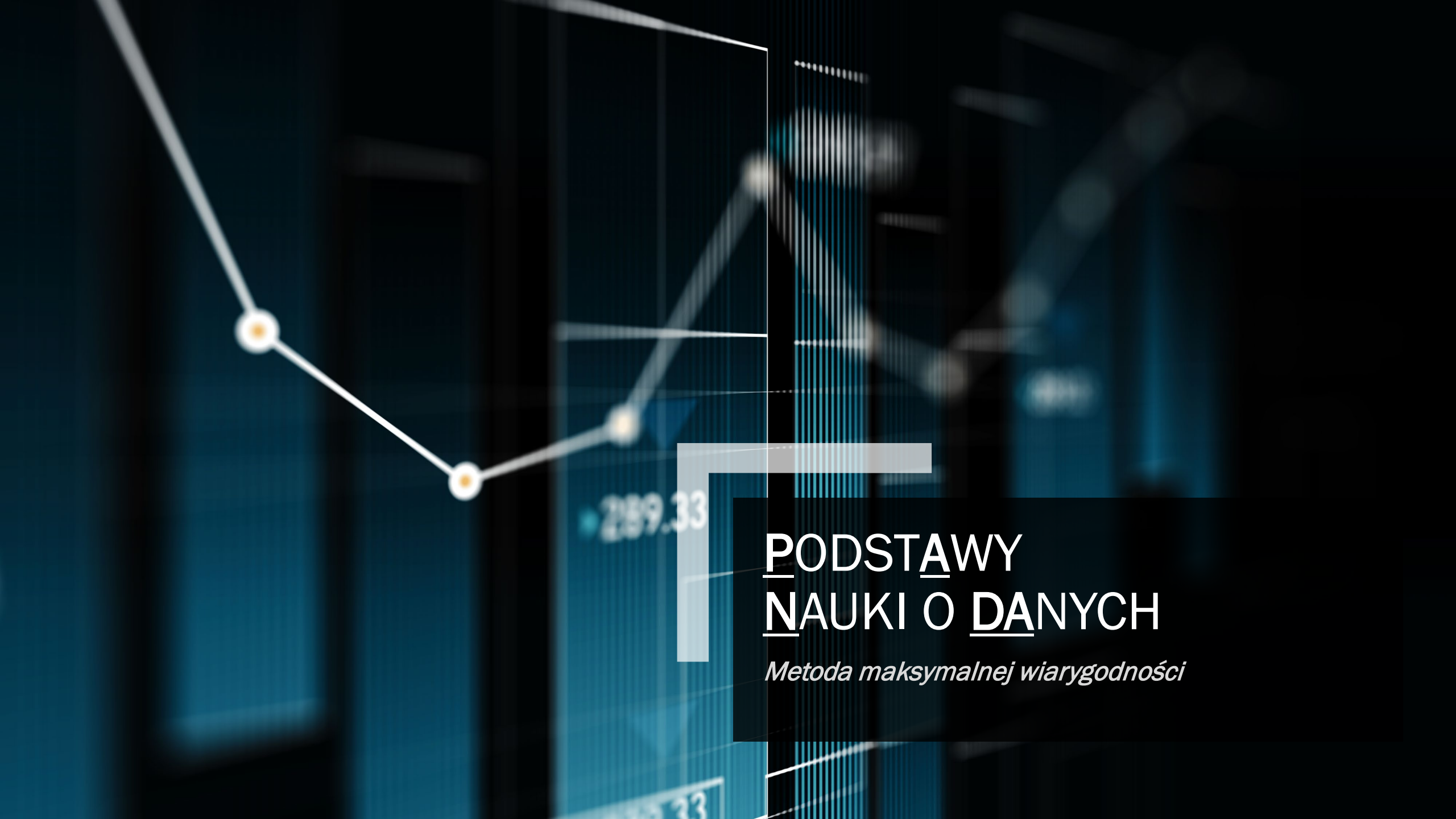




PODSTAWY NAUKI O DANYCH

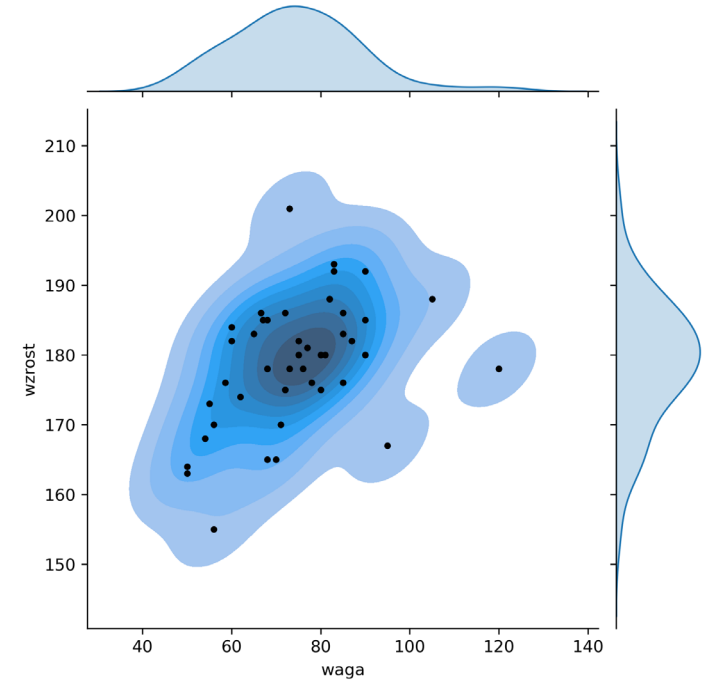
Metoda maksymalnej wiarygodności

Przypominajka statystyczna

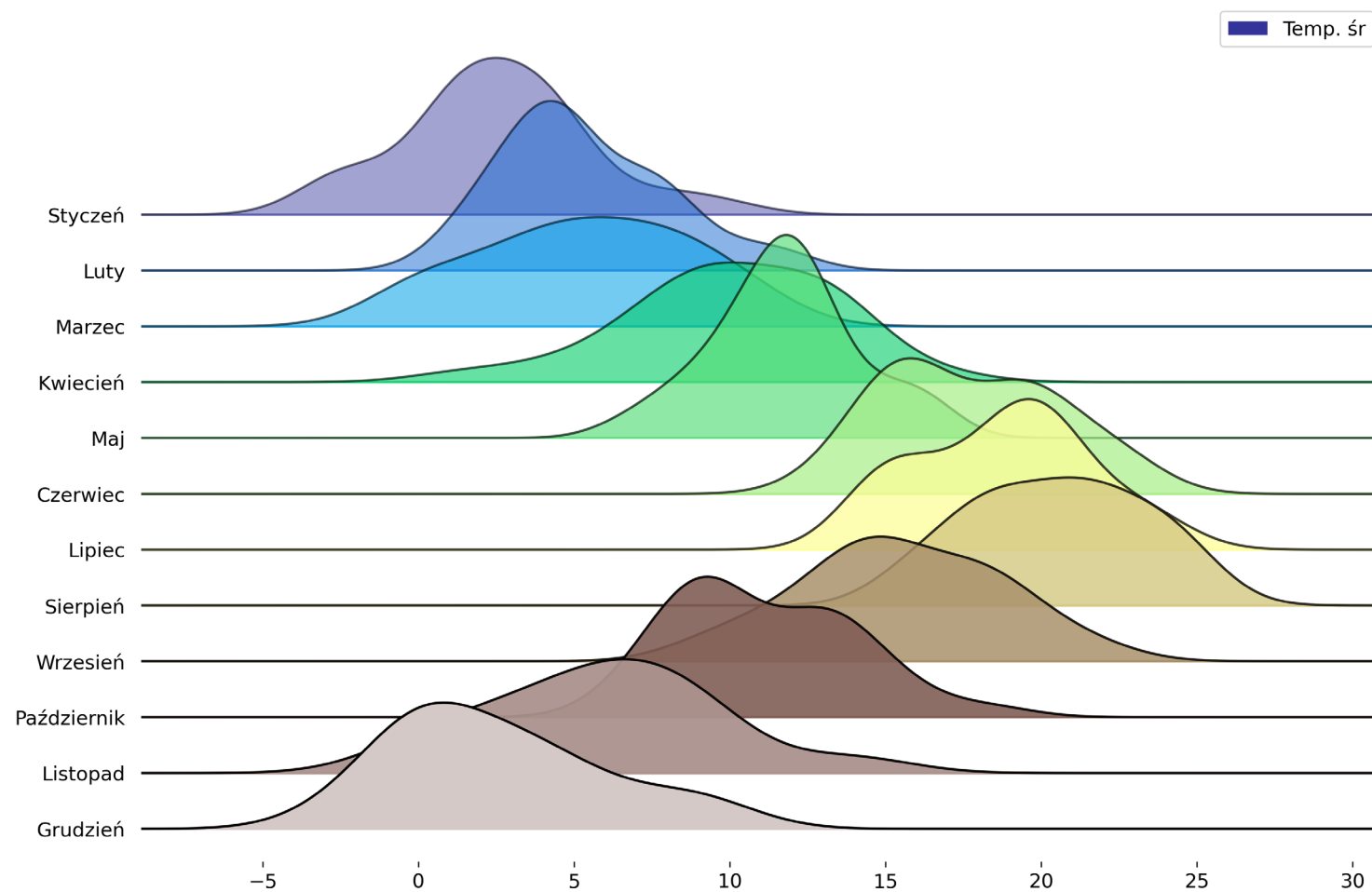
- Rozkład łączny – pełna informacja probabilistyczna
- Jak z rozkładu łącznego $p(X, \theta)$ wyznaczyć inne rozkłady?
- **Rozkłady brzegowe** (ang. *marginal PDFs*)

$$p(X) = \int_{D_\theta} p(X, \theta) d\theta \quad p(\theta) = \int_{D_X} p(X, \theta) dX$$

$$p(\theta|X) = \frac{p(X, \theta)}{p(X)} \quad p(X|\theta) = \frac{p(X, \theta)}{p(\theta)}$$

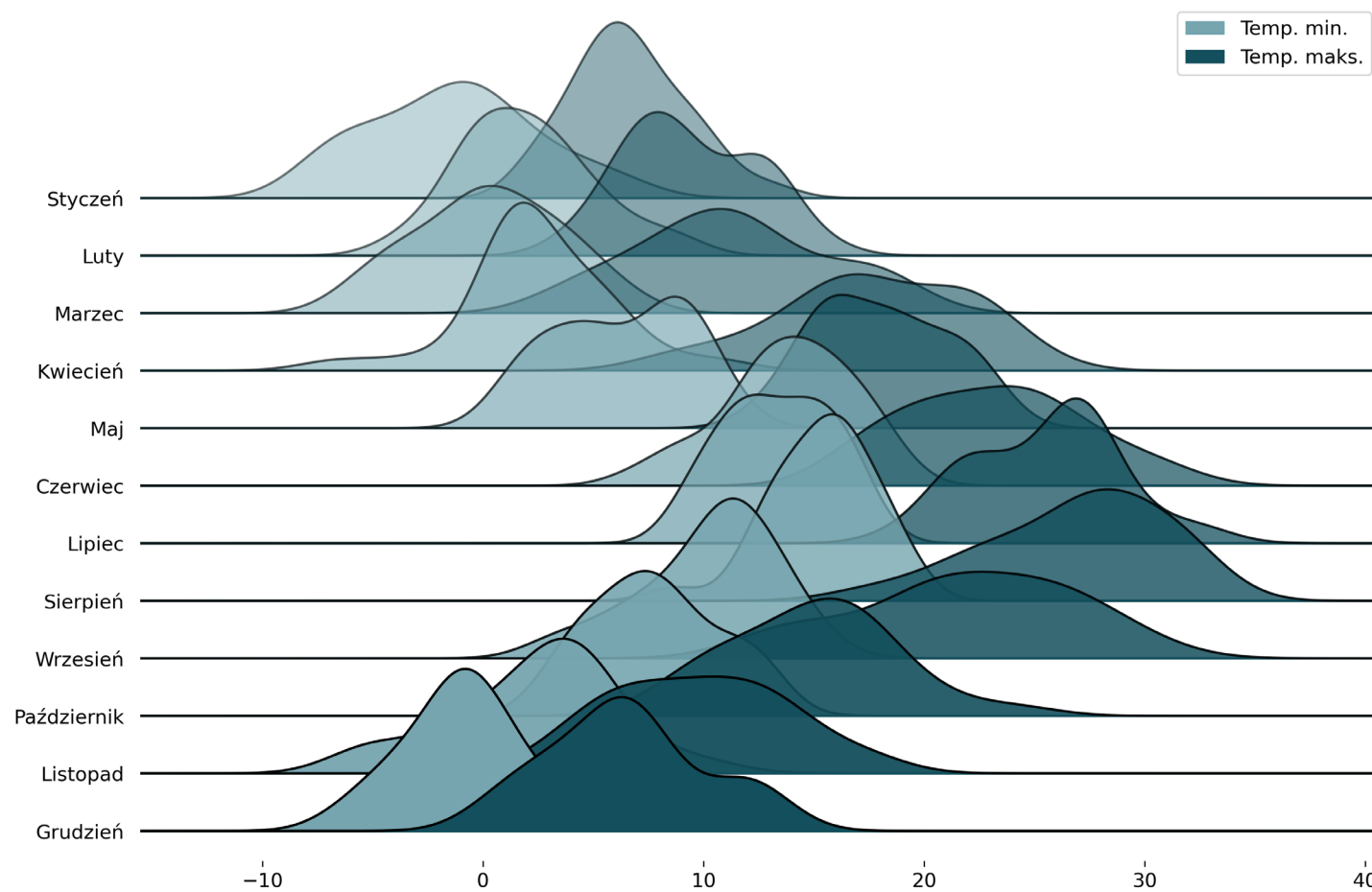


Przypominajka statystyczna



$$P(\text{Temp śr}|\text{miesiąc})$$

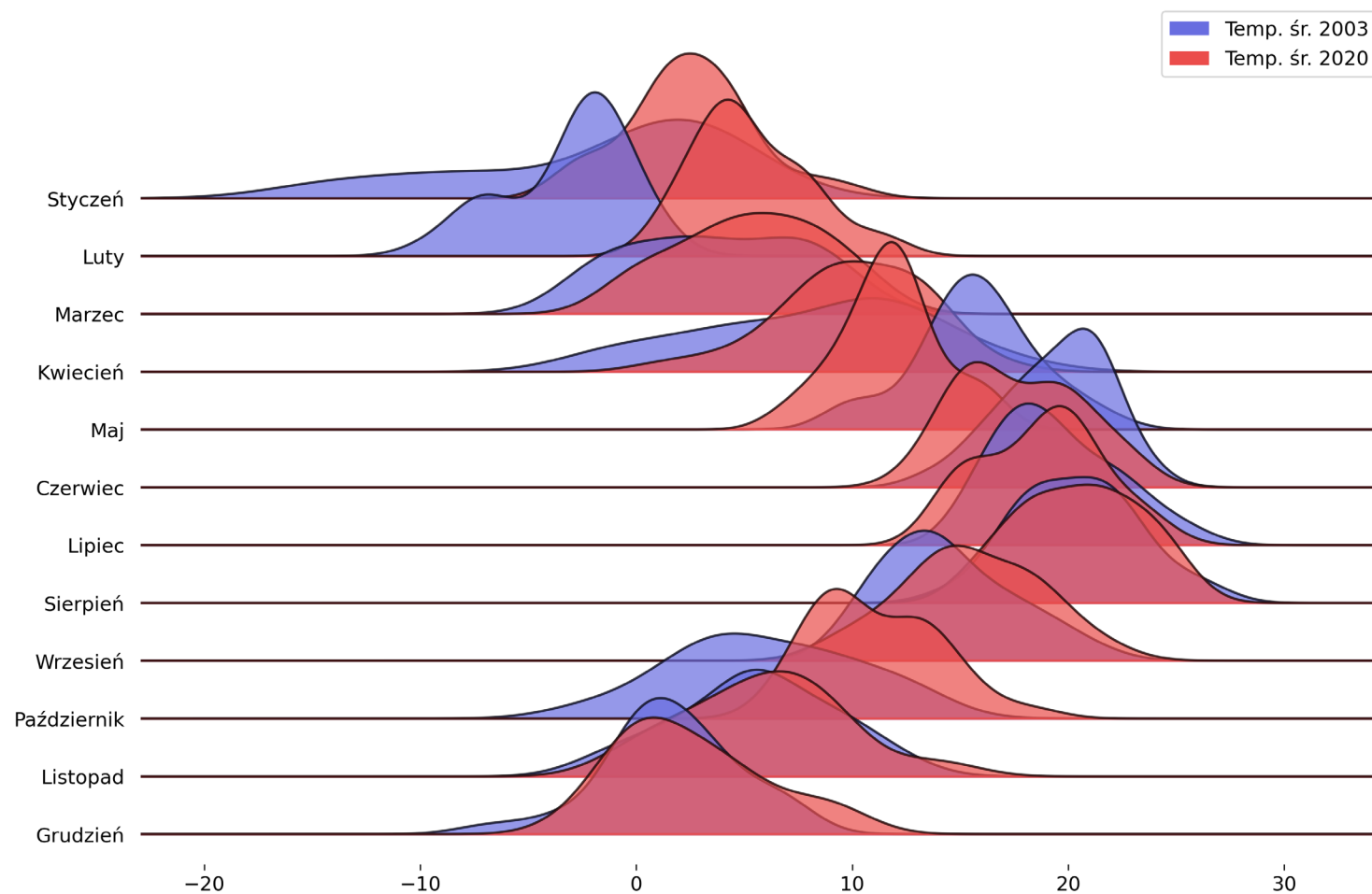
Przypominajka statystyczna



$$P(\text{Temp min}|\text{miesiac})$$

$$P(\text{Temp maks}|\text{miesiac})$$

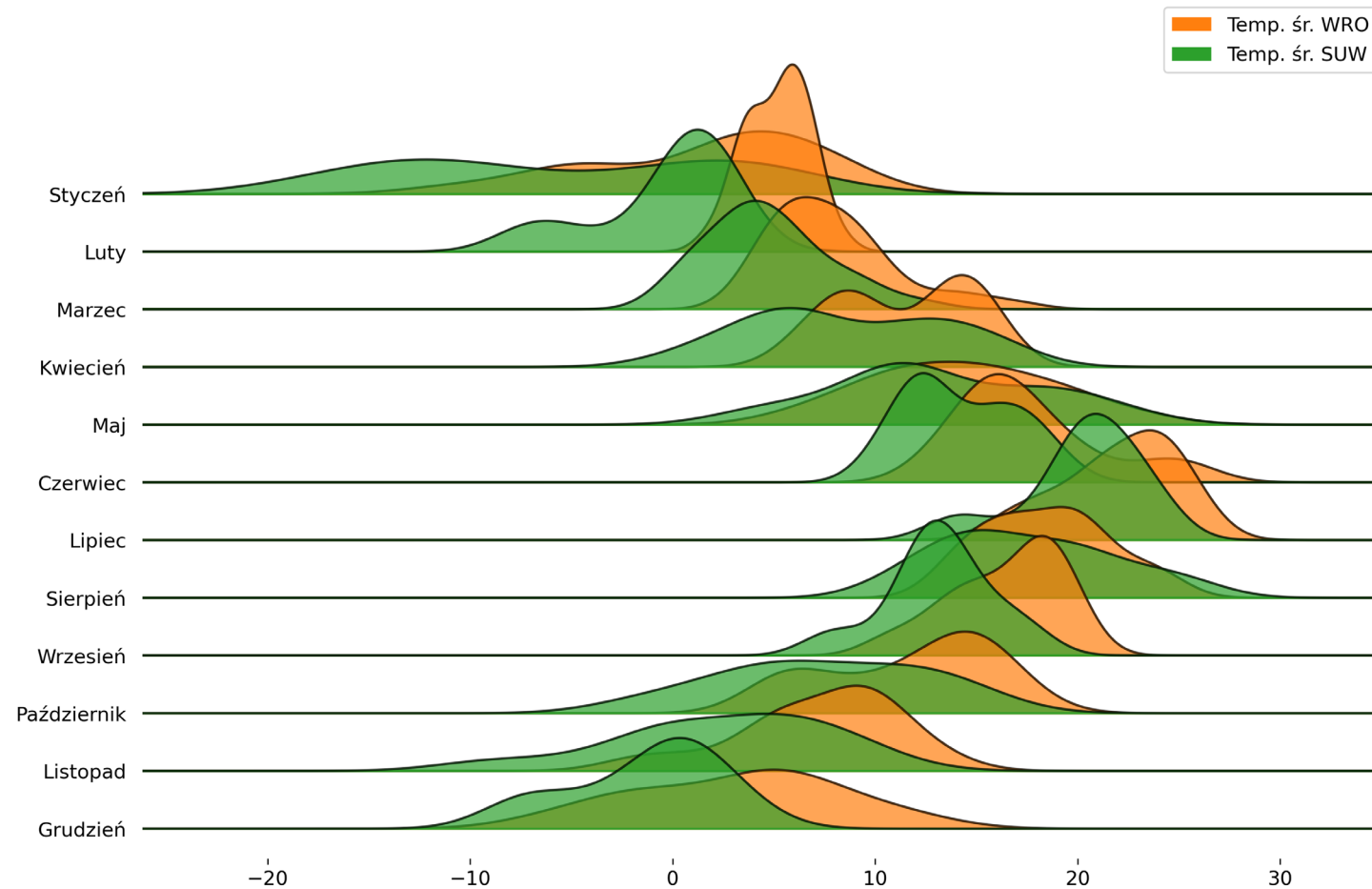
Przypominajka statystyczna



$$P(\text{Temp śr 2003}|\text{miesiac})$$

$$P(\text{Temp śr 2020}|\text{miesiac})$$

Przypominajka statystyczna

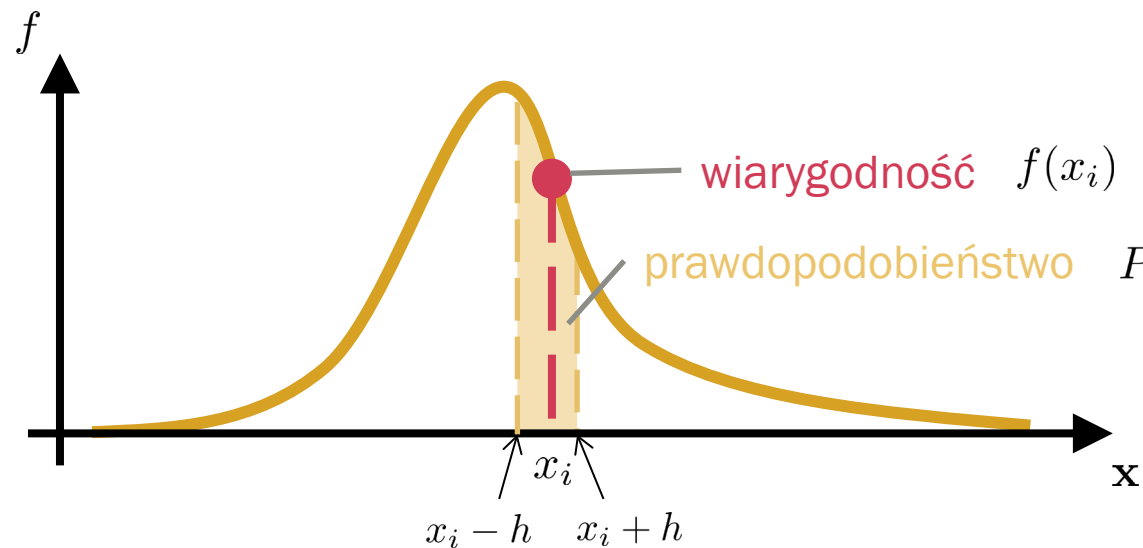


$P(\text{Temp śr Wrocław}|\text{miesiąc})$

$P(\text{Temp śr Suwałki}|\text{miesiąc})$

Prawdopodobieństwo a wiarygodność

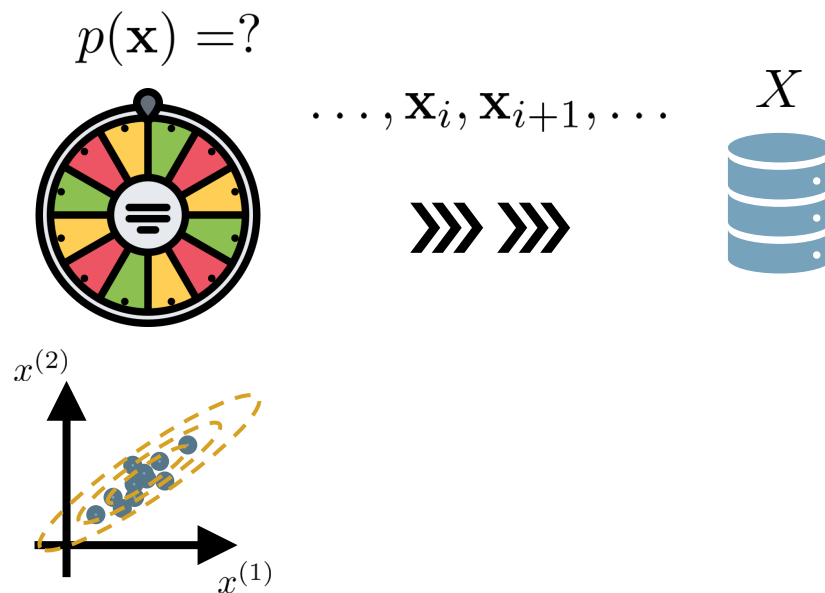
ang. *probability vs likelihood*



$$P(x_i - h \leq x \leq x_i + h) = \int_{x_i - h}^{x_i + h} f(x) dx$$

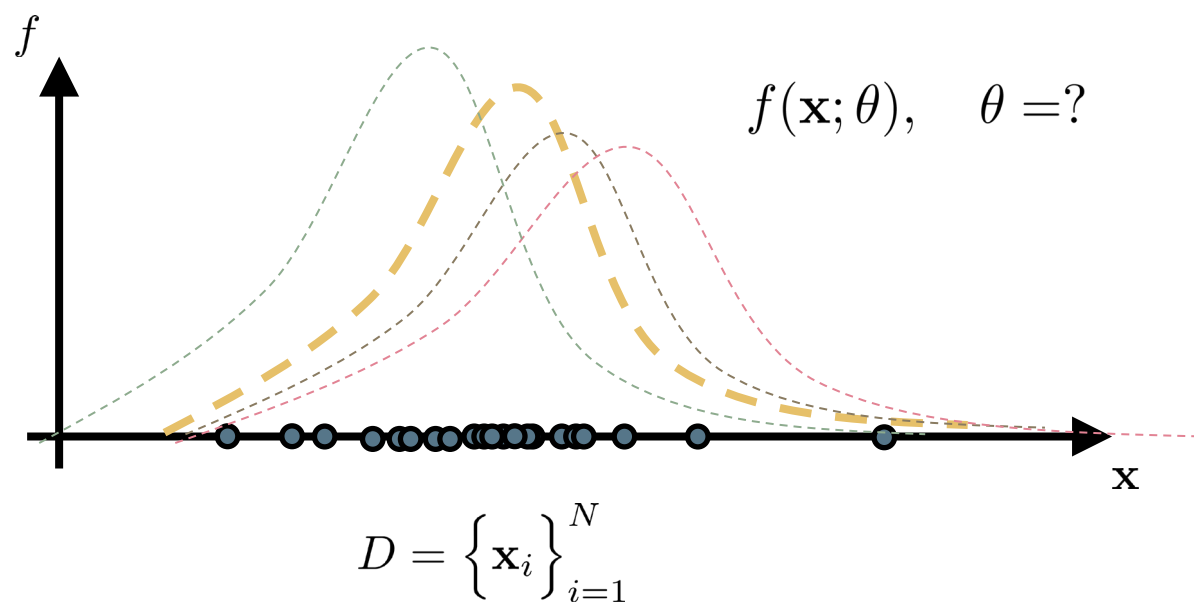
$$\int_{-\infty}^{\infty} f(x) dx = 1$$

Rozkład prawdopodobieństwa jako model generatywny



Dane pomiarowe traktujemy, jakby pochodziły z odpowiednio przygotowanego generatora liczb losowych

Dopasowanie rozkładu do danych



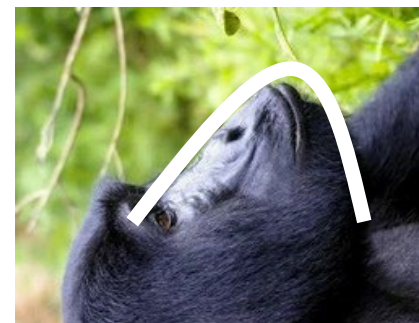
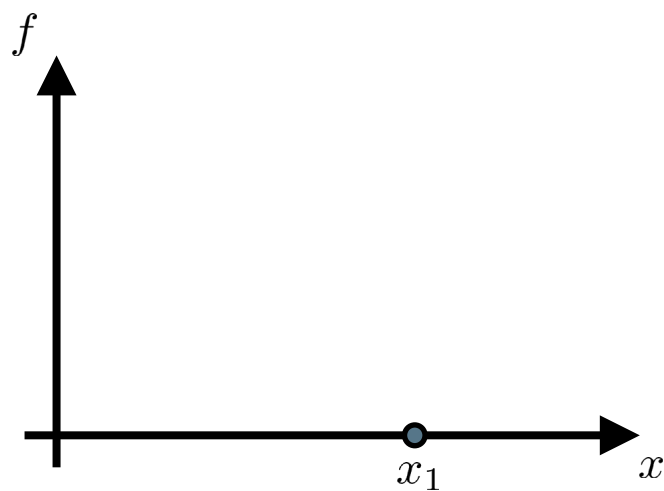
Dopasowanie rozkładu do danych

Zrób to sam



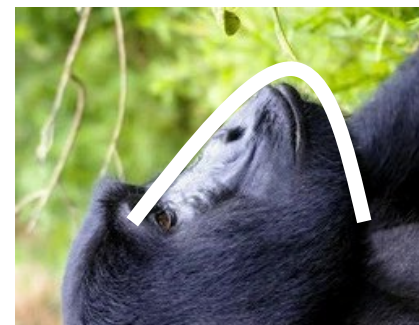
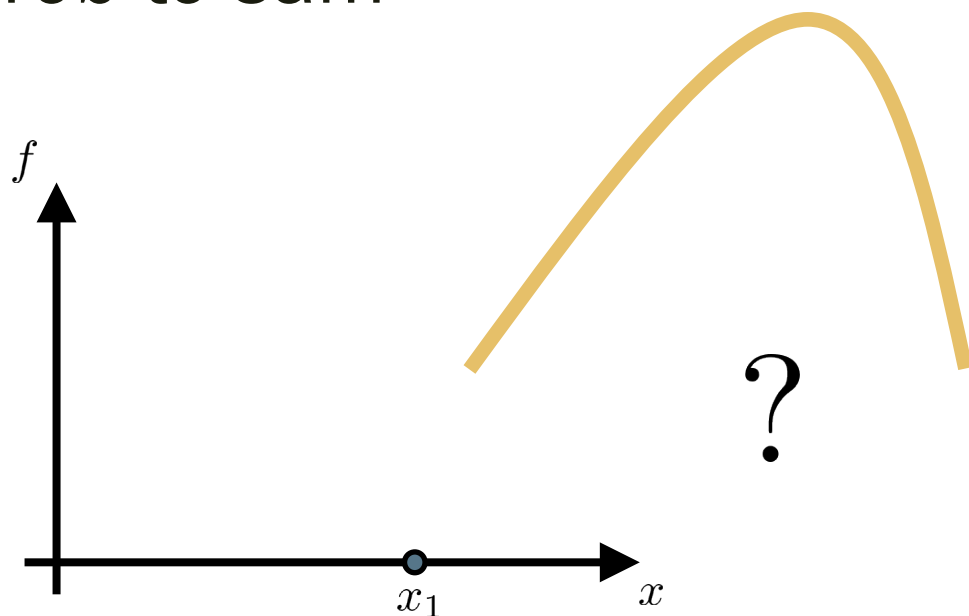
Dopasowanie rozkładu do danych

Zrób to sam



Dopasowanie rozkładu do danych

Zrób to sam



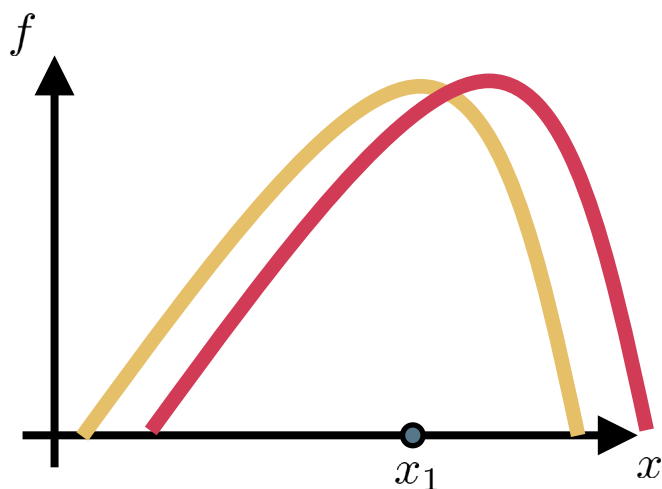
to jest pysk gorylcy

Gdzie go umieścisz na tle pojedynczego pomiaru?

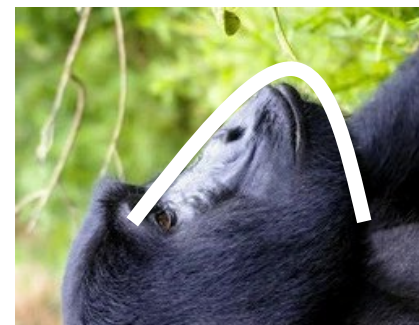
Czym się będziesz kierować?

Dopasowanie rozkładu do danych

Zrób to sam



?



to jest pysk gorylcy

Gdzie go umieścisz na tle pojedynczego pomiaru?

Czym się będziesz kierować?

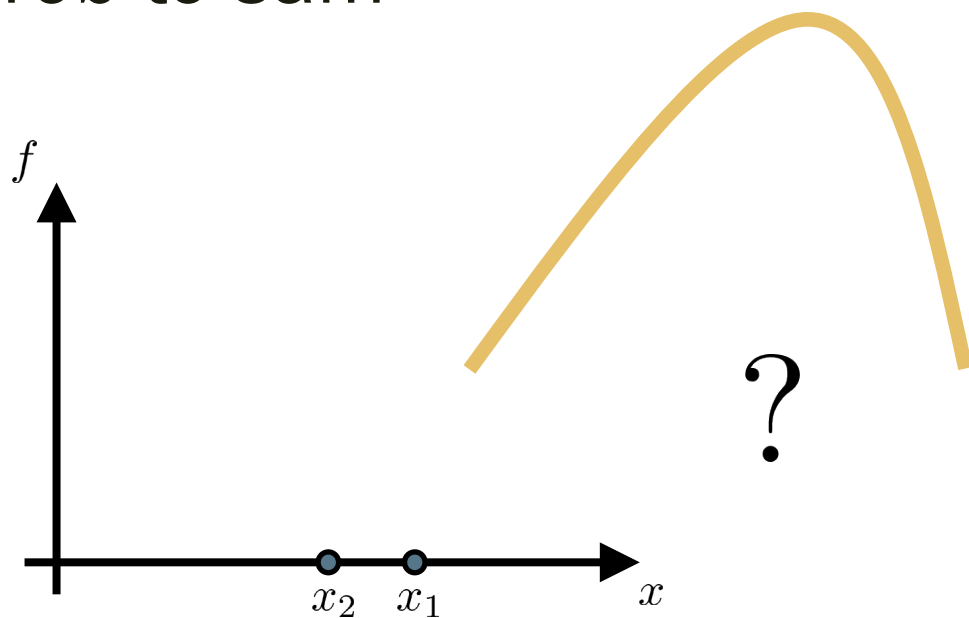
$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} f(x_1; \theta)$$

Metoda maksymalnej wiarygodności to **heurystyka**, czyli **zalgorytmizowana intuicja** / **droga na skróty**

Dopasowanie rozkładu do danych

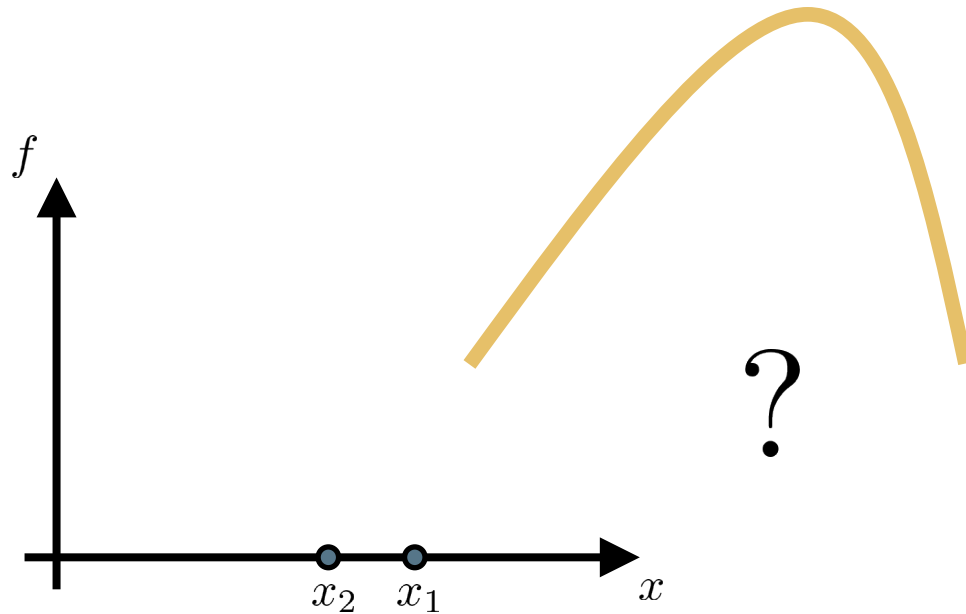
Zrób to sam

A co zrobisz teraz?



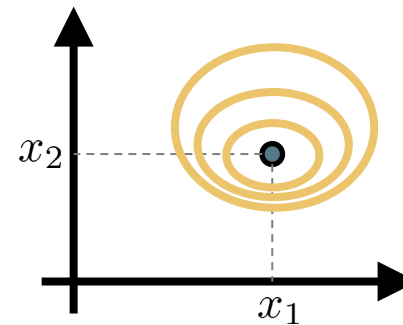
Dopasowanie rozkładu do danych

Zrób to sam



A co zrobisz teraz?

Może raczej tak:

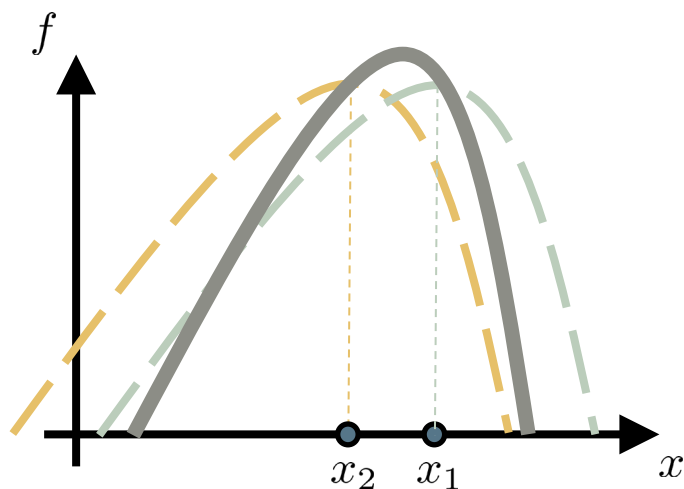


$$f(x_1, x_2; \theta), \quad \theta = ?$$

Dopasowanie rozkładu do danych

Zrób to sam

- To może lepiej tak



Zamiast $f(x_1, x_2; \theta)$

upraszczamy $f(x_1|\theta) \cdot f(x_2|\theta) \equiv p(X|\theta)$

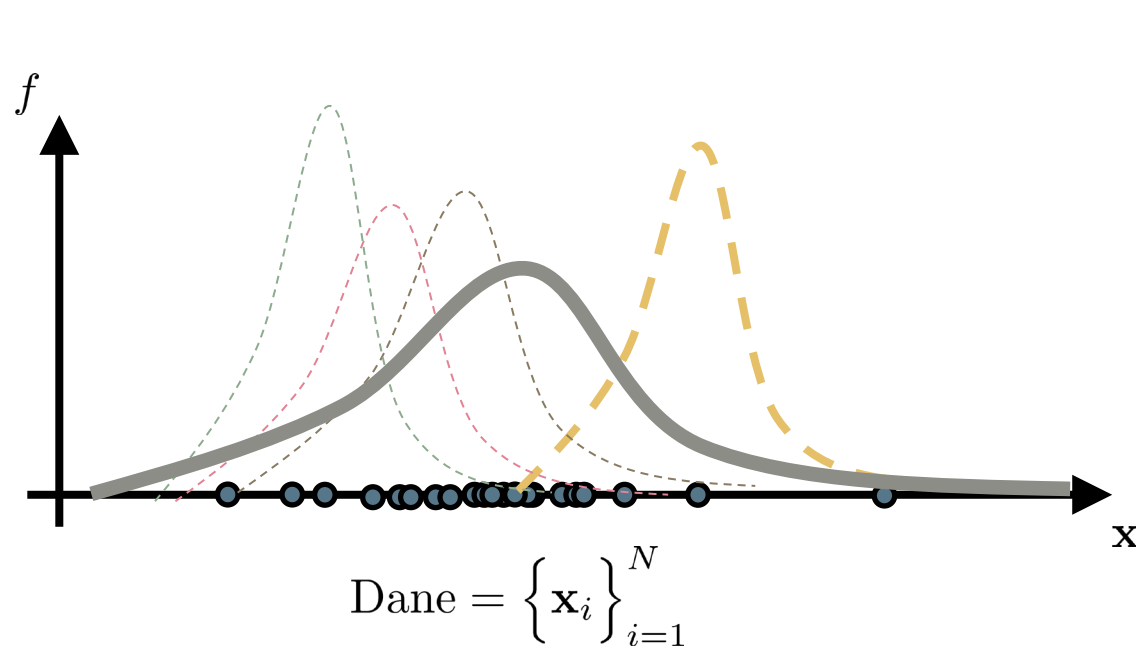
zachowujemy się więc, tak, jakby pomiary były niezależne:

$$\text{Cov}(x_1, x_2) = 0$$

... może są?

Funkcja wiarygodności

- Dla dowolnego zbioru danych



$$L(\theta) \equiv p(X|\theta) = \prod_{i=1}^N p(\mathbf{x}_i|\theta)$$

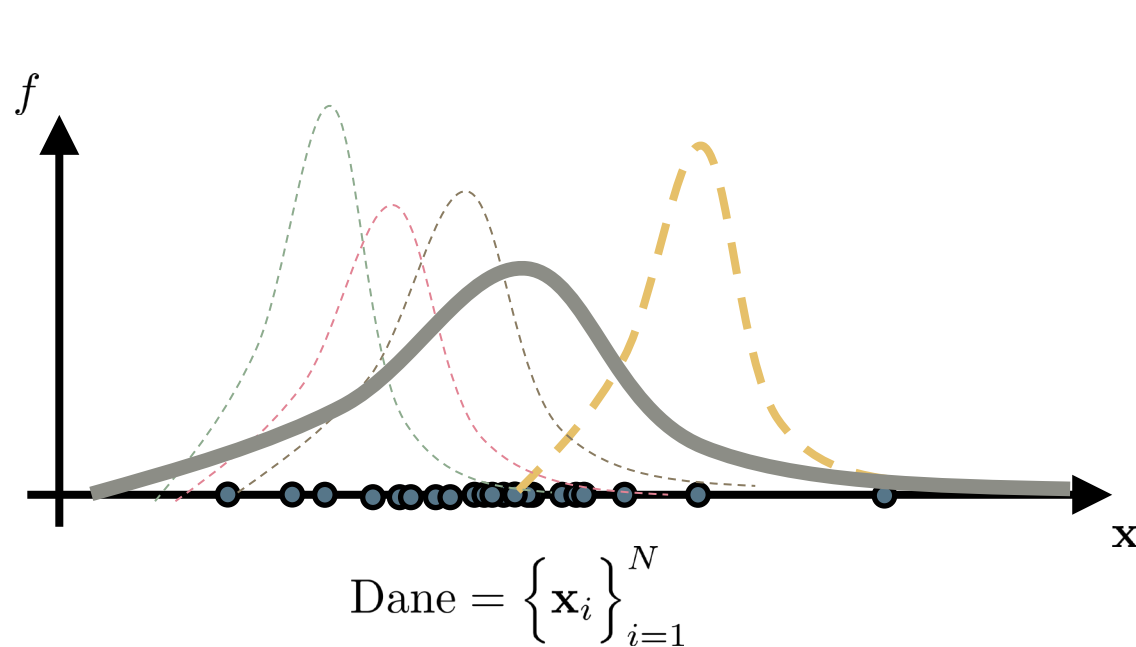
$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} L(\theta)$$

Dwa problemy:

- niedomiar (ang. *underflow error*)

Funkcja wiarygodności

- Dla dowolnego zbioru danych



$$L(\theta) \equiv p(X|\theta) = \prod_{i=1}^N p(\mathbf{x}_i|\theta)$$

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} L(\theta)$$

Dwa problemy:

- niedomiar (ang. *underflow error*)
- pochodna iloczynu funkcji

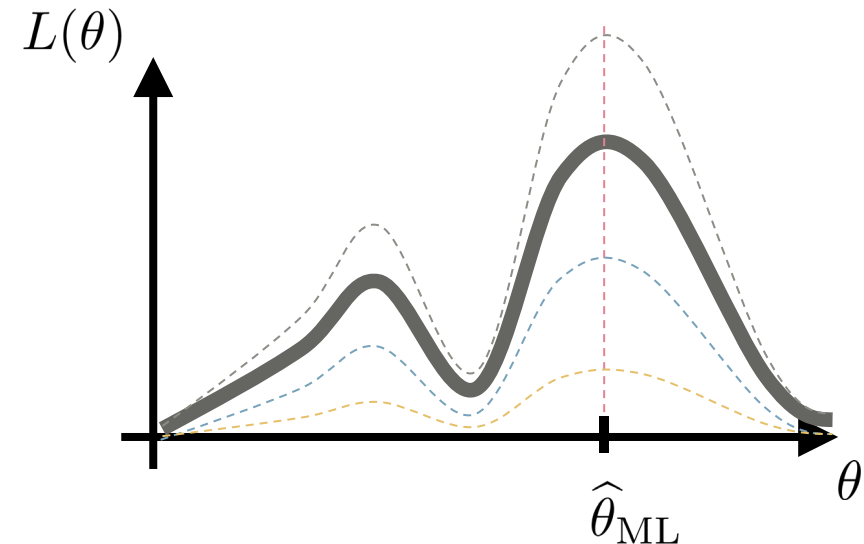
Funkcja wiarygodności

- Zmiana skali na logarytmiczną

$$\log \prod_{i=1}^N \heartsuit_i = \sum_{i=1}^N \log \heartsuit_i$$

$$\log L(\theta) = \sum_{i=1}^N \log p(\mathbf{x}_i|\theta)$$

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \log L(\theta)$$



Przykład zadania

- Dane pochodzą z rozkładu normalnego, ale nie znamy jego parametrów

$$\text{Dane} = \left\{ x_i \right\}_{i=1}^N$$

$$p(x; \theta) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right], \quad \theta = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$$

- Należy wyznaczyć wzór analityczny na parametry rozkładu tak, by maksymalizować funkcję wiarygodności



THE END