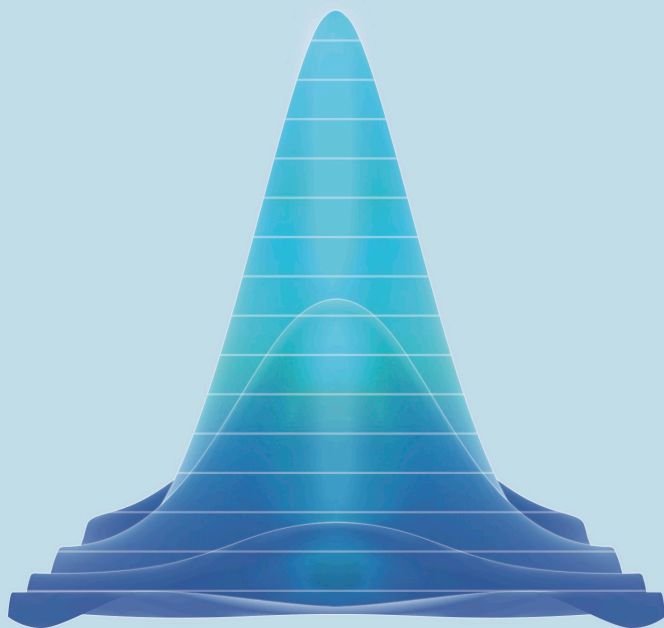# THE
DATA SCIENCE
HANDBOOK

ADVICE AND INSIGHTS FROM
25 AMAZING DATA SCIENTISTS

DJ **Patil,** Hilary **Mason,** Pete **Skomoroch,** Riley **Newman,** Jonathan **Goldman,** Michael **Hochster,** George **Roumeliotis,** Kevin **Novak,** Jace **Kohlmeier,** Chris **Moody,** Erich **Owens,** Luis **Sanchez,** Eithon **Cadag,** Sean **Gourley,** Clare **Corthell,** Diane **Wu,** Joe **Blitzstein,** Josh **Wills,** Bradley **Voytek,** Michelangelo **D'Agostino,** Mike **Dewar,** Kunal **Punera,** William **Chen,** John **Foreman,** Drew **Conway**

CARL **SHAN**   HENRY **WANG**   WILLIAM **CHEN**   MAX **SONG**

The Data Science Handbook

Advice and Insights from 25 Amazing Data Scientists

Carl Shan, Henry Wang, William Chen and Max Song

This book is for sale at http://leanpub.com/datasciencehandbook

This version was published on 2015-05-04

To our parents, siblings, friends and mentors. Your support and encouragement has been the fuel for our fire.

Contents

Foreword by Jake Klamka

In the past five years, data science has gone from a nascent, tech industry competency to a field that is having a global, cross-industry impact in almost every major area of human endeavour. From education, to energy, to government, to non-profits and, of course, software and the Internet, data science is creating immense value for companies and organizations across the world. In fact, in early 2015, the President of the United States announced the creation of the new role of Chief Data Scientist to the White House, appointing one of the interviewees of this book, DJ Patil.

Like many innovations in the world, the birth and growth of this industry was started by a few motivated people. Over the last few years, they founded, developed and advocated for the value that data analytics can bring to every industry around the world. In The Data Science Handbook, you will have the opportunity to meet many of these founding data scientists, hear first hand accounts of the incredible journeys they took, and where they think the field is headed.

The road to becoming a data scientist is not always an easy one. When I tried to transition from experimental particle physics to industry, resources were few and far between. In fact, although a need for data science existed in companies, the job title had not been created yet. I spent a lot of time learning and teaching myself, working on various startup projects, and later saw many of my friends from academia run into the same challenges.

I saw a groundswell of incredibly gifted and highly trained researchers who were excited about moving into data-driven roles, yet they were missing key pieces of knowledge, and had trouble transferring the incredible quantitative and data analysis skills they

had gained in their research to a career in industry. Meanwhile, having lived and worked in Silicon Valley, I also saw that there was a very strong demand from the technology companies who wanted to hire these people.

To help others bridge the gap between academia and industry, I founded the Insight Data Science Fellows Program[1] in 2012. Insight is a training fellowship that helps quantitative PhDs transition from academia to industry. Over the last few years, we've helped hundreds of Insight Fellows, from fields like physics, computational biology, neuroscience, math, and engineering transition from a background in academia to become leading data scientists at companies like Facebook, Airbnb, LinkedIn, New York Times, Memorial Sloan Kettering Cancer Center and nearly a hundred other companies, with a strong alumni network on both the East and West Coast.

In my personal journey to enter the technology field, and creating a community for others to do the same, one key resource I found to be tremendously useful was conversations with others who had successfully made the transition themselves. As I developed Insight, I have had the chance to engage with some of Silicon Valley's best data scientists who are mentors to the program:

Jonathan Goldman created one of the first data products at LinkedIn–People You May Know–which transformed the growth trajectory of the company. DJ Patil build and grew the data science team at LinkedIn into a powerhouse and in the process co-coined the term "Data Scientist." Riley Newman worked on developing product analytics that was instrumental in Airbnb's growth. Jace Kohlmeier led the data team at Khan Academy that helped to define how to optimize learning at a scale of millions of students.

Unfortunately, face-to-face time with people has trouble scaling. At Insight, to maintain an exceptional high quality and personal time with its mentors, we accept a small group of talented scientists and engineers three times per year. The Data Science Handbook

---

[1]http://insightdatascience.com/

provides readers with a way to have that in-depth conversation at scale. By reading the interviews in The Data Science Handbook, you will have the experience of learning from the leaders in data science at your own pace, no matter where you are in the world. Each interview is an in-depth conversation, covering the personal stories of these data scientists from their initial experiences that helped them find their own path to a career in data science.

It's not just the early data science leaders who can have a big impact on the field. There is also new talent entering the field, with the opportunity for each and every new member to push the field forward. When I met the authors of this book, they were still college students and aspiring data scientists, full of the same questions that those beginning in data science have. Through 18 months of hard work, they have gone and done the legwork for all those interested, seeking out some of the best data scientists around the country, and asking them for their advice and guidance. This book is the result of that work, containing over 100 hours of collected wisdom with people otherwise inaccessible to talk to (imagine having to compete with President Obama to talk with DJ Patil!). In the meantime, these young authors also have gone on to earn their own stripes as data scientists, working at some well-known companies.

By reading these extended, informal interviews, you will get to sit down with industry trailblazers like DJ Patil, Jonathan Goldman and Pete Skomoroch, who were all part of the core, early LinkedIn data science teams. You will meet with Hilary Mason and Drew Conway, who were instrumental in creating the thriving New York data science community. You will hear advice from the next generation of data science leaders, like Diane Wu and Chris Moody, both former PhDs and Insight Alumni, who are now blazing new trails at MetaMinds and Stitch Fix. You will meet data scientists who are having a big impact in academia, including Bradley Voytek from UCSD and Joe Blitzstein from Harvard. You will meet data scientists in startups like Clare Corthell from Mattermark and Kunal Punera of Bento Labs, who will share how they use data science and

analytics as a core competitive advantage.

The data scientists in the Data Science Handbook, along with dozens of others, have helped create the very industry that is now having such a tremendous impact on the world. Here, in this book, they discuss the mindset that allowed them to create this industry, address misconceptions about the field, share stories of specific challenges and victories, and talk about what skills they look for when building their teams. By reading their stories, hearing how they think and learning about where they see the future of data science going, you will gain the context to think of ways you can both have an impact and perhaps advance the field yourself in the years to come.

Jake Klamka

Founder

Insight Data Science Fellows Program[2]

Insight Data Engineering Fellows Program[3]

Insight Health Data Science Fellows Program[4]

---

[2]http://insightdatascience.com/
[3]http://insightdataengineering.com/
[4]http://insighthealthdata.com/