Innovation and Your Career in Rapidly Changing World of Job Titles: How Different is Data Scientist from What I do?

Richelle Serrano, Clindata Insight Inc., Moraga, CA

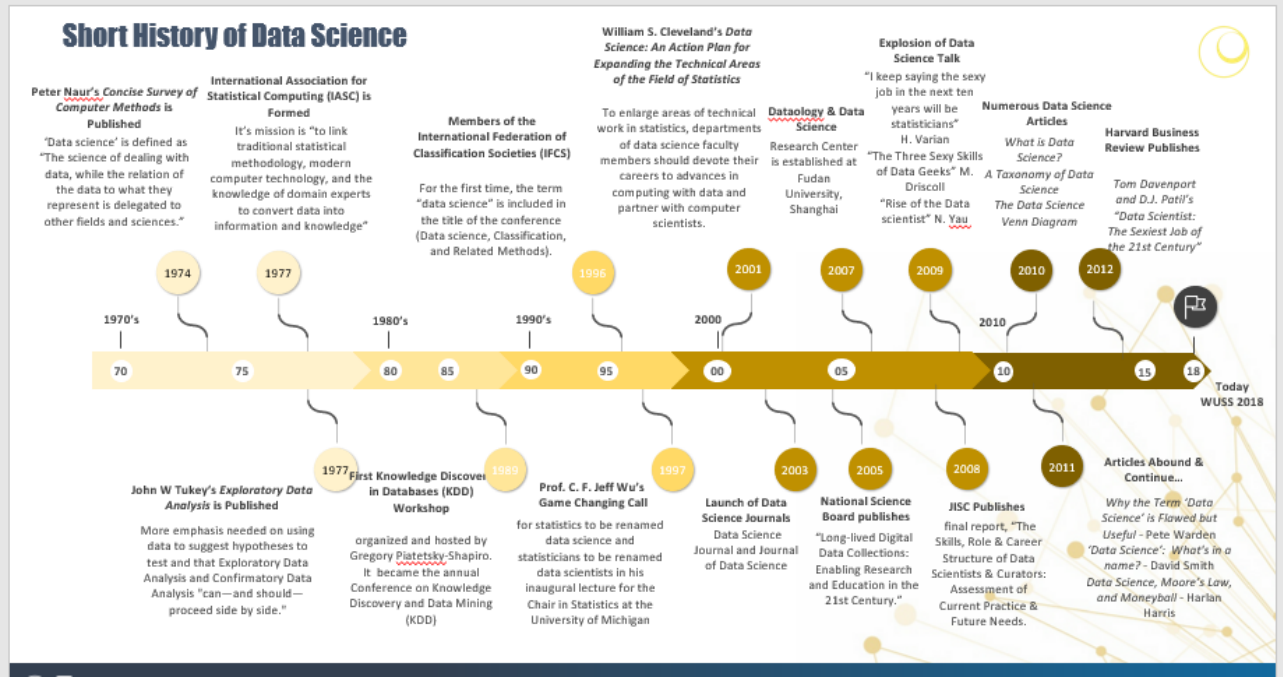
ABSTRACT

Data science employs techniques and theories drawn from mathematics, statistics, information science, and computer science to understand and analyze a fact or situation. In the last 15-20 years data science has become an independent discipline. In 2012 the Harvard Business Review published the article "Data Scientist: The Sexiest Job of the 21st Century". It's no surprise the job of Data Scientist is growing faster than Statistician or Statistical Programmer. This paper will explore the role of a data scientist as defined today and the overlap with statisticians and statistical programmer's traditional roles. Next the paper will discuss the skill set and knowledge needed for a statistical programmer to stay current in the evolving world of scientific data and information becoming increasingly available and accessible through open science and open source.

INTRODUCTION

It's practically impossible not to have noticed the explosion of the term Data Science. Not only if you are working in fields that benefit most from big data, healthcare, science, retail, insurance, agriculture but every business in the world needs data to thrive. Data is what tells you who your customers are and how they operate, and it's what can guide you to new insights and new innovations. Generally, the more data you have, the more specific and accurate insights you'll be able to generate. To show where we are now, I looked back into the history of the term, Data Science. One article in particular, "A Very Short History of Data Science written by Gil Press, published in Forbes, May 2013 traces the evolutions of the term "Data Science" from the 1960's to the current decade.

What I liked about this article is the premise that making sense of data has a long history and has been discussed by scientists, statisticians, librarians, computer scientists and others for decades. What's new is the combining of a mature discipline of statistics with a young one, computer science and the emergence of a new profession "the Data Scientist".



DATA SCIENTIST AND STATISTICAL PROGRAMMER- COMPARING THE NUMBER OF JOBS

Numerous articles, publications, workshops, and courses have been written and developed to evaluate the emergence of data science as a discipline and its alignment with the challenges of addressing complex and dynamic digital data in science and society. This talk will not address this broader issue and instead discuss the overlap of data scientists with statisticians and statistical programmers, in pharma, and other industries working on clinical trials, i.e. medical device, diagnostics, biotech.

I began my research for this paper by making a simple comparison of the number of open jobs for data scientist (pharmaceutical) and statistical programmer (pharmaceutical) on Glassdoor. I included the term 'pharmaceutical' to the search string to reduce the number of employers outside our industry. It was not entirely effective, but the results were more relevant with the term, than without it.

Consistently, the data suggested 5 Data Scientist jobs for every 1 Statistical Programmer job. Note this ratio does not correspond to all industries. Across all industries the number of data scientists outpaces programmers/software developers more than 10-fold.

As you would expect the Pharmaceutical Statistical Programmer in regulated environments does differ from a Programmer/Software Developer in other industries for several reasons.

Pharmaceutical programmers have industry specific knowledge.

a. CDISC Standards - With increasing adoption of CDISC standards, the pharmaceutical programmer must have ability to implement CDISC SDTM and ADaM data standards for job. Adherence to standards aligns with how software developers/programmers look at things with building blocks in mind for reusability, consistency and efficiency.

b. Pharmaceutical programmers have specific therapeutic area knowledge and are aware that other therapeutic areas may each require a different approach.

a. Data scientists focus on building predictive models and developing supervised and unsupervised machine learning capabilities to analyze the data. Pharmaceutical Statistical Programmers focus more on applying a set of statistical models to clinical trial data.

b. Pharmaceutical programmers work in a regulated environment, following Good Clinical Practice (GCP) and ICH guidelines. Data scientist in other disciplines may face much less regulation.

DATA SCIENTIST AND STATISTICAL PROGRAMMER- COMPARING JOBS

One comparison I've heard is that a Data Scientist job is work that takes more programming skills than most statisticians have, and more statistics skills than a programmer has. Despite many comparisons there exists chronic confusion as to what the distinction is between the average programmer and a data scientist, which is totally understandable, considering both jobs do involve programming.

I created a sample job description for a Data Scientist and Statistical Programmer. And then looked at them from a recruiter perspective on how the responsibilities and qualifications differed in the function of the job.

SAMPLE JOB DESCRIPTIONS: Data Scientist and Statistical Programmer

Data Scientist

Data Scientist will execute advanced computational approaches to accelerate and optimize evidence-based pharmaceutical product development. Will leverage high-dimensional population health data to support R&D, Medical, Health Economics, commercial product development and business strategy.

Advanced Analytics Activities

- Generating insights by merging diverse data streams e.g. health, surveillance, trend data, sensor, imaging.
- Machine/Deep Learning to elucidate disease trajectories, patient subtypes, define underdiagnosed conditions, and unmet health needs.

Job Responsibilities

- Apply a broad array of analytics skills including machine learning, statistics, text-mining/NLP, and modeling to extract insights to structured and unstructured healthcare data sources, pre-clinical, clinical trial and complementary real-world & digital information streams.
- Work on a variety of team-based projects providing expertise in analytical and computational approaches.
- Work with the latest tools and technologies to impact drug development
- Contribute to the Advanced Analytics plans for projects across R&D, Medical Affairs, Health Economics and Market Access Strategies and Plans.
- Leverage analytics involving large datasets to refine and improve data models.
- Build and construct prototypes of Advanced Analytic work-flows.

Qualifications

- PhD or ScD in quantitative field such as Medical Informatics, Mathematics, Biostatistics, or Computer Science, Engineering or related field
- Relevant Master's Degree, with at least 2 years of industry experience. Bachelor's degree with at least 3 years in a relevant discipline and commensurate analytics experience will be considered.
- Experience with open source technologies, ML libraries, and programming languages (R, python)
- An ability to interact with a variety of large-scale data structures e.g. HDFS, SQL, noSQL
- Experience working across multiple compute environments to create workflows and pipelines (e.g. HPC, cloud, Linux systems)
- Experience with biomedical data types/population health data/real world data/novel data streams.
- Experience with reproducible and collaborative technology platforms (e.g. github, containers, jupyter notebooks)
- Experience with big data analytics platforms and/or workflow tools
- Exposure to NLP technologies and analyses
- Knowledge of some data visualization technologies (ggplot2, shiny, plotly, d3, Tableau or Spotfire)

Senior Statistical Programmer

Statistical Programmer will design, develop, evaluate and modify computer programs to analyze and evaluate clinical data. Generate study specific and ad-hoc clinical data listings, summary tables and figures.

Job Responsibilities

- Performs data analysis using primarily the SAS programming language for the summary and interpretation of clinical trial data.
- Create CDISC SDTM and ADaM files and SAS export files to use for Electronic Submission from the data received in non-standard form from various sources.
- Performs Data analysis, statistical analysis, generate safety and efficacy tables, listings and graphs using Base SAS, SAS Macros, SAS/STAT, SAS/Graph, SAS/SQL and SAS/ODS.
- Review Data Management Plan, Data validation plan and edit check specifications
- Interact with Statisticians and other clinical team, perform ad hoc analysis and generate outputs.
- Recognizes inconsistencies and initiates resolution of data problems.
- Acts as a liaison between statistical programming, subcommittees and project teams as needed.
- Exercises independent judgment in developing methods, techniques and evaluation criteria for obtaining results.
- Works on significant and unique issues where analysis of situations or data requires evaluation of intangibles.
- Acts independently to determine methods and procedures on new assignments.
- May provide guidance to other lower level personnel.
- Support lead programmer in generation of analysis datasets, generation Tables, Listings, and Figures validation of data and reports, and maintenance of study documentation
- Generate and QC summary tables, data listings and figures for in-house analyses of study data or publications using SAS standard coding practices.
- Validate work of other programmer/analysts at CRO or in-house.
- Maintain complete and auditable documentation of all programming activities.

Qualifications

- BS in Statistics, Math or Scientific Discipline.
- Ability to use professional concepts to achieve objectives in creative and effective ways.
- Experience in the analysis of specific therapeutic areas clinical trial data a plus.
- Knowledge of CDISC standard (SDTM & ADaM).
- Minimum 5 years Pharmaceutical/Biotech programming experience.
- NDA submission experience is a plus
- Strong SAS programming and Statistical background along with experience with SAS Base, SAS/Macros, SAS/Graph and SAS/Stat.

COMPARISON TABLE: Data Scientist and Statistical Programmer

	Data Scientist	Statistical Programmer
Summary	-Select and apply a computational approach to advance the product development. -Identify data that supports multi-functional groups progress and objectives.	-Designs, develops, modifies programs to analyze & evaluate clinical trial data. -Create study specific and ad hoc listings summary tables figures.
Skills	-Apply a variety of analytics to identify trends or insights from multiple data sources and info streams. -Knowledge across many prog. languages and technologies, less of any expert in any one. -Using analytics to evaluate retrospective data vs prospective data. -Construct prototypes of analytics workflows -Design and construct new processes for data modeling and production using prototypes, algorithms, predictive models, and custom analysis. -Communicates to multiple functional groups across the organization.	-Statistical programming expertise (SAS, R) -CDISC hands on experience -Study lifecycle from Protocol, SAP, eCRF, SDTM, ADaM, Define.xml to TFL -Therapeutic area expertise -Regulatory support – ISS/ISE data integration, reports, SDRG and ADRG prep and review. -Ability to address questions from other functions/partners: CRO's, Medical Affairs and statisticians. -Manage and audit programming documentation -Communicates with Biometrics groups, CRO's, Data Managers, Medical Affairs and statisticians.
Software	Python, R, SAS, Java, Scala Programming SQL, Oracle, Hadoop, NoSQL Jupyter notebook Tensorflow, sci-kit learn, Keras, H2O	SAS, R Programming, SAS datasets, CSV files
Education	MS or PhD in Mathematics, Statistics, Data Science	BS or MS in Computer Science, Mathematics, Statistics
Other Observations	Data scientists communicate in statistics backed with mathematics	More lenient on education in exchange for experience Ability to understand data coming in and expected out. Programmers communicate under the convention of other programmers written code.

STATISTICAL PROGRAMMER- KEEPING UP TO SPEED

From my perspective, sitting in an office in the San Francisco Bay Area and supporting pharmaceutical, biotech, and medical devices clients with their hiring needs for biostatisticians, statistical programmers and data management roles, your jobs are not only in high demand now, it is still one of the best jobs for the future. There is no question there is more data than ever before. More data was created in 2017 than the previous 5,000 years of humanity. Your experience and skills to manage, understand, and communicate findings is crucial to making sense of the rapidly changing environment of today.

Advancement in any field relies on continual learning and to challenge yourself to think out of the box. If I were to suggest what more, you could do to enhance your skills and advance your career it would be:

- Learn what you can about the full lifecycle of the study. Pay attention to the study challenges even if they don't apply to your work.

- Get hands on experience across therapeutic areas. This skill cannot be replaced by data science workshops or masters in statistics.

- Stay abreast of latest cutting-edge techniques in stats, be open minded to open source programs (R, Python, etc.), and pay attention to technologies that offer better ways to communicate and share data (Tableau, Spotfire, ggplot2, shiny, etc.).

- Be open to the challenge of incorporating data from many sources (structured data, unstructured data) and propose solutions to handle difficult and new types of data.

- Be aware of Real-World Evidence to enrich the understanding of the data.
- Learn more statistics – take courses, go to workshops, and study on your own.
- Always work on communication skills. It is not only a life skill, but essential skill for a statistical programmer to communicate findings and make recommendations in ways the organization's leadership can understand and act on. This is key to career advancement in any job.

CONCLUSION

Much of what you do overlaps with the new and upcoming role of a Data Science. Your specialized work and industry specific knowledge cannot be replaced with what a fresh data science graduate learns. The key to staying competitive in your job is to embrace todays challenges in data analysis and to not shy away from new approaches.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Richelle Serrano
Clindata Insight
(925) 385-3017
rserrano@clindatainsight.com
LinkedIn: https://www.linkedin.com/in/richelle-serrano-4b00102/

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies.