# Информационная стеганография & машинное обучение

Слипенчук
Павел Владимирович
PavelSlipenchuk@stego.su
аспирант каф. ИУ-8
«Информационная безопасность»
МГТУ. им.Баумана

Москва
29.09.2016

# План

- История. Заметки
- Стеганография сегодня
- Цели стеганографии, практическое применение
- Примеры методов современной стеганографии
- Роль математики; машинного обучения

- Модель Грушо
- Модель стегоаналитического классификатора
- Актуальные проблемы в ML и стеганографии

- Ответы на вопросы

# История. Заметки.

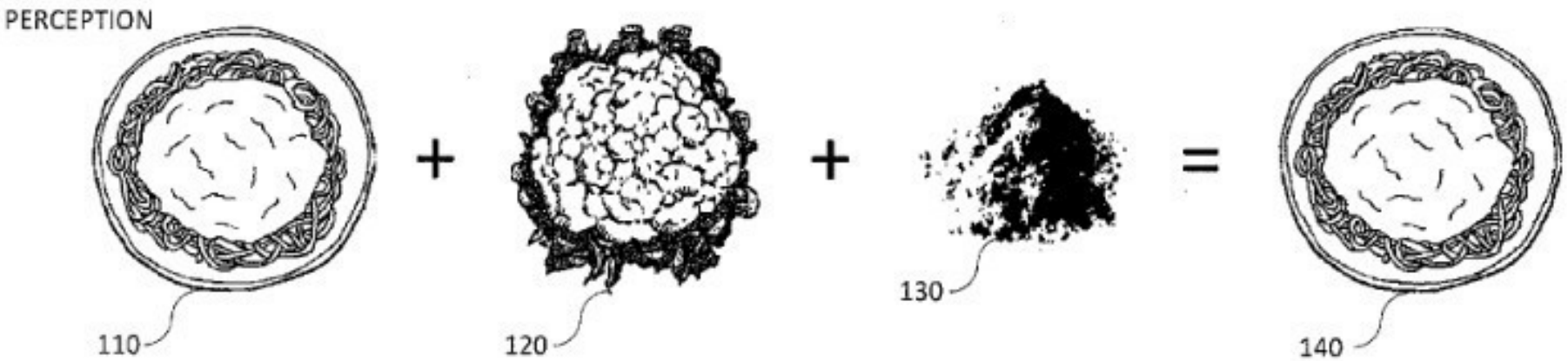
- Геродот «История» (V в. до Р.Х.)
- Эней Тактик «О перенесении Осады» глава «О тайных письмах» (IV в. до Р.Х)
- Иоанн Тритемий пишет книгу «Steganographia» (1499). **τεγανός** (скрытый) + **γράφω** (пишу) = «*Скрытопись*»*
- Симпатические чернила (Филон Александрийский I в.)
- Микроточки (XX в)
- Появление вычислительной техники и её повсеместное распространение.

* Слово «тайнопись» уже занято ***криптографией***

# Среда

- Среда – это канал (в самом широком смысле), внутри которого происходит передача стеганографического сообщения
- Среда может иметь совершенно различную природу. Поэтому стеганография – это *междисциплинарная* наука и искусство.
- Примеры сред: изображения; файловая система; коды, исправляющие ошибки; акротексты; просодии человеческой речи; и т.д.

# Пример среды.
# Just4fun Food steganography




Kursh K. и Lav R. Varchney «Продовольственная стеганография»

# Just4fun

# Food steganography

(19) **United States**

(12) **Patent Application Publication**

**Varshney et al.**

(10) **Pub. No.: US 2015/0059438 A1**

(43) **Pub. Date: Mar. 5, 2015**

(54) **FOOD STEGANOGRAPHY**

(71) Applicant: **International Business Machines Corporation**, Armonk, NY (US)

(72) Inventors: **Kush R. Varshney**, Ossining, NY (US); **Lav R. Varshney**, Yorktown Heights, NY (US)

(73) Assignee: **International Business Machines Corporation**, Armonk, NY (US)

(21) Appl. No.: **14/011,421**

(22) Filed: **Aug. 27, 2013**

**Publication Classification**

(51) **Int. Cl.**

***G01N 33/00*** (2006.01)

***G01N 33/02*** (2006.01)

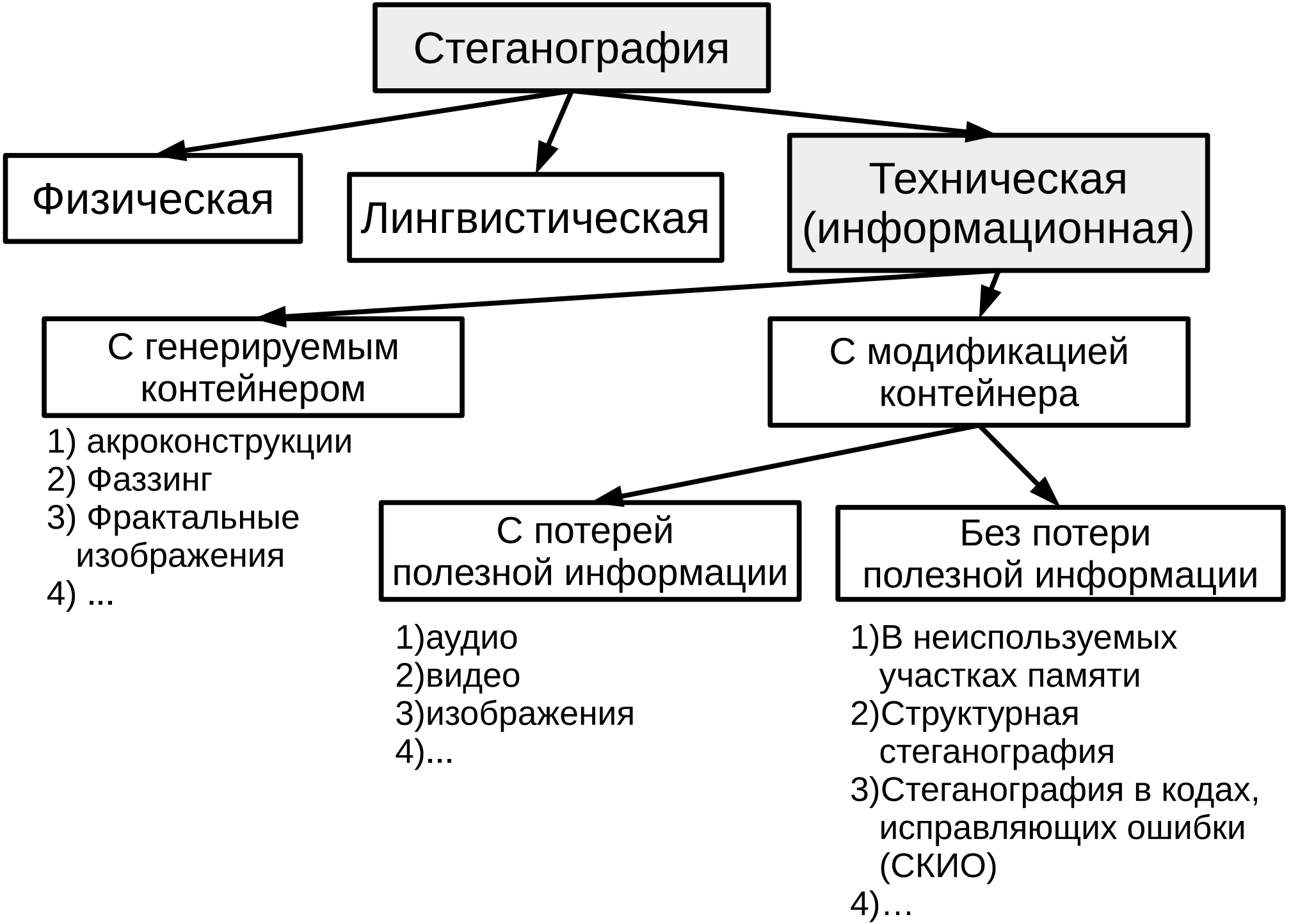
(52) **U.S. Cl.**

CPC ..... ***G01N 33/0001*** (2013.01); ***G01N 33/02*** (2013.01)

USPC ..... **73/23.34**

(57) **ABSTRACT**

The present disclosure relates to methods and systems for calculating a food additive. A first method includes identifying chemical compounds of an averse food ingredient, identifying chemical compounds of a flavorful food ingredient and calculating a set of chemical compounds for the food additive such that an olfactory perception of a mixture of the averse food ingredient, the flavorful food ingredient and the food additive is the same as an olfactory perception of only the flavorful food ingredient. A first device includes a database storing information identifying chemical compounds of an averse food ingredient and identifying chemical compounds of a flavorful food ingredient, and a processor for calculating a food additive such that an olfactory perception of flavors of a mixture of the averse food ingredient, the flavorful food ingredient and the food additive is the same as an olfactory perception of only the flavorful food ingredient.


Стеганография
Физическая
Лингвистическая
Техническая (информационная)
С генерируемым контейнером
С модификацией контейнера
1) акроконструкции
2) Фаззинг
3) Фрактальные изображения
4) ...
С потерей полезной информации
Без потери полезной информации
1)аудио
2)видео
3)изображения
4)...
1)В неиспользуемых участках памяти
2)Структурная стеганография
3)Стеганография в кодах, исправляющих ошибки (СКИО)
4)…

# Стеганография сегодня

# Стеганография сегодня

- Big Data – очень, очень много данных
- Огромное количество протоколов передачи и хранения данных, файловых систем, операционных систем
- Интернет
- Интернет – IoT
- Интернет – большое количество участников, каналов передачи данных

Все вышеперечисленное создает хорошее подспорье для стеганографии.

# Три цели информационного сокрытия

- **Скрытая передачи или хранения данных (СПД)** – только это стеганография («скрытопись») в строгом смысле
- **Водяные знаки (ВЗ, «digital watermarking»)** – определенные метки, *одинаковые* для каждой копии
- **Цифоровые отпечатки (ЦО, «stego fingerprinting»)** – определенные метки, *различные* для каждой копии.

# ЦО vs ВЗ

- Необходимо различать ЦО и ВЗ!

  Хотя бы по причине *атаки сговором*.

  ВЗ может быть стеганографией: стеганографический ВЗ (СВЗ), ЦО – нет.

- **Атака сговором**.

  Беруться **n** копий контейнера и из них создается одна копия – побитовая XOR каждой из них.

**Замечание 1.** Термин «информационное сокрытие» не устоялся в рускоязычной литературе и часто ЦО, ВЗ так же называют «стеганографией»

**Замечание 2**. Есть термин «ЦВЗ» ( цифровой водяной знак).
Это то ЦО, то ВЗ. Иногда и то и другое одновременно в одной статье ;)

# Практическое применение

1. Незаметная передача информации (СПД)
2. Скрытое хранение информации (СПД)
3. Недекларированное хранения информации (СПД)
4. Защита исключительного права (ЦО)
5. Защита авторского права (ВЗ)

# Практическое применение

6. Защита подлинности документов (ВЗ)
7. Индивидуальный отпечаток в СЭДО (ЦО)
8. Водяной знак в DLP системах (ВЗ)
9. Скрытая передача управляющего сигнала (СПД)
10. Стеганографические botnet-сети (СПД)

# Практическое применение

11. Неотчуждаемость информации (ВЗ)
12. Подтверждение достоверности переданной информации(ЦО)
13. Funkspiel («Радиоигра») (СПД)
14. Стеганографическое отслеживание (СПД)
15. Стеганографическое отвлечение (?)

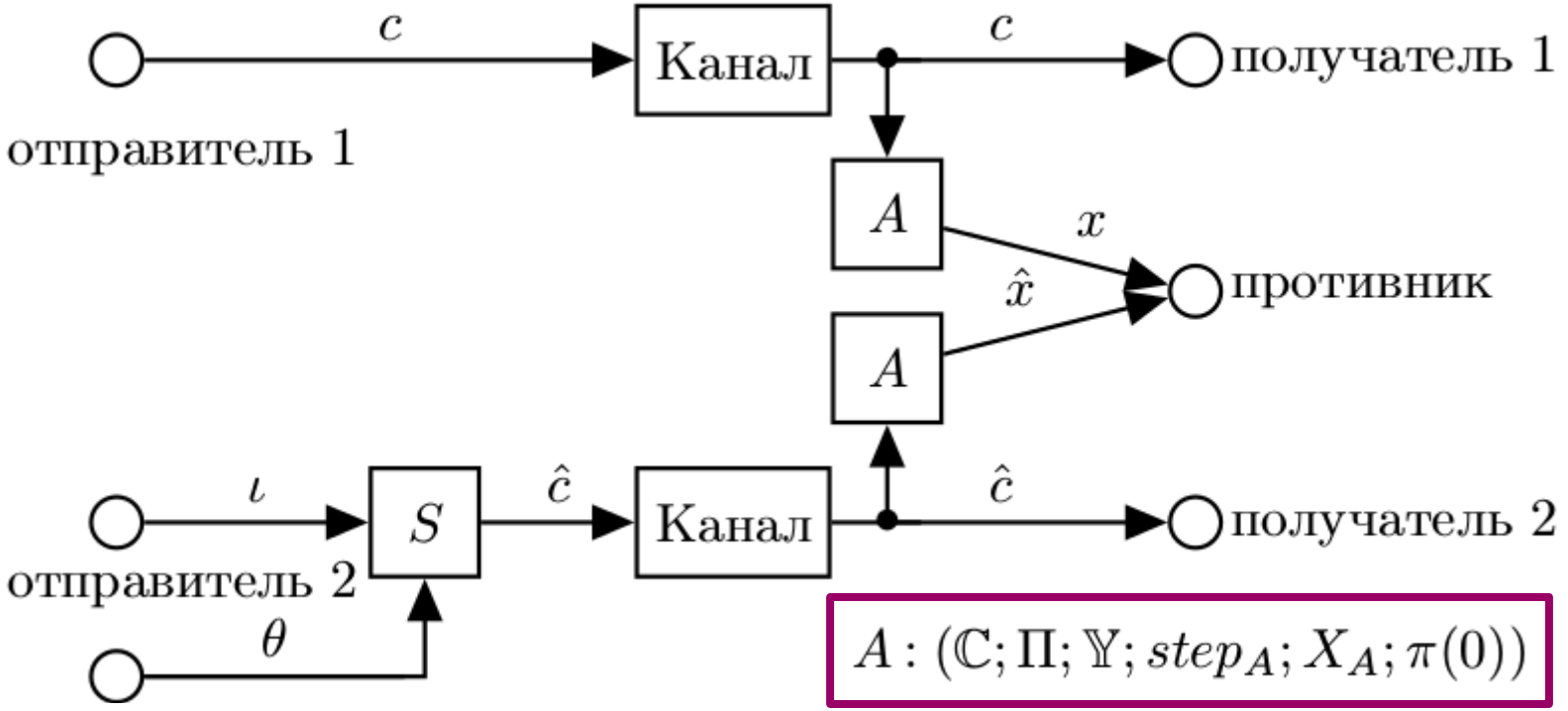
# Machine Learning + стегоанализ

**Как скрестить ужа с ежом?**

Варианты:

- Извлечение признаков (Data Mining) из стеганографических контейнеров и решение задачи классификации.
- Искусственные нейронные сети

# Модель Грушо




1. $\mathbb{C}$ – множество допустимых контейнеров $c$ и $\hat{c}$ в канале;
2. $\mathbb{X}$ – конечное множество допустимых выходных значений $x$ передаваемые противнику
3. $\Pi = \{\pi_1, \pi_2, ..., \pi_{\|\Pi\|}\}$ – конечное множество всевозможных состояний автомата $A$;
4. $step_A : \Pi \times \mathbb{C} \longmapsto \Pi$ – функция переходов, переводящая из состояния $\pi(i)$ в состояние $\pi(i+1)$ в зависимости от прошлого состояния $(\pi(i))$ и полученного контейнера $c$ или $\hat{c}$;
5. $X_A : \Pi \times \mathbb{C} \longmapsto \mathbb{X}$ – функция возвращающее значение $x$ противнику в зависимости от состояния системы и в зависимости от полученного $c$ или $\hat{c}$.
6. $\pi(0) \in \Pi$ – начальное состояние автомата $A$.

# Стойкая стегосистема в сильном смысле и в слабом смысле

- Вероятность выбрать $c$ для отправителя 1

$$P_{\mathbb{C}}(c) > 0 \qquad \sum_{\forall c \in \mathbb{C}} P_{\mathbb{C}}(c) = 1$$

- Вероятность выбрать $\iota$ для отправителя 2

$$P_{\mathbb{I}}(\iota) > 0 \qquad \sum_{\forall \iota \in \mathbb{I}} P_{\mathbb{I}}(\iota) = 1$$

$$S : \mathbb{I} \times \Theta \longmapsto \mathbb{C}$$

$$A : (\mathbb{C}; \Pi; \mathbb{Y}; step_A; X_A; \pi(0))$$

# Стойкая стегосистема в сильном смысле и в слабом смысле

$$Q_\pi(x) \stackrel{\text{def}}{=} P_{\mathbb{C}}\{c : X_A(c,\pi) = x\}$$

$$\hat{Q}_\pi(x) \stackrel{\text{def}}{=} P_{\mathbb{C}}\{\hat{c} : X_A(\hat{c},\pi) = x\}$$

Введем функцию $\zeta(A) = 1$ если верно высказывание $A$, иначе $\zeta(A) = 0$.

$$Q_\pi(x) = \sum_{\forall c \in \mathbb{C}} P_{\mathbb{C}}(c) \cdot \zeta\left(X_A(c,\pi) = x\right)$$

$$\hat{Q}_\pi(x) = \sum_{\forall \iota \in \mathbb{I}} P_{\mathbb{I}}(\iota) \cdot \zeta\left(\left(X_A\left(S(\iota,\theta),\pi\right)\right) = x\right)$$

# Стойкая стегосистема в сильном смысле и в слабом смысле

Будем называть систему ***стойкой для состояния**** $\pi$ если:

$$\hat{Q}_\pi(x) = Q_\pi(x)$$

Система ***стойкая в слабом смысле***, если:

$$\forall x \in \mathbb{X} \Longrightarrow \hat{Q}_\pi(x) = Q_\pi(x)$$

* В оригинальной работе Грушо: «противник не видит канал в стостоянии $\pi$»

# Стойкая стегосистема в сильном смысле и в слабом смысле

Определим функцию $\overline{X}_A(c)$ как отображение из $\mathbb{C}$ в $\mathbb{X}^{\|\Pi\|}$, полагая, что:

$$\overline{X}_A(c) \stackrel{\text{def}}{=} \left(X_A(c|\pi_1), X_A(c|\pi_2), ..., X_A(c|\pi_{\|\Pi\|})\right)$$

Определим функции

$$\mathbf{Q}(\overline{x}) \stackrel{\text{def}}{=} P_{\mathbb{C}}\{c : \overline{X}_A(c) = \overline{x}\} \quad \hat{\mathbf{Q}}(\overline{x}) \stackrel{\text{def}}{=} P_{\mathbb{C}}\{\hat{c} : \overline{X}_A(\hat{c}) = \overline{x}\}$$

Аналогично:

$$\mathbf{Q}(\overline{x}) = \sum_{\forall c \in \mathbb{C}} P_{\mathbb{C}}(c) \cdot \zeta\left(\overline{X}_A(c) = \overline{x}\right)$$

$$\hat{\mathbf{Q}}(\overline{x}) = \sum_{\forall \iota \in \mathbb{I}} P_{\mathbb{I}}(\iota) \cdot \zeta\left(\left(\overline{X}_A\left(S(\iota, \theta)\right)\right) = \overline{x}\right)$$

Стеганографическую систему будем называть ***стойкой в сильном смысле***, если:

$$\forall \overline{x} \in \mathbb{X}^{\|\Pi\|} \Longrightarrow \hat{\mathbf{Q}}(\overline{x}) = \mathbf{Q}(\overline{x})$$

# Теорема Грушо (1999)

Обозначим за $domain(\overline{x})$ множество таких $c \in \mathbb{C}$, для которых $\overline{X}_A(c) = \overline{x}$.

$$domain_A(\overline{x}) = \{c \in \mathbb{C} : \overline{X}_A(c) = \overline{x}\}$$

**Теорема (Грушо, 1999)** . Если для любого $\overline{x} \in \mathbb{X}^{\|\Pi\|}$ количество элементов в $domain_A(\overline{x})$ не меньше мощности конечного множества $\mathbb{I}$, то существует стегосистема, стойкая в сильном смысле. Иначе говоря, если верно выражение

$$\forall \overline{x} \in \mathbb{X}^{\|\Pi\|} \Longrightarrow \|domain_A(\overline{x})\| \geqslant \|\mathbb{I}\|$$

то существует хотя бы одна функция $S$
создающая *стойкую в сильном смысле* стегосистему.

**Следствие**. Если стегосистема стойкая в сильном смысле, то:

$$\|\Pi\| \cdot \|\mathbb{I}\| \cdot \|\mathbb{X}\| \geqslant \|\mathbb{C}\|$$

# Скрещивание

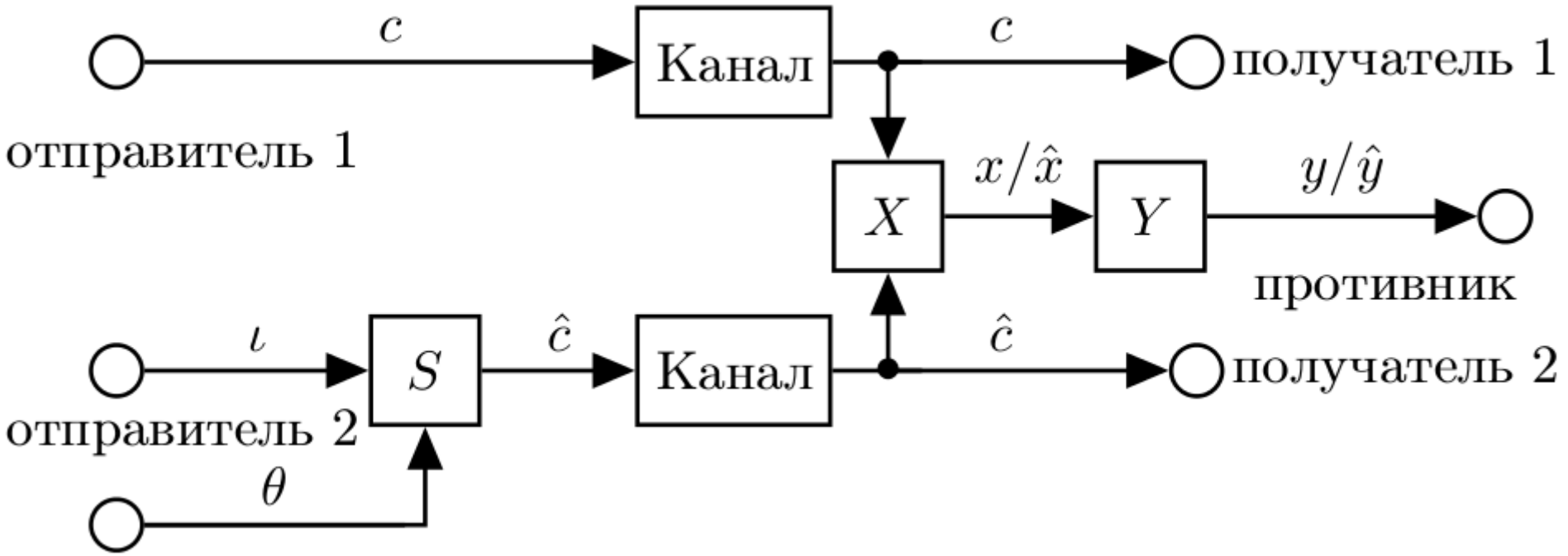
**Идеи**:

1) Количество состояний – одно. **(!)**
2) X(c) заменяем на Y(X(c)) – Функция извлечения признаков + классификатор
3) Задача стеганографии – классификатор должен быть «плохим»
4) Задача стегоанализа – классификатор должен быть «хорошим»

# Скрещивание

- Модель Грушо и теорема Грушо.
- Стойкая в сильном смысле = стойкая в слабом смысле (т. к. всего 1 состояние)
- Можно использовать мат.аппарат и практические результаты Machine Learning.
- Система работает даже если неизвестен алгоритм стеганографии, если признаки информативны.

# «Модель классификатора без памяти»




**X** – функция извлечения признаков
**Y** – решающий классификатор

# Обозначения

$y = +1$ (наличие стеганографии)

$y = -1$ (отсутствие стеганографии)

$P_Y(y_a|y = y_b)$ – вероятность того, что контейнер $y_a$ когда классификатор $Y$ присвоил контейнеру класс $y_b$

Формально можно определить формулой Байеса:

$$P_Y(-1|y = y_0) \stackrel{\text{def}}{=} \frac{\sum\limits_{\forall c \in \mathbb{C}} P(c)\zeta\left(Y(X(c)) = y_0\right)}{\sum\limits_{\forall c \in \mathbb{C}} P(c)\zeta\left(Y(X(c)) = y_0\right) + \sum\limits_{\forall \hat{c} \in \hat{\mathbb{C}}} P(\hat{c})\zeta\left(Y(X(\hat{c})) = y_0\right)}$$

$$P_Y(+1|y = y_0) \stackrel{\text{def}}{=} \frac{\sum\limits_{\forall \hat{c} \in \hat{\mathbb{C}}} P(\hat{c})\zeta\left(Y(X(\hat{c})) = y_0\right)}{\sum\limits_{\forall c \in \mathbb{C}} P(c)\zeta\left(Y(X(c)) = y_0\right) + \sum\limits_{\forall \hat{c} \in \hat{\mathbb{C}}} P(\hat{c})\zeta\left(Y(X(\hat{c})) = y_0\right)}$$

$P(c)$ – вероятность пустого контейнера (передача между отправителем 1 и получателем 1)

$P(\hat{c})$ – вероятность стегоконтейнера (передача между отправителем 2 и получателем 2)

# Обозначения

При равновероятных контейнерах и стегоконтейнерах ( $P(c) \equiv P(\hat{c}) \equiv const$)

$$P_Y(-1|y=y_0) \stackrel{\text{def}}{=} \frac{\sum\limits_{\forall c \in \mathbb{C}} \zeta\left(Y(X(c))=y_0\right)}{\sum\limits_{\forall c \in \mathbb{C}} \zeta\left(Y(X(c))=y_0\right) + \sum\limits_{\forall \hat{c} \in \hat{\mathbb{C}}} \zeta\left(Y(X(\hat{c}))=y_0\right)}$$

$$P_Y(+1|y=y_0) \stackrel{\text{def}}{=} \frac{\sum\limits_{\forall \hat{c} \in \hat{\mathbb{C}}} \zeta\left(Y(X(\hat{c}))=y_0\right)}{\sum\limits_{\forall c \in \mathbb{C}} \zeta\left(Y(X(c))=y_0\right) + \sum\limits_{\forall \hat{c} \in \hat{\mathbb{C}}} \zeta\left(Y(X(\hat{c}))=y_0\right)}$$

Индекс Джини классификатора $Y$ для класса $y_0 \in \mathbb{Y}$ определяется по формуле:

$$\delta_Y(y_0) \stackrel{\text{def}}{=} 1 - \sum_{\forall y_i \in \mathbb{Y}} \left(P_Y(y_i|y=y_0)\right)^2$$

В частности при $Y = \{-1, +1\}$ имеем:

$$\delta_Y(+1) = 1 - \left(P_Y(-1|y=+1)\right)^2 - \left(P_Y(+1|y=+1)\right)^2$$

$$\delta_Y(-1) = 1 - \left(P_Y(-1|y=-1)\right)^2 - \left(P_Y(+1|y=-1)\right)^2$$

# Стойкость по Джини и по Грушо

**Определение.** Стеганографическую систему будем называть ***стойкой по Джини относительно функции извлечения признаков*** $X$ ***и классификатора*** $Y$, если

$$\delta_Y(+1) = \delta_Y(-1) = \frac{1}{2}$$

**Определение.** Стеганографическую систему будем называть $\epsilon$***-стойкой по Джини относительно функции извлечения признаков*** $X$ ***и классификатора*** $Y$, если

$$\forall y \in \mathbb{Y} = \{-1, +1\} \Longrightarrow \delta_Y(y) \in \left[\frac{1}{2} - \frac{\epsilon}{2}, \frac{1}{2} + \frac{\epsilon}{2}\right]$$

Определения

$$Q(y) \stackrel{\text{def}}{=} P_{\mathbb{C}}\{c : Y(X(c)) = y\} \qquad \hat{Q}(y) \stackrel{\text{def}}{=} P_{\mathbb{C}}\{\hat{c} : Y(X(\hat{c})) = y\}$$

стегосистему будем называть ***стойкой (по Грушо)*** если:

$$\forall y \in \{+1, -1\} \Longrightarrow \hat{Q}(y) = Q(y)$$

# Леммы

- Любая стегосистема, стойкая по Джини так же стойкая по Грушо.
- Если классы сбалансированы, то любая стойкая по Грушо, стойкая по Джини.

# От «теорвера» к «статистике»

- Проводим исследование и находим доли. Рассчитываем Индекс Джини.

$$\delta_Y(y_0) \stackrel{\text{def}}{=} 1 - \sum_{\forall y_i \in \mathbb{Y}} \left(\Delta_Y(y_i|y = y_0)\right)^2$$

Аналогично для Q можно вычислить статистически:

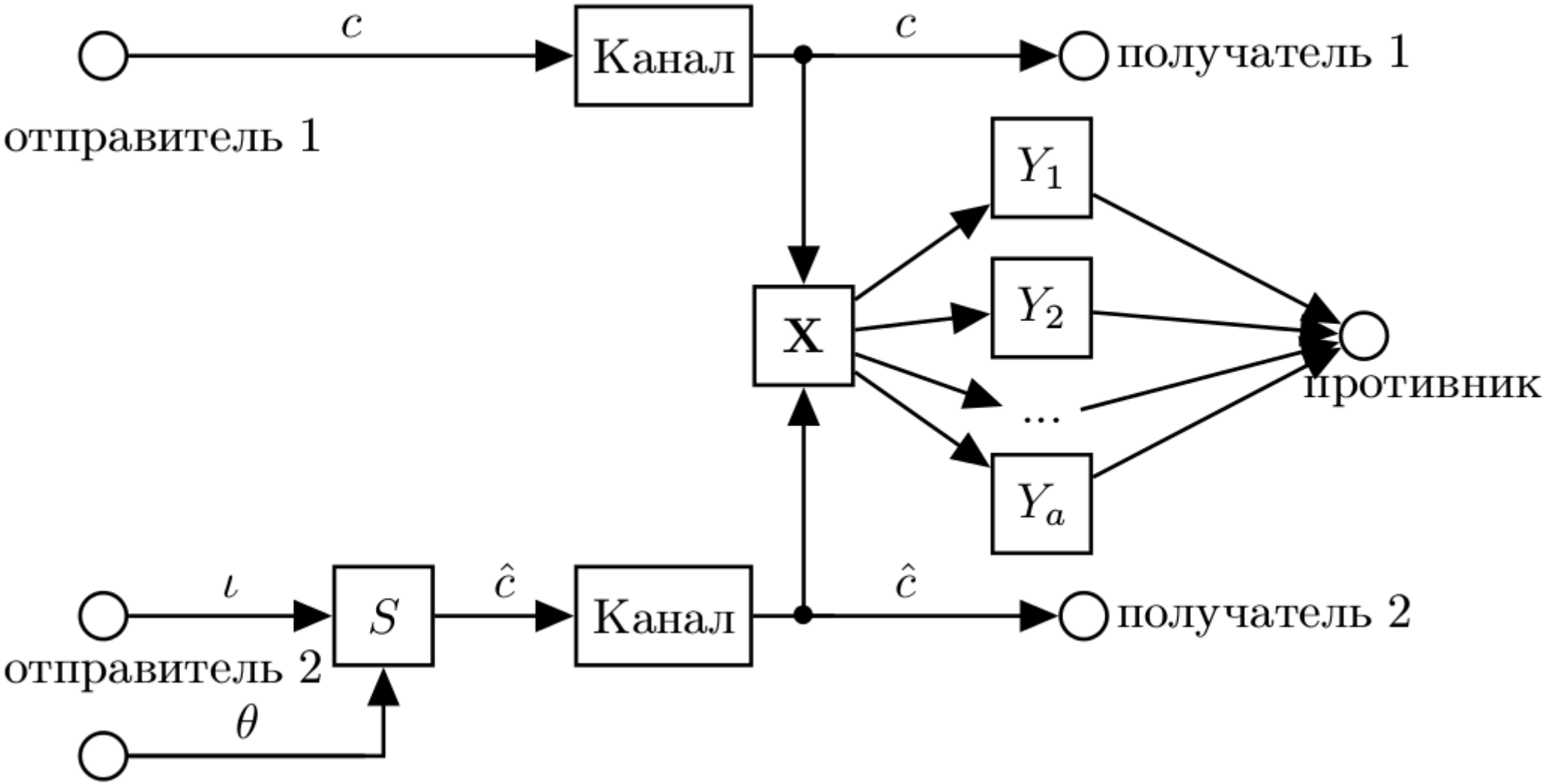
$Q(-1)$ – вероятность поместить пустой контейнер в класс $y = -1$

$Q(+1)$ – вероятность поместить пустой контейнер в класс $y = +1$

$\hat{Q}(-1)$ – вероятность поместить стегоконтейнер в класс $y = -1$

$\hat{Q}(+1)$ – вероятность поместить стегоконтейнер в класс $y = +1$

# Модель классификатора без памяти



# Стойкость

$$\mathbf{y} \stackrel{\text{def}}{=} \mathbf{Y}(\mathbf{x}) \stackrel{\text{def}}{=} (Y_1(\mathbf{x}), Y_2(\mathbf{x}), ..., Y_a(\mathbf{x}))$$

$$\mathbf{Q}(\mathbf{y}) \stackrel{\text{def}}{=} P_{\mathbb{C}}\{c : \mathbf{Y}(\mathbf{X}(c)) = \mathbf{y}\}$$

$$\hat{\mathbf{Q}}(\mathbf{y}) \stackrel{\text{def}}{=} P_{\hat{\mathbb{C}}}\{\hat{c} : \mathbf{Y}(\mathbf{X}(\hat{c})) = \mathbf{y}\}$$

Стегосистему будем называть ***стойкой относительно функции извлечения признаков*** $\mathbf{X}$ ***и множества классификаторов*** $Y_1$, $Y_2$, ... $Y_a$, если:

$$\forall \mathbf{y} \in \{-1, +1\}^a \Longrightarrow \mathbf{Q}(\mathbf{y}) = \hat{\mathbf{Q}}(\mathbf{y})$$

Обозначение: $$\varphi = \sum_{\forall \mathbf{y} \in \{-1,+1\}^a} \left| \mathbf{Q}(\mathbf{y}) - \hat{\mathbf{Q}}(\mathbf{y}) \right|$$

**Лемма о не увеличении $\varphi$ при уменьшении количества классификаторов** . Для любого множества классификаторов $\{Y_1, Y_2, ... Y_a\}$ величина $\varphi$ не увеличиться для классификаторов $\{Y_1, Y_2, ..., Y_{j-1}, Y_{j+1}, ..., Y_a\}$ при любом $j$ от 1 до $a$.

# «Новый подход» к стеганографии и стегоанализу

- **Старый 1** («наследие криптографии»). Предлагаем алгоритм. Доказываем, что *известные методы* его не ломают. Ждем **n** лет. Никто не взломал → шифр надежен.
- **Старый 2.** («поиск аномалии»). Предлагаем алгоритм. Для всех *известных распределений* проводим исследование и убеждаемся, что нет аномалий.
- **Новый** («машинное обучение»). Предлагаем алгоритм. Создаем множество пустых контейнеров (не содержащие стеганографию) и стегоконтейнеров – это выборка. Экспертными оценками создаем функцию X. Пытаемся решить задачу классификации (функция Y).
- Старые *методы* и *распределения* можно определить как признаки для нового.

# Области исследований

- **Научная работа.**

  Практическая. Попытаться взломать различные алгоритмы стеганографии (LSB, patchwork, F5 и т. д.) методом машинного обучения.

  Теоретическая. Актуальность и сложность моделей. Энтропийные модели в стеганографии, их актуальность для ML-подхода. Задача Data Mining

- **Инженерная работа.**

  Создание  универсальной системы ML для обнаружения стеганографии в определенной среде. (хорошая дипломная работа)

# Резюме.

- Информационная стеганография – это актуальная, наукоемкая дисциплина.
- Задачи стегоанализа можно (и нужно) сводить к задачам ML
- Не рассматривалась последовательность контейнеров («память», состояние конечного автомата). Имеет смысл?
- Модель Грушо. Теорема Грушо.
- Модель классификатора без памяти
- ML подход стеганографии и стегоанализа

# Вопросы?

Слипенчук
Павел
PavelSlipenchuk@stego.su
аспирант каф. ИУ-8
«Информационная безопасность»
МГТУ. им.Баумана