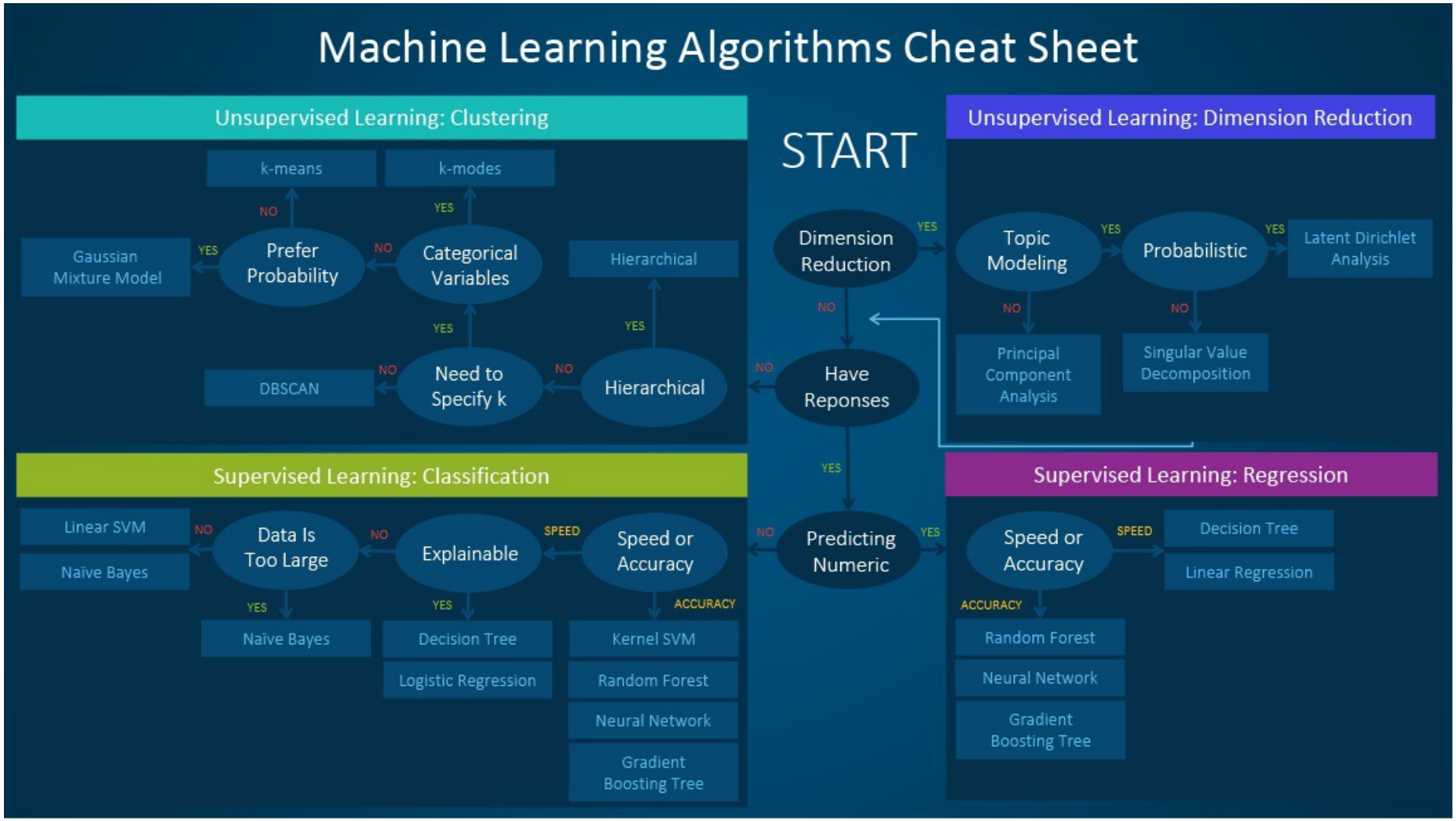
Machine Learning Algorithmen



DATA SCIENCE

MACHINE LEARNING		
Tag	Beschreibung	Beispiel
• **CRISP-DM** (immer im Kreislauf von den Arbeitsschritten vo 1.-8.) cross-industry-standard-process--for--data-minding	• **<u>Struktur von Data Science Projekten:</u>** •	1. Fragestellung erkennen 2. Business Understanding (Recherche in den Bereichen/Branchen): einzelne Problemstellungen 3. Data Understanding (Daten beziehen von...) 4. Austausch: Business Understanding - Data Understanding 5. Data Preparation (Daten vorbereiten, bei Defiziten z.B. Mittelwerte) 6. Modelling (auch in Rücksprache mit Data Preparation) 7. Evaluation (Daten werden ausgewertet, hohe Wahrscheinlichkeit für die Aussagekraft des Modells, z.B. 99%) auch in Rücksprache mit Business Understanding 8. Deployment (Einsatz/Ausführen des Modells) auch in Rücksprache mit Business Understanding
• Datensatz mit: • Features(Spalten, 1-3 oder mehr) • 1 Label (z.B. Spezies bei Tieren) • viele Samples (Zeilen, z.B. 1000)	• Wie **Algorithmen** lernen • Bsp.: Welche Spezies ist der Pinguin? • Wichtig: gute Verteillung von Daten, z.b. gleichviele Daten von den verschiedenen	• Man geht von einem Datensatz aus (.csv Datei) • ca. 70 – 98% davon wird als Trainingsdatensatz -**Training Data-** verwendet: --->Struktur finden, trainiertes Modell entwickelt aus den Features (Train_Test_Split (Datensatz splitten),Fitting(Datensatz um zu trainieren)) • ca. 30 – 2% Testdatensatz (=Rest vom Trainingsdatensatz) -**Validation Data-**: --->Struktur testen (Validierungs-Datensatz) • nach und nach Samples (Features ohne Labels) aus dem

	Labels, um eine gute Vorhersage zu generieren	Testdatensatz werden im trainierten Modell eingesetzt ---> Prediction (Vorhersage), also eines der Labels • nach und nach Samples (Labels) aus dem Testdatensatz wird im trainierten Modell -**Test Data**- eingesetzt----> korrektes Label ▪ ---> Fehler/Treffer Tabelle: ○ Prediction = <u>korrektes</u> Label (Ja/Nein) •
• • **Accuracy** • • **Confusion-Matrix**	• **Evaluieren** •	• Trefferquote über richtige Vorhersage • trefferanzahl / Gesamtanzahl (=accuracy) • • Abgleichen mit realen Werten (actual values) ○ Möglichkeiten: richtig (true) als Wert/Kategorie (Value/Klasse) und falsch(false) als Wert/Kategorie
• Zyklus zwischen Daten-Produkt-Kunden-Daten	• **Selbst lernendes Produkt** • wenn alle 3 Elemente steigen (Produkt-Kunden-Daten) profitieren alle untereinander	• Auto: selbstfahrend • Daten (wenn nicht vorhanden, dann manuelle verwenden oder z.B captcha[als Nutzer die Ampel in 9 Bildern finden,Text erkennen]) • Kunden (bezahlen für deren Daten zu sammeln) • Produkt (wenn nicht vorhanden, dann als gegeben betrachten, aber manuell bedienen) •
• Fundamentale Arbeit in der Data Science	• **Fragestellungen**	• 3 Typen: 1. Descriptive (klare, beschreibende Fragen) 2. Exploratory (offene, untersuchende Fragen) 3. Prediction (Vorhersagen, klar, aber auch offen)
• Algorithmen werden mit Labels trainiert	• **Supervised Learning** Voraussetzung: historische Daten mit hohen Warscheinlichkeiten für Zukunftsvorhersage	• Input, bei dem der Output klar ist (Output ist das Label, z.B. Hund oder Katze als Vorhersage, Spam: Ja/Nein) • Algorithmus lernt, indem er das Ergebnis mit dem Label vergleicht ---->Fehler wird minimiert, Modell wird modifiziert

	• Regression • Klassifizierung	• Vorhersage von kontinuierlichen Werten • Aussage JA / NEIN (Treffer oder nicht Treffer)
•	• **Unsupervised Learning**	• Keine vorhandenen Labels ◦ exploratory data analysis z.B. Geschäfte einer Stadt ohne Kenntnis über die Kategorie - Bar,Cafe,Friseur,etc. ◦ Markt segmentation (innerhalb eines business,z.B. Versicherungsbereiche) ◦ social network ◦ clustering businesses in a given area •
• Overfitting:(zu viele passende Traingsdaten vorhanden) • Underfitting (zu wenig passende Traingsdaten vorhanden)	• **Typissche Probleme bei Machine Learning**	◦ Nahezu perfekte Vorhersage (100%), Fehler bei Trainingsdaten nach validierung immer gering und bei einsatz in Testdaten groß ◦ ----> für Vorhersage eher z.B. eher über eine Gerade im einem Diagramm die Trainingsdaten interpretieren als über jeden einzelnen vorhandenenen Datenpunkt (Sample)