Elevating your career through scientific computing and the cloud

Ryan Elkins

Twitter: @ryanelkins

Email: ryan-elkins@outlook.com

GitHub: https://github.com/brevityinmotion

Information Security Architect

Ryan Elkins

Twitter: @ryanelkins

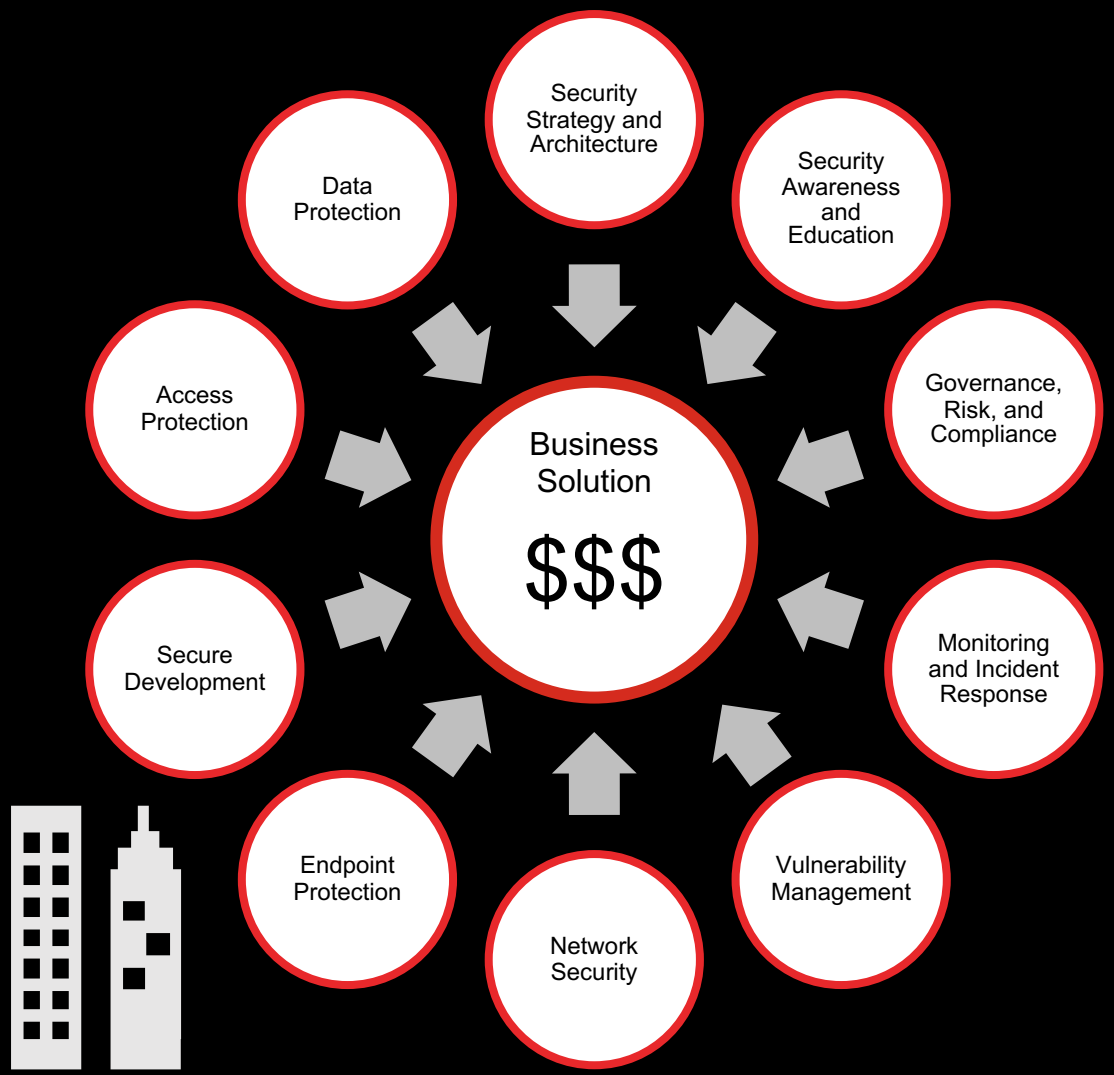
My views are my own.

Security Researcher

David Rowan – Wired – Atech 2017

Security by Design

Modular software, unopinionated deployment.

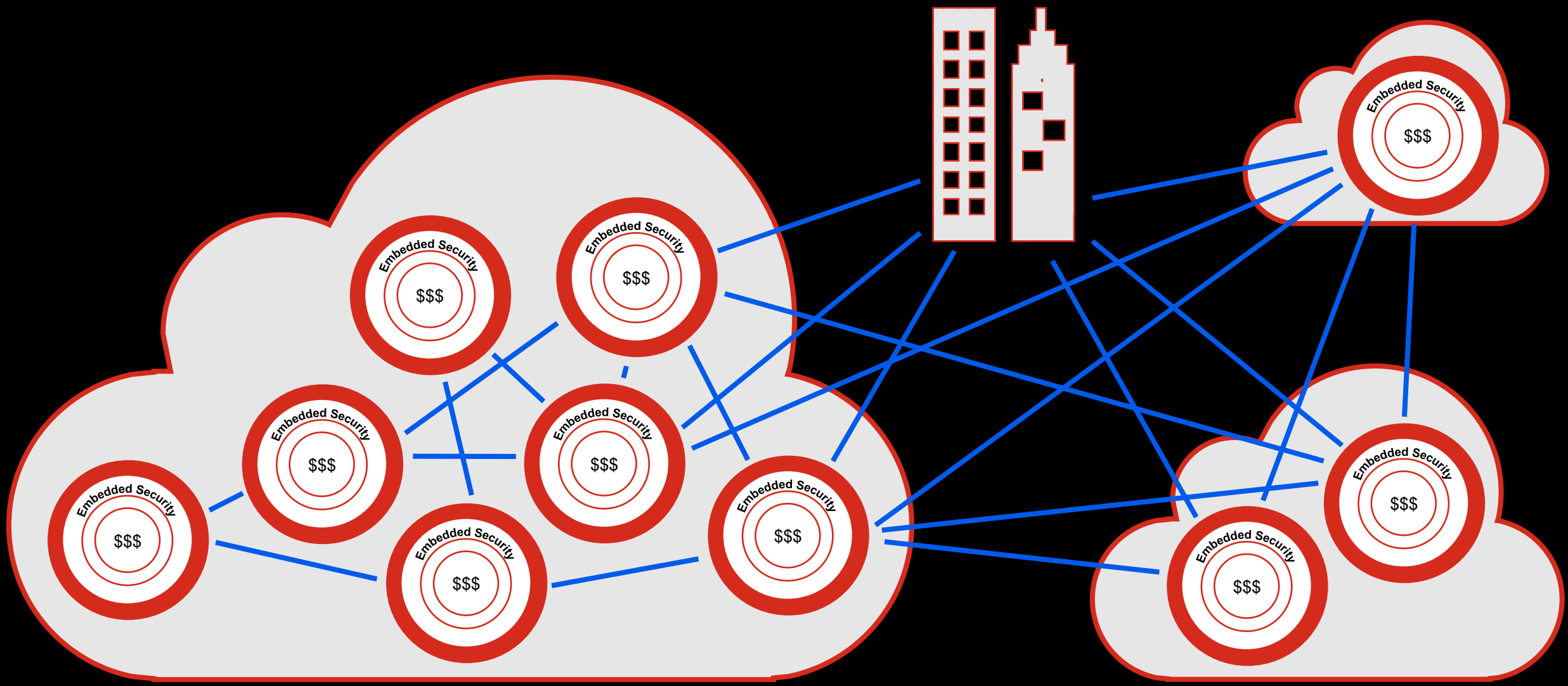


Bolt-on Security

Embedded Security

Hybrid Multi-Cloud

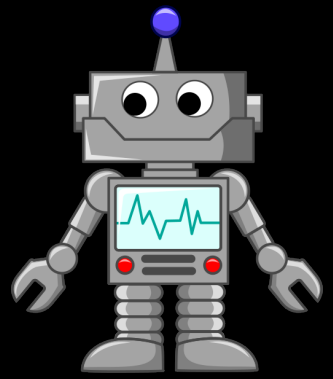
An interconnected, decentralized ecosystem.



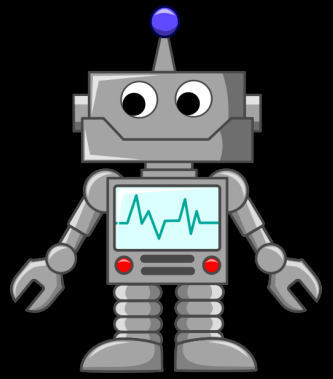
Robotic Process Automation

Eliminate manual processes.

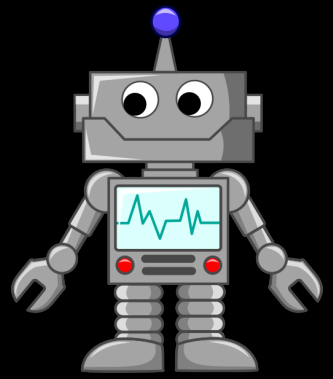
Systems Administration



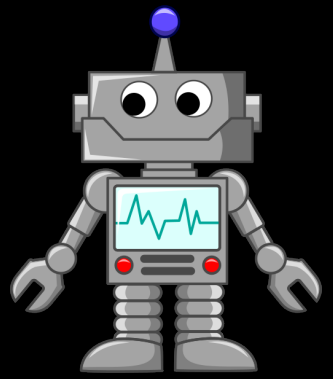
DevOps



Engineering

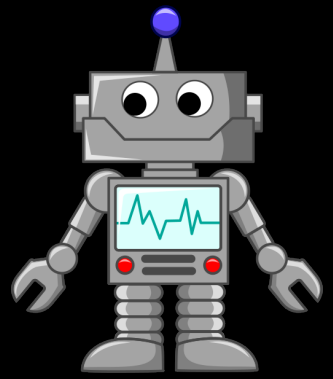


Quality Assurance

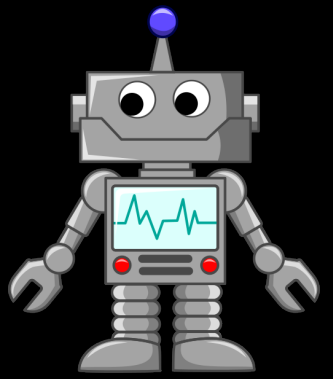


Information Security

Production Support



Lifecycle Management

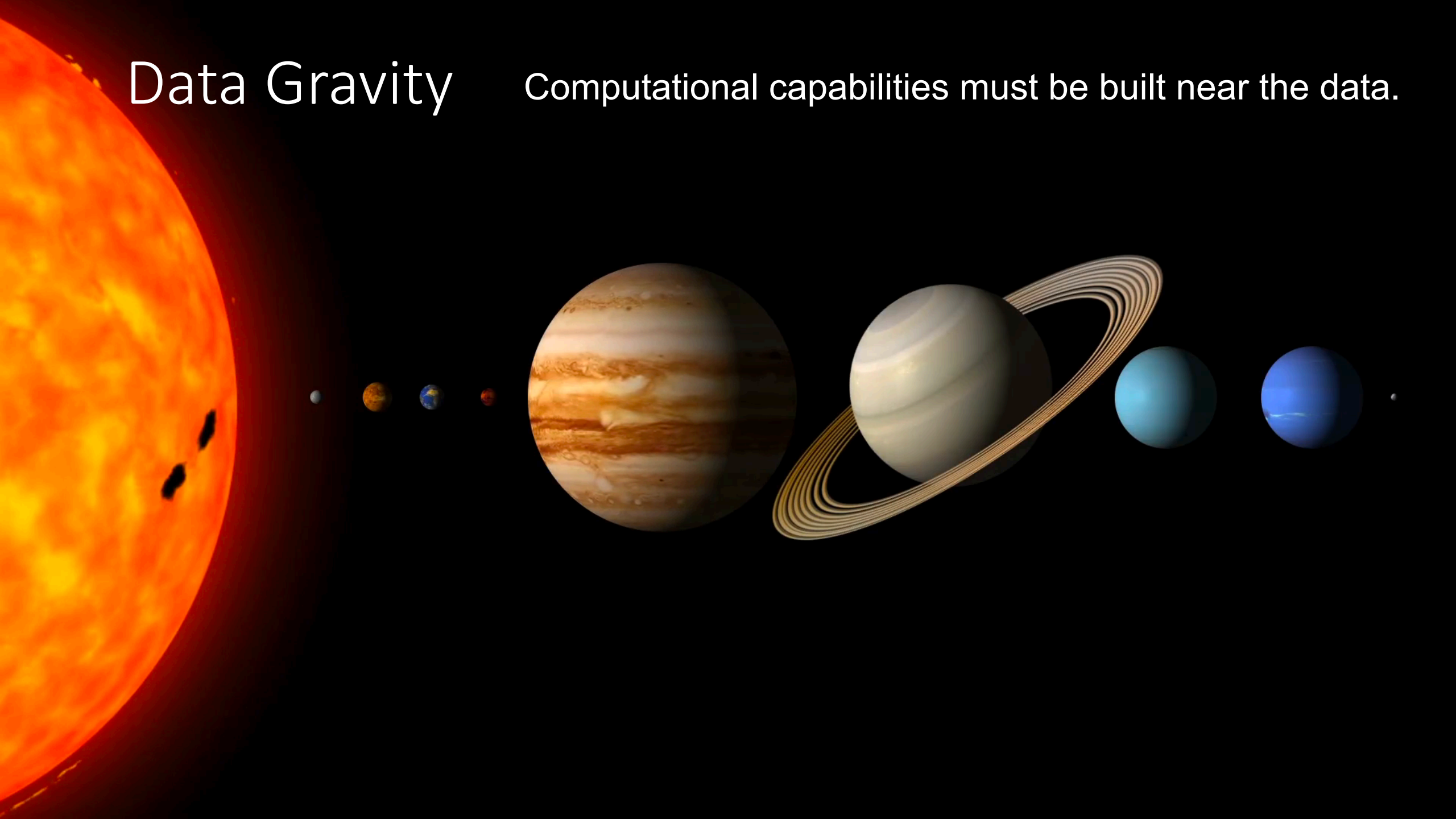


Data Driven Everything

Three major shifts driving mass adoption.

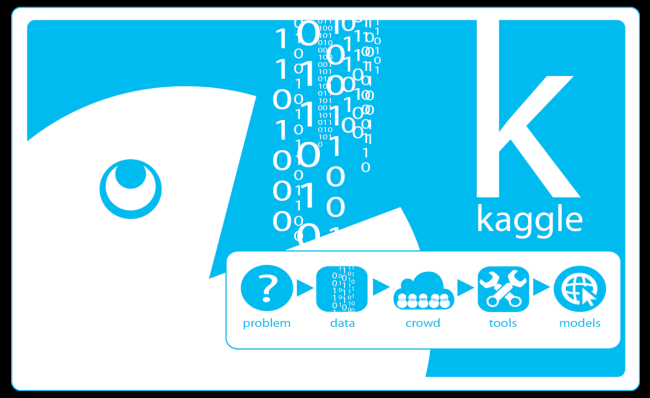
Data Gravity

Computational capabilities must be built near the data.



Democratized Accessibility

Increasing number of public data sets.



Google Dataset Search Beta

Search for Datasets

RAPID7 Open Data

Open Data

Offering researchers and community members open access to data from Project Sonar, which conducts internet-wide surveys to gain insights into global exposure to common vulnerabilities.

DATASETS: 13 FILES: 14,263 TOTAL SIZE: 27.4 TB

AWS Public Dataset Program

The AWS Public Dataset Program covers the cost of storage for publicly available high-value cloud-optimized datasets.

GitHub Repos

Github

5 months 3 TB 6.5 BigQuery

Common Crawl

encyclopedic internet machine learning natural language processing

A corpus of web crawl data composed of over 25 billion web pages.

Collectively solving the world's largest challenges

HARVARD Dataverse

Add Data

Resources for COVID-19 (China Data Lab)

Harvard Dataverse > China Data Lab Dataverse > **Resources for COVID-19**

NIH National Institutes of Health
Office of Data Science Strategy

Home | Strategic Plan | Resources | Research

COVID-19 is an emerging, rapidly evolving situation.

Get the latest public health information from CDC: https://www.coronavirus.gov
Get the latest research information from NIH: https://www.nih.gov/coronavirus
NIH and other federal agencies have made COVID-19 data available through several Open-Ac
Resources

Open-Access Data and Computational Resources to Address COVID-19

DATA ANALYTICS

COVID-19 public dataset program: Making data freely accessible for better public outcomes

HOME / DATASETS / **NOVEL CORONAVIRUS (COVID-19) CASES DATA**

Novel Coronavirus (COVID-19) Cases Data

This dataset is part of COVID-19 Pandemic

Novel Corona Virus (COVID-19) epidemiological data since 22 January 2020. The data is compiled by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) from various sources including the World Health Organization (WHO), DXY.cn, BNO News, National Health Commission of the People's Republic of China (NHC), China CDC (CCDC), Hong Kong ... More

249000+ Downloads | This dataset updates: Live

BROAD INSTITUTE

ABOUT US PEOPLE SCIENCE

HOME » DATA SOFTWARE AND TOOLS

DATA, SOFTWARE AND TOOLS

FILTER BY

Select options

All

Dataset

7496

COVID-19 Open Research Dataset Challenge (CORD-19)

An AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House

Ai2 Allen Institute For AI and 8 collaborators • updated 4 days ago (Version 32)

Data | Tasks (17) | Kernels (1,549) | Discussion (353) | Activity | Metad | Download (13 GB) | New Notebook

Financially Viable

No longer cost prohibitive.

Google colab

Welcome to Colaboratory!

Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud.

With Colaboratory you can write and execute code, save and share your analyses, and access powerful computing resources, all for free from your browser.

Amazon SageMaker

Machine learning for every developer and data scientist

Get Started with SageMaker

Anaconda Distribution

The open-source Anaconda Distribution is the easiest way to perform Python/R data science and machine learning on Linux, Windows, and Mac OS X. With over 15 million users worldwide, it is the industry standard for developing, testing, and training on a single machine, enabling *individual data scientists* to:

amazon web services

AWS Free Tier

Gain free, hands-on experience with the AWS platform, products, and services

Let’s get started!

Jupyter Notebook?

?

?

Jupyter Notebook?

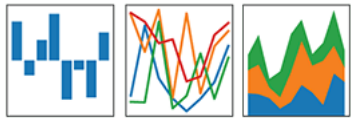
Jupyter – Julia, Python, and R

- Open source web application for interactive computing.
- Supports over 40 programming languages.
- Notebooks contain live code, equations, visualizations, and narrative text.

Seaborn

pandas

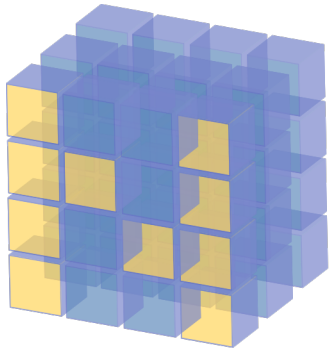
$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



jupyter

IP[y]:

IPython

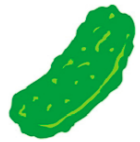


NumPy

bokeh

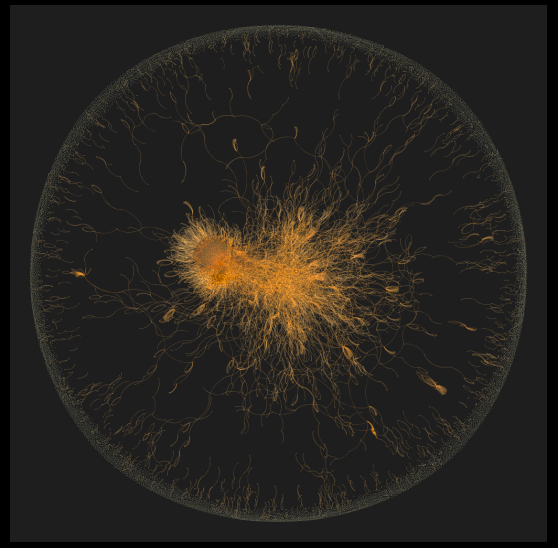
DASK

matplotlib

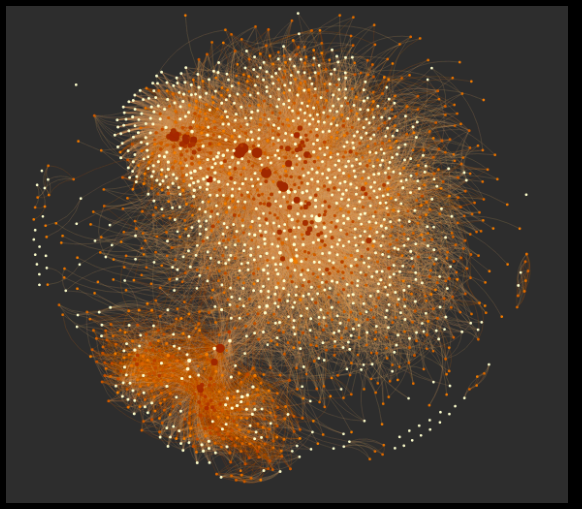


pickleDB

Security experts love to tell stories.



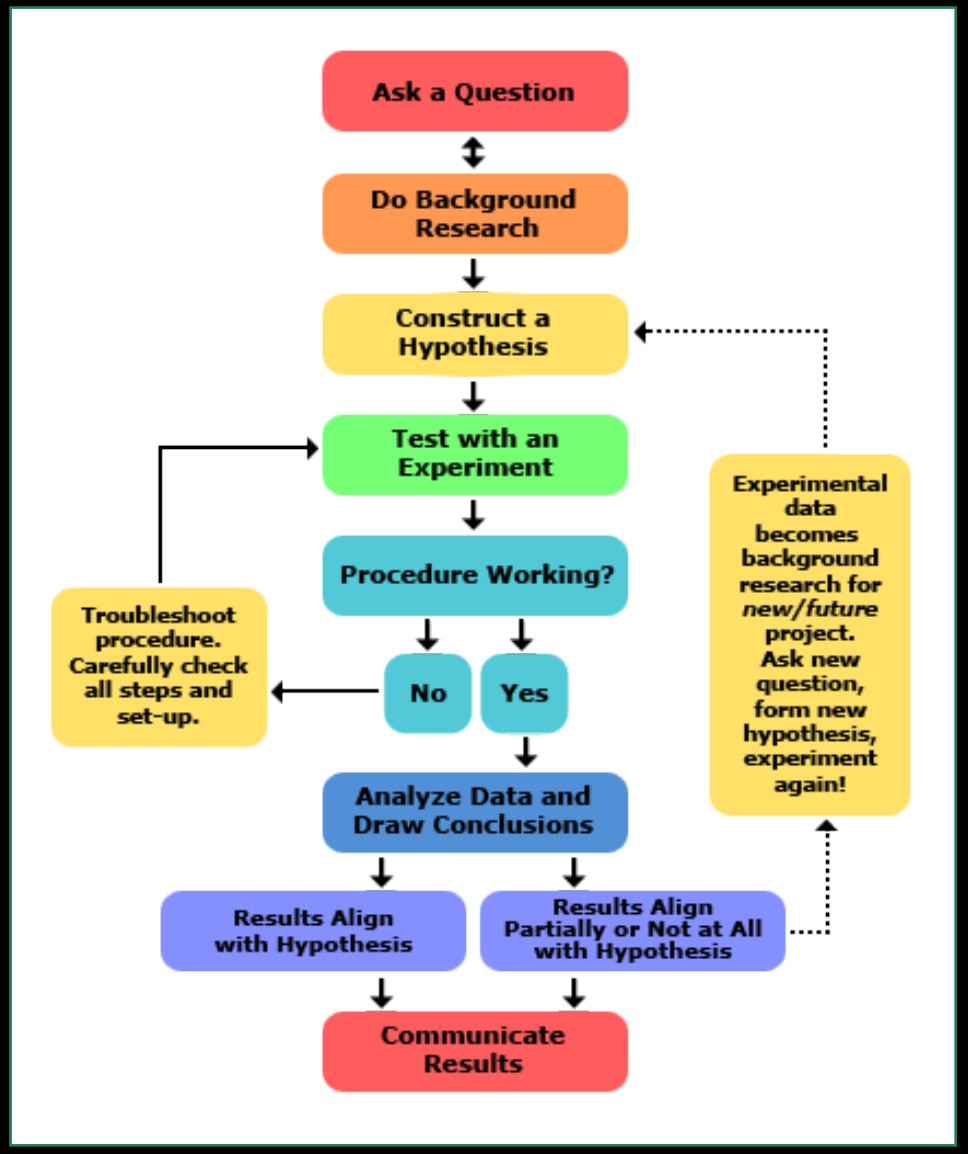
Data visualizations help us tell these stories.



5th Grade Science

Scientific Method

*The **scientific method** is an **empirical** method of acquiring knowledge that has characterized the development of **science** since at least the 17th century.*



Asking the right questions.

Sharing the right insights.

Asking the right questions.

- Adversary: How can I perform passive reconnaissance without touching the target?

Sharing the right insights.

Asking the right questions.

- Adversary: How can I perform passive reconnaissance without touching the target?
- CISO: What does our external attack surface look like?

Sharing the right insights.

Asking the right questions.

- Adversary: How can I perform passive reconnaissance without touching the target?
- CISO: What does our external attack surface look like?
- CEO: What does our company look like to our shareholders, customers, and competitors?

Sharing the right insights.

Passive Recon Demo

RAPID7 Open Data

Open Data

Offering researchers and community members open access to data from Project Sonar, which conducts internet-wide surveys to gain insights into global exposure to common vulnerabilities.

DATASETS: 13 FILES: 14,263 TOTAL SIZE: 27.4 TB

https://medium.com/@brevityinmotion/external-ip-domain-reconnaissance-and-attack-surface-visualization-in-under-2-minutes-b2ab06105def?sk=45a029919647bd3214e6dd1e8526ca25

17766	2019-07-27 16:42:43.000
17767	2019-07-27 16:45:10.000
17768	2019-07-27 16:36:42.000
17769	2019-07-27 16:38:06.000
17770	2019-07-27 16:36:53.000
17771	2019-07-27 04:33:12.000

17772 rows × 4 columns

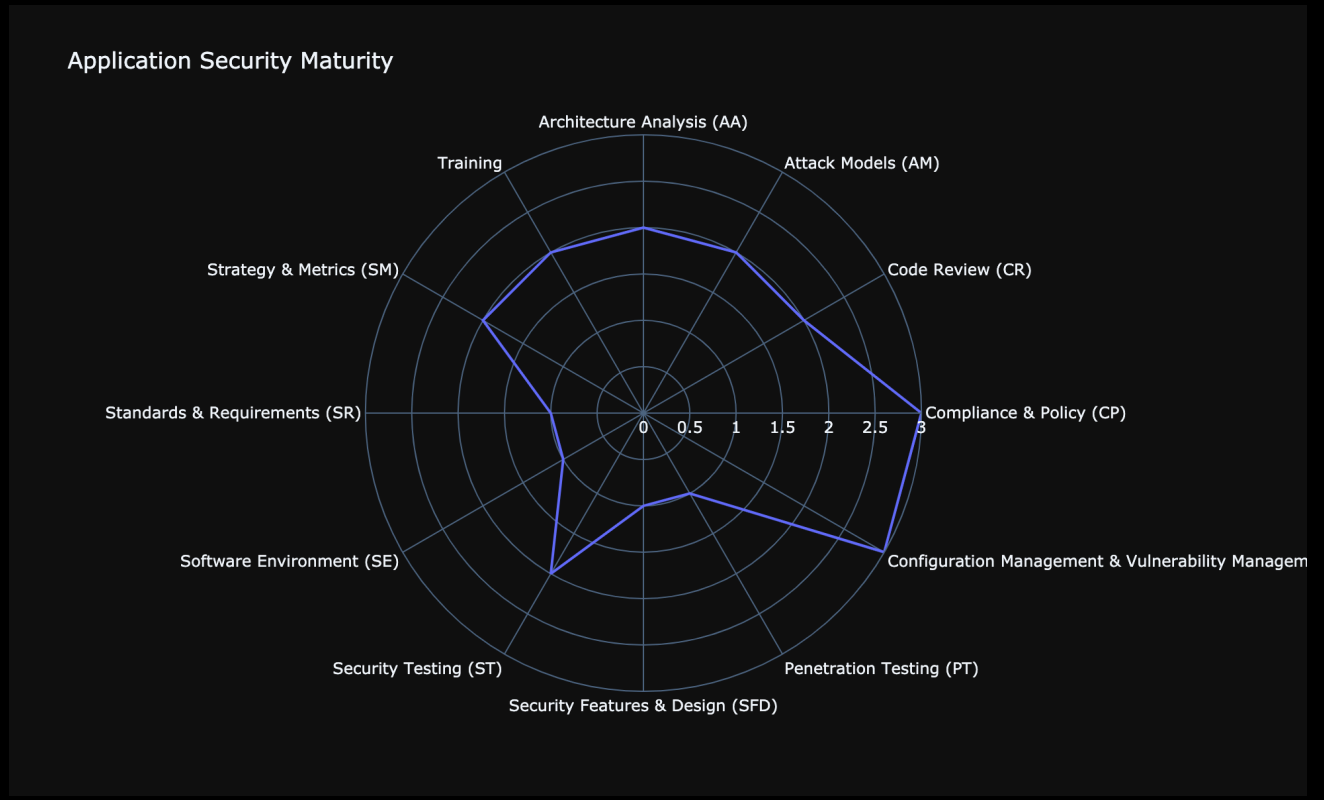
	latitude	longitude	count
0	-37.807	144.952	1
1	-36.851	174.768	1
2	-33.859	151.200	2
3	-33.810	151.131	4
4	-33.494	143.210	7
5	-22.831	-43.219	8
6	1.293	103.855	229
7	1.367	103.800	6
8	13.084	80.281	1
9	17.378	78.471	13
10	18.533	73.863	2
11	18.972	72.825	1
12	20.000	77.000	23
13	22.291	114.150	145
14	29.425	-98.493	335

Name	date_201907
Description	
Database	rapid7fqdns
Classification	parquet
Location	s3://rapid7-opendata/fdns/any/v1/date=201907/
Connection	
Deprecated	No
Last updated	Sat Aug 31 23:48:48 GMT-400 2019
Input format	org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat
Output format	org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat
Serde serialization lib	org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe
Serde parameters	serialization.format 1
Table properties	sizeKey 45714202983 objectCount 153 UPDATED_BY_ recordCount 2945369954 averageRecordSize 48 Craw

Program Value/Measurement

"Security technology management keeps the system fine tuned. But the secret sauce? That's in data enrichment. That's where the magic happens."

- Robert Herjavec



Category	Section	Activity Description	Activity	Level	Participant Percentage		Current	Target
SSDL Touchpoints	Security Testing (ST)	Ensure QA supports edge/boundary value cond	ST1.1	1	83.3%		TRUE	TRUE
SSDL Touchpoints	Security Testing (ST)	Drive tests with security requirements and secu	ST1.3	1	73.3%		TRUE	TRUE
SSDL Touchpoints	Security Testing (ST)	Integrate black-box security tools into the QA p	ST2.1	2	25.0%		FALSE	TRUE
SSDL Touchpoints	Security Testing (ST)	Share security results with QA.	ST2.4	2	11.7%		FALSE	TRUE
SSDL Touchpoints	Security Testing (ST)	Include security tests in QA automation.	ST2.5	2	10.0%		FALSE	FALSE
SSDL Touchpoints	Security Testing (ST)	Perform fuzz testing customized to application	ST2.6	2	10.8%		TRUE	TRUE
SSDL Touchpoints	Security Testing (ST)	Drive tests with risk analysis results.	ST3.3	3	3.3%		FALSE	FALSE
SSDL Touchpoints	Security Testing (ST)	Leverage coverage analysis.	ST3.4	3	2.5%		FALSE	FALSE
SSDL Touchpoints	Security Testing (ST)	Begin to build and apply adversarial security te	ST3.5	3	2.5%		FALSE	FALSE
Deployment	Penetration Testing (PT)	Use external penetration testers to nd problem	PT1.1	1	87.5%		TRUE	TRUE
Deployment	Penetration Testing (PT)	Feed results to the defect management and mit	PT1.2	1	74.2%		TRUE	TRUE

Asking the right questions.

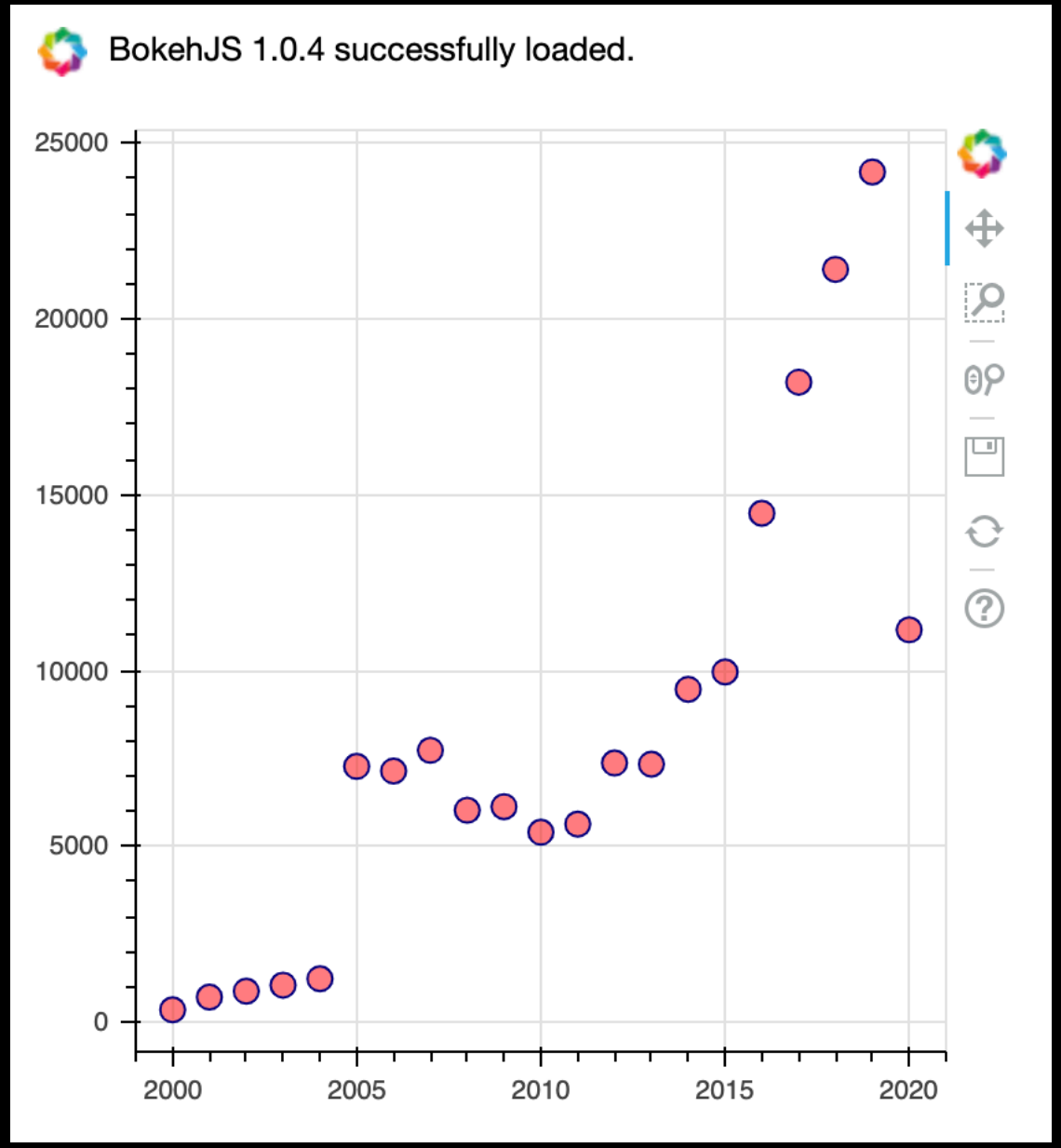
- Patch Management Team: Why are we so busy over the holidays?
- Security Manager: At what times do we need to increase staffing?
- CISO: Why are our vulnerability counts increasing?

Sharing the right insights.

Program Value/Measurement

Common Vulnerabilities & Exposures (CVEs)

- Growth over the past 20 years.



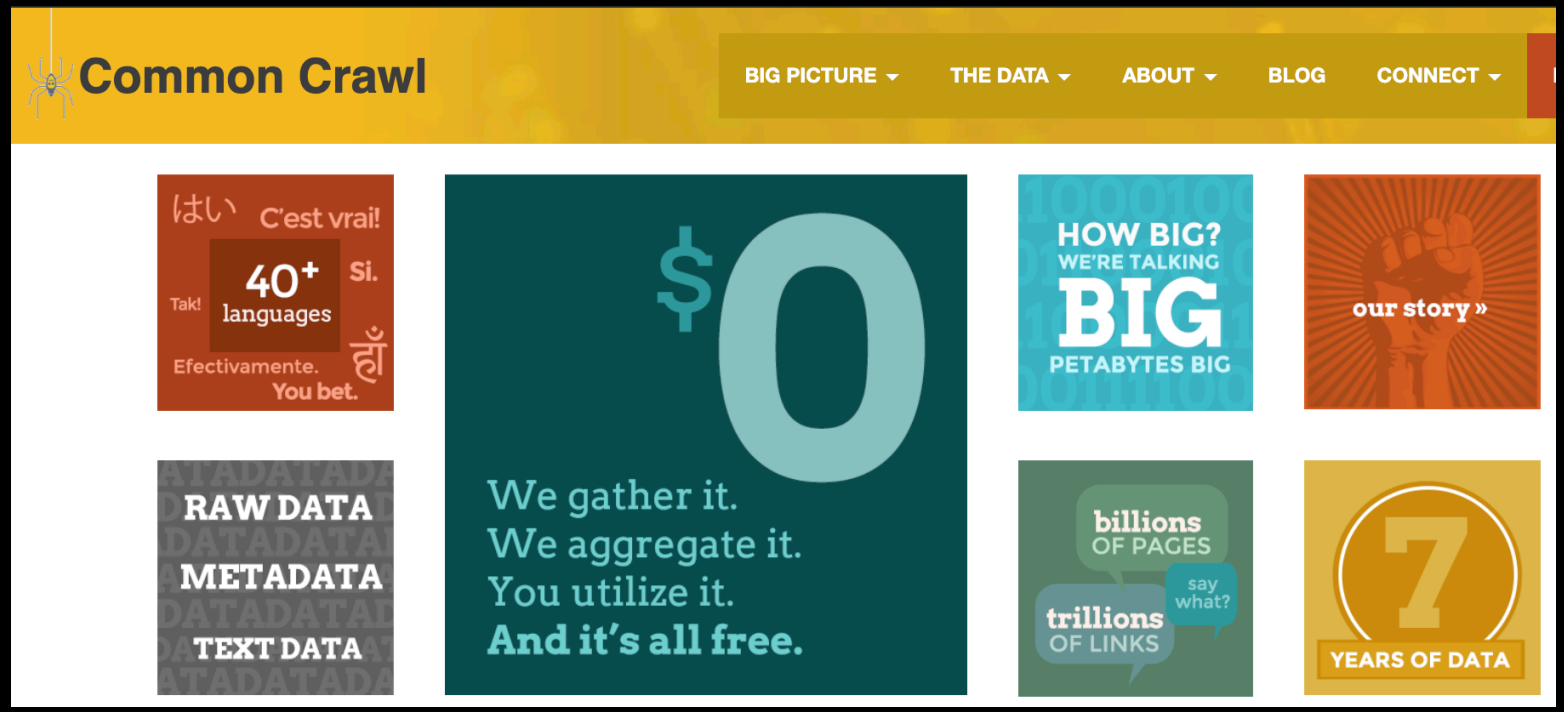
	Month	Count
0	1	20184
1	2	12733
2	3	14568
3	4	12006
4	5	11882
5	6	13866
6	7	11862
7	8	13621
8	9	12698
9	10	12492
10	11	14476
11	12	22983

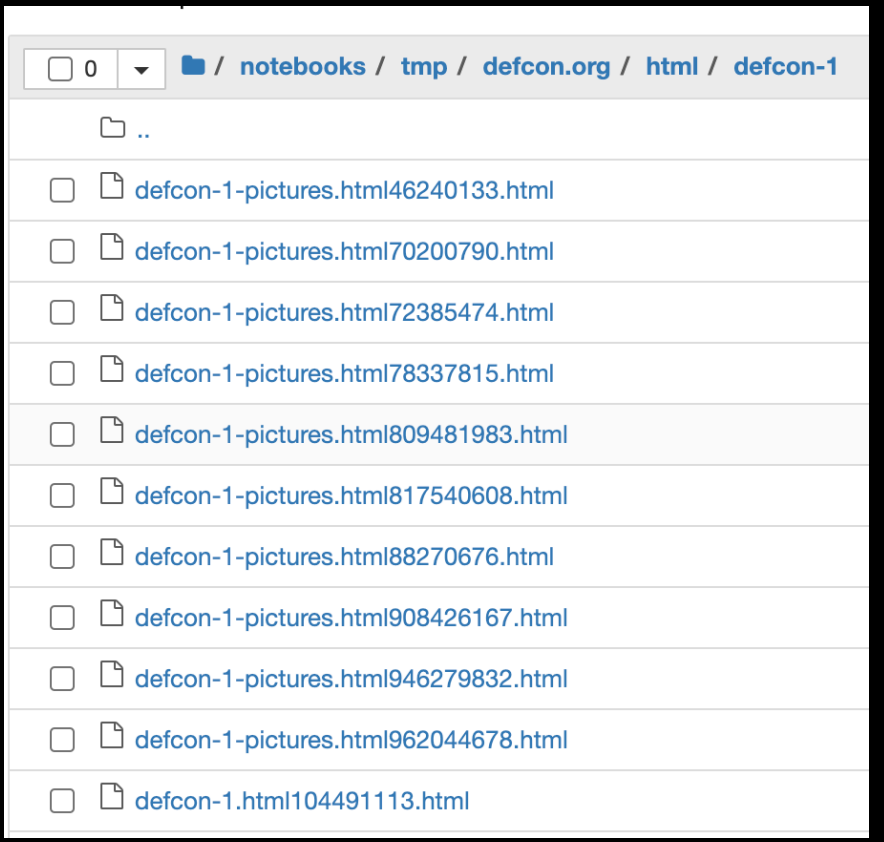
	Year	Count
0	1999	249
1	2000	351
2	2001	713
3	2002	881
4	2003	1053
5	2004	1234
6	2005	7274
7	2006	7142
8	2007	7731
9	2008	6026
10	2009	6129
11	2010	5402
12	2011	5632
13	2012	7371
14	2013	7337
15	2014	9466
16	2015	9960
17	2016	14473
18	2017	18202
19	2018	21409
20	2019	24179
21	2020	11157

Common Crawl Demo

69599	https://media.defcon.org/
69600	https://media.defcon.org/DEF%20CON%2021/
69601	https://media.defcon.org/DEF%20CON%2021/DEF%20...
69602	https://media.defcon.org/DEF%20CON%2021/DEF%20...
69603	https://media.defcon.org/DEF%20CON%2021/DEF%20...
69604	https://media.defcon.org/DEF%20CON%20Conferenc...
69605	https://media.defcon.org/Hacking%20Related%20D...

69606 rows × 4 columns





Challenges to Solve

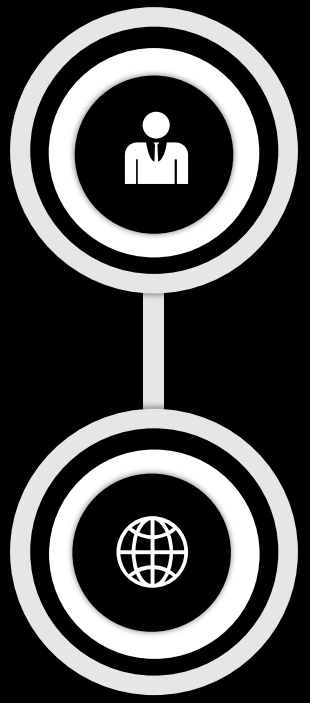
Expertise must keep up with an evolving attack surface and a dynamic nature of threats.

Competition for talent

"the average CISO tenure is only about 24 to 48 months"

- 38% leave for higher compensation
- 36% leave for lack of security culture
- 34% leave for not being active members with executive management or board of directors
- 31% leave from lack of budget

Challenges to Solve



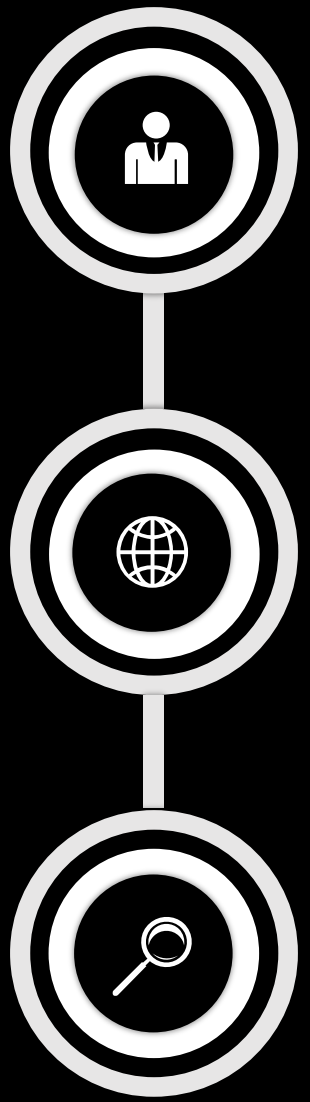
Expertise must keep up with an evolving attack surface and a dynamic nature of threats.

Competition for talent

Programs are reactive and rarely maintain a multi-year roadmap.

Program continuity

Challenges to Solve



Expertise must keep up with an evolving attack surface and a dynamic nature of threats.

Competition for talent

Programs are reactive and rarely maintain a multi-year roadmap.

Program continuity

Visibility and speed of execution must improve with automation.

Continuous risk measurement

Thank you!

All code and slides from this presentation are at:
https://github.com/brevityinmotion/straylight

- Name: Ryan Elkins
- Twitter: @ryanelkins
- Email: ryan-elkins@outlook.com
- Github: brevityinmotion
- Medium: brevityinmotion

Brevity In Motion