Санкт-Петербургский
Государственный
Политехнический
Университет

Институт прикладной
математики и механики

КАФЕДРА ТЕЛЕМАТИКА

История и методология математики и компьютерных наук

Не бойтесь
расти медленно,
бойтесь остановиться
/Будда/

Лекция 10

Вопросы вычислимости: от логических ошибок программных автоматов к суперпозиции несовместных состояний квантовых систем

17 ноября 2021 г.

ПОЛИТЕХ
Санкт-Петербургский
политехнический университет
Петра Великого

COMPUTO ERGO SUM

НАУЧНЫЙ СОВЕТ ПО ИНФОРМАТИЗАЦИИ САНКТ-ПЕТЕРБУРГА

О ПРИМЕНЕНИИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РАЗВИТИЯ ТЕХНОЛОГИЙ СУПЕРКОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ

17 ноября 2021 г.
СПб

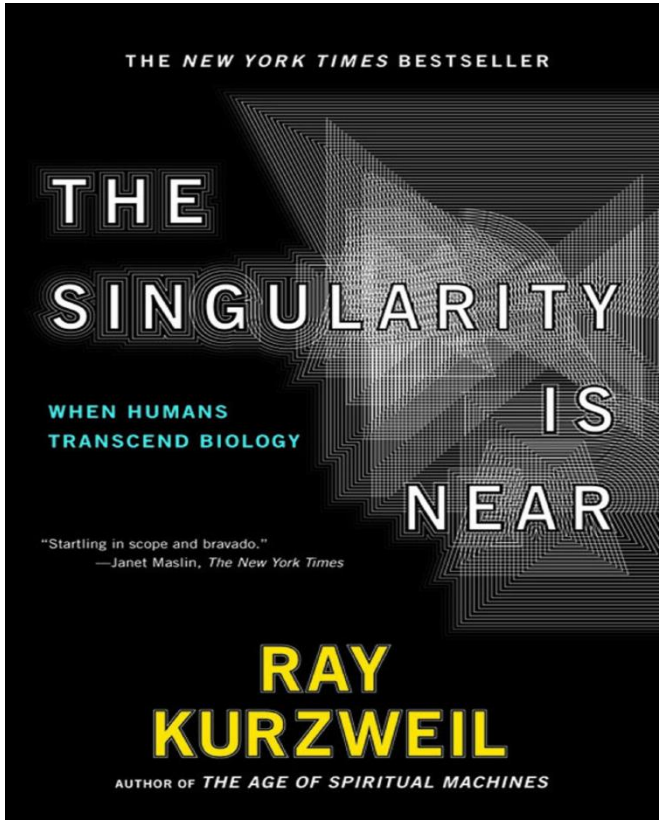
Заборовский В.С. , Л. В. Уткин
СКЦ «Политехнический»
Институт компьютерных наук и технологий

ПОЛИТЕХ

Содержание

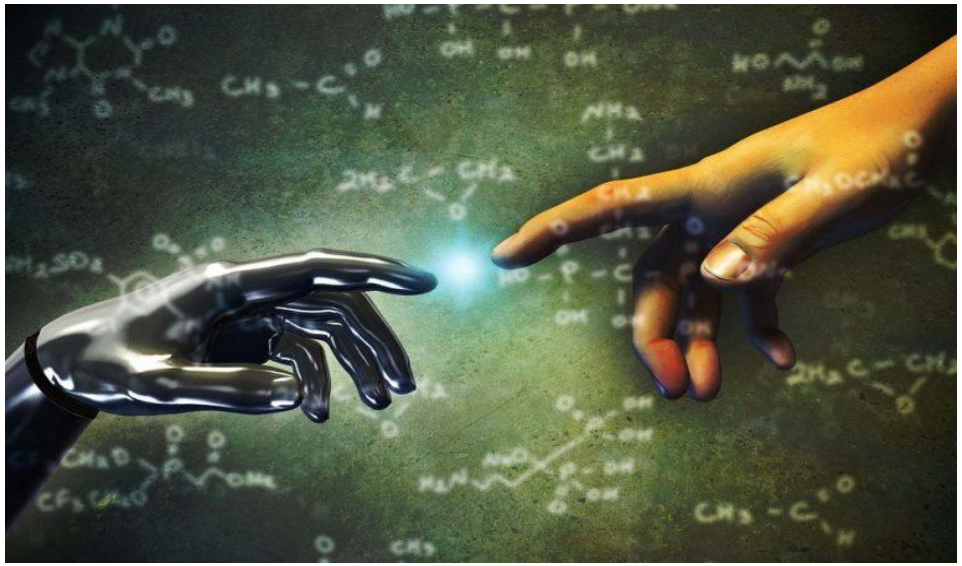
- Введение. Оценки эффективности процессов информатизации
- Перспективы развития и применения СКТ: ЭКБ vs ПО ?
- Аудит концепции ИТ: чему можно «научить» современный СК ?
- Предложение в проект решения

Введение в проблему (слайд презентации доклада. Сделанного на НСИ СПб в октябре 2018 г.)



Технический прогресс становится настолько **быстрым и сложным**, что может оказываться недоступным для понимания, но при этом открывая фантастические перспективы симбиоза возможностей человека и машины.

Р. Курцвейл



Возможен ли такой симбиоз технологически, если человек хорошо «понимает» слова, а компьютер - только числа…???

Траектория эволюции развития компьютерных технологий

Эра
часов автоматов,
управляемых одной
программной

«сила»
мысли -
алгоритм

эра
«**арифмометров**,
управляемых человеком
1900-1960

механическая
энергия-
вычисления

эра
«программных автоматов,
вычисляющих решения под
управлением программ
1960 – 2020

«сила»
алгоритма

Эл.энергия-
вычисления

результат-
объяснение

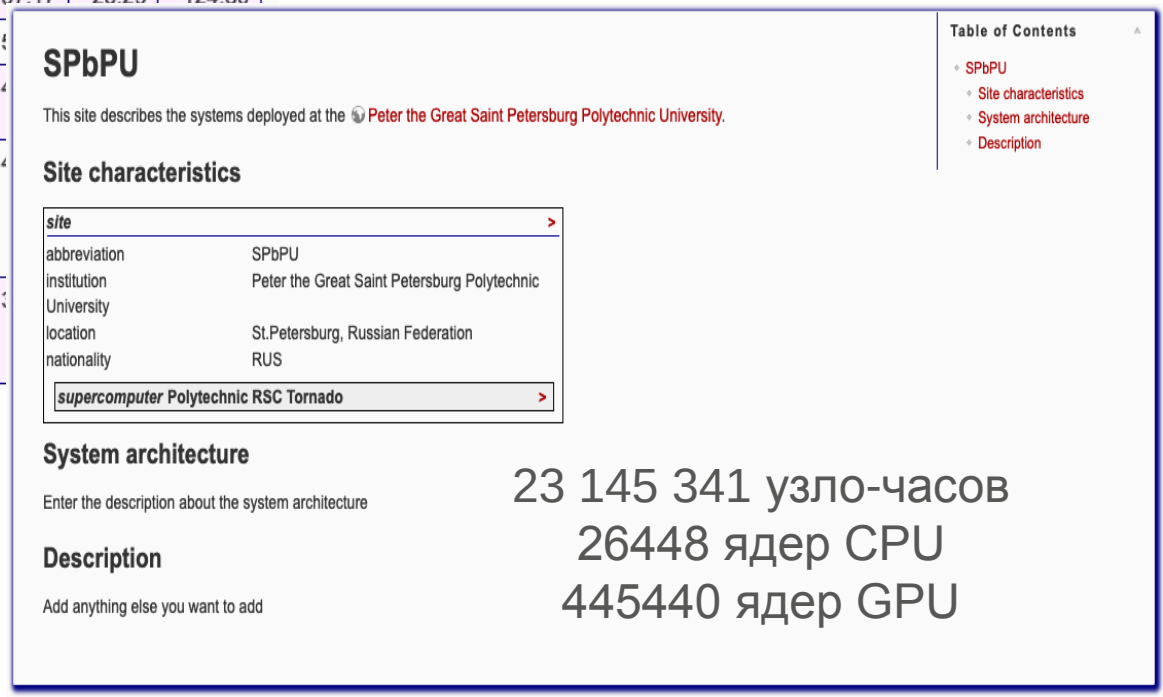
эра
гетерогенных «нейроморфных»
компьютерных систем, вычисляющих
решения под управлением
эмпирических **данных** и
объяснением результатов
2020 >

«знание -
сила»

Чего добились за 3 года:
СКЦ «Политехнический» в мировом рейтинге TOP IO500 (22-е место)

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw GiB/s	md kIOP/s
21	isc20		Officinalis	Red Hat, Intel, QCT	CephFS	8	256	zip	66.88	28.58	156.48
22	isc20	SPbPU	Polytechnic RSC Tornado	RSC Group	Lustre	59	944	zip	64.29	21.56	191.73
23	sc19	DDN	AI400	DDN	Lustre	10	240	zip	63.88	19.65	207.63
24	isc20	Red Hat	EC2-10xi3en.metal	Red Hat	CephFS	10	320	zip	57.17	26.29	124.30
25	sc19	Google Cloud	EXA5-GCP-PD-STD	Google Cloud	Lustre	200	1600	zip			
26	sc19	Janelia Research Campus, HHMI	Weka	WekaIO	wekaio	18	1368	zip			
27	sc19	Oracle Cloud Infrastructure	Oracle Cloud Infrastructure with Block Volume Service running Spectrum Scale	Oracle Cloud Infrastructure Block Volumes Service	Spectrum Scale	30	480	zip			
28	sc19	Penguin Computing Benchmarking and Innovation Lab	Penguin-ASG-NVBeeOne	Penguin Computing/Excelero	BeeGFS	10	320	zip			

на 15.11.2020



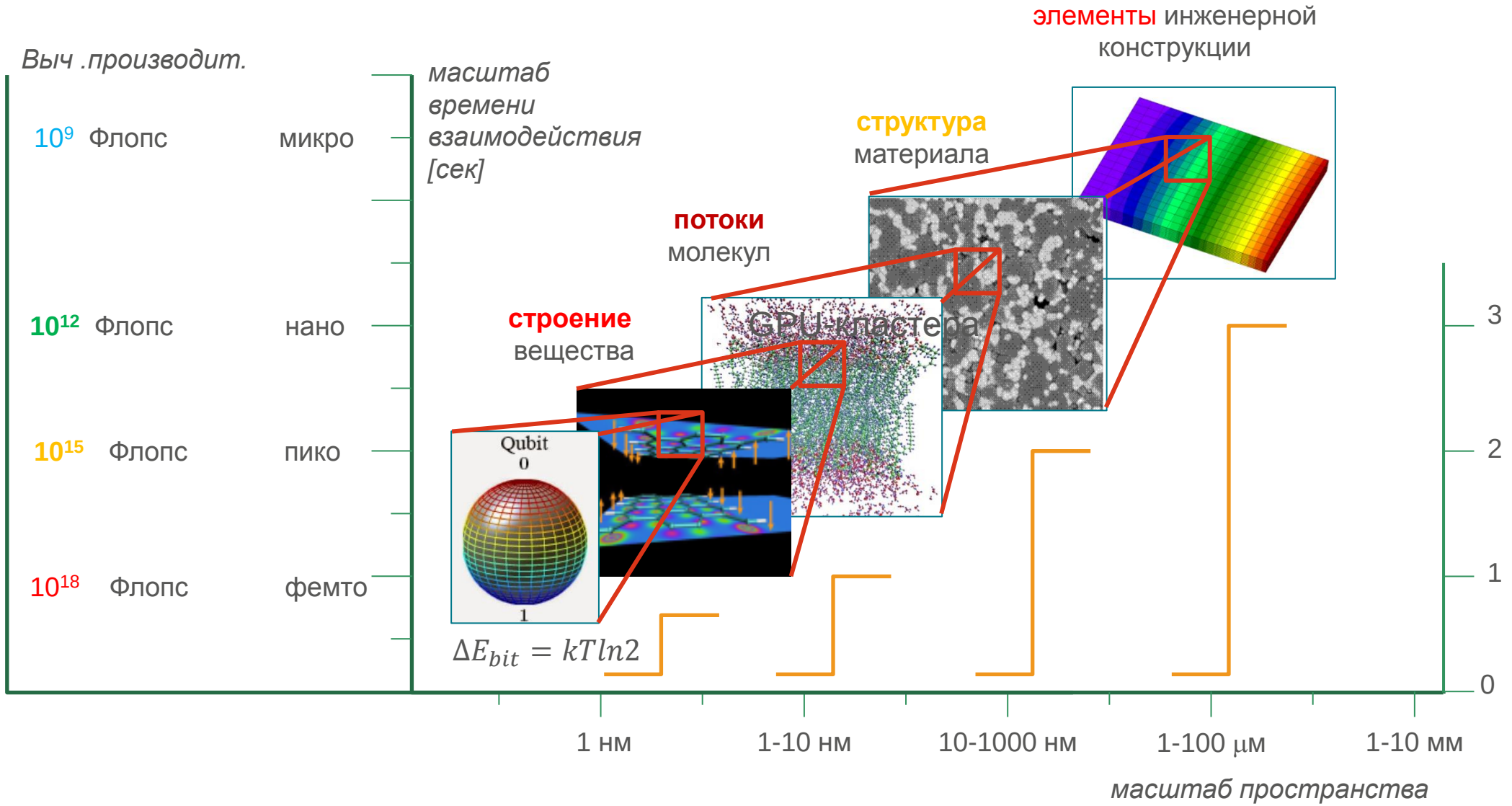
1 570 442
выполненных задач

23 145 341 узло-часов
26448 ядер CPU
445440 ядер GPU

В подведомственных организациях **Минобрнауки** СКЦ «Политехнический» самый производительный.
гетерогенный кластер
Пиковая суммарная производительность > **2 Флопс**.

ЦКП http://ckp-rf.ru/ckp/500675/
УНУ https://ckp-rf.ru/usu/507708/

Специализация СКЦ: расчеты мультимасштабных физических процессов и инженерных конструкций



Тренды 20-х: новые ИТ-приложения, требующие суперкомпьютеров

В ноябре 2021 **три GPU-кластера компании Яндекс** заняли **19, 36 и 40** места в рейтинге суперкомпьютеров Top500.

ссылка:

https://habr.com/ru/company/yandex/blog/589363/

Характеристики кластера -19 место:

AMD EPYC 7702 64C 2 ГГц, NVIDIA A100 80GB, Infiniband, IPE, Nvidia,

- Число ядер 193,440 шт.
- Реальная производительность 21,530.0 Тфлопс (21, 5 *10^{15} Флопс)
- Пиковая производительность 29,415.2 Тфлопс (29,4 * 10^{15} Флопс)

(10^{15} флопс - миллион миллиардов операций с плавающей точкой в секунду)

Невые приложения и… сервисы:

- задачи настройки **нейроморфных** алгоритмов поиска в Интернет
- алгоритмы машинного обучения **нейросетей-трансформеров** анализа текстов

Зачем же Яндексу GPU-кластеры: «цена» обучения нейронных сетей

На примере задачи обучения н/с распознавать 3D изображения компьютерной томографии (КТ)

СК: **1Пфлопс = 1×10^{15} Флопс**

Обучающая выборка **10^4 КТ снимков**

p=10^{10} - число настраиваемых весов нейронной сети

КТ легких

Вход x:

Выход y:

сверочная нейронная сеть

окно:

100 ×100

Вектор признаков ≈ **10**

n – число слоев КТ (n=100)

«Вес» данных КТ одного слоя **2 ×10^6 Байт**

Входной слой нейронной сети размером **10^3 нейронов**

Число слоев н/с 10^2

Скользящее окно алгоритма свертки

Для обучения н/с с точностью классификации **90%** надо:

Выполнить вычислительных операций ≈ **10^{17}**

Время обучения (настройки) ИНС

на СК ≈ **10^2 сек**

на ПК ≈ **2×10^6 сек**

(200 Гфлопс)

Нужен ли такой «черный ящик» врачам ?

Подробности: для обучения глубокой сверточной нейронной сети используется градиентный алгоритм :

- функция ошибки классификации $F = \|y^* - y\|^2$, y* - эталонный вектор признаков
- **Алгоритм вычисляет 10^{10}** частных **производных** функции F по всем настраиваемым параметрам ИНС;
- Число операций численного **дифференцирования** на одну итерацию (эпоху) обучения: **Q= 10^{15}**

Что мешает реализовать «фантастические возможности» СКТ?

Несмотря на постоянное развитие ИТ- технологий, даже наиболее современные системы несовершенны – цикл эволюции не завершен. Почему ?

Технические аспекты:

- Квалификация пользователей в сфере информационных технологий, различна. Одно и то же программное обеспечение используют люди разной квалификации и возраста.
- Пользователи ИТ консервативны, поэтому обновления и изменения ИТ нередко вызывают у них не радость, а раздражение.

Экономические аспекты:

- Совокупная стоимость владения **растет**
- Коэффициент возврата инвестиций определить **сложно**
- Референтная модель: качество сервисов, расширение возможностей, увеличение объема транзакций и пр. **противоречива**

ПОЛИТЕХ

Научные аспекты

- Современные компьютеры это «очень быстрые» автоматы, которые умеют выполнять программный код «слово в слово», но не **понимают** суть выполняемых операций. За них это «делают»: программисты, а результат объясняют – «хороший» специалист.
- Может ли компьютер исполнять роль и программиста и хорошего специалиста ? В приципе, да, но… для этого его надо наделить возможностью
  - «настраиваться» на решение новых задач,
  - накапливать опыт, чтобы затем использовать его для решения аналогичных задач.

Всего этого современные компьютеры пока умеют делать плохо!

Надо … «научить» компьютер не только вычислять, но и объяснять полученный результат

Если вы не можете **объяснить** что-либо простыми словами, вы это не **понимаете**

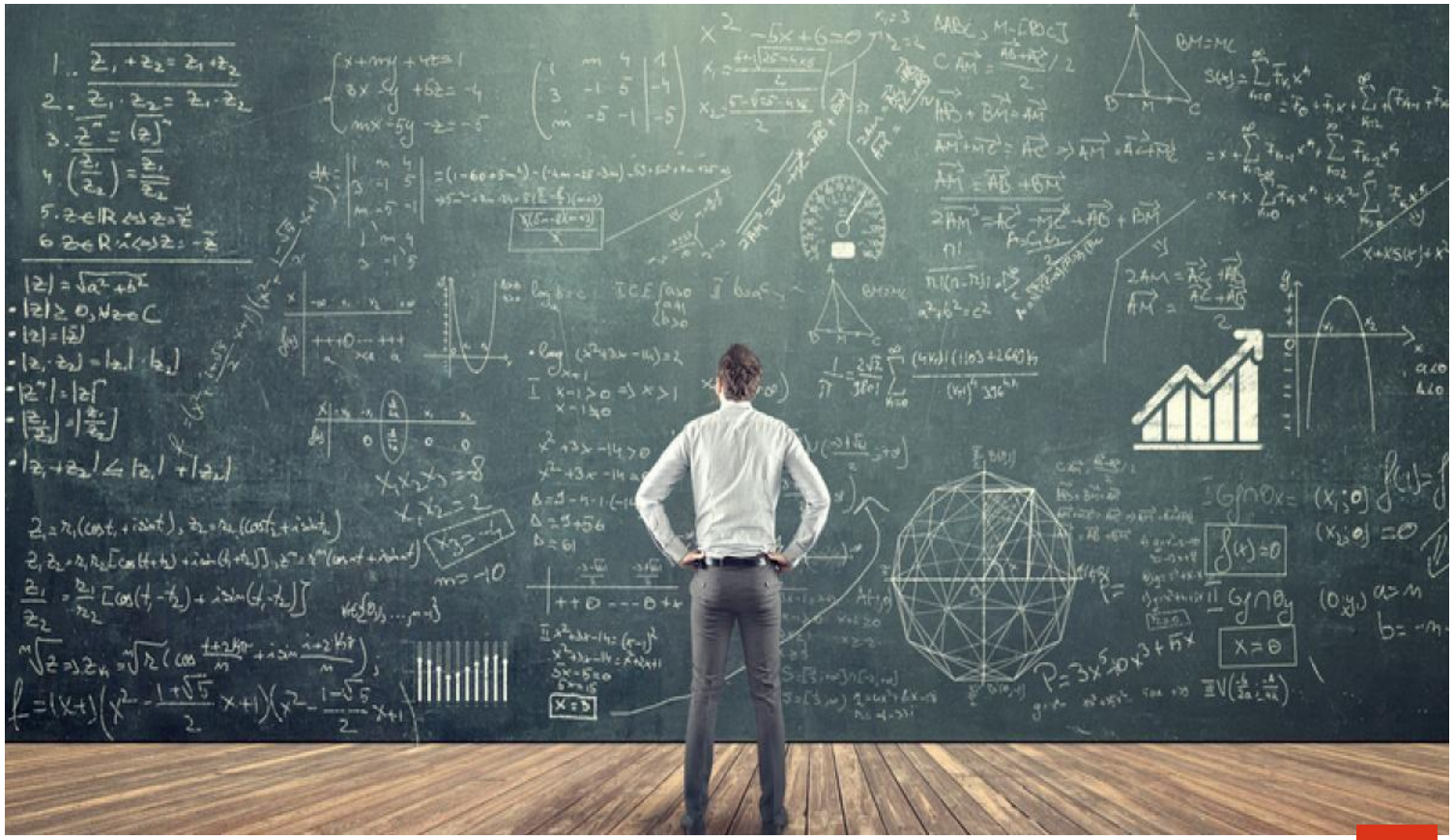
Р. Фейнман

Современные компьютеры, особенно суперкомпьютеры, похожи на «черные ящики».

а

технологии «программирования» - на **манипуляцию «past and copy» кодов из открытых**, **не** известно кем написанных **библиотечных функций**.

Надо «разорвать» этот порочный круг не впадая в нумерологию, а совершенствуя используемый инструментарий и «подкрепляя вычисленное компьютером число – интерпретацией полученного результата»….

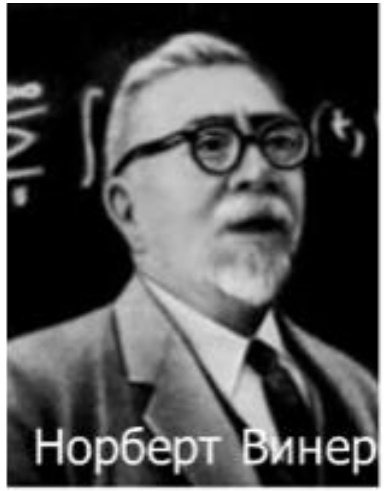


ПОЛИТЕХ

Развитая технология неотличима от магии
Артур Кларк

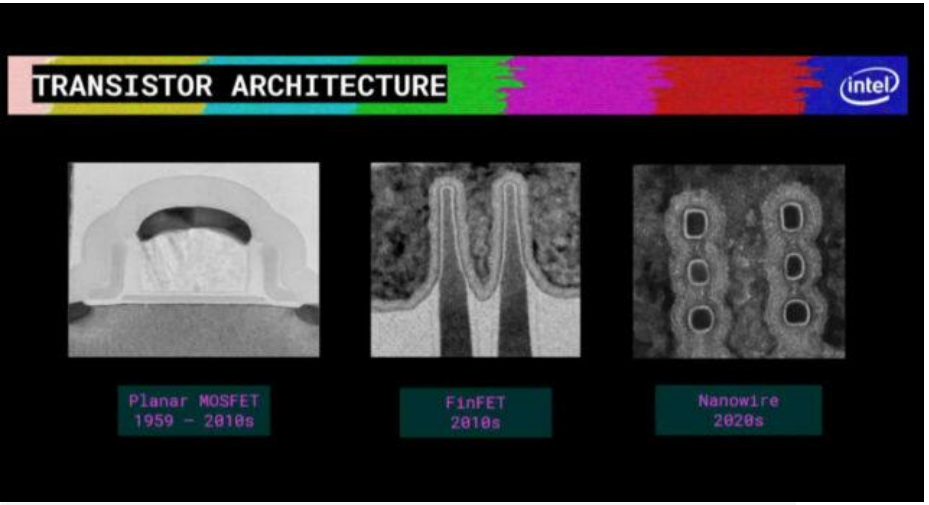
Фундаментальная проблема компьютерных наук: как сформулировать прикладную задачу так, чтобы она всегда **имела решение**, которое компьютер может …не только **вычислить**, но проверить и «**объяснить**» ?

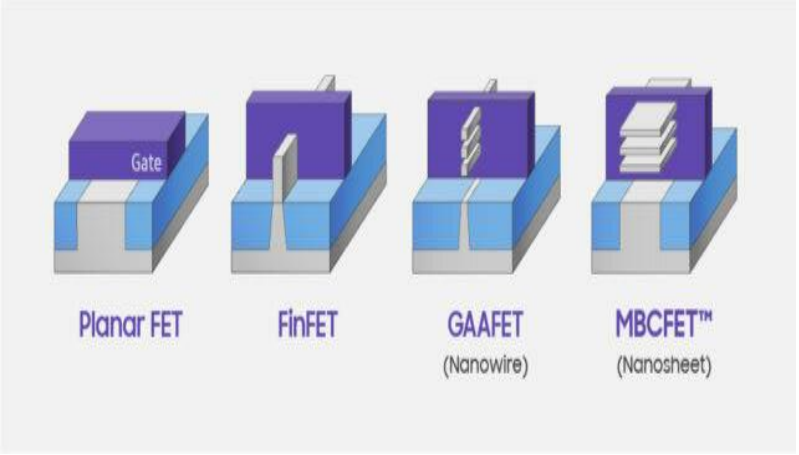
«знание -сила»



Если магия вообще способна даровáть что-либо, то она дарýет именно то, что вы **попросили**, а не то, что вы **подразумевали**. **Н. Винер**

Решения в пространстве технологических возможностей *: 3-5 нм ЭКБ + МО*

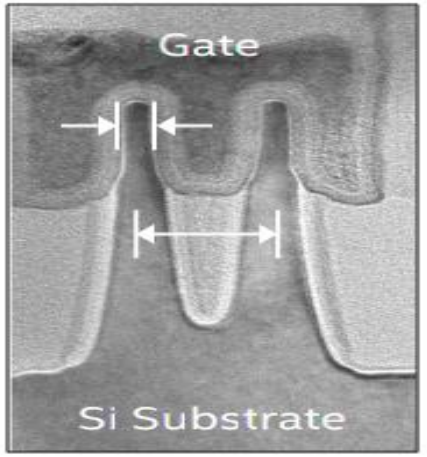




эволюция конструкций планарных транзисторов

8 nm Fin Width

42 nm Fin Pitch



3-5 нм МП содержит от 8 до 50 млрд. транзисторов это >>, чем все население Земли

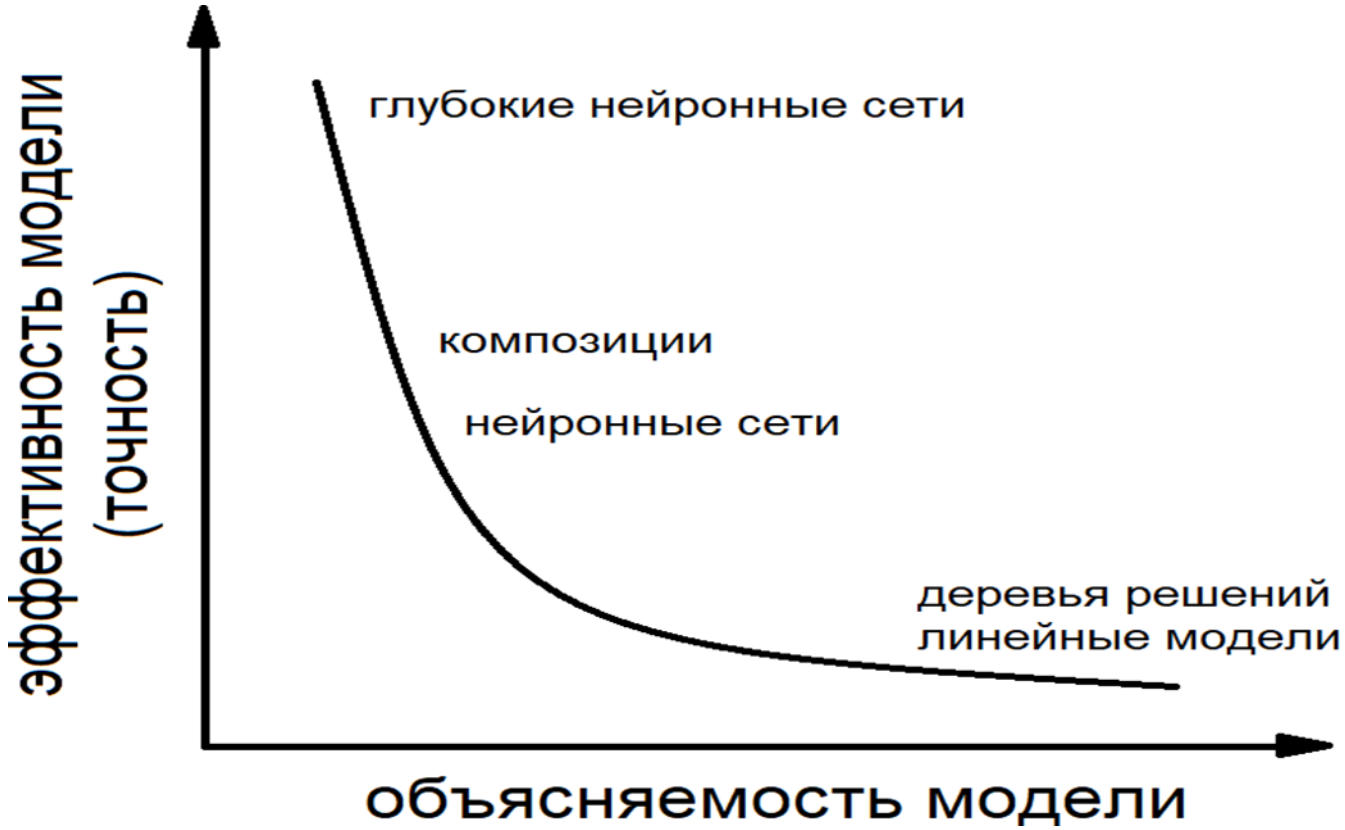
В «лабиринтах» нового пространства вычислительных возможностей можно найти «траекторию» для реализации любого эффективного алгоритма. Вопрос: на что же лучше «потратить» имеющиеся ресурсы ? Варианты:

1. (было) **Гетерогенность** архитектуры МП и расширение **шины данных** между компонентами

2. (есть) **Реконфигурируемость** структуры МП «под задачу»

3. (д. быть) **Встроенная в МП система «машинного обучения»**, которая обеспечить «накопление» знаний как решать задачи

Чему «учить» МП и нужно ли открывать «черный ящик »

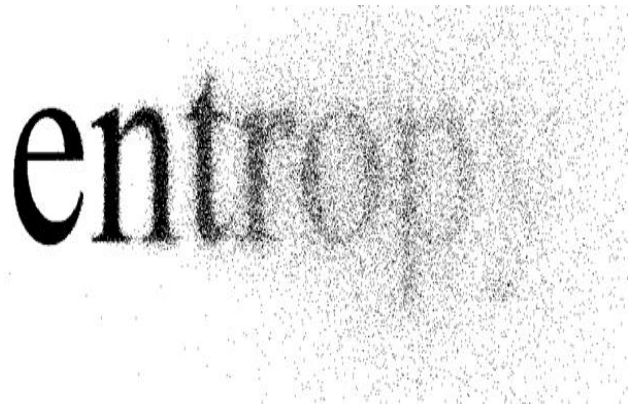


- Чему «учить»: как гетерогенный МП из 50 млрд транзисторов может управлять своими ресурсами
- Нужно ли «открывать»: да, так как **эффективные** вычислители + программы как правило сложны для понимания, не реализуют функционал **самообъяснения, поэтому …** не робастны к определенному классу ошибок (например, **атака «одного пикселя»**)

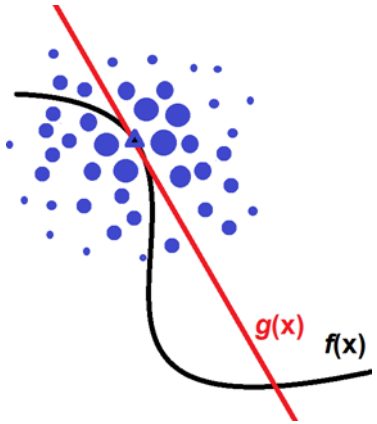
Направление исследований: от GPU вычислений – MIMD структур к системам объяснительного интеллекта – «правильному» ИИ

Суть интеллектуализации компьютерных технологий:

1) Способность **накапливать информацию о произведенных вычисленных** для объяснения полученных результатов и поиска вариантов используя этих результатов…



Неопределенность данных и сложность алгоритмов

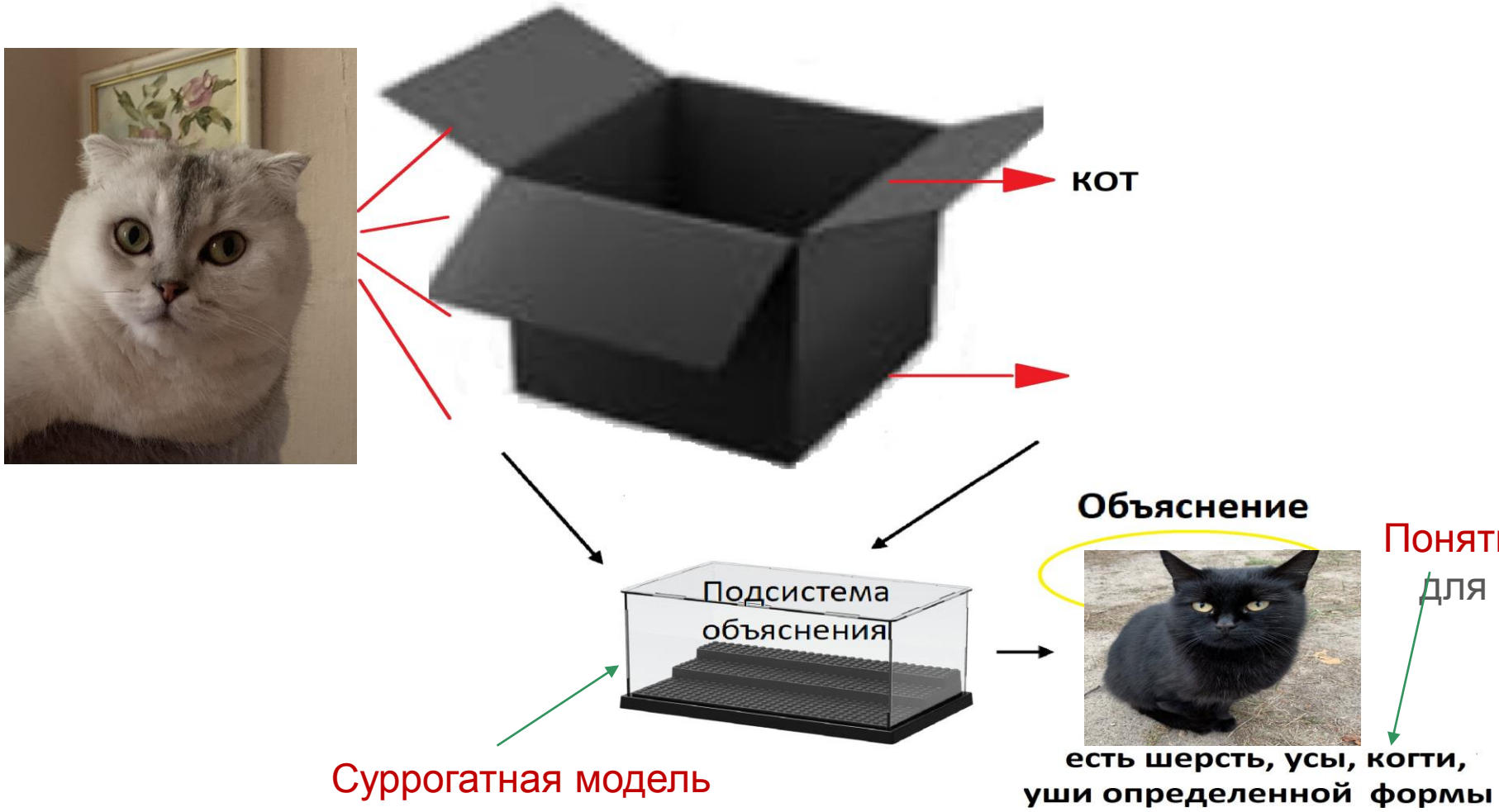


Модели объяснения решения

2) **«объяснения»** – суть новый метод повышения «доверия» к результатам вычислений или **способ регуляризации решения «обратных алгоритмических задач»** (использование знаний не отраженных в самом алгоритме решения)

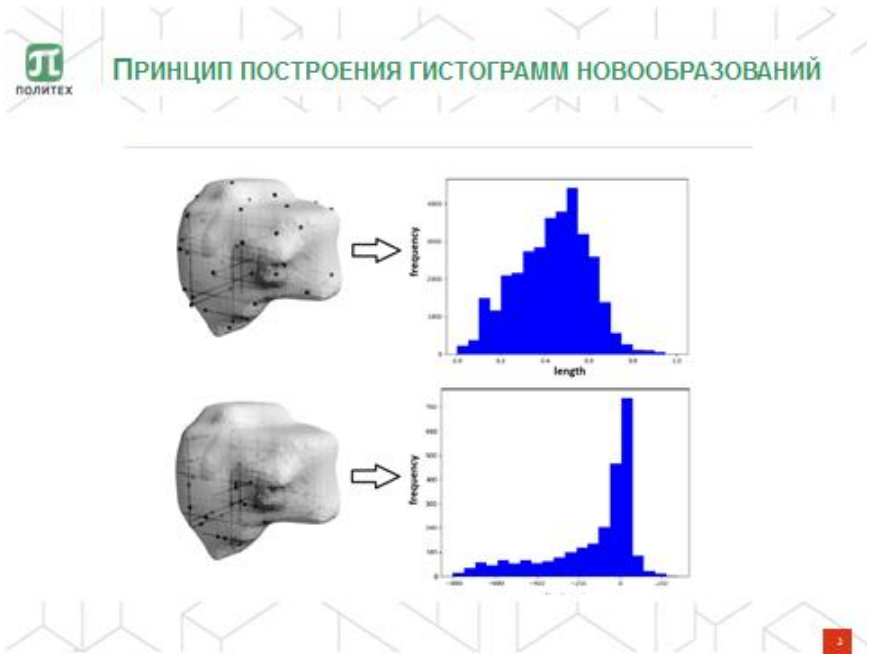
Итак: умение решать обратные задачи – основа «правильного» ИИ, с помощью которого «поиск» решений реализуется в «пространстве возможностей», в котором **вычислитель может сформировать интерпретацию** полученного численного результата

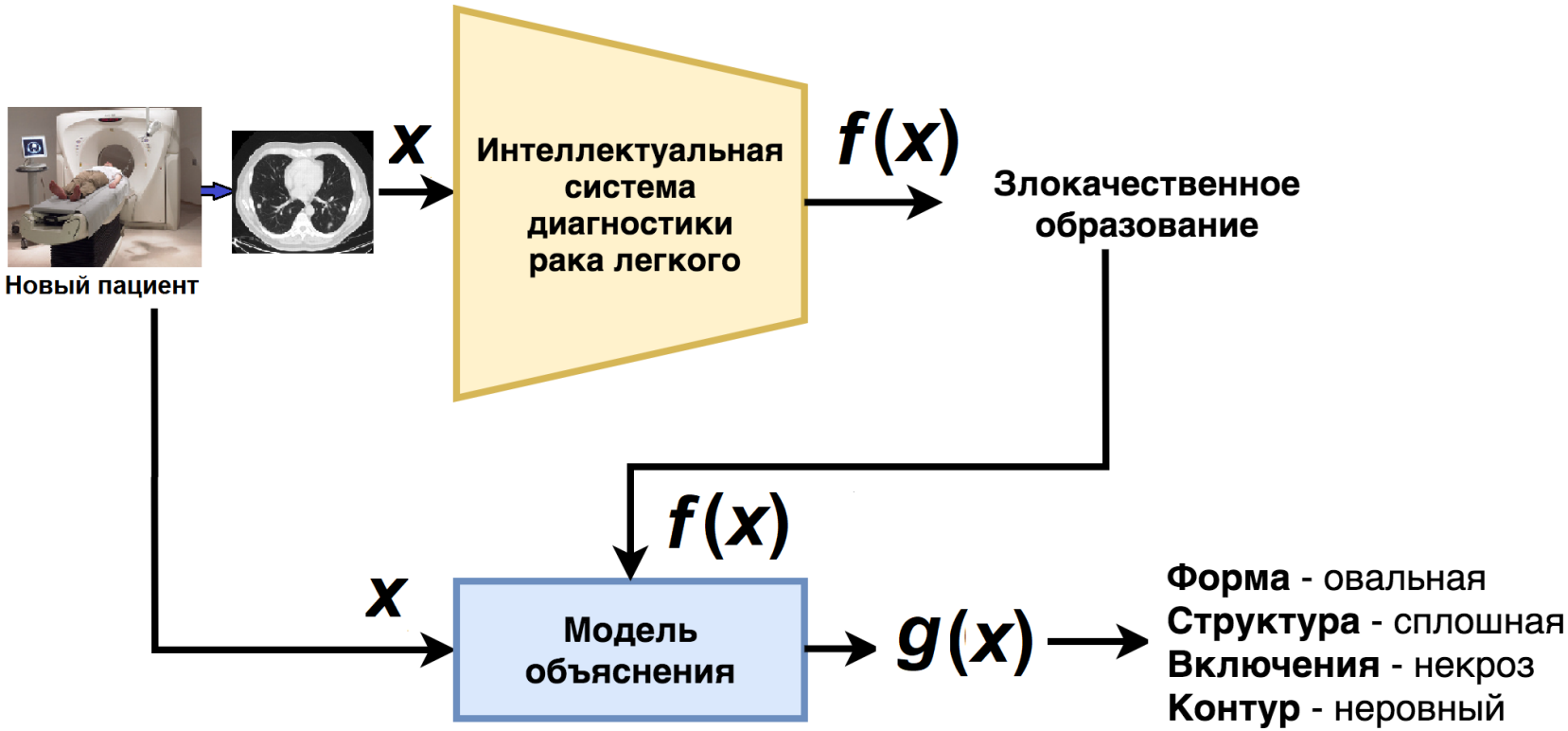
Пример: модель машинного обучения с подсистемой объяснения



Пример: система объяснительного интеллекта для медицинской диагностики

Необходимо построить объяснения на основе объективных (не зависящих от данных) **инвариантов-паттернов** (например, топологических) и модели МО (глубокая нейронная сеть, случайный лес, SVM и т.д.), которая «аппроксимирует» свойства сложной модели реальности **в окрестности конкретного примера**





Реализуемое направления научения СК

Создание систем **интеллектуального диспетчерского управления** ресурсами гетерогенных СК, включая возможность их реконфигурации **с учетом** метрик/целевых требований:

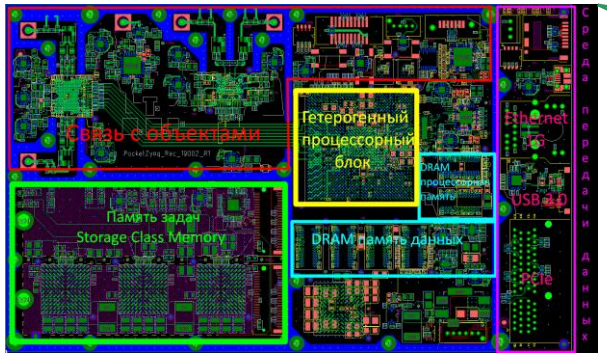
1. минимальное среднее время ожидания $\mathbf{mat_i}$ – интервал между временем постановки в очередь и началом исполнения прикладной задачи I
2. минимальный цикл ответа $\mathbf{art_i}$ – среднее время ожидания и время проведения расчетов
3. минимальное среднее замедление выполнение **mas** $=\mathbf{(mat_i+art_i)/art_i}$, i=1,2…..N
4. максимальное использование ресурсов – средняя доля используемых гетерогенных узлов к общему числу узлов суперкомпьютерной платформы за заданный период времени
5. максимизация отношения энерго-вычислительной эффективности вычислений (Гфлопс/Вт), включая интерпретацию полученной оценки для различных классов прикладных задач и выработка рекомендаций по оптимизации параметров вычислительной платформы с использованием методов объяснительного интеллекта XAI

Работа выполняется в рамках Государственного задания СПбПУ на проведение -прикладных научных исследований 075-01018-21-04 «Разработка мультиагентного диспетчера управления ресурсами гетерогенной суперкомпьютерной платформы с использованием методов машинного обучения и искусственного интеллекта

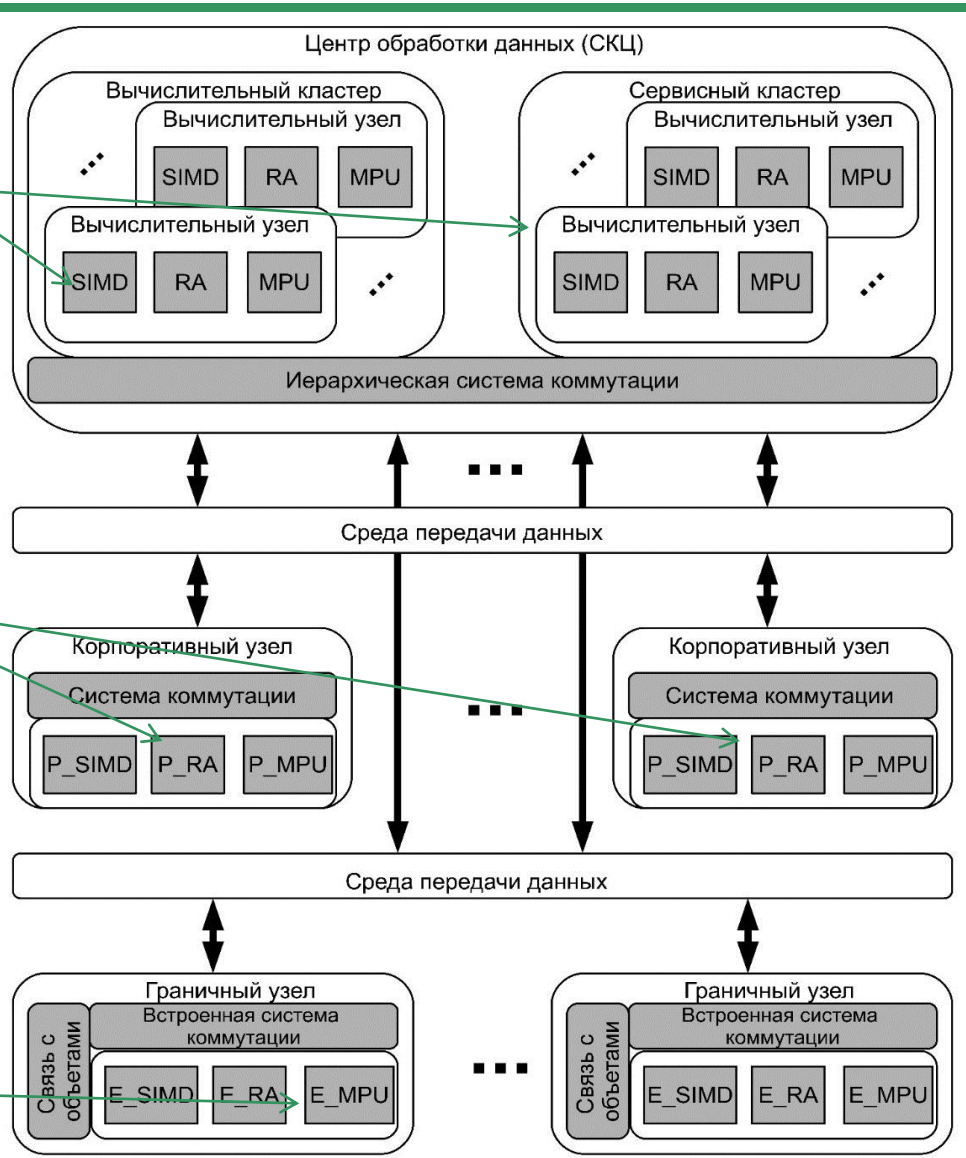
Суть предлагаемого решения: новая иерархия «уровней и функций» перспективной вычислительной платформы СКЦ «Политехнический»

Уровень «**объяснения**» результатов ; функция оценка парамтеров >4 Гфлопс/Вт

Уровень **обобщения результатов ; функция** «машинного обучения» >10 Гфлопс/Вт



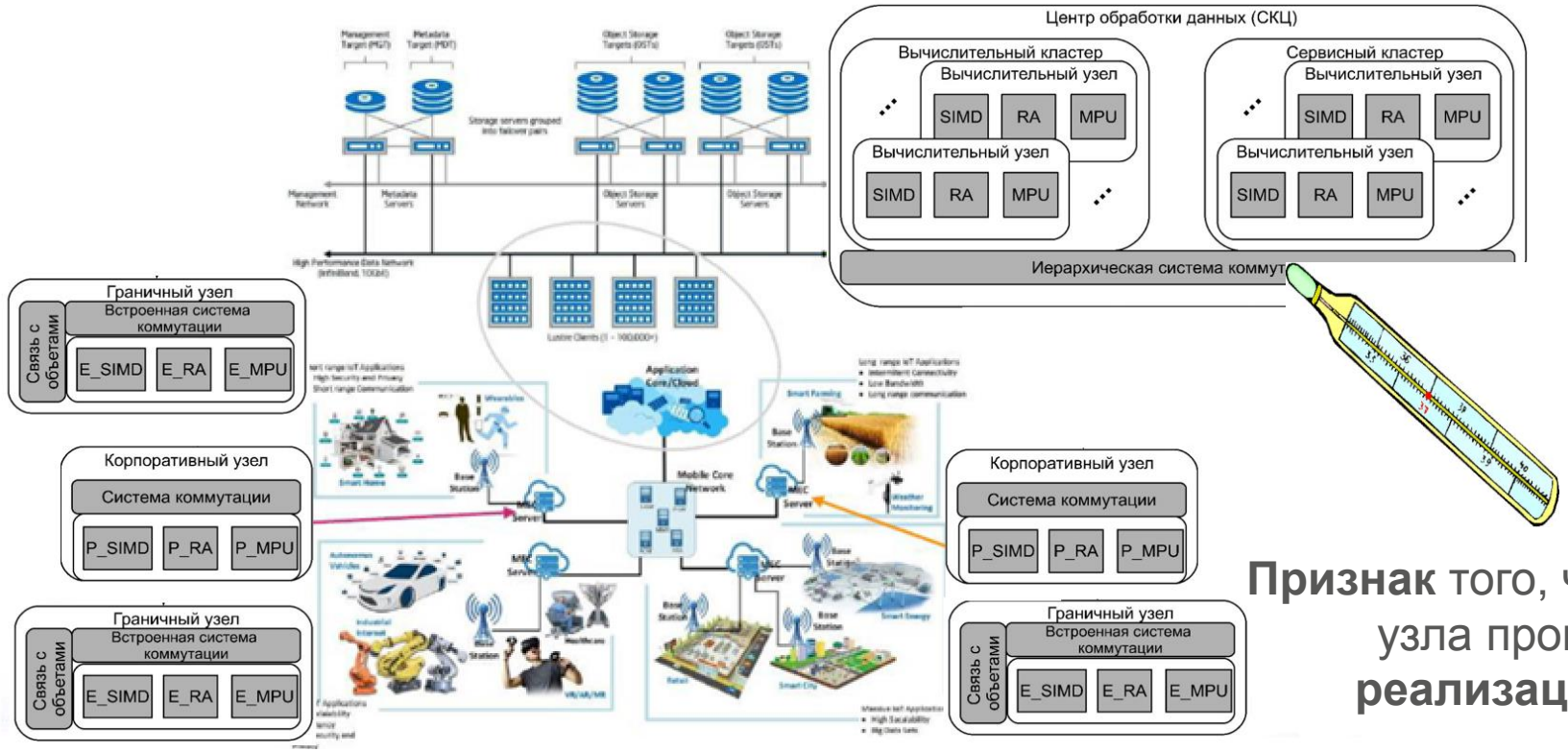
Уровень **моделирования**; функция алгоритм вычисления , > Гфлопс/Вт

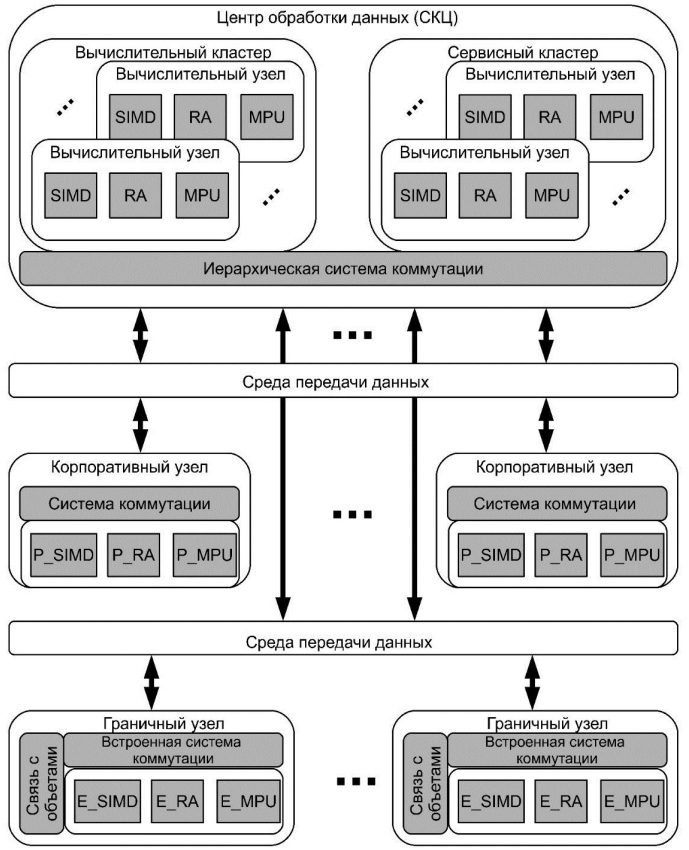


Перспективы: создание Петербургского узла национальной супер компьютерной инфраструктуры производительностью *20 Пфлопс*

Создание такого СК узла в структуре НСИ обеспечит:

- *лидирующие позиции* ***СПб*** *в процессах цифровой трансформации экономики*
- *новые возможности в управлении городской инфраструктурой*
- *повышение качества медицинского обслуживания*
- *ускоренное развитие наукоемких производств*
- *высокую точность прогноза инвестиций*





Признак того, что нужна «адаптация» АО и ПО - узла прогноз диссипации тепла **при реализации прикладного алгоритма**

Предложения в проект решения НСИ

Не бойтесь расти медленно, бойтесь остановиться

/Будда/

- **Одобрить** деятельность СПбПУ Петра Великого по разработке методов машинного обучения для повышения эффективности технологий суперкомпьютерного моделирования.
- **Создать** на базе СКЦ «Политехнический» и Института компьютерных наук региональный центр исследований и подготовки кадров в области использования перспективных компьютерных технологий
- **Поддержать** предложение СПбПУ по созданию регионального центра национальной суперкомпьютерной инфраструктуры производительностью 20 Пфлопс

Приложение
«Вычисление» диагноза на основе извлечения знаний из данных

клиники,
мед. организации

Этап 1 - получения и обработка данных

МРТ

КТ/ПЭТ

УЗИ

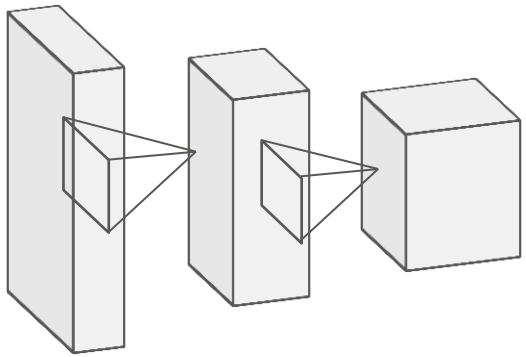
Омиксные данные

Цикл 4 - персональный цифровой «двойник» пациента

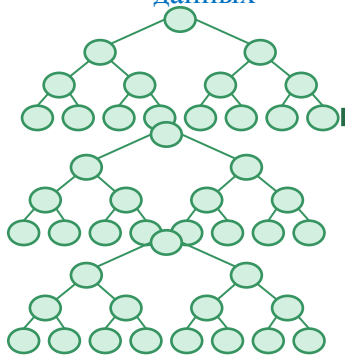
вход в систему классификации и диагностики

Этап 2 - интеллектуальная обработки данных, предсказательное моделирование и анализ

СКЦ «Политехнический»: нейронные сети, системы объяснительного ИИ, машинного обучения и моделирования



Компьютерные платформы анализа медицинских данных



Этап 3 - точная и ранняя диагностика, профилактика и лечение

Цифровой диагностический центр:
база знаний, компьютерный атлас патологий, профессиональная аналитика и экспертный контроль, диагностика в режиме онлайн, ранняя диагностика социально значимых заболеваний, прогноз эпидемий

Центр омиксных исследований компьютерной геномики

Этап 5
Проведение доклинических и клинических исследований, разработка методик

Партнеры:
мед. академ. организации