Solingen GEOINFORMATION

Kreuzentropie

-- Verlustfunktion in Neuronales Netzwerk

SD62-1 Dr. Lingquan Zhao

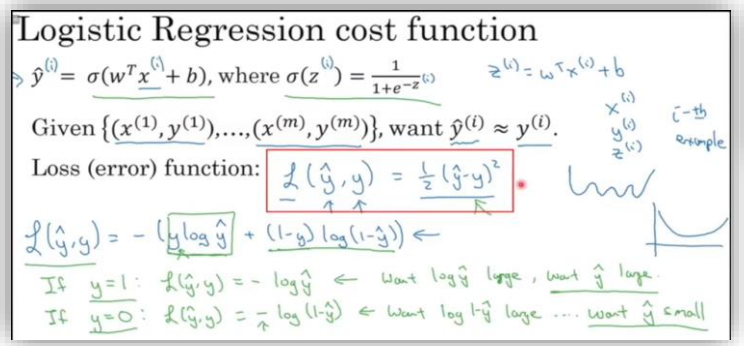
26.06.2024

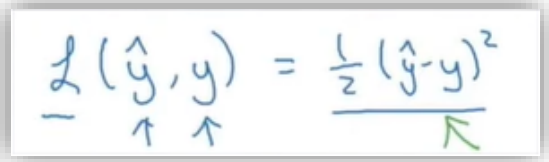
Wie die Verlustfunktion gestaltet wird?

$L(y,\hat{y})$

- **Methode der kleinsten Quadrate**
- Maximum-Likelihood-Schätzung

Methode der kleinsten Quadrate





Prof. Andrew Ng, 2018

$$L_i(y, \hat{y}) = \min \sum_{i=1}^{n} | y - \hat{y} |$$

$$L(y, \hat{y}) = \min \sum_{i=1}^{n} \frac{1}{2}(y - \hat{y})^2$$

$$L_{ii}(y, \hat{y}) = \min \sum_{i=1}^{n} (y - \hat{y})^2$$

Wie die Verlustfunktion gestaltet wird?

$L(y,\hat{y})$

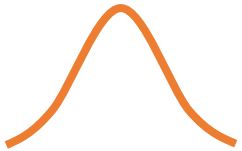
- Methode der kleinsten Quadrate
- **Maximum-Likelihood-Schätzung**

Maximum-Likelihood-Schätzung

nicht in der Realität existieren

y_i

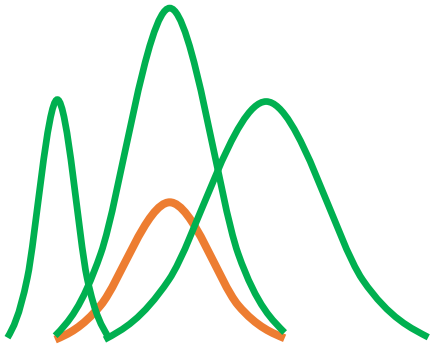
N (μ, σ²)



- Echter Wert: $y_i \in \{0, 1\}$
- Vorhersagewert: $\hat{y}_i \sim (0, 1)$

quantitativ beurteilen

Annäherungsweise Darstellung des Modells



N (μ, σ²)

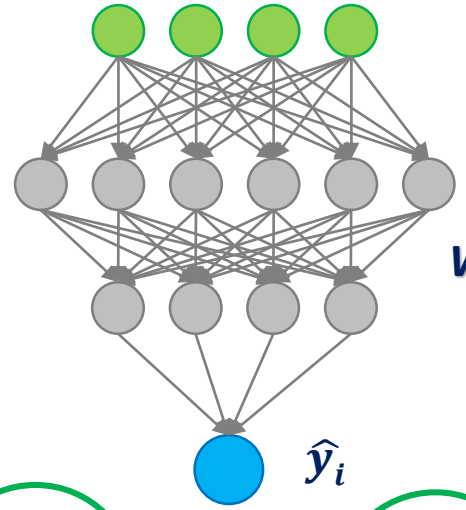
Statistisches Moell

- Gauß-Verteilung
- Poisson-Verteilung

Kommensurabilität

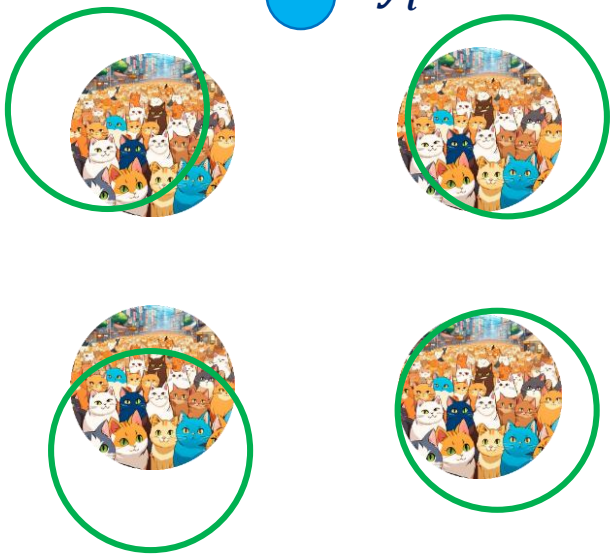
$$Min\ (-\sum_{i=1}^{n} (\ y_i \log \hat{y}_i + (1-yi)\ log(\ 1 - \hat{y}_i)))$$

Neuronales Netzwerk



W, b

$\hat{y}_i$

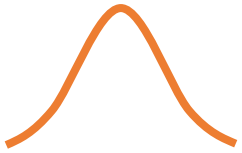


Maximum-Likelihood-Schätzung

nicht in der Realität existieren

y_i

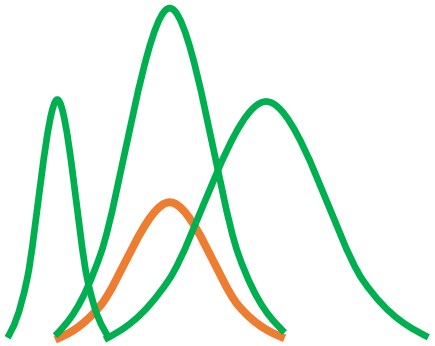
$N(\mu, \sigma^2)$



- Echter Wert: $y_i \in \{0, 1\}$
- Vorhersagewert: $\hat{y}_i \sim (0, 1)$

quantitativ beurteilen

Annäherungsweise Darstellung des Modells



$N(\mu, \sigma^2)$

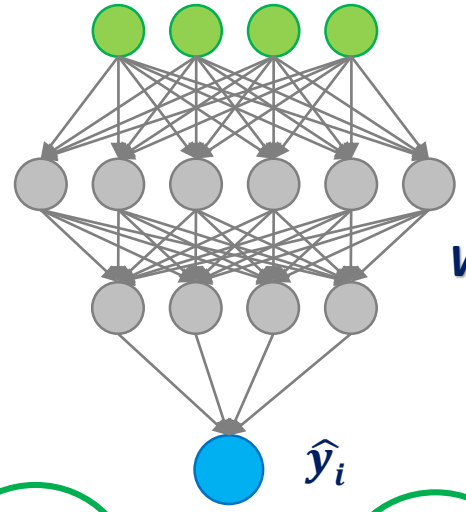
Statistisches Moell

- Gauß-Verteilung
- Poisson-Verteilung

~~**Inkommensurabilität**~~

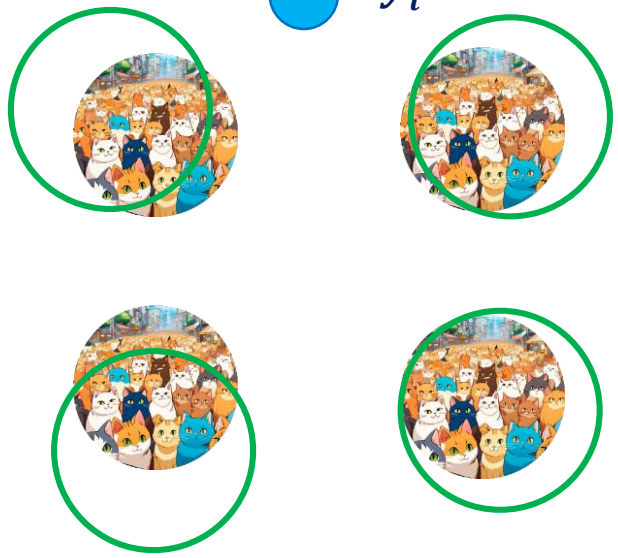
$$Min\ (-\sum_{i=1}^{n} (\ y_i \log \hat{y}_i + (1-yi)\ log(\ 1 - \hat{y}_i)))$$

Neuronales Netzwerk



W, b

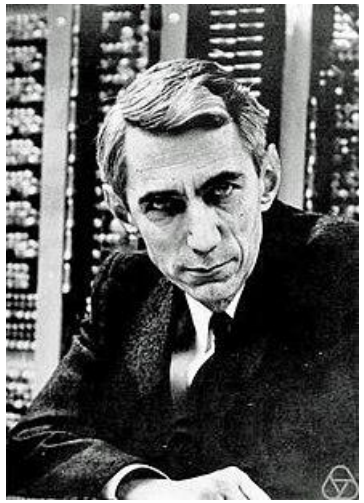
$\hat{y}_i$



Wie die Verlustfunktion gestaltet wird?

$L(y,\hat{y})$

- Methode der kleinsten Quadrate
- Maximum-Likelihood-Schätzung
- **Kreuzentropie** *(Informationstheorie / Thermodynamik)*

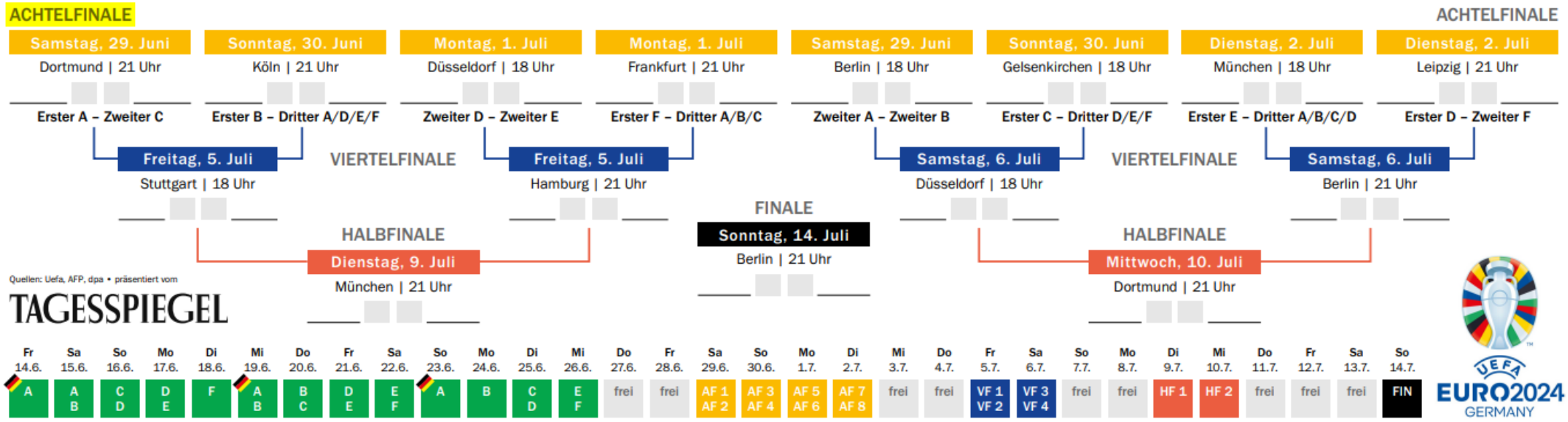


Claude Elwood Shannon (1916-2001) war ein US-amerikanischer Mathematiker und Elektrotechniker. Er gilt als Begründer der Informationstheorie.

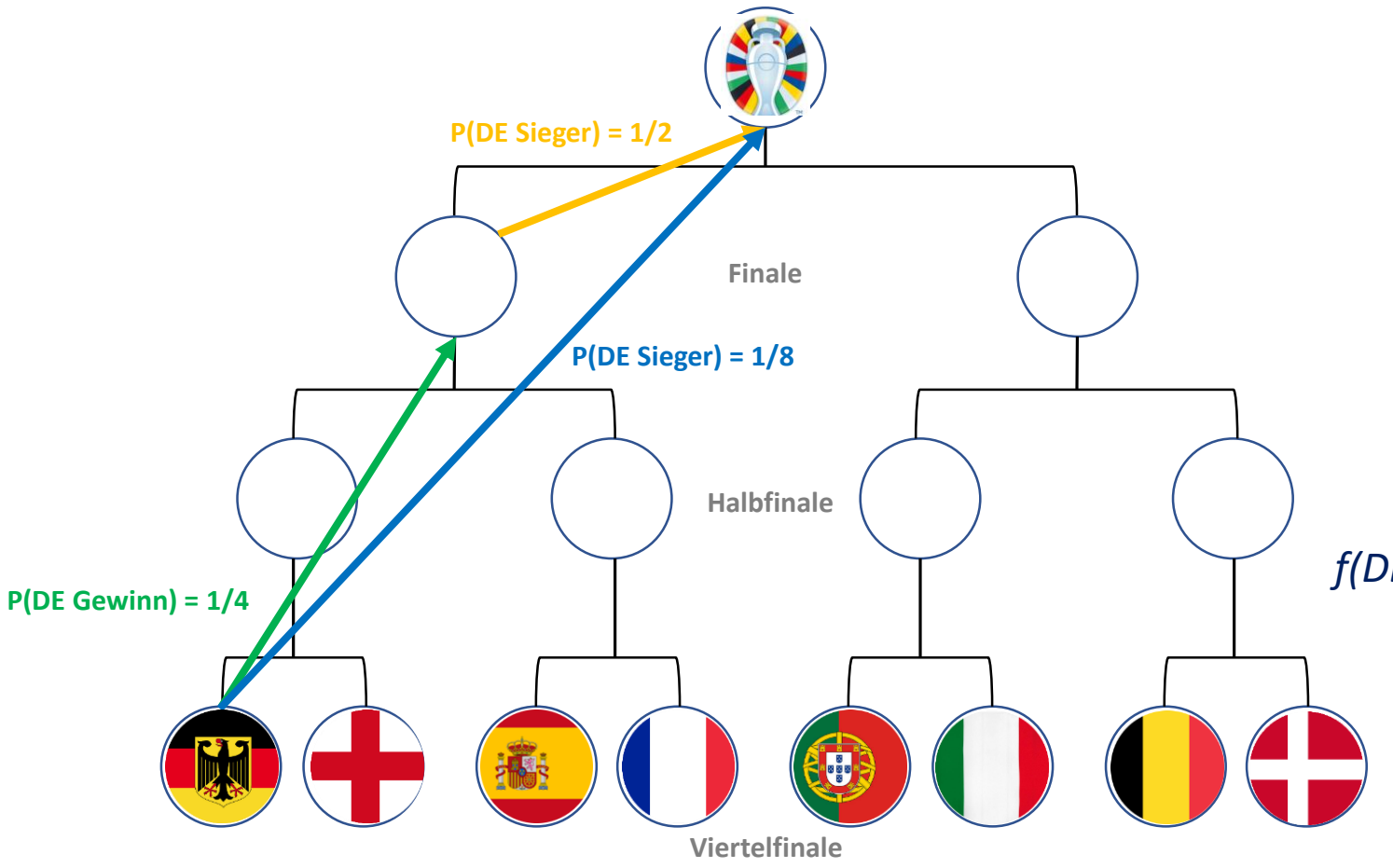
Informationsgehalt

Informationsgehalt

f(x) := Informationsgehalt ?



Informationsgehalt

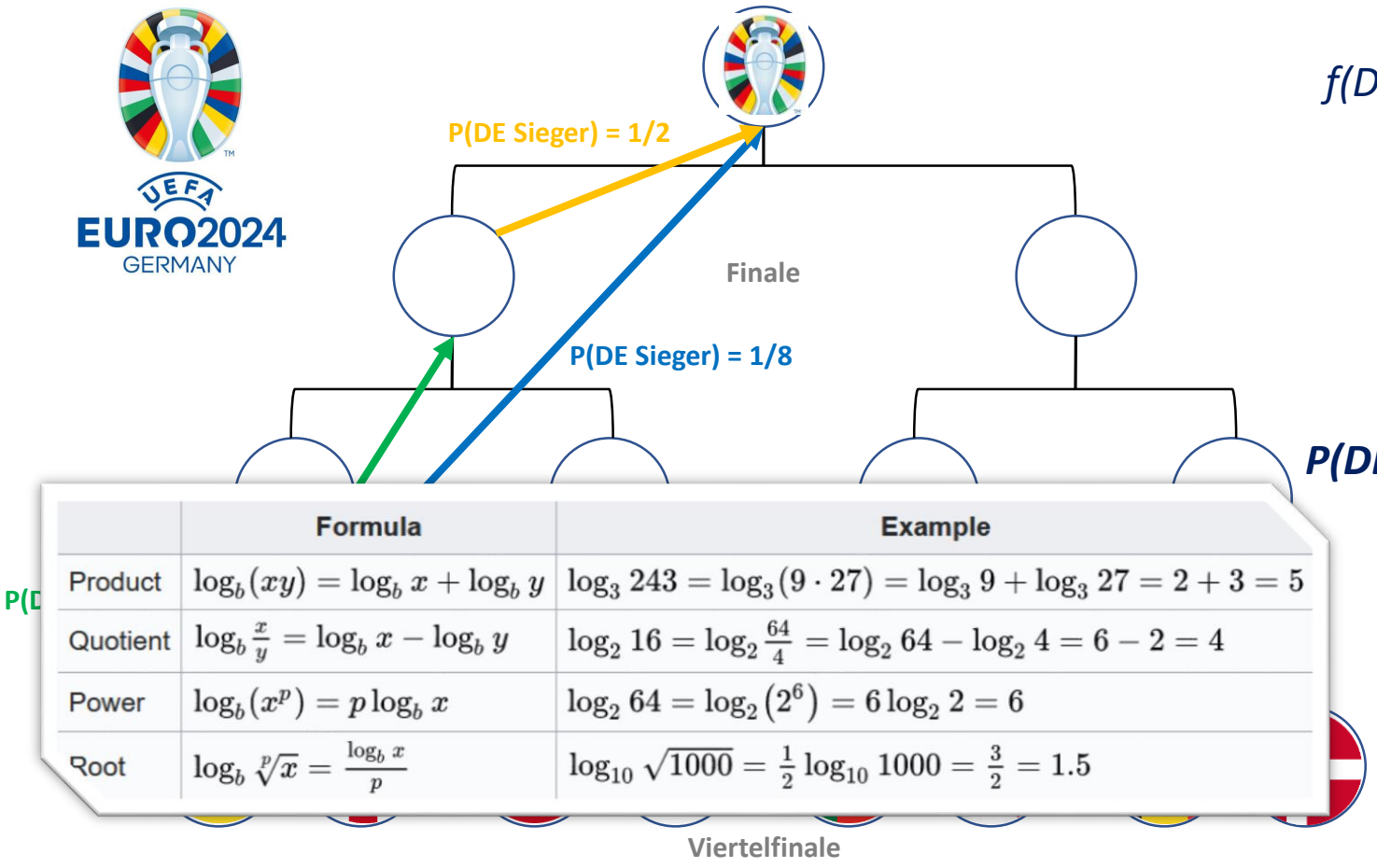


f(x) **:=** Informationsgehalt

f(DE Sieger) **=** *f(DE Finale)* + *f(DE Finale Gewinn)*

Ungewissheit (1/8) --> Gewissheit (1)

Informationsgehalt



	Formula	Example
Product	$\log_b(xy) = \log_b x + \log_b y$	$\log_3 243 = \log_3(9 \cdot 27) = \log_3 9 + \log_3 27 = 2 + 3 = 5$
Quotient	$\log_b \frac{x}{y} = \log_b x - \log_b y$	$\log_2 16 = \log_2 \frac{64}{4} = \log_2 64 - \log_2 4 = 6 - 2 = 4$
Power	$\log_b(x^p) = p \log_b x$	$\log_2 64 = \log_2(2^6) = 6 \log_2 2 = 6$
Root	$\log_b \sqrt[p]{x} = \frac{\log_b x}{p}$	$\log_{10} \sqrt{1000} = \frac{1}{2} \log_{10} 1000 = \frac{3}{2} = 1.5$

f(x) := Informationsgehalt

f(DE Sieger) = f(DE Finale) + f(DE Finale Gewinn)

Ungewissheit (1/8) --> Gewissheit (1)

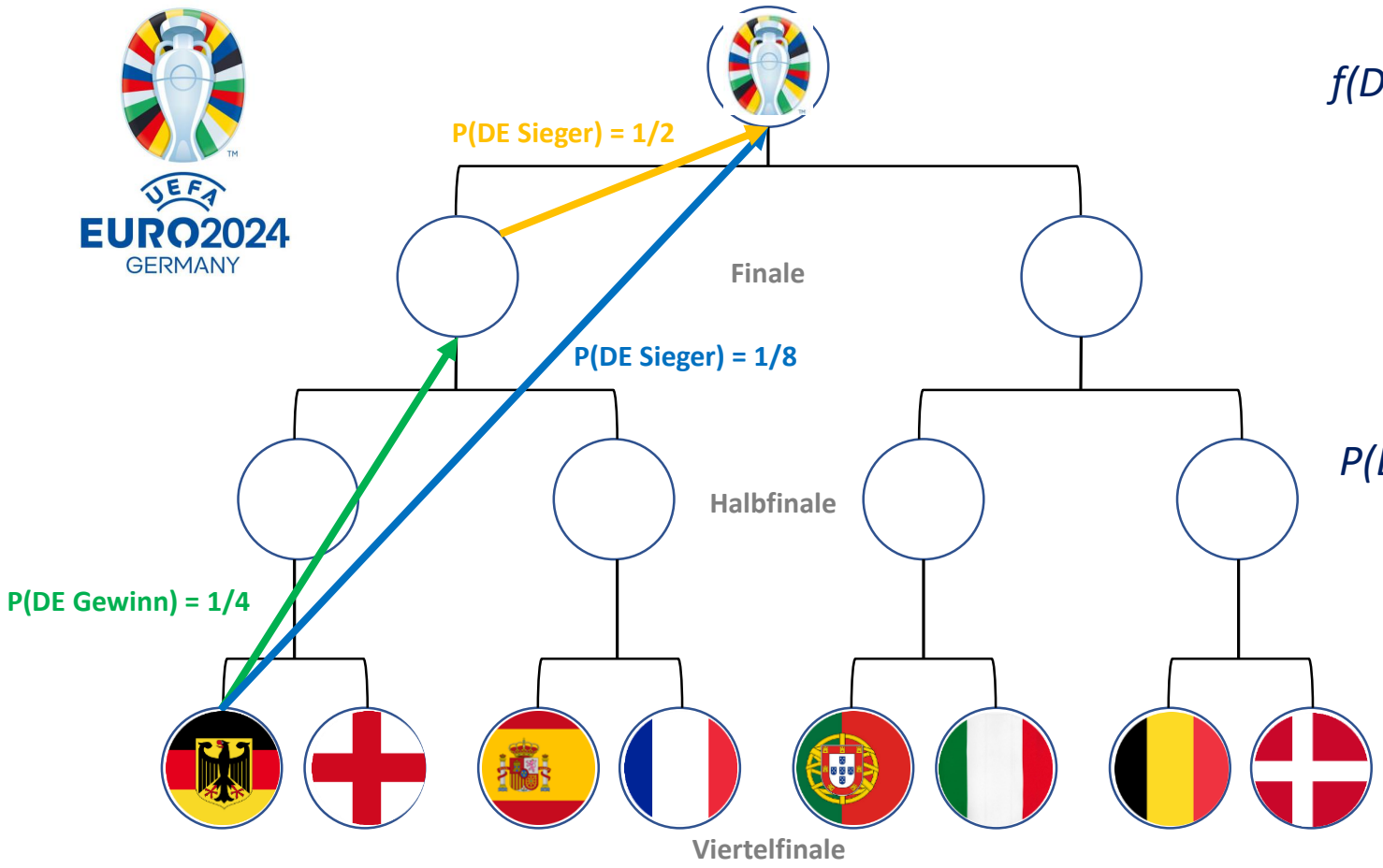
f(1/8) = f(1/4) + f(1/2)

P(DE Sieger) = P(DE Finale) · P(DE Finale Gewinn)

$f(x_1 \cdot x_2) = f(x_1) + f(x_2)$

$f(x) := ?\log_? x$

Informationsgehalt



f(x) := Informationsgehalt

f(DE Sieger) = f(DE Finale) + f(DE Finale Gewinn)

Ungewissheit (1/8) --> Gewissheit (1)

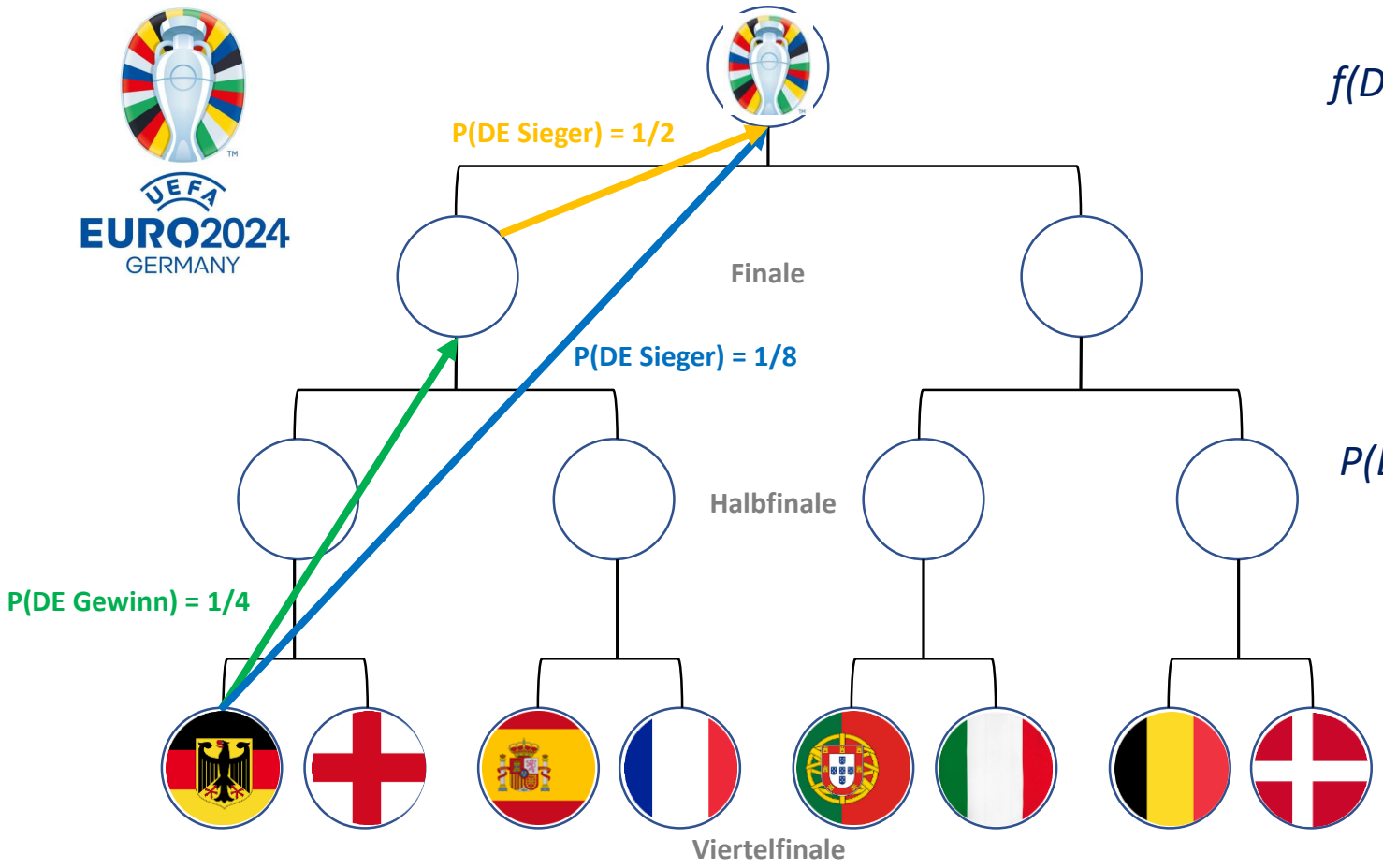
f(1/8) = f(1/4) + f(1/2)

P(DE Sieger) = P(DE Finale) · P(DE Finale Gewinn)

$f(x_1 \cdot x_2) = f(x_1) + f(x_2)$

$f(x) := -\log_{?} x$

Informationsgehalt



$f(x)$:= Informationsgehalt

$f(\text{DE Sieger}) = f(\text{DE Finale}) + f(\text{DE Finale Gewinn})$

Ungewissheit (1/8) --> Gewissheit (1)

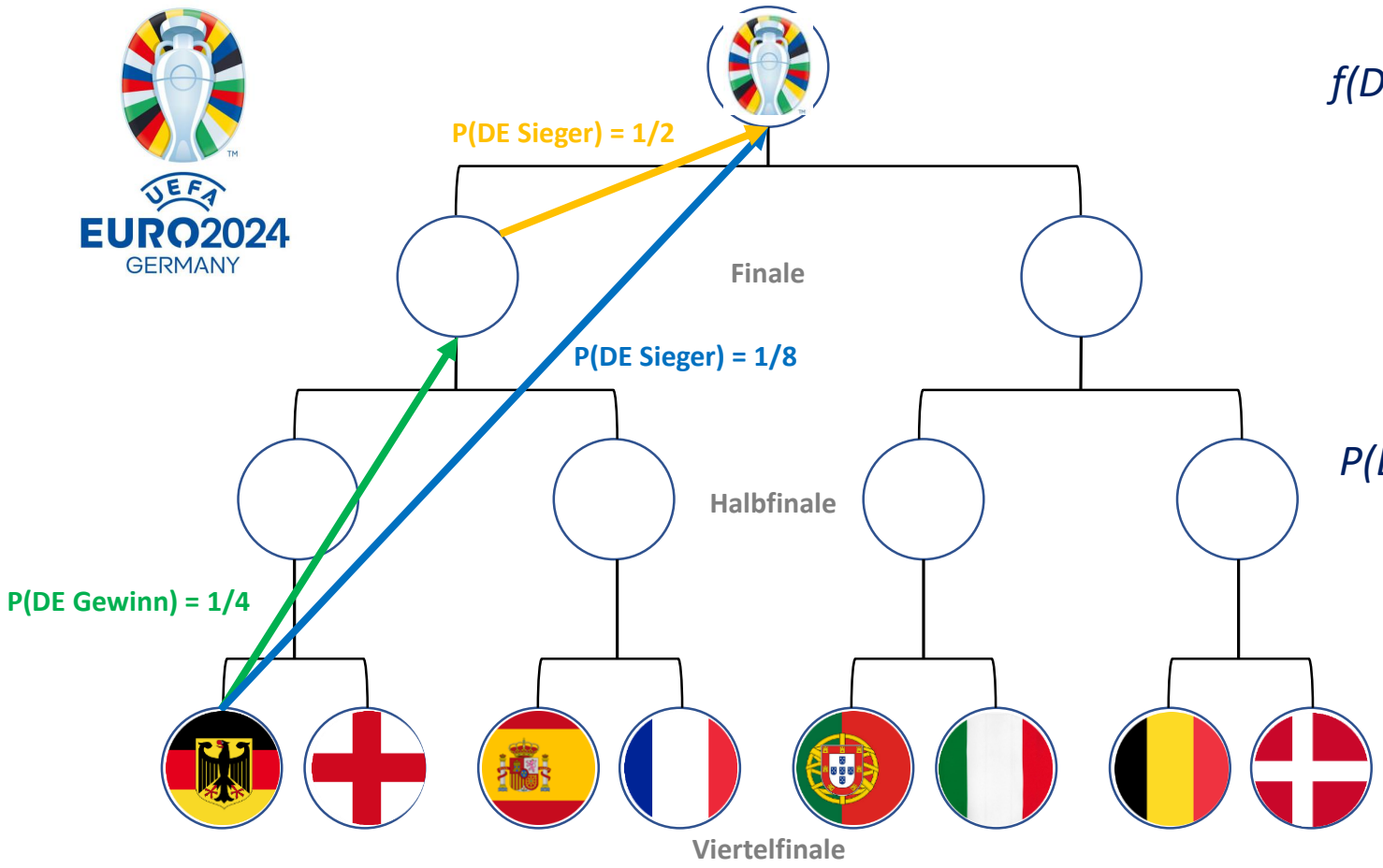
$f(1/8) = f(1/4) + f(1/2)$

$P(\text{DE Sieger}) = P(\text{DE Finale}) \cdot P(\text{DE Finale Gewinn})$

$f(x_1 \cdot x_2) = f(x_1) + f(x_2)$

$f(x) := -\log_2 x$

Informationsgehalt



$f(x)$:= Informationsgehalt

$f(\text{DE Sieger}) = f(\text{DE Finale}) + f(\text{DE Finale Gewinn})$

Ungewissheit (1/8) --> Gewissheit (1)

$f(1/8) = f(1/4) + f(1/2)$

$P(\text{DE Sieger}) = P(\text{DE Finale}) \cdot P(\text{DE Finale Gewinn})$

$f(x_1 \cdot x_2) = f(x_1) + f(x_2)$

$f(x) := -\log_2 x$

$f(1/8) = f(1/4) + f(1/2) = 3$

Informationsgehalt

$P = 1/1024$

$f(x) := -\log_2 x$

$f(P) = -\log_2 P = 10\ (bit)$

Informationsgehalt:

Schwierigkeit (Event): Ungewissheit --> Gewissheit

Entropie:

Schwierigkeit (System): Ungewissheit --> Gewissheit

Einheit (**Informationsgehalt**) = Einheit (**Entropie**)

$f(x)$:= Informationsgehalt

$f(\text{DE Sieger}) = f(\text{DE Finale}) + f(\text{DE Finale Gewinn})$

Ungewissheit (1/8) --> Gewissheit (1)

$f(1/8) = f(1/4) + f(1/2)$

$P(\text{DE Sieger}) = P(\text{DE Finale}) \cdot P(\text{DE Finale Gewinn})$

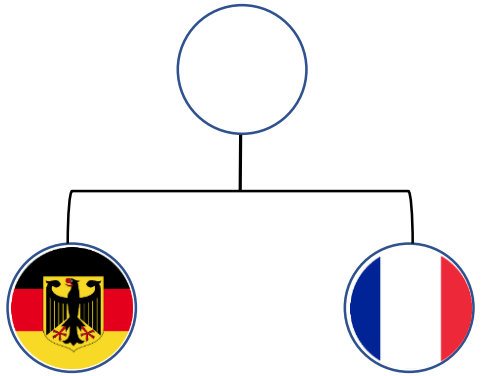
$f(x_1 \cdot x_2) = f(x_1) + f(x_2)$

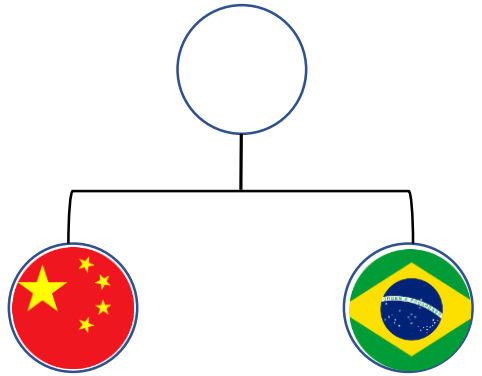
$f(x) := -\log_2 x$

$f(1/8) = f(1/4) + f(1/2) = 3$

Informationsgehalt

$f(x) := -\log_2 x$



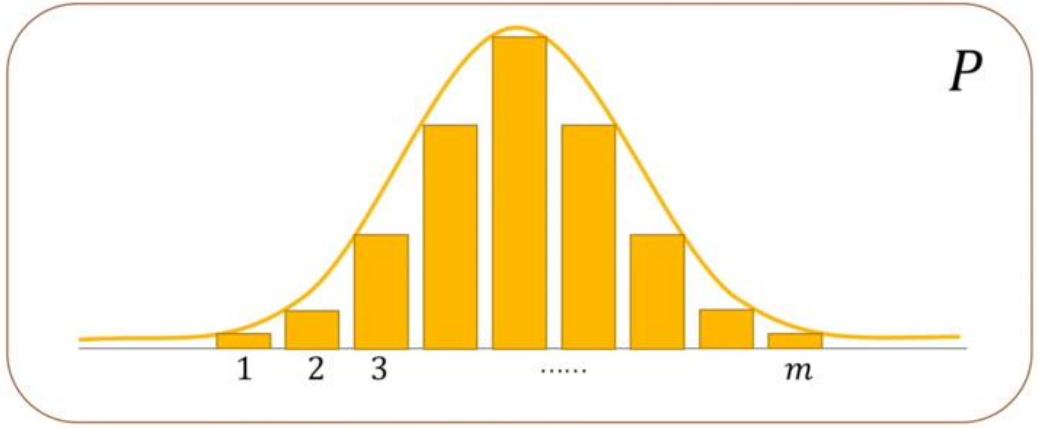


P:	1/2	1/2	1/100	99/100
Informationsgehalt:	$-\log_2 1/2 = 1$	$-\log_2 1/2 = 1$	$-\log_2 1\% = 6{,}6439$	$-\log_2 99\% = 0{,}0145$
Entropie (Event):	$½ \cdot (-\log_2 1/2) = 1/2$	$½ \cdot (-\log_2 1/2) = 1/2$	$1\ \% \cdot (-\log_2 1\%) = 0{,}066439$	$99\ \% \cdot (-\log_2 99\%) = 0{,}014355$
Entropie (System):	**½ + ½ = 1**		$0{,}066439 + 0{,}014355 =$ **0,080794**	

Entropie (System): Erwartungswert des Informationsgehalts

Entropie (Informationstheorie)

Entropie (System): Erwartungswert des Informationsgehalts



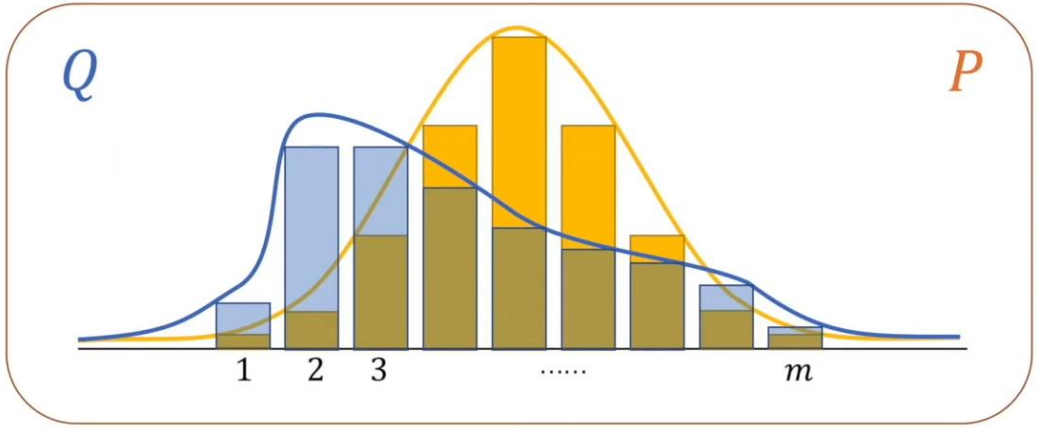
$$H(P) := E(P_I)$$

$$= \sum_{i=1}^{m} P_i \cdot f(P_i) = \sum_{i=1}^{m} P_i \cdot (-\log_2 P_i) = -\sum_{i=1}^{m} P_i \cdot \log_2 P_i$$

Kullback-Leibler-Divergenz (relative Entropie)

Kullback-Leibler-Divergenz (relative Entropie)

Kullback-Leibler-Abstand und relative Entropie bezeichnen ein Maß für die Unterschiedlichkeit zweier Wahrscheinlichkeitsverteilungen.



Informationsgehalt: $f_Q(q_i)$ $f_P(p_i)$

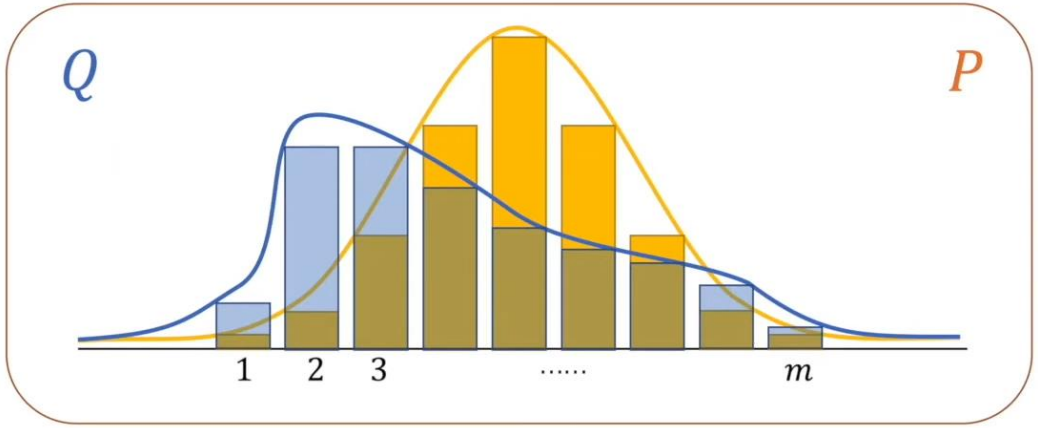
$$D_{KL}(P||Q) \text{ basierend auf } P := \sum_{i=1}^{m} p_i \cdot (f_Q(q_i) - f_P(p_i))$$

$$= \sum_{i=1}^{m} p_i \cdot ((-\log_2 q_i) - (-\log_2 p_i))$$

$$= \sum_{i=1}^{m} p_i \cdot (-\log_2 q_i) - \sum_{i=1}^{m} p_i \cdot (-\log_2 p_i)$$

Kullback-Leibler-Divergenz (relative Entropie)

Kullback-Leibler-Abstand und relative Entropie bezeichnen ein Maß für die Unterschiedlichkeit zweier Wahrscheinlichkeitsverteilungen.



Informationsgehalt: $f_Q(q_i)$ $f_P(p_i)$

D_{KL} (P||Q)
basierend auf **P**

$$:= \sum_{i=1}^{m} p_i \cdot (f_Q(q_i) - f_P(p_i))$$

$$= \sum_{i=1}^{m} p_i \cdot ((-\log_2 q_i) - (-\log_2 p_i))$$

$$= \boxed{\sum_{i=1}^{m} p_i \cdot (-\log_2 q_i)} - \boxed{\sum_{i=1}^{m} p_i \cdot (-\log_2 p_i)}$$

Entropie (System P)

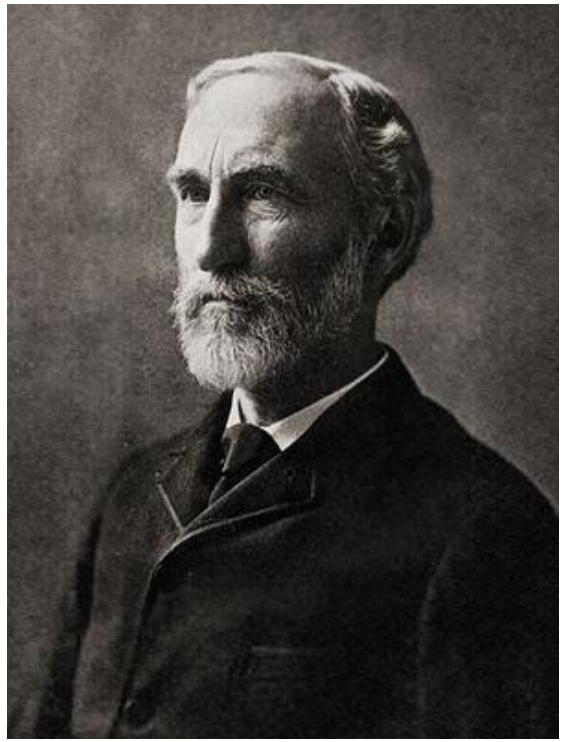
Kreuzentropie H(P, Q)

Gibbs-Ungleichung

Es seien $p = (p_1, \ldots, p_n)$ und $q = (q_1, \ldots, q_n)$ diskrete Wahrscheinlichkeitsverteilungen, d. h. $p_i, q_i > 0$ für alle i und $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} q_i = 1$. Dann gilt:

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2 q_i$$

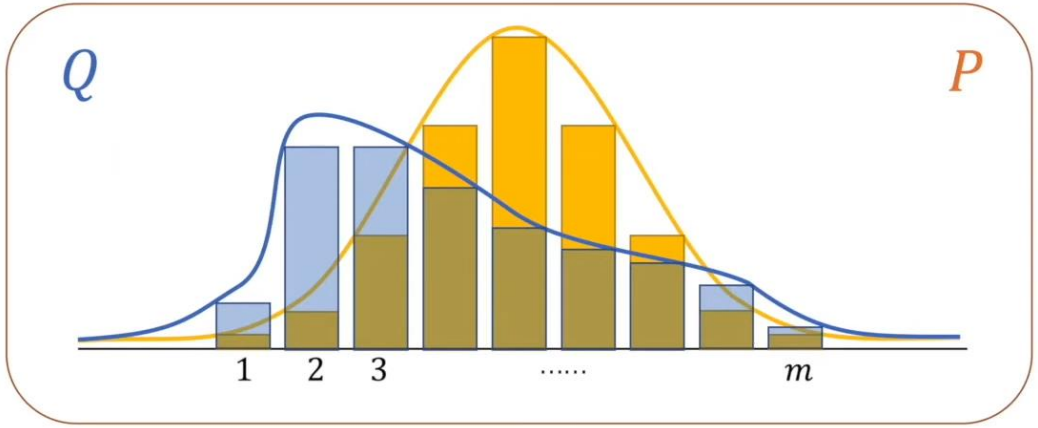
Gleichheit tritt genau dann auf, wenn $p_i = q_i$ für alle i.



Josiah Willard Gibbs (1839-1903) war ein **amerikanischer** Wissenschaftler, der bedeutende theoretische Beiträge zur Physik, Chemie und Mathematik leistete.

Kullback-Leibler-Divergenz (relative Entropie)

Kullback-Leibler-Abstand und relative Entropie bezeichnen ein Maß für die Unterschiedlichkeit zweier Wahrscheinlichkeitsverteilungen.



Informationsgehalt: $f_Q(q_i)$ $f_p(p_i)$

$$D_{KL}(P||Q) := \sum_{i=1}^{m} p_i \cdot (f_Q(q_i) - f_P(p_i))$$
basierend auf **P**

$$= \sum_{i=1}^{m} p_i \cdot ((-\log_2 q_i) - (-\log_2 p_i))$$

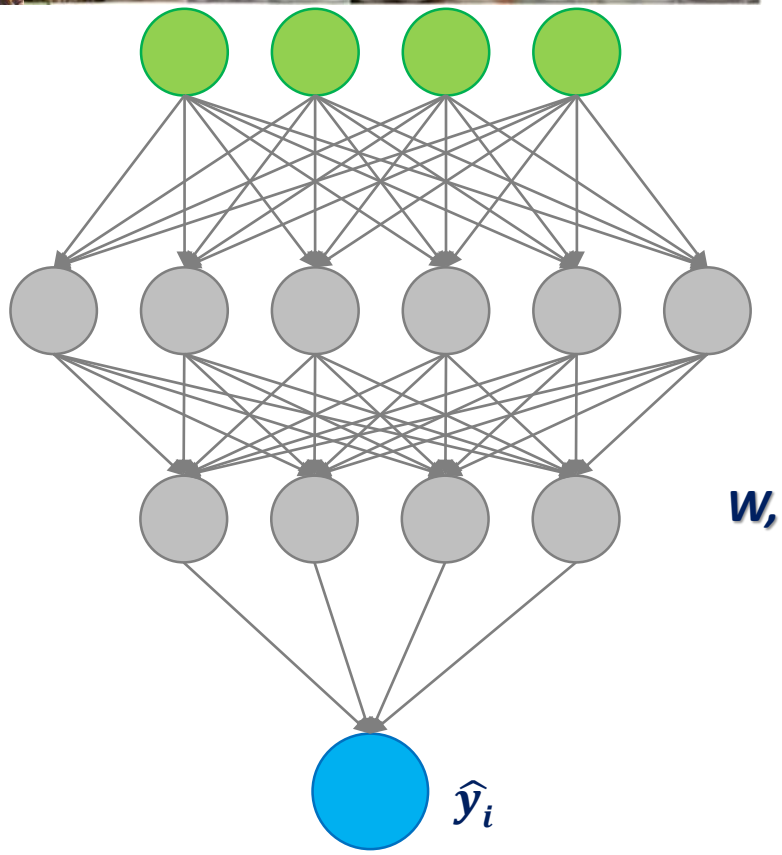
$$= \boxed{\sum_{i=1}^{m} p_i \cdot (-\log_2 q_i)} - \boxed{\sum_{i=1}^{m} p_i \cdot (-\log_2 p_i)}$$

Entropie (System P)

Kreuzentropie H(P, Q)

Kreuzentropie

y_i



W, b

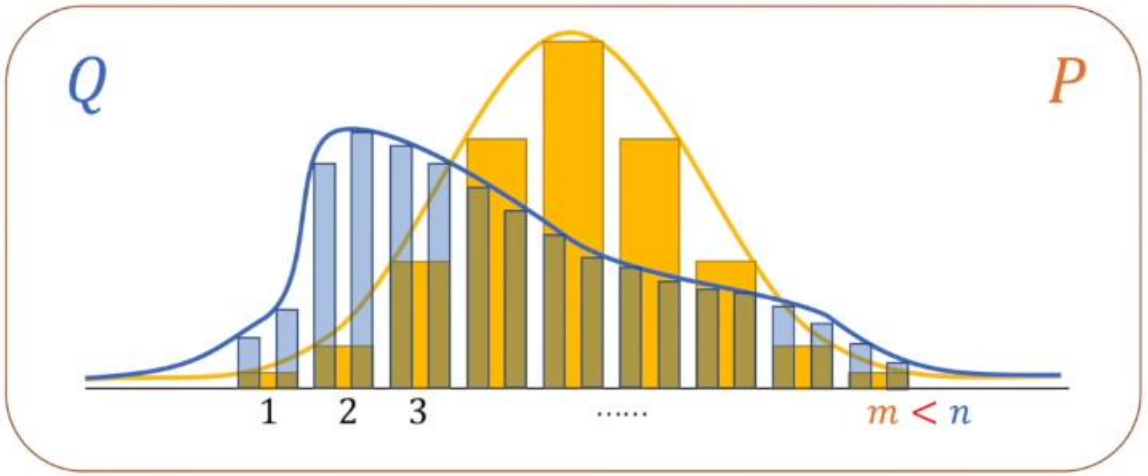
$\hat{y}_i$

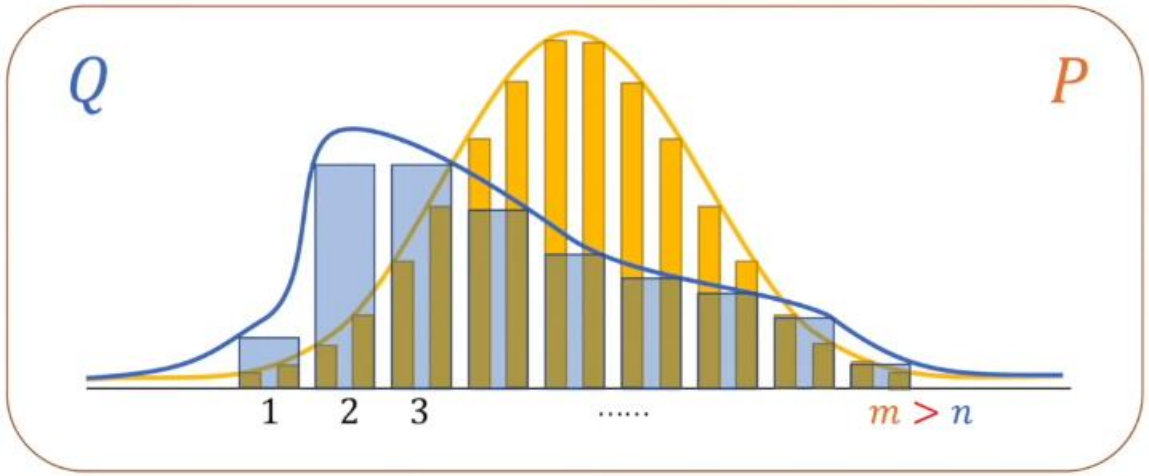
Kreuzentropie H(P, Q)

$$= \sum_{i=1}^{m} p_i \cdot (-\log_2 q_i)$$

- **Echter Wert:** $y_i \in \{0, 1\}$
- **Vorhersagewert:** $\hat{y}_i \sim (0, 1)$

Kreuzentropie





$$D_{KL}(P||Q) := \sum_{i=1}^{n} p_i \cdot (f_Q(q_i) - f_P(p_i))$$

$$= \sum_{i=1}^{n} p_i \cdot ((-\log_2 q_i) - (-\log_2 p_i))$$

$$= \sum_{i=1}^{n} p_i \cdot (-\log_2 q_i) - \sum_{i=1}^{m} p_i \cdot (-\log_2 p_i)$$

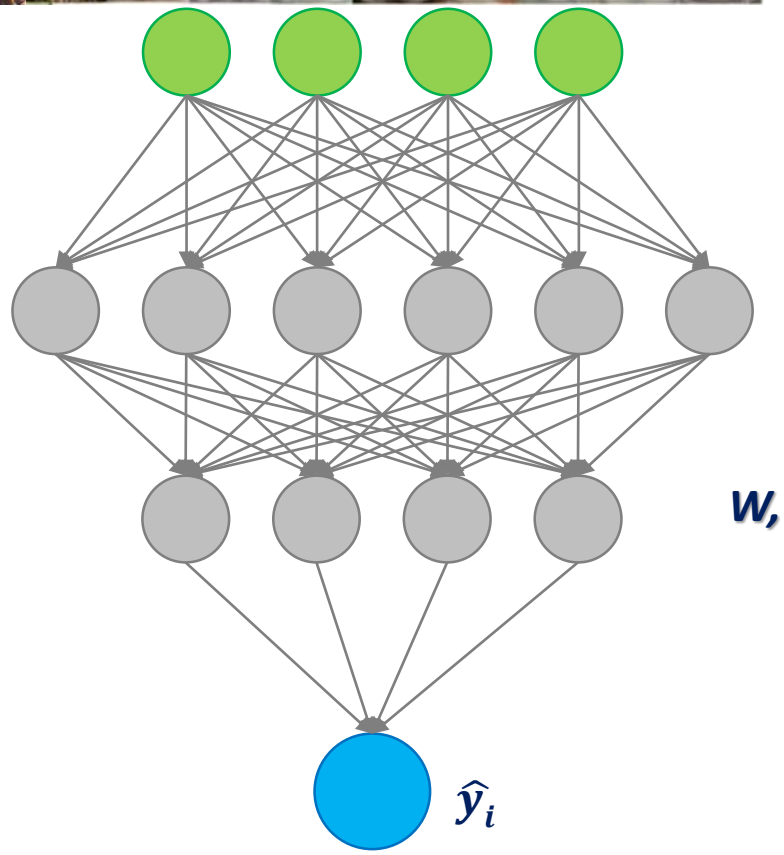
$$D_{KL}(P||Q) := \sum_{i=1}^{m} p_i \cdot (f_Q(q_i) - f_P(p_i))$$

$$= \sum_{i=1}^{m} p_i \cdot ((-\log_2 q_i) - (-\log_2 p_i))$$

$$= \sum_{i=1}^{m} p_i \cdot (-\log_2 q_i) - \sum_{i=1}^{m} p_i \cdot (-\log_2 p_i)$$

Kreuzentropie

y_i



$\hat{y}_i$

W, b

Kreuzentropie H(P, Q)

$$= \sum_{i=1}^{m} p_i \cdot (-\log_2 q_i)$$

$$= \sum_{i=1}^{n} y_i \cdot (-\log_2 q_i)$$

$$= -\sum_{i=1}^{n} (y_i \cdot (\log_2 \hat{y}_i) + (1-y_i) \cdot \log_2(1 - \hat{y}_i))$$

$$\mathcal{L}(\hat{y}, y) = -(y \log \hat{y} + (1-y) \log(1-\hat{y}))$$

Prof. Andrew Ng, 2018

- **Echter Wert:** $y_i \in \{0, 1\}$
- **Vorhersagewert:** $\hat{y}_i \sim (0, 1)$

Take Home Messages

- **Der Informationsgehalt bewertet das Ausmaß der Schwierigkeiten eines Events, das sich von Ungewissheit zu Gewissheit entwickelt. Je höher der Informationsgehalt, desto größer die Schwierigkeiten.**

- **Die Entropie bewertet das Ausmaß der Schwierigkeiten aller Ereignisse in einem System, das sich von der Ungewissheit zur Gewissheit entwickelt.**

- **Die Kreuzentropie kann eine Verlustfunktion sein, die Formel ist die gleiche wie bei der Maximum-Likelihood, aber die physikalischen Angaben sind unterschiedlich.**

Nächste Schritte:

Gradientenverfahren

Vielen herzlichen Dank für eure Aufmerksamkeit!