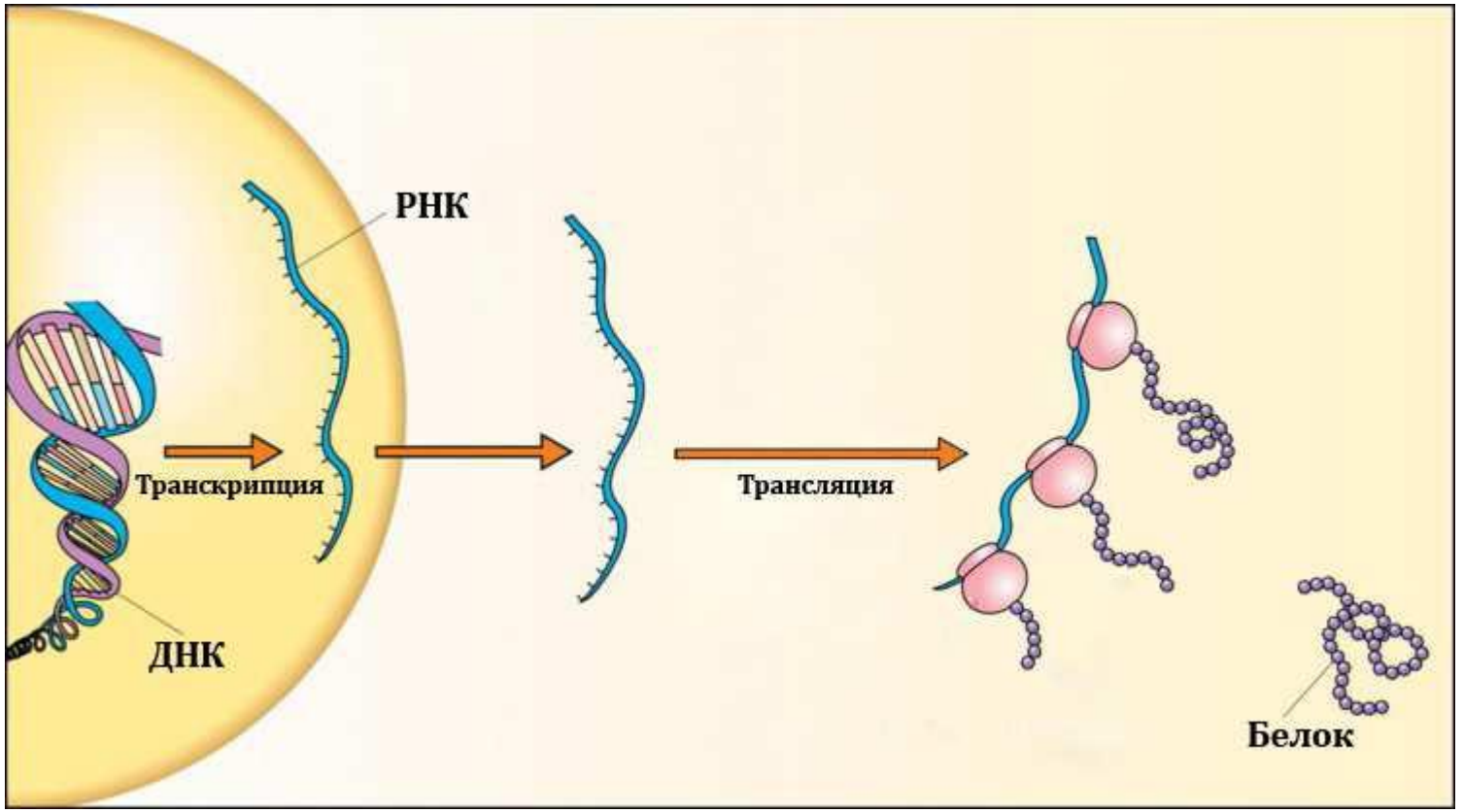
Статистический анализ генов, влияющих на развитие рака груди

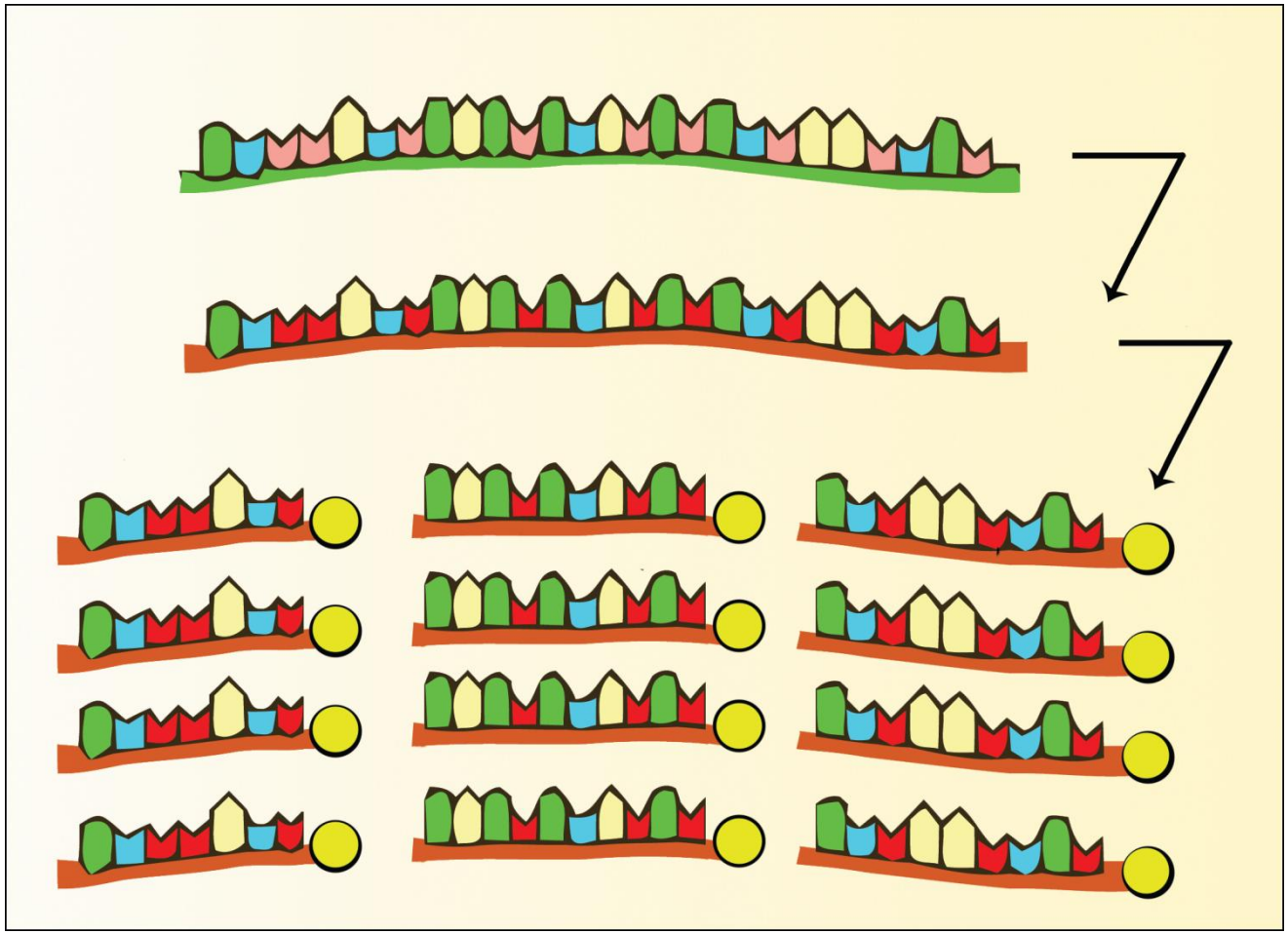
Константин Некрасов

Центральная догма молекулярной биологии

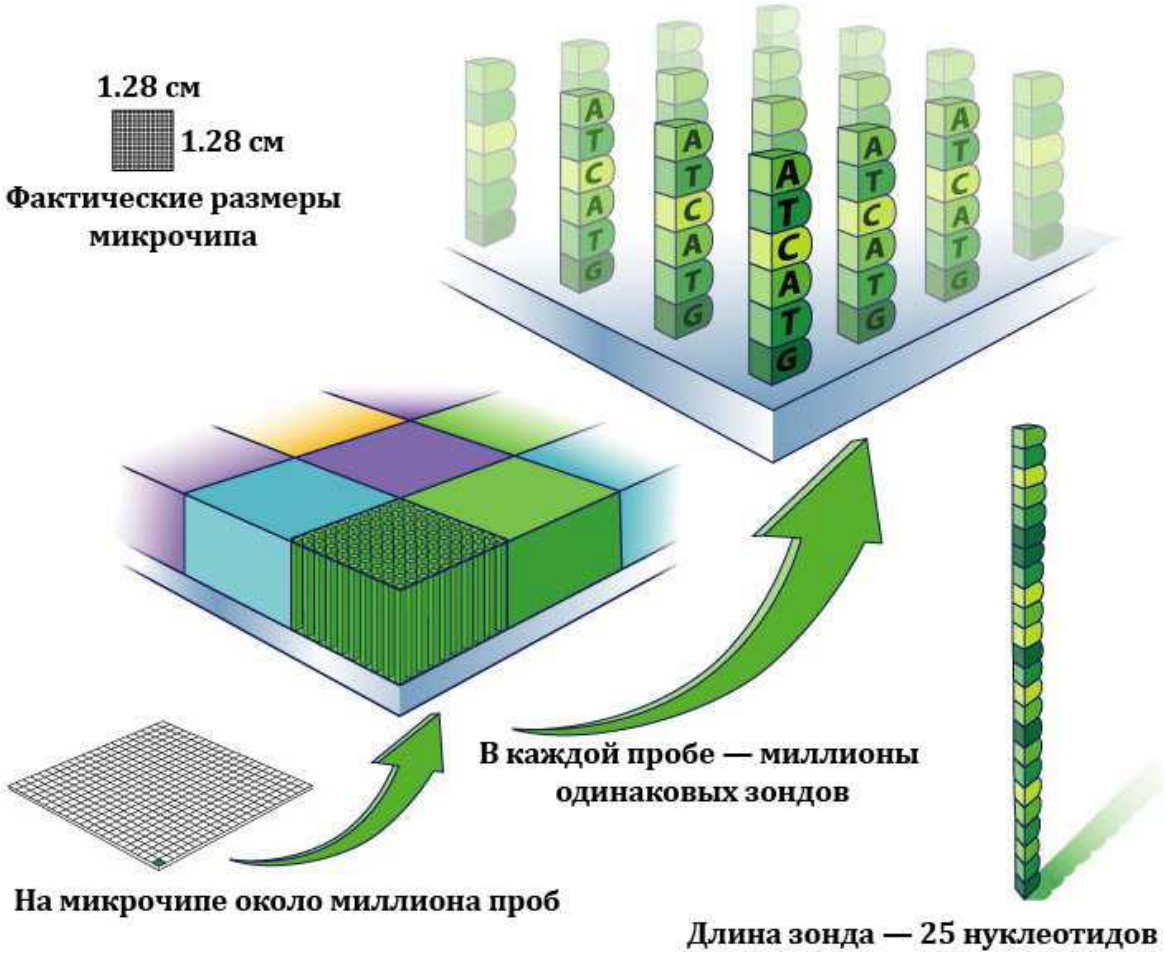


Картинки про ДНК взяты из презентации Евгения Рябенко

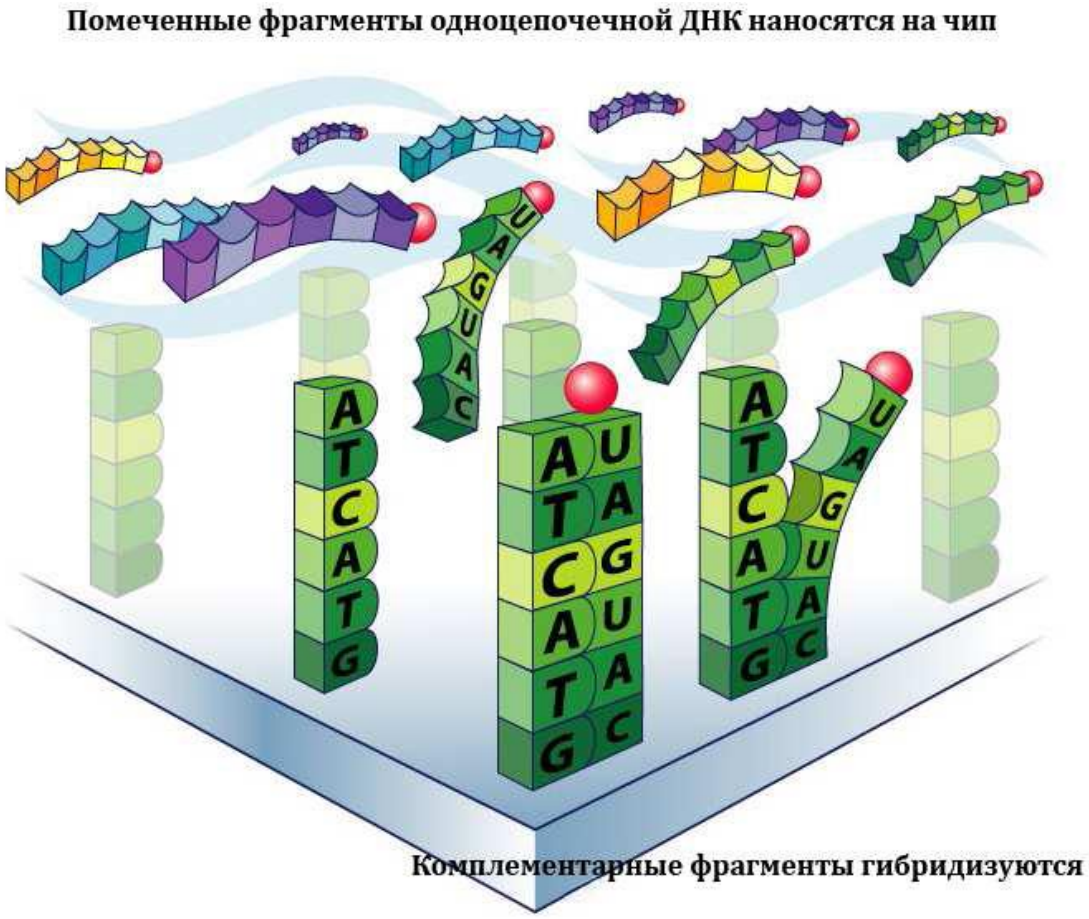
Нарезка ДНК



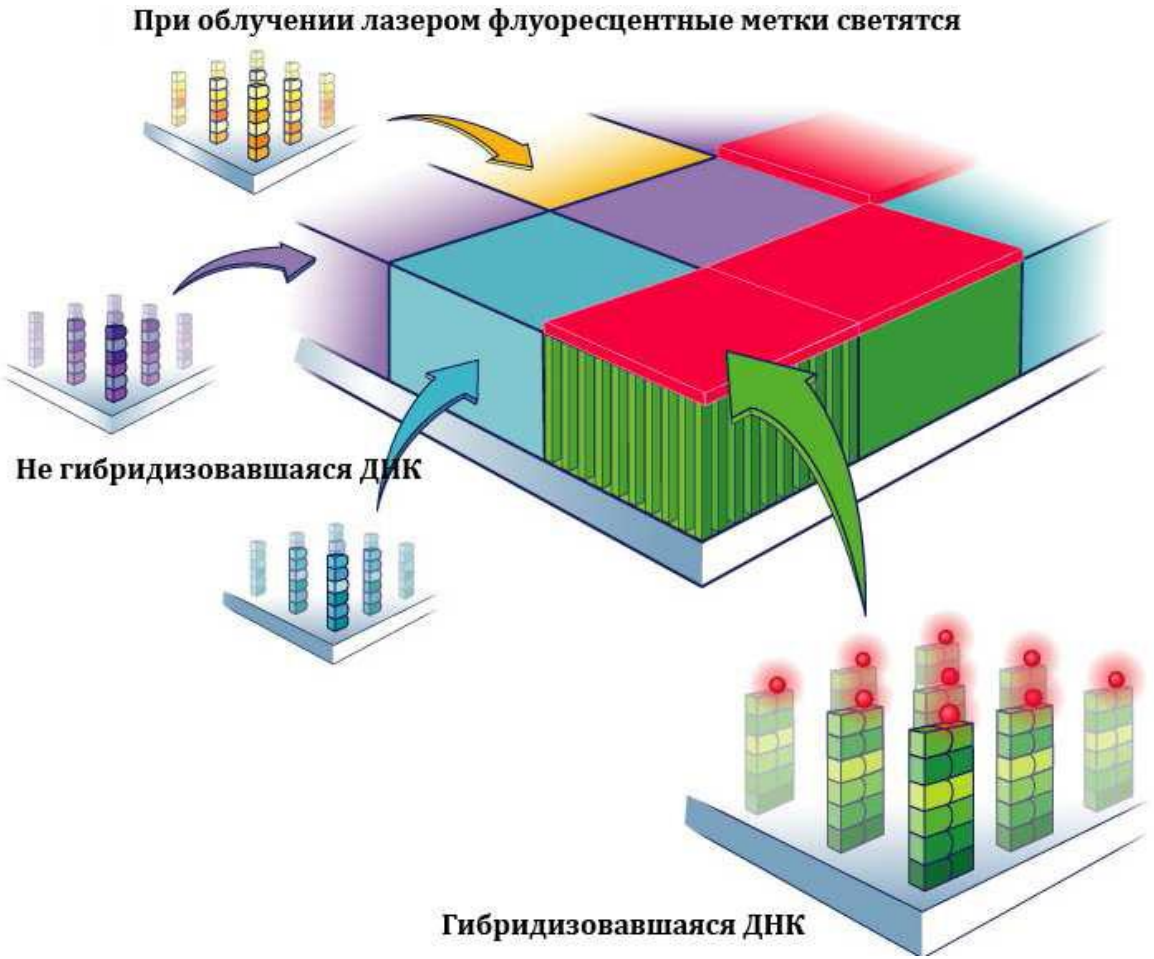
ДНК-микрочип



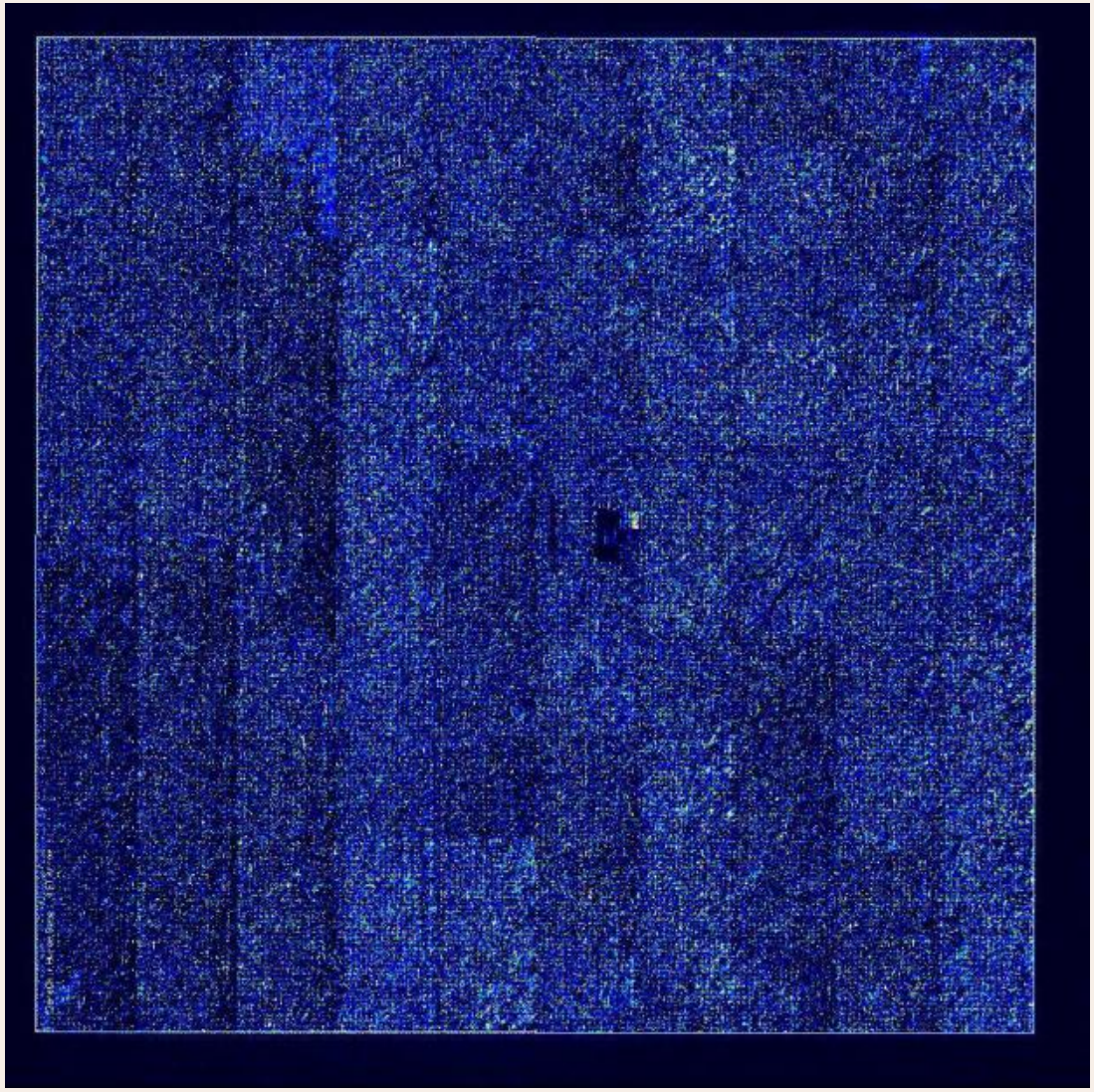
ДНК-микрочип



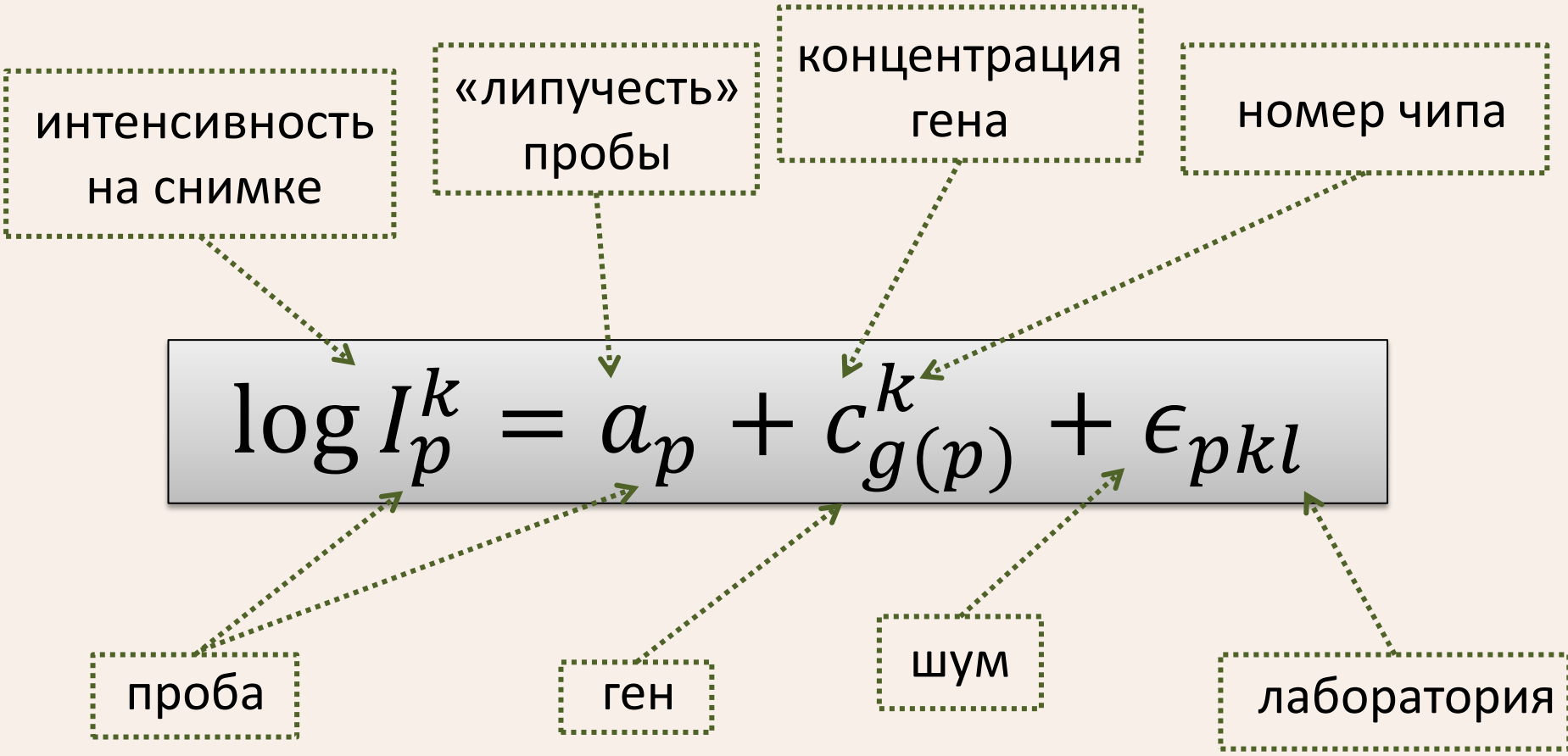
ДНК-микрочип



ДНК-микрочип



Экспрессия гена



Восстановление концентрации

$$\sum_p \left(\log I_p^k - a_p - c_{g(p)}^k\right)^2 \to \min_{\{c_g^k\}}$$

- «Липучести» $\boldsymbol{a_p}$ и интенсивности $\boldsymbol{I_{p,g}^k}$ известны
- Концентрации генов $\boldsymbol{c_g^k \geq 0}$ и $\boldsymbol{\sum_g c_g^k = 1}$
- $\boldsymbol{p}$ – номер пробы (много для одного гена!)
- $\boldsymbol{g}$ – номер гена
- $\boldsymbol{k}$ – номер чипа

Где используется экспрессия

Болезни

Рак

Способности

Наклонности

Характер

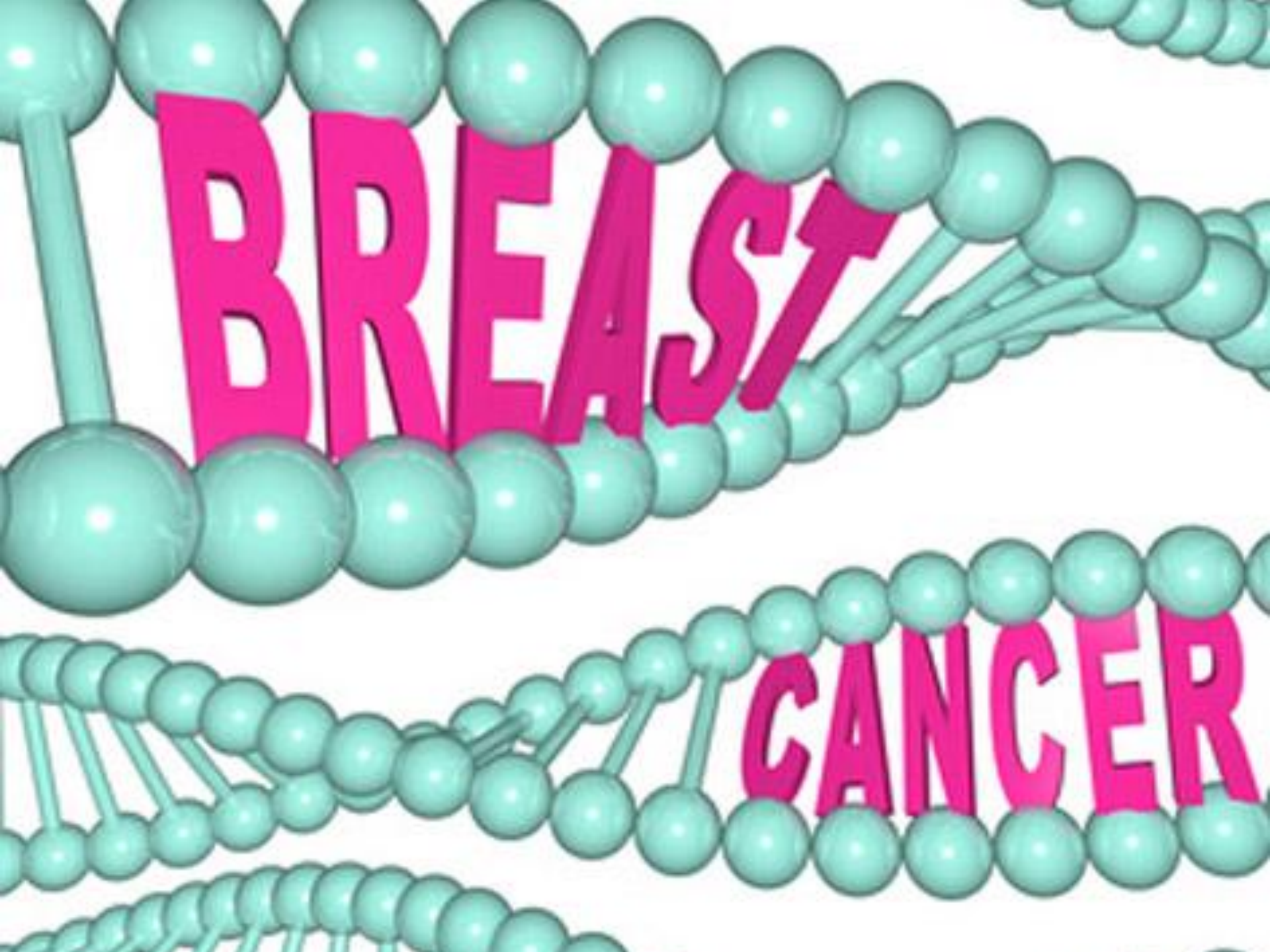
Предрасполо-
женности

Химиотерапия

Gattaca!

Вид лечения

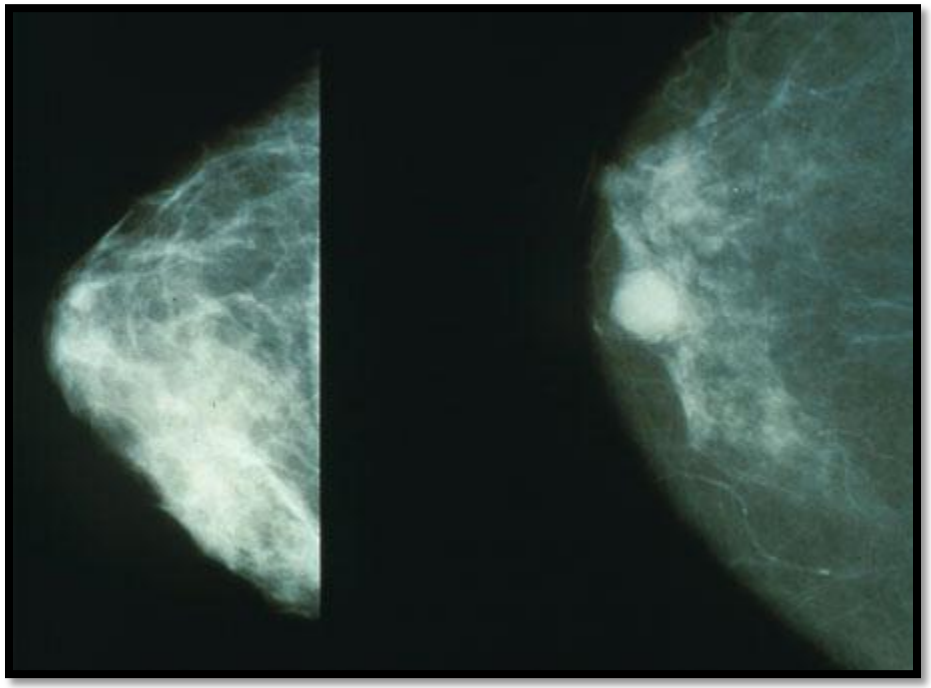
Продолжительность
жизни


BREAST
CANCER

Диагностика

С чем мы имеем дело

Здоровая грудь



Раковая опухоль

Лечение

Обычно в молочных железах

Вырезают опухоль

В России в год выявляется 55 000

У женщин в 100 раз чаще

Рецидив

Выздоровление

Из них умирает 22 000

В 2008 году в мире погибло 458 503

Зато у мужчин смертность выше!

Химиотерапия

- Для повышения вероятности выздоровления, после операции применяют химиотерапию.
- Это вредно для здоровья: портятся почки, печень, сердце, иммунитет, выпадают волосы и т. д.
- Мы не хотим прописывать химию тем, кто и так выздоровеет

Задача классификации

Что случится с больным после операции?

Выздоровление

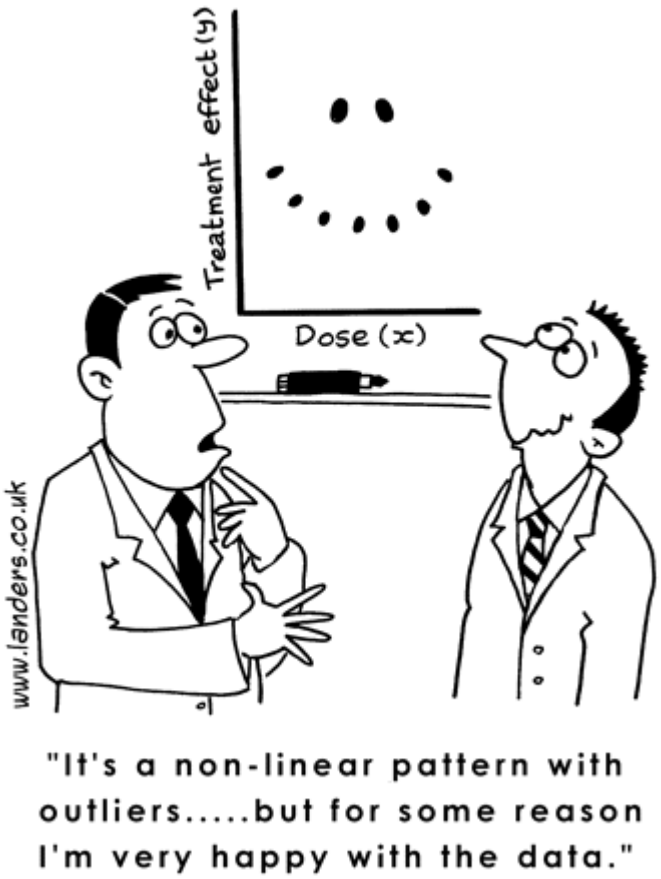
VS.

Рецидив

Мотивация

- Хотим по 20 генам научиться распознавать рецидив
- 20 генов позволят проводить дешевые тесты
- Сможем судить о том, нужна ли химиотерапия
- Нет конкуренции

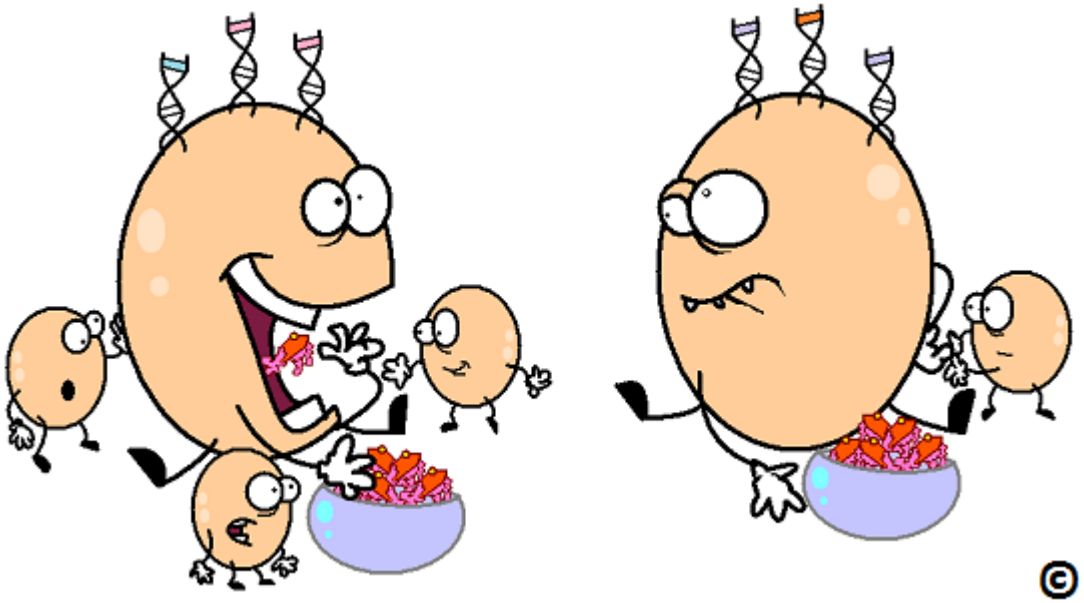
Данные



- Три разные лаборатории предоставляют уже вычисленные по пациентам экспрессии генов
- В них 79, 58 и 201 человек по 22 тыс. генов
- Всем пациентам была вырезана опухоль
- Есть два типа пациентов: в течении 5 лет случился рецидив, в течении 7 лет рецидива не было.
- Других пациентов не рассматриваем.

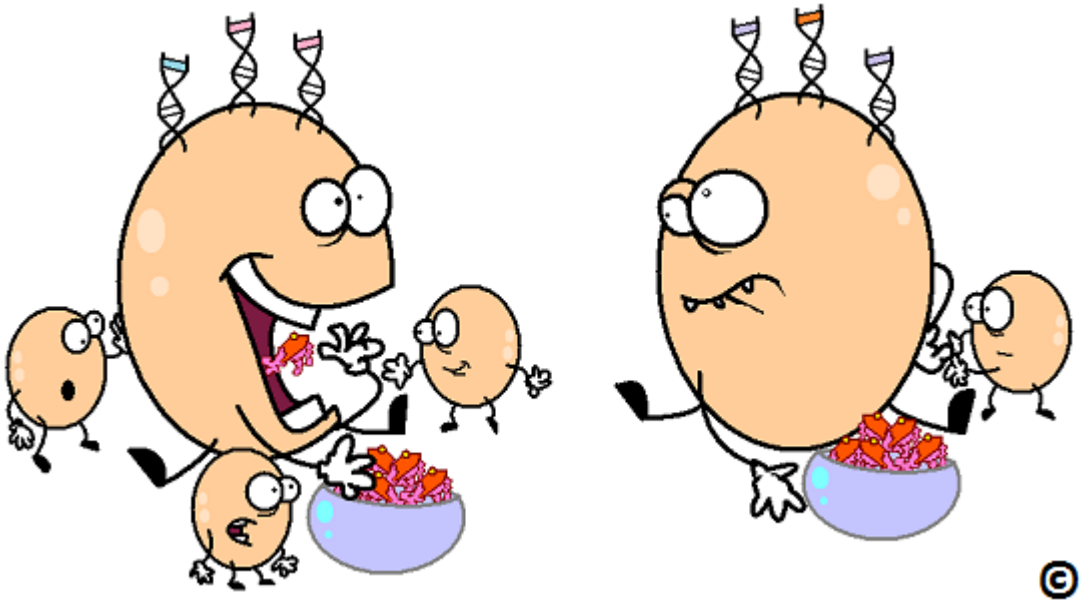
Эвристики от генетиков

- Концентрация гена не слишком маленькая/большая
- Имеет достаточно большой разброс
- Итоговый классификатор будем применяться на совсем других чипах и на данных от другой лаборатории
- Нужно узнать, насколько результат испортится, если тестировать на выборке пациентов из совсем другой лаборатории



Эвристики от генетиков

- Используем лог. шкалу для концентраций
- Медиана по каждому из классов лежит на $[5, 12]$
- Разница между 95%- и 5%-квантилем больше 2
- Применив эти эвристики остается 470 генов



Случайный лес

Строим решающее дерево

- Выбираем очередной признак и помещаем его в корень
- Разделяем выборку по этому признаку
- Для каждой части рекурсивно строим дерево
- Пока не иссякнут признаки или останутся объекты одного класса

Information Gain

$$I(p) = -p \log_2 p$$

$$H(X) = \sum_{k=1}^{K} I\left(\frac{\text{число прецедентов класса } k}{\text{всего прецедентов}}\right)$$

Обучающая выборка $X = (x_i, y_i)_{i=1}^{N}$

$$IG(X, f) = H(X) - H(X|f)$$

Признак объектов

$$H(X|f) = \sum_{v \in vals(f)} \frac{\text{число прецедентов } x_f = v}{\text{всего прецедентов}} \cdot H\left(\text{выборка прецедентов с } x_f = v\right)$$

I'd like to be a tree!

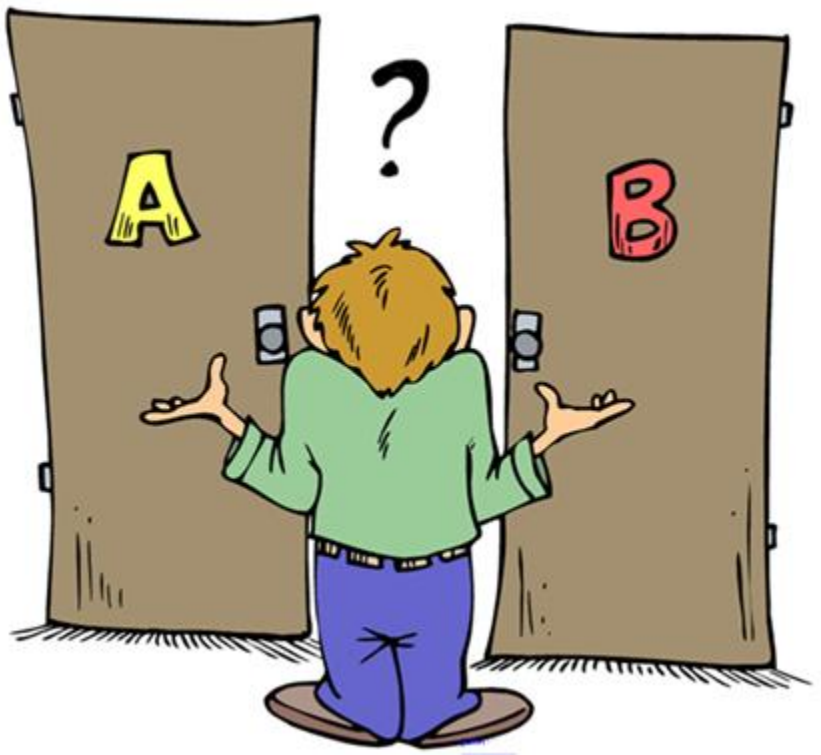
Random Forest

- Число прецедентов N
- Число признаков M
- Для каждого дерева случайно с повторением выбираем N объектов
- И случайную подвыборку из $\sqrt{M}$ признаков

Out-of-bag (OOB)

- В каждое дерево не попадает примерно N/3 прецедентов
- Можно вычислять ошибку случайного леса, посылая в каждое дерево только объекты, которые не участвовали при его обучении
- Такая ошибка будет несмещенной оценкой ошибки на генеральной совокупности

Feature importance



- Пусть X – обучающая выборка
- Пусть X^i – та же выборка, но значенияi-го признака случайно перемешаны по всем прецедентам
- Важность i-го признака определяется как разница между ошибками леса на OOB между X и X^i

Достоинства/недостатки

Оценка важности признаков

Сравнимо с SVM и бустингом

Высокое качество

Одинаково хорошо обрабатываются дискретные и непрерывные данные

Лучше нейросетей

Внутренняя оценка обобщающей способности (out-of-bag)

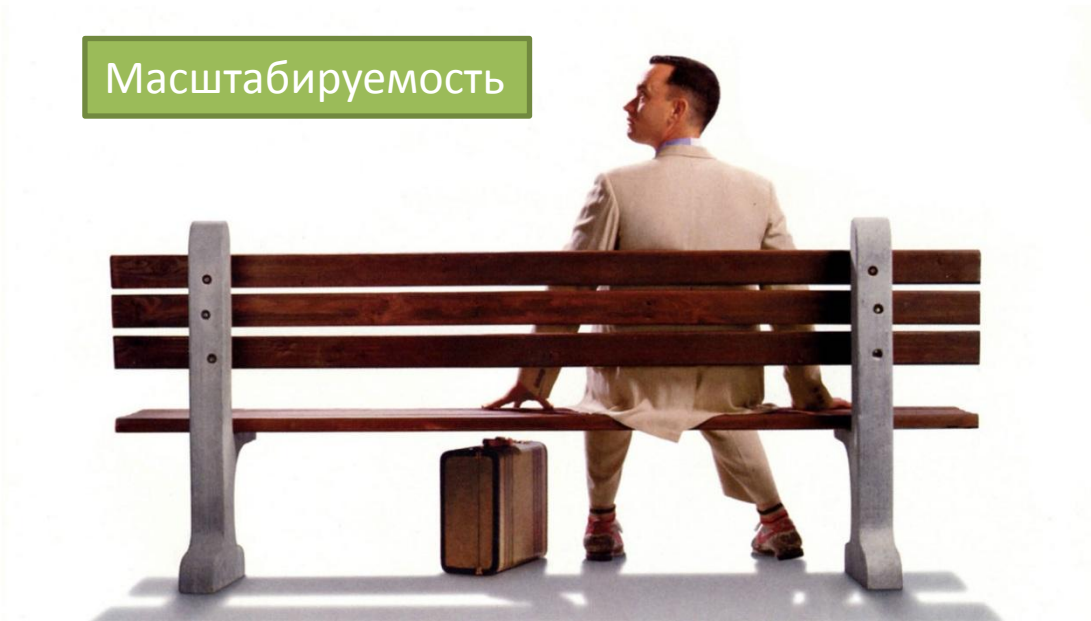
Эффективная обработка большого числа признаков и классов

Параллелизм

Склонен к переобучению на зашумленных данных

Масштабируемость

Большой размер получающихся моделей




EXPERIMENT

Число деревьев на обучении

Motivation

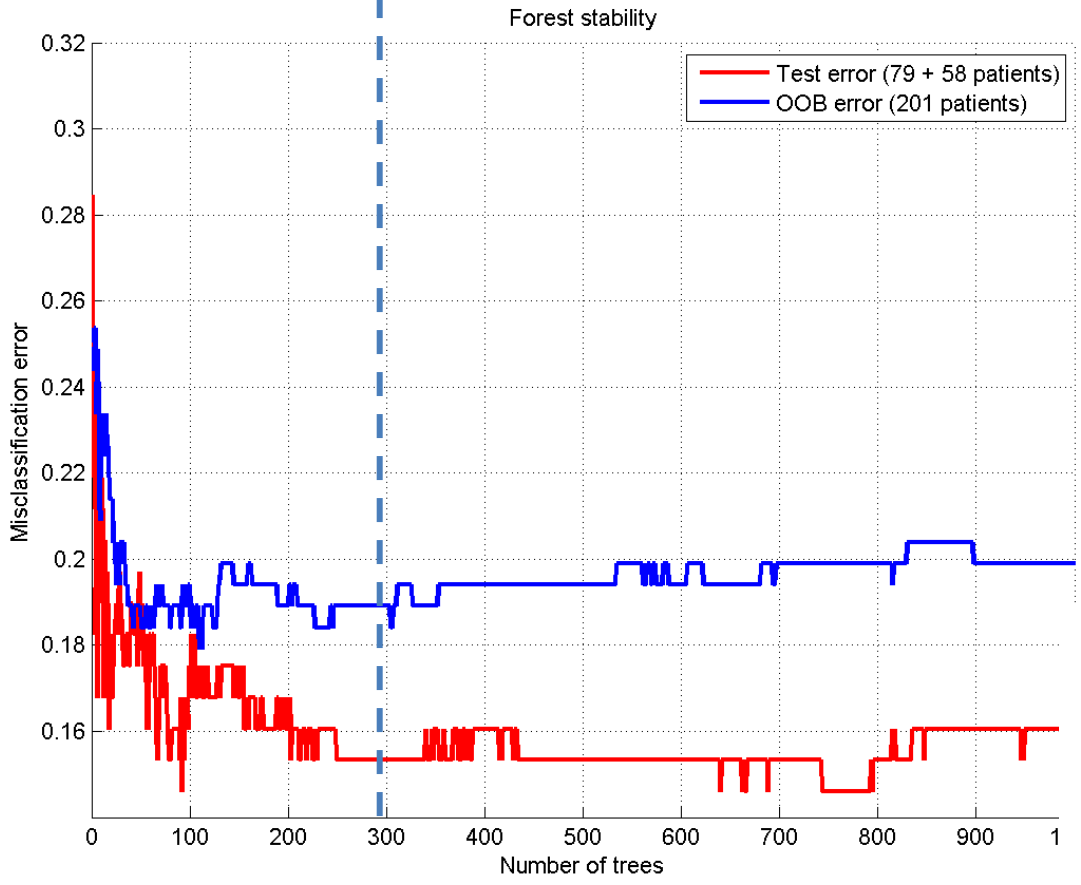
- Обучение леса отчасти случайно
- Обучая небольшое число деревьев мы можем получать нестабильный результат
- Сколько нужно деревьев для получения стабильности?

Experiment

- Будем наращивать число деревьев на обучении
- Сравнивать ошибку на обучении и тесте
- Нужно найти число деревьев, на которых стабилизируется ошибка
- Обучение на 3й (201)
- Тест на 1 и 2й (79+58)

Число деревьев на обучении

Стабилизация



Все 470 признаков

Обучение по 3й

Тест на 1 и 2й

Будем использовать 300 деревьев

Смещенные данные

Motivation

- Нам предоставляются данные из трех различных лабораторий
- В каждой съемка чипов проводилась в разных условиях
- Что если данные не однородны и задача не корректна?
- Насколько разную природу имеют данные?

Experiment

- Будем обучаться на одной лаборатории и тестировать на других
- Так для каждой лаборатории
- Построим ROC для каждого случая
- По ROC можно судить об однородности данных

Смещены ли данные?

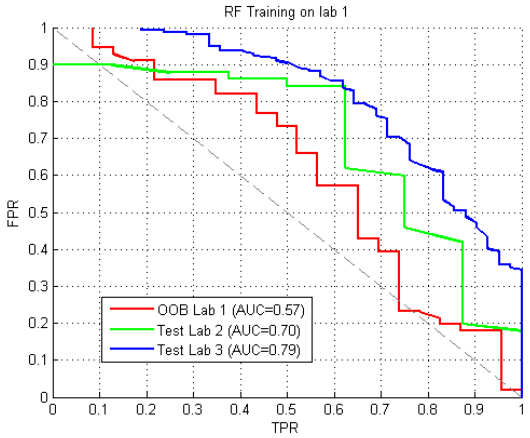
79 человек

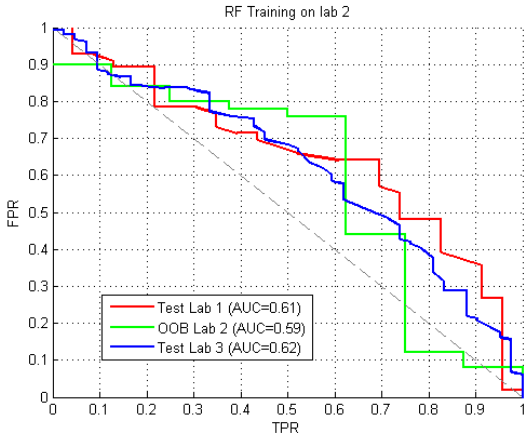
58 человек

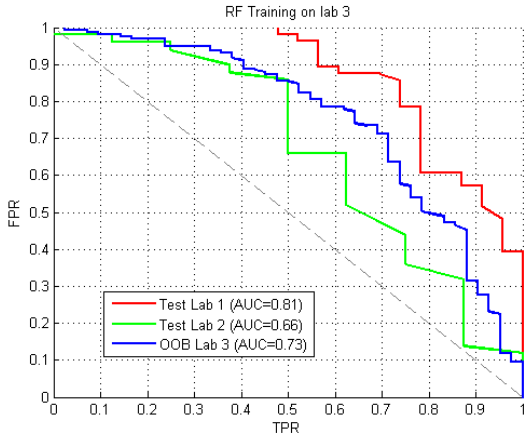
201 человек

Самая маленькая — Лучше не обучаться

Недообучается — Можно тестировать







Смотрите, данные одной природы!

Устойчивость отбора признаков

Motivation

- Случайный лес позволяет оценить информативность признака
- Лес обучается случайно и разные ансамбли будут возвращать разную информативность
- Насколько устойчив показатель информативности, который возвращает случайный лес из 300 деревьев?

Experiment

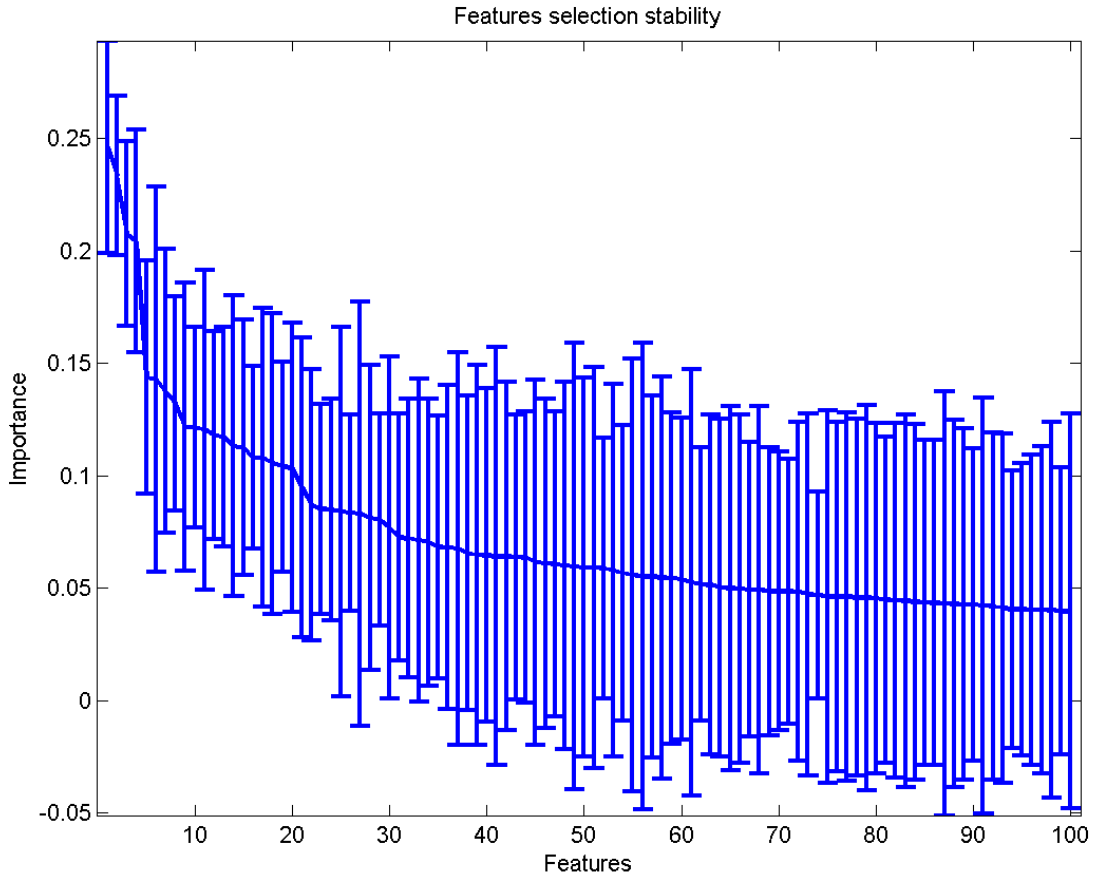
- Обучим 50 лесов по 300 деревьев
- Вычислим по каждому признаку медиану и интерквантильный размах информативности
- Хорошо, если вверху окажется несколько генов с высокой медианой и низким размахом

Устойчивость отбора признаков

× 50 лесов

× 300 деревьев

Значимость



Обучение на 3й лаборатории

Медиана

Размах

Признаки отсортированы по медиане значимости

Качество отобранных признаков

Motivation

- Наконец, мы готовы посмотреть на качество отбираемых признаков
- Будем смотреть на зависимость качества классификатора от числа добавляемых признаков
- Признаки добавляем в порядке убывания значимости

Experiment

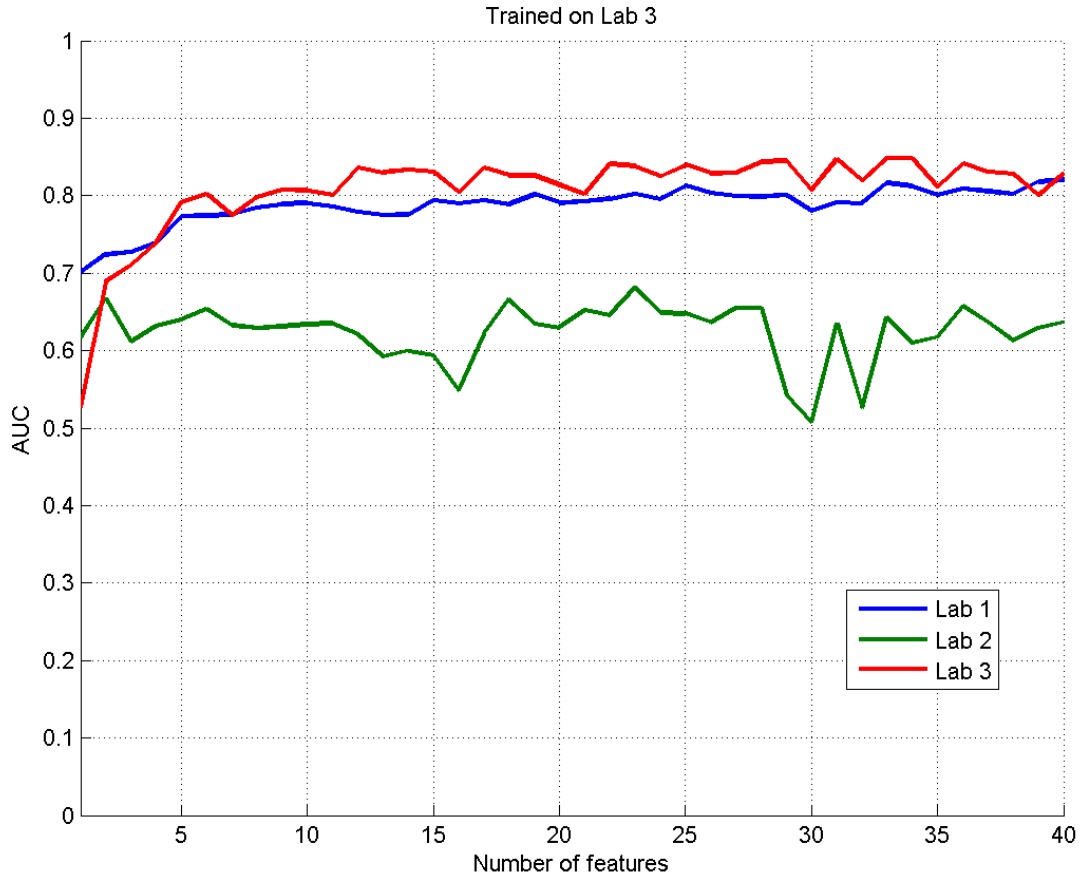
- Значимость определяем по величине медианы
- Качество будем измерять по AUC
- Обучаем отдельные леса для каждого набора признаков
- Обучение на 3й
- Тест на 1 и 2й отдельно

Качество отобранных признаков

Тест

Тест

Обучение



79 человек

58 человек

201 человек

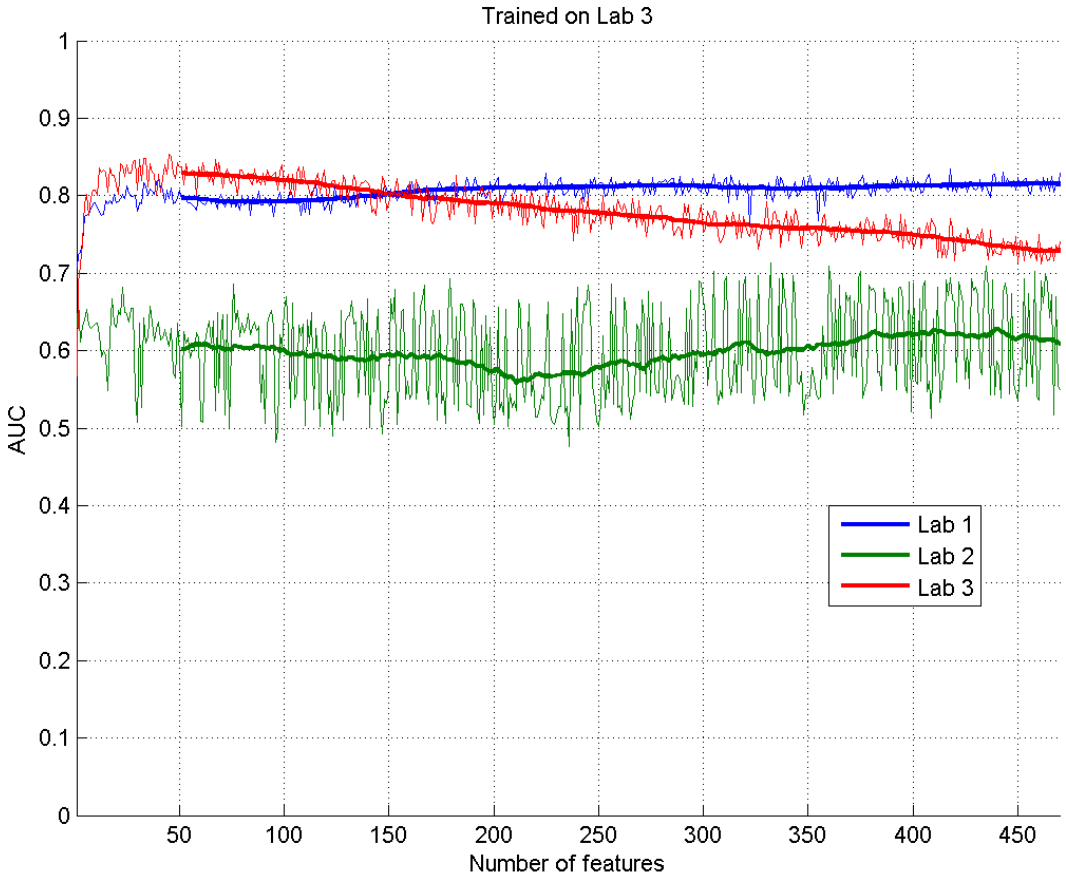
Добавляем по одному признаку и строим лес заново

Качество отобранных признаков

Тест

Тест

Обучение



79 человек

58 человек

201 человек

Добавляем по одному признаку и строим лес заново

Итоговый классификатор

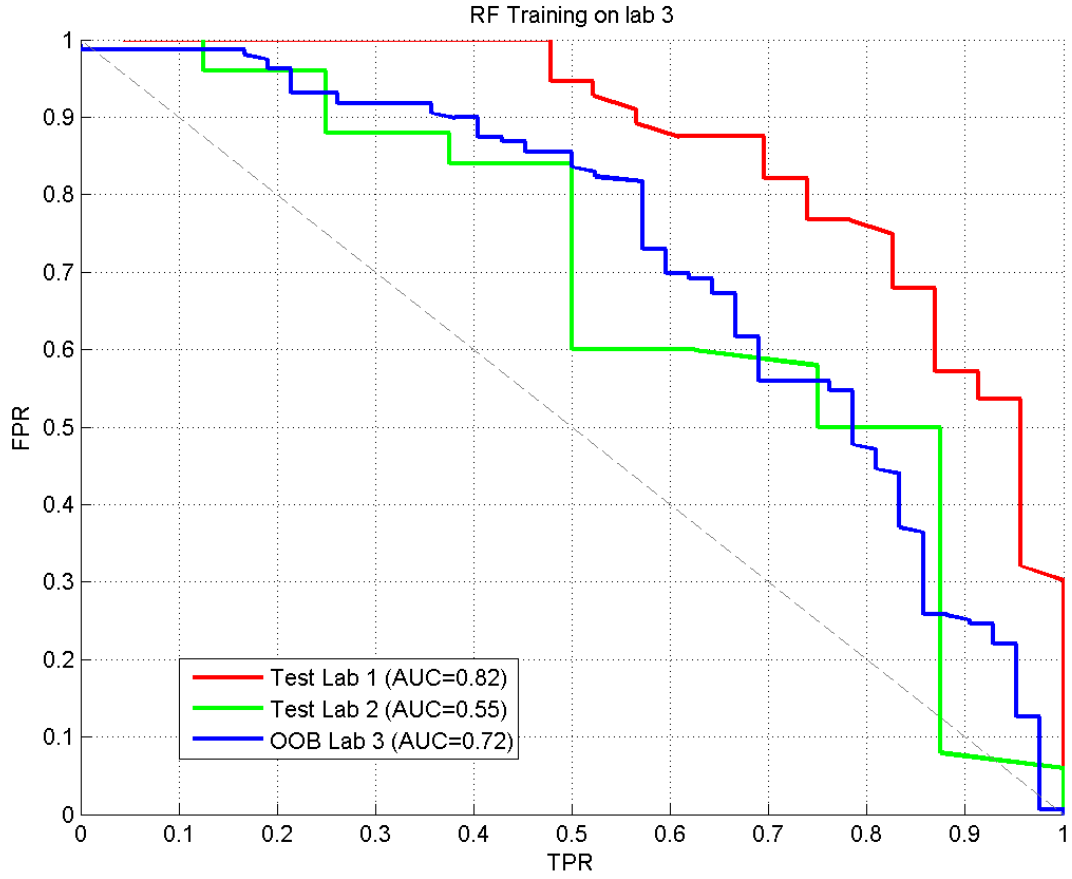
Motivation

- Давайте, наконец, посмотрим на ROC итогового классификатора
- Которым будем прогнозировать выздоровление/рецидив!
- Заодно посмотрим, насколько качество скачет на разных лабораториях

Experiment

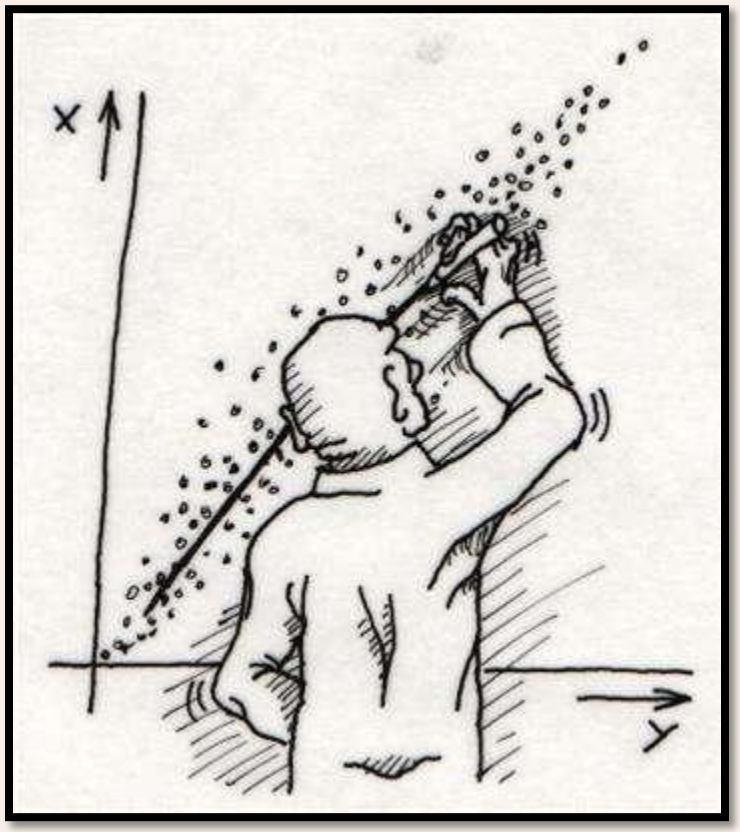
- Берем первые 20 признаков
- Обучение на 3й
- Тест на 1 и 2й отдельно

Качество отобранных признаков



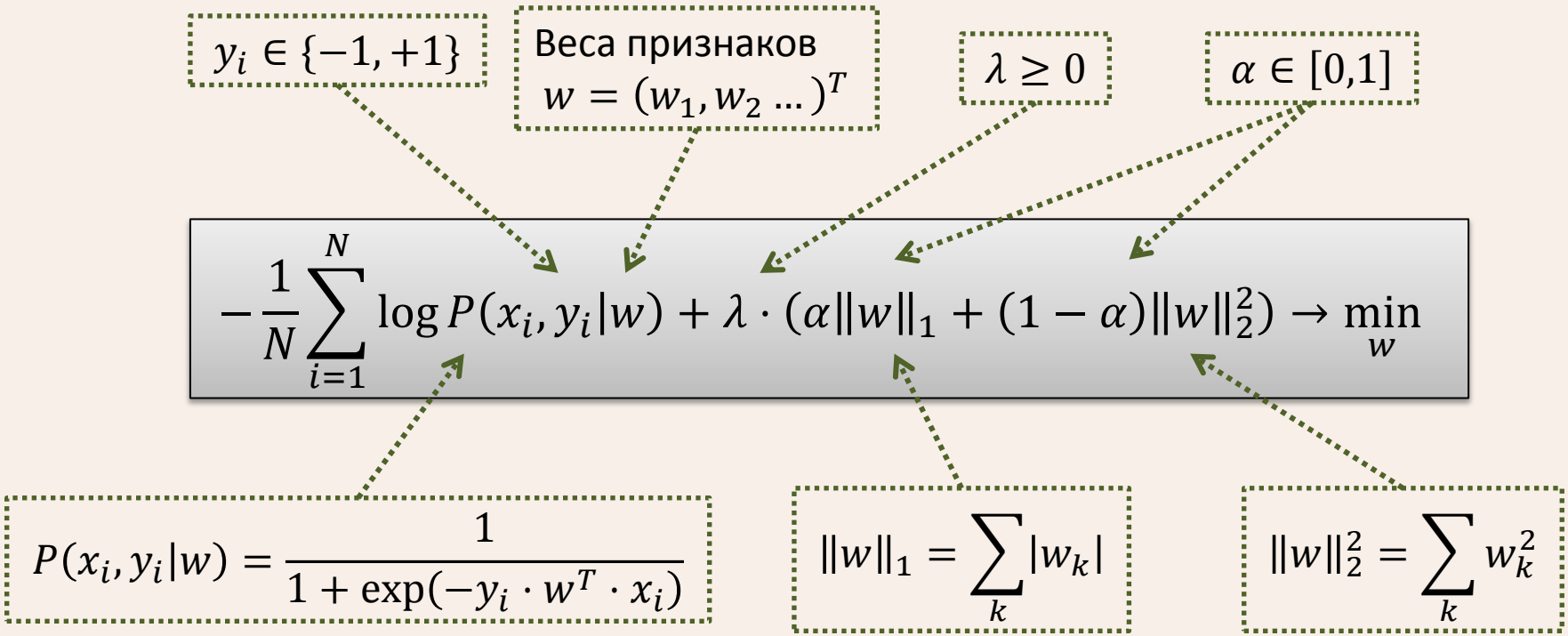
Логистическая регрессия

Логистическая регрессия с L_1 и L_2 регуляризацией



- Настраивает вероятностную модель для бинарной классификации
- Регуляризация против переобучения
- L_1 (Lasso) отбирает небольшое число информативных признаков
- L_2 (Ridge) учитывает коррелирующие признаки
- Вписывается с помощью модификации LARS
- Применяется для обработки DNA-microarray

Максимизация правдоподобия



EXPERIMENT

Подбор параметров

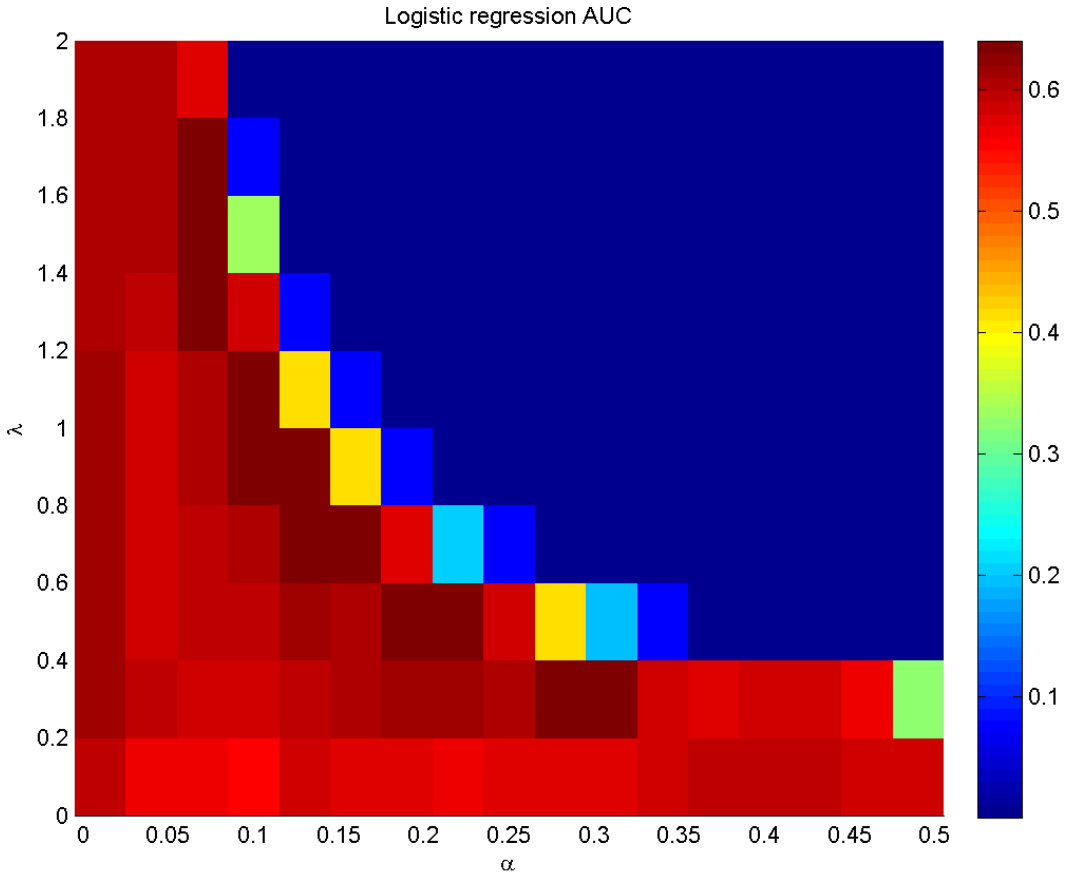
Motivation

- Нужно настроить «эластичную сеть»
- Параметр λ указывает общую силу регуляризации
- Параметр α определяет компромисс между L_1- и L_2-регуляризацией
- Также нужно учесть, что в конечной модели мы хотим получить не более 20 весов

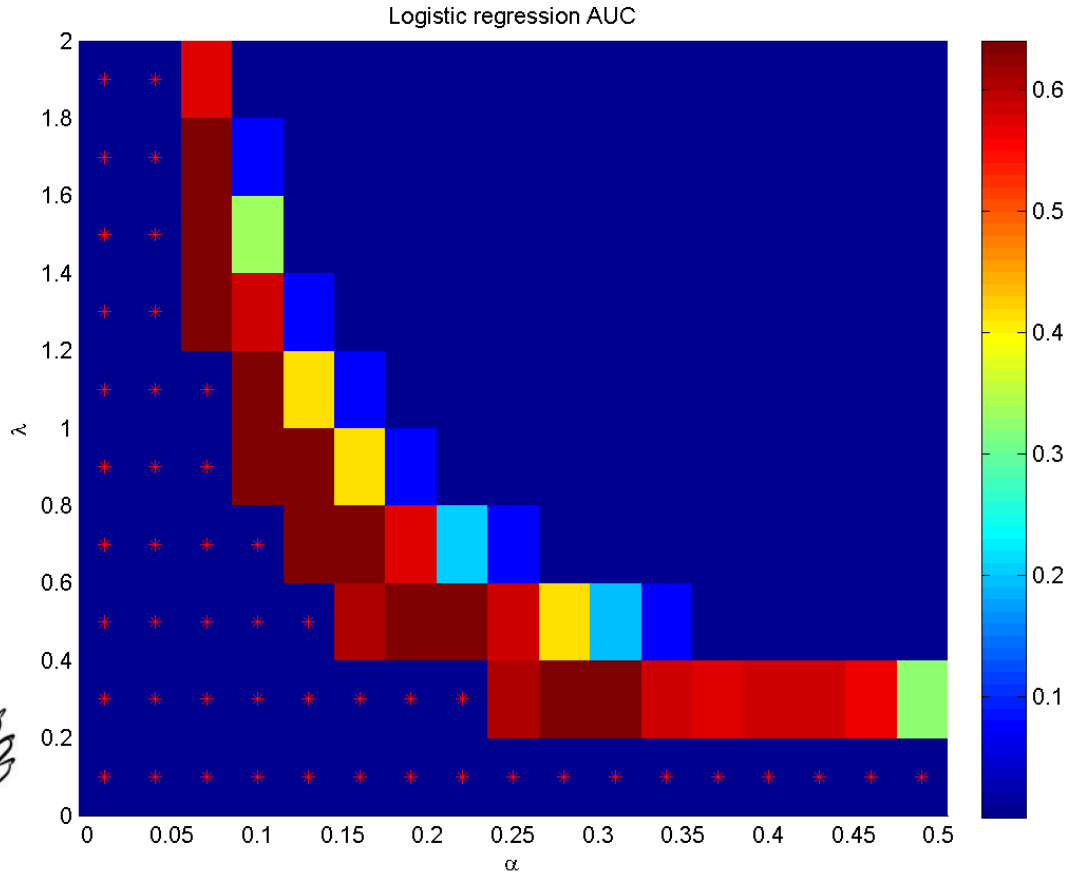
Experiment

- Пройдемся фиксированной сеткой альф и лямбд
- Вычислим средний AUC по пяти фолдам на 3й лаборатории
- Посчитаем среднее количество отбираемых весов на фолдах
- Выберем точку с максимальным AUC, со средним числом весов не больше 20

AUC на кроссвалидации



Обнуляем точки с числом признаков более 20

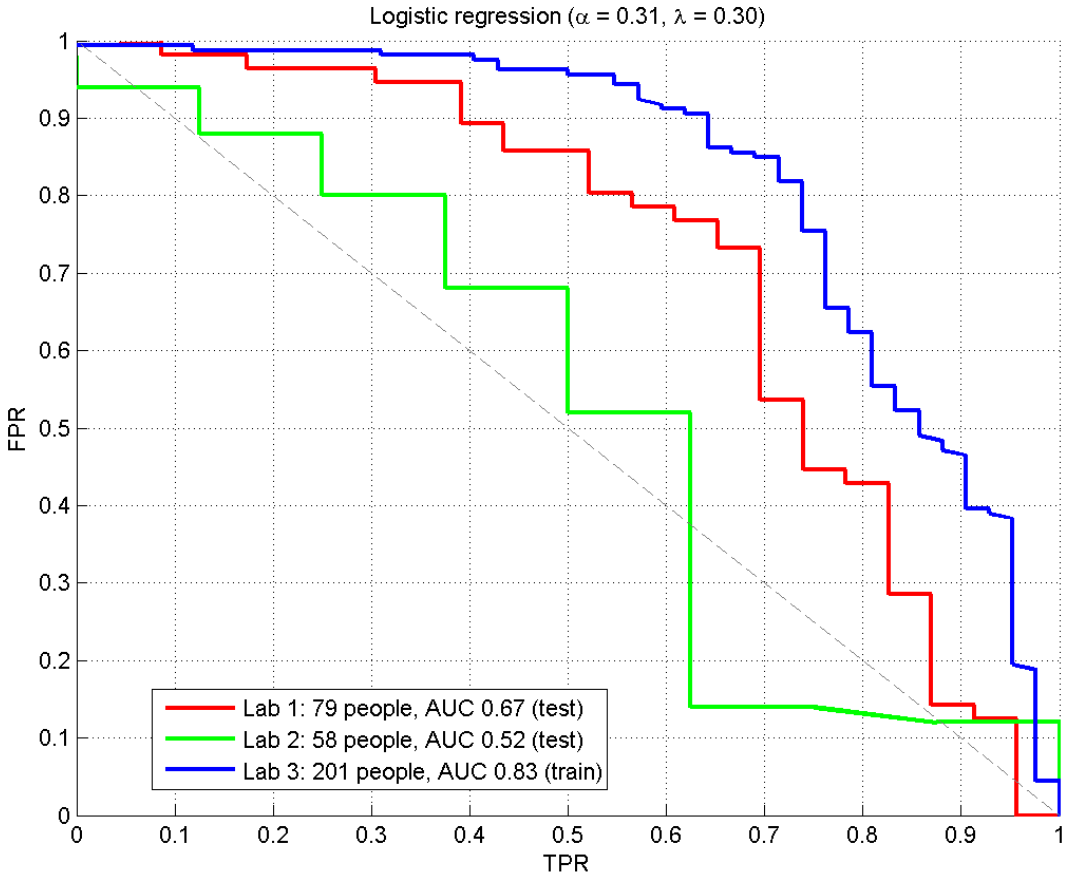


Красные точки – число признаков больше 20

Логистическая регрессия

Отобраны гены:

1. **RAD21**
2. **CCT2**
3. **LYPLA1**
4. **S100P**
5. **CX3CR1**
6. **SQLE**

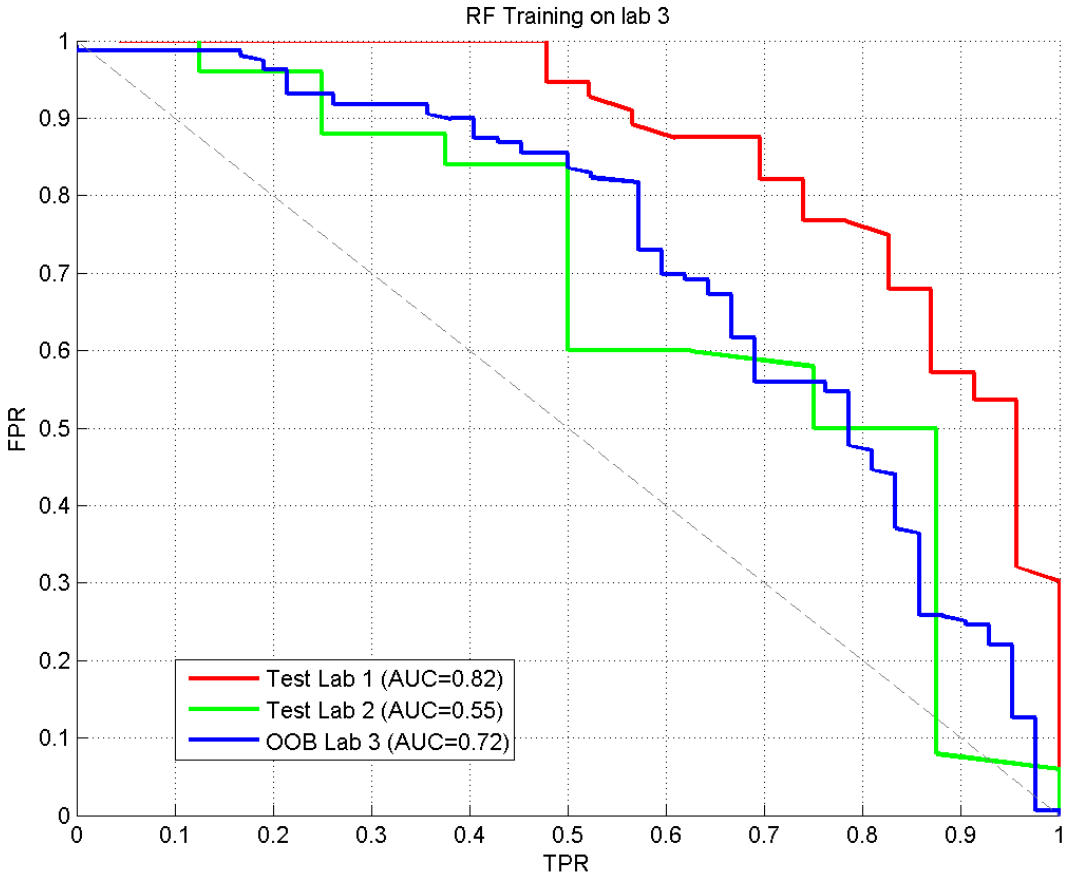


Переобучение!

Случайный лес

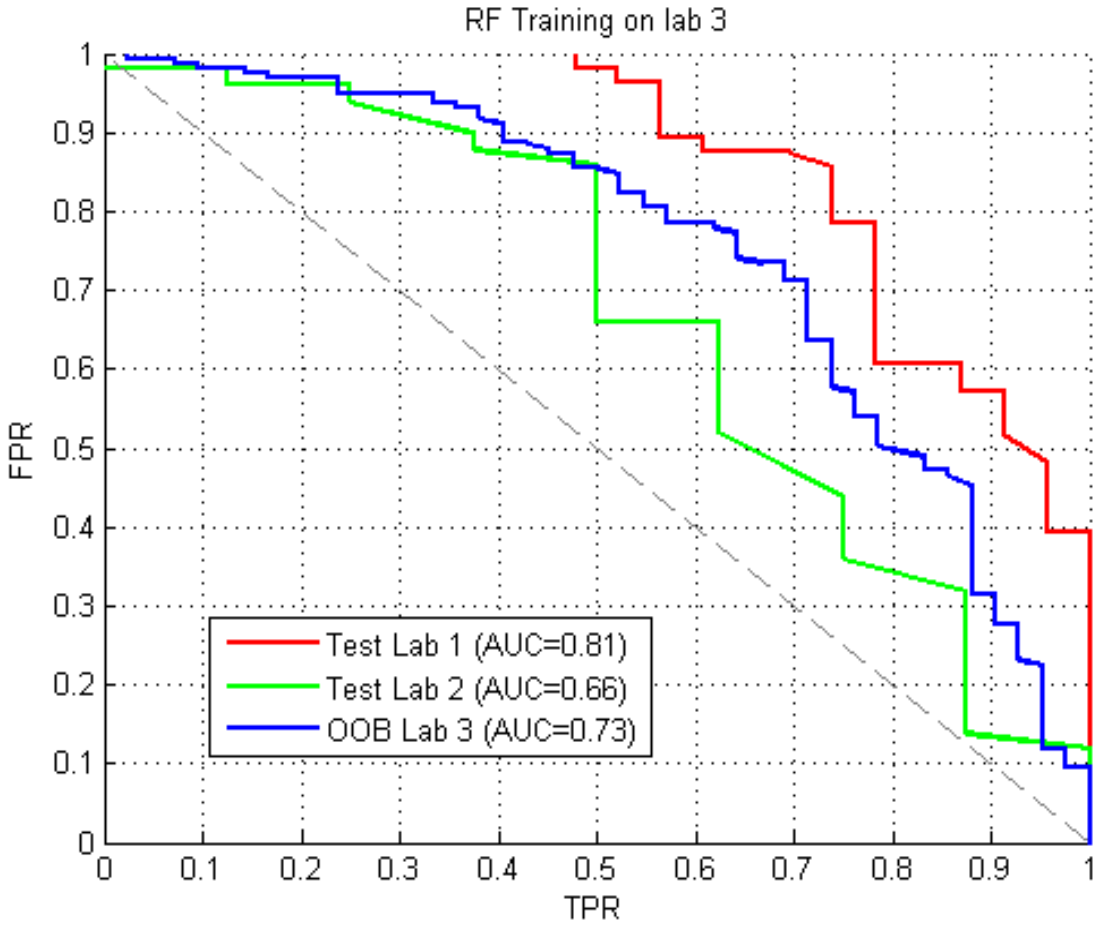
Отобраны гены:

1. **S100P**
2. **SQLE**
3. MTDH
4. PGK1
5. C3orf14
6. **CX3CR1**
7. **RAD21**
8. MLF1IP
9. LTF
10. **CCT2**
11. FCGBP
12. CXCL12
13. TMEM70
14. HMGB2
15. TPD52
16. **LYPLA1**
17. TFRC
18. MUC1
19. PERP
20. MRPS12



Судя по AUC – RF лучше, чем LR

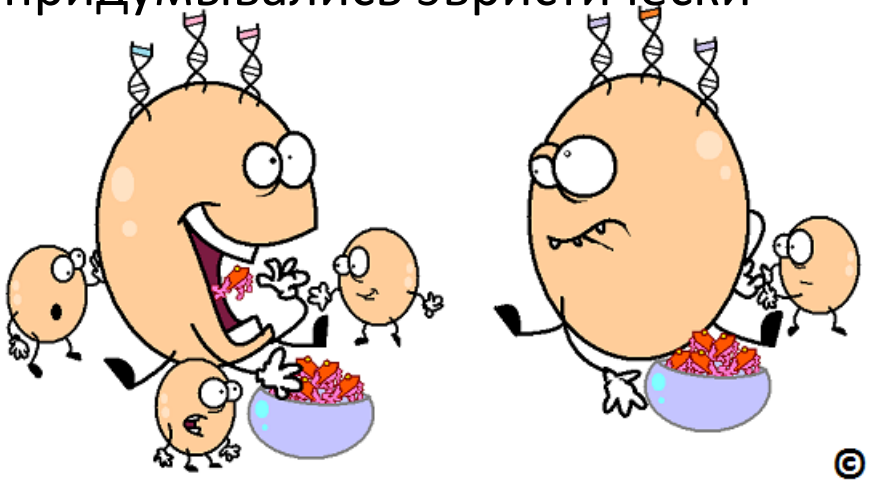
RF на всех признаках



Смотрите – одно и то же!

Критерий качества для генетиков

- Требуется отобрать 20 наиболее информативных генов и научиться предсказывать рецидив
- Точность по выздоровевшим пациентам (нулевой класс) $\mathrm{P}_0 \geq 90\%$
- Точность по пациентам с рецидивом (первый класс) $\mathrm{P}_1 \to \max$
- Сейчас в мире $P_1 \approx 30..50\%, P_0 \approx 90\%$
- Низкое качество, потому что придумывались эвристически



Финальная точность

	LR		RF		World	
	P_0	P_1	P_0	P_1	P_0	P_1
Train/OOB (lab 3)	0.90	0.65	0.90	0.36	0.8-0.9	0.3-0.5
Test (lab 1)	0.81	0.60	0.90	0.62		
Test (lab 2)	0.87	0.17	0.93	0.20		

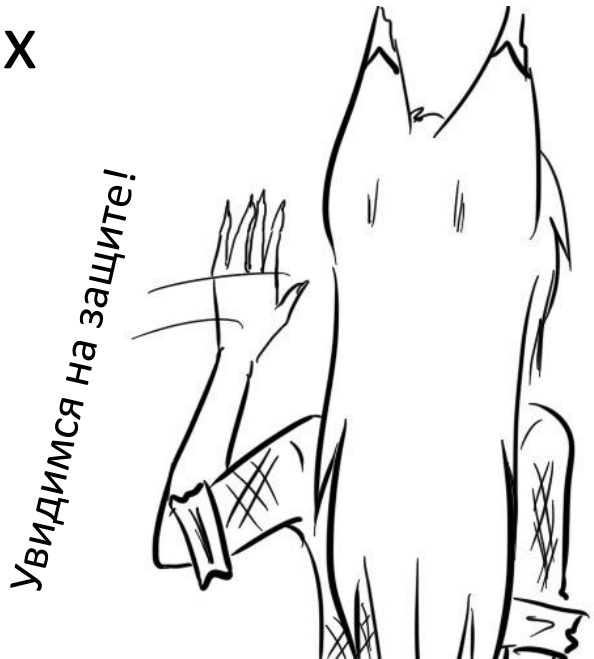
Заказчик оперировал точностью!

Пора её вычислить

По результатам в таблице
и графикам AUC – RF побеждает

Итог

1. Мы устойчиво отобрали несколько генов
2. Обучили случайный лес и логистическую регрессию с «эластичной сетью»
3. Полученное качество варьирует от 20% до 60% на разных лабораториях



Спасибо за внимание!

КОНЕЦ