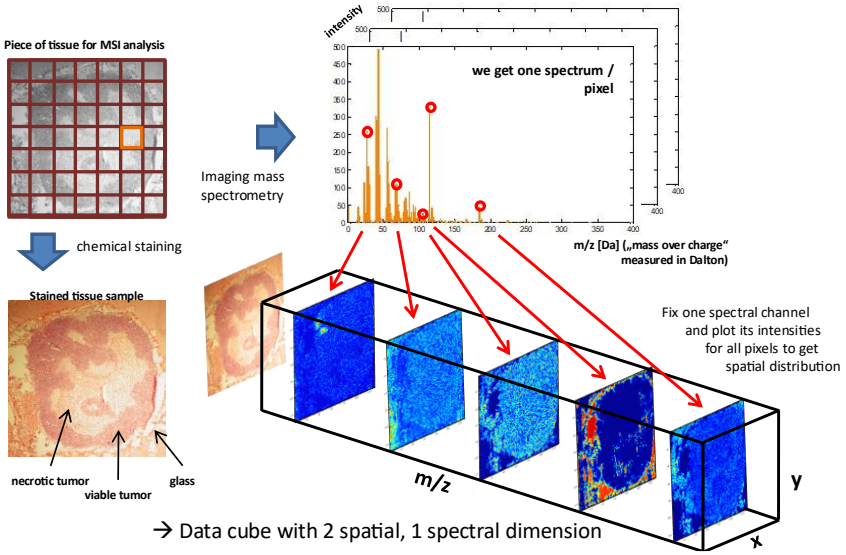
Programmentwurf Data Science

Gegeben ist der aus den Vorlesungsfolien bekannte Datensatz zur bildgebenden Massenspektrometrie (mass spectrometry imaging - MSI). Sie können die Daten von moodle herunterladen. Wie in der Vorlesung besprochen handelt es sich um einen Datenwürfel mit zwei räumlichen (128x128) und einer spektralen Dimension (ursprünglich mehrere tausend Merkmale, hier schon reduziert auf 191 Features). Der Datensatz zeigt eine Gewebeprobe, in der verschiedene Regionen vorkommen (siehe unten, z.B., Krebsgewebe, Gelatine, durchscheinender Glasträger).

Für diese Aufgabe gehen wir davon aus, dass uns diese (durch histochemische Färbung) erhaltene Labels *nicht* für die Analyse bekannt sind, stattdessen verwenden wir unüberwachte Methoden. Sie dürfen das Wissen über die Labels aber im Rahmen der Diskussion ihrer Ergebnisse verwenden.

Data coutesy of Ron Heeren, see Hanselmann 09

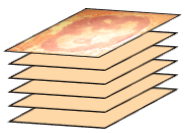
Mass Spectrometry Imaging (MSI)

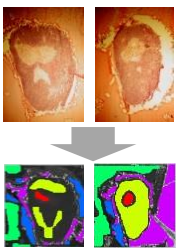


MSI Dataset of a Rat Brain

- Sectioning of rat tissue into parallel slices
- Stain each 2nd slice and label data five classes of interest:

  necrotic tumor, viable tumor, interface region, gelatin, glass
- Use remaining slices for MSI





1. Data Understanding (5 Punkte)

Laden Sie den Datensatz und machen Sie sich mit den Daten vertraut:

```
ims_cube = scipy.io.loadmat("ims_cube.mat")['ims_cube']
```

- Plotten Sie die räumliche Verteilung einzelner Features/Kanäle (das sind also Bilder analog zu den bläulich gefärbten Bildern in der Grafik oben)
- Plotten Sie weiterhin die in den Daten enthaltenen Spektren (mit jeweils 191 Features) und analysieren Sie deren Eigenschaften. Nutzen Sie dazu auch Visualisierungstechniken aus der Vorlesung und adaptieren Sie sie in geeigneter Weise (falls notwendig).

Beschreiben Sie, welche Erkenntnisse Sie aus diesen Untersuchungen über die Daten ziehen können. Welche Konsequenzen ergeben sich, wenn Sie die Daten mit den unten eingesetzten Verfahren weiter analysieren möchten?

2. Data Preparation (2 Punkte)

Bereiten Sie die Daten gemäß obiger Erkenntnisse für die weitere Verarbeitung auf.

3. Modeling & Evaluation: Clustering (7 Punkte)

Wie anhand der oben gezeigten Labels ersichtlich ist, sind in den Daten verschiedene (Gewebe-)regionen enthalten. Die Erstellung solcher Labels ist jedoch sehr aufwändig und erfordert typischerweise das chemische Anfärben von Gewebeschichten. Zudem kann aus technischen Gründen meist nicht dieselbe Schicht angefärbt und mittels MSI analysiert werden, so dass auch kein exaktes Labeling möglich ist. Folglich sind häufig nur die aufgenommenen Daten, aber keine Labels verfügbar. Zeigen sie nun, wie Sie in solchen Fällen mit unüberwachten Methoden einen Biologen oder Arzt unterstützen können, indem Sie Datenanalyse betreiben.

- Wählen Sie ein geeignetes Clusteringverfahren aus, um Pixel in dem 128x128 großen Bild zu identifizieren, die bezüglich ihrer (spektralen) Features ähnlich sind. Begründen Sie Ihre Wahl und beschreiben Sie kurz in *eigenen* Worten, wie das gewählte Verfahren arbeitet.
- Führen Sie das Clustering mit geeignete(r/n) Parametrierung(en) durch.
- Dokumentieren Sie Ihre Ergebnisse (Visualisierungen, quantitative Evaluation) und beschreiben Sie, welche Erkenntnisse Sie aus der Analyse ziehen (in dieser Diskussion dürfen Sie sich auch auf die gezeigten Labels beziehen).

4. Modeling & Evaluation: PCA und Alternativen (10 Punkte)

Eine andere Möglichkeit ist es, die aufgenommenen Spektren an den Pixelpositionen als Überlagerung der charakteristischen Spektren der dort auftretenden (Gewebe-)arten aufzufassen. Beispielsweise könnte sich hinter einem Pixel eine Mischung aus gesunden und Krebszellen verbergen. Wir sind nun an zwei Dingen interessiert (siehe dazu auch das Skript aus der Vorlesung):

- Identifikation von charakteristischen Spektren, mit deren Hilfe die Daten erklärt werden können (jeweils 1x191)
- Plotten der räumlichen Verteilung für diese charakteristischen Spektren (jeweils 128x128)

Verwenden Sie im Folgenden eine Principal Component Analysis (PCA) und eine zweite Methode nach Wahl (s.u.). Sie dürfen gerne verfügbare Implementierungen verwenden, z.B. aus scikit learn:

- Independent Component Analysis (ICA)
- Non-negative Matrix Factorization (NMF), eine Methode, die eng mit der in der Vorlesung besprochenen probabilistic latent semantic analysis (pLSA) verwandt ist und ebenfalls eine nicht-negative Faktorisierung durchführt

In beiden Fällen:

- Bestimmen Sie die Komponentenzerlegung, d.h. im Fall der PCA die Hauptkomponenten (d.h. hier die charakteristischen Spektren) und die zugehörige räumliche Verteilung (die sich aus der Projektion der Daten auf diese Spektren ergibt)
- Visualisieren Sie die Ergebnisse auf geeignete Weise und diskutieren Sie die Ergebnisse! Wie viele Komponenten sind sinnvoll und warum?
- Welche Vor- und Nachteile sehen Sie bei den Ansätzen?

  Hinweis: Bei ICA oder NMF können Sie ähnlich wie bei PCA (scores, loadings) eine Zerlegung in charakteristische Spektren und Mischverhältnisse durchführen.

5. Business Understanding und Deployment (6 Punkte)

Entwickeln Sie ein Geschäftsmodell für ein Produkt oder eine Dienstleistung, die Sie mit oben entwickelten Algorithmen und Erkenntnisse anbieten könnten. Gehen Sie dabei strukturiert vor und nutzen eine der in der Vorlesung vorgestellten Frameworks, um alle relevanten Aspekte zu bewerten. Stellen Sie Ihre Erkenntnisse als Markup in Jupyter dar, fokussieren Sie sich dabei auf das Wesentliche (Gedankenmodell: Basis für einen kurzen aber knackigen Pitch bei einem Venture Capitalist oder einem firmeninternen Manager).

Bewertungskriterien

1. Fachliche Bewertung (50%): Vollständigkeit, Korrektheit, Lösungsqualität und Eleganz sowie Klarheit und Umfang der Betrachtung, Einreichung von lesbarem und kompilierenden Python Code, korrekte Verwendung von wichtigen Funktionen / Bibliotheken, Güte
2. Dokumentation (50%): Dokumentation des Vorgehens der Datenauswertung im Sinne von Data Science, Codekommentare wie in der Informatik üblich wo notwendig, Qualität der Diagramme, Markup, Texte, pdf.

Neu: Beachten Sie unsere Regeln zur Nutzung von ChatGPT und anderen generativen KIs. Analog zu normalen Internetquellen (dort Datum und Link angeben) dürfen Sie ChatGPT nutzen, sofern Sie sich an unsere Regeln halten. Die Datendatei darf aus rechtlichen Gründen ***nicht*** online hochgeladen werden.

Abgabe bis zum 06.11.2025, 18 Uhr

Bearbeitung findet in Gruppen mit jeweils **genau 2 Personen** statt oder als freiwillige Einzelarbeit. Alle Ergebnisse sind einzureichen als **Jupyter Notebook** über **Moodle**.