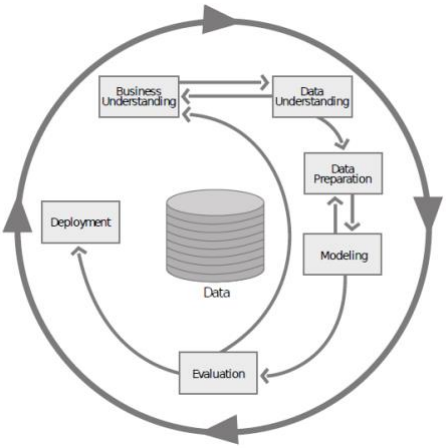
Summary Lectures DSS

1. Regular lecture: Methodology, Statistics and Pitfalls 1 2
Prediction 3
Inference and causality 4
Data quality 4
2. Regular lecture: Methodology, Statistics and Pitfalls 2 4
3. Regular lecture: SQL & Spark 5
SQL DDL 6
4. Regular lecture: Synthesis & Trends 8
5. Guest lecture: Menopause and cardiometabolic disease risk 9
Genetics 9
6. Guest lecture: Big spatial data 10
7. Guest lecture: Core Life Analytics 11
8. Guest lecture: Ethics and law 12
Anonymity 12
Consent 13
Research exemption 13
9. Guest lecture: Innovation in psychiatry 14
10. Guest lecture: Natural language processing in psychiatry 15

1. Regular lecture: Methodology, Statistics and Pitfalls 1

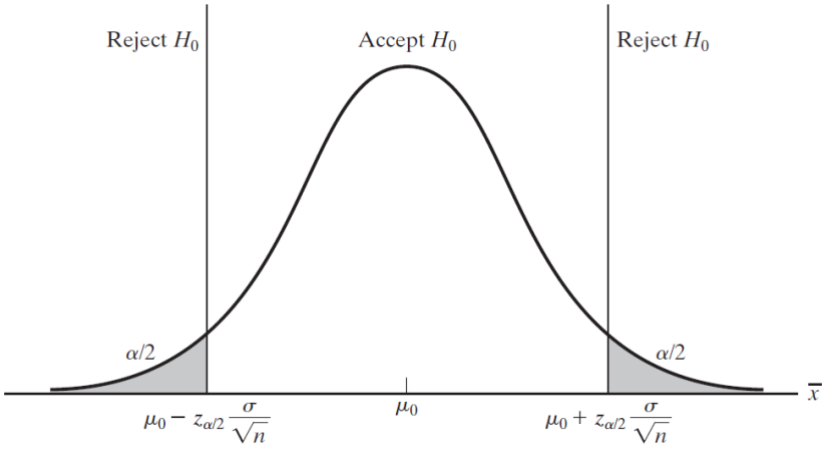
Structured methods help in preventing methodological errors, like CRISP-DM:



1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

Null hypothesis significance testing, traditional steps:

1. Formulate H_0 and H_a
2. Use sample distribution of the appropriate test statistics, determine critical region of size α
3. Determine value of the test statistics from the sample data
4. Check if the value falls in the critical region (reject H_0) or outside (retain H_a).



<- the grey is the 'critical' region.

α = significance level (p-value)

If α >= p, the null hypothesis must be rejected. Does not prove that the tested hypothesis is true. Guarantees that the Type 1 error (false positive) rate is at most α.

P-value:

- Frequentist: limiting relative frequency, if you could repeat the experiment.
- Bayesian: subjective, degree of belief, personally defined.

There is a problem with the traditional p <= 0.05, since the large number of n.

The problem with multiple testing: When pursuing multiple inferences, if you select only the significant ones for reporting, you can produce a greatly increased false positive rate. That is: the increasement of coming to a significant result by chance.

The Bonferroni correction can help mitigate the type 1 error, this by dividing the original α-value by the number of analyses on the dependent variable. However, it is vulnerable to Type 2 errors (failing to reject the null hypothesis when in fact you should).

Prediction

$$Y = f(X) + \epsilon$$

Where f is an unknown function of X, and ε is the random error, independent of X. You can estimate f in two ways:

- Prediction (black box)
- Inference (interest in associations, what is the type of relationship etc.)
    - There is a trade-off between prediction accuracy (black box) and model interpretability (inference).

In prediction, what do you predict against?

- The role of testing data
- How much is there to gain? Is the baseline (no prediction) as good as a prediction? Example: burglary, how high is the chance that someone breaks in your house?
- From 90% certainty to 100% certainty can be a long way (reducible vs irreducible error).
- How to quantify prediction quality? (model accuracy)
- Dynamic prediction (feedback loops, fraud prediction and Netflix challenge) → training data becomes outdated when you put your algorithm in production, so does your algorithm. Solution: inference instead of black box.

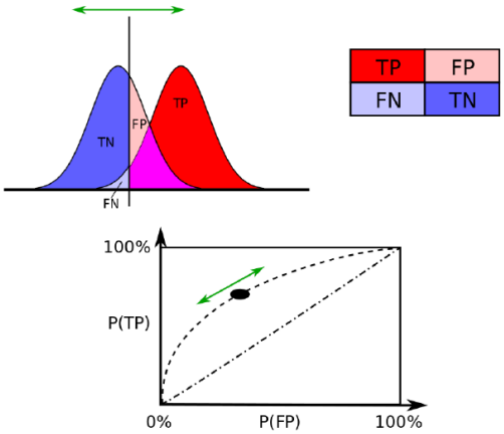
Validation set approach:
Split your data set in a training set (to fit the model) and a validation set (to evaluate the model). If you only use one subset, the test error can be variable, different cross-validation approaches are possible to mitigate this (leave-one-out, k-fold, bootstrap).

Measuring the quality of fit:

- Mean squared error: $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$ ← for continuous variables
- Error rate: $\frac{1}{n}\sum_{i=1}^{n}(y_i \neq \hat{y}_i)$ ← for categories
- There are many more, especially in classification

		True class: Positive	True class: Negative	Measures
Predicted class	Positive	True positive *TP*	False positive *FP*	Positive predictive value (PPV) $\frac{TP}{TP+FP}$
Predicted class	Negative	False negative *FN*	True negative *TN*	Negative predictive value (NPV) $\frac{TN}{FN+TN}$
Measures		Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{FP+TN}$	Accuracy $\frac{TP+TN}{TP+FP+FN+TN}$

← Confusion matrix



ROC analysis →

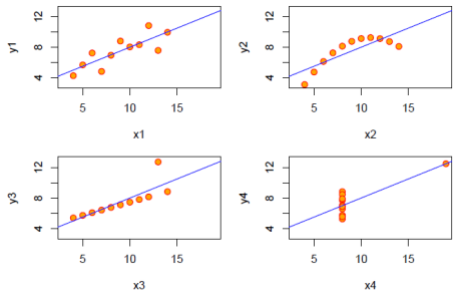
Inference and causality

There is an important distinction in exploratory vs. confirmatory analysis in statistics. It is important not to confuse exploratory with explanatory. Related concepts are:

- Spurious relations
    - A correlation in exploratory analysis should be tested causally in explanatory, otherwise it is a spurious relation.
- Coincidence
- Causal relationships (or lack thereof)
- Fishing

So, beware of causal statements, pitfalls include:

- Simpson's paradox
    - When a trend appears from one variable or group but disappears when multiple variables or groups are combined. Or when another variable is added into account.
- 3rd variable
- Anscombe's quartet



    - 4 data sets, all with the same summary data, but different patterns.
- Lack of theory
- Lack of experiment

Data quality

- Garbage-in garbage-out principle
- Examples of measures: completeness, validity, accuracy, consistency, availability and timeliness.
- Beware of missing data:
    - What deleting or disregarding data can do to your research
    - But what to do then?
        - Explicit assumptions
        - Model when needed

2. Regular lecture: Methodology, Statistics and Pitfalls 2

Research on whether search or social media can predict x has become commonplace and is often put in sharp contrast with traditional methods and hypotheses.

Big Data Hubris

Quantity of data does not mean that one can ignore foundational issues of measurement, construct validity and reliability.

- Overfitting
    - Simply put a lot of variables against the flu instead of hypothesis testing.
- Flu vs winter detector
- Specific model misfit
    - To prevent: inference instead of blind modelling.
- CDC data does better
- Dynamically recalibration needed

Algorithm dynamics
Dynamics and pitfalls: Search algorithm itself changes

- Complexity due to changes in algorithm
- Replication problems (even using Google Correlate)
- Changes on commercial aspects and suggesting searches using trends

Transparency, granularity, and all-data
The GFT parable is important as a case study where we can learn critical lessons as we move forward in the age of big data analysis:

- Transparency and replicability
- Use big data to understand the unknown
- Study the algorithm
- It's not just about size of the data

> *Instead, traditional "small data" often offer information that is not contained (or containable) in big data […]. Instead of focusing on a "big data revolution," perhaps it is time we were focused on an "all data revolution," where we recognize that the critical change in the world has been innovative analytics, using data from all traditional and new sources, and providing a deeper, clearer understanding of our world.*

Reply by Broniatowski, Paul, and Dredze:
Many limitations overcome:

- By including Twitter data (namely: replicability, overfitting, construct validity, granularity, and temporal confounds)
- Separation of awareness and infection on Twitter
- Now we are doing great (great correlation, externally validated)
- Do not generalize GFT to big data analyses.

Reply by Lazer et al.,:

- Not all problems occur in the entire field
- Problems with data quality (changes by all sorts of parties) remains
- Still related to Zombies
- Build strong collaboration to *learn* from such data

3. Regular lecture: SQL & Spark

NoSQL has 6 key features:

1. The ability to horizontally scale "simple operation" throughput over many servers
2. The ability to replicate and to distribute data over many servers
3. A simple call level interface or protocol
4. A weaker concurrency model than the ACID transactions of most relational database systems
5. Efficient use of distributed indexes and RAM for data storage
6. The ability to dynamically add new attributes to data records

- A key feature of NoSQL systems is "shared nothing" horizontal scaling → replicating and portioning data over many servers.
- NoSQL has no ACID transactional properties: updates are eventually propagated, but there are limited guarantees on the consistency of reads. NoSQL is more BASE, not ACID:
    - BASE = Basically Available, Soft state, Eventually consistent
    - ACID = Atomicity, Consistency, Isolation, and Durability
- The idea of giving up ACID constraints, one can achieve higher performance and scalability.

The arguments for SQL (relational):

1. New relational systems can do everything that a NoSQL system
2. Relational DBMSs have taken and retained majority of the market share
3. Successful relational DBMSs have been built to handle other specific application loads in the past.
4. There is no 'one size fits all', but there is a common interface with SQL, that gives advantages in training, continuity, and data interchange.

The arguments for NoSQL:

1. There are no benchmarks showing that RDBMSs can achieve scaling comparable with NoSQL
2. If you only require a lookup of objects based on a single key, a key-value store is adequate and easier to understand than a relational DBMS.
3. Some applications require a flexible schema, allowing each object in a collection to have different attributes.
4. A relational DBMS makes multimode multi-table operations too easy. You don't want this to happen, NoSQL systems make them impossible.
5. Other products than RDBMSs have established markets in areas where there is a need for capabilities.

SQL DDL

Data definition language

Spark implements subset of SQL:2003.

SQL is no 'classical' programming language, it is more a 'data sub language'. Core parts of SQL:

- Data definition language (DDL): used to define database structures
- Data manipulation language (DML): define, update, and request data (queries)

SQL statements:

```
–CREATE TABLE EMP_SKILL(
    EmpID       Integer     NOT NULL,
    SkillID     Integer     NOT NULL,
    SkillLevel  Integer     NULL,
    CONSTRAINT  EmpSkill_PK PRIMARY KEY (EmpID, SkillID),
    CONSTRAINT  Emp_FK      FOREIGN KEY(EmpID)
                            REFERENCES EMPLOYEE(EmpID),
    CONSTRAINT  Skill_FK    FOREIGN KEY(SkillID)
                            REFERENCES SKILL(SkillID)
    );
```

- CREATE TABLE
  - Creation of tables
- PRIMARY KEY
  - Composite primary keys within table creation
- FOREIGN KEY
  - Composite foreign keys within table creation
- DROP
  - Deletion of unwanted database objects
  - Warning, DROP permanently deletes the data objects along with the data related to that object
- INSERT
  - Add a row in a table
  - Non-numerical data needs to be within single quotes: 'test'
- UPDATE
  - Modify values in an existing row
- DELETE
  - Delete a row or a set of rows
- SELECT
  - Is a query, retrieves data from one or more tables and creates a new (temporary) table.
  - Use an asterisk (*) to show all values in a table that match certain criteria

- DISTINCT
  - Add to SELECT to prevent that duplicate rows are shown
- WHERE
  - Provide criteria which the shown records should meet
  - May include:
    - Equals “=”
    - Not equals “<>”
    - Greater than “>”
    - Less than “<”
    - Greater than or equal to “>=”
    - Less than or equal to “<=”
  - AND- OR-operators
    - Multiple matching criteria can be specified
  - IN
    - Make sure that certain column values are included in the query result
  - Logical NOT-operator
    - Used to prevent records from being part of the resulting set of records
  - BETWEEN
    - Show records where the values in a column are between a minimum and maximum value
  - LIKE
    - Partially complete values can be searched for
    - Use wildcards to find records
      - % for multiple arbitrary characters
      - _ for a single arbitrary character
- ORDER BY
  - Query results are sorted with ORDER BY
- Built-in SQL functions:
  - COUNT: counts the number of rows that match the criteria as posed in the query
  - MIN: search for the minimum value in a certain column
  - MAX: search for the maximum value in a certain column
  - SUM: computes the sum of values in a certain column
  - AVG: computes the average of the values in a certain column
- GROUP BY
  - Sub totals are computed with the GROUP BY clause
  - HAVING
    - Used to limit how much data is shown

Data from multiple tables:

- Subqueries
  - The result of a query is a subset of the data which can be used as input for another query. This is called a subquery
- Joins
  - A different way to combine data can be realized with a join
  - Move the JOIN syntax to the FROM clause with ON

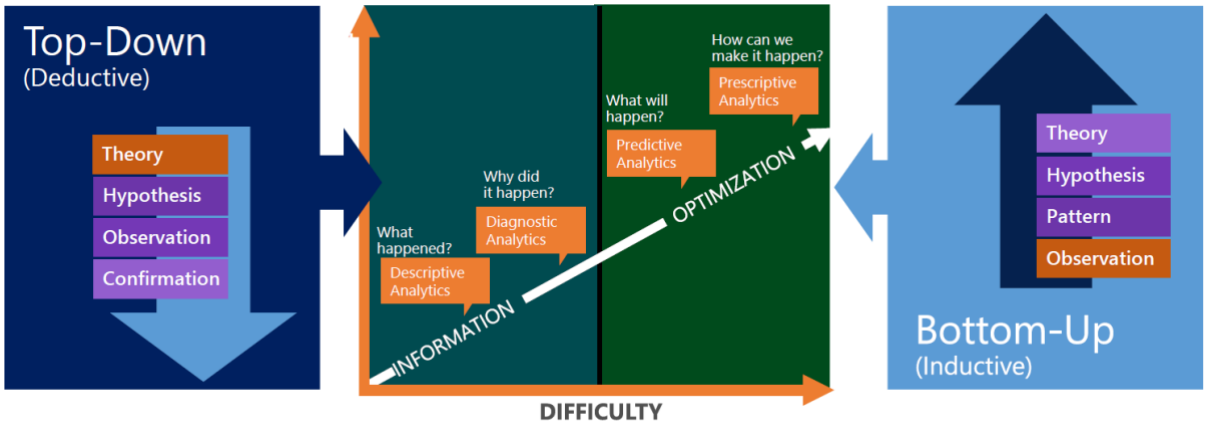
SQL in Spark

- First, it was Apache Hive
- Now users use Spark SQL, it is still compatible with Hive, and it is interoperable with DataFrames.

4. Regular lecture: Synthesis & Trends

The V's:

- Volume
- Velocity
- Variety
- Veracity
- Variability
- Value
- Volatility



- Hot path: real time, predictive model
- Cold path: historical, historical data science

AlphaGo Zero:

- Computer program AlphaGo beat world champions, it relied largely on supervised learning from millions of human expert moves.
- AlphaGo Zero is purely based on reinforcement learning and learn solely from self-play. Starting from random moves, it reached superhuman level in just a couple of days, beating all previous versions of AlphaGo.
- Because the machine independently discovers the same fundamental principles of the game that took humans millennia to conceptualize, the work suggests that such principles have some universal character, beyond human bias.

Trending in 2017-2018:

1. The success of AlphaGo
2. Deep learning mania
3. Self-driving cars
4. TensorFlow's influence on the commoditization of neural network technology

Expert opinions:

- Self-_ & scalability
  - Deep reinforcement learning → new approach?
  - Self-play is old idea in ML
  - AI support in the cloud → scalability
- Deployability
  - Self-driving cars and virtual assistants
  - AI increasingly for competitive advantage
  - Shortage of data scientists who know AI/DL
  - Meta-learning

- Applicability
    - Machine-transcription of telephone conversations
    - User privacy in deep learning applications
    - Buzzword: Artificial General Intelligence (AGI) vs AI
- Explainability
    - Ethics, accountability, and explainability
    - Transparency
    - Explainable AI as an emerging discipline
- Manageability
    - Increased developer usability
    - Docker for data science
    - Governance in data science

5. Guest lecture: Menopause and cardiometabolic disease risk

Cardiometabolic diseases are a combination of different diseases:

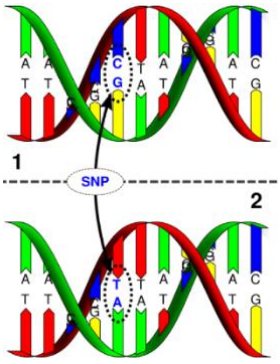
- Cardiovascular disease
    - Class of diseases that involve the heart or blood vessels, stroke, heart failure etc.
    - Leading cause of death in the world and underlying mechanisms are unclear
- Type 2 diabetes
- Common risk factors are:
    - Obesity
    - High tiglyceride levels
    - Low HDL cholesterol (good cholesterol)
    - Elevated blood pressure

Menopause is when a woman has her final menstrual period. From menopause, there is a change in hormones, one of the results is higher FSH. Earlier menopause = higher risk of cardiovascular disease (CVD) mortality. Smoking accelerates menopause. Surgical menopause increases risk for CVD.

Concluding:

- Menopause is associated with higher cardiometabolic disease risk
- Causality is unclear
- Mechanism is unclear
- Anti-Müllerian hormone may constitute a new possible mechanism

Genetics



During DNA replication → mutations occur. Single Nucleotide Polymorphism (SNP) is the point of mutation. Some mutations (SNPs) lead to diseases. You have common SNPs and rare SNPs. Two projects:

- Hapmap project:
    - To build genome-wide inventory of 3M human SNPs
- 1000 genome project
    - Sequencing to identify nearly all variants

Understanding the genetics of a disease may help understanding mechanisms causing that disease, which may lead to new therapies.

Genome-wide association studies identify genetic variants associated with diseases. There are a lot of SNPs in genetics, therefore big data analysis is needed to find these associations.

6. Guest lecture: Big spatial data

There are different ways to describe reality:

- Thematic description (what)
- Geometric description (where)
- Temporal description (when)

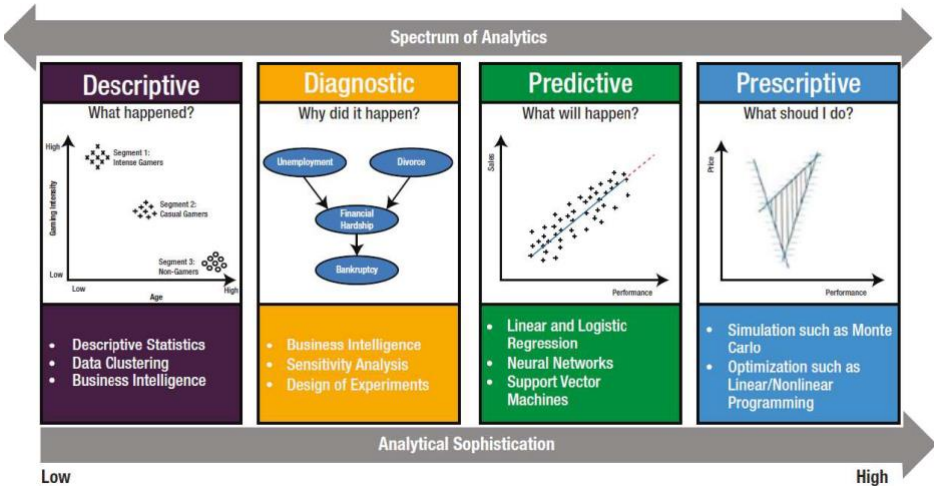
How GIS works:

- Data is stored as a collection of thematic layers
- Two data-models
    - Two components:
        - Geometric data
        - Attribute data
    - Two representation models:
        - Vector data model, for discrete objects (trees, railroad tracks)
        - Raster data model, for continuous phenomena (temperature, elevation)

Data to action process:

1. Data management and integration
2. Visualization and mapping
3. Analysis and modeling
4. Planning and design
5. Decision making
6. Action

A geographic information systems (GIS) lets us: 1) visualize, 2) question, 3) analyze, and 4) interpret our data to understand relationships, patterns, and trends to obtain location intelligence.



Statistical tools in ArcGIS:

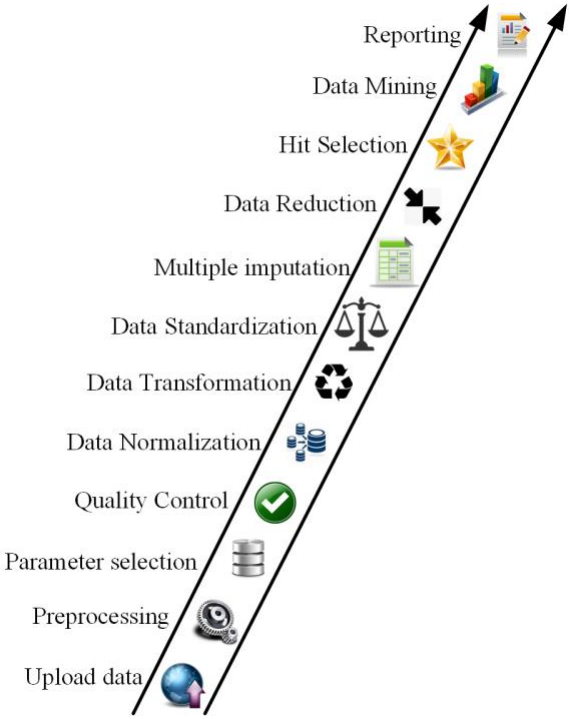
- Classification
- Clustering
- Prediction

7. Guest lecture: Core Life Analytics

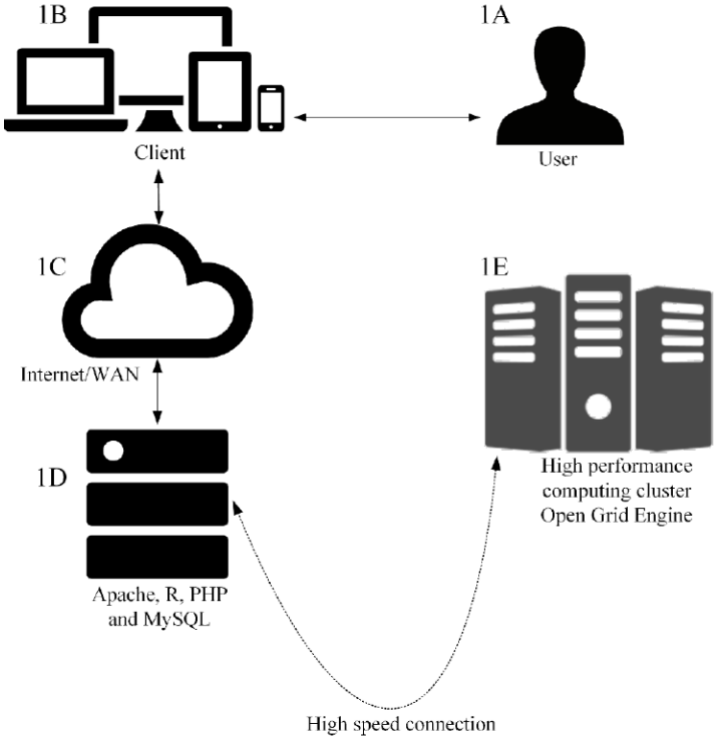
High content screening is a subset of High Throughput Screening, which is a process in which large libraries of chemical or biological reagents can be tested for activity in assays using automated methods.

HC StratoMinder provides a workflow in which big data can be analyzed:

1) Upload data
2) Preprocessing
3) Parameter selection
4) Quality control
5) Data normalization
6) Data transformation
7) Data standardization
8) Multiple imputation
9) Data reduction
10) Hit selection
11) Data mining
12) Reporting



Architecture:



Using the cloud for:

- Scalability
- Flexibility (in programming language)
- Cost-effectiveness
- Processing: f.e. EC2
- Storage: f.e. S3
- Security

8. Guest lecture: Ethics and law

Privacy is a difficult topic in big data. There is not a single definition for privacy. How is big data in research regulated? Through:

- Fundamental rights
  - Privacy, data protection, non-discrimination, etc.
- Regulations
  - International treaties
  - General Data Protection Regulation (GDPR)
  - National laws
- Non-binding guidelines
  - Declaration of Taipei (health research)
  - CIOMS guidelines (health research)

The GDPR article 9 claims:

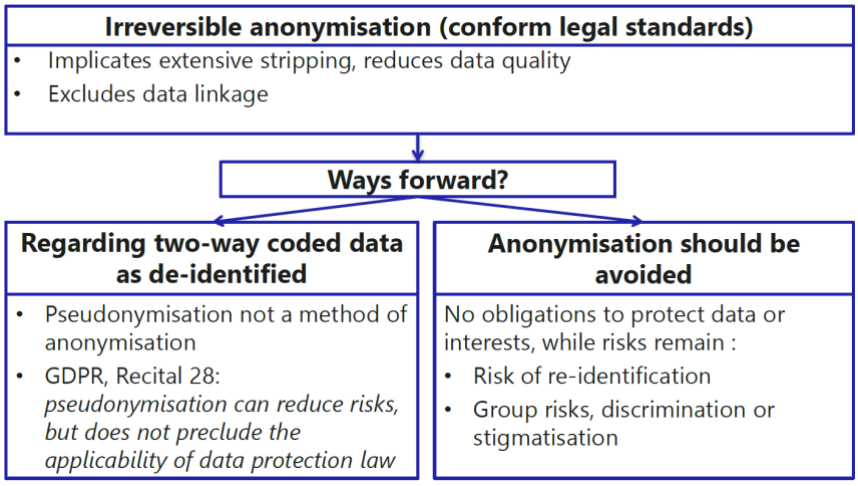
- Processing of special categories of personal data shall be prohibited. (e.g. health-related and genetic data)
- This prohibition does not apply
  - If the data subject has given **explicit consent**
  - If an appropriate **research exemption** from this prohibition is laid down in national (or EU) law

To do research, you can do three things to use data:

1. Provide anonymity
2. Get consent
3. Have research exemption

Anonymity

It might be impossible to guarantee absolute anonymity. The GDPR says: single out everything which can determine whether a natural person is directly or indirectly identifiable.



Key messages:

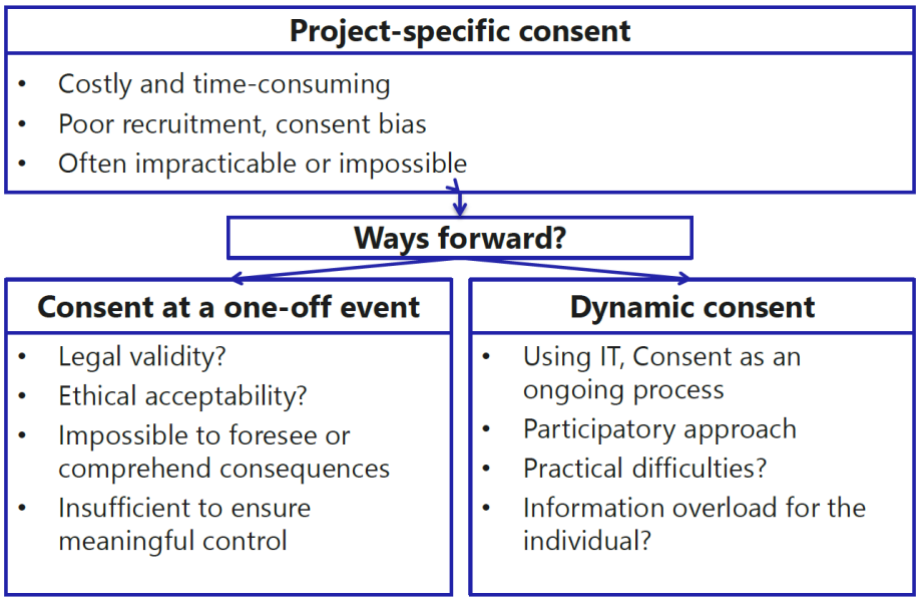
1. Anonymization is, on its own, an unsafe strategy for ensuring protection of individual rights and interests.
2. In addition, anonymization is difficult in practice without compromising the utility of the data set. In Big Data research, this problem increases.
3. Pseudonymization is a useful method to reduce privacy risks, but it does not preclude the applicability of the law.

Consent

Processing of personal data concerning health which is necessary for scientific research purposes, shall be permitted only with the consent of the data subject. And it leaves limited room for exceptions.

Informational self-determination problems:

- Cognitive:
  - People do not read privacy policies
  - If people read them, they do not understand them
  - If people read and understand them, they often lack enough background knowledge to make an informed choice
  - If people read them, understand them, and can make an informed choice, their choice might be skewed by various decision-making difficulties.
- Structural:
  - Information overload



Broad consent, GDPR: It is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping which recognized ethical standards for scientific research.

Key messages:

1. Seeking (broad) consent remains an important requirement to show respect for persons, also in Big Data research
2. Difficulties in seeking consent could hamper scientific progress
3. Consent does not, in itself, reduce risks to individuals

Research exemption

The importance of appropriate safeguards:

- Literature:
  - Limiting data access and use
  - Opt-out registration
  - Pseudonymization
  - Authorization by ethics committee
  - Engaging in public participation, etc.

- GDPR:
    - Research exemption should be subject to certain additional appropriate safeguards
    - Privacy by design, good information governance

General principles:
- Lawfulness
- Fairness
- Transparency
- Purpose limitation
- Data minimization
- Data accuracy
- Storage limitation
- Data security
- Accountability

9. Guest lecture: Innovation in psychiatry

What is changing in healthcare? The patient → citizen, doctor → ICT, treatment based on trials → based on clinical intelligence.

E-health is changing healthcare on 3 levels:
1. Research (real-time sensor/wearable data, big data)
2. Content (applications and games)
3. Organization of care (IT connections between different institutions)

Applications:
- Collect:
    - Research data platform (RDP)
    - Lokale opslag
- Pseudonymize prepare
    - RDP
    - Python
    - Jupyter
    - Excel
- Integrate
    - SynerScope
    - Postgres
- Explore & Analuse
    - SynerScope
    - R

Nowadays, the data process uses a lot more public datasets, code repositories, and pipelines to empower data scientists. These generate infographic reports which improve the connection between data scientist and beside employees.

Examples: Prediction of medication effect of antipsychotics and antidepressants, prediction of aggression based on text mining.

Blended psychiatry; a data driven learning organization: Data knowledge → Patient subjective knowledge → Professional knowledge.

10. Guest lecture: Natural language processing in psychiatry

Free text in the psychiatric electronic health record → 80% of data in unstructured.

It is domain-specific language, with spelling errors, abbreviations etc. The focus however lay on making the data useful, rather than redefining registration protocols.

But first, de-identification was needed. The free text holds a lot of privacy sensitive data. Through the DE-identification method for Dutch medical text they de-identified the patients. For example: names, locations and institutions.

Example topics for NLP in psychiatry other than de-identification:

- Sentiment analysis
- Clustering
- Information retrieval
- Text classification
    - Example: assessing violence risk

Automatic violence risk assessment:

Word2Vec: how to represent words numerically? Each word gets represented by a dense vector. The vectors are collected in a 'bag-of-words'.

Vector representation then allows mathematical operations such as addition or subtraction. You can use this to get similarities for the words for example.

An extension for word2vec is doc2vec, where texts are represented in the same way as in word2vec, in a 'bag-of-words'.

Using doc2vec, they made bag-of-words from text notes from psychiatric admissions. Them, by using a Recurrent Neural Network, they processed each note, classifying the admission. By adjusting internal weights, based on right/wrong prediction on violence, they improved the model.

Results:

Cross validation is used to estimate the performance (with folds, training and testing sets). Performance is then measured using Area Under Curve of the Receiver Operator Curve.

They showed that assessment of violence risk is possible using text only.