Yandex

Yandex

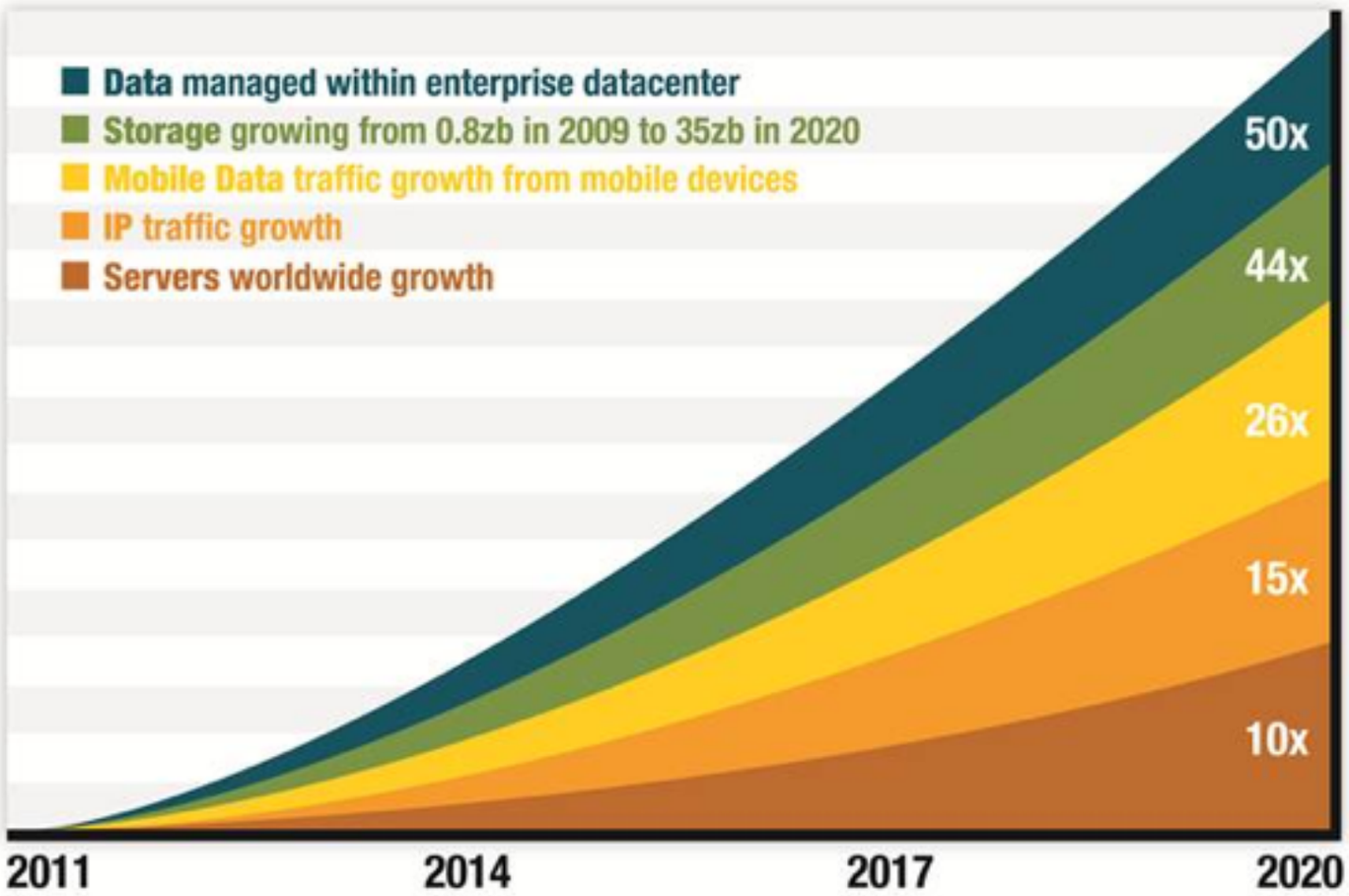
Машинное обучение как инструмент современного ученого

Андрей Устюжанин

Руководитель совместных проектов Яндекс-ЦЕРН

Рост потоков данных

- Текст
- Фото
- Аудио
- Видео
- GPS-треки

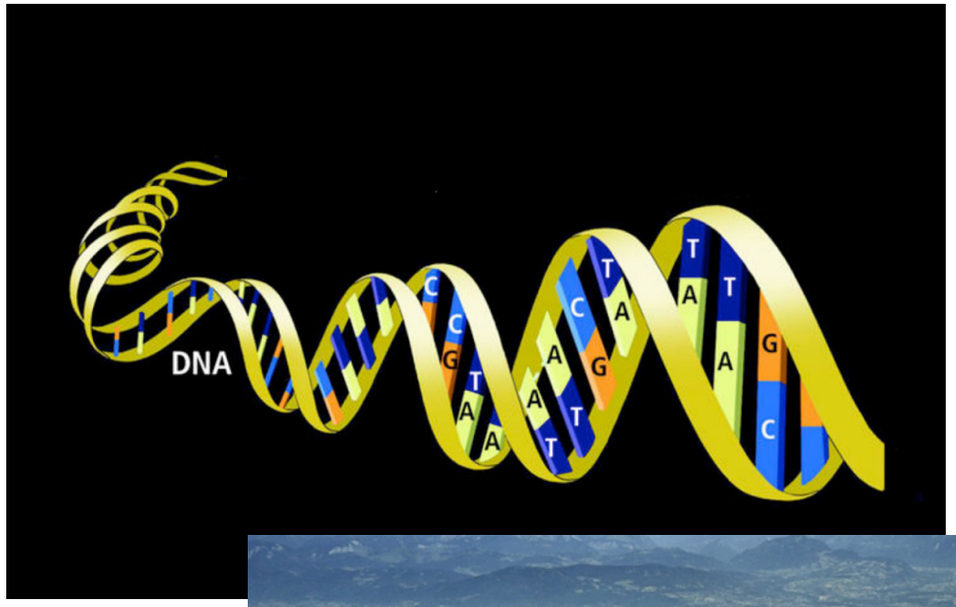


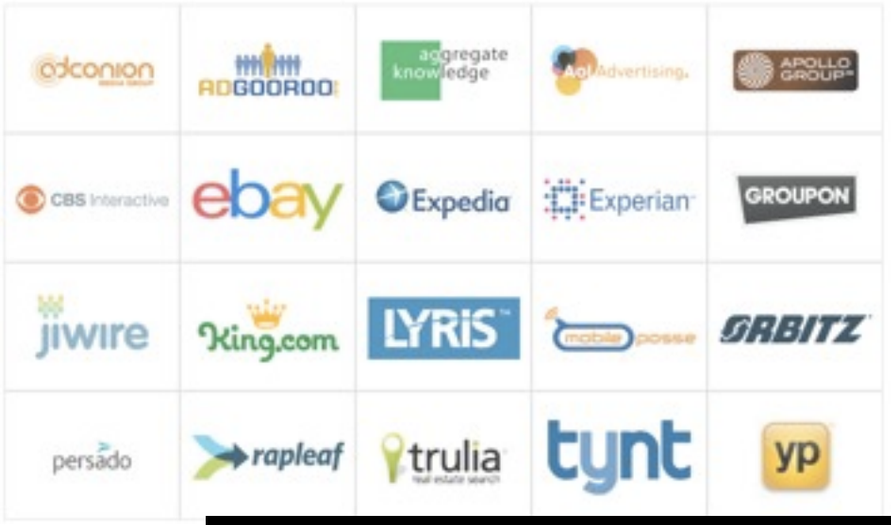
Source: 2011 and 2012 Cisco VNI, EMC and IDC

Компании

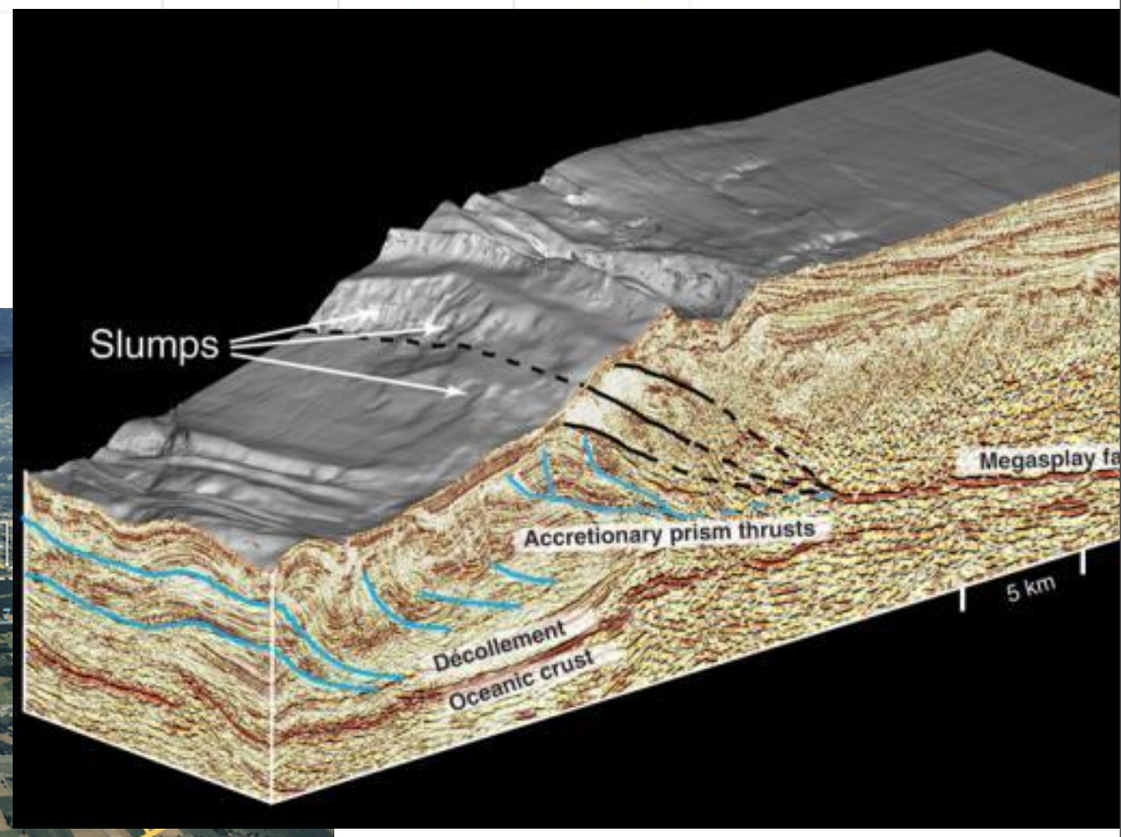

adconion MEDIA GROUP
ADGOOROO
aggregate knowledge
Aol Advertising.
APOLLO GROUP
CBS Interactive
ebay
Expedia
Experian
GROUPON
jiwire
King.com
LYRIS
mobile posse
ORBITZ
persado
rapleaf
trulia real estate search
tynt
yp

Наука


DNA


adconion
ADGOOROO
aggregate knowledge
Advertising.
APOLLO GROUP
CBS Interactive
ebay
Expedia
Experian
GROUPON
jiwire
King.com
LYRIS
mobile posse
ORBITZ
persado
rapleaf
trulia
tynt
yp

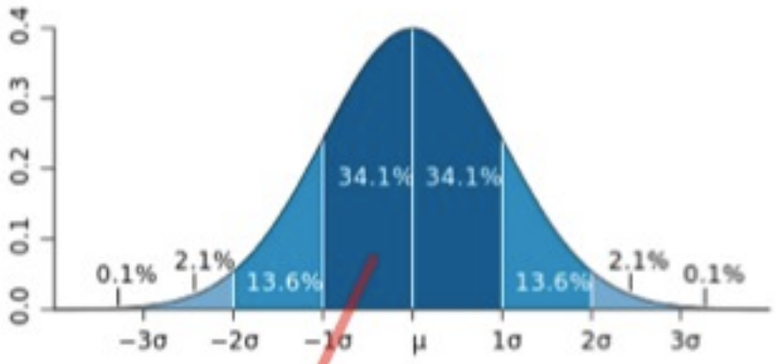

LHCb
ATLAS
CERN
SUISSE
FRANCE
CMS
LHC 27 km


Slumps
Megasplay fa
Accretionary prism thrusts
5 km
Décollement
Oceanic crust

How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?"

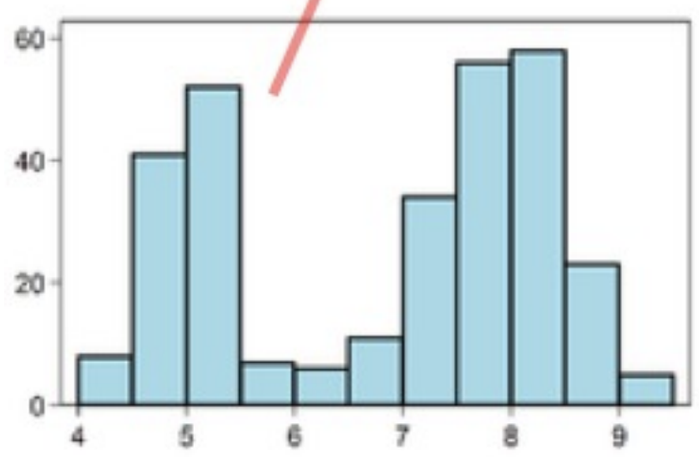
Tom Mitchell, CMU

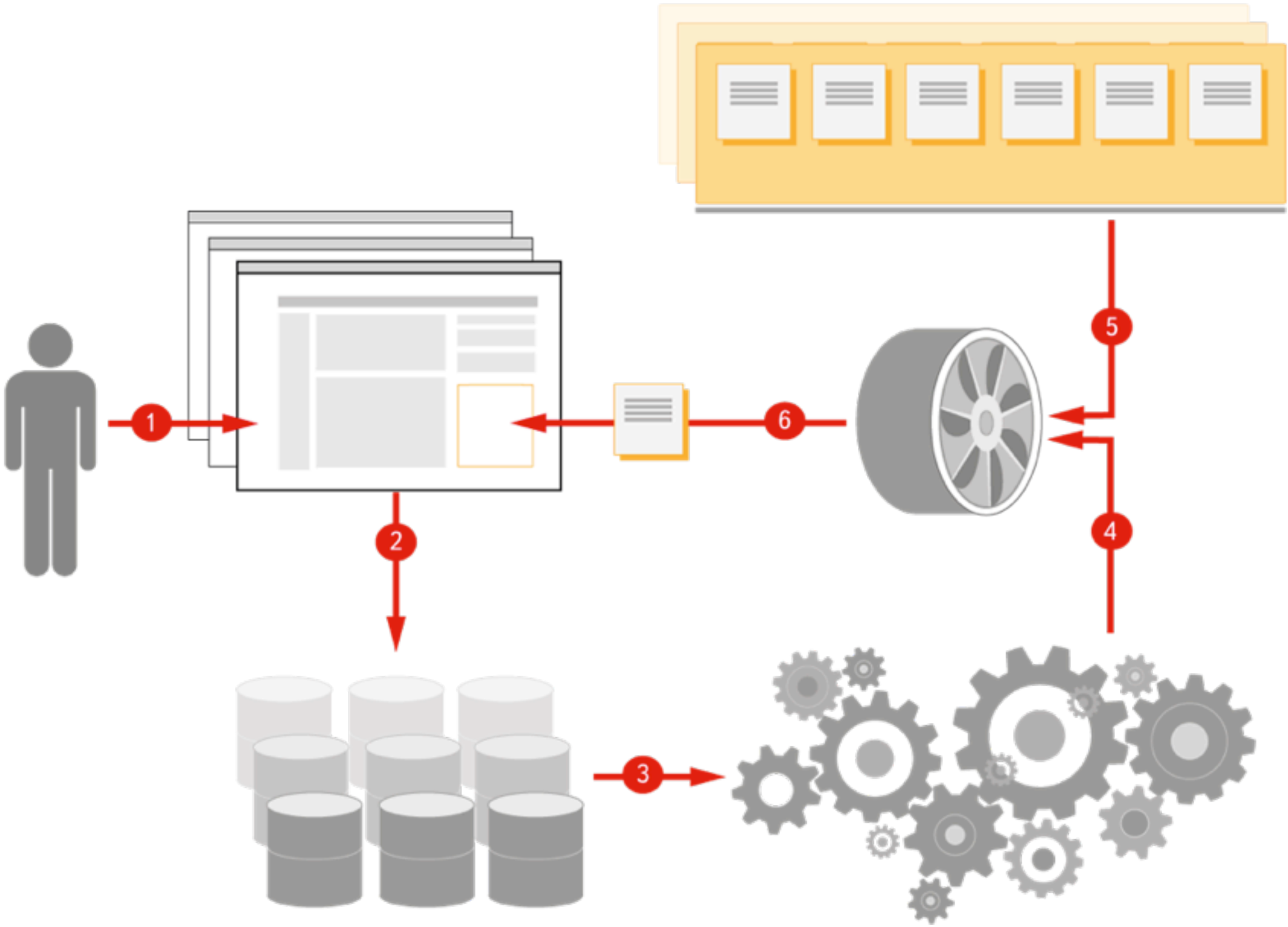
Метапереход: от статистики к машинному обучению



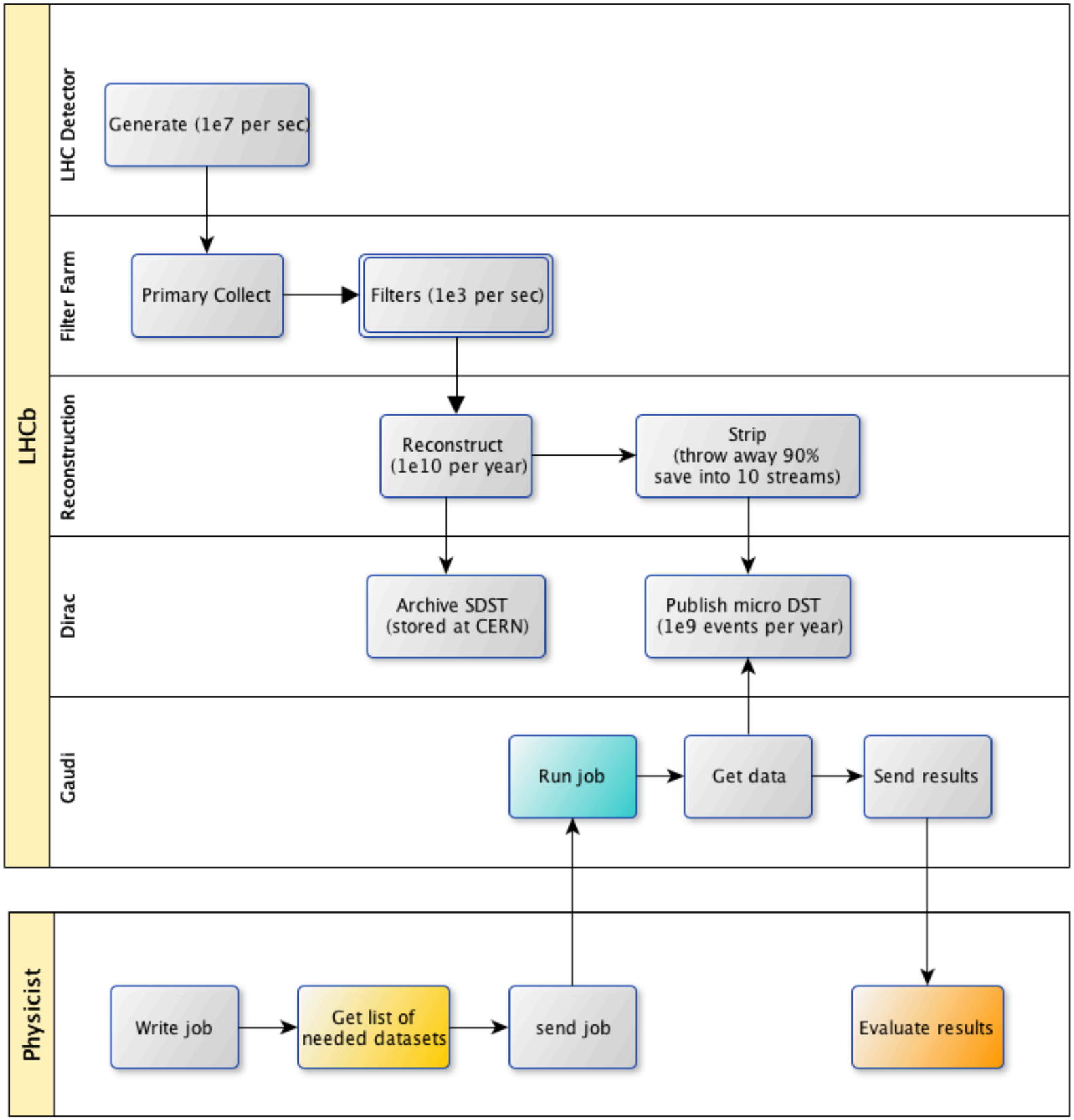
Мир
глазами аналитика

Реальный мир

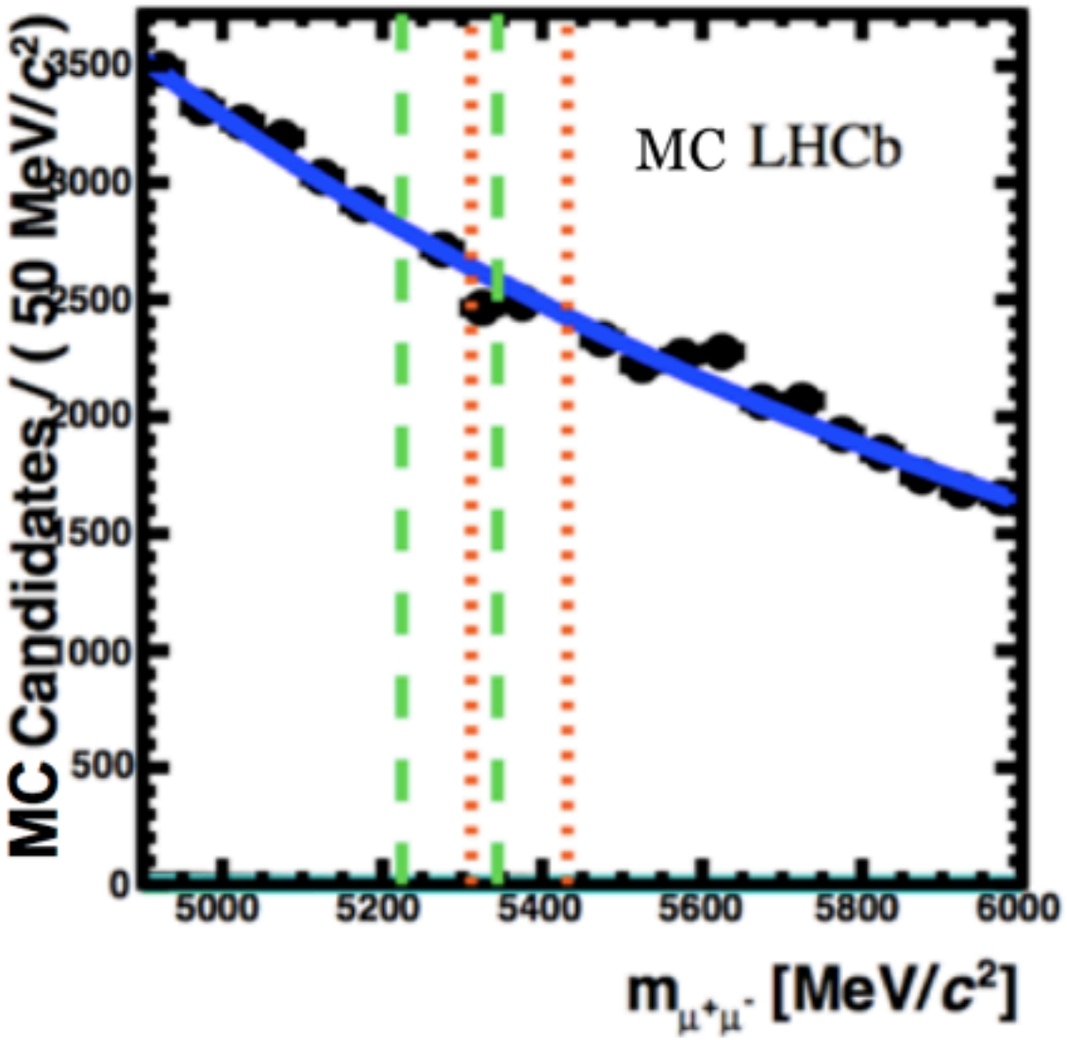




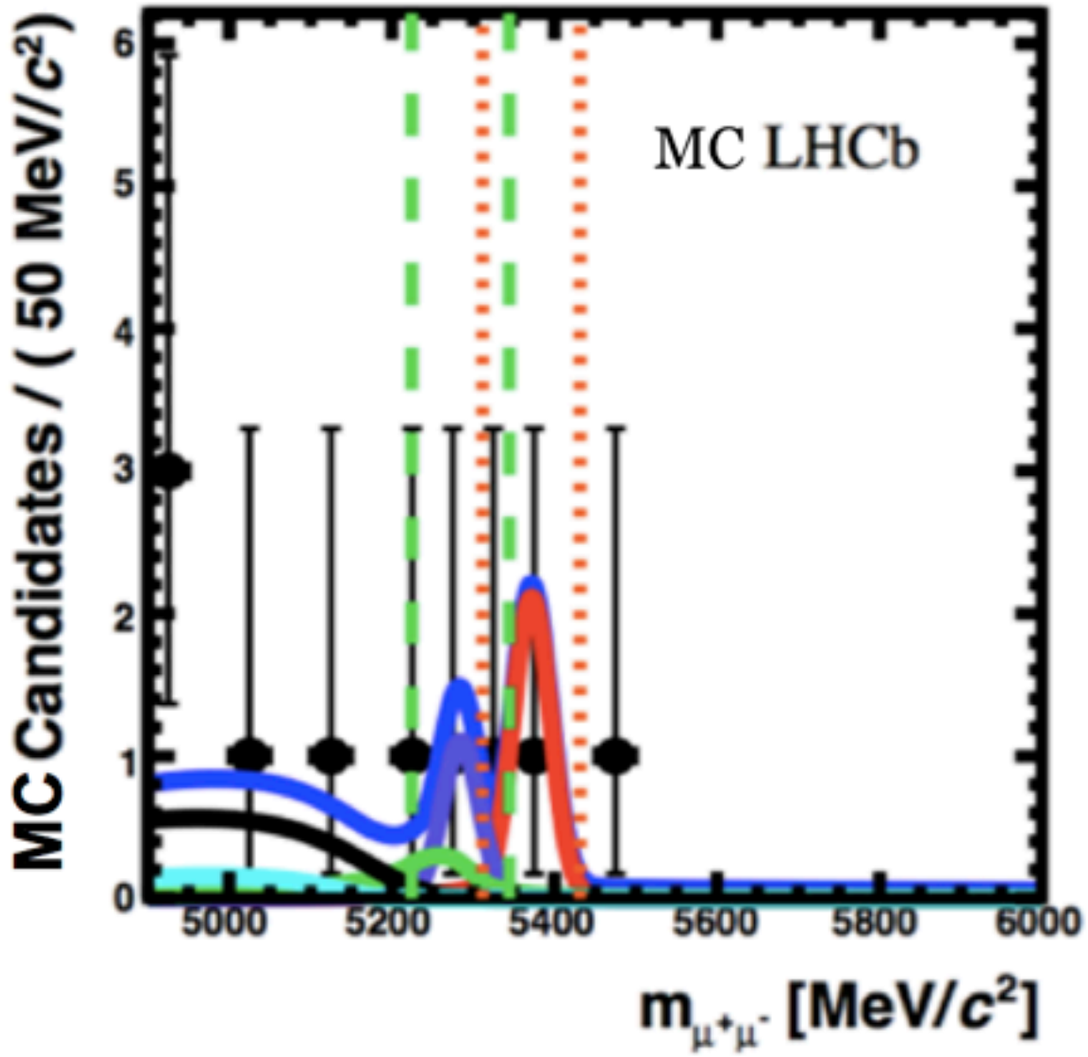
Баннерная реклама



Физический анализ



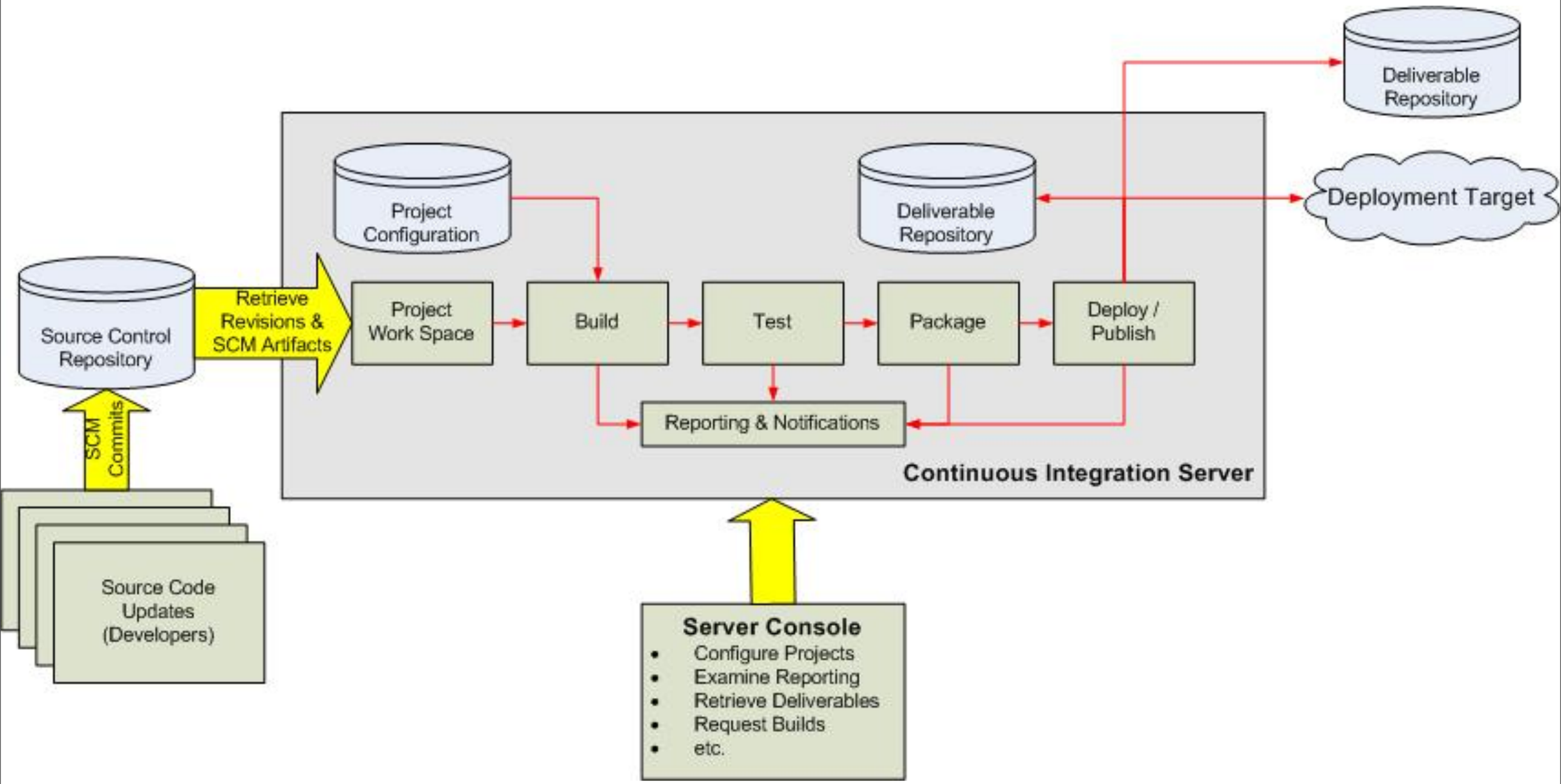
FML



Event Filtering (Bs→μ+μ-)

Выборка реальных данных
Подготовка виртуальных данных
Проектирование / вычисление признаков
Обучение модели
Проверка
Применение модели
Нормализация
Фитирование
Вычисление частоты распада/достоврености

Выборка реальных данных
Подготовка виртуальных данных
Проектирование / вычисление признаков
Обучение модели
Проверка
Применение модели
Нормализация
Фитирование
Вычисление частоты распада/достоврености



http://bit.ly/130e2X3

Основные шаги

- Поиск/сбор данных
- Подготовка признаков
- Обучение модели
- Проверка
- Использование
- Анализ

Платформы машинного обучения

word2vec by Google http://code.google.com/p/word2vec

h2o by 0xdata http://0xdata.com/h2O

“The Berkeley Stack” by AMPLab (see Spark, MLBase, etc.)http://amplab.cs.berkeley.edu

Vorpal Wabbit by John Langford http://github.com/JohnLangford/vowpal_wabbit

KNIME http://knime.org

scikit-learn for Python http://scikit-learn.org

Платформы машинного обучения

word2vec by Google http://code.google.com/p/word2vec

h2o by 0xdata http://0xdata.com/h2O

“The Berkeley Stack” by AMPLab (see Spark, MLBase, etc.)http://amplab.cs.berkeley.edu

Vorpal Wabbit by John Langford http://github.com/JohnLangford/vowpal_wabbit

KNIME http://knime.org

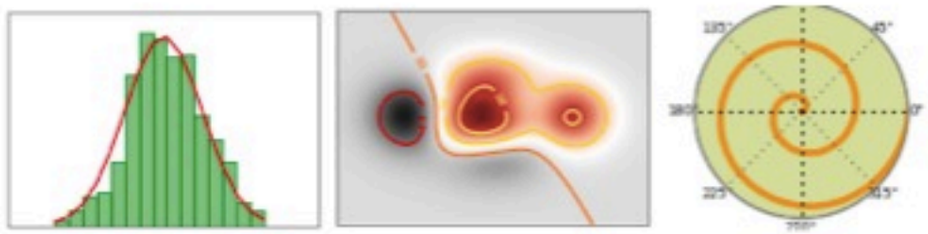
scikit-learn for Python http://scikit-learn.org

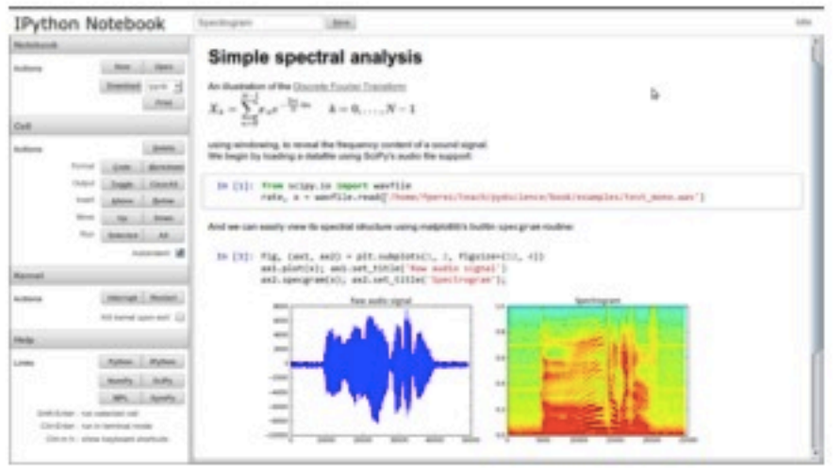
astroml.org/sklearn_tutorial/general_concepts.html#features-and-feature

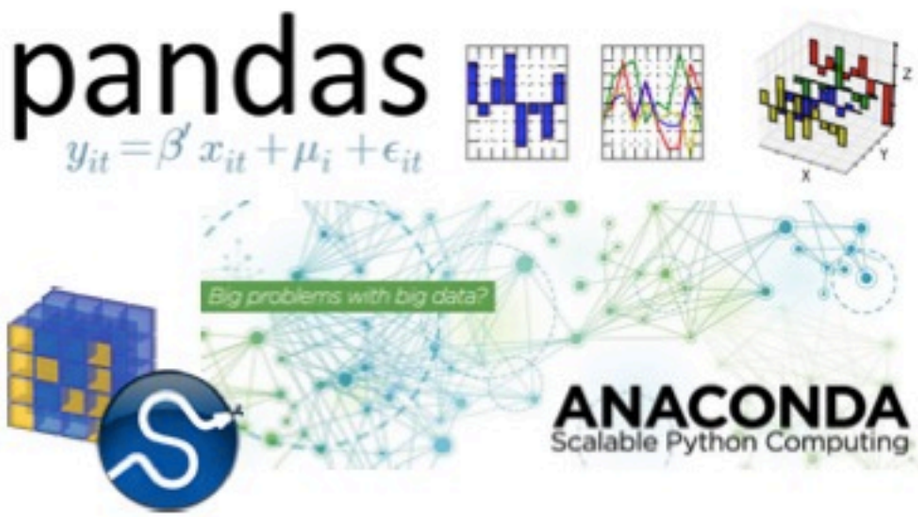
```
from sklearn.datasets import load_iris from sklearn.svm import
LinearSVC
iris = load_iris()
n_samples, n_features = iris.data.shape
    n_samples
    n_features
iris.target list(iris.target_names)
X, y = iris.data, iris.target clf = LinearSVC()
clf = clf.fit(X, y)
clf.coef_ clf.intercept_
X_new = [[ 5.0, 3.6, 1.3, 0.25]]
l = clf.predict(X_new)
map(lambda x: iris.target_names[x], l)
```

Python

- matplotlib.org
- ipython.org
- pandas.pydata.org
- numpy.org
- scipy.org
- continuum.io
- nltk.org
- scikit-learn.org
- beta.graphlab.com







iPython notebook

- Literation programming
- Literate computation

IPython Notebook

Spectrogram | Save

Notebook

Actions: New | Open | Download | ipynb | Print

Cell

Actions: Delete

Format: Code | Markdown

Output: Toggle | ClearAll

Insert: Above | Below

Move: Up | Down

Run: Selected | All

Autoindent: ☑

Kernel

Actions: Interrupt | Restart

Kill kernel upon exit: ☐

Help

Links: Python | IPython | NumPy | SciPy | MPL | SymPy

Shift-Enter : run selected cell

Ctrl-Enter : run in terminal mode

Ctrl-m h : show keyboard shortcuts

Simple spectral analysis

An illustration of the Discrete Fourier Transform

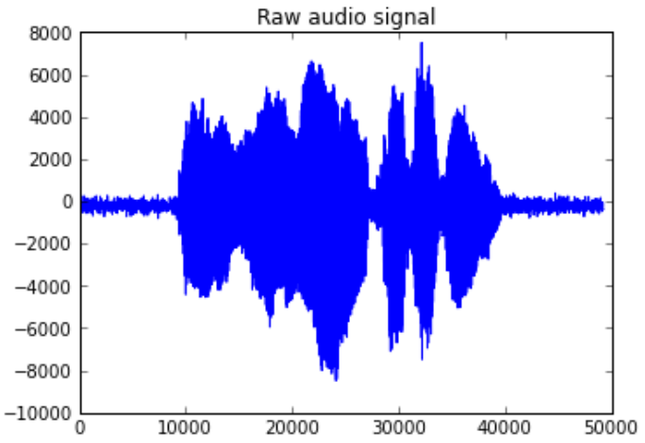
$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0, \ldots, N-1$$

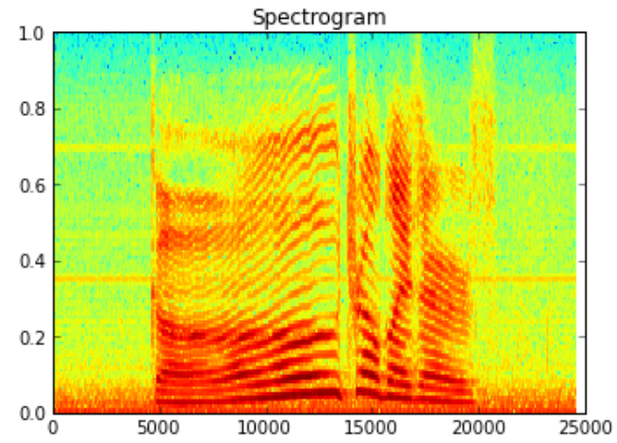
using windowing, to reveal the frequency content of a sound signal.
We begin by loading a datafile using SciPy's audio file support:

```
In [1]: from scipy.io import wavfile
        rate, x = wavfile.read('/home/fperez/teach/py4science/book/examples/test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin `specgram` routine:

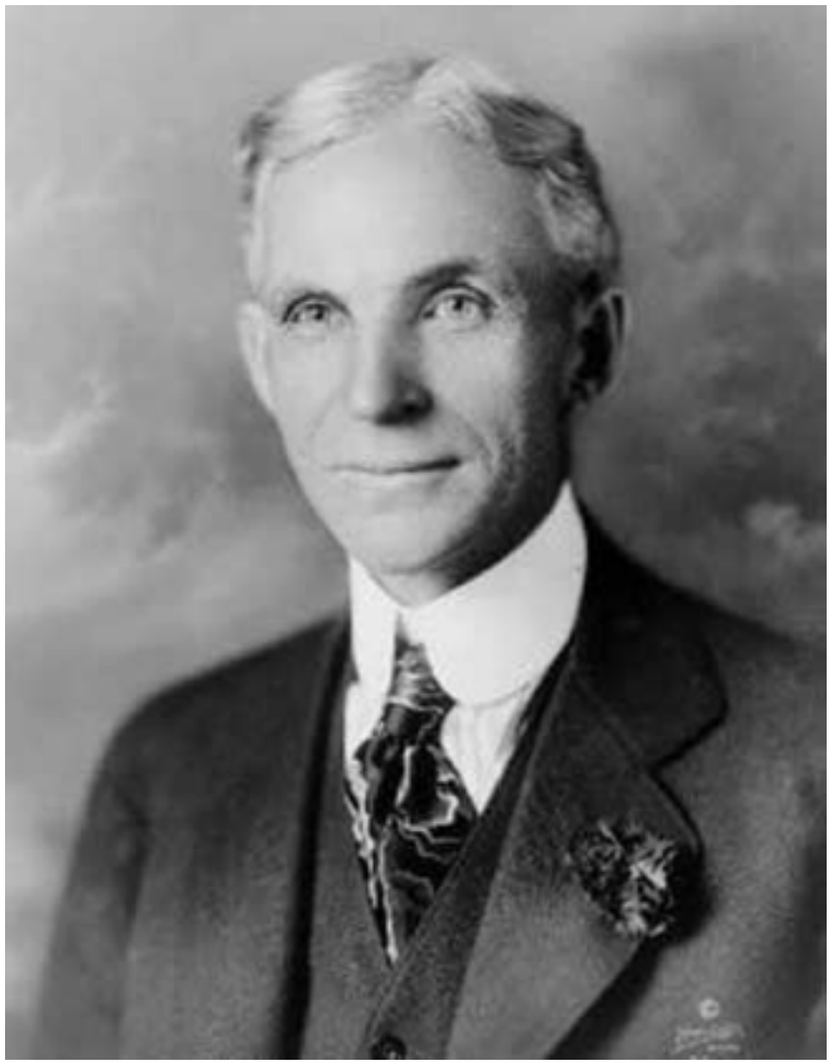
```
In [3]: fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))
        ax1.plot(x); ax1.set_title('Raw audio signal')
        ax2.specgram(x); ax2.set_title('Spectrogram');
```

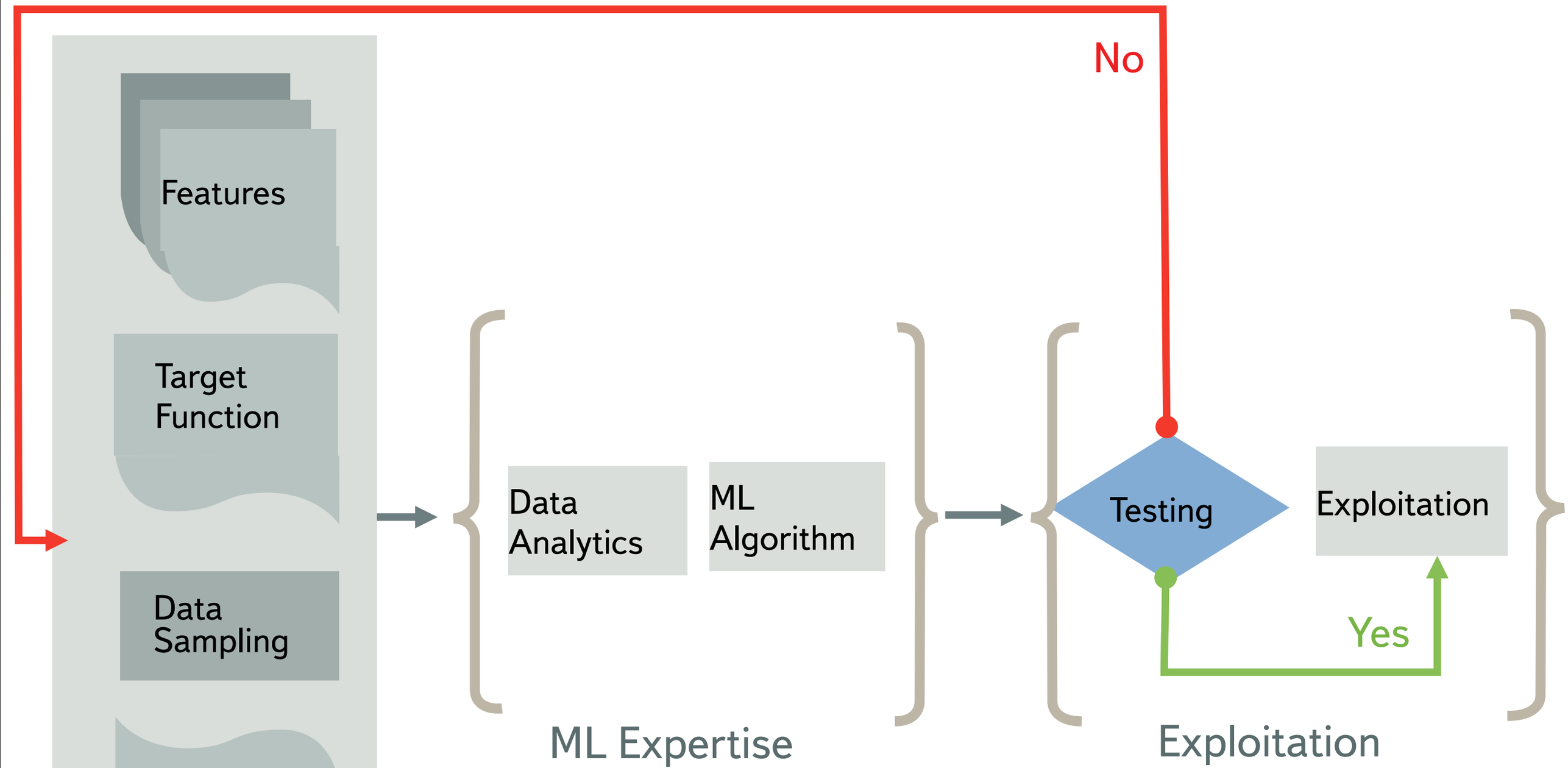




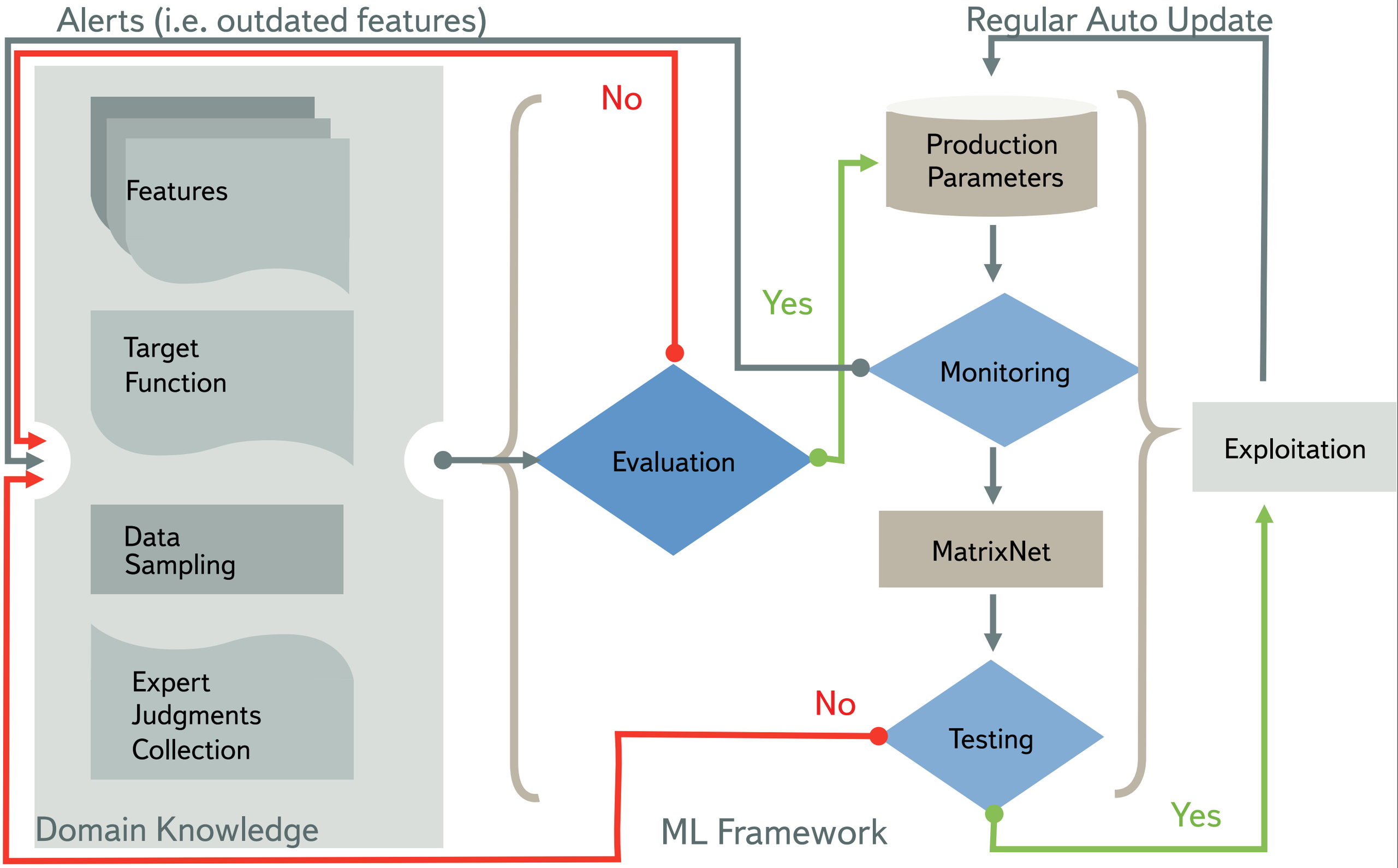
Мета-системный переход

«Как автоматизировать ~~производство автомобилей~~ обучение?»





Обычный цикл обучения



Итеративное обучение

Today

13:08 test01 Mitchell

13:08 Success Event classification
- completed Calculating metrics
- started Calculating metrics
- completed Training test formula
- completed Training baseline formula
- started Training test formula
- started Training baseline formula
- completed Feature selection test formula
- completed Feature selection baseline formula
- started Feature selection test formula
- started Feature selection baseline formula
- completed Preparing dataset
- started Preparing dataset

13:07 Created Event classification

28.09.2013

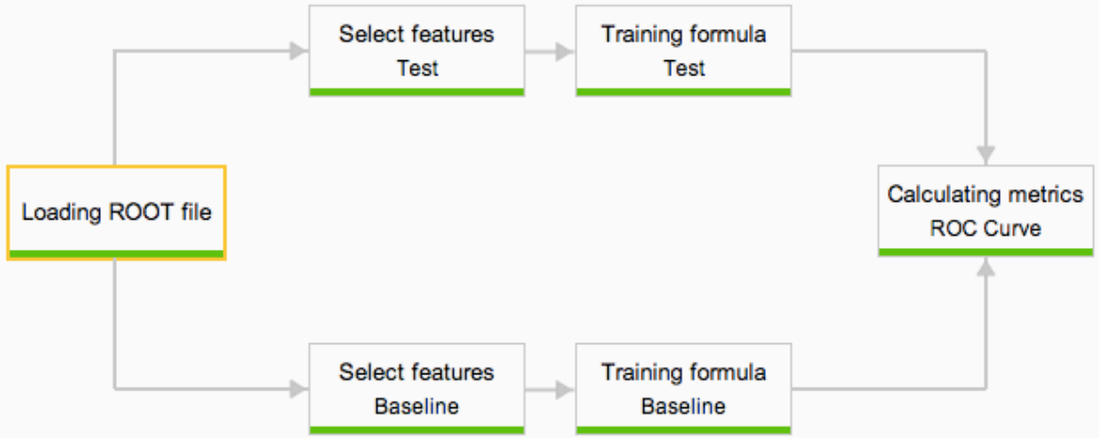
18:18 Test Newton

18:18 Success Event classification
- completed Calculating metrics
- started Calculating metrics
- completed Training test formula
- completed Training baseline formula
- started Training test formula
- started Training baseline formula
- completed Feature selection test formula
- completed Feature selection baseline formula
- started Feature selection test formula
- started Feature selection baseline formula
- completed Preparing dataset

Event Classification

Mitchell

13:07 test01



Loading ROOT file (Baseline)

Dataset will be split into training set and validation set. Training set will be used for MatrixNet learning. Validation set will be used to measure quality of results.

Validation set size (%)

20

Подпись под картинкой

Вычислительные эксперименты

Модульность

Читаемость/прозрачность

Измеримость

Совместная работа

Воспроизводимость

Автоматизация

Андрей Устюжанин
Руководитель совместных проектов
Яндекс-ЦЕРН

anaderi@yandex-team.ru

Спасибо!