Разработка методов и средств контроля достоверности и актуальности фактографического наполнения информационных систем

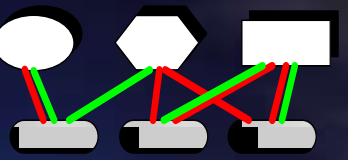
Серый А.С.

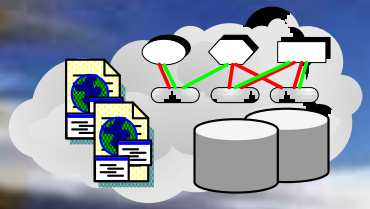
Институт систем информатики им. А.П. Ершова

Новосибирск

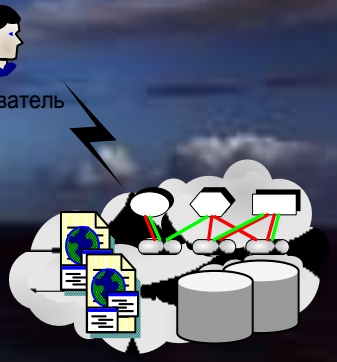
Информационная система

- Онтология - целостное представление предметной области и различных ее аспектов





- интеграция знаний и информационных ресурсов по заданной теме в единое информационное пространство



- обеспечение содержательного доступа и удобной навигации по информационному пространству

- обновление информации в соответствии с изменениями в предметной области

Информационный объект

Просмотр объекта

Проект	
Название деятельности	Проект AGILE
Описание деятельности	Automatic Generation of Instructions in Languages of Eastern Europe
Дата начала	1 января 1998
Дата окончания	31 декабря 2000

Связи объекта

Результат-Деятельности

Научный Результат
Система AGILE

Направление деятельности

Раздел Науки
Генерация текста

Ссылки на объект

Персона-Участник-Деятельности

Персона	Роль Участника Деятельности
Bateman, J.A.	исполнитель
Hana, J.	исполнитель
Hartley, T.	исполнитель
Kruijff, G.-J.	исполнитель
Kruijff-Korbayová, I.	исполнитель
(Всего: 10)	

Организация-Участник-Деятельности

Организация
Information Technology Research Institute, University of Brighton, ITRI
Institute for Applied Linguistics, University of the Saarland
Institute of Formal and Applied Linguistics(Charles University), ÚFAL
Institute of Information Technology, Bulgarian Academy of Sciences
РосНИИ искусственного интеллекта, РосНИИ ИИ

Публикация о Деятельности

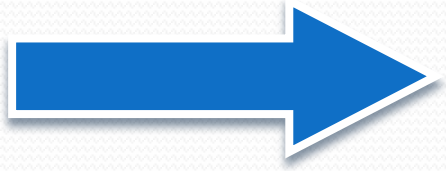
Публикация
Bateman, J.A., Hana, J., Hartley, T., Kruijff, G.-J., Kruijff-Korbayová, I., Skoumalová, H., Staykova, K., Teich, E., Соколова, Е.Г., Шаров, С.А., A multilingual system for text generation in three slavic languages, 2000, статья
Bateman, J.A., Kruijff, G.-J., Kruijff-Korbayová, I., Skoumalová, H., Teich, E., Шаров, С.А., Resources for multilingual text generation in three Slavic languages, 2000, статья

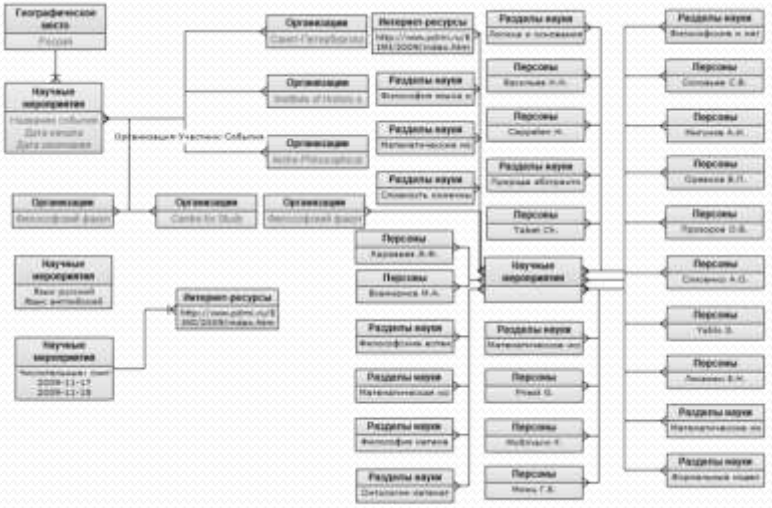
Ресурс-Деятельности

Ресурс
Сайт проекта AGILE

Входящий поток данных

- Новые данные поступают в виде семантической сети информационных объектов, извлекаемых из текстовых документов



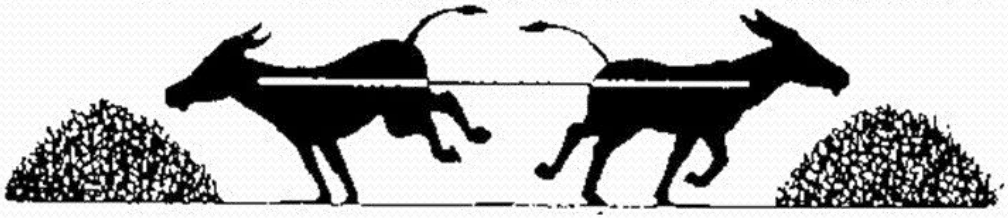


Динамический характер информации

- Накапливаемая в системе информация может оказаться не соответствующей действительности

- Может устареть

- Также, новая информация может вступить в противоречие с уже содержащейся в базе данных



Факты и доверие к ним

Что такое факт

Факт - минимальное знание об информационном объекте.

Для экземпляров понятий это будут значения его атрибутов и связи с другими объектами,

Для экземпляров отношений - значения атрибутов и аргументы.

Уровень доверия к фактам

Трастовая метрика учитывает характеристики источника фактов и время его существования в самой информационной системе.

Обработка входящих данных

Для оценки надежности фактов необходимо различать в поступающем потоке факты, эквивалентные уже содержащимся в системе. Для этого предусмотрен подготовительный этап обработки.

Обработка входящих данных

Референциальная эквивалентность объектов

- Вычисление сходства текущего объекта со всеми объектами, лежащими в его контексте,
- Построение множества потенциальных эквивалентов текущего объекта,
- Определение эквивалента текущего объекта,
- Объединение референциально эквивалентных объектов.

Идентификация объектов

- Идентификация по точному совпадению,
- Построение множеств похожих объектов,
- Построение фокусных множеств,
- Определение эквивалентных объектов.

Достоверность информации

Достоверность факта *F* зависит от известных источников, в которых он упоминается и от времени его существования в информационной системе.

$$Trust_F = Trust_F(D, t)$$

D – множество источников

t- время

Характеристики фактов и источников

Факт *F* упоминается в *N* источниках

$$F \subset \{d^1, d^2, ..., d^N\}$$

i-му источнику соответствует экспертная оценка

$$x^i \in [-1; 1]$$

Вероятность получения достоверного знания из *i*-го источника

$$\delta^i = \varphi\left(\frac{x^i + 1}{2}\right)^{\mathfrak{M}}$$

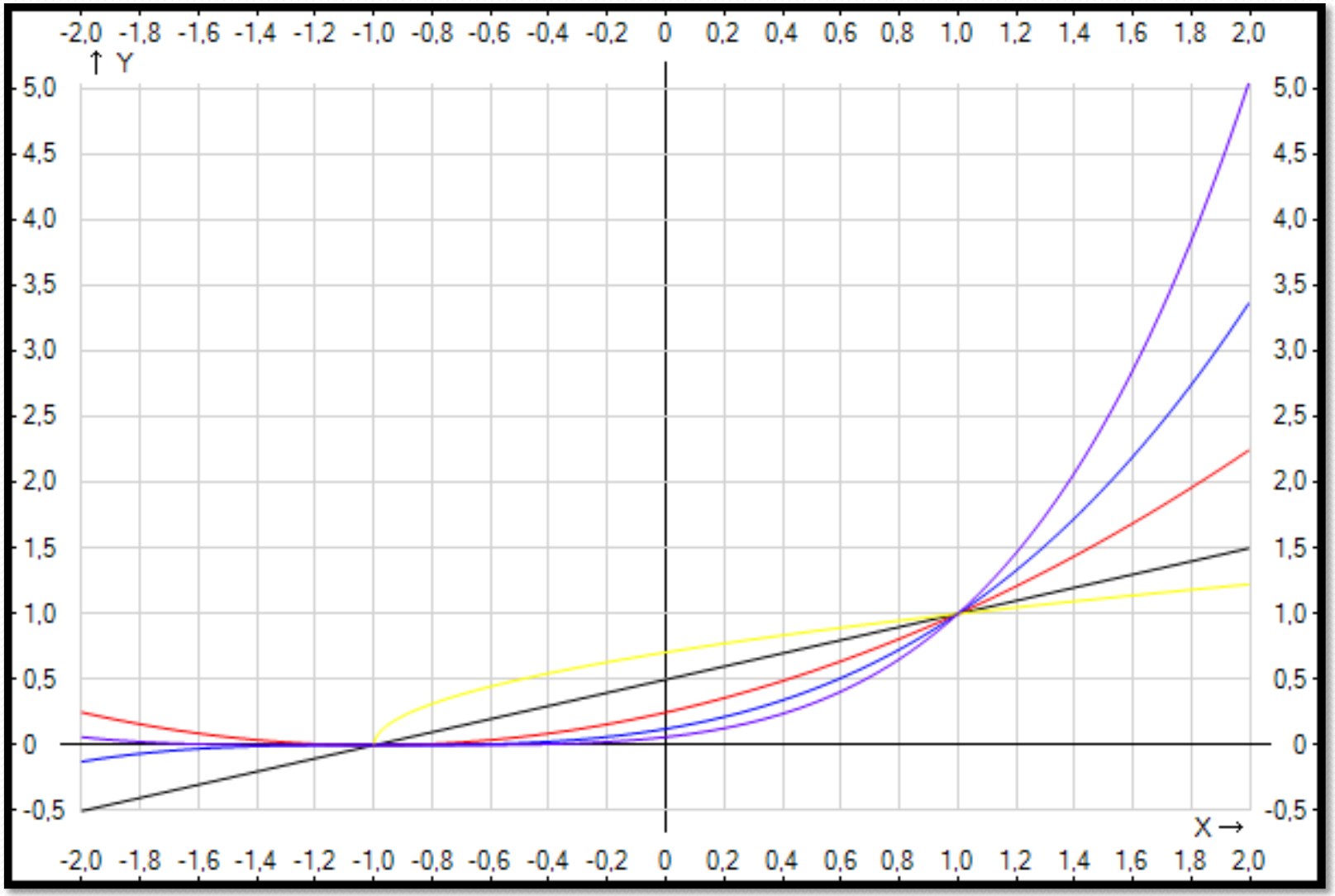
$$f(x) = \left(\frac{x+1}{2}\right)^{\mathfrak{M}}$$

Семейство характеристик *i*-го источника

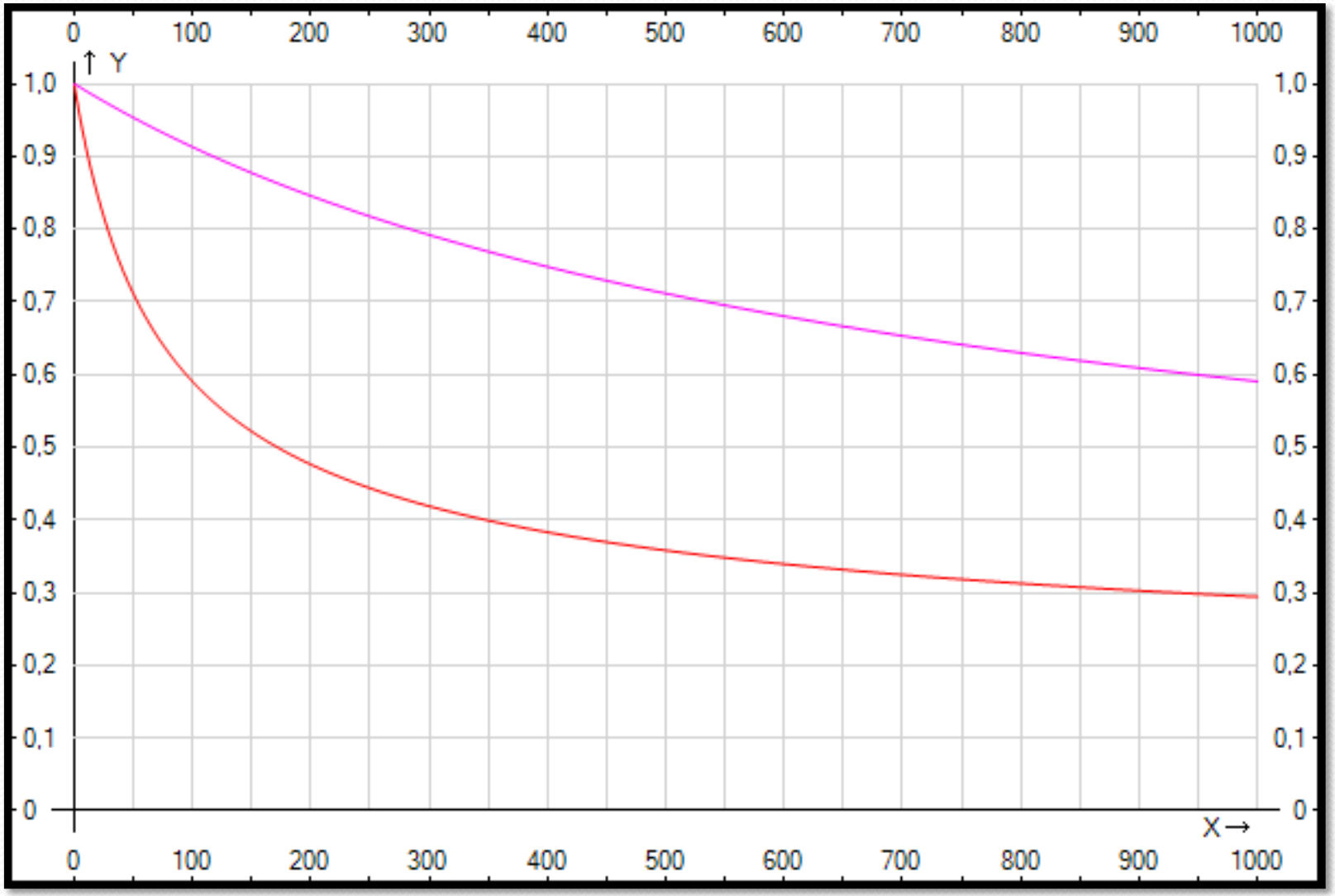
Функция от *t* для учета времени существования факта *F* в системе

$$h(t) = \frac{1}{1 + \ln\left(\frac{t}{\mathcal{M}} + 1\right)}$$

Семейство функций $f_{\mathfrak{M}}(x)$



Семейство функций $h(t)$



Случайный процесс

За основу модели достоверности факта взят неоднородный марковский процесс с тремя состояниями

E_1 – состояние недоверия

E_2 – состояние неопределенности

E_3 – состояние доверия

X_n – случайная величина, равная номеру состояния в момент времени n

Оценкой достоверности факта *F* считается вероятность того, что процесс находится в состоянии E_3

Вектор распределения X_n $\quad \pi^n = \pi^0 \cdot \mathbb{P}^{(n)}$ $\quad \pi^0 = (0,1,0), \pi_i^0 = P(X_0 = i)$

$$\mathbb{P}(n, n+1) = \begin{bmatrix} 1-\delta^{n+1} & 0 & \delta^{n+1} \\ 1-\delta^{n+1} & 0 & \delta^{n+1} \\ \frac{1-p_{33}}{2} & \frac{1-p_{33}}{2} & p_{33} \end{bmatrix}$$

$$p_{33} = \frac{1}{2}(p_{23}(0,n) + \delta^{n+1})$$

Влияние времени

Со временем вектор распределения изменяется

$$\vec{\pi} = (\pi_1, \pi_2, \pi_3) \quad \Rightarrow \quad \overrightarrow{\pi_t} = (\pi_1^t, \pi_2^t, \pi_3^t)$$

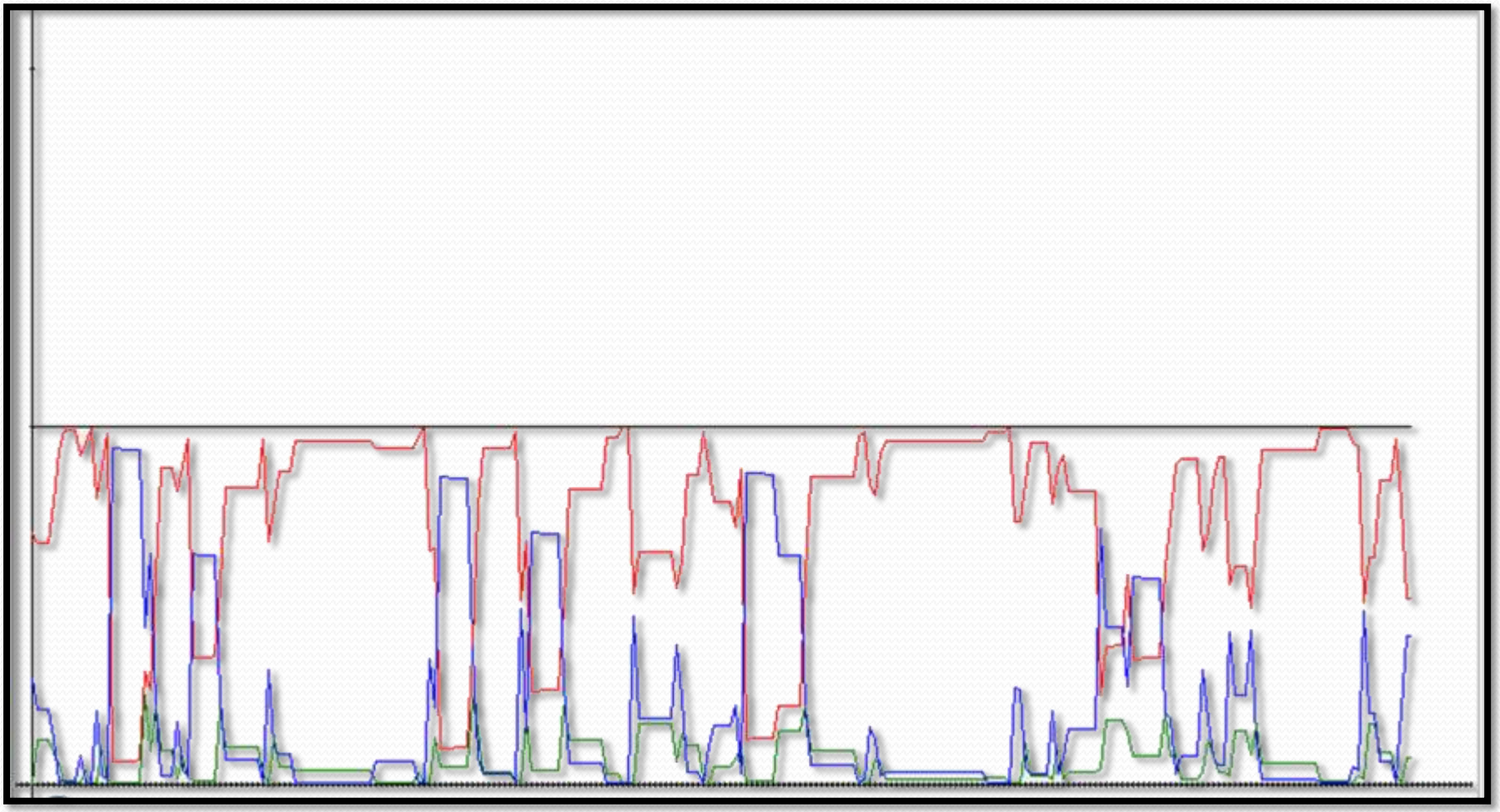
Потребуем, чтобы

$$\begin{cases} \pi_3^t = h(t) \cdot \pi_3 \\ \pi_1^t = \pi_1 + (1 - h(t)) \cdot \pi_3 \end{cases}$$

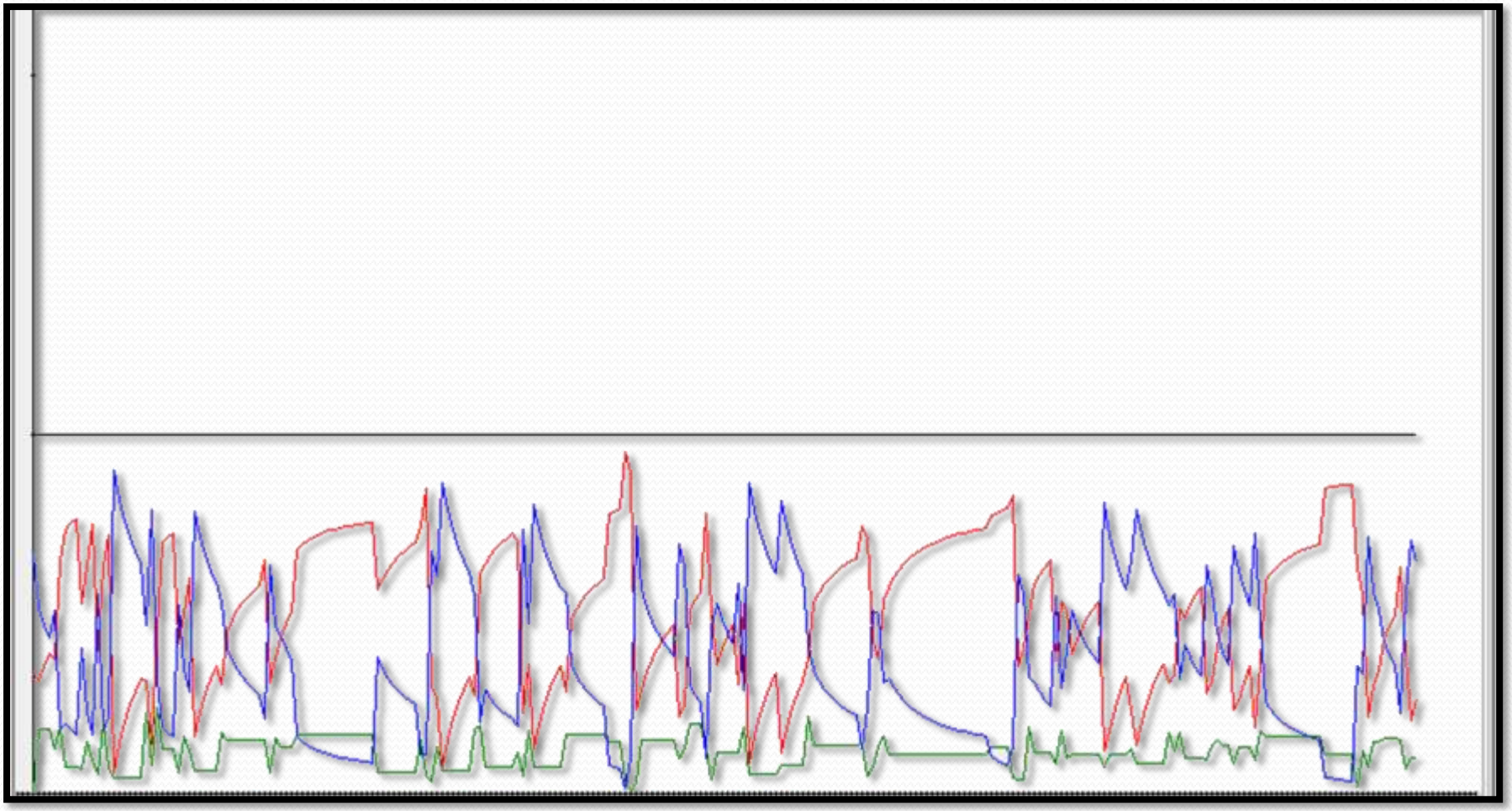
T_t – линейный оператор

$$T_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 - h(t) & 0 & h(t) \end{bmatrix}$$

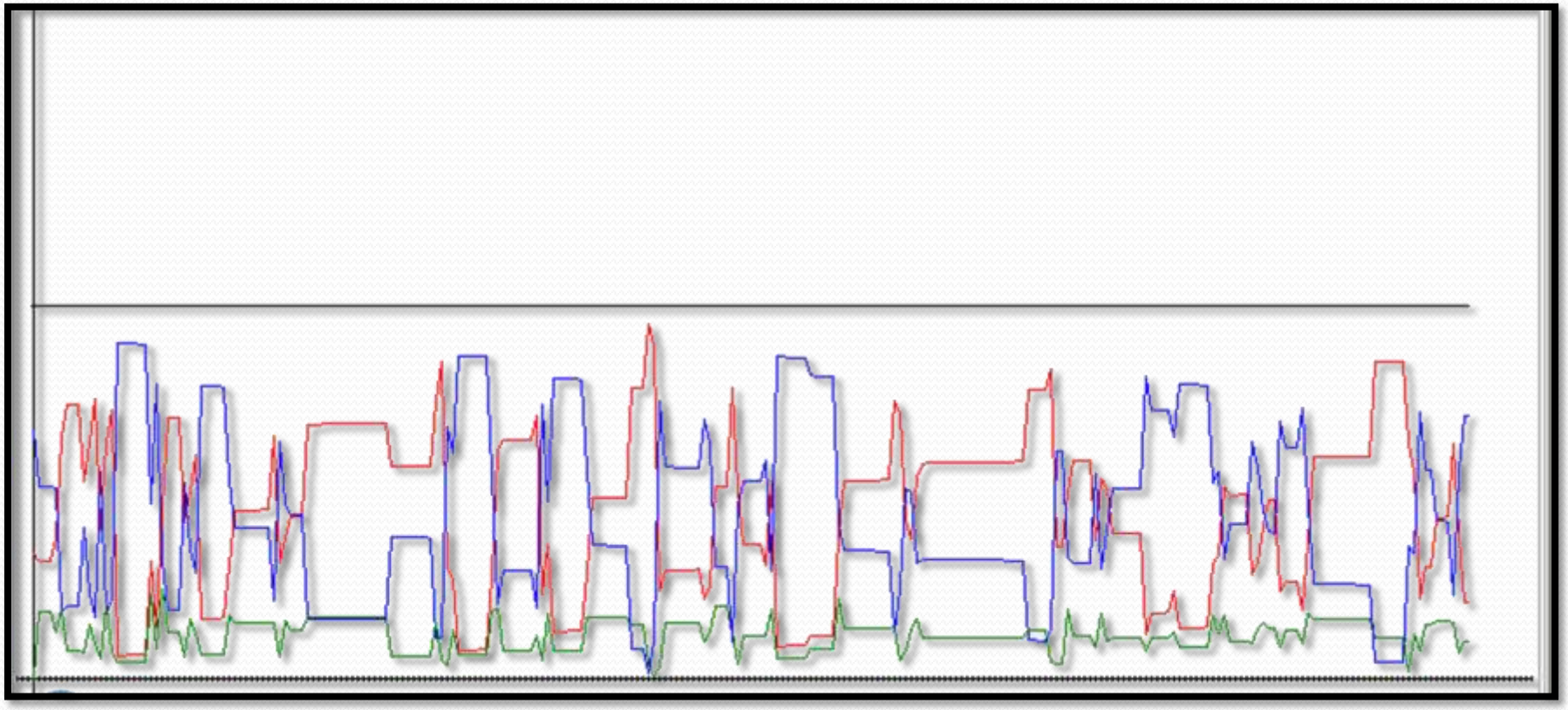
Распределение при $\mathcal{M} = 1000\ \mathfrak{M} = 3$



Распределение при $\mathcal{M} = 10\ \mathfrak{M} = 1$



Распределение при $\mathcal{M} = 1000\ \mathfrak{M} = 1$



Что делать с ненадежными данными?

Изменение достоверности факта *F* описывается дискретным множеством <*j*, $Trust_j$>, где *j* – момент времени, $Trust_j$ – достоверность на момент *j*.
Достоверность *F* может опуститься ниже минимально допустимого порога вследствие ряда причин.
Необходимо оценить значения в промежуточных точках путем аппроксимации дискретного множества.
Отдельные сегменты кривой аппроксимации позволяют оценить достоверность в окрестности заданного момента времени.

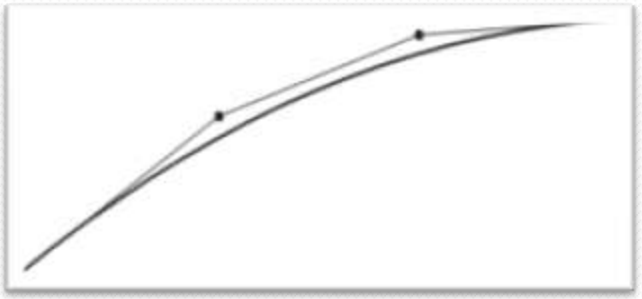
$$\boldsymbol{P}(u) = \sum_{i=0}^{n} \boldsymbol{P}_i N_{i,k}(u) \ (t_{min} \le t \le t_{max}) \qquad p = n + k + 1$$

$$N_{i,k}(u) = \frac{(u - t_i)N_{i,k-1}(u)}{t_{i+k-1} - t_i} + \frac{(t_{i+k} - u)N_{i+1,k-1}(u)}{t_{i+k} - t_{i+1}}$$

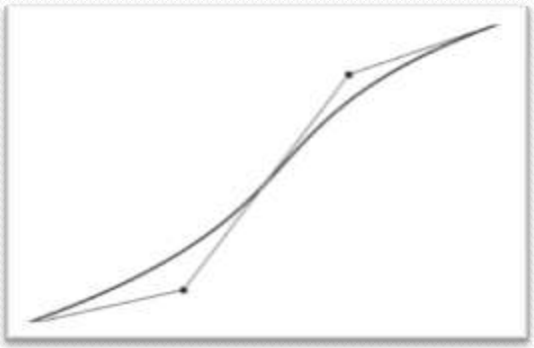
$$N_{i,1} = \begin{cases} 1, & t_i \le u \le t_{i+1} \\ 0, \text{ в противном случае} \end{cases}$$

Возможный вид хвоста

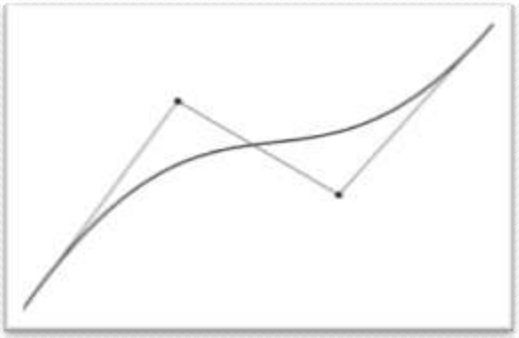
2.1.1



2.1.2

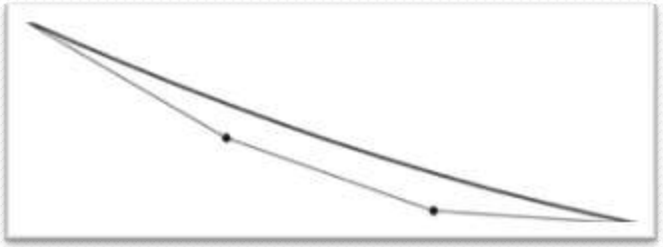


1.1.1



1.1.2

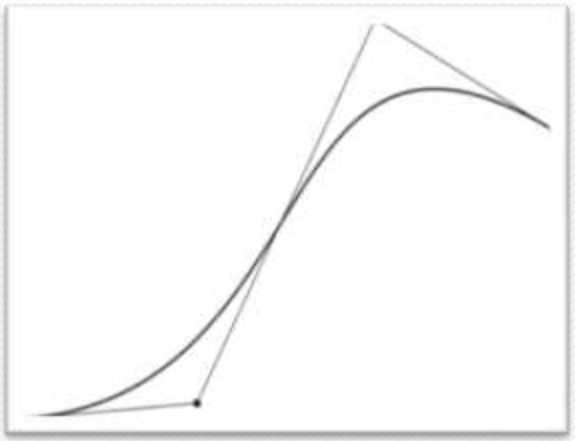
2.2.1



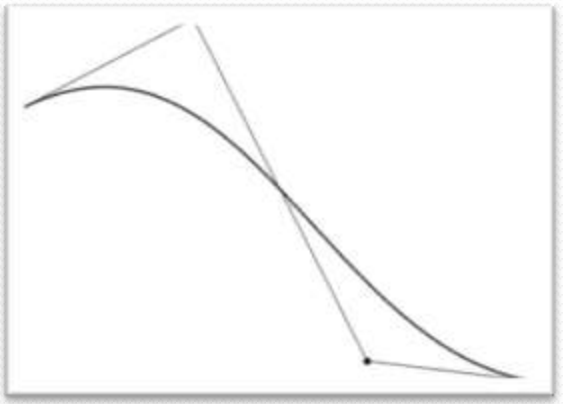
2.2.2

Возможный вид хвоста

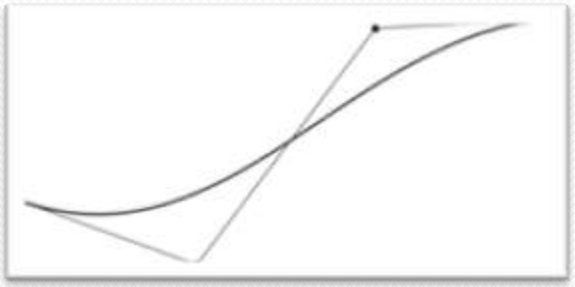
1.3.1



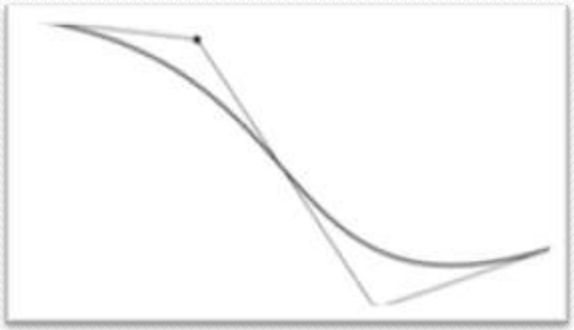
1.3.2



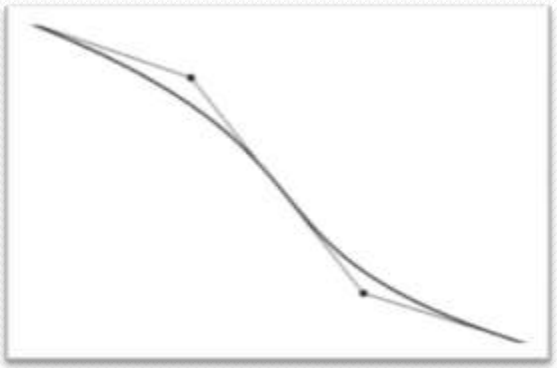
1.4.1



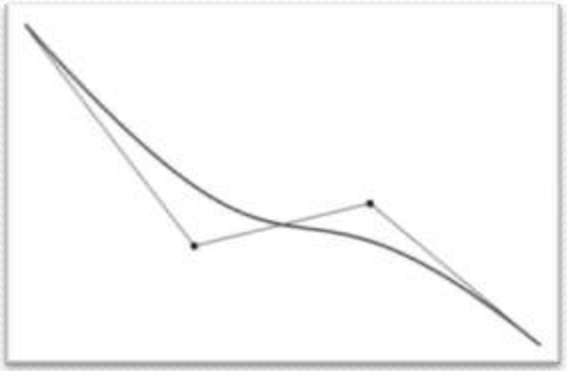
1.4.2



1.2.1

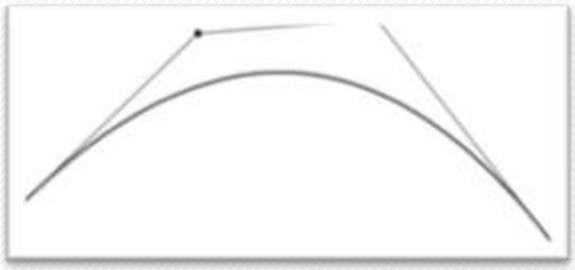


1.2.2



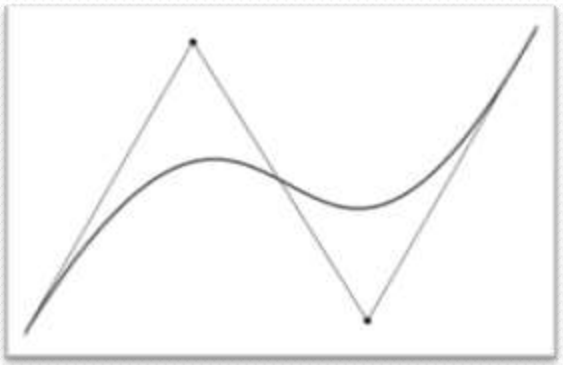
Возможный вид хвоста

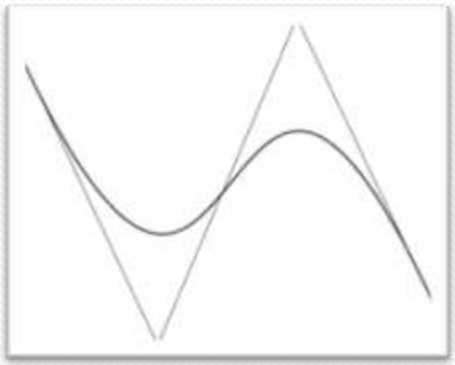
2.3.2



2.3.1

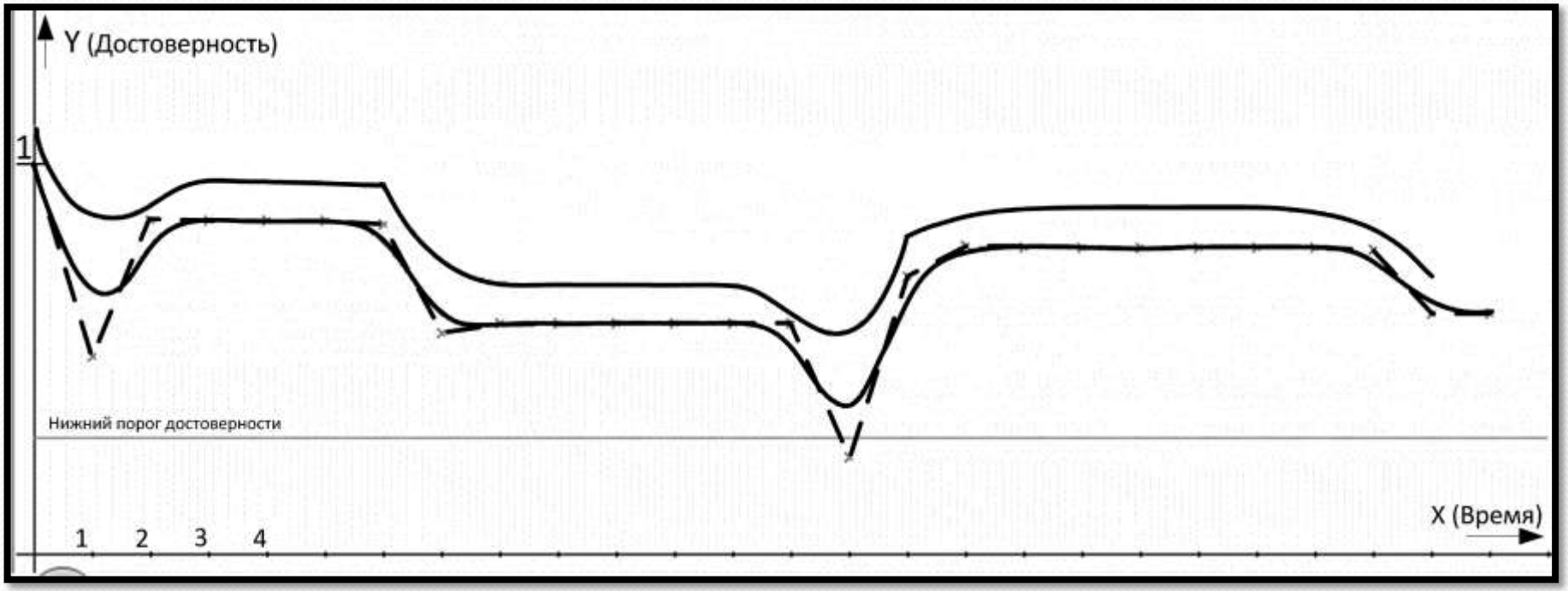
1.6.2





1.5.1

Пример кусочной аппроксимации



В заключение

Данный подход позволяет

- Провести предварительную обработку входящих данных в информационной системе, предметная область которой формально задана онтологической моделью
- Оценить достоверность содержащихся в информационной системе данных
- Удалить недостоверную информацию

Благодарю за внимание !