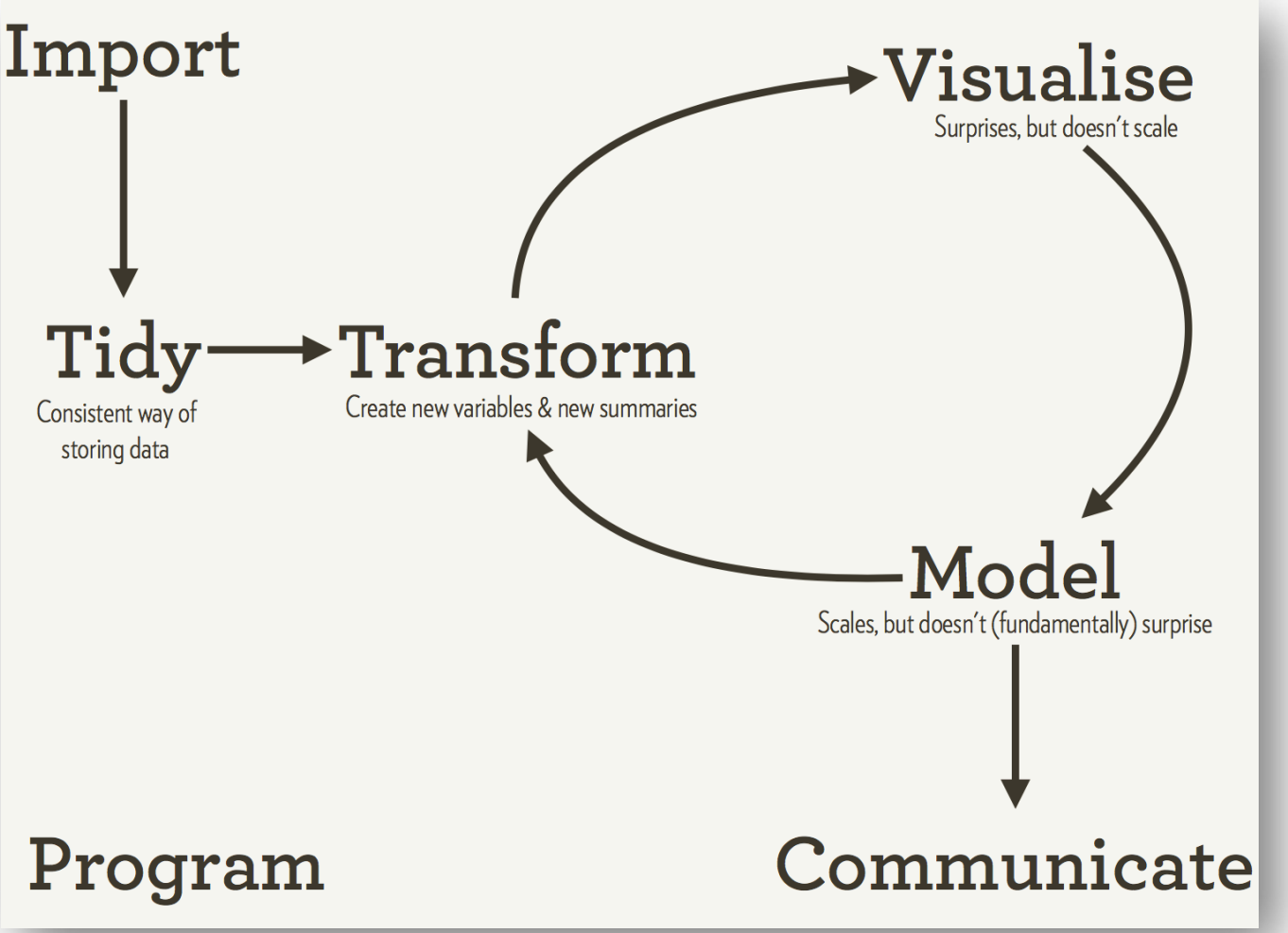
Modeling with R, TensorFlow, and Spark

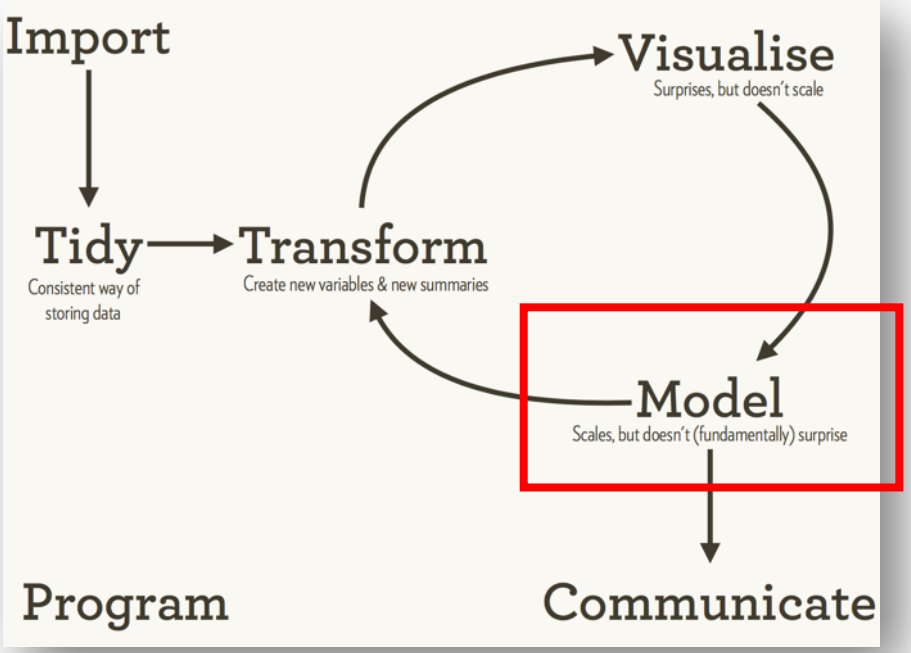
An Introduction to the Landscape for Machine Learning with R

How we think about data science



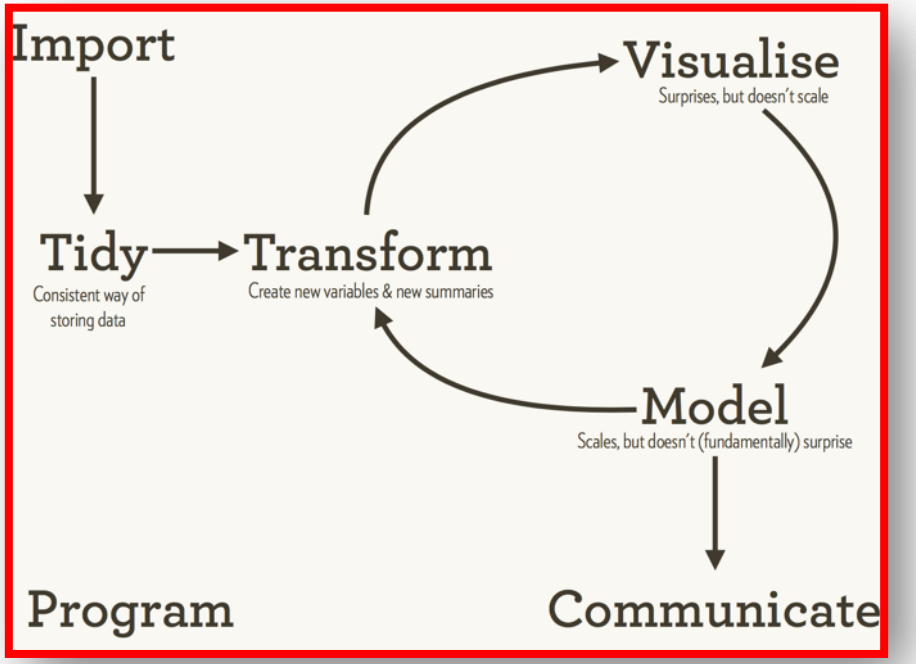
Models and methods

- Regression models
  - Logistic regression (stats)
  - Lasso and Elastic Net Regularized GLM (glmnet)
  - Gradient Boosting Machine (gbm)
- Convolutional neural networks
  - Tensorflow (tensorflow)
- Helpers
  - Keras (keras)
  - Classification and regression training (caret)



Workflow - Census Data

- Data wrangling
  - Tidy + Transformation
- Visualization
  - Exploratory data analysis
- Modeling
  - Logistic regression (stats)
  - Lasso and Elastic Net Regularized GLM (glmnet)
  - Gradient Boosting Machine (gbm & caret)
- Communicate
  - Flex Dashboard
  - Shiny



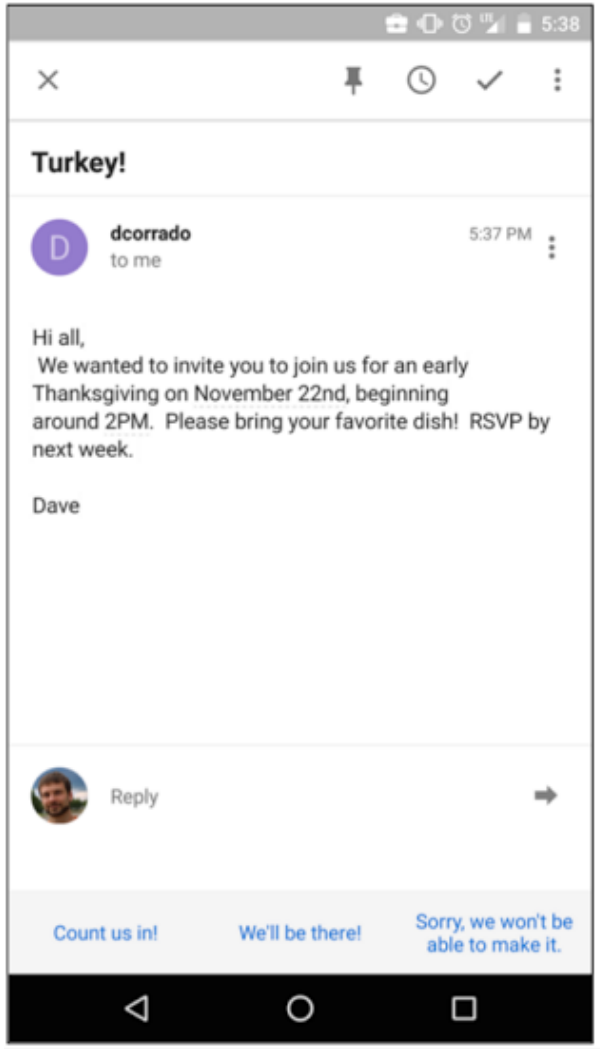
Census Data

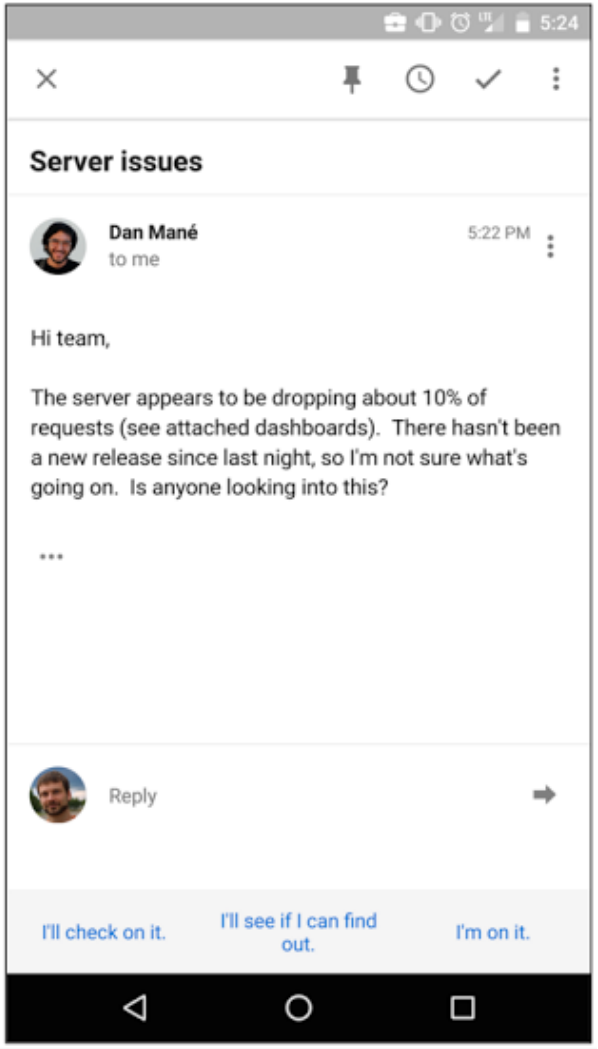
- Prediction task is to determine whether a person makes over 50K a year in the year 1994.
- https://archive.ics.uci.edu/ml/datasets/Census+Income

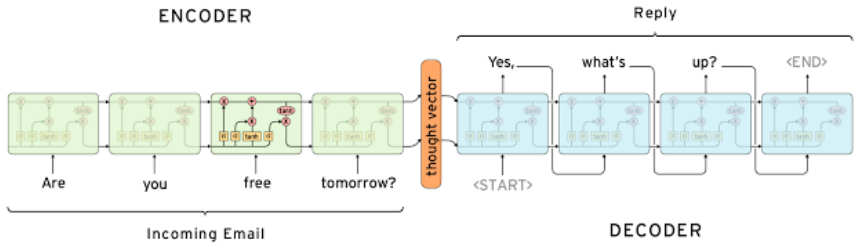
	label	gender	native_country	education	education_num	occupation	workclass	marital_status	race	age_buckets
1	0	Male	United-States	Bachelors	13	Adm-clerical	State-gov	Never-married	White	(35,40]
2	0	Male	United-States	Bachelors	13	Exec-managerial	Self-emp-not-inc	Married-civ-spouse	White	(45,50]
3	0	Male	United-States	HS-grad	9	Handlers-cleaners	Private	Divorced	White	(35,40]
4	0	Male	United-States	11th	7	Handlers-cleaners	Private	Married-civ-spouse	Black	(50,55]
5	0	Female	Cuba	Bachelors	13	Prof-specialty	Private	Married-civ-spouse	Black	(25,30]
6	0	Female	United-States	Masters	14	Exec-managerial	Private	Married-civ-spouse	White	(35,40]
7	0	Female	Jamaica	9th	5	Other-service	Private	Married-spouse-absent	Black	(45,50]
8	1	Male	United-States	HS-grad	9	Exec-managerial	Self-emp-not-inc	Married-civ-spouse	White	(50,55]
9	1	Female	United-States	Masters	14	Prof-specialty	Private	Never-married	White	(30,35]
10	1	Male	United-States	Bachelors	13	Exec-managerial	Private	Married-civ-spouse	White	(40,45]
11	1	Male	United-States	Some-college	10	Exec-managerial	Private	Married-civ-spouse	Black	(35,40]
12	1	Male	India	Bachelors	13	Prof-specialty	State-gov	Married-civ-spouse	Asian-Pac-Islander	(25,30]
13	0	Female	United-States	Bachelors	13	Adm-clerical	Private	Never-married	White	(18,25]

Census Demo

TensorFlow







Smart Reply

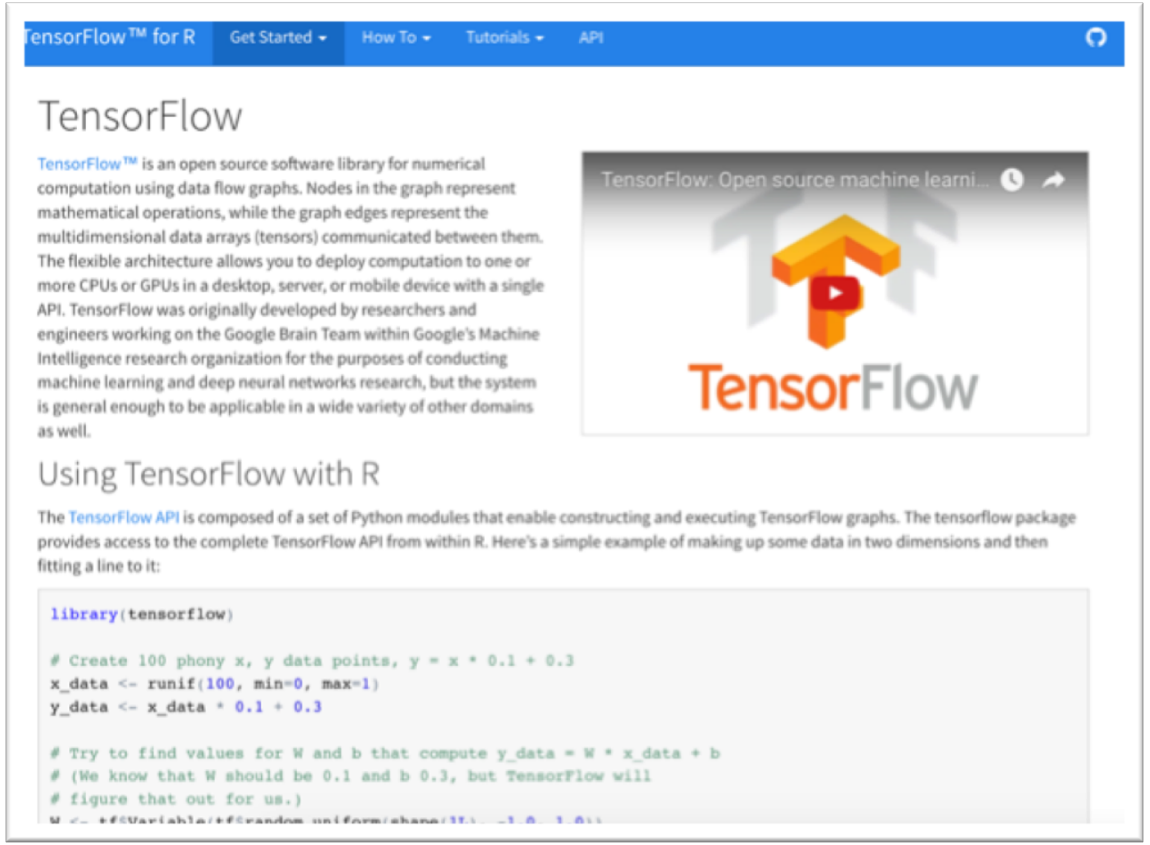
Are you free tomorrow?

Yes, what's up?

https://www.tensorflow.org/about/uses

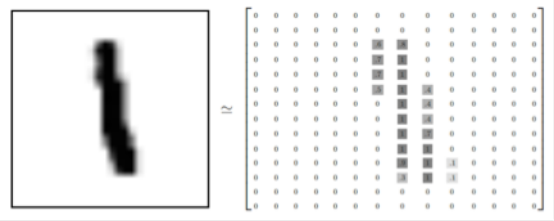
TensorFlow & RStudio

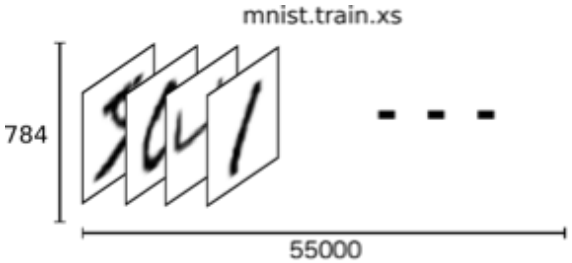
- tensorflow R package
- https://tensorflow.rstudio.com/
- Connects to TensorFlow API
- Runs in standalone or cluster mode
- Portable across servers and devices
- Works with Keras
- Easy to get started!

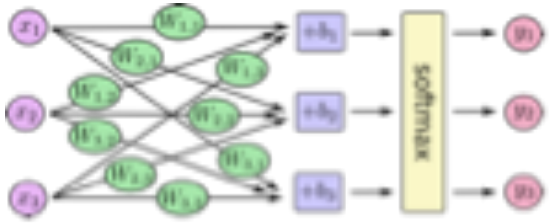


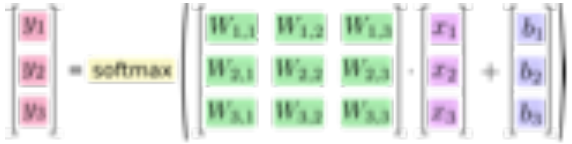
TensorFlow – MNIST Example

- Define your model
  - y = softmax(Wx + b)
- Set up the loss function
  - cross entropy = -sum(y*log(yhat))
- Optimize using stochastic gradient decent
  - Enhanced by backpropogation









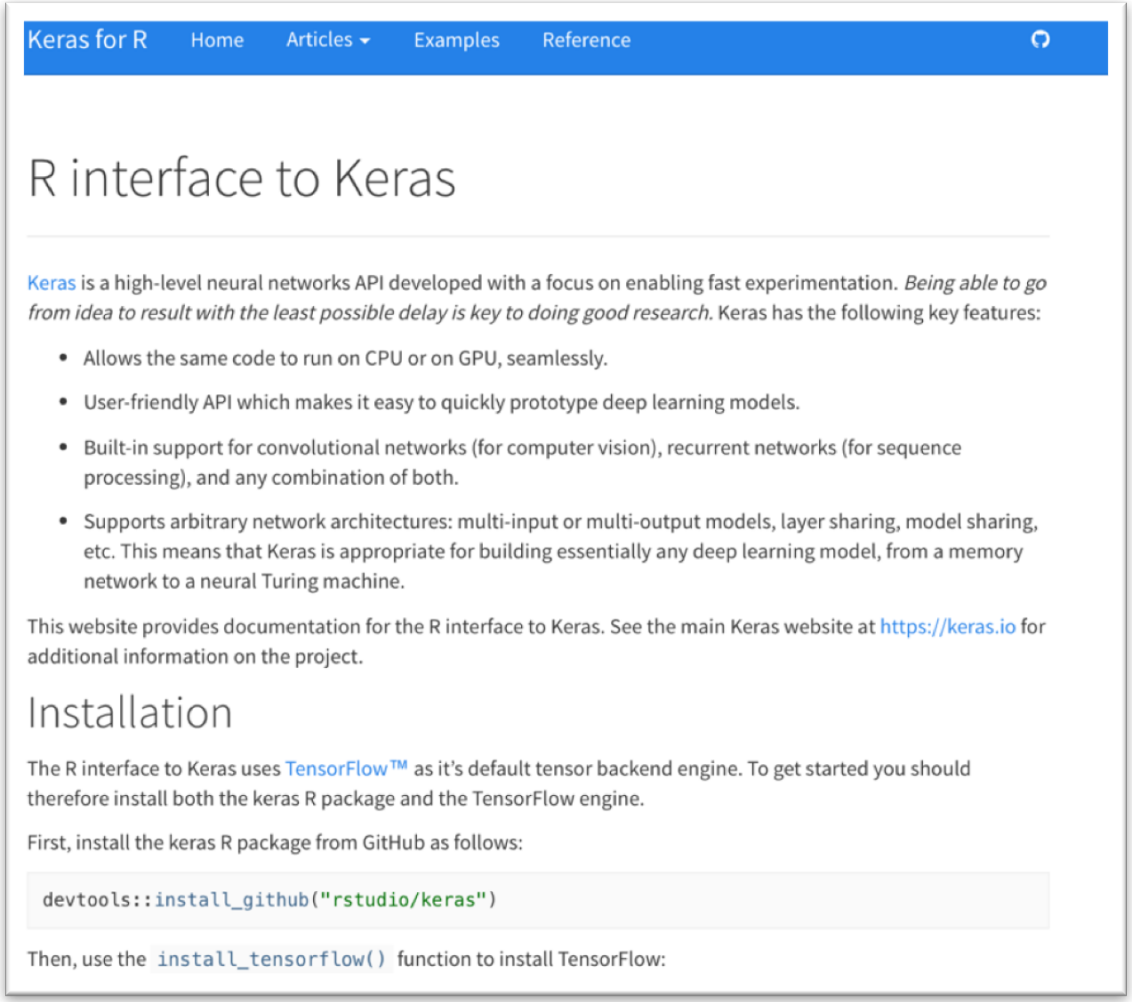
TensorFlow Demo

Keras

- Allows the **same code** to run on CPU or on GPU, seamlessly.
- **User-friendly** API which makes it easy to quickly prototype deep learning models.
- **Built-in support** for convolutional networks (for computer vision), recurrent networks (for sequence processing), and any combination of both.
- Supports arbitrary network architectures: multi-input or multi-output models, layer sharing, model sharing, etc. This means that Keras is appropriate for building essentially **any deep learning model**, from a memory network to a neural Turing machine.

Keras & RStudio

- keras R package
- https://rstudio.github.io/keras
- Uses friendly magrittr syntax
- Runs in standalone or cluster mode
- Uses tensorflow on the backend
- Portable across servers and devices
- Even easier to get started!



Keras for R Home Articles Examples Reference

R interface to Keras

Keras is a high-level neural networks API developed with a focus on enabling fast experimentation. *Being able to go from idea to result with the least possible delay is key to doing good research.* Keras has the following key features:

- Allows the same code to run on CPU or on GPU, seamlessly.
- User-friendly API which makes it easy to quickly prototype deep learning models.
- Built-in support for convolutional networks (for computer vision), recurrent networks (for sequence processing), and any combination of both.
- Supports arbitrary network architectures: multi-input or multi-output models, layer sharing, model sharing, etc. This means that Keras is appropriate for building essentially any deep learning model, from a memory network to a neural Turing machine.

This website provides documentation for the R interface to Keras. See the main Keras website at https://keras.io for additional information on the project.

Installation

The R interface to Keras uses TensorFlow™ as it's default tensor backend engine. To get started you should therefore install both the keras R package and the TensorFlow engine.

First, install the keras R package from GitHub as follows:

```
devtools::install_github("rstudio/keras")
```

Then, use the `install_tensorflow()` function to install TensorFlow:

Keras – Reuters Example

- Sequential model
- Layers
  - Densely connected NN
  - ReLU activation
  - Dropout regularization
  - Densely connected NN
  - Softmax activation
- Optimization
  - Categorical cross entropy
  - Adam
  - Accuracy
- Fit Model
- Assess Performance

Reuters Text Categorization data set (Reuters-21578) document

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

&#3;</BODY></TEXT></REUTERS>

Keras Demo

If you are investing in Spark, then there is nothing stopping you from using it with the full power of R.

Apache Spark

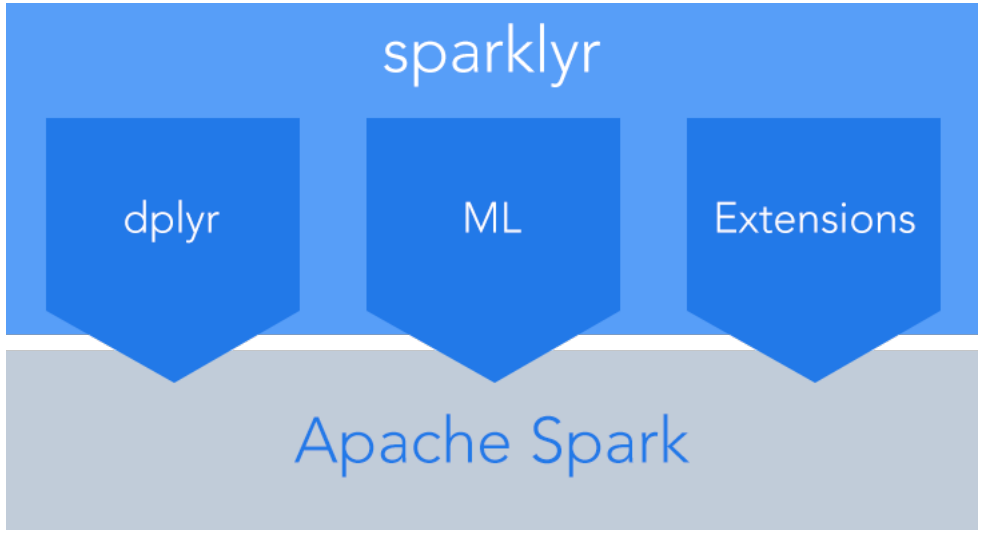
Fast and general engine for large-scale data processing

- *Can integrate with the Hadoop ecosystem*
- *Supports Spark SQL (HiveQL)*
- *Built-in machine learning*
- *Designed for performance*
- *Extensible*

The sparklyr package lets use Spark with R

APACHE Spark™

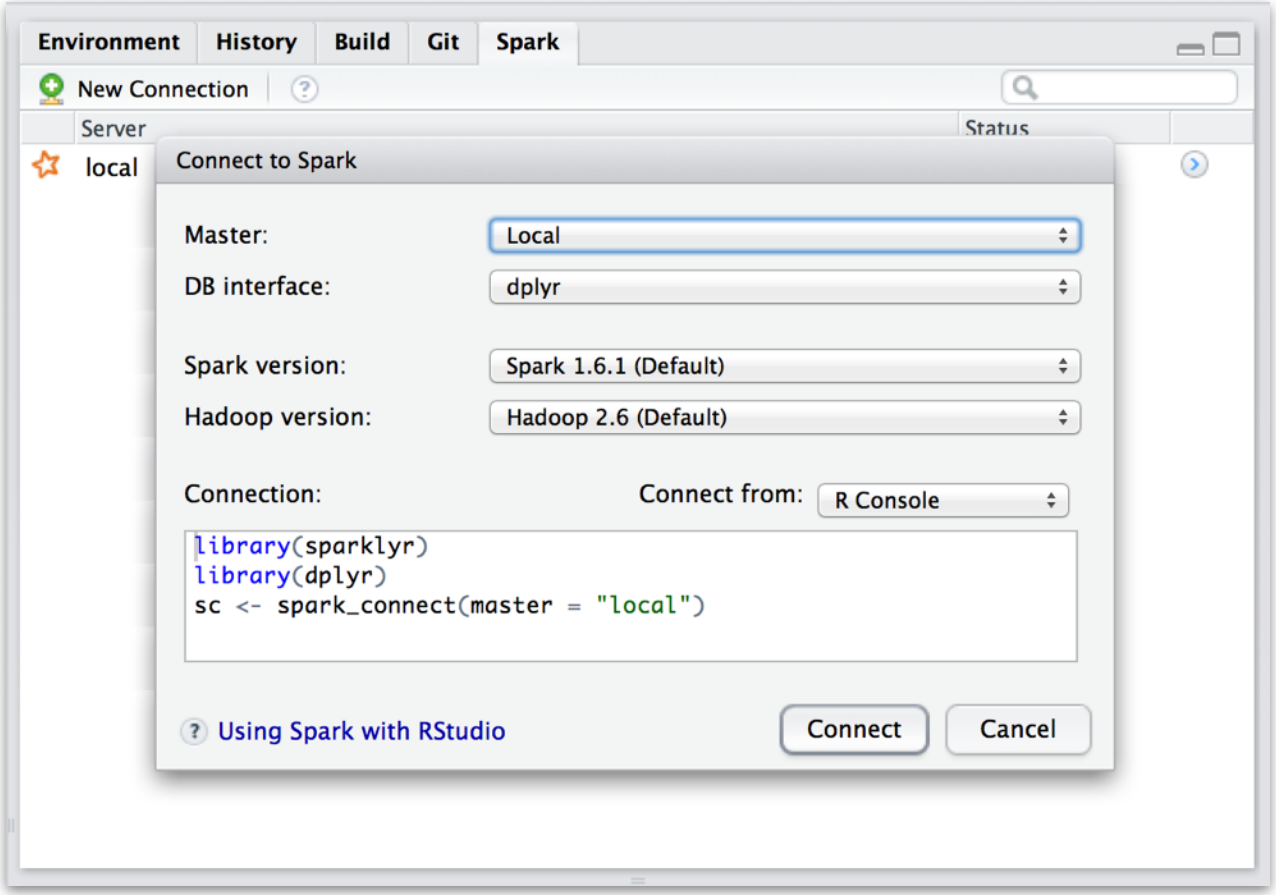
sparklyr has full support for



- dplyr syntax
- Spark ML
- Third party extensions

- *And the RStudio IDE*

Integrated with the RStudio IDE

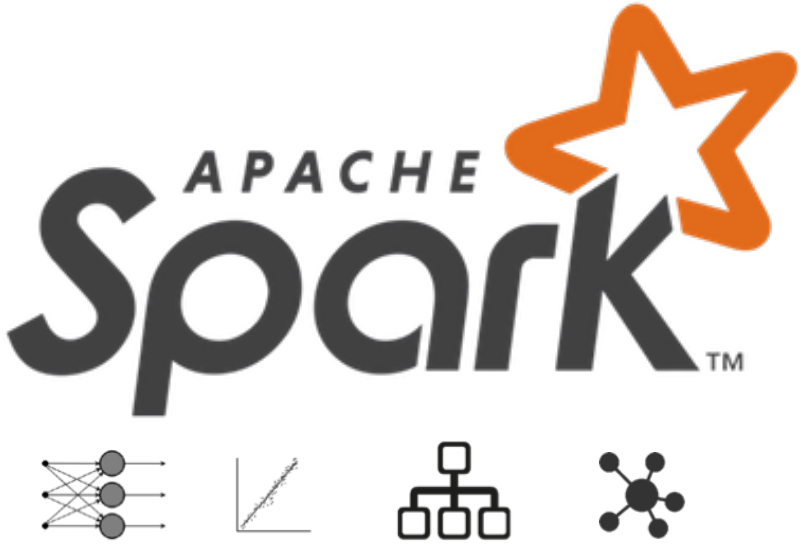


dplyr backend

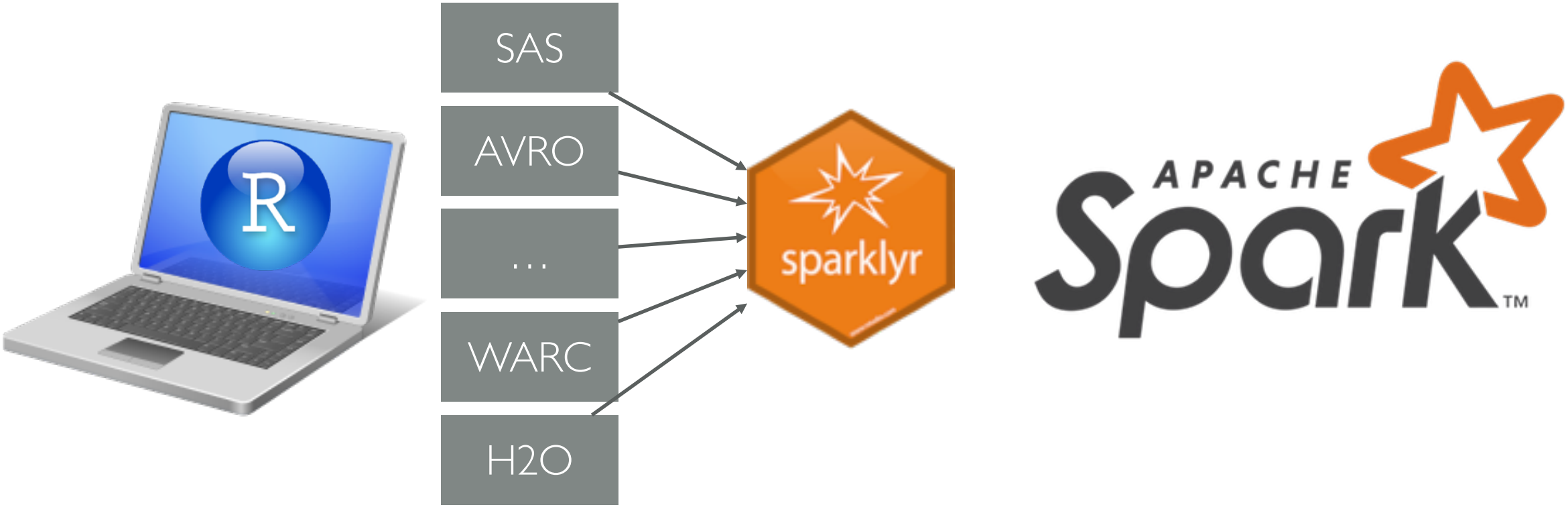

```
flights %>%
  select(arr_delay) %>%
  filter(distance > 100)
```

```
select arr_delay
from flights
where distance > 100
```

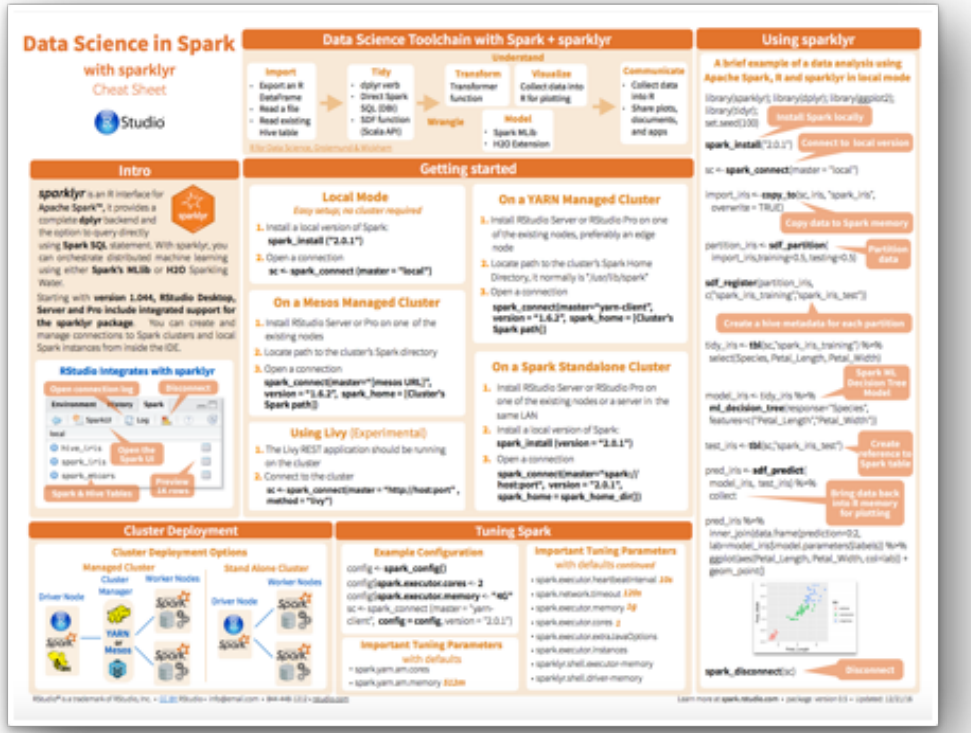
Analyze at scale with Spark ML



Create your own extensions



Run in local mode



Connect to your cluster

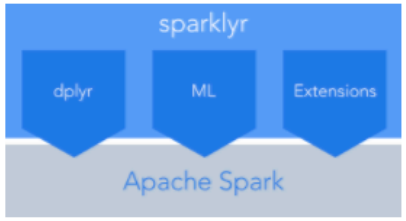
For more information see: spark.rstudio.com

spark.rstudio.com

sparklyr | Home | dplyr | ML | Extensions | Deployment | Examples | Reference

sparklyr: R interface for Apache Spark

- Connect to Spark from R. The sparklyr package provides a complete dplyr backend.
- Filter and aggregate Spark datasets then bring them into R for analysis and visualization.
- Use Sparks distributed machine learning library from R.
- Create extensions that call the full Spark API and provide interfaces to Spark packages.



Installation

You can install the **sparklyr** package from CRAN as follows:

```
install.packages("sparklyr")
```

You should also install a local version of Spark for development purposes:

```
library(sparklyr)
spark_install(version = "1.6.2")
```

To upgrade to the latest version of sparklyr, run the following command and restart your r session:

```
devtools::install_github("rstudio/sparklyr")
```

If you use the RStudio IDE, you should also download the latest preview release of the IDE which includes several enhancements for interacting with Spark (see the RStudio IDE section below for more details).

Connecting to Spark

Links

Download from CRAN at https://cran.r-project.org/package=sparklyr
Report a bug at https://github.com/rstudio/sparklyr/issues

License

Apache License 2.0 | file LICENSE

Developers

Javier Luraschi
Author, maintainer
Kevin Ushey
Author
JJ Allaire
Author
The Apache Software Foundation
Author, copyrightholder
All authors...

Dev status

build passing
CRAN 0.5.4

sparklyr Demo