DataCampus

www.DataCampus.eu

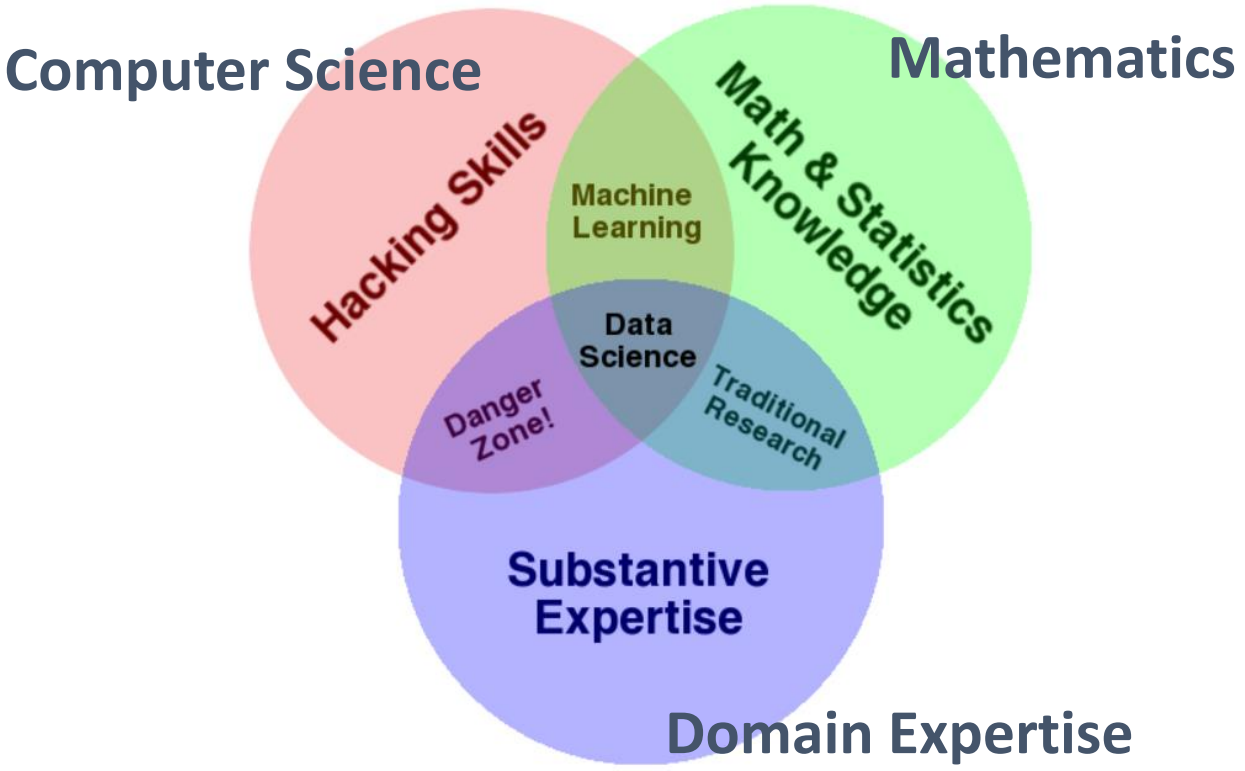
Data Science Lifecycle

Prof. Dr.-Ing. Marina Tropmann-Frick

6th & 7th KnowHow-Meeting (ONLINE) 26/27 Aug 2020 - in collaboration with ...

COMPUTERFUTURES

Motivation – What is Data Science?



Drew Conway
Data Science Venn Diagram

- Statistics
- Exploratory Data Analysis and Visualization
- High-Performance Computing technologies for dealing with Big Data
- Machine Learning
- Predictive Analytics
- Mixed Reality
- …
- Domain Knowledge

Motivation – Appreciating Data

- Daten für Informatiker – oft nur Material (*just stuff to run through a program*)

- Der übliche Weg, die Performanz der Algorithmen zu testen: “random data”

- Schaffen von sauberen und organisierten Umgebungen – der Fokus liegt auf Technologien, nicht auf Daten

➢ Unsere natürliche Umgebung ist oft kompliziert und unordentlich
➢ Fast nichts ist vollständig wahr oder falsch (1 oder 0)

- Naturwissenschaftler fühlen sich wohl bei dem Gedanken, dass Daten fehlerhaft sind; Informatiker nicht
- Interessante Datensätze sind eine knappe Ressource; verbunden mit harter Arbeit und Vorstellungskraft

Data Science Lifecycle

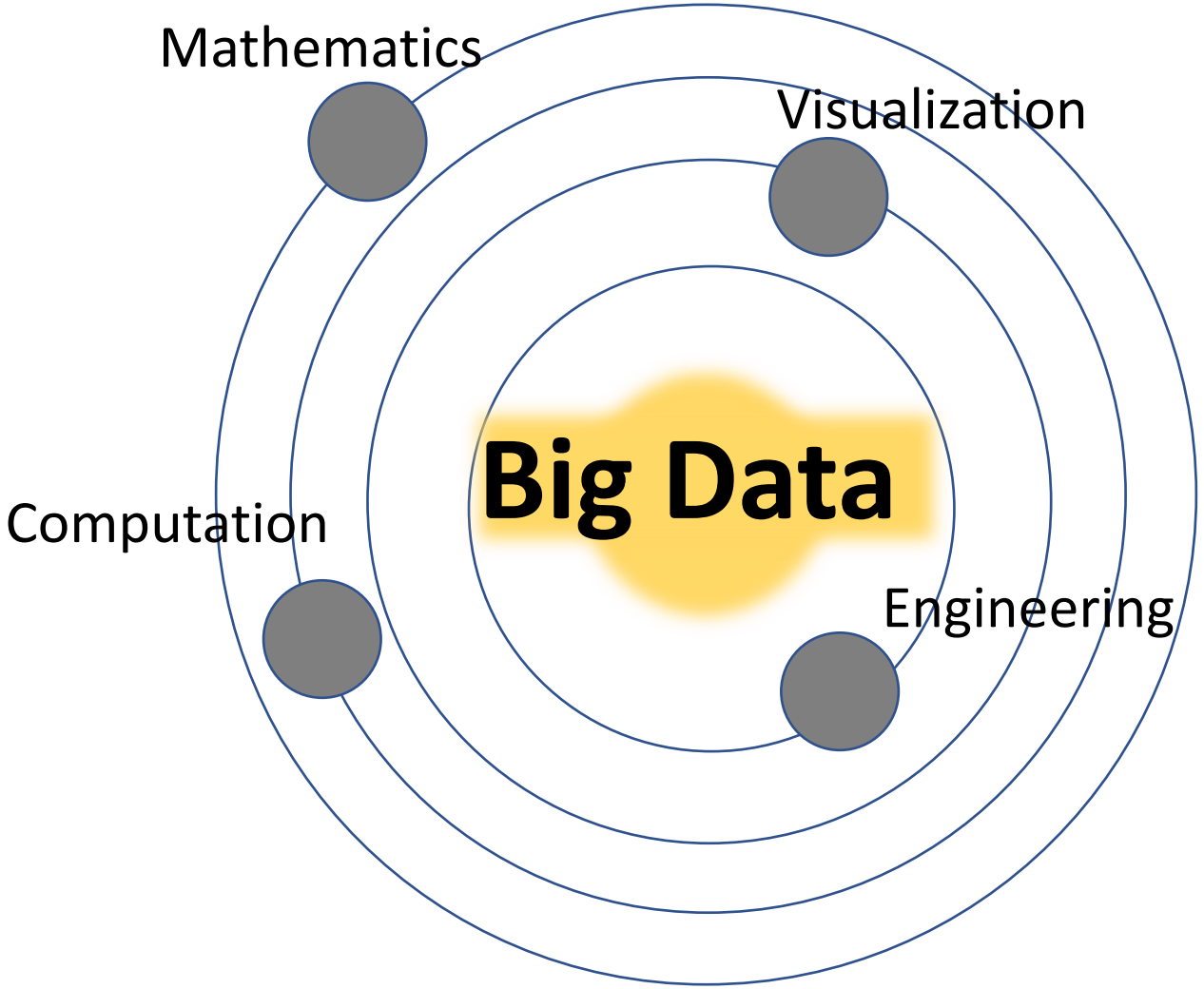
It is all about data:

- Wissen extrahieren
- Richtige Fragen stellen
- Antworten / Erklärungen finden
- Daten in Erkenntnisse verwandeln

Engineering/Computation

Effizientes:

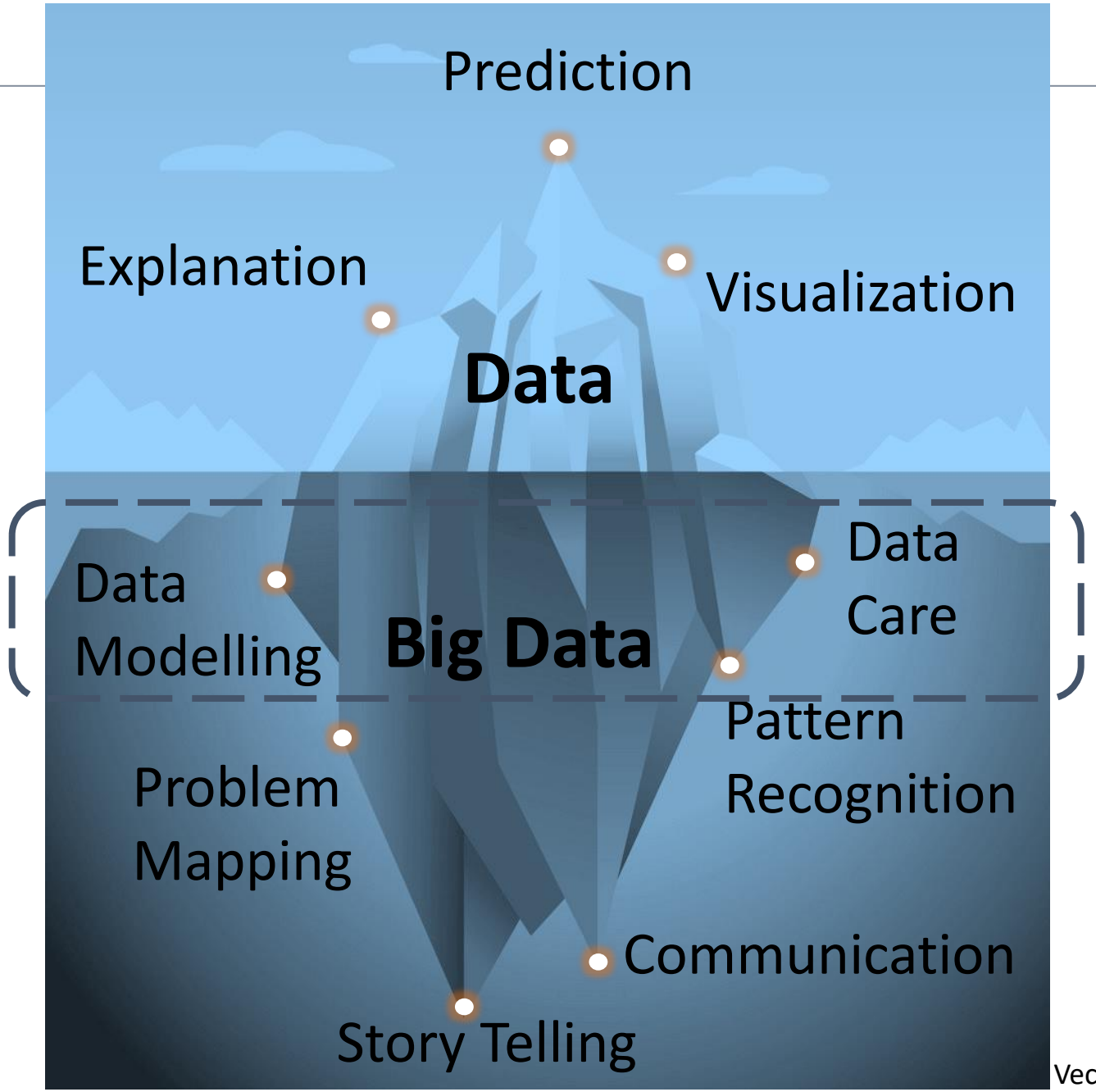
- Datenmanagement
- Datenverarbeitung



Data Science Lifecycle

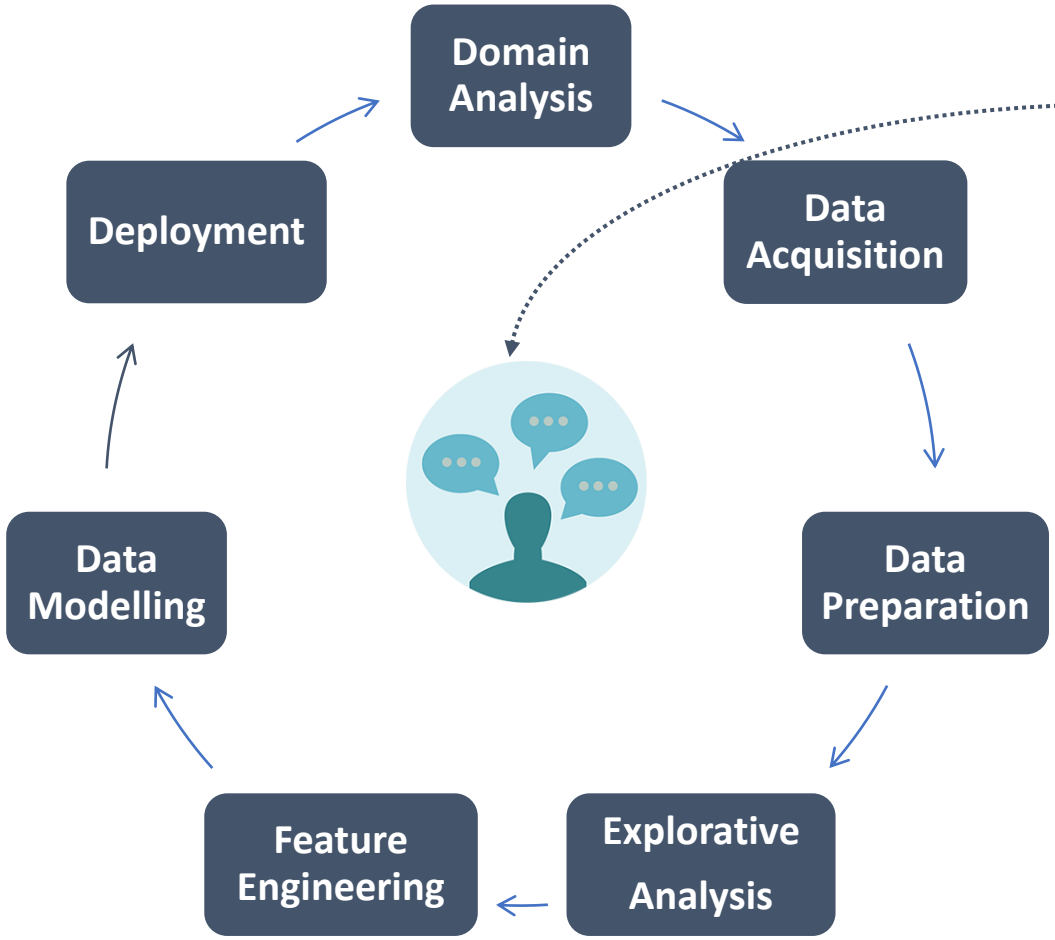
80% der Zeit

- finden
- bereinigen
- transformieren
- mergen / integrieren
- …



VectorStock.com

Data Science Lifecycle



Human in the Loop

7 stages of a data science lifecycle

Data Science Lifecycle

Domain Analysis

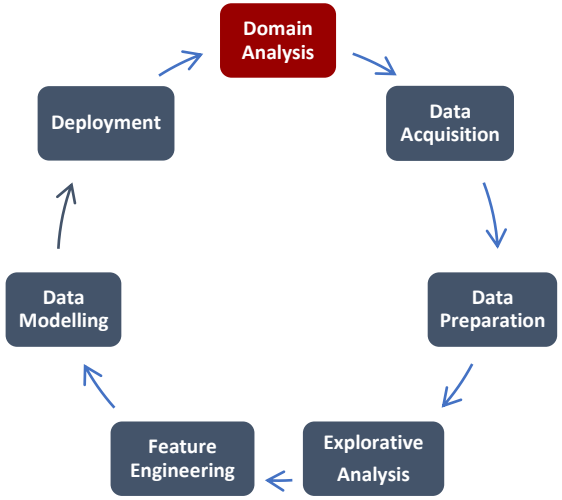
1. Mit Stakeholdern sprechen

2. Ziele des Projekts / der Analyse klären
   - Umsatzprognose?
   - Anomalie-Erkennung?
   - …

3. Das Problem verstehen
   - Expertenmeinungen erfragen
   - richtige Fragen stellen

4. Use Cases definieren

5. Entscheidungen dokumentieren

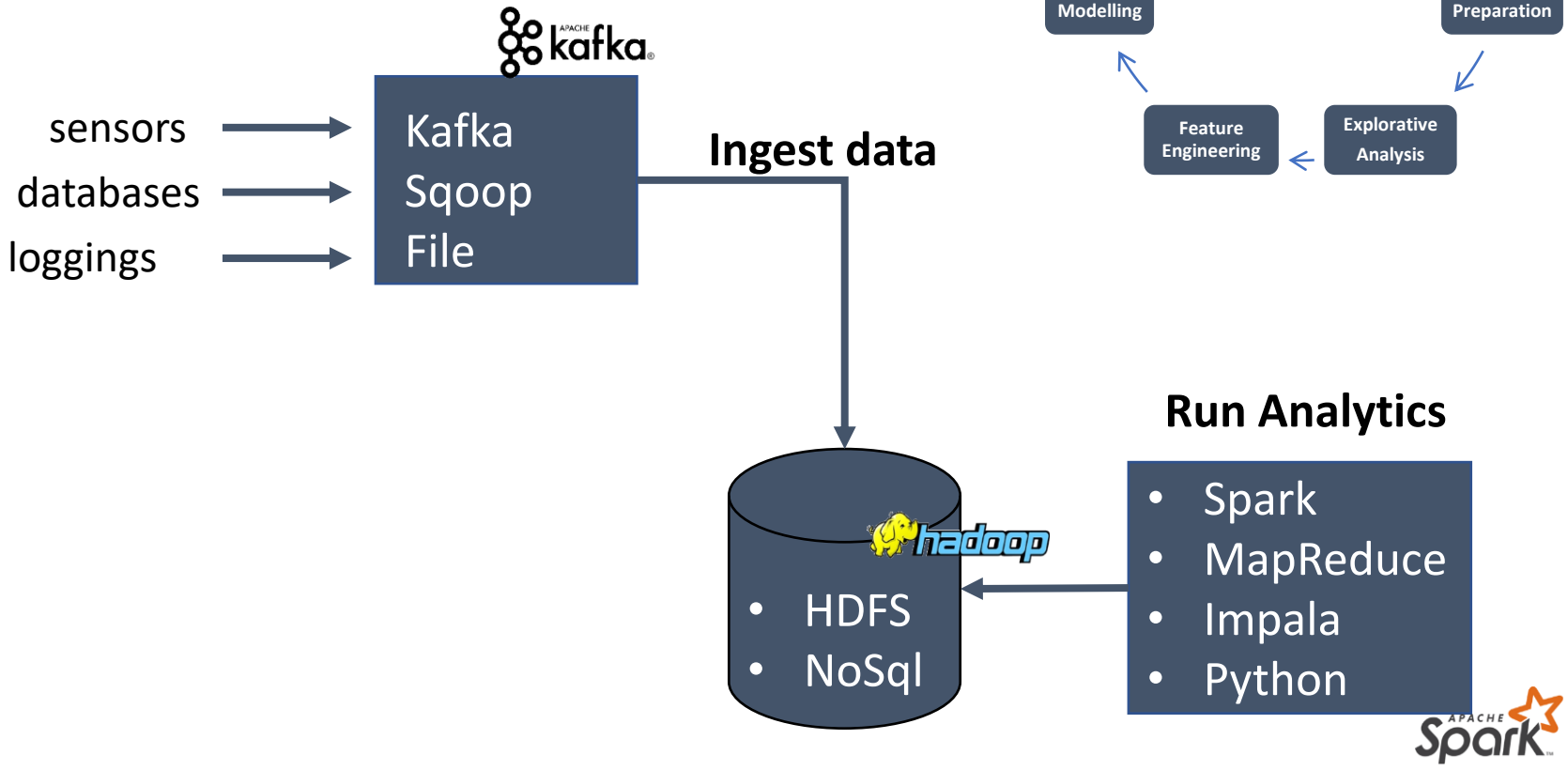


Data Science Lifecycle

Data Acquisition

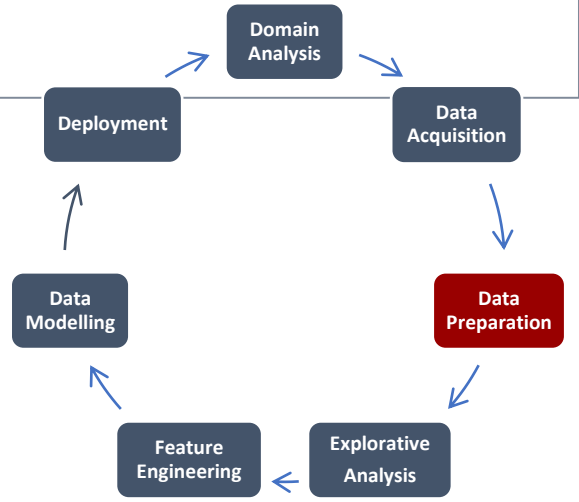
1. Datenquellen ermitteln
   - databases
   - sensors
   - social media
   - internet
   - loggings
2. Datenmanagement

3. Achten auf:
   - ausreichend Daten für die Modellierung?
   - Datenqualität?
   - Aktualität der Daten?



Architektur? Rechenpower? Speicheranforderungen?

Data Science Lifecycle



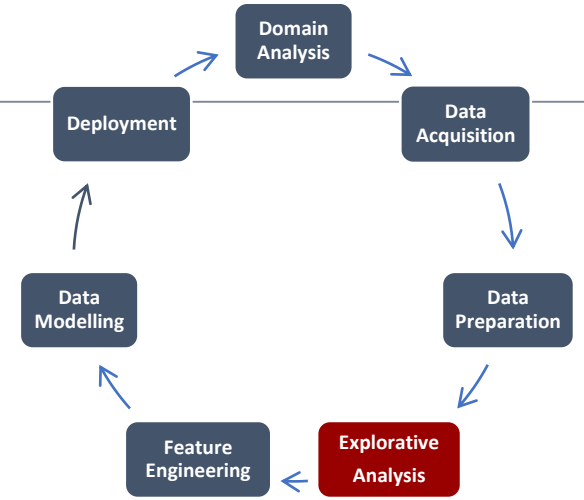
Data Preparation

1. Datenbereinigung (80 % der Zeit)
   - Fehlende Werte
   - Ausreißer
   - Unterschiedliche Formate
   - Errors
     - Groß-/Kleinschreibung
     - Abkürzungen
     - Rechtschreibfehler
   - Entfernen irrelevanter Daten
     - Formatierung von Artefakten
2. Datenfusion
   - Daten verknüpfen
   - Annotieren
3. Daten-Indizierung

In 1996:
Ariane 5 exploded after 40 s
after launch because of conversion of
64 bit float to 16 bit integer

In 1999:
NASA lost 125$ million Mars Climate Orbiter space mission due to a metric-to-English conversion

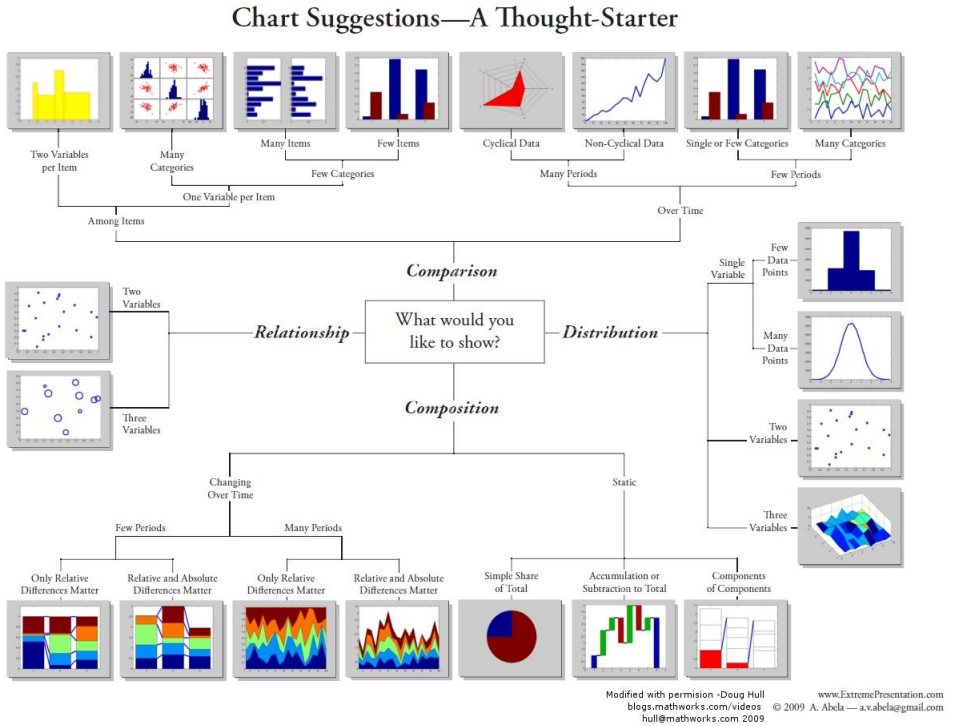
Data Science Lifecycle



Explorative Analysis

1. Tiefes Verständnis von Daten entwickeln
2. Visualisierung von Datenbeziehungen und Trends
3. Statistiken zusammenfassen und visualisieren
4. Erkennen von Mustern
5. Erkennen von Anomalien und Ausreißern
6. Unzureichende Datenbereinigung erkennen
7. Wichtige Merkmale hervorheben

Welches Diagramm?



Wertschätzung der Kunst / Darstellung – was sieht besser aus?

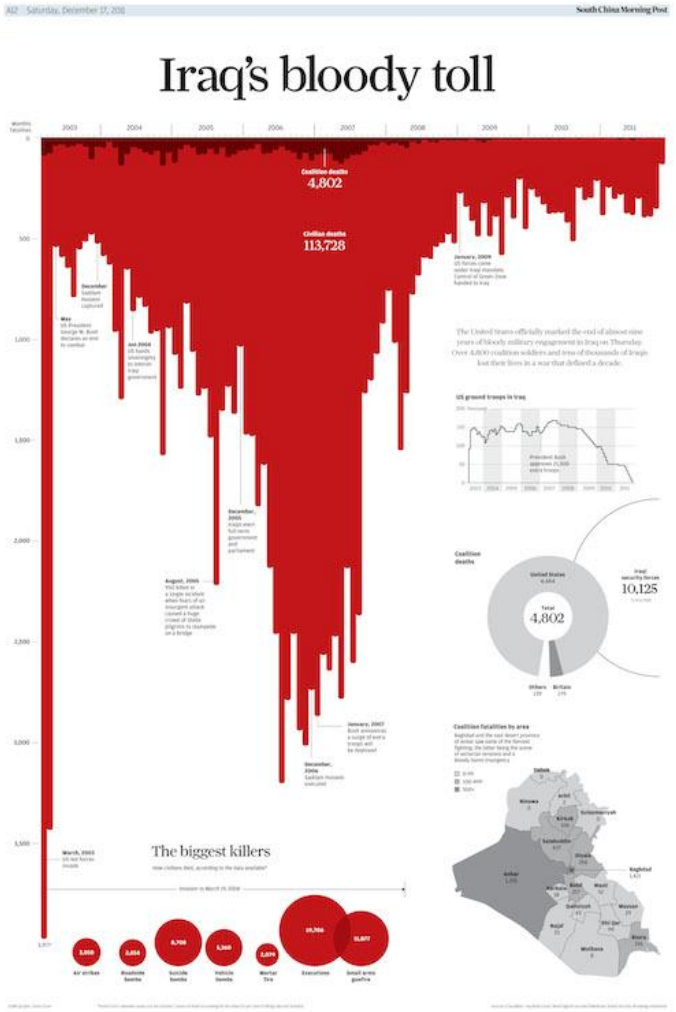
→ Entwicklung einer besonderen visuellen Ästhetik

images: Wikipedia

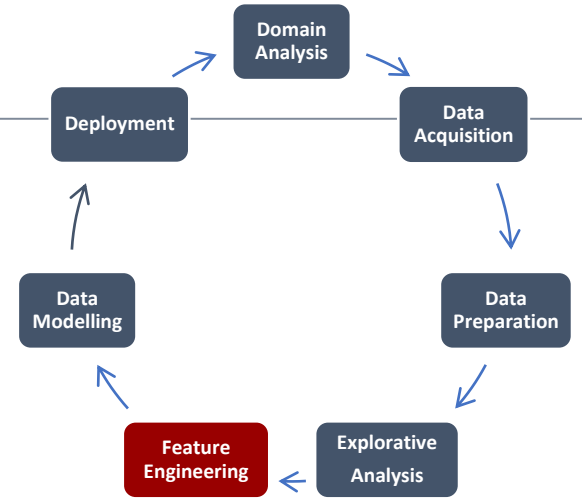
1854


50 0 50 100 150 200 Yards
× Pump • Deaths from cholera
OXFORD STREET
GT. MARLBOROUGH ST.
REGENT STREET
CONDUIT STREET
KING STREET
BROAD STREET
WARDOUR STREET
DEAN STREET
GOLDEN SQUARE
SAVILLE ROW
BREWER STREET
NEW BOND STREET
REGENTS QUADRANT
PICCADILLY

2011

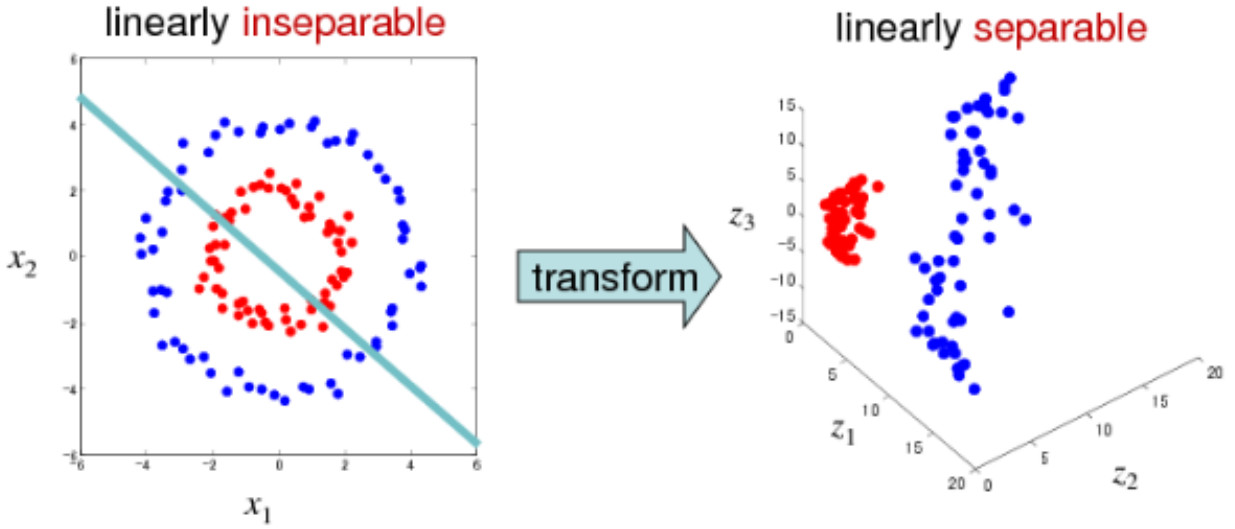

Iraq's bloody toll
4,802
113,728
The biggest killers
10,125

Data Science Lifecycle



Feature Engineering

1. Normalisierung
2. Korrektur fehlender Werte
3. Auswahl von Merkmalen (features)
   - relevante Merkmale und Zielvariablen bestimmen
   - Dimensionsanpassung (Reduktion)
     - irrelevante Merkmale entfernen
     - Regularisierung
4. Merkmalsanpassung
   - Domänenspezifische Features hinzugügen
   - Feature-Transformation
     - Summe, Differenz, Produkt, …

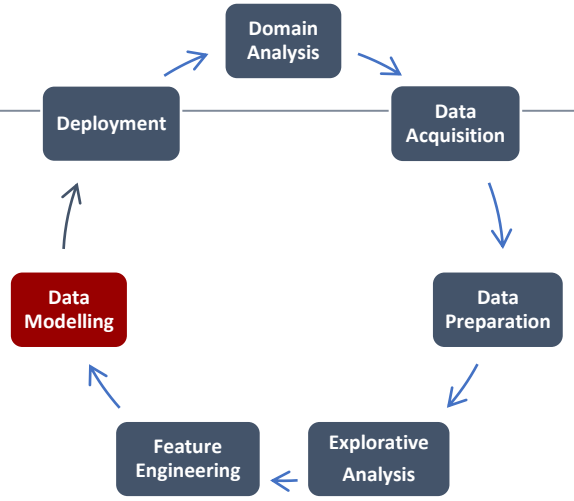


$$(z_1, z_2, z_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

http://yosinski.com/mlss12/MLSS-2012-Fukumizu-Kernel-Methods-for-Statistical-Learning/

Data Science Lifecycle

Data Modelling



1. Herangehensweise?
    1. Supervised
    2. Unsupervised
    3. Semi-supervised
    4. Self-supervised

2. Evaluation unterschiedlicher Modelle
    - geeignete Trainings- und Testsätze auswählen

3. Das “beste” Model für die Daten finden

4. Unterschiedliche Metriken vergleichen (accuracy, …)

- Computational analysis readily finds patterns and correlations in large data sets

    But when is a pattern significant?

“All models are wrong,

- George Box, 1976

…but some are useful”

Data Modelling

- Modellierung ist der Prozess der *Kodierung / Einkapselung* von Informationen zu einem Model, das für unterschiedliche Zwecke als Instrument benutzt werden kann

- Data Science: *building, fitting, and validating the model*

- The *"Occam's Razor"* philosophical principle: "the simplest explanation is the best"

- In Bezug auf die Modellierung (in Data Science) bedeutet dies oft eine Minimierung der Parameterzahl in einem Modell

→ First principle models: basieren auf der Grundlage einer theoretischen Erklärung der Funktionsweise des Systems

→ Data-driven models: basieren auf beobachteten Datenkorrelationen zwischen Eingabeparametern und Ergebnisvariablen

➔ Gute Modelle sind in der Regel eine Mischung aus beidem

Data Science Lifecycle

Deployment

1. Echtzeit- oder Batch-Analyse

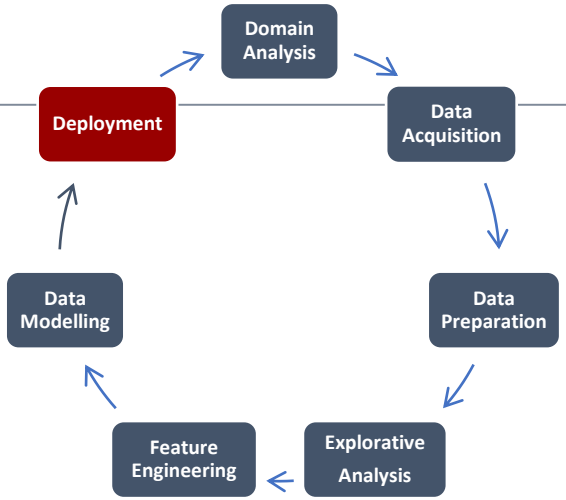
2. Technologieauswahl
   - Phython, R, …
   - Scala, TensorFlow, …
   - GPU - Beschleunigung?

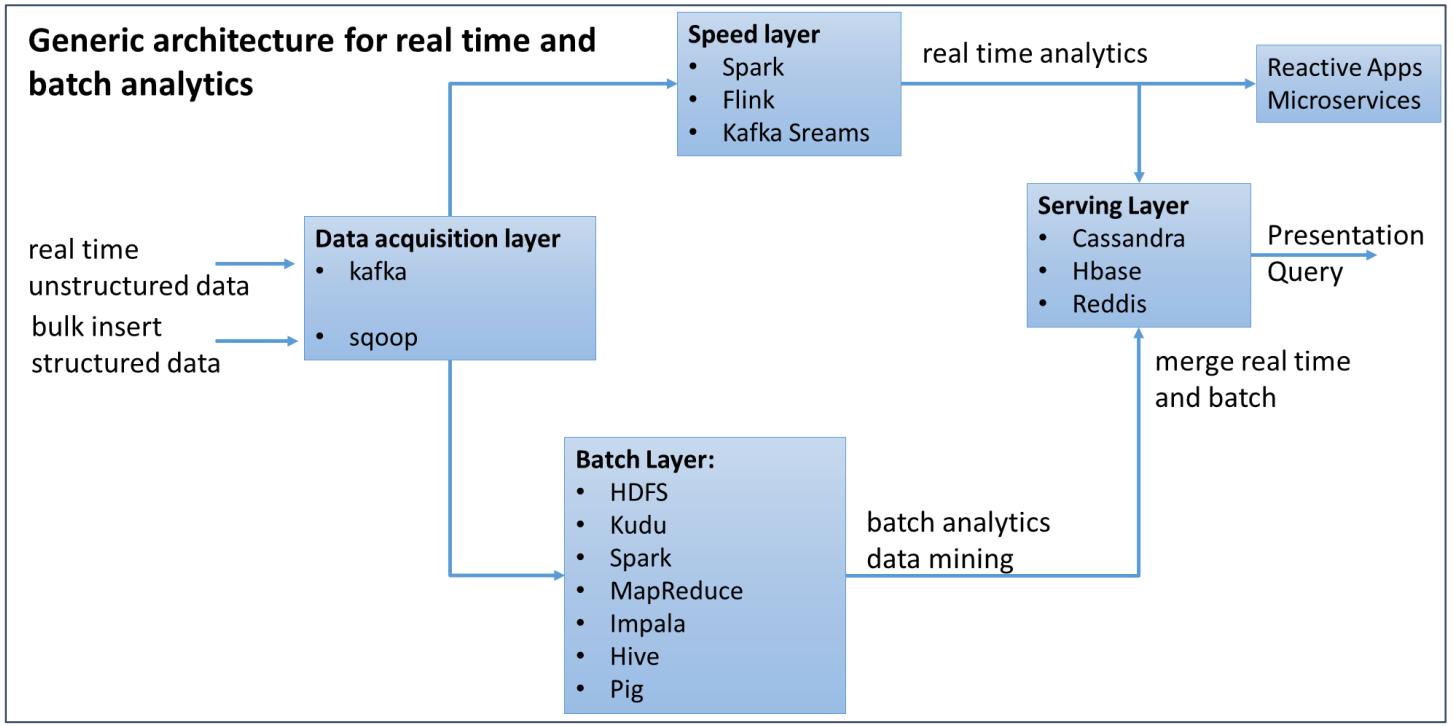
3. BigData?

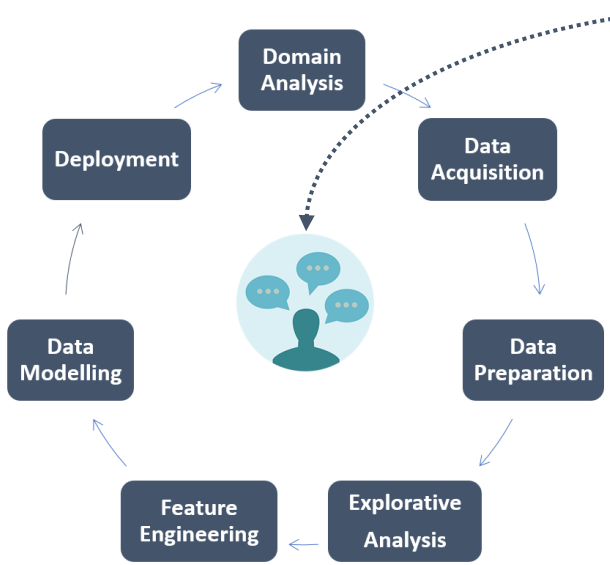
Z.B. high data frequency (sensors)

4. Service-Bereitstellung
   - Reporting, Visualisierung
   - Interaktive Exploration

5. Monitoring und Anpassung







Human in the Loop

Explainability

Explainability

- ✓ Nachvollziebarkeit
- ✓ Verständlichkeit
- ✓ Akzeptanz
- ✓ Legitimierung
- ✓ Vertrauen
- ✓ Sicherheit
- ✓ Ethik

Explainability

Komplexität der Problemstellungen ➔ undurchsichtige Modelle (Black-Box)

1. **Confidence**

Zuversicht in ein System wird geschaffen, indem es explizit zeigt, dass seine Entscheidung durch Erkennung von Mustern entsteht, die für den Nutzer nachvollziehbar sind und logisch erscheinen

2. **Trust**

3. **Safety**

4. **Ethics**

Xie, N., Ras, G., van Gerven, M., Doran, D.: Explainable deep learning: A field guide for the uninitiated (2020)

Explainability

Komplexität der Problemstellungen ➔ undurchsichtige Modelle (Black-Box)

1. **Confidence**
2. **Trust**
3. **Safety**
4. **Ethics**

Ein System wird als vertrauenswürdig eingestuft, wenn die Ausgaben vor ihrer tatsächlichen Anwendung nicht validiert werden müssen

Xie, N., Ras, G., van Gerven, M., Doran, D.: Explainable deep learning: A field guide for the uninitiated (2020)

Explainability

Komplexität der Problemstellungen → undurchsichtige Modelle (Black-Box)

1. **Confidence**
2. **Trust**
3. **Safety**
4. **Ethics**

i. konsistent
ii. keine negativen Auswirkungen
iii. hohe Verlässlichkeit auch unter außerordentlichen Bedingungen
iv. Rückmeldung bei neuen Einflussfaktoren

Xie, N., Ras, G., van Gerven, M., Doran, D.: Explainable deep learning: A field guide for the uninitiated (2020)

Explainability

Komplexität der Problemstellungen → undurchsichtige Modelle (Black-Box)

1. **Confidence**
2. **Trust**
3. **Safety**
4. **Ethics**

kulturelle / individuelle / rechtliche / moralische Unterschiede

Xie, N., Ras, G., van Gerven, M., Doran, D.: Explainable deep learning: A field guide for the uninitiated (2020)

Explainability

Ansätze

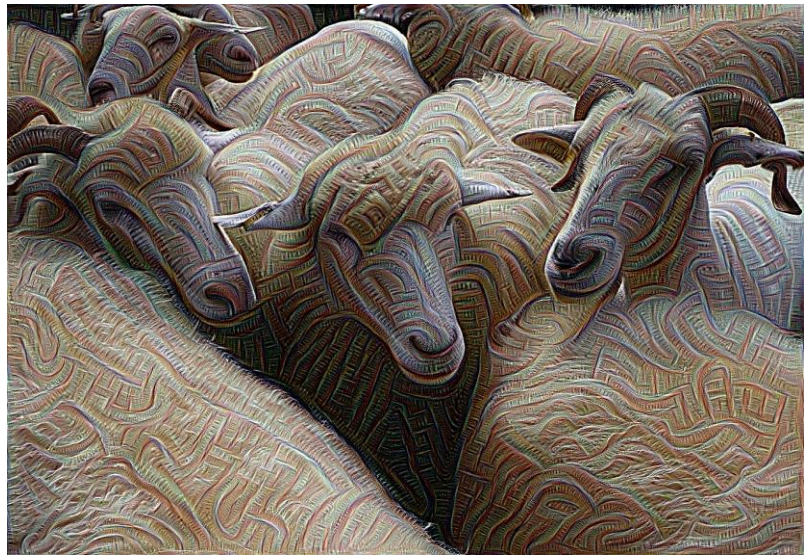
1. Visual Analytics
2. Model Distillation
3. Intrinsische Methoden

Explainability

Visual Analytics

- Interaktive Beteiligung des Nutzers / Experten in der Analysephase
- Untersuchung der Gradientensignale
- Variation des Inputs

1. Aktivierungsmaximierung



http://calhounpress.net

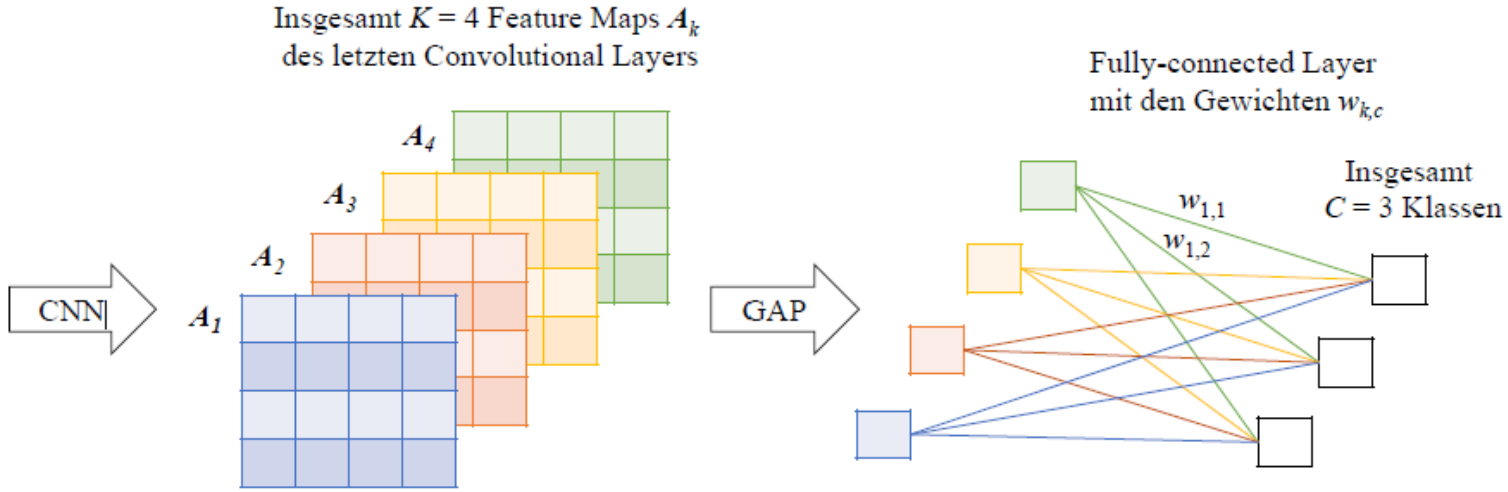
https://deepdreamgenerator.com/

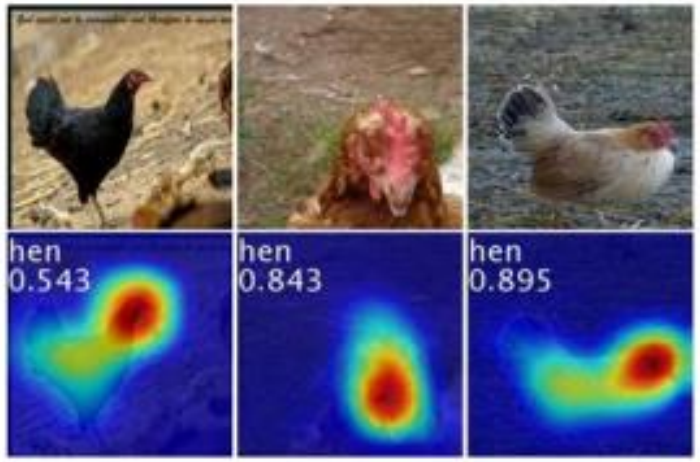
Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing higher-layer features of a deep network. Technical Report, Univeristé de Montréal (2009)

Explainability

Visual Analytics

2. Class Activation Maps





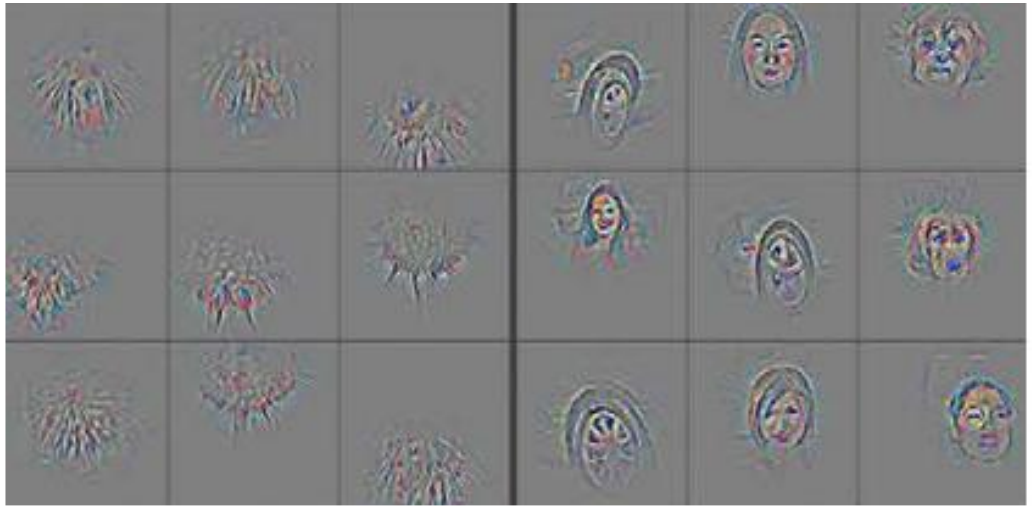
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization (2015)

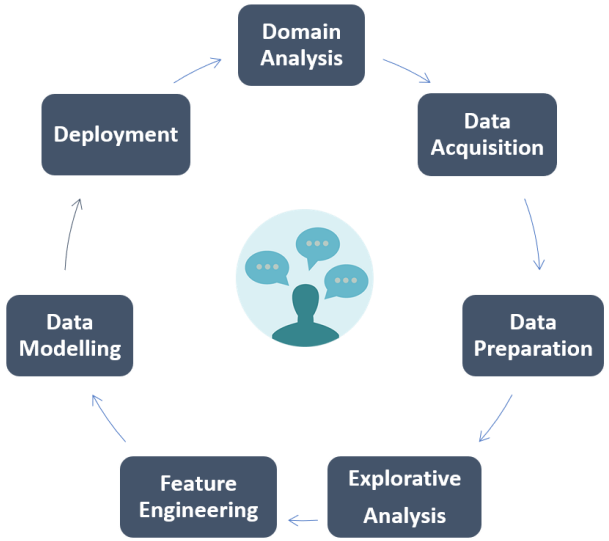
Explainability

Visual Analytics

Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks (2013)
33. Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning (2011)

3. Deconvolution





Privacy & Ethics

Privacy and Ethics

- Es ist wichtig, den potenziellen Schaden zu verstehen, den datengesteuerte Modelle verursachen können.
- *Correlation is not causation*, aber so trainierte Modelle können Aktionen und Feedback-Mechanismen auslösen, was zu sich selbst erfüllenden Prophezeiungen führt
- Jeder Data Scientist muss sich die Zeit nehmen, konstruktiv über gesellschaftliche Fragen / Folgen nachzudenken
- Totale Online-Privacy ist grundsätzlich unmöglich - aber wahrscheinlich auch nicht das, was angestrebt wird

Privacy vs. Security vs. Anonymity

- **Privacy – freedom from intrusion, observation or attention**
  - no one can see you
- **Security/Safety – freedom from danger or harm**
  - you are safe
- **Anonymity – freedom from identification or recognition**
  - nobody knows who you are

Ethics

Tradeoff → Kompromiss zwischen dem Nutzen des Teilens und dem Risiko der Offenlegung

- Menschenwürde
  - nicht auf eine Zahl zu reduzieren, nicht zu kaufen und zu verkaufen
- Vertrauen
  - wichtige Währung in der Informationswirtschaft

GI-Fachgruppe „Informatik und Ethik“ – https://gewissensbits.gi.de

Buch „Gewissensbisse – Ethische Probleme der Informatik. Biometrie – Datenschutz – geistiges Eigentum“. 2009.

DataCampus

www.DataCampus.eu

Vielen Dank!