Data Challenges

How to handle data?

Kim Hee, Todor Ivanov and Karsten Tolle

http://www.bigdata.uni-frankfurt.de/
Overall process ...

Start?

Market Research / Thinking

Idea!

How to handle: primary, secondary and 3rd data
How to store, process, analyse, visualize ...

Idea in theory

more Market Research / Thinking / Business Model

Implementation of App, Service, Application, Analysis ...

How to handle: primary, secondary and 3rd data
How to store, process, analyse, visualize ...

get data
APIs: Rest, Stream, Files, ...
formats: csv, xml, rdf, json, ...
examples for Open Data

clean, integrate data (ETL?)
... do you have a model? (might be done later)

store data
... if needed ...
types:
relational Databases
noSQL:
    Key-Value
    Wide-Column
    Graph
    Document
    ...
Hadoop (HDFS)

analyse/process data
languages:
Python, Java, Scala
platforms:
Spark, Flink, Hadoop Map-Reduce ...

visualize
Tableau, ...

tools:
Scrappy
Flume
Openrefine
Talend
...
Offenes Wissen
für die digitale Gesellschaft

Die Open Knowledge Foundation Deutschland ist ein gemeinnütziger Verein, der sich für offenes Wissen, offene Daten, Transparenz und Beteiligung einsetzt.

Hier kannst du mehr über unsere Organisation, unsere Ziele und unsere Projekte erfahren.

25 OK Labs
entwickeln Tools für digitale Städte in ganz Deutschland.

325 000 Kulturdaten
wurden bei Coding da Vinci für alle zugänglich und nutzbar gemacht.

11 000+ Anfragen
wurden bei FrgDenStaat.de von Bürgern an Behörden gestellt.

300 Jugend-Hacker
experimentieren bei Jugend hackt mit Code und Freundschaft.
### Datenauswahl

<table>
<thead>
<tr>
<th>Schnittstellen/Dateiformate</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>API</td>
<td><a href="https://codingdavinci.de/daten/">https://codingdavinci.de/daten/</a></td>
</tr>
<tr>
<td>CSV</td>
<td></td>
</tr>
<tr>
<td>JSON</td>
<td></td>
</tr>
<tr>
<td>XML</td>
<td></td>
</tr>
<tr>
<td>SOAP</td>
<td></td>
</tr>
<tr>
<td>PL</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
</tr>
<tr>
<td>Inline Static Web</td>
<td></td>
</tr>
<tr>
<td>PDF</td>
<td></td>
</tr>
<tr>
<td>PNG</td>
<td></td>
</tr>
<tr>
<td>OpenGLAM Collections</td>
<td></td>
</tr>
<tr>
<td>JPEG</td>
<td></td>
</tr>
<tr>
<td>JSON</td>
<td></td>
</tr>
<tr>
<td>XML</td>
<td></td>
</tr>
<tr>
<td>SOAP</td>
<td></td>
</tr>
<tr>
<td>PL</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
</tr>
<tr>
<td>Inline Static Web</td>
<td></td>
</tr>
<tr>
<td>PDF</td>
<td></td>
</tr>
<tr>
<td>PNG</td>
<td></td>
</tr>
<tr>
<td>OpenGLAM Collections</td>
<td></td>
</tr>
</tbody>
</table>

### GLAM Daten
- Liste mit schweizer Kulturdaten
- APIs & Tools für Bibliotheken
- Europäische Datenbanken
- Public Domain Werke
- Public Domain Review
- Flickr Commons
- OpenGLAM Collections

### Offene Daten Hamburg
- Diagramm der bundesweiten Regionalstatistik
- Offene Daten Köln
- Offene Daten Berlin
- OpenData Portal des Freistaats Bayern
- Portal für offene Daten der Hanseat Stadt Rostock
- Bundeshaus, Ministerium für Bildung und Kultur
- Portal Offene Daten Baden-Württemberg
- Statistik der Raumplanung in den Landkreisen
- Offene Daten der Stadt Bremen
- Aktuelle Luftdaten des Umweltbundesamtes
- Offene Daten des Umweltbundesamtes
- Dresden Lokalpunkt als Download
- Offenbad.de

### Andere Daten
- Free weather data API
- A comprehensive overview of open catalogues curated by experts from around the world
- DataCatalogue.org/dataset
Offene Daten Frankfurt

65 Datensätze gefunden

Parkdaten dynamisch
Die Publikation enthält dynamische Parkdaten der Parkhäuser der Stadt Frankfurt am Main. Das sind z.B. der Status des Parkhauses (offen oder geschlossen) und die jeweiligen...

Verkehrsmeldungen
Diese Publikation enthält Informationen zu den Baustellen und Veranstaltungen der Stadt Frankfurt am Main, die Einfluss auf den Verkehrsaufkommen haben.

Parkdaten statisch
Die Publikation enthält statistische Versorgungsdaten der Parkhäuser der Stadt Frankfurt am Main. Das sind z.B. der Name und der Ort des Parkhauses.

Datensätze suchen

Organisationen

1. Bürgeramt, Stadtrat (53)
2. Stadtverwaltungsamt (5)
3. Straßenverkehrsamt (3)
4. Umweltamt (1)
5. Standesamt (1)
6. Stadtammern (1)
7. Grundstücksamt (1)

Gruppen

1. Kommunalwahlen 2016 (17)
2. Bevölkerung (12)
3. Infrastruktur, Bau (11)
4. Wahllist 2016 (9)
5. Geographie, Geologie (8)
6. Wirtschaft und Arbeit (5)
7. Soziales (5)
OpenRefine – insbesondere zum bearbeiten einer Datei
gleiches Arbeiten mit verschiedenen Formaten: TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents
Support for other formats can be added with OpenRefine extensions.
Talend Open Studio
REST - Representational State Transfer

- Client-Server-communication
- Usage of standard HTML methods: Get, Put, Post, ...
How to handle: primary, secondary and 3rd data
How to store, process, analyse, visualize ...

- **get data**
  - APIs: Rest, Stream, Files, ...
  - formats: csv, xml, rdf, json, ...
  - examples for Open Data

- **clean, integrate data (ETL?)**
  - ... do you have a model? (might be done later)

- **store data**
  - ... if needed ...
  - types:
    - relational Databases
    - noSQL:
      - Key-Value
      - Wide-Column
      - Graph
      - Document
      - ...
  - Hadoop (HDFS)

- **analyse/process data**
  - languages: Python, Java, Scala
  - platforms: Spark, Flink, Hadoop Map-Reduce
  - ...

- **visualize**
  - Tableau, ...

**tools:**
- Scrapy
- Flume
- Openrefine
- Talend
- ...
Big Data Landscape 2016 (Version 3.0)

Infrastructure
- Hadoop
- Hadoop in the Cloud
- Spark
- Cluster Services
- Neo4j
- Cassandra
- MongoDB
- IBM Infosphere
- Red Hat

Analytics
- Data Integration
- Data Warehousing
- Business Intelligence
- Data Preparation
- Reporting
- Predictive Analytics
- Data Mining
- Machine Learning

Applications
- Sales & Marketing
- Customer Service
- Human Capital
- Legal

Cross-Infrastructure/Analytics
- Graph Databases
- MPP Databases
- Object Databases
- Data Integration
- Data Warehousing
- Business Intelligence
- Data Preparation
- Reporting
- Predictive Analytics
- Data Mining
- Machine Learning

Open Source
- Apache
- Google
- Microsoft
- IBM
- SAP
- NVIDIA
- VMware
- TIBCO
- TRM
- Oracle

Data Sources & APIs
- Framework
- Query/Data Flow
- Data Access
- Coordination
- Real-Time
- Statistical
- Machine Learning
- Search
- Security

Incubators & Schools
- JAWBONE
- Garmin
- MIT
- UC Berkeley
- Stanford
- Columbia
- Carnegie Mellon
- Northeastern

Last Updated 3/23/2016
© Matt Turk (mattturk), Jim Hao (jimhao), & FirstMark Capital (firstmarkcap)
Storage Systems

- Relational Databases Management Systems (RDMS)
  - MySQL, MariaDB

- NoSQL (Not-Only-SQL)
  - Examples on the next slides

- Apache Hadoop (Distributed file system)

- Other types of engines
  - In-Memory - VoltDB, Aerospike
  - Time series databases - InfluxDB
  - Search platforms (full text search) - Solr, ElasticSearch
NoSQL Systems Classification

- **Properties**
  - Non-relational
  - Open-Source
  - Schema-less (*schema-free*)
  - Optimized for distribution (clusters)
  - Tunable consistency, availability and partitioning

- **Data Model**
  - Key-value
  - (Wide-)Column
  - Document
  - Graph
NoSQL Landscape
Key-Value Stores

- Data model: key → value
- Interface: CRUD (Create, Read, Update, Delete Operations)
- Examples: Amazon Dynamo, Redis, Riak, RocksDB (Facebook)

Source: http://docs.aws.amazon.com/amazondynamodb/latest/developerguide/WorkingWithItems.html
Column Stores

- Data model: (rowkey, column, timestamp) → value
- Interface: CRUD (Create, Read, Update, Delete Operations), Scan
- Examples: Cassandra, Hbase, Google BigTable

Source: https://pandaforme.gitbooks.io/introduction-to-cassandra/content/understand_the_cassandra_data_model.html
Document Stores

- Data model: (key, collection) → document
- Interface: CRUD (Create, Read, Update, Delete Operations), Queries
- Examples: MongoDB, CouchDB, RethinkDB

Source: https://www.mongodb.com/json-and-bson
Graph Stores

- Data model: $G = (V, E)$: Graph-Property Modell
- Interface:Traversal algorithms, queries, transactions
- Examples: Neo4j, InfiniteGraph, OrientDB

Source: https://neo4j.com/developer/guide-data-modeling/
Hadoop

- **Hadoop** is an open-source framework for storing, processing, and analyzing massive amounts of distributed, unstructured data.
  - Distributed cluster system
  - Platform for massively scalable applications
  - Enables parallel data processing
  - Built in replication
  - Stores petabytes of unstructured data
  - **Move computation to data**
  - Automatically handles node failures
  - Written in Java (cross-platform portability)
  - Open Source Apache License
Big Data & Hadoop Ecosystem
Apache Spark is a fast and general engine for large-scale data processing.

Spark has several high-level tools, including:
- **MLlib** is Apache Spark's scalable machine learning library.
- **Spark Streaming** makes it easy to build scalable fault-tolerant streaming applications.
- **Spark SQL** is Apache Spark's module for working with structured data and enables HiveQL/SQL-like queries.
- **GraphX** is Apache Spark's API for graphs and graph-parallel computation.

Spark can be executed in two ways:
- Independent processes on a cluster
- As a YARN application

**Apache Flink** is an open source platform for distributed stream and batch data processing.

Flink includes **several APIs** for creating applications that use the Flink engine:
- **DataStream API** for unbounded streams embedded in Java and Scala, and
- **DataSet API** for static data embedded in Java, Scala, and Python,
- **Table API** with a SQL-like expression language embedded in Java and Scala.

Flink also bundles **libraries for domain-specific use cases**:
- **CEP**, a complex event processing library,
- **Machine Learning library**, and
- **Gelly**, a graph processing API and library.

More on [https://flink.apache.org/](https://flink.apache.org/)
Big Data Resource & More

• Many Online Courses on Big Data + Certificates - https://bigdatauniversity.com/courses/
• Coursera - Data Science courses
• edX - Data Analytics courses

• Hadoop & Spark Books
  – Hadoop Operations 2012
• Spark Examples - http://ampcamp.berkeley.edu/big-data-mini-course/
• Cloudera - Online resources - http://www.cloudera.com/training/library.html ; Quick VM
• HortonWorks tutorials - http://hortonworks.com/tutorials/ ; SandBox VM
• MongoDB videos - https://www.youtube.com/user/sergeymsg/playlists
• NoSQL Data Stores in Research and Practice - ICDE 2016 Tutorial - Extended Version
• A Little Riak Book - http://www.littleriakbook.com/
CRISP-DM
Cross Industry Standard Process for Data Mining

Data Mining Life Cycle

Business Understanding
- Identify project objectives

Data Understanding
- Collect and review data

Data Preparation
- Select and cleanse data

Modeling
- Manipulate data and draw conclusions

Evaluation
- Evaluate model and conclusions

Deployment
- Apply conclusions to business

Determine Business Objectives
- Background
- Business Objectives
- Business Success Criteria

Assess Situation
- Inventory of Resources, Requirements, Assumptions, and Constraints
- Risks and Contingencies
- Terminology
- Costs and Benefits

Determine Data Mining Goals
- Data Mining Goals
- Data Mining Success Criteria

Produce Project Plan
- Project Plan
- Initial Assessment of Tools and Techniques

Collect Initial Data
- Initial Data Collection Report
- (Log and Report Process)

Describe Data
- Data Description Report
- (Log and Report Process)

Explore Data
- Data Exploration Report
- (Log and Report Process)

Verify Data Quality
- Data Quality Report
- (Log and Report Process)

Data Set
- Data Set Description
- (Log and Report Process)

Select Data
- Reason for Inclusion/Exclusion
- (Log and Report Process)

Clean Data
- Data Cleaning Report
- (Log and Report Process)

Construct Data
- Derived Attributes
- Generated Records
- (Log and Report Process)

Integrate Data
- Merged Data
- (Log and Report Process)

Format Data
- Reformatted Data
- (Log and Report Process)

Select Modeling Technique
- Modeling Technique
- (Log and Report Process)

Generate Test Design
- Test Design
- (Log and Report Process)

Build Model Parameter Settings
- Models
- (Log and Report Process)

Assess Model
- Model Assessment
- Revised Parameter
- (Log and Report Process)

Evaluate Results
- Align Assessment of Data Mining Results with Business Success Criteria
- (Log and Report Process)

Plan Deployment
- Deployment Plan
- (Log and Report Process)

Plan Monitoring and Maintenance
- Monitoring and Maintenance Plan
- (Log and Report Process)

Produce Final Report
- Final Report
- (Log and Report Process)

Review Project
- Experience
- Documentation
- (Log and Report Process)
Python was originally a general purpose language. Python got dedicated library for data analysis and predictive modeling with strong community support over the years.

Python libraries
- numpy - numerical library
- scipy - advanced math, signal processing, optimization, statistics
- matplotlib - python plotting
- scikits-learn, MDP, mlpy, Orange, pandas, pybrain - a collection of learning algorithms, provide solutions to learning problems
- NLTK - Natural Language Toolkit
- OpenCV - library for image processing and computer vision
- tweepy - easy-to-use Python library for accessing the Twitter API
- (l) scrapy - open source web scraping framework

IPython notebook is an interactive shell the Python which also provides a browser based interface.

Anaconda is a Python distribution including over 100 of the most popular Python, R and Scala packages for data science.
KNIME and RapidMiner

- Visual programming environment (programming free) for building complete predictive analytic workflows
- Dragging and dropping icons onto a drawing window that represent steps of the analysis
Deployment

- **Django**
  - a Python web framework that encourages rapid development and clean, pragmatic design

- **GeoDjango**
  - a geospatial framework built on top of Django, that enables storage, manipulation and display of spatial data.

- **PhoneGap**
  - a framework that builds cross-platform mobile apps using web technology like HTML, CSS and Javascript.

GeoDjango example: [http://groundedinphilly.org/](http://groundedinphilly.org/)
Data visualization

- **BOOK AT HOME HOW TO VISUALIZE DATA**
- **TOOLS**
Appendix - Type of questions machine learning can answer

1. **Classification**: is this A or B?
2. **Regression**: how much or how many?
3. **Anomaly detection**: is this weird?
4. **Clustering**: how is this organized?

**Quiz**
- How many of XX smartphone 8 will sell over next 3 months?
- Discover market segments automatically and group customers into different market segments
- Is this tumor malignant or benign?
Appendix - Statistical Method for Data mining

classification

- SVC
- Ensemble Classifiers
- KNeighbors Classifier
- SGD Classifier
- Naive Bayes

Text Data
Linear SVC

<100K samples

predicating a category

>50 samples

no more data

regression

- SGD Regressor
- Lasso
- ElasticNet
- SVR(kernel='rbf')
- SVR(kernel='linear')

<100K samples

predicating a quantity

few features should be important

dimensionality reduction

- Randomized PCA
- Isomap
- Spectral Embedding
- LLE

<10K samples

just looking

tough luck

clustering

- Spectral Clustering
- GMM
- KMeans
- Minibatch KMeans
- MeanShift
- VBGMM

<10K samples

number of categories known

START

Appendix - Causation and Association

- The suicide rates comparison between people have taken psychotherapy and not.
- Association is not causation
- Causation can be identified by controlled experiments (comparison between treatment group and control group. Two groups are similar in all ways except a treatment. If there is a difference in the outcome, treatment is evidence of causality)
- Quiz
  - According to FBI statistics, 25% of home burglaries occur during the summer holiday seasons between Memorial day (May 30, 2016) and Labor day (Sep 5, 2016). Is it causation or association?
Fr., 25.11.2016, 18:00 Uhr – So., 27.11.2016, 18:00 Uhr

The goal is to develop banking solutions based on cognitive IBM Technology with voice and gesture control that better and easier access to simplify your daily financial business.

The special focus covers people with handicap to empower them doing their financial services like transactions, ledger queries, investment consulting and stock-exchange overviews without any difficulty.

https://www.eventbrite.de/e/meet-watson-the-platform-for-cognitive-businessibm-hackathon-frankfurt-tickets-27807524053
Appendix - 33 Brilliant and Free Data Sources – 1/2

- [http://data.gov](http://data.gov) The US Government pledged last year to make all government data available freely online. This site is the first stage and acts as a portal to all sorts of amazing information on everything from climate to crime.
- [Socrata](http://socrata.com) is another interesting place to explore government-related data, with some visualisation tools built-in.
- [Canada Open Data](http://open.canada.ca) is a pilot project with many government and geospatial datasets.
- [Datacatalogs.org](http://www.datacatalogs.org) offers open government data from US, EU, Canada, CKAN, and more.
- [https://www.healthdata.gov/](https://www.healthdata.gov/) 125 years of US healthcare data including claim-level Medicare data, epidemiology and population statistics.
- NHS [Health and Social Care Information Centre](http://www.hscic.gov.uk/home) Health data sets from the UK National Health Service.
- [UNICEF](http://www.unicef.org) offers statistics on the situation of women and children worldwide.
- [World Health Organization](http://www.who.int) offers world hunger, health, and disease statistics.
- [http://aws.amazon.com/datasets](http://aws.amazon.com/datasets) Huge resource of public data, including the 1000 Genome Project, an attempt to build the most comprehensive database of human genetic information and NASA’s database of satellite imagery of Earth.
- [https://developers.facebook.com/docs/graph-api](https://developers.facebook.com/docs/graph-api) Although much of the information on users' Facebook profile is private, a lot isn't – Facebook provide the Graph API as a way of querying the huge amount of information that its users are happy to share with the world (or can't hide because they don't know the privacy settings).
- [Face.com](http://www.face.com) A fascinating tool for facial recognition data.
- [UCLA](http://ucla.edu) makes some of the data from its courses public.
Appendix - 33 Brilliant and Free Data Sources – 2/2

- **Data Market** is a place to check out data related to economics, healthcare, food and agriculture, and the automotive industry.
- **Google Public data explorer** includes data from world development indicators, OECD, and human development indicators, mostly related to economics and the world.
- **Junar** is a data scraping service that also includes data feeds.
- **Buzzdata** is a social data sharing service that allows you to upload your own data and connect with others who are uploading their data.
- **http://www.gapminder.org/data/** Compilation of data from sources including the World Health Organization and World Bank covering economic, medical and social statistics from around the world.
- **http://www.google.com/trends/explore** Statistics on search volume (as a proportion of total search) for any given term, since 2004.
- **https://www.google.com/finance** 40 years’ worth of stock market data, updated in real time.
- **http://storage.googleapis.com/books/ngrams/books/datasetsv2.html** Google Books Ngrams. Search and analyze the full text of any of the millions of books digitised as part of the Google Books project.
- **http://www.ncdc.noaa.gov/data-access/quick-links#loc-clim** Huge collection of environmental, meteorological and climate data sets from the US National Climatic Data Center. The world’s largest archive of weather data.
- **http://wiki.dbpedia.org** Wikipedia is comprised of millions of pieces of data, structured and unstructured on every subject under the sun. DBPedia is an ambitious project to catalogue and create a public, freely distributable database allowing anyone to analyze this data.
- **http://www.freebase.com/** A community-compiled database of structured data about people, places and things, with over 45 million entries.
- **UCI Machine Learning Repository** is a dataset specifically pre-processed for machine learning.
- **Financial Data Finder at OSU** offers a large catalog of financial data sets.
- **Pew Research Center** offers its raw data from its fascinating research into American life.
- **The BROAD Institute** offers a number of cancer-related datasets.