DATA SOCIETY®

Introduction to Data science

What can data science do for you?

TABLE OF CONTENTS

Data driven stages assessment

Instructions: circle the answer that best applies to you and your organization. Then, score yourself by tallying up the points you receive for each answer

a = 0 b = 1 c = 2 d = 3

Component: Data infrastructure		
		Points
Data access	1. I can easily access the data I need without asking others for help a) Not at all b) Only for some colleagues c) Only for some teams d) Organization-wide	
Data access	2. I can easily access the data I need in a timely manner a) Not at all b) Only for some data c) Only for data in my team / related teams d) Organization-wide	
Data collection	3. Data is automatically collected and stored on a continuous basis a) Not at all b) Only at someone's request c) Regularly, a few times a year d) There is continuous data collection	
Data collection	4. The data we have is accurate and good quality (few missing entries, few duplicates, accurate measurements) a) Not at all b) Only for some data c) Only for data in my team / related teams d) Organization-wide	
Data storage	5. Our data is stored securely either internally or offsite a) Not at all b) Only for some data c) Only for data in my team / related teams d) Organization-wide	

Total: _____

Data driven stages assessment, cont'd

Instructions: circle the answer that best applies to you and your organization. Then, score yourself by tallying up the points you receive for each answer

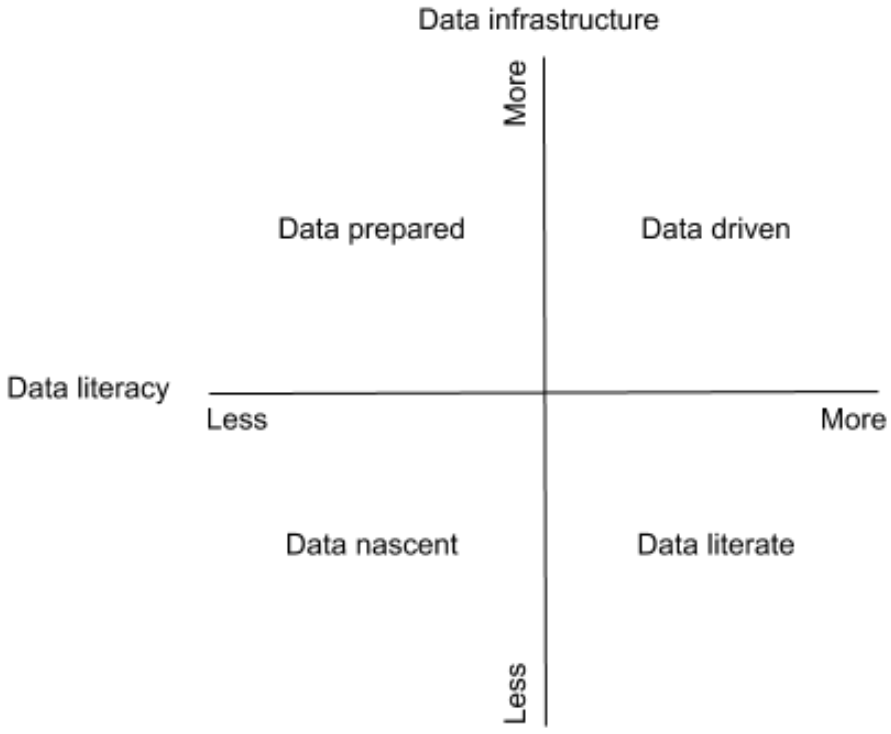
a = 0 b = 1 c = 2 d = 3

Component: Data literacy		
		Points
Data literacy	1. My company routinely offers data trainings and other educational opportunities a) Not at all b) Occasionally c) Regularly, a few times a year d) There are continuous learning opportunities	
Data literacy	2. Most of my colleagues understand the importance of data a) Not at all b) Only for some colleagues c) Only for some teams d) Organization-wide	
Data governance	3. Our organization has a set of data standards that reviews how data should be collected, stored, and analyzed a) Not at all b) Only for some colleagues c) Only for some teams d) Company-wide	
Data leadership	4. My organization emphasizes the importance of using data to track initiatives a) No one b) A few people across the company c) Some teams across the company d) Organization-wide	
Data leadership	5. I am expected to present data metrics when I explain conclusions and decisions a) Not at all b) Only for some colleagues c) Only for some teams d) Organization-wide	

Total: _____

Data driven stages: results

Circle or shade the quadrant that matches your score. If you scored below 8 points on either the data infrastructure or data literacy sections, you are on the left-hand side or bottom half. If you scored above, you are on the right-hand side or upper half of the graph.



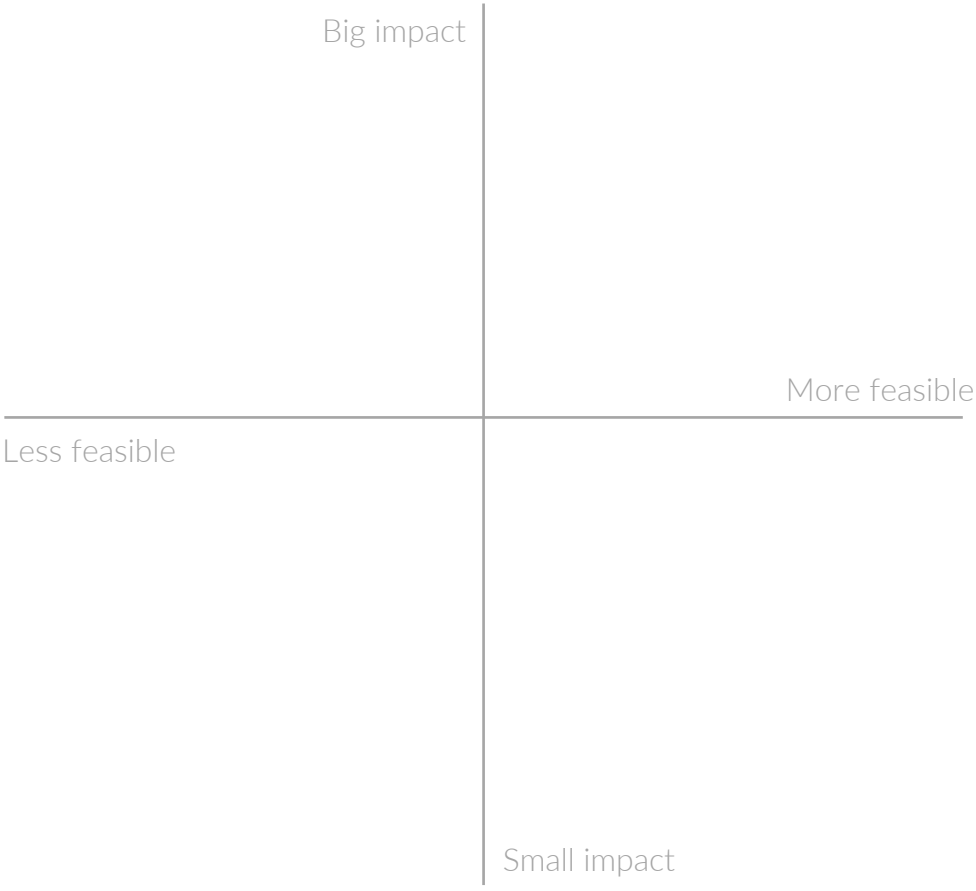
1. Which quadrant is your organization in?
2. What is the strongest area that your team / organization has?
3. Which area do you want to focus on first?

Project brainstorm

Take the best ideas that you and your group came up with and write them in the appropriate quadrants below.

Data project ideas

How can you use data to solve problems?



Additional notes:

Re-form your questions

Take 3 – 5 ideas from the previous and rephrase them into objectives that are **Specific**, **Measurable**, and **Actionable.**

Objectives

Initial problem	Re-phrased objectives
Initial problem	Re-phrased objectives
Initial problem	Re-phrased objectives

Additional notes:

Analyzing visualizations

Look at the following visualizations – do they follow the 10 principles? Write down your thoughts, then discuss with your group

Chart 1:

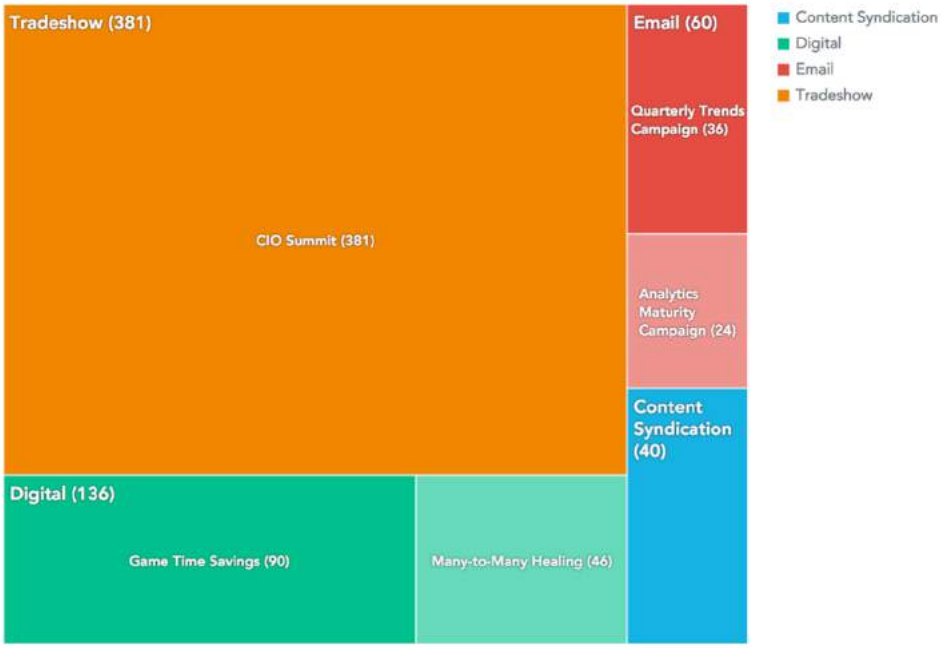
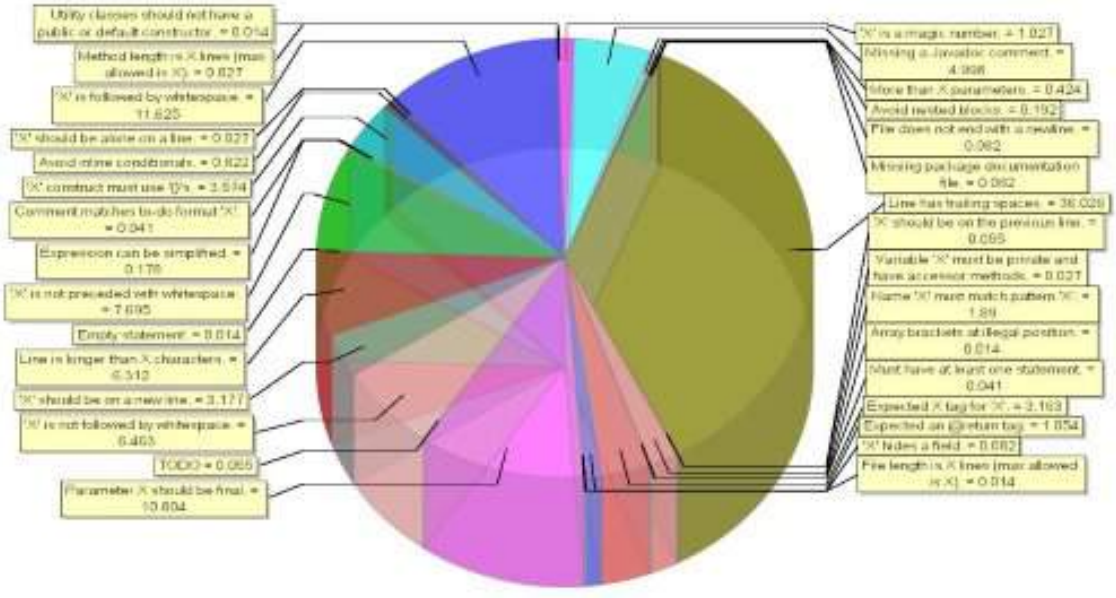


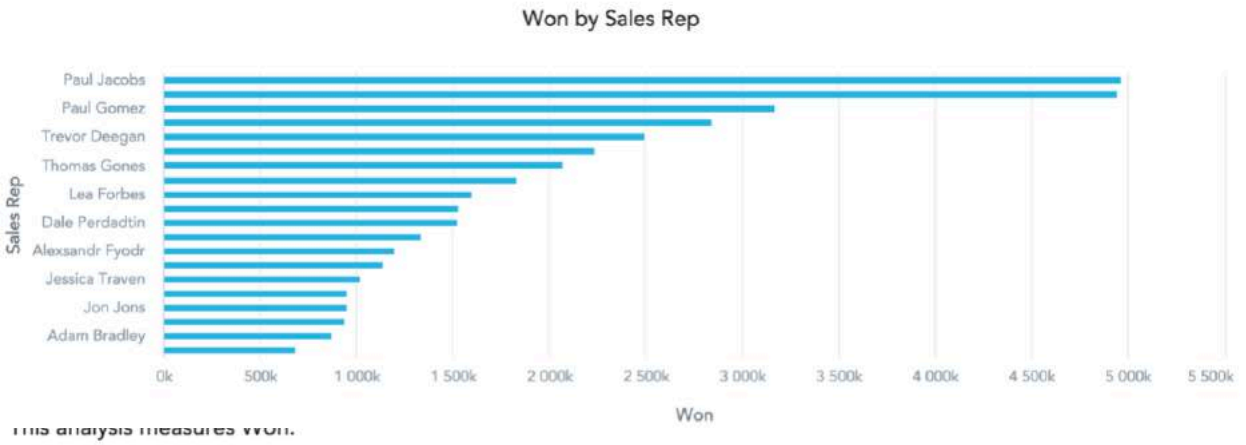
Chart 2:



Analyzing visualizations

Look at the following visualizations – do they follow the 10 principles? Write down your thoughts, then discuss with your group

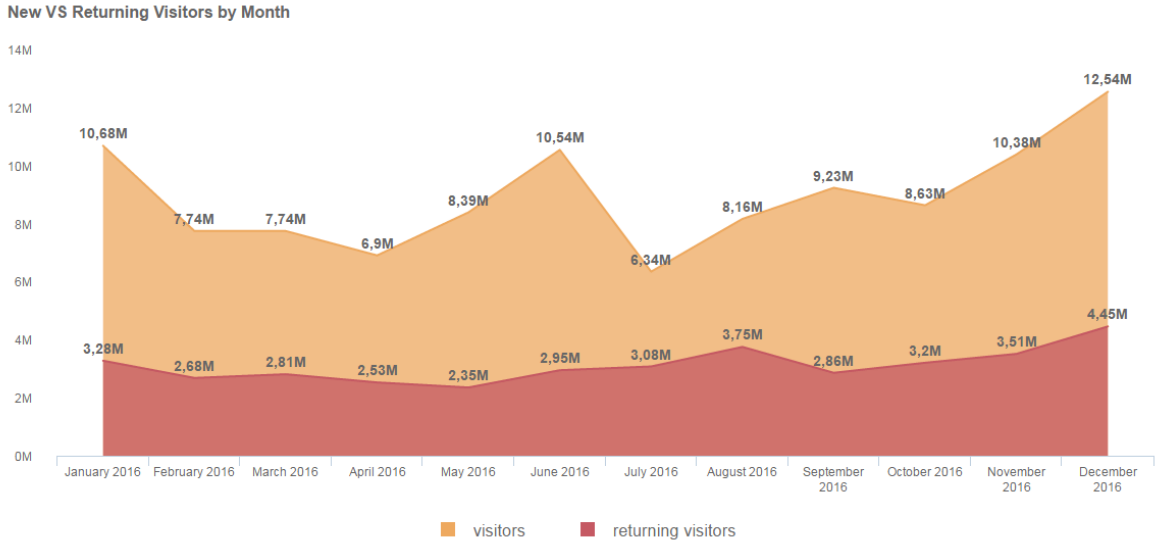
Chart 3:



This analysis measures Won.

- Overall Won is 38,310,753.45 across all twenty sales rep.
- The distribution ranges from 684,740.93 (Dave Bostadt) to 4,968,600.86 (Paul Jacobs), a difference of 4,283,859.93.
- The top three sales rep (Paul Jacobs, Ravi Deetri and Paul Gomez) represent 34% of overall Won.
- The lowest three sales rep (Dave Bostadt, Adam Bradley and Cory Owens) represent 7% of overall Won.

Chart 4:



Additional notes:

Evaluating ethics

Read over the following article about building predictions that can predict terrorism – discuss the following questions with your group.

1. What are the ethical implications to this algorithm?

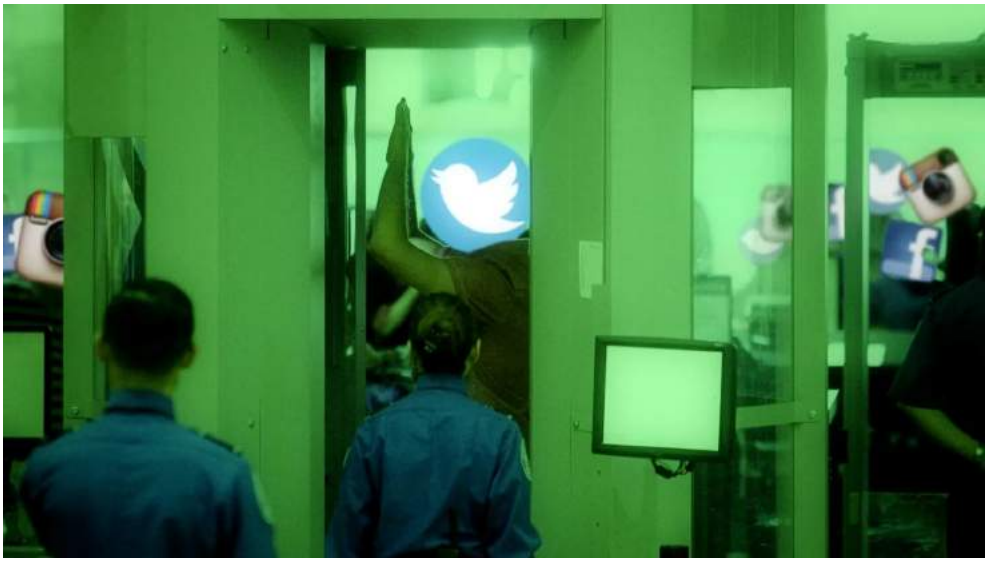
2. What are some safeguards that you would put in place to avoid ethical mishaps?

3. Are there other situations that you would want to apply the algorithm to?

Additional notes:

The government wants Silicon Valley to build terrorist-spotting algorithms. But is it possible?

Kashmir Hill
1/14/16 1:31pm



Aaron Cornette/FUSION

Last week, a bunch of important people from Washington, D.C. packed their bags and flew to California to meet with a bunch of important people from Silicon Valley. The occasion was not the usual fundraising dinners or donor-wooing — it was a terrorism summit.

Government types are freaked out about the role of technology in how groups like ISIS recruit members and plan attacks. They think the heads of tech companies like Facebook, Apple, Twitter, and Google can do more to help them keep the world safe. And so counter-terrorism officials got tech executives to spend a day with them in San Jose last Friday. Among the topics

on the agenda were consumers' access to encrypted communications that aren't easily intercepted by the government (a horse that's so beaten to death that it's been zombified) and a new idea posited by the policymakers: some kind of technological system that could detect, measure, and flag "radicalization."

A terrorist-hunting algorithm isn't a completely off-the-wall idea. Financial companies have proposed scanning Facebook postings to help determine people's credit worthiness. There are already products for police departments that dole out "threat scores" to individuals based on scanning public social media activity and looking for key words; one police department's use of a "beware algorithm" was recently revealed by the ACLU of Northern California.

MEDIA SONAR

Keywords List

The following key words can be extremely effective in pro-active policing. If you have any improvements, corrections or additions, please let us know. *Some of the words below are shown in groups for simplicity, but may not be directly related. Those we've had particular success with have been highlighted. Our thanks to Arman Refugee Asylum group for the new human trafficking words.*

WEAPONS
2 piece
40, 40cal
45cal
380
357
9mm
9piece
nine, nina
Mac 10, mac10, mack10
Ak47
AR15
ARpistol
Tech9, tek9, tek 9
12gauge
KG9
m4
260rem
SR9
44
Glock
Pistol
Rifle
Shotgun
Bullet(s)
Carbine
Clap
Luger
Magnum
Remington
Ruger
Semiauto
Semi
smallarms
smithandwesson
sniper
piece
strapped up

GANGS
Gang
Cartel
gangsta
playa(z)
Boyz
da boyz
da streets(z) fight
beat
hot boy
jumped
knock out
lick(s)
makin moves
meeting
mpd(k)
on da 4 (or 5)
peelie
pillie
gangsta
thug
Snitch
RIP
cutthroat
busta cap
Bangin
beatdown
beefin
boyz
bussin
bust
bustin
CEO
choppa
chopper

NARCOTICS
420
smoke
green
goldcaps
hash
herb
hightimes
indica
jetlife
marijuana
mary jane
oil
pipe
plant
pot
purple
sativa
shatter
strain
toke
wakeandbake
wax
weedporn
weedstagram
mac
buda
shwag
mota
dodo
green
fire
water
bluntos
dro
sticky
Acid
CID
Ecstasy
Molly
Shrooms
SpecialK
Vicodin
Bake(d)
Stoner
Stoned
High
Blow
Dope
Smack
Coke
Cocaine
Cheese
Crack
Crystal
Ice
Yay
Yayo
Dealer
Money
Growop
Dime
Dimebag
dimesack
sawbuck
nickel
nickelbag
nickelsack
elbow
dub
quarter
eightball

POLICE EVASION & CRIMES AGAINST POLICE
Cops
Feds
Popo

The words that Media Sonar, a product used by Fresno's police department, uses to identify threats

SELF-HARM, CYBERBULLYING, SUICIDE
Abuse
Alone
Bully
Bullied
Bullying
Crying
Depressed
Selfhate
Selfharm
Self-injury
Die
emo
Forgotten
Hate
Help
Hurt
Kill
Killme
Lonely
misunderstood
Pain
Suffer
Ugly
Unheard
Unloved
Unseen
Unworthy
Anorexia
Ana
Starving
Bulimia
Mia
Binge
Purge
Starve
Blades
Cutting
Cut
Cutter
deepcuts
Burn
Overdose
Pill
scars
schiz
shame
Suicide
Suicidal
scabs
triggerwarning
trigger
tw
cat
Secret Society
secretsociety123

YOUTH / SEXTING / CHILD LURING
IWSN (I want sex now)
CU46 (see you for sex)
53X (sex)
1174 (party meeting place)
THOT (that hoe over there)
Sugarpic (erotic photo)
(L)MIRL (lets meet in real life)
PRON (porn)
8 (oral sex)
IPN (Im posting naked)
LH6 (lets have sex)
GYPO (get your pants off)
PIR (parent in room)
9 (parent watching)
99 (parent gone)
POS (parent over shoulder)
CD9 (parent around, code 9)
KPC (keeping parents clueless)

PROPERTY CRIMES
Borrow
Gank
Cheat
Free
Grab
Jack
Lift
Nick
Pilfer
Pinch
Rob
Snatch
Steal
stole(n)
theft
thief

HUMAN TRAFFICKING
(new and relatively untested)
Baller
bitch(es)
bottom bitc
brother
bunny ranch
cunt
escort
fuck4cash
hoes
pussy
skank
sleeper leaper
slut
sprankles
tramp
titty
trick
Lolita
wife in law
Hooker
Hookerbait
hooker baiting
hooker barbie
hooker binge
hooker booker
hookerrated
hookerbeast
hooker bopping
hooker boppers
hookerbelle
hookerbot
Boss
Bosslife
John
King
king pimp
king pin
lot lizard
pimp(s)
pimp hand
pimpin
pimping
pimpinmaster
Mattress maiden
money whore
pavement princess
porn
porn acting
porn addict(ion)
porn again
porn star
pornabe
pornacopia
pornagize
pornographer
porno
prostitot
prostitott
prostitoy
prostitute
webcam

MIKE BROWN RELATED
Mikebrown
justiceformike
ripmikebrown
mikebrowncampaign
iammikebrown
Nojusticenopeace
Blacklivesmatter
Weorganize
Riot
boycott
rebel
Handsup
imunarmed
Beheard
itstimeforchange
Dayofrage
Wewantanswers
wewantjustice

Via ACLU FOIA

But this proposal that tech companies might give their own users "radicalism scores" is more novel, and comes on the heels of the San Bernardino shootings, after which law enforcement discovered that one of the shooters had posted to Facebook advocating jihad.

A White House memo that went out to summit participants before Friday's briefing acknowledged that such a system would raise privacy and civil liberties concerns, and that it's "unclear" whether radicalization is as easily

measurable as credit scores. But the memo said that "such a measurement would be extremely useful to help shape and target counter-messaging and efforts focused on countering violent extremism." The Guardian reported that it was compared during the meeting to Facebook's attempts to prevent users from committing suicide:

> The social network's chief operating officer, Sheryl Sandberg, walked government officials through how Facebook currently enables users to flag people who appear to be posting suicidal thoughts, a person familiar with the conversation said. The government officials in the room wondered if such a system could be used to flag terrorist content or detect a user who appears to be radicalizing, added the person, declining to be quoted on the record.

Facebook suicide prevention system is the digital equivalent of "See something, say something"—except it asks users to "See something, flag something." (A more appropriate comparison might be Facebook's automated scanning of Facebook activity to bust sexual predators.) But is flagging radical thoughts as easy as suicidal ones?

"It's tricky to measure radicalism when someone hasn't committed a crime," said Gary LaFree, director of University of Maryland's academic center START, which studies terrorism. "When you dive into that, it gets very controversial."

START built a database of 1,500 radicalized individuals—people arrested, killed or convicted while in pursuit of extremist ideologies—to try to figure out how they became far-left, far-right, or Islamist extremists. LaFree said when START first started looking at radicalization 12 years ago, researchers didn't have a good understanding of how radicalism happens. They thought it was a gradual process, in which someone steadily becomes more and more radicalized, and then decides to act. They thought there were strong ties between radical thoughts and radical action. But now they realize that's not always the case, he says.

"Sometimes very radical thought doesn't lead to action," he said. "But then someone else will be loosely connected to a radical group, played soccer with someone for example, and then they're willing to do something very radical. We understand the different pathways to radicalization better now, but it makes it even more complicated."

LaFree says the science of radicalization predication is still developing, and so far, hasn't really involved social media analysis. Most radicalization measurement focuses instead on people's psychological and sociological backgrounds. Lone terror actors, for example, says LaFree, are more likely to have criminal activity in their pasts, psychological issues and military training.

"We're learning more about the background things that predict radicalization, more than we knew 10 years ago. But we're not at Minority Report yet," LaFree said.

An algorithm designed to spot radicalized individuals, at this point, would generate a considerable number of false positives. There's also the question of whether the tech giants the White House met with—Google, Facebook, Dropbox, LinkedIn, Twitter, Apple and Cloudflare—could actually *build* what policymakers want. As vaunted as "big data" is, it still struggles with sentiment, image and word analysis. Facebook and LinkedIn try to force us to friend our exes. Google's artificial intelligence for photos labeled black people as gorillas. Are these same companies up to the task of identifying potential terrorists?

Akli Adjaoute, the CEO and founder of Brighterion, a company that initially used artificial intelligence to detect credit card fraud, says his firm has developed a tool that will be rolled out in Europe next month for use in transportation and border security. Adjaoute, who has a PhD in artificial intelligence, says his firm developed a tool that works by doing contextual word analysis and relationship analysis, and tracks social media users over time, doling out alerts if they pass a certain level of radicalization.

"Instead of processing credit card transactions as fraudulent, we process every tweet and message as radical or not," he said. "Safety is number one for me. People talking about privacy don't understand the risks."

Automated pre-crime analysis like this has historically been controversial. And at this point, there's not clear evidence to suggest it works. But with terrorist groups like ISIS making greater use of social media, there's great pressure on the government to address their activity there.

"On the internet, the advantage is on the side of the terrorists," says LaFree. "They have hundreds of free workers. We have to pay [federal workers] to look at everything and are going to look like Big Brother if we do it."

Or, the government can try to get technology companies to do it instead, so that it's Facebook and Google doing the watching, instead of Big Brother.

Internal data science capacity

Now that you know the main types of roles on data science teams, it's time to evaluate your own staff to identify the data skills in your department. Fill out the sections below and then discuss with your group.

Data science needs

What data science skills do you use / need?

Current talent

Which of your team members work with data?

Current skills

What skills do they have?

Internal data science capacity

Skills gaps

What skills gaps did you identify?

Organizational chart

How do you envision the new organizational chart?

Once you complete your analysis, share your results with your group. What similarities do you have? What are some differences?

Additional notes:

Data governance assessment

The Basic Maturity Assessment is a condensed version of the Stanford Maturity Measurement Tool that model focuses both on foundational and project aspects of DG. The foundational components (Awareness, Formalization and Metadata) of the maturity model focus on measuring core DG competencies and development of critical program resources. The project components (Stewardship, Data Quality and Master Data) measure how effectively DG concepts are applied in the course of funded projects.

Additionally, it includes the three dimensions (People, Policies and Capabilities) which further subdivide each of the six maturity components, focusing on specific aspects of component maturation.

Whether your organization uses the Stanford Maturity Measurement Tool or the Basic Maturity Assessment, it is imperative that the maturity model you choose is finalized and adopted early in the rollout of the DG program. Depending on where your organization is in the process of standing up the data governance program, it may be most appropriate to use the Basic Maturity Assessment to measure the baseline maturity of and resources available to the organization. Then, as the data governance program is fleshed out, perhaps you will find that a more robust maturity assessment is needed. In that case, because they are both based on the same component-dimensions, you can easily transition from using the Basic Maturity Assessment to using the full Stanford Maturity Measurement Tool.

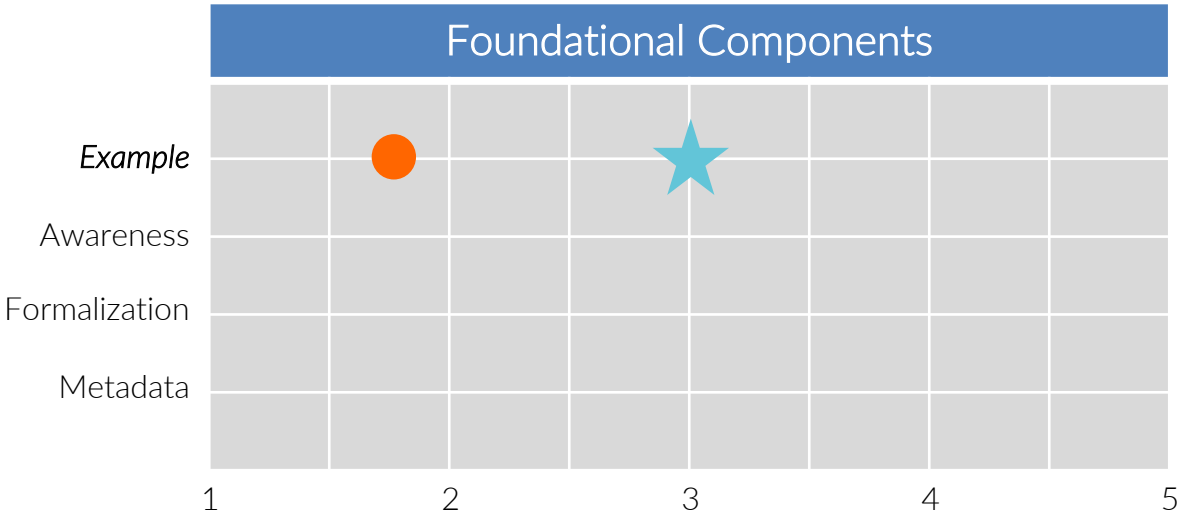
Regardless of which tool you choose to use, or if you choose to use a combination of both, thoughtful input from across the organization will help assure the model's usefulness and long-term fitness.

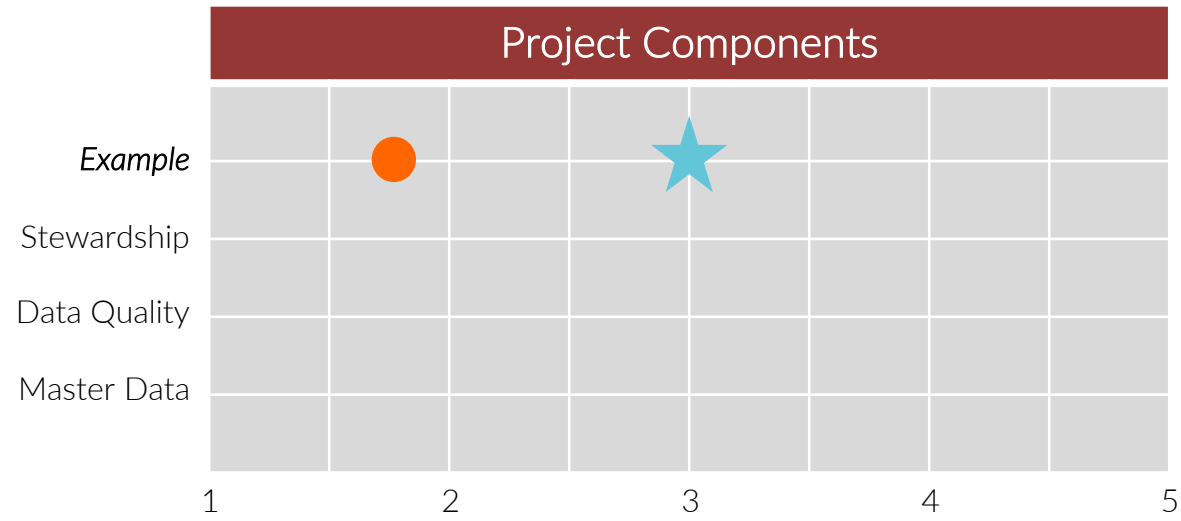
Data governance assessment

Data Governance Foundational Components Maturity		
Component: Awareness - The extent to which individuals within the organization have knowledge of the roles, rules and technologies associated with the data governance program.		
	Objective	Rating
People	Are executives, employees and stakeholders aware of the purpose or value of the DG program?	1 2 3 4 5
Policies	Are existing data policies documented, consistently maintained and available to stakeholders?	1 2 3 4 5
Capabilities	Are stakeholders aware of the specific DG capabilities that are available at the organization?	1 2 3 4 5
Component: Formalization - The extent to which roles are structured in an organization and the activities of the employees are governed by rules and procedures.		
	Objective	Rating
People	Have DG roles and responsibilities been defined and vetted with program sponsors?	1 2 3 4 5
Policies	Are data polices around the governance of specific data defined as best practices?	1 2 3 4 5
Capabilities	Are classes of DG capabilities defined and is there an available solution?	1 2 3 4 5
Component: Metadata - Technical metadata describes data elements and other IT assets as well as their use, representation, context and interrelations. Business metadata answers who, what, where, when, why and how for users of the data and other IT assets.		
	Objective	Rating
People	Do executives, employees or stakeholders have understanding of types and values of metadata?	1 2 3 4 5
Policies	Are metadata best practices produced and made available?	1 2 3 4 5
Capabilities	Is metadata consistently collected, consolidated and available from a single portal?	1 2 3 4 5
Data Governance Project Components Maturity		
Component: Stewardship - The formalization of accountability for the definition, usage and quality standards of specific data assets within a defined organizational scope.		
	Objective	Rating
People	Have DG or stewardship roles and responsibilities been defined within the organization?	1 2 3 4 5
Policies	Have policies around data stewardship been defined within a functional area?	1 2 3 4 5
Capabilities	Does a centralized location exist for consolidation of and/or access to stewardship related documentation?	1 2 3 4 5
Component: Data Quality - The continuous process for defining the parameters for specifying acceptable levels of data quality to meet business needs, and for ensuring that data quality meets these levels.		
	Objective	Rating
People	Are people assigned to assess and ensure data quality within the scope of each project?	1 2 3 4 5
Policies	Have data quality best practices been defined and adopted as official organizational data policies?	1 2 3 4 5
Capabilities	Have basic data profiling tools been made available for use anywhere in the system development lifecycle?	1 2 3 4 5
Component: Master Data - Business-critical data that is highly shared across the organization. Master data are often codified data, data describing the structure of the organization or key data entities.		
	Objective	Rating
People	Is there consistent understanding among stakeholders of the concepts and benefits of master data?	1 2 3 4 5
Policies	Are there formal policies that define what data are considered institutional master data?	1 2 3 4 5
Capabilities	Are master data identified, managed and provisioned?	1 2 3 4 5

Data governance assessment – evaluation

Average your scores for each section and put a circle / dot for where you are currently. Then, take some time to think about your goals by putting stars where you want your score to be in the future.





Discuss your results with your group – are there any patterns that you notice between your scores? If you're part of the same team / department, did you get similar results?

Additional notes:

Data science tools

Now that you have an overview of existing tools, as well as their purpose, what tools do you currently use? And who uses them for their analyses?

Data science tools

What data science tools do you have?

Data analysis	Data storage
Data visualization	Others

Current talent

Which of your team members use these tools?

Cost

How much are you spending on these tools?

Do you need all the tools you have? What tools would you be interested in using more? Discuss your findings with your group.

Additional notes:

Data event plan

Now that you have an overview of how to build awareness and support, fill out the template below to plan out your first event!

Data event plan

What objective do you want to achieve with your data event?

What data event would you want to plan?

Who would you involve in the planning and management of the event?

Who would you involve in the planning and management of the event?

What obstacles may come up as you plan and execute your event?

Data event plan

What is the name of your event?

Who would you invite to the event?

What would be the motivation for them to attend the data event?

Share your event details with your group - did you come up with similar events? How else do you think you can improve awareness about data and its uses?

Additional notes: