Grafische Darstellung und Editierung sowie versionierbare Speicherung von Massendaten (Big Data)

Dr. Tom Gutzke
Dr. Frank Reußner

WHO'S WHO

envi-systems

envi-systems GmbH
Ing.-Büro für Umweltinformatik

Dr.-Ing. Thomas Gutzke
Geschäftsführender Gesellschafter

An der Eschollmühle 28
64297 Darmstadt

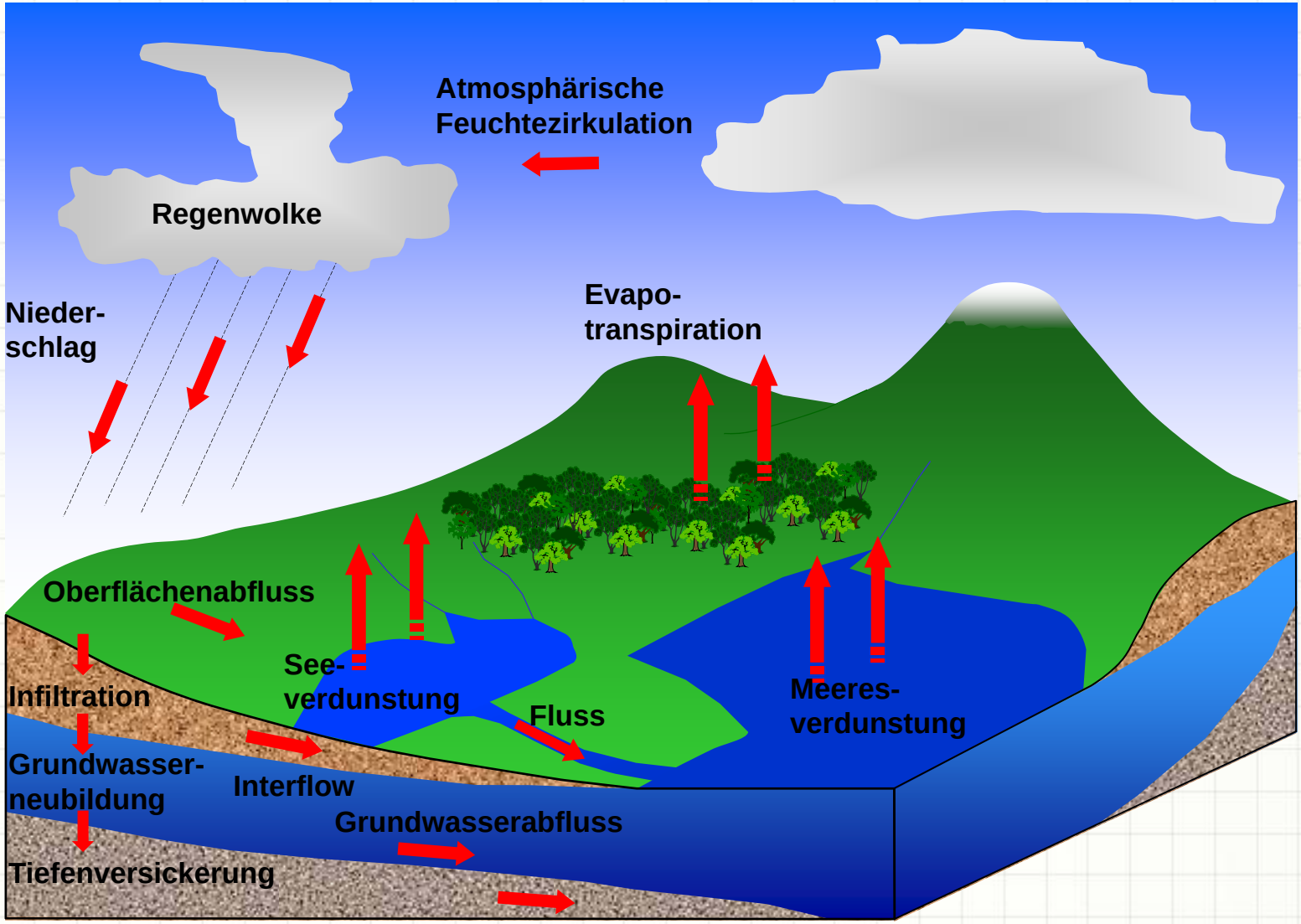
Tel.: 06151 - 9456-30
Fax.: 06151 - 9456-80
Mail: gutzke@envi-systems.com
www.envi-systems.com

Dr.-Ing.
Frank Reußner

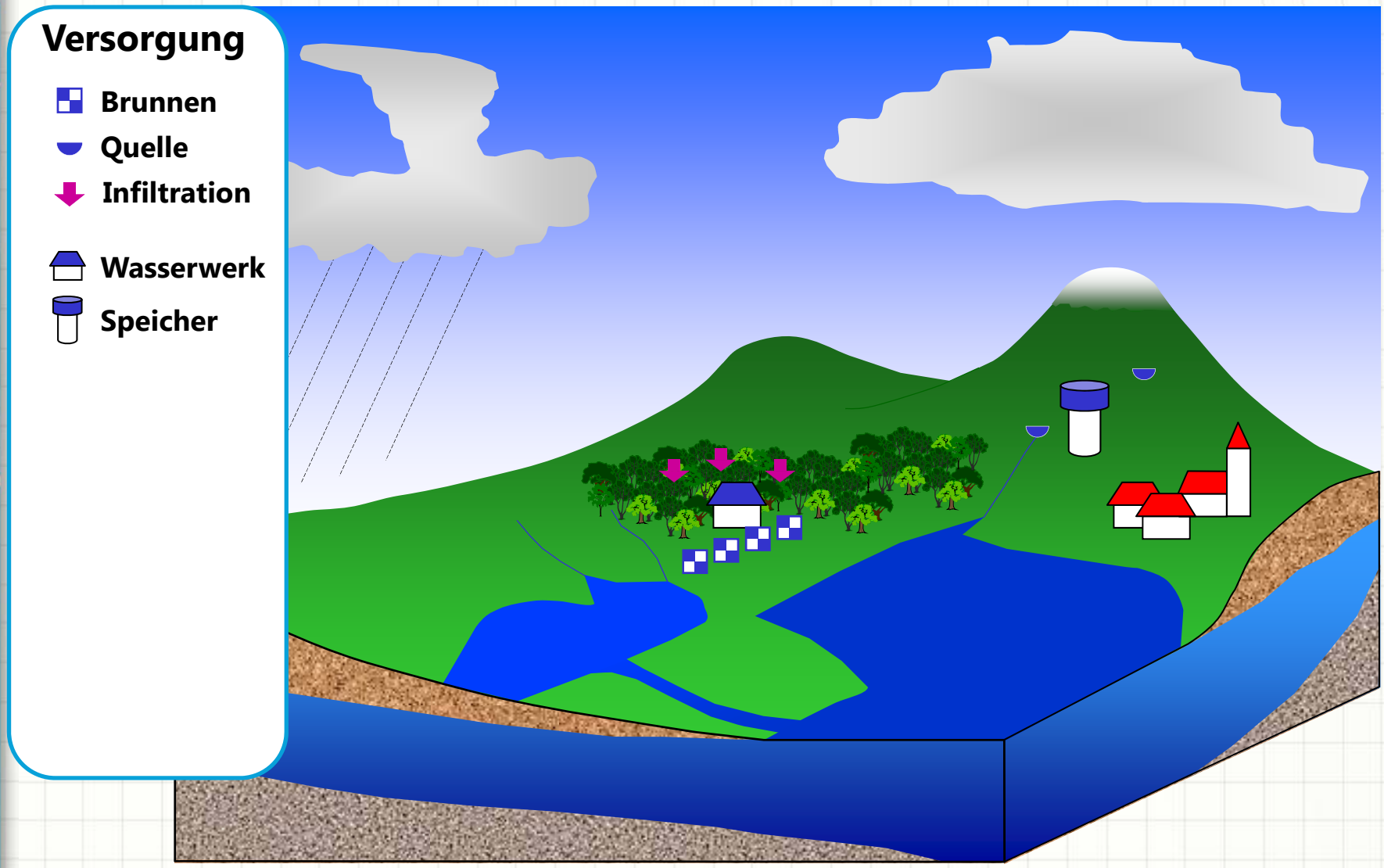
Ingenieurberatung
& Umweltinformatik

Lindenweg 45
64291 Darmstadt
Mobil: 0176 233 49 200
Email: frank@reussner.org
www.reussner.org

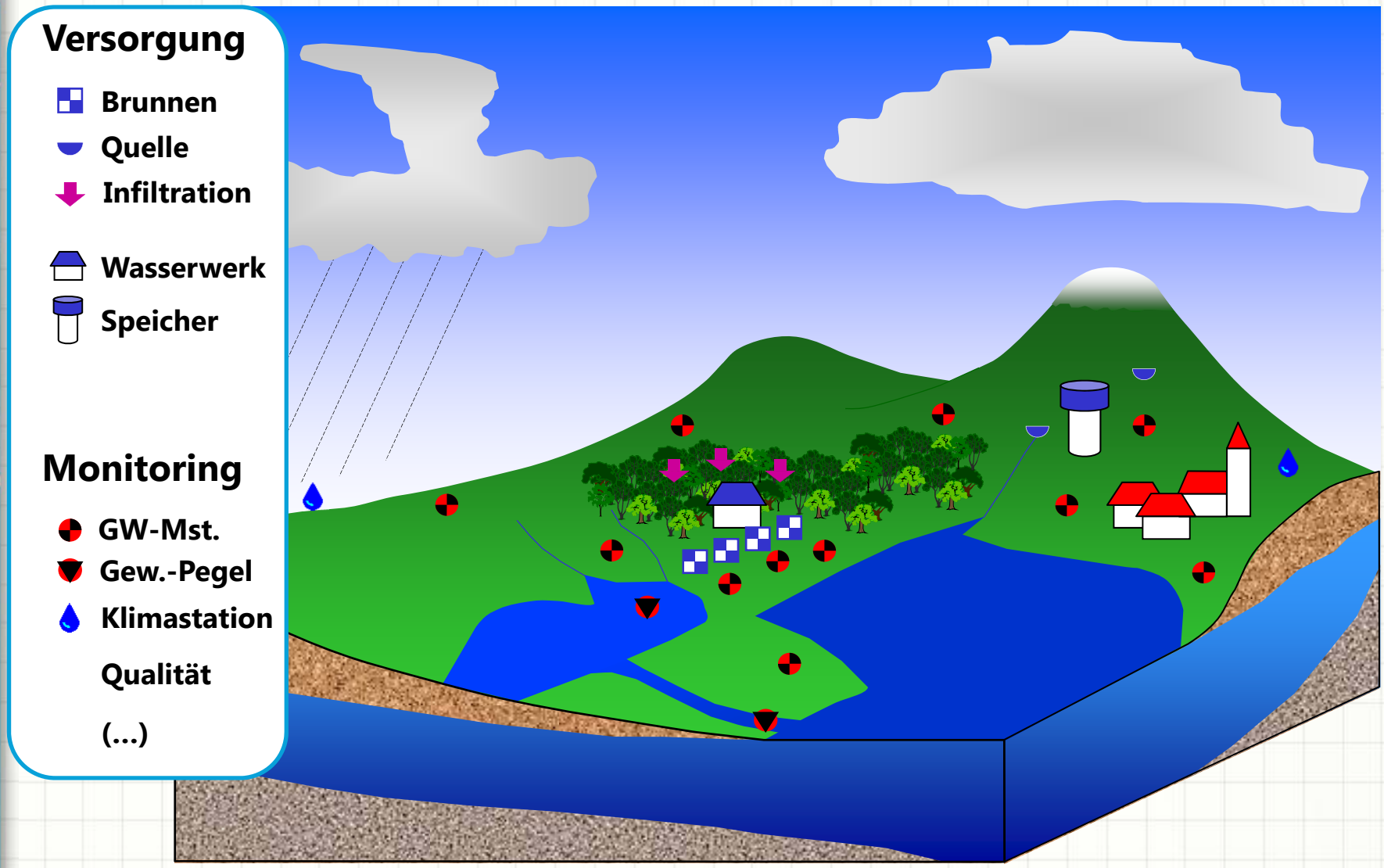
Wasserwirtschaftliche Daten



Wasserwirtschaftliche Daten



Wasserwirtschaftliche Daten



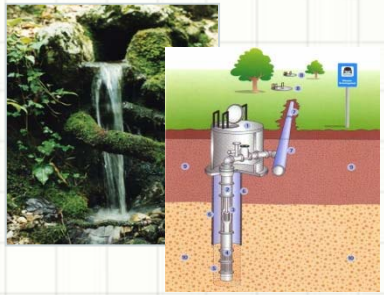
Daten - Bandbreite

Logger/Leitsysteme

Stammdaten

Wasserstände

Altdaten



Fördermengen
Quellschüttungen

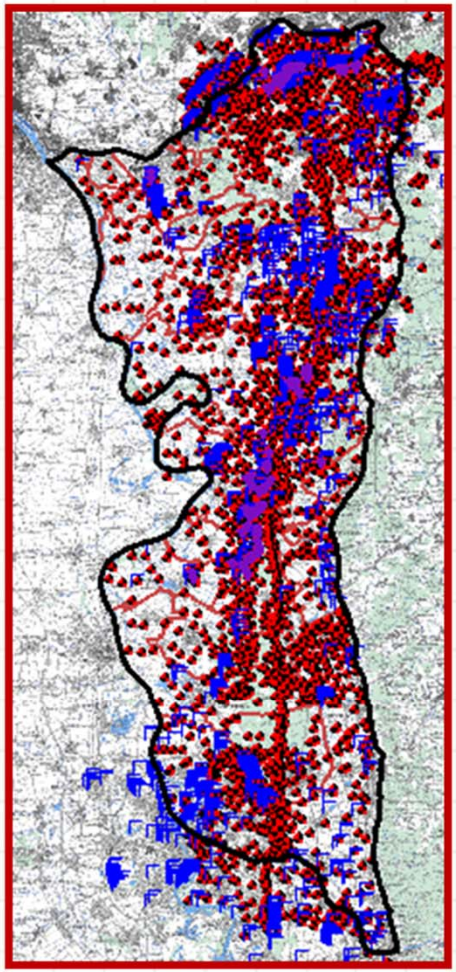
Niederschlag
Versickerung

Analysen

Wasserzähler

Daten - Objektanzahl

Projektgebiet



Externe Betreiber

- EWR
- Fa. Deutsche Bahn
- Fa. Megal
- Gerauer Land
- Hessenwasser
- HLUG
- Riedgruppe Ost
- RP Darmstadt
- Stadt Griesheim
- Stadt Weiterstadt
- Stadtwerke Dreieich
- Stadtwerke Heppenheim
- Stadtwerke Langen
- Stadtwerke Mainz
- Stadtwerke Neu-Isenburg
- Stadtwerke Pfungstadt

⋮

Objekt-Übersicht

GWM	6.006
Brunnen	1.650
Infiltration	153
Klimastation	6
Pegel	31
Quellen	3
QualitätsMst.	140
...	
Σ	7.991

Stand: Nov. 2012

Daten - Messwerte

Datensammler

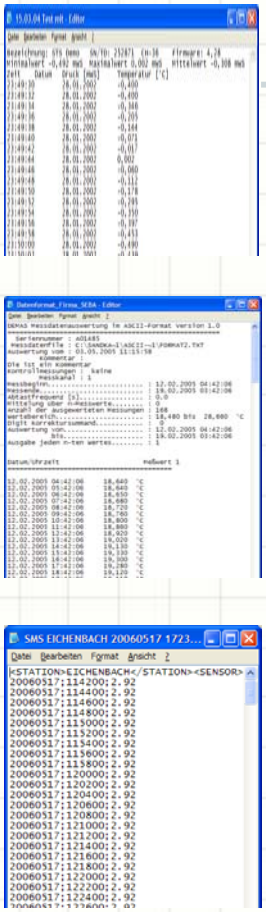
Beispielrechnung:

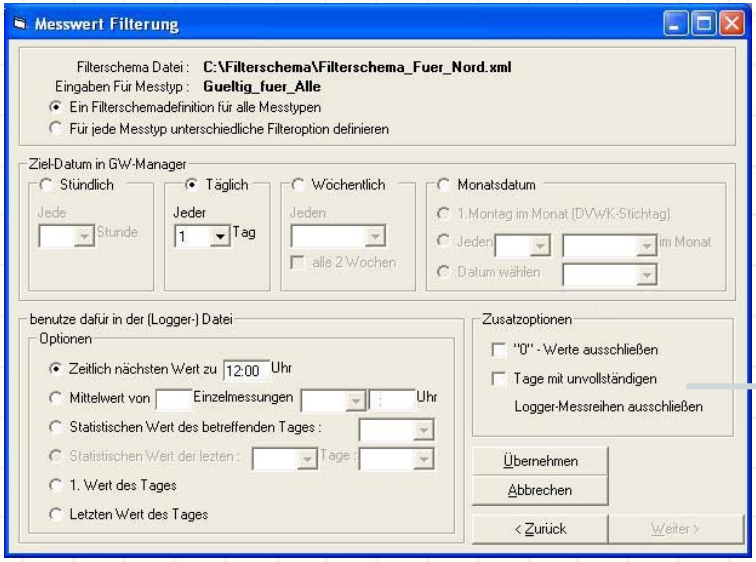
- 8.000 Objekte
- 2 Messtypen pro Objekt
- Messintervall 5 sek.

→ ca. 100 Mio. Messwerte pro Jahr

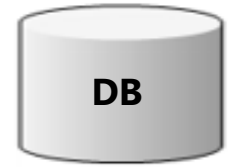
Datenmanagement

Datenübernahme



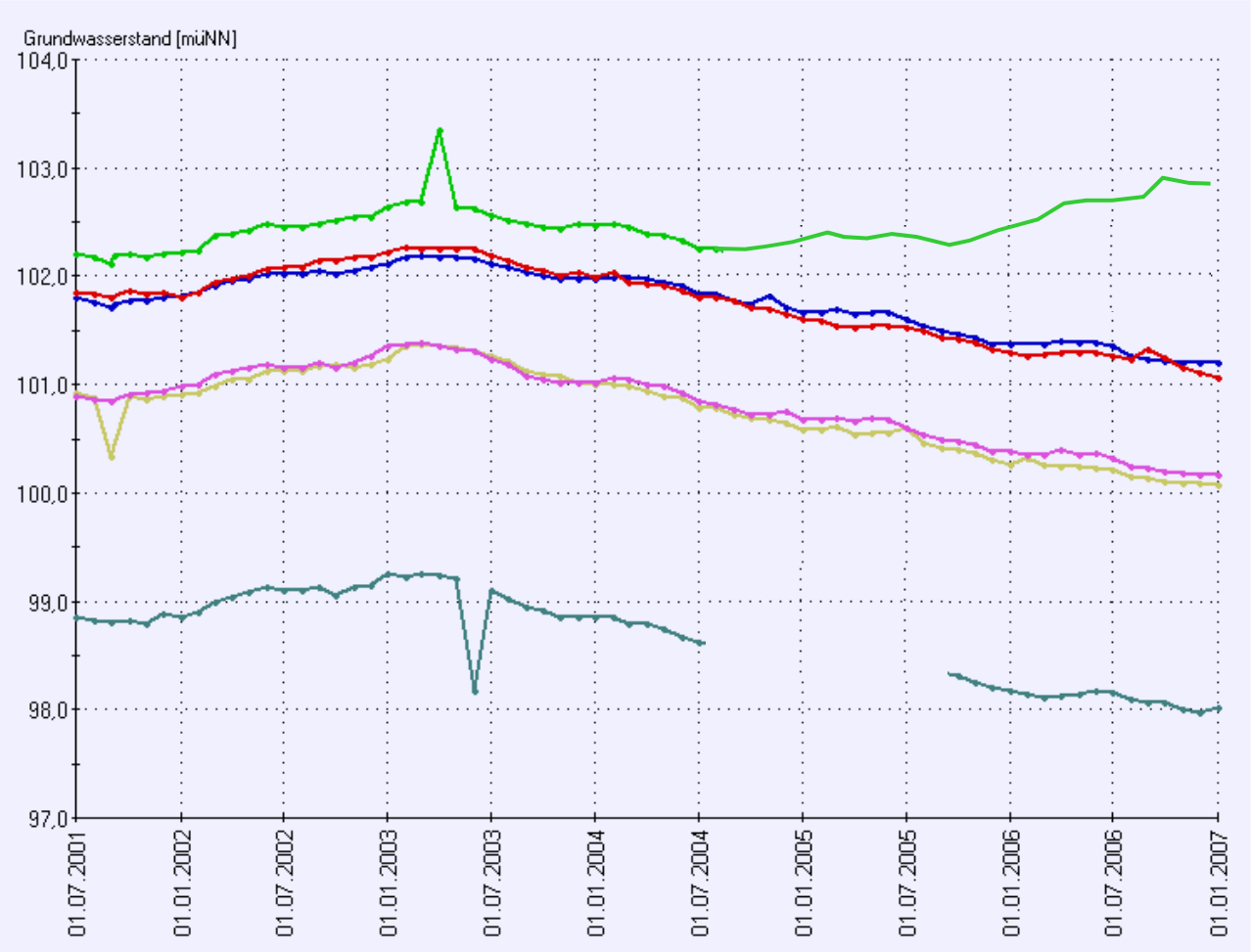


Archivierung

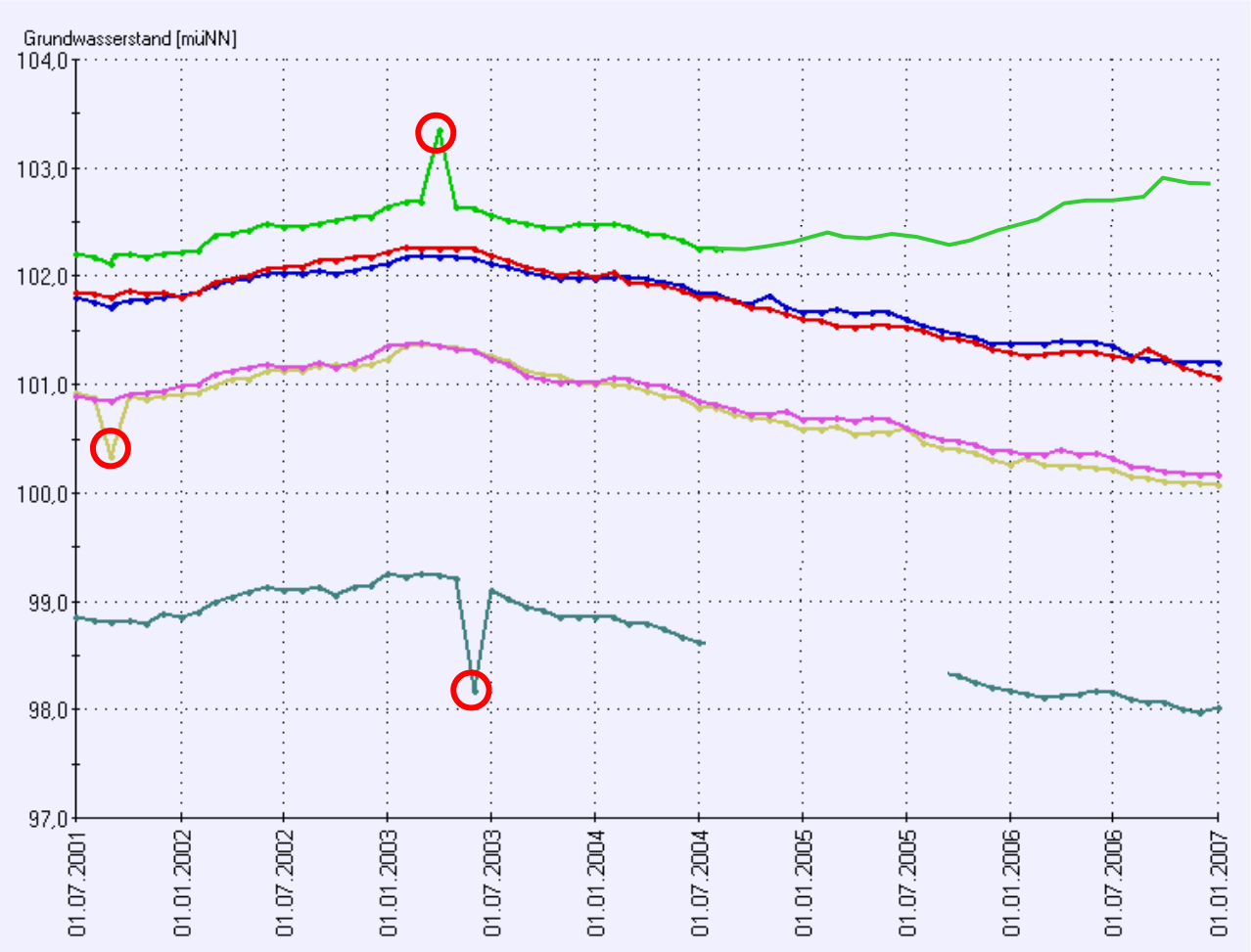


Dateiablage | **Formate** | **Prüfung + Filterung** | **Importierung**

Probleme

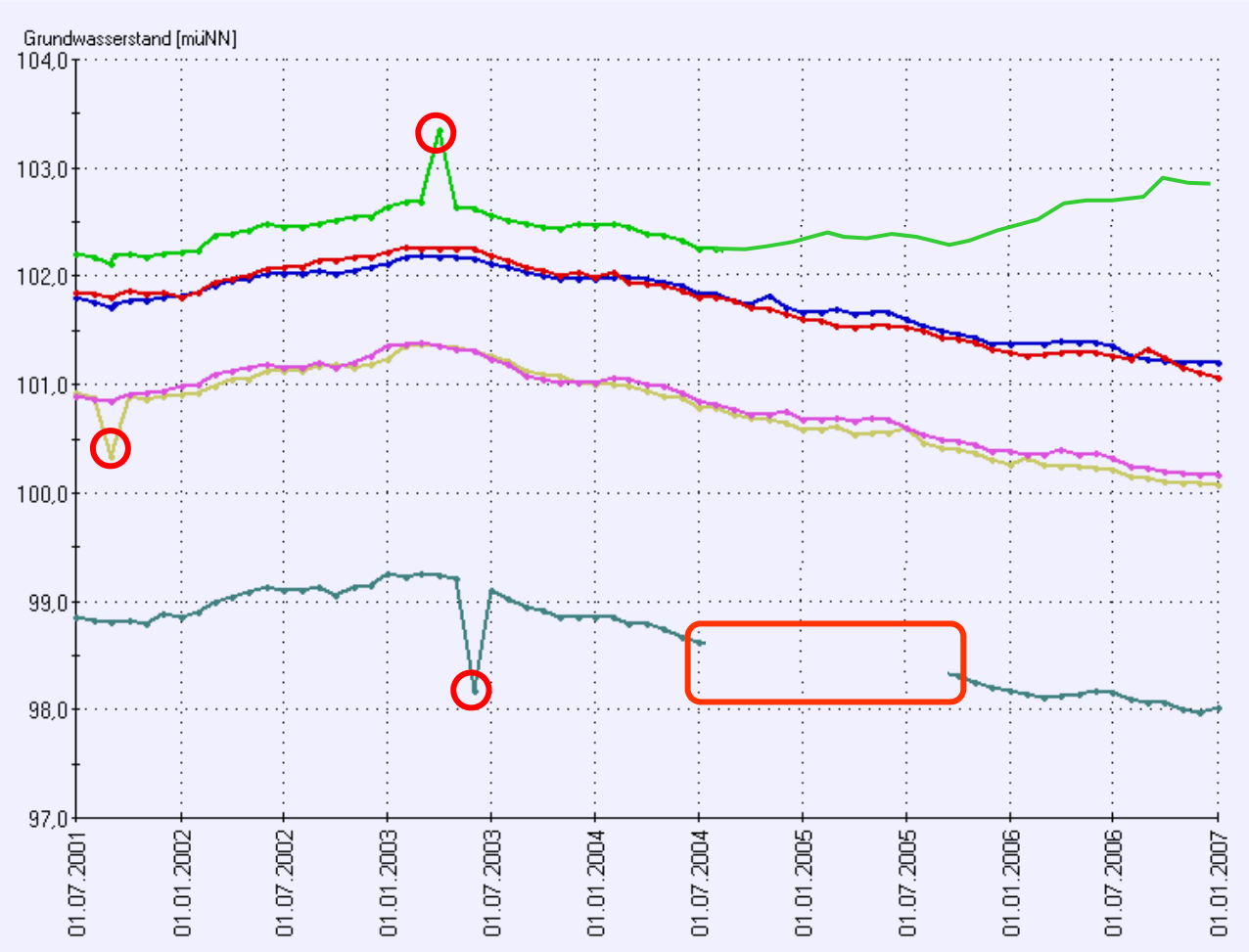

Grundwasserstand [müNN]
104,0
103,0
102,0
101,0
100,0
99,0
98,0
97,0
01.07.2001
01.01.2002
01.07.2002
01.01.2003
01.07.2003
01.01.2004
01.07.2004
01.01.2005
01.07.2005
01.01.2006
01.07.2006
01.01.2007

Probleme



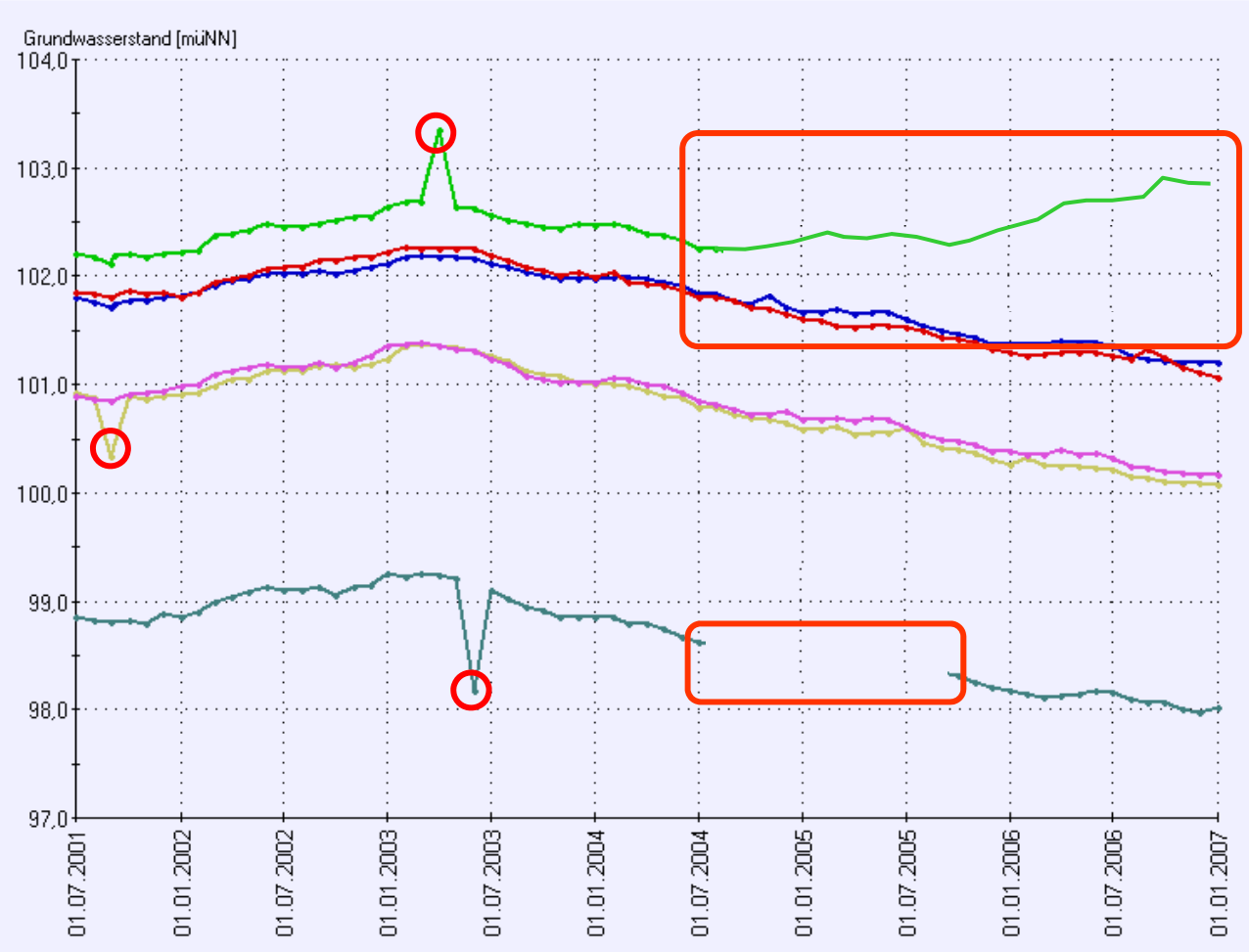
- **Ausreißer**

Probleme



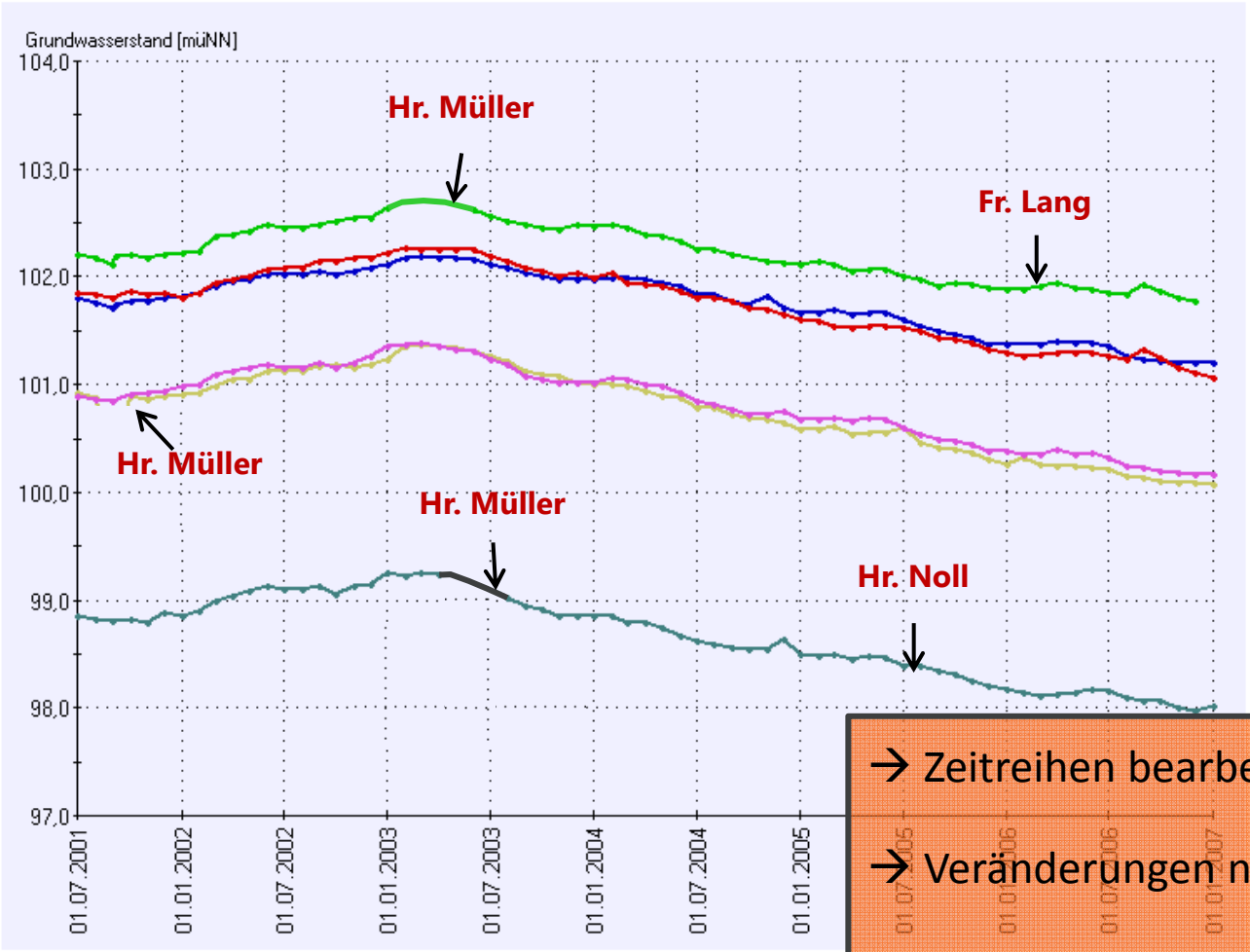
- **Ausreißer**
- **Datenlücken**

Probleme



- **Ausreißer**
- **Datenlücken**
- **Drifts**

Aufgaben



- **Ausreißer**
- **Datenlücken**
- **Drifts**

→ Zeitreihen bearbeiten

→ Veränderungen nachvollziehen

→ Performance !!!

Techniken und Umsetzung

Anforderungen

Zeitreihen

- Messintervall < 10 Minuten
- Zeitraum > 10 Jahre

Datenveredlung

- Visuelle Datenkontrolle und Bearbeitung
- Ausreißer, Lücken, Trend-/Driftbereinigung

Trennung Rohdaten und veredelte Daten

- Versionierung

Analyse des bestehenden Systems

Performance

- Anzeige
- Verarbeitung

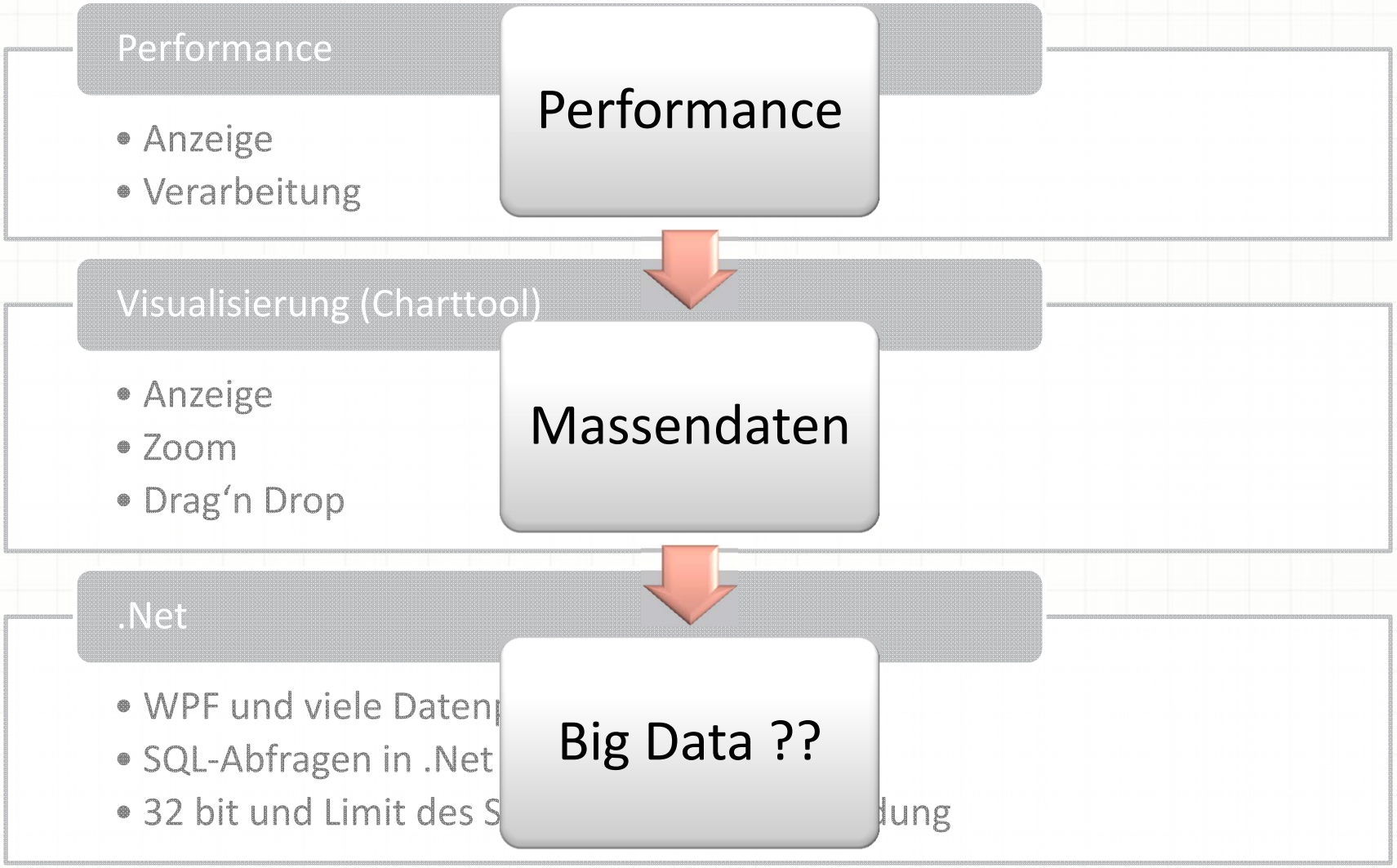
Visualisierung (Charttool)

- Anzeige
- Zoom
- Drag‘n Drop

.Net

- WPF und viele Datenpunkte auf Canvas
- SQL-Abfragen in .Net vs. Stored procedure
- 32 bit und Limit des Speichers einer Anwendung

Analyse des bestehenden Systems



Big Data

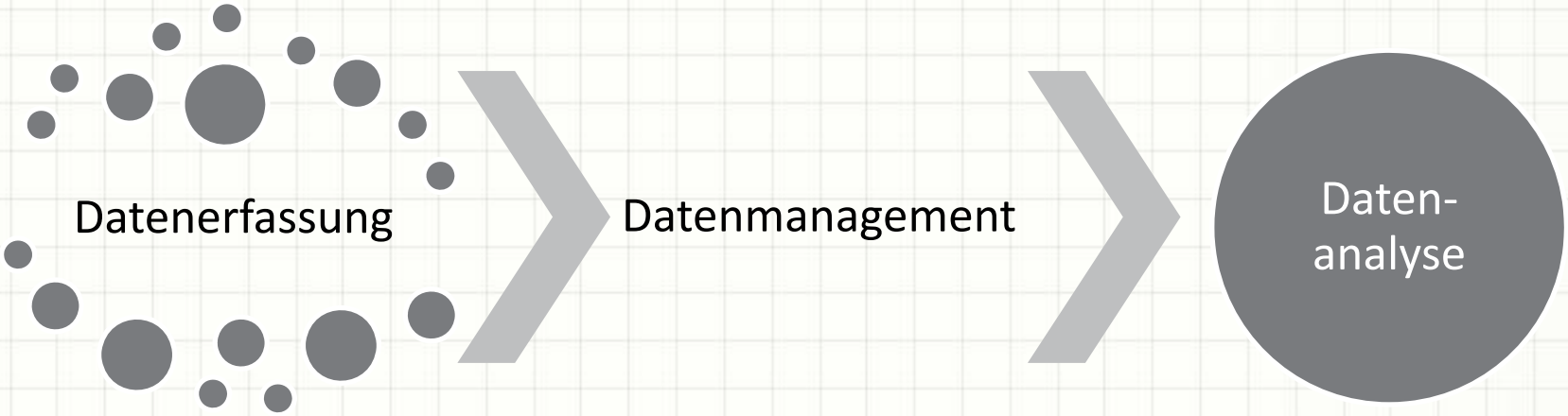
Was ist Big Data?

- Buzzword ?
- Neu ?

Begriffe

- NoSQL
- In-Memory
- Unstrukturierte Daten
- Hadoop

Big Data Übersicht



- Datenbanken
- Sensoren
- Logger
- Dateien

- DBMS
- NoSQL
- In-Memory

- Data Mining
- Text Mining
- Data Discovery

Projekt

Herausforderungen

- Visuelle Interaktion
  - Darstellung Massendaten
    - > 500.000 Datenpunkte in einer Zeitreihe
- Speicherung in Datenbank
  - Varianten (Kopien)

Projektspezifikationen

- Auch 32bit Betriebssysteme unterstützen
- Integration in bestehendes Projekt

Umsetzungen und Arbeiten

Visuelle Interaktion

- Wechsel auf geeignetes Diagramm/Chartwerkzeug

Datenbank

- Indexierung / Indizierung
- Analyse aller SQL-Abfragen
- Umstieg auf stored procedures (auch für Select)

Programmierung

- Neues Modul in .Net 4.0
  - GUI als WPF
  - MVVM
- Auch manuelle Kontrolle der Garbage Collection (beim Löschen von Zeitreihen)

Aktueller Projektstand

Massendaten-erfassung

- Import in Relationale DB
- Anbindung PLS

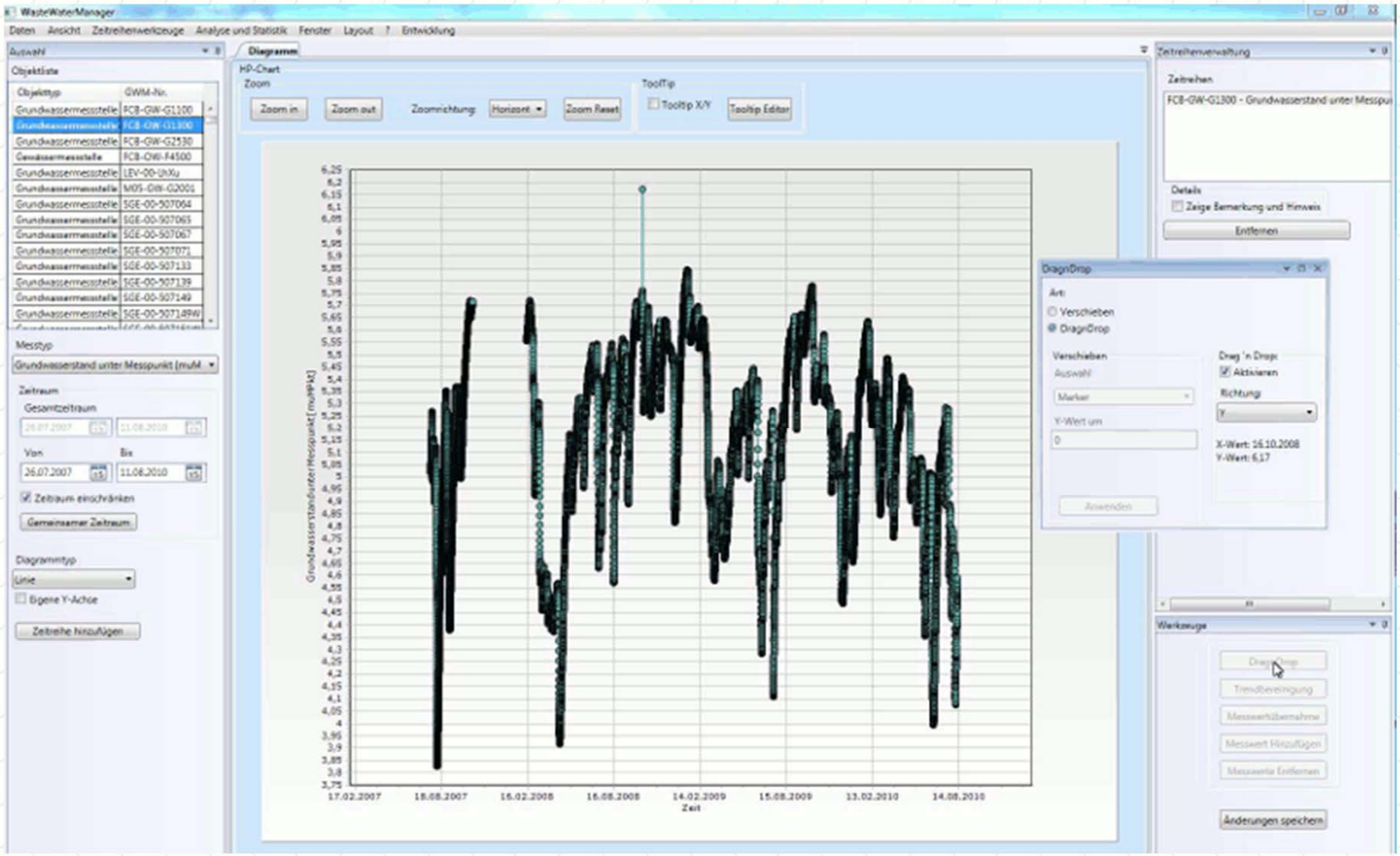
Massendaten-management

- Grafische Datenbearbeitung
- Drift-/Trendbereinigung
- Versionierung von Zeitreihen

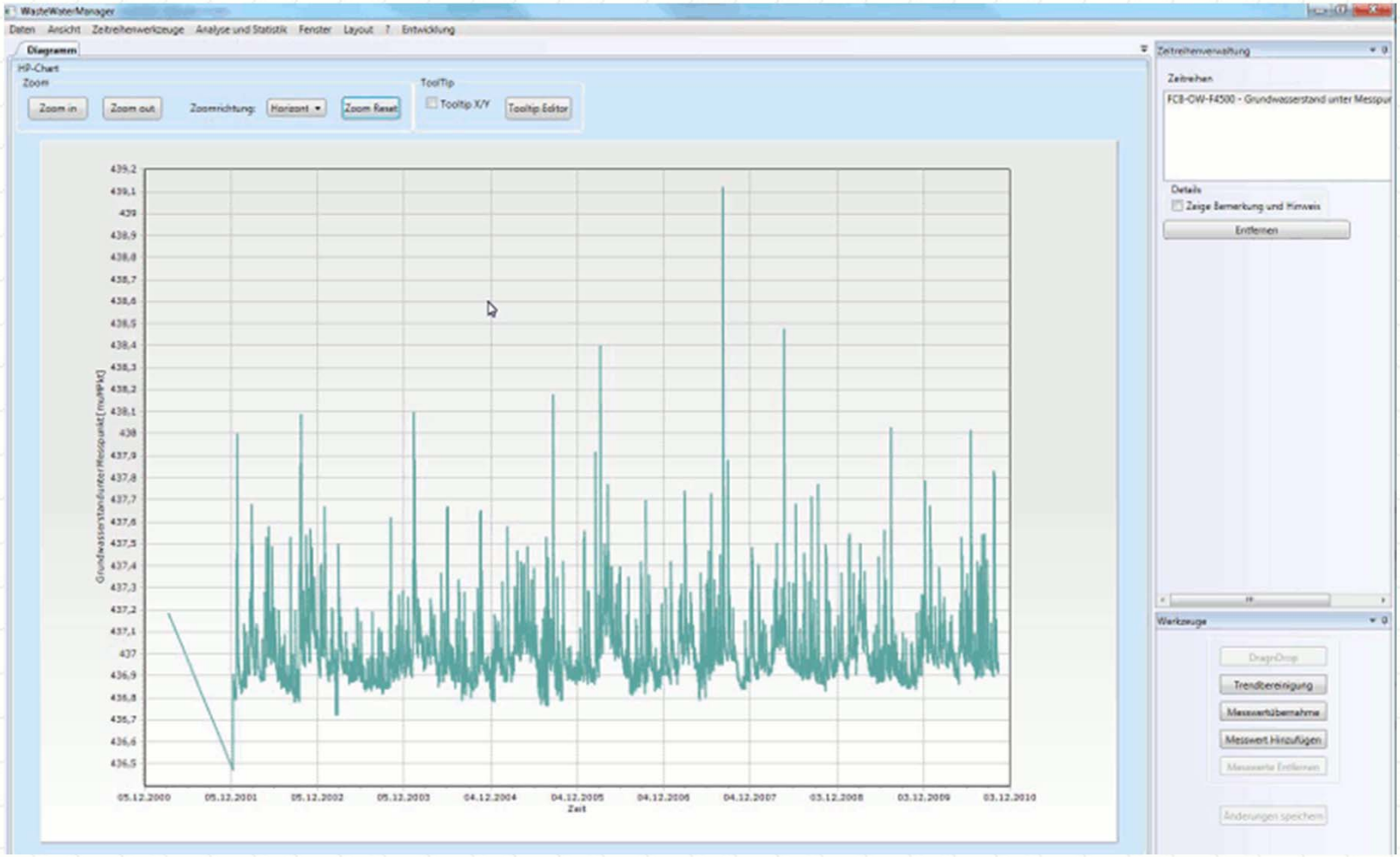
Massendaten-analyse

- Einfache Statistik (Mittelwert, gleitendes Mittel etc.)
- Data Discovery

Screenshot Drag‘n Drop & Zoom



Screenshot Zeitreihe (> 300.000)



Versionierung von Zeitreihen

Versionierung

- Trennung der Rohdaten von verarbeiteten Daten
- Bearbeitungsschritte

Herausforderung

- Performance
- Fortlaufende Zeitreihen (stetig neue Daten)

Konzept zur Versionierung

Einführung eines statischen Systems

- Qualitätsstufen der Daten
  - Wenige Versionen
- Historie der Bearbeitung
  - Wer
  - Was
  - Warum

Zusammenfassung

Massendatenmanagement und -veredlung

- Big Data Techniken meist nicht für Bearbeitung
- Performance im Projekt analysieren und steigern
  - SQL => stored procedures
  - Speicherlöcher (.net und Garbage collector)

Versionierung

- Fortlaufende Zeitreihen machen Versionierung schwierig
- Einführung von Qualitätsstufen

envi-systems

envi-systems GmbH
Ing.-Büro für Umweltinformatik

Dr.-Ing. Thomas Gutzke
Geschäftsführender Gesellschafter

An der Eschollmühle 28
64297 Darmstadt

Tel.: 06151 - 9456-30
Fax.: 06151 - 9456-80
Mail: gutzke@envi-systems.com
www.envi-systems.com

Dr.-Ing.
Frank Reußner

Ingenieurberatung
& Umweltinformatik

Lindenweg 45
64291 Darmstadt
Mobil: 0176 233 49 200
Email: frank@reussner.org
www.reussner.org