

Глубинное обучение

Введение. Backpropagation.

Ирина Сапарина

Организационные вопросы

- Материалы курса: github.com/fintech-dl-hse/course
- 12 основных тем + дополнительные
- Формула оценки $O_{\text{итог}} = 0.7 \times O_{\text{накоп}} + 0.3 \times O_{\text{экз}}$, автомат с 6.0
- Telegram чат
- Формы обратной связи - пишите свои вопросы!

Дополнительные материалы

- github.com/aosokin/dl_cshse_ami
- github.com/yandexdataschool/Practical_DL
- github.com/yandexdataschool/nlp_course
- CS231N CV with DL
- CS224N NLP with DL
- CS224W ML with Graphs
- [PyTorch Tutorials](https://pytorch.org/tutorials) и многие другие - гуглите!

О чём курс?

Глубинное обучение: приложения



Глубинное обучение: приложения

≡ Google Translate ⋮

Text Documents

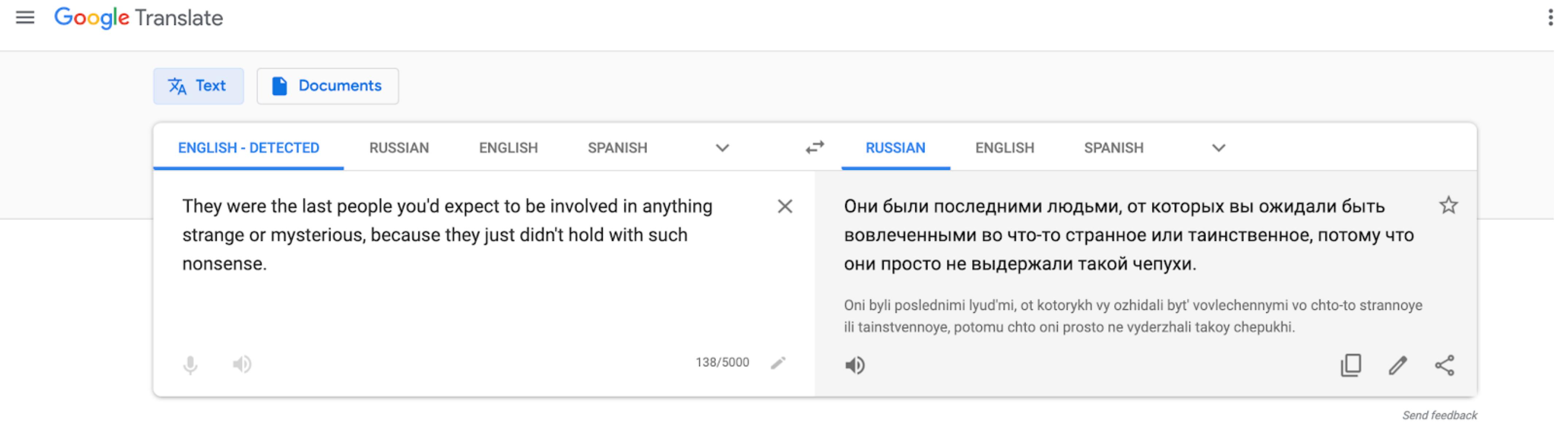
ENGLISH - DETECTED RUSSIAN ENGLISH SPANISH RUSSIAN ENGLISH SPANISH

They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

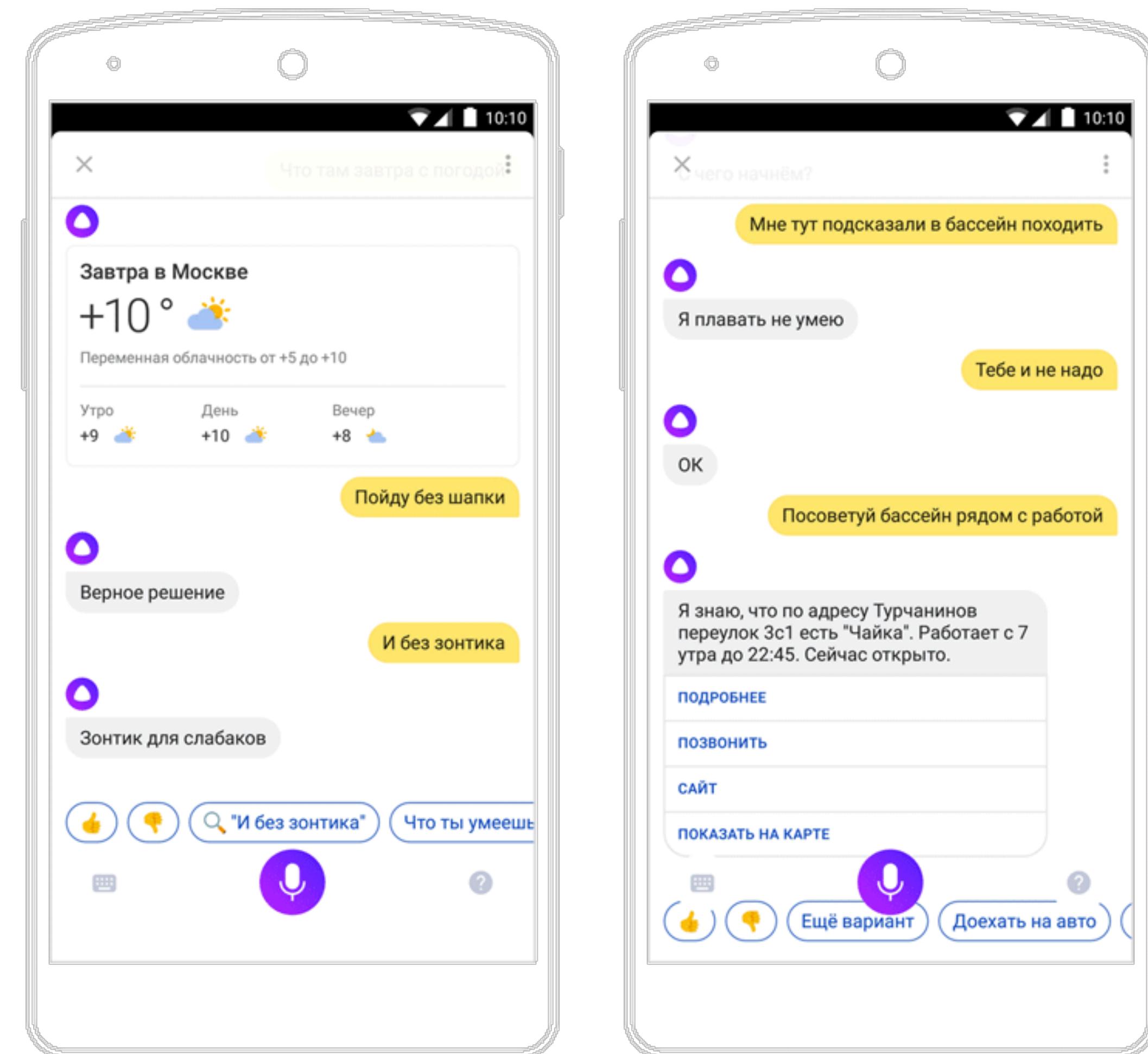
Oни были последними людьми, от которых вы ожидали быть вовлечеными во что-то странное или таинственное, потому что они просто не выдержали такой чепухи.

Oni byli poslednimi lyud'mi, ot kotorikh vy ozhidali byt' vovlechennymi vo chto-to strannoye ili tainstvennoye, potomu chto oni prosto ne vyderzhali takoy chepukhi.

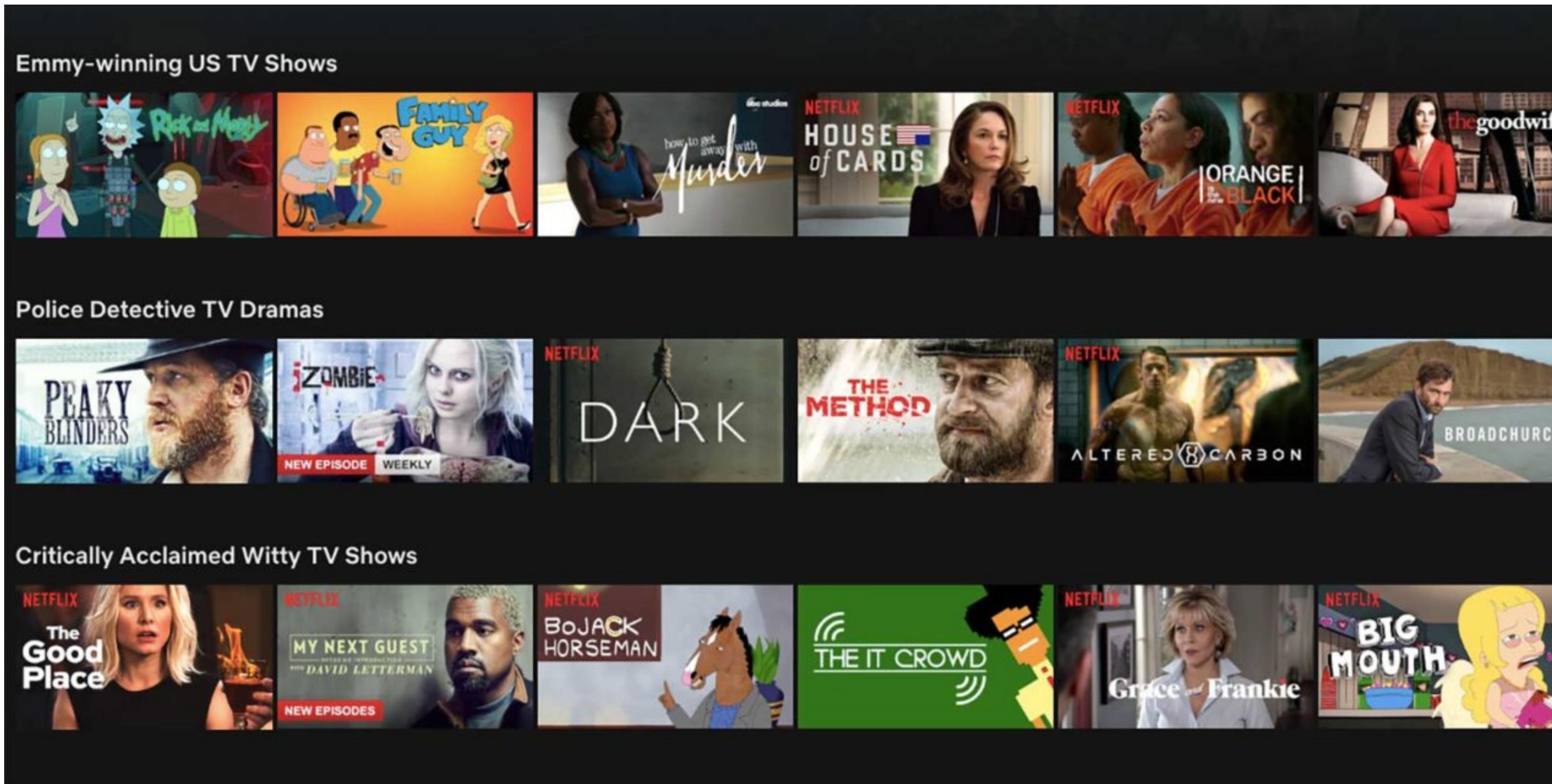
Send feedback



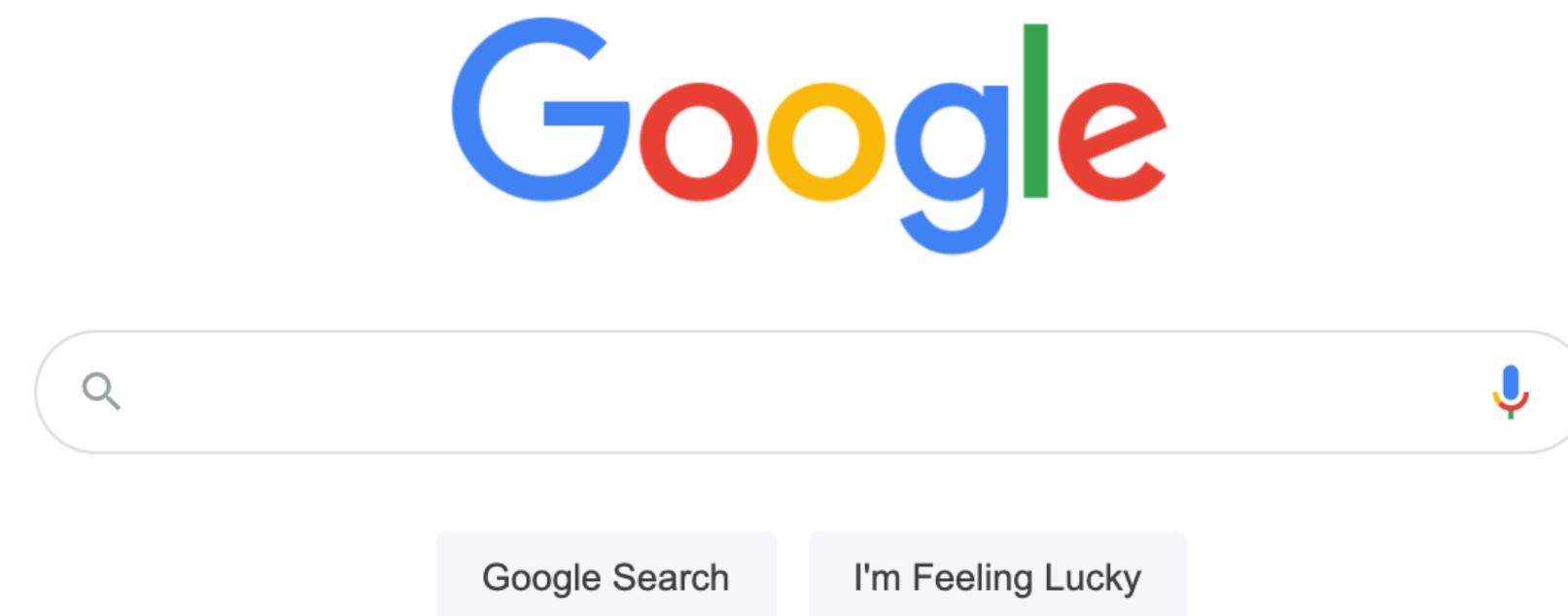
Глубинное обучение: приложения



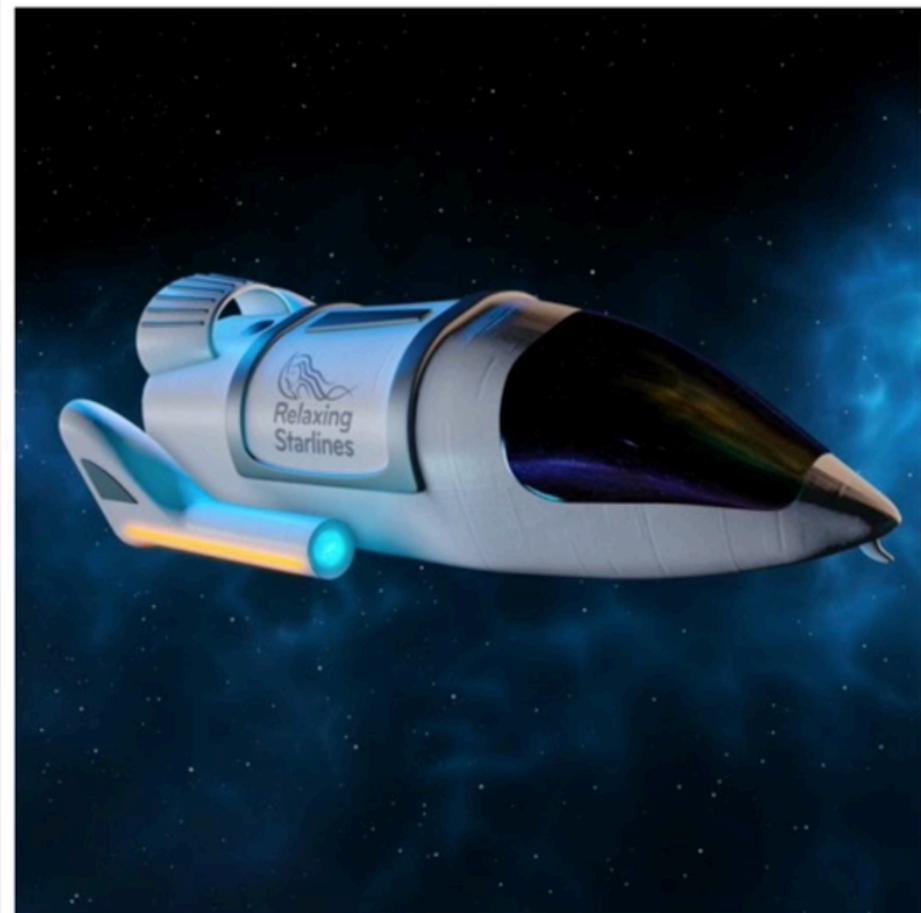
Глубинное обучение: приложения



Глубинное обучение: приложения



Глубинное обучение: приложения



Add this image of a rocketship:

<https://i1.sndcdn.com/artworks-j8xjG7zc1wmTe07b-06l83w-t500x500.jpg>



```
/* Add this image of a
rocketship:
https://i1.sndcdn.com/artworks
-j8xjG7zc1wmTe07b-06l83w-
t500x500.jpg */
var rocketship =
document.createElement('img');
rocketship.src =
'https://i1.sndcdn.com/artwork
s-j8xjG7zc1wmTe07b-06l83w-
t500x500.jpg';
document.body.appendChild(rock
etship);
```

```
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return ''.join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return ''.join(groups)
```

Глубинное обучение: о курсе

- Начнем с самых основ и дойдем до современных SOTA моделей
 - Как обучать нейросети, когда и какие архитектуры использовать
 - Практическое применение - CV, NLP

ML recap

Постановка задачи: обучение с учителем

Объекты x_1, \dots, x_n

Ответы y_1, \dots, y_n

Постановка задачи: обучение с учителем

Объекты x_1, \dots, x_n

Ответы y_1, \dots, y_n

Функция потерь $L(\mathbf{y}, \hat{\mathbf{y}})$

Постановка задачи: обучение с учителем

Объекты x_1, \dots, x_n

Ответы y_1, \dots, y_n

Алгоритм предсказания $f(x, \theta)$

Функция потерь $L(y, \hat{y})$

Задача: настроить θ

Постановка задачи: обучение с учителем

Объекты x_1, \dots, x_n

Ответы y_1, \dots, y_n

Алгоритм предсказания $f(x, \theta)$

Функция потерь $L(\mathbf{y}, \hat{\mathbf{y}})$

Задача: настроить θ

$$\mathbb{E}_{(x,y)} L(\mathbf{y}, f(x, \theta)) \rightarrow \min_{\theta}$$

Постановка задачи: обучение с учителем

Объекты x_1, \dots, x_n

Ответы y_1, \dots, y_n

Алгоритм предсказания $f(x, \theta)$

Функция потерь $L(\mathbf{y}, \hat{\mathbf{y}})$

Задача: настроить θ

$$\mathbb{E}_{(x,y)} L(\mathbf{y}, f(x, \theta)) \rightarrow \min_{\theta}$$

$$\frac{1}{n} \sum_i L(y_i, f(x_i, \theta)) \rightarrow \min_{\theta}$$

эмпирический риск

Пример: бинарная классификация

Объекты x_1, \dots, x_n

Алгоритм предсказания $f(x, \theta)$

Функция потерь $L(y, \hat{y})$

Ответы $y_1, \dots, y_n \in [0,1]$



Задача: настроить θ

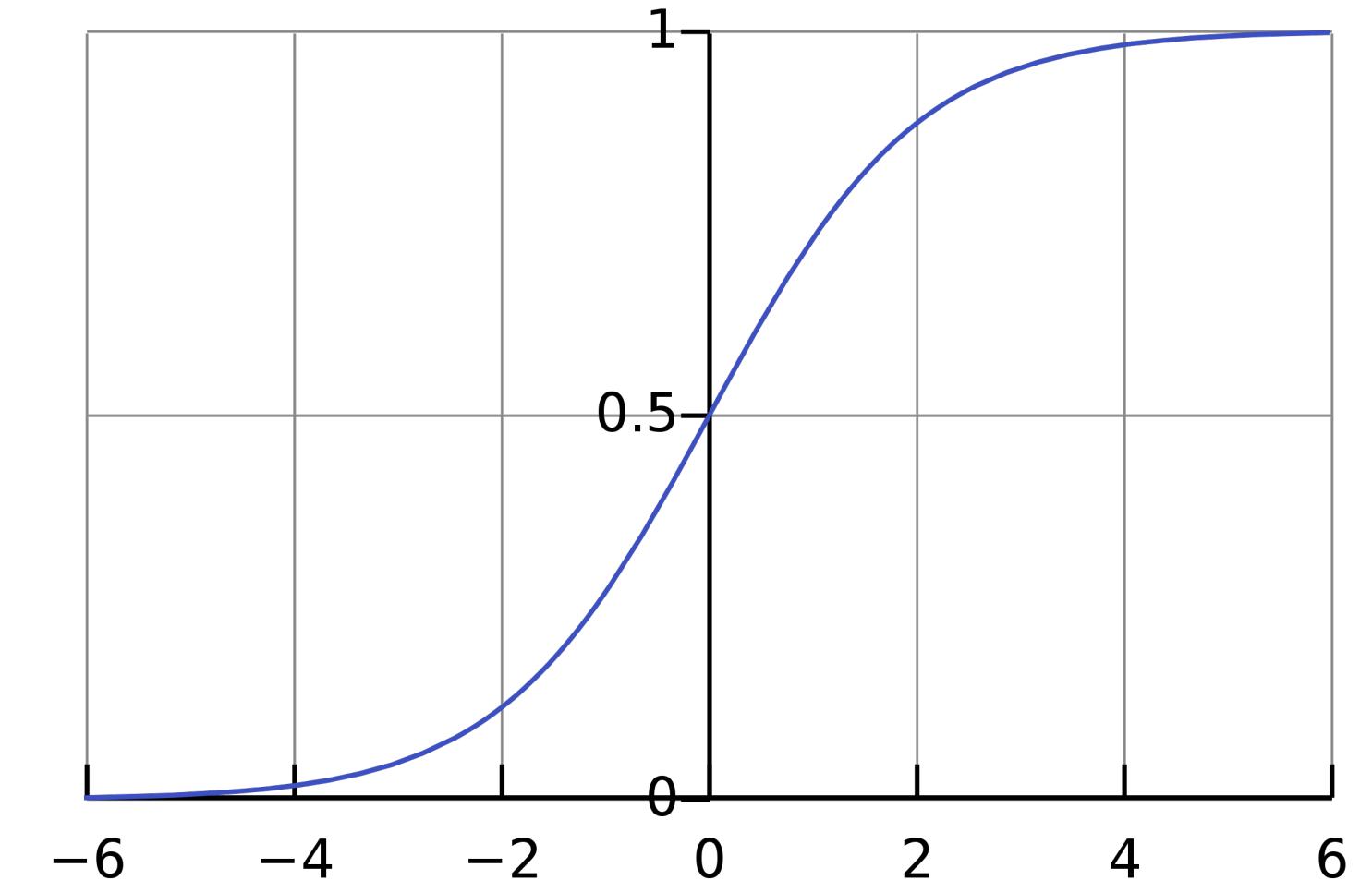
Логистическая регрессия

Объекты x_1, \dots, x_n

Ответы $y_1, \dots, y_n \in [0,1]$

Алгоритм предсказания $f(x, \theta) = P(y = 1 | x, \theta) > \frac{1}{2}$

$$P(y = 1 | x, \theta) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$



Логистическая регрессия

Объекты x_1, \dots, x_n

Алгоритм предсказания $f(x, \theta) = P(y = 1 | x, \theta) > \frac{1}{2}$

Ответы $y_1, \dots, y_n \in [0,1]$

$$P(y = 1 | x, \theta) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Функция потерь

$$L(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_i [y_i \log P(y = 1 | X, \theta) + (1 - y_i) \log(1 - P(y = 1 | X, \theta))]$$



Логистическая регрессия

	какие признаки?	
Объекты	x_1, \dots, x_n	
Ответы	$y_1, \dots, y_n \in [0,1]$	
Алгоритм предсказания	$f(x, \theta) = P(y = 1 x, \theta) > \frac{1}{2}$	
		
	$P(y = 1 x, \theta) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$	
Функция потерь		

$$L(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_i [y_i \log P(y = 1 | X, \theta) + (1 - y_i) \log(1 - P(y = 1 | X, \theta))]$$

Логистическая регрессия

Объекты x_1, \dots, x_n

Алгоритм предсказания $f(x, \theta) = P(y = 1 | x, \theta) > \frac{1}{2}$

Ответы $y_1, \dots, y_n \in [0,1]$



$$P(y = 1 | x, \theta) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Функция потерь

$$L(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_i [y_i \log P(y = 1 | X, \theta) + (1 - y_i) \log(1 - P(y = 1 | X, \theta))]$$

Логистическая регрессия

Объекты x_1, \dots, x_n

Алгоритм предсказания $f(x, \theta) = P(y = 1 | x, \theta) > \frac{1}{2}$

Ответы $y_1, \dots, y_n \in [0,1]$



$$P(y = 1 | x, \theta) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Функция потерь

$$L(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_i [y_i \log P(y = 1 | X, \theta) + (1 - y_i) \log(1 - P(y = 1 | X, \theta))]$$

для линейно разделяемых классов

Логистическая регрессия

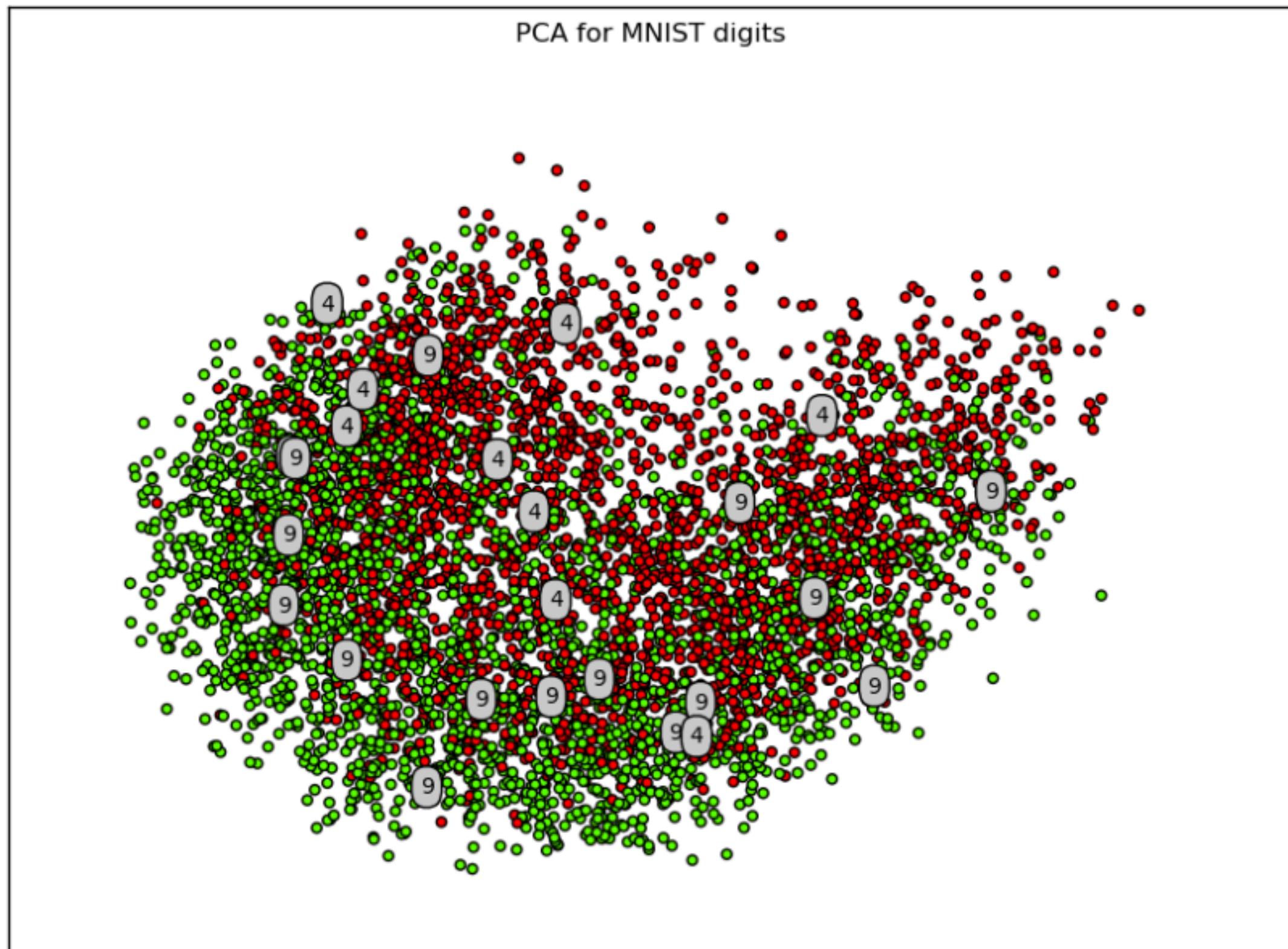


Image credit



Логистическая регрессия

Объекты x_1, \dots, x_n

Ответы $y_1, \dots, y_n \in [0,1]$

Алгоритм предсказания $f(x, \theta) = P(y = 1 | x, \theta) > \frac{1}{2}$



$$P(y = 1 | x, \theta) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Последовательное применение операций:

$$x \longrightarrow \theta^T x \longrightarrow \sigma(\theta^T x) \longrightarrow P(y = 1 | x, \theta)$$

Логистическая регрессия

Объекты x_1, \dots, x_n

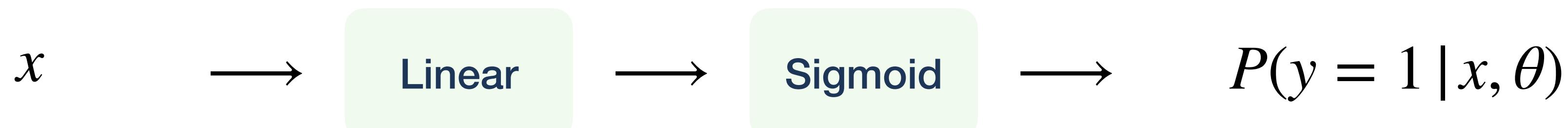
Ответы $y_1, \dots, y_n \in [0,1]$

Алгоритм предсказания $f(x, \theta) = P(y = 1 | x, \theta) > \frac{1}{2}$



$$P(y = 1 | x, \theta) = \sigma(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Последовательное применение операций:



Простейшая нейросеть

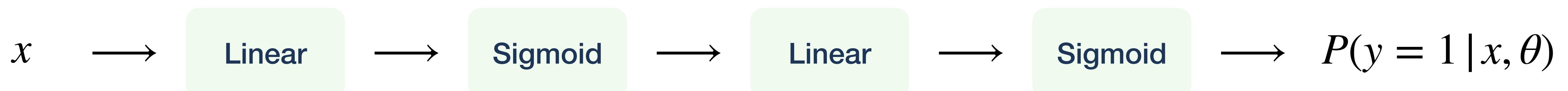
Объекты x_1, \dots, x_n

Ответы $y_1, \dots, y_n \in [0,1]$

Алгоритм предсказания $f(x, \theta) = P(y = 1 | x, \theta) > \frac{1}{2}$



Последовательное применение операций:

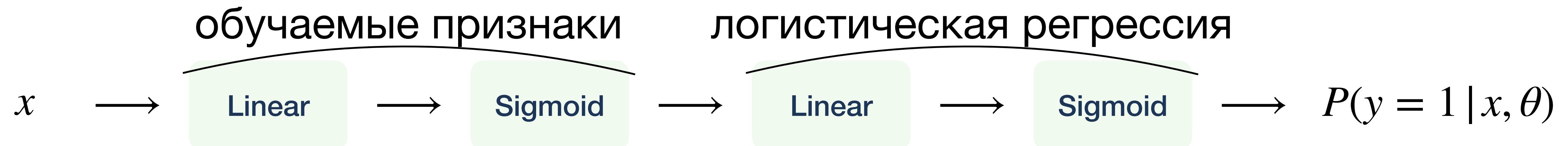


**Нейросеть - это сложная
функция**

Простейшая нейросеть



Простейшая нейросеть



Нейросеть - это сложная функция

Можно не придумывать сложные признаки (feature engineering)

Простейшая нейросеть

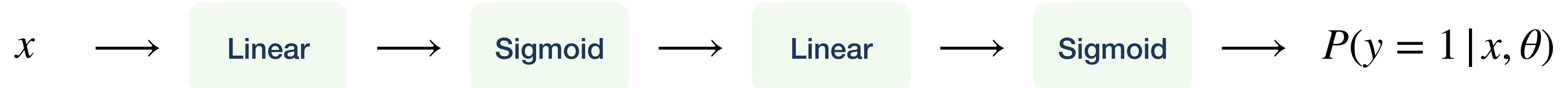


Нейросеть - это сложная функция

Можно представить в виде вычислительного графа:

Блок - слой нейросети, вычисляющий определенную функцию

Простейшая нейросеть



Нейросеть - это сложная функция

Можно представить в виде вычислительного графа:

Блок - слой нейросети, вычисляющий определенную функцию

Linear

линейный слой

$$x_{output} = W \cdot x_{input}$$

параметры: W - матрица размера $d1 \times d2$ - веса линейного слоя

Простейшая нейросеть



Нейросеть - это сложная функция

Можно представить в виде вычислительного графа:

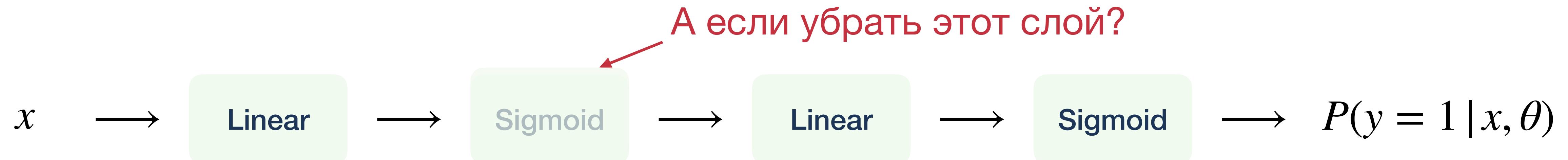
Блок - слой нейросети, вычисляющий определенную функцию

Sigmoid

нелинейность или функция активации (сигмоида)

$$x_{output} = \sigma(x_{input}) = \frac{1}{1 + e^{-x_{input}}}$$

Простейшая нейросеть



Нейросеть - это сложная функция

Можно представить в виде вычислительного графа:

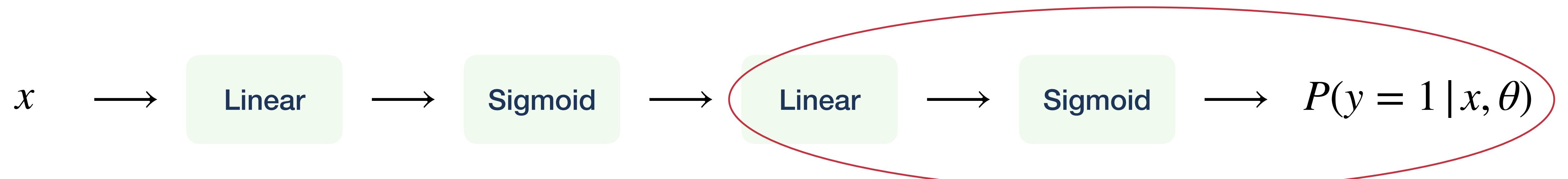
Блок - слой нейросети, вычисляющий определенную функцию

Sigmoid

нелинейность или функция активации (сигмоида)

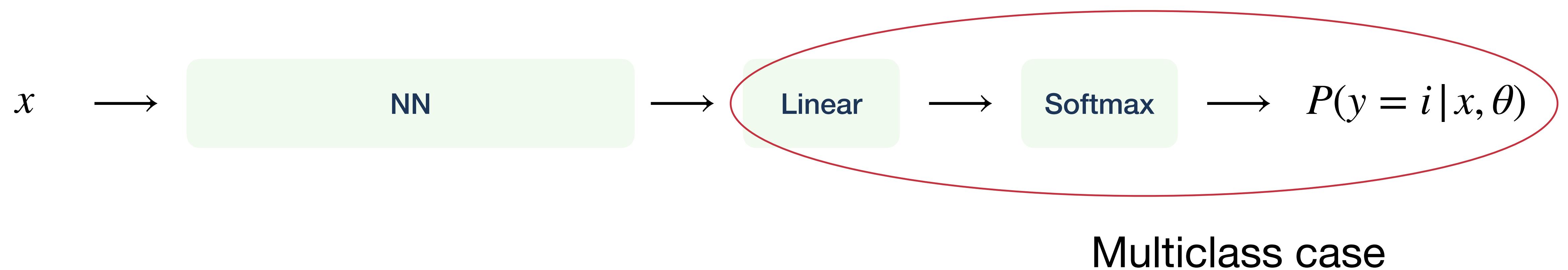
$$x_{output} = \sigma(x_{input}) = \frac{1}{1 + e^{-x_{input}}}$$

Простейшая нейросеть



Эта часть есть в почти всех современных нейросетях!

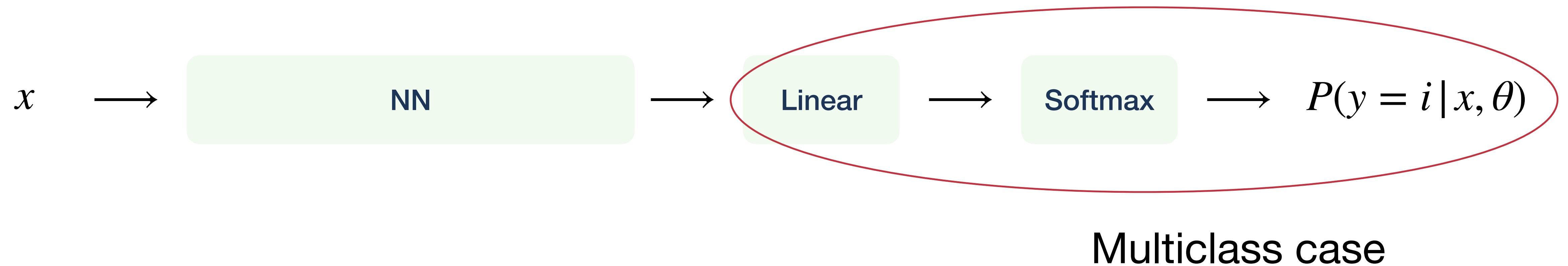
Простейшая нейросеть



Softmax

$$x_{output_i} = \frac{e^{x_{input_i}}}{\sum_j e^{x_{input_j}}}$$

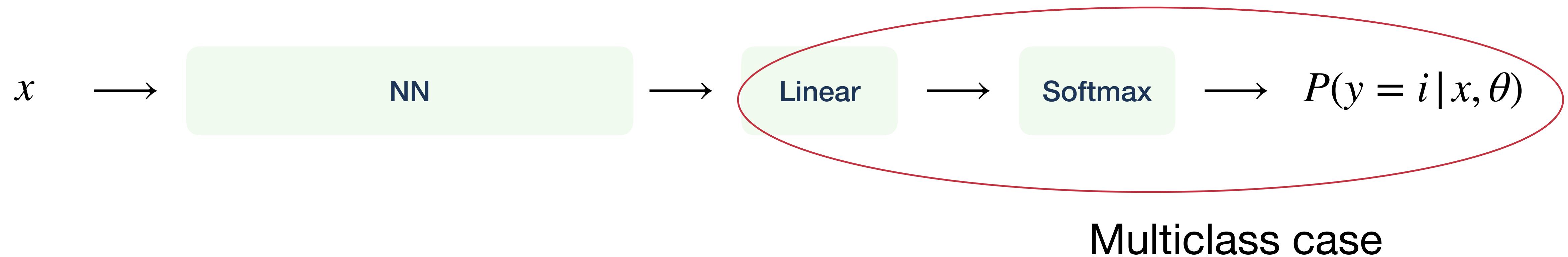
Простейшая нейросеть



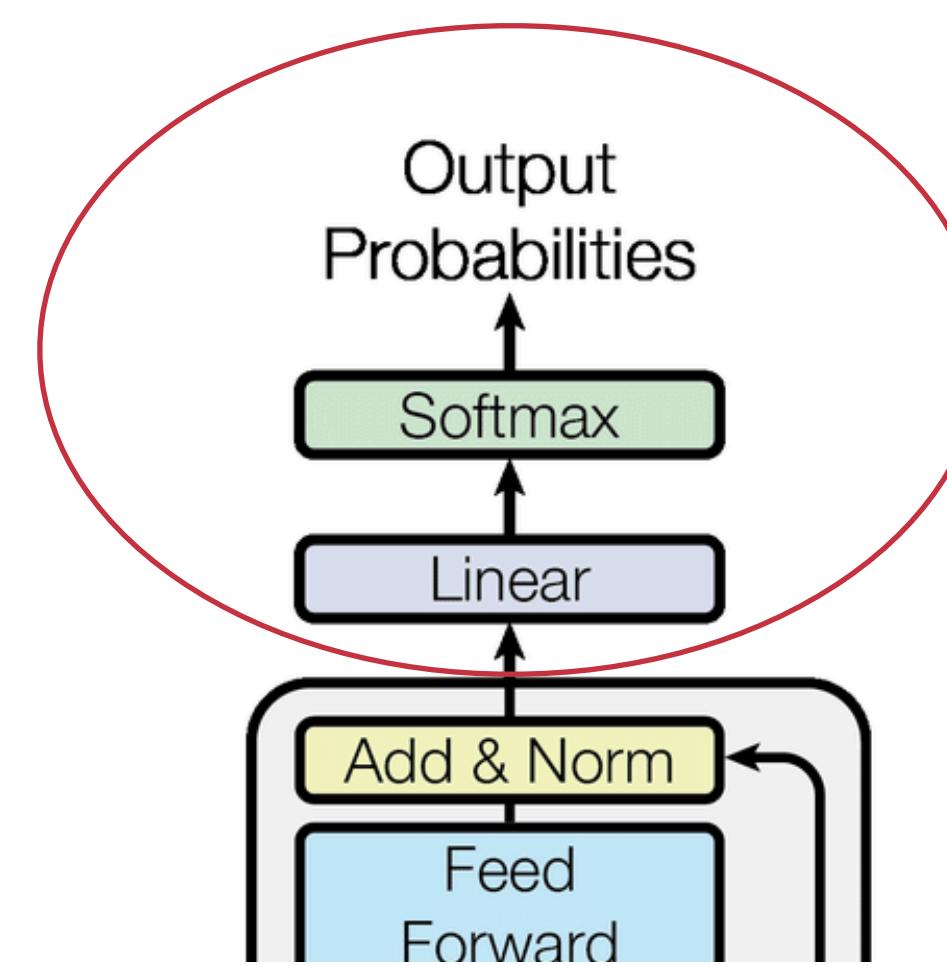
Linear

параметры: W - матрица размера $d \times n_classes$

Простейшая нейросеть



Transformer last layers



Простейшая нейросеть



Нейросеть - это сложная функция

Можно представить в виде вычислительного графа:

Блок - слой нейросети, вычисляющий определенную функцию

Как обучать?

Простейшая нейросеть



Нейросеть - это сложная функция

Можно представить в виде вычислительного графа:

Блок - слой нейросети, вычисляющий определенную функцию

Как обучать?

$$\frac{1}{n} \sum_i L(y_i, f(x_i, \theta)) \rightarrow \min_{\theta}$$

задача оптимизации

Простейшая нейросеть



Нейросеть - это сложная функция

Можно представить в виде вычислительного графа:

Блок - слой нейросети, вычисляющий определенную функцию

Как обучать?

Градиентный спуск:

$$\frac{1}{n} \sum_i L(y_i, f(x_i, \theta)) \rightarrow \min_{\theta}$$
$$\theta_{t+1} = \theta_t - \alpha \frac{dL}{d\theta}$$

задача оптимизации

методы оптимизации
1го порядка

Простейшая нейросеть



Нейросеть - это сложная функция

Можно представить в виде вычислительного графа:

Блок - слой нейросети, вычисляющий определенную функцию

Как обучать?

Градиентный спуск:

$$\frac{1}{n} \sum_i L(y_i, f(x_i, \theta)) \rightarrow \min_{\theta}$$
$$\theta_{t+1} = \theta_t - \alpha \frac{dL}{d\theta}$$

Как посчитать градиенты?

Обучение нейросети

$$\frac{1}{n} \sum_i L(y_i, f(x_i, \theta)) \rightarrow \min_{\theta}$$

Численные методы:

$$\frac{df(x, \theta)}{d\theta} = \frac{f(x, \theta + eps) - f(x, \theta - eps)}{2eps}$$

Градиентный спуск:

$$\theta_{t+1} = \theta_t - \alpha \frac{dL}{d\theta}$$

- конечные разности

Обучение нейросети

$$\frac{1}{n} \sum_i L(y_i, f(x_i, \theta)) \rightarrow \min_{\theta}$$

Численные методы:

$$\frac{df(x, \theta)}{d\theta} = \frac{f(x, \theta + eps) - f(x, \theta - eps)}{2eps}$$

Градиентный спуск:

$$\theta_{t+1} = \theta_t - \alpha \frac{dL}{d\theta}$$

- конечные разности
хороший unit test

Обучение нейросети

$$\frac{1}{n} \sum_i L(y_i, f(x_i, \theta)) \rightarrow \min_{\theta}$$

Градиентный спуск:

$$\theta_{t+1} = \theta_t - \alpha \frac{dL}{d\theta}$$

Обычно блоки дифференцируемые - можно посчитать “на бумаге”

Обучение нейросети

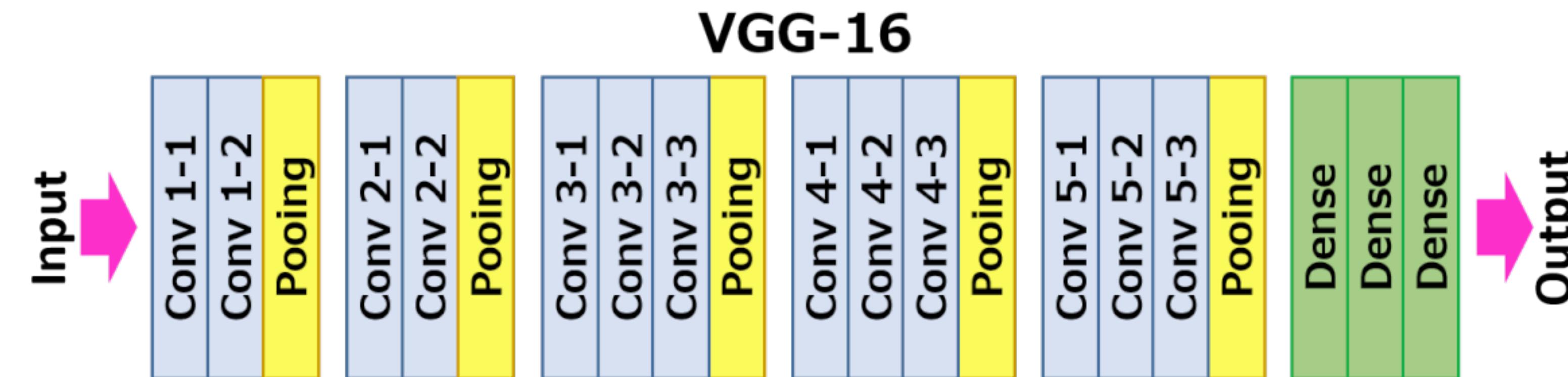
$$\frac{1}{n} \sum_i L(y_i, f(x_i, \theta)) \rightarrow \min_{\theta}$$

Градиентный спуск:

$$\theta_{t+1} = \theta_t - \alpha \frac{dL}{d\theta}$$

Обычно блоки дифференцируемые - можно посчитать “на бумаге”

БЛОКОВ МОЖЕТ БЫТЬ ОЧЕНЬ МНОГО



Backpropagation

$$\frac{1}{n} \sum_i L(y_i, f(x_i, \theta)) \rightarrow \min_{\theta}$$

Градиентный спуск:

$$\theta_{t+1} = \theta_t - \alpha \frac{dL}{d\theta}$$

Chain rule (дифференцирование сложной функции): $\frac{dL}{d\theta} = \frac{dL}{dz} \frac{dz}{d\theta}$

Backpropagation: простой пример

Дано: $f(x, y, z) = (x + y)z$

Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

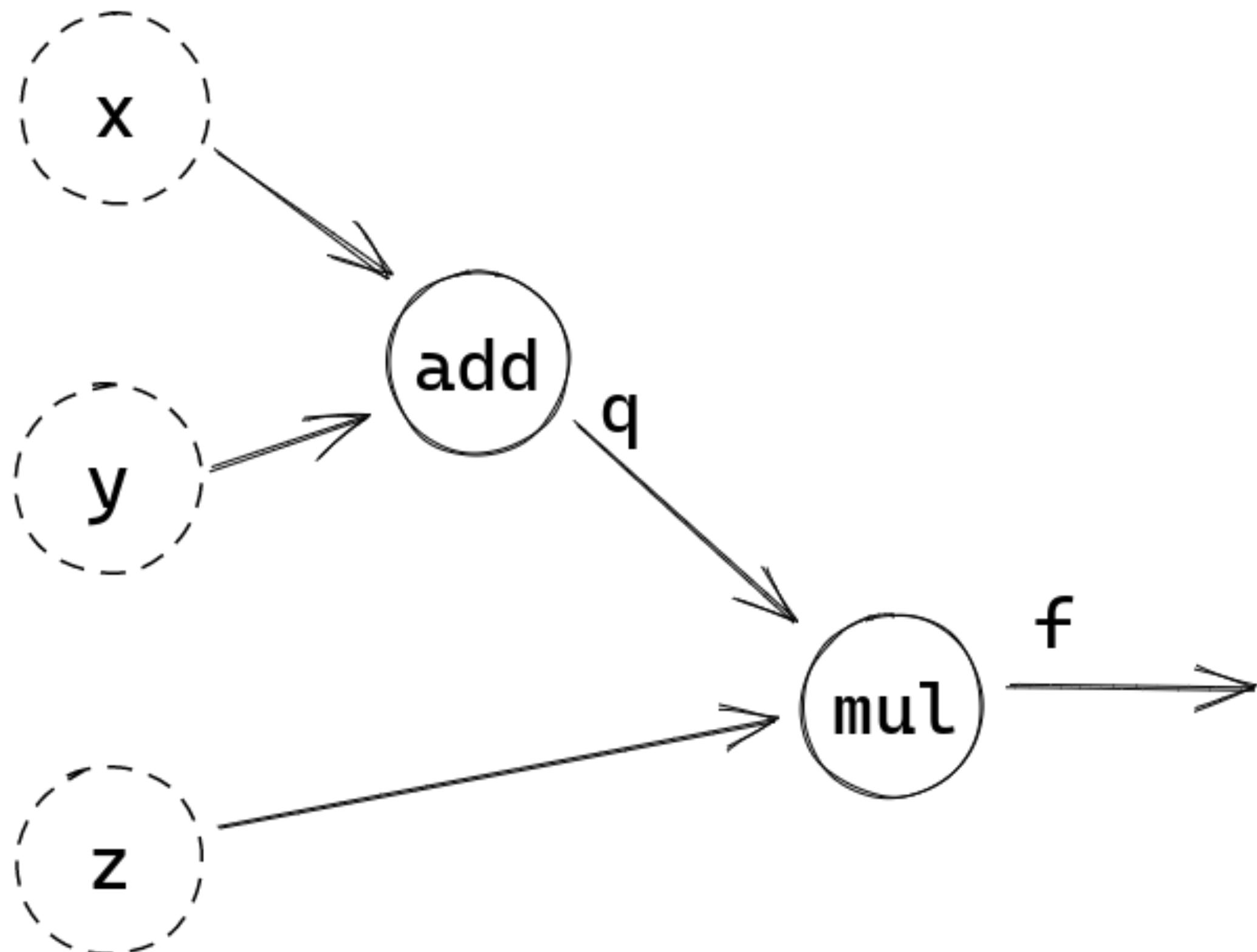
Backpropagation: простой пример

Дано: $f(x, y, z) = (x + y)z$

Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y$$

$$f = qz$$



Backpropagation: простой пример

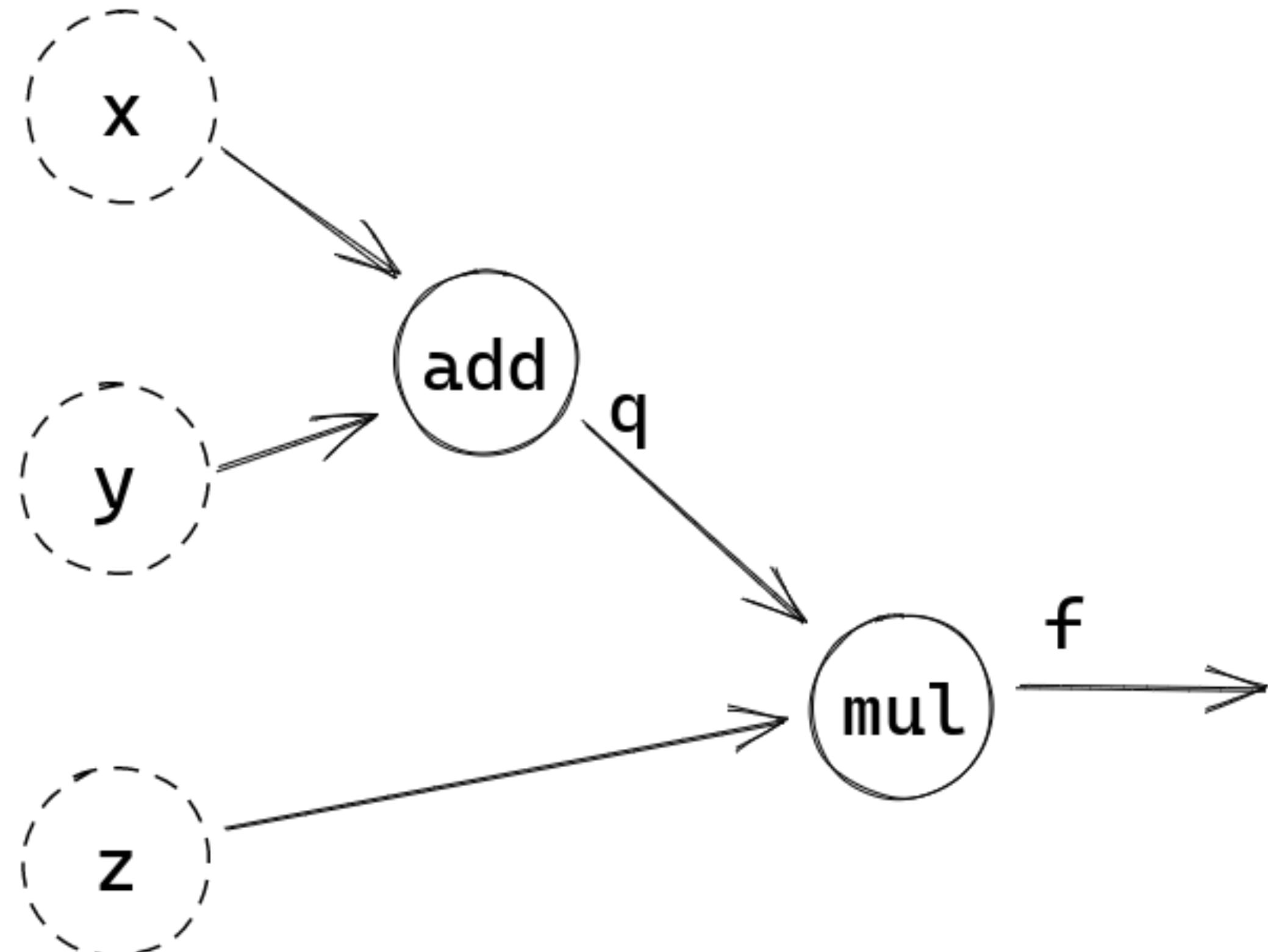
Дано: $f(x, y, z) = (x + y)z$

Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y$$

$$f = qz$$

$$x = 1, y = -2, z = 3$$



Backpropagation: простой пример

Дано: $f(x, y, z) = (x + y)z$

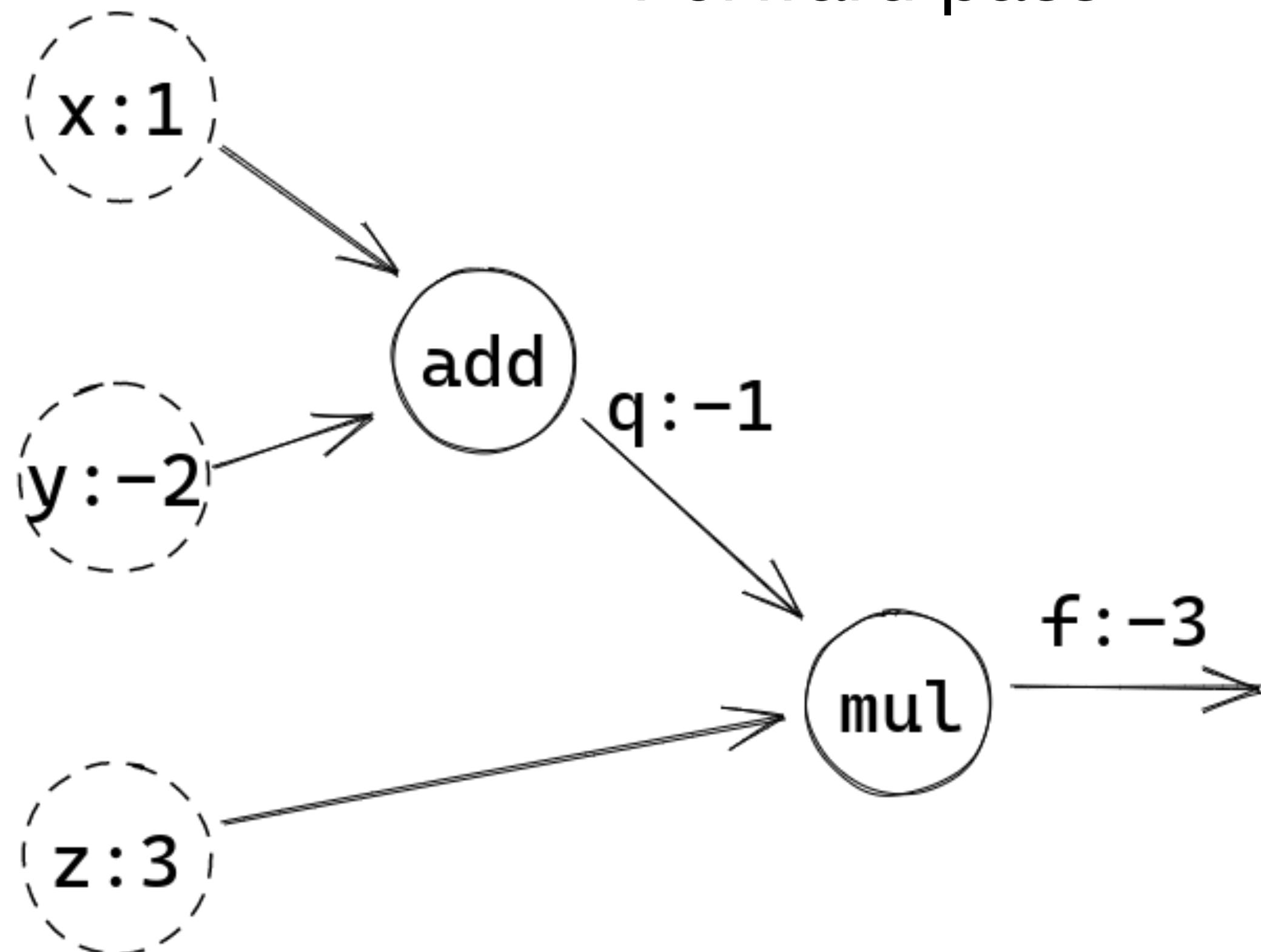
Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y = 1 - 2 = -1$$

$$f = qz = -1 \cdot 3 = -3$$

$$x = 1, y = -2, z = 3$$

Forward pass



Backpropagation: простой пример

Дано: $f(x, y, z) = (x + y)z$

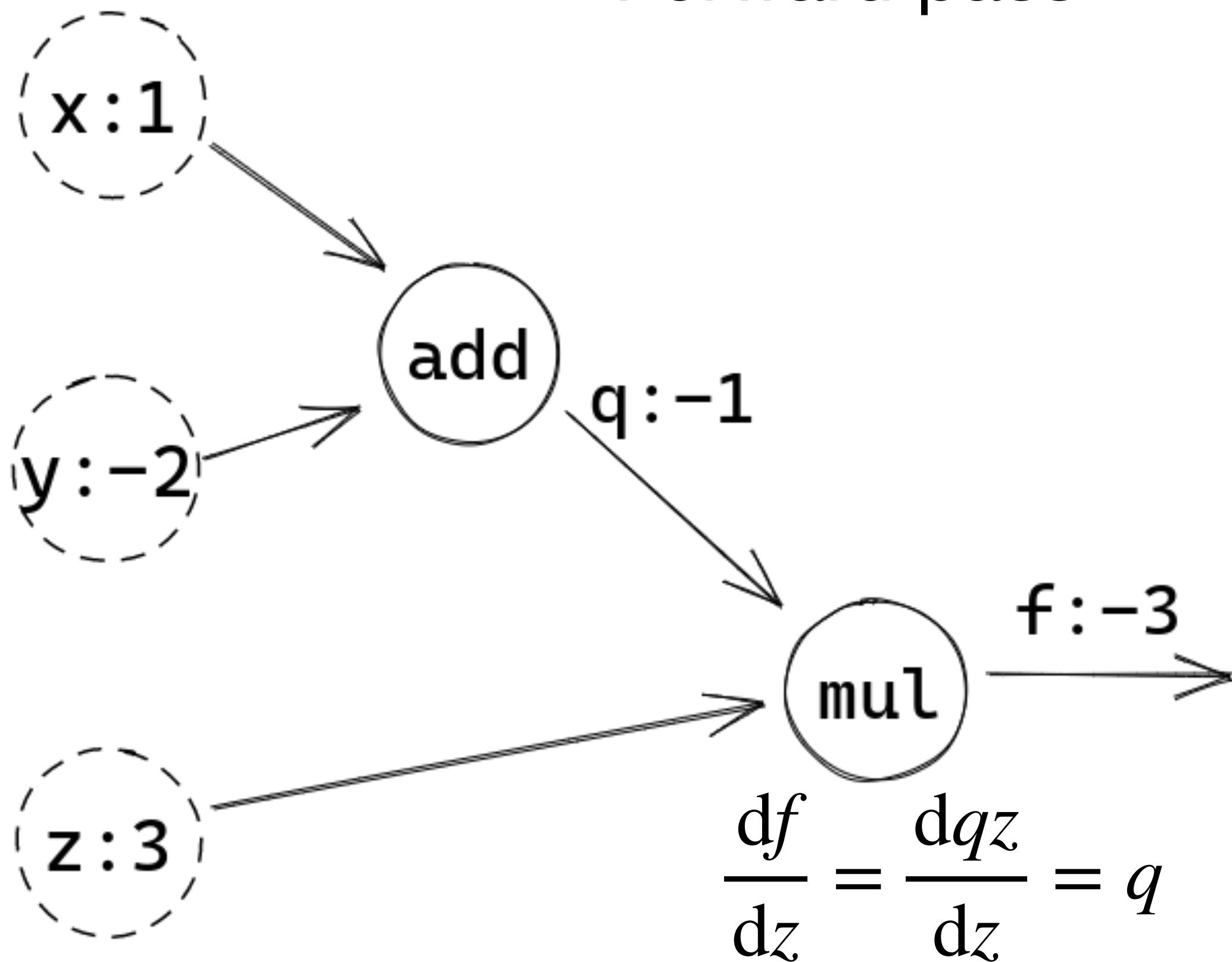
Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y = 1 - 2 = -1$$

$$f = qz = -1 \cdot 3 = -3$$

$$x = 1, y = -2, z = 3$$

Forward pass



Backpropagation: простой пример

Дано: $f(x, y, z) = (x + y)z$

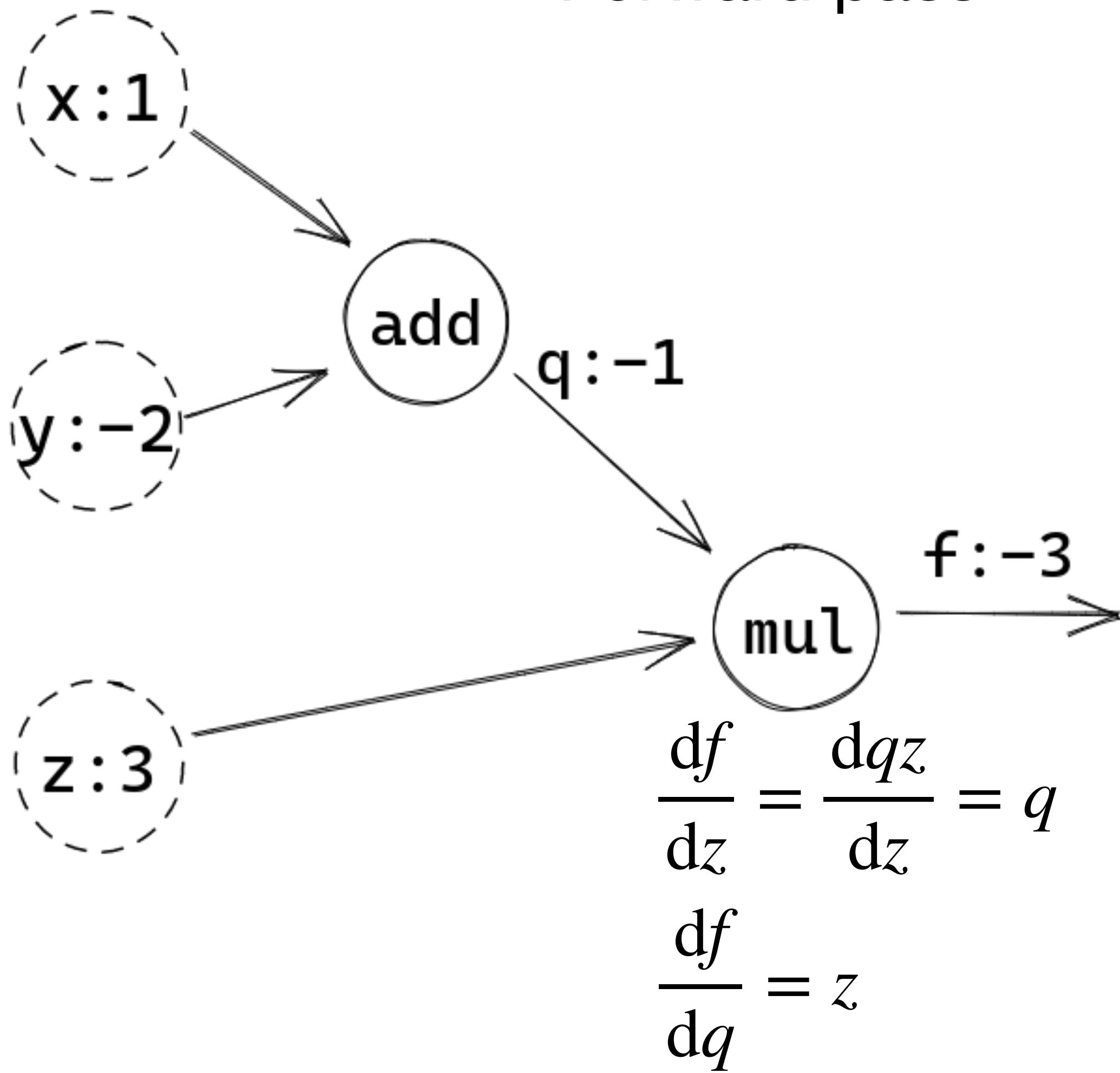
Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y = 1 - 2 = -1$$

$$f = qz = -1 \cdot 3 = -3$$

$$x = 1, y = -2, z = 3$$

Forward pass



Backpropagation: простой пример

Дано: $f(x, y, z) = (x + y)z$

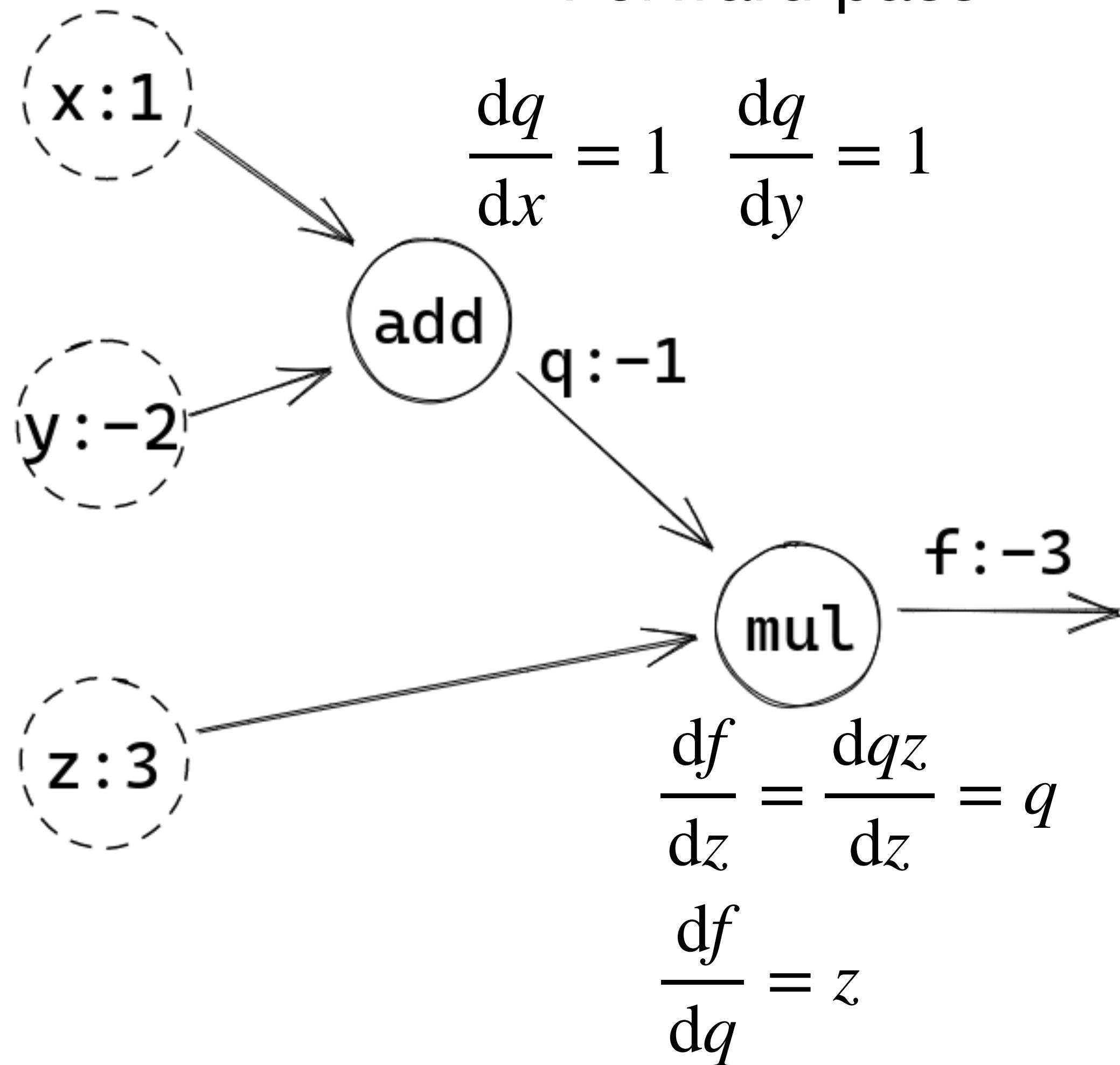
Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y = 1 - 2 = -1$$

$$f = qz = -1 \cdot 3 = -3$$

$$x = 1, y = -2, z = 3$$

Forward pass



Backpropagation: простой пример

Дано: $f(x, y, z) = (x + y)z$

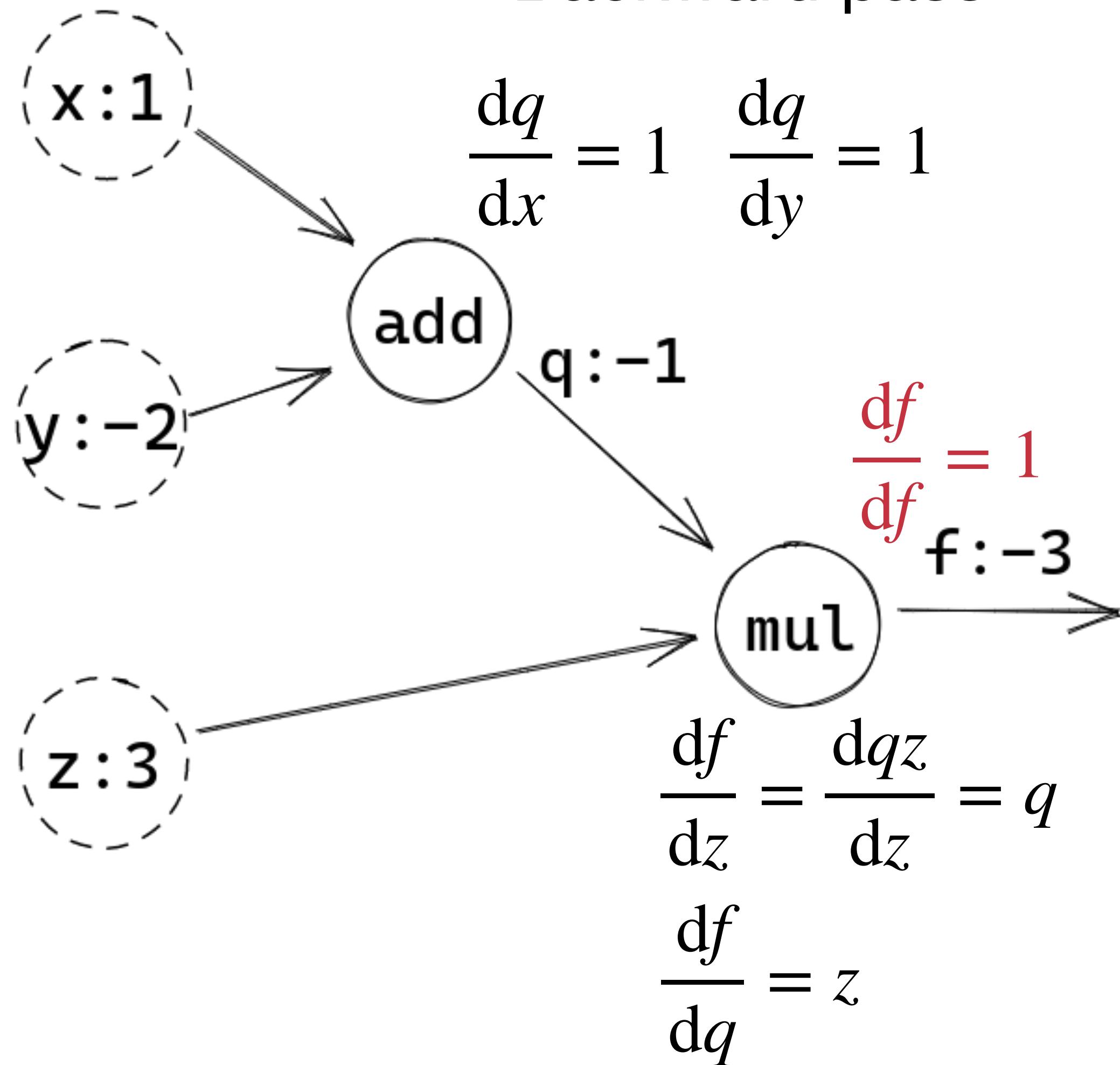
Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y = 1 - 2 = -1$$

$$f = qz = -1 \cdot 3 = -3$$

$$x = 1, y = -2, z = 3$$

Backward pass



Backpropagation: простой пример

Дано: $f(x, y, z) = (x + y)z$

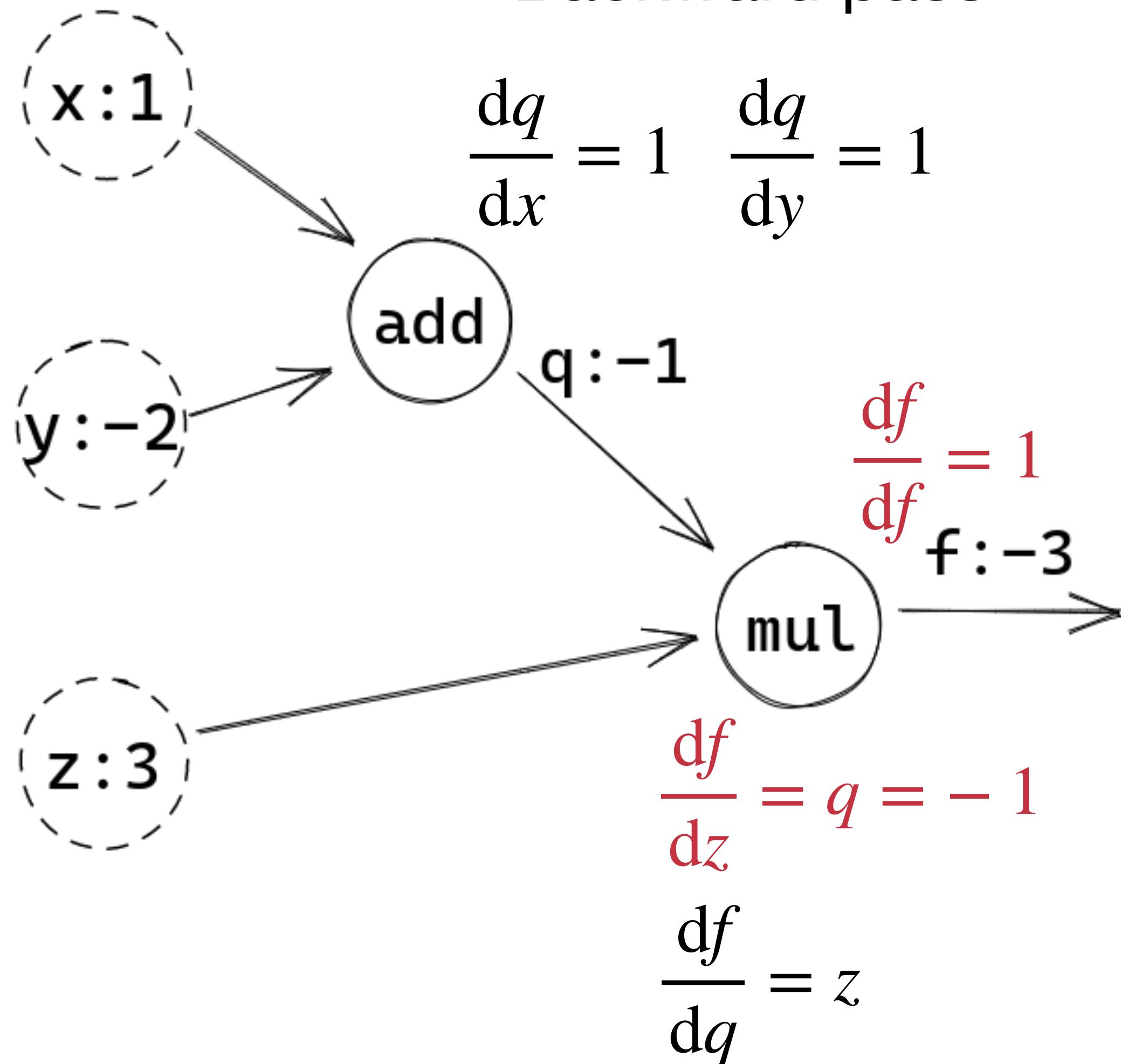
Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y = 1 - 2 = -1$$

$$f = qz = -1 \cdot 3 = -3$$

$$x = 1, y = -2, z = 3$$

Backward pass



Backpropagation: простой пример

Дано: $f(x, y, z) = (x + y)z$

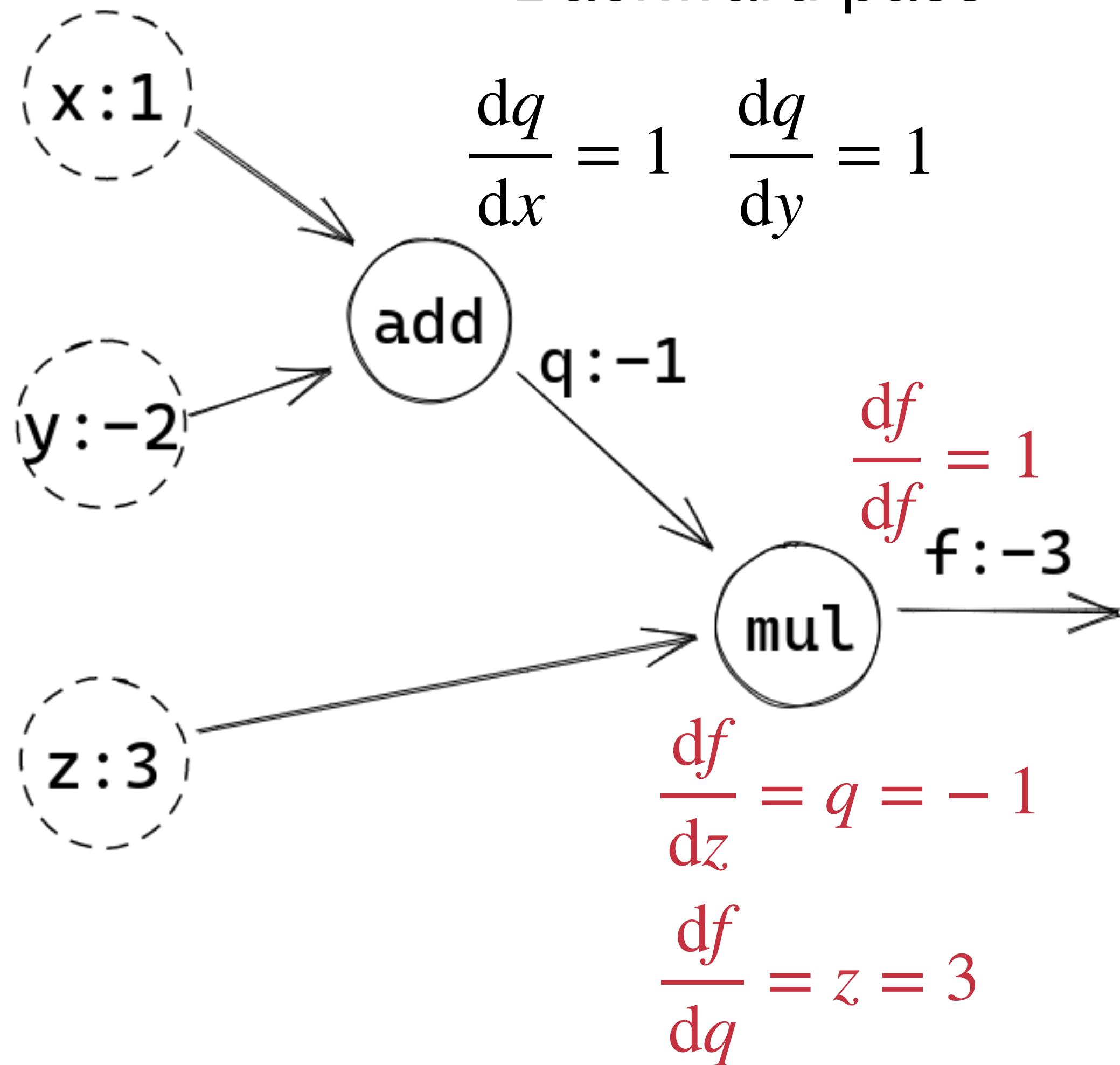
Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y = 1 - 2 = -1$$

$$f = qz = -1 \cdot 3 = -3$$

$$x = 1, y = -2, z = 3$$

Backward pass



Backpropagation: простой пример

Дано: $f(x, y, z) = (x + y)z$

Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y = 1 - 2 = -1$$

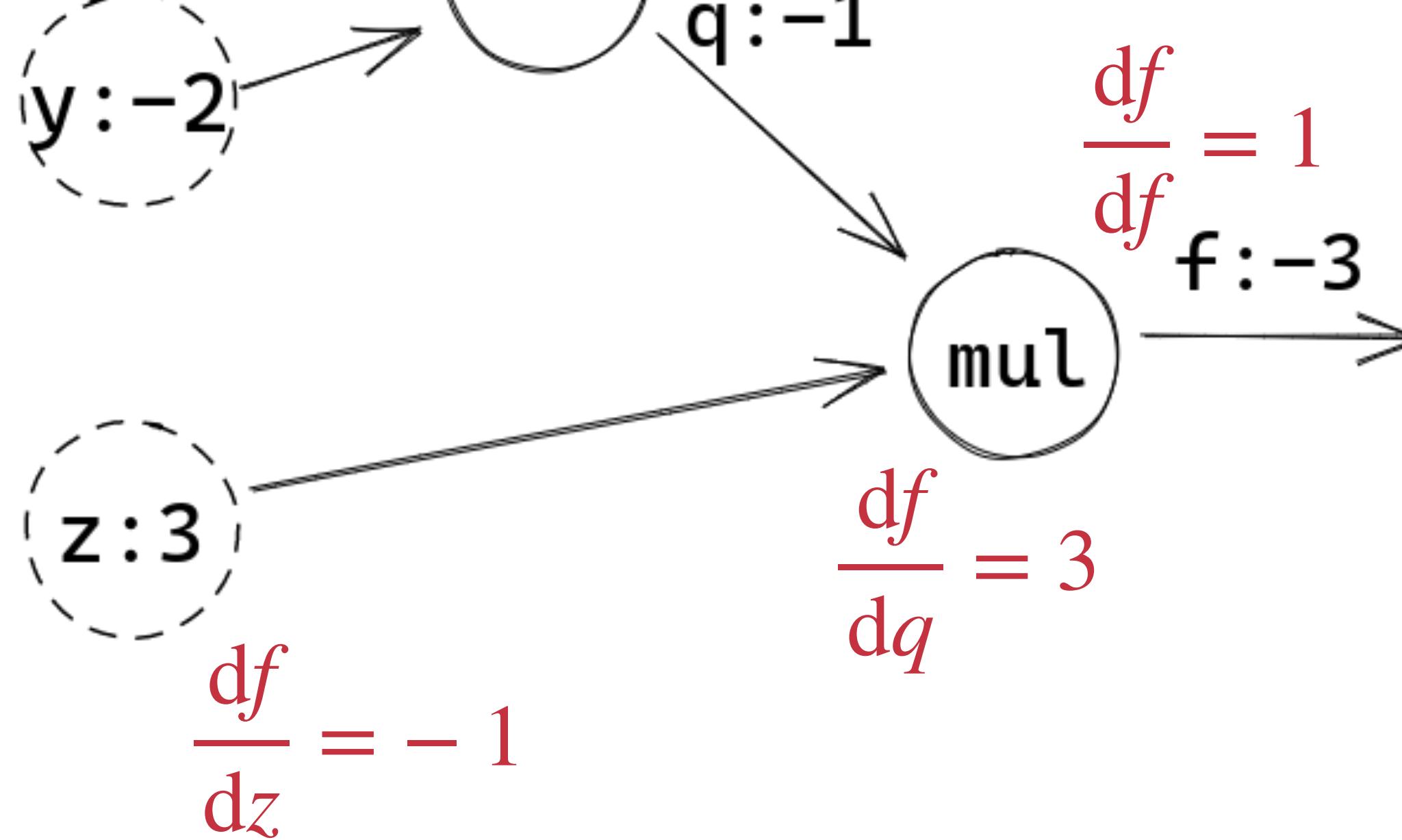
$$f = qz = -1 \cdot 3 = -3$$

$$x = 1, y = -2, z = 3$$

$$\frac{df}{dx} = \frac{df}{dq} \frac{dq}{dx}$$

Backward pass

$$\frac{dq}{dx} = 1 \quad \frac{dq}{dy} = 1$$



Backpropagation: простой пример

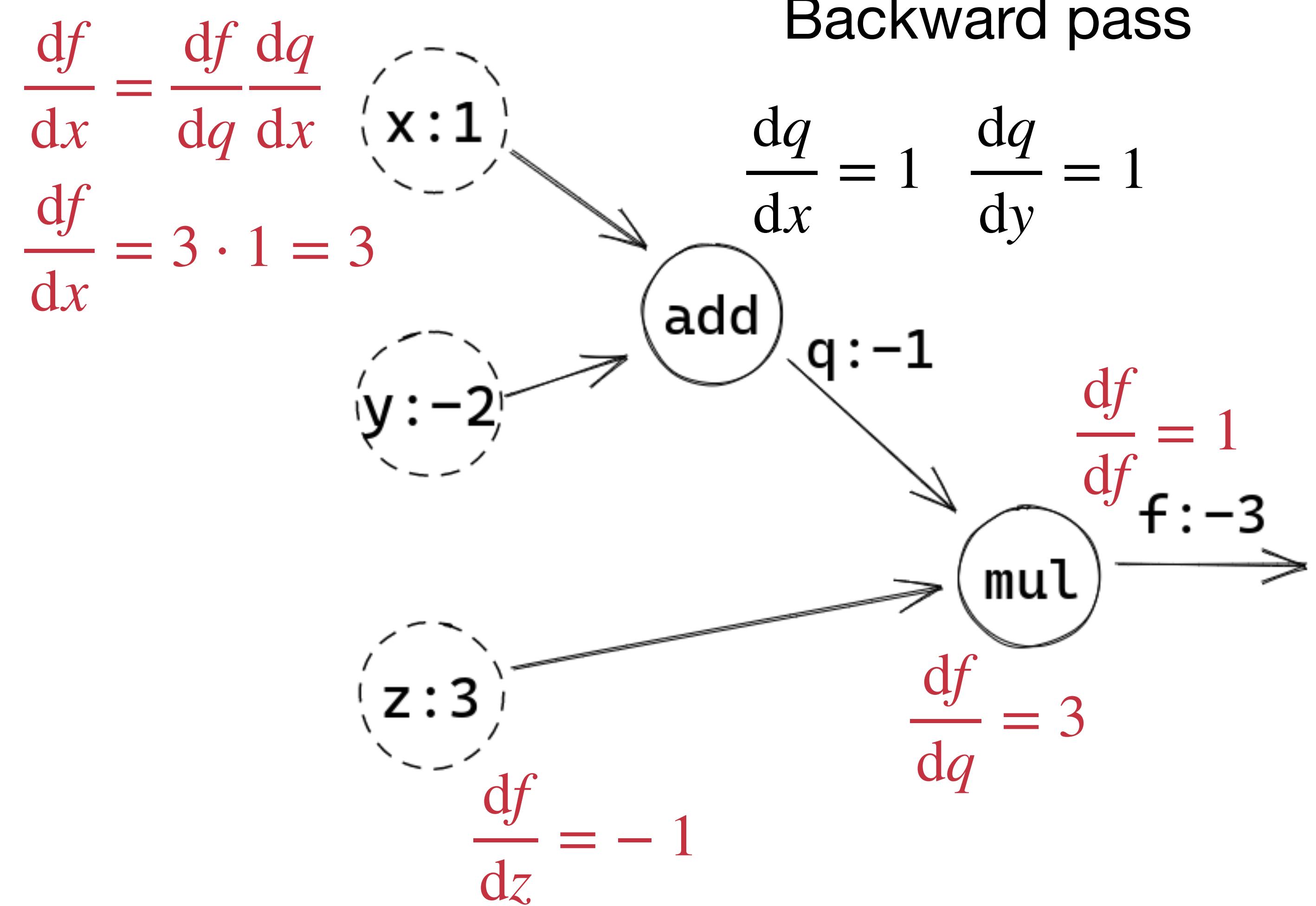
Дано: $f(x, y, z) = (x + y)z$

Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y = 1 - 2 = -1$$

$$f = qz = -1 \cdot 3 = -3$$

$$x = 1, y = -2, z = 3$$



Backpropagation: простой пример

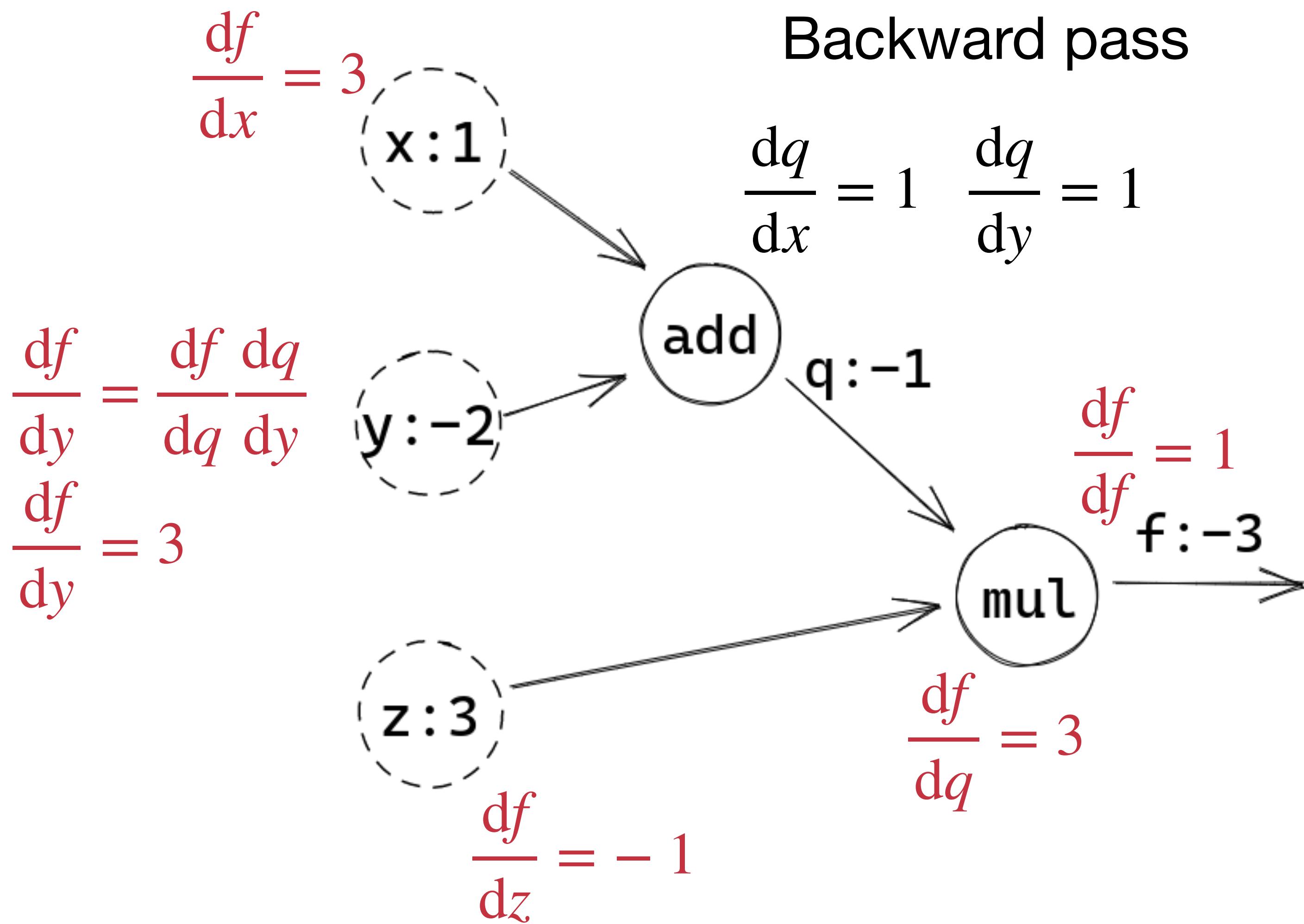
Дано: $f(x, y, z) = (x + y)z$

Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y = 1 - 2 = -1$$

$$f = qz = -1 \cdot 3 = -3$$

$$x = 1, y = -2, z = 3$$



Backpropagation: простой пример

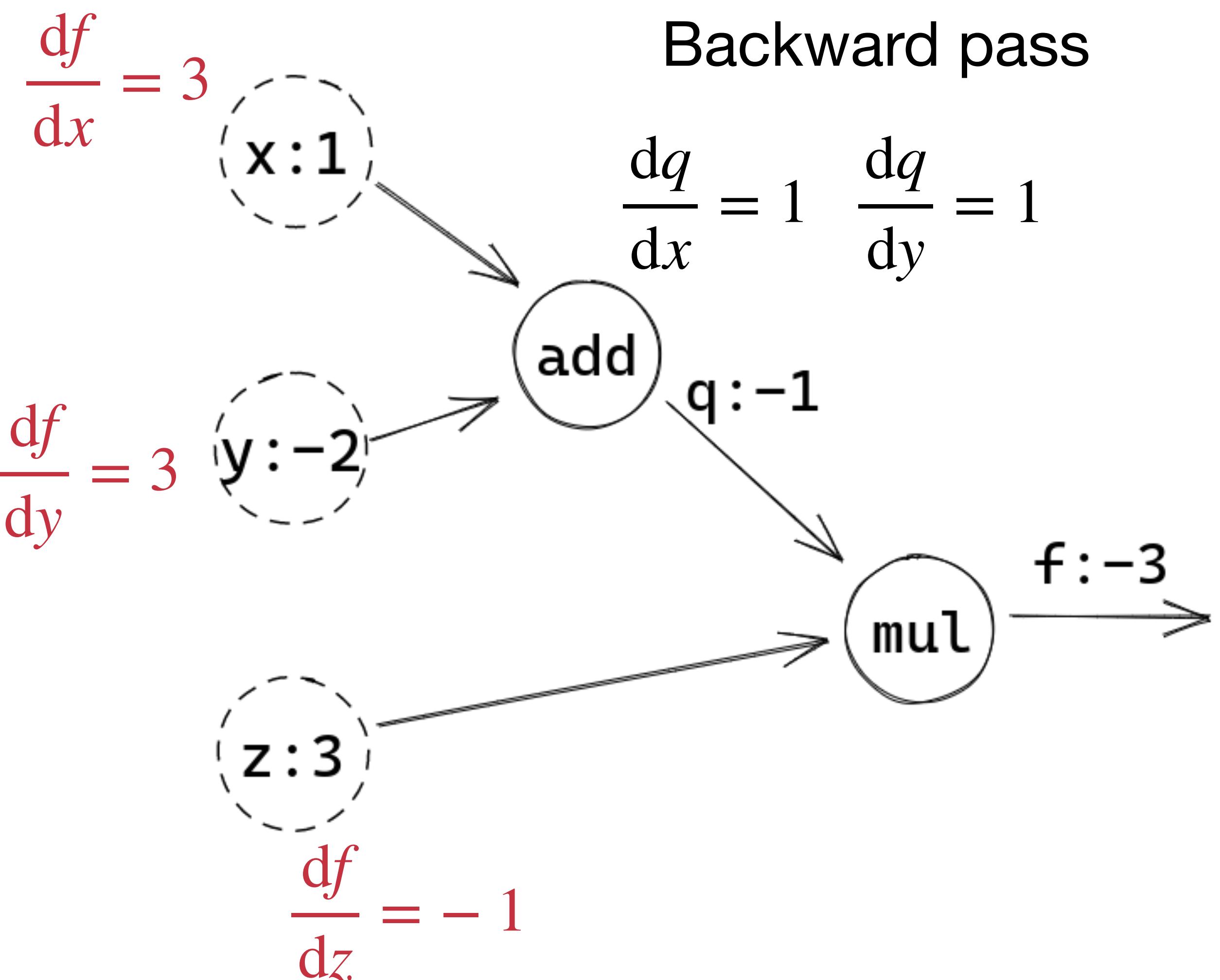
Дано: $f(x, y, z) = (x + y)z$

Вычислить: $\frac{df}{dx}, \frac{df}{dy}, \frac{df}{dz}$

$$q = x + y = 1 - 2 = -1$$

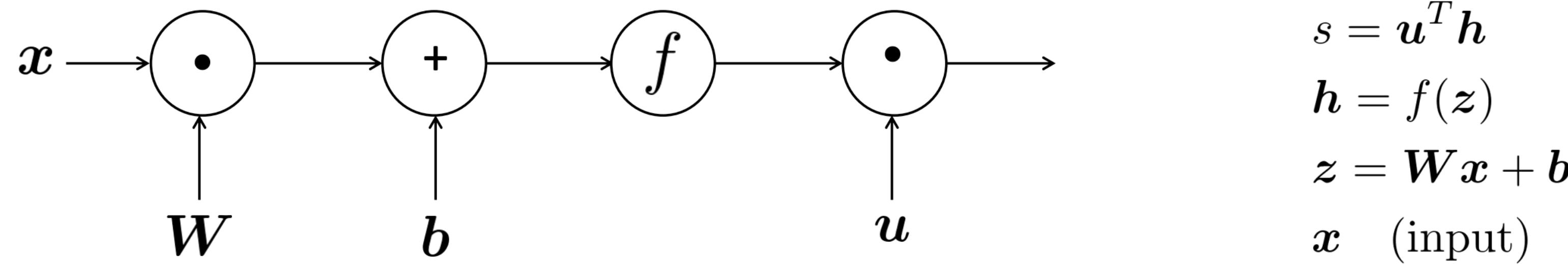
$$f = qz = -1 \cdot 3 = -3$$

$$x = 1, y = -2, z = 3$$



Backpropagation в нейросетях

Представление нейросети в виде вычислительного графа



[Image credit](#)

Backpropagation в нейросетях

Представление нейросети в виде вычислительного графа

Forward pass: вычисление результатов всех операций

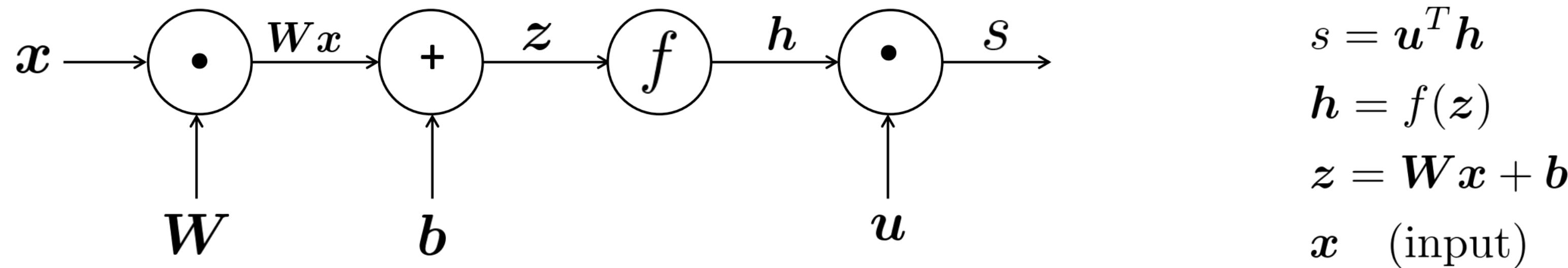


Image credit

Backpropagation в нейросетях

Представление нейросети в виде вычислительного графа

Forward pass: вычисление результатов всех операций

Backward pass: вычисление всех градиентов

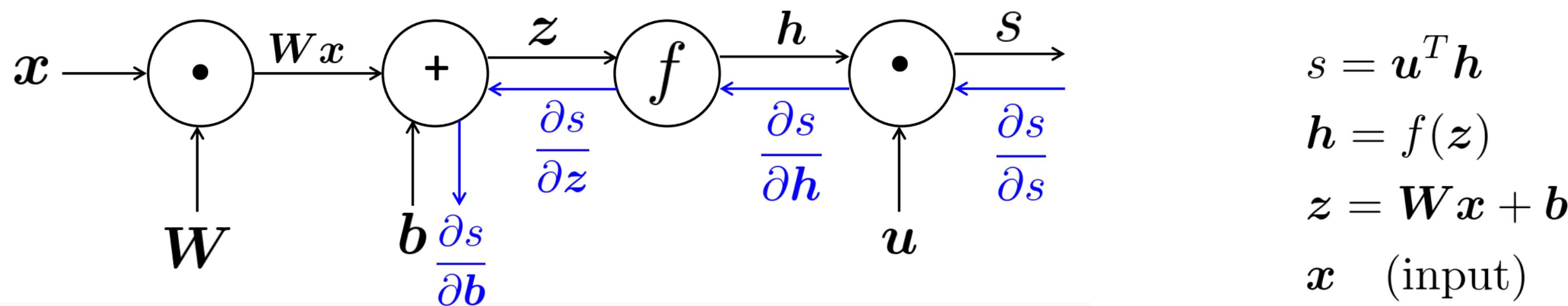
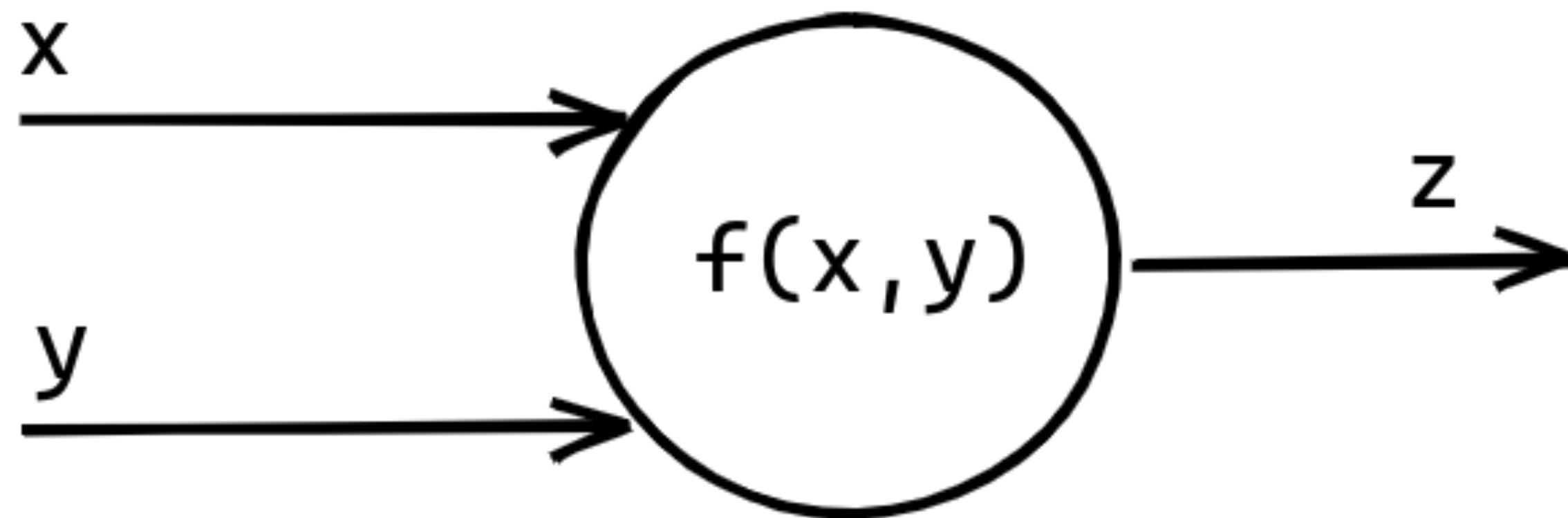


Image credit

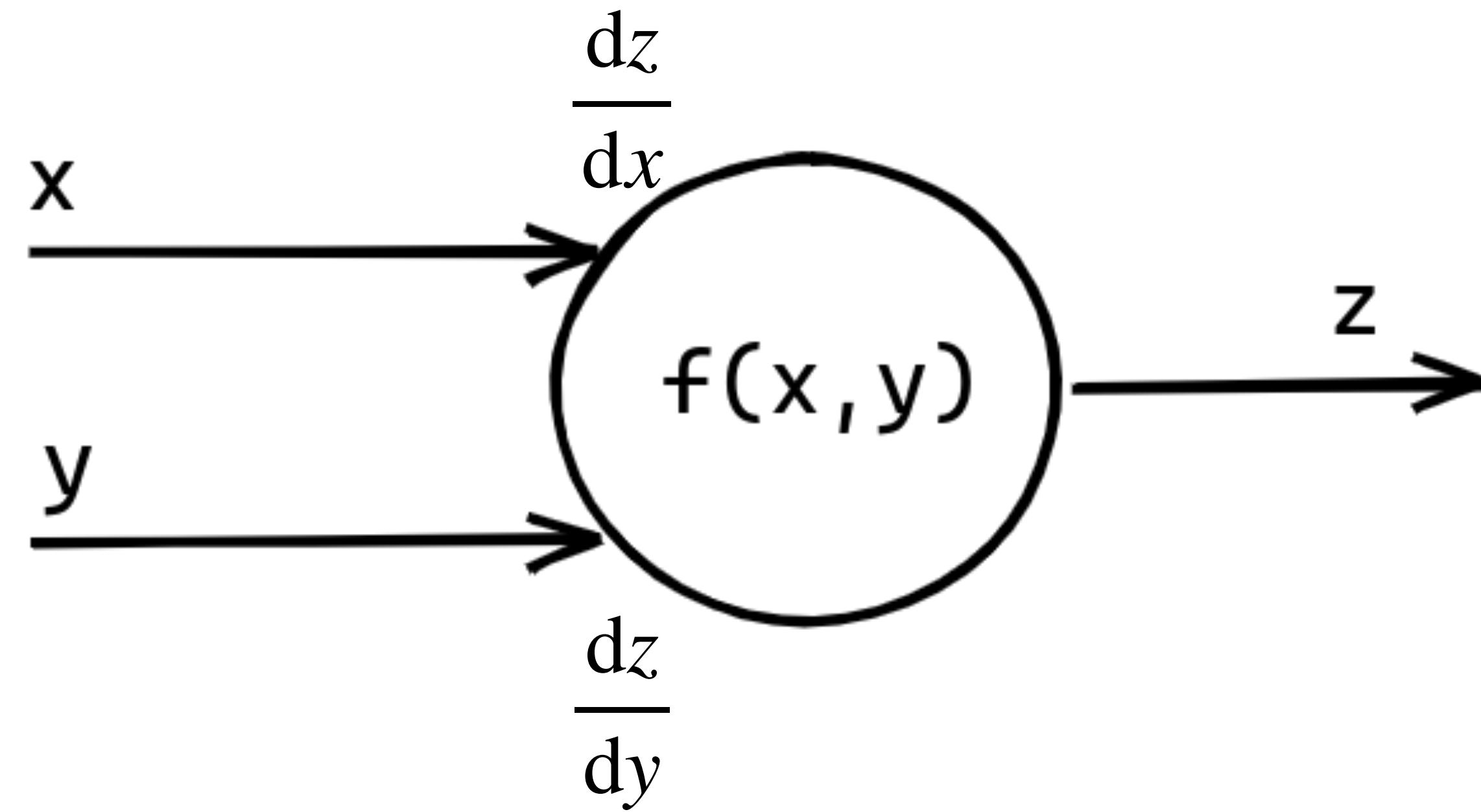
Backpropagation: одна вершина

$$z = f(x, y)$$



Backpropagation: одна вершина

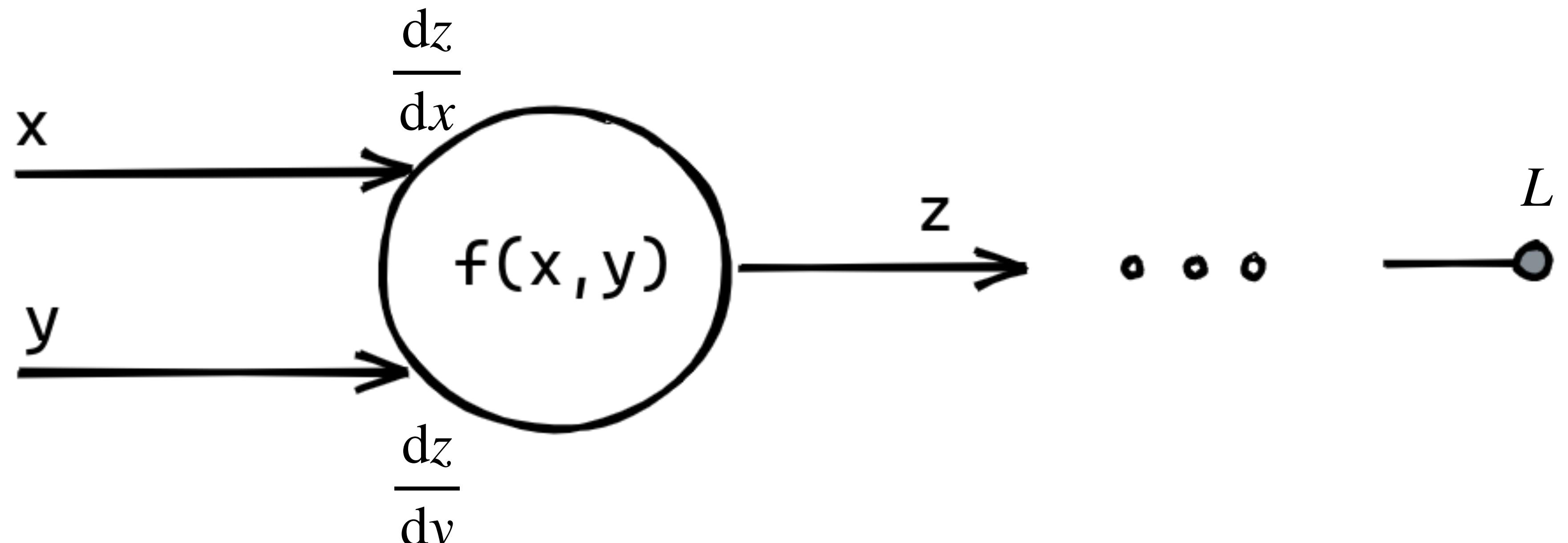
$$z = f(x, y)$$



Forward pass: вычисление z , сохраняем локальные градиенты

Backpropagation: одна вершина

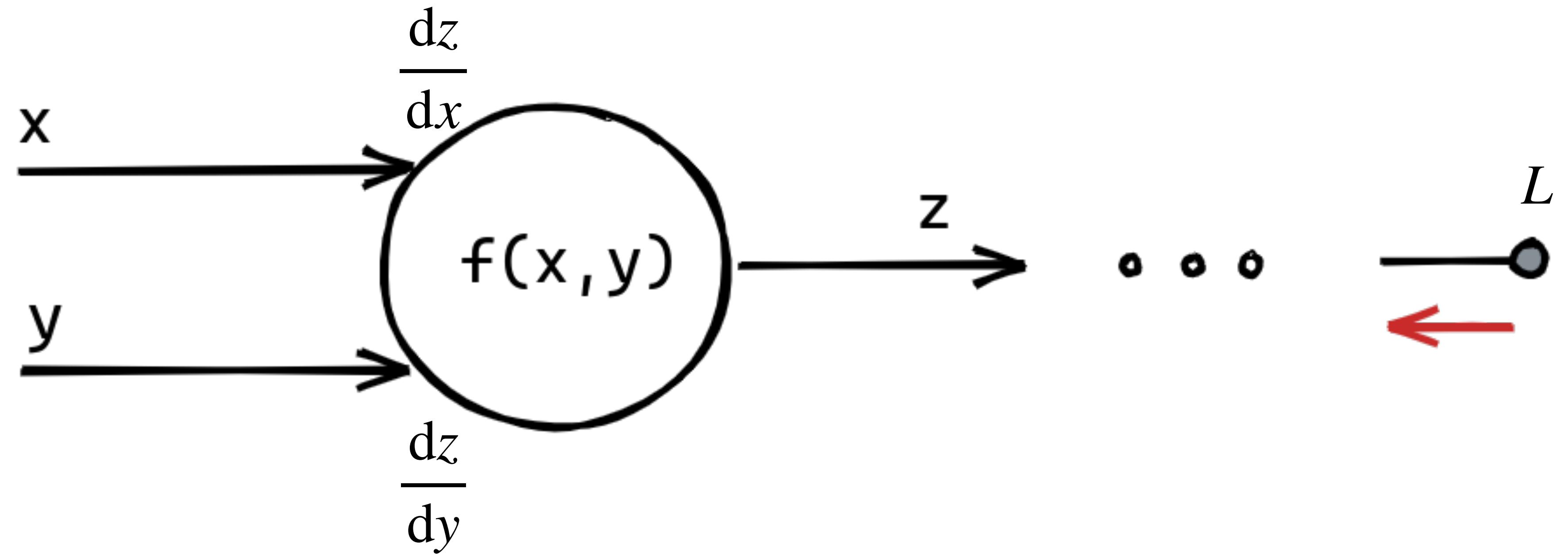
$$z = f(x, y)$$



...посчитался лосс

Backpropagation: одна вершина

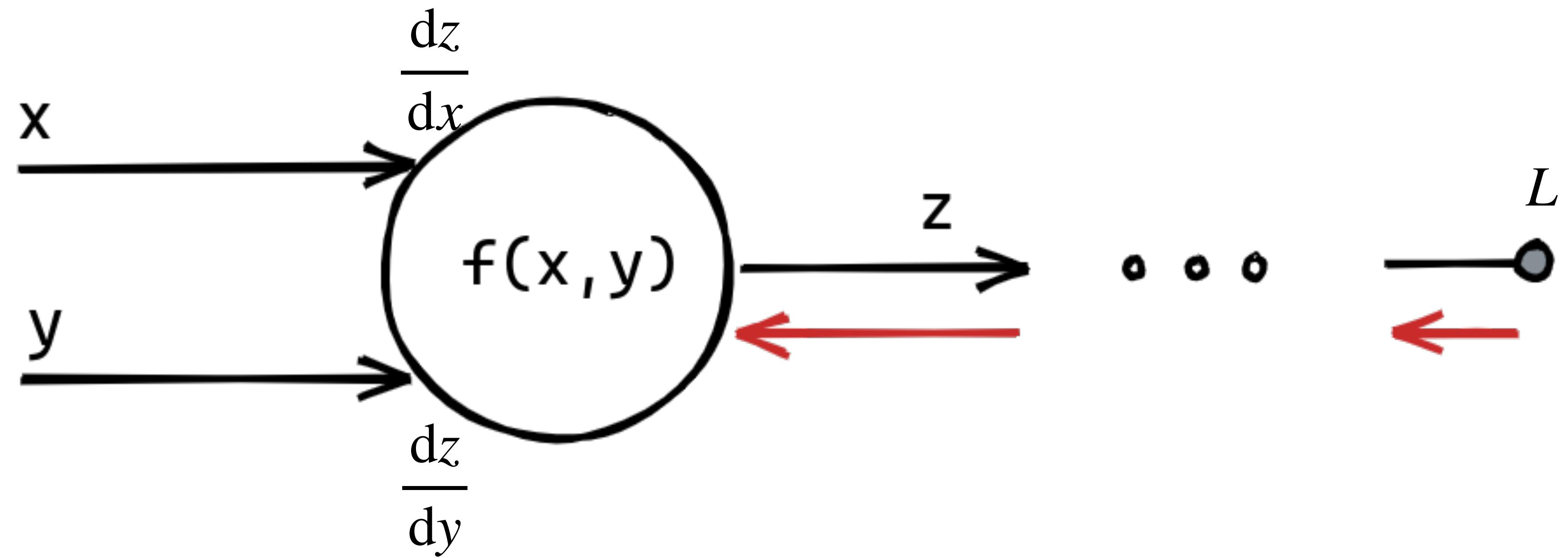
$$z = f(x, y)$$



Backward pass: итеративно вычисляем глобальные градиенты

Backpropagation: одна вершина

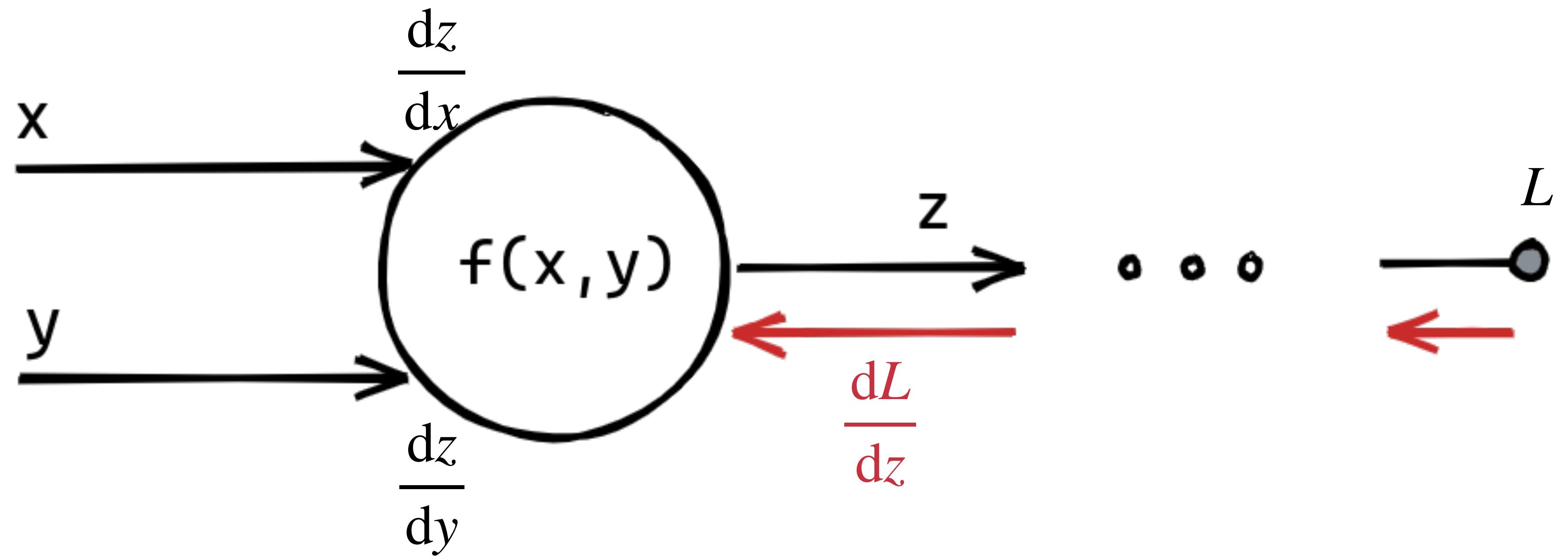
$$z = f(x, y)$$



Backward pass: итеративно вычисляем глобальные градиенты

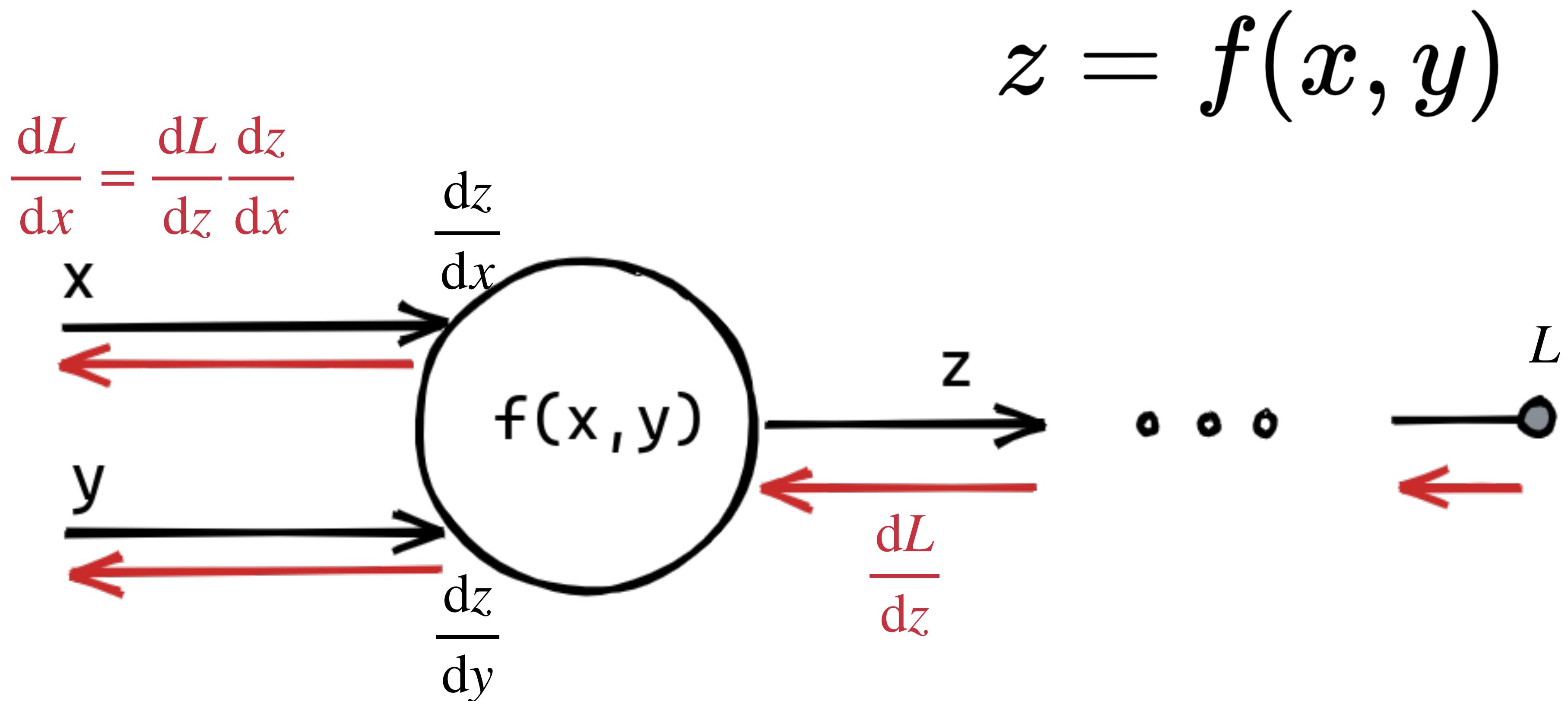
Backpropagation: одна вершина

$$z = f(x, y)$$



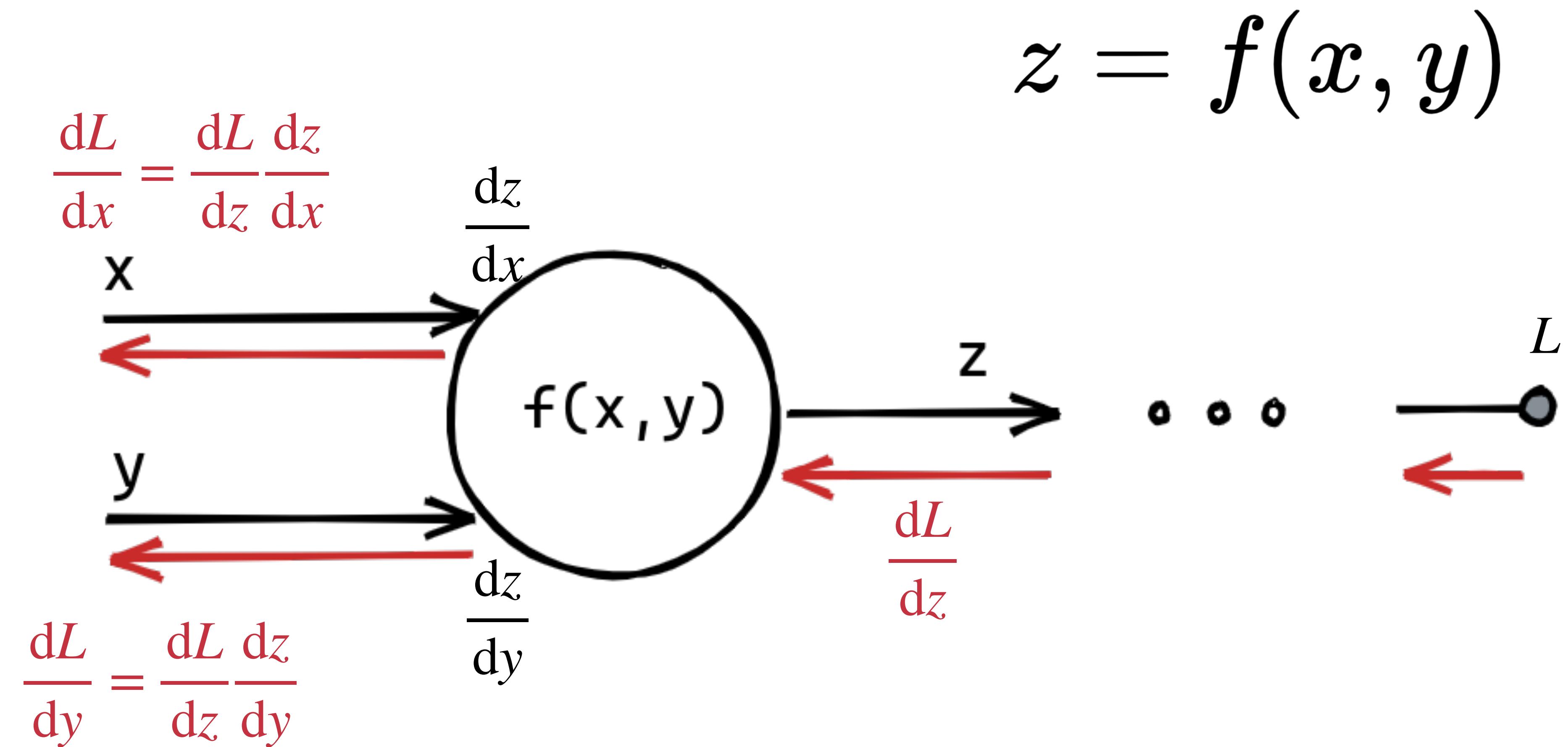
Backward pass: итеративно вычисляем глобальные градиенты

Backpropagation: одна вершина



Backward pass: итеративно вычисляем глобальные градиенты

Backpropagation: одна вершина



Backward pass: итеративно вычисляем глобальные градиенты

Backpropagation: как считать?

Можно оперировать векторной формой

$$x \in \mathbb{R}, y \in \mathbb{R}$$

$$\frac{dy}{dx} \in \mathbb{R}$$

Производная

$$x \in \mathbb{R}^n, y \in \mathbb{R}$$

$$\frac{dy}{dx} \in \mathbb{R}^n$$

$$\left(\frac{dy}{dx} \right)_i = \frac{dy}{dx_i}$$

Градиент

$$x \in \mathbb{R}^n, y \in \mathbb{R}^m$$

$$\frac{dy}{dx} \in \mathbb{R}^{n \times m}$$

$$\left(\frac{dy}{dx} \right)_{i,j} = \frac{dy_j}{dx_i}$$

Якобиан

Backpropagation: как считать?

Можно оперировать векторной формой

$$x \in \mathbb{R}, y \in \mathbb{R}$$

$$\frac{dy}{dx} \in \mathbb{R}$$

Производная

$$x \in \mathbb{R}^n, y \in \mathbb{R}$$

$$\frac{dy}{dx} \in \mathbb{R}^n$$

$$\left(\frac{dy}{dx} \right)_i = \frac{dy}{dx_i}$$

Градиент

$$x \in \mathbb{R}^n, y \in \mathbb{R}^m$$

$$\frac{dy}{dx} \in \mathbb{R}^{n \times m}$$

$$\left(\frac{dy}{dx} \right)_{i,j} = \frac{dy_j}{dx_i}$$

Якобиан

DL frameworks

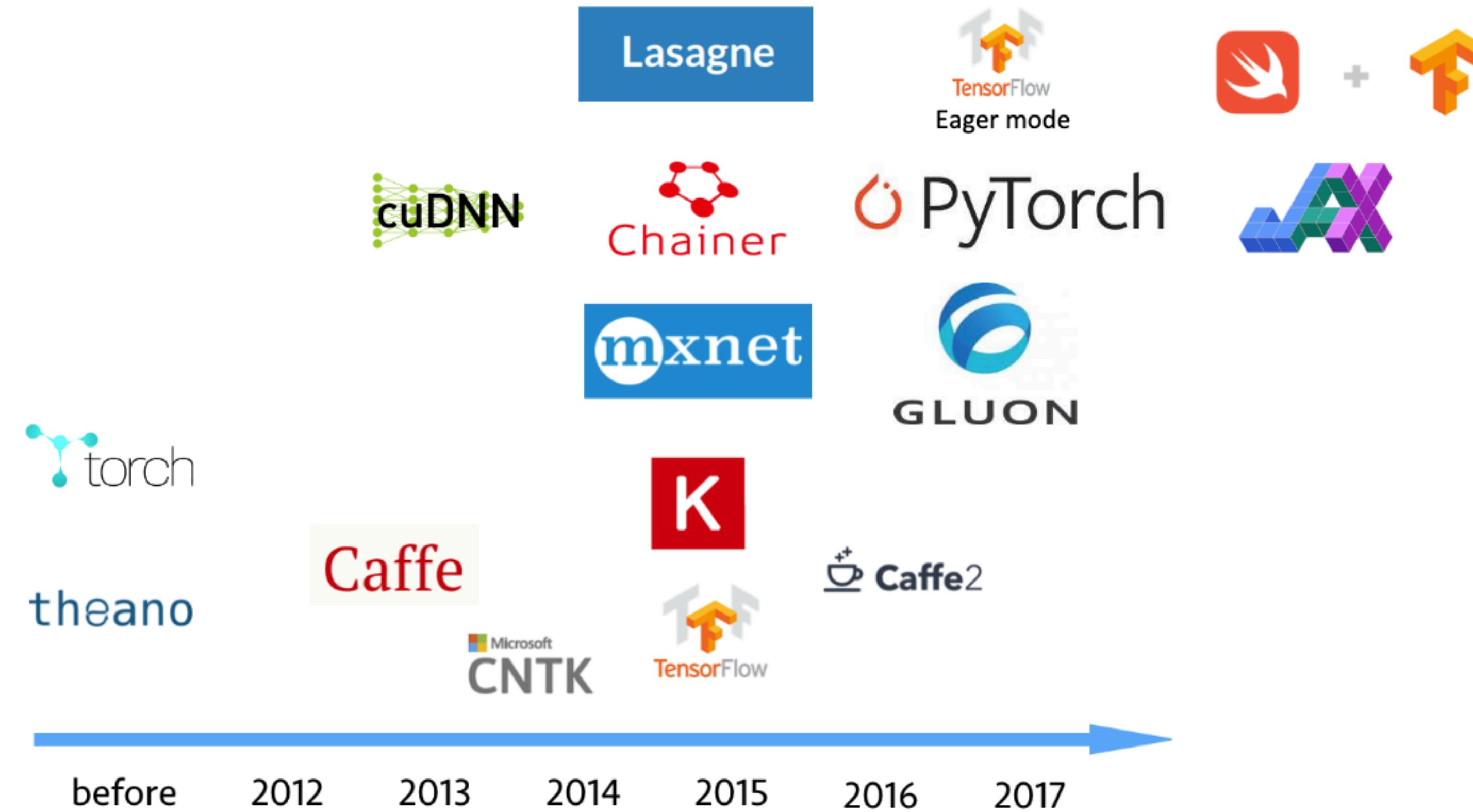


Image credit

DL frameworks



Josh Tobin
@josh_tobin_

Why do people always ask what ML framework to use? It's easy:

- jax is for researchers
- pytorch is for engineers
- tensorflow is for boomers

...



Josh Tobin @josh_tobin_ · 12 map.
Keras is for infants

...



Josh Tobin @josh_tobin_ · 12 map.
mxnet is for no one

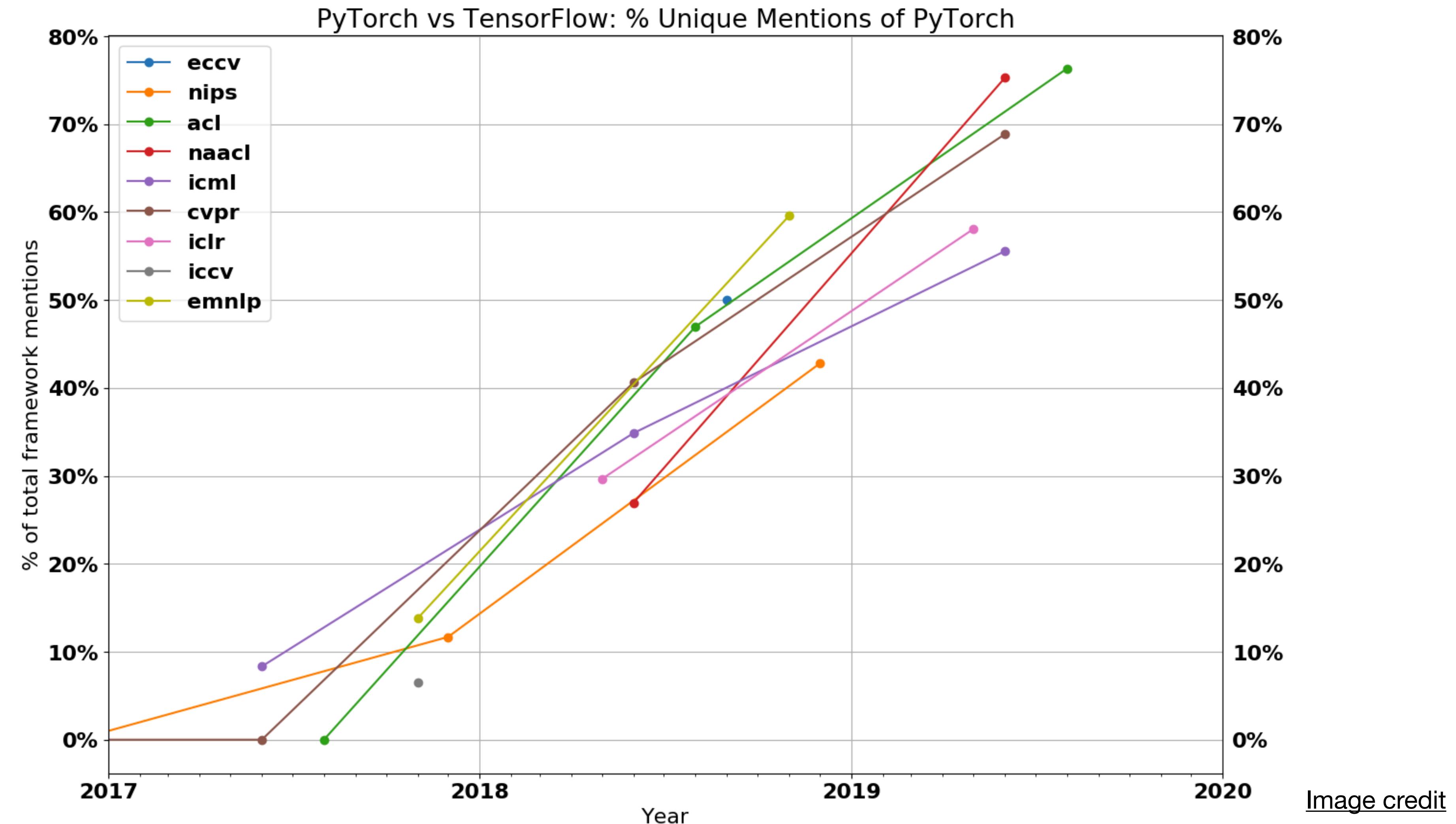
...



Josh Tobin @josh_tobin_ · 12 map.
matlab is for professors

...

DL frameworks



Итог

- Нейросеть - сложная функция
- Обучение - методы оптимизации 1го порядка
 - Градиентный спуск
- Backpropagation - способ подсчета градиентов в нейросетях
- Linear + Softmax = вероятности классов на выходе нейросети