

Математическое моделирование, численные методы и комплексы программ

Судаков Владимир Анатольевич

vasudakov@fa.com

Литература

- Хемди Таха. Исследование операций
- Олег Ларичев. Теория и методы принятия решений, а также Хроника событий в Волшебных странах
- Губанов Д.А., Новиков Д.А., Чхартишвили А.Г. Социальные сети: модели информационного влияния, управления и противоборства
- Ричард Саттон, Эндрю Барто. Обучение с подкреплением
- Аллен Б. Дауни. Изучение сложных систем с помощью Python
- Джоэл Грас. Data Science. Наука о данных с нуля

Содержание

1. Модели исследования операций
2. Программные пакеты оптимизации
3. Модели клеточных автоматов
4. Дискретно-событийное моделирование
5. Машинное обучение с подкреплением
6. Системы поддержки принятия решений
7. Мультиагентное моделирование
8. Мультидисциплинарная оптимизация
9. Моделирование на базе фреймов

Математическое моделирование.

Определение

- **Математическая модель** — это приближённое описание какого-либо класса явлений внешнего мира, выраженное *математическими символами*
- **Математическое моделирование** — это опосредованное практическое или теоретическое исследование объекта, при котором непосредственно изучается не сам интересующий нас объект, а некоторая вспомогательная искусственная или естественная система (модель), находящаяся в некотором объективном соответствии с познаваемым объектом, способная замещать его в определённых отношениях и дающая при её исследовании, в конечном счёте, информацию о самом моделируемом объекте

Дискуссия

- Почему мы исследуем модели объектов, а не сами объекты?
-

Некоторые причины для моделирования

- Натурные испытания – это дорого/долго
- *Суть моделирования в абстрагировании/избирательности*
 - Абстрагирование – это одни из способов борьбы со сложностями

Опасности математического моделирования

Ферми сказал, что в теоретической физике есть лишь два подхода к вычислениям:

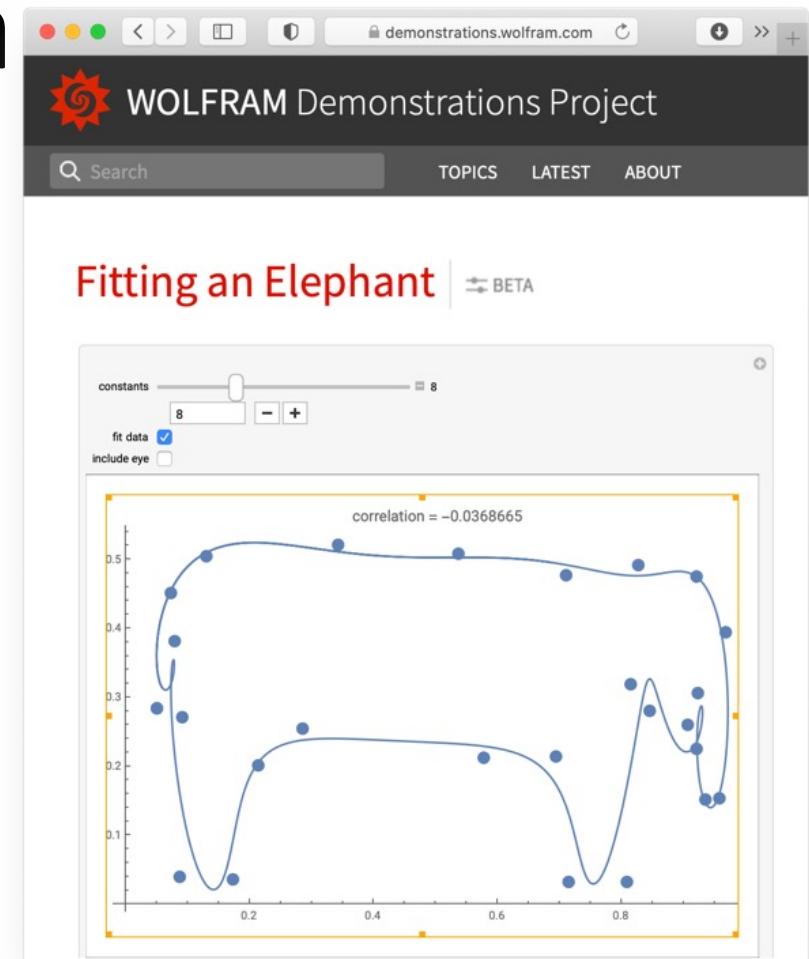
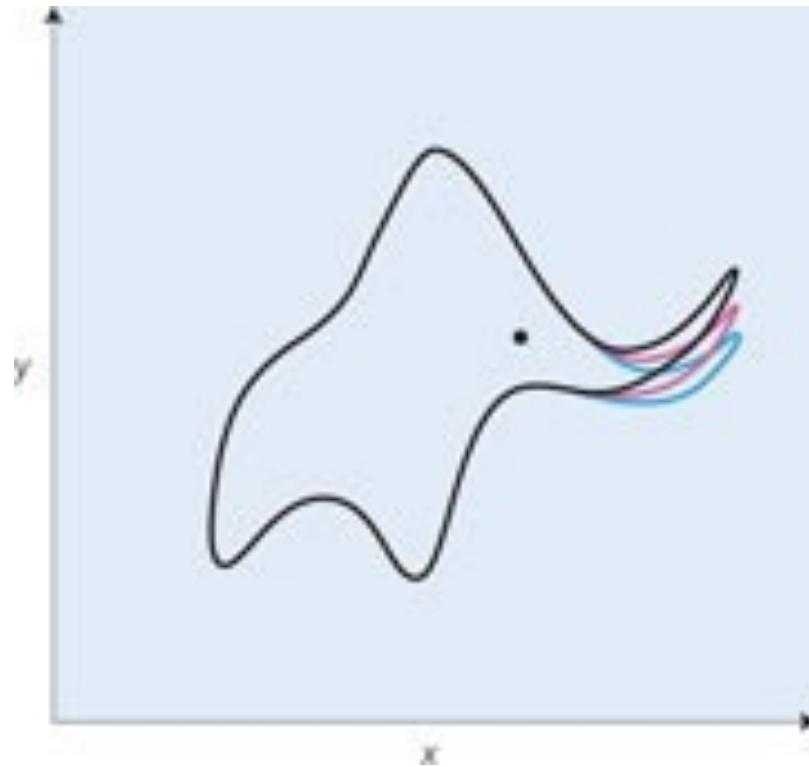
- понимание физической природы процесса
- или
- наличие точного математического формализма,

и работа Дайсона не идёт ни по одному из этих путей.

Когда обескураженный Дайсон спросил Ферми, почему тому не кажется убедительным совпадение результатов вычислений и эксперимента, Ферми указал на наличие произвольных параметров в модели Дайсона и отметил:

мой друг Джонни фон Нейман говорил, что с четырьмя параметрами он может описать слона, а с пятым — заставить его махать хоботом

Слон фон Неймана



- Jürgen Mayer, Khaled Khairy and Jonathon Howard. [Drawing an elephant with four complex parameters](#). // Am. J. Phys. 78, 648 (2010).

Классификация моделей

- Линейные или нелинейные модели
- Дискретные или непрерывные
- Детерминированные или стохастические или нечеткие или игровые
- Статические или динамические
- Структурные или функциональные модели

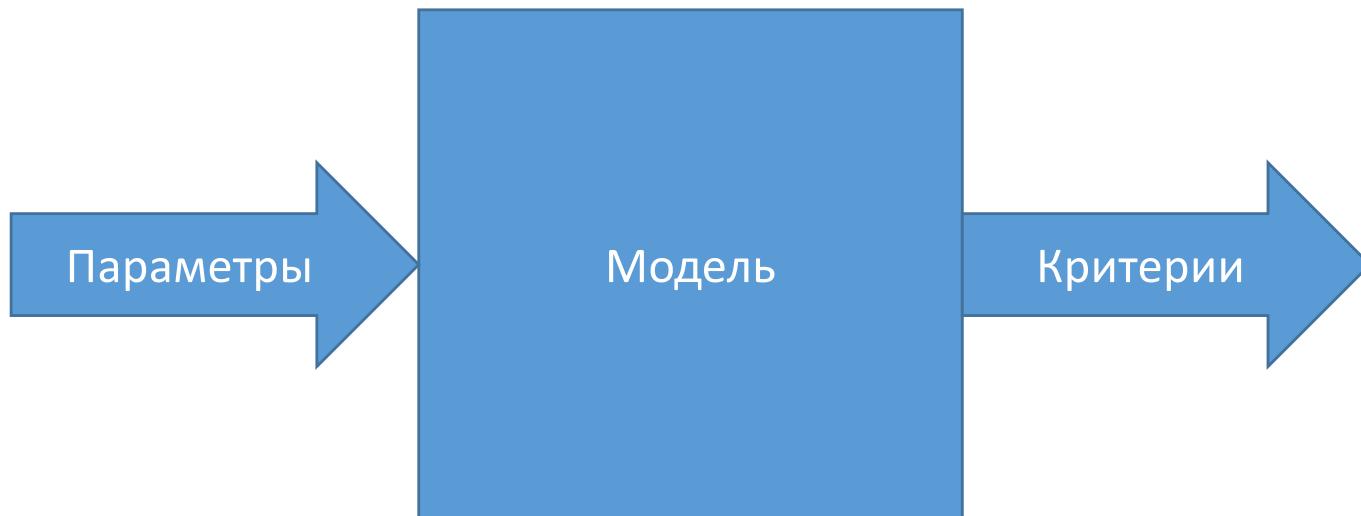
Классификация задач

- **Прямая задача:** структура модели и все её параметры считаются известными, необходимо провести исследование модели для извлечения полезного знания об объекте.
- **Обратная задача:** известно множество возможных моделей, надо выбрать конкретную модель (параметры модели) на основании дополнительных данных об объекте.
- Если модель (или параметры) выбираются под требованиях к объекту, то это *задача проектирования*.

Этапы построения модели для решения практических задач

1. Предпроектное обследование (мониторинг ситуации)
 - Как же сейчас обходятся без модели или какие модели используют?
2. Определение целей моделирования
3. Определение критериев оценки результатов моделирования
4. Определение параметров модели (важен компромисс между универсальностью и сложностью)
5. Математическая формализация (на этом этапе определяется класс модели)
6. Выбор метода решения задачи
7. Разработка алгоритмов моделирования (опционально)
8. Выбор инструментальных средств или языков моделирования
9. Проведение вычислительных экспериментов
10. Анализ результатов
11. При необходимости возврат на предыдущие шаги

Модель



Дискуссия

- Но можно ли описать вертолет одной моделью?
- А есть ли польза от одной модели?
- Что такая система?

Анализ систем

Система –

- множество элементов и
- отношений между ними,
- объединенных для достижения цели

Моделей много, их нужно
объединять в систему

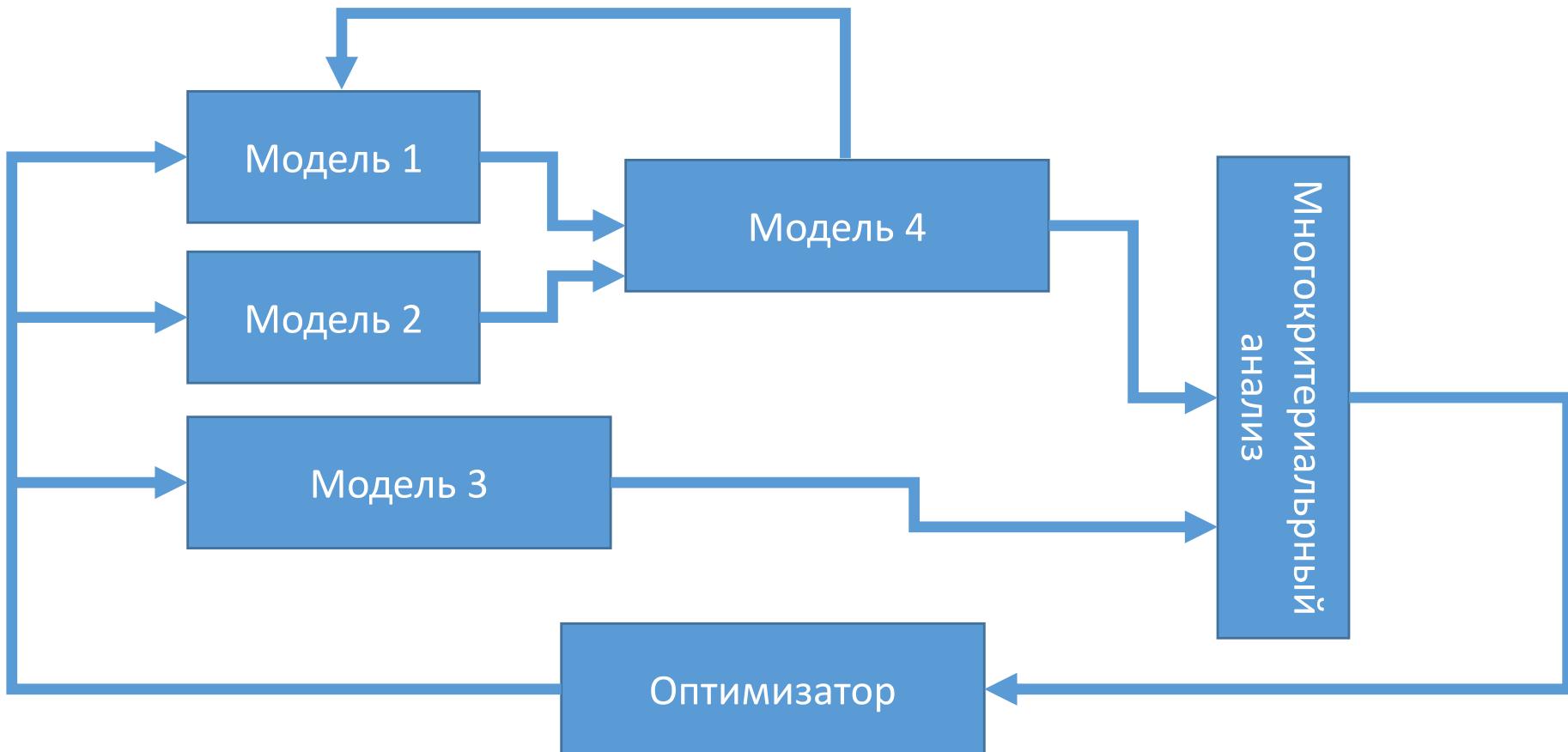
В качестве элементов могут быть другие системы

- Система взаимодействует со средой как единое целое

Свойства:

- Интегративность (ограниченность от среды)
- Синергичность
- Эмерджентность
- Ингерентность

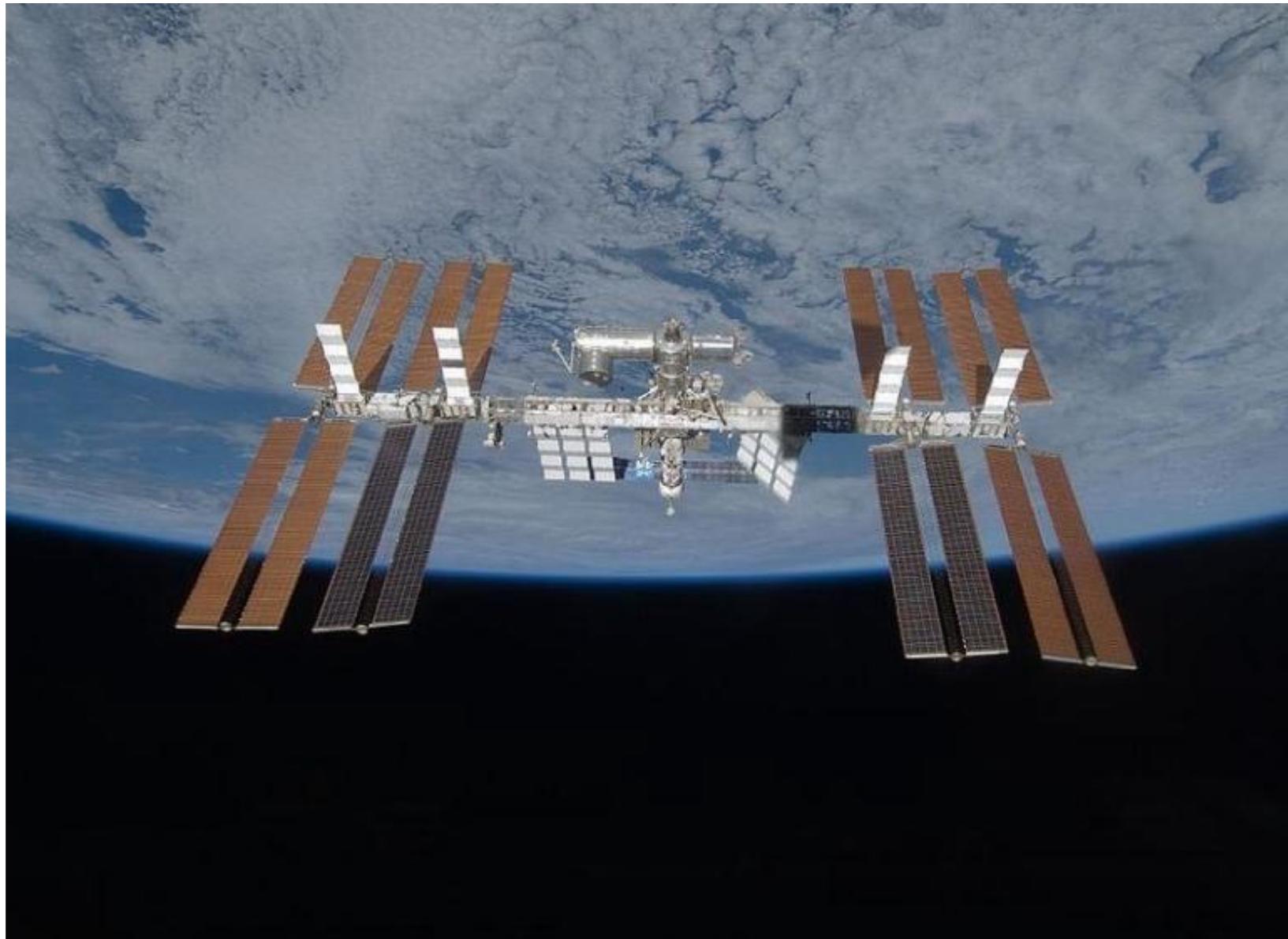
Использование моделей



Модели исследования операций

- *Операция* — всякое мероприятие (система действий), объединённое единым замыслом и направленное к достижению какой-то цели
- Решение — всякий определённый набор зависящих от человека значений параметров
- Оптимальное — решение, которое по тем или другим признакам предпочтительнее других
- Цель исследования операций — предварительное количественное обоснование оптимальных решений с опорой на показатель эффективности

Планирование космических экспериментов (КЭ) на РС МКС



Редактор критериев

The screenshot shows a web-based application titled 'Редактор критериев' (Criteria Editor) running in a browser window. The URL is knts.tsniimash.ru/DSS/Criteria.aspx. The interface includes a sidebar with user information ('Пользователь tester'), navigation links ('Настройки', 'Выход'), and a main content area.

Дерево критериев (Criterion Tree):

- Оценка КЭ
 - Значимость эксперимента
 - Актуальность КЭ
 - Значимость результатов КЭ
 - Научный эффект КЭ
 - Прикладной эффект КЭ
 - Безопасность
 - Готовность КЭ к включению программы
 - Доступность информации по КЭ
 - Космическая деятельность
 - КЭ является международным
 - Наличие заинтересованного заказчика
 - Наличие КЭ в подпрограмме исследований
 - Соответствие КЭ направлениям других программ
 - Соответствие КЭ направлениям ФКП
 - Социальный эффект
 - Экологический эффект
 - Экономический эффект
 - Технологическая инновационность
 - Использование существующей техники
 - Наличие международного патента

Параметры (Parameters):

Наименование	Описание
Актуальность КЭ	

Родитель: Значимость эксперимента

Шкала (Scale):

№	Градация	Ранг	Удалить
1	Оценка невозможна	1	X
2	Эксперимент в принципе полезен, однако его задержка не повлияет на темп исследований в данной области	2	X
3	Задержка проведения КЭ затормозит, но не остановит дальнейшие исследования в данной области.	3	X
4	Задержка проведения КЭ приведет к соответствующей остановке исследования в данной области.	4	X
5	Срыв проведения эксперимента в заданные сроки приведет к принципиальной невозможности в течение многих лет получить требуемую информацию (например, вследствие изменения активности Солнца, или если аппаратура входит в состав МЛМ и, таким образом, сроки её	5	X

Направление улучшений: Чем больше, тем лучше

Числовой показатель:

Порядковый номер:

Ресурсные ограничения

Редактор ресурсов

192.168.56.101:8080/DSS/Resource.aspx

Редактор ресурсов

Задача: Оценка КЭ
Пользователь: Администратор

Настройки Выход

- Система поддержки принятия решений
- Задачи
- Альтернативы
- Критерии
- Ресурсы
- Экспертные оценки
- Ранжирование
- Планирование
- Пользователи

Ресурсы

Наименование	Описание	Значение	Период	Единица изменения периода	Критерий	Удалить
Масса доставляемой на борт НА (кг)	Масса доставляемой на борт НА (кг)	0	2	квартал	Масса доставляемой на борт НА (кг)	<input type="checkbox"/>
Объём доставляемой на борт НА (м³)	Объём доставляемой на борт НА (м ³)	0	2	квартал	Объём доставляемой на борт НА (м ³)	<input type="checkbox"/>
Масса возвращаемых блоков (кг)	Масса возвращаемых блоков (кг)	0	2	квартал	Масса возвращаемых блоков (кг)	<input type="checkbox"/>
Объём возвращаемых блоков (м³)	Объём возвращаемых блоков (м ³)	0	2	квартал	Объём возвращаемых блоков (м ³)	<input type="checkbox"/>
Мощность НА (кВт)	Потребляемая мощность НА (кВт)	0	2	квартал	Мощность НА (кВт)	<input type="checkbox"/>
Энергопотребление НА (кВт·ч)	Суммарное энергопотребление НА (кВт·ч)	0	2	квартал	Энергопотребление НА (кВт·ч)	<input type="checkbox"/>
Время реализации сеансов КЭ (мин)	Суммарное время реализации сеансов КЭ (мин)	0	2	квартал	Время реализации сеансов КЭ (мин)	<input type="checkbox"/>
Требуемое рабочее время экипажа (мин)	В том числе требуемое рабочее время экипажа на проведение КЭ (мин)	0	2	квартал	Требуемое рабочее время экипажа (мин)	<input type="checkbox"/>
Объём передаваемой информации (Гбайт)	Необходимый за время проведения КЭ объём передаваемой информации по радиолинии связи в (Гбайт)	0	1	день	Объём передаваемой информации (Гбайт)	<input type="checkbox"/>

Формализация задачи планирования КЭ

- Целевая функция

$$\max_{x_j, t_j (j \in J)} \left\{ \sum_{j=1}^J x_j p_j; \sum_{j=1}^J x_j p_j (T - t_j) / T \right\}$$

- Ограничения

$$\sum_{j=1}^J x_j \left\{ \begin{array}{c} \sum_{t=(n-1)\Delta+1}^{n\Delta} f_j(t - t_j) \\ \tau_j \end{array} \right\} q_{jk} \leq Q_k, \quad (k=1,2,\dots,K; n=1,2,\dots,N),$$

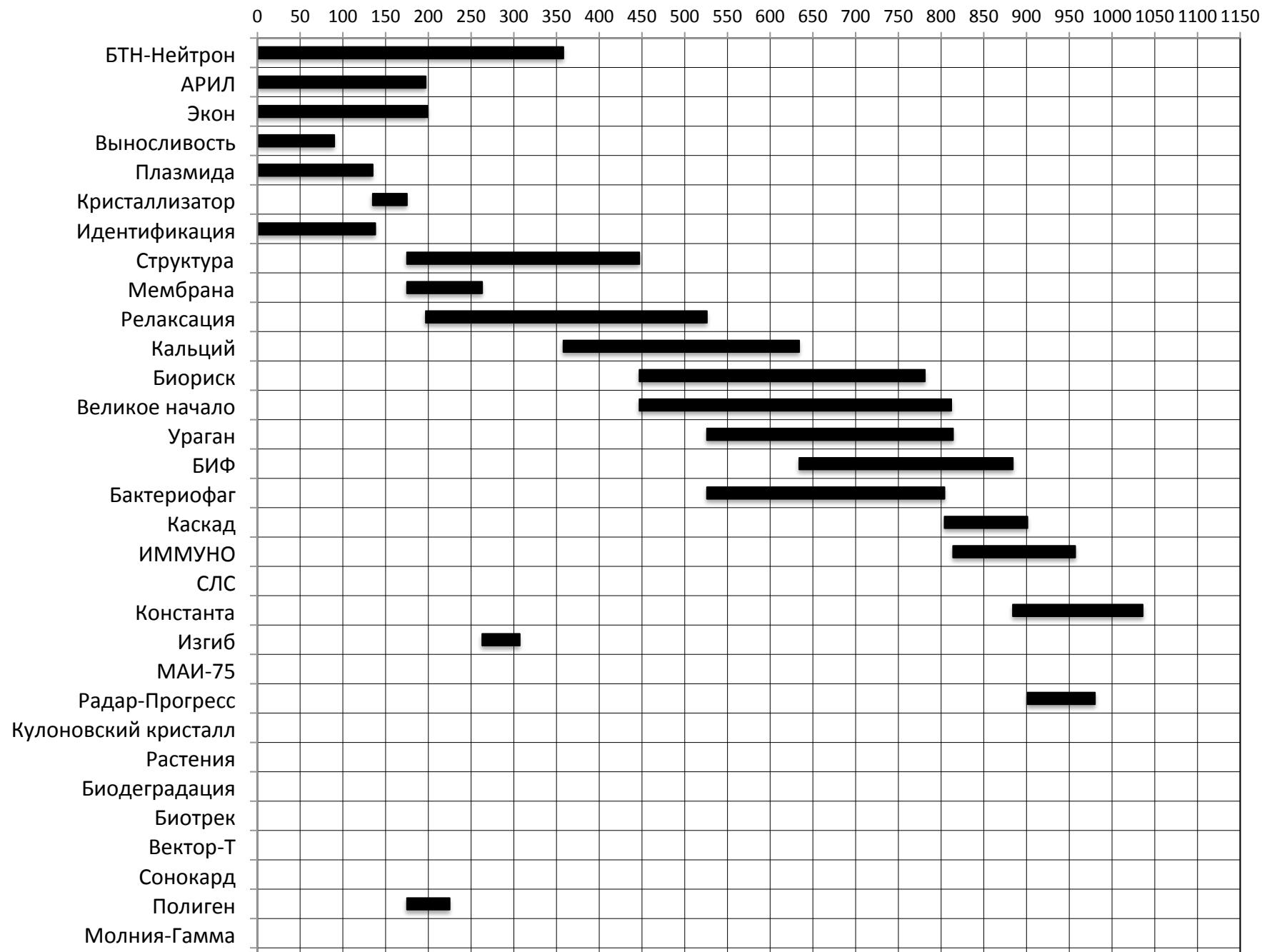
$$\sum_{j=1}^J x_j f_j(t - t_j) v_{jl} \leq V_l, \quad (l=1,2,\dots,L; t=1,2,\dots,T),$$

$$t_j \geq \max_s \{t_{ps[j,s]} + \tau_{ps[j,s]}\}, \quad (j=1, 2, \dots, J),$$

$$x_j \leq t_j \leq T x_j - \tau_j, \quad (j=1, 2, \dots, J),$$

$$x_j = 0 \text{ или } 1, \quad (j=1, 2, \dots, J).$$

Диаграмма Ганта



Оптимальное размещение грузов на борту воздушных судов

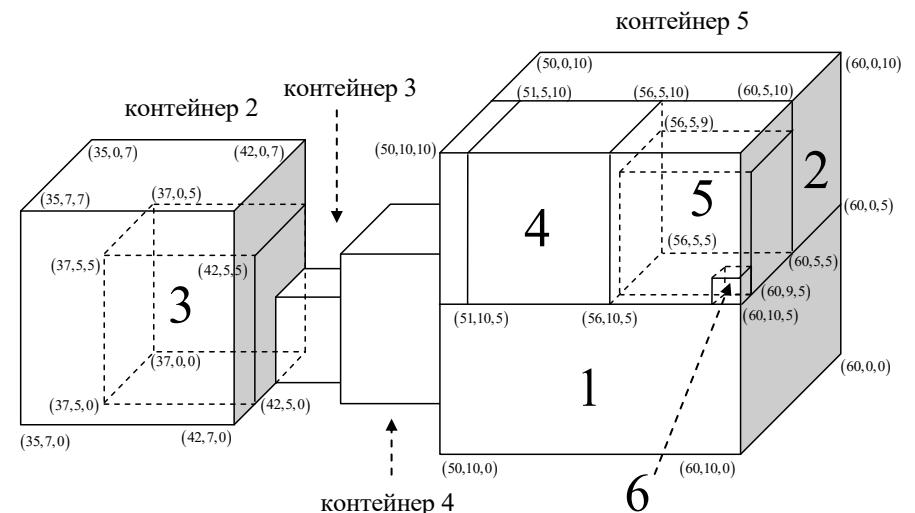
$$\max \alpha \left(\sum_{i=1}^n x_i^1 \right) + \beta \left(\sum_{i=1}^n \sum_{j=2}^g m_i r_{ij} \right)$$

$$(-1)^e x_v^{\left\lfloor \frac{e}{2} \right\rfloor} + (-1)^{e+1} x_i^{\left\lfloor \frac{e}{2} \right\rfloor} + \sum_{d=1}^D (-1)^{f(e,d)+1} M a_{vi}^d$$

$$\leq -l_{u(e,v,i)}^{\left\lfloor \frac{e}{2} \right\rfloor} + \sum_{d=1}^D f(e, d) M$$

$$\sum_{v=1}^n (m_v q_{vij}^d) - m_i p_{ij}^d - \frac{1}{2} m_i l_i^d r_{ij} = 0$$

$$\sum_{i=1}^n \prod_{d=1}^D l_i^d r_{ij} \leq \prod_{d=1}^D L_j^d .$$



Примеры

- Задача о формировании производственного плана предприятия
- Задача целераспределения средств ПВО по объектам нападения
- Задача выбора оптимального маршрута на транспортной сети.
- Задача оптимального управления многоэтапной программой работ

Общая постановка

$$\min_{x \in D} (\max) \{z = f(x)\} \quad D = \{x \in R^n : g_i(x) \leq (=, \geq) b[i], i = \overline{1, m}\}.$$

z – целевая функция,

$x = (x[1], x[2], \dots, x[n])^T$ -вектор оптимизационных переменных,

\min или \max – направление оптимизации,

$\min_{x \in D} \{z = f(x)\}$ - критерий оптимизации,

D – множество допустимых решений оптимизационной задачи,

$x \in R^n$ - задание типа пространства, на котором определены оптимизационные переменные (

$x \in Z^n, x \in N^n$),

$g_i(x) \leq b[i]$ - ограничение оптимизационной задачи,

$g_i(x)$ - левая часть ограничения,

$b[i]$ - правая часть ограничения.

Классификация

- Задачи линейного программирования
- Задачи дискретного программирования
- Задачи смешанного линейно-целочисленного программирования
- Задачи динамического программирования
- Задачи нелинейного программирования
- Задачи оптимального управления
- Задачи стохастического программирования

Производственная задача

Рассматривается некоторая производственная система, способная производить несколько видов продукции. Для производства используется ряд сырьевых ресурсов, имеющихся в системе в ограниченном количестве. От реализации произведенной продукции система получает прибыль. Требуется так составить производственный план (определить, какие виды продукции и в каком количестве производить), чтобы при имеющихся ограничениях на сырьевые ресурсы получить максимальную прибыль.

Формализованная постановка

n - количество видов продукции, которую может производить система;

m - количество видов сырья, используемого при производстве продукции;

$c[j]$, ($j = 1, 2, \dots, n$) - прибыль, получаемая от реализации произведенной единицы j -ого вида продукции;

$b[i]$, ($i = \overline{1, m}$) - количество имеющегося в наличии сырья i -ого вида;

$a[i, j]$, ($i = \overline{1, m}; j = \overline{1, n}$) - технологический коэффициент затрат i -ого вида сырья на производство единицы продукции j -ого вида;

$x[j]$, ($j = \overline{1, n}$) - планируемого количества производимой продукции j -ого вида (оптимизационная переменная).

С учетом введенных обозначений формализованная запись имеет вид:

$$\max_x z = \sum_{j=1, n} c[j] x[j],$$

$$\sum_{j=1}^n a[i, j] x[j] \leq b[i], \quad (i = \overline{1, m}), \quad (1.1)$$

$$x[j] \geq 0, \quad (j = \overline{1, n}).$$

Здесь z - суммарная прибыль от реализации произведенной продукции, $\sum_{j=1}^n a[i, j] x[j]$ - затраты i -ого вида сырьевого ресурса на реализацию всего производственного плана.

Задача о поднятии плит

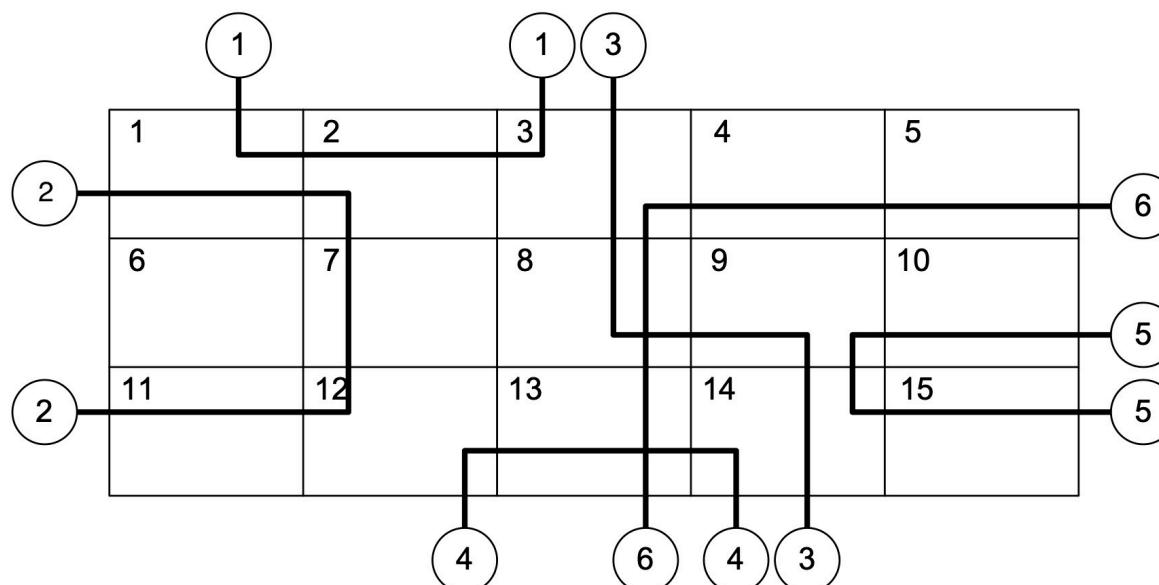
Необходимо проверить состояние кабельных линий, которые находятся под полом, состоящим из плит. Для проверки каждого кабеля достаточно получить к нему доступ в любом месте, для чего нужно поднять соответствующую плиту. На поле расположено оборудование. Поэтому с поднятием каждой плиты пола связан определенный объем работ по демонтажу и перемещению оборудования, задаваемый в человеко-часах. Необходимо определить плиты, которые нужно поднять таким образом, чтобы обеспечить доступ ко всем кабелям, а суммарный объем работ, связанный с поднятием плит фальшпола, был бы минимальным.

Вход: стоимость поднятия плит, какие кабели под какими плитами лежат.

Выход: Какие плиты поднимаем.

Программа должна работать с произвольными корректными данными.

Пример исходных данных:



Трудноразрешимые задачи, NP-полнота

Функция временной сложности	Размер n					
	10	20	30	40	50	60
n	0,00001 сек	0,00002 сек	0,00003 сек	0,00004 сек	0,00005 сек	0,00006 сек
n^2	0,0001 сек	0,0004 сек	0,0009 сек	0,0016 сек	0,0025 сек	0,0036 сек
n^3	0,001 сек	0,008 сек	0,027 сек	0,064 сек	0,125 сек	0,216 сек
n^5	0,1 сек	3,2 сек	24,3 сек	1,7 мин	5,2 мин	13,0 мин
2^n	0,001 сек	1,0 сек	17,9 мин	12,7 дней	35,7 лет	366 столетий
3^n	0,059 сек	58 мин	6,5 лет	3855 столетий	2×10^8 столетий	$1,3 \times 10^{13}$ столетий

Влияние быстродействия ЭВМ

*Размеры наибольшей задачи,
разрешимой за один час*

<i>Функция временной сложности</i>	<i>На современных ЭВМ</i>	<i>На ЭВМ, в 100 раз более быстрых</i>	<i>На ЭВМ, в 1000 раз более быстрых</i>
n	N_1	$100 N_1$	$1000 N_1$
n^2	N_2	$10 N_2$	$31,6 N_2$
n^3	N_3	$4,64 N_3$	$10 N_3$
n^5	N_4	$2,5 N_4$	$3,98 N_4$
2^n	N_5	$N_5 + 6,64$	$N_5 + 9,97$
3^n	N_6	$N_6 + 4,19$	$N_6 + 6,29$

Экспоненциальная сложность на практике

- Некоторые задачи имеют экспоненциальную сложность, но это сложность для *наихудшего* случая
- Во многих практических задачах эта сложность не проявляется
- Проблема – нет способа спрогнозировать как поведет себя экспоненциальный алгоритм на тех или иных данных

Где решать задачи

Пакеты

- SCIP
- IBM ILOG
- GAMS
- FRODO
- GUROBI

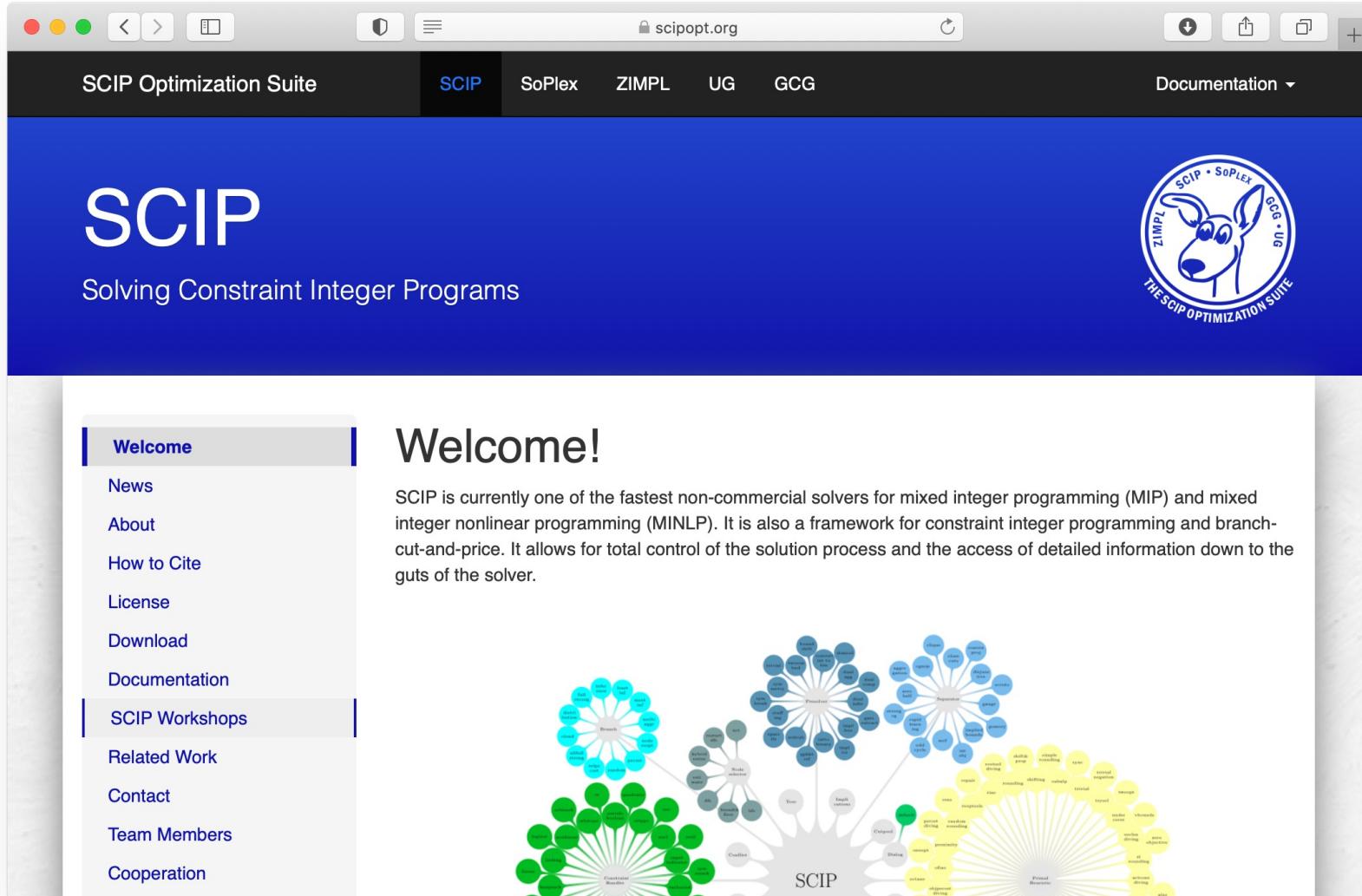
Языки/Форматы

- AMPL
- MPS

Современные средства решения задач оптимизации

- AMPL - A Modeling Language for Mathematical Programming — язык моделирования для математического программирования
- GAMS - General Algebraic Modeling System – система моделирования для математического программирования и оптимизации
- JULIA - высокоуровневый высокопроизводительный свободный язык программирования с динамической типизацией, созданный для математических вычислений.
- ILOG CPLEX - это решение, предназначенное для быстрой разработки и развертывания моделей математического программирования и программирования в ограничениях. Сочетает в себе полнофункциональную интегрированную среду разработки с поддержкой языка OPL и высокопроизводительные модули решений CPLEX и CP Optimizer.
- SCIP - один из самых быстрых некоммерческих решателей для смешанного целочисленного программирования (MIP) и смешанного целочисленного нелинейного программирования (MINLP).

SCIP



The screenshot shows the SCIP Optimization Suite website. The top navigation bar includes links for "SCIP Optimization Suite", "SCIP", "SoPlex", "ZIMPL", "UG", "GCG", and "Documentation". The main title "SCIP" is prominently displayed, followed by the subtitle "Solving Constraint Integer Programs". A logo featuring a cartoon dog is on the right. The left sidebar has a "Welcome" tab selected, with other options like "News", "About", "How to Cite", "License", "Download", "Documentation", "SCIP Workshops" (which is highlighted), "Related Work", "Contact", "Team Members", and "Cooperation". The main content area features a large "Welcome!" heading and a paragraph about SCIP's capabilities. Below this is a complex network diagram illustrating the solver's internal structure and components.

SCIP

Solving Constraint Integer Programs

SCIP • SoPlex • ZIMPL • GCG • UG

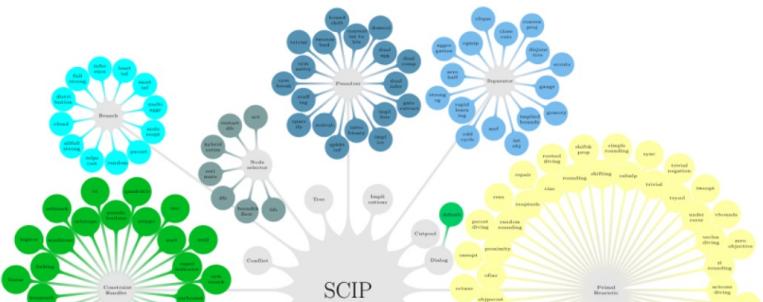
THE SCIP OPTIMIZATION SUITE

Welcome

- News
- About
- How to Cite
- License
- Download
- Documentation
- SCIP Workshops
- Related Work
- Contact
- Team Members
- Cooperation

Welcome!

SCIP is currently one of the fastest non-commercial solvers for mixed integer programming (MIP) and mixed integer nonlinear programming (MINLP). It is also a framework for constraint integer programming and branch-cut-and-price. It allows for total control of the solution process and the access of detailed information down to the guts of the solver.



Решение оптимизационных задач в ws-dss



```
{  
    "trace": 1,  
    "kn": 1,  
    "kria": 1,  
    "kpp": 1,  
    "dir": 1,  
    "c": [-3,-7, 1,-1],  
    "h": [1, 1, 1, 1],  
    "asb": [  
        [2, -1, 1, -1, 2, 1],  
        [-1, 1, -6, -4, 0, -8],  
        [-5, -3, 0, -1, 0, -5]  
    ]  
}
```

Выходные данные:

Метод неявного перебора по векторной решетке

SVL Версия 2.0

КН: ВКЛЮЧЕН

КПИА: ВКЛЮЧЕН

КПП: ВКЛЮЧЕН

Данные по задаче в исходной форме

Направление оптимизации: MAX

Коэффициенты целевой функции: -3 -7 1 -1

Ограничения на опт. переменные сверху: 1 1 1 1

Ограничения:

```
2      -1      1      -1 >= 1  
-1      1      -6      -4 <= -8  
-5      -3      0      -1 <= -5
```

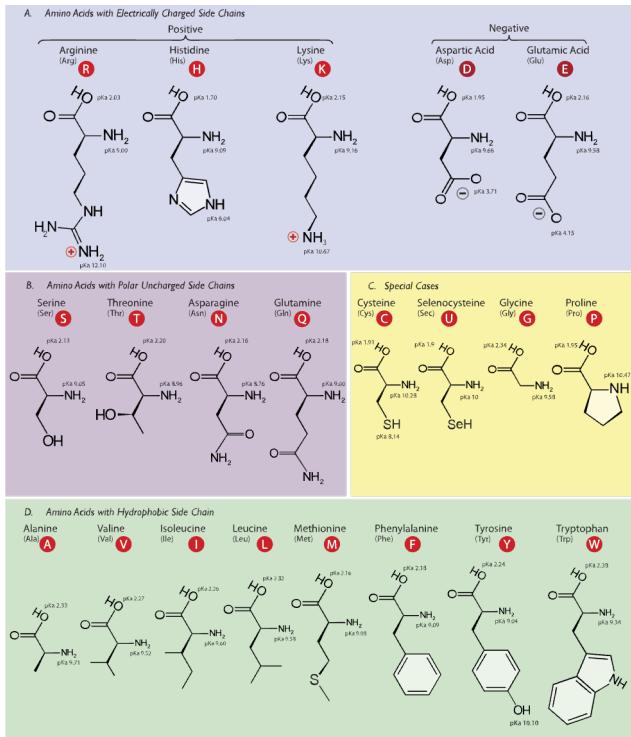
Время начала работы: 11/04/2016 22:10:35

Преобразованная задача:

Направление оптимизации: MIN

Коэффициенты целевой функции: 3 7 1 1

Постановка задачи оптимизации белковых компонент пищевых продуктов



Дано: u_j - масса вещества j -го типа - оптимационная переменная (в 100 г.)

$$j = 0..2$$

0 - черный (КМБ)

1 - красный (КСБ)

2 - зеленый (ПШ)

$$u_j \geq 0$$

d_i - масса аминокислоты i -го вида в идеальном случае (в мг.)

$$i = 0..7$$

a_{ij} - сколько мг i -й аминокислоты содержится в 100 г j -го вещества

$$\min_{\{u_j: j \in J\}} \left(\sum_{i=0}^7 |d_i - \sum_{j \in J} (a_{ij} u_j)| \right)$$

Сведение к каноническому виду

Целевая функция:

$$\min_{\{u_j: j \in J\}} \left(\sum_{i=0}^7 |d_i - \sum_{j \in J} (a_{ij} u_j)| \right)$$

Переход к линейной задаче

$$\begin{aligned} y_i^+ - y_i^- &= d_i - \sum_{j \in J} (a_{ij} u_j) \\ y_i^+ - y_i^- + \sum_{j \in J} (a_{ij} u_j) &= d_i \\ y_i^+ \geq 0, y_i^- \geq 0 \end{aligned}$$

Общий вид задачи

$$\min_{y_i^+, y_i^-, \{u_j: j \in J\}} \left(\sum_{i=0}^7 (y_i^+ + y_i^-) \right)$$

Ограничения:

$$\begin{aligned} y_i^+ - y_i^- + \sum_{j \in J} (a_{ij} u_j) &= d_i, i = 0..7 \\ y_i^+ \geq 0, y_i^- \geq 0, i &= 0..7 \end{aligned}$$

$$u_j \geq 0, j \in J.$$

Оптимизация планирования экипажей

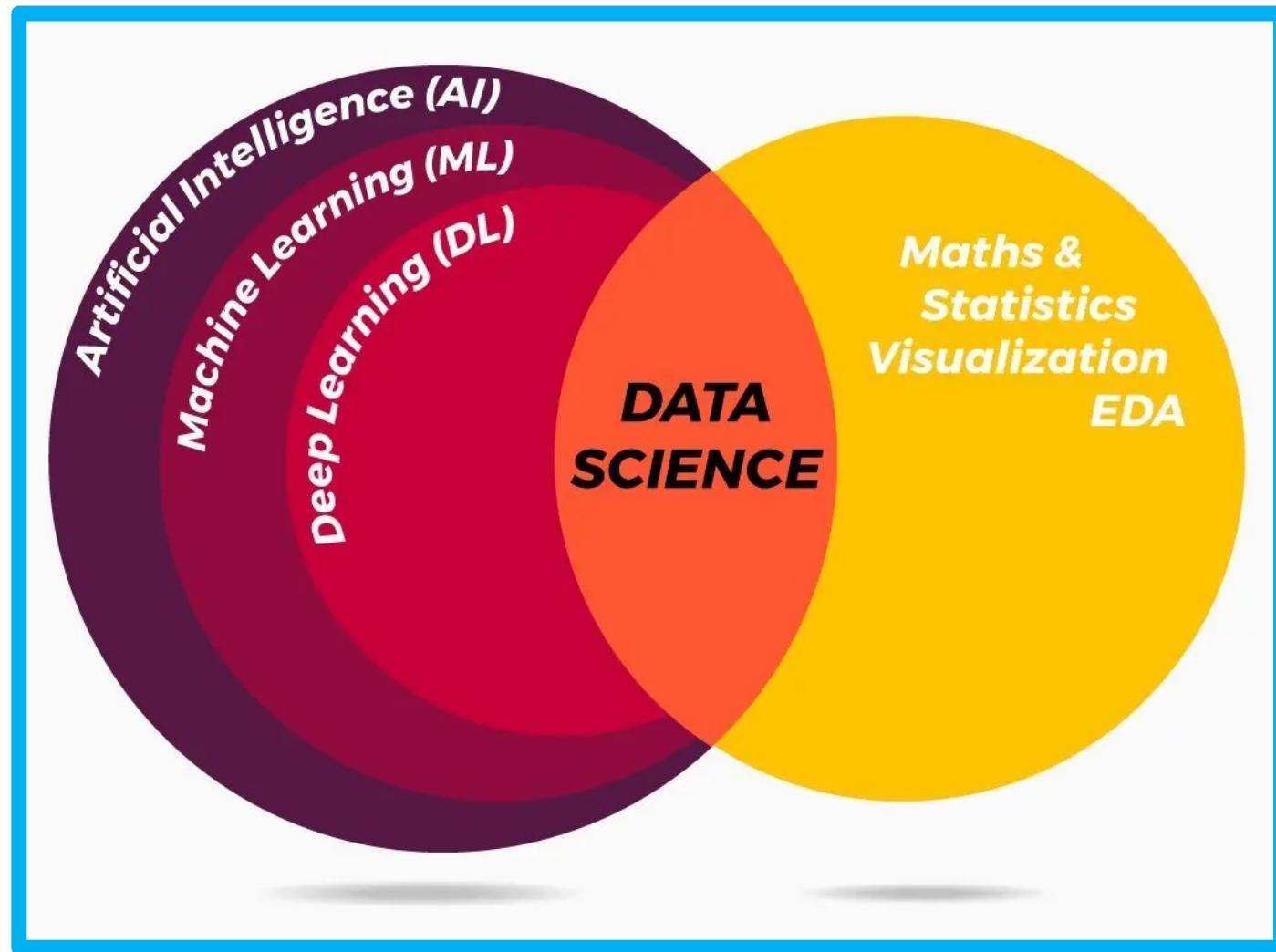
Дано:

- Множество рейсов (время и аэропорт для вылета/прилета)
- Стоимость полета экипажа экипажем и стоимость полета экипажа пассажирами
- Стоимость простоя
- Число экипажей в аэропортах
- Экипаж должен вернуться в пункт вылета
- Предполетные операции 1 час, послеполетные 15 минут

Требуется:

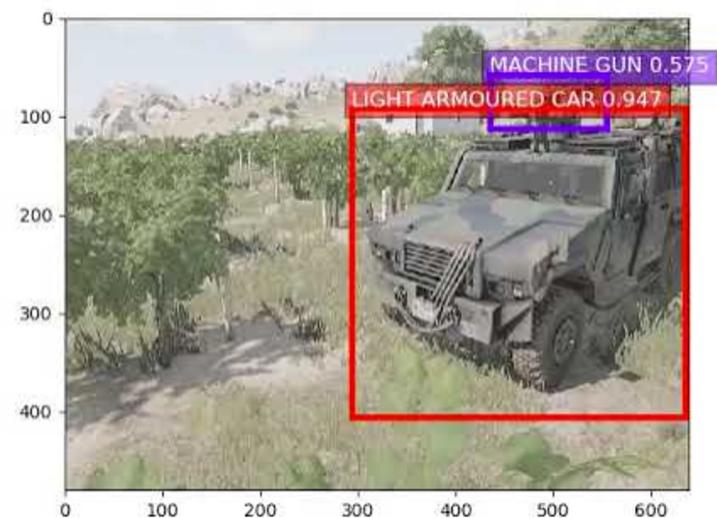
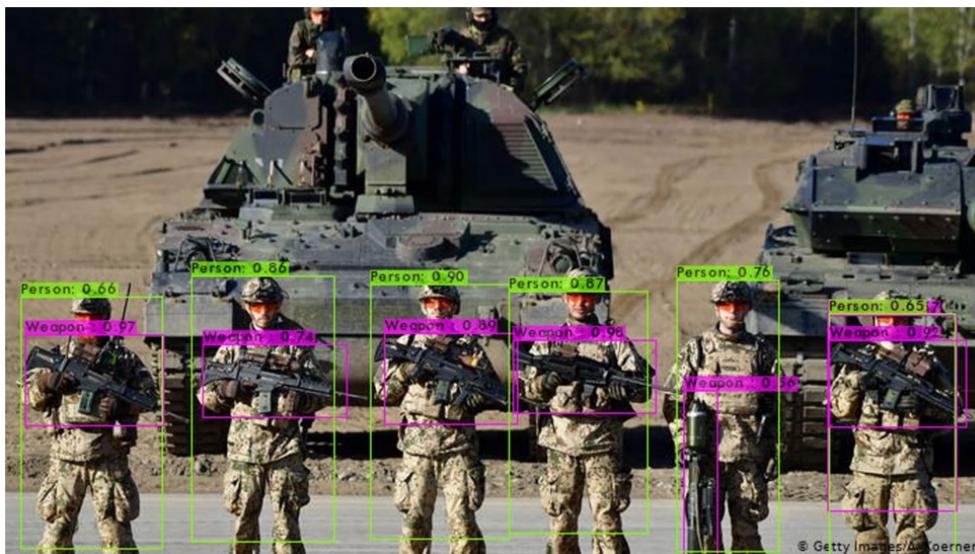
- Составить расписание для экипажей

Взаимосвязи между науками



Некоторые популярные на сегодня задачи ИИ

- Обработка естественного языка (NLP), в том числе аннотирование научной-технической информации
- Анализ изображений



Машинное обучение. Определение

- Машинное обучение (англ. machine learning, ML) — это исследование компьютерных алгоритмов, которые автоматически улучшаются благодаря опыту и использованию данных.
- Алгоритмы машинного обучения создают модель на основе выборочных данных, известных как «обучающие данные», чтобы делать прогнозы или предлагать решения, не будучи явно запрограммированными на это.

Вопрос для обсуждения

Многие методы Data Science и Machine Learning
появились достаточно давно - 50-70 года
прошлого века, но активно использоваться в
бизнесе начали только сейчас

С чем это связано?

Что такого случилось?

Что есть сейчас и чего не было тогда?

Искусственный интеллект

ГОСТ Р 59277— 2020:

Искусственный интеллект (*artificial intelligence*):

комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение, поиск решений без за ранее заданного алгоритма и достижение инсайта) и получать при выполнении конкретных практически значимых задач обработки данных результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека

Машинное обучение

- Машинное обучение (англ. machine learning, ML) — это исследование компьютерных алгоритмов, которые автоматически улучшаются благодаря опыту и использованию данных
- Алгоритмы машинного обучения создают модель на основе выборочных данных, известных как «обучающие данные», чтобы делать прогнозы или предлагать решения, не будучи явно запрограммированными на это

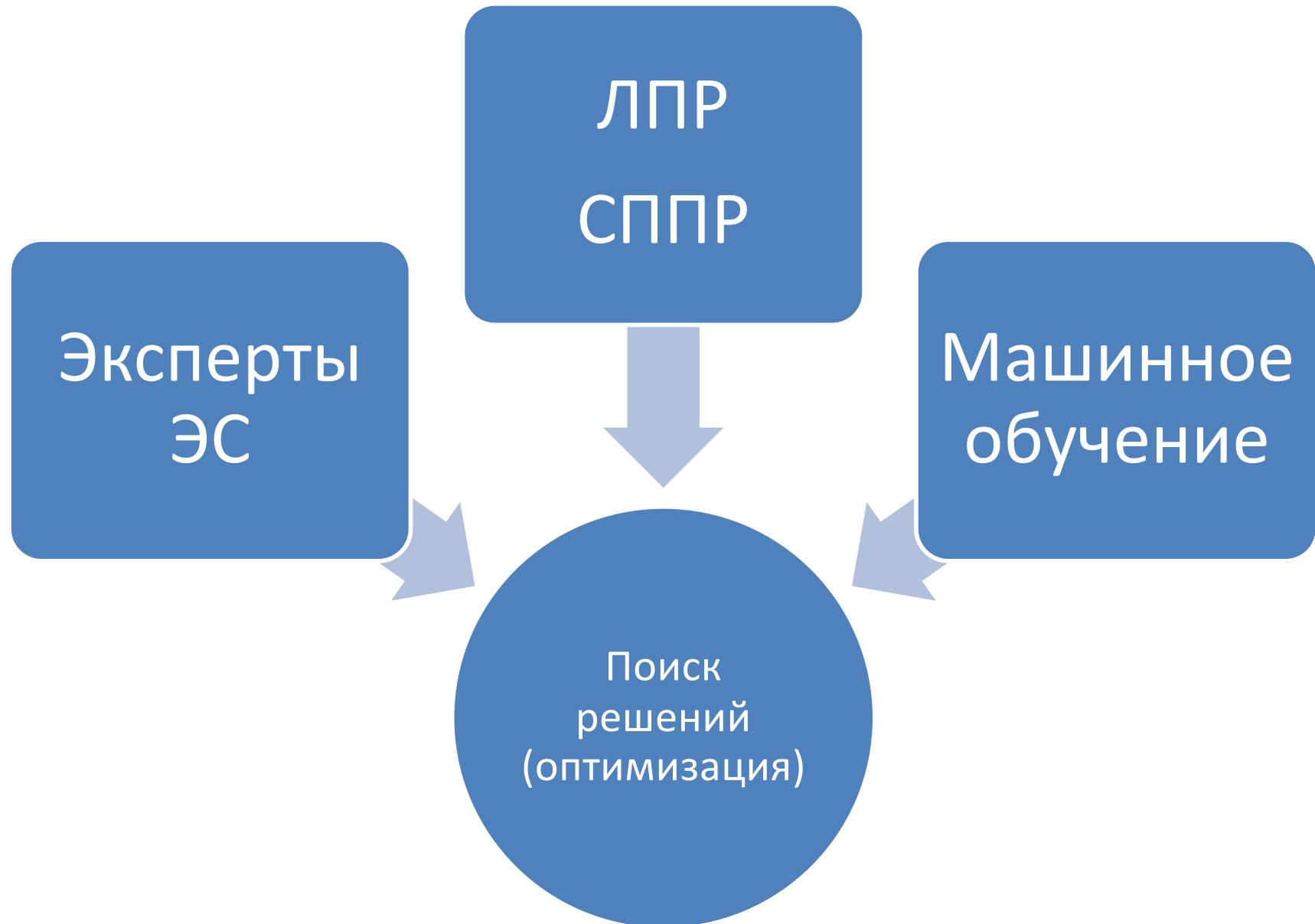
Давайте подумаем про цели

- Зачем мы учим машины? Например, получить прогноз продаж
- Какие у разработчика моделей?
- А какие цели у бизнеса?

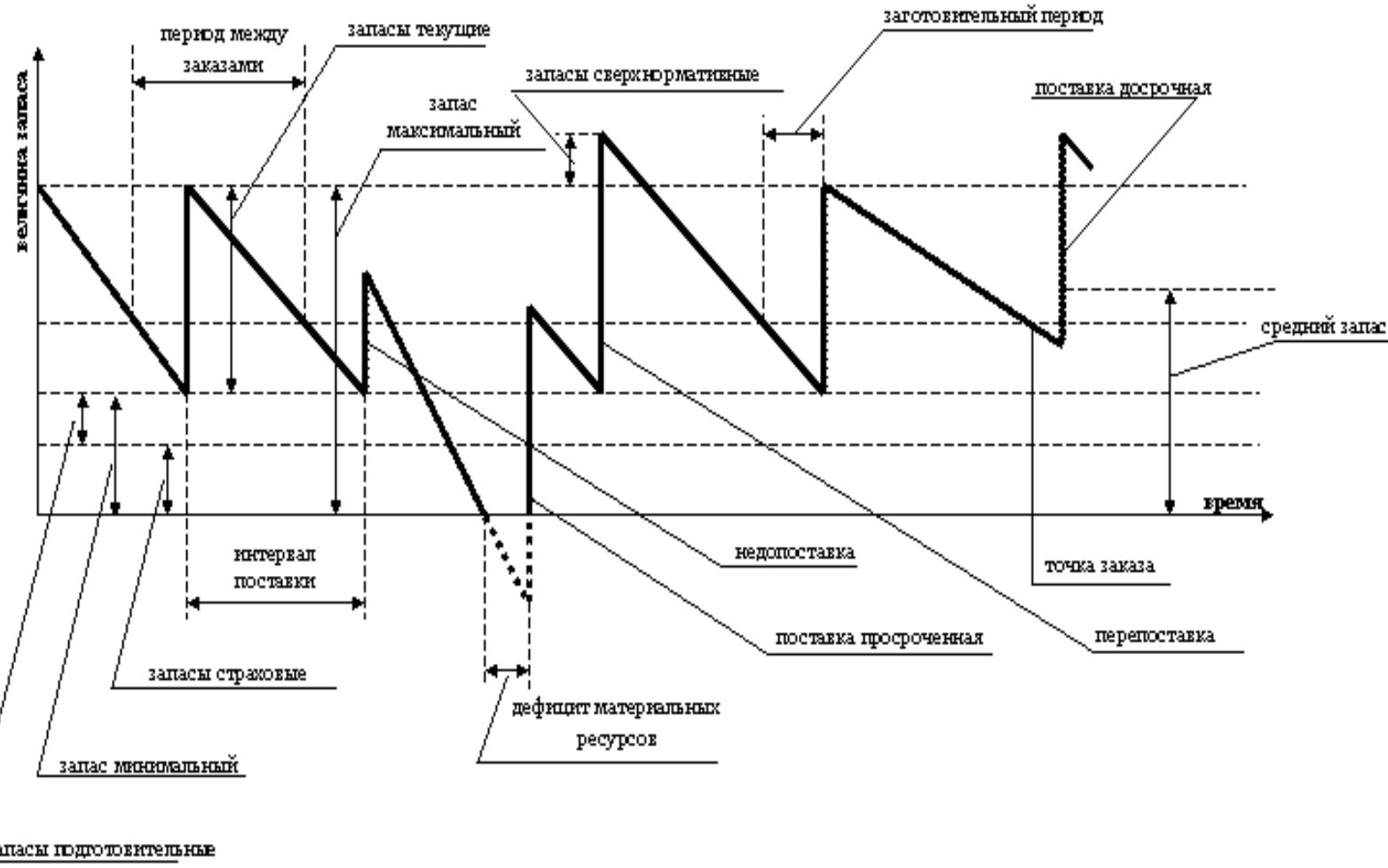
Цели

- Прибыль (часто важность не велика в краткосрочной перспективе)
- Доля рынка (важность велика)
- Сделать людей счастливыми
- Прославится
- Обеспечить долгую, стабильную жизнь компании
- Не знаю....

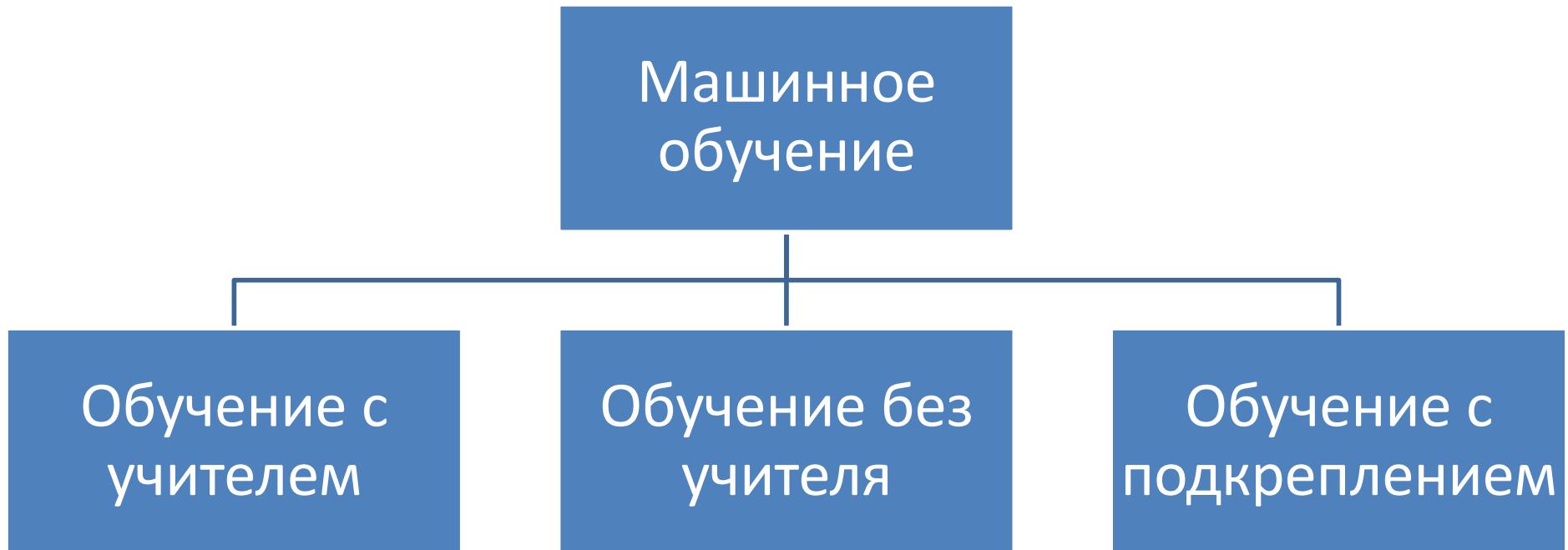
Подходы ИИ



Пример из бизнеса FMCG



Машинное обучение



Задачи машинного обучения

- Регрессия
- Классификация
- Ранжирование
- Кластеризация
- Понижение размерности

Итоги

Нужно справиться с тремя сложностями:



Критерии и цели не
определенны



Нет данных



Пространство поиска решений
колossalno – даже
суперкомпьютер не справится

Методы машинного обучения с учителем

Дано

Множество объектов X .

Множество допустимых ответов Y .

Целевая функция (target function) $y^* : X \rightarrow Y$, значения которой $y_i = y^*(x_i)$ известны только на конечном подмножестве объектов $\{x_1, \dots, x_l\} \subset X$.

Пары «объект– ответ» (x_i, y_i) называются прецедентами.

Совокупность пар $X^l = (x_i, y_i)_{i=1}^l$ называется обучающей выборкой (training sample).

Требуется найти

зависимость y^* по выборке X' , то есть построить решающую функцию (decision function)

$$a: X \rightarrow Y,$$

которая приближала бы целевую функцию $y^*(x)$, причём не только на объектах обучающей выборки, но и на всём множестве X .

Решающая функция a должна допускать эффективную компьютерную реализацию.

Типы задач

Задача регрессии – прогноз на основе выборки объектов с различными признаками. На выходе - вещественное число (2, 35, 76.454 и др.). Например, цена квартиры, стоимость ценной бумаги через неделю, ожидаемый доход магазина на следующий месяц, качество вина. Множество Y – бесконечное.

Задача классификации – получение категориального ответа на основе набора признаков. Имеет конечное количество ответов (часто, в формате «да» или «нет»): является ли изображение человеческим лицом, давать ли клиенту кредит, к какой категории отнести товар. Множество Y – конечное.

Пример задачи

Вход X :

4, 8, 9, 26

Выход Y :

39.2, 46.4, 48.2, 78.8

Задача:

Если $x = 256$,
то чему равен y ?

Подготовка к решению

jupyter ml1 Last Checkpoint: 2 минуты назад (autosaved)

Файл Редактировать Вид Вставка Ячейка Ядро Widgets Справка

File Cell Kernel Help

In [1]:

```
1 from sklearn import linear_model
2 import numpy as np
3 import matplotlib.pyplot as plt
```

In [2]:

```
1 x = np.array([4, 8, 9, 26])
2 y = np.array([39.2, 46.4, 48.2, 78.8])
```

In [3]:

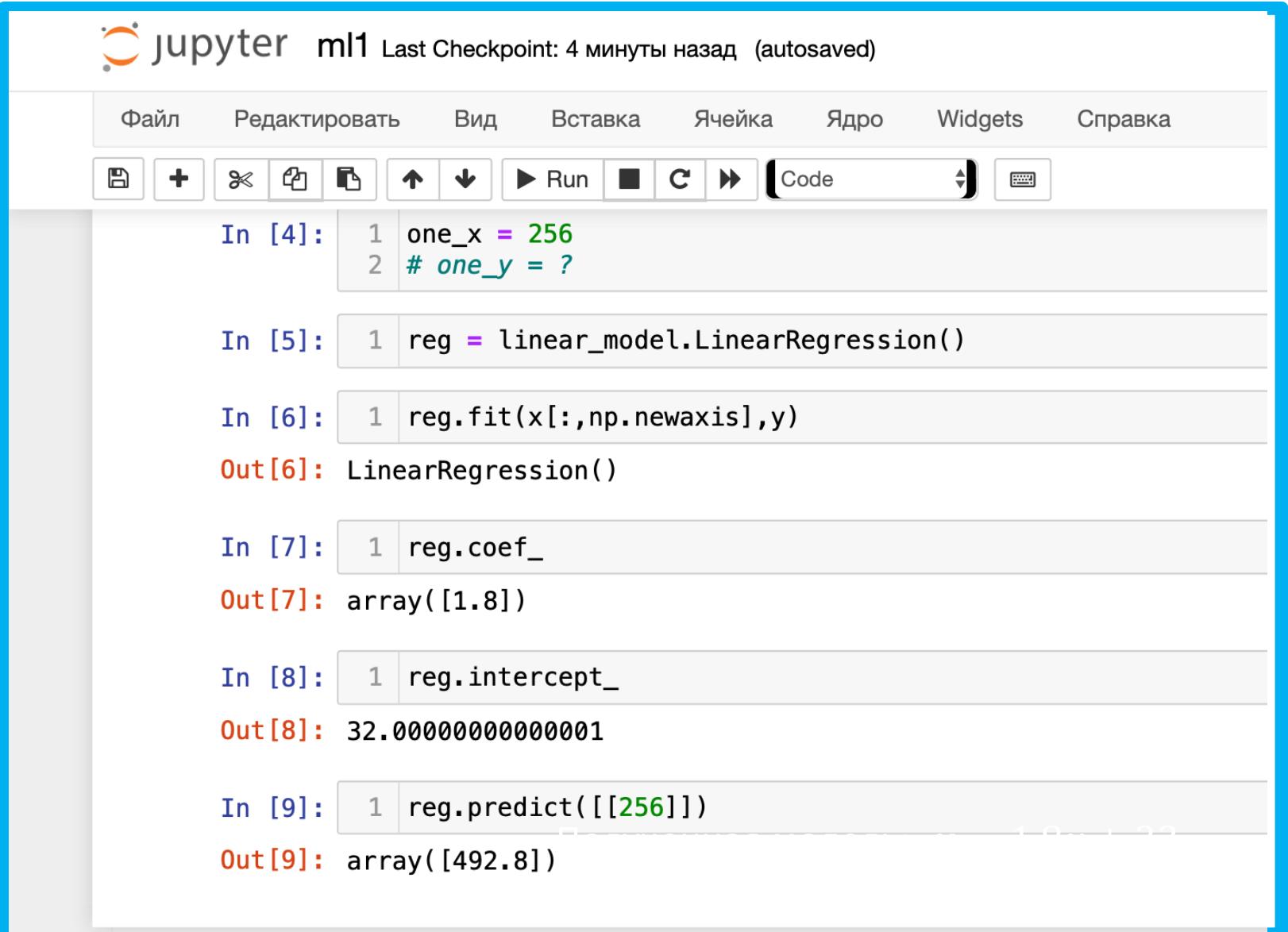
```
1 plt.plot(x, y)
```

Out[3]: [<matplotlib.lines.Line2D at 0x7fbe6b2144f0>]

In [4]:

```
1 one_x = 256
2 # one_y = ?
```

Решение в scikit-learn



The screenshot shows a Jupyter Notebook interface with a blue border. At the top, it displays the logo and the notebook name "jupyter ml1 Last Checkpoint: 4 минуты назад (autosaved)". Below the title bar is a menu bar with Russian labels: Файл, Редактировать, Вид, Вставка, Ячейка, Ядро, Widgets, Справка. Underneath the menu bar is a toolbar with various icons for file operations like saving, running cells, and cell selection.

The main area contains the following code snippets:

```
In [4]: 1 one_x = 256
         2 # one_y = ?

In [5]: 1 reg = linear_model.LinearRegression()

In [6]: 1 reg.fit(x[:,np.newaxis],y)

Out[6]: LinearRegression()

In [7]: 1 reg.coef_

Out[7]: array([1.8])

In [8]: 1 reg.intercept_

Out[8]: 32.00000000000001

In [9]: 1 reg.predict([[256]])

Out[9]: array([492.8])
```

Решение в многомерном случае

x1	x2	x3	x4	y
18	71	69	60	77
27	74	97	27	337
5	42	33	41	-27
49	38	44	56	123
83	76	42	37	196
91	81	34	42	155
78	93	35	49	76
97	41	89	93	
	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение
Y-пересечение	2,2737E-13	1,4927E-13	1,52328045	0,26714387
Переменная X 1	3	1,45E-15	2,0689E+15	2,3362E-31
Переменная X 2	-2	3,6581E-15	-5,467E+14	3,3454E-30
Переменная X 3	5	1,6099E-15	3,1058E+15	1,0367E-31
Переменная X 4	-3	3,2317E-15	-9,283E+14	1,1604E-30

Математическая постановка

Дана обучающая выборка: $X^l = \{(x_i, y_i)\}_{i=1}^l$

x_i – вектор размерности n .

Необходимо построить такую функцию:

$$a(x, w) = \sum_{j=1}^n w_j f_j(x_j)$$

Для поиска w используем метод наименьших квадратов:

$$\min_w \sum_{i=1}^l (a(x_i, w) - y_i)^2$$

Вопросы для контроля

Мы начали с простейшей модели:

$$y = 1.8x + 32$$

А как в формуле общего вида

$$y = \sum_{j=1}^n w_j x_j$$

Отразить свободный член 32?

А можно ли используя линейную регрессию получить квадратичную модель вида:

$$y = ax^2 + bx + c ?$$

Как оценить качество решения?

Перед началом обучения выборку разбивают на обучающую и тестовую.

Качество следует оценить на тестовой выборке.

Меры качества:

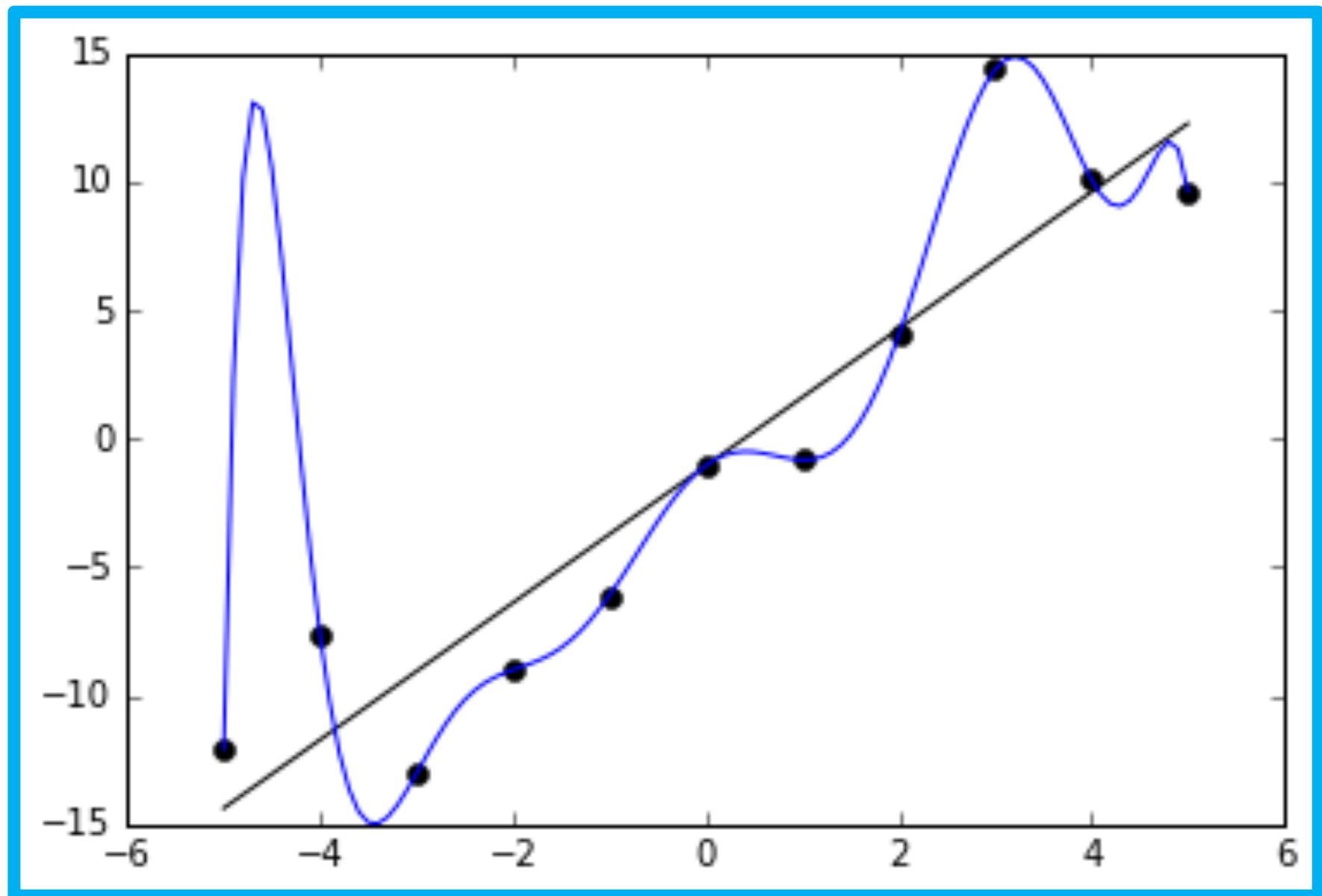
Средняя квадратичная ошибка (англ. Mean Squared Error, MSE).

Средняя абсолютная ошибка (англ. Mean Absolute Error, MAE).

Коэффициент детерминации (R^2). Коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Если она близка к единице, то модель хорошо объясняет данные, если же она близка к нулю, то прогнозы сопоставимы по качеству с константным предсказанием.

$$R^2 = 1 - \frac{\sum_{i=1}^n (a(x_i) - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Переобучение



Вопросы для контроля

Чему равен y для последней строки в примере с многомерной регрессией?

x1	x2	x3	x4	y
97	41	89	93	???

Постройте линейную регрессию для предсказания качества вина по набору данных:

<https://archive.ics.uci.edu/ml/datasets/wine+quality>

Оцените качество решения

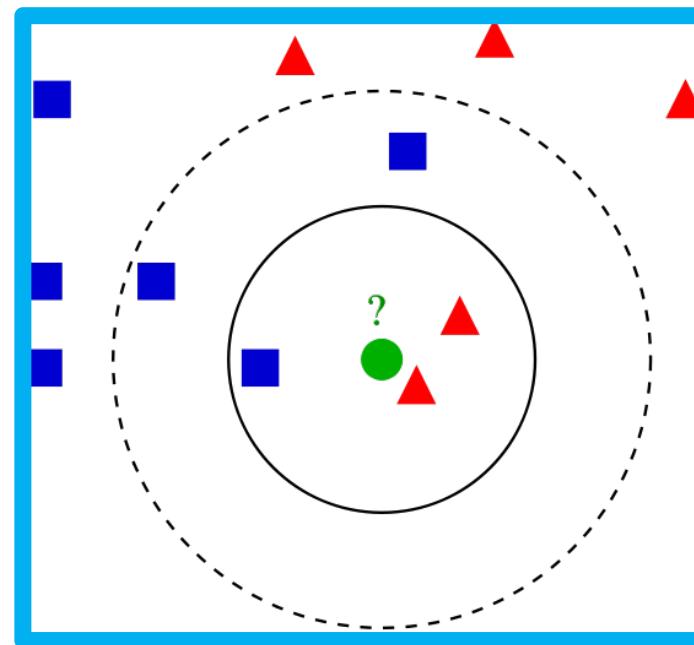
Метод ближайших соседей

Правило классификации: объект принадлежит тому же классу что и его k-ближайший соседей.

Близость определяется в пространстве признаков.

Для применения метода необходимо решить задачи:

- нормализация признаков,
- выбор метрики,
- выбор k.



Оценка качества классификации

Пусть есть два класса $Y=\{0,1\}$. Пусть банк использует систему классификации заёмщиков на кредитоспособных и некредитоспособных. Обнаружение некредитоспособного заёмщика ($y=1$) можно рассматривать как "сигнал тревоги", сообщающий о возможных рисках.

Возможны следующие исходы классификации:

- Некредитоспособный заёмщик классифицирован как некредитоспособный, т.е. положительный класс распознан как положительный (True Positive — TP).
- Кредитоспособный заёмщик классифицирован как кредитоспособный, т.е. отрицательный класс распознан как отрицательный. (True Negative — TN).
- Кредитоспособный заёмщик классифицирован как некредитоспособный, т.е. имела место ошибка, в результате которой отрицательный класс был распознан как положительный (False Positive — FP) – это ошибка I рода (ложная тревога).
- Некредитоспособный заёмщик распознан как кредитоспособный, т.е. имела место ошибка, в результате которой положительный класс был распознан как отрицательный (False Negative — FN) – это ошибка II рода (пропуск цели).

Вопрос для самоконтрол

Где ошибка первого рода и где ошибка второго рода?



Метрики качества классификатора

Аккуратность (англ. Accuracy) – доля правильных ответов.

Бесполезна в задачах с неравными классами.

Точность (англ. Precision) - доля правильных ответов модели в пределах класса:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Полнота (англ. Recall) - это доля истинно положительных классификаций:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F-мера (англ. F-score) – гармоническое среднее между точностью и полнотой.

Итоги

- Машинное обучение с учителем решает задачи регрессии и классификации
- Нужно сформировать обучающую и проверочную выборки
- Визуализируйте данные, выбирайте признаки
- Оценивайте метрики результата
- Совершенствуйте модели

Методы машинного обучения без учителя

Отличия

Обучение с учителем (supervised) vs

Обучение без учителя (unsupervised)



Задачи и приложения

- Задачи кластеризации
- Задачи обобщения
- Задачи обнаружения аномалий
- Задачи поиска правил ассоциации
- Задачи сокращения размерности
- Маркетинговые исследования: разбиение множества всех клиентов на кластеры для выявления типичных предпочтений.
- Анализ рыночных корзин: выявление сочетаний товаров, часто встречающихся вместе в покупках клиентов.

Метод k-средних

Задача: разбить выборку на кластеры

$$S = \{S_1, S_2, \dots, S_k\}$$

Метод стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:

$$\arg \min_S \sum_{i=1}^k \sum_{x \in S_i} (x - \mu_i)^2$$

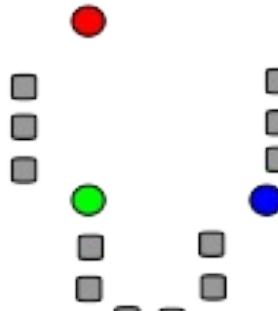
μ_i - центр кластера S_i .

Алгоритм k-средних

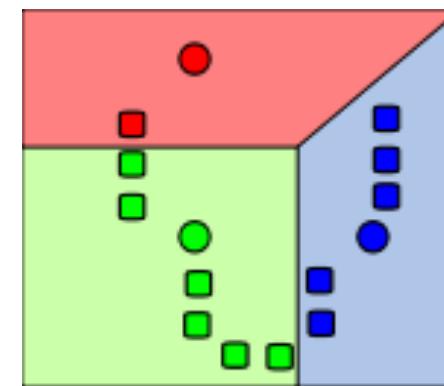
1. Начальные μ_i выбираются случайно.
2. Относим наблюдения к тем кластерам, чье μ_i к ним ближе всего.
3. Затем μ_i перевычисляется.
4. Если μ_i существенно изменились, то возврат к шагу 2.

Визуализация алгоритма k-средних

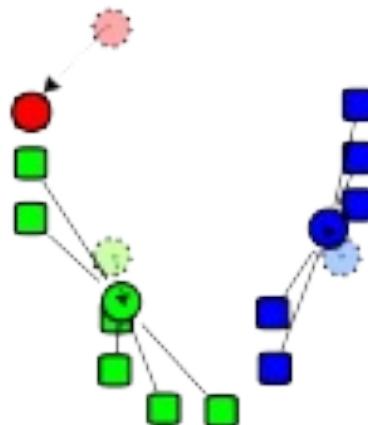
1



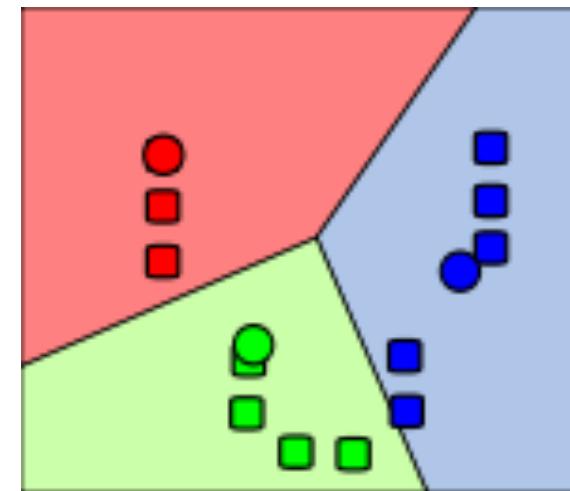
2



3



4



Пример задачи

Опасное вождение



Есть данные по трекам, скорости и ускорению
Нет (и не будет) данных о тому что такое
опасное вождение
А что делать?

Кластеризация вождения

Вход модели:

- скорости и ускорения конкретных ТС по конкретным водителям
- число кластеров
- уровень иерархии объектов для прогноза

Выход модели:

- принадлежность водителя определённому кластеру

По результатам работы модели нужно:

- выбрать несколько водителей из каждого кластера – отследить насколько хорошо они водят.
- присвоить кластерам категории опасности.
- на регулярной основе информировать о попадании водителя в «опасный» кластер.

Оценка качества кластеризации

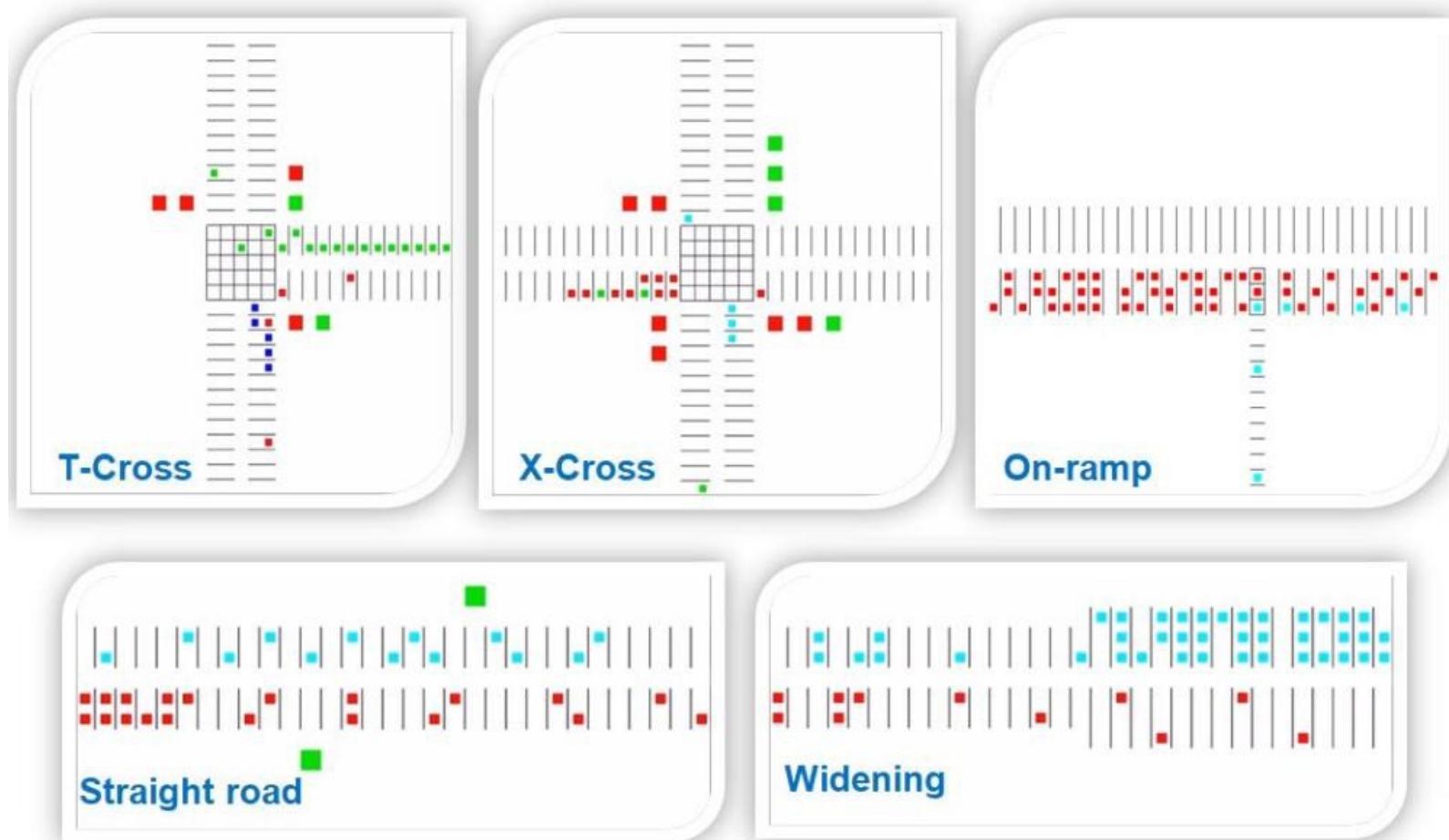
Внешние (англ. External) меры основаны на сравнении результата кластеризации с априори известным разделением на классы.

Внутренние (англ. Internal) меры отображают качество кластеризации только по информации в данных.

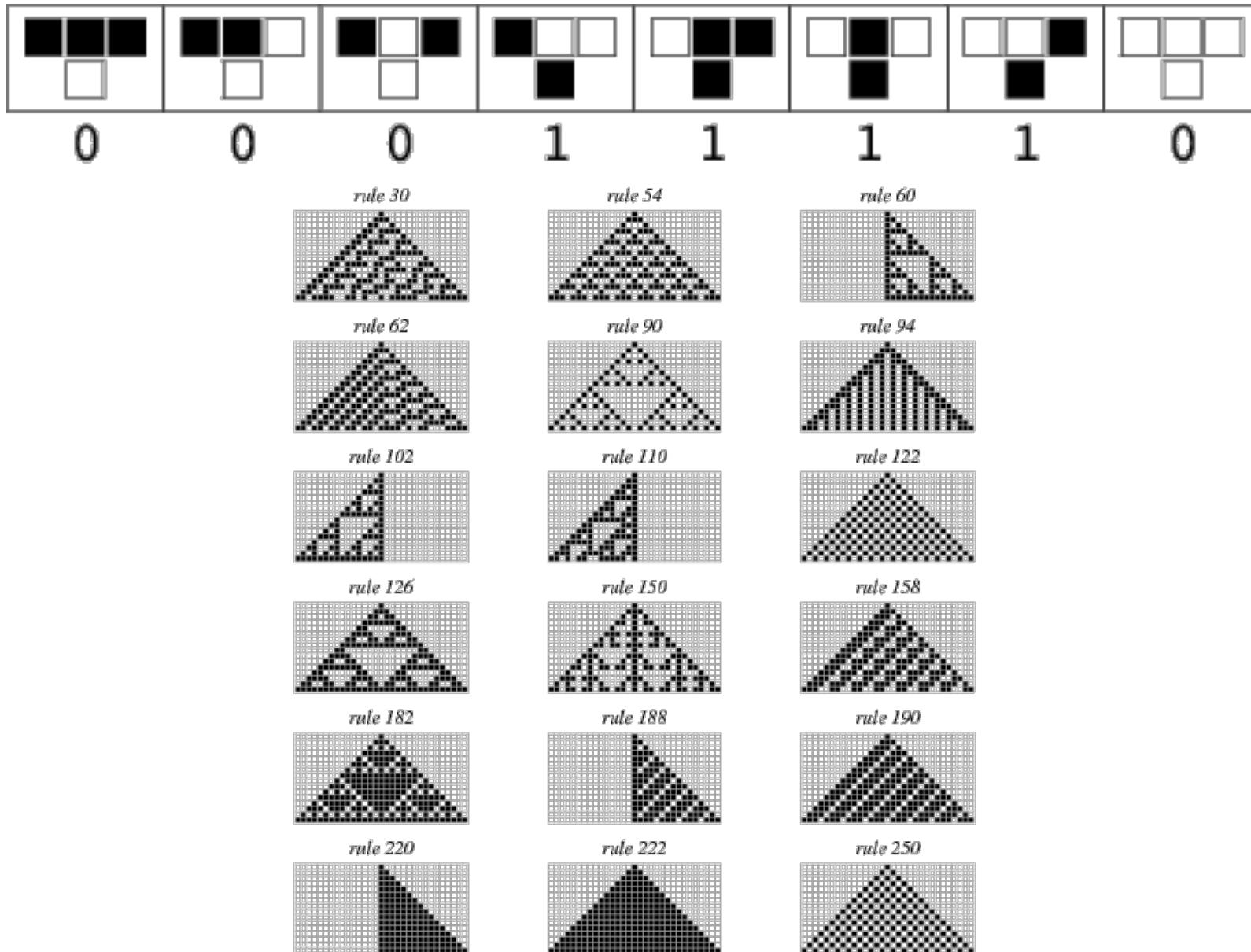
Из внутренних мер часто применяют:

Силуэт (англ. Silhouette) – он показывает, насколько объект похож на свой кластер по сравнению с другими кластерами.

Модели клеточных автоматов



Модель Вольфрама



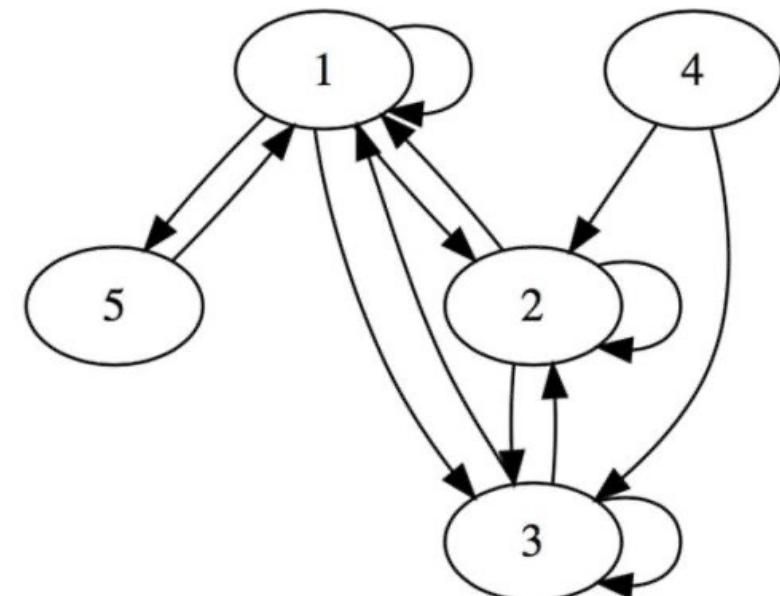
Дискретно-событийные модели

Планируемые события заносятся в список будущих событий (СБС). Его структура:

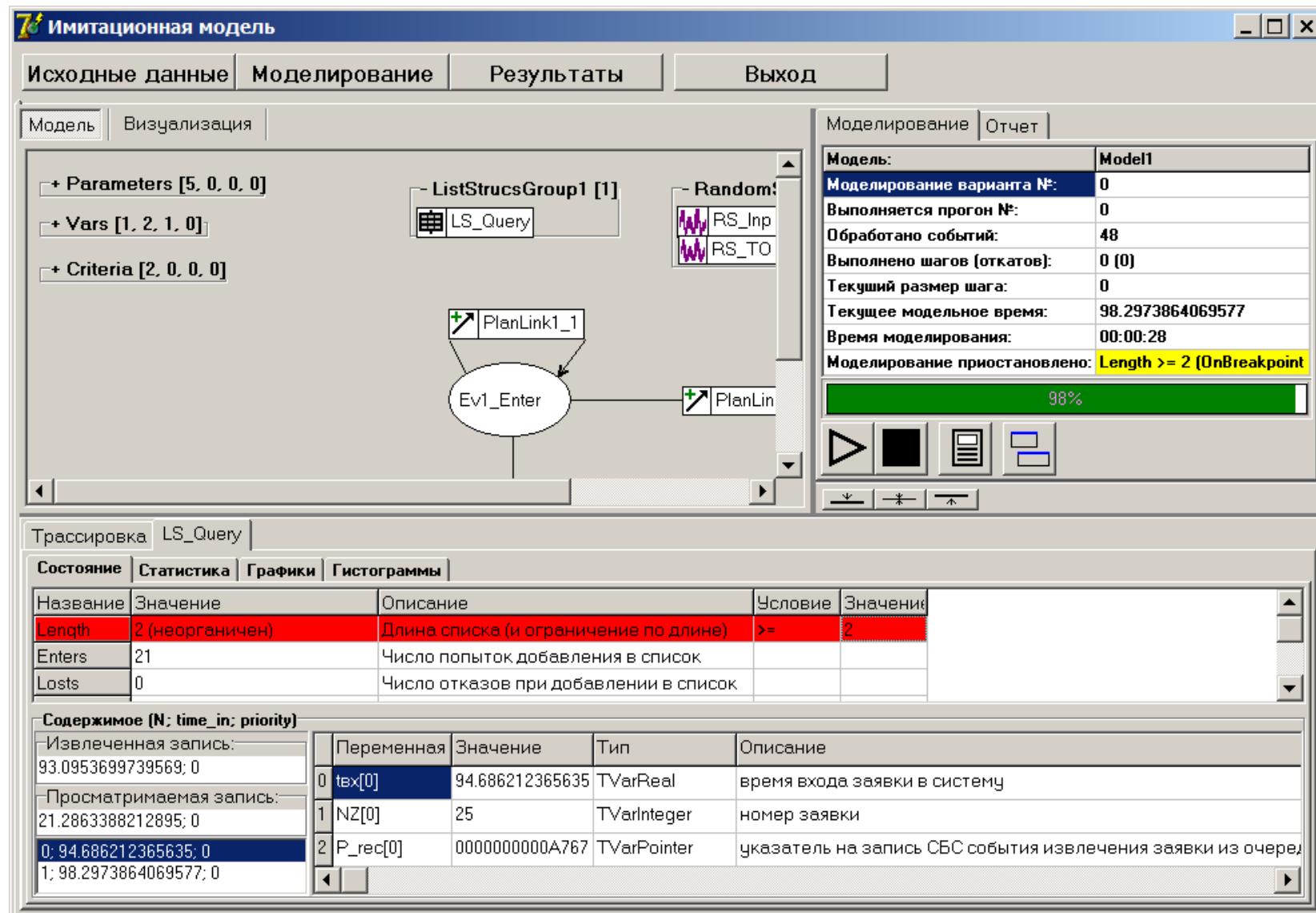
- время свершения события
- код события
- место события
- массив идентификаторов ЛА, к которым относится событие.

ПО:

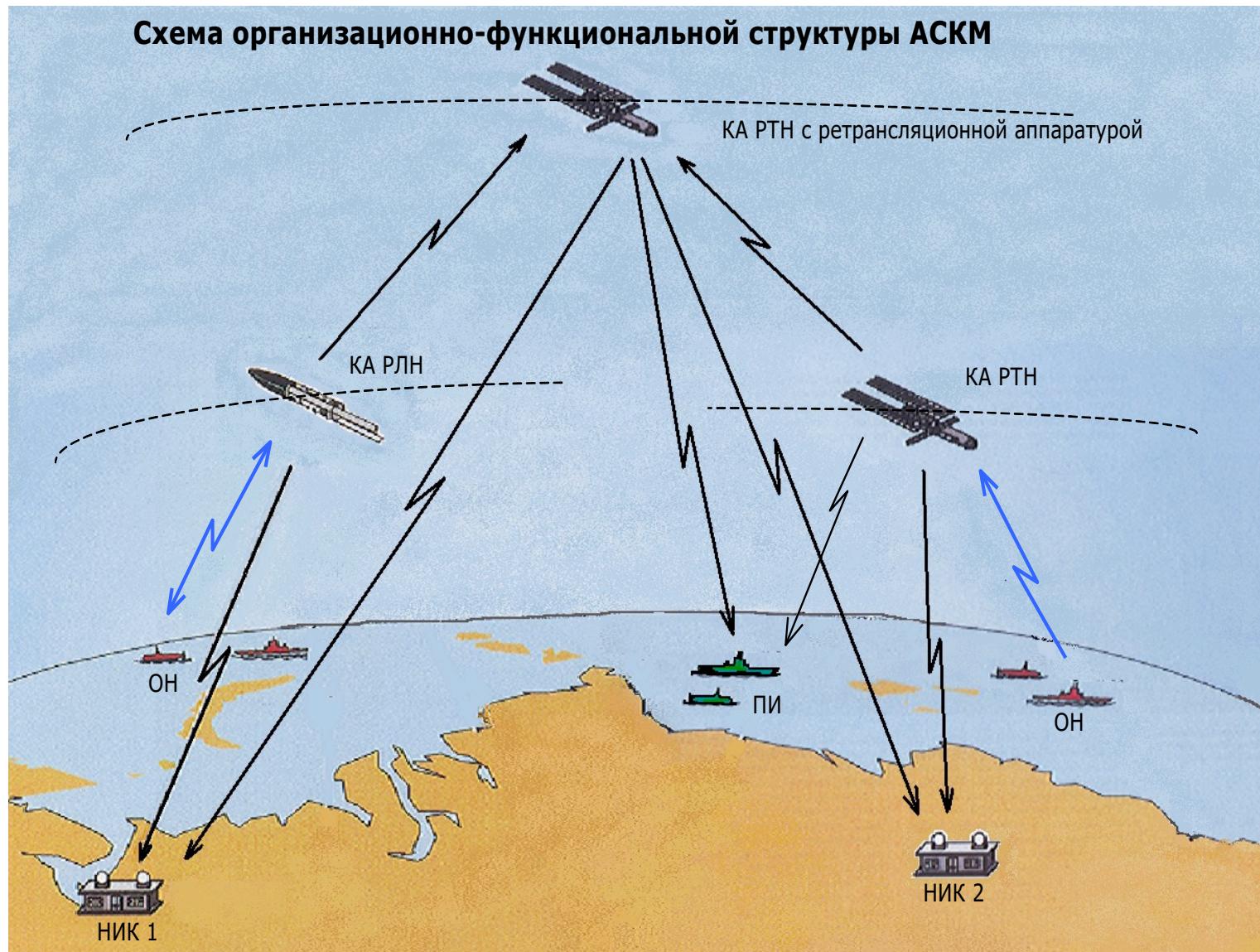
AnyLogic, GPSS, SimPy



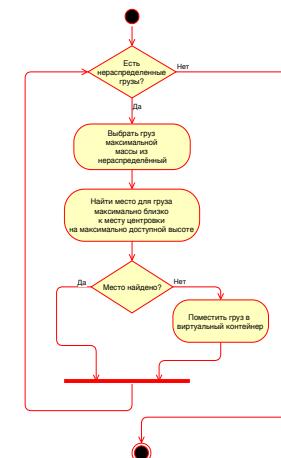
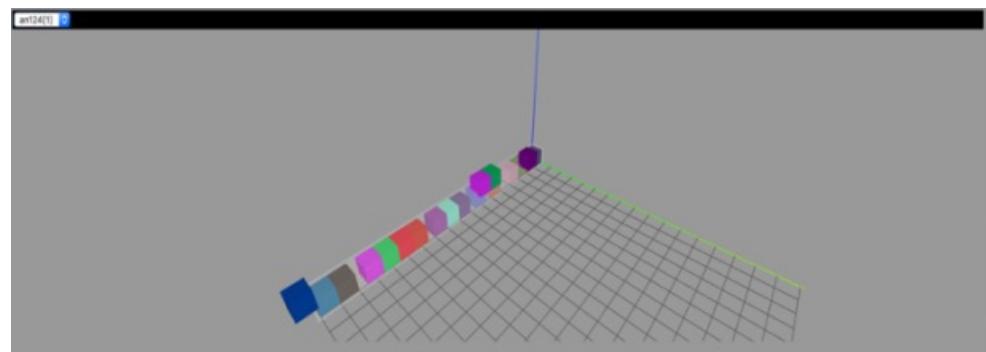
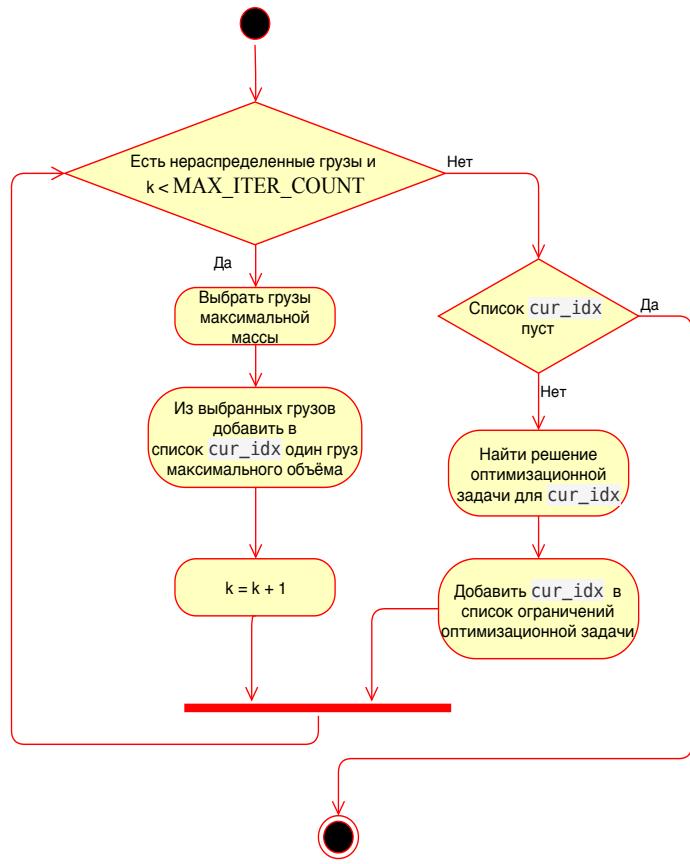
Имитационное моделирование



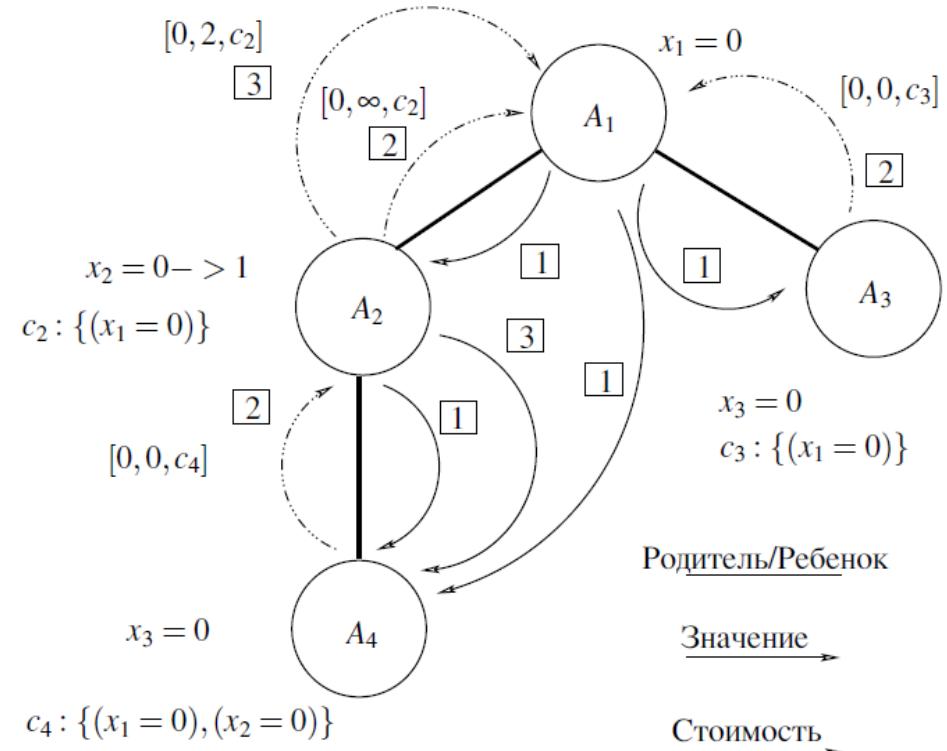
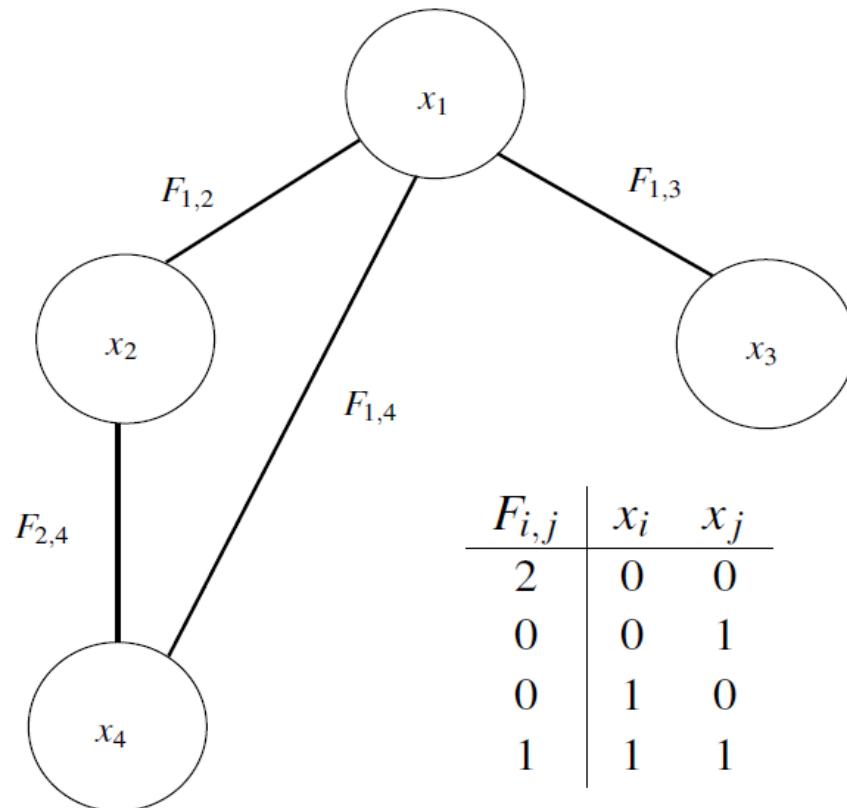
Планирование космических наблюдений



Алгоритм декомпозиции, эвристика, мультиагентный подход

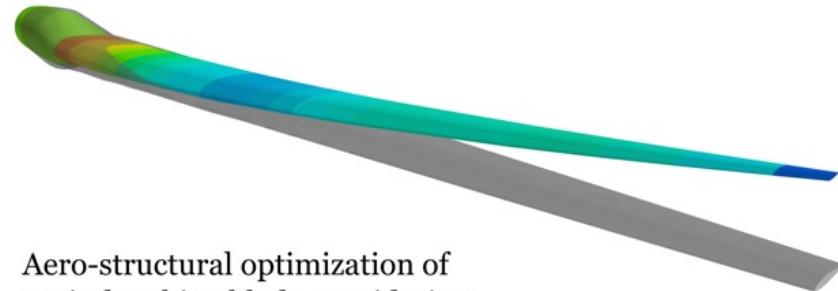
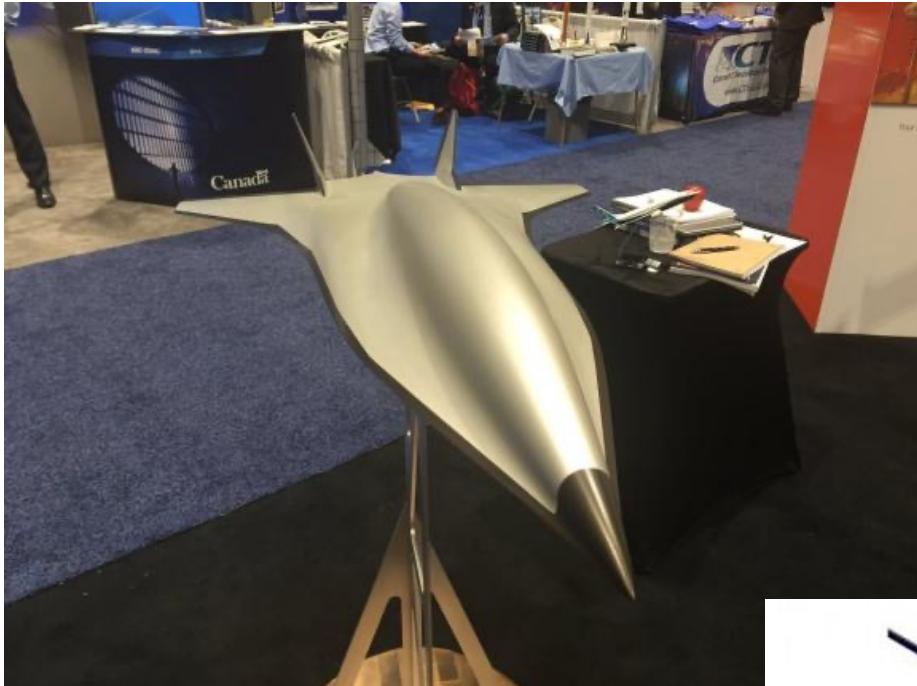


Мультиагентное моделирование в задачах DCOP

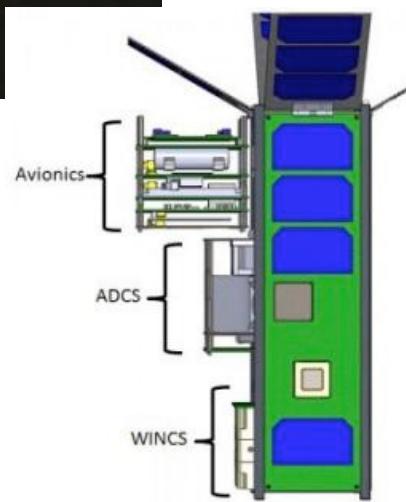


Пакет pyDCOP

Мультидисциплинарная оптимизация



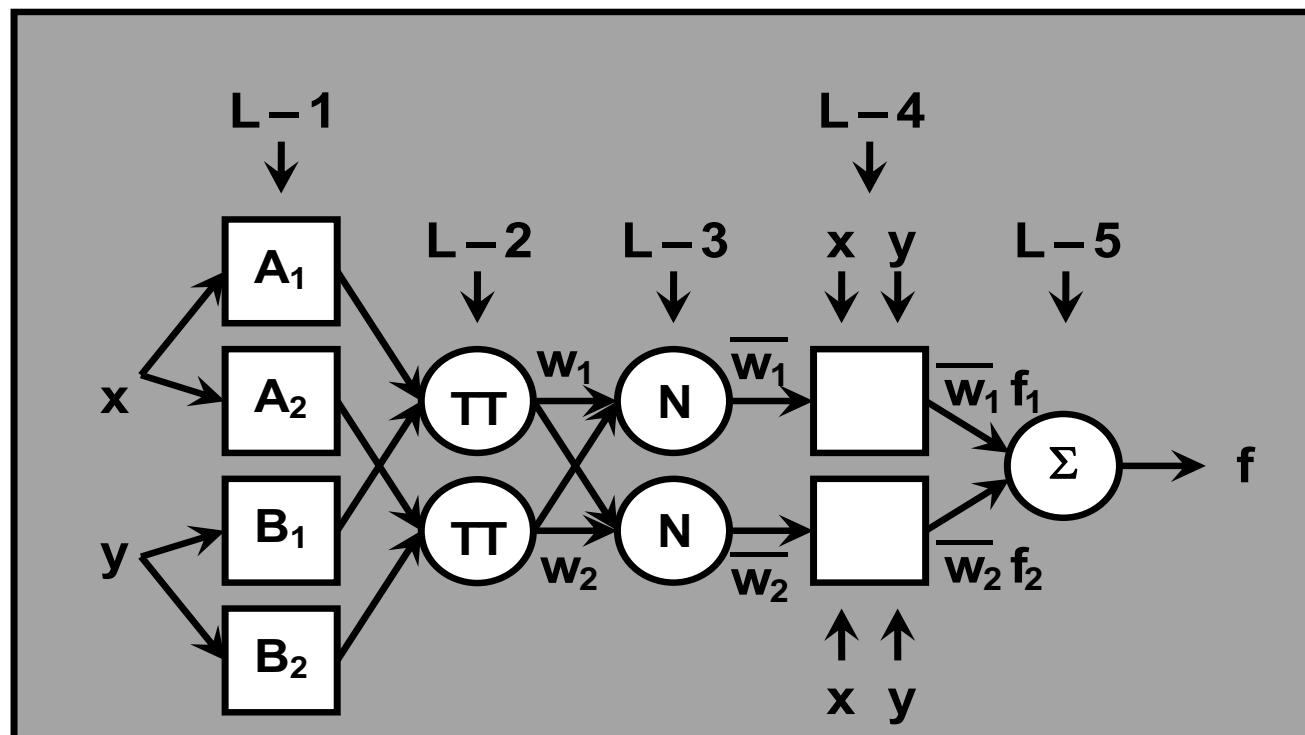
Aero-structural optimization of
a wind turbine blade considering
pre-curve jig shape



Пакет OpenMDAO

NEURO-FUZZY модель

Нейронная сеть обучается на базе нечетких предпочтений ЛПР:



Многокритериальная оценка ЛА

Критерии:

Технические характеристики	Экономичность	Импортозависимость
Хорошие 1	Хорошая 1	Высокая 1
Удовл. 0.5	Удовл. 0.5	Низкая 0
Неудовл. 0	Неудовл. 0	

Альтернативы:

АН-124



ИЛ-76



Тех. характеристики	Экономичность	Импортозависимость
0.25	0.8	1

Тех. характеристики	Экономичность	Импортозависимость
0.5	0.35	0

Области предпочтения:

Уровень предпочтения	Низкий	Средний	Высокий
	Значения критериев	Тех. характеристики: 0.5, 0 Экономичность: 0.5, 0 Импорт.: 1	Тех. характеристики: 0.5 Экономичность: 1, 0.5 Иморт.: 1

Ранги:

Ранг АН-124: 0.8256

Ранг ИЛ-76: 0.3484

Использование областей предпочтений

Области предпочтения

Удалить Уровень предпочтения Неуд 2.0

Технические характеристики	Хорошие 0.5	Плохие 0.3	Неуд 0.0
Экономичность	Уд 0.3	Неуд 0.0	Хор 0.5

Удалить Уровень предпочтения Уд 3.0

Технические характеристики	Хорошие 0.5	Плохие 0.3	Неуд 0.0
Экономичность	Уд 0.3	Неуд 0.0	Хор 0.5

Удалить Уровень предпочтения Хорошо 4.0

Технические характеристики	Хорошие 0.5	Плохие 0.3	Неуд 0.0
Экономичность	Уд 0.3	Неуд 0.0	Хор 0.5

Удалить Уровень предпочтения Неуд 2.0

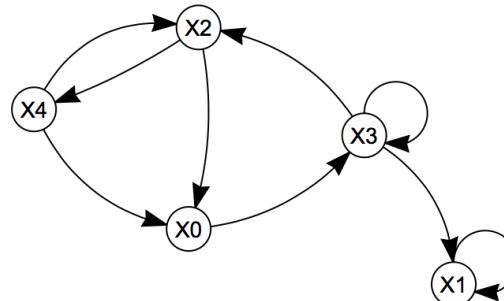
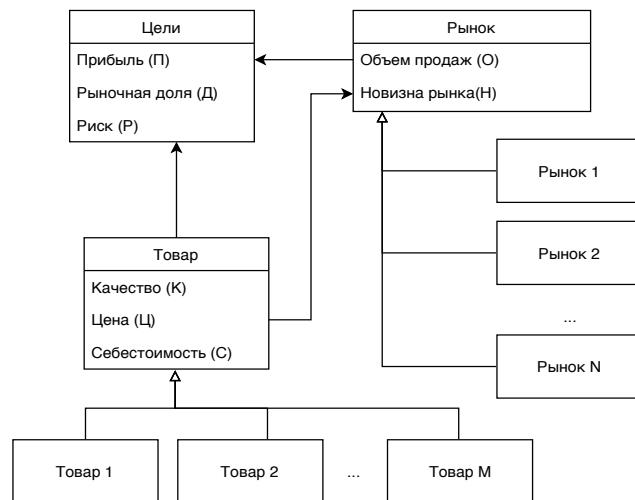
Технические характеристики	Хорошие 0.5	Плохие 0.3	Неуд 0.0
Экономичность	Уд 0.3	Неуд 0.0	Хор 0.5

[Добавить область предпочтения](#)

[Сохранить](#)

Факторная модель для исследования сложных процессов

- Анализ сложных систем включая технические, экономические и социальные, требует одновременного учета большого количества разнородных показателей.
- Во многих случаях приходится сталкиваться с высокой размерностью решаемой задачи, когда число факторов велико. Объем вычислений резко возрастает. Для разрешения данной ситуации разработана процедура моделирования факторов на основе концепции фреймов.
- Технология:
 - Библиотека матричных вычислений LAPACK.
 - Интерфейс вызова на языке Ruby
- Результаты:
 - Исследована модель инновационной деятельности предприятий.



$$U_{i,j^*} = \begin{vmatrix} 1 & u_{12}^{j^*} & u_{13}^{j^*} & u_{14}^{j^*} & \dots & u_{1q_1}^{j^*} \\ 1/u_{12}^{j^*} & 1 & u_{23}^{j^*} & u_{24}^{j^*} & \dots & u_{2q_1}^{j^*} \\ 1/u_{13}^{j^*} & 1/u_{23}^{j^*} & 1 & u_{34}^{j^*} & \dots & u_{3q_1}^{j^*} \\ 1/u_{14}^{j^*} & 1/u_{24}^{j^*} & 1/u_{34}^{j^*} & 1 & \dots & u_{4q_1}^{j^*} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1/u_{1q_1}^{j^*} & 1/u_{2q_1}^{j^*} & 1/u_{3q_1}^{j^*} & 1/u_{4q_1}^{j^*} & \dots & 1 \end{vmatrix}$$

$$a_{ij} = \begin{cases} w_{i,j}^{j^*}, & \text{если } \exists i_1 \exists i_2 : ((i_1, i_2) \in G) \wedge (i \in J_{i_1}) \wedge (j \in J_{i_2}) \\ 0, & \text{иначе} \end{cases}$$

ГОСТ 19.101-77

По видам программы делят на:

- компонент — программа, рассматриваемая как единое целое, выполняющая законченную функцию и применяемая самостоятельно или в составе комплекса;
- комплекс — программа, состоящая из двух или более компонентов и (или) комплексов, выполняющих взаимосвязанные функции, и применяемая самостоятельно или в составе другого комплекса.