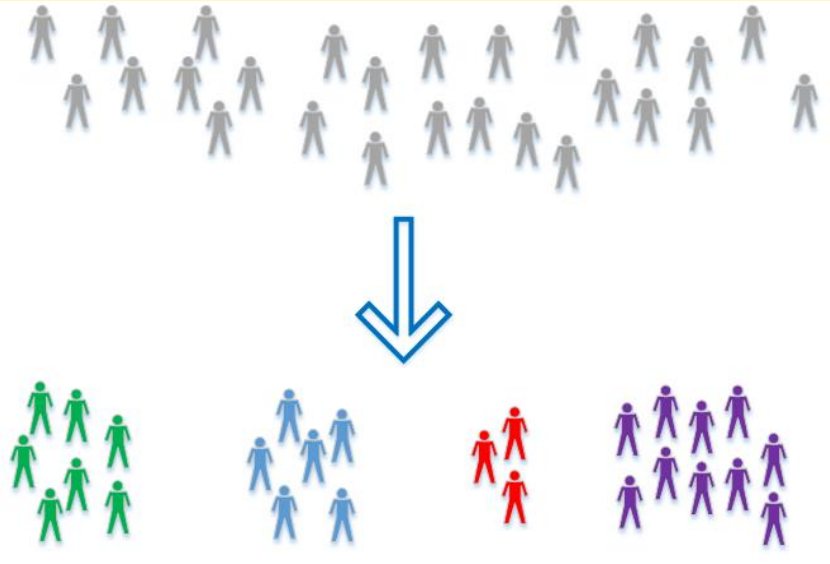
Clustering Example

- **Data**: Water Treatment Plant (1993)
- **Source**: https://archive.ics.uci.edu/ml/datasets/water+treatment+plant

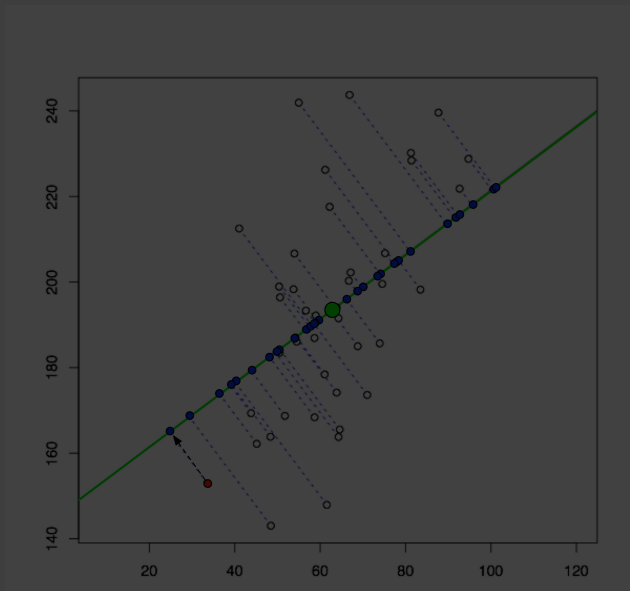
Unsupervised Learning – Discover Hidden Patterns

Clustering:
Customer Segmentation



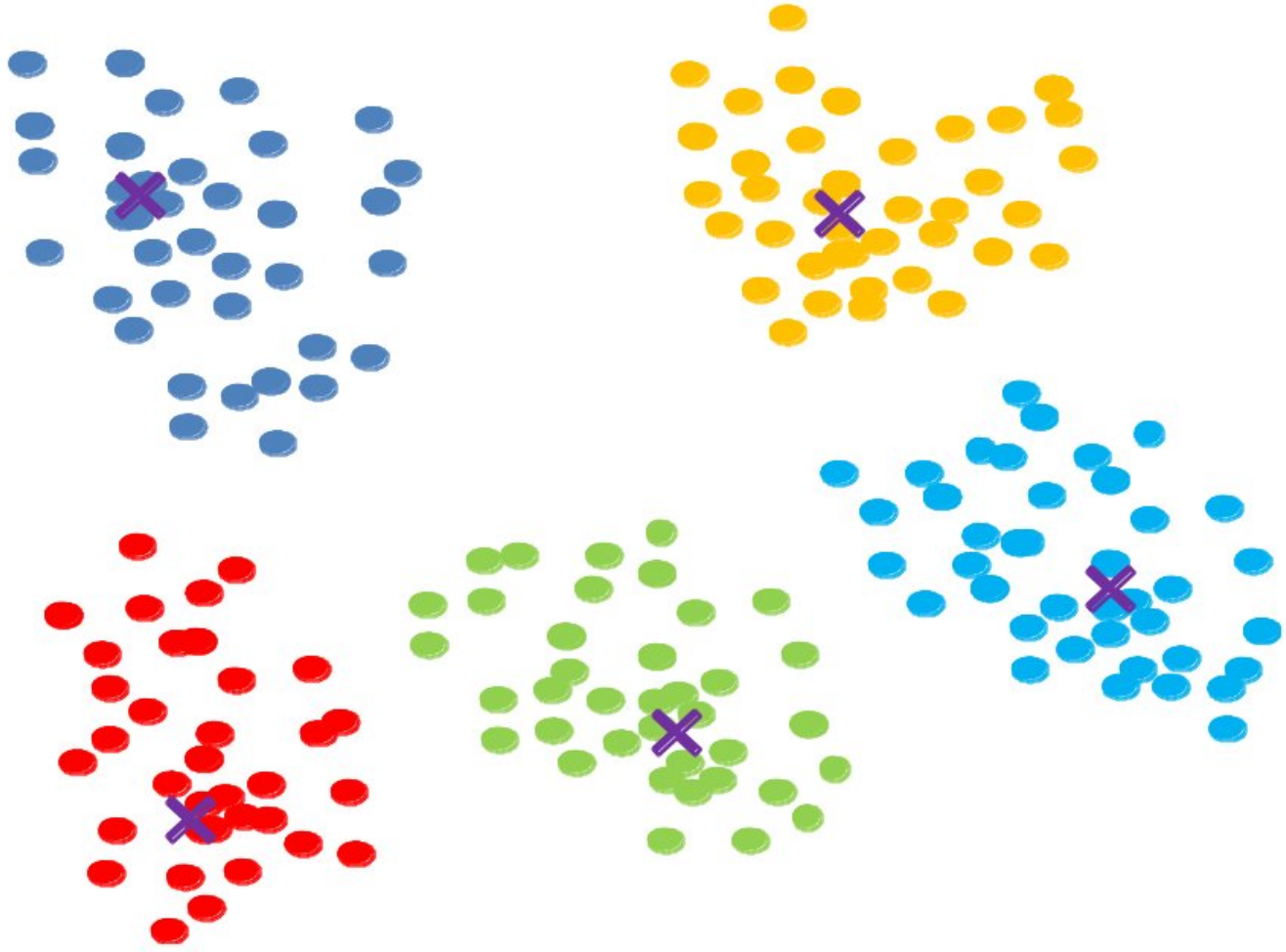
H_2O algos:
K-Means
Generalised Low Rank Model

Dimensionality Reduction:
Linear Transformation of Variables



H_2O algos:
Principal Component Analysis
Generalised Low Rank Model

K-Means Clustering



UCI Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Water Treatment Plant Data Set

Download: Data Folder, Data Set Description

Abstract: Multiple classes predict plant state

Data Set Characteristics:	Multivariate	**Number of Instances:**	527	**Area:**	Physical
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	38	**Date Donated**	1993-06-01
Associated Tasks:	Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	91704

Source:

Creators:

Manel Poch (igte2 '@' cc.uab.es)
Unitat d'Enginyeria Quimica
Universitat Autonoma de Barcelona. Bellaterra. Barcelona; Spain

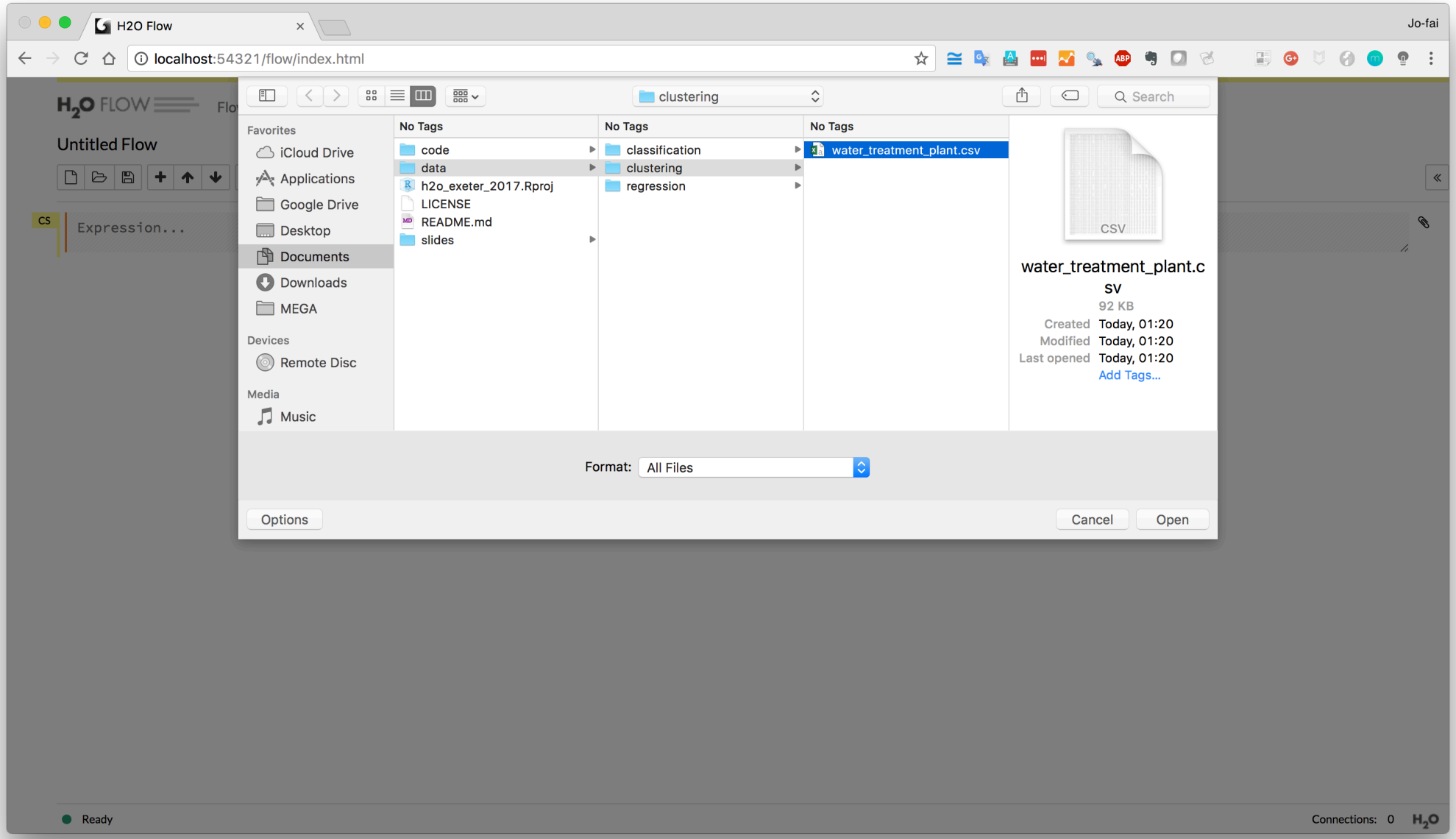
Donor:

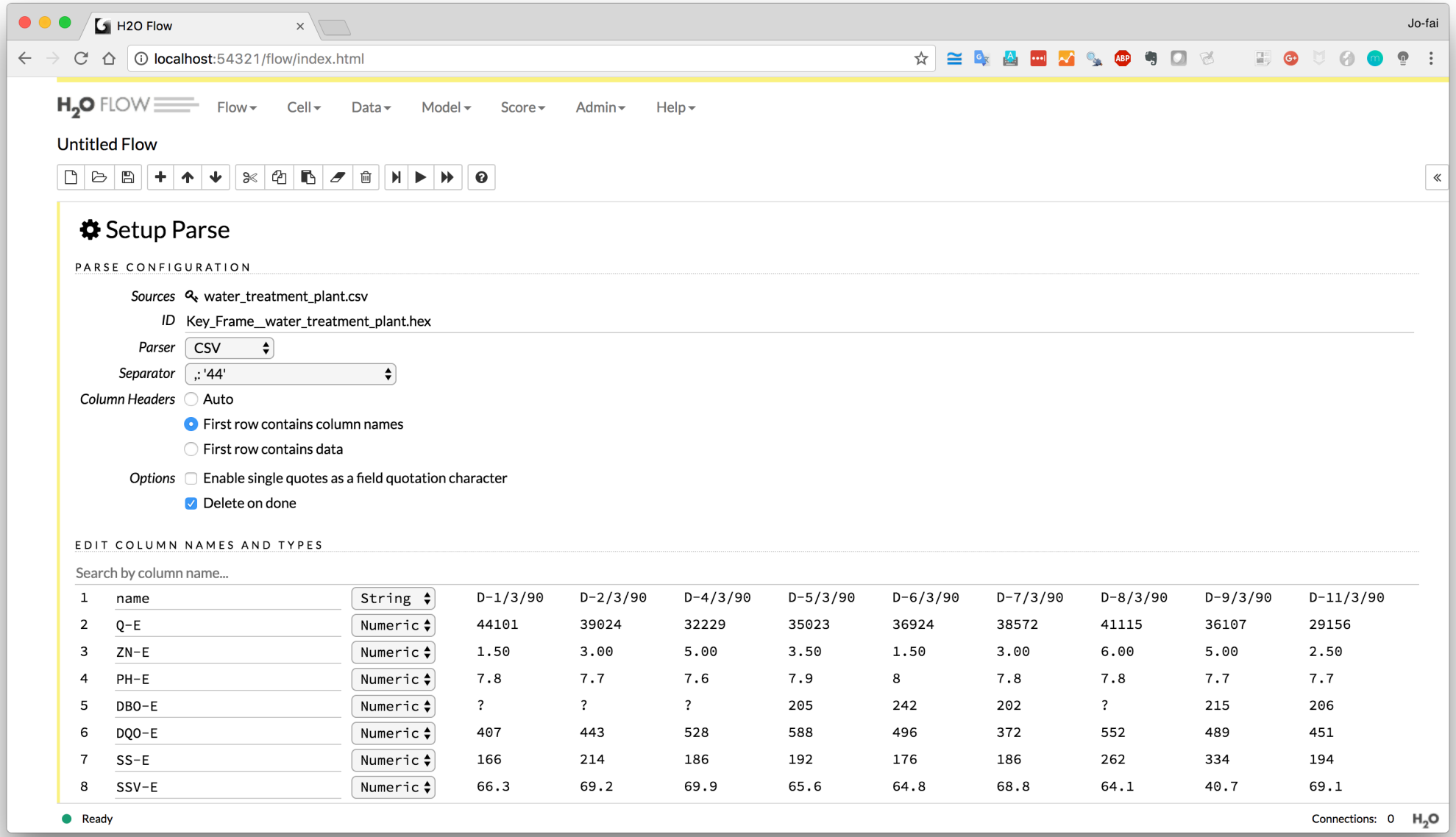
Javier Bejar and Ulises Cortes (bejar '@' lsi.upc.es)
Dept. Llenguatges i Sistemes Informatics;
Universitat Politecnica de Catalunya. Barcelona; Spain

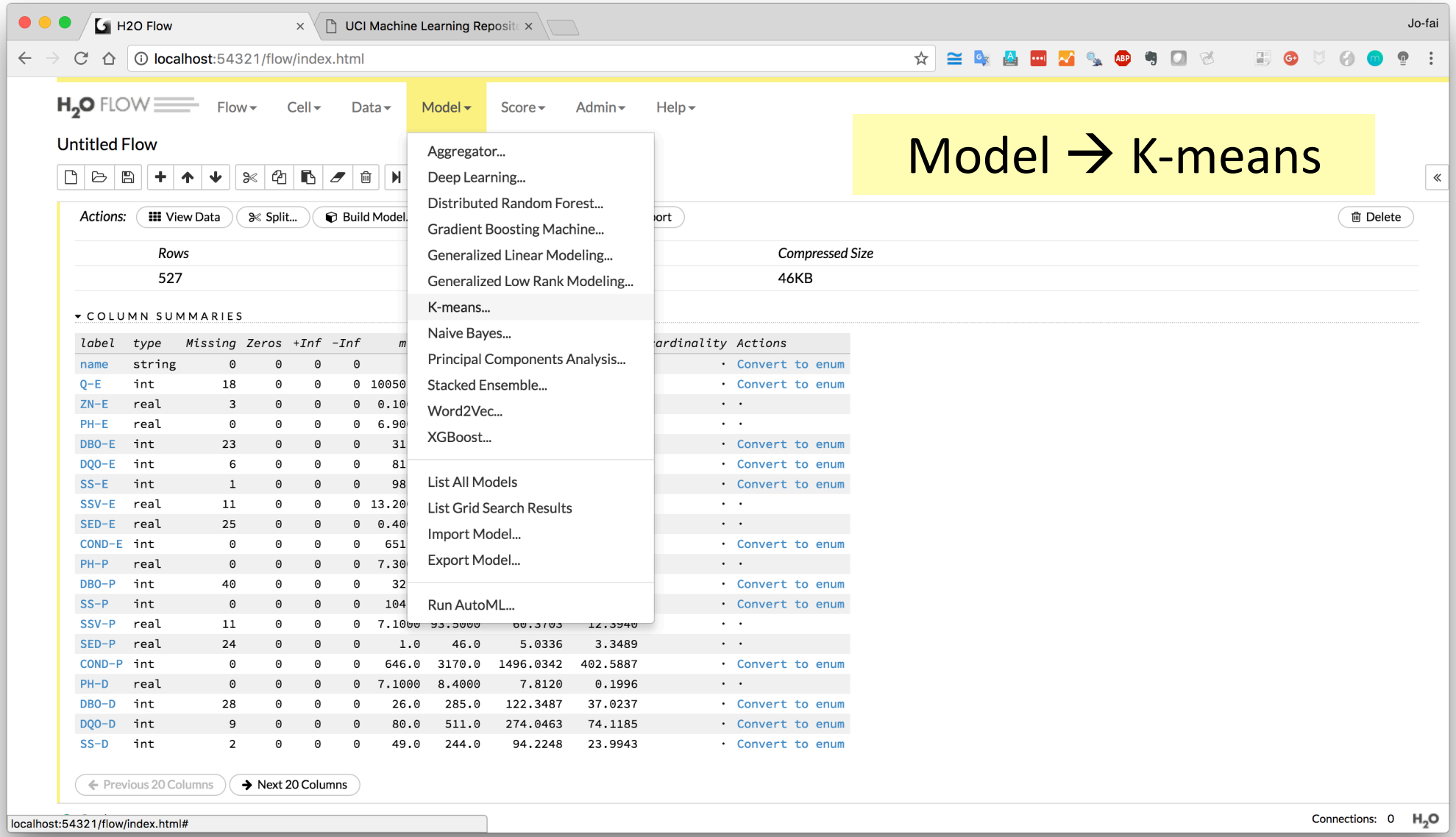
Attribute Information:

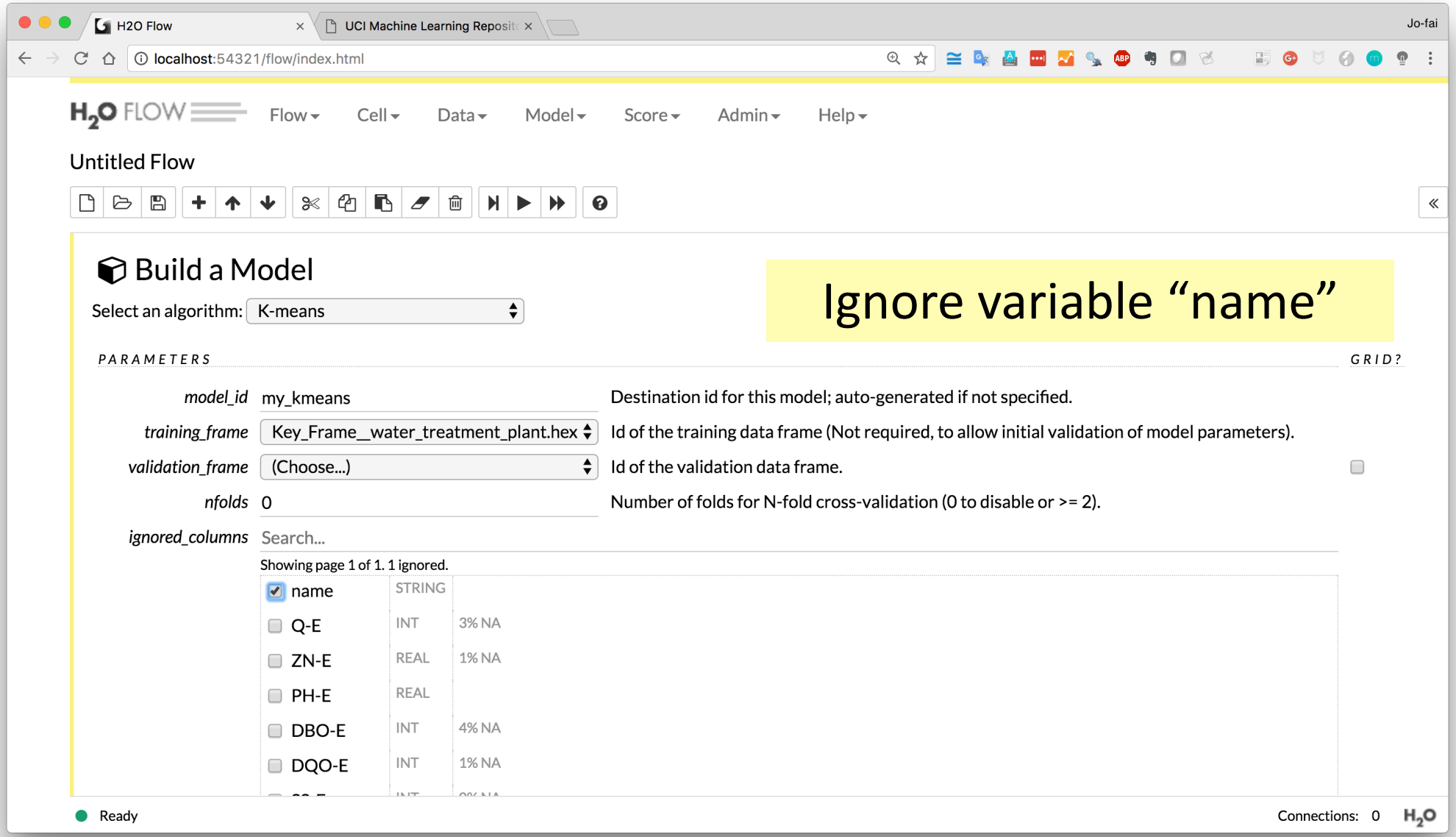
All atrributes are numeric and continuous

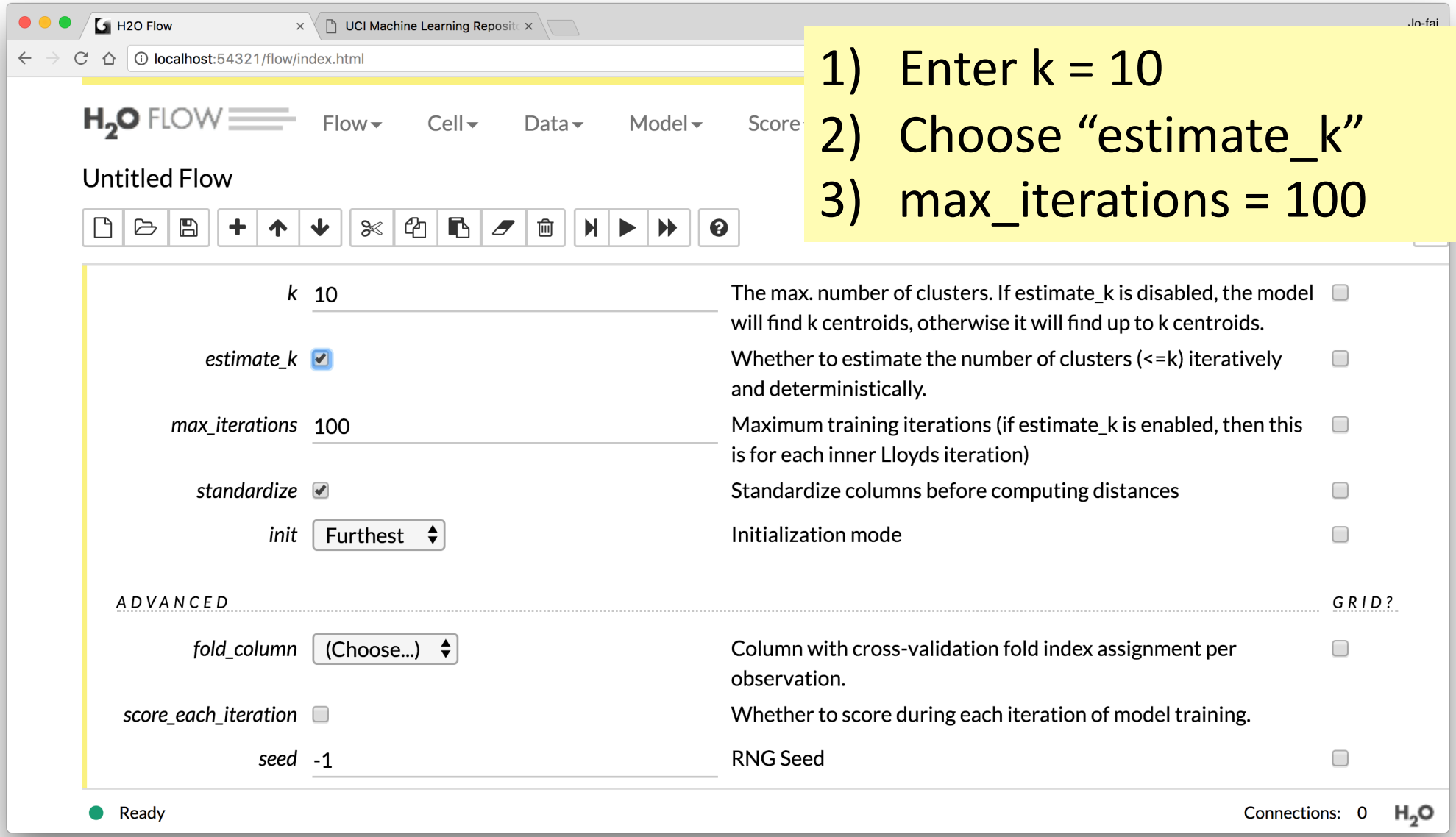
N. Attrib.
1 Q-E (input flow to plant)
2 ZN-E (input Zinc to plant)
3 PH-E (input pH to plant)
4 DBO-E (input Biological demand of oxygen to plant)
5 DQO-E (input chemical demand of oxygen to plant)
6 SS-E (input suspended solids to plant)
7 SSV-E (input volatile supended solids to plant)
8 SED-E (input sediments to plant)
9 COND-E (input conductivity to plant)
10 PH-P (input pH to primary settler)
11 DBO-P (input Biological demand of oxygen to primary settler)
12 SS-P (input suspended solids to primary settler)
13 SSV-P (input volatile supended solids to primary settler)
14 SED-P (input sediments to primary settler)
15 COND-P (input conductivity to primary settler)
16 PH-D (input pH to secondary settler)
17 DBO-D (input Biological demand of oxygen to secondary settler)
18 DQO-D (input chemical demand of oxygen to secondary settler)
19 SS-D (input suspended solids to secondary settler)
20 SSV-D (input volatile supended solids to secondary settler)
21 SED-D (input sediments to secondary settler)
22 COND-D (input conductivity to secondary settler)
23 PH-S (output pH)
24 DBO-S (output Biological demand of oxygen)
25 DQO-S (output chemical demand of oxygen)
26 SS-S (output suspended solids)
27 SSV-S (output volatile supended solids)
28 SED-S (output sediments)
29 COND-S (output conductivity)
30 RD-DBO-P (performance input Biological demand of oxygen in primary settler)
31 RD-SS-P (performance input suspended solids to primary settler)
32 RD-SED-P (performance input sediments to primary settler)
33 RD-DBO-S (performance input Biological demand of oxygen to secondary settler)
34 RD-DQO-S (performance input chemical demand of oxygen to secondary settler)
35 RD-DBO-G (global performance input Biological demand of oxygen)
36 RD-DQO-G (global performance input chemical demand of oxygen)
37 RD-SS-G (global performance input suspended solids)
38 RD-SED-G (global performance input sediments)


H2O Flow
Jo-fai
localhost:54321/flow/index.html
H2O FLOW
Untitled Flow
CS
Expression...
clustering
Search
Favorites
iCloud Drive
Applications
Google Drive
Desktop
Documents
Downloads
MEGA
Devices
Remote Disc
Media
Music
No Tags
code
data
h2o_exeter_2017.Rproj
LICENSE
README.md
slides
No Tags
classification
clustering
regression
No Tags
water_treatment_plant.csv
CSV
water_treatment_plant.csv
92 KB
Created Today, 01:20
Modified Today, 01:20
Last opened Today, 01:20
Add Tags...
Format: All Files
Options
Cancel
Open
Ready
Connections: 0

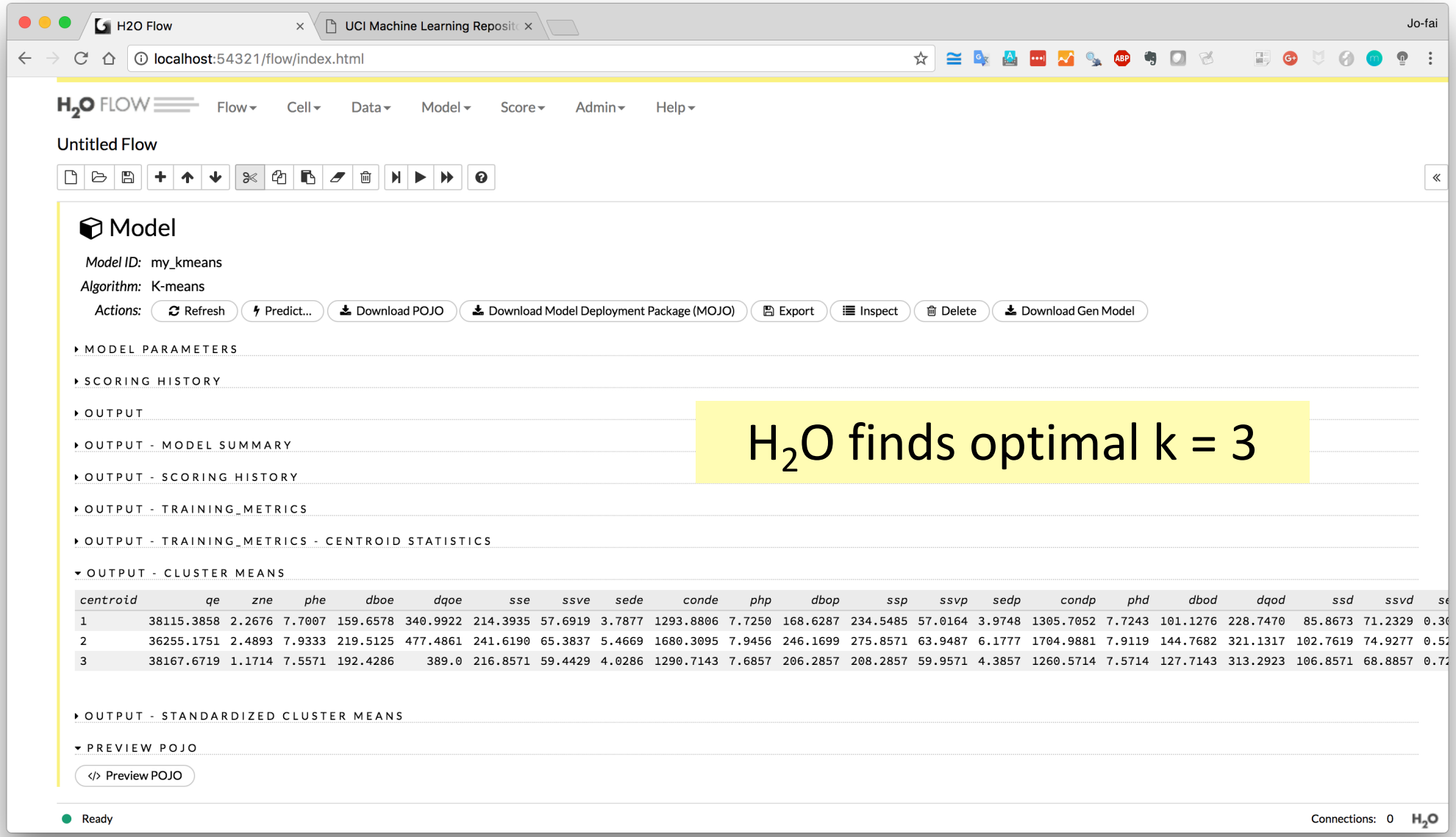

H2O Flow
Jo-fai
localhost:54321/flow/index.html
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
Setup Parse
PARSE CONFIGURATION
Sources water_treatment_plant.csv
ID Key_Frame__water_treatment_plant.hex
Parser CSV
Separator ,: '44'
Column Headers Auto
First row contains column names
First 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 name String D-1/3/90 D-2/3/90 D-4/3/90 D-5/3/90 D-6/3/90 D-7/3/90 D-8/3/90 D-9/3/90 D-11/3/90
2 Q-E Numeric 44101 39024 32229 35023 36924 38572 41115 36107 29156
3 ZN-E Numeric 1.50 3.00 5.00 3.50 1.50 3.00 6.00 5.00 2.50
4 PH-E Numeric 7.8 7.7 7.6 7.9 8 7.8 7.8 7.7 7.7
5 DBO-E Numeric ? ? ? 205 242 202 ? 215 206
6 DQO-E Numeric 407 443 528 588 496 372 552 489 451
7 SS-E Numeric 166 214 186 192 176 186 262 334 194
8 SSV-E Numeric 66.3 69.2 69.9 65.6 64.8 68.8 64.1 40.7 69.1
Ready
Connections: 0
H2O


H2O Flow
UCI Machine Learning Reposit
Jo-fai
localhost:54321/flow/index.html
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
Model → K-means
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...
List All Models
List Grid Search Results
Import Model...
Export Model...
Run AutoML...
Actions:
View Data
Split...
Build Model...
Delete
Rows
527
Compressed Size
46KB
COLUMN SUMMARIES
label type Missing Zeros +Inf -Inf Actions
name string 0 0 0 0 Convert to enum
Q-E int 18 0 0 0 10050 Convert to enum
ZN-E real 3 0 0 0 0.10 · ·
PH-E real 0 0 0 0 6.90 · ·
DBO-E int 23 0 0 0 31 Convert to enum
DQO-E int 6 0 0 0 81 Convert to enum
SS-E int 1 0 0 0 98 Convert to enum
SSV-E real 11 0 0 0 13.20 · ·
SED-E real 25 0 0 0 0.40 · ·
COND-E int 0 0 0 0 651 Convert to enum
PH-P real 0 0 0 0 7.30 · ·
DBO-P int 40 0 0 0 32 Convert to enum
SS-P int 0 0 0 0 104 Convert to enum
SSV-P real 11 0 0 0 7.1000 93.5000 60.3703 12.3940 · ·
SED-P real 24 0 0 0 1.0 46.0 5.0336 3.3489 · ·
COND-P int 0 0 0 0 646.0 3170.0 1496.0342 402.5887 Convert to enum
PH-D real 0 0 0 0 7.1000 8.4000 7.8120 0.1996 · ·
DBO-D int 28 0 0 0 26.0 285.0 122.3487 37.0237 Convert to enum
DQO-D int 9 0 0 0 80.0 511.0 274.0463 74.1185 Convert to enum
SS-D int 2 0 0 0 49.0 244.0 94.2248 23.9943 Convert to enum
Previous 20 Columns
Next 20 Columns
localhost:54321/flow/index.html#
Connections: 0
H2O

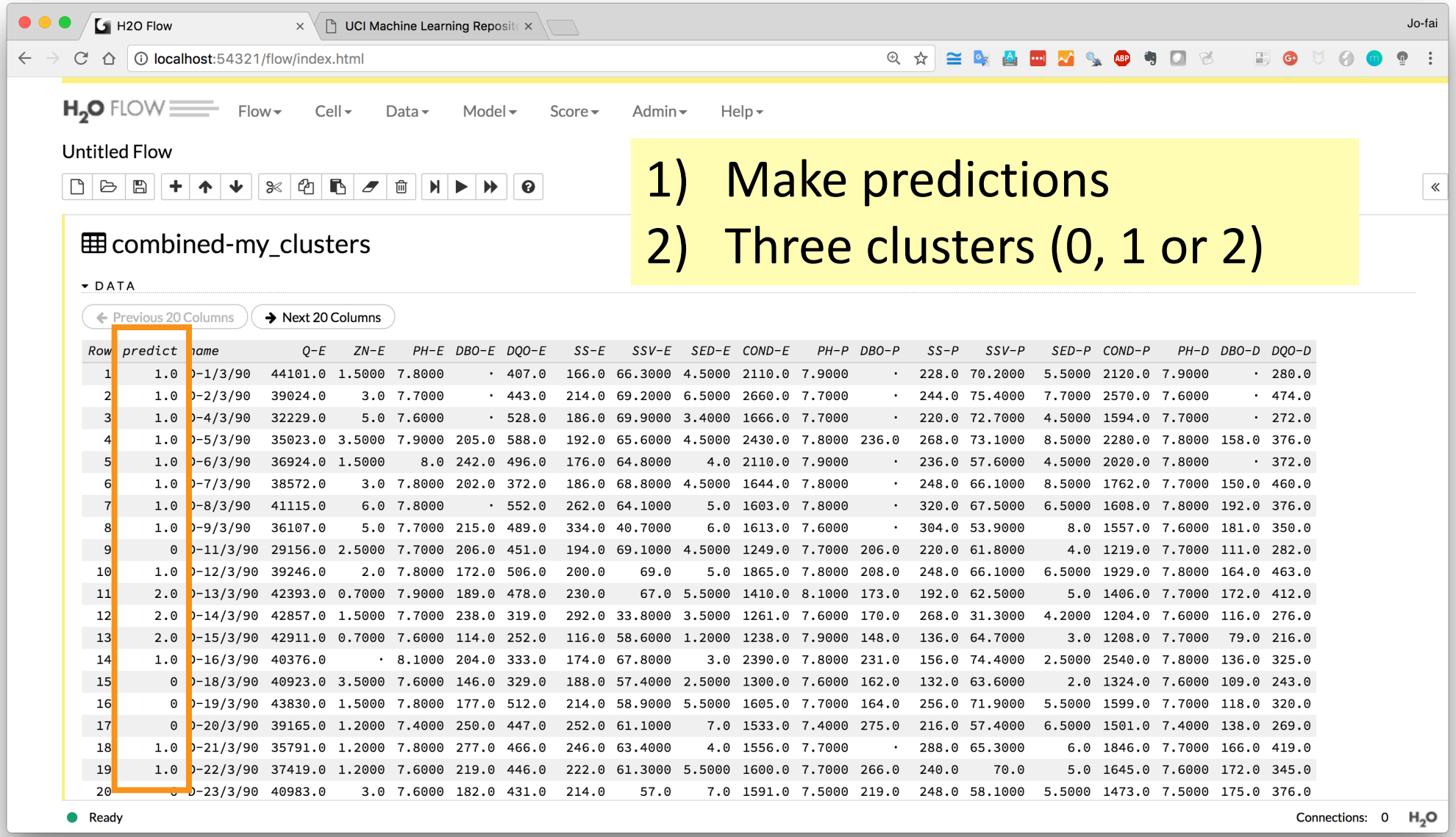

H2O Flow
UCI Machine Learning Reposit
Jo-fai
localhost:54321/flow/index.html
H2O FLOW
Flow
Cell
Data
Model
Score
Admin
Help
Untitled Flow
Build a Model
Select an algorithm: K-means
Ignore variable “name”
PARAMETERS
GRID?
model_id my_kmeans
Destination id for this model; auto-generated if not specified.
training_frame Key_Frame__water_treatment_plant.hex
Id of the training data frame (Not required, to allow initial validation of model parameters).
validation_frame (Choose...)
Id of the validation data frame.
nfolds 0
Number of folds for N-fold cross-validation (0 to disable or >= 2).
ignored_columns Search...
Showing page 1 of 1. 1 ignored.
name STRING
Q-E INT 3% NA
ZN-E REAL 1% NA
PH-E REAL
DBO-E INT 4% NA
DQO-E INT 1% NA
Ready
Connections: 0
H2O



1) Enter k = 10
2) Choose “estimate_k”
3) max_iterations = 100



H_2O finds optimal k = 3



Row	predict	name	Q-E	ZN-E	PH-E	DBO-E	DQO-E	SS-E	SSV-E	SED-E	COND-E	PH-P	DBO-P	SS-P	SSV-P	SED-P	COND-P	PH-D	DBO-D	DQO-D
1	1.0	D-1/3/90	44101.0	1.5000	7.8000	·	407.0	166.0	66.3000	4.5000	2110.0	7.9000	·	228.0	70.2000	5.5000	2120.0	7.9000	·	280.0
2	1.0	D-2/3/90	39024.0	3.0	7.7000	·	443.0	214.0	69.2000	6.5000	2660.0	7.7000	·	244.0	75.4000	7.7000	2570.0	7.6000	·	474.0
3	1.0	D-4/3/90	32229.0	5.0	7.6000	·	528.0	186.0	69.9000	3.4000	1666.0	7.7000	·	220.0	72.7000	4.5000	1594.0	7.7000	·	272.0
4	1.0	D-5/3/90	35023.0	3.5000	7.9000	205.0	588.0	192.0	65.6000	4.5000	2430.0	7.8000	236.0	268.0	73.1000	8.5000	2280.0	7.8000	158.0	376.0
5	1.0	D-6/3/90	36924.0	1.5000	8.0	242.0	496.0	176.0	64.8000	4.0	2110.0	7.9000	·	236.0	57.6000	4.5000	2020.0	7.8000	·	372.0
6	1.0	D-7/3/90	38572.0	3.0	7.8000	202.0	372.0	186.0	68.8000	4.5000	1644.0	7.8000	·	248.0	66.1000	8.5000	1762.0	7.7000	150.0	460.0
7	1.0	D-8/3/90	41115.0	6.0	7.8000	·	552.0	262.0	64.1000	5.0	1603.0	7.8000	·	320.0	67.5000	6.5000	1608.0	7.8000	192.0	376.0
8	1.0	D-9/3/90	36107.0	5.0	7.7000	215.0	489.0	334.0	40.7000	6.0	1613.0	7.6000	·	304.0	53.9000	8.0	1557.0	7.6000	181.0	350.0
9	0	D-11/3/90	29156.0	2.5000	7.7000	206.0	451.0	194.0	69.1000	4.5000	1249.0	7.7000	206.0	220.0	61.8000	4.0	1219.0	7.7000	111.0	282.0
10	1.0	D-12/3/90	39246.0	2.0	7.8000	172.0	506.0	200.0	69.0	5.0	1865.0	7.8000	208.0	248.0	66.1000	6.5000	1929.0	7.8000	164.0	463.0
11	2.0	D-13/3/90	42393.0	0.7000	7.9000	189.0	478.0	230.0	67.0	5.5000	1410.0	8.1000	173.0	192.0	62.5000	5.0	1406.0	7.7000	172.0	412.0
12	2.0	D-14/3/90	42857.0	1.5000	7.7000	238.0	319.0	292.0	33.8000	3.5000	1261.0	7.6000	170.0	268.0	31.3000	4.2000	1204.0	7.6000	116.0	276.0
13	2.0	D-15/3/90	42911.0	0.7000	7.6000	114.0	252.0	116.0	58.6000	1.2000	1238.0	7.9000	148.0	136.0	64.7000	3.0	1208.0	7.7000	79.0	216.0
14	1.0	D-16/3/90	40376.0	·	8.1000	204.0	333.0	174.0	67.8000	3.0	2390.0	7.8000	231.0	156.0	74.4000	2.5000	2540.0	7.8000	136.0	325.0
15	0	D-18/3/90	40923.0	3.5000	7.6000	146.0	329.0	188.0	57.4000	2.5000	1300.0	7.6000	162.0	132.0	63.6000	2.0	1324.0	7.6000	109.0	243.0
16	0	D-19/3/90	43830.0	1.5000	7.8000	177.0	512.0	214.0	58.9000	5.5000	1605.0	7.7000	164.0	256.0	71.9000	5.5000	1599.0	7.7000	118.0	320.0
17	0	D-20/3/90	39165.0	1.2000	7.4000	250.0	447.0	252.0	61.1000	7.0	1533.0	7.4000	275.0	216.0	57.4000	6.5000	1501.0	7.4000	138.0	269.0
18	1.0	D-21/3/90	35791.0	1.2000	7.8000	277.0	466.0	246.0	63.4000	4.0	1556.0	7.7000	·	288.0	65.3000	6.0	1846.0	7.7000	166.0	419.0
19	1.0	D-22/3/90	37419.0	1.2000	7.6000	219.0	446.0	222.0	61.3000	5.5000	1600.0	7.7000	266.0	240.0	70.0	5.0	1645.0	7.6000	172.0	345.0
20	0	D-23/3/90	40983.0	3.0	7.6000	182.0	431.0	214.0	57.0	7.0	1591.0	7.5000	219.0	248.0	58.1000	5.5000	1473.0	7.5000	175.0	376.0