АВТОСТРУКТУРИЗАЦИЯ НЕПРЕРЫВНОГО ТЕКСТОВОГО ПОТОКА

(Априорно неопределенной предметной области)

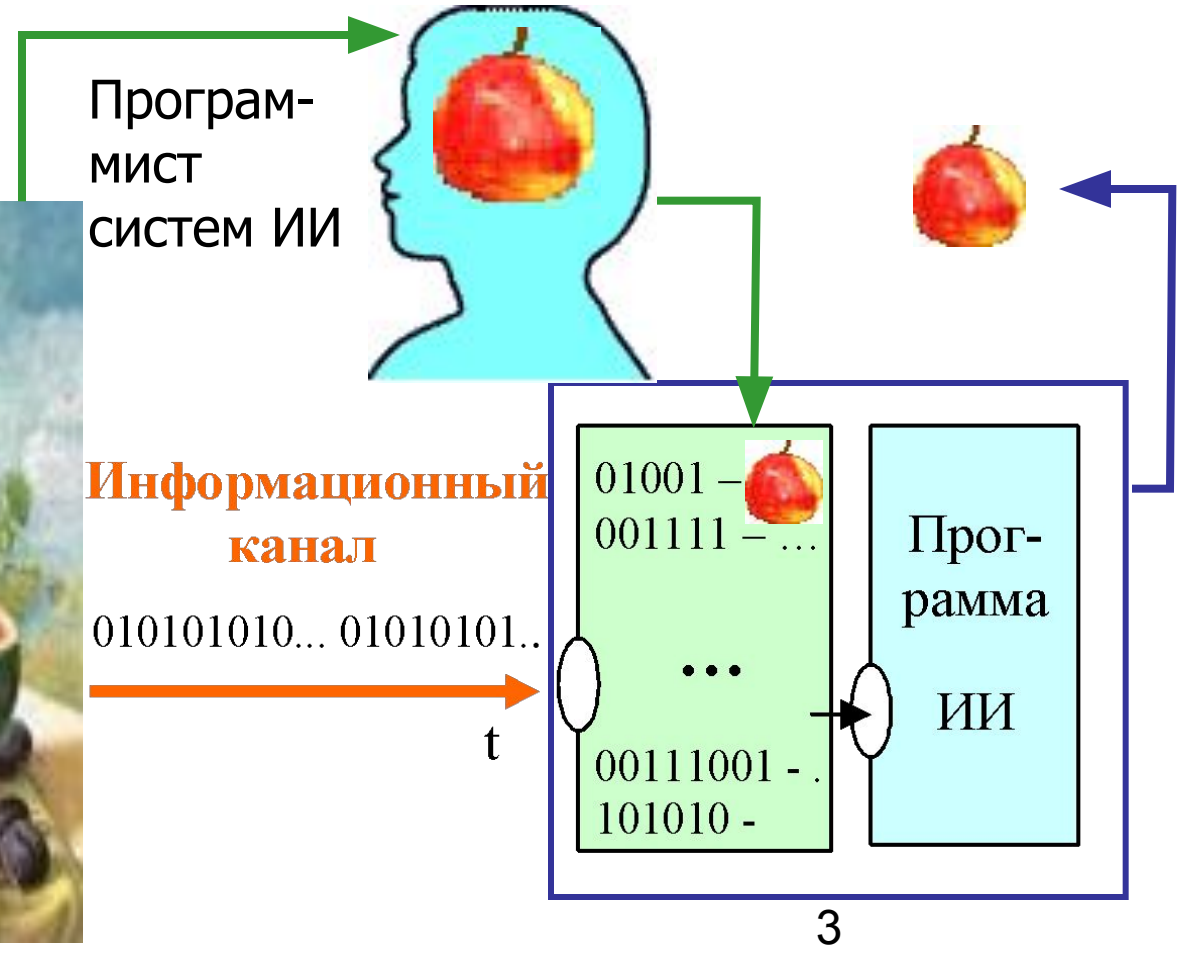
Бодякин В.И. к.ф.-м.н. с.н.с.
Институт проблем управления РАН
им. В.А. Трапезникова, Москва

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

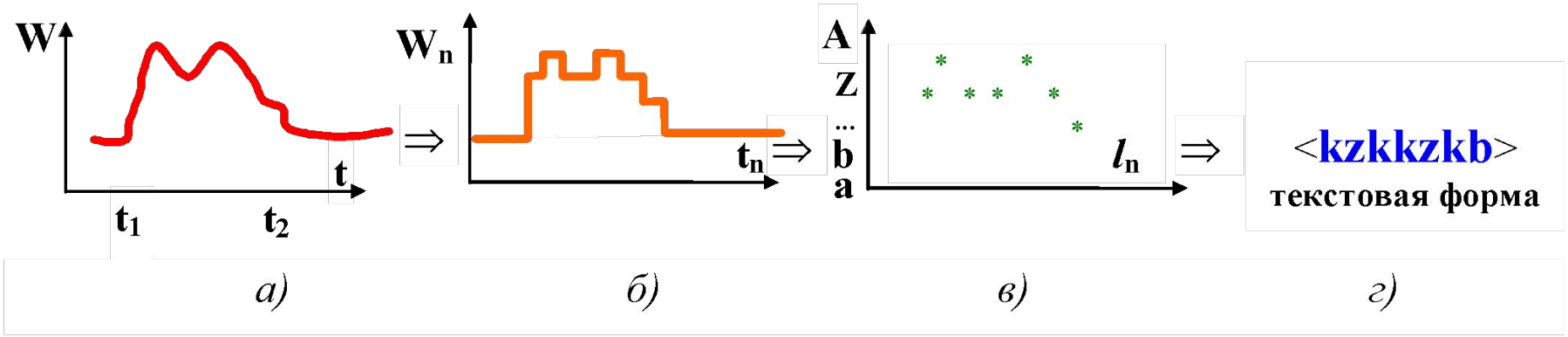
Предметная область



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

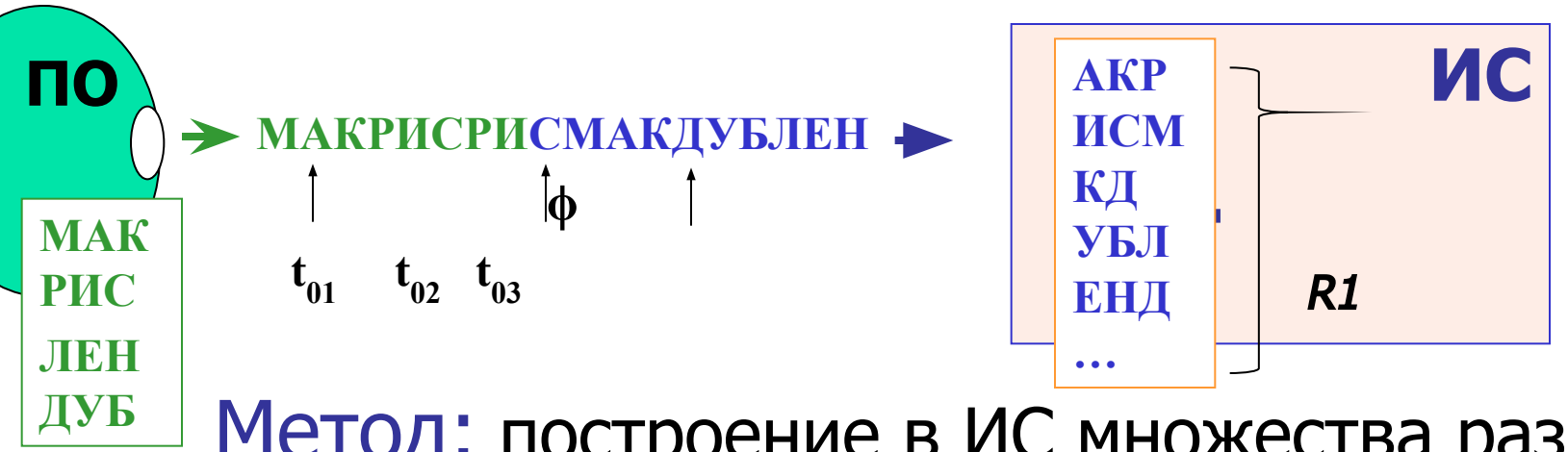
Процесс – независимое от времени и пространства детерминированное изменение некоторого физического параметра во времени.

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



многомерный физический процесс **<-> текстовая форма**

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



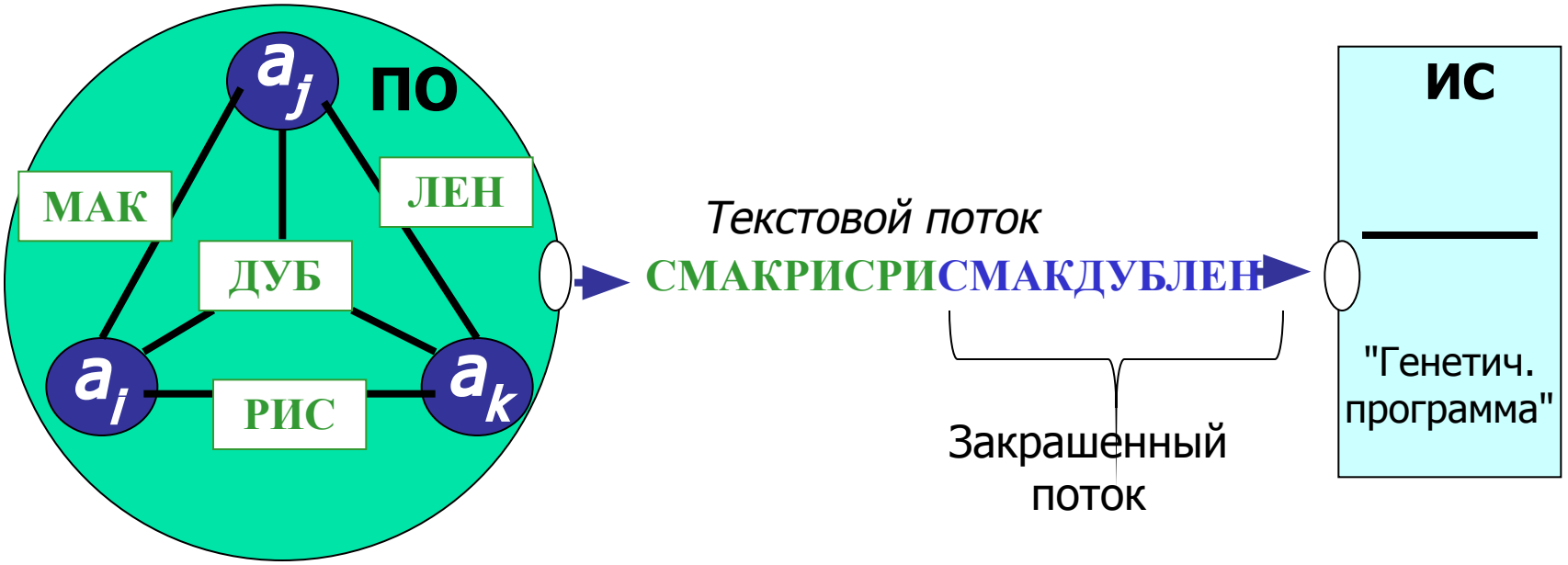
Метод: построение в ИС множества различных словарей и выбор минимального (R_i)

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(обработка) -1/2N(хранение) + (12-N)(прогнозирование) = **(12 - 5/2N)*E**

Результаты эксперимента

Первая структуризация словаря ИС

Форма словаря: (наибольший размер образа один символ):

- <М><А><К><Р><И><С><Л><Е><Н><Д><У><Б>,
- R(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(ИС)=8R =4E-, прогнозирование = 0,5T (на образ),
- затраты энергии на распознавание = 8E-,
- контролирование потенциальной энергии ТФ = 4E+.
- *Итог а)* 4E- + (0,5T * 8(образов на ТФ=12) = 4E+) + 8E- = 8E- .

- б) Максимальная форма (без полного пересечения):
  <МА><КЛ><КР><КД><КМ> <ЛЕ>…<БД>:
- R(ИС)= 20R=10E-, прогнозирование = 0,2T (на образ),
- затраты энергии на распознавание = 20E-,
- контролирование потенциальной энергии ТФ = 4E+.
- *Итог б)* 10E- + (0,2T * 20(образов на ТФ=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 ТФ = 8E+.
- *Итог а) 2*E- + (2T * 4(образов на ТФ=12) = 8E+) + 4E- = 2E+.

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

- Лучший итоговый эволюционный потенциал = 2E+ , 50E- → 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 ТФ = 8E+.
- *Итог а) 2*E- + (2T * 4(образов на ТФ=12) = 8E+) + 4E- = 2E+.

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

- Лучший итоговый эволюционный потенциал = 2E+ , 78E- → 2E+

Пятая и другие структуризации словаря

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

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

Худший - ∞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) !!!

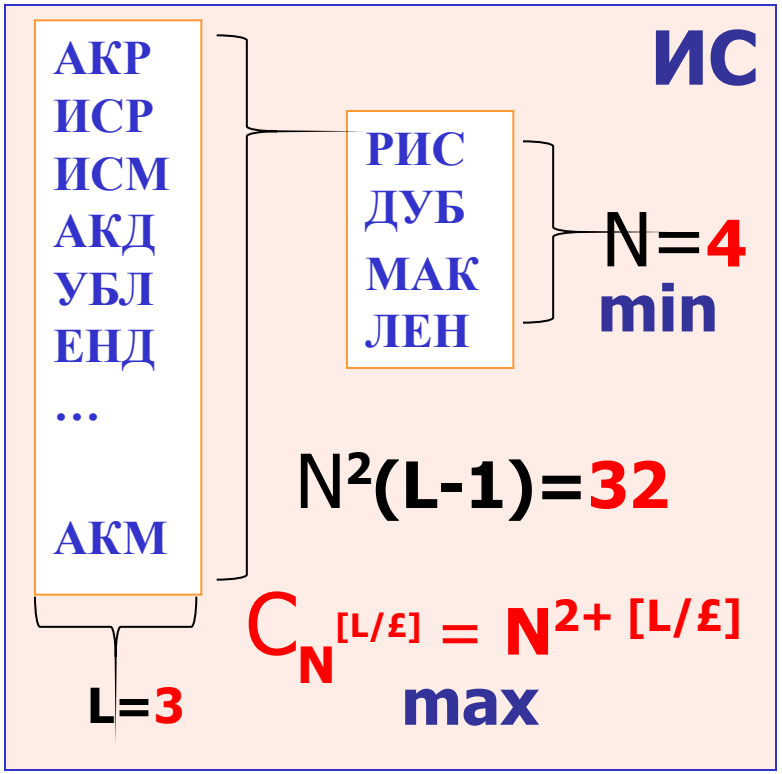
где: N – максимальный размер образа словаря

Автоструктуризация информации в ИС

Минимальный словарь ИС

-----> ! ! !

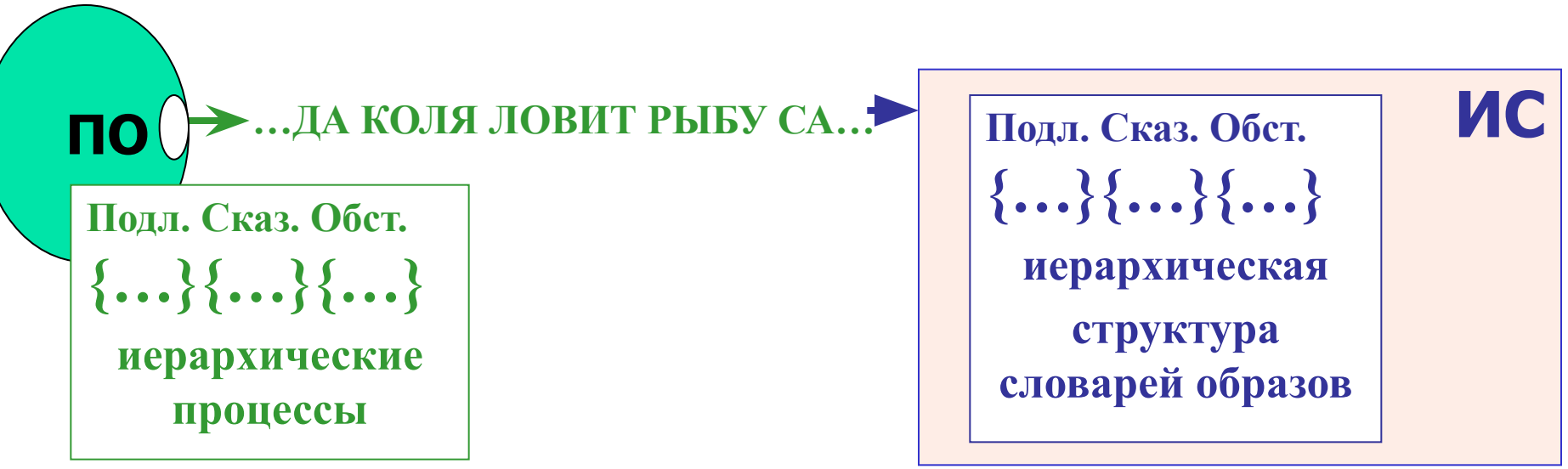
Образы ИС = процессам ПО



$o(N^3)$ >> $o(N)$!!!

max **min**

Автоструктуризация иерархических процессов

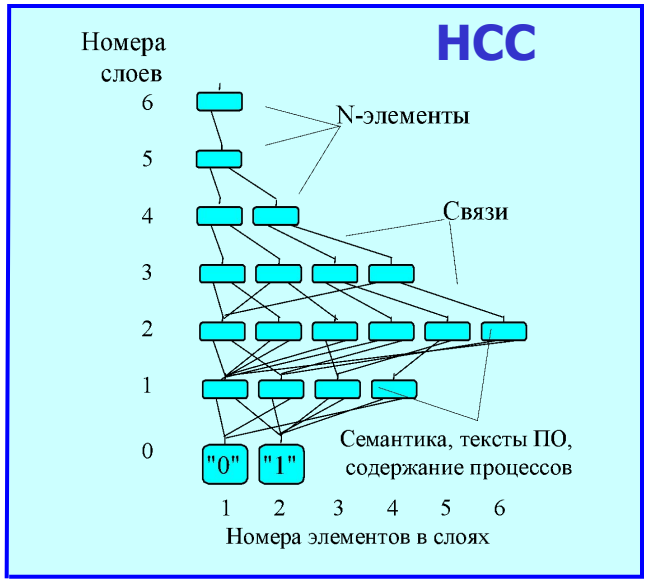


При минимизации словаря на нейросемантических структурах, его топология гомоморфна структуре исходных процессов ПО

Нейросемантическая форма представления информации

Алгоритм НСС

1101101000110110100 1□-□



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(число N-элементов, число связей)** *в битах*

1/P (компрессия) = ---------- → **0**

при t → ∞ **$ТФ_{ИС}$ = объем текстовой информации в ИС** *в битах*

или **$\Delta R_{ИС}$ / $\Delta ТФ_{ИС}$** → **0** и **$\Delta R_{ИС}$** → **const**, при t → **M**

при t → ∞

Примеры:

а) <RISMAKDUBLENLENDUBMAKMAKLENRISRISLENMAKRISDUBRIS>
правильно выделяются все **процессы**: <MAK><RIS><MAK><DUB>;

б)<ДОМЗЕБРЫСКИТНАДОМДОМВНАДОМВСКИТВНАСКИТВВЗЕБРЫНАВНА>,
правильно выделяются все **процессы**:<ЗЕБРЫ><СКИТ><ДОМ><НА><В>.

сдвиг алфавита А в кодах ASCII в примере б)

<ЕПНИЖВСЬТЛЙУОБЕПНЕПНГОБЕПНГТЛЙУГОБТЛЙУГЖВСЬОБГОБ> на +1

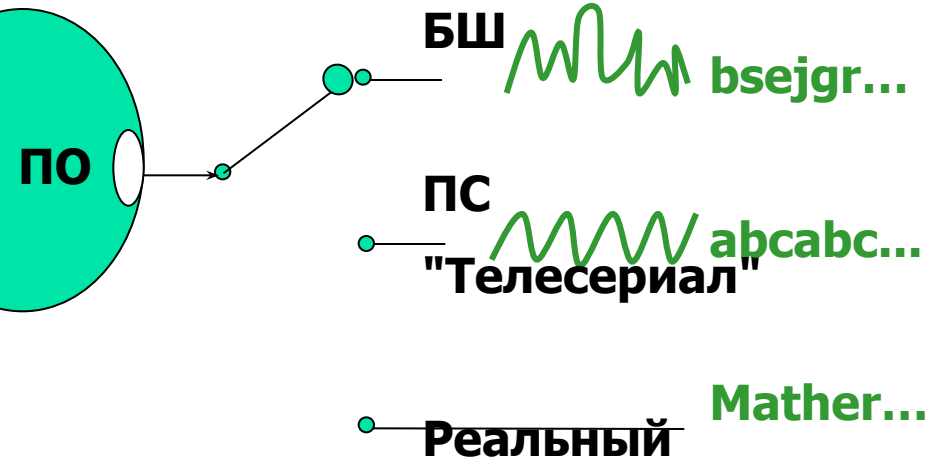
<?IGB@<KVLECMH;?IG?IG=H;?IG=LECM =H;LECM==B<KVH;=H;{> на -133.

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 тексто- вой 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

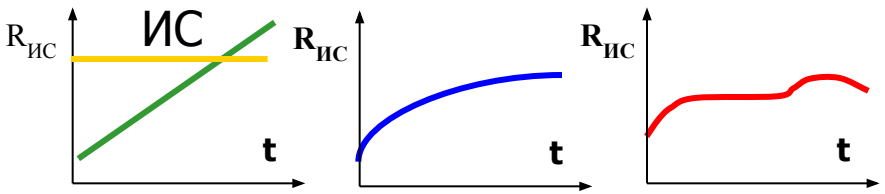


Автомат "животн." ИС-человек

Сигнал	С	С
С	Инфор-мация	И
С	И	Знание

При $T_{ИС} = const\ (t)$

$f = (R_{ИС}\ (t))$:



линейная; логарифмическая; const;

функции затрат Rис

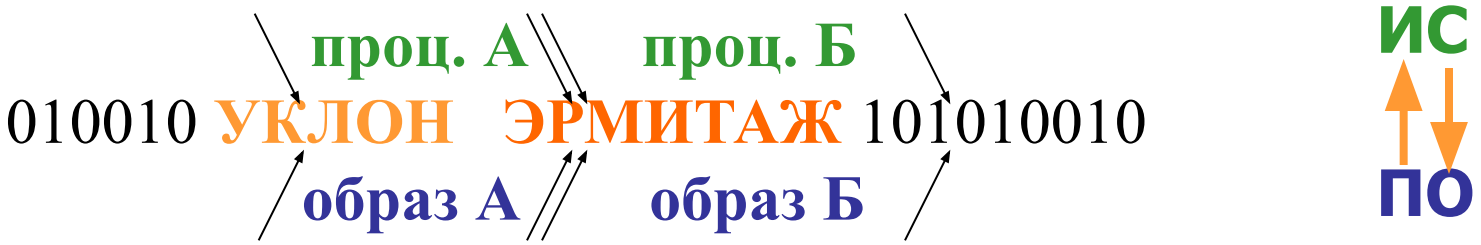
Определения:

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 генерирующих 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".