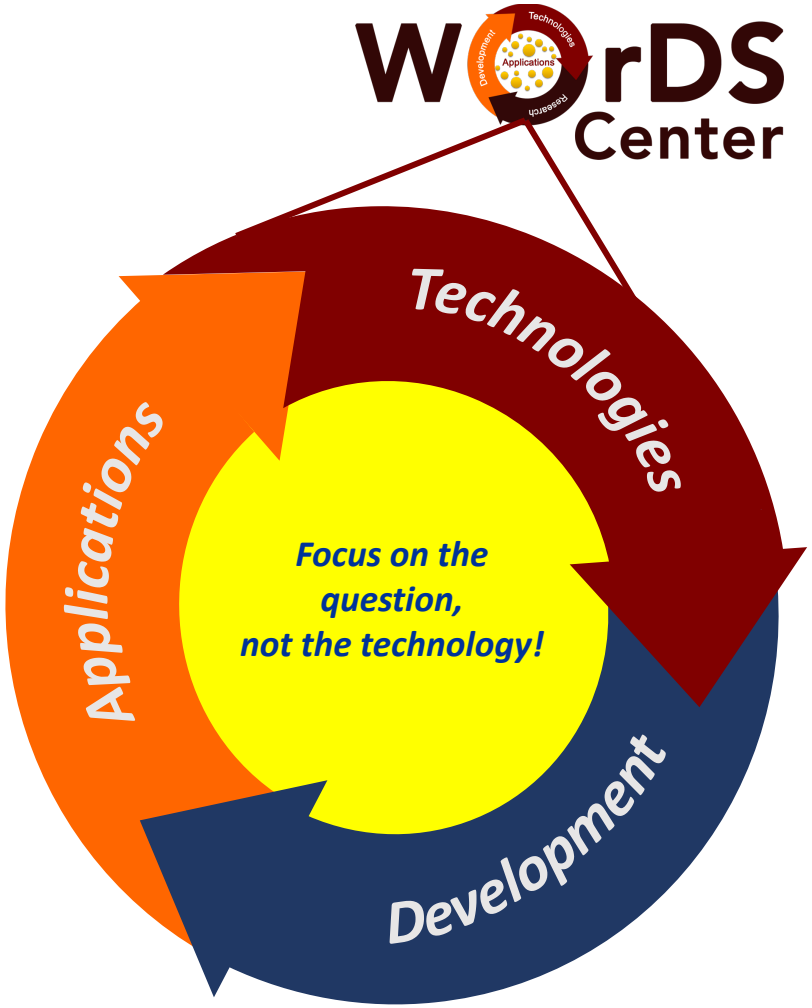
A Short Introduction to Data Science and its Applications

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.*

SDSC SAN DIEGO SUPERCOMPUTER CENTER

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

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

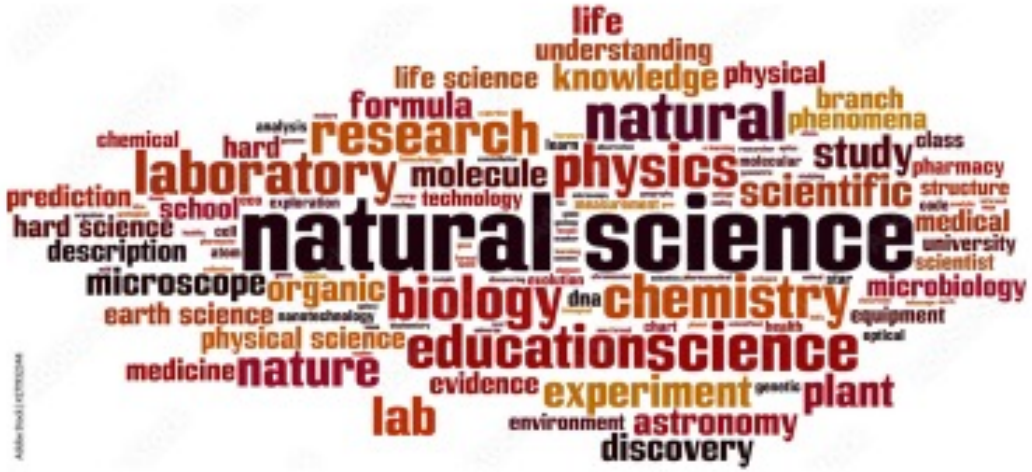
“Big” Data, Data Science and “AI”, Cyberinfrastructure, and Their Applications

SDSC SAN DIEGO SUPERCOMPUTER CENTER

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

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Knowledge Management for Scientific Application Using Polystore



Subhasis Dasgupta, Ph.D.

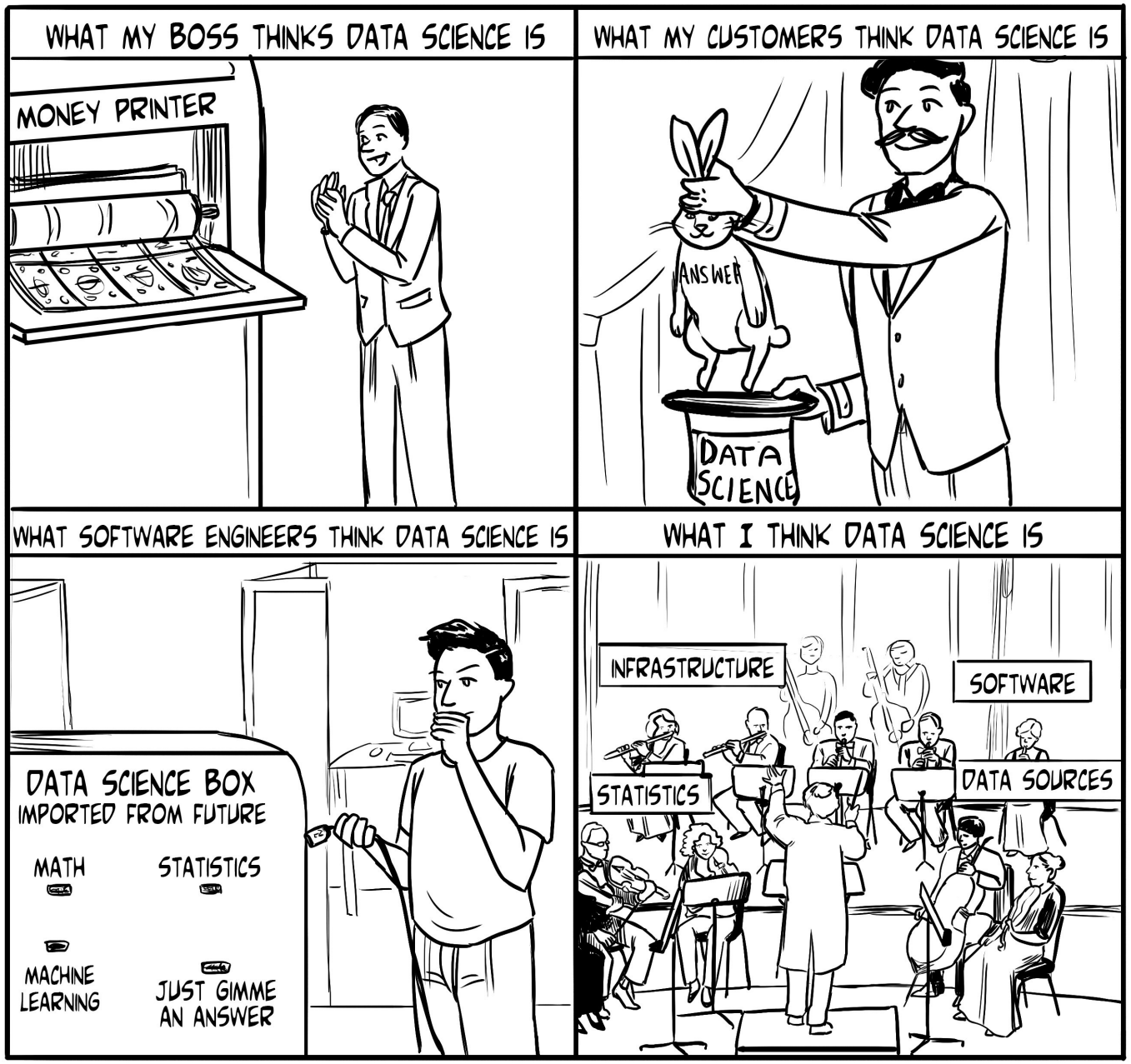
Assistant Scientist, University of California San Diego

Spatial Data Science for COVID-19 analysis and modeling.

Johnny Lei
Computational and Data Researcher,
UCSD

What is Data Science?

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



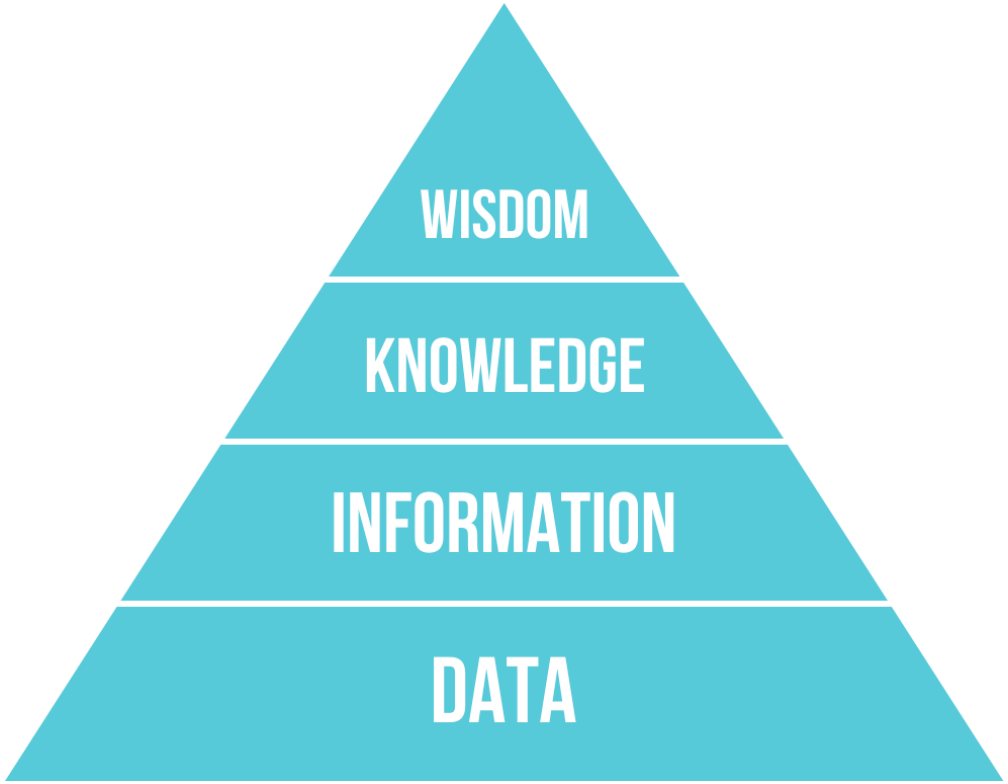
So, what exactly is data science?

Image Source:

Think Like a Data Scientist,
a book by Brian Godsey.
https://www.manning.com/books/think-like-a-data-scientist

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

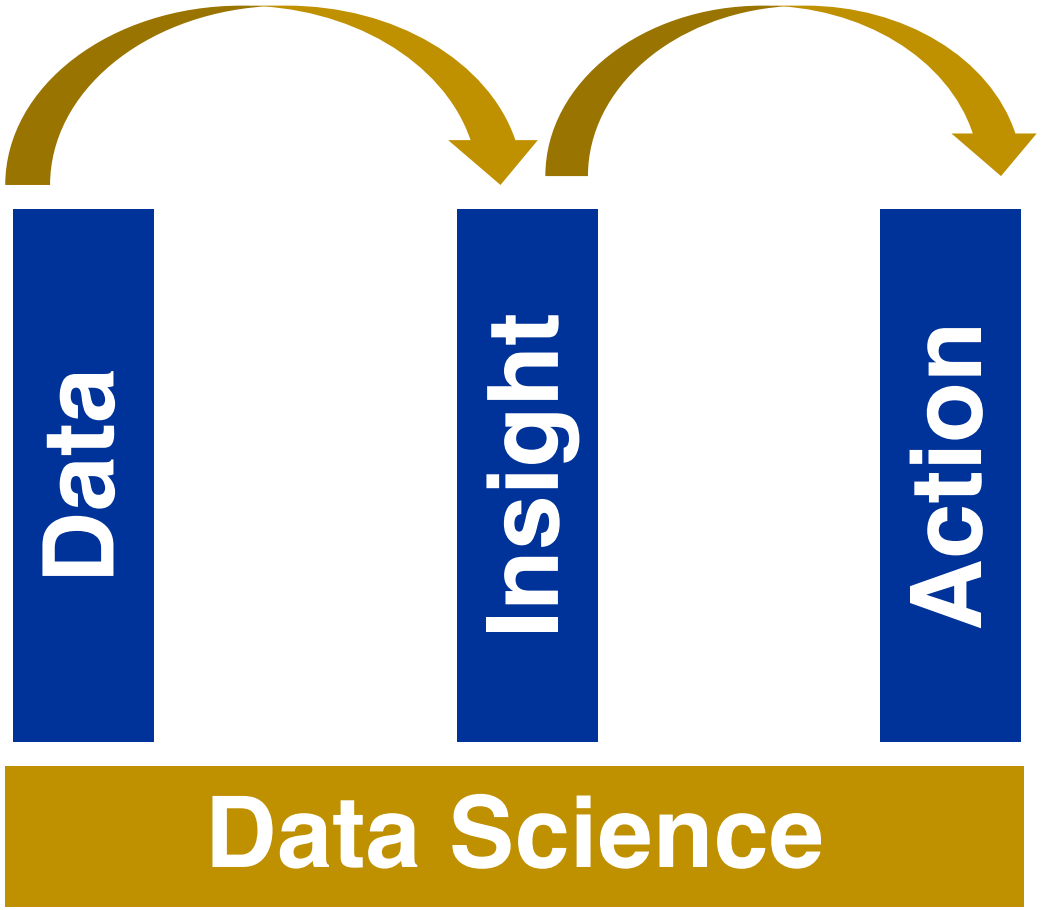
What is data?



data: */ˈdadə,ˈdādə/*
facts and statistics collected together for reference or analysis.

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

The Ultimate Goal of Data Science

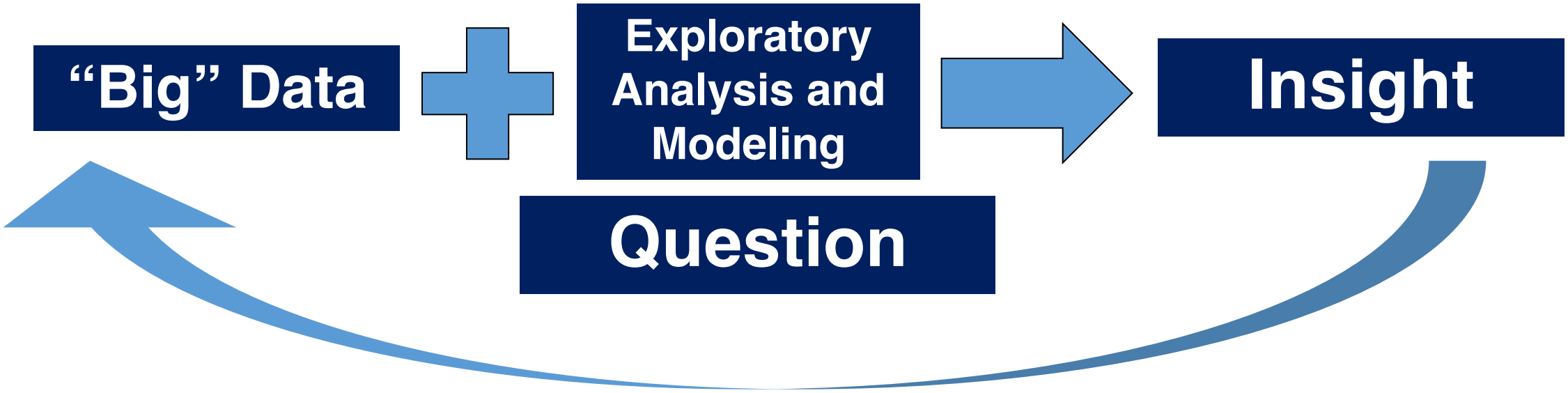


Shweta Purawat (shpurawat@ucsd.edu)
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Where is the science in data science?

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

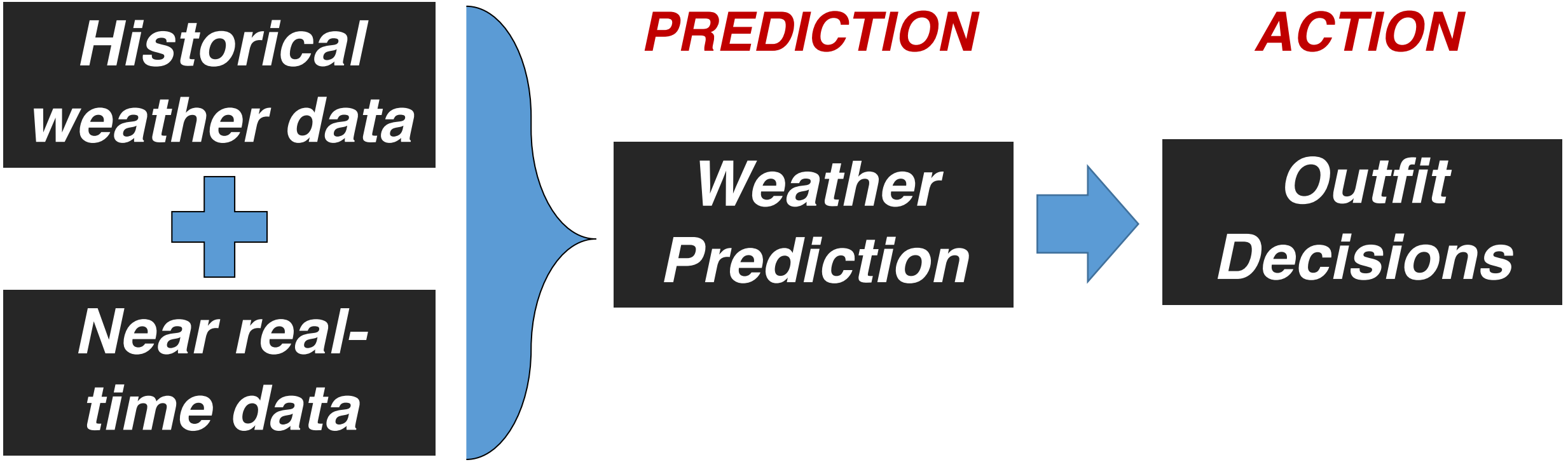
Generating Insights from Data



Insight = Data Product

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

Example: Weather Forecast to Actions



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

Systems and models

that help us to

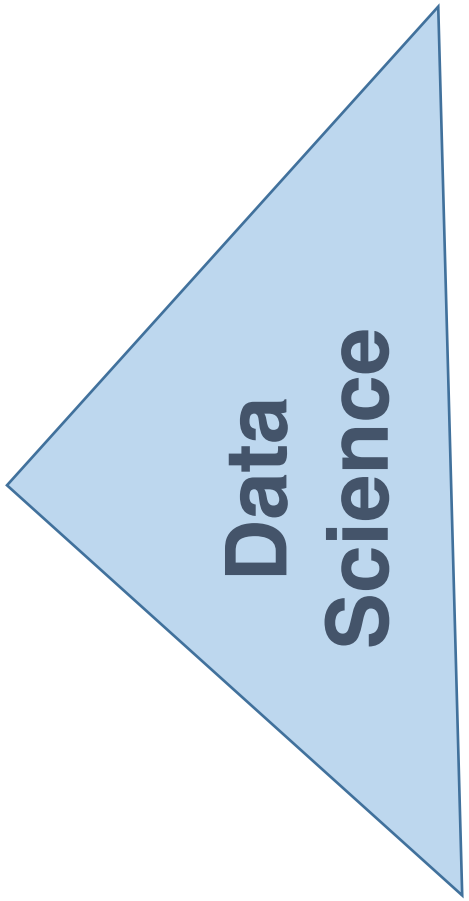
understand data

in order to

gain insights *and*

make predictions

leading to action for impact.



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

Data Science is “IMPACT” Science!

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

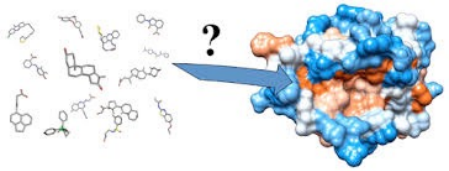
Data Science in Everyday Life

- Finance
- Healthcare
- Entertainment
- Travel
- Shopping
- …

Personalized Precision Medicine

Smart Manufacturing

Computer-Aided Drug Discovery



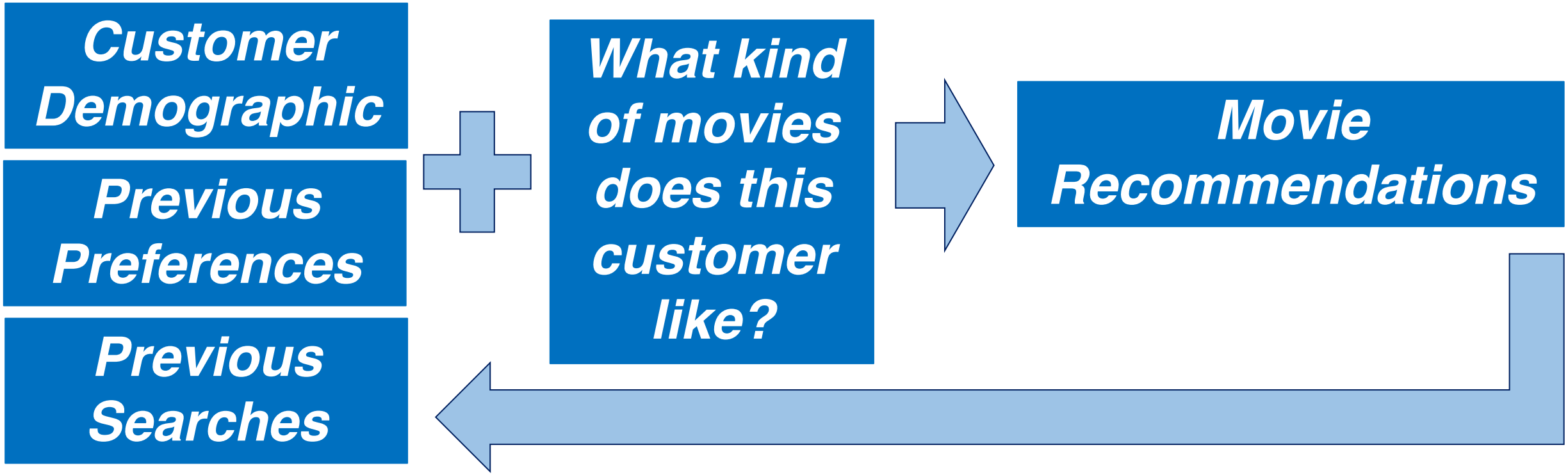
Smart Cities

Disaster Resilience and Response

Smart Grid and Energy Management

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

Entertainment Industry: Movie Recommendations

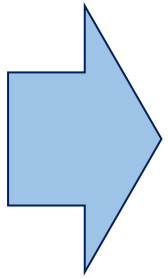


Shweta Purawat (shpurawat@ucsd.edu)
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Entertainment Industry: Movie Recommendations

PREDICTION

What kind of movies do customers like?



ACTION

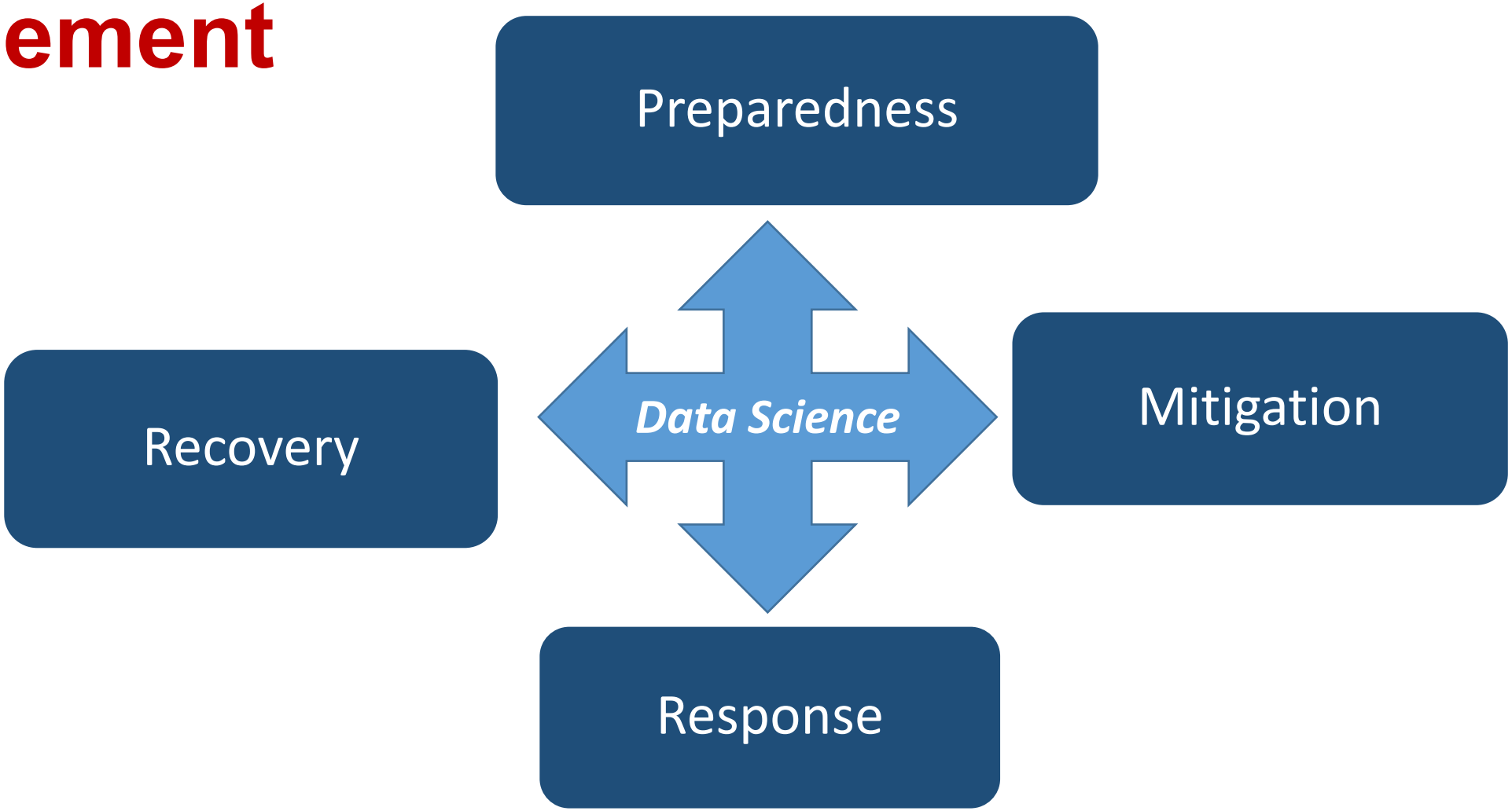
What movies to produce and market?

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

Example Industry: Smart Healthcare

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

Safety: Data Science for Emergency Management



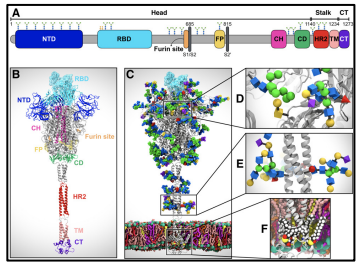
SDSC SAN DIEGO SUPERCOMPUTER CENTER

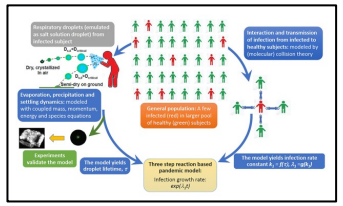
Shweta Purawat (shpurawat@ucsd.edu)
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

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
- …

SDSC SAN DIEGO SUPERCOMPUTER CENTER

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

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Why are we talking about data science and AI now?

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

Why is data science so popular?

The Computing Continuum

IoT/Edge → Fog → HPC/Cloud/Instrument

Size	Nano	Micro	Milli	Server	Fog	Campus	Facility
Example	Adafruit Trinket	Particle.io Boron	Array of Things	Linux Box	Co-located Blades	1000-node cluster	Datacenter
Memory	0.5K	256K	8GB	32GB	256G	32TB	16PB
Network	BLE	WiFi/LTE	WiFi/LTE	1 GigE	10GigE	40GigE	N*100GigE
Cost	$5	$30	$600	$3K	$50K	$2M	$1000M

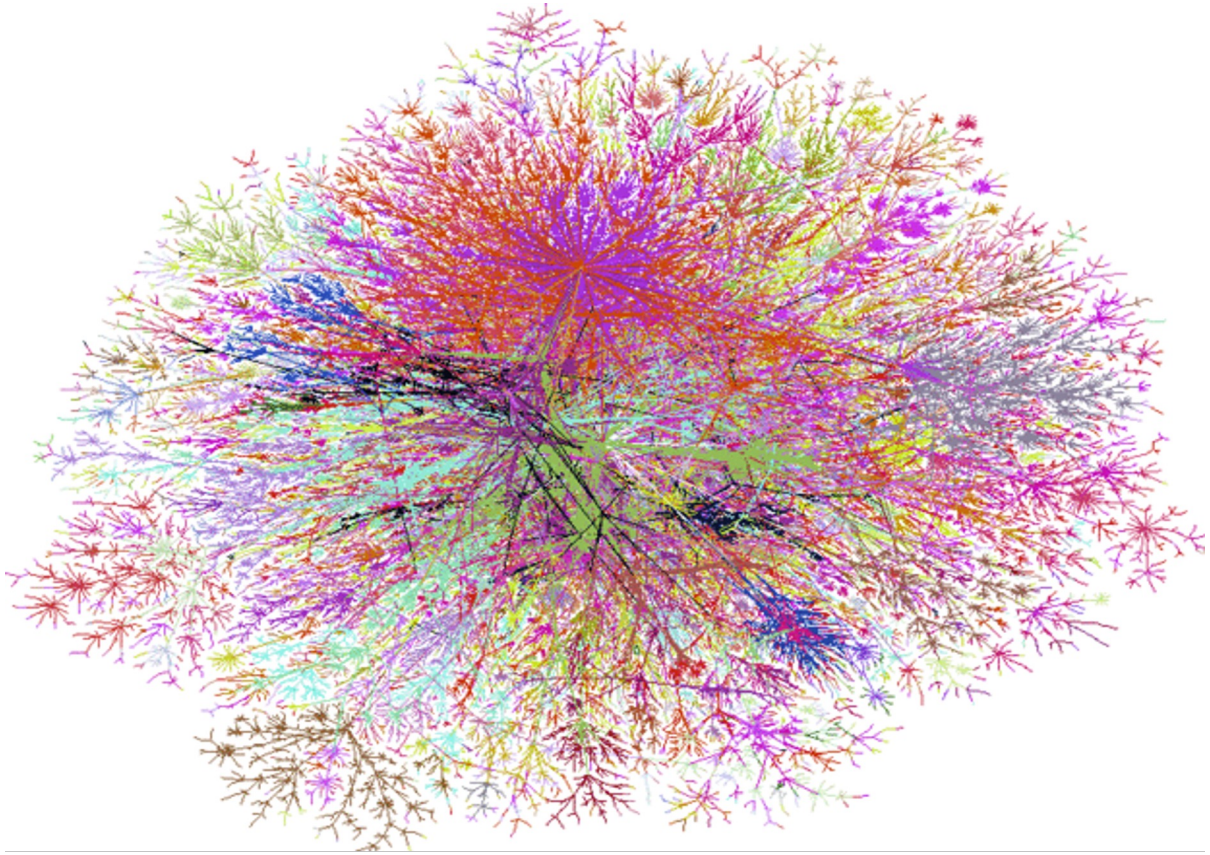
Count = 10^9 Size = 10^1 … Count = 10^1 Size = 10^9

SDSC SAN DIEGO SUPERCOMPUTER CENTER

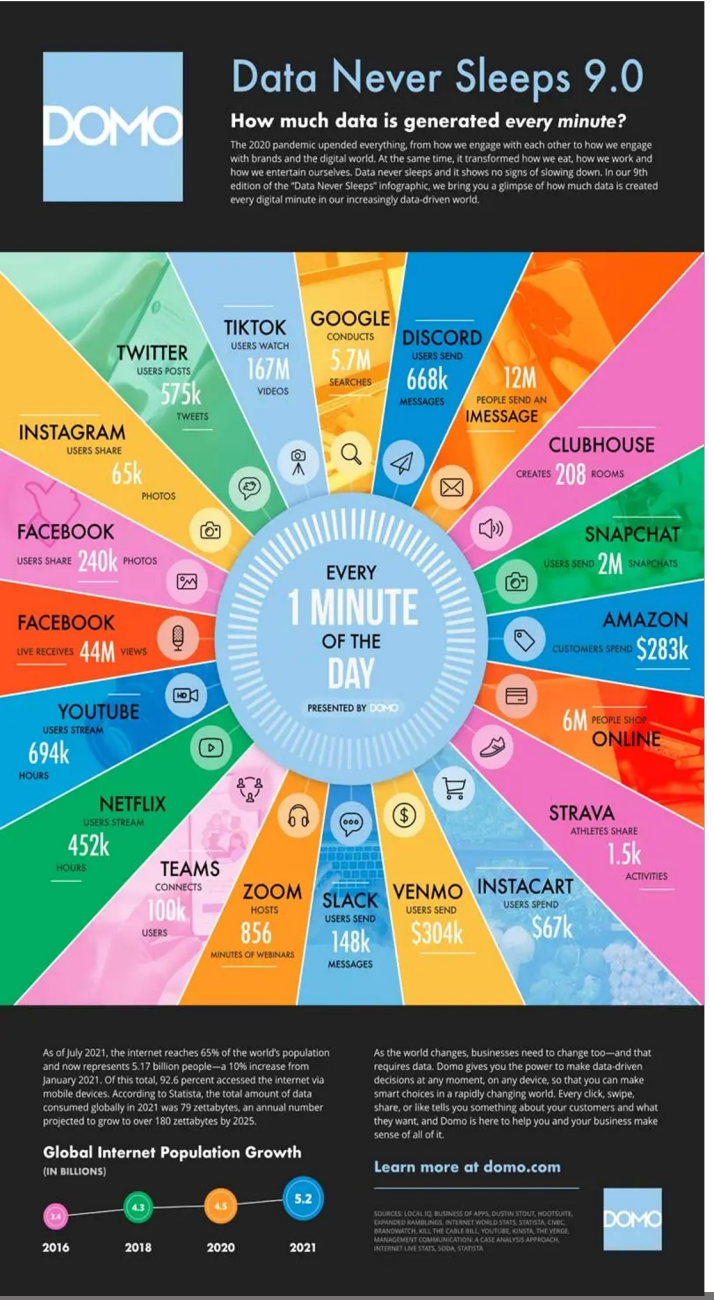
Shweta Purawat (shpurawat@ucsd.edu)
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

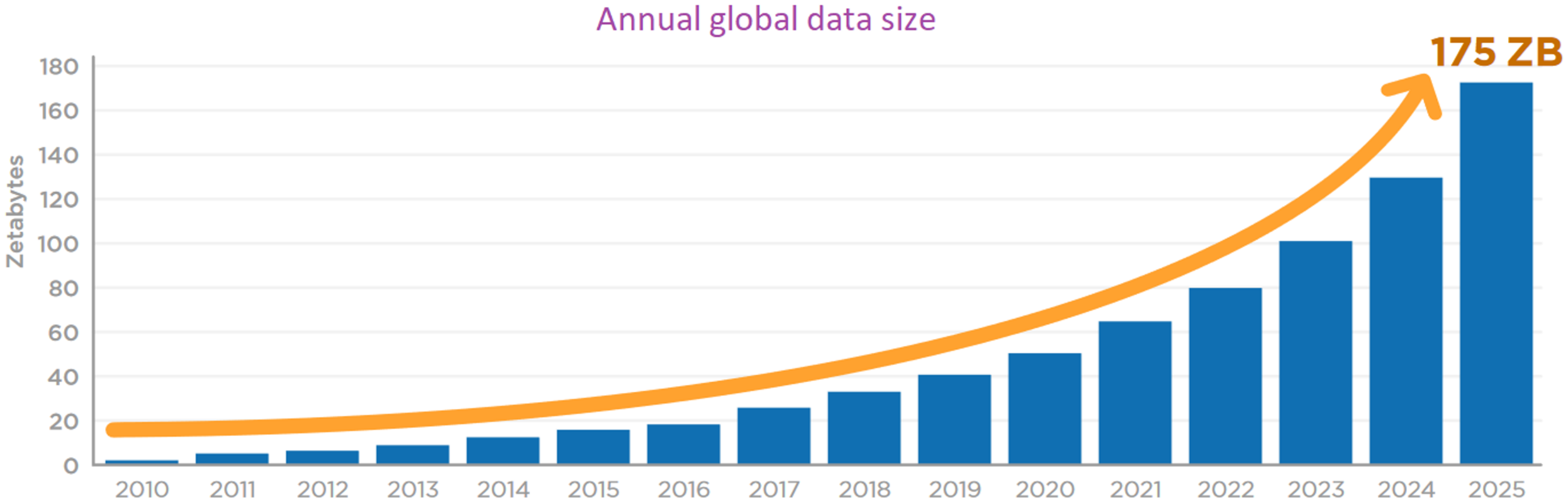
What is Big Data?



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



SDSC SAN DIEGO SUPERCOMPUTER CENTER
Shweta Purawat (shpurawat@ucsd.edu)
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)
UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

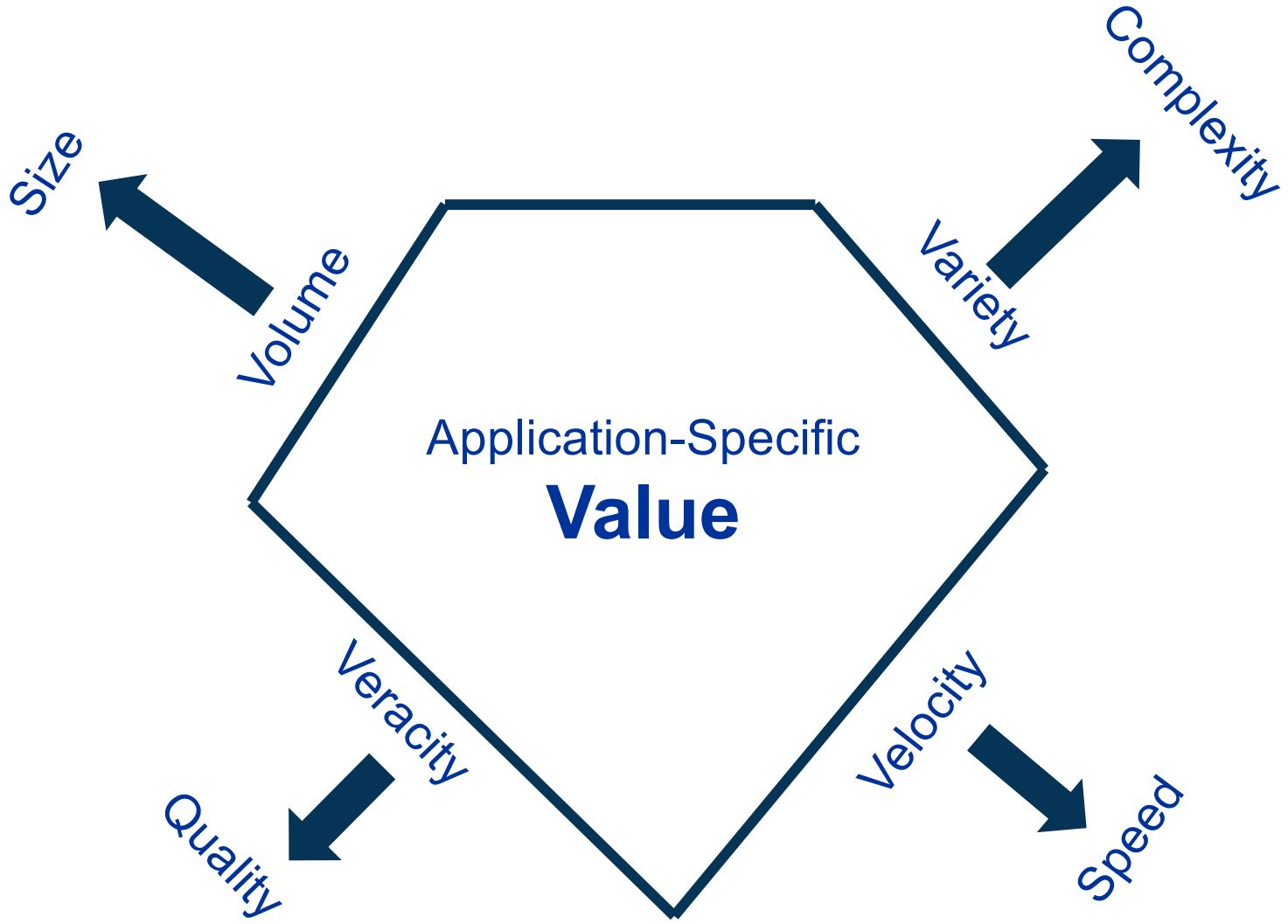


https://www.datanami.com/2022/01/11/big-growth-forecasted-for-big-data/

175 ZB = 175 Trillion GB!

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

V's of Big Data



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

Data Science:

Finding connections in big data…

… and applications using the connections for a purpose.

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

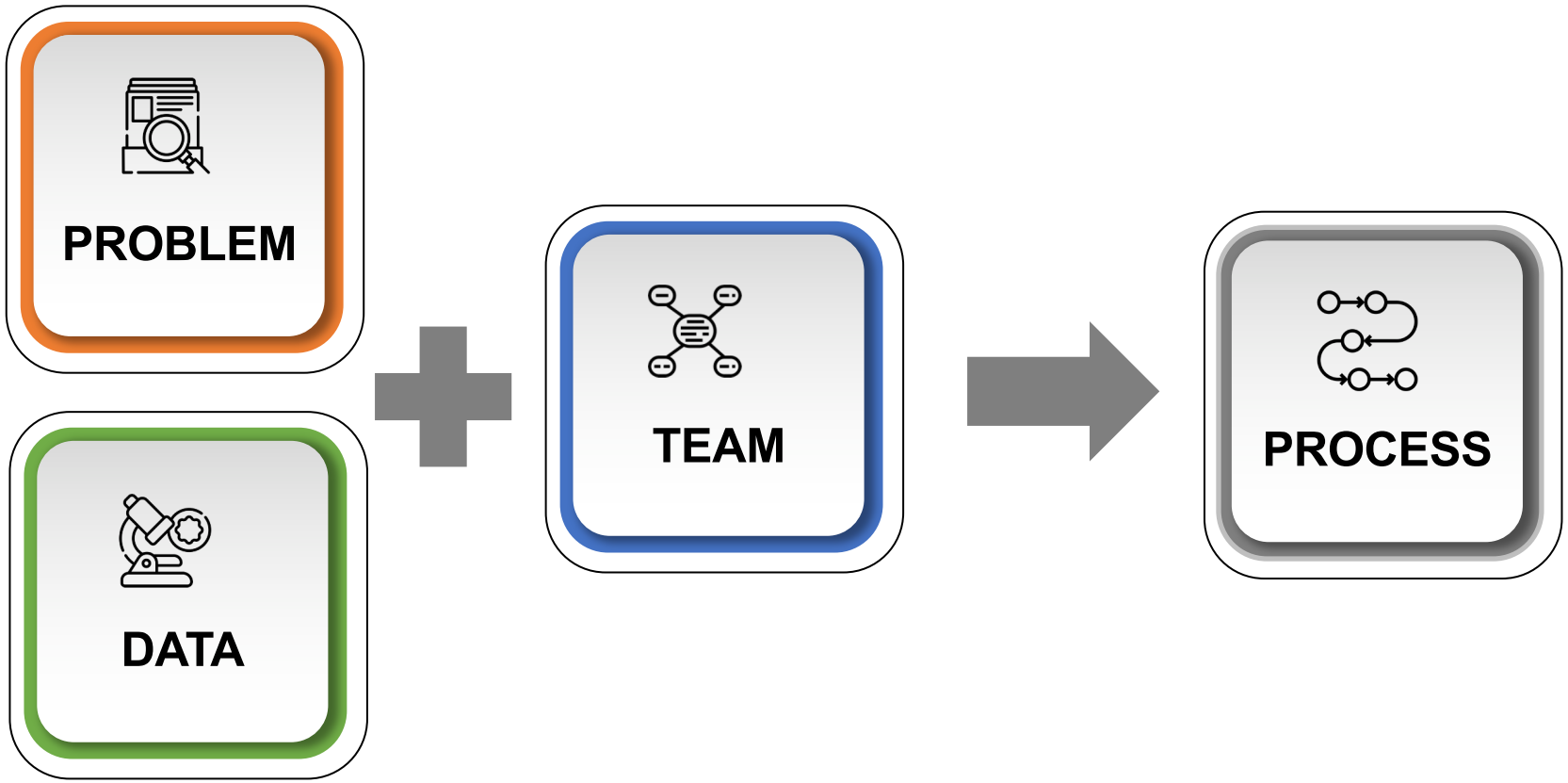
How does data science happen?

or

How to manage data science projects?

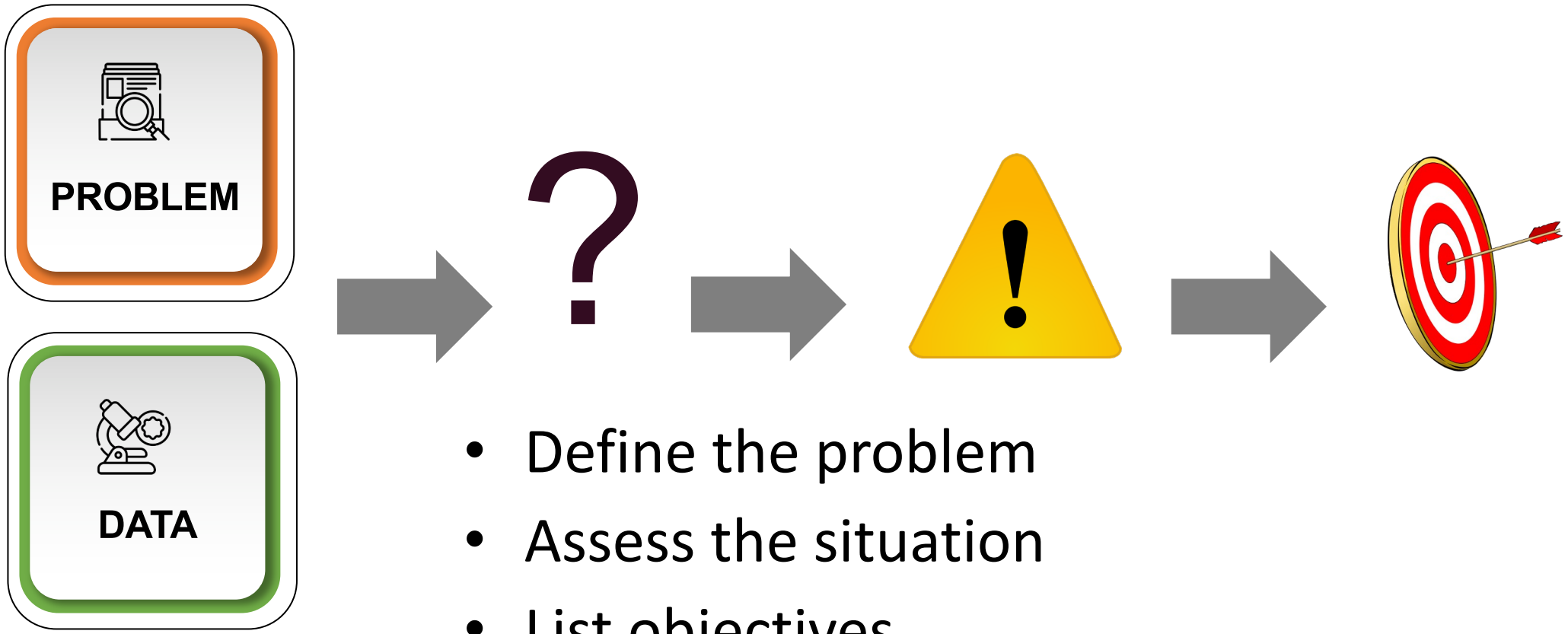
Shweta Purawat (shpurawat@ucsd.edu)
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Dimensions of Data Science



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

Asking the Right Question



- Define the problem
- Assess the situation
- List objectives

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

Data Science Process

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

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

Phases of the Data Science Process

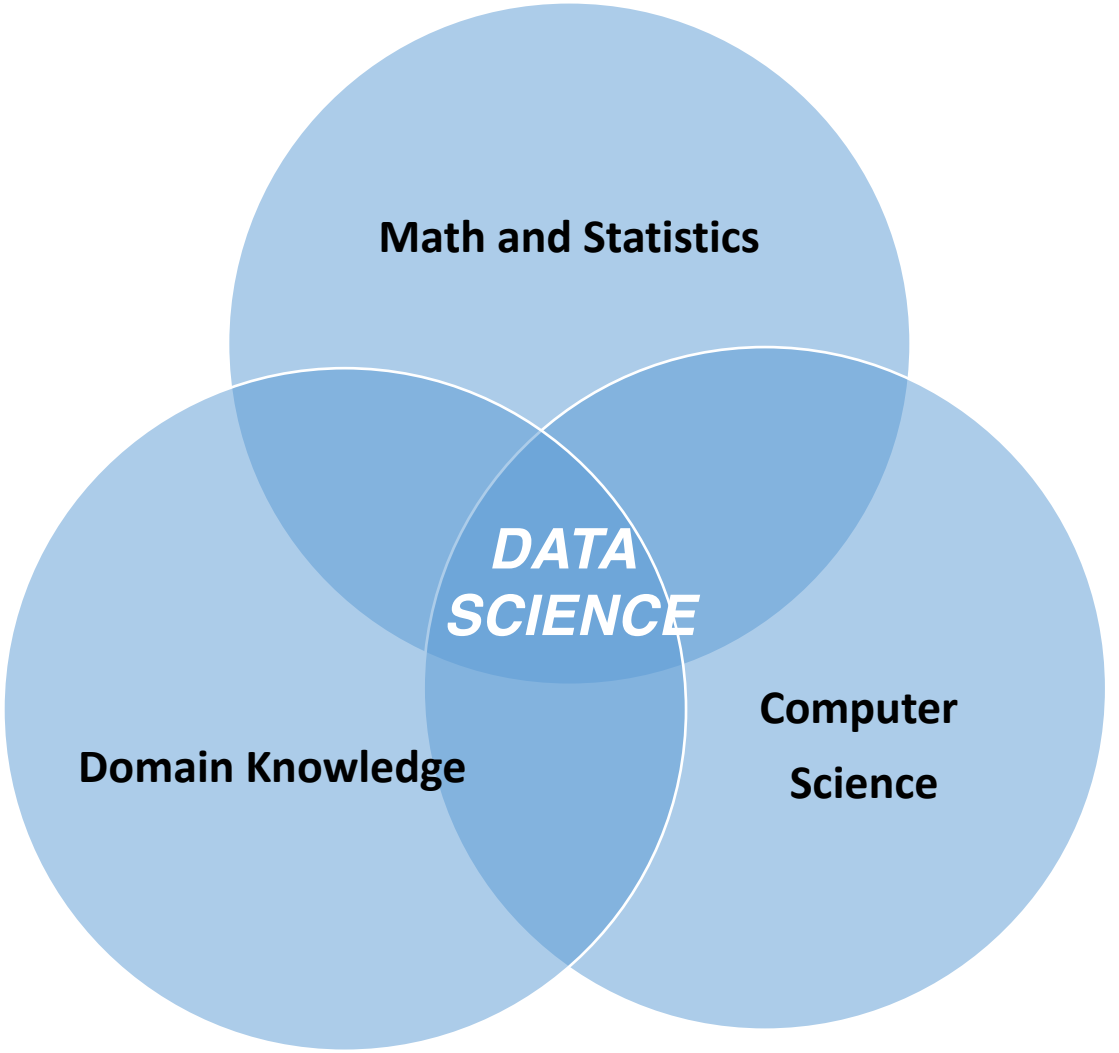
- **BUILD**
- **EXPLORE**
- **SCALE**

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

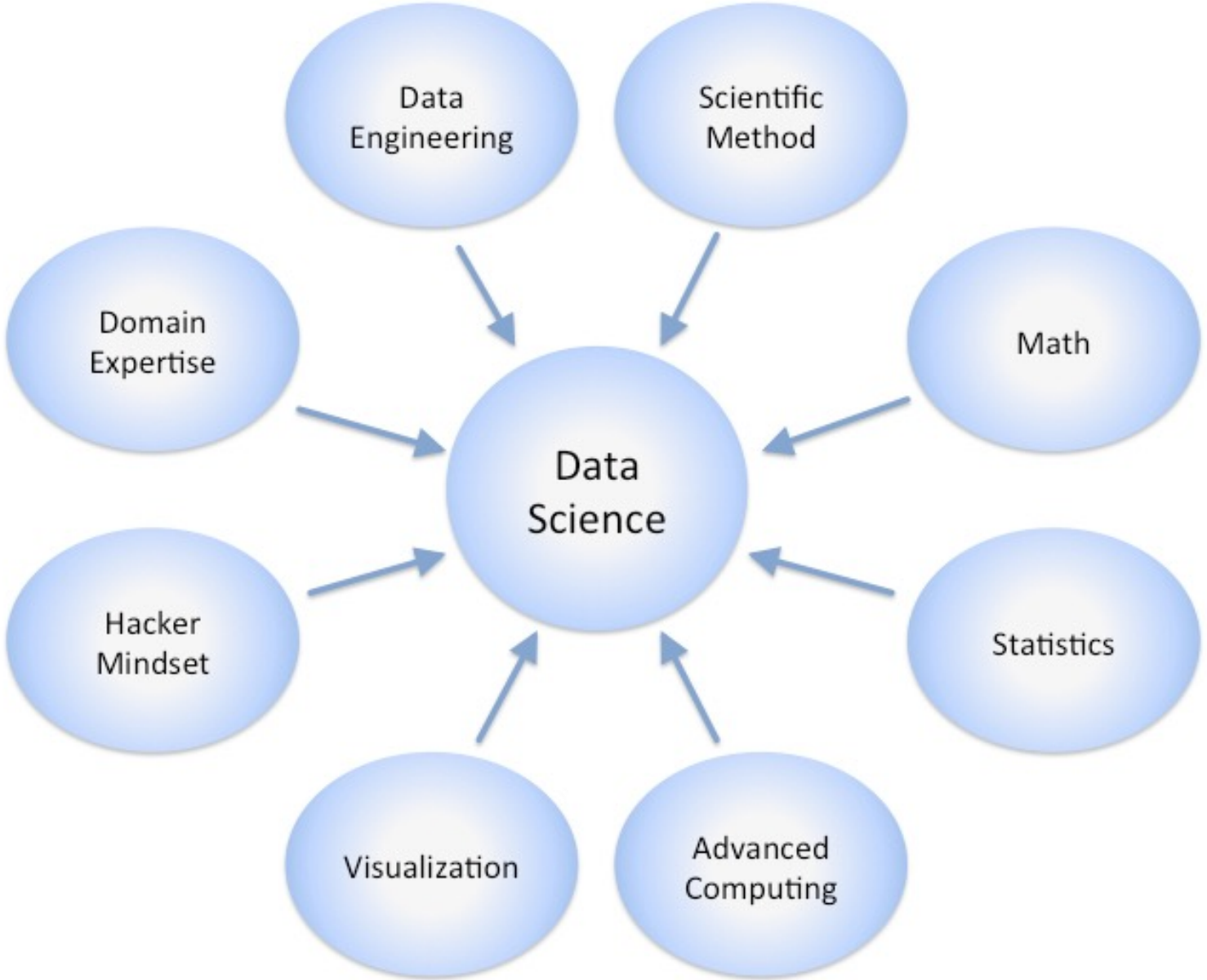
What is a Data Scientist?

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

Data science is multidisciplinary.



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



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

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

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

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

SDSC SAN DIEGO SUPERCOMPUTER CENTER

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

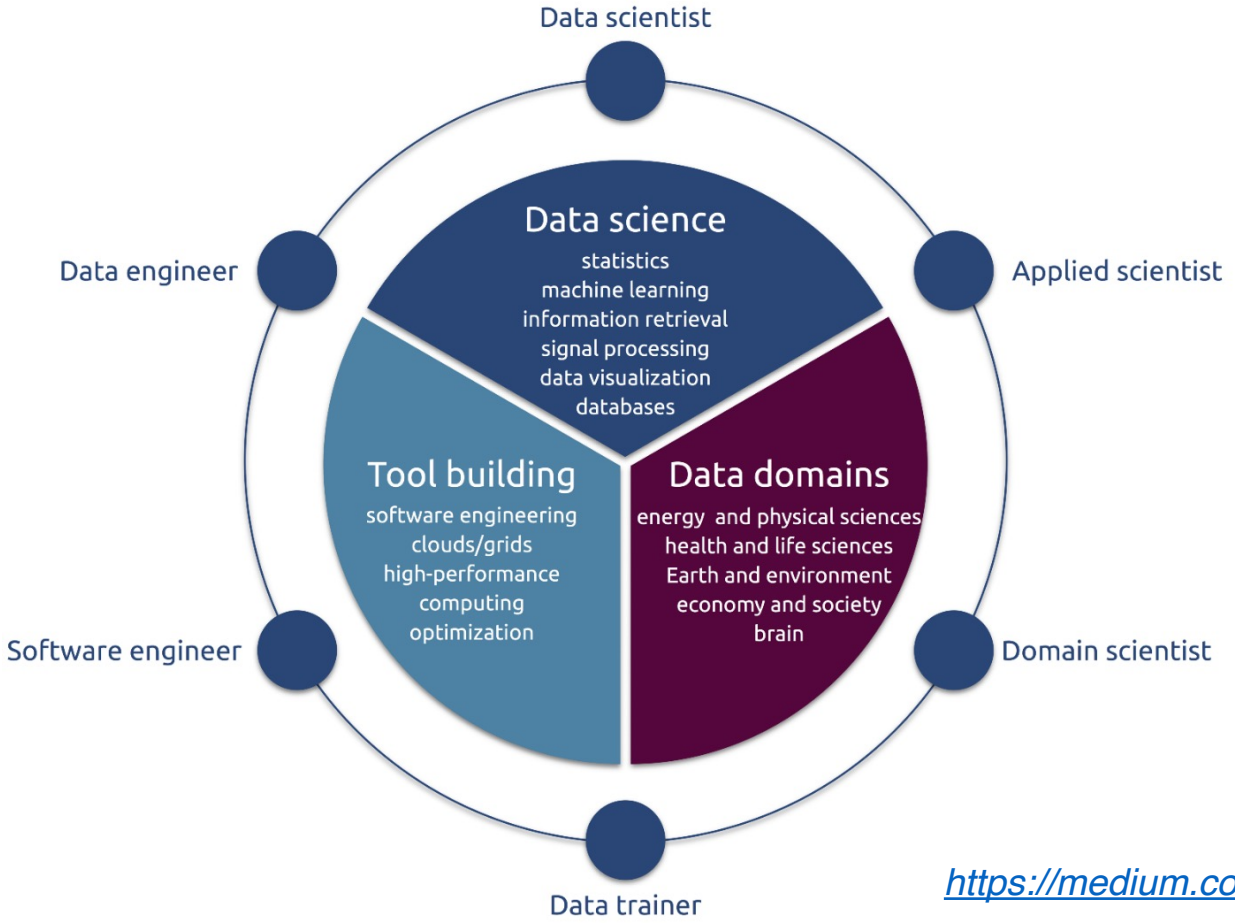
UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Data science is “WE” science!

- Team collaboration
- Curiosity
- Respect

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

Expertise and Skills often overlap between different data science roles.



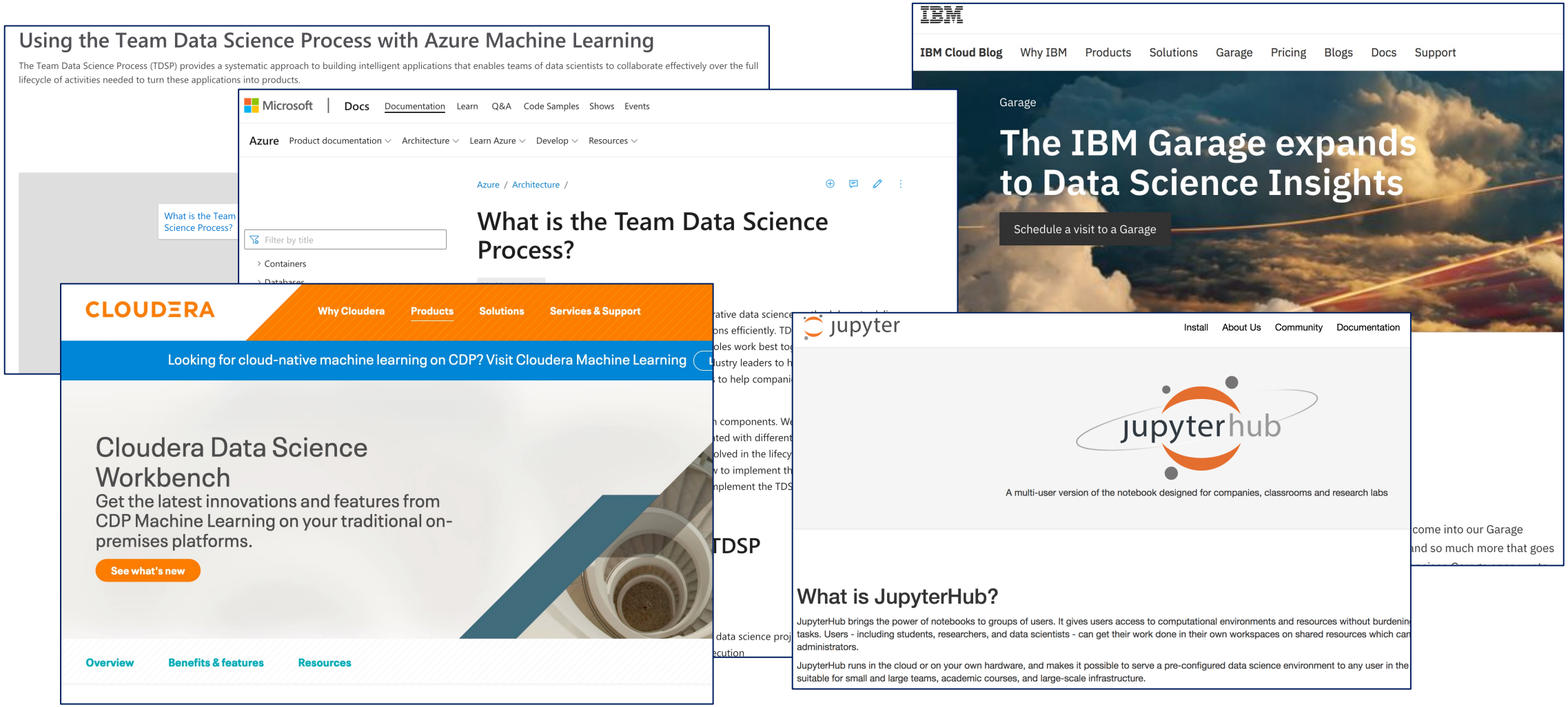
The data science ecosystem: activities and actors

- 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

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

Team Data Science Environments



SDSC SAN DIEGO SUPERCOMPUTER CENTER

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

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

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

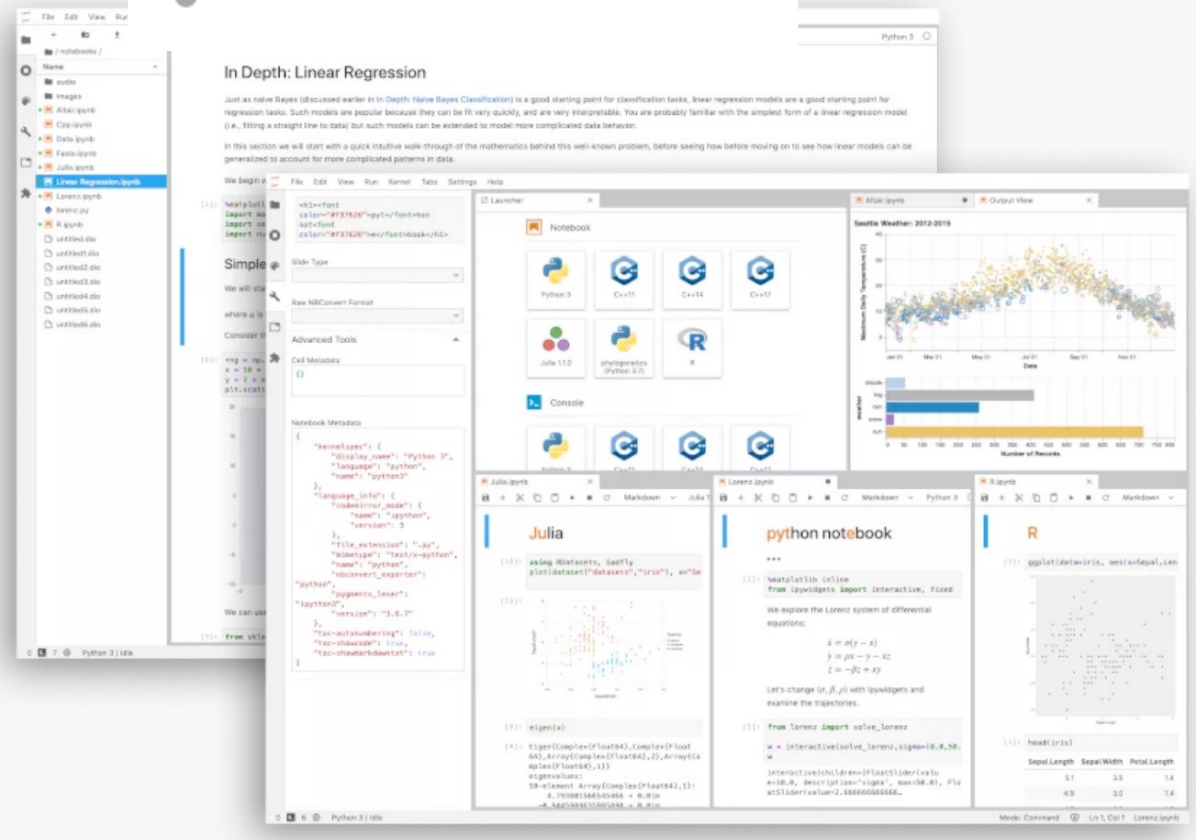
Source: PYPL

Top Data Science Programming Languages

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

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

Jupyter Notebooks and JupyterLab



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

SDSC SAN DIEGO SUPERCOMPUTER CENTER

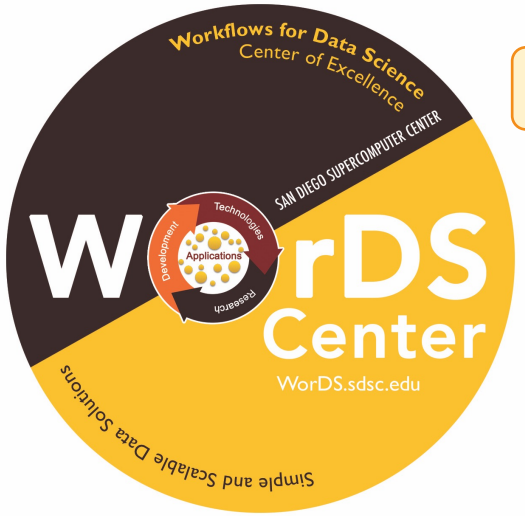
Shweta Purawat (shpurawat@ucsd.edu)
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Contact:

Shweta Purawat *Email:shpurawat@ucsd.edu*

Ilkay Altintas, Ph.D. *Email: ialtintas@ucsd.edu*



https://words.sdsc.edu/

https://wifire.ucsd.edu/

WIFIRE

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

Questions?

NIH National Institutes of Health

NSF

U.S. DEPARTMENT OF ENERGY | 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.