Adapting to the evolving nature of data through governance

Julie Hollek

Summary

Data Products & Data Science

Consumer and Regulatory Concerns

Case Study: Revenue Data Access Initiative

Background + Acknowledgements

Senior DS + ML manager at Mozilla

- Metrics, Revenue, ML/Data Products, Subscription Services
- Previously: internet health, ad tech

Thank you to the Mozilla Revenue Data Group and Xuan Luo, Arkadiusz Komarzewski

We’re hiring!

careers.mozilla.org

Product Thinking and Data

Data Product

“*A product that facilitates an end goal through the use of data*”

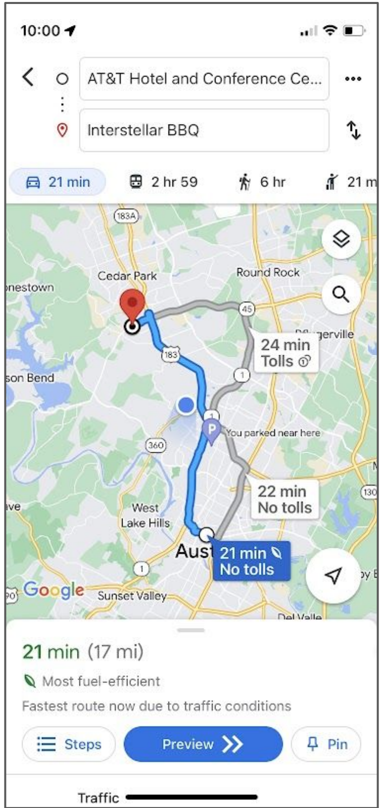
DJ Patil, Data Jujitsu: The Art of Turning Data into Product

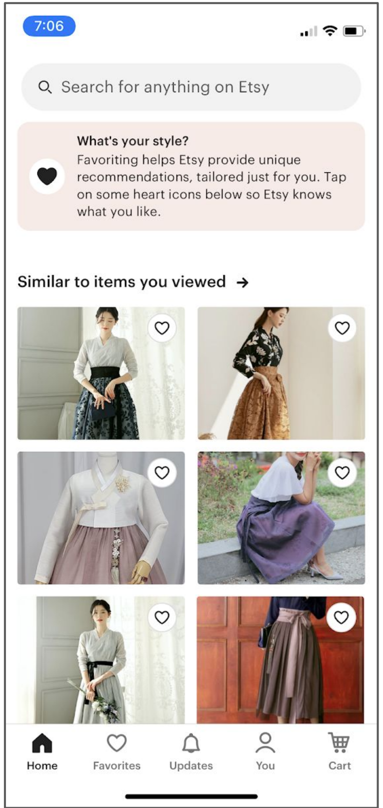
Data as a Product

“...data teams must apply product thinking [...] to the datasets that they provide; considering their data assets as their products and the rest of the organization’s data scientists, ML and data engineers as their customers.”

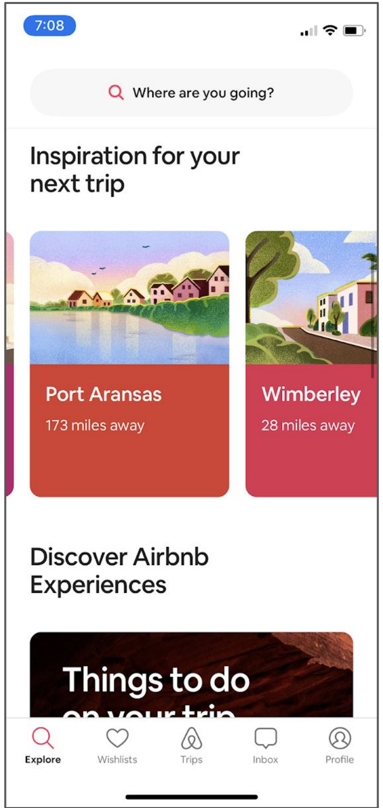
Zhamak Dehghani, How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh

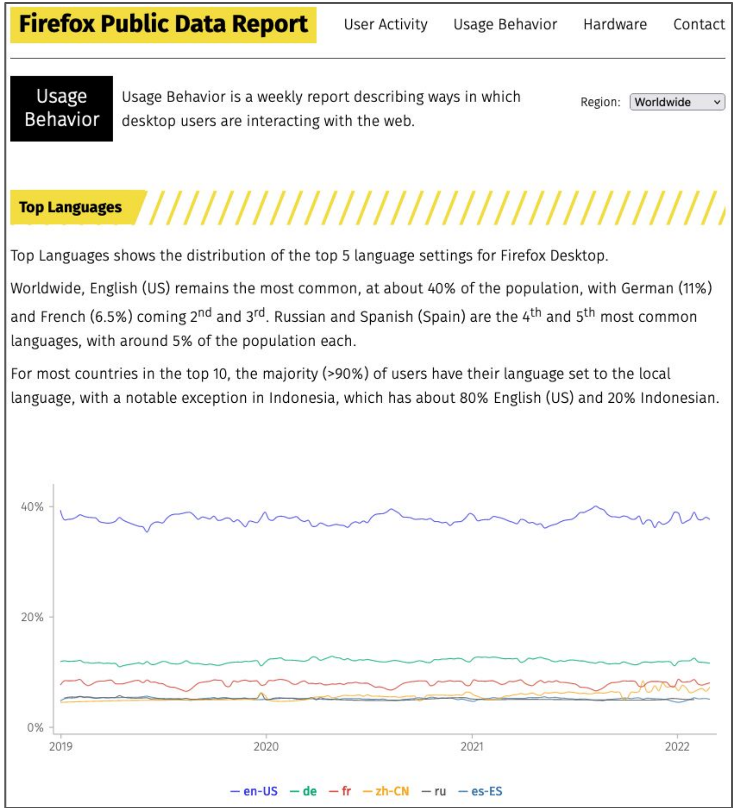
Data Products in the Wild





What's your style?
Favoriting helps Etsy provide unique recommendations, tailored just for you. Tap on some heart icons below so Etsy knows what you like.





Usage Behavior is a weekly report describing ways in which desktop users are interacting with the web.

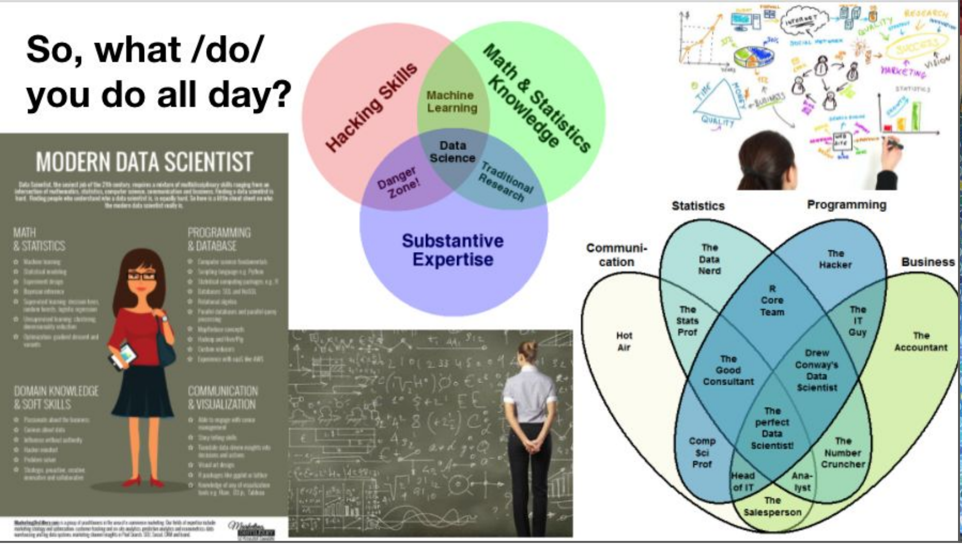
Top Languages

Top Languages shows the distribution of the top 5 language settings for Firefox Desktop.

Worldwide, English (US) remains the most common, at about 40% of the population, with German (11%) and French (6.5%) coming 2nd and 3rd. Russian and Spanish (Spain) are the 4th and 5th most common languages, with around 5% of the population each.

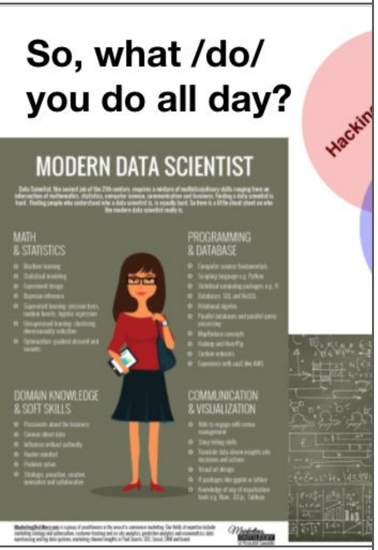
For most countries in the top 10, the majority (>90%) of users have their language set to the local language, with a notable exception in Indonesia, which has about 80% English (US) and 20% Indonesian.

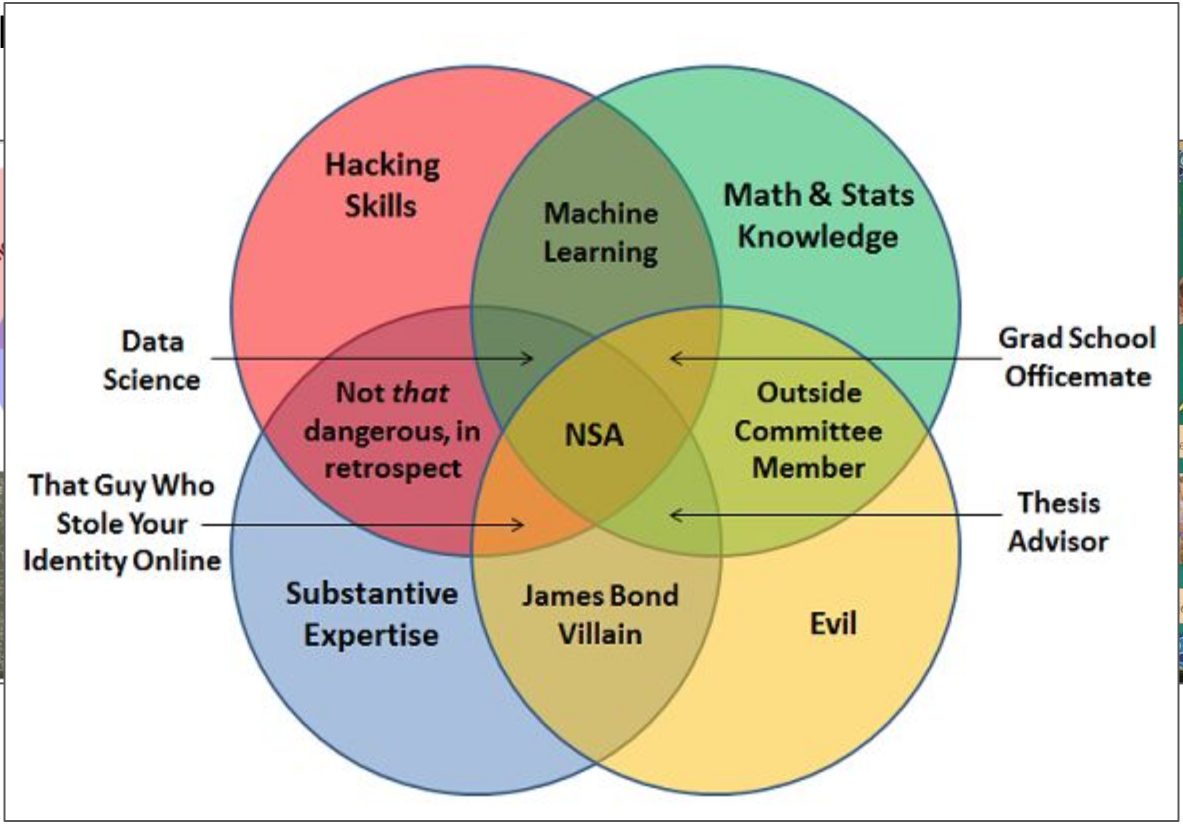
Data Science

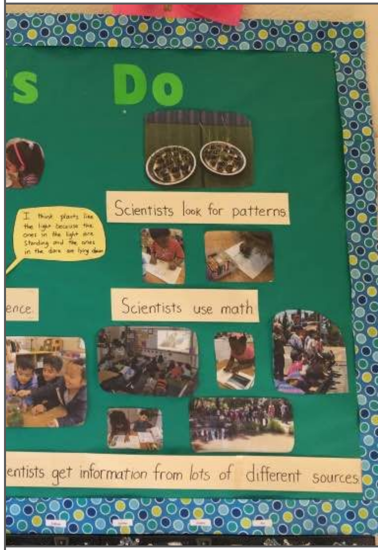


What Scientists Do
Scientists communicate information
Scientists make observations
Scientists look for patterns
Scientists ask questions
Scientists support their ideas with evidence
Scientists use math
Scientists make and use models
Scientists do investigations
Scientists get information from lots of different sources

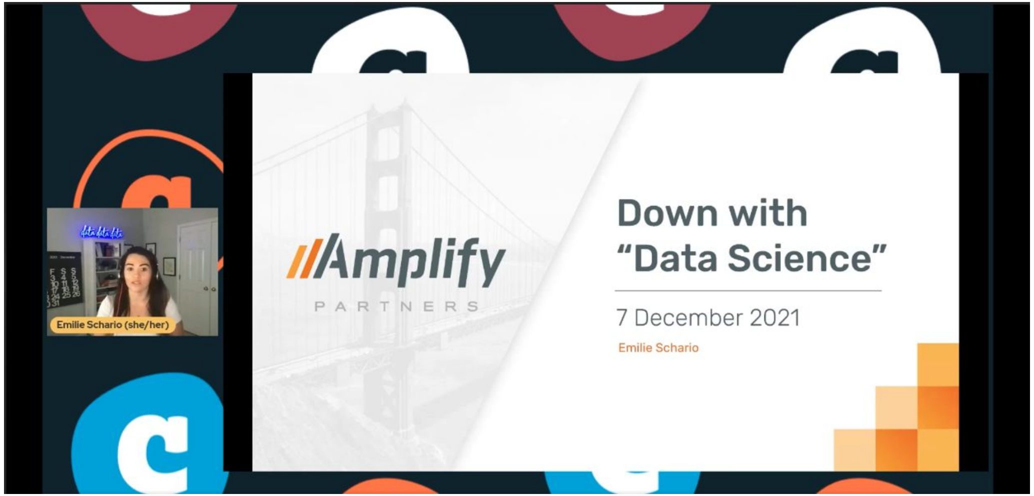
Data Scie


So, what /do/ you do all day?
MODERN DATA SCIENTIST
MATH & STATISTICS
PROGRAMMING & DATABASE
DOMAIN KNOWLEDGE & SOFT SKILLS
COMMUNICATION & VISUALIZATION
Hackin


Hacking Skills
Machine Learning
Math & Stats Knowledge
Data Science
Grad School Officemate
Not *that* dangerous, in retrospect
NSA
Outside Committee Member
That Guy Who Stole Your Identity Online
Thesis Advisor
Substantive Expertise
James Bond Villain
Evil


s Do
Scientists look for patterns
I think plants like the light because the ones in the light are standing and the ones in the dark are lying down
ence
Scientists use math
entists get information from lots of different sources

Beyond the Venn Diagram



Down with “Data Science”

7 December 2021

Emilie Schario

Emilie Schario - Down with “Data Science”

- Data Engineers
- Analytics Engineers
- Data Analysts
- Machine Learning Engineers

Analytics

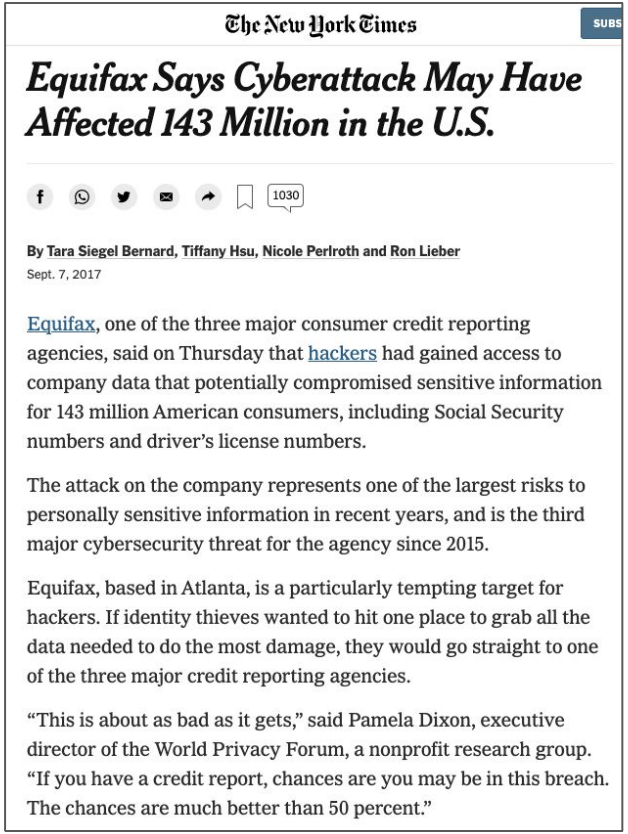
Insight as output

- Play the *objective* voice of your customers
- Metrics + measurements frame how your company views its health
- Looks like
  - Opportunity Sizing/Prototyping
  - Experimentation
  - Impact Analyses

What is the secret of ~~Soylent Green~~ analytics?

(data)

It's the year 2022...
People are still the same.
They'll do anything to get what they need.
And they need SOYLENT GREEN.
RIOT CONTROL
SOYLENT GREEN
MGM Presents
CHARLTON HESTON · LEIGH TAYLOR-YOUNG in SOYLENT GREEN
Co-Starring
CHUCK CONNORS · JOSEPH COTTEN · BROCK PETERS · PAULA KELLY and EDWARD G. ROBINSON
Screenplay by STANLEY R. GREENBERG · Based on a novel by HARRY HARRISON · Produced by WALTER SELTZER and RUSSELL THACHER · Directed by RICHARD FLEISCHER
METROCOLOR · PANAVISION®
MGM



The New York Times

SUBS

Equifax Says Cyberattack May Have Affected 143 Million in the U.S.

1030

By Tara Siegel Bernard, Tiffany Hsu, Nicole Perlroth and Ron Lieber

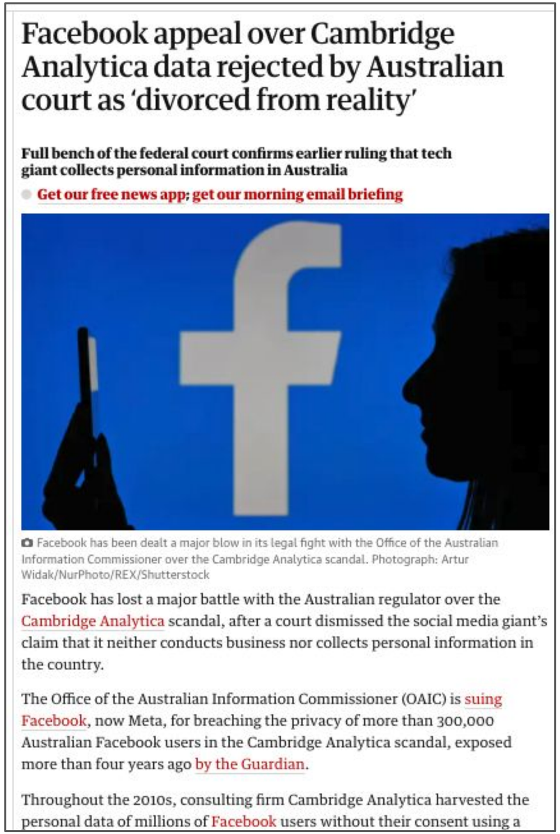
Sept. 7, 2017

Equifax, one of the three major consumer credit reporting agencies, said on Thursday that hackers had gained access to company data that potentially compromised sensitive information for 143 million American consumers, including Social Security numbers and driver's license numbers.

The attack on the company represents one of the largest risks to personally sensitive information in recent years, and is the third major cybersecurity threat for the agency since 2015.

Equifax, based in Atlanta, is a particularly tempting target for hackers. If identity thieves wanted to hit one place to grab all the data needed to do the most damage, they would go straight to one of the three major credit reporting agencies.

"This is about as bad as it gets," said Pamela Dixon, executive director of the World Privacy Forum, a nonprofit research group. "If you have a credit report, chances are you may be in this breach. The chances are much better than 50 percent."



Facebook appeal over Cambridge Analytica data rejected by Australian court as 'divorced from reality'

Full bench of the federal court confirms earlier ruling that tech giant collects personal information in Australia

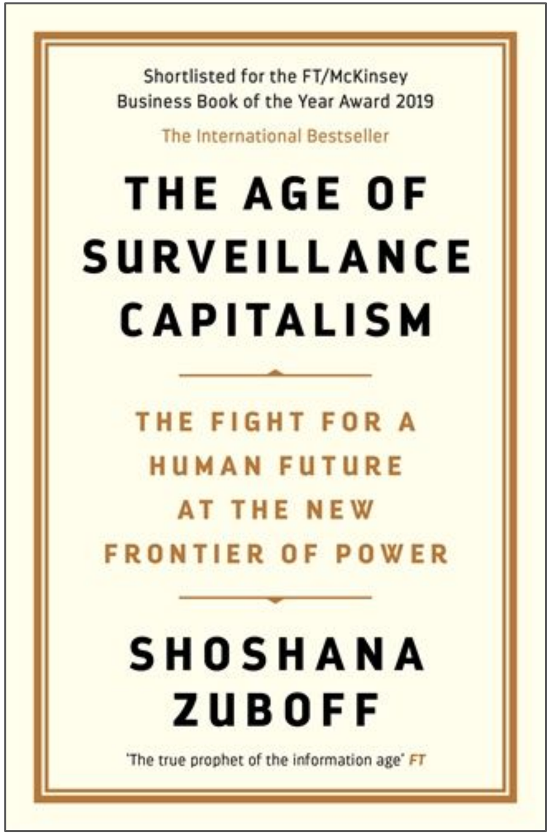
- **Get our free news app; get our morning email briefing**

Facebook has been dealt a major blow in its legal fight with the Office of the Australian Information Commissioner over the Cambridge Analytica scandal. Photograph: Artur Widak/NurPhoto/REX/Shutterstock

Facebook has lost a major battle with the Australian regulator over the Cambridge Analytica scandal, after a court dismissed the social media giant's claim that it neither conducts business nor collects personal information in the country.

The Office of the Australian Information Commissioner (OAIC) is suing Facebook, now Meta, for breaching the privacy of more than 300,000 Australian Facebook users in the Cambridge Analytica scandal, exposed more than four years ago by the Guardian.

Throughout the 2010s, consulting firm Cambridge Analytica harvested the personal data of millions of Facebook users without their consent using a



Shortlisted for the FT/McKinsey Business Book of the Year Award 2019

The International Bestseller

THE AGE OF SURVEILLANCE CAPITALISM

THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER

SHOSHANA ZUBOFF

'The true prophet of the information age' *FT*

The **General Data Privacy Regulation** or GDPR is part of the privacy and human rights laws of the EU that set the standards of how companies collect, handle, and protect personal data for EU citizens.

- Users can know how their data are used, what data companies have about them, correct mistakes in the data, have their data deleted, and opt out
- Companies must pay fines for non-compliance such as data breaches or lack of user consent

This is the first of many regulatory standards worldwide. The **California Consumer Privacy Act** is another that is US-based.

Data governance is the set of roles, policies, processes, and technologies that empower an organization to consistently and appropriately handle its data.

Why is this important? It ensures compliance, security, privacy, quality, availability, and usability. It ultimately provides the foundation for an organization’s data strategy.

Case Study: Revenue Data Access Initiative

How do we leverage sensitive data for insight and understanding?

Revenue Data Threat Model

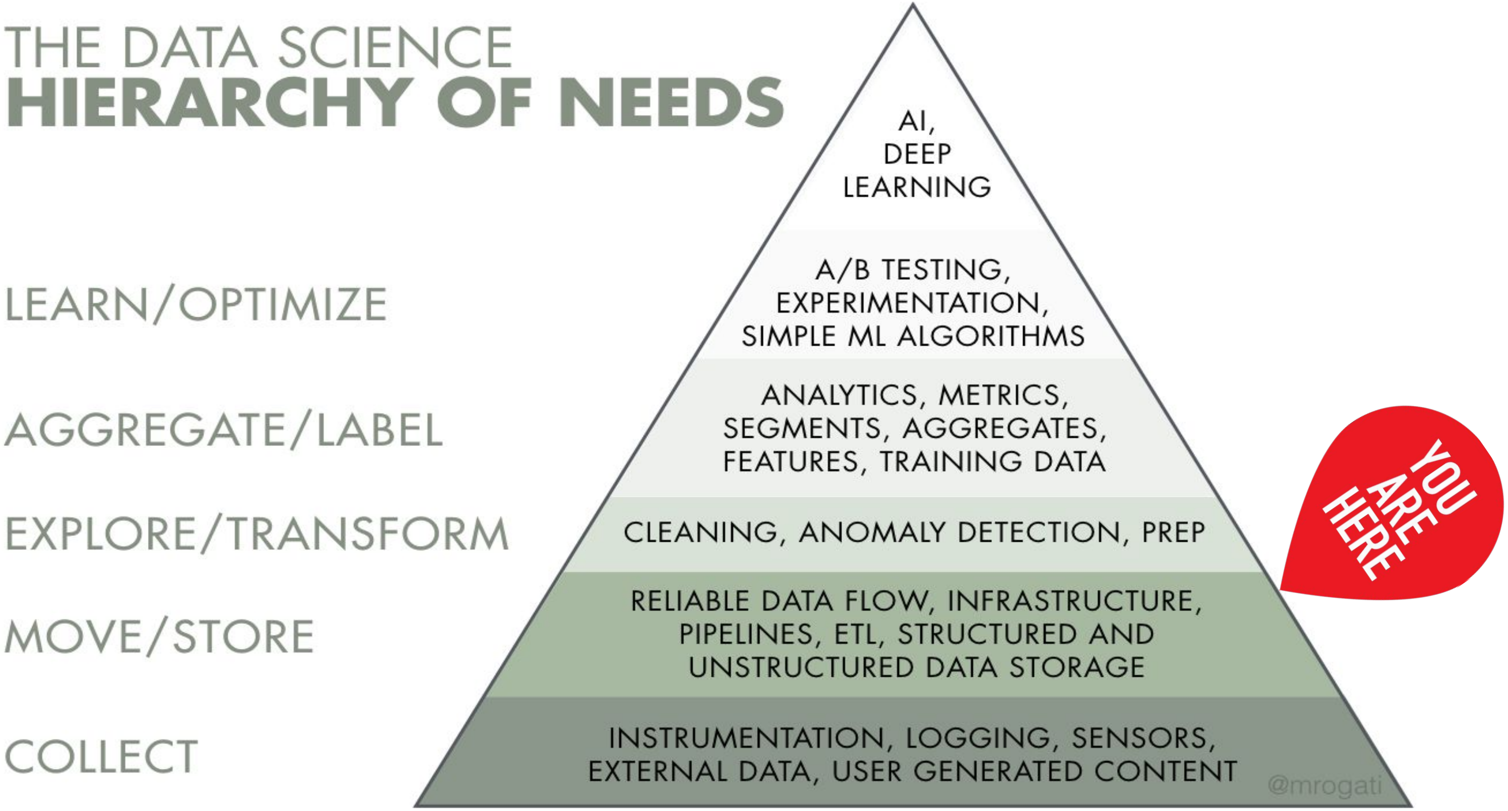
What are we making?
What threats are we concerned about?
What can we do to mitigate these threats?
Do these mitigations work?

Adapted from Toreon threat modeling materials

Revenue data are

- important to analyze because they will allow us to make better choices about our business
- sensitive because they contain confidential information that present risk to the business
  - Do not contain personal information
- in need of system designed to
  - grant access to only those who need access to it for processing, evaluation, or decision-making purposes
  - restrict access from the rest of the company

Monica Rogati’s Hierarchy of Needs


THE DATA SCIENCE
HIERARCHY OF NEEDS
LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
AI, DEEP LEARNING
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
YOU ARE HERE
@mrogati

Revenue Data

- PDFs and Spreadsheets and APIs, oh my!
- Sometimes hand-curated, from non-Mozillians, mostly maintained by hand by non-technologists
- *sensitive*
- All-or-nothing access, but difficult to use
- safeguarded by the CFO herself

Revenue Data Science @ Mozilla

Revenue Forecasting

Product Data Science for our monetizable surface areas

Data Help for Finance and Business Operations Analysts

- Methods
- Data

Given that this is what they do, what do Rev DS look like?

Data Scientist

- Tend to have advanced degrees (Ph.D, MS) in a STEM field
- Advanced skills in SQL and scripting language (usually Python or R)

Finance/Business Analyst

- Subject matter expert
- Simple SQL skills, proficient in Excel
- Straightforward domain-relevant modeling

Revenue Data Access Initiative

- Framework
  - Policy
  - Process
- Technical Infrastructure
  - Differential Access Implementation
  - Data Pipeline Migration & Improvements
- Empowerment
  - Visualization Layer

Policy

Principles-first approach to understand who should get access to sensitive data

Spell out why you need these particular principles

Categories of Data

1	Data that are sensitive but extremely difficult or impossible to calculate sensitive quantities
2	Data that allow someone to back-calculate sensitive quantities
3	Highly sensitive, restricted, and rarely shared data that must be kept confidential

Policy

Principles-first approach to understand who should get access to sensitive data

Role-based Access

- Permanent - you have a job at the company that requires you to deal with these data regularly
- Project - you're working on a project that requires these data but this is due to the project and not your position

Policy

Principles-first approach to understand who should get access to sensitive data

Compliance

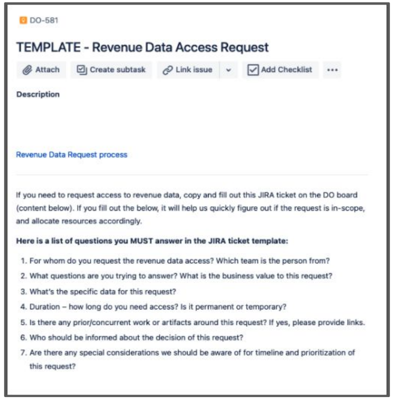
People with access to these data must take a test to demonstrate that they have read and understood the sensitive information training and sign an acknowledgement that they will comply

Process

Practical, standardized workflow to apply our policy

Request and Evaluation Flow

Jira Software



DO-581

TEMPLATE - Revenue Data Access Request

Description

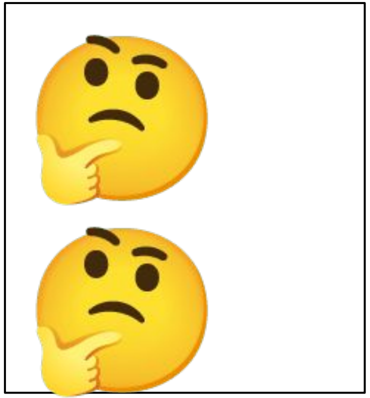
Revenue Data Request process

If you need to request access to revenue data, copy and fill out this JIRA ticket on the DO board (content below). If you fill out the below, it will help us quickly figure out if the request is in-scope, and allocate resources accordingly.

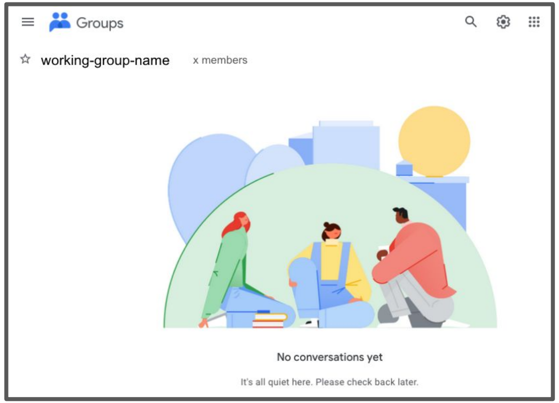
Here is a list of questions you MUST answer in the JIRA ticket template:

1. For whom do you request the revenue data access? Which team is the person from?
2. What questions are you trying to answer? What is the business value to this request?
3. What's the specific data for this request?
4. Duration – how long do you need access? Is it permanent or temporary?
5. Is there any prior/concurrent work or artifacts around this request? If yes, please provide links.
6. Who should be informed about the decision of this request?
7. Are there any special considerations we should be aware of for timeline and prioritization of this request?

Data stewards



Google Groups



working-group-name x members

No conversations yet

It's all quiet here. Please check back later.

Framework

Technical

Empowerment

Process

Practical, standardized workflow to apply our policy

Auditing

- Quarterly audits on permanent access
  - Requires manager and access steward approval
- Extension evaluation for temporary access if needed for project
  - Request is evaluated by access stewards

Technical Infrastructure

Differential Access Implementation

give access to those who need it and restrict access from those who don’t

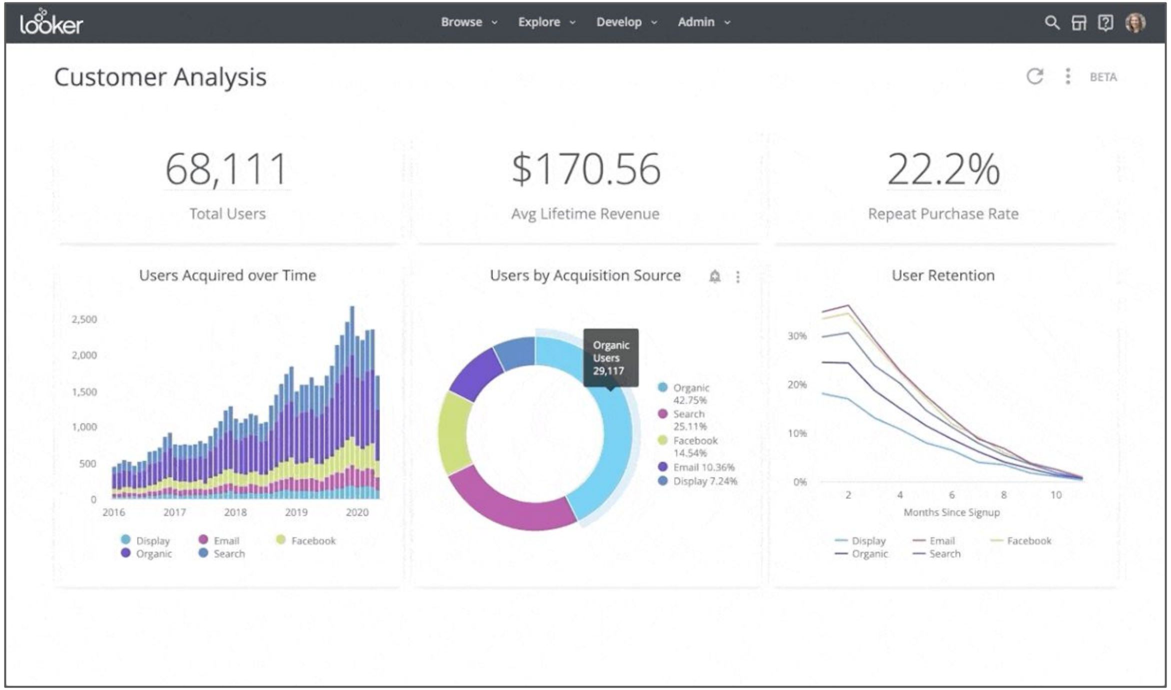
- Leverages BigQuery’s authorized views to create differential access based on revenue access policy specifications

Data Pipeline Migration & Improvement

contain all of the revenue data in one place to make easier access configuration, more robust datasets, SRE support

- Syndication of data, new ETL, new connectors that standardize and stabilize pipelines

Visualization Layer



Customer Analysis

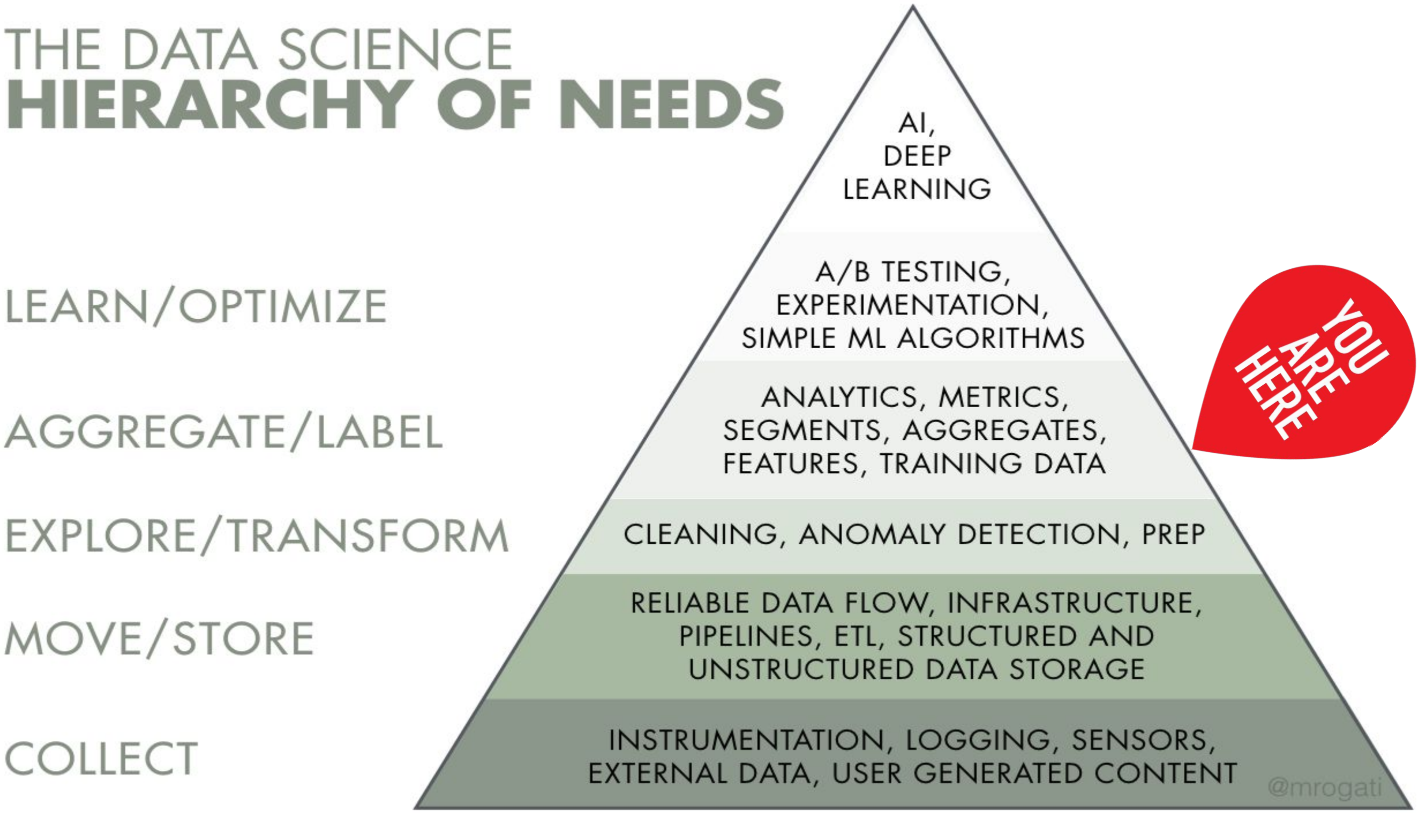
68,111
Total Users

$170.56
Avg Lifetime Revenue

22.2%
Repeat Purchase Rate

Framework Technical **Empowerment**

Monica Rogati’s Hierarchy of Needs



THE DATA SCIENCE
HIERARCHY OF NEEDS
AI, DEEP LEARNING
LEARN/OPTIMIZE
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
AGGREGATE/LABEL
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
YOU ARE HERE
EXPLORE/TRANSFORM
CLEANING, ANOMALY DETECTION, PREP
MOVE/STORE
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
COLLECT
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
@mrogati

How does this tie back to data privacy?

Domain	Revenue Data	Personal Data
Framework	Policy based on business risk	Policy based on user privacy
Technical Infrastructure	Differential access, data warehouse	Differential access, data warehouse
Empowerment	Insights	Insights, user-facing data product