

Машинное обучение и глубокое обучение в науках об атмосфере, океане и климате: достижения и перспективы в 2024 году

Михаил Криницкий, к.т.н.

Лаборатория машинного обучения в науках о Земле МФТИ,

Институт океанологии РАН

НИВЦ МГУ

Михаил Криницкий

разработка и анализ алгоритмов, основанных на методах машинного обучения и глубокого обучения, в применении к задачам естественных наук

канд. техн. наук,

Зав. Лабораторией машинного обучения в науках о Земле (МФТИ)

с.н.с. ИО РАН, доцент МФТИ, с.н.с. НИВЦ МГУ, с.н.с. ИФА РАН



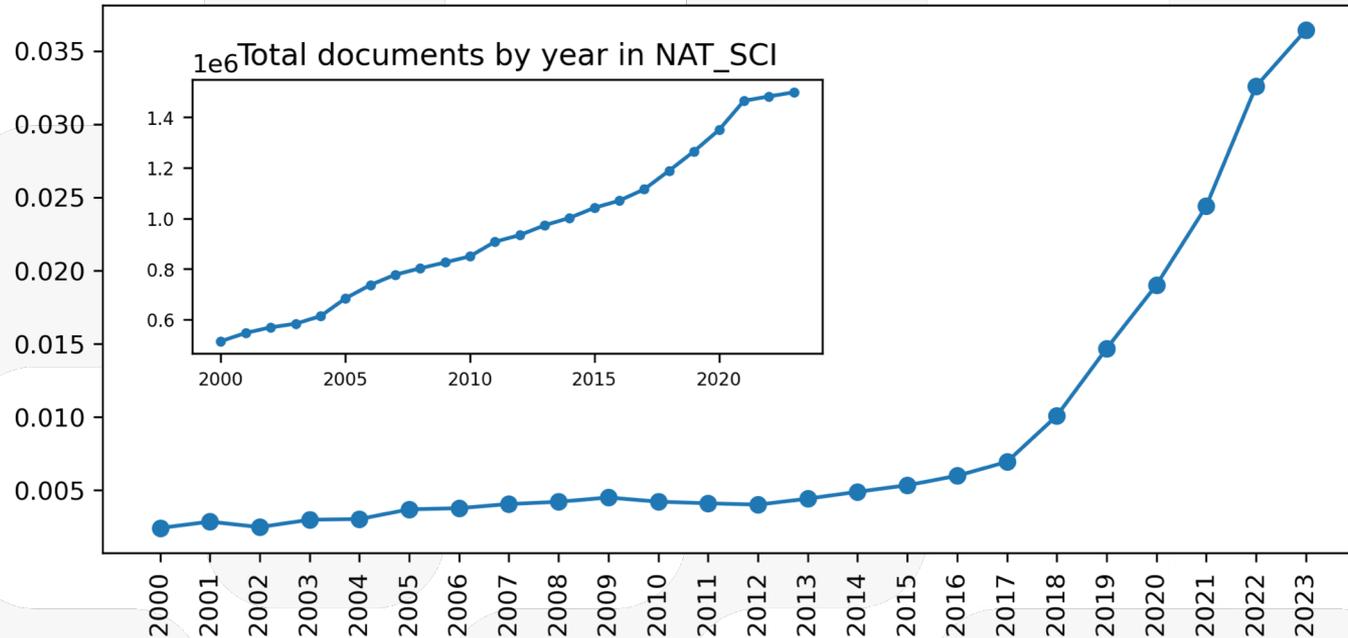
МО в естественных науках

1. Машинное обучение неочевидный выбор при решении задач в науках о Земле
2. Примеры задач в науках о Земле с применением МО
3. Базовые модели: тихая революция 2023 г. в статистическом моделировании погоды и климата;
4. Искусственные нейронные сети (ИНС)
 - ИНС как дифференцируемая функция; примеры эксплуатации
 - Дискриминативные модели vs. порождающие модели: принципы, примеры эксплуатации
5. Вызовы и перспективы

МО в естественных науках

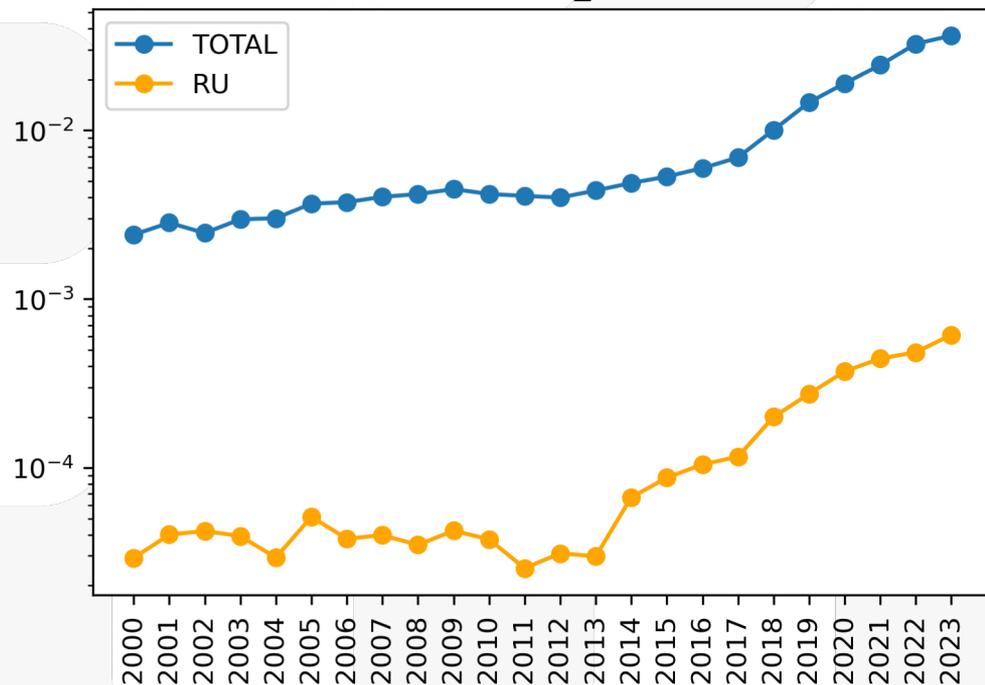
Статьи в рецензируемых журналах по тематикам естественных наук с применением методов машинного обучения и ИИ. По данным Scopus (2000 – 2023 г.).

Documents by year,
relative to total NAT_SCI documents

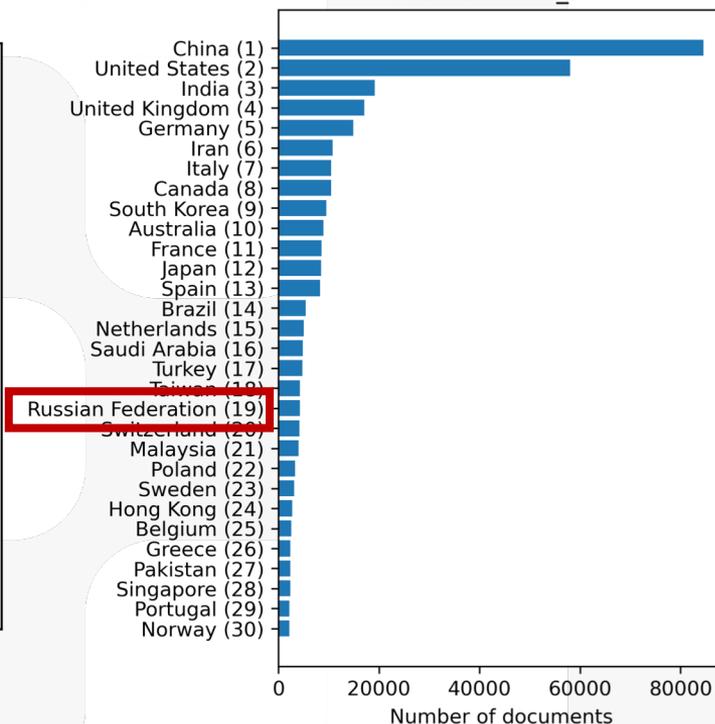


МО в естественных науках

Documents by year,
relative to total NAT_SCI documents

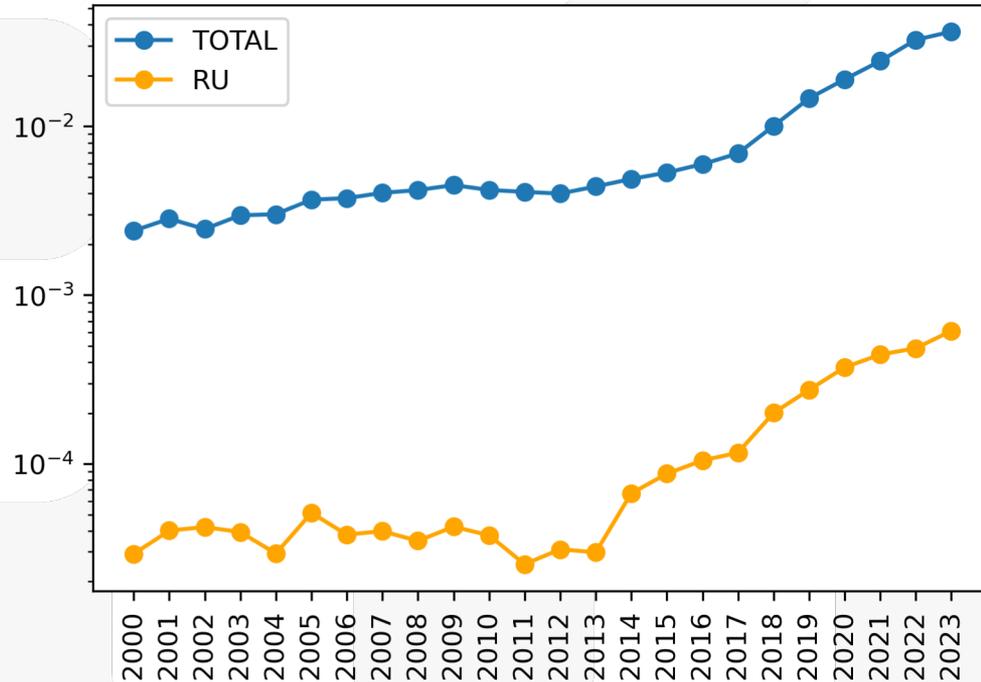


Documents, 2000-2023
AI&ML in NAT_SCI

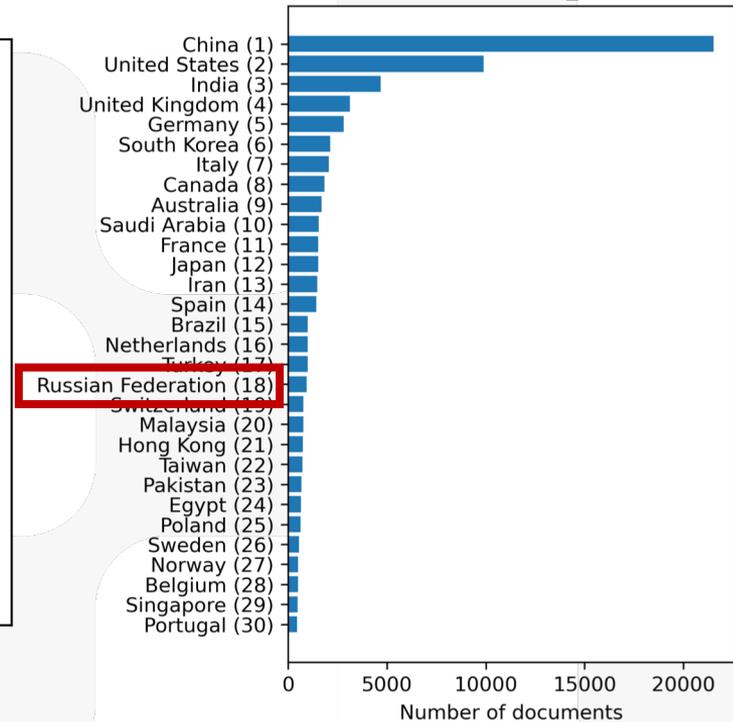


МО в естественных науках

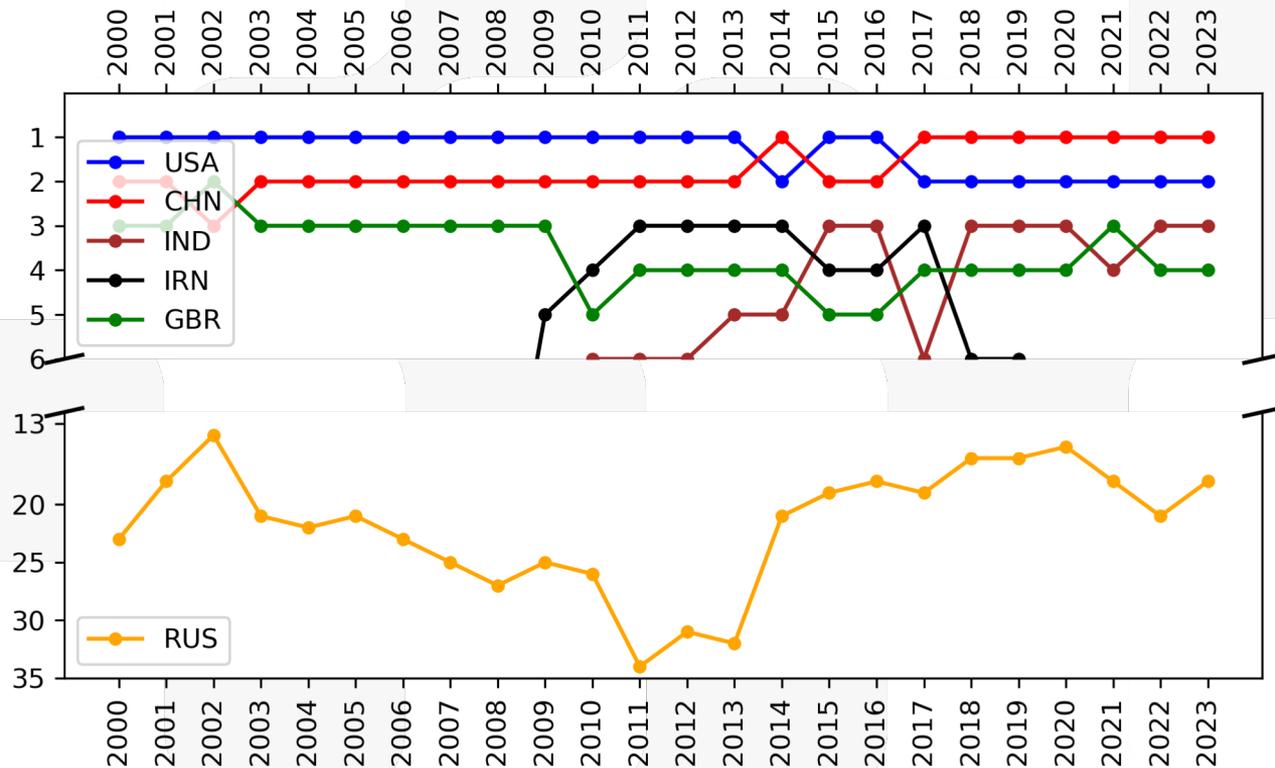
Documents by year,
relative to total NAT_SCI documents



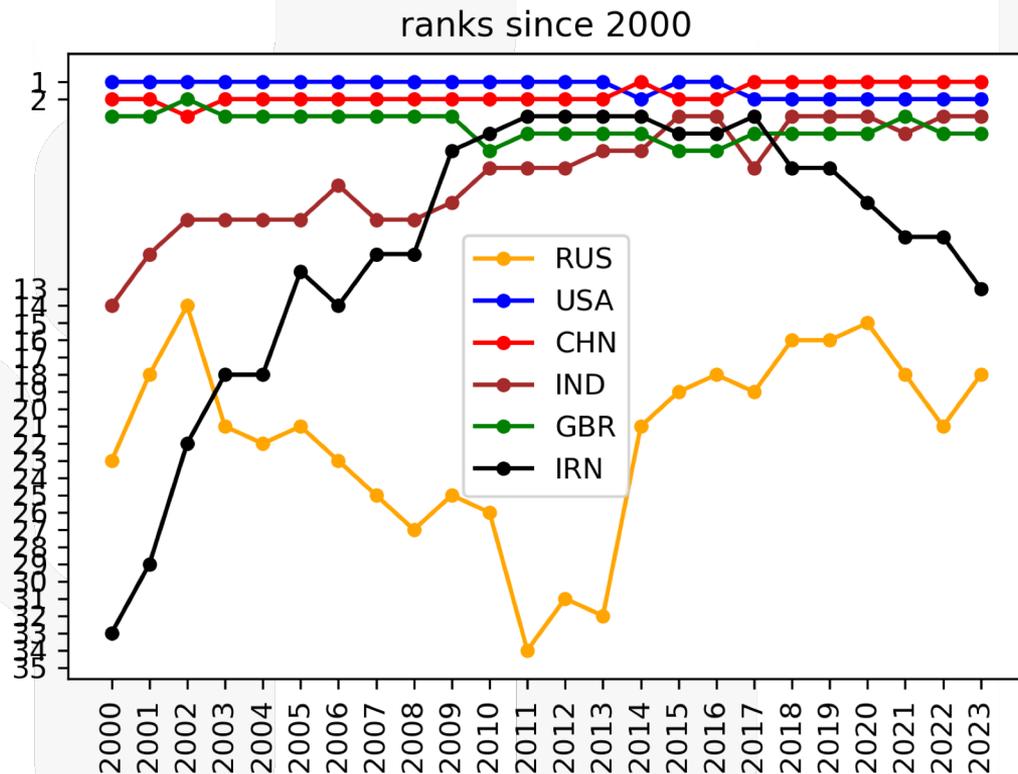
Documents, 2023
AI&ML in NAT_SCI



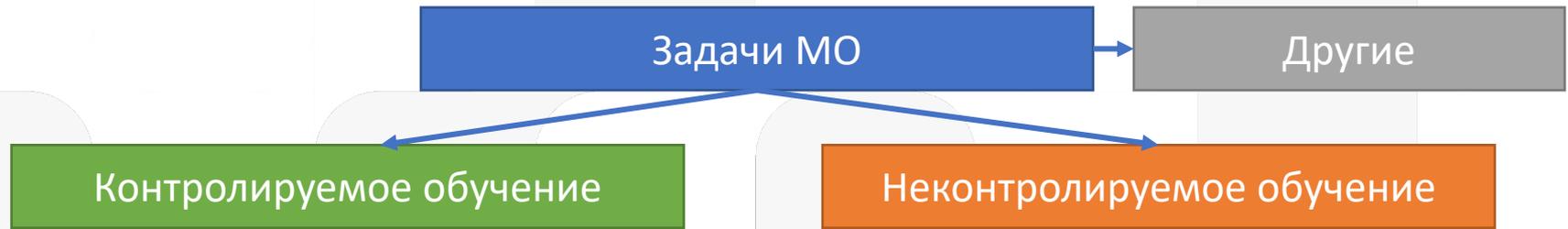
МО в естественных науках



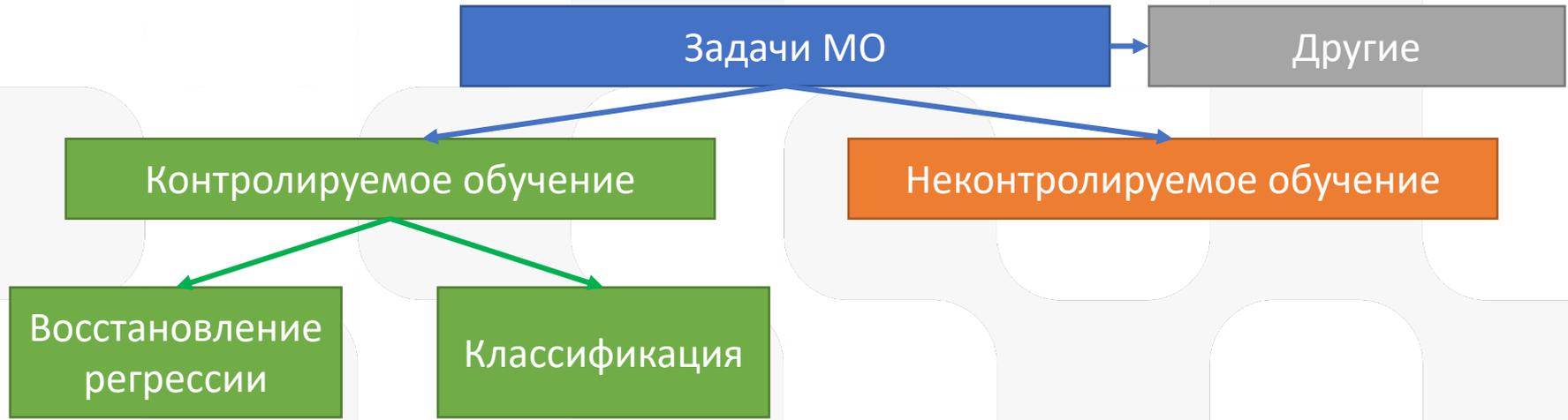
МО в естественных науках



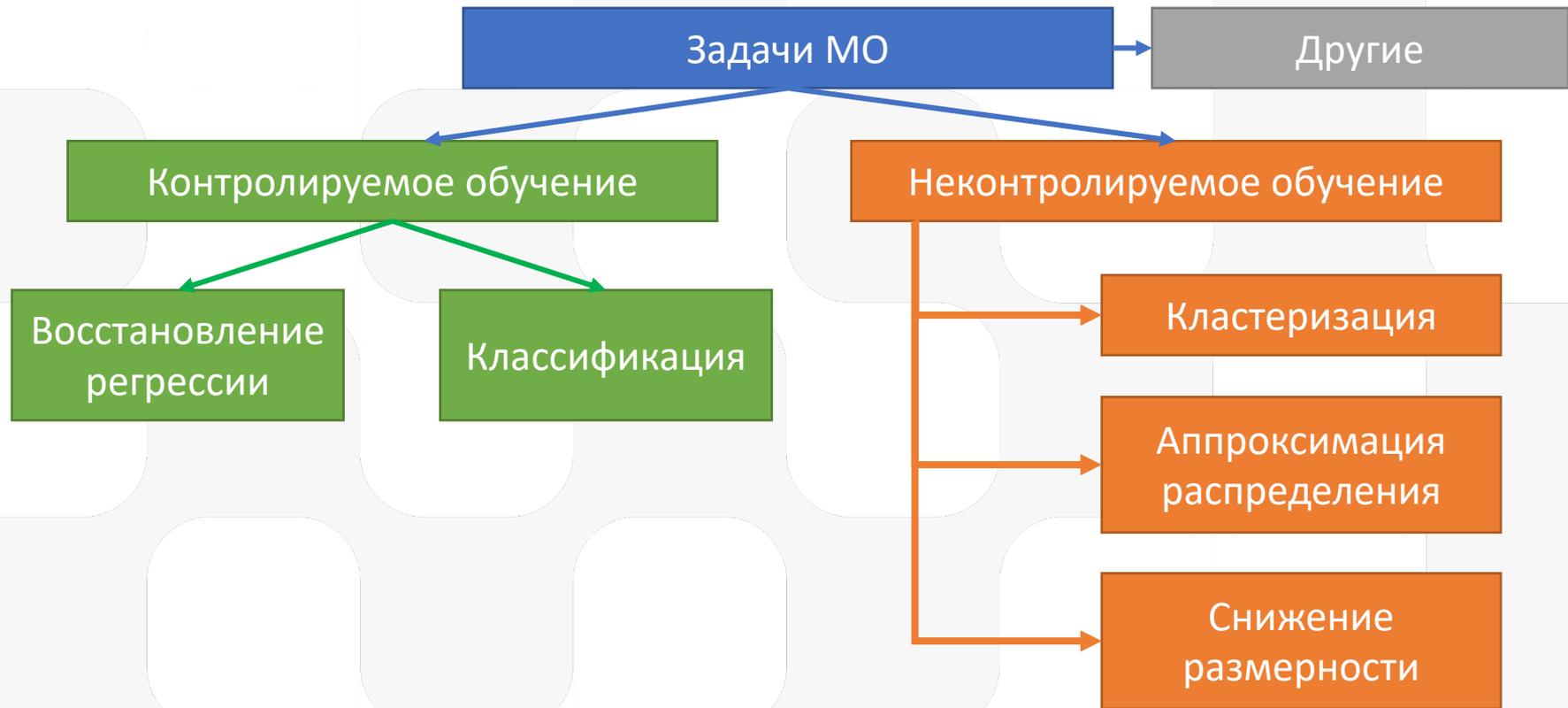
Машинное обучение



Машинное обучение

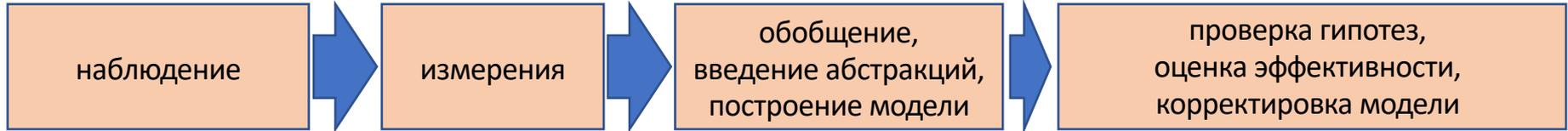


Машинное обучение

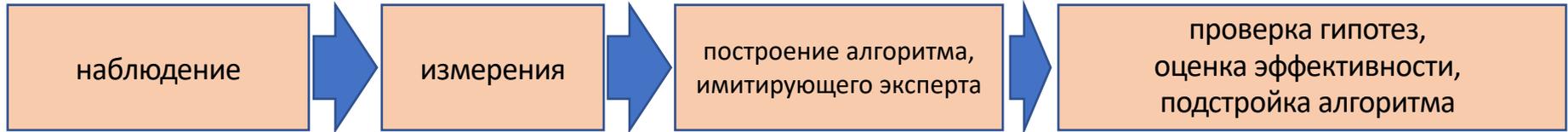


Машинное обучение: место в методологии

«классический» путь развития физических моделей



эвристический способ решения задач (решения, основанные на правилах)



статистический способ (науки, основанные на данных, МО, Data Science)

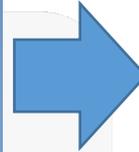


методы машинного обучения

Примеры решения задач в подходе контролируемого обучения

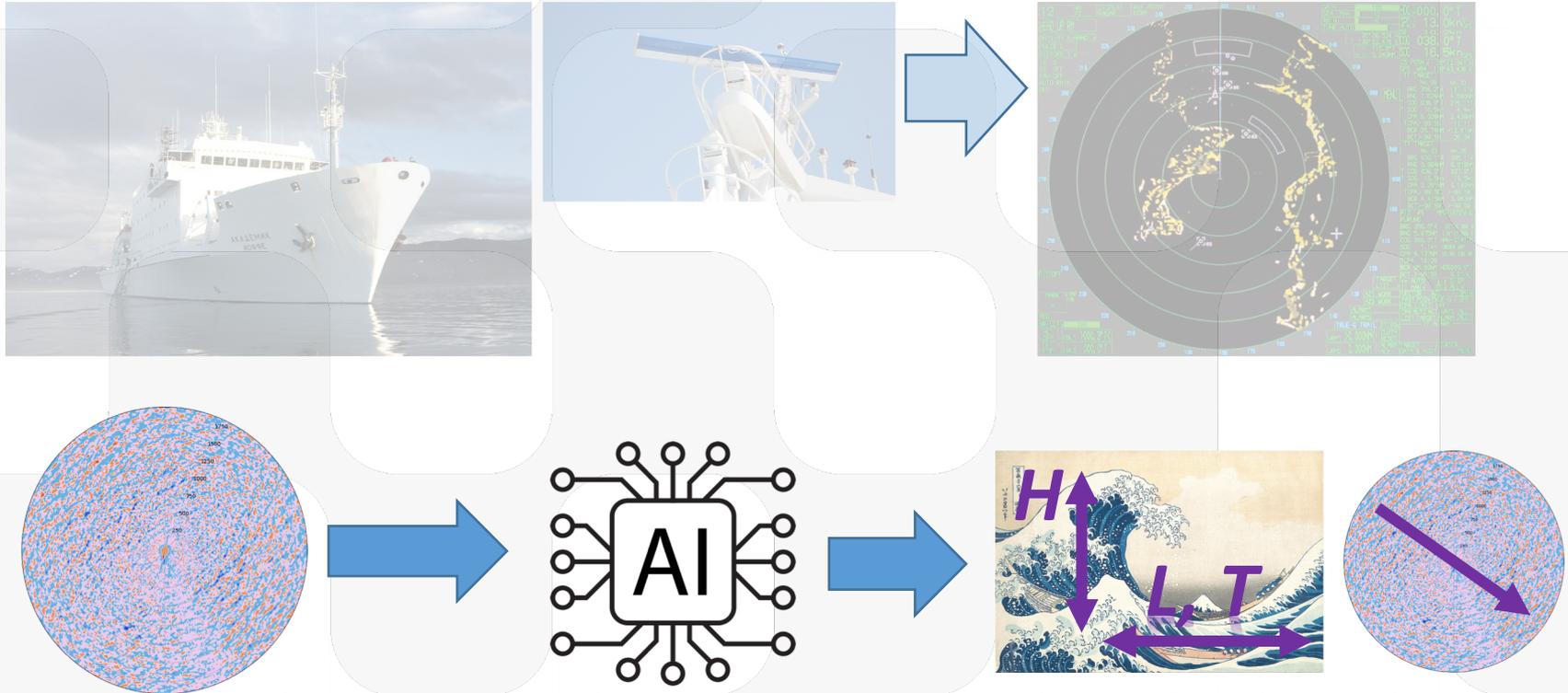
Контролируемое обучение

Характеристики ветрового волнения по данным навигационного радара



Контролируемое обучение

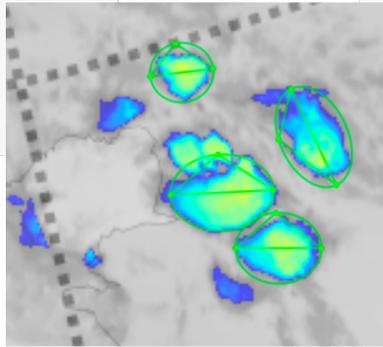
Характеристики ветрового волнения по данным навигационного радара



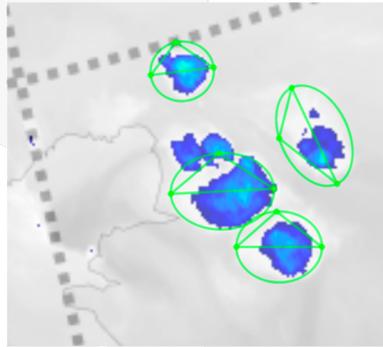
Контролируемое обучение

Идентификация мезомасштабных конвективных систем по данным ДЗЗ

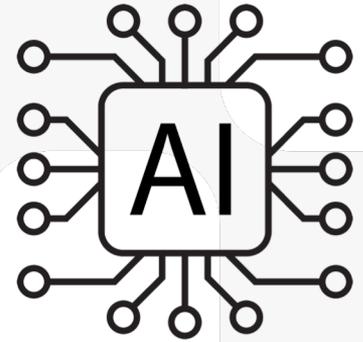
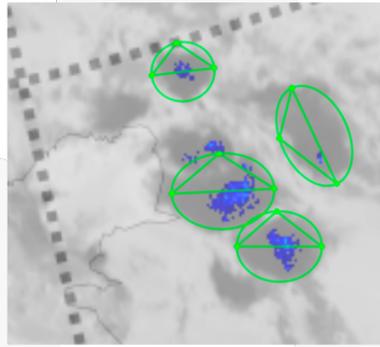
ch9 (10.8 μm)



ch5 (6.25 μm)



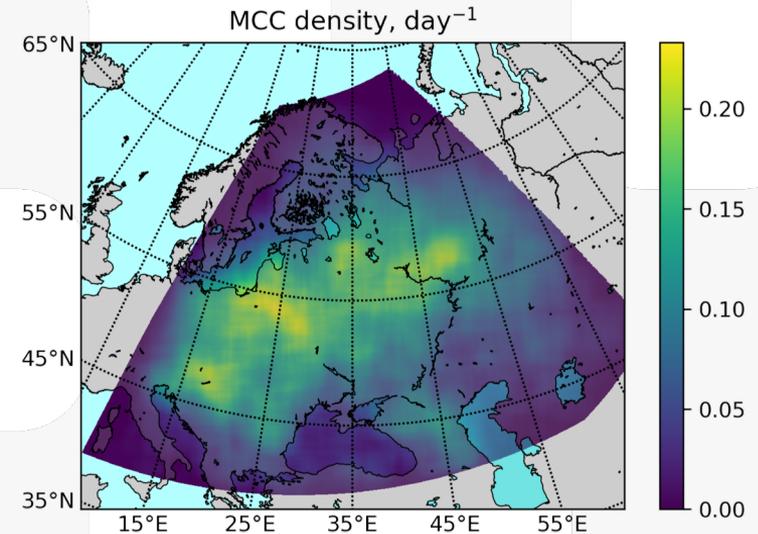
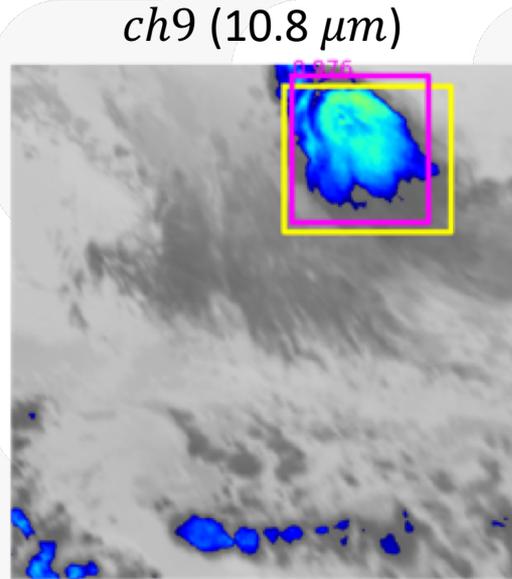
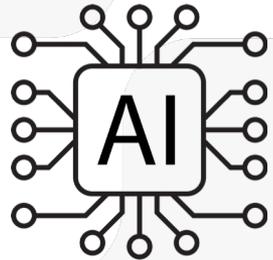
ch5 - ch9



Данные ДЗЗ (Meteosat, MSG4),
Европейская территория России

Контролируемое обучение

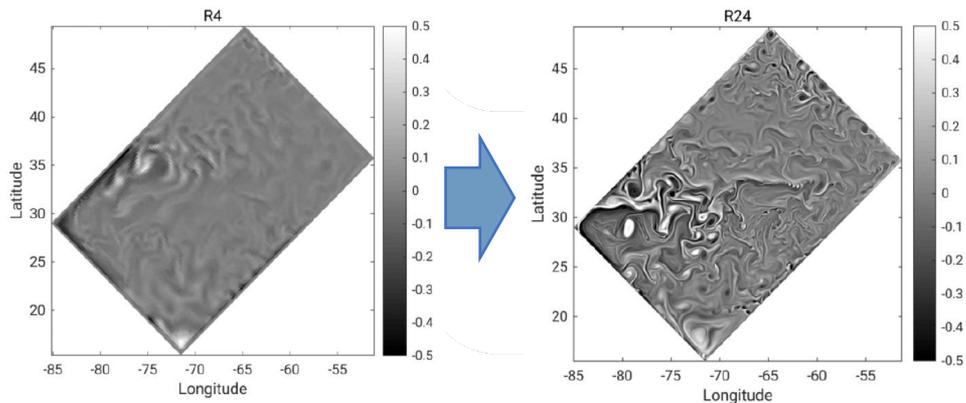
Идентификация мезомасштабных конвективных систем по данным ДЗЗ



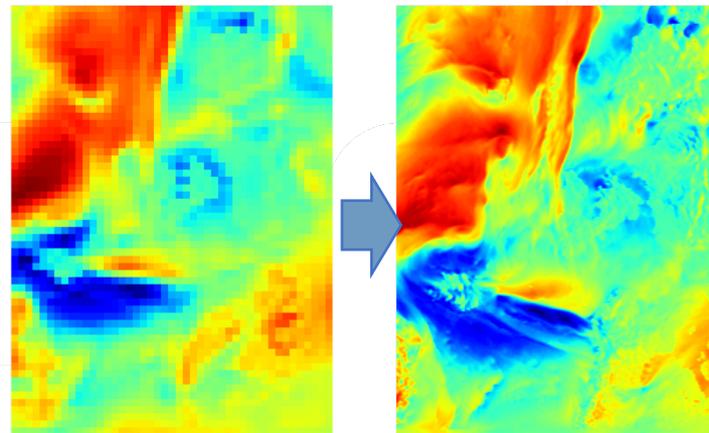
Контролируемое обучение

Статистическое масштабирование приповерхностного ветра в сложном рельефе

Задача статистического масштабирования (даунскейлинга)



Масштабирование динамики течений



Масштабирование скорости ветра

Контролируемое обучение

Статистическое масштабирование приповерхностного ветра в сложном рельефе

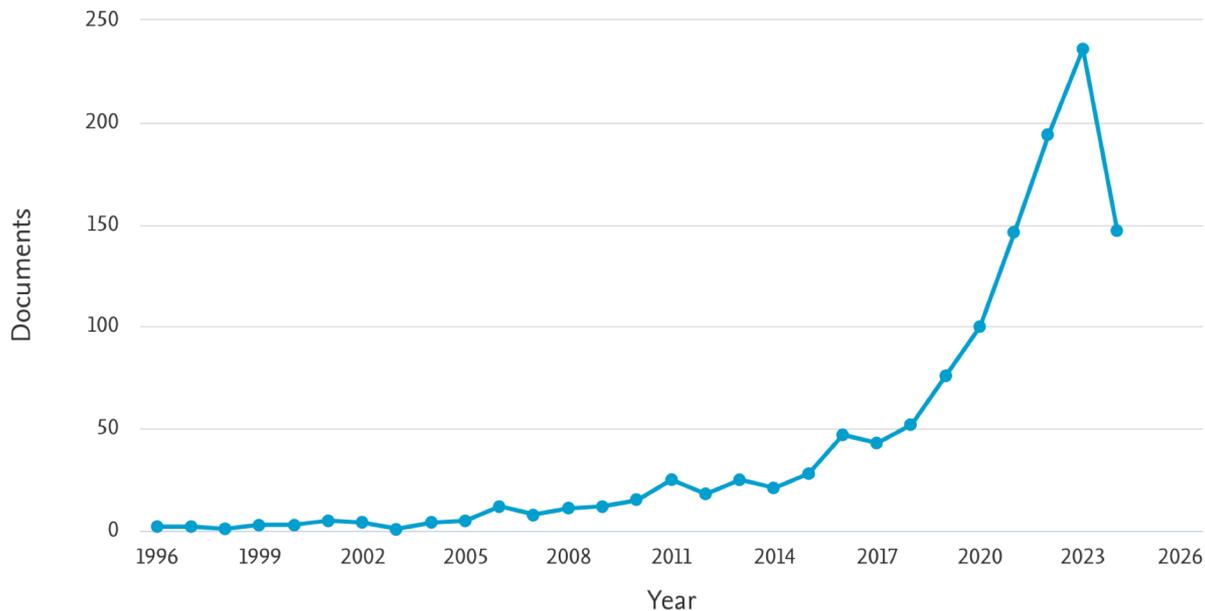
Виды даунскейлинга

1. Масштабирование в точку (уточнение модельных результатов для отдельной локации)
2. Повышение разрешения результатов моделирования 2D+t (3D+t)
3. Коррекция результатов моделирования 2D+t (3D+t) без изменения разрешения
4. Аппроксимация **характеристик распределений** геофизических полей, характерных сложной пространственной структурой
5. Масштабирование результатов климатических моделей

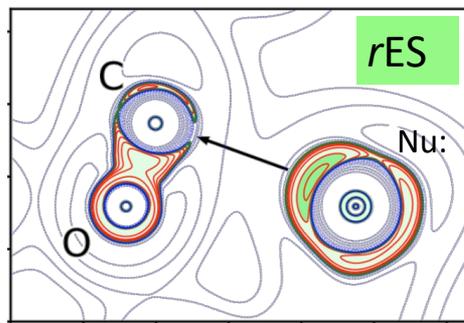
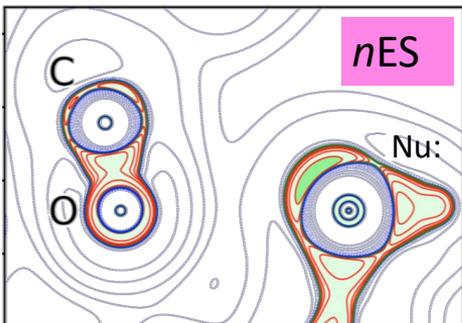
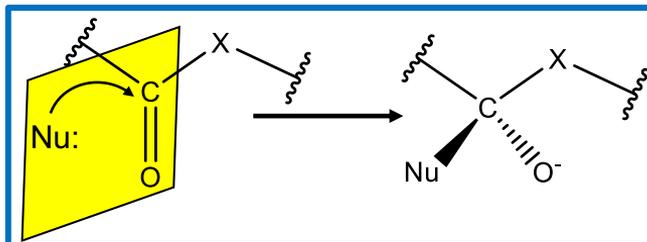
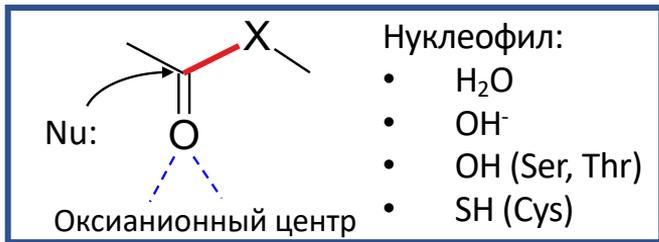
Контролируемое обучение

Статистическое масштабирование приповерхностного ветра в сложном рельефе

Documents by year



downscaling AND ((neural AND networks) OR (machine AND learning))



$$\nabla^2 \rho(\mathbf{r}) = \frac{\partial^2 \rho(\mathbf{r})}{\partial x^2} + \frac{\partial^2 \rho(\mathbf{r})}{\partial y^2} + \frac{\partial^2 \rho(\mathbf{r})}{\partial z^2}$$

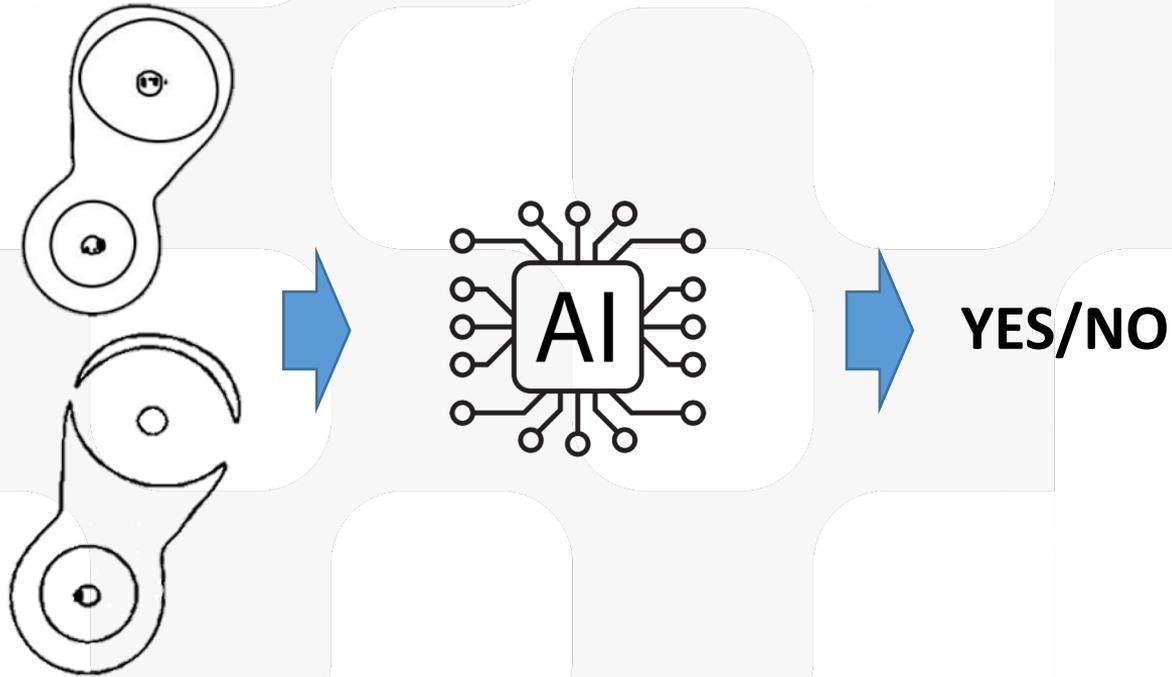
$\nabla^2 \rho > 0$ – область деконцентрации ЭП
 $\nabla^2 \rho < 0$ – область концентрации ЭП



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.

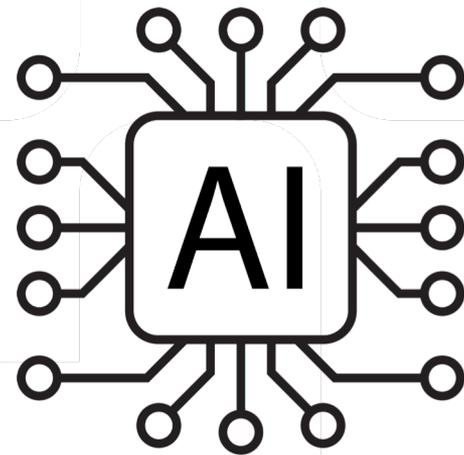
Контролируемое обучение

Реактоспособность субстрата по картам лапласиана эл. плотности



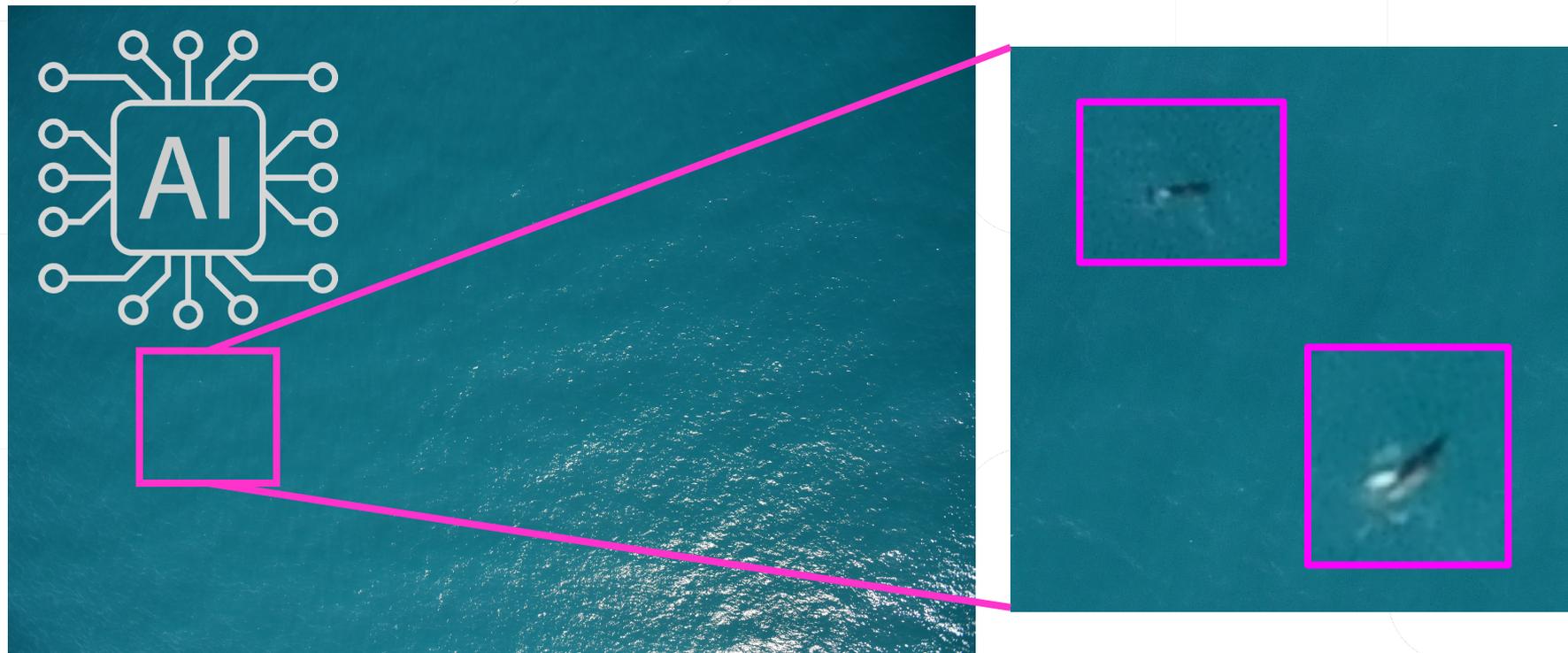
Пример: новые методы исследований

Обнаружение дельфинов в высокоразрешенных учетных фотографиях



Пример: новые методы исследований

Обнаружение дельфинов в высокоразрешенных учетных фотографиях

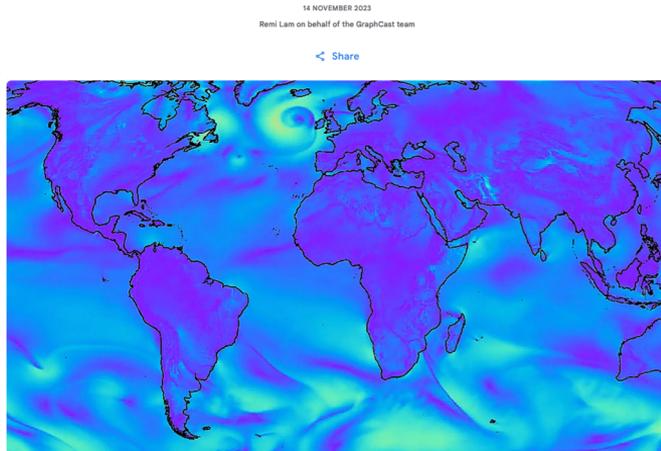


Контролируемое обучение

Нейросетевой прогноз погоды

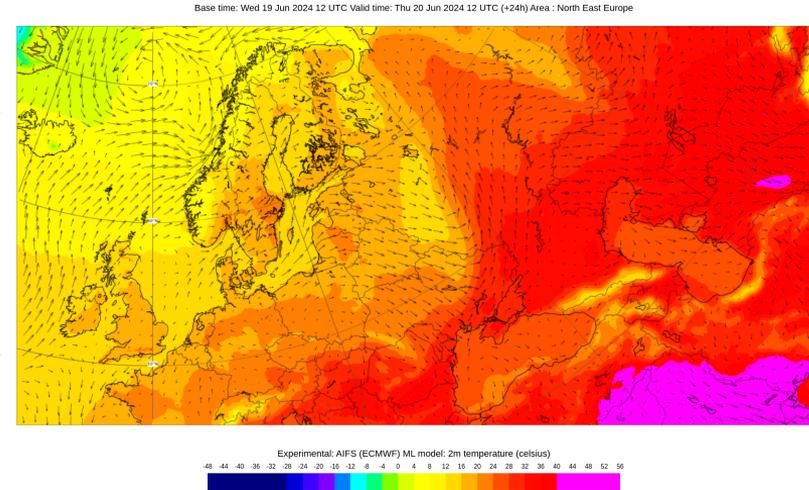
GraphCast (Google)¹

GraphCast: AI model for faster and more accurate global weather forecasting



AIFS (ECMWF)²

Experimental: AIFS (ECMWF) ML model: 2 m temperature and 10 m wind



¹ Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, eadi2336

² <https://www.ecmwf.int/en/newsletter/178/news/aifs-new-ecmwf-forecasting-system>

Контролируемое обучение

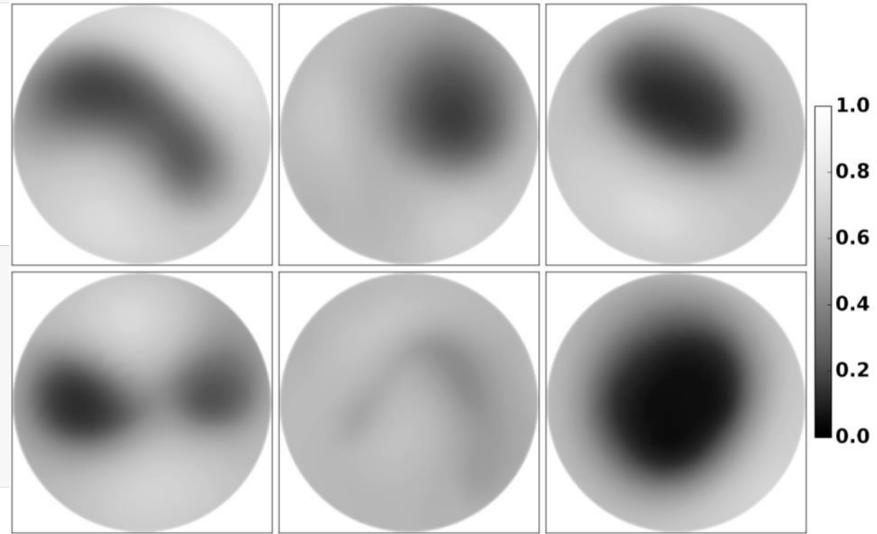
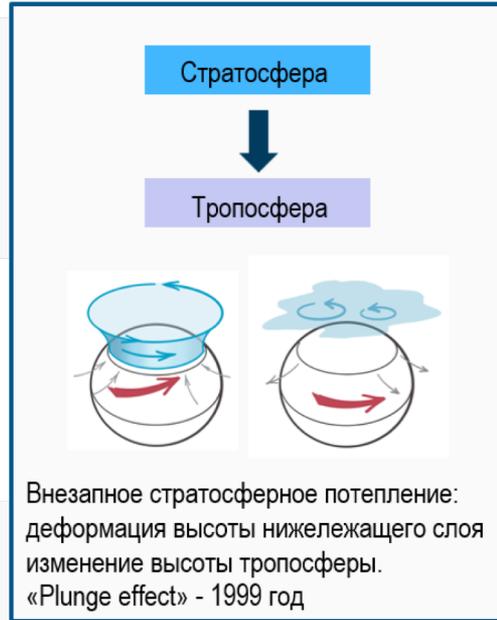
- Автоматизация наблюдений, измерений;
- Исключение человеческого фактора из процесса измерений;
- Расширение сети наблюдений/измерений.

Вызовы:

- Малые объемы обучающих данных, высокая стоимость получения данных;
- Шум в исходных данных, зависимость достоверности данных от чистоты эксперимента;
- Вопрос доверия: классические методы обработки данных традиционно считаются более достоверными

Неконтролируемое обучение

Кластеризация состояний стратосферного полярного вихря

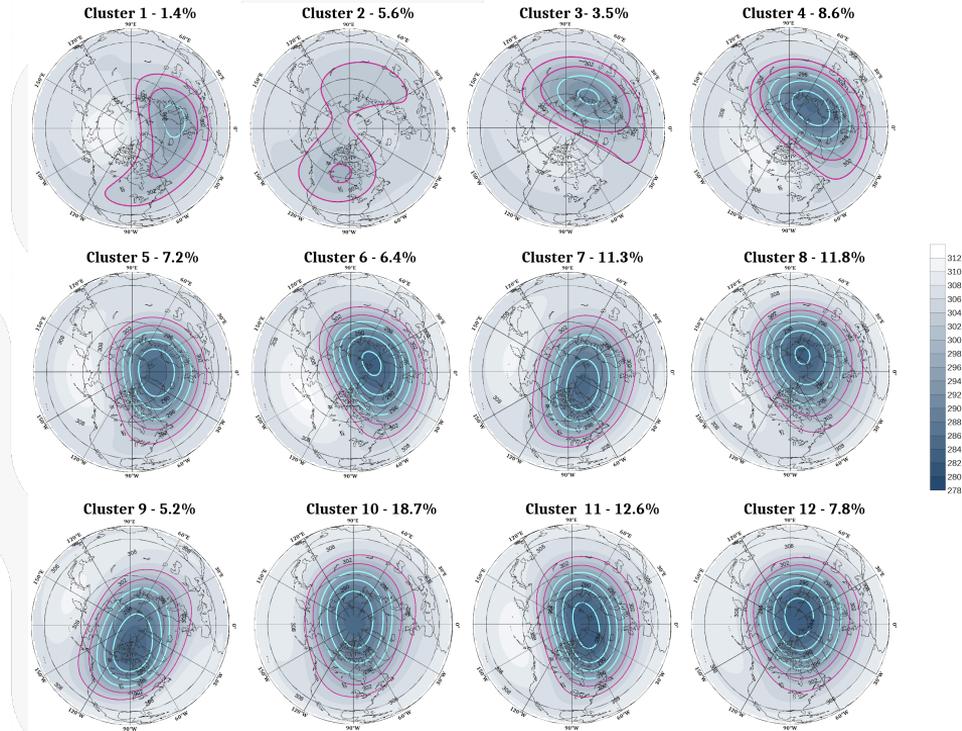


(a)

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

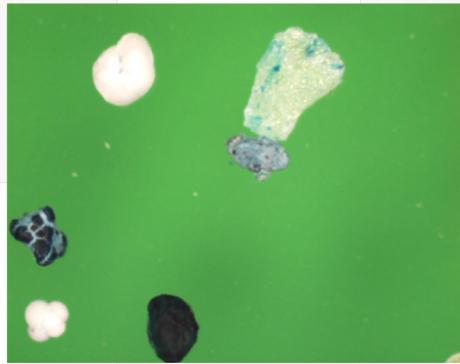
Неконтролируемое обучение

Кластеризация состояний стратосферного полярного вихря

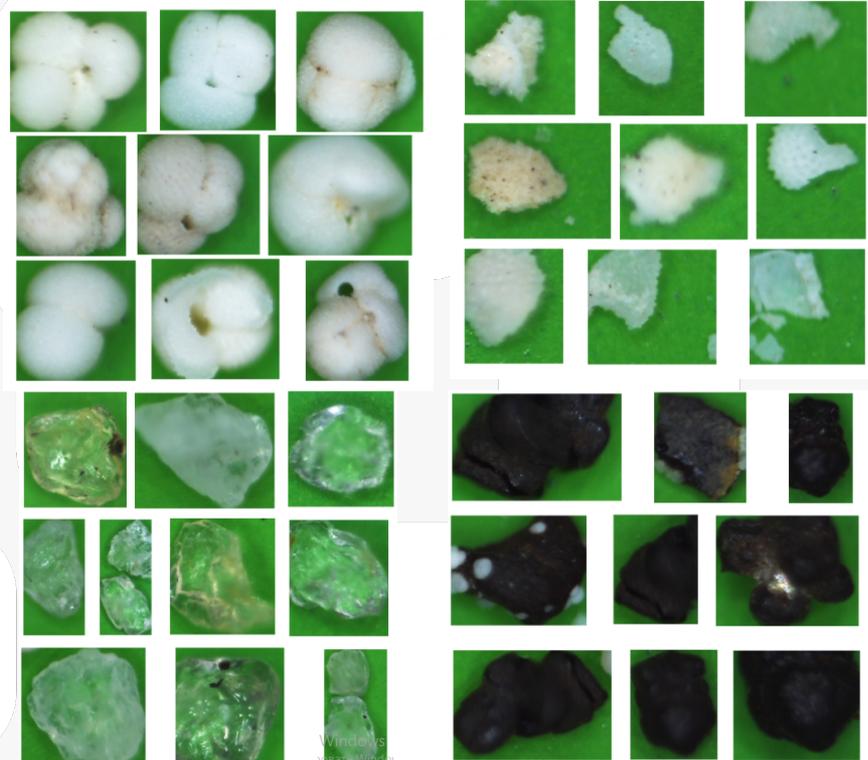
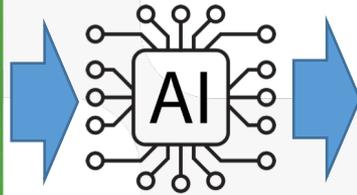


Неконтролируемое обучение

Кластеризация зерен донного осадка на микрофотографиях



×N фотографий



Неконтролируемое обучение

- Обнаружение структуры в данных;
- Повышение объективности классификации объектов/событий по сравнению с экспертной таксономией;

Вызовы:

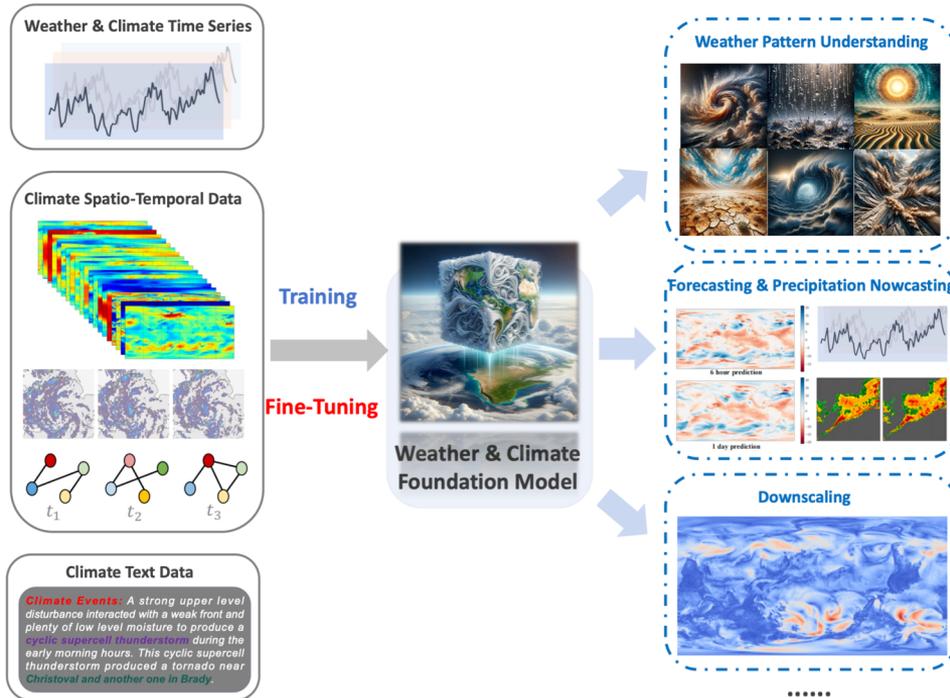
- Некорректность задачи кластеризации
 - отсутствие строгой постановки задачи;
 - множественность решений;
 - отсутствие обоснованной меры качества;
 - решение задачи сильно зависит от определения метрики;
- => вопрос доверия, надежности интерпретации результатов

Самоконтролируемое обучение

Базовые нейросетевые модели:
тихая революция
в мире нейросетевого прогнозирования
погоды и климата

Самоконтролируемое обучение

Базовые нейросетевые модели



Chen, S., Long, G., Jiang, J., Liu, D., & Zhang, C. (2023). Foundation models for weather and climate data understanding: A comprehensive survey. *arXiv preprint arXiv:2312.03014*.

Самоконтролируемое обучение

Большие базовые модели для климатических и погодных данных

1. Microsoft ClimaX ^{1,*} – Jan'2023, UCLA (USA)	1.40625°	T
2. FengWu ^{2,*} – Apr'2023, China (6 организаций)	0.25°	T
3. PanGu ^{3,*} – July'2023, Huawei (China)	0.25°	T
4. FuXi ^{4,*} – Jun'2023, Fudan University (China)	0.25°	T
5. FourCastNet ^{5,*} – Feb'2022, NVIDIA	0.25°	FNO
6. GraphCast ^{6,*} – Nov'2023, Google	0.25°	G
7. W-MAE ^{7,*} – Apr'2023, UEST (China)	0.25°	T

* доступны для скачивания, использования, код и веса в свободном доступе

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.

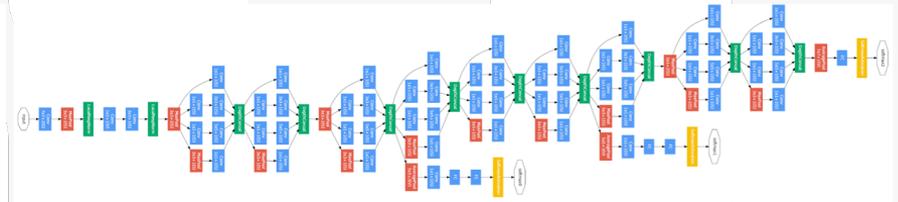
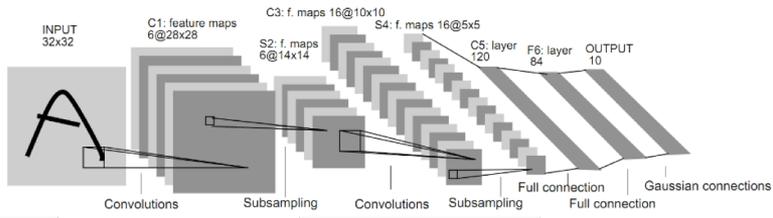
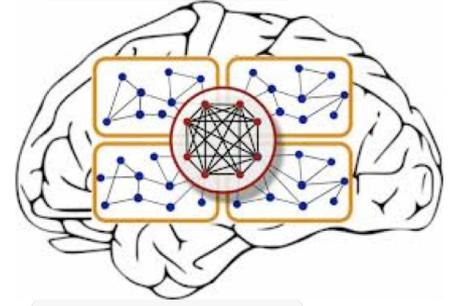
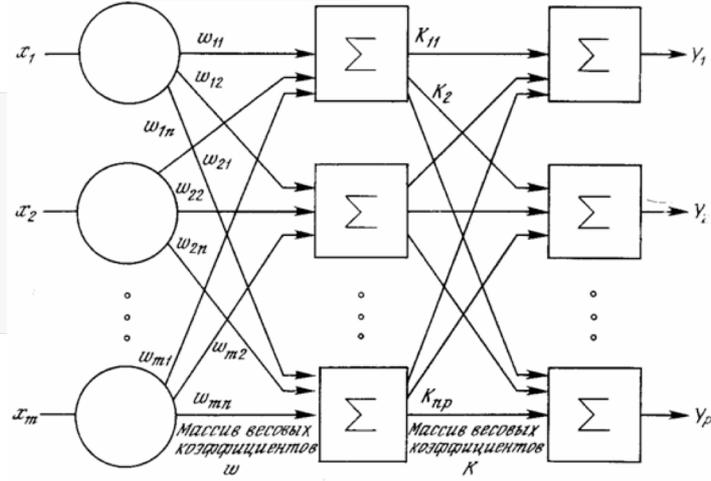
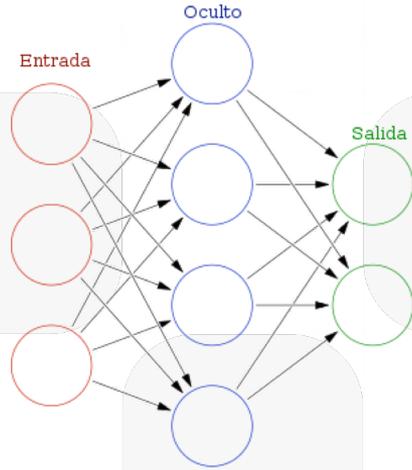
⁴ Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., & Li, H. (2023). FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1), 190.

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*.

Искусственные нейронные сети

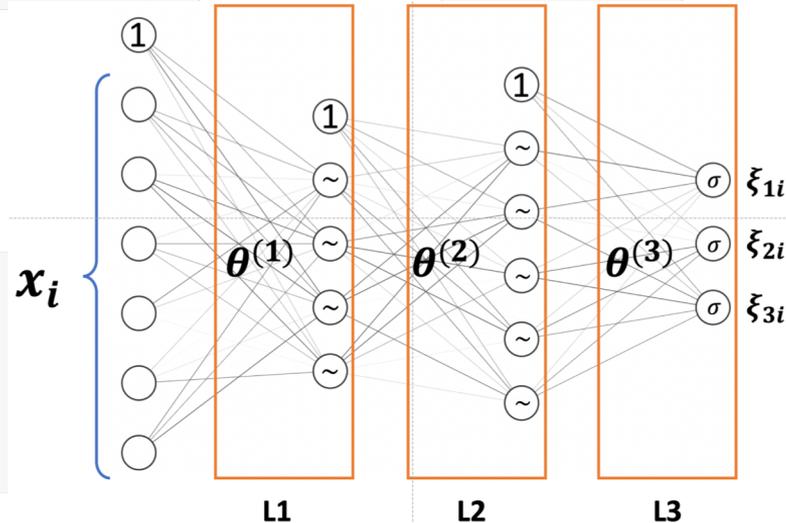


Искусственные нейронные сети

- ИНС - функция

$$\xi = \mathcal{F}(\theta, x)$$

$$\xi_i = \sigma \left(\theta_0^{(3)} + \theta^{(3)} \cdot \phi \left(\theta_0^{(2)} + \theta^{(2)} \cdot \phi \left(\theta_0^{(1)} + \theta^{(1)} \cdot x_i \right) \right) \right)$$



Искусственные нейронные сети

- ИНС – сложная нелинейная дифференцируемая функция

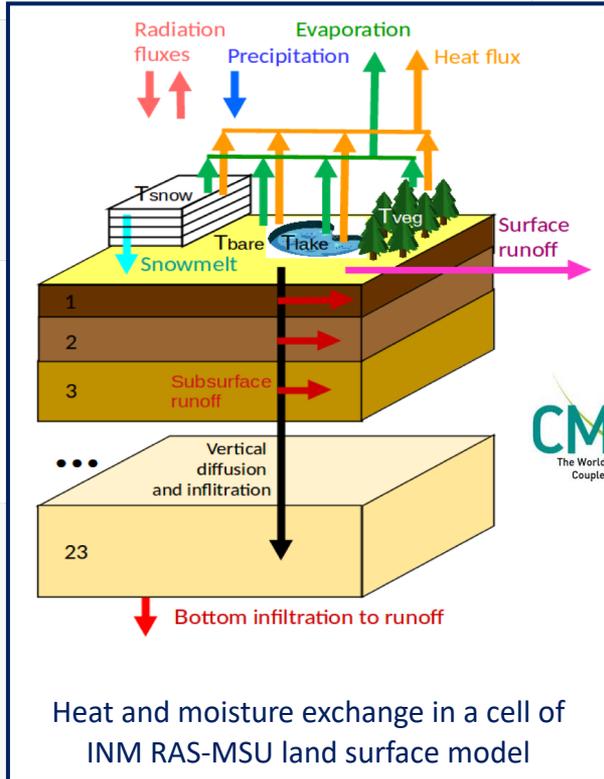
$$\xi = \mathcal{F}(\theta, x)$$

$$\mathcal{L} = \mathcal{L}(\mathcal{F}(\theta, \{x_i, y_i\}))$$

$\nabla_{\theta} \mathcal{L} \Rightarrow$ оптимизация $\mathcal{F}(\theta, x)$
для решения задачи \mathcal{L}

Искусственные нейронные сети

Идентификация ДУЧП, описывающих динамику влаги и тепла в почве



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

Искусственные нейронные сети

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

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

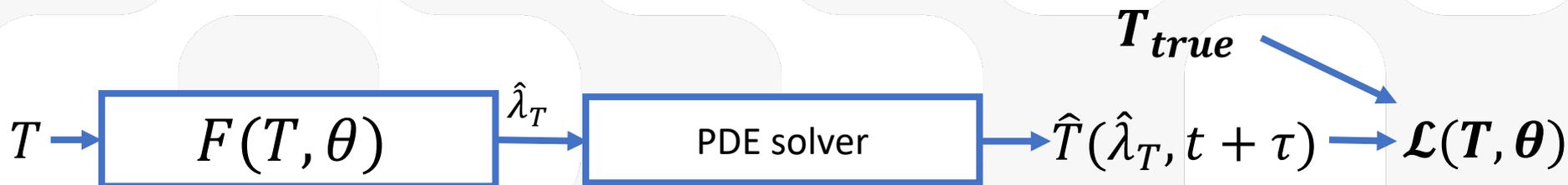
Искусственные нейронные сети

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

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

θ are the parameters
 T is the PDE solution

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



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

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

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

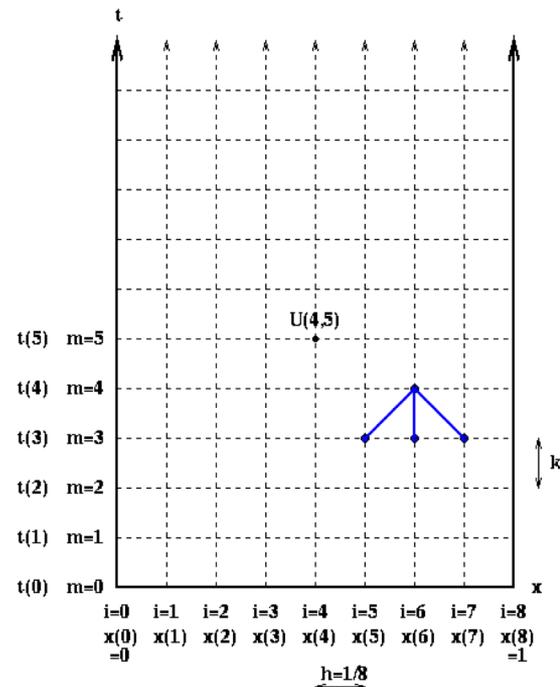
$$T_i^{t+1} = T_i^t + \left(\frac{\partial T}{\partial t}\right)_i \Delta t$$

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

Discretization of the 1D Heat Equation



Искусственные нейронные сети

ИНС – дифференцируемая сложная функция:

- Можно использовать для идентификации ДУЧП;
- (ПЕРСПЕКТИВА) можно заменять отдельные блоки/параметризации геофизических моделей; обучать - на данных наблюдений измеряемых характеристик, которые легче сравнивать с модельными переменными

Вызовы:

- Идентификация ДУЧП – обратная задача => решение может быть неустойчивым;
- Для реального применения нужно дифференцировать оператор интегрирования на несколько шагов по времени;
- Для замены параметризации нужно дифференцировать оператор интегрирования модели и все дальнейшие операции в графе вычислений геофизической модели;

Классы вероятностных моделей

Вероятностные модели (ИНС)

```
graph TD; A[Вероятностные модели (ИНС)] --> B[Дискриминативные]; A --> C[Порождающие (генеративные)];
```

Дискриминативные

Порождающие
(генеративные)

Классы вероятностных моделей

Дискриминативные

$$\hat{y}_i = \underset{Y}{\operatorname{argmax}} \mathcal{P}(y|\theta, x_i)$$

\mathcal{P} – из некоторого семейства **по выбору исследователя**

Например:

$$\left\{ \begin{array}{l} y \sim \mathcal{N}(\mu_i, \sigma^2) \\ \mu_i = \mathcal{F}(\theta, x_i) \end{array} \right. \Rightarrow \text{MLE} \Rightarrow \mathcal{L} = \sum_i (y_i - \mathcal{F}(\theta, x_i))$$

Классы вероятностных моделей

Порождающие
(генеративные)

$$\hat{y}_i \sim \mathcal{P}(y|\theta, x_i)$$

\mathcal{P} – моделируется (нейросетью)

Например:

$$\begin{cases} z \sim \mathcal{N}(0, 1) \\ \hat{y}_i = \mathcal{F}(z, \theta, x_i) \end{cases}$$

Обучение \mathcal{F} :

- GAN
- Diffusion-based
- Normalizing Flows

Классы вероятностных моделей

Вероятностные модели (ИНС)

Дискриминативные

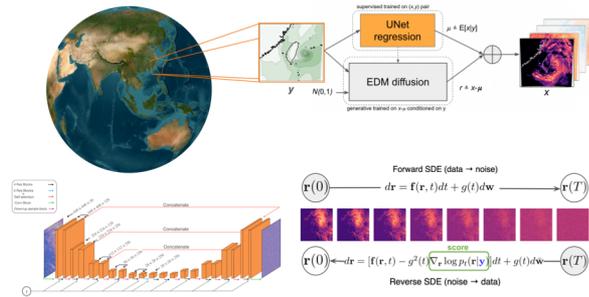
- Обучение быстрое и эффективное;
- Одно значение y для одного x ;
- Оценка неопределенностей на основе ансамбля моделей (долго, дорого).

Порождающие
(генеративные)

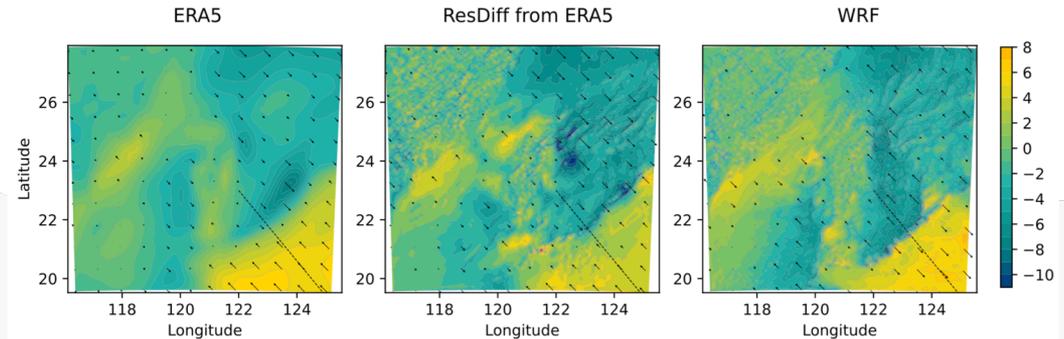
- Обучение менее стабильное и более долгое и дорогое;
- Для каждого x_i – распределение y , из которого можно (быстро и недорого) породить y_i ;
- Оценка неопределенностей на основе ансамбля реализаций y_i .

Порождающие нейросети

Статистическое масштабирование (даунскейлинг) полей скорости жидкости



«Классические» нейросети
+ генеративные нейросети



чуть менее точно,
быстро

Точно, но дорого и долго

Порождающие нейросети

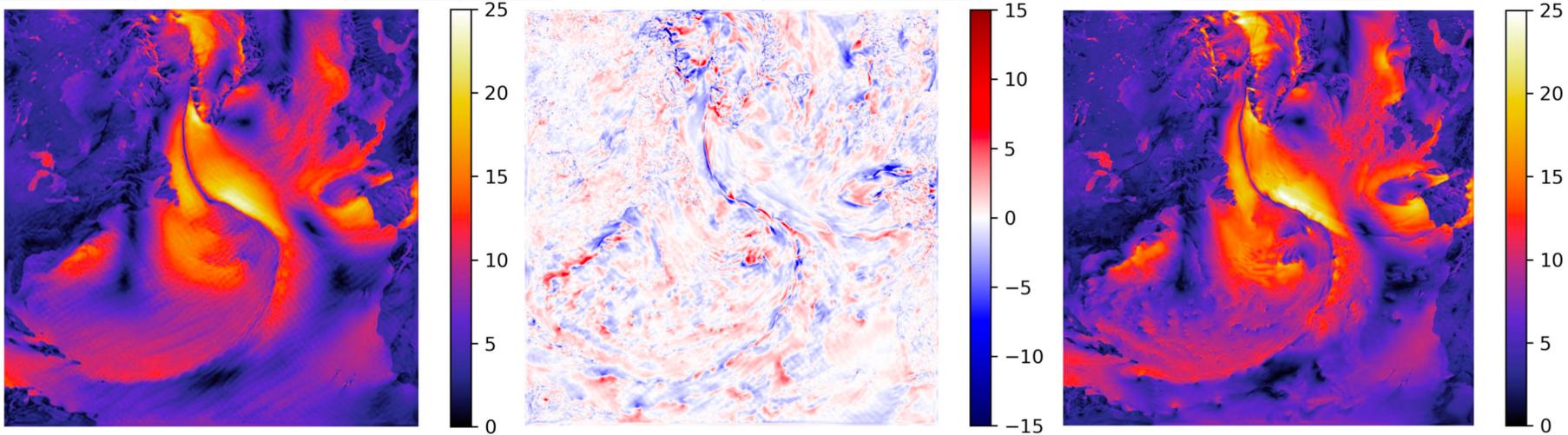
Статистическое масштабирование (даунскейлинг) геофизических полей

SR-GAN для стат. масштабирования приповерхностного ветра*

Low-res

error

Downscaled



Качество выше бикубической интерполяции*

* Rezvov, V. Y., M. Krinitskiy, and S. Gulev. "Approximation of high-resolution surface wind speed in the North Atlantic using discriminative and generative neural models based on RAS-NAAD 40-year hindcast." *The 6th International Workshop on Deep Learning in Computational Physics*. 2022.

Порождающие нейросети

- Статистическое масштабирование возможно;
- Статистическое масштабирование предоставляет возможности ускорения моделирования;
- «Суррогатное» моделирование – аппроксимация эволюции геофизической системы нейросетью, – может использоваться для:
 - подбора параметров модели;
 - оценки неопределенностей;
 - нейросетевого усвоения данных.

Вызовы:

- Необходимо учитывать ограничения, диктуемые физикой процессов.

МО в естественных науках

Возможности:

- **Автоматизация** мониторинговых наблюдений, исключение человеческого фактора, расширение базы наблюдений;
- **Ускорение** камеральных исследований, обработки данных;
- **Поиск структуры** в данных измерений, обнаружение закономерностей и феноменов;
- **Выявление неочевидных механизмов** взаимодействий, нелинейных связей, поиск структуры в данных;
- **Расширение возможностей** мониторинговых и исследовательских измерений в поле;
- **Решение различных конечных задач** с использованием **единой** базовой нейросетевой модели

МО в естественных науках

Вызовы:

- Слишком малые / слишком большие объемы обучающих данных;
- Недостаток вычислительных ресурсов
- Высокая стоимость получения данных (разметки);
- Шум в исходных данных, зависимость достоверности данных от чистоты эксперимента, мнения эксперта;
- Попытки внедрения физики в стат. модели (PINN) пока не увенчались оглушительным успехом;
- Остро стоит вопрос доверия к результатам.

МО в естественных науках

Перспективы:

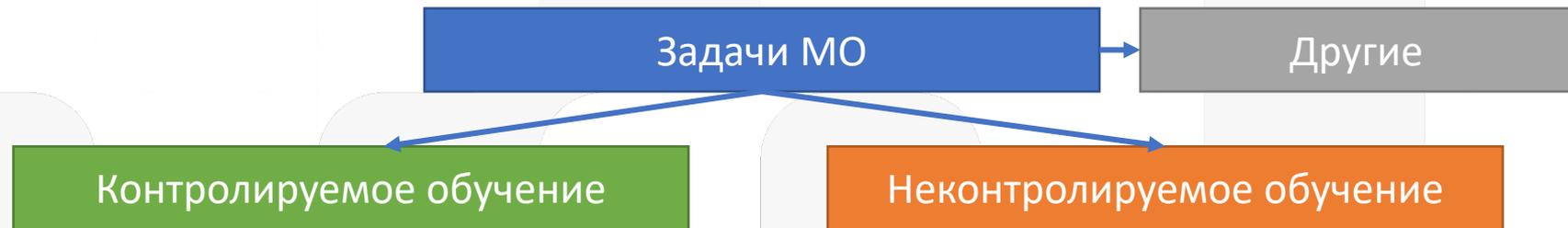
- **Эволюционное развитие** приложений дискриминативных моделей МО в физике;
- **Генеративные модели** дадут **эволюционное развитие** МО в физике;
- Замена эмпирических соотношений в физических моделях на нейросетевые аппроксимации => **уточнение + ускорение** моделирования;
- Суррогатное нейросетевое моделирование (генеративными моделями) => **оценка неопределенностей** моделирования;
- **Внедрение физических ограничений** в стат. модели (PINN) – путь к повышению качества моделирования физических процессов;
- Развитие базовых моделей – путь к решению **сразу многих задач**.

Спасибо за внимание

Михаил Криницкий, к.т.н.

МФТИ, Институт океанологии РАН, МГУ

Машинное обучение



$x \in X$ — объекты / события

$y \in Y$ — «ответы», «метки»

$\mathcal{F}: X \rightarrow Y$ — искомая закономерность

$\mathcal{T}: \{x_i; y_i\}$ — прецеденты

Найти: $\hat{\mathcal{F}}: \{x_i\} \rightarrow \{y_i\}$

Контролируемое обучение

«Обучение»

Объекты

признаки

	1	2	3	4	5	6	7
1	-0.0683	-0.0676	-0.0643	-0.0617	-0.0550	-0.0488	-0.0365
2	-0.0706	-0.0726	-0.0633	-0.0579	-0.0490	-0.0390	-0.0267
3	-0.0691	-0.0766	-0.0676	-0.0589	-0.0484	-0.0350	-0.0172
4	-0.0631	-0.0719	-0.0662	-0.0557	-0.0441	-0.0317	-0.0144
5	-0.0500	-0.0634	-0.0603	-0.0502	-0.0384	-0.0263	-0.0115
6	-0.0377	-0.0491	-0.0449	-0.0352	-0.0214	-0.0062	0.0056
7	-0.0375	-0.0319	-0.0235	-0.0133	0.0012	0.0169	0.0282
8	-0.0540	-0.0310	-0.0169	6.9177e...	0.0181	0.0339	0.0445
9	-0.0725	-0.0550	-0.0158	0.0145	0.0323	0.0501	0.0641
10	-0.0789	-0.0844	-0.0512	-0.0010	0.0387	0.0612	0.0810
11	-0.0803	-0.0965	-0.0789	-0.0543	-0.0036	0.0405	0.0685
12	-0.0995	-0.1248	-0.1134	-0.0950	-0.0699	-0.0281	0.0261
13	-0.1026	-0.1306	-0.1283	-0.1171	-0.0982	-0.0806	-0.0412
14	-0.0977	-0.1284	-0.1301	-0.1188	-0.1054	-0.0932	-0.0677
15	-0.0840	-0.1133	-0.1138	-0.1040	-0.0895	-0.0750	-0.0511
16	-0.0820	-0.1089	-0.1072	-0.0967	-0.0814	-0.0669	-0.0464
17	-0.0655	-0.0685	-0.0617	-0.0595	-0.0549	-0.0499	-0.0396
18	-0.0640	-0.0734	-0.0653	-0.0566	-0.0510	-0.0440	-0.0330

Целевая переменная

	1
1	-0.0683
2	-0.0706
3	-0.0691
4	-0.0631
5	-0.0500
6	-0.0377
7	-0.0375
8	-0.0540
9	-0.0725
10	-0.0789
11	-0.0803
12	-0.0995
13	-0.1026
14	-0.0977
15	-0.0840
16	-0.0820
17	-0.0655
18	-0.0640

объекты

Настройка параметров модели

НАСТРОЕННАЯ МОДЕЛЬ

Модель - метод, структура, значения параметров

$$\hat{y}_i = f(\vec{p}, x_i)$$

Режим использования

новые объекты

	1	2	3	4	5	6	7	8	9
1	-0.0683	-0.0676	-0.0643	-0.0617	-0.0550	-0.0488	-0.0365	-0.0235	-0.0133
2	-0.0706	-0.0726	-0.0633	-0.0579	-0.0490	-0.0390	-0.0267	-0.0136	-0.0044
3	-0.0691	-0.0766	-0.0676	-0.0589	-0.0484	-0.0350	-0.0172	-0.0038	0.0054
4	-0.0631	-0.0719	-0.0662	-0.0557	-0.0441	-0.0317	-0.0144	-0.0022	0.0084
5	-0.0500	-0.0634	-0.0603	-0.0502	-0.0384	-0.0263	-0.0115	0.0020	0.0152
6	-0.0377	-0.0491	-0.0449	-0.0352	-0.0214	-0.0062	0.0056	0.0145	0.0238
7	-0.0375	-0.0319	-0.0235	-0.0133	0.0012	0.0169	0.0282	0.0355	0.0403

НАСТРОЕННАЯ МОДЕЛЬ

Результат:
значения целевой
переменной

Контролируемое обучение

НЕОБХОДИМЫЕ КОМПОНЕНТЫ

- Обучающий набор данных
- Модель МО (предположение о виде связи между целевой переменной и признаками)
- Способ настройки модели МО (алгоритм обучения)
- Способ сохранения и восстановления структуры и полученных настроек модели
- Способ применения модели МО на новых данных (вычисления целевой переменной)
- Способ оценки качества модели МО

Машинное обучение

Задачи МО

Другие

Контролируемое обучение

$x \in \mathbb{X}$ — объекты / события

$y \in \mathbb{Y}$ — «ответы», «метки»

$\mathcal{F}: \mathbb{X} \rightarrow \mathbb{Y}$ — искомая закономерность

$\mathcal{T}: \{x_i; y_i\}$ — прецеденты

Найти: $\hat{\mathcal{F}}: \{x_i\} \rightarrow \{y_i\}$

Неконтролируемое обучение

$x \in \mathbb{X}$ — объекты / события

$\mathcal{T}: \{x_i\}$ — прецеденты

Найти:

- $\{y_i\}$ — метки кластеров
- $\mathcal{P}(x)$ — плотность вероятности
- $\mathcal{F}: \{x_i\} \rightarrow \{x'_i\}$ — снижение размерности