КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ ИМЕНИ АЛЬ-ФАРАБИ · КАФЕДРА КОМПЬЮТЕРНЫХ НАУК

Научное машинное обучение

Лекция 1. Введение в научное машинное обучение

Физика, данные и вычисления: где встречаются уравнения и нейронные сети

Кенжебек Ержан Ғалымжанұлы

PhD, старший преподаватель кафедры компьютерных наук

Цель и план лекции

Цель

Сформировать представление о научном машинном обучении (SciML) как о сочетании физического моделирования и методов обучения по данным; понять, какие задачи оно решает и какими методами.

План

1. Что такое SciML
2. Ограничения чисто данных-ориентированного ML
3. Три источника смещения (inductive bias)
4. Прямые и обратные задачи
5. Обзор методов: PINN, Neural ODE, операторы, GNN, SINDy
6. Приложения, трудности, инструменты

Что такое научное машинное обучение

Научное машинное обучение (SciML) — область на стыке вычислительной науки, прикладной математики и машинного обучения. Модели строятся так, чтобы учитывать и данные, и известные физические законы.

Физика

Дифференциальные уравнения, законы сохранения, симметрии — знания, накопленные наукой.

Данные

Измерения, наблюдения и результаты расчётов — основа для обучения и проверки модели.

Вычисления

Автоматическое дифференцирование, GPU, дифференцируемое программирование.

Спектр подходов: от уравнений к данным



Чем левее метод, тем больше знаний о системе закладывается в явном виде; чем правее — тем больше информации берётся из данных.

Где чисто данные-ориентированное ML недостаточно

Мало данных

Эксперименты и численные расчёты дороги, поэтому наблюдений обычно немного.

Нет физики

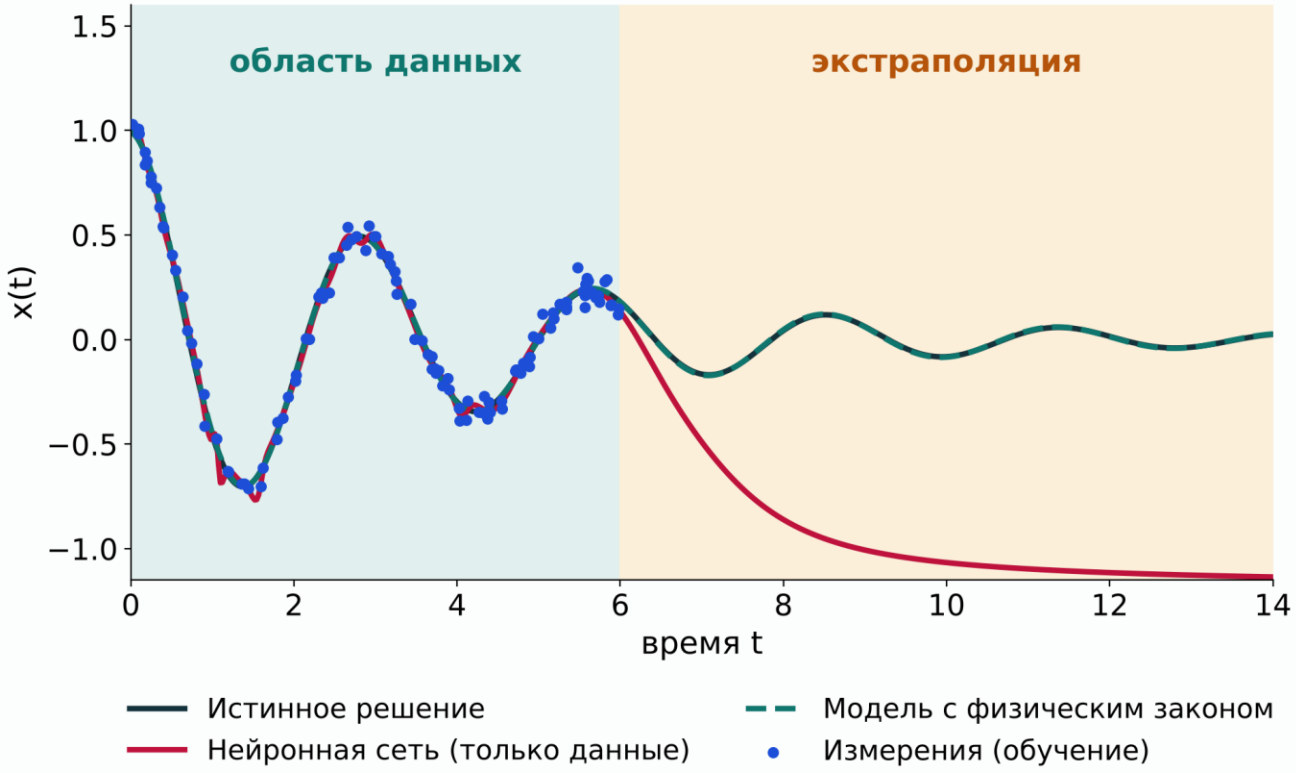
Модель может нарушать законы сохранения и давать физически невозможные ответы.

Плохая экстраполяция

За пределами обучающей области предсказания быстро теряют смысл.

Научные задачи требуют решений, которым можно доверять и которые согласуются с теорией.

Пример: экстраполяция затухающих колебаний



Обе модели хорошо приближают измерения в области обучения (t от 0 до 6).

За её пределами нейросеть без знания закона уходит от истинного решения, а модель с физическим законом совпадает с ним.

Рисунок получен расчётом: затухающий косинус с шумом, MLP и подгонка параметров закона.

Три способа встроить знания в модель

Наблюдательное смещение

Больше и разнообразнее данных: измерения, симуляции, аугментация.

Индуктивное смещение

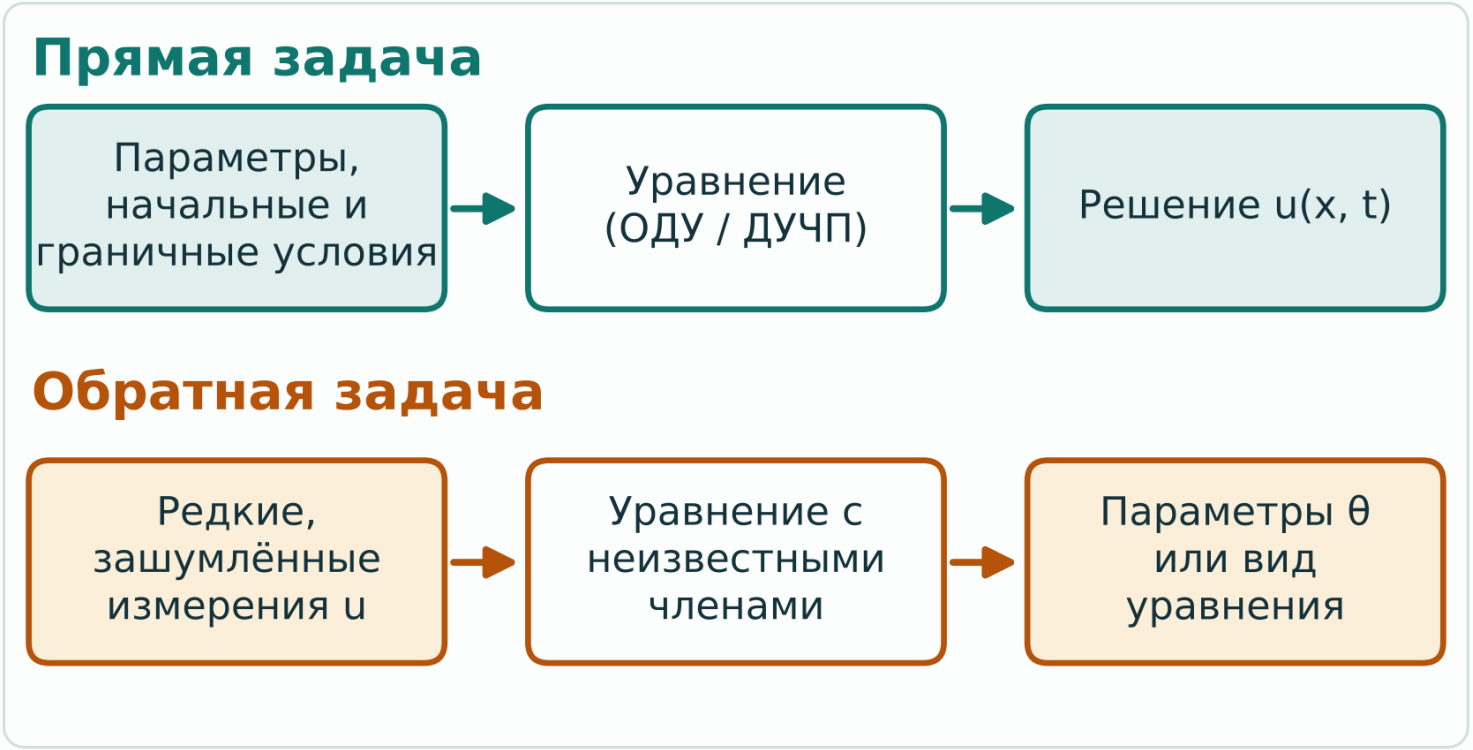
Архитектура уже кодирует симметрии и структуру: свёртки, графы, уравнения внутри сети.

Смещение обучения

Функция потерь содержит физику: невязка уравнения, граничные условия, законы сохранения.

Классификация по работе Дж. Карниадакиса и соавторов «Physics-informed machine learning» (Nature Reviews Physics, 2021).

Прямые и обратные задачи



Прямая: известны уравнение и условия — нужно найти решение.

Обратная: известны измерения — нужно восстановить параметры или сам вид уравнения.

Обратные задачи обычно некорректны: данные редки и зашумлены.

Типовые классы задач

Задача	Что известно	Что ищем	Пример
Решение уравнения	Уравнение, условия	Поле u(x, t)	Теплопроводность, течение жидкости
Идентификация параметров	Модель, измерения	Коэффициенты θ	Вязкость, скорость реакции
Открытие уравнений	Только измерения	Вид уравнения	Динамика популяций, механика
Суррогатное моделирование	Набор расчётов	Быстрая модель	Погода, инженерные расчёты
Оценка неопределённости	Модель, данные	Доверительный интервал	Прогнозы с рисками

Обзор основных методов курса

PINN

Сеть обучается так, чтобы удовлетворять уравнению и данным одновременно.

Neural ODE

Динамика задаётся нейросетью, а решение получается интегрированием ОДУ.

Нейронные операторы

DeepONet и FNO учатся отображению «входная функция → решение».

Графовые сети

Сеть на графах и сетках для частиц, молекул и неструктурированных расчётных областей.

SINDy

Разреженная регрессия по библиотеке функций находит компактное уравнение.

Автоматическое дифференцирование — основа SciML

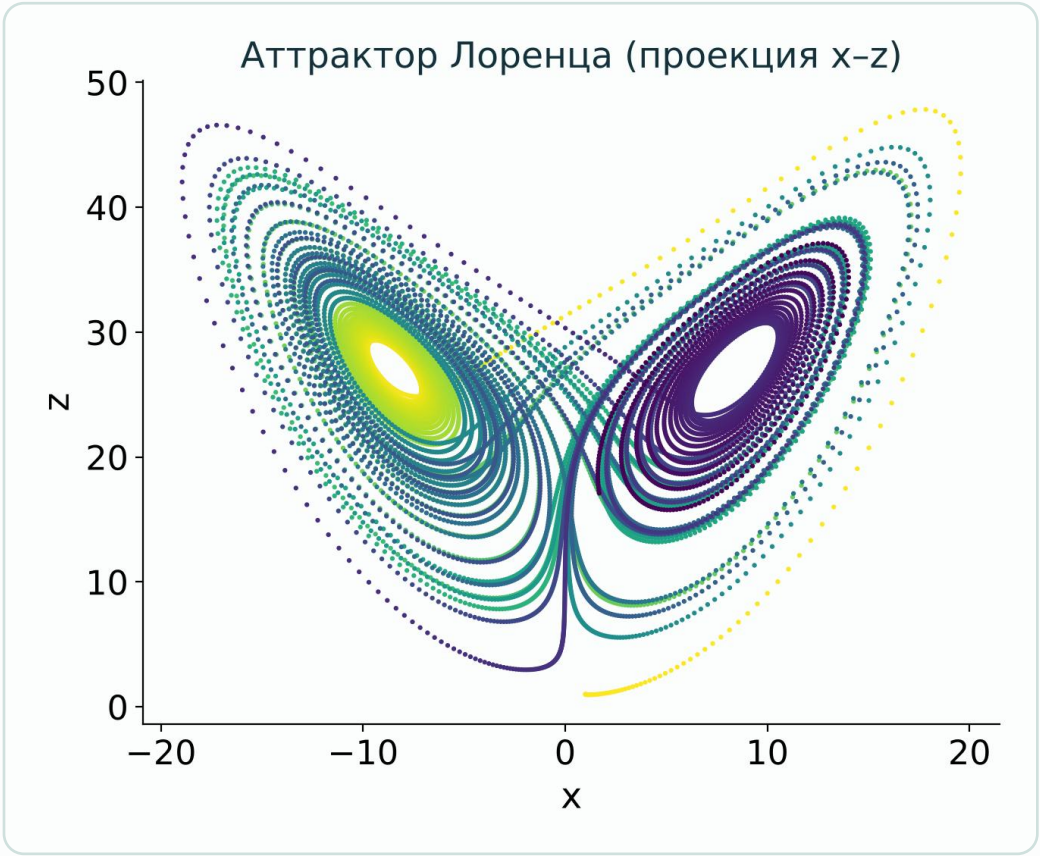
Уравнения содержат производные. Фреймворки (PyTorch, JAX) вычисляют их точно, а не конечными разностями, и позволяют включать в функцию потерь.

```
u = сеть(t)
du_dt = d(u)/dt (autograd)
потеря = среднее (du_dt + k·u)²
```

Дифференцируемость всей цепочки вычислений позволяет обучать модель градиентным спуском.

Тот же механизм используется при обратных задачах: градиент по параметрам θ.

Хаос и пределы предсказуемости



Система Лоренца — простая модель конвекции с тремя переменными. Малое различие в начальных условиях со временем растёт экспоненциально.

Для хаотических систем долгосрочный точный прогноз невозможен ни для численных методов, ни для нейросетей; оценивают статистические свойства и неопределённость.

Рисунок: траектория при σ = 10, ρ = 28, β = 8/3.

Приложения

Погода и климат

Модели GraphCast (Google DeepMind) и FourCastNet (NVIDIA) дают быстрые глобальные прогнозы, обученные на реанализе ERA5.

Материалы и химия

Нейросетевые потенциалы ускоряют молекулярную динамику и поиск материалов.

Инженерия

Суррогатные модели заменяют дорогие расчёты в аэродинамике, энергетике и медицине.

Во всех случаях ценность — ускорение расчётов при контроле физической согласованности.

Трудности и открытые вопросы

- **Жёсткие уравнения** и многомасштабность затрудняют обучение PINN.
- **Обобщение** на новые геометрии, режимы и параметры.
- **Оценка неопределённости:** какой точности можно доверять.
- **Баланс членов** функции потерь: данные и физика конкурируют.
- **Воспроизводимость** и сравнение с классическими численными методами.

Типовой цикл исследования в SciML

1. Постановка

Какую величину предсказываем, какие данные и законы известны.

2. Модель

Выбор метода и способа встраивания физики.

3. Обучение

Функция потерь, оптимизатор, контроль сходимости.

4. Проверка

Сравнение с эталоном, анализ ошибки и неопределённости.

Инструменты курса

Инструмент	Назначение
Python, NumPy, SciPy	Базовые вычисления, численное решение ОДУ
PyTorch, JAX	Нейросети и автоматическое дифференцирование
DeepXDE	Библиотека для PINN и операторных сетей
torchdiffeq	Neural ODE в PyTorch
PySINDy	Идентификация уравнений по данным

Выводы

- SciML сочетает физическое моделирование и обучение по данным.
- Физические знания повышают точность при малых данных и улучшают экстраполяцию.
- Знания встраиваются через данные, архитектуру и функцию потерь.
- Прямые и обратные задачи, суррогатное моделирование и открытие уравнений — основные классы задач.
- Остаются трудности: жёсткость, обобщение, неопределённость.

Контрольные вопросы

1. Что изучает научное машинное обучение и чем оно отличается от «классического» ML?
2. Почему чисто данные-ориентированная модель плохо экстраполирует?
3. Назовите три вида смещения и приведите по примеру.
4. Чем прямая задача отличается от обратной?
5. Какую роль играет автоматическое дифференцирование?
6. Почему хаотические системы ограничивают горизонт прогноза?

Литература

1. Karniadakis G.E. et al. Physics-informed machine learning // Nature Reviews Physics. 2021. Vol. 3. P. 422–440.
2. Raissi M., Perdikaris P., Karniadakis G.E. Physics-informed neural networks // Journal of Computational Physics. 2019. Vol. 378. P. 686–707.
3. Chen R.T.Q. et al. Neural Ordinary Differential Equations // NeurIPS. 2018.
4. Brunton S.L., Proctor J.L., Kutz J.N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems // PNAS. 2016.
5. Lam R. et al. Learning skillful medium-range global weather forecasting // Science. 2023.
6. Brunton S.L., Kutz J.N. Data-Driven Science and Engineering. Cambridge University Press, 2019.