# Применение машинного обучения К задаче обнаружения спама

# SPAM (**SP**iced h**AM**)

- 1936 *Hormel Foods Corporation*

# SPAM

# SPAM (**SP**iced h**AM**)

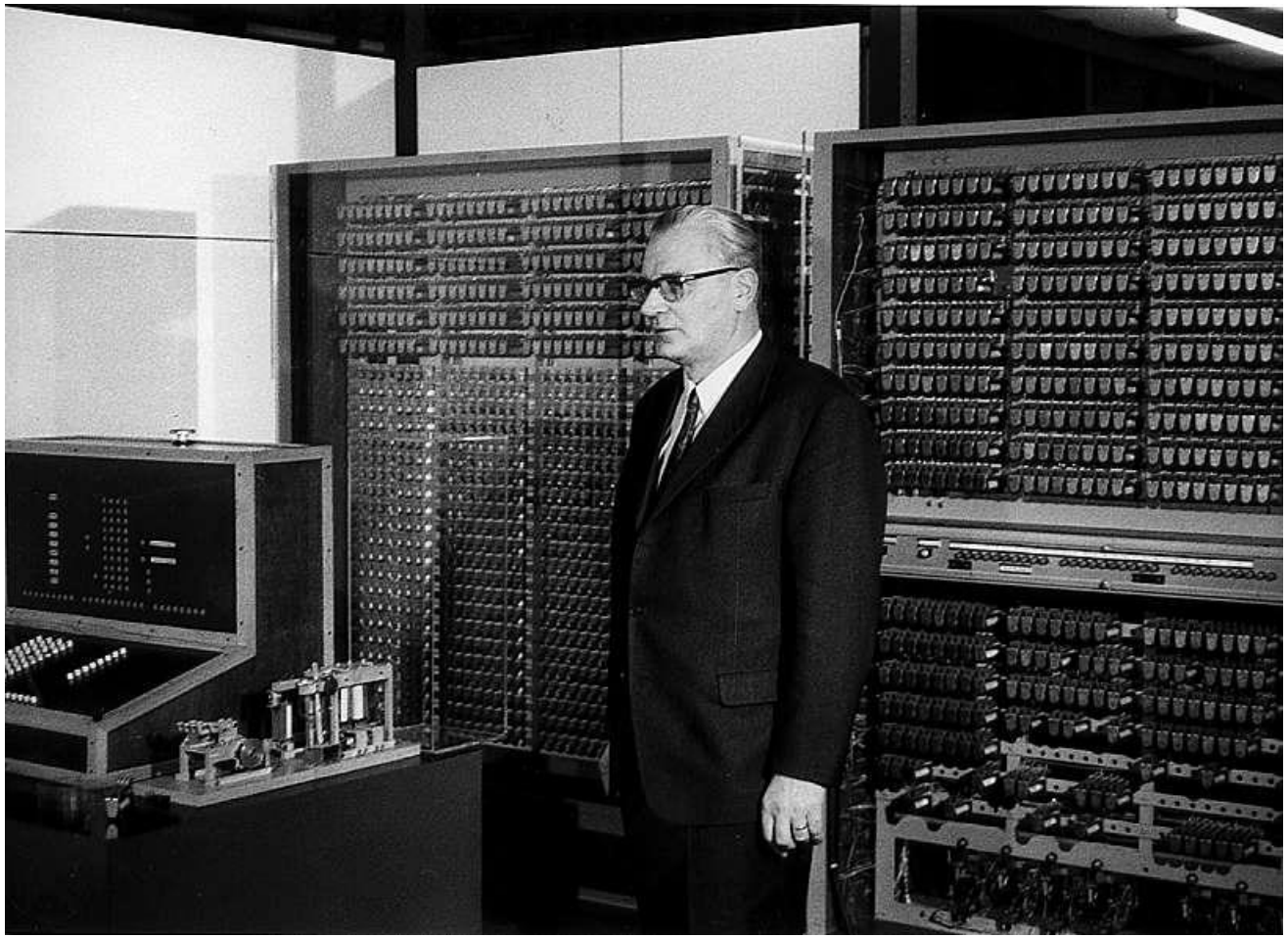
- 1936 *Hormel Foods Corporation*
- 1939 *World War II*

# SPAM (**SP**iced h**AM**)

# SPAM (**SP**iced h**AM**)

- 1936 *Hormel Foods Corporation*
- 1939 *World War II*
- 1944 *Z4, Конрад Цузе, высокоуровневый язык Планкалкюль*

# SPAM (**SP**iced h**AM**)



# SPAM (**SP**iced h**AM**)

- 1936 *Hormel Foods Corporation*
- 1939 *World War II*
- 1944 *Z4, Конрад Цюзе, высокоуровневый язык Планкалкюль*
- 1969 *Monty Python's Flying Circus*

# SPAM (**SP**iced h**AM**)

# SPAM (**SP**iced h**AM**)

- 1936 *Hormel Foods Corporation*
- 1939 *World War II*
- 1944 *Z4, Конрад Цюзе, высокоуровневый язык Планкалкюль*
- 1969 *Monty Python's Flying Circus*
- 1980s *BBS, MUD, Usenet*

# SPAM (**SP**iced h**AM**)

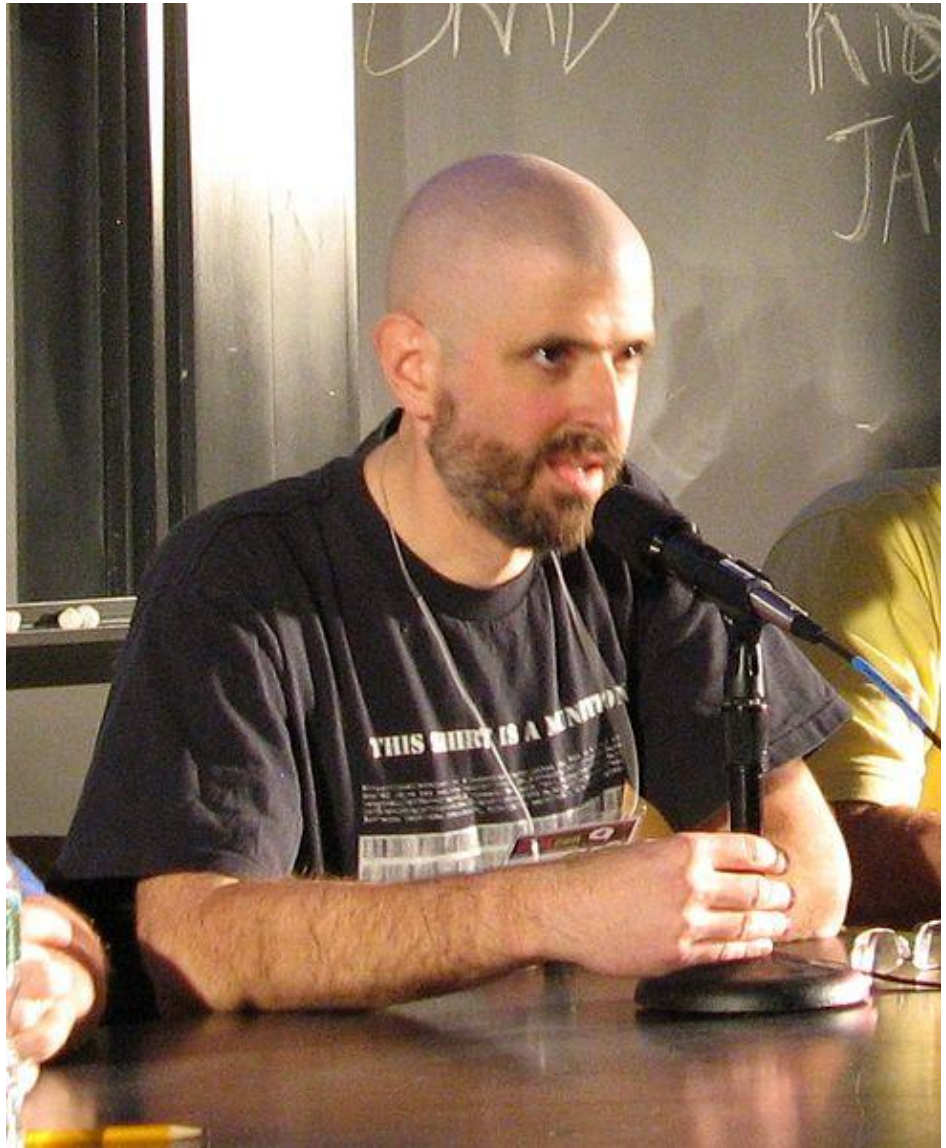
- 1936 *Hormel Foods Corporation*
- 1939 *World War II*
- 1944 *Z4, Конрад Цюзе, высокоуровневый язык Планкалкюль*
- 1969 *Monty Python's Flying Circus*
- 1980s *BBS, MUD, Usenet, Star Wars vs Star Trek*

# SPAM (**SP**iced h**AM**)

# SPAM (**SP**iced h**AM**)

- 1936 *Hormel Foods Corporation*
- 1939 *World War II*
- 1944 *Z4, Конрад Цюзе, высокоуровневый язык Планкалкюль*
- 1969 *Monty Python's Flying Circus*
- 1980s *BBS, MUD, Usenet, Star Wars vs Star Trek*
- 1993 *Джоел Фурр*

# SPAM (**SP**iced h**AM**)



# SPAM

# SPAM

# SPAM

# Актуальна ли проблема?

# Актуальна ли проблема?

# Актуальна ли проблема?

- *you're a clown. kill yourself.*
- *if you delete flappy bird I will literally kill myself. It's my drug and I am so addicted!! PLEASE DO NOT DO THIS TO MEEE PLEASEE*
- *I think someone will kill you idk just saying*
- *NO ONLY 1 HOUR I WILL KILL YOU IF U TAKE IT DOWN*
- *…*

# Актуальна ли проблема?

- *мою почту атакует сайт занкомств twoo и в частности Игорь из Луганск 34 года*
- *<вырезано цензурой>*
- *сайт знакомст украина польша и че то еще*
- *аааааа*
- *бесит спам*
- *бесит*
- *еще гугл почему-то его не в спам, не в соц сети, не в рекламу не сортирует, а в главный ящик*

# Актуальна ли проблема?

Да

# Актуальна ли проблема?

# Актуальна ли проблема?


# Актуальна ли проблема?



# Актуальна ли проблема?




# Актуальна ли проблема?





# Актуальна ли проблема?






# Актуальна ли проблема?







# Актуальна ли проблема?








# Актуальна ли проблема?









# Актуална ли проблема?











# Актуальна ли проблема?

Symantec

SpamAssassin

LIBRA ESVA
email security virtual appliance

TREND MICRO

KASPERSKY lab

F-Secure

bitdefender
secure your every bit

OnlyMyEmail

SPAM TITAN

Google

# Актуальна ли проблема?









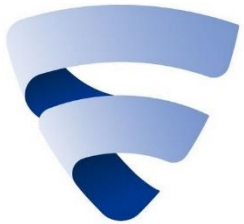



# Актуальна ли проблема?









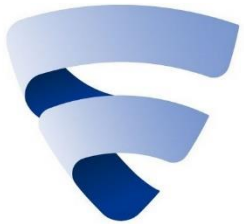



# Актуальна ли проблема?











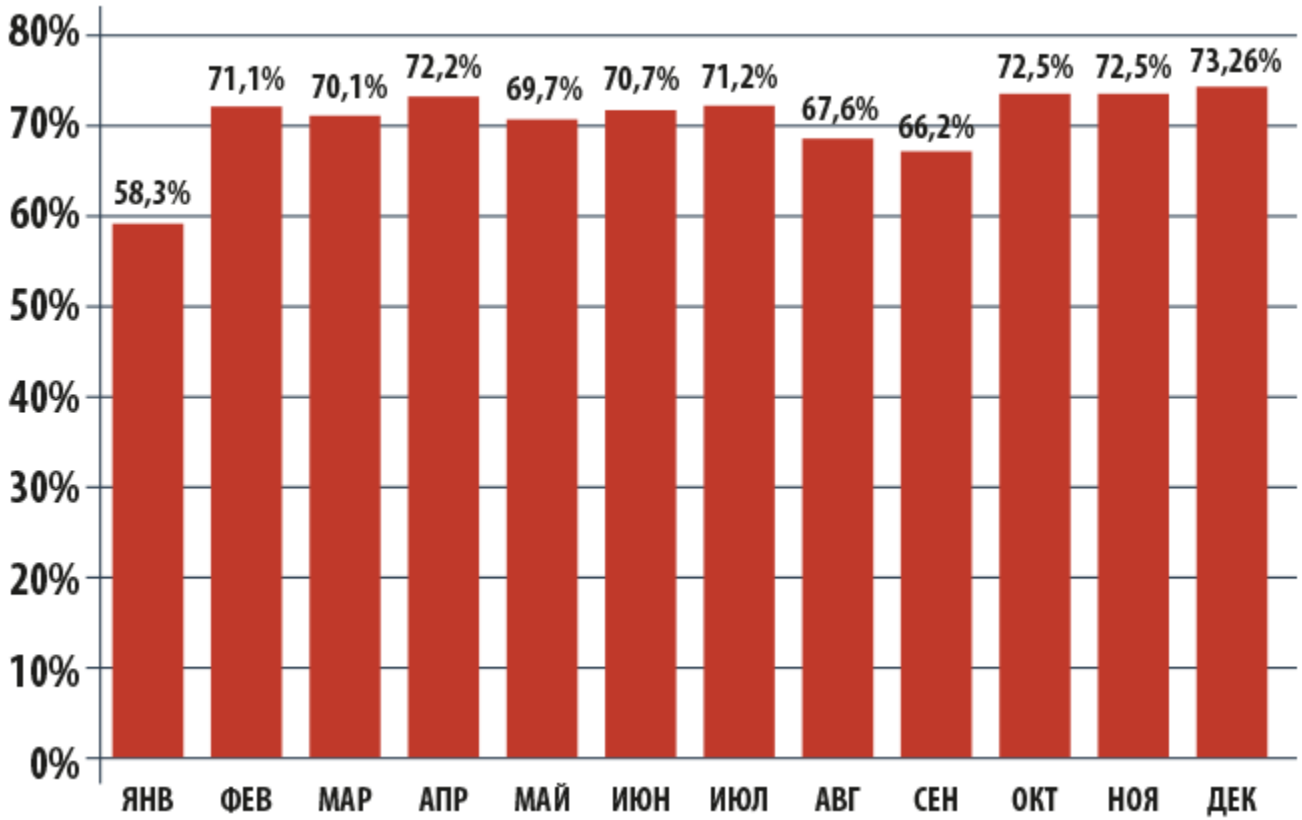


# Актуальна ли проблема?

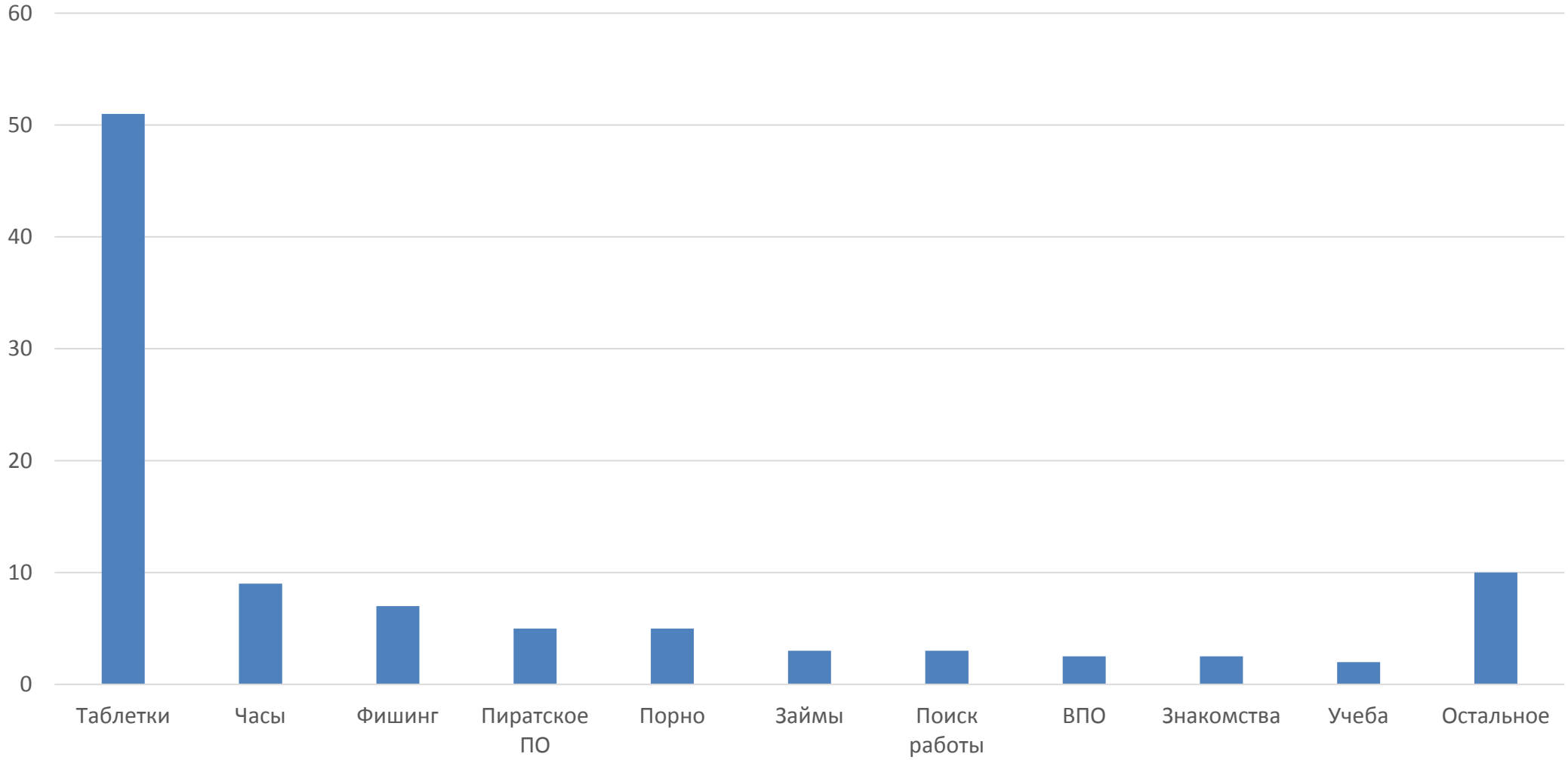
- Годовая прибыль спаммеров ~200.000.000$
- Годовые затраты на борьбу и финансовые потери ~20.000.000.000$

*http://blogs.wsj.com/ideas-market/2012/08/13/the-economics-of-spam/*

# Актуальна ли проблема?



# АКТУАЛЬНА ЛИ ПРОБЛЕМА?



# Зачем?

- Относительно малоэффективно, но дешево

*After 26 days, and almost 350 million e-mail messages, only 28 sales resulted*

*http://www.icsi.berkeley.edu/pubs/networking/2008-ccs-spamalytics.pdf*

# Зачем?

- Относительно малоэффективно, но дешево
- Распространение вредоносного ПО

# Зачем?

- Относительно малоэффективно, но дешево
- Распространение вредоносного ПО
- Таргетированный спам

# Зачем?

- Относительно малоэффективно, но дешево
- Распространение вредоносного ПО
- Таргетированный спам
- Очернение конкурентов

# Зачем?

- Относительно малоэффективно, но дешево
- Распространение вредоносного ПО
- Таргетированный спам
- Очернение конкурентов
- Just because we can

# Зачем?

- Относительно малоэффективно, но дешево
- Распространение вредоносного ПО
- Таргетированный спам
- Очернение конкурентов
- Just because we can
- …

# Методы борьбы

- Ручной анализ и фильтрация

# Методы борьбы

- Ручной анализ и фильтрация

| + | - |
| --- | --- |
| Высокая точность | Человеческий фактор |
| | Трудозатраты |
| | Стоимость |

# Методы борьбы

- Ручной анализ и фильтрация
- **Автоматический анализ и фильтрация**

# Методы борьбы

- Ручной анализ и фильтрация
- **Автоматический анализ и фильтрация**

| + | - |
|---|---|
| Нетрудозатрано | Сложные письма |
| Недорого | |
| Роботы, а не человеки | |

# Методы борьбы

- Ручной анализ и фильтрация
- **Автоматический анализ и фильтрация**

| + | - |
|---|---|
| Нетрудозатрано | Сложные письма |
| Недорого | |
| Роботы, а не человеки | |

# Определение

Спам = электронная почта, **...?**

# Определение

Спам = электронная почта, обладающая двумя свойствами:

- Нежелательная
- Массовая

# Машинное обучение

Какую задачу мы решаем?

- Кластеризация – обучение без учителя
- Классификация – обучение с учителем

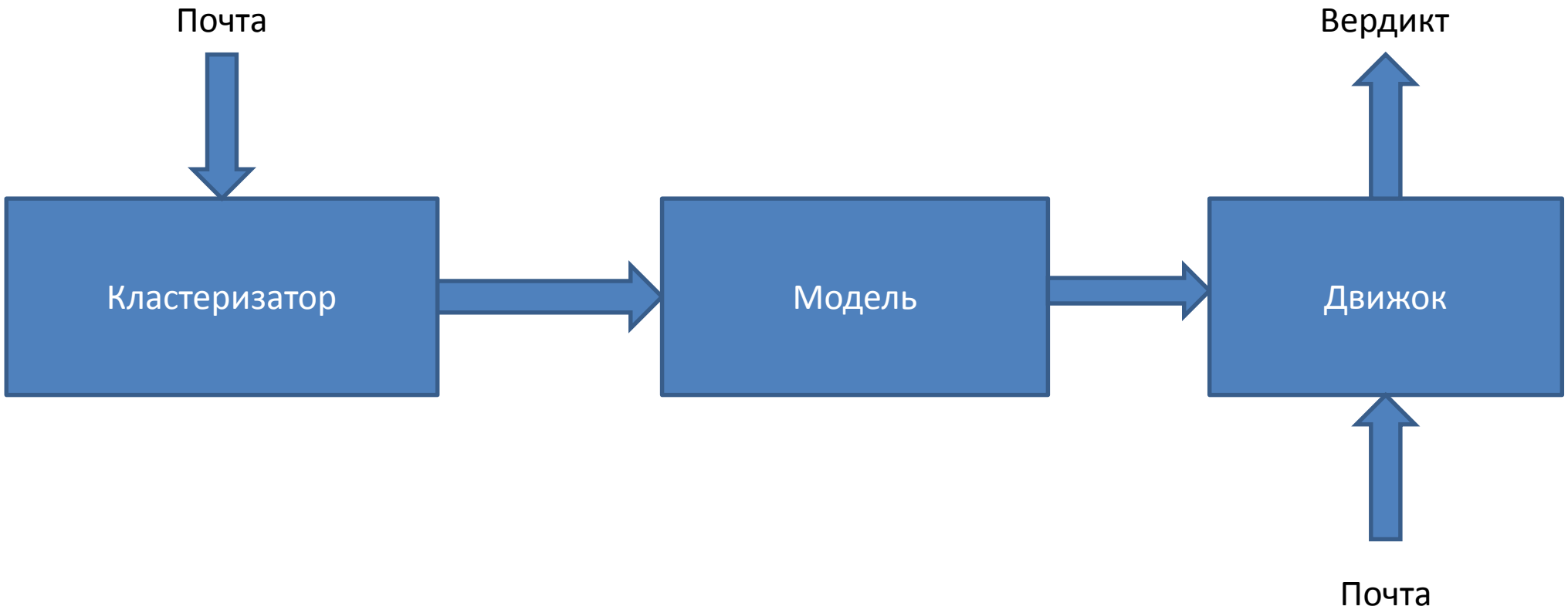
# Машинное обучение

Какую задачу мы решаем?

- **Кластеризация – обучение без учителя**
- Классификация – обучение с учителем

# Постановка задачи

# Постановка задачи

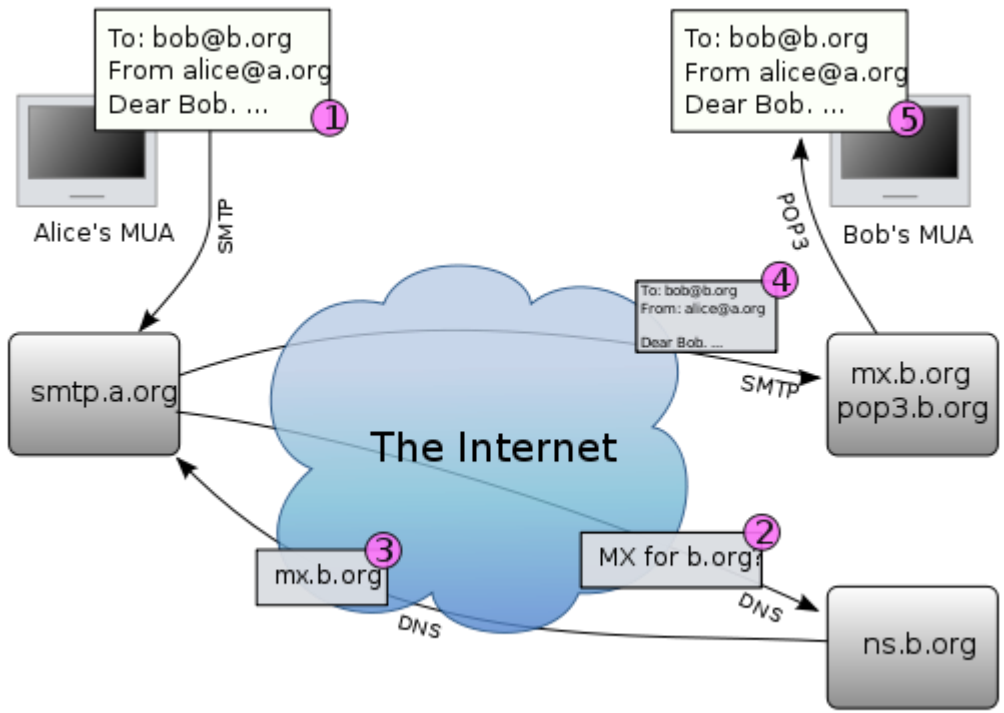


# Модель

- На выходе кластеризатора: множество скластеризованных рассылок
- Для каждой мы построим структуру, ее описывающую
- Построенную структуру сможет интерпретировать наш движок, сканируя поступающую в него почту

# Выбор признаков

# Простейшая схема отправки почты



# Simple Mail Transfer Protocol (SMTP)

- RFC 821, RFC 5321
- Простой протокол, операции команда/ответ

# SMTP: пример сессии

- S: 220 smtp.kremlin.ru ESMTP Postfix
- C: HELO relay.msu.ru
- S: 250 Hello relay.msu.ru

# SMTP: пример сессии

- C: MAIL FROM:<student@msu.ru>
- S: 250 Ok

# SMTP: пример сессии

- C: RCPT TO:<putin@kremlin.ru>
- S: 250 Ok

# SMTP: пример сессии

- C: DATA
- S: 354 End data with <CR><LF>.<CR><LF>
- C: From: “MSU Student" <student@msu.ru>
- C: To: “Vladimir Putin" <putin@kremlin.ru>
- C: Date: Th, 20 February 2014 18:00:00
- C: Subject: Test message
- C:
- C: TESTTESTTEST

# Electronic Mail (Email)

- RFC 5322, RFC 2045-2049
- Multipurpose Internet Mail Extensions (MIME)
- Конверта, заголовок и тело сообщения.
- Поддерживает вложения произвольных типов

# Mime: пример

From: MSU Student <student@msu.ru>

MIME-Version: 1.0

Content-Type: multipart/mixed; boundary=“--THIS_IS_SPARTA"

**Это сообщение из нескольких частей в MIME формате.**

--THIS_IS_SPARTA

Content-Type: text/plain

**Это тело сообщения**

--THIS_IS_SPARTA

Content-Type: text/plain;

Content-Disposition: attachment; filename=“diploma.txt”

**Это текст вложения**

--THIS_IS_SPARTA

# За что нам зацепиться?

# За что нам зацепиться?

- Аномалии SMTP сессии

# За что нам зацепиться?

- Аномалии SMTP сесии
- Аномалии MIME

# За что нам зацепиться?

- Аномалии SMTP сессии
- Аномалии MIME
- Контент письма

# За что нам зацепиться?

- Аномалии SMTP сессии
- Аномалии MIME
- Контент письма

# MIME

- Пустое поле To

*From: "2014-01-13 14:44:30" <2079965388@qq.com>*

***To:***

*Subject:*
*=?gb2312?B?xOO6w6OhzOEguakgyKsgufogu/ogtPKhts2oINPDobehtg==?=*

*….*

# MIME

- Нераскрывшаяся переменная в Message-ID

*Message-ID:*

*<%**RNDDIGIT1025**.%**RNDDIGIT15**%**RNDLCCHAR15**%**RNDDIGIT110**%**RNDLCCHAR13**@yahoo.com>*

# MIME

- Отсутствие поля Date при заголовке Microsoft Outlook Express

*Message-ID: <201421019429.22097@dtp.com>*
*From: PayPal  <mail@dtp.com>*
*Subject: Account Reviewed*
***X-Mailer:*** *Microsoft Outlook Express 15.0.1823*
*X-MimeOLE: Produced By Microsoft MimeOLE V5.00.2553.956*
*MIME-Version: 1.0*
*Content-Type: multipart/mixed;*
*boundary="--12940«*
*…*

# MIME

- Аномалии HTML:
  - Белый текст
  - Битые теги
  - Замена одинаковых букв в различных раскладках (c/c)
- Картинки
- …

# Представление аномалий MIME

- Булевый вектор с 0/1 (не встречается/встречается) на месте соответствующей аномалии
- Каждое письмо – точка в N-мерном пространстве, где N – количество используемых нами аномалий
- Но есть характеристики, не лежащие в булевом диапазоне 0/1:
  - Размер письма в байтах
  - Размер вложений в байтах
  - Количество ссылок в письме
  - …

# Представление аномалий MIME

- Возможные решения:
  - Нормализация: логарифмическая шкала, …
  - Представление в виде квантилей
  - …

# Контент письма

- Существует множество способов анализа текстов:
  - TF/IDF
  - N-Grams
  - Различные статистические метрики схожести
  - LDA
  - ...

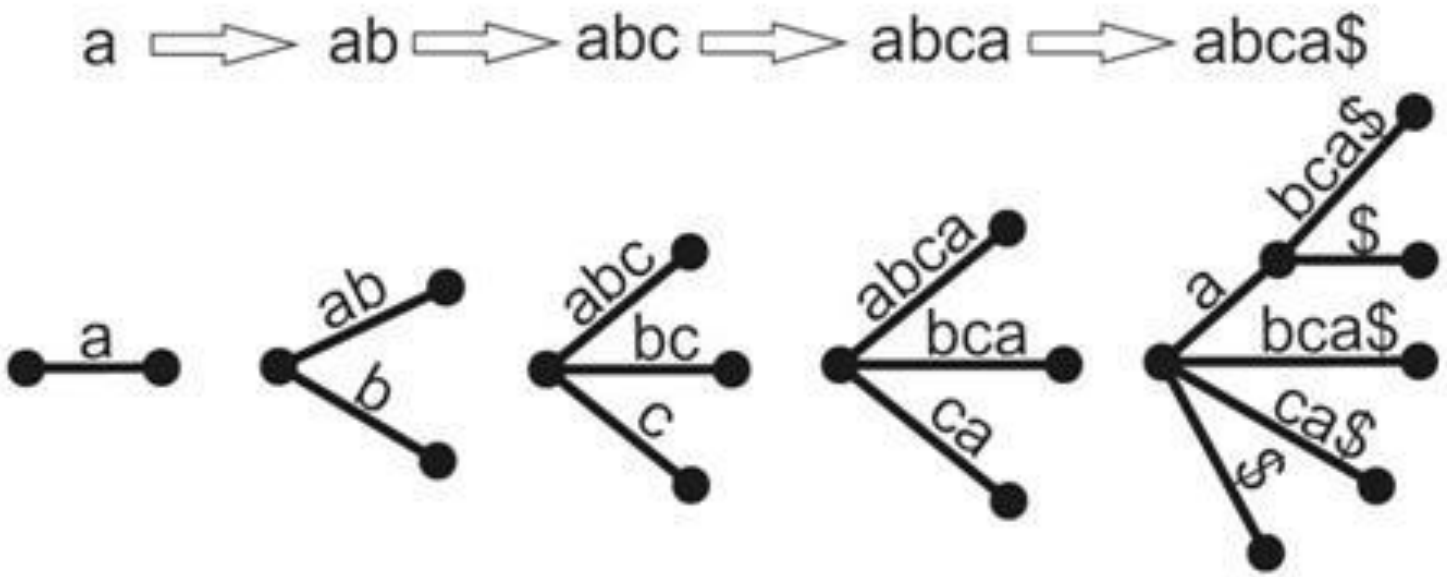
# Контент письма

- В рамках одной рассылки часто бывают письма с полностью не пересекающимися текстами (возможны различные языки в рамках одной рассылки)
- Дополнительная вычислительная сложность для нашего кластеризатора

# Контент письма

- Задача поиска наибольшей общей подстроки
  - Суффиксные массивы
  - Суффиксные деревья

# Контент письма



# Контент письма

- 1995 *Esko Ukkonen, "On-line construction of suffix trees“*
- Сложность:
  - O(n) при заранее известном размере алфавита
  - O(n*log n) при произвольном алфавите

# Контент письма

# Контент письма

- Будем строить минимальное множество терминов, максимально покрывающих скластеризованную рассылку:
  - Терминов должно быть минимальное возможное количество (сложность проверки)
  - Термины не должны быть чересчур большими (сложность проверки)
  - Термины не должны быть чересчур маленькими (возможность ошибок)

# Модель

- Список аномалий MIME, которым удовлетворяют все письма в рассылке
- Список терминов по контенту письма, максимально полно покрывающий рассылку

# Модель

<0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0 >

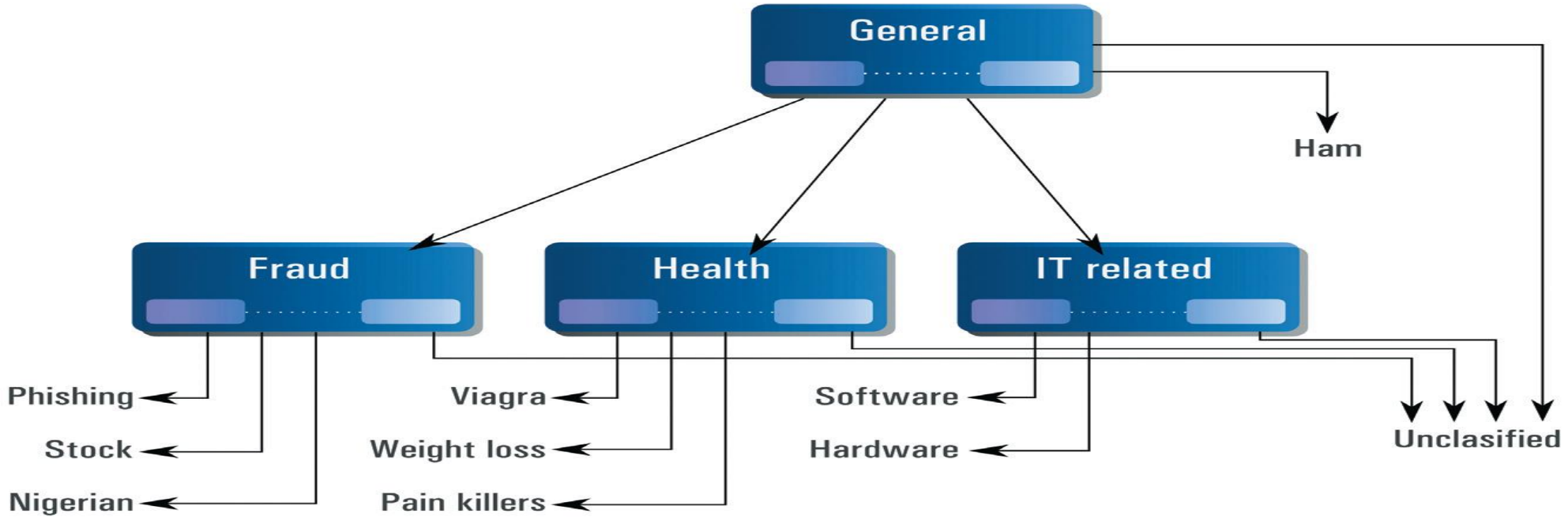
Это первый термин
Это второй
Это третий
…

# Выбор метода кластеризации

# BitDefender

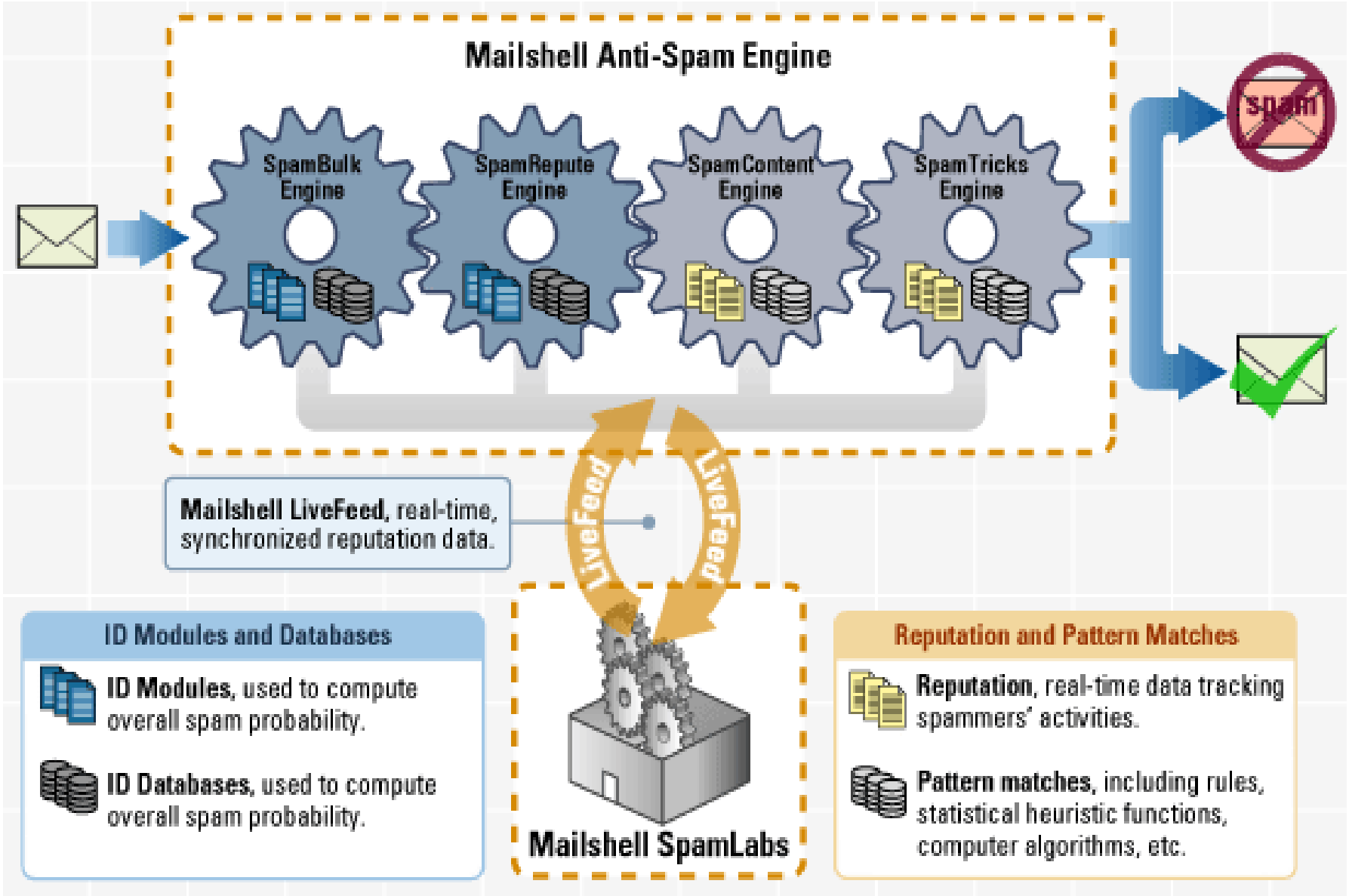
- Сканирование траффика
- Байесовский фильтр
- Эвристический фильтр
- Дерево нейросетей

# BitDefender




1. Run requested heuristics
2. Extract input vector
3. Reduce noise
4. Pass to the neural network
5. Classify

# Mailshell


Mailshell Anti-Spam Engine
SpamBulk Engine
SpamRepute Engine
SpamContent Engine
SpamTricks Engine
spam
Mailshell LiveFeed, real-time, synchronized reputation data.
LiveFeed
LiveFeed
ID Modules and Databases
ID Modules, used to compute overall spam probability.
ID Databases, used to compute overall spam probability.
Mailshell SpamLabs
Reputation and Pattern Matches
Reputation, real-time data tracking spammers' activities.
Pattern matches, including rules, statistical heuristic functions, computer algorithms, etc.

# Mailshell

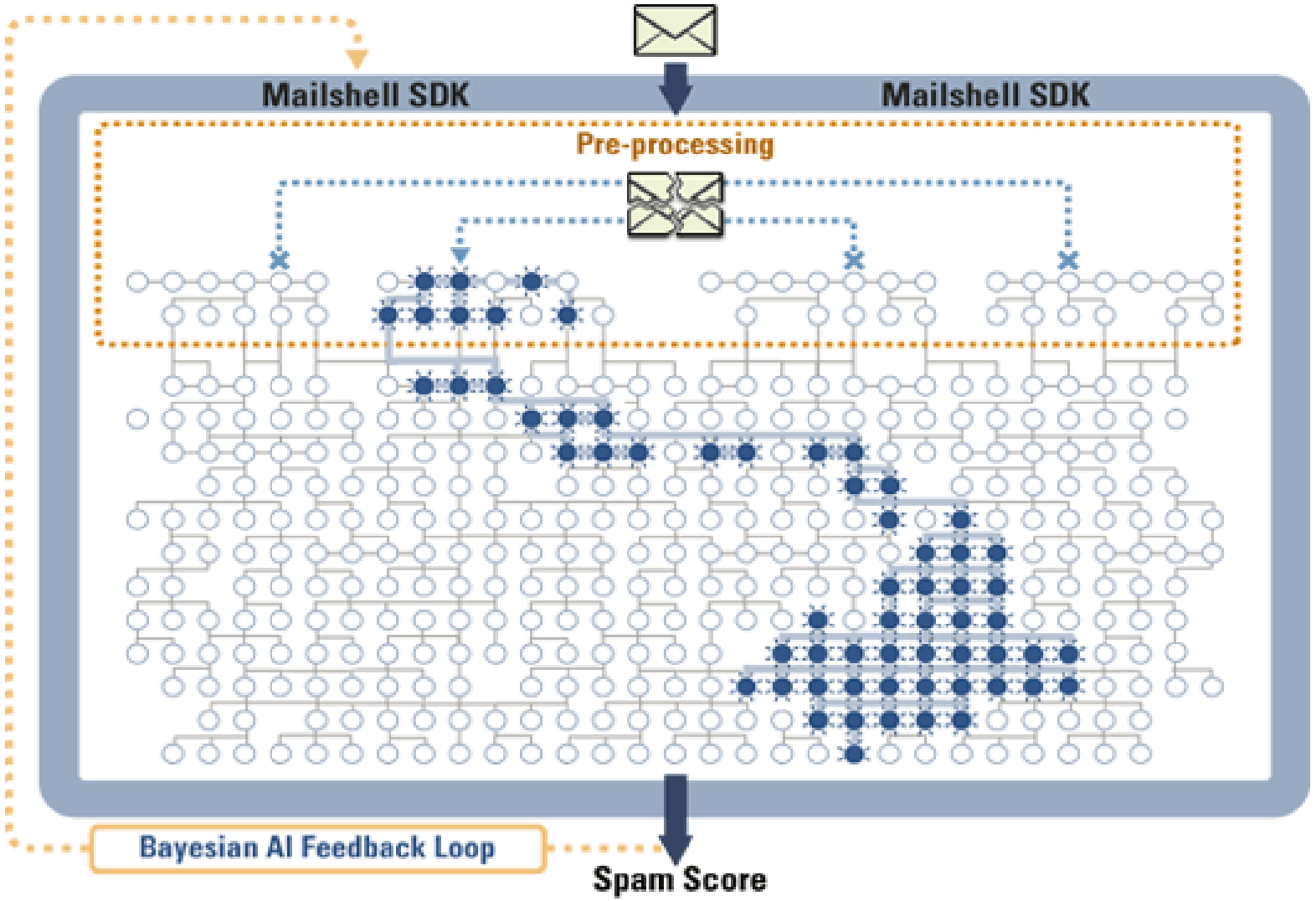
- Байесовский фильтр
- Вычисление репутации письма по трудно подделываемым характеристикам
- N-Grams
- SpamAdapt AI (patent) – real time переобучение используя **Genetic/Neural Fuzzy Logic Algorithm, Content-based Bayesian learning.**

# Mailshell

- Байесовский фильтр
- Вычисление репутации письма по трудно подделываемым характеристикам
- N-Grams
- SpamAdapt AI (patent) – real time переобучение, используя **Genetic/Neural Fuzzy Logic Algorithm, Content-based Bayesian learning.**

# WAT?

# Mailshell


Mailshell SDK
Mailshell SDK
Pre-processing
Bayesian AI Feedback Loop
Spam Score

# Выбор метода кластеризации

# Выбор метода кластеризации

- Онлайн обучение
- Итерационное

# Выбор метода кластеризации

| Онлайн | Итерационное |
| --- | --- |
| Проблема “запоминания” и “переобучения” | Немгновенная реакция (время обучения) |
| Большой объем хранимых данных | |
| Трудность “калибровки” | |

*Стоит посмотреть: BIRCH, E-SOINN*

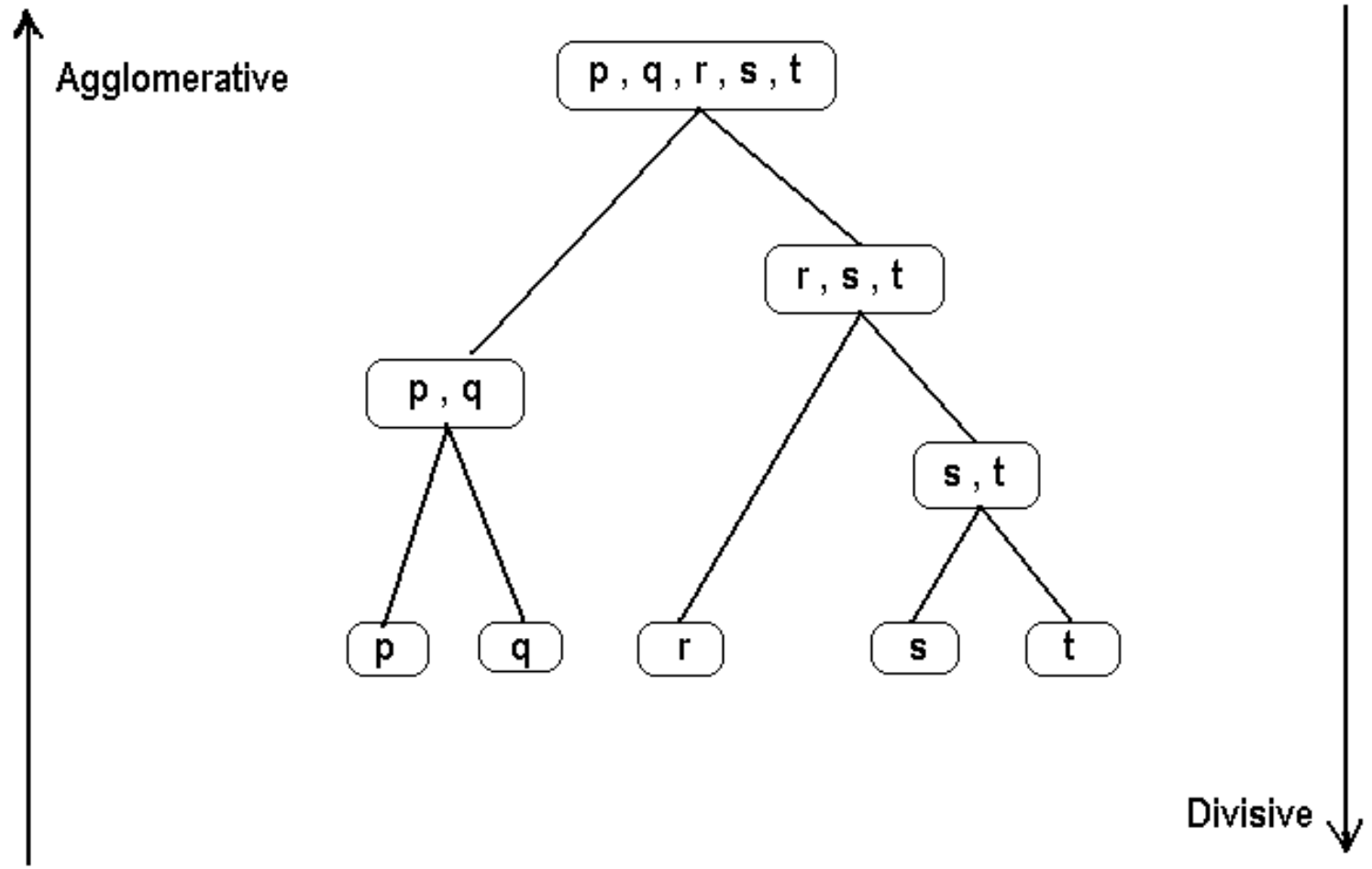
# Итерационное обучение

- С фиксированным количеством разбиений (K-Means)
- С нефиксированным количеством разбиений

# Итерационное обучение

- С фиксированным количеством разбиений
- С нефиксированным количеством разбиений:
  - Иерархические: последовательное построение кластеров на основании предыдущего построения
  - Неиерархические: оптимизация целевой функции

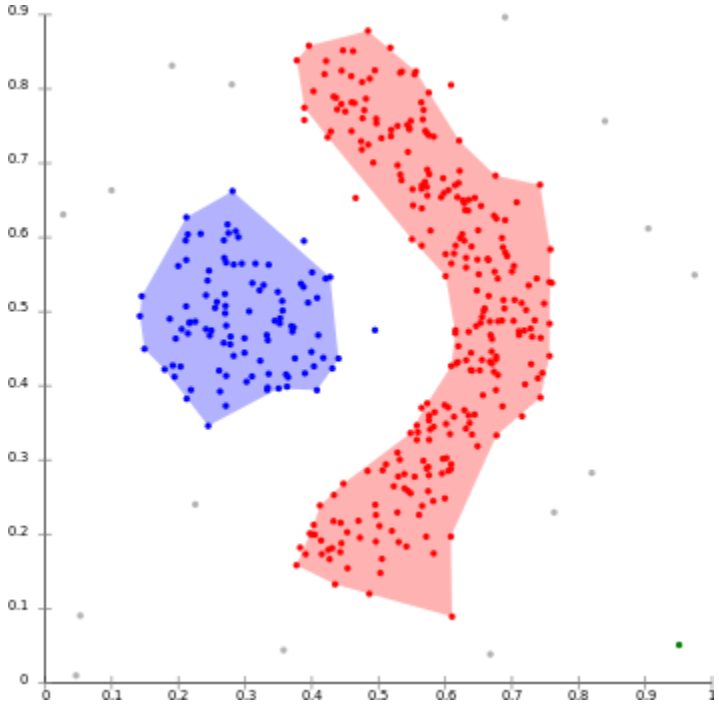
# Иерархическое обучение



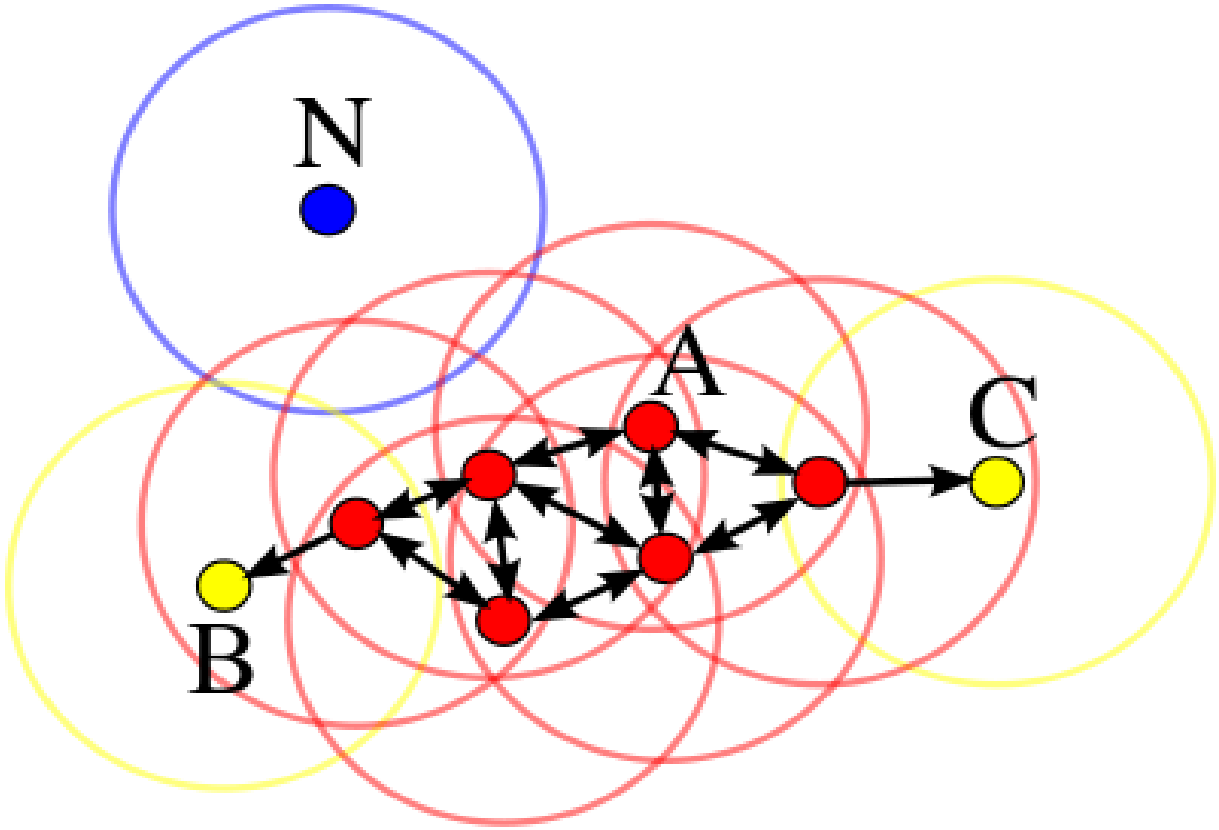
# Иерархическое обучение

- На каком моменте остановиться:
  - Подбор и фиксация значения
  - Попытка эвристической подборки

# Неиерархическое обучение: DBSCAN



# Неиерархическое обучение: DBSCAN



# Неиерархическое обучение: DBSCAN

- Два параметра, значение которых нам заранее неизвестно:
  - Размер рассматриваемой окрестности каждой точки
  - Минимальное количество необходимых точек в окрестности
- Различные модификации алгоритма, позволяющие эвристически вычислять значения (HDBSCAN)

# Метрика расстояния

- Метрики расстояний – взвешенная функция
- Подбор весов:
  - Градиентный спуск
  - Генетические алгоритмы
  - …

# Борьба с фолсами

- “Пропуск” спама менее значителен, чем блокирование легальной почты
- 100% точность недостижима

# Борьба с фолсами

- Формирование выборки белых писем, максимально полно покрывающих все возможные MIME-аномалии. Построенная нами модель никогда не должна блокировать их

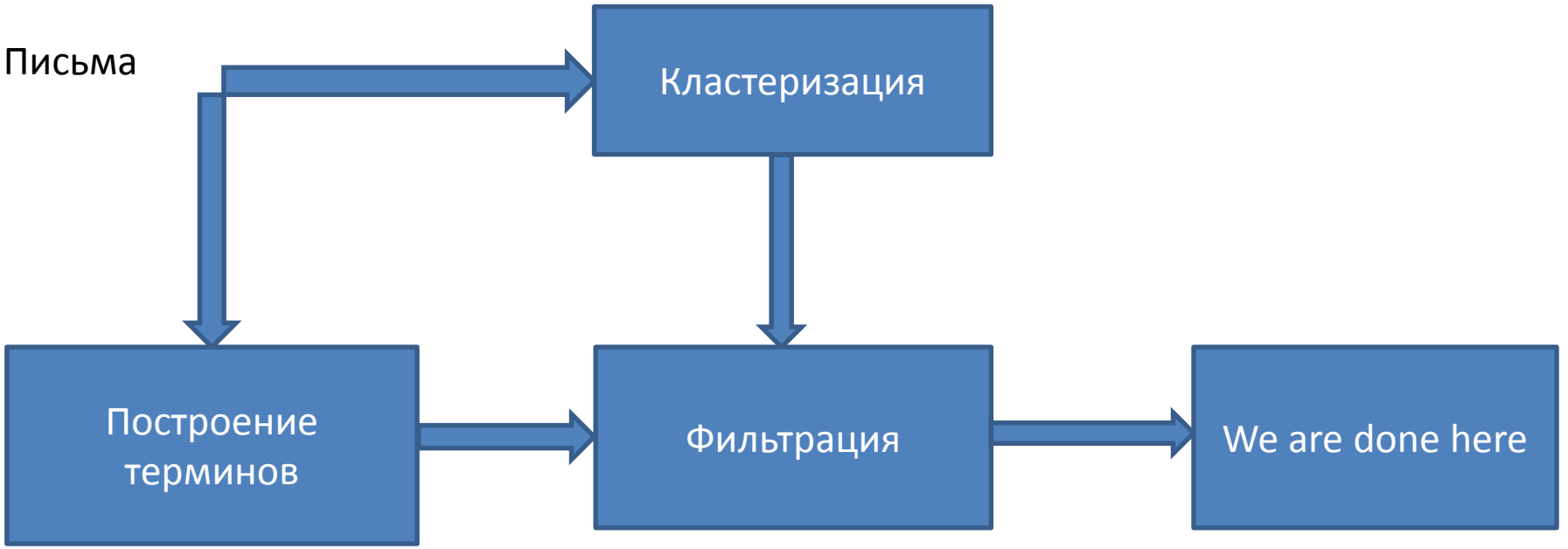
# Борьба с фолсами

- Формирование выборки белых писем, максимально полно покрывающих все возможные MIME-аномалии. Построенная нами модель никогда не должна блокировать их
- IP-репутация

# Борьба с фолсами

- Формирование выборки белых писем, максимально полно покрывающих все возможные MIME-аномалии. Построенная нами модель никогда не должна блокировать их
- IP-репутация
- Подмешивание в кластеризацию заранее известных белых писем

# Результаты


Письма
Кластеризация
Построение терминов
Фильтрация
We are done here

# Toolchain

- Python:
  - Numpy
  - Scipy
  - Scikit-Learn
  - Pandas
- Rapidminer
- Weka
- R

# Вопросы?


Я не понимаю... Я отправил им миллионы писем. Эти деньги принадлежат им, черт возьми!
© 2014 seemikedraw.com.au
Где-то в Нигерии
comicsbook.ru