

Custom Publishing From:

Sponsored By:

TheScientist
EXPLORING LIFE, INSPIRING INNOVATION


PerkinElmer
For the Better



FINDING VALUE IN VOLUME: APPLICATIONS OF BIG DATA ANALYTICS

Page 3

Predictive Data Analytics:
Driving Research
Breakthroughs

Page 4

The R&D Road to
Progress: Accelerating
Data Analysis

Page 5

Analyze More, Worry
Less: Big Data Solutions

Page 6

Deriving Insights from
Publicly Available Data

Copyright © 2017 PerkinElmer, Inc. 400312A_16 All rights reserved. PerkinElmer® is a registered trademark of PerkinElmer, Inc. All other trademarks are the property of their respective owners.



WHEN DATA CONNECTS,
YOU CAN TAKE
YOUR SCIENCE IN
THE RIGHT DIRECTION

Informatics Solutions:

TIBCO Spotfire®

ChemDraw®

PerkinElmer Signals™

There's *so much* great science buried in your data – problem is, much of it remains isolated and undiscovered in Research, in Clinical, in Manufacturing, in scientists' lab notebooks themselves. Our Informatics solutions connect *all* your disparate islands of information through a unique, comprehensive combination of instrument control, scientific-based applications, and enhanced services – so you can not only manage your data, you can help drive the discovery process. This is how unconnected data becomes *smart* data. PerkinElmer Informatics solutions: Because when your data connects, you can go where your science takes you.

www.perkinelmer.com/informatics



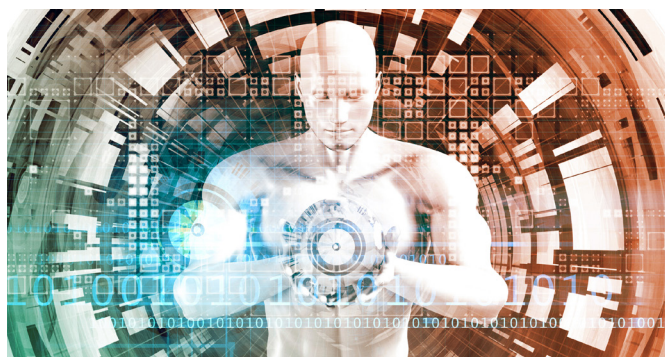
PerkinElmer
For the Better

Predictive Data Analytics: Driving Research Breakthroughs

The ability to gain future insights and make informed predictions is invaluable in the life sciences industry for providing a rapid response to events like global epidemics, vaccine needs, and bringing therapies to market in a time- and cost-effective way. The power to do so comes with the availability of reliable data sets and the ability to gain insights quickly and accurately. This requires statistical methods comprising predictive analytics, namely predictive modeling, machine learning, and data mining.

Predictive modeling, as its name suggests, predicts the probability of an outcome from a given input data set. Predictive models can often be built on regression analysis, which is used to identify relationships between two (or more) variables: a dependent, or target variable, and an independent, or predictive variable.¹ Based on the predictive model, researchers can detect relationships between variables and the impact strength of those relationships to forecast similar occurrences and/or inform future decisions.² Several types of regression analyses exist, each based on their own predefined assumptions, but three are most commonly employed in science and medicine: linear, logistic, and Cox regressions.³ For example, to predict a rise in blood pressure with age, one may assume a linear relationship and therefore work with a linear regression model. Whereas if the predictive probability is a measure of a yes or no outcome (e.g., Was a drug treatment successful?), based on predefined groups, a logistic regression analysis may be appropriate.⁴ Finally, Cox regression models (i.e., proportional hazard regression) are frequently used to model clinical survival data, from diagnosis to death.

Machine learning, like predictive modeling, utilizes mathematical methods to build predictive models from user-supplied inputs. However, machine learning depends on the machine to learn from previously inputted data sets to correctly predict future outcomes. Unlike regression analysis, machine learning does not require the user to make a priori assumptions about the best way to model the data. Instead,



the automated component of machine learning places a black box around the structural relationship between variable(s) and outcome(s).³ The ability of today's machines to be trained to identify repeatable patterns based on prior knowledge (e.g., publications, patient data, experimentally-validated datasets) and then make predictions has numerous applications in scientific research. For example, previously unknown cell-signaling pathway connections and novel prognostic and diagnostic biomarkers can be derived from data, giving researchers new hypotheses as starting points for further validation.

Machine learning is an important addition to the toolbox of data scientists. The goal of data mining in general is to identify novel patterns in data, or to discover knowledge, as it is more commonly known by data scientists.⁴ With machine learning, knowledge is gained from previously known information. This type of analytics is especially powerful in bioinformatics, where the search for patterns based on DNA sequence datasets can uncover powerful evolutionary and disease-related genomic profiles. Machine learning can also improve clinical decision making by learning from vast troves of patient records, which can detail diagnosis and prognosis measurements, particularly in cancer.

References

1. S. Ray, "7 types of regression techniques you should know!" *Analytics Vidhya*, August 14 2015. Available online at <https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>. Accessed September 28 2017.
2. A. Schneider, "Linear Regression Analysis," *Dtsch Arztebl Int* 107(44):776-782, 2010.
3. A. Mayr, "An update on statistical boosting in medicine," *Comp Math Methods Med* 2017:1-12, 2017.
4. B. Read, "Data mining and science?" CLRC Rutherford Appleton Laboratory, *European Research Consortium for Informatics and Mathematics* 1-8, 1999.

The R&D Road to Progress: Accelerating Data Analysis

The research and development journey for a novel compound is a lengthy and exhaustive one. Learn how Big Data approaches can help you cut down on time-consuming validation experiments, accelerating the early discovery stage of the process.



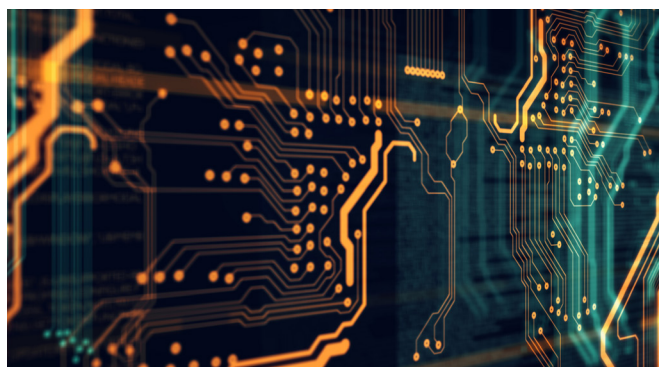
Analyze More, Worry Less: Big Data Solutions

Better answers result from a deeper understanding of your data—essential for taking the right action at the right time. In the field of immuno-oncology, big data technologies handle massive amounts of information, generated by diverse applications such as fluorescence-activated cell sorting cytometry (FACS), genomic analysis, and tissue imaging, while meeting the demand for real-time analysis and data traceability.

Big data is a popular term for data having the Four Vs: volume, variety, velocity, and veracity.¹ As such, large-scale datasets (e.g., patient records, medical images, genomic sequencing, proteomic structure alignment) have created the need for a new generation of computing infrastructure that can outperform the average single or even multi-core processor.² Data-distributing and -processing frameworks like Hadoop and Spark bring together complete analytics in near-real time. Hadoop, an open-source software framework, works to store and distribute massive amounts of data, which, among other available engines, Spark takes on the continuous workload of processing and learning.

Explosive growth in DNA sequencing technologies has brought about massive amounts of genomics data with wide ranging applications in pharmacogenomics, genotype/phenotype association, and patient stratification. With this growth came the need for efficient and accurate analysis, particularly in cancer research. Machine learning methods, based on clustering algorithms or the cancer outlier profile analysis algorithm, for example, mine data and group gene expression (i.e., variables) according to similarities and differences, thereby uncovering patterns in large data sets or patient cohorts not previously recognized.⁵ Spark data-processing algorithms have made this possible at a rate previously unthinkable; it is now possible to convert billions of genomic variants to hundreds of genetic disease drivers in mere minutes.⁵

Biomarker discovery and selection begins with the pinpointing of molecular differences between samples of different cohorts that are usually defined by differences in phenotypes. Over 20 public databases comprising genomic and proteomic information exist and their automated analysis can aid in both the understanding



of disease and biomarker identification.⁵ Meta-analyses of such reference databases can be leveraged through the conception of a target list before performing time-consuming expression and function validation experiments. With machine-performed integration of different data source types (e.g., pharmacological drug profiles, genomic/proteomic databases), researchers can also transform existing data to be predictive of biomarker drug sensitivity.^{2,6}

Such insights, data connectivity, the creation of actionable and repeatable workflows, and visual data analytics help scientists to recognize relationships, patterns, trends, and correlations that lead to discovery and drive future decisions. Computational skills may no longer be an obstacle that each one of us needs to conquer. Instead, scientific domain expertise takes precedence, allowing researchers to ask questions, explore hypothetical scenarios, conduct deeper research, and find more actionable insights.

References

1. C. Ho Lee, et al., "Medical big data: Promise and challenges," *Kidney Res Clin Pract* 36:3-11, 2017.
2. A.E. Mohammed, et al., "Applications of the MapReduce programming framework to clinical big data analysis: Current landscape and future trends," *BigData Mining* 7(22):1-23, 2014.
3. J. Hyrkas, et al., "Scalable clustering algorithms for continuous environmental flow cytometry," *Bioinformatics* 2015:1-7, 2015.
4. E. Lugli, et al., "Data analysis in flow cytometry: The future just started," *Cytometry A* 77(7):705-713, 2010.
5. A. Woodie, "How Spark and Hadoop are advancing cancer research," *datanami*, May 23, 2016. Available online at <https://www.datanami.com/2016/05/23/spark-hadoop-advancing-cancer-research/>. Accessed October 3 2017.
6. B. Chen and A.J. Butte, "Leveraging big data to transform target selection and drug discovery," *CPT* 99(3):285-297, 2016.

Deriving Insights from Publicly Available Data



In an effort to make discoveries, scientific organizations had developed a tendency to internalize and view data as proprietary. With drug pipelines increasingly depleted, that inward focus has shifted as technology advanced, data exploded, and publicly accessible databases began to appear online. The website re3data.org is the largest registry of public repositories, with 1,500 listed. In addition, organizations such as the Bill and Melinda Gates Foundation started to mandate that data produced from their funding must be “promptly and broadly disseminated.”

The benefits of public access are numerous, from researchers verifying findings using broader data sets to enabling web-scale analytics.¹ Big data is helping life science researchers tap into real-world health and clinical data to identify unmet needs and to develop new therapies. In addition to biotech and pharma industries, such benefits have been recognized by clinical researchers and decision makers. Electronic medical records, high-resolution medical imaging, and next-generation genomics are best represented, analyzed, and stored via the ever-evolving digital data ecosystem.²

When offered, publicly accessible databases can help shape new understandings and insights. For example, the Broad Institute, in cooperation with the Novartis Institutes for BioMedical Research and the Genomics Institute of the Novartis Research Foundation, has compiled the Cancer Cell Line Encyclopedia (CCLE) to conduct a detailed genetic characterization of a large panel of human cancer cell lines.

CCLE is a publicly available collection containing gene expression, chromosomal copy number, and sequencing data across 947 human cancer cell lines, as well as pharmacologic profiles for 24 anticancer drugs across 479 of the lines. Mining this database greatly aids in the identification and profiling of human cancers.

Alexandra Vamvakidou, PhD, a field application scientist at PerkinElmer, has demonstrated how the CCLE cancer cell line models can be used to explore the genetic characterization and drug sensitivity of tumor biology using highly specialized analysis and visualization software. Commercially available software

solutions create dynamic representations of massive databases. For example, Dr. Vamvakidou was able to compare pharmacological data across different cancer cell lines and determine the gene signatures across a variety of tumor types. Computational and integrated algorithms measure significance on trends observed, such as statistical analyses to determine significant changes in CCLE profiles, and the probing of OMICS and structure-activity relationship (SAR) data.

Software solutions can quickly visualize the CCLE data set for distribution of cancers by tissue type. Dr. Vamvakidou, for example, was interested in seeing how heterogenous samples behave and if any common relationships exist across multiple tissue types. Once identified, researchers can also explore the biological testing data of drugs used on a specific tumor cell line against data downloaded from the Drug Bank, an open data–drug and drug–target database. Other features of visual analytics taken advantage of by Dr. Vamvakidou include the ability to drill down, zoom in, overlay, filter, create hierarchies, and otherwise interact with CCLE and Drug Bank data to explore hypotheses on the fly.

For researchers facing the challenges of easily accessing and comparing all available data; searching, analyzing, and reporting their findings; and being able to store information for large studies, visual analytics platforms can leverage publicly accessible databases and present data dynamically to deliver the fresh insights that lead to important discoveries.

References

1. M. Taylor, “Should Research Data Be Publicly Available?” Elsevier, 30 May 2013. Available online at elsevier.com/connect/should-research-data-be-publicly-available. Accessed 30 October 2017
2. B. Srinivasan, et al., “2016 Technology Vision for Life Sciences.” *Accenture*, 2017. Available online at accenture.com/us-en/insight-life-sciences-tech-vision-2016. Accessed October 30, 2017
This article is based on the following webinar:
3. A. Vamvakidou, “Deriving Insights from Publicly Available Data: A Case Study on Exploring CCLE Profiles for Validated Oncology Research,” PerkinElmer Webinar, June 15, 2017. Available online at <http://info.perkinelmer.com/2017-Website-DerivingInsightsFromPubliclyWebinar-LP>. Accessed October 30, 2017.



BECAUSE BREAKTHROUGHS CAN'T WAIT

SCREENING SOLUTIONS

Automated Liquid Handling

Assays and Reagents

Imaging and Detection Systems

Informatics

Imaging and detection systems, automation, informatics, and assays and reagents for high-throughput and phenotypic screening – together, they spell *discovery*.

Drug discovery is critical to our health and well-being. And to get therapies to market that much sooner, you need to accelerate your workflow, upstream and downstream.

Our screening solutions are a crucial component in that effort: State-of-the art imaging and detection instruments, assay technologies and reagents, and versatile automation systems, working together to ensure consistent, accurate, physiologically relevant results – with one-of-a-kind analytics to wrest more meaning from your findings.

The *next* big breakthrough? It begins with you.

To see how *our* screening solutions can lead to *your* next big breakthrough, visit www.perkinelmer.com/screeningsolutions


PerkinElmer[®]
For the Better