

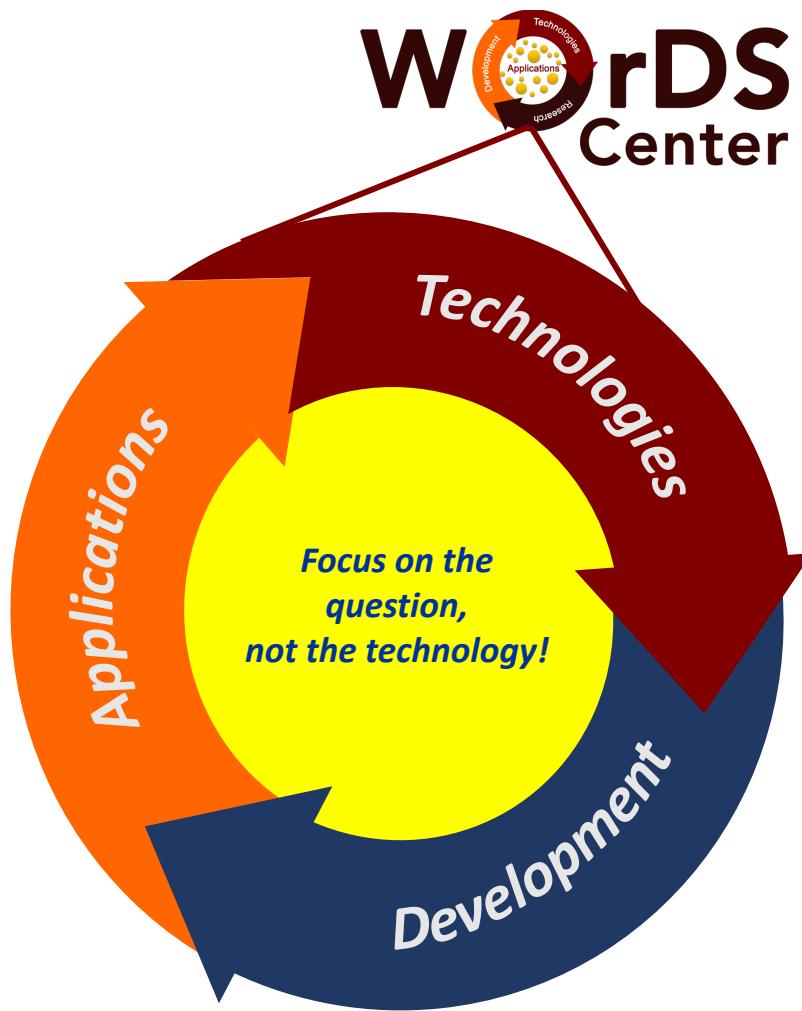
A Short Introduction to Data Science and its Applications

Aug 11, 2023

Shweta Purawat

Director of Product Management

Workflows for Data Science Center of Excellence and WIFIRE Lab
University of California, San Diego



Workflows for Data Science Center of Excellence at SDSC

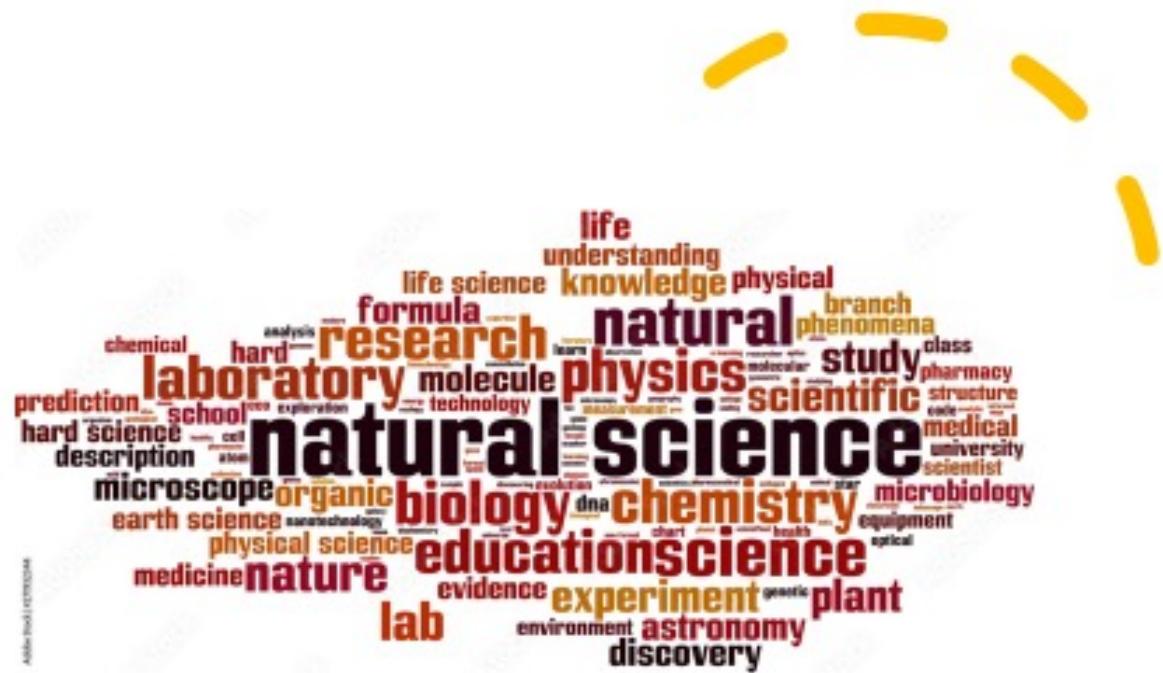
<http://WorDS.sdsc.edu>

Research and Development Mission:

- *Methodology and tool innovation to enable collaborative workflow-driven science*
- *Create solutions on top of big data and advanced computing platforms.*

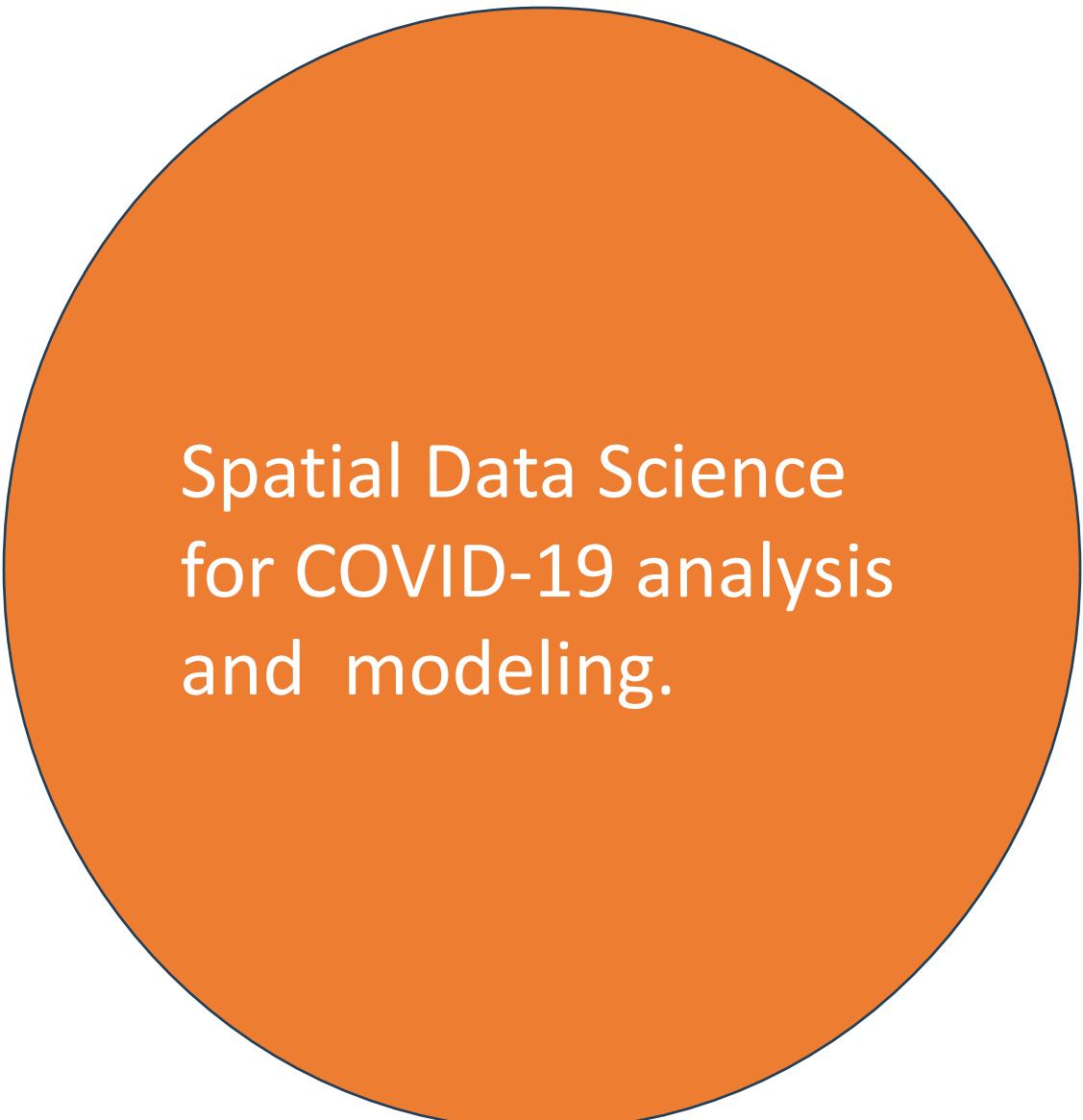
“Big” Data, Data Science and “AI”, Cyberinfrastructure, and Their Applications

Knowledge Management for Scientific Application Using Polystore



Subhasis Dasgupta, Ph.D.

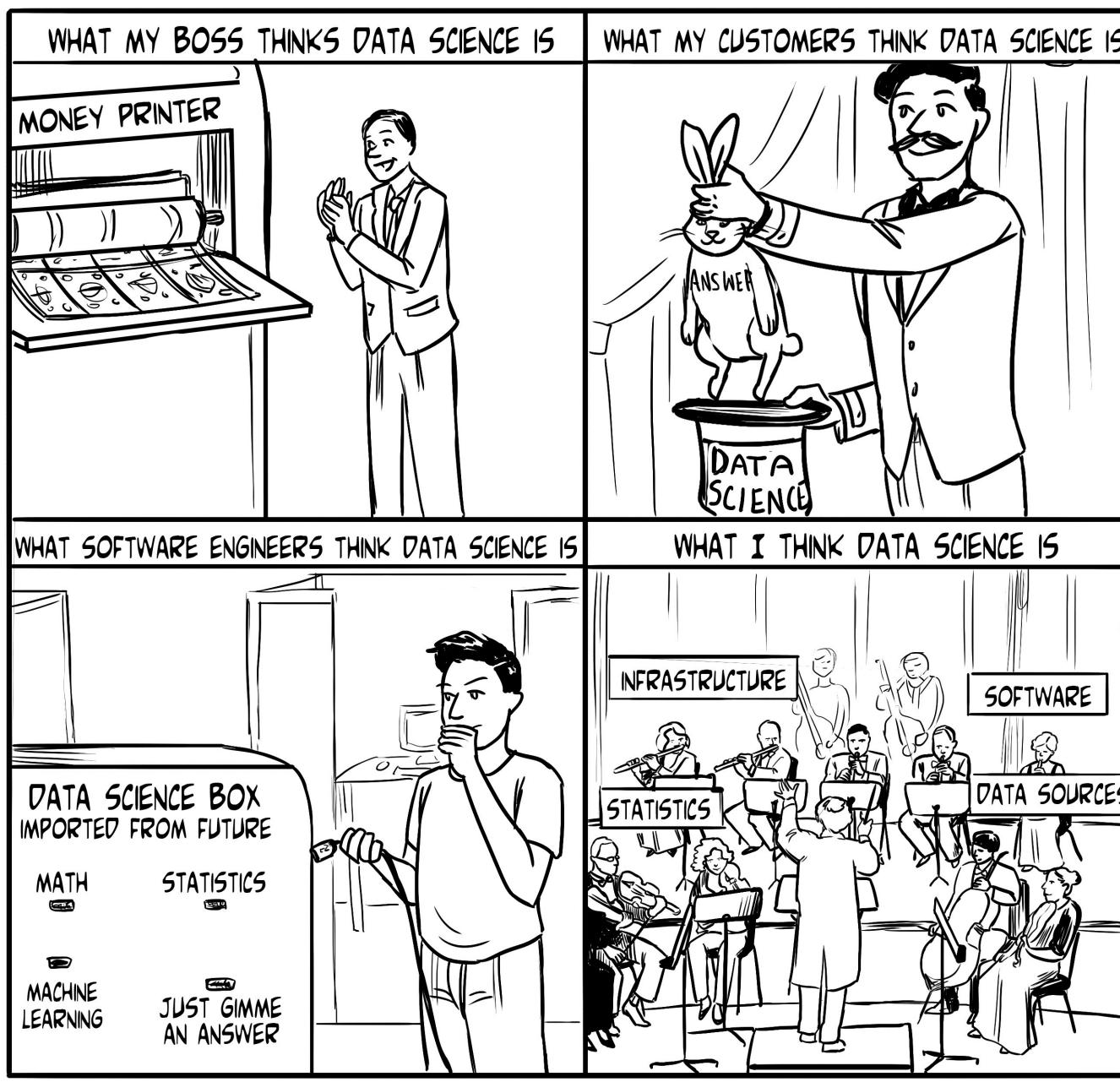
Assistant Scientist, University of California San Diego



Spatial Data Science for COVID-19 analysis and modeling.

Johnny Lei
Computational and Data Researcher,
UCSD

What is Data Science?



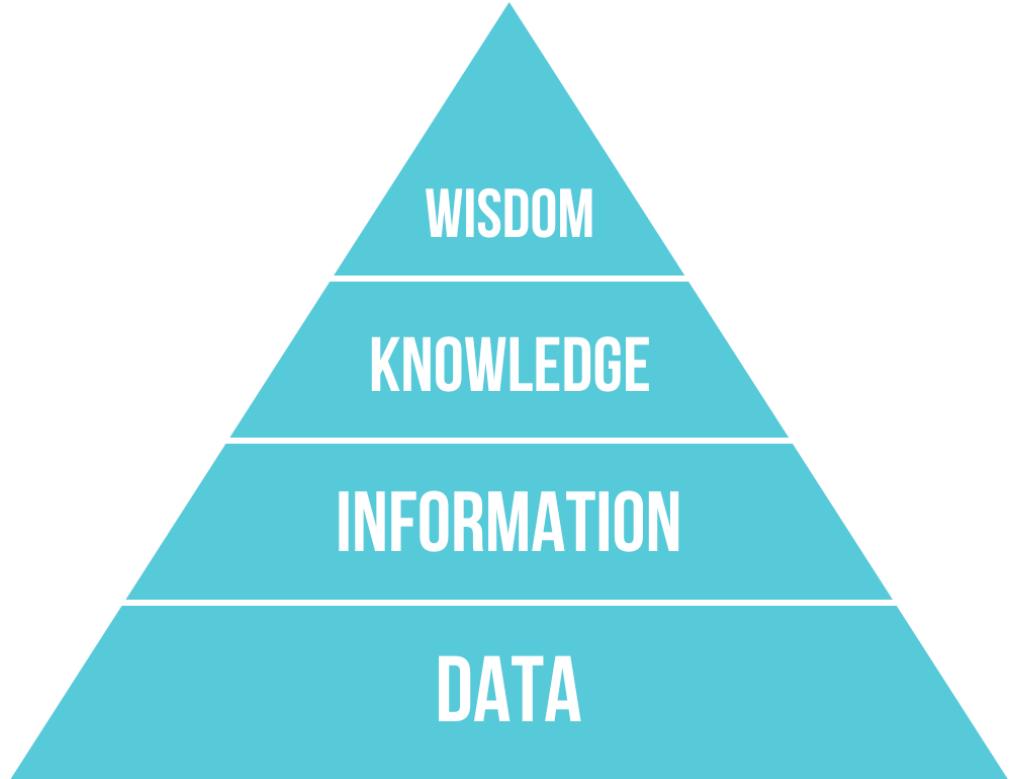
So, what exactly is data science?

Image Source:

Think Like a Data Scientist,
a book by Brian Godsey.

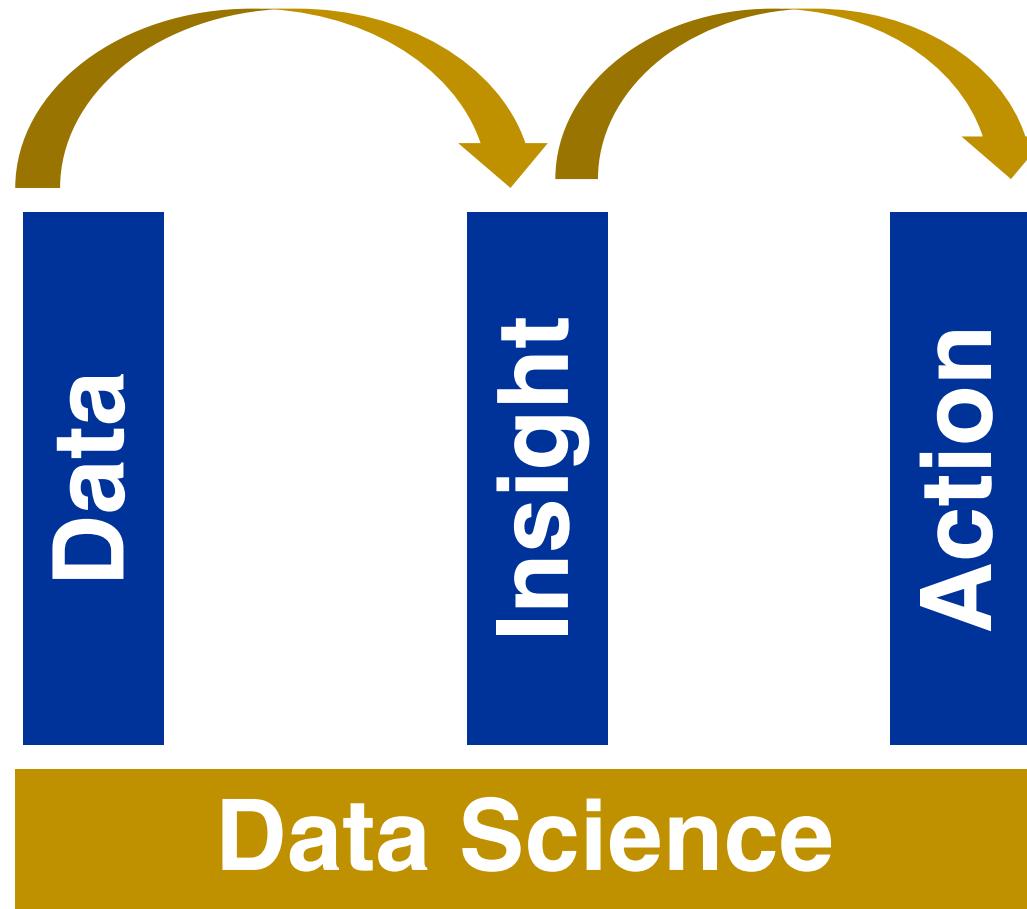
<https://www.manning.com/books/think-like-a-data-scientist>

What is data?



data: /'dædə, 'dādə/
facts and statistics collected together for reference or analysis.

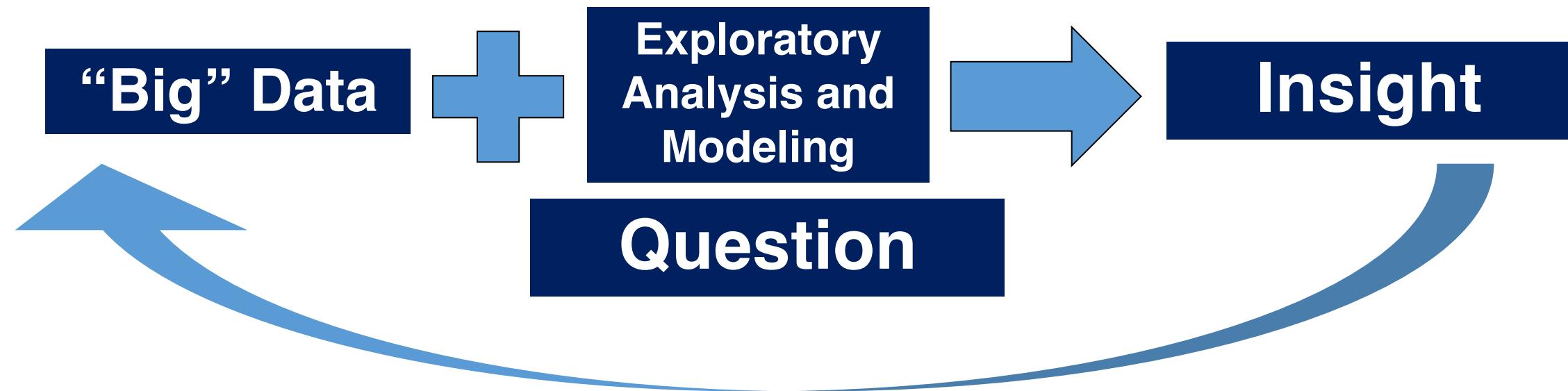
The Ultimate Goal of Data Science



Where is the science in data science?

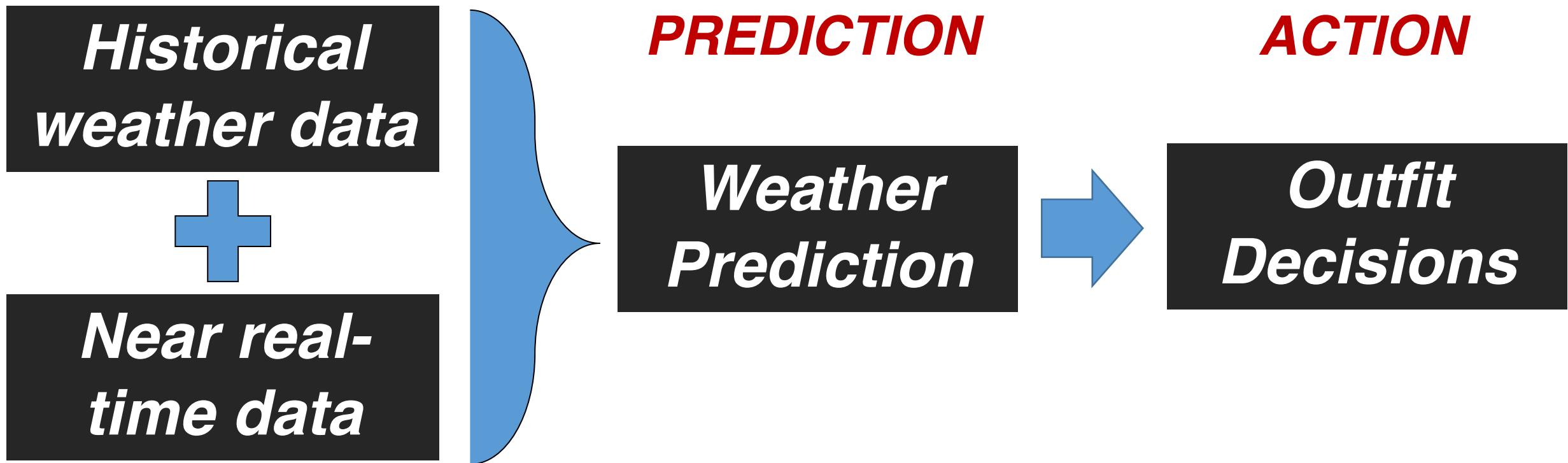


Generating Insights from Data

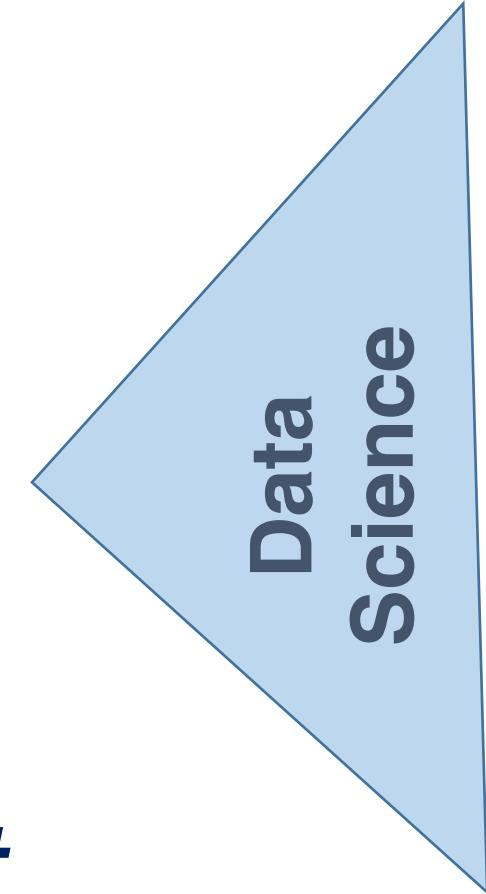


Insight Data Product

Example: Weather Forecast to Actions



***Systems and models
that help us to
understand data
in order to
gain insights and
make predictions
leading to action for impact.***



Data Science is “IMPACT” Science!

Data Science in Everyday Life

- Finance
- Healthcare
- Entertainment
- Travel
- Shopping
- ...



Smart Cities

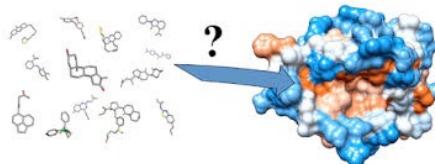


Personalized Precision Medicine



Smart Manufacturing

Computer-Aided Drug Discovery

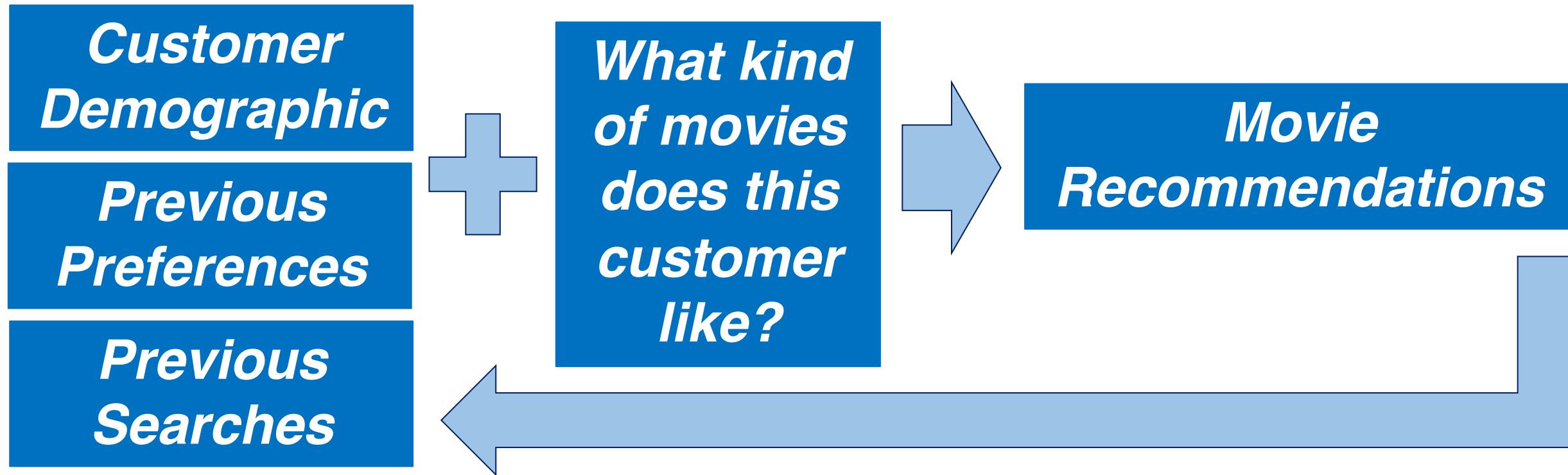


Disaster Resilience and Response



Smart Grid and Energy Management

Entertainment Industry: Movie Recommendations



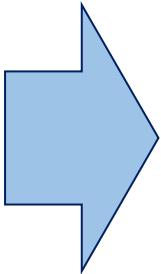
Entertainment Industry: Movie Recommendations

PREDICTION

What kind of movies do customers like?

ACTION

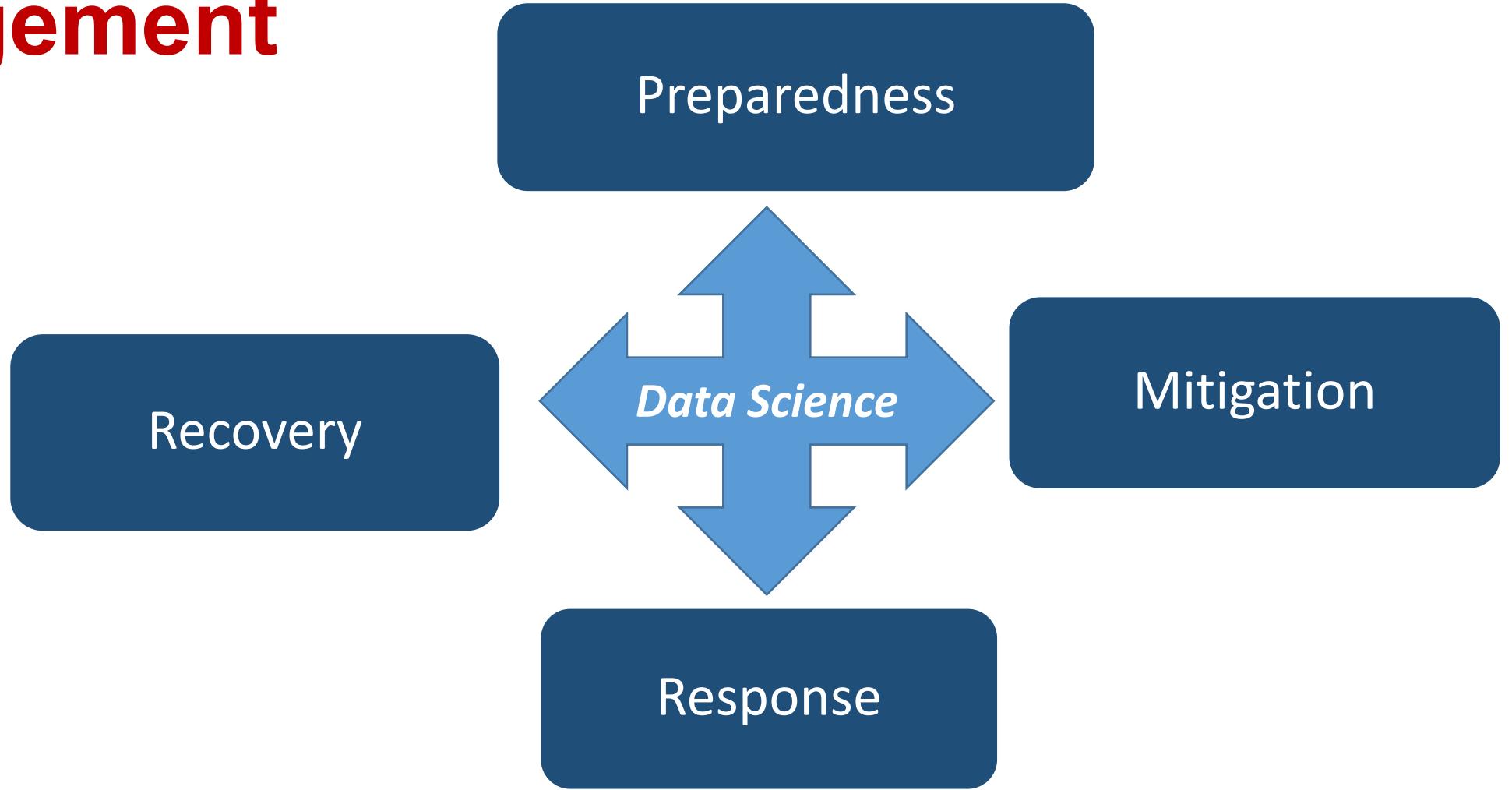
What movies to produce and market?



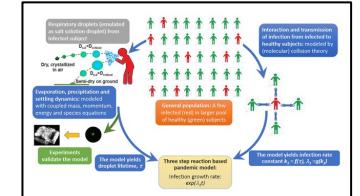
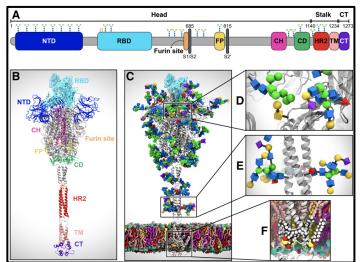
Example Industry: Smart Healthcare



Safety: Data Science for Emergency Management



Data Science Impact to Fight COVID-19



- Disease data
- Virus models
- Wearables
- Public health surveys
- Social media
- Contact data
- ...
- Spread modeling
- Understanding the virus
- Drug design
- Vaccine development
- Economical impact
- Mental health

<https://www.tacc.utexas.edu/~sugar-coating-locks-and-loads-coronavirus-for-infection>

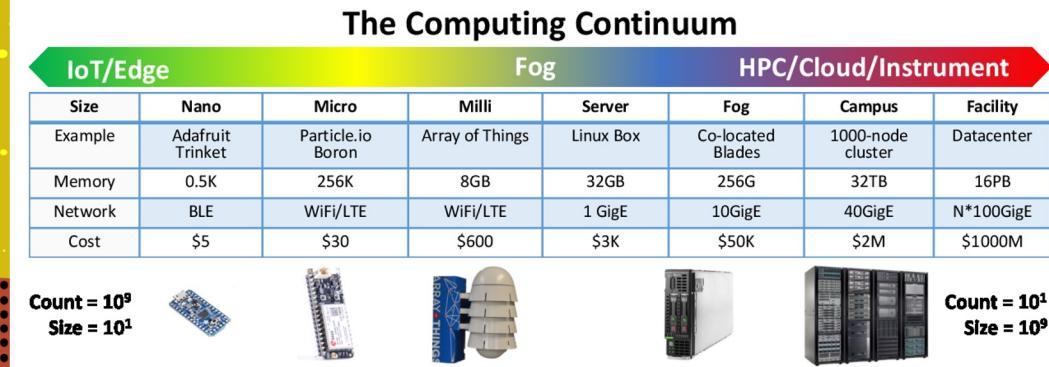
<https://phys.org/news/2020-06-respiratory-droplet-motion-evaporation-covid-type.html>

- Decision-support
- Policy making
- Vaccine
- Treatment
- Cure
- ...

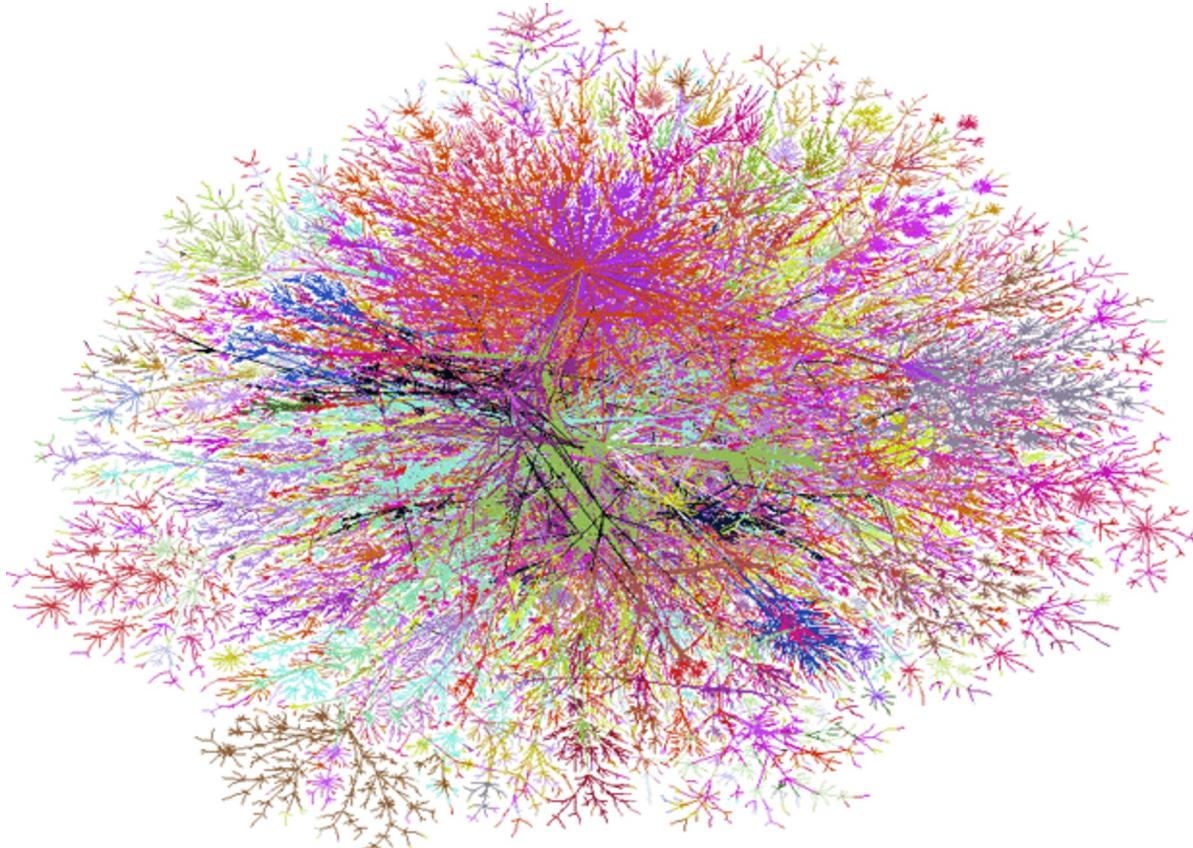
Why are we talking about data science and AI now?



Why is data science so popular?



What is Big Data?

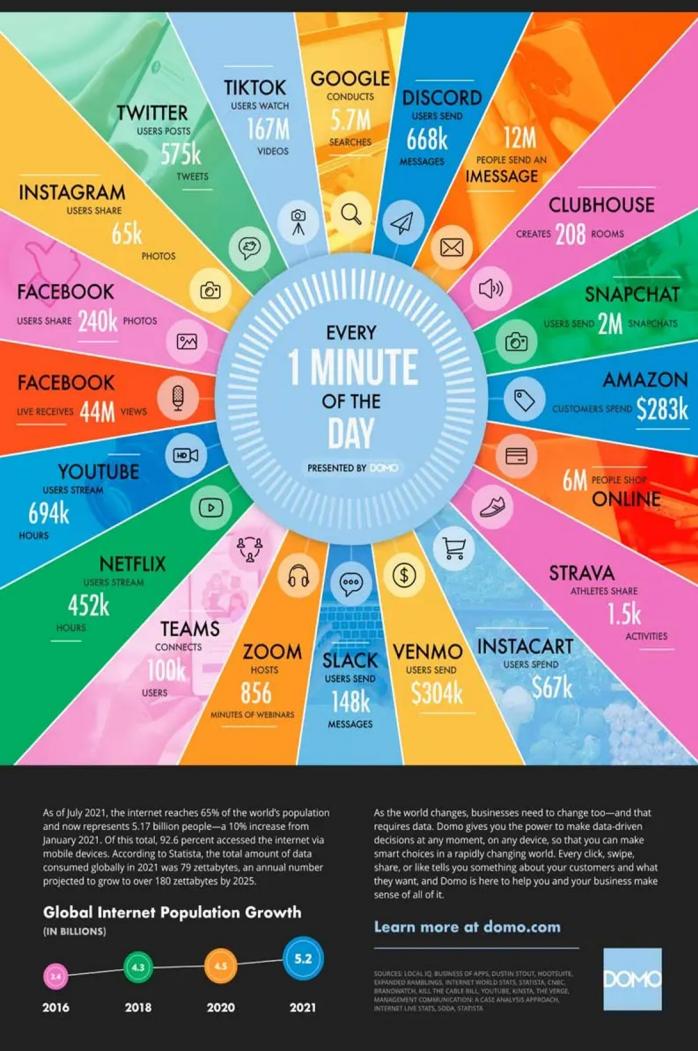




Data Never Sleeps 9.0

How much data is generated every minute?

The 2020 pandemic upended everything, from how we engage with each other to how we engage with brands and the digital world. At the same time, it transformed how we eat, how we work and how we entertain ourselves. Data never sleeps and it shows no signs of slowing down. In our 9th edition of the "Data Never Sleeps" infographic, we bring you a glimpse of how much data is created every digital minute in our increasingly data-driven world.



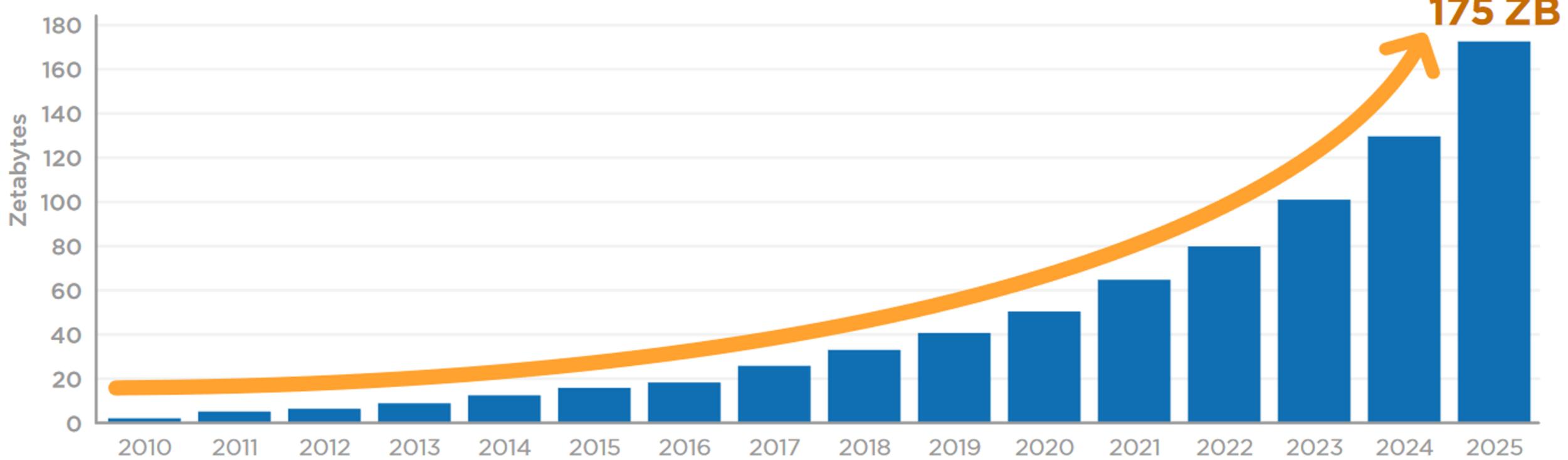
THE INTERNET IN 2023 EVERY MINUTE



Created by: eDiscovery Today & LTMG

Shweta Purawat (shpurawat@ucsd.edu)
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

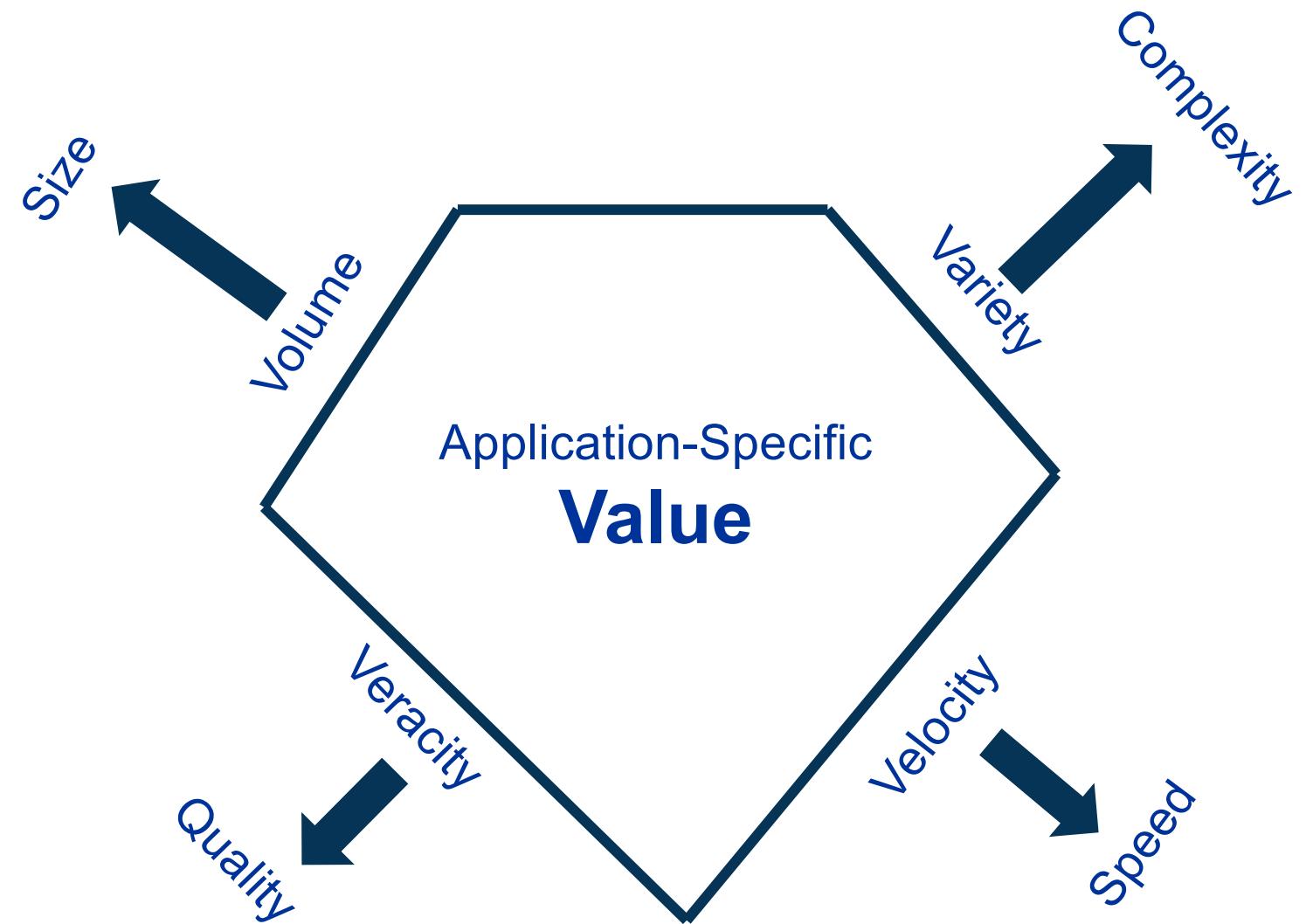
Annual global data size



<https://www.datanami.com/2022/01/11/big-growth-forecasted-for-big-data/>

175 ZB = 175 Trillion GB!

V's of Big Data



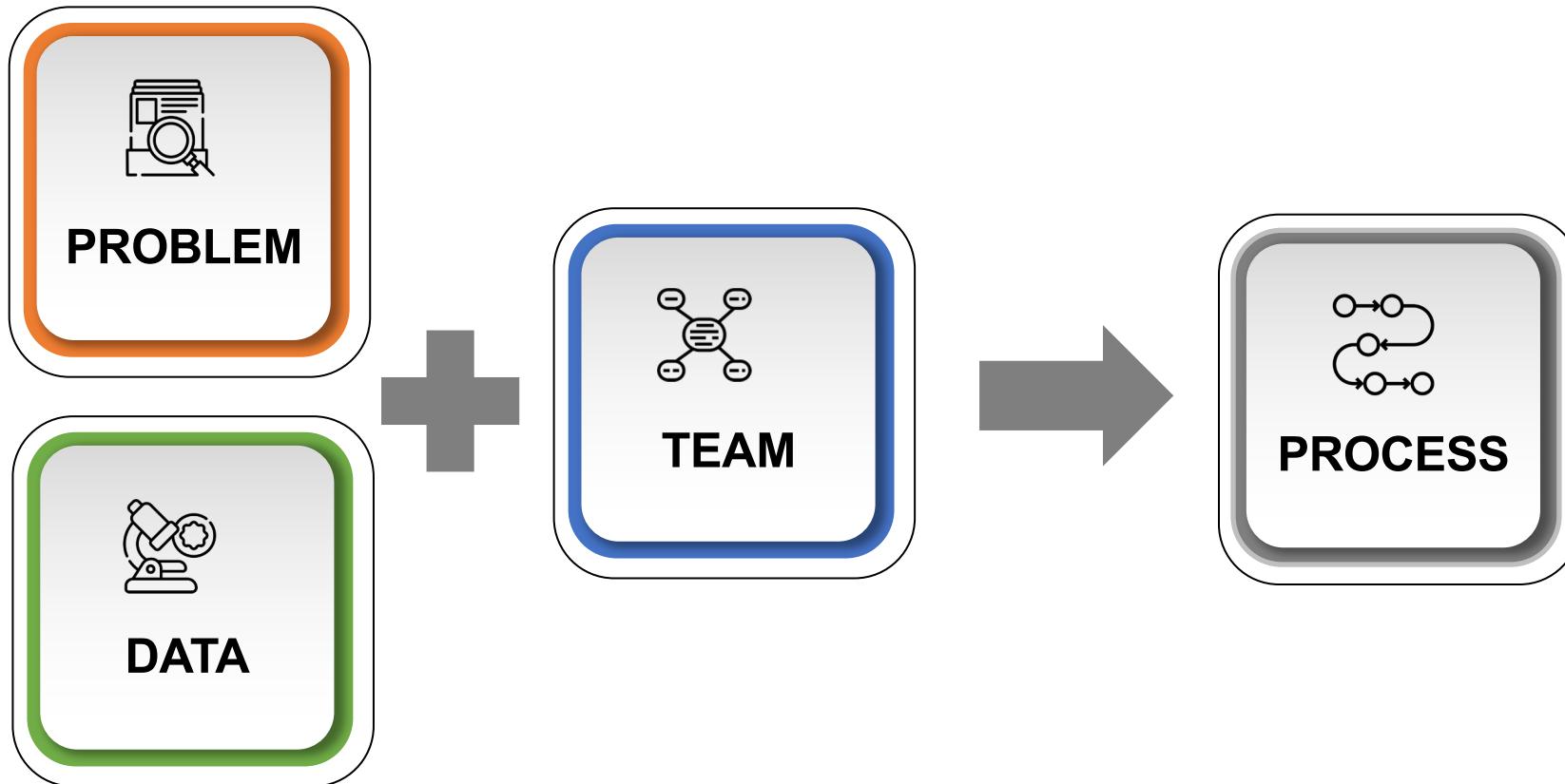
Data Science: Finding connections in big data... ... and applications using the connections for a purpose.

How does data science happen?

or

How to manage data science projects?

Dimensions of Data Science



Asking the Right Question



- Define the problem
- Assess the situation
- List objectives

Data Science Process



- Data engineering
- Data analysis
- Machine learning
- Scalable computing

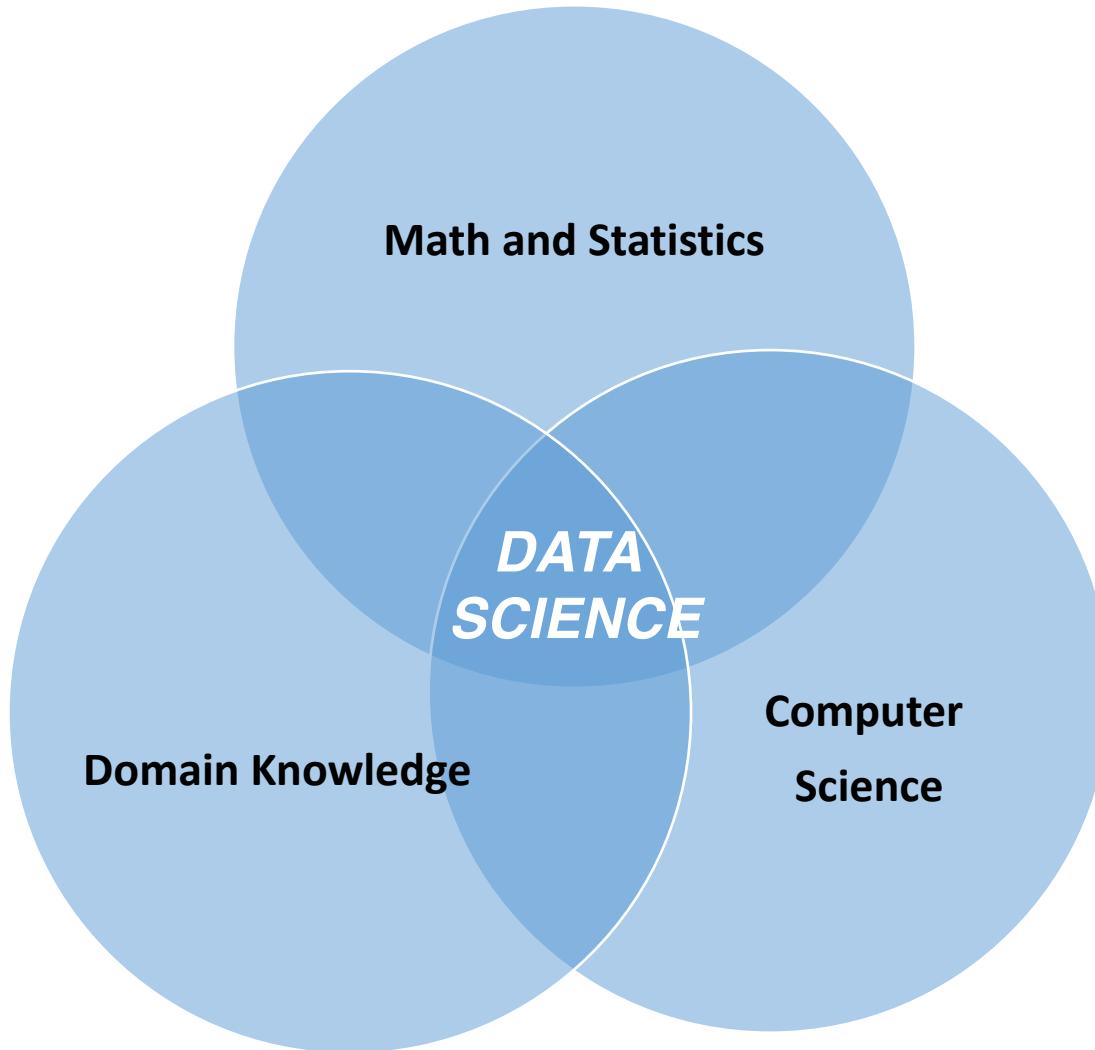
Phases of the Data Science Process

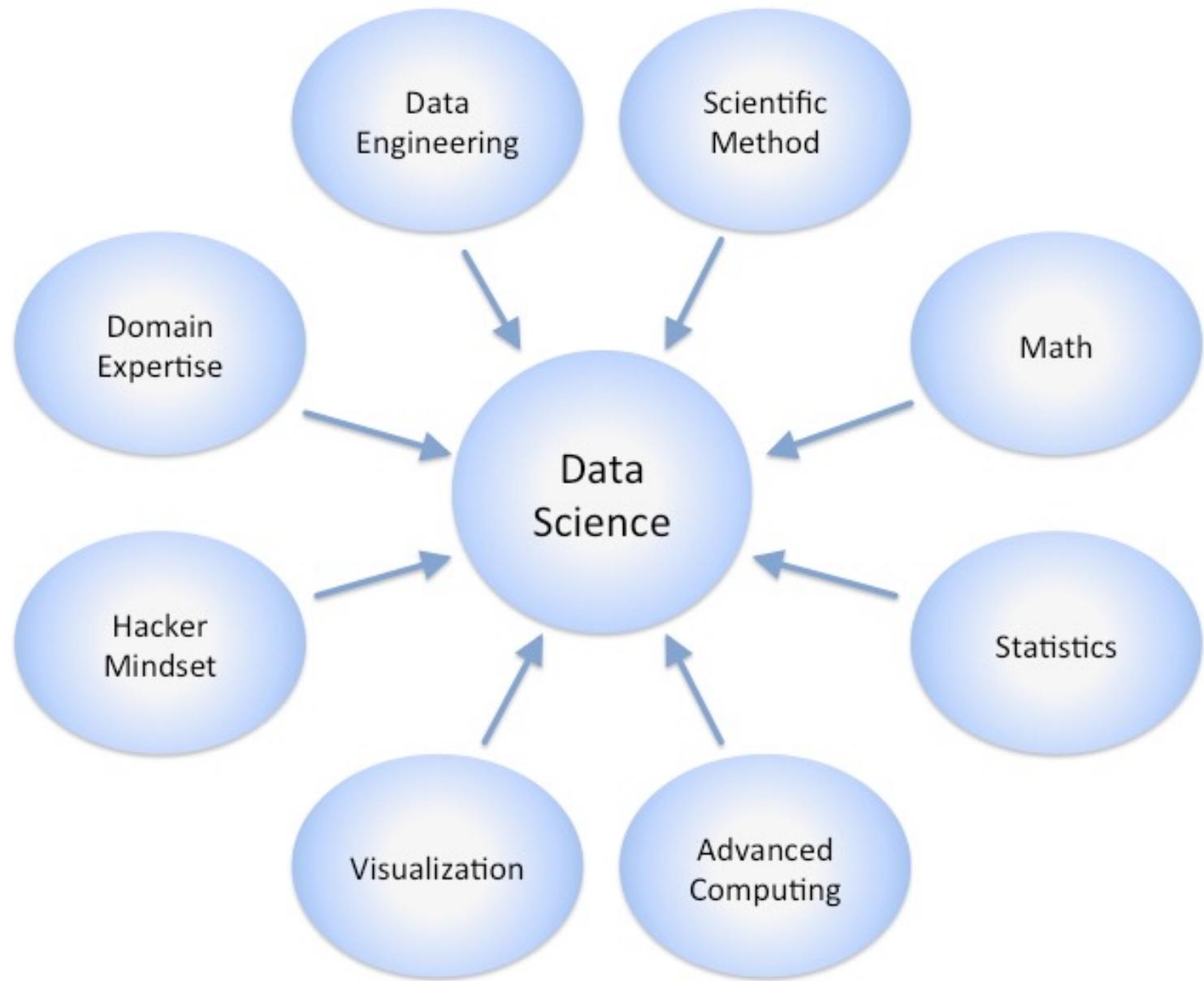


- **BUILD**
- **EXPLORE**
- **SCALE**

What is a Data Scientist?

Data science is multidisciplinary.





MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants



PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing package e.g. R
- ★ Databases SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants



PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing packages, e.g., R
- ★ Databases: SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing
DISTILLERY

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

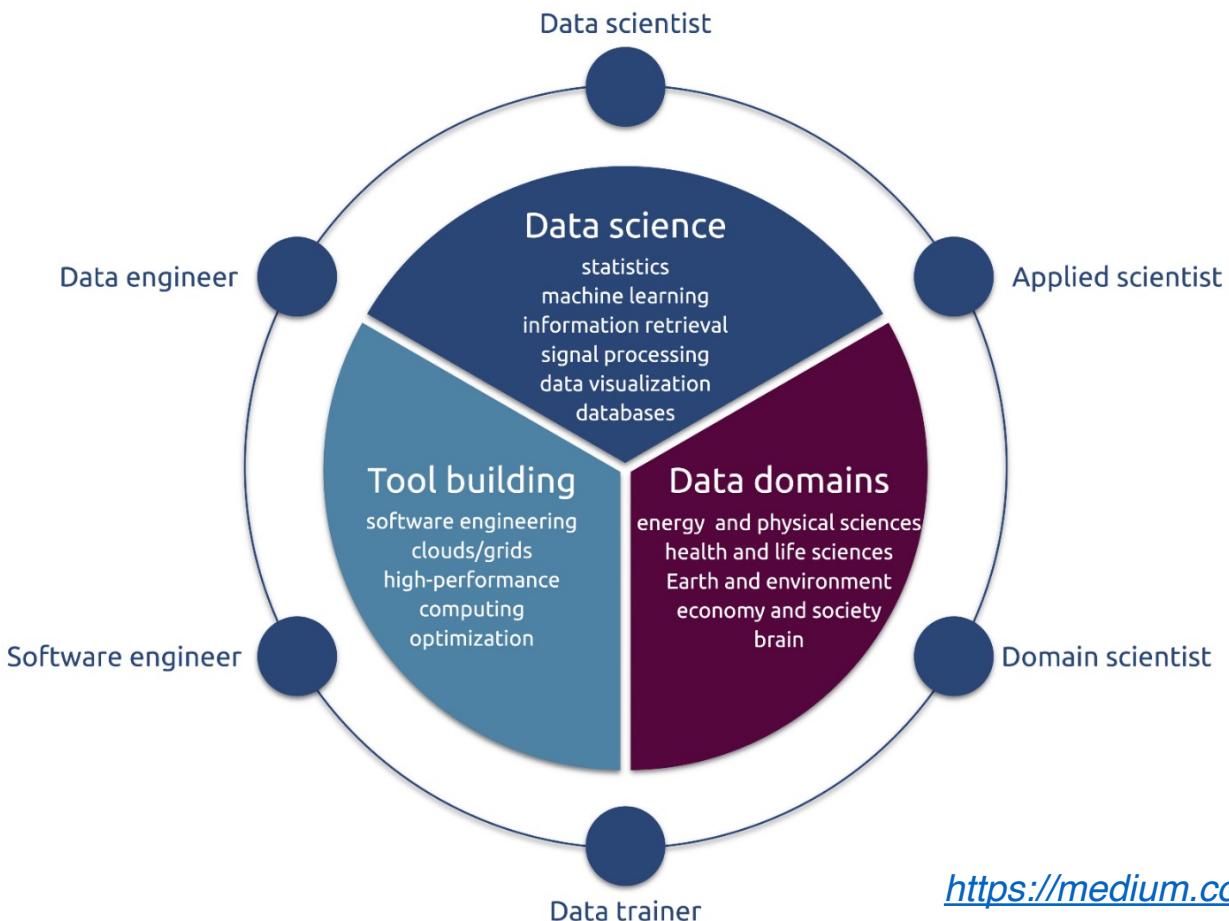
Marketing
DISTILLERY
(c) Krzysztof Zawadzki

Data science is “WE” science!



- Team collaboration
- Curiosity
- Respect

Expertise and Skills often overlap between different data science roles.



The data science ecosystem: activities and actors

- Data engineer
- Data analyst
- Methods expert
- Scalability and operations expert
- Business manager
- Business analyst
- Scientist
- Visualization and dashboard developer
- Solution architect
- Story teller/coordinator
- Project manager

<https://medium.com/@balazskegl/the-data-science-ecosystem-678459ba6013>

Team Data Science Environments

Using the Team Data Science Process with Azure Machine Learning

The Team Data Science Process (TDSP) provides a systematic approach to building intelligent applications that enables teams of data scientists to collaborate effectively over the full lifecycle of activities needed to turn these applications into products.

This screenshot shows a Microsoft Azure documentation page. The top navigation bar includes 'Microsoft' logo, 'Docs', 'Documentation' (which is underlined), 'Learn', 'Q&A', 'Code Samples', 'Shows', and 'Events'. Below this is another navigation bar for 'Azure' with links like 'Product documentation', 'Architecture', 'Learn Azure', 'Develop', and 'Resources'. On the left, there's a sidebar with a link to 'What is the Team Data Science Process?'. The main content area has a heading 'What is the Team Data Science Process?' and a search bar labeled 'Filter by title' with options for 'Containers' and 'Databases'.

This screenshot shows the IBM Garage website. The header includes the 'IBM' logo and links for 'IBM Cloud Blog', 'Why IBM', 'Products', 'Solutions', 'Garage', 'Pricing', 'Blogs', 'Docs', and 'Support'. The main section features a large image of a sunset with the text 'The IBM Garage expands to Data Science Insights' overlaid. A button at the bottom says 'Schedule a visit to a Garage'.

This screenshot shows the Cloudera Data Science Workbench homepage. It features a large image of a spiral staircase. The top navigation bar includes 'CLOUDERA', 'Why Cloudera', 'Products', 'Solutions', and 'Services & Support'. A banner says 'Looking for cloud-native machine learning on CDP? Visit Cloudera Machine Learning'. Below this is a section titled 'Cloudera Data Science Workbench' with a 'See what's new' button. Another section discusses 'TDSP' (Team Data Science Process) and its benefits. At the bottom, there are links for 'Overview', 'Benefits & features', and 'Resources'.

This screenshot shows the JupyterHub website. It features the Jupyter logo and the text 'jupyterhub'. Below this is a description: 'A multi-user version of the notebook designed for companies, classrooms and research labs'. The main content area has a heading 'What is JupyterHub?' and a detailed description of its features, mentioning it runs in the cloud or on hardware and serves pre-configured data science environments.

Worldwide, Mar 2023 compared to a year ago:

Rank	Change	Language
1		Python
2		Java
3		JavaScript
4		C/C++
5		C#
6		PHP
7		R
8	↑	TypeScript
9	↑	Swift
10	↓↓	Objective-C

Top Data Science Programming Languages

<https://pypl.github.io/PYPL.html>

Source: PYPL

Shweta Purawat (shpurawat@ucsd.edu)
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Jupyter Notebooks and JupyterLab



The screenshot shows the Jupyter Notebook interface. On the left, there's a file tree with notebooks like 'Linear Regression.ipynb' and 'Lorenz.ipynb'. The main area displays a notebook titled 'In Depth: Linear Regression' with code and text. Below it, a Python 3 IDE window shows code related to linear regression. To the right, there's a 'Launcher' window showing icons for various kernels (Python 3, C++11, C++14, C++17, Julia 1.1, Python 3.7, R) and a 'Console' window showing a Julia session.

<https://jupyter.org/>

JupyterLab: A Next-Generation Notebook Interface

JupyterLab is the latest web-based interactive development environment for notebooks, code, and data. Its flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning. A modular design invites extensions to expand and enrich functionality.

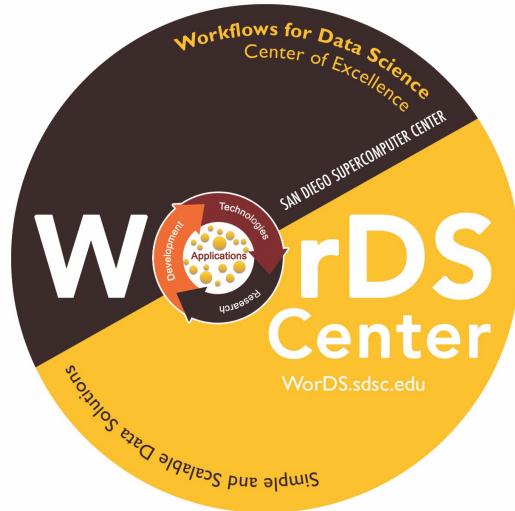
Try it in your browser

Install JupyterLab

*Shweta Purawat
Ilkay Altintas, Ph.D.*

Contact:

*Email: shpurawat@ucsd.edu
Email: ialtintas@ucsd.edu*



<https://words.sdsc.edu/>

<https://wifire.ucsd.edu/>



We are hiring! -- <https://words.sdsc.edu/careers>

Questions?



Office of
Science

The presented work is collaborative work with many wonderful individuals, and parts of it are funded by NSF, DOE, NIH, UC San Diego and various industry, government and foundation partners.