The Emerging Role of Data Scientists on Software Development Teams

Miryung Kim, Thomas Zimmermann, Robert DeLine, Andrew Begel

UFMG

ICSE 2016

aSERG
APPLIED SOFTWARE ENGINEERING RESEARCH GROUP

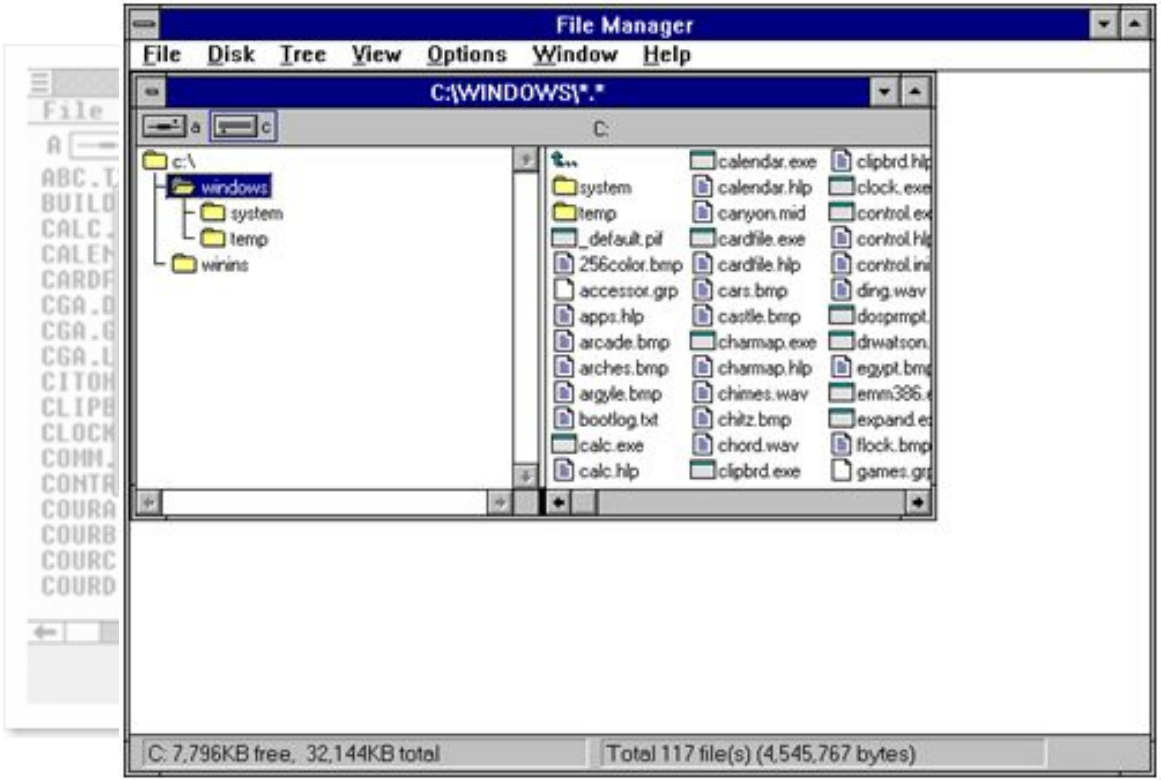
Introduction

- Software teams are increasingly using **data analysis**
  - Informing their engineering and business **decisions**
  - **Building** data solutions

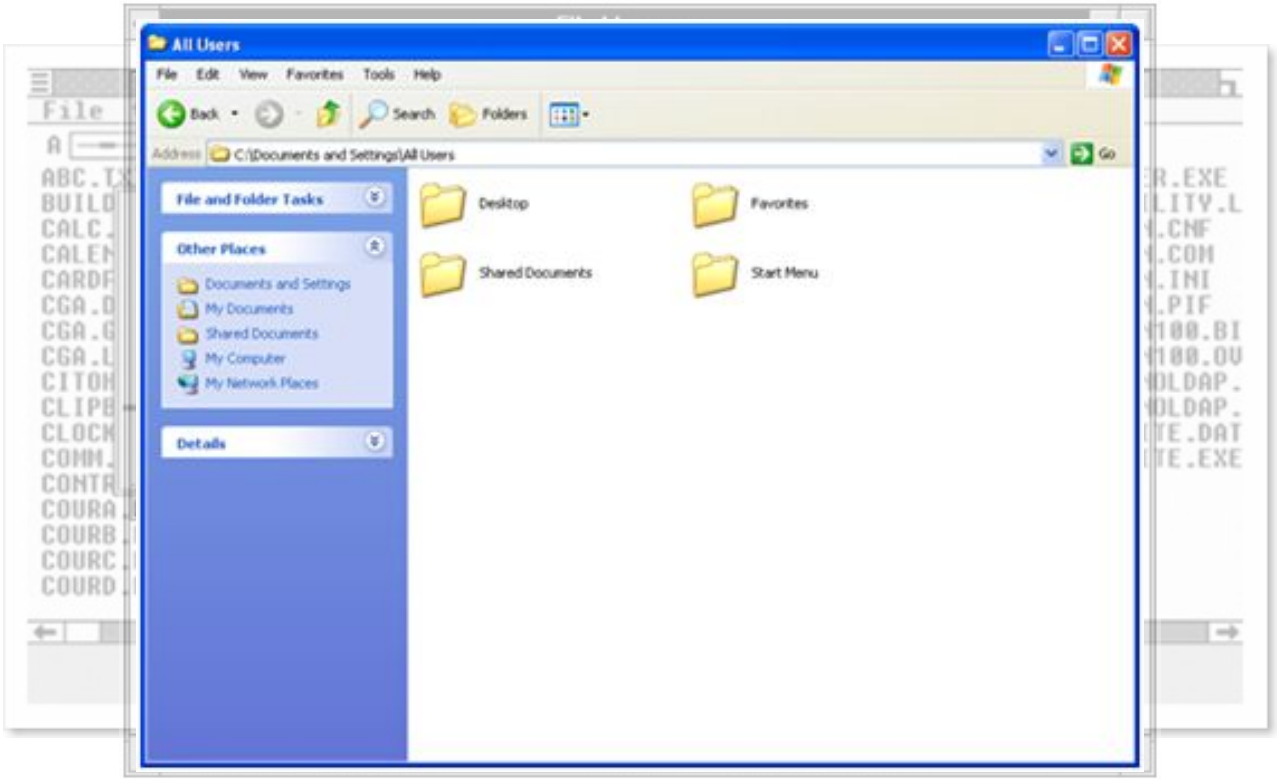
- The role of ***data scientist*** is becoming standard on development teams, alongside existing roles like **developers**, **testers**, and **program managers**

- Teams also need effective ways to **operationalize** data analytics

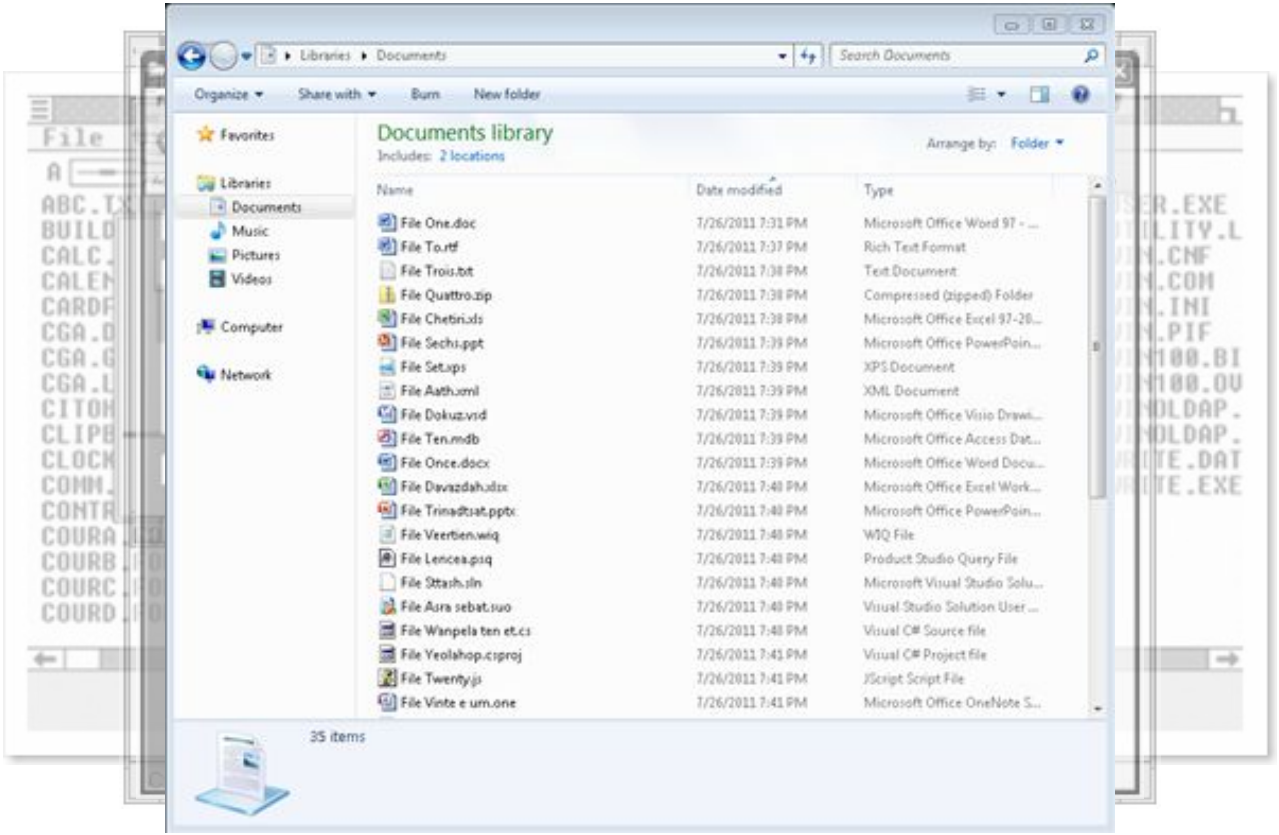
MS-DOS Executive - Windows 1.0



File Manager - Windows 3.1

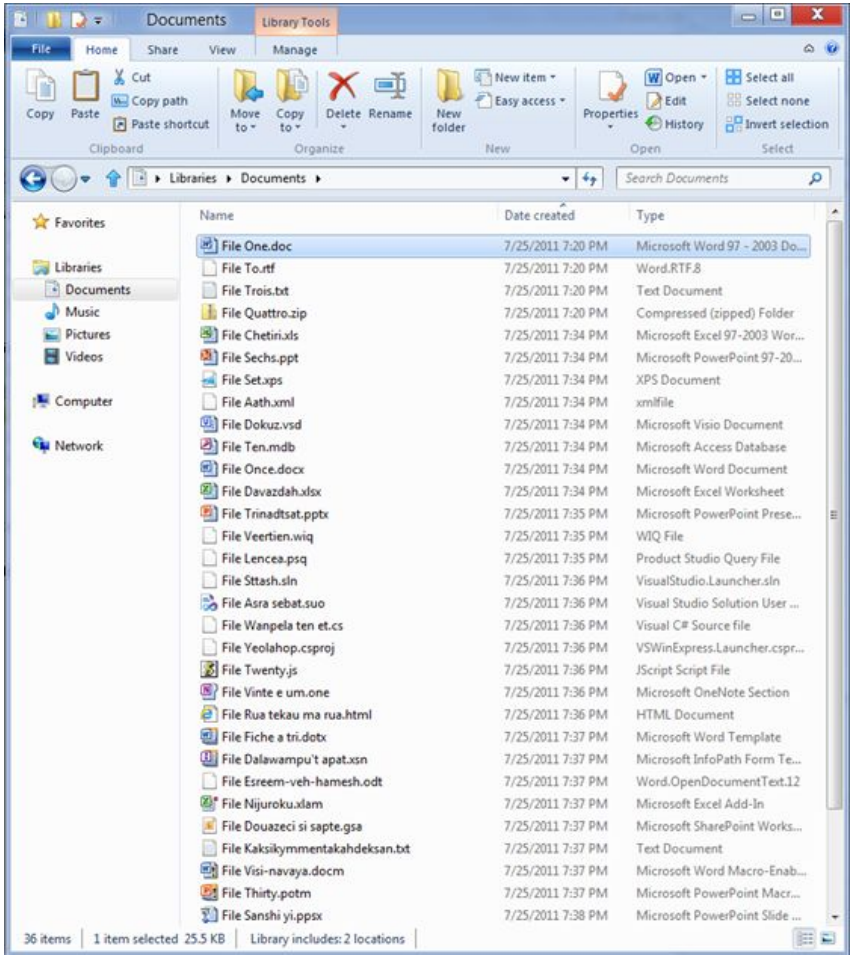


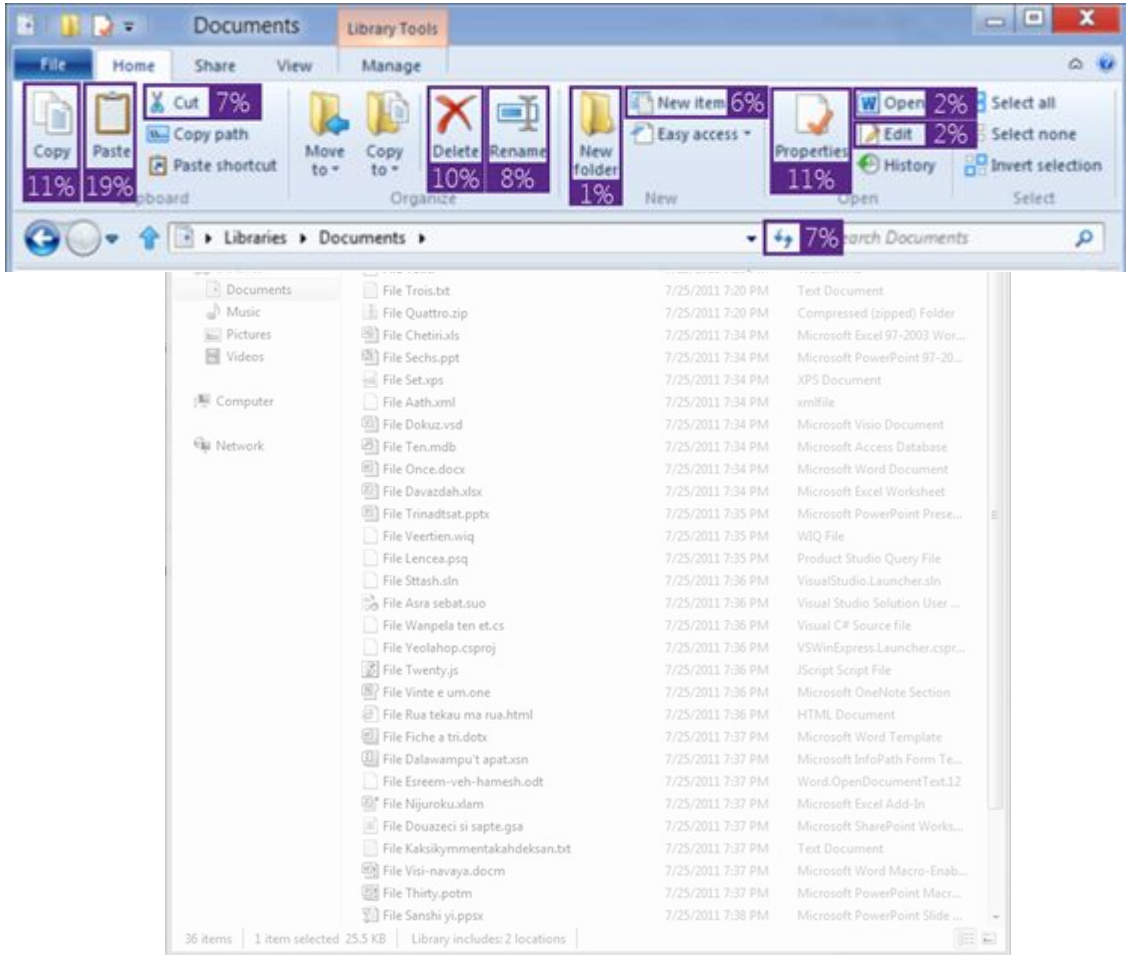
Explorer - Windows XP



Explorer - Windows 7

Top 10 commands are 81.8% of Explorer Command Use!


Documents
Library Tools
File
Home
Share
View
Manage
Copy
Paste
Cut
Copy path
Paste shortcut
Move to
Copy to
Delete
Rename
New folder
New item
Easy access
Properties
Open
Edit
History
Select all
Select none
Invert selection
Clipboard
Organize
New
Open
Select
Libraries ▸ Documents ▸
Search Documents
Favorites
Libraries
Documents
Music
Pictures
Videos
Computer
Network
Name
Date created
Type
File One.doc 7/25/2011 7:20 PM Microsoft Word 97 - 2003 Do...
File To.rtf 7/25/2011 7:20 PM Word.RTF.8
File Trois.txt 7/25/2011 7:20 PM Text Document
File Quattro.zip 7/25/2011 7:20 PM Compressed (zipped) Folder
File Chetiri.xls 7/25/2011 7:34 PM Microsoft Excel 97-2003 Wor...
File Sechs.ppt 7/25/2011 7:34 PM Microsoft PowerPoint 97-20...
File Set.xps 7/25/2011 7:34 PM XPS Document
File Aath.xml 7/25/2011 7:34 PM xmlfile
File Dokuz.vsd 7/25/2011 7:34 PM Microsoft Visio Document
File Ten.mdb 7/25/2011 7:34 PM Microsoft Access Database
File Once.docx 7/25/2011 7:34 PM Microsoft Word Document
File Davazdah.xlsx 7/25/2011 7:34 PM Microsoft Excel Worksheet
File Trinadtsat.pptx 7/25/2011 7:35 PM Microsoft PowerPoint Prese...
File Veertien.wiq 7/25/2011 7:35 PM WIQ File
File Lencea.psq 7/25/2011 7:35 PM Product Studio Query File
File Sttash.sln 7/25/2011 7:36 PM VisualStudio.Launcher.sln
File Asra sebat.suo 7/25/2011 7:36 PM Visual Studio Solution User ...
File Wanpela ten et.cs 7/25/2011 7:36 PM Visual C# Source file
File Yeolahop.csproj 7/25/2011 7:36 PM VSWinExpress.Launcher.cspr...
File Twenty.js 7/25/2011 7:36 PM JScript Script File
File Vinte e um.one 7/25/2011 7:36 PM Microsoft OneNote Section
File Rua tekau ma rua.html 7/25/2011 7:36 PM HTML Document
File Fiche a tri.dotx 7/25/2011 7:37 PM Microsoft Word Template
File Dalawampu't apat.xsn 7/25/2011 7:37 PM Microsoft InfoPath Form Te...
File Esreem-veh-hamesh.odt 7/25/2011 7:37 PM Word.OpenDocumentText.12
File Nijuroku.xlam 7/25/2011 7:37 PM Microsoft Excel Add-In
File Douazeci si sapte.gsa 7/25/2011 7:37 PM Microsoft SharePoint Works...
File Kaksikymmentakahdeksan.txt 7/25/2011 7:37 PM Text Document
File Visi-navaya.docm 7/25/2011 7:37 PM Microsoft Word Macro-Enab...
File Thirty.potm 7/25/2011 7:37 PM Microsoft PowerPoint Macr...
File Sanshi yi.ppsx 7/25/2011 7:38 PM Microsoft PowerPoint Slide ...
36 items
1 item selected 25.5 KB
Library includes: 2 locations


Documents
Library Tools
File
Home
Share
View
Manage
Copy 11%
Paste 19%
Cut 7%
Copy path
Paste shortcut
Move to
Copy to
Delete 10%
Rename 8%
New folder 1%
New item 6%
Easy access
Properties 11%
Open 2%
Edit 2%
History
Select all
Select none
Invert selection
Organize
New
Open
Select
Libraries ▸ Documents ▸
7%
Documents
Music
Pictures
Videos
Computer
Network
File Trois.txt 7/25/2011 7:20 PM Text Document
File Quattro.zip 7/25/2011 7:20 PM Compressed (zipped) Folder
File Chetiri.xls 7/25/2011 7:34 PM Microsoft Excel 97-2003 Wor...
File Sechs.ppt 7/25/2011 7:34 PM Microsoft PowerPoint 97-20...
File Set.xps 7/25/2011 7:34 PM XPS Document
File Aath.xml 7/25/2011 7:34 PM xmlfile
File Dokuz.vsd 7/25/2011 7:34 PM Microsoft Visio Document
File Ten.mdb 7/25/2011 7:34 PM Microsoft Access Database
File Once.docx 7/25/2011 7:34 PM Microsoft Word Document
File Davazdah.xlsx 7/25/2011 7:34 PM Microsoft Excel Worksheet
File Trinadtsat.pptx 7/25/2011 7:35 PM Microsoft PowerPoint Prese...
File Veertien.wiq 7/25/2011 7:35 PM WIQ File
File Lencea.psq 7/25/2011 7:35 PM Product Studio Query File
File Sttash.sln 7/25/2011 7:36 PM VisualStudio.Launcher.sln
File Asra sebat.suo 7/25/2011 7:36 PM Visual Studio Solution User ...
File Wanpela ten et.cs 7/25/2011 7:36 PM Visual C# Source file
File Yeolahop.csproj 7/25/2011 7:36 PM VSWinExpress.Launcher.cspr...
File Twenty.js 7/25/2011 7:36 PM JScript Script File
File Vinte e um.one 7/25/2011 7:36 PM Microsoft OneNote Section
File Rua tekau ma rua.html 7/25/2011 7:36 PM HTML Document
File Fiche a tri.dotx 7/25/2011 7:37 PM Microsoft Word Template
File Dalawampu't apat.xsn 7/25/2011 7:37 PM Microsoft InfoPath Form Te...
File Esreem-veh-hamesh.odt 7/25/2011 7:37 PM Word.OpenDocumentText.12
File Nijuroku.xlam 7/25/2011 7:37 PM Microsoft Excel Add-In
File Douazeci si sapte.gsa 7/25/2011 7:37 PM Microsoft SharePoint Works...
File Kaksikymmentakahdeksan.txt 7/25/2011 7:37 PM Text Document
File Visi-navaya.docm 7/25/2011 7:37 PM Microsoft Word Macro-Enab...
File Thirty.potm 7/25/2011 7:37 PM Microsoft PowerPoint Macr...
File Sanshi yi.ppsx 7/25/2011 7:38 PM Microsoft PowerPoint Slide ...
36 items
1 item selected 25.5 KB
Library includes: 2 locations

What is the emerging role that data scientists play within software development teams?

***Q1*: Why are data scientists needed in software development teams and what competencies are important?**

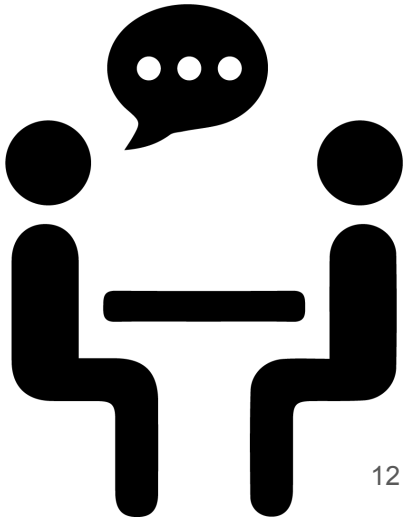
***Q2*: What are the educational and training backgrounds of data scientists in software development teams?**

***Q3*: What kinds of problems and activities do data scientists work on in software development teams?**

***Q4*: What are the working styles of data scientists in software development teams?**

Methodology (1/2)

- One-hour, semi-structured **interviews**
- Conducted by **two** authors
  - One interviews and the second takes notes
- Interviews **audio-taped** and later **transcribed** for analysis



Methodology (2/2)

- Participants selections (*snowball sampling*):
  - Identified **presenters** and **technical leaders** responsible for sharing best practices
  - Additional data scientists by **word-of-mouth**

- Data Analysis:
  - Atlas.TI **qualitative** coding tool
  - Identification of **themes** and **codes**

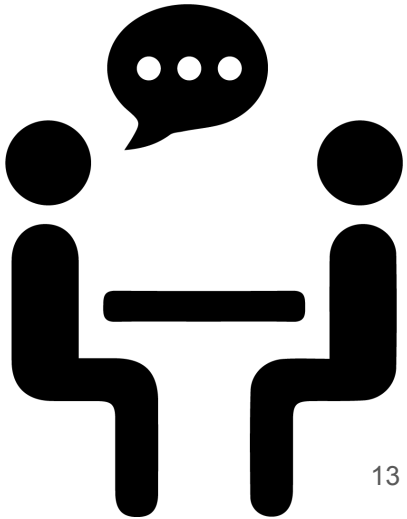


TABLE 1. PARTICIPANT INFORMATION

	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

16 Participants (5 women, 11 men)

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

8 Different Organizations at Microsoft
Different Job Titles

Why are Data Scientists Needed in Software Development Teams?

- Demand for **Experimentation**
  - Designing experiments with real user data (tests)

- Demand for **Statistical Rigor**
  - Hypothesis testing, Confidence intervals, and Baselines through normalization

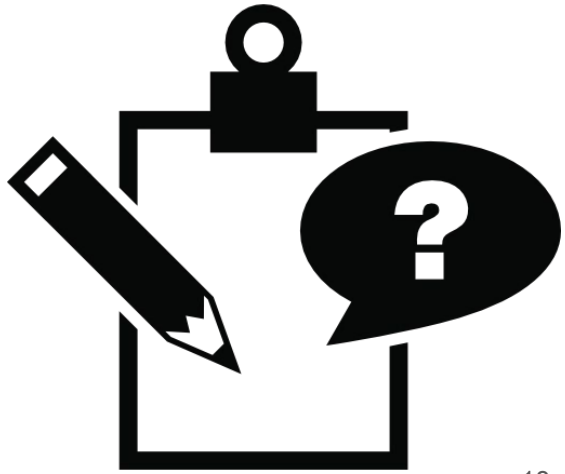
- Demand for **Data Collection** Rigor
  - Shaping and cleaning data

Background of Data Scientists

- **Higher** education degree
- Strong **passion for data**
- Capacity for identifying and refining **questions**

Problems that Data Scientists Work on

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



Activities of Data Scientists

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	■	■	■	■	■				■	■	■	■	■	□	■	■

Activities of Data Scientists

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	■	■	■	■	■				■	■	■	■	■	□	■	■

Activities of Data Scientists

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	■	■	■	■	■				■	■	■	■	■	□	■	■

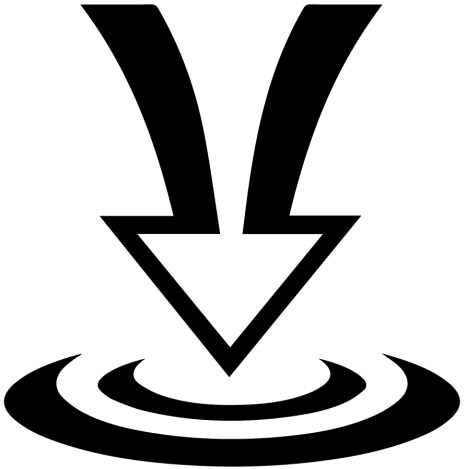
Activities of Data Scientists

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

- New Features
  - Repetitive sequences of **user action**
- Deprecate unused features
  - Large amount of code of features that **nobody used**
- Defect prediction
  - **Releases earlier** than expected

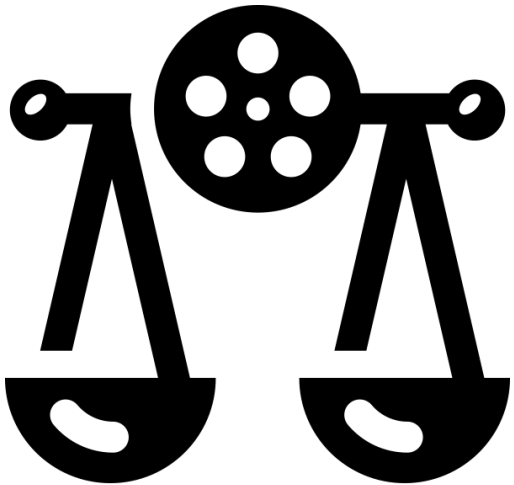


Data Scientists Working Styles

- Insight Providers
  - Generate **insights** to support and guide managers decisions
- Modeling Specialists
  - Build **predictive models** (new features or make decisions)
- Platform Builders
  - Build **data** engineering **platform**
- Polymaths
  - “Do it all”
- Team Leaders

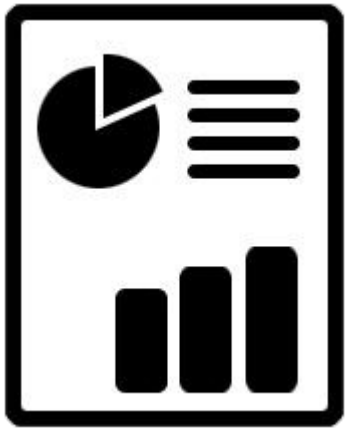
Implications

- **Research**
  - New data available, new role in development teams...
- **Practice**
  - Monetization, new job positions...
- **Education**
  - New skills to be taught



Conclusion

- Data Scientists **Characterization**
  - Activities, problems, roles, background, working style...
- Redesigning Experiments
  - Real user data and report results
- **Future Work**: large scale survey



The Emerging Role of Data Scientists on Software Development Teams

Miryung Kim, Thomas Zimmermann, Robert DeLine, Andrew Begel

UFMG

ICSE 2016

aSERG
APPLIED SOFTWARE ENGINEERING RESEARCH GROUP