



Машинное обучение в задачах наук о Земле

М.А. Криницкий

Михаил Криницкий

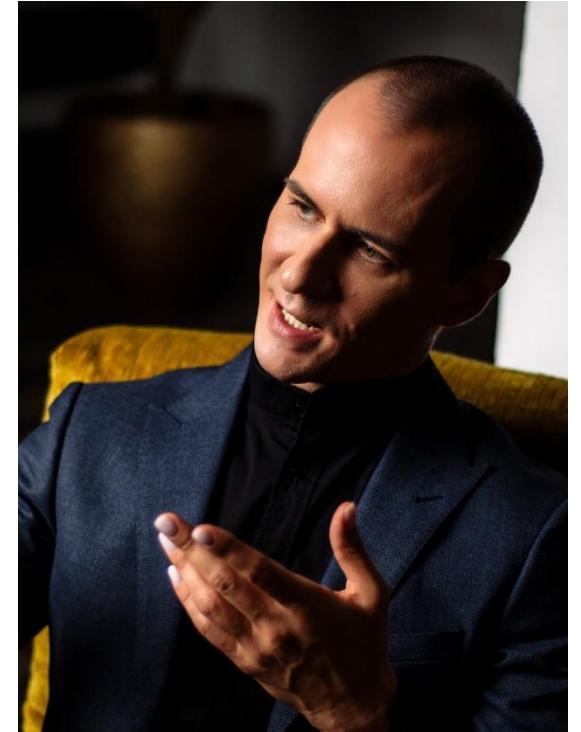
к.т.н.

Зав. Лабораторией машинного обучения в науках о Земле МФТИ

с.н.с. ИО РАН

доцент МФТИ

разработка и анализ алгоритмов, основанных на методах
машииного обучения и глубокого обучения, в применении к
задачам наук о Земле



krinitsky@sail.msk.ru

<https://t.me/mkrinitskiy>



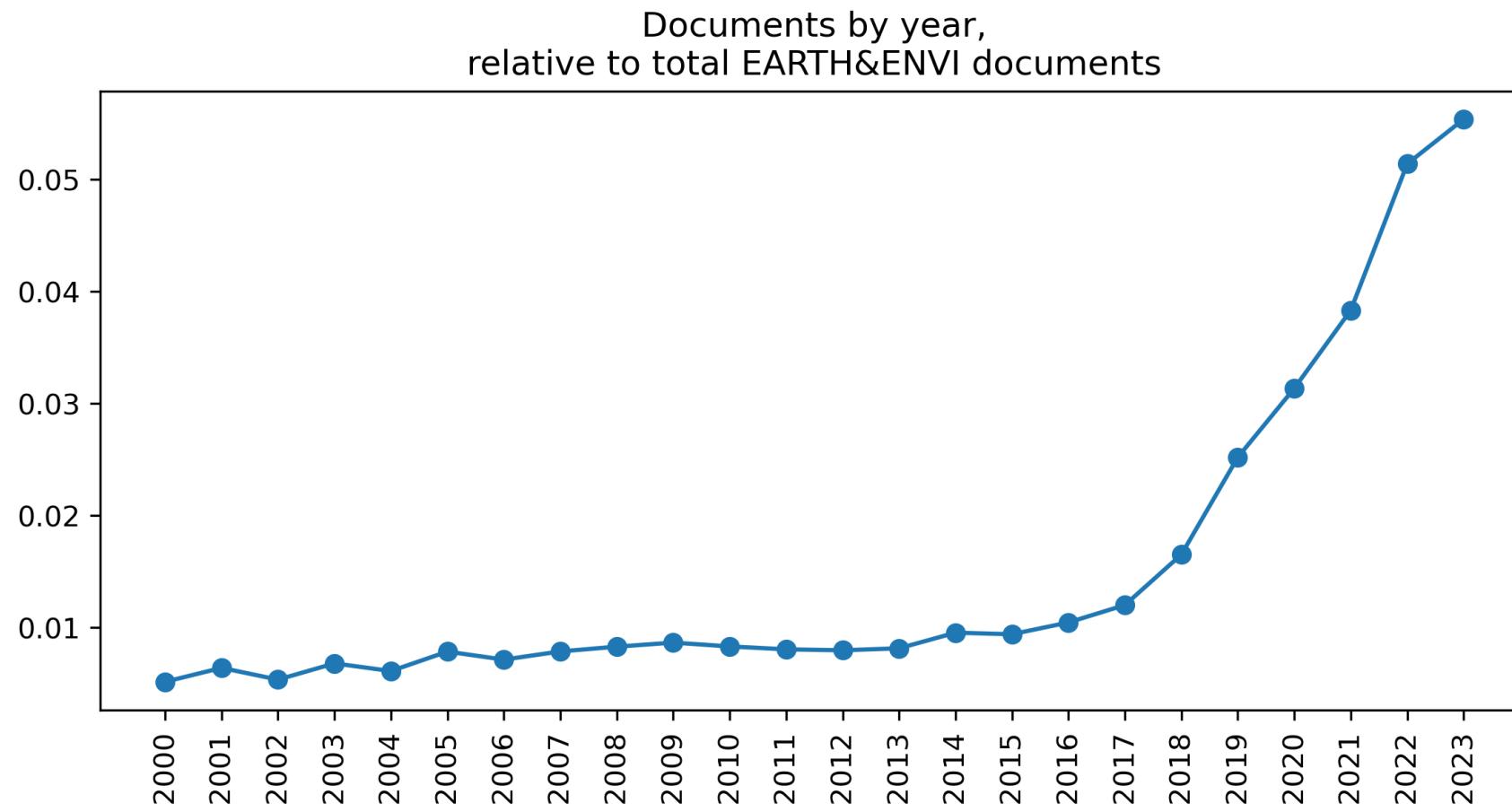


Машинное обучение: Введение

Осторожно, демотиватор

МО в науках о Земле

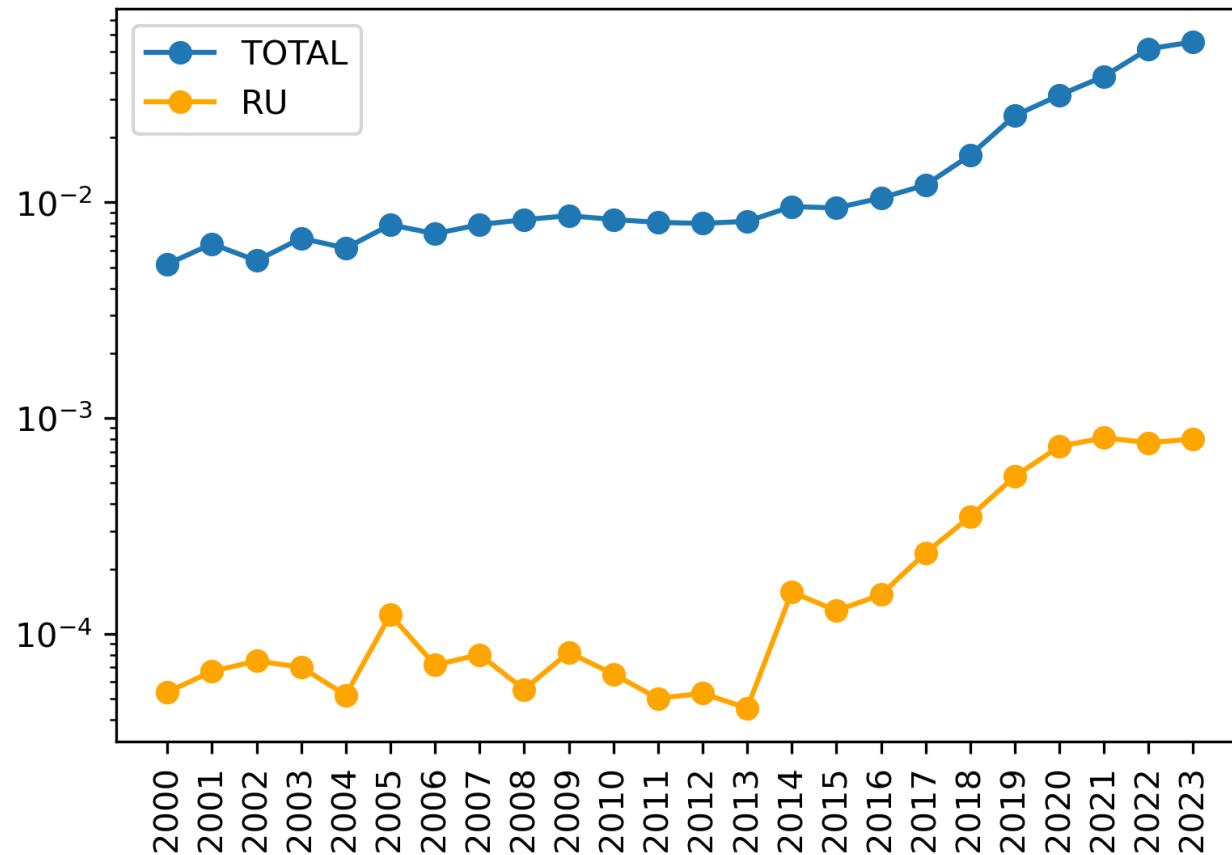
Статьи в рецензируемых журналах по тематике “Environmental sciences” и “Earth sciences” с применением методов машинного обучения. По данным Scopus.



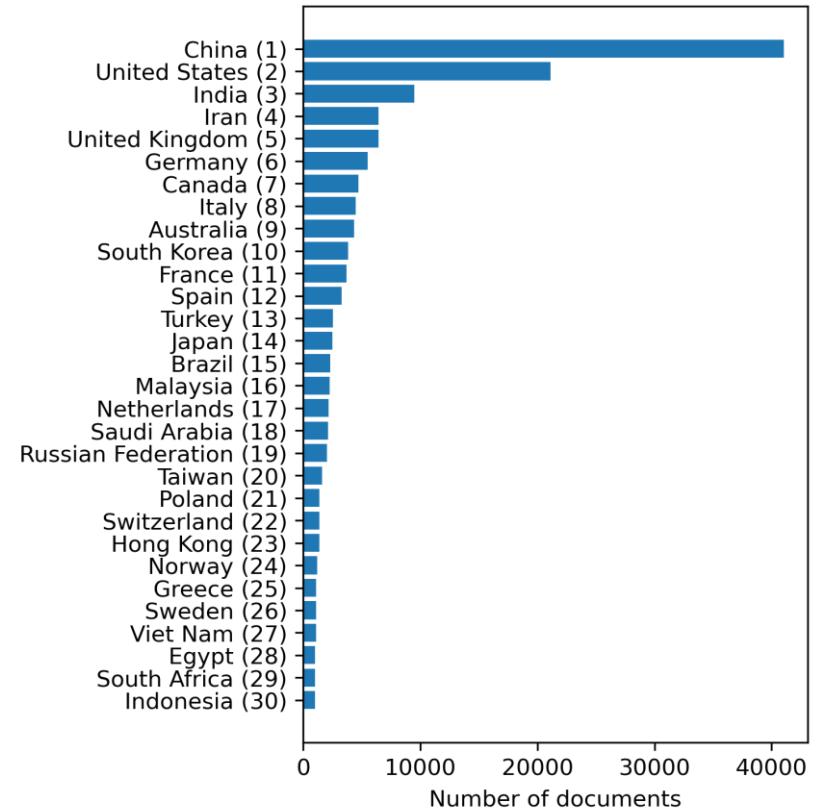
МО в науках о Земле

Статьи в рецензируемых журналах по тематике “Environmental sciences” и “Earth sciences” с применением методов машинного обучения. По данным Scopus.

Documents by year,
relative to total EARTH&ENVI documents

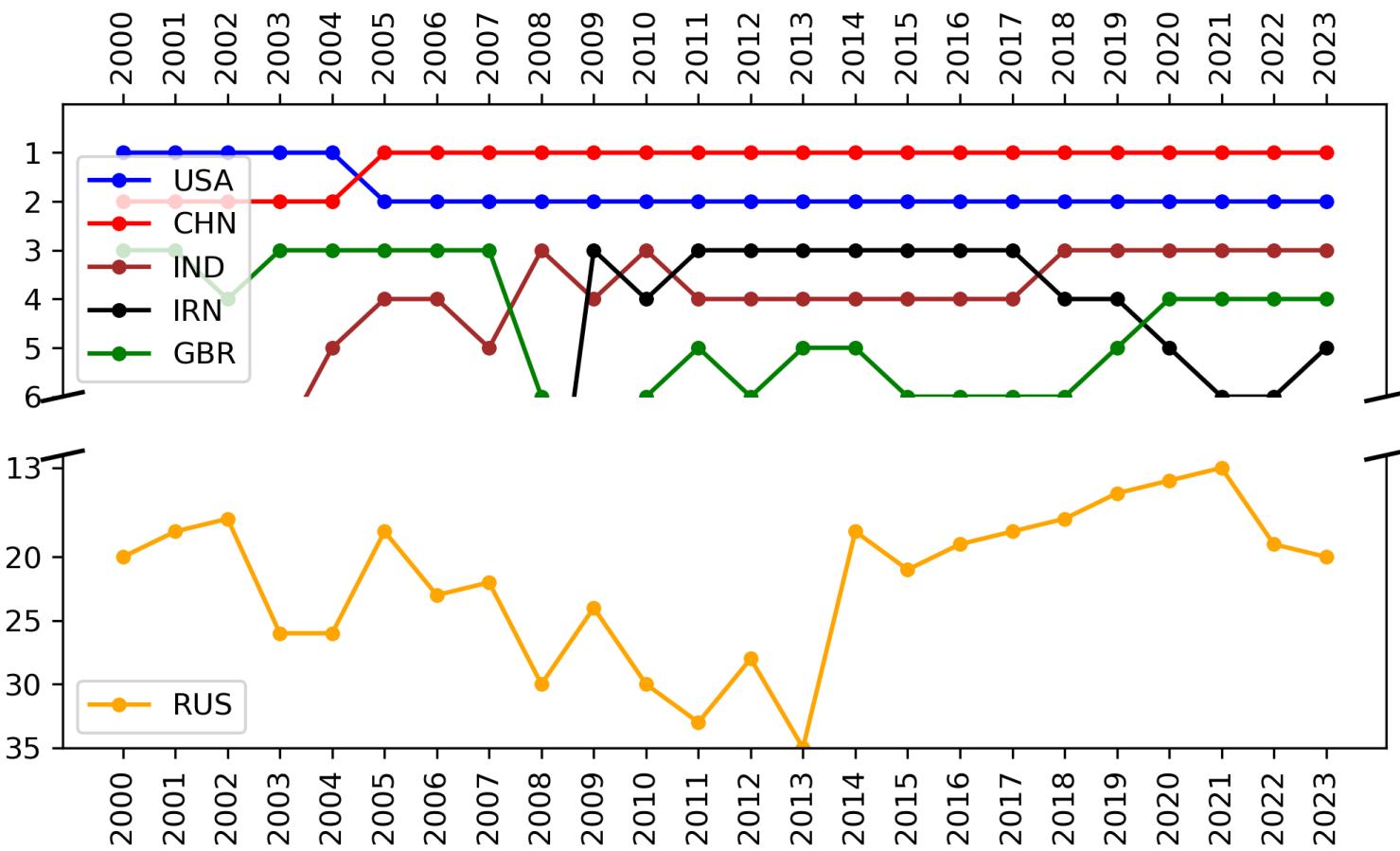


Documents, 2000-2023
AI&ML in EART&ENVI



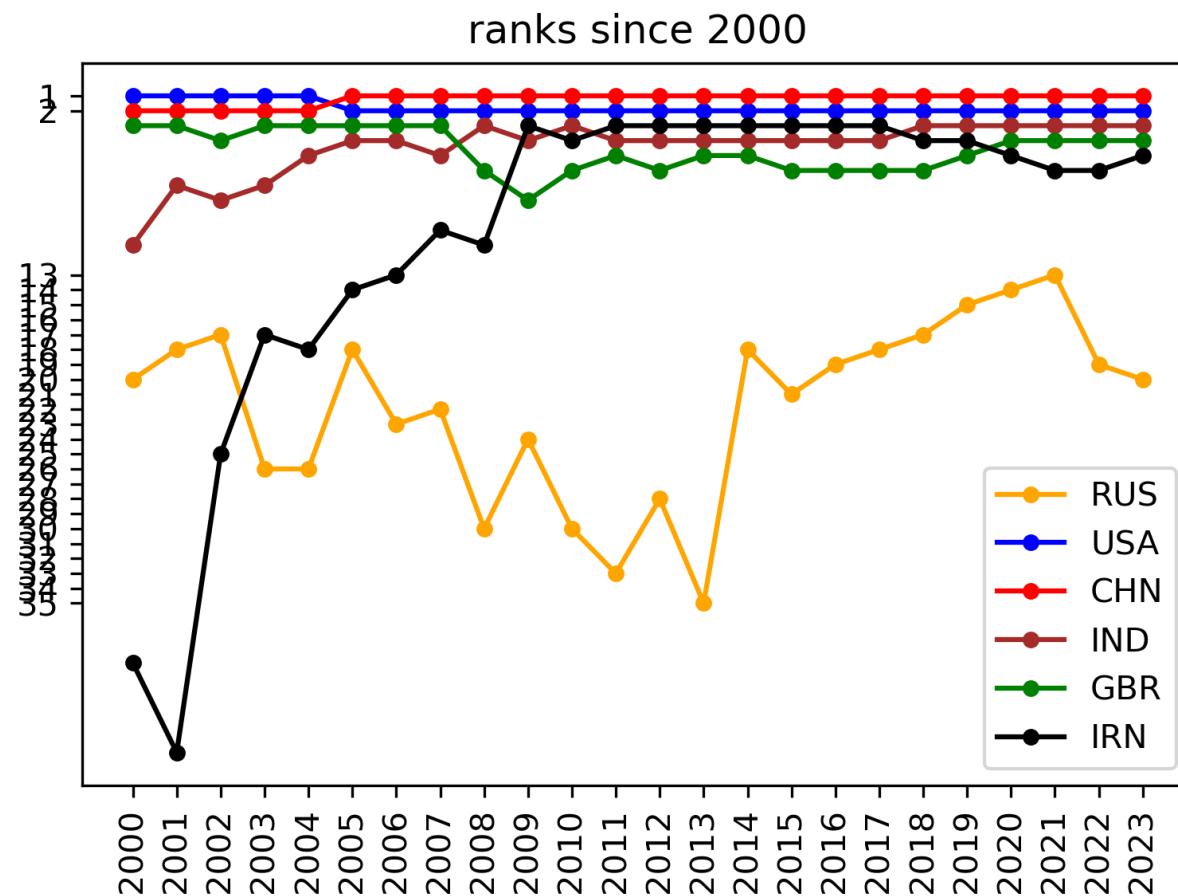
МО в науках о Земле

Статьи в рецензируемых журналах по тематике “Environmental sciences” и “Earth sciences” с применением методов машинного обучения. По данным Scopus.



МО в науках о Земле

Статьи в рецензируемых журналах по тематике “Environmental sciences” и “Earth sciences” с применением методов машинного обучения. По данным Scopus.



ОЧЕНЬ КРАТКОЕ
ВВЕДЕНИЕ
В МЕТОДОЛОГИЮ МАШИННОГО ОБУЧЕНИЯ

на контрасте с другими методологиями

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

ЧЕМ ЗАНИМАЮТСЯ ФИЗИКИ?

атмосферы

океана

взаимодействия океана и атмосферы

цунами

ядерщики, высоких энергий

био-

...

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

ЧЕМ ЗАНИМАЮТСЯ ФИЗИКИ?

ИЩУТ ИСТИНУ (об устройстве Вселенной) ?

ОТКРЫВАЮТ ЗАКОНЫ МИРОЗДАНИЯ ?

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

ЧЕМ ЗАНИМАЮТСЯ ФИЗИКИ?

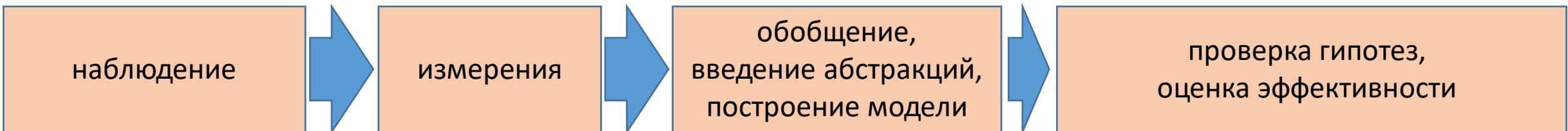
ПРЕДЛАГАЮТ МОДЕЛИ наблюдаемых процессов (**какие?..**)

ПРОВЕРЯЮТ - ОЦЕНИВАЮТ ЭФФЕКТИВНОСТЬ (**как?..**)

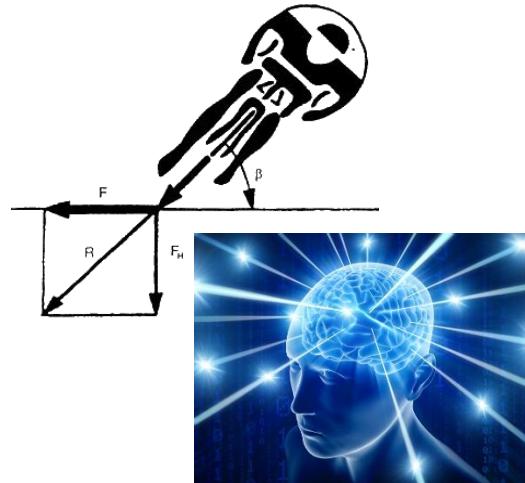
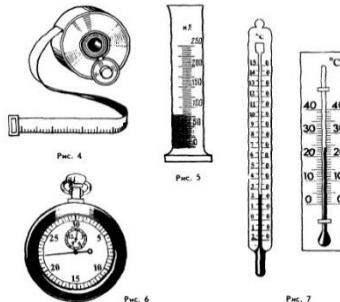
ПРИМЕНЯЮТ НА ПРАКТИКЕ

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

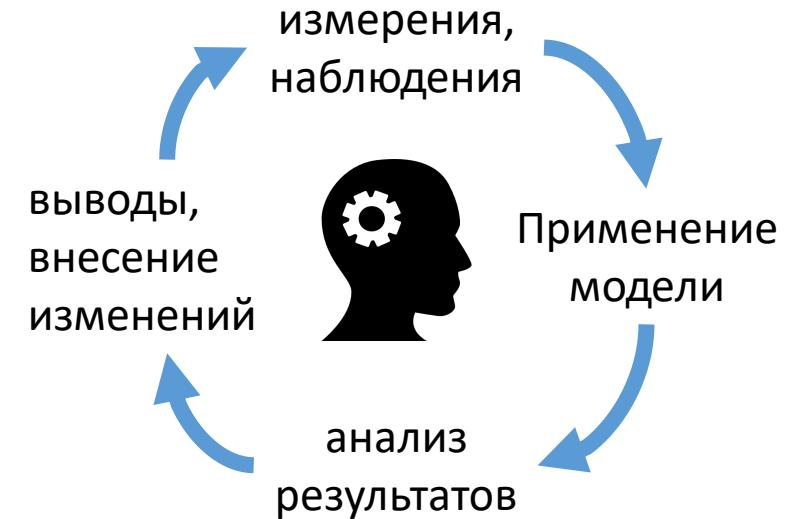
КАК проводятся физические исследования?



by Frits Ahlefeldt



Настоящая наука начинается с тех
пор, как начинают измерять.
Точная наука немыслима без меры.
Д.И. Менделеев



ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

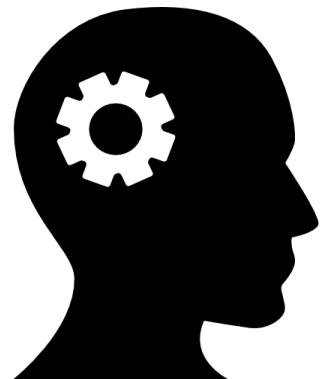
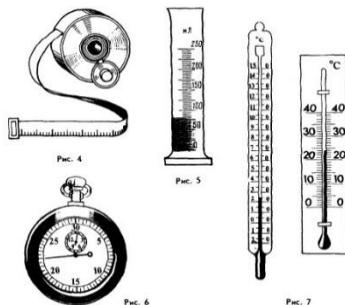
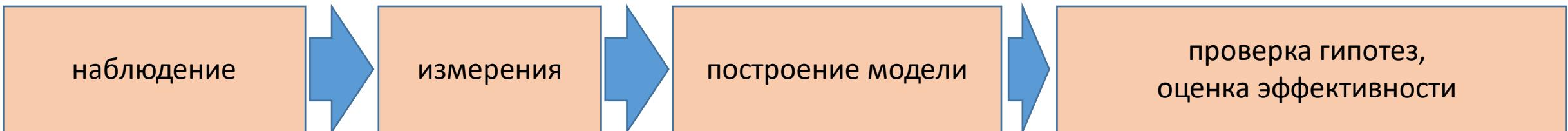
Когда (человеку) непонятно, что происходит
но делать-то что-то надо

наблюдение

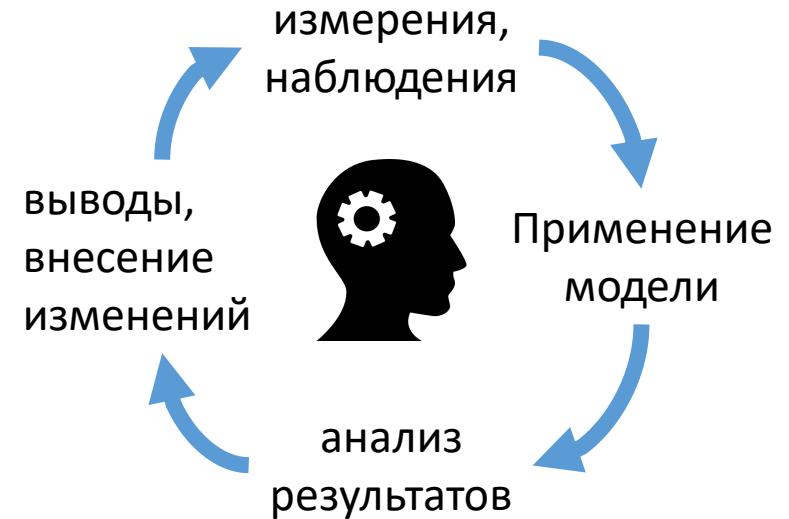


ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Когда (человеку) непонятно, что происходит
все равно строим модель



~~обобщение ?
введение абстракций ?~~



ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Когда (человеку) непонятно, что происходит
все равно строим модель

- Для чего? Какова цель?
- Что у нас для этого есть?
- Какого рода модель?
- Какая должна быть модель?

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Когда (человеку) непонятно, что происходит
все равно строим модель

- Для чего? Какова цель?
- Что у нас для этого есть?
- Какого рода модель?
- Какая должна быть модель?
- Оценить неизвестную(ые) величину(ы) $\{y_i\}$
-
-
-

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Когда (человеку) непонятно, что происходит
все равно строим модель

- Для чего? Какова цель?
- Что у нас для этого есть?
- Какого рода модель?
- Какая должна быть модель?
- Оценить неизвестную(ые) величину(ы) $\{y_i\}$
- Данные измерений $\{x_i\}$ (желательно релевантных задаче)

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Когда (человеку) непонятно, что происходит
все равно строим модель

- Для чего? Какова цель?
- Что у нас для этого есть?
- Какого рода модель?
- Какая должна быть модель?
- Оценить неизвестную(ые) величину(ы) $\{y_i\}$
- Данные измерений $\{x_i\}$
- $\mathcal{F}: \mathbb{X} \rightarrow \mathbb{Y}$

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Когда (человеку) непонятно, что происходит
все равно строим модель

- Для чего? Какова цель?
- Что у нас для этого есть?
- Какого рода модель?
- Какая должна быть модель?
- Оценить неизвестную(ые) величину(ы) $\{y_i\}$
- Данные измерений $\{x_i\}$
- $\mathcal{F}: \mathbb{X} \rightarrow \mathbb{Y}$
- Обобщающая. Достоверная (в каком смысле?)

Применимая.

Понятия в МО

- Объекты/события
- Признаковое описание объектов/событий \mathbf{x} – случайная величина
- Реализация признакового описания для i -го объекта/события x_i
- Целевая переменная y – случайная величина
- Реализация целевой переменной для i -го объекта/события y_i
- Множество возможных векторов признакового описания \mathbb{X}
- Множество возможных значений (исходов) целевой переменной \mathbb{Y}
- Отображение $\mathcal{F}: \mathbb{X} \rightarrow \mathbb{Y}$ – модель МО, иногда статистической или вероятностной природы

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Когда (человеку) непонятно, что происходит
все равно строим модель

КАК?

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Когда (человеку) непонятно, что происходит
все равно строим модель

КАК?

Методы машинного обучения

Искусственный интеллект

Теория Вапника-Червоненкиса

Статистическая теория восстановления
зависимостей по эмпирическим данным

Машинный интеллект

ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

строим модель для решения задачи

типы задач МО

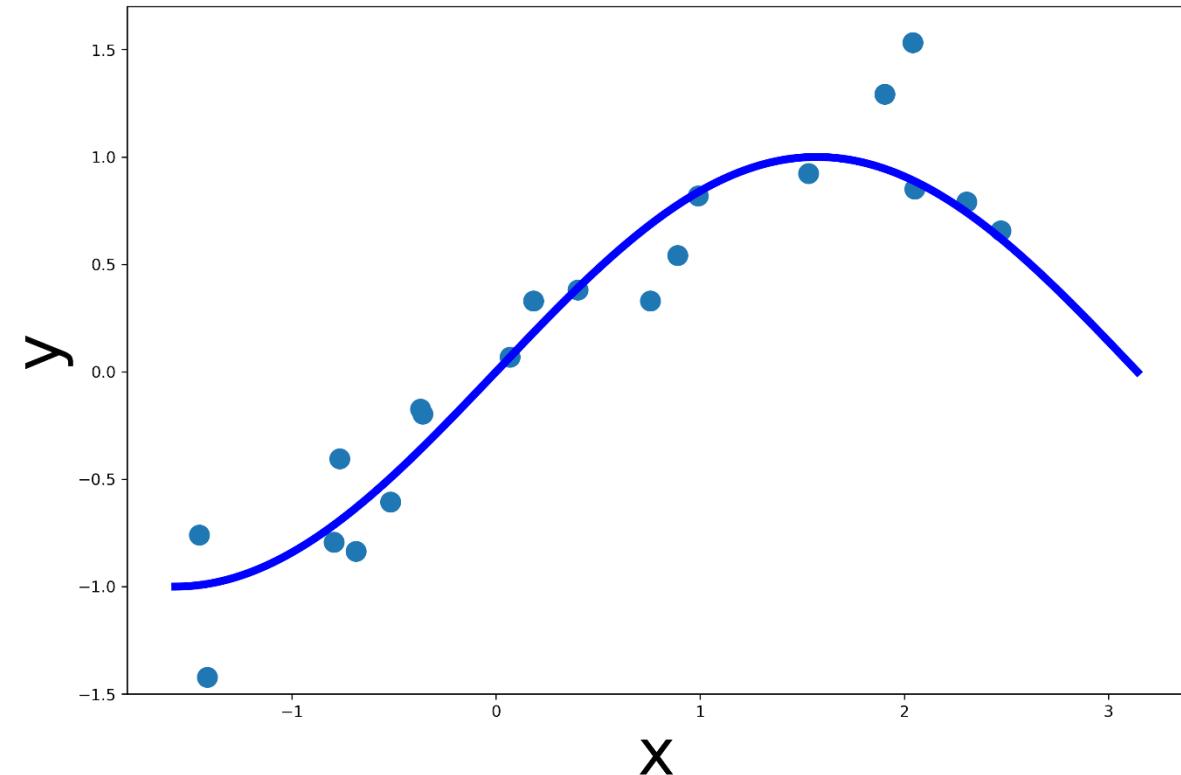
ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

строим модель для решения задачи

типы задач:

- «Обучение с учителем»
 - восстановление регрессии

что я хочу? – значение y



ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

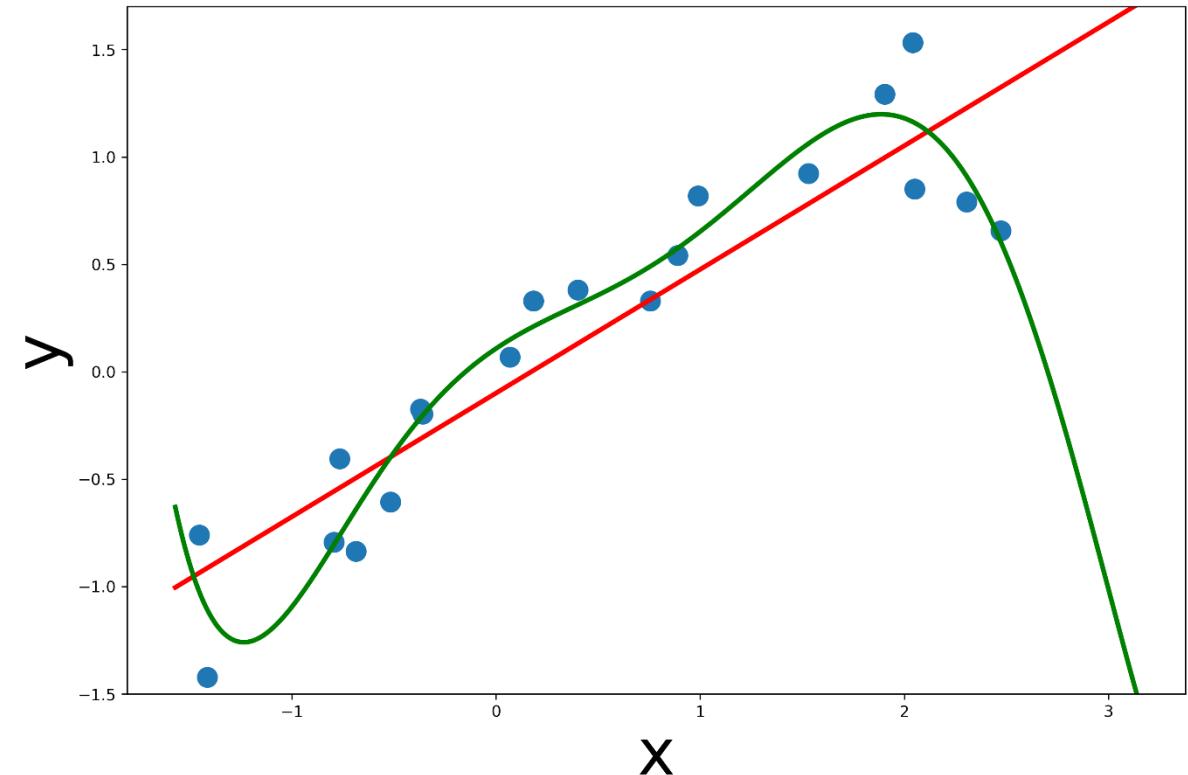
строим модель для решения задачи

типы задач:

- «Обучение с учителем»
 - восстановление регрессии

$$\hat{y} = ax + b$$

$$\hat{y} = p^{(6)}(x)$$



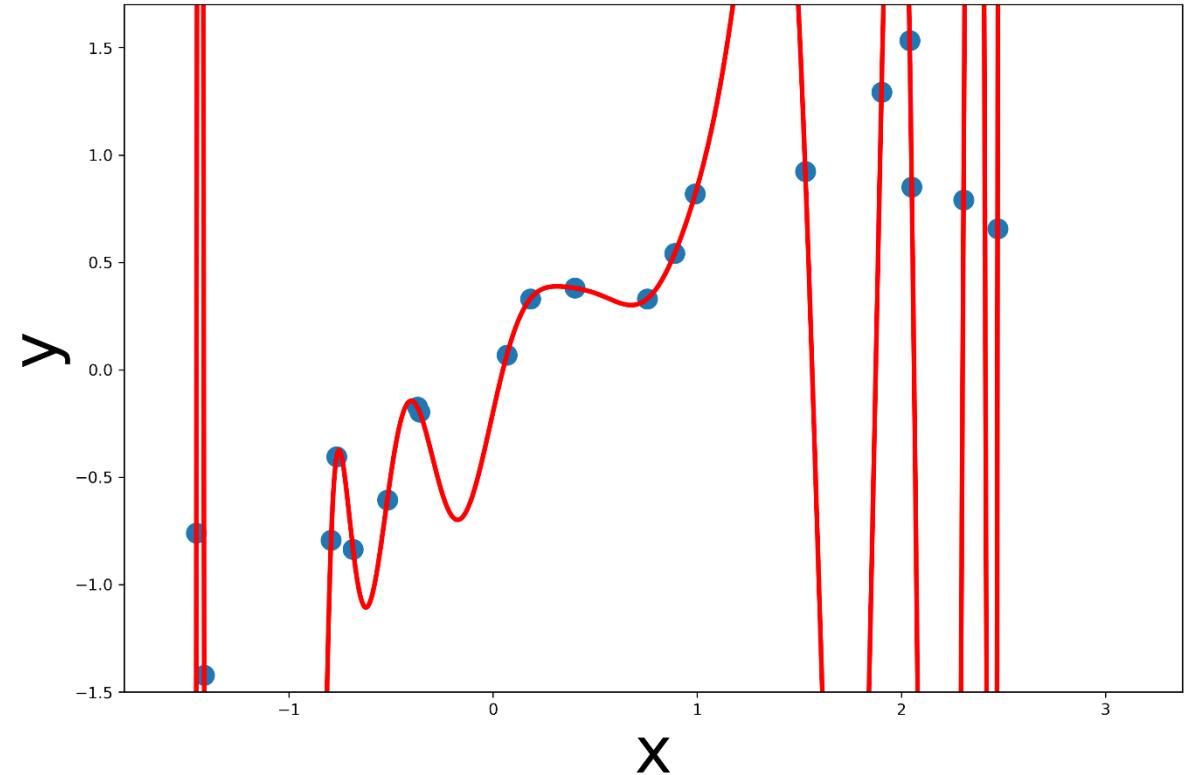
ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

строим модель для решения задачи

типы задач:

- «Обучение с учителем»
 - восстановление регрессии

$$\hat{y} = p^{(20)}(x)$$



«переобучение»

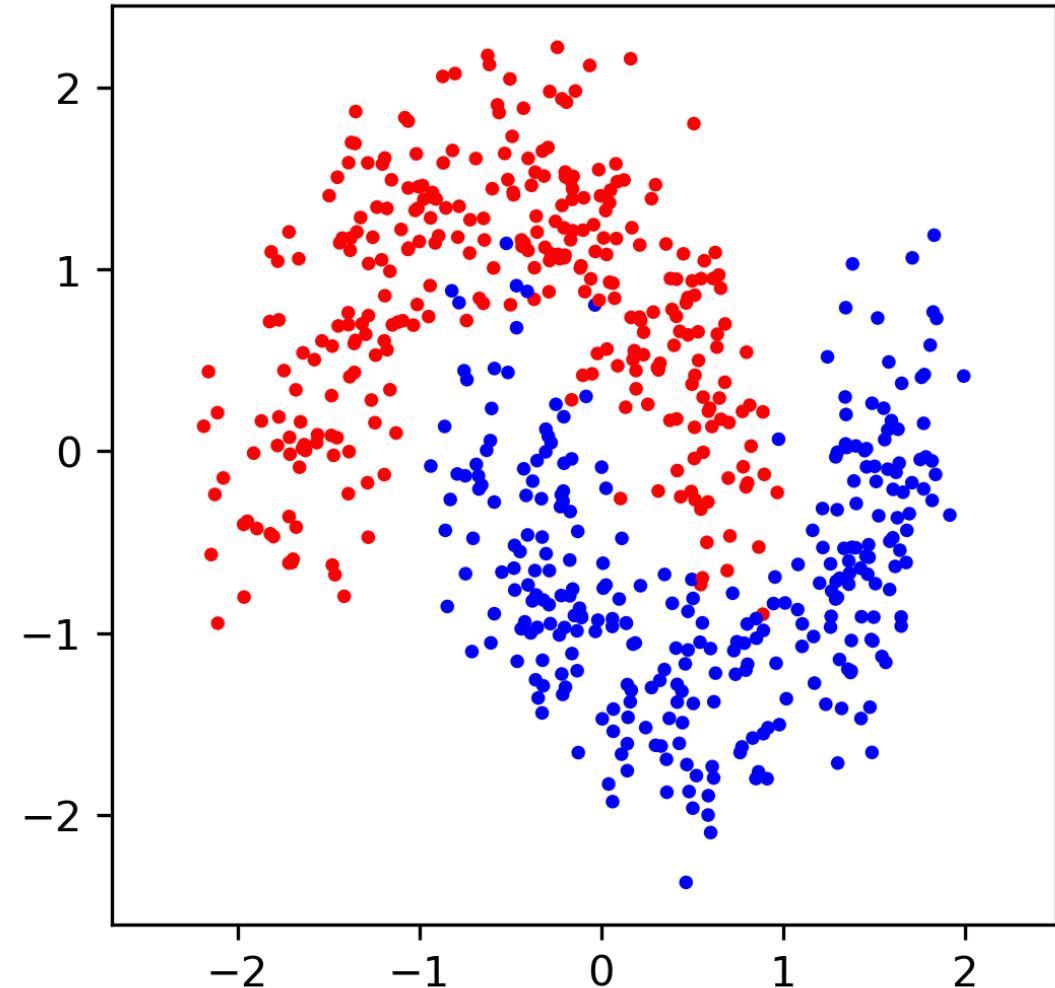
ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

строим модель для решения задачи

типы задач:

- «Обучение с учителем»
 - классификация

что я хочу? – метку класса
(красный или синий?)



ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

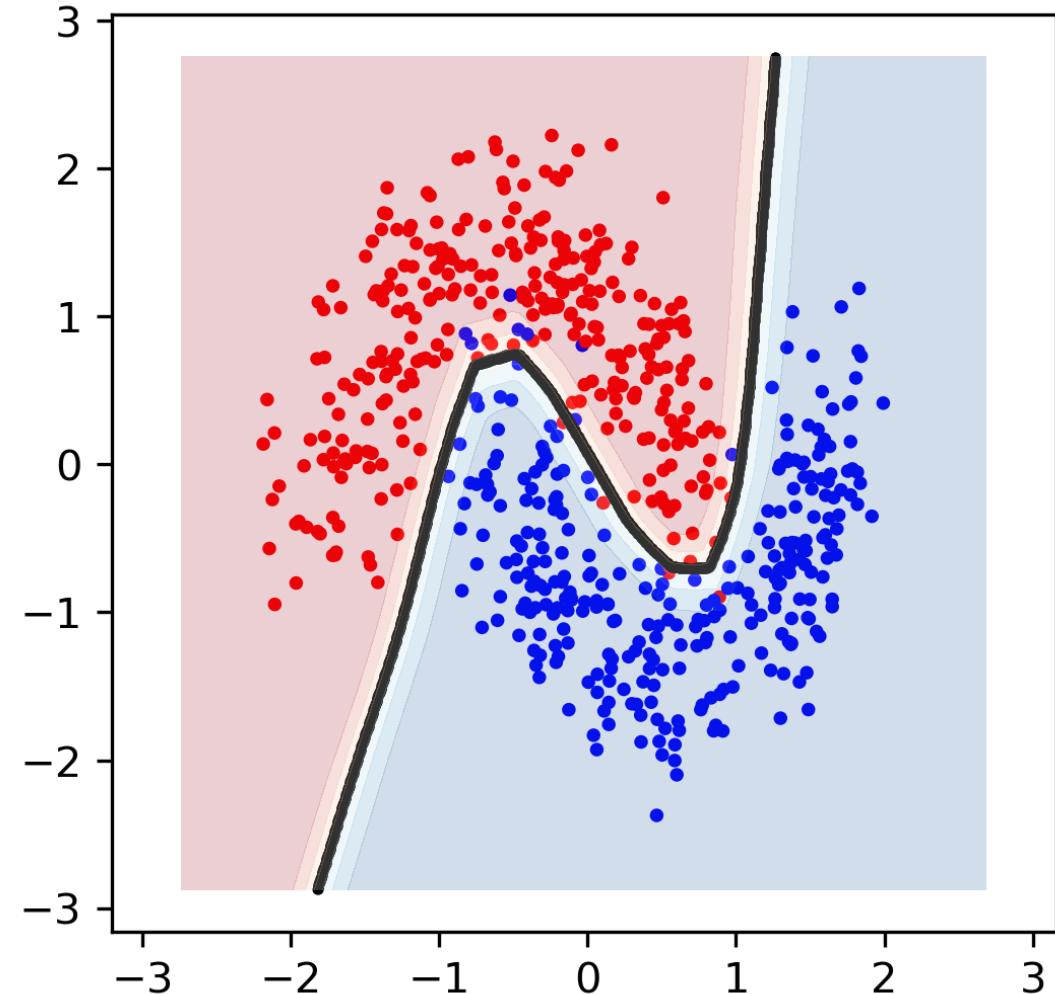
строим модель для решения задачи

типы задач:

- «Обучение с учителем»
 - классификация

$$\hat{p}(\text{красный}) = f(x)$$

$$\hat{p}(\text{синий}) = 1 - f(x)$$

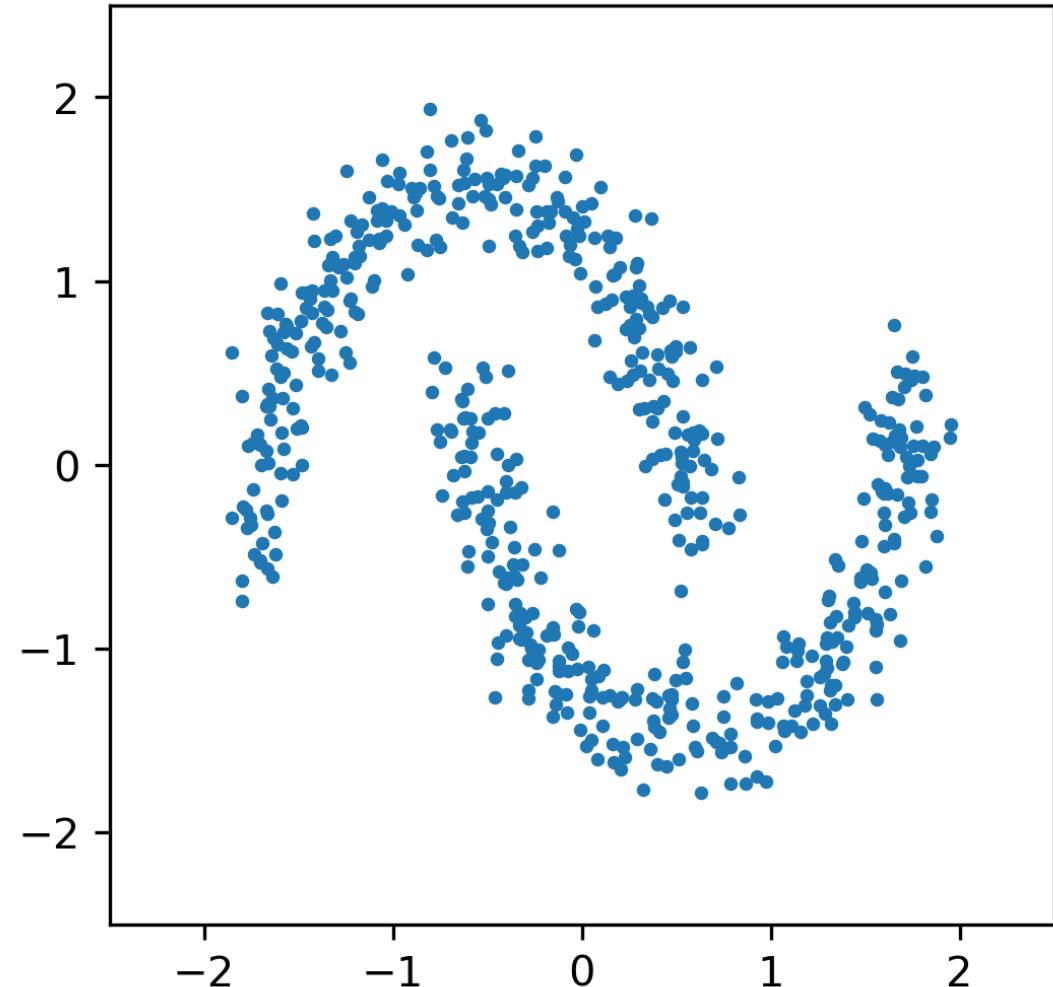


ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

строим модель для решения задачи

типы задач:

- «Обучение без учителя»
 - поиск структуры в данных
- что я хочу?**
- метки групп
 - знать, есть ли группы?
 - сколько групп?



ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

строим модель для решения задачи

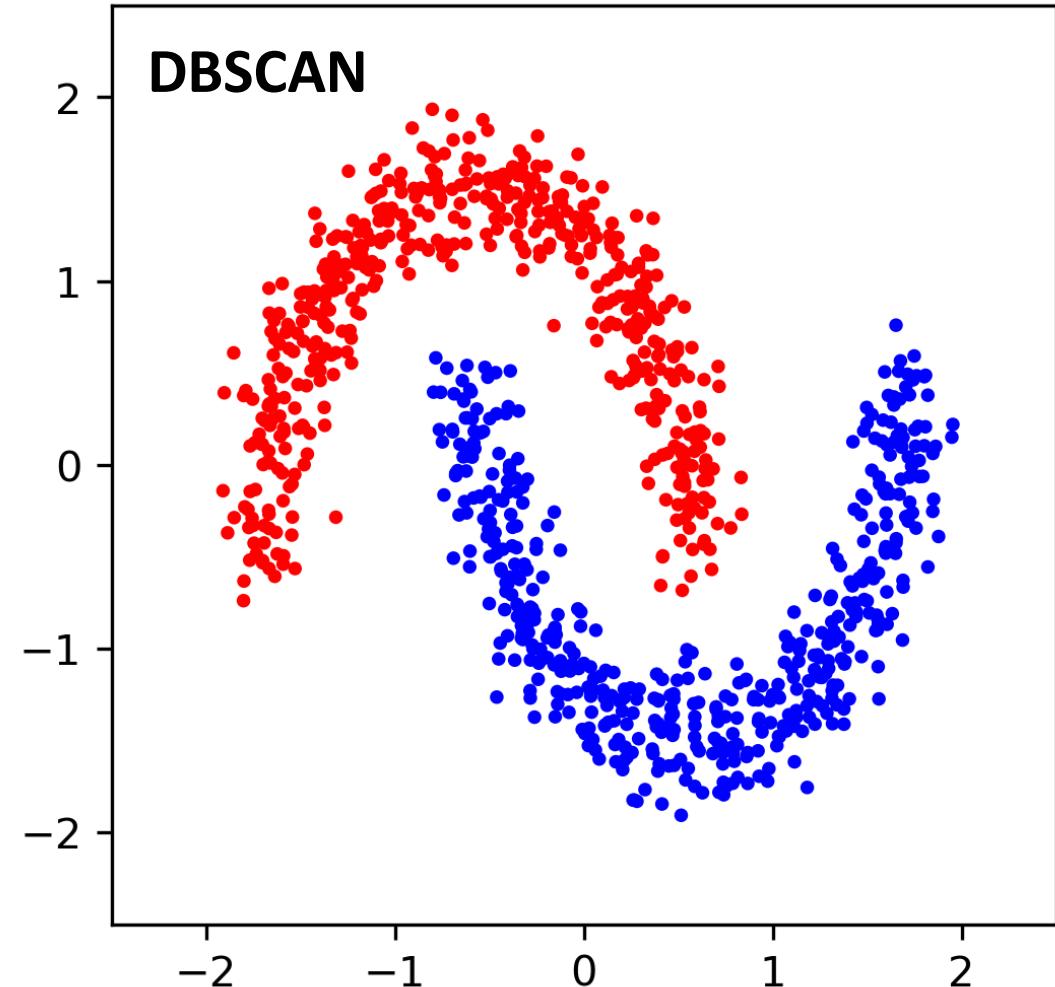
типы задач:

○ «Обучение без учителя»

- кластеризация

ЧТО Я ХОЧУ?

- метки групп
- знать, есть ли группы?
- сколько групп?



ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

строим модель для решения задачи

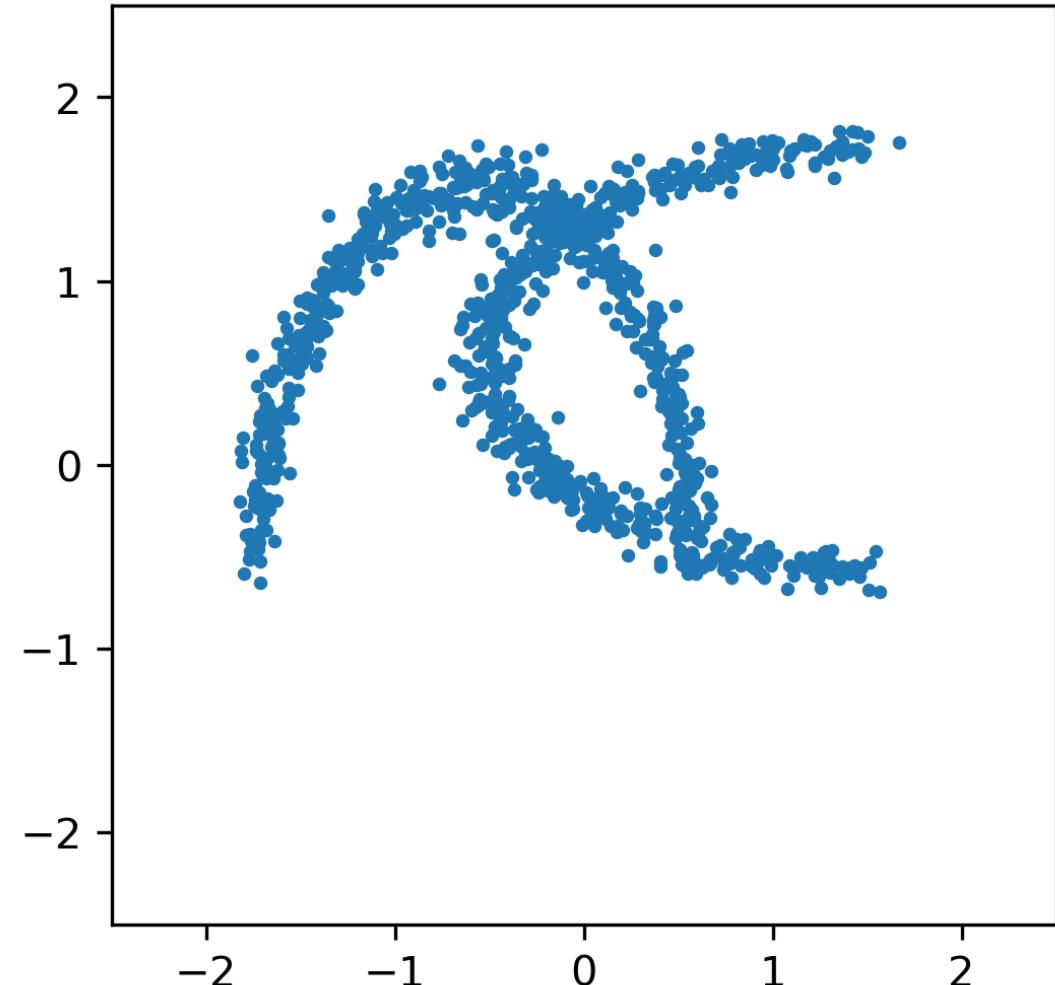
типы задач:

- «Обучение без учителя»
 - кластеризация

Всегда ли есть решение?

хоть какое-нибудь

ДА



ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

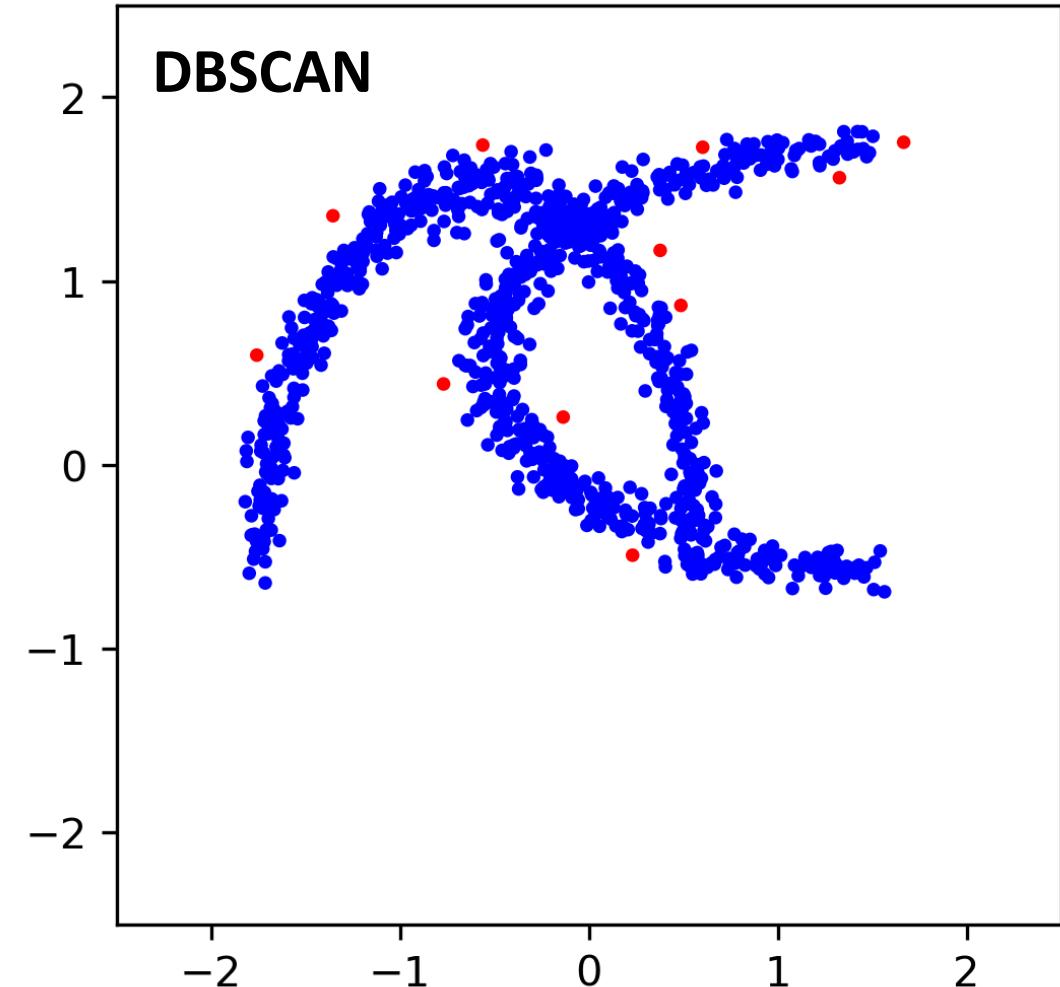
строим модель для решения задачи

типы задач:

○ «Обучение без учителя»

- кластеризация

Всегда ли есть решение,
которое мне понравится?



ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

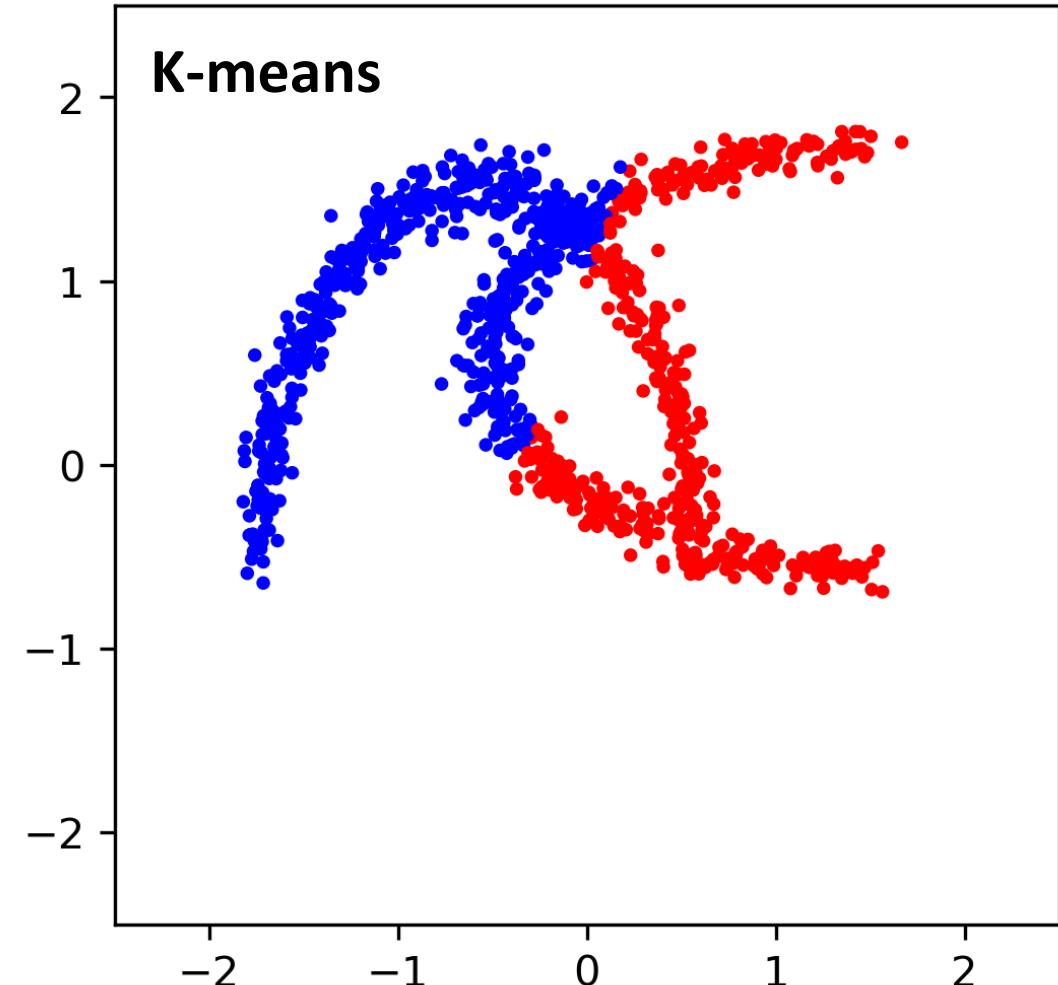
строим модель для решения задачи

типы задач:

○ «Обучение без учителя»

- кластеризация

Всегда ли есть решение,
которое мне понравится?



ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

строим модель для решения задачи

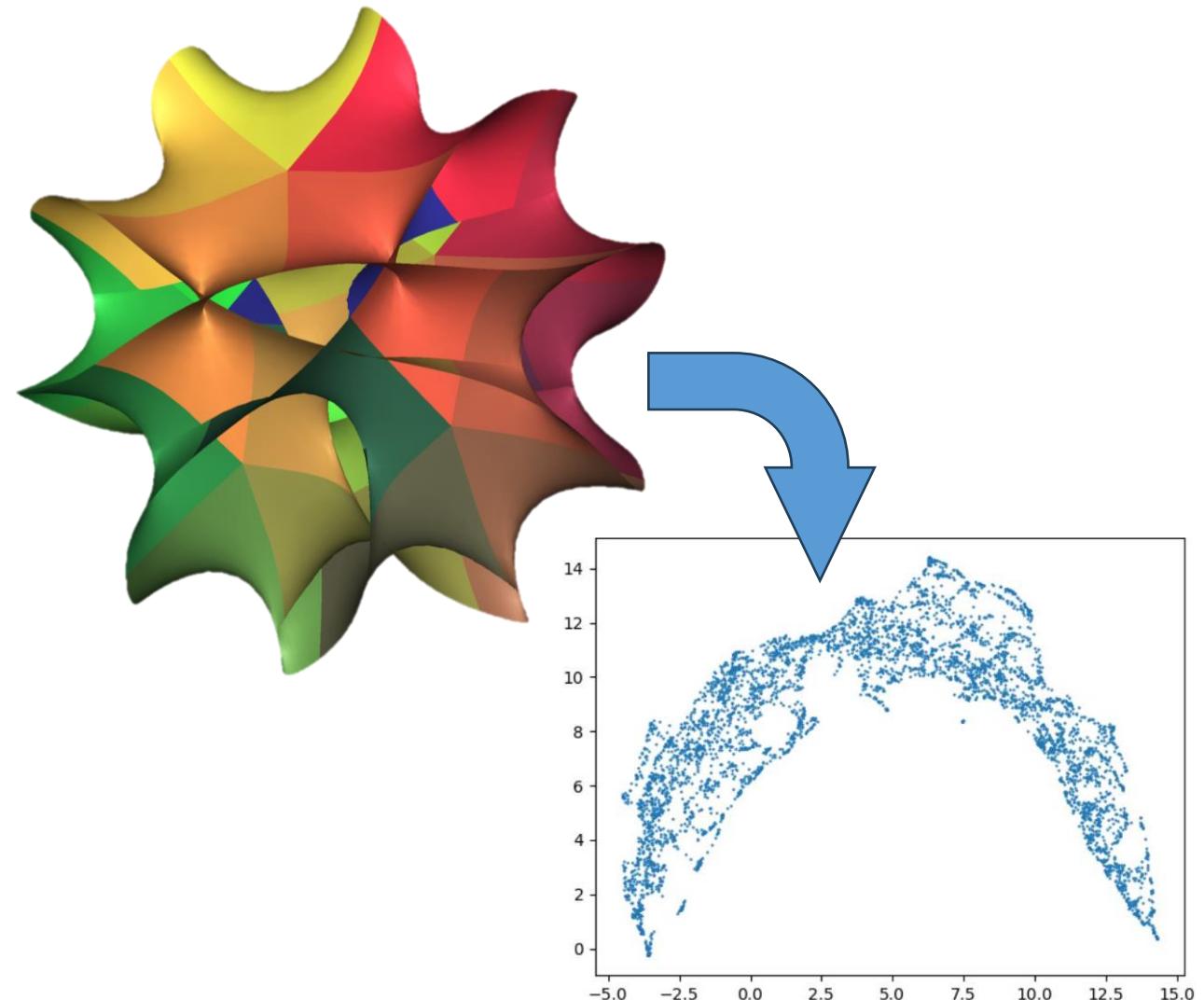
типы задач:

○ «Обучение без учителя»

- снижение размерности

что я хочу?

признаковое описание
сниженной размерности



ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

строим модель для решения задачи

типы задач:

○ «Обучение без учителя»

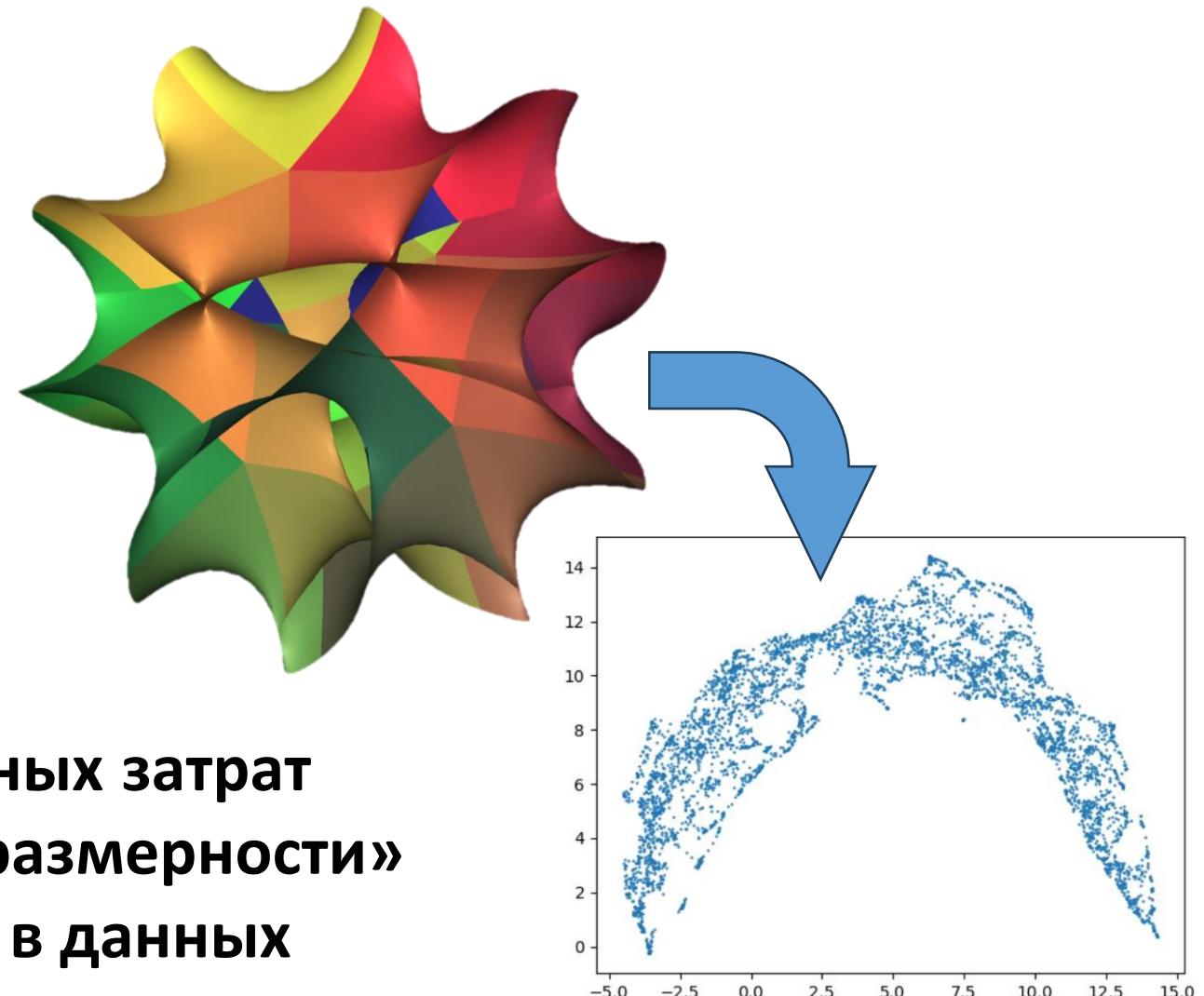
- снижение размерности

что я хочу?

признаковое описание
сниженной размерности

зачем?

- **визуализация данных**
- **снижение вычислительных затрат**
- **борьба с «проклятием размерности»**
- **снижение уровня шума в данных**



ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

строим модель для решения задачи

типы задач:

- «Обучение с учителем»
 - восстановление регрессии
 - классификация
- «Обучение без учителя»
 - Кластеризация
 - Снижение размерности
 - Апроксимация распределения данных
- Другие задачи: смежные, редкие, специальные.
 - самоконтролируемое обучение
 - с частичным привлечением учителя
 - обучение с подкреплением
 - выучивание меры различия (дистанции)
 - ...

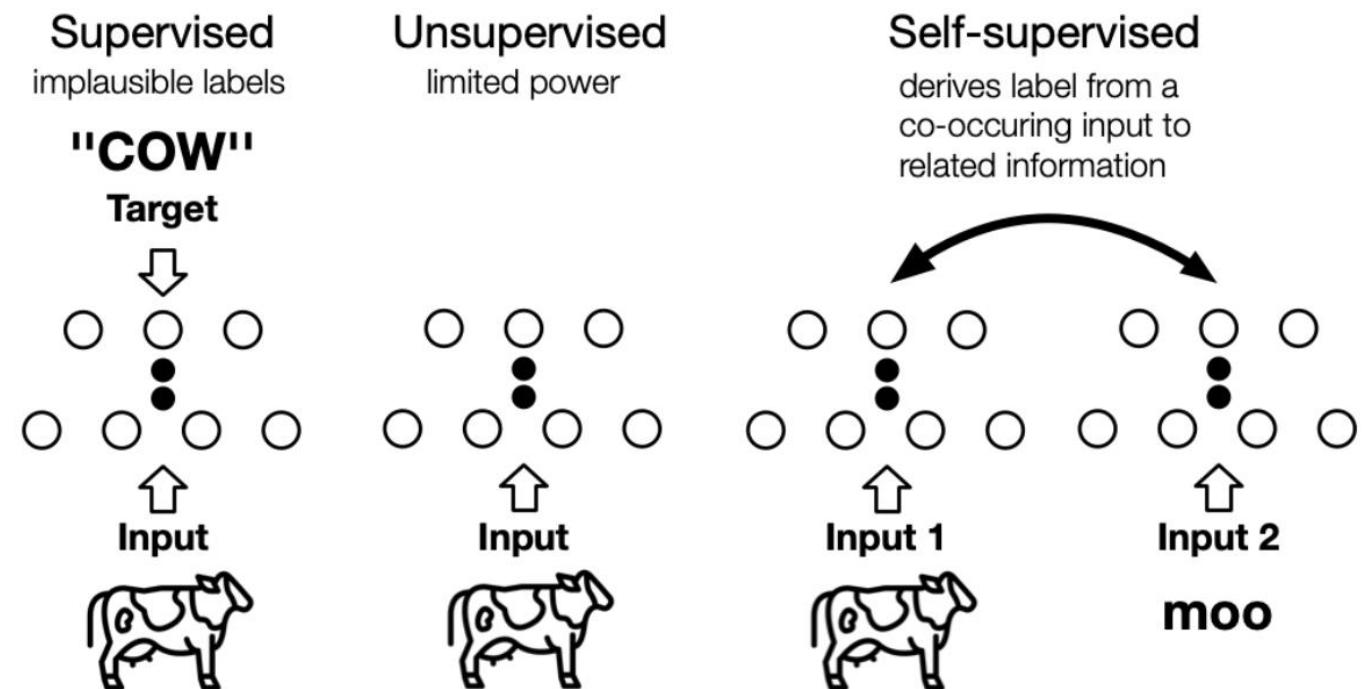
ОЧЕНЬ КРАТКОЕ ВВЕДЕНИЕ В МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

типы задач:

- «Обучение без учителя»
 - самоконтролируемое обучение

ЧТО Я ХОЧУ?

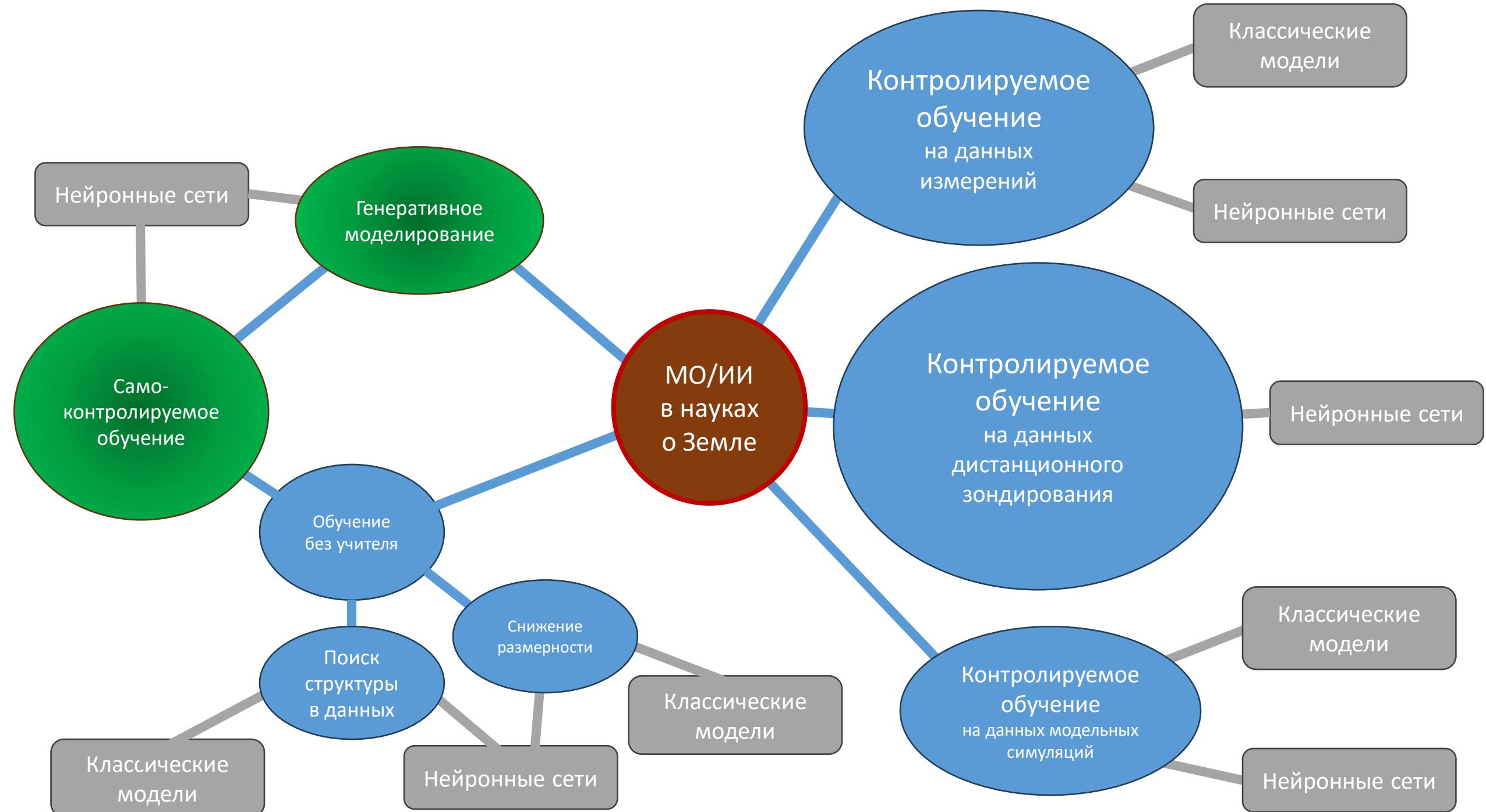
модель, которая «знает,
какие бывают объекты в
коллекции данных»
для применения ее в
далнейшем в широком
круге задач





Машинное обучение в задачах наук о Земле

М.А. Криницкий



Поисковые исследования



Прикладные исследования



Анализ результатов моделирования

Прикладные

- Статистический даунскейлинг
- Краткосрочный и сверхкраткосрочный нейросетевой прогноз (погоды, погоды в океане, ледовой обстановки, рядов натурных измерений)

Поисковые

(Поиск структуры в данных)

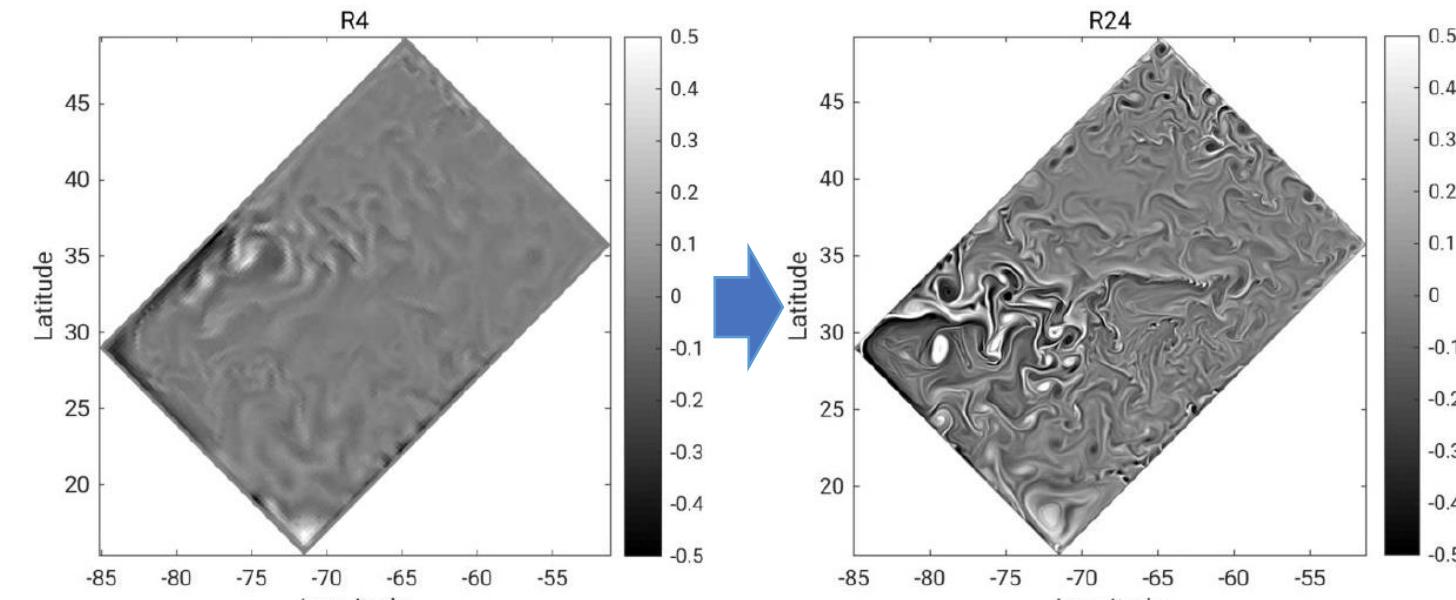
- Выявление экстремальных и аномальных событий, объектов
- Оценка тенденций их возникновения

(Генеративные модели)

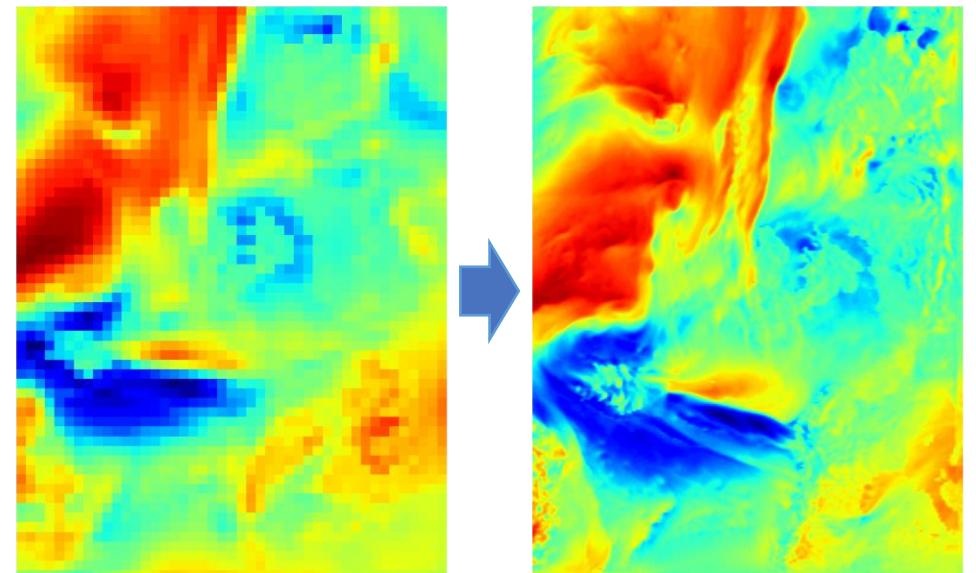
- Нейросетевая оценка качества гидродинамического моделирования

Анализ результатов моделирования

- Статистический даунскейлинг



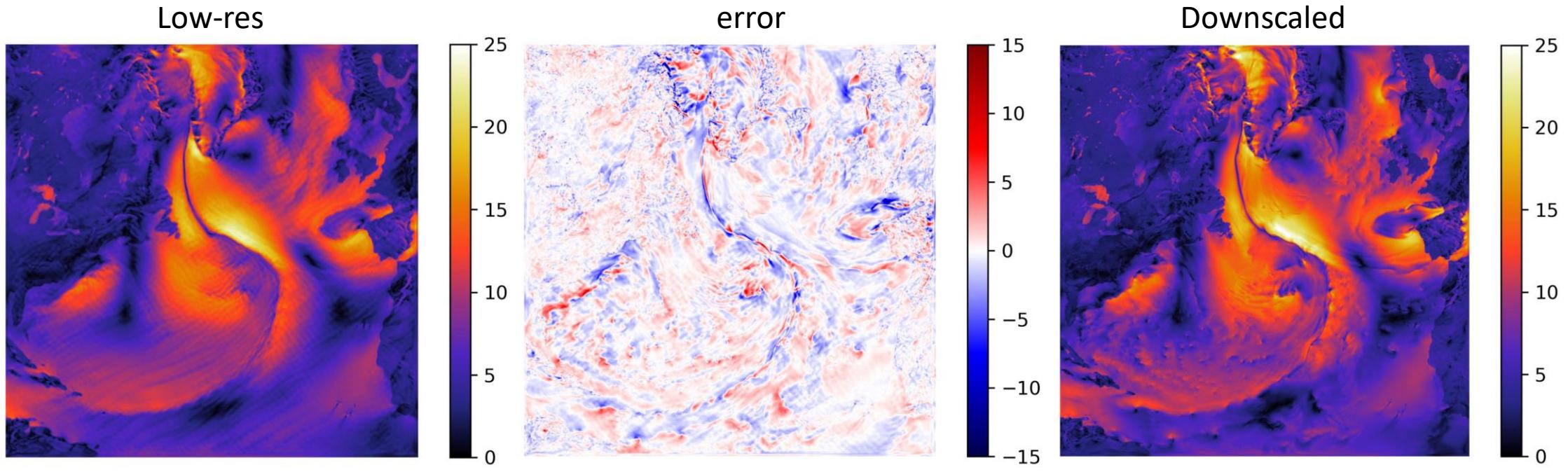
Масштабирование динамики течений



Масштабирование скорости ветра

Анализ результатов моделирования

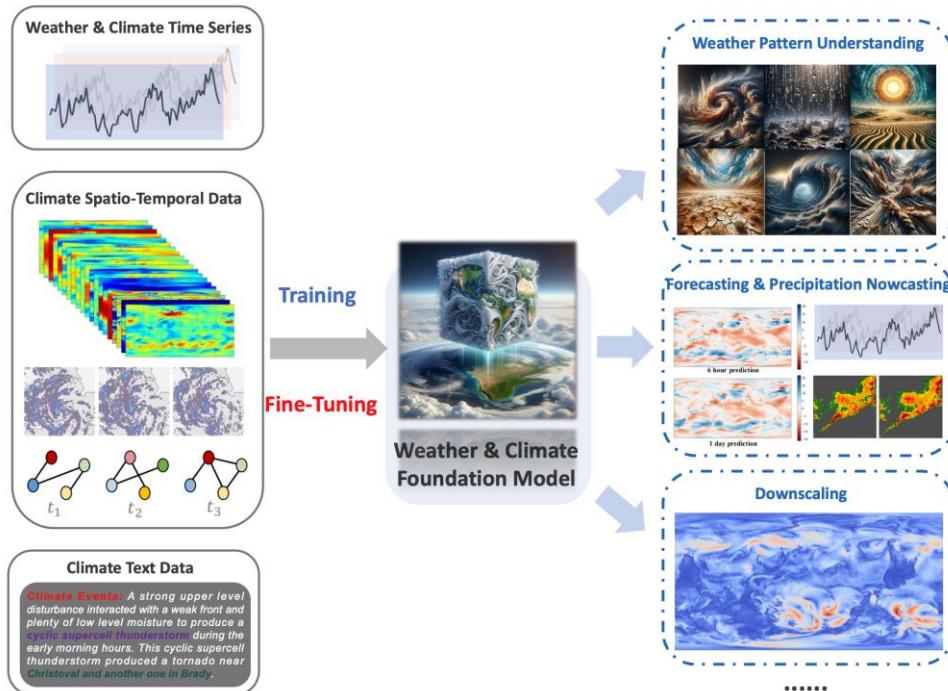
- Статистический даунскейлинг



Масштабирование скорости ветра

Анализ результатов моделирования

- Статистический даунскейлинг с использованием фундаментальных нейросетевых моделей



1. Microsoft ClimaX^{1,*} – Jan'2023, UCLA (USA) 1.40625°
2. FengWu^{2,*} – Apr'2023, China (6 организаций) 0.25°
3. PanGu^{3,*} – July'2023, Huawei (China) 0.25°
4. FuXi^{4,*} – Jun'2023, Fudan University (China) 0.25°
5. FourCastNet^{5,*} – Feb'2022, NVIDIA 0.25°
6. GraphCast^{6,*} – Nov'2023, Google 0.25°
7. W-MAE^{7,*} – Apr'2023, UEST (China) 0.25°

Поисковые исследования

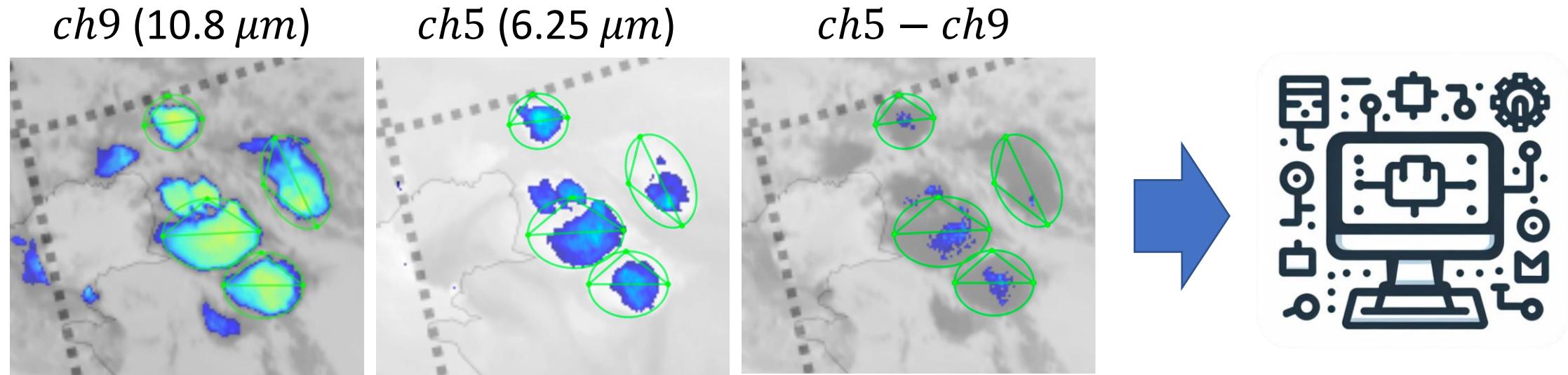


Прикладные исследования



Контролируемое обучение: пример

Идентификация мезомасштабных конвективных систем

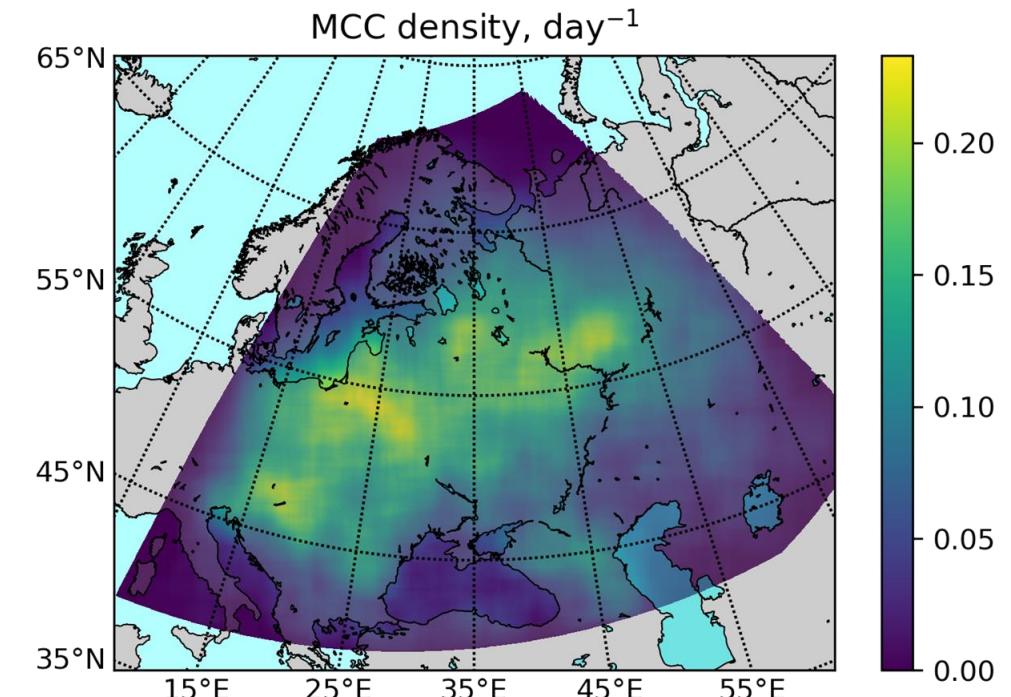
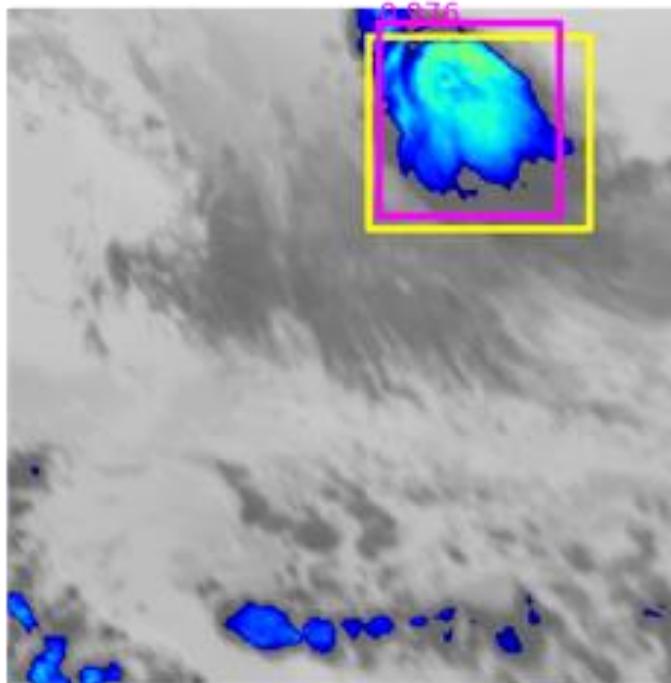


Данные ДЗЗ (Meteosat, MSG4),
Европейская территория России

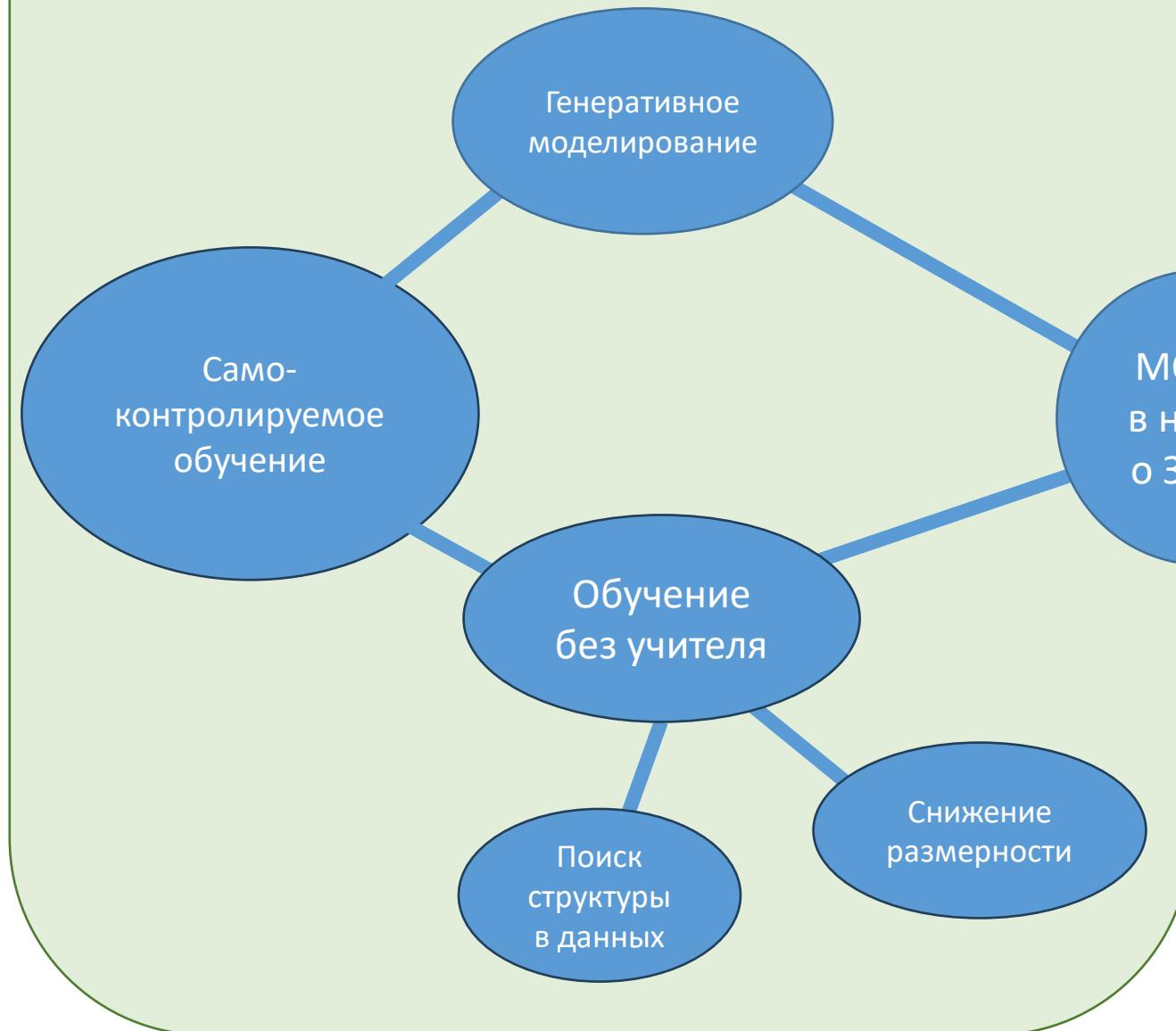
Контролируемое обучение: пример

Идентификация мезомасштабных конвективных систем

ch9 (10.8 μm)



Поисковые исследования



Прикладные исследования



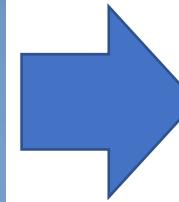
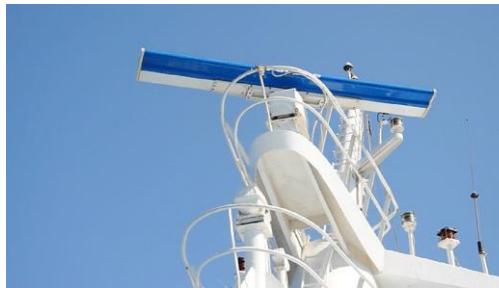
Анализ и интерпретация данных наблюдений

Прикладные

- Определение характеристик ветрового волнения по данным судового радара
 - Определение характеристик облачности по данным визуальной съемки
 - Прогноз характеристик возвратной миграции нерки
-
- Определение реактоспособности субстрата в активных центрах гидролаз для определения эффективности активации различных систем: комплексов бактериальной металло-β-лактамазы NDM-1 и L1 с антибиотиком имипенема, а также комплексов капралактама и капралактона с липазой CALB

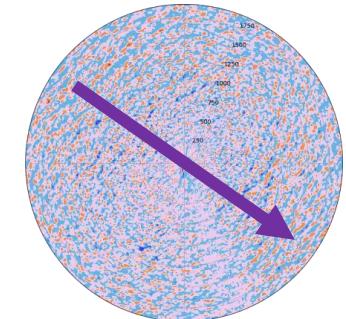
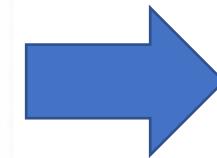
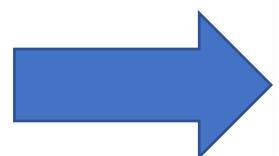
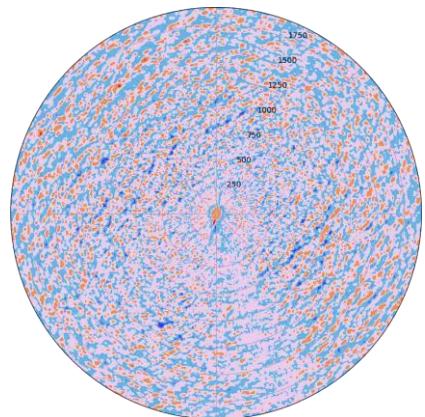
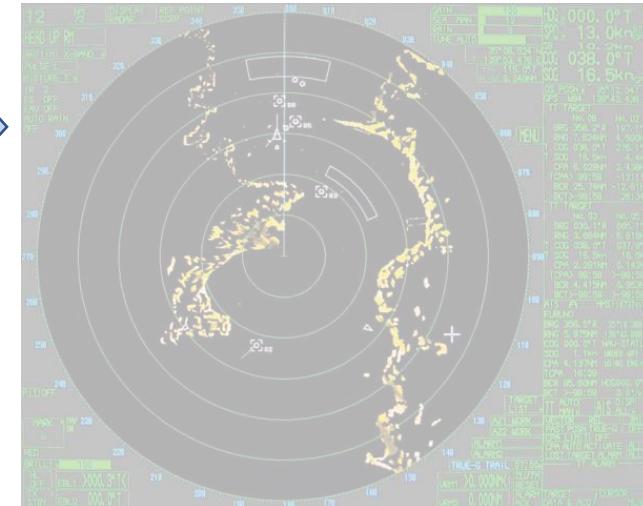
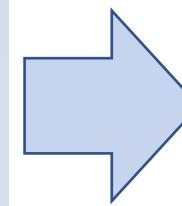
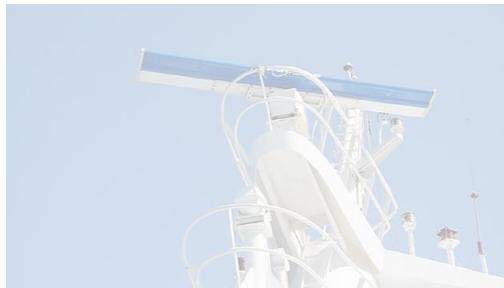
Анализ и интерпретация данных наблюдений

Характеристики ветрового волнения по данным навигационного радара



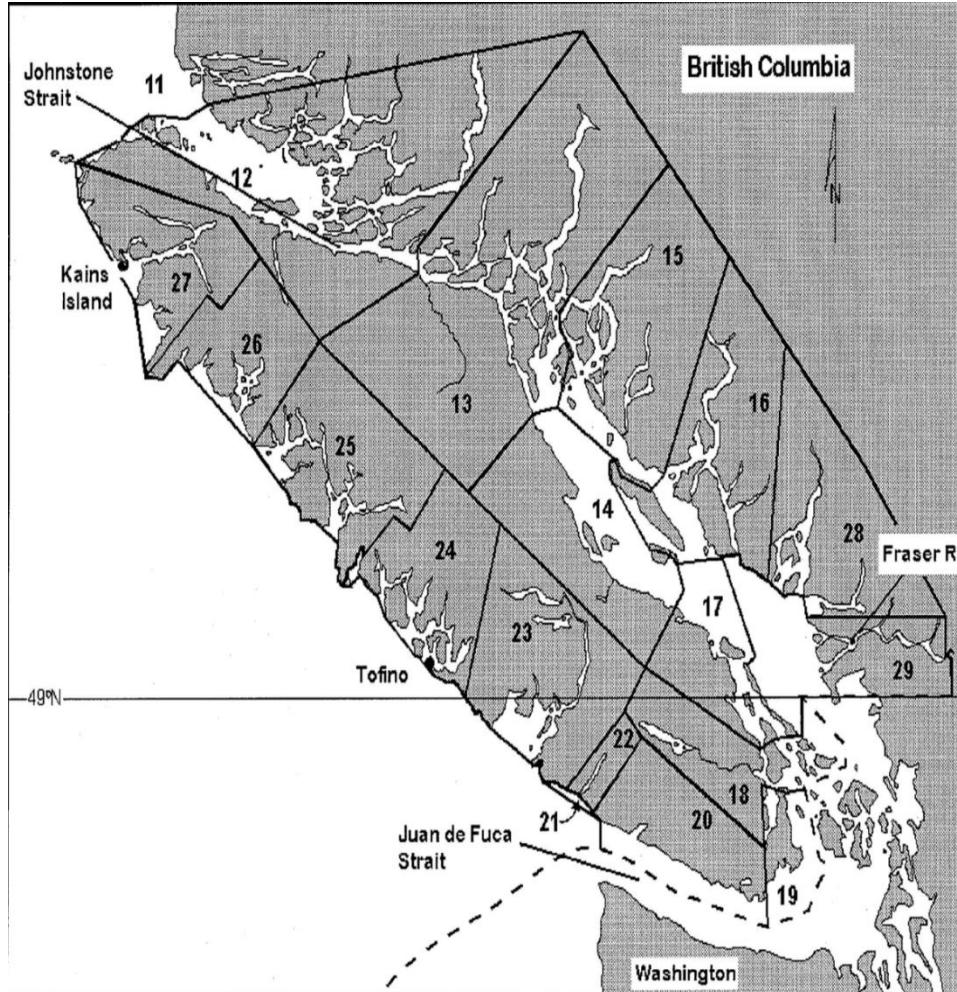
Анализ и интерпретация данных наблюдений

Характеристики ветрового волнения по данным навигационного радара

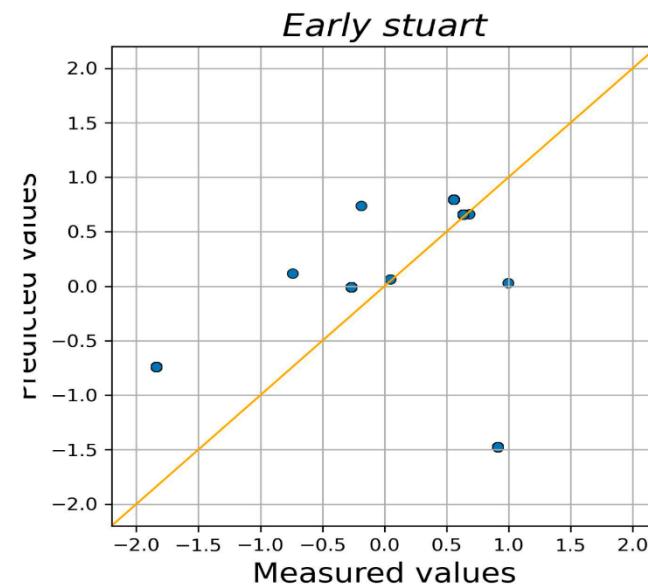


Прогноз рядов измерений

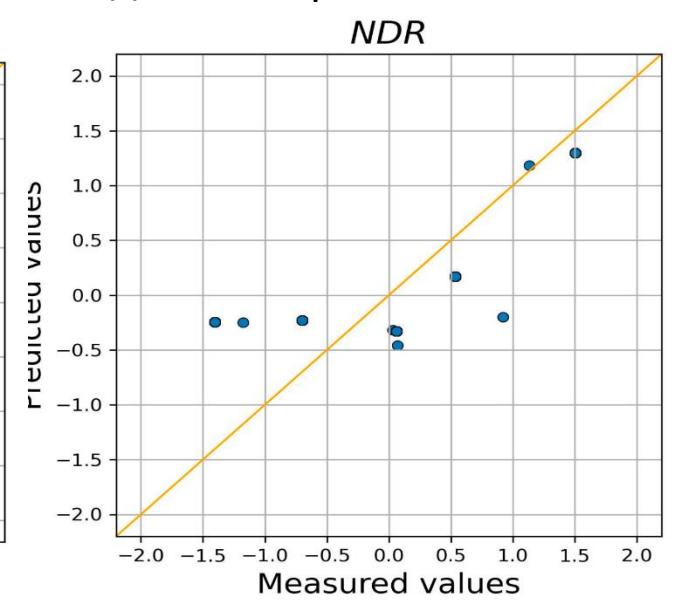
Прогноз характеристик возвратной миграции нерки в устье р. Фрейзер



дата медианного возврата



доля северного отклонения



Перспективные типы задач: прикладные

- Анализ данных дистанционного зондирования (спутникового базирования, судового, БПЛА) - решение прикладных задач в постановке контролируемого обучения
- Анализ данных полевых измерений
 - заполнение пропусков во временных рядах, в пространственно распределенных данных
 - восстановление климатических рядов по косвенным измерениям
- Статистическое масштабирование геофизических полей
- Статистический прогноз (краткосрочный: погоды, ледовой обстановки, рядов измерений)

Перспективные типы задач: поисковые

- **Выявление паттернов** (динамических, пространственных) в климатических и погодных данных
- **Выявление аномалий** в климатических, погодных данных
- **Предварительное самоконтролируемое обучение** нейросетевых моделей для последующего решения широкого круга прикладных задач
- (Обучение?) исследование свойств **фундаментальных климатических, погодных моделей**
- **Усвоение данных** в моделировании атмосферы, океана
- **Выучивание** нейросетевой **меры качества** воспроизведения динамики климата, динамики атмосферы, океана
- **Идентификация ДУЧП** климатических моделей с использованием нейросетей
- **Внедрение физических ограничений** в нейросетевое моделирование климатических процессов в атмосфере, океане

Неконтролируемое обучение

Выявление структуры состояний стратосферного полярного вихря

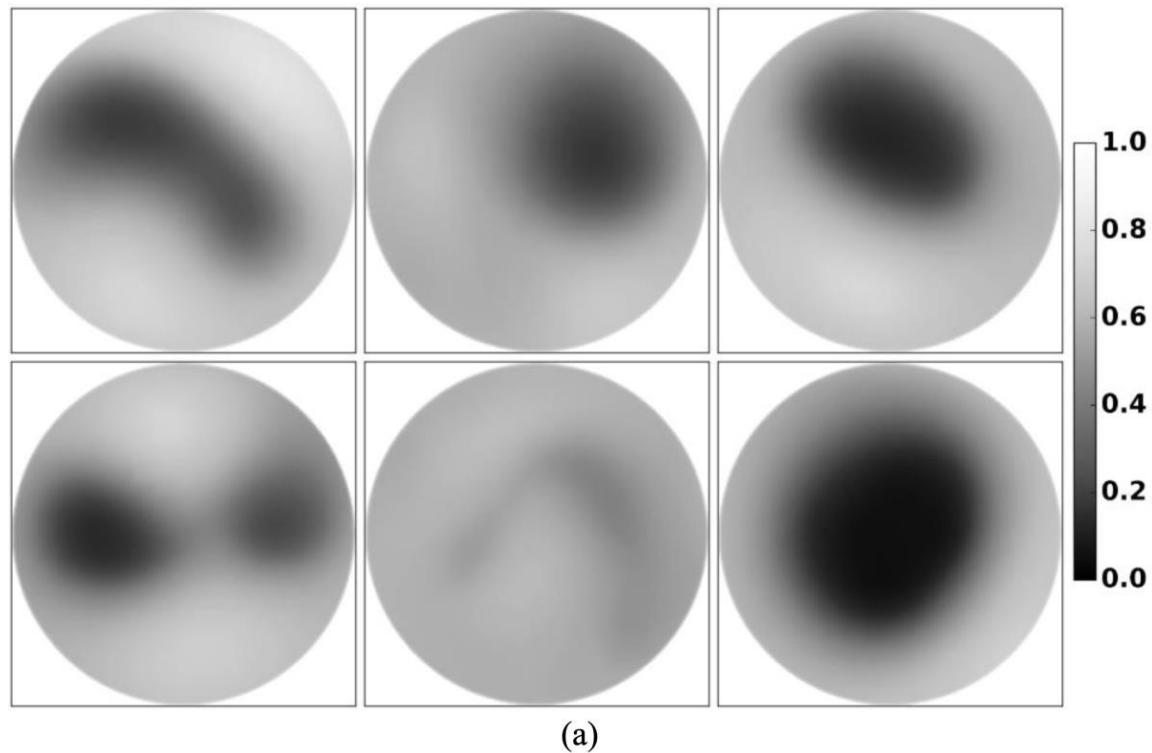
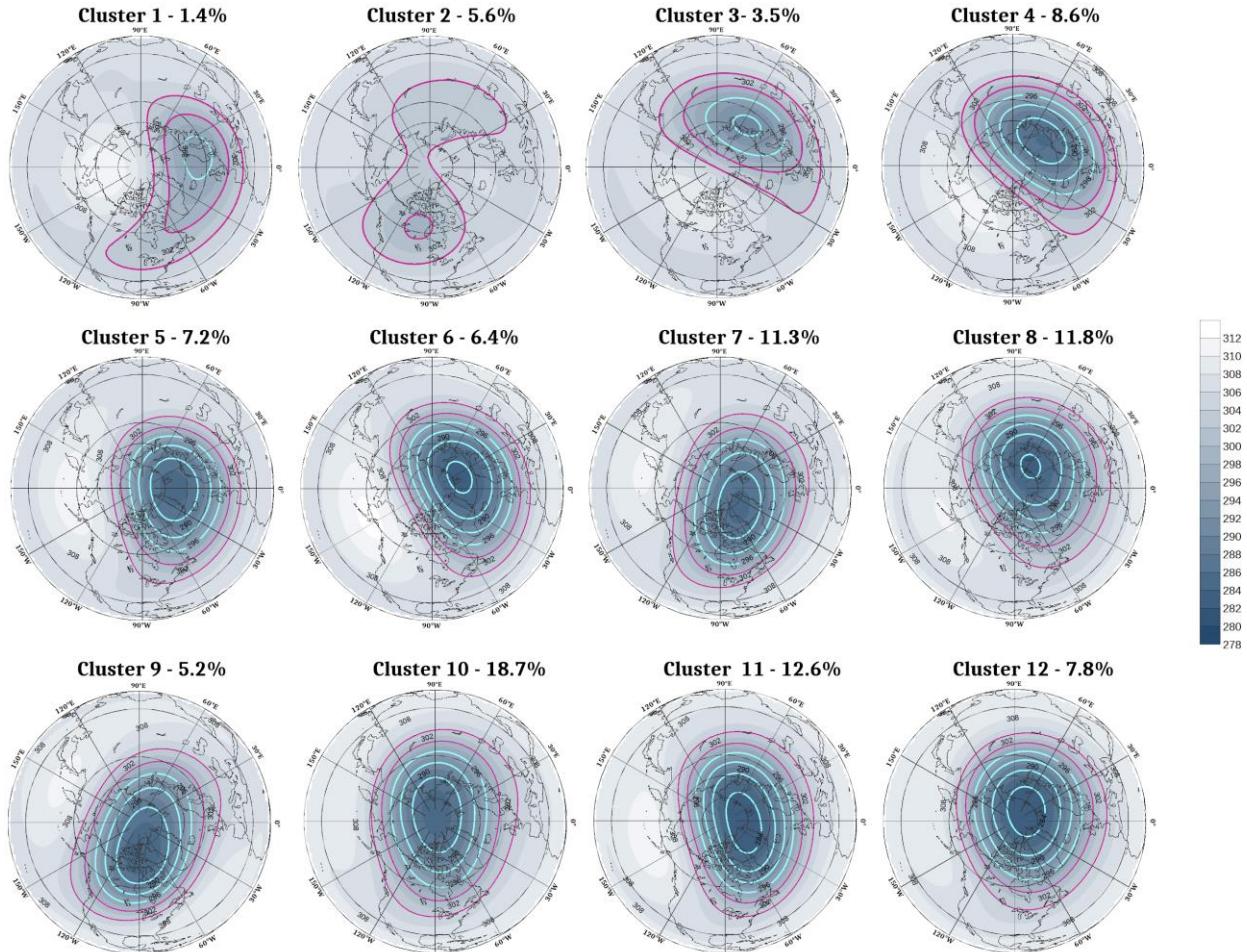


Figure 1: (a) Examples from the dataset of PV states (HGT values only, normalized)

Неконтролируемое обучение

Выявление структуры состояний стратосферного полярного вихря



Самоконтролируемое обучение

Большие базовые модели для климатических и погодных данных

ClimaX¹ FengWu² PanGu³ FuXi⁴ FourCastNet⁵ GraphCast⁶ W-MAE⁷

¹ Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*.

² Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J. J. & Ouyang, W. (2023). Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*.

³ Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970), 533-538.

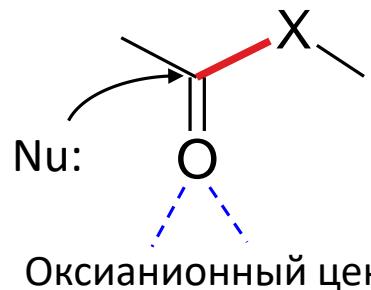
⁴ Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., & Li, H. (2023). FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1), 190.

⁵ Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., ... & Anandkumar, A. (2022). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.

⁶ Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, 382(6677), 1416-1421.

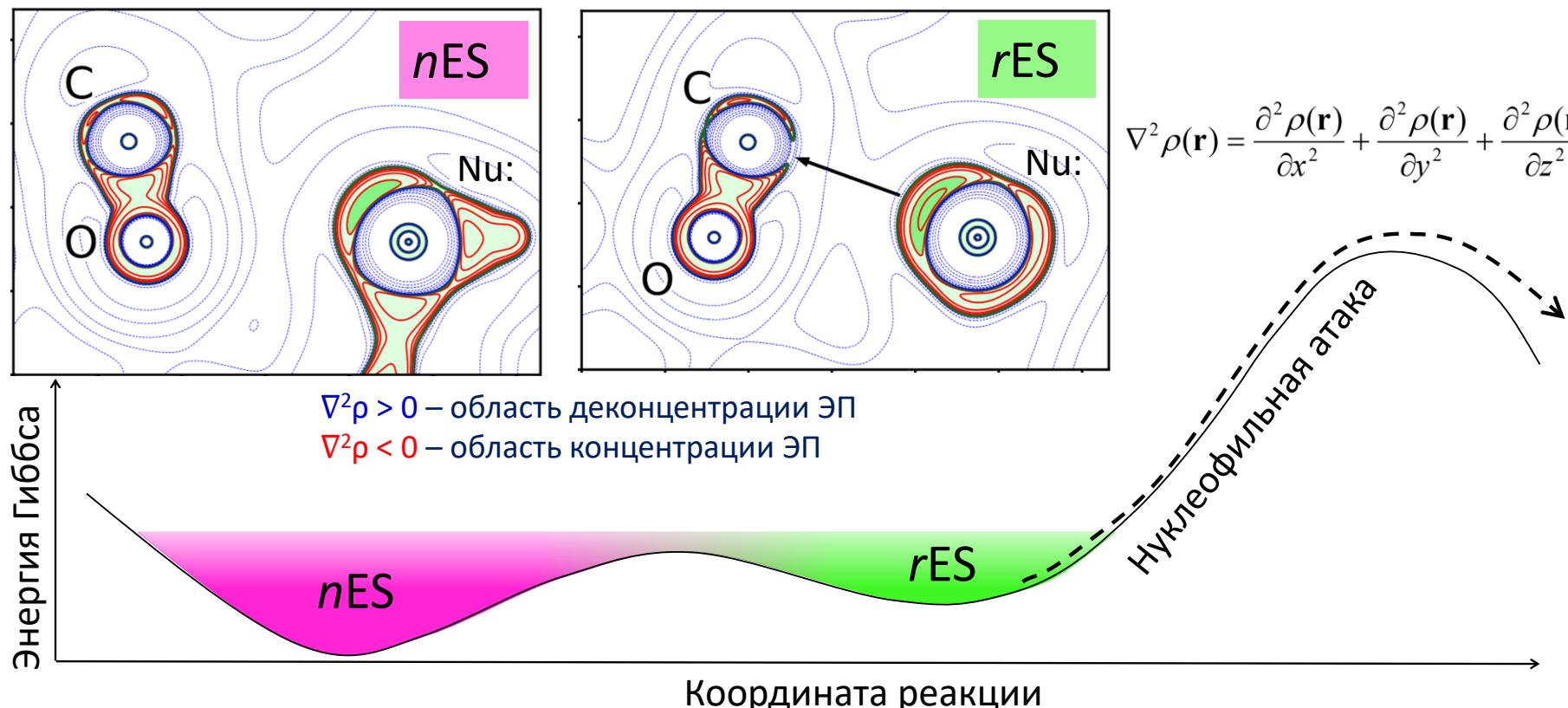
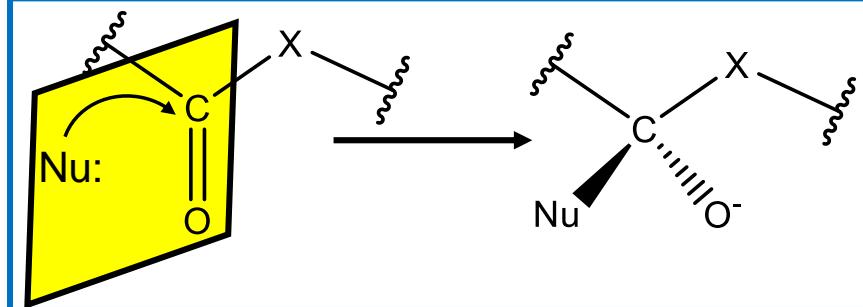
⁷ Man, X., Zhang, C., Feng, J., Li, C., & Shao, J. (2023). W-mae: Pre-trained weather model with masked autoencoder for multi-variable weather forecasting. *arXiv preprint arXiv:2304.08754*.

Активация субстрата



Нуклеофил:

- H₂O
- OH⁻
- OH (Ser, Thr)
- SH (Cys)

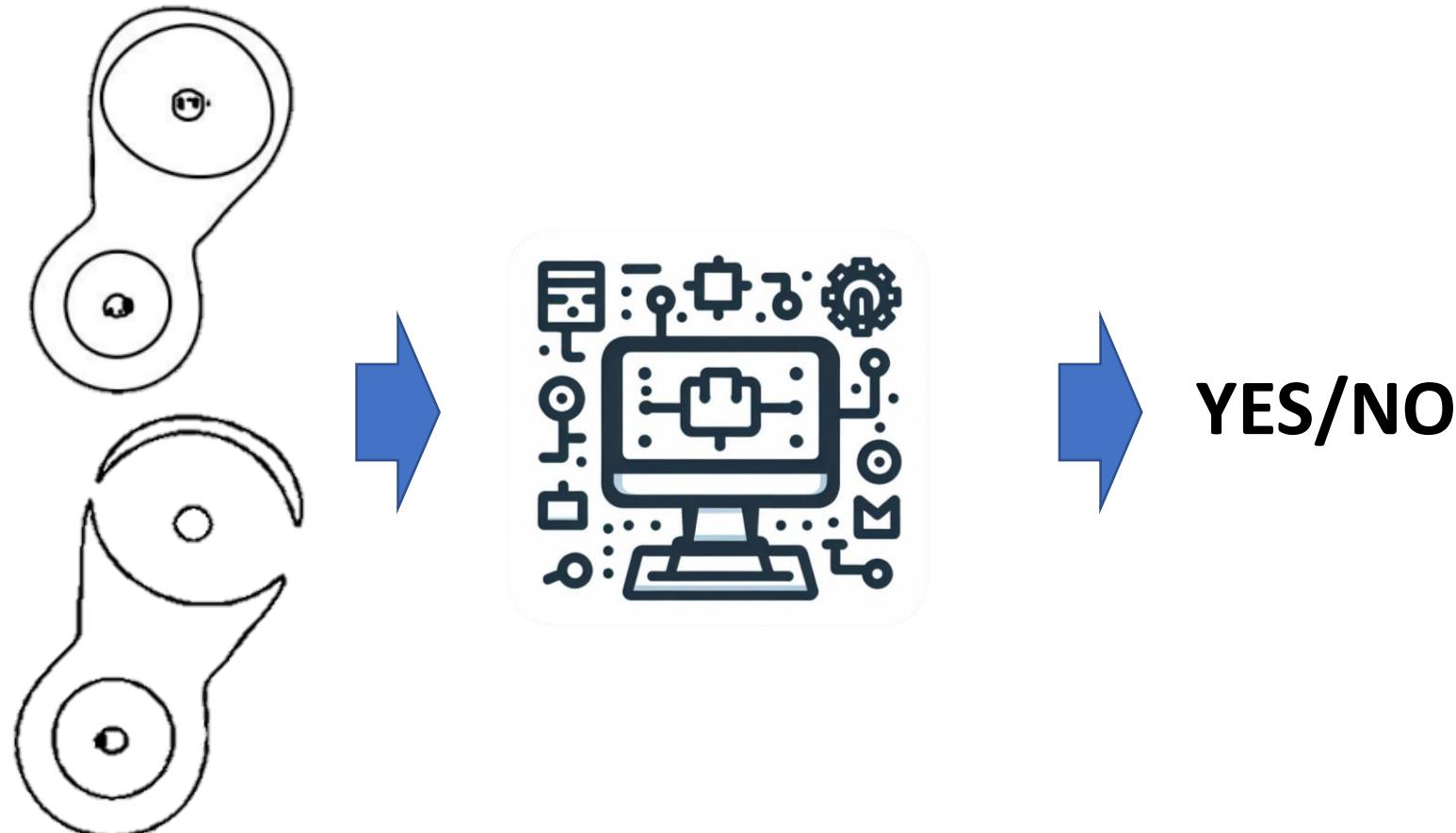


Цель работы:

Разработка свёрточной нейронной сети, проводящей бинарную классификацию наличия активации субстрата в активных центрах гидролаз, и ее применение для определения эффективности активации различных систем: комплексов бактериальной металло-β-лактамазы NDM-1 и L1 с антибиотиком имипенема, а также комплексов капралактама и капралактона с липазой CALB.

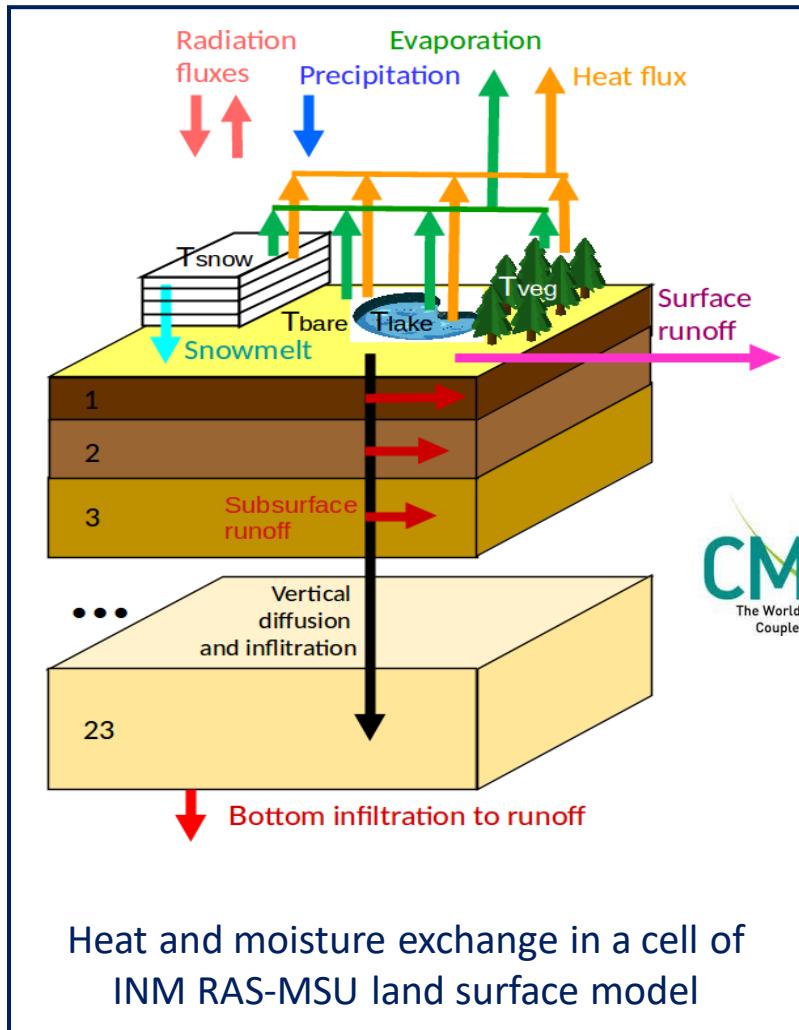
Контролируемое обучение

Реактоспособность субстрата по картам лапласиана эл. плотности



Идентификация ДУЧП

Идентификация ДУЧП, описывающих динамику влаги и тепла в почве



The World Climate Research Programme's
Coupled Model Intercomparison Project

INM RAS-MSU land surface scheme

Climate model of
Institute of
Numerical
Mathematics
(INMCM)

Global numerical
weather
prediction SL-AV
model



- Heat, moisture, water vapor and ice dynamics in soil (23 layers)
- Snow cover with liquid water treatment (4 layers)
- Transpiration and photosynthesis by vegetation
- Wetland CH₄ model
- LAKE model
- River routing scheme
- Standalone version resolution is 0.5 x 0.5 deg.
- GLCC global land cover data

Идентификация ДУЧП

Идентификация ДУЧП, описывающих динамику влаги и тепла в почве

Уравнение теплопроводности

$$\rho c \frac{\partial T}{\partial t} = \frac{\partial}{\partial z} \lambda_T \frac{\partial T}{\partial z} + \dots$$

$$\frac{\partial T}{\partial t} = \frac{\partial}{\partial z} \lambda_T \frac{\partial T}{\partial z}$$

Коэффициент теплопроводности

$$\lambda_T = \lambda_T(W, T)$$

Уравнение Ричардса для динамики жидкой влаги в почве

$$\frac{\partial W}{\partial t} = \frac{\partial}{\partial z} \left(\lambda_W \frac{\partial W}{\partial z} \right) + \frac{\partial \gamma_W}{\partial z} + \dots$$

$$\frac{\partial W}{\partial t} = \frac{\partial}{\partial z} \left(\lambda_W \frac{\partial W}{\partial z} \right) + \frac{\partial \gamma_W}{\partial z}$$

коэффициент влагопроводности

$$\lambda_W = \lambda_W(W, T)$$

коэффициент гидравлической проводимости

$$\gamma_W = \gamma_W(W, T)$$

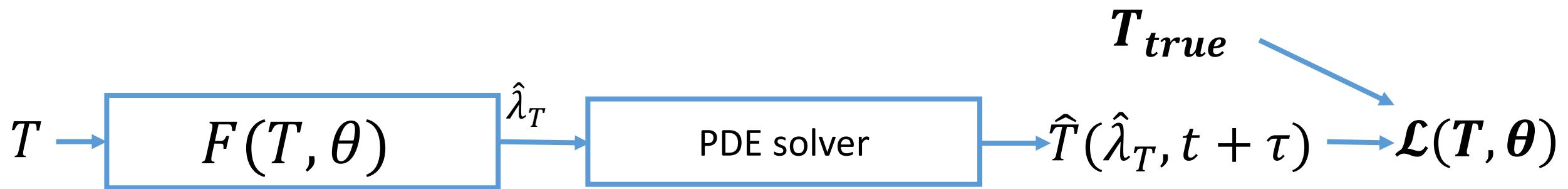
Идентификация дучп

$$\frac{\partial T}{\partial t} = \frac{\partial}{\partial z} \lambda_T \frac{\partial T}{\partial z}$$

$$\lambda_T = F(T, \theta)$$

θ are the parameters
 T is the PDE solution

$$\hat{\theta} = \underset{\Theta}{\operatorname{argmin}} \mathcal{L}(T_{true}, F(T, \theta))$$



$$\mathcal{L}(T, \theta) = MSE(\hat{T}(F(T, \theta)), T_{true})$$

Chain rule:

$$\frac{\partial \mathcal{L}(T, \theta)}{\partial \theta} = 2(\hat{T}(F(T, \theta)) - T_{true}) * \frac{\partial \hat{T}}{\partial F} * \frac{\partial F}{\partial \theta}$$

PDE identification: simulated evolution

$$\frac{\partial T}{\partial t} = \frac{\partial}{\partial z} \lambda_T \frac{\partial T}{\partial z}$$

PDE solver:
explicit scheme

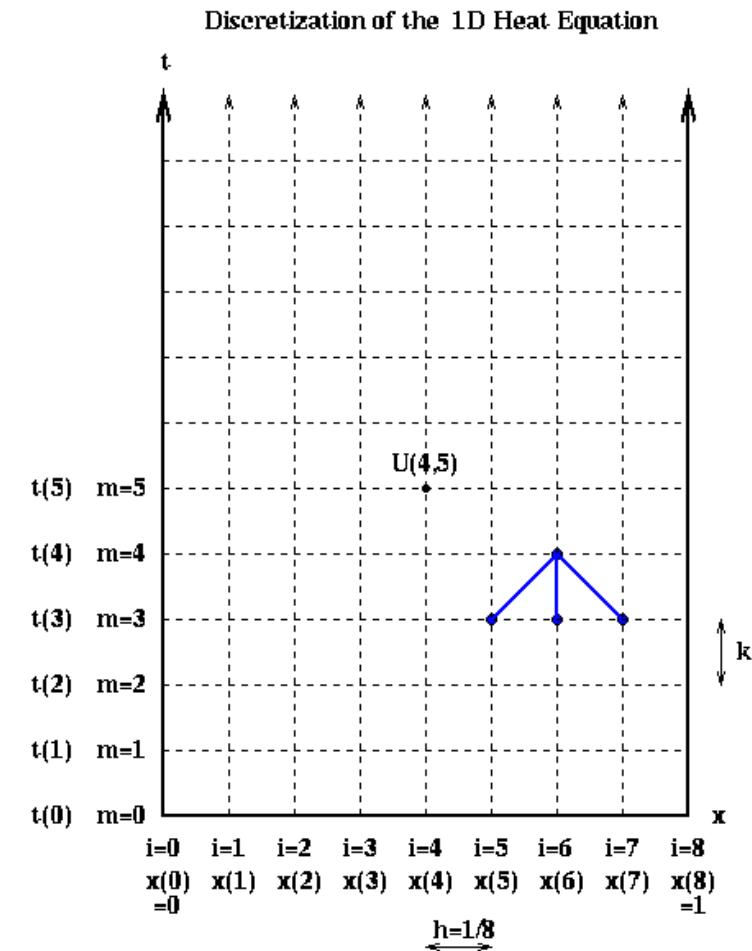
$$\left(\frac{\partial T}{\partial t} \right)_i = \frac{1}{\Delta z_i} \left(\frac{\lambda \left(T_{i+\frac{1}{2}}^t \right) (T_{i+1}^t - T_i^t) - \lambda \left(T_{i-\frac{1}{2}}^t \right) (T_i^t - T_{i-1}^t)}{\Delta z_i + \Delta z_{i+1}} \right)$$

$$T_i^{t+1} = T_i^t + \left(\frac{\partial T}{\partial t} \right)_i \Delta t$$

remember the goal:

$$\frac{\partial \mathcal{L}(T, \theta)}{\partial \theta} = 2(\hat{T} - T_{true}) * \frac{\partial \hat{T}}{\partial \lambda_T} * \frac{\partial F_{NN}}{\partial \theta}$$

and gradient optimization starts here



Results: Richards equation

λ_W

