

Realtime predictive analytics

using scikit-learn & RabbitMQ
Michael Becker

Who Is This Guy?

Data guy @ AWeber

@beckerfuffle

beckerfuffle.com

These slides and more @ github.com/mdbecker



What My Coworkers Think I Do

$$h_{w,b}(x) = g(w^T x + b)$$

$$\hat{\gamma}^{(i)} = y^{(i)} (w^T x + b)$$

$$\hat{\gamma} = \min_{i=1,\dots,m} \hat{\gamma}^{(i)}$$

$$w^T \left(x^{(i)} - \gamma^{(i)} \frac{w}{\|w\|} \right) + b = 0$$

What I Actually Do

```
from sklearn.svm import SVC
```


What I'll Cover

- Scikit-learn overview

What I'll Cover

- Scikit-learn overview
- Model Distribution

What I'll Cover

- Scikit-learn overview
- Model Distribution
- Data flow

What I'll Cover

- Scikit-learn overview
- Model Distribution
- Data flow
- RabbitMQ

What I'll Cover

- Scikit-learn overview
- Model Distribution
- Data flow
- RabbitMQ
- Demo

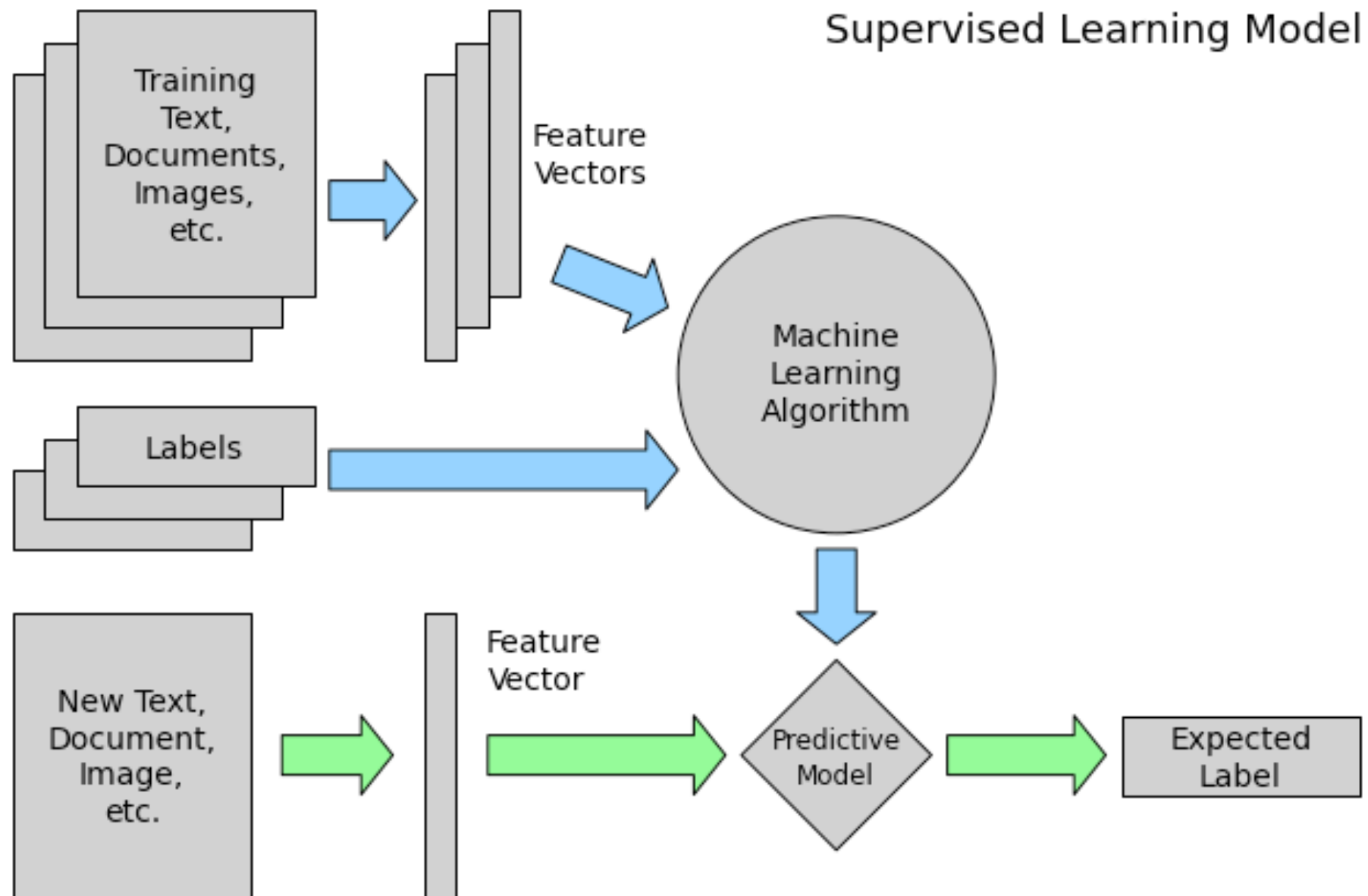
What I'll Cover

- Scikit-learn overview
- Model Distribution
- Data flow
- RabbitMQ
- Demo
- Scalability

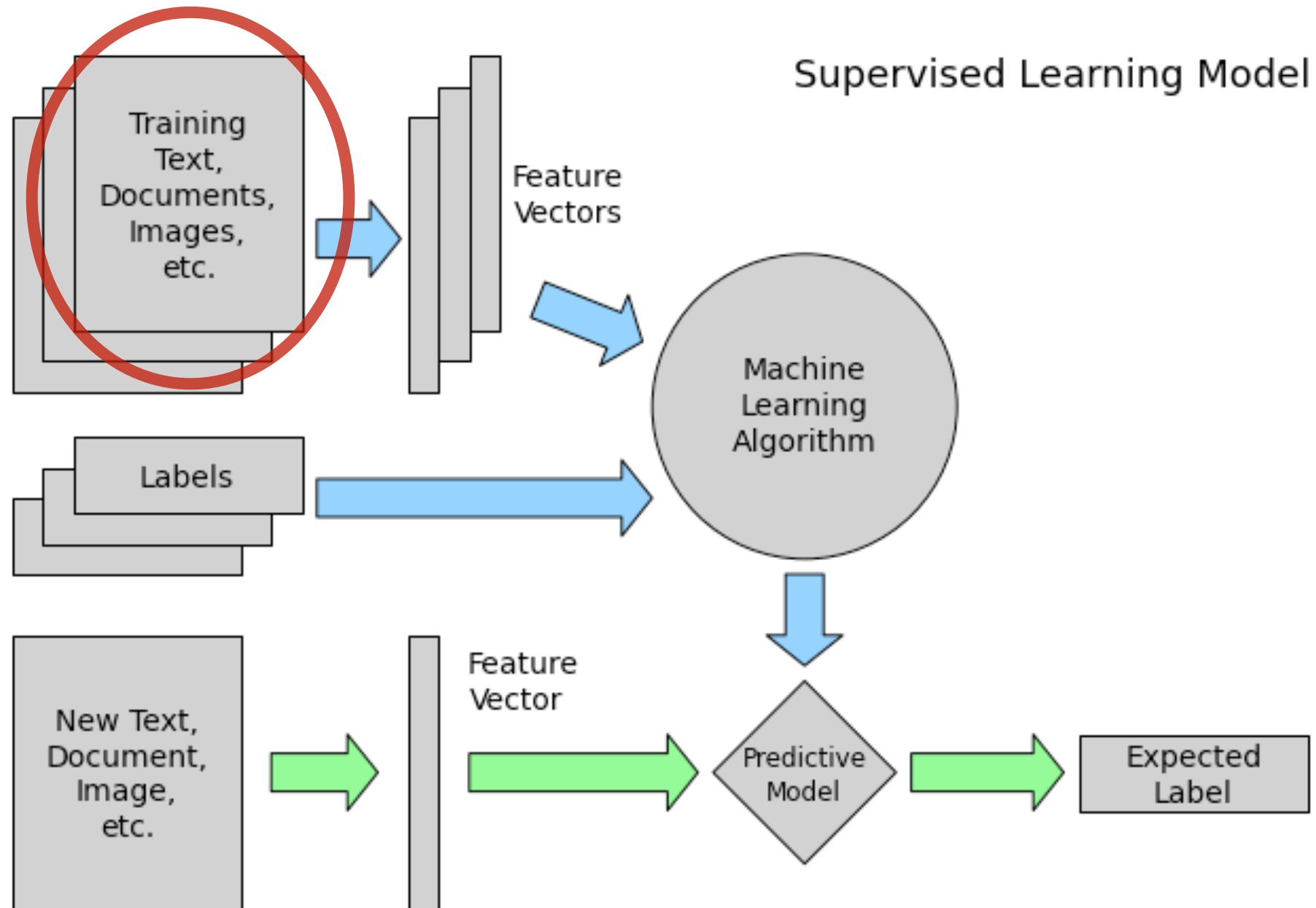
What I'll Cover

- Scikit-learn overview
- Model Distribution
- Data flow
- RabbitMQ
- Demo
- Scalability
- Other considerations

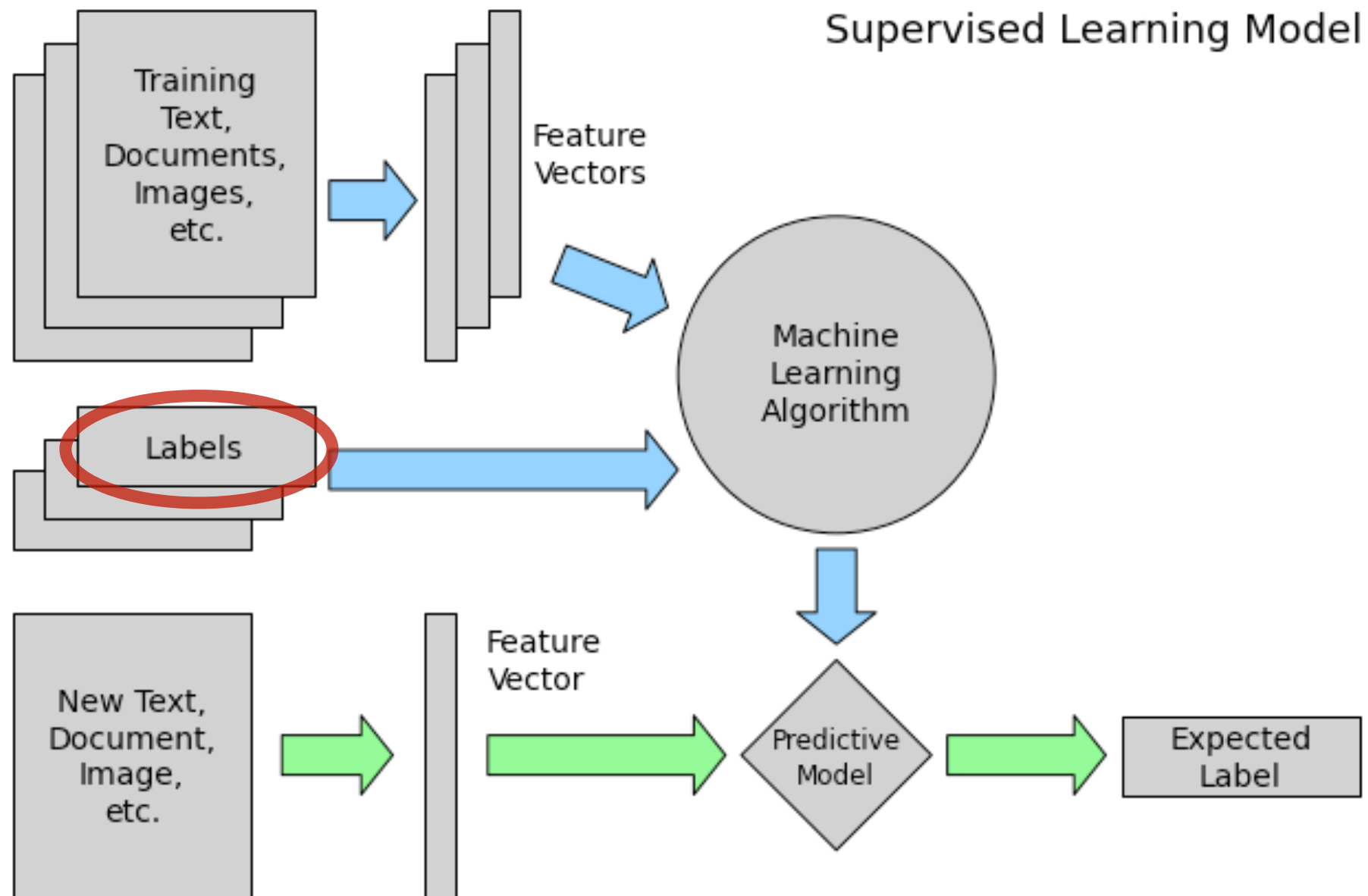
Supervised Learning



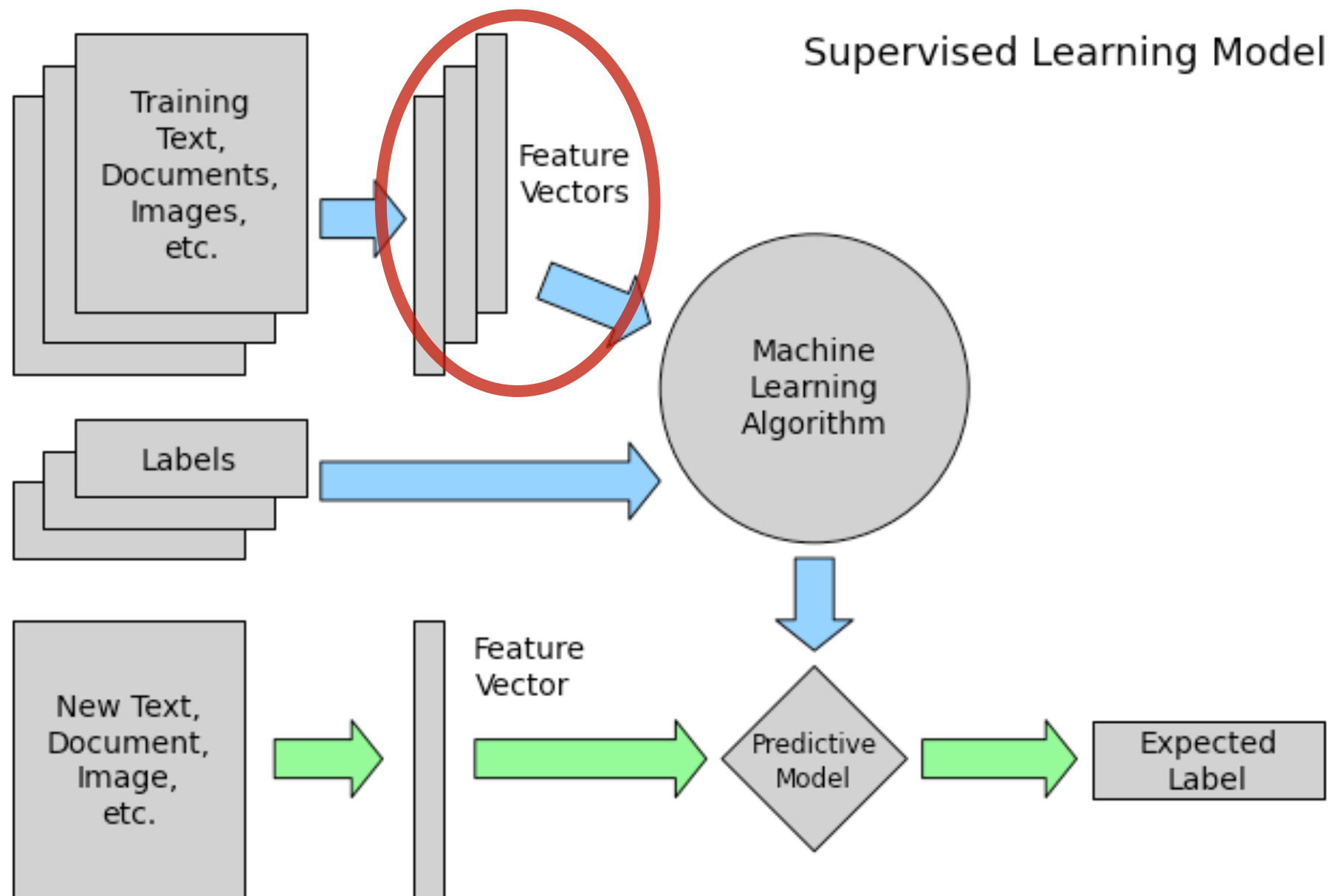
Supervised Learning



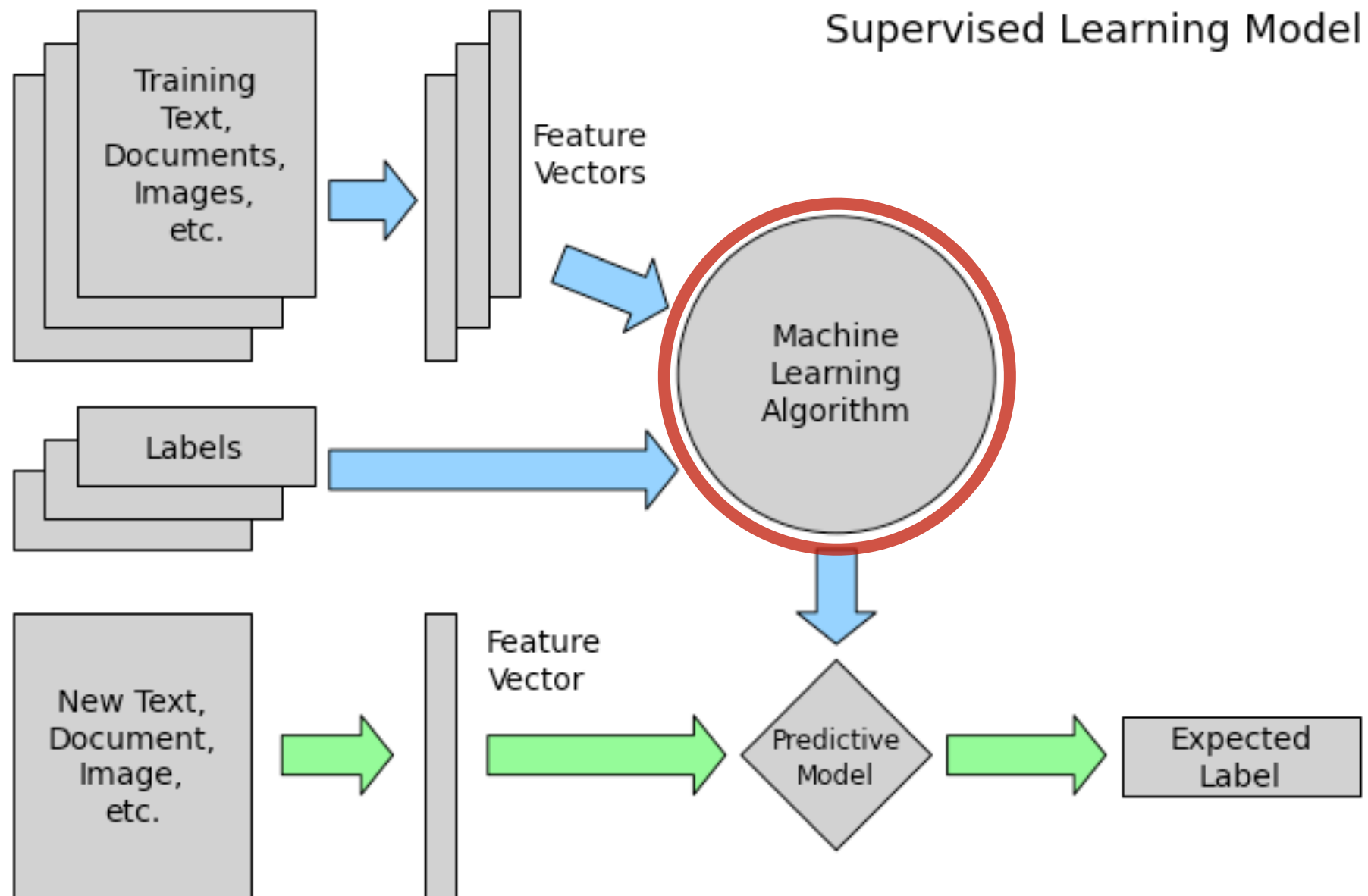
Supervised Learning



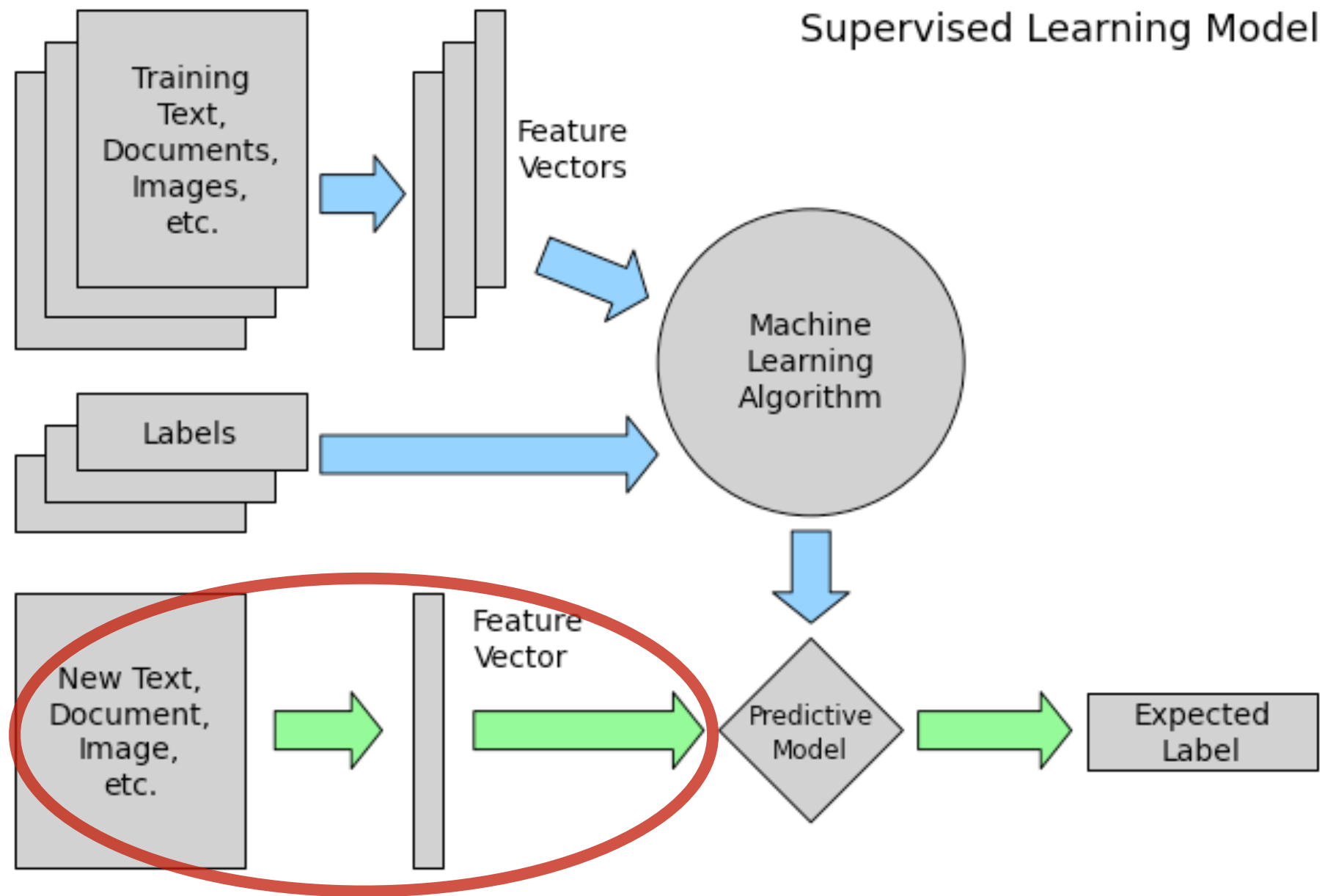
Supervised Learning



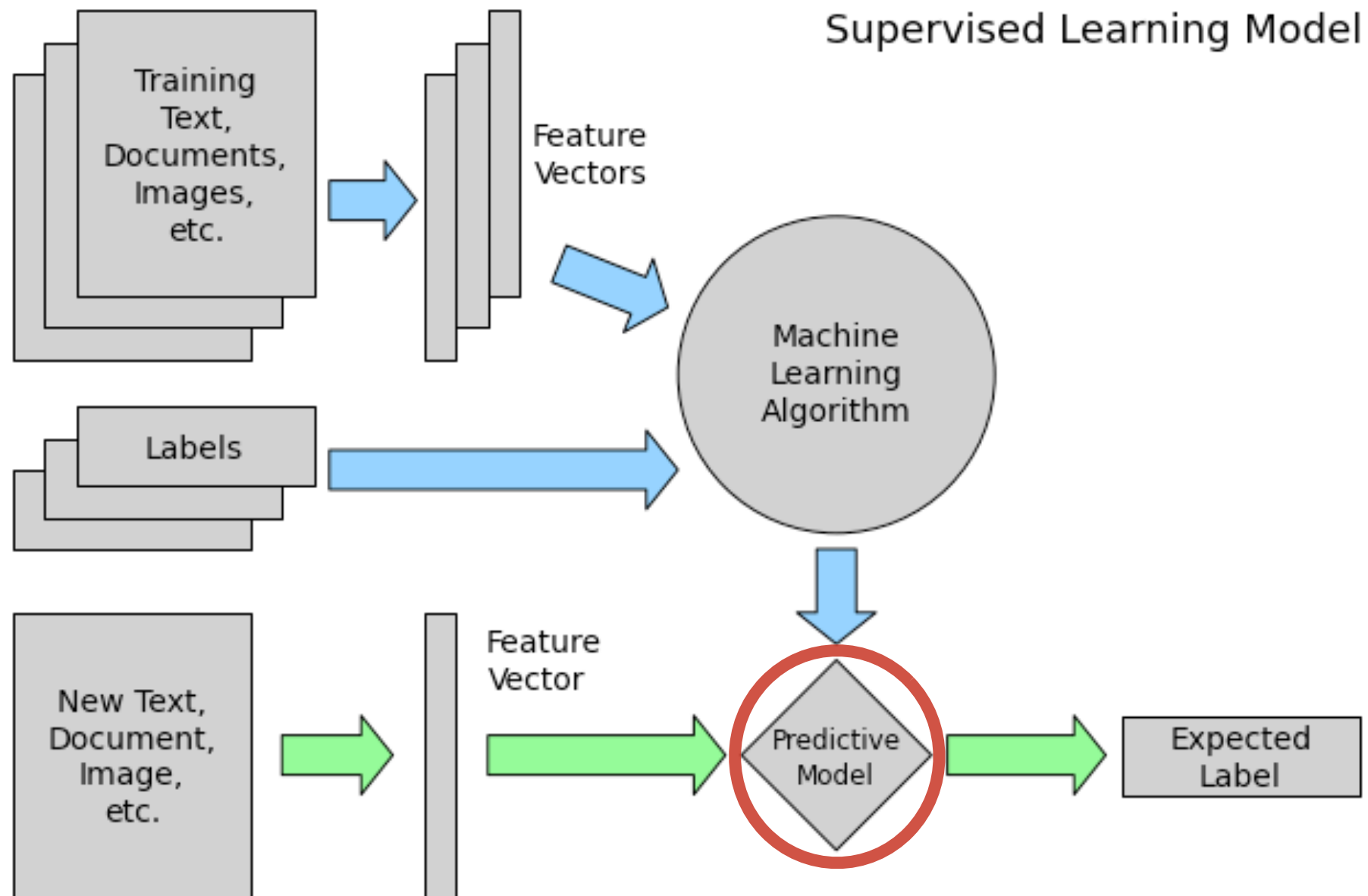
Supervised Learning



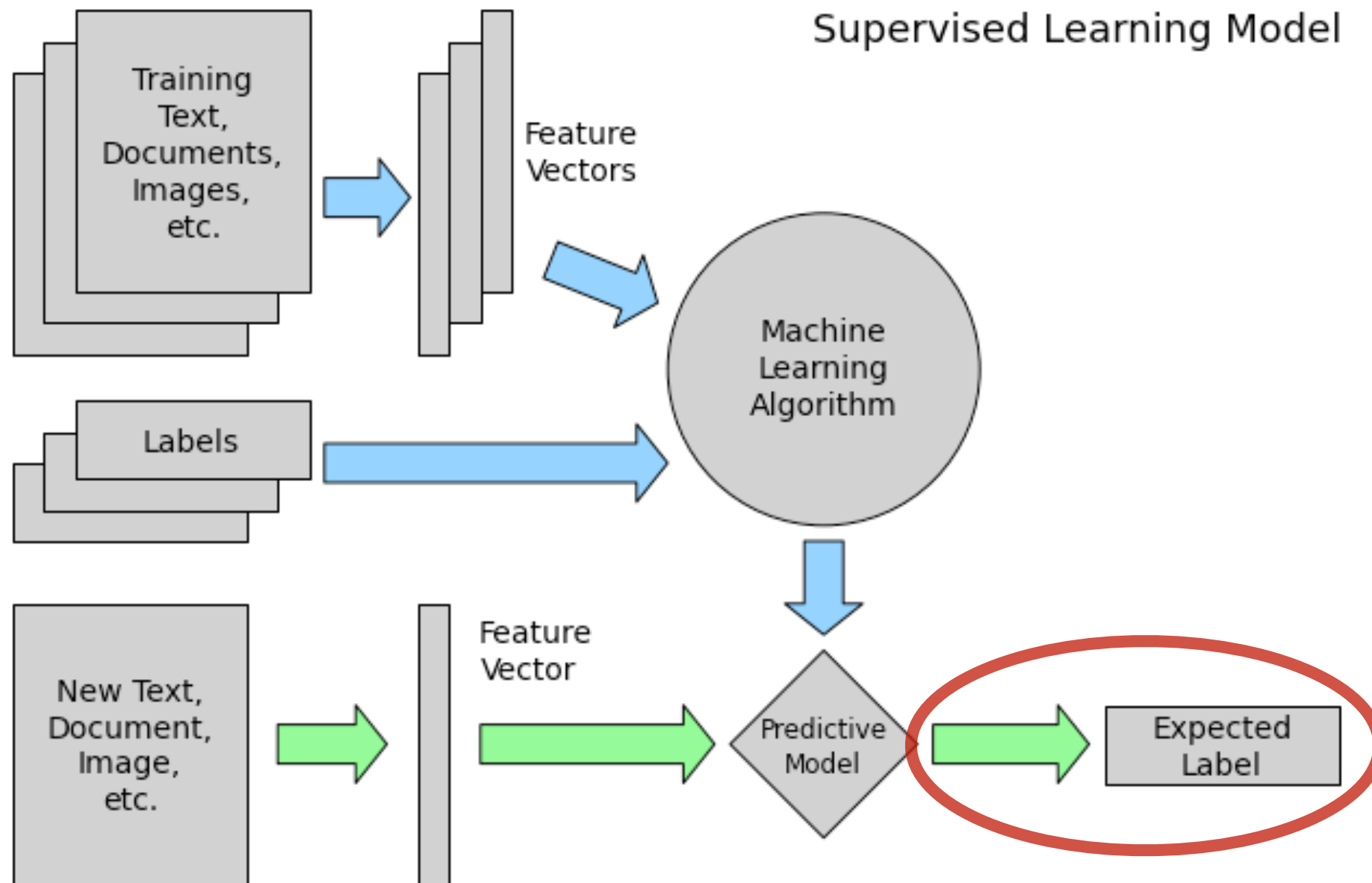
Supervised Learning



Supervised Learning



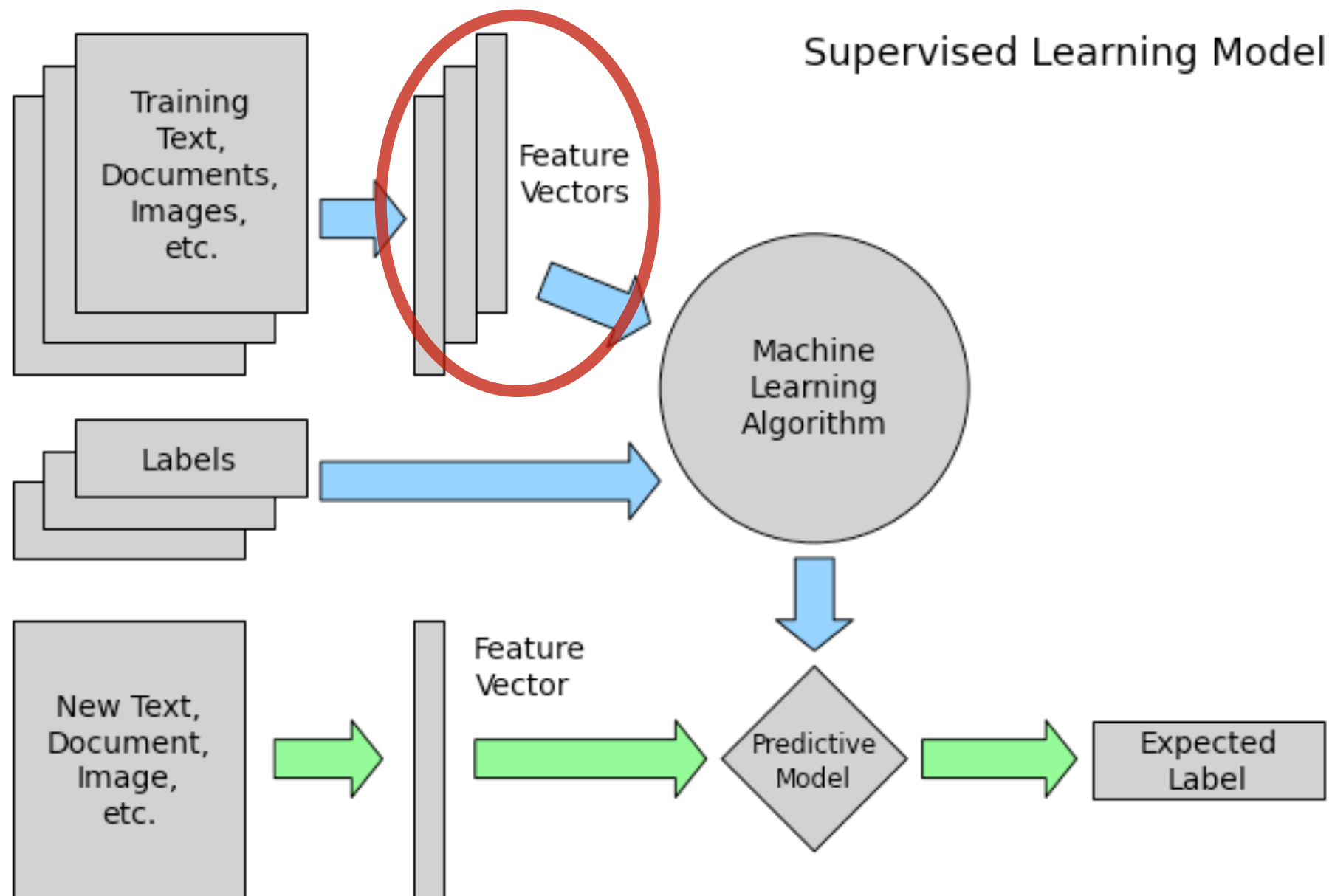
Supervised Learning



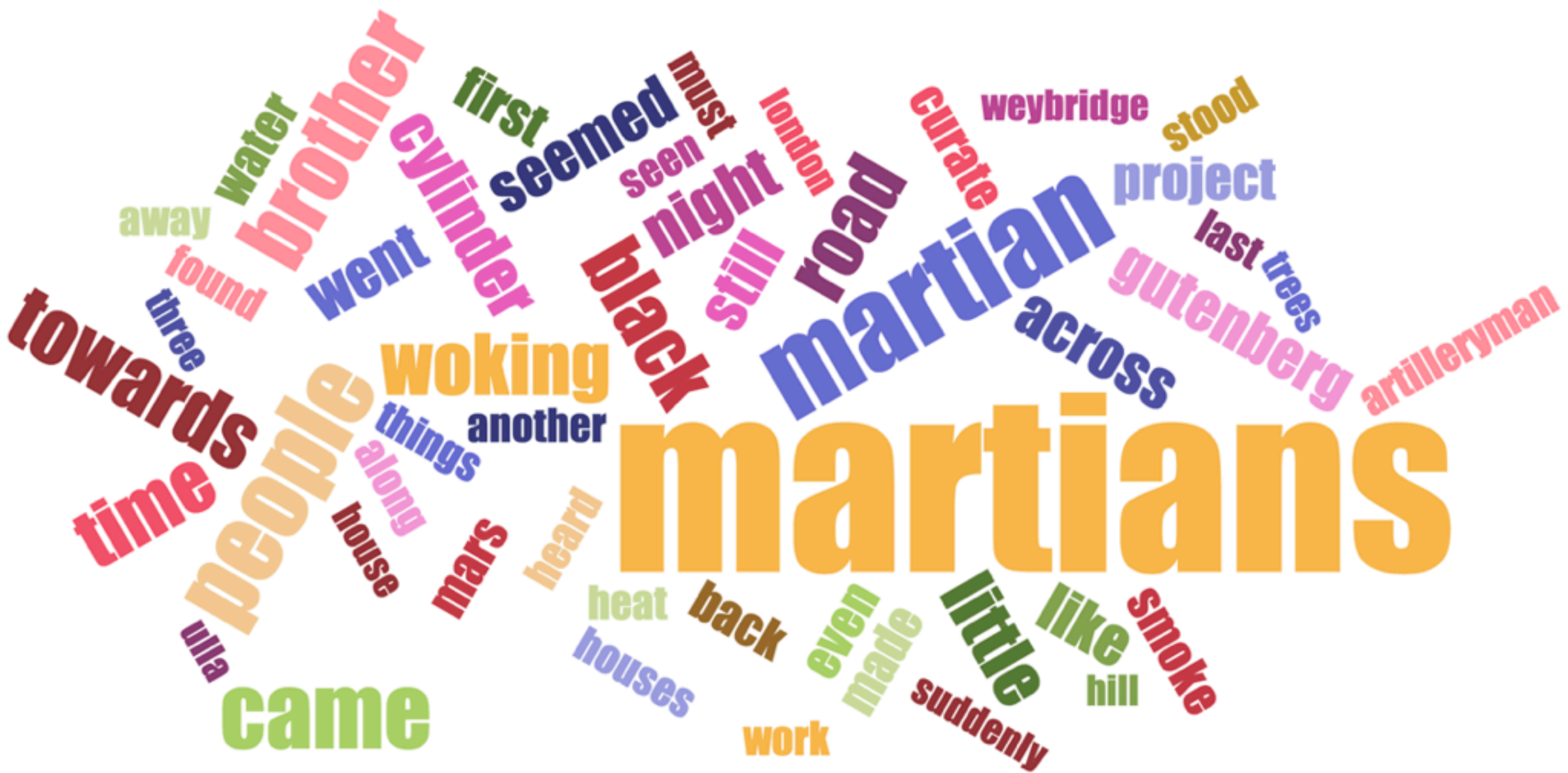
38 Top Wikipedias

Arabic: العربية	Japanese: 日本語
Bulgarian: Български	Kazakh: Қазақша
Catalan: Català	Korean: 한국어
Czech: Čeština	Lithuanian: Lietuvių
Danish: Dansk	Malay Bahasa: Melayu
German: Deutsch	Dutch: Nederlands
English: English	Norwegian: Norsk (Bokmål)
Spanish: Español	Polish: Polski
Estonian: Eesti	Portuguese: Português
Basque: Euskara	Romanian: Română
Persian: فارسی	Russian: Русский
Finnish: Suomi	Slovak: Slovenčina
French: Français	Slovenian: Slovenščina
Hebrew: עברית	Serbian: Српски / Srpski
Hindi: हिन्दी	Swedish: Svenska
Croatian: Hrvatski	Turkish: Türkçe
Hungarian: Magyar	Ukrainian: Українська
Indonesian: Bahasa Indonesia	Vietnamese: Tiếng Việt
Italian: Italiano	Waray-Waray: Winaray

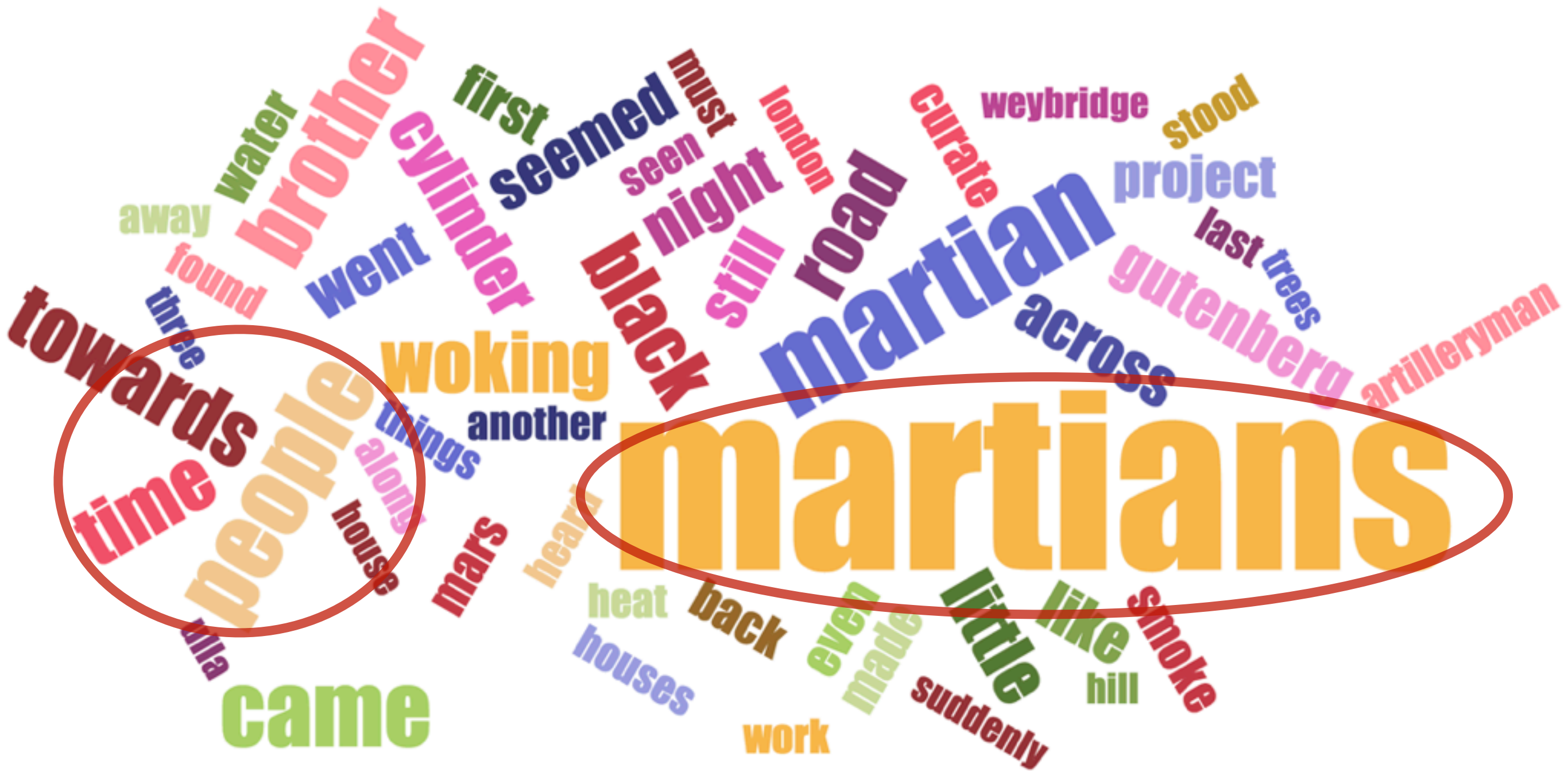
Vectorizing Text








```
TfidfVectorizer(analyzer='word', ngram_range=(1, 1))
```

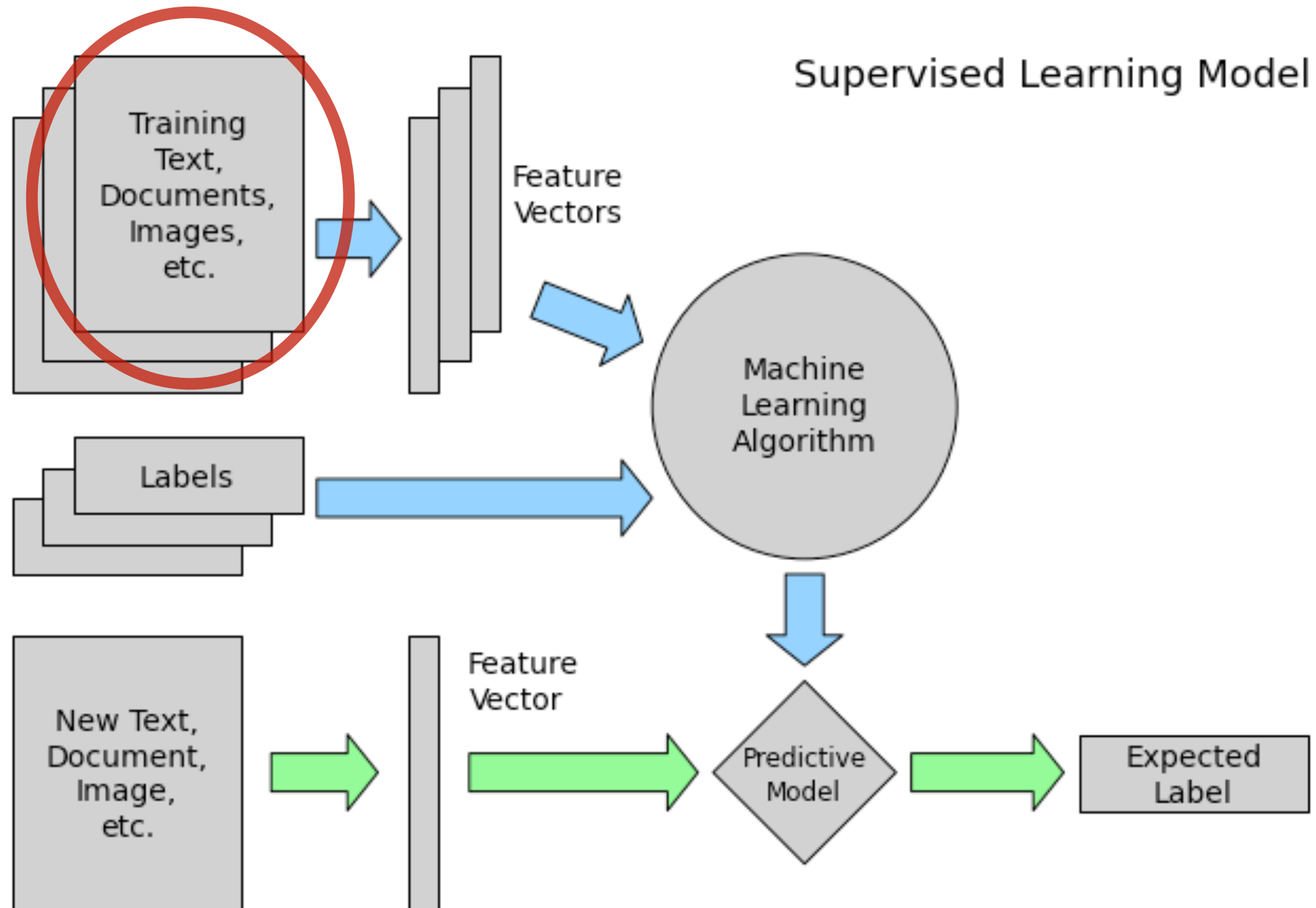


The Model

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.pipeline import Pipeline
from sklearn.svm import SVC
```

```
vect = TfidfVectorizer(analyzer='char', ngram_range=(2, 3))
clf = SVC()
text_clf = Pipeline([
    ('vect', vect),
    ('clf', clf),
])
text_clf = text_clf.fit_transform(X_train, y_train)
```

Supervised Learning

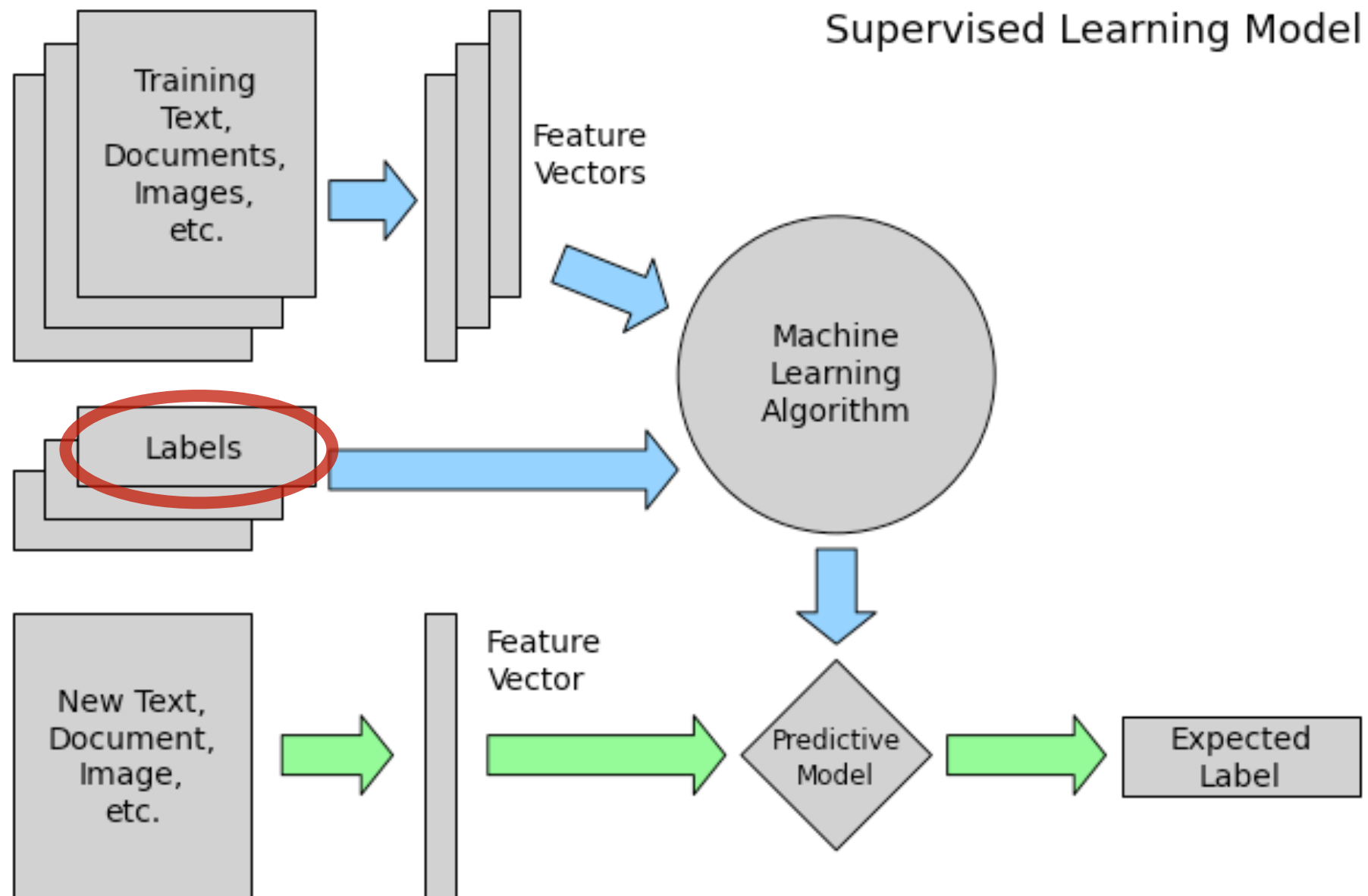


X_train

```
for x in X_train[:10]:  
    print ' '.join(unicode(x, 'utf8')[:70].split())
```

Szöul Szöul (서울 특별시 Söul T'ükpyölsi, szoros átírásban: Szoul thukpjol
Sean Connery Cecil B. DeMille Award (1996) | baftaawards Beste hov
Nemecká demokratická republika Nemecká demokratická republika (NDR, h
Plastic A plastic material is any of a wide range of synthetic or sem
Phaistose ketas Phaistose ketas on pōletatud savist ketas, mille leid
Soustava SI Soustava SI (zkratka z francouzského Le Système Internati
Баку Баку () је главни град Азербејџана. Налази се у јужном делу полу
Gottlob Frege nume Friedrich Ludwig Gottlob Frege
Belgrado Belgrado (Београд / Beograd em servo-croata ouça) é a capi
Cálculo infinitesimal El cálculo infinitesimal o cálculo de infinites

Supervised Learning



y_train

```
y
```

```
['uk', 'cs', 'ko', 'es', 'war', 'de', 'de', 'pl', 'ar', 'cs']
```

```
from sklearn.preprocessing import LabelEncoder  
le = LabelEncoder()  
y_train = le.fit_transform(y)  
y_train
```

```
array([35,  3, 21, ...,  8, 18, 30])
```


Dimensionality Reduction

```
from sklearn.decomposition import RandomizedPCA
```

```
vect = TfidfVectorizer(analyzer='char', ngram_range=(2, 3))  
pca = RandomizedPCA(n_components=50, whiten=True)  
clf = SVC()  
text_clf = Pipeline([  
    ('vect', vect),  
    ('pca', pca),  
    ('clf', clf),  
])  
text_clf = text_clf.fit_transform(X_train, y_train)
```

Dimensionality Reduction

```
vect = TfidfVectorizer(analyzer='char', ngram_range=(2, 3))  
X1_train = vect.fit_transform(X_train)  
X1_train.shape
```

```
(47221, 1048576)
```


Dimensionality Reduction

```
vect = TfidfVectorizer(analyzer='char', ngram_range=(2, 3))  
X1_train = vect.fit_transform(X_train)  
X1_train.shape
```

```
(47221, 1048576)
```

```
pca = RandomizedPCA(n_components=50, whiten=True)  
X2_train = pca.fit_transform(X1_train)  
X2_train.shape
```

```
(47221, 50)
```

Dimensionality Reduction

```
pca.explained_variance_ratio_
```

```
array([ 0.36366957,  0.0984124 ,  0.04520123,  0.04459129,  0.03954548,  
        0.03202782,  0.02943691,  0.02429984,  0.01907511,  0.01890865,  
        0.01858178,  0.0171858 ,  0.016482   ,  0.01596849,  0.01584665,  
        0.01499711,  0.01378617,  0.01368058,  0.01190577,  0.01173788,  
        0.01131731,  0.01045798,  0.01029921,  0.01004662,  0.00951628,  
        0.00803825,  0.00728402,  0.00700877,  0.00667922,  0.00664124,  
        0.00574287,  0.00548161,  0.00500291,  0.00433605,  0.003948   ,  
        0.00283822,  0.00241628,  0.00189858,  0.00168776,  0.00160769,  
        0.001431   ,  0.00137901,  0.00132682,  0.00129177,  0.00126926,  
        0.0012235  ,  0.00117239,  0.00114645,  0.00110811,  0.00106229])
```


Dimensionality Reduction

```
pca.explained_variance_ratio_
```

```
array([ 0.36366957,  0.0984124 ,  0.04520123,  0.04459129,  0.03954548,  
        0.03202782,  0.02943691,  0.02429984,  0.01907511,  0.01890865,  
        0.01858178,  0.0171858 ,  0.016482   ,  0.01596849,  0.01584665,  
        0.01499711,  0.01378617,  0.01368058,  0.01190577,  0.01173788,  
        0.01131731,  0.01045798,  0.01029921,  0.01004662,  0.00951628,  
        0.00803825,  0.00728402,  0.00700877,  0.00667922,  0.00664124,  
        0.00574287,  0.00548161,  0.00500291,  0.00433605,  0.003948   ,  
        0.00283822,  0.00241628,  0.00189858,  0.00168776,  0.00160769,  
        0.001431   ,  0.00137901,  0.00132682,  0.00129177,  0.00126926,  
        0.0012235 ,  0.00117239,  0.00114645,  0.00110811,  0.00106229])
```

```
pca.explained_variance_ratio_.sum()
```

```
0.9999999999999999
```


Don't Do What I Just Did!

sklearn.decomposition.TruncatedSVD

```
class sklearn.decomposition.TruncatedSVD(n_components=2, algorithm='randomized', n_iter=5,  
random_state=None, tol=0.0)
```

Dimensionality reduction using truncated SVD (aka LSA).

This transformer performs linear dimensionality reduction by means of truncated singular value decomposition (SVD). It is very similar to PCA, but operates on sample vectors directly, instead of on a covariance matrix. This means it can work with `scipy.sparse` matrices efficiently.

In particular, truncated SVD works on term count/tf-idf matrices as returned by the vectorizers in `sklearn.feature_extraction.text`. In that context, it is known as latent semantic analysis (LSA).

Tuning

sklearn.svm.SVC

```
class sklearn.svm.SVC(C=1.0, kernel='rbf', degree=3, gamma=0.0, coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1, random_state=None) ¶
```

C-Support Vector Classification.

The implementation is based on libsvm. The fit time complexity is more than quadratic with the number of samples which makes it hard to scale to dataset with more than a couple of 10000 samples.

The multiclass support is handled according to a one-vs-one scheme.

For details on the precise mathematical formulation of the provided kernel functions and how gamma, coef0 and degree affect each, see the corresponding section in the narrative documentation: [Kernel functions](#).

RandomizedSearchCV

```
from sklearn.grid_search import RandomizedSearchCV
from scipy.stats import uniform as sp_uniform
```

```
clf = SVC()
param_dist = {
    'C': sp_uniform(1, 4),
    'kernel': ['linear', 'rbf'],
}

random_search = RandomizedSearchCV(
    clf, param_distributions=param_dist, n_jobs=9)
random_search.fit(X_train, y_train)
```

RandomizedSearchCV

```
random_search.best_score_
```

```
0.96647677939899623
```

```
random_search.best_params_
```

```
{'C': 3.3357996909950347, 'kernel': 'rbf'}
```


Distributing The Model



Data Input

The Client

Language Prediction

Input

Result

Message Loss



Enter RabbitMQ

Reliability

Flexible Routing

Clustering



RabbitMQ

HA Queues

Many clients

Enter RabbitMQ

Reliability

Flexible Routing

Clustering



RabbitMQ

HA Queues

Many clients

Using the Blocking Connection to consume messages from RabbitMQ

The `BlockingChannel.basic_consume` method assign a callback method to be called every time that RabbitMQ delivers messages to your consuming application.

When pika calls your method, it will pass in the channel, a `pika.spec.Basic.Deliver` object with the delivery tag, the redelivered flag, the routing key that was used to put the message in the queue, and the exchange the message was published to. The third argument will be a `pika.spec.BasicProperties` object and the last will be the message body.

Example of consuming messages and acknowledging them:

```
import pika

def on_message(channel, method_frame, header_frame, body):
    print method_frame.delivery_tag
    print body
    print
    channel.basic_ack(delivery_tag=method_frame.delivery_tag)

connection = pika.BlockingConnection()
channel = connection.channel()
channel.basic_consume(on_message, 'test')
try:
    channel.start_consuming()
except KeyboardInterrupt:
    channel.stop_consuming()
```

These tutorials cover the basics of creating messaging applications using RabbitMQ. You need to have the RabbitMQ server installed to run the tutorials – please see the **installation guide**.

1 "Hello World!"

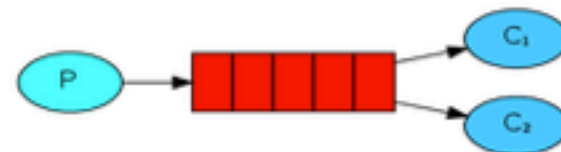
The simplest thing that does *something*



Python | **Java** | **Ruby** | **PHP**
| **C#**

2 Work queues

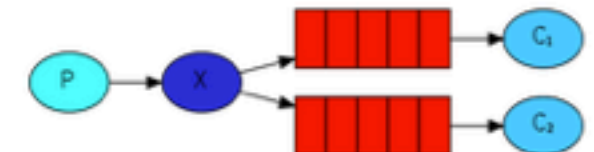
Distributing tasks among workers



Python | **Java** | **Ruby** | **PHP**
| **C#**

3 Publish/Subscribe

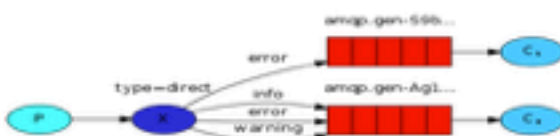
Sending messages to many consumers at once



Python | **Java** | **Ruby** | **PHP**
| **C#**

4 Routing

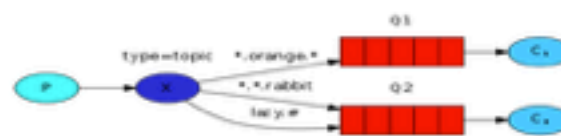
Receiving messages selectively



Python | **Java** | **Ruby** | **PHP**
| **C#**

5 Topics

Receiving messages based on a pattern



Python | **Java** | **Ruby** | **PHP**
| **C#**

6 RPC

Remote procedure call implementation



Python | **Java** | **Ruby** | **PHP**
| **C#**



RabbitMQ

IN DEPTH

Gavin M. Roy

All the things!



Storm

Distributed and fault-tolerant realtime computation

beanstalkd



ØMQ

Data Processing



#pycon2014

@beckerfuffle

The Consumer

```
class LanguagePredictorWorker(object):

    classifier, label_encoder = load_pickled_files(clf)
    def __init__(self):
        subscribe_to_queue()
        self.language_coll = get_db_collection()

    def process_event(self, body, message):
        text = body['text']
        _id = body['id']
        language = self.predict_language_for_text(text)
        result = self.language_coll.update(
            {'_id': ObjectId(_id), 'text_input': text},
            {'$set': {'language': language}},
        )
        message.ack()

    def predict_language_for_text(self, text):
        lang_vector = self.classifier.predict([text])
        lang_labels = self.label_encoder.inverse_transform(lang_vector)
        return lang_labels[0]

worker = LanguagePredictorWorker()
worker.main()
```

The Consumer

```
class LanguagePredictorWorker(object):  
  
    classifier, label_encoder = load_pickled_files(clf)  
    def __init__(self):  
        subscribe_to_queue()  
        self.language_coll = get_db_collection()  
  
    def process_event(self, body, message):  
        text = body['text']  
        _id = body['id']  
        language = self.predict_language_for_text(text)  
        result = self.language_coll.update(  
            {'_id': ObjectId(_id), 'text_input': text},  
            {'$set': {'language': language}},  
        )  
        message.ack()  
  
    def predict_language_for_text(self, text):  
        lang_vector = self.classifier.predict([text])  
        lang_labels = self.label_encoder.inverse_transform(lang_vector)  
        return lang_labels[0]  
  
worker = LanguagePredictorWorker()  
worker.main()
```


The Consumer

```
class LanguagePredictorWorker(object):

    classifier, label_encoder = load_pickled_files(clf)
    def __init__(self):
        subscribe_to_queue()
        self.language_coll = get_db_collection()

    def process_event(self, body, message):
        text = body['text']
        _id = body['id']
        language = self.predict_language_for_text(text)
        result = self.language_coll.update(
            {'_id': ObjectId(_id), 'text_input': text},
            {'$set': {'language': language}},
        )
        message.ack()

    def predict_language_for_text(self, text):
        lang_vector = self.classifier.predict([text])
        lang_labels = self.label_encoder.inverse_transform(lang_vector)
        return lang_labels[0]
```

```
worker = LanguagePredictorWorker()
worker.main()
```


The Consumer

```
class LanguagePredictorWorker(object):

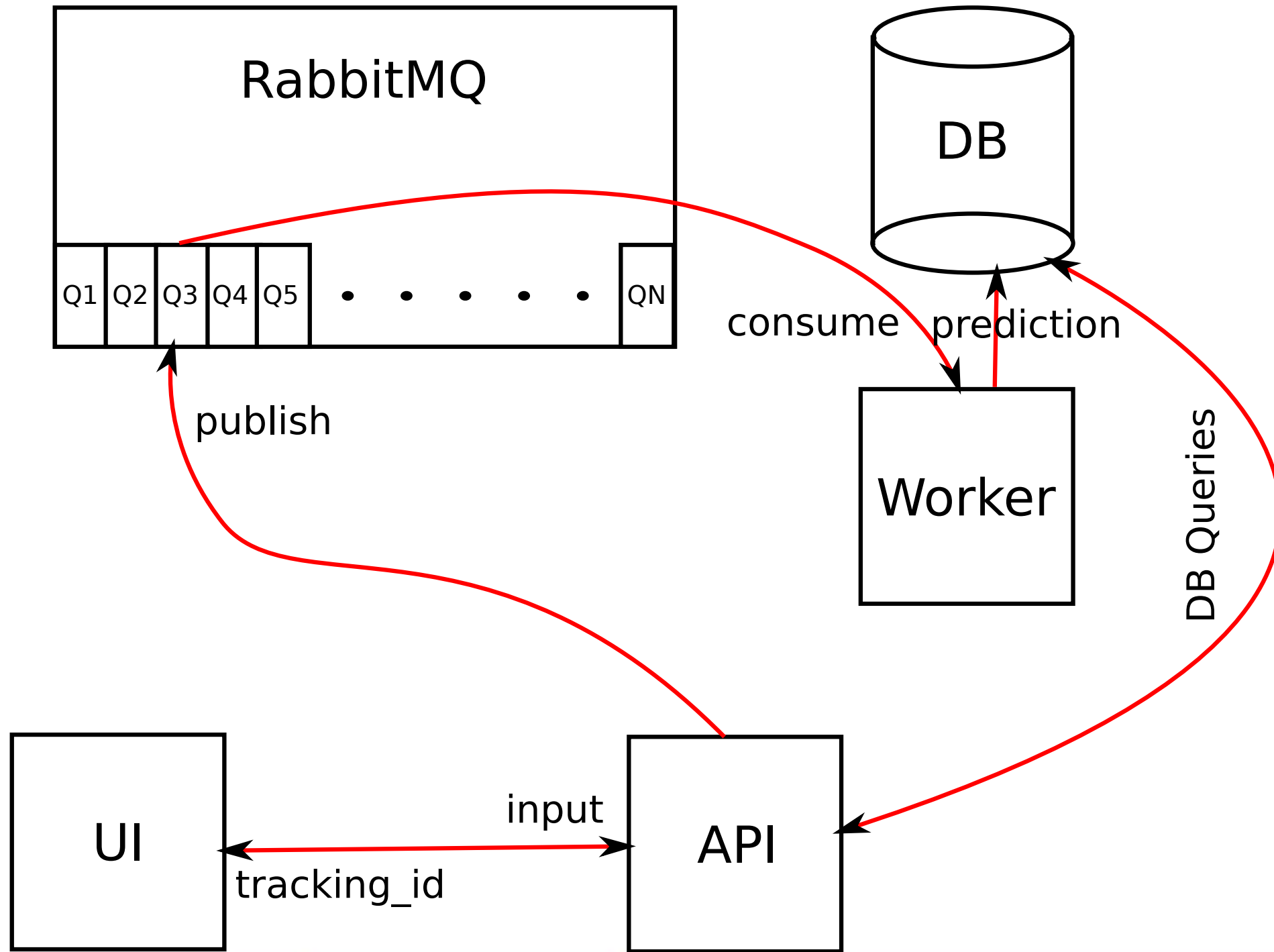
    classifier, label_encoder = load_pickled_files(clf)
    def __init__(self):
        subscribe_to_queue()
        self.language_coll = get_db_collection()

    def process_event(self, body, message):
        text = body['text']
        _id = body['id']
        language = self.predict_language_for_text(text)
        result = self.language_coll.update(
            {'_id': ObjectId(_id), 'text_input': text},
            {'$set': {'language': language}},
        )
        message.ack()

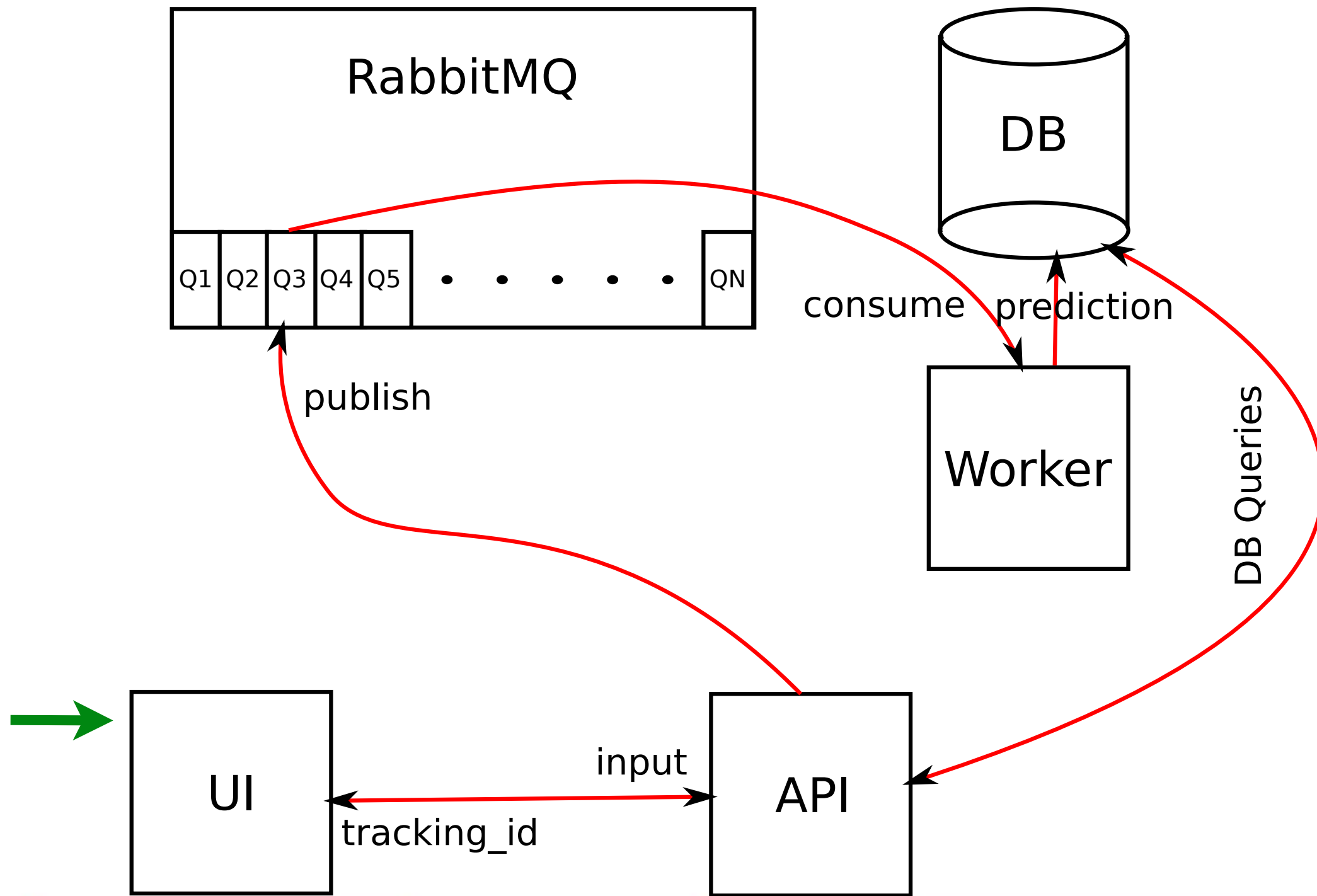
    def predict_language_for_text(self, text):
        lang_vector = self.classifier.predict([text])
        lang_labels = self.label_encoder.inverse_transform(lang_vector)
        return lang_labels[0]

worker = LanguagePredictorWorker()
worker.main()
```

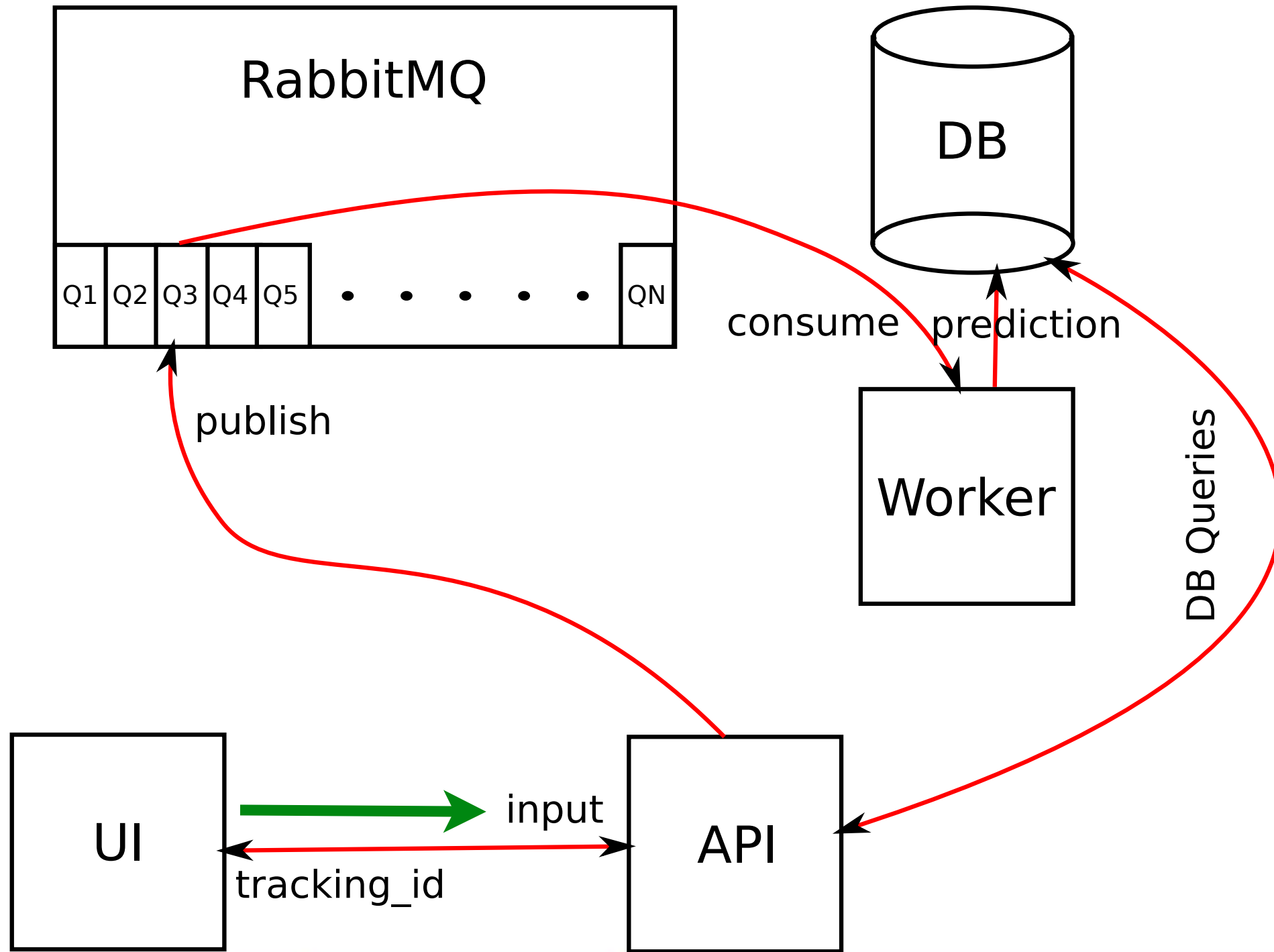
The Design



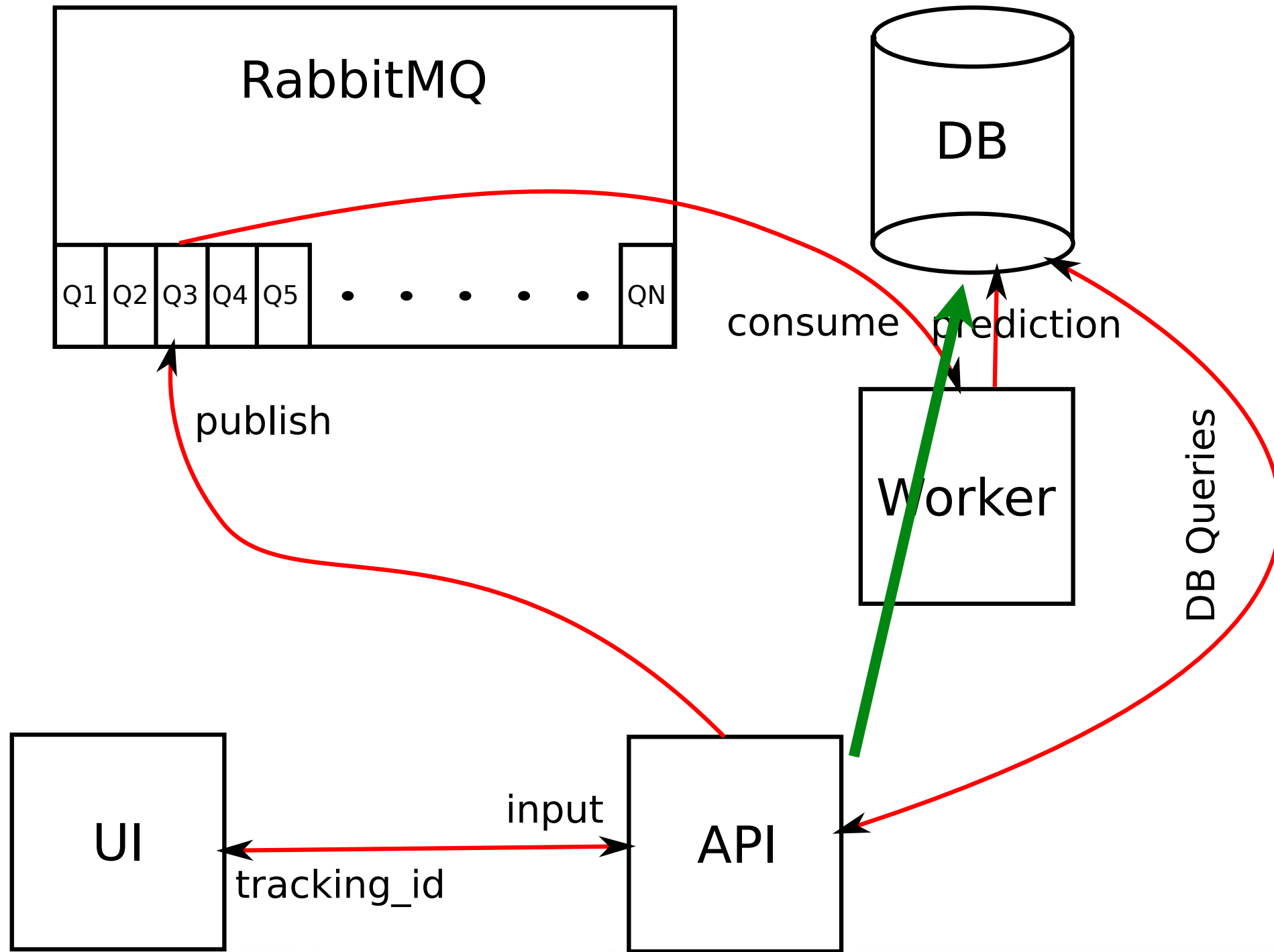
The Design



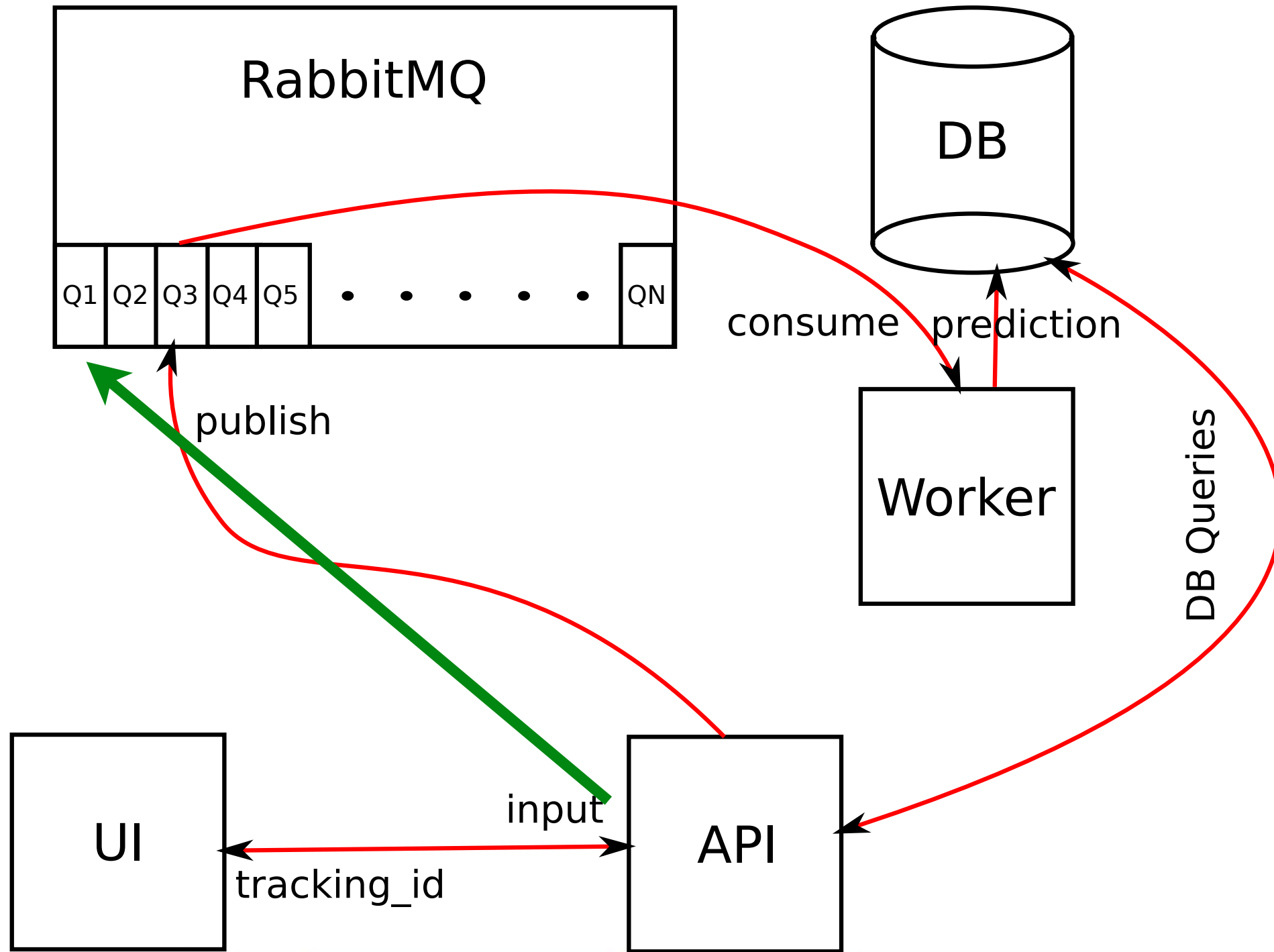
The Design



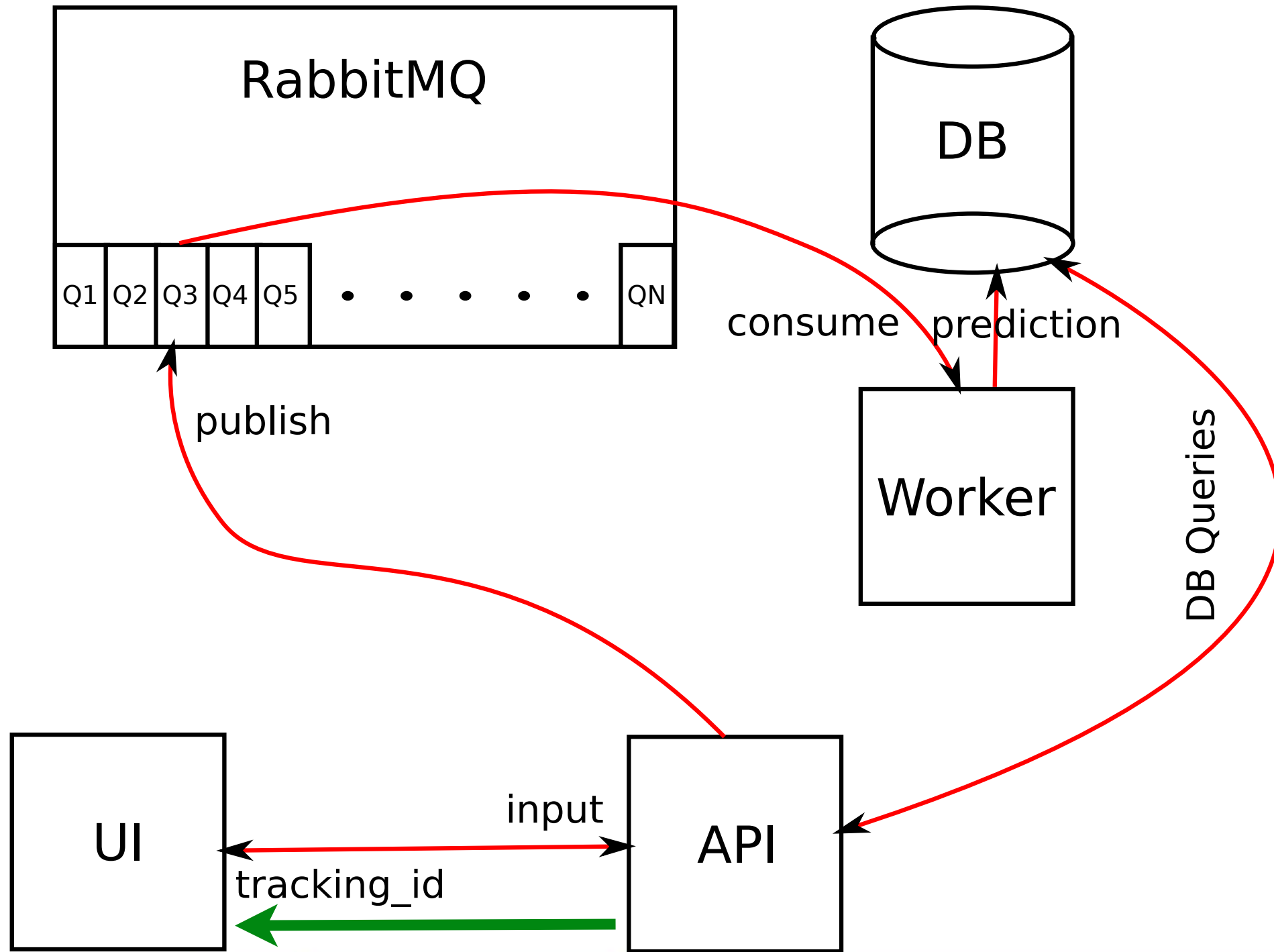
The Design



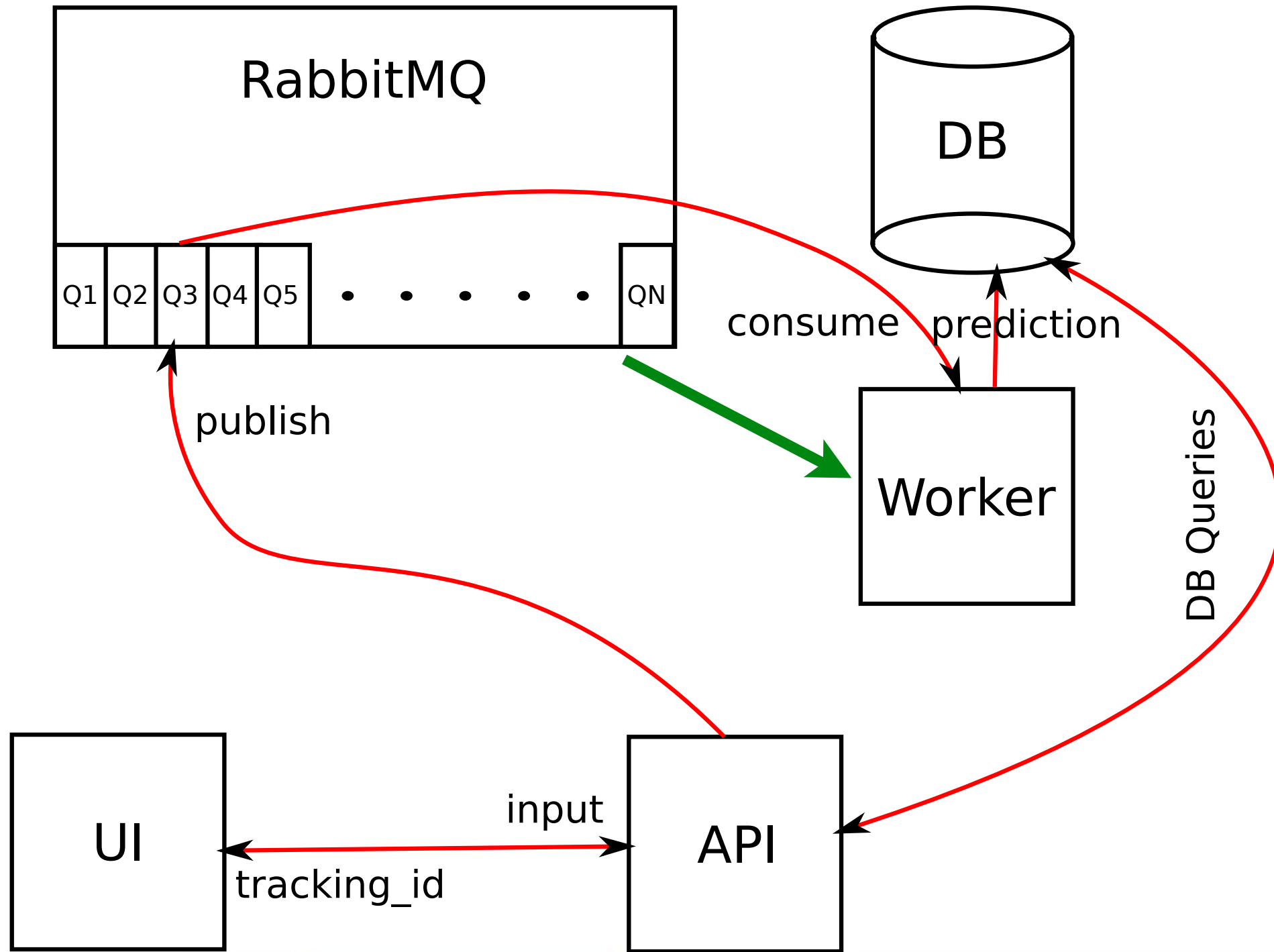
The Design



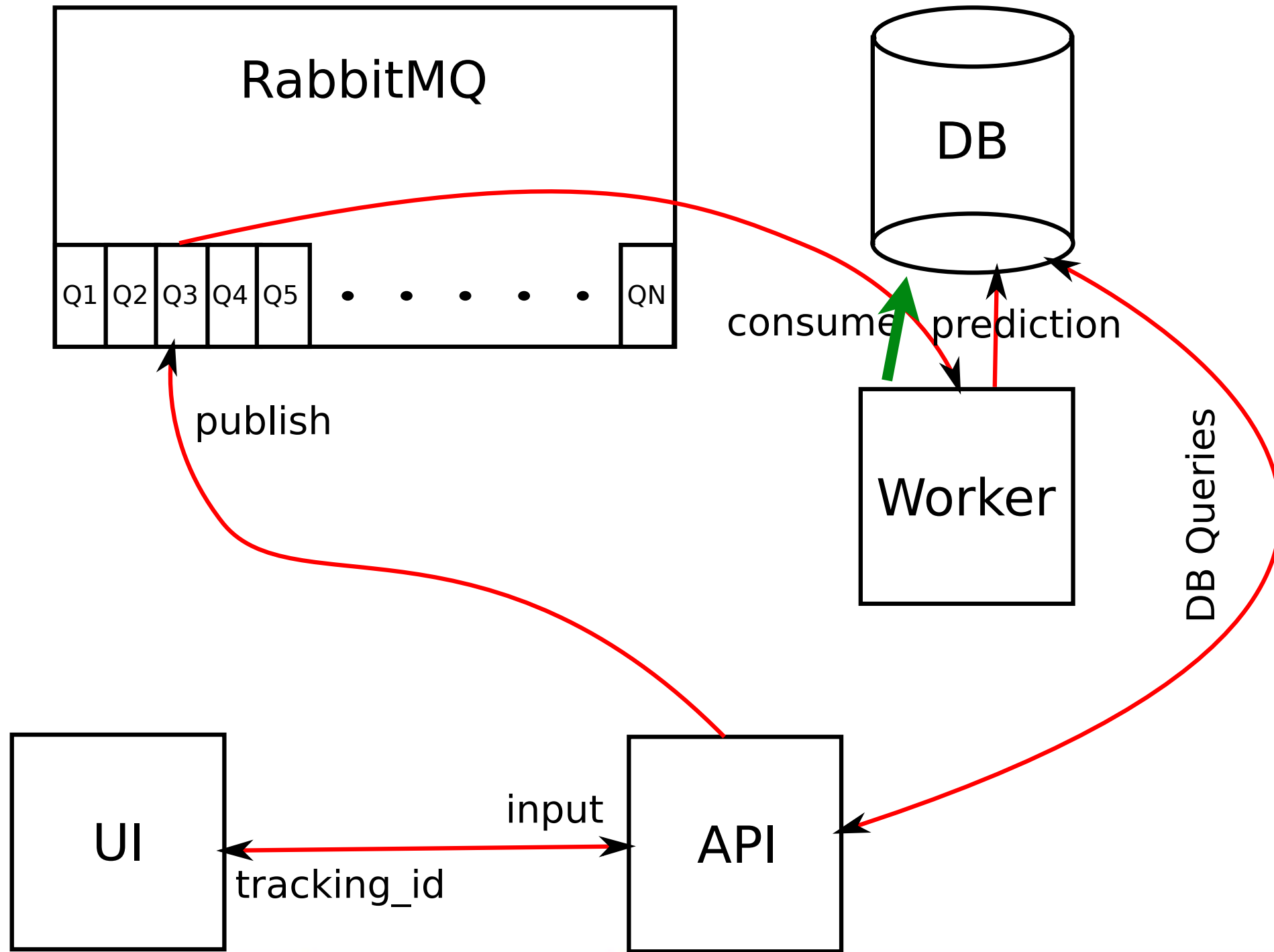
The Design



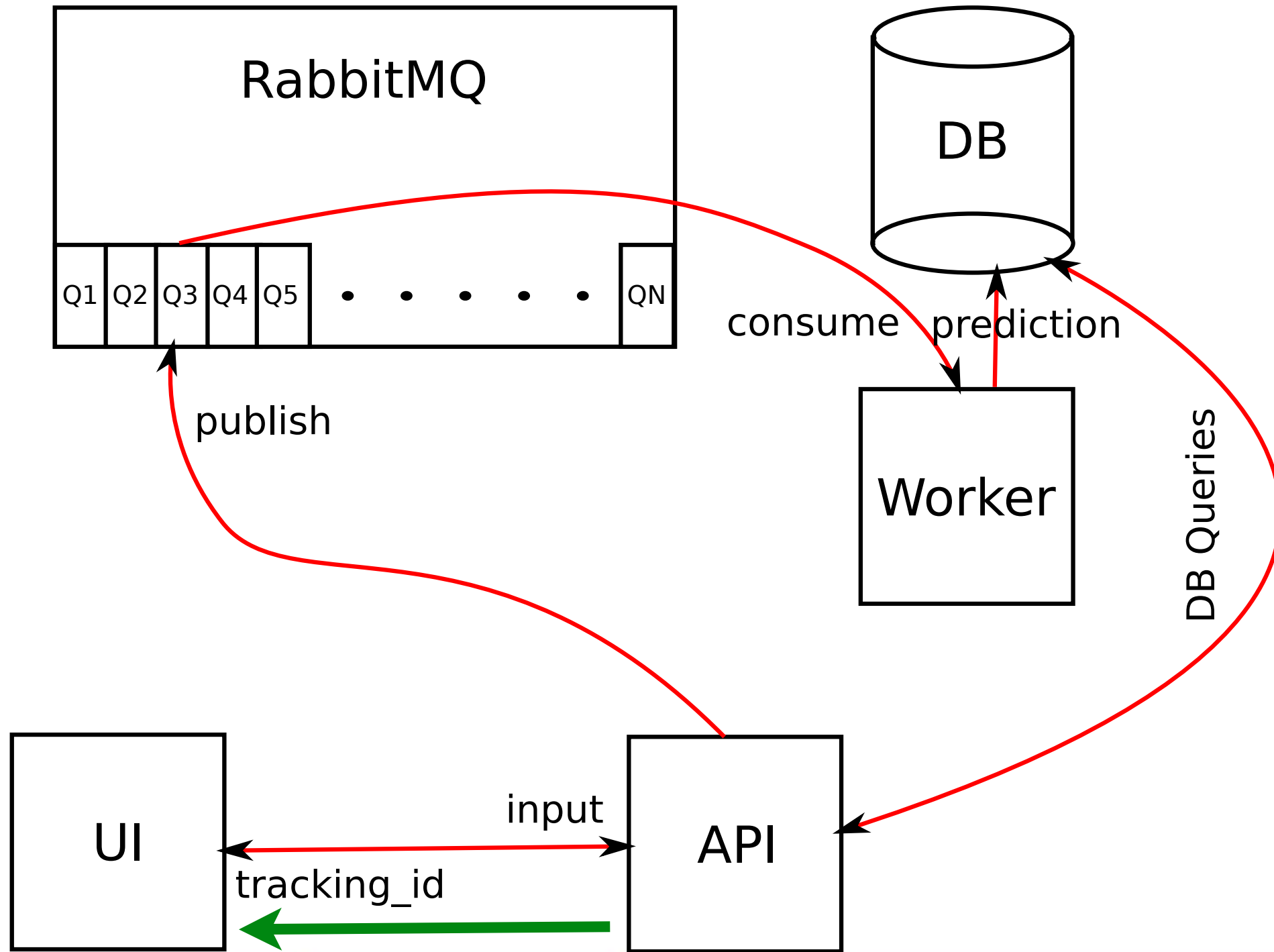
The Design



The Design



The Design



Demo Time!

Language Prediction

Input

Adolf II av Holstein

Adolf II av Holstein, född 1128, död 6 juli 1164 (stupad vid [Verchen](#) i [Demmin](#)), begravd i Minden, var greve av Holstein 1131–1164. Son till greve Adolf I av Holstein (död 1131) och [Hildewa](#).

Biografi[redigera]

Adolf II efterträdde fadern i Schauenburg och [Holstein-Wagrien](#). I flera år låg dock [Wagrien](#) under kontroll av [Pribislaw](#) av Mecklenburg. I tyska tronkriget stod Adolf på [welfisk](#) sida, och han var 1138–1142 förjagad av [Albrekt Björnen](#) då Adolf vägrade erkänna denne som sachsisk hertig. Som ny greve i Holstein och [Stormarn](#) insatte [Albrekt](#) då [Heinrich von Badwide](#), men denne kunde Adolf senare driva ut med [welfisk](#) hjälp. 1143 återfick Adolf av Henrik Lejonet sitt tidigare grevskap mot en stor summa pengar, och detta år förenades landskapet [Wagrien](#) slutgiltigt med Holstein.

Adolf grundade 1134 [Segeberg](#) vars borg senare brändes ned av den av Adolf på flykt [fördrivne](#) [Heinrich von Badwide](#). Borgen återuppbyggdes och blev Adolfs viktigaste stödjepunkt. 1143 grundade Adolf även [Alt-Lübeck](#) som förstördes 1147 av furst [Niklot](#) av [obotriterna](#). Området överlämnades slutligen till Henrik Lejonet vilken 1158 nygrundade Lübeck.

Submit

Result

Swedish Svenska

Scaling



Realtime vs Batch



Monitoring



Load



Re-verify



Re-verify



Thank You!

API & Consumer: Kelly O'Brien ([linkedin.com/in/kellyobie](https://www.linkedin.com/in/kellyobie))

UI: Matt Parke (ordinaryrobot.com)

Classifier: Michael Becker (github.com/mdbecker)

Images: Wikipedia; flickr.com users: kohsa, spenceyc, statefarm, klara, rh2ox, nasahqphoto, frickfrack



My Info

Twitter: @beckerfuffle

Blog: beckerfuffle.com

These slides and more @ github.com/mdbecker



Demo time!

Language Prediction

Input

Adolf II av Holstein

Adolf II av Holstein, född 1128, död 6 juli 1164 (stupad vid [Verchen](#) i [Demmin](#)), begravd i Minden, var greve av Holstein 1131–1164. Son till greve Adolf I av Holstein (död 1131) och [Hildewa](#).

Biografi[redigera]

Adolf II efterträdde fadern i Schauenburg och [Holstein-Wagrien](#). I flera år låg dock [Wagrien](#) under kontroll av [Pribislaw](#) av Mecklenburg. I tyska tronkriget stod Adolf på [welfisk](#) sida, och han var 1138–1142 förjagad av [Albrekt Björnen](#) då Adolf vägrade erkänna denne som sachsisk hertig. Som ny greve i Holstein och [Stormarn](#) insatte [Albrekt](#) då [Heinrich von Badwide](#), men denne kunde Adolf senare driva ut med [welfisk](#) hjälp. 1143 återfick Adolf av Henrik Lejonet sitt tidigare grevskap mot en stor summa pengar, och detta år förenades landskapet [Wagrien](#) slutgiltigt med Holstein.

Adolf grundade 1134 [Segeberg](#) vars borg senare brändes ned av den av Adolf på flykt [fördrivne](#) [Heinrich von Badwide](#). Borgen återuppbyggdes och blev Adolfs viktigaste stödjepunkt. 1143 grundade Adolf även [Alt-Lübeck](#) som förstördes 1147 av furst [Niklot](#) av [obotriterna](#). Området överlämnades slutligen till Henrik Lejonet vilken 1158 nygrundade Lübeck.

Submit

Result

Swedish Svenska

Demo time!

Language Prediction

Input

Manhattanprosjektet (engelsk: Manhattan Project) var et forsknings- og utviklingsprogram, som under amerikansk ledelse og med deltakelse av Storbritannia og Canada førte til fremstillingen av de første atombombene under andre verdenskrig. Fra 1942 til 1946 ble prosjektet ledet av generalmajor Leslie Groves fra den amerikanske hærens ingeniørkorps (US Army Corps of Engineers). Hærens del av prosjektet fikk betegnelsen Manhattan District. Manhattan avløste etter hvert det offisielle kodenavnet Development of Substitute Materials som betegnelse for hele prosjektet. Underveis slukte Manhattanprosjektet også den tidligere britiske motparten, kalt Tube Alloys. Blant fysikerne som tok del i Manhattanprosjektet var Albert Einstein, Edward Teller og danske Niels Bohr. Manhattanprosjektet begynte i det små i 1938, men det vokste raskt og tilsammen beskjeftiget det over 130 000 personer og kostet nesten 2 milliarder dollar (tilsvarende ca. 150 milliarder i 2012[1]). Over 90 % av pengene gikk til byggingen av fabrikker og produksjonen av spaltbart materiale, mens under 10 % gikk til selve utviklingen av våpnene. Forskning og utvikling foregikk på mer enn 30 forskjellige steder rundt om i USA, Storbritannia og Canada, og noen av disse var hemmelige. Det ble bygget to typer atombomber under andre verdenskrig. En forholdsvis enkel uranbombe som benyttet uran-235, som er en relativt sjelden uranisotop som kun utgjør 0,7 % av naturlig

Submit

Result

Norwegian (Bokmål) Norsk (Bokmål)

Demo time!

Language Prediction

Input

مره أخرى لأنه ينفذ أوامر القادة ولد (بول تيببتس) في عام 1915 م واسم والدته "اينولا جاي" اشتهر تيببتس بأنه أفضل طياري الجيش الأمريكي وقتها وكان تيببتس، الذي اشتهر بأنه أفضل طياري الجيش الأمريكي وقتها، وفي عام 1945 م ترقى إلى رتبة كولونيل وهي تعادل رتبة (عقيد) في سلاح الجو الأمريكي في 6-8-1945 قاد الطائرة الأمريكية التي ألقت القنبلة الذرية على هيروشيما في اليابان وهي قنبلة تحتوي على 60 كيلوغراما (130 رطلا) من مادة اليورانيوم - 235 هيروشيما هي مدينة في اليابان، تقع في جزيرة "هونشو"، وتشرف على "خليج هيروشيما". وكان تعداد سكانها عام 1945 م وأما الطائرة التي تحمل القنبلة فكانت قاذفة من (350,000 Little Boy نسمة وكان اسم الطائرة الكودي (بالشفرة السرية) هو (الولد الصغير النوع (بي 29) وكان بول تيببتس أول من اختبر هذه الطائرة وأطلق عليها اسم أمه "اينولا جاي". وتقرر تنفيذ مهمة قصف هيروشيما يوم 6-8-1945 م حيث كان الطقس موافقا لتنفيذ المهمة الخطيرة بعد أيام من السحب التي تجمعت فوق هيروشيما، فانطلق بول تيببتس بطائرته وهي تحمل القنبلة الذرية من قاعدة «نورث فيلد» في جزيرة تينيان، غرب المحيط الهادئ، مصحوبا بطائرتين أخريين. وقبل القصف بساعة اكتشف نظام الإنذار المبكر الياباني دخول الطائرات للمجال الجوي الياباني فنبه السلطات في كبرى المدن بما فيها هيروشيما. لكن بول تيببتس كان في طريقه للمدينة المستهدفة يقود الطائرة القاذفة، واستطاع أن يتهرب من الدفاعات اليابانية ليصل مدينة هيروشيما وحوالي الساعة الثامنة صباحا تمكنت أجهزة الرادار في هيروشيما من تحديد الطائرات الأمريكية لكن المسؤولين العسكريين قرروا أن عددها الصغير لا يستدعي التصدي لها بطائرات مضادة على ضوء سياستهم الرامية لتوفير وقود الطائرات وفي تمام الساعة الثامنة والرابع قصف بول تيببتس القنبلة الرهيبة من طائرته «بي 29» على

Submit

Result

العربية Arabic