Informatics Challenges for Startups
ACS National Meeting San Diego, 3/13/05

Capturing and aggregating large-scale discovery data in a start-up environment

Susan Baxter, PhD
Chief Operating Officer
National Center for Genome Resources
Santa Fe, New Mexico

NCGR

Focus on biotech companies aimed at drug discovery

- Introduction – Best Practices
- Challenges particular to chemical & biological informatics
- Challenges particular to biotech start-ups
- Summary

NCGR

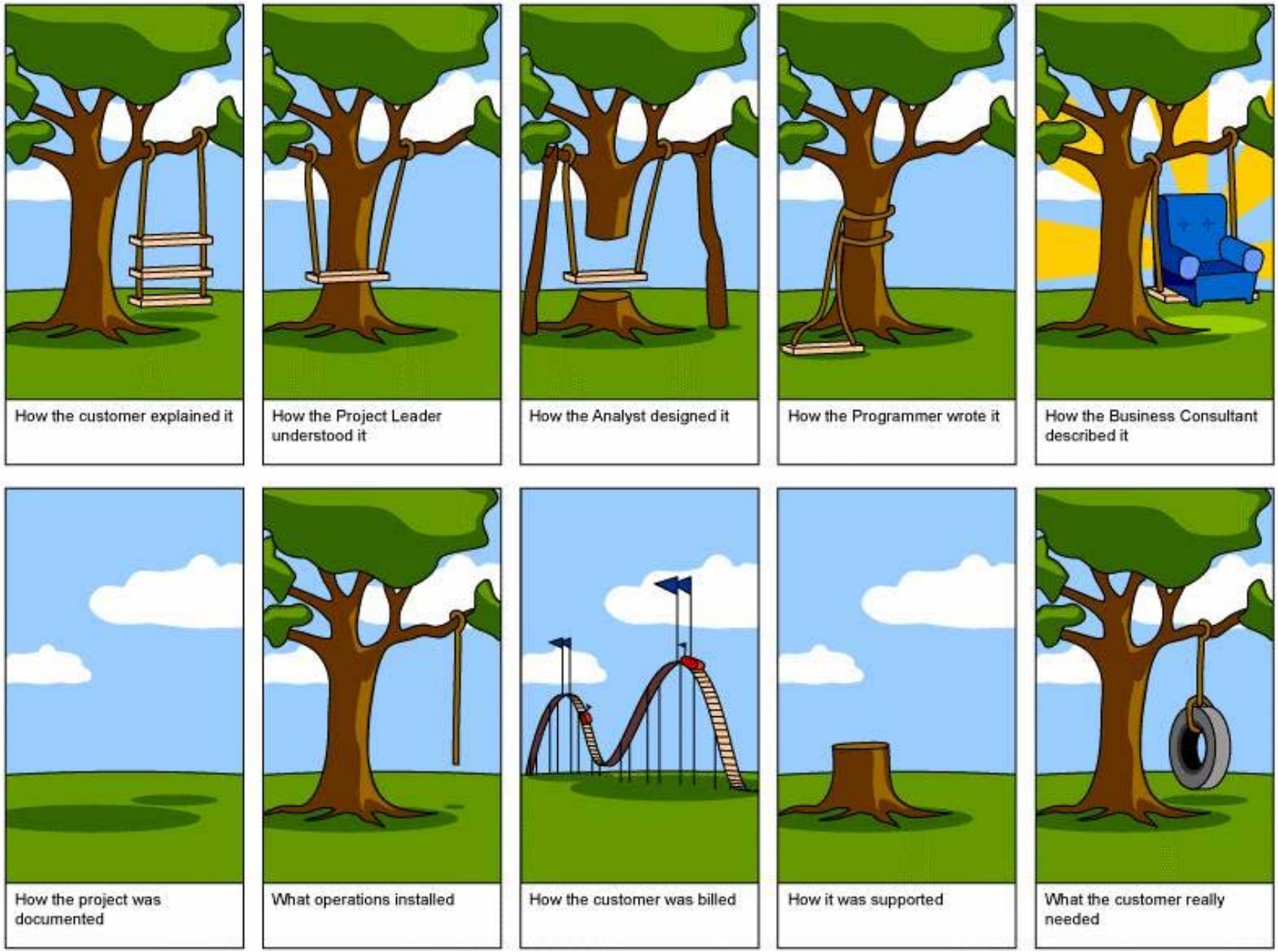
Best Practices: Informatics Infrastructure

- Staff (IT, computational science, engineers, scientists)
- Facilities
- Compute platform (RedHat, Solaris, Windows, Purchase time)
- Solutions provider (Build-it, local IT provider, established Big Dog provider…)
- Storage & recovery solution (SAN, NFS, tape)
- Database solution (DB2, SyBase, Oracle, MySQL)
- Security
- Scalability

The goal is to implement a solution that can be sustained for a period of time, and provides appropriate ROI, and…..

makes a contribution to drug discovery workflow

NCGR


How the customer explained it
How the Project Leader understood it
How the Analyst designed it
How the Programmer wrote it
How the Business Consultant described it
How the project was documented
What operations installed
How the customer was billed
How it was supported
What the customer really needed

From Genome to Function to Understanding

The Grand Challenge ahead is to interpret the **influence of individual genes on the function** of proteins, cells and organisms (systems).

The Operational Challenge ahead is to learn how to **manage and analyze large scale information** and integrate disparate data

"...in some ways we've overgrown our capabilities and need more of an industrial model, both in engineering as well as in data management."*

-- **Edward Rubin,*
Director, Joint Genome Institute
BioInform, Feb. 2, 2004

NCGR

Challenges particular to chemical & biological informatics

- Cross-disciplinary tension
- Current state of data
- Data integration and federation

Natural "tension" between science and engineering / IT

- **Defined standards, processes, procedures vs. free thinking discovery**
  - More than the Mac vs PC controversy
- **Cutting edge may be goal of science, but probably shouldn't be the goal of IT**
  - Productivity depends on infrastructure, arguing for tried and true vs. latest open source solution
- **Large scale science forcing the issue**
  - science + computer science + software engineering + IT = today's highly effective informatics teams

Information at the chemistry-biology interface

- **Incomplete Knowledge** - Fragmented information, always changing
- **Cutting Edge Biotechnologies Evolving** - Requiring re-analysis, reassessment, and versioning of data and resources
- **Experimental Confirmation Difficult** - Dependent on well-behaved factors or precisely defined assays and clinical phenotypes
- **Lack of Accurate Predictive or Computational Methods -** Limited training sets, statistical power, force fields and models
- **Rampant False Positives** - Low signal-to-noise, incomplete, or sparse datasets skew conclusions
- **Volume of Data Growing Exponentially** – Unprecedented in this domain; database standards and interoperability lacking
- **Cross-disciplinary Teams needed** – No one can know it all – “Chemical Biology – what is it anyway?”

NCGR

Data Available

- **Multiple data types, data formats**
- **Raw, Metadata, Results**
- **Some data is publicly available:**
  - **Commercially available small molecules (ZINC, USCF)**
  - **Human protein sequences (RefSeq, NCBI)**
  - **Protein structures (PDB, www.rcsb.org)**
- **In-house data from company processes**
- **Data from vendors & partners**

Data Access and Data Comparison are the two main functionalities desired by scientists working in high-throughput environment

- **Multi-disciplinary teams, geographically dispersed, require data access (first); sharing (second)**
  - **Tools for *integrative*, comparative analysis are required : each dataset provides an angle on the biology; each team interprets it with their own bias (chemistry, pharmacology, biology…)**
  - **Pattern detection tools compress data and highlight important features to aid expert interpretation**

Integration: Understanding Syntax and Semantics

- **Syntax: format and structure of data**
  - Underlying storage architecture: Relational, XML, Text/ASCII, SAS, proprietary format, etc
  - "Shape" of field (numeric, text, decimal, date, units)
- **Semantics: meaning, interpretation, intended use**
  - Homonyms / Synonyms
    - Headache: Migraine, Vascular Headache, Tension Headache, Raeder Syndrome, Bing-Horton neuralgia, etc….
  - Different underlying objectives of dataset:
    - Clinical trials: narrow protocol-based focus
    - Epidemiological studies: "Real World" practice
  - Data collected in different eras (era effect)
    - New standards of care
    - New assays, methods of testing
  - Different geographies / demographics
    - Lifestyle and diet

Integration Strategies

- **Multidisciplinary approach**
  - Scientific expertise (semantics)
  - Software and database engineering (syntax)
  - Advanced analytic capabilities (no false signals or missed signals)
- **Use of Standards**
  - Common understanding of underlying semantics (many industry standards emerging)
  - XML can be used as a container for both the data and the underlying assumptions and definitions
- **Vendor solutions / enterprise systems**

Need to harness the energy from science-engineering tension for integration projects that successfully contribute to drug discovery

Advanced analytics to validate integrated datasets

- **Integration achieved by centralized RDBMS, or web services or semantic web technologies**

- **Advanced analytics to provide understanding of inherent differences (semantics) between source datasets**
  - **Statistical models**
  - **Comparative analysis**
  - **Integrity checking**

Analysis: Informatics Challenges

- **What computational processes will be run? (virtual screening, protein structure prediction, BLAST, phylogenetic clustering)**
- **How many users? Where are the users?**
- **How much data will be captured in the corporate database & why?**
- **Will the database need to support analytics capability?**
- **Will clinical data or human population data be collected?**
- **What is the big picture? Can the process be plugged into the drug discovery workflow?**

Challenges particular to start-up companies

- Limited resources
- When to grow/how to grow

– $$$$$

IT/Process ROI in a discovery environment

- **Can't always be tied to new "widgets"**
- **Data Integration and Workflow automation can**:
  - Provide greater efficiencies (islands of automation, processing on genomic scale)
  - Increase accuracy (avoid costly mistakes)
  - Enable collaboration (discoveries that would otherwise not happen)

NCGR

What are the motivations behind data management?

- Accuracy and reproducibility
- Intellectual Property
- Regulatory compliance
- Customer/client contracts
- Company culture (collaborative, sharing, iterative)
- Science (large scale biology)
- Efficiency

Intellectual Property

- Many times, data is “value” of company
- Buy-out; merger activity – informatics can facilitate due diligence & affect value
- Program partnering – especially preclinical and early phase development
- E-Notebooks

Industry regulations and guidance documents (IT/Software specific) need to be taken seriously - early

- **21 CFR Part 11** – Electronic Records, Electronic Signatures
  - Data Security, Integrity, Traceability
  - Software Validation
  - Electronic signatures
- **21 CFR 312** – Investigational New Drug Application
- **21 CFR Part 820** - Quality System Regulation
  - System Quality
- **FDA Guidance Documents**
  - Computerized systems used in clinical trials
  - General principals for software validation
- **HIPAA** – Health Insurance Portability and Accountability Act of 1996
  - Patient Confidentiality

NCGR

Data management processes don't always save money…at first

- Well-known cost associated with implementing process or building infrastructure before gains in efficiency or cost savings seen
- Cost barrier to implementing CMM or process automation in start-up environments
- Customer driven
- Putting it off until later dramatically increases cost and difficulty

NCGR

Particular Challenges

- **It's not Academia; it's not Wall Street**
  - At an early stage, convincing scientific founders, drug discovery management, & VCs that infrastructure is important can be challenging
- **Once you get a green light, you can't count on starting over:**
  - "Under the hood" iteration needed; unlikely that you'll get a cap ex budget (for a while)
- **It takes talent to herd cats and talk babel:**
  - Strong, open project management
  - Iterative development – deliver early, exceed expectations
- **Reducing entropy requires energy ($$):**
  - Lab notebooks or a corporate database?
  - LIMS build vs. buy?
  - Free, open source or enterprise class?
- **Rigor required:**
  - Corporate infrastructure isn't an open source development project – hardware & OS standardization, password protection, documentation, QC needed to avoid brittleness, and remain robust

Lessons Learned

- **Don't reinvent the wheel**
  - Build vs. Buy
  - Modular development vs. Enterprise solution
- **Federation / Integration allows new functionality**
  - Collaboration
  - Iteration
- **Scientists want everything in one place**
  - Desktop solutions
  - Who gets what information?
- **Chem / Bio-informatics still challenging**
  - Evolving standards
  - Database legacy issues
  - No one out of the box solution – but startups can adopt and build
- **Communicate, communicate, communicate**
  - Tactical plans
  - Requirements gathering
  - Stakeholders identified early

NCGR

Acknowledgements

- **Stephanie Reisinger**
  Vice President, Development
  ProSanos Corporation
  Harrisburg, PA

- **Jacquelyn Fetrow, PhD**
  Reynolds Professor of Computational Biophysics
  Departments of Physics and Computer Science
  Wake Forest University
  Winston-Salem, NC

- Vickie Hall (NCGR)
- John Utsey (NCGR)