

Блок

# Feature Engineering

Занятие № 5

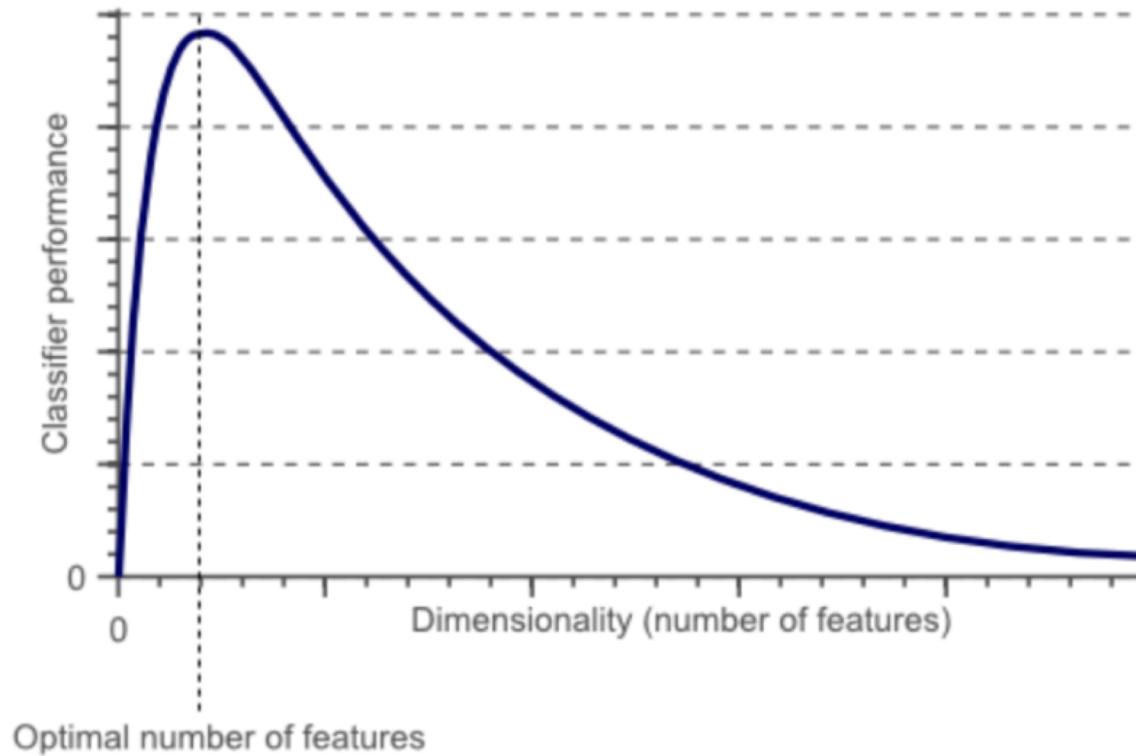
# Feature Selection

- Вспомним методы статистики
- Узнаем про методы декомпозиции данных
- Понять принцип РСА на практике

## Цели занятия

**Зачем всё это?**

# ПРОКЛЯТЬЕ РАЗМЕРНОСТИ

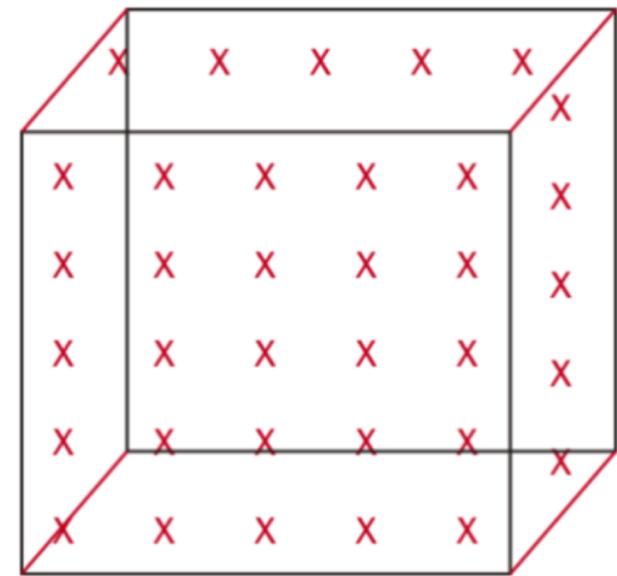


# ПРОКЛЯТЬЕ РАЗМЕРНОСТИ

          x      x      x      x      x



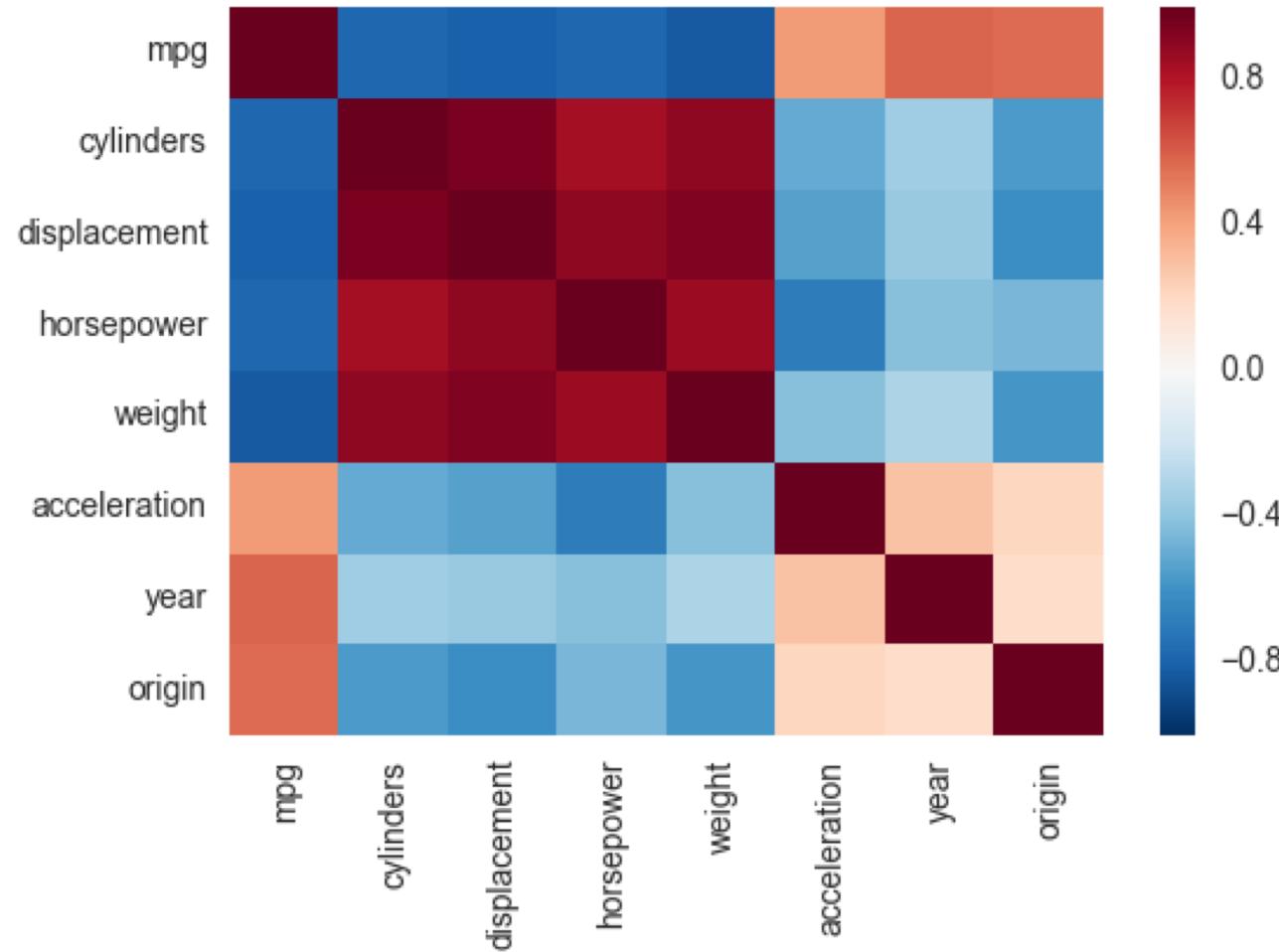
Одно измерение - 5 точек



Три измерения - 125 точек

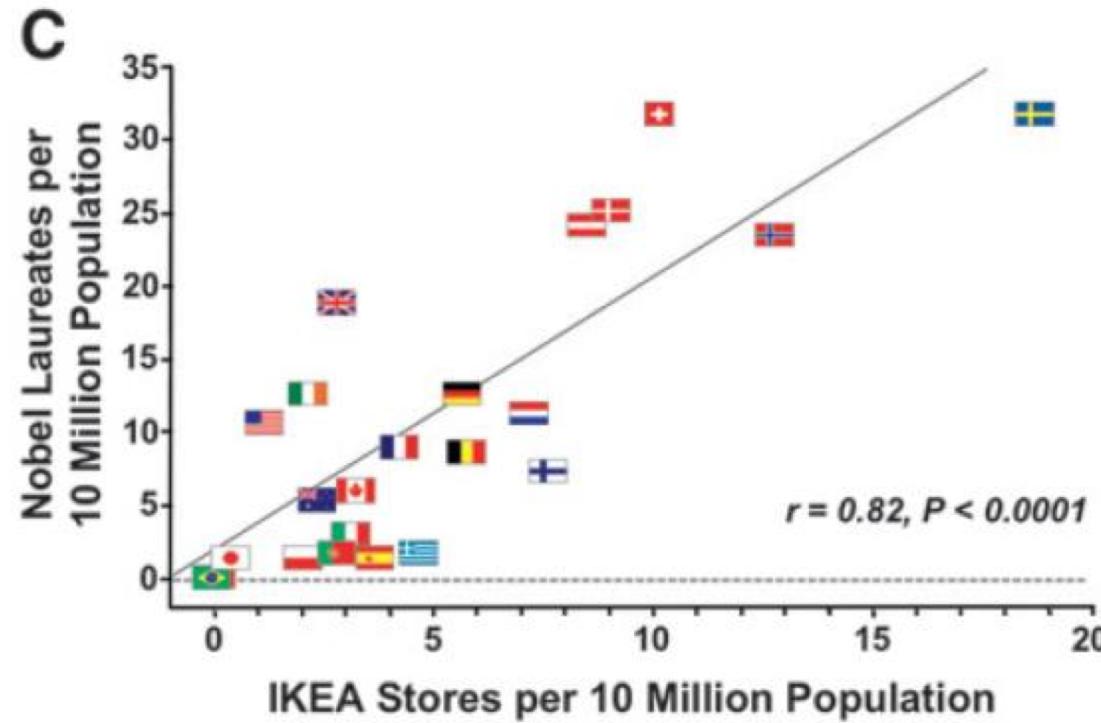
# Статистика в отборе признаков

# Корреляция



# Корреляция

Статистическая зависимость двух и более величин



ForexAW.com

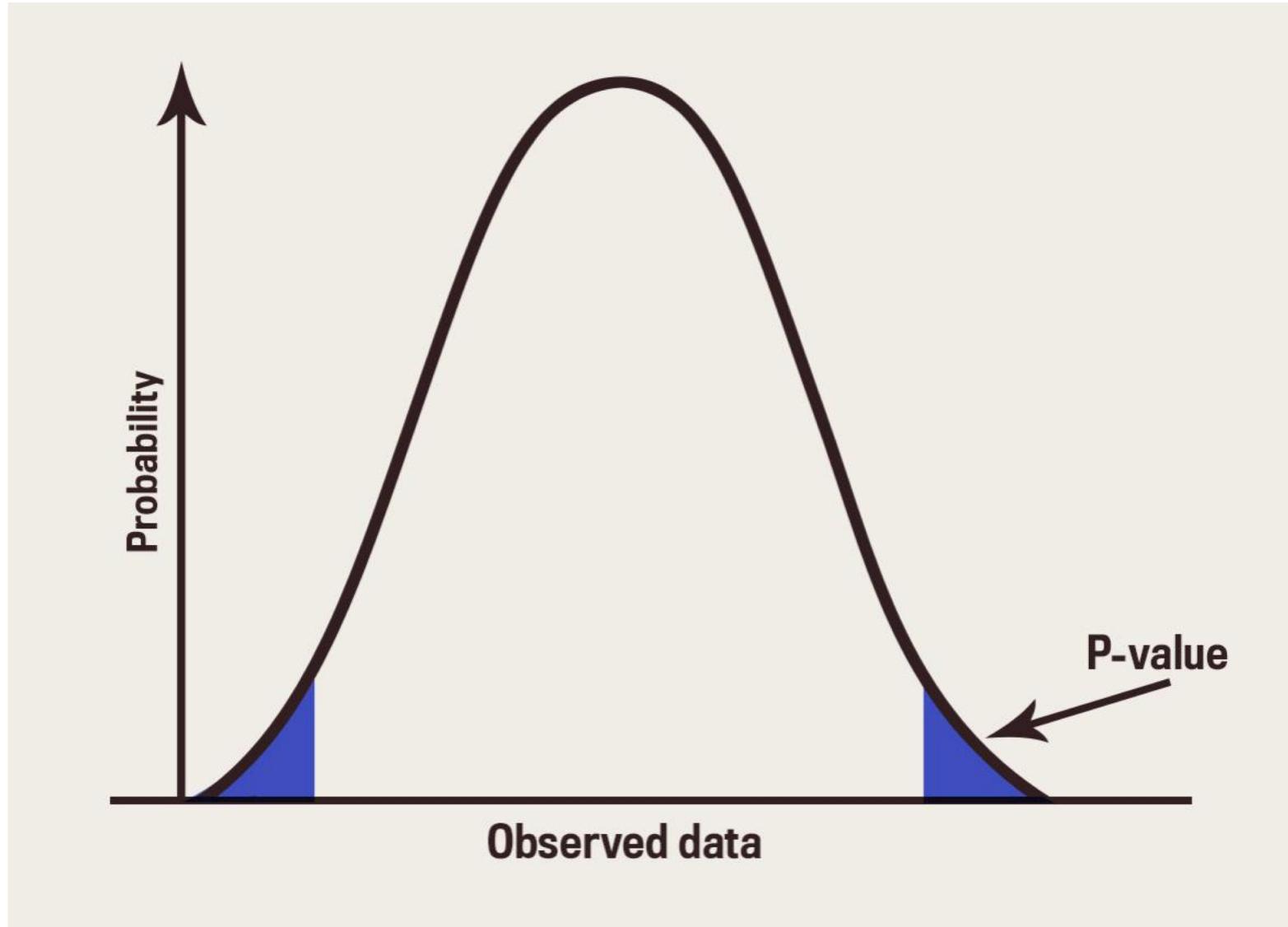
# T-статистика

$$t = \frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)}$$

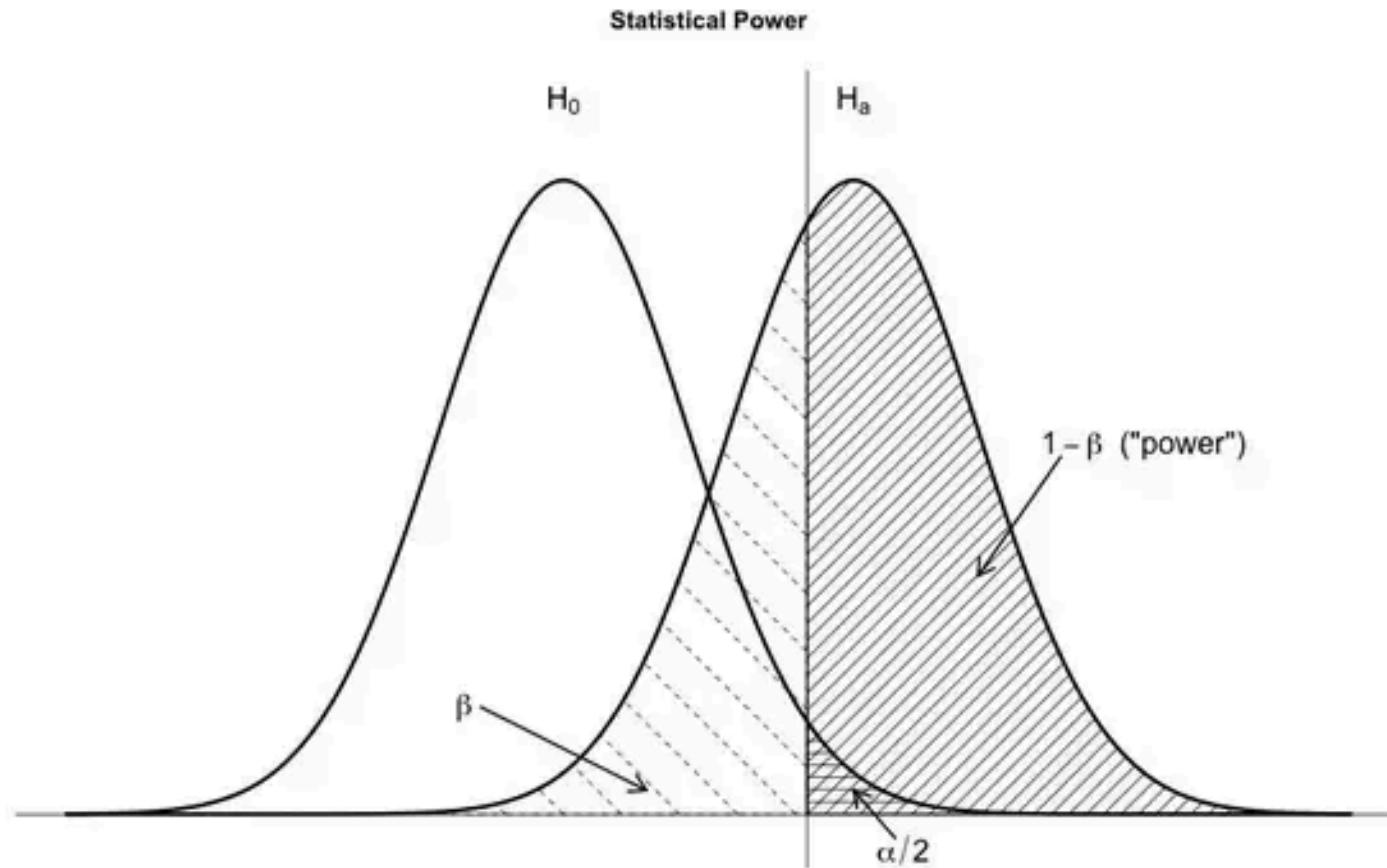
- Если между  $x_i$  и  $y$  нет зависимости, то  $t$  соответствует  $t$ -распределению с  $n-2$  степенями свободы
- $p\text{-value}$  - вероятность того, что при известном распределении наблюдаемое значение  $\geq |t|$  (при условии, что  $\beta_i = 0$ )
- Если  $p\text{-value}$  достаточно маленький ( $< 1\%$ ), то мы можем отклонить  $H_0$

--

# P-value



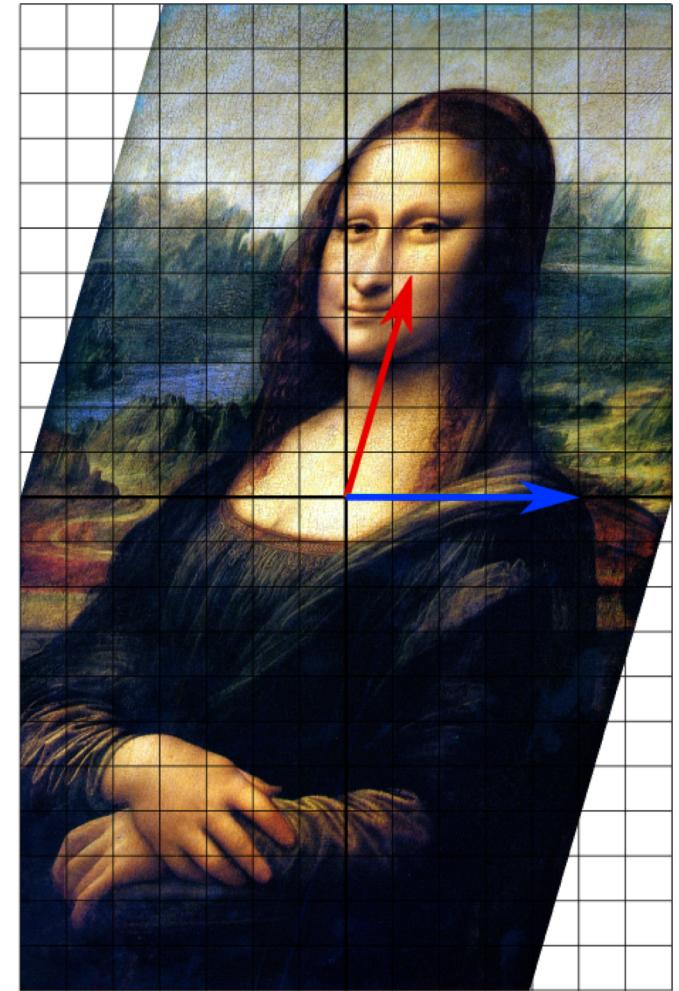
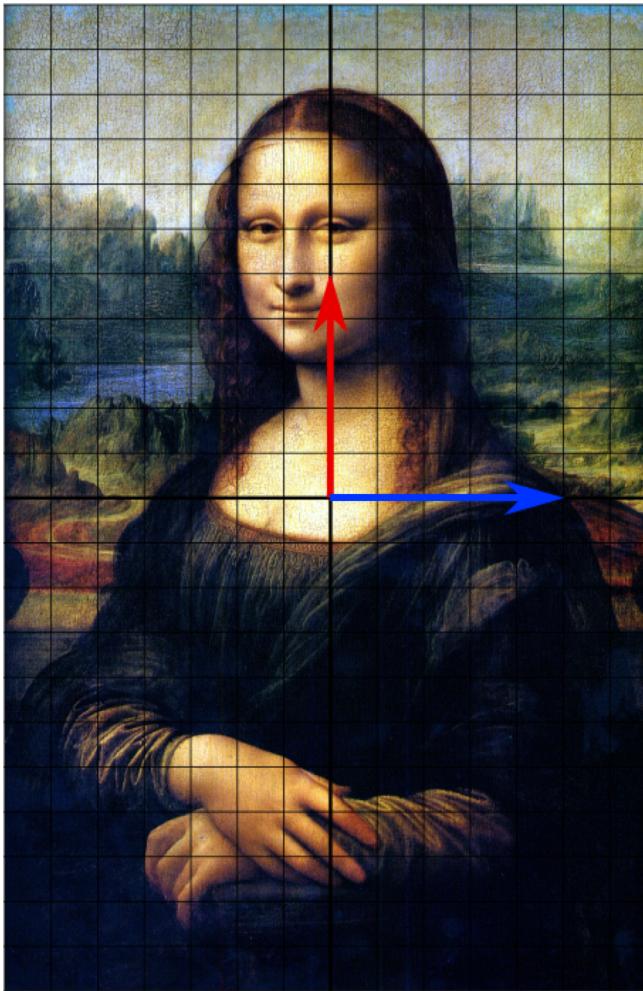
# P-value



# **Декомпозиция данных**

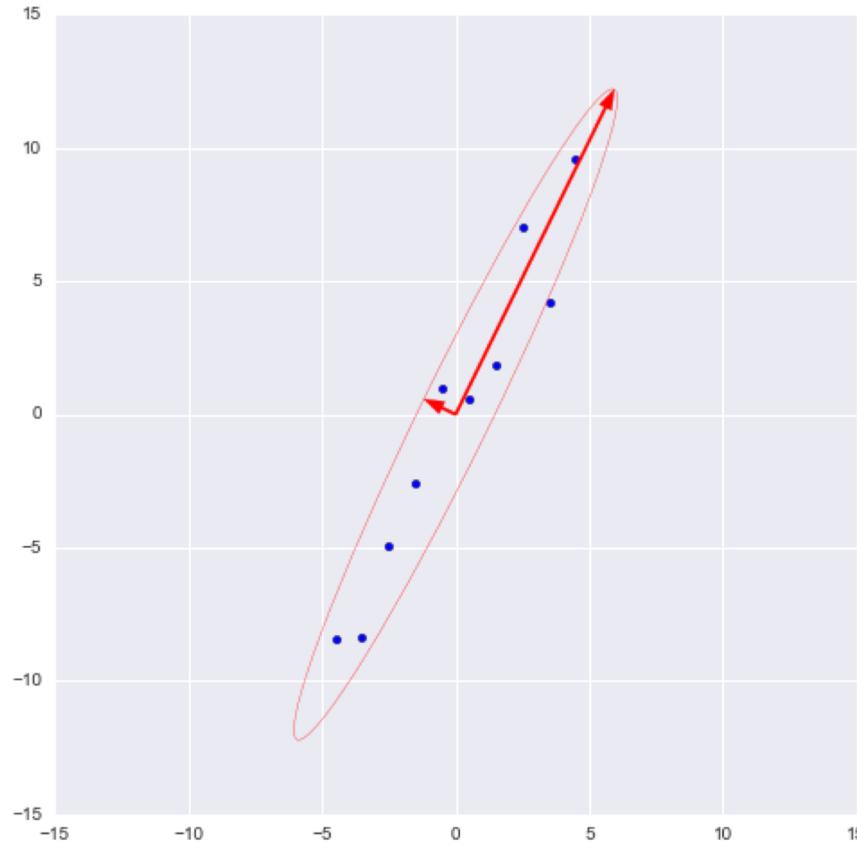
# Собственный вектор

$$M\vec{x} = \lambda\vec{x}$$



# PCA

Зачем он нужен? Он уменьшает размерность 😊



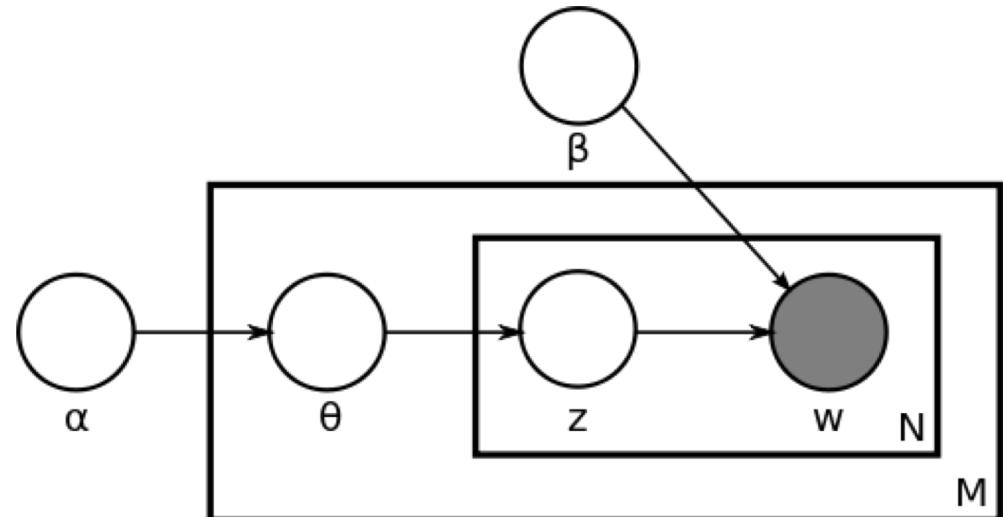
# PCA

$$Cov(X_i, X_j) = E[(X_i - E(X_i)) \cdot (X_j - E(X_j))] = E(X_i X_j) - E(X_i) \cdot E(X_j)$$

$$\begin{aligned}Var(X^*) &= \Sigma^* = E(X^* \cdot X^{*T}) = E((\vec{v}^T X) \cdot (\vec{v}^T X)^T) = \\&= E(\vec{v}^T X \cdot X^T \vec{v}) = \vec{v}^T E(X \cdot X^T) \vec{v} = \vec{v}^T \Sigma \vec{v}\end{aligned}$$

# LDA

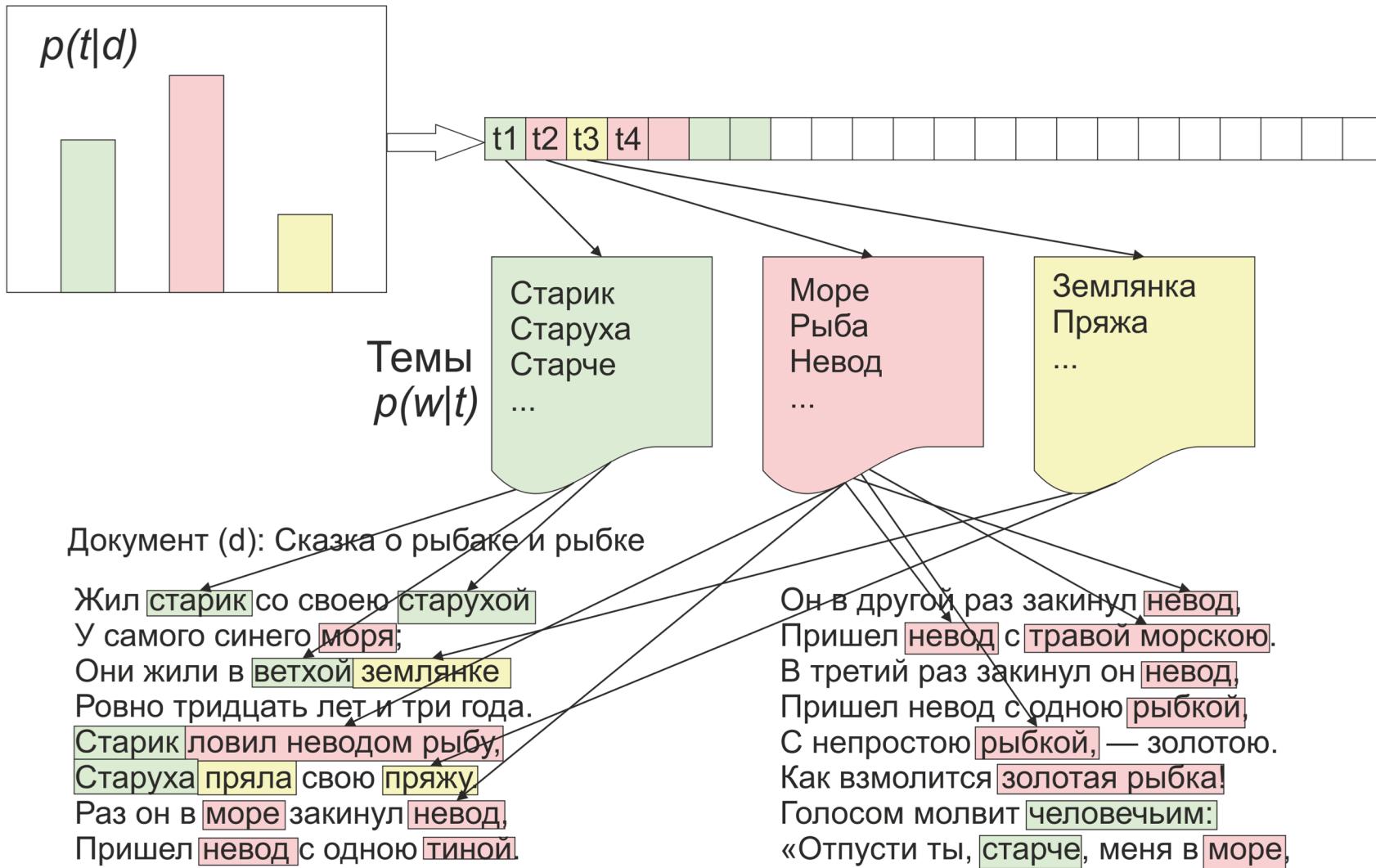
это иерархическая байесовская модель, состоящая из двух уровней:  
на первом уровне – смесь, компоненты которой соответствуют «темам»;  
на втором уровне – мультиномиальная переменная с априорным распределением Дирихле, которое задаёт «распределение тем» в документе.



$$p(\theta, \dots, N \mid \alpha, \beta) = p(N \mid \xi)p(\theta \mid \alpha) \prod_{n=1}^N p(z_n \mid \theta)p(w_n \mid z_n, \beta).$$

[https://en.wikipedia.org/wiki/Latent\\_Dirichlet\\_allocation](https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation)

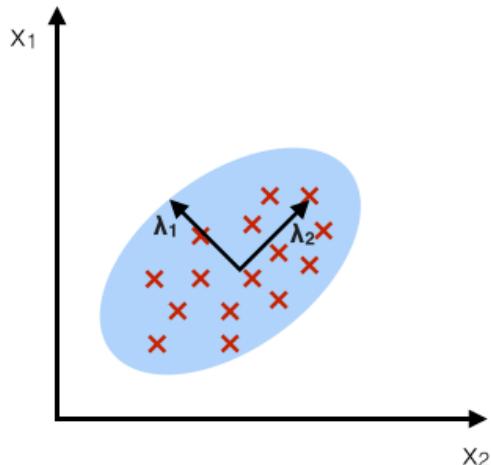
# LDA



# Сравнение

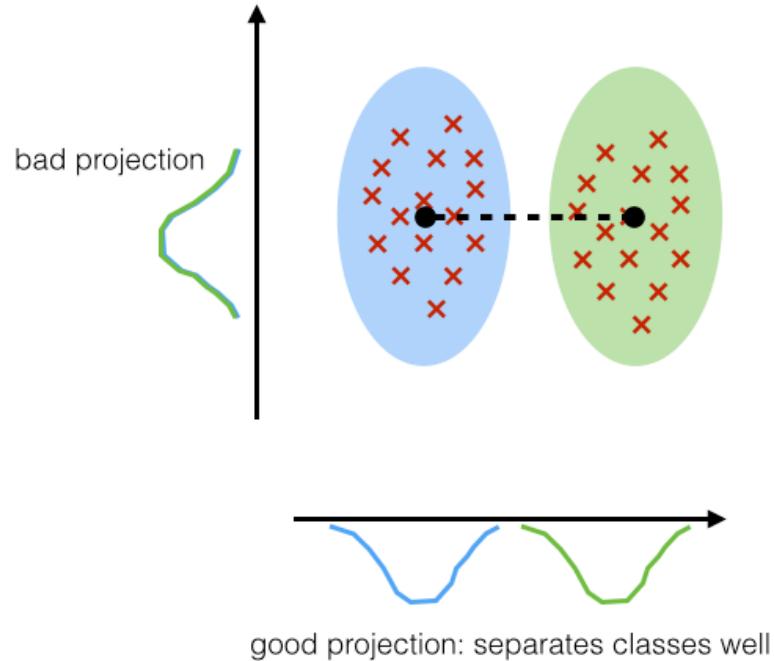
## PCA:

component axes that  
maximize the variance

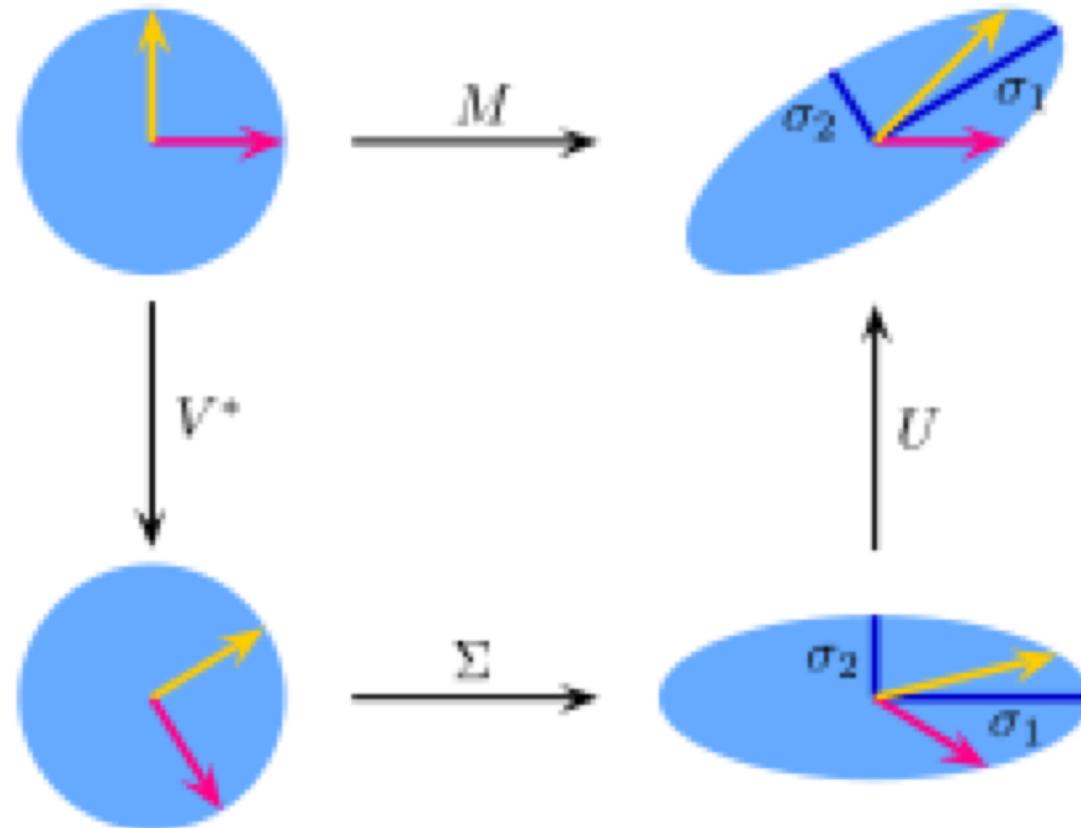


## LDA:

maximizing the component  
axes for class-separation



# SVD



$$M = U \cdot \Sigma \cdot V^*$$

# SVD

$$A_{n \times d} = \hat{U}_{n \times r} \Sigma_{r \times r} \hat{V}^T_{r \times d}$$

$U_{n \times n}$        $\Sigma_{n \times d}$        $V^T_{d \times d}$

# SVD

	Трактористы	Свинаяка и пастух	Once Upon a Tractor	Tractor, Love & Rock'n Roll	Babe
Вася	?	3	4	5	2
Пётр	3	5	2	2	5
Валерик	5	3		4	3
Жанночка	5	5	5		4
Петрович	2	3		2	2

mu: 2.54559533638261  
User base: 0.7271 0.1626 0.7139 1.9097 - 0.9677  
Item base: 0.8450 0.6593 0.2731 0.7328 0.0354  
**User features:**  
user 0: -0.5087 -0.8326  
user 1: 1.0220 1.2826  
user 2: -0.9509 0.2792  
user 3: 0.1031 -0.4814  
user 4: 0.6095 0.0557  
**Item features:**  
item 0: -0.8368 0.2511  
item 1: 1.1101 0.4120  
item 2: -0.4159 -0.4073  
item 3: -0.3130 -0.9115  
item 4: 0.6408 1.2205

# **ПРАКТИЧЕСКАЯ ЧАСТЬ**

# ВОПРОСЫ