# Aufbereitung und Analyse von Daten in phänotypischen und genotypischen Forschungs-datenbanken mit tranSMART

5. August 2016 | Berlin

# Programm

*(Stand: 03.08.2016)*

## 10.15 Uhr Grundlegende Konzepte

- Begrüßung

5' *Dr. Johannes Drepper (TMF), Prof. Dr. Ulrich Sax (Universitätsmedizin Göttingen), Matthias Löbe (Universität Leipzig)*

- Einführung in die tranSMART-Plattform

15'+3' *Prof. Dr. Ulrich Sax (Universitätsmedizin Göttingen)*

- tranSMART Architecture and Roadmap

45'+15' *Kees van Bochove (CIO The Hyve, tranSMART-RoadMap-Presentation at tranSMART-Meeting 2015)*

- Open Data Analytics – Gefahren durch p-Hacking

15'+3' *Prof. Dr. Ulrich Sax (Universitätsmedizin Göttingen)*

## 12.00 Uhr Mittagspause

## 13.00 Uhr Praktische Arbeit

- Klick-a-thon: Durchgehen eines vorbereiteten praktischen Szenarios durch die Workshopteilnehmer
- Lösung von Aufgaben in Eigenarbeit

# Programm

*(Stand: 03.08.2016)*

## 13.00 Uhr Praktische Arbeit

- Klick-a-thon: Durchgehen eines vorbereiteten praktischen Szenarios durch die Workshopteilnehmer
- Lösung von Aufgaben in Eigenarbeit

## 14.15 Uhr Kaffeepause

## 14.45 Uhr Fortgeschrittene Themen

- tranSMART and beyond

**20'+5'** *Reinhard Schneider / Sascha Herzinger (Universität Luxemburg)*

- SmartR – Dynamic Visual Analytics in TranSMART

**20'+5'** *Reinhard Schneider / Sascha Herzinger (Universität Luxemburg)*

- Import und Repräsentation hochdimensionaler Daten

**10'+3'** *Christoph Knell (Friedrich-Alexander-Universität Erlangen-Nürnberg)*

# Programm

*(Stand: 03.08.2016)*

## 14.45 Uhr Fortgeschrittene Themen

- tranSMART and beyond

20'+5' *Reinhard Schneider / Sascha Herzinger (Universität Luxemburg)*

- SmartR – Dynamic Visual Analytics in TranSMART

20'+5' *Reinhard Schneider / Sascha Herzinger (Universität Luxemburg)*

- Import und Repräsentation hochdimensionaler Daten

10'+3' *Christoph Knell (Friedrich-Alexander-Universität Erlangen-Nürnberg)*

- Beispiele zu Daten aus der Routineversorgung (§ 21 KHEntG, Intensivmedizin) und aus der Forschung (Mikrobiom)

10'+3' *Benjamin Baum (Universitätsmedizin Göttingen)*

- Zusammenfassung und Ausblick

*Prof. Dr. Ulrich Sax (Universitätsmedizin Göttingen)*

## 16.15 Uhr Ende des Workshops

UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

# Einführung in die tranSMART-Plattform

**TMF-Workshop**

Aufbereitung und Analyse von Daten in phänotypischen und genotypischen Forschungsdatenbanken mit tranSMART

Berlin, 05.08.2016

Christian Bauer[1], Benjamin Baum[1], Ulrich Sax[1]

[1]Department of Medical Informatics, WG Infrastructure for Translational Research

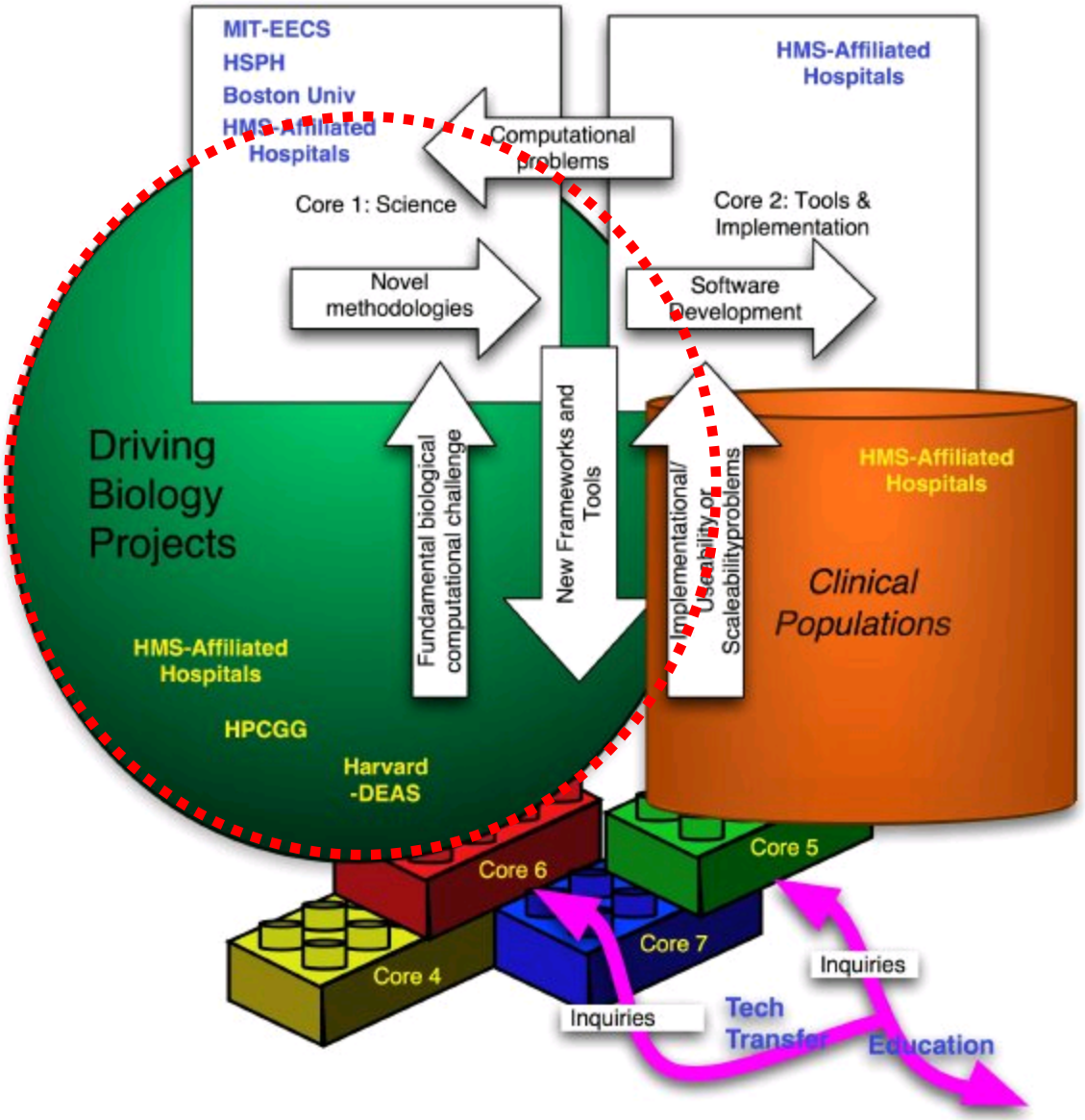
GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

# Informatics for Integrating Biology and the Bedside ($I^2B^2$)

- IT-Infrastruktur zur Unterstützung genetischer bzw. genomischer Forschung mit dem Schwerpunkt der Integration klinischer und genomischer Daten
- Eines von vier Zentren in USA

# Informatics For Integrating Biology And The Bedside ($I^2B^2$)




**Vernetzte Forschung:**

- IT-Infrastruktur zur Unterstützung genetischer bzw. genomischer Forschung
- Schwerpunkt: Integration klinischer und genomischer Daten
- Eines von vier Zentren in USA

www.I2B2.org
Proposal to Establish an NIH-Supported National Center for Biomedical Computing
PI: Isaac Kohane (2004)

Collaboration between the i2b2 (Informatics for Integrating Biology and the Bedside) National Center for Biomedical Computing, the Chair of Medical Informatics, University of Erlangen-Nuremberg and the Department of Information Technology, Göttingen University Hospital

**MEMORANDUM OF UNDERSTANDING**

Date: 19.11.2009

Prof. Dr. Hans-Ulrich Prokosch

Head of Department

Chair of Medical Informatics
University of Erlangen-Nuremberg

Prof. Dr. Ulrich Sax

Head of Department

Department of Information Technology
Göttingen University Hospital

Isaac Kohane, M.D., Ph.D.

Director

i2b2 National Center for Biomedical Computing
Brigham Women's Hospital Boston

UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG

Isaac Kohane, M.D., Ph.D.
Director
i2b2 National Center for Biomedical Computing
77 Avenue Louis Pasteur
Boston, MA 02115

USA

37099 Göttingen **Briefpost**
Robert-Koch-Straße 40, 37075 Göttingen, Germany **Adresse**
0551 / 39-13148 **Telefon**
0551 / 39-8234 **Fax**
usax@med.uni-goettingen.de **E-Mail**

15. Dezember 2009 **Datum**

**Letter of Support for the „i2b2 National Center of Biomedical Computing" call RFA-RM-09-002**

in response to the Call of the Department of Health and Human Services and the National Institutes of Health (NIH) on the topic: RFA-RM-09-002 National Centers for Biomedical Computing (U54).

Sincerely

Prof. Dr. Hans-Ulrich Prokosch  Prof. Dr. Ulrich Sax  Sebastian C. Semler

Research Article

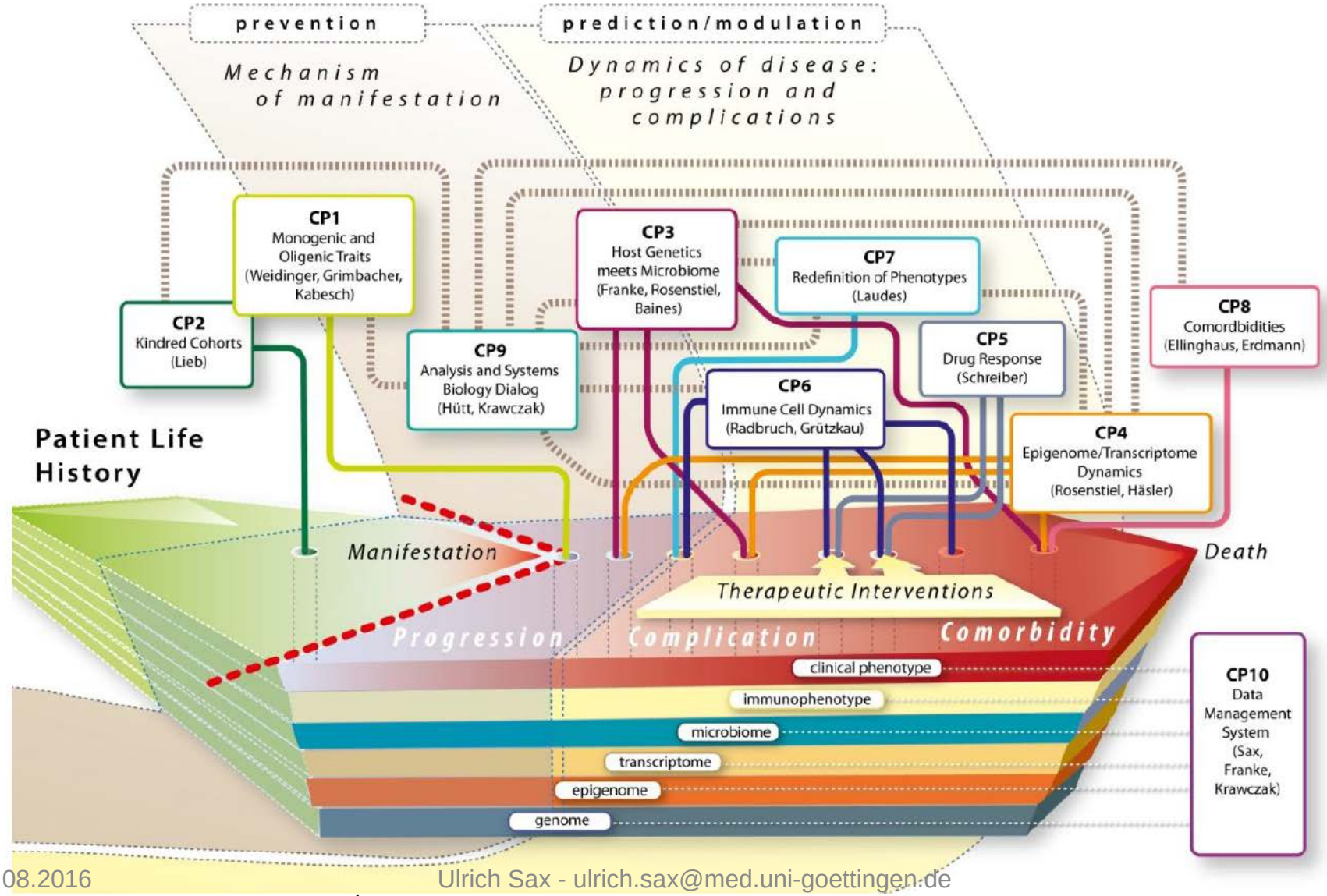
Applied Clinical Informatics

# TMF Special Issue: Unlocking Data for Clinical Research - The German i2b2 Experience

T. Ganslandt[1]; S. Mate[2], Helbing K[3], U. Sax[3,4], H.U. Prokosch[1,2]

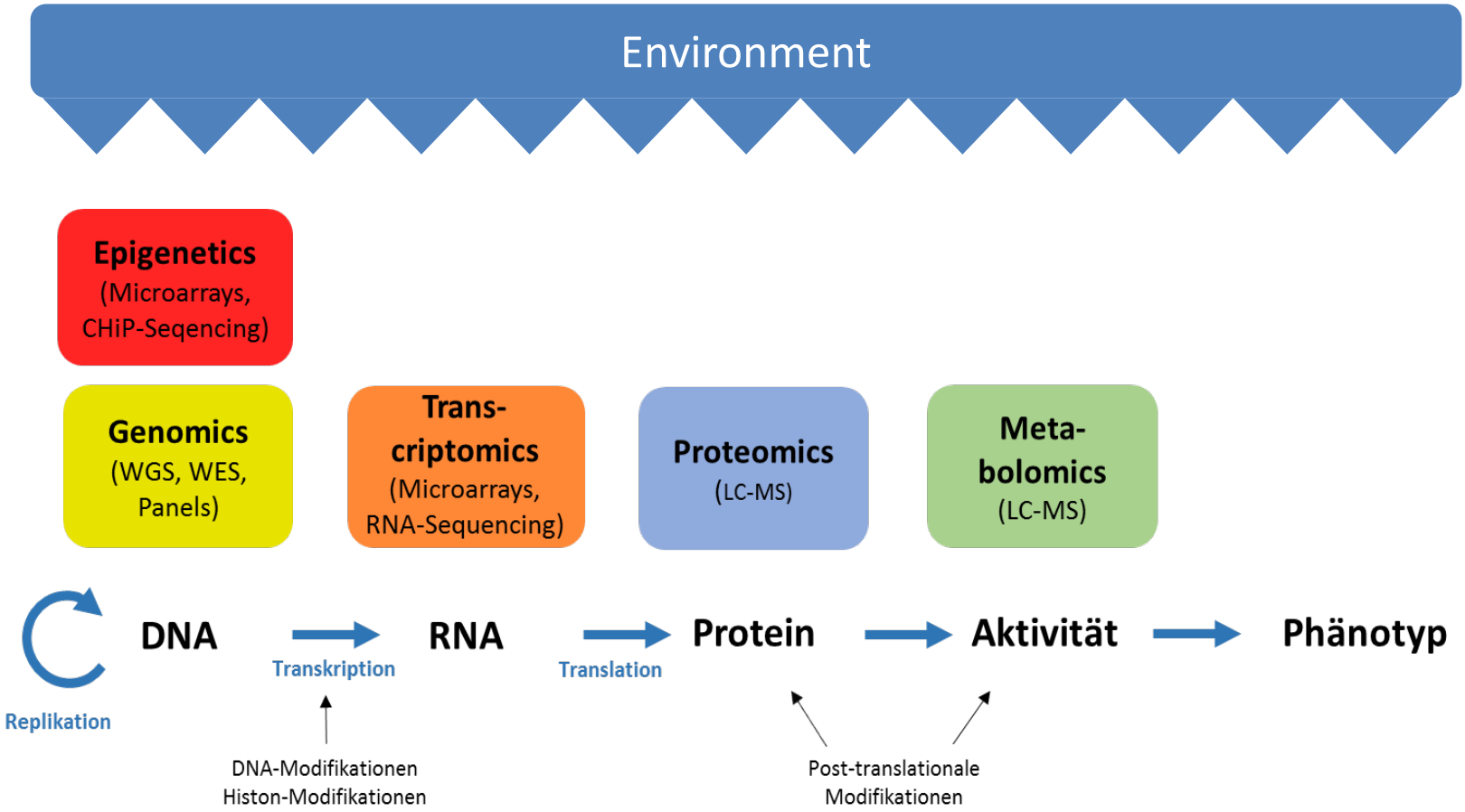
[1] Center for Medical Information and Communication, Erlangen University Hospital, Erlangen, Germany
[2] Chair of Medical Informatics, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany
[3] Department of Medical Informatics, University Medical Center Göttingen, Göttingen, Germany
[4] Division of Information Technology, University Medical Center Göttingen, Göttingen, Germany

# e:Med - sysINFLAME



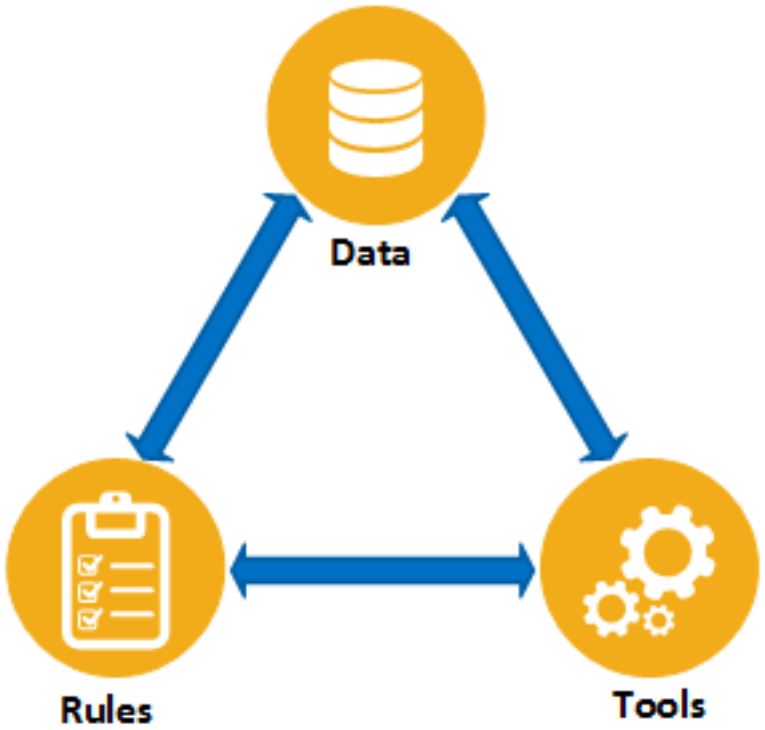



sysINFLAME proposal 2014: consortium






http://www.tmf-ev.de/Produkte/ITReport.aspx


Data
Rules
Tools

# WP3: Use and Access Policy (UAP)



**Nutzungsordnung für Daten im Verbund sysINFLAME**

v1.5 vom 17.06.2015

1. Ziel dieser Nutzungsordnung ist die Regelung der wissenschaftlichen Verwendung der im Verbund sysINFLAME verfügbaren Daten. Gegenstand der Nutzungsordnung sind allein die in Anhang 1 beschriebenen, bereits existierenden bzw. noch zu erhebenden Datensätze (fortan: „Daten“). Diese Daten werden in einer gemeinsamen Infrastruktur des sysINFLAME-Verbundes zusammengeführt.

2. Jeder Wissenschaftler, der Daten in sysINFLAME einbringt oder nutzt, ist für die Einholung eines eventuell erforderlichen Ethikvotums selbst verantwortlich. In jedem Schritt der Datennutzung sind alle einschlägigen Datenschutzbestimmungen einzuhalten. Daten dürfen auch innerhalb des Verbundes nur in pseudonymisierter Form weitergegeben werden. Alle Einverständniserklärungen und Patientenlisten verbleiben beim Datenhalter.

3. Das Steering Committee (STC) des Verbundes sysINFLAME setzt ein dreiköpfiges Use and Access Committee (UAC) ein, das auf Antrag und nach transparenten Kriterien über die Nutzung der Daten aus Anhang 1 entscheidet. Das UAC wählt aus dem Kreis seiner Mitglieder einen Vorsitzenden. Die Entscheidungskriterien des UAC sind mit dem STC abzustimmen, schriftlich niederzulegen und orientieren sich an den Regeln der guten wissenschaftlichen Praxis. Mit der Unterzeichnung dieser Nutzungsordnung erkennen die Datenhalter in sysINFLAME die Entscheidungskompetenz des UAC über die Nutzung von Daten verbindlich an.

4. Jede Nutzung der Daten aus Anhang 1 bedarf grundsätzlich eines positiven Votums des UAC sowie der Genehmigung des STC, sofern die Nutzung nicht bereits durch eine andere Nutzungsordnung geregelt ist. In diesem Fall ist die Nutzungsordnung dem UAC zur Kenntnis zu geben und ihr entsprechend zu verfahren. Die Nutzung darf weder vom UAC noch vom Datenhalter unbillig verweigert werden. Streitfälle entscheidet das STC mit einfacher Mehrheit seiner Mitglieder.

5. Das UAC hat seine Aufgaben zügig zu erfüllen und spätestens 4 Wochen nach Eingang eines Nutzungsantrags über diesen zu entscheiden. Das UAC beschließt mit einfacher Mehrheit seiner Mitglieder.

6. Jeder Antrag auf Datennutzung enthält die folgenden Angaben:
- verantwortlicher Wissenschaftler
- Vertragspartner
- Titel des Vorhabens
- beabsichtigter Zeitraum
- Ziel des Vorhabens
- wissenschaftlicher Hintergrund
- Vorhabenbeschreibung
- Begründung der Machbarkeit
- Beschreibung möglicher Risiken
- zur Verfügung stehende (materielle und personelle) Ressourcen
- Einzelheiten zu den angeforderten Daten

7. Das UAC prüft den Antrag hinsichtlich folgender Kriterien:
- schlüssige wissenschaftliche Begründung
- Verfügbarkeit eines ausreichenden Datenbestandes (Fallzahl, Einwilligung)
- Erreichbarkeit des Ziels der Nutzung

8. Nach Prüfung des Antrags gibt das UAC schriftlich eine der folgenden drei Empfehlungen:
- Der Antrag soll genehmigt werden
- Der Antrag soll nur unter Auflagen oder nach bestimmten Modifikationen genehmigt werden
- Der Antrag soll abgelehnt werden

Eine Ablehnung oder Genehmigung unter Auflagen ist kurz zu begründen. Das STC entscheidet final über den Antrag und veranlasst ggf. die Bereitstellung der Daten.

BRIEFINGS IN BIOINFORMATICS. VOL 16. NO 2. 280–290
Advance Access published on 7 March 2014
doi:10.1093/bib/bbu006

# Translational research platforms integrating clinical and omics data: a review of publicly available solutions

*Vincent Canuel*, Bastien Rance*, Paul Avillach, Patrice Degoulet and Anita Burgun*

Submitted: 11th November 2013; Received (in revised form): 3rd February 2014

## Briefings in Bioinformatics

- International forum for researchers and educators
- For users of databases and analytical tools
  - Contemporary genetics
  - Molecular and systems biology
  - Provides practical help
- Impact factor: 9.6 (2014)

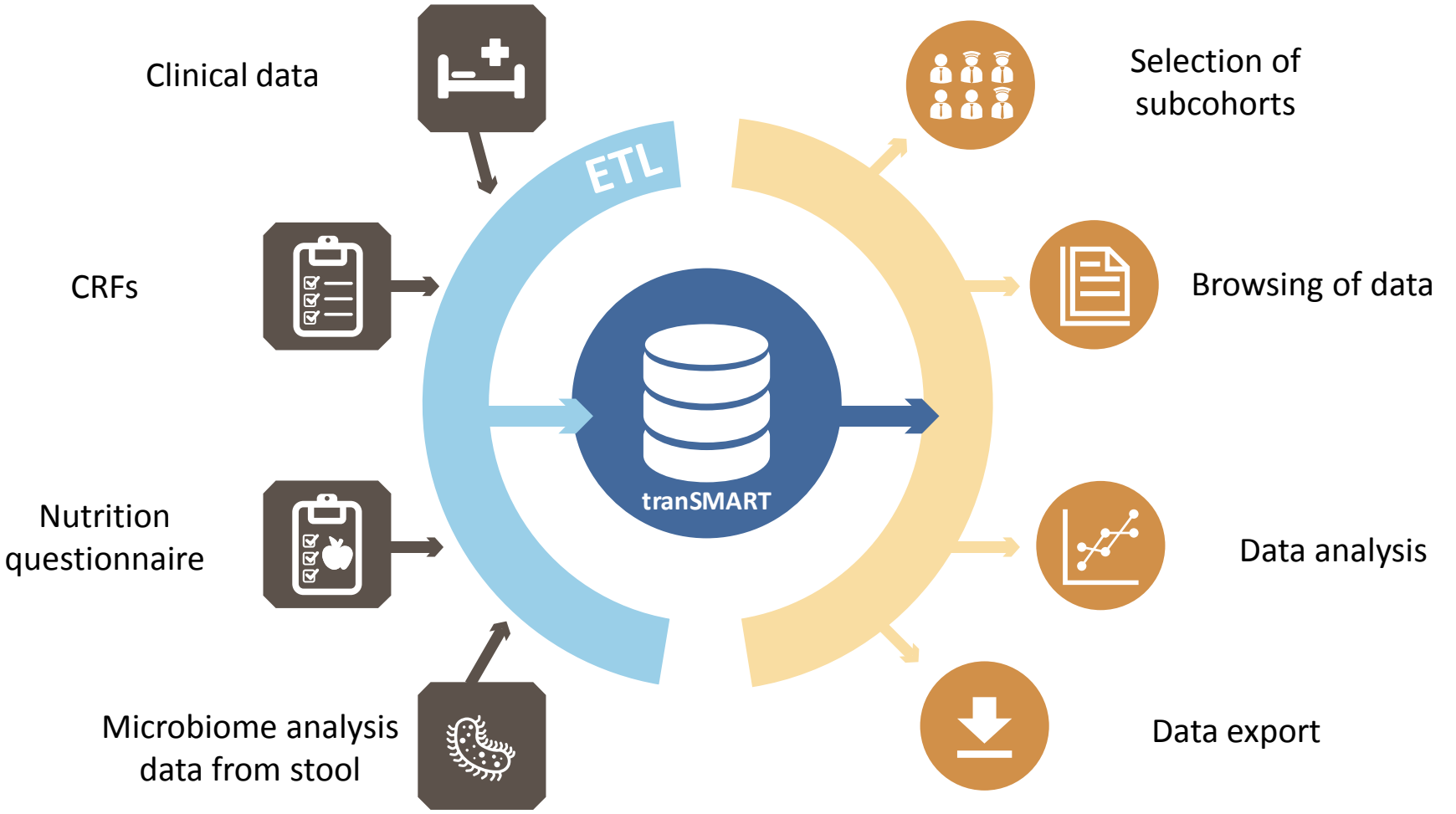
# Results

| Platform | tranSMART |
|---|---|
| Community | |
| Institution/project initiators | Johnson & Johnson, USA |
| Funding | Initially Johnson & Johnson funded—now public-private consortia |
| Reference | Szalma *et al.* 2010 |
| PMID | 20642836 |
| Software availability | Open source |
| Licensing | GPL v3 |
| User mailing list or support | Yes |
| URL | http://transmartfoundation.org |
| Information content | |
| Clinical data | |
| Demographics | Yes |
| Outcomes | Yes |
| Biological results | Yes |
| Images | No |
| Structured clinical research data | Yes |
| 'omics' data | |
| mRNA expression | Yes |
| miRNA expression | No |
| SNPs | Yes |
| Copy number variations | Yes |
| DNA methylation | No |
| Protein/phosphoprotein expression | Yes |
| Structural rearrangements | No |

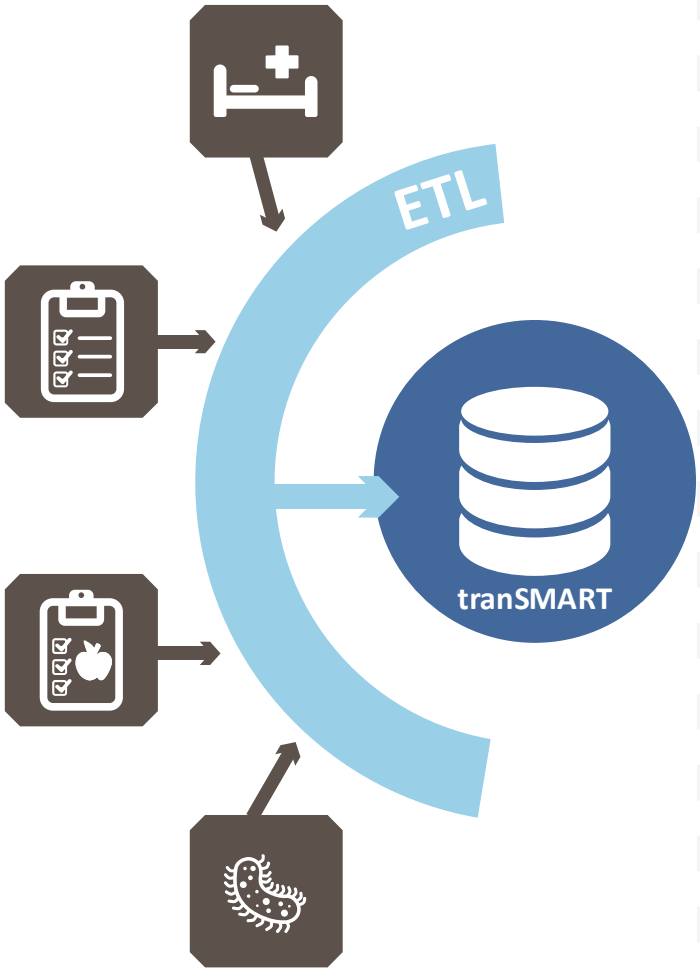
# Results

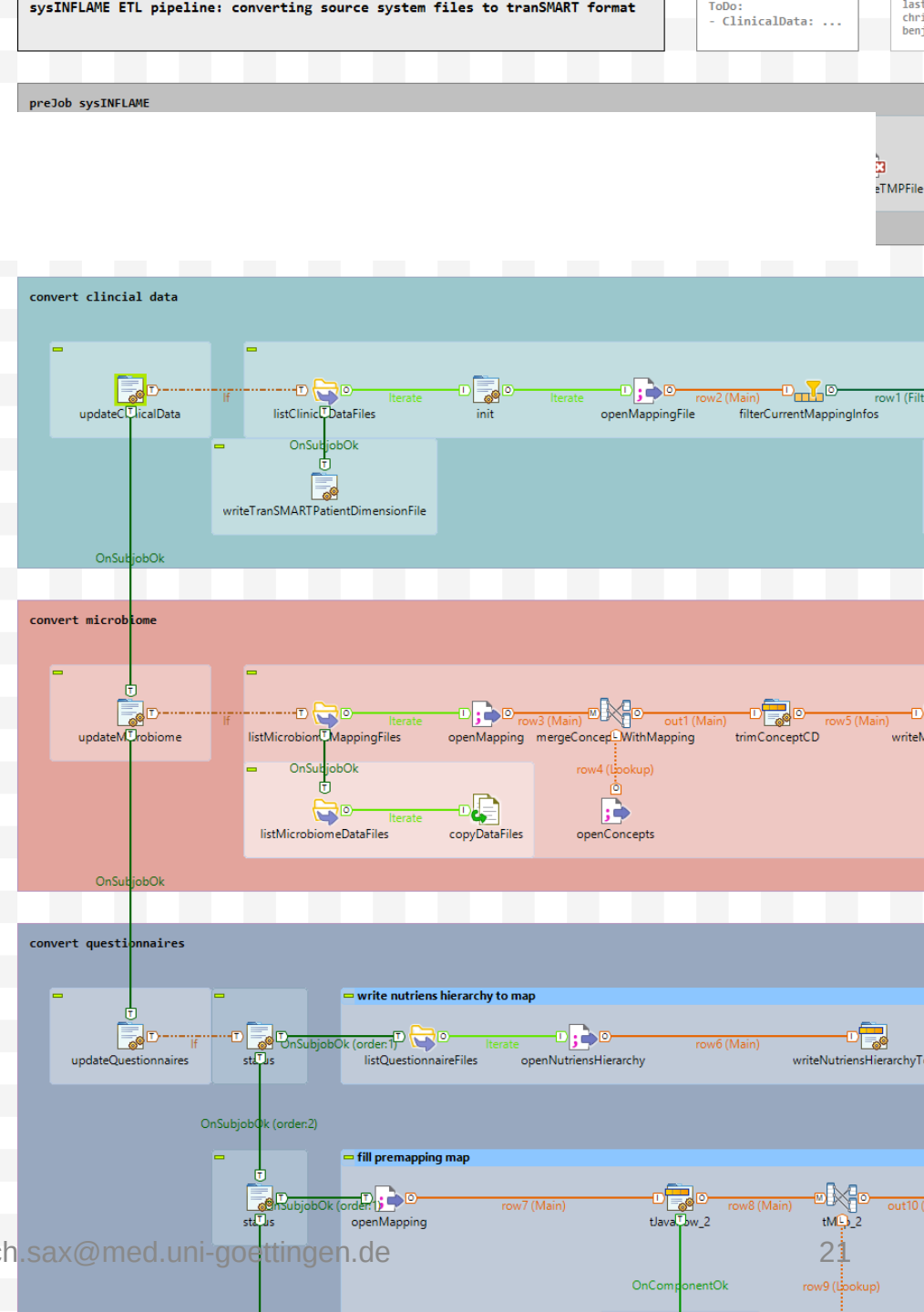
| Platform | tranSMART |
|---|---|
| Privacy management environment | |
| Anonymization/de-anonymization | No |
| Analysis supports | |
| Statistical framework | R |
| Analytical features | GenePattern, Bioconductor Plink |
| Visualization tools | Haploview, IGV and output of analytical tools |
| Interoperability support | |
| Ontologies/standard terminologies | i2b2 ontology features |
| Collaborative environment | Yes |
| Secure environment | Yes |
| Support of multisite requests | No |
| APIs/web services interface | JBoss |
| System requirements | |
| Operating system | Linux |
| Database management system | Oracle II/PostgreSQL |
| Database type | Centralized |
| Software dependencies | SOLR, R, i2b2 |
| Main programming language | Java/GRAILS |
| Server side | Tomcat/JBoss |
| Client-side interface | Web browser |
| Platform support | |
| Installation procedures | Yes |
| Configuration documentation | Yes |
| User documentation | Yes |

# ShowCase tranSMART

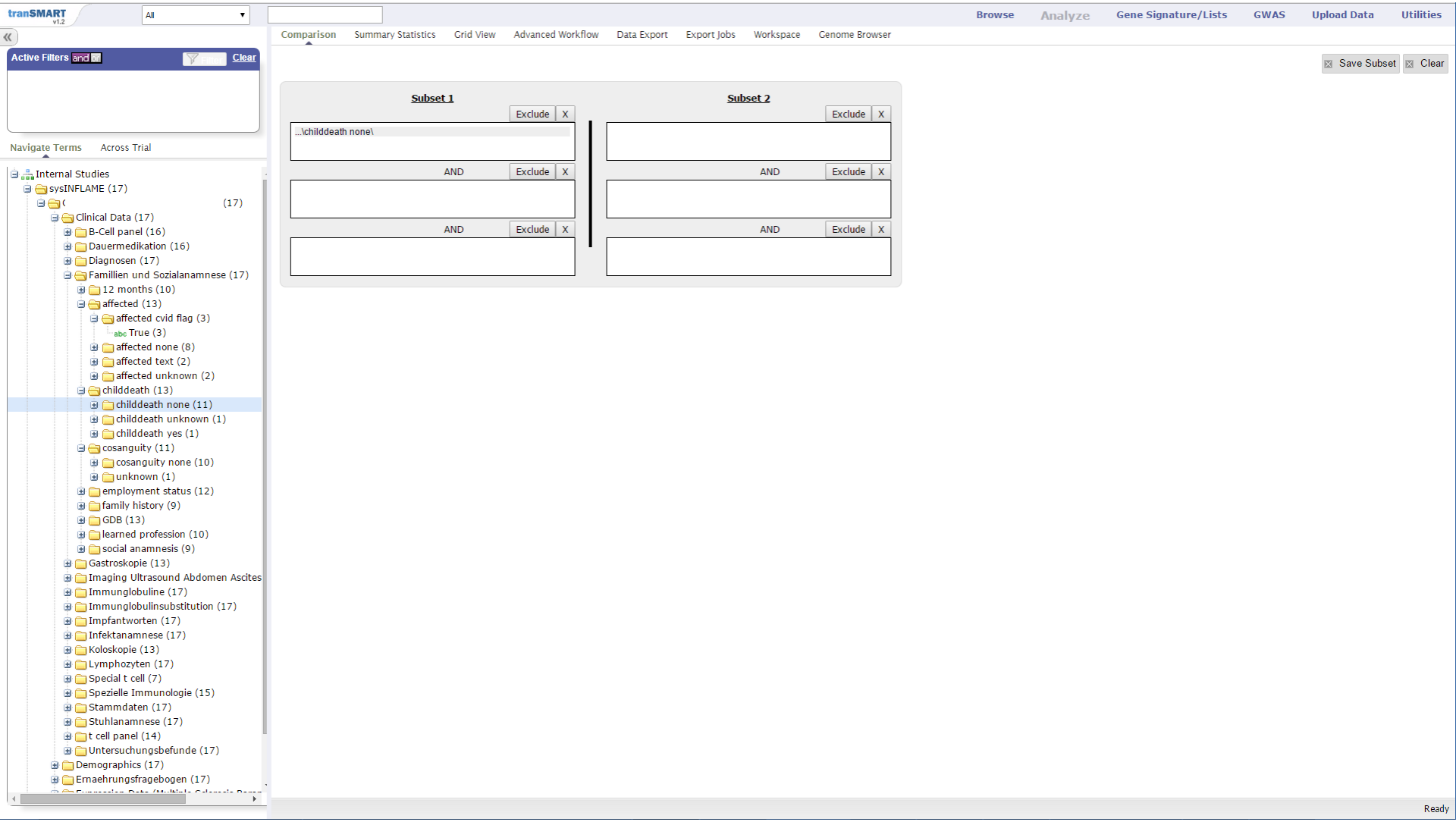


# ShowCase tranSMART

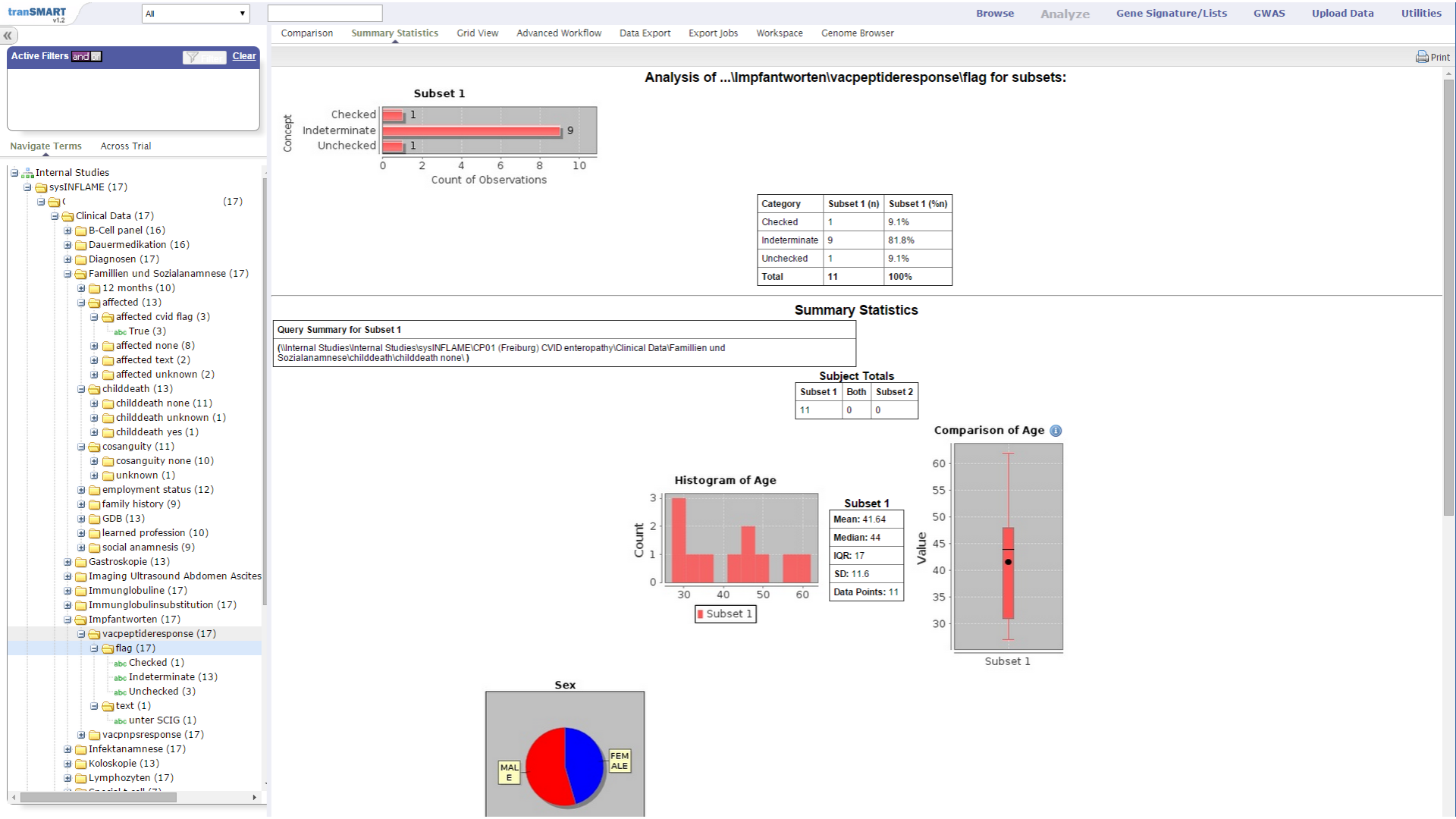


# ShowCase tranSMART



# ShowCase tranSMART



# ShowCase tranSMART

tranSMART v1.2 | All | Browse | Analyze | Gene Signature/Lists | GWAS | Upload Data | Utilities

Comparison | Summary Statistics | Grid View | Advanced Workflow | Data Export | Export Jobs | Workspace | Genome Browser

Active Filters and | Clear

Navigate Terms | Across Trial

- Internal Studies
  - sysINFLAME (17)
    - ( (17)
      - Clinical Data (17)
      - Demographics (17)
      - Ernaehrungsfragebogen (17)
        - Food Groups (17)
        - Naehrstoffe (17)
          - Alkohol (17)
            - 123 Alkohol [g/Tag] (17)
          - Ballaststoffe (17)
            - Wasserloesl Ballaststoffe (17)
            - wasserunloesl. Ballaststoffe (17)
            - 123 Ballaststoffe [g/Tag] (17)
          - Eiweiss (17)
          - Fett (17)
          - Gesamtenergie (17)
          - Kochsalz (17)
            - 123 Kochsalz [g/Tag] (17)
          - Kohlenhydrate (17)
          - Kuechenabfall (17)
          - Mineralstoffe (17)
          - Organische Saeuren (Vitamine) (17)
          - Wasser (17)
            - 123 Wasser [g/Tag] (17)
      - Expression Data (Multiple Sclerosis Baranz
      - Fragebogen (generierte Daten) (17)
      - Microbiome Mice (sample data) (17)
  - sysINFLAME_Test (17)

Analysis

Analysis: Correlation Analysis

Cohorts:
Subset 1: (\Internal Studies\sysINFLAME Demographics\ )

**Variable Selection**

Drag two or more **numerical** concepts from the tree into the box below that you wish to generate correlation statistics on.

...\Alkohol [g/Tag]\
...\Kochsalz [g/Tag]\

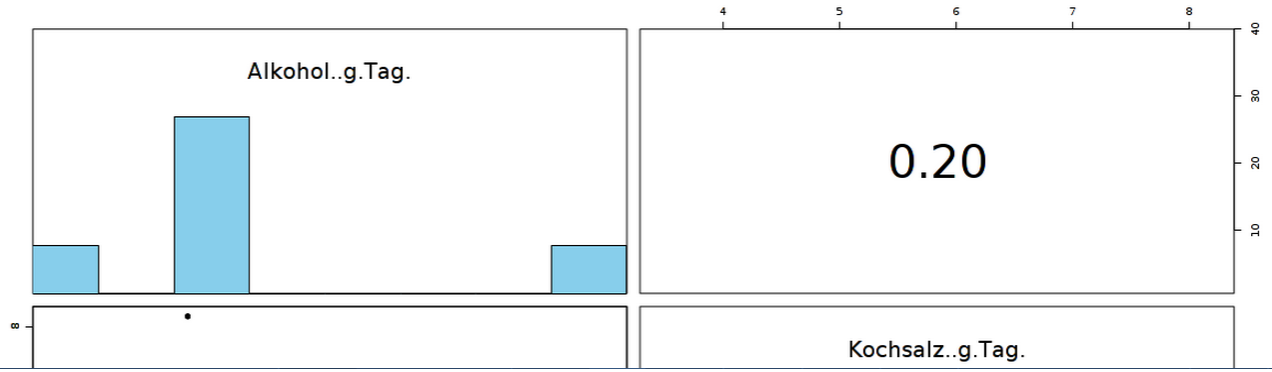
Clear

Run Correlation: By variable
Correlation Type: Spearman

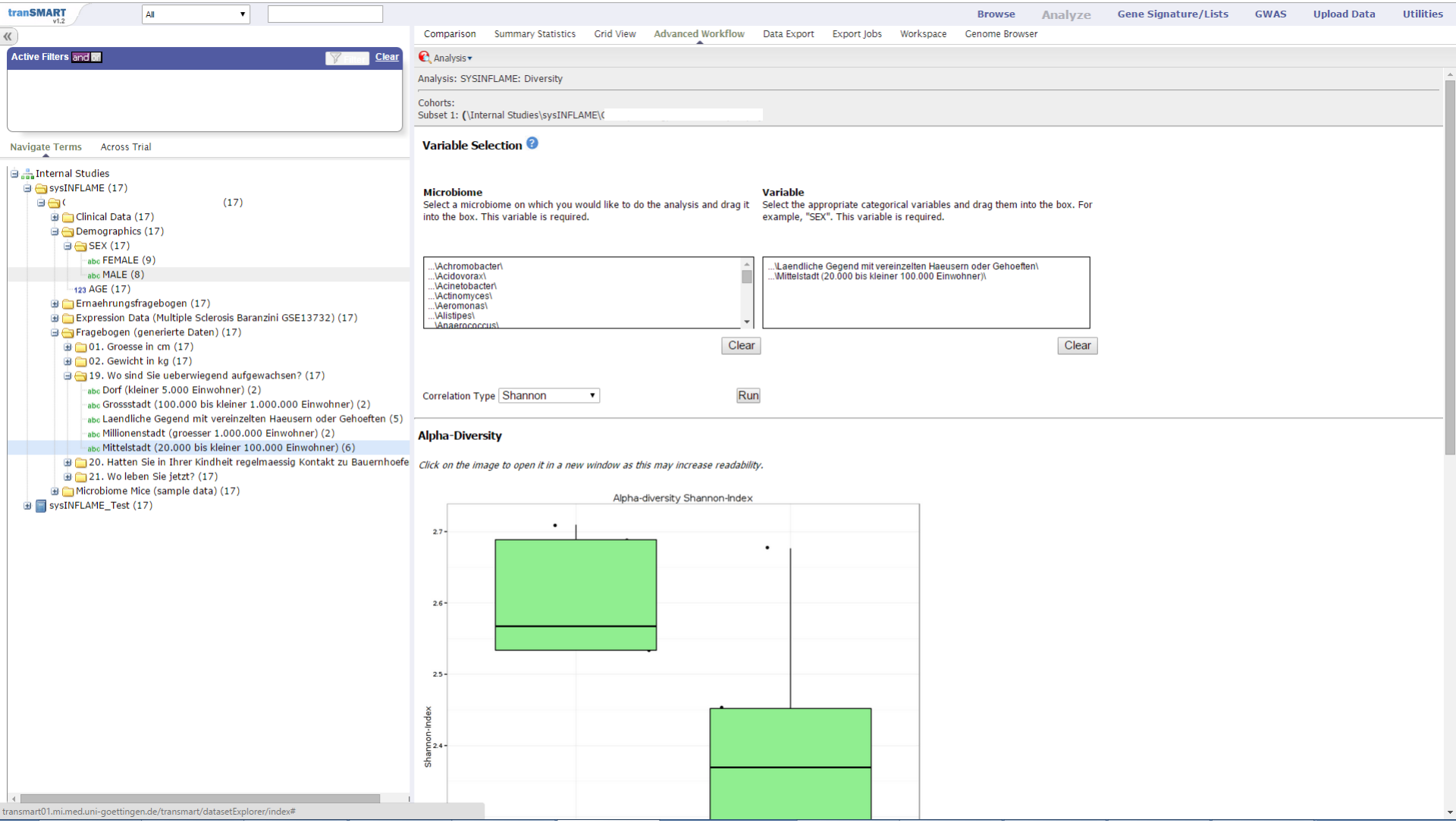
Run

**Correlation Table (p-values on top right half, correlation coefficient on bottom left)**

| | Alkohol g Tag | Kochsalz g Tag |
|---|---|---|
| **Alkohol g Tag** | 1.00000 | 0.44988 |
| **Kochsalz g Tag** | 0.19643 | 1.00000 |



# ShowCase tranSMART



tranSMART v1.2

Browse | Analyze | Gene Signature/Lists | GWAS | Upload Data | Adr

Comparison | Summary Statistics | Grid View | Advanced Workflow | Data Export | Export Jobs | Workspace | Genome Browser

Active Filters and/or | Clear

Navigate Terms | Across Trial

- Internal Studies
  - sysINFLAME (17)
    - (17)
      - Clinical Data (17)
      - Demographics (17)
      - Ernaehrungsfragebogen (17)
      - Fragebogen (generierte Daten) (17)
        - 01. Groesse in cm (17)
          - 123 01. Groesse in cm (17)
        - 02. Gewicht in kg (17)
        - 19. Wo sind Sie ueberwiegend aufgev
        - 20. Hatten Sie in Ihrer Kindheit regelı
        - 21. Wo leben Sie jetzt? (17)
      - Microbiome Mice (sample data) (17)

Analysis

Analysis: sysINFLAME: Outlier

Cohorts:
Subset 1: (\Internal Studies\sysINFLAME\ )

**Variable Selection**

Drag **numerical** concepts from the tree into the box below that you wish to generate outlier statistics on.

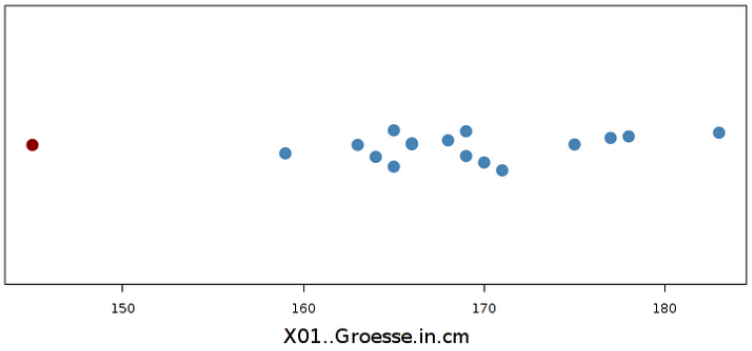
...\01. Groesse in cm\

Clear

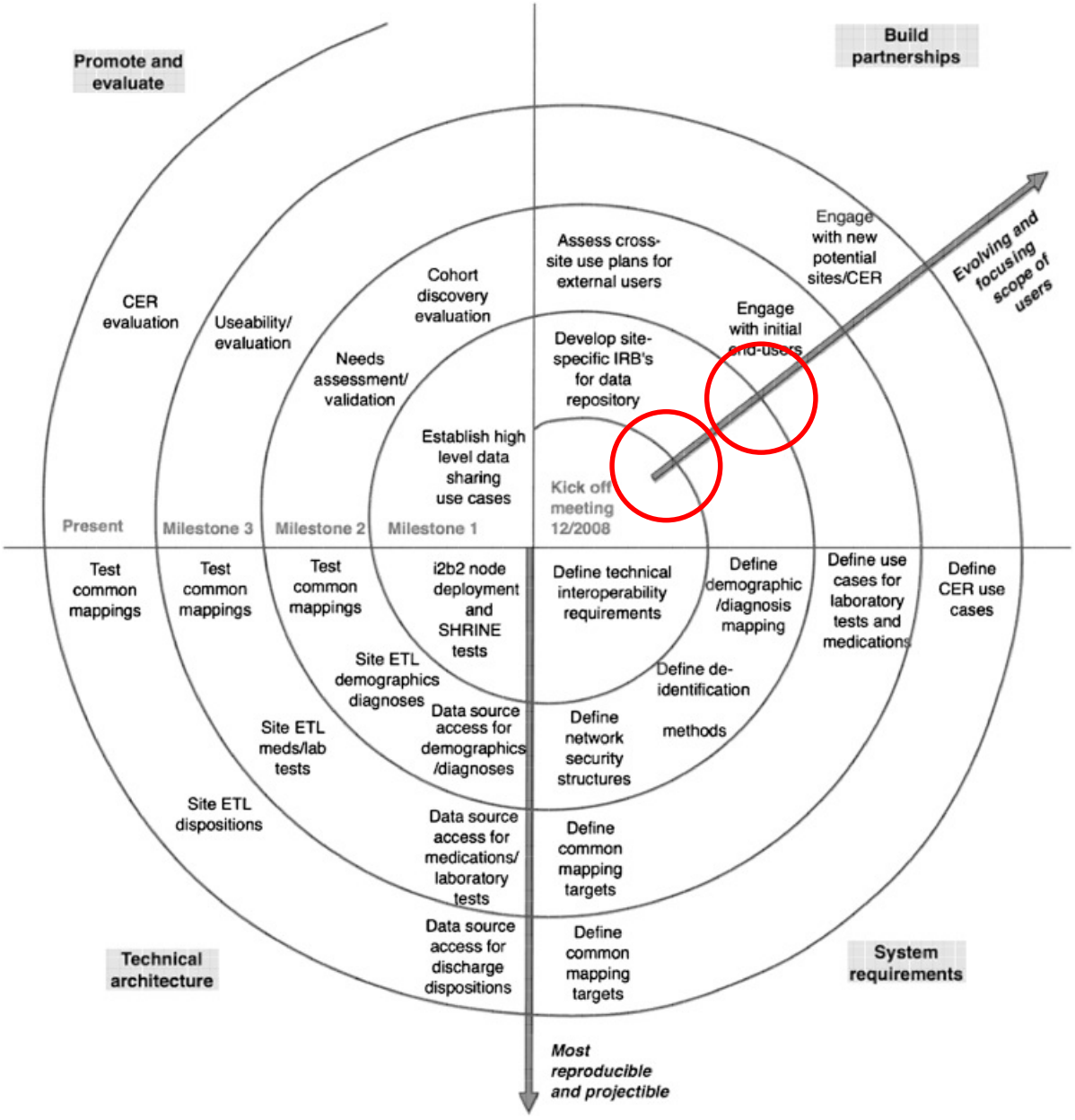
Run

**Outlier**

| Spaltenname | Samples | Extremwerte | Methode | Bemerkung |
|---|---|---|---|---|
| X01..Groesse.in.cm | 1001 | 145 | Grubbs | |

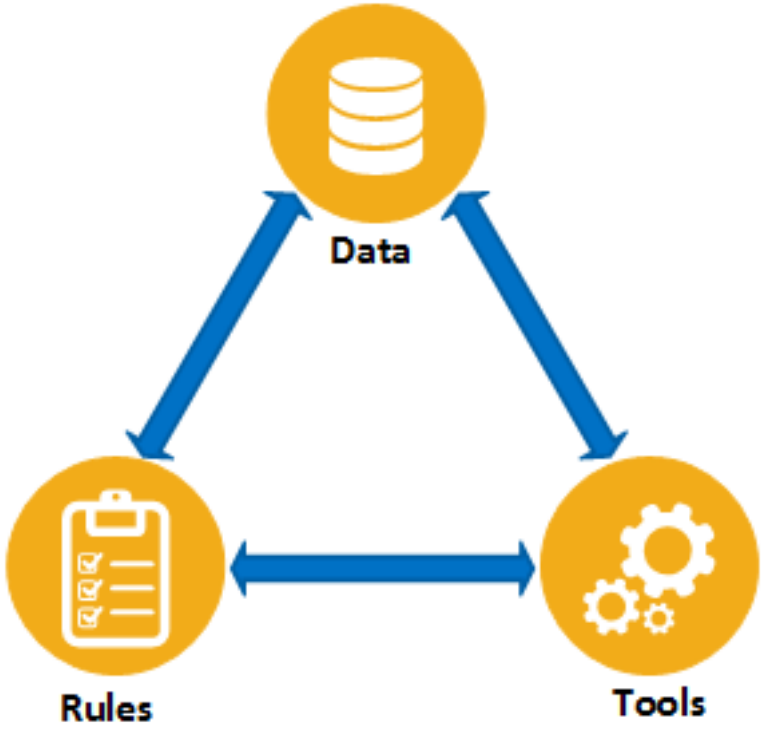



Download raw R data




Anderson, N., A. Abend, et al. (2012). "Implementation of a deidentified federated data network for population-based cohort discovery." J Am Med Inform Assoc **19**(e1): e60-67.

Sifting and Cleaning
Phenotyping!




Use and Access
Privacy Policy

Network/Server/Storage
Integrate
Ontology support
**Tool sustainability**

# i2b2 und tranSMART

- I2b2-TMF Kooperation seit 2009, Toolkits
- Entwicklung tranSMART auf älterem i2b2-Stand
- „Cousins“
- tranSMART Foundation ermöglicht externe Entwicklungsaufträge, schnelle Releasezyklen

ABER

- Modifier fehlen (!!!) derzeit in tranSMART
- Re-Integration angestrebt (Avillach, Murphy)
- Zeitschiene?

# Programm

*(Stand: 03.08.2016)*

## 10.15 Uhr Grundlegende Konzepte

- Begrüßung

5' *Dr. Johannes Drepper (TMF), Prof. Dr. Ulrich Sax (Universitätsmedizin Göttingen), Matthias Löbe (Universität Leipzig)*

- Einführung in die tranSMART-Plattform

15'+3' *Prof. Dr. Ulrich Sax (Universitätsmedizin Göttingen)*

- tranSMART Architecture and Roadmap

45'+15' *Kees van Bochove (CIO The Hyve, tranSMART-RoadMap-Presentation at tranSMART-Meeting 2015)*

- Open Data Analytics – Gefahren durch p-Hacking

15'+3' *Prof. Dr. Ulrich Sax (Universitätsmedizin Göttingen)*

## 12.00 Uhr Mittagspause

## 13.00 Uhr Praktische Arbeit

- Klick-a-thon: Durchgehen eines vorbereiteten praktischen Szenarios durch die Workshopteilnehmer
- Lösung von Aufgaben in Eigenarbeit

UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

# Open Data Analytics – Gefahren durch p-Hacking

**TMF-Workshop**

Aufbereitung und Analyse von Daten in phänotypischen und genotypischen Forschungsdatenbanken mit tranSMART

Berlin, 05.08.2016

Prof. Dr. Ulrich Sax

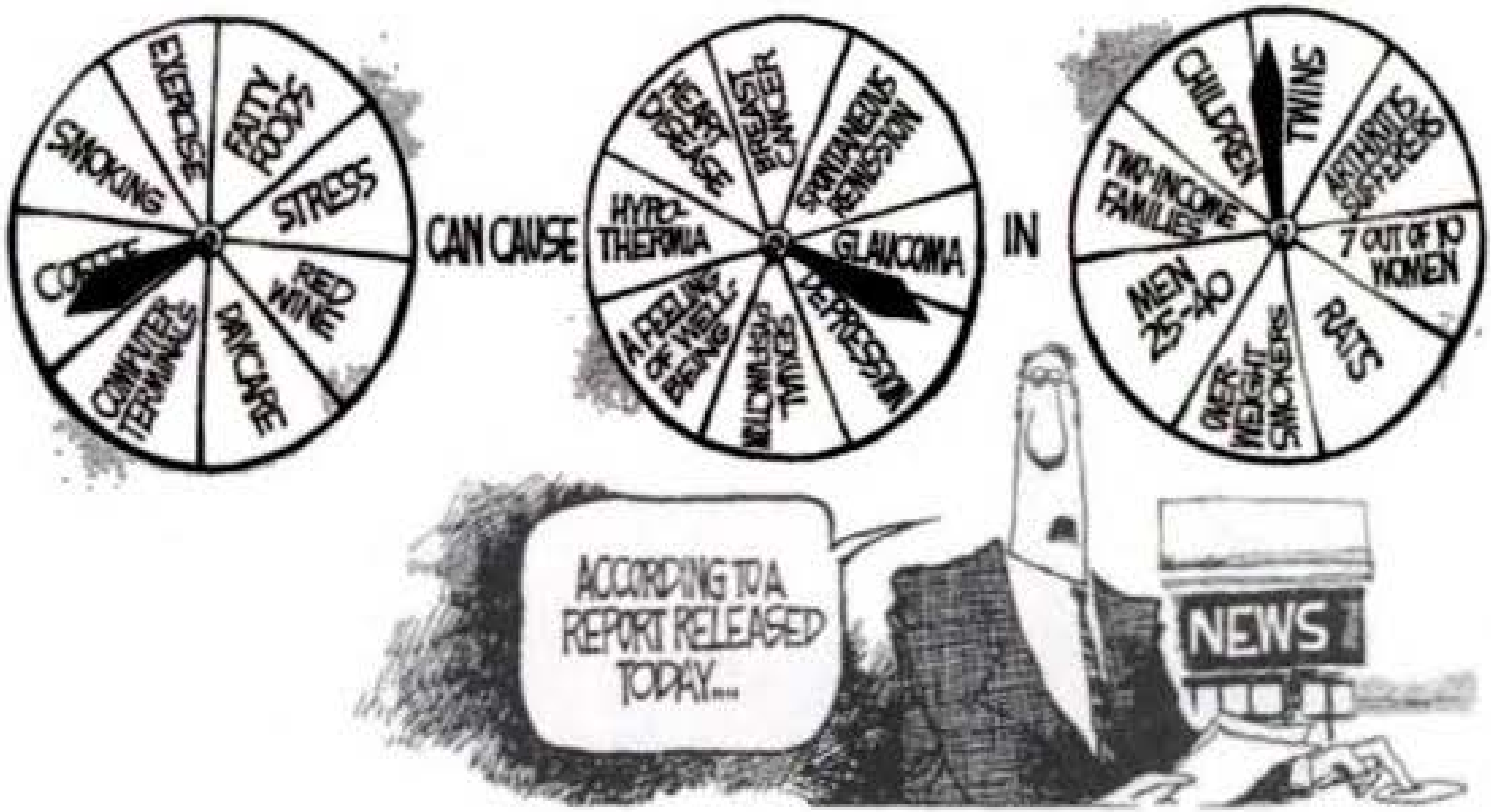
Department of Medical Informatics, WG Infrastructure for Translational Research

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

# Today's Random Medical News

from the New England Journal of Panic-Inducing Gobbledygook

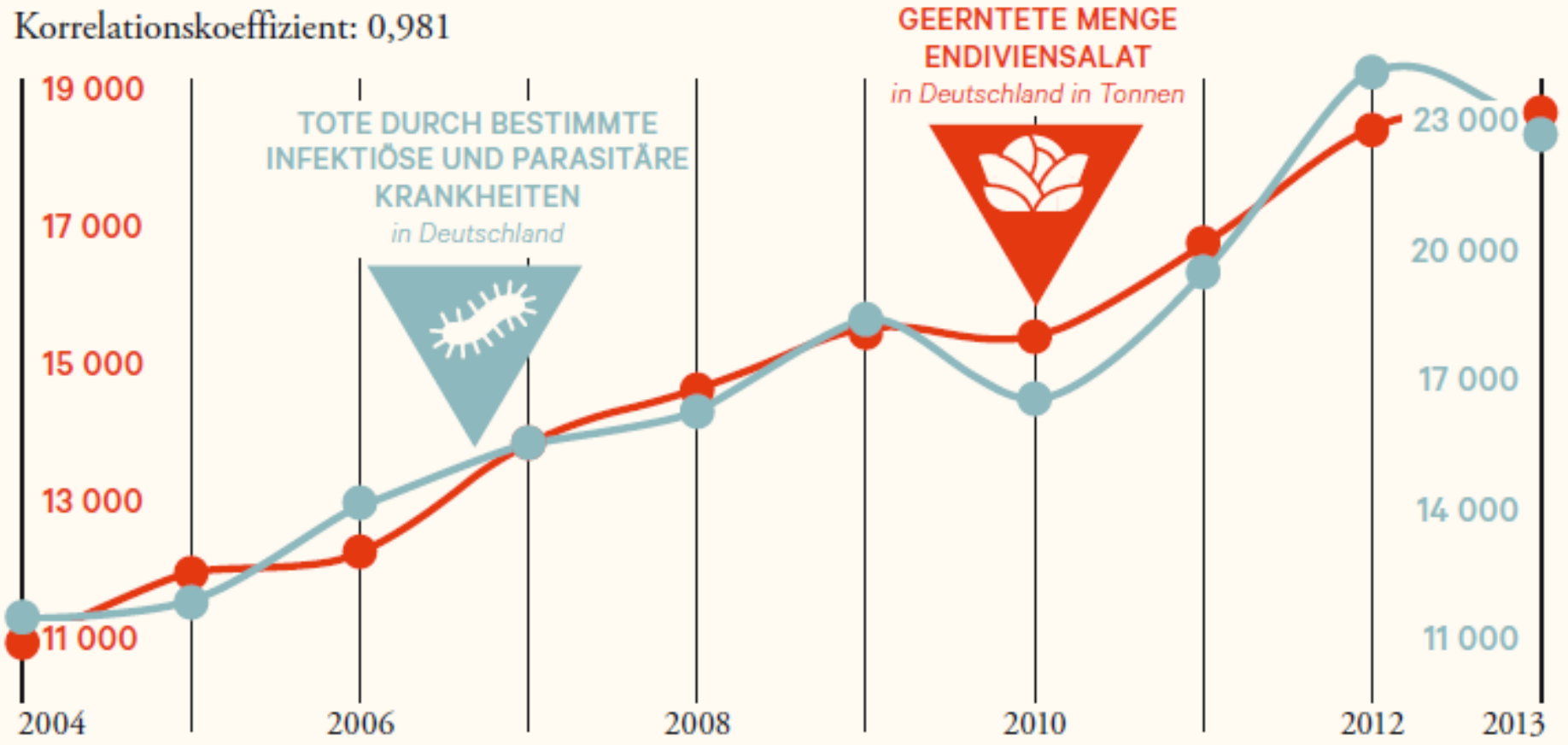



**Jim Borgman**
**The Cincinnati Enquirer**
King Features Syndicate

# GIFTIGES GRÜN

Ist Endiviensalat verantwortlich für Infektionskrankheiten?

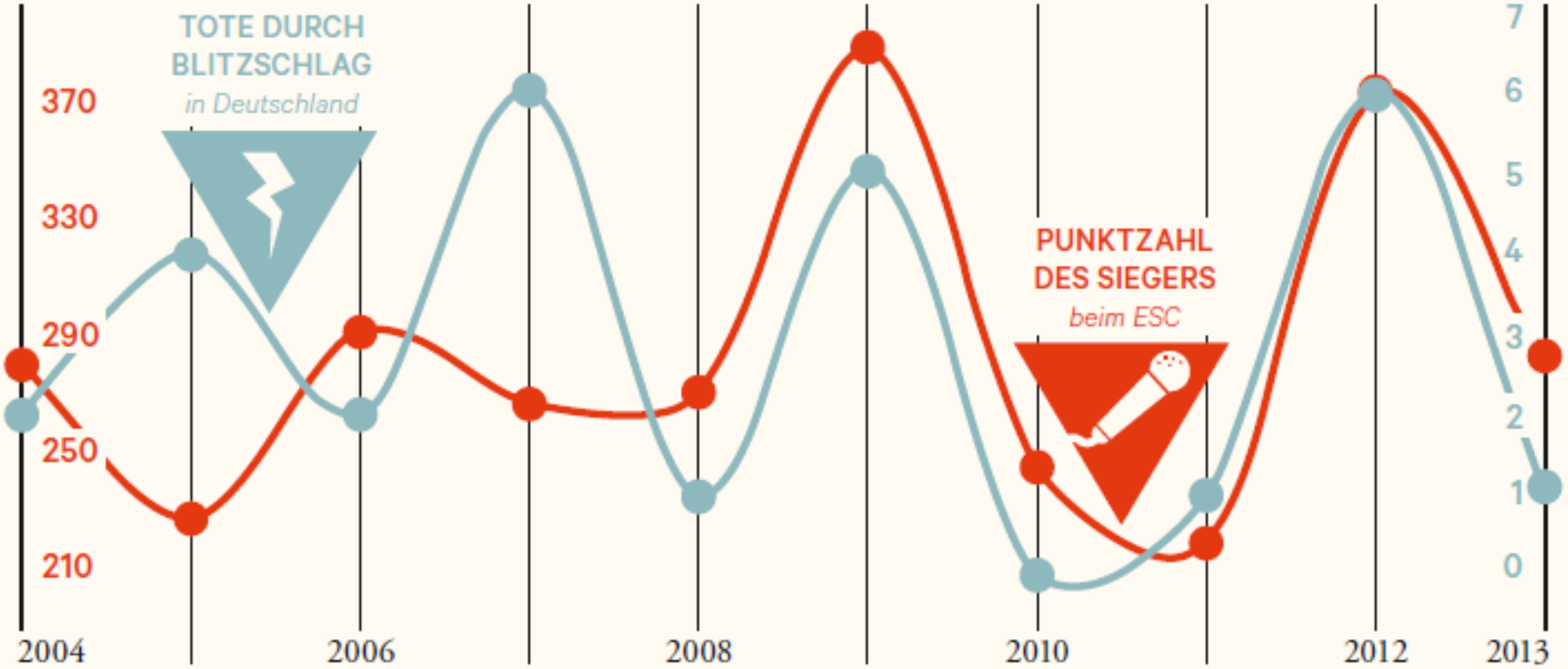
Korrelationskoeffizient: 0,981



# EINSCHLAGENDER ERFOLG

Was hat die Punktzahl des Siegers beim Eurovision Song Contest mit Toten durch Blitzschlag zu tun?

Korrelationskoeffizient: 0,571

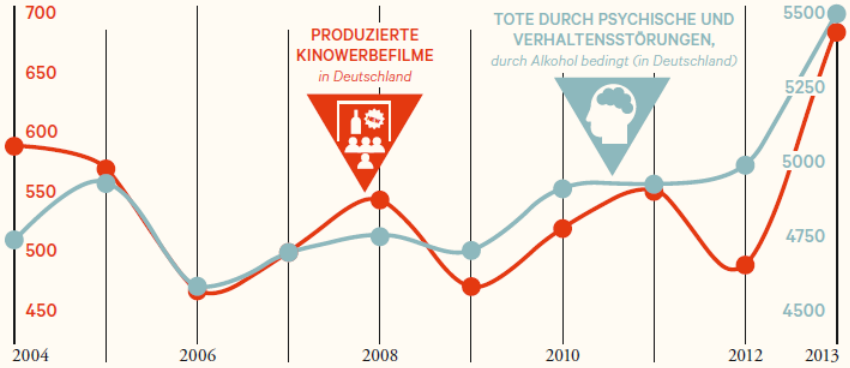


# Es ist nicht so, wie es scheint

Guck mal, die Grafik! Wie ähnlich die beiden Kurven verlaufen, da muss es doch einen Zusammenhang geben ... Nein, muss es nicht. Aber solche Trugschlüsse sind ein häufiges Phänomen. Wir haben in den unterschiedlichsten Statistiken nach solchen Scheinkorrelationen gesucht. Und viel dabei gelacht

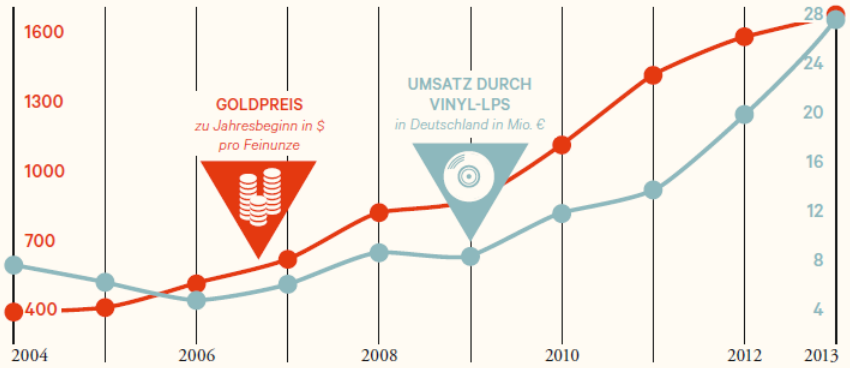
## IRREFÜHRENDE WERBUNG

Können Werbefilme verantwortlich sein für den Tod von Menschen?
Korrelationskoeffizient: 0,792 (siehe Erklärung unten rechts)




## GOLDENE SCHALLPLATTEN

Treibt der Verkauf von Vinylplatten den Goldpreis in die Höhe?
Korrelationskoeffizient: 0,905



# FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION

# FOOLING OURSELVES

## HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION. BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.

BY REGINA NUZZO

In 2013, five years after he co-authored a paper showing that Democratic candidates in the United States could get more votes by moving slightly to the right on economic policy[1], Andrew Gelman, a statistician at Columbia University in New York City, was chagrined to learn of an error in the data analysis. In trying to replicate the work, an undergraduate student named Yang Yang Hu had discovered that Gelman had got the sign wrong on one of the variables.

Gelman immediately published a three-sentence correction, declaring that everything in the paper's crucial section should be considered wrong until proved otherwise.

# Gefahren durch p-Hacking

- Was ist p-Hacking?
- Warum ist es gefährlich?

# p-Wert, statistische Signifikanz

UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

- Statistisch signifikant, auch statistisch bedeutsam, wird das Ergebnis einer Untersuchung genannt, wenn die statistische Auswertung der Daten ergibt, dass die Wahrscheinlichkeit für die Annahme, die festgestellten Unterschiede zwischen Messgrößen oder Variablen seien durch Zufall derart zustande gekommen, einen zuvor als Signifikanzniveau festgelegten Schwellenwert nicht überschreitet.

Quelle: https://de.wikipedia.org/wiki/Statistische_Signifikanz

1. Stephen Stigler: *Fisher and the 5% level.* In: *CHANCE.* Band 21, Nr. 4, Springer, New York Dezember 2008, S. 12.

# p-Wert, statistische Signifikanz

UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG

- Die obere Grenze für die Irrtumswahrscheinlichkeit, also jener Wert, den man für die Wahrscheinlichkeit eines Fehlers 1. Art noch eben zu akzeptieren bereit ist, heißt Signifikanzniveau.

- **Grundsätzlich ist dies frei wählbar; häufig wird ein Signifikanzniveau von 5 % verwendet ($p \leq 0,05$)** [1]

- In der Praxis bedeutet dieses Kriterium, dass im Schnitt eine von 20 Untersuchungen, bei denen die Nullhypothese richtig ist (z. B. ein Medikament tatsächlich wirkungslos ist), zu dem Schluss kommt, sie sei falsch (z. B. behauptet, das Medikament erhöhe die Heilungschancen).

Quelle: https://de.wikipedia.org/wiki/Statistische_Signifikanz

1. Stephen Stigler: *Fisher and the 5% level.* In: *CHANCE.* Band 21, Nr. 4, Springer, New York Dezember 2008, S. 12.

https://de.wikipedia.org/wiki/Statistische_Signifikanz

- Eine heuristische Motivation des Wertes 5 % ist wie folgt: Eine normalverteilte Zufallsgröße nimmt nur mit einer Wahrscheinlichkeit von weniger als (≤) 5 % einen Wert an, der sich vom Erwartungswert um mehr als die zweifache Standardabweichung unterscheidet:
- Bei einem p-Wert von kleiner oder gleich 5 % **[p ≤ 0,05]** spricht man von einem signifikanten […] Ergebnis.



# SPURIOUS CORRELATION: A CAUSAL INTERPRETATION*

HERBERT A. SIMON
*Carnegie Institute of Technology*

To test whether a correlation between two variables is genuine or spurious, additional variables and equations must be introduced, and sufficient assumptions must be made to identify the parameters of this wider system. If the two original variables are causally related in the wider system, the correlation is "genuine."

"SCIENCE IS AN ONGOING RACE BETWEEN OUR INVENTING WAYS TO FOOL OURSELVES, AND OUR INVENTING WAYS TO AVOID FOOLING OURSELVES."

Saul Perlmutter, an astrophysicist at the University of California, Berkeley

Failure to understand our own biases has helped to create a crisis of confidence about the reproducibility of published results, says statistician John Ioannidis, co-director of the Meta-Research Innovation Center at Stanford University in Palo Alto, California. The issue goes well beyond cases of fraud. Earlier this year, a large project that attempted to replicate 100 psychology studies managed to reproduce only slightly more than one-third[2]. In 2012, researchers at biotechnology firm Amgen in Thousand Oaks, California, reported that they could replicate only 6 out of 53 landmark studies in oncology and haematology[3]. And in 2009, Ioannidis and his colleagues described how they had been able to fully reproduce only 2 out of 18 microarray-based gene-expression studies[4].

**Regina Nuzzo** *is a freelance writer in Washington DC.*

1. Gelman, A. & Cai, C. J. *Ann. Appl. Stat.* **2,** 536–549 (2008).
2. Open Science Collaboration. *Science* http://dx.doi.org/10.1126/science.aac4716 (2015).
3. Begley, C. G. & Ellis, L. M. *Nature* **483,** 531–533 (2012).
4. Ioannidis, J. P. A. *et al. Nature Genet.* **41,** 149–155 (2009).

From: **Finding the Missing Link for Big Biomedical Data**

JAMA. 2014;311(24):2479-2480. doi:10.1001/jama.2014.4228

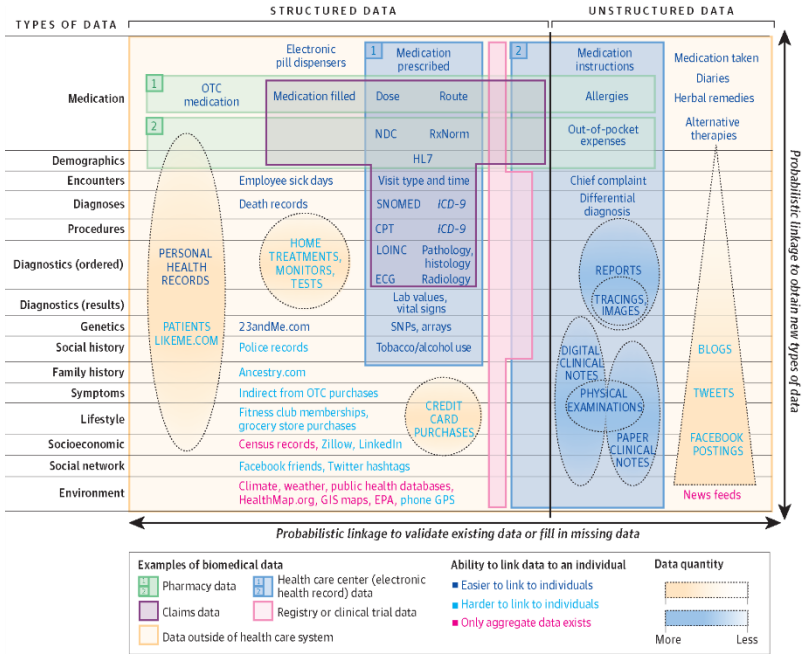
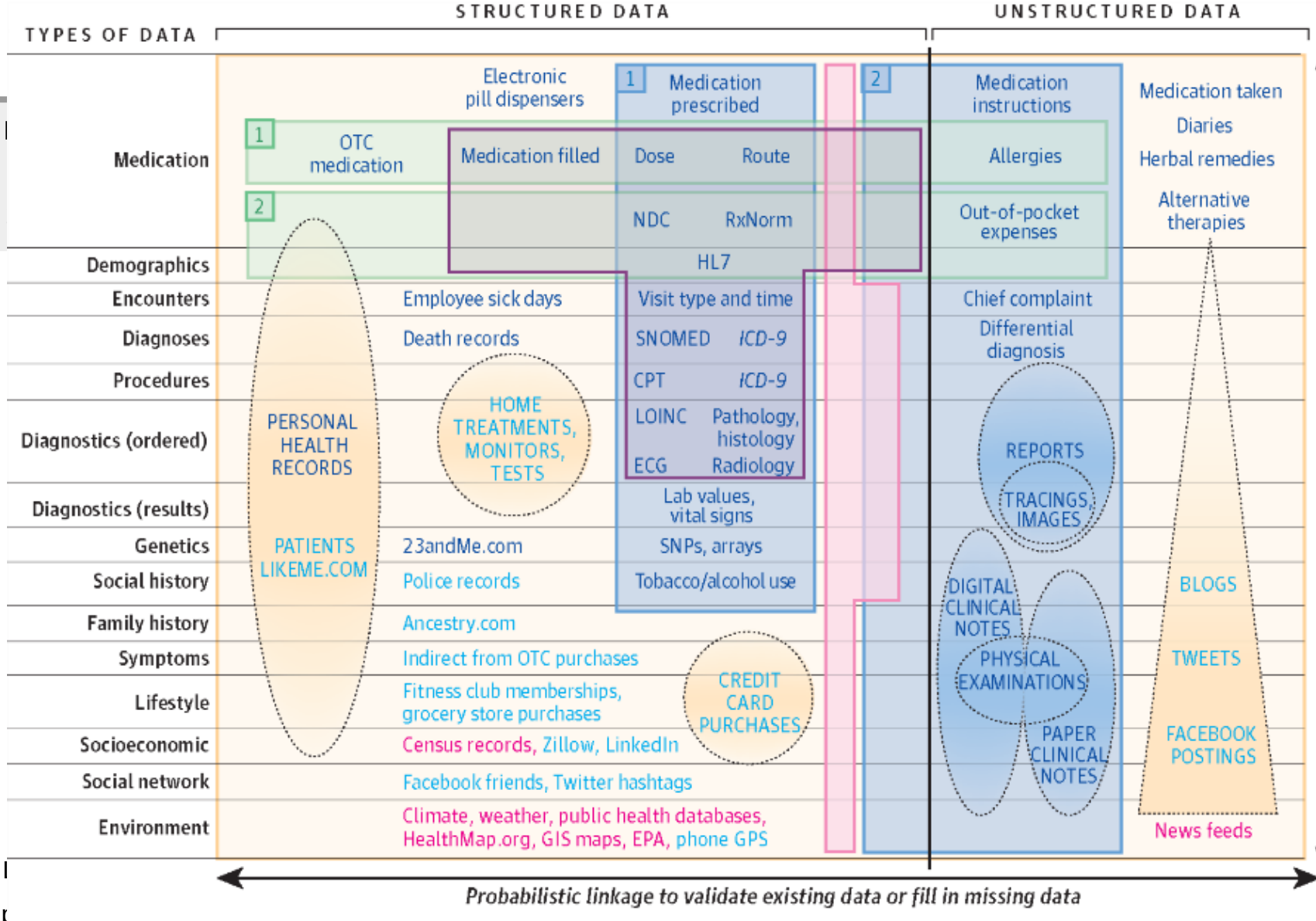



Figure Legend:

The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health CareCPT indicates current procedural terminology; ECG, electrocardiography; EPA, US Environmental Protection Agency; GIS, geographic information systems; GPS, global positioning system; HL7, Health Level 7 coding standard; ICD-9, Institutional Classification of Diseases, Ninth Revision; LOINC, Logical Observation Identifiers Names and Codes; NDC, National Drug Code; OTC, over-the-counter; SNOMED, Systematized Nomenclature of Medicine; SNP, single-nucleotide polymorphism.

From:

JAMA.




Figure

The Tap
indicate
informa
Disease
counter

# Lösungsansätze

- **Correlation does not replace causation!**
- R. Nuzzo Nature 2015:
  - "One solution that is piquing interest revives an old tradition: explicitly considering competing hypotheses, and if possible working to develop experiments that can distinguish between them. "
  - "Furthermore, when scientists make themselves explicitly list alternative explanations for their observations, they can reduce their tendency to tell just-so stories."
- Zur Vertiefung:
  - http://www.laborjournal.de/editorials/981.lasso
  - http://www.labtimes.org/editorial/e_654.lasso

# Eine neue Wissenschaft-(lichkeit)?

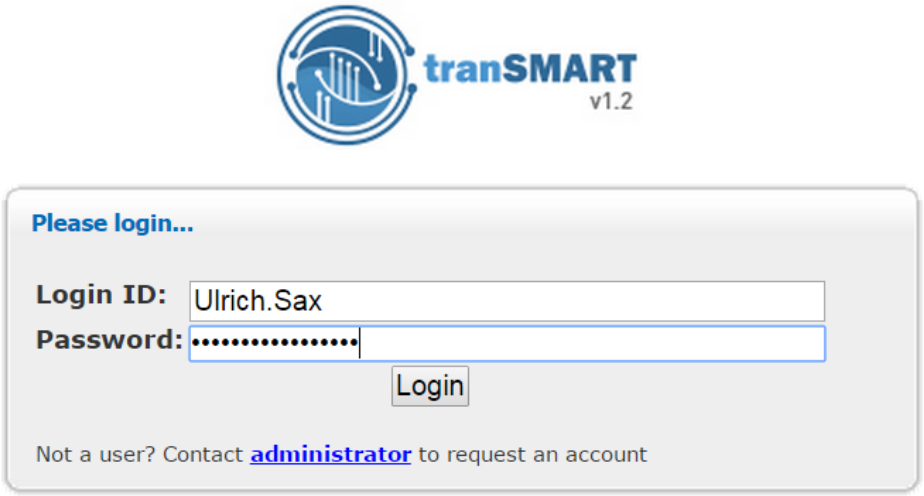
**Big Data, Innovation, Personalisierte Medizin und Co. – Sind dies die Markenzeichen einer neuen Wissenschaft-(lichkeit) in der Medizin? Ein Essay von Gerd Antes, Freiburg.**



Wenn man Editorials, Kommentare oder Meinungsartikel in wissenschaftlichen Zeitschriften liest und ihnen glaubt, so stehen wir am Beginn eines goldenen Zeitalters für Patienten und Gesunde. Patienten werden viel früher und fehlerfrei diagnostiziert und dann mit personalisierter Medizin zielgenau, wirksam und nebenwirkungsfrei behandelt. Gesunde kommen erst gar nicht in die Gefahr, weil sie durch perfekte Vorsorge vor dem Schritt geschützt werden, überhaupt erst krank zu werden.

Erreicht wird dies mit „Systemmedizin", in der die biologischen Mechanismen der Krankheitsentstehung unter Nutzung der Methoden von „Omics"-Forschung, Systembiologie, Informatik und Netzwerktheorie besser verstanden werden – was dann wiederum durch „Translation" scheinbar wie von selbst in der personalisierten oder individualisierten Medizin genutzt wird. Überall dabei ist „Big Data" als Universalwerkzeug; Hindernisse und Barrieren gibt es praktisch nicht. Erforderlich sind nur unbegrenzte Rechnerleistung, Datenerfassung ohne Behinderung samt deren Speicherung in grenzenlosen Clouds, sowie

# ShowCase tranSMART




**DISCLAIMER**

By logging in to this application I acknowledge that, according to the **DFG-Memorandum on Safeguarding Good Scientific Practice**, I am obliged to ensure good scientific practice by consulting a (bio)medical informaticist (data management) and a statistician or epidemiologist (analysis) prior to publishing any results obtained (partly or in full) through the use of this application.

Noncompliance could lead to a severe lack in reproducibility and may constitute a case of scientific misconduct.

# Programm

(Stand: 03.08.2016)

## 10.15 Uhr Grundlegende Konzepte

- Begrüßung

5' *Dr. Johannes Drepper (TMF), Prof. Dr. Ulrich Sax (Universitätsmedizin Göttingen), Matthias Löbe (Universität Leipzig)*

- Einführung in die tranSMART-Plattform

15'+3' *Prof. Dr. Ulrich Sax (Universitätsmedizin Göttingen)*

- tranSMART Architecture and Roadmap

45'+15' *Kees van Bochove (CIO The Hyve, tranSMART-RoadMap-Presentation at tranSMART-Meeting 2015)*

- Open Data Analytics – Gefahren durch p-Hacking

15'+3' *Prof. Dr. Ulrich Sax (Universitätsmedizin Göttingen)*

## 12.00 Uhr Mittagspause

## 13.00 Uhr Praktische Arbeit

- Klick-a-thon: Durchgehen eines vorbereiteten praktischen Szenarios durch die Workshopteilnehmer
- Lösung von Aufgaben in Eigenarbeit

# Programm

(Stand: 03.08.2016)

## 13.00 Uhr Praktische Arbeit

- Klick-a-thon: Durchgehen eines vorbereiteten praktischen Szenarios durch die Workshopteilnehmer
- Lösung von Aufgaben in Eigenarbeit

## 14.15 Uhr Kaffeepause

## 14.45 Uhr Fortgeschrittene Themen

- tranSMART and beyond

20'+5' *Reinhard Schneider / Sascha Herzinger (Universität Luxemburg)*

- SmartR – Dynamic Visual Analytics in TranSMART

20'+5' *Reinhard Schneider / Sascha Herzinger (Universität Luxemburg)*

- Import und Repräsentation hochdimensionaler Daten

10'+3' *Christoph Knell (Friedrich-Alexander-Universität Erlangen-Nürnberg)*

UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

# Diskussion
# Zusammenfassung und Ausblick

**TMF-Workshop**

Aufbereitung und Analyse von Daten in phänotypischen und genotypischen Forschungsdatenbanken mit tranSMART

Berlin, 05.08.2016

Prof. Dr. Ulrich Sax

Department of Medical Informatics, WG Infrastructure for Translational Research

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

# Wrap up and Future directions

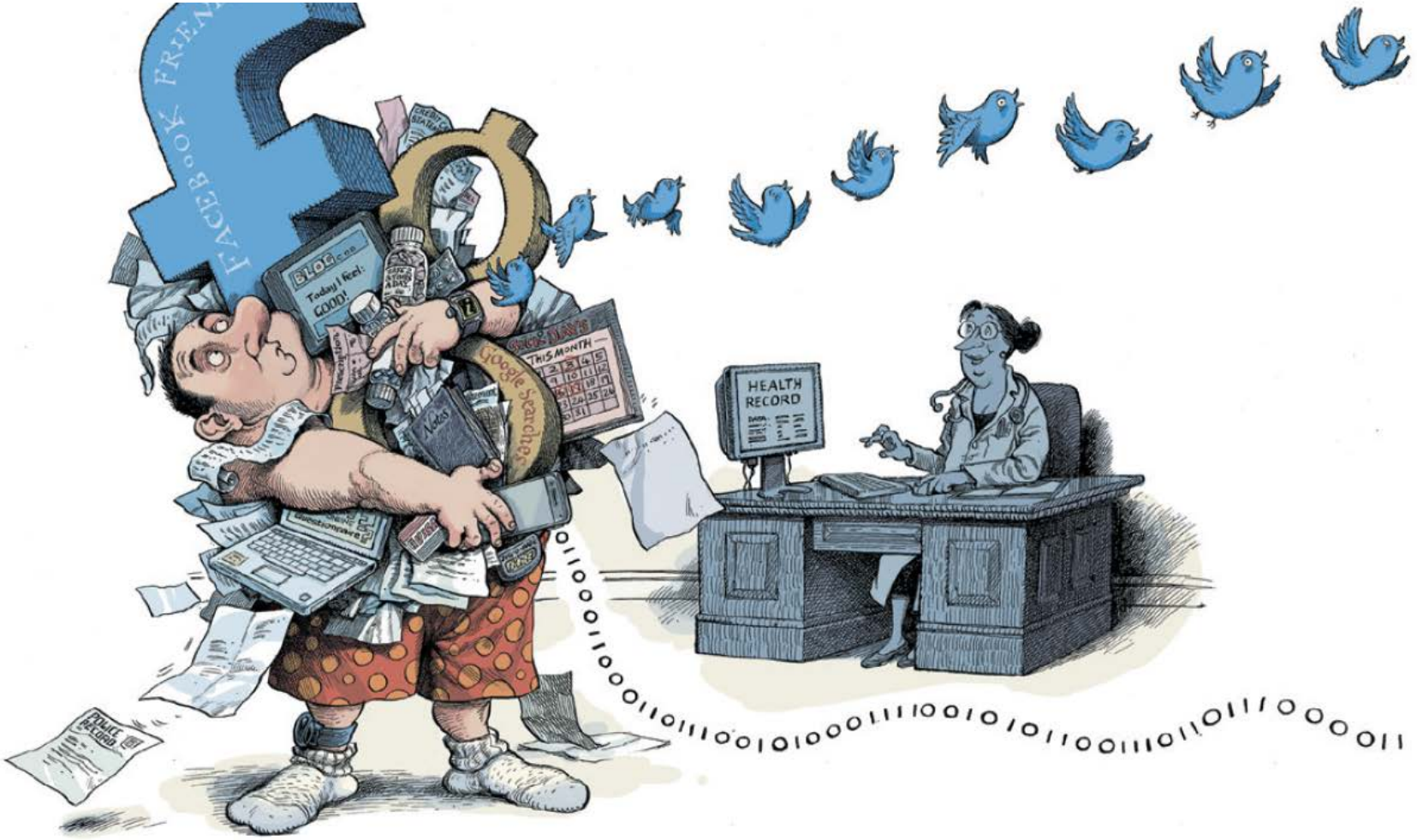
**Was nehmen Sie mit nach Hause?**

**Lösungsansätze für die Praxis**

**neue Fragen**

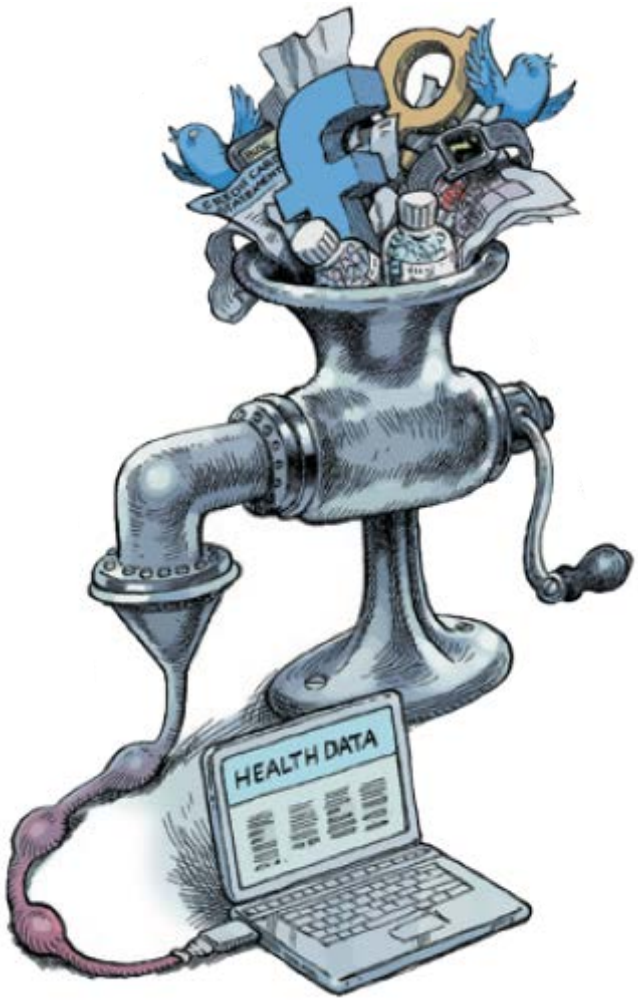
**Ideen und Ansprechpartner**

ILLUSTRATIONS BY DAVID PAR




# Make sense of health data

Develop the science of data synthesis to join up the myriad varieties of health information, insist **Julian H. Elliott**, **Jeremy Grimshaw** and colleagues.




- Pooling data
- Managing Bias
- Joining the Dots

# Make sense of health data

Develop the science of data synthesis to join up the myriad varieties of health information, insist **Julian H. Elliott**, **Jeremy Grimshaw** and colleagues.

# Data Provenance: a forgotten concept?

## A Survey of Data Provenance in e-Science

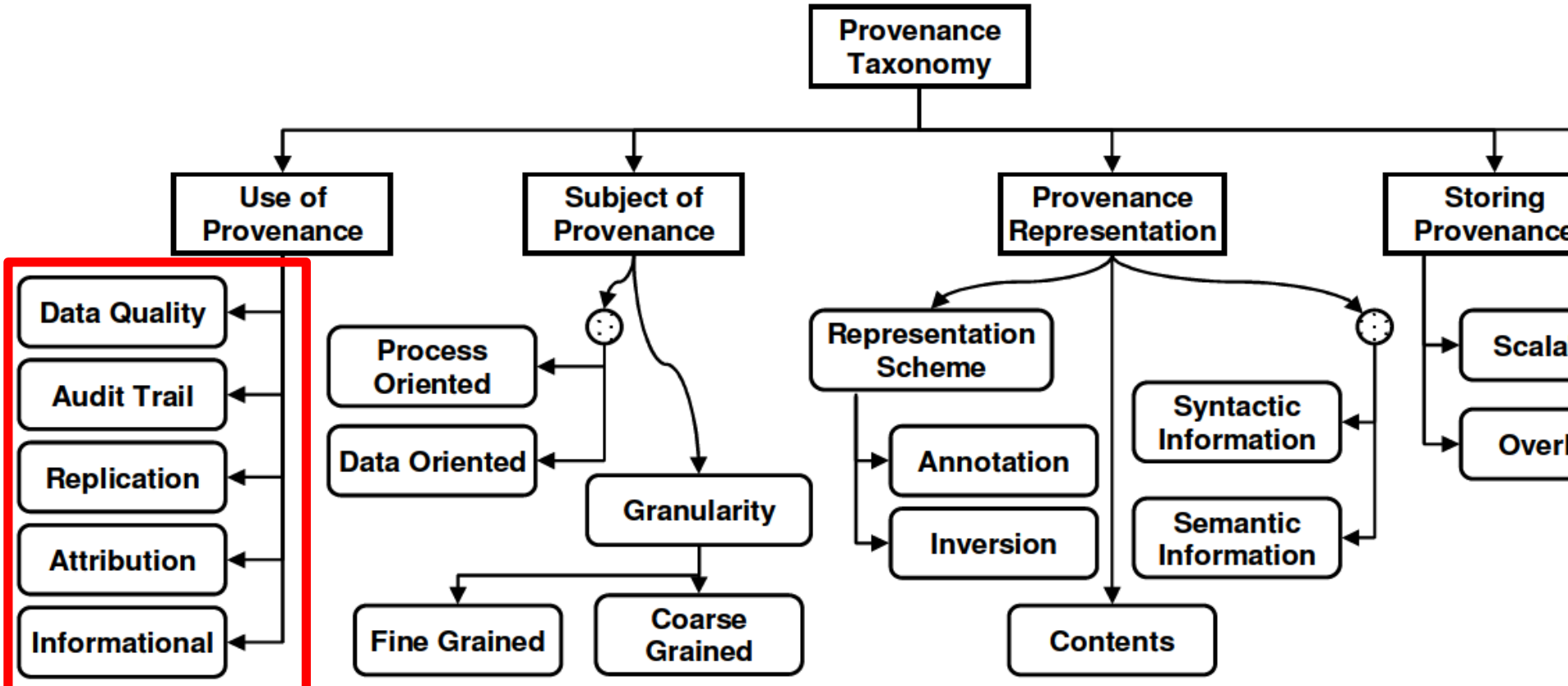
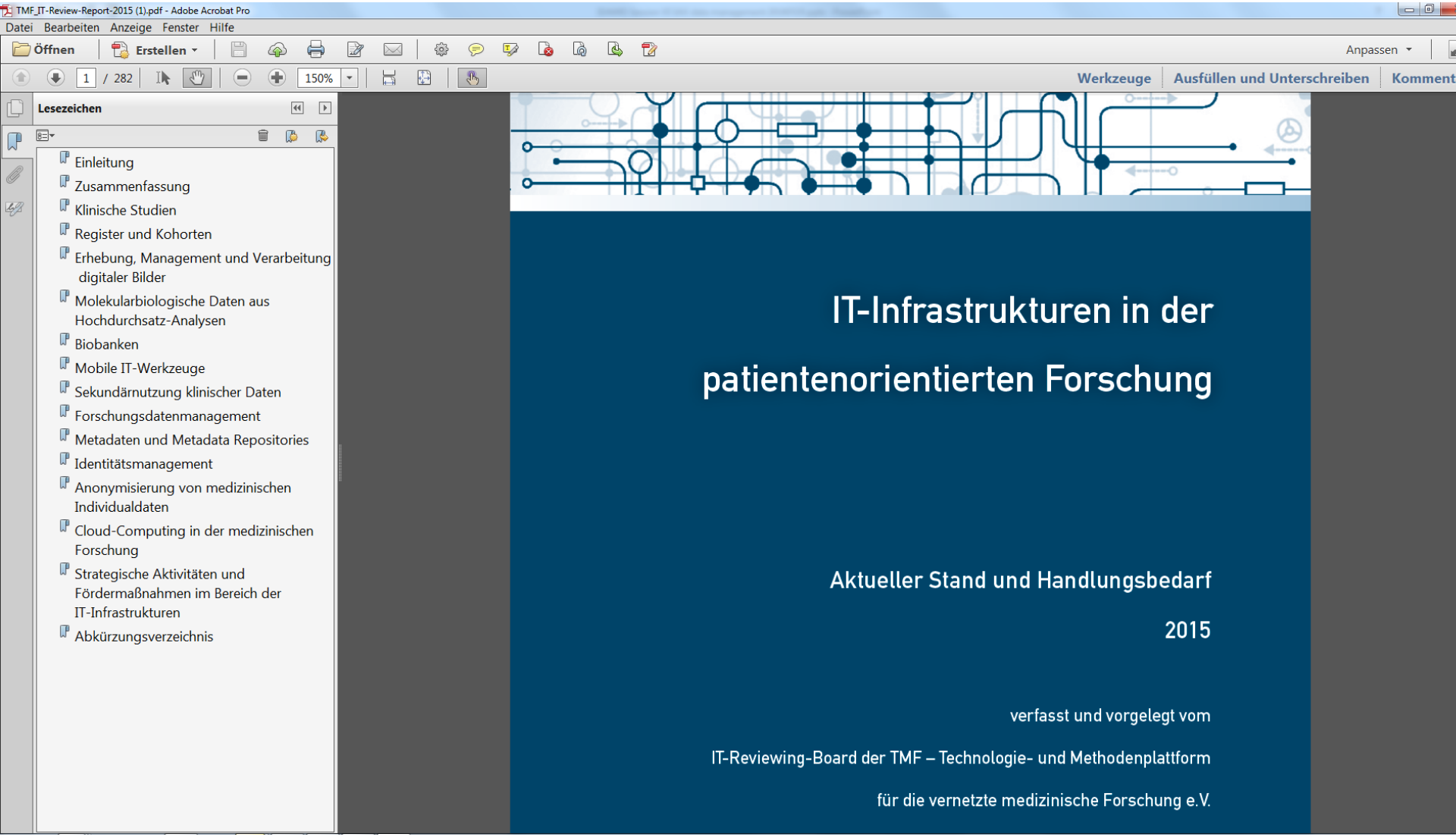



Figure 1 Taxonomy of Provenance

SIGMOD Record, Vol. 34, No. 3, Sept. 2005

# TMF IT-Report 2015/6

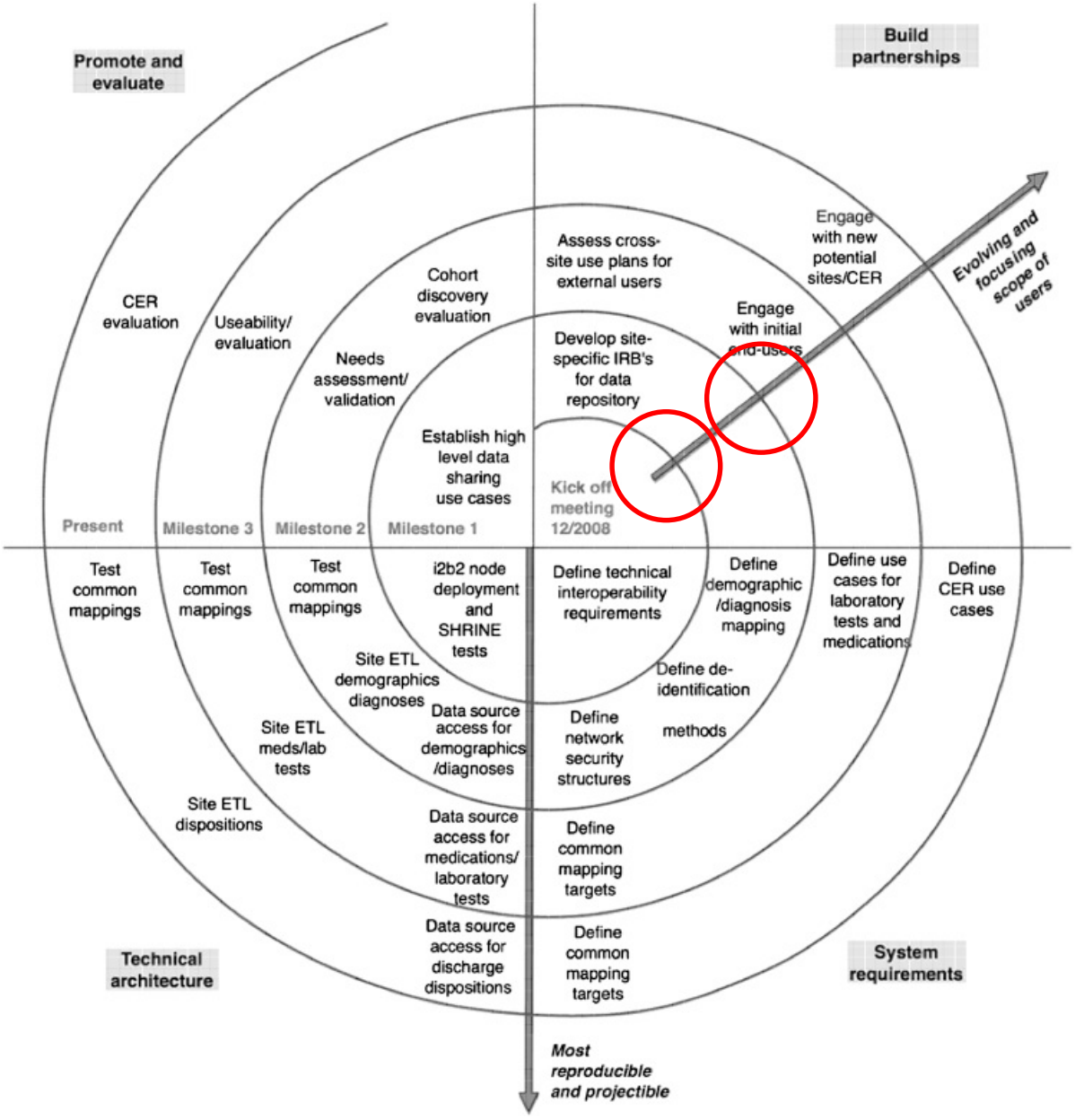


# Wrap up and Future directions

**Was nehmen Sie mit nach Hause?**

**Lösungsansätze für die Praxis**

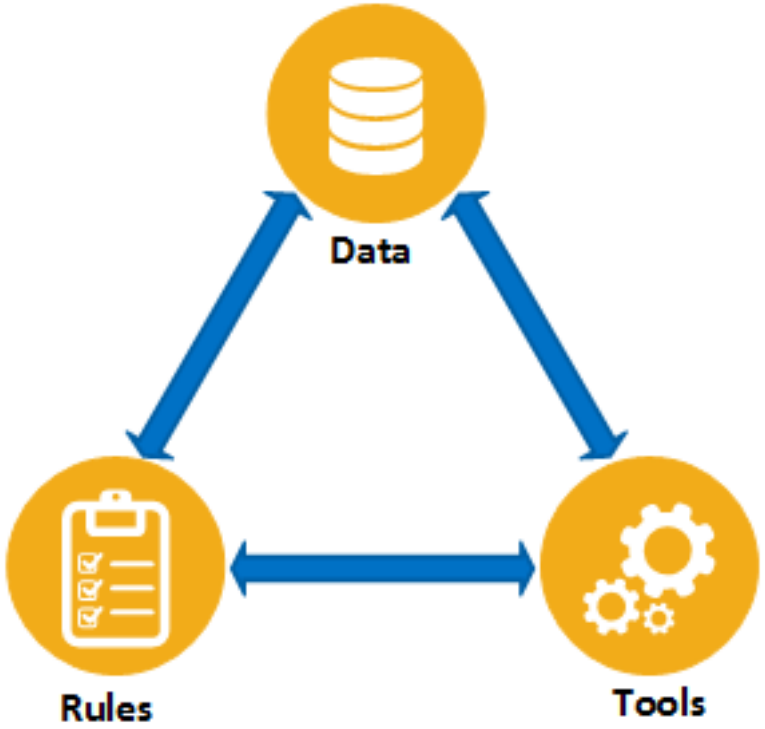



Anderson, N., A. Abend, et al. (2012). "Implementation of a deidentified federated data network for population-based cohort discovery." J Am Med Inform Assoc **19**(e1): e60-67.

# Discussion

Sifting and Cleaning
Phenotyping!




Use and Access
Privacy Policy

Network/Server/Storage
Integrate
Ontology support
**Tool sustainability**

# Aufbereitung und Analyse von Daten in phänotypischen und genotypischen Forschungs-datenbanken mit tranSMART

5. August 2016 | Berlin