Big Data - und nun?

Was kann die Bioinformatik?

Jochen Kruppa

Institut für Biometrie und Klinische Epidemiologie

jochen.kruppa@charite.de

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

Vorstellung

Wer spricht heute zu Ihnen?

Lübeck

Göttingen

- Studium der Pflanzenbiotechnologie
- PhD Universität zu Lübeck. Institut für medizinische Biometrie und Statistik (2009 - 2013)
- Postdoc Universität Göttingen. Department of Animal Breeding (2013 - 2014)
- Postdoc Universitätsmedizin Göttingen. Department of Medical Statistics (2014 - 2015)
- Postdoc TiHo Hannover. AG Genomics and Bioinformatics of Infectious Diseases (2016 - 2018)
- AG-Leiter Statistische Bioinformatik. Berliner Charité (2018 - heute)

"Das ist die Logik der Forschung, die nie verifizieren, sondern immer nur jene Erklärungen beibehalten kann, die beim derzeitigen Erkenntnisstand am wenigsten falsifiziert sind."

– Bildungsökonom Ludger Wößmann

http://www.faz.net/aktuell/wirtschaft/bildungsoekonom-woessmann-hat-preis-gewonnen-15183142.html

"Das ist die Logik der Forschung, die nie verifizieren, sondern immer nur jene Erklärungen beibehalten kann, die beim derzeitigen Erkenntnisstand am wenigsten falsifiziert sind."

– Bildungsökonom Ludger Wößmann

http://www.faz.net/aktuell/wirtschaft/bildungsoekonom-woessmann-hat-preis-gewonnen-15183142.html

"Das ist die Logik der Forschung, die nie verifizieren, sondern immer nur jene Modelle beibehalten kann, die beim derzeitigen Erkenntnisstand am wenigsten falsifiziert sind."

– Bildungsökonom Ludger Wößmann

http://www.faz.net/aktuell/wirtschaft/bildungsoekonom-woessmann-hat-preis-gewonnen-15183142.html

Small data

Small Data vs. Big Data

$$\begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \\ \vdots \\ \\ \\ \\ \\ n_s \end{array} \begin{array}{c} \begin{array}{ccc} p_1 & p_2 & p_3 \end{array} \\ \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \\ \vdots & \vdots & \vdots \\ \\ \\ \\ \\ n_{s1} & n_{s2} & n_{s3} \end{pmatrix} \end{array}$$

$$n > p$$

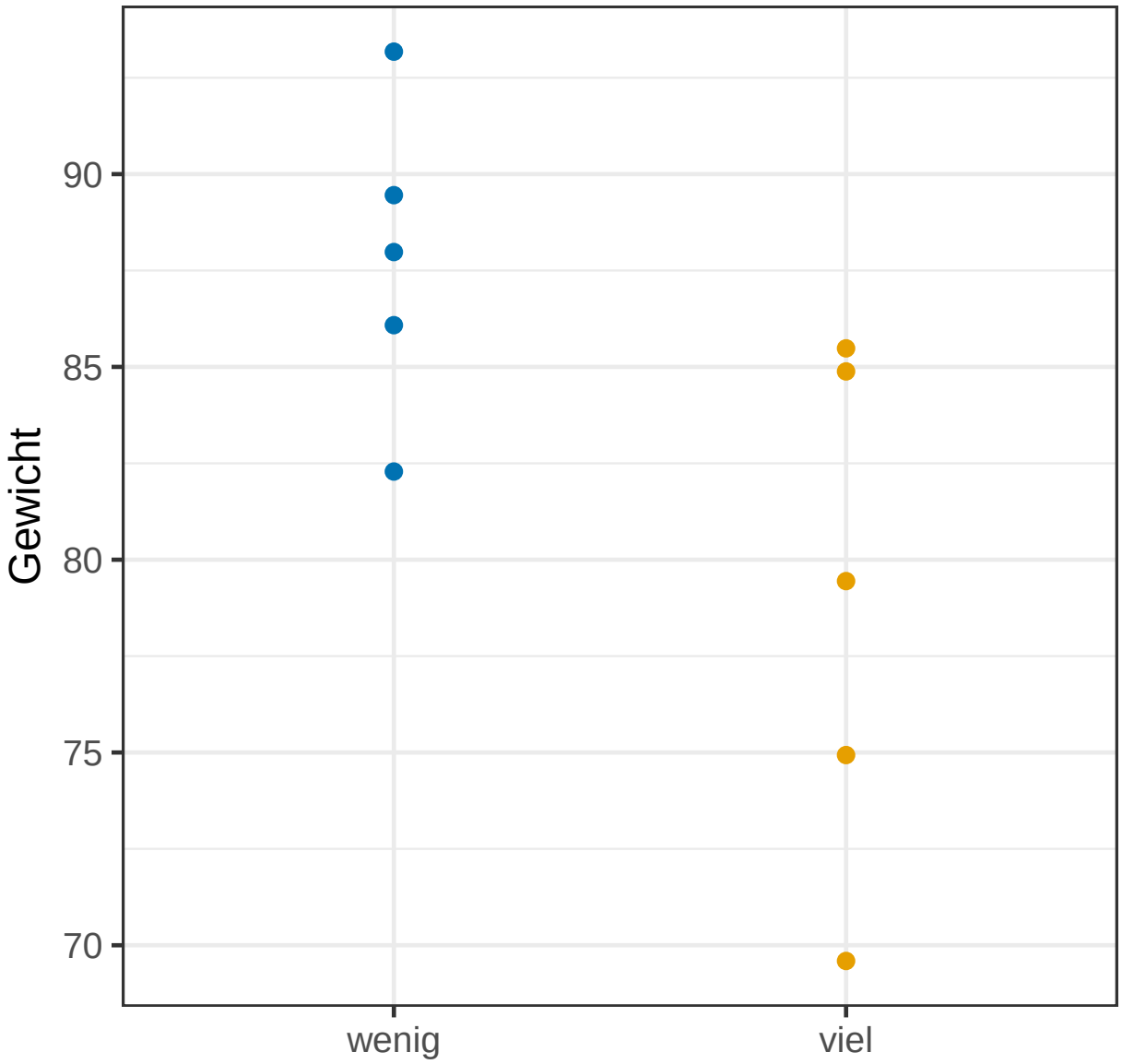
$$\begin{array}{c} \\ n_1 \\ n_2 \\ n_3 \\ \vdots \\ \\ \\ \\ \\ n_b \end{array} \begin{array}{c} \begin{array}{ccccc} p_1 & p_2 & p_3 & \cdots & p_b \end{array} \\ \begin{pmatrix} a_1 & a_2 & a_3 & \cdots & a_4 \\ b_1 & b_2 & b_3 & \cdots & b_4 \\ c_1 & c_2 & c_3 & \cdots & c_4 \\ \vdots & \vdots & \vdots & & \vdots \\ \\ \\ \\ \\ n_{b1} & n_{b2} & n_{b3} & \cdots & n_{bp} \end{pmatrix} \end{array}$$

$$n \ll p \quad \text{oder} \quad n \gg p$$

Small Data – Ein Model von Gewicht und Sport

Gewicht	Sport
80.50	wenig
83.03	wenig
92.99	wenig
94.13	wenig
84.90	wenig
76.68	viel
70.49	viel
67.02	viel
72.14	viel
73.38	viel

Small Data – Ein Model von Gewicht und Sport



Small Data – Ein Model von Gewicht und Sport

Ein simpler t-Test

$$T = \frac{Sport_{wenig} - Sport_{viel}}{SE_{Sport}}$$

Small Data – Ein Model von Gewicht und Sport

Ein simpler t-Test

$$T = \frac{Sport_{wenig} - Sport_{viel}}{SE_{Sport}}$$

$$T = \frac{\qquad\qquad\qquad\qquad}{\qquad\qquad\qquad\qquad}$$

Small Data – Ein Model von Gewicht und Sport

Ein simpler t-Test

$$T = \frac{Sport_{wenig} - Sport_{viel}}{SE_{Sport}}$$

$$T = \frac{87.80}{}$$

Small Data – Ein Model von Gewicht und Sport

Ein simpler t-Test

$$T = \frac{Sport_{wenig} - Sport_{viel}}{SE_{Sport}}$$

$$T = \frac{87.80 - 78.87}{}$$

Small Data – Ein Model von Gewicht und Sport

Ein simpler t-Test

$$T = \frac{Sport_{wenig} - Sport_{viel}}{SE_{Sport}}$$

$$T = \frac{87.80 - 78.87}{3.57}$$

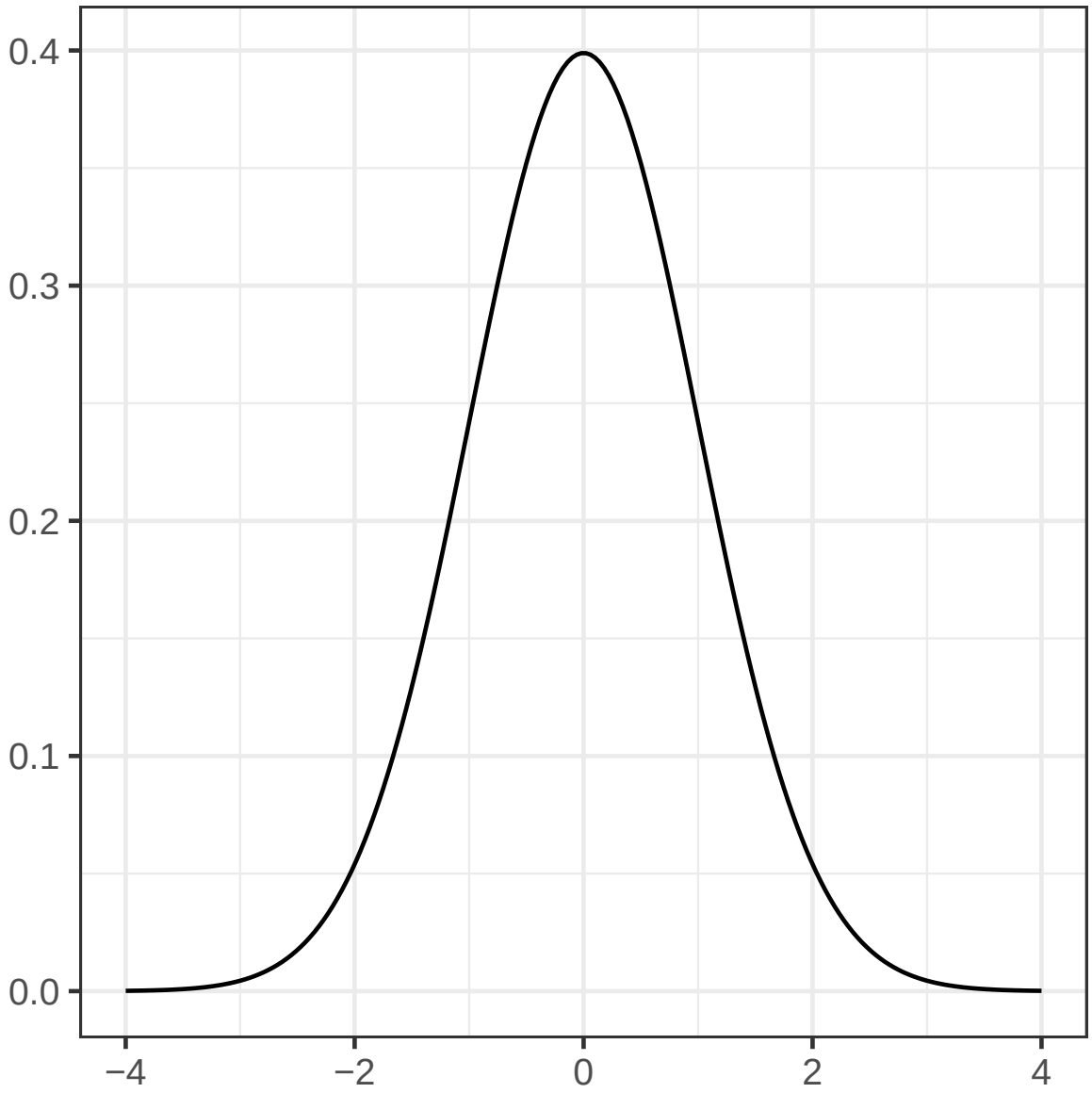
Small Data – Ein Model von Gewicht und Sport

Ein simpler t-Test

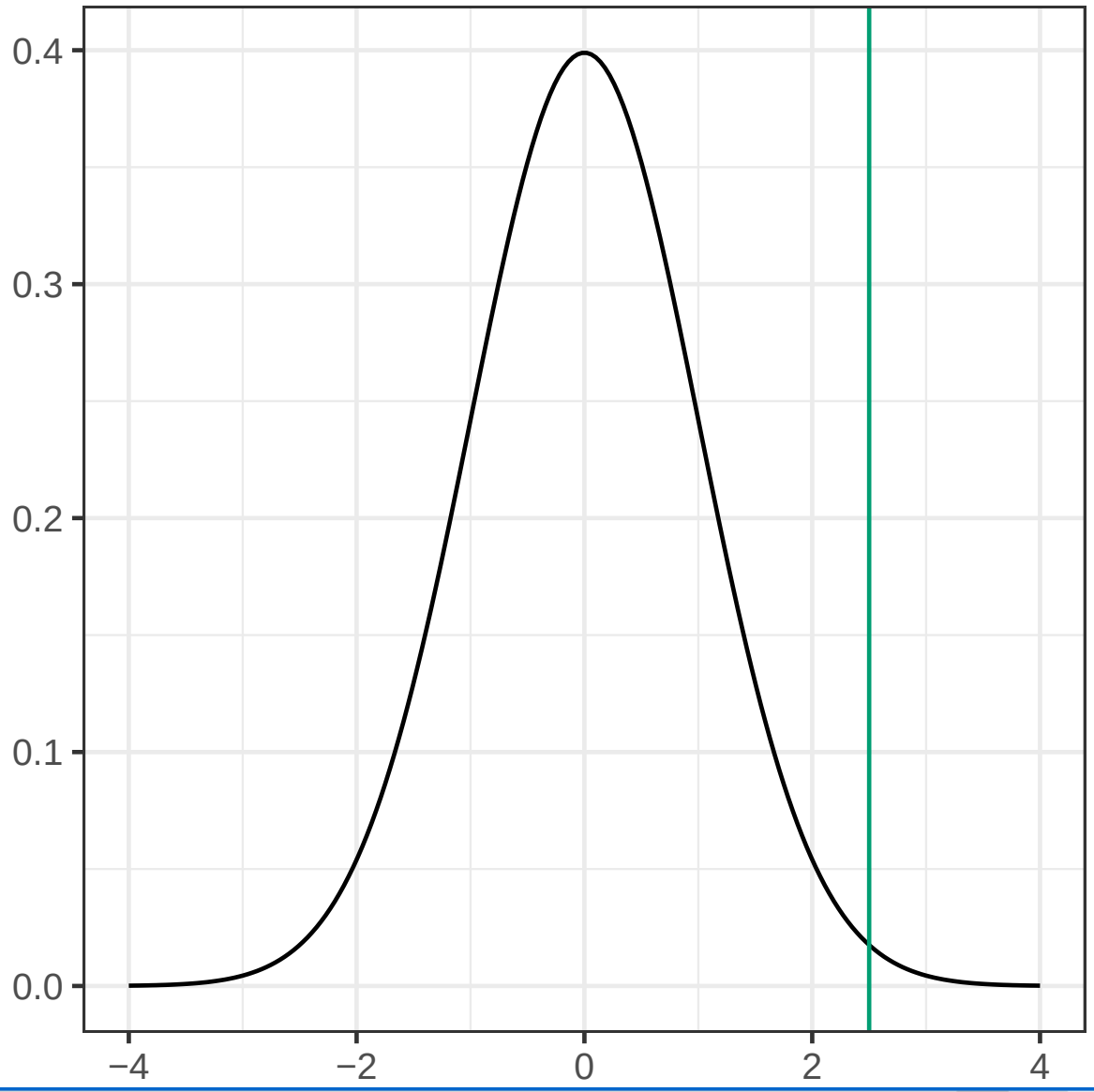
$$T = \frac{Sport_{wenig} - Sport_{viel}}{SE_{Sport}}$$

$$T = \frac{87.80 - 78.87}{3.57} = 2.5$$

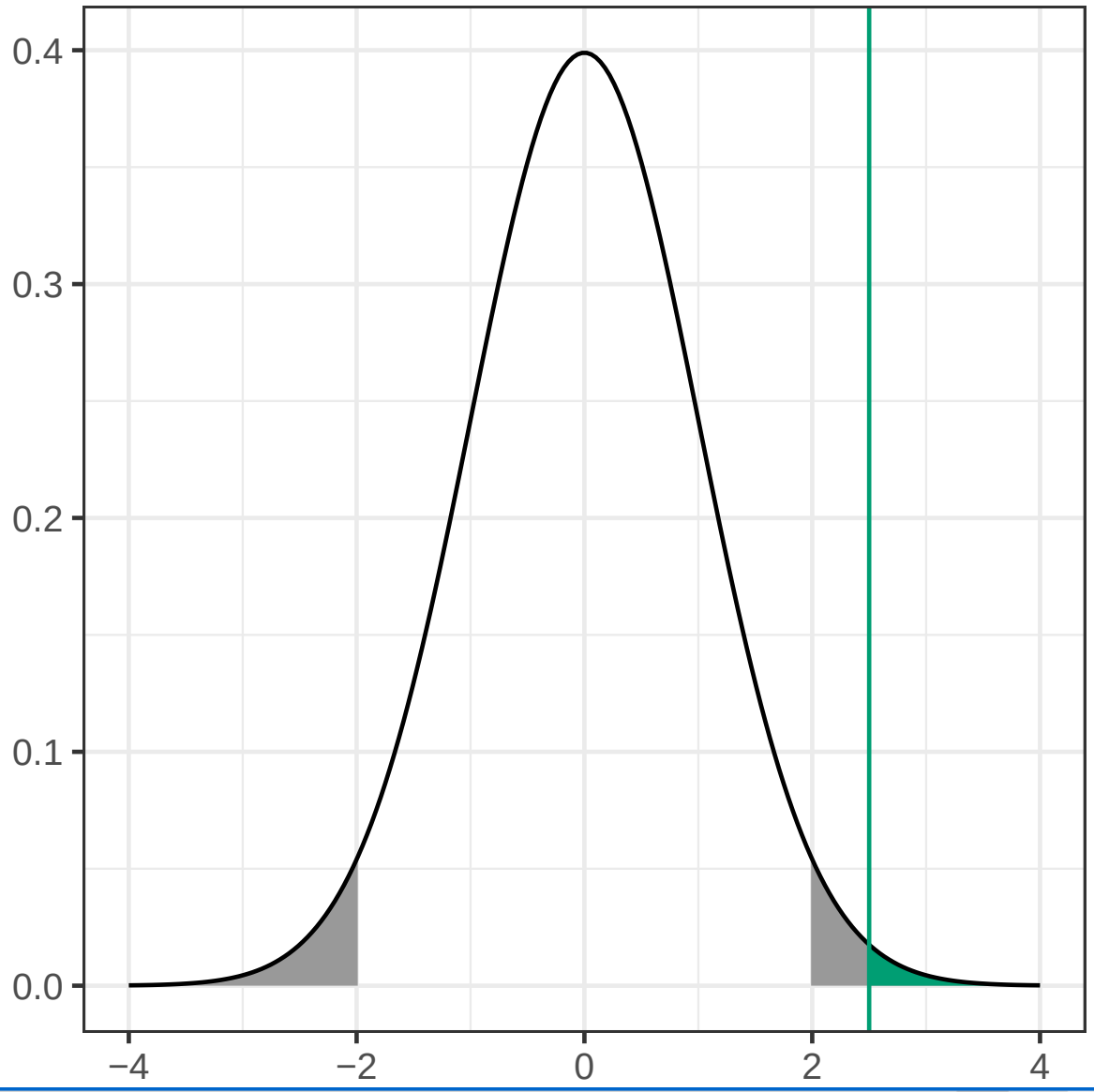
Small Data – Ein Model von Gewicht und Sport



Small Data – Ein Model von Gewicht und Sport



Small Data – Ein Model von Gewicht und Sport



Small Data – Der Stern mit dem p-Wert

- p-Werte werden meist falsch interpretiert und daher falsch genutzt
- **The ASA's Statement on p-Values**

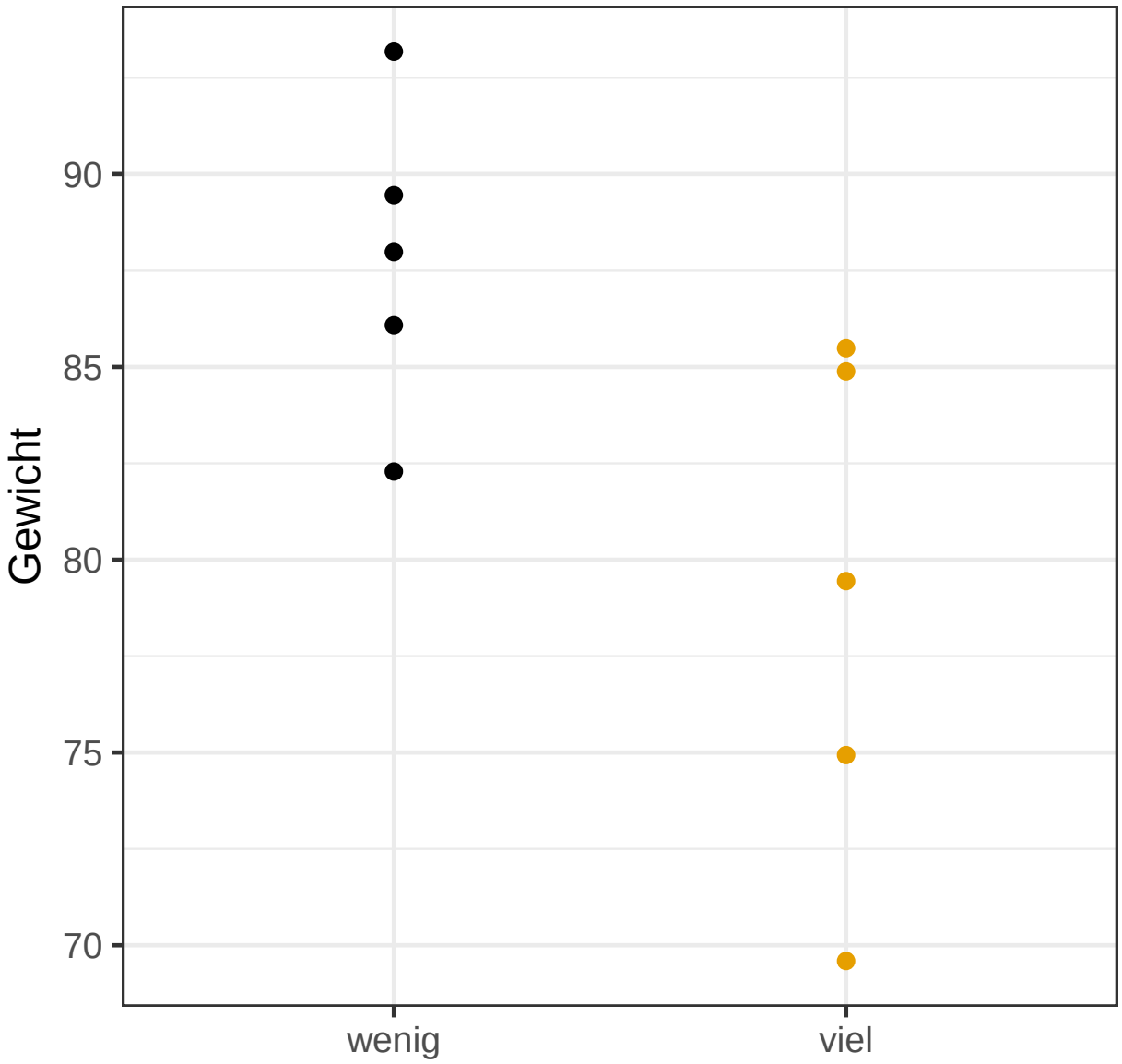
  Q: Why do so many colleges and grad schools teach p = 0.05?
  A: Because that's still what the scientific community and journal editors use.
  Q: Why do so many people still use p = 0.05?
  A: Because that's what they were taught in college or grad school.

- p-Werte waren nie so von Fisher gedacht, wie p-Werte genutzt werden
- p-Werte sind bedingte Wahrscheinlichkeiten
- p(Meine Daten | Null Hypothese)

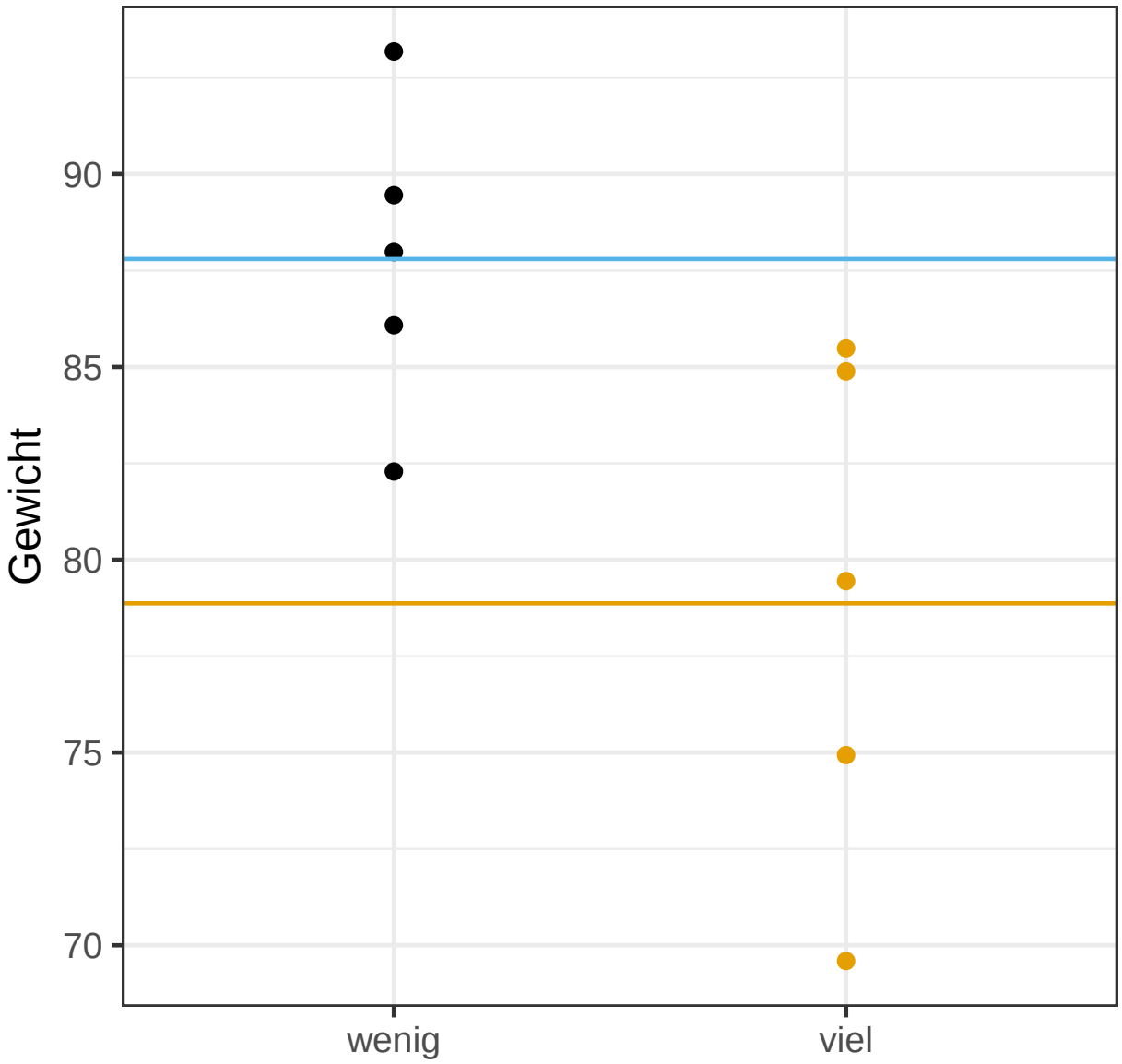
Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, https://doi.org/10.1080/00031305.2016.1154108

Small data und confounder

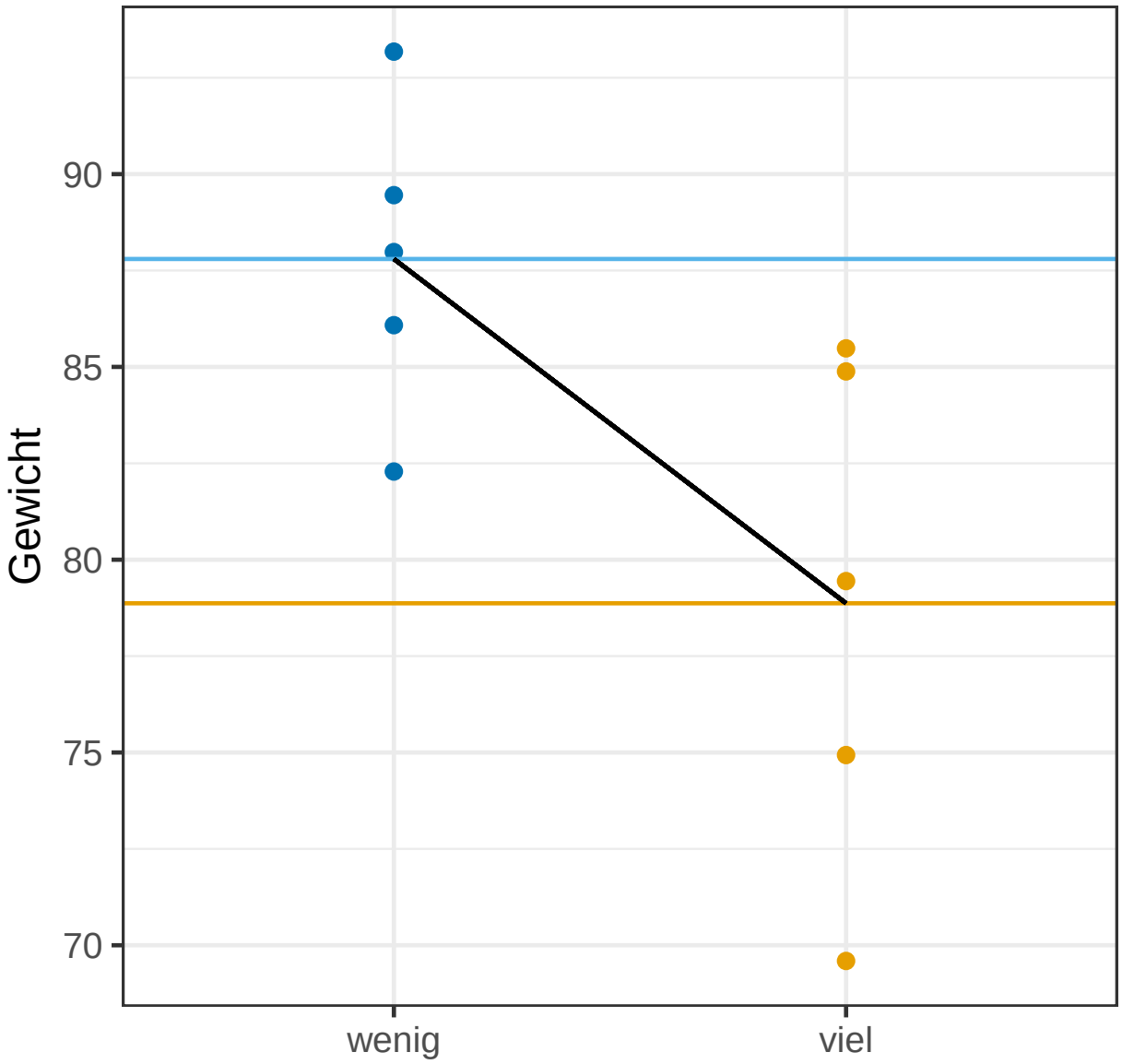
Small Data – Ein Model von Gewicht und Sport



Small Data – Ein Model von Gewicht und Sport

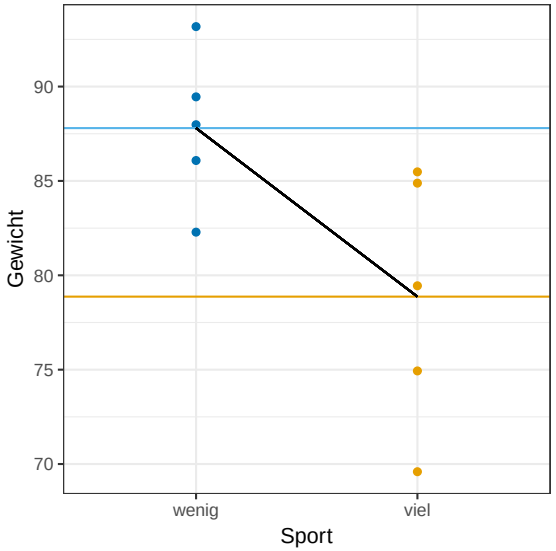


Small Data – Ein Model von Gewicht und Sport



Small Data – Ein Model von Gewicht und Sport

Lineare Regression



$$f(Gewicht) = \beta_0 + \beta_1 \cdot Sport$$

- β_0 gleich dem Y-Achsenabschnitt
- β_1 gleich der Steigung und dem Mittelwertsunterschied

Small Data – Ein Model von Gewicht und Sport

$$T = \frac{Sport_{wenig} - Sport_{viel}}{SE_{Sport}} \qquad f(Gewicht) = \beta_0 + \beta_1 \cdot Sport$$

Wir können (Gruppen)variablen testen

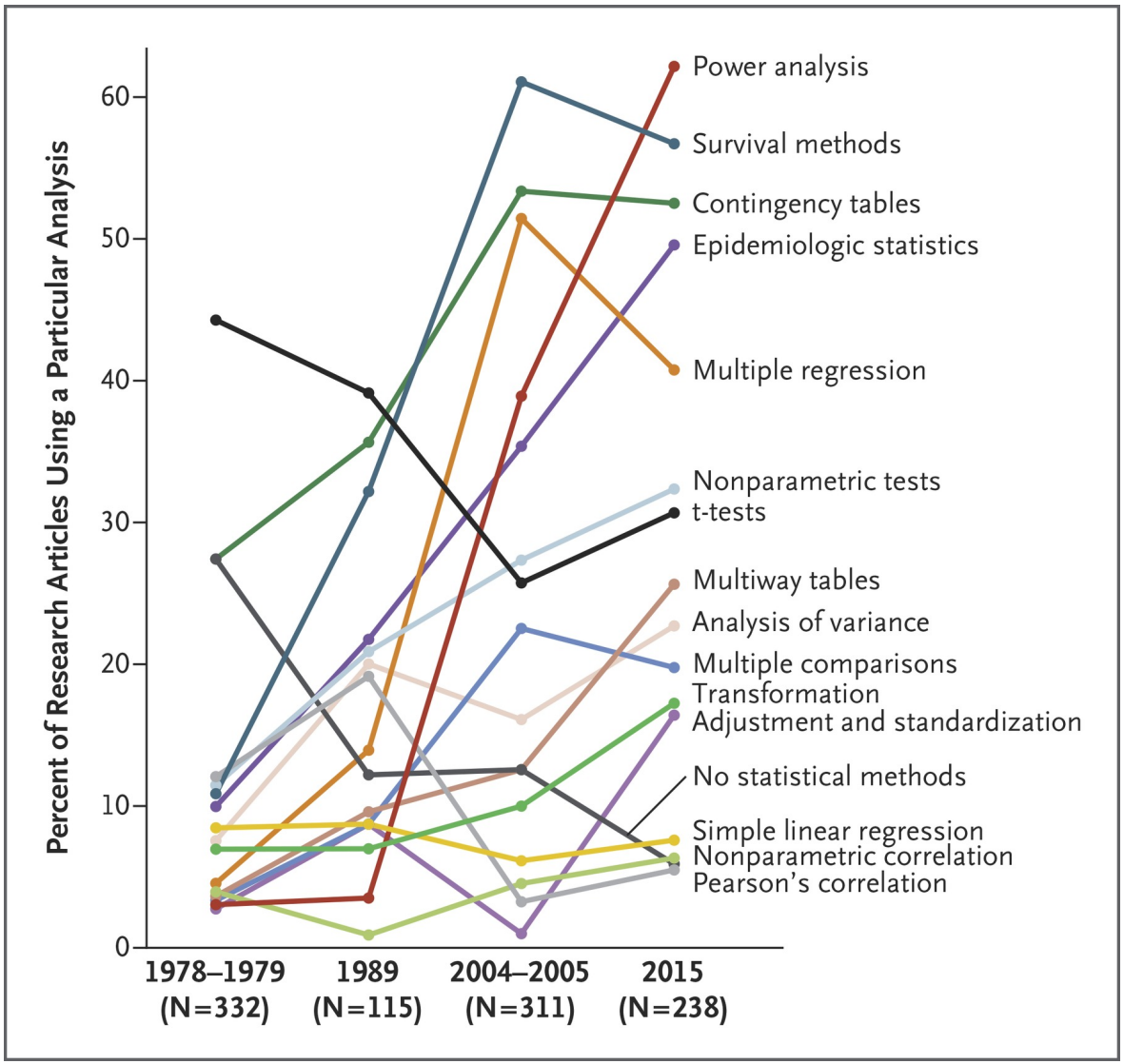
- und Covariablen hinzufügen (die Varianz erklären)
- und $f()$ ändern, wenn die z.B. (0,1) vorliegt
- und erhalten einen p-Wert für die Signifkanz ($p \leq \alpha$)

Small Data – Ein Model von Gewicht und Sport

		p_1	p_2	p_3	p_4
	Gewicht	Sport	Geschlecht	Rauchen	Kalorien
n_1	86.08	wenig	weiblich	1	2770.41
n_2	82.29	wenig	weiblich	0	2082.58
n_3	89.46	wenig	männlich	1	2653.81
n_4	87.98	wenig	männlich	0	2644.13
n_5	93.18	wenig	weiblich	1	2296.45
n_6	79.45	viel	weiblich	1	1858.53
n_7	69.59	viel	männlich	0	2627.94
n_8	74.93	viel	männlich	0	2454.92
n_9	85.48	viel	weiblich	0	1535.99
n_{10}	84.88	viel	männlich	1	2620.09

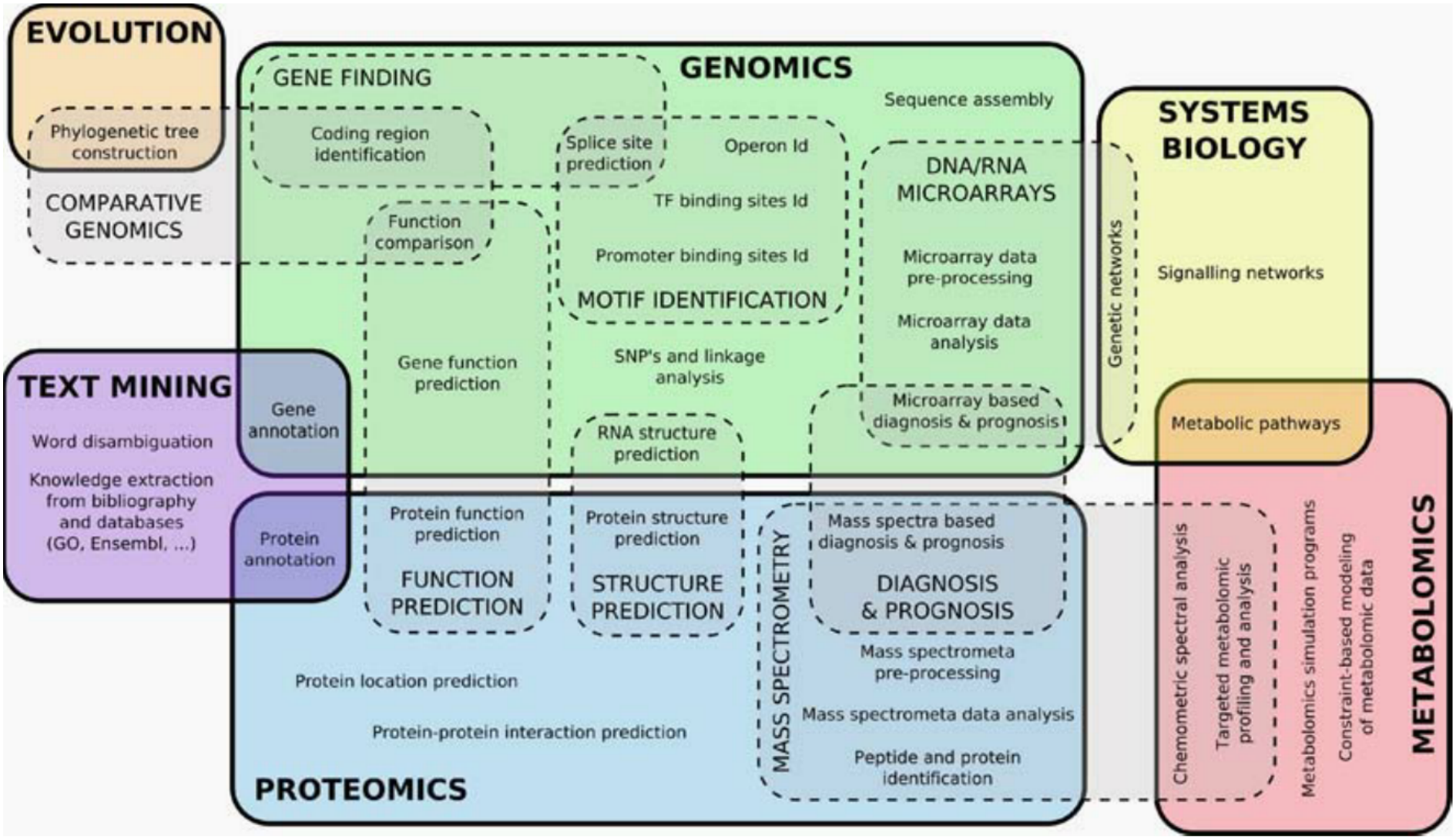
n > p

t-Test und 2x2 Tafeln...



Sato, Y. (2017) Statistical Methods in the Journal - An Update. N Engl J Med.

Überblick



Inza (2010) Met Mol Biol 593:25

Big data in der Informationstechnologie

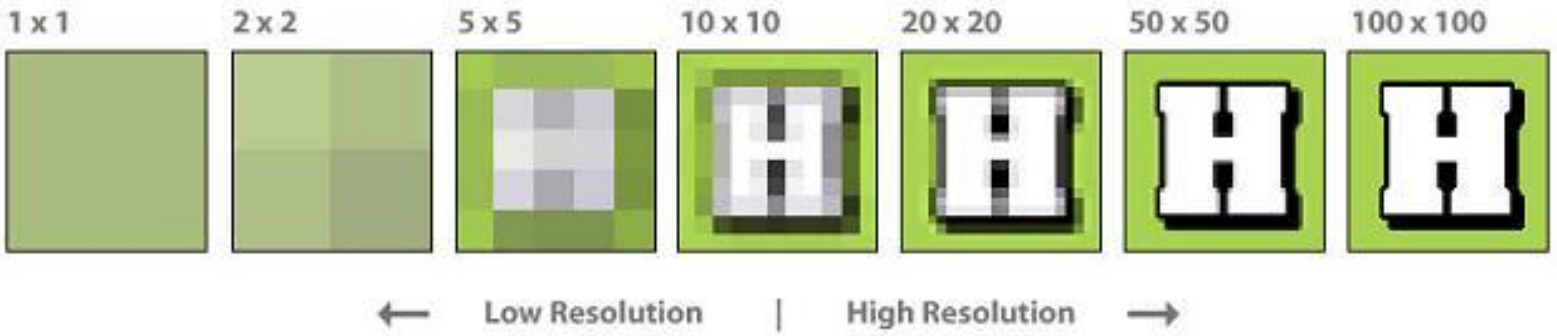
Big Data und Informationen: Text Mining

Einfache Frage: Welchen Buchstaben sehen Sie?

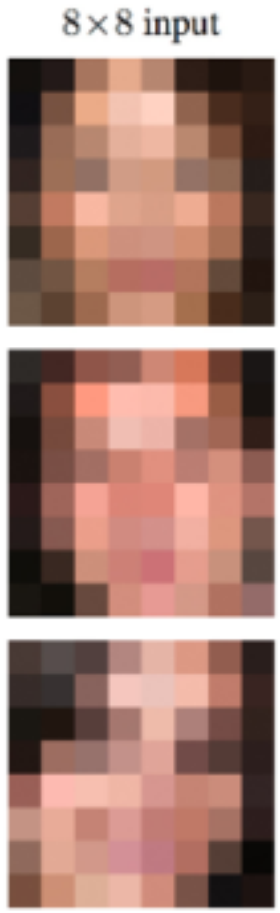
Big Data und Informationen: Text Mining

Big Data und Informationen: Text Mining

Big Data und Informationen: Text Mining

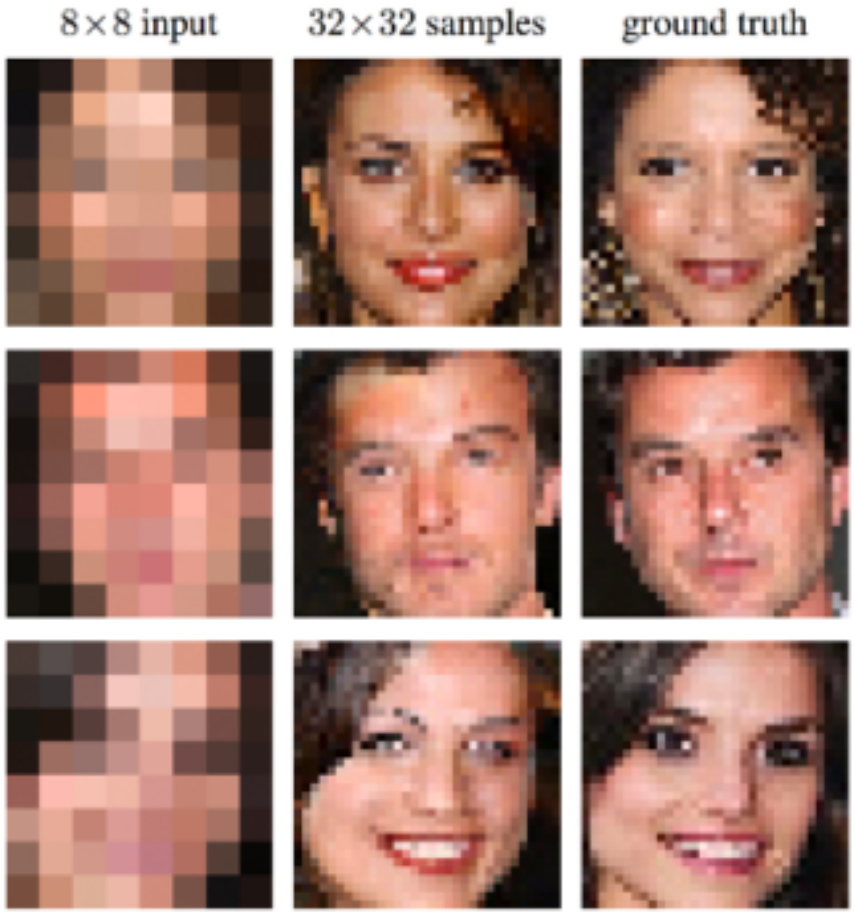


Big Data und Informationen: Bilderkennung



Einfache Frage: Welche Geschlechter sehen Sie?

Big Data und Informationen: Bilderkennung



Big Data und Informationen: Medizin

- Im Februar bei einem Allgemeinarzt

Big Data und Informationen: Medizin

- Im Februar bei einem Allgemeinarzt
- Patient mit Fieber steht an der Anmeldung

Big Data und Informationen: Medizin

- Im Februar bei einem Allgemeinarzt
- Patient mit Fieber steht an der Anmeldung
- Liefert jeder weitere Parameter mehr Informationen?

Big Data und Informationen: Medizin

- Im Februar bei einem Allgemeinarzt
- Patient mit Fieber steht an der Anmeldung
- Liefert jeder weitere Parameter mehr Informationen?

- Eine wichtige Annahme ist, dass mehr Parameter auch mehr Informationen bedeuten
- Das ist in der Medizin nicht so gegeben, wie in dem Bereich wo maschinelle Lernverfahren entwickelt werden

Maschinelle Lernverfahren

- Inza, I., Calvo, B, Armananzas, R., Bengoetxea, E., Larranaga, P. and Lozano, J,A. (2010) *Machine learning: an indispensable tool in bioinformatics.* Methods in Molecular Biology, 593:25
- Libbrecht M. W. and Noble W. S. (2015) *Machine learning applications in genetics and genomics.* Nature Review Genetics, 16:321-32
- Yip, K. Y., Cheng, C. and Gerstein, M. (2013) *Machine learning and genome annotation: a match meant to be?* Genome Biology, 14:205

Maschinelle Lernverfahren

- Inza, I., Calvo, B, Armananzas, R., Bengoetxea, E., Larranaga, P. and Lozano, J,A. (2010) *Machine learning: an indispensable tool in bioinformatics.* Methods in Molecular Biology, 593:25
- Libbrecht M. W. and Noble W. S. (2015) *Machine learning applications in genetics and genomics.* Nature Review Genetics, 16:321-32
- Yip, K. Y., Cheng, C. and Gerstein, M. (2013) *Machine learning and genome annotation: a match meant to be?* Genome Biology, 14:205

Maschinelle Lernverfahren

Vokabular

Modellbeschreibung

$$y \sim a + b + c$$

wobei

- y ist die response oder Endpoint/Endpunkt oder gemessene Variable/Outcome: Krebs ja/nein
- a, b, c sind die Covariablen, Risk factors oder Variablen: sex, age, dose level

Maschinelle Lernverfahren

Vokabular

Modellbeschreibung

$$y \sim a + b + c$$

wobei

- y ist die response oder Endpoint/Endpunkt oder gemessene Variable/Outcome: Krebs ja/nein
- a, b, c sind die Covariablen, Risk factors oder Variablen: sex, age, dose level

$$Krebs\,[ja/nein] \text{ hängt ab } sex + age + dose$$

label hängt ab *features*

Maschinelle Lernverfahren

Vokabular

Modellbeschreibung

label hängt ab *features*

label ist das Ziel der Vorhersage

feature einzelne Varibalen, die für das maschinelle Lernverfahren verwendet werden

Ziel von maschinellen Lernverfahren

Etwas Vorhersagen (test data) mit der Hilfe von etwas Anderem (training data)

Libbrecht (2015) Nat Rev Genet 16:321

Maschinelle Lernverfahren

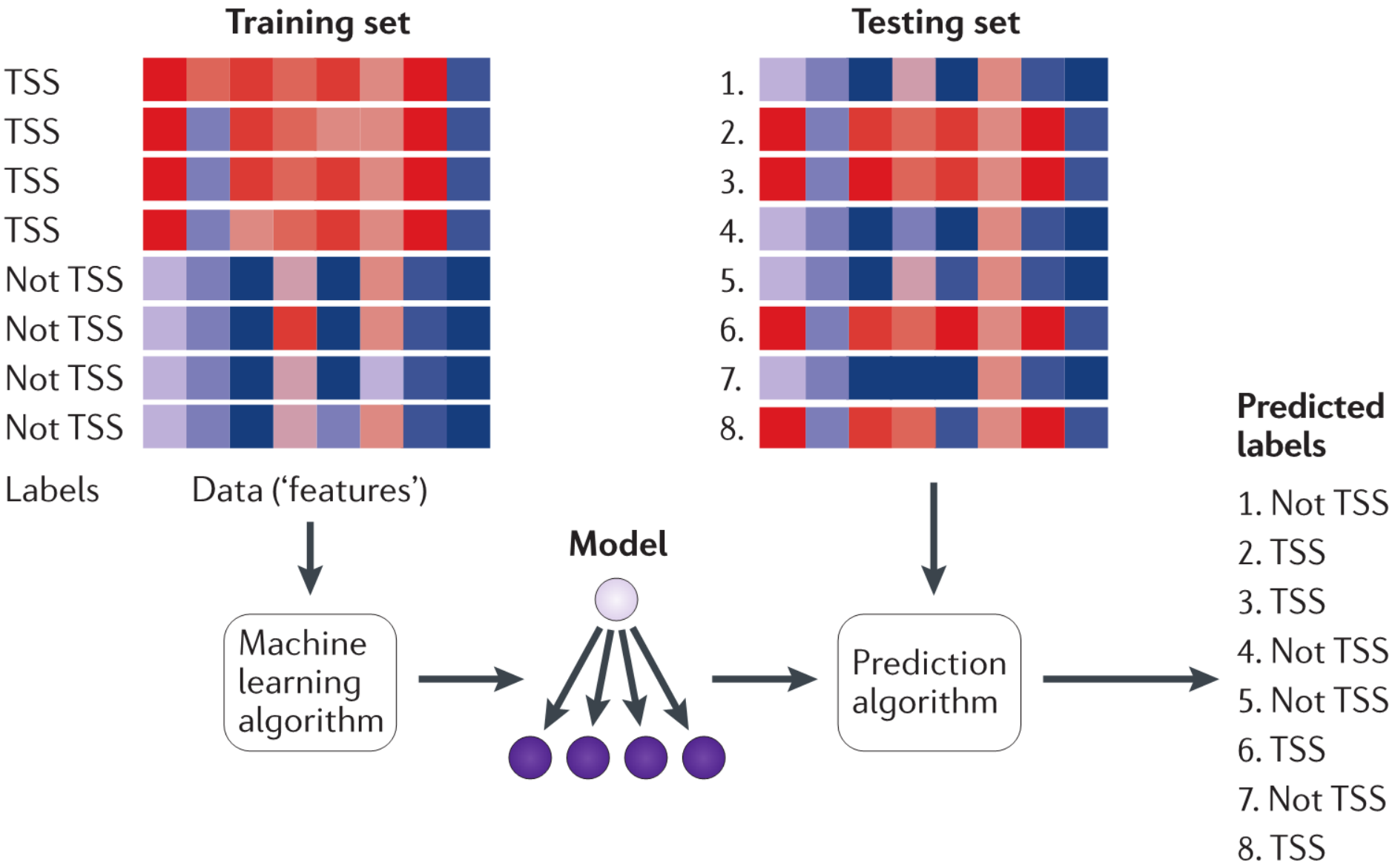
Ziel von maschinellen Lernverfahren

Etwas Vorhersagen (test data) mit der Hilfe von etwas Anderem (training data)

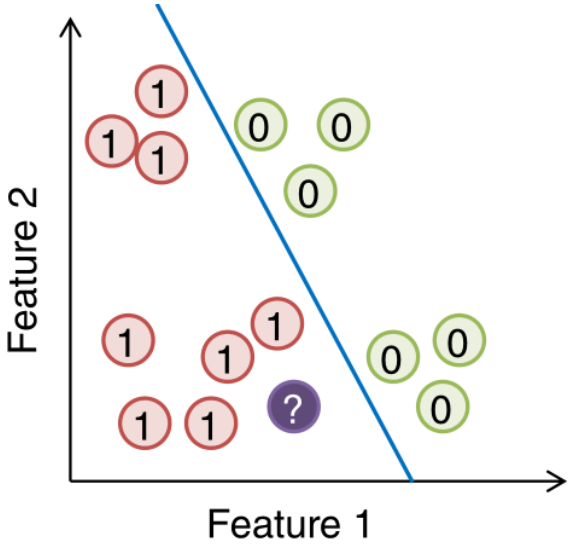
- Maschinelle Lernverfahren behandeln Klassifikation
  - Ist der Patient ein Krebspatient gegeben der Feature?
  - In welche Gruppe gehört eine Maus?

Ein p-Wert ist nicht zu berechnen

Maschinelle Lernverfahren

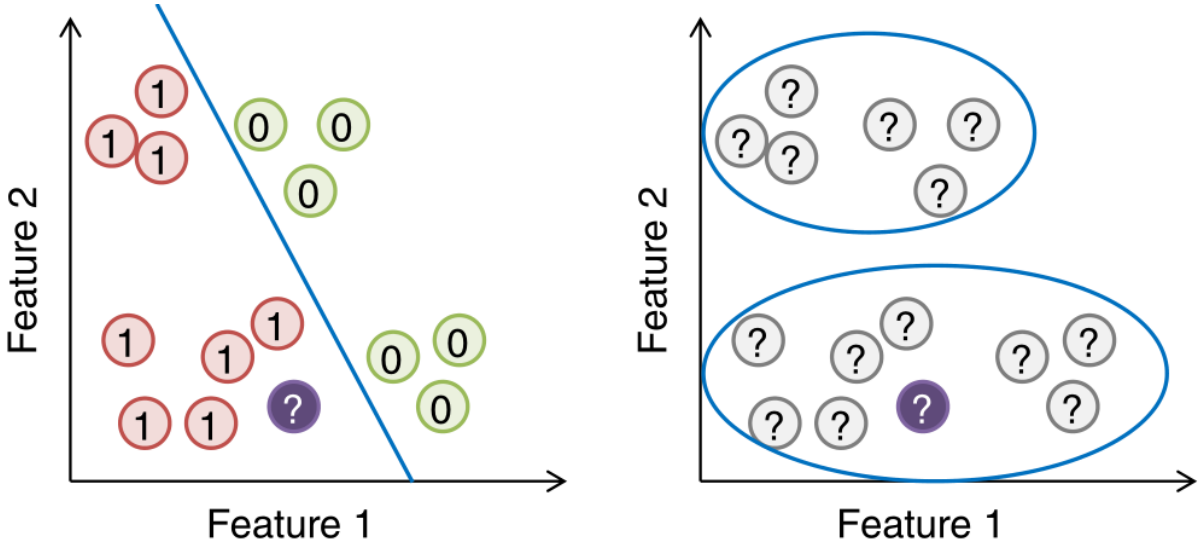


Maschinelle Lernverfahren



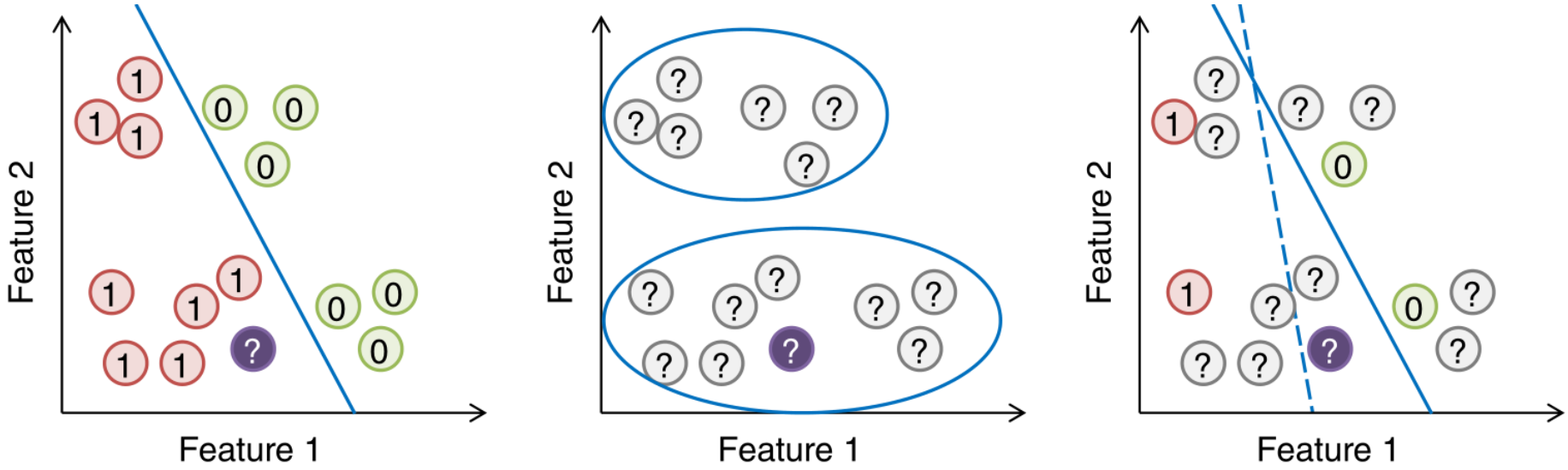
- **Supervised learning:** Maschinelle Lernverfahren basierend auf gelabelten Patienten. Diese werden genutzt um nicht bekannte Labels vorherzusagen.

Maschinelle Lernverfahren



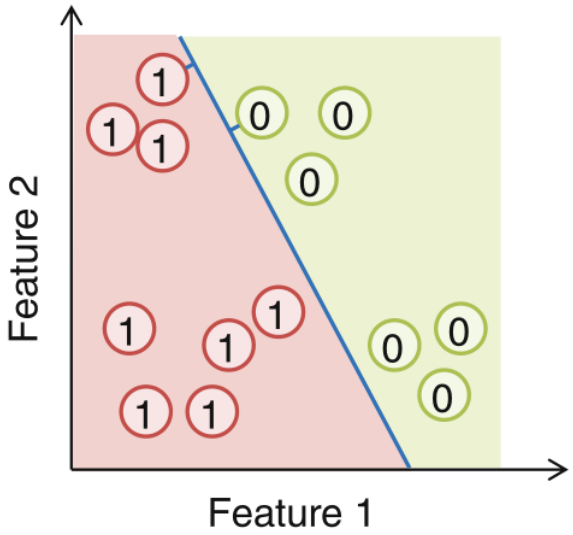
- **Supervised learning:** Maschinelle Lernverfahren basierend auf gelabelten Patienten. Diese werden genutzt um nicht bekannte Labels vorherzusagen.
- **Unsupervised learning:** Maschinelle Lernverfahren ohne bekannte Label

Maschinelle Lernverfahren



- **Supervised learning:** Maschinelle Lernverfahren basierend auf gelabelten Patienten. Diese werden genutzt um nicht bekannte Labels vorherzusagen.
- **Unsupervised learning:** Maschinelle Lernverfahren ohne bekannte Label
- **Semi-supervised learning:** Maschinelle Lernverfahren nutzen eine Mischung

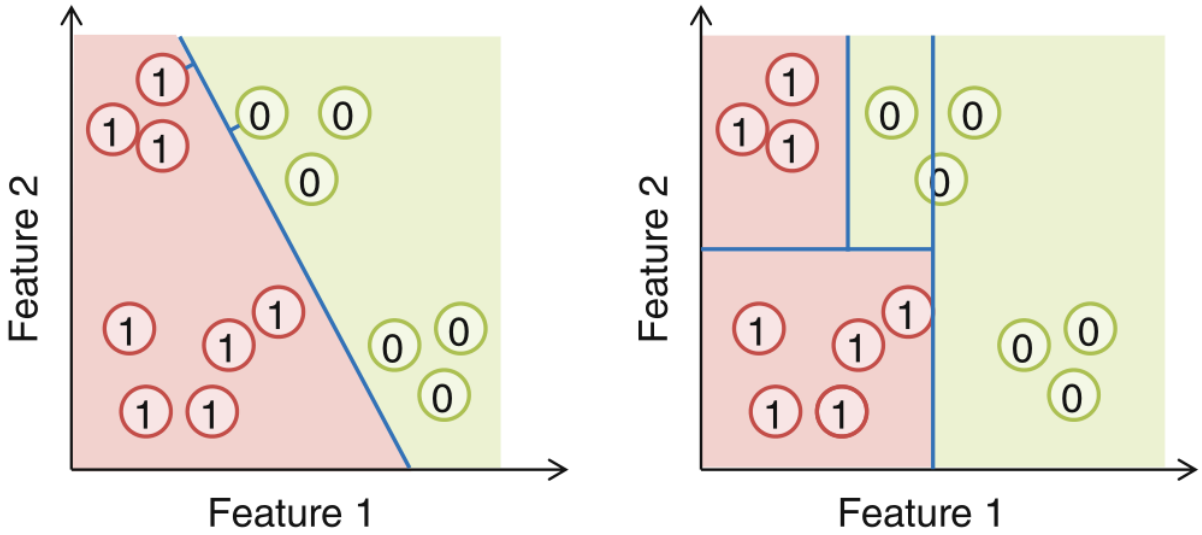
Die drei meist verbreitesten Maschinen



- **Support vector machine:** Zeichne eine Linie durch Punkte

Yip (2013) Genom Biol 14:205

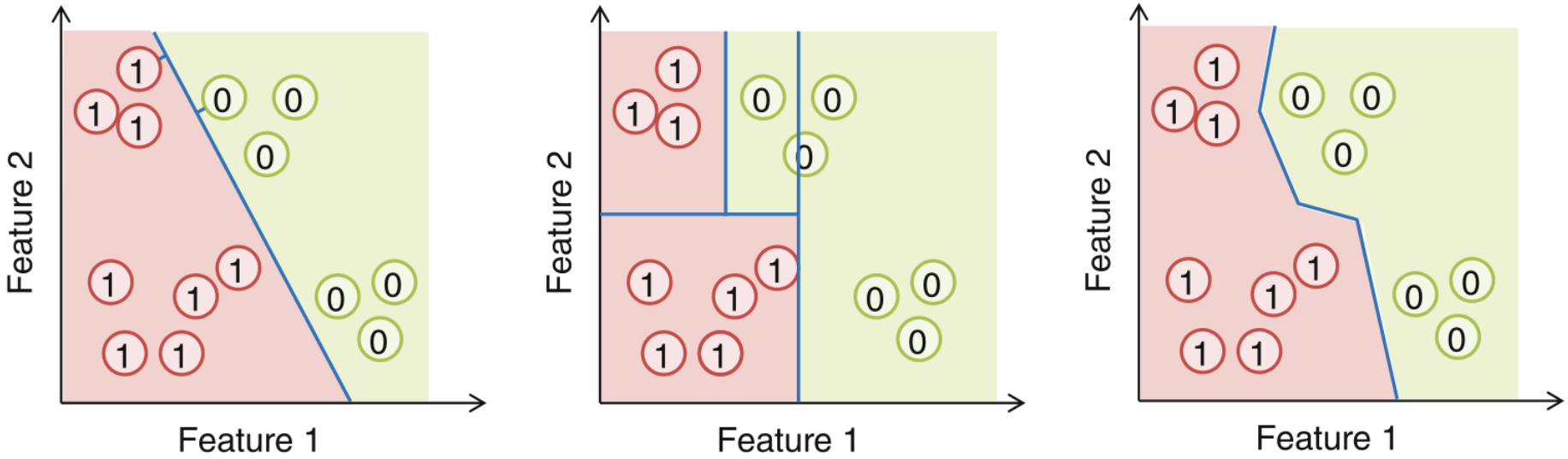
Die drei meist verbreitesten Maschinen



- **Support vector machine:** Zeichne eine Linie durch Punkte
- **Random Forest:** Baue eine Mischung aus verschiedenen Entscheidungsbäumen

Yip (2013) Genom Biol 14:205

Die drei meist verbreitesten Maschinen

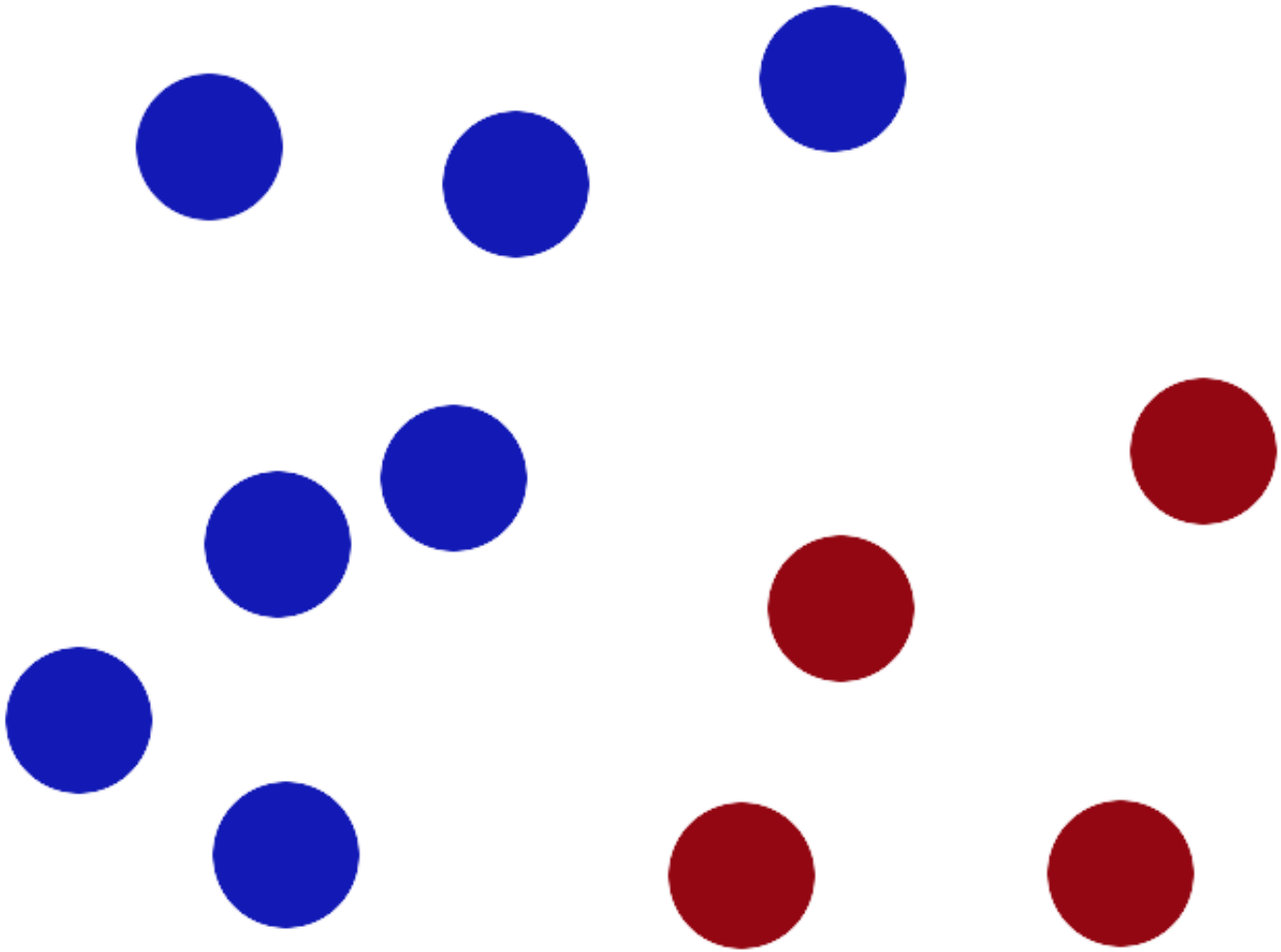


- **Support vector machine:** Zeichne eine Linie durch Punkte
- **Random Forest:** Baue eine Mischung aus verschiedenen Entscheidungsbäumen
- *k* **nearest neighbor:** Ich mache was mein Nachbar macht

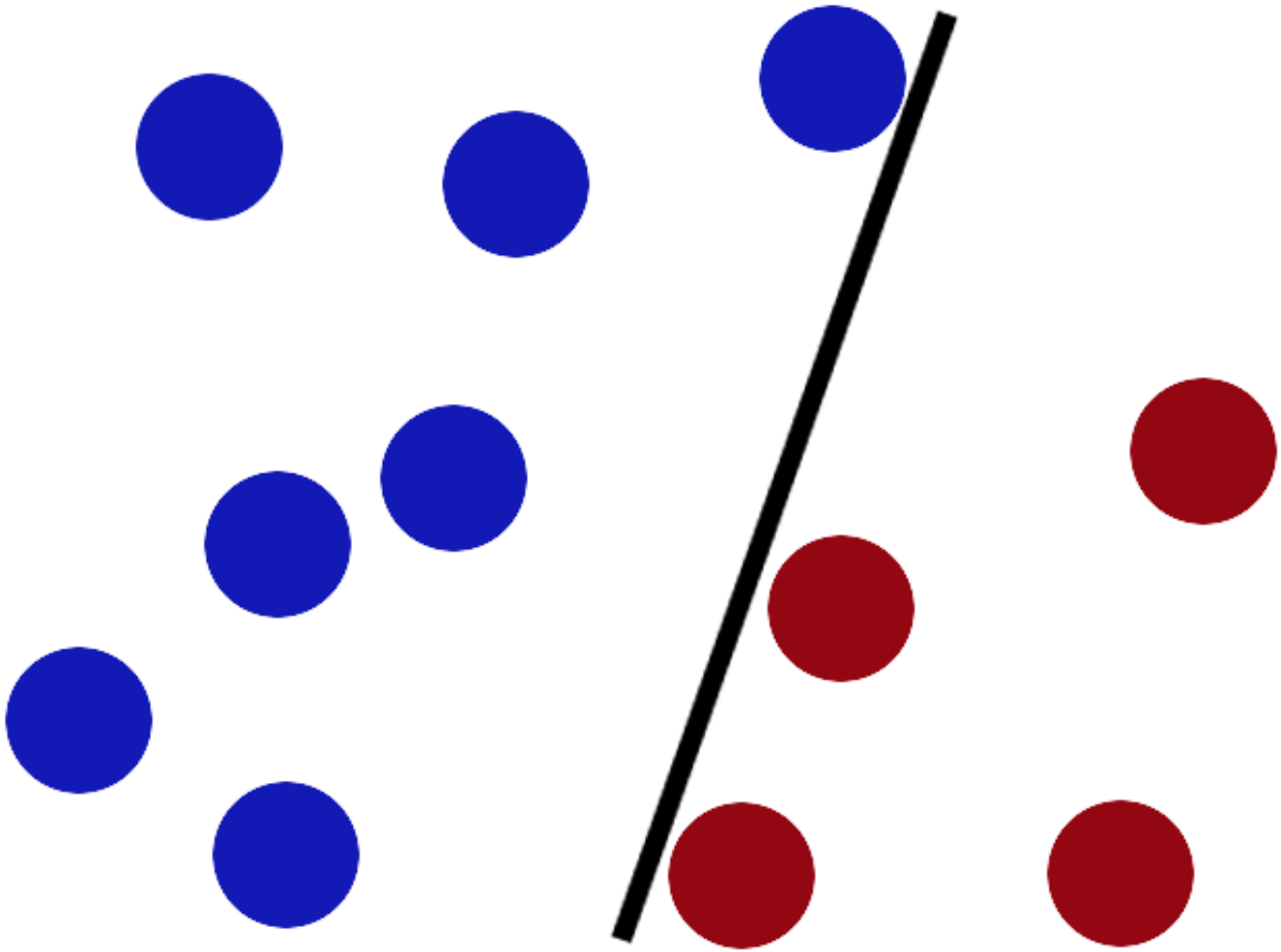
Yip (2013) Genom Biol 14:205

Support vector machines (SVM)

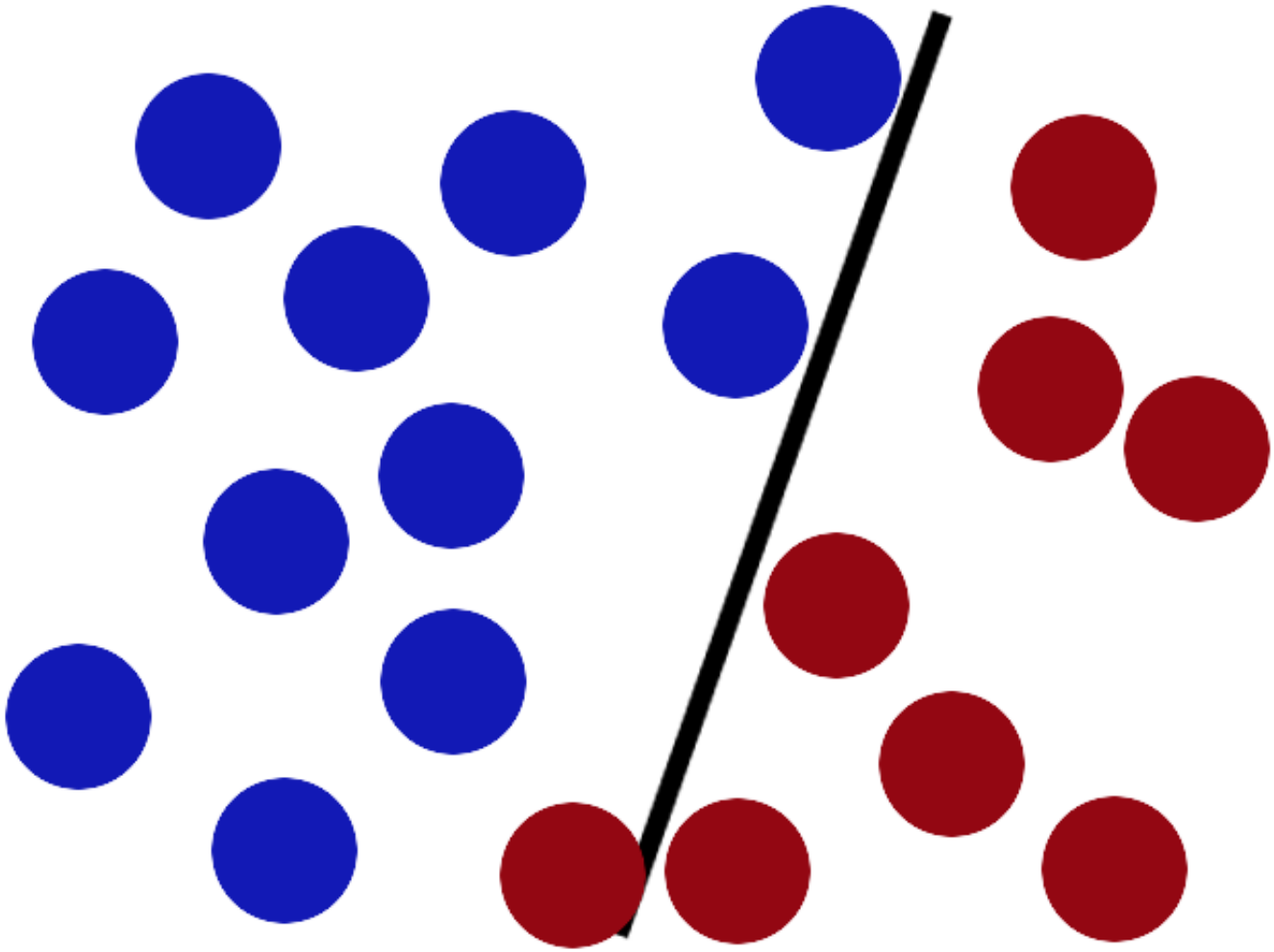
Die farbigen Bälle sollen getrennt werden



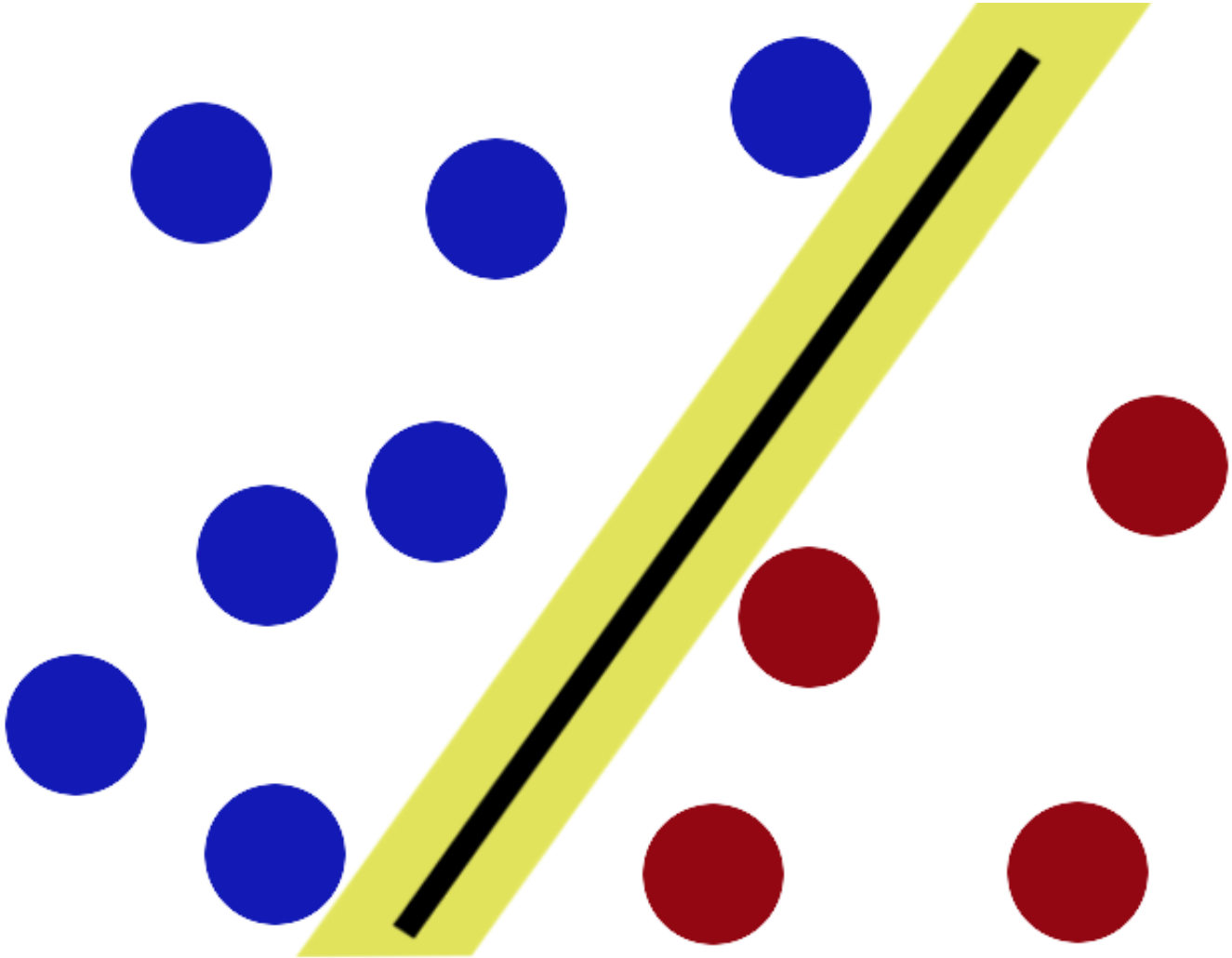
Eine einfache Linie macht den Job



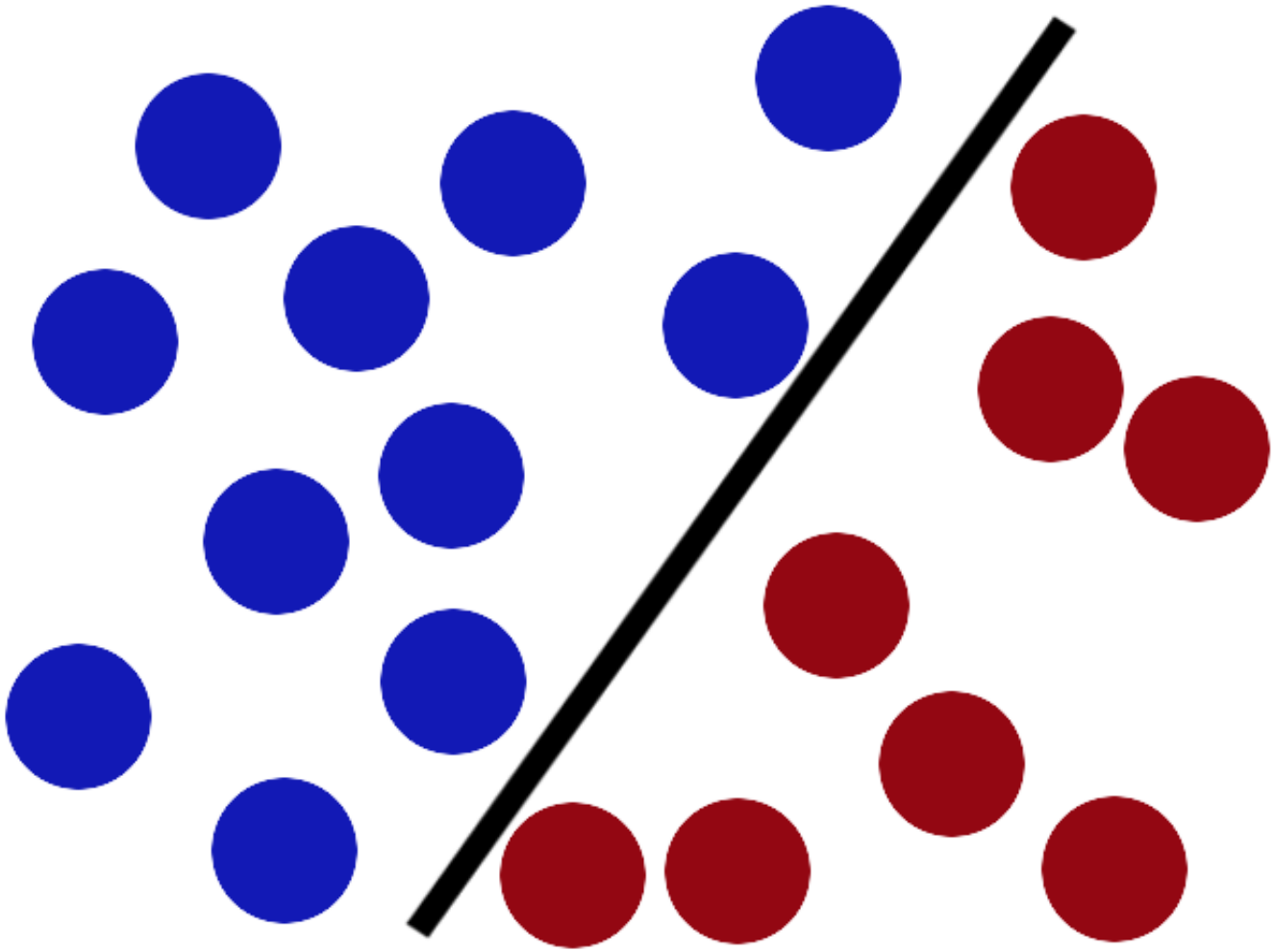
Wir erhöhen die Anzahl an Bällen nachträglich



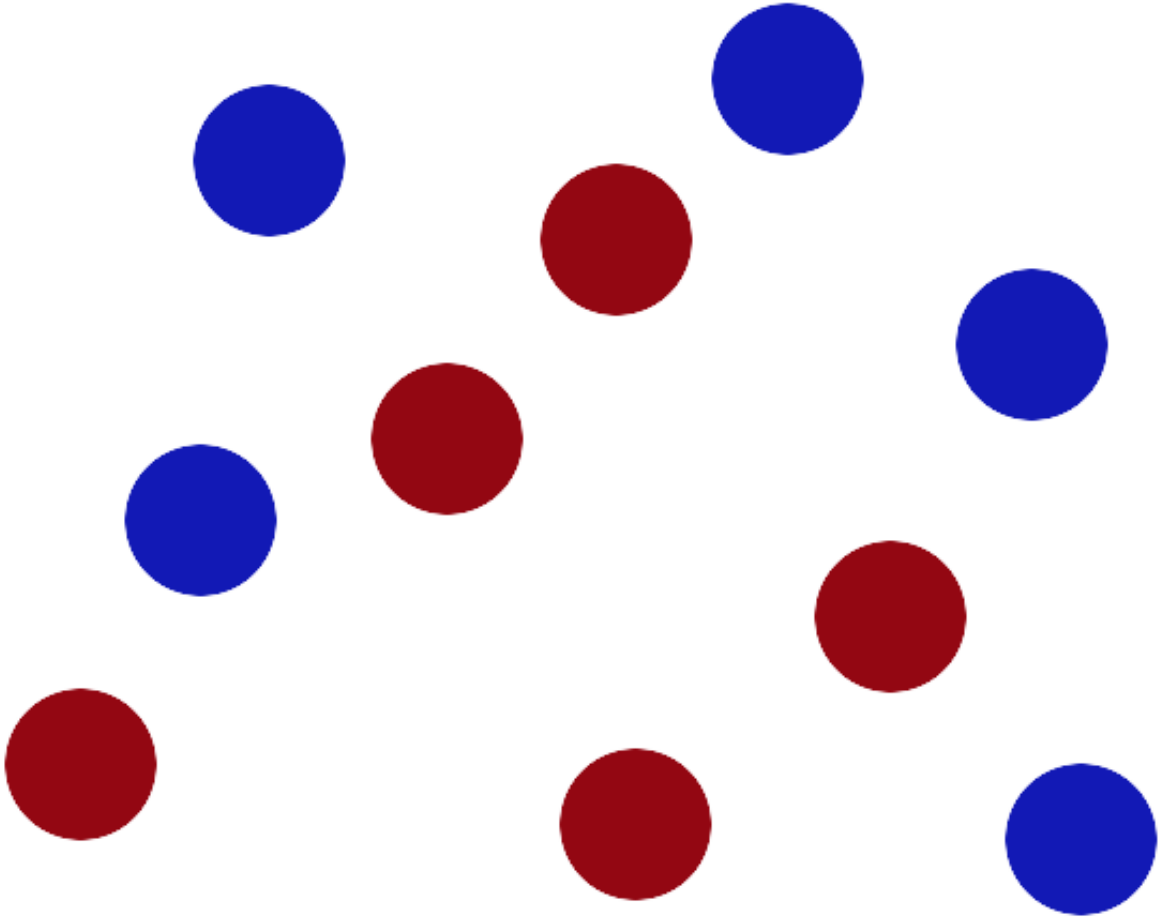
SVM's versuchen so viel Platz wie möglich zwischen der Linie zu generieren



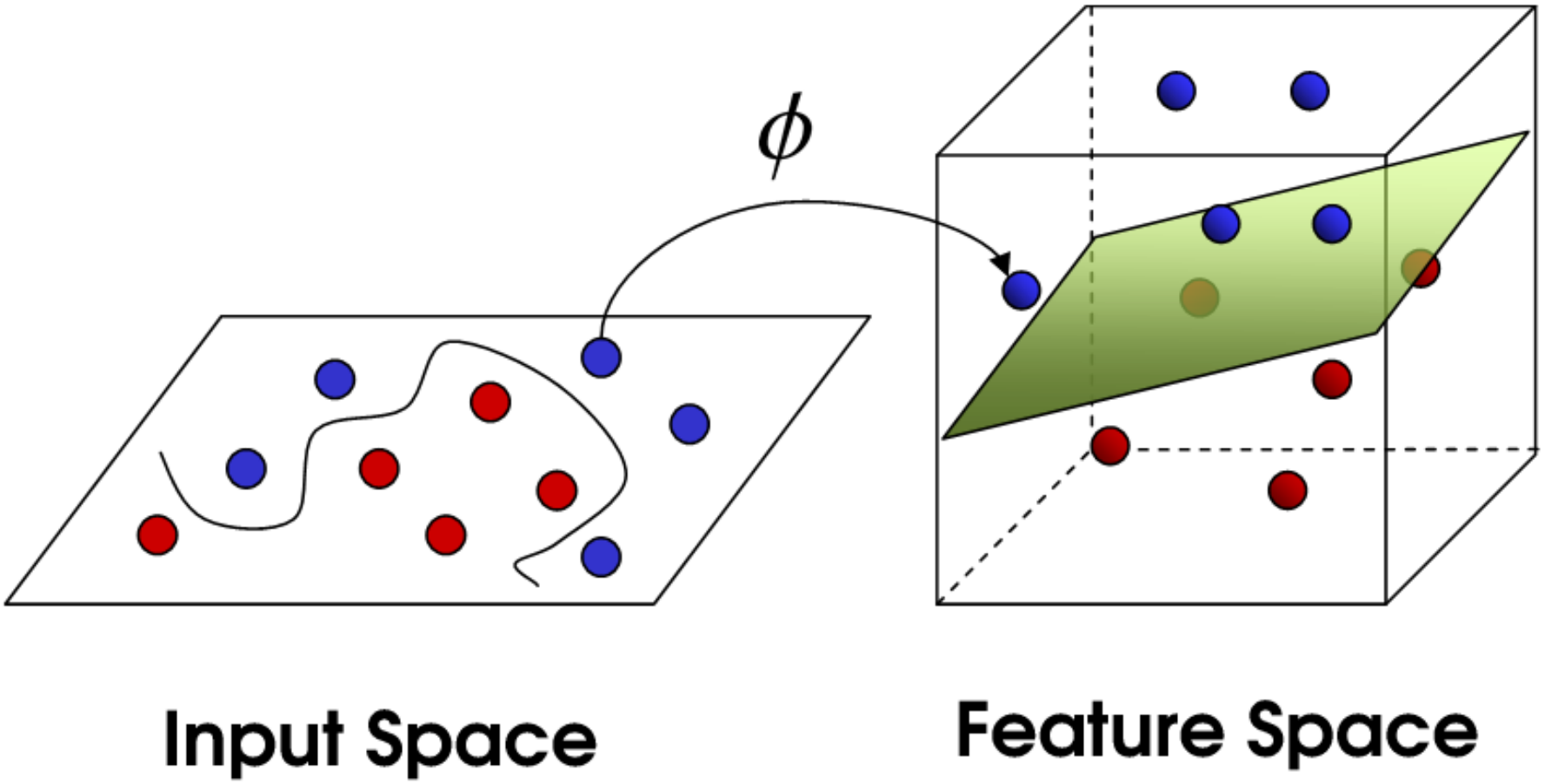
Mehr Bälle passen hinein...



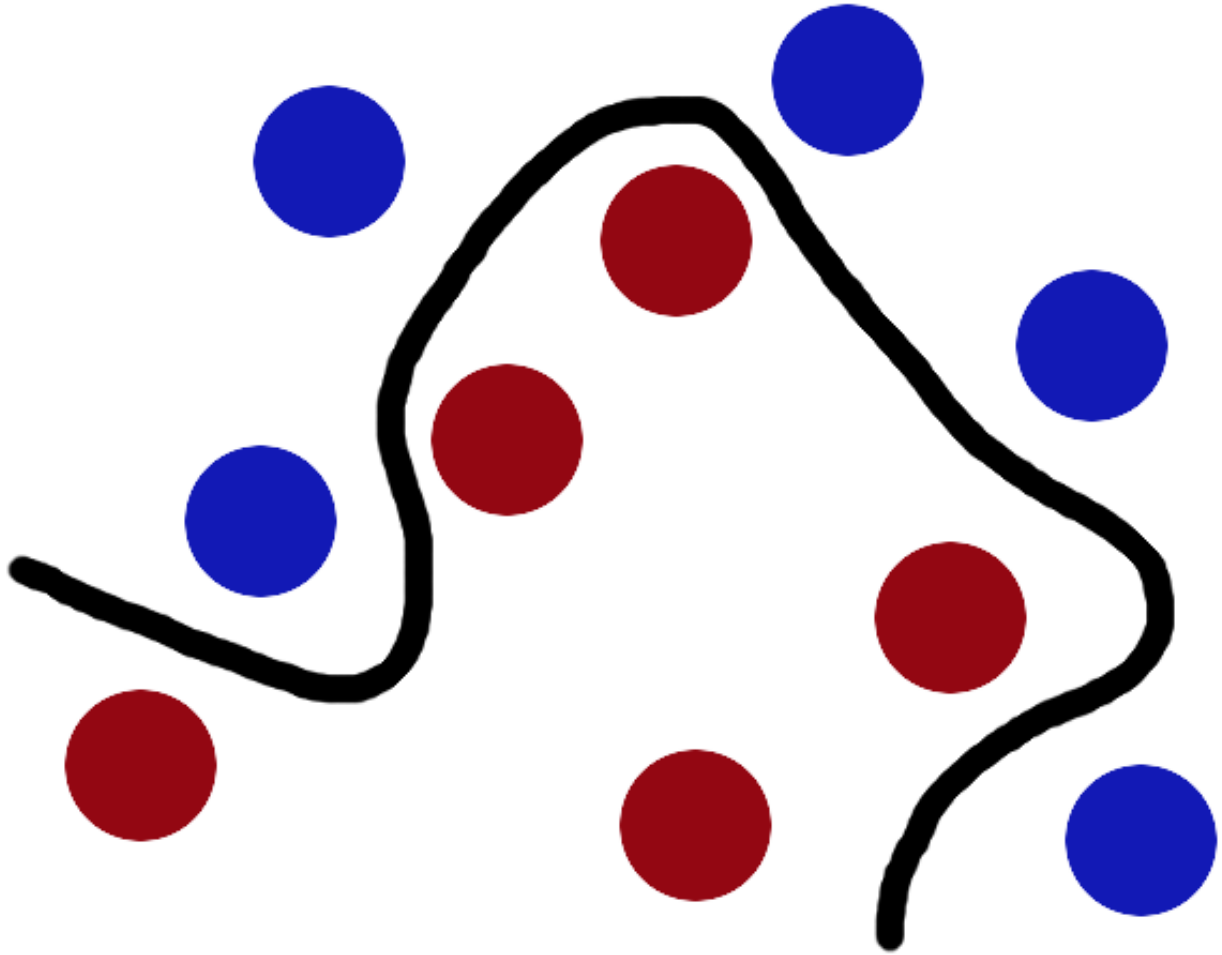
Es gibt auch komplexe Situationen



Transformieren von 2D zu 3D

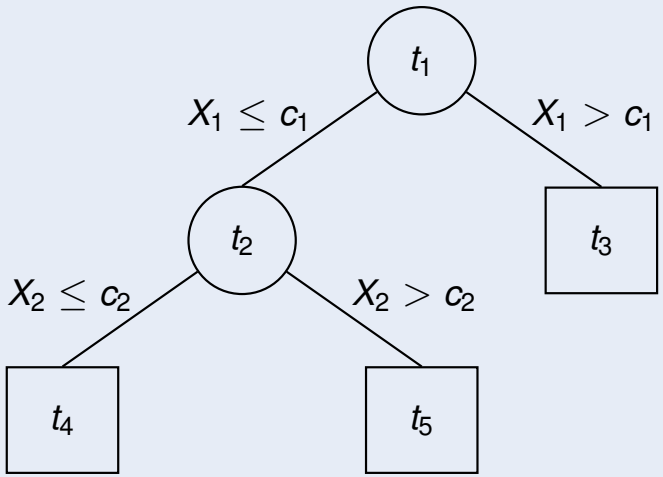


Rücktransformation von 3D zu 2D

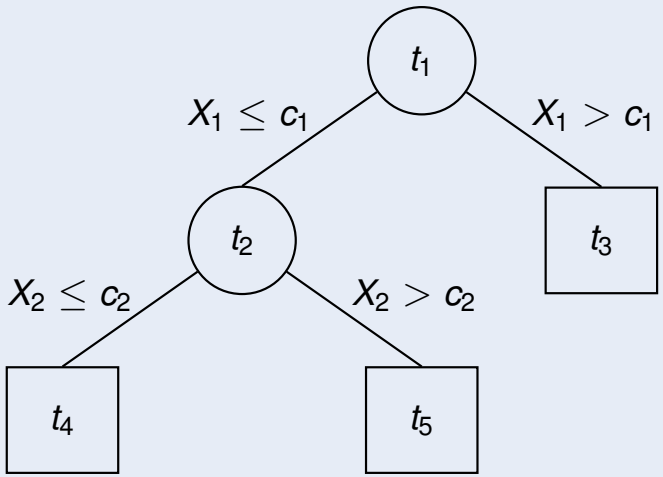


Random Forest

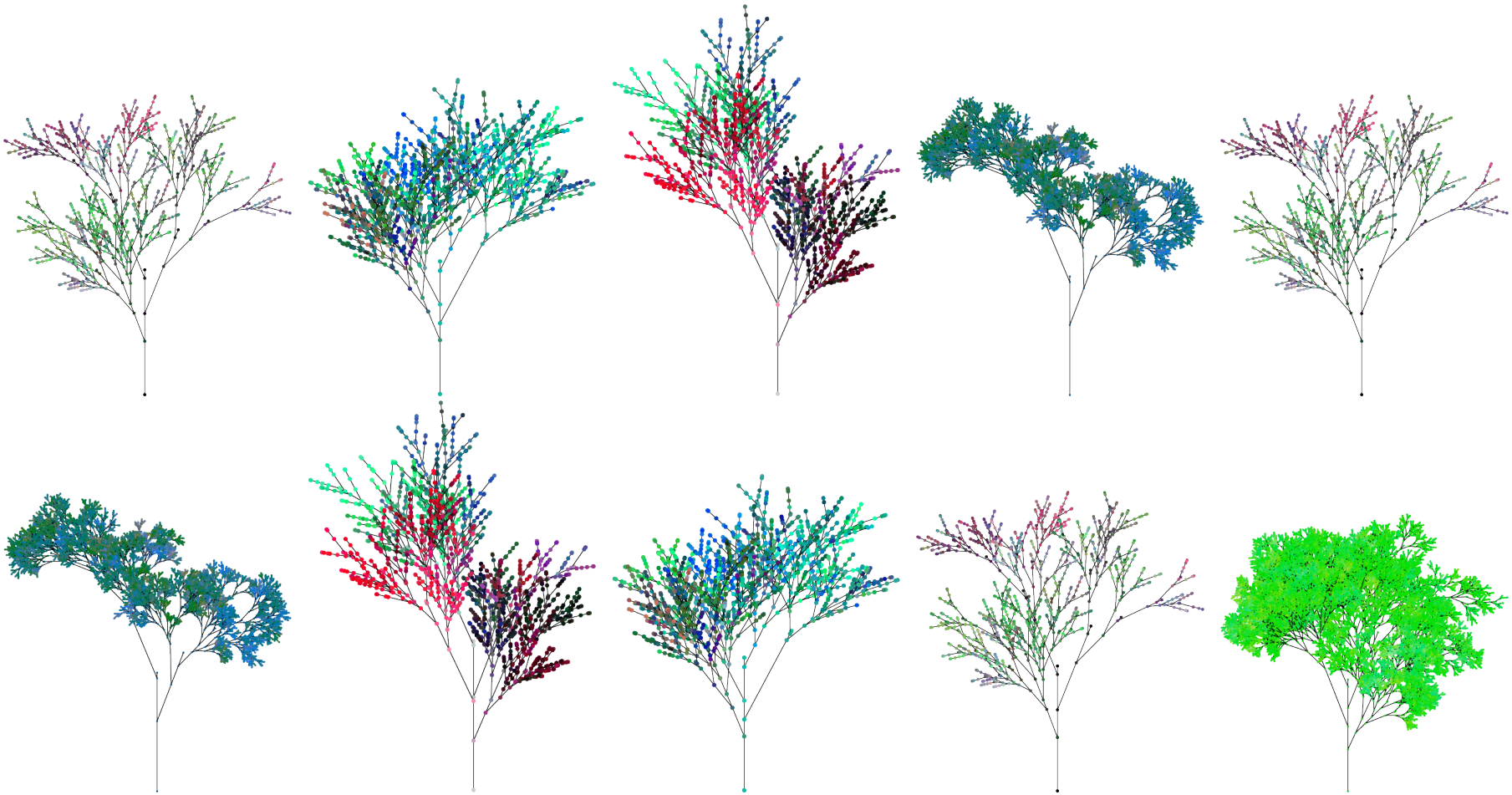
Random Forest basiert auf Entscheidungsbäumen



Random Forest basiert auf Entscheidungsbäumen



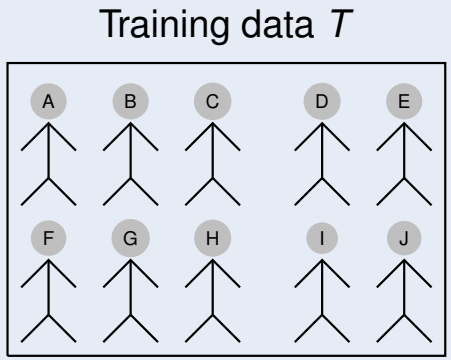
Random Forest a ensemble of trees



- Random forest ist ein Wald aus bis zu 1000 Bäumen

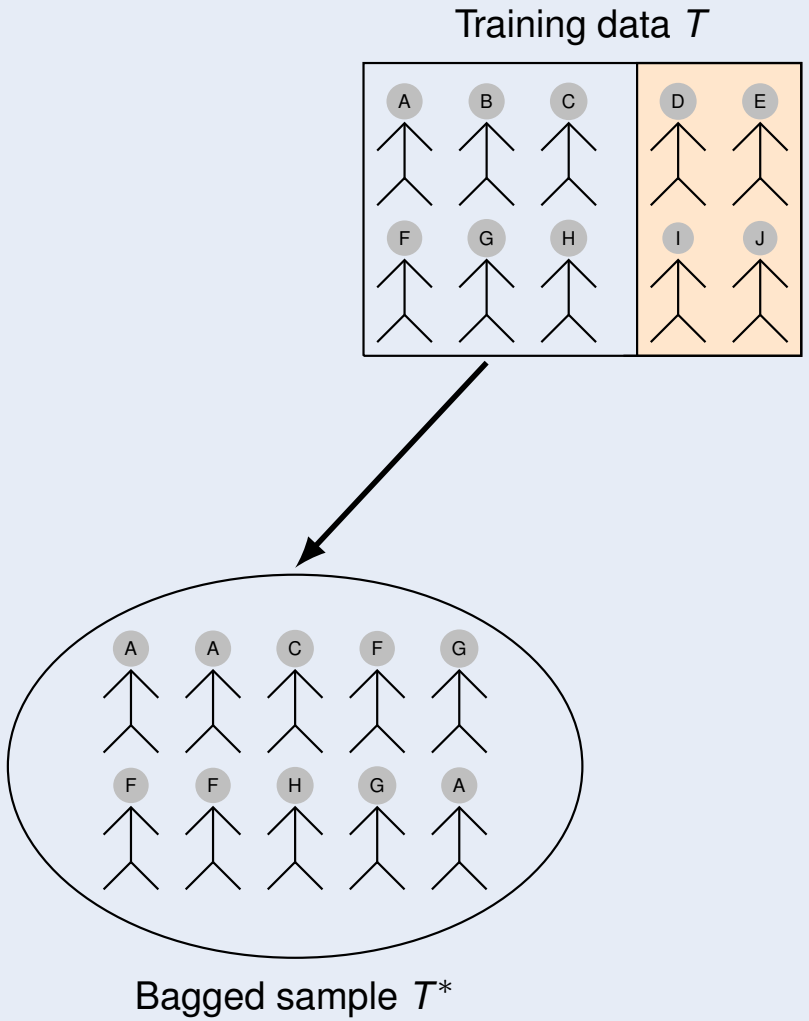
Random Forest: Wie Varibilität generieren?

Für einen Baum nutzen wir ein Bootstrap Sample



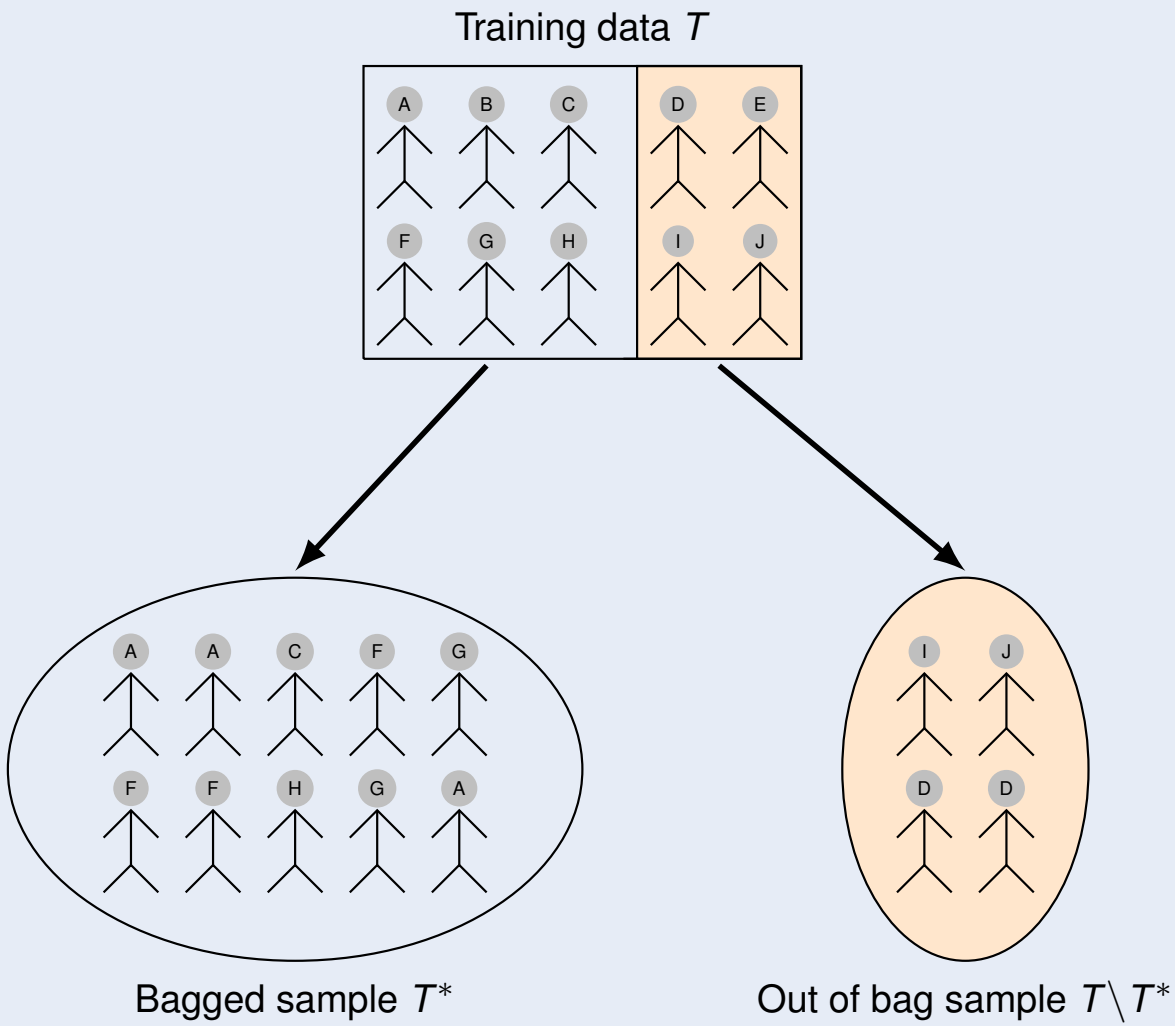
Random Forest: Wie Varibilität generieren?

Für einen Baum nutzen wir ein Bootstrap Sample



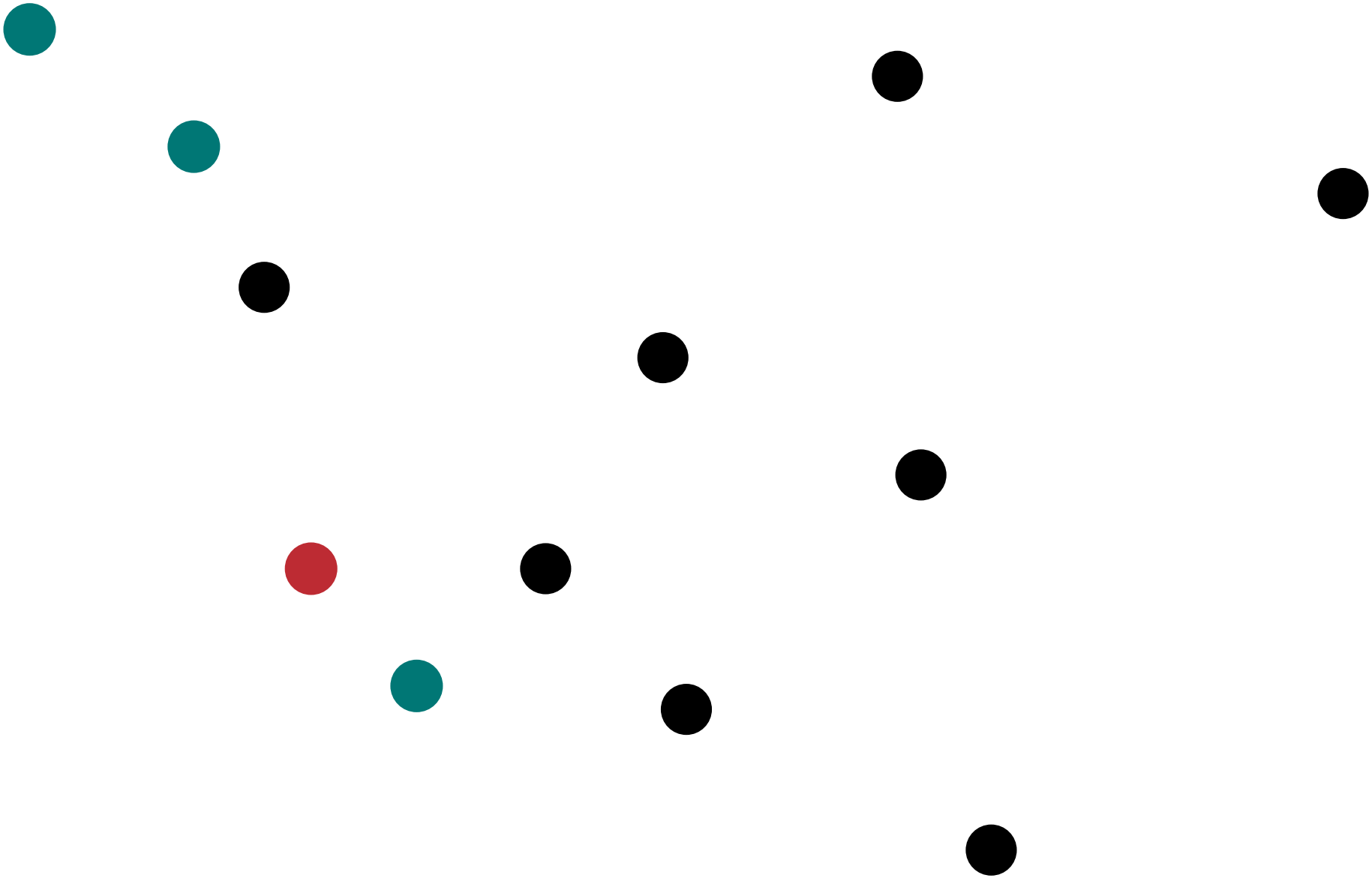
Random Forest: Wie Varibilität generieren?

Für einen Baum nutzen wir ein Bootstrap Sample

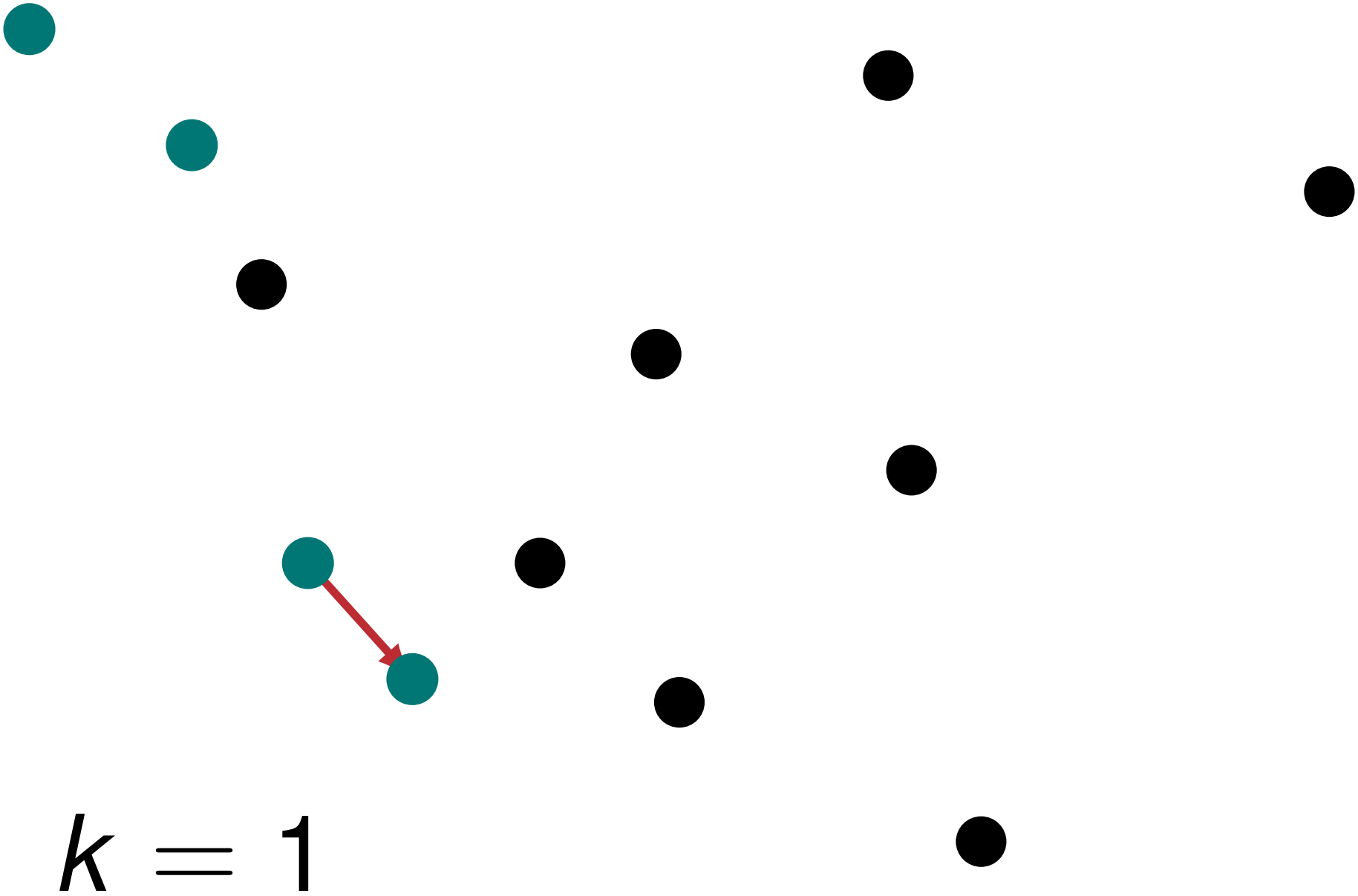


k nearest neighbors (k-NN)

k-NN: Algorithm

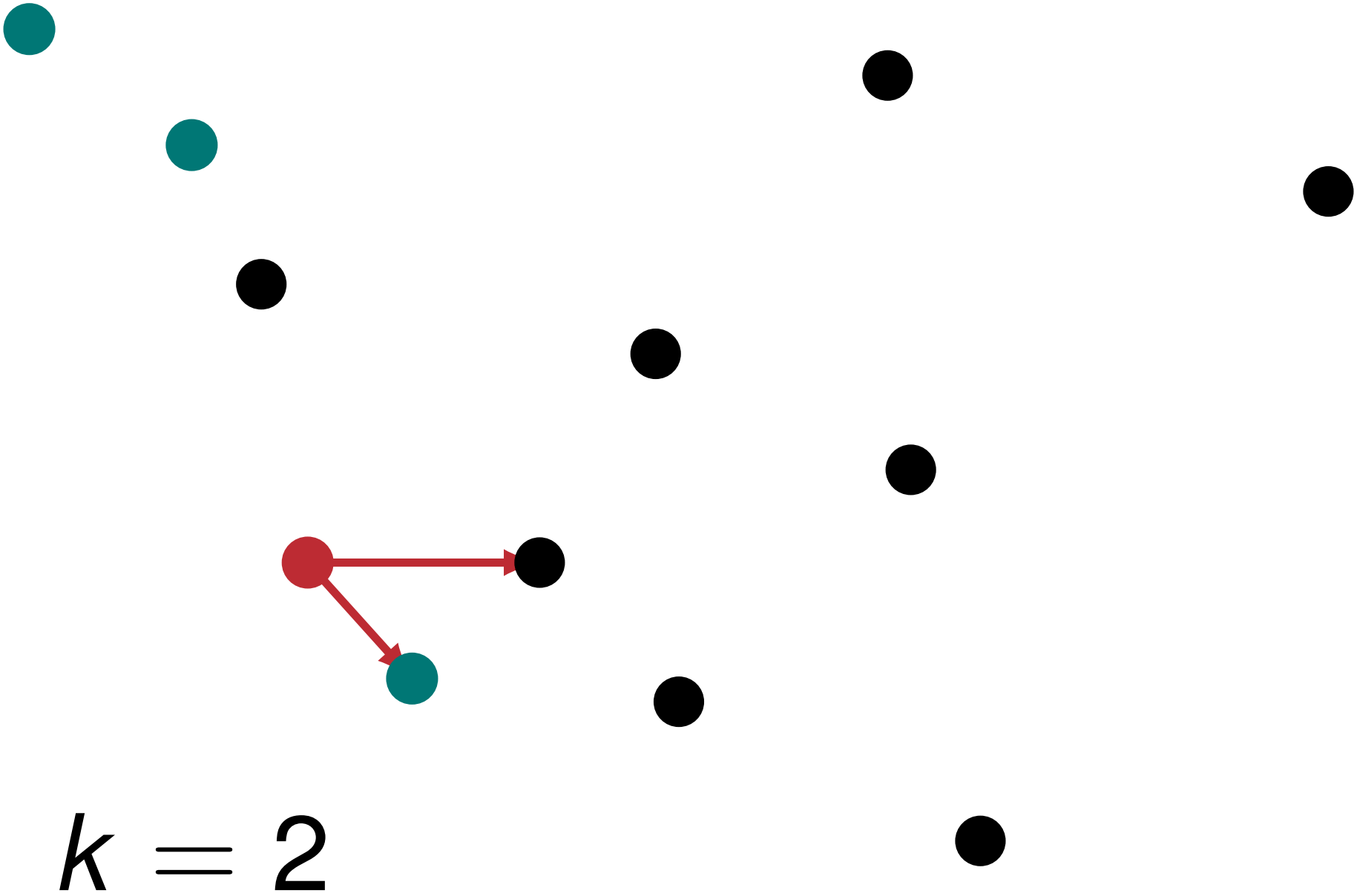


k-NN: Algorithm



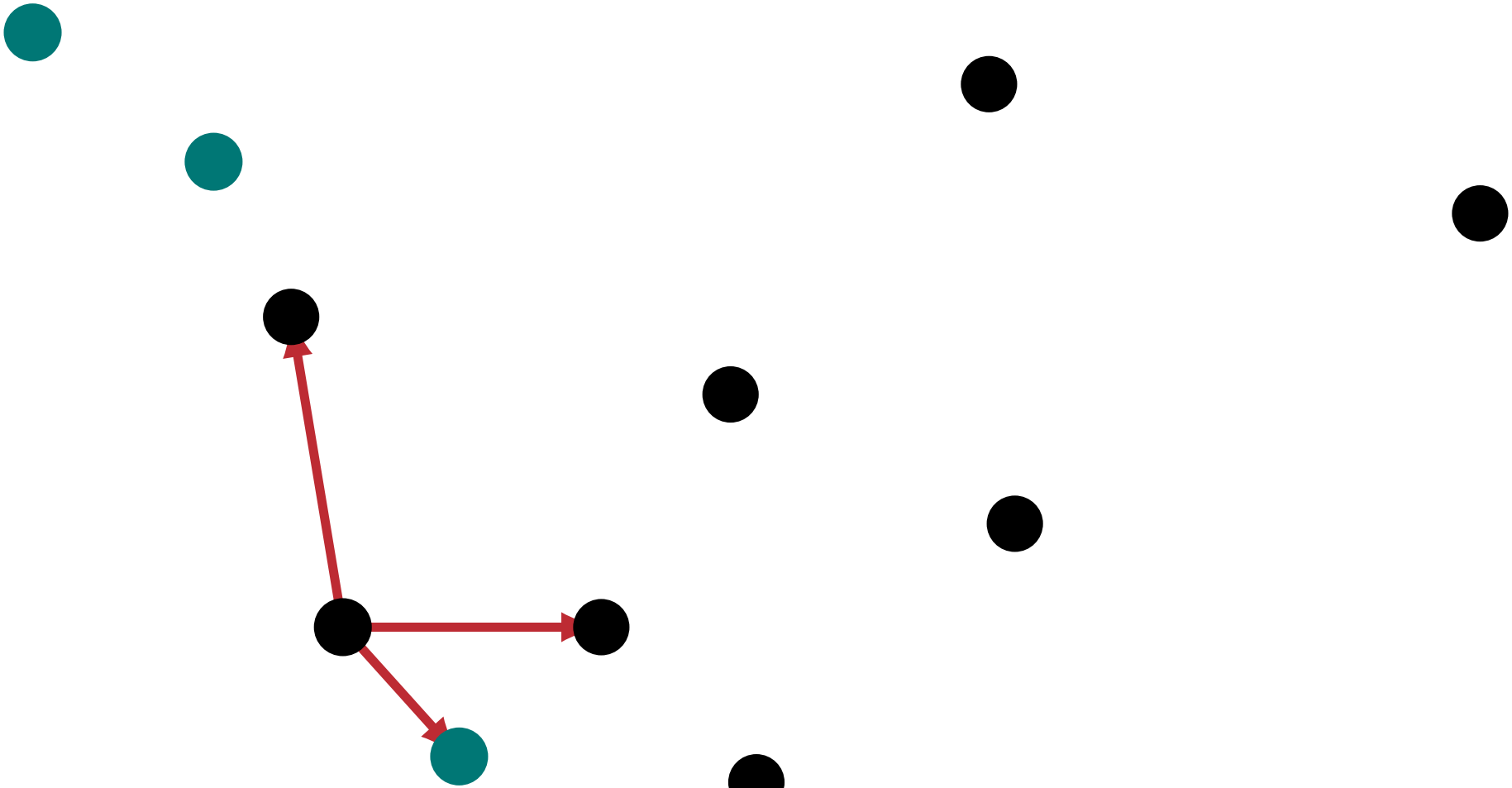
$k = 1$

k-NN: Algorithm



$k = 2$

k-NN: Algorithm



$k = 3$

Probleme mit den Daten

Unterschiedliche Klassengröße

- Die Daten bestehen aus $\sim$ 2000 enhancer sites und 3000000 non enhancer sites

Missing data

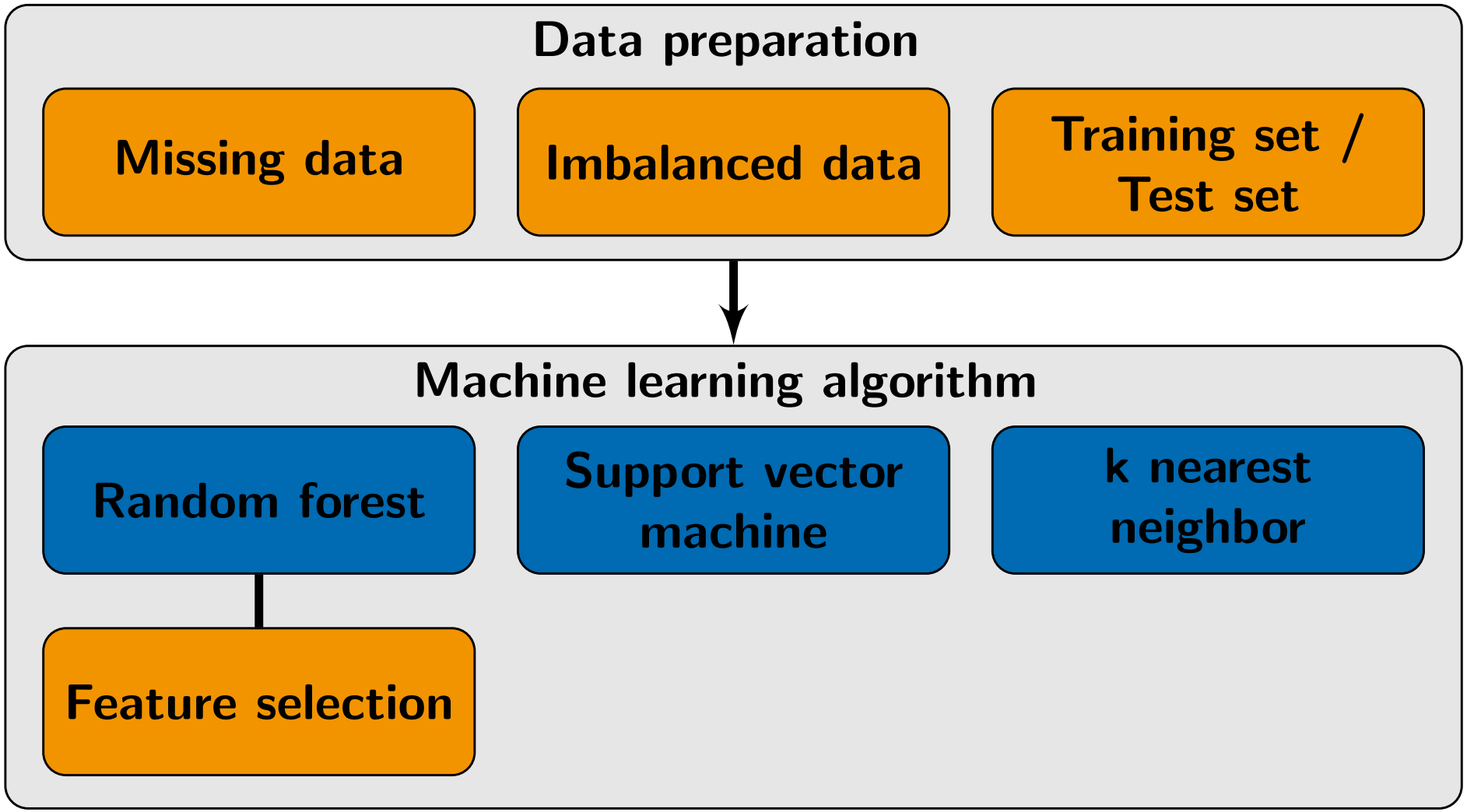
- Die Daten haben fehlende Werte (*NA*, *NaN*, oder ähnliches)

Stark korrelierte Variablen

- BMI und Gewicht wird Probleme verursachen
- Wenn zwei Variablen gleich gut sind, welche bevorzugen?

Libbrecht (2015) Nat Rev Genet 16:321

Zusammenfassung



Regulierung und Anwendung

FDA und EDA

- Maschinelle Lernverfahren sind bekannt und werden auch berücksichtigt
- Mangel an Experten auf dem Gebiet für die Regulierung (Risikoabschätzung)

Vorausetzungen

- Wie auch bei normalen Tests, haben maschinelle Lernverfahren Annahmen
- Fehlende Werte, Unbalanzierte Daten oder starke Korrelationen, können die Klassifikation beeinflussen

Kontakt

Institut für Biometrie und Klinische Epidemiologie

DE

Dr.rer.hum.biol. Jochen Kruppa

AG-Leiter Statistische Bioinformatik

+49 30 450 562 189

Kontakt aufnehmen

Route / Geländeplan

Charité - Universitätsmedizin Berlin
Charitéplatz 1
10117 Berlin

Campus- bzw. interne Geländeadresse:
Reinhardtstraße 58