

Hands-on: Sparkling Water

Spark  + H₂O

**SPARKLING
WATER**

Michal
Malohlava

**Can I call H2O's
algorithms from
my Spark
workflow?**

Spark

Open-source distributed execution platform

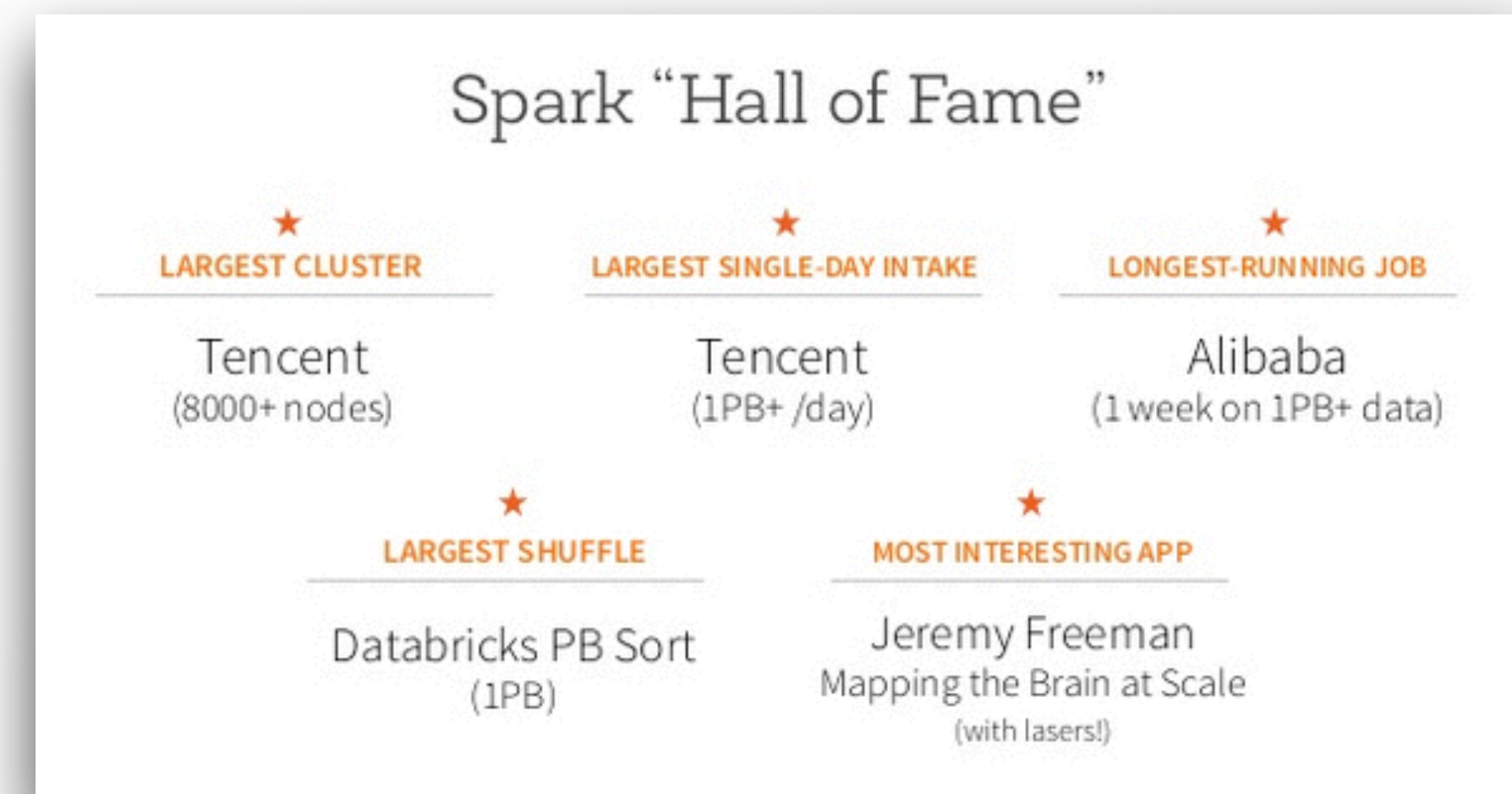


User-friendly API for data transformation based on DataFrames

Platform components - SQL, MLlib, NLP

Multitenancy

Large and active community



**Can I call H2O's
algorithms from
my Spark
workflow?**

**YES,
You can!**

Sparkling Water

Provides

Transparent integration of H2O with Spark ecosystem

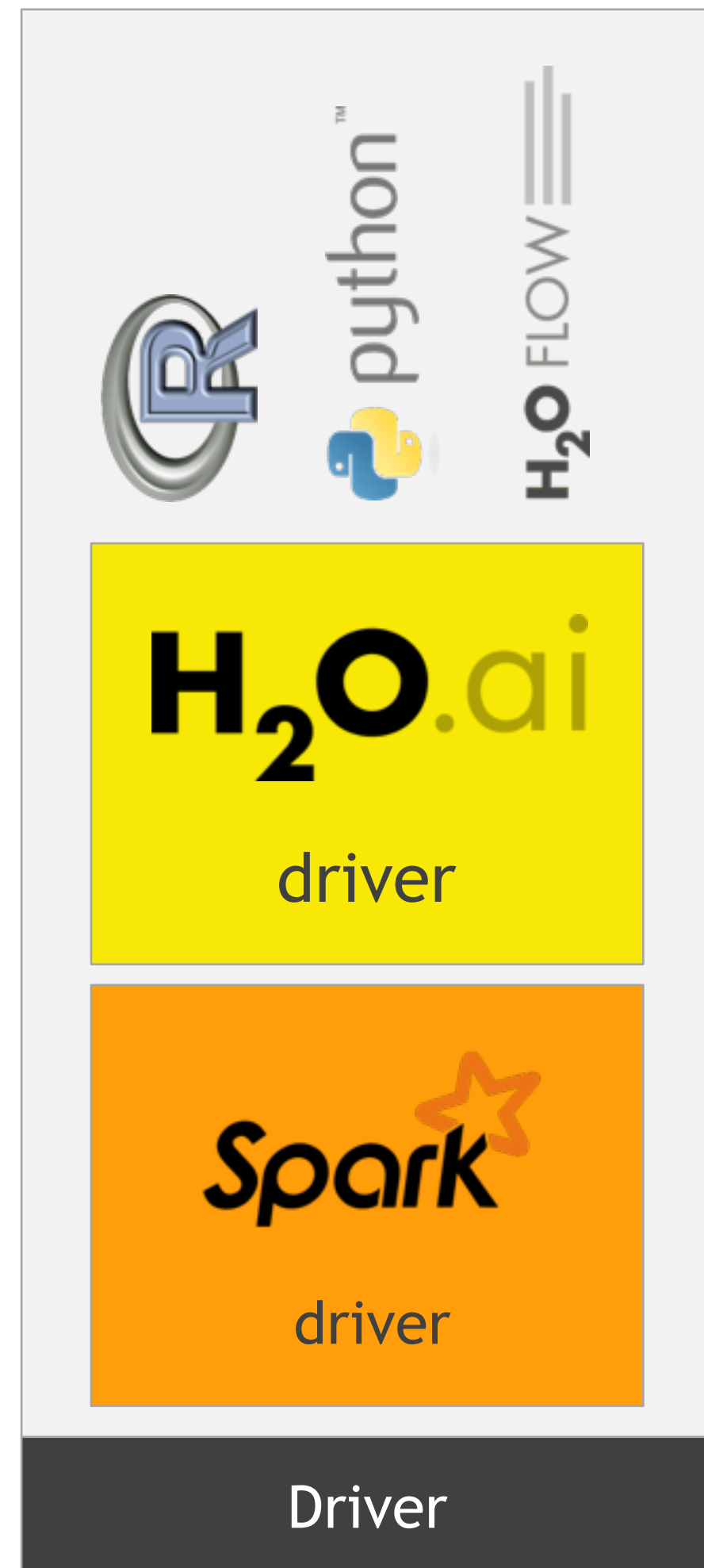
Transparent use of H2O data structures and algorithms with Spark API

Platform for building Smarter Applications

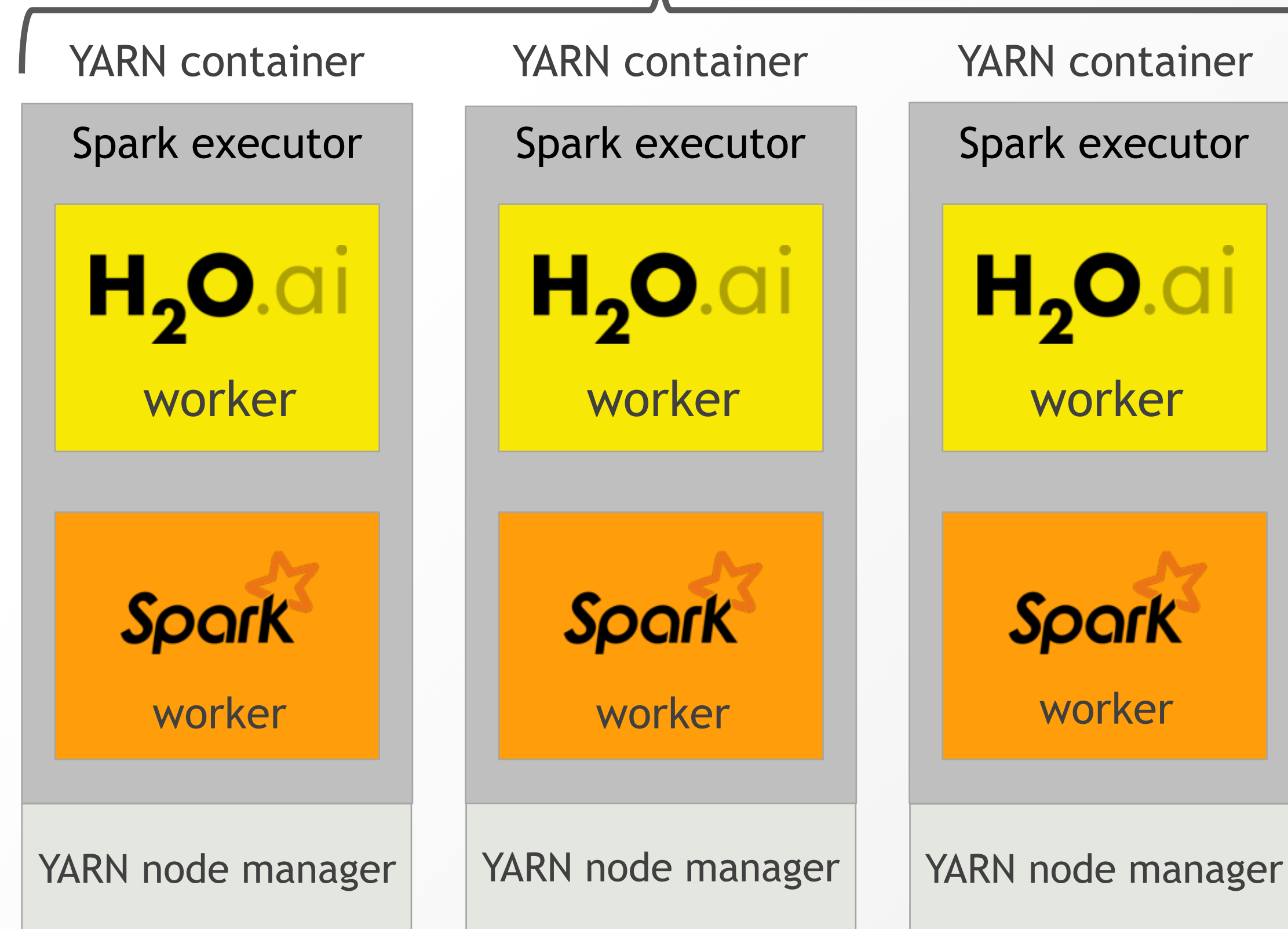
Excels in existing Spark workflows requiring advanced Machine Learning algorithms

Sparkling Water cluster of size 3 on YARN

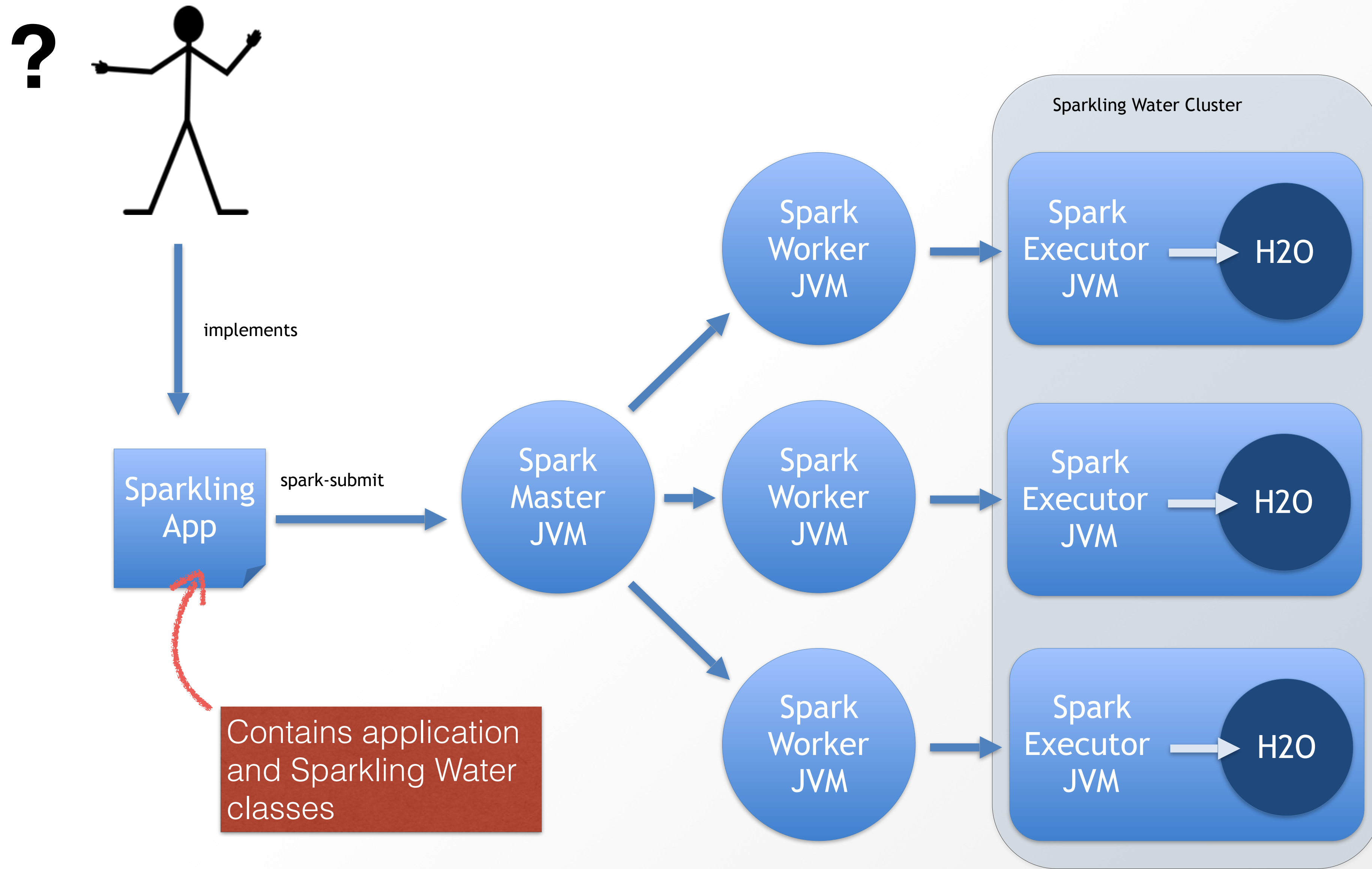
Scala/Py main program



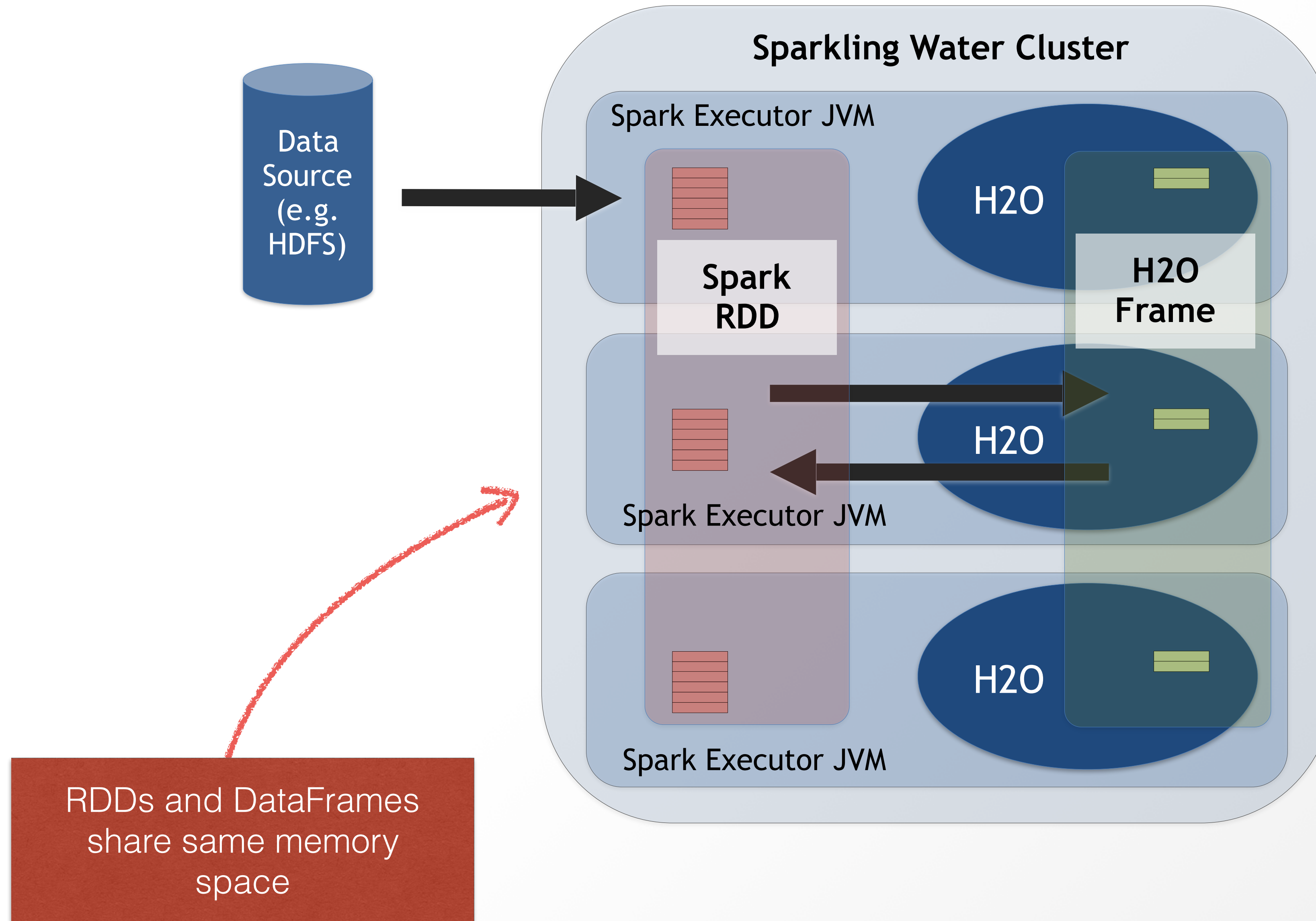
Spark + H₂O
SPARKLING
WATER



Sparkling Water Application Lifecycle



Data Sharing



**Lets play
with it!**

OR



Detect spam text messages

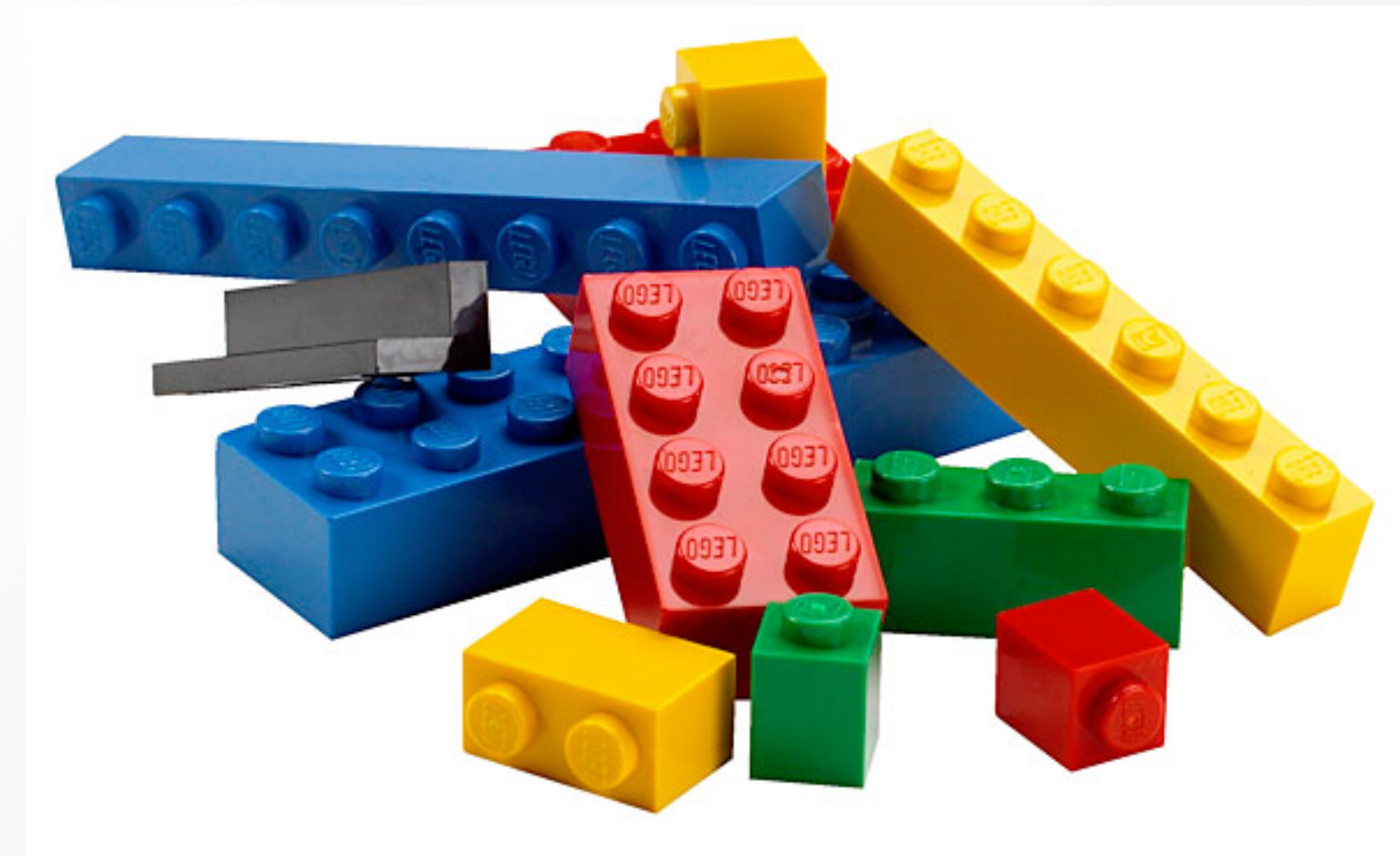
Data sample

	A	B
1	ham	Ok... But they said i've got wisdom teeth hidden inside n mayb need 2 remove.
2	ham	U thk of wat to eat tonight.
3	ham	I dunno until when... Lets go learn pilates...
4	spam	Someoneone you know is trying to contact you via our dating service! To find out who it could be call from your mobile or landline 09064015307 BOX334SK38ch
5	ham	Ok c u then.
6	spam	URGENT! We are trying to contact U. Todays draw shows that you have won a £800 prize GUARANTEED. Call 09050003091 from land line. Claim C52. Valid12hrs only
7	spam	Not heard from U4 a while. Call 4 rude chat private line 01223585334 to cum. Wan 2C pics of me gettin shagged then text PIX to 8552. 2End send STOP 8552 SAM xxx
8	ham	staff.science.nus.edu.sg/~phyhcmk/teaching/pc1323
9	ham	Thank god they are in bed!
10	ham	Hey tmr meet at bugis 930 ?
11	spam	You are a winner you have been specially selected to receive £1000 cash or a £2000 award. Speak to a live operator to claim call 087123002209am-7pm. Cost 10p
12	spam	URGENT! Your Mobile No. was awarded £2000 Bonus Caller Prize on 5/9/03 This is our final try to contact U! Call from Landline 09064019788 BOX42WR29C, 150PPM
13	spam	Loan for any purpose £500 - £75,000. Homeowners + Tenants welcome. Have you been previously refused? We can still help. Call Free 0800 1956669 or text back 'help'
14	ham	Haha... Sounds crazy, dunno can tahan anot...
15	spam	You have won ?spam 000 cash or a ?2,000 prize! To claim, call09050000327
16	ham	Sorry i din lock my keypad.
17	ham	Thanx but my birthday is over already.
18	spam	FREE for 1st week! No1 Nokia tone 4 ur mobile every week just txt NOKIA to 8077 Get txting and tell ur mates. www.getzed.co.uk POBox 36504 W45WQ 16+ norm150p/tono
19	spam	Congratulations - Thanks to a good friend U have WON the £2,000 Xmas prize. 2 claim is easy, just call 08712103738 NOW! Only 10p per minute. BT-national-rate
20	ham	Me n him so funny...
21	spam	pdate_Now - Double mins and 1000 txts on Orange tariffs. Latest Motorola, SonyEricsson & Nokia & Bluetooth FREE! Call MobileUpd8 on 08000839402 or call2optout/!YHL
22	ham	Ok...
23	ham	Yup no more already... Thanx 4 printing n handing it up.
24	ham	Anything lor. Juz both of us lor.
25	ham	It's é only \$140 ard...É rest all ard \$180 at least...Which is é price 4 é 2 bedrm (\$900)
26	ham	Oh oh... Den muz change plan liao... Go back have to yan jiu again...
27	ham	Ok lor then we go tog lor...
28	ham	Okay lor... Wah... like that def they wont let us go... Haha... What did they say in the terms and conditions?
29	ham	Dunno lei... I thk mum lazy to go out... I neva ask her yet...
30	ham	THATS ALRITE GIRL, U KNOW GAIL IS NEVA WRONG!!TAKE CARE SWEET AND DONT WORRY.C U L8TR HUN!LOVE Yaxxx

ML Workflow

Goal: For a given text message identify if it is spam or not

1. Extract data
2. Transform, tokenize messages
3. Build Tf-IDF model
4. Create and evaluate **Deep Learning** model
5. Use the model to detect spam



Application environment



Lego #1: Data load

```
// Data load
def load(dataFile: String): RDD[Array[String]] = {
  sc.textFile(dataFile).map(l => l.split("\t"))
                           .filter(r => !r(0).isEmpty)
}
```

	A	B
1	ham	Ok... But they said i've got wisdom teeth hidden inside n mayb need 2 remove.
2	ham	U thk of wat to eat tonight.
3	ham	I dunno until when... Lets go learn pilates...
4	spam	Someoneone you know is trying to contact you via our dating service! To find out who it could be call from your mobile or landline 09064015307 BOX334SK38ch
5	ham	Ok c u then.
6	spam	URGENT! We are trying to contact U. Todays draw shows that you have won a £800 prize GUARANTEED. Call 09050003091 from land line. Claim C52. Valid12hrs only
7	spam	Not heard from U4 a while. Call 4 rude chat private line 01223585334 to cum. Wan 2C pics of me gettin shagged then text PIX to 8552. 2End send STOP 8552 SAM xxx
8	ham	staff.science.nus.edu.sg/~phyhcmk/teaching/pc1323
9	ham	Thank god they are in bed!
10	ham	Hey tmr meet at bugis 930 ?
11	spam	You are a winner you have been specially selected to receive £1000 cash or a £2000 award. Speak to a live operator to claim call 087123002209am-7pm. Cost 10p
12	spam	URGENT! Your Mobile No. was awarded £2000 Bonus Caller Prize on 5/9/03 This is our final try to contact U! Call from Landline 09064019788 BOX42WR29C, 150PPM
13	spam	Loan for any purpose £500 - £75,000. Homeowners + Tenants welcome. Have you been previously refused? We can still help. Call Free 0800 1956669 or text back 'help'
14	ham	Haha... Sounds crazy, dunno can tahan anot...
15	spam	You have won ?spam 000 cash or a ?2,000 prize! To claim, call09050000327

Lego #2: Ad-hoc Tokenization

```
def tokenize(data: RDD[String]): RDD[Seq[String]] = {  
  val ignoredWords = Seq("the", "a", ...)  
  val ignoredChars = Seq(',', ' ', ':', ...)  
  
  val texts = data.map( r => {  
    var smsText = r.toLowerCase  
    for( c <- ignoredChars) {  
      smsText = smsText.replace(c, ' ')  
    }  
  
    val words = smsText.split(" ").filter(w =>  
      !ignoredWords.contains(w) && w.length>2).distinct  
    words.toSeq  
  })  
  texts  
}
```

Lego #3: Tf-IDF

```
def buildIDFModel(tokens: RDD[Seq[String]],
                  minDocFreq: Int = 4,
                  hashSpaceSize: Int = 1 << 10):
    (HashingTF, IDFModel, RDD[Vector]) = {
        // Hash strings into the given space
        val hashingTF = new HashingTF(hashSpaceSize)
        val tf = hashingTF.transform(tokens)

        // Build term frequency-inverse document frequency
        val idfModel = new IDF(minDocFreq=minDocFreq).fit(tf)
        val expandedText = idfModel.transform(tf)
        (hashingTF, idfModel, expandedText)
    }
```

Hash words
into large
space

Term freq scale

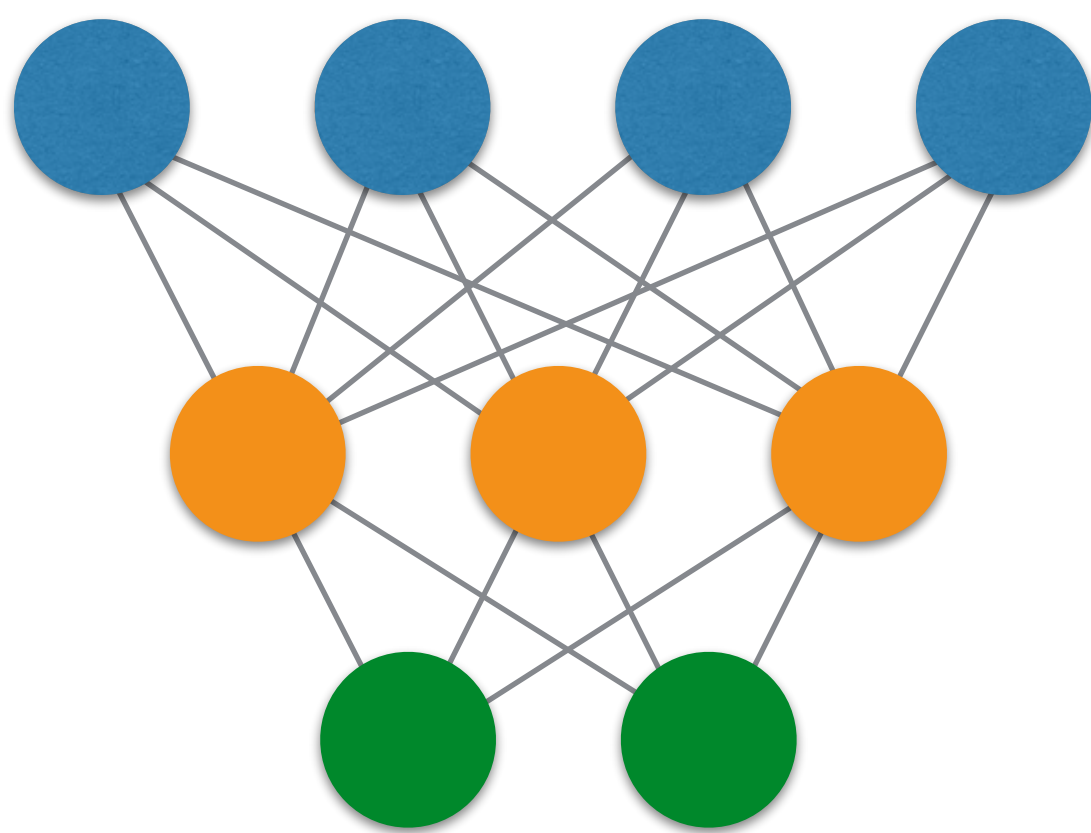
“Thank for the order...”



[..., 0, **3.5**, 0, 1, 0, **0.3**, 0, 1.3, 0, 0,...]
Thank **Order**

Lego #4: Build a model

Deep Learning: Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations



```
def buildDLModel(train: Frame, valid: Frame, epochs: Int = 10,
                  l1: Double = 0.001, l2: Double = 0.0,
                  hidden: Array[Int] = Array[Int](200, 200))
  (implicit h2oContext: H2OContext): DeepLearningModel = {
  import h2oContext._
  // Build a model
  val dlParams = new DeepLearningParameters()
  dlParams._destination_key = Key.make("dlModel.hex")
  dlParams._train = train
  dlParams._valid = valid
  dlParams._response_column = 'target
  dlParams._epochs = epochs
  dlParams._l1 = l1
  dlParams._hidden = hidden

  // Create a job
  val dl = new DeepLearning(dlParams)
  val dlModel = dl.trainModel.get

  // Compute metrics on both datasets
  dlModel.score(train).delete()
  dlModel.score(valid).delete()

  dlModel
}
```

Assembly final workflow

```
// Data load
val data = load(DATAFILE)
// Extract response spam or ham
val hamSpam = data.map( r => r(0))
val message = data.map( r => r(1))
// Tokenize message content
val tokens = tokenize(message)
```

← Data munging
in Spark

```
// Build IDF model
var (hashingTF, idfModel, tfidf) = buildIDFModel(tokens)

// Merge response with extracted vectors
val resultDF = hamSpam.zip(tfidf).map(v => SMS(v._1, v._2))
val tableHF:H2OFrame = resultDF
```

```
// Split table
val keys = Array[String]("train.hex", "valid.hex")
val ratios = Array[Double](0.8)
val frs = split(table, keys, ratios)
val (train, valid) = (frs(0), frs(1))
table.delete()
```

← H2O split dataset

```
// Build a model
val dlModel = buildDLModel(train, valid)
```

Build
H2O model

H2O Flow: Data exploration

H₂O FLOW

Flow Edit View Format Run Admin Help

Untitled Flow

train.hex

ACTIONS: View Data Inspect Build Model... Predict Download Delete

ROWS	COLUMNS	COMPRESSED SIZE
1059	1025	704KB

LABEL	MISSING_COUNT	ZERO_COUNT	POSITIVE_INFINITY_COUNT	NEGATIVE_INFINITY_COUNT	MIN	MAX
target	0	792	0	0	0	1
a0	0	1059	0	0	0	0
a1	0	1059	0	0	0	0
a2	0	636	0	0	0	3.451498120136954
a3	0	0	0	0	4.550110408805064	4.550110408805064
a4	0	397	0	0	0	5.579729825986222
a5	0	0	0	0	4.624218	4.624218

OUTLINE FLOWS CLIPS HELP

Help

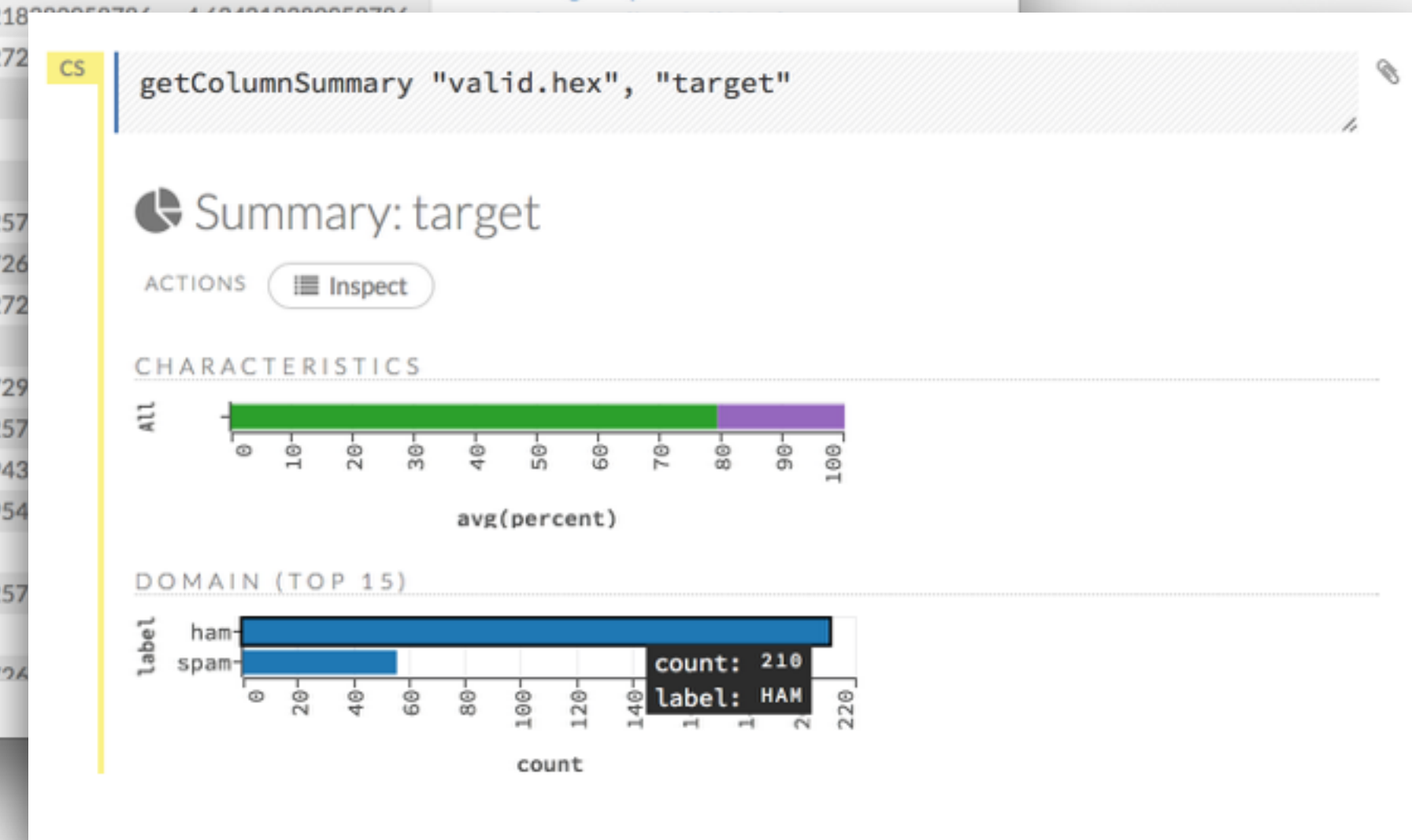
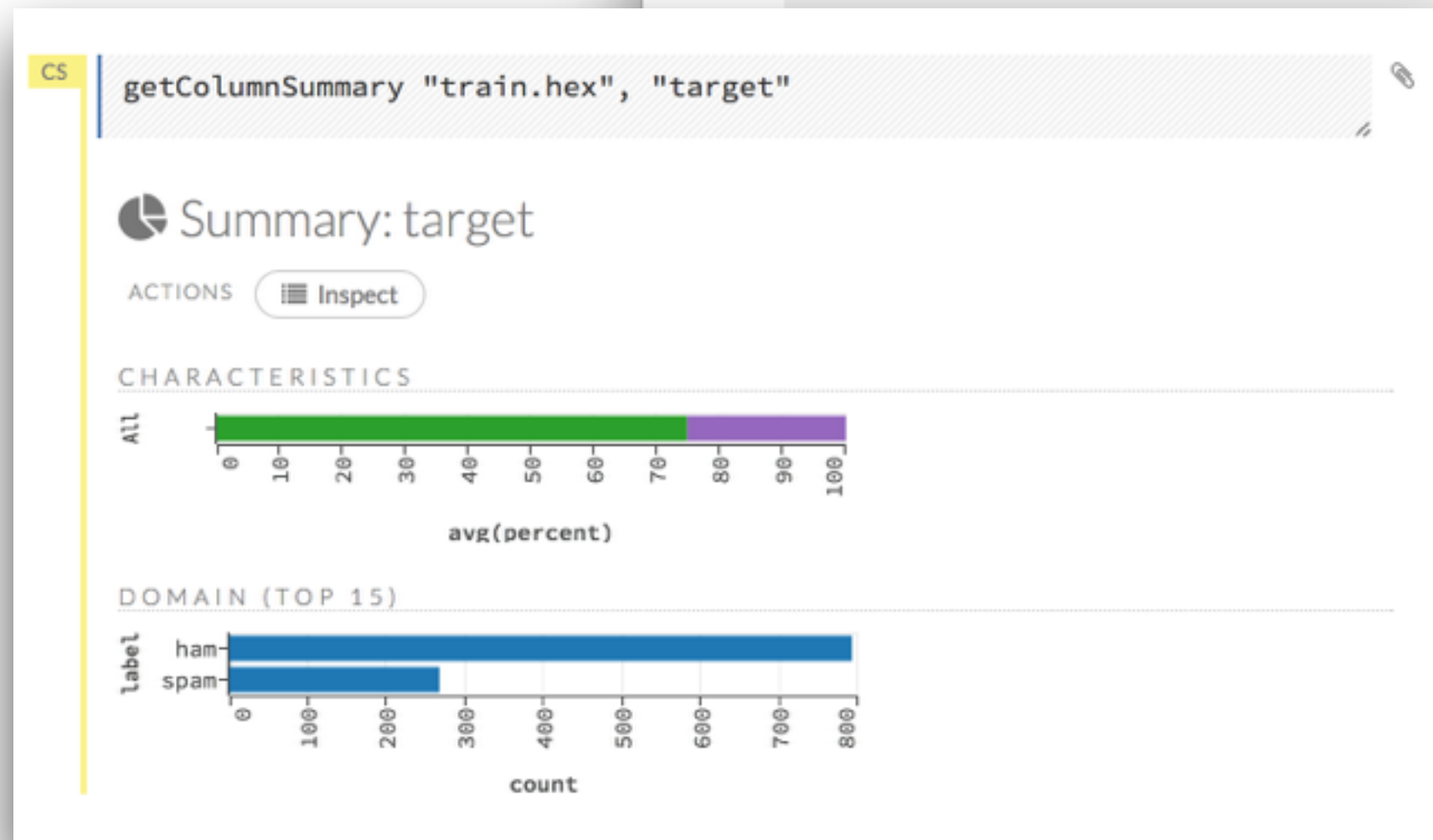
Using Flow for the first time?

Assist Me!

Help Topics

GENERAL

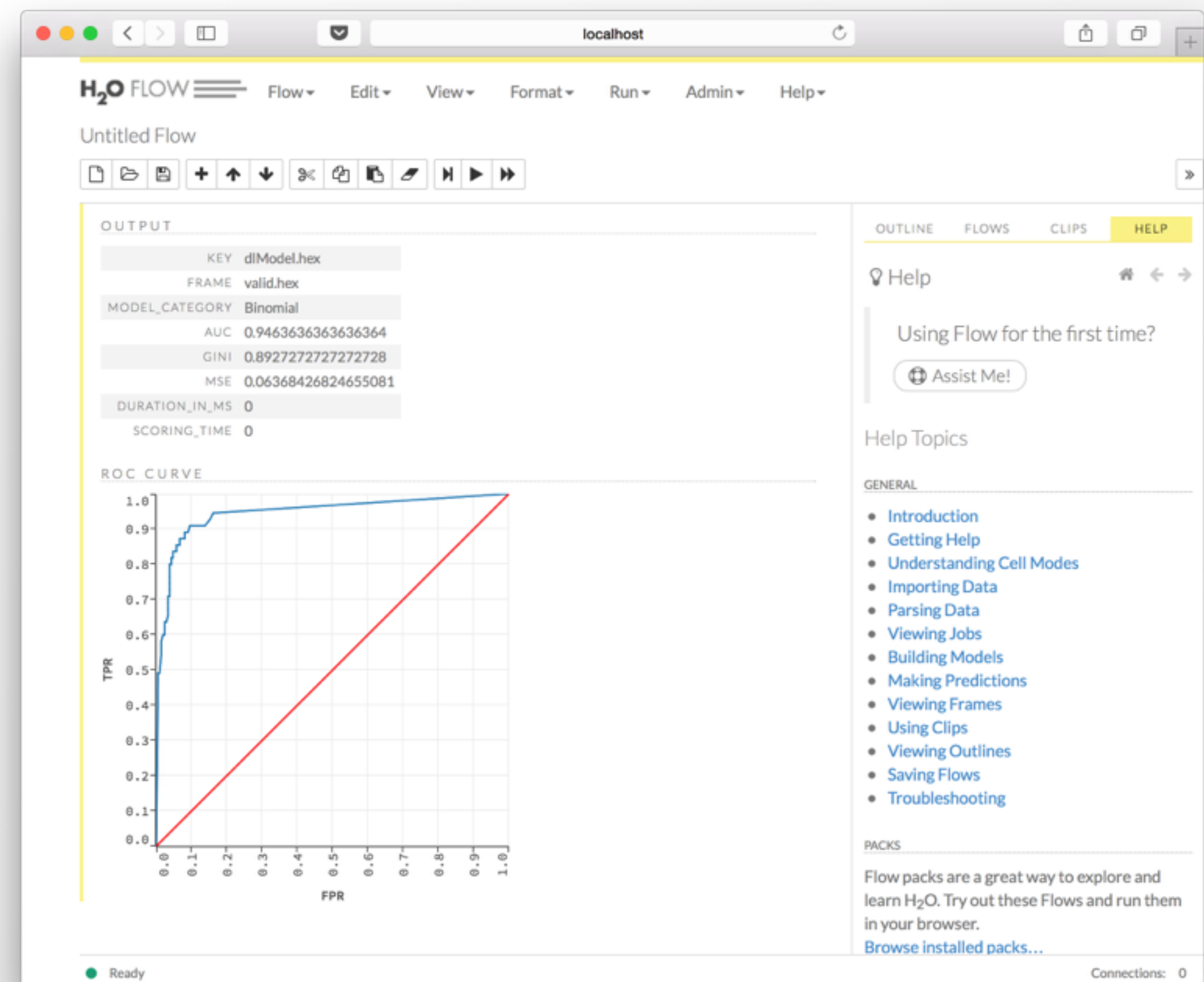
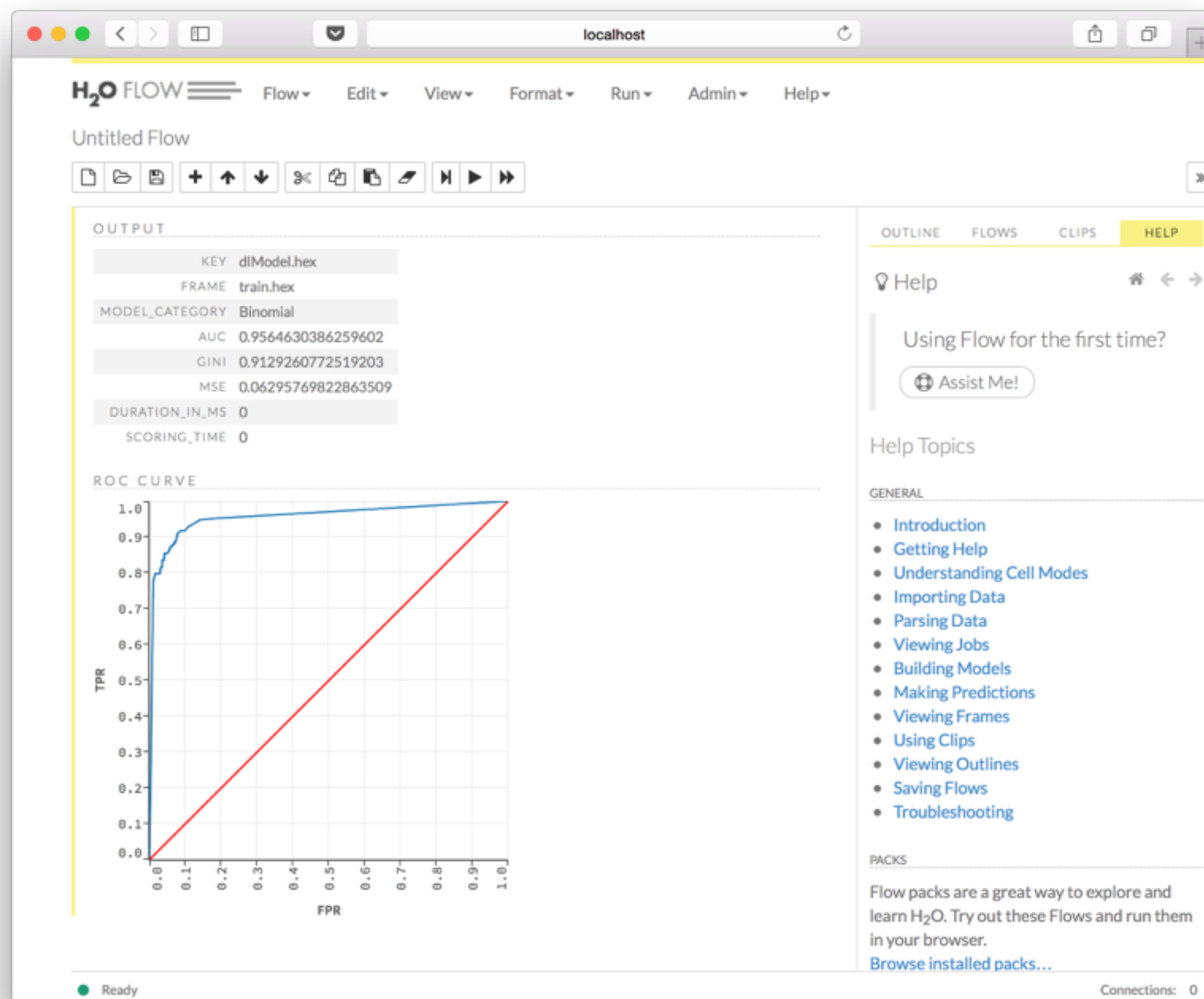
- Introduction
- Getting Help



H2O Flow: Model evaluation

```
val trainMetrics = binomialMM(dlModel, train)
val validMetrics = binomialMM(dlModel, valid)
```

Collect model metrics



Spam predictor

```
def isSpam(msg: String,  
          dlModel: DeepLearningModel, ← Prepared models  
          hashingTF: HashingTF, ← Default decision threshold  
          idfModel: IDFModel,  
          hamThreshold: Double = 0.5): Boolean = {  
  val msgRdd = sc.parallelize(Seq(msg))  
  val msgVector: SchemaRDD = idfModel.transform(  
    hashingTF.transform(  
      tokenize(msgRdd))  
    .map(v => SMS("?", v))  
  )  
  val msgTable: DataFrame = msgVector  
  msgTable.remove(0) // remove first column  
  val prediction = dlModel.score(msgTable) ← Model scoring  
  prediction.vecs()(1).at(0) < hamThreshold  
}
```

Predict spam

isSpam("Michal, H2OWorld
party tomorrow in MV?")



isSpam("We tried to contact
you re your reply
to our offer of a Video
Handset? 750
anytime any networks mins?
UNLIMITED TEXT?")



Thank you!

Sparkling Water is
open-source
ML application platform
combining
power of Spark and H2O

Learn more at h2o.ai
Follow us at [@h2oai](https://twitter.com/h2oai)

