Universität Basel
Departement Altertumswissenschaften
Alte Geschichte

Authorship Attribution der umstrittenen Paulusbriefe. Mit maschinellem Lernen auf der Spur von Mister X

Vortrag von Johannes Nussbaum (Universität Basel)
im Rahmen des Kolloquiums «Zur neueren Forschung in der Alten Geschichte»
30.03.2021

Diese Präsentation inkl. Paper & Programmcode ist verfügbar auf: https://github.com/jnussbaum/authorship-attribution

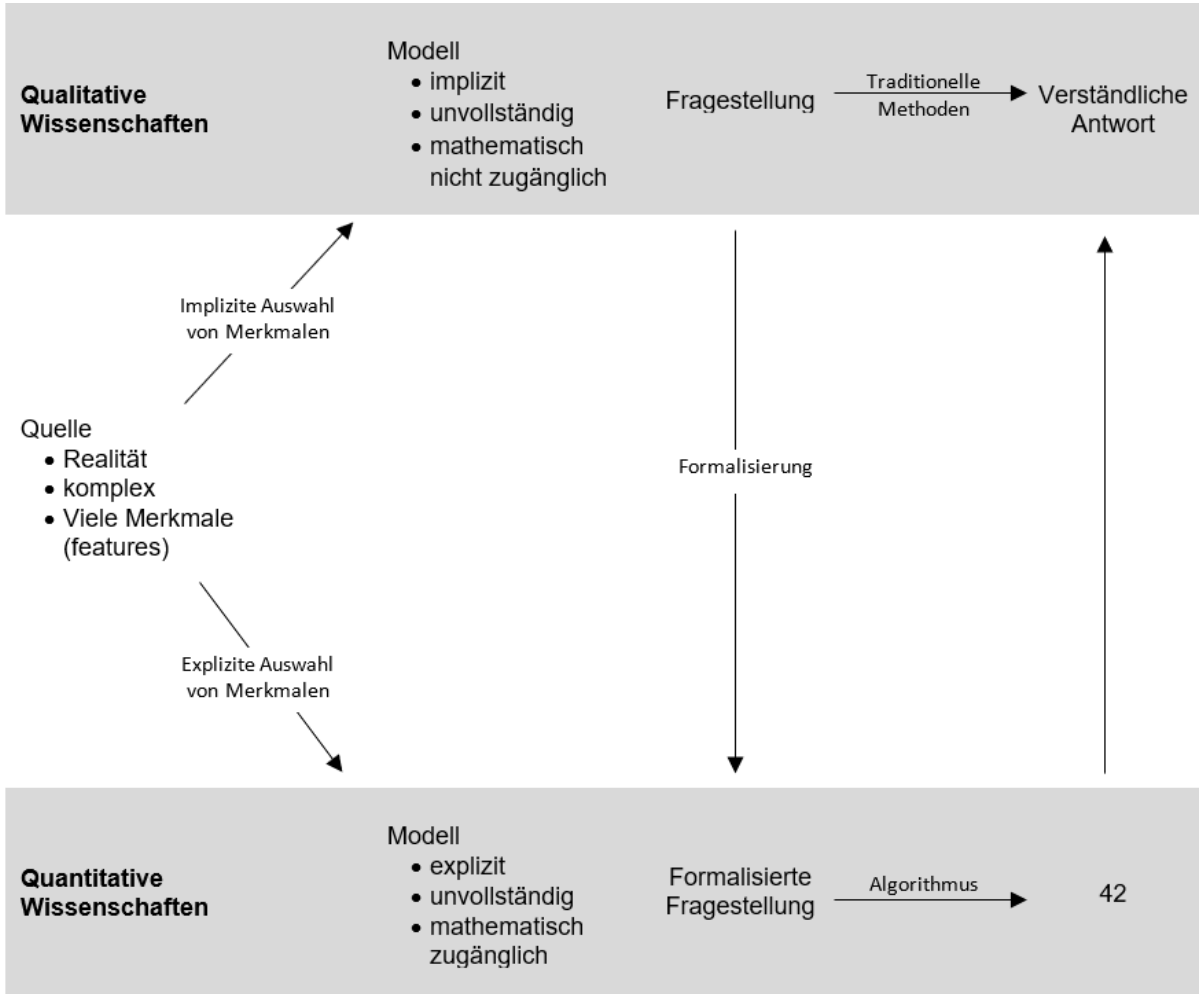
Einstieg: Um was geht es heute Abend?

– Seminararbeit
– Frage: Welche Paulusbriefe im Neuen Testament wurden wirklich vom Apostel Paulus geschrieben?
– Methode: Statistik, Machine Learning, Künstliche Intelligenz, …

…was ist das denn genau?

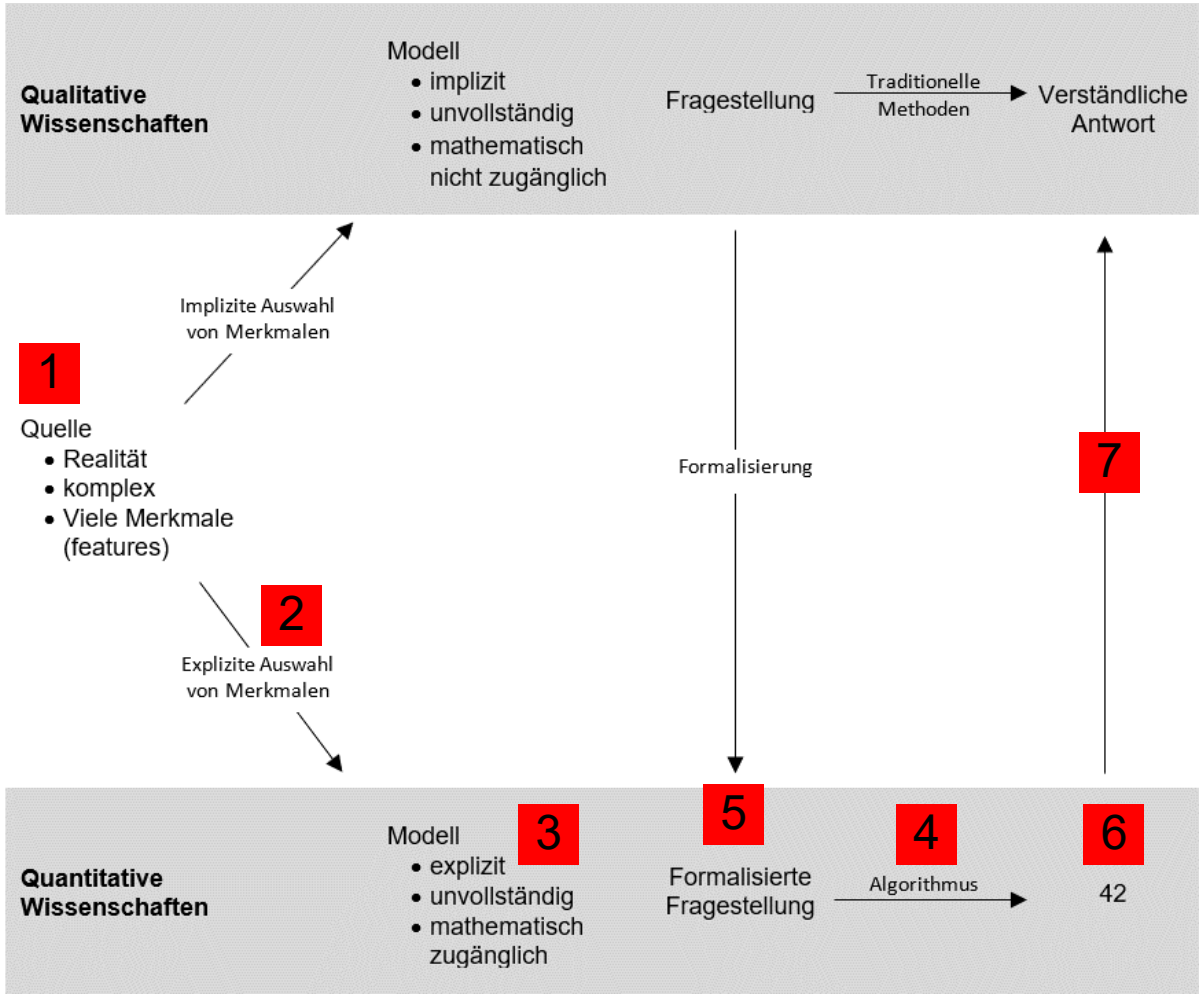
– Erklärung von wichtigen Konzepten
– Praxisbericht, wie ich vorgegangen bin
– Anhand der Fallstudie zu den Paulusbriefen betrachten wir die Arbeitsschritte.

Epistemologische Grundlage: Modellbildung



– Digital Humanities formalisieren den Prozess der Erkenntnisgewinnung.
– Streng logische Operationen können am besten an einem Modell durchgeführt werden.
– Modell: Vereinfachung der Realität
– Es gibt viele verschiedene Arten, die Realität im Modell abzubilden.

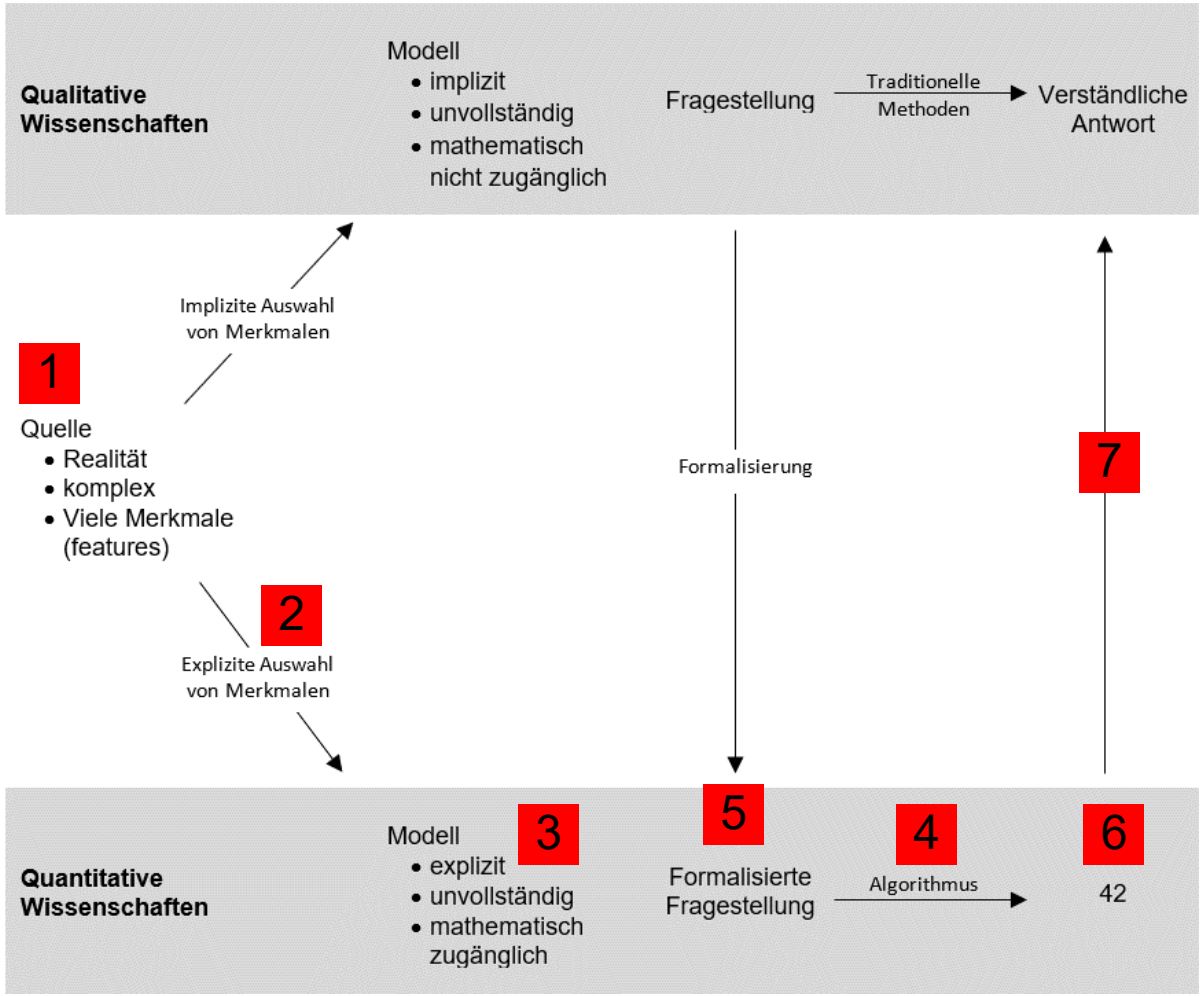
Ablauf



Anhand der Fallstudie zu den Paulusbriefen betrachten wir die Arbeitsschritte:

1. Der griechische Text des Neuen Testaments
   …und seine digitale Verfügbarkeit
2. Textmerkmale
3. Textrepräsentation (Modell)
4. Textklassifikation
5. Studiendesign (Formalisierte Fragestellung)
6. Resultate
7. Diskussion

Ablauf



Anhand der Fallstudie zu den Paulusbriefen betrachten wir die Arbeitsschritte:

1. **Der griechische Text des Neuen Testaments …und seine digitale Verfügbarkeit**
2. Textmerkmale
3. Textrepräsentation (Modell)
4. Textklassifikation
5. Studiendesign (Formalisierte Fragestellung)
6. Resultate
7. Diskussion

1. Der griechische Text des Neuen Testaments …und seine digitale Verfügbarkeit

– Neues Testament: 27 griechische Schriften
– 1. Jh. n.Chr., Osthälfte des Römischen Reiches
– verschiedene Autoren
– unterschiedliche Textgattungen
– Einige nennen ihren Autor, andere sind anonym.
– Rasch Konsens in der Verfasserfrage: traditionelle Zuschreibungen
– Röm 16,22: «Ich, Tertius, der ich diesen Brief geschrieben habe, grüße euch in dem Herrn.»
– Gal 6,11: «Seht, mit wie großen Buchstaben ich euch schreibe mit eigener Hand!»
– 1. Thess 1,1: «Paulus und Silvanus und Timotheus an die Gemeinde der Thessalonicher…»
  (im ganzen Brief oft wir-Form)

1. Der griechische Text des Neuen Testaments …und seine digitale Verfügbarkeit

- Problem der genauen Textgestalt (Textkritik)
- Bibelwissenschaftlich breit abgestützter Standardtext: Nestle-Aland (27. Aufl.)
- Daraus die wahrscheinlichen Interpolationen entfernen
- https://github.com/getbible/Unbound-Biola:
  - TXT-Datei in UTF-8 Unicode, alles in Kleinbuchstaben und ohne Akzente
  - Westcott-Hort 1881 mit den Varianten von NA 27
  - Wenn alle mit VAR1 gekennzeichneten Passagen daraus entfernt werden, resultiert NA27 daraus.

1. Der griechische Text des Neuen Testaments …und seine digitale Verfügbarkeit

Matthäus 3,15 in der Datei
https://github.com/getbible/Unbound-Biola/blob/master/Greek__NT_Westcott_Hort_UBS4_variants_Parsed__westcotthort__LTR.txt:

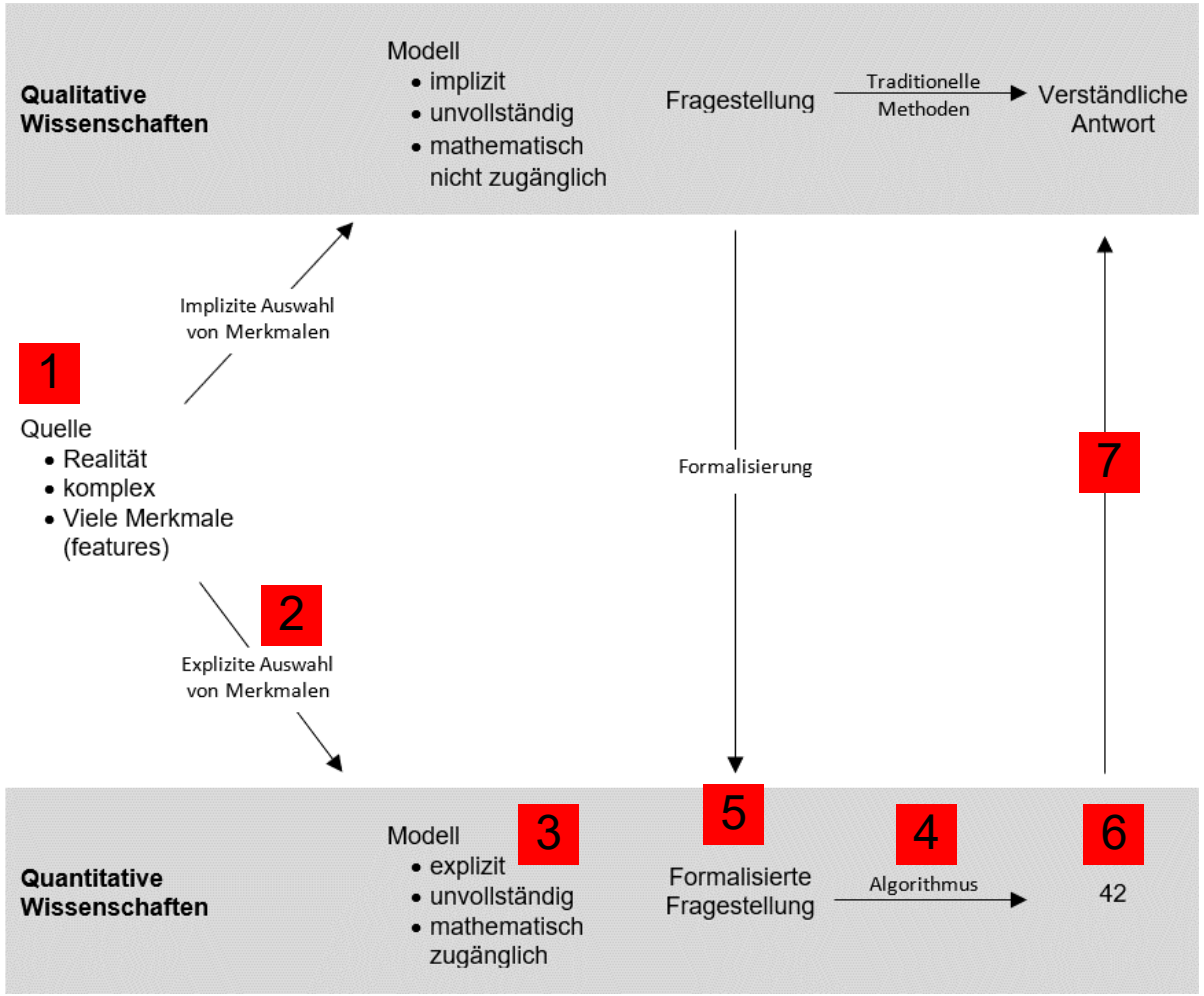
Code für Bibelbuch *Strong-Nummer* *POS-Tag: **V**erb – **A**orist Passivdep**o**nens **P**artizip – **N**ominativ **S**ingular **M**askulin* *Lesarten*

```
40N||3||15||αποκριθεις G611 G5679 V-AOP-NSM δε G1161 CONJ ο G3588 T-NSM ιησους G2424 N-NSM
ειπεν G2036 G5627 V-2AAI-3S {VAR1: αυτω G846 P-DSM } {VAR2: προς G4314 PREP αυτον G846 P-
ASM } αφες G863 G5628 V-2AAM-2S αρτι G737 ADV ουτως G3779 ADV γαρ G1063 CONJ πρεπον G4241
G5901 V-PQP-NSN εστιν G2076 G5748 V-PXI-3S ημιν G2254 P-1DP πληρωσαι G4137 G5658 V-AAN
πασαν G3956 A-ASF δικαιοσυνην G1343 N-ASF τοτε G5119 ADV αφιησιν G863 G5719 V-PAI-3S αυτον
G846 P-ASM
40N||3||16||...
40N||3||17||...
```

1. Der griechische Text des Neuen Testaments …und seine digitale Verfügbarkeit

Griechischer Text:	αποκριθεις δε ο ιησους ειπεν προς αυτον αφες αρτι ουτως γαρ πρεπον εστιν ημιν πληρωσαι πασαν δικαιοσυνην τοτε αφιησιν αυτον.
Strong-Nummern:	G611 G1161 G3588 G2424 G2036 G4314 G846 G863 G737 G3779 G1063 G4241 G5901 G2076 G2254 G4137 G3956 G1343 G5119 G863 G846
POS-Tags:	V-AOP-NSM T-NSM N-NSM V-2AAI-3S PREP P-ASM V-2AAM-2S ADV ADV CONJ V-PQP-NSN V-PXI-3S P-1DP V-AAN A-ASF N-ASF ADV V-PAI-3S P-ASM
POS-Tags (nur erster Teil):	V T N V PREP P V ADV ADV CONJ V V P V A N ADV V P

Ablauf



Anhand der Fallstudie zu den Paulusbriefen betrachten wir die Arbeitsschritte:

1. Der griechische Text des Neuen Testaments
   …und seine digitale Verfügbarkeit
2. **Textmerkmale**
3. Textrepräsentation (Modell)
4. Textklassifikation
5. Studiendesign (Formalisierte Fragestellung)
6. Resultate
7. Diskussion

2. Textmerkmale

Auswahl der Merkmale / feature engineering: komplex

Traditionelle Fragestellung: Gibt es Hinweise/Indizien in den umstrittenen Briefen?

– Methode der konventionellen Stilistik: Gelehrte kennen den Stil eines (antiken) Autors.

Erster Formalisierungsschritt der Fragestellung: Gibt es Ähnlichkeitsverhältnisse zwischen den Briefen?

– Methoden der Stilometrie: Statistik

Statistik braucht etwas Zählbares. → Unterteilung des Textes in zählbare Einheiten

2. Textmerkmale

Tokenisierung: Unterteilung des Textes in kleine Einheiten (Tokens)

Text A: “the the and and and and and”

Text A hat

- 2 Types
- 7 Tokens

Der Type “the” tritt 2 Mal auf.

Der Type “and” tritt 5 Mal auf.

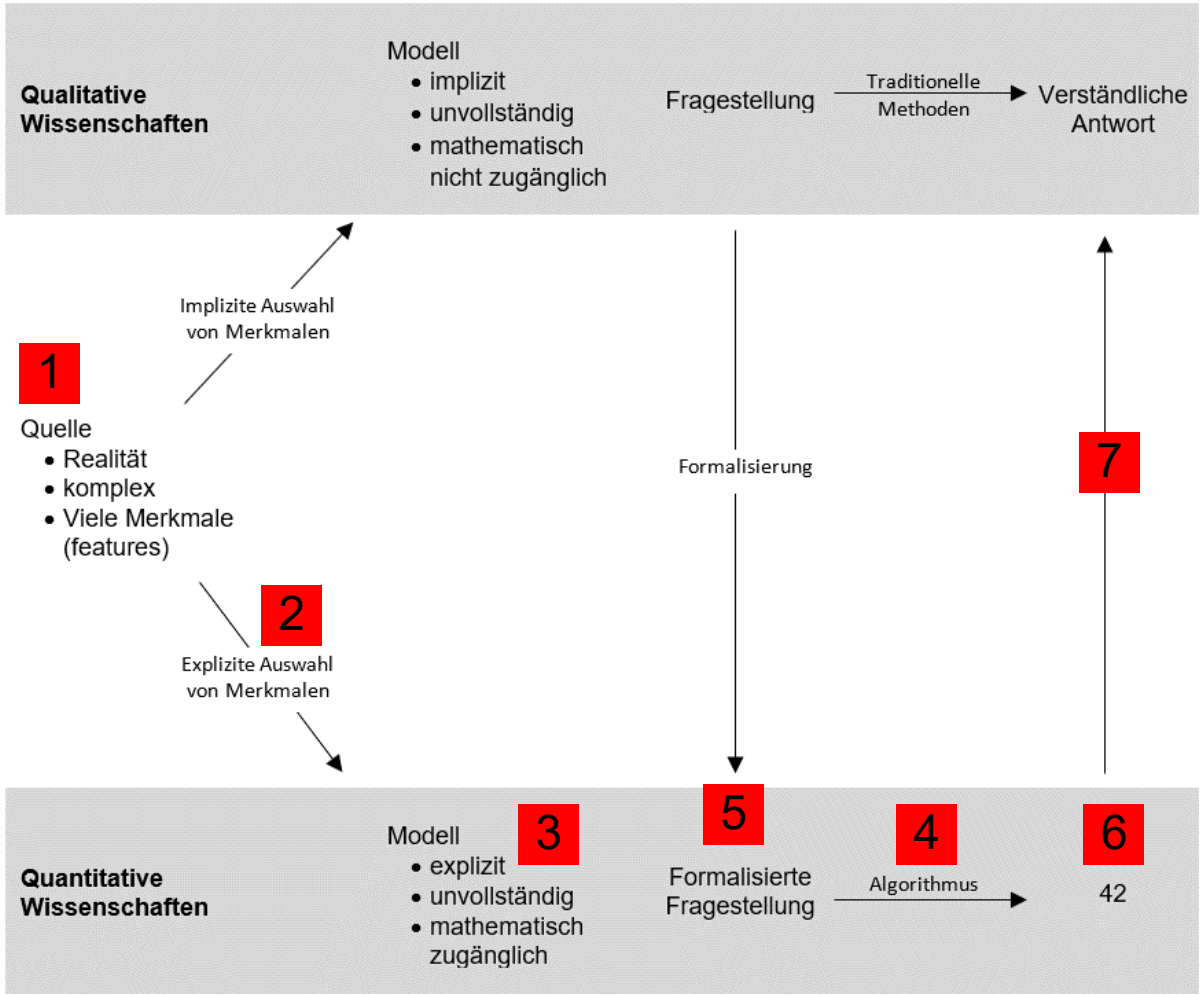
2. Textmerkmale

Tokenisierung muss nicht auf Basis von Wörtern geschehen. Möglich sind auch:

- n-Gramme von Wörtern:
  - n = 1 (bag of words): αποκριθεις, δε, ο, ιησους, ειπεν, προς, αυτον,…
  - n = 2 (Bigramme): αποκριθεις δε, δε ο, ο ιησους, ιησους ειπεν, ειπεν προς,…
  - n = 3 (Trigramme): αποκριθεις δε ο, δε ο ιησους, ο ιησους ειπεν, ιησους ειπεν προς,…

- n-Gramme von Buchstaben:
  - n = 1: α, π, ο, κ, ρ, ι, θ, ε, ι, σ, δ, ε, ο, ι, η, ς, …
  - n = 2 (Bigramme): απ, πο, οκ, κρ, ρι, ιθ, θε, ει, ισ, σδ, δε, εο, οι, ιη, ης, …
  - n = 3 (Trigramme): απο, ποκ, οκρ, κρι, ριθ, ιθε, θει, εισ, ισδ, σδε, δεο, εοι, …

Ablauf

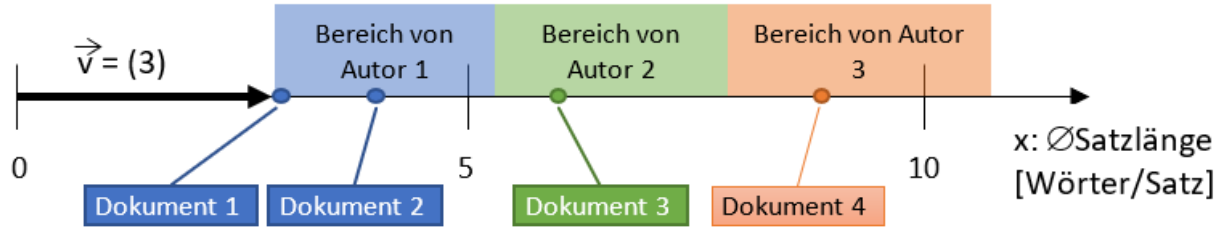


Anhand der Fallstudie zu den Paulusbriefen betrachten wir die Arbeitsschritte:

1. Der griechische Text des Neuen Testaments
   …und seine digitale Verfügbarkeit
2. Textmerkmale
3. **Textrepräsentation (Modell)**
4. Textklassifikation
5. Studiendesign (Formalisierte Fragestellung)
6. Resultate
7. Diskussion

3. Textrepräsentation (Modell)

		Variable(n)
		∅Satzlänge [Wörter/Satz]
Beobachtungen	Dokument 1	3
	Dokument 2	4
	Dokument 3	6
	Dokument 4	9



Univariate Analyse im Koordinatensystem
© Eigene Darstellung

Wie können diese Tokens in einem Modell dargestellt werden?

Darstellung in einem Koordinatensystem (Vektorraummodell)

Univariate Statistik:

Eine einzige Variable wird ausgewertet.

Im Vektorraummodell:

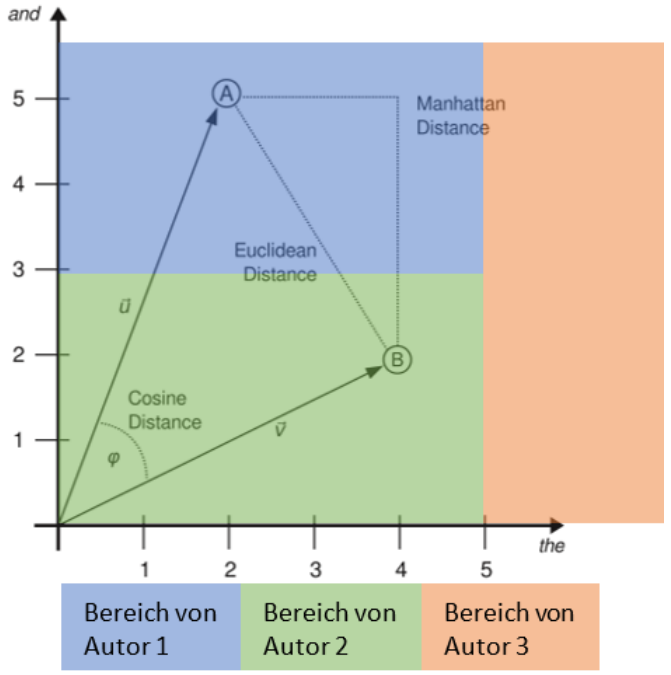
Eine einzige Dimension

3. Textrepräsentation (Modell)

Korpus
Text A: “the the and and and and and”
Text B: “the the the the and and”

Häufigkeitstabelle der Wort-1-Gramme	
and	7
the	6

Textrepräsentationen		
	x-Variable: «the»	y-Variable: «and»
Text A	2	5
Text B	4	2

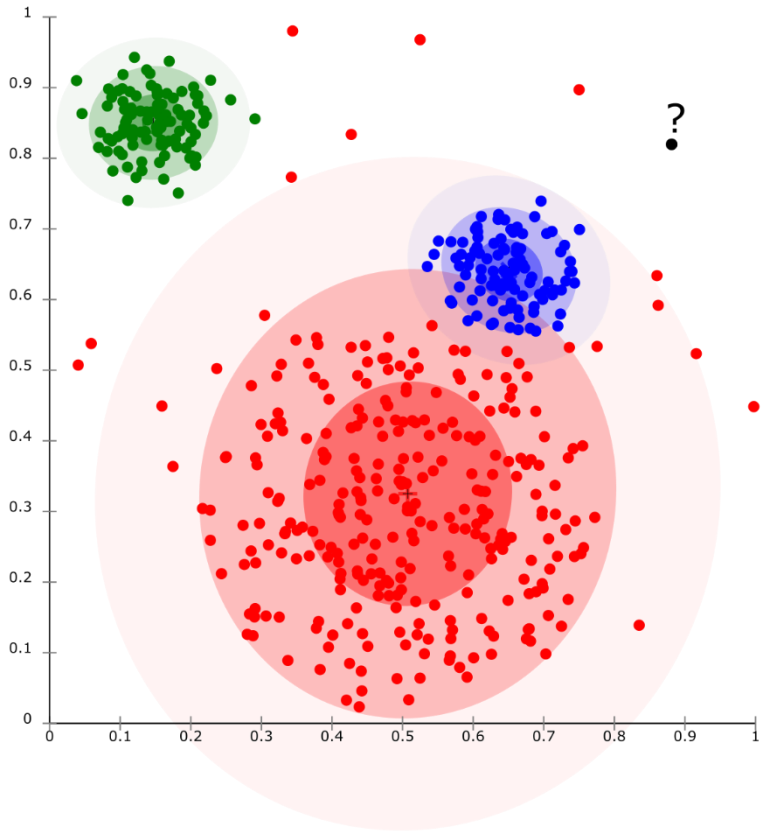


Multivariate Analyse im Koordinatensystem
© Eigene Darstellung, basierend auf Evert et al. 2017, Understanding and explaining Delta measures for authorship attribution, in: Digital Scholarship in the Humanities 32 (suppl_2), S. ii7.

- Multivariate Statistik:
  Mehrere Variablen werden ausgewertet.
- Das Vektorraummodell wird mehrdimensional.

- Häufigkeitstabelle: Wie oft kommt jeder Type vor im Korpus?
- Textrepräsentationen: Häufigkeit von jedem Type in diesem Text

3. Textrepräsentation (Modell)



Clustering im zweidimensionalen Raum

Was machen wir jetzt mit diesem Modell?

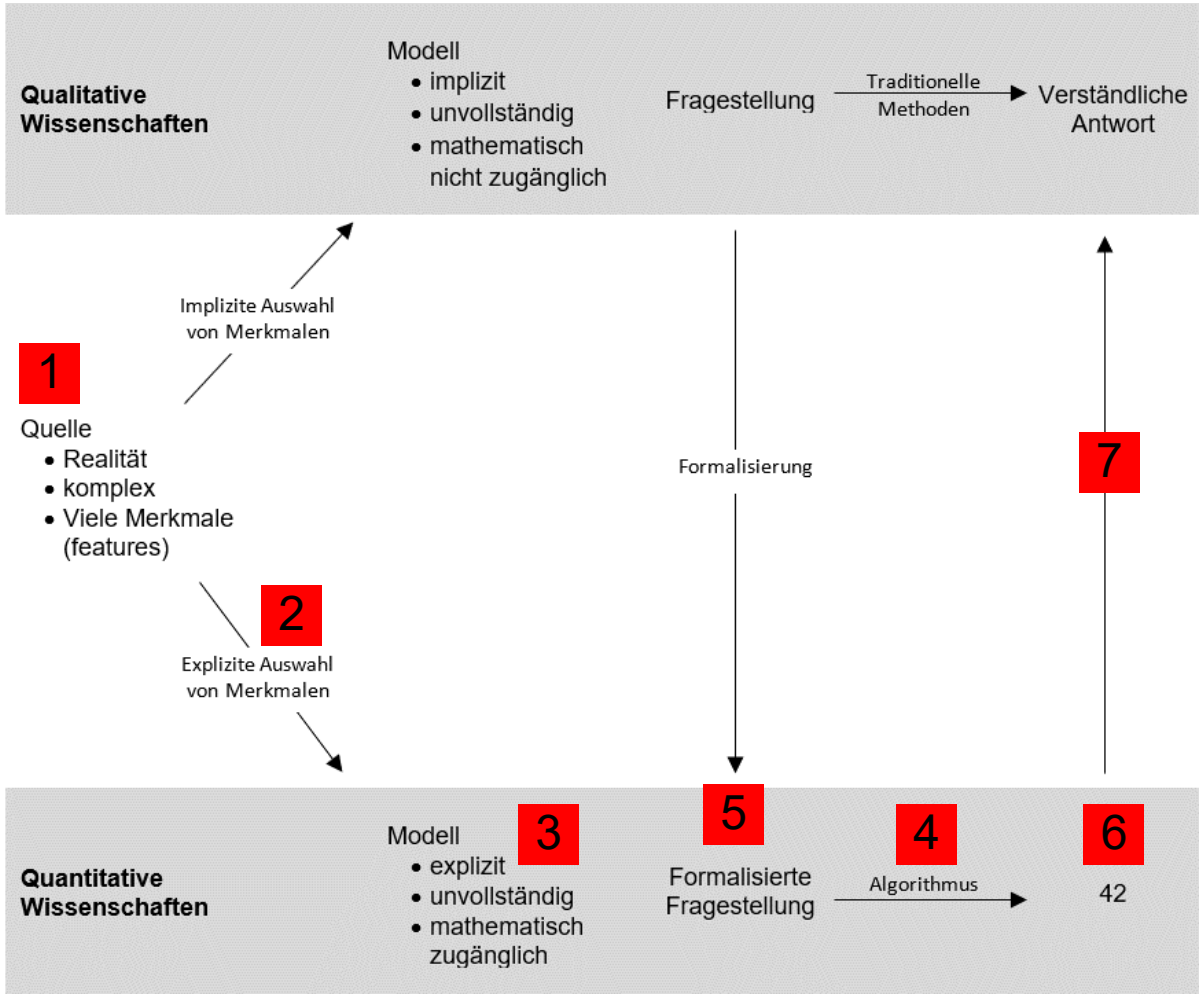
Komplexere Textklassifikations-Probleme haben viele verschiedene Punkte in einem hochdimensionalen Raum.

– Doch was ist mit überlappenden Bereichen?
– Und was ist bei hunderten von Dimensionen?

Allgemein formuliert:

– Wie können in einem geometrischen Raum Klassengrenzen etabliert werden?

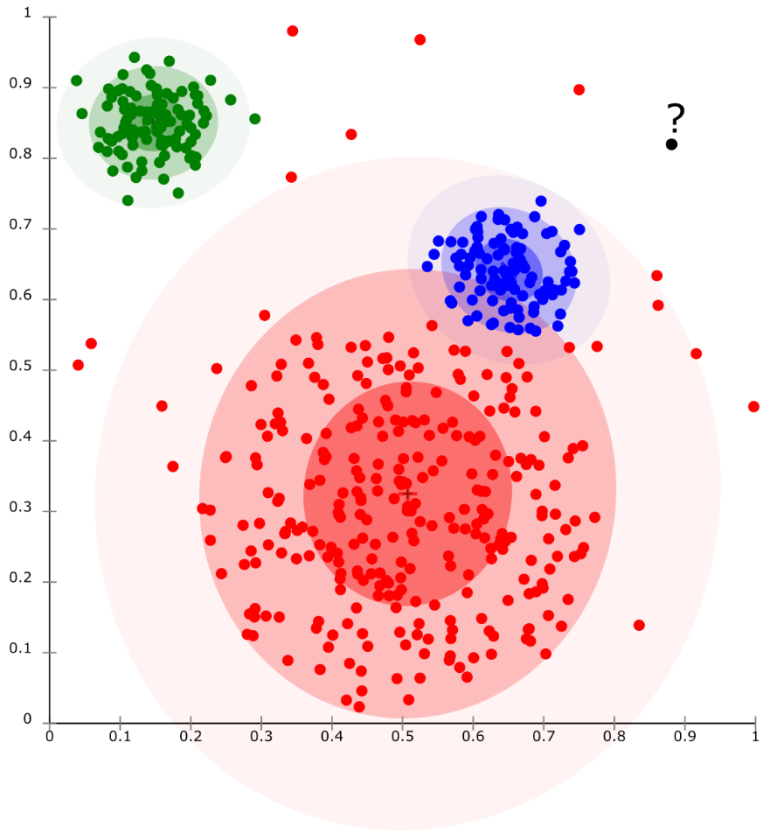
Ablauf



Anhand der Fallstudie zu den Paulusbriefen betrachten wir die Arbeitsschritte:

1. Der griechische Text des Neuen Testaments
   …und seine digitale Verfügbarkeit
2. Textmerkmale
3. Textrepräsentation (Modell)
4. **Textklassifikation**
5. Studiendesign (Formalisierte Fragestellung)
6. Resultate
7. Diskussion

4. Textklassifikation



Clustering im zweidimensionalen Raum

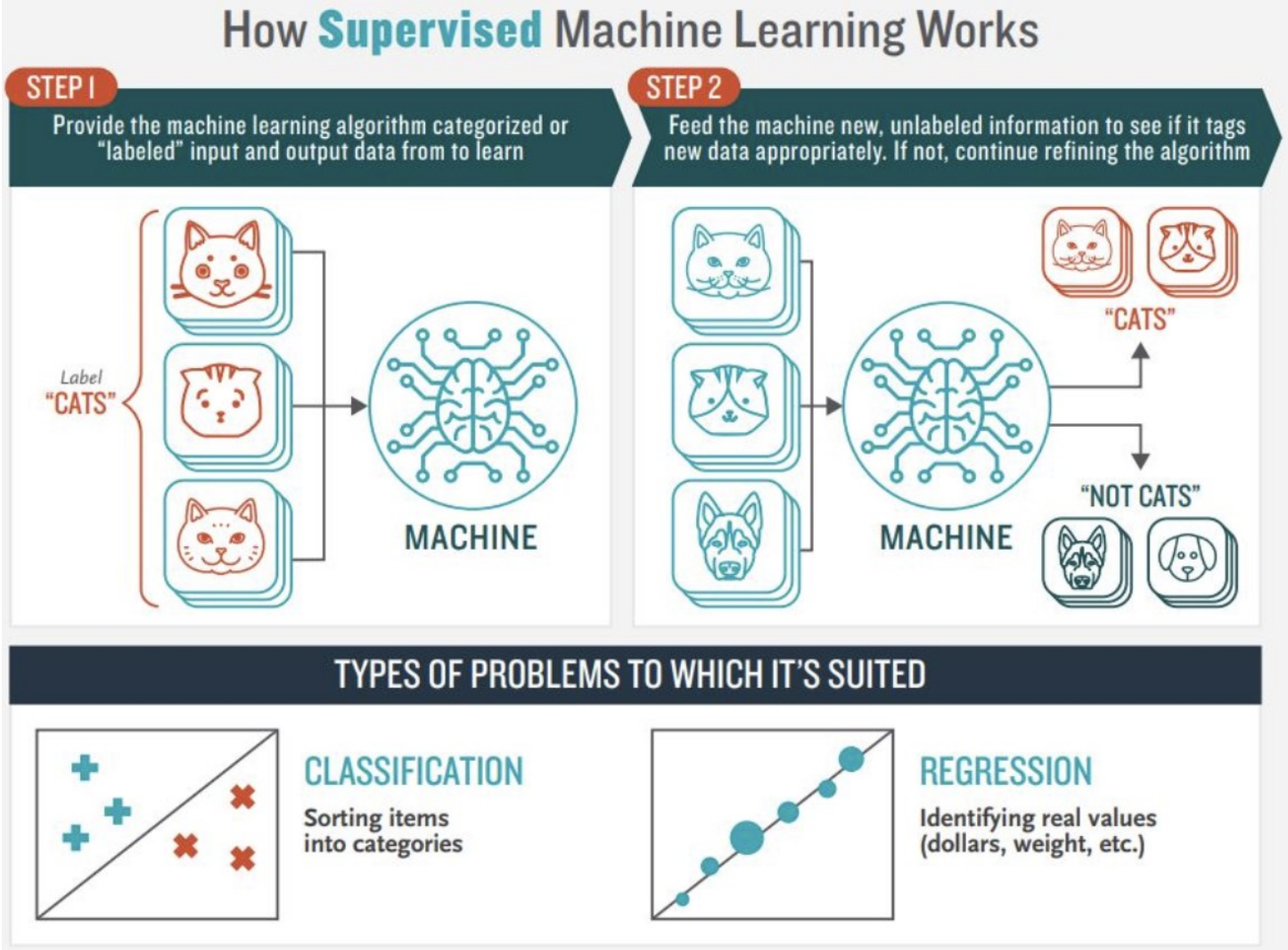
Klassifikation:

– Wir haben viele Punkte, deren Klasse bekannt ist.
– Farbe = Label für Klassenzugehörigkeit.
– Einige Punkte (Texte) sind von einem unbekannten Autor.
– Sie sind schwarz = ungelabelt.

Das ist ein Supervised Machine Learning Problem.

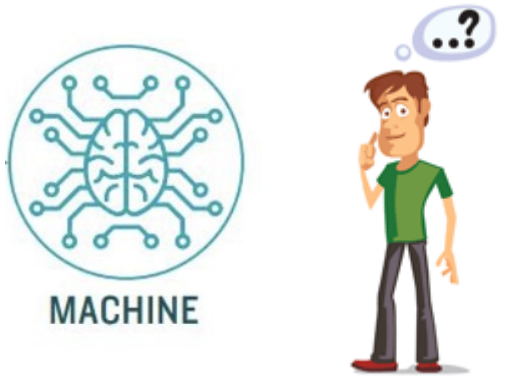
– Trainingsdaten: Bekannte Texte
– Testdaten: Unbekannte Texte

4. Textklassifikation



Supervised Learning
© https://medium.com/@jorgesleonel/supervised-learning-c16823b00c13.

4. Textklassifikation

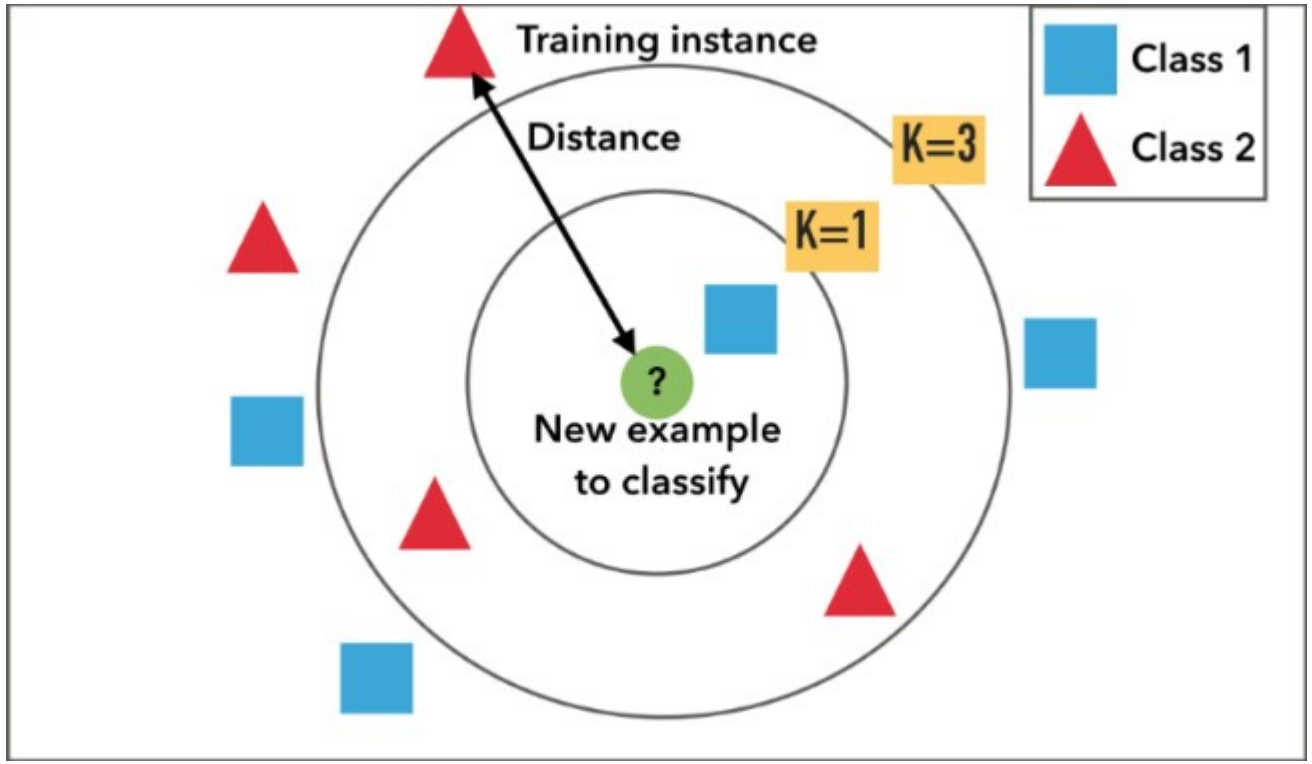


- Was geschieht im Innern der Maschine?
- Nicht Magie, sondern Mathematik!
- Wir als Geisteswissenschaftler wollen wegkommen von der Dichotomie zwischen *Euphorie – Dysphorie* :
  - Euphorie: Machine Learning = wundersame Blackbox / deus ex machina
  - Dysphorie: Machine Learning = Hype für Naivlinge
- Dafür müssen wir die Grundlagen verstehen!
- Klassifikator: Algorithmus = Folge von genauen Anweisungen, wie der Input (Text) in einen Output (Autor) transformiert werden soll.
- Der Algorithmus ist bloss im Grundgerüst (grundsätzliche Herangehensweise) vorhanden.
- Der Algorithmus muss durch die Trainingsdaten präzisiert werden.

4. Textklassifikation

- Es gibt viele verschiedene Algorithmen.
- Für Autorschaftsstudien oft gewählt:
  kNN = k nearest neighbors
- Zugehörigkeit wird bestimmt durch die nahe gelegenen Nachbarn.
- k ist frei wählbar, $k \in \mathbb{N}$
- Wenn k = 1 → ? = Klasse 1
- Wenn k = 3 → ? = Klasse 2

- Autorschaftsstudien verwenden oft kNN mit k=1 (aka. Delta-Methode)



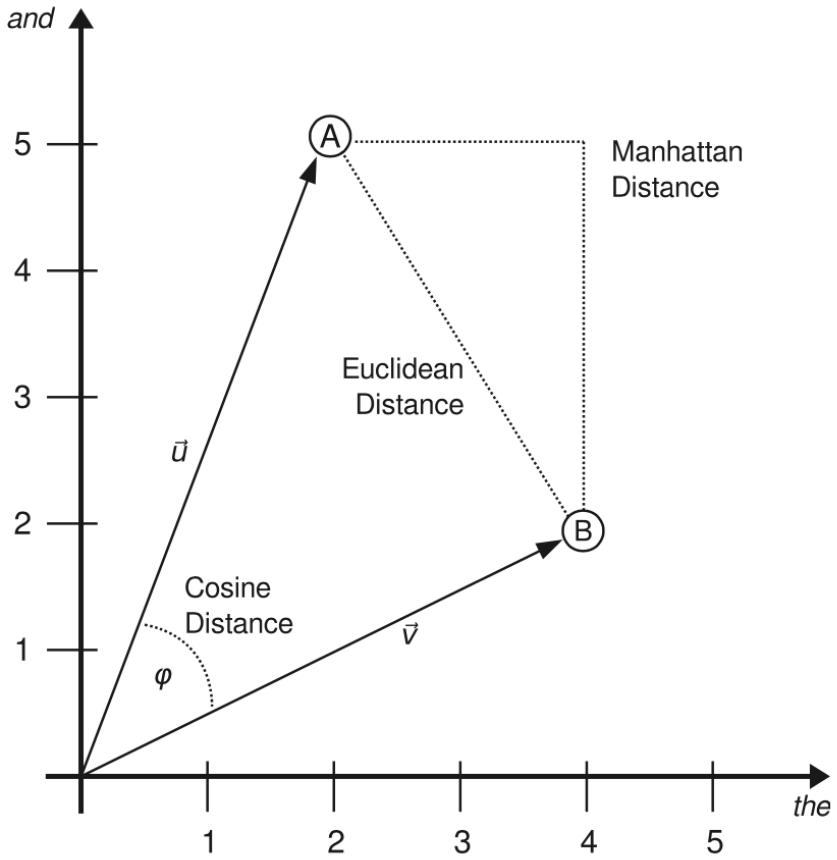
K Nearest Neighbors
© https://medium.com/@anuuz.soni/advantages-and-disadvantages-of-knn-ee06599b9336.

4. Textklassifikation

- Wenn die Distanzen der Texte zueinander bekannt sind, lässt sich eine Distanzmatrix erstellen.
- Daraus lässt sich eine probabilistische Reihenfolge der Kandidaten ablesen.
- Das neue Dokument hat denselben Autor wie …
  - Doc3 (am wahrscheinlichsten)
  - Doc2 (am 2. wahrscheinlichsten)
  - Doc1 (am wenigsten wahrscheinlich)

	Doc1	Doc2	Doc3	New Doc
Doc1		1.8	6.9	6.7
Doc2	1.8		5.1	6.1
Doc3	6.9	5.1		4.8
New Doc	6.7	6.1	4.8	

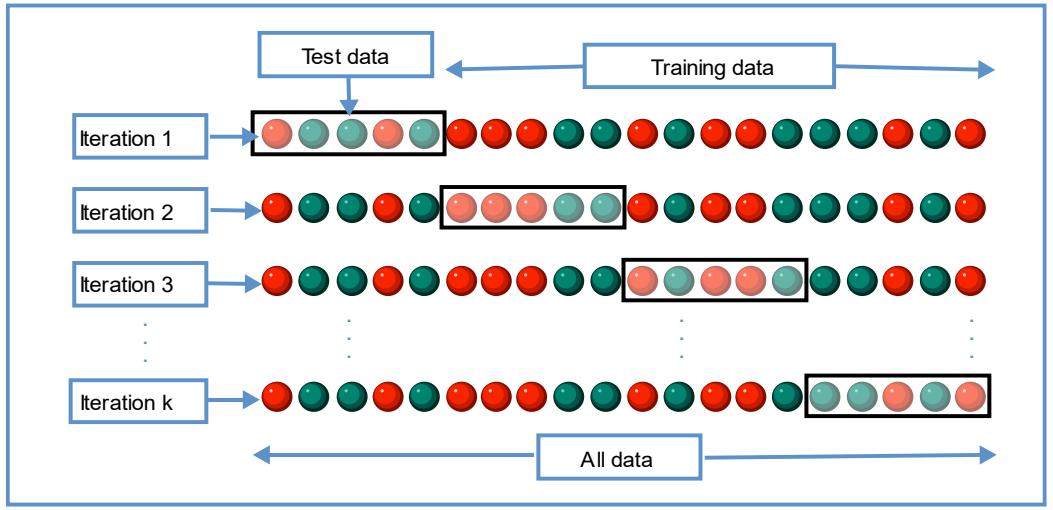
4. Textklassifikation



Verschiedene Distanzmasse
© Evert et al. 2017, Understanding and explaining Delta measures for authorship attribution, in: Digital Scholarship in the Humanities 32 (suppl_2), S. ii7.

- Wir müssen also bloss die Distanzen der Texte zueinander errechnen. Ganz einfach, oder…?
- Problem 1: Es gibt viele verschiedene Distanzmasse!
- Problem 2: Je nach mathematischem Verfahren sind die Punkte anders angeordnet im Raum!
- Woher kann man wissen, dass ein bestimmtes Modell gute Resultate liefert?

4. Textklassifikation



K-fold cross validation

Kreuzvalidierung:

– Nicht alle gelabelten Daten werden benutzt, um das Modell zu trainieren.

– Einige werden zum Testen übrig gelassen.

– Dies geschieht reihum viele Male.

– Dadurch bekommt man eine Einschätzung, ob das gewählte Modell eine gute Vorhersagekraft hat.

– Wenn ein Modell in der Kreuzvalidierung gut abschneidet, kann man den Vorhersagen des Modells vertrauen.

Zwischenstand

– Wir können Modelle erstellen, und sogar rausfinden, ob sie taugen oder nicht.

– Wir können alle Schriften des Neuen Testaments in einem Modell abbilden, und schauen, ob die umstrittenen Paulusbriefe nahe bei den echten sind.

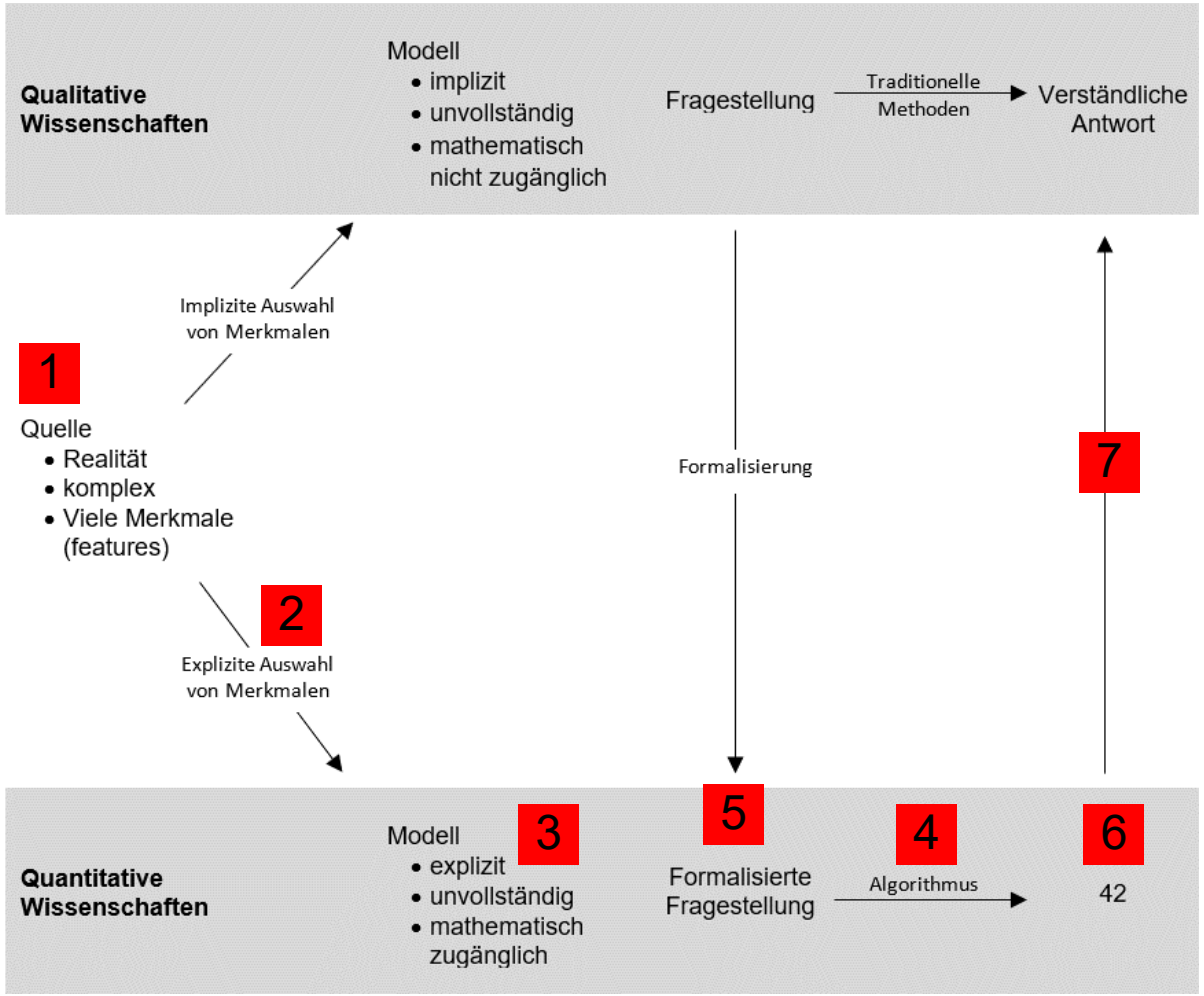
– Doch das geht noch wesentlich smarter.

Doch zuvor…

Verschnaufpause!

Dringende Fragen, um nicht abgehängt zu werden?

Ablauf



Anhand der Fallstudie zu den Paulusbriefen betrachten wir die Arbeitsschritte:

1. Der griechische Text des Neuen Testaments
   …und seine digitale Verfügbarkeit
2. Textmerkmale
3. Textrepräsentation (Modell)
4. Textklassifikation
5. **Studiendesign (Formalisierte Fragestellung)**
6. Resultate
7. Diskussion

5. Studiendesign (Formalisierte Fragestellung)

– offene Kandidatengruppe → Authorship Verification

– Paulus könnte der Autor sein sein, aber auch irgendeine unbekannte Drittperson.

– General Imposters Framework (GI):

"[t]he general intuition behind the GI, is not to assess whether two documents are simply similar in writing style, given a static feature vocabulary, but rather, it aims to assess

  – **whether two documents are significantly more similar to one another than other documents**,

  – across a variety of stochastically impaired feature spaces,

  – and compared to random selections of so-called distractor authors, also called 'imposters'."

  (Kestemont et al. 2016, S. 88)

5. Studiendesign (Formalisierte Fragestellung)

Drei Textgruppen:

- Sichere Paulusbriefe (P_S)
- Impostors (I) (übrige Schriften des NT)
- Umstrittene Paulusbriefe (P_{US})

Diese Berechnung wird mehrmals durchgeführt,

- mit je zufällig ausgewählten Textmerkmalen
- mit je zufällig ausgewählten Vergleichstexten aus den Impostors.

Wieso mehrmals?

- Weil die zur Verfügung stehenden Textmerkmale und Vergleichstexte wahrscheinlich nicht repräsentativ für die Grundgesamtheit sind.

Fehlerrate verkleinert durch Bootstrapping, d.h. Durchschnitt aus allen Berechnungen.

5. Studiendesign (Formalisierte Fragestellung)

Die Implementation des General Imposters Framework im R-Package stylo:

4 Hyperparameter:

- Distanzmass (z.B. Cosinus-Distanz)
- Anzahl Iterationen (z.B. 100 Iterationen)
- prozentualer Anteil der Textmerkmale (features) (z.B. 50% der Textmerkmale)
- prozentualer Anteil der Imposter-Texte (z.B. 50% der Imposter-Texte)

Dann werden anhand der ausgewählten 50% der Textmerkmale sämtliche Cosinus-Distanzen des zu untersuchenden Textes zu allen Texten des vermuteten Autors und zu den ausgewählten 50% der Imposter-Texte berechnet.

5. Studiendesign (Formalisierte Fragestellung)

Falls der nächste Nachbar des zu untersuchenden Textes ein Imposter ist, lautet das Resultat dieser Iteration «anderer Autor».

Ist hingegen der nächste Nachbar ein Text des vermuteten Autors, so lautet das Resultat dieser Iteration «vermuteter Autor».

Angenommen, dass 90 der 100 Iterationen «vermuteter Autor» liefern, dann liegt die Wahrscheinlichkeit, dass der vermutete Autor den zu untersuchenden Text geschrieben hat, bei 90%.

5. Studiendesign (Formalisierte Fragestellung)

- Die Methode benötigt vier Hyperparameter. Doch wie kann man herausfinden, welche Werte diese Hyperparameter haben sollen?
- Um dies herauszufinden, stellt das stylo-Package die Funktion imposters.optimize() zur Verfügung.
- Wenn man diese Funktion mit einer bestimmten Textrepräsentation der Trainingstexte und einer bestimmten Kombination aus Hyperparametern aufruft, errechnet sie die beiden Werte p1 und p2.
- Diese Werte liegen zwischen 0 und 1 und begrenzen dasjenige Intervall, in dem die Resultate unzuverlässig sein werden, wenn man sie mit genau derselben Kombination aus Textrepräsentation und Hyperparametern aufruft.
- Wenn beispielsweise die Wahrscheinlichkeit, dass Paulus den 1. Timotheusbrief geschrieben hat, p = 0.9 beträgt, aber das Unsicherheitsintervall [p1 = 0.3, p2 = 0.95] ist, so ist das vermeintlich klare Resultat wertlos.

5. Studiendesign (Formalisierte Fragestellung)

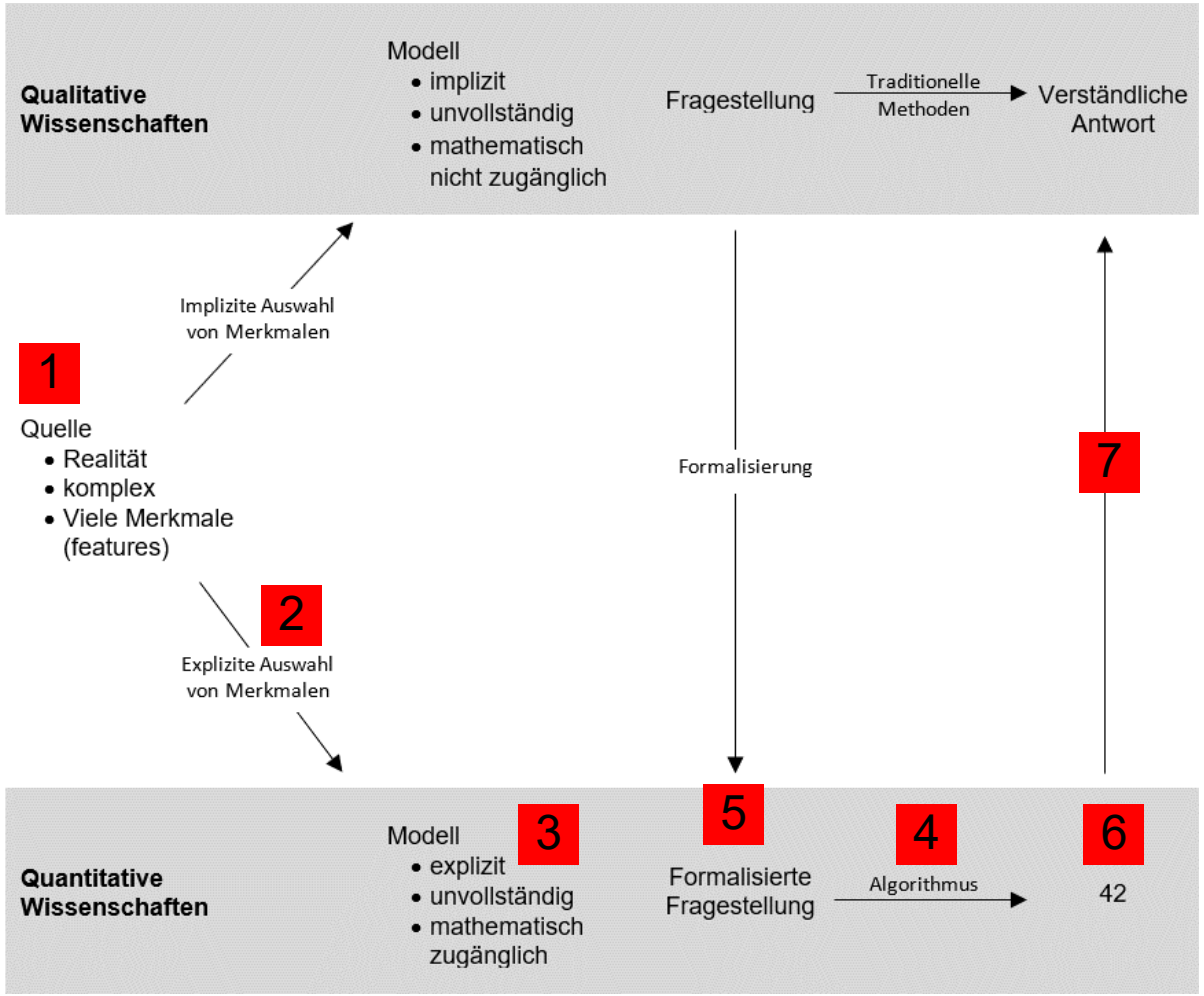
Aus diesem Grund wurden für die vorliegende Studie in einem aufwändigen Suchverfahren diejenigen Textrepräsentationen und Hyperparameter bestimmt, welche erfolgversprechend sind, weil sie ein kleines Unsicherheitsintervall haben.

Nach einem aufwändigen Optimierungsverfahren wurden nur diejenigen Textrepräsentationen und Hyperparameter berücksichtigt, welche ein sehr kleines Unsicherheitsintervall haben.

Entsprechend resultierte für jede Testtext-Autorkandidat-Kombination 46 Einschätzungen, mit welcher Wahrscheinlichkeit dieser Autorkandidat der tatsächliche Autor des Testtextes ist.

Von diesen Einschätzungen wurden nur diejenigen behalten, welche ausserhalb ihres jeweiligen Unsicherheitsintervalls liegen.

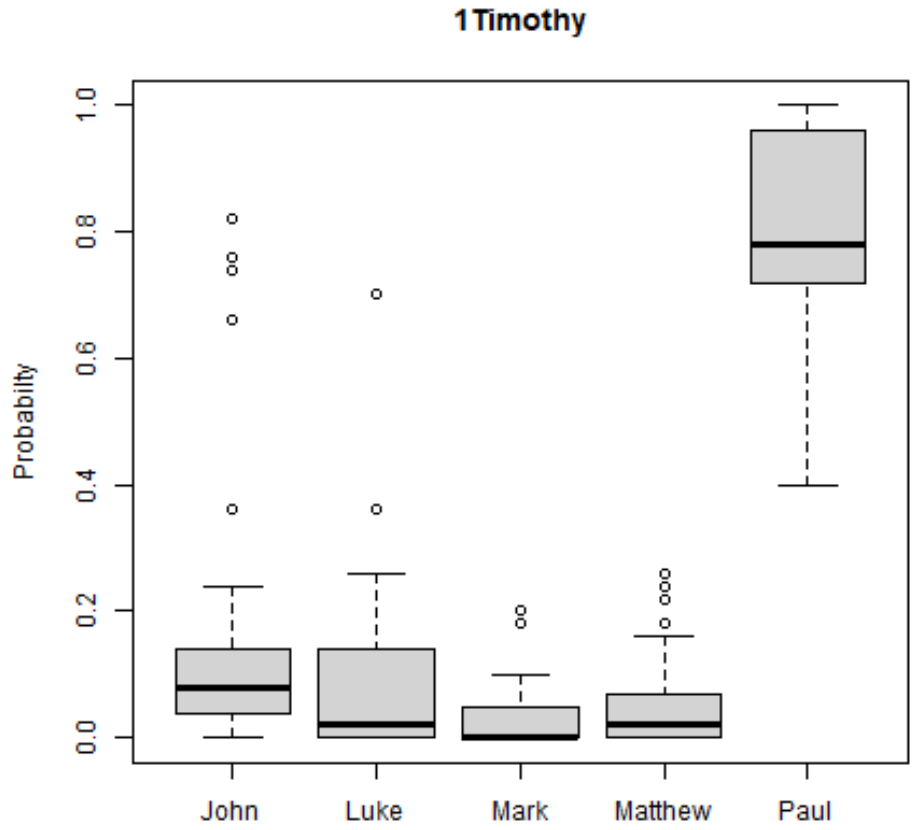
Ablauf



Anhand der Fallstudie zu den Paulusbriefen betrachten wir die Arbeitsschritte:

1. Der griechische Text des Neuen Testaments
   …und seine digitale Verfügbarkeit
2. Textmerkmale
3. Textrepräsentation (Modell)
4. Textklassifikation
5. Studiendesign (Formalisierte Fragestellung)
6. **Resultate**
7. Diskussion

6. Resultate

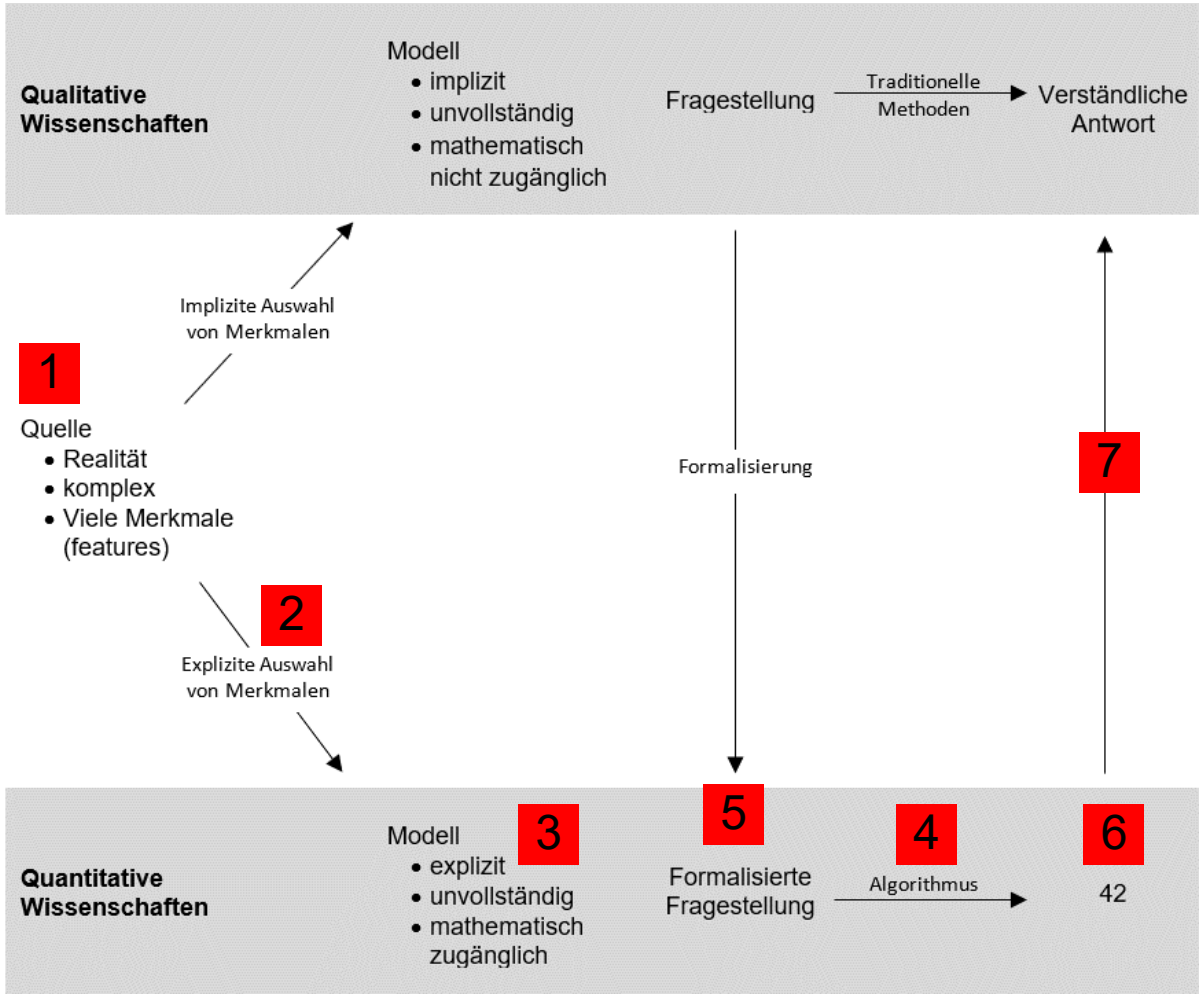


1. Timotheus-Brief, Überblick über die Resultate aller 46 Iterationen, die je eine Wahrscheinlichkeit pro Autorkandidat lieferten.
© Eigene Darstellung

Lesehilfe:

- Die 46 Einschätzungen, mit welcher Wahrscheinlichkeit Johannes den 1. Timotheusbrief geschrieben hat, sind verteilt zwischen 0 und ca. 0.8.
- Der Median aller Einschätzungen liegt bei ca. 0.1, wobei die Hälfte aller Einschätzungen zwischen 0.05–0.15 liegen.
- Die kleinen Kreise sind Datenpunkte, die besonders weit abweichen von den übrigen.

Ablauf



Anhand der Fallstudie zu den Paulusbriefen betrachten wir die Arbeitsschritte:

1. Der griechische Text des Neuen Testaments
   …und seine digitale Verfügbarkeit
2. Textmerkmale
3. Textrepräsentation (Modell)
4. Textklassifikation
5. Studiendesign (Formalisierte Fragestellung)
6. Resultate
7. **Diskussion**

7. Diskussion

Die Resultate besagen mit überraschender Klarheit, dass die umstrittenen Paulusbriefe mit einer sehr viel höheren Wahrscheinlichkeit von Paulus verfasst wurden als von einem anderen Autor des Neuen Testaments.

Einwände:

- Niemand behauptet, dass einer der vier Evangelisten die umstrittenen Paulusbriefe geschrieben hat.
  - Doch das ist gerade der springende Punkt des General Imposters Framework: Die Hochstapler sollen Texte sein, die dem zu untersuchenden Text zwar ähnlich sind, aber von anderen Autoren stammen.

7. Diskussion

2. Einwand:

- Nicht die Autorschaft, sondern die Textgattung und das Thema haben zu diesem Resultat geführt.
  - Es ist Konsens in der Authorship Attribution-Community, dass die präsentierten Methoden funktionieren.
  - Dies wurde in vielen Studien ausprobiert und nachgewiesen.

3. Einwand:

- Sind die Werte für Paulus genug hoch, um signifikant zu sein? Es ist nämlich denkbar, dass die betroffenen Dokumente aus P_{US} von einem Autor X stammen, dessen Stil näher an Paulus' Stil liegt als an den Stilen der anderen neutestamentlichen Autoren.

Fazit

Authorship Attribution kann nicht eine abschliessende Wahrheit finden, sondern nur neue Argumente in einer komplexen Diskussion.

Ein mangelndes Verständnis der Vorgänge führt dazu, dass digitale Methoden als eine wundersame Blackbox gelten, deren Resultate nicht genügend reflektiert werden. Es ist dringend notwendig, genau erklären zu können, was die eigenen Resultate bedeuten, was sie können, und was sie nicht können.

Universität Basel

Vielen Dank
für Ihre Aufmerksamkeit.