Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Maschinelles Lernen

Niels Landwehr, Jules Rasetaharison,
Christoph Sawade, Tobias Scheffer

Organisation

- Vorlesung/Übung, praktische Informatik.
- 4 SWS.
- Übung:
  - Mo 10:00-11:30 1.02.
- Vorlesung:
  - Mo 12:00-13:30 1.02.

Organisation

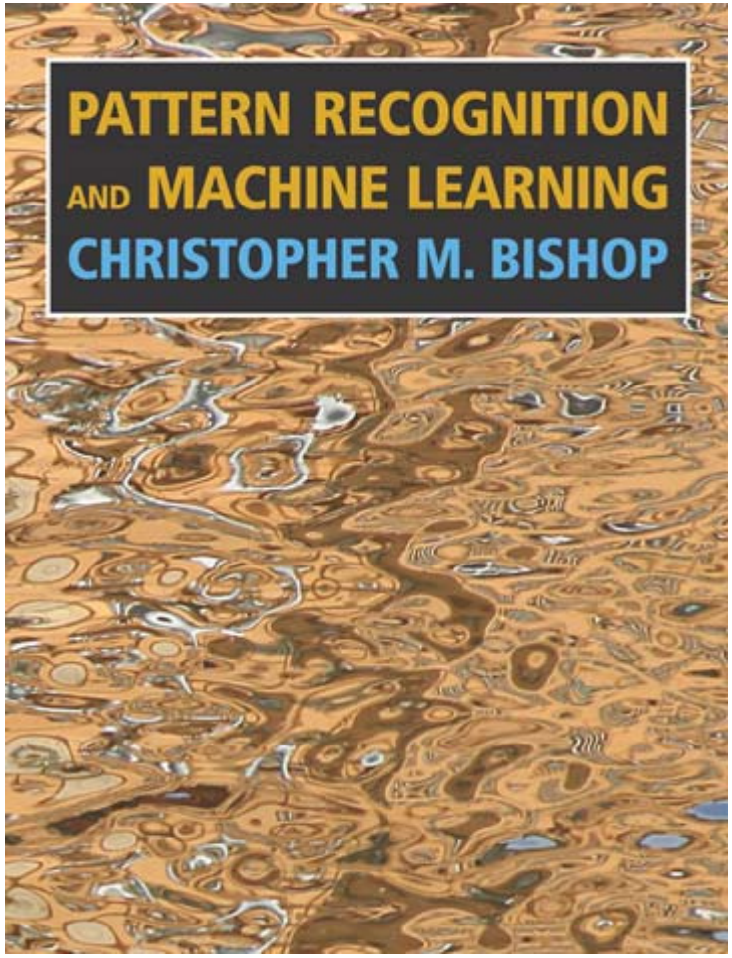
- Webseite.
- Kalender.
  - Vorlesungs- und Übungstermine.
- Blog:
  - Ihre Fragen, Kommentare.
- Folien:
  - Am Tag nach der Vorlesung im Netz.

Organisation

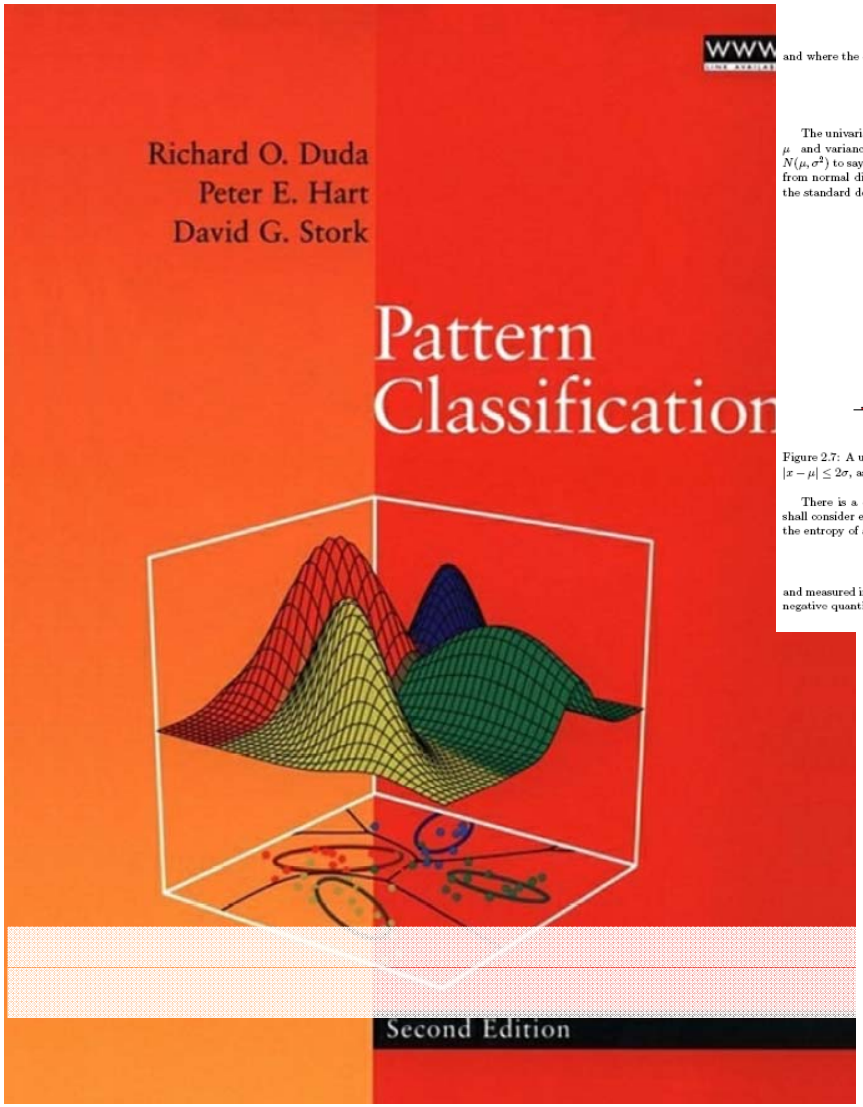
- Übungsaufgaben:
  - Am Tag nach der Vorlesung im Netz.
  - Werden in der darauffolgenden Übung besprochen.
  - Sie können für einzelne Aufgaben votieren.
  - Sie müssen für 2/3 der Aufgaben des Semesters votieren, um die Prüfung abzulegen.
  - Sie rechnen votierte Aufgaben vor.
- Mündliche Prüfung am Ende des Semesters.

Literatur

- Chris Bishop: Pattern Recognition and Machine Learning.
- 30 Exemplare in der Bibliothek.



Literatur



2.5.1 Univariate Density

We begin with the continuous univariate normal or Gaussian density,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \tag{33}$$

for which the *expected value* of x (an average, here taken over the feature space) is

$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x)\, dx, \tag{34}$$

and where the expected squared deviation or *variance* is

$$\sigma^2 \equiv \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx. \tag{35}$$

The univariate normal density is completely specified by two parameters: its mean μ and variance σ^2. For simplicity, we often abbreviate Eq. 33 by writing $p(x) \sim N(\mu, \sigma^2)$ to say that x is distributed normally with mean μ and variance σ^2. Samples from normal distributions tend to cluster about the mean, with a spread related to the standard deviation σ (Fig. 2.7).

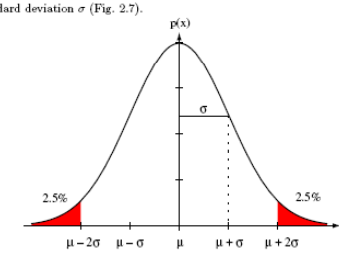


Figure 2.7: A univariate normal distribution has roughly 95% of its area $|x-\mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1$,

There is a deep relationship between the normal distribution and shall consider entropy in greater detail in Chap. ??, but for now we mere the entropy of a distribution is given by

$$H(p(x)) = -\int p(x) \ln p(x)\, dx,$$

and measured in *nats*. If a $\log_2$ is used instead, the unit is the *bit*. The entr negative quantity that describes the fundamental uncertainty in the val

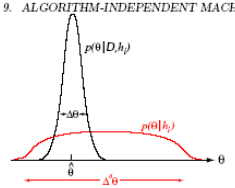


Figure 9.13: In the absence of training data, a particular model h has available a large range of possible values of its parameters, denoted $\Delta^0\theta$. In the presence of a particular training set $\mathcal{D}$, a smaller range is available. The Occam factor, $\Delta\theta/\Delta^0\theta$, measures the fractional decrease in the volume of the model's parameter space due to the presence of training data $\mathcal{D}$. In practice, the Occam factor can be calculated fairly easily if the evidence is approximated as a k-dimensional Gaussian, centered on the maximum-likelihood value $\hat{\theta}$.

Naturally, once the posteriors for different models have been calculated by Eq. 42 & 40, we select the single one having the highest such posterior. (Ironically, the Bayesian model selection procedure is itself not truly Bayesian, since a Bayesian procedure would average over *all* possible models when making a decision.)

The evidence for h_i, i.e., $P(\mathcal{D}|h_i)$, was ignored in a maximum-likelihood setting of parameters $\hat{\theta}$; nevertheless it is the central term in our comparison of models. As mentioned, in practice the evidence term in Eq. 40 dominates the prior term, and it is traditional to ignore such priors, which are often highly subjective or problematic anyway (Problem 38, Computer exercise 7). This procedure represents an inherent bias towards simple models (small $\Delta\theta$); models that are overly complex (large $\Delta\theta$) are automatically self-penalizing where "overly complex" is a data-dependent concept.

In the general case, the full integral of Eq. 41 is too difficult to calculate analytically or even numerically. Nevertheless, if θ is k-dimensional and the posterior can be assumed to be a Gaussian, then the Occam factor can be calculated directly (Problem 37), yielding:

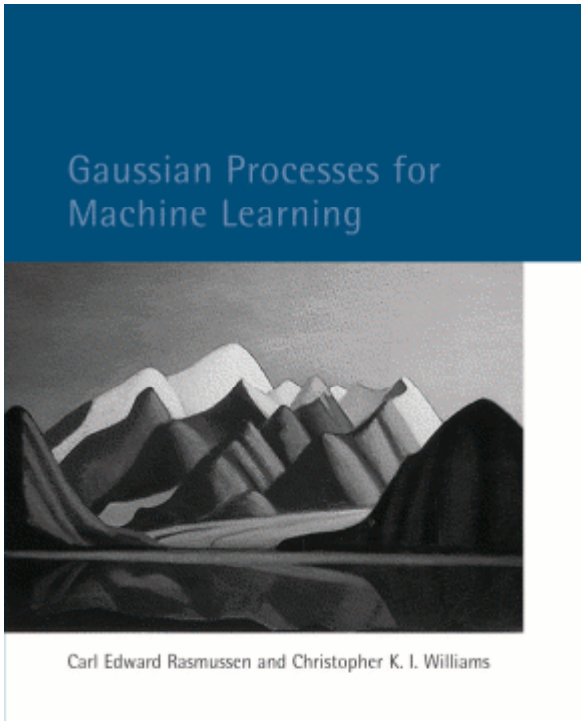
$$P(\mathcal{D}|h_i) \simeq \underbrace{P(\mathcal{D}|\hat{\theta}, h_i)}_{\text{best fit likelihood}} \underbrace{p(\hat{\theta}|h_i)(2\pi)^{k/2}|\mathbf{H}|^{-1/2}}_{\text{Occam factor}}. \tag{44}$$

where

$$\mathbf{H} = \frac{\partial^2 \ln p(\theta|\mathcal{D}, h_i)}{\partial\theta^2} \tag{45}$$

is a Hessian matrix — a matrix of second-order derivatives — and measures how "peaked" the posterior is around the value $\hat{\theta}$. Note that this Gaussian approximation does not rely on the fact that the underlying model of the distribution of the *data* in feature space is or is not Gaussian. Rather, it is based on the assumption that the evidence distribution arises from a large number of independent uncorrelated processes and is governed by the Law of Large Numbers. The integration inherent

Literatur



2.2 Function-space View

An alternative and equivalent way of reaching identical results to the previous section is possible by considering inference directly in function space. We use a Gaussian process (GP) to describe a distribution over functions. Formally:

Definition 2.1 *A* Gaussian process *is a collection of random variables, any finite number of which have a joint Gaussian distribution.* □ — Gaussian process

A Gaussian process is completely specified by its mean function and covariance function. We define mean function $m(\mathbf{x})$ and the covariance function $k(\mathbf{x},\mathbf{x}')$ of a real process $f(\mathbf{x})$ as — covariance and mean function

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})], \\ k(\mathbf{x},\mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))], \end{aligned} \tag{2.13}$$

and will write the Gaussian process as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x},\mathbf{x}')). \tag{2.14}$$

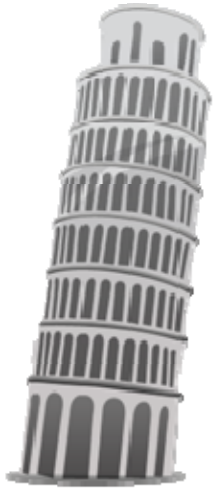
Usually, for notational simplicity we will take the mean function to be zero, although this need not be done, see section 2.7.

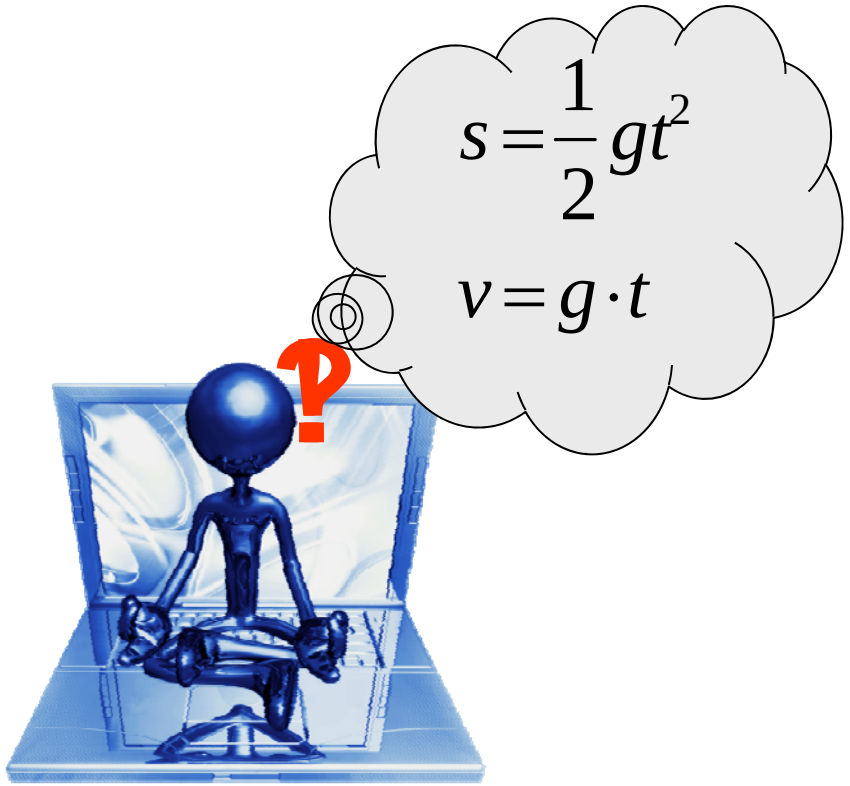
In our case the random variables represent the value of the function $f(\mathbf{x})$ at location $\mathbf{x}$. Often, Gaussian processes are defined over time, i.e. where the index set of the random variables is time. This is not (normally) the case in our use of GPs; here the index set $\mathcal{X}$ is the set of possible inputs, which could be more general, e.g. $\mathbb{R}^D$. For notational convenience we use the (arbitrary) enumeration of the cases in the training set to identify the random variables such that $f_i \triangleq f(\mathbf{x}_i)$ is the random variable corresponding to the case $(\mathbf{x}_i, y_i)$ as would be expected. — index set ≡ input domain

A Gaussian process is defined as a collection of random variables. Thus, the definition automatically implies a *consistency* requirement, which is also sometimes known as the marginalization property. This property simply means that if the GP e.g. specifies $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$, then it must also specify $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$ where Σ_{11} is the relevant submatrix of Σ, see eq. (A.6). In other words, examination of a larger set of variables does not change the distribution of the smaller set. Notice that the consistency requirement is automatically fulfilled if the covariance function specifies entries of the covariance matrix.[5] The definition does not exclude Gaussian processes with finite index sets (which would be simply Gaussian *distributions*), but these are not particularly interesting for our purposes. — marginalization property; finite index set

[5] Note, however, that if you instead specified e.g. a function for the entries of the *inverse* covariance matrix, then the marginalization property would no longer be fulfilled, and one could not think of this as a consistent collection of random variables—this would not qualify as a Gaussian process.

Maschinelles Lernen





System

Daten

Lern-Algorithmus

Modell

Maschinelles Lernen und Data Mining

Datenbank

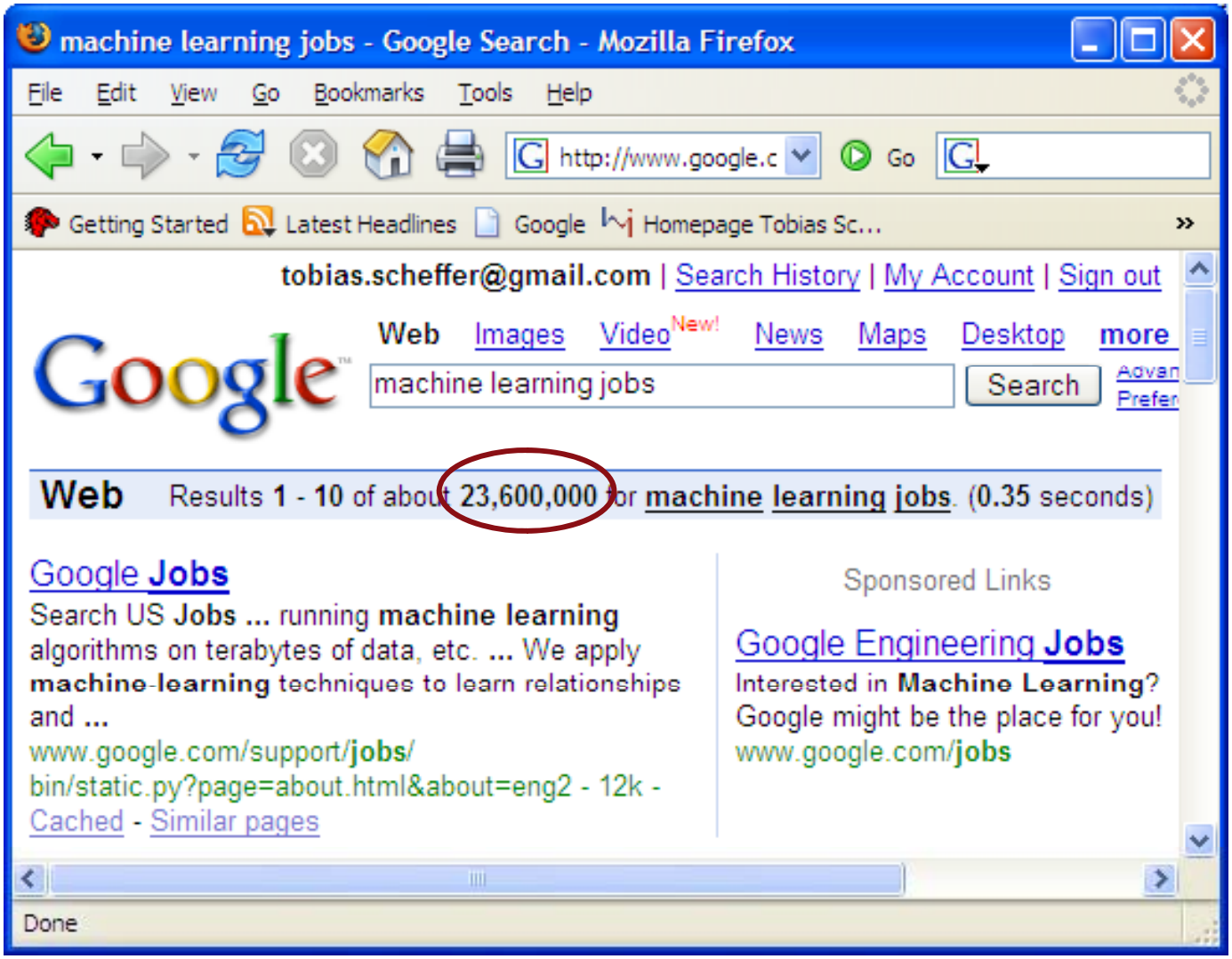
Defekte bestimmter Gene beeinträchtigen Zellstoffwechsel-prozesse.

In Ländern in denen im Winter Salz gestreut wird häufen sich Defekte der neuen Lichtmaschine.

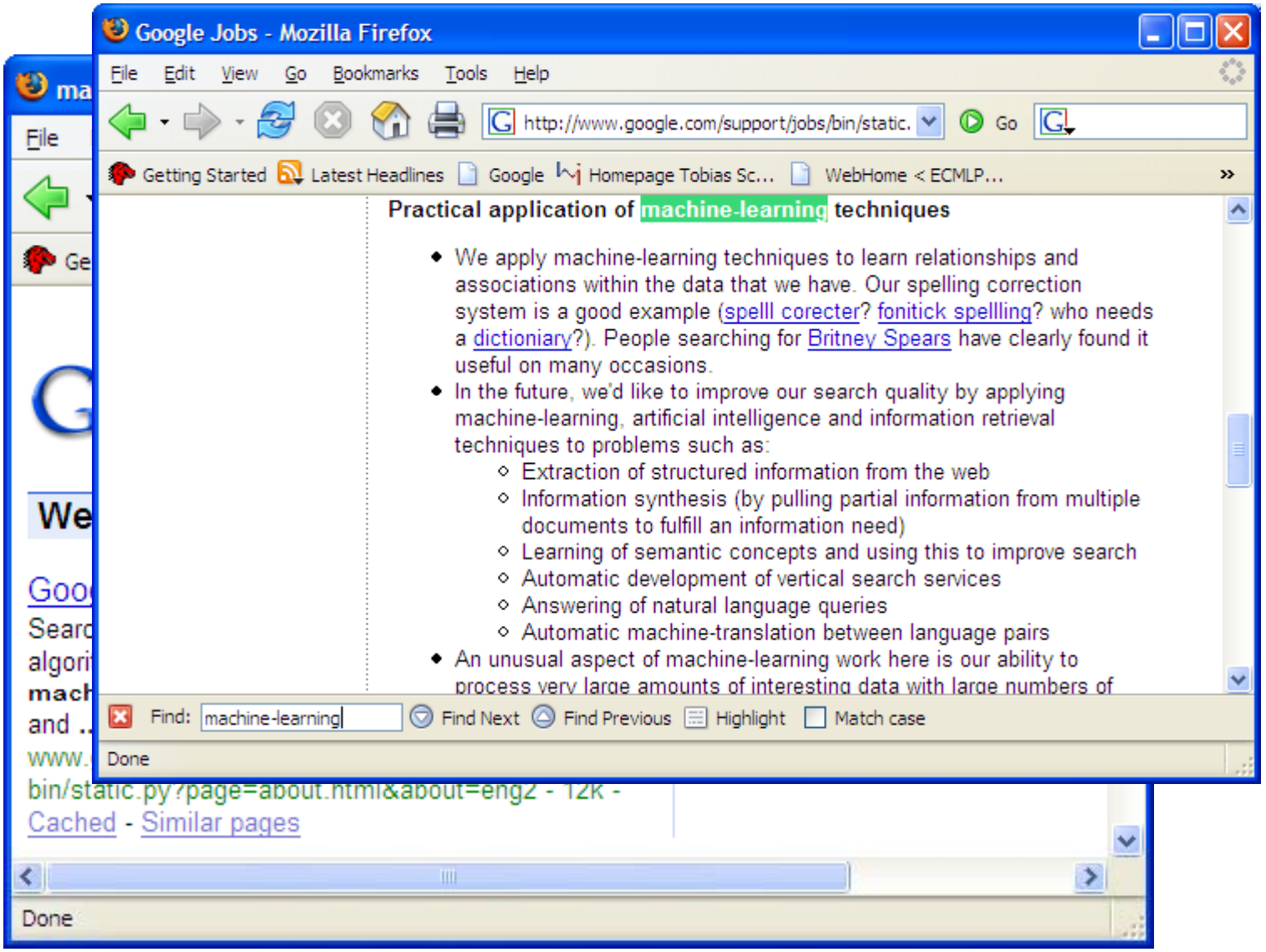
Bestimmte Muster in der Kommunikation deuten auf Hackerangriffe auf Server hin.

Lern-Algorithmus

Maschinelles Lernen und Data Mining



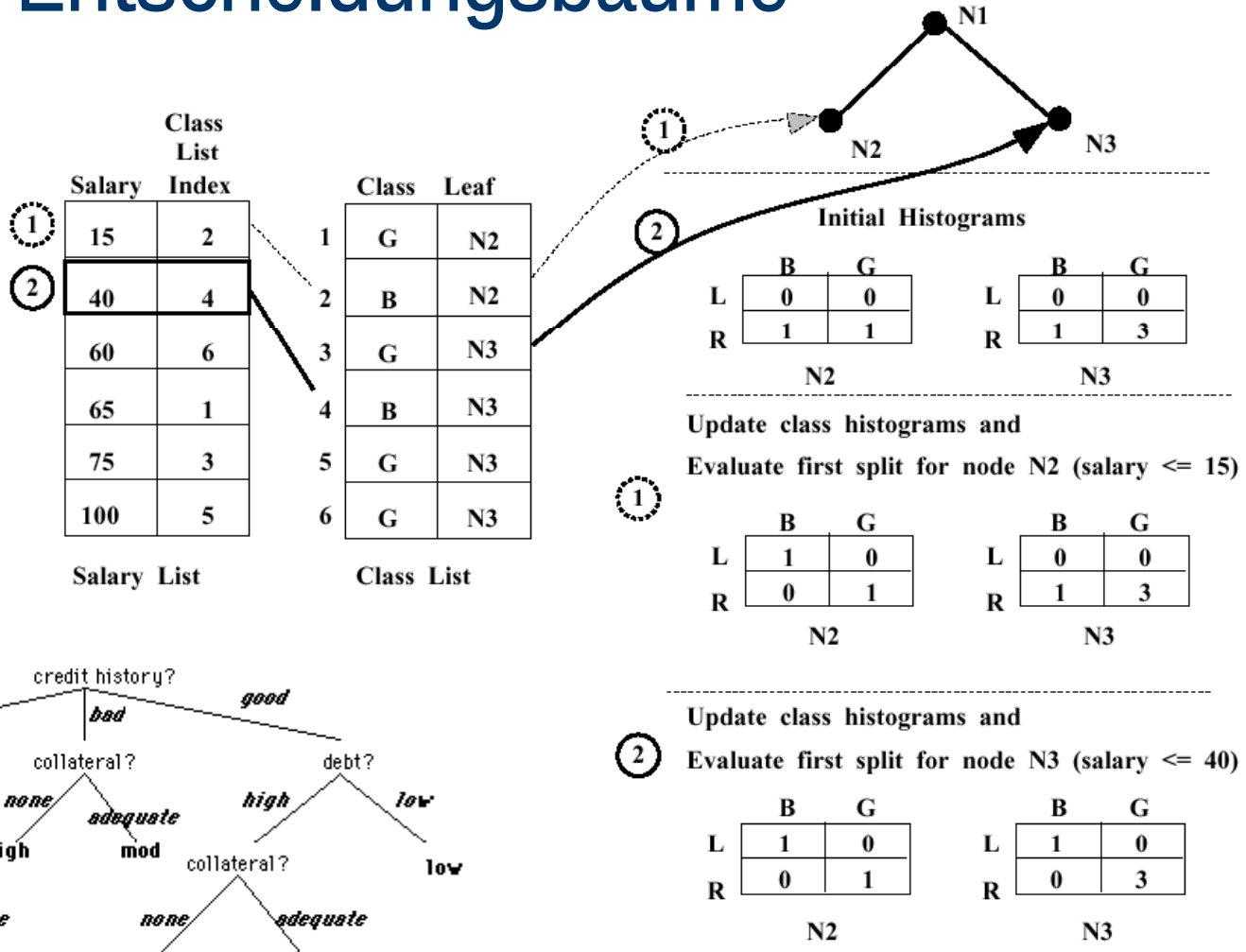
Maschinelles Lernen und Data Mining

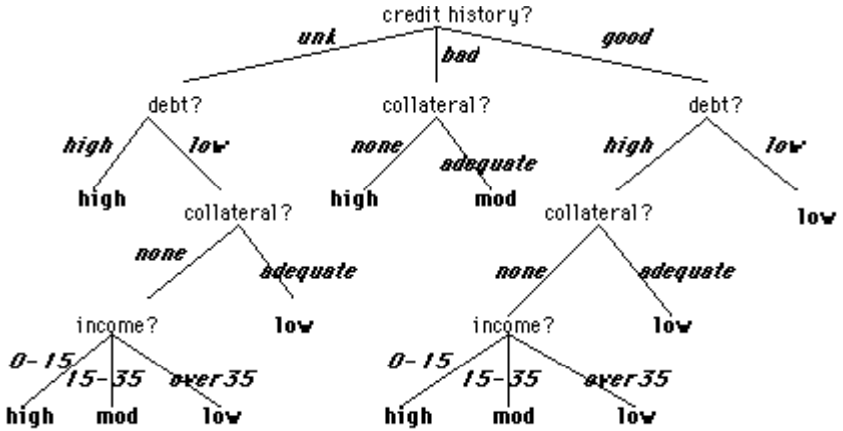


Anwendungen: Modellierung von Risiken

- Kredit-Risiken, Versicherungs-Risiken.

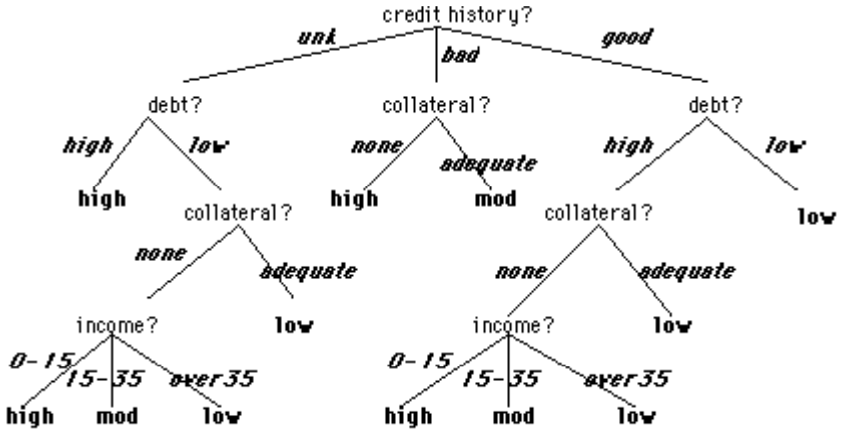
Methoden: Entscheidungsbäume

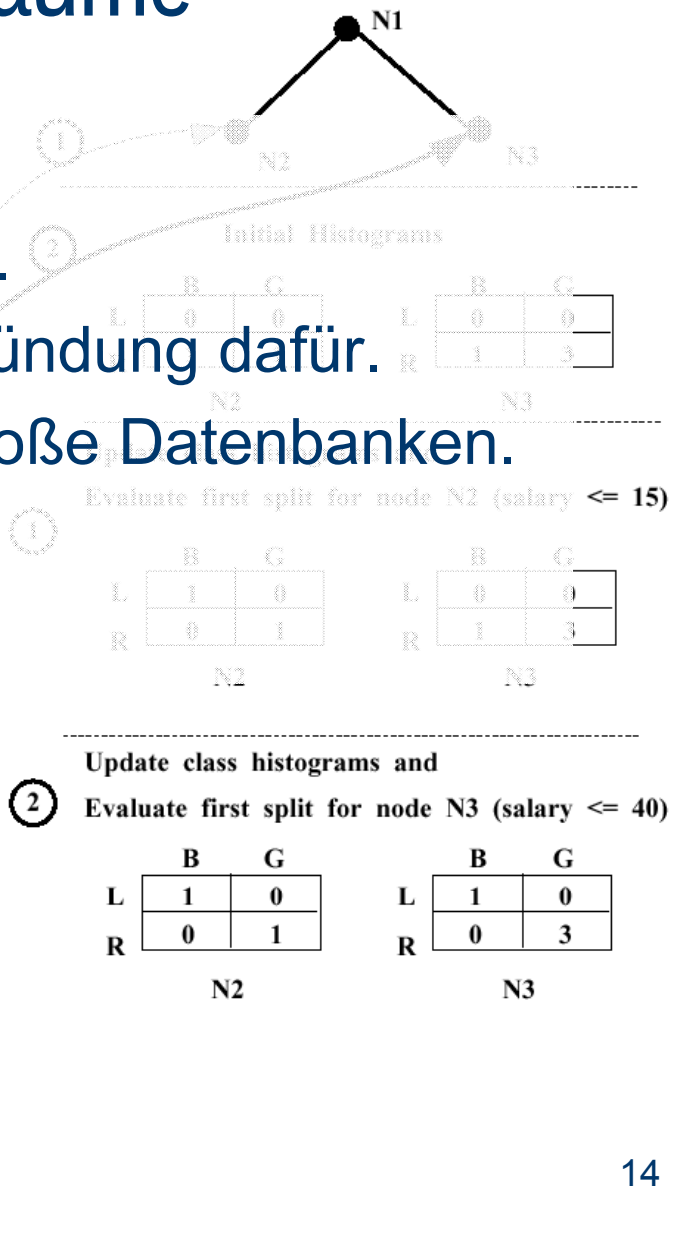




Methoden: Entscheidungsbäume

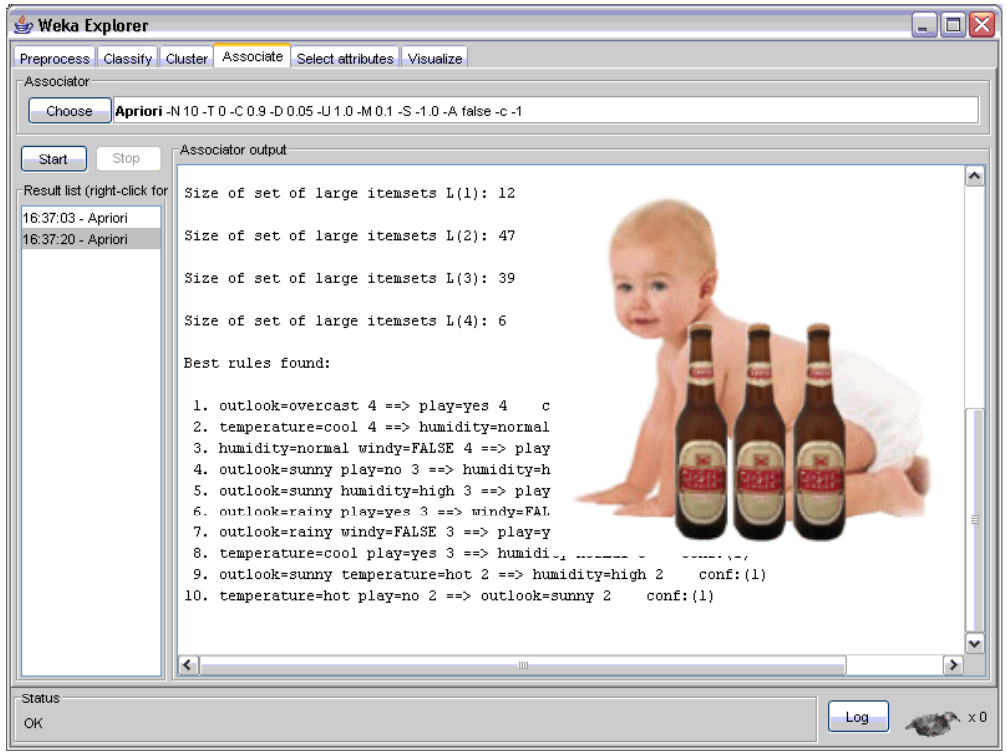
- Einfach, effektiv, effizient.
- Schön interpretierbare Regeln.
- Liefert Klassifikation und Begründung dafür.
- Skalierbare Algorithmen für große Datenbanken.
- Klassifikation und Regression.





Anwendungen: Cross-/Upselling

- Entdecken von Mustern in Datenbanken.
- Welche Produkte ins Sortiment und wohin stellen?



Anwendungen: Empfehlungen

- Nutzer-Item-Empfehlungen.
- Zentrales Element vieler Geschäftsmodelle.

Anwendungen: Empfehlungen

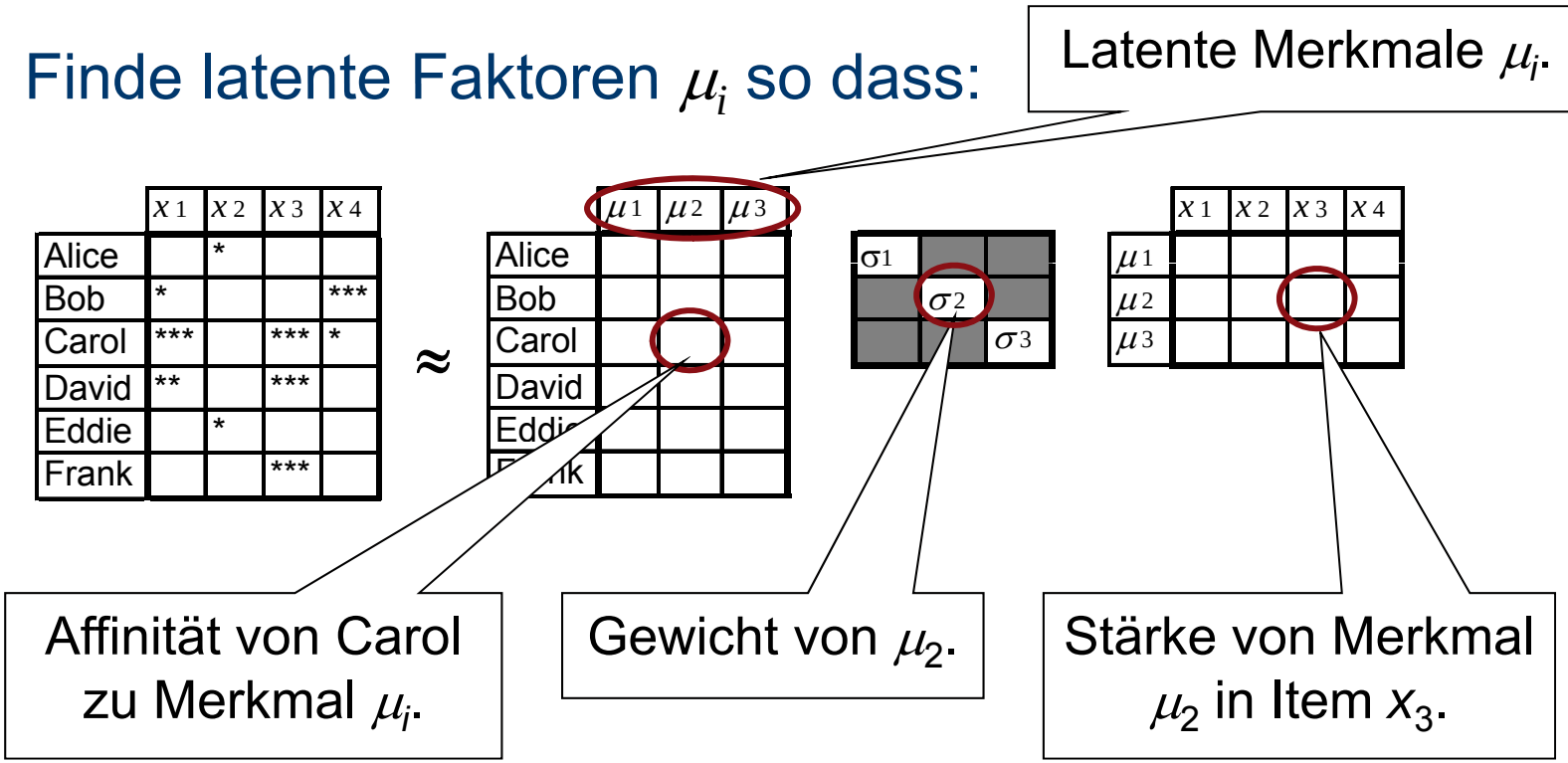
- Netflix Prize: $1.000.000.
- 10% Verbesserung gegenüber aktuellem Modell.

Anwendungen: Empfehlungen

- Long-Tail-Produkte.
- Mass Customization.

Methoden: Matrix-Faktorisierung

- Finde latente Faktoren μ_i so dass:



- Lernalgorithmus: fülle Matrizen mit Merkmalen, die Trainingsdaten möglichst dut rekonstruieren.
- Tracenorm-Regularisierung.

Anwendungen: Spam, Phishing, Angriffe

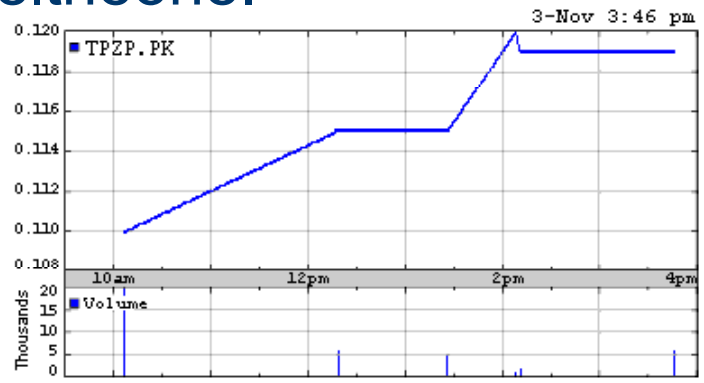
- Klassifikationsprobleme mit Gegenspieler.
- Gegenspieler verändert Verhalten in Reaktion auf gelernte Modelle.
- Maschinelles Lernen + Spieltheorie.

The Bullish Investor
Daily Trading Bulletin
PPI (Potential Profit Index): 8/10 (Med-High)

Trade Entry Date: MONDAY, NOVEMBER 6, 2006

Company Name: **ThermaFreeze Products**
Symbol: TFZP
Current Price: $.119 (Up .014 since last alert!)
5-Day Target: $.600

Price forecasted to **increase** within 72 hours.
BI Recommendation: Buy ASAP!



Press Release:
THEODORE, AL--(MARKET WIRE)
ThermaFreeze Products Corporation released today temperature control performance results in the handling of retail meats by Honey Baked Ham, Inc. Honey Baked Ham is a highly successful forty-year-old national franchise chain with over 400 retail stores. Their delicious smoked, glazed and spiral cut hams are standard holiday, party and family gathering fare for millions of Americans.

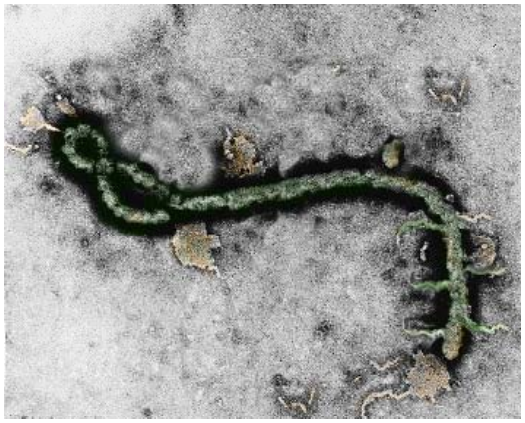
Information within this report contains forward lookingstatements within the meaning of Section 27A of the Securities Act of 1933 and Section 21B of the SEC Act of 1934. Statements that involve discussions with respect to projections of future events. Don't rely on them. This company doesn't report. Past performance isn't indicative of future results. We received 150,000 free trading shares and one thousand dollars in the past All those shares have been sold. We have received an additional one hundred fifty thousand free trading shares now. All shares were received from the same third party, not an officer, director or affiliate. We intend to sell all 150,000 shares now, which could cause the stock to go down. This company has: no revenue in its most recent quarter, nominal cash,an accumulated deficit and a reliance on loans from related parties. It is not a revenue producing company. These factors raise doubt about its ability to continue as a going concern. A failure to finance could cause the company to go out of business. This is high risk stock. This report shall not be taken as any kind of investment advice or solicitation. Read the company's information statement before you invest.

Anwendungen: Fraud, Intrusion Detection

- Fraud Detection: erkennen betrügerischer Transaktionen.

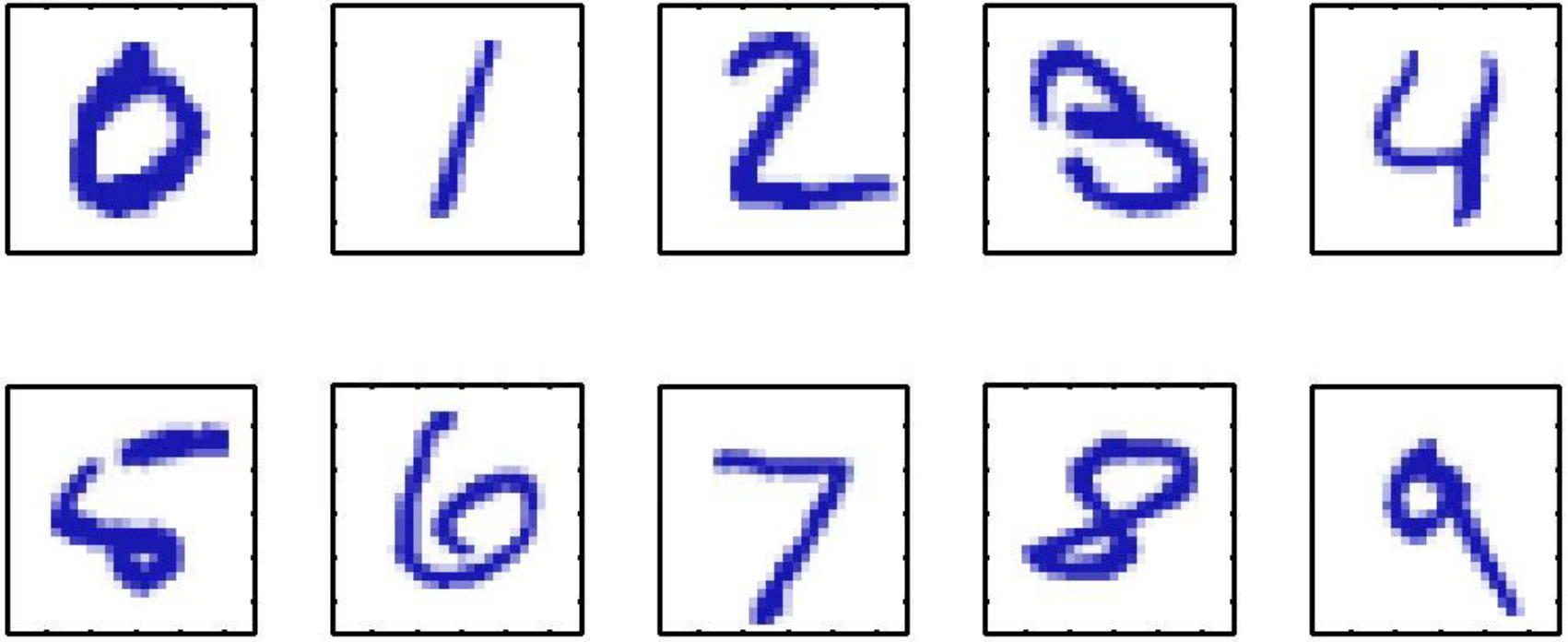
Anwendungen: Gesundheit

- Vorhersage: Medikament gegen gegebene Virus-Version wirksam?
- Vorhersage: Krankheitsverlauf in Abhängigkeit von Behandlungsparametern.



Anwendungen: Mustererkennung

- Z.B. Handschrifterkennung.



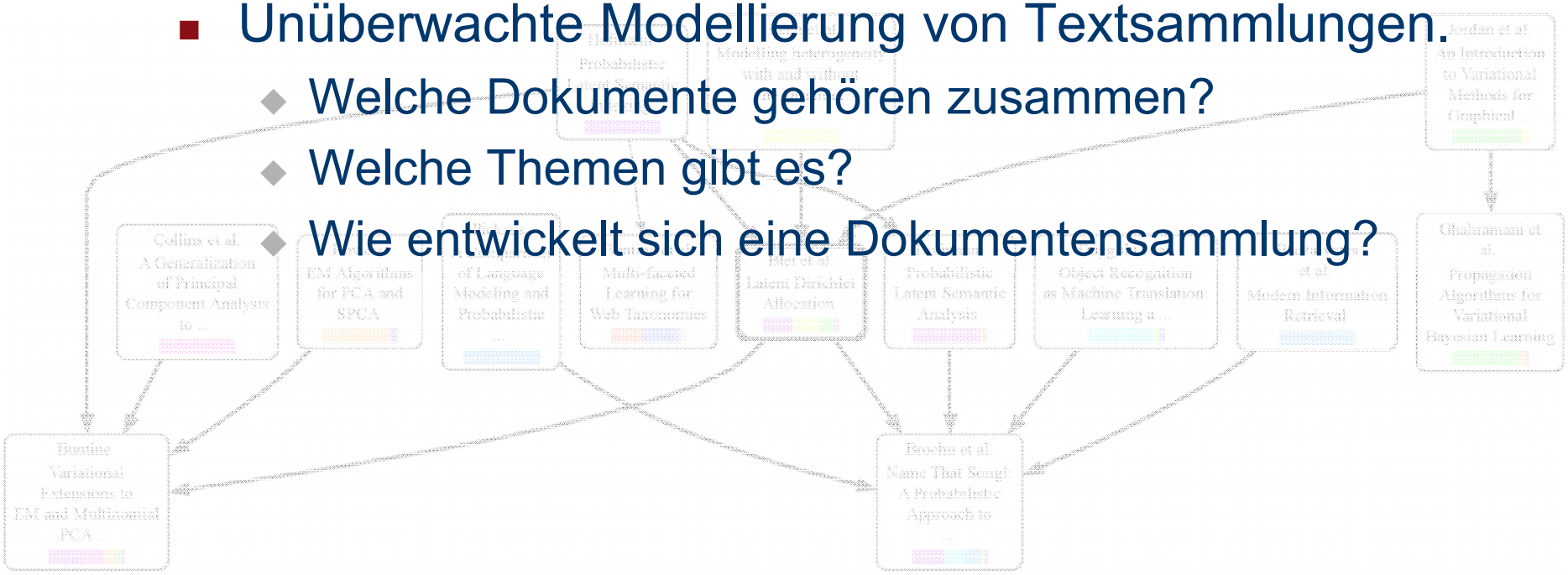
Anwendungen: Mustererkennung

- Z.B. Luftbild-/Radarbilder.

Anwendungen: natürliche Sprache

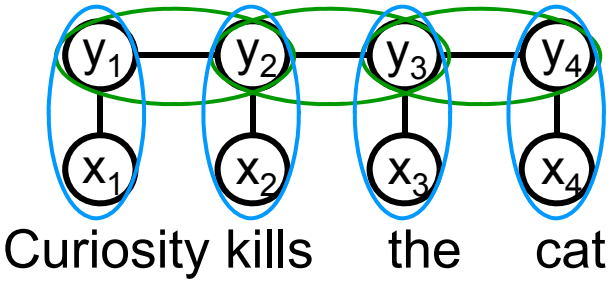
- Vorhersage, überwachtes Lernen.
  - Textklassifikation.
  - Informationsextraktion.
  - Wortarterkennung, Parsen.
  - Übersetzung.

- Unüberwachte Modellierung von Textsammlungen.
  - Welche Dokumente gehören zusammen?
  - Welche Themen gibt es?
  - Wie entwickelt sich eine Dokumentensammlung?



Methoden: Kernel-Methoden

- Familie von Methoden für diskriminatives Lernen.
  - Klassifikation,
  - Verarbeitung von Sequenzen (Text),
  - Bäumen (Parsen), Graphen (Web).



$$\Pi_t \Phi(y_t,y_{t+1}) = \exp \{ \sum_i w_i \sum_t \varphi_i(y_t,y_{t+1}) \}.$$

$$\Pi_t \Phi(x_t,y_t) = \exp \{ \sum_i w_i \sum_t \varphi_i(x_t,y_t) \}.$$

$$\min \quad \tfrac{1}{2} |\mathbf{w}|^2 + C \left(\sum_i \xi_i + \mathbf{c_u} \sum_i \gamma_i \xi_i \right)$$

$$\text{s.t.} \quad \mathbf{w}^T \left(\boldsymbol{\varphi}(\mathbf{x}_i,\mathbf{y}_i) - \boldsymbol{\varphi}(\mathbf{x}_i,\bar{\mathbf{y}}) \right) \geq 1 - \xi_i \qquad \forall \bar{\mathbf{y}} \neq \mathbf{y}_i \ \forall i=1,\ldots,n$$

$$\xi_i \geq 0 \qquad \forall i=1,\ldots,n$$

Streifzug: Kernel-Methoden

- Familie von Methoden für diskriminatives Lernen.
  - Klassifikation,
  - Verarbeitung von Sequenzen (Text),
  - Bäumen (Parsen), Graphen (Web).
- Beste bekannte Methoden für
  - Parsieren, Eigennamenerkennung, Textklassifikation.
  - viele Mustererkennungsaufgaben.

$$\prod_t \Phi(y_t, y_{t-1}) = \exp\{\sum_i w_i \sum_t \varphi_i(y_t, y_{t-1})\}$$

$$\sum_t \langle w_i, y_t \rangle \cdot \exp\{\sum_i w_i \sum_t \varphi_i(x_t, y_t)\}.$$

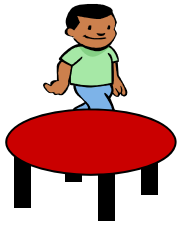
$$\min \; \tfrac{1}{2}|w|^2 + C\left(\sum_i \xi_i + C_u \sum_i \gamma_i \xi_i\right)$$

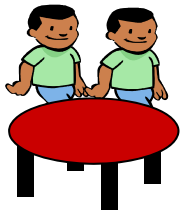
$$\text{s.t.} \quad w^T(\varphi(x_i, y_i) - \varphi(x_i, \bar{y})) \geq 1 - \xi_i \qquad \forall \bar{y} \neq y_i \; \forall i = 1, \ldots, n$$

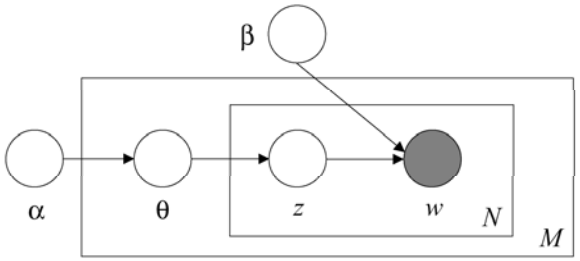
$$\xi_i \geq 0 \qquad \forall i = 1, \ldots, n$$

Streifzug: Statistische Modelle

- Bayessche Statistik:
  - Vorwissen + Beobachtungen → Modell, dass verbleibende Ungewissheit charakterisiert.
  - Sauberes, probabilistisches Modell.







$$\phi_i \mid \phi_1, \ldots, \phi_{i-1}, \alpha_0, G_0 \sim \sum_{l=1}^{i-1} \frac{1}{i-1+\alpha_0} \delta_{\phi_l} + \frac{\alpha_0}{i-1+\alpha_0} G_0$$

$\frac{1}{2+\alpha} \cdot \delta_{\phi_l}$	$\frac{1}{2+\alpha} \cdot \delta_{\phi_l}$	$\frac{\alpha}{2+\alpha} \cdot G_0$
$\frac{1}{3+\alpha} \cdot \delta_{\phi_l}$	$\frac{2}{3+\alpha} \cdot \delta_{\phi_l}$	$\frac{\alpha}{3+\alpha} \cdot G_0$

Streifzug: Statistische Modelle

- Bayessche Statistik:
  - Vorwissen + Beobachtungen → Modell, dass verbleibende Ungewissheit charakterisiert.
  - Sauberes, probabilistisches Modell.
- Flexibel, in extrem vielen Gebieten einsetzbar.
  - Sprachverarbeitung, Übersetzung,
  - Roboterlokalisierung, …
- Elegante Modelle, schöne Algorithmen.
  - Chinese Restaurant Processes.
  - Dirichlet-Prozesse, Gaußsche Prozesse.
  - Gibbs Sampling.

Wir stellen ein

- Studentische Mitarbeiter
  - Gern im Zusammenhang mit Studien-/Diplomarbeit.
  - Bitte sprechen Sie uns an.
- Wissenschaftliche Mitarbeiter,
- Promotionsstipendiaten.
  - z.T. Drittmmittelprojekten mit Industriepartnern.
  - Schreiben Sie am besten eine Diplomarbeit bei uns.