Ruby without Rails

Doing Data Science and Machine Learning
on the Ruby stack

Q: What do I need to learn
to become Data Scientist?

1992: Learn SQL

2001: Learn R

2008: Learn Python

Spark, Julia, Tidyverse...

2019: Try Ruby (and all above)

Q: How do I start doing Data Science?

A: You open a new Jupyter notebook...
in Ruby!

Ruby without Rails

Ruby On Ice 2019 | Lightning talk

Code: https://github.com/arbox/Ruby_without_Rails_2019

Install dependencies

```
$ apt install gnuplot
$ gem install specific_install
$ gem specific_install https://github.com/SciRuby/iruby
$ gem install iruby-dependencies
```

In [4]:

```ruby
require 'iruby/dependencies'

dependencies do
  gem "red-datasets"
  gem "rdatasets"
  gem "numo-narray"
  gem "numo-gnuplot" # Gnuplot shold be installed already.
  gem "svmkit"
end
```

```
Dependencies installing. This could take a minute ...
```

```
Dependencies successfully installed.
```

Take a well known dataset…

Time Series plots

In [6]:
```
require "daru/view"
require "rdatasets"
lynx = Daru::DataFrame.from_rdatasets :datasets, :lynx
```

Out[6]: **Daru::DataFrame(114x2)**

	time	value
1	1821	269
2	1822	321
3	1823	585
4	1824	871
5	1825	1475
6	1826	2821
7	1827	3928
8	1828	5943
9	1829	4950
10	1830	2577
11	1831	523
12	1832	98
13	1833	184

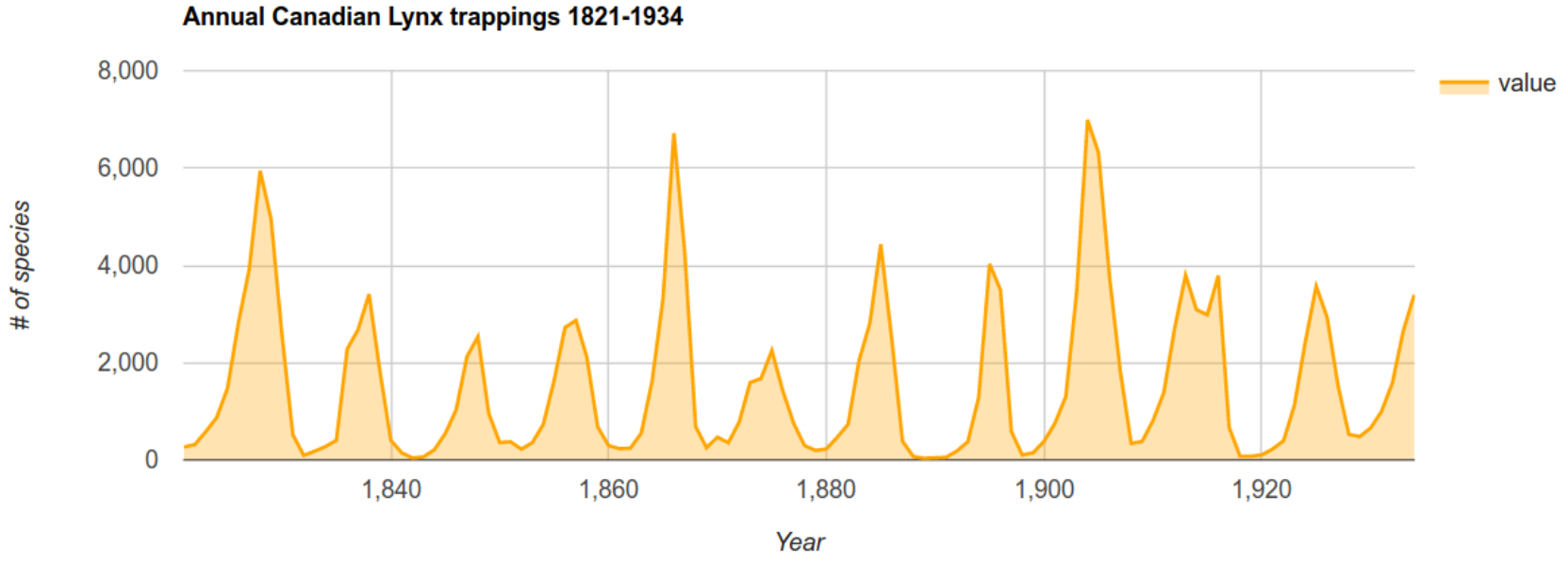
… explore it graphically ...

In [3]:
```
# Annual Canadian Lynx trappings 1821-1934
# Credits kojix2 (https://qiita.com/kojix2), main contributor of RDatasets.

Daru::View.plotting_library = :googlecharts
chart = Daru::View::Plot.new(lynx,
    type: :area,
    title: "Annual Canadian Lynx trappings 1821-1934",
    hAxis: {title: "Year"},
    vAxis: {title: "# of species"},
    colors: ["Orange"],
    height: 360)

chart.show_in_iruby
```

Out[3]:



… and train a model!

Support Vector Machines: Size prediction on the Iris Dataset

In [2]:

```
# Credits @joshoku (https://yoshoku.hatenablog.com), main author of SVMKit.
# https://en.wikipedia.org/wiki/Iris_flower_data_set

require "datasets"
require "svmkit"
require "numo/narray"

# Read in Iris Dataset.
iris = Datasets::Iris.new

# Convert to a tabular representation.
iris_table = iris.to_table

# Divide into labels and feature quantities.
iris_labels = iris_table[:label]
iris_attrs = iris_table.fetch_values(
  :sepal_length, :sepal_width, :petal_length, :petal_width).transpose

# Encode String labels in the Iris Dataset as Integer values (`Numo::Int32`).
encoder = SVMKit::Preprocessing::LabelEncoder.new
labels = encoder.fit_transform(iris_labels)

# Convert feature values of the Iris Dataset to `Numo::DFloat`.
samples = Numo::DFloat[*iris_attrs]

# Define a cross validation by 5-fold division of linear SVM.
svc = SVMKit::LinearModel::SVC.new(
  reg_param: 0.0001, fit_bias: true, max_iter: 3000, random_seed: 1)
kf = SVMKit::ModelSelection::StratifiedKFold.new(n_splits: 5, random_seed: 1)
cv = SVMKit::ModelSelection::CrossValidation.new(estimator: svc, splitter: kf)

# Perform cross validation.
report = cv.perform(samples, labels)

# Output the average Accuracy.
mean_accuracy = report[:test_score].inject(:+) / kf.n_splits
puts("Mean Accuracy: %.1f%%" % (100.0 * mean_accuracy))
```

```
Mean Accuracy: 94.7%
```

Q: Do you want more?

RubyData

Or just talk to us!

GitHub: @arbox