TECHNISCHE UNIVERSITÄT
BERGAKADEMIE FREIBERG
Die Ressourcenuniversität. Seit 1765.

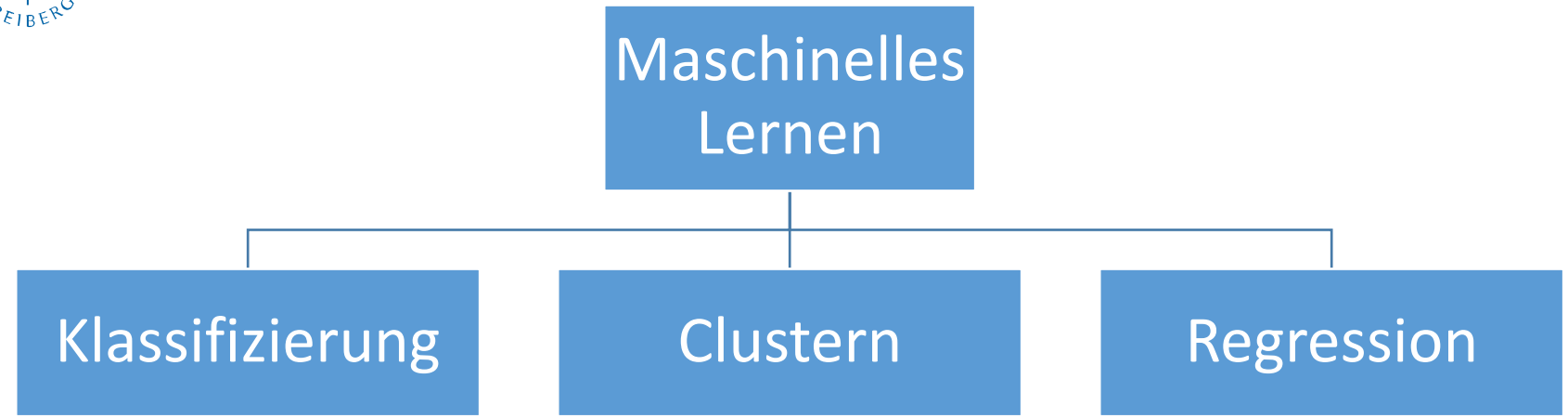
Samuel Kost
kosts@mailbox.tu-freiberg.de
Institut für Numerische Mathematik und Optimierung

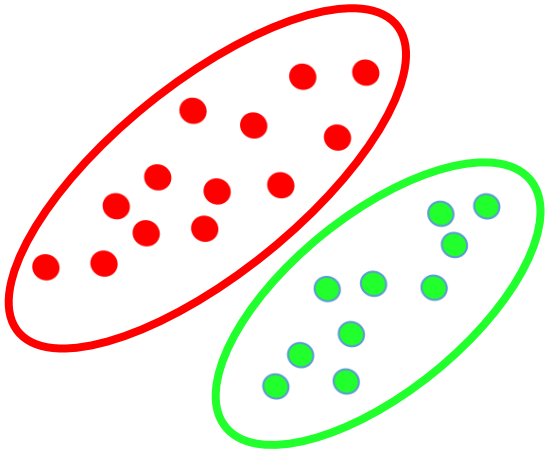
Modellierung mit künstlicher Intelligenz

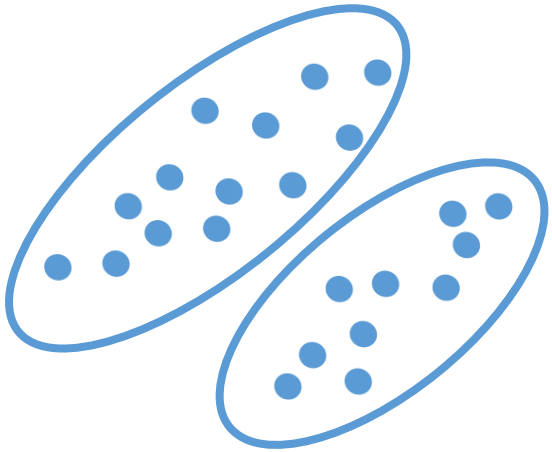
Ein Überblick über existierende Methoden des maschinellen Lernens

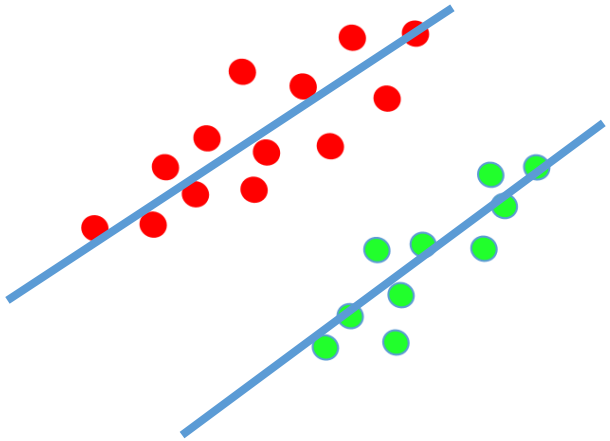
13. Sächsisches GIS-Forum, Dresden, 27.01.2016

Was versteht man unter maschinellem Lernen?









Vorbemerkungen

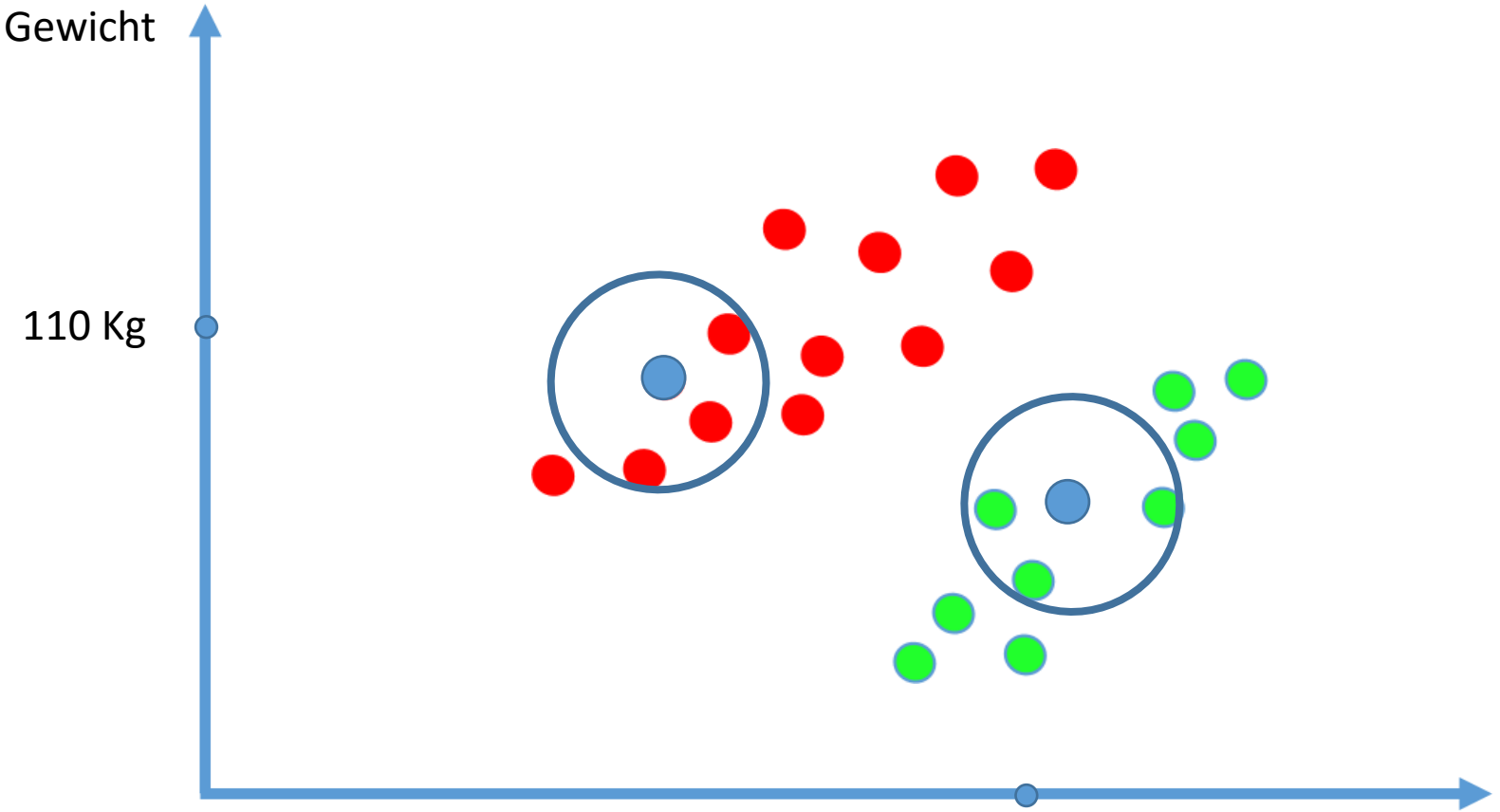
Instanz	Merkmale
	(198,98)
	(179,74)

Vorbemerkungen

Instanz	Merkmale
	(198,98)
	(179,74,120,80,125,75,115,65)

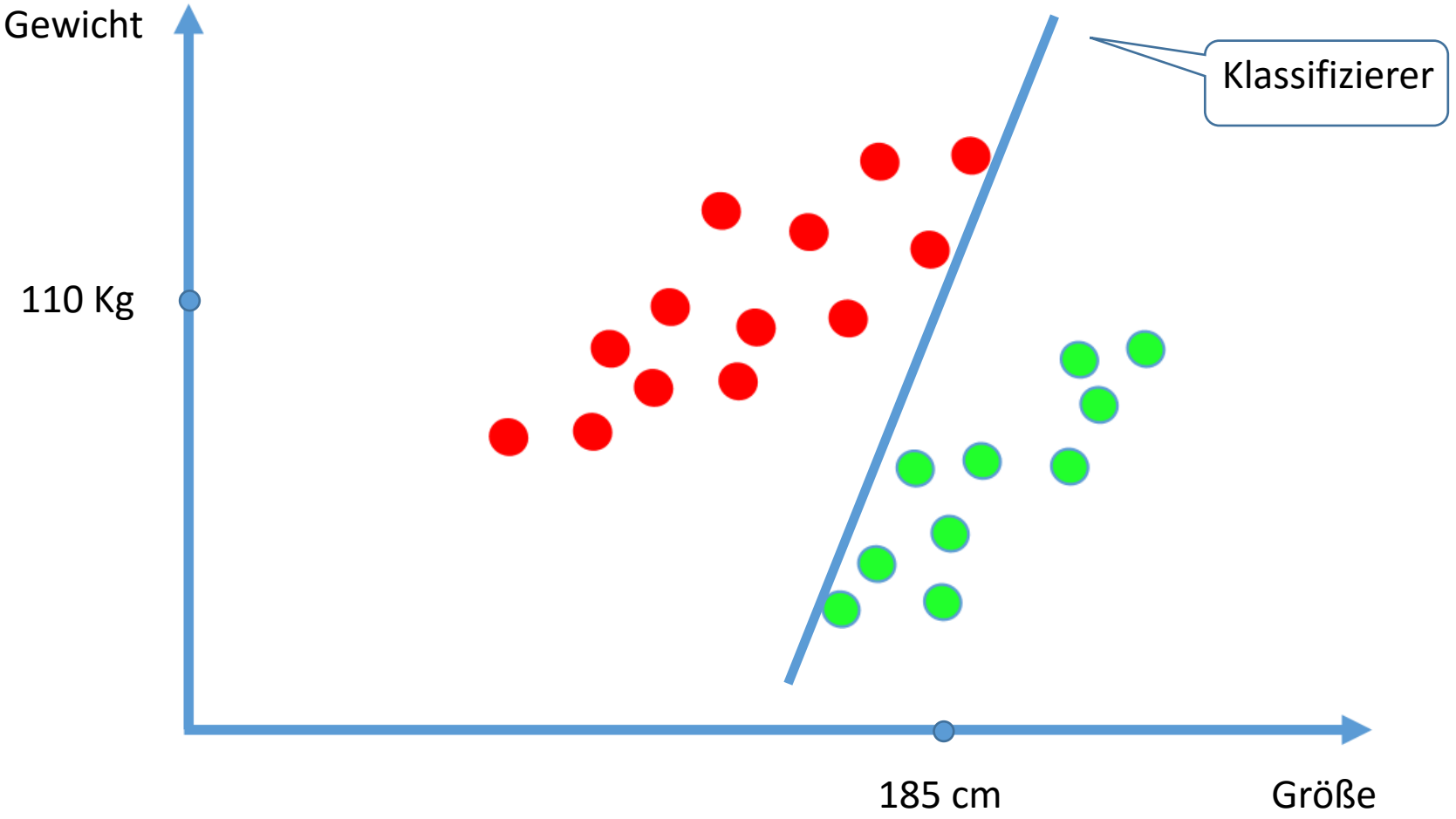
- Auswahl der Merkmale sollte wenn möglich sorgfältig per Hand durchgeführt werden
- Es gilt die Regel: **Nonsense als Input → Nonsense als Output**

k-Nearest Neighbors

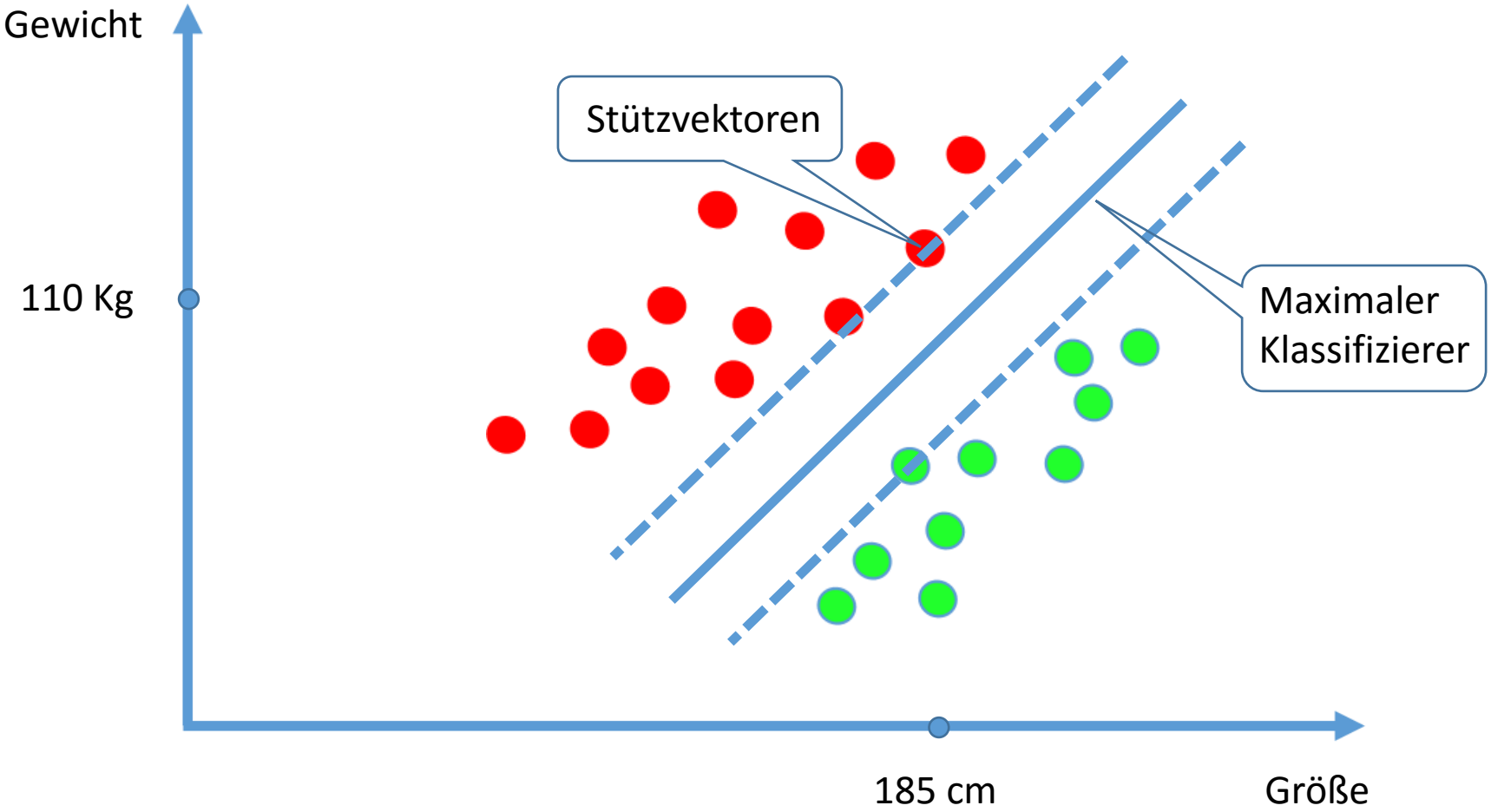


Wo liegt die Schwierigkeit?

Support Vector Machine (SVM)



Support Vector Machine (SVM)



Support Vector Machine (SVM)

Wann sollte SVM verwendet werden?

- Bei ungleichen Klassengrößen
- (Bei nur nichtlinear trennbaren Klassen)
- Bei hochdimensionalen Daten
- Wenn hohe Genauigkeit notwendig ist
- Wenn nicht zu viele Instanzen vorhanden sind ($<10^6$)
- Falls die Daten eine geometrische Interpretation haben

Logistische Regression

- Klassifizierungsmethode, keine Regression!

- Es gilt folgender Zusammenhang:

$$\frac{p(gut\ in\ Basketball)}{p(schlecht\ in\ Basketball)} \sim C_1 * Größe + C_2 * Gewicht$$

Logistische Regression

- Klassifizierungsmethode, keine Regression!
- Es gilt folgender Zusammenhang:

$$log\frac{p(gut\ in\ Basketball)}{p(schlecht\ in\ Basketball)} = C_1 * Größe + C_2 * Gewicht$$

- Die zu lösende Aufgabe ist das Lernen der Koeffizienten C
  - C_1 wäre hier positiv, C_2 wäre negativ
- Mathematische Interpretation der Aufgabe:
  - Finden des globalen Minimums einer nichtlinearen Funktion

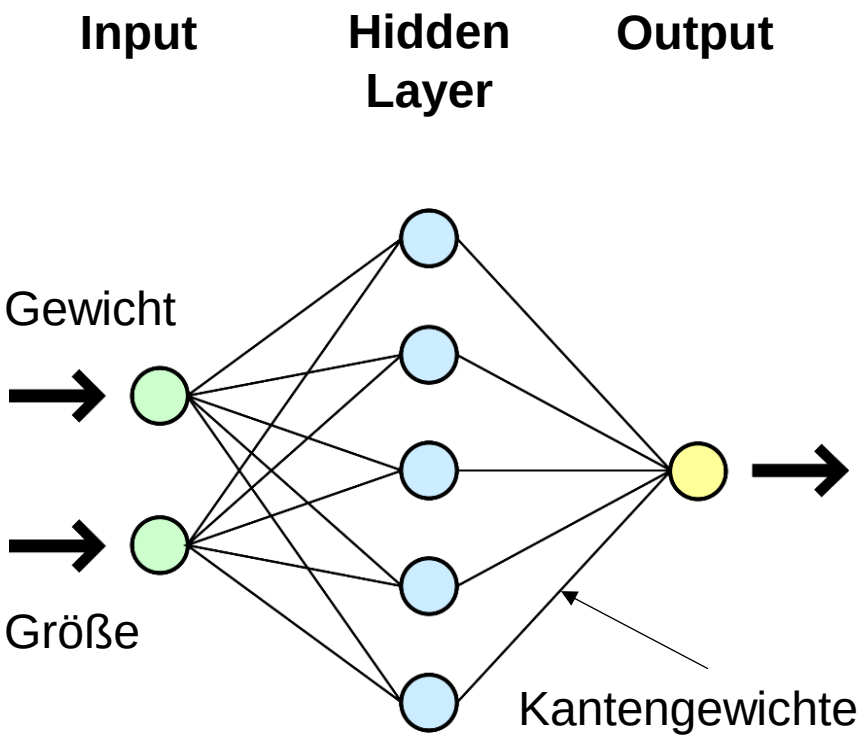
Logistische Regression

Wann sollte Logistische Regression verwendet werden?

- Falls Wahrscheinlichkeiten eine Rolle spielen
  - Vorkenntnisse sind leicht in das Modell zu integrieren
- Falls die Anzahl der Merkmale nicht zu groß ist
  - Statistische Größen verfügbar für Wichtung der Merkmale
  - Sehr gute Interpretierbarkeit
- Wenn die Trainingsgeschwindigkeit eine Rolle spielt
  - Logistische Regression trainiert sehr schnell
- Wenn keine sehr hohe Genauigkeit benötigt wird
- Wenn Klassen linear trennbar sind

Künstliche Neuronale Netze

- Verallgemeinerung von Logistischer Regression
- Aktivierungsfunktion der Summe gewichteter Outputs als Input für neue Schicht
- Training heißt einstellen der Gewichte an den Kanten
- Mathematisch: Finden eines Minimums einer nichtlinearen Funktion (viele lokale Minima)

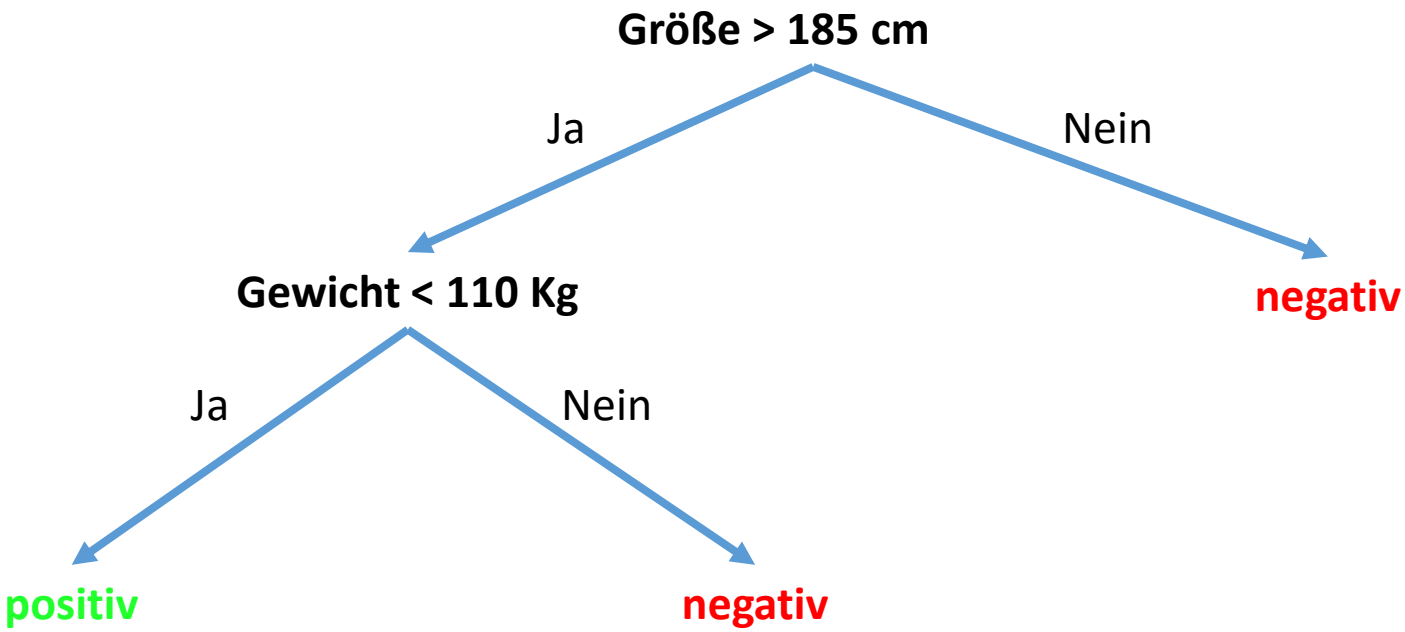


Künstliche Neuronale Netze

Wann sollten Neuronale Netze verwendet werden?

- Falls keine statistischen Kenngrößen benötigt werden
- Wenn die Trainingsgeschwindigkeit keine Rolle spielt
  - Training gewöhnlich sehr langsam und aufwendig
- Wenn eine hohe Genauigkeit benötigt wird
- Wenn Klassen nichtlinear trennbar sind
- Für alle möglichen Problemklassen einsetzbar

Entscheidungsbaum



Random Forest

- Forest besteht aus vielen flachen Entscheidungsbäumen
- Alle werden auf neue Daten angewendet
- Ein Entscheidungsmechanismus wird für Ergebnis verwendet

Verstärkte Entscheidungsbäume

- Entscheidungsbäume werden iterativ aufgebaut
  - Neue Bäume für Fälle die bisher noch nicht im aktuellen Random Forest betrachtet wurde
- Sehr starke Methode
  - Aktuell state-of-the-art bei der Websuche

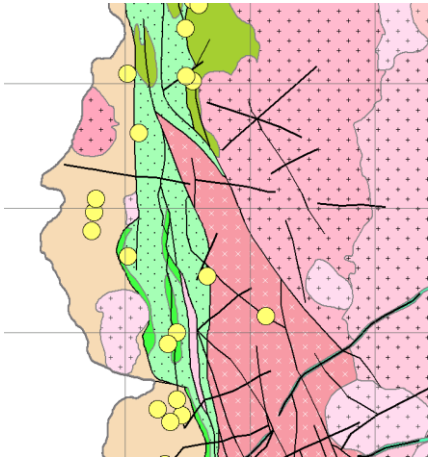
Random Forest

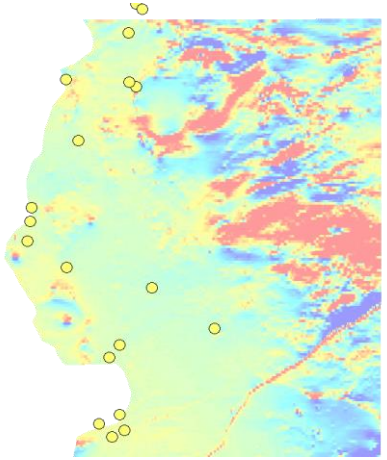
Wann sollte Random Forest verwendet werden?

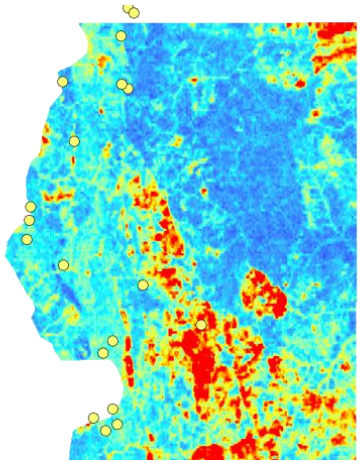
- Wenn die Klassen ausgeglichen sind
  - Bei 10 positiven und 1M negativen Instanzen werden alle Bäume für negativ stimmen
- Bei sehr vielen Klassen
  - Man kann so viele Klassen wie Blätter in den Bäumen haben
- Bei hochdimensionalen Daten
  - Z.B. Textklassifizierung

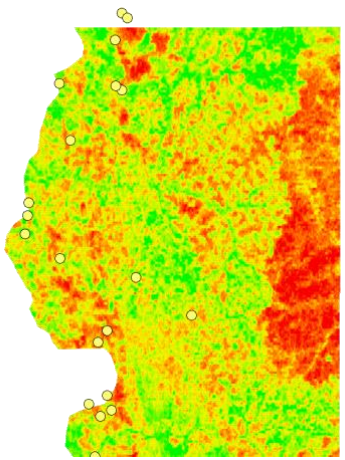
Beispieldatensatz: Goldvorkommen in Ghana

Potentielle Einflussfaktoren / Merkmale









Geophysikalische Befliegung

Geologie und Tektonik

Magnetik

Kalium

Thorium

Beispieldatensatz: Goldvorkommen in Ghana

- Datensatz mit 17 Merkmalen
  - binär und kontinuierlich
- Fast 1 Mio. Instanzen (99,5% negativ)

Verfahren	Zeit in Sekunden	Präzision (AUC-score)
Neuronale Netze	451	0.84
Logistische Regression	2	0.78
Random Forest (100)	884	0.87
Support Vector Machine	11563	0.64

Weitere Anwendungsbeispiele

- Erosionserscheinungen
- Hangbewegungen
- Manganknollen Belegungsdichte
- Kohlefeuer in China
- Befall von Wäldern durch Borkenkäfer
- Regionalisierung von Schadstoffen
- Lagerstätten versch. Genese
- Brutplätze von Vogelarten
- Geochemischer Atlas (Regionalisierung von Punktdaten)
- Schadstoffe in Siedlungsgebieten
- Stabilität von Kippen stillgelegter Tagebaue