Seminar: Beiträge zum Software Engineering

Data Science in Praxis: Umfang und Techniken

Oleksandr Tepliak

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

MENU

Harvard Business Review

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY | SAVE | SHARE | 4 COMMENT | TEXT SIZE | PRINT | $8.95 BUY COPIES

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join.

Fragen:

1. Welche Fertigkeiten und Kompetenzen machen einen Data Scientist aus?

2. Welche Methoden und Technicken benutzt bzw. bietet Data Science an (sind das die numerischen Methoden, lineare Algebra, Statistik und Softwaretechnik, in welchen Proportionen sind die in Data Science Praxis vertreten)?

3. Welche Tools werden von Datasceintists eingesetzt?

4. Wie ist der Zusammenhang Data Science, Machine Learning, Big Data?

5. In welche Phasen der Softwareentwicklung werden die Data Scientists eingesetzt?

6. Welche Rolle spielen die Data Scientists bei Softwareentwicklung?

Begriffe:

Telemetry is the automatic recording and transmission of data from remote or inaccessible sources to an IT system in a different location for monitoring and analysis.

Data Mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.. In essential process where intelligent methods are applied to extract data patterns. It is an interdisciplinary subfield of computer science.

Basis-Quellen:

“The emerging role of data scientists on software development teams”

Published in: · Proceeding ICSE '16 Proceedings of the 38th International Conference on Software Engineering Pages 96-107

Autoren:

Miryung Kim UCLA Los Angeles, CA (University of California)

Thomas Zimmermann Microsoft Research, Redmond, WA

Robert DeLine Microsoft Research, Redmond, WA

Andrew Begel Microsoft Research, Redmond, WA

Basis-Quellen:

“Data Science: A Comprehensive Overview”

Journal ACM Computing Surveys (CSUR) Surveys Homepage archive Volume 50 Issue 3, October 2017 Article No. 43

Autor : Longbing Cao, University of Technology Sydney (Member, Association for Computing Machinery Senior Member, Institution of Electrical and Electronic Engineers)

Data Science-Bereiche nach Harris et. al

- Data business people
- Data creatives
- Data developers
- Data researchers

Buch:

H. D. Harris, S. P. Murphy and M. Vaisman, Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work,2013

Fisher et. al.

Challenges in big data computing platforms in:

- ➔ data integration;
- ➔ cloud computing cost estimation;
- ➔ difficulties shaping data to the computing platform;
- ➔ the need for fast iteration on the analysis results

DATA SCIENTISTS IN SOFTWARE DEVELOPMENT TEAMS

Anmerkungen:

1. Microsoft (ein sehr großes Unternehmen)
2. Data-driven decision making: Softwareentwicklung bzw. Evolution basiert auf Data aus z.B. Windows Error Reporting tool
3. Die Rolle war im Prozess der Definierung z.Z. der Studie
4. Die Daten, die bei der Microsoft-Produkten Telemetry entstehen, werden als Ersatz des Testens betrachtet! So genannte “ test-in -production paradigm”

The emerging role of data scientists… Die Befragten:

TABLE 1. PARTICIPANT INFORMATION

	Title	Education
P1	Data Scientist II	BS in CS / Statistics, MS in SE, currently pursuing PhD in Informatics
P2	Director, App Statistics Engineer	MS in Physics
P3	Principal Data Scientist	MBA, BS in Physics / CS, currently pursuing PhD in Statistics
P4	Principal Quality Manager	BS in CS
P5	Partner Data Science Architect	PhD in Applied Mathematics
P6	Principal Data Scientist	PhD in Physics
P7	Research Software Design Engineer II	MS in Computer Science, MS in Statistics
P8	Program Manager	BS in Cognitive Science
P9	Senior Program Manager	BSE in CS and BAS in Economics/Finance
P10	Director of Test	BS in CS
P11	Principal Dev Manager	MS in CS
P12	Data Scientist	PhD in CS / Machine Learning
P13	Applied Scientist	PhD in CS / Machine Learning and Database
P14	Principal Group Program Manager	BS in business
P15	Director of Data Science	PhD in CS / Machine Learning
P16	Senior Data Scientist	PhD in CS / Machine Learning

Warum braucht man Data Scientists in SW-Teams:

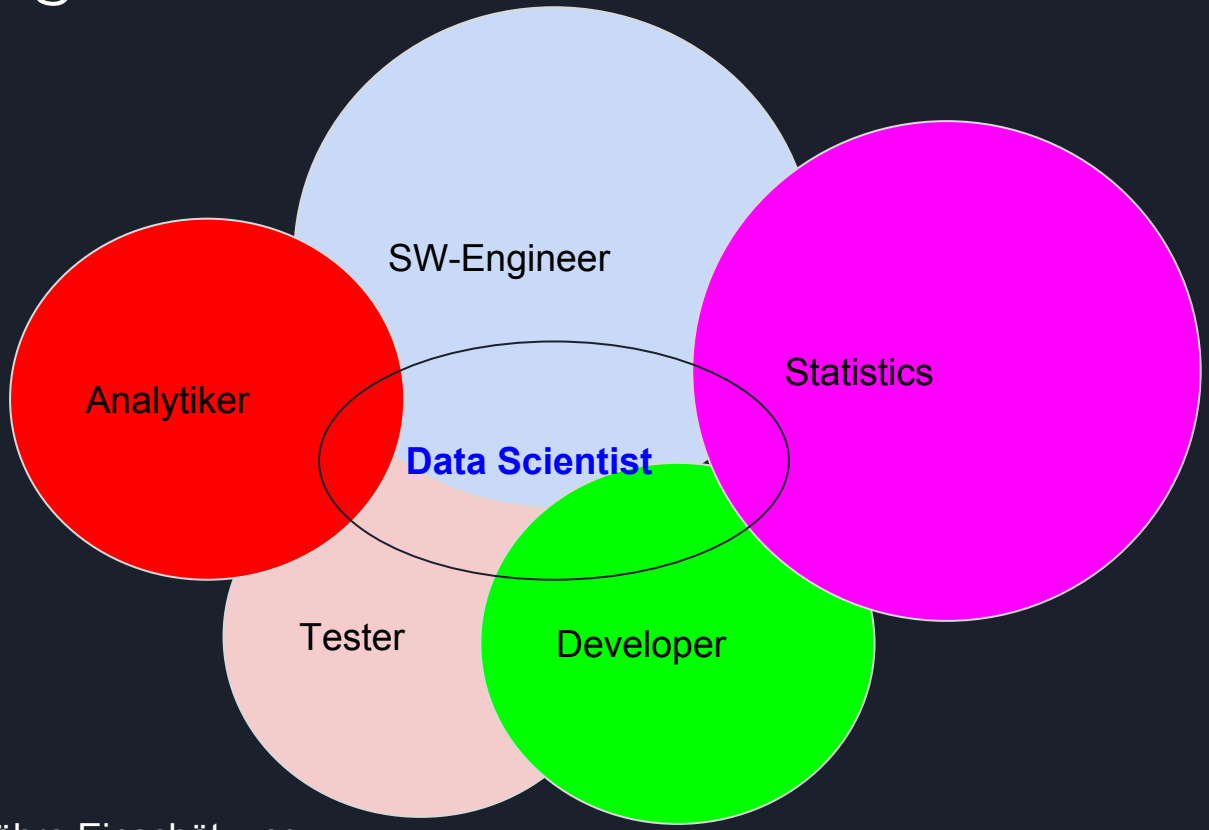
1.Demand for Experimentation (z.B. Feature hinzufügen und nach einer Zeit wegnehmen)

3. Demand for Data Collection Rigor (Welche Daten sammeln, wie kann man die sauber behalten und schnell bearbeiten -80% “janitar work”)

2. Demand for Statistical Rigor (Die Frage welche Zustände, telemetrische Daten normal sind, und welche Max, Min, Anomalie sind)

Background von Data Scientists

Fertigkeiten*:



*ungefähre Einschätzung

Aufgaben eines Data Scientists:

- Performance Regression
- Requirements Identification
- Fault Localization and Root Cause Analysis.
- Bug Prioritization
- Server Anomaly Detection.Cost Benefit Analysis
- Failure Rate Estimation
- Customer Understanding

Aktivitäten:
Collection:

- ★ Data engineering platform;
- ★ Telemetry injection;
- ★ Experimentation platform: “building inherent capability for experimentation with alternative software designs”

Aktivitäten:
Analysieren:

- ★ Data merging and cleaning;
- ★ Sampling:
- ★ Data shaping including selecting and creating features:
- ★ Defining sensible metrics
- ★ Building predictive models:
- ★ Defining ground truths;
- ★ Hypothesis testing:

Aktivitäten:
Use and Dissemination:

- ★ Operationalizing predictive models;
- ★ Defining actions and triggers;
- ★ Translating insights and models to business values;

Aktivitäten pro Person:

TABLE 2. ACTIVITIES THAT PARTICIPANTS STATED THEY DID THEMSELVES (■) OR MANAGED (□)

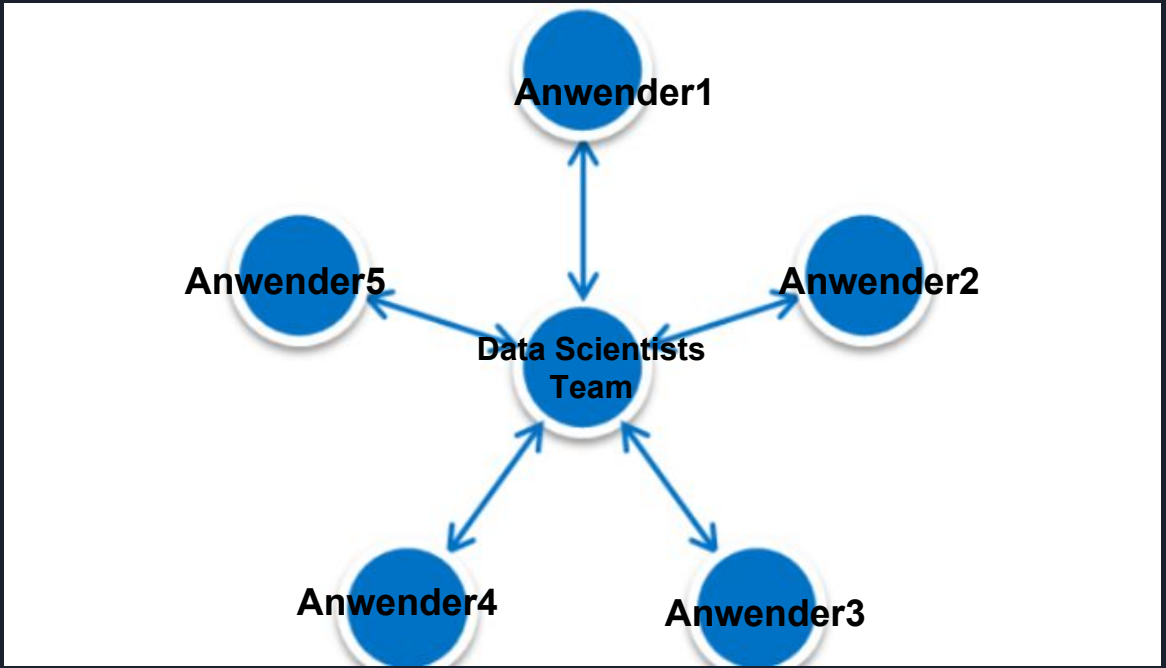
		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
Collecting	Building the data collection platform	■			■				■			■		■	□		
	Injecting telemetry	■	□		■				■		□	■		■	□		
	Building the experimentation platform	■													□		
Analyzing	Data merging and cleaning	■	■	■	■	■	■	■	■	■	□	■		■	□		
	Sampling	■	■	■	■	■	■	■		■	□	■	■	■	□	■	■
	Shaping, feature selection		■	■	■	■	■	■			□		■	■	□	■	■
	Defining sensible metrics	■			■	■	■	■			□	■		■	□		■
	Building predictive models		■	■		■	■	■		■	□		■	■	□	■	■
	Defining ground truth							■		■			■	■	□	■	■
	Hypothesis testing		■	■		■	■				□				□		■
Using and Disseminating	Operationalizing models						■	■		■	□		■	■	□	■	
	Defining actions and triggers									■	■	■		■	□		
	Applying insights/models to business	■	■	■	■	■				■	■	■	■	■	□	■	■

Impact <-> Actionabilaty

Data Scientists and Developers' Teams:
Das "Dreieck-"Model

Data Scientists and Developers' Teams:

The
"Hub and Spoke" model.



Data Scientists and Developers' Teams:
Other models

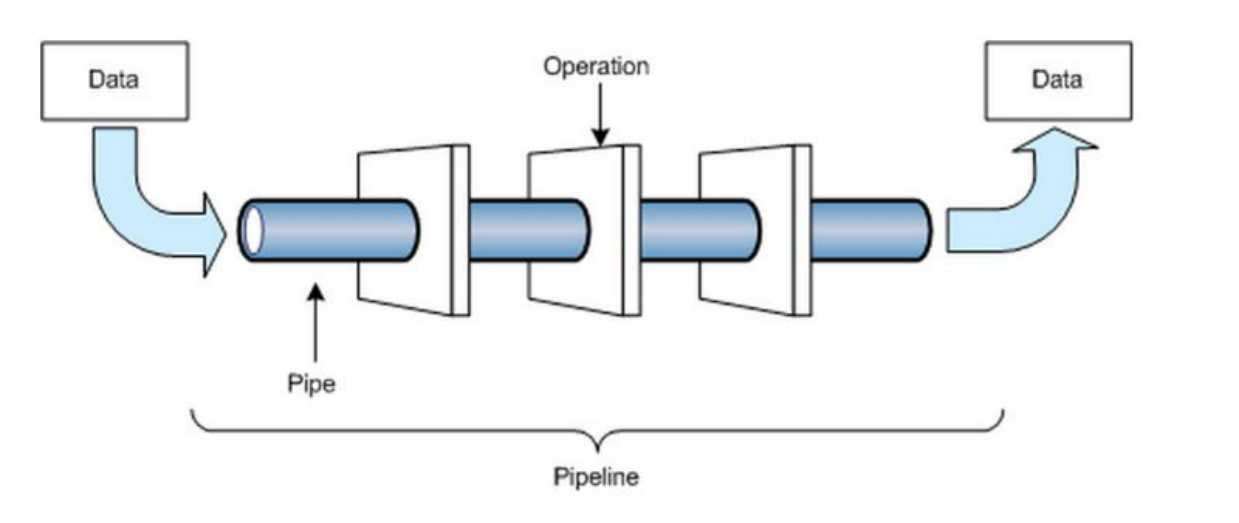
The "Consulting" model.

The "Individual Contributor".

The "Virtual Team"

Arbeitsorganisation:

Data engineering pipe-line:



Die Funktionen:

Insight Providers:

‘You need to think about, “ If you find this anomaly, then what? ” Just finding an anomaly is not very actionable. What I do also involves thinking, “ These are the anomalies I want them to detect. Based on these anomalies, I’m going to stop the build. I’m going to communicate to the customer and ask them to fix something on their side.’ [P9]

Die Funktionen:

Modeling Specialists:

“They accepted [the model] and they understood all the results and they were very excited about it. Then, there’s a phase that comes in where the actual model has to go into production. ... You really need to have somebody who is confident enough to take this from a dev side of things”. [P12]

Die Funktionen:

Platform Builders:

“If you could survey everybody every ten minutes, you don't need telemetry. The most accurate is to ask everybody all the time. The only reason we do telemetry is that [asking people all the time] is slow and by the time you got it, you ’ re too late. So you can consider telemetry and data an optimization. So what we do typically is 10% are surveyed and we get telemetry. And then we calibrate and infer what the other 90% have said.” [P4]

Die Funktionen:

Polymaths:

“So I am the only scientist on this team. I'm the only scientist on sort of sibling teams and everybody else around me are like just straight-up engineers.

For months at a time I'll wear a dev hat and I actually really enjoy that, too. ... I spend maybe three months doing some analysis and maybe three months doing some coding that is to integrate whatever I did into the product. ... I do really, really like my role. I love the flexibility that I can go from being developer to being an analyst and kind of go back and forth.”[P13]

Team Leaders: pushing results through

Implications:

Vielen Dank.