Appendix to
The Emerging Role of Data Scientists on Software Development Teams

February 12, 2016
Technical Report
MSR-TR-2016-04

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Appendix to the Emerging Role of Data Scientists on Software Development Teams

Miryung Kim	Thomas Zimmermann	Robert DeLine	Andrew Begel
UCLA, USA	Microsoft Research, Redmond, WA, USA
miryung@cs.ucla.edu	{tzimmer, rdeline, andrew.begel}@microsoft.com

Hierarchical taxonomy of codes corresponding to ICSE 2016 paper

- Why are Data Scientists Needed in Software Development Teams? (Section 4.1)
    - Demand for experimentation
        - ABtesting
        - Flighting
        - Experimentation
        - design of experiments
        - need of experimentation with real user data
        - testing is diminishing
    - Demand for statistical rigor
        - hypothesis testing
        - confidence interval
        - normalization
        - sampling
        - statistical rigor
    - Demand for data collection rigor
        - data quality matters
        - data collection scalability
        - factors to consider in data collection
        - big data characteristics
        - quality matters

- Background of data scientists (Section 4.2)
    - Skill set / skill set
        - A quantitative reasoning, B data manipulation, C users of machine learning, D hacking skill for integration
        - communication skill
    - Data science is a research
        - DS is research
        - a science project
        - analogy to research
        - no given questions. how can I be useful?
    - Data science is a logical role, not a physical title

- Problems that data scientists work on (Section 4.3)
    - anomaly detection in time series data of on-line services

  - automated bug filing
  - antipiracy
  - bug prioritization
  - bug reproduction - how do we reproduce this error?
  - fault localization
  - feature creation
  - feature deprecation
  - prioritization- which bugs should I fix first?
  - quality estimation
  - question development process
  - regression
  - resource rebalancing
  - user engagement

- Activities of data scientists (Section 4.4)
  - Collection
    - Infrastructure for data pipeline / platform
    - Agile data collection
    - data democratization
  - Analysis
    - data cleaning
      - incomplete data
    - data collection
    - data shaping
  - Use and Dissemination
    - brown bag lunch
    - weekly data meet up
    - on-going communication channel

- Impact (Section 4.5)
  - Actionability
    - actionality matters
    - define actions and triggers
    - prove for a while
    - prove now
  - Data Science Impact Examples
    - bing tool bar being deprecated
    - data ruled

- Organization of Data Science Teams (Section 4.6)
  - Triangle model
    - 3 parts: data platform, flighting, mini data science
    - triangle-scientist, data steward, telemetry
  - Hub and Spoke
    - hub and spoke
  - Consulting
    - collaboration with a feature team
  - Individual Contributor

- Virtual Team
  - virtual team

- Data Scientist Working Styles (Section 5)
  - Insight providers (Section 5.1)
    - high level direction coming from managers
    - lack of understanding stats
    - justification and confirmation
    - influence data collection
    - insight- predictability of crashes in Windows
    - insights- update annoyed customers
    - talk simple
  - Modeling specialists (Section 5.2)
    - Operationalization
    - cost benefit analysis
    - define ground truth with feature teams
    - integration and operationalization
    - limitations of ML alones
    - precision and recall do not matter
    - translate results into business values
  - Platform builders (Section 5.3)
    - data engineering
    - health in the market report
    - knowledge engine
      - need of knowledge engine
    - need of central data engineering infrastructure
    - scalability
    - scalability, storage cost
    - sensible metric
    - tension between collection and analysis
  - Polymaths (Section 5.4)
    - does everything
  - Team Leaders (Section 5.5)
    - culture
      - culture-advocate
    - data advocate
    - data PM
    - dumb down to 7th grade
    - explain
    - questions
    - simplicity
    - people should be accountable for bad decisions

Implications and Guidelines (Section 6)

- human in the loop
    - human in the loop in defining scenarios to focus on
- involve data scientist early
    - involve data scientist early before human capital
- Iteration and refinement
- Value of insights
    - insights - knowledge pyramid, data, information, insights, actions
- How to choose questions: three conditions - priority, actionable, and commitment

Other Miscellaneous Codes

3 features of StreamML- storage, visualization, operationalization
agile approach
currently there are about 400-500 data scientists at MS
data science characterstics
domain knowledge
easy integration
fork instrumentation
heterogenous data platform in the past
history-DS existed before
o the last thing that I did wa..
piracy prevention
planning
proof of the concept, flighted, production
quality
reusable analysis
rise of big data
role- design of knowledge engine
role- mixed role of data engineering infrastructure and data science
tools
tuesday security update
updating existing instrumentation code is hard
validation
value of data scientist
vignetts scenarios
we need to connect all these d..
weighting

Selected quotes from the Interview Study

The quotes are organized by the taxonomy described in the beginning of this technical report.

- **Why are Data Scientists Needed in Software Development Teams? (Section 4.1)**
  - **Demand for experimentation**
    - **ABtesting**

P1
Codes: [ABtesting]

it's this idea that sort of tries to borrow the idea from a lot of the current industry practices around the service industry about quickly sort of saying stuff out, trying it, get results back from it and then implementing changes if necessary, but certainly reverting it if that fails. So, the part of it that I am primarily interested in is the experimentation part of that particular system.

P5
Codes: [ABtesting] [flighting]

You create an environment where, for example in search, where I can actually experiment based on a mockup, if you will, of the idea. I can actually come up with a set of ideas, broad ideas about my work, and I can actually deploy them in some easy way.

- **Flighting**

P5
Codes: [ABtesting] [flighting]

You create an environment where, for example in search, where I can actually experiment based on a mockup, if you will, of the idea. I can actually come up with a set of ideas, broad ideas about my work, and I can actually deploy them in some easy way.

- **Experimentation**

P16
Codes: [experimentation] [proof of the concept, flighted, production]

We create prototypes and proof-of-concept, so we show that if we do things the way we propose there is going to be lifting some credible metric or, you know, things are going to be better.

But then putting it into production is a separate process, and it usually doesn't go straight into production. Even if we prove offline that the prototype works well and it does what it was intended to do there is still a process of controlled experiments. So if – so the future team builds that into the production system, but then it doesn't get what it needs to all the users, it gets, as we call it in-bank flighted, which means that it's going to a controlled experiment, and there is a small group that is being exposed to the new way of doing things and then there is a control group and we are comparing whether users, in fact, like the new version. And then it gets released to everybody.

design of experiments

P3
Codes: [design of experiments]

Well, is it because people don't like using it or is it because they don't know it's there?" So we create a game that gets people to use it and repetitively use it so that they get proficient at it, and then we watch what happens when we take the game away. Did it stick or did it not stick? If it doesn't stick, then chances are, they just don't like that feature, which is possibly what we have here. So, "Should we deprecate that feature?" becomes the question then. We certainly should not do any improvements on that feature. Maybe leave it there because there's probably some company out there that's put it into their process and removing it will make them angry, but we certainly won't do any more work on it

- **need of experimentation with real user data**

P10
Codes: [need of experimentation with real user data] [testing is diminishing]

instead of having an army of testers to go off and generate a bunch of data about [application], that data's already here, and it's actually more authentic because it's real customers who are generating it, on real machines, real networks. You no longer have to simulate and anticipate what you think the customer's gonna do; you actually have data that tells you exactly what the customer's

- **testing is diminishing**

P10
Codes: [need of experimentation with real user data] [testing is diminishing]

instead of having an army of testers to go off and generate a bunch of data about [application], that data's already here, and it's actually more authentic because it's real customers who are generating it, on real machines, real networks. You no longer have to simulate and anticipate what you think the customer's gonna do; you actually have data that tells you exactly what the customer's

- **Demand for statistical rigor**
  - **hypothesis testing**

P3
Codes: [hypothesis testing] [statistical rigor]

The statistics is – that's sort of a very small percentage of it and very few people actually do hypothesis testing, formal. There's a few of us, but not very many. But I tend to be a stickler on that point. It is valid. When I do my analyses, I always have my null and alternative hypothesis just because that was how I was trained.

- **confidence interval**

P2
Codes: [confidence interval] [statistical rigor]

How confident are you that this is the number you will get?" And I'm looking and go, "What do you mean? It's 95 percent.

P2
Codes: [confidence interval] [statistical rigor]

And he goes, "Wow, this is the first time I'm getting this answer," because every time I would ask him – I would ask. Some people kinda reach for the guns and say 75 percent, 80 percent, something like this. And then this is the first time that he saw this incorporated into the process and actually providing me the real number with confidence level and the statistics behind that and science to support.

- **Normalization**

P3
Codes: [normalization] [statistical rigor]

"If we were to force everybody to upgrade to at least this version, it will reduce the number of crashes that we see by X percent." And that's just an easy sort of normalization equation ________. So it's a fairly easy, simple thing. It's not complex data science. It's just a simple thing.

- **Sampling**

P11
Codes: [sampling] [weighting]

I put more weight on the heavy beta users and less weight on the light beta users so that I have a beta profile that's matching the RTM profile.

- **statistical rigor**

P3
Codes: [statistical rigor]

We got a lot of people who know how to take the mean and standard – well, not even standard deviation *[laughs]*, but there's a lot of people who know how to produce the mean of a set of numbers, but that's – I don't really call that data science *[Laughs]*.

P3
Codes: [hypothesis testing] [statistical rigor]

The statistics is – that's sort of a very small percentage of it and very few people actually do hypothesis testing, formal. There's a few of us, but not very many. But I tend to be a stickler on that point. It is valid. When I do my analyses, I always have my null and alternative hypothesis just because that was how I was trained.

- **Demand for data collection rigor**
  - **data quality matters**

P4
Codes: [data quality matters] [limitations of ML alones]

And then once we have the more data flowing nicely and neatly, yeah you should machine learning, machine classification what you want on that one. It is nice to do. We try it. We took raw screen data. We have like what, X screen data points. Any machine learning you do it says too much noise.

- **data collection scalability**

P8
Codes: [data collection scalability] [tension between collection and analysis]

how can we effortlessly and automatically scale to meet the growing demands of our business?

- **factors to consider in data collection**

P1
Codes: [factors to consider in data collection]

What about storage, what about speed. There is like, what about legal, what about privacy. There is like a whole entire gamit of things that you need to like jump through all those hoops in order to collect the instrumentation.

- **big data characteristics**

P2
Codes: [big data characteristcs]

There’s the volume. There’s the variety and the velocity.

- **quality matters**

P11
Codes: [data cleaning] [quality matters]

We are decoding to regional format, and then we need to cleanse the data, because there are all sorts of data quality issues, imperfect instrumentation and also some weirdness in the ecosystem because there’s all sorts of machines, and then you’re bound to have some weirdness in there, right

- **Background of data scientists (Section 4.2)**
  - **Skill set / skill set**

P15
Codes: [skill set]

So the typical guys on my team have, right now, pretty much all of them, some PhD in a quantitative field with some machine learning in background and, also importantly, the ability to code, and some domain expertise

- **A quantitative reasoning, B data manipulation, C users of machine learning, D hacking skill for integration**

P15
Codes: [A quantitative reasoning, B data manipulation, C users of machine learning, D hacking skill for integration] [skill set]

but the PhD is not always directly in machine learning.   I say a quantitative field.

- **communication skill**

P15
Codes: [communication skill] [skill set]

it's good to have communication ability, because you're often meeting with business owners and communicating results

- **Data science is a research**
  - **DS is research**

P11
Codes: [DS is research]

it was kind of a science project, and it got out of hand and then became something that got into the product.   And then, there was also really cool stuff that came out that particular project –

P13
Codes: [DS is research] [identifying questions]

Probably that part of your PhD when you're figuring out what is the most important question.   Yeah, that probably is very relevant.   And I like that part.

- **analogy to research**

P5
Codes: [analogy to research]

Doing data science is kind of like doing research

- **no given questions. how can I be useful?**

P13
Codes: [no given questions. how can I be useful?]

It has never ever been in my four years here I don't think it's ever been the case that somebody has come to me and said hey, can you answer this question. I mostly sit around thinking how can I be helpful? And I think like this is the most important thing I can do on this project.

- **Data science is a logical role, not a physical title**

P5
Codes: [Data science is a logical role, not a physical title]

So up to now the people that we've had in the company over the last handful – two, three, four years that have been data science type people have usually been grouped with PMs. That is their test, depending on where they happen to be. So you had some people that were classified as devs or PMs or test, and yet they did analytics for a living.

- **Problems that data scientists work on (Section 4.3)**
  - **anomaly detection in time series data of on-line services**

P7
Codes: [anomaly detection in time series data of on-line services] [questions]

We’re interested in anomaly detection on time servers in general.

- **automated bug filing**

P11
Codes: [actionability] [automated bug filing] [define actions and triggers]

the notion of actionability is we will not stop until we find these bugs, yeah, and then we’re headed off to the feature teams, because the bugs don’t lie, right.

- **antipiracy**

P2
Codes: [antipiracy] [questions]

And help the business actually – antipiracy initiative, support the antipiracy initiative at the same time without impacting negatively the console _______.

- **bug prioritization**

P11
Codes: [bug prioritization] [questions]

So, then our goal is to basically say, “Hey, these are all the bugs that contribute to these two crashes a day, and if you want to reduce to one crash a day, that means you have to reduce the volume by 50 percent, so let’s start from the head of the curve

- **bug reproduction - how do we reproduce this error?**

P6
Codes: [bug reprodduce - how do we reproduce this error?] [prioritization- which bugs should I fix first?] [questions]

"Oh, cool. Now we know what the bugs we should fix first." So this enabled them to prioritize which bugs were more important, at least to some extent. There's a lot of things that are disconnected, so you can't really say that's the best bug to fix. Okay? But then the next question then is, okay, how do we reproduce this error?

- **fault localization**

P2
Codes: [fault localization] [questions]

So the business impact was that you were able to actually detect defect or prevent this defective update

P3
Codes: [fault localization] [questions]

what areas of the product are failing? And why?" That type of thing.

- **feature creation**

P3
Codes: [feature creation] [impact]

– if you see the repetitive pattern where people don't recognize –

Interviewee: That the feature is there.

- **feature deprecation**

P3
Codes: [feature deprecation] [impact]

It costs about X lines of code to make that happen. Nobody uses it. So here we had this whizz-bang feature that – and a significant amount of money went into both creating it and testing it and nobody knows it's even there is what our – what we seem to think.

P5
Codes: [feature deprecation] [impact]

And so we stopped deployment, which was a big deal because all the energy and inertia was to go full steam with this thing.

P13
Codes: [feature deprecation]

The cutting of the feature was a big deal and that was actually a good success story.

- **prioritization- which bugs should I fix first?**

P6
Codes: [bug reproduce - how do we reproduce this error?] [prioritization- which bugs should I fix first?] [questions]

"Oh, cool. Now we know what the bugs we should fix first." So this enabled them to prioritize which bugs were more important, at least to some extent. There's a lot of things that are disconnected, so you can't really say that's the best bug to fix. Okay? But then the next question then is, okay, how do we reproduce this error?

- **quality estimation**

P8
Codes: [planning] [quality estimation] [questions]

“Is [application] ready to ship?” And for making a beta they say, “Is the beta ready to ship? Is the self-host version ready to ship?

- **question development process**

P3
Codes: [high level direciton coming from managers] [question development process]

So yeah, it could come from – mostly it's gonna come from management, it's gonna come from the manager wants to know something about it. And often it starts as, I don't know, sort of a conjecture. I don't know if I'd call it conjecture. Rumor. So there's a rumor that our quality is getting worse over time, people have this feeling that it is. "Okay, let's not depend on feeling. Let's go and actually measure and find it."

- **Regression**

P2

Codes: [questions] [regression]

It should be as good as before. It should not deteriorate any performance, customer ________ experience that they have

- **resource rebalancing**

P10
Codes: [impact] [resource rebalancing]

When the leadership saw this gap, here, the allocation of developers towards new features versus stabilization shifted away from features toward stabilization. Not 100 percent, but we shifted some of the focus toward stabilization, to get this number back.

- **user engagement**

P16
Codes: [user engagement]

looking at their past history and which stories they tend to click on and that was, actually, a combination of using client history and user history and the clicks that similar users made in the last few hours on stories that are current now. You predict what the user may be interested in.

- **Activities of data scientists (Section 4.4)**
  - **Collection**
    - **Infrastructure for data pipeline / platform**

P14
Codes: [an infrastructure for data pipeline / platform]

we would really rather have a single technology pipeline that allows for you to describe data, capture and instrument data, send that data to the service for collection. And then to manage that data through a pipeline that allows for you to make sense of it, including making sure that it's high quality, dealing with data loss and data issues, data quality issues, data suppression, where you don't want people to have access to data if it's sensitive data.

- **Agile data collection**

P14
Codes: [agile data collection] [iteration and refinement]

Until you actually have a thousand devices or a million devices out in the wild of people using them, and you're now seeing the data flowing out those devices, you didn't realize that you had that bug. You didn't realize you needed that other data point. You didn't realize. So until you actually start writing queries that look at the real data, you're not actually gonna have the best opinion of what that data should be. The faster you get people to having any starting point of that data is the amazing power.

- **data democratization**

P14
Codes: [data democratization]

I had to use a phrase that everybody uses, but democratization of data, it really allows for people whose job is I'm a data scientist. I am a person who loves and plays with data. I care passionate about data. If they can't understand or discover or find or make sense of data, it almost begs the question of why should we even bother logging if we can't then use it. Such is one area where between our two teams, trying to make sure that that's fundamental.

- **Analysis**
  - **data cleaning**

P11
Codes: [data cleaning] [quality matters]

We are decoding to regional format, and then we need to cleanse the data, because there are all sorts of data quality issues, imperfect instrumentation and also some weirdness in the ecosystem because there's all sorts of machines, and then you're bound to have some weirdness in there, right

- **incomplete data**

P6
Codes: [incomplete data] [influence data collection]

Problem was that we ran into, then, is two things. One is the instrumentation of the usage database was incomplete.

- **data collection**

P8
Codes: [data collection] [fork instrumentation]

A great example I can give here is referring to the latency requirement, the 30 milliseconds requirements. That's not a real requirement. There was a group however that had a 15-minute requirement to get information back, and at the time [service] was running about—I don't know, let's say 30, 75 minutes or so latency. So they said, "Hmm, that doesn't look like that'll meet our needs. We're just gonna make our own."

- **data shaping**

P11
Codes: [data shaping]

there’s also the data cooking, shaping, right, so you essentially want to transform the data from the instrumentation view into analysis view,

P13
Codes: [data shaping]

A lot of what I do is helping the data to become the features. Identify the features and then doing the data processing 'cause the data is really dirty a lot of times there's missing values

- **Use and Dissemination**
  - **brown bag lunch**

P13
Codes: [brown bag lunch] [on-going communication channel]

I talk to my manager but I always make sure to inform the broader team they don't necessarily need to understand the details of how I got there but I think it's jarring when somebody says oh, stop working on this without explaining why. And so I would always to make sure to do like a brown bag or something saying like here's what I tested, here's how I tested it, here's the results.

- **weekly data meet up**

P3
Codes: [culture-advocate] [DS team structure] [weekly data meet up]

So he takes it to a group, basically, with his boss and his peers and they go through the data. So [name] is my boss' boss and he runs sort of the data meet-up weekly with a whole group of people. So [another name] can definitely give you all sorts of insights as to what goes on in that meeting. And that's – that'll be kind of interesting insights as to how they use or don't use it, as the case may be, which is still, unfortunately, true to an extent greater than I'd like to see

- **Impact (Section 4.5)**
  - **Actionability**

P9
Codes: [actionability] [define actions and triggers] [impact]

They’re more interested in impact. So, if you come and say, “I have this anomaly detection that finds these anomalies,” they’re like, “So what. What’s the impact that has had?” It’s not even about precision and ___________ your algorithm. It’s about, “Hey. How did this make the lives of our end users better? How did this help us understand our tenants better? What did this do for this service?” Right? That’s kinds of the things that they care about.

P10
Codes: [impact] [resource rebalancing]

When the leadership saw this gap, here, the allocation of developers towards new features versus stabilization shifted away from features toward stabilization.   Not 100 percent, but we shifted some of the focus toward stabilization, to get this number back.

- **actionality matters**

Code: actionality matters {1-0}

P11
Codes: [actionability] [actionality matters]

They're not used, so actionability is actually a big thing, but if it's not actionable, the engineers then look at you, say, "I don't know what to do with this, so don't even bother me,"

- **define actions and triggers**

P4
Codes: [actionability] [define actions and triggers]

We found in the source code who touch last time this function.   He gets the bug. And we've done one more thing.   To compute the impact, we associated number of minutes lost in my life because of that trash and because we track number of minutes active on the machine, we find out which bucket is responsible for the most minutes of lost opportunity in the wall.   Clever, right?

- **prove for a while**

P10
Codes: [prove for a while]

And so it takes a couple releases to demonstrate, like, "No, look, this is accurate."   And then as people start to see – "Okay, wow, that does look concerning; I have no idea how you did that, but I know that when you did it last time, you were right, so I'm gonna have more confidence that this is right, and I'm – looks to be like we're off, so maybe we should do something."

- **prove now**

P9
Codes: [prove now]

Yeah.   That's the whole thing of our pipeline.   We adjust all of this data as it's coming from servers.   Within 5 minutes, I can go and analyze it and make all kinds of decisions on it.

○ **Data Science Impact Examples**

P5

Codes: [feature deprecation] [impact]

And so we stopped deployment, which was a big deal because all the energy and inertia was to go full steam with this thing.

P3
Codes: [feature deprecation] [impact]

It costs about X million lines of code to make that happen. Nobody uses it. So here we had this whizz-bang feature that – and a significant amount of money went into both creating it and testing it and nobody knows it's even there is what our – what we seem to think.

P2
Codes: [confidence interval] [impact] [statistical rigor]

"We're bored. Nothing happens." So we planned for the waves rolling out. We ended up rolling it to the whole world two weeks earlier than planned, 'cause it was just smooth. Why? Because we worked with confidence intervals, because we didn't say what is the expected value. We said, "What's the upper 95 confidence interval bound of the expected value?" so the actual was lower than we said. That was the worst case we predicted, and everything was just very –

        - **data ruled**

P5
Codes: [data ruled] [feature deprecation] [impact]

So that's one case where the insight was counter to the expectations and the overall culture and approach, and yet it persevered because at the end of the day the data ruled and people stopped something that they thought they should continue doing, and they found out that indeed they should have stopped it.

- **Organization of Data Science Teams (Section 4.6)**
    - **Triangle model**
        - **3 parts: data platform, flighting, mini data science**

P14
Codes: [3 parts: data platform, flighting, mini data science] [DS team structure]

Like our team, we actually have like three halves of our brain – three-thirds of our brain; one part that is built in the data platform, so doing all of the engineering work to build this platform. Another part that is all about flighting because we've believe that there is a huge and core fundamental element around data and data-led learning that requires you to experiment, so we have a flighting investment. But then the third part is where we actually have a mini data science team in our own team.

        - **triangle-scientist, data steward, telemetry**

P2

Codes: [DS team structure] [triangle-scientist, data steward, telemetry]

you can see this is one exception. ______ and I really wanna hire, I need to convert that we're in hiring phase. So, in order, this is a statistician with statistical master's degree. This is my physicist. This is my mathematician. These are all people that are analyzing the data. I'm considering one of them, but this is my hardware telemetry statistician. He has his master's in statistics. That's this information, still statistics. So, one, two, three, four, five, six people are doing analysis. Now, this guy here, database, SQL, that's my data steward, if that makes sense.

- **Hub and Spoke**
  - **hub and spoke**

P4
Codes: [DS team structure] [hub and spoke]

So the Hub and Spoke means there is a central team that builds common pieces of platform. They are pretty low level right now. They simply move data from one place to another

- **Consulting**

P12
Codes: [consulting] [DS team structure]

This team works with both internal and external customers in Microsoft and offsite. Within Microsoft we work with other teams in terms of solving their data problems.

P12
Codes: [consulting]

In terms of having a data science team and having work, this team is pretty different from the other teams that I have seen in terms of going around and helping other teams and bringing them to a sample product and making sure that we solve common problems and use the same platform.

- **collaboration with a feature team**

P5
Codes: [collaboration with a feature team]

– to have the data scientists work as part of the feature team that's doing whatever the feature is doing that the data scientist is working on.

- **Individual Contributor**
  - **IC- solo scientist**

P13
Codes: [IC - solo scientist]

As far as I understand there's kind of two different models of data scientist at Microsoft. There's one, a data scientist who works on a team of data scientists and then there's data scientists who are sort of solo sitting on an actual product team and I am the latter. So I am the only scientist on this team.

  - Virtual Team
    - virtual team

P3
Codes: [DS team structure] [virtual team]

Well, it's more of a virtual team. There's data scientists all over the place and we work together.

- Data Scientist Working Styles (Section 5)
  - Insight providers (Section 5.1)
    - high level direction coming from managers

P3
Codes: [high level direciton coming from managers] [question development process]

So yeah, it could come from – mostly it's gonna come from management, it's gonna come from the manager wants to know something about it. And often it starts as, I don't know, sort of a conjecture. I don't know if I'd call it conjecture. Rumor. So there's a rumor that our quality is getting worse over time, people have this feeling that it is. "Okay, let's not depend on feeling. Let's go and actually measure and find it."

    - lack of understanding stats

P3
Codes: [lack of understanding stats] [statistical rigor]

So I think part of the problem is that a lot of the people aren't given training in statistics, so they don't know what questions to ask and they don't necessarily know why some particular piece of information is important. So you got some P value of .01. "Oh, gee, should I be happy or sad? What does that mean?" And they don't necessarily know that. So I have to explain things in terms that might not be forceful enough. Maybe they just don't get it, necessarily. "You got a statistical probability of winning the lottery here, so you should be really, really mindful of this dataset." And they don't necessarily have a feeling for that since data never went really into decisions being made by software people before.

    - justification and confirmation

P2
Codes: [justification and confirmation]

Honestly, to be fair, the decision was made before we were able to do the analysis, the analysis more like backing up their already made decision, but it's okay.

P3
Codes: [justification and confirmation]

So if the data confirm what they already believed, then they'll listen to them. If they do not confirm, then they'll find any excuse in the boo k to not use it

P4
Codes: [justification and confirmation]

So as long as you come with data that kind of confirms the ____ it is great data. The second you start contradicting the ____, then you have to be able to document it or explain pretty well data model.

- **influence data collection**

P6
Codes: [influence data collection]

we need to connect all these databases more efficiently so that you're not spending hours and hours of time crawling through the stack and trying to figure out what it is that's caused a crash.

P8
Codes: [influence data collection]

We explained to him what options we already had aggregated against, and when it turned out that he needed something different we built him the specific aggregation. However, in the long term the model that we like to approach

- **insight- predictability of crashes in Windows**

P5
Codes: [actionability] [insight- predictability of crashes in Windows]

big aha was that what people cared about more than anything else is the unpredictability of an event, not the fact that the event happens.

- **insights- update annoyed customers**

P5
Codes: [impact] [insights- update annoyed customers] [tuesday security update]

So it wasn't that they lost security. What they were really annoyed about was the almost daily reboots required by security updates, so it was the updates that was annoying them, not the lack of security

- **talk simple**

P10
Codes: [dumb down to 7th grade] [talk simple]

I'll just say a super smart data scientist – their understanding and presentation of their findings is usually way over the head of the managers

- **Modeling specialists (Section 5.2)**
  - **Operationalization**

P7
Codes: [integration and operationalization]

The big problem is not to operationalize the algorithms because they're most of the time not that complicated. But getting everything in their infrastructure to run and set up is most – and the climate itself is the big hassle and that's where they're supporting us. Like, writing up the algorithm and implementing it in their infrastructure is not so much the issue. It's really the stuff that surrounds it.

P12
Codes: [integration and operationalization]

what we told them is using their process, what is the difference from themselves and the monitoring? Having a sliding window in the training cycle, we told them, "This is the main key game that you want to get." Then, we told them the parameters for the model that before and after parameter training, that will give you the best results _____________. This was one thing. In terms of the actual work on the model, they were a little bit unsure about doing it themselves.

- **cost benefit analysis**

P12
Codes: [precision and recall do not matter] [translate results into business values]

So, in terms of convincing, what we have seen is that if you just prevent to them all of these numbers like positioning ________ and __________ factors, that is important from the knowledge sharing model transfer perspective, but if you are out there to sell your model or idea, this will not work because the people who will be in the decision-making seat will not be the ones doing the model transfer. So, for those people, what we did is cost benefit analysis where we showed how our model was adding the new _______ of what they already had. So, we showed in terms of dollar amount how the picture would look like if you had this model applied right now.

- **define ground truth with feature teams**

P7
Codes: [define groudtruth with feature teams] [iteration and refinement]

You have communication going back and forth where you will find what you're actually looking for, what is anomalous and what is not anomalous in the set of data that they looked at.

P9
Codes: [define groudtruth with feature teams]

When you're seeing this part of the data, this one's good versus here's setting that ground truth. Here's where you should have alerted. Here's where you shouldn't have done anything. That's something that we are continuing to integrate on, but that's something that was fairly labor-intensive

P9
Codes: [define groudtruth with feature teams]

We've analyzed over thousands of time series to form these. So, they were segregated by different types of errors and then for different types of proponents and protocols. I knew what the incidents were for each of those based on the different data sets.

- **limitations of ML alones**

P4
Codes: [data quality matters] [limitations of ML alones]

And then once we have the more data flowing nicely and neatly, yeah you should machine learning, machine classification what you want on that one. It is nice to do. We try it. We took raw screen data. We have like what, 13,000 screen data points. Any machine learning you do it says too much noise.

- **precision and recall do not matter**

P12
Codes: [precision and recall do not matter] [translate results into business values]

So, in terms of convincing, what we have seen is that if you just prevent to them all of these numbers like positioning ________ and ___________ factors, that is important from the knowledge sharing model transfer perspective, but if you are out there to sell your model or idea, this will not work because the people who will be in the decision-making seat will not be the ones doing the model transfer. So, for those people, what we did is cost benefit analysis where we showed how our model was adding the new ________ of what they already had. So, we showed in terms of dollar amount how the picture would look like if you had this model applied right now.

- **translate results into business values**

P12
Codes: [cost benefit analysis] [translate results into business values]

how much the dollar amount that you gain would look like." So, we assured them that as long as the humans are more than 55% accurate, there's a positive cash flow.

P15
Codes: [translate results into business values]

Unless you're working with a data scientist on the – if the user is – if the consumer is a data scientist who's been working on a similar problem and you want to show like a better way to do it, then you could show technical metrics like ROC curves and so on. But in many cases the user is more like a business user who wants a problem solved, so there's a lot of effort.

- **Platform builders (Section 5.3)**
  - **data engineering**

P4
Codes: [data engineering]

I personally think when you want to do that at scale, you won't, I mean if they have 100 people who do scientific matter, you don't get the 100 X better resolve.

- **knowledge engine**

P4
Codes: [impact] [knowledge engine]

So then what to do is you created the knowledge engine for what questions you want and then you can get the variety of answers. When you can fundamental higher level questions like where should I apply my resources?

- **need of knowledge engine**

P4
Codes: [need of knowledge engine]

– that leads to a knowledge engine and the knowledge engine is continually maintained by the telemetric system. It just feeds it and keeps it accurate and it queried that knowledge engine to get the high level answers you want.

- **need of central data engineering infrastructure**

P6
Codes: [need of central data engineering infrastructure]

where most things are in NCBI, National Center for Bioinformatic Information maybe, or Biological Information, I don't know exact, but the thing was at some point in the last 10 to 15 years, all the data in biology, or at least a huge portion of the data in biology, got centralized in one place. So pretty much all you needed to do was to know just some snippet of the sequence of the DNA of interest, and then you plug that into NCBI and then out comes a whole bunch of other information that you can then know about this particular sequence.

P6
Codes: [need of central data engineering infrastructure]

Just-in-Time, Experimentation, Telemetry and Services, something like that – was they were trying to get the infrastructure to enable the collection of all these different signals from [applications], and any other piece of software out here, collect whatever information is needed through some infrastructural, through some pipeline, upload it into [database] so that then you can actually get, one, large amounts of data and, two, make it easy for others to manipulate that data and then make all those nice conclusions that they have. So coming here, that was all very new, okay, so it was –

- **scalability**

P8
Codes: [scalability] [tension between collection and analysis]

"All right, [name], you've got a good scenario. We care about your scenario. We want to give this to you, but we can't give it to you in the volume that you require. Is it possible for us to work with you about the requirements of the data that you have in order for—by managing retention policy, by managing the number of events that we'll collect from each console, things like that to reduce it down to a size that we can manage?"

- **scalability, storage cost**

P8
Codes: [scalability, storage cost] [tension between collection and analysis]

"We want X. What will it take to give us X?" And the manager replied, "$X million a year."

- **sensible metric**

P11
Codes: [normalization] [sensible metric]

We're able to normalize by ever more sophisticated denominators, and, hopefully, the metrics makes more and more sense.

- **tension between collection and analysis**

P1
Codes: [tension between collection and analysis]

How do you construct a sample, what do you sample in such a way that enables you to do the types of analysis that you want to do.

P2
Codes: [tension between collection and analysis]

“We collected all that data. Just go figure it out.” And I said no. They’re like, “What do you mean?” And I said, “Let’s sit down and understand. What does this data mean?”

- **Polymaths (Section 5.4)**
  - **does everything**

P13
Codes: [IC - solo scientist] [Julie does everything]

Oh yeah. I contribute to it all the time. In fact, remember earlier I said that I use the role of applied scientist to kind of do be useful. How to be useful and I can shape that pretty liberally. Our new project is the whole team, all 10 of us, are working on something called [name]. And so this has been a really fun project for me because I have for the first time been very critical to it. My name shows up in all of the planning like because I am responsible for that core core part. And so I've been involved in all the architecting of it. My manager was just away for five weeks, he just got back last week, but the five weeks that he was gone I kind of just ended up running the project. I ran all the team meetings, I ran the architecture stuff. It was a really cool opportunity for me and the stuff that I was actually doing was not data science but I was given the opportunity to do it.

- **Team Leaders (Section 5.5)**
  - **culture**
    - **culture-advocate**

P3
Codes: [culture-advocate] [DS team structure] [weekly data meet up]

So he takes it to a group, basically, with his boss and his peers and they go through the data. So X is my boss' boss and he runs sort of the data meet-up weekly with a whole group of people. So X can definitely give you all sorts of insights as to what goes on in that meeting. And that's – that'll be kind of interesting insights as to how they use or don't use it, as the case may be, which is still, unfortunately, true to an extent greater than I'd like to see

- **data advocate**

P5
Codes: [culture] [data advocate]

But now all the stars were aligned. The company as a whole, if you look at Satya [Microsoft CEO], he's actually been extremely good about vocalizing the need for this, and that's because he spent time in search and obviously online services, and those guys are at the forefront of doing this work.

- **data PM**

P10
Codes: [data PM] [translate results into business values]

Hey," you know, "keep everything at, like, seventh-grade level." And that – but then there's also sort of this role that we're gonna be experimenting with, which we'll call a "Data PM." And if you imagine – how do I say this? – that sometimes – [laughs] – sometimes people who are real good with numbers are not as good with words.

P10
Codes: [data PM] [translate results into business values]

so, one is that the scientists can dumb things down; another way is to put sort of an intermediary in there. And that – we think what that will do is allow sort of both directions. So, if they dumb things down and present it, then managers know how to consume it; that's great. But this intermediary can do that part, as well as go – you know, if there's a – like, [application], and there's [the application] team, like, to go interface with the [application] team and say, "Okay, we need access; here's this; here's this," I think we'll be more effective than having the data scientist do that, because it's a lot of work to really – to get in and understand the first time and get things going

- **dumb down to 7th grade**

P10
Codes: [dumb down to 7th grade] [talk simple]

I'll just say a super smart data scientist – their understanding and presentation of their findings is usually way over the head of the managers

P10
Codes: [dumb down to 7th grade]

dumb everything down to seventh-grade level, right? And whether you're writing or you're presenting charts, you know, keep it simple

- **explain**

P9
Codes: [define actions and triggers] [explain]

I think the one challenge of data scientists is that often times even if they have this amazing algorithm that finds anomalies is explaining what these are and what to do with them

- **simplicity**

P1
Codes: [simplicity]

And most of the time, the other interesting thing about it is that one of the things that I always try to tell people is that fancy is not good. Fancy is bad. Fancy requires you to be there in order for the person reading it to figure out what's going on, so the easier, dumber, the better.

P3
Codes: [simplicity]

So we tend to go with very simple things to start with and then as they start to understand it, you can go better and better and better models.

- **people should be accountable for bad decisions**

P3
Codes: [culture] [people should be accountable for bad decisions]

It's just people – we don't have a culture where people are held accountable to their – to the process used to make a decision. So I can see if you make a bad decision based on all the right data, that's just bad luck.

Implications and Guidelines (Section 6)

- **human in the loop**

P4
Codes: [human in the loop] [validation]

So I am a believer that you have to apply some human thing.

P4
Codes: [human in the loop]

To the few that have the knowledge, the knowledge just in the data is not enough. You will need a person who knows the space.

- **human in the loop in defining scenarios to focus on**

P4
Codes: [human in the loop in definining scenarios to focus on]

They put the report out and nobody believed it. And I told them that look, if you did not instrument upfront all the features and you have some sense that this a complete set of features, people are not going to believe the report and they didn't believe it because we had a couple of features, but people intuitively expected something seems to be in the least at all and they didn't see it and they said well, what is that? Well, we don't have instrumentation right. Okay. Fine.

- **involve data scientist early**

P5
Codes: [involve data scientist early] [recommendation]

So the problem is, is that for you to really be able to be experimentation focused and data driven, you need to be able to validate your hypothesis before you've done all that work. Because once you've invested all that effort, you're gonna ship this thing no matter what.

- **involve data scientist early before human capital**

P5
Codes: [involve data scientist early before human capital] [recommendation]

so what's really important for services like that is to have the ability to test your hypothesis before you've invested all that human capital in it, ideally in the planning phase

P5
Codes: [involve data scientist early before human capital] [recommendation]

What it is, is by the time you commit to do – by the time you've done an effort, whatever that effort is, there's just some cost, and as a human you're not gonna give it up, especially when your review depends on it.

- **Iteration and refinement**

P5
Codes: [iteration and refinement]

Then you need to have the iteration with the – both with the customer like we said on a regular basis and also internally within us to make sure your project is on a good path and leading to the end.

P7
Codes: [define groudtruth with feature teams] [iteration and refinement]

You have communication going back and forth where you will find what you're actually looking for, what is anomalous and what is not anomalous in the set of data that they looked at.

P11
Codes: [iteration and refinement]

It was a very iterative approach, so don't remember if we had explicit session with the analyst, say, "Hey, let's sit down. Let's design this thing." It was more of, "Let's try this. Let's try that," and then – and see how people respond to things, and then let's find out how we should convey

P14
Codes: [agile data collection] [iteration and refinement]

Until you actually have a thousand devices or a million devices out in the wild of people using them, and you're now seeing the data flowing out those devices, you didn't realize that you had that bug. You didn't realize you needed that other data point. You didn't realize. So until you actually start writing queries that look at the real data, you're not actually gonna have the best opinion of what that data should be. The faster you get people to having any starting point of that data is the amazing power.

- **Value of insights**
  - **insights - knowledge pyramid, data, information, insights, actions**

P5
Codes: [insights - knowledge pyramid, data, information, insights, actions]

This is data. It could be log data, could be anything you're collecting. This is action. Something's actually happened because of that. This is information, and this is insight. Information I think of as metrics, averages, whatever. Scorecards. Graphs.

All those things are information, and what I would always have to spend time impressing on my data scientists is that their job was really to get there and to influence that, and it is really hard for people to get there. It is real easy to get there. They can do all kinds of graphs, all kinds of analysis, they can show me all kinds of tables or all kinds of numbers. See, [my name]? See? See? Look, look, look. But the point is not that. The point is okay, great. Given all that, so what's the insight? What's the aha? Oh, aha, that then someone can actually take an action on.

- **How to choose questions: three conditions - priority, actionable, and commitment**

P5
Codes: [question development process] [three conditions - priority, actionable, and commitment]

A, is it a priority for the organization, B, is it actionable – if I get an answer to this, is this something someone can do something with? – and, C, are you as the feature team, if you're coming to me or if I'm going to you, telling you this is a good opportunity, are you committing resources to deliver a change? If those things are not true, then it's not worth us talking anymore