My first scientific paper

Week 4

Plan the experiment

Vadim Strijov

Moscow Institute of Physics and Technology

2022

Вычислительный эксперимент

Цель эксперимента

Восстановить плотность пространственной конфигурации пары аминокислота-лиганд с помощью предложенного монотонного многослойного перцептрона.

Данные

В базе данных 20 аминокислотных остатков. Аминокислотные остатки взаимодействуют с лигандами, их 40. Взаимодействие описывается 5 признаками: тип аминокислотного остатка, тип лиганда, расстояние и 2 угла θ и φ.

Гипотеза

Максимумы полученной плотности распределения соответствуют устойчивым пространственным конфигурациям.

Вычислительный эксперимент

Цель

Восстановить плотности распределения пространственных ориентаций различных пар вида аминокислота-лиганд.

Данные

Данные представляют собой 47916041 пятерку значений, элементы каждой пятерки: a — индекс аминокислоты, b — индекс лиганда и тройка r, θ, φ. Индексы аминокислоты и лиганды образуют 840 пар и используются для разделения данных на 840 выборок (r, θ, φ), каждая из которых соответсвует своей взаимодействующей паре.

Описание эксперимента

Для каждой из 840 выборок строится восстановленная плотность $\hat{p}^{a,b}(r, \theta, \varphi) = p(r, \theta, \varphi | \boldsymbol{w}^*, \mathbf{U}^*)$.

Вычислительный эксперимент

Наборы данных

Набор	Нативные структуры	Модели структур	Разбиение
CASP 9	117	35963	Train, Validation
CASP 10	103	15450	
CASP 11	84	12291	
CASP 12	37	5501	Test

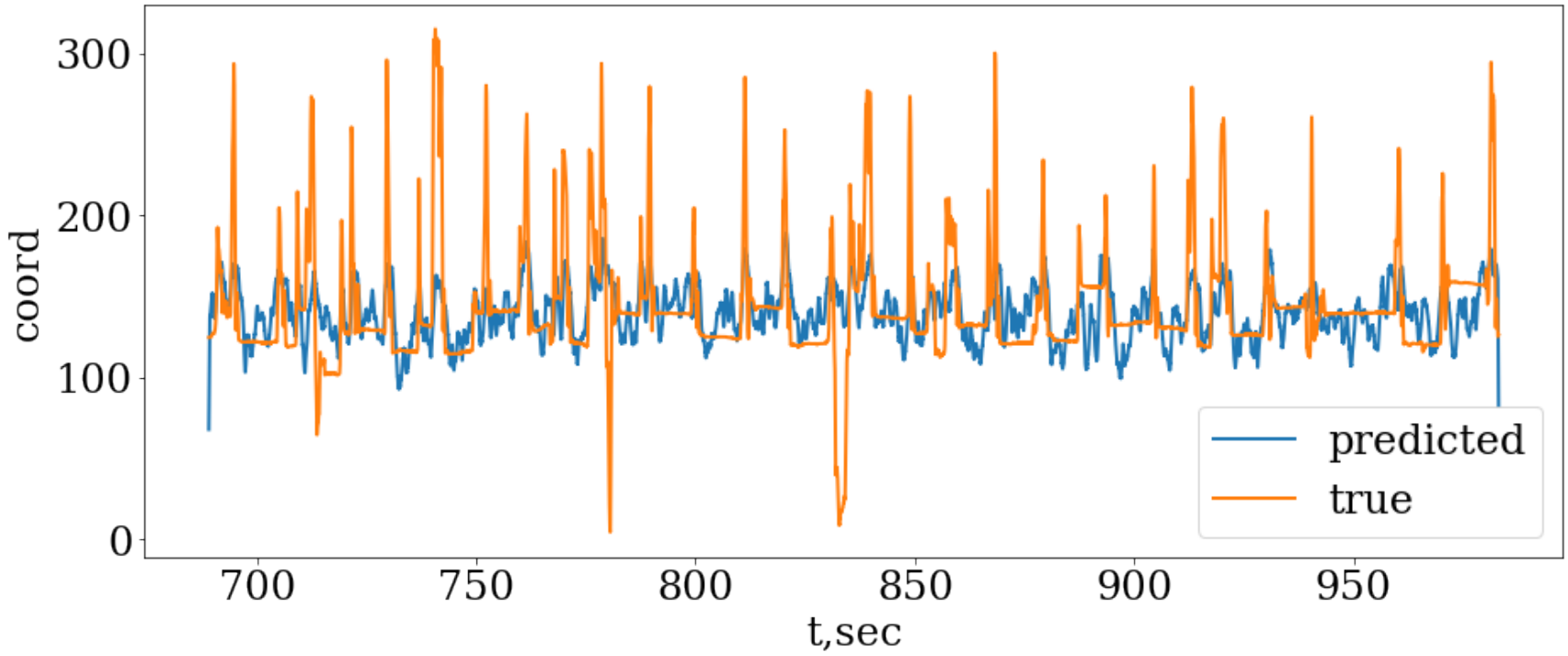
При обучении нейросети анализируются усредненные по T нативным структурам коэффициенты корреляции Пирсона и Спирмена

$$R = R(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{T}\sum_{i=1}^{T} R_i^{\text{target}} = \frac{1}{T}\sum_{i=1}^{T} \text{PEARSON}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

$$\rho = \rho(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{T}\sum_{i=1}^{T} \rho_i^{\text{target}} = \frac{1}{T}\sum_{i=1}^{T} \text{SPEARMAN}(\mathbf{y}_i, \hat{\mathbf{y}}_i)$$

Описание экспериментов

- Требуется восстановить координату движения конечности по сигналу электрокортикограммы.
- 5 временных рядов по 20 минут, первые 15 минут обучение, остальные 5 минут — тест.
- Критерий качества: коэффициент корреляции между предсказанной траекторией и истинной.



Зависимость предсказанной и истинной траекторий от времени

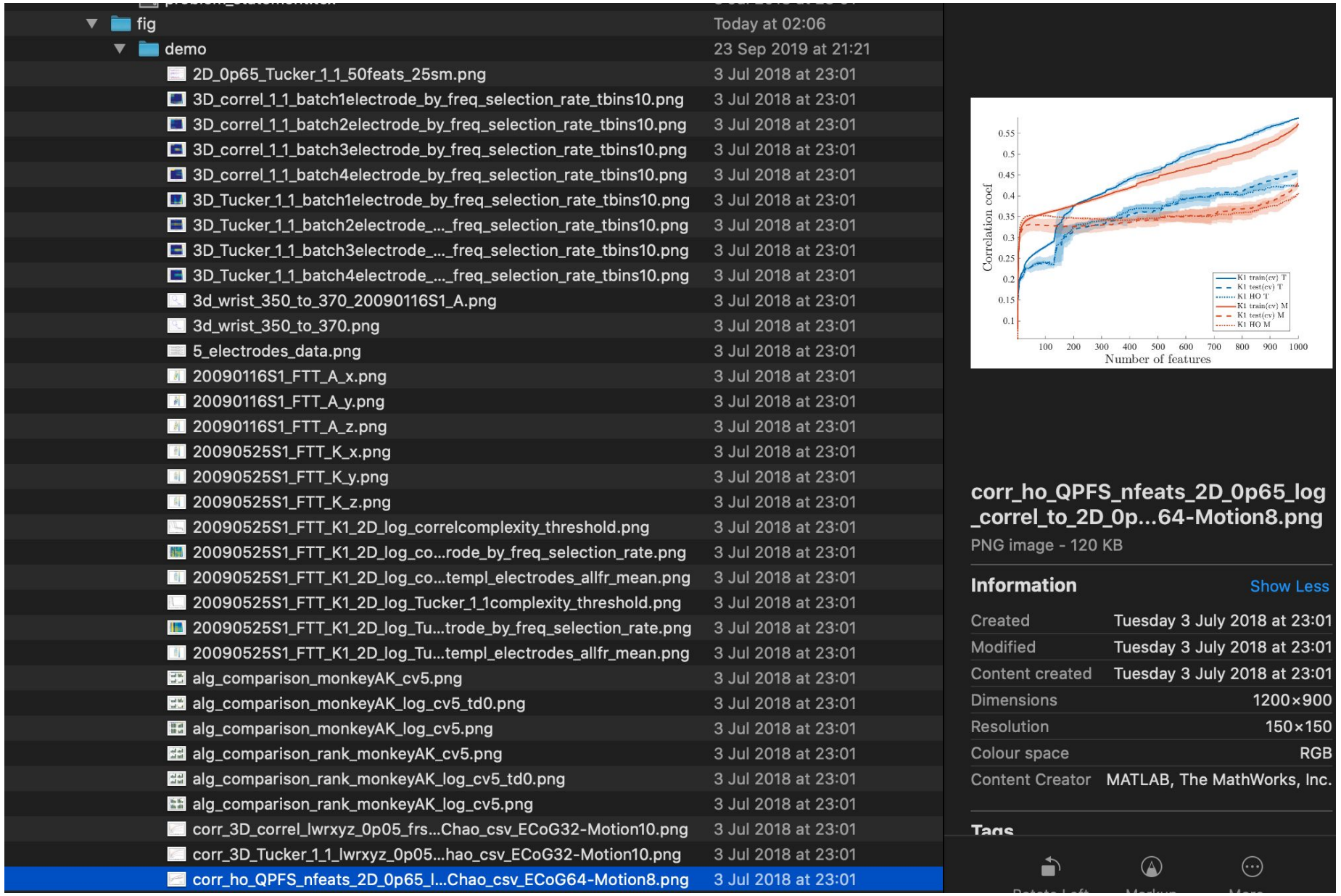


fig
Today at 02:06
demo
23 Sep 2019 at 21:21
2D_0p65_Tucker_1_1_50feats_25sm.png
3 Jul 2018 at 23:01
3D_correl_1_1_batch1electrode_by_freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_correl_1_1_batch2electrode_by_freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_correl_1_1_batch3electrode_by_freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_correl_1_1_batch4electrode_by_freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_Tucker_1_1_batch1electrode_by_freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_Tucker_1_1_batch2electrode_..._freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_Tucker_1_1_batch3electrode_..._freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3D_Tucker_1_1_batch4electrode_..._freq_selection_rate_tbins10.png
3 Jul 2018 at 23:01
3d_wrist_350_to_370_20090116S1_A.png
3 Jul 2018 at 23:01
3d_wrist_350_to_370.png
3 Jul 2018 at 23:01
5_electrodes_data.png
3 Jul 2018 at 23:01
20090116S1_FTT_A_x.png
3 Jul 2018 at 23:01
20090116S1_FTT_A_y.png
3 Jul 2018 at 23:01
20090116S1_FTT_A_z.png
3 Jul 2018 at 23:01
20090525S1_FTT_K_x.png
3 Jul 2018 at 23:01
20090525S1_FTT_K_y.png
3 Jul 2018 at 23:01
20090525S1_FTT_K_z.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_correlcomplexity_threshold.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_co...rode_by_freq_selection_rate.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_co...templ_electrodes_allfr_mean.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_Tucker_1_1complexity_threshold.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_Tu...trode_by_freq_selection_rate.png
3 Jul 2018 at 23:01
20090525S1_FTT_K1_2D_log_Tu...templ_electrodes_allfr_mean.png
3 Jul 2018 at 23:01
alg_comparison_monkeyAK_cv5.png
3 Jul 2018 at 23:01
alg_comparison_monkeyAK_log_cv5_td0.png
3 Jul 2018 at 23:01
alg_comparison_monkeyAK_log_cv5.png
3 Jul 2018 at 23:01
alg_comparison_rank_monkeyAK_cv5.png
3 Jul 2018 at 23:01
alg_comparison_rank_monkeyAK_log_cv5_td0.png
3 Jul 2018 at 23:01
alg_comparison_rank_monkeyAK_log_cv5.png
3 Jul 2018 at 23:01
corr_3D_correl_lwrxyz_0p05_frs...Chao_csv_ECoG32-Motion10.png
3 Jul 2018 at 23:01
corr_3D_Tucker_1_1_lwrxyz_0p05...hao_csv_ECoG32-Motion10.png
3 Jul 2018 at 23:01
corr_ho_QPFS_nfeats_2D_0p65_l...Chao_csv_ECoG64-Motion8.png
3 Jul 2018 at 23:01
Correlation coef
Number of features
K1 train(cv) T
K1 test(cv) T
K1 HO T
K1 train(cv) M
K1 test(cv) M
K1 HO M
corr_ho_QPFS_nfeats_2D_0p65_log_correl_to_2D_0p...64-Motion8.png
PNG image - 120 KB
Information
Show Less
Created
Tuesday 3 July 2018 at 23:01
Modified
Tuesday 3 July 2018 at 23:01
Content created
Tuesday 3 July 2018 at 23:01
Dimensions
1200×900
Resolution
150×150
Colour space
RGB
Content Creator
MATLAB, The MathWorks, Inc.
Tags

Планирование коммерческих проектов (m1p.org)

Научный проект: исследует свойства объекта, решает узкую задачу

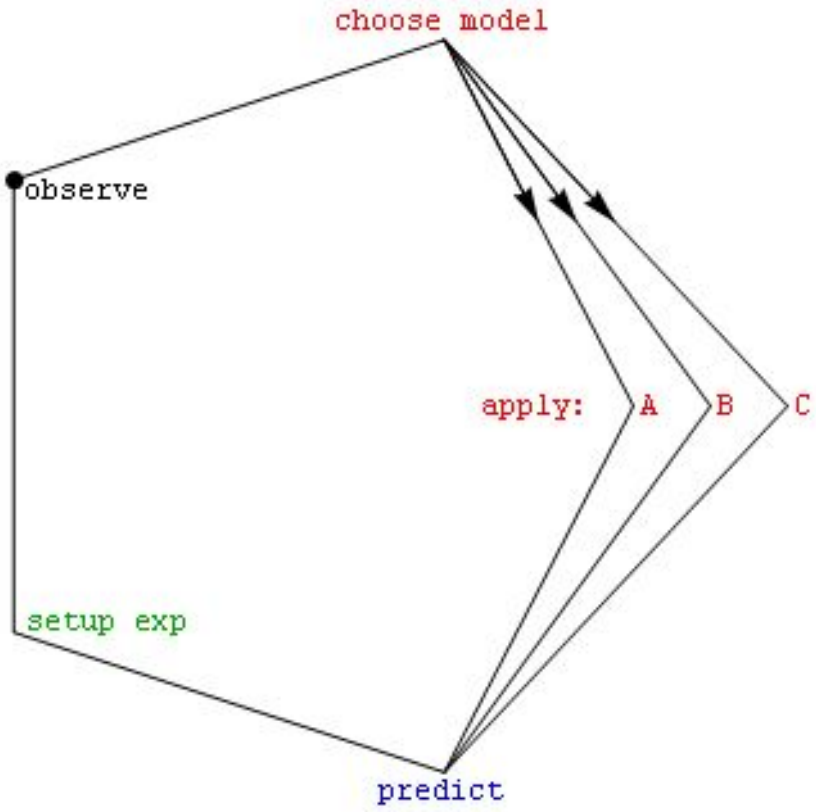
Коммерческий проект: создает систему, продукт

Правила проектировния:

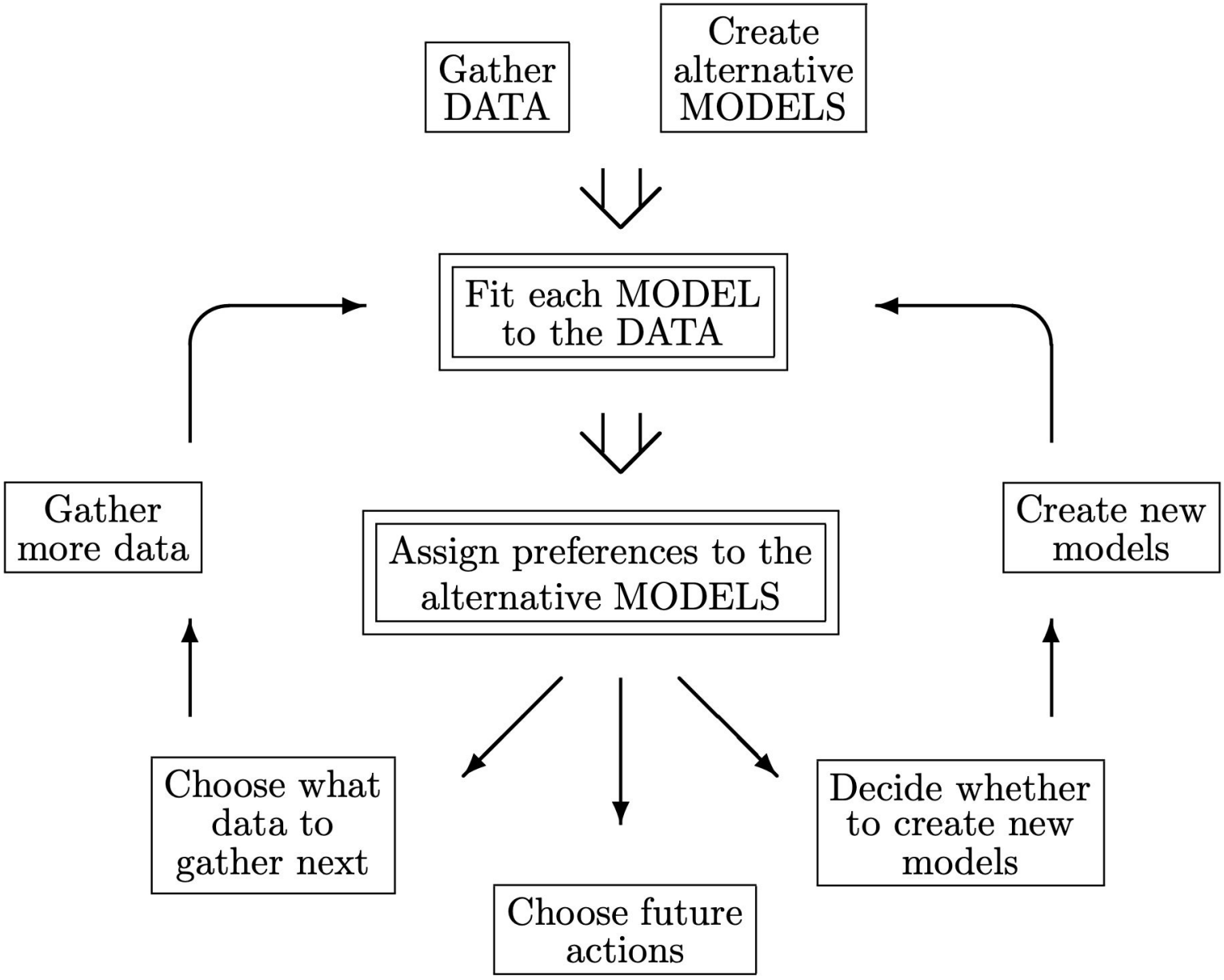
1) система полна,
2) система иерархична,
3) система имеет обратную связь

https://docs.google.com/presentation/d/1Y8uRX_T_RBIQtEcboEND6-kTJ-c2uqu6-BzVItSQ_-c/edit#slide=id.g71982c3020_0_0

The scientific observation cycle (model selection)



wikipedia.org



Hour by Hour Energy Forecasting

Data:

- historical consumption and prices, multivariate time series.

To forecast:

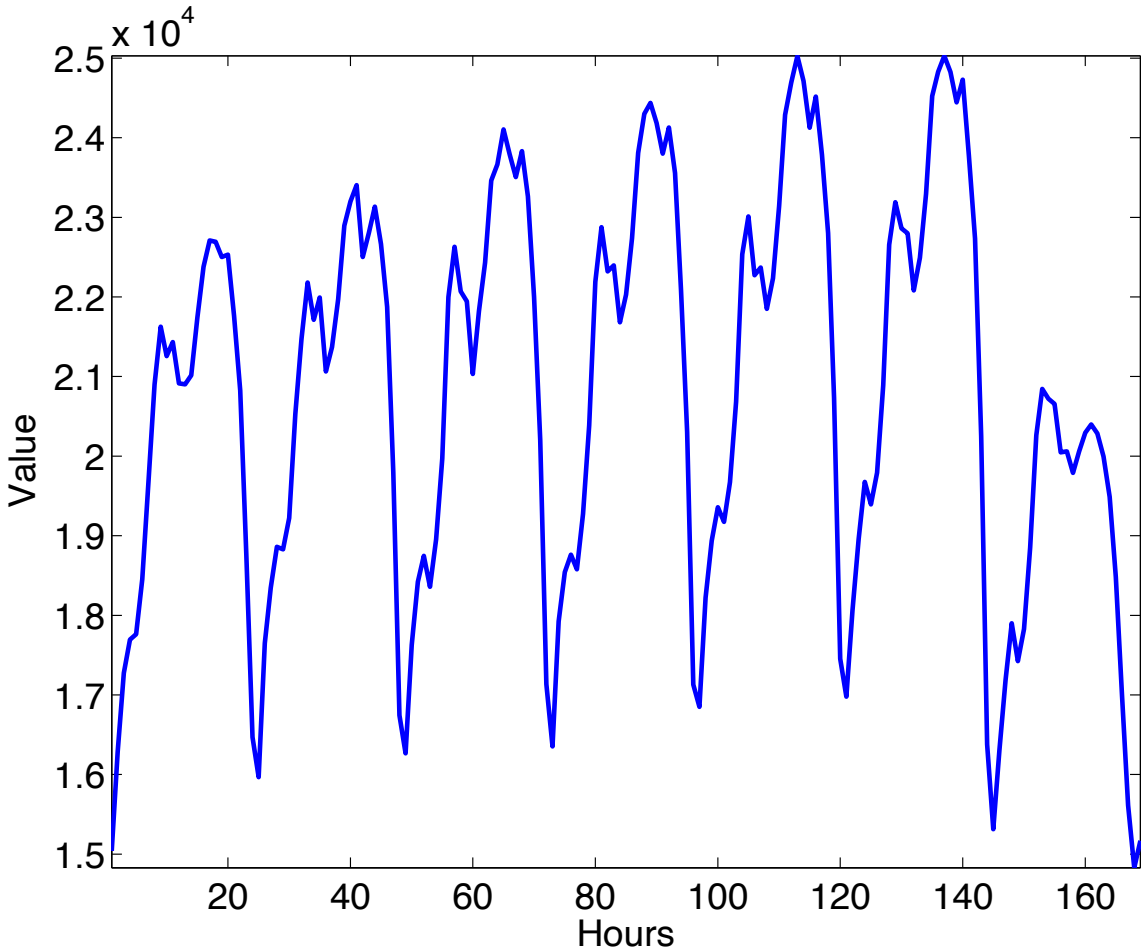
- hour-by-hour, the next day

  ✓ consumption and
  ✓ price.

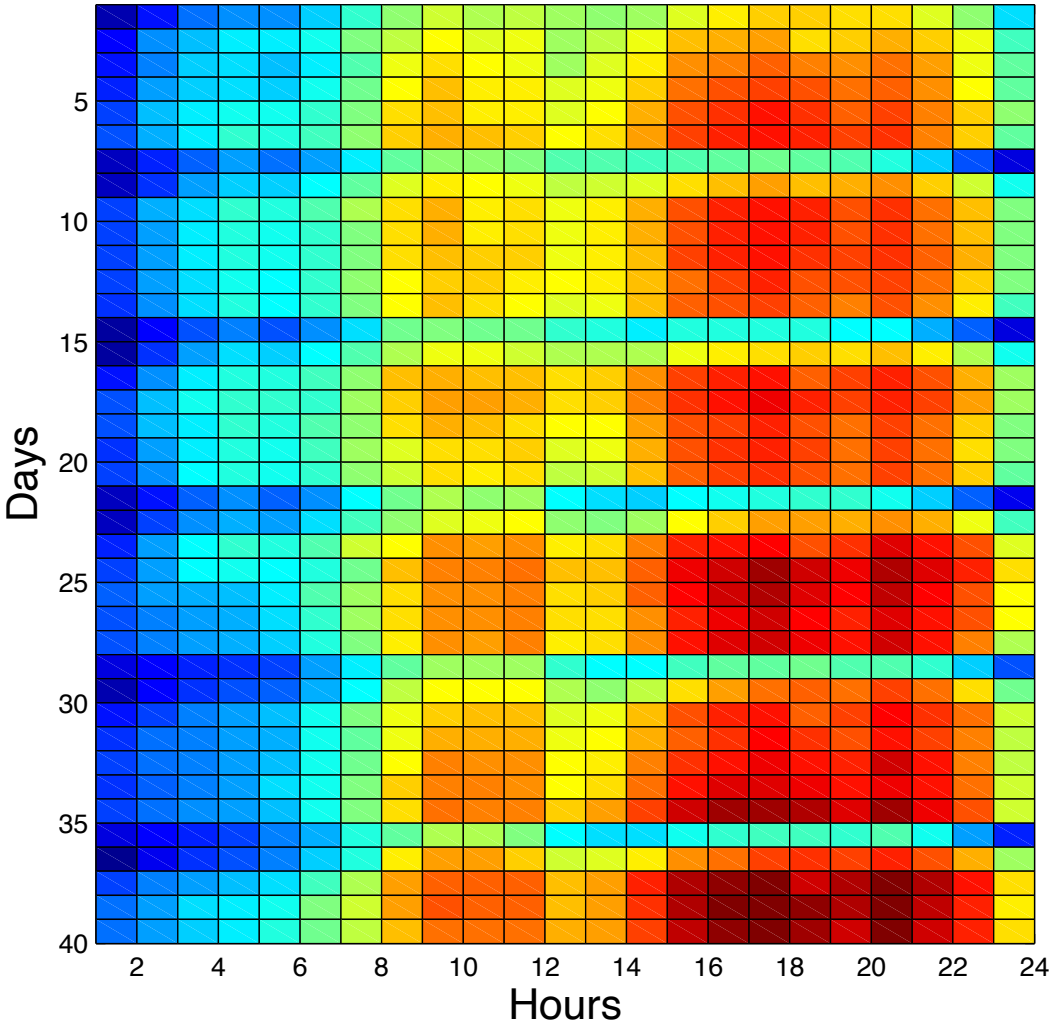
Solution:

- the autoregressive model generation and model selection.

Source time series, one week



The autoregressive matrix, five week-ends



The autoregressive matrix and the linear model

$$
\underset{(m+1)\times(n+1)}{X^*} = \left(\begin{array}{c|ccc}
s_T & s_{T-1} & \cdots & s_{T-\kappa+1} \\
\hline
s_{(m-1)\kappa} & s_{(m-1)\kappa-1} & \cdots & s_{(m-2)\kappa+1} \\
\cdots & \cdots & \cdots & \cdots \\
s_{n\kappa} & s_{n\kappa-1} & \cdots & s_{n(\kappa-1)+1} \\
\cdots & \cdots & \cdots & \cdots \\
s_{\kappa} & s_{\kappa-1} & \cdots & s_1
\end{array}\right).
$$

In a nutshell,

$$
X^* = \left[\begin{array}{c|c}
\underset{1\times 1}{s_T} & \underset{1\times n}{\mathbf{x}_{m+1}} \\
\hline
\underset{m\times 1}{\mathbf{y}} & \underset{m\times n}{X}
\end{array}\right].
$$

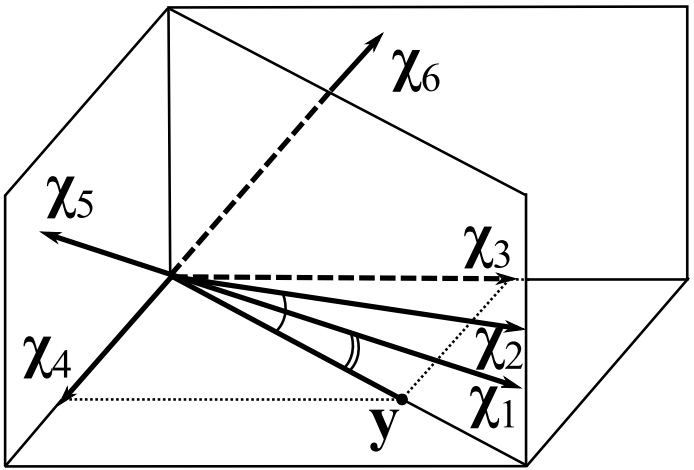
In terms of linear regression:

$$
\mathbf{y} = X\mathbf{w},
$$

$$
y_{m+1} = s_T = \mathbf{w}^{\mathsf{T}}\mathbf{x}_{m+1}^{\mathsf{T}}.
$$

Выбор устойчивой и точной модели

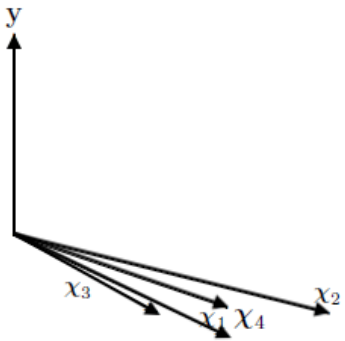
Выборка содержит мультикоррелирующие χ_1, χ_2 и устойчивые χ_5, χ_6 признаки — столбцы матрицы «объект-признак» $\mathbf{X}$. Требуется выбрать два признака из шести.



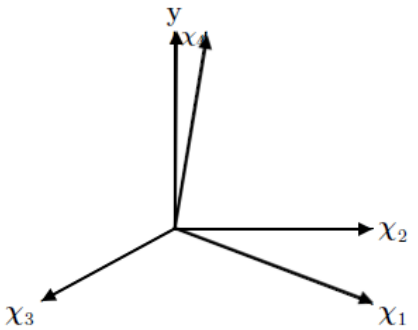
Точность и устойчивость при заданной сложности

Решение: χ_3, χ_4 — набор ортогональных признаков с наименьшим значением функции ошибки.

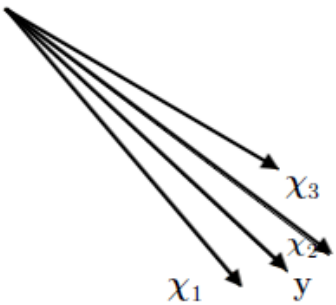
Configurations of design space



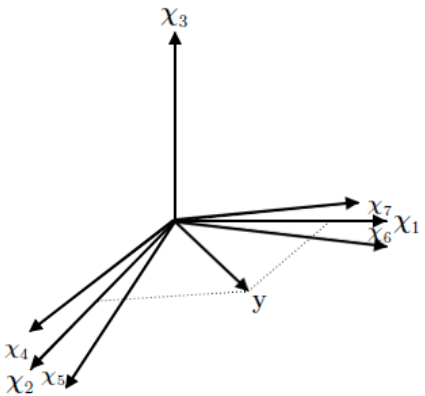
Non-adequate correlated



Adequate random



Adequate redundant



Adequate correlated

Katrutsa, Strijov. 2017. Comprehensive study of feature selection methods to solve multicollinearity problem // Expert Systems with Applications

Задача выбора оптимального набора признаков

- Задана выборка $D = \{(\mathrm{x}_i, y_i)\}$, $i \in \mathcal{I}$.
- Задано случайное разбиение множество индексов элементов выборки $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$.
- Множество независимых переменных $\mathrm{x} = [x_1, \ldots, x_j, \ldots, x_n]$ проиндексировано $j \in \mathcal{J} = \{1, \ldots, n\}$.
- Задано множество моделей-претендентов $\mathfrak{F} = \{f(\mathrm{w}, \mathrm{x})\}$.
- Модель — параметрическое семейство функций $f(\mathrm{w}, \mathrm{x}) = \mu(\mathrm{w}^\mathsf{T}\mathrm{x})$, где μ — функция связи (в случае регрессии $\mu = \mathrm{id}$, в случае классификации $\mu = \frac{1}{1+\exp(-\mathrm{w}^\mathsf{T}\mathrm{x})}$.
- Структура модели $f_\mathcal{A}$ задана множеством индексов $\mathcal{A} \subseteq \mathcal{J}$ и означает включение переменных $\mathrm{x}_\mathcal{A}$. Иначе, используются только признаки-столбцы матрицы X с индексами из множества $\mathcal{A}$.
- Задана функция ошибки S.

Задача выбора оптимального набора признаков

Требуется найти такое подмножество индексов $\mathcal{A} \subseteq \mathcal{J}$, которое бы доставляло минимум функции

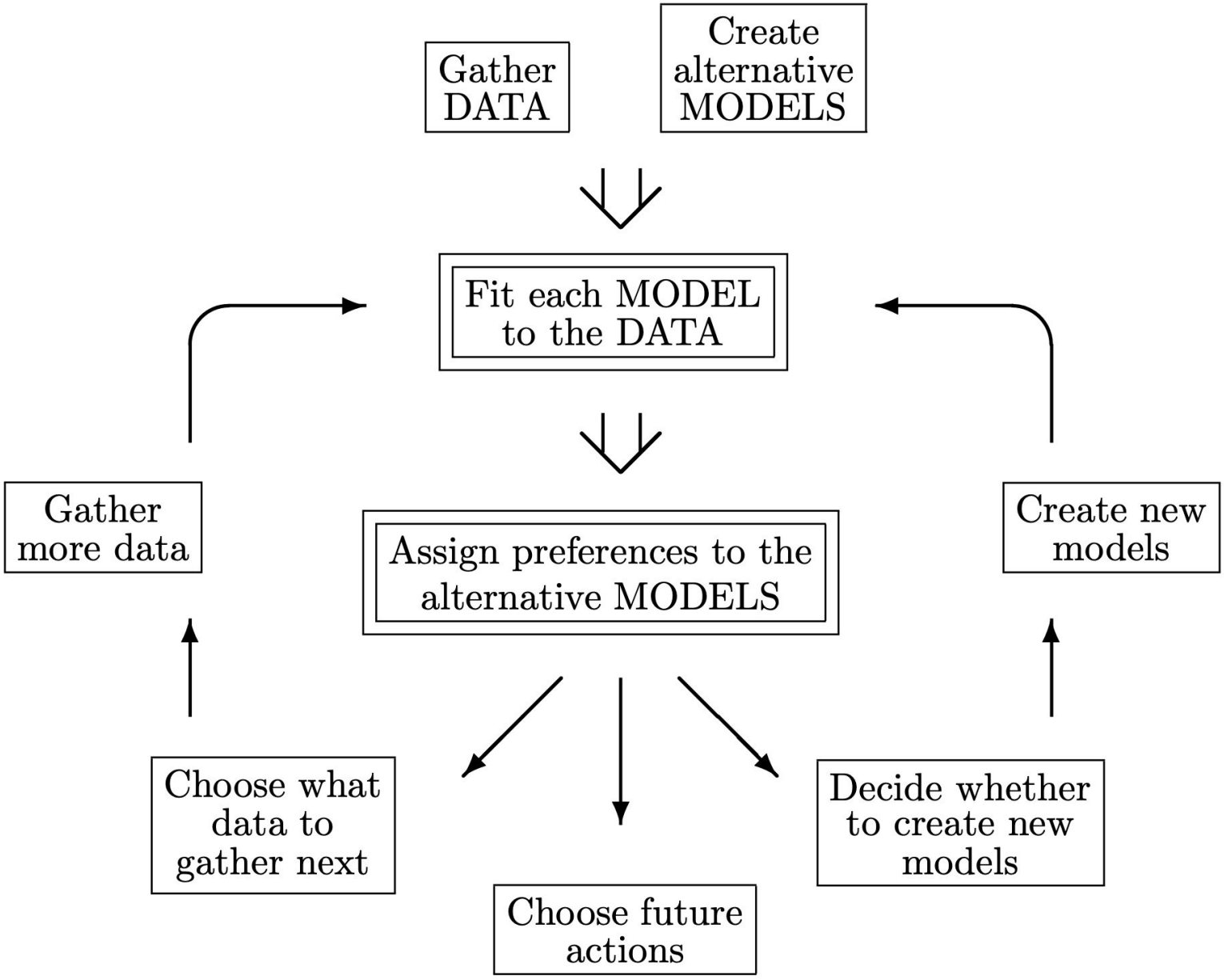
$$\mathcal{A}^* = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S(f_{\mathcal{A}} | \mathrm{w}^*, D_{\mathcal{C}})$$

на разбиении выборки D, определенном множеством индексов $\mathcal{C}$.

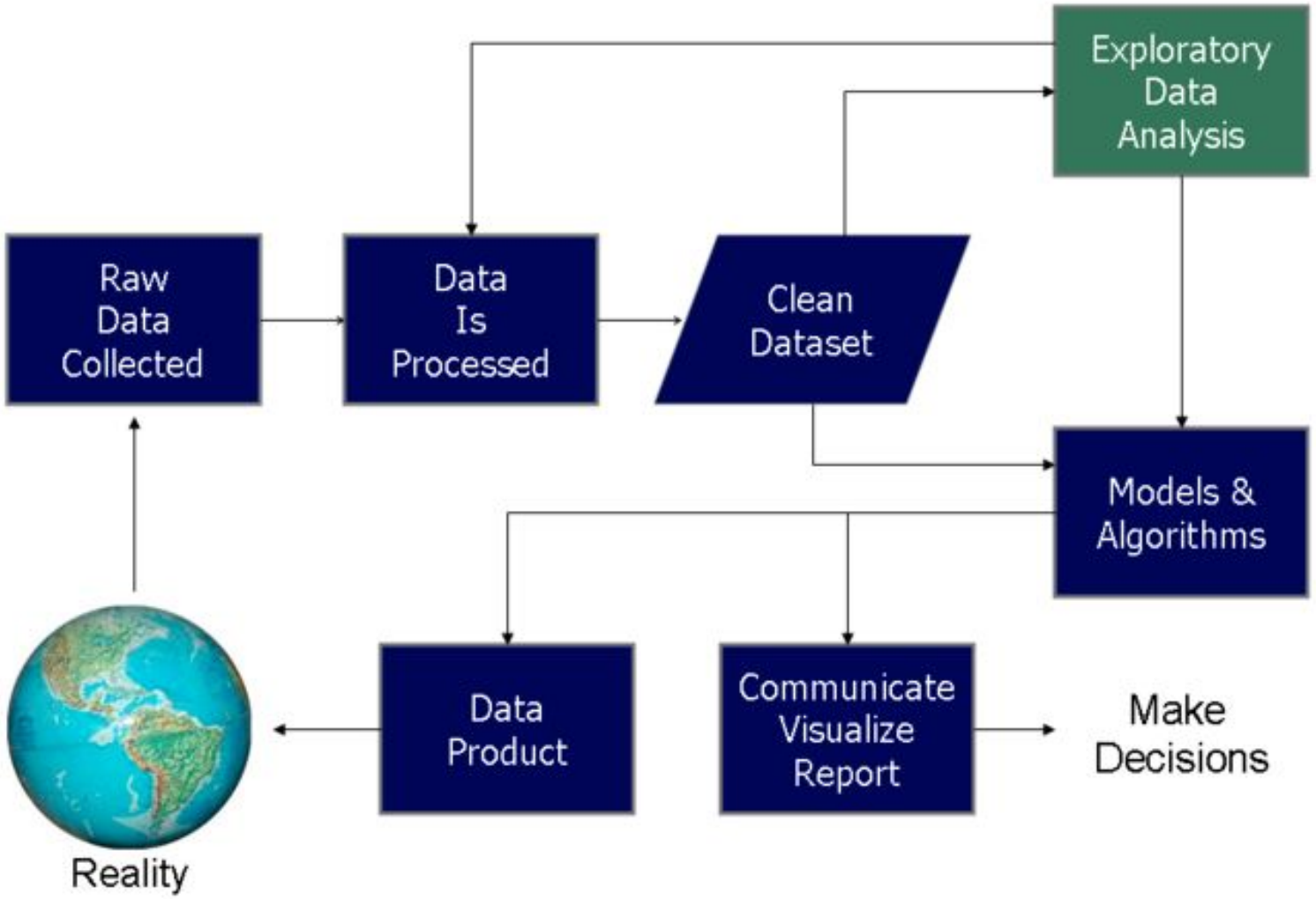
При этом параметры w^* модели должны доставлять минимум функции

$$\mathrm{w}^* = \arg\min_{\mathrm{w} \in \mathbb{W}} S(\mathrm{w} | D_{\mathcal{L}}, f_{\mathcal{A}})$$

на разбиении выборки, определенном множеством $\mathcal{L}$.



Data Science Process

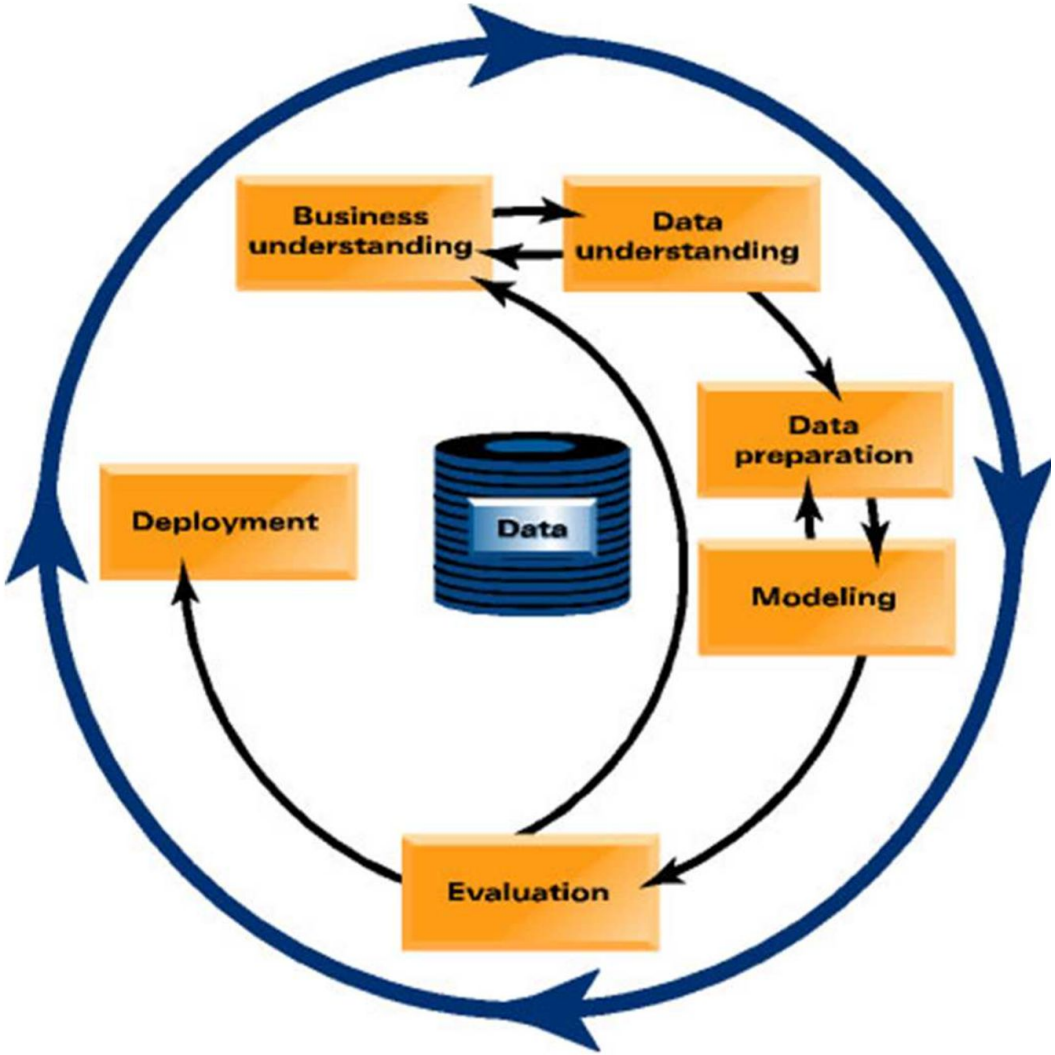


wikipedia.org

Cross-industry standard process for data mining (CRISPR-DM)

Six major phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment



Project life cycle

Р 50.1.028—2001

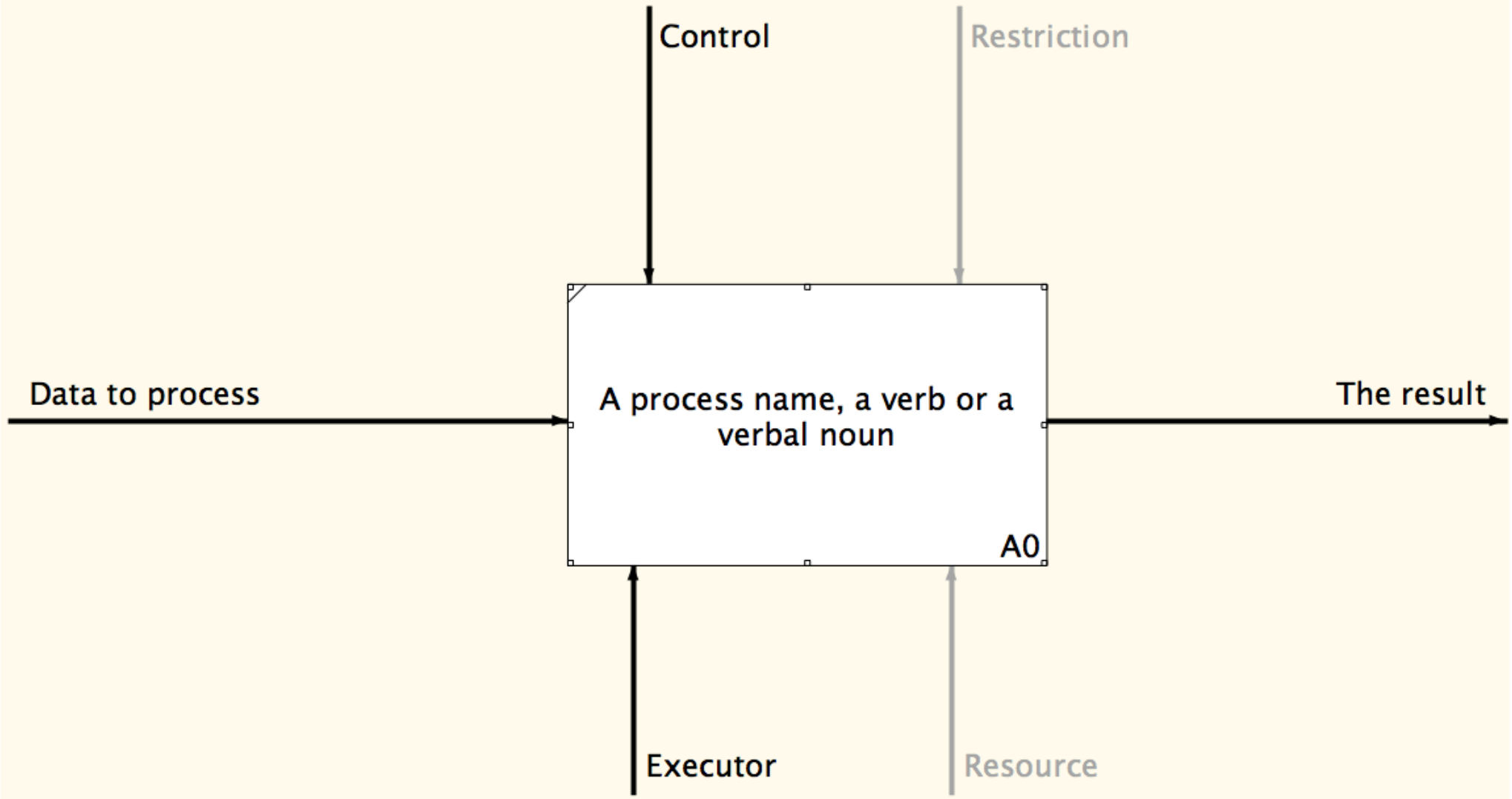
РЕКОМЕНДАЦИИ ПО СТАНДАРТИЗАЦИИ

Информационные технологии поддержки жизненного цикла продукции

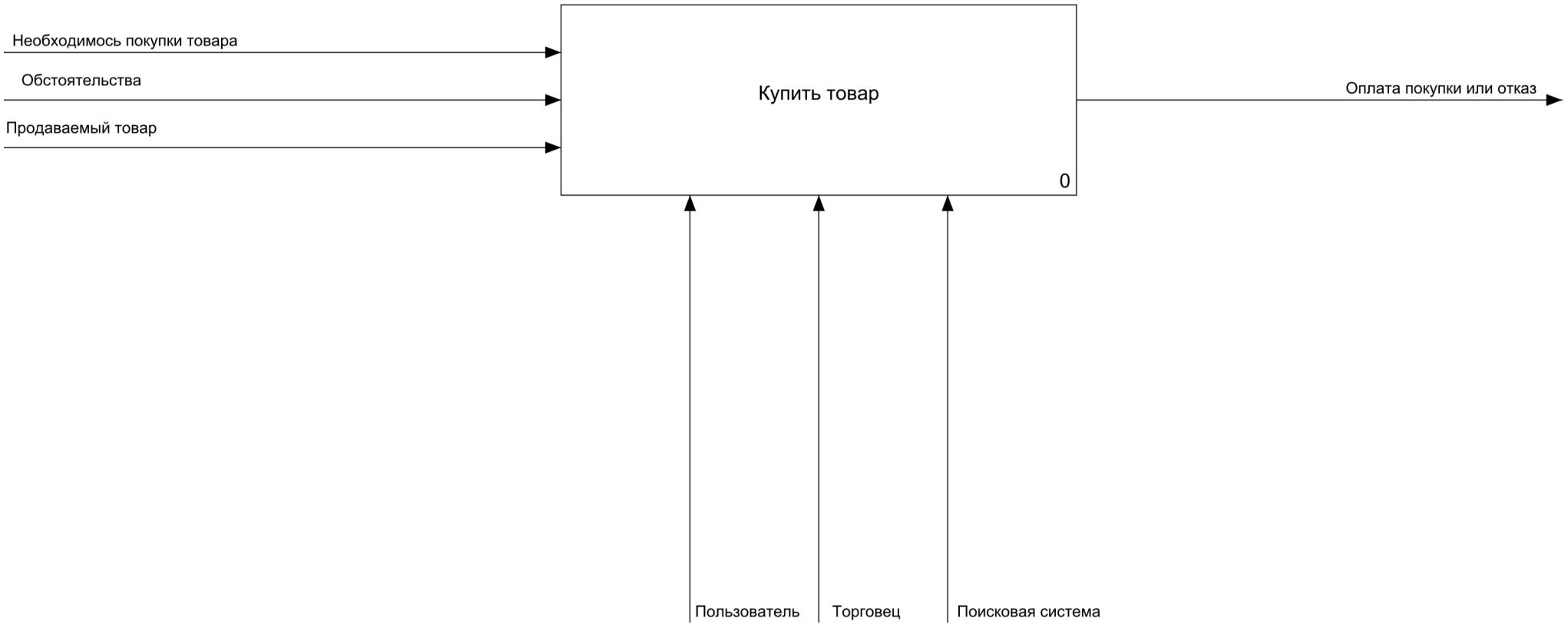
МЕТОДОЛОГИЯ ФУНКЦИОНАЛЬНОГО МОДЕЛИРОВАНИЯ

Издание официальное

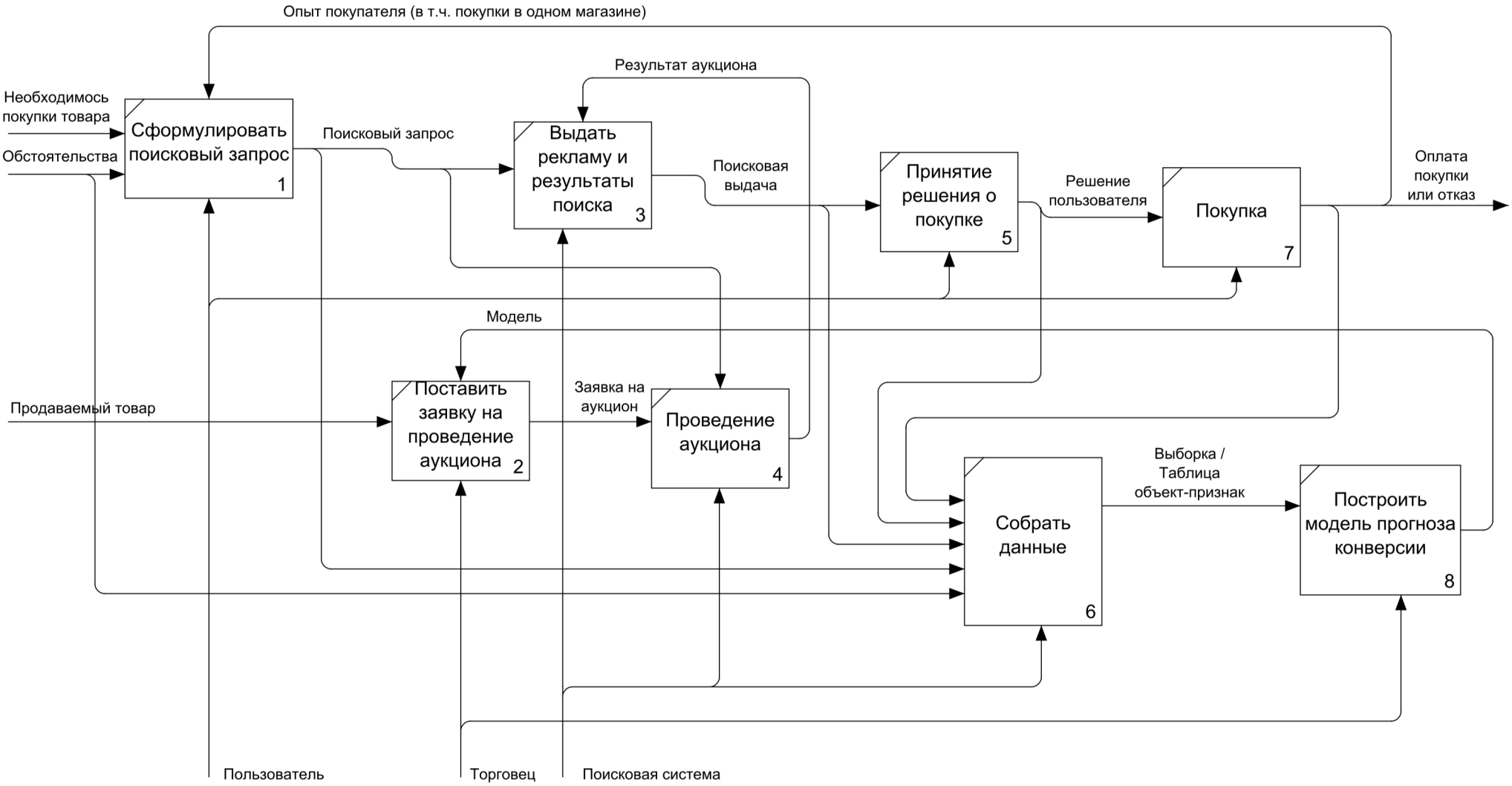
Функциональная схема IDEF0, лист A-0



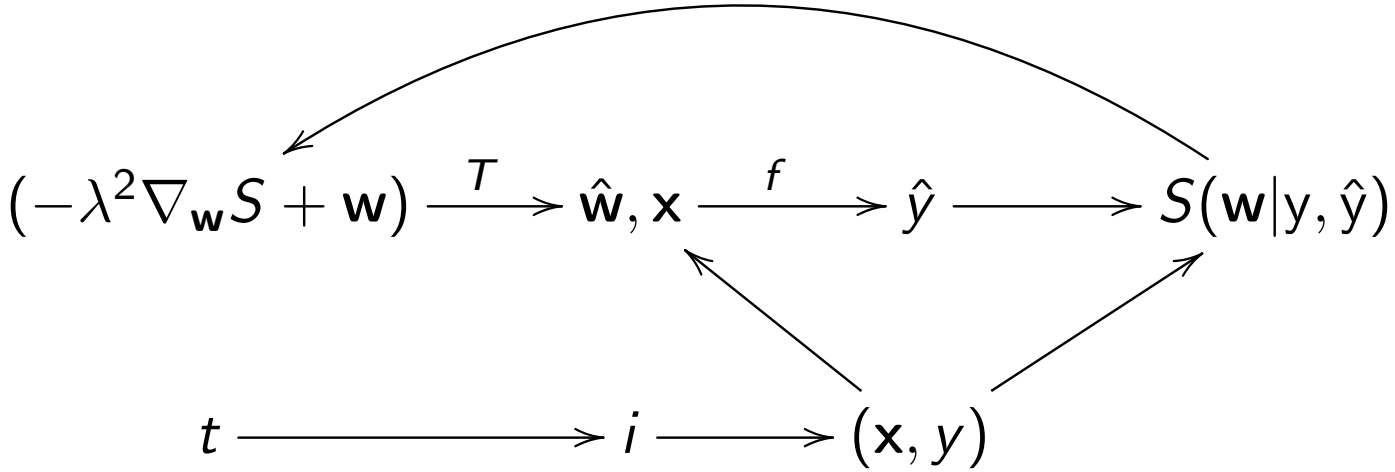
Непрерывный аукцион в поисковых системах, лист A-0



Непрерывный аукцион в поисковых системах, лист A0



The simplest problem statement in machine learning



f is the forecasting model,

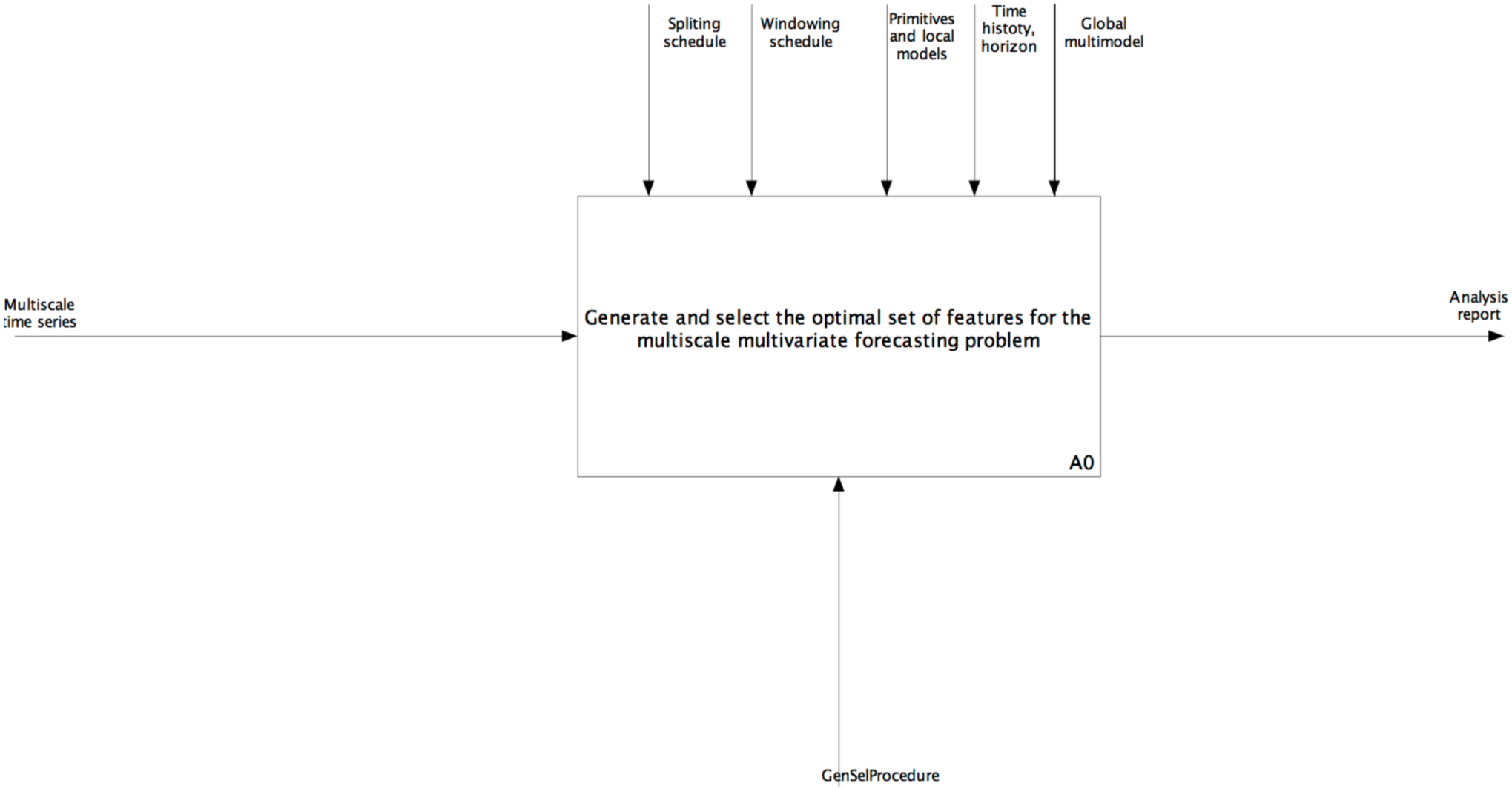
S is the criterion,

T is an optimization algorithm,

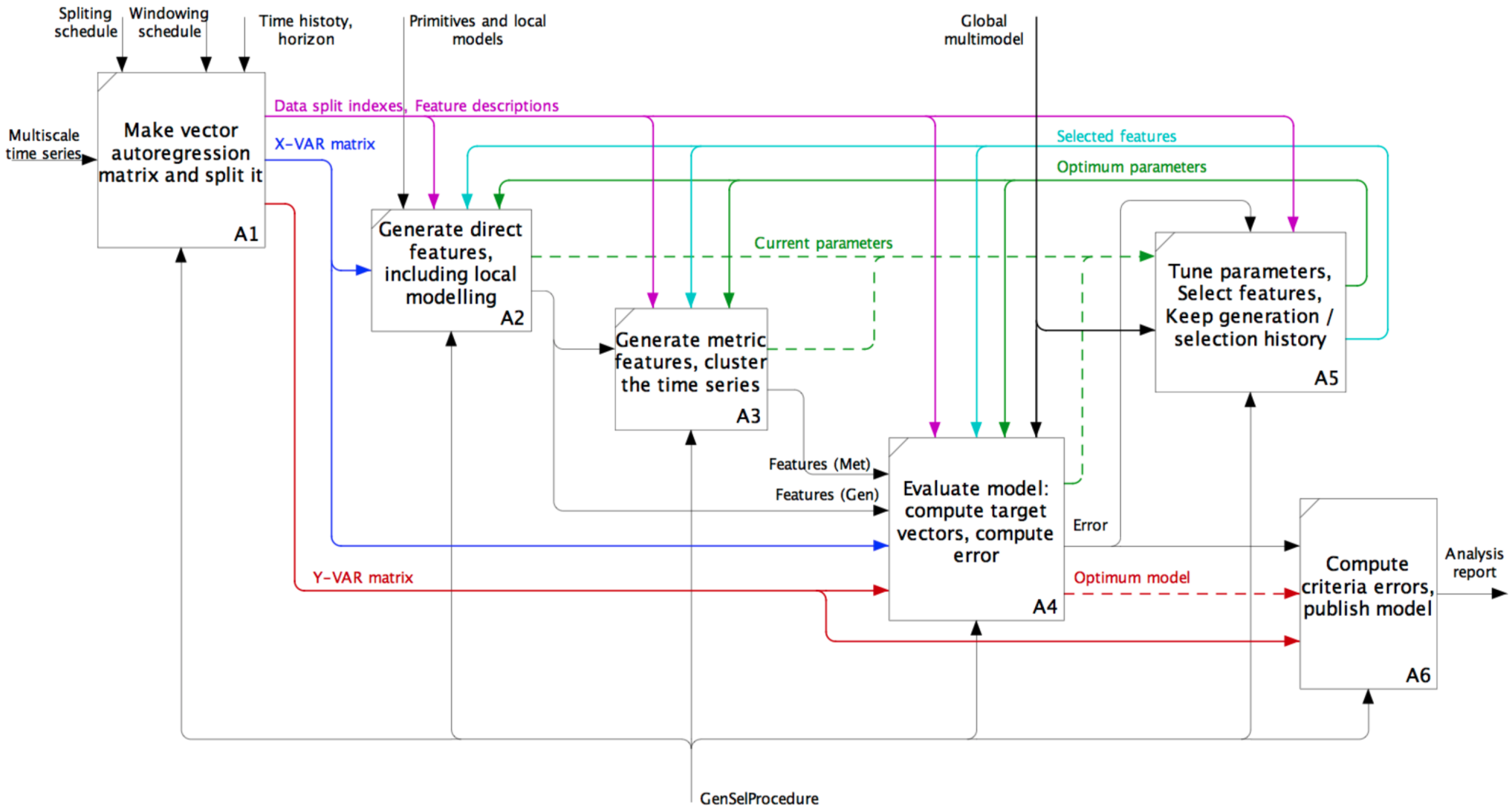
$\hat{\mathbf{w}}$ is some solution,

$$\hat{\mathbf{w}} = \arg\min S(\mathbf{w}|y, f).$$

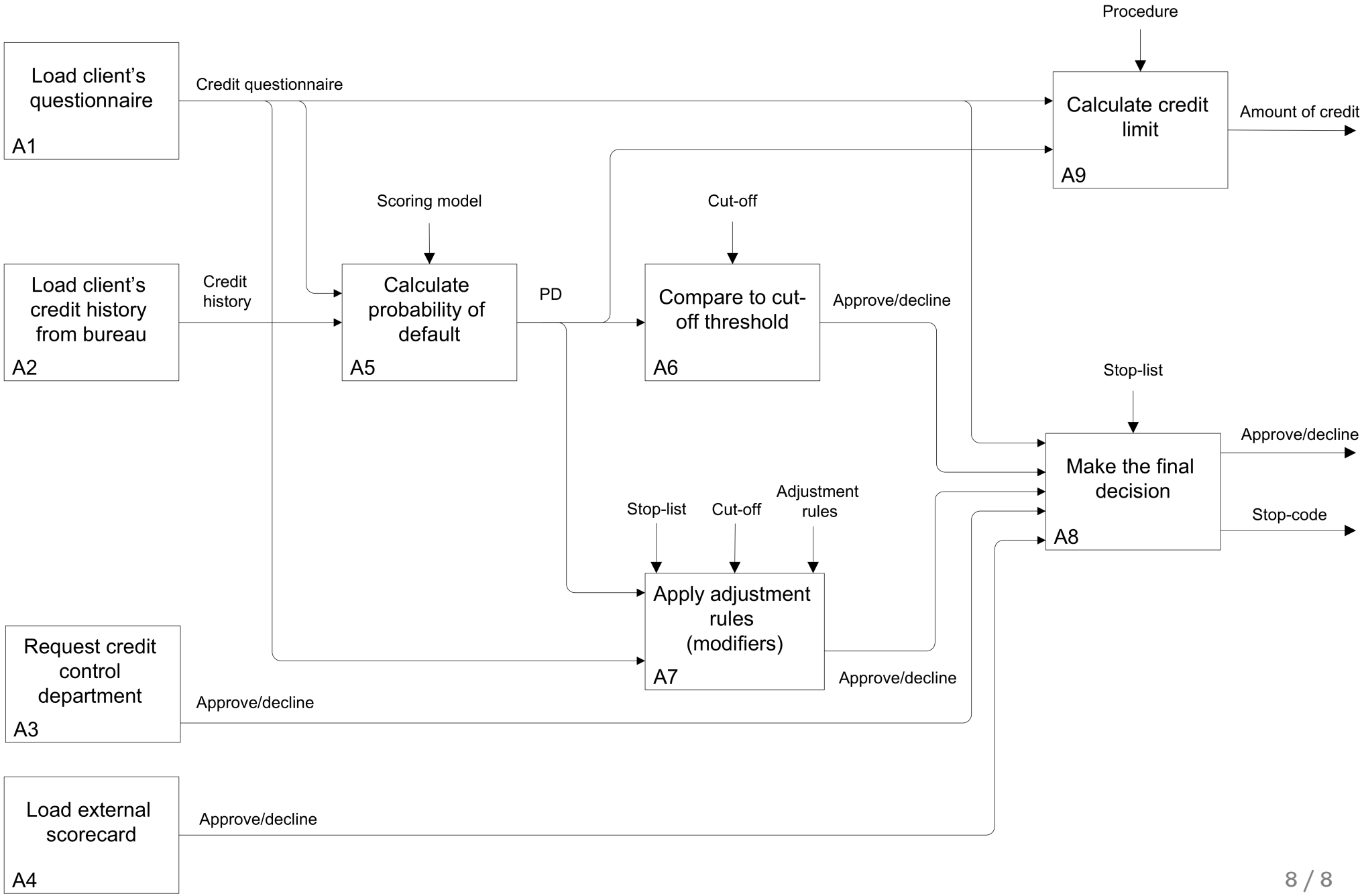
Порождение и выбор моделей, лист A-0



Порождение и выбор моделей, лист A0



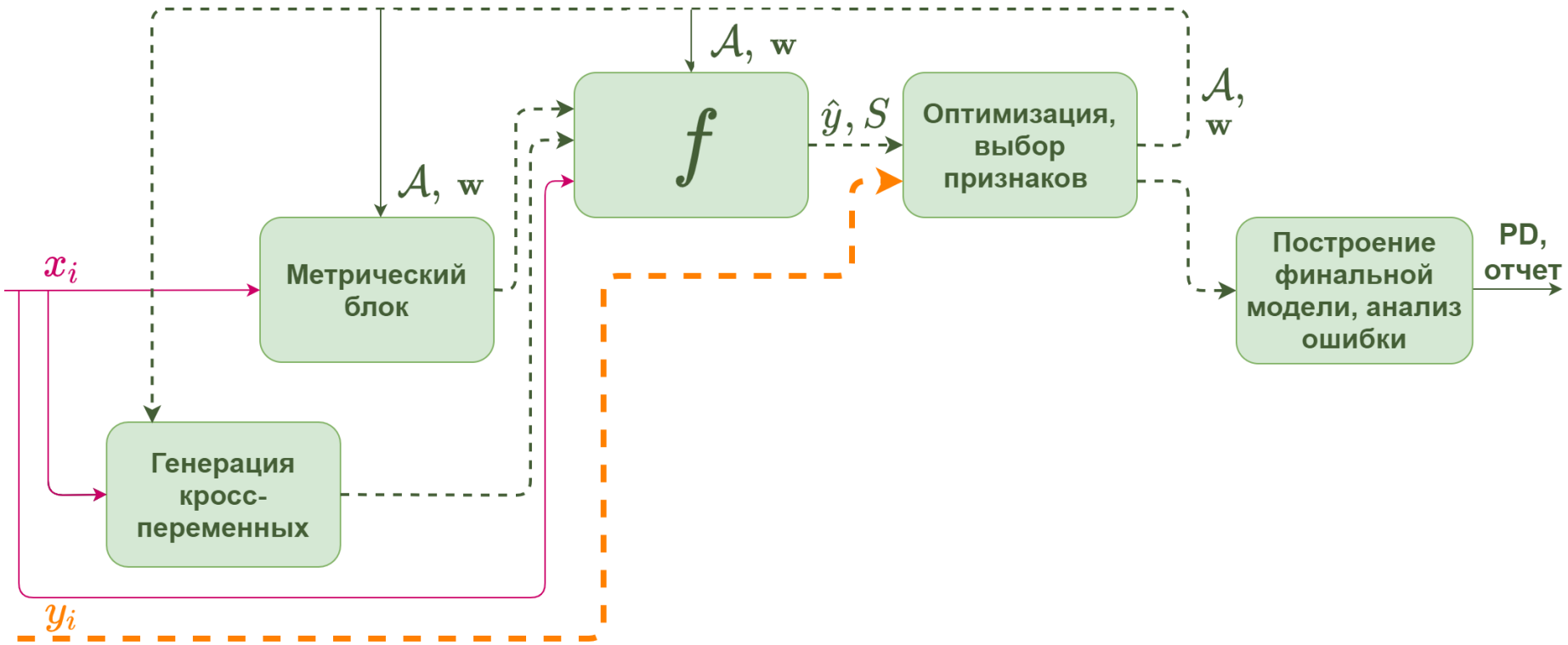
Функциональная схема скоринговой карты



Универсальная модель как суперпозиция локальных моделей

Идея. Новизна

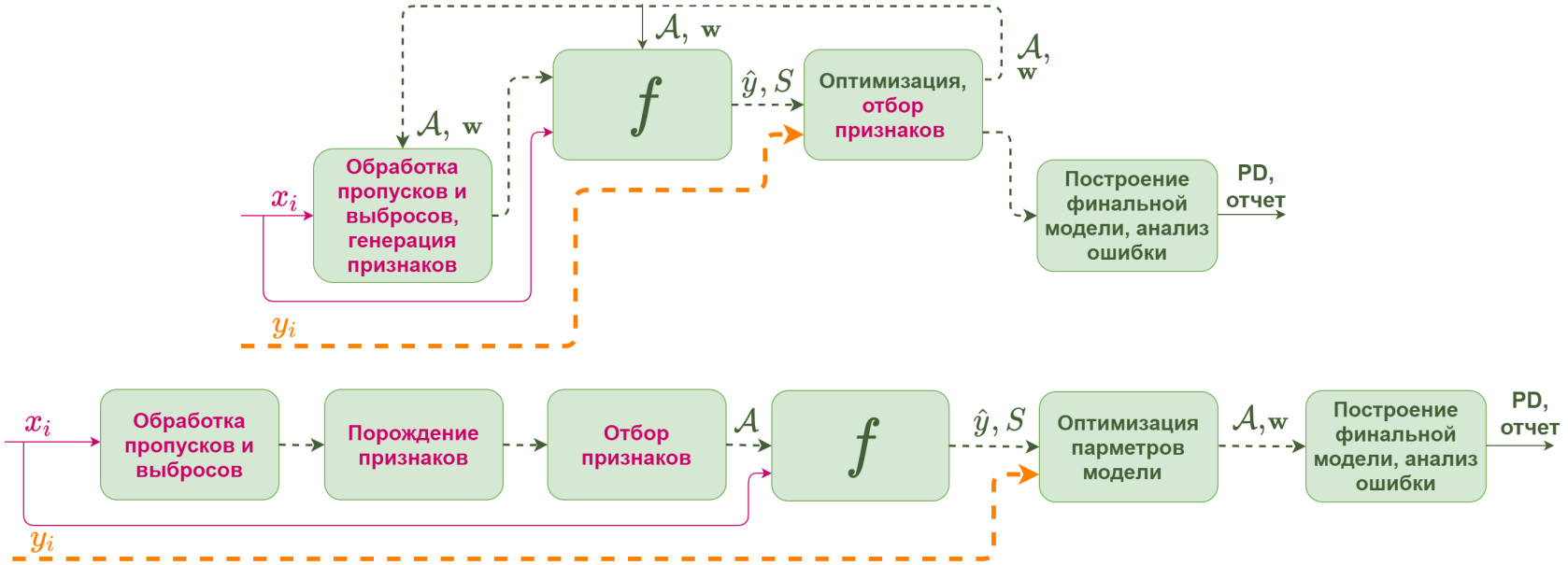
Система построена в единой оптимизационной цепочке с использованием модуля построения модели: процедуры обработки пропусков и выбросов, порождения и отбора признаков выполняются не на этапах предобработки, а в ходе оптимизации



Универсальная модель как суперпозиция локальных моделей

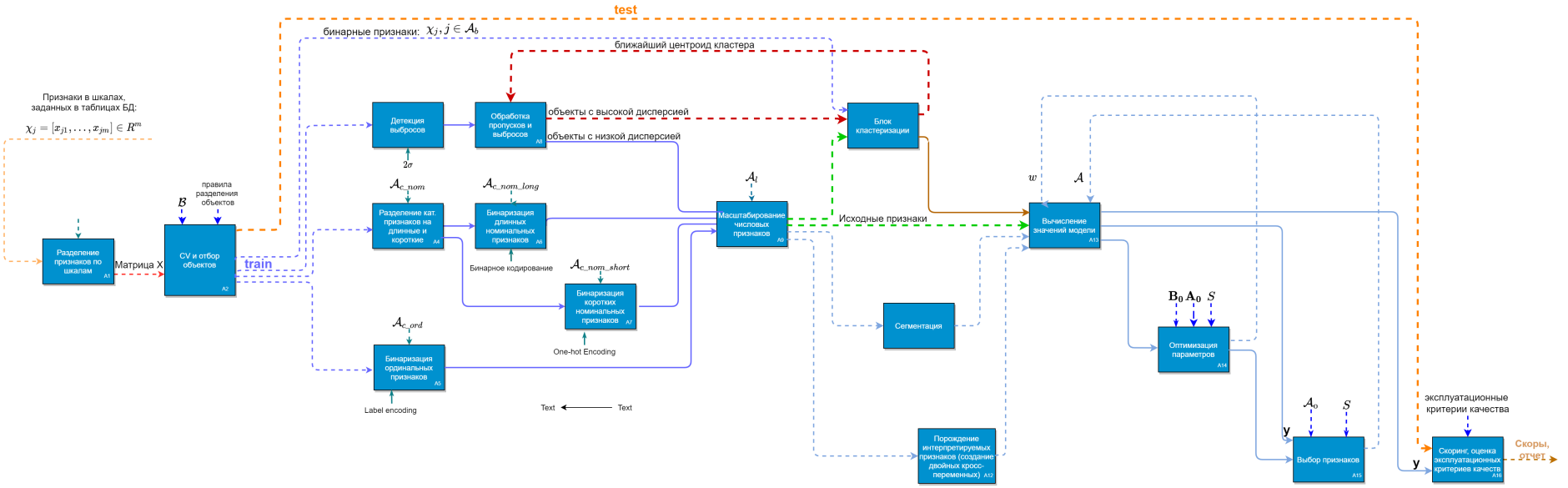
Уникальность модели как скоринговой системы

- Модель как суперпозиция локальных моделей
- Адаптивная, итеративная система (формирование $\mathcal{A}$ на каждой итерации обеспечивает адаптивность)
- Структура модели определяет способность системы накапливать информацию об опорных клиентах и информативных признаках
- Модель не требует разбиения данных на обучающую и контрольную выборку



Сравнение предложенной модели (сверху) и базового решения (снизу)

Универсальная модель как суперпозиция локальных моделей



Список модулей для построения модели

1. Семплирование выборки
2. Процедура скользящего контроля
3. Заполнение пропусков
4. Фильтрация выбросов
5. Бинаризация признаков
6. Сегментация признаков
7. Эмбеддинг
8. Порождение признаков, прямое
9. Порождение признаков, метрическое
10. Прогноз, вычисление значений прогностической модели
11. Оптимизация параметров модели
12. Выбор признаков
13. Анализ ошибки
14. Вычисление внешних критериев, отчет

План действий

1. Распределить блоки
2. Самостоятельно для своего блока (10 минут)
   a. Нарисовать схему в стандарте IDEF0
   b. Внутри блока написать действие
   c. Стрелки назвать и продублировать обозначениями
   d. Рядом кратко описать алгоритм действия
   e. Загрузить в http://bit.ly/m1p_file2discuss
3. В парах (10 минут)
   a. Согласовать общую схему для одного или нескольких блоков
   b. Нарисовать схему
   c. Загрузить
4. Общая схема (15 минут)
   a. Нарисовать общую схему на большом листе
   b. Загрузить
5. Обсуждение

Подсказка

Типы стрелок:

- независимая переменная
- целевая переменная
- индексы объектов
- индексы признаков
- ошибки
- прогнозы
- модели
- …

Вместе с порожденными данными (объектами и признаками) порождаются и их описания (индексы)