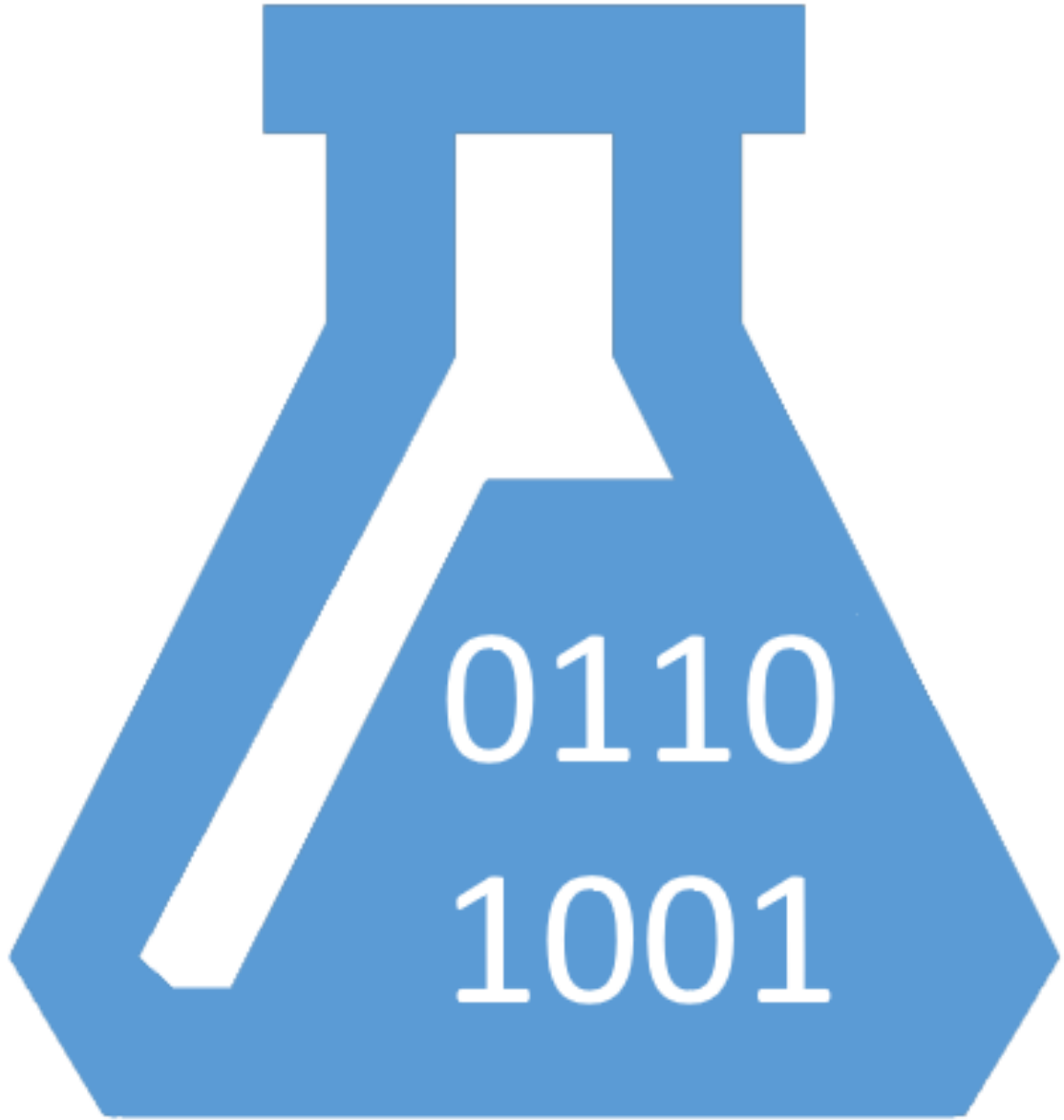
Data Science for Developers

@MatthewRenze

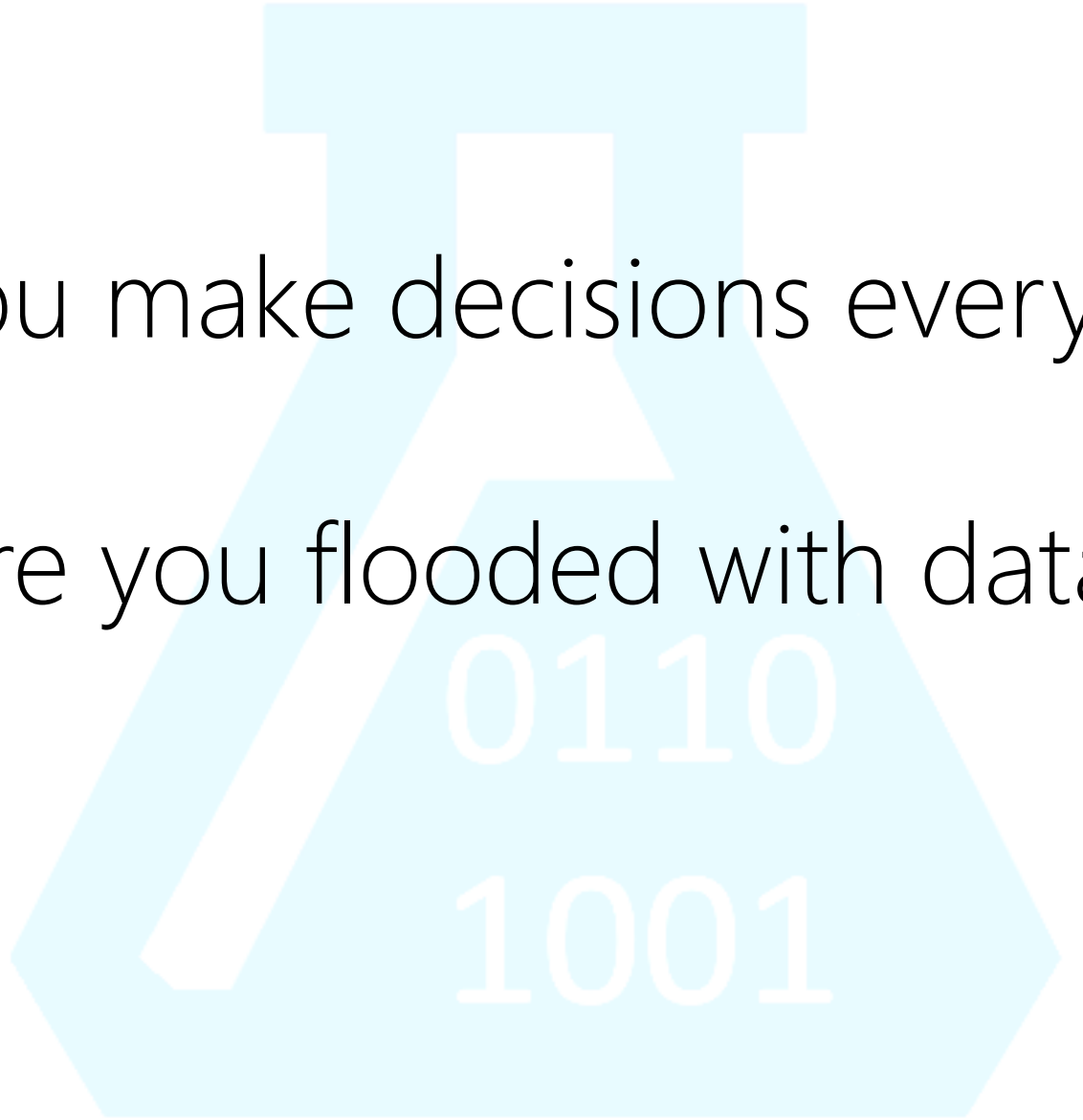
#Cigna


0110
1001

Do you make decisions every day?

Do you make decisions every day?

Are you flooded with data?



Do you make decisions every day?

Are you flooded with data?

Are you using data to make decisions?

What is data science?

Why is it important?

How do I get started?

The New York Times

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

TWITTER
LINKEDIN
COMMENTS (58)
SIGN IN TO E-MAIL

Data Scientist:
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

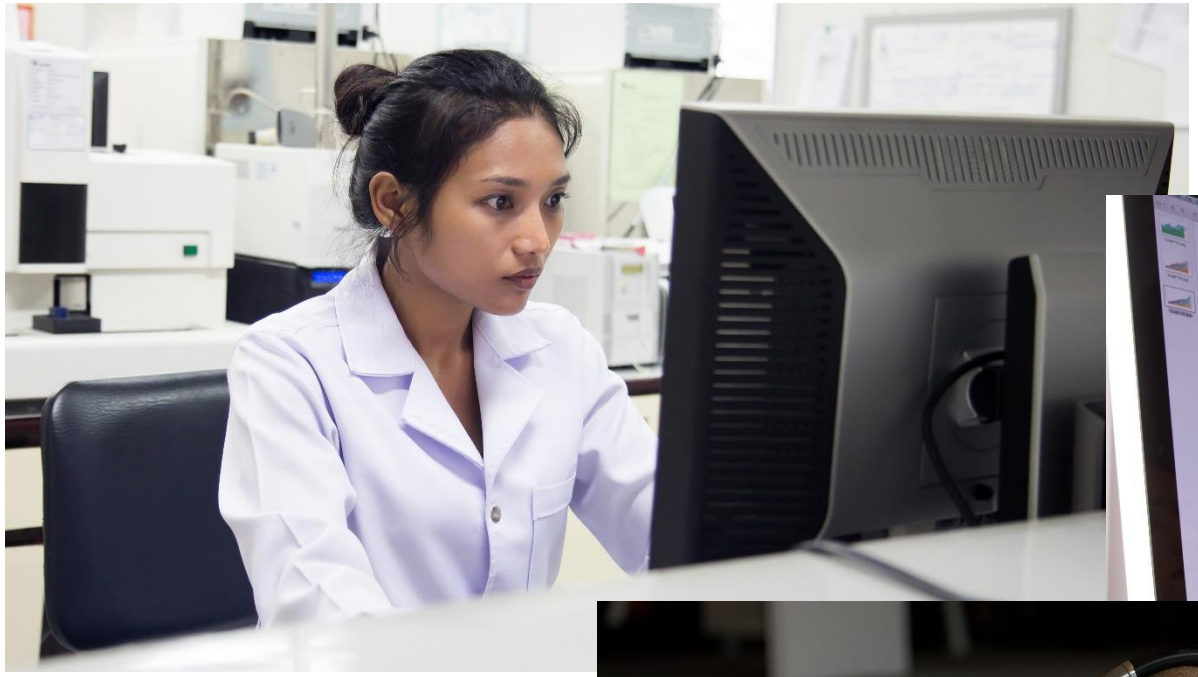
Job Postings for Data Scientists

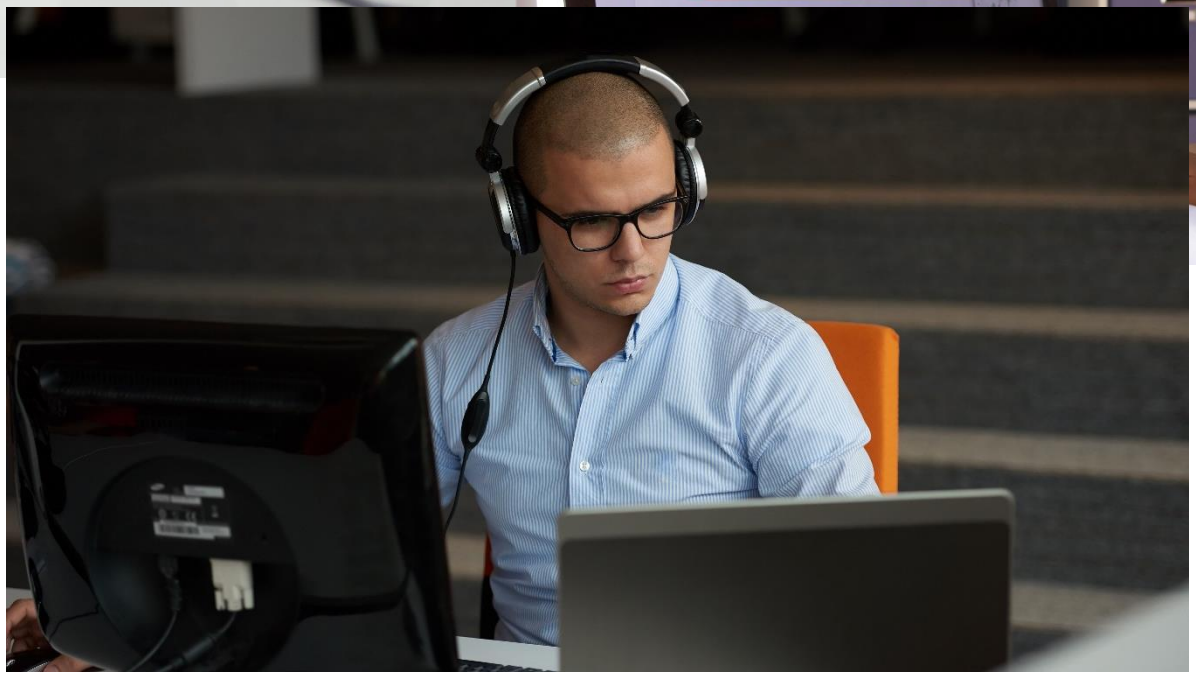
Source: Thinknum Job Listings

Top-paying Tech Skills

Skill	2019	Change
Apache Kafka	$134,557	▲ 5.5%
HANA	$134,462	▲ 9.4%
Cloudera	$133,695	▲ 8.9%
MapReduce	$132,708	▲ 7.9%
Cassandra	$132,497	▲ 6.7%
Chef	$132,136	▲ 19.5%
Mokito	$131,772	▲ 11.5%
Service Oriented Architecture (SOA)	$131,556	▲ 7.6%
Amazon Redshift	$130,723	▲ 4.5%
PAAS	$130,669	▲ 6.3%
Deep Learning	$129,978	— N/A
Elasticsearch	$129,938	▲ 4.8%
Zookeeper	$129,833	▲ 8.2%
Amazon Route 53	$129,295	▲ 7.7%
Dynamo DB	$129,255	▲ 2.9%
Jetty	$128,751	▲ 14.4%
NoSQL	$127,741	▲ 4.3%
Redis	$127,441	▲ 4.0%
Spring Framework	$127,286	▲ 16.0%
Containers	$127,110	▲ 5.7%

Source: Dice Salary Survey 2019






IDEAL
MODEL SMALL
STACK 100h
DATASEG
HW DB "Danger"
CODESEG
Begin:
MOV AX, @data
MOV DS, AX
MOV DX, OFFSET HW
MOV AH, 09H
INT 21H
MOV AX, 4C00H
INT 21H
END Begin

Overview

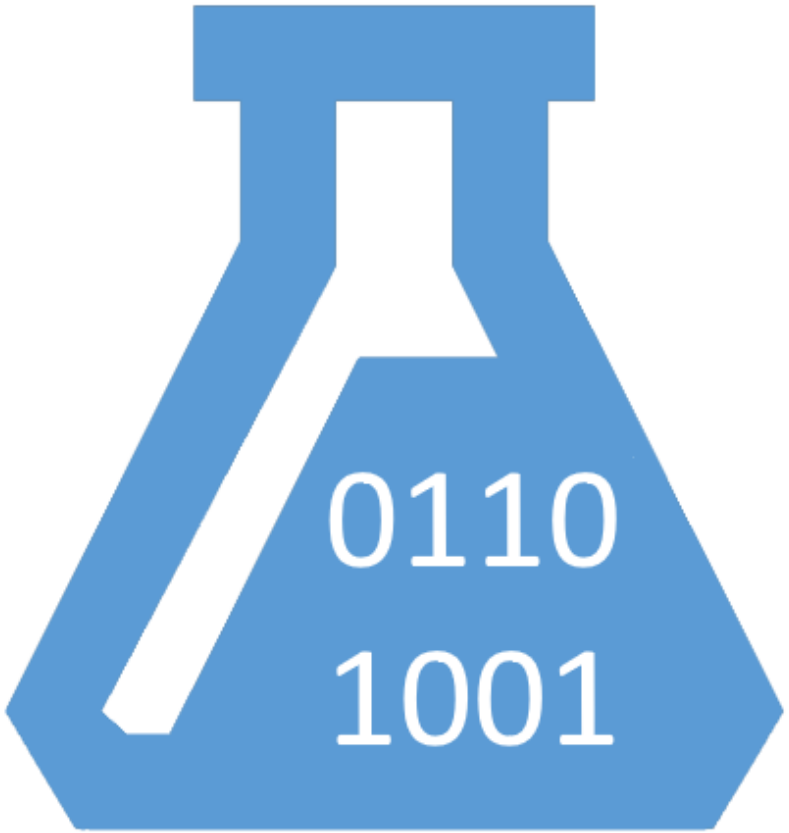
What Is Data Science?

Why Is It Important?

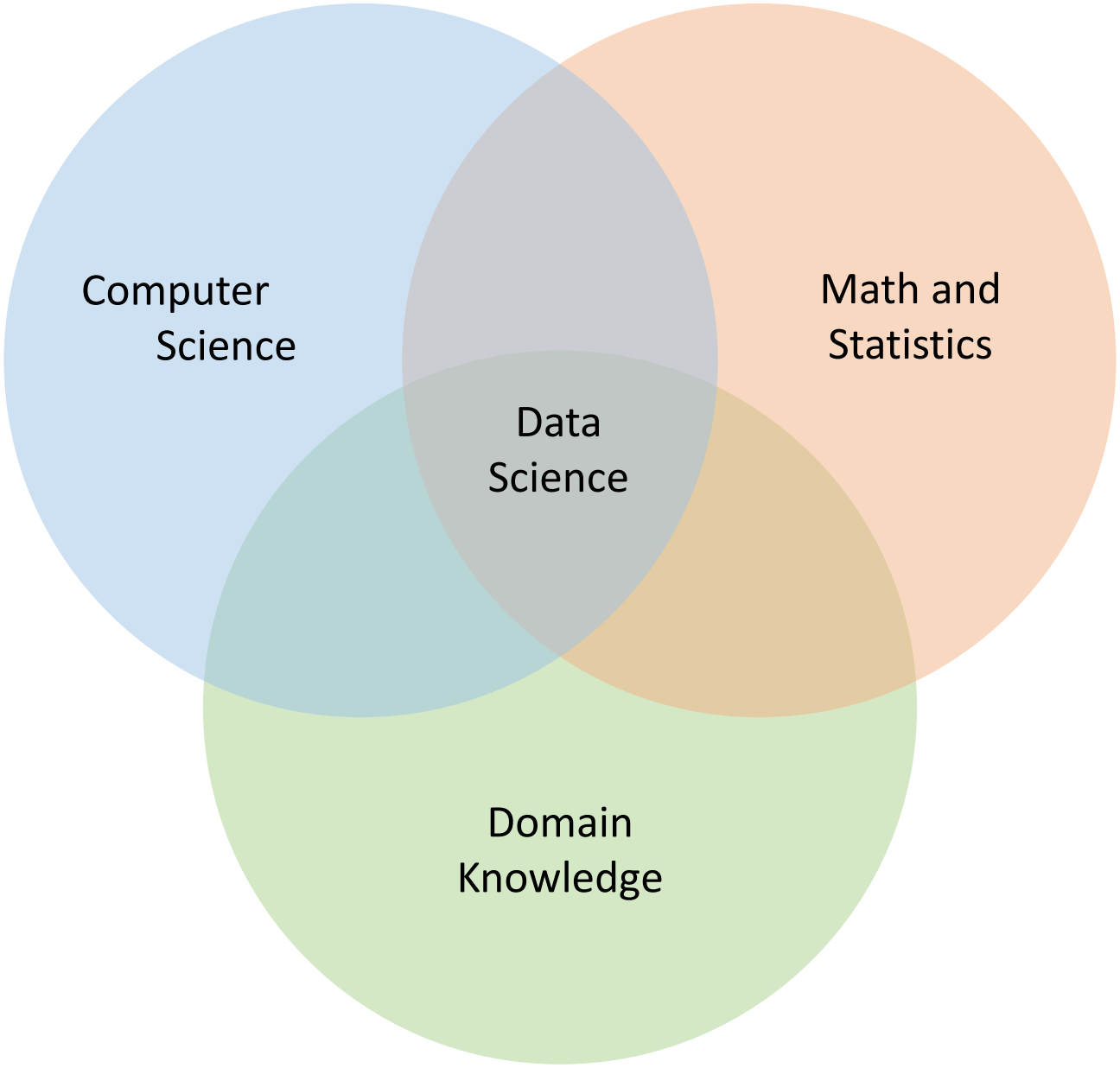
Building Intelligent Software

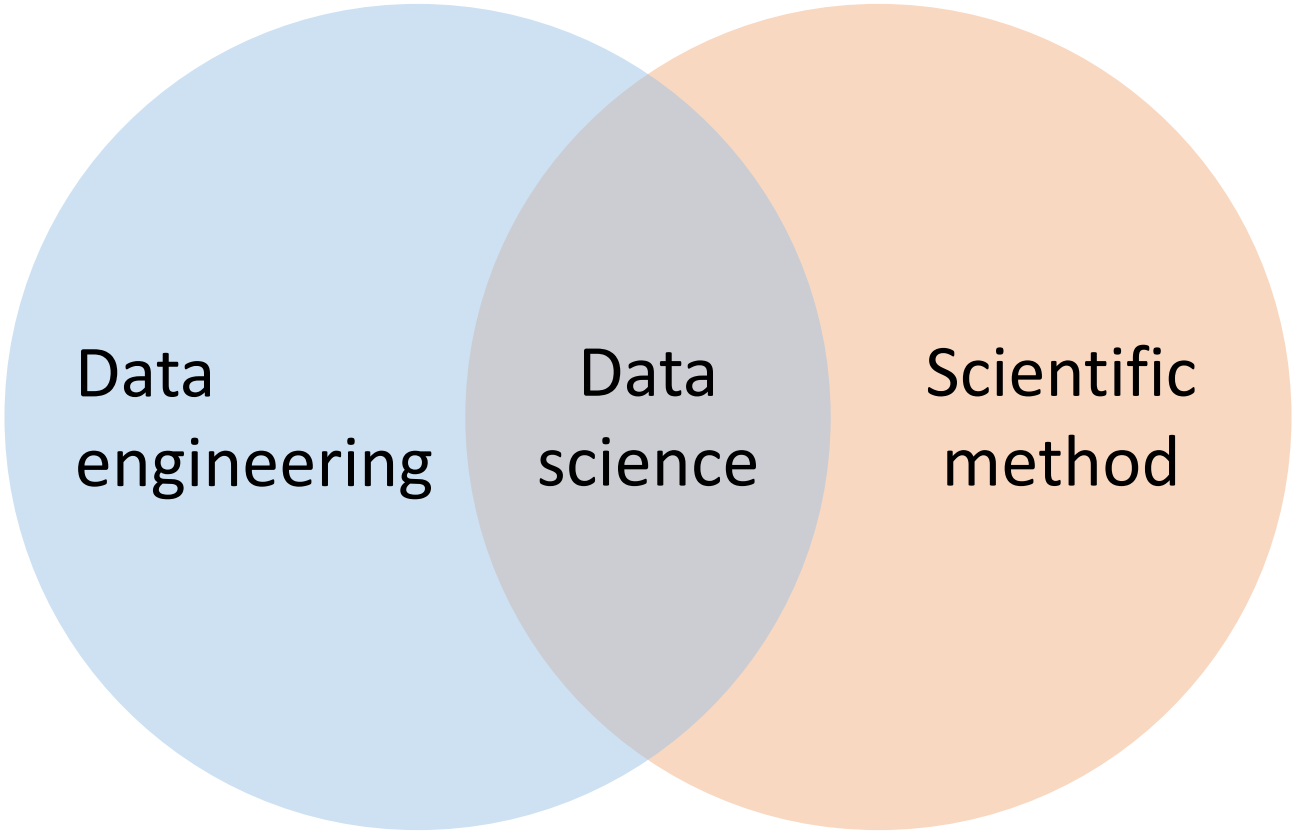
Improving Development Practices

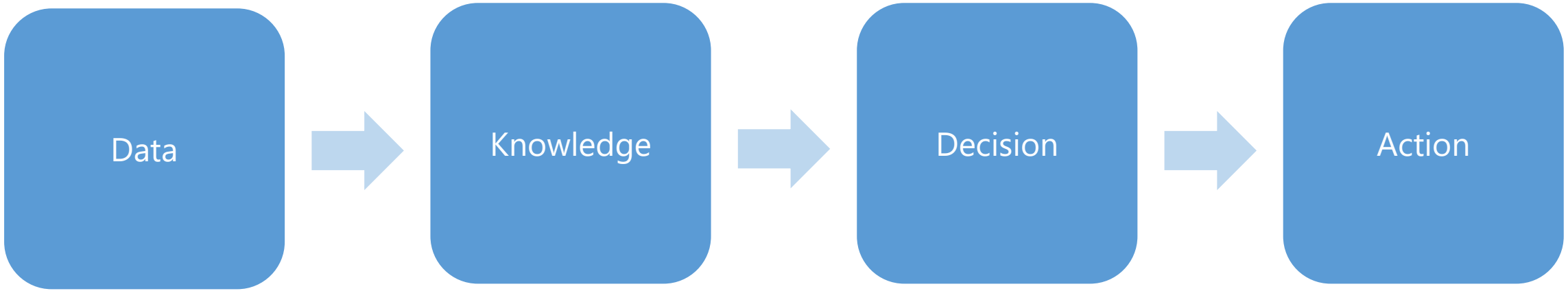
How Do I Get Started?



What Is Data Science?


Computer Science
Math and Statistics
Data Science
Domain Knowledge


Data engineering
Data science
Scientific method


Data
Knowledge
Decision
Action

What Is a Data Scientist?

Performs data science

More than a scientist

More than an analyst

More than a developer

Data Science Team

Data Science Team

Data Engineer

Domain Expert

Data Scientist

Data Science Team

Data Engineer

Domain Expert

Data Scientist

Data Science Team

Data Engineer

Domain Expert

Data Scientist

What skills are necessary?

Data Science Skills

Programming

Working with data

Descriptive statistics

Data visualization

Data Science Skills

Programming

Working with data

Descriptive statistics

Data visualization

Statistical modeling

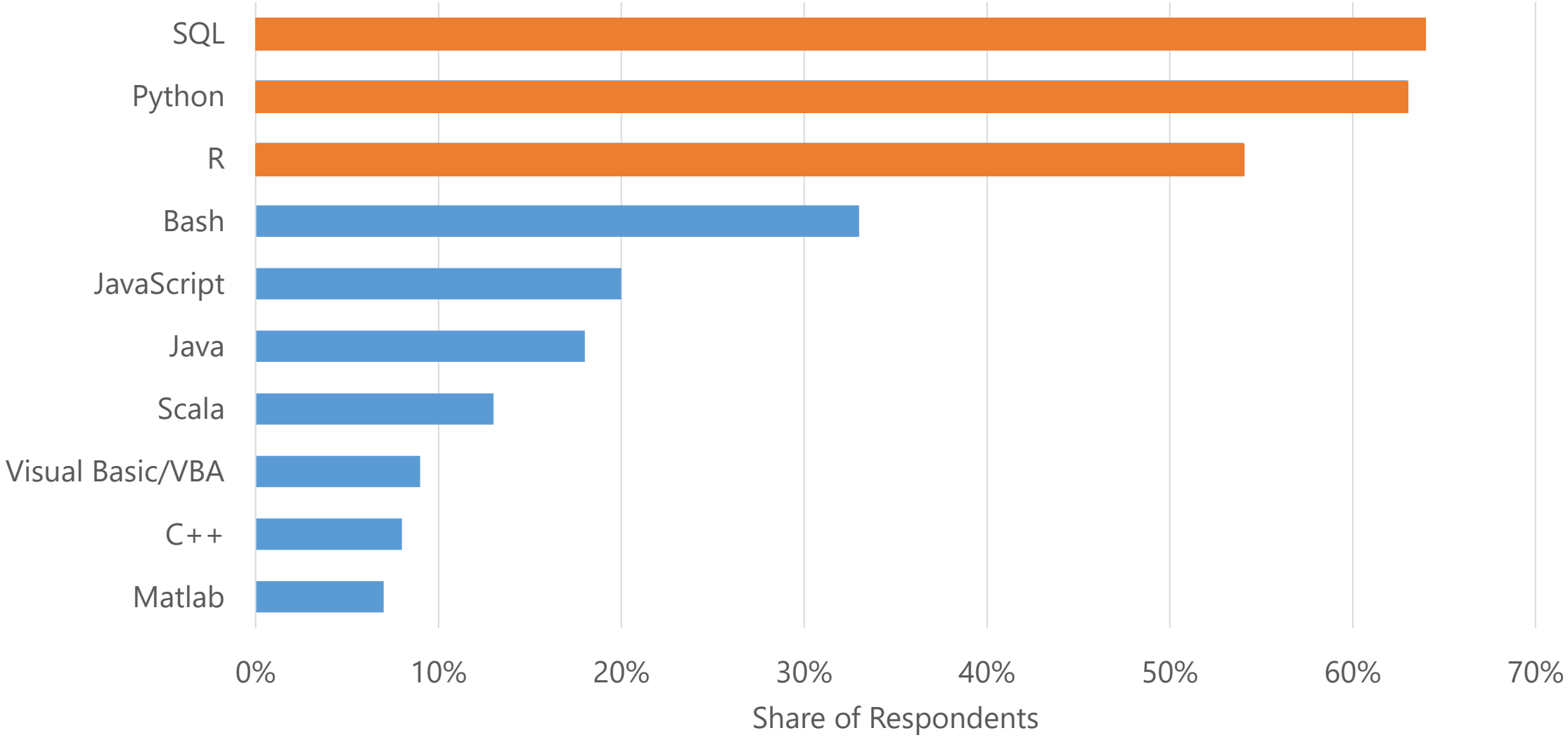
Handling Big Data

Machine learning

Deploying to production

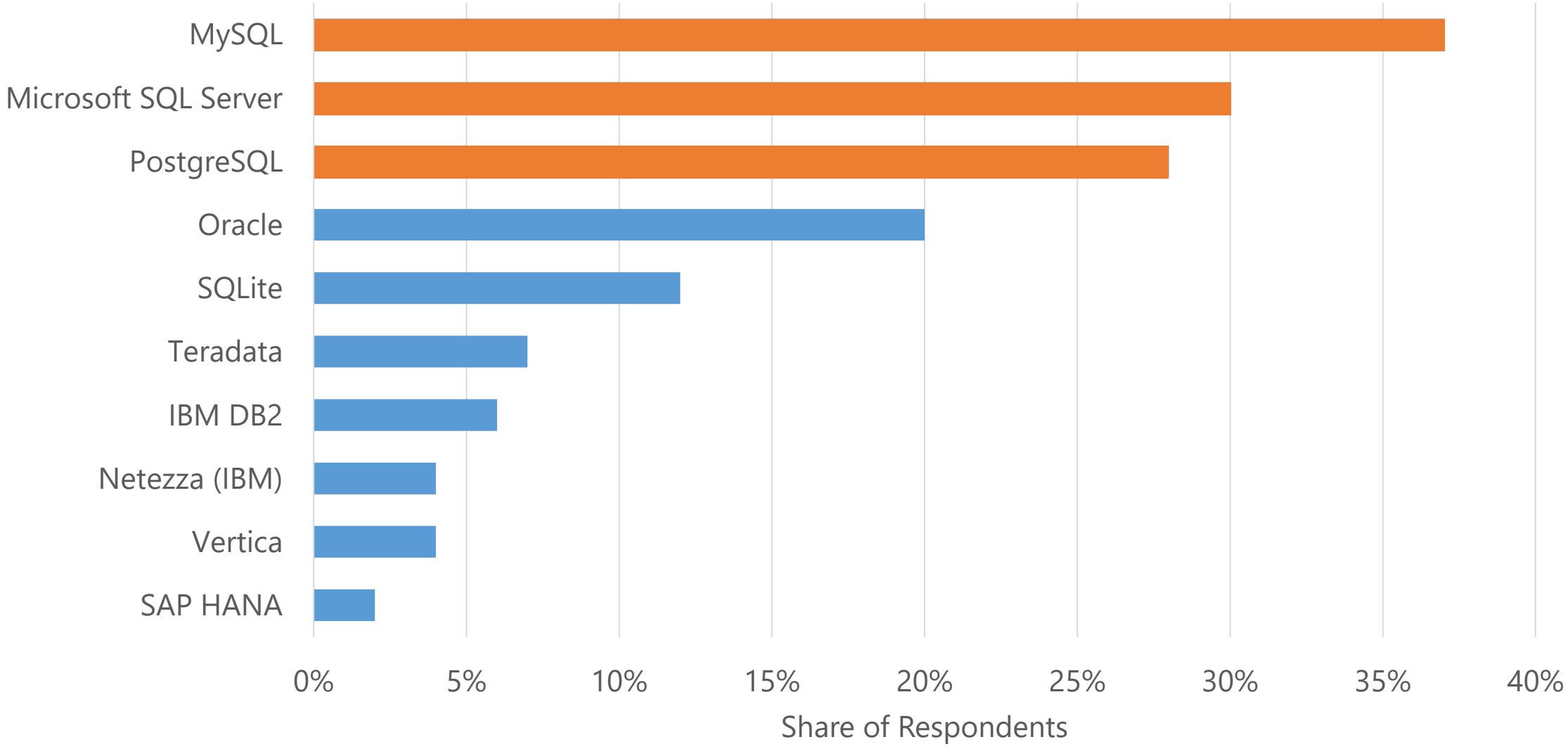
What tools are used?

Programming Languages



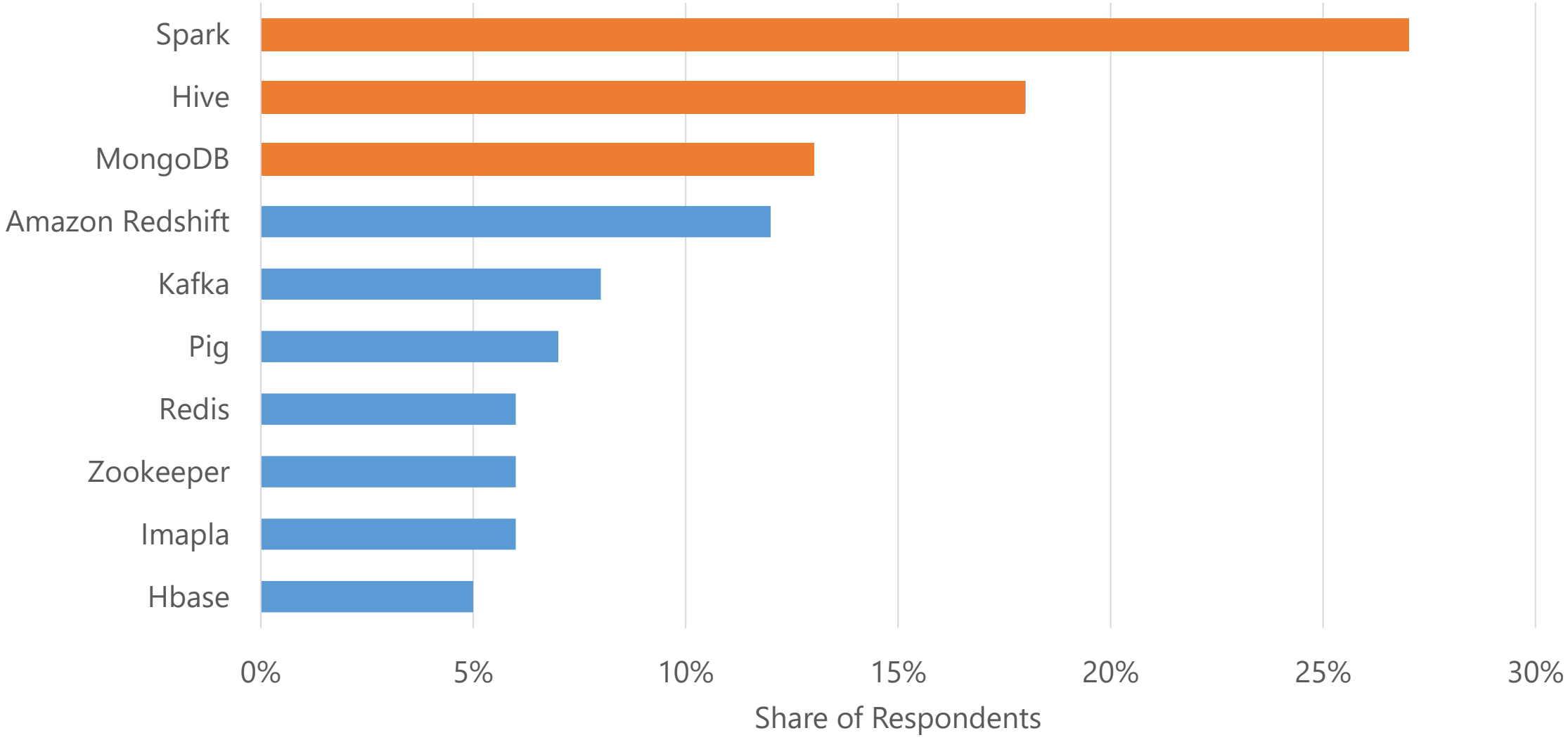
Source: O'Reilly 2017 Data Science Salary Survey

Relational Databases



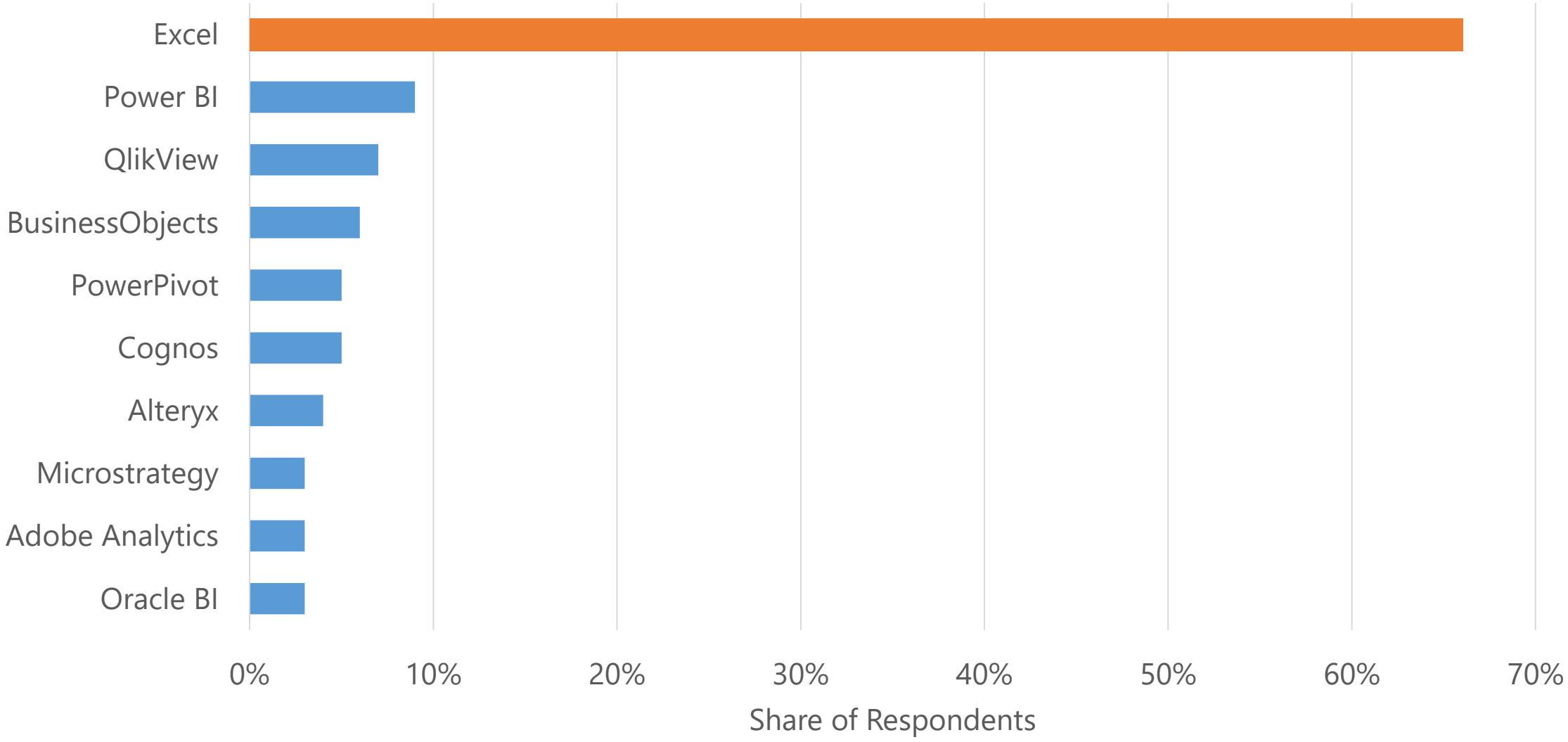
Source: O'Reilly 2017 Data Science Salary Survey

Big Data Platforms



Source: O'Reilly 2017 Data Science Salary Survey

Spreadsheets, BI, Reporting



Source: O'Reilly 2017 Data Science Salary Survey

How is data science performed?

The Data Science Process

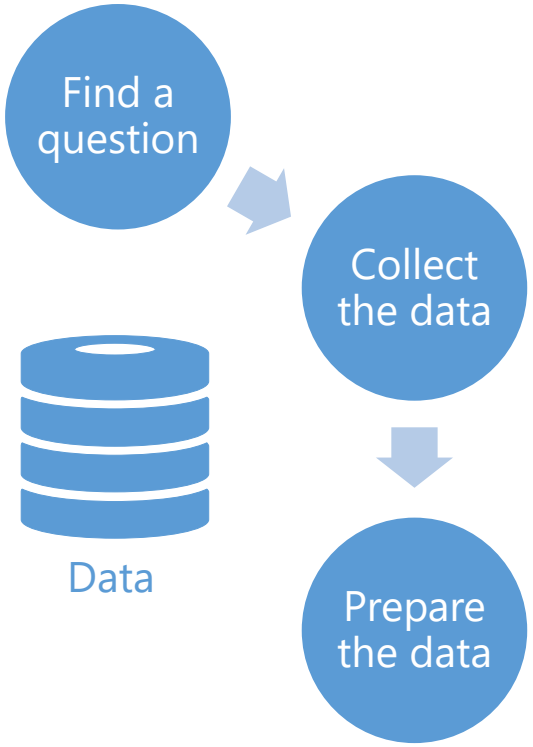
Data

The Data Science Process

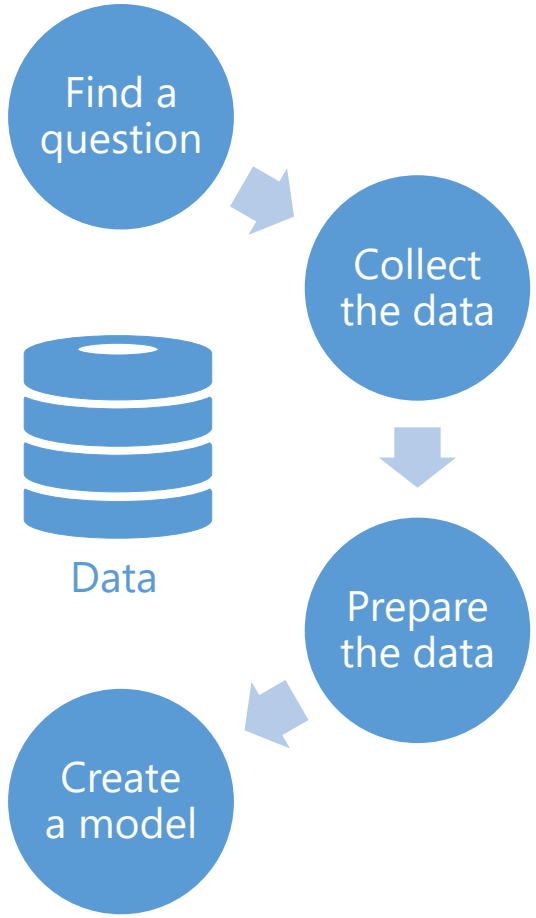
Data

The Data Science Process

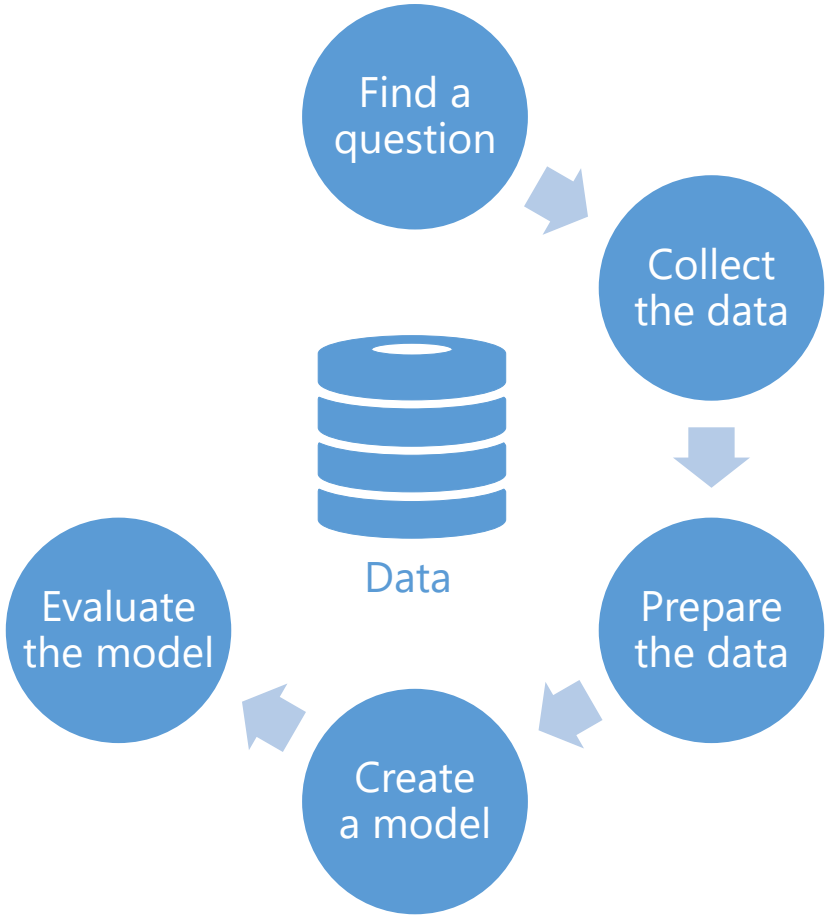
The Data Science Process



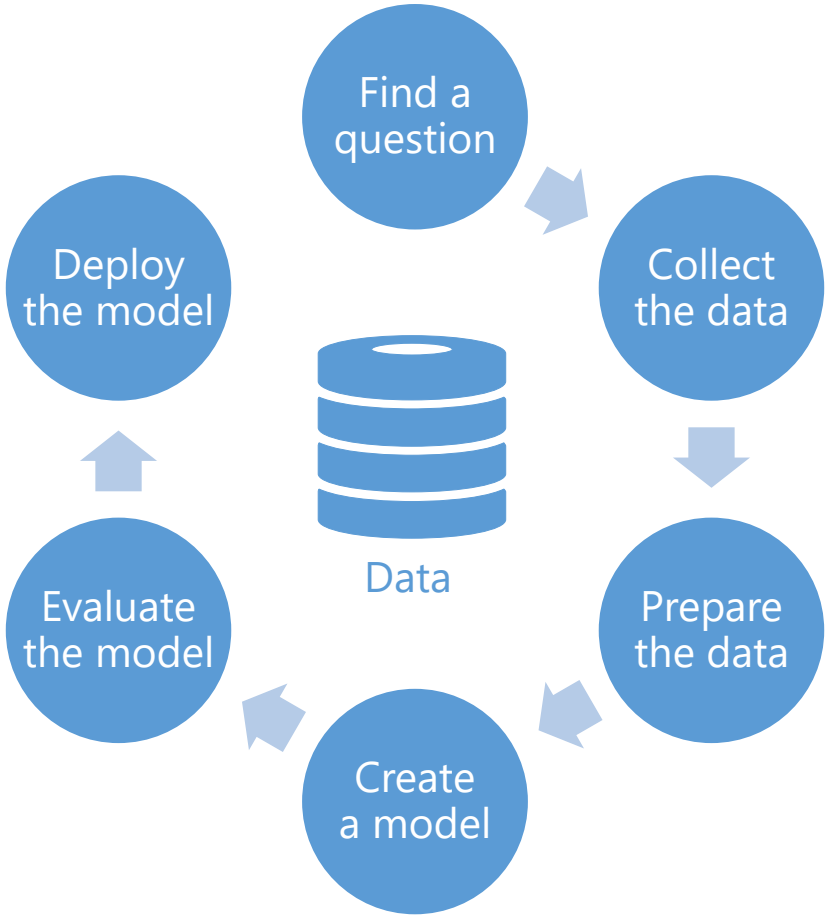
The Data Science Process



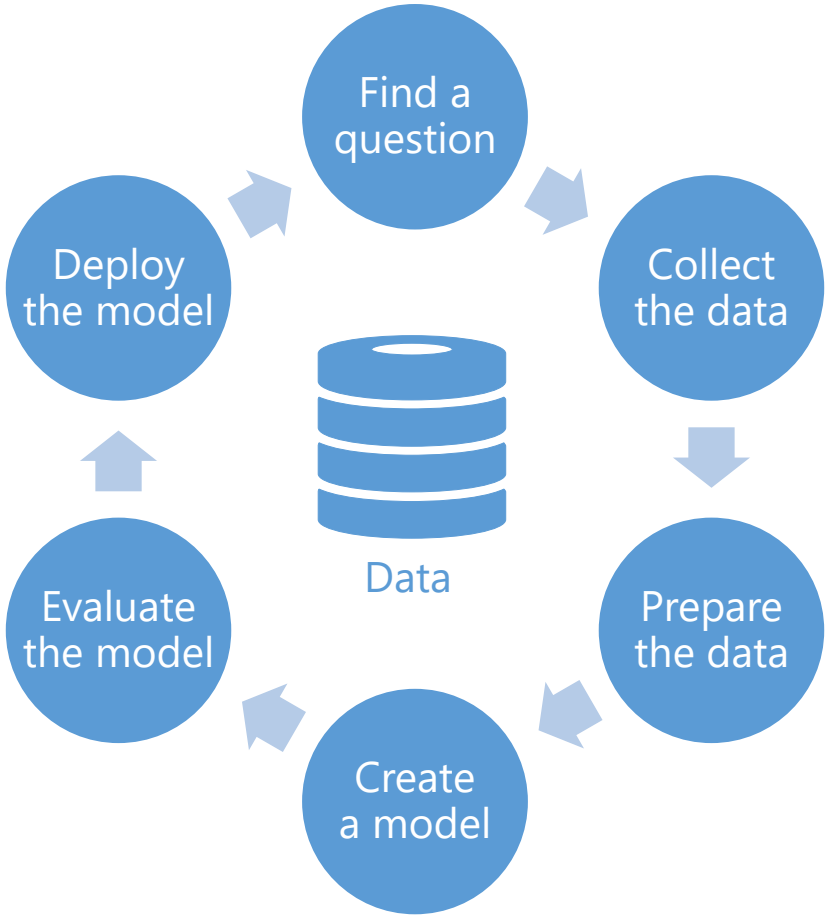
The Data Science Process



The Data Science Process

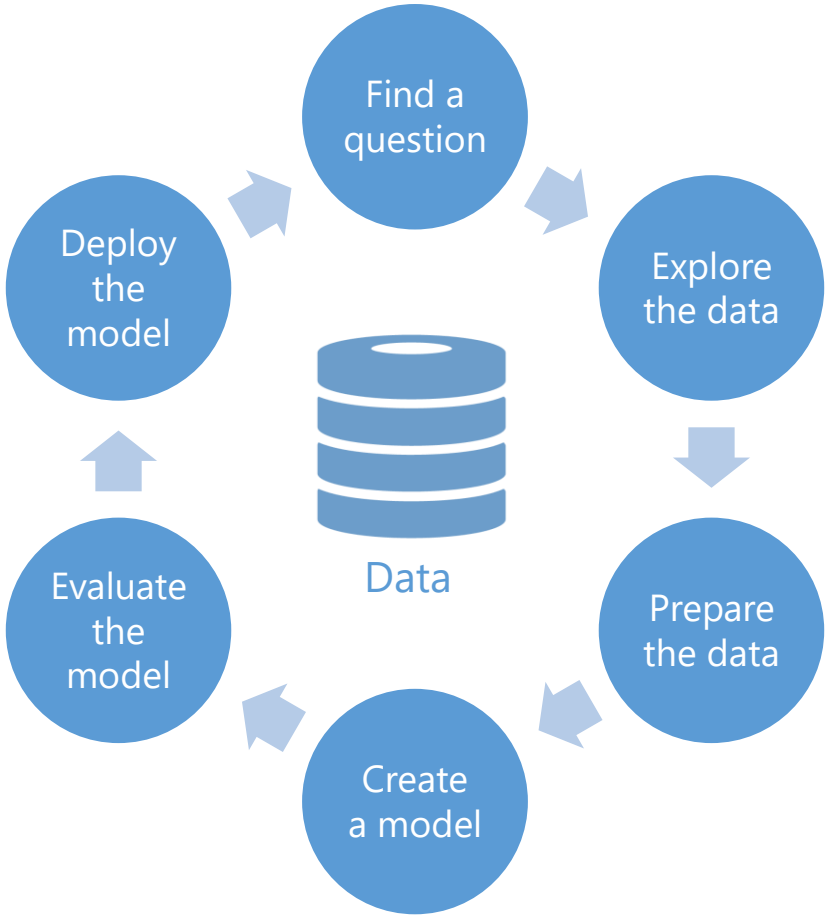


The Data Science Process



The Data Science Process

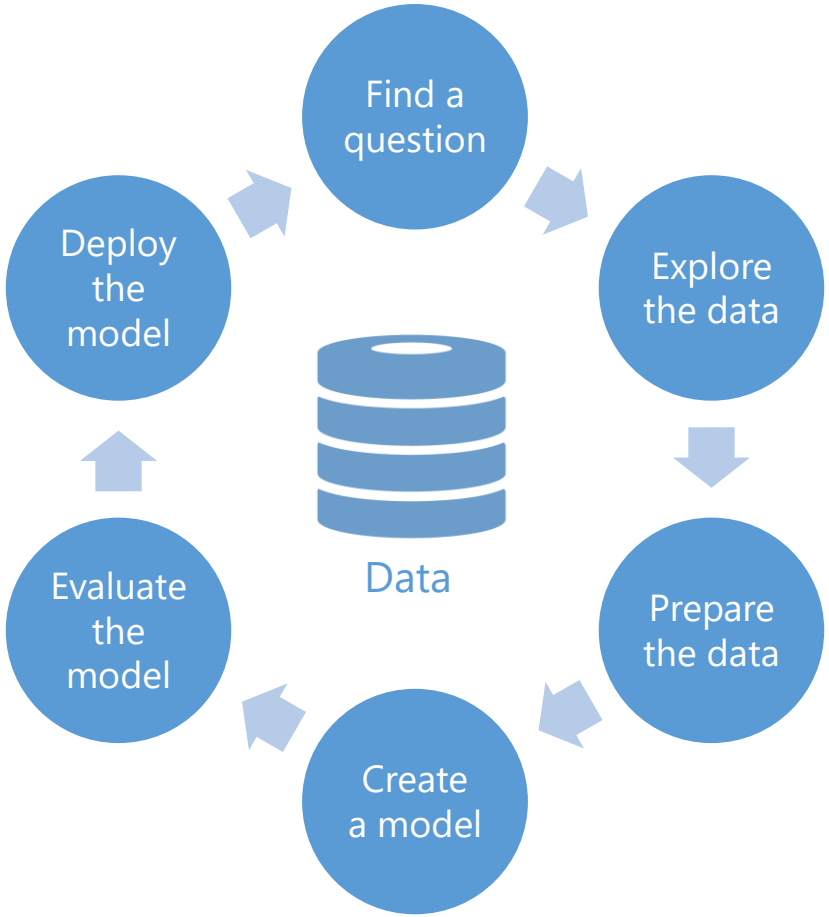
Iterative process



The Data Science Process

Iterative process

Non-sequential

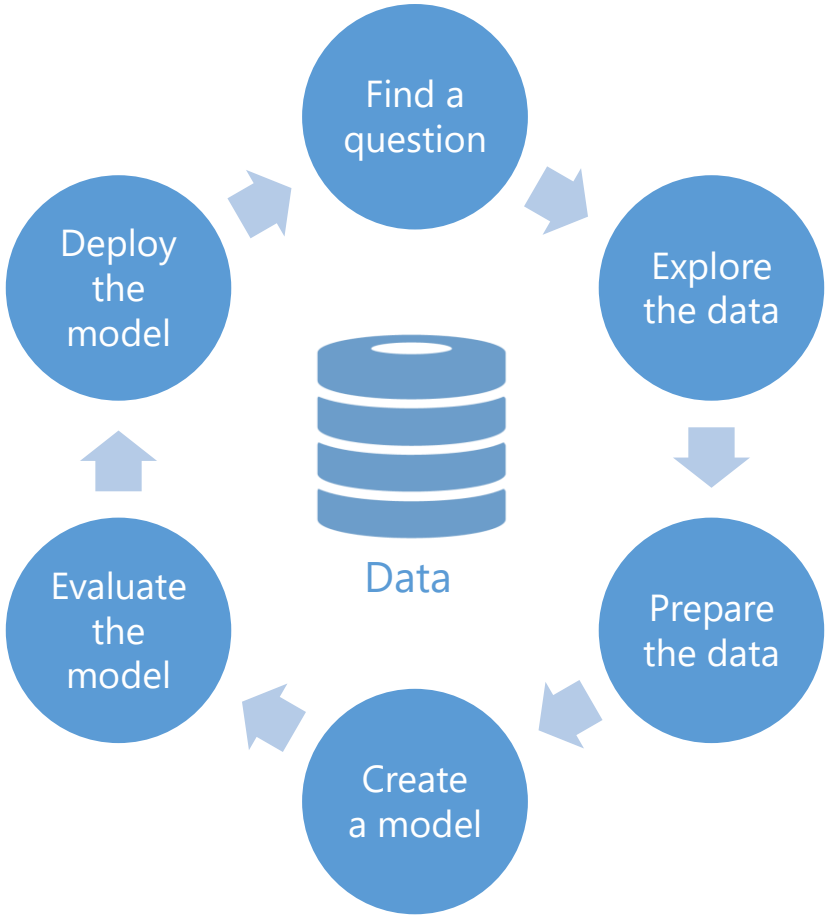


The Data Science Process

Iterative process

Non-sequential

Early termination



Data science is the discipline of converting data into actionable insight

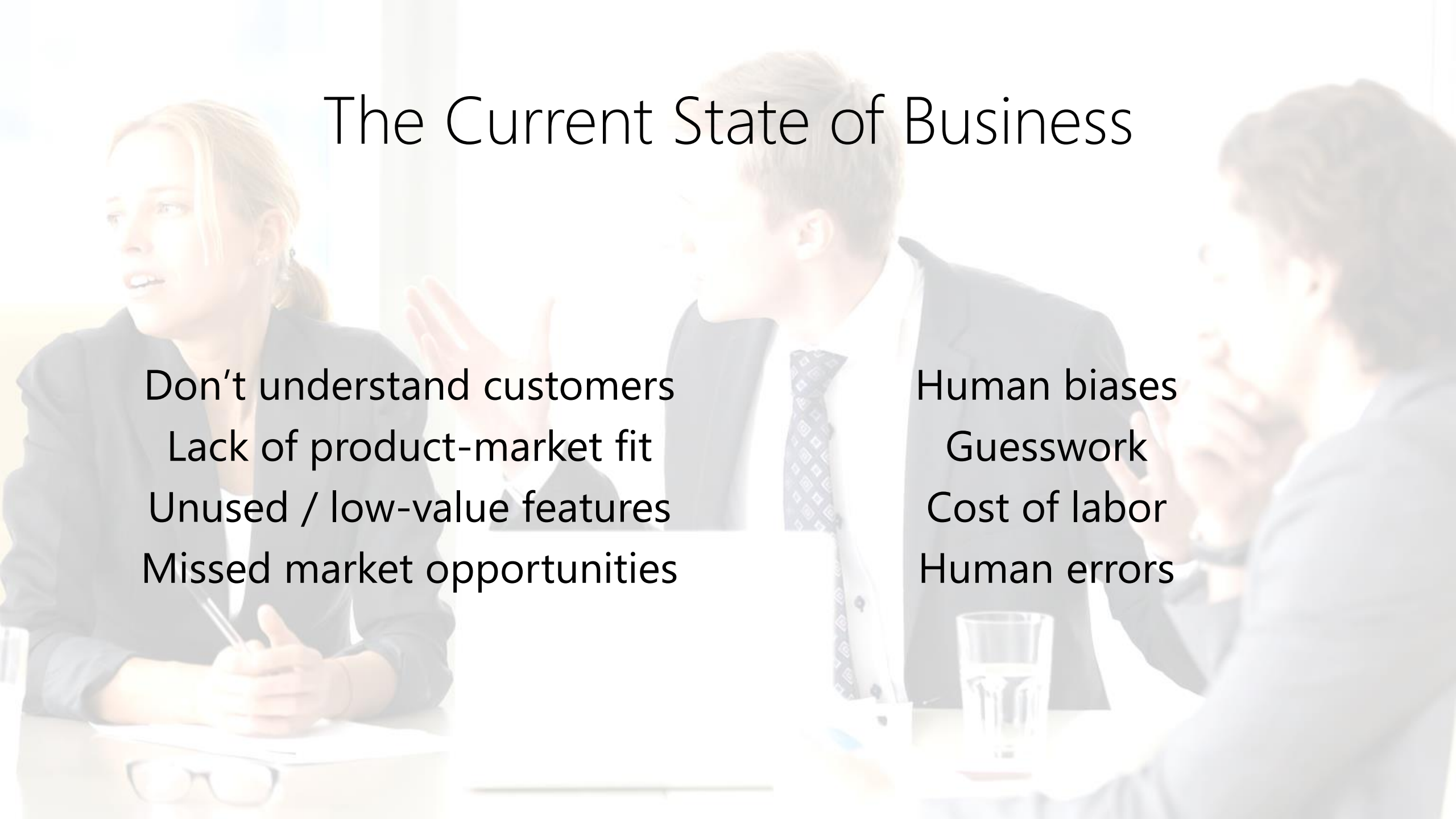
Why Is Data Science Important?

What's the problem?


CLIENT / OPTIONS
1
2
3
4


CONTRACT

The Current State of Business



Don’t understand customers

Lack of product-market fit

Unused / low-value features

Missed market opportunities

Human biases

Guesswork

Cost of labor

Human errors


BUSINESS SOLUTION
GLOBAL NETWORK
MARKETING STRATEGIC

4% higher productivity

6% higher profits

Source: https://hbr.org/2012/10/big-data-the-management-revolution


BUSINESS SOLUTION
GLOBAL NETWORK
MARKETING STRATEGIC


8,427.14
31,246.04
(+270.78)
24,413.84
(-21.87)
26,275.30
(+7.62)
30,463.58
(+15.94)
1,014.12
333.20
(+20.20)
342.71
(+2.85)
137.04
(-60.01)
60.44
(-55.90)
60.30
(-0.23)
3.65
351.38
511.22
(+45.49)
598.71
685.65
(-7.74)
937.99
(-57.22)
233.88
(-75.07)
142.09
154.12
(-7.84)
393.13
97.55
(+23.57)
33.13
(-76.72)
102.97
171.57
220.19
93.52
(-57.53)
75.41
(-19.36)
132.89
(+76.22)
142,684.54
(+12.06)
143,653.64
(+0.68)
150,028.94
(+4.44)
156,015.25
(+3.99)
50.44
726.98
741.27
48.64

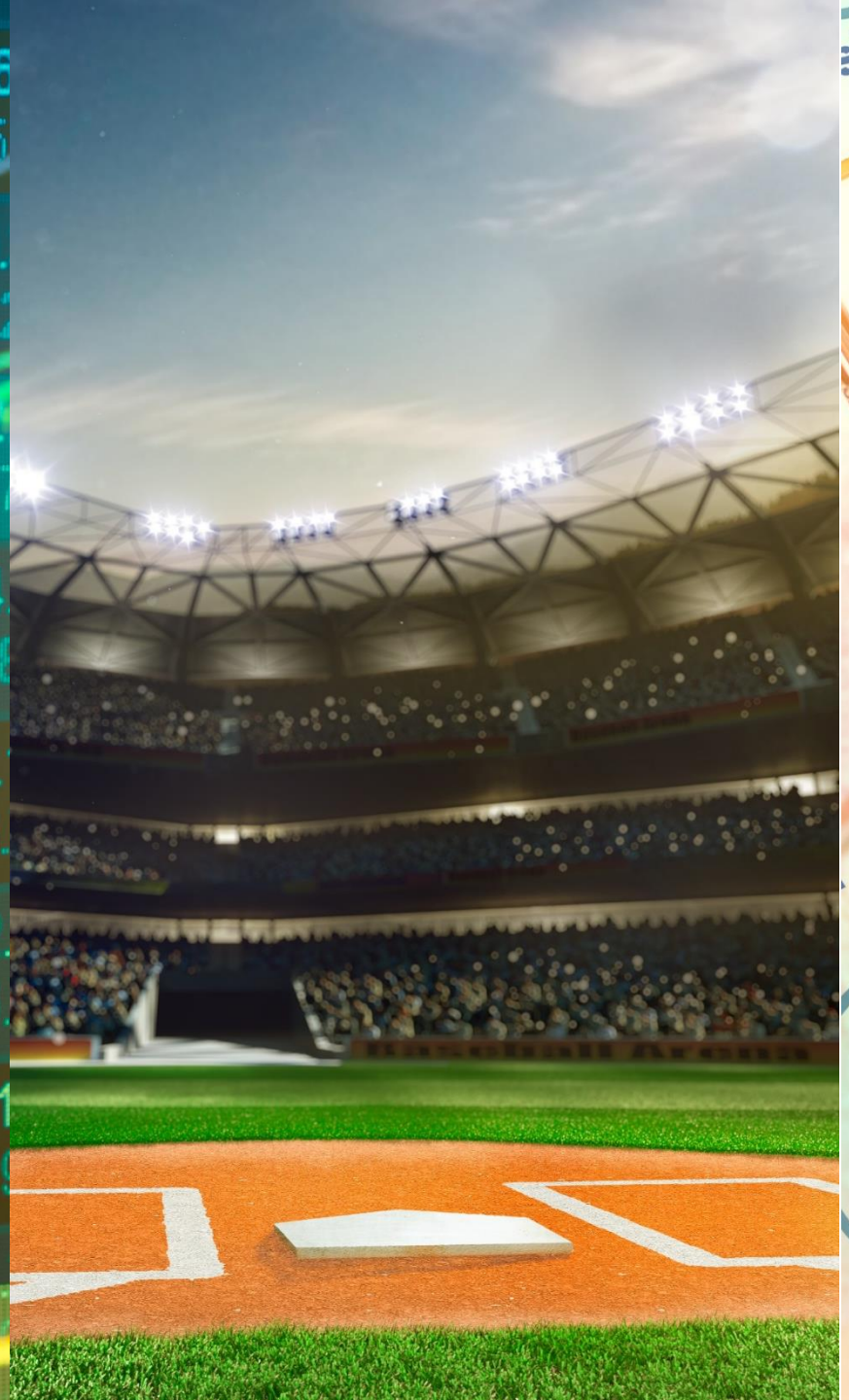

CPK METER


25


Baseball Arena
Baseball Arena
Baseball Arena


Residence
Car
Family life
Factory
InsurTECH
Medical
Travel
Hospital


24,413.84
(-21.87)
26,275.30
(+7.62)
137.04
(-60.01)
60.44
(-55.90)
598.71
685.65
142.09
(+23.57)
93.52
(-57.53)
75.41
(-19.36)
143,653.64
(+0.68)
150,028.94
(+4.44)
726.98
50.44




InsurTECH
Car
Hospital

How does data science *actually* help us build better software?

Two Main Approaches

Build intelligent software

Improve development practices

Building Intelligent Software

Retail Analysis Sample

Share Dashboard

Ask a question about the data on this dashboard

Total Stores

104

This Year's Sales

$22.05M

This Year's Sales

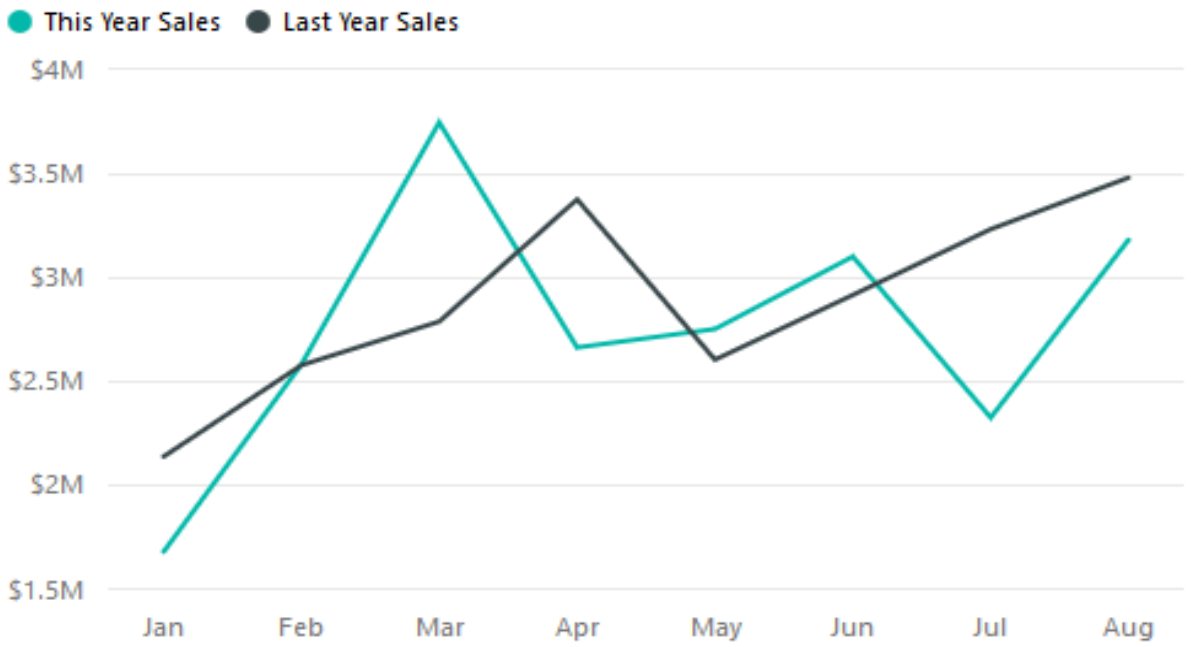
BY CHAIN

New Stores Opened This Year

10

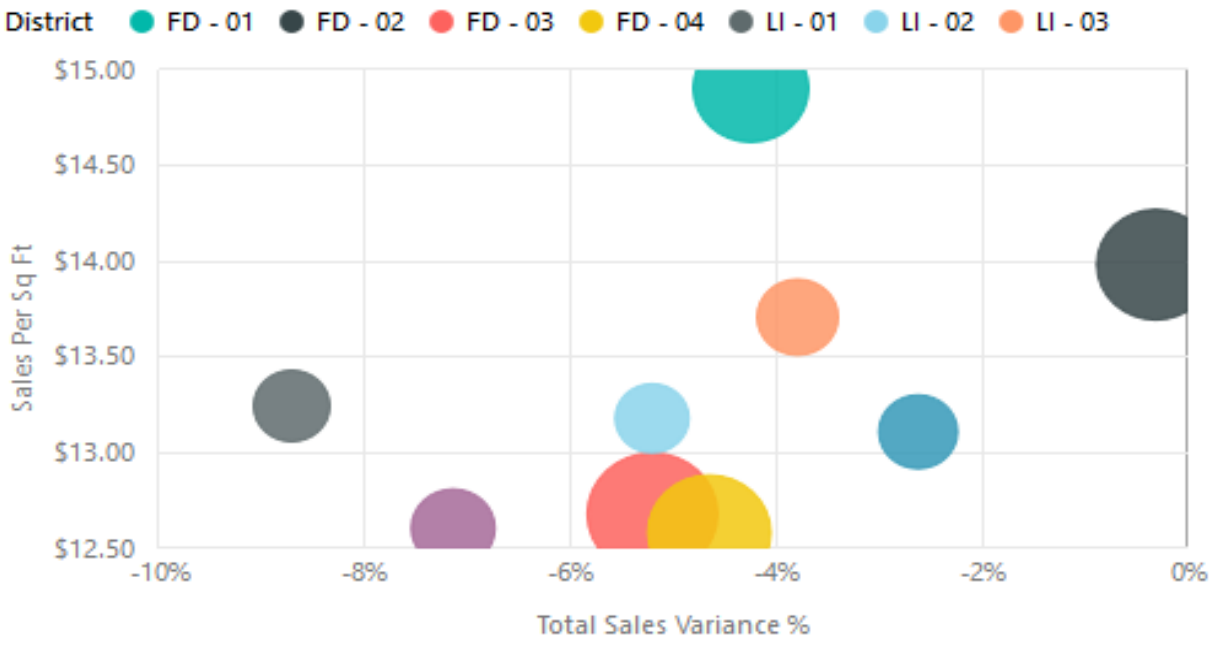
This Year's Sales, Last Year's Sales

BY FISCAL MONTH



Sales Per Sq Ft, Total Sales Variance %, This Year's Sales

BY DISTRICT

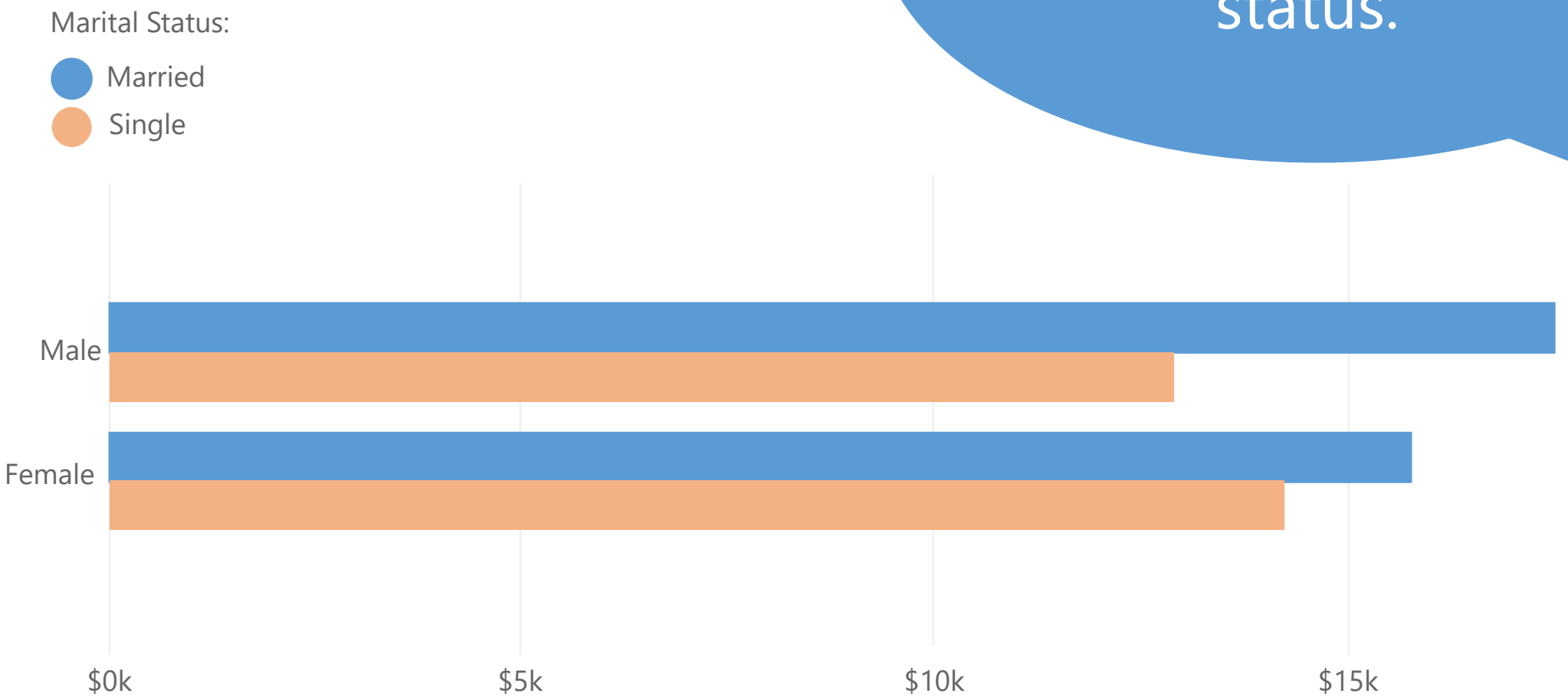


Internet Sales

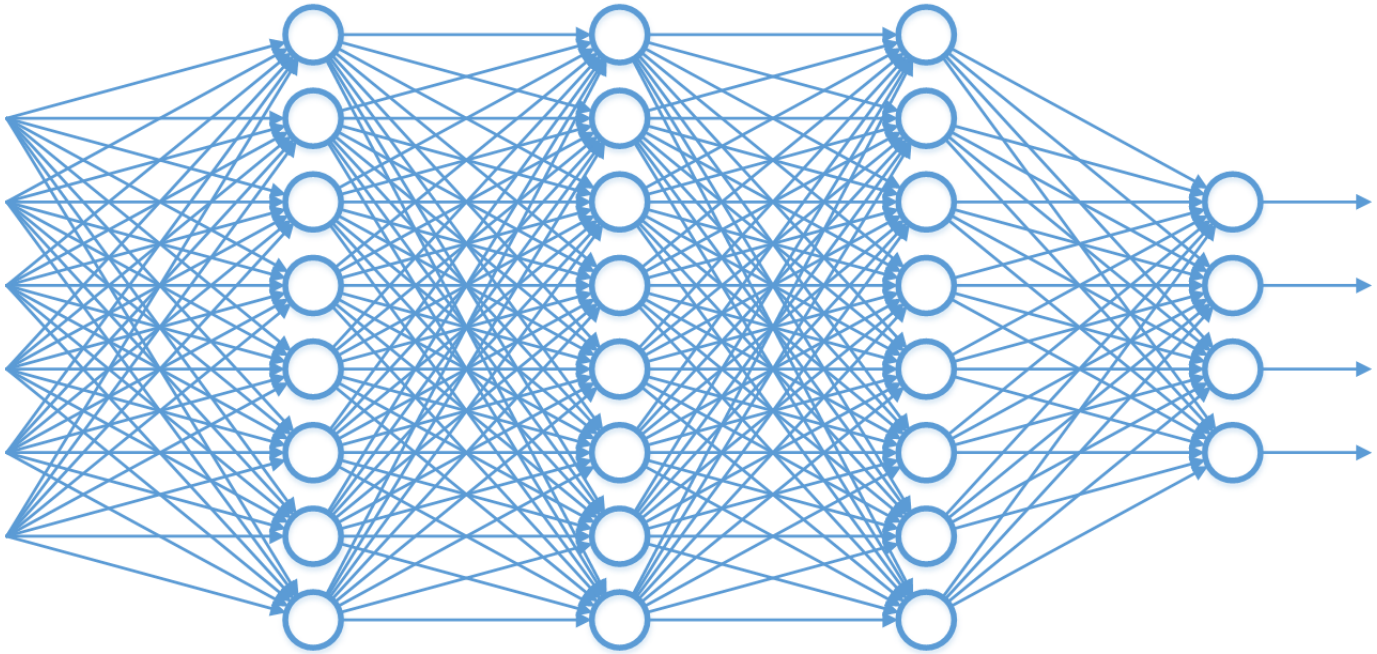
Show me sales by gender and marital status.

"Show me sales by gender and marital status."

Displaying sum of sales by gender and marital status



Machine Learning

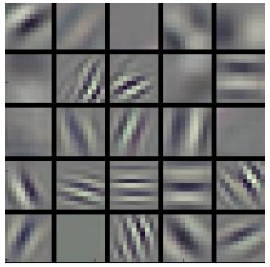


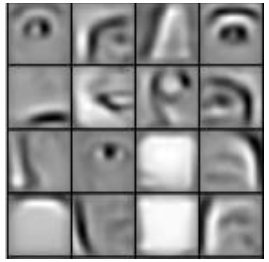
Human

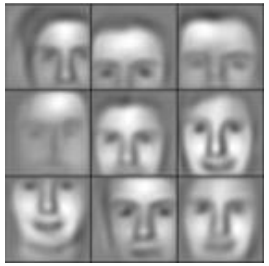
Cat

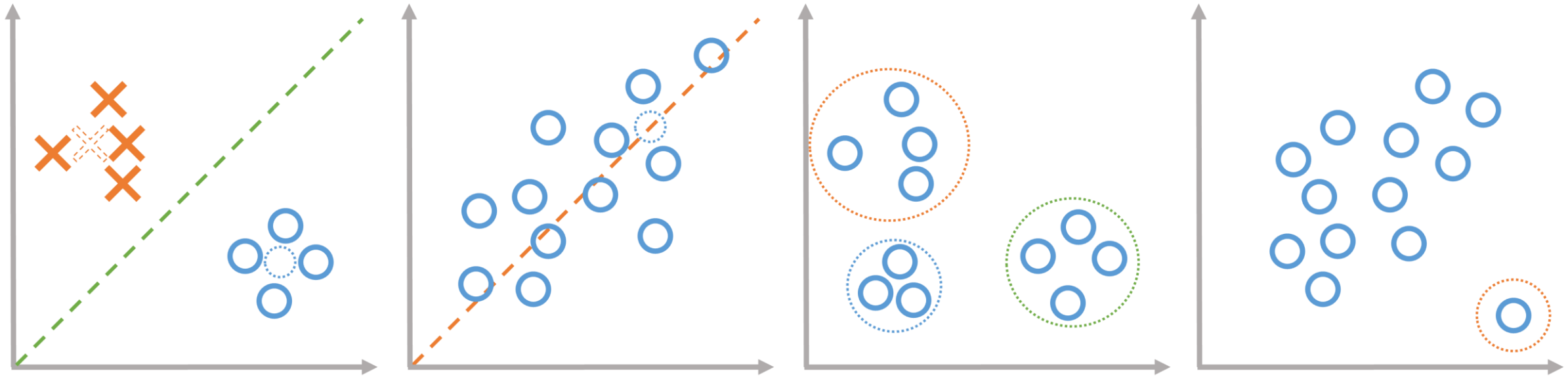
Dog

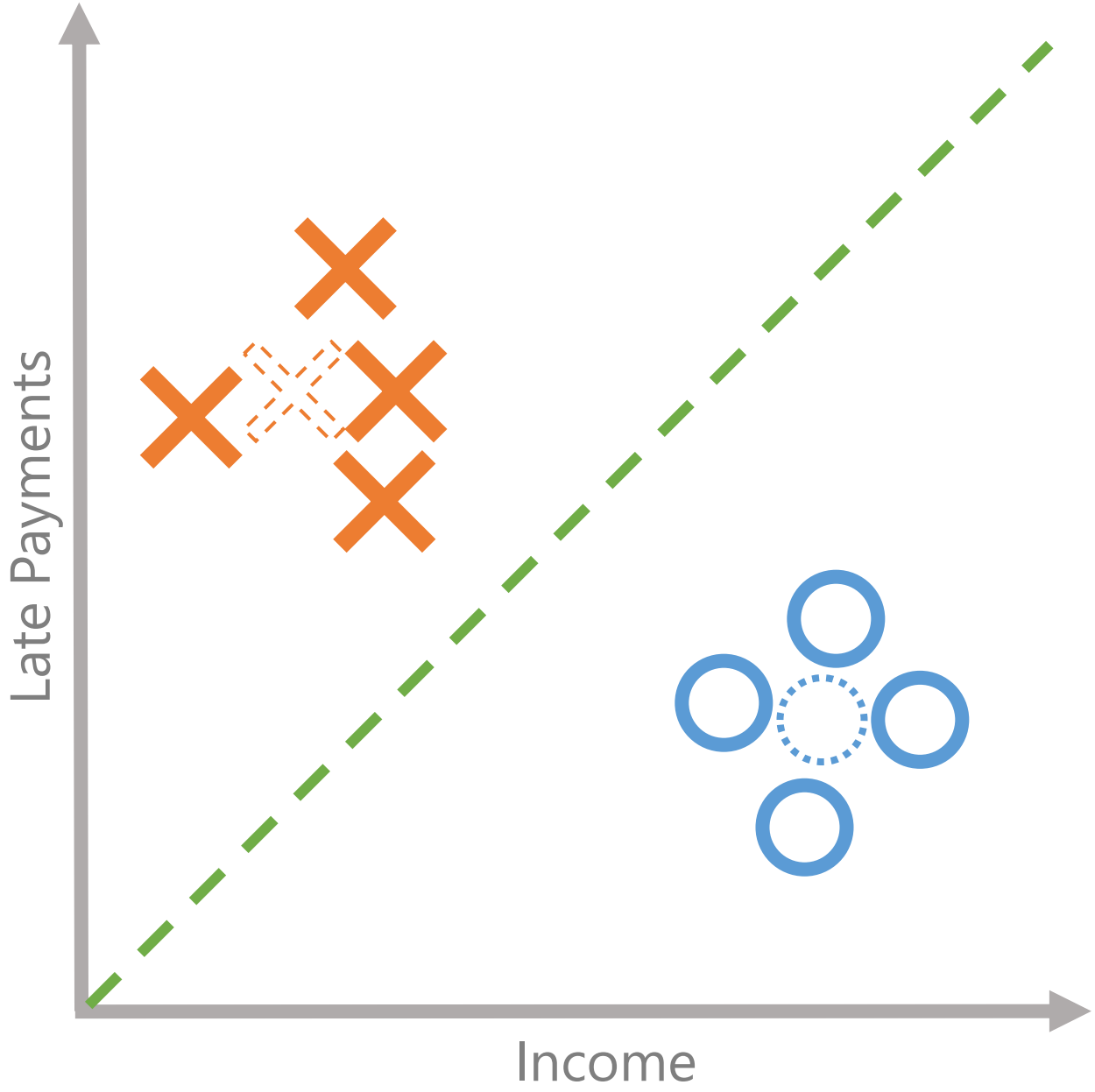
Car

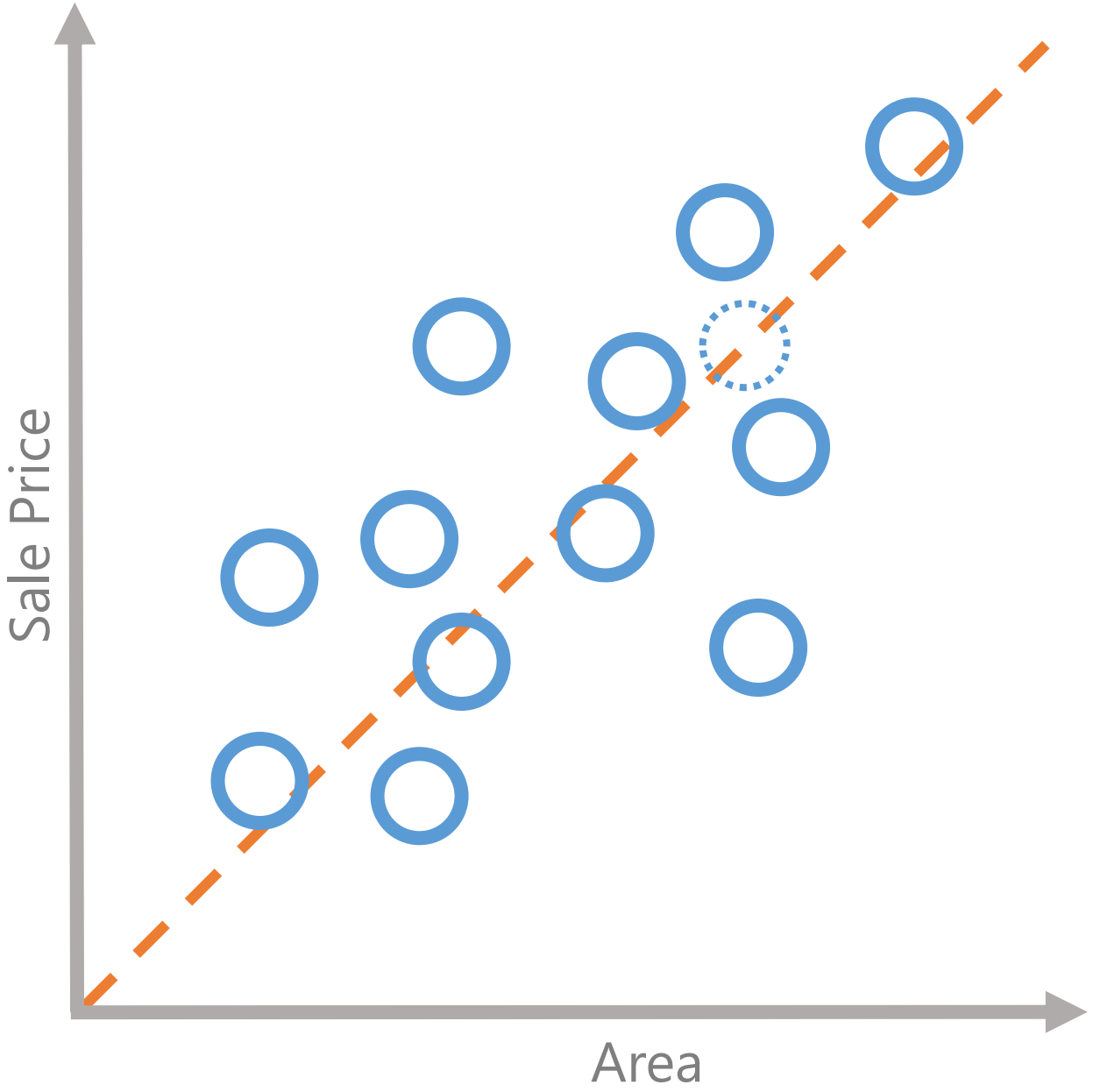


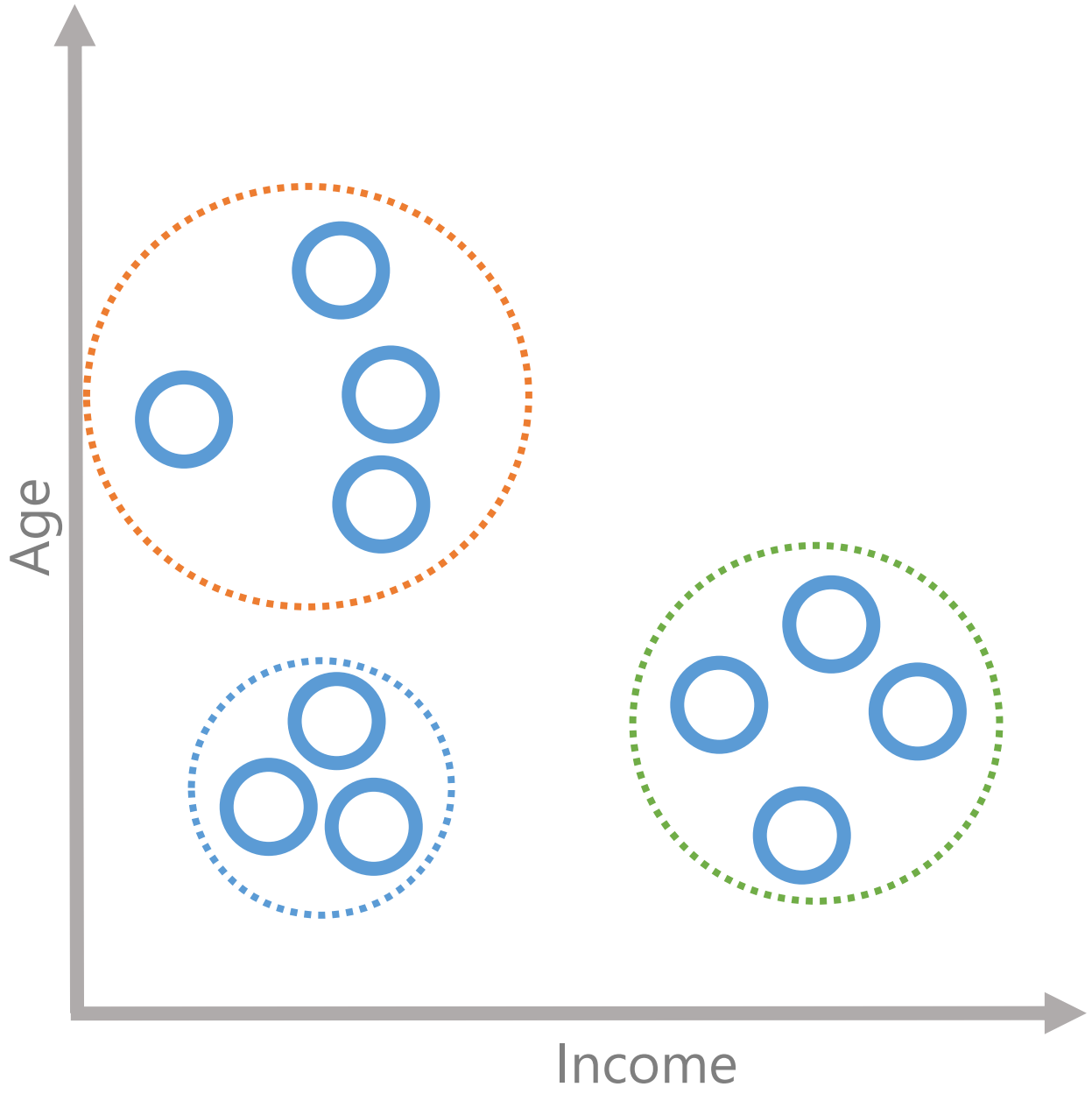


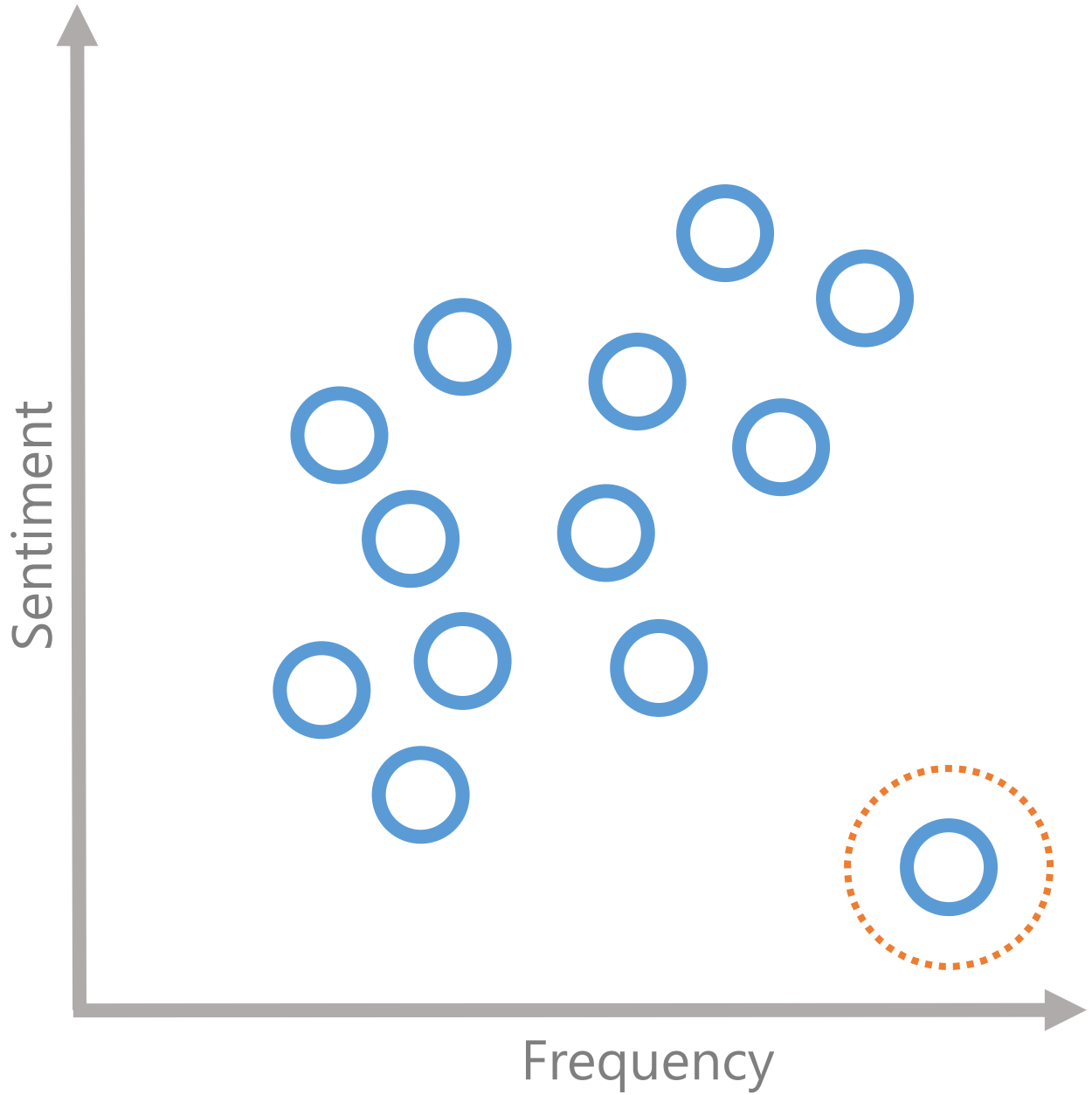



Classification
Regression
Clustering
Anomaly Detection

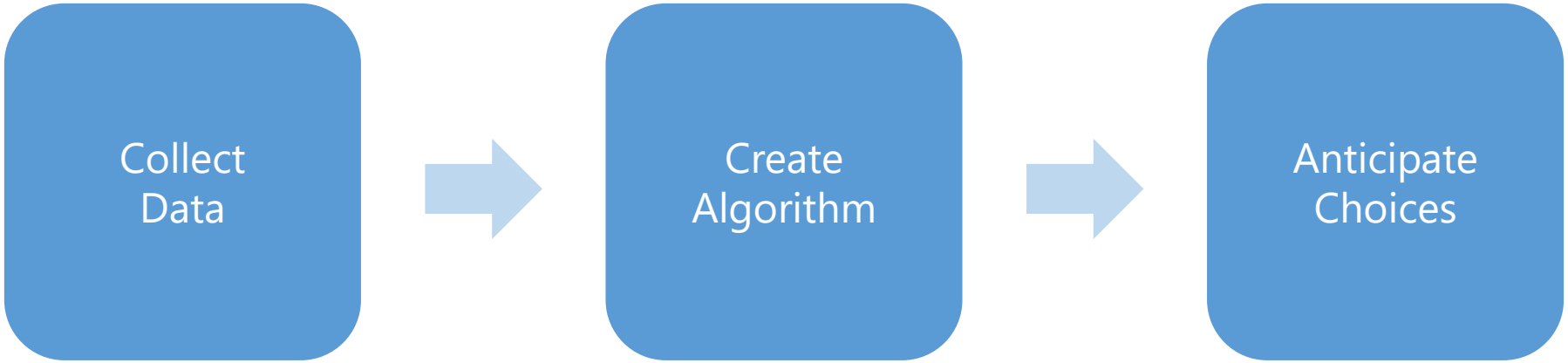

Late Payments
Income


Sale Price
Area


Age
Income


Sentiment
Frequency

Anticipatory Design



Browse
Radio

YOUR LIBRARY
Your Daily Mix
Recently Played
Songs
Albums
Artists
Stations
Local Files
Videos
Podcasts

PLAYLISTS
Discover Weekly
Closer
One Man Puppet Show
A Puppet Named Julio
This Will Destroy You
Explosions in the Sky

New Playlist

Your Daily Mixes

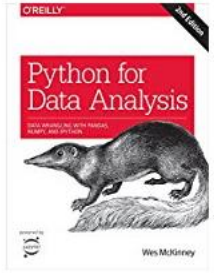
Play the music you love, without the effort. Packed with your favorites and new discoveries.

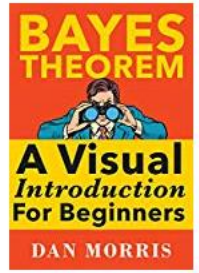
Daily Mix 1
Bonobo, The Album Leaf, Thievery Corporation and more
MADE FOR RENZE1577

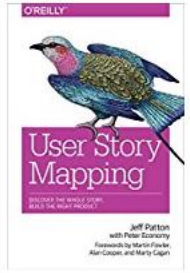
Daily Mix 2
Epic45, Sigur Rós, múm and more
MADE FOR RENZE1577

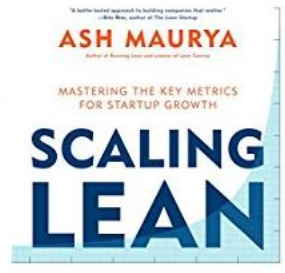
Daily Mix 3
Joey Fehrenbach, The Album Leaf, The Abbasi Brothers and more
MADE FOR RENZE1577

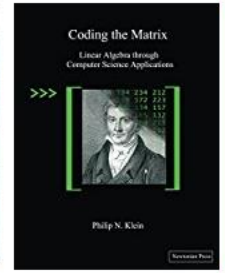
Recommendations for you in Kindle Store



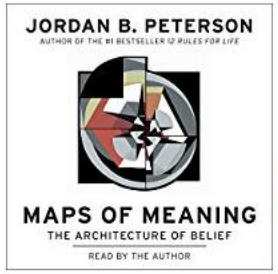


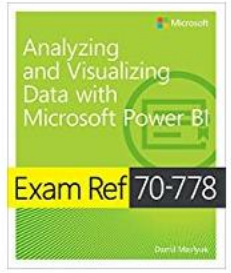


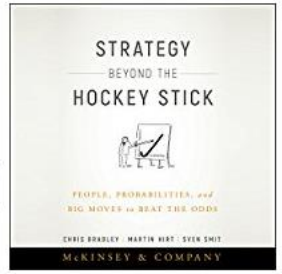


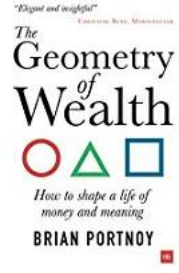


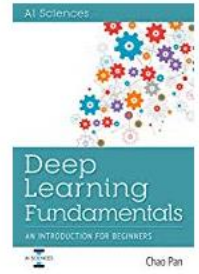
New for you See more

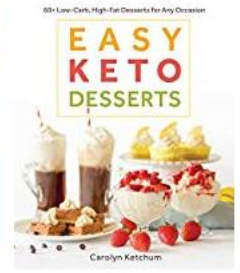












Buy it again


LOREM IPSUM

Lorem Ipsum
Dolor Sit Amet

SEARCH

MENU

- Suspendisse
- Velit Sit Amet
- Pellentesque
- Volutpat
- Pretium Nulla
- Eget Turpis

PROJECTS

- Nunc venenatis
- Maecenas
- Curabitur

Dashboard › Ornare Orci Velit

Consectetur Adipiscing
Donec sollicitudin, ipsum non

$1,310

Nunc et metus magna

Aenean Porta
Donec sollicitudin, ipsum non

$24,815

Nunc et metus magna

Aenean Porta
Donec sollicitudin, ipsum non

94.2%

Nunc et metus magna

Aenean Porta
Donec sollicitudin, ipsum non

Nunc et metus magna

Integer

Nam ultrices

Phasellus vulputate

Nullam Ac Leo Vitae

Vivamus

NULLAM SAPIEN

Mon	Tue	Wed	Thu	Fri	Sat	Sun
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2

Lorem Ipsum
Nunc faucibus dignissim nibh

Cras Pharetra

Aliquam
Nunc faucibus dignissim nibh

Donec orci nulla	872
Lobortis non nisi quis	650

Phasellus Interdum
Dignissim nibh

Bibendum

Nunc Leo Tortor
Nunc faucibus dignissim nibh

Felis ligula rutrum orci 2016

Phasellus interdum ligula eget turpis

Sed id egestas 2016

salesforce.com - Enterpr | Opportunity: Gamma Servic | salesforce.com - Enterprise E

na15.salesforce.com/home/home.jsp

salesforce.com - | Power BI | opportunity Gamma

Home | Chatter | Contacts | Accounts | Leads | Opportunities | Metrics | Reports | Dashboards | +

Gerald Frazer

Wednesday September 16, 2015

Show Feed

Dashboard

Refresh

Customize Page

As of 9/16/2015 5:29 PM. Displaying data as Gerald Frazer

Current Quarter Sales Predictions

Sales Won	$3.3MM		↑
Team Estimate	$6.3MM		↑
Predicted Sales	$5MM		↓
Predicted Upside	$1.3MM		↑
Predicted Risk	$2.2MM		↑

AlpineMetrics | Powered by Microsoft Cortana Analytics Suite

Predicted Risk by Team

Sort: By Highest Value

Changes Over | 7 Days

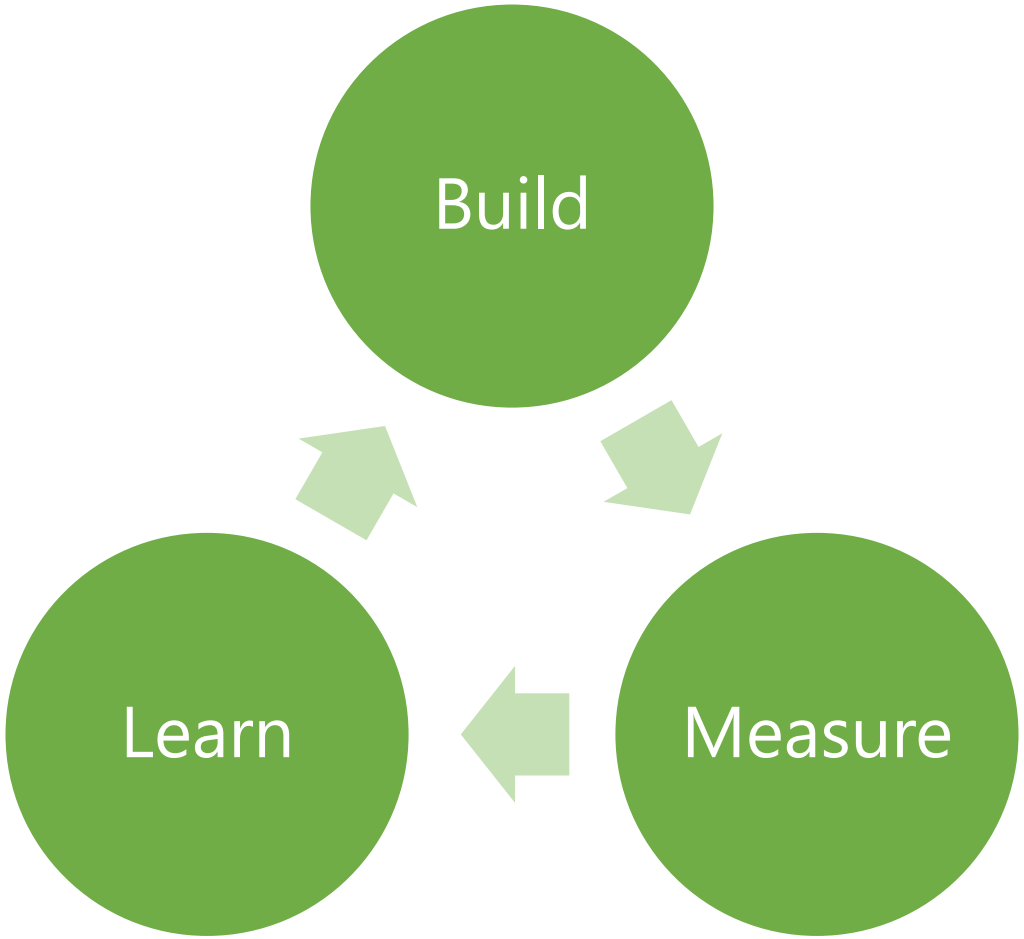
Chat

Ask me anything

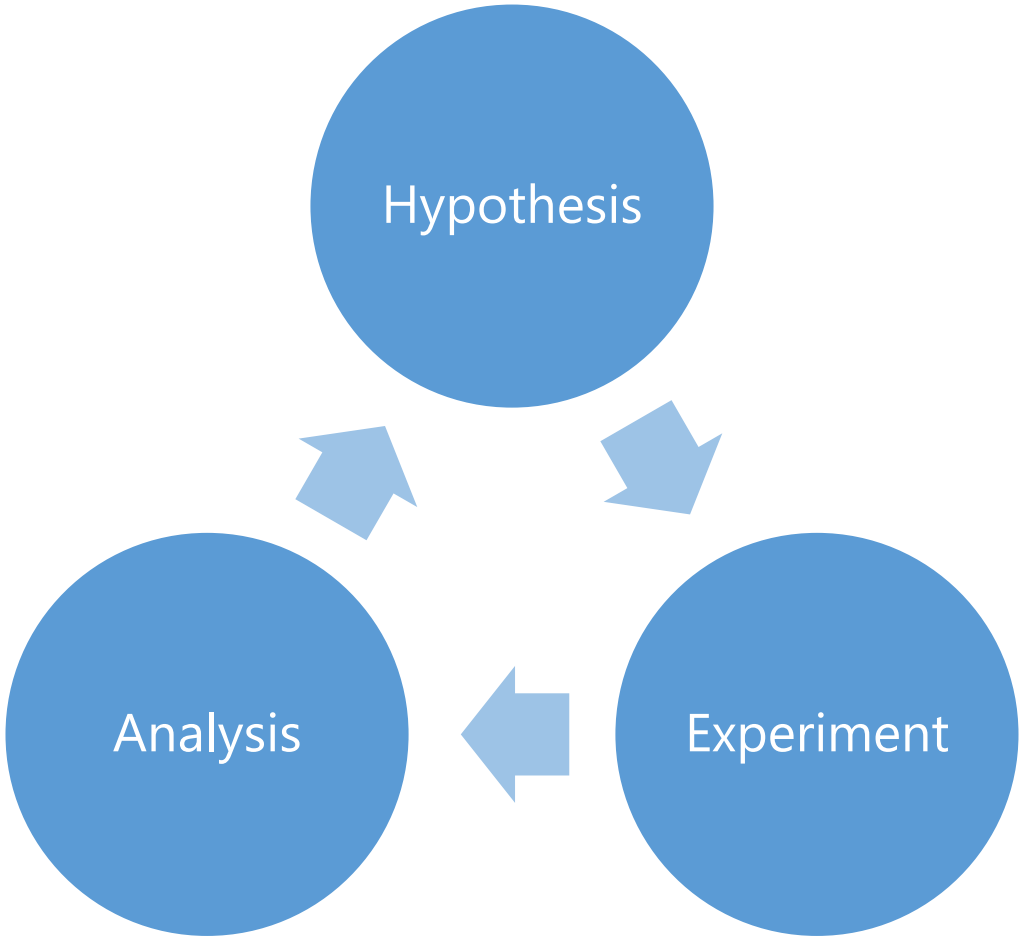
5:42 PM
9/16/2015

Improving Development Practices

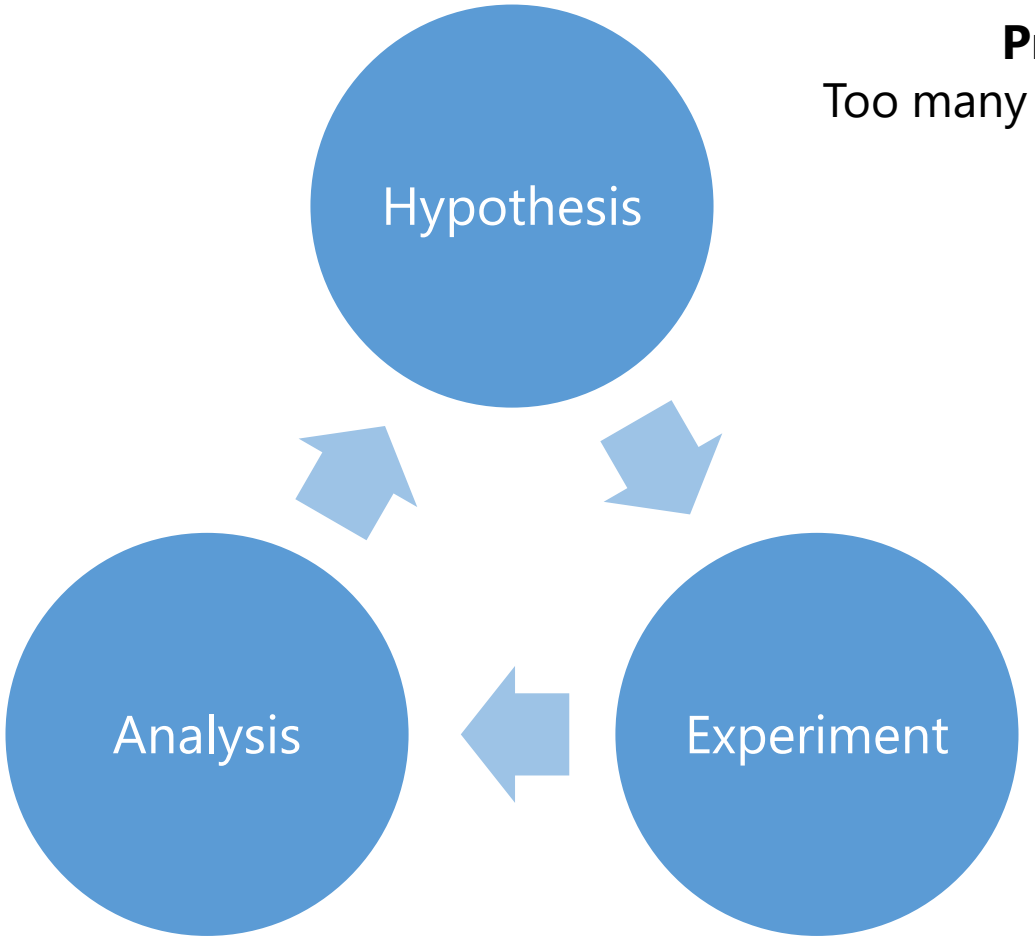
Data-Driven Decision Making



Hypothesis-Driven Development

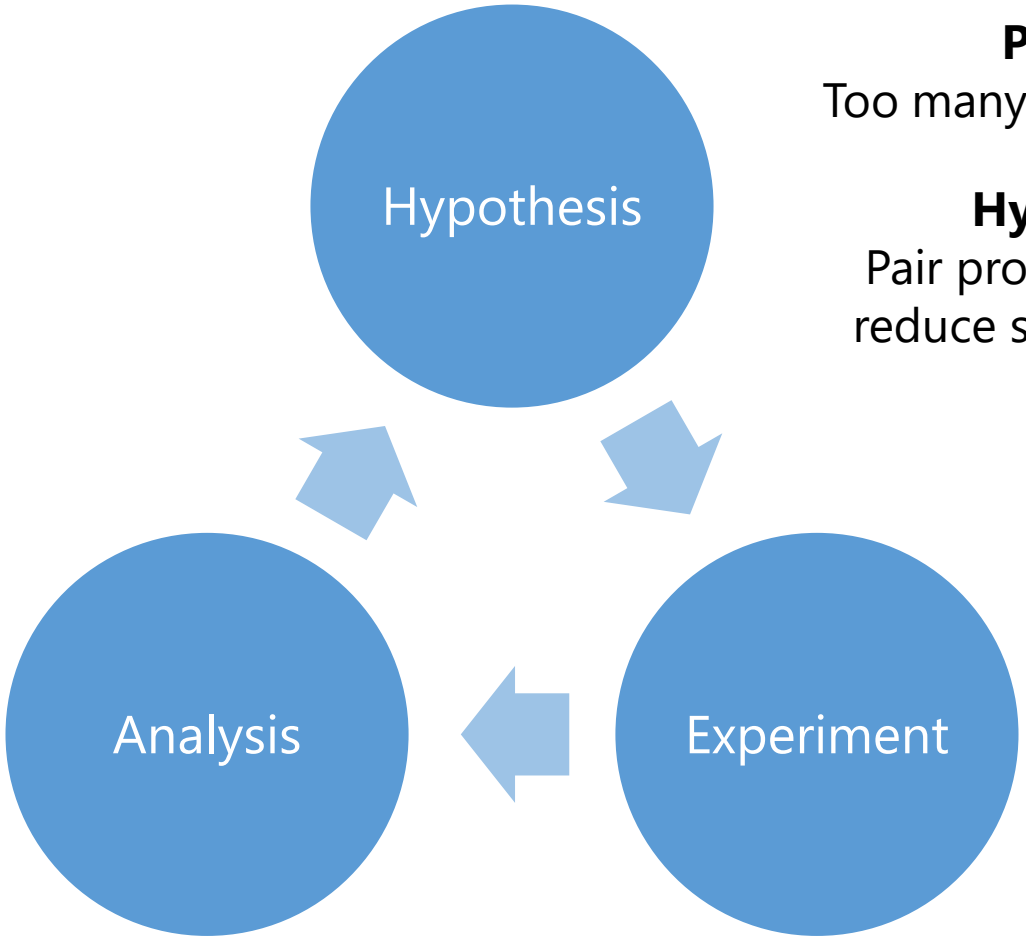


Hypothesis-Driven Development



Problem:
Too many software defects

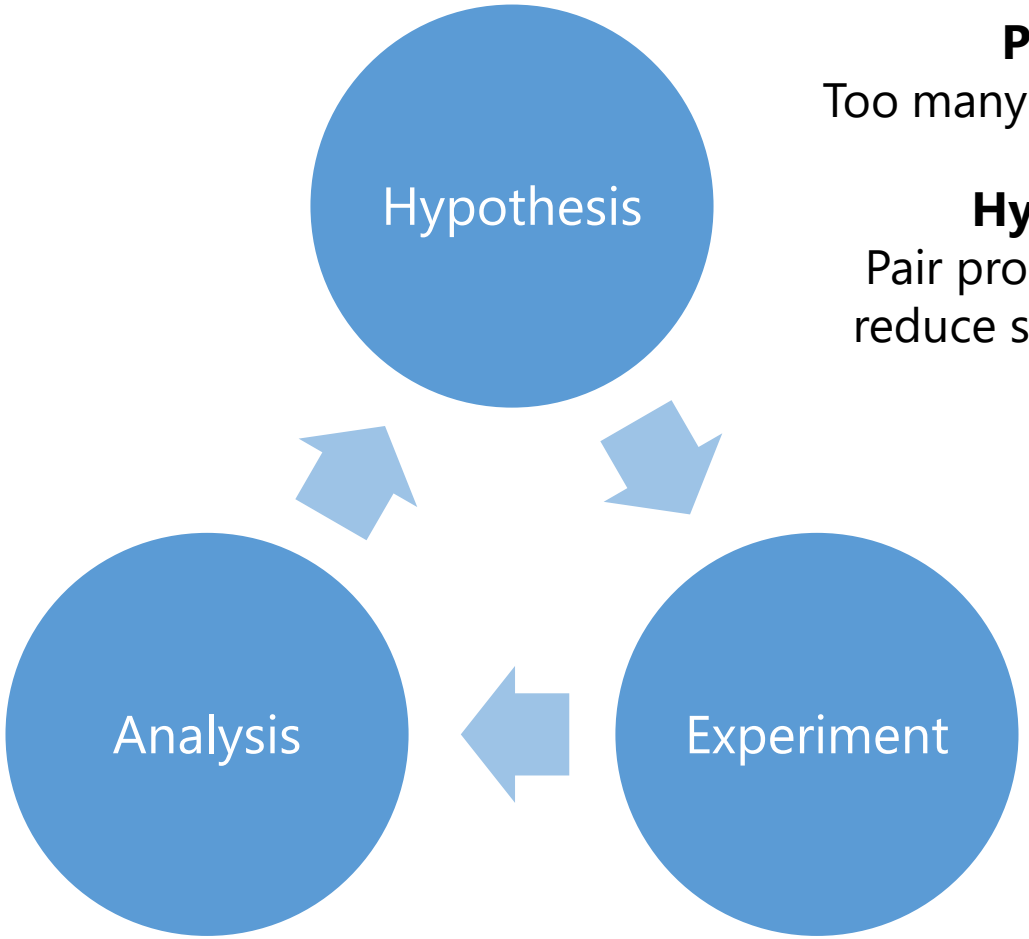
Hypothesis-Driven Development



Problem:
Too many software defects

Hypothesis:
Pair programming will reduce software defects

Hypothesis-Driven Development

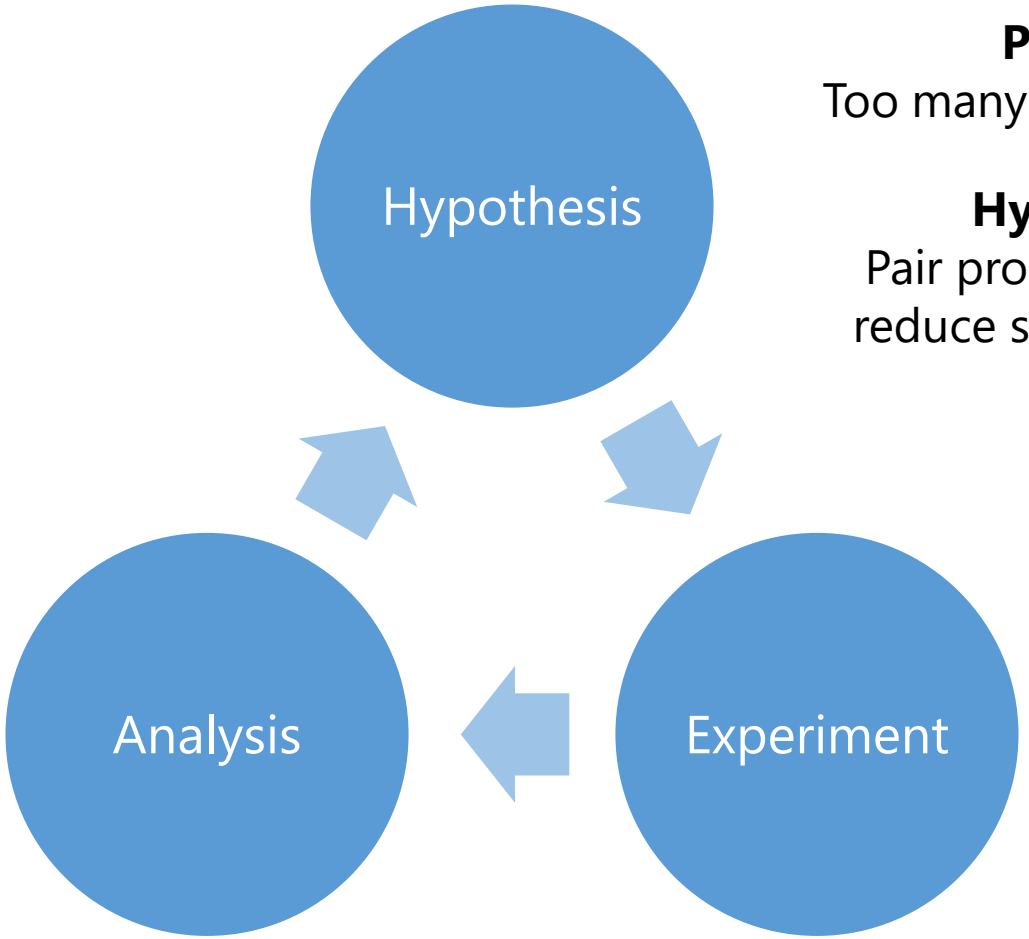


Problem:
Too many software defects

Hypothesis:
Pair programming will reduce software defects

Experiment:
Pair for 4 sprints and track defects

Hypothesis-Driven Development



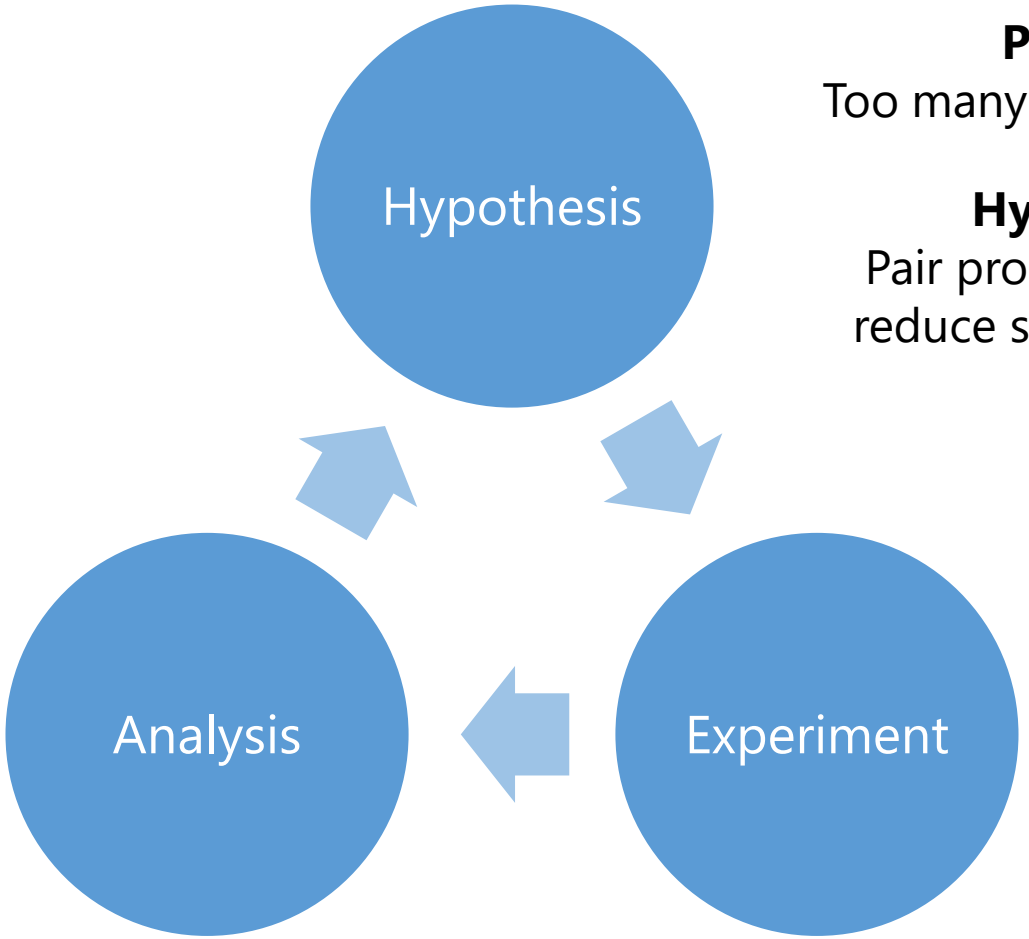
Problem:
Too many software defects

Hypothesis:
Pair programming will reduce software defects

Experiment:
Pair for 4 sprints and track defects

Level of Support:
20% decrease in defects

Hypothesis-Driven Development



Problem:
Too many software defects

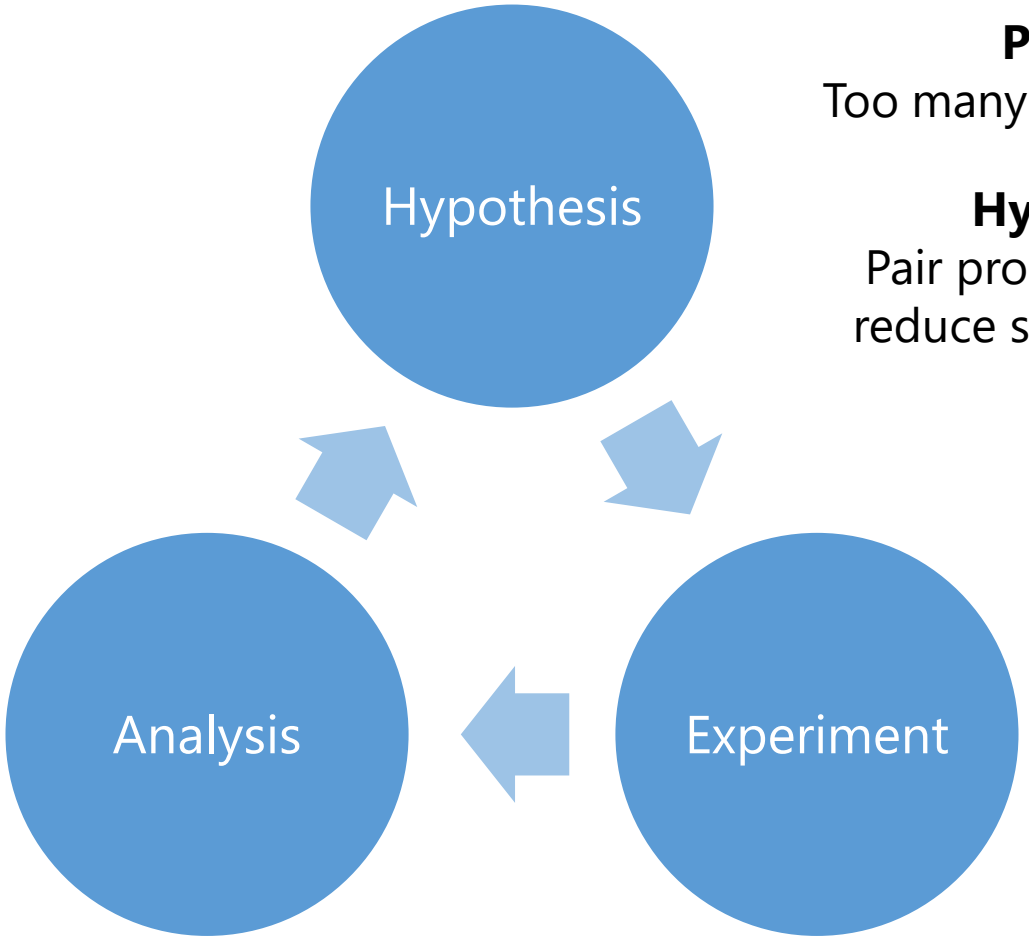
Hypothesis:
Pair programming will reduce software defects

Experiment:
Pair for 4 sprints and track defects

Level of Support:
20% decrease in defects

Analysis:
Defects decreased by 30% per sprint

Hypothesis-Driven Development



Problem:
Too many software defects

Hypothesis:
Pair programming will reduce software defects

Experiment:
Pair for 4 sprints and track defects

Level of Support:
20% decrease in defects

Analysis:
Defects decreased by 30% per sprint

Decision:
Continue pairing

Hypothesis Stories

<Hypothesis>
We assume that <hypothesis>
Will result in<outcome>
We will have succeeded when <measurable result>

Hypothesis Stories

Pair Programming Hypothesis
We assume that pair programming
Will result in less software defects
We will have succeeded when we have seen a 20% or greater decrease in defects after 4 sprints.

Not all hypotheses are testable
or should be tested.

A/B Testing

A/B Testing


50 % visitors
see variation A
Variation A
23%
conversion
50 % visitors
see variation B
Variation B
11%
conversion

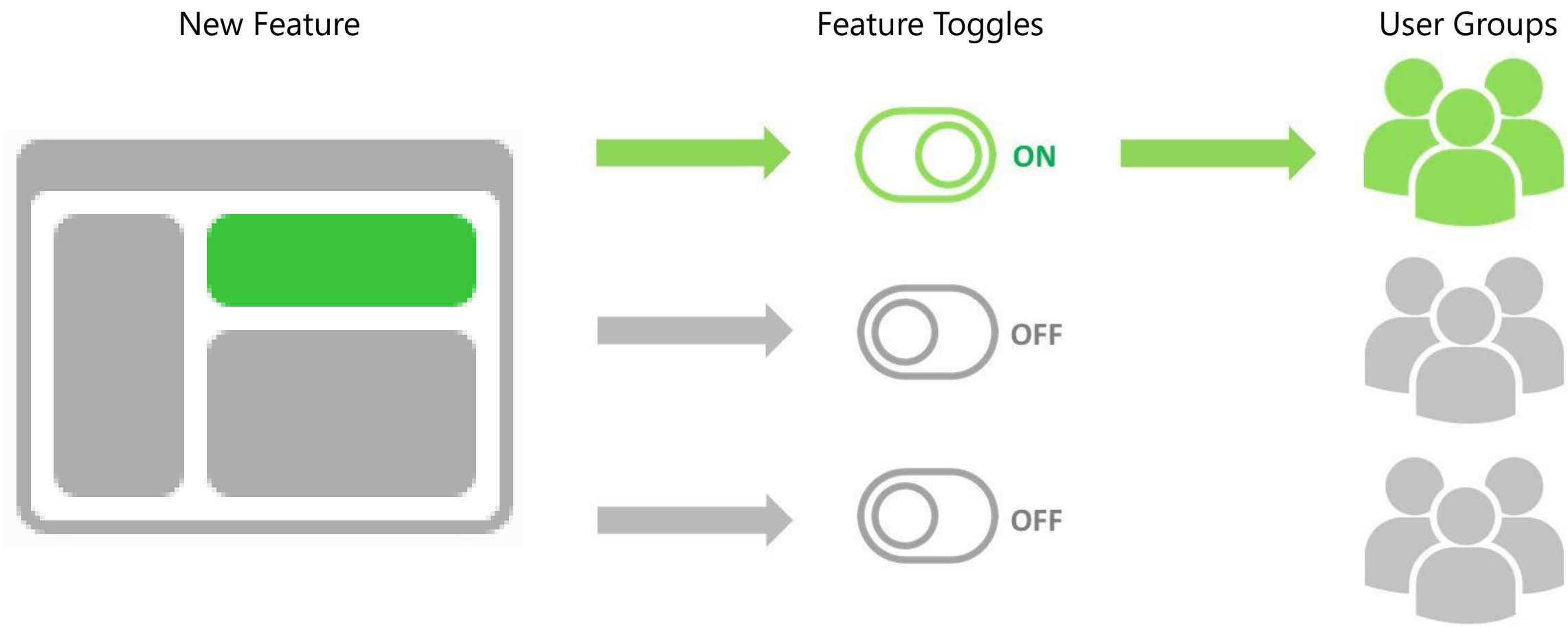
Feature Toggles

New Feature

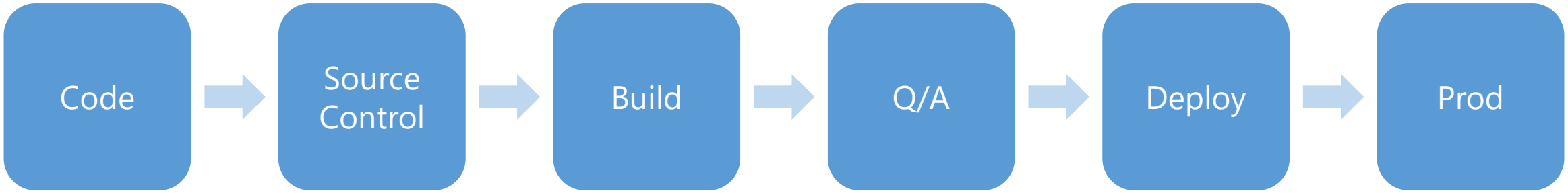
Feature Toggles

User Groups

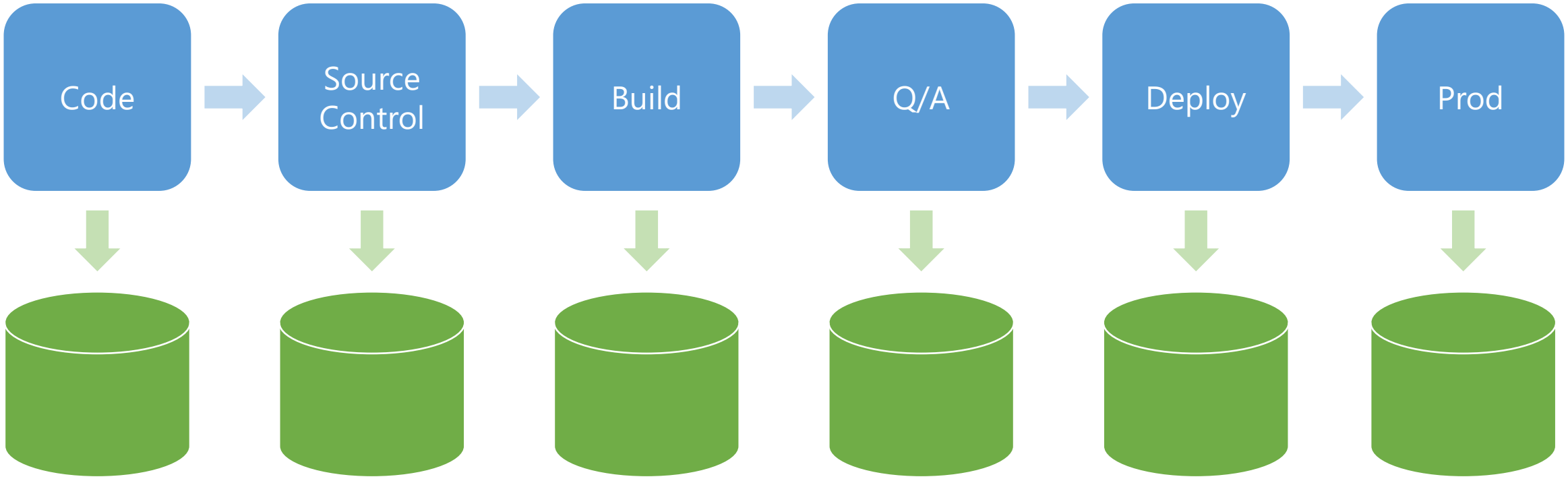
Feature Toggles



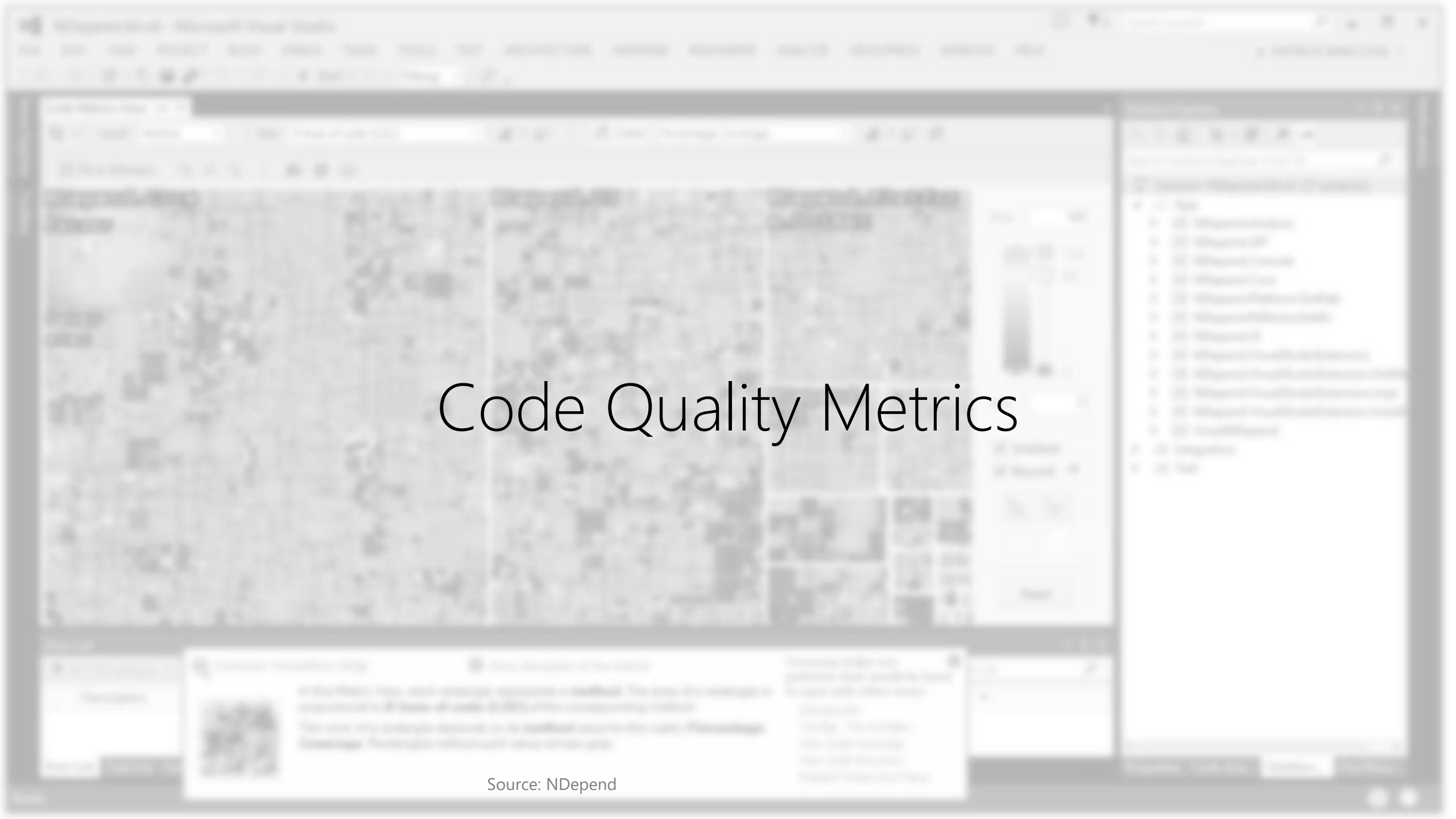
DevOps Pipeline



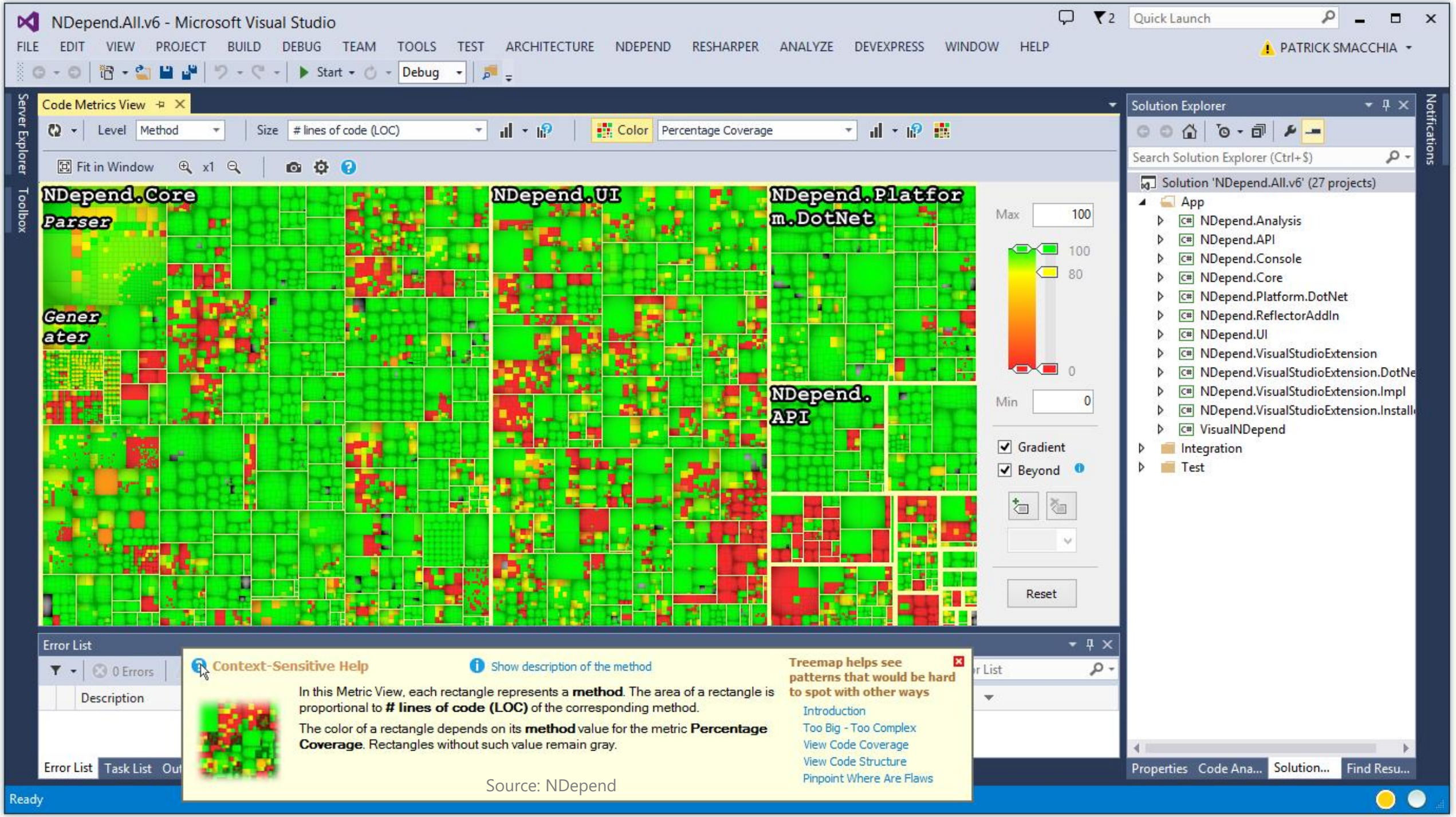
DevOps Pipeline



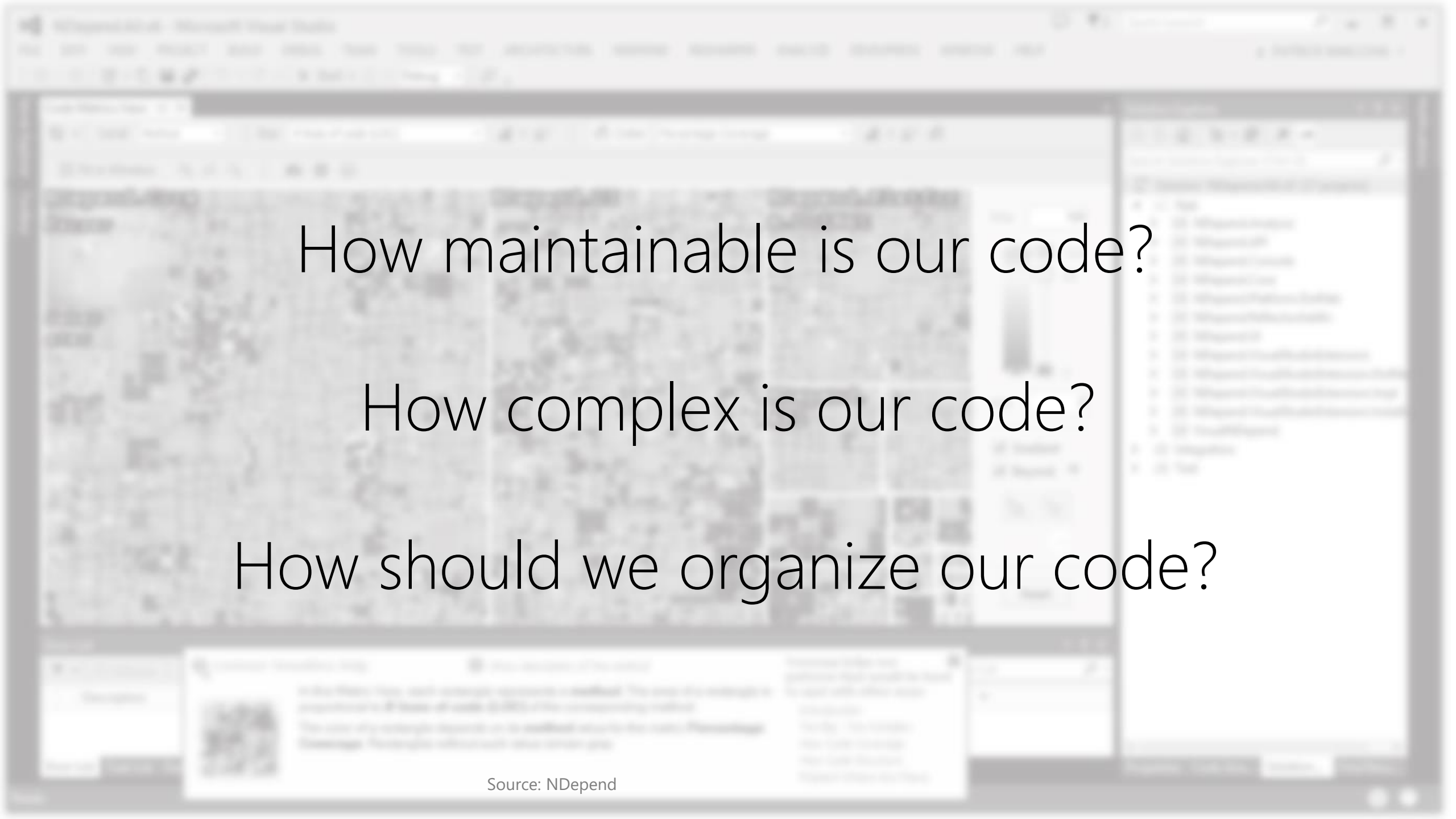
Code Quality Metrics

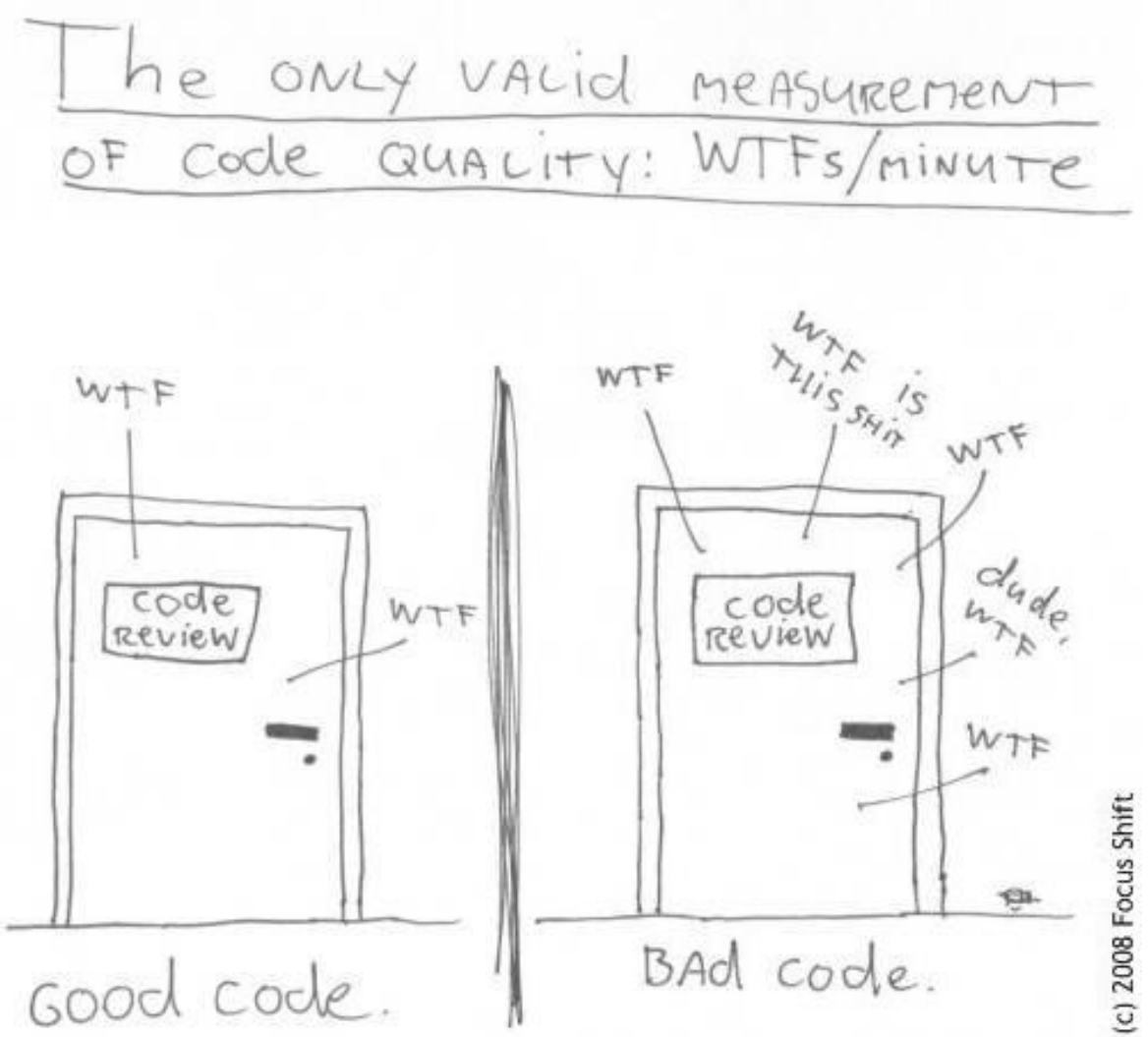


Source: NDepend



Source: NDepend


How maintainable is our code?
How complex is our code?
How should we organize our code?
Source: NDepend



Source: http://www.bernie-eng.com/blog/wp-content/uploads/2011/03/code.jpg

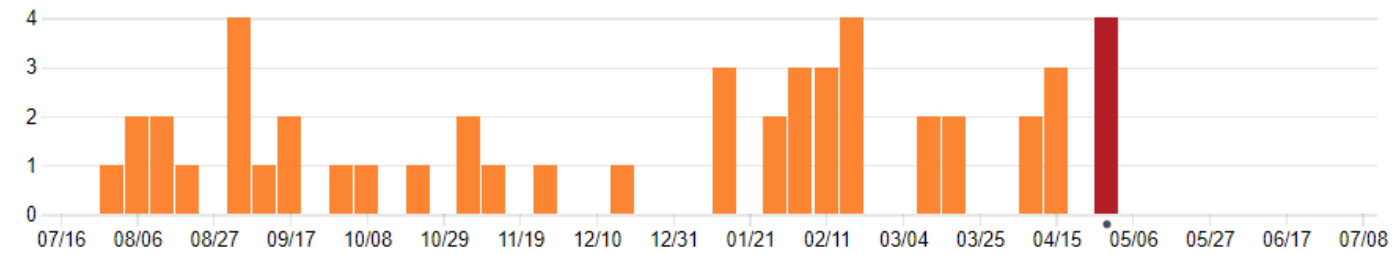
Source Control Metrics

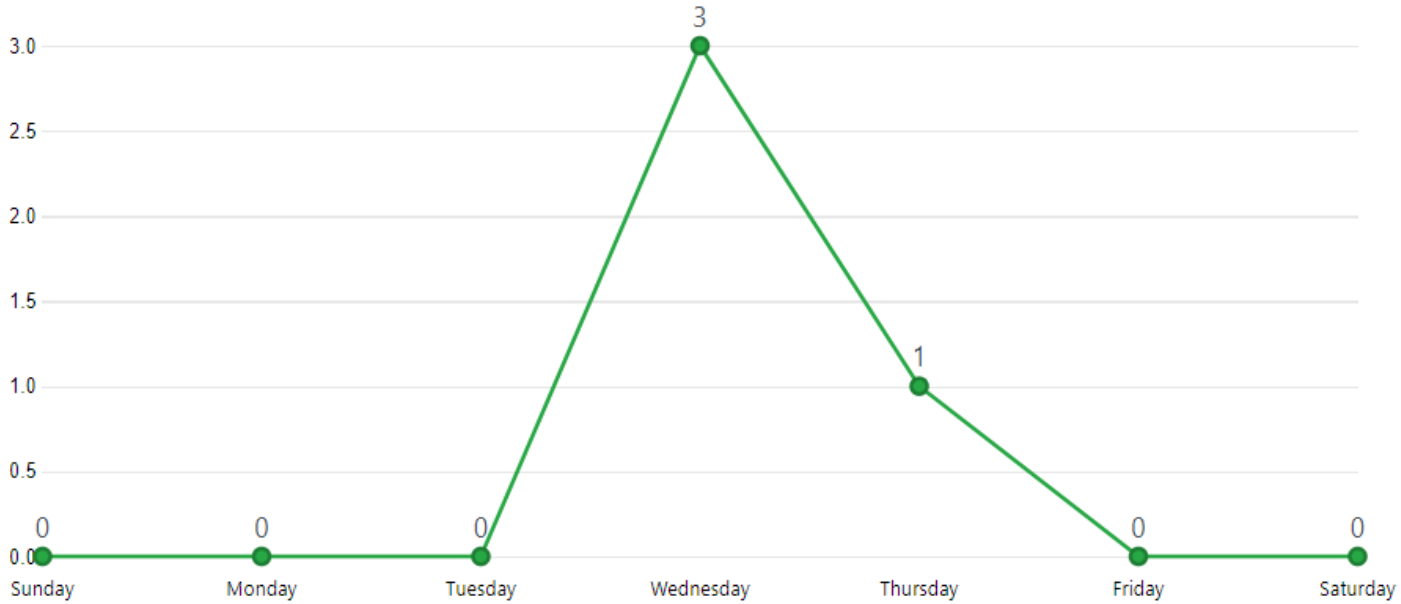
matthewrenze / matthewrenze.com

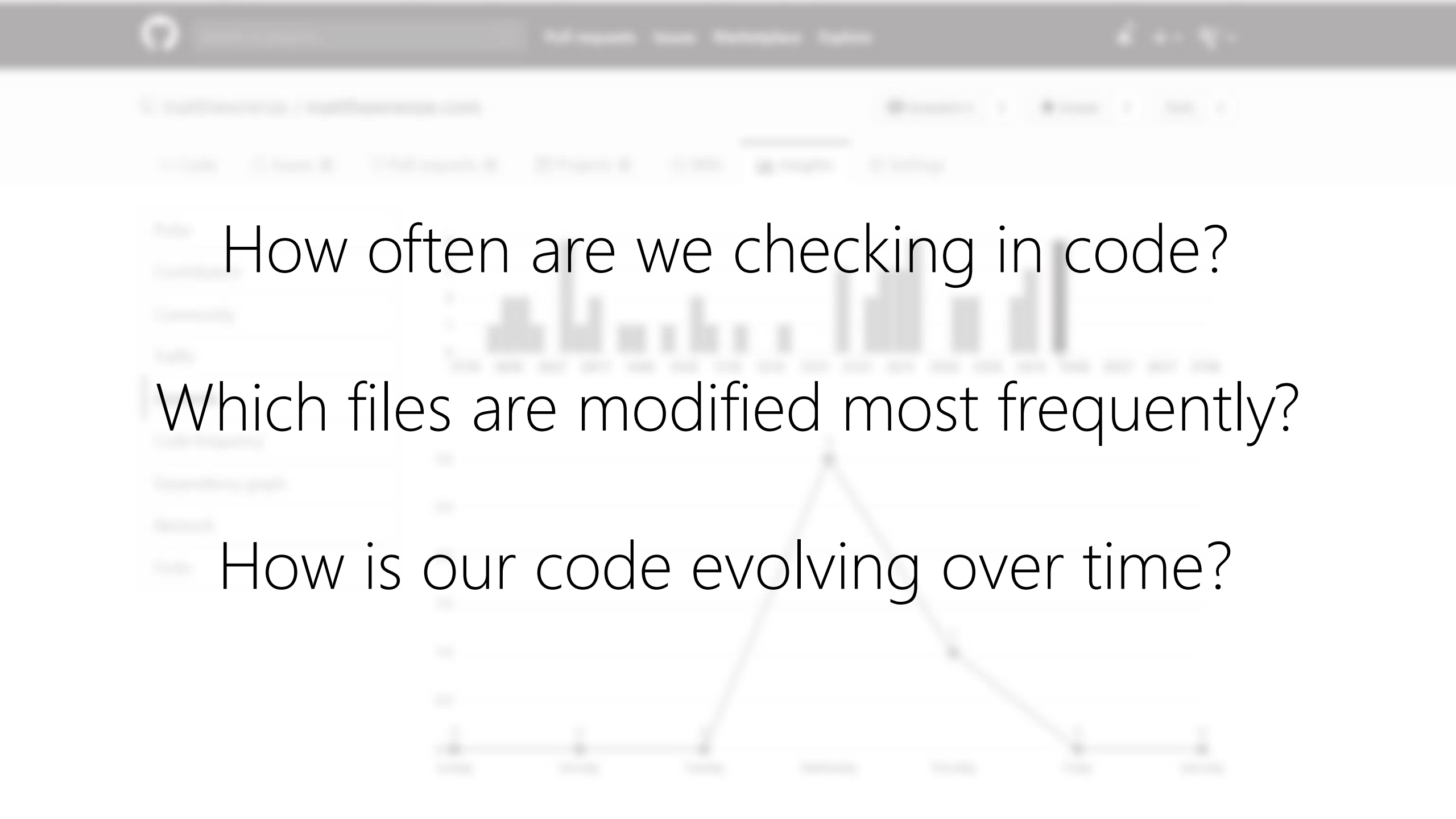
Unwatch 3 | Unstar 3 | Fork 2

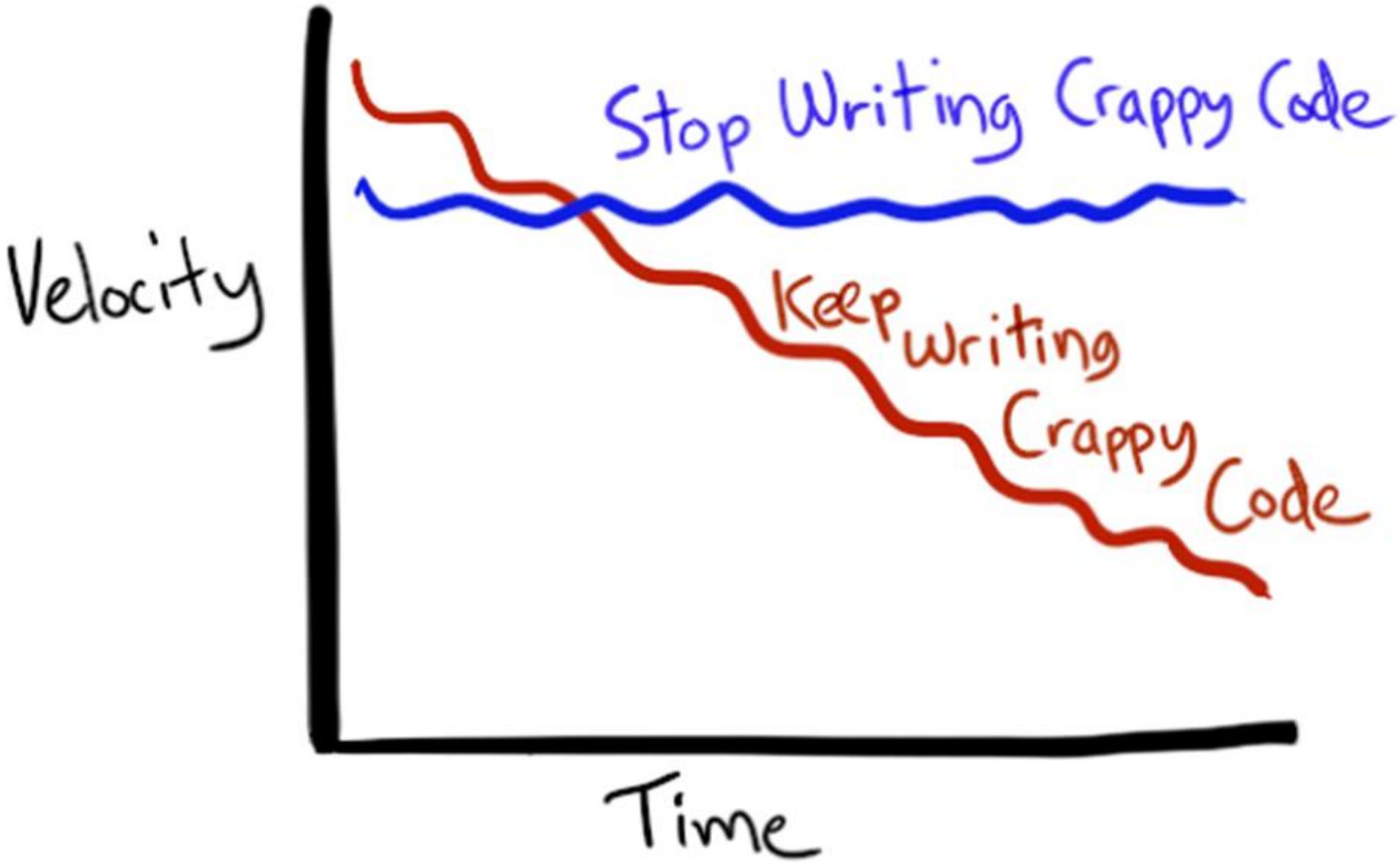
Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

- Pulse
- Contributors
- Community
- Traffic
- **Commits**
- Code frequency
- Dependency graph
- Network
- Forks






How often are we checking in code?
Which files are modified most frequently?
How is our code evolving over time?



Source: http://talkingincode.com/wp-content/uploads/2014/12/crapcode.png

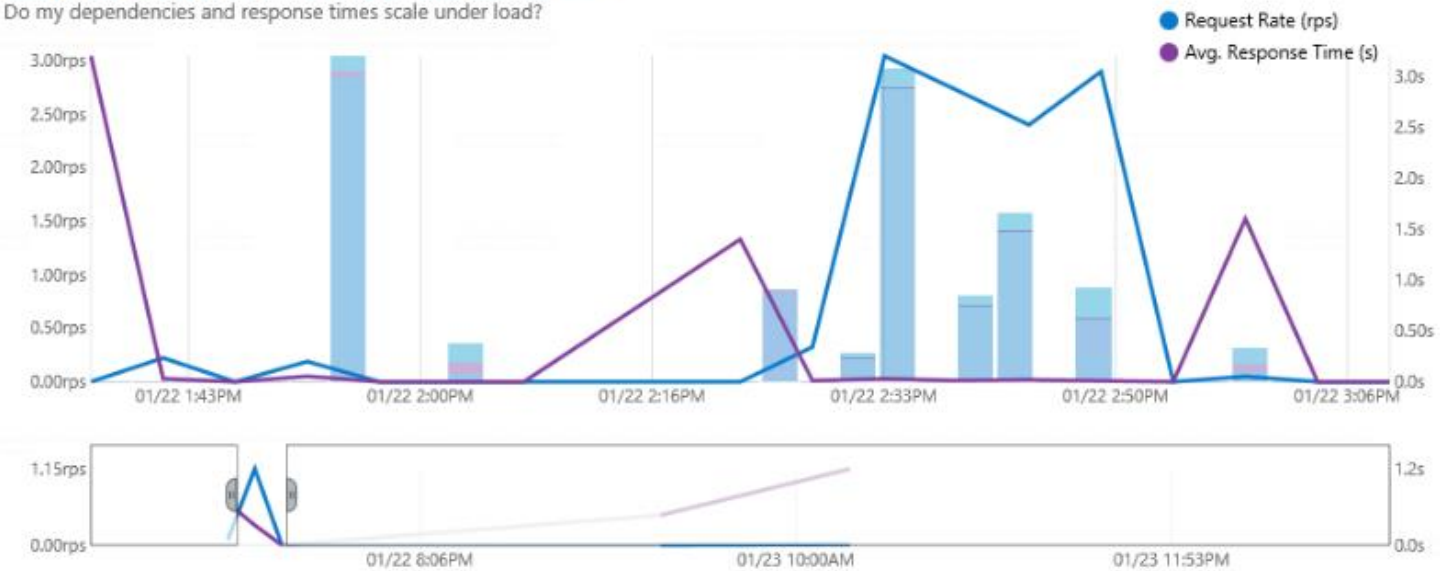
Software Telemetry

FF-Intranet-Test ▾ Server Performance

Selected date range: Last 48 hours

Response Time and Load vs. Dependencies

Do my dependencies and response times scale under load?



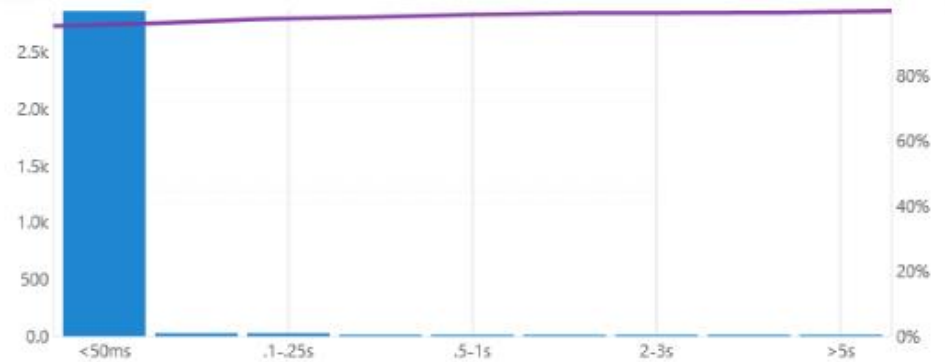
Resource	Type	Calls	Weight	Avg	Reqs/Sec
(localdb)\v11.0\FabrikamFiber-Express	Sql	2.58k	86.72%	20ms	0.3932
(localdb)\v11.0\master	Sql	18.0	2.1%	68ms	0.0027
https://dc.services.visualstudio.com/	HTTP	30.0	11.18%	219ms	0.0045

Top 10 slowest requests by issue count

Description	Avg	Max	Count	Last Date
MVC page : Home.Index [28,270 ms] slow at FabrikamFiber.DAL.Data.ServiceTicketRepository.AllIncluding() [...	26.51s	28.23s	7	2014/01/22 02:13:03 PM
MVC View rendering: Index [11,461 ms] slow at MVC View rendering: Index [11,461 ms]	10.09s	11.46s	7	2014/01/22 01:32:04 PM
MVC page : Home.Index [15,514 ms] slow at MVC page : Home.Index [15,514 ms]	15.51s	15.51s	1	2014/01/22 02:50:00 PM

Response time distribution

How many requests have high latency?



Exceptions Rate

How many exceptions are thrown per second?

CPU

How much CPU does my service consume?

Network

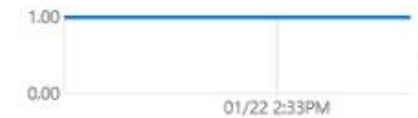
What is the network I/O rate for my service?

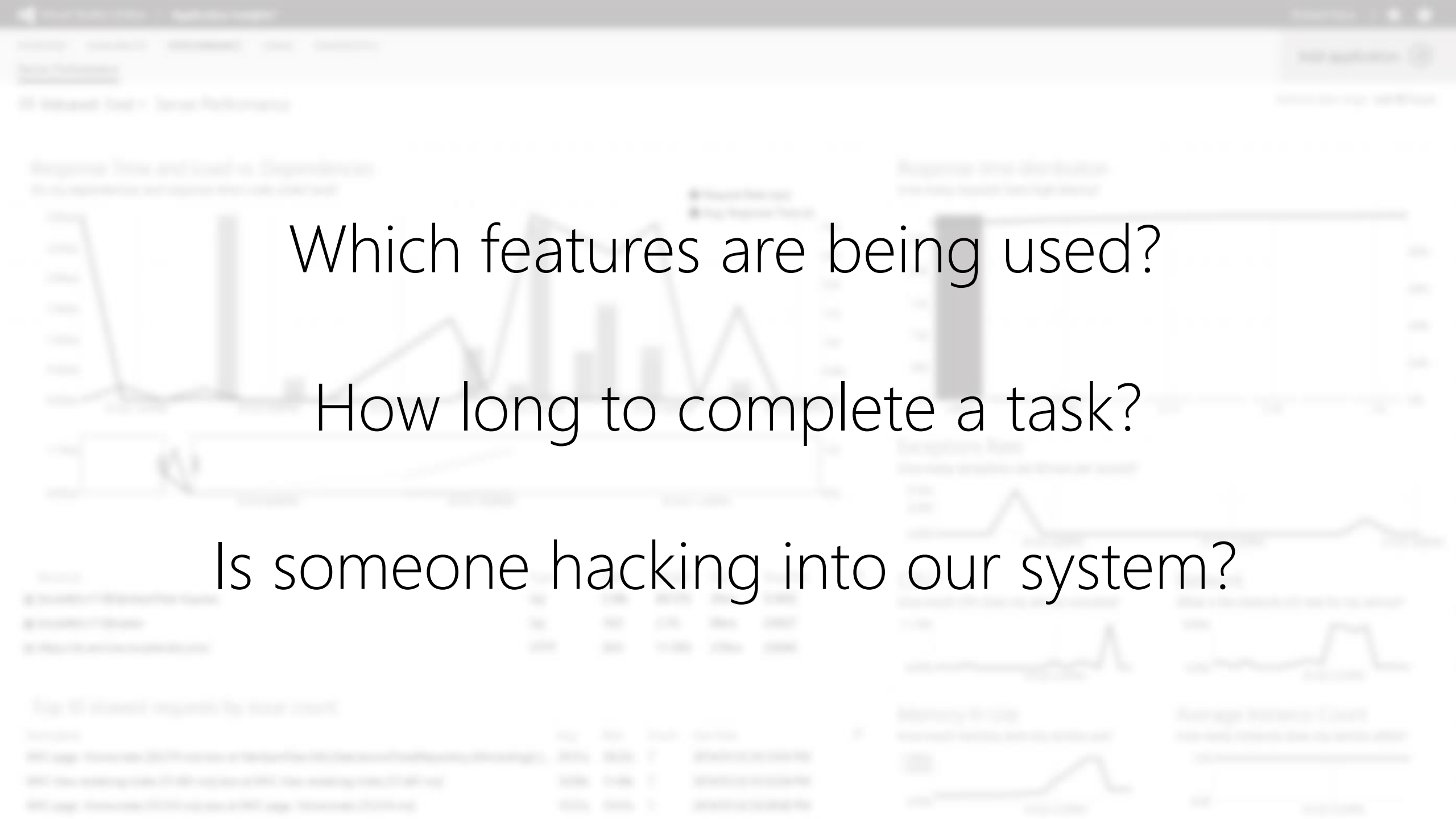
Memory In Use

How much memory does my service use?

Average Instance Count

How many instances does my service utilize?





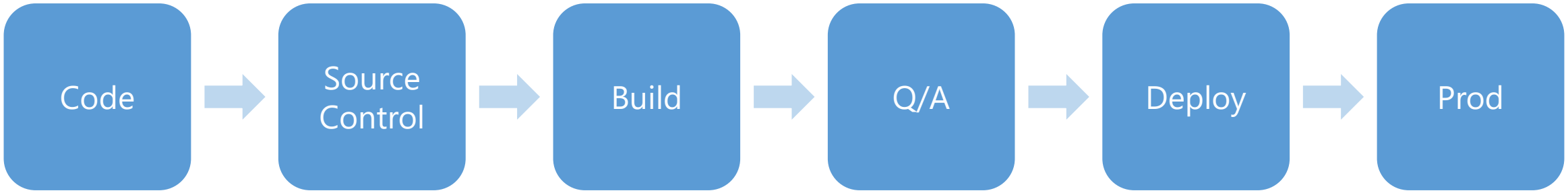
Which features are being used?

How long to complete a task?

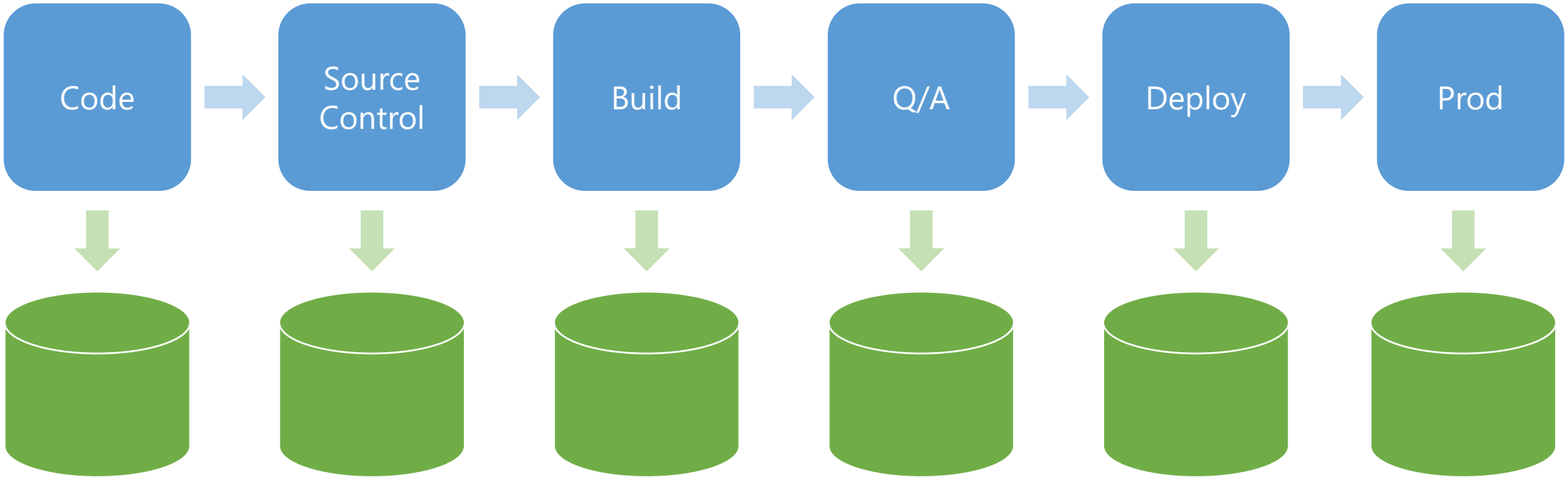
Is someone hacking into our system?


HACKERMAN

DevOps Pipeline

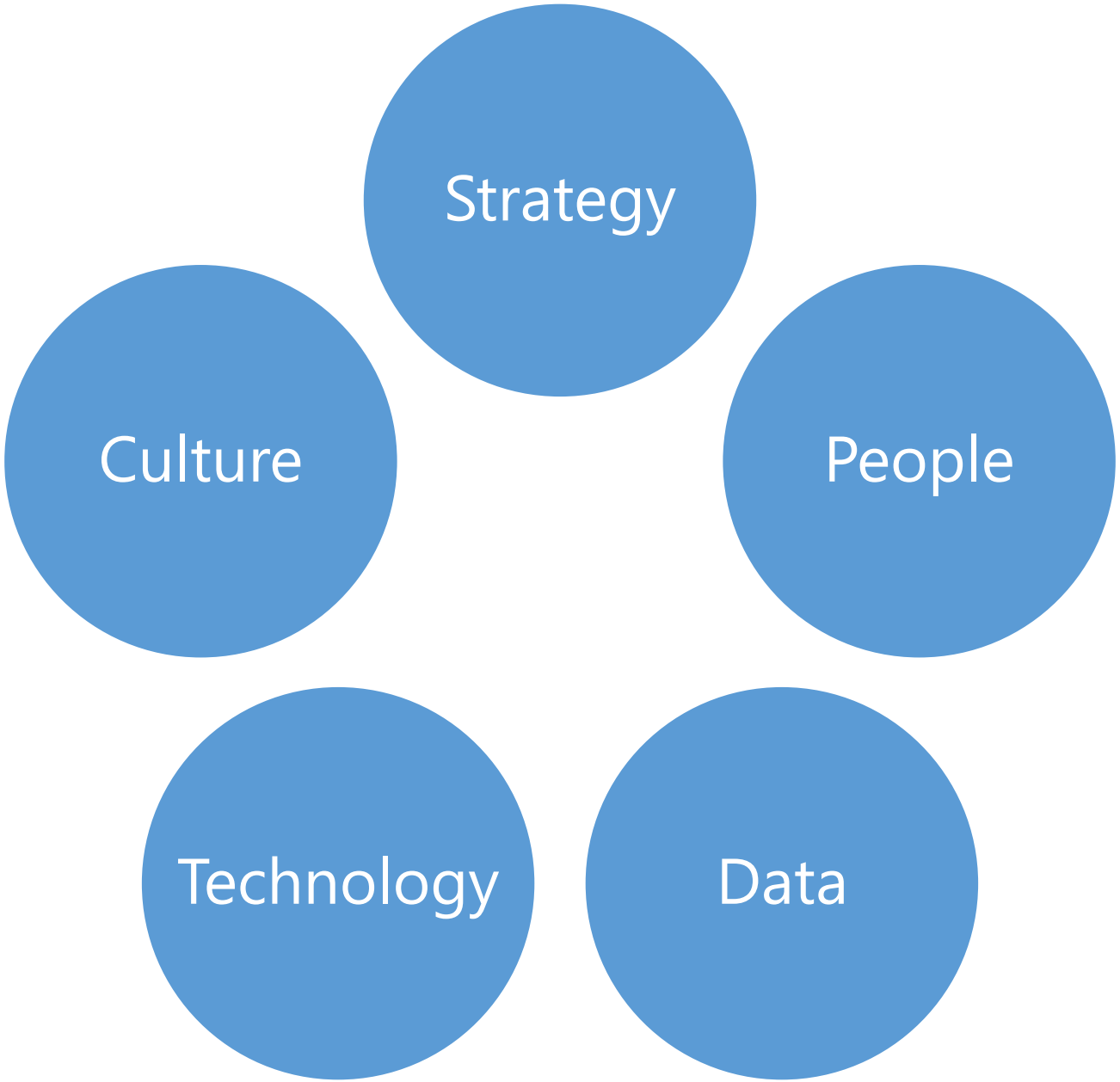


DevOps Pipeline

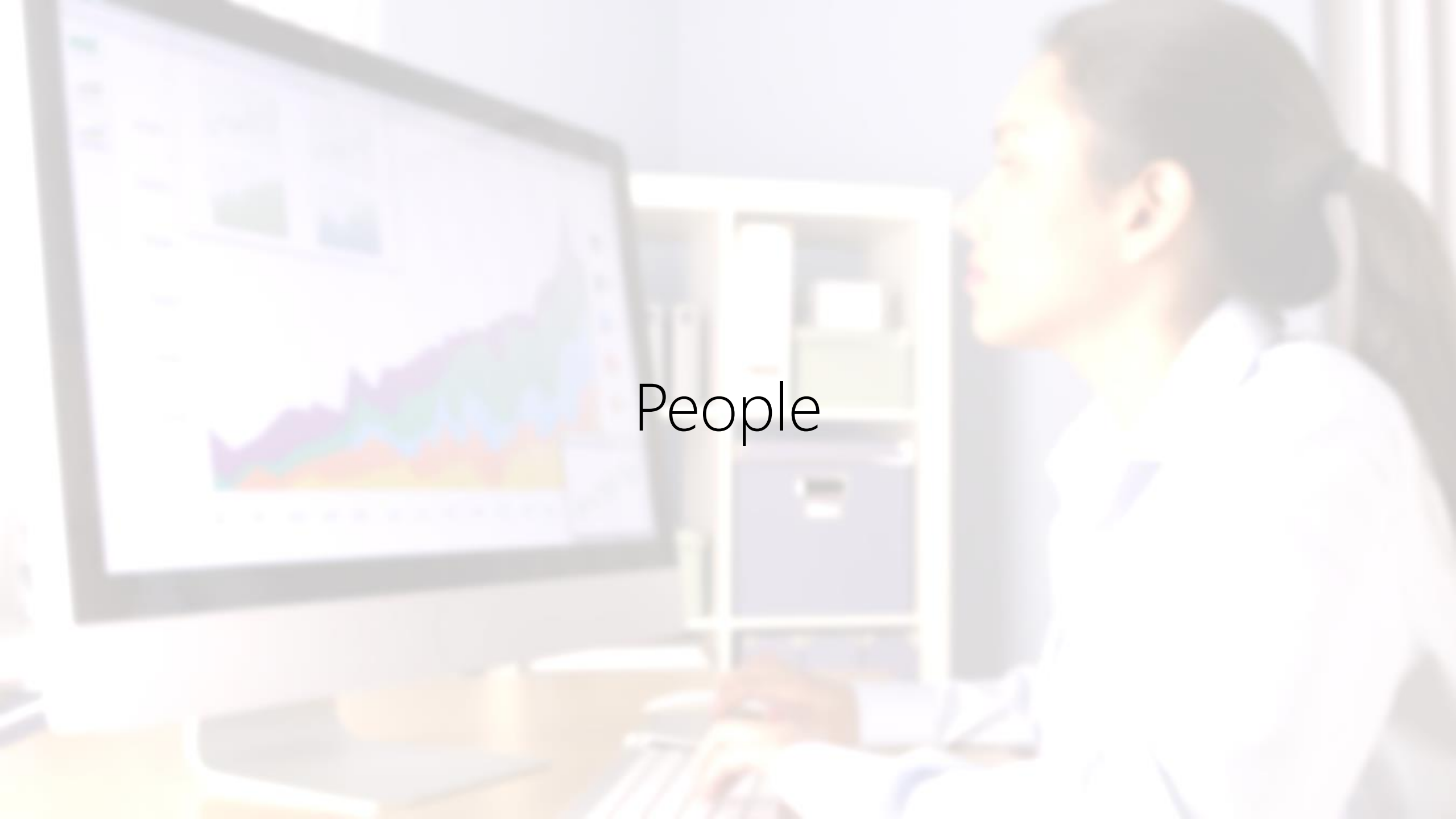


How Do I Get Started?

What are the ingredients of a data-driven enterprise?


Strategy
Culture
People
Technology
Data

Strategy



People

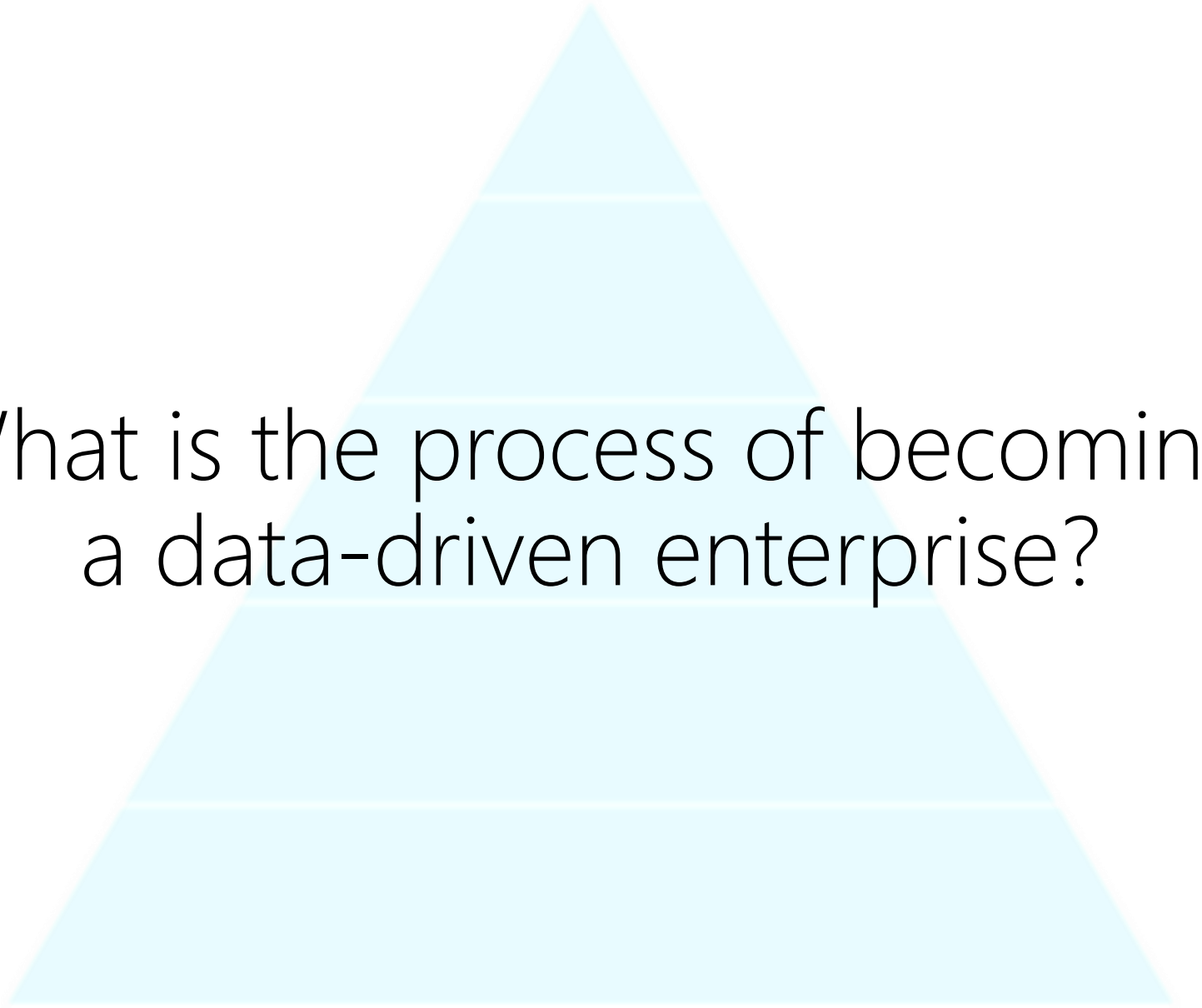

350,000
250,000
200,000
150,000
100,000
50,000
JAN
FEB
MAR
APR
MAY
JUN

Data

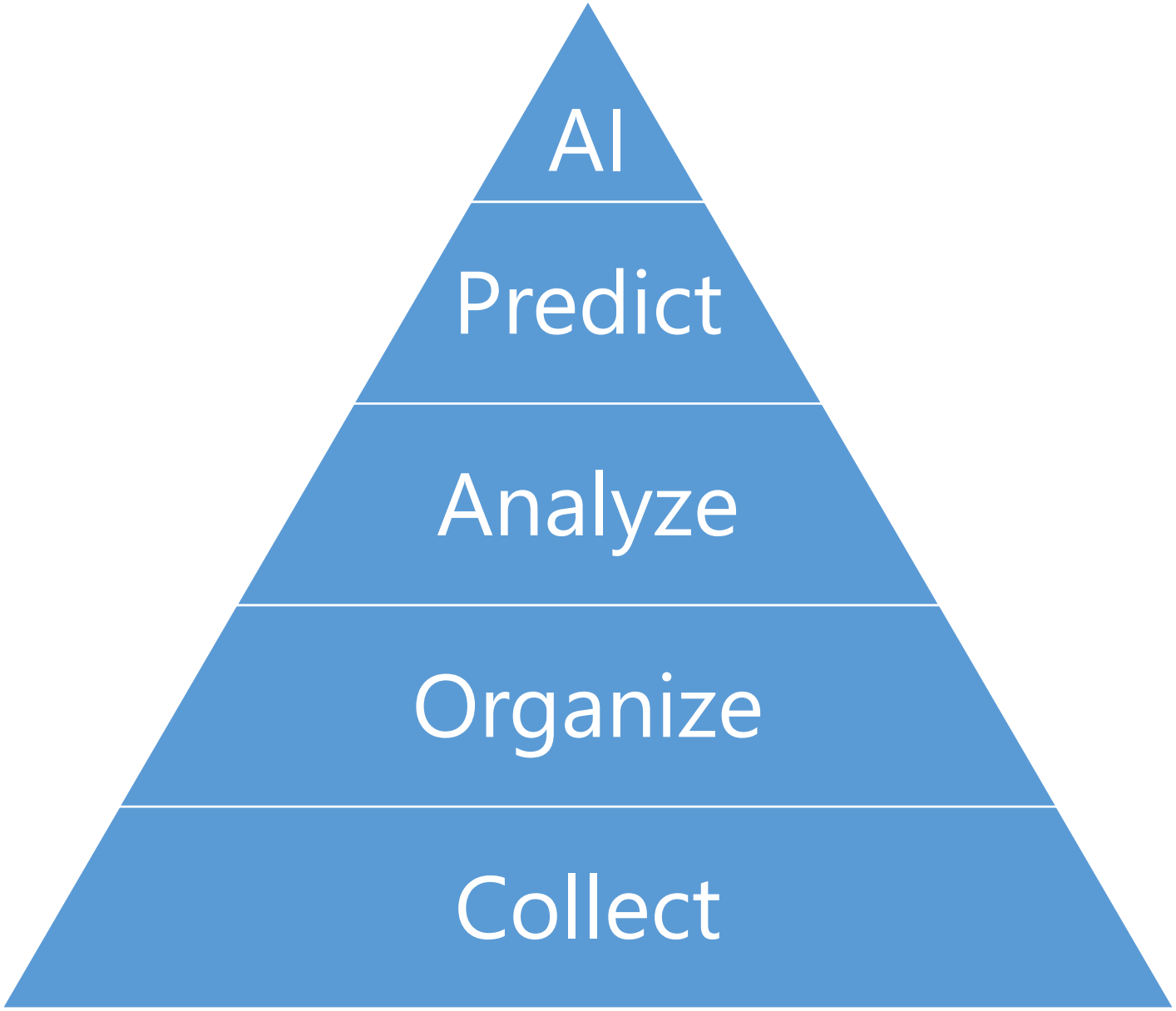

Technology

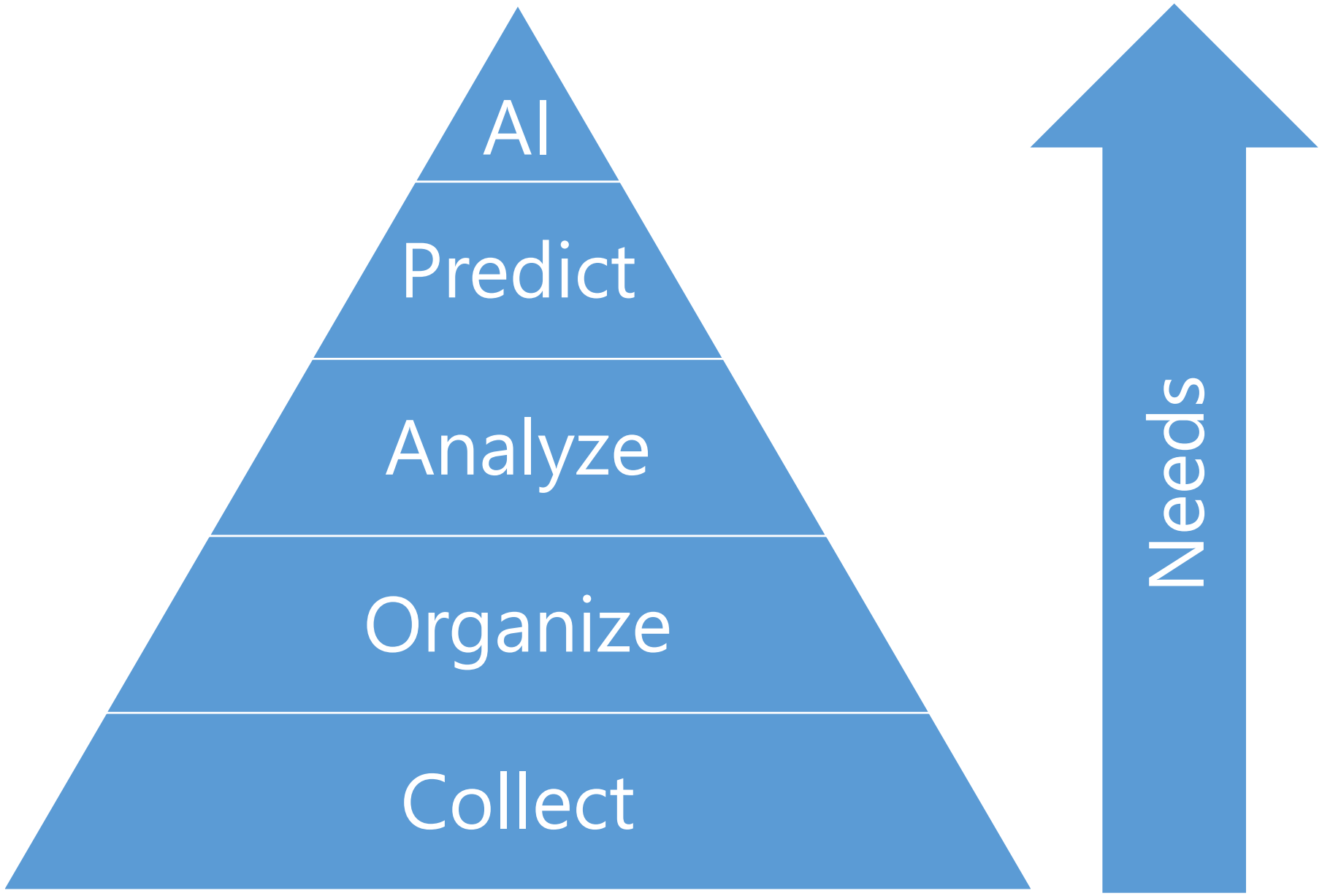

Power
Active
Power
Active
Power
Active
Power
Active
Power
Active
Power
Active
Power
Active
Power
Active

Culture



What is the process of becoming a data-driven enterprise?


AI
Predict
Analyze
Organize
Collect


AI
Predict
Analyze
Organize
Collect
Needs

1. Collect

1. Collect

Transactions

Logging

Digitization

1. Collect

Transactions
Logging
Digitization
Telemetry
Experiments
External data

2. Organize

2. Organize

Transform

Clean

Store

2. Organize

Transform

Clean

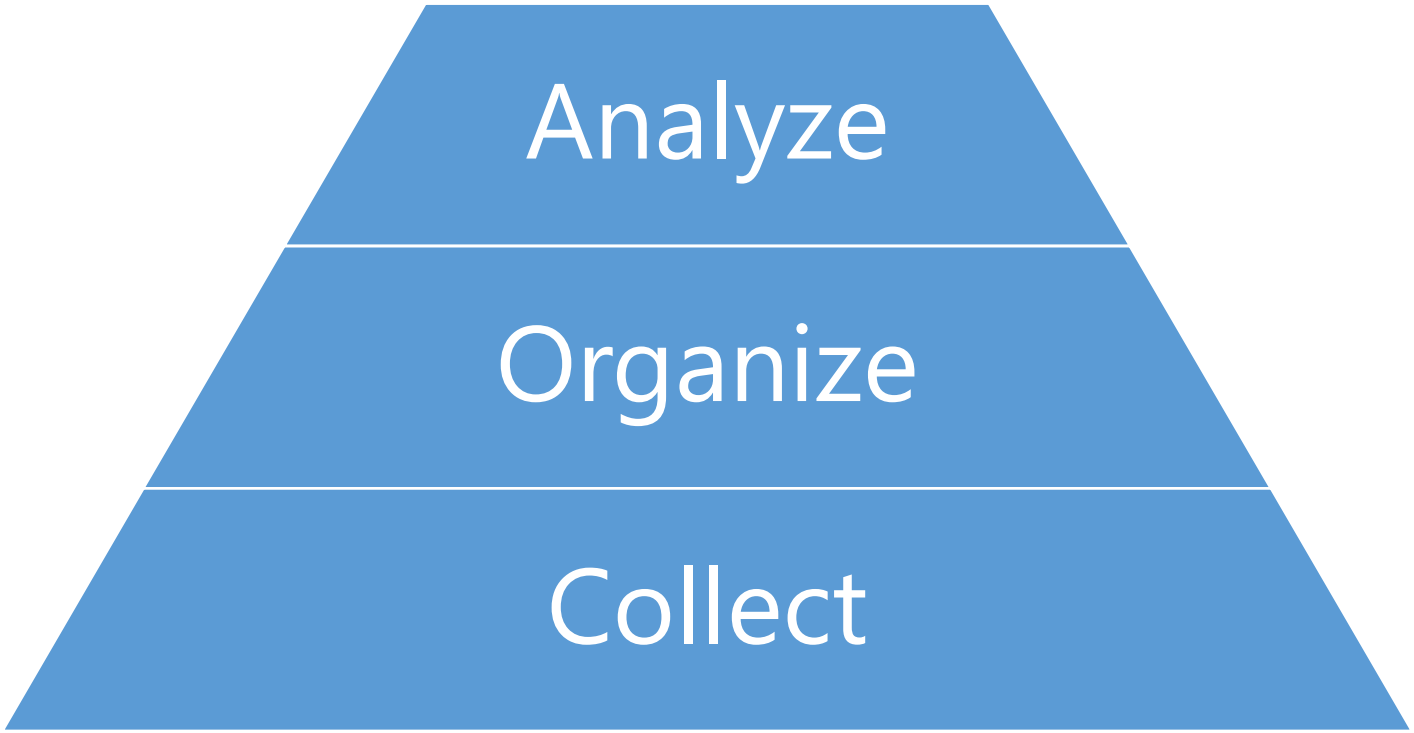
Store

Data ETL

Data Warehouse

Data Lake

3. Analyze

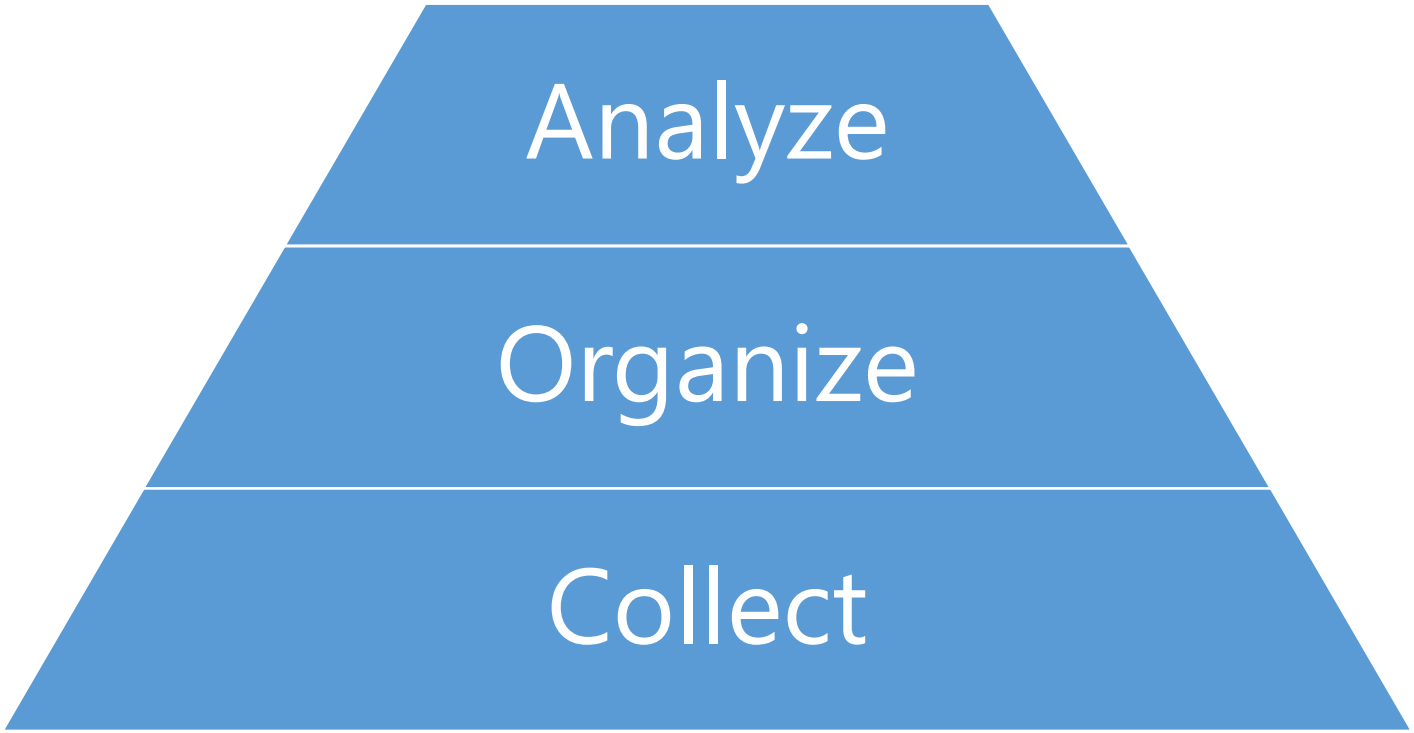


3. Analyze

Reports

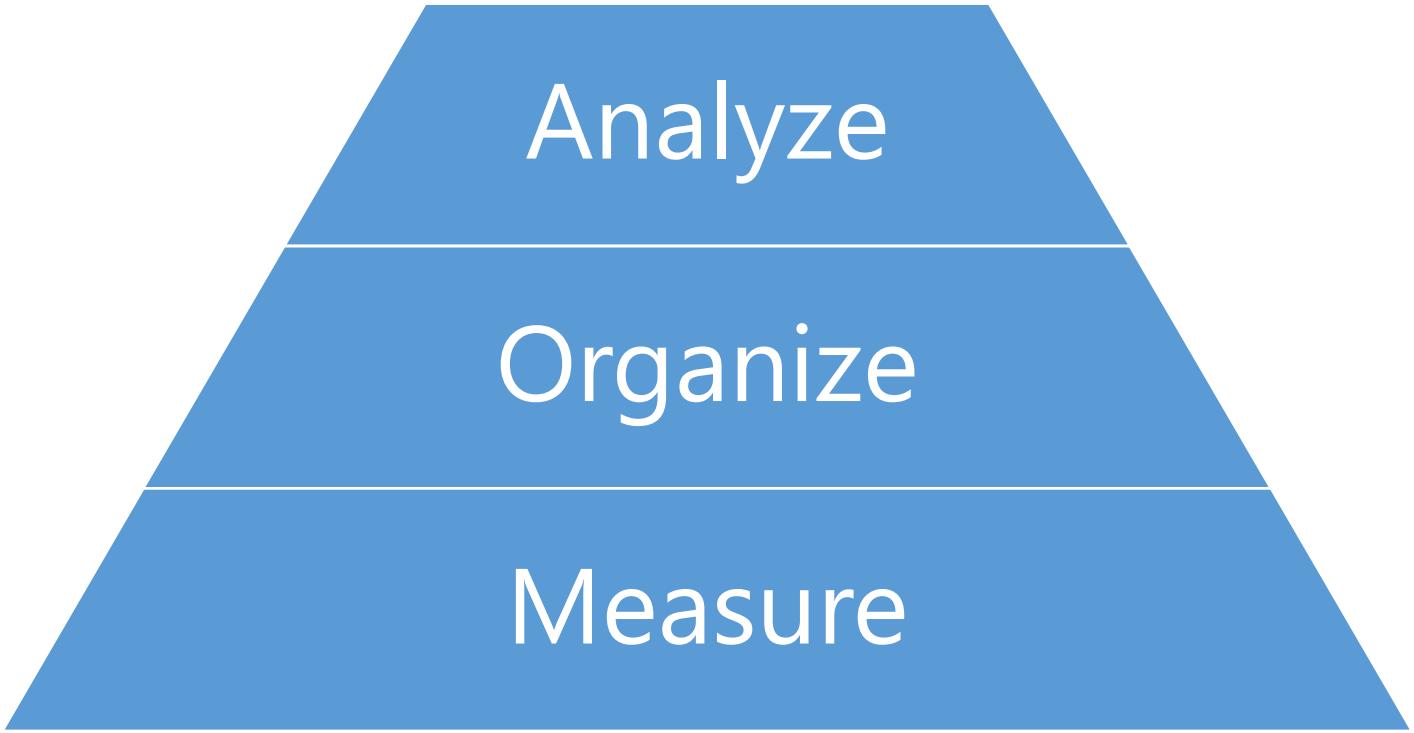
Dashboards

KPI monitors

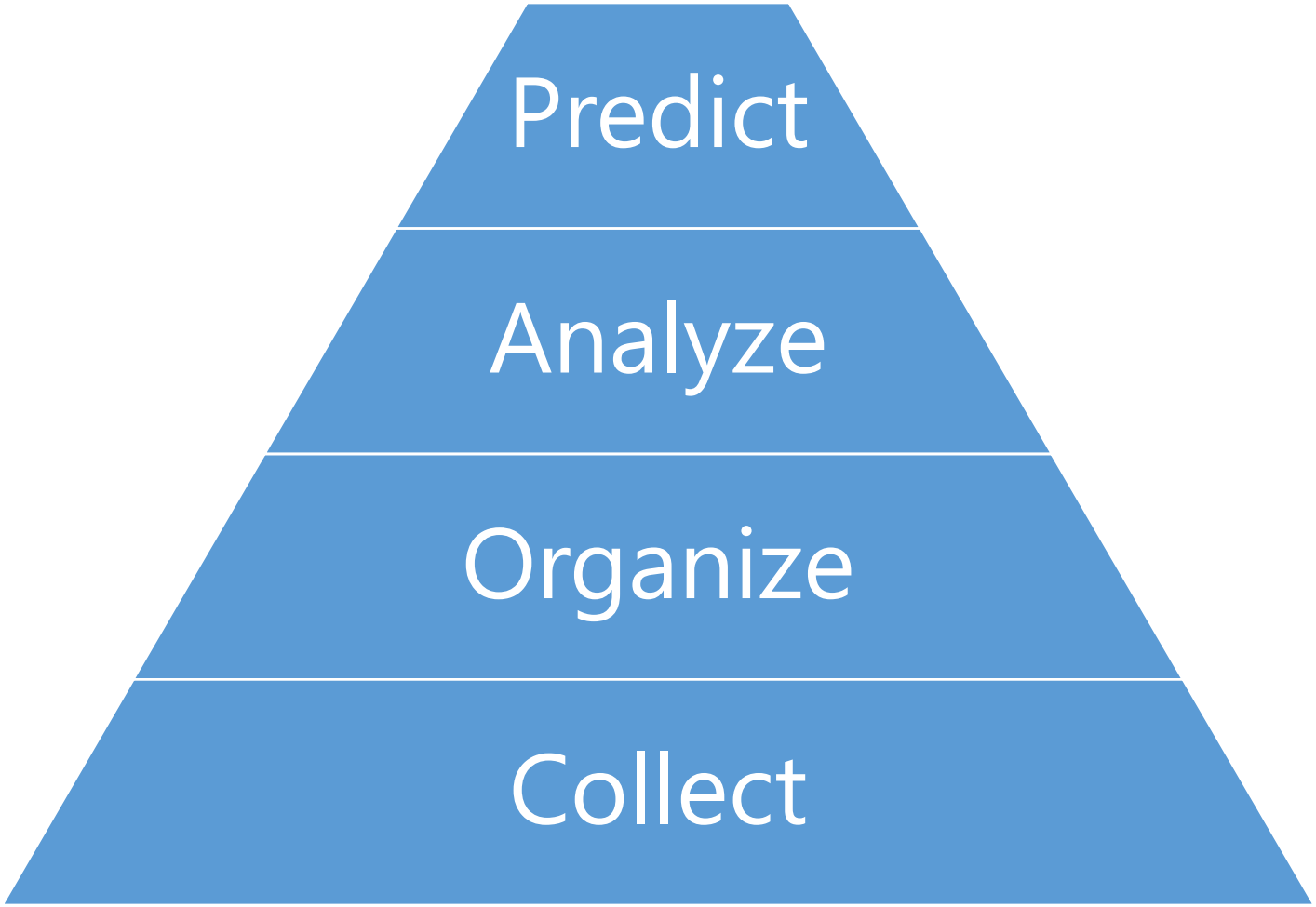


3. Analyze

Reports
Dashboards
KPI monitors
Data mining
Descriptive analytics
Diagnostic analytics



4. Predict

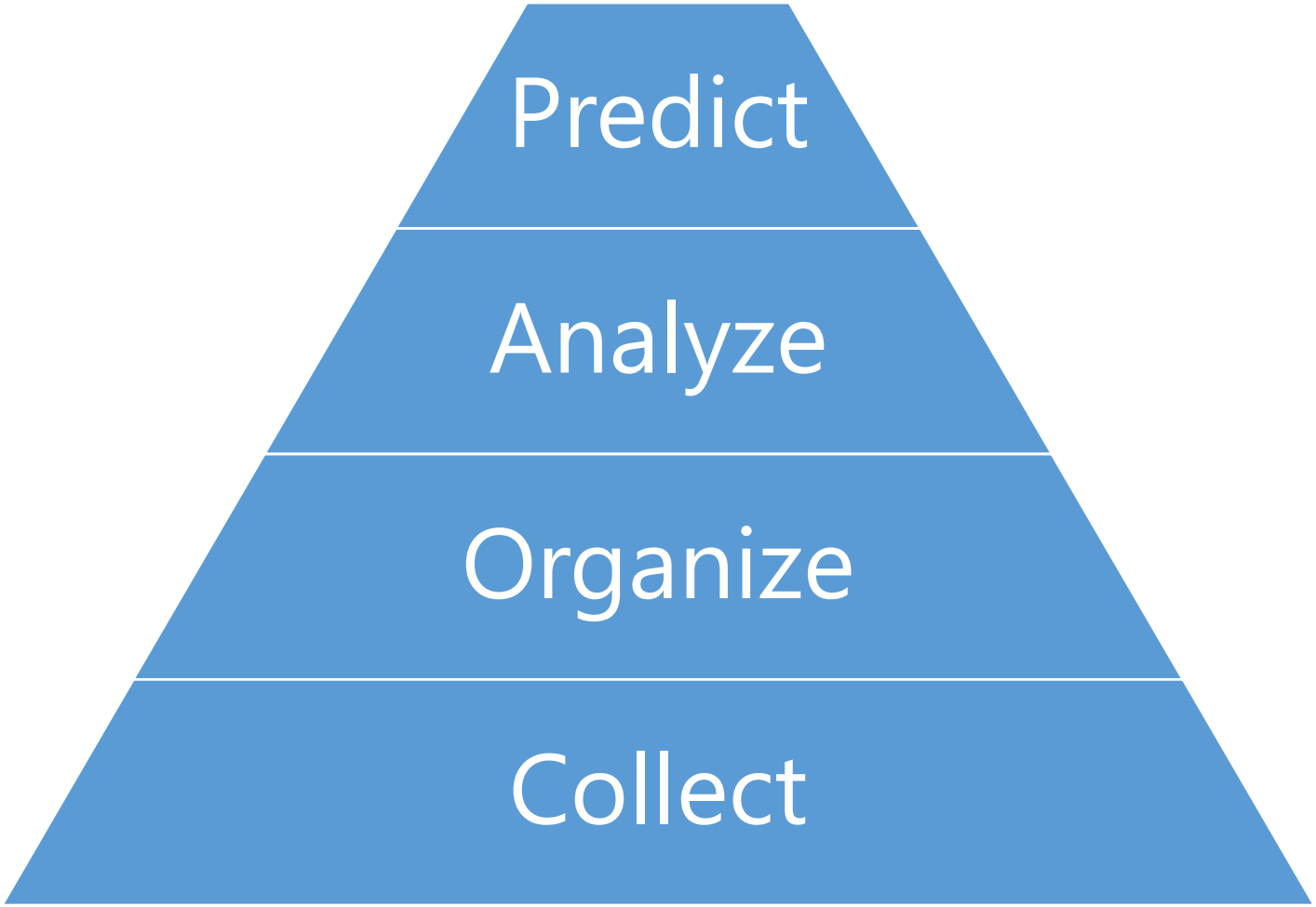


4. Predict

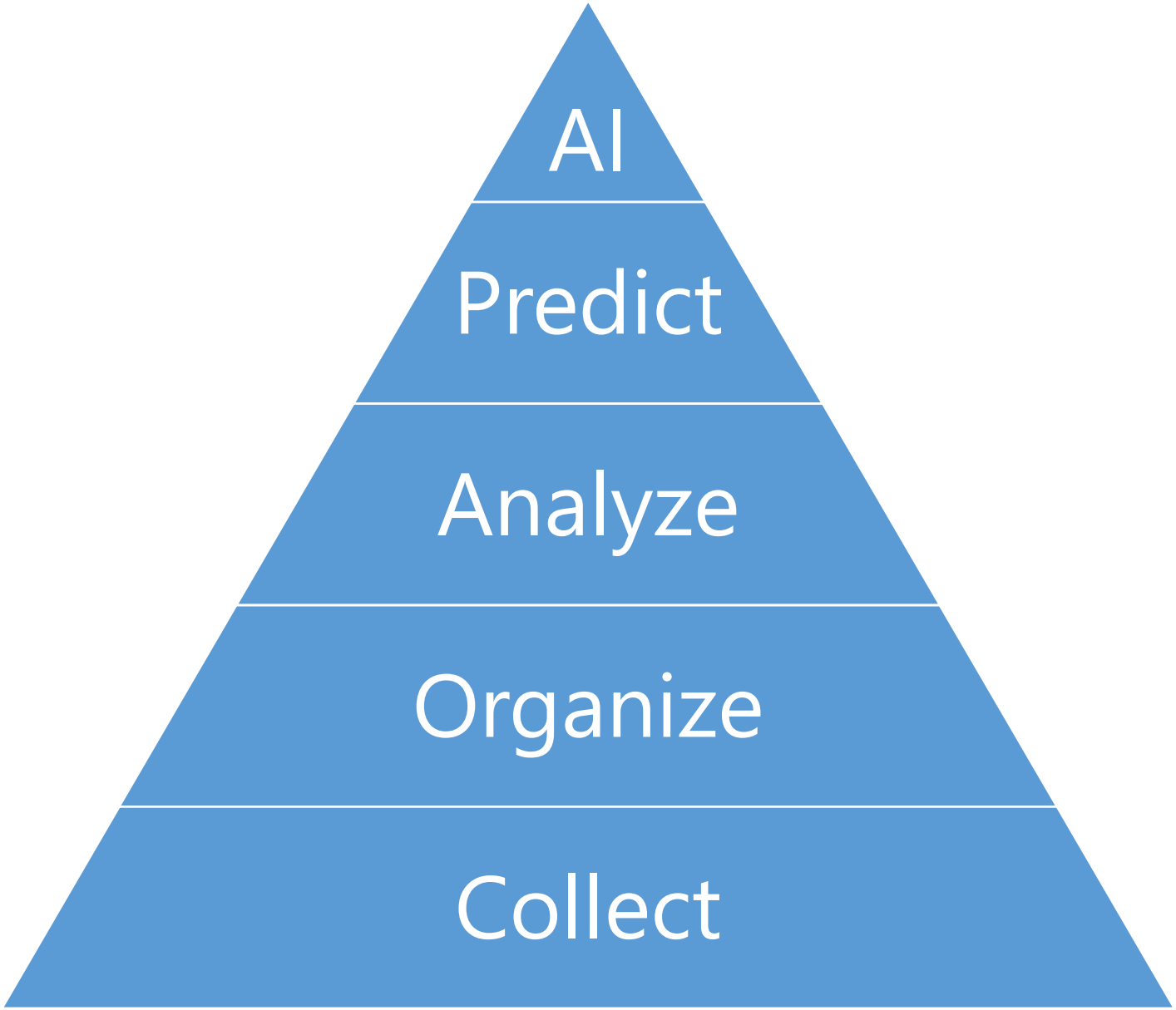
Predictive analytics

Prescriptive analytics

Machine learning



5. Automate

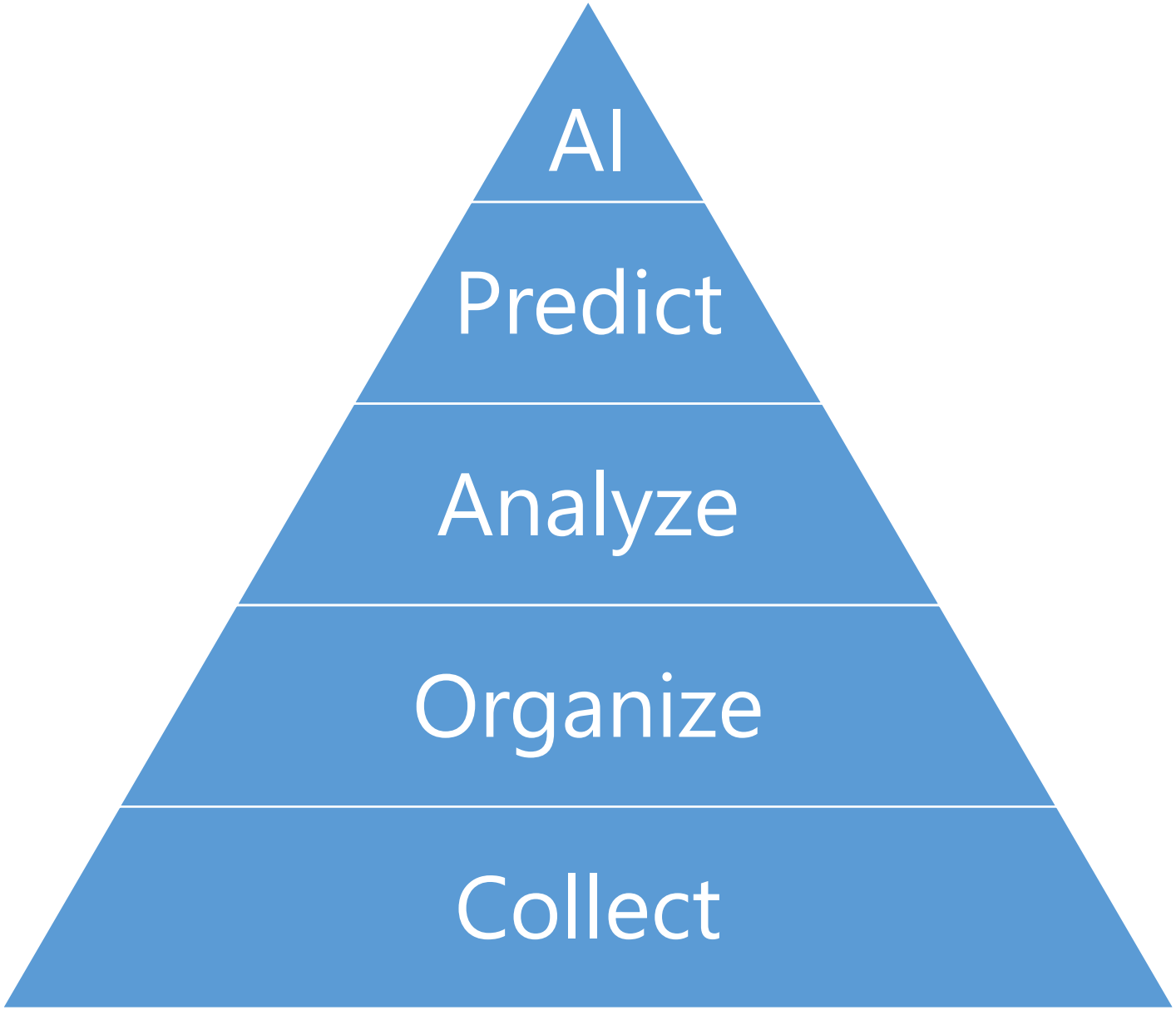


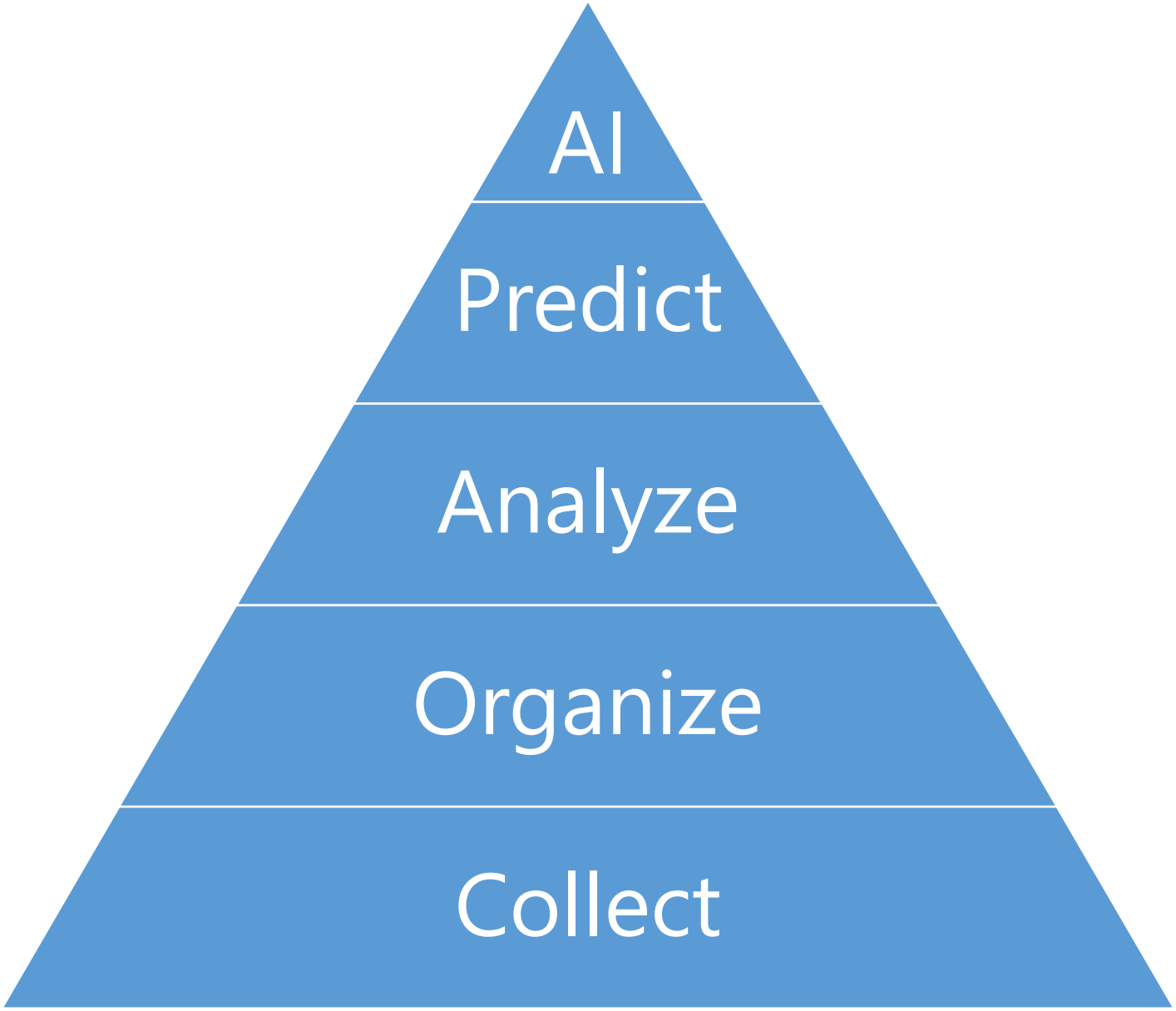
5. Automate

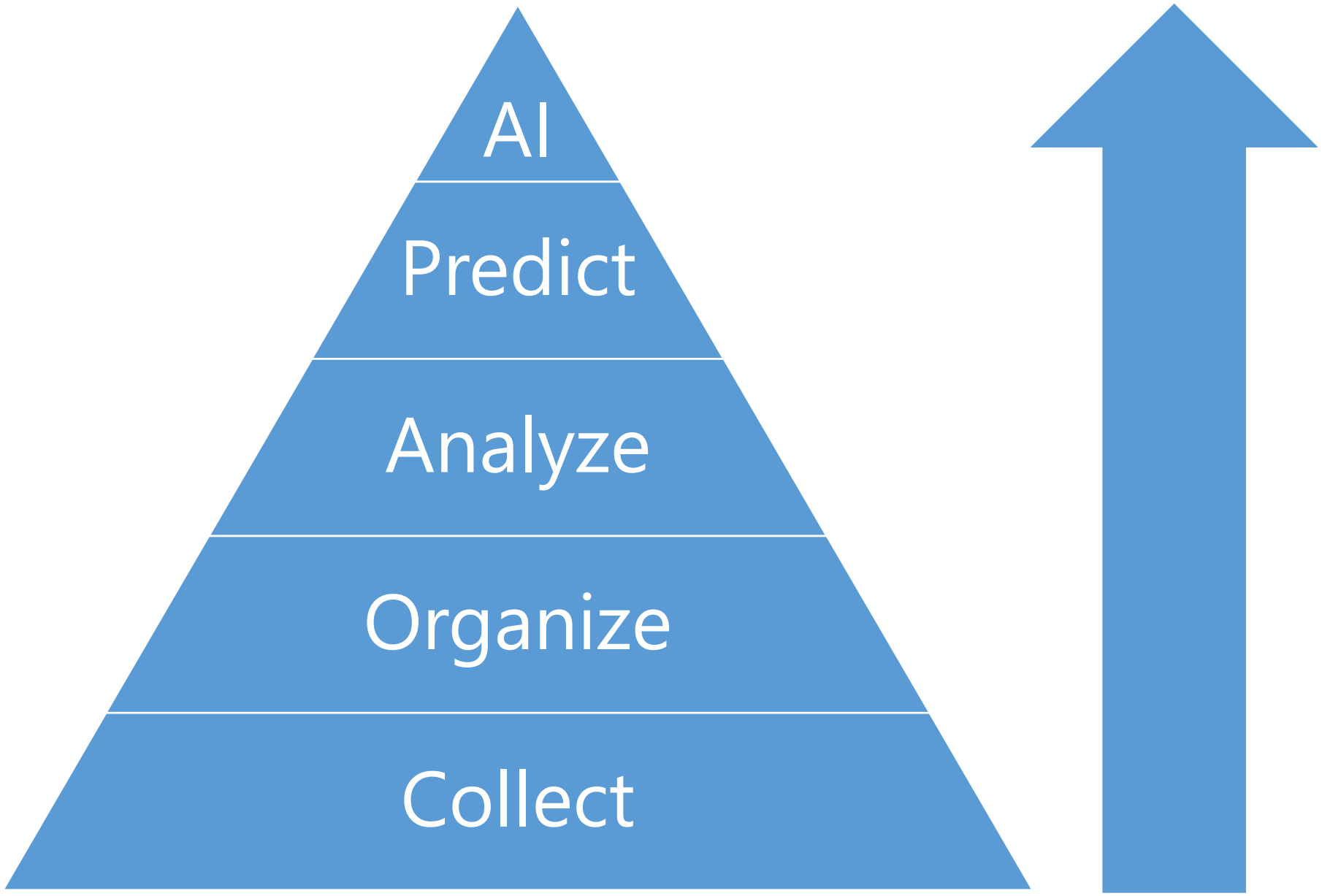
Artificial intelligence

Deep learning

Reinforcement learning




AI
Predict
Analyze
Organize
Collect


AI
Predict
Analyze
Organize
Collect

Advice for Success

Get buy-in from leadership

Focus on low-hanging fruit

Don’t silo data science teams

Democratize your data

Advice for Success

Get buy-in from leadership
Focus on low-hanging fruit
Don’t silo data science teams
Democratize your data

Embrace smart failure
Focus on feedback
Embed data collection
Avoid the Observer Effect

Conclusion

Where to Go Next?

Pluralsight Courses

Data Science: The Big Picture

Data Science with R

Exploratory Data Analysis with R

Data Visualization with R (3-part)

Deep Learning: The Big Picture

https://www.pluralsight.com/authors/matthew-renze

Matthew Renze

Home Articles Courses Presentations Software About Contact

News

2017-08-25 - Invitation to Speak at Devoxx Morocco

Very excited to announce that I've been invited to give a keynote in Casablanca at Devoxx Morocco in November. My keynote presentation will be on Artificial Intelligence.

2017-08-16 - Invitation to Speak at Microsoft Ignite

I've been invited to speak at Microsoft Ignite in Orlando, Florida in September. This will be my first time speaking at Ignite. Talks will include both Data Science and Machine Learning with R.

2017-08-14 - Dev on Fire Interview

Matthew is a data science consultant, author for Pluralsight, international public speaker, a Microsoft MVP, ASPInsider, and open-source software contributor.

www.matthewrenze.com

Engage

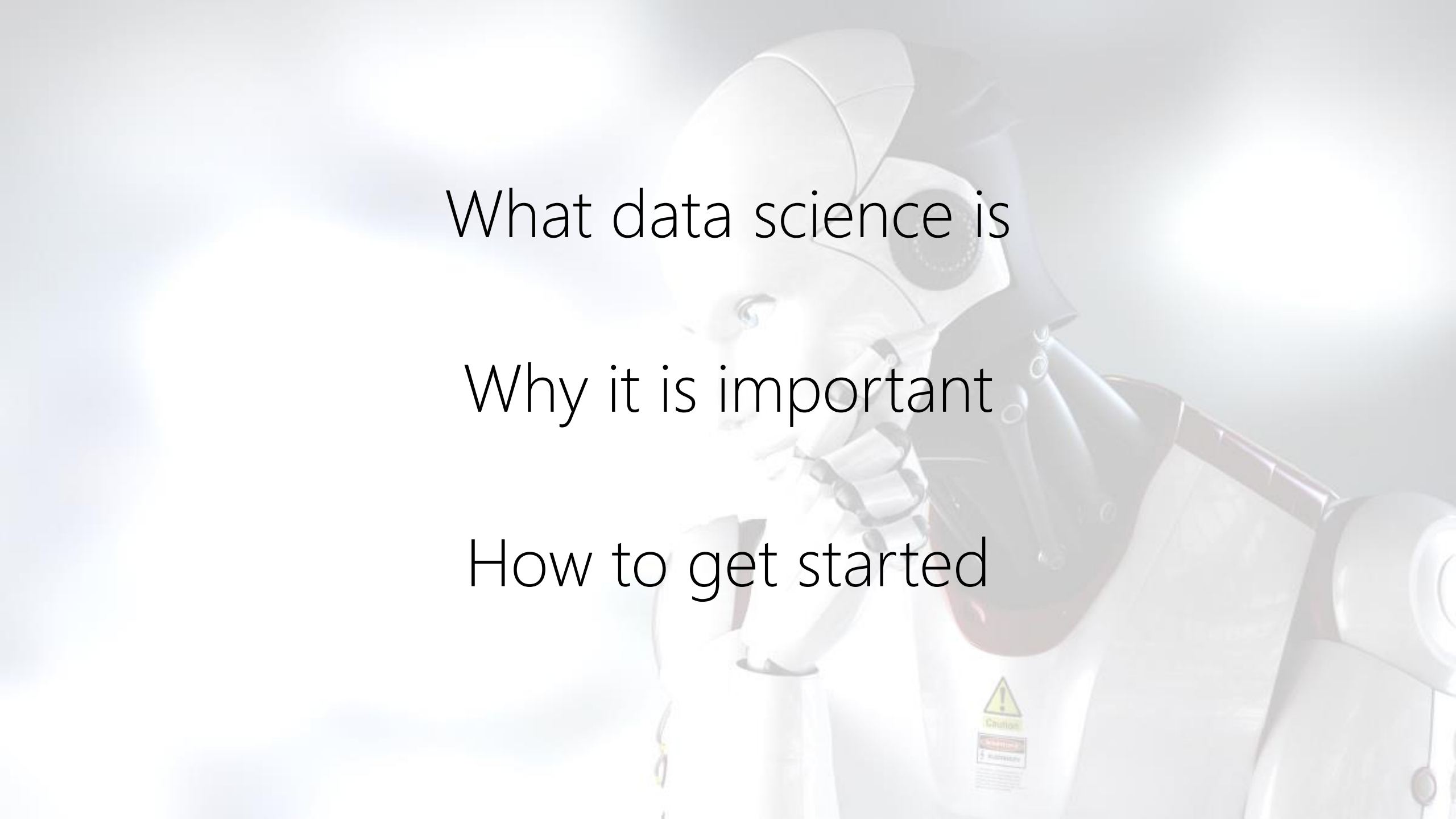
Rate

Question

Comment

Feedback


Caution
WARNING
ELECTRICITY

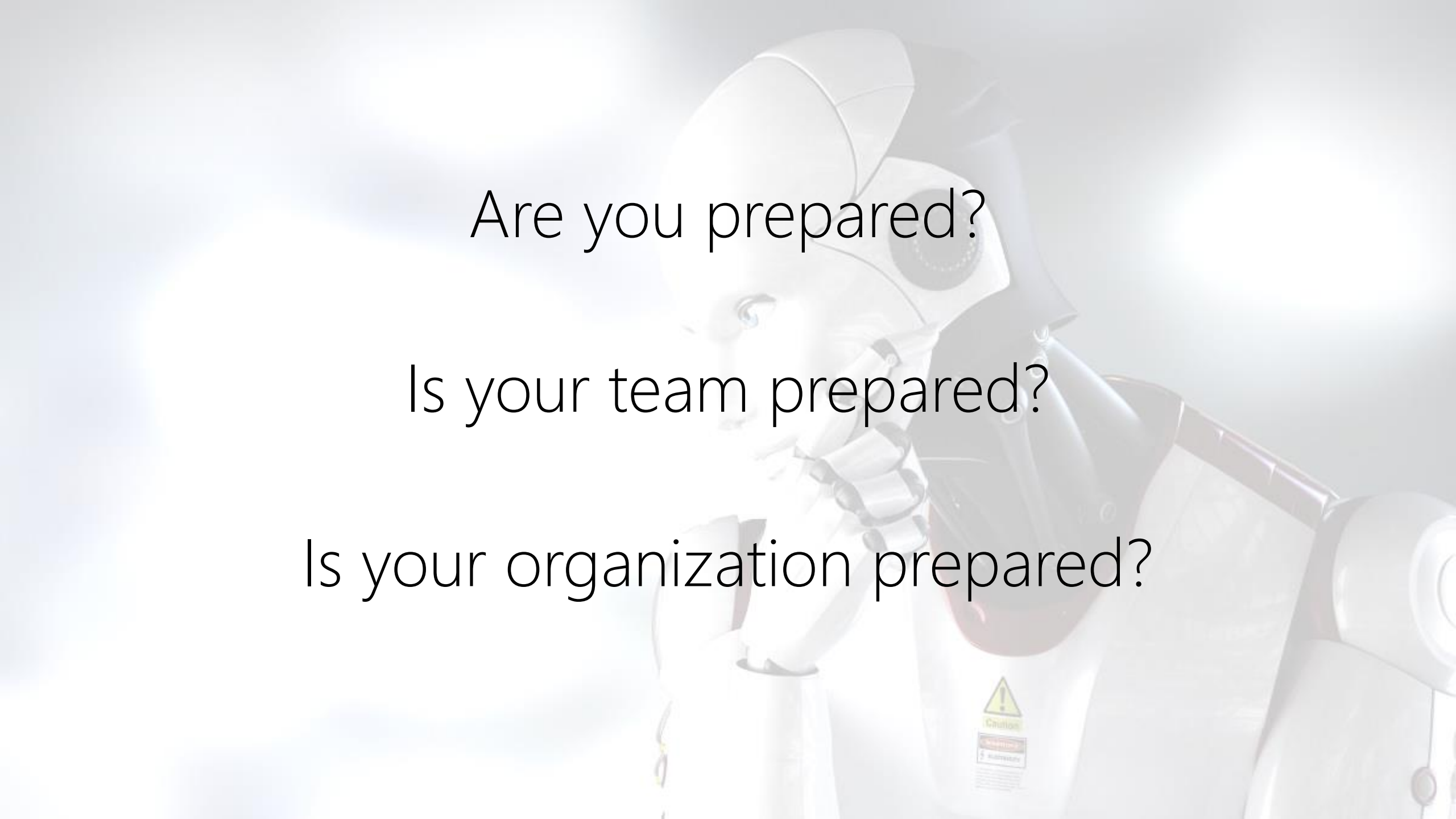


What data science is

Why it is important

How to get started


Caution
WARNING
ELECTRICITY



Are you prepared?

Is your team prepared?

Is your organization prepared?


Caution
WARNING
ELECTRICITY

Thank You!

Matthew Renze
Data Science Consultant
Renze Consulting

Twitter: @matthewrenze
Email: info@matthewrenze.com
Website: www.matthewrenze.com