# Maschinelles Lernen im Labor

BVL Symposium 2021

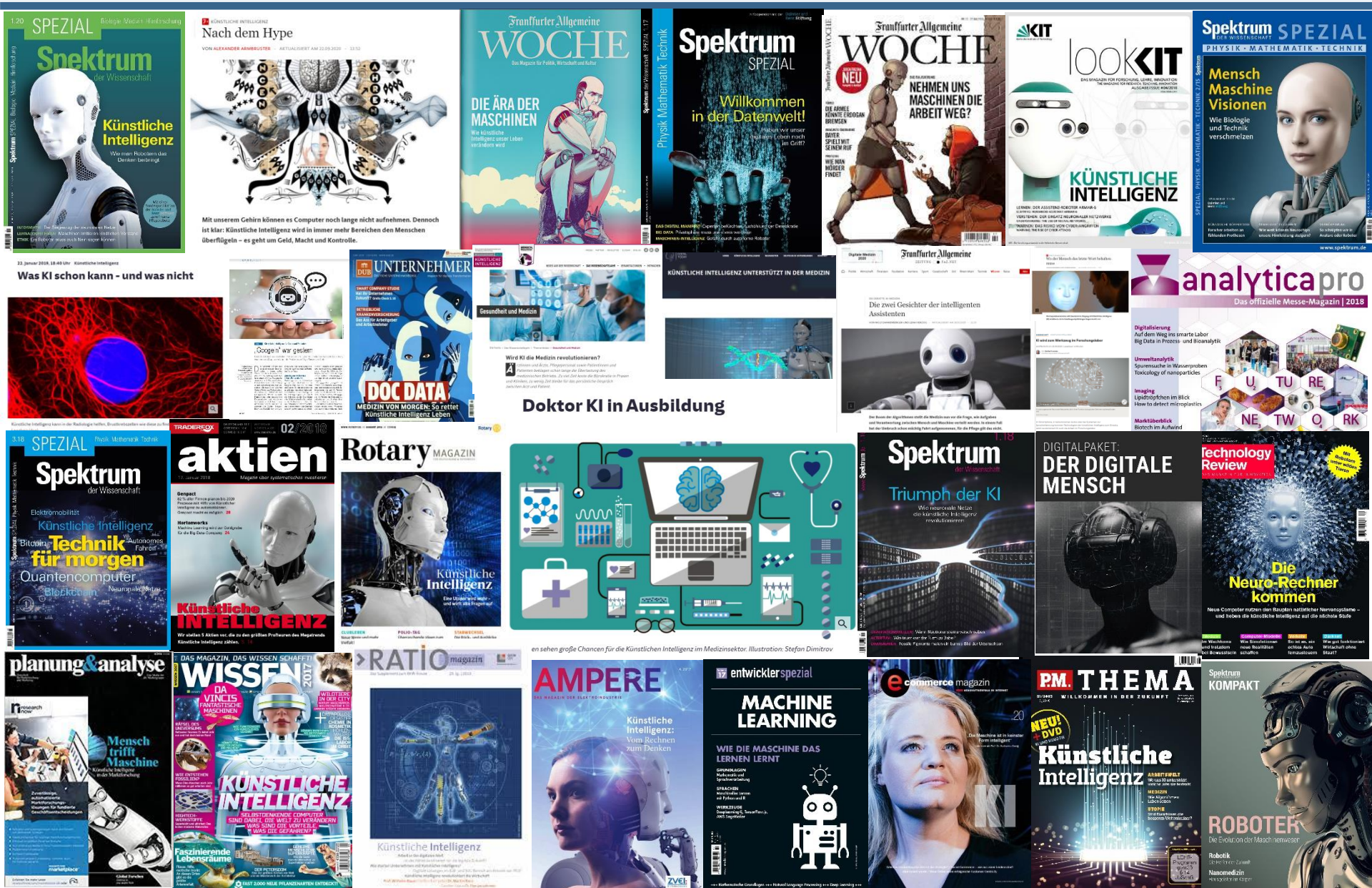
Data Science / Labor 4.0 - Neue Formen der Datengewinnung und –analyse

06.10.2020

MSc. Kapil Nichani
kapil.nichani@quodata.de

quo data
*QUALITY & STATISTICS!

# KI auf den Titelseiten

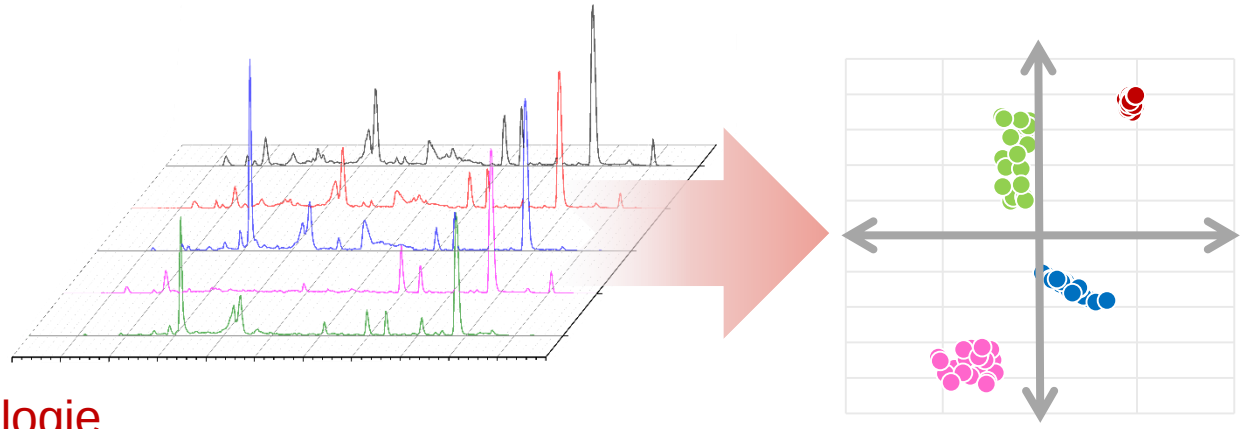


# Was heißt Maschinelles Lernen (ML)?

## Wie lernen Maschinen?

Einerseits durch
nicht beaufsichtigtes Lernen

- Dimensionsreduktion von Spektren
- Korrelationen



## Analogie

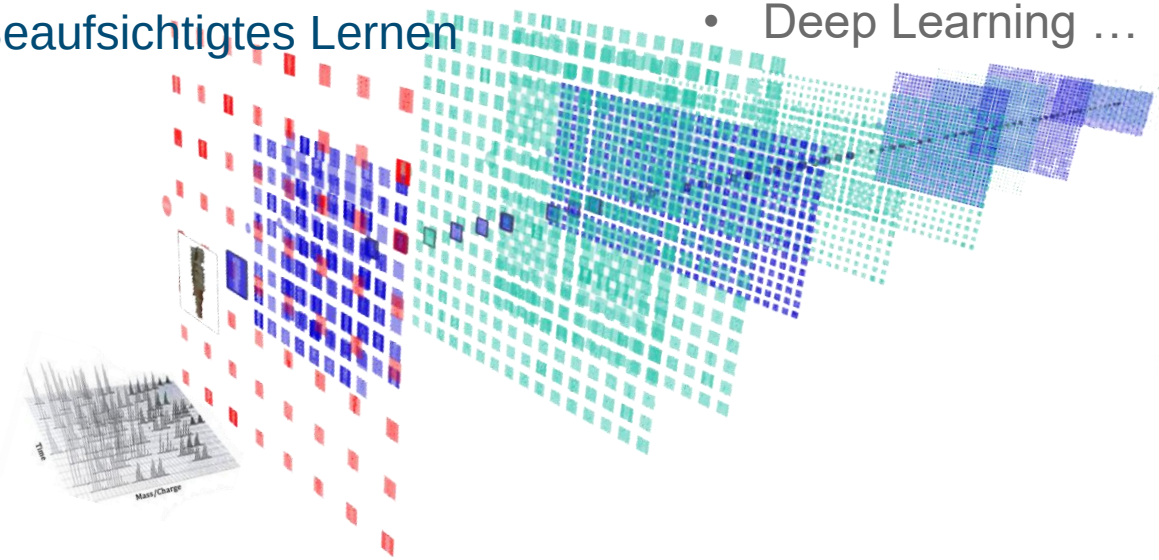
Aus Beobachtungen alleine können keine Ursache-Wirkungs-Beziehungen gelernt werden

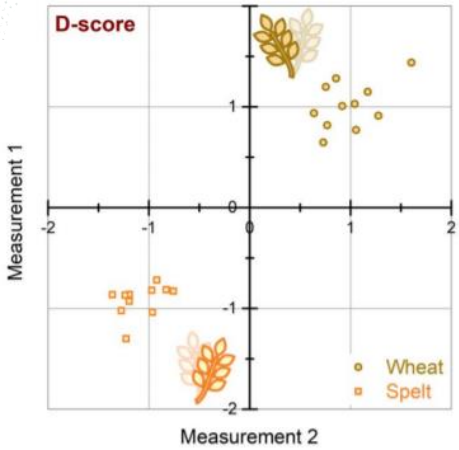
# Was heißt Maschinelles Lernen (ML)?

## Wie lernen Maschinen?

Andererseits durch
Beaufsichtigtes Lernen

- Regression, Klassifizierung, Regression Trees, LASSO..
- Deep Learning …






## Analogie:

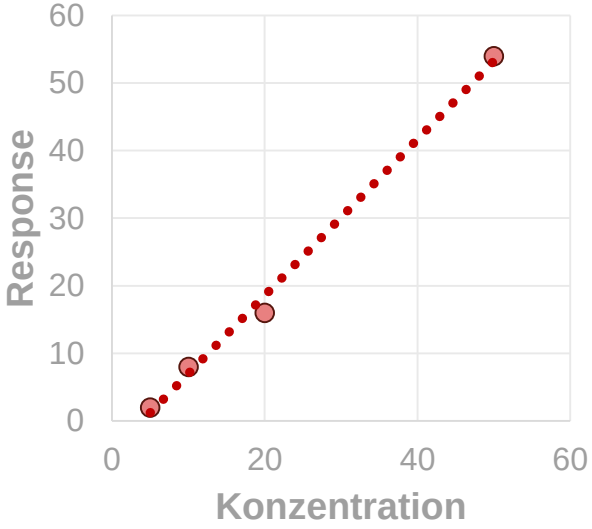
Training mit einer Auswahl von Proben/Mustern, bei denen die zu erlernenden Eigenschaften (Label) bekannt sind.

Kapil Nichani; Steffen Uhlig; Bertrand Colson; Karina Hettwer; Kirsten Simon; Josephine Bönick; Carsten Uhlig et al. "AI-based identification of grain cultivars via non-target mass spectrometry." bioRxiv (2020).

Image adapted from tensorspace.org

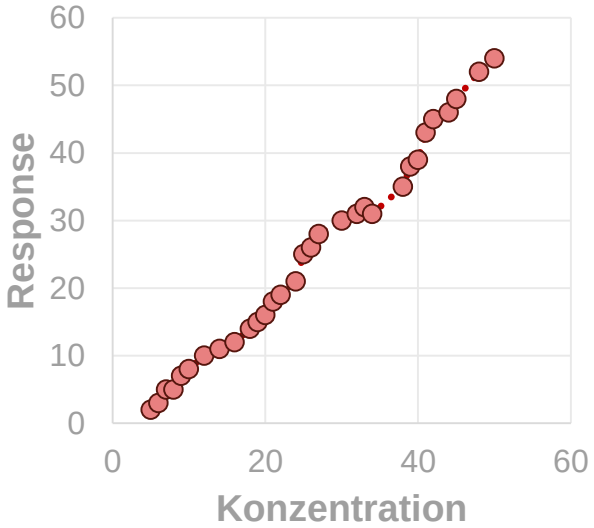
# Wie bringt man der Maschine die Naturgesetze bei?

ML, welches die eine einfache lineare Kalibrierkurve kennt




Benötigt nur wenige Proben

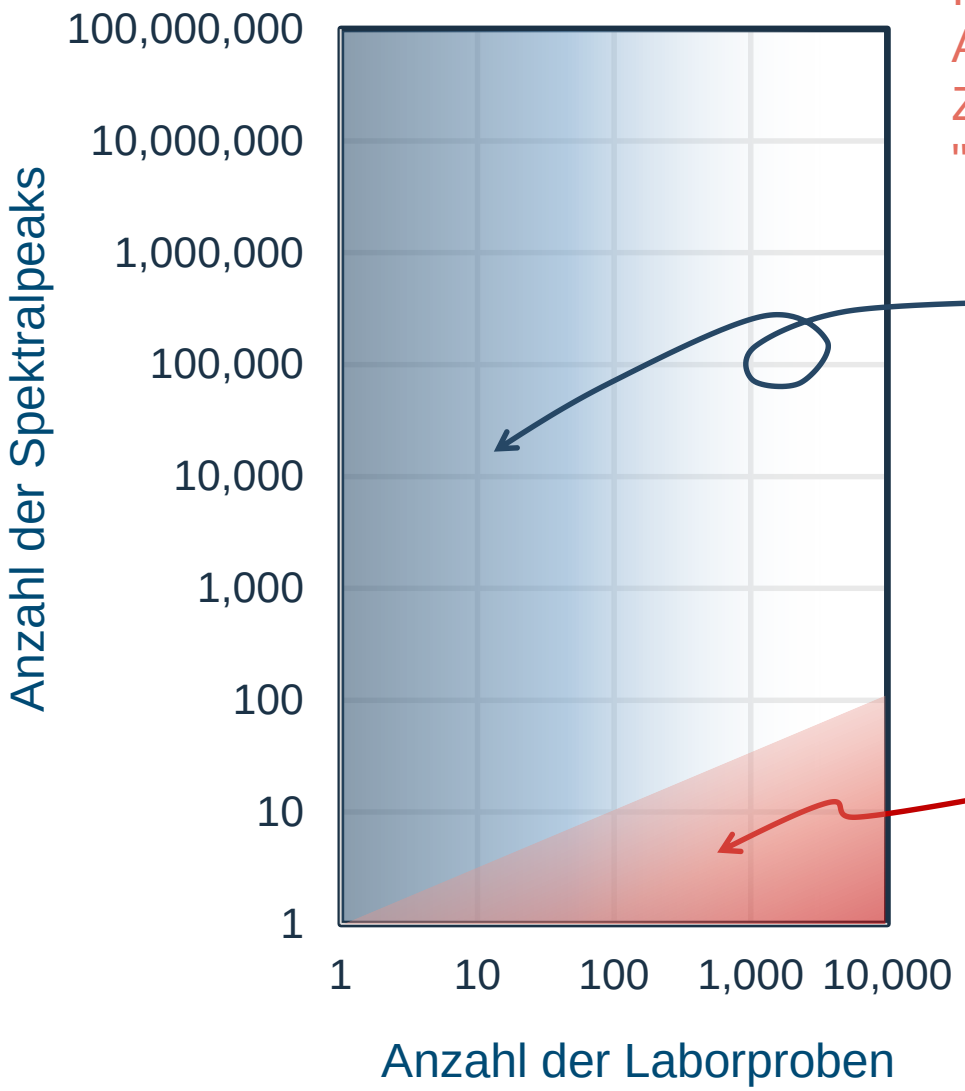
ML, welches die lineare Kalibrierkurve nicht kennt




Benötigt Hunderte von Proben

ML steht noch am Anfang. So gibt es bislang noch kein ML, welche – ohne explizite Unterstützung – Naturgesetze erlernt und anzuwenden versteht

# Größe des einzelnen Datensatzes (Stichprobe) vs Anzahl der Datensätze aus verschiedenen Stichproben




Im Vergleich zu anderen Anwendungsbereichen ist die Zahl der Datensätze im Labor "endlich" und "begrenzt".

Typische Non-target Methoden

Klassische Statistik: Anzahl der Features muss kleiner sein als die Anzahl der Probe

# Trainieren des Neuronalen Netzes

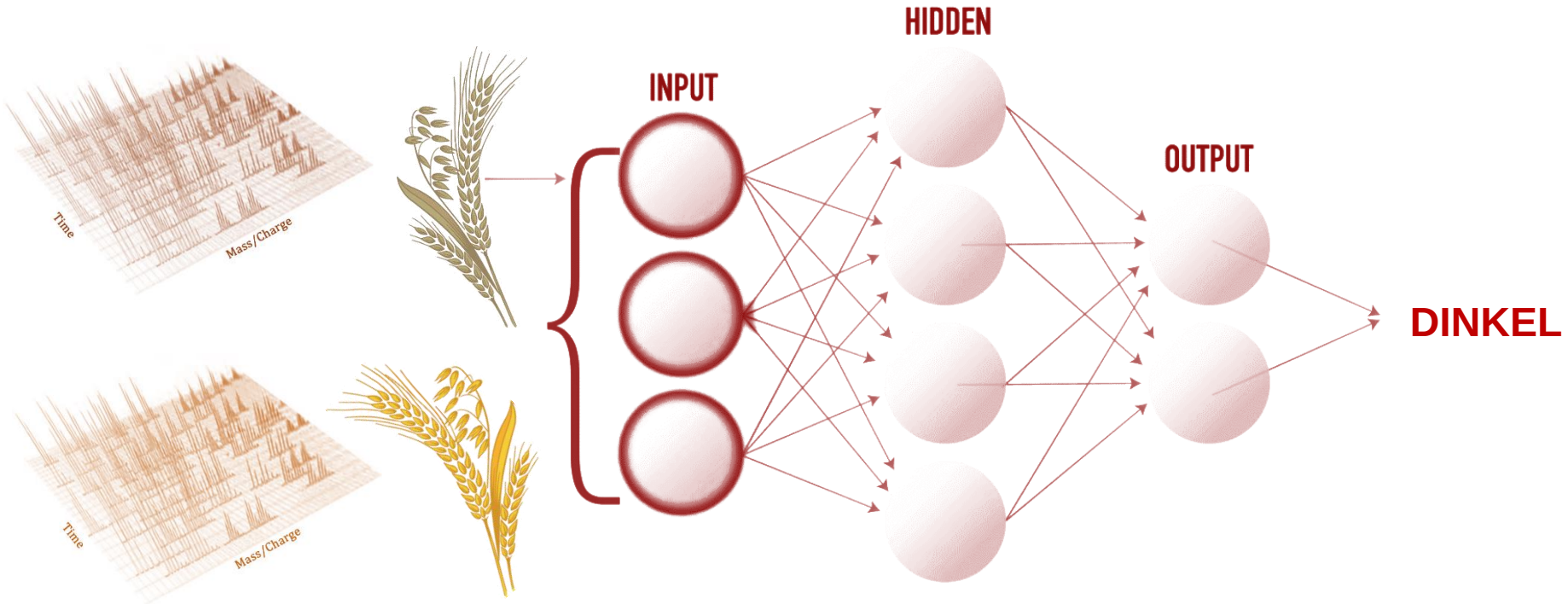



Image modified & adapted. Original image source: CMU Dept of ML

# Maschinelles Lernen erfordert akribisches Sammeln von Daten

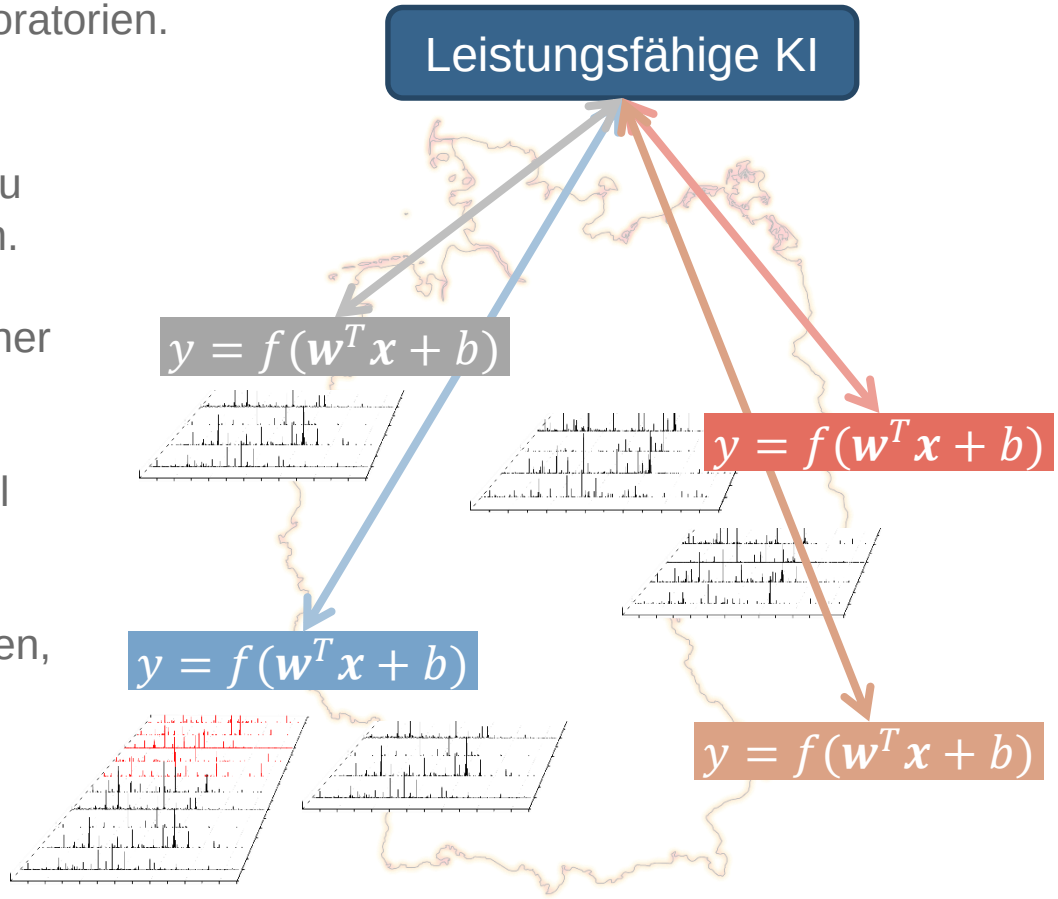
Um KI zu trainieren, müssen wir zunächst über möglichst viele Proben möglichst viel wissen:

- Dinkelgehalt im Dinkelbrot mindestens 90%?
- Ist die Glutenkonzentration im Dinkelbrot > 0?
- Handelt es sich bei dem Dinkel im Brot tatsächlich um Dinkel?
- Was genau ist Dinkel?

Validierung der zur Training der KI benötigten Daten erforderlich

# Notwendig:Kooperation der Laboratorien und Entwicklung von Datenbanken

- Weil in aller Regel die Verfügbarkeit von vollständig charakterisierten Proben begrenzt ist, empfiehlt sich für die Methodenentwicklung die Analyse der Proben jeweils in 2 oder 3 Laboren, weil von Labor zu Labor mit erheblichen Unterschieden zu rechnen ist.
- Aufgrund der benötigten Probenzahlen empfiehlt sich zugleich die Aufteilung der Arbeiten auf mehrere Schultern/Laboratorien.
- Die Qualität der Metadaten ist von fundamentaler Bedeutung, weil fehlerhafte Angaben immer wieder zu falschen Ergebnissen führen können.
- Das gemeinsame Lernen von einer Gruppe von Laboratorien führt zu einer leistungsfähigen KI, von der alle Laboratorien profitieren können.
- Die sehr aufwändige KI kann parallel (und ohne hohe zusätzliche Kosten) auf unterschiedliche Problemstellungen angewandt werden, wenn entsprechende Metadaten zur Verfügung stehen.



# Diskussion

Maschinelles Lernen wird seit vielen Jahren in der Chemometrie in Form von nicht beaufsichtigtem Lernen z.B. zur Dimensionsreduktion mittels PCA angewandt.

Einige neuere Verfahren des beaufsichtigten Lernens haben sich in ersten Anwendungen als äußerst vielversprechend erwiesen, zum Beispiel für die Ermittlung von Fischspezies, Getreidearten und Bakterienresistenz

Gleichwohl müssen ML-Verfahren weiter verbessert und angepasst werden, um den besonderen Bedürfnissen des Labors gerecht zu werden.

Dabei ist einerseits die regelmäßig sehr geringe Anzahl von Proben bzw. Datensätzen zu beachten, andererseits die Möglichkeit, bei der Herstellung der Proben und der Durchführung der Messungen Methoden der experimentellen optimalen Versuchsplanung heranzuziehen.

Aufgrund der geringen Probenzahlen erscheint die Integration von Expertenwissen und insbesondere von Naturgesetzen in das ML ganz wesentlich, um neue effiziente Untersuchungsmethoden zu entwickeln.

Umgekehrt ist es von zentraler Bedeutung, dass das ML befähigt wird, ihr neu erworbenes Wissen mit den menschlichen Experten zu teilen. Es geht also um neue Formen der Zusammenarbeit zwischen Mensch und ML.

ML hat für das analytische Labor enormes Potenzial. Weil aber nie ausgeschlossen werden kann, dass das ML auch Fehler macht, müssen mehr Ressourcen für die Validierung verfügbar gemacht werden.

**Kapil Nichani**

kapil.nichani@quodata.de

quodata.de

@Quodata

Vielen Dank!

quo data

*QUALITY & STATISTICS!