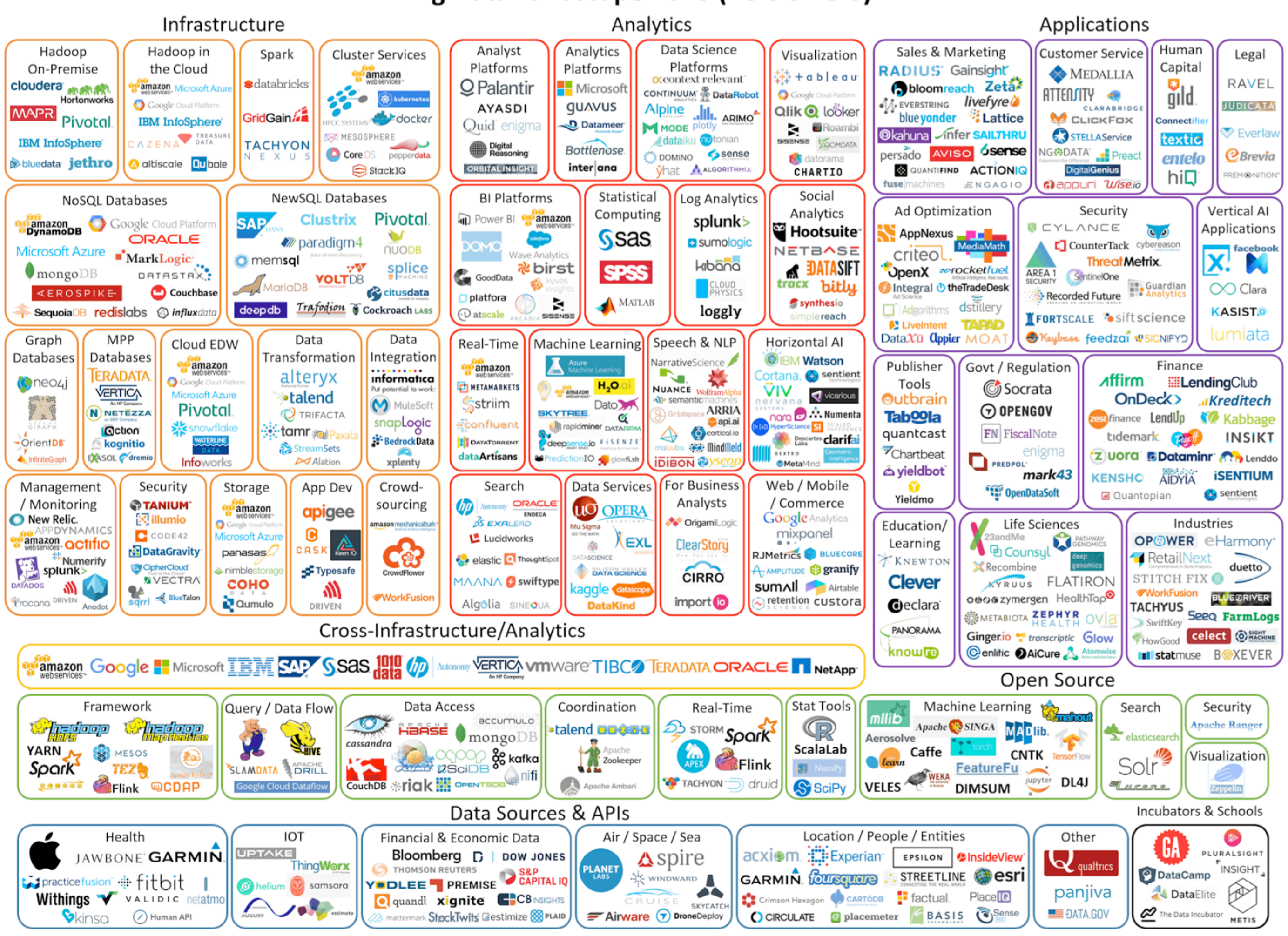
# Changes in data ownership and usage

Preliminary thoughts

Katrina Ligett and Kobbi Nissim

amazon
UBER
Microsoft
Google
f
TESLA
David Parkins

# Big Data Landscape 2016 (Version 3.0)



## Infrastructure

### Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, bluedata, jethro

### Hadoop in the Cloud
amazon web services, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, TREASURE DATA, altiscale, Qubole

### Spark
databricks, GridGain, TACHYON NEXUS

### Cluster Services
amazon web services, kubernetes, HPCC SYSTEMS, docker, MESOSPHERE, CoreOS, pepperdata, StackIQ

### NoSQL Databases
amazon DynamoDB, Google Cloud Platform, ORACLE, Microsoft Azure, MarkLogic, mongoDB, DATASTAX, AEROSPIKE, Couchbase, SequoiaDB, redislabs, influxdata

### NewSQL Databases
SAP HANA, Clustrix, Pivotal, paradigm4, NuoDB, memsql, splice machine, MariaDB, VOLTDB, citusdata, deepdb, Trafodion, Cockroach LABS

### Graph Databases
neo4j, TITAN, OrientDB, InfiniteGraph

### MPP Databases
TERADATA, VERTICA, NETEZZA, Actian, kognitio, EXASOL, dremio

### Cloud EDW
amazon web services, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, WATERLINE DATA, Infoworks

### Data Transformation
alteryx, talend, TRIFACTA, tamr, Paxata, StreamSets, Alation

### Data Integration
informatica, MuleSoft, snapLogic, BedrockData, xplenty

### Management / Monitoring
New Relic, APPDYNAMICS, amazon web services, actifio, Numerify, splunk, DATADOG, Anodot, DRIVEN

### Security
TANIUM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrrl, BlueTalon

### Storage
amazon web services, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, COHO DATA, Qumulo

### App Dev
apigee, CASK, Keen IO, Typesafe, DRIVEN

### Crowd-sourcing
amazon mechanical turk, CrowdFlower, WorkFusion

## Analytics

### Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITAL INSIGHT

### Analytics Platforms
Microsoft, GUAVUS, Datameer, Bottlenose, interana

### Data Science Platforms
context relevant, CONTINUUM ANALYTICS, DataRobot, Alpine, ARIMO, MODE, plotly, dataiku, nutonian, DOMINO, sense, yhat, ALGORITHMIA

### Visualization
tableau, Google Cloud Platform, Qlik, looker, SISENSE, Roambi, ZOOMDATA, datorama, CHARTIO

### BI Platforms
Power BI, amazon web services, DOMO, salesforce Wave Analytics, GoodData, birst, kyvos insights, platfora, ARCADIA, SISENSE, atscale

### Statistical Computing
sas, SPSS, MATLAB

### Log Analytics
splunk, sumologic, kibana, CLOUD PHYSICS, loggly

### Social Analytics
Hootsuite, NETBASE, DATASIFT, traackr, bitly, synthesio, simplereach

### Real-Time
amazon web services, METAMARKETS, striim, confluent, DATATORRENT, dataArtisans

### Machine Learning
Azure Machine Learning, H2O.ai, amazon web services, Dato, SKYTREE, rapidminer, DATARPM, deepsense.io, VISENZE, PredictionIO, glowfish

### Speech & NLP
NarrativeScience, NUANCE, WolframAlpha, semantic machines, Gridspace, ARRIA, api.ai, cortical.io, maluuba, MindMeld, IDIBON, yseop

### Horizontal AI
IBM Watson, Cortana, sentient, VIV, vicarious, nervana, nara, HyperScience, Numenta, SCALED INFERENCE, Descartes Labs, clarifai, DEXTRO, MetaMind, Geometric Intelligence

### Search
Autonomy, ORACLE ENDECA, EXALEAD, Lucidworks, elastic, ThoughtSpot, MAANA, swiftype, Algolia, SINEQUA

### Data Services
Mu Sigma, OPERA, EXL, DATA SCIENCE, kaggle, datascope, DataKind

### For Business Analysts
OrigamiLogic, ClearStory, CIRRO, import.io

### Web / Mobile / Commerce
Google Analytics, mixpanel, RJMetrics, BLUECORE, AMPLITUDE, granify, SumAll, Airtable, retention science, custora

## Applications

### Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, EVERSTRING, livefyre, blue yonder, Lattice, kahuna, infer, SAILTHRU, persado, AVISO, 6sense, QUANTIFIND, ACTIONIQ, fuse machines, ENGAGIO

### Customer Service
MEDALLIA, ATTENSITY, CLARABRIDGE, CLICKFOX, STELLAService, NGDATA, Preact, DigitalGenius, appuri, Wise.io

### Human Capital
gild, Connectifier, textio, entelo, hiQ

### Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

### Ad Optimization
AppNexus, MediaMath, criteo, OpenX, rocketfuel, Integral Ad Science, theTradeDesk, Adgorithms, distillery, LiveIntent, TAPAD, DataXu, Appier, MOAT

### Security
CYLANCE, cybereason, CounterTack, ThreatMetrix, AREA 1 SECURITY, SentinelOne, Guardian Analytics, Recorded Future, FORTSCALE, sift science, Keybase, feedzai, SIGNIFYD

### Vertical AI Applications
x.ai, facebook M, Clara, KASISTO, lumiata

### Publisher Tools
outbrain, Taboola, quantcast, Chartbeat, yieldbot, Yieldmo

### Govt / Regulation
Socrata, OPENGOV, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

### Finance
Affirm, LendingClub, OnDeck, Kreditech, zestfinance, LendUp, Kabbage, tidemark, Payoff, INSIKT, Zuora, Dataminr, Lenddo, KENSHO, AIDYIA, iSENTIUM, Quantopian, sentient

### Education / Learning
KNEWTON, Clever, declara, PANORAMA, knowre

### Life Sciences
23andMe, PATHWAY GENOMICS, Counsyl, deep genomics, Recombine, KYRUUS, FLATIRON, zymergen, HealthTap, METABIOTA, ZEPHYR HEALTH, ovia, Ginger.io, transcriptic, Glow, enlitic, AiCure, Atomwise

### Industries
OPOWER, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, TACHYUS, BLUE RIVER, SwiftKey, Seeq, FarmLogs, HowGood, celect, SIGHT MACHINE, statmuse, BOXEVER

## Cross-Infrastructure/Analytics
amazon web services, Google, Microsoft, IBM, SAP, sas, 1010data, hp, Autonomy, VERTICA, vmware, TIBCO, TERADATA, ORACLE, NetApp

## Open Source

### Framework
hadoop HDFS, hadoop MapReduce, YARN, MESOS, Spark, TEZ, Flink, CDAP

### Query / Data Flow
Pig, HIVE, SLAMDATA, APACHE DRILL, Google Cloud Dataflow

### Data Access
cassandra, HBASE, mongoDB, accumulo, kafka, CouchDB, riak, SciDB, nifi, OPENTSDB

### Coordination
talend, Apache Zookeeper, Apache Ambari

### Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

### Stat Tools
R, ScalaLab, NumPy, SciPy

### Machine Learning
mllib, mahout, Apache SINGA, Aerosolve, MADlib, scikit-learn, Caffe, torch, CNTK, TensorFlow, FeatureFu, VELES, WEKA, DIMSUM, jupyter, DL4J

### Search
elasticsearch, Solr, Lucene

### Security
Apache Ranger

### Visualization
Zeppelin

## Data Sources & APIs

### Health
JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

### IOT
UPTAKE, ThingWorx, helium, samsara, estimote

### Financial & Economic Data
Bloomberg, DOW JONES, THOMSON REUTERS, S&P CAPITAL IQ, YODLEE, PREMISE, quandl, xignite, CBINSIGHTS, mattermark, StockTwits, estimize, PLAID

### Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

### Location / People / Entities
acxiom, Experian, EPSILON, InsideView, GARMIN, foursquare, STREETLINE, esri, Crimson Hexagon, CARTODB, factual, PlaceIQ, CIRCULATE, placemeter, BASIS, Sense360

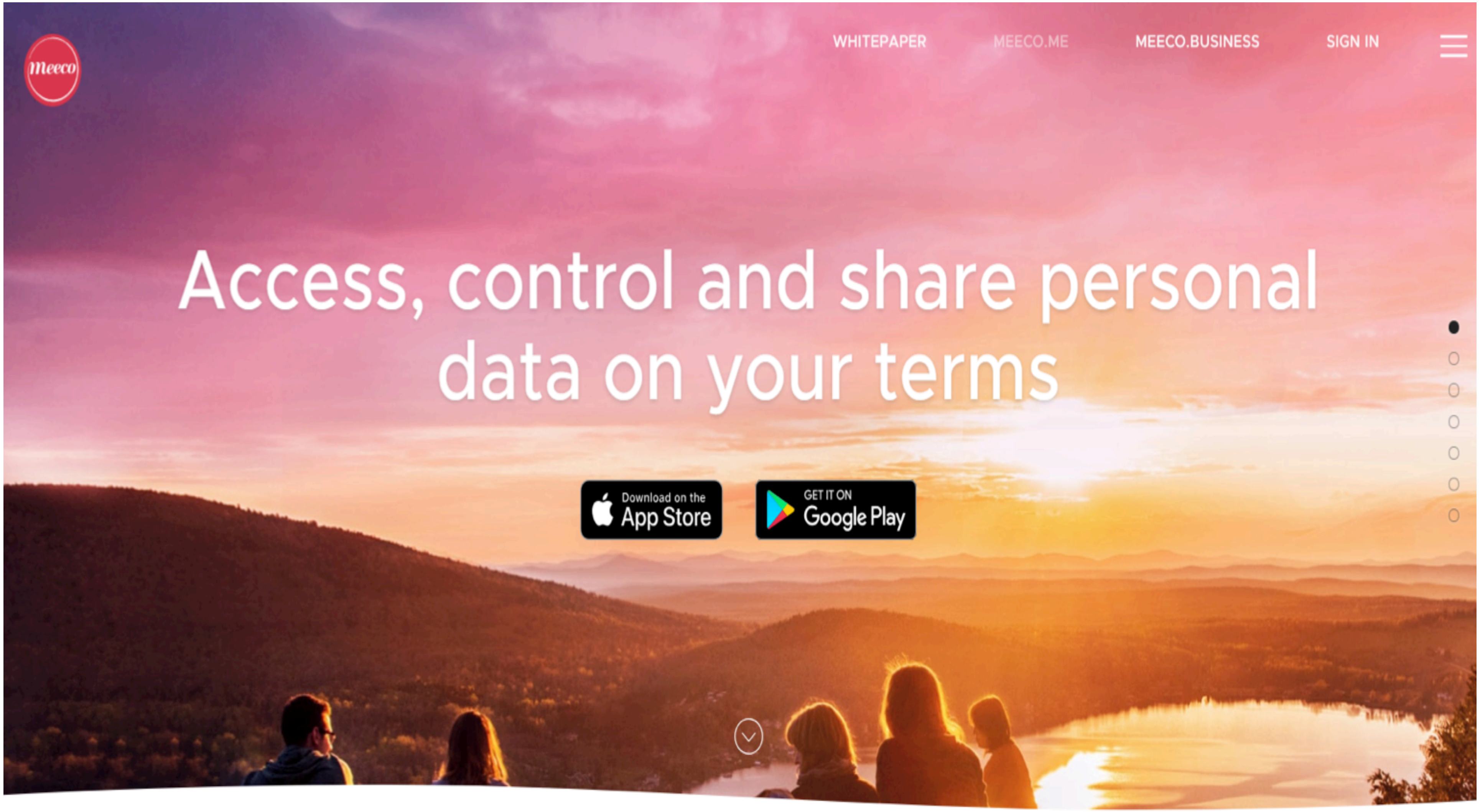
### Other
qualtrics, panjiva, DATA.GOV

### Incubators & Schools
GA, PLURALSIGHT, DataCamp, INSIGHT, DataElite, The Data Incubator, METIS

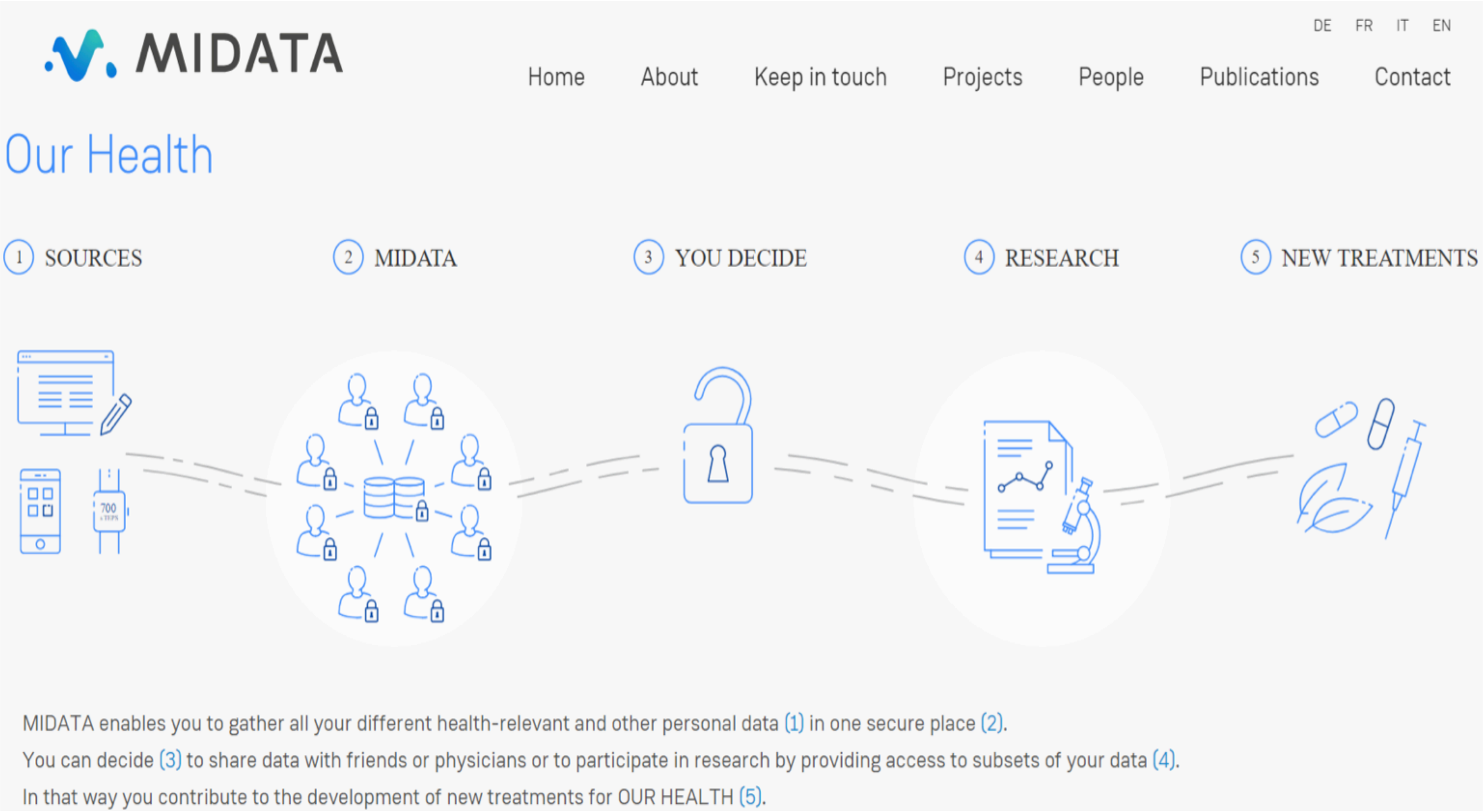
*Last Updated 3/23/2016*

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK



Access, control and share personal data on your terms

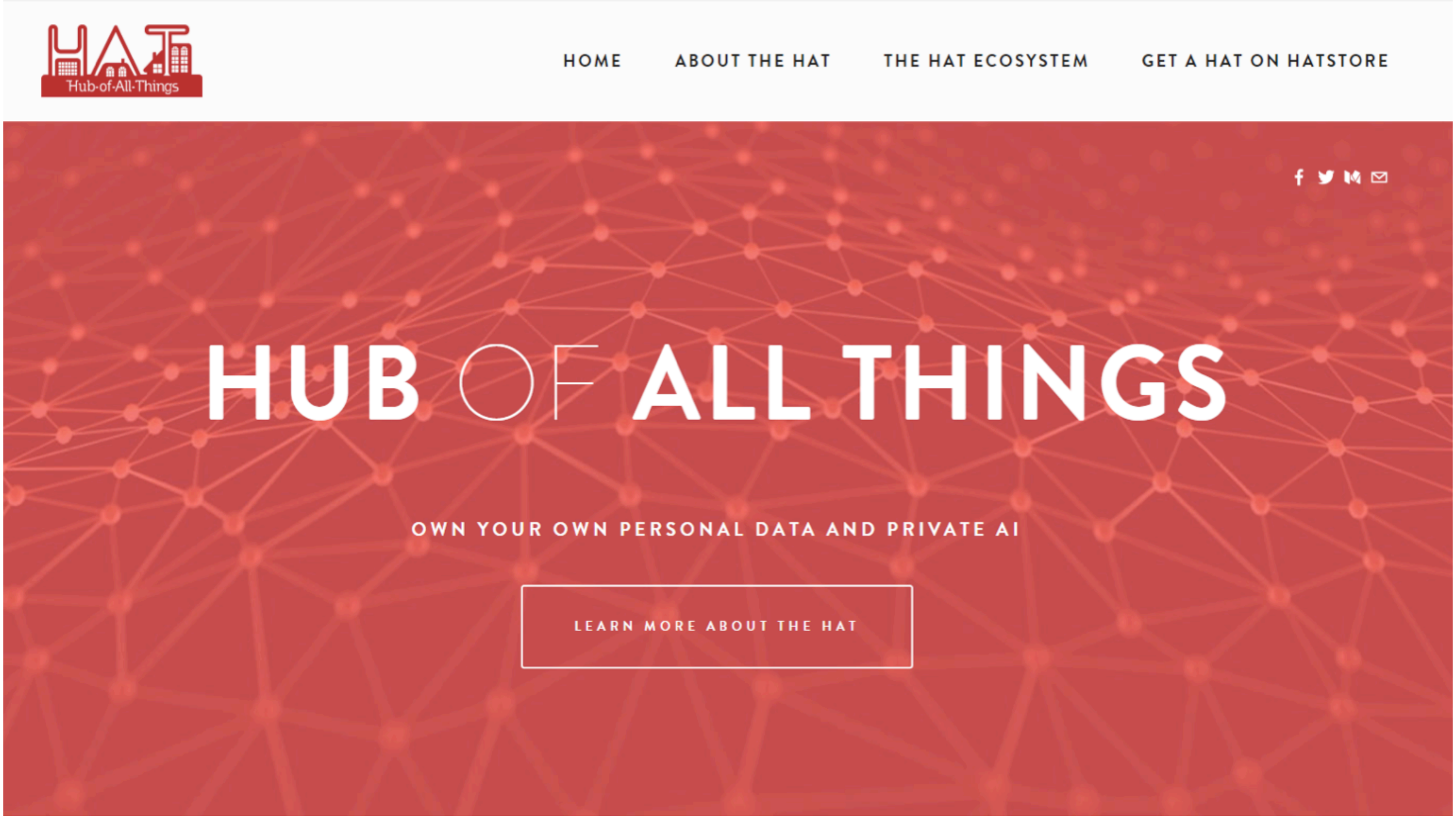


# Our Health

MIDATA enables you to gather all your different health-relevant and other personal data (1) in one secure place (2).

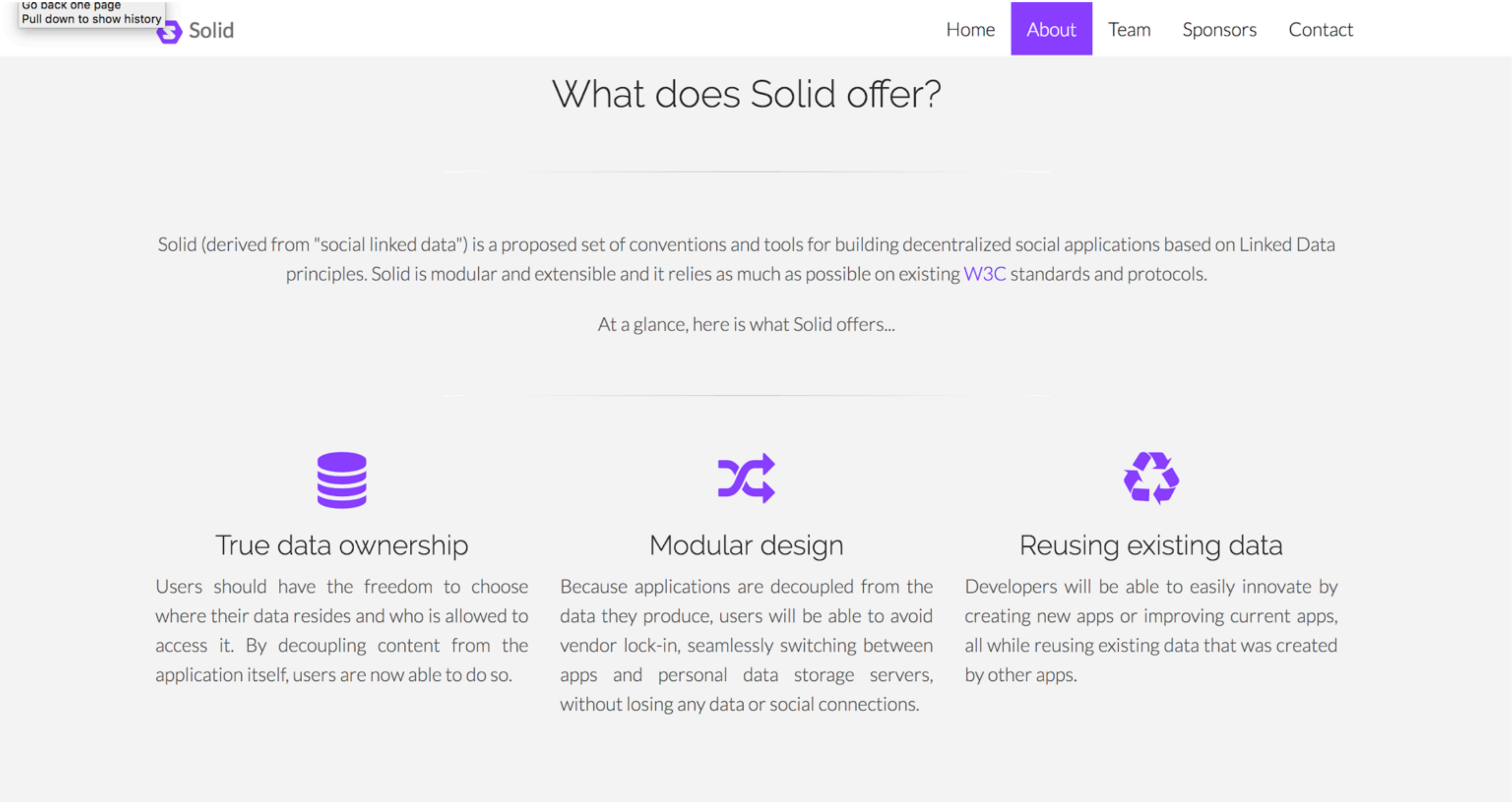
You can decide (3) to share data with friends or physicians or to participate in research by providing access to subsets of your data (4).

In that way you contribute to the development of new treatments for OUR HEALTH (5).



# HUB OF ALL THINGS

OWN YOUR OWN PERSONAL DATA AND PRIVATE AI



## What does Solid offer?

Solid (derived from "social linked data") is a proposed set of conventions and tools for building decentralized social applications based on Linked Data principles. Solid is modular and extensible and it relies as much as possible on existing W3C standards and protocols.

At a glance, here is what Solid offers...

### True data ownership

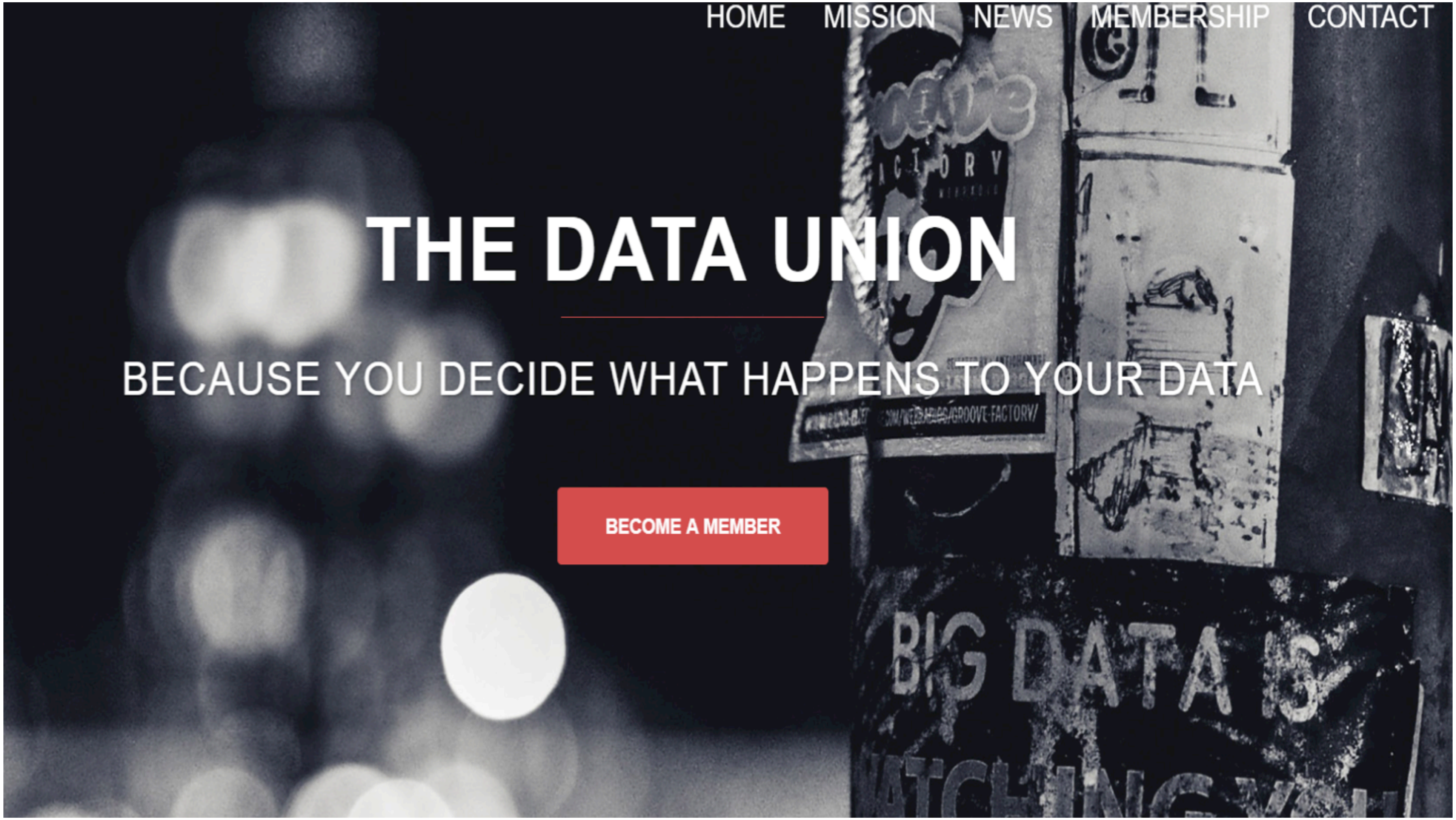
Users should have the freedom to choose where their data resides and who is allowed to access it. By decoupling content from the application itself, users are now able to do so.

### Modular design

Because applications are decoupled from the data they produce, users will be able to avoid vendor lock-in, seamlessly switching between apps and personal data storage servers, without losing any data or social connections.

### Reusing existing data

Developers will be able to easily innovate by creating new apps or improving current apps, all while reusing existing data that was created by other apps.



# THE DATA UNION

BECAUSE YOU DECIDE WHAT HAPPENS TO YOUR DATA

What problems do these projects respond to?

# Today’s internet

“*That feeling of individual control, that empowerment, is something we’ve lost*”

– Brewster Kahale, on today’s internet

A few dominant platforms = a few potential points of massive failure

- Security vulnerabilities
- Failure of incentives

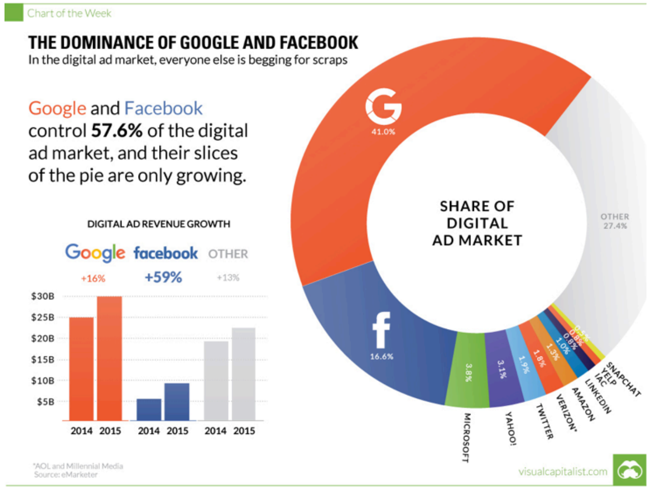


Chart of the Week

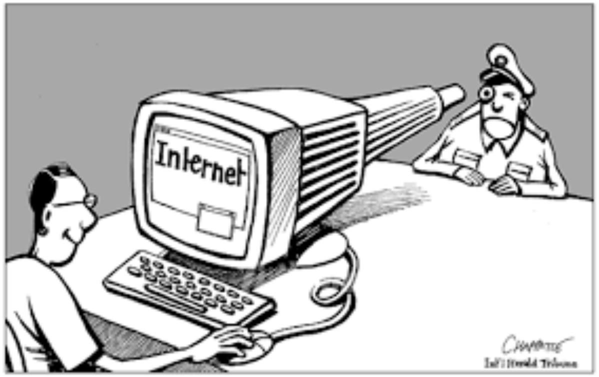
THE DOMINANCE OF GOOGLE AND FACEBOOK

In the digital ad market, everyone else is begging for scraps

Google and Facebook control **57.6%** of the digital ad market, and their slices of the pie are only growing.

*AOL and Millennial Media
Source: eMarketer

visualcapitalist.com



# Missed opportunities for innovation

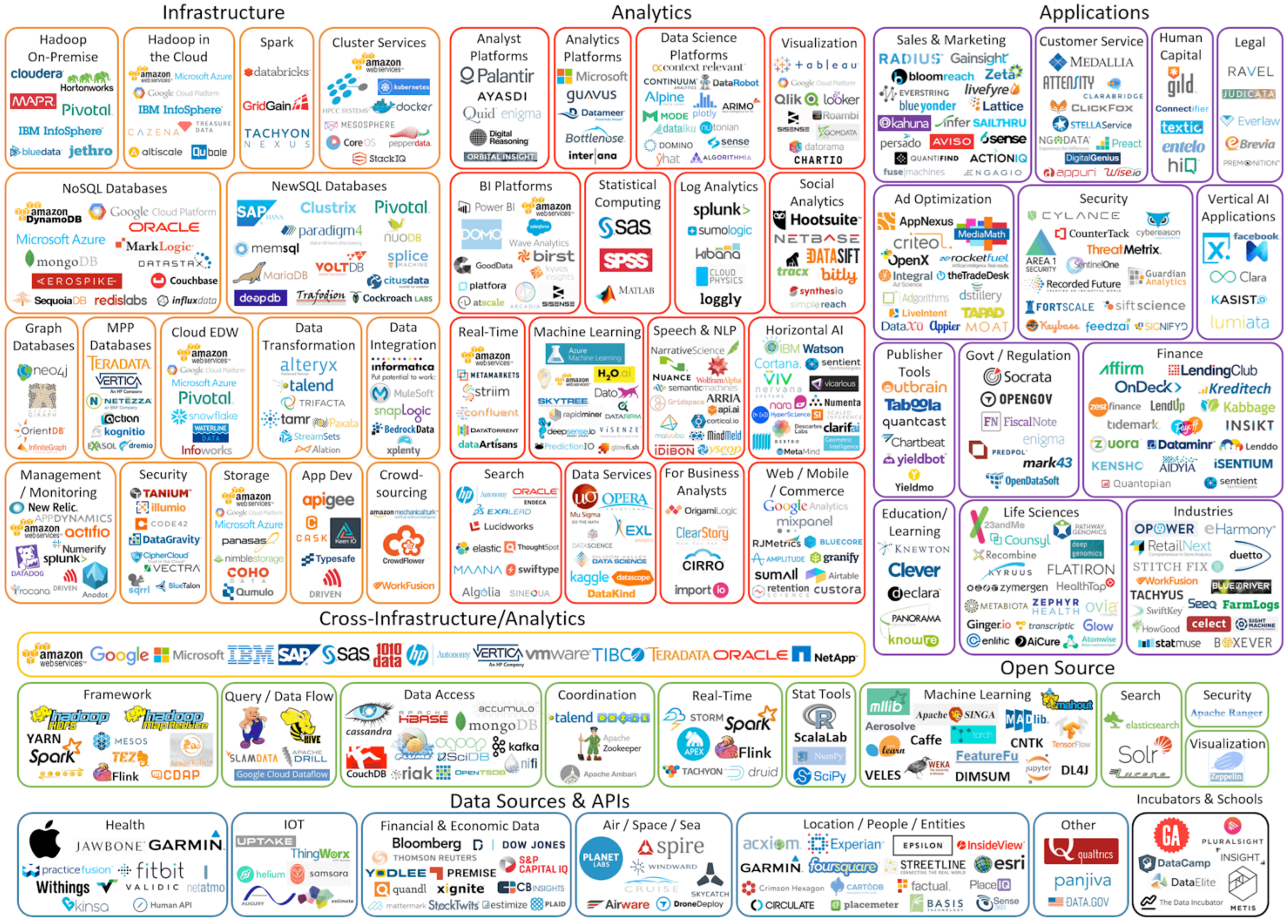
How can a small startup

- get access to data relating X and Y?
  *purchasing habits and mental health information*
- get interactive querying of highly tailored subpopulations?
  *people with predisposition to sickle cell anemia who live within 10 miles of a coal mine*
- deliver a highly personalized experience without needing highly personal data?
  *provide medical advice without having a person's medical record*

Why would high-quality data be produced/collected [c.f. Frauke's talk] when currently

- there's no market for it (chicken and egg problem)?
- quality is not compensated?
- the risks of breaches/subpoenas/GDPR violations are so costly?

# Big Data Landscape 2016 (Version 3.0)



## Infrastructure

Hadoop On-Premise

Hadoop in the Cloud

Spark

Cluster Services

NoSQL Databases

NewSQL Databases

Graph Databases

MPP Databases

Cloud EDW

Data Transformation

Data Integration

Management / Monitoring

Security

Storage

App Dev

Crowd-sourcing

## Analytics

Analyst Platforms

Analytics Platforms

Data Science Platforms

Visualization

BI Platforms

Statistical Computing

Log Analytics

Social Analytics

Real-Time

Machine Learning

Speech & NLP

Horizontal AI

Search

Data Services

For Business Analysts

Web / Mobile / Commerce

## Applications

Sales & Marketing

Customer Service

Human Capital

Legal

Ad Optimization

Security

Vertical AI Applications

Publisher Tools

Govt / Regulation

Finance

Education/ Learning

Life Sciences

Industries

## Cross-Infrastructure/Analytics

## Open Source

Framework

Query / Data Flow

Data Access

Coordination

Real-Time

Stat Tools

Machine Learning

Search

Security

Visualization

## Data Sources & APIs

Health

IOT

Financial & Economic Data

Air / Space / Sea

Location / People / Entities

Other

Incubators & Schools

*Last Updated 3/23/2016*

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcapital)

FIRSTMARK

# Lack of individual autonomy

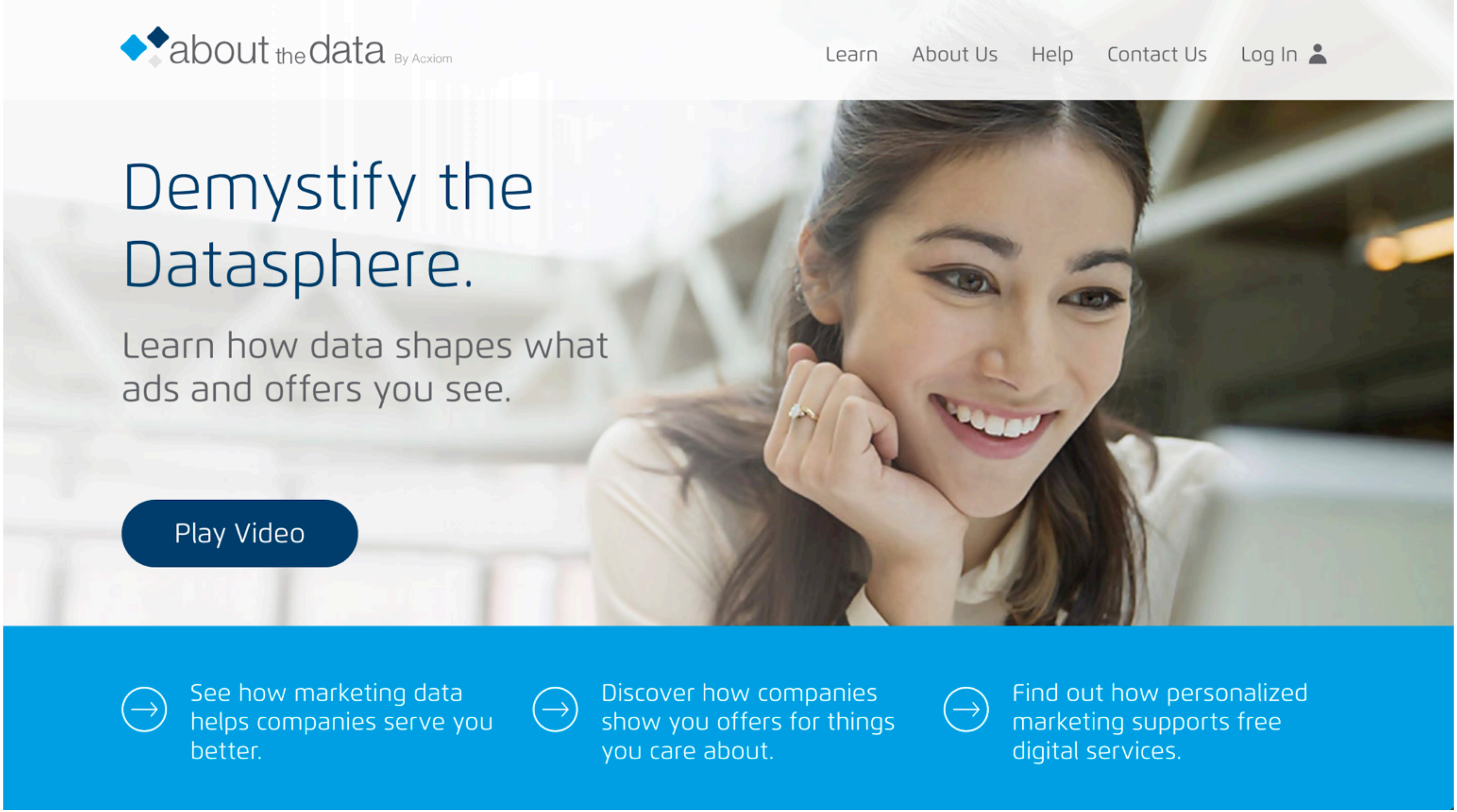
I don’t choose who gets what data of mine, for what purpose

I don’t **know** who gets what data of mine, for what purpose

I’m trading my data for something (services, etc.), but the transaction isn’t transparent

The potential future consequences aren’t transparent, either

Everybody is making money off my data, except for me



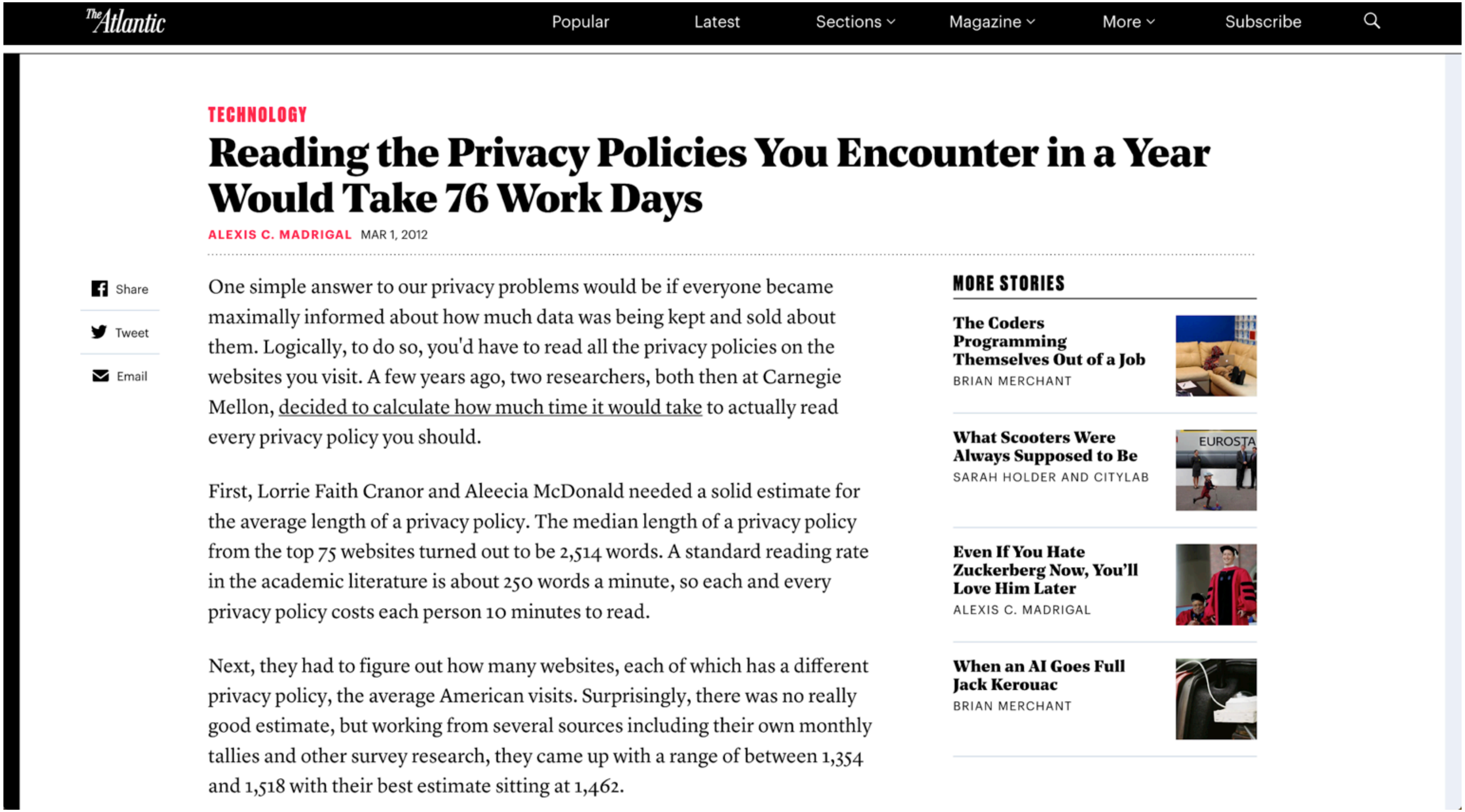
# Demystify the Datasphere.

Learn how data shapes what ads and offers you see.

See how marketing data helps companies serve you better.

Discover how companies show you offers for things you care about.

Find out how personalized marketing supports free digital services.



TECHNOLOGY

# Reading the Privacy Policies You Encounter in a Year Would Take 76 Work Days

ALEXIS C. MADRIGAL MAR 1, 2012

One simple answer to our privacy problems would be if everyone became maximally informed about how much data was being kept and sold about them. Logically, to do so, you'd have to read all the privacy policies on the websites you visit. A few years ago, two researchers, both then at Carnegie Mellon, decided to calculate how much time it would take to actually read every privacy policy you should.

First, Lorrie Faith Cranor and Aleecia McDonald needed a solid estimate for the average length of a privacy policy. The median length of a privacy policy from the top 75 websites turned out to be 2,514 words. A standard reading rate in the academic literature is about 250 words a minute, so each and every privacy policy costs each person 10 minutes to read.

Next, they had to figure out how many websites, each of which has a different privacy policy, the average American visits. Surprisingly, there was no really good estimate, but working from several sources including their own monthly tallies and other survey research, they came up with a range of between 1,354 and 1,518 with their best estimate sitting at 1,462.

MORE STORIES

**The Coders Programming Themselves Out of a Job**
BRIAN MERCHANT

**What Scooters Were Always Supposed to Be**
SARAH HOLDER AND CITYLAB

**Even If You Hate Zuckerberg Now, You'll Love Him Later**
ALEXIS C. MADRIGAL

**When an AI Goes Full Jack Kerouac**
BRIAN MERCHANT

# Many ideas out there…

*Monetize personal data*

- **Meeco** - personal life management platform; users can exchange data for special offers from brands
- **CitizenMe** - information management service; users can sell/donate data (to academic researchers)
- **Datacoup** - data exchange platform; users decide who buys the data, in exchange for small fees

*Make data available for research*

- **Midata.coop** - nonprofit around health-related data; users donate their medical data for research, in exchange for data analysis and interpretation tools offered by the platform
- **The common data project** - nonprofit data trust managed by a community of volunteers, nonprofit; users donate their data

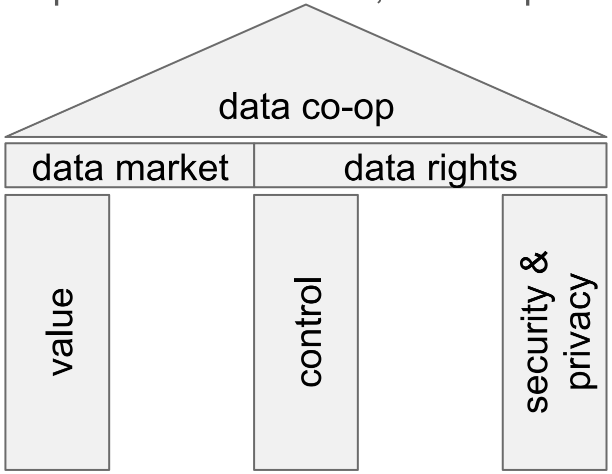
*Restore individual control*

- **Hub of All Things** - initiative of 6 UK universities; personal data micro-server and database, wrapped with microservices that protect user data
- **Enigma** - decentralized and secure data infrastructure
- **TheDataUnion** - aim to form a union that can negotiate with the big tech companies
- **Data as Labor movement** - advocate for protection and compensation for personal data as labor

# Opportunity to formalize and organize the problem

# Three pillars of a solution

- Creation of value
  - Transform personal data into new products traded in a data market, maintaining info validity
- Meaningful individual control
  - Transparent, varied choices with options for delegated decision-making
- Security and privacy
  - State-of-the-art data protection solutions, with emphasis on provable guarantees



# What are we trying to do?

Catalyze a discussion about alternative paths, force a reckoning with the path society is currently on

Build a research community around data co-op idea, understand the research challenges

- Product: tools that should serve many different sorts of initiatives that share commonalities with co-ops

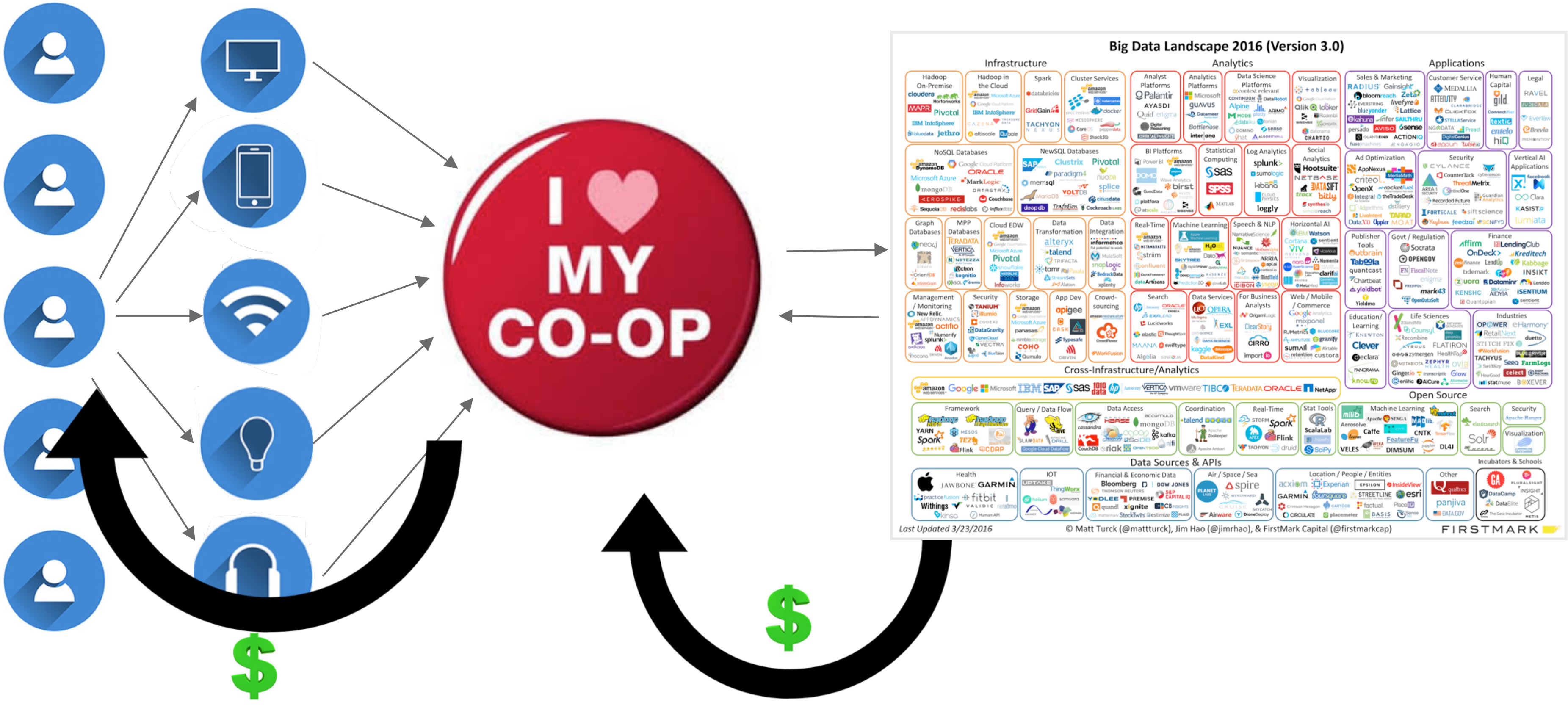
Understand the principles and ground rules

We are not committing to a single model

# This talk

- **How might a co-op provide value, control, and security?**
- What might this future look like?
- The multidisciplinary research agenda

# One vision of a data co-op



# Co-ops could create value by...

- creating new data products
  - new combinations of data sources/individuals’ data
  - access to currently unavailable aggregate statistics and sanitized datasets
  - interaction with targeted audience
- incentivizing the capture of more and better data
- taking the pain out of working with personal data
  - providing data science tools to facilitate discovery and ensure statistical validity
  - ensuring compliance with data protection regulations
- eliminating market frictions and inefficiencies

# Co-ops could enable control by...

- supporting delegated decision-making
- providing individuals with a variety of meaningful, understandable choices to control how data are used and by whom
- allowing choices to be changed or revoked
- providing an overview of data uses and risks

# Co-ops could support security and privacy by...

- giving individuals access to meaningful trade-offs between data use and risks
- modeling data usage in a concrete and rigorous manner
- implementing state-of-the-art data protection technologies
- vigilantly protecting their members against data theft, auditing for unauthorized use, advocating for their members' rights, and pursuing legal action when necessary

# Co-op non-negotiables

**Value**

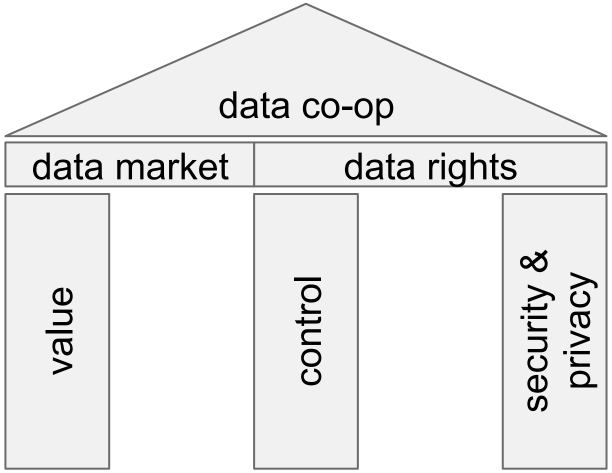
- Market-based individual compensation for both risks and value in data
- Align incentives of the co-op with the incentives of the members
- Ability to ensure high-quality data products (bots, fraud, statistical validity)

**Control**

- Delegation
- Expressive controls
- Usability by non-experts
- Right to withdraw or change permissions

**Security & Privacy**

- State-of-the-art protection against data theft
- Mathematically rigorous data rights guarantees in all use cases where possible



# Some concrete questions

What types of data would be held by a co-op? Medical? Financial? Geo-location? Genetic? Browsing history? Purchase history?

What would happen to my data that’s currently held by companies? Hospitals? Other organizations?

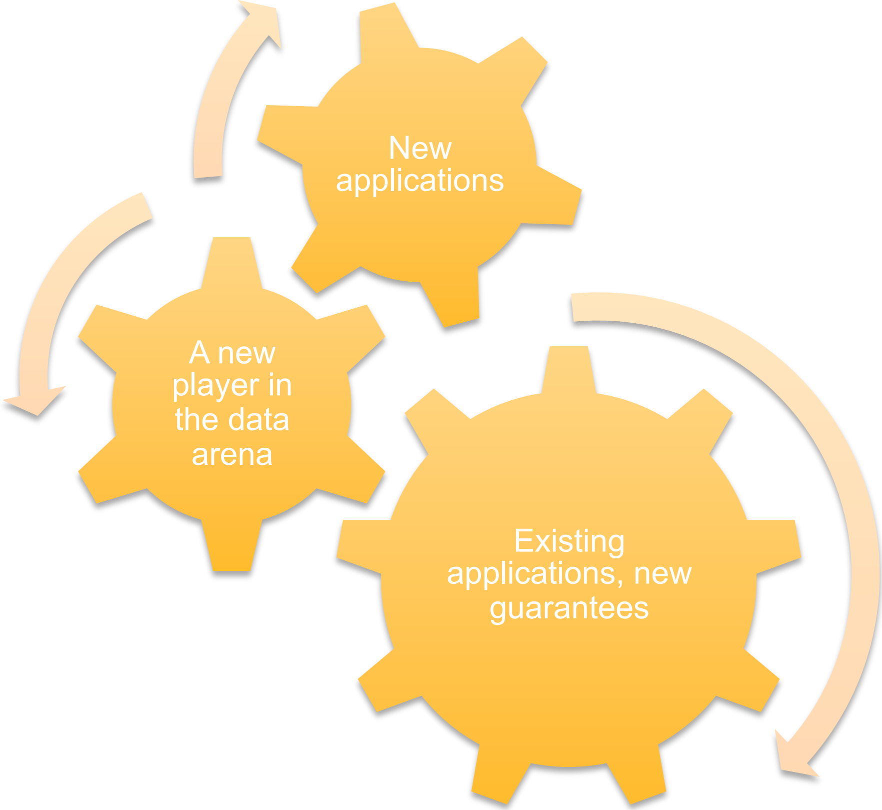
What would happen to “free” services that I currently receive in exchange for my data?

Would companies be buying my data (e.g., my entire browsing history) from the co-op outright?

# This talk

- How might a co-op provide value, control, and security?
- **What might this future look like?**
- The multidisciplinary research agenda

# What might this future look like?



New applications
A new player in the data arena
Existing applications, new guarantees

# Rethinking existing applications

- Modernization gradual. Maybe more legal protections first, technical tools as they mature (e.g., secure multi-party computation)

- What of value is currently being exchanged? (Unshackle from pay-with-data.)

# Rethinking existing applications

- Is transmission of personal data needed?
    - e.g., Email. Encryption for contents, mixing network for hiding source-destination pairs. Stronger guarantees for data that stays within co-op. Pay for usage or flat-fee models.

# Rethinking existing applications

- Is transmission of personal data needed?
  - e.g., Email. Encryption for contents, mixing network for hiding source-destination pairs. Stronger guarantees for data that stays within co-op. Pay for usage or flat-fee models.
- Does the service need to learn from aggregate user data?
  - e.g., Recommendation systems, navigation software. Learning the recommendation/route could happen “inside” the co-op. Co-op could act as proxy for delivery of goods and payments. Recommendations would be much better! Strict controls on privacy risk from tailoring of recommendations.

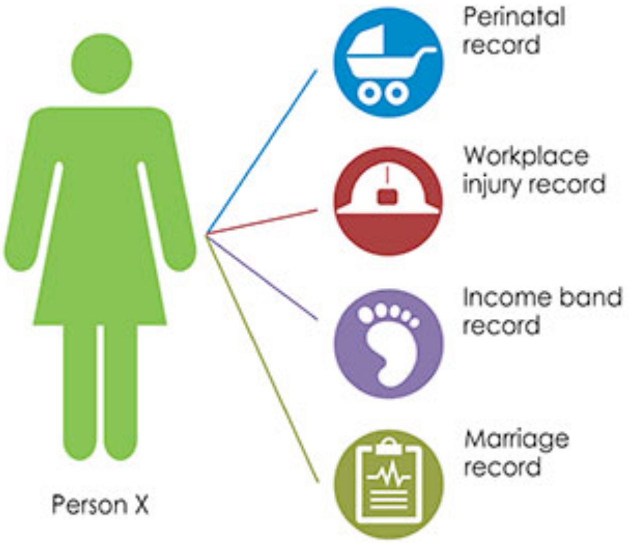
.

# Rethinking existing applications

- Is transmission of personal data needed?
    - e.g., Email. Encryption for contents, mixing network for hiding source-destination pairs. Stronger guarantees for data that stays within co-op. Pay for usage or flat-fee models.
- Does the service need to learn from aggregate user data?
    - e.g., Recommendation systems, navigation software. Learning the recommendation/route could happen “inside” the co-op. Co-op could act as proxy for delivery of goods and payments. Recommendations would be much better! Strict controls on privacy risk from tailoring of recommendations.
- Does the service involve extraordinary computational requirements?
    - e.g., Web search. Strip queries of unnecessary identifiers, bundle them, act as proxy for payment. Contractual limits on how information in queries is used.

# Enabling innovation

- Nano-targeting ads without invading privacy
- All your data needs, under one roof
- Secondary uses
- Data linkage, made easy
- The niche application



# Potential roles of the co-op

- Consumer protection
- Protecting political discourse in democratic society
- Letting consumers vote with their feet (choice, interoperability, no data lock-in)
- Providing protections currently unavailable (e.g., informed consent)
- Setting a higher standard for algorithms that use data
- Promoting priorities of the members (e.g., encourage small businesses)
- Promoting public interest (e.g., academic research)

# Pitfalls and risks

- Fat target for data theft, disruption of society, harming individuals
- Subpoena risk
- Great harm could come from not acting/negotiating/advocating for best interests of members, obeying constraints/preferences on data use, reporting clearly and truthfully about uses
- Must respect confidentiality for data purchasers, provide high-quality data
- Could be “taken over” legally or technically
- Could evolve in a bad direction, or programmed directives could have unintended consequences
- Need for control of incentives, oversight, legal constraints, technical constraints

# This talk

- How might a co-op provide value, control, and security?
- What might this future look like?
- **The multidisciplinary research agenda**

# Ground rules

Focused on understanding. Mathematical rigor where possible. Scrutiny by the community.

Timely sharing of results within the community. Code made public. Collectively developing a good for the public.

Openness about funding, competing interests.

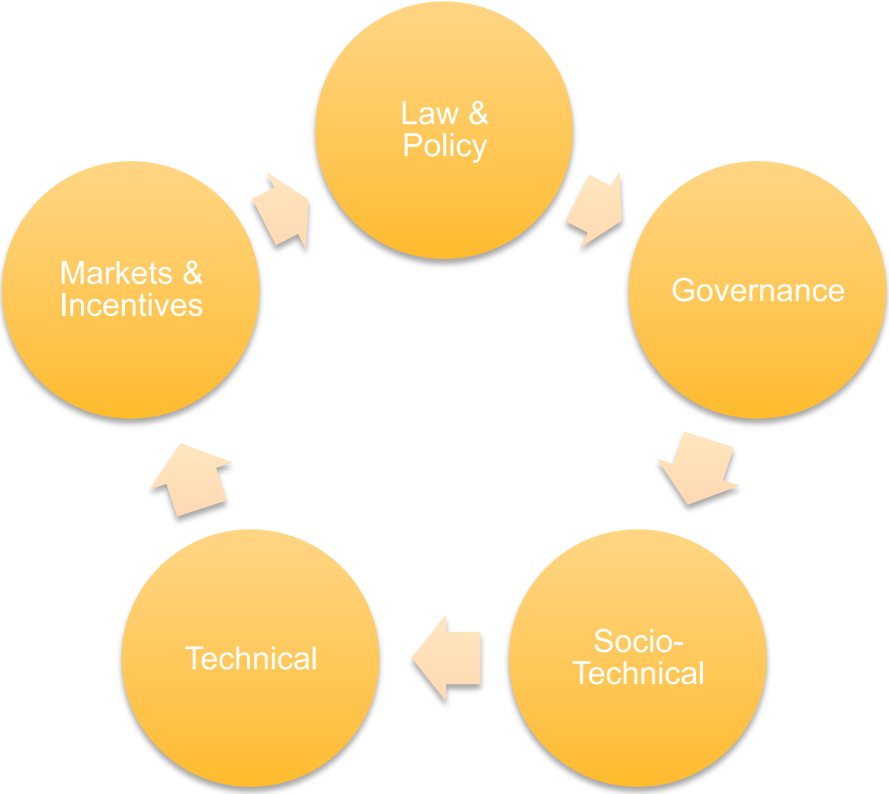
The community should have a careful eye on the research agenda, and on the overall incentives that drive the project.

Open design: Research should support a large variety of implementations of concepts related to data co-ops and not settle on a particular architecture.

# Creating a toolkit, not a single solution

- Value in diversity among co-op models, friendly competition among models
  - more and better co-op design ideas
  - wider set of participants in the research and development
  - more data and more types will be made available, supporting innovation
- Value in co-existence of multiple co-ops
  - avoid creating data monopoly
  - ensure reasonable pricing and access
  - pressure co-ops to serve their constituents, meet their obligations
  - different communities may have different priorities, values

# Relevant research areas



# Law and policy: research agenda

How do current legal and policy frameworks around the world fit with the co-op idea? What needs to be changed? Mechanisms for change?

Current practices may infringe individual rights w.r.t. privacy, data, labor (data as labor). Who owns data?

Where are the relevant rights enshrined? Some data rights have not yet been enshrined.

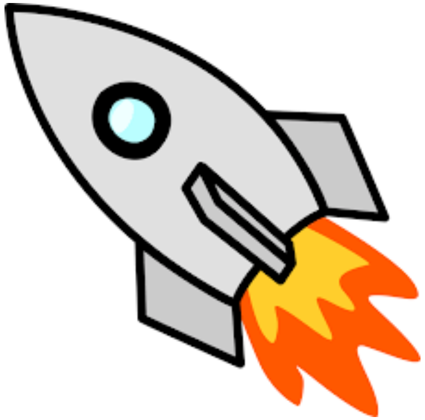
# Regulation and governance: research agenda

What type(s) of legal entity could a co-op be? Can legal aspects of the co-op structure enforce and reinforce its goals? (“Information fiduciaries”)

Not self-regulation, not government-regulation. (But also an opportunity for regulation.)

Success of the coop will rely on trust of and protections for the information-buyers. Technical or legal safeguards to protect data users and analysts from the coop.

# Socio-technical: research agenda



Launch: How to ramp up to large enough user base?

Who will oppose this kind of solution? Who would be in favor?

Could a co-op emerge from existing organizations?

Safeguards against drift and hard fork of behavior/intentions of project. Mission creep. Subpoena risk.

Creating a language/framework for discussing the different (computational, privacy, economic, legal, ethical, social) aspects of the system in a way that is consistent across disciplines.

# Technical, system design: research agenda

HCI: give a comprehensive understanding of how data is used, tools for reasoning about choices, language for individuals to express their values and preferences [people don’t understand Randomized Response! c.f. Frauke’s talk]

Provide a global view of the information landscape: what information is out there? what uses?

Data protection: we are creating a new attack surface! Merging state-of-the-art privacy, crypto, security

System design: centralized vs. decentralized data storage, computation

Formal verification: how to ensure the system operates and evolves as intended

# Incentives, mechanism design: research agenda

How to set/control incentives of the co-op organization?

Managing privacy budgets

Mission creep

How to detect and protect against data being used to manipulate (e.g., children, Cambridge Analytica)

Much more tomorrow!

Cambridge
Analytica

# Our goals

- Catalyze a research community that will provide answers, guidance, toolkit to existing/future efforts in this vein
- Explore paths forward, both in technical design decisions and practical launch decisions
- Raise awareness of potential pitfalls, failure modes, mission creep
- Maximize chances of successes

# Wanted: your engagement

- Thoughts about our documents
- Pointers to people and literature
- Participation in fora
- Your research efforts

# Value in data: research agenda

Co-ops should increase welfare

Economic growth: universalized data access, greater competition, interoperability

Data products have value greater than the sum of their parts; more and better data should be captured

Insurance for risk vs. compensation for value; how to allocate the surplus

Statistical validity: preserving data quality by mitigating selection bias, overfitting

Data resale, re-purposing: how to price it? how to control it?

# Mechanism design for information: research agenda

What are the goods exchanged?

How to price them, how to split revenues fairly?

How to incentivize participation on both sides of the market?

Multiple co-ops? Competing?