

DATA SCIENCE with SMALL TEAMS

Drew Fustin

PhD, Physics

Lead Data Scientist



S P O T
H E R O

drewfustin@gmail.com | 7.5.2016 | PyData Chicago

DATA SCIENCE

with

SMALL TEAMS

Or: Lots of Advice I Need to
Hear and Apply Myself



DATA SCIENCE

according to Jeremy Stanley, VP of Data Science at Instacart

DECISION SCIENCE

use data to analyze business metrics – such as growth, engagement, profitability drivers, and user feedback – to inform strategy and key business decisions.

DATA PRODUCTS

use data and engineering to improve product performance, typically in the form of better search results, recommendations, and automated decisions.

DATA SCIENCE

according to Jeremy Stanley, VP of Data Science at Instacart

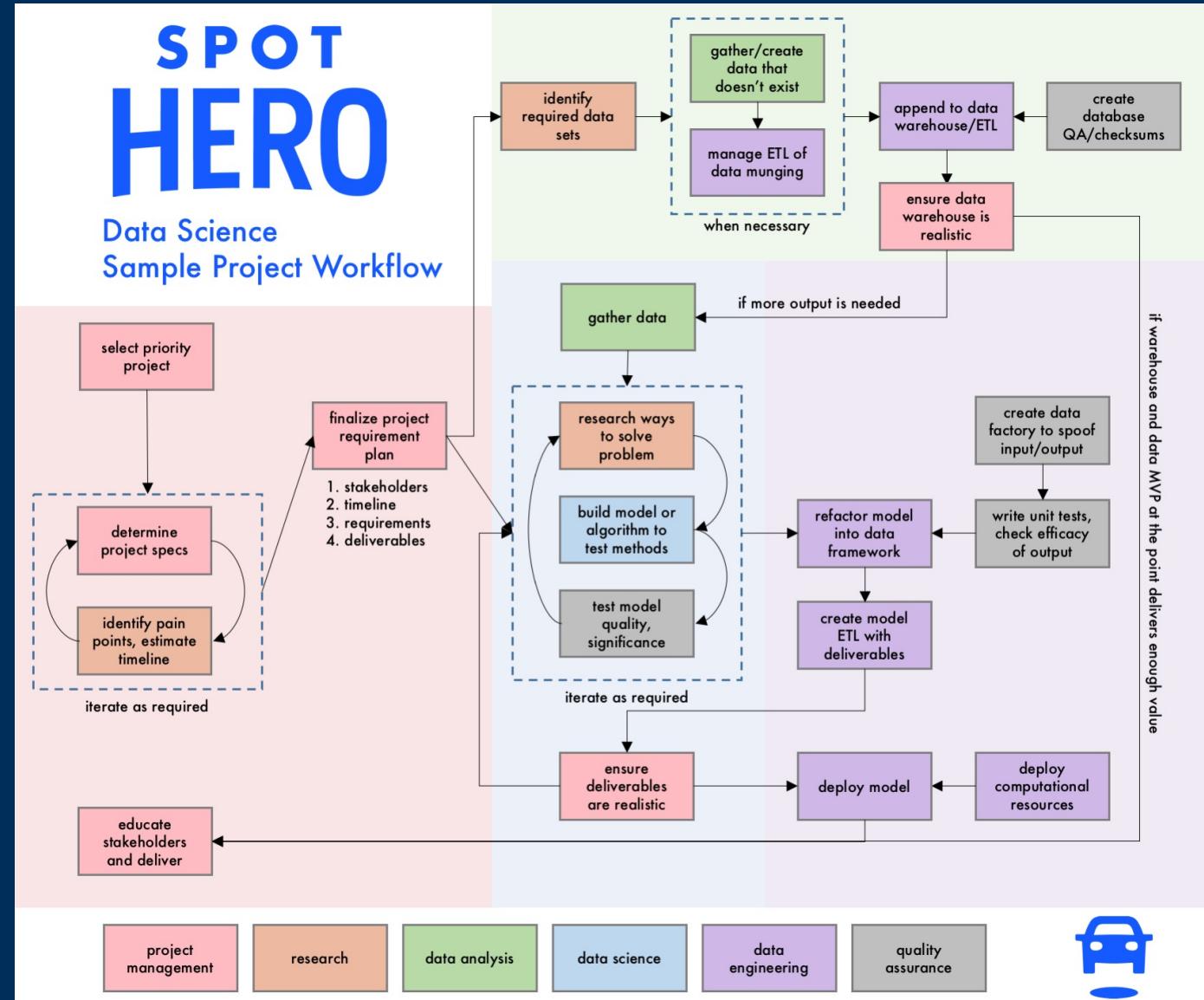
“While decision science and data products call for some of the same skills, it’s rare for data scientists to excel at both. Decision science depends on business and product sense, systems thinking, and strong communication skills. Data products require machine learning knowledge and production-level engineering skills. If you have a small data science team, you may need to find the rare superstars who can do both. But you’ll benefit from specialization as you scale your team.”

Doing Data Science Right – Your Most Common Questions Answered

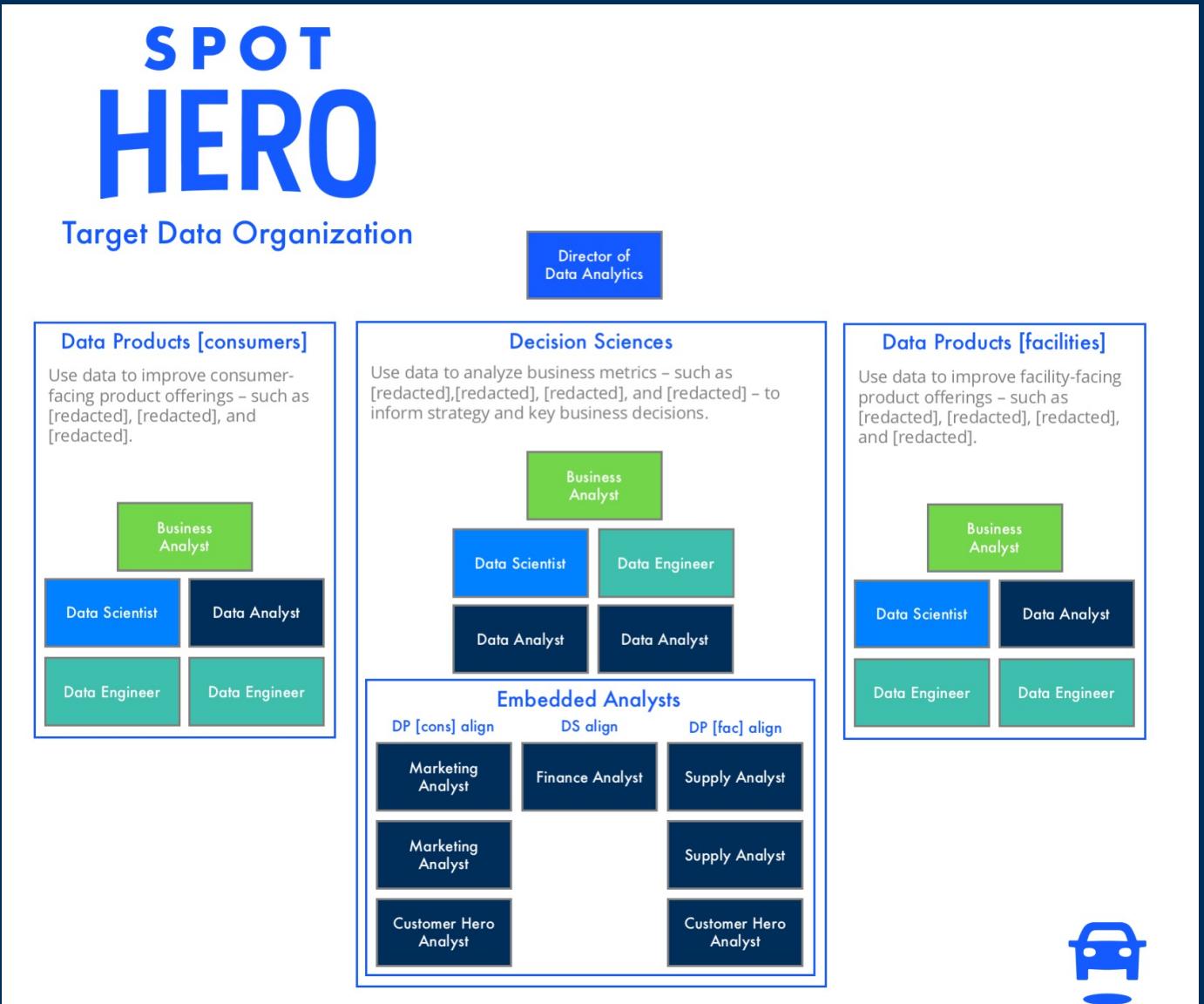
Jeremy Stanley and Daniel Tunkelang

<http://firstround.com/review/doing-data-science-right-your-most-common-questions-answered/>

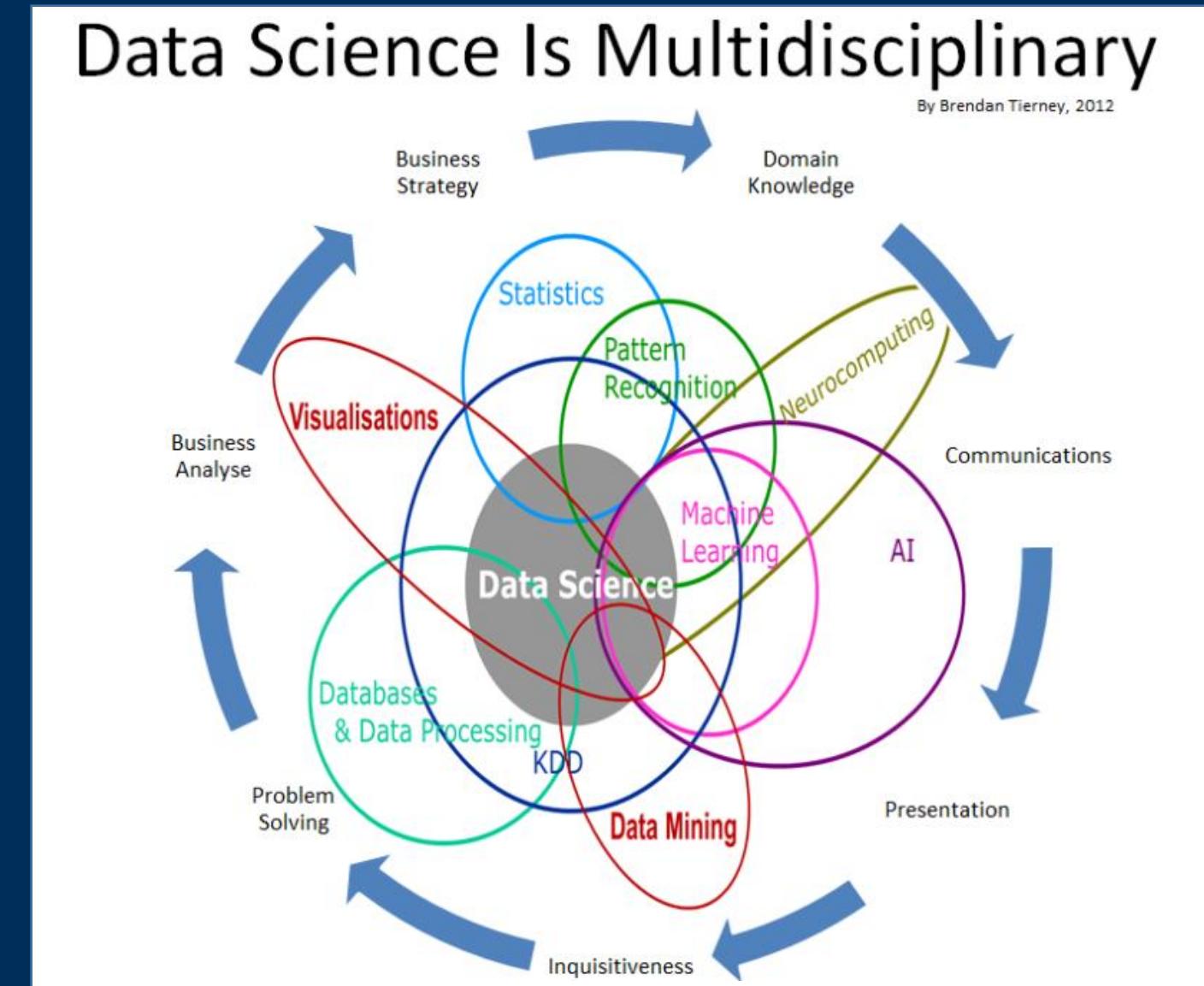
DATA SCIENCE IS COMPLICATED



DATA SCIENCE IS EASIEST WITH A TEAM



DATA SCIENCE IS EASIEST WITH A TEAM

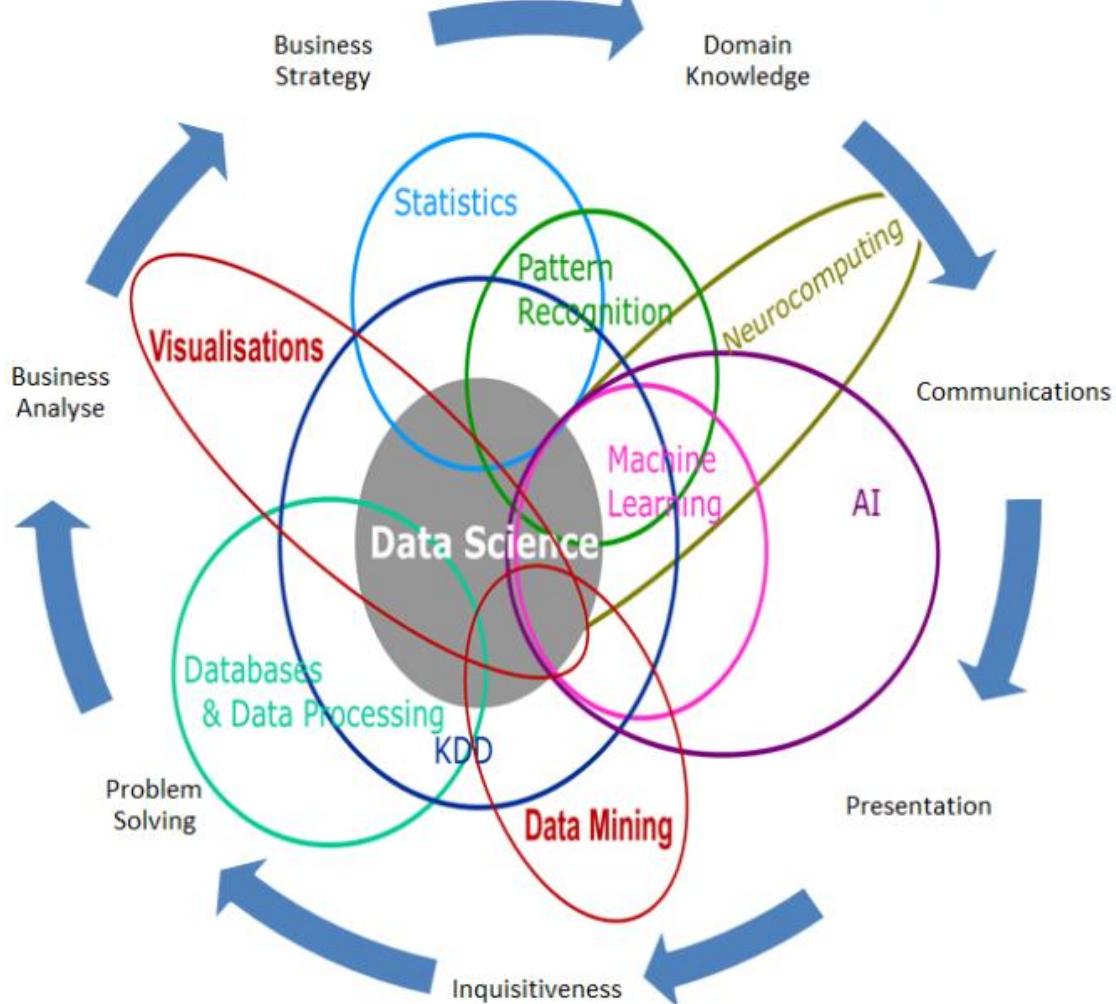


DATA SCIENCE

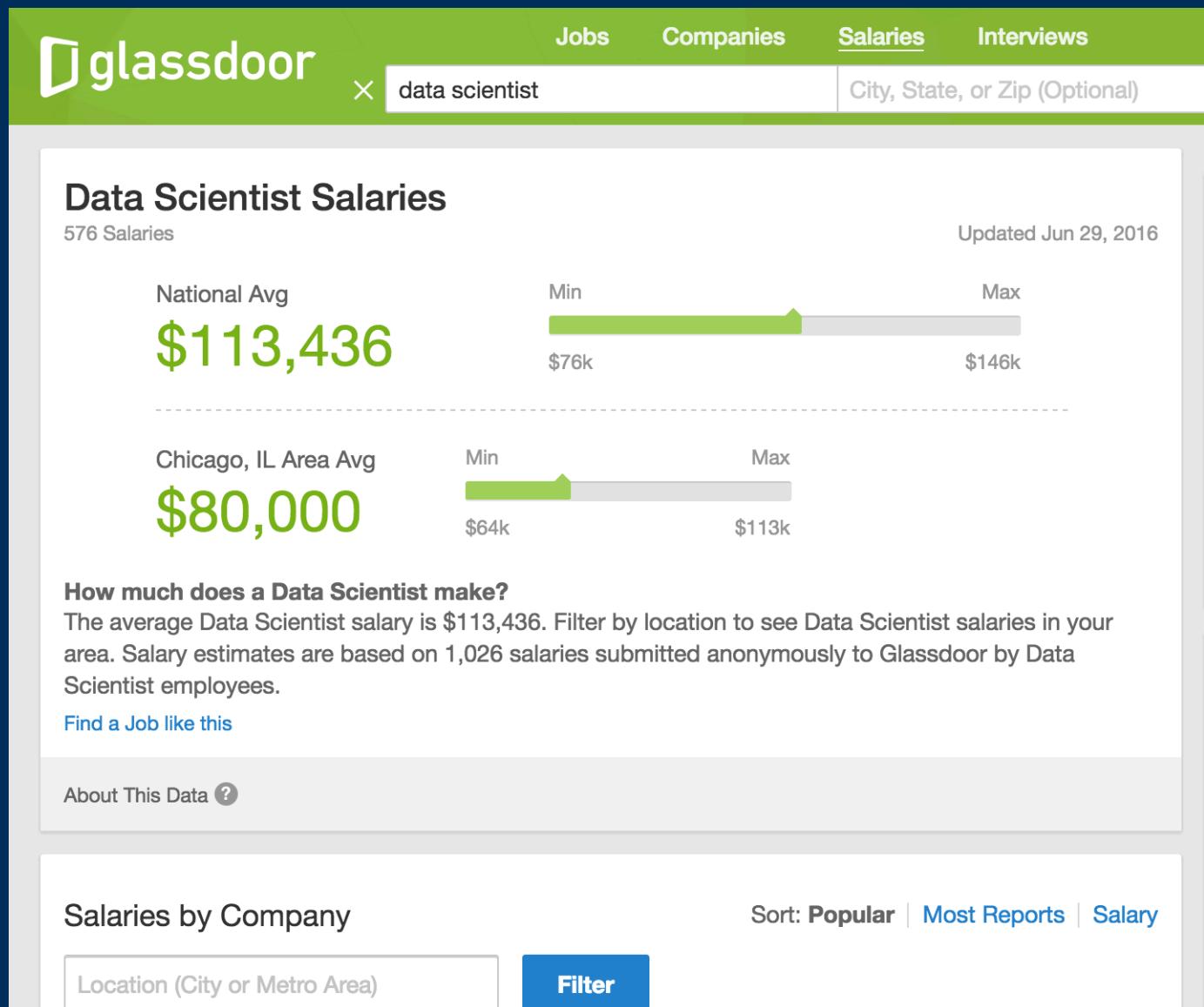


Data Science Is Multidisciplinary

By Brendan Tierney, 2012



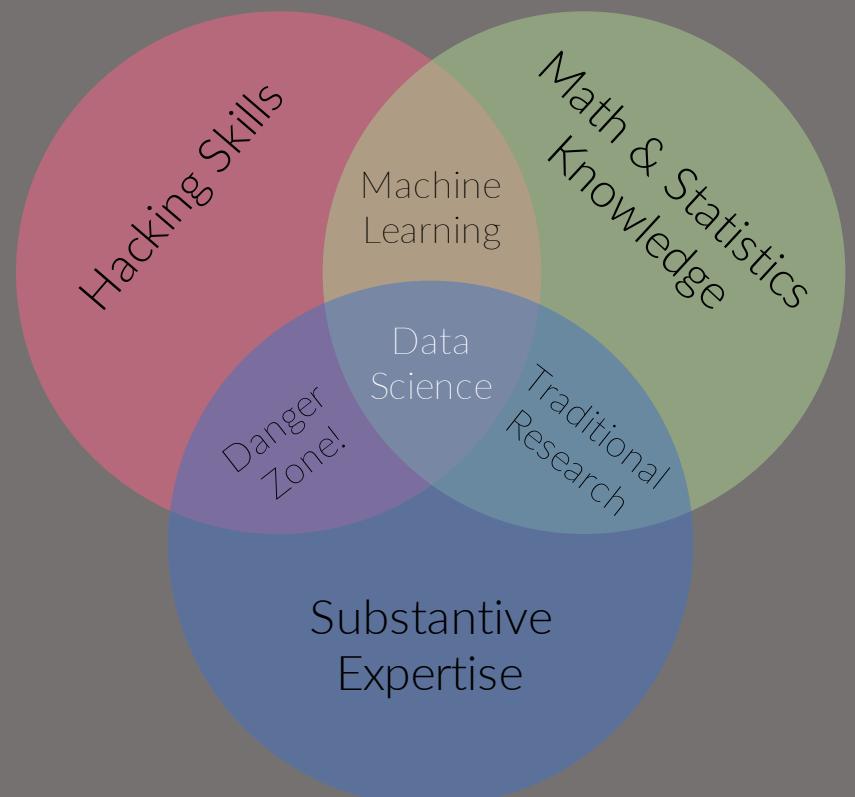
DATA SCIENCE IS EXPENSIVE



DATA SCIENCE

according to Drew Conway, CEO of Alluvium

NOT ALL
VENN DIAGRAMS
ARE USELESS

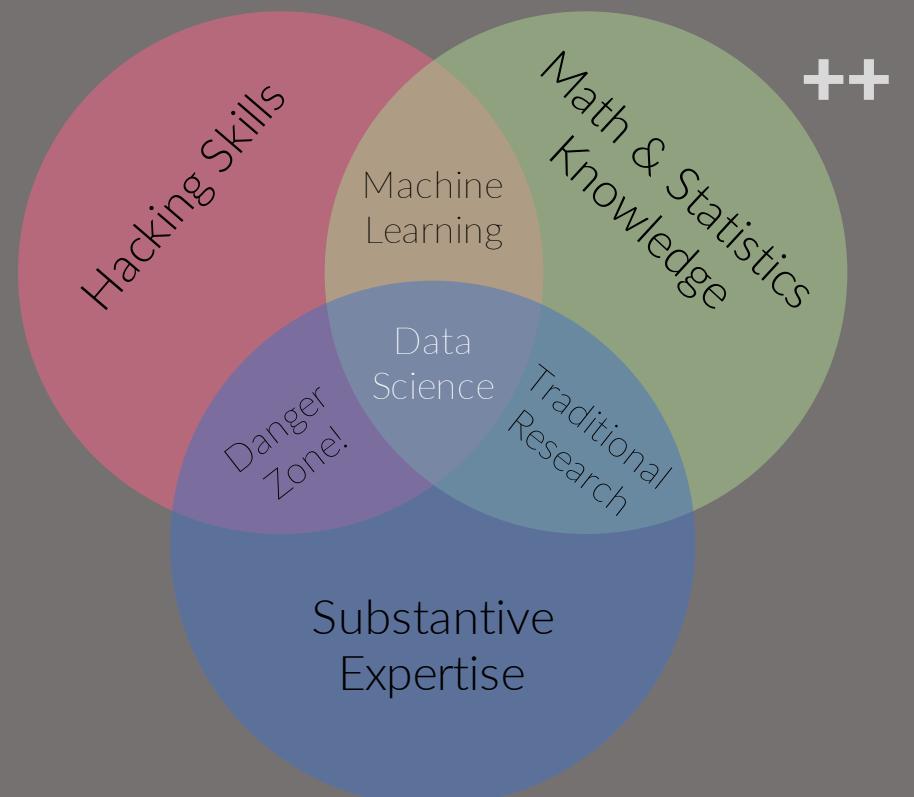


DATA SCIENCE

according to Drew Conway, CEO of Alluvium

**SMALL TEAM?
YOU CAN'T JUST BE
A DATA SCIENTIST**

You must be your own Product Manager
and User Experience Researcher



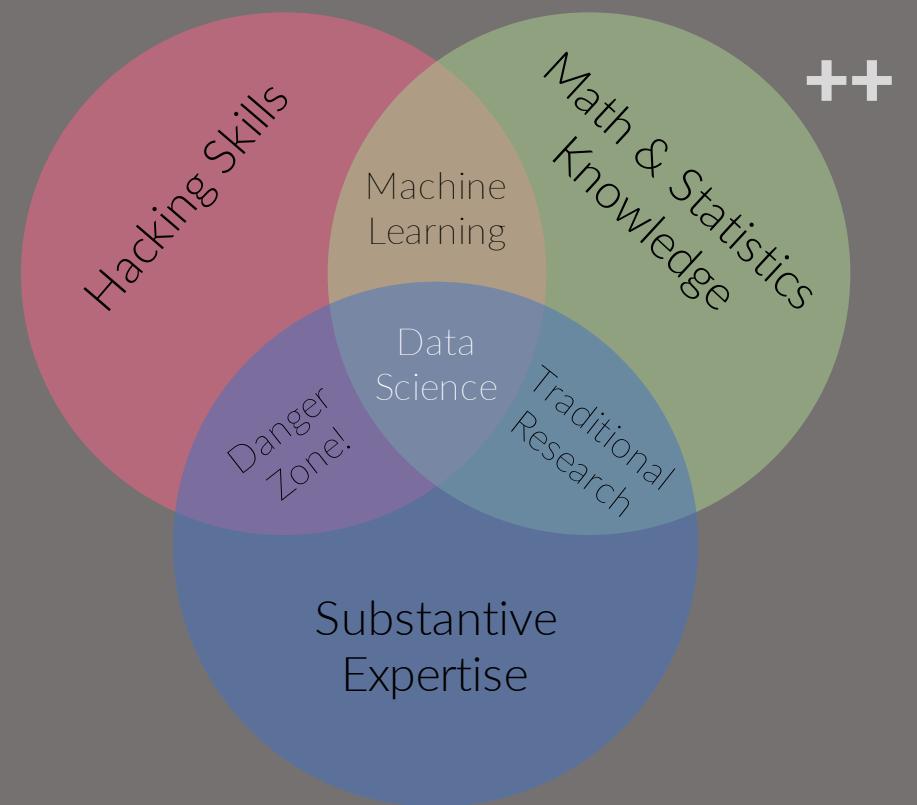
DATA SCIENCE

according to Drew Conway, CEO of Alluvium

**SMALL TEAM?
YOU CAN'T JUST BE
A DATA SCIENTIST**

You must be your own Product Manager
and User Experience Researcher

Sorry, but it's probably wise to spend ~20% of your time
OVER-COMMUNICATING with stakeholders
and UNDERSTANDING their needs

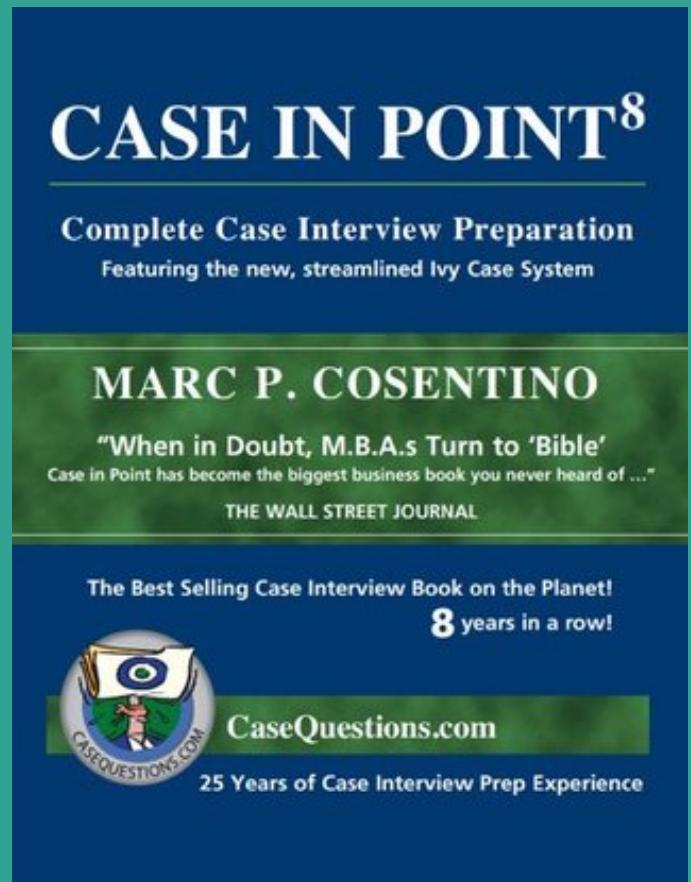


PM + UX

FROM ACADEMIA?

Learn to speak
strategy and business

GROSS MARGIN, KPIs, LTV, CAC,
ATTRIBUTION, ROI, RETENTION,
ATTRITION, MONTHS-TO-PAYBACK,
blah blah blah



drewfustin@gmail.com | 7.5.2016 | PyData Chicago

PM + UX

according to Arthur Geoffrion, UCLA Anderson School of Management

RESERVE THE RIGHT OF PROBLEM FORMULATION

Maxims for Modelers

Arthur Geoffrion

<http://www.anderson.ucla.edu/faculty/art.geoffrion/home/docs/Gudmdlg2.htm>

PM + UX

according to Arthur Geoffrion, UCLA Anderson School of Management

DEVELOP A CLEAR CHARTER AND PROJECT PLAN

Maxims for Modelers

Arthur Geoffrion

<http://www.anderson.ucla.edu/faculty/art.geoffrion/home/docs/Gudmdlg2.htm>

PM + UX

according to Arthur Geoffrion, UCLA Anderson School of Management

FIND A HIGH-LEVEL CHAMPION

Maxims for Modelers

Arthur Geoffrion

<http://www.anderson.ucla.edu/faculty/art.geoffrion/home/docs/Gudmdlg2.htm>

PM + UX

according to Arthur Geoffrion, UCLA Anderson School of Management

ESTABLISH PERSONAL CREDIBILITY AND PRODUCE RESULTS EARLY

Maxims for Modelers
Arthur Geoffrion
<http://www.anderson.ucla.edu/faculty/art.geoffrion/home/docs/Gudmdlg2.htm>

drewfustin@gmail.com | 7.5.2016 | PyData Chicago

PM + UX

according to Arthur Geoffrion, UCLA Anderson School of Management

INVOLVE THE STAKEHOLDER AND FUTURE USERS AT ALL STAGES

Maxims for Modelers
Arthur Geoffrion
<http://www.anderson.ucla.edu/faculty/art.geoffrion/home/docs/Gudmdlg2.htm>

PM + UX

according to Arthur Geoffrion, UCLA Anderson School of Management

COMMUNICATE OFTEN AND WELL

Maxims for Modelers

Arthur Geoffrion

<http://www.anderson.ucla.edu/faculty/art.geoffrion/home/docs/Gudmdlg2.htm>

OUR GOAL

FOCUS

when it works to stay small
provide a customer solution that
is usually similar each time,
and iteratively add features to
this solution so the product
expands over time

or

EXPAND

when scale is needed to be impactful
get more team members that
complement your own skills,
eventually growing a team to allow for
increased scope that includes both
Decision Science and Data Products

FOCUS

e.g. SaaS Data Product or Marketing Optimization Decision Science

CRITICAL: UX FEEDBACK LOOP

Understand (and help drive) expectations
and deliver each project excellently and completely.

FOCUS

e.g. SaaS Data Product or Marketing Optimization Decision Science

UNIT TEST ALL THE THINGS

While unit tests are always important, when a product is built iteratively, they are necessary. Have data factories to spoof typical/edge case data.

FOCUS

e.g. SaaS Data Product or Marketing Optimization Decision Science

DOUBLE THE DOCUMENTATION

Explain the process technically (for future you)
and accurately but simply (for the consumer).

EXPAND

e.g. Recommendation Engines + Attribution Models + Routing Algorithms + ...

CRITICAL: QUICK TO “GOOD ENOUGH”

You need to scale the team to be effective, so produce small, easy wins to earn favor and budget. Think Pareto’s principle: 80% quality at 20% time cost.

EXPAND

e.g. Recommendation Engines + Attribution Models + Routing Algorithms + ...

GET DATA MVPs INTO PRODUCTION

Many Data Products require lots of user data,
so get an MVP out to start, and wait to improve.

EXPAND

e.g. Recommendation Engines + Attribution Models + Routing Algorithms + ...

BE AN EXCELLENT DATA ENGINEER

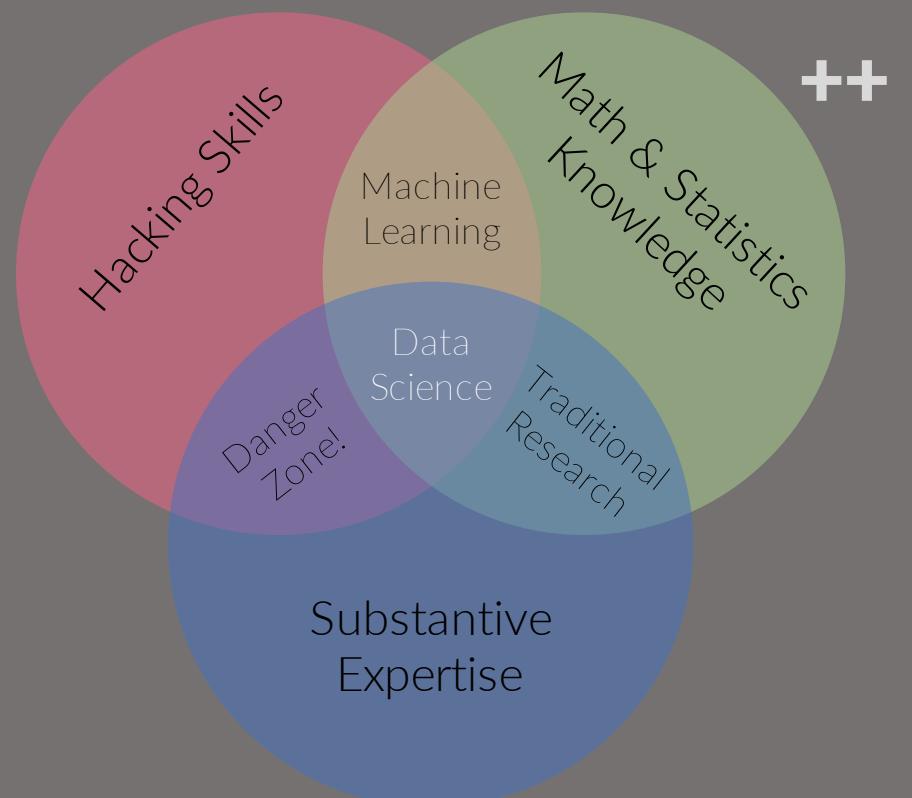
Projects are often standalone,
so have a reliable ETL process to reduce upkeep.
Upkeep time budget will increase over time.

DATA SCIENCE

according to Drew Conway, CEO of Alluvium

**SMALL TEAM?
YOU CAN'T JUST BE
A DATA SCIENTIST**

You must be your own Product Manager
and User Experience Researcher

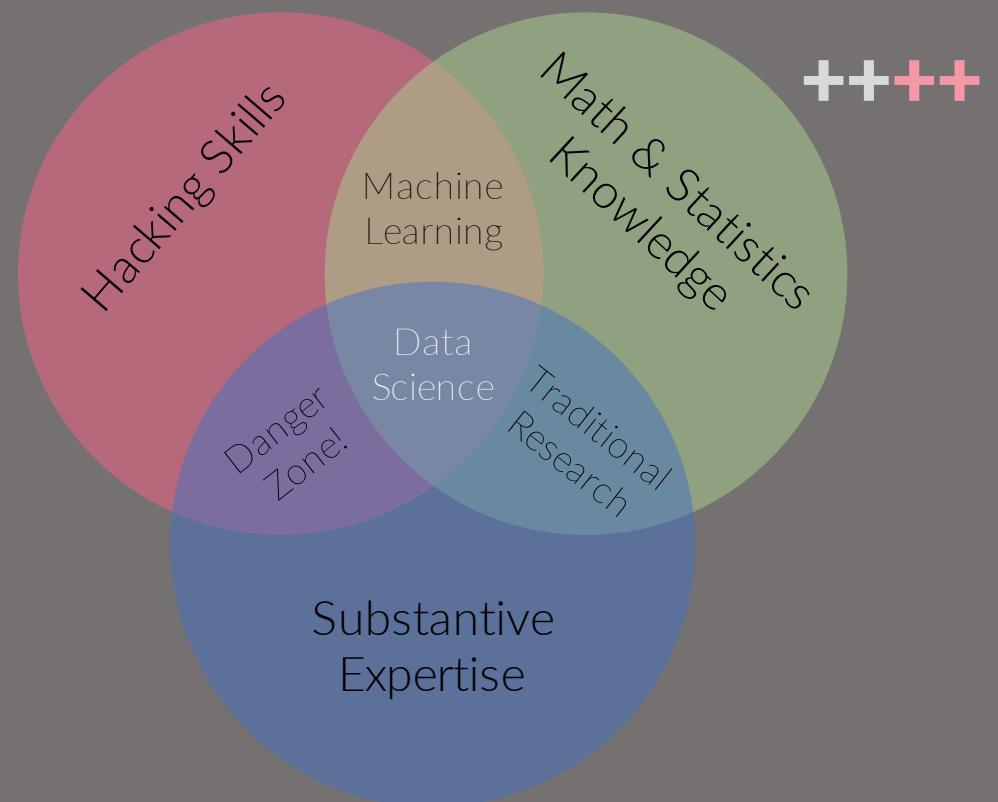


DATA SCIENCE

according to Drew Conway, CEO of Alluvium

**SMALL TEAM?
YOU CAN'T JUST BE
A DATA SCIENTIST**

You must be your own Product Manager
and User Experience Researcher
and Data Engineer and QA Test Engineer

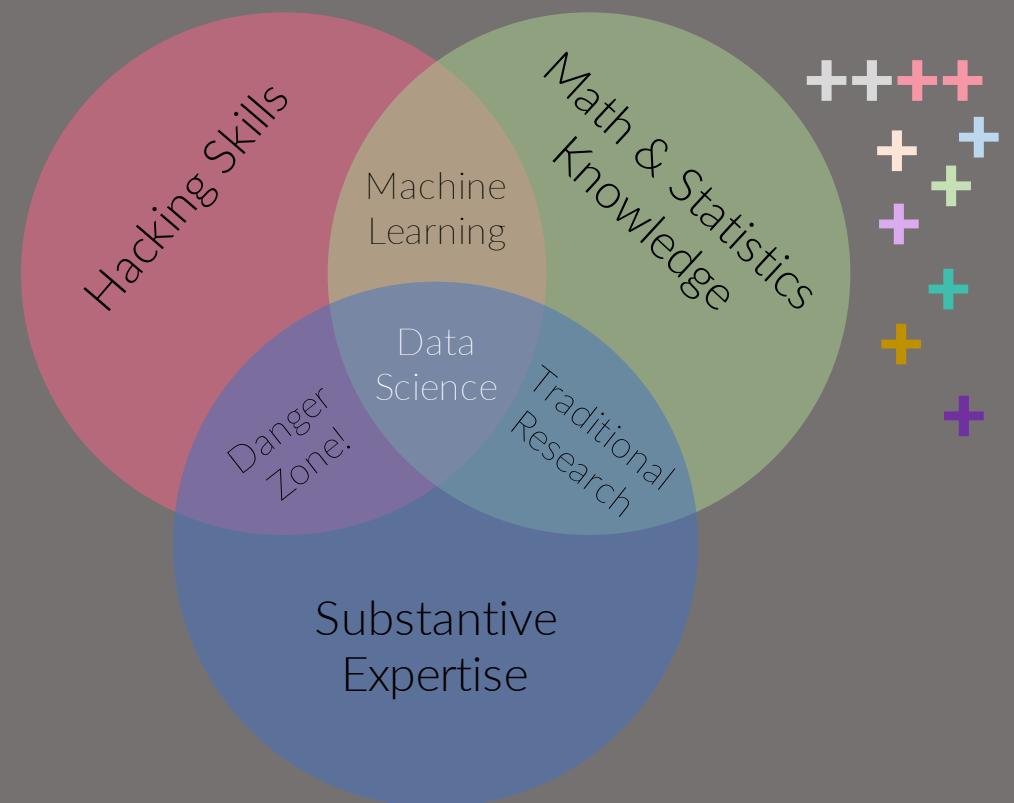


DATA SCIENCE

according to Drew Conway, CEO of Alluvium

**SMALL TEAM?
YOU CAN'T JUST BE
A DATA SCIENTIST**

You must be your own Product Manager
and User Experience Researcher
and Data Engineer and QA Test Engineer



AN ASIDE

drewfustin@gmail.com | 7.5.2016 | PyData Chicago

AN ASIDE

Lest you feel overwhelmed:
You don't have to be *great* at all of this.

AN ASIDE

Lest you feel overwhelmed:
You don't have to be *great* at all of this.

The beauty of being part of a small team is that what you contribute is probably going to be significantly better than what already exists.

ALSO... NO.

DON'T BE A PHYSICIST.
PEOPLE WHO ACTUALLY DO
THESE JOBS?

THEY'RE BETTER THAN YOU AT
DOING THEM.



<https://xkcd.com/793/>

drewfustin@gmail.com | 7.5.2016 | PyData Chicago

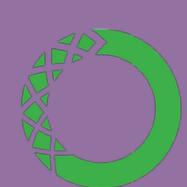
EXAMPLE

MY EVERYDAY TOOLKIT

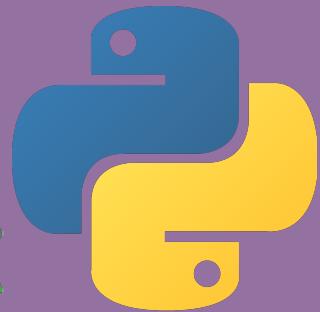
EXAMPLE

MY EVERYDAY TOOLKIT

PYTHON 3 + ANACONDA



ANACONDA®

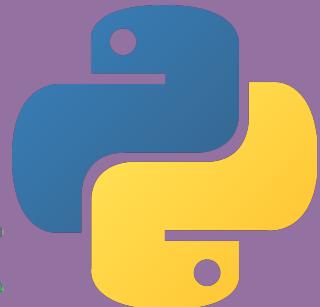


drewfustin@gmail.com | 7.5.2016 | PyData Chicago

EXAMPLE

MY EVERYDAY TOOLKIT

**PYTHON 3 + ANACONDA
+ JUPYTER + PLOTLY**



drewfustin@gmail.com | 7.5.2016 | PyData Chicago

EXAMPLE

MY EVERYDAY TOOLKIT

PYTHON 3 + ANACONDA
+ JUPYTER + PLOTLY
+ PANDAS

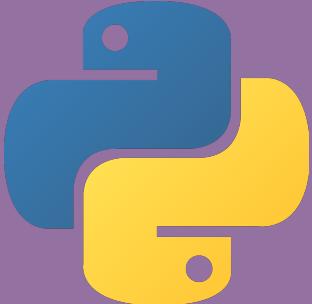
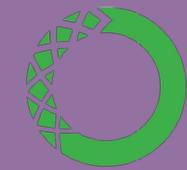
pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



plotly



ANACONDA®

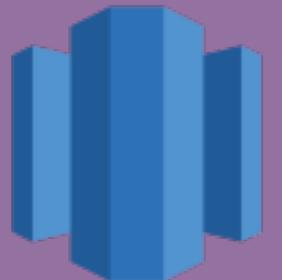


EXAMPLE

MARKETING OPTIMIZATION AT SPOTHERO

DATA COLLECTION

Transaction data in [AWS Redshift](#)



drewfustin@gmail.com | 7.5.2016 | PyData Chicago

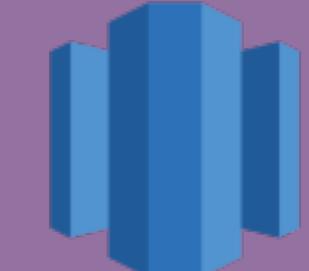
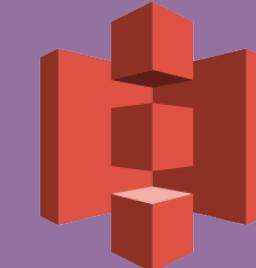
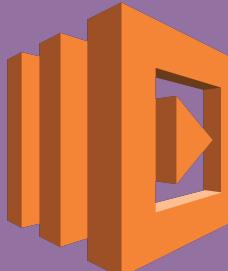
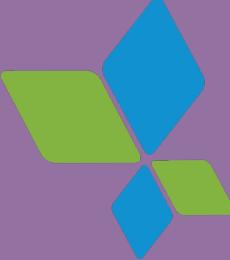
EXAMPLE

MARKETING OPTIMIZATION AT SPOTHERO

DATA COLLECTION

Transaction data in [AWS Redshift](#)

User acquisition data from [Google Analytics](#), [AppsFlyer](#), [Branch](#), [Typeform](#), etc
to Redshift via in-house/external ETL or webhooks + [AWS S3](#) + [AWS Lambda](#)

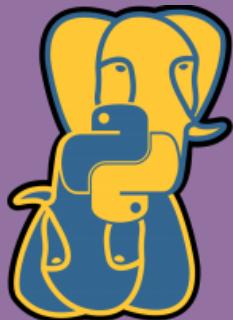


EXAMPLE

MARKETING OPTIMIZATION AT SPOTHERO

DATA QUERYING

Query customer data using [Psycopg2](#) against Redshift into [pandas](#) dataframes



pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



EXAMPLE

MARKETING OPTIMIZATION AT SPOTHERO

CHANNEL ATTRIBUTION

Develop an algorithm in `pandas` to assign for each customer the likelihood that a particular marketing channel was used in their acquisition

pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



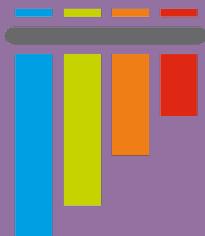
EXAMPLE

MARKETING OPTIMIZATION AT SPOTHERO

CHANNEL ATTRIBUTION

Develop an algorithm in `pandas` to assign for each customer the likelihood that a particular marketing channel was used in their acquisition

Data factories spoof customer data for `pytest` unit tests of acquisition model



pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



EXAMPLE

MARKETING OPTIMIZATION AT SPOTHERO

COHORT VALUE FORECASTS

Use historical rental trends to predict future revenue in [pandas](#)

pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



EXAMPLE

MARKETING OPTIMIZATION AT SPOTHERO

COHORT VALUE FORECASTS

Use historical rental trends to predict future revenue in `pandas`

Regression forecasting + attrition modeling + backtesting in `scikit-learn`

ARIMA forecasting + tests for autocorrelation, stationarity, etc in `statsmodels`



pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

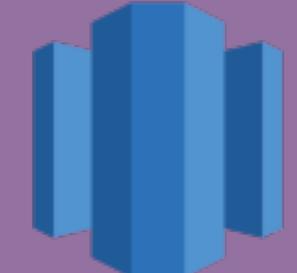
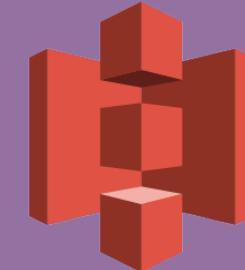


EXAMPLE

MARKETING OPTIMIZATION AT SPOTHERO

MARKETING SPEND

Move marketing spend csv to [Redshift](#) via [S3](#) using [Boto 3](#) + [Psycopg2](#)



drewfustin@gmail.com | 7.5.2016 | PyData Chicago

EXAMPLE

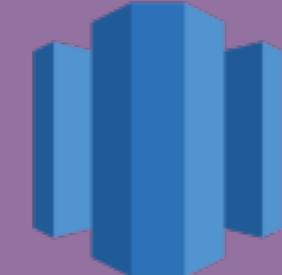
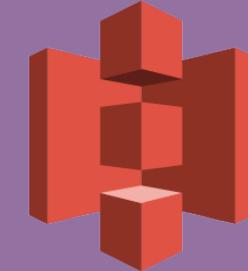
MARKETING OPTIMIZATION AT SPOTHERO

MARKETING SPEND

Move marketing spend csv to [Redshift](#) via [S3](#) using [Boto 3](#) + [Psycopg2](#)

Assign spend to customers (and cohorts) with a divvying algorithm
applied using (you guessed it) [pandas](#)

pandas
 $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

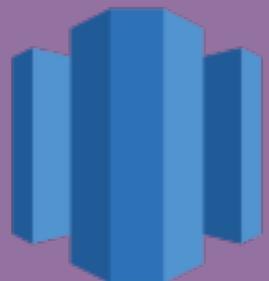


EXAMPLE

MARKETING OPTIMIZATION AT SPOTHERO

CAC/LTV RATIO OPTIMIZATION

All customer acquisition costs (CAC) and lifetime value (LTV) forecasts are written to a data store in [Redshift](#)



EXAMPLE

MARKETING OPTIMIZATION AT SPOTHERO

CAC/LTV RATIO OPTIMIZATION

All customer acquisition costs (CAC) and lifetime value (LTV) forecasts are written to a data store in [Redshift](#)

Marketing analysts use this data in the [Looker](#) BI tool to find efficient channels

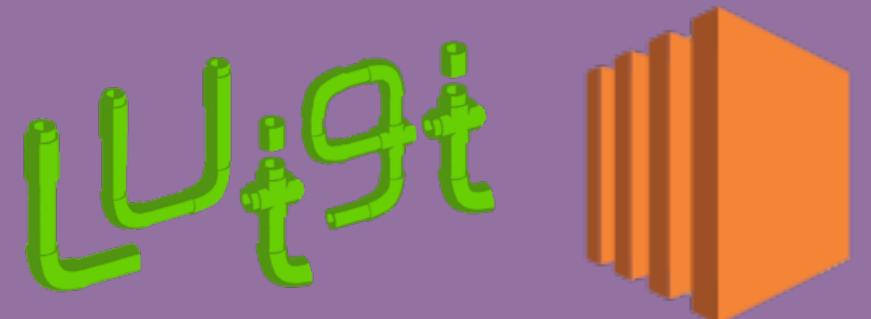


EXAMPLE

MARKETING OPTIMIZATION AT SPOTHERO

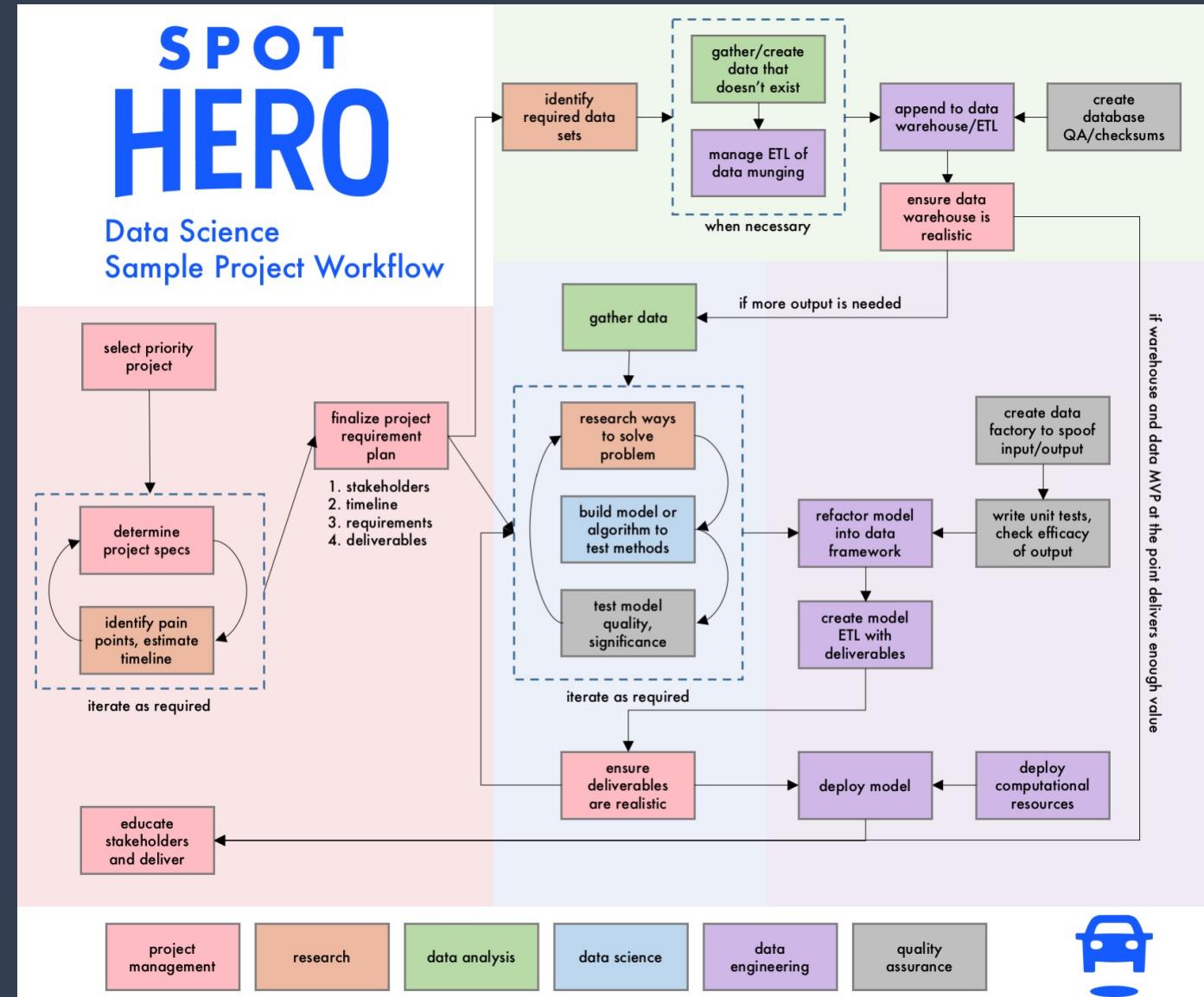
ETL PROCESS

Task dependencies on each branch of this process are handled by [Luigi](#) on an [AWS EC2](#) instance



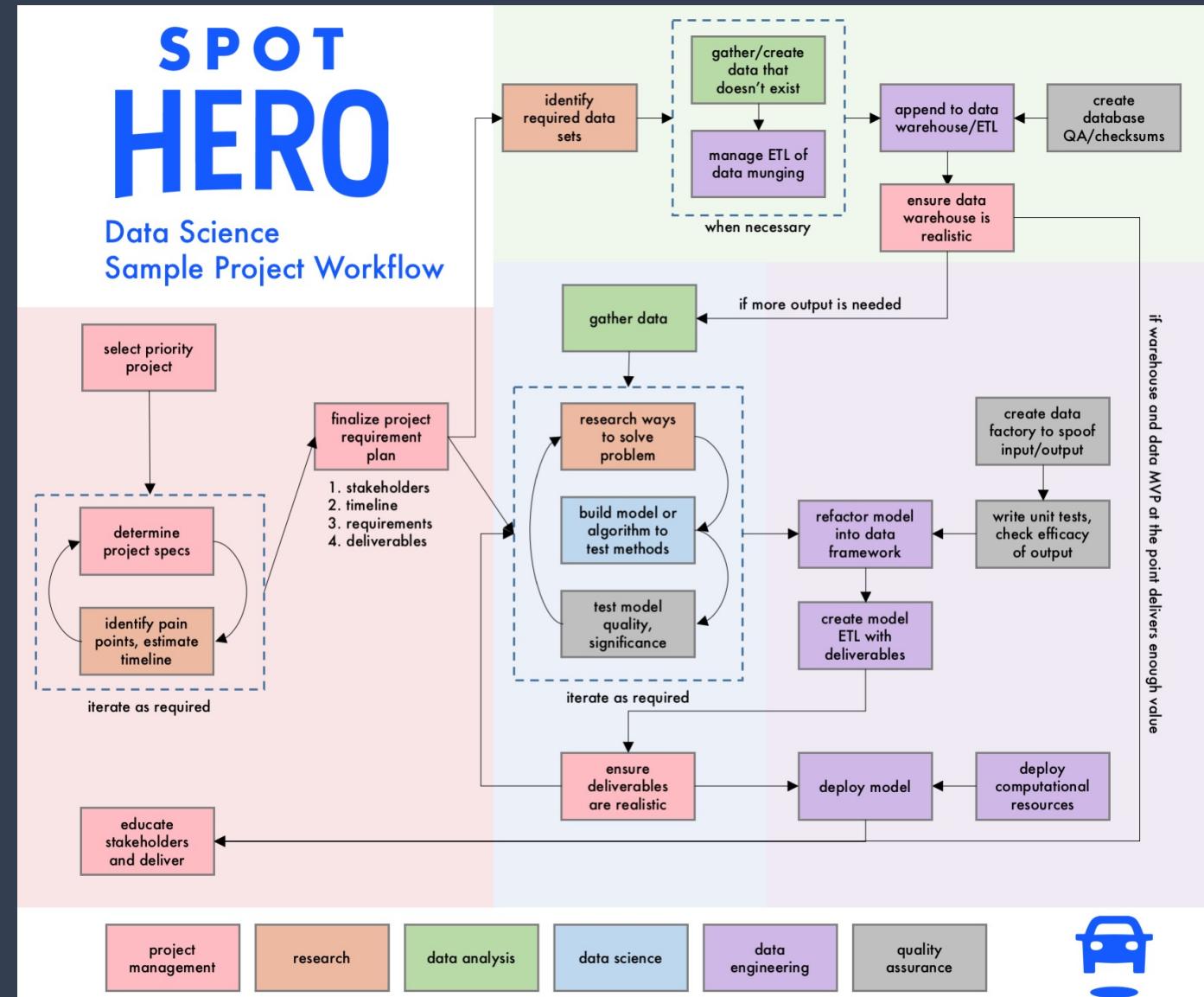
drewfustin@gmail.com | 7.5.2016 | PyData Chicago

REMINDER: DATA SCIENCE IS COMPLICATED



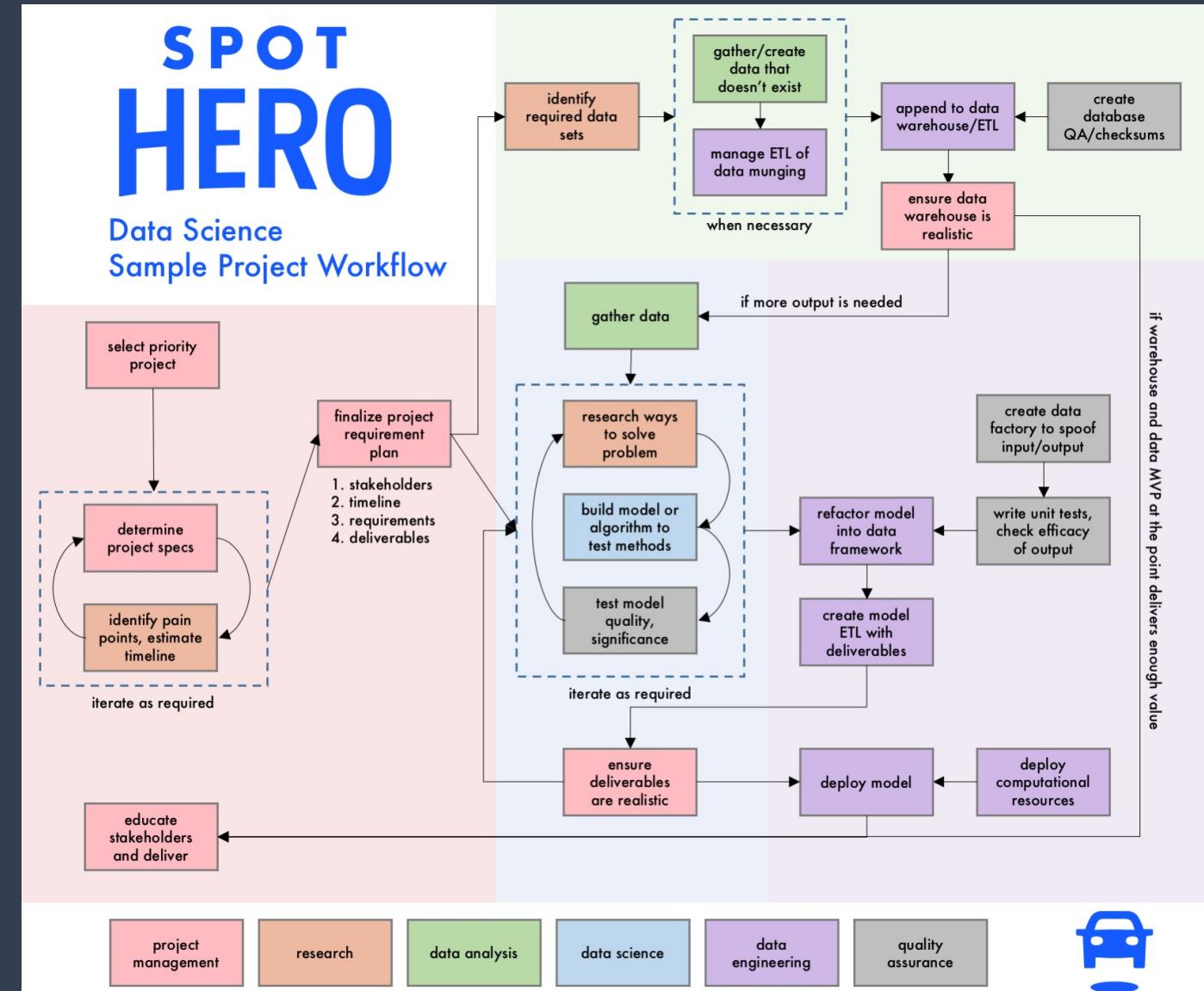
REMINDER: DATA SCIENCE IS COMPLICATED

TRY YOUR BEST TO BE YOUR OWN:
DATA SCIENTIST
DATA ENGINEER
QA TEST ENGINEER
PROJECT MANAGER
UX RESEARCHER



REMINDER: DATA SCIENCE IS COMPLICATED

TRY YOUR BEST TO BE YOUR OWN:
DATA SCIENTIST
DATA ENGINEER
QA TEST ENGINEER
PROJECT MANAGER
UX RESEARCHER



DATA SCIENCE with **SMALL TEAMS**

“While decision science and data products call for some of the same skills, it’s rare for data scientists to excel at both. Decision science depends on business and product sense, systems thinking, and strong communication skills. Data products require machine learning knowledge and production-level engineering skills. If you have a small data science team, you may need to find the rare superstars who can do both. But you’ll benefit from specialization as you scale your team.” Jeremy Stanley and Daniel Tunkelang

FURTHER READING

[Doing Data Science Right – Your Most Common Questions Answered \[Jeremy Stanley and Daniel Tunkelang\]](#)

<http://firstround.com/review/doing-data-science-right-your-most-common-questions-answered/>

[Maxims for Modelers \[Arthur Geoffrion\]](#)

<http://www.anderson.ucla.edu/faculty/art.geoffrion/home/docs/Gudmdl2.htm>

[Highly Effective Data Science Teams \[Drew Harry\]](#)

<https://medium.com/mit-media-lab/highly-effective-data-science-teams-e90bb13bb709>

[Data Engineering Architecture at Simple \[Rob Story\]](#)

https://github.com/wrobstory/DataEngArchSimple/blob/master/2016_03_29_SimpleDataArch_with_notes.pdf

[Data Science Team-Building and Optimization \(talk\) \[Jeremy Stanley\]](#)

https://www.youtube.com/watch?v=CqQyrkEvh_8