Google Prediction API

Machine Learning as a Service on the Cloud

Max Lin
Software Engineer
Google Research NYC

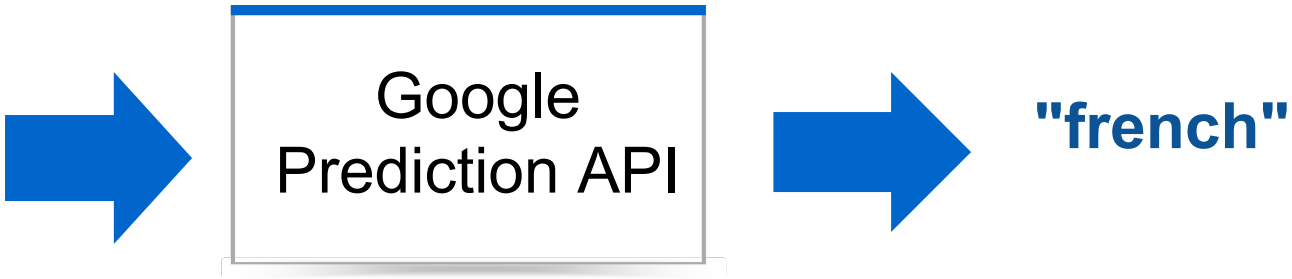
Google Prediction API

Google

Machine Learning as a Service

- Google's machine learning algorithms and infrastructure
- HTTP RESTful web service
- Train asynchronously, predict in real time

"Tous pour un, un pour tous, c'est notre devise."



"french"

How does it work?

Google

The Prediction API finds relevant **features** in the data during **training**.

"english"	**The** quick brown fox jumped over **the** lazy dog.
"english"	**To** err **is** human, but **to** really foul things up you need a computer.
"spanish"	**No** hay mal **que por** bien **no** venga.
"spanish"	**La** tercera **es la** vencida.

The Prediction API later searches for those **features** during **prediction**.

?	**To** be or not **to** be, that **is the** question.
?	**La** fe mueve montañas.

A virtually endless number of applications

Google

Customer Sentiment

Transaction Risk

Species Identification

Message Routing

Diagnostics

Churn Prediction

Legal Docket Classification

Suspicious Activity

Work Roster Assignment

Inappropriate Content

Recommend Products

Political Bias

Uplift Marketing

Email Filtering

Career Counselling

... and many more ...

Three steps to use the Prediction API

Google

1. Upload

Upload your training data to Google Storage

Use the Google Storage API or gsutil to upload your {data} to Google Storage

2. Train

Build a model from your data

POST prediction/v1/training/{data}
make a training request

GET prediction/v1/training/{data}
get training status

3. Predict

Make new predictions

POST prediction/v1/training/{data}/predict
make a prediction request

Analyzing Big Data

Google

Challenges

- Provision large number of machines
- Develop scalable machine learning algorithms
- Integrate and deploy models with existing apps

Make Big Data Analysis Easy

Google

With machine learning as a web service

- Don't need to provision large number of machines
- Don't require substantial investment upfront
- Don't require deep machine learning expertise
- Easy to integrate with existing apps and deploy models

Google Prediction API

Automatically categorize and respond to emails by language

- Customer: ACME, a multinational organization
- Goal: Respond to customer emails in their languages
- Data: Many emails, tagged with their languages
- Outcome: Predict language and respond accordingly

Message Routing

Step 1: Upload

Google

Upload your training data to Google Storage

- Training data: outputs and input features
- Data format: comma separated value format (CSV)

```
$ head -n 2 ${data}
"english","To err is human, but to really ..."
"spanish","No hay mal que por bien no venga."
```

Upload to Google Storage

```
$ gsutil cp ${data} gs://${bucket}/${object}
```

Step 2: Train

Google

Create a new model by training on data
To train a model:

POST prediction/v1/training/${data}

Training runs asynchronously.
To see if it finishes:
GET prediction/v1/training/${data}

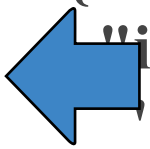
```
{"data": {
    "data": "${data}",
    "modelinfo": "estimated accuracy: ${acc}"}}}
```

Step 3: Predict

Google

Make predictions on new data in JSON format

POST prediction/v1/training/${data}/predict



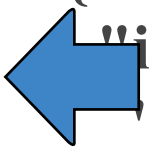
```
{"data": {
  "input": {
    "text": ["Cada niño es un artista"]}}}
```

Step 3: Predict

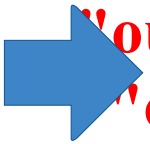
Google

Apply the trained model to make predictions on new data

POST prediction/v1/training/${data}/predict



```
{"data": {
  "input": {
    "text": ["Cada niño es un artista"]}}}
```



```
{"data": {
  "output": {
    "output_label": "Spanish" }}}
```

Step 3: Predict

Google

Make predictions on new data in Python

```
import urllib2

headers = {"Content-Type": "application/json"}test_data
= {"data":   {"input":     {"text": "Cada nino es un artista"}}}
urllib2.Request("https://www.googleapis.com" +
"/prediction/v1/training/%s/prediction" % data, data = simplejson.
dumps(test_data),headers = headers)
```

Google Prediction API Capabilities

Google

Data

- Input Features: numeric or unstructured text
- Output: up to 100s of discrete categories

Training

- Many machine learning techniques
- Automatically selected
- Performed asynchronously

Access from many environments

- Web app, mobile app, desktop app
- Programming language independent

Supervised Learning on Large Data

Google

Shard data to overcome memory limitation

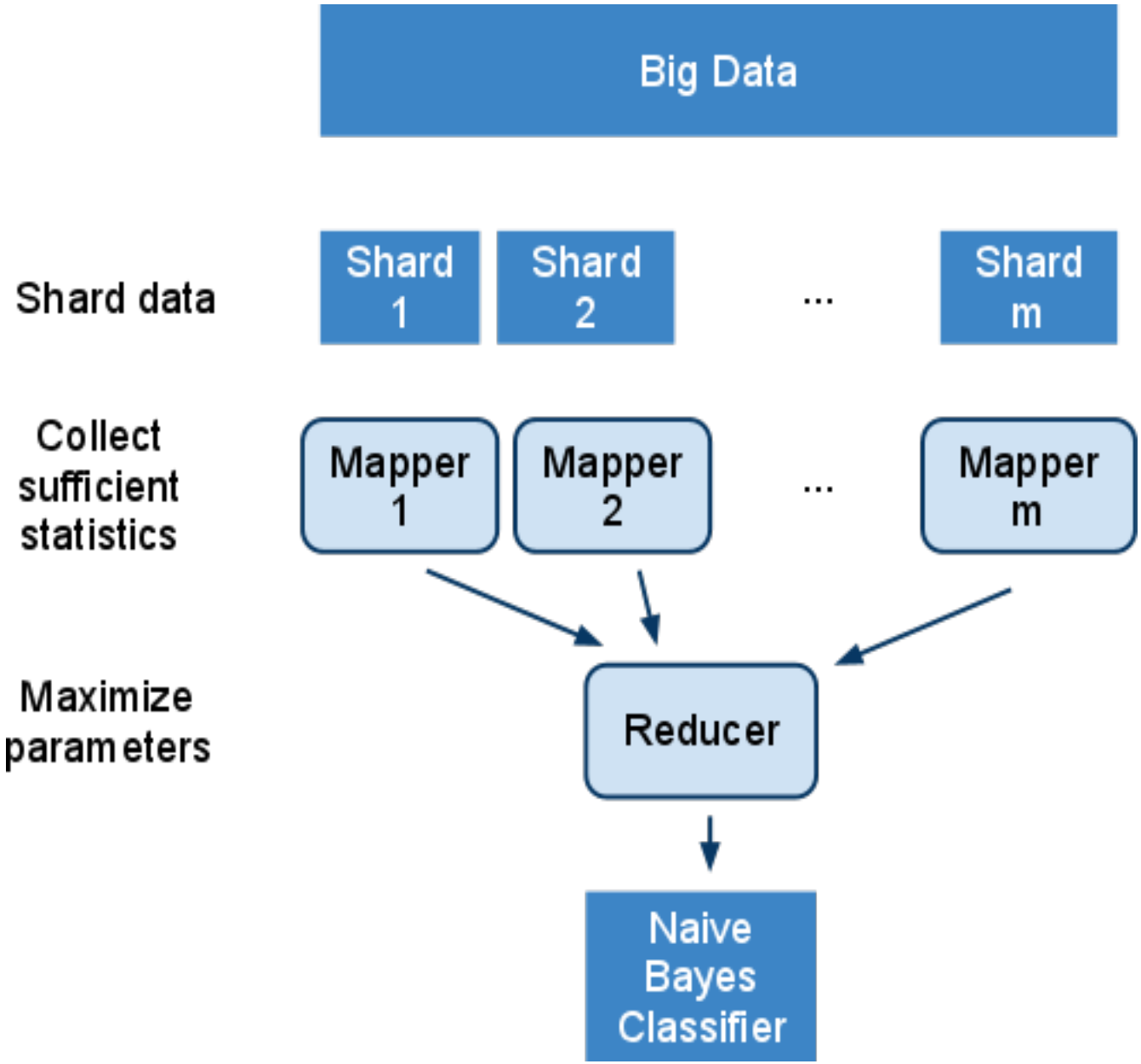
MapReduce as building blocks
Accuracy and resource trade-off

Some strategies for large-scale supervised learning:

1. Sub-sample
2. Embarrassingly parallelize some algorithms
3. Distributed gradient descent
4. Majority Vote
5. Parameter mixture
6. Iterative parameter mixture

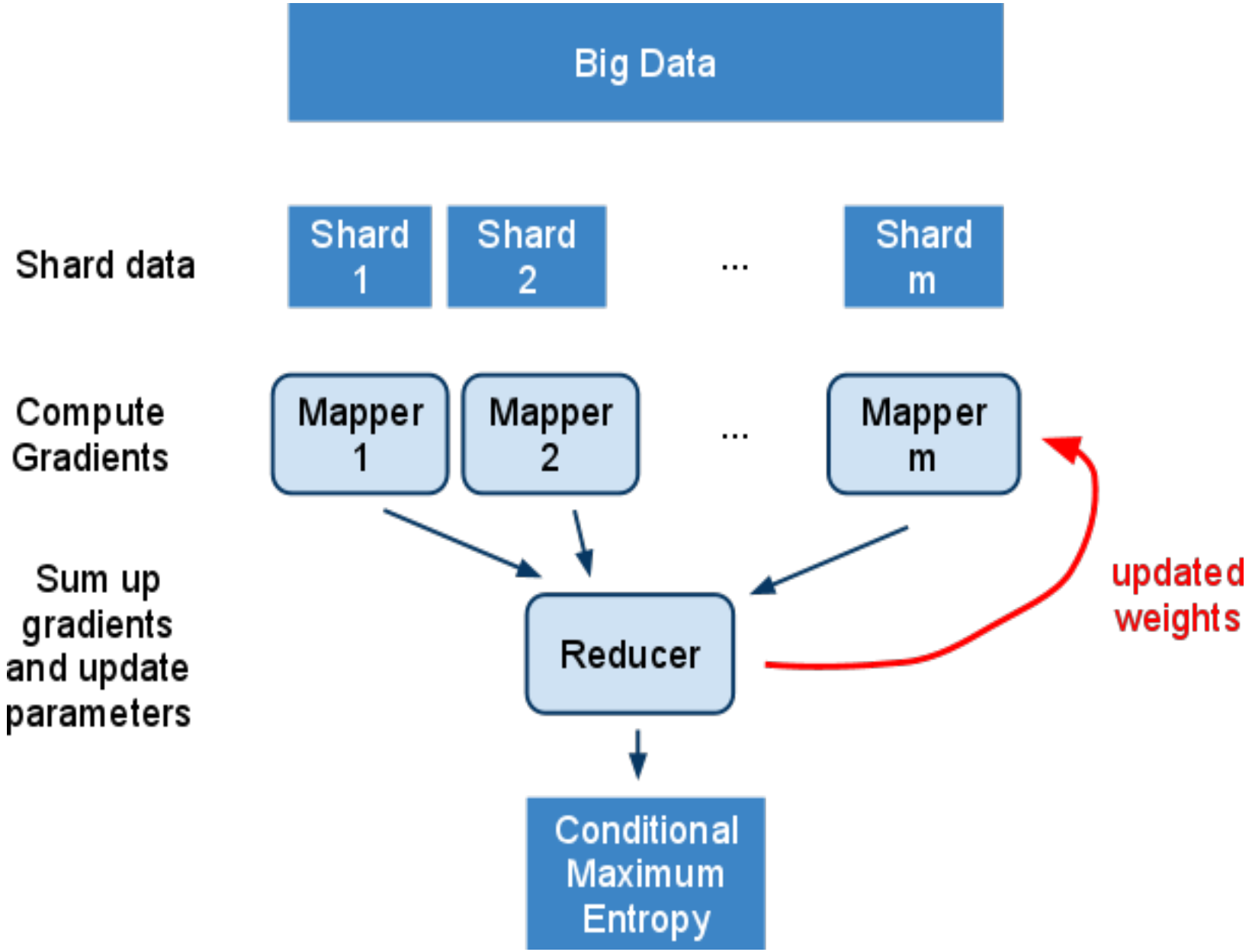
2. Distributed Naive Bayes Training

Google



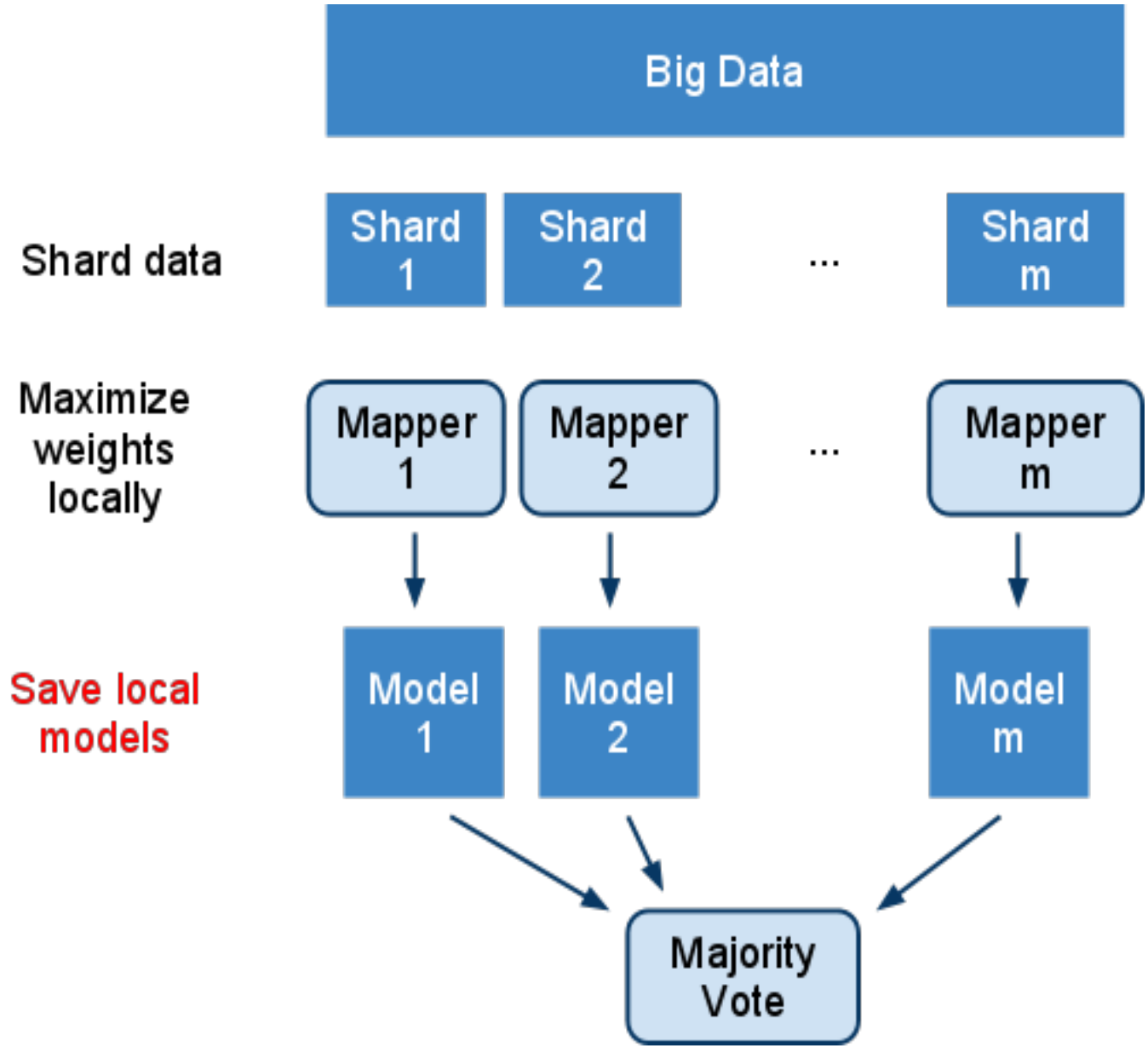
3. Distributed Gradient Descent

Google



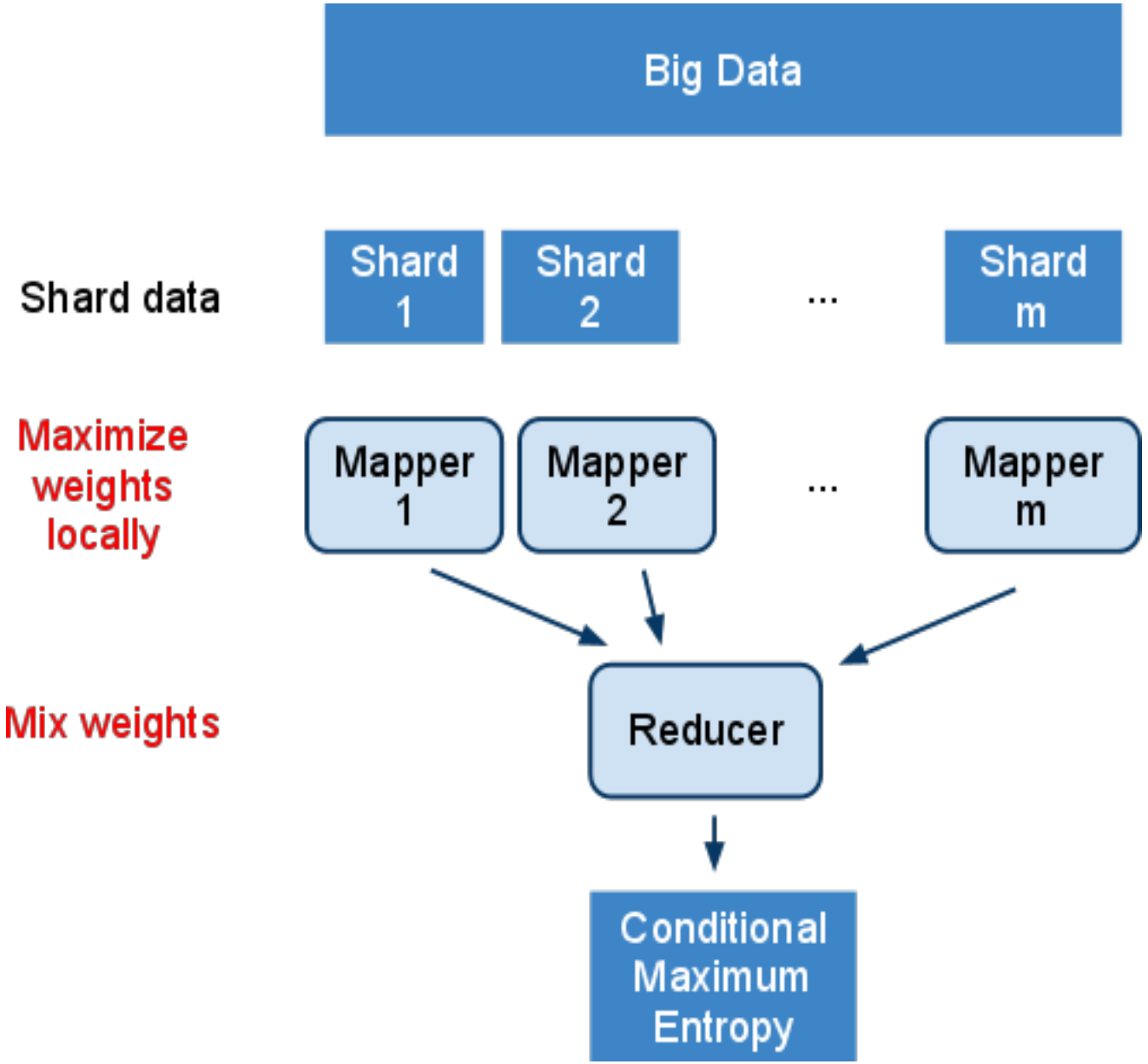
4. Majority Vote

Google



5. Parameter Mixture

Google



Compare Different Training Strategies

Google

Parameter mixture vs. distributed gradient descents

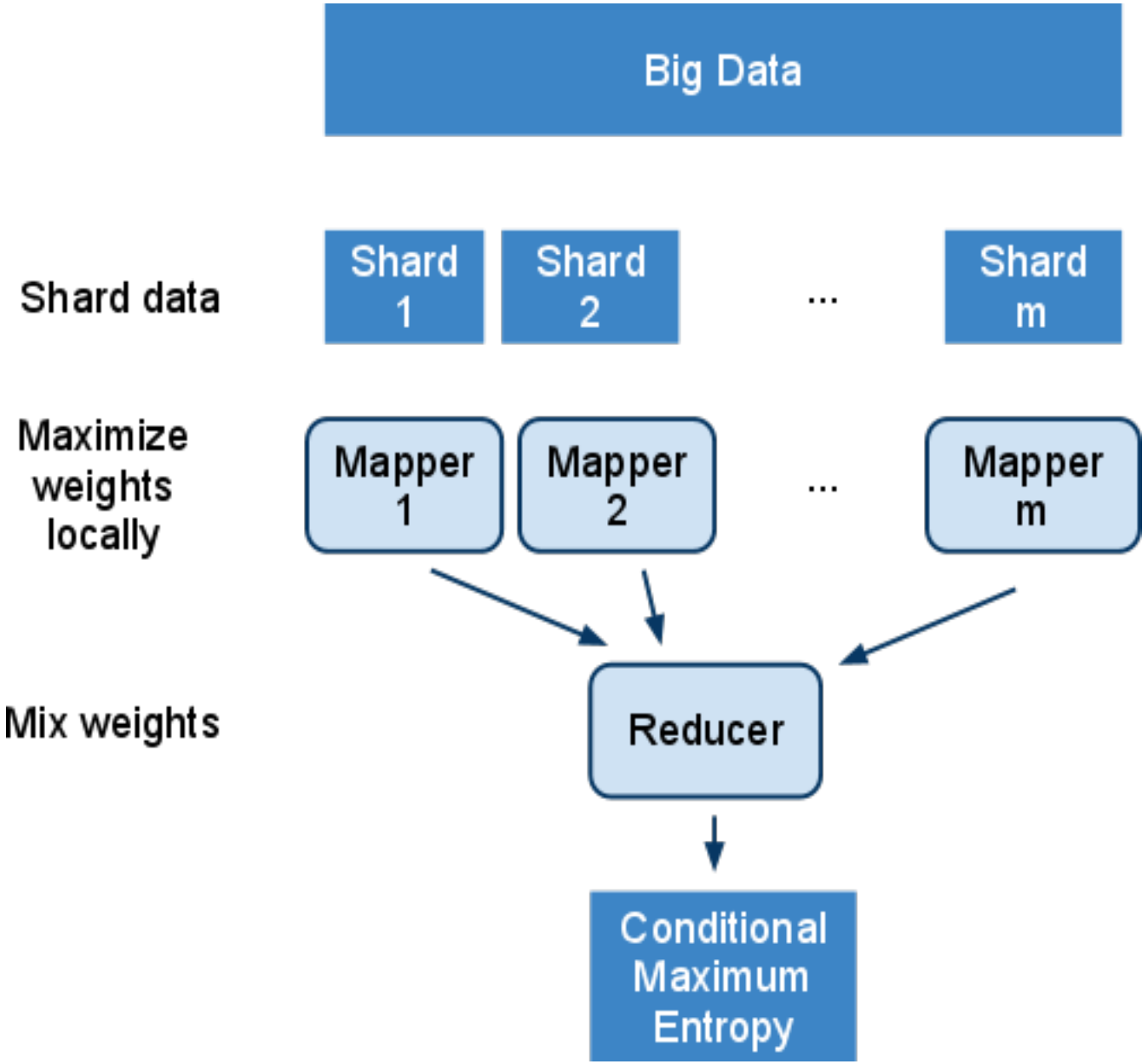
- Parameter mixture has significantly reduced network use
- Converges to the same point and at a comparable rate

On 7 data sets ranging 1M to 1B instances

- parameter mixture achieves comparable or higher accuracies
- three orders of magnitude less network usage
- modestly reduce wall clock time by 15+%

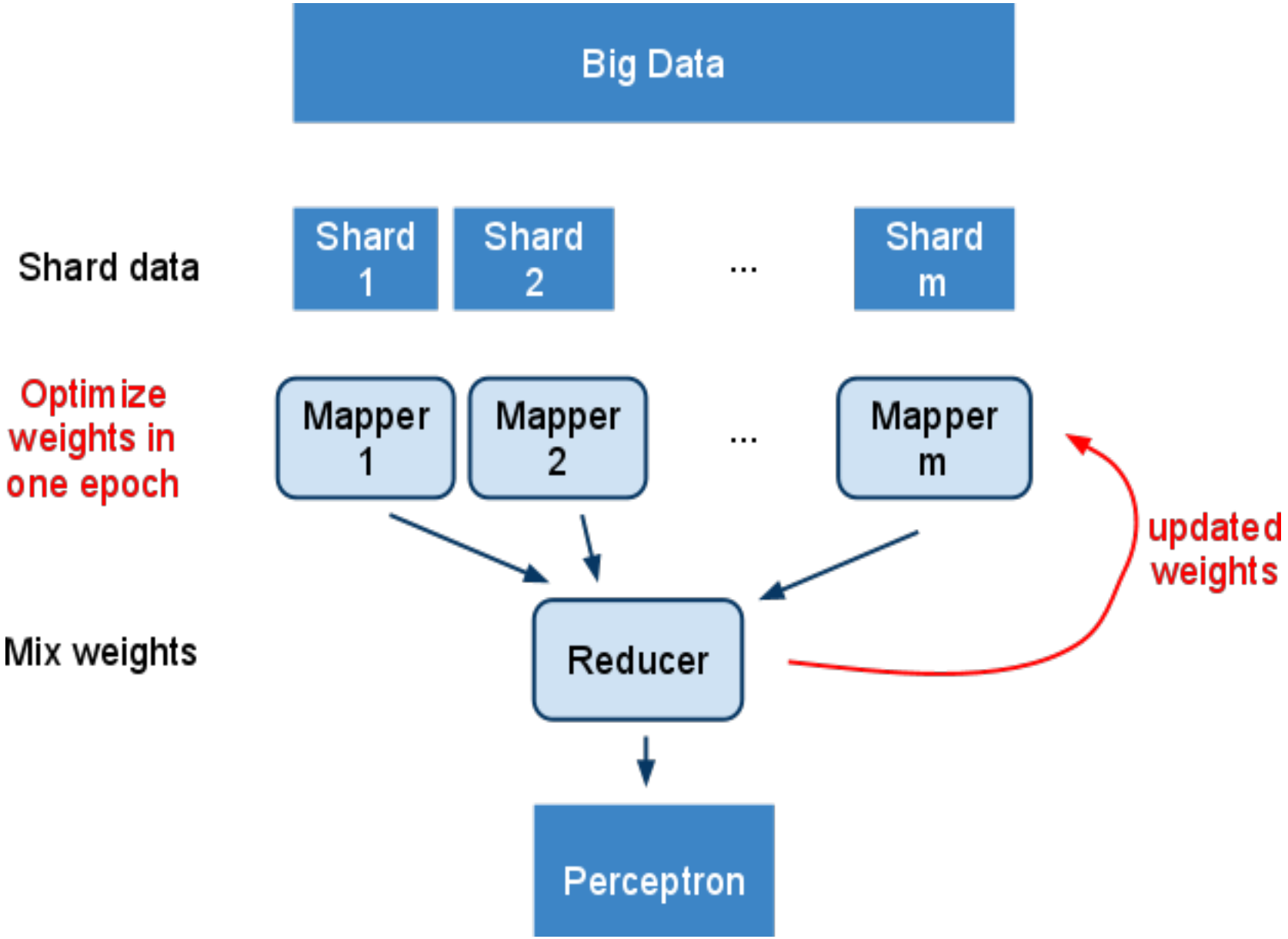
See Mann, et al., "Efficient Large-Scale Distributed Training of Conditional Maximum Entropy Models", NIPS 2009

5. Parameter Mixture Recap

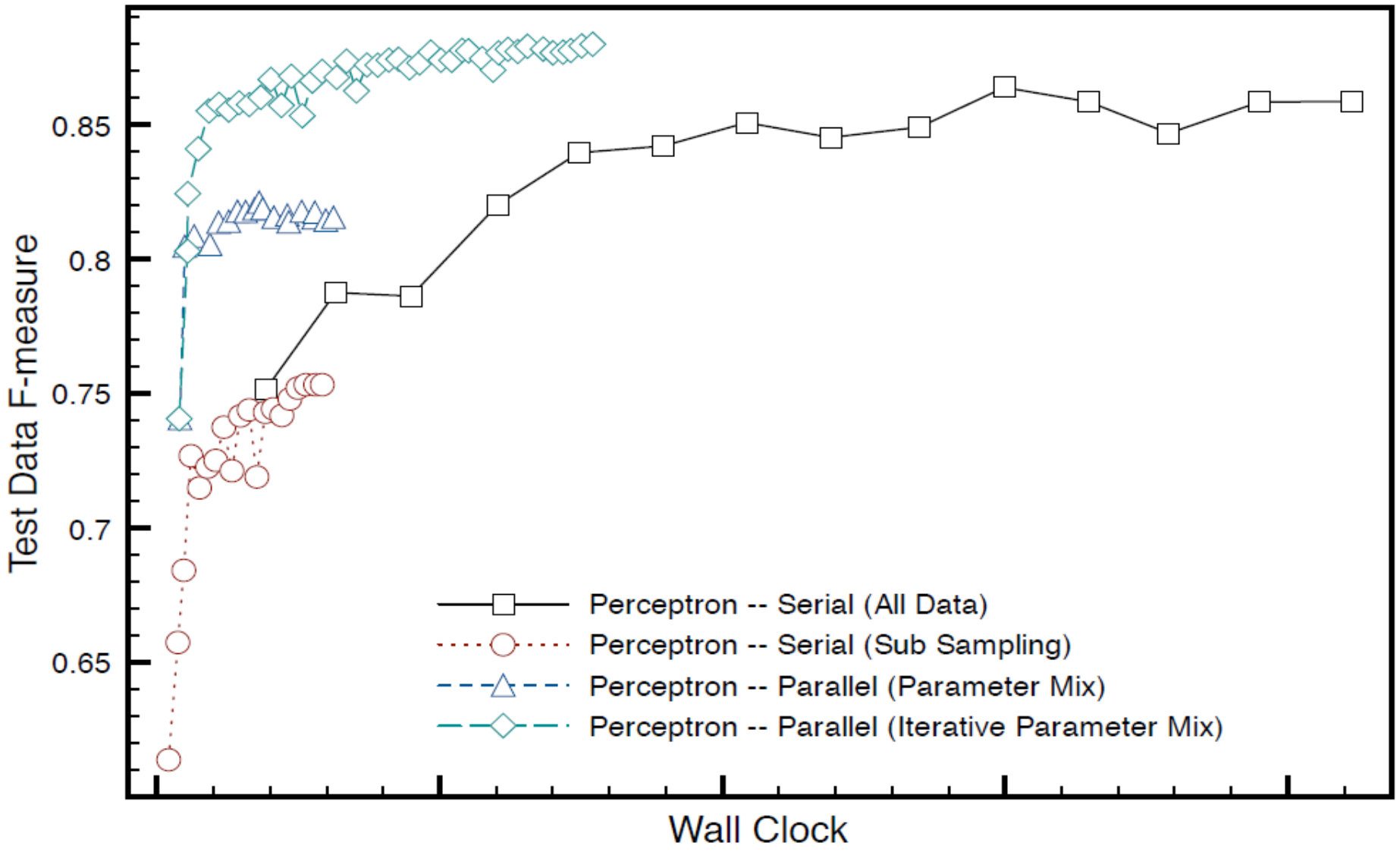


6. Iterative Parameter Mixture

Google



Compare Parameter Mixture Strategies



Compare Parameter Mixture Strategies

Google

- Sub-sampling provides inferior performance
- Parameter mixture improves, but not as good as all data
- Iterative parameter mixture achieves as good as all data
- Distributed algorithms return better classifiers quicker

See McDonald, et al., "Distributed Training Strategies for the Structured Perceptron", NAACL 2010

Google Prediction API

Google

- Machine learning as a service on the cloud
- Make ML easy to use; make every app smarter
- A web service accessible from many platforms
- Strategies for large scale machine learning

To request access and get more information about the Google Prediction API, please go to

http://code.google.com/apis/predict