Entimo AG

Big and Bigger:

A Big Data Project in in the Pharma Industry, and Emerging Trends Beyond

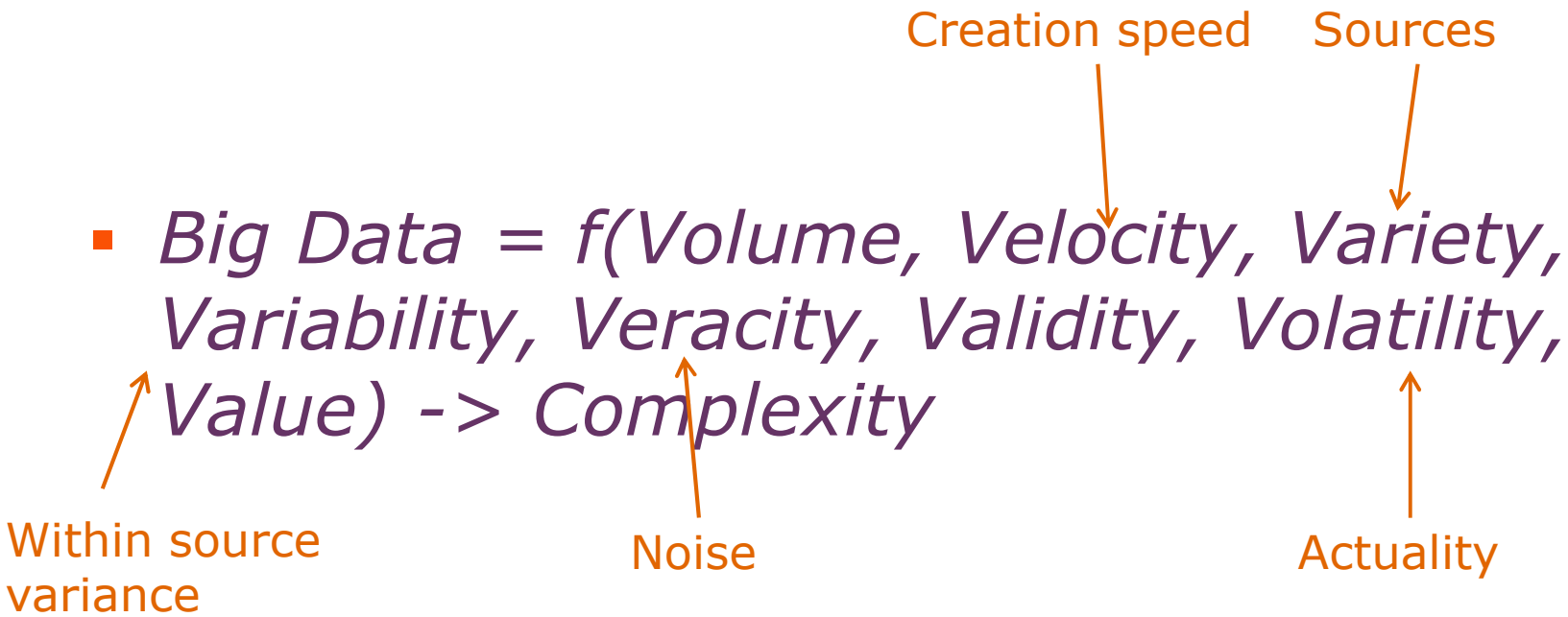
Agenda

- Big Data Characteristics
- Big Data in Pharma
- A Project:
  - Project Specifics
  - Process
  - Analytical/Coding Approaches and Insights
  - Constrains and Limitations
  - Achievements
- New Age
- Conclusions

Definitions and Characteristics

- Wikipedia:
  - "Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time"
- Gartner:
  - "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"

Big Data Dimensions



Big Data is a moving target!

How big is Big Data?

- Other Industries:
  - Lufthansa: 40 years of traffic information available for online queries
  - Facebook: 300 petabytes**
  - Ebay.com: 47.5 PB data warehouse + 40 PB Hadoop cluster*

*Source: http://en.wikipedia.org/wiki/Big_data
** Source: https://www.facebook.com/notes/facebook-engineering/presto-interacting-with-petabytes-of-data-at-facebook/10151786197

Why Big Data Processing in Pharma?

- Driving Factors:
  - Pressure for faster innovation/controlled risks
  - Interconnected world with fast-spreading diseases
  - Outcomes-based reimbursement
  - Protection from drug theft
  - Better monitoring of the social media for brand protection…

Big Data Sources in Pharma

- Investigator and patient portals
- EHR systems
- Insurance and claim data
- Prescription data
- Population health surveys
- Safety data
- PK/PD data
- Lab data
- ePRO and EDC data
- Imaging data
- *Social media?*

Source: Greg Moody „Powering Holistic Data Review and Decision Making Using Visual Analytics "

A Big Data Project (RWE)

- Objectives:
  - Implement a unified system to manage the business workflows including program code and data for research and analytics
  - Enable collaboration, knowledge sharing, and reuse of artifacts
  - Provides annotation capabilities, project metadata management, search, indexing, auditing, security, and versioning
  - Handle TB data volumes with good performance

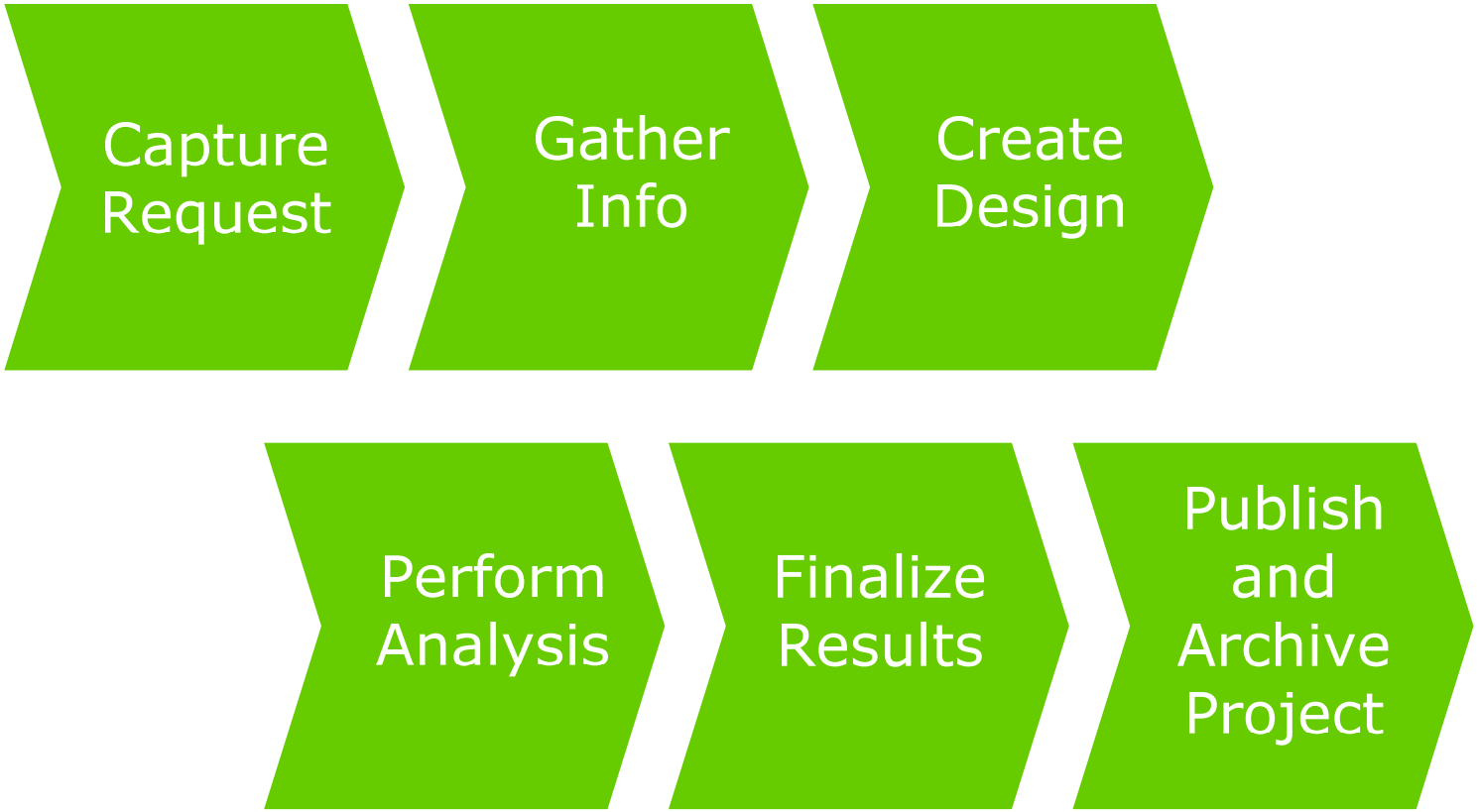
Common Project Traits

- Duration: 1 day to several months
- Stakeholders: Clinical, Medical, Marketing, Manufacturing…

High Project Variation

- Data Volumes:
  - Identify and retrieve study patients - TB
  - Processing of study data - GB
  - Retrieve relevant data from other source - TB
  - Process data to create analytic files - GB
  - Analytic datasets - MB/GB
- Outputs: Data, graphics

Top-Level Scientific Process

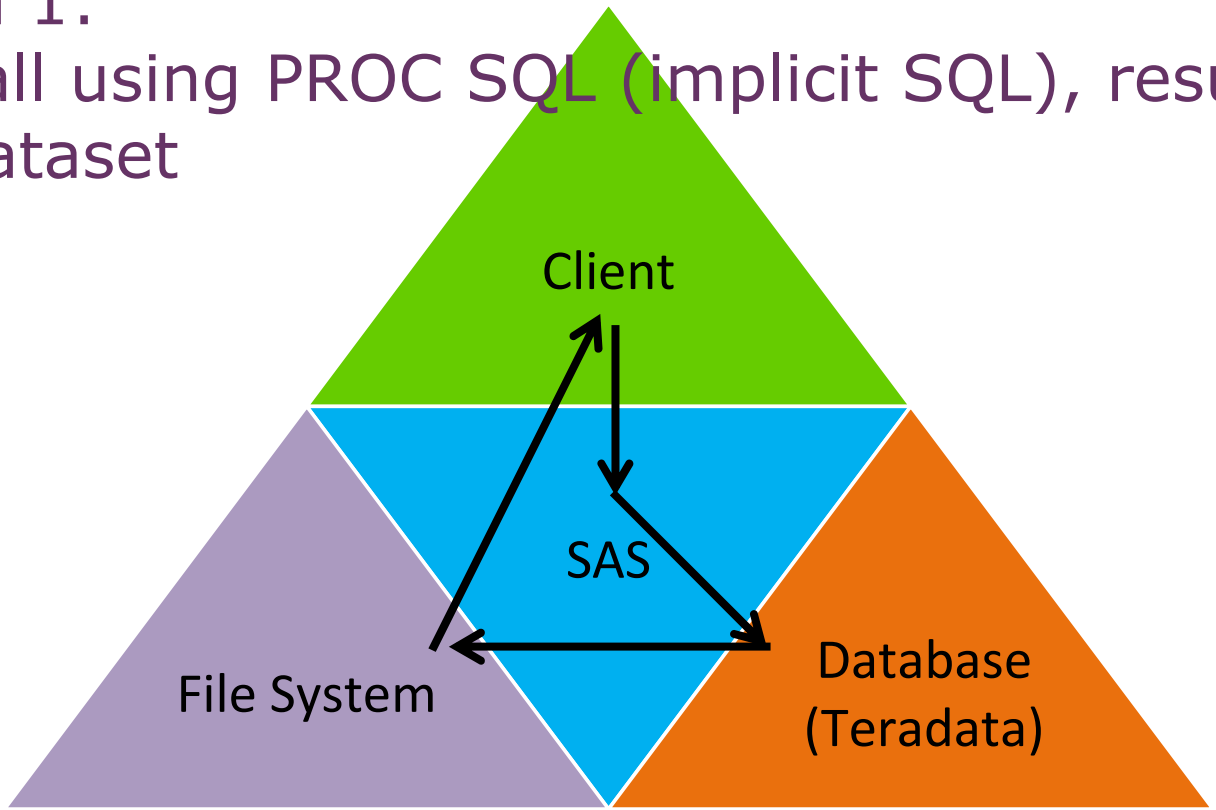


Metadata Layer - Project Attributes

- **Request:**
  - Project Status, Project Description, Project Focus, Project Type, Rationale, Research Question, Business Purpose…
- **Design:**
  - Protocol, Study Objectives, Study Rationale, Study Design, Study Population, Data Collection Methods, Sample Size/Power, Limitations/Strengths…
- **Analysis:**
  - Drug Codes, Diagnoses, Treatment Patterns, Conditions, Trends, Code Comments, Methods Used
- **Outcomes:**
  - Methods, Code…

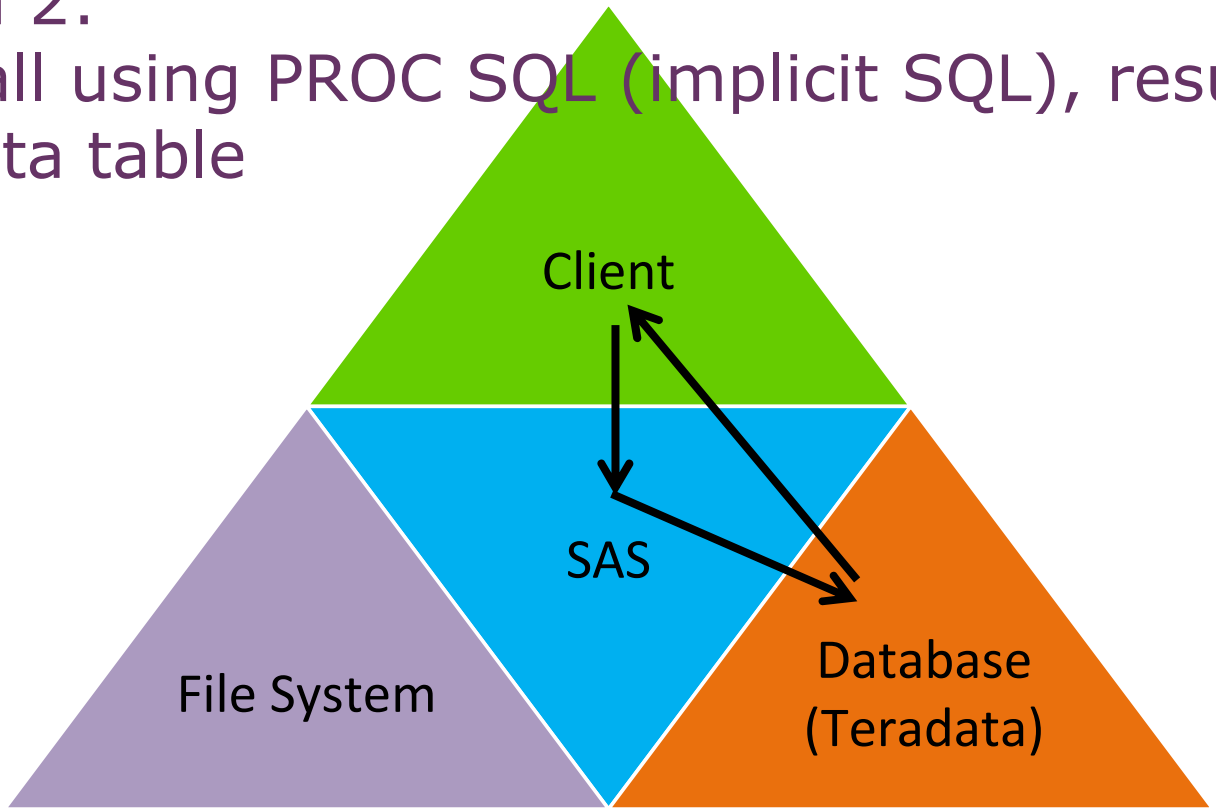
Access Approaches / Coding Practice (1)

Design 1:
SAS call using PROC SQL (implicit SQL), result as SAS dataset



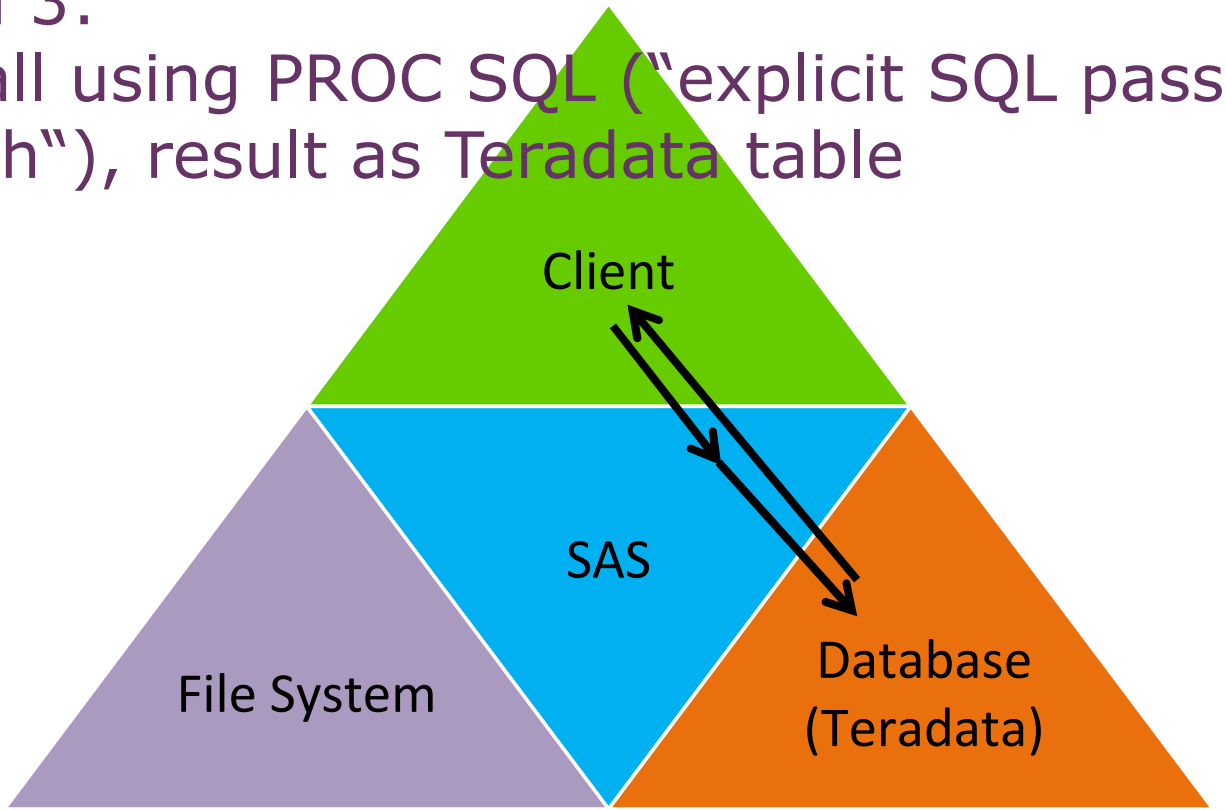
Access Approaches / Coding Practice (2)

Design 2:
SAS call using PROC SQL (implicit SQL), result as Teradata table



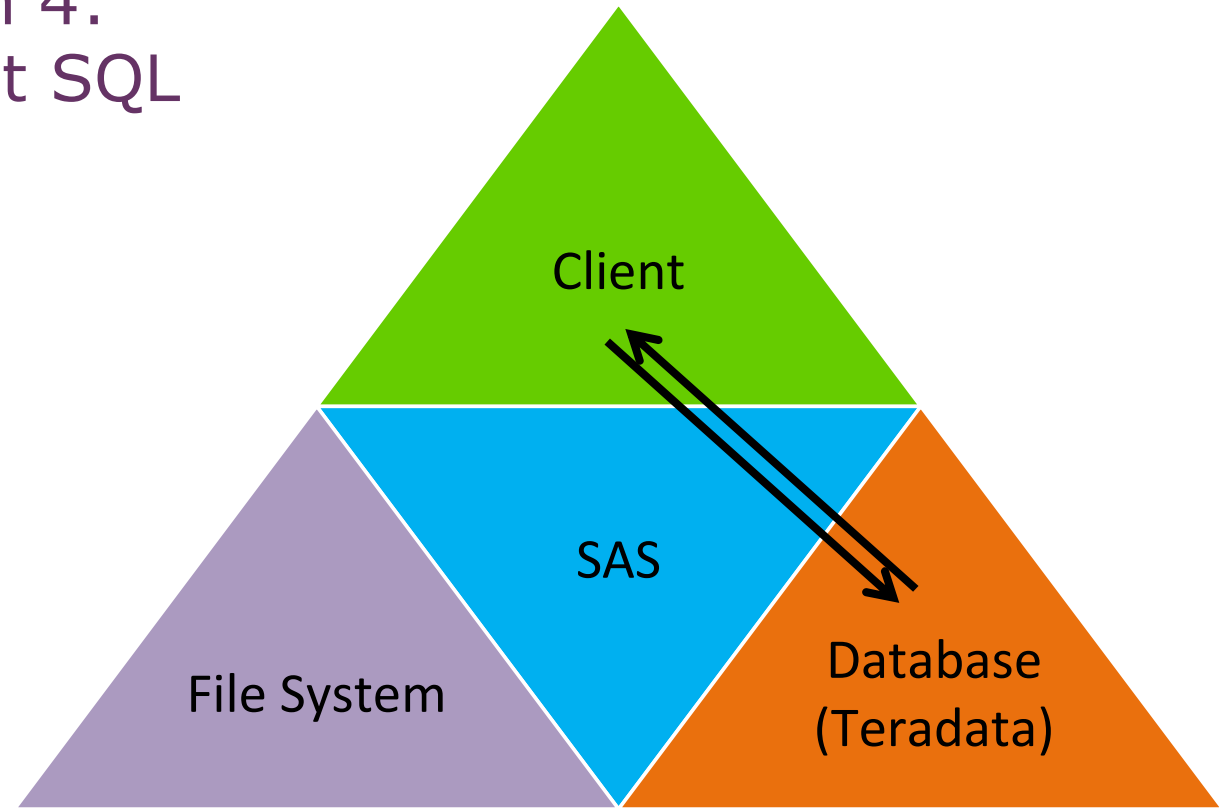
Access Approaches / Coding Practice (3)

Design 3:
SAS call using PROC SQL ("explicit SQL pass-through"), result as Teradata table



Access Approaches / Coding Practice (4)

Design 4: Explicit SQL



SAS vs SQL Insights

- Test Design (extract a large dataset from TD):
  1. Run a SAS call using PROC SQL and store the result in a new SAS dataset
  2. Run a SAS call using PROC SQL and store the result as a new Teradata table
  3. Run a SAS call using PROC SQL with "SQL pass-through"
  4. Run SQL directly via the Teradata interface
- Results:
  - #4 is the fastest
  - #2 and #4 are almost as fast if SAS "knows" specific SQL statements
  - #2 does not support all SQL flavors
  - #3 supports almost all SQL flavors

Constrains and Performance Factors

- Technology:
  - Network Proximity – client cluster, database cluster, SAS, file server)
  - Storage concept - shared storage (SAN, NAS) vs. direct-attached storage (SSD, SATA)
- Social:
  - Existing skill set (SAS vs PL/SQL)

Project Achievements

- **Project Repository** enables RWE analysts to:
  - Search previous studies/projects to leverage method(s)
  - Define project/program ontology to refine search /reuse
  - Store and retrieve project documentation, and publications
  - Archive of protocol, analysis plans, programs, datasets…
  - Achieve traceability and reproducibility
  - Track and document project decisions in real-time
- **Analysis Library** enables RWE analysts to:
  - Create a reusable library of standard programs
  - Maintain the central library of code lists and standard listings, templates of table shells
  - Design templates for easy project creation

New? Technological Age in Pharma

- Trends:
  - Grids
  - Data Streaming
  - MapReduce (e.g. Hadoop)
  - NoSQL Databases (Key-Value,Graph,Document)
  - Transactor based architectures (e.g. Scala/Akka)
  - Presto
- Challenge:
  - Mapping of non-structured to structured data sources

New Initiatives (1)

- Nextgov (FDA)*:
  - Tender "to crawl 20 million biomedical journal abstracts and citations housed on a National Library of Medicine database to uncover drugs that are disproportionately associated with "adverse events.""
  - "… a program to monitor the Web for illegal sales of drugs, medical devices, cosmetics and veterinary products as well as counterfeit food and adulterated or misbranded vaccines"

* source: http://www.fedtechmagazine.com/article/2013/04/fda-launches-three-big-data-initiatives

New Initiatives (2)

- Roche*:
  - "Social analytics"
  - "Data exploration" - the analysis and detection of meaningful structures within unstructured text
- Alliances**:
  - TransCelerate
  - Roche and Astra Zeneca (AZ)
  - GSK and Pfizer
  - AZ and BMS…

Sources:
- * http://www.roche.com/de/media/roche_stories/roche-stories-2014-07-21.htm
- ** http://online.wsj.com/news/articles/SB10001424127887323998604578567682985587790

Conclusions

- Current Pharmaceutical Situation:
  - BIG ambitions
  - “Rigid” technological attitude
- Future Developments:

Point of Singularity

The End… and just Beginning

Big Thanks for Your Attention!