lifescale analytics

analytics + insights for life science

10 Trends

Driving a Changing Data Landscape in Life Sciences

a lifescale leadership brief


The Life Sciences data landscape is growing, accelerating, diversifying on some fronts and converging on others. How can those in Life Science leadership...
identify the challenges
+
recognize the business opportunities
in this changing data landscape?

lifescale analytics analytics + insights for life science

10 Trends Driving a Changing Data Landscape in Life Sciences

Decreasing Cost-Per-Observation

Over the last couple of decades CPO (cost-per-observation) has inexorably declined across a very wide range of biological measures. In 2001, J. Craig Venter, iconic figure in the race between Celera Genomics and the Human Genome Project, predicted that the cost of sequencing a human genome would drop from $2.7 billion for the recently concluded dollar Human Genome Project, to $1000. We stand on the verge of realizing that $1000 milestone in early 2013. High throughput and high content screening have been on similar trajectories. This decrease in cost-per-observation, driven by automation, parallelism and miniaturization, has been used by researchers to expand the scope of experimental coverage driving up data volume and velocity within the lab. Acquiring life science data is cheaper and faster than ever before and the trends suggest that costs will drop further before stabilizing.

opportunity

Increase in volume and velocity of data can accelerate time and reduce cost in getting to market, as well as increase confidence in analytic results and ability to address new and different questions.

challenge

In order to retain higher volumes of data, organizations must adopt new collection and processing techniques.

In Situ Measurement on the Rise

Wearable and implanted physiological sensors, wirelessly connected to smart phones provide spatial and time context to ambulatory real-life measurements. This combination will revolutionize medical research. Consumers aren't waiting. A small vanguard is starting to live the "Quantified Self" life. These innovators are tracking food choices, calories and timing of meals. They are recording volume, type and intensity of physical activity ranging from formal workouts to how many steps they accumulate through the day. They may track progress against fitness goals and benchmarks along with daily fluctuations in weight and body-fat. Some are tracking blood glucose levels even though they are not diabetic. Others go so far as to order, pay for and track their own lab work on blood, urine and stool samples. Eventually these smart sensor technologies will transform health care delivery for obesity and chronic diseases such as diabetes and asthma. In situ measurement technology will impact clinical trial data collection moving it into the field in ways not previously practical. The data challenges from this trend include sampling or aggregating from very granular time-series and coping with data that is simultaneously polluted and enriched by real world messiness.

opportunity

Collecting data "in context" enables deeper analytic insights. Typically results in higher numbers of observations resulting in the ability to investigate fine-grained temporal and locational relationships.

challenge

Requires a specialized data collection infrastructure and adoption of collection and processing techniques as well as new technologies to process and retain higher volumes of data.

6200 Shingle Creek Parkway, Suite 580 Minneapolis, Minnesota 55430 | 763.585.5870 | info@lifescaleanalytics.com | lifescaleanalytics.com

lifescale analytics analytics + insights for life science

Increasing Integration Across Domains

Evidence-based medicine (EBM) and its cousin Translational medicine are busy merging data and analytical techniques from bioinformatics and medical informatics. Biobanks bring together patient descriptions like age, gender, and ethnicity with medical records of symptoms, diagnosis, treatment and outcomes. And, when research is conducted on preserved "banked" samples from those patients, molecular analysis is merged with clinical outcomes all without having to recruit participants. In agricultural life sciences we see crop research translated into the practice of "precision agriculture." Will precision agriculture growers and seed retailers close the loop by sharing granular data back with seed companies? This is just a sampling of places in the life sciences where boundaries are blurring and data is merging. Although data volume relative to a specific research focus does increase the real issue is a dramatic increase in data diversity. Distinctive semantics and conceptual frameworks create barriers to integration along with complex human factors in bringing together professional and academic practices with unique heritages, techniques and goals.

opportunity

Gain greater insight by analyzing a richer set of relationships thus reducing reliance on inference. A common view of information across the enterprise enables consistent interpretation and decision-making.

challenge

Requires enhanced information processing techniques for correlating disparate data.

Accelerating Automated Analysis and Decision Loops

Large high throughput screening facilities are now using Design-of-Experiment (DOE) software to automate the development and optimization of assays. Some labs are moving to Adaptive DOE, an iterative approach that starts with an initial experimental design, executes the initial experimental run, analyzes the interim outcomes and uses that interim result data to design the next iteration of experiments. Statistically-driven, iterative, automated experiments are difficult to comprehend or follow without specialized training and analytical tools. Many labs will use just the outcomes and end up treating the intervening analytical/experimental iterations like a "black box." This change in the data landscape impacts, data velocity, volume and ultimately analytical complexity.

opportunity

Results in richer experimental data sets, enabling deeper insights.

challenge

Assumes knowledge of complex experimental designs requiring more sophisticated analytical techniques.

Social Graph and Networks Generate new Classes of Data

Online social networks provide patients with a disease-specific support and create new classes of data for the life sciences. Text data shared in online forums among patients will be analyzed and used to segment patients into smaller clusters based on similarity of symptoms and clinical outcomes. Sharing and discussion among patients has already generated new research hypotheses and documented patient (market) demand. Self-organization among patients has in some cases greatly accelerated finding and enrolling patients in clinical trials. To the degree that this trend combines with the "Quantified Self" movement cited above, we will find that

opportunity

Potential to spot trends and associations not possible in traditional experimental data.

challenge

Requires non-traditional infrastructure for efficient storage and access to information as well as new analytical techniques to derive meaning from non-controlled data.

6200 Shingle Creek Parkway, Suite 580 Minneapolis, Minnesota 55430 | 763.585.5870 | info@lifescaleanalytics.com | lifescaleanalytics.com

the data shared online in disease forums will include lab reports, physiological measurements, and images not just text descriptions of symptoms and outcomes. How will biotech and medical researchers use self-measured, self-contributed medical data? How do we assess reliability and validity in an environment rich in granular observations but lacking in formal experimental design and controls?

Computable Text Becoming Mainstream

Text continues to become more machine accessible, semantically retrievable and computable. Ontologies and taxonomies contribute not only structure but relationships between entities. Statistical text analytics reveal latent semantic structures. Statistical and rules based natural language processing (NLP) techniques used in an integrated system can lead to accurate and sophisticated tagging of text. This tagging/indexing enables the creation of query-friendly, meta-data for text "objects" with little or no human intervention at indexing time. Textual analysis applied to large collections of abstracts and full text of research papers (i.e. corpus analytics) enrich and extend the traditional research literature review. In medicine, real-time monitoring of text on Twitter, Facebook, Google and other social and search sites have been demonstrated to provide early detection of contagious outbreaks. IBM's Watson, the Jeopardy winning, natural language question answering system is now being "trained" to be a medical diagnostics savant. These changes convert vast archives of textual information into more structured data and raise new challenges from mastering unfamiliar tools to data quality. Integration or federation between text sources and databases also pose formidable tasks to information architects.

opportunity

Ability to derive meaning from unformatted textual information. Increased potential for capturing richer observations and information sets. Increased likelihood that observations are documented.

challenge

Requires new techniques and technologies for processing and deriving meaning from information.

Digitized Images Dominating

Light, magnetic, tomography, x-ray and sound; whatever the waveform, the capture, storage, and management of images are all going digital. This transition to digital transforms the ways that organizations manage, secure, store, share and analyze biological and medical images. Algorithms are coming to the aid of human eye/brain pattern recognition. Edge detection helps isolate features and algorithmic pattern and anomaly detection aid radiologists in identifying key regions of an image for closer examination. In biotech, we see the rise of HCS (High Content Screening). HCS automates fluorescence and visible light microscopy, using robots to handle samples and slides, and automates digital capture and analysis of images to speed discovery research. GPUs (Graphical Processing Units), whose rapid advance is driven by the PC gaming industry, have significant capabilities for image analysis. GPU

opportunity

Ability to derive meaningful data from images, increasing the potential for capturing richer observations and information sets.

challenge

Requires sophisticated techniques and technologies for processing and deriving data from images.

clusters are being used to accelerate and enhance digital MRI and CT imaging. Digital imaging of all types creates terabyte and petabyte scale data volume challenges and places high demand on infrastructure ranging from more robust data networks to more capable workstations and higher resolution displays.

Open Science & Public Data Growing Rapidly

The Wikipedia page for "biological databases" lists over 130, ranging from nucleotide sequences, to protein structure, metabolic pathways and more. Most of those are maintained by academic institutions or government agencies and are open to the "public" with minimal or no licensing fees. The data is freely available for querying, merging with, and enriching private internal data. Some segments of the scientific community are pressuring governments to mandate that all research projects receiving government funding be required to release data sets along with their papers and reports. This move toward more extensive and diverse, publicly available data has numerous opportunities for skillful merging of datasets to find new insights. How much could be learned simply through more skillful integration and analysis of publicly available data? Will this be the key contribution of "data scientists" in life sciences?

opportunity

Reduce cost of data collection while increasing access to proven, high-quality information (compared to generating data internally).

challenge

Techniques will be required to correlate public data with internal information and licensing restrictions may limit use of open data.

Collaborative Research Growing

The lone scientist in his lab, working for years before a big breakthrough has given way to multi-disciplinary teams, big science, robotic labs and more. Many organizations increasingly seek to collaborate outside their walls. This externalization may range from outsourcing a high-throughput screening run to full execution of a clinical trial. In other cases collaboration is between a biotech and academic labs. Whatever the particular form of the externalization, the methods, controls and technology for sharing data for that collaborative work brings challenges for data transport, access control and regulatory audits.

opportunity

Avails data generation techniques that may not be available internally and leverages shared expertise in creating and interpreting information - all while reducing costs.

challenge

The methods, controls and technology for sharing data and work brings challenges for data transport, collection, access control and regulatory audits.

Competitive Data Analysis Emerging

The other end of the spectrum of externalization is competitive analysis. Competitions sponsored by companies like Netflix, GE and others provide prizes for the best analytic results for a defined input data set and a well characterized ideal outcome. Kaggle has been gaining attention as a platform for organizations to field and manage competitive/prize funded collaboration. While still in its infancy this approach may develop into significant feature of the data landscape and collaborative milieu for life science companies.

lifescale analytics analytics + insights for life science

Major Challenges for Life Science R&D Teams

The changes in the data landscape for Life Science all point toward increasing diversity of data sources, data formats, complexity and granularity of data that threatens to overwhelm. Velocity of data is so high that much of the raw data from instruments is discarded once its first purpose in the workflow has been completed. Many executives live with a nagging worry that valuable insight is being lost as data is taken off-line into archives or purged completely. "Have we learned all we could from this stream of data? " These increased volumes of data, variety of data structure and velocity of data streams are the locus of what has recently come to be called "Big Data." While the entities, sensors and objectives are different for the Life Sciences than from an Internet business, many of the same changes to the data landscape articulated above are happening across multiple industries. What are some of these Big Data challenges?

Classic data management issues such as information and data architecture, data governance, quality control, data aggregation and transformations don't go away. If anything, solving those issues well and automating them to the greatest extent becomes essential so that users ranging from scientists to executives can trust the data for query and analysis. This is the data plumbing that lets a scientist get to the analysis that will actually yield insight. Unfortunately, many "data scientists" will admit that they still spend most of their time "wrangling data" and a minority of actually building and refining models.

Although servers and workstations are more capable today then ever before, it is still true that transformations and analytical methods that work well with a few gigabytes of data will often choke systems when the scale is expanded to hundreds of gigabytes and require dramatically different, that is, distributed parallel approaches when expanded to terabyte scale. Fortunately, the prevalence of these Big Data problems across multiple industries means that technology companies and open source projects are busy building distributed data management tools and deploying them across public and private cloud infrastructure to meet the challenges.

Unfortunately, many "data scientists" will admit that they still spend most of their time "wrangling data" and a minority of actually building and refining models.

6200 Shingle Creek Parkway, Suite 580 Minneapolis, Minnesota 55430 | 763.585.5870 | info@lifescaleanalytics.com | lifescaleanalytics.com

lifescale analytics | analytics + insights for life science

Impact on Life Science Companies

The changing data landscape impacts many dimensions of operations within a life sciences company, a sampling of issues include:

Business Strategy	How much should we keep internal versus outsourcing or collaboration?
Capital Investments	How do we schedule major lab equipment upgrades in an environment of dropping cost-per-observation?
Team Profiles	What kind of people should we be hiring to take advantage of an increasingly complex and larger data environment? Can we create competitive advantage by being more aggressive in staffing for informatics analysts and "data scientists?"
R&D Team Organization	How do we structure our teams with new data analysis capabilities? Do we embed data scientists within each team? Or, do we have a center of excellence approach, or internal consultants that a team can call on as needed?
Allocation of Resources	How do we best reallocate limited research budgets to invest in new data-centric capabilities?
Supplier Selection	How does one find the right vendors in a rapidly changing landscape? Which of the new tool entrants in Big Data will thrive? How do we minimize risks while trying to capture the benefits of emerging technology tools and new data sources?
Emerging Technology Adoption	How aggressive should our organization be with adopting the newest technologies and following nascent trends?

At LifeScale Analytics, we help life science organizations architect the process, collect, manage and analyze of research data. We understand that there is no one right set of answers for the questions listed above. Your legacy environment and particular business goals may warrant a unique subset of technologies and methodologies. That's where we can help. LifeScale Analytics is business biased not vendor biased, allowing us to help in the selection of tools and vendors that best meet your organization's needs.

Let us help you navigate through this changing data landscape with a focus on improving your return-on-data and ultimately accelerating your speed to market.

6200 Shingle Creek Parkway, Suite 580 Minneapolis, Minnesota 55430 | 763.585.5870 | info@lifescaleanalytics.com | lifescaleanalytics.com

References

Decreasing Cost-Per-Observation
$1000 Genome, http://en.wikipedia.org/wiki/$1,000_genome
High Throughput Screening Retools for the Future, http://www.bio-itworld.com/BioIT_Article.aspx?id=86836
Microfluidics Advances, http://phys.org/news186758376.html

In Situ Measurement on the Rise
Quantified Life, Larry Smarr the measured man
- http://www.theatlantic.com/magazine/archive/2012/07/the-measured-man/309018/
- http://en.wikipedia.org/wiki/Quantified_Self

Ambulatory Stress Monitoring with Minimally-Invasive Wearable Sensors, http://www.cs.tamu.edu/academics/tr/2010-11-1
Data Driven Asthma Management, http://www.healthcare-informatics.com/article/data-driven-asthma-management
Big Data in your Blood, http://bits.blogs.nytimes.com/2012/09/07/big-data-in-your-blood/

Increasing Integration Across Domains
Evidence-based Medicine, http://en.wikipedia.org/wiki/Evidence-based_medicine
Translational Medicine, http://en.wikipedia.org/wiki/Translational_medicine
Directory of Biobanks, http://specimencentral.com/biobank-directory.aspx
Overview of Biobanking, http://mayoresearch.mayo.edu/mayo/research/biobank/about-biobanking.cfm
Monsanto's Plan Puts Focus on Data, http://www.precisionag.com/article/32147/monsantos-plan-puts-focus-on-data

Accelerating Automated Analysis and Decision Loops
High Throughput Screening Retools for the Future, http://www.bio-itworld.com/BioIT_Article.aspx?id=86836
Robot Scientist - iterative Experimentation, http://www.aber.ac.uk/en/cs/research/cb/projects/robotscientist/

Social Graph and Networks Generate New Classes of Data
Social Media Spurs More Clinical Research, http://www.ihealthbeat.org/features/2012/patients-use-of-social-media-spurs-more-clinical-research.aspx
Patient Recruitment via Social Media Lessons Learned, http://blog.pharmexec.com/2012/02/13/patient-recruitment-via-social-media-lessons-learned/

Computable Text Becoming Mainstream
The Open Biological and Biomedical Ontologies, http://www.obofoundry.org/
Mining the Biological Literature, http://www.ebi.ac.uk/2can/databases/TextMiningandBioinformatics.html
Predicting Epidemics via Social Networks, http://blogs.plos.org/everyone/2010/09/17/predicting-epidemics-via-social-networks-an-author-spotlight-on-james-h-fowler-and-nicholas-a-christakis/
Google Flu Trends, http://www.google.org/flutrends/
IBM Supercomputer, Watson, is headed to Medical School - in Cleveland, http://www.cleveland.com/healthfit/index.ssf/2012/10/ibm_supercomputer_watson_is_he.html

Digitized Images Dominating
High Content Screening http://en.wikipedia.org/wiki/High-content_screening
fMRI analysis on the GPU – Possibilities and challenges http://gpuscience.com/articles/fmri-analysis-on-the-gpu-possibilities-and-challenges/
Medical Imaging Advances lead to Storage Headaches
http://www.informationweek.com/healthcare/clinical-systems/medical-imaging-advances-lead-to-storage/219100499

Open Science & Public Data Growing Rapidly
List of Biological Databases, http://en.wikipedia.org/wiki/List_of_biological_databases
Open Access Right to Research, http://www.righttoresearch.org/about/statement/index.shtml
Open Your Minds and Share Your Results, http://www.nature.com/news/open-your-minds-and-share-your-results-1.10895

Collaborative Research Growing
Theoretical Consideration of Collaboration in Scientific Research, http://www.aaas.org/spp/rcp/policy/strategies_book/str6.pdf
Session Descriptions on Externalization, http://www.triconference.com/mmtc_content.aspx?id=118870&libID=118818

Competitive Data Analysis Emerging
Kaggle, http://www.wired.com/business/2012/12/kaggle-crowdsourced-startups
Heritage Health Prize, http://www.heritagehealthprize.com/c/hhp
Netflix Prize, http://www.netflixprize.com//index