ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

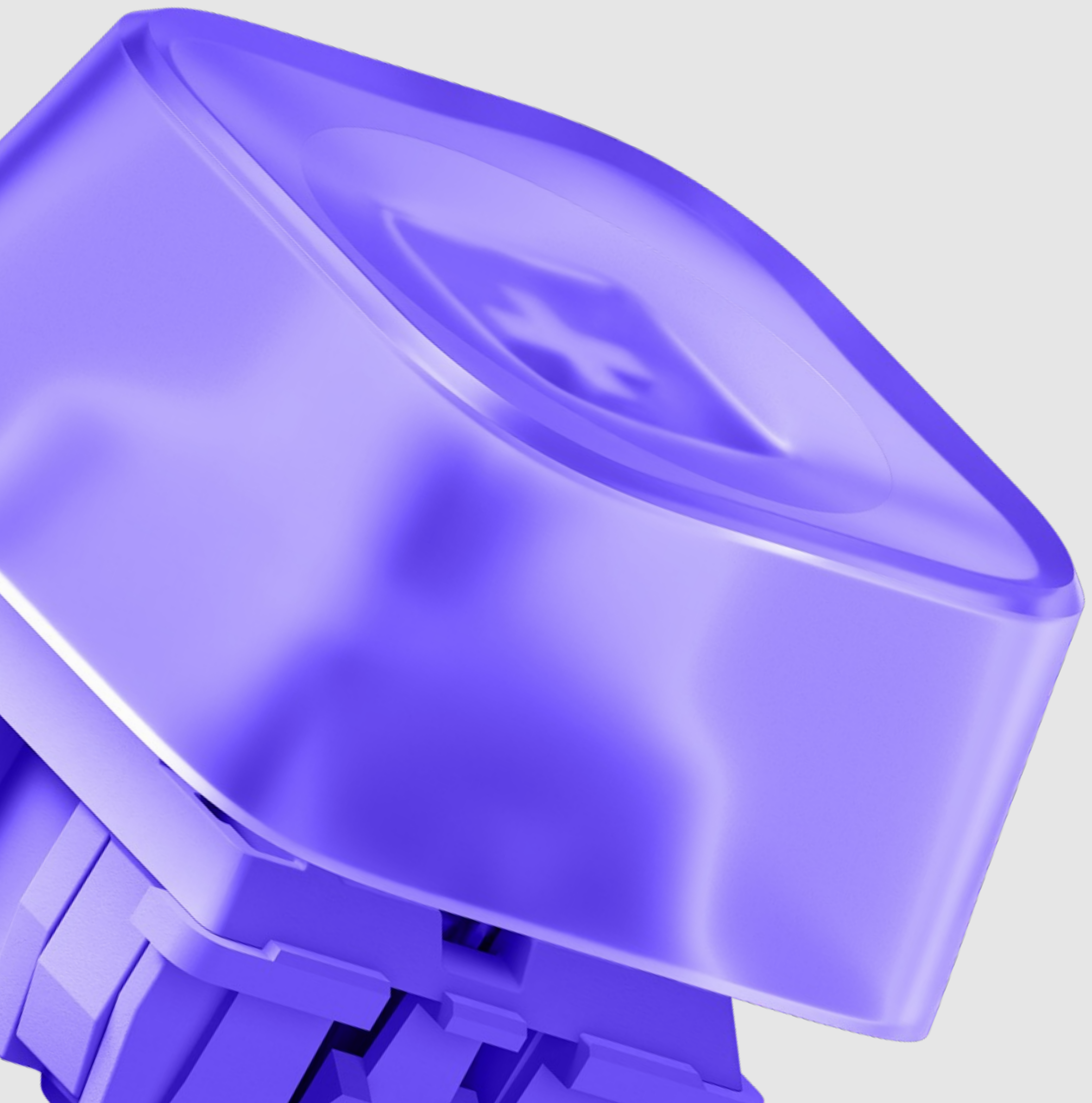
Искусственный интеллект и математические законы

Владимир Спокойный

- Научный руководитель лаборатории теоретических основ моделей искусственного интеллекта ВШЭ
- Руководитель образовательной программы «Математика машинного обучения» ВШЭ & Сколтеха
- Академический консультант Центрального университета

17.04.2026

Наш план на сегодня



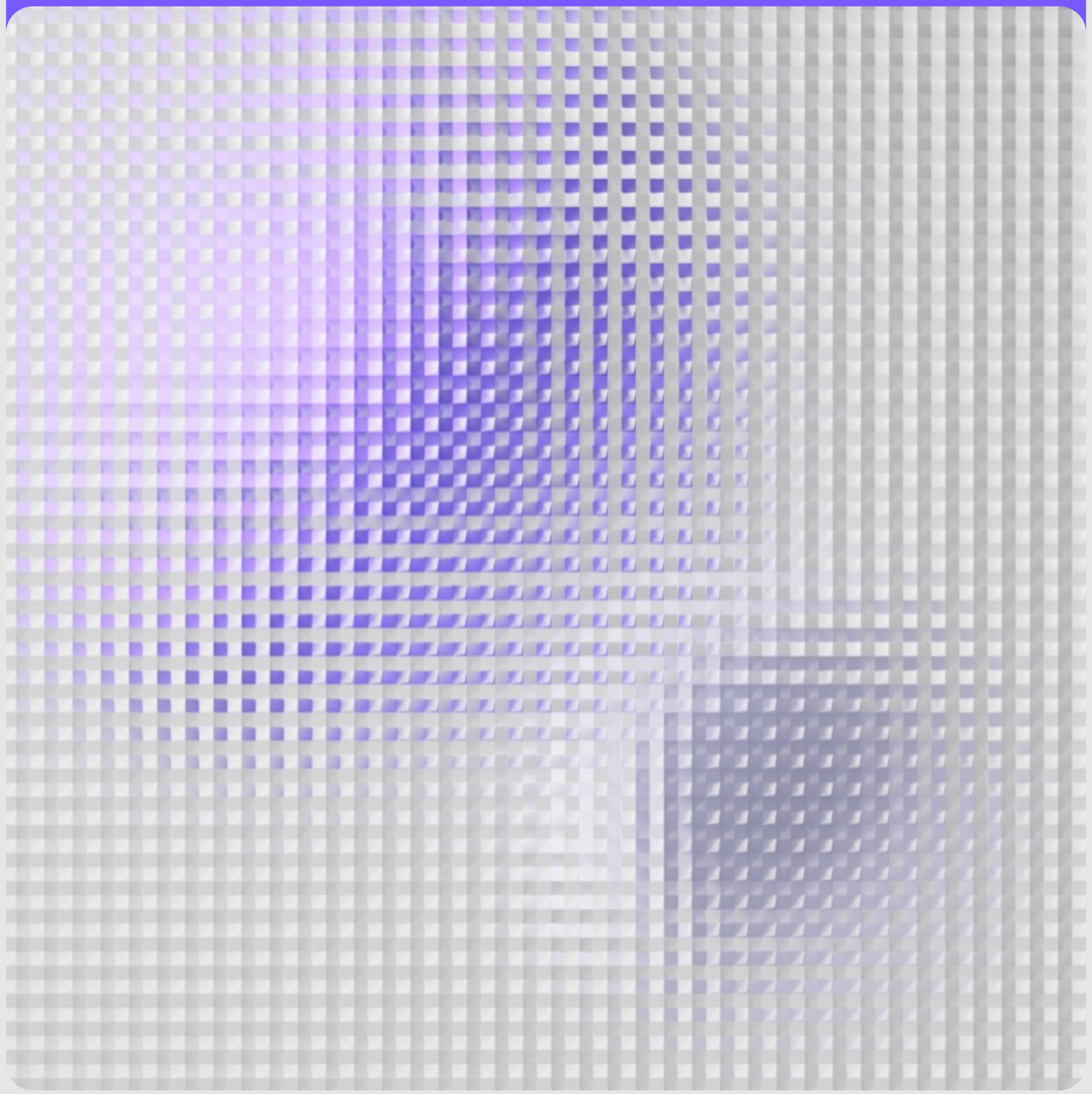
→ Можно ли просто и понятно объяснить, что именно ИИ делает?

→ Подчиняются ли сложные ИИ-модели, такие как глубинные сети, генеративные и большие языковые модели, математическим законам?

→ Если да, можно ли просто их сформулировать, объяснить, проверить?

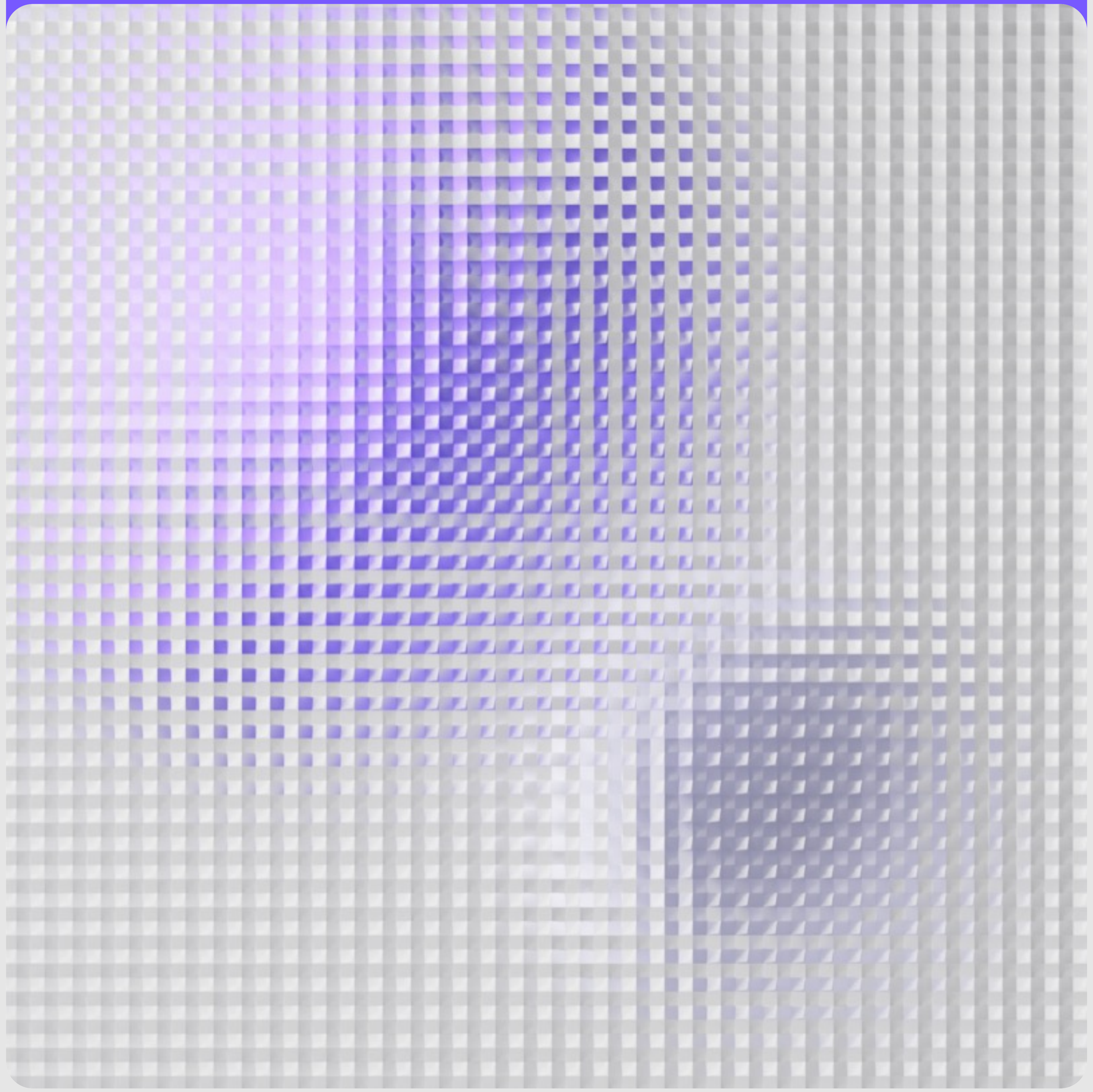
→ Могут ли эти законы помочь в понимании пределов, сферы применимости, надежности, уровня доверия и других свойств ИИ?

Что есть ИИ?



Что есть ИИ на самом деле?

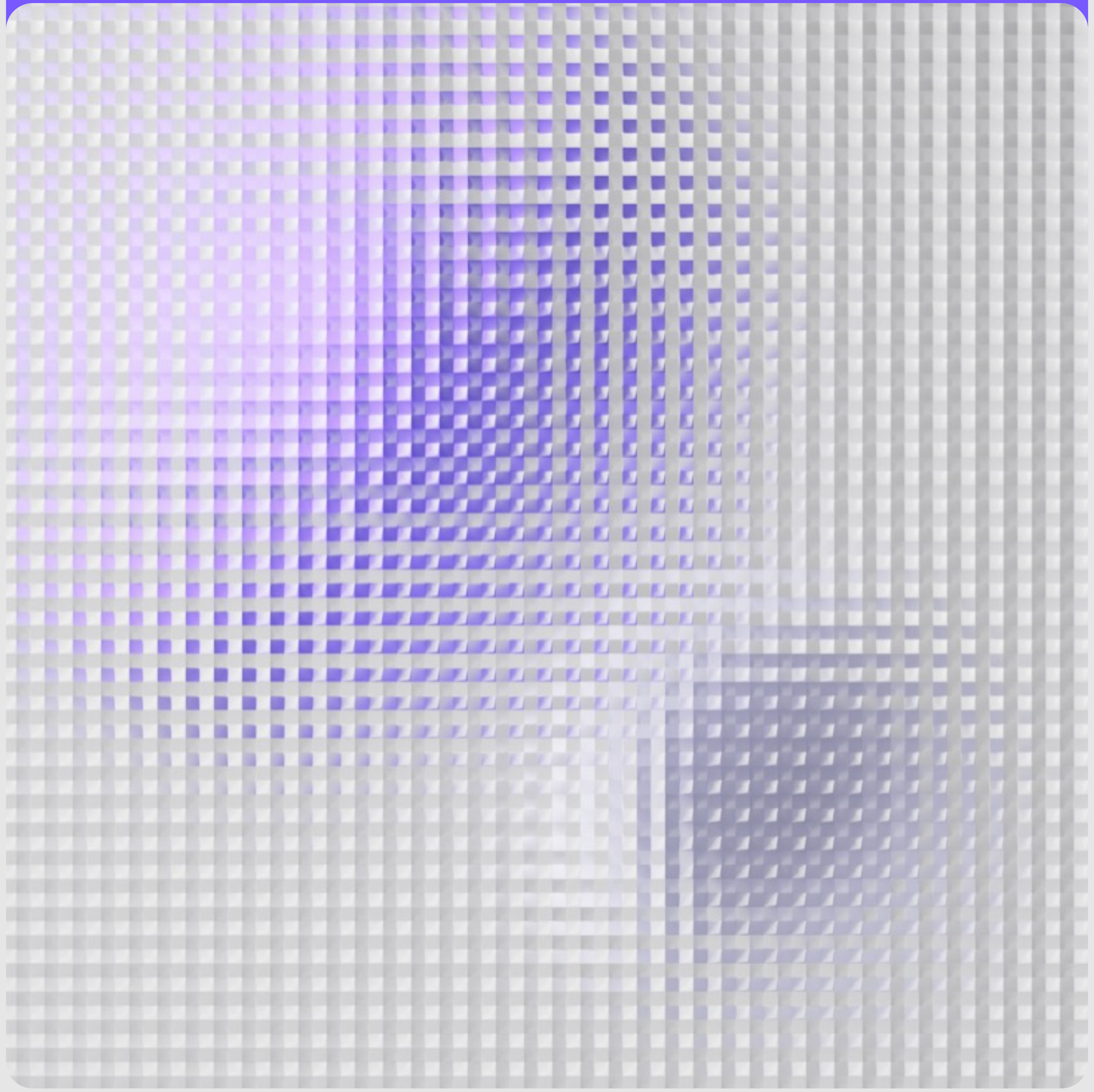
Что есть ИИ?



Что есть ИИ на самом деле?

Очень упрощенный ответ

Что есть ИИ?

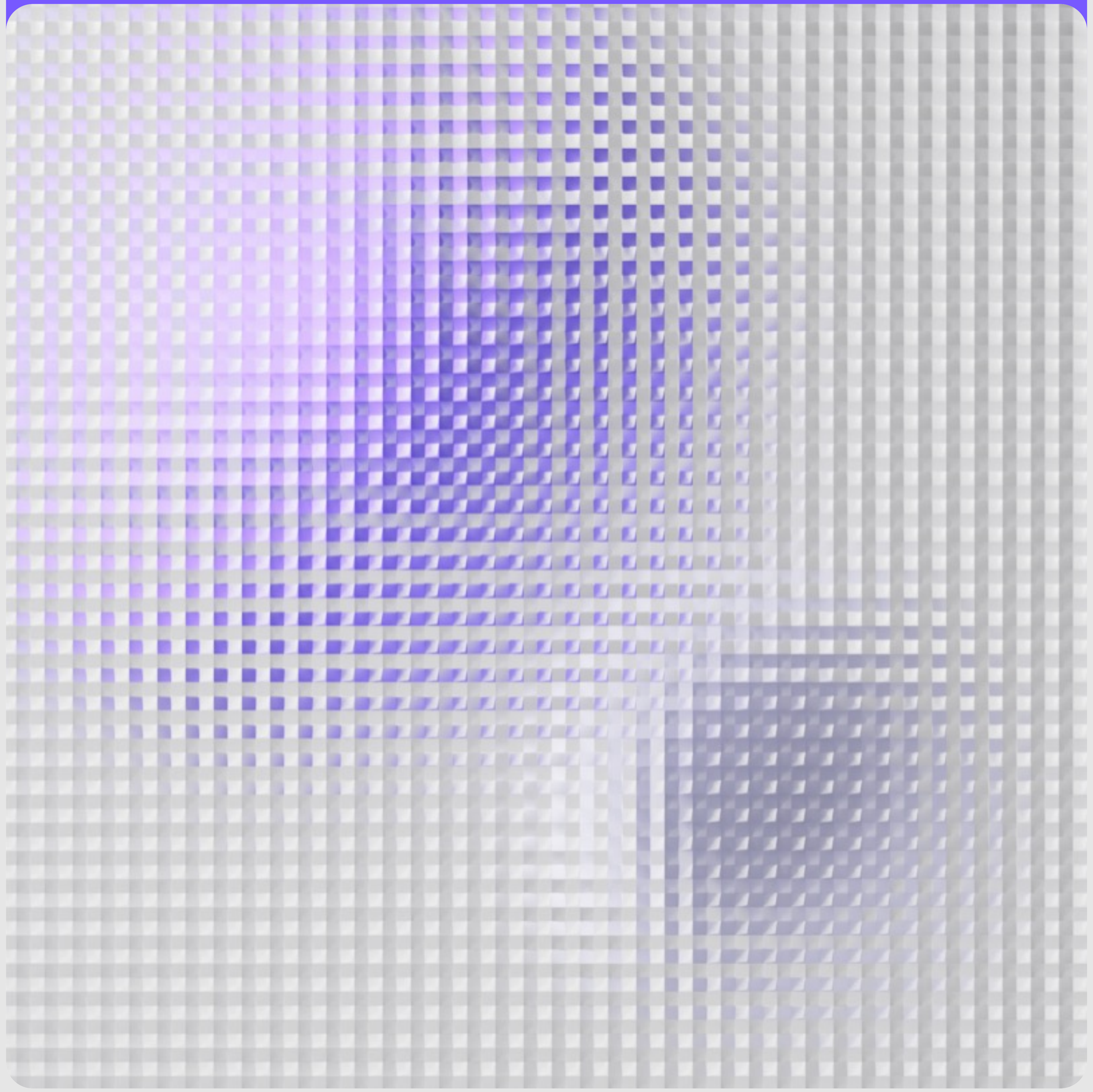


Что есть ИИ на самом деле?

Очень упрощенный ответ

ИИ — это компьютерная программа, располагающая большой вычислительной мощностью, гигантской памятью и доступом к большим массивам данных

Что есть ИИ?



Что есть ИИ на самом деле?

Очень упрощенный ответ

ИИ — это компьютерная программа, располагающая большой вычислительной мощностью, гигантской памятью и доступом к большим массивам данных

Что ИИ делает?

Что есть ИИ?

Что есть ИИ на самом деле?

Очень упрощенный ответ

ИИ — это компьютерная программа, располагающая большой вычислительной мощностью, гигантской памятью и доступом к большим массивам данных

Что ИИ делает?

Упрощенный ответ

Что есть ИИ?

Что есть ИИ на самом деле?

Очень упрощенный ответ

ИИ — это компьютерная программа, располагающая большой вычислительной мощностью, гигантской памятью и доступом к большим массивам данных

Что ИИ делает?

Упрощенный ответ

Любое действие ИИ — это решение определенной задачи по оптимизации, специально сформулированной по входным данным

Ученые, внесшие основной вклад, и методы ИИ

Ученые, внесшие основной вклад, и методы ИИ

Середина XVIII века

Ньютон и Лейбниц

Общие методы оптимизации

Методы первого и второго порядка

Начало XIX века

Гаусс и Лежандр

Метод наименьших квадратов

Около 1920

Фишер

Метод максимума правдоподобия как универсальное решение задачи по тренировке модели

Метод максимума правдоподобия

Пусть $L(v)$ — случайная функция (потери, эмпирический риск, функция правдоподобия, ...). Рассмотрим:

$$\underbrace{\tilde{v} = \operatorname*{argmin}_{v} L(v)}_{trained} \; ; \qquad \underbrace{v^* = \operatorname*{argmin}_{v} \mathbb{E}L(v)}_{truth} \; ;$$

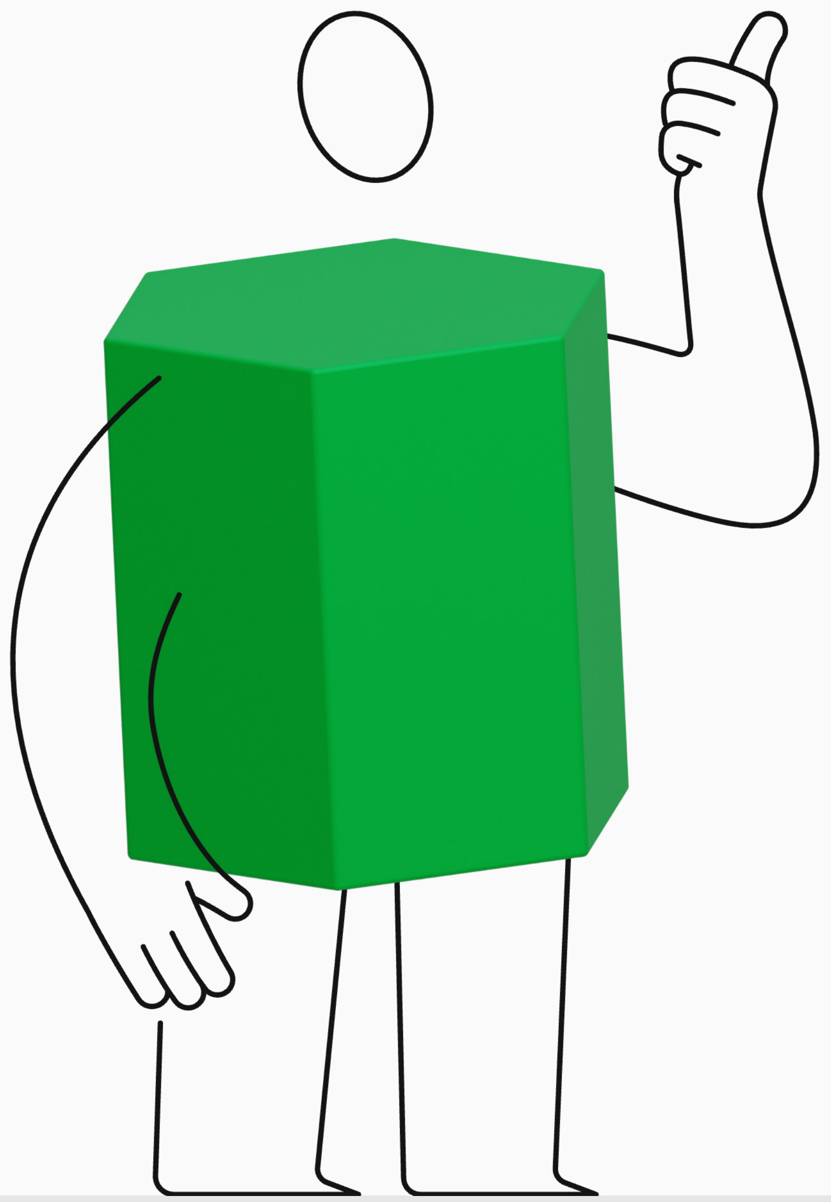
Этот подход включает основные процедуры трейнинга:
MLE, LSE, LAD, Minimum Contrast, ...

Математическая проблема: описать $\tilde{v} - v^*$(ошибка трейнинга) и $L(\tilde{v}) - L(v^*)$ (эксцесс).

Теория возмущенной оптимизации: $L(v)$ — это возмущение $\mathrm{f}(v) = \mathbb{E}L(v)$ случайной компонентой

$$\zeta(v) = L(v) - \mathbb{E}L(v)$$

Особенности современной теории обучения



Большие объемы данных

01

Очень сложные модели
с огромным числом параметров

02

Важна вычислимость решений (за короткое время) и скалируемость методов.
Время вычислений растет линейно с размерностью

03

Классическая асимптотика (Фишер)

Если размерность модели (число параметров p) фиксированна,
а объем выборки растет к бесконечности ($n \to \infty$), то для ОМП $\tilde{v}$ верно:

$$\sqrt{n}(\tilde{v} - v) \xrightarrow{w} \mathcal{N}(0, F_v),$$

См. любой учебник по статистике, например:

— Боровков Л. Л. Математическая статистика. 1984.

— Lehmann and Romano. Testing statistical hypotheses. 2006.

— Van der Vaart. Asymptotic statistics. 2000.

Современные приложения: новые феномены

- Огромная размерность моделей $p \gg n$, проблема переобучаемости, необходимость регуляризации

  [Cheng and Montanari, 2022]

- Случайный план и плохо обусловленная матрица Фишера

  [Bartlett et al., 2020, Montanari et al., 2025, Kuchelmeister and van de Geer, 2024]

- Смещение, вызванное случайным планом и высокой размерностью задачи. Теория Фишера неприменима

  [Sur and Candès, 2019, Candès and Sur, 2020]

- Требуются новые подходы, позволяющие получить гарантии точности для конечных объемов выборок и высокой размерности модели

  [Cheng and Montanari, 2022]

Новые феномены

в простейшей линейной модели

- Феномен «двойного спуска» (double decent).
  Фазовый переход в переобучаемости

  [Bach, 2024]

- Преодоление переобучаемости (benighing overfitting).
  Перепараметризация может быть даже полезной.
  Оптимальная обучаемость и отличное качество предсказания

  [Bartlett et al., 2020, Montanari et al., 2025,
  Kuchelmeister and van de Geer, 2024]

- Методы оптимизации второго порядка неприменимы
  в практических постановках с плохо обусловленным Гессианом.
  Методы первого порядка + «ранняя остановка» хорошо работают

  [Wu et al., 2025].

  Методы нулевого порядка (без использования производных)
  все более популярны

Математические дисциплины

Теоретические выводы даже для простейших моделей в неоклассической постановке требуют развития математических дисциплин, таких как:

- Возмущенная оптимизация
- Теория случайных матриц
- Неравенства больших уклонений для случайных тензоров
- Методы регуляризации для нелинейных обратных задач

Некоторые законы в современном звучании

Закон Фишера $p \ll n$ (для обучения нужно много наблюдений на каждый параметр).

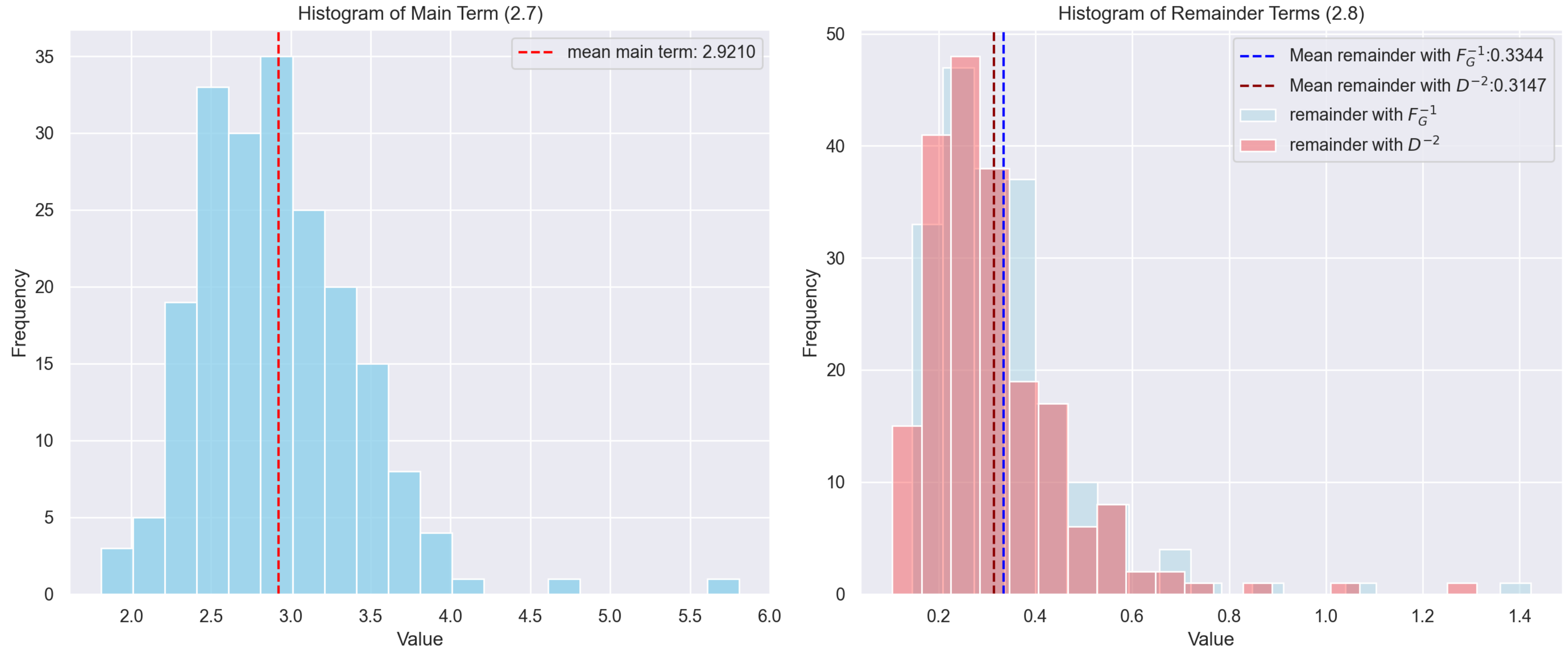
Уточненная версия закона Фишера: важны эффективная размерность $\mathbb{p}$ Место для уравнения.и эффективный объем выборки $\mathbb{N}$ matter.

Нео-Фишер: $\mathbb{p} \ll \mathbb{N}$.

Эффективная размерность $\mathbb{p}$ может контролироваться путем подбора пенализации или методами редукции модели.

Эффективный объем выборки $\mathbb{N}$ определяется нормой и обусловленностью матрицы Фишера $\mathbb{F}$. Неформально это объем имеющейся информации.

Distribution of the leading term and the remainder

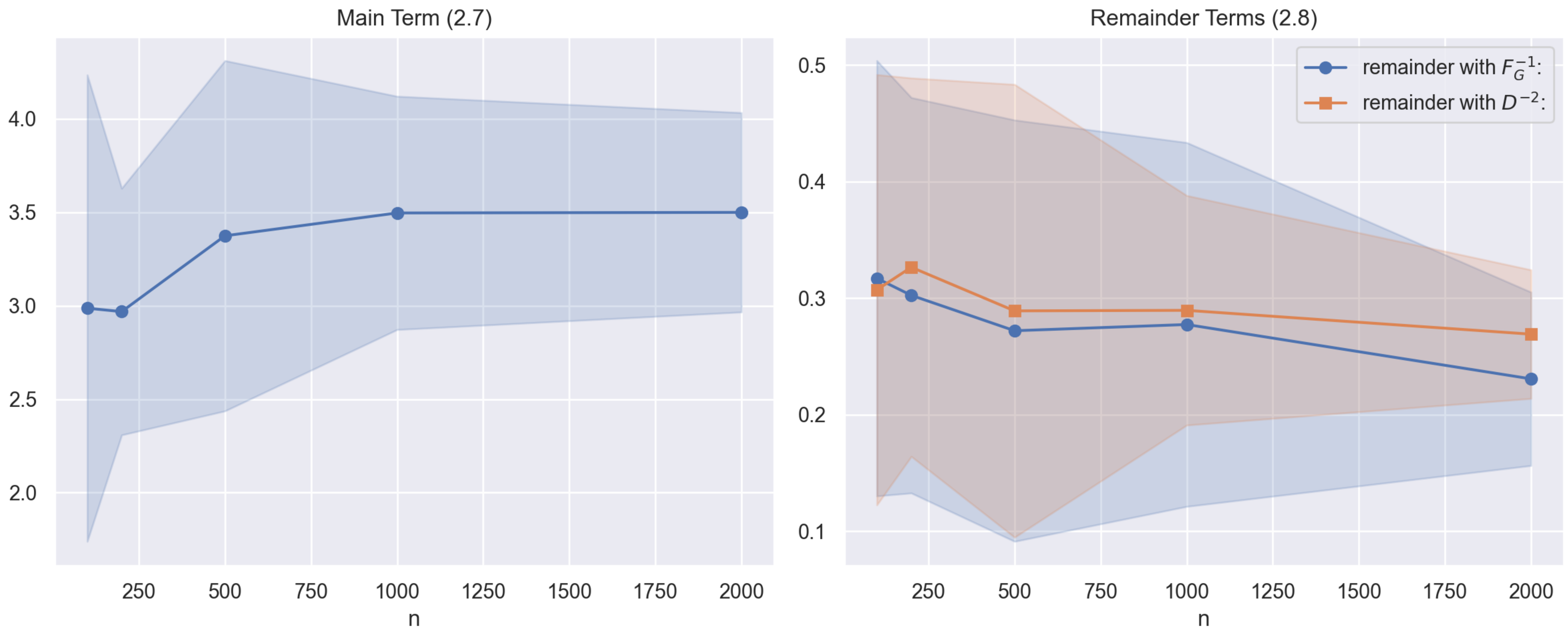


Distribution of the leading term and the remainder for $n = 100$.

Left: ошибка тренировки $\left\|\mathbb{F}_G^{-1}\nabla\zeta\right\|_\infty$ и $\|D^{-2}\nabla\zeta\|_\infty$. Right: ошибка после коррекции Фишера $\|\tilde{v} - v^* + \mathbb{F}^{-1}\nabla\zeta\|_\infty$ и $\|\tilde{v} - v^* + D^{-2}\nabla\zeta\|_\infty$ для $n = 100$.

Ошибка справа примерно в 10 раз меньше ошибки слева: иллюстрирует качество работы математического закона.

Результаты для n $n \in \{100, 200, 500, 1000, 2000\}$



Comparison of the leading term and the remainder for different n.

DNN

Данные (Y_i, X_i), где Y_i — отклик (метка) и регрессор $X_i \in \mathbb{R}^d$.

Цель: построить сеть с малой ошибкой предсказания $Y_i \ - m(X_i)$.

Двухуровневая (Shallow) сеть: $\mathbb{X} = \sigma(WX)$.

Глубинная сеть (DNN): $\mathbb{X} = \ \mathbb{X}^{(K)} = DNN(X) \in \ \mathbb{R}^M$

$$DNN(X) = \sigma\left(W^{(K)}\sigma\left(W^{(K-1)} \ldots \sigma\left(W^{(1)}X\right)\right)\right),$$

где $M = M^{(K)}$ — число узлов в последнем слое, $W^{(k)} = \left(w_{mj}\right) - \ M_k{\times}M_{k-1}$ — матрица весов со строками $W_m^{(k)}$, σ — функция активации.

Предсказание на основе линейной регрессии на выход последнего слоя $\mathbb{X}$:

$$m(X_i) = \mathbb{X}_i^T \beta = \underbrace{\sigma(WX_i)^T \beta}_{\text{shallow}}$$

Обучение глубинных сетей

Двухуровневая (shallow) модель: $DNN(X_i|W) = \sigma(WX_i)$.

Глубинная (deep) модель: $W = \left(W^{(1)}, \dots, W^{(K)}\right)$.

$$DNN(X|W) = \sigma\left(W^{(K)}\sigma\left(W^{(K-1)} \dots \sigma\left(W^{(1)}X\right)\right)\right),$$

Метод обучения (ОМП):

$$m(X_i|W,\beta) = DNN(X_i|W)^T\beta$$

$$L(W,\beta) = \sum_{i=1}^{n} |Y_i - m(X_i|W,\beta)|^2$$

$$\left(\widehat{W}, \hat{\beta}\right) = \operatorname*{argmin}_{W,\beta} L(W,\beta)$$

Сложность: задача невыпуклая, неидентифицируемая, перепараметризованная.

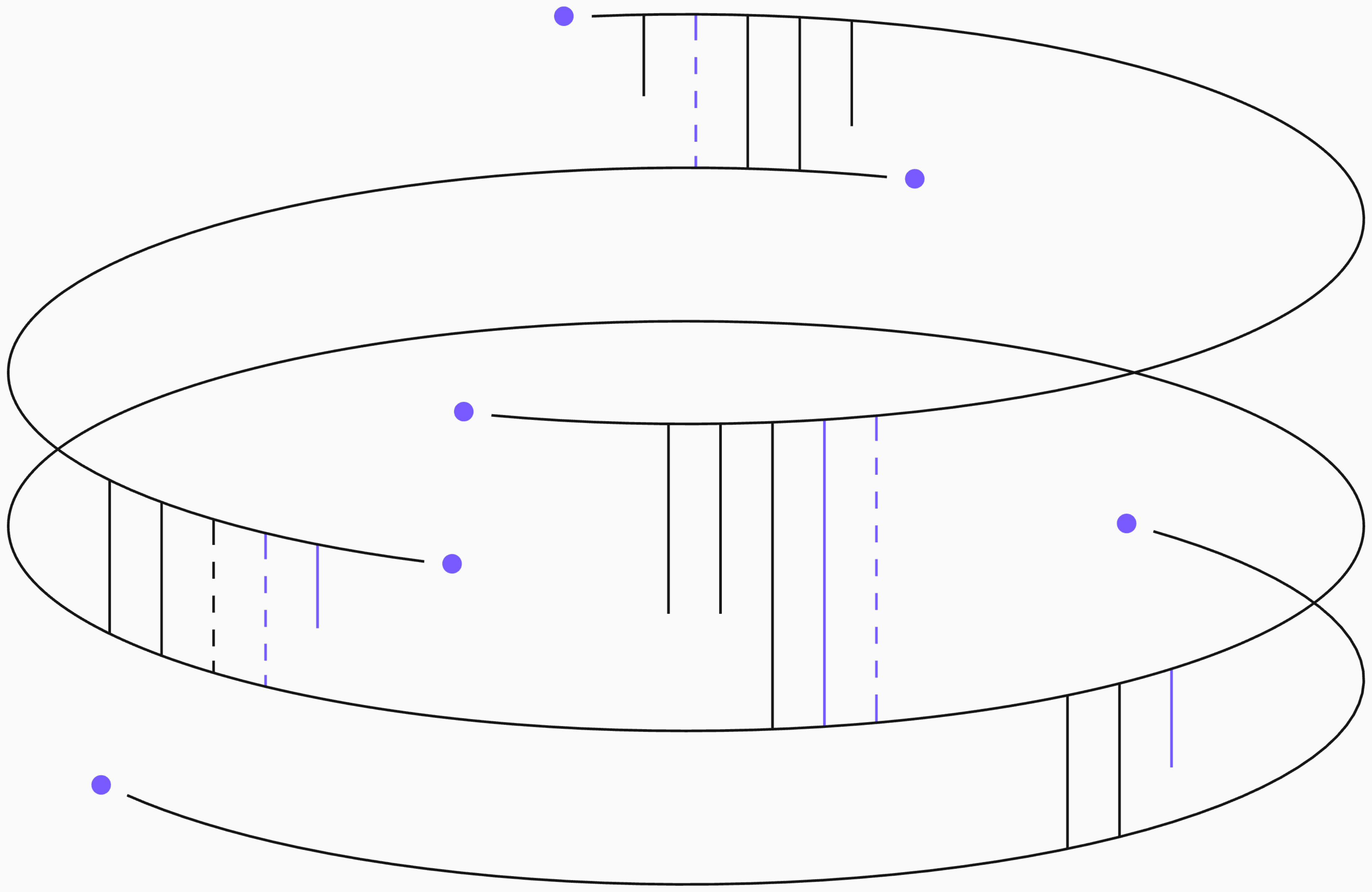
Some statistical literature on DNN regression

- [Schmidt-Hieber, 2020]. Nonparametric regression using deep neural networks with ReLU activation function, Kolmogorov-Arnold representation.
- [Zuowei Shen et al., 2020], Deep Network Approximation Characterized by Number of Neurons.
- [Fan et al., 2024], How do noise tails impact on deep ReLU networks? polynomial noise, tails, Huber, DNN approximation.
- [Kohler and Krzyzak, 2022], Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent.
- [Shen et al., 2022], Approximation with CNNs in Sobolev Space: with Applications to Classification.
- [Liu et al., 2022], Benefits of Overparameterized Convolutional Residual Networks: Function Approximation under Smoothness Constraint.
- [Simionescu-Badea, 2022], Analysis of the rate of convergence of fully connected deep neural network regression estimates with smooth activation.

Manifold regression

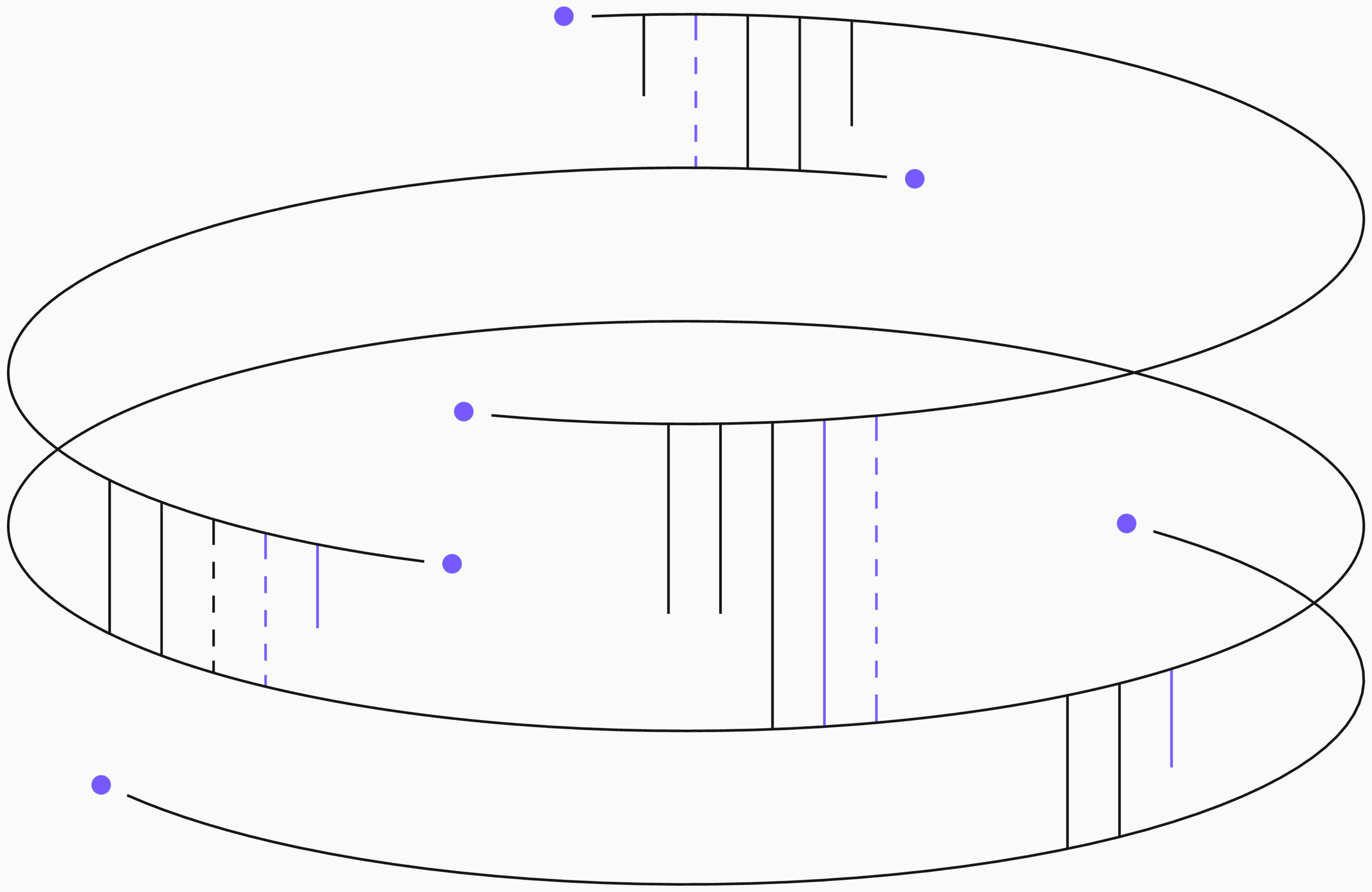
- [Chen et al., 2019], Efficient Approximation of Deep ReLU Networks for Functions on Low Dimensional Manifolds.
- [Kohler et al., 2019], Estimation of a Function of Low Local Dimensionality by DNN.
- [Jiao et al., 2021], Deep Nonparametric Regression on Approximately Low-dim Manifolds.
- [Liu et al., 2021], Besov Function Approximation …on Low-Dim Manifolds …
- [Cloninger and Klock, 2021], A deep network construction that adapts to intrinsic dimensionality beyond the domain.
- [Zhang et al., 2023], Effective Minkowski Dimension of Deep Nonparametric Regression: Function Approximation and Statistical Theories.
- [Shen et al., 2023], RePU DNN, manifolds, Differentiable Neural Networks with RePU Activation: with Applications…
- [Jiao et al., 2023], Deep nonparametric regression on approximate manifolds.

Проклятие размерности



Размерность?

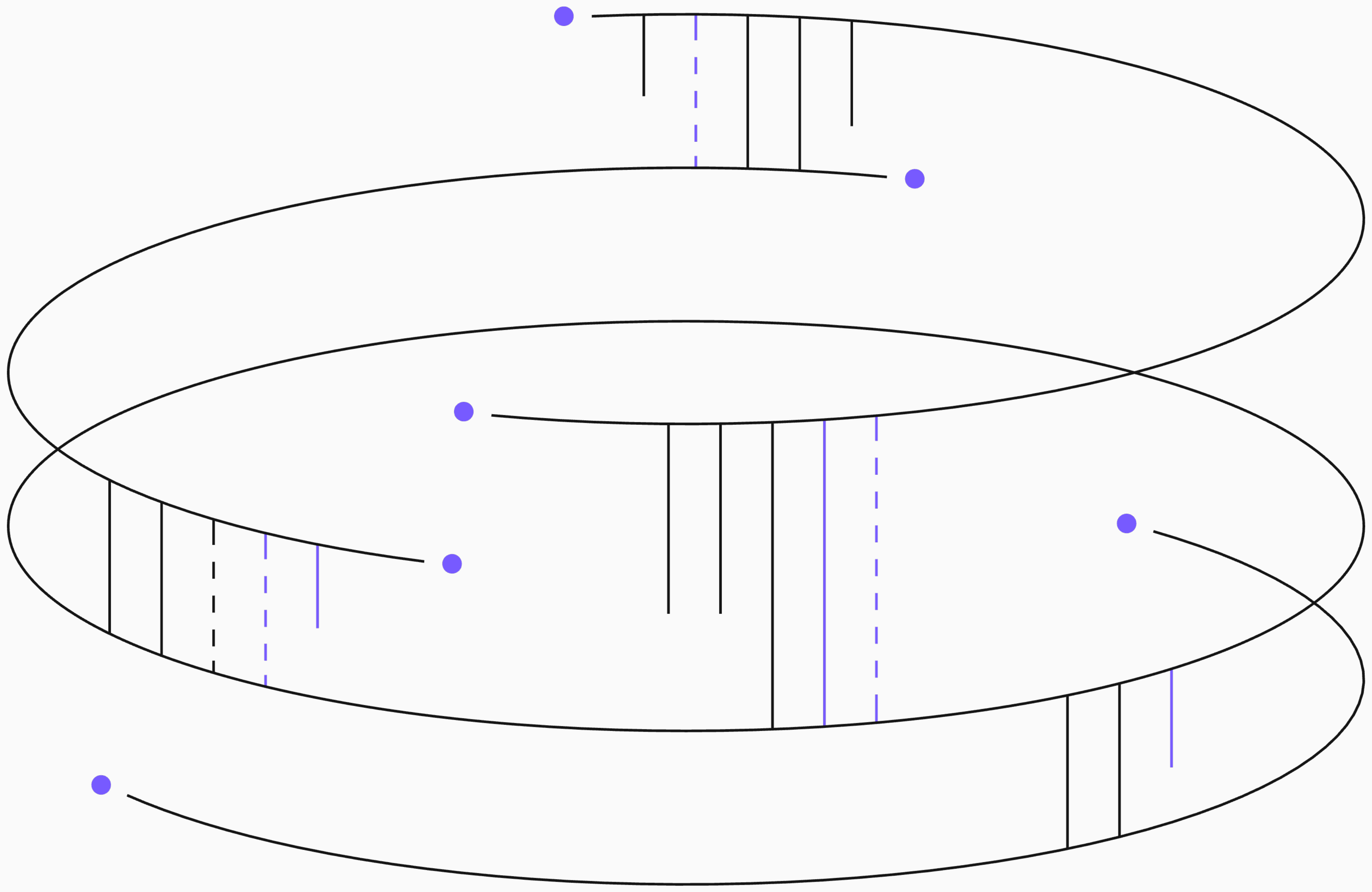
Проклятие размерности



Размерность?

Число регрессоров d?

Проклятие размерности

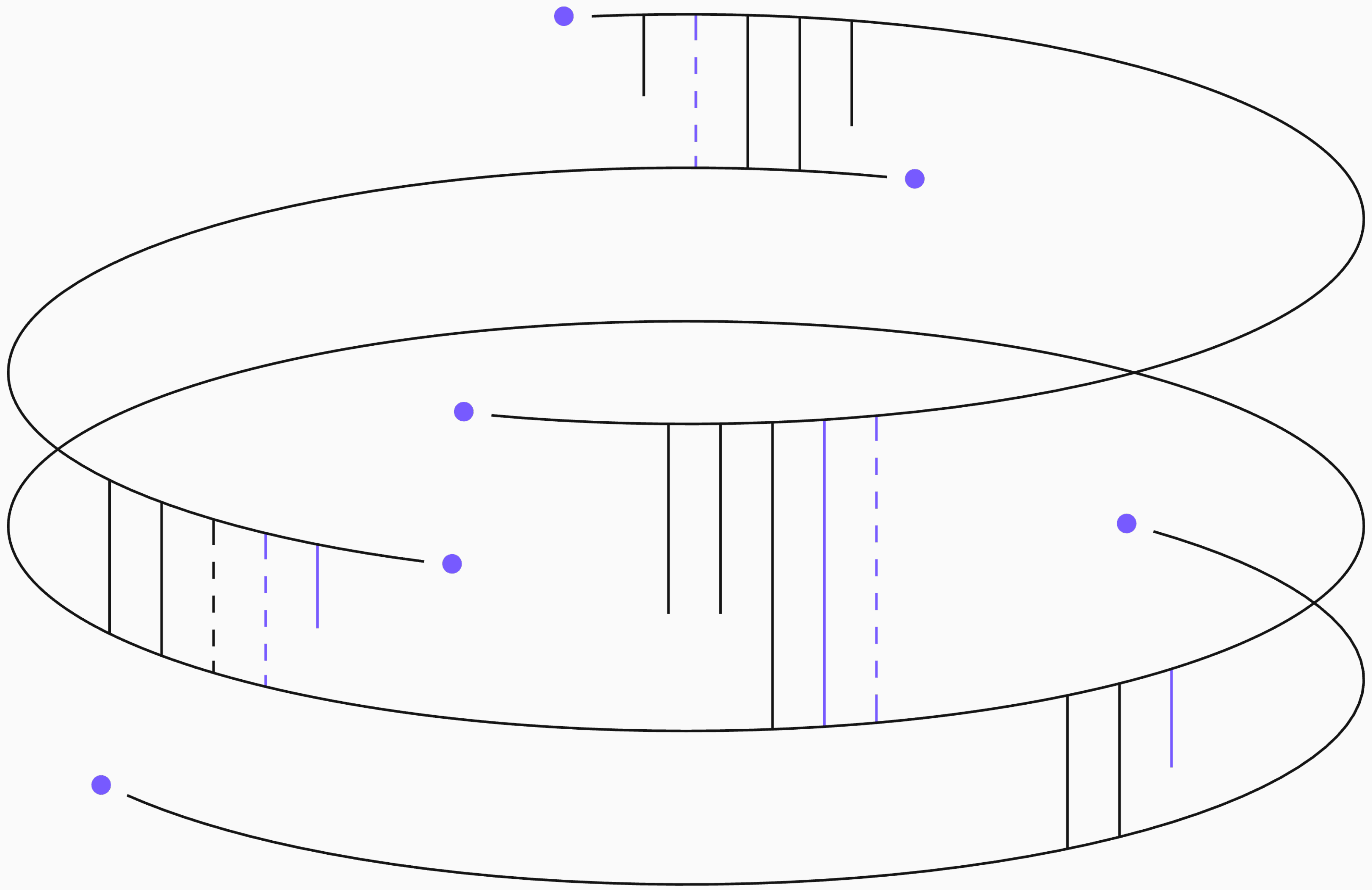


Размерность?

Число регрессоров d?

Или

Проклятие размерности



Размерность?

Число регрессоров d?

Или

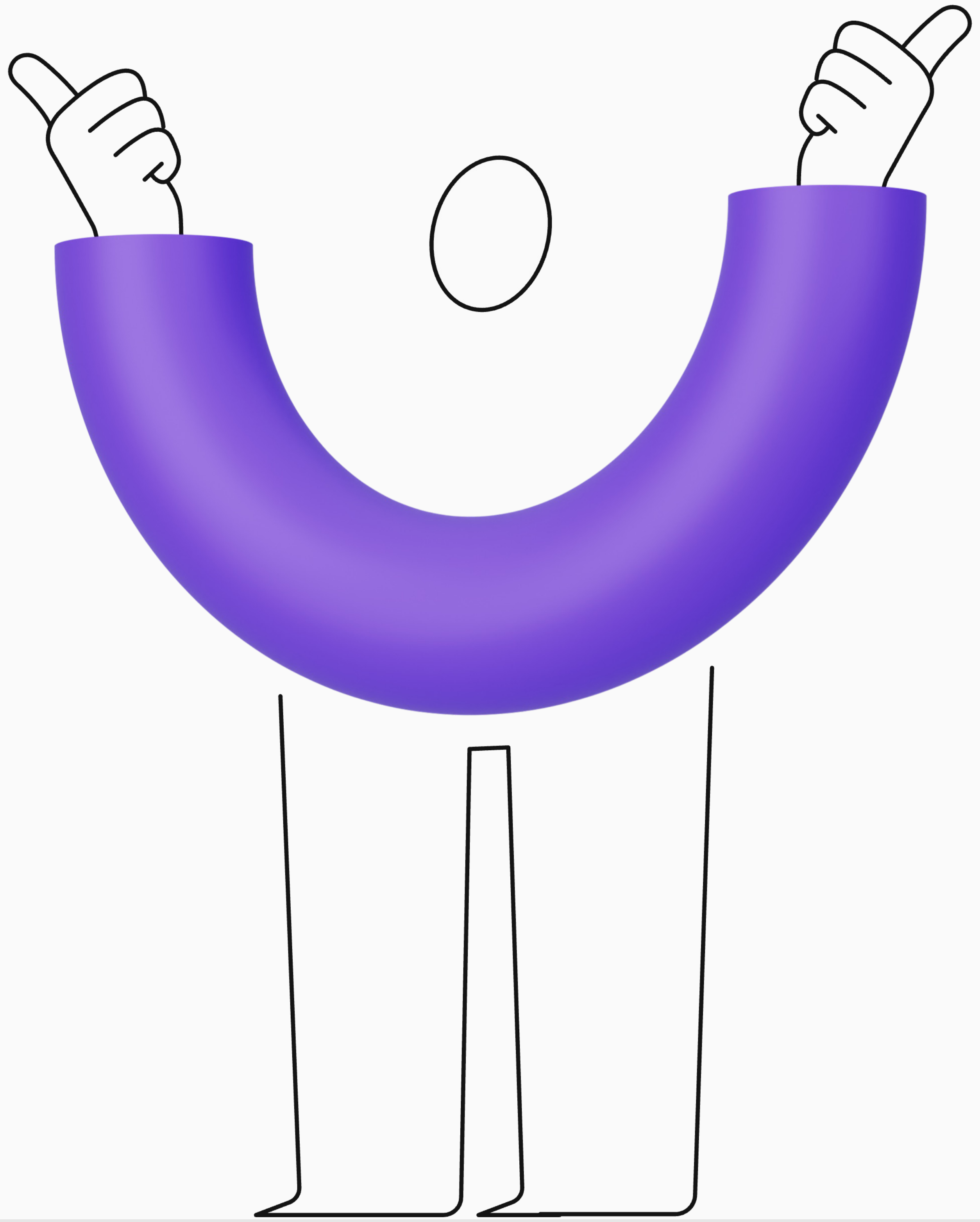
Число параметров p модели?

Совпадение для линейных моделей

Благословение размерности: идея

Число параметров p не важно
и не входит в оценки точности

А что важно?

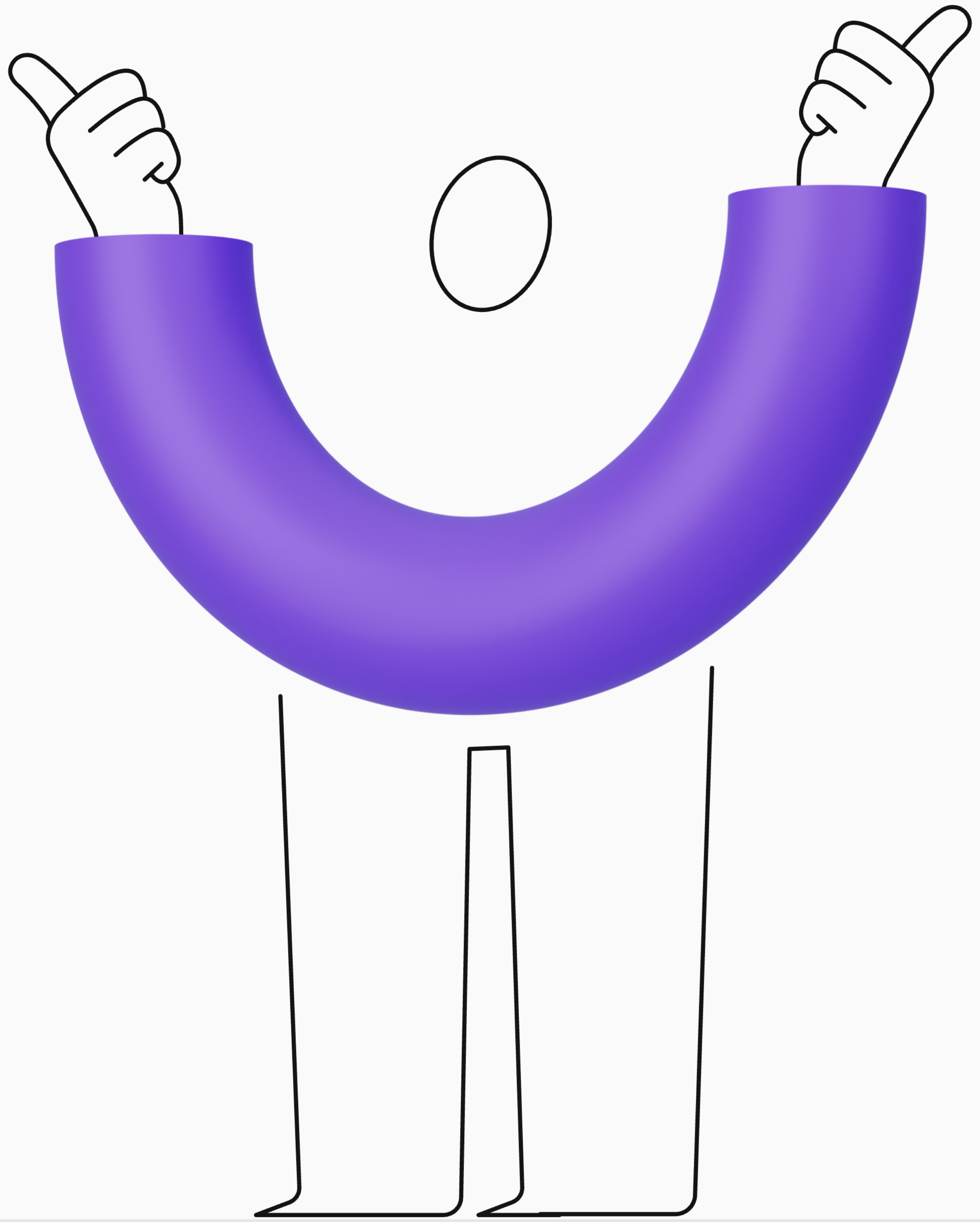


Благословение размерности: идея

Число параметров p не важно
и не входит в оценки точности

А что важно?

- Эффективная размерность $\mathbb{p}$
- Геометрические свойства (такие как строгая выпуклость) целевой функции



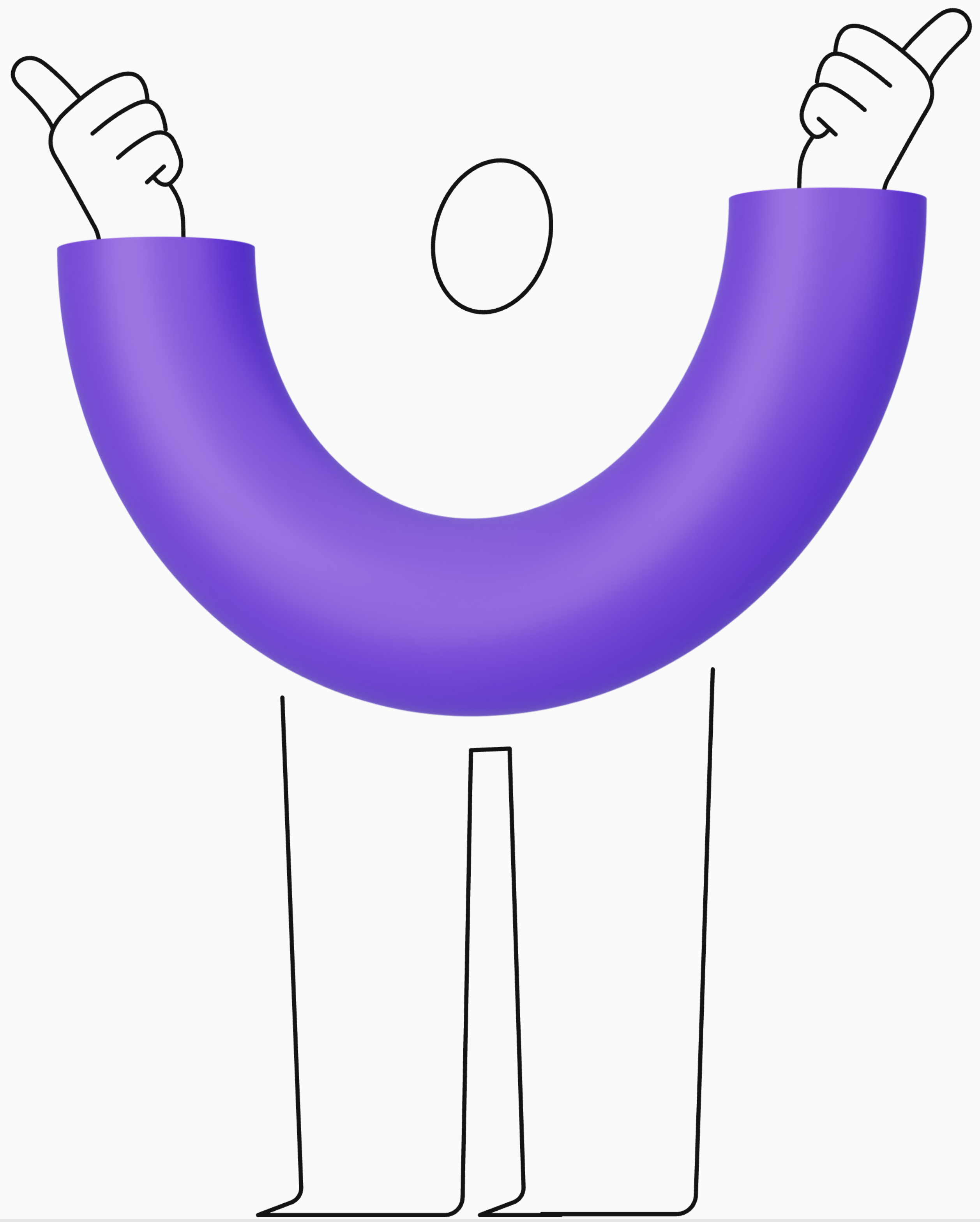
Благословение размерности: идея

Число параметров p не важно
и не входит в оценки точности

А что важно?

- Эффективная размерность $\mathbb{p}$
- Геометрические свойства (такие как строгая выпуклость) целевой функции

Подходящее расширение модели может быть полезным
и может улучшить геометрические свойства целевой функции
без значительного увеличения эффективной размерности



Снижение размерности многомерных распределений данных

Высокая размерность — проклятие для моделей данных.

Например, для генеративных моделей нужно семплировать из таких распределений.

Наблюдение 1

Любое многомерное облако данных выглядит как шарик (из нормального распределения). Этот нетривиальный факт может быть обоснован применением центральной предельной теоремы в асимптотике растущей Размерности.

Наблюдение 2

Нормальное распределение неинформативно и отражает многомерный «шум».

Наблюдение 3

Информативная негауссовская компонента распределения сосредоточена в подпространстве малой размерности.

Идея снижения размерности:

восстановить по данным информативное подпространство и использовать для моделирования (предсказания).

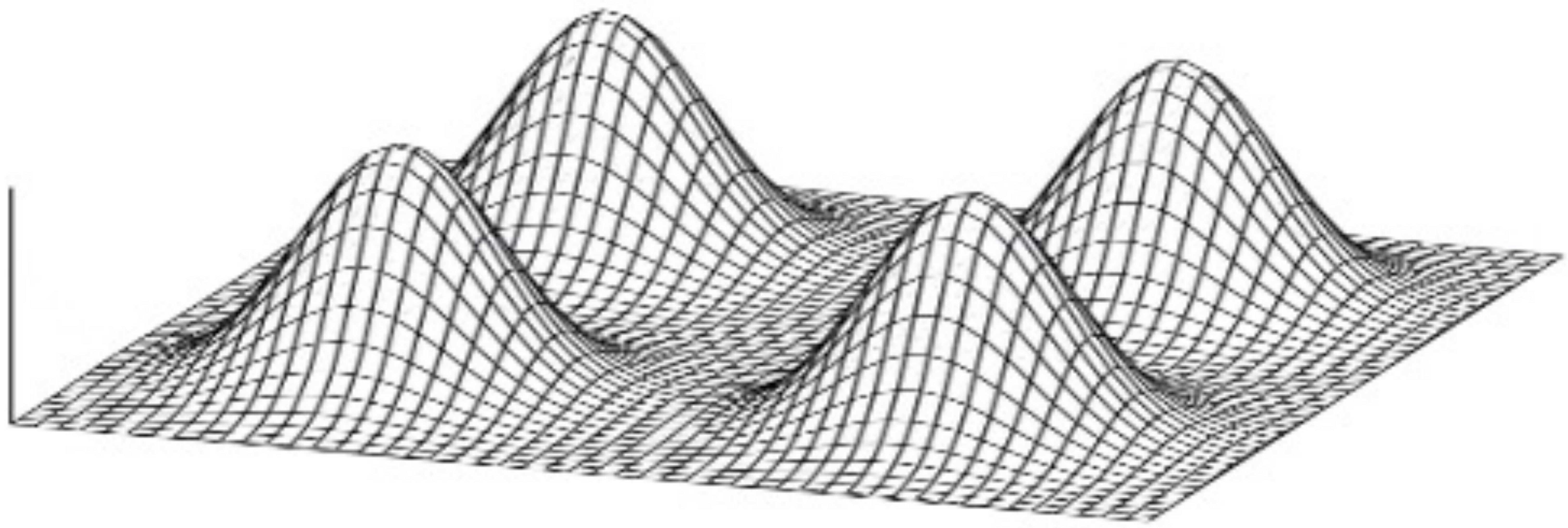
Моделирование многомерных распределений

Информативная модель должна быть мультимодальной

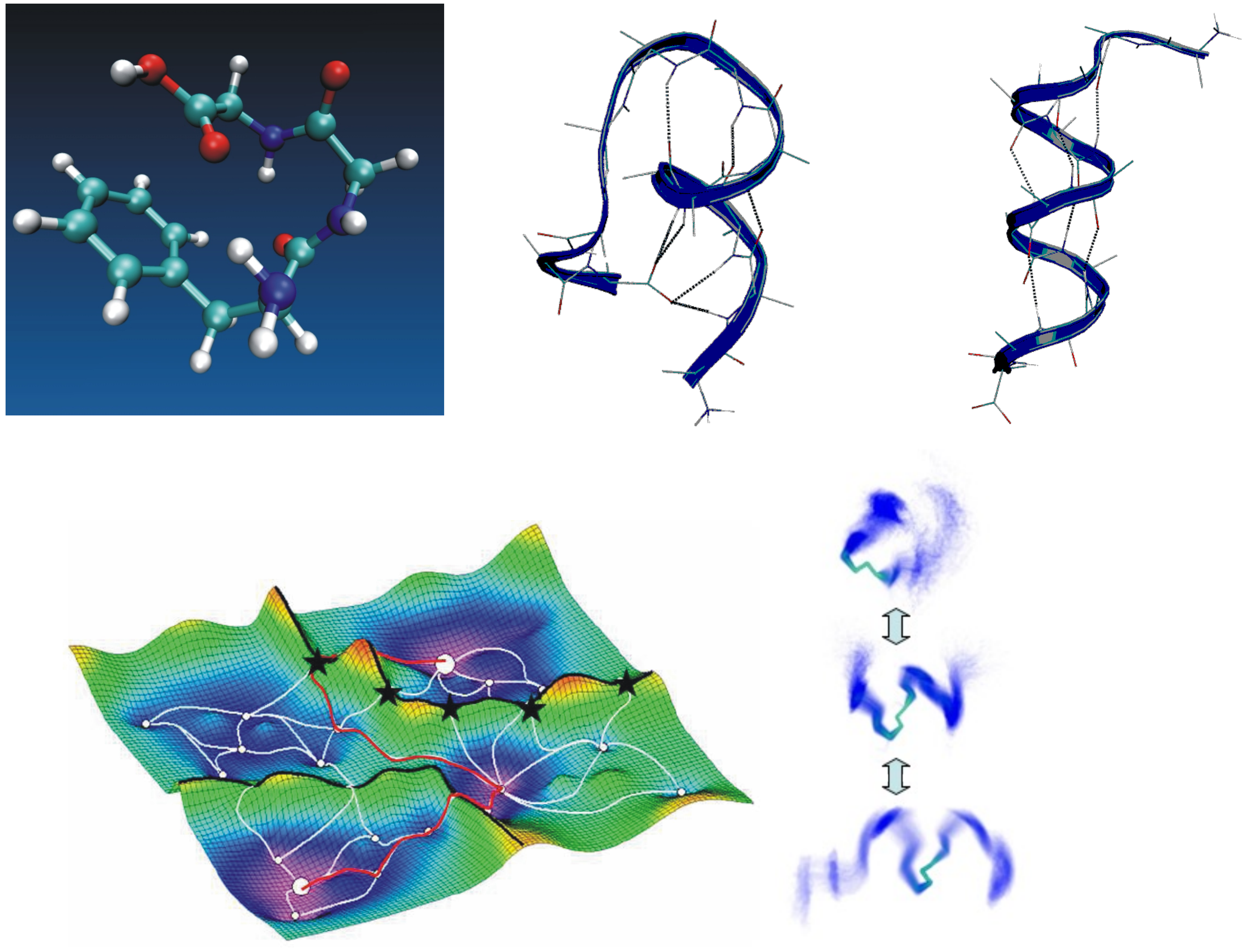
Типичный пример — гауссовские смеси

Общая (гибкая, быстрая, хорошо настраиваемая, скалируемая) модель многомерного распределения:

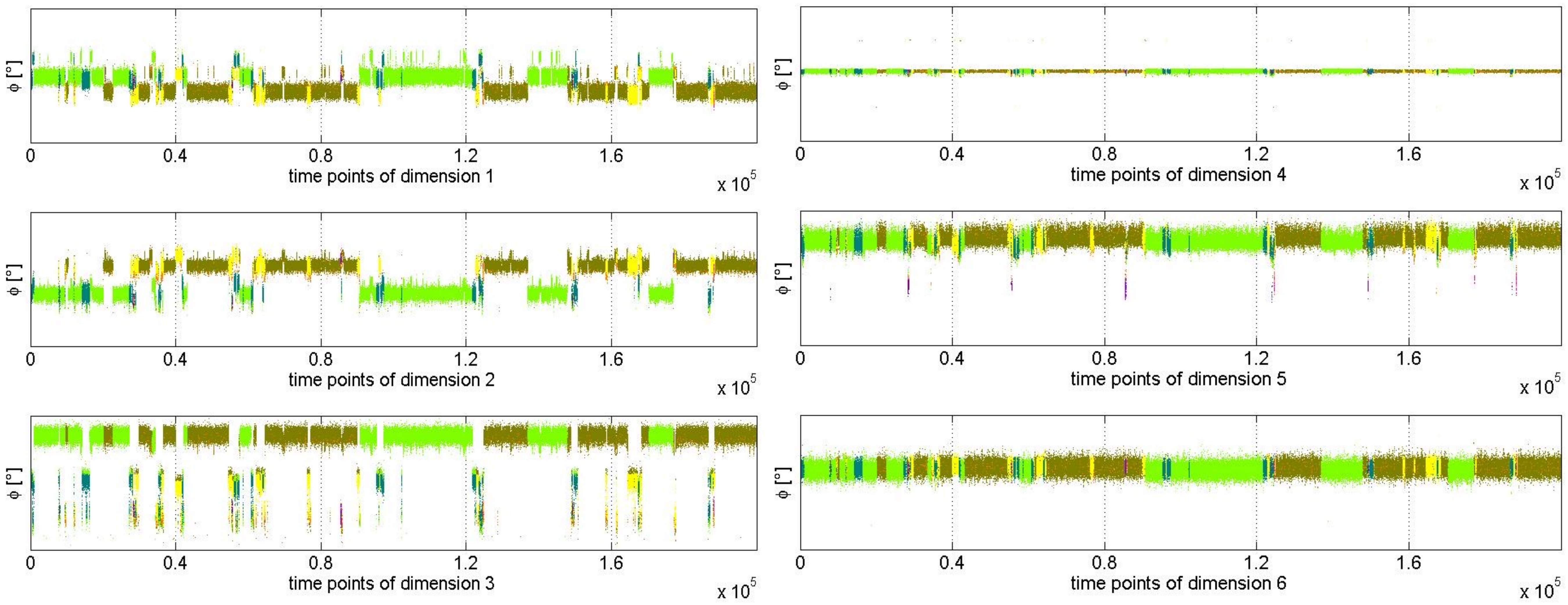
Маломерная гауссовская смесь +
Многомерный гауссовский шум



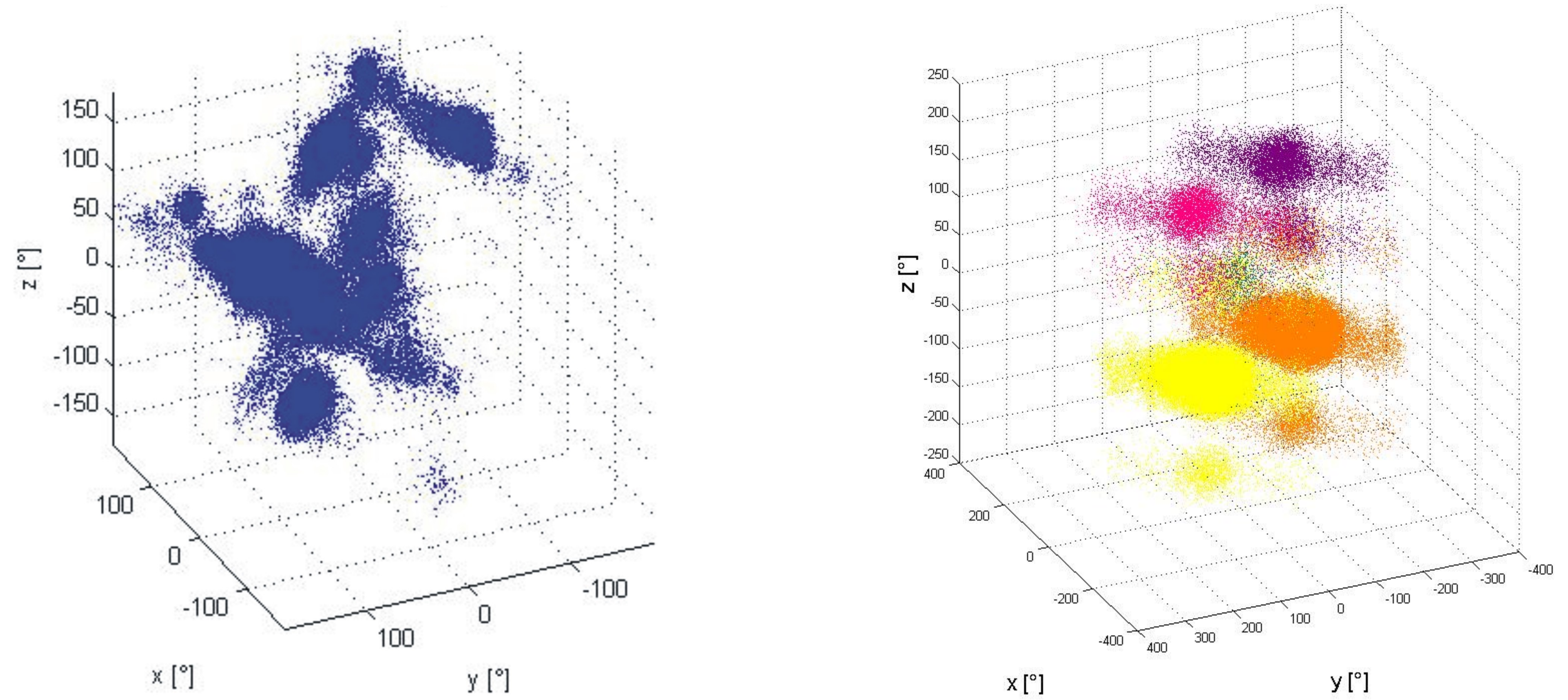
Моделирование протеинов



Моделирование протеинов



Моделирование протеинов



Выводы и заключения

Понимание математических законов важно для моделей ИИ, определения их надежности

Основные понятия в формулировках — эффективная размерность и эффективный объем выборки

Методы и идеи снижения размерности могут быть эффективно использованы в современных моделях ИИ

Расширение параметрического пространства может быть полезным (благословление размерности), при условии что эффективная размерность не «взрывается»

References I

- Al-Ghattas, O., Chen, J., and Sanz-Alonso, D. (2025).
  Sharp concentration of simple random tensors.
  arxiv.org/abs/2502.16916.

- Bach, F. (2024).
  High-dimensional analysis of double descent for linear regression with random projections.
  SIAM Journal on Mathematics of Data Science, 6(1):26—50.

- Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020).
  Benign overfitting in linear regression.
  Proceedings of the National Academy of Sciences, 117(48):30063—30070.

- Candès, E. J. and Sur, P. (2020).
  The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression.
  The Annals of Statistics, 48(1):27—42.

- Chen, M., Jiang, H., Liao, W., and Zhao, T. (2019).
  Efficient approximation of deep relu networks for functions on low dimensional manifolds.
  ArXiv, abs/1908.01842:null.

- Cheng, C. and Montanari, A. (2022).
  Dimension free ridge regression.
  https://arxiv.org/abs/2210.08571.

- Cloninger, A. and Klock, T. (2021).
  A deep network construction that adapts to intrinsic dimensionality beyond the domain.
  Neural Networks, 141:404—419.

References II

- Fan, J., Gu, Y., and Zhou, W.-X. (2024).
  How do noise tails impact on deep ReLU networks?
  The Annals of Statistics, 52(4):1845—1871.

- Jiao, Y., Shen, G., Lin, Y., and Huang, J. (2021).
  Deep nonparametric regression on approximately low-dimensional manifolds.
  arXiv: Statistics Theory, page null.

- Jiao, Y., Shen, G., Lin, Y., and Huang, J. (2023).
  Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors.
  The Annals of Statistics, 51:691—716.

- Kohler, M. and Krzyzak, A. (2022).
  Analysis of the rate of convergence of an over-parametrized deep neural network estimate learned by gradient descent.

- Kohler, M., KrzyÅŒak, A., and Langer, S. (2019).
  Estimation of a function of low local dimensionality by deep neural networks.
  IEEE Transactions on Information Theory, 68:4032—4042.

- Koltchinskii, V. and Lounici, K. (2017).
  Concentration inequalities and moment bounds for sample covariance operators.
  Bernoulli, 23(1):110—133.

- Kuchelmeister, F. and van de Geer, S. (2024).
  Finite sample rates for logistic regression with small noise or few samples.
  Sankhya A.

References III

- Liu, H., Chen, M., Er, S., Liao, W., Zhang, T.-M., and Zhao, T. (2022).
  Benefits of overparameterized convolutional residual networks: Function approximation under smoothness constraint.
  ArXiv, abs/2206.04569:null.

- Liu, H., Chen, M., Zhao, T., and Liao, W. (2021).
  Besov function approximation and binary classification on low-dimensional manifolds using convolutional residual networks.

- Montanari, A., Ruan, F., Sohn, Y., and Yan, J. (2025).
  The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime.
  The Annals of Statistics, 53(2):822—853.

- Nesterov, Y. and Nemirovskii, A. (1994).
  Interior-Point Polynomial Algorithms in Convex Programming.
  Society for Industrial and Applied Mathematics.

- Noskov, F., Puchkin, N., and Spokoiny, V. (2025).
  Dimension-free bounds in high-dimensional linear regression via error-in-operator approach.
  arxiv.org/abs/2502.15437.

- Ostrovskii, D. M. and Bach, F. (2021).
  Finite-sample analysis of M-estimators using self-concordance.
  Electronic Journal of Statistics, 15(1):326—391.

- Schmidt-Hieber, J. (2020).
  Nonparametric regression using deep neural networks with ReLU activation function.
  The Annals of Statistics, 48(4):1875—1897.

References IV

- Shen, G., Jiao, Y., Lin, Y., and Huang, J. (2022).
  Approximation with cnns in sobolev space: with applications to classification.
- Shen, G., Jiao, Y., Lin, Y., and Huang, J. (2023).
  Differentiable neural networks with repu activation: with applications to score estimation and isotonic regression.
  ArXiv, abs/2305.00608:null.
- Simionescu-Badea, C. (2022).
  Analysis of the rate of convergence of fully connected deep neural network regression estimates with smooth activation function.
- Spokoiny, V. (2025a).
  Marginal minimization and sup-norm expansions in perturbed optimization.
  arxiv.org/2505.02562.
- Spokoiny, V. (2025b).
  Sharp bounds in perturbed smooth optimization.
  arxiv.org/2504.11834.
- Sur, P. and Candès, E. J. (2019).
  A modern maximum-likelihood theory for high-dimensional logistic regression.
  Proceedings of the National Academy of Sciences, 116(29):14516—14525.
- Tropp, J. A. (2015).
  An introduction to matrix concentration inequalities.
  Foundations and Trends in Machine Learning, 8(1-2):1—230.

References V

- Vershynin, R. (2018).
  High-Dimensional Probability: An Introduction with Applications in Data Science.
  Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

- Wu, J., Marion, P., and Bartlett, P. (2025).
  Large stepsizes accelerate gradient descent for regularized logistic regression.
  arXiv preprint arXiv:2506.02336.

- Zhang, Z., Chen, M., Wang, M., Liao, W., and Zhao, T. (2023).
  Effective minkowski dimension of deep nonparametric regression: Function approximation and statistical theories.
  ArXiv, abs/2306.14859:null.

- Zhivotovskiy, N. (2024).
  Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle.
  Electronic Journal of Probability, 29(none):1—28.

- Zuowei Shen, Z. S., Haizhao Yang, H. Y., and Shijun Zhang, S. Z. (2020).
  Deep network approximation characterized by number of neurons.
  Communications in Computational Physics, 28(5):1768—1811.

ЦЕНТРАЛЬНЫЙ
УНИВЕРСИТЕТ

Спасибо
за внимание!