Informationsveranstaltung „Qualitätsmanagement für Hochdurchsatz-Genotypisierung”

TP5: „Primäre Datenstrukturen, Speicherung und Transfer von Genotyp-Daten“

21. Juni 2010, Berlin

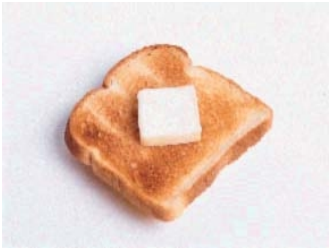
Andreas Wolf, Olaf Junge, Tim Lu, Michael Krawczak
Institut für Medizinische Informatik und Statistik
Universitätsklinikum Schleswig-Holstein, Campus Kiel

SPONSORED BY THE
Federal Ministry of Education and Research

Motivation I

Fehler sind nicht vermeidbar,
kommen zum ungünstigsten Zeitpunkt vor,
und sie sind teuer!

- Murphy's law: “If anything can go wrong, it will!“
- McGillicuddy's corollary: „...at the most inopportune time!“
- Murphy's constant: “Matter will be damaged in direct proportion to its value!”



Hier: *IT-Qualitätsmanagement zur Verlustminimierung*

Motivation II

- Transfer und Speicherung von Hochdurchsatz-Genotypisierungdaten stellen eine Herausforderung dar:
  - *Volumen* (GB, TB)
  - *Inhalt* (personenbezogene Daten)

- Hohes Konfliktpotential:
  - *Technische Beschränkungen* (Kapazitäten,...)
  - *Datenschutz* (re-identifizierbare Daten,...)
  - *Wissenschaftliche Fragestellung* (Klinische und genetische Daten,...)

Inhalt

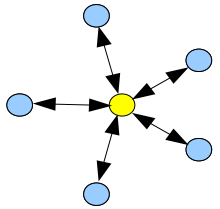
Bestandsaufnahme, **Evaluation** und **Empfehlungen** zu den Themen

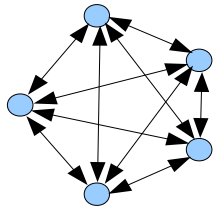
- Qualitätskontrollierter Datenaustausch (Formate, Transfer)
- Effiziente und sichere Speicherung (Architektur)
- Gesetzeskonformer Zugriff (rechtliche Aspekte)
- ...

Datenaustausch: Formate

Einheitliches Datenformat?

- *Pro*:
  - Als Standard durch strikte Regeln gut dokumentiert
  - Einfache Import / Export Routinen für Standard
  - Weniger Konverter nötig





- *Contra*:
  - Ein Standard sollte *alle* features *aller* Formate berücksichtigen
  - *Zukünftige Formate* (neue Technologien) schwer / nicht integrierbar, da Standard hierfür nicht konzipiert
  - Ständige *Wartung*

Nicht nötig!

Datenaustausch: Formate

Strukturparameter (“wie”): Syntax eines Formats

- Text *vs.* binär (Lesbarkeit *vs.* Platzbedarf)
- Feldtrenner (z.B. `[ \t\n\r\f]+`)
- End-Of-Line (abhängig vom OS)
- Character coding (z.B. ISO, UTF)
- Spalten- und Zeileninhalte (explizit oder implizit)

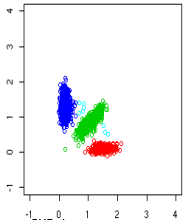
```
0 9682877SPC 0 0 2 2
0 3649838SPC 0 0 2 2
0 14639SPC 0 0 2 2
0 781250SPC 0 0 2 2
```


```
0 9682877SPC 0 0 2 2^M0 3649838SPC 0 0 2 2^M0 14639SPC 0 0 2 2^M0 781250SPC 0 0 2 2^M
```

Datenaustausch: Formate

Inhaltsparameter (“was”): Semantik, Metainformationen

- *Probe* (DNA): Probe annotations, genome build,...
- *Chip* (Substrat, auf dem DNA Probe analysiert wird): Chip ID, probe coordinates,...
- *Target* (Individuum, vom dem DNA analysiert wird): PedID, IndID, FaID, MoID, Geschlecht, Krankheitsstatus, Phänotypen,...
- *Assay* (Biochemische Reaktionen): Assay protocols, Variationen,...
- *Readout* (Messungen): Allele calling software, Genotyp-Formate,...
- *Analysis* (Auswertung des readout): Methoden, Software,...



Datenaustausch: Formate

Datentypen zum Speichern von Genotypen: Text *vs.* binär

Speicherplatz für einen Genotyp:

- *Character*: 2 Bytes / Genotyp = 16 Bit (z.B. “`AC`“ explizit)
- *Binary*: 4 Bits / Genotyp = 4 Bit (z.B. “`0010`” als Codierung für “`AC`”)
- *Binary*: 2 Bits / Genotyp = 2 Bit (z.B. “`10`” für Heterozygote)

Lesbarkeit *vs.* Platzbedarf *vs.* Verfügbarkeit von Metainformationen...

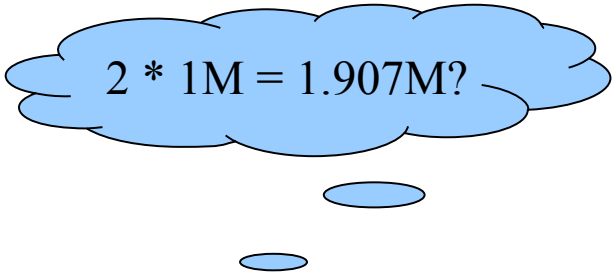
Datenaustausch: Formate

NB: International Standard Units *vs.*
International Electrotechnical Commission!

IEC Empfehlungen (1996):

- 1 Kibibyte (1KiB) = 1024 Byte
- 1 Kilobyte (1KB) =1000 Byte

Geringe Akzeptanz!



Name	Symbol	SI (Basis 10)	IEC (Basis 2)	Differenz
kilo	k/K	10^3	2^{10}	+2.4%
mega	M	10^6	2^{20}	+4.9%
giga	G	10^9	2^{30}	+7.4%
tera	T	10^{12}	2^{40}	+10.0%
peta	P	10^{15}	2^{50}	+12.6%

Datenaustausch: Transfer

Datentransport

- *Portionieren* in “Archiven” oder “Container”
- Optionale *Kompression*
- *Datenintegrität* (Fehlererkennung und -korrektur)
  - Anspruchsvolle Prüfsummen (cyclic redundancy check, CRC)
  - Kryptographische Hashes, z.B. SHA-Familie (secure hash algorithms, NSA), MD-Familie (message-digest algorithms, MIT)
  - Andere, mit Fehlerkorrektur und Schutz vor mutwilligen Modifikationen
- *Datenaustausch* über
  - Nicht-zuverlässige aber effiziente Kanäle (UDP) mit Check
  - Sichere Kanäle, z.B. VPN, SSL, Verschlüsselung (PGP)

Datenaustausch: Transfer

Genotyp- und Sequenz- spezifische Kompression

- Ausnutzen der inhärenten Struktur von DNA
  - <1% Variation von einem Referenz-Genom
  - “SNP mapping”, Variation von DBSNP
  - Kopien von Genen, Wiederholungen (STRs), Palindrome,...
- Effizientere Kompression als generische Algorithmen, dennoch verlustfrei

Datenspeicherung

Relationale Datenbanken auch für genomweite Daten?

Pro:

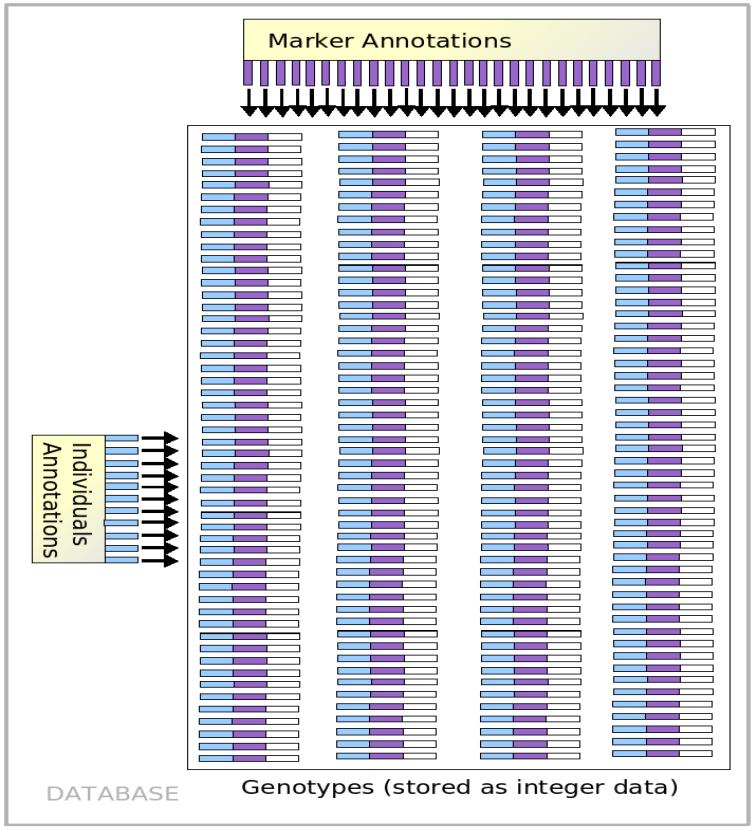
- Organisierte Speicherung
- Indizierung bei Speicherung
- Normalisierungen zur Fehler-Erkennung und -Vermeidung
- Strukturierter Zugriff
- GUI

Contra:

- Limitierter Speicherplatz
- Begrenzter Hauptspeicher
- Strukturelle Ineffizienzen, besonders bei Verwendung eines LIMS, das nicht zu diesem Zweck designed wurde

Datenspeicherung

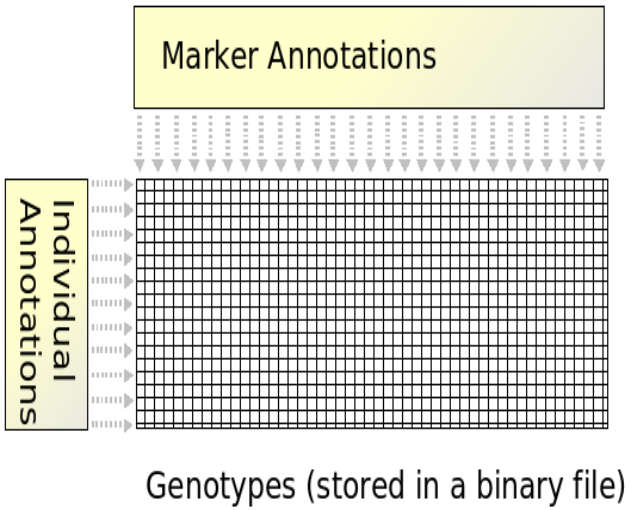
Speichern von Genotypen: **Integer** *vs.* Binary *vs.* BLOB.



- Normalisiert
- Keine Redundanzen
- Genotypen explizit indiziert (Individuum und Marker)
- Einfacher Zugriff
- Hoher Speicherbedarf ☹

Datenspeicherung

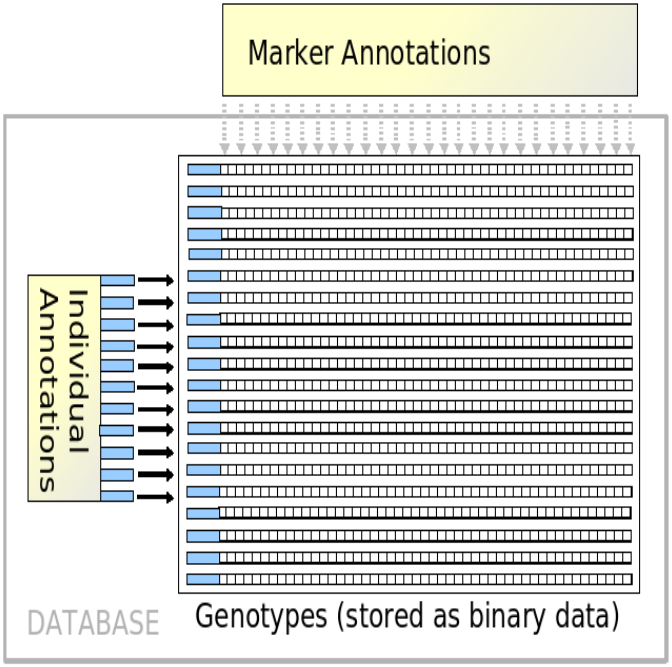
Speichern von Genotypen: Integer *vs.* **Binary** *vs.* BLOB.



- Normalisiert
- Komprimiert / binär
- Geringer Speicherbedarf
- Genotypen implizit durch Position indiziert (Zeile & Spalte)
- Schwieriger Zugriff auf Genotypen ☹

Datenspeicherung

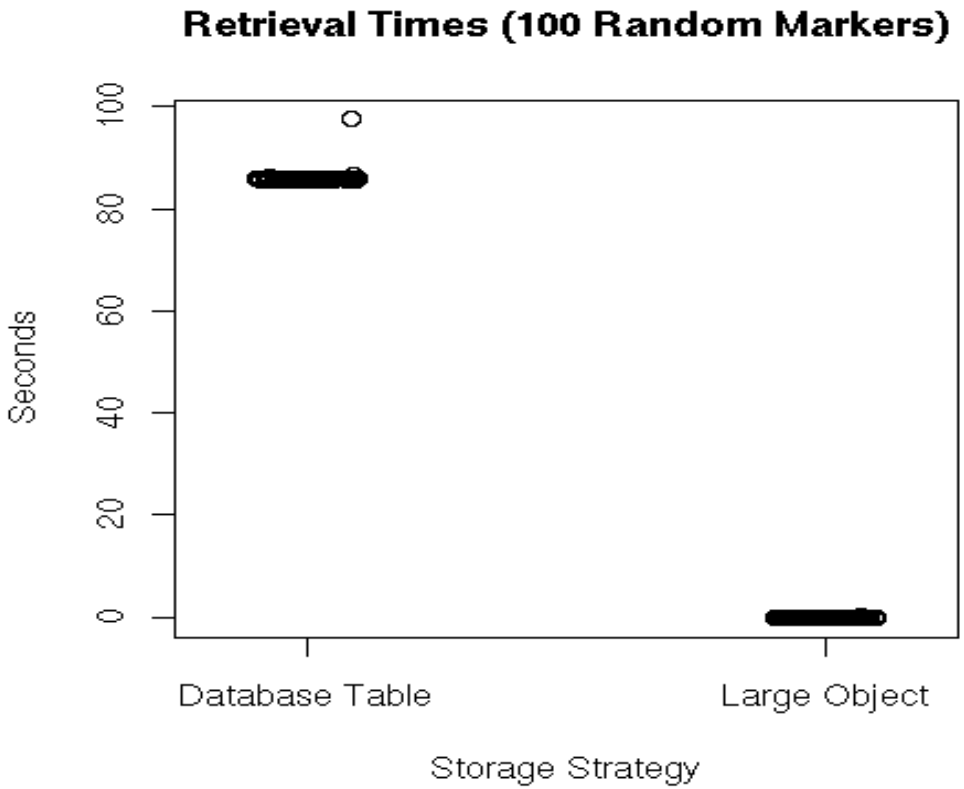
Speichern von Genotypen: Integer *vs.* Binary *vs.* **BLOB**.



- Hybridlösung
- **Binary large objects**, BLOBs
- De-normalisiert
- Genotypen explizit indiziert an Individuum und implizit an Marker
- Moderater Speicherbedarf
- Einfacher Zugriff, z.B., `lo_import()` und `lo_read()` in PostgreSQL

Datenspeicherung

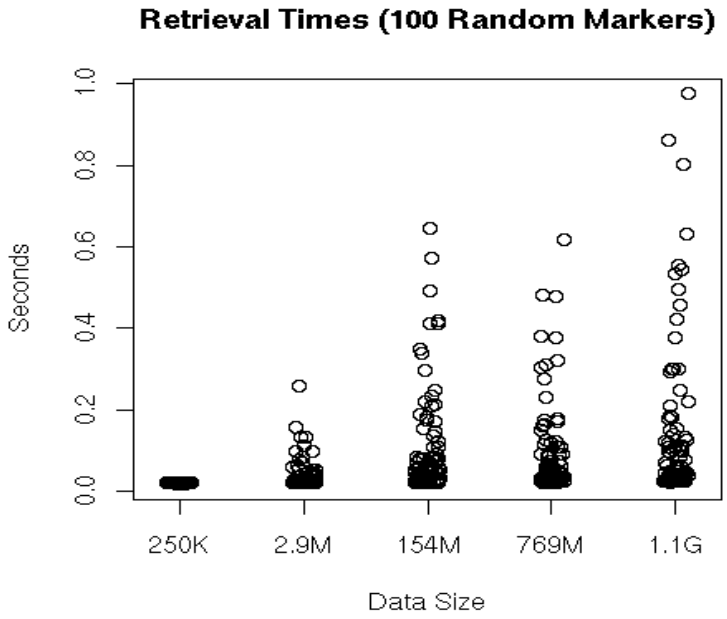
Benchmark: Normalisiert *vs.* de-normalisiert (BLOB)

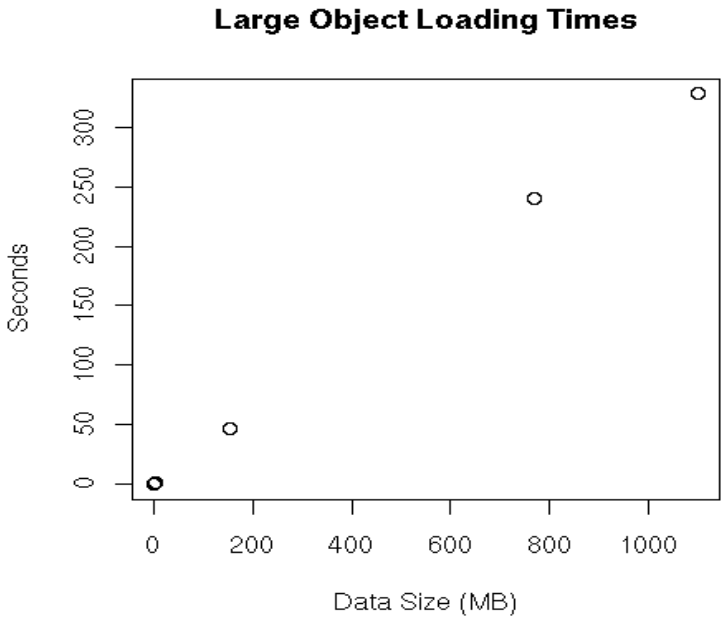


(1M SNPs, 250 KB)

Datenspeicherung

Benchmark: Lesen und Schreiben von BLOBs





Datenschutz

Rechtliche Aspekte:

- Genetische Daten berühren *per se* verschiedene Rechtsbereiche, z.B. Datenschutzrecht, Persönlichkeitsrecht, Eigentumsrecht
- Der Umgang mit solchen Daten wird durch Gesetze geregelt
- Gerade genetische Daten stehen unter strenger öffentlicher Beobachtung

Essenzielle Anforderungen:

- Die Rechte der Besitzer / Spender müssen respektiert werden
- Geltende Gesetze müssen befolgt werden
- Zustimmung öffentlicher Datenschutzbehörden sollte eingeholt werden

Datenschutz

Zu beachtende Rechtskriterien:

- Datenminimierung und Datenvermeidung
- Anonymisierung und Pseudonymisierung
- Schutz vor Datenverlust („data safety“)
- Schutz vor Einbruch und Umgehung („data security“)
- Transparente Beschreibung der Prozesse (SOPs)
- Zustimmung zu Daten-Erhebung, -Transport und -Verarbeitung
- ...

Datenschutz

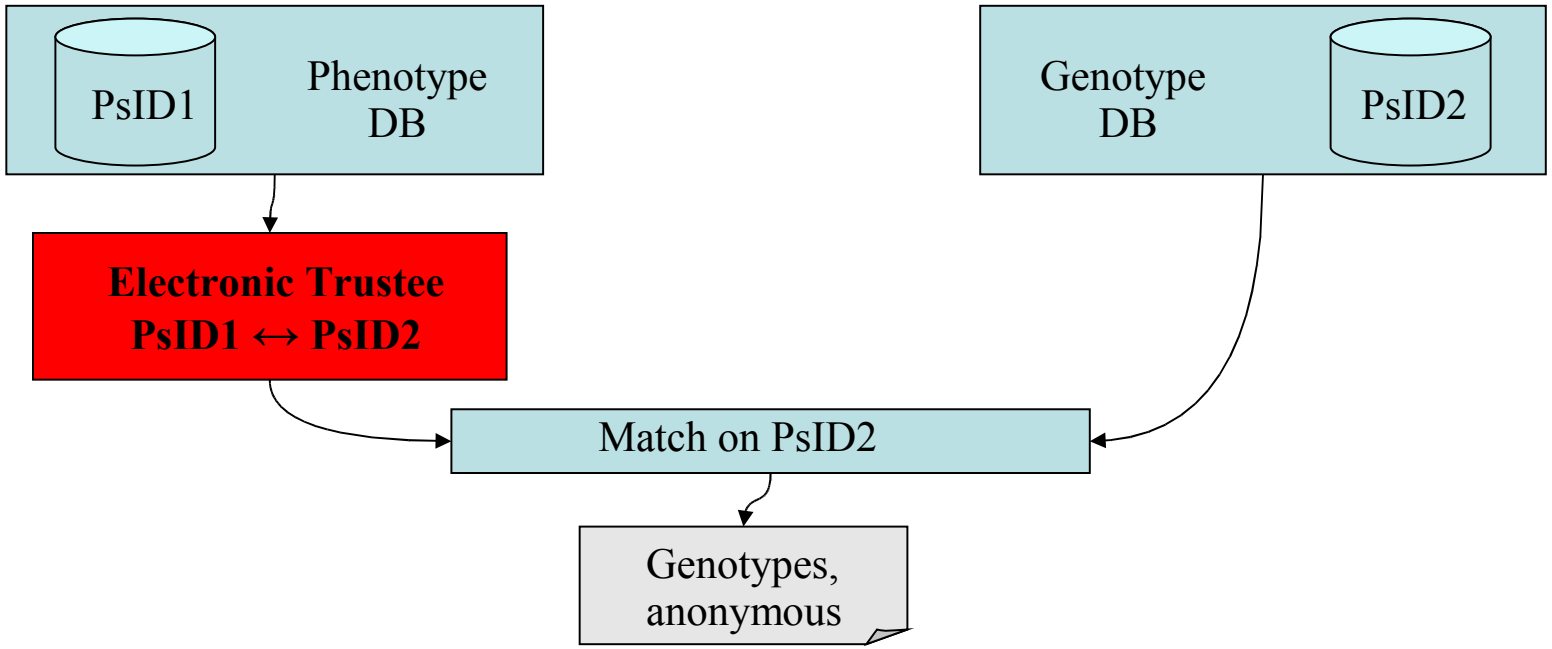
Mechanismen zum Erreichen dieser Ziele:

- *Rollenkonzpte (multilevel access)*
  - Definition einzelner Prozesse
  - Klassifizierung der Daten gemäß eines Sicherheitsmodells gestaffelt nach Sensitivität
  - Daten mit Prozessen und Prozesse mit Rollen verknüpfen
  - Zugriffsrechte an Rollen mit gewissen Freigaben vergeben
- *Multilevel security (MLS)*
  - Unterstützung von Rollenkonzepten durch Hardware (z.B. “asymmetric Isolation”) und Software (z.B. “trusted OS”)
- *Datentreuhänder (data custodianship, electronic trustees)*

Datenschutz

Beispiel (popgen-Trustee):

- formale, organisatorische und räumliche Trennung
- Pseudonymisierung, (faktische) Anonymisierung
- Vertrauenswürdiger Datentreuhänder



Datenschutz

Probleme und Einwände:

- Administrativ komplex
- Zeitaufwand und eingeschränkter Zugriff
- Datensätze sind (faktisch) anonym, sind nicht wiederverwendbar (in anschließenden Analysen oder Metaanalysen)

Dennoch:

- Datentreuhänder garantieren Einhaltung wesentlicher Kriterien (Datenminimierung, Schutz vor Re-Identifizierung und Schattendatenbanken)
- Zustimmung von öffentlichen Datenschutzbehörden erhöht das Ansehen und die Vertrauenswürdigkeit (Einwerben von Spendern und Fördermitteln)

Weitere Themen des TP

- **Hardwareanforderungen**
  - CPU, RAM, HD, Optische Medien, USB Flash
  - Rasante Entwicklung, wie Genotypisierungstechnologien
- **Archivierung / Backups**
  - Backup Systeme und Version control software
  - Wahl abhängig von diversen Parametern (Effizienz, Vernetzung,...)
- **Testen von Programmen und Formatversionen**
  - Referenzdaten und -programme zum Testen, program checking
  - Softwaredesign inclusive Testen (EP, AGILE), Generatoren

Zusammenfassung

“Take home messages“

- *Das* Supertool gibt es nicht, und es ist auch nicht erstrebenswert
- Ansprüche → Anforderungen ← Rahmenbedingungen
- Universelle Empfehlungen nicht sinnvoll / möglich
- Datenstrukturen und Datentransfer
  - Lesbarkeit *vs.* Speicherplatzbedarf, d.h. Text *vs.* binär
  - Datenintegrität
- Datenspeicherung und Datenzugriff
  - Relationale Datenbanken und BLOBs
  - Datenschutz und Datentreuhänder