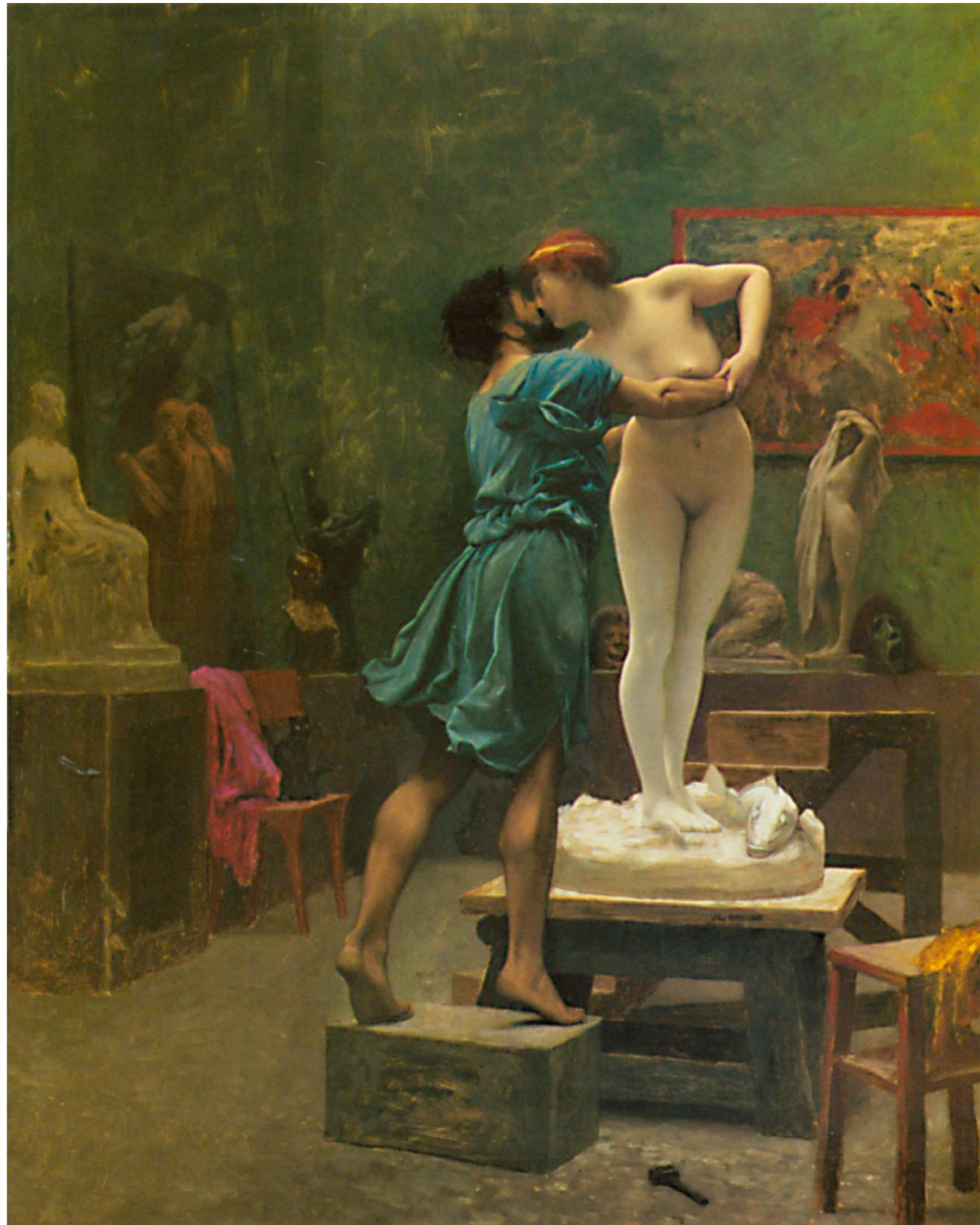


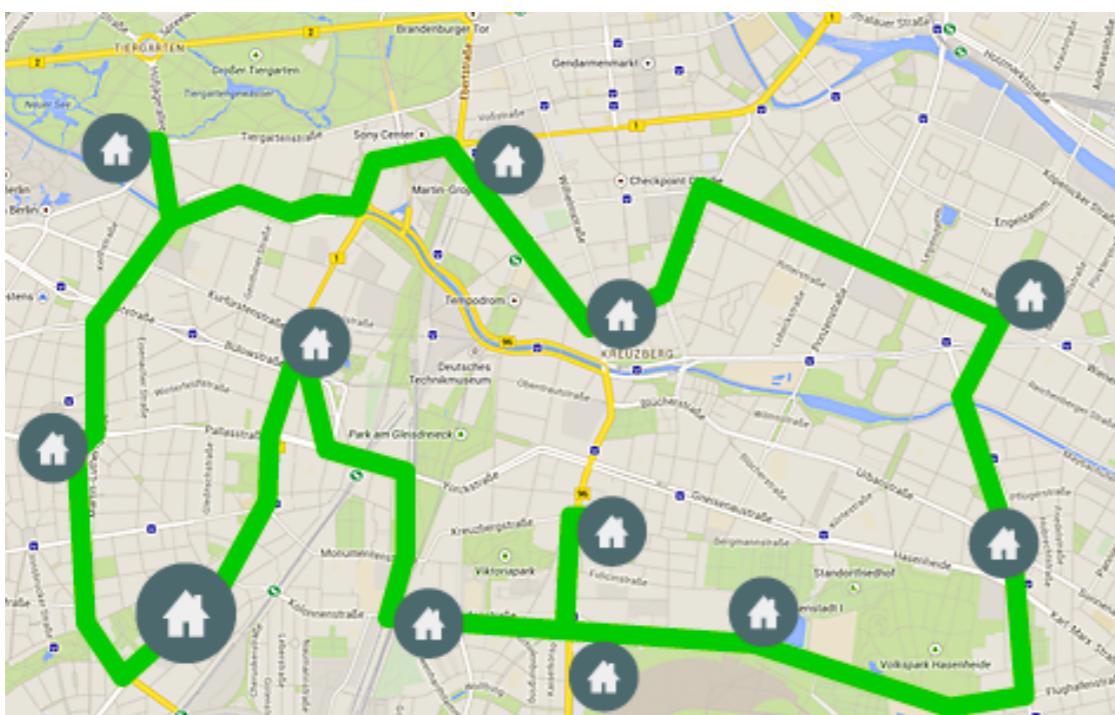
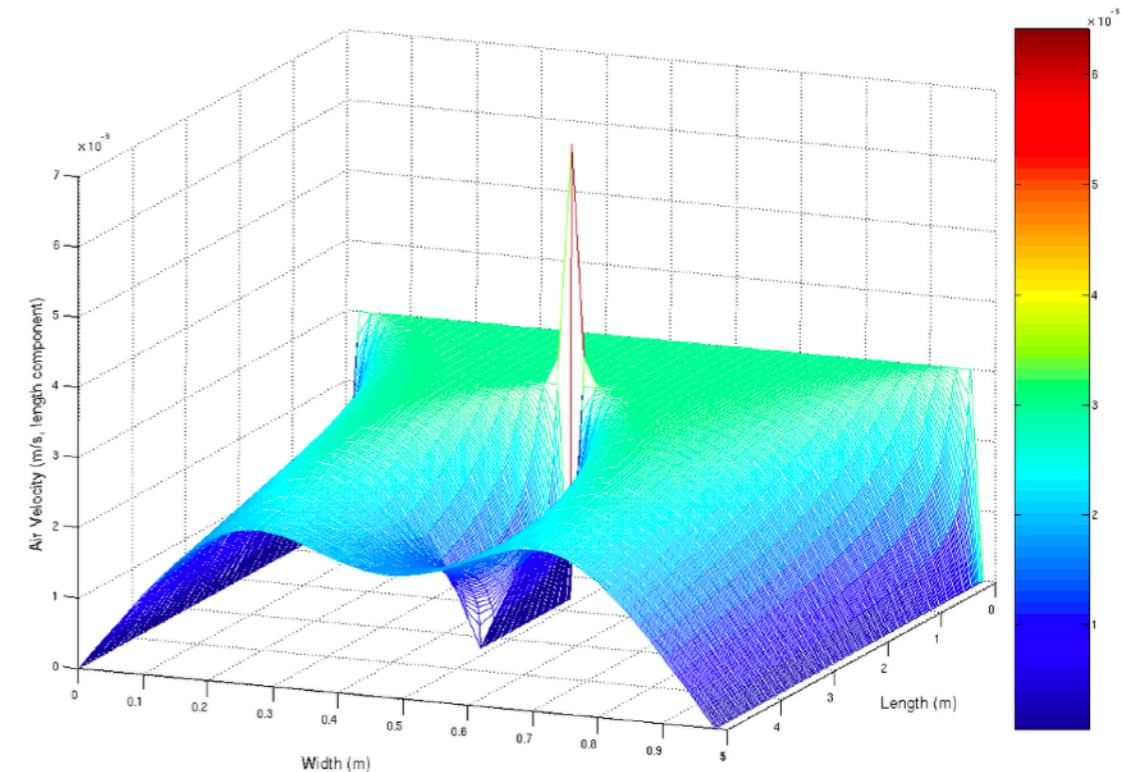
Машинное обучение

Лекция 1

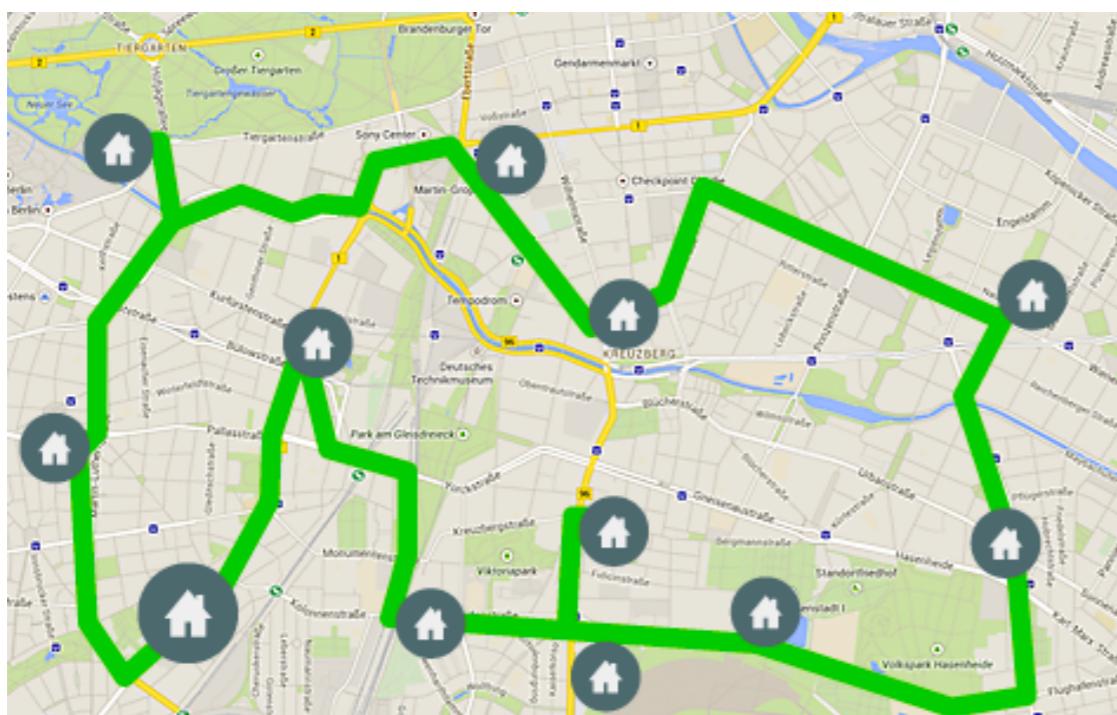
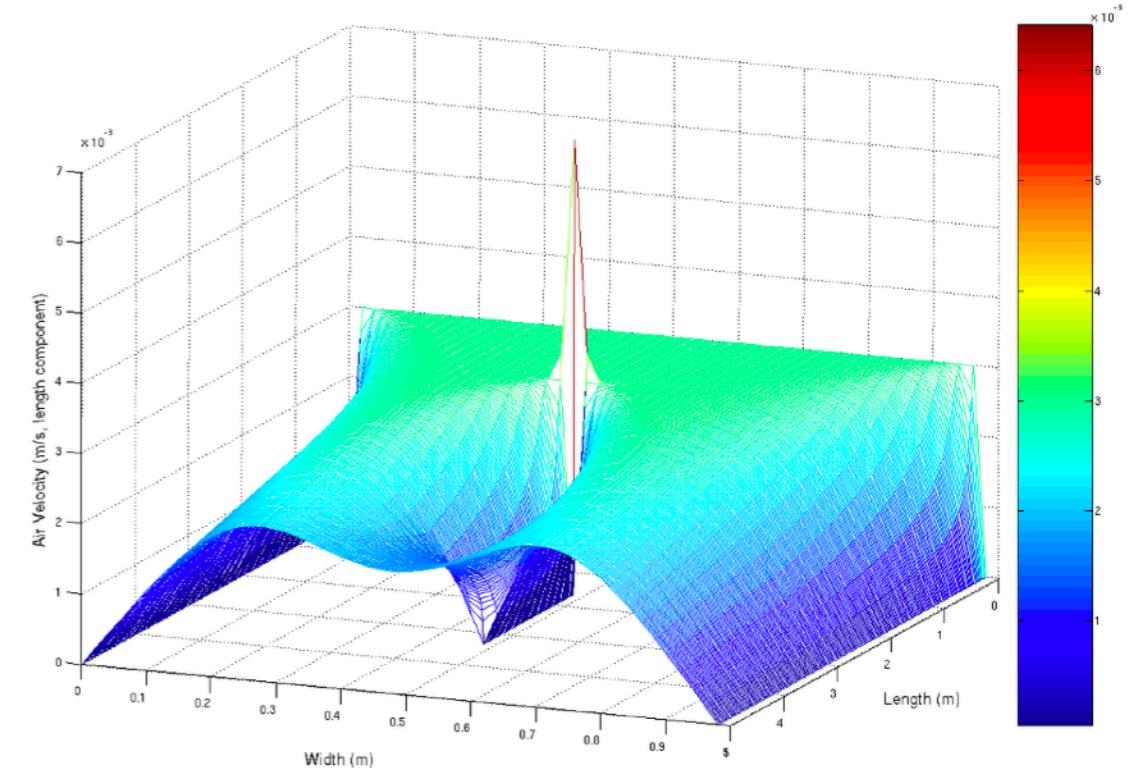
Галатея



Сложные для человека проблемы

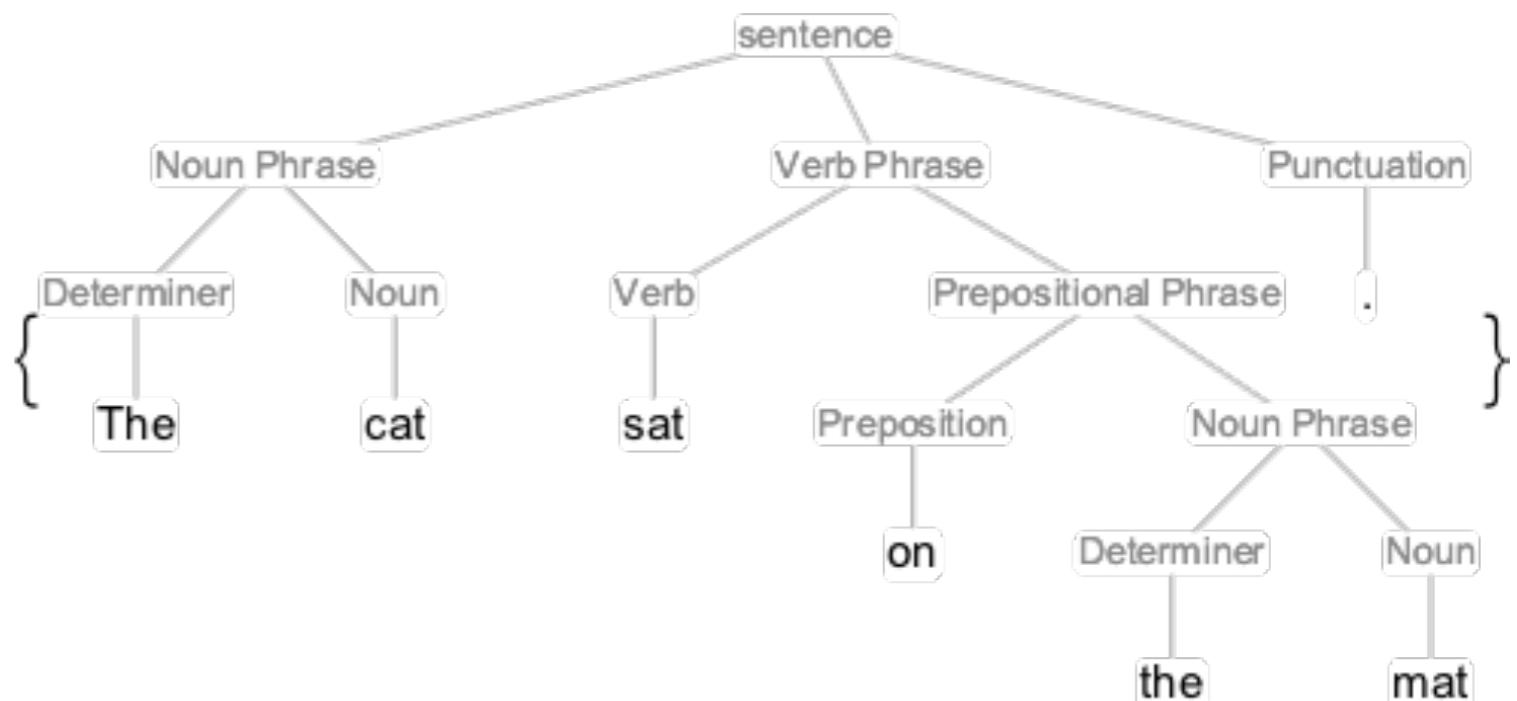
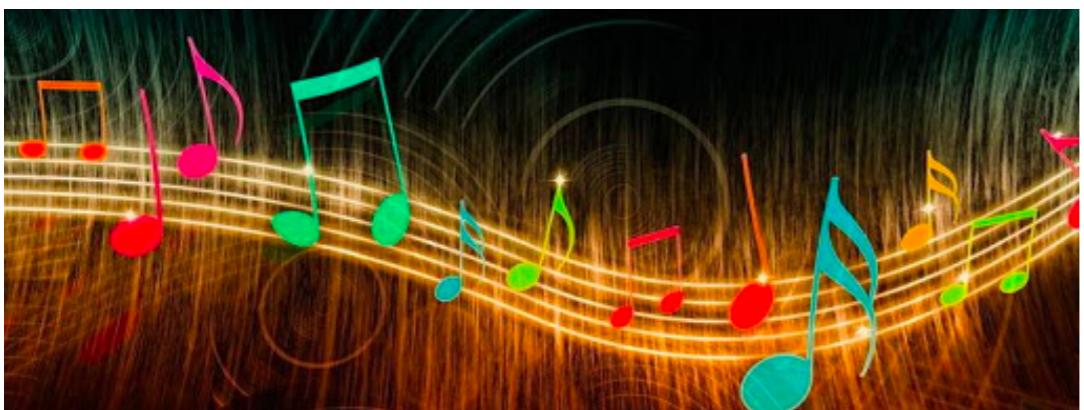
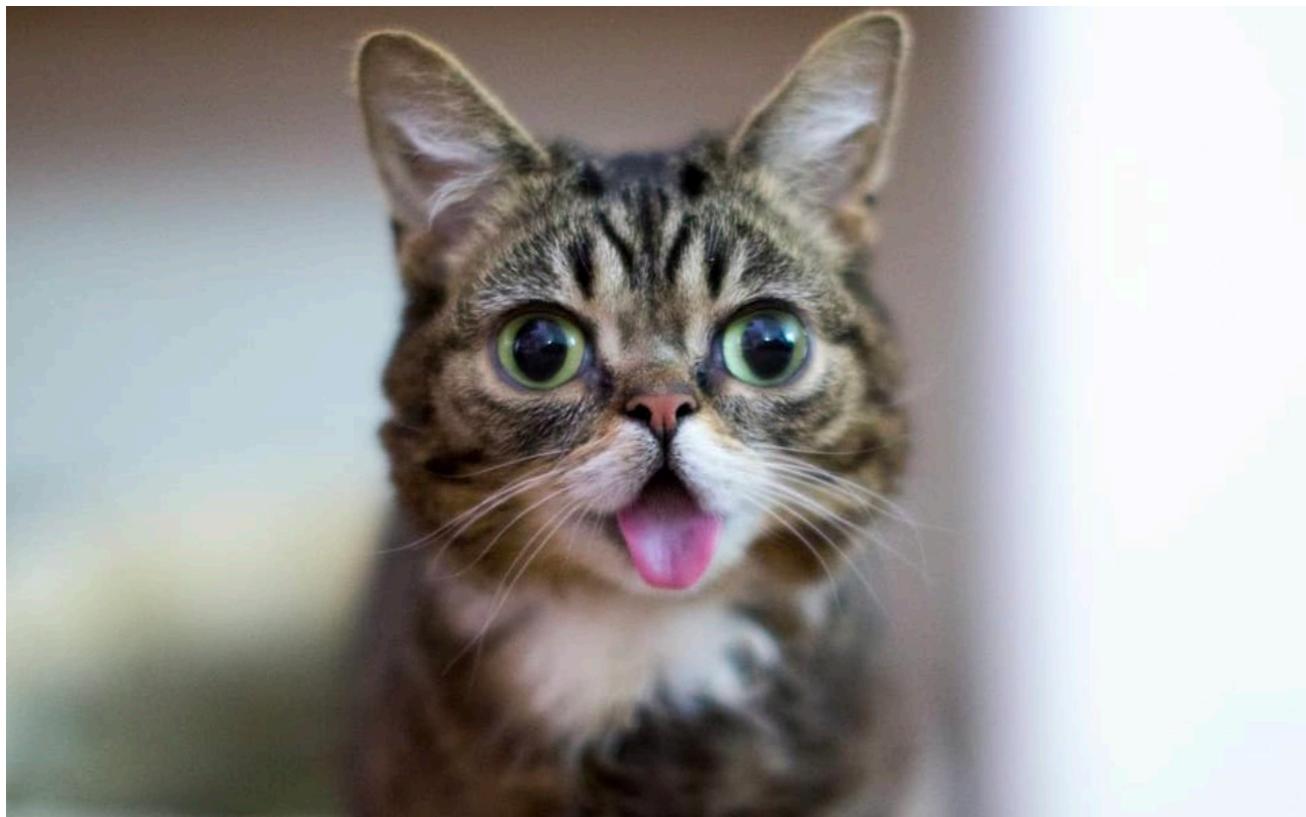


Сложные для человека проблемы



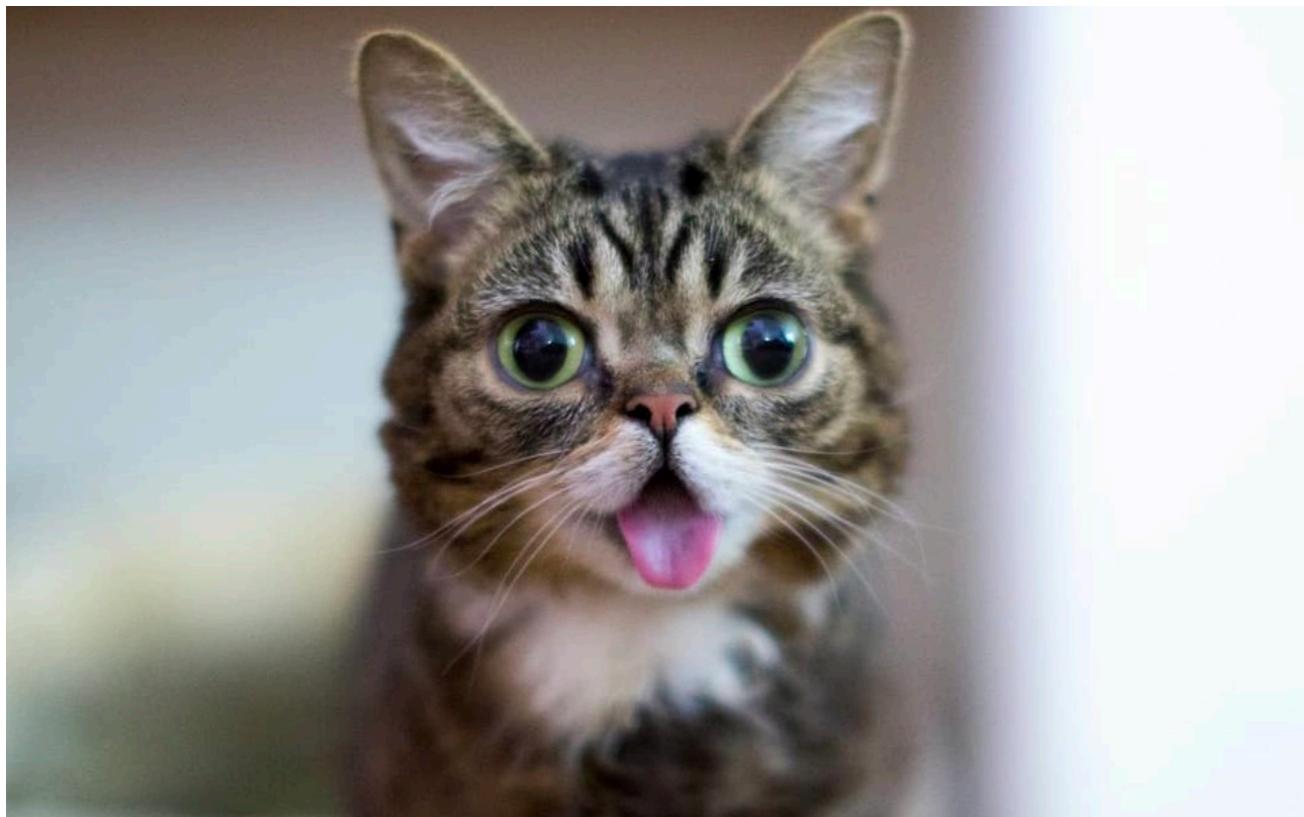
Для компьютера легки!

Легкие для человека

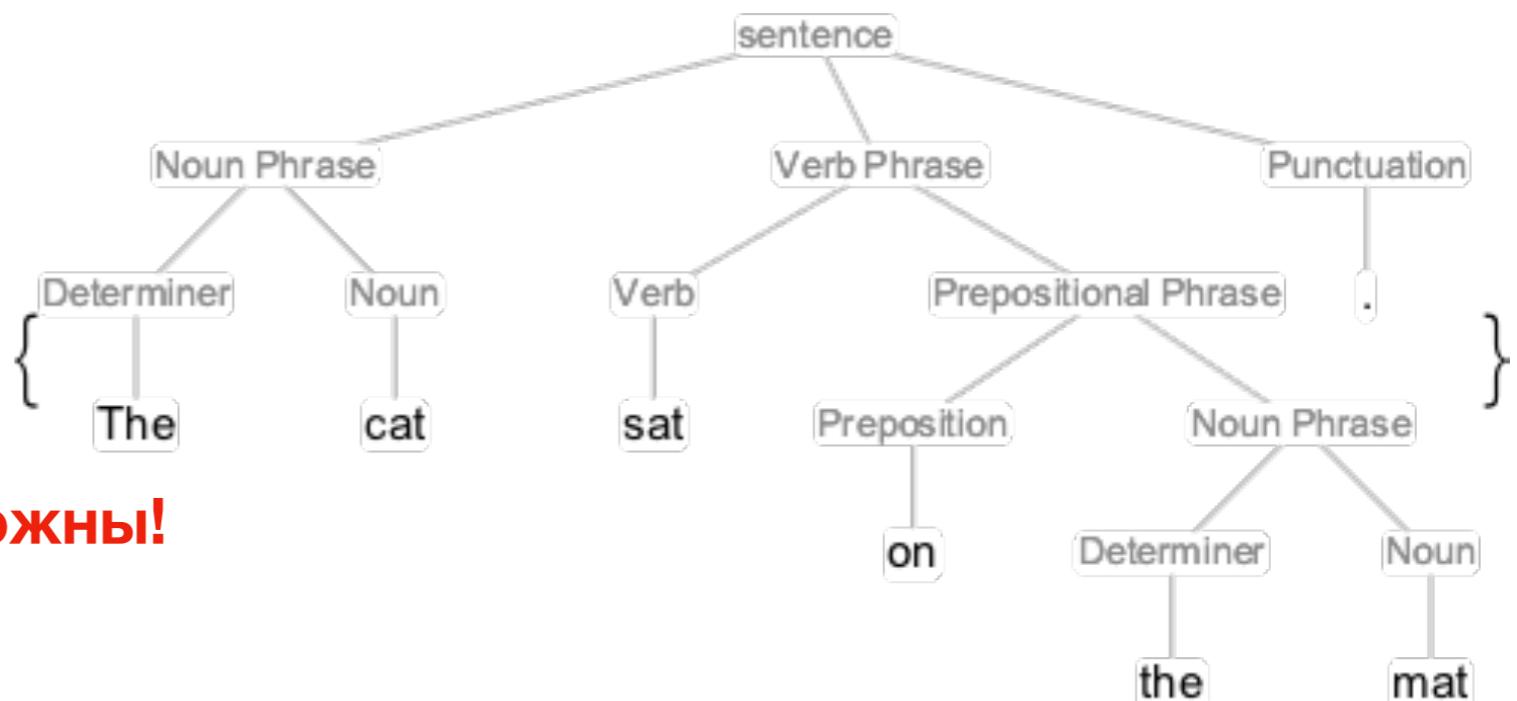


及锂离子电池用聚合物隔膜的制备方法。提供的制法包括如下步骤：将聚烯烃进行单向或双向拉伸形成微孔，制得聚合物采用聚烯烃与聚偏氟乙烯的混合物中制法，是将聚偏氟乙烯或者聚偏氟聚物用有机溶剂溶解均匀后，涂敷在之上，干燥后再进行单向或双向拉伸形貌与电极的粘接性能，提高电池的导电性。孔隙率较高，利用此隔膜制备的电池其制法可操作性强，工艺简单，易于推广。提供的聚合物对海松酸乙烯酯丙烯酰胺及制法。现阶段中，而纸张增韧剂则使用丙烯酰胺树脂将优良的增强剂、留剂的聚苯乙烯乙酸乙酯

Легкие для человека



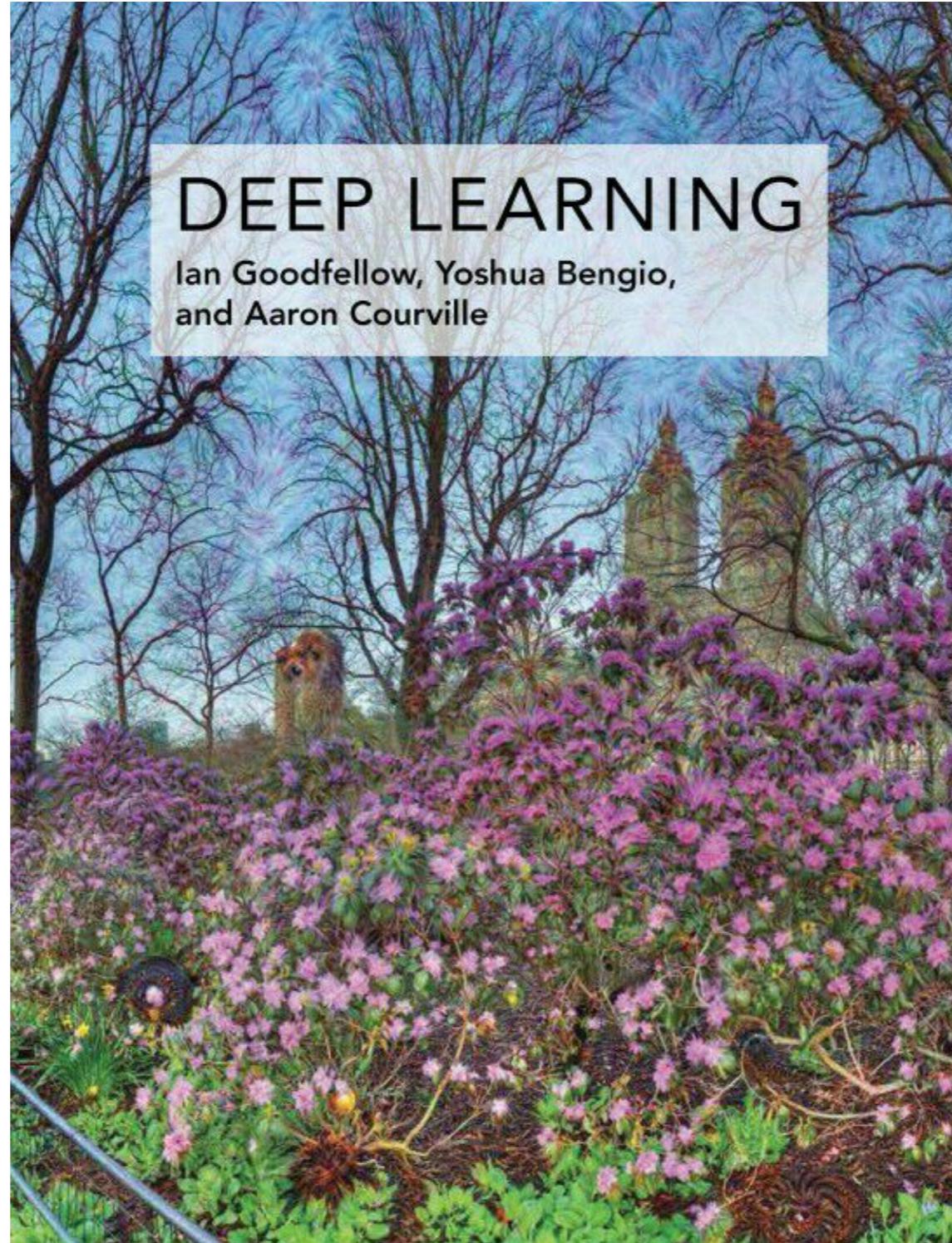
Для компьютера необычайно сложны!



及锂离子电池用聚合物隔膜的制备方法
是提供的制法包括如下步骤：将聚烯烃进
行单向或双向拉伸形成微孔，制得聚合
烃采用聚烯烃与聚偏氟乙烯的混合物
中制法，是将聚偏氟乙烯或者聚偏氟
聚物用有机溶剂溶解均匀后，涂敷在
之，干燥后再进行单向或双向拉伸形
成与电极的粘接性能，提高电池的导
电性。孔隙率较高，利用此隔膜制备的电
池其制法可操作性强，工艺简单，易于产
生。是提供的聚
物对
进
极
也
氏
容
解
均
化
度
在
聚合反应条件进
海松酸乙烯
酯丙烯酰胺
及制法。现
技术中，
而纸张增
剂则使用
酰胺树脂
将优良的
增强剂、
留剂的聚
差的
性。

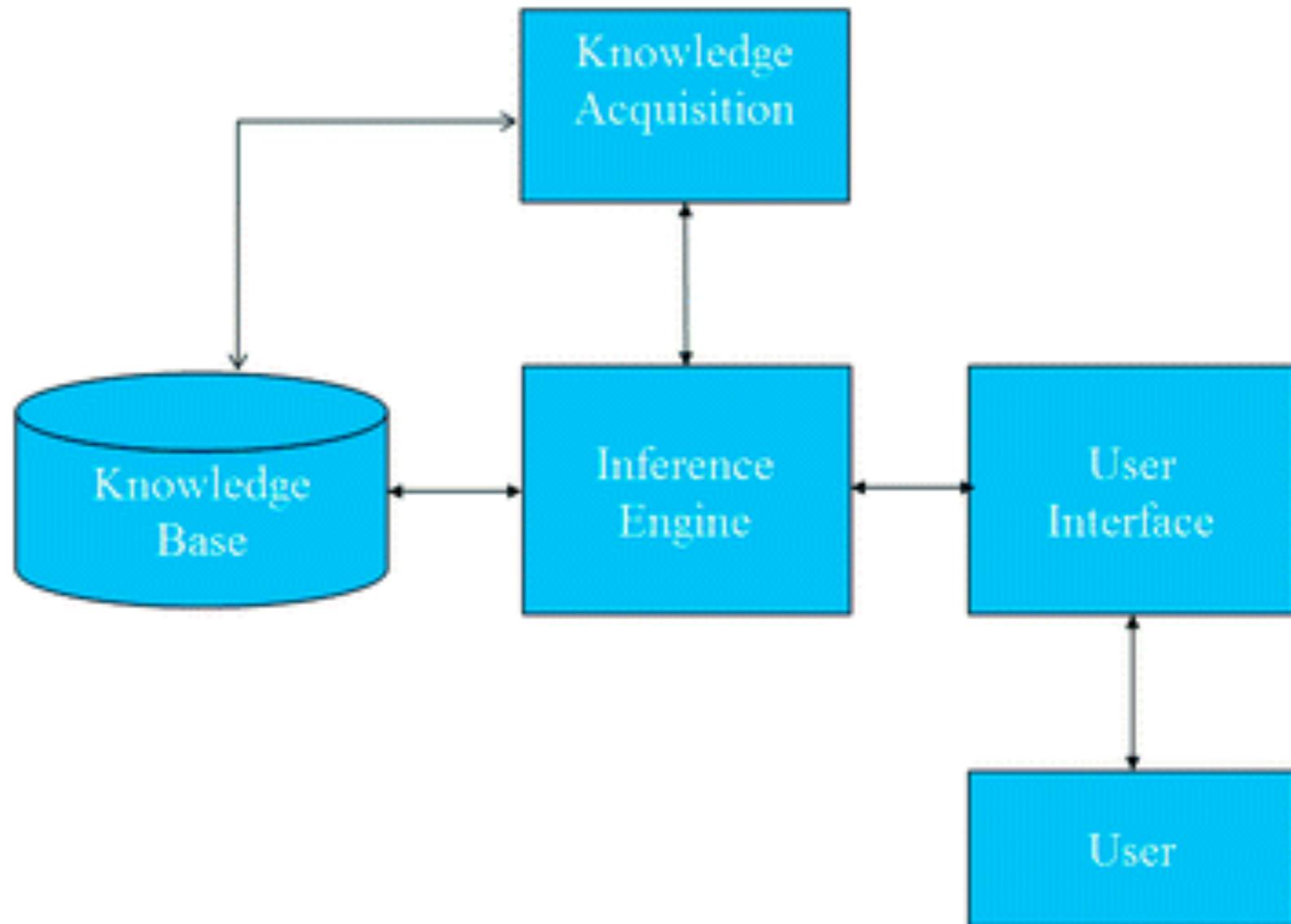
The true challenge to artificial intelligence proved to be solving the tasks that are easy for people to perform but hard for people to describe formally - problems the we solve intuitively..

Ian Goodfellow



Первая попытка - knowledge base approach

Основная идея - давайте все знания напишем на каком-то формальном языке и научим компьютер из этих знаний получать новые.



Prolog



Декларативный язык программирования
Вносим какую-то информацию и
отношения между объектами - благодаря развитой **системе**
доказательства теорем получаем новые знания

```
parent("Tom", "Jake").  
parent("Janna", "Jake").  
parent("Tom", "Tim").  
male("Tom").  
male("Tim").  
male("Jake").  
female("Janna").  
  
brother(X, Y) :-  
parent(Z, X), parent(Z, Y), male(X), male(Y), X \= Y.
```

Вывод:(Jake, Tim) (Tim, Jake)

Предполагалось, что станет языком компьютеров пятого поколения - устройств,
способных к имитации мышления.

И Prolog, и другие подобные языки не смогли выполнить поставленных задач.
Сейчас продолжает существовать, но сильно менее пафосно

Machine learning (ML)

The solution is to allow computers to learn from experience...

Ian Goodfellow

 ПОИСК машиналье обучение без регистрации и смс Найти :

Интернет Картинки Видео Новости Ответы

 Конференция по искусствен. интеллект / opentalks.ai
opentalks.ai/Конференция-ИИ Директ
19-21 февраля, лучшие российские докладчики на международных топ-конференциях
Контактные данные пн-пт 10:00-19:00

 ТОИР системы на базе 1С / 1cbit.ru
1cbit.ru/TOIR Директ
Внедряем 1С:ТОИР 1.3 и 1С:ТОИР 2 КОРП. Настроим, доработаем, проведем обучение!
1С:ERP Внедрение 1С Комплексная автоматизация Блог
Контактные данные м. Преображенская площадь круглосуточно 18+

 Машинное обучение в промышленности / proizvodstvo.zyfra.com
proizvodstvo.zyfra.com/ЦифровоеПроизводство Директ
Получите знания о стратегиях применения больших данных в промышленности!
Для инженеров Для студентов тех.вузов Для IT - Руководства Сертификат
Контактные данные пн-пт 9:00-18:00

 AWS Machine Learning / aws.amazon.com
aws.amazon.com/AWS-MI Директ
Машинное обучение на платформах TensorFlow, Caffe2, Apache MXNet и др. Выберите свою!
Обзор Статьи Начало работы Блог-EN

Машинное обучение

Алгоритм машинного обучения - это алгоритм, который способен обучаться, используя данные.

Компьютерная программа обучается на **опыте E** по отношению к какому-то **классу задач T**, если качество поведения этой программы, измеряемое при помощи **метрики P**, при учете программой опыта E, увеличивается.

Опыт E



Design matrix (матрица признаков)

	Признак 1	Признак 2	Признак п
Объект 1	x_{11}	x_{12}	x_{1m}
Объект 2	x_{21}	x_{22}	x_{2m}
.....
Объект m	x_{m1}	x_{m2}	x_{mm}

Согласно принятому в ML соглашению, строкам матрицы соответствуют объекты, а столбцам - признаки.

Supervised (с учителем) vs Unsupervised (без учителя)

В случае unsupervised learning у нас есть датасет с признаками. Задача алгоритма - выучить особенности структуры наших данных. Более формально - наш алгоритм должен выучить многомерное вероятностное распределение, из которого происходит наш датасет.

В случае supervised learning кроме у каждого объекта в датасете есть метка. Задача алгоритма эту метку научиться предсказывать (для любого объекта из вероятностного распределения, из которого происходит наш датасет).

Supervised (с учителем) vs Unsupervised (без учителя)

В случае unsupervised learning мы учим $p(x)$ (плотность распределения)

В случае supervised learning мы учим условную плотность $p(y|x)$ (нам надо для каждого x определить вероятные для него y)

Из формулы условной вероятности:

$$p(x) = p(x_1) \cdot p(x_2, \dots, x_n | x_1) = p(x_1) \cdot p(x_2 | x_1) \cdot p(x_3, \dots, x_n | x_1, x_2) = \dots = \prod_{i=1}^n p(x_i | x_1, \dots, x_{i-1})$$

Потому решение unsupervised-задачи можно свести к решению supervised-задач (x_i - признаки объекта)

Supervised (с учителем) vs Unsupervised (без учителя)

В случае unsupervised learning мы учим $p(x)$ (плотность распределения)

В случае supervised learning мы учим условную плотность $p(y|x)$ (нам надо для каждого x определить вероятные для него y)

При этом из теоремы Байеса

$$p(y|x) = \frac{p(x,y)}{\sum_{y'} p(x|y')}$$

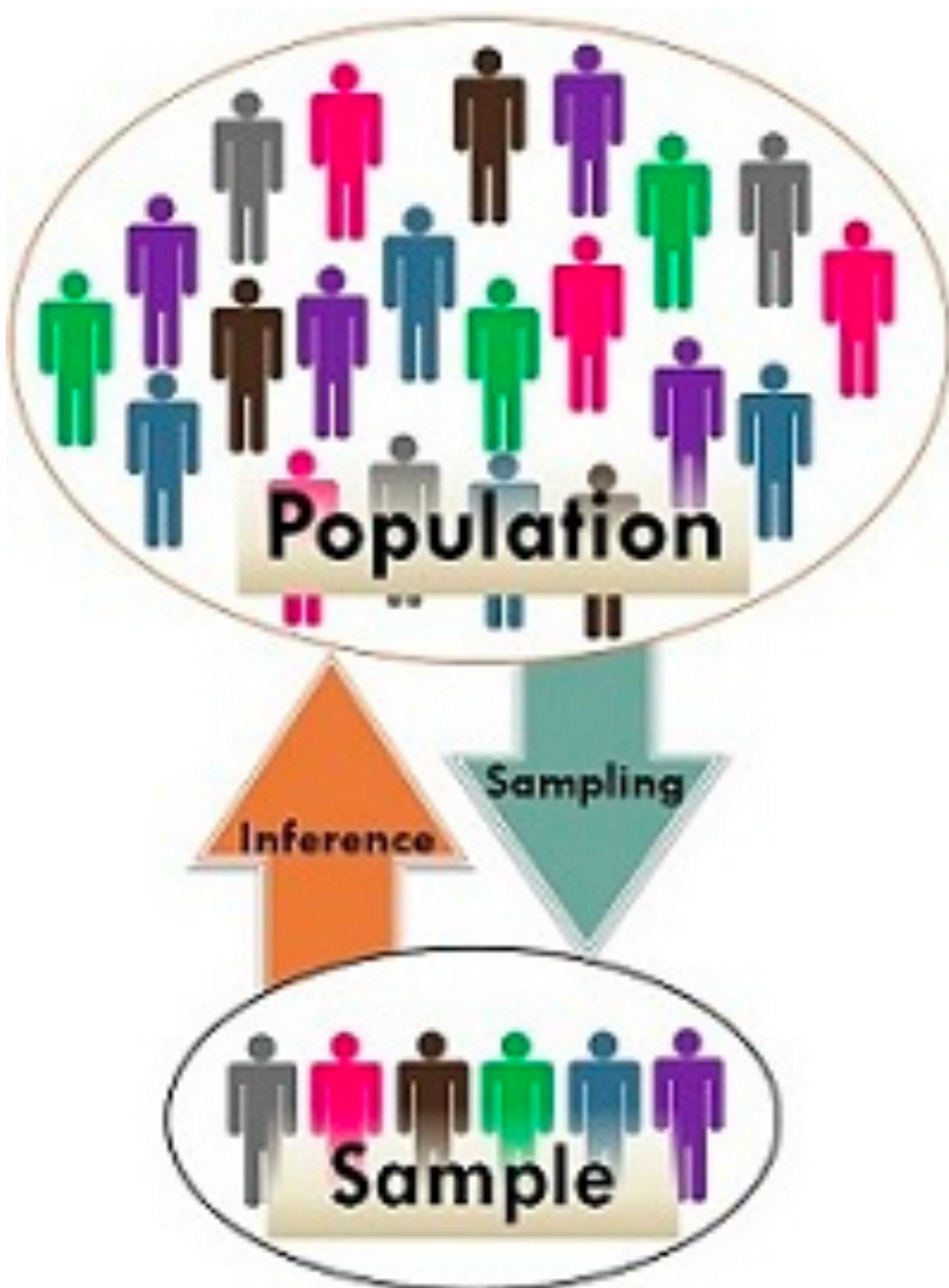
Потому решение supervised задачи можно свести к решению unsupervised задачи

Машинное обучение - часть статистики

Machine learning is essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions

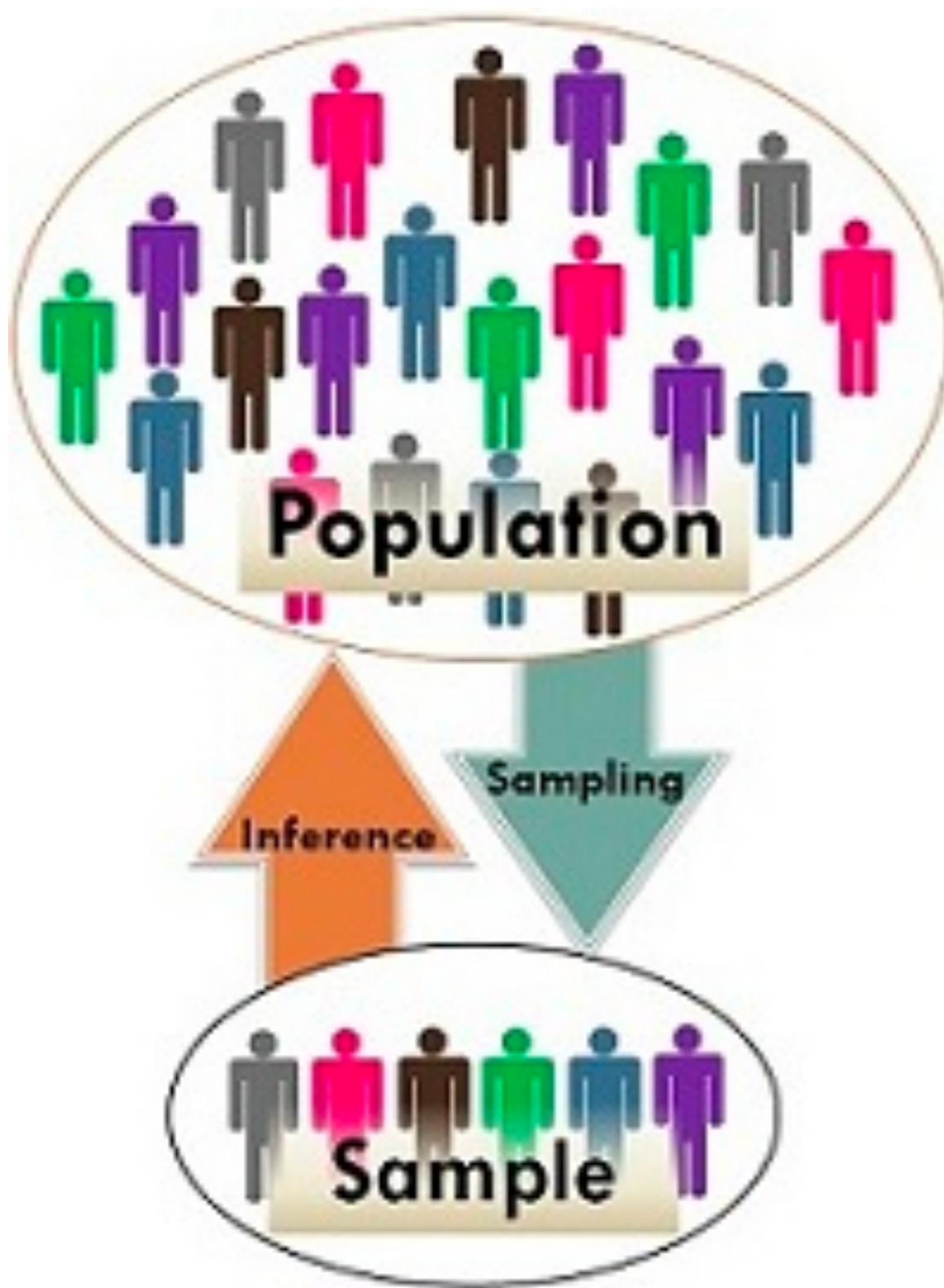
Ian Goodfellow

О чём вообще статистика



Имея выборку из генеральной совокупности, сделать выводы (inference) о генеральной совокупности

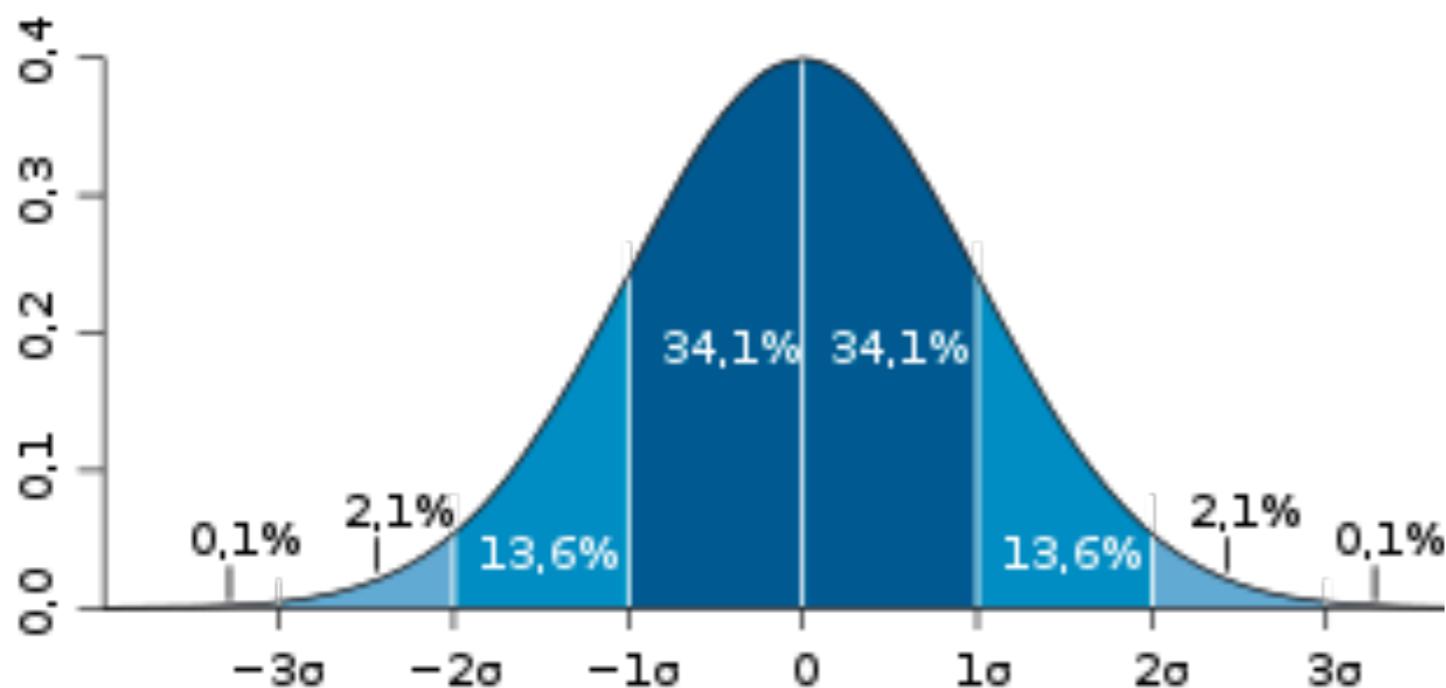
О чём вообще статистика



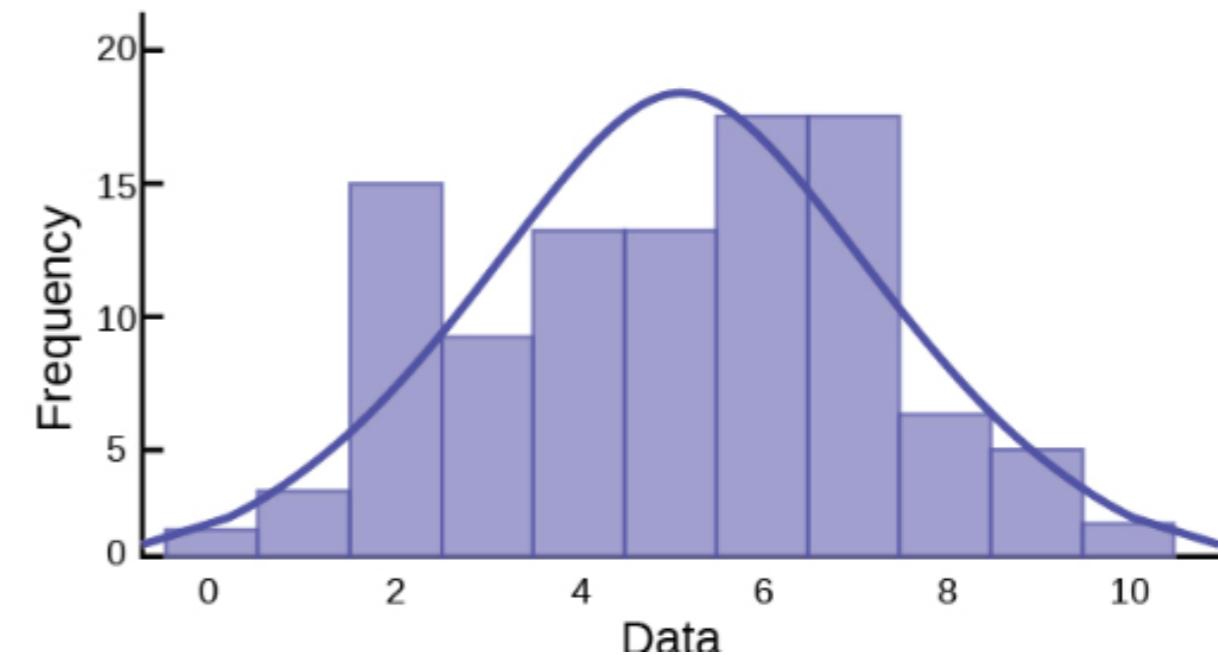
Имея выборку из генеральной совокупности, сделать выводы (inference) о генеральной совокупности

В машинном обучении мы точно также внутри модели формируем наше представление о генеральной совокупности

Оценка параметра



Генеральная совокупность



Параметры



Оценки параметров

Репрезентативность и все-все-все.

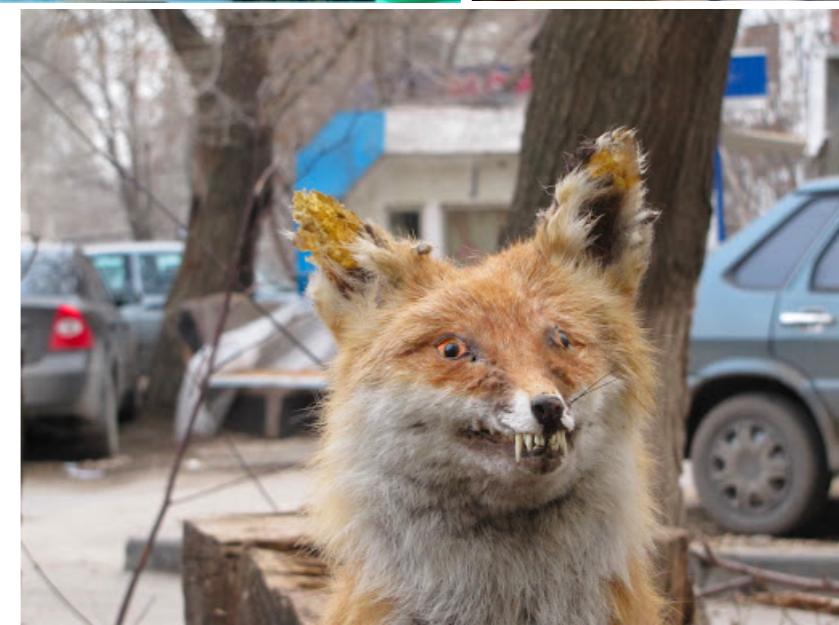
Машинное обучение - не панацея. Если набранные данные ничего не говорят о генеральной совокупности - получится что получится. Trash in - trash out.

Репрезентативность и все-все-все.

Генеральная
совокупность



Моя выборка



Классы задач (Т)

Probability is not a mere computation of odds on the dice or more complicated variants; it is the acceptance of the lack of certainty in our knowledge and the development of methods for dealing with our ignorance.

Outside of textbooks and casinos, probability almost never presents itself as a mathematical problem or a brain teaser. Mother nature does not tell you how many holes there are on the roulette table, nor does she deliver problems in a textbook way (in the real world one has to guess the problem more than the solution).

Fooled by Randomness. Nassim Nicholas Taleb

Классификация

Необходимо определить, к какой из К категорий принадлежит объект.

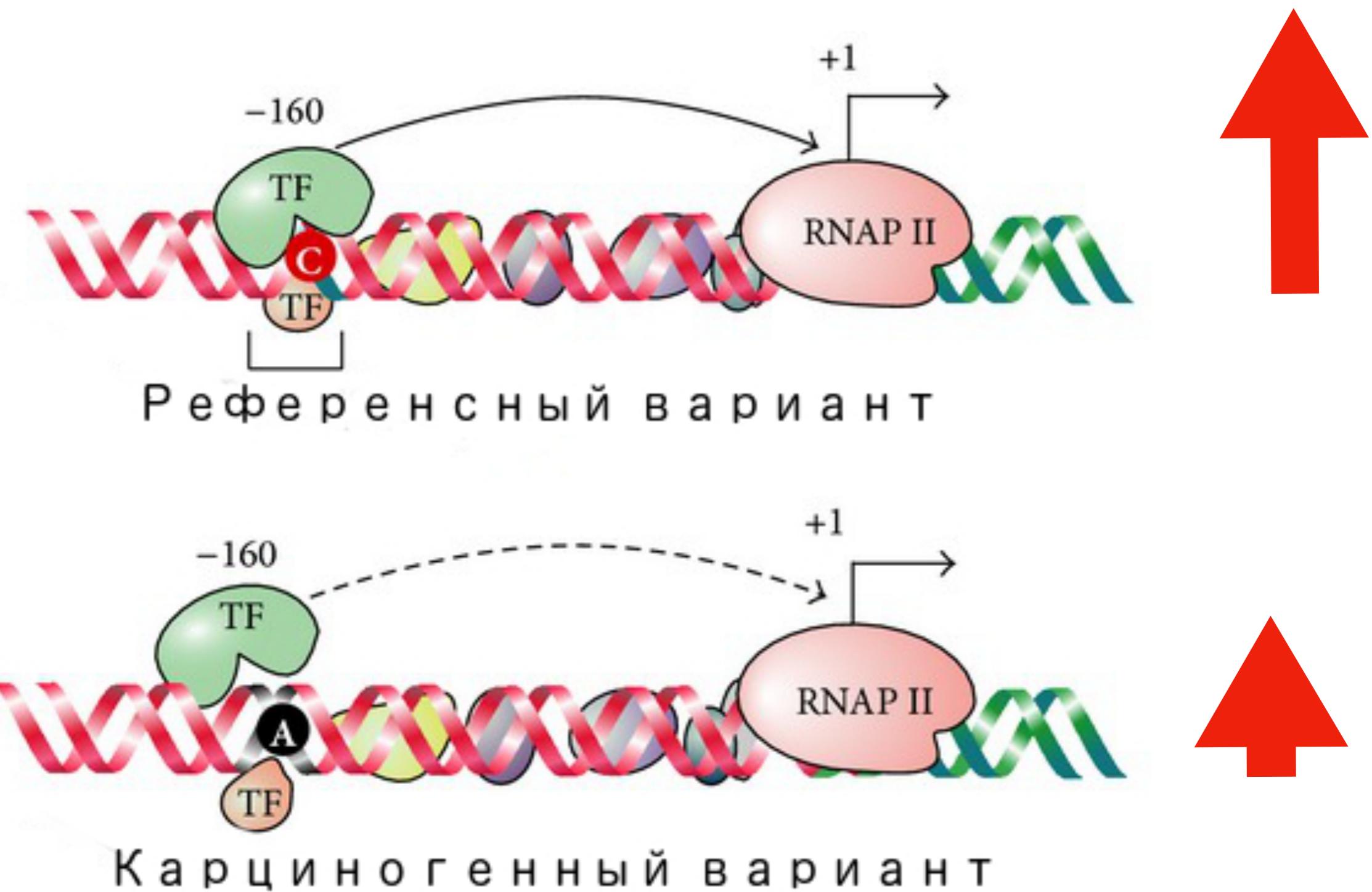
Более формально: хотим получить функцию $y = f(x)$, которая для каждого объекта x из некоего пространства выдает номер от 1 до K , обозначающий, к какой категории он принадлежит.

Еще более формально:

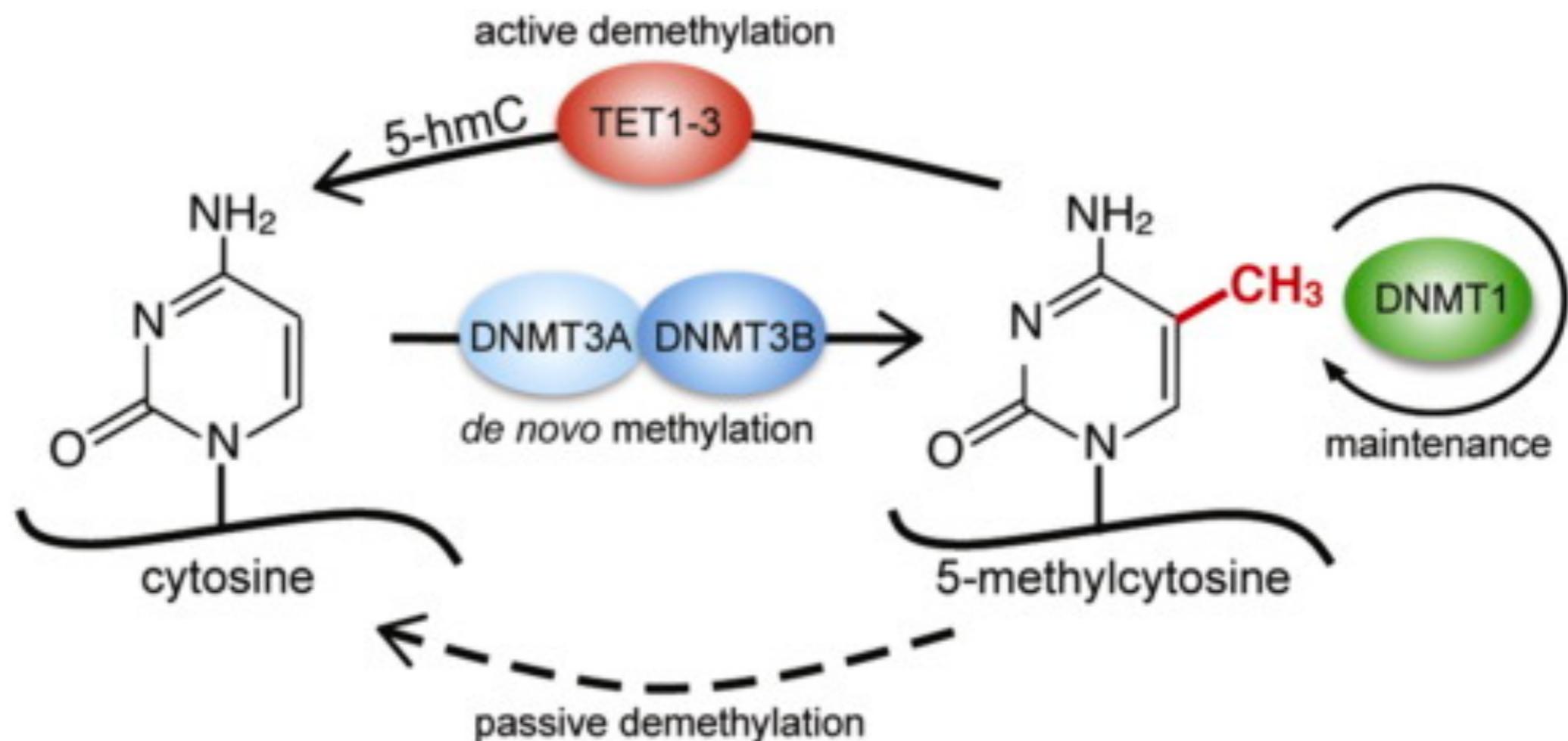
Есть некое пространство X . Существует некая функция $y=f(x)$, которая каждому объекту из X ставит в соответствие число от 1 до K .

Мы хотим найти функцию $h(x)$, которая наиболее точно аппроксимирует данную функцию

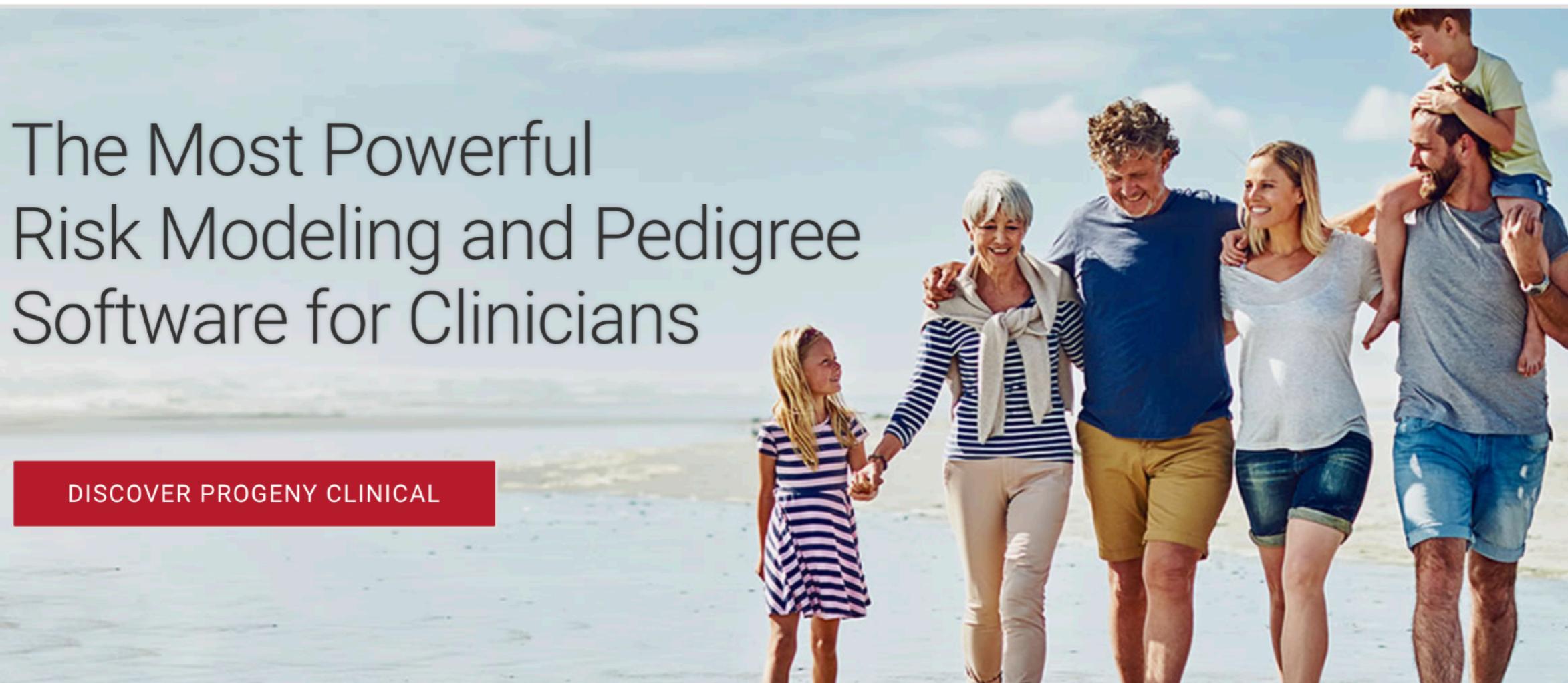
Предсказание класса SNV



Предсказание метилированных участков



Предсказание пациентов с опухолями

[PRODUCTS](#)[SERVICES](#)[SUPPORT](#)[COMPANY](#)[CONTACT](#)[DISCOVER PROGENY CLINICAL](#)

New High Risk Triage Screening Tools

See our quick screening tools to identify high risk patients for breast, colorectal and other cancers...

[See Features](#)

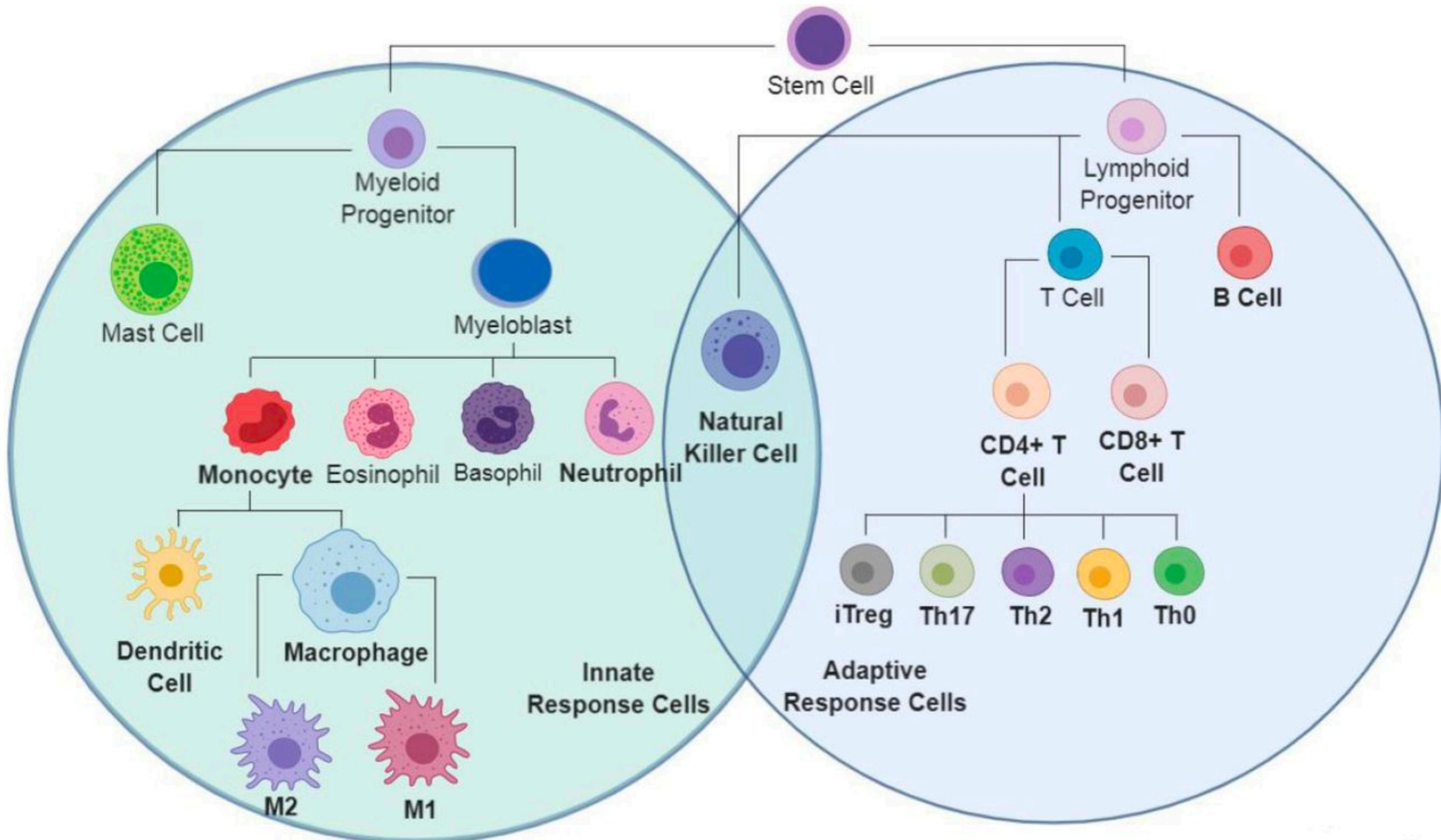
Классификация с пропущенными значениями

Теперь мы не знаем какой-то части информации про наши объекты (для каждого объекта часть признаков неизвестна)

Проблема та же самая, **но есть один нюанс**

Теперь не получится сделать **одну функцию**, решающую нашу задачу. Нам необходимо иметь на каждый случай пропуска данных свою функцию. Один из способов эффективно этого достичь - это научиться по известным признакам восстанавливать неизвестные, а потом уже все это подавать на вход одной функции

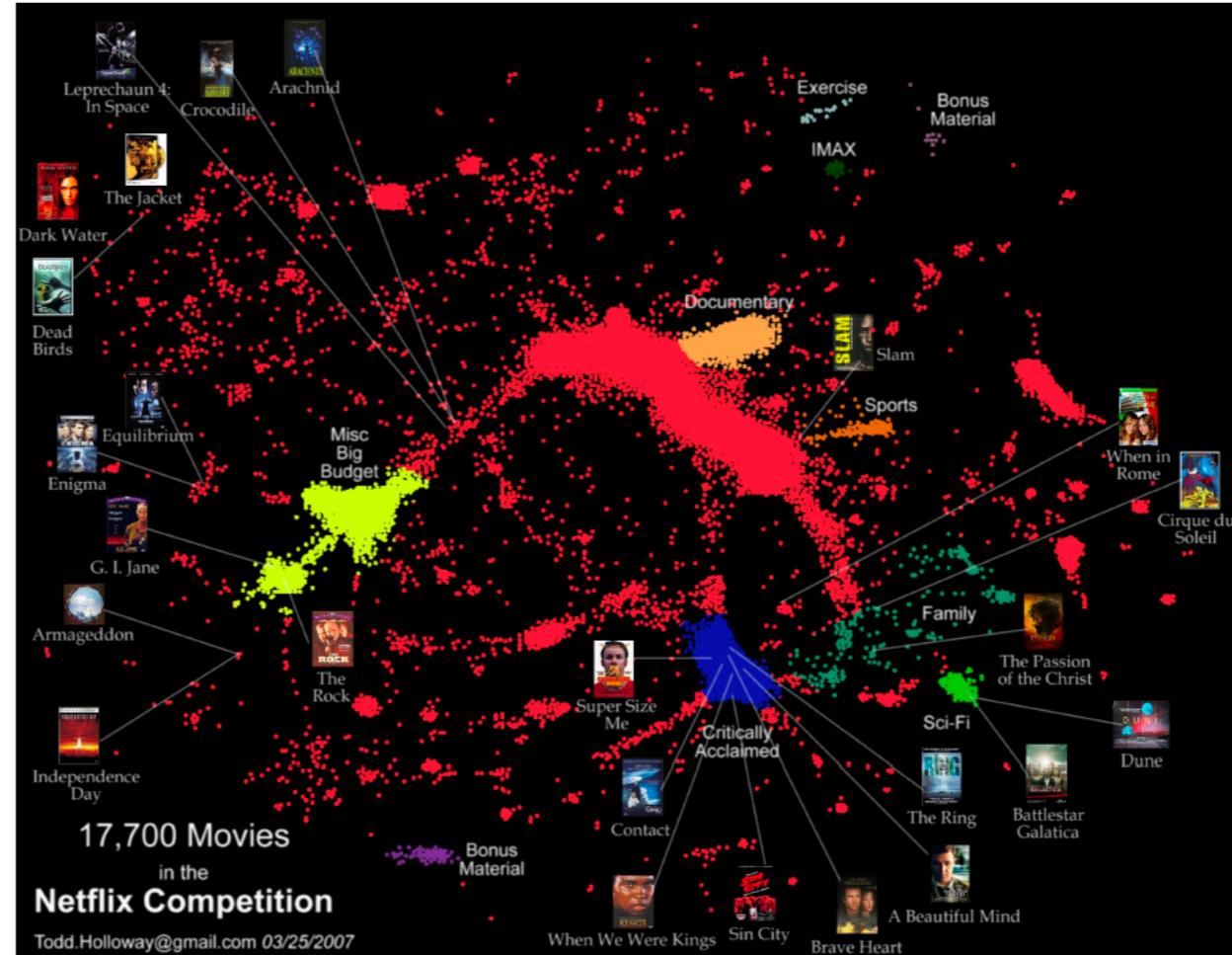
Предсказание типа клетки по данным single cell



Регрессия

Похоже на задачу классификации, но теперь $f(x)$ возвращает не метку от 1 до K, а какое-то вещественное число

Netflix Challenge



- ▶ Задача предсказания оценки фильму
- ▶ ПФ - 1000000\$
- ▶ Победил Stacking на “зоопарке” алгоритмов

ДОКИНГ

Target

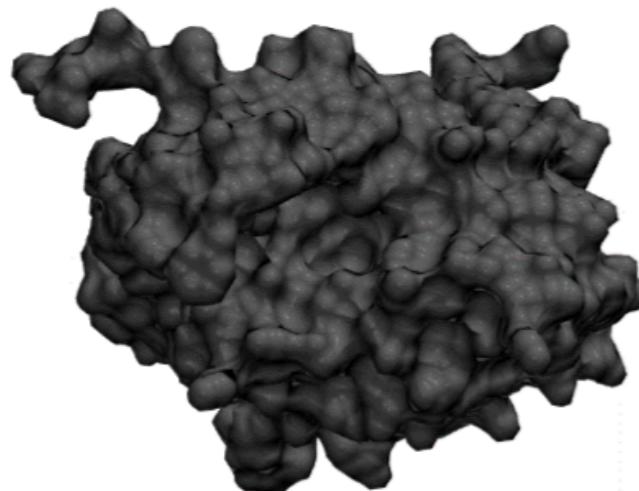
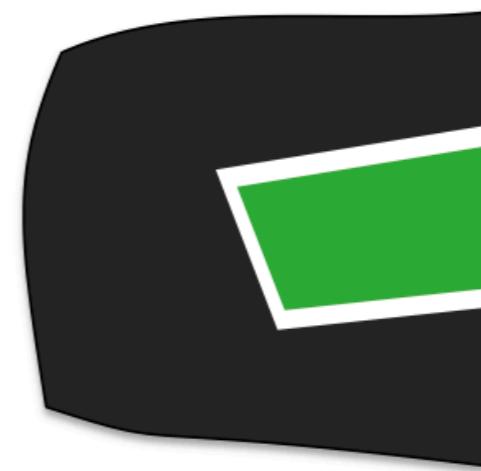


Ligand

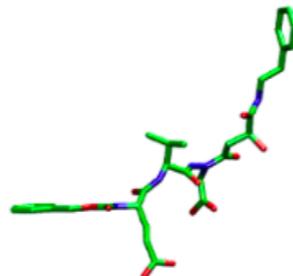


docking

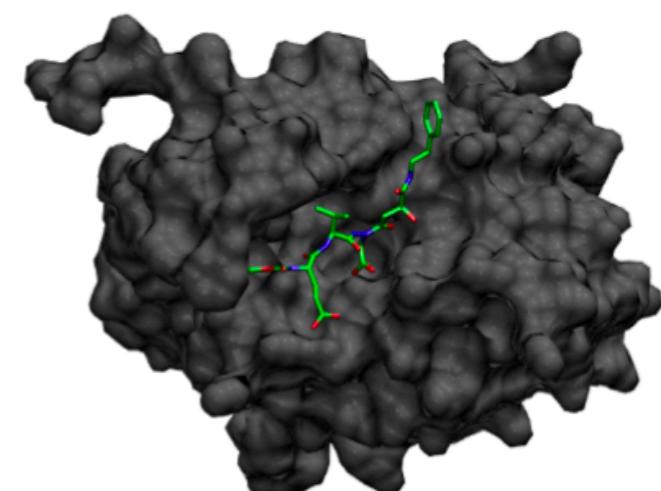
Complex



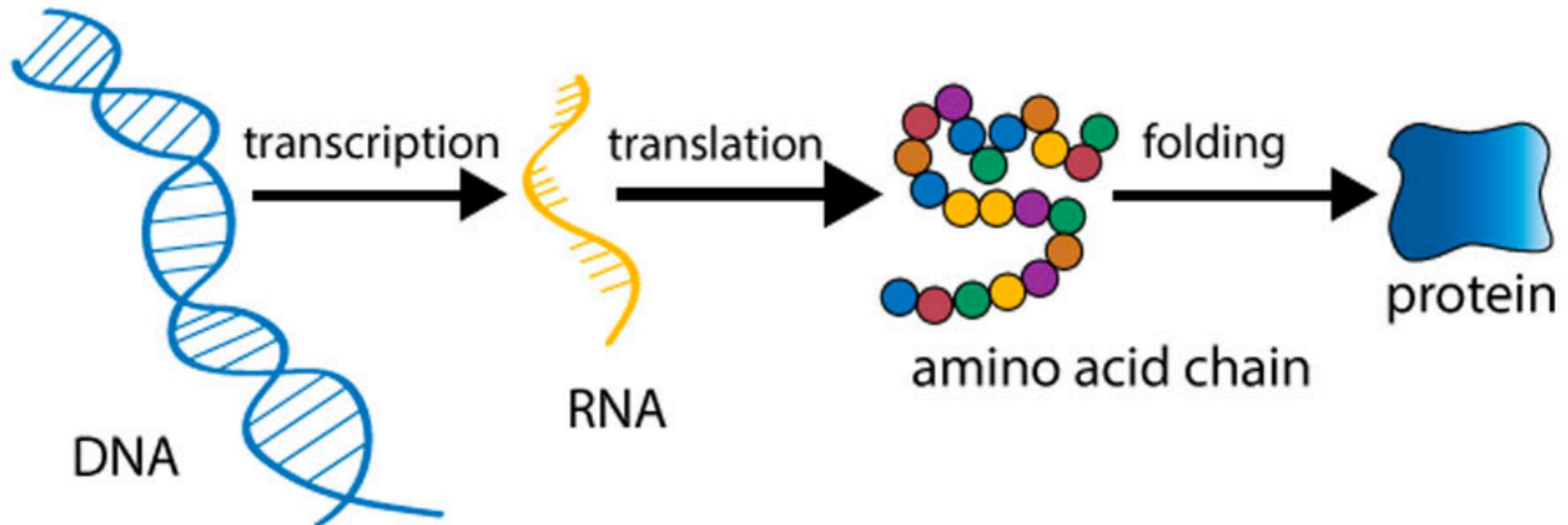
+



docking



Предсказание экспрессии гена



МОЁ ХОББИ: ЭКСТРАПОЛИРОВАТЬ



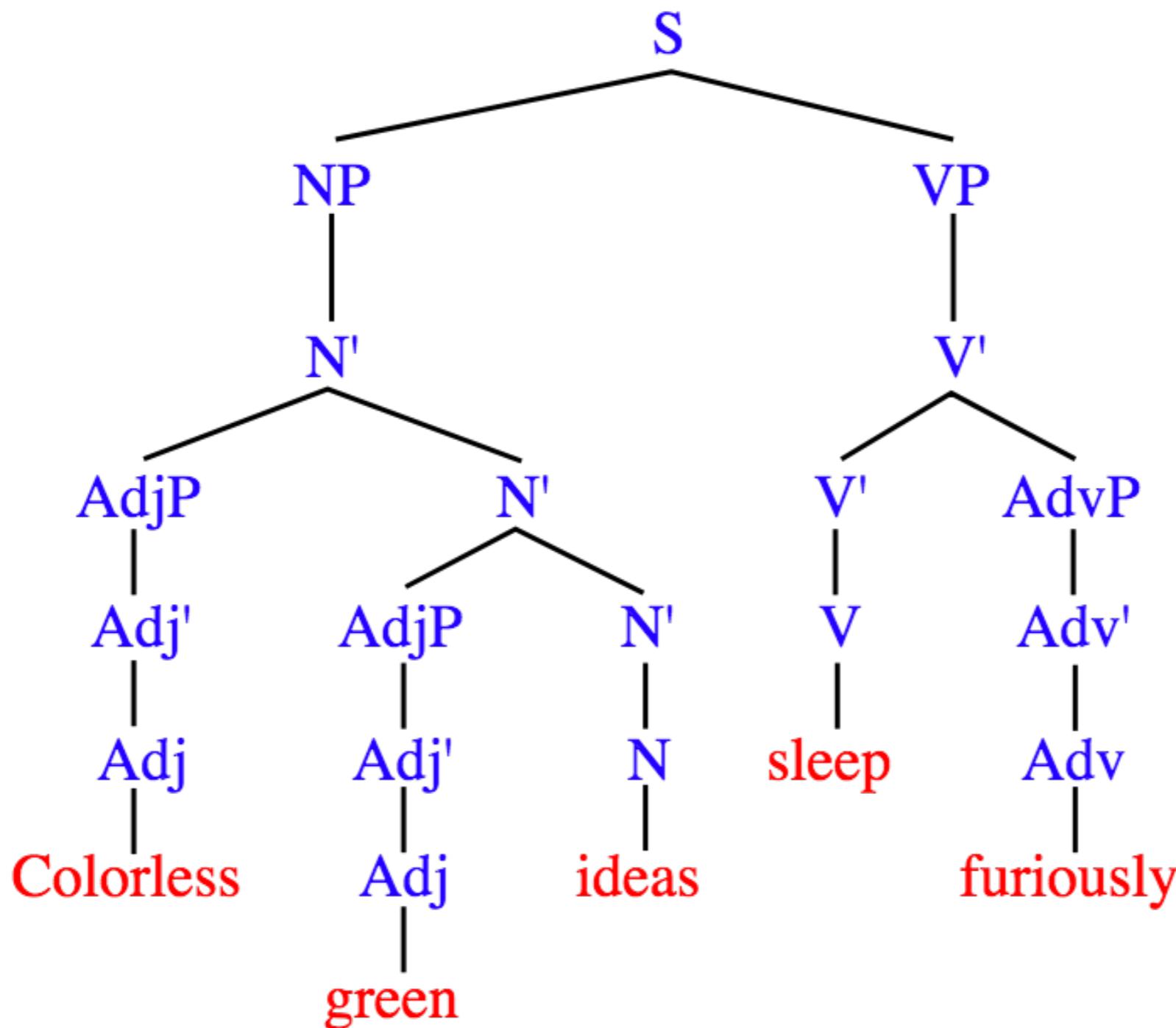
Выдача структурированного вывода

На вход дается сравнительно неструктурированная информация - на выход надо выдать информацию в структурированном виде.

Выдать номер машины



Парсинг структуры предложения



Перевести информацию из статьи в граф

Published online 13 August 2007

Nucleic Acids Research, 2007, Vol. 35, No. 16 5393–5401
doi:10.1093/nar/gkm584

Protein p56 from the *Bacillus subtilis* phage φ29 inhibits DNA-binding ability of uracil-DNA glycosylase

Gemma Serrano-Heras¹, José A. Ruiz-Masó², Gloria del Solar², Manuel Espinosa², Alicia Bravo² and Margarita Salas^{1,*}

¹Instituto de Biología Molecular 'Eladio Viñuela' (CSIC), Centro de Biología Molecular 'Severo Ochoa' (CSIC-UAM), Universidad Autónoma, Cantoblanco, 28049 Madrid and ²Centro de Investigaciones Biológicas (CSIC), Ramiro de Maeztu 9, 28040 Madrid, Spain

Received April 27, 2007; Revised July 16, 2007; Accepted July 17, 2007

ABSTRACT

Protein p56 (56 amino acids) from the *Bacillus subtilis* phage φ29 inactivates the host uracil-DNA glycosylase (UDG), an enzyme involved in the base excision repair pathway. At present, p56 is the only known example of a UDG inhibitor encoded by a non-uracil containing viral DNA. Using analytical ultracentrifugation methods, we found that protein p56 formed dimers at physiological concentrations. In addition, circular dichroism spectroscopic analyses revealed that protein p56 had a high content of β-strands (around 40%). To understand the mechanism underlying UDG inhibition by p56, we carried out *in vitro* experiments using the *Escherichia coli* UDG enzyme. The highly acidic protein p56 was able to compete with DNA for binding to UDG. Moreover, the interaction between p56 and UDG blocked DNA binding by UDG. We also demonstrated that Ugi, a protein that interacts with the DNA-binding domain of UDG, was able to replace protein p56 previously bound to the UDG enzyme. These results suggest that protein p56 could be a novel naturally occurring DNA mimicry.

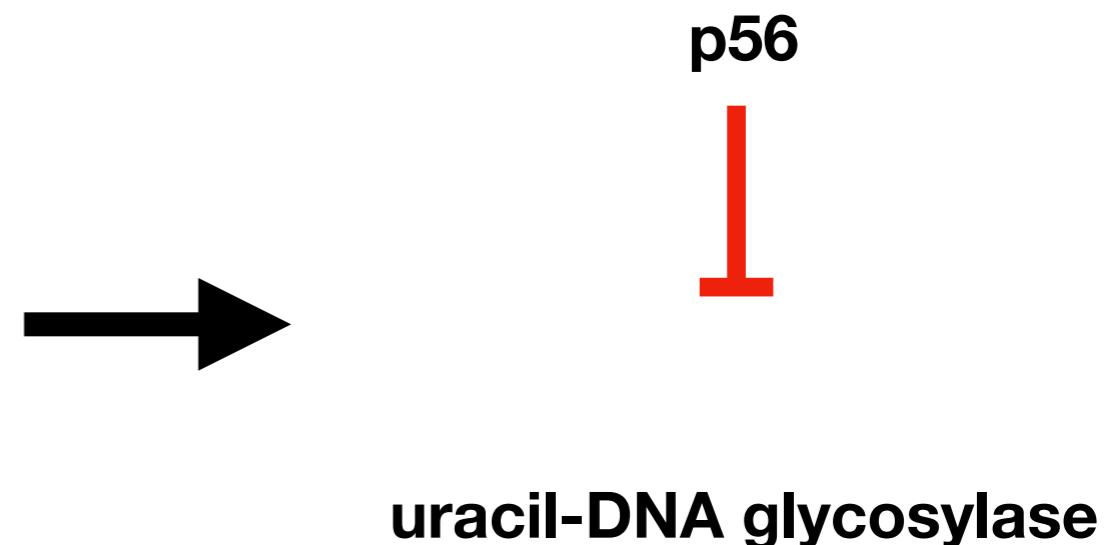
INTRODUCTION

Uracil in DNA may arise from the occasional use of dUTP during DNA replication and from spontaneous deamination of cytosine, which is one of the major mutagenic events in DNA. To maintain the integrity of the genetic information, most prokaryotic and eukaryotic cells encode uracil-DNA glycosylases (UDGs). These enzymes recognize and remove uracil residues from DNA by the base excision repair (BER) pathway. In human cells, five distinct UDG activities have been

identified namely UNG1, UNG2, TDG, MBD4 and SMUG (1). UNG2 is known to enter the nucleus while the isoform UNG1 enters the mitochondria (2). Moreover, UNG2 plays an important role in immunoglobulin gene diversification (3) and is incorporated into virions of the human immunodeficiency virus type-1 (4,5). Some DNA viruses, such as herpesviruses and poxviruses, also encode a UDG activity. In these instances, the UDG activity appears to have an important role in virus replication (6).

The first UDG activity reported was purified from *Escherichia coli* cells. Since then, enzymes highly homologous to the archetypal *E. coli* UDG have been purified from numerous organisms, including herpes simplex virus type-1 and human cells (UNG1 and UNG2 enzymes). These UDGs (Family-1) are able to eliminate uracil bases efficiently from both single-stranded (ss) and double-stranded (ds) DNAs regardless of the partner base, U:A or U:G (7). However, in some cases, a preference for the ssDNA substrates has been reported (8,9). Furthermore, a mismatch-specific uracil-DNA glycosylase (MUG) was purified from *E. coli* cells (10). This enzyme, which is related to human thymine-DNA glycosylase (TDG) (11), is exclusively active against U:G mismatches. Both MUG and TDG are members of the Family-2 UDGs (7).

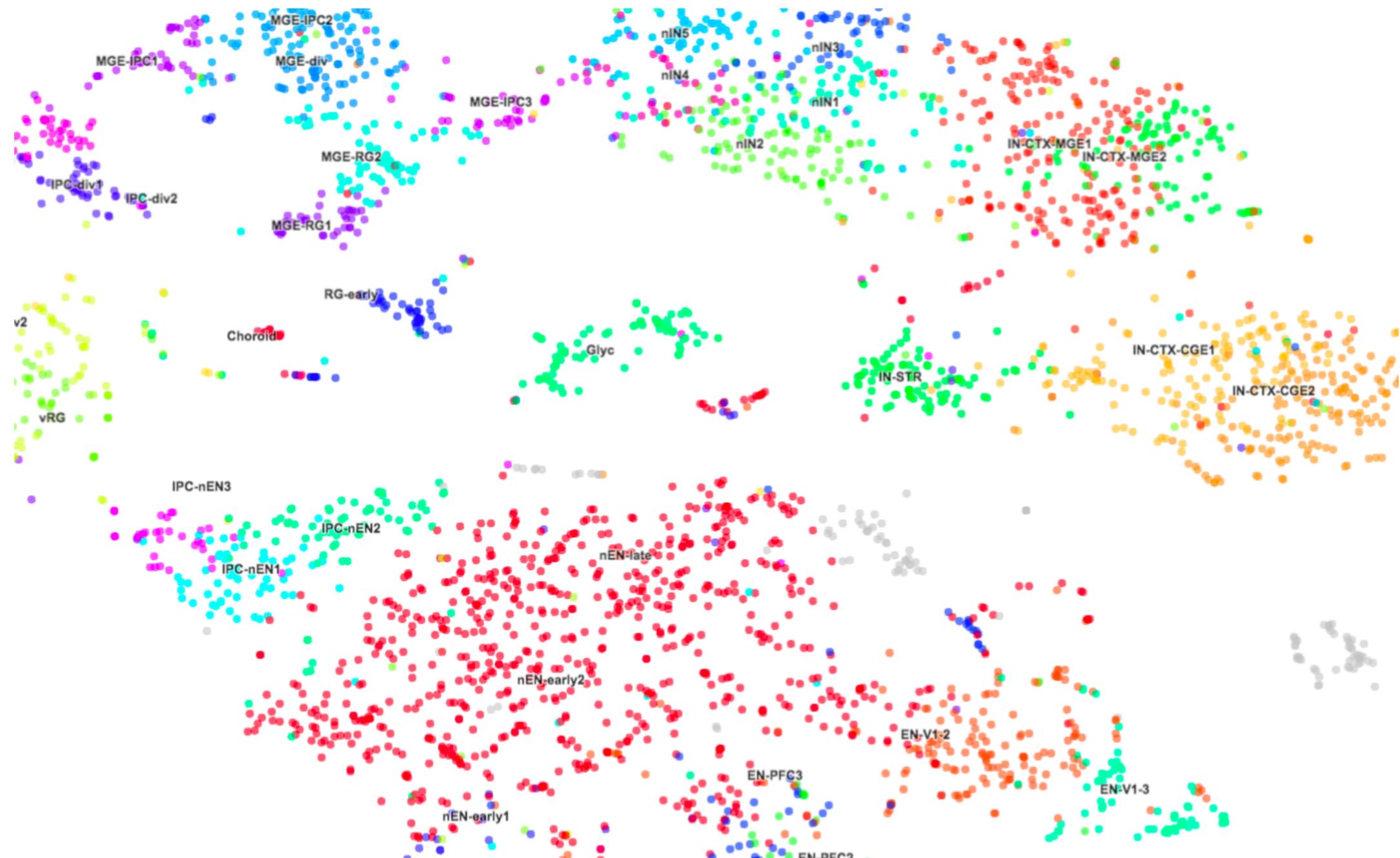
During the last years, UDGs are emerging as attractive therapeutic targets due to their role in a wide range of biological processes. Hence, the discovery of small molecules able to inhibit the activity of particular UDGs has a great interest. In addition, the knowledge generated by studying new UDG inhibitors should provide further insights into the process of substrate recognition and catalysis by UDGs. The first natural UDG inhibitor reported was Ugi, a highly acidic protein (84 amino acids) encoded by the *Bacillus subtilis* phage PBS2, whose DNA genome is unusual in that it contains uracil instead of thymine (12). Ugi inactivates Family-1 UDGs from *B. subtilis*, *E. coli*, *Micrococcus luteus*,



Машинный перевод

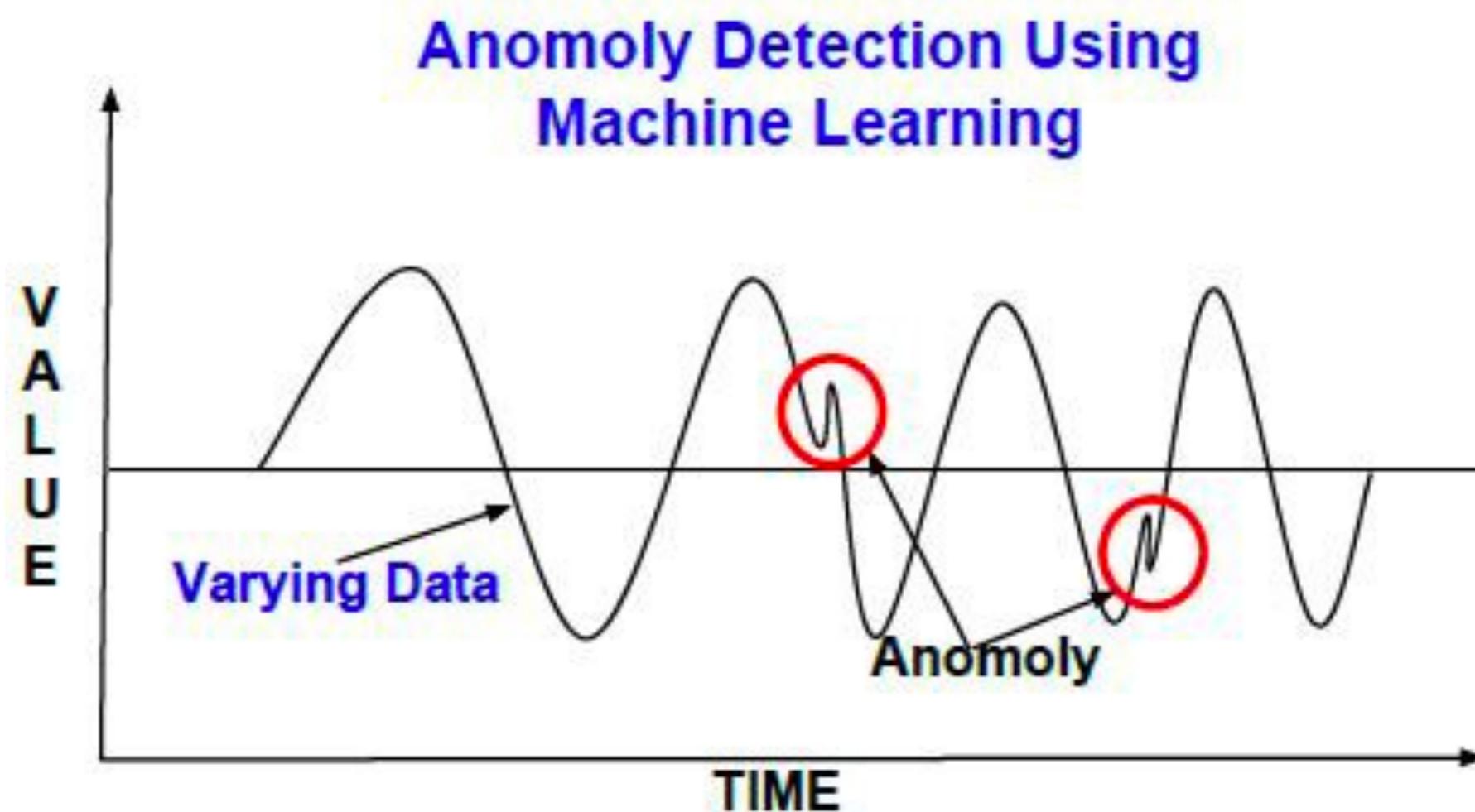


“Перевести” данные экспрессии генов в одном клеточном типе в данные экспрессии в другом клеточном типе



Детекция аномалий

Программа отмечает объекты, которые видит, либо как нормальные (похожие на те, что демонстрировали ей при обучении), либо как аномальные.



Детекция мошенничества с банковскими картами

У каждого человека есть определенный паттерн покупок и т.д. У мошенника паттерн будет совсем другой - программа может заметить это и заблокировать карту.

Детекция рака

BRIEF RESEARCH REPORT ARTICLE

Front. Genet., 02 July 2019 | <https://doi.org/10.3389/fgene.2019.00599>

Cancer as a Tissue Anomaly: Classifying Tumor Transcriptomes Based Only on Healthy Data

 Thomas P. Quinn^{1,2,3*},  Thin Nguyen¹,  Samuel C. Lee¹ and  Svetha Venkatesh¹

¹Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, VIC, Australia

²Centre for Molecular and Medical Research, Deakin University, Geelong, VIC, Australia

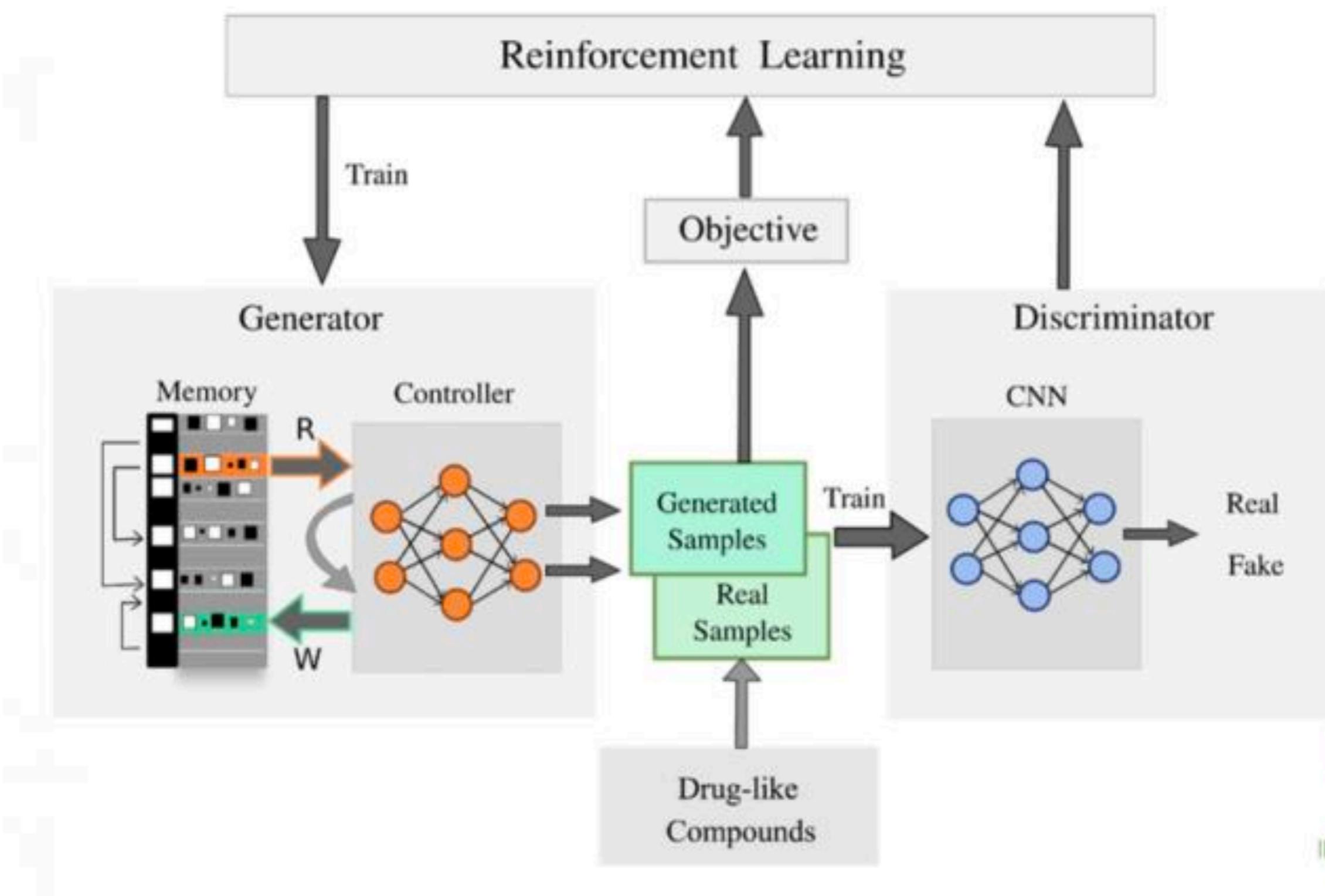
³Bioinformatics Core Research Group, Deakin University, Geelong, VIC, Australia

Детекция аномальных последовательностей ДНК

Генерация и сэмплирование

Алгоритм должен научиться генерировать данные,
похожие на те, что были в обучающей выборке

Генерация новых лекарств



INSILICO MEDICINE
insilico.com

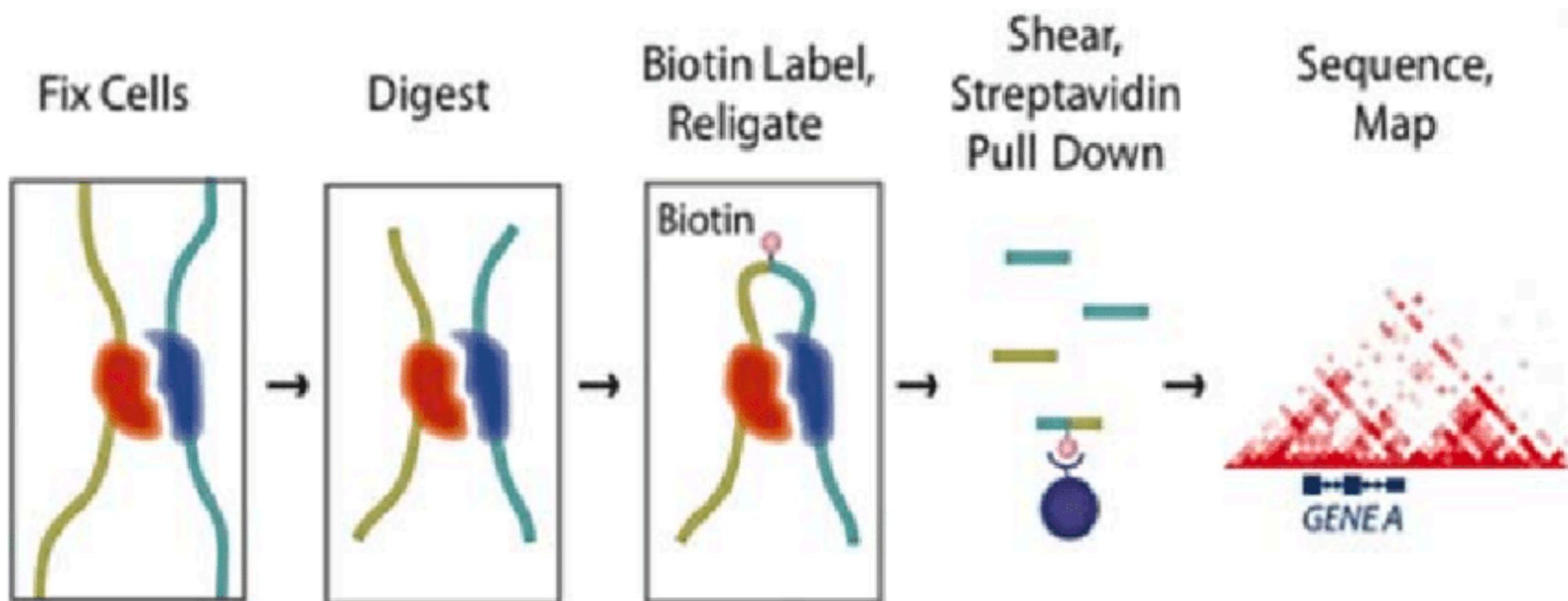
Вставка пропущенных значений и удаление шума

Дается объект x . Надо угадать недостающие значения/
удалить шум из объекта.

Повышение разрешения фотографий



Повышение разрешения результатов Hi-C



Оценка качества

Специфична для каждого задания. Не всегда просто подобрать, приходится комбинировать разные метрики.

Модели

