

АВТОСТРУКТУРИЗАЦИЯ НЕПРЕРЫВНОГО ТЕКСТОВОГО ПОТОКА

(Априорно неопределенной предметной области)

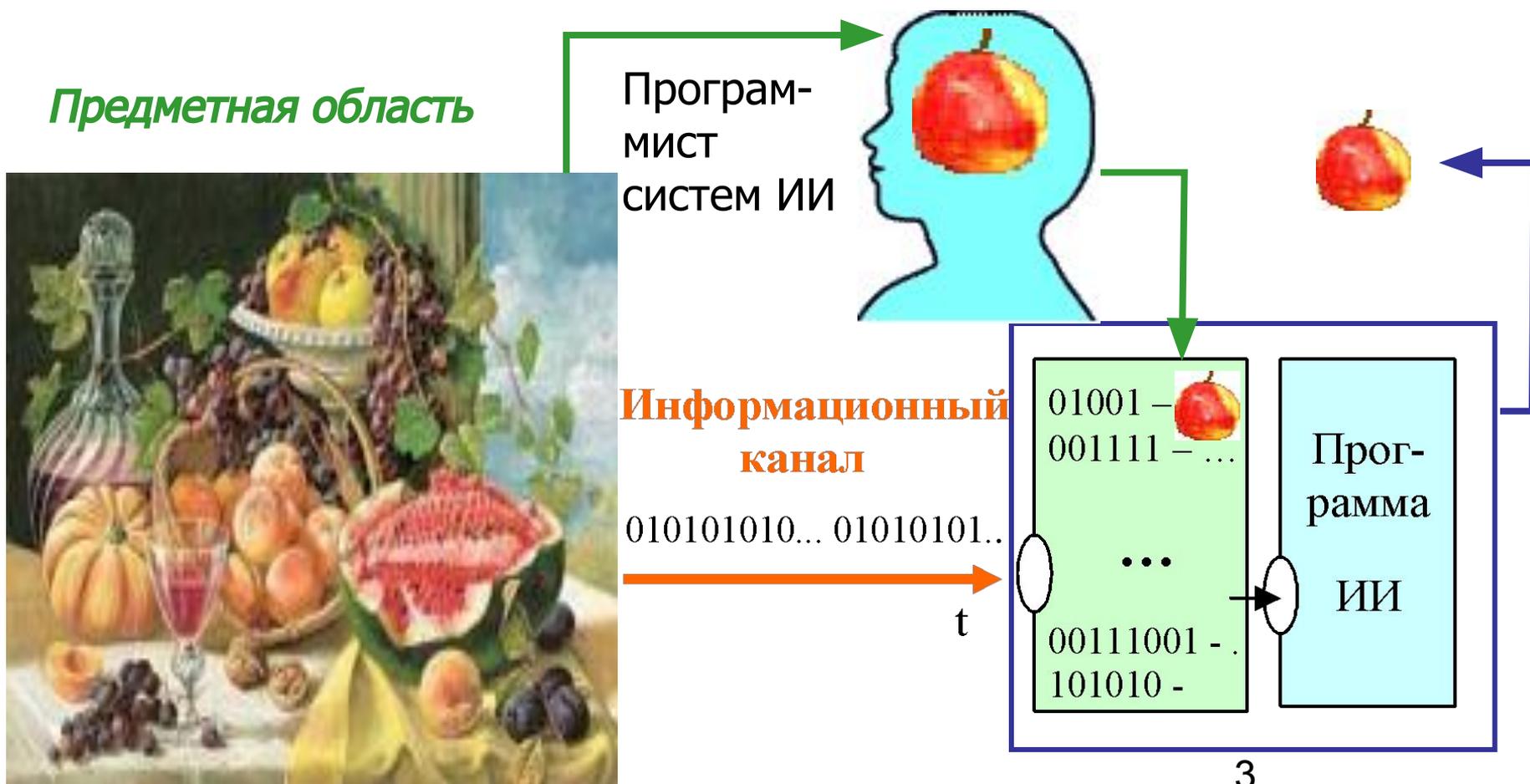
Бодякин В.И. к.ф.-м.н. с.н.с.
Институт проблем управления РАН
им. В.А. Трапезникова, Москва

E-mail: body@ipu.ru ,
<http://www.informograd.narod.ru> ,
служ.тел.:334-92-39

ПОЧЕМУ ВСЕ ОРГАНИЗМЫ "ЕДИНОДУШНЫ" В КЛАСТЕРИЗАЦИИ ОКРУЖАЮЩЕГО НАС МИРА НА ОТДЕЛЬНЫЕ ОБРАЗЫ ?



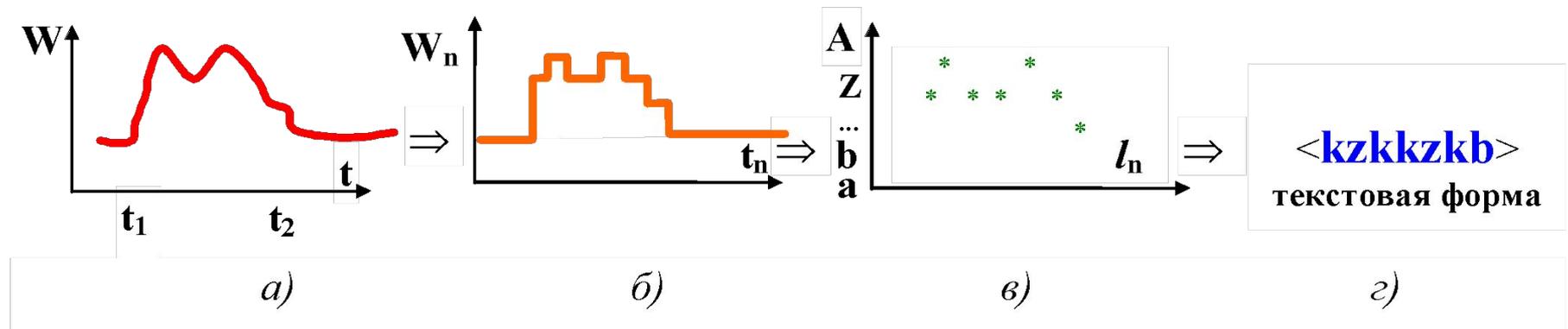
Традиционный способ структуризации в системах ИИ



Предметная область (ПО) – причинно-связанная совокупность физических процессов.

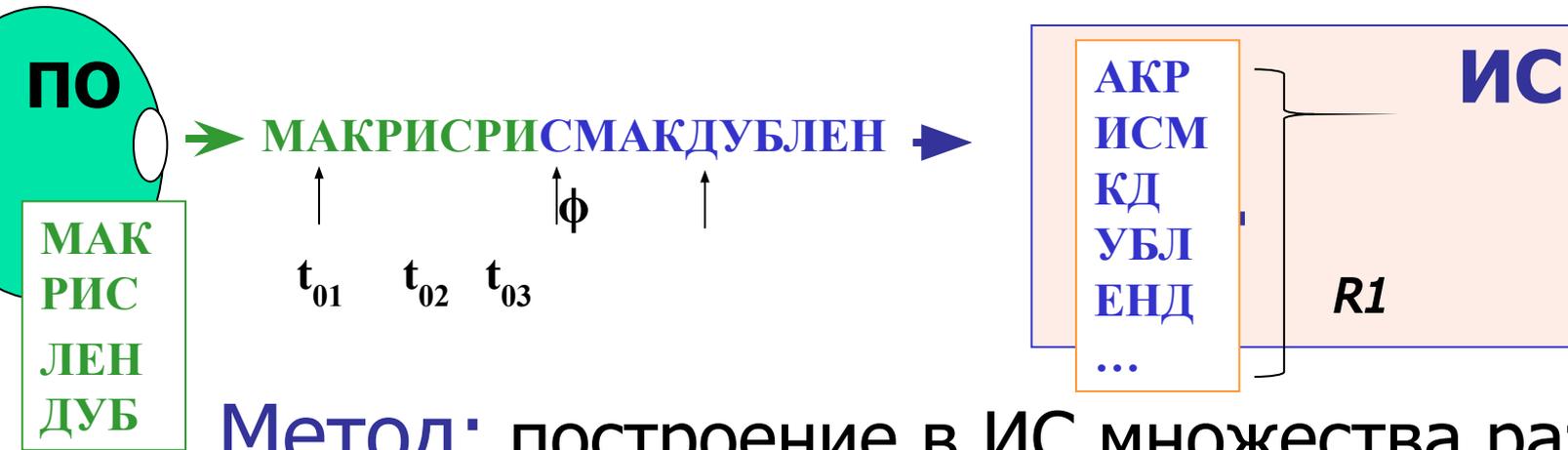
Процесс – независимое от времени и пространства детерминированное изменение некоторого физического параметра во времени.

Любой процесс может быть преобразован в **текстовую форму**.



многомерный физический процесс \leftrightarrow **текстовая форма**

Задача: В непрерывном потоке ТФ необходимо выделить образы, соответствующие процессам любой ПО



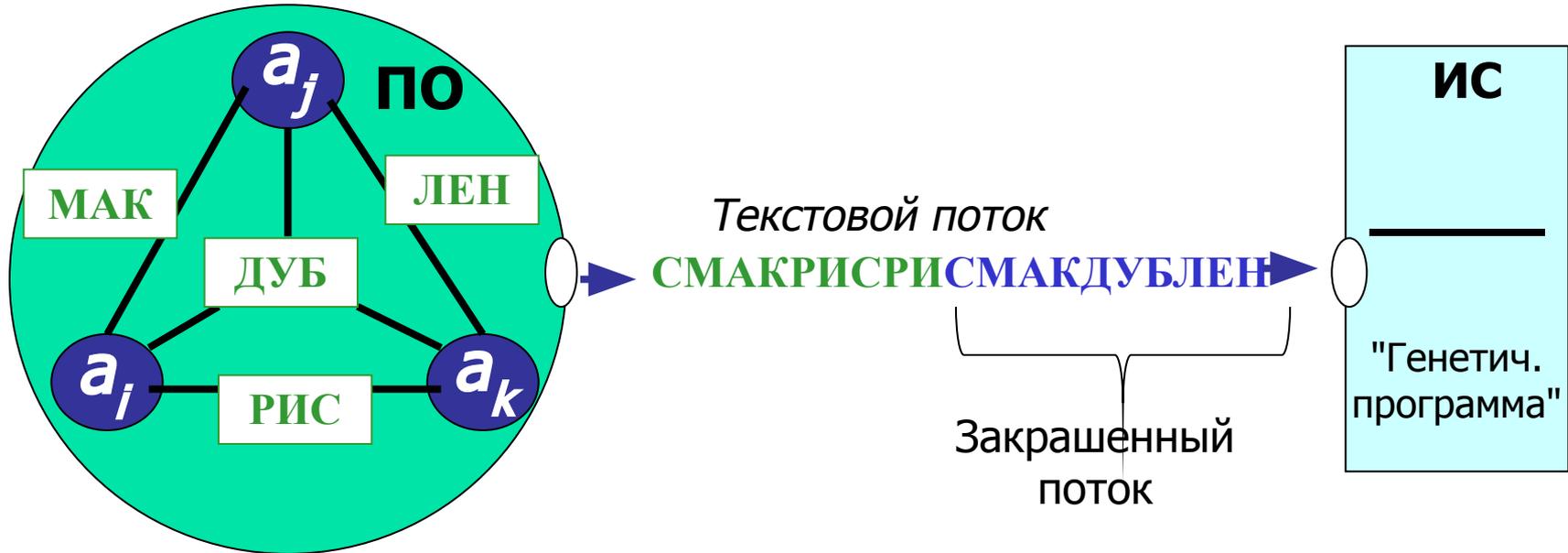
Метод: построение в ИС множества различных словарей и выбор минимального (R_i)

Цель: Минимальный словарь ИС
(гомоморфен процессам ПО)

Инструментарий: нейросемантические структуры

Демонстрационный пример

(четыре равновероятных процесса формируют непрерывный поток ТФ)



Необходимо построить словарь в N образов, **полностью покрывающий ТФ**. Примем что:

Энергетические затраты на обработку одного образа в ИС равна $1E-$.

Время обработки символа и образа – **один такт T** ,

Энергетические затраты на хранение одного образа в памяти $1/24 E-$.

Энергетика каждого прогнозируемого ИС символа равна $1E+$,

Усредненный на 12 тактов энергетический баланс ИС =

$$-N(\text{обработка}) - 1/2N(\text{хранение}) + (12-N)(\text{прогнозирование}) = \mathbf{(12 - 5/2N)*E}$$

Результаты эксперимента

Первая структуризация словаря ИС

Форма словаря: (наибольший размер образа один символ):

- $\langle M \rangle \langle A \rangle \langle K \rangle \langle P \rangle \langle I \rangle \langle C \rangle \langle L \rangle \langle E \rangle \langle H \rangle \langle D \rangle \langle U \rangle \langle B \rangle$,
- $R(\text{ИС}) = 12R * (0,5E^- / R) = 6E^-$,
- прогнозирование (Т) = 0Т (т.к. у образа только один символ),
- затраты энергии на распознавание = $12E^-$
- контролирование потенциальной энергии ТФ = $0E^+$.

Итог: для односимвольного словаря усредненный энергетический баланс = $18E^-$.

Усредненный энергетический баланс **ИС** назовем
ЭВОЛЮЦИОННЫМ ПОТЕНЦИАЛОМ ИС

Вторая структуризация словаря ИС

(наибольший размер образа два символа):

- а) Минимальная форма словаря:
<МА><К><РИ><С><ЛЕ><Н><ДУ><Б>:
- $R(\text{ИС})=8R = 4E^-$, прогнозирование = $0,5T$ (на образ),
- затраты энергии на распознавание = $8E^-$,
- контролирование потенциальной энергии $T\Phi = 4E^+$.
- *Итог а)* $4E^- + (0,5T * 8(\text{образов на } T\Phi=12) = 4E^+) + 8E^- = 8E^-$.

- б) Максимальная форма (без полного пересечения):
<МА><КЛ><КР><КД><КМ> <ЛЕ>...<БД>:
- $R(\text{ИС})= 20R=10E^-$, прогнозирование = $0,2T$ (на образ),
- затраты энергии на распознавание = $20E^-$,
- контролирование потенциальной энергии $T\Phi = 4E^+$.
- *Итог б)* $10E^- + (0,2T * 20(\text{образов на } T\Phi=12) = 4E^+) + 20E^- = 26E^-$.

- **Итоговый лучший эволюционный потенциал = $8E^-$.**

Третья структуризация словаря

(наибольший размер образа в три символа).

- а) Минимальная форма словаря: <МАК><РИС><ЛЕН><ДУБ>:
- $R(ИС) = 4R = 2E^-$, прогнозирование = $2T$ (на образ),
- затраты энергии на распознавание = $4E^-$,
- контролирование потенциальной энергии $T\Phi = 8E^+$.
- *Итог а) $2E^- + (2T * 4(\text{образов на } T\Phi=12) = 8E^+) + 4E^- = 2E^+$.*

- б) Максимальная форма (без дублирования, т.е. без полного пересечения): <АКР><ИСП><ИСМ><АКД> <УБЛ><ЕНД> ... <АКМ>:
- $R(ИС) = 36R = 18E^-$,
- прогнозирование (T) = $+4E$,
- затраты энергии на распознавание = $36E^-$,
- контролирование потенциальной энергии $T\Phi = 0E^+$.
- *Итог б) = $50E^-$.*

- Лучший итоговый эволюционный потенциал = $2E^+$, $50E^- \rightarrow 2E^+(!!)$.

Четвертая структуризация словаря

(наибольший размер образа в четыре символа)

- а) Минимальная форма словаря: <МАК><РИС><ЛЕН><ДУБ>:
- $R(ИС)=4R = 2E^-$, прогнозирование = $2T$ (на образ),
- затраты энергии на распознавание = $4E^-$,
- контролирование потенциальной энергии $T\Phi = 8E^+$.
- *Итог а)* $2E^- + (2T * 4(\text{образов на } T\Phi=12) = 8E^+) + 4E^- = 2E^+$.

- б) Максимальная форма (без дублирования):
<АКРИ><ИСРИ><ИСМА> <АКДУ><УБЛЕ><ЕНДУ> ... <АКМА>:
- $R(ИС)= 48R=24E^-$, прогнозирование (T) = $0T$ (на образ),
- затраты энергии на распознавание = $48E^-$,
- контролирование потенциальной энергии $T\Phi = 0,5E^+$.
- *Итог б)* = $78E^-$.

- Лучший итоговый эволюционный потенциал = $2E^+$, $78E^- \rightarrow 2E^+$

Пятая и другие структуризации словаря

Лучший итоговый эволюционный потенциал =
 $2E+$, $204E-$ → $2E+$

Шестая структуризация словаря, седьмая ... и т.д. → $2E+$!!!

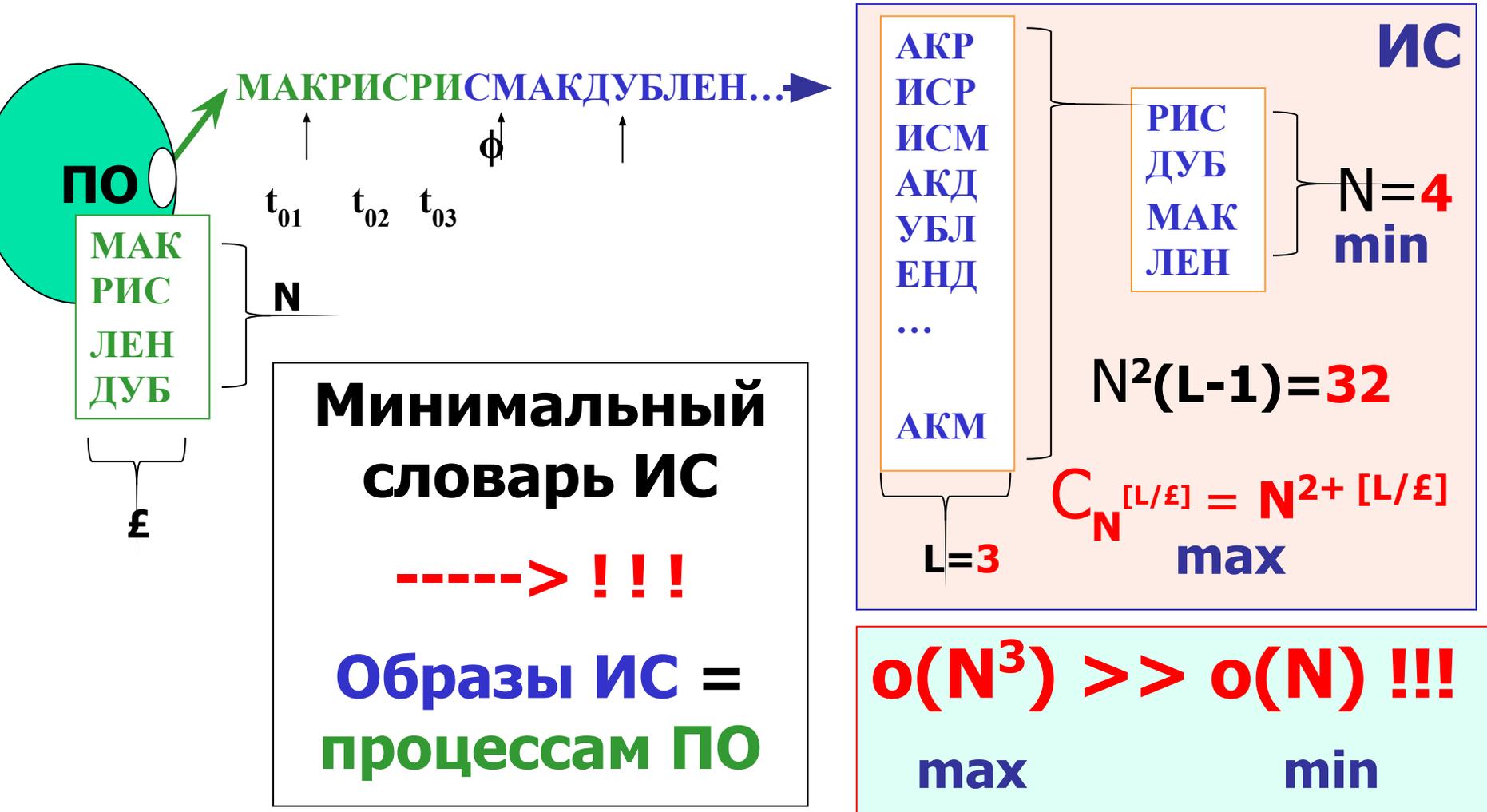
Худший - $\infty E-$

- Теоретический *анализ результатов эксперимента* показывает, что **эволюционный потенциал ИС** *обратно пропорционален размеру словаря*,
- Размеры минимальных и максимальных словарей ИС соотносятся как:

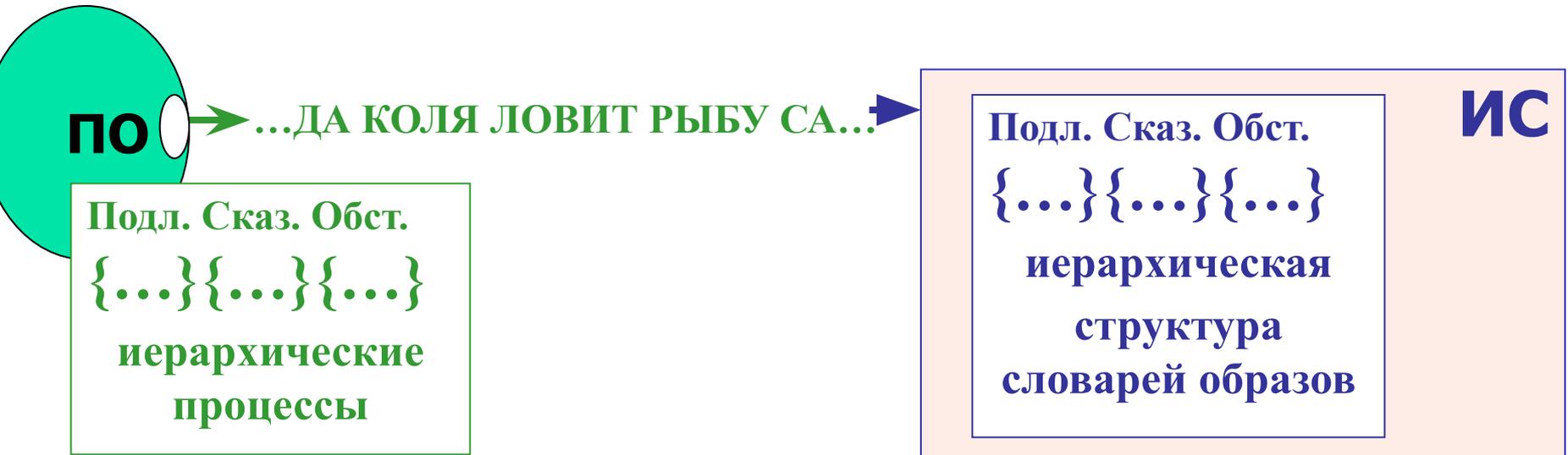
$O(N)$ и $O(N^3)$!!!

где: N – максимальный размер образа словаря

Автоструктуризация информации в ИС



Автоструктуризация иерархических процессов



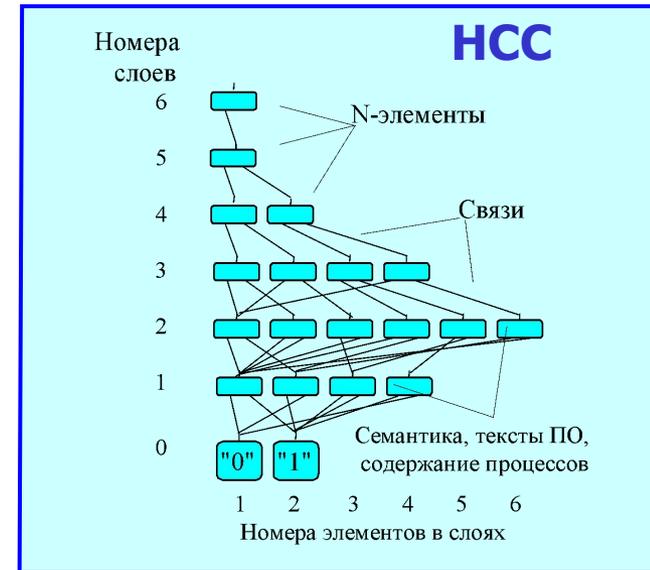
При минимизации словаря на нейросемантических структурах, его топология гомоморфна структуре исходных процессов ПО

Нейросемантическая форма представления информации



Алгоритм
НСС

11011010001101101001 □-□



Автоматическое выделение образов-процессов из предметной области в нейроподобные элементы НСС при минимизации ресурсных затрат (памяти)

N-элемент (образ НСС) ↔ процесс предметной области

Автоструктуризация на нейросемантических структурах

$R_{ис} = f(\text{число } N\text{-элементов, число связей})$ в битах

$1/P$ (компрессия) = ----- $\rightarrow 0$
 при $t \rightarrow \infty$ $T\Phi_{ис} = \text{объем текстовой информации в ИС}$ в битах
 или $\Delta R_{ис} / \Delta T\Phi_{ис} \rightarrow 0$ и $\Delta R_{ис} \rightarrow \text{const}$, при $t \rightarrow M$
 при $t \rightarrow \infty$

Примеры:

а) <RISMAKDUBLENLENDUBMAKMAKLENRISRISLENMAKRISDUBRIS>

правильно выделяются все **процессы**: <МАК><RIS><МАК><DUB>;

б) <ДОМЗЕБРЫСКИТНАДОМДОМВНАДОМВСКИТВНАСКИТВВЗЕБРЫНАВНА> ,

правильно выделяются все **процессы**: <ЗЕБРЫ><СКИТ><ДОМ><НА>

<В>.

сдвиг алфавита А в кодах ASCII в примере б)

<ЕПНИЖВСЬТЛЙУОБЕПНЕПНГОБЕПНГТЛЙУГОБТЛЙУГЖВСЬОБГОБ> на +1

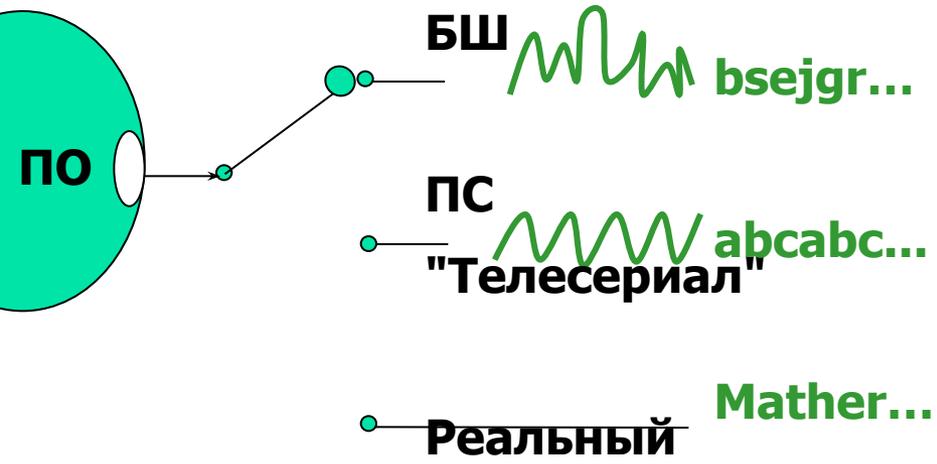
<?IGB@<KVLECMH;?IG?IG=H;?IG=LECM =H;LECM==B<KVH;=H;{> на -133.

НСС – это пример **1-го формального преобразования** количественной текстовой формы представления информации **в качественно новую форму** – **структуру образов ИС**



Критерии достаточности: а) все пространство состояний;
 б) если человек может правильно структурировать данный текстовой материал в непривычной, но взаимнооднозначной нотации,
 в) наличие характерных особенностей динамического процесса при минимизации ресурса $R_{ИС}$

По минимальной функции затрат ресурсов ИС $R_{ИС}$ можно объективно определять процессы ПО

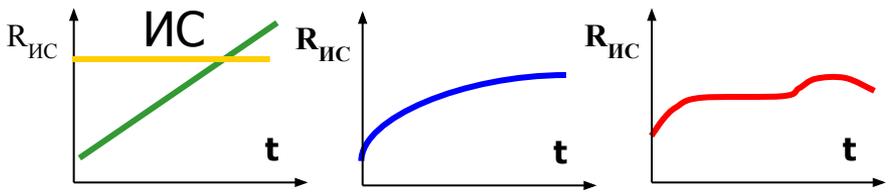


Автомат "животн." ИС-человек

| | | |
|--------|------------|--------|
| Сигнал | С | С |
| С | Информация | И |
| С | И | Знание |

При $T_{ИС} = \text{const} (t)$

$f = (R_{ИС} (t)) :$



линейная; логарифмическая; const; функции затрат $R_{ИС}$

Определения:

Информация – знаковая последовательность на языке системы, соответствующая целому числу причинно-связанных процессов ПО

...

Теорема: минимальная форма словаря ИС может достигаться только при ее изоморфности исходной структуре процессов ПО

Доказательство: Если представить формирование текстовой формы двумя независимыми и непересекающимися процессами А и Б, то становится очевидным,



что минимальным словарем образов закрашивающим эти два процесса могут быть только образы совпадающие по текстовой форме с генерирующими их процессами.

На вопрос: "Почему все организмы 'единодушны' в кластеризации окружающего нас мира на отдельные образы?"

Вытекает ответ: "Т.к. минимальный словарь, дает эволюционные преимущества, то все ИС данной ПО выбирают его, а соответственно, и его образы".