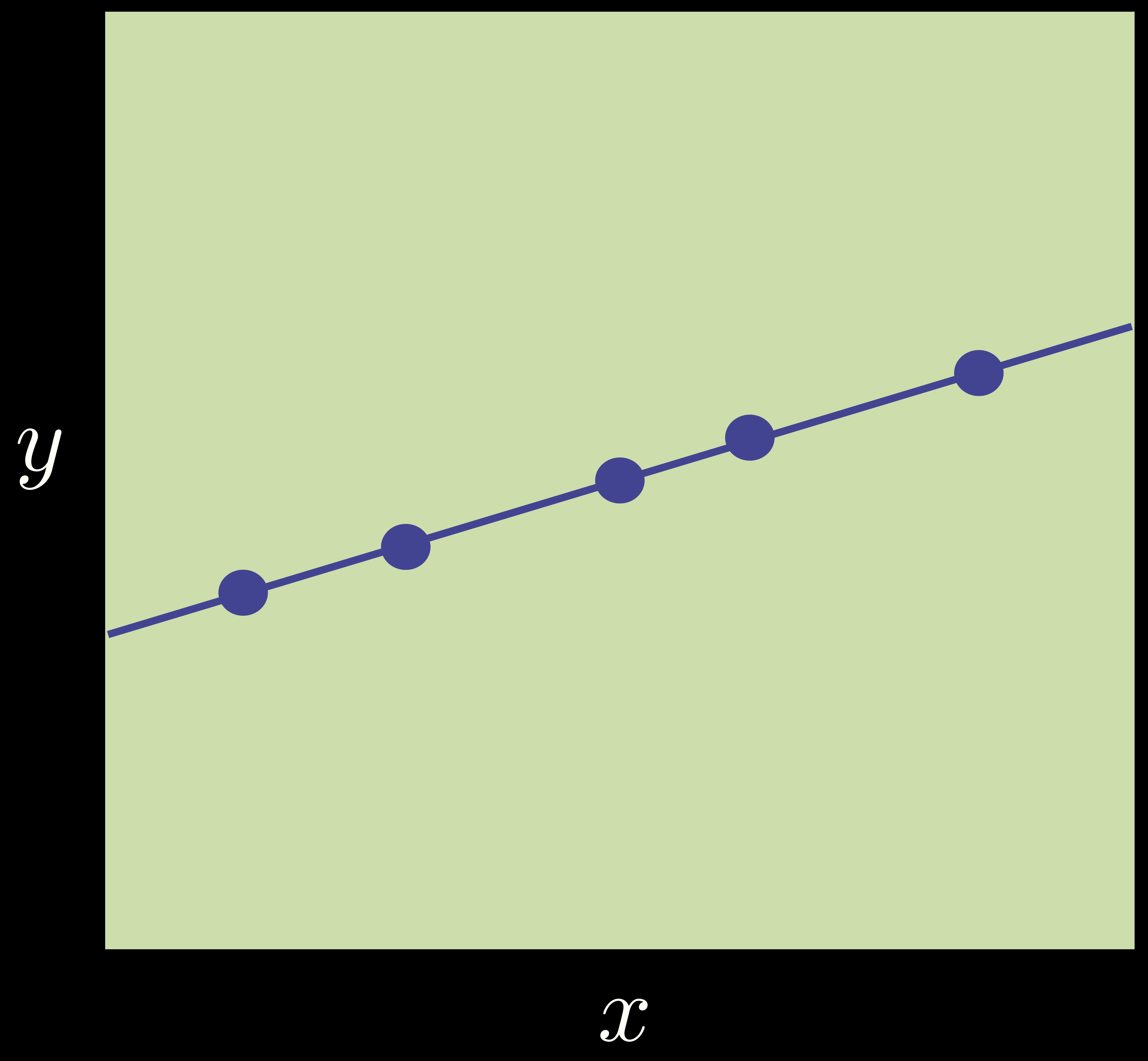
TESTING AND VALIDATION

- Testing
  - Pure unbiased testing
- Cross Validation
  - Not unbiased
  - More efficient method, you can use all data for both training and validation

NOISE

- The world is an ugly place …
- The target function is maybe not a function but a probability distribution because of noise.

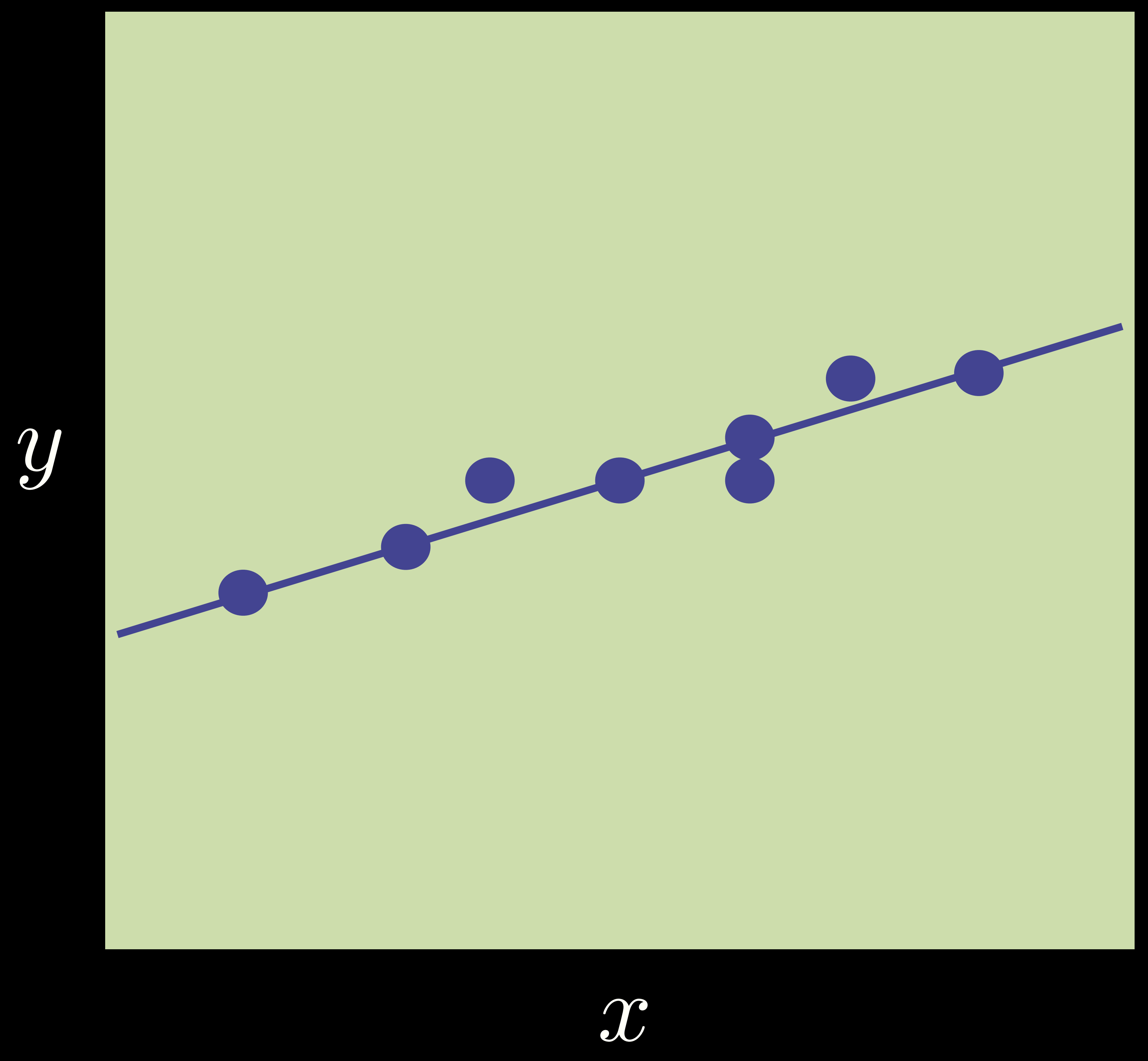
$$P(y|\mathbf{x}) = f + noise$$



NOISE

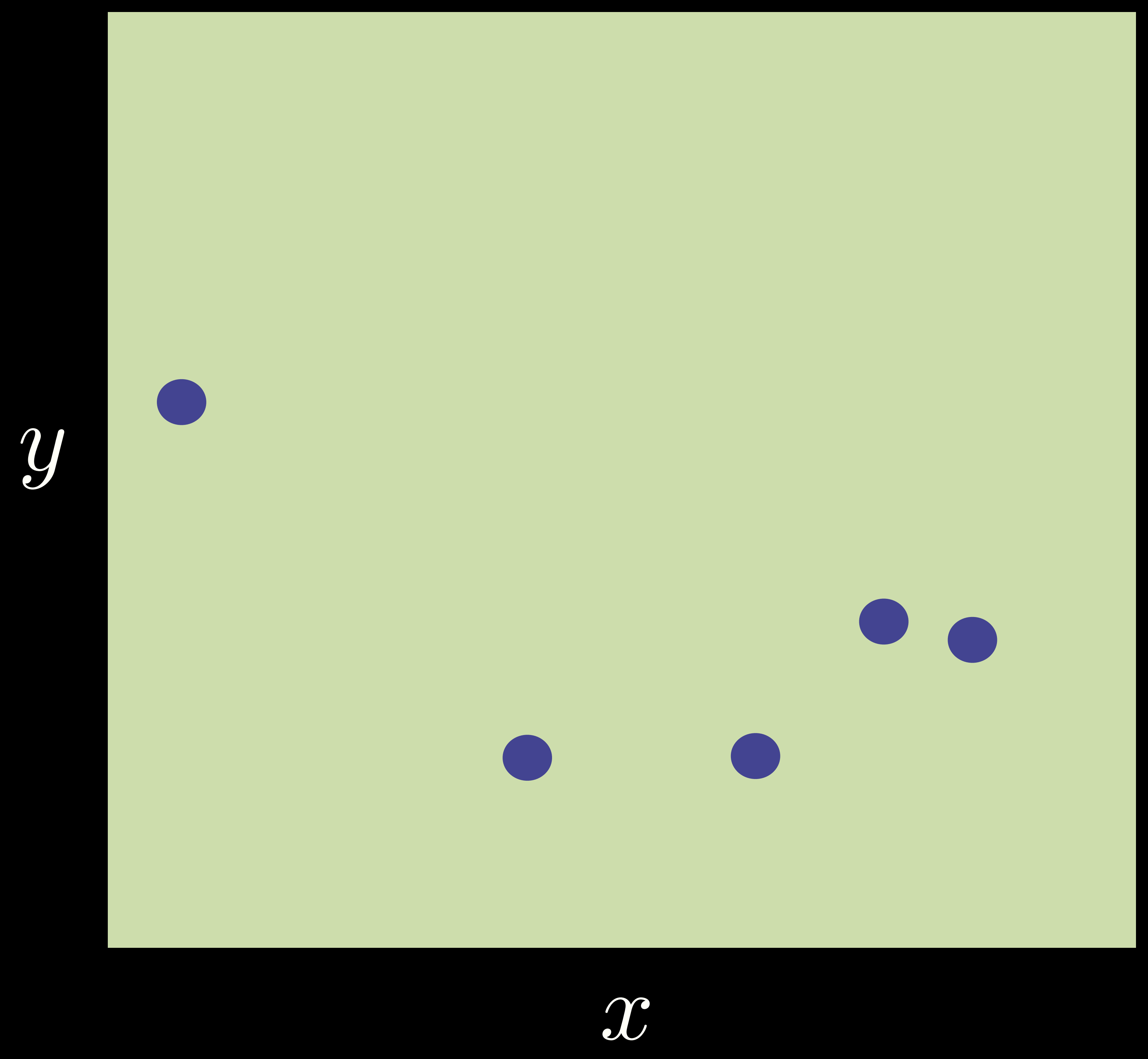
- The world is an ugly place …
- The target function is maybe not a function but a probability distribution because of noise.

$$P(y|\mathbf{x}) = f + noise$$



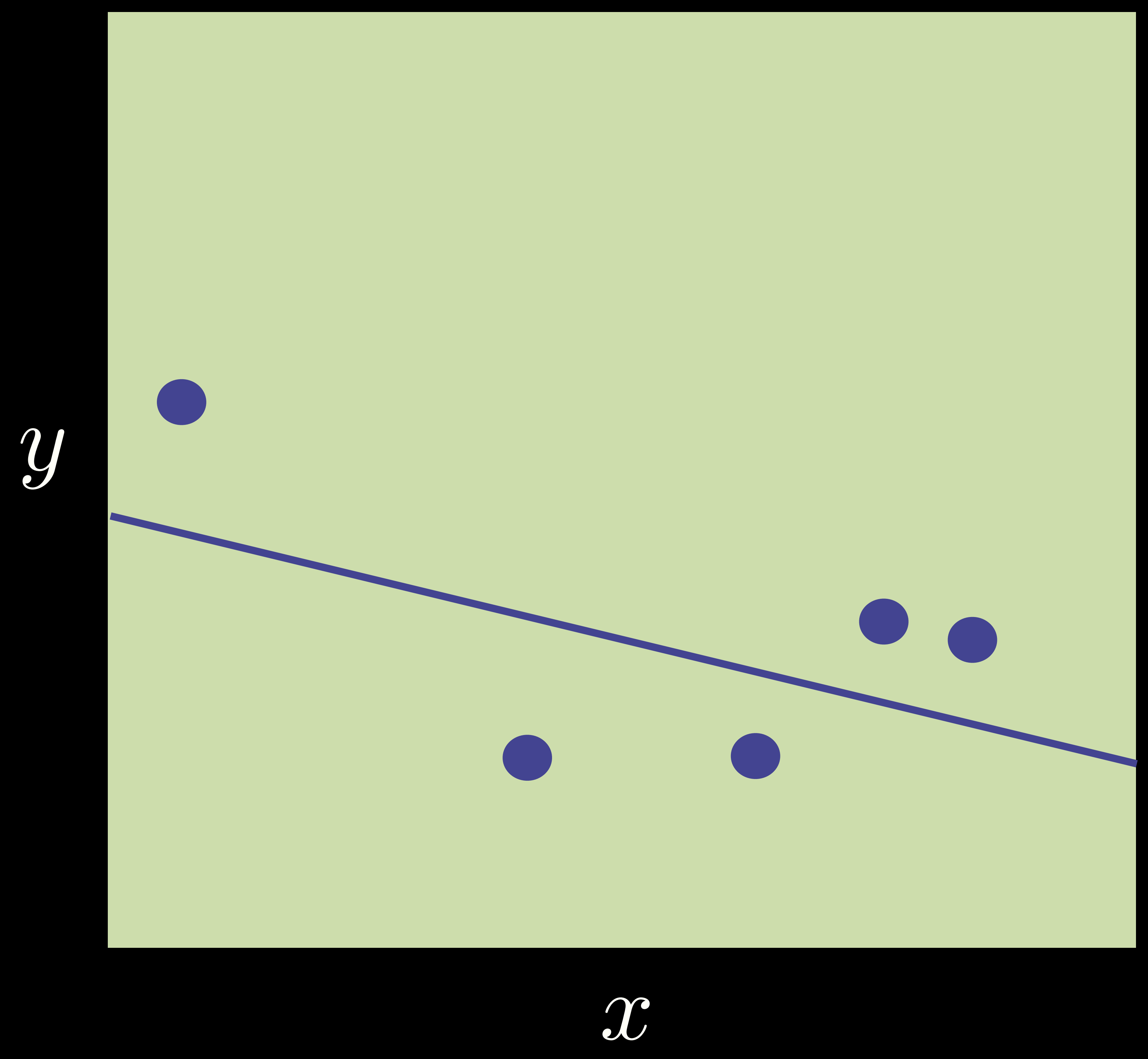
OVERFITTING

- Some training data



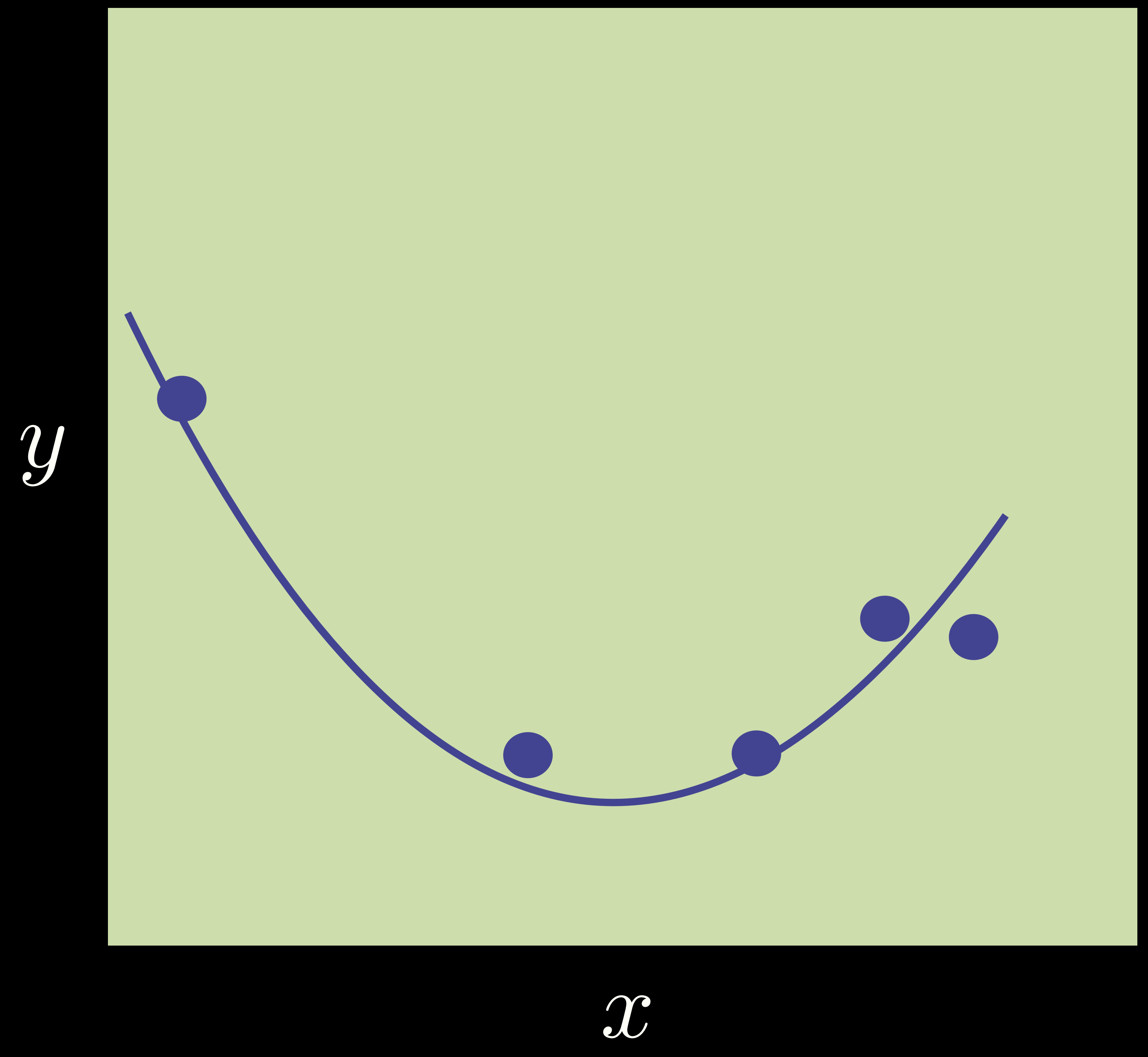
OVERFITTING

- Ein > Large, no good hypothesis



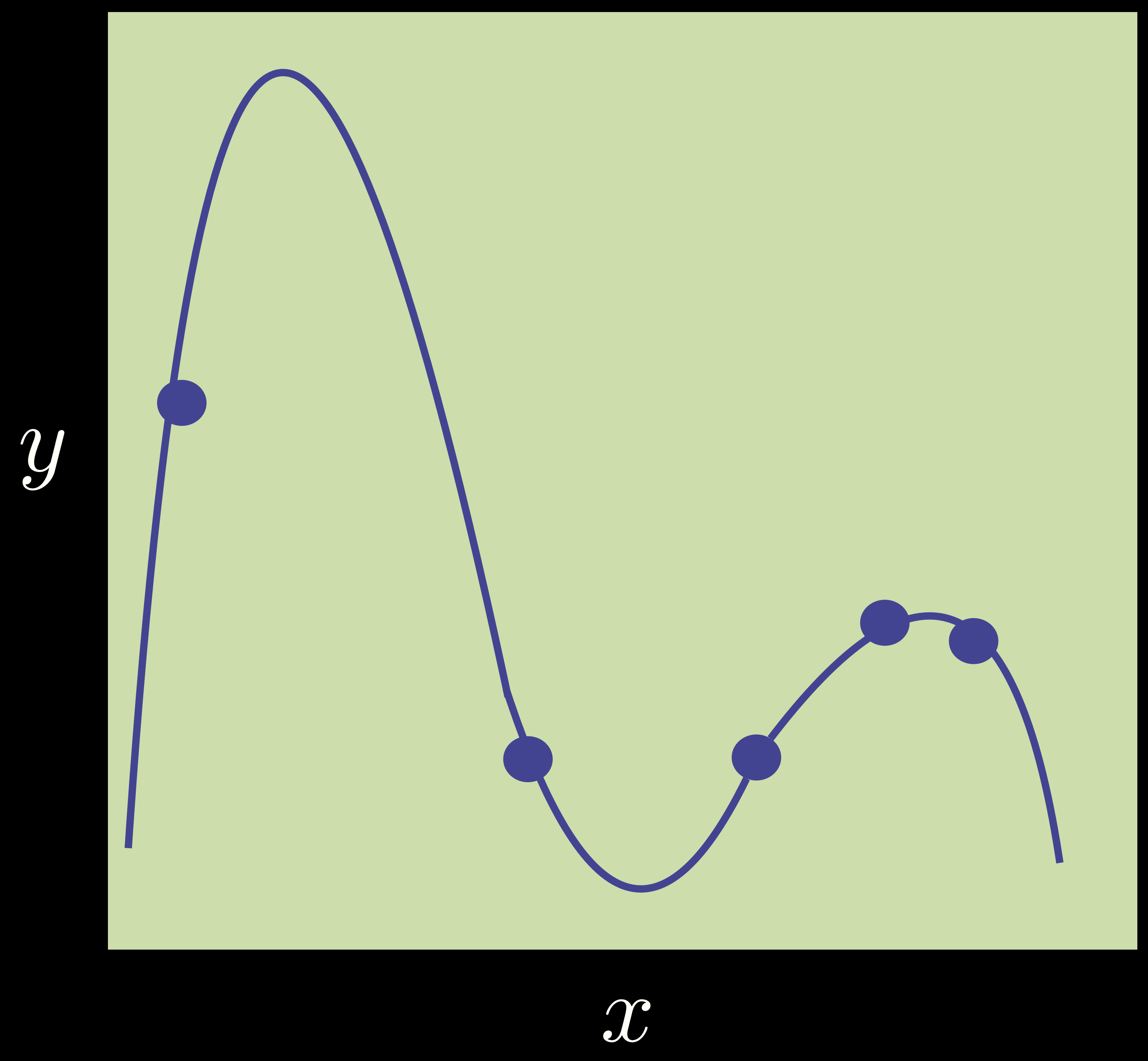
OVERFITTING

- Ein > 0, not perfect fit



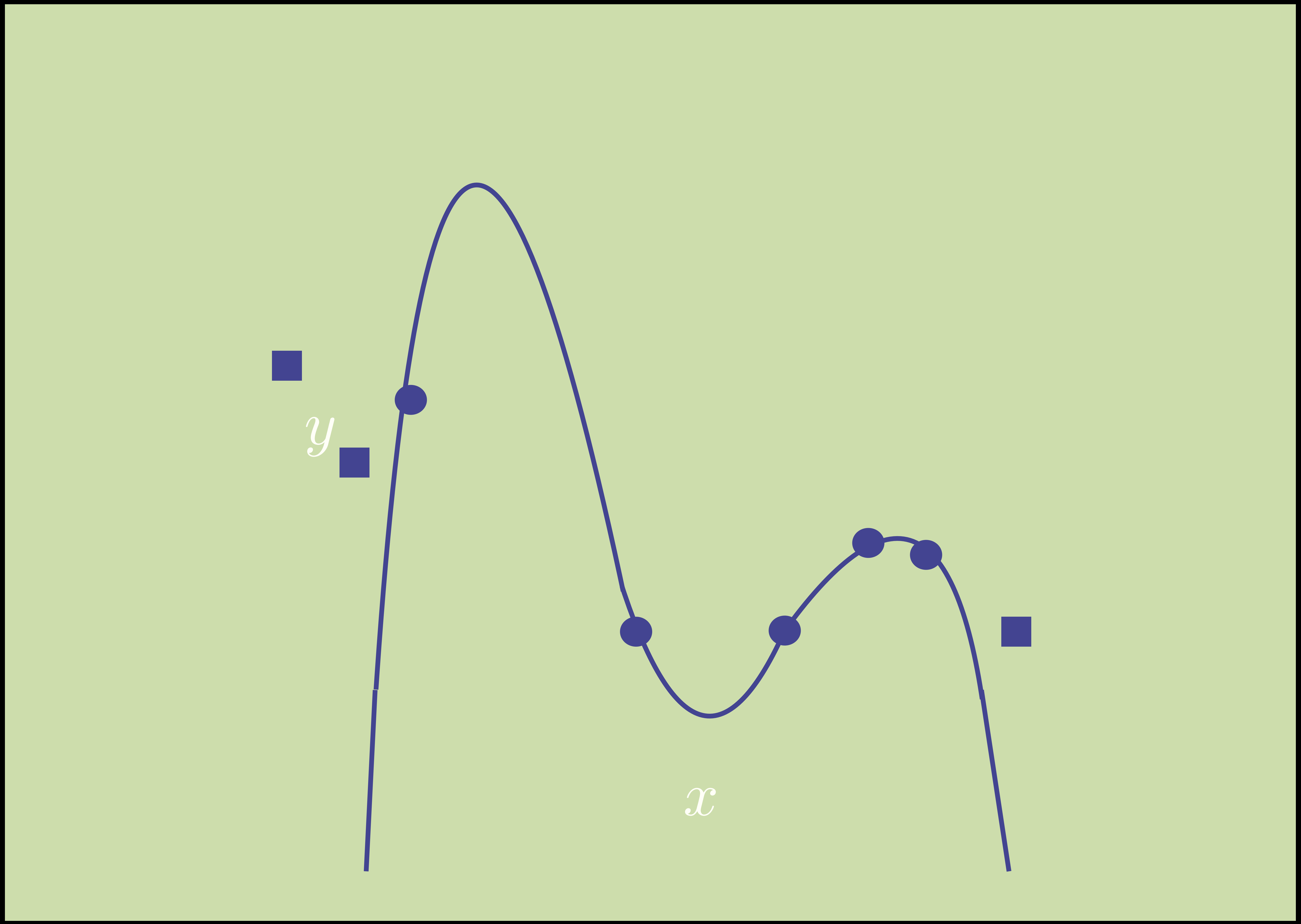
OVERFITTING

- Ein = 0 , fits perfect on training data
- Success! Or?



OVERFITTING

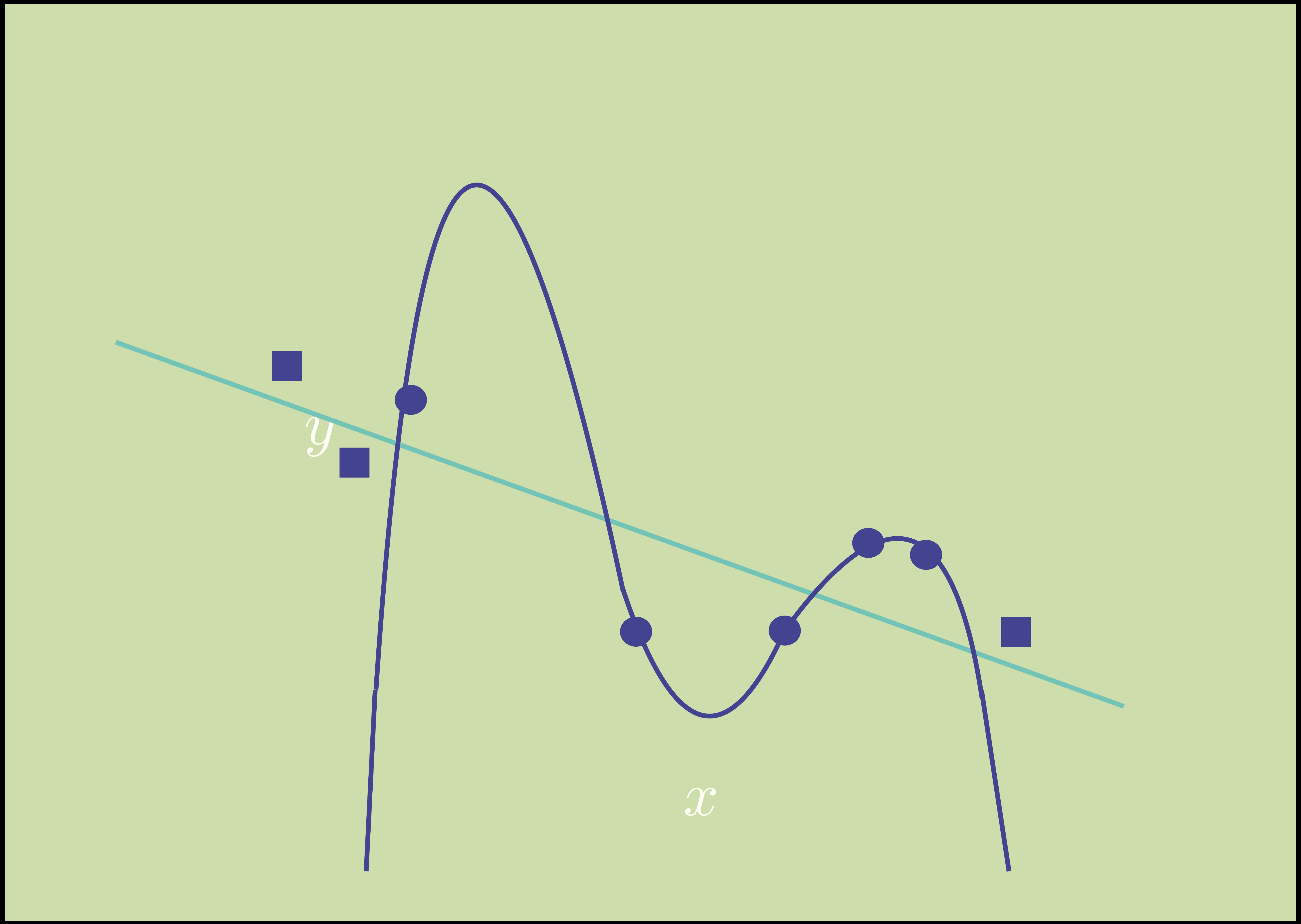
- Ein = 0 , fits perfect on training data
- Eout = Really Big!



OVERFITTING

- Ein = 0 , fits perfect on training data
- Eout = Really Big!

- We have fitted the noise!!!



REGULARISATION

- One of the main solutions to Overfitting
- You try to smoothen the fit with “breaks” on the weights

$$\lambda$$

TO SUMMARISE

- Overfitting is the problem
- Noise is the cause
- We detect it with Validation
- We cure it with Regularisation

AGENDA

- Introduction and context
- The work process
- The learning problem
- Validation and overfitting
- **Tools**
- Risks and ethics
- Demo

TOOLS

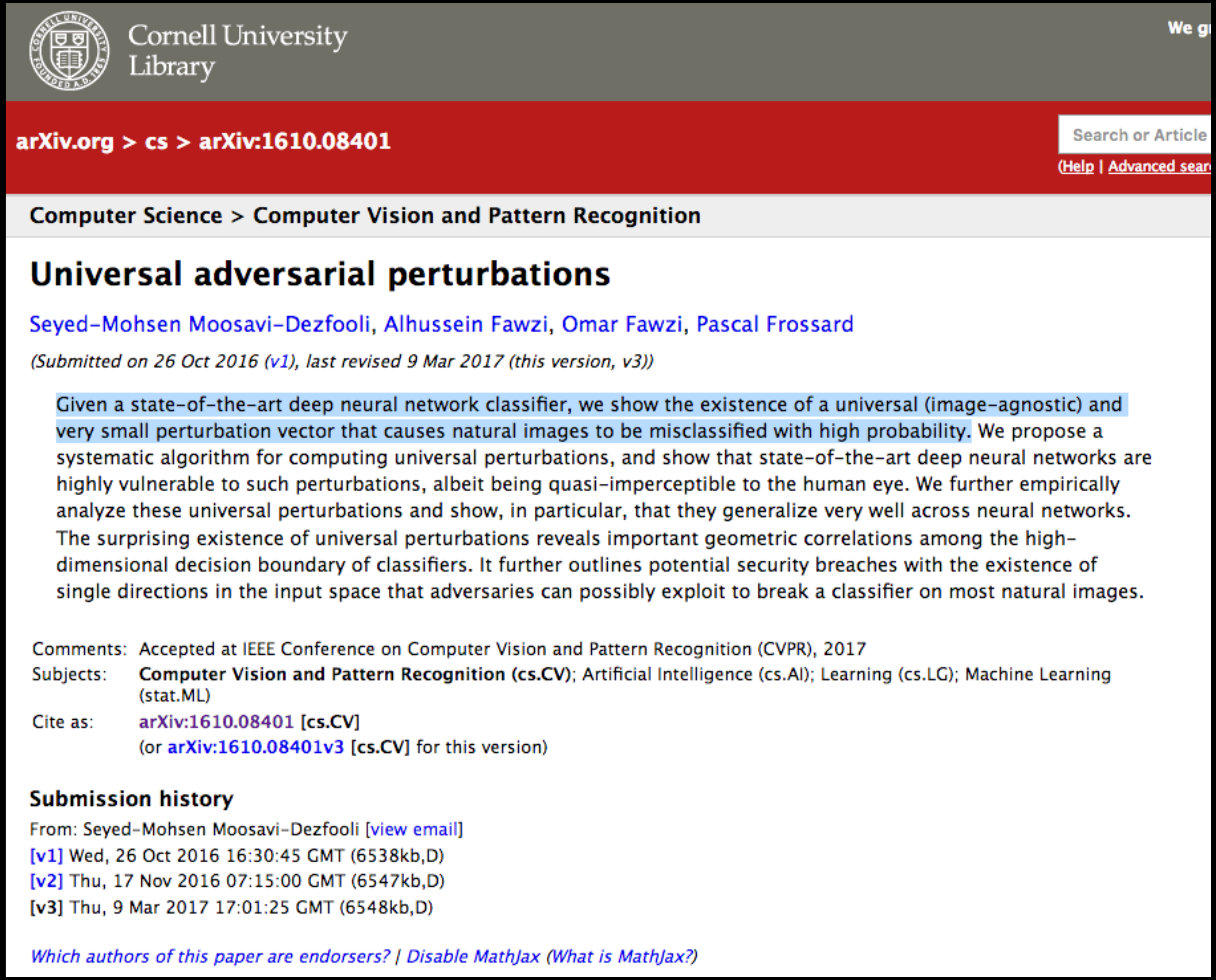
- Languages
  - Matlab, R, Python, Javascript, Julia men även Java
- Frameworks
  - Low level: Tensor Flow, Theano, MXNet
  - High Level: Keras, DeepLearning4J
- Hardware: Cuda
- End 2 End: H20

AGENDA

- Introduction and context
- The work process
- The learning problem
- Validation and overfitting
- Tools
- **Risks and ethics**
- Demo

ADVERSARIAL PERTURBATIONS

- Anomaly detection
- Self-driving cars



Cornell University Library

We g

arXiv.org > cs > arXiv:1610.08401

Search or Article

(Help | Advanced sear

Computer Science > Computer Vision and Pattern Recognition

Universal adversarial perturbations

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, Pascal Frossard

(Submitted on 26 Oct 2016 (v1), last revised 9 Mar 2017 (this version, v3))

Given a state-of-the-art deep neural network classifier, we show the existence of a universal (image-agnostic) and very small perturbation vector that causes natural images to be misclassified with high probability. We propose a systematic algorithm for computing universal perturbations, and show that state-of-the-art deep neural networks are highly vulnerable to such perturbations, albeit being quasi-imperceptible to the human eye. We further empirically analyze these universal perturbations and show, in particular, that they generalize very well across neural networks. The surprising existence of universal perturbations reveals important geometric correlations among the high-dimensional decision boundary of classifiers. It further outlines potential security breaches with the existence of single directions in the input space that adversaries can possibly exploit to break a classifier on most natural images.

Comments: Accepted at IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Artificial Intelligence (cs.AI); Learning (cs.LG); Machine Learning (stat.ML)
Cite as: arXiv:1610.08401 **[cs.CV]**
(or arXiv:1610.08401v3 **[cs.CV]** for this version)

Submission history

From: Seyed-Mohsen Moosavi-Dezfooli [view email]
[v1] Wed, 26 Oct 2016 16:30:45 GMT (6538kb,D)
[v2] Thu, 17 Nov 2016 07:15:00 GMT (6547kb,D)
[v3] Thu, 9 Mar 2017 17:01:25 GMT (6548kb,D)

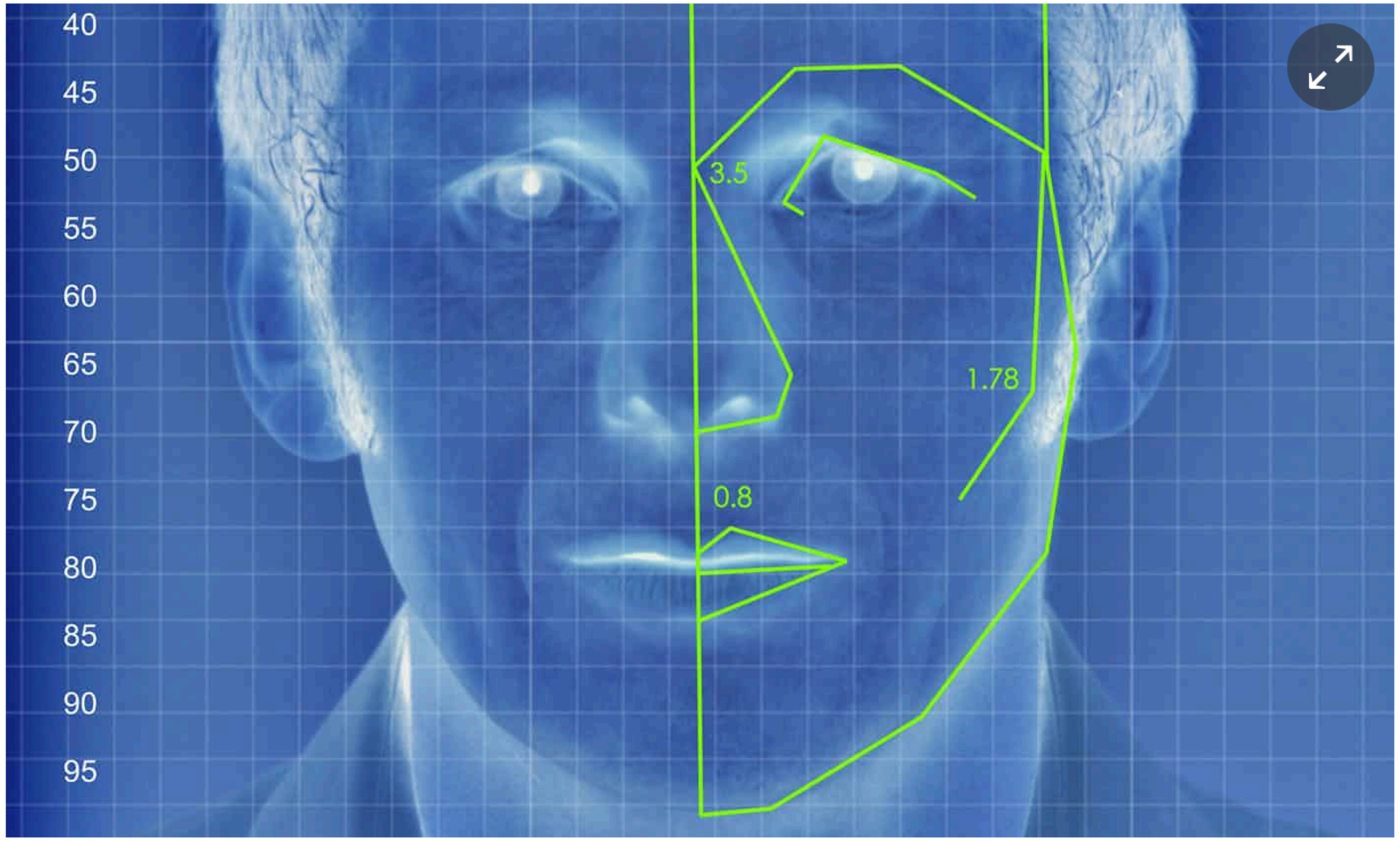
Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Given a state-of-the-art deep neural network classifier, we show the existence of a universal (image-agnostic) and very small perturbation vector that causes natural images to be misclassified with high probability. We propose a

Artificial intelligence (AI)

New AI can guess whether you're gay or straight from a photograph

An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions



An illustrated depiction of facial analysis technology similar to that used in the experiment. Illustration: Alamy

"- The primitive forms of artificial intelligence we already have have proved very useful. But I think the development of full artificial intelligence could spell the end of the human race."

Stephen Hawking, 2015

AGENDA

- Introduction and context
- The work process
- The learning problem
- Validation and overfitting
- Tools
- Risks and ethics
- **Demo**

H20

- End 2 End tool covering the whole workflow
- Nice GUI (Notebook Style)
- Both REST, Python, R, Scala API's
- Versions for Deep Learning, GPU etc etc …
- Clustering of compute nodes
- Apache 2.0 License

THE PROCESS

%

BUSINESS TARGET

WINE

3840 sorts of wine where tasted and graded and then sent to physiochemical analysis.

Create a formula that can determine the wine quality from the physiochemical attributes

Data from UCI Machine Learning Data Set repository

THE DATA

INPUT DATA

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol

OUTPUT DATA

Quality Score from 0.0 to 10.0

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Browse Through: 416 Data Sets

Table View List View

Default Task
- Classification (308)
- Regression (78)
- Clustering (69)
- Other (54)

Attribute Type
- Categorical (37)
- Numerical (265)
- Mixed (55)

Data Type
- Multivariate (317)
- Univariate (18)
- Sequential (42)
- Time-Series (77)
- Text (42)
- Domain-Theory (22)
- Other (21)

Area
- Life Sciences (97)
- Physical Sciences (47)
- CS / Engineering (140)
- Social Sciences (24)
- Business (26)
- Game (10)
- Other (68)

 # Attributes
- Less than 10 (97)
- 10 to 100 (191)
- Greater than 100 (73)

 # Instances
- Less than 100 (23)
- 100 to 1000 (146)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
Audiology (Original)	Multivariate	Classification	Categorical	226		1987
Audiology (Standardized)	Multivariate	Classification	Categorical	226	69	1992
	Multivariate	Regression	Categorical, Real	398	8	1993

Index of /ml/machine-learning-data

https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

Index of /ml/machine-learning-databases/wine-quality

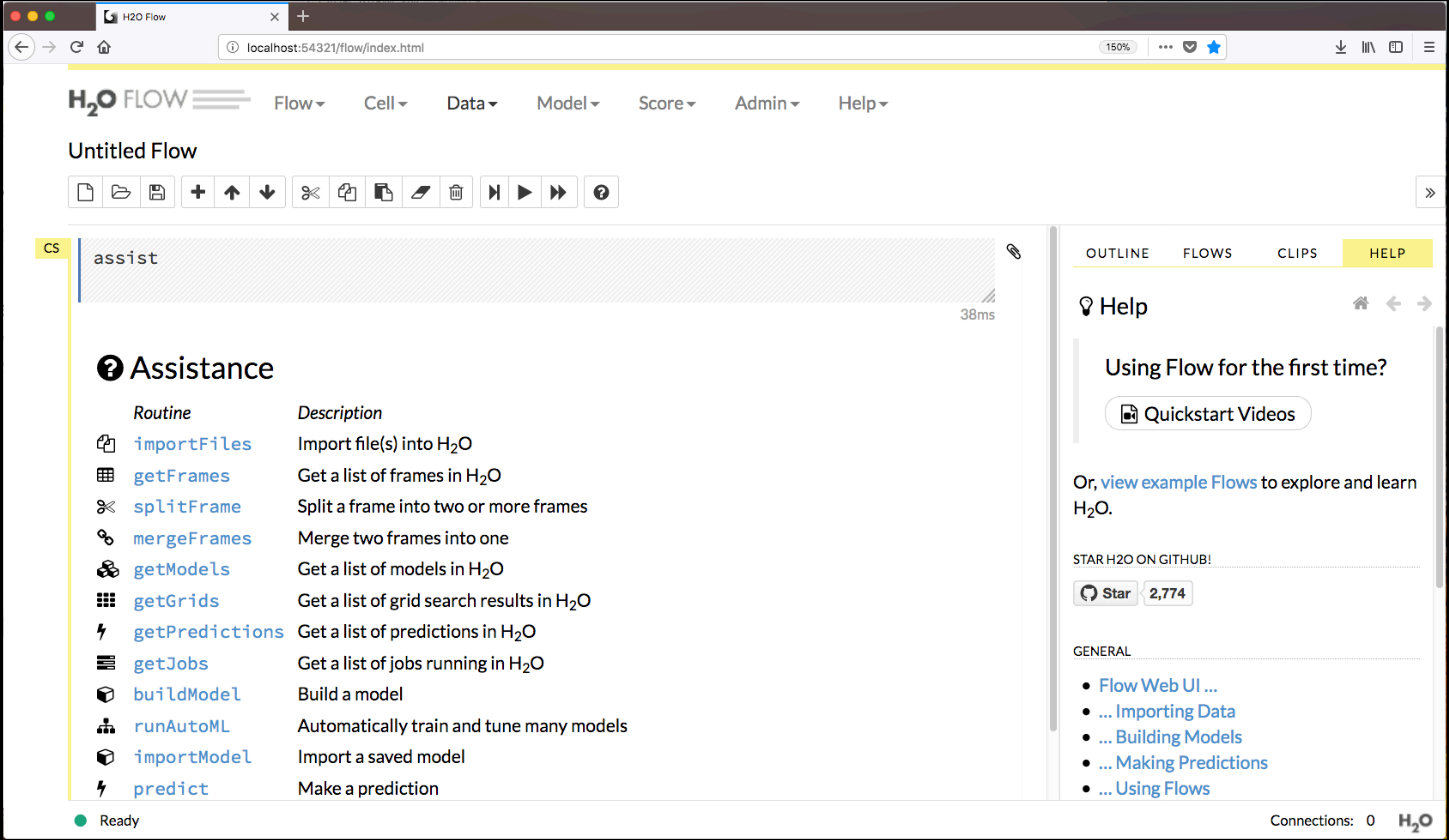
Name	Last modified	Size	Description
Parent Directory		-	
winequality-red.csv	16-Oct-2009 14:36	82K	
winequality-white.csv	16-Oct-2009 14:36	258K	
winequality.names	21-Oct-2009 11:00	3.2K	

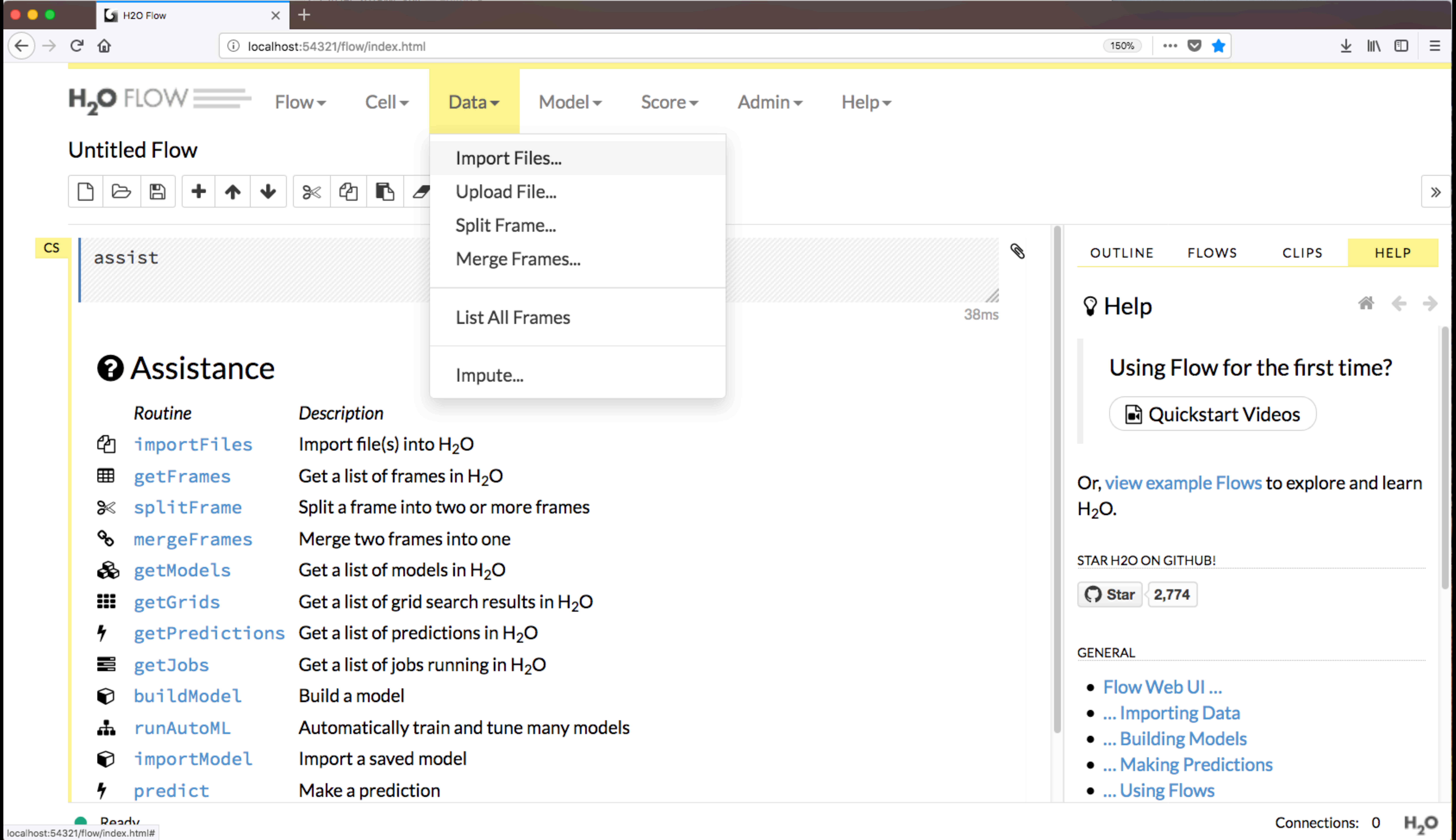
Apache/2.2.15 (CentOS) Server at archive.ics.uci.edu Port 443 h2o — -bash — 82×22

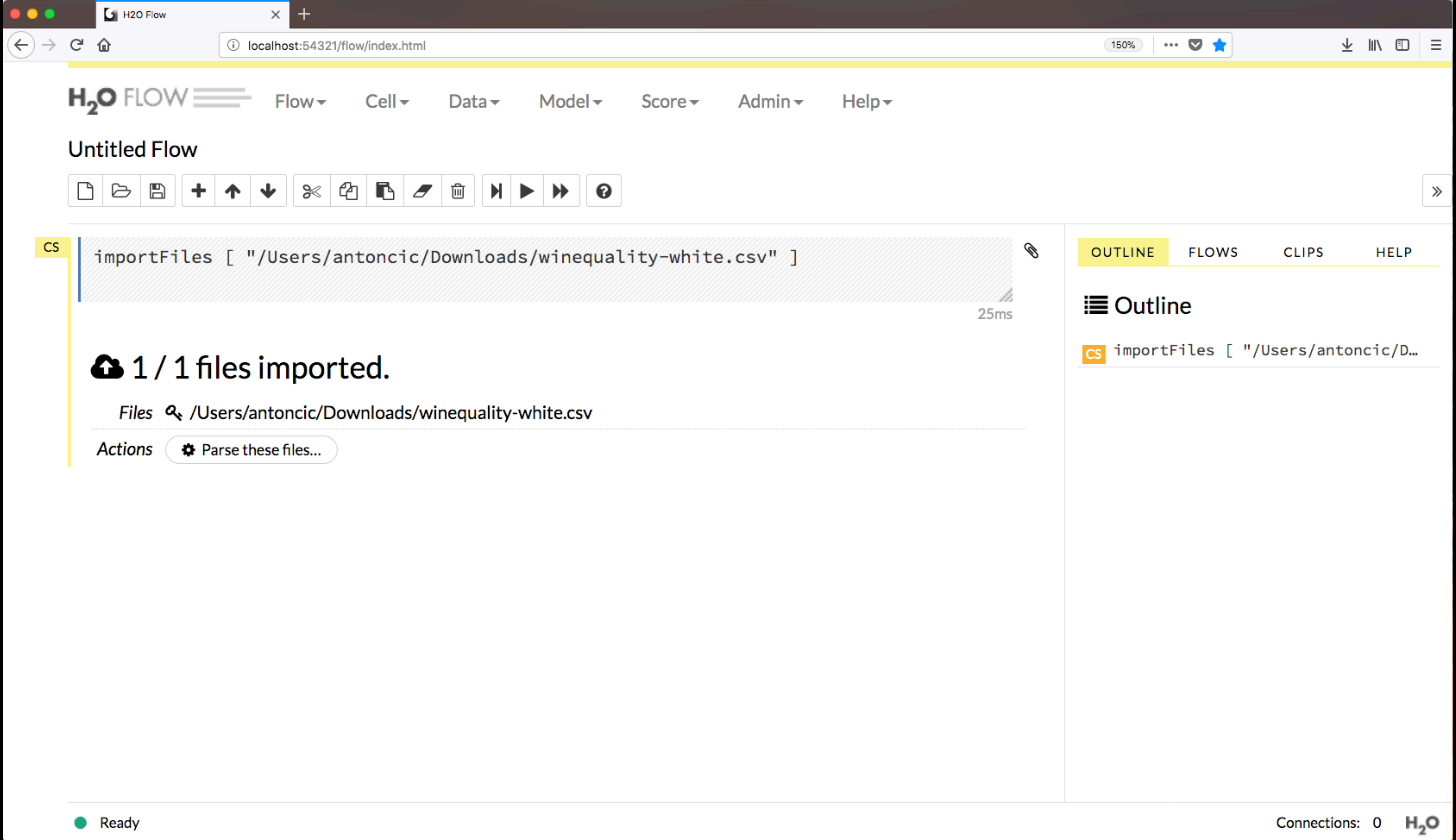
```
Last login: Tue Jan 23 10:48:44 on ttys000
antoncic@NiklasMBP:~$ cd MLTools/h2o
antoncic@NiklasMBP:~/MLTools/h2o$ java -jar h2o.jar
```

h2o — java -jar h2o.jar — 126×22

```
01-23 10:57:16.900 192.168.0.129:54321   27951  main      INFO: Cur dir: '/Users/antoncic/MLTools/h2o'
01-23 10:57:16.903 192.168.0.129:54321   27951  main      INFO: HDFS subsystem successfully initialized
01-23 10:57:16.905 192.168.0.129:54321   27951  main      INFO: S3 subsystem successfully initialized
01-23 10:57:16.905 192.168.0.129:54321   27951  main      INFO: Flow dir: '/Users/antoncic/h2oflows'
01-23 10:57:16.921 192.168.0.129:54321   27951  main      INFO: Cloud of size 1 formed [/192.168.0.129:54321]
01-23 10:57:16.928 192.168.0.129:54321   27951  main      INFO: Registered parsers: [GUESS, ARFF, XLS, SVMLight, AVRO, PARQUET
, CSV]
01-23 10:57:16.928 192.168.0.129:54321   27951  main      INFO: Watchdog extension initialized
01-23 10:57:16.928 192.168.0.129:54321   27951  main      INFO: XGBoost extension initialized
01-23 10:57:16.928 192.168.0.129:54321   27951  main      INFO: KrbStandalone extension initialized
01-23 10:57:16.928 192.168.0.129:54321   27951  main      INFO: Registered 3 core extensions in: 83ms
01-23 10:57:16.928 192.168.0.129:54321   27951  main      INFO: Registered H2O core extensions: [Watchdog, XGBoost, KrbStandal
one]
01-23 10:57:17.136 192.168.0.129:54321   27951  main      INFO: Registered: 162 REST APIs in: 207ms
01-23 10:57:17.136 192.168.0.129:54321   27951  main      INFO: Registered REST API extensions: [XGBoost, Algos, AutoML, Core
V3, Core V4]
01-23 10:57:17.226 192.168.0.129:54321   27951  main      INFO: Registered: 232 schemas in 90ms
01-23 10:57:17.226 192.168.0.129:54321   27951  main      INFO: H2O started in 2262ms
01-23 10:57:17.226 192.168.0.129:54321   27951  main      INFO:
01-23 10:57:17.226 192.168.0.129:54321   27951  main      INFO: Open H2O Flow in your web browser: http://192.168.0.129:54321
01-23 10:57:17.226 192.168.0.129:54321   27951  main      INFO:
```


H2O Flow
localhost:54321/flow/index.html
150%
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
CS
assist
38ms
Assistance
Routine
Description
importFiles
Import file(s) into H2O
getFrames
Get a list of frames in H2O
splitFrame
Split a frame into two or more frames
mergeFrames
Merge two frames into one
getModels
Get a list of models in H2O
getGrids
Get a list of grid search results in H2O
getPredictions
Get a list of predictions in H2O
getJobs
Get a list of jobs running in H2O
buildModel
Build a model
runAutoML
Automatically train and tune many models
importModel
Import a saved model
predict
Make a prediction
OUTLINE
FLOWS
CLIPS
HELP
Help
Using Flow for the first time?
Quickstart Videos
Or, view example Flows to explore and learn H2O.
STAR H2O ON GITHUB!
Star
2,774
GENERAL
Flow Web UI ...
... Importing Data
... Building Models
... Making Predictions
... Using Flows
Ready
Connections: 0
H2O


H2O Flow
localhost:54321/flow/index.html
150%
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
Import Files...
Upload File...
Split Frame...
Merge Frames...
List All Frames
Impute...
CS
assist
38ms
Assistance
Routine
Description
importFiles
Import file(s) into H2O
getFrames
Get a list of frames in H2O
splitFrame
Split a frame into two or more frames
mergeFrames
Merge two frames into one
getModels
Get a list of models in H2O
getGrids
Get a list of grid search results in H2O
getPredictions
Get a list of predictions in H2O
getJobs
Get a list of jobs running in H2O
buildModel
Build a model
runAutoML
Automatically train and tune many models
importModel
Import a saved model
predict
Make a prediction
OUTLINE
FLOWS
CLIPS
HELP
Help
Using Flow for the first time?
Quickstart Videos
Or, view example Flows to explore and learn H2O.
STAR H2O ON GITHUB!
Star
2,774
GENERAL
Flow Web UI ...
... Importing Data
... Building Models
... Making Predictions
... Using Flows
Ready
Connections: 0
localhost:54321/flow/index.html#


H2O Flow
localhost:54321/flow/index.html
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
CS
importFiles ["/Users/antoncic/Downloads/winequality-white.csv"]
25ms
1 / 1 files imported.
Files
/Users/antoncic/Downloads/winequality-white.csv
Actions
Parse these files...
OUTLINE
FLOWS
CLIPS
HELP
Outline
CS
importFiles ["/Users/antoncic/D…
Ready
Connections: 0

H2O Flow

localhost:54321/flow/index.html

H2O FLOW   Flow ▾   Cell ▾   Data ▾   Model ▾   Score ▾   Admin ▾   Help ▾

Untitled Flow

Setup Parse

PARSE CONFIGURATION

Sources nfs://Users/antoncic/Downloads/winequality-white.csv

ID winequality_white.hex

Parser CSV

Separator ;: '59'

Column Headers
- ○ Auto
- ● First row contains column names
- ○ First row contains data

Options
- ☐ Enable single quotes as a field quotation character
- ☑ Delete on done

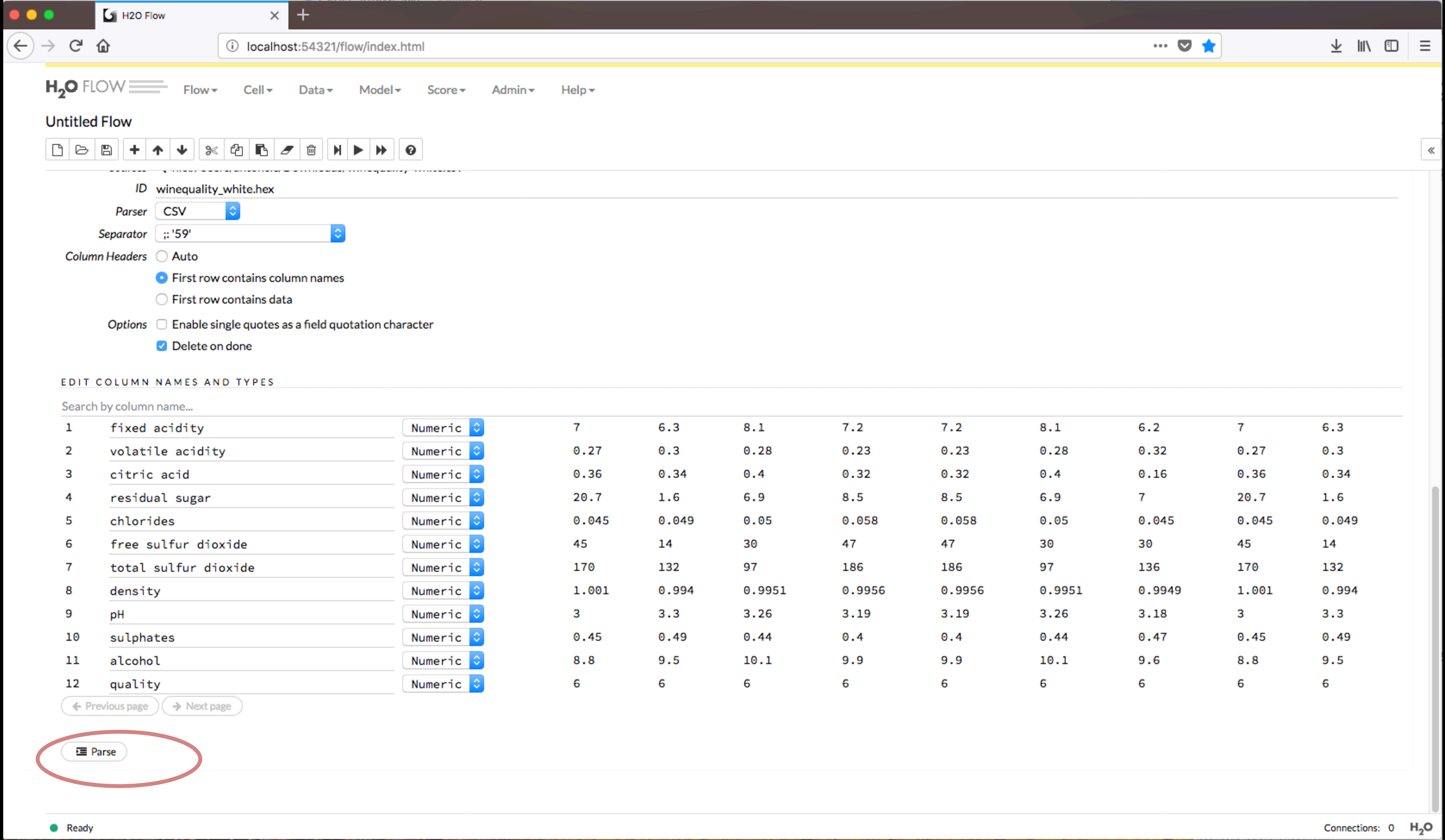
EDIT COLUMN NAMES AND TYPES

Search by column name...

1	fixed acidity	Numeric	7	6.3	8.1	7.2	7.2	8.1	6.2	7	6.3
2	volatile acidity	Numeric	0.27	0.3	0.28	0.23	0.23	0.28	0.32	0.27	0.3
3	citric acid	Numeric	0.36	0.34	0.4	0.32	0.32	0.4	0.16	0.36	0.34
4	residual sugar	Numeric	20.7	1.6	6.9	8.5	8.5	6.9	7	20.7	1.6
5	chlorides	Numeric	0.045	0.049	0.05	0.058	0.058	0.05	0.045	0.045	0.049
6	free sulfur dioxide	Numeric	45	14	30	47	47	30	30	45	14
7	total sulfur dioxide	Numeric	170	132	97	186	186	97	136	170	132
8	density	Numeric	1.001	0.994	0.9951	0.9956	0.9956	0.9951	0.9949	1.001	0.994
9	pH	Numeric	3	3.3	3.26	3.19	3.19	3.26	3.18	3	3.3
10	sulphates	Numeric	0.45	0.49	0.44	0.4	0.4	0.44	0.47	0.45	0.49
11	alcohol	Numeric	8.8	9.5	10.1	9.9	9.9	10.1	9.6	8.8	9.5
12	quality	Numeric	6	6	6	6	6	6	6	6	6

← Previous page   → Next page

● Ready   Connections: 0   H2O


H2O Flow
localhost:54321/flow/index.html
H2O FLOW
Flow Cell Data Model Score Admin Help
Untitled Flow
ID winequality_white.hex
Parser CSV
Separator ;: '59'
Column Headers Auto
First row contains column names
First row contains data
Options Enable single quotes as a field quotation character
Delete on done
EDIT COLUMN NAMES AND TYPES
Search by column name...
1 fixed acidity Numeric 7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3
2 volatile acidity Numeric 0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3
3 citric acid Numeric 0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34
4 residual sugar Numeric 20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6
5 chlorides Numeric 0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049
6 free sulfur dioxide Numeric 45 14 30 47 47 30 30 45 14
7 total sulfur dioxide Numeric 170 132 97 186 186 97 136 170 132
8 density Numeric 1.001 0.994 0.9951 0.9956 0.9956 0.9951 0.9949 1.001 0.994
9 pH Numeric 3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3
10 sulphates Numeric 0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49
11 alcohol Numeric 8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5
12 quality Numeric 6 6 6 6 6 6 6 6 6
Previous page
Next page
Parse
Ready
Connections: 0

H2O Flow

localhost:54321/flow/index.html

H2O FLOW Flow Cell Data Model Score Admin Help

Untitled Flow

Parse

CS

```
parseFiles
  source_frames: ["nfs://Users/antoncic/Downloads/winequality-white.csv"]
  destination_frame: "winequality_white.hex"
  parse_type: "CSV"
  separator: 59
  number_columns: 12
  single_quotes: false
  column_names: ["fixed acidity","volatile acidity","citric acid","residual sugar","chlorides","free sulfur dioxide","total sulfur
dioxide","density","pH","sulphates","alcohol","quality"]
  column_types: ["Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric"]
  delete_on_done: true
  check_header: 1
  chunk_size: 8264
```

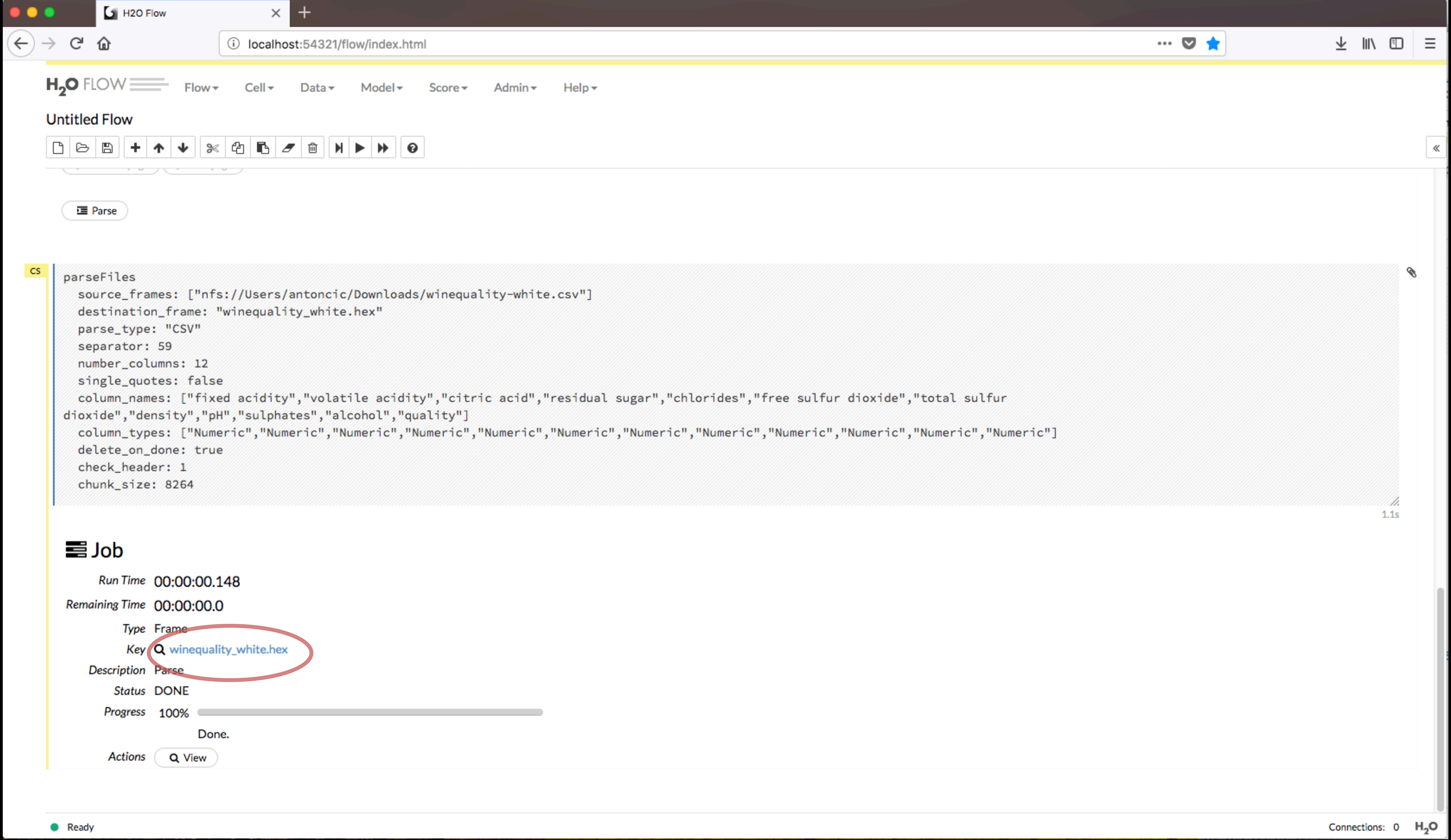
1.1s

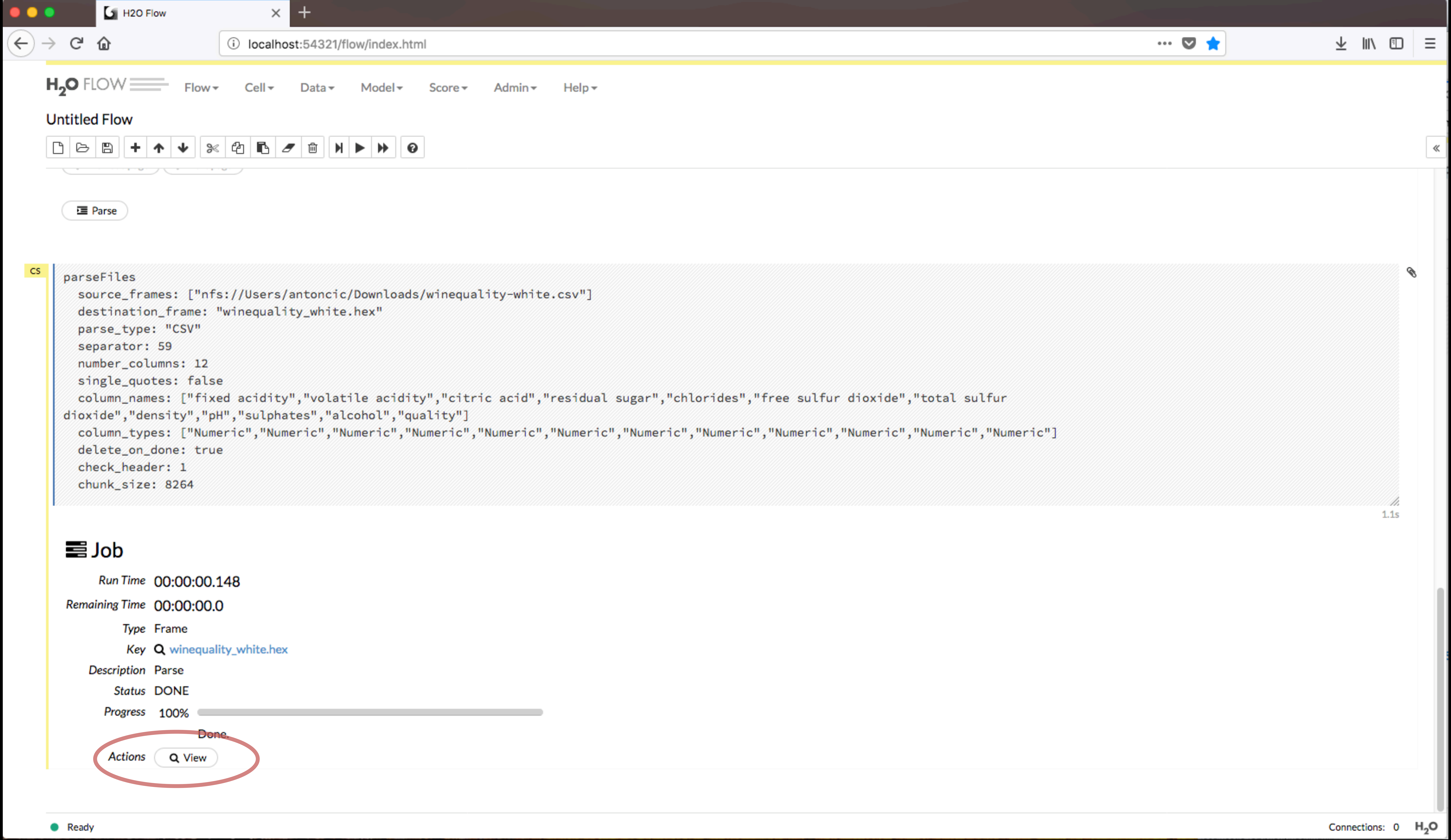
Job

Run Time	00:00:00.148
Remaining Time	00:00:00.0
Type	Frame
Key	winequality_white.hex
Description	Parse
Status	DONE
Progress	100%
	Done.
Actions	View

Ready

Connections: 0 H2O


H2O Flow
localhost:54321/flow/index.html
H2O FLOW
Flow Cell Data Model Score Admin Help
Untitled Flow
Parse
CS
parseFiles
source_frames: ["nfs://Users/antoncic/Downloads/winequality-white.csv"]
destination_frame: "winequality_white.hex"
parse_type: "CSV"
separator: 59
number_columns: 12
single_quotes: false
column_names: ["fixed acidity","volatile acidity","citric acid","residual sugar","chlorides","free sulfur dioxide","total sulfur dioxide","density","pH","sulphates","alcohol","quality"]
column_types: ["Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric"]
delete_on_done: true
check_header: 1
chunk_size: 8264
1.1s
Job
Run Time 00:00:00.148
Remaining Time 00:00:00.0
Type Frame
Key winequality_white.hex
Description Parse
Status DONE
Progress 100%
Done.
Actions View
Ready
Connections: 0
H2O

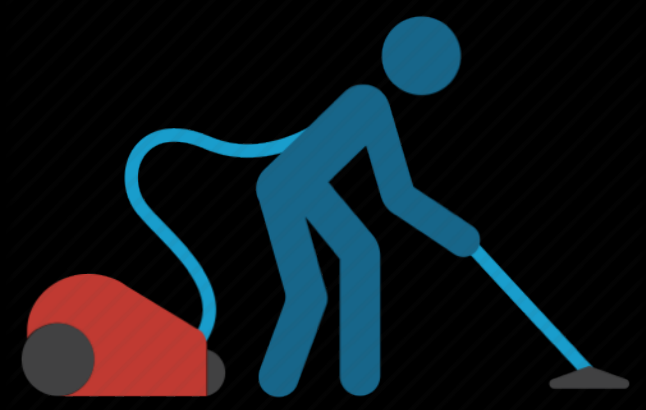

H2O Flow
localhost:54321/flow/index.html
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
Parse
CS
parseFiles
source_frames: ["nfs://Users/antoncic/Downloads/winequality-white.csv"]
destination_frame: "winequality_white.hex"
parse_type: "CSV"
separator: 59
number_columns: 12
single_quotes: false
column_names: ["fixed acidity","volatile acidity","citric acid","residual sugar","chlorides","free sulfur dioxide","total sulfur dioxide","density","pH","sulphates","alcohol","quality"]
column_types: ["Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric","Numeric"]
delete_on_done: true
check_header: 1
chunk_size: 8264
1.1s
Job
Run Time 00:00:00.148
Remaining Time 00:00:00.0
Type Frame
Key winequality_white.hex
Description Parse
Status DONE
Progress 100%
Done.
Actions View
Ready
Connections: 0
H2O

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA



PRE PROCESS

H2O Flow

localhost:54321/flow/index.html

H2O FLOW Flow▾ Cell▾ Data▾ Model▾ Score▾ Admin▾ Help▾

Untitled Flow

97ms

winequality_white.hex

Actions: View Data | Split... | Build Model... | Predict | Download | Export | Delete

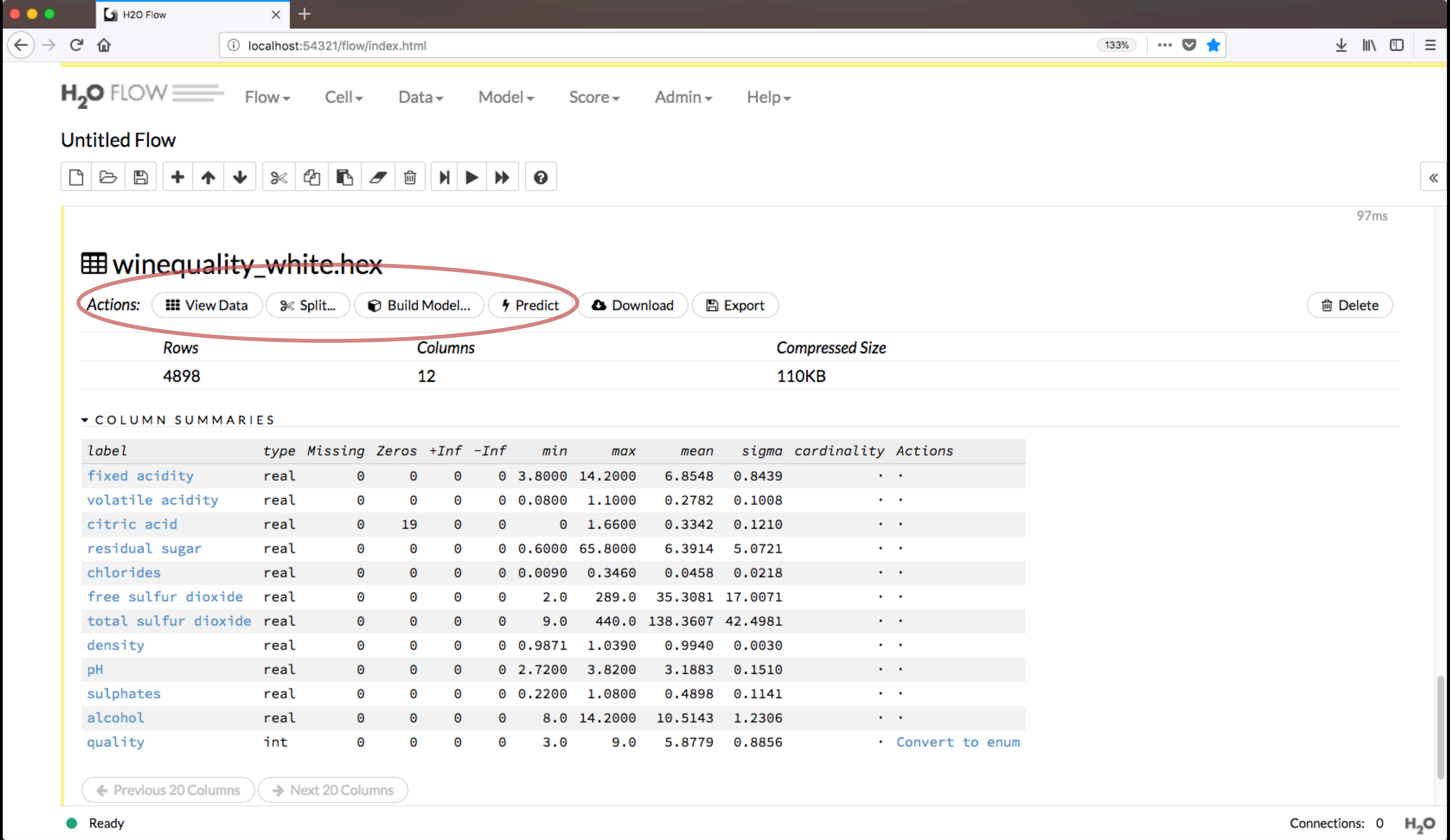
Rows	*Columns*	*Compressed Size*
4898	12	110KB

▾ COLUMN SUMMARIES

label	*type*	*Missing*	*Zeros*	*+Inf*	*-Inf*	*min*	*max*	*mean*	*sigma*	*cardinality*	*Actions*
fixed acidity	real	0	0	0	0	3.8000	14.2000	6.8548	0.8439	.	.
volatile acidity	real	0	0	0	0	0.0800	1.1000	0.2782	0.1008	.	.
citric acid	real	0	19	0	0	0	1.6600	0.3342	0.1210	.	.
residual sugar	real	0	0	0	0	0.6000	65.8000	6.3914	5.0721	.	.
chlorides	real	0	0	0	0	0.0090	0.3460	0.0458	0.0218	.	.
free sulfur dioxide	real	0	0	0	0	2.0	289.0	35.3081	17.0071	.	.
total sulfur dioxide	real	0	0	0	0	9.0	440.0	138.3607	42.4981	.	.
density	real	0	0	0	0	0.9871	1.0390	0.9940	0.0030	.	.
pH	real	0	0	0	0	2.7200	3.8200	3.1883	0.1510	.	.
sulphates	real	0	0	0	0	0.2200	1.0800	0.4898	0.1141	.	.
alcohol	real	0	0	0	0	8.0	14.2000	10.5143	1.2306	.	.
quality	int	0	0	0	0	3.0	9.0	5.8779	0.8856	.	Convert to enum

← Previous 20 Columns | → Next 20 Columns

Ready Connections: 0 H2O



label	type	Missing	Zeros	+Inf	-Inf	min	max	mean	sigma	cardinality	Actions
fixed acidity	real	0	0	0	0	3.8000	14.2000	6.8548	0.8439	.	.
volatile acidity	real	0	0	0	0	0.0800	1.1000	0.2782	0.1008	.	.
citric acid	real	0	19	0	0	0	1.6600	0.3342	0.1210	.	.
residual sugar	real	0	0	0	0	0.6000	65.8000	6.3914	5.0721	.	.
chlorides	real	0	0	0	0	0.0090	0.3460	0.0458	0.0218	.	.
free sulfur dioxide	real	0	0	0	0	2.0	289.0	35.3081	17.0071	.	.
total sulfur dioxide	real	0	0	0	0	9.0	440.0	138.3607	42.4981	.	.
density	real	0	0	0	0	0.9871	1.0390	0.9940	0.0030	.	.
pH	real	0	0	0	0	2.7200	3.8200	3.1883	0.1510	.	.
sulphates	real	0	0	0	0	0.2200	1.0800	0.4898	0.1141	.	.
alcohol	real	0	0	0	0	8.0	14.2000	10.5143	1.2306	.	.
quality	int	0	0	0	0	3.0	9.0	5.8779	0.8856	.	Convert to enum

H2O Flow

localhost:54321/flow/index.html

H2O FLOW   Flow ▾   Cell ▾   Data ▾   Model ▾   Score ▾   Admin ▾   Help ▾

Untitled Flow

winequality_white.hex

▾ DATA

← Previous 20 Columns   → Next 20 Columns

Row	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
1	7.0	0.2700	0.3600	20.7000	0.0450	45.0	170.0	1.0010	3.0	0.4500	8.8000	6.0
2	6.3000	0.3000	0.3400	1.6000	0.0490	14.0	132.0	0.9940	3.3000	0.4900	9.5000	6.0
3	8.1000	0.2800	0.4000	6.9000	0.0500	30.0	97.0	0.9951	3.2600	0.4400	10.1000	6.0
4	7.2000	0.2300	0.3200	8.5000	0.0580	47.0	186.0	0.9956	3.1900	0.4000	9.9000	6.0
5	7.2000	0.2300	0.3200	8.5000	0.0580	47.0	186.0	0.9956	3.1900	0.4000	9.9000	6.0
6	8.1000	0.2800	0.4000	6.9000	0.0500	30.0	97.0	0.9951	3.2600	0.4400	10.1000	6.0
7	6.2000	0.3200	0.1600	7.0	0.0450	30.0	136.0	0.9949	3.1800	0.4700	9.6000	6.0
8	7.0	0.2700	0.3600	20.7000	0.0450	45.0	170.0	1.0010	3.0	0.4500	8.8000	6.0
9	6.3000	0.3000	0.3400	1.6000	0.0490	14.0	132.0	0.9940	3.3000	0.4900	9.5000	6.0
10	8.1000	0.2200	0.4300	1.5000	0.0440	28.0	129.0	0.9938	3.2200	0.4500	11.0	6.0
11	8.1000	0.2700	0.4100	1.4500	0.0330	11.0	63.0	0.9908	2.9900	0.5600	12.0	5.0
12	8.6000	0.2300	0.4000	4.2000	0.0350	17.0	109.0	0.9947	3.1400	0.5300	9.7000	5.0
13	7.9000	0.1800	0.3700	1.2000	0.0400	16.0	75.0	0.9920	3.1800	0.6300	10.8000	5.0
14	6.6000	0.1600	0.4000	1.5000	0.0440	48.0	143.0	0.9912	3.5400	0.5200	12.4000	7.0
15	8.3000	0.4200	0.6200	19.2500	0.0400	41.0	172.0	1.0002	2.9800	0.6700	9.7000	5.0
16	6.6000	0.1700	0.3800	1.5000	0.0320	28.0	112.0	0.9914	3.2500	0.5500	11.4000	7.0
17	6.3000	0.4800	0.0400	1.1000	0.0460	30.0	99.0	0.9928	3.2400	0.3600	9.6000	6.0
18	6.2000	0.6600	0.4800	1.2000	0.0290	29.0	75.0	0.9892	3.3300	0.3900	12.8000	8.0
19	7.4000	0.3400	0.4200	1.1000	0.0330	17.0	171.0	0.9917	3.1200	0.5300	11.3000	6.0

Ready   Connections: 0   H2O

H2O Flow

localhost:54321/flow/index.html 133%

H2O FLOW Flow ▾ Cell ▾ Data ▾ Model ▾ Score ▾ Admin ▾ Help ▾

Untitled Flow

97ms

winequality_white.hex

Actions: View Data | Split... | Build Model... | Predict | Download | Export | Delete

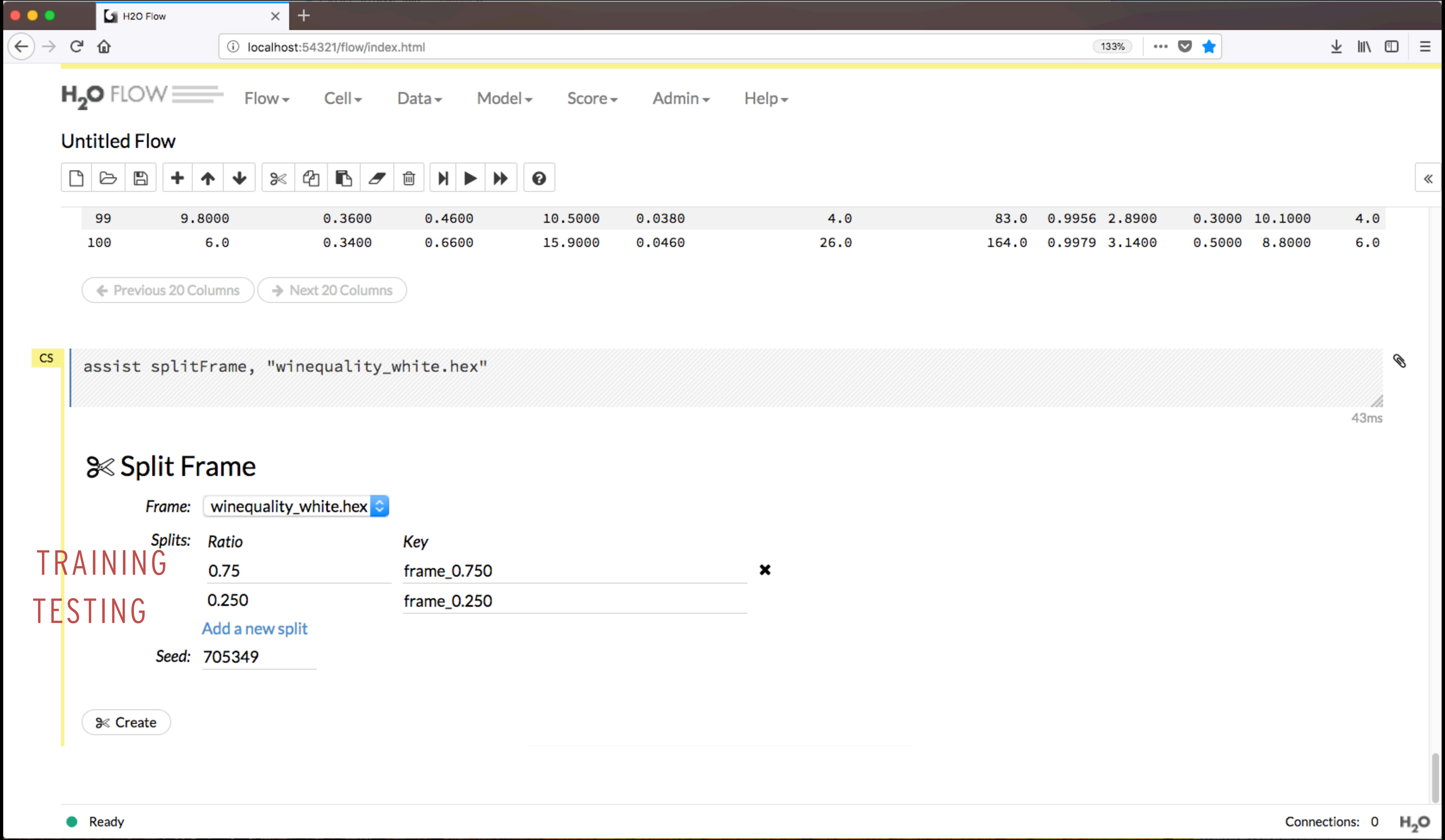
Rows	*Columns*	*Compressed Size*
4898	12	110KB

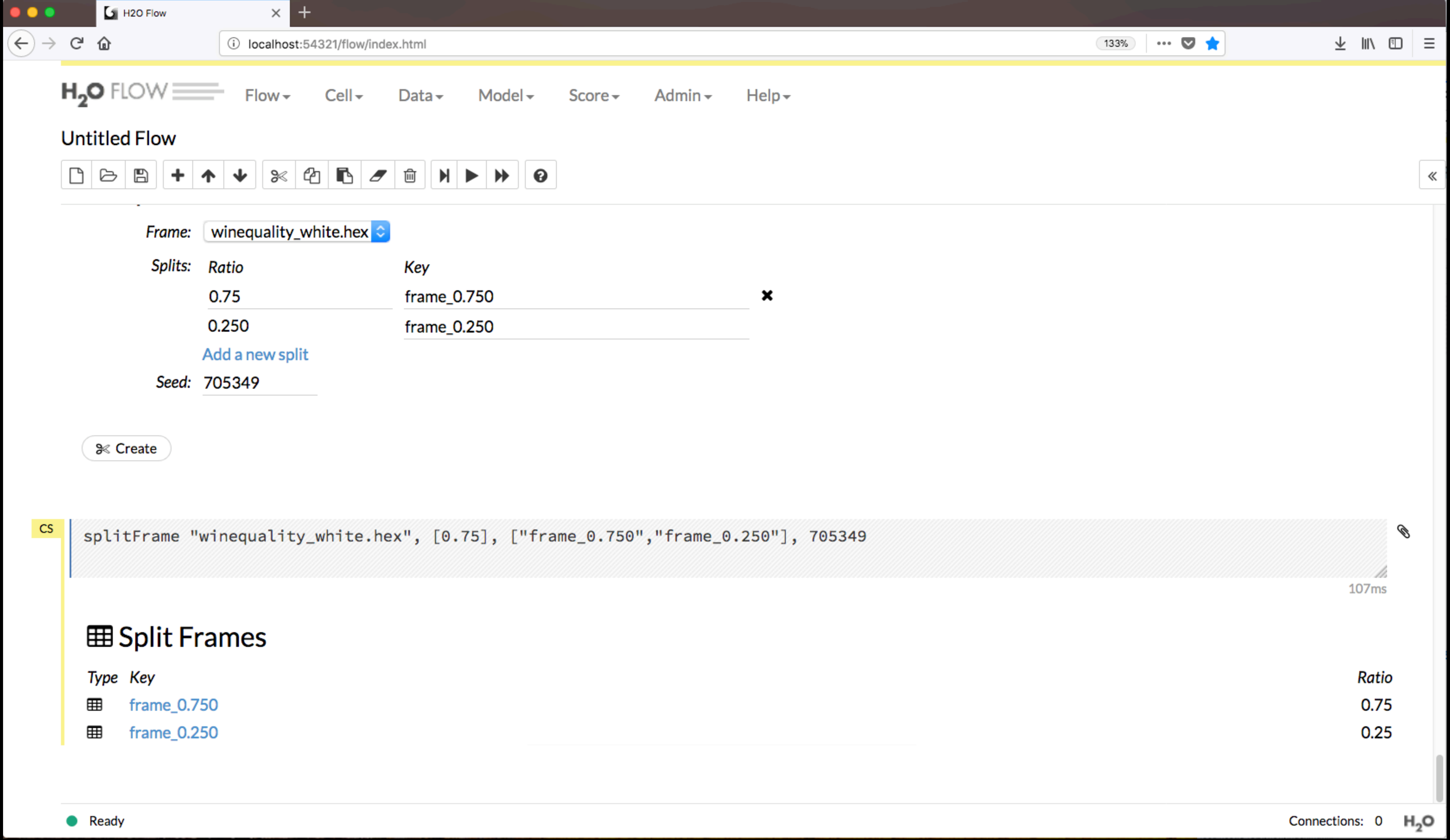
▾ COLUMN SUMMARIES

label	type	Missing	Zeros	+Inf	-Inf	min	max	mean	sigma	cardinality	Actions
fixed acidity	real	0	0	0	0	3.8000	14.2000	6.8548	0.8439	.	.
volatile acidity	real	0	0	0	0	0.0800	1.1000	0.2782	0.1008	.	.
citric acid	real	0	19	0	0	0	1.6600	0.3342	0.1210	.	.
residual sugar	real	0	0	0	0	0.6000	65.8000	6.3914	5.0721	.	.
chlorides	real	0	0	0	0	0.0090	0.3460	0.0458	0.0218	.	.
free sulfur dioxide	real	0	0	0	0	2.0	289.0	35.3081	17.0071	.	.
total sulfur dioxide	real	0	0	0	0	9.0	440.0	138.3607	42.4981	.	.
density	real	0	0	0	0	0.9871	1.0390	0.9940	0.0030	.	.
pH	real	0	0	0	0	2.7200	3.8200	3.1883	0.1510	.	.
sulphates	real	0	0	0	0	0.2200	1.0800	0.4898	0.1141	.	.
alcohol	real	0	0	0	0	8.0	14.2000	10.5143	1.2306	.	.
quality	int	0	0	0	0	3.0	9.0	5.8779	0.8856	.	Convert to enum

← Previous 20 Columns → Next 20 Columns

Ready Connections: 0 H2O


H2O Flow
localhost:54321/flow/index.html
133%
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
99 9.8000 0.3600 0.4600 10.5000 0.0380 4.0 83.0 0.9956 2.8900 0.3000 10.1000 4.0
100 6.0 0.3400 0.6600 15.9000 0.0460 26.0 164.0 0.9979 3.1400 0.5000 8.8000 6.0
Previous 20 Columns
Next 20 Columns
CS
assist splitFrame, "winequality_white.hex"
43ms
Split Frame
Frame: winequality_white.hex
Splits: Ratio Key
TRAINING
0.75 frame_0.750
TESTING
0.250 frame_0.250
Add a new split
Seed: 705349
Create
Ready
Connections: 0
H2O


H2O Flow
localhost:54321/flow/index.html
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
Frame: winequality_white.hex
Splits: Ratio Key
0.75 frame_0.750
0.250 frame_0.250
Add a new split
Seed: 705349
Create
CS
splitFrame "winequality_white.hex", [0.75], ["frame_0.750","frame_0.250"], 705349
107ms
Split Frames
Type Key Ratio
frame_0.750 0.75
frame_0.250 0.25
Ready
Connections: 0

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA

PRE PROCESS

SELECT MODEL

H2O Flow

localhost:54321/flow/index.html

H2O FLOW   Flow   Cell   Data   Model   Score   Admin   Help

Untitled Flow

97ms

winequality_white.hex

Actions: View Data   Split...   Build Model...   Predict   Download   Export   Delete

Rows	*Columns*	*Compressed Size*
4898	12	110KB

COLUMN SUMMARIES

label	type	Missing	Zeros	+Inf	-Inf	min	max	mean	sigma	cardinality	Actions
fixed acidity	real	0	0	0	0	3.8000	14.2000	6.8548	0.8439	.	.
volatile acidity	real	0	0	0	0	0.0800	1.1000	0.2782	0.1008	.	.
citric acid	real	0	19	0	0	0	1.6600	0.3342	0.1210	.	.
residual sugar	real	0	0	0	0	0.6000	65.8000	6.3914	5.0721	.	.
chlorides	real	0	0	0	0	0.0090	0.3460	0.0458	0.0218	.	.
free sulfur dioxide	real	0	0	0	0	2.0	289.0	35.3081	17.0071	.	.
total sulfur dioxide	real	0	0	0	0	9.0	440.0	138.3607	42.4981	.	.
density	real	0	0	0	0	0.9871	1.0390	0.9940	0.0030	.	.
pH	real	0	0	0	0	2.7200	3.8200	3.1883	0.1510	.	.
sulphates	real	0	0	0	0	0.2200	1.0800	0.4898	0.1141	.	.
alcohol	real	0	0	0	0	8.0	14.2000	10.5143	1.2306	.	.
quality	int	0	0	0	0	3.0	9.0	5.8779	0.8856	.	Convert to enum

Previous 20 Columns   Next 20 Columns

Ready   Connections: 0   H2O

H2O Flow

localhost:54321/flow/index.html

H2O FLOW Flow▾ Cell▾ Data▾ Model▾ Score▾ Admin▾ Help▾

Untitled Flow

```
splitFrame "winequality_white.hex", [0.75], ["frame_0.750","frame_0.250"], 705349
```

107ms

Split Frames

Type	Key	Ratio
	frame_0.750	0.75
	frame_0.250	0.25

CS

```
assist buildModel, null, training_frame: "winequality_white.hex"
```

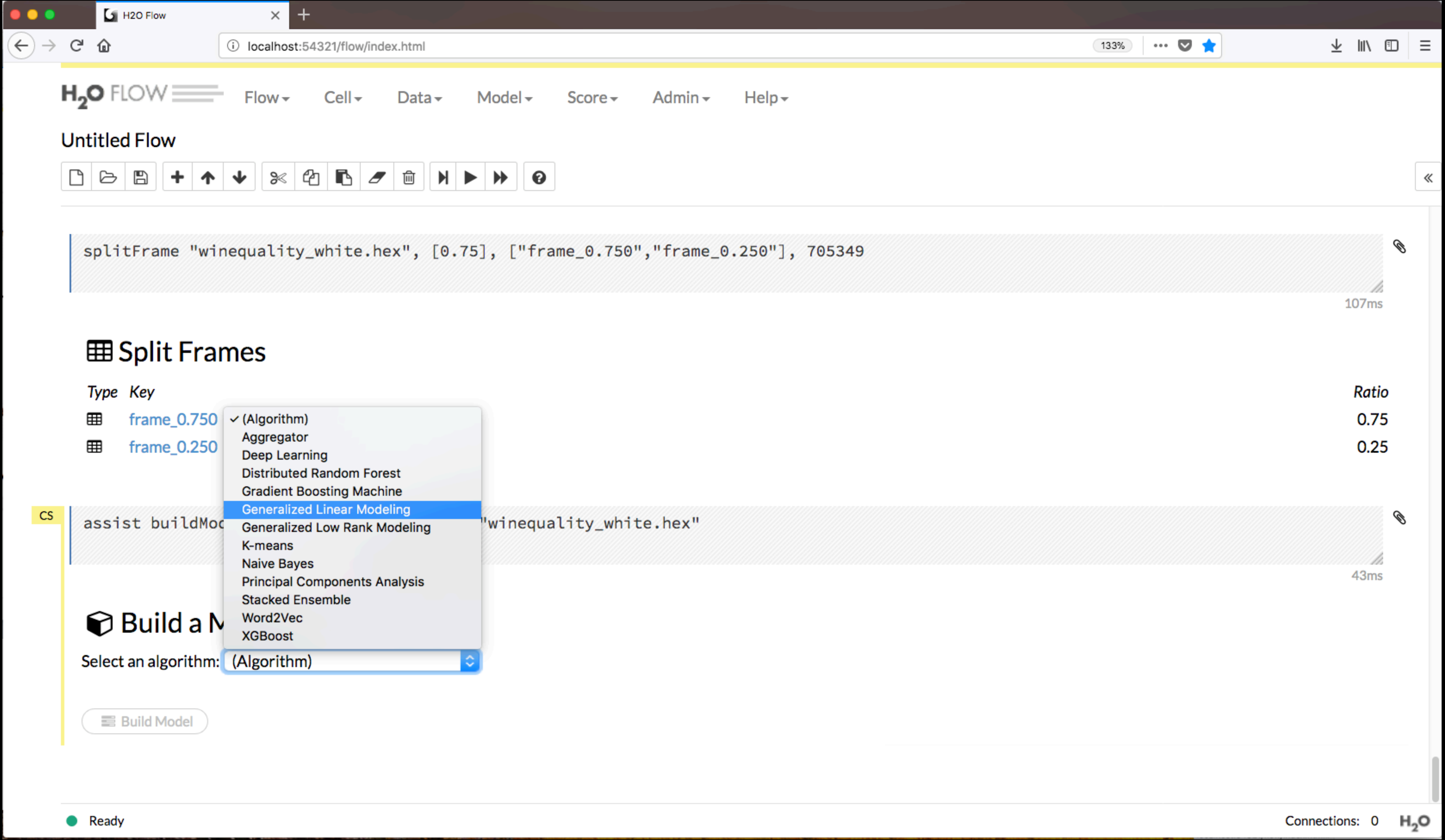
43ms

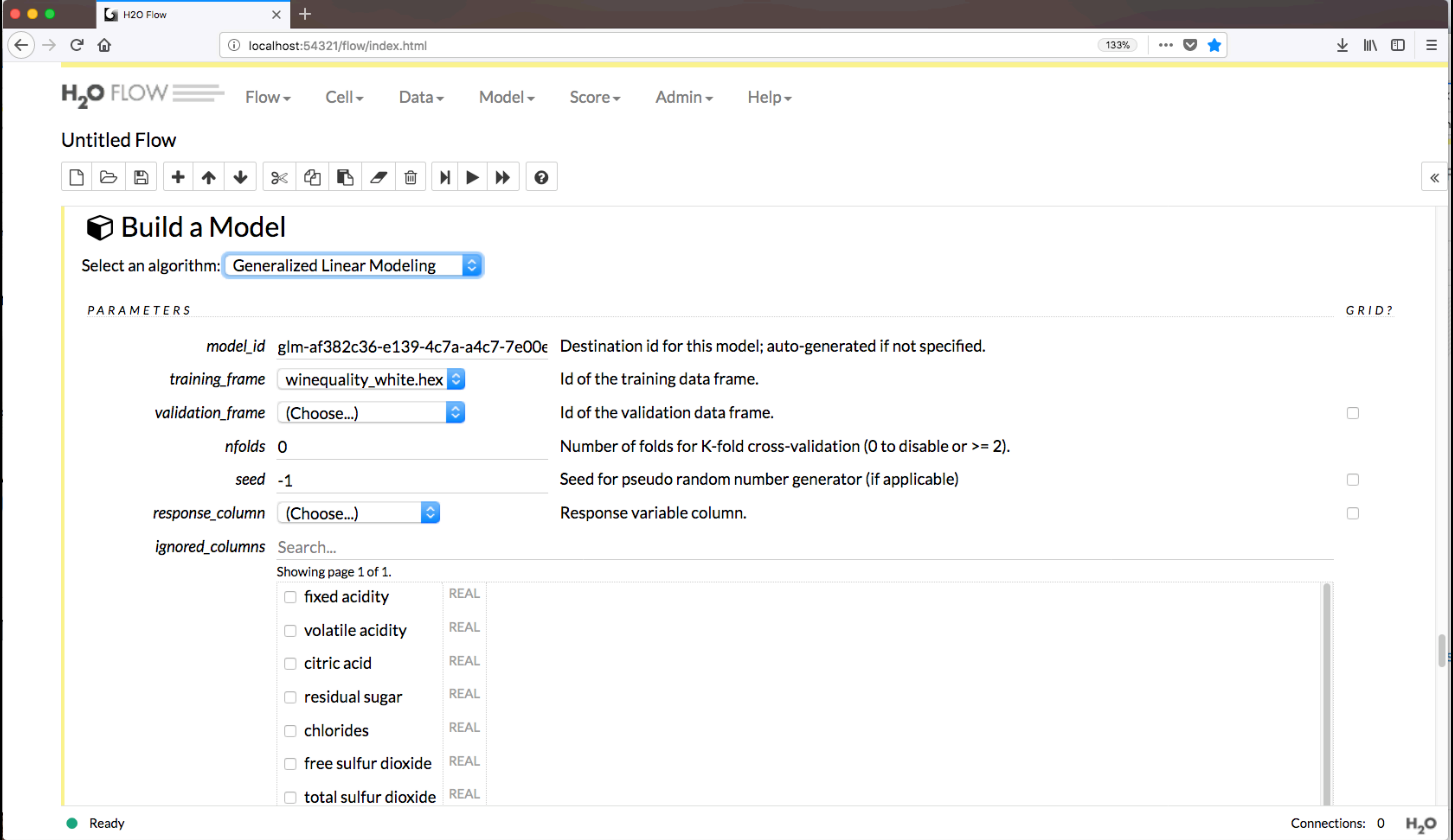
Build a Model

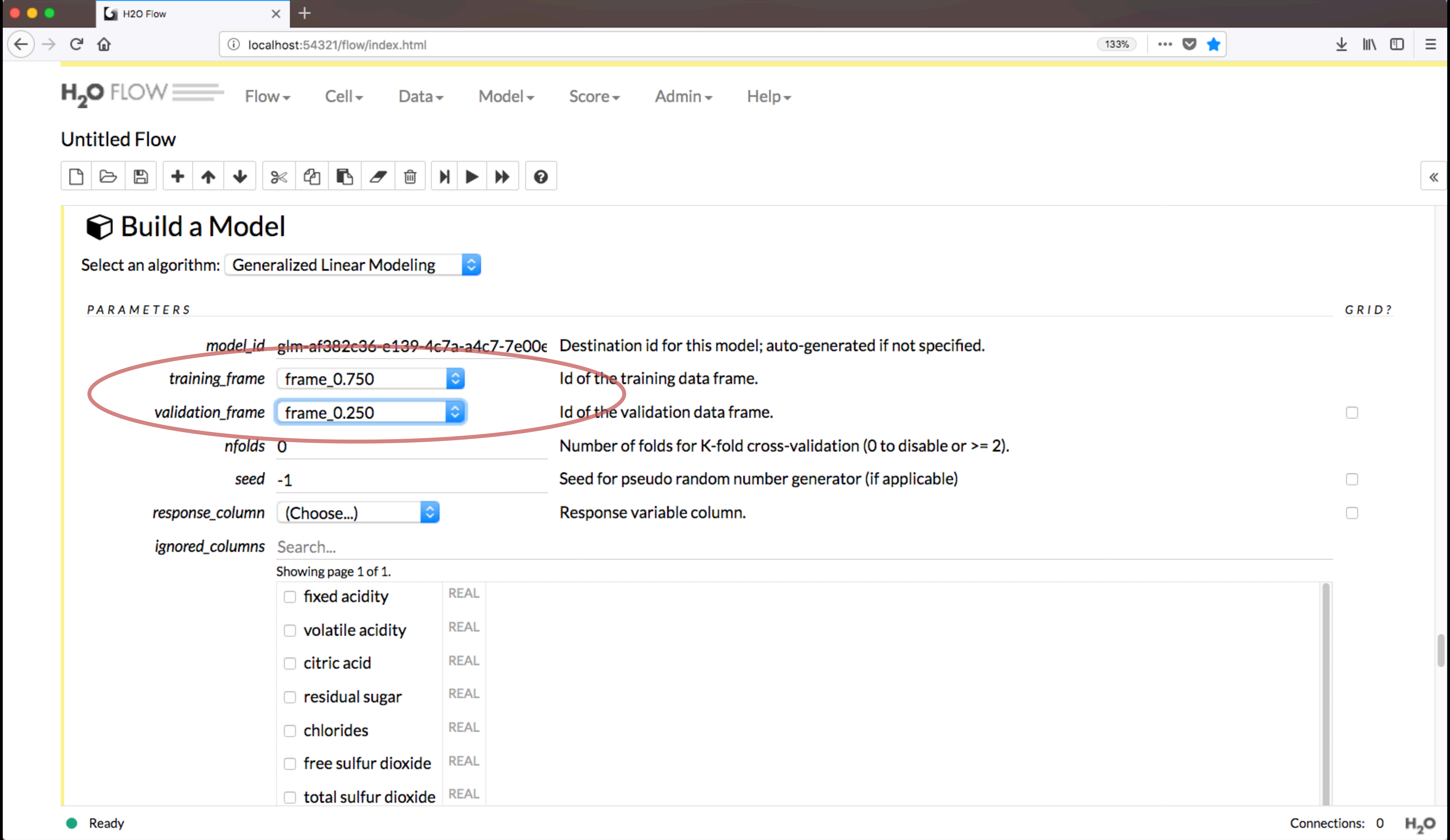
Select an algorithm: (Algorithm)

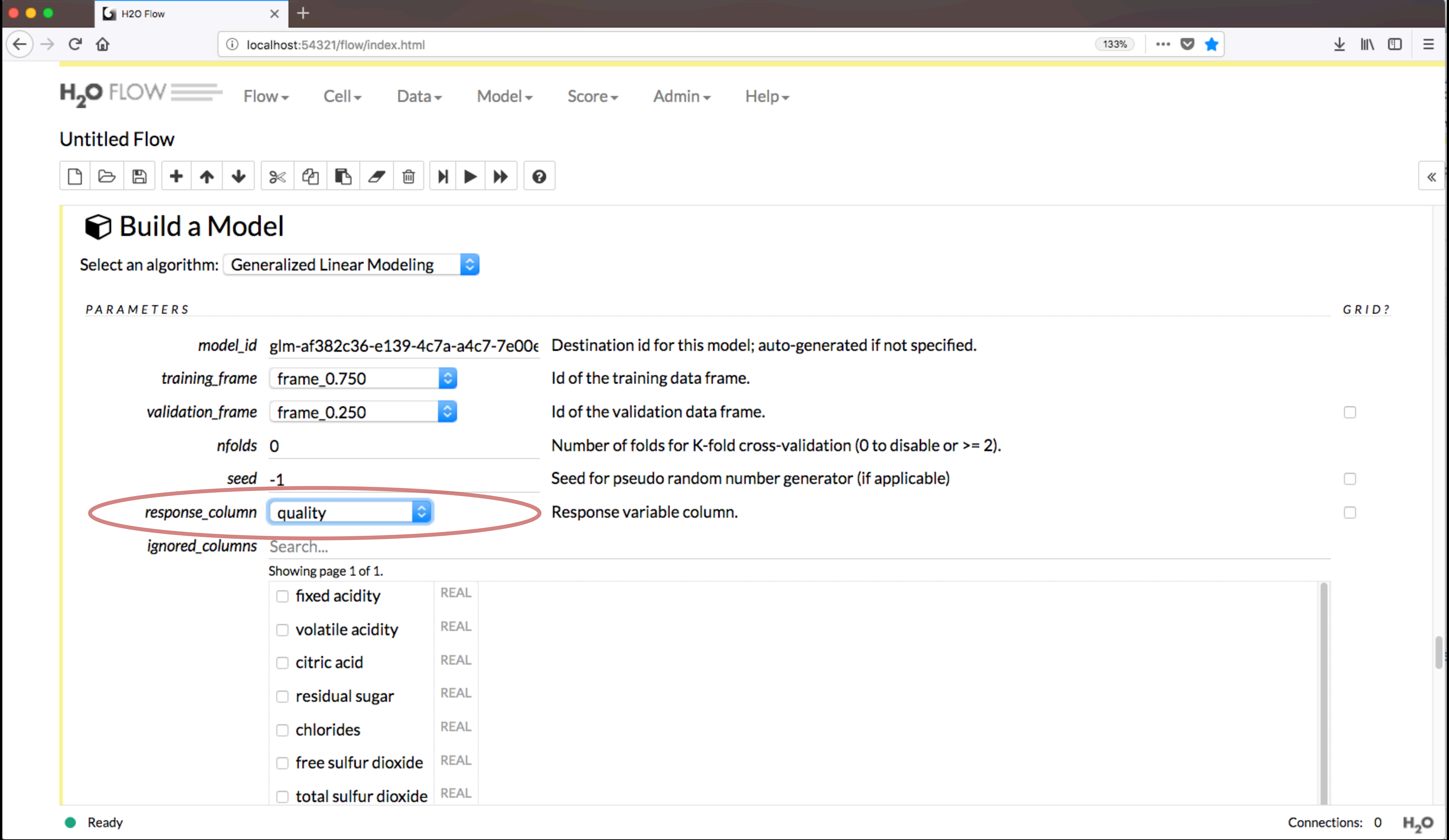
Build Model

Ready Connections: 0 H2O


H2O Flow
localhost:54321/flow/index.html
133%
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
splitFrame "winequality_white.hex", [0.75], ["frame_0.750","frame_0.250"], 705349
107ms
Split Frames
Type
Key
Ratio
frame_0.750
0.75
frame_0.250
0.25
(Algorithm)
Aggregator
Deep Learning
Distributed Random Forest
Gradient Boosting Machine
Generalized Linear Modeling
Generalized Low Rank Modeling
K-means
Naive Bayes
Principal Components Analysis
Stacked Ensemble
Word2Vec
XGBoost
CS
assist buildMod
"winequality_white.hex"
43ms
Build a M
Select an algorithm:
(Algorithm)
Build Model
Ready
Connections: 0
H2O


H2O Flow
localhost:54321/flow/index.html
133%
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
Build a Model
Select an algorithm: Generalized Linear Modeling
PARAMETERS
GRID?
model_id glm-af382c36-e139-4c7a-a4c7-7e00e
Destination id for this model; auto-generated if not specified.
training_frame winequality_white.hex
Id of the training data frame.
validation_frame (Choose...)
Id of the validation data frame.
nfolds 0
Number of folds for K-fold cross-validation (0 to disable or >= 2).
seed -1
Seed for pseudo random number generator (if applicable)
response_column (Choose...)
Response variable column.
ignored_columns Search...
Showing page 1 of 1.
fixed acidity REAL
volatile acidity REAL
citric acid REAL
residual sugar REAL
chlorides REAL
free sulfur dioxide REAL
total sulfur dioxide REAL
Ready
Connections: 0


H2O Flow
localhost:54321/flow/index.html
133%
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
Build a Model
Select an algorithm: Generalized Linear Modeling
PARAMETERS
GRID?
model_id glm-af382c36-e139-4c7a-a4c7-7e00e
Destination id for this model; auto-generated if not specified.
training_frame frame_0.750
Id of the training data frame.
validation_frame frame_0.250
Id of the validation data frame.
nfolds 0
Number of folds for K-fold cross-validation (0 to disable or >= 2).
seed -1
Seed for pseudo random number generator (if applicable)
response_column (Choose...)
Response variable column.
ignored_columns Search...
Showing page 1 of 1.
fixed acidity REAL
volatile acidity REAL
citric acid REAL
residual sugar REAL
chlorides REAL
free sulfur dioxide REAL
total sulfur dioxide REAL
Ready
Connections: 0
H2O


H2O Flow
localhost:54321/flow/index.html
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
Build a Model
Select an algorithm: Generalized Linear Modeling
PARAMETERS
GRID?
model_id glm-af382c36-e139-4c7a-a4c7-7e00e
Destination id for this model; auto-generated if not specified.
training_frame frame_0.750
Id of the training data frame.
validation_frame frame_0.250
Id of the validation data frame.
nfolds 0
Number of folds for K-fold cross-validation (0 to disable or >= 2).
seed -1
Seed for pseudo random number generator (if applicable)
response_column quality
Response variable column.
ignored_columns Search...
Showing page 1 of 1.
fixed acidity REAL
volatile acidity REAL
citric acid REAL
residual sugar REAL
chlorides REAL
free sulfur dioxide REAL
total sulfur dioxide REAL
Ready
Connections: 0

H2O Flow

localhost:54321/flow/index.html

H2O FLOW — Flow ▾ Cell ▾ Data ▾ Model ▾ Score ▾ Admin ▾ Help ▾

Untitled Flow

All | None — Previous 100 | Next 100

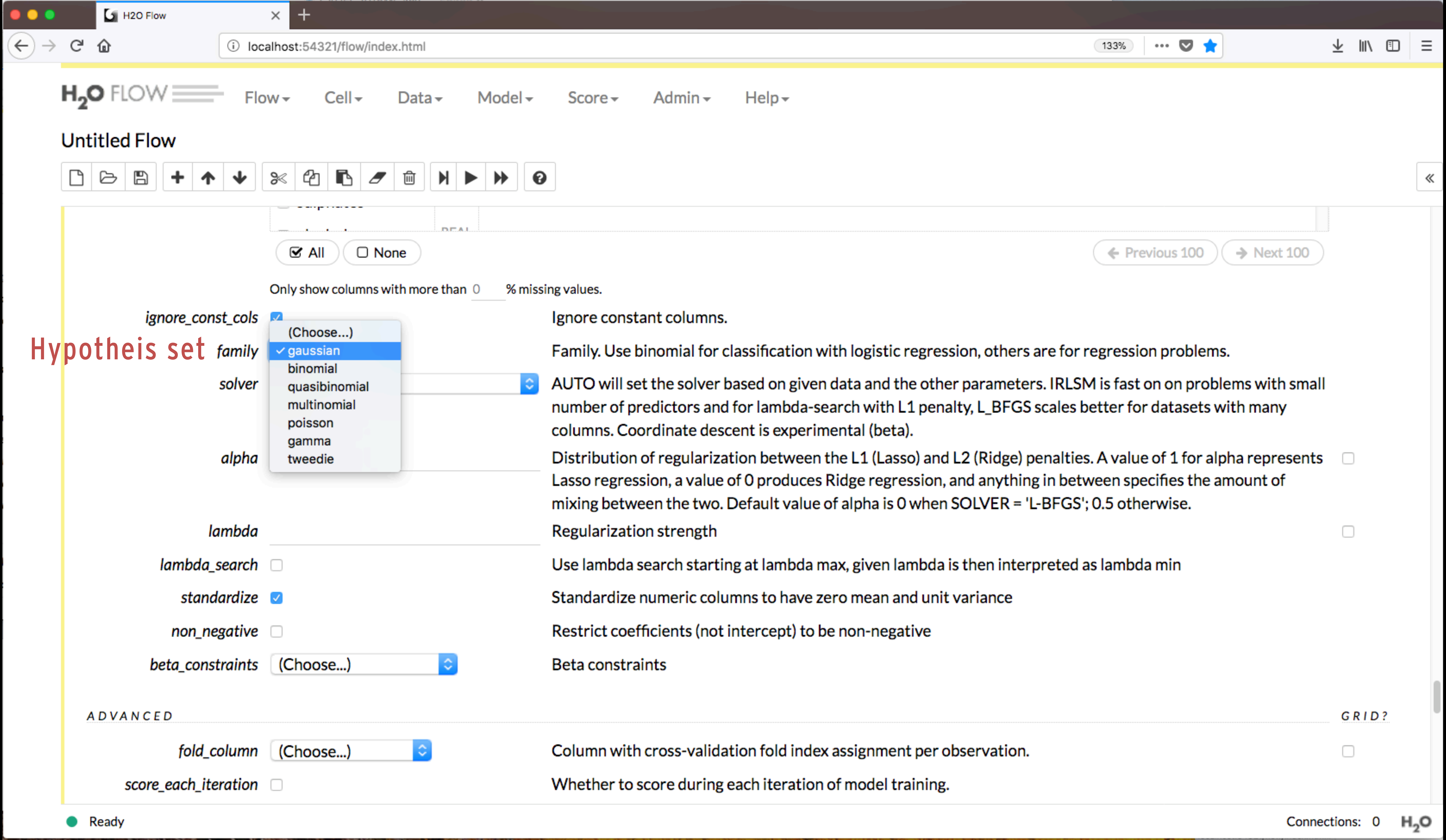
Only show columns with more than 0 % missing values.

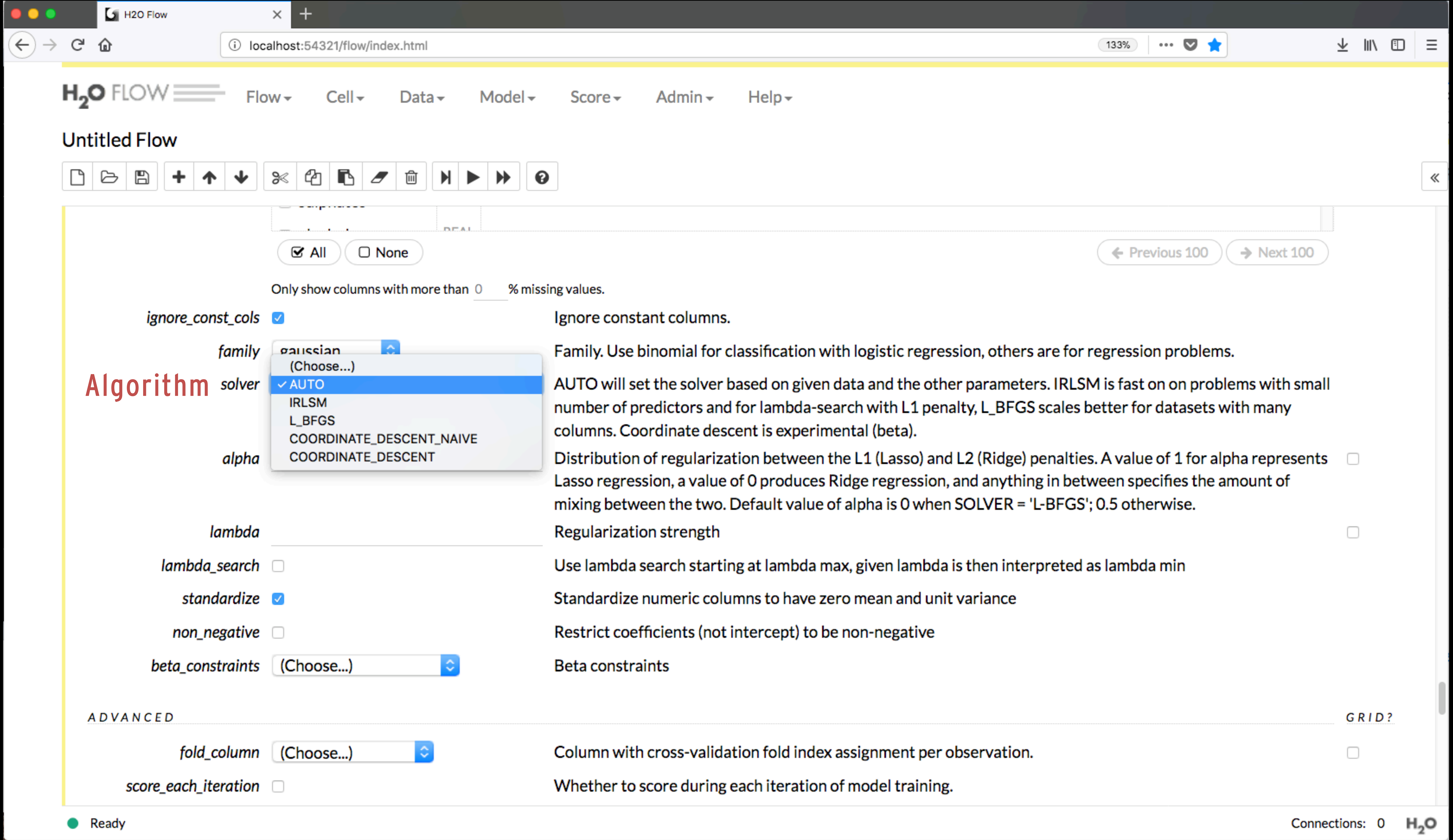
ignore_const_cols	☑	Ignore constant columns.
family	gaussian	Family. Use binomial for classification with logistic regression, others are for regression problems.
solver	AUTO	AUTO will set the solver based on given data and the other parameters. IRLSM is fast on on problems with small number of predictors and for lambda-search with L1 penalty, L_BFGS scales better for datasets with many columns. Coordinate descent is experimental (beta).
alpha		Distribution of regularization between the L1 (Lasso) and L2 (Ridge) penalties. A value of 1 for alpha represents Lasso regression, a value of 0 produces Ridge regression, and anything in between specifies the amount of mixing between the two. Default value of alpha is 0 when SOLVER = 'L-BFGS'; 0.5 otherwise.
lambda		Regularization strength
lambda_search	☐	Use lambda search starting at lambda max, given lambda is then interpreted as lambda min
standardize	☑	Standardize numeric columns to have zero mean and unit variance
non_negative	☐	Restrict coefficients (not intercept) to be non-negative
beta_constraints	(Choose...)	Beta constraints

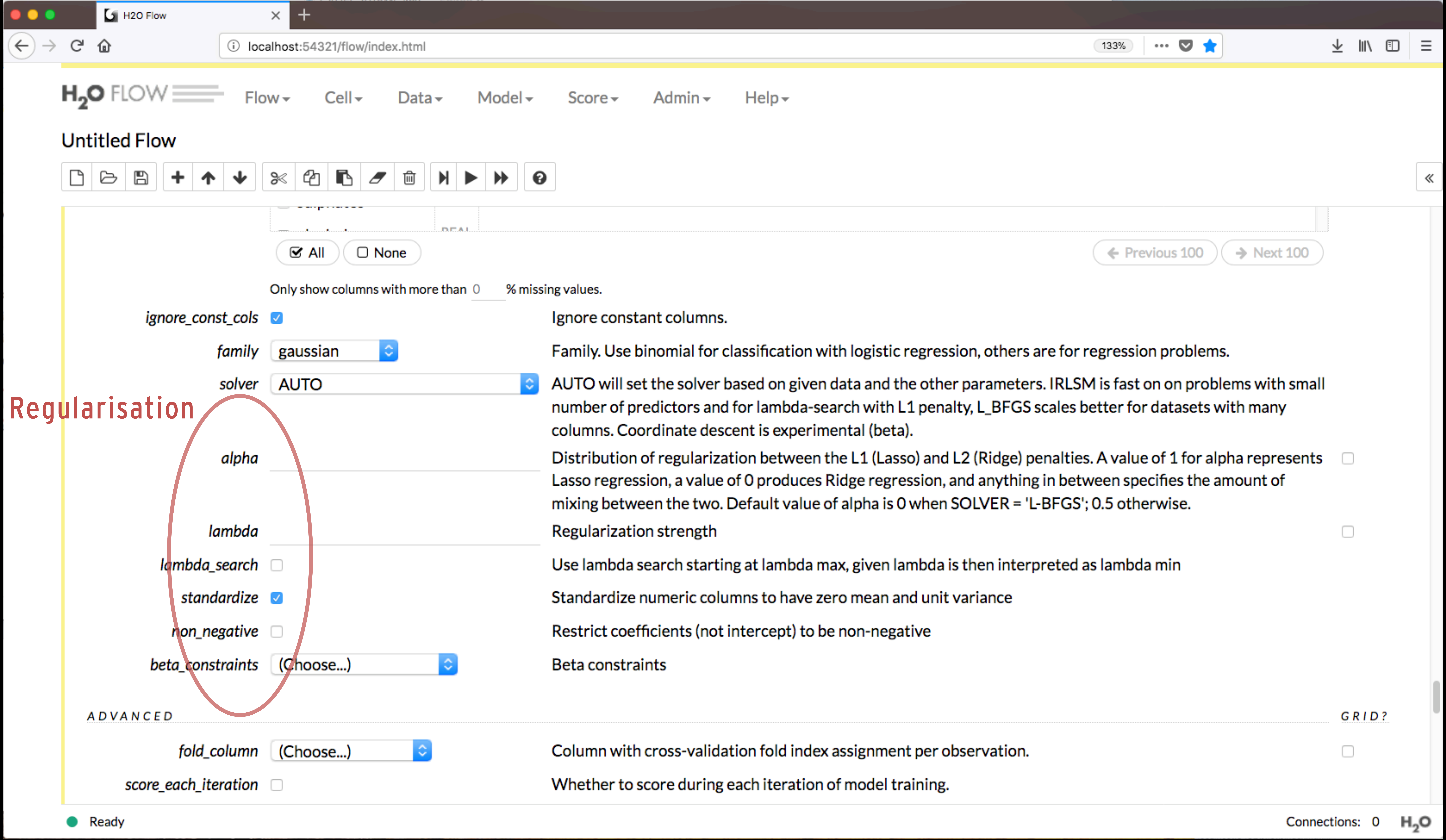
ADVANCED — GRID?

fold_column	(Choose...)	Column with cross-validation fold index assignment per observation.
score_each_iteration	☐	Whether to score during each iteration of model training.

Ready — Connections: 0


H2O Flow
localhost:54321/flow/index.html
133%
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
All
None
Previous 100
Next 100
Only show columns with more than 0 % missing values.
ignore_const_cols
Ignore constant columns.
Hypotheis set
family
(Choose...)
gaussian
binomial
quasibinomial
multinomial
poisson
gamma
tweedie
Family. Use binomial for classification with logistic regression, others are for regression problems.
solver
AUTO will set the solver based on given data and the other parameters. IRLSM is fast on on problems with small number of predictors and for lambda-search with L1 penalty, L_BFGS scales better for datasets with many columns. Coordinate descent is experimental (beta).
alpha
Distribution of regularization between the L1 (Lasso) and L2 (Ridge) penalties. A value of 1 for alpha represents Lasso regression, a value of 0 produces Ridge regression, and anything in between specifies the amount of mixing between the two. Default value of alpha is 0 when SOLVER = 'L-BFGS'; 0.5 otherwise.
lambda
Regularization strength
lambda_search
Use lambda search starting at lambda max, given lambda is then interpreted as lambda min
standardize
Standardize numeric columns to have zero mean and unit variance
non_negative
Restrict coefficients (not intercept) to be non-negative
beta_constraints
(Choose...)
Beta constraints
ADVANCED
GRID?
fold_column
(Choose...)
Column with cross-validation fold index assignment per observation.
score_each_iteration
Whether to score during each iteration of model training.
Ready
Connections: 0
H2O


H2O Flow
localhost:54321/flow/index.html
133%
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
All
None
Previous 100
Next 100
Only show columns with more than 0 % missing values.
ignore_const_cols
Ignore constant columns.
family
gaussian
Family. Use binomial for classification with logistic regression, others are for regression problems.
Algorithm
solver
(Choose...)
AUTO
IRLSM
L_BFGS
COORDINATE_DESCENT_NAIVE
COORDINATE_DESCENT
AUTO will set the solver based on given data and the other parameters. IRLSM is fast on on problems with small number of predictors and for lambda-search with L1 penalty, L_BFGS scales better for datasets with many columns. Coordinate descent is experimental (beta).
alpha
Distribution of regularization between the L1 (Lasso) and L2 (Ridge) penalties. A value of 1 for alpha represents Lasso regression, a value of 0 produces Ridge regression, and anything in between specifies the amount of mixing between the two. Default value of alpha is 0 when SOLVER = 'L-BFGS'; 0.5 otherwise.
lambda
Regularization strength
lambda_search
Use lambda search starting at lambda max, given lambda is then interpreted as lambda min
standardize
Standardize numeric columns to have zero mean and unit variance
non_negative
Restrict coefficients (not intercept) to be non-negative
beta_constraints
(Choose...)
Beta constraints
ADVANCED
GRID?
fold_column
(Choose...)
Column with cross-validation fold index assignment per observation.
score_each_iteration
Whether to score during each iteration of model training.
Ready
Connections: 0
H2O


H2O Flow
localhost:54321/flow/index.html
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
All
None
Previous 100
Next 100
Only show columns with more than 0 % missing values.
ignore_const_cols
Ignore constant columns.
family
gaussian
Family. Use binomial for classification with logistic regression, others are for regression problems.
solver
AUTO
AUTO will set the solver based on given data and the other parameters. IRLSM is fast on on problems with small number of predictors and for lambda-search with L1 penalty, L_BFGS scales better for datasets with many columns. Coordinate descent is experimental (beta).
Regularisation
alpha
Distribution of regularization between the L1 (Lasso) and L2 (Ridge) penalties. A value of 1 for alpha represents Lasso regression, a value of 0 produces Ridge regression, and anything in between specifies the amount of mixing between the two. Default value of alpha is 0 when SOLVER = 'L-BFGS'; 0.5 otherwise.
lambda
Regularization strength
lambda_search
Use lambda search starting at lambda max, given lambda is then interpreted as lambda min
standardize
Standardize numeric columns to have zero mean and unit variance
non_negative
Restrict coefficients (not intercept) to be non-negative
beta_constraints
(Choose...)
Beta constraints
ADVANCED
GRID?
fold_column
(Choose...)
Column with cross-validation fold index assignment per observation.
score_each_iteration
Whether to score during each iteration of model training.
Ready
Connections: 0

THE PROCESS

%

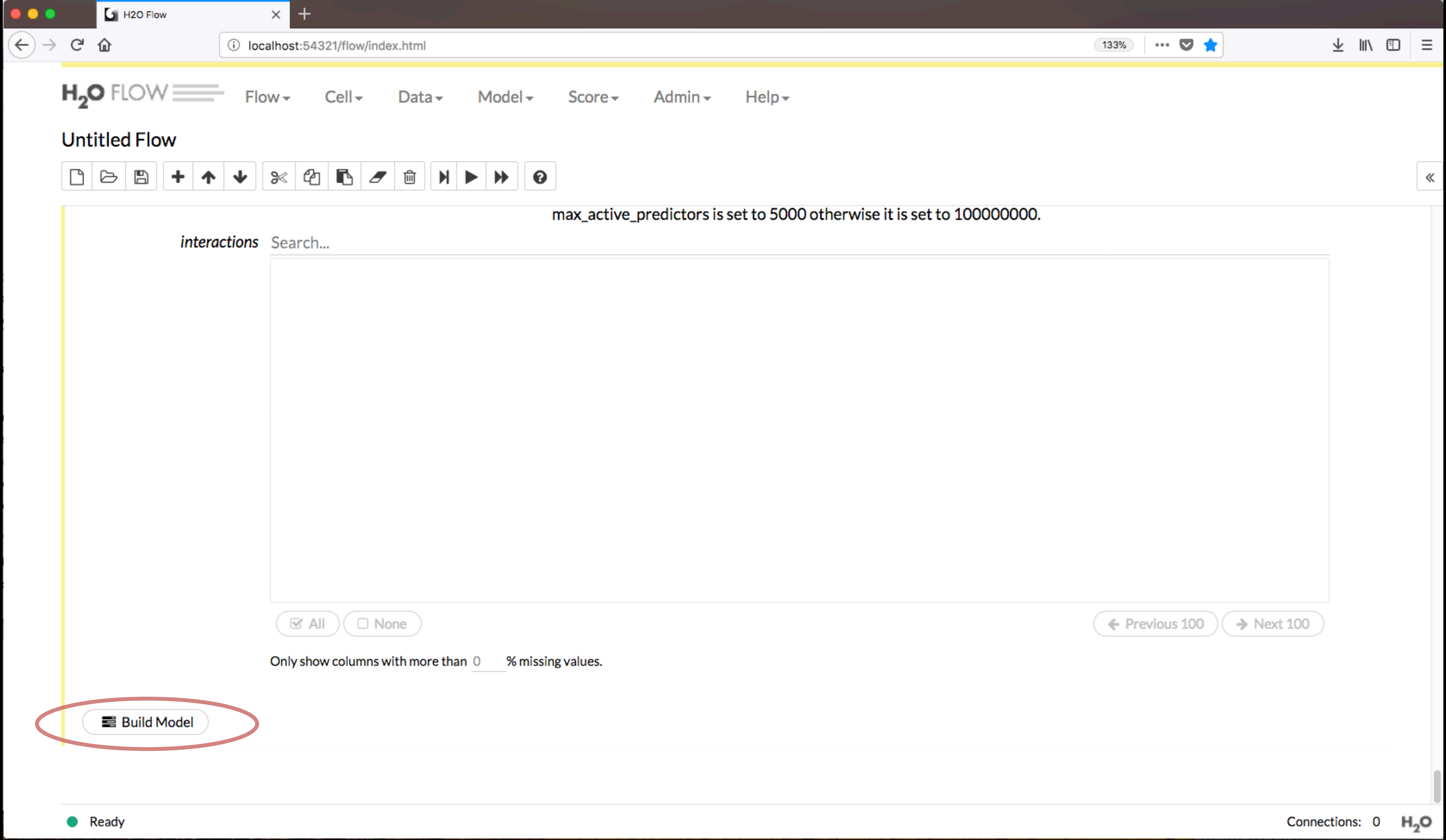
BUSINESS TARGET

AQUIRE RAW DATA

PRE PROCESS

SELECT MODEL

TRAIN


H2O Flow
localhost:54321/flow/index.html
133%
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
max_active_predictors is set to 5000 otherwise it is set to 100000000.
interactions
Search...
All
None
Previous 100
Next 100
Only show columns with more than 0 % missing values.
Build Model
Ready
Connections: 0
H2O

H2O Flow

localhost:54321/flow/index.html

H2O FLOW Flow Cell Data Model Score Admin Help

Untitled Flow

CS

```
buildModel 'glm', {"model_id":"glm-af382c36-e139-4c7a-a4c7-7e00eb4950a2","training_frame":"frame_0.750","validation_frame":"frame_0.250","nfolds":0,"seed":-1,"response_column":"quality","ignored_columns":[],"ignore_const_cols":true,"family":"gaussian","solver":"AUTO","alpha":[],"lambda":[],"lambda_search":false,"standardize":true,"non_negative":false,"score_each_iteration":false,"compute_p_values":false,"remove_collinear_columns":false,"max_iterations":-1,"link":"family_default","max_runtime_secs":0,"custom_metric_func":"","missing_values_handling":"MeanImputation","intercept":true,"objective_epsilon":-1,"beta_epsilon":0.0001,"gradient_epsilon":-1,"prior":-1,"max_active_predictors":-1}
```

1.1s

Job

Run Time	00:00:00.117
Remaining Time	00:00:00.0
Type	Model
Key	glm-af382c36-e139-4c7a-a4c7-7e00eb4950a2
Description	GLM
Status	DONE
Progress	100% Done.
Actions	View

Ready

Connections: 0 H2O

H2O Flow

localhost:54321/flow/index.html

H2O FLOW Flow Cell Data Model Score Admin Help

Untitled Flow

CS

```
buildModel 'glm', {"model_id":"glm-af382c36-e139-4c7a-a4c7-7e00eb4950a2","training_frame":"frame_0.750","validation_frame":"frame_0.250","nfolds":0,"seed":-1,"response_column":"quality","ignored_columns":[],"ignore_const_cols":true,"family":"gaussian","solver":"AUTO","alpha":[],"lambda":[],"lambda_search":false,"standardize":true,"non_negative":false,"score_each_iteration":false,"compute_p_values":false,"remove_collinear_columns":false,"max_iterations":-1,"link":"family_default","max_runtime_secs":0,"custom_metric_func":"","missing_values_handling":"MeanImputation","intercept":true,"objective_epsilon":-1,"beta_epsilon":0.0001,"gradient_epsilon":-1,"prior":-1,"max_active_predictors":-1}
```

1.1s

Job

Run Time	00:00:00.117
Remaining Time	00:00:00.0
Type	Model
Key	glm-af382c36-e139-4c7a-a4c7-7e00eb4950a2
Description	GLM
Status	DONE
Progress	100% Done.
Actions	View

Ready

Connections: 0 H2O

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA

PRE PROCESS

SELECT MODEL

TRAIN

FINAL HYPOTHESIS

H2O Flow

localhost:54321/flow/index.html

H2O FLOW Flow Cell Data Model Score Admin Help

Untitled Flow

▸ OUTPUT - VALIDATION_METRICS

▾ OUTPUT - COEFFICIENTS (GLM COEFFICIENTS)

names	*coefficients*	*standardized_coefficients*
Intercept	127.0896	5.8763
fixed acidity	0.0645	0.0545
volatile acidity	-1.9234	-0.1932
citric acid	-0.0222	-0.0027
residual sugar	0.0729	0.3684
chlorides	-0.6587	-0.0148
free sulfur dioxide	0.0034	0.0589
total sulfur dioxide	-0.0004	-0.0174
density	-126.9049	-0.3809
pH	0.5911	0.0892
sulphates	0.6208	0.0713
alcohol	0.2229	0.2747

▸ OUTPUT - STANDARDIZED COEFFICIENT MAGNITUDES (STANDARDIZED COEFFICIENT MAGNITUDES)

▾ PREVIEW POJO

</> Preview POJO

Ready Connections: 0 H2O

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA

PRE PROCESS

SELECT MODEL

TRAIN

VALIDATE RESULT

FINAL HYPOTHESIS

H2O Flow

localhost:54321/flow/index.html

H2O FLOW Flow Cell Data Model Score Admin Help

Untitled Flow

OUTPUT - VALIDATION_METRICS

model	glm-dc2c16f4-a4a6-44fd-9d3e-fea14aeb82a9
model_checksum	6460191994285201408
frame	frame_0.250
frame_checksum	5832198383805547584
description	·
model_category	Regression
scoring_time	1516729266917
predictions	·
MSE	0.624444
RMSE	0.790218
nobs	1213
custom_metric_name	·
custom_metric_value	0
r2	0.254020
mean_residual_deviance	0.624444
mae	0.608048
rmsle	0.118016
residual_deviance	757.450856
null_deviance	1015.430875
AIC	2897.151369
null_degrees_of_freedom	1212
residual_degrees_of_freedom	1201

Ready

Connections: 0

THE PROCESS

%

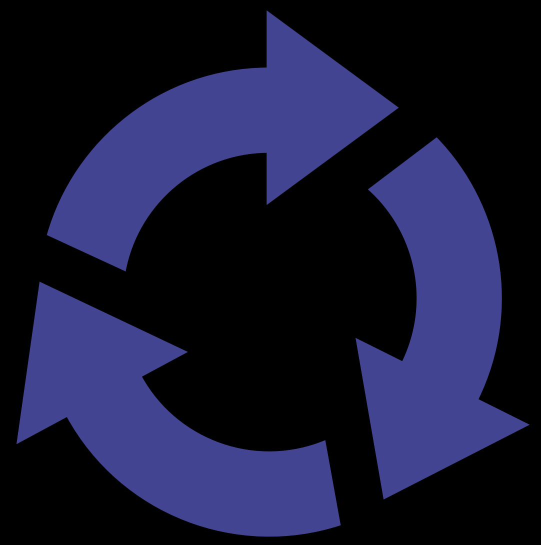
BUSINESS TARGET

AQUIRE RAW DATA

PRE PROCESS

SELECT MODEL

TRIM OR CHANGE MODEL



TRAIN

VALIDATE RESULT

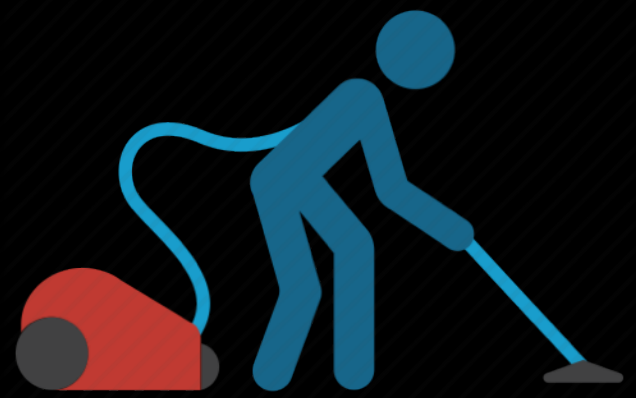
FINAL HYPOTHESIS

THE PROCESS

%

BUSINESS TARGET

AQUIRE RAW DATA



PRE PROCESS

SELECT MODEL

TRAIN

€

FINAL HYPOTHESIS

VALIDATE RESULT

FINAL HYPOTHESIS

H2O Flow

localhost:54321/flow/index.html

H2O FLOW   Flow ▾   Cell ▾   Data ▾   Model ▾   Score ▾   Admin ▾   Help ▾

Untitled Flow

▸ OUTPUT - VALIDATION_METRICS

▾ OUTPUT - COEFFICIENTS (GLM COEFFICIENTS)

names	*coefficients*	*standardized_coefficients*
Intercept	127.0896	5.8763
fixed acidity	0.0645	0.0545
volatile acidity	-1.9234	-0.1932
citric acid	-0.0222	-0.0027
residual sugar	0.0729	0.3684
chlorides	-0.6587	-0.0148
free sulfur dioxide	0.0034	0.0589
total sulfur dioxide	-0.0004	-0.0174
density	-126.9049	-0.3809
pH	0.5911	0.0892
sulphates	0.6208	0.0713
alcohol	0.2229	0.2747

▸ OUTPUT - STANDARDIZED COEFFICIENT MAGNITUDES (STANDARDIZED COEFFICIENT MAGNITUDES)

▾ PREVIEW POJO

</> Preview POJO

Ready   Connections: 0   H2O

H2O Flow

localhost:54321/flow/index.html

H2O FLOW Flow Cell Data Model Score Admin Help

Untitled Flow

```
import java.util.Map;
import hex.genmodel.GenModel;
import hex.genmodel.annotations.ModelPojo;

@ModelPojo(name="glm_dc2c16f4_a4a6_44fd_9d3e_fea14aeb82a9", algorithm="glm")
public class glm_dc2c16f4_a4a6_44fd_9d3e_fea14aeb82a9 extends GenModel {
  public hex.ModelCategory getModelCategory() { return hex.ModelCategory.Regression; }

  public boolean isSupervised() { return true; }
  public int nfeatures() { return 11; }
  public int nclasses() { return 1; }

  // Names of columns used by model.
  public static final String[] NAMES = NamesHolder_glm_dc2c16f4_a4a6_44fd_9d3e_fea14aeb82a9.VALUES;

  // Column domains. The last array contains domain of response column.
  public static final String[][] DOMAINS = new String[][] {
    /* fixed acidity */ null,
    /* volatile acidity */ null,
    /* citric acid */ null,
    /* residual sugar */ null,
    /* chlorides */ null,
    /* free sulfur dioxide */ null,
    /* total sulfur dioxide */ null,
    /* density */ null,
    /* pH */ null,
    /* sulphates */ null,
    /* alcohol */ null,
```

Ready Connections: 0 H2O

THE PROCESS


BUSINESS TARGET
AQUIRE RAW DATA
PRE PROCESS
SELECT MODEL
TRIM OR CHANGE MODEL
TRAIN
IMPLEMENT
FINAL HYPOTHESIS
VALIDATE RESULT
FINAL HYPOTHESIS

FINAL POINT

The tools are here!

Read the theory!

Have fun!

- Big thanks to Yaser Abu-Mostafa of CalTech for the extremely inspiring teaching in the online course Learning From Data (see links on next slide), that has greatly inspired the theory parts of this presentation. Buy the book!

LINKS

- Learning From Data, CalTech Course http://work.caltech.edu/telecourse.html
- Learning From Data, book https://www.amazon.com/gp/product/1600490069
- H2O https://www.h2o.ai/
- UCI ML Data Set repository http://archive.ics.uci.edu/ml/datasets.html
- Apple https://machinelearning.apple.com/
- Kaggle ML community: https://www.kaggle.com/
- Cross Validated https://stats.stackexchange.com