# FlinkML: Vision and Roadmap

## Vision

The Machine Learning (ML) library for Flink is a new effort to bring scalable ML tools to the Flink community. Our goal is is to design and implement a system that is scalable and can deal with problems of various sizes, whether your data size is measured in megabytes or terabytes and beyond. We call this library FlinkML.

An important concern for developers of ML systems is the amount of glue code that developers are forced to write [1] in the process of implementing an end-to-end ML system. Our goal with FlinkML is to help developers keep glue code to a minimum. The Flink ecosystem provides a great setting to tackle this problem, with its scalable ETL capabilities that can be easily combined inside the same program with FlinkML, allowing the development of robust pipelines without the need to use yet another technology for data ingestion and data munging.

Another goal for FlinkML is to make the library easy to use. To that end we will be providing detailed documentation along with examples for every part of the system. Our aim is that developers will be able to get started with writing their ML pipelines quickly, using familiar programming concepts and terminology.

Contrary to other data-processing systems, Flink exploits in-memory data streaming, and natively executes iterative processing algorithms which are common in ML. We plan to exploit the streaming nature of Flink, and provide functionality designed specifically for data streams.

FlinkML will allow data scientists to test their models locally and using subsets of data, and then use the same code to run their algorithms at a much larger scale in a cluster setting.

We are inspired by other open source efforts to provide ML systems, in particular scikit-learn for cleanly specifying ML pipelines, and Spark's MLLib for providing ML algorithms that scale with problem and cluster sizes.

## Roadmap

The roadmap below can provide an indication of the algorithms we aim to implement for the library. Items in **bold** have already been implemented:

- **Pipelines of transformers and learners**
- Data pre-processing
  - **Feature scaling**
  - **Polynomial feature base mapper**
  - Feature hashing
  - Feature extraction for text
  - Dimensionality reduction
- Model selection and performance evaluation
  - Model evaluation using a variety of scoring functions
  - Cross-validation for model selection and evaluation
  - Hyper-parameter optimization
- Supervised learning
  - Optimization framework
    - **Stochastic Gradient Descent**
    - L-BFGS
  - Generalized Linear Models
    - **Multiple linear regression**
    - LASSO, Ridge regression
    - Multi-class Logistic regression
    - Random forests
    - **Support Vector Machines**
  - Decision trees
- Unsupervised learning
  - Clustering
    - K-means clustering
  - Principal Components Analysis
- Recommendation
  - **ALS**
- Text analytics
  - LDA
- Statistical estimation tools
- Distributed linear algebra
- Streaming ML

# How can I help?

Any contribution to the above roadmap is welcome! You can also check out the list of open issues for FlinkML on JIRA, or send a message with your idea to the Flink developers list.

We recommend reading the FlinkML contribution guide before starting out, and definitely subscribe and post a message on the Flink developers mailing list to introduce yourself!

**References:**

D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, and M. Young. *Machine learning: The high interest credit card of technical debt.* In SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop), 2014.