Introduction to Data Science and Its Applications

Presentation for the SDSC Summer Institute
August 3, 2021 (Delivered remotely due to COVID-19.)

İlkay ALTINTAŞ, Ph.D.

Chief Data Science Officer & Division Director of Cyberinfrastructure Research, Education and Development, **San Diego Supercomputer Center**

Founding Fellow, **Halıcıoğlu Data Science Institute**

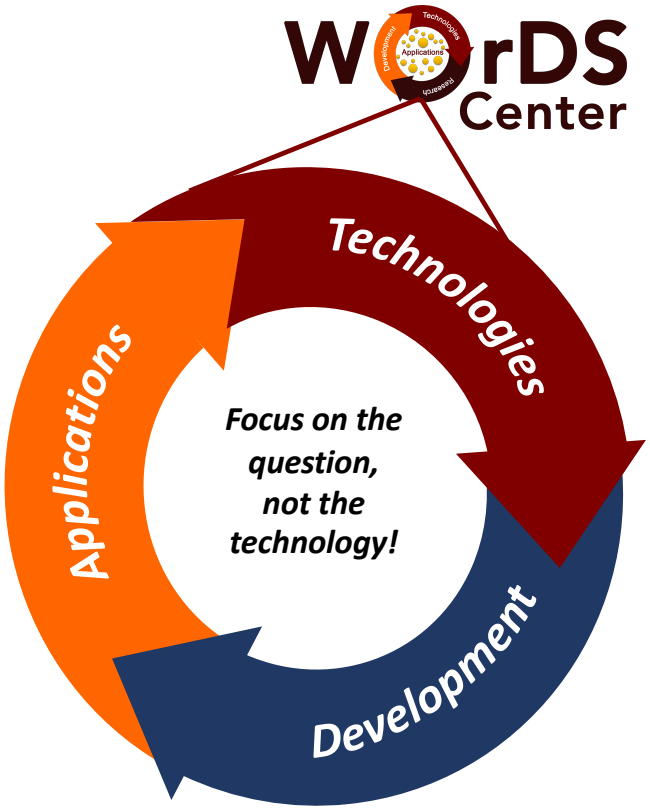
Founding Director**, Workflows for Data Science Center of Excellence**

Founding Director, **WIFIRE Lab**

Part 1:

Data-Driven Problem Solving

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)



Workflows for Data Science Center of Excellence at SDSC

http://WorDS.sdsc.edu

Mission:
Methodology and tool development
to enable collaborative workflow-driven science
and create solution architectures
op top of big data and advanced computing platforms.

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Common Theme…

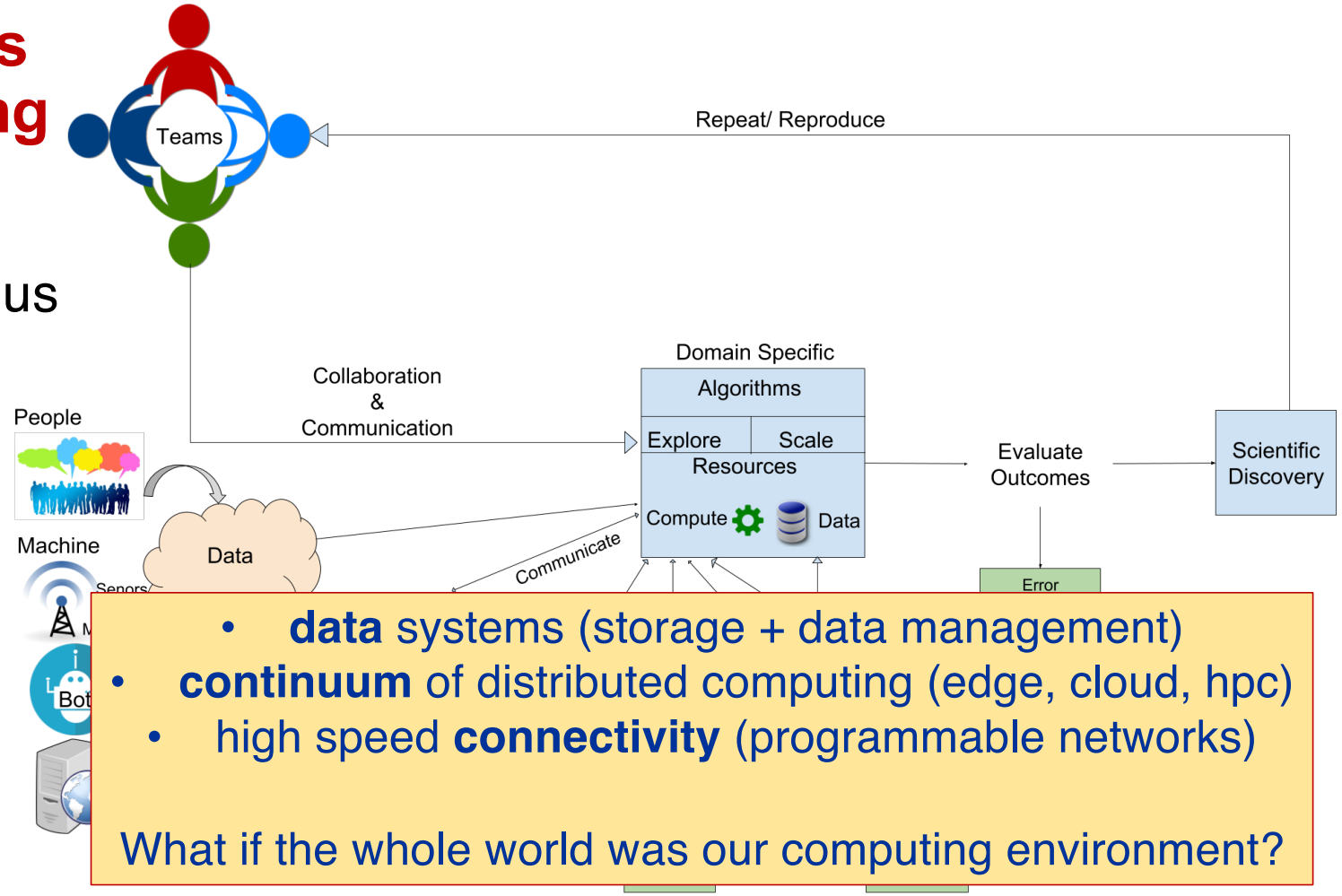
"Big" Data, Computational Science, Data Science, Cyberinfrastructure,

and Their Applications

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

The problems we are solving are …

- data-driven
- heterogeneous
- collaborative

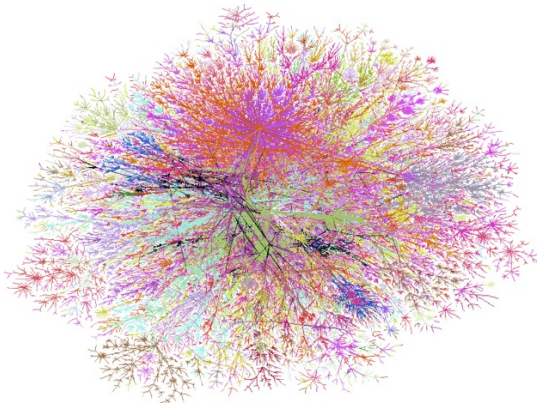


- **data** systems (storage + data management)
- **continuum** of distributed computing (edge, cloud, hpc)
- high speed **connectivity** (programmable networks)

What if the whole world was our computing environment?

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Big Data combined with Scalable Computing can be very valuable.



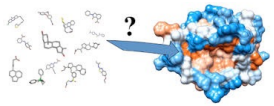
COMPUTING AT DIVERSE SCALES

"BIG" DATA

Enables dynamic data-driven applications

Smart Manufacturing

Computer-Aided Drug Discovery



Personalized Precision Medicine

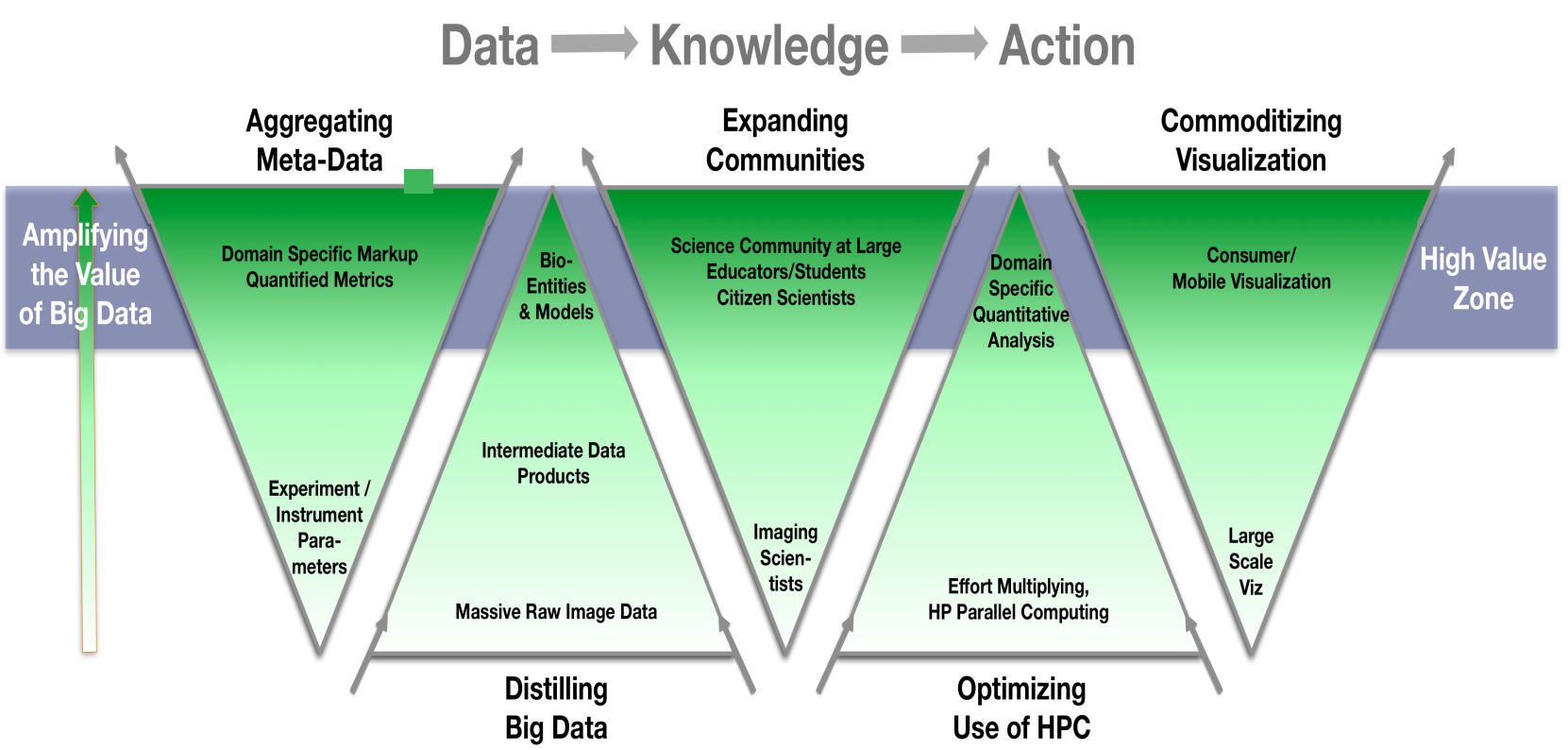
Smart Cities

Disaster Resilience and Response

Smart Grid and Energy Management

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

How do we amplify the value of Big Data?



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

How do we find the connections and impactful answers to our questions, i.e., turn it into benefit?

"We are drowning in information and starving for knowledge"

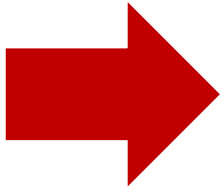
– John Naisbitt

Source: Megatrends, 1982

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Going from Data to Discovery to Impact

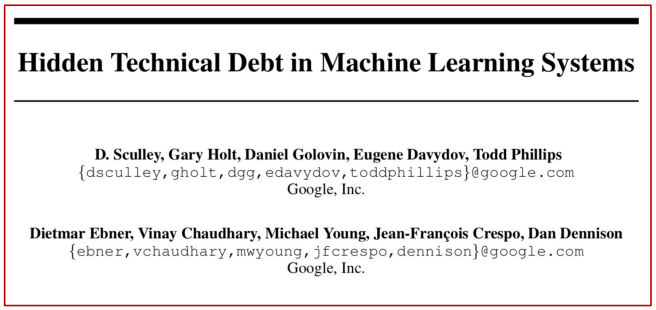
Amplifying the Value of Data Related to **X**



Benefit **Y** for Business, Science, Society, …

We need to focus on the problems to solve.

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)



Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

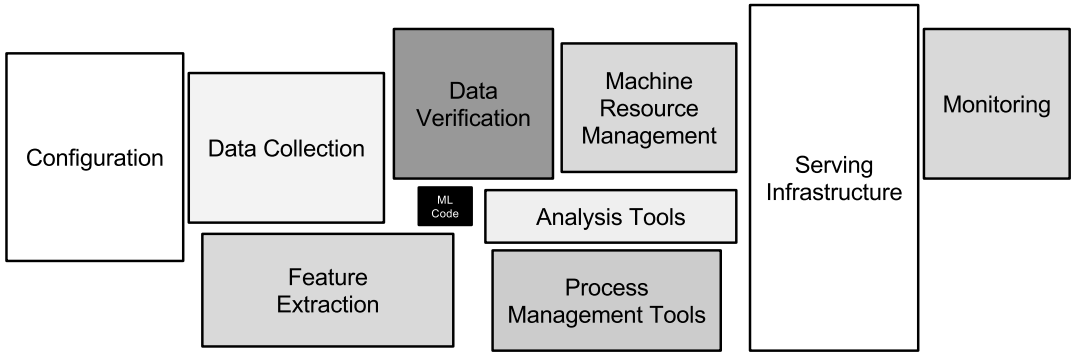


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

The problem is:

"solutions are complicated"!

- Heterogenous systems and infrastructure
- Data management
- Machine learning, statistics and analytical methods
- Scalable process management
- Dynamic coordination and resource optimization
- Skilled interdisciplinary team
- Collaborative culture and communication tools

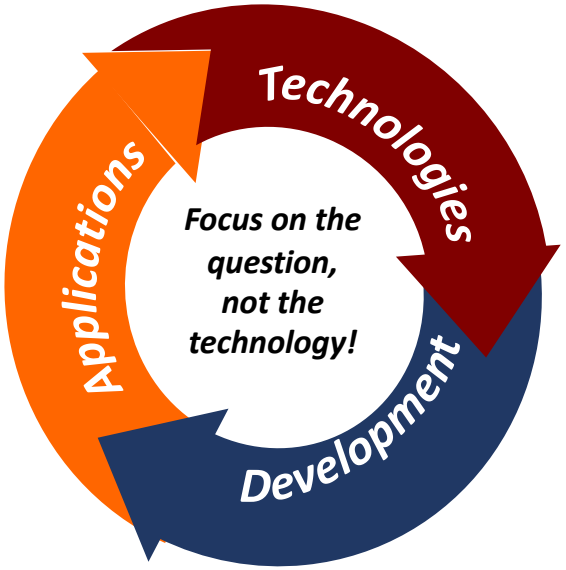
Research Goal:
Create a Data Science Solution Ecosystem that Enables Needs and Best Practices

- data-driven
- scalable
- dynamic
- process-driven
- heterogeneous
- accountable
- reproducible
- interactive
- multidisciplinary
- collaborative

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Team Data Science

How can I get smart people to collaborate and communicate?



…to utilize data and computing to generate insights and solve a question.

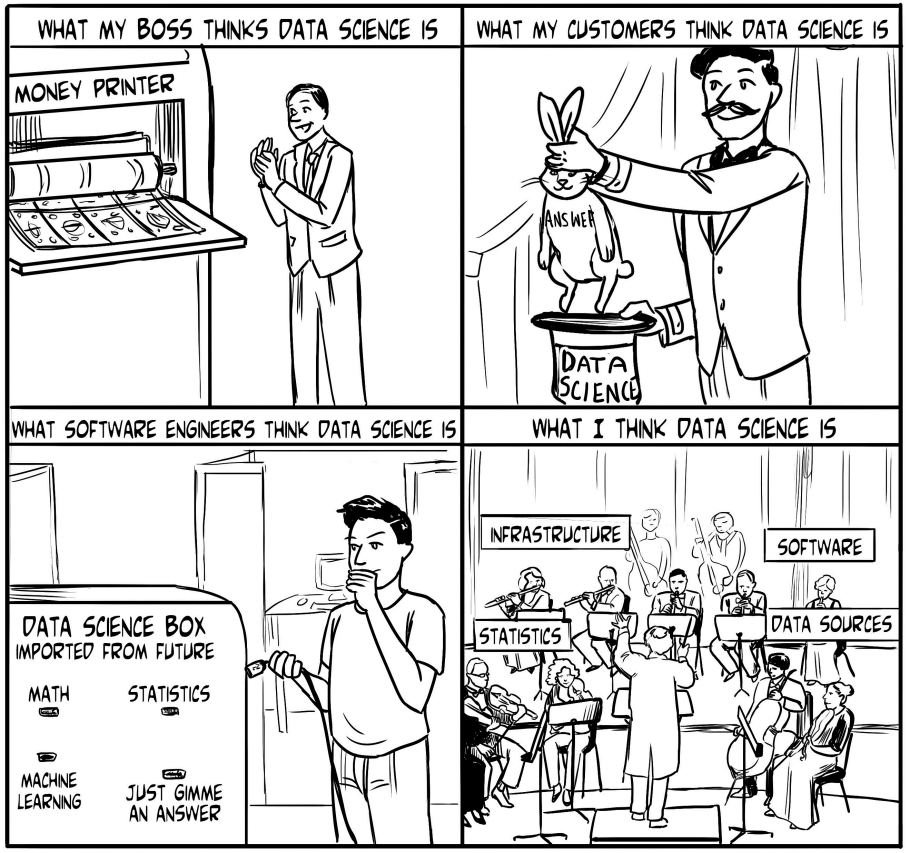
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Part 2:

What is Data Science?

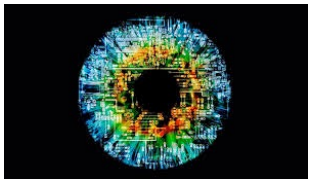
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

So what is data science?



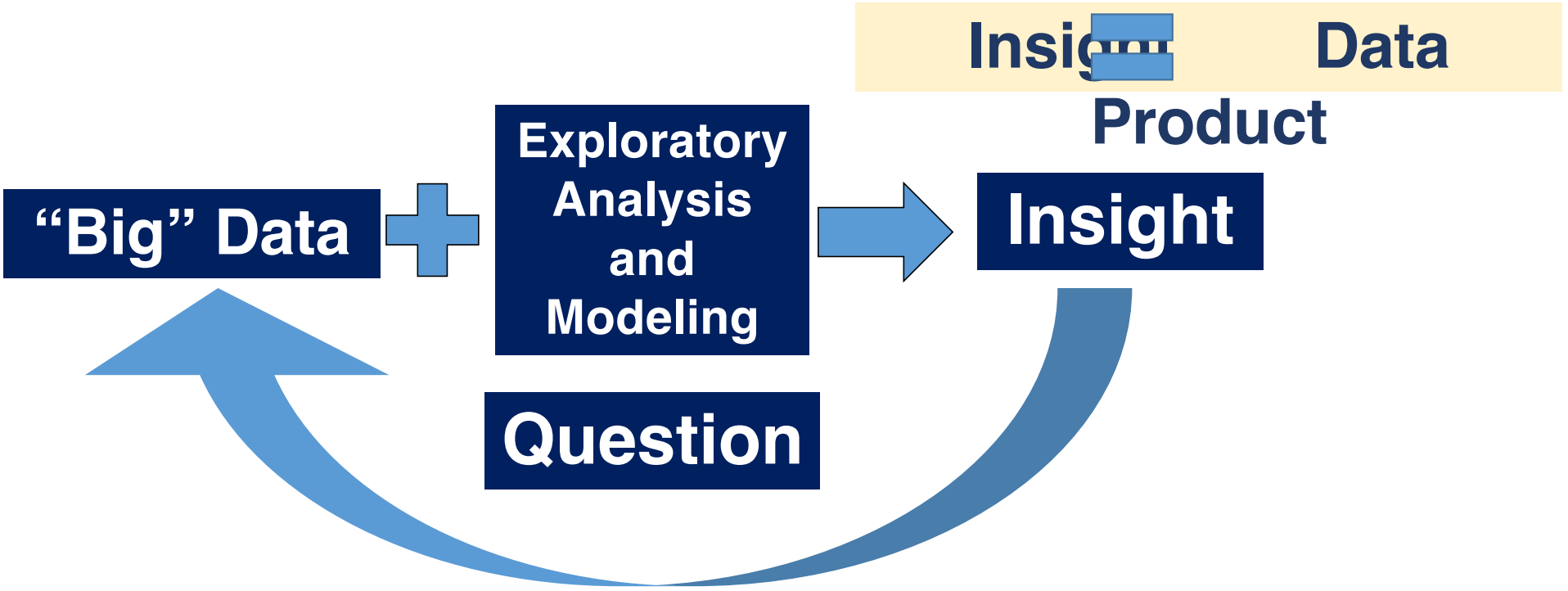
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Ultimate Goal of Data Science



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

How does successful data science happen

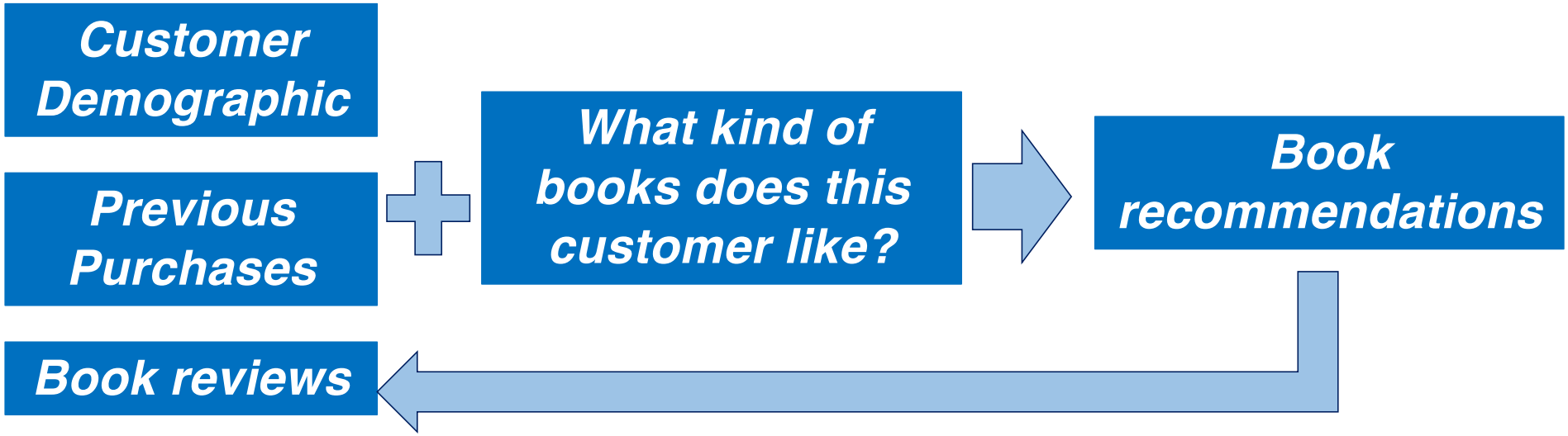


Data Products

"… a product that facilitates an end goal through the use of data"

-- DJ Patil, Former U.S. Chief Data Scientist in Data Jujitsu

Book Recommendations



amazon.com

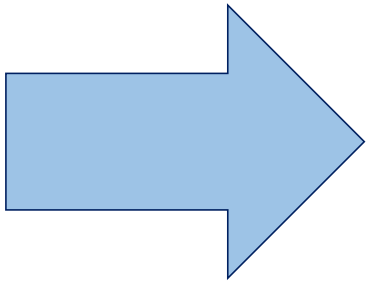
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Find Potential Audience for a

Model of customer's book

+

New book information

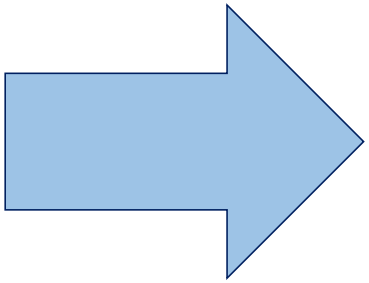


Who is likely to like this book?

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Market a New Book

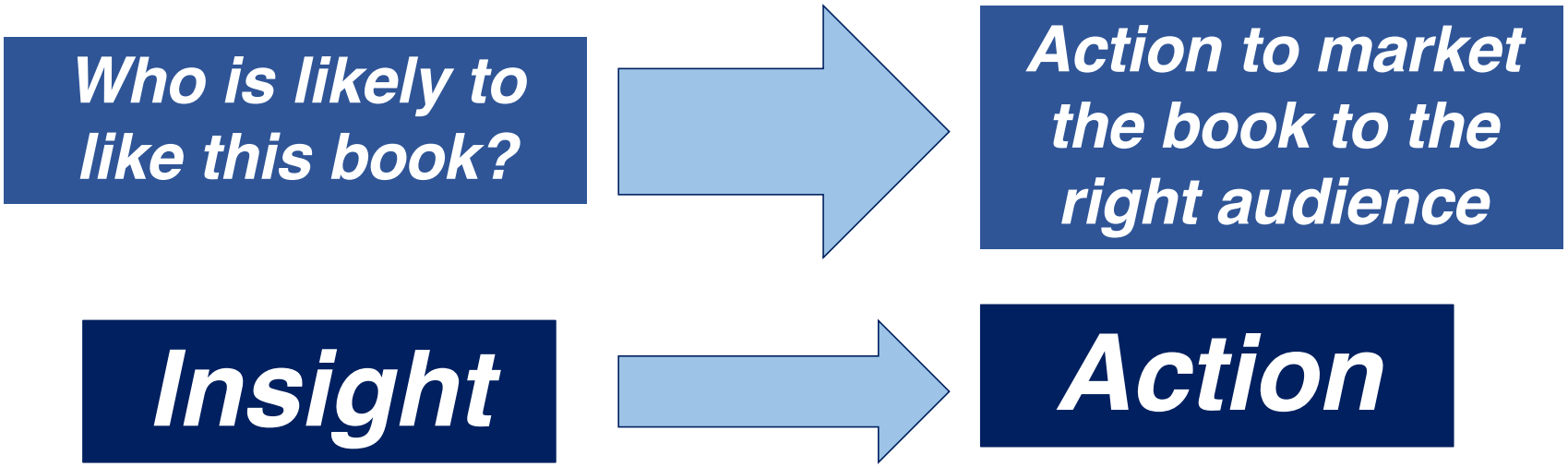
Who is likely to like this book?



Action to market the book to the right audience

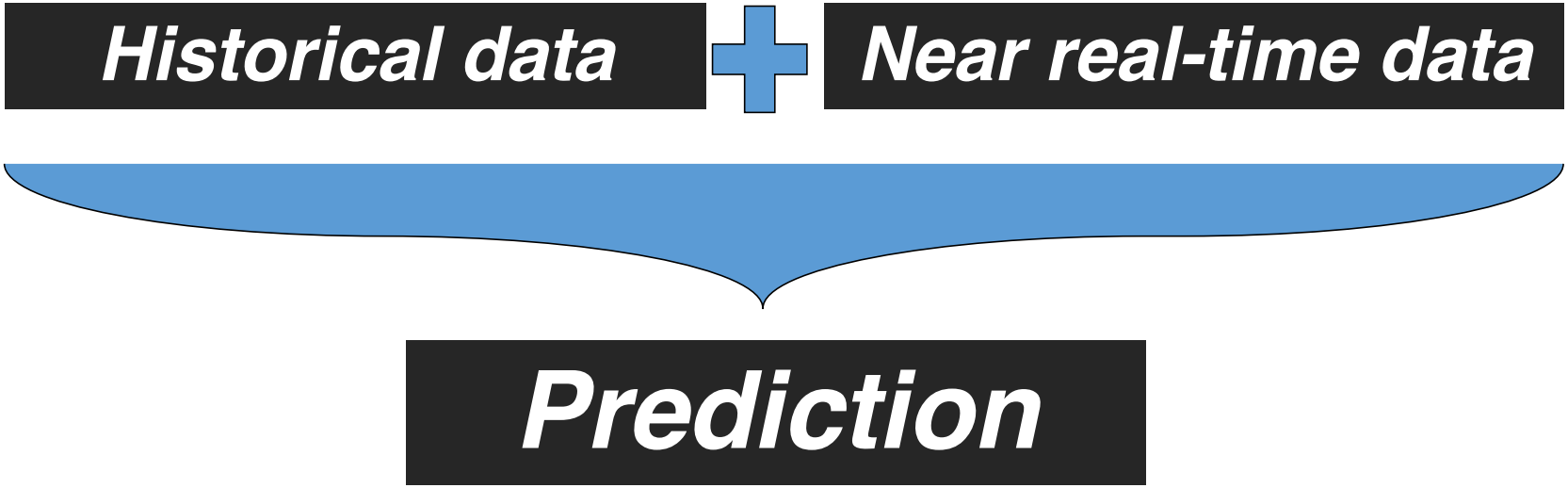
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Market a New Book



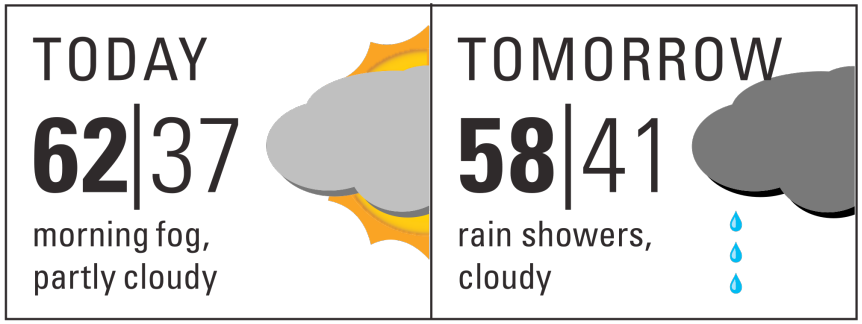
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Actionable Information



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Prediction



Action

SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

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

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Data Science is “IMPACT” Science!

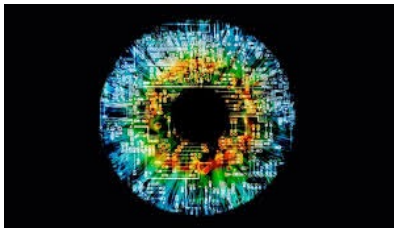
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Part 3:

Big Data Science Today

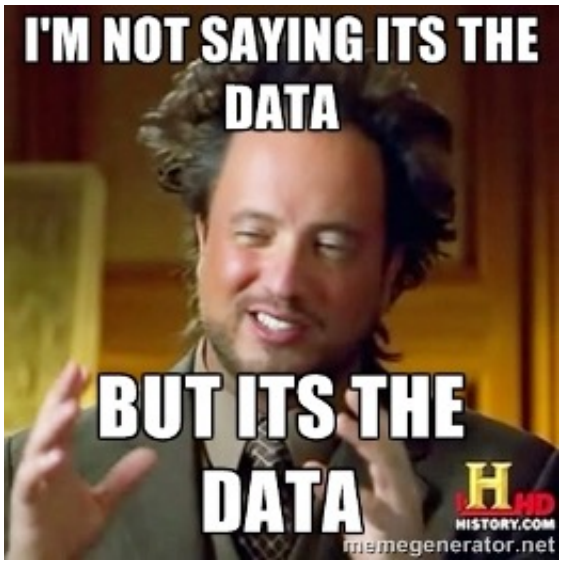
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

The Disruptors:



Why are we talking about Data Science now?

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)



+

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

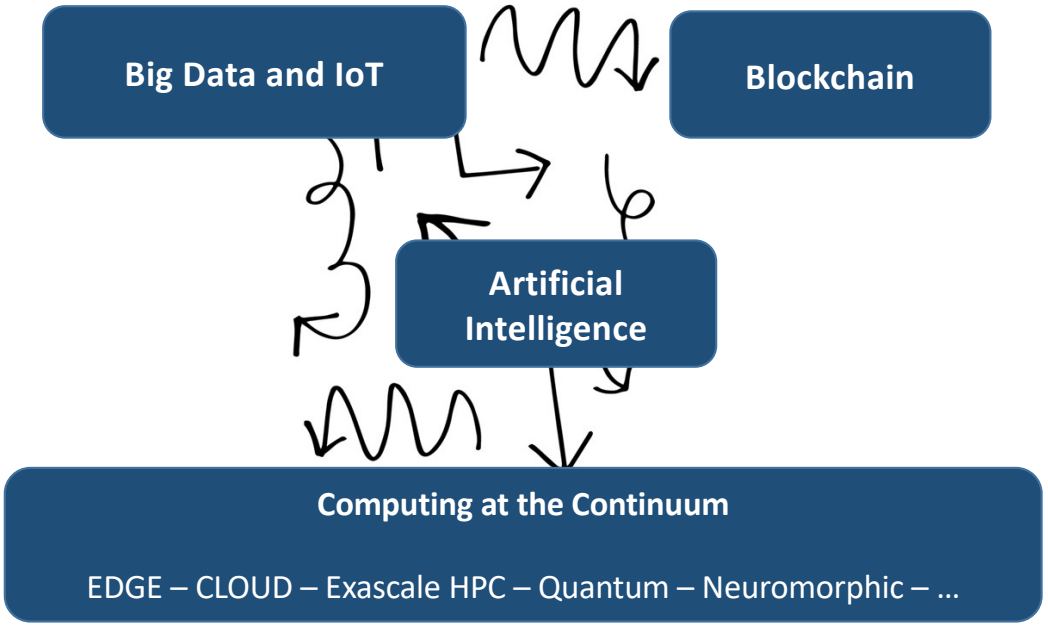
Count = 10^1 Size = 10^9

Big Data **Computing at the Continuum**

Changed the way we solve problems!

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

New kids on the block…



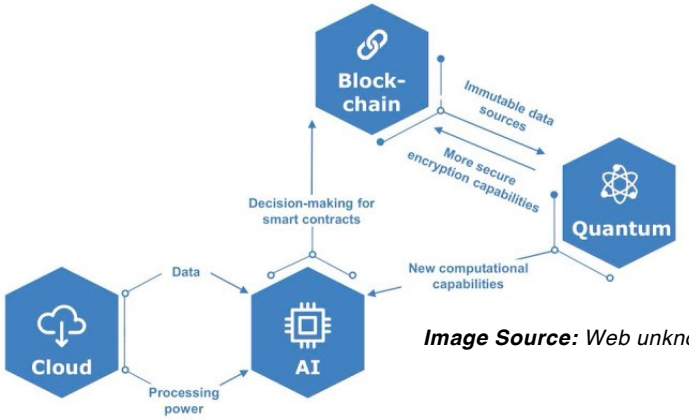
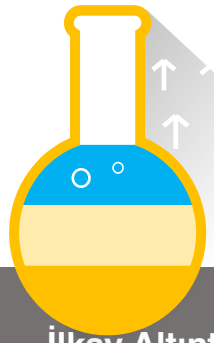


Image Source: *Web unknown*

How Much Data Is Big Data?



İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

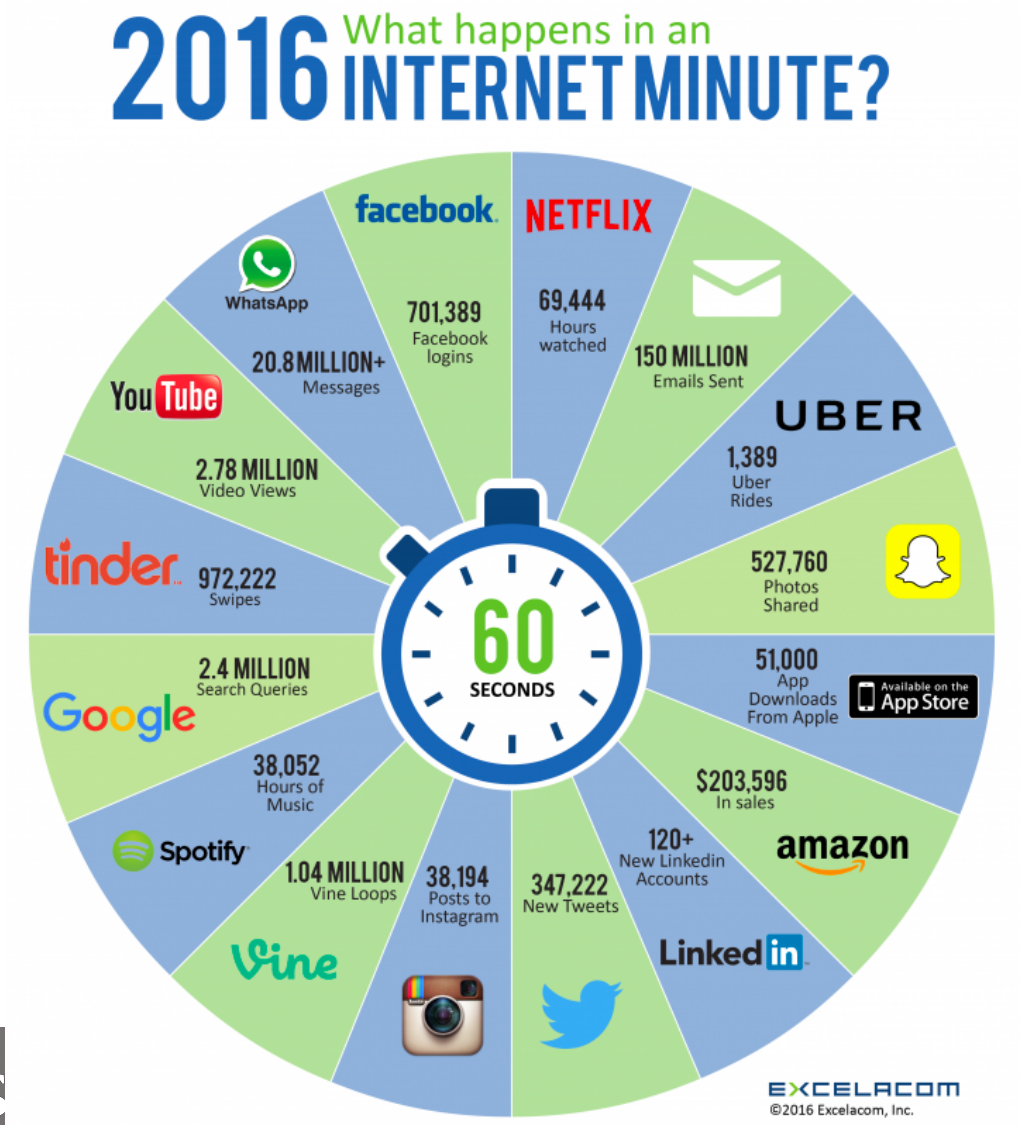
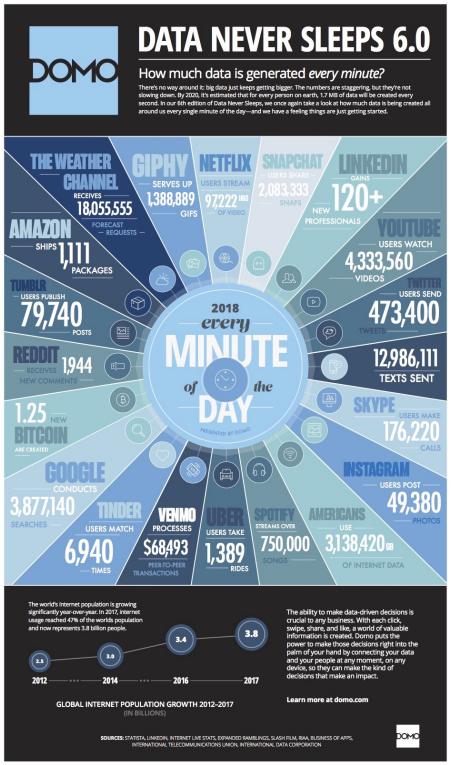


Image Source: http://www.marketwatch.com/story/one-chart-shows-everything-that-happens-on-the-internet-in-just-one-minute-2016-04-26

Data Generated Every Minute on Internet

http://www.visualcapitalist.com/internet-minute-2018/

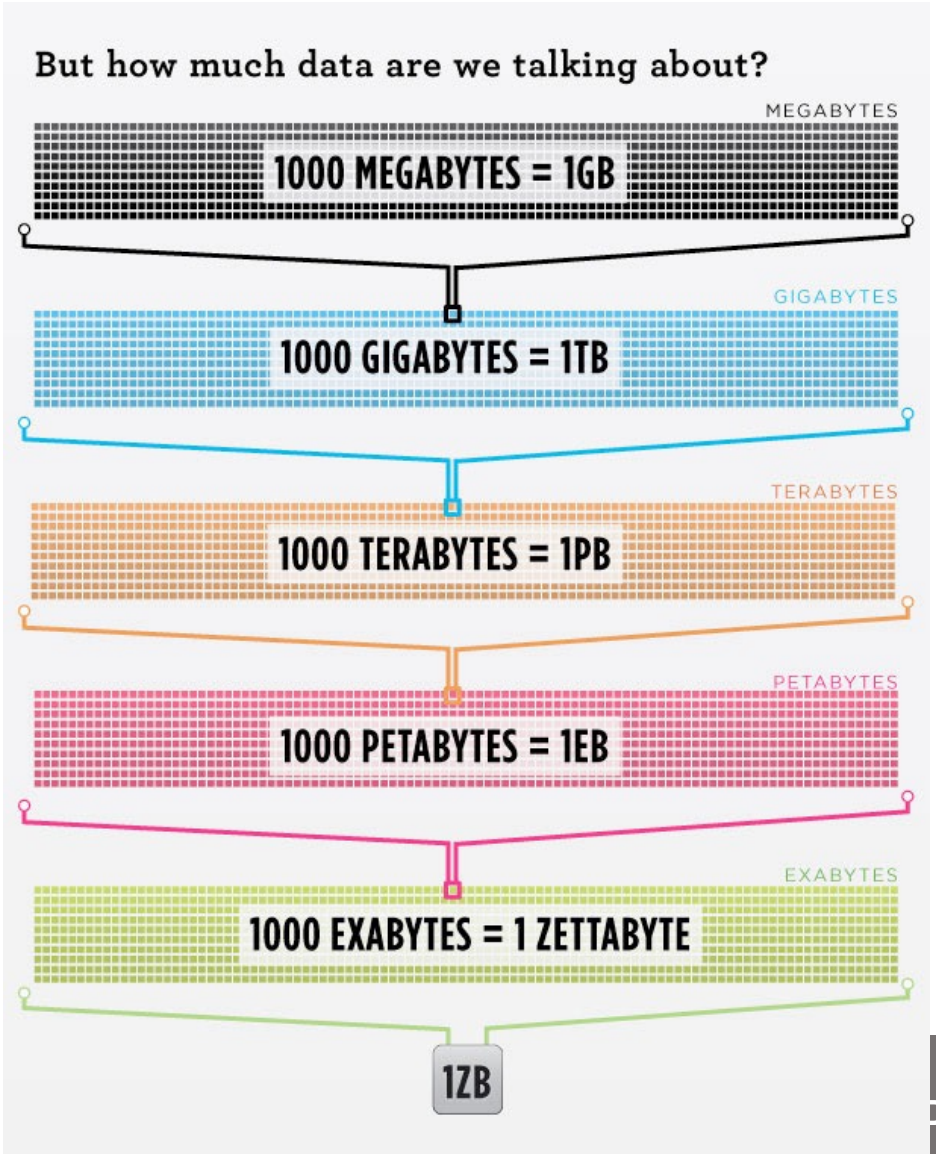


https://www.adweek.com/digital/domo-data-never-sleeps-5/

Scientific Data Management and Analysis

- HPWREN: hpwren.ucsd.edu
  - 30 TB of data annually
- MODIS: modis.gsfc.nasa.gov
  - 219 TB of data annually
- Precision Medicine
  - 4 EB (10^{18} bytes) of data in 2016 (www.fastcompany.com)
- LIGO, Deep Space Network, Protein Data Bank, …

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)



100 MBs ~= couple of volumes of Encyclopedias

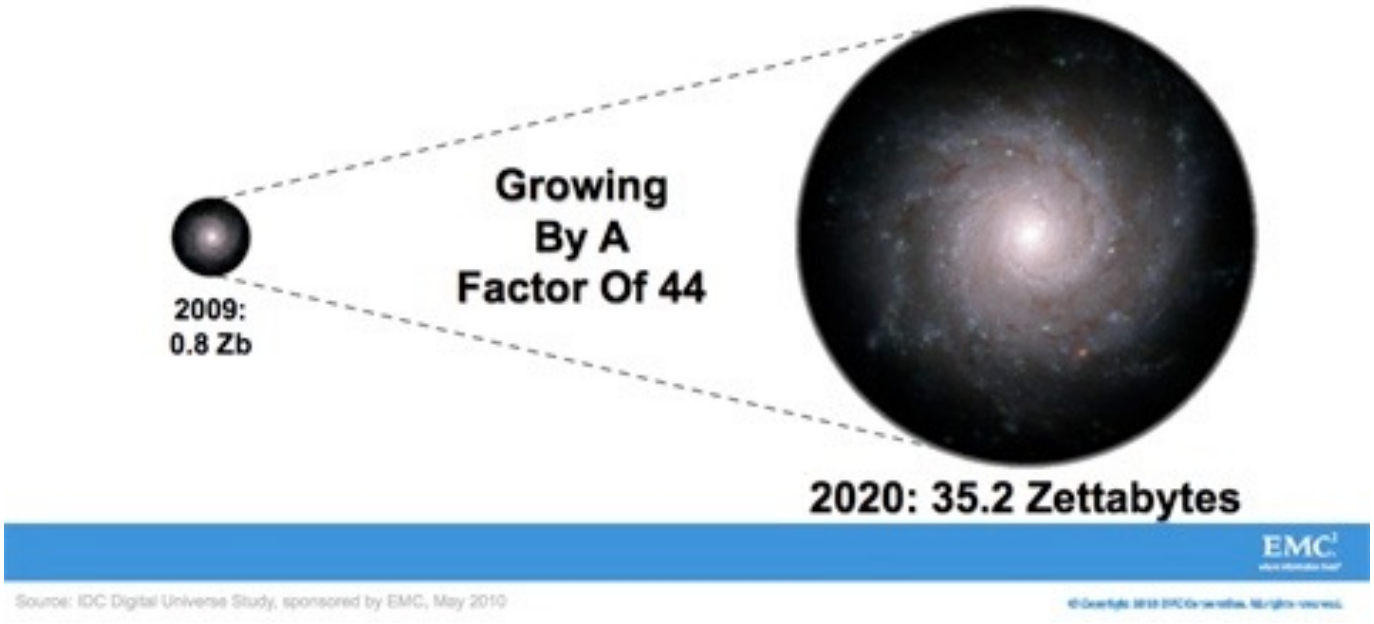
A DVD ~= 5 GBs

1 TB ~= 300 hours good quality video

LHC ~= 15 PBs a y

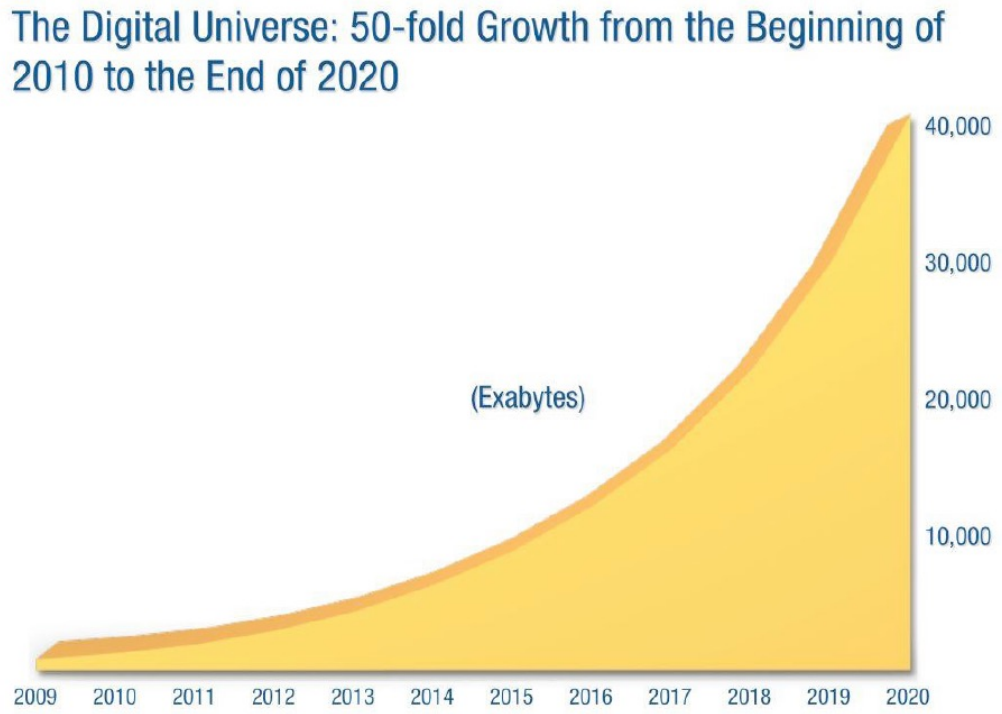
(ialtintas@ucsd.edu)

The Digital Universe 2009-2020



Source: IDC Digital Universe Study, sponsored by EMC, May 2010

Exponential data growth!



This IDC graph predicts exponential growth of data from around 3 zettabytes in 2013 to approximately 40 zettabytes by 2020. An exabyte equals 1,000,000,000,000,000,000 bytes and 1,000 exabytes equals one zettabyte. Source: IDC's Digital Universe Study, December 2012, http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

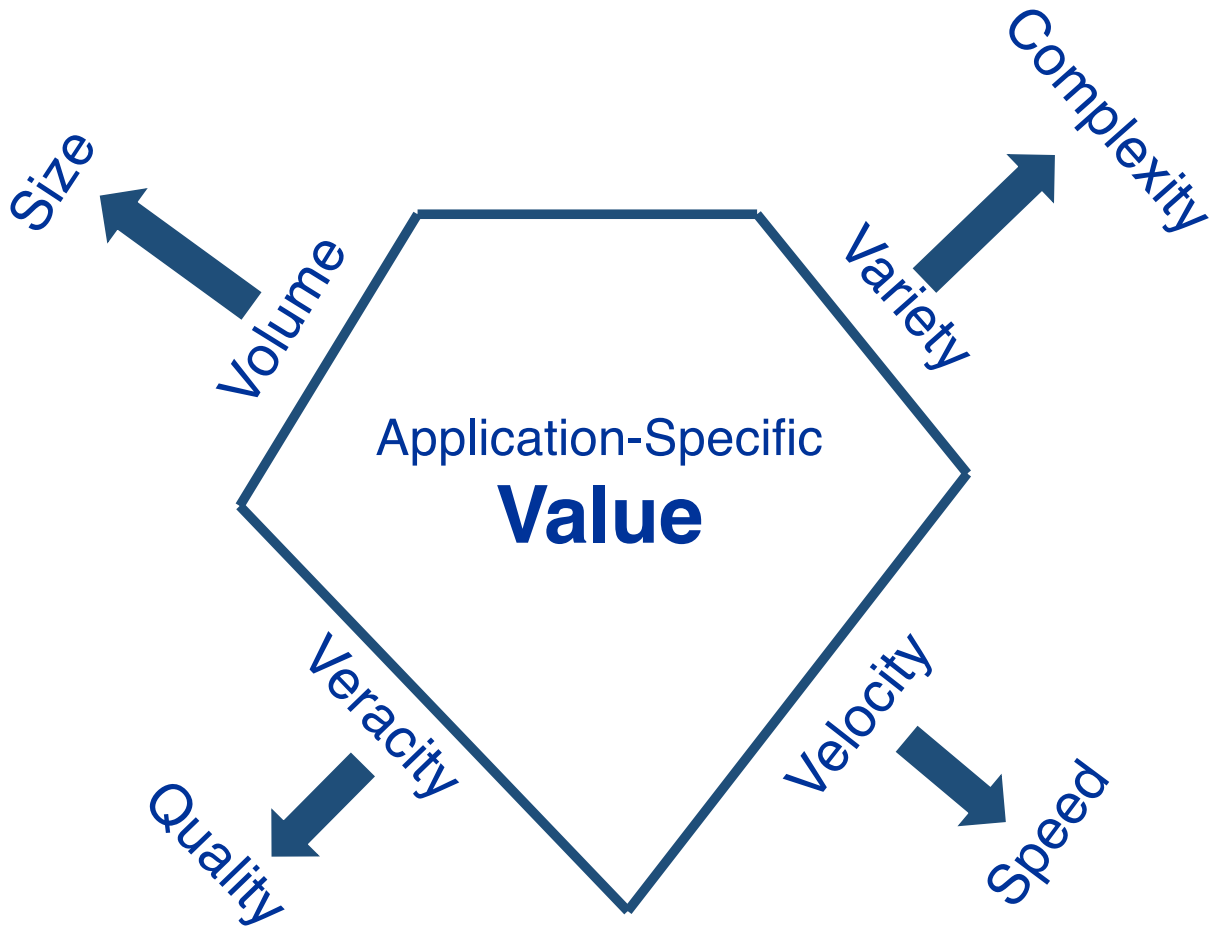
Data Deluge

"We are drowning in information and starving for knowledge"

– John Naisbitt

Source: Megatrends, 1982

Image Source: http://www.digitalzenway.com/2011/12/data-diet-a-resolution-you-can-stick-to/


Size
Volume
Complexity
Variety
Application-Specific
Value
Veracity
Quality
Velocity
Speed

What problem are these disruptors helping to solve?

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

New Opportunities for AI-Driven Approaches and Cyberinfrastructure

- Closing the loop between observation, experimentation and simulation
- Dynamic data-ng data and AI
- Real-time data otation
- Reactive syster
- Coupling the e computations
- New software and tools for AI-driven computing
- Intelligent security, integrity and privacy practices

All require getting value from data as fast as possible!

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

How do we find the connections?

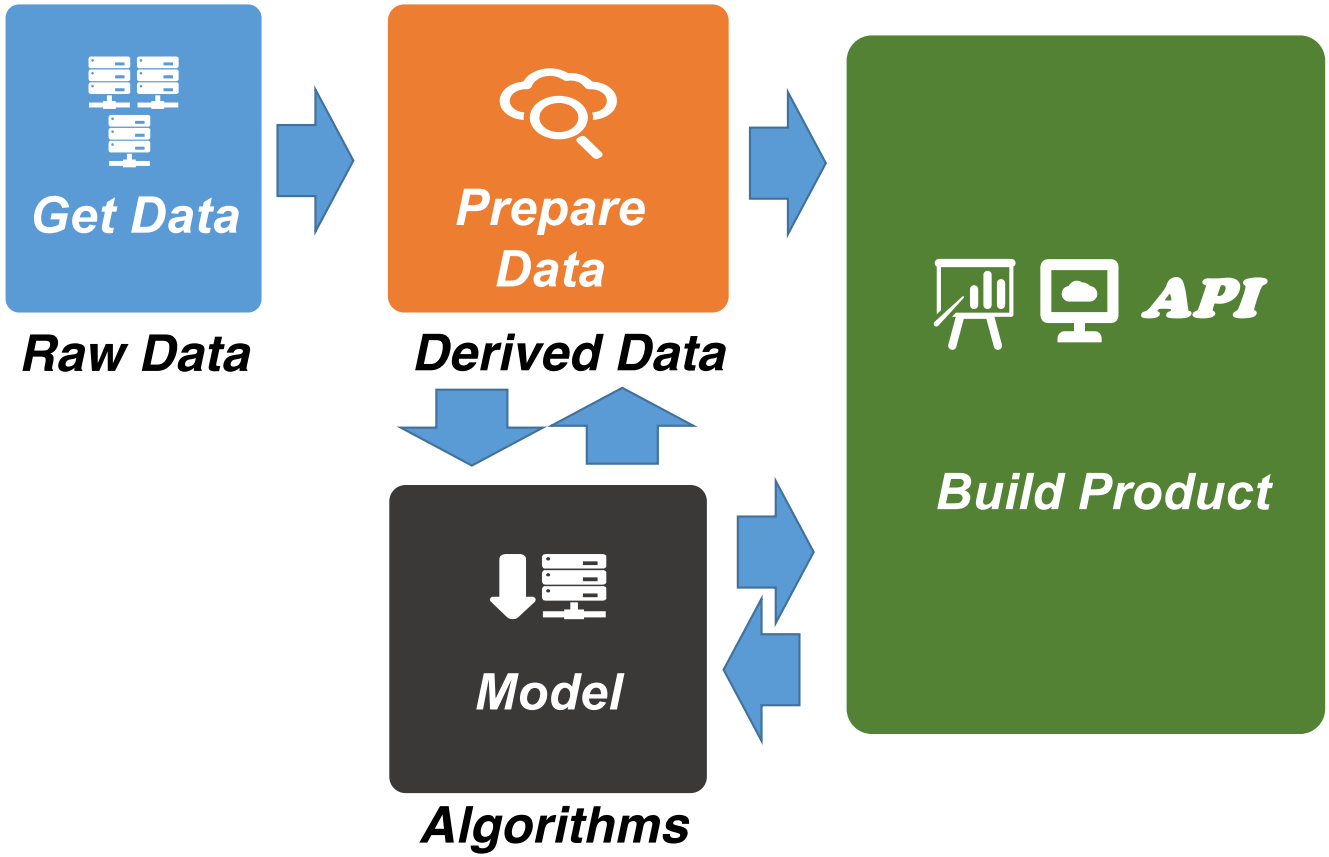
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Part 4:

Scalable Data Science Process

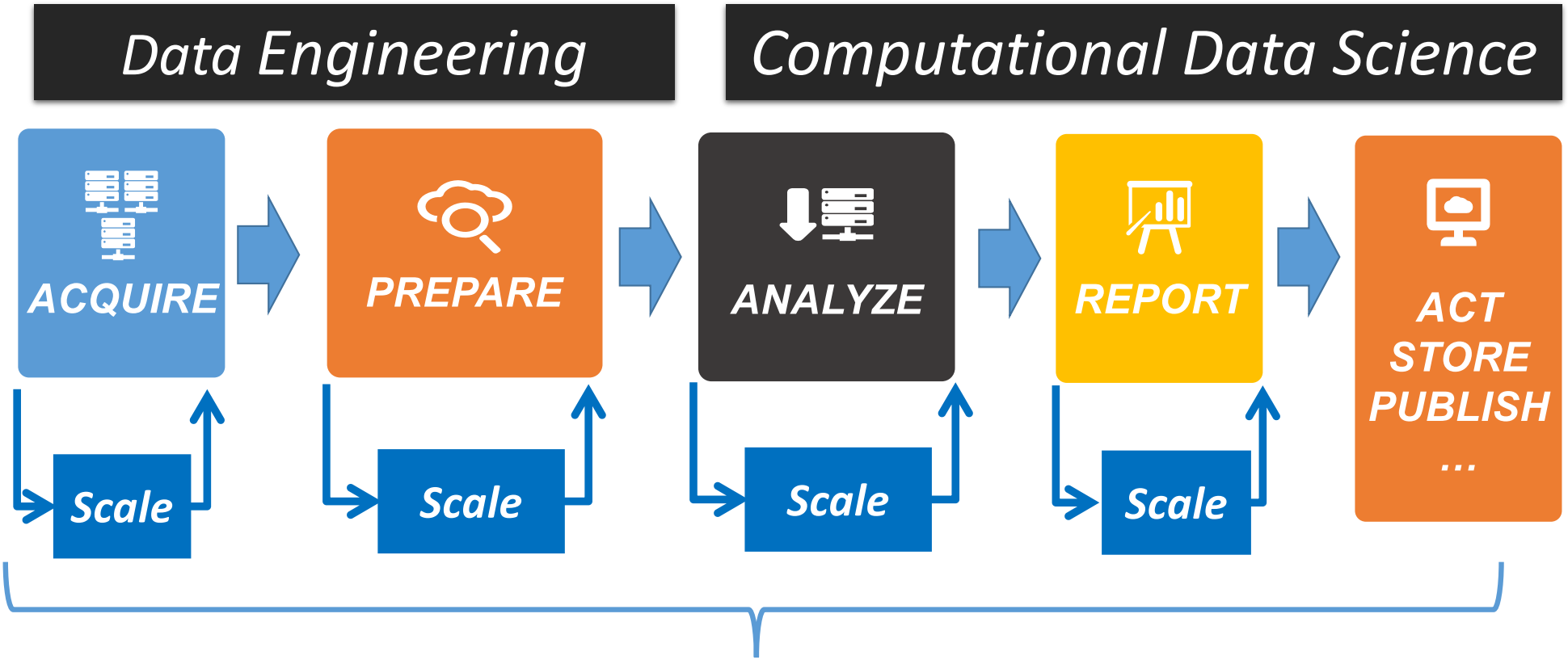
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Going from raw data to a model using data science…



- *Visual Dashboards*
- *Web Interfaces*
- *Programming Interfa*
- *Robotics platforms*

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

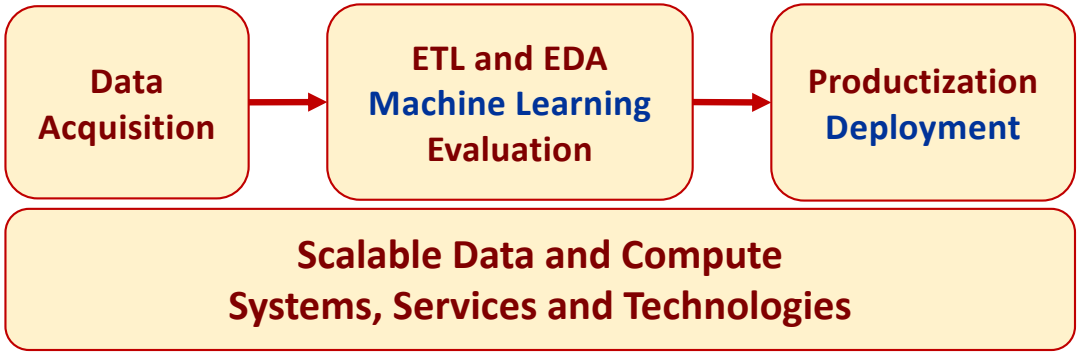


Continuous Iteration, Integration, Programmability, Measurement and Scalability

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Scalable Data Science Process

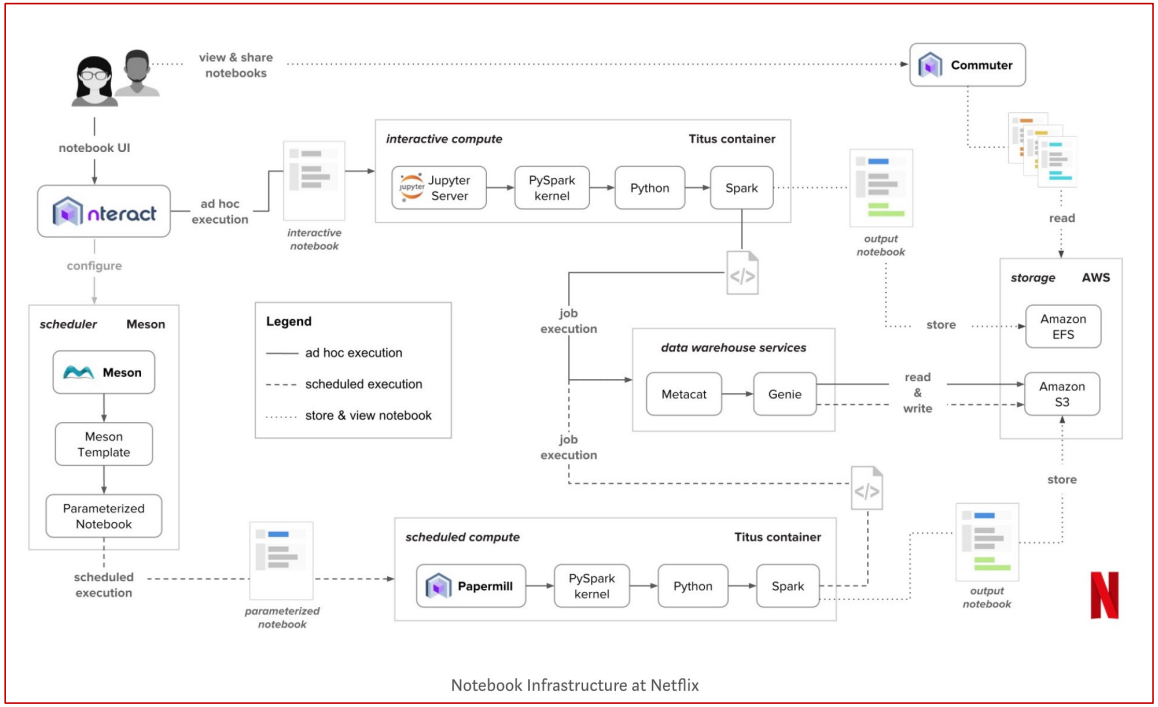
together with



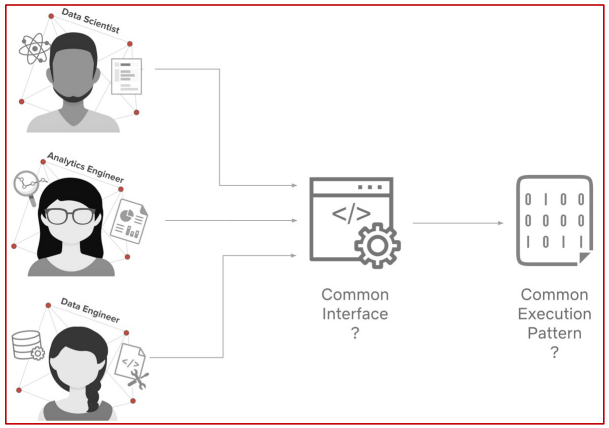
CRISP-DM: Cross-industry standard process for data mining

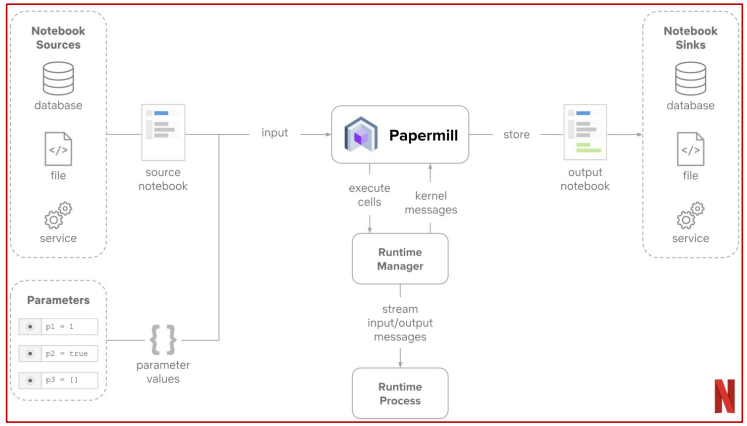
Example: Notebook Innovation at Netflix

- https://medium.com/netflix-techblog/notebook-innovation-591ee3221233
- https://medium.com/netflix-techblog/scheduling-notebooks-348e6c14cfd6



Notebook Infrastructure at Netflix





İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Pieces of the Solution

- Stakeholders
- Datasets and data lifecycle
- Compliance requirements
- Defined actions
- Analytical methods
- Technical infrastructure

Part 5:

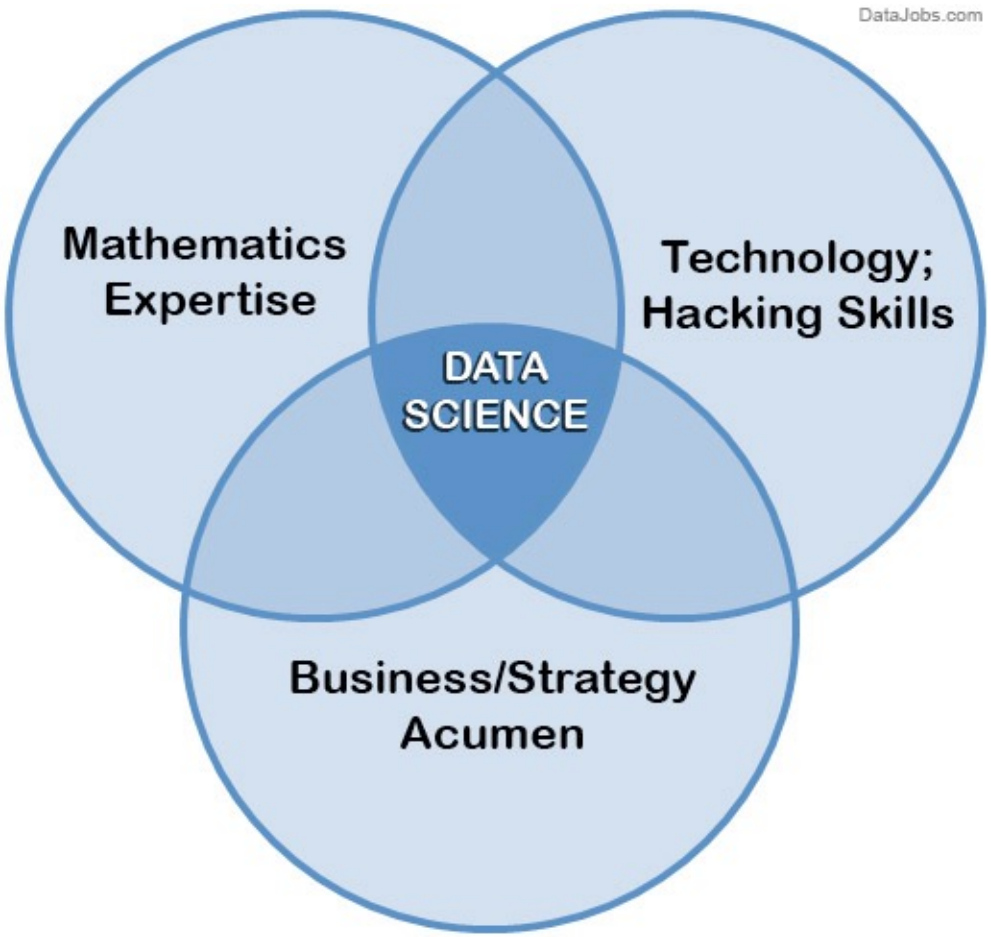
Team Data Science

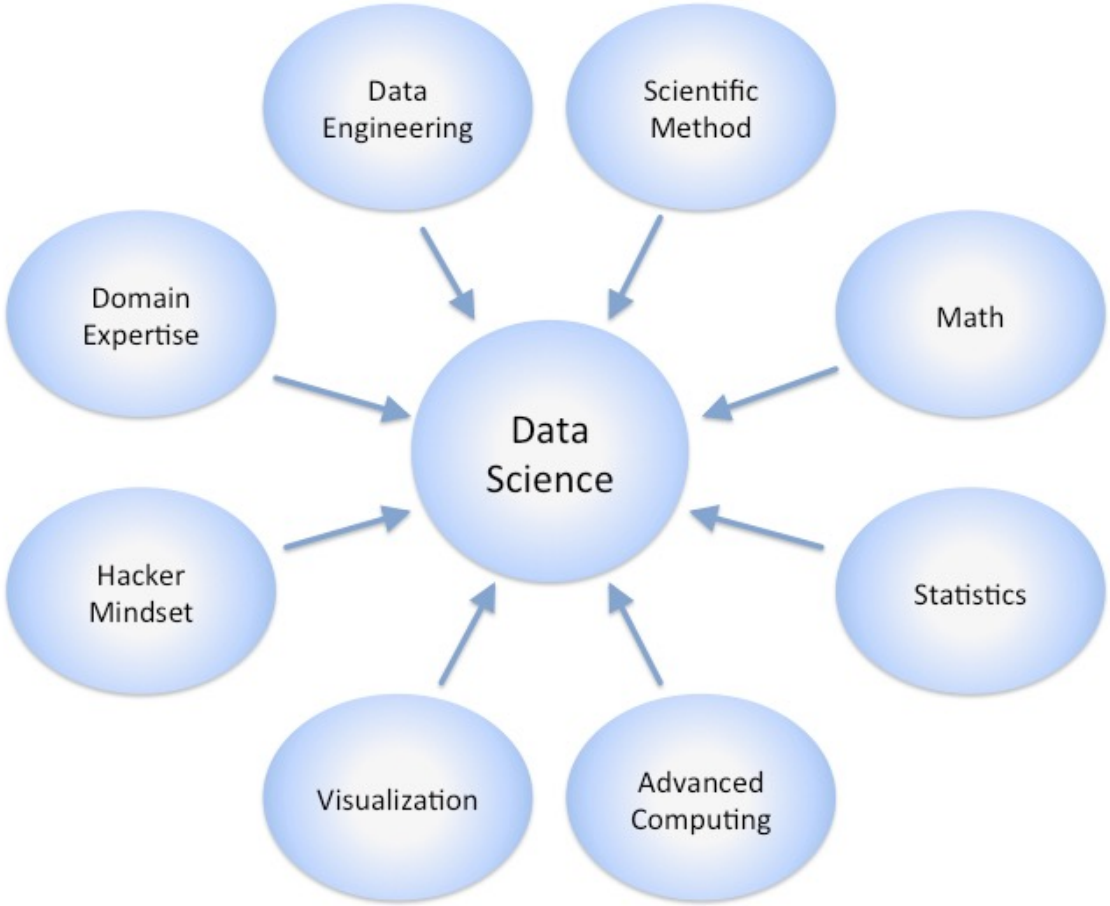
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Modern Data Science Skills

- Programming in Python
- Statistics
- Machine Learning
- Scalable Big Data Analysis

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)


DataJobs.com
Mathematics Expertise
Technology; Hacking Skills
DATA SCIENCE
Business/Strategy Acumen


Data Engineering
Scientific Method
Domain Expertise
Math
Data Science
Hacker Mindset
Statistics
Visualization
Advanced Computing

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

Are data scientists unicorns?

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

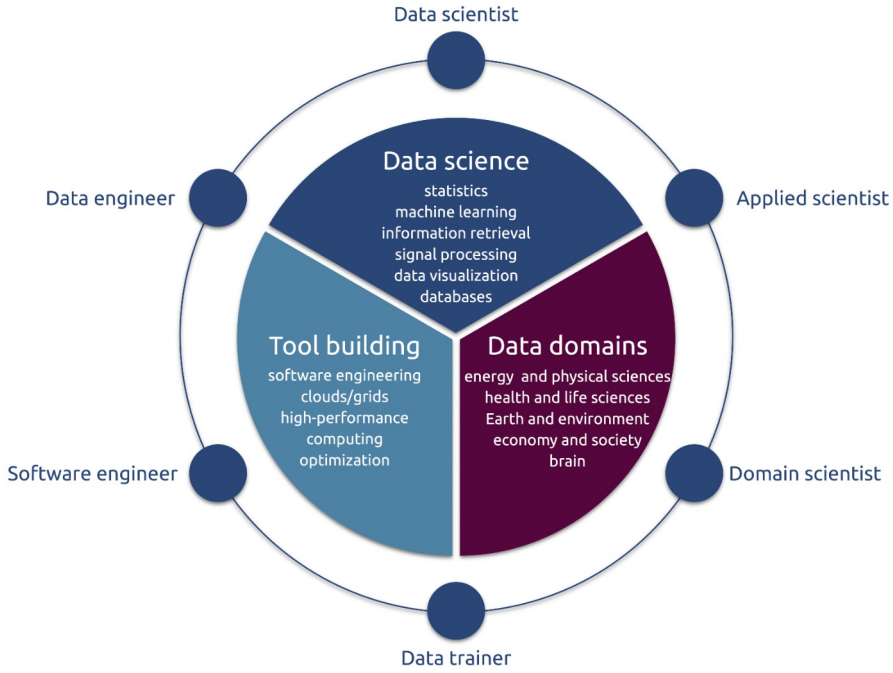
Data science is team sport!

Data Science is “WE” Science!

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Data Science Team

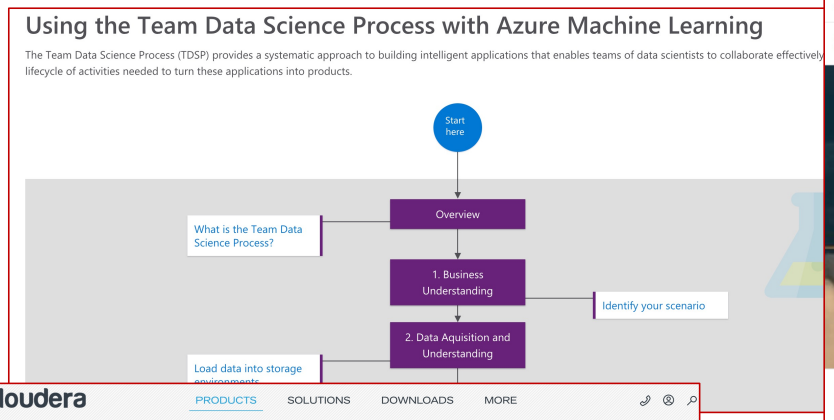
- Data engineer
- Data analyst
- Methods expert
- Scalability and operations expert
- Business manager
- Business analyst
- Scientist
- Visualization and dashboard developer
- Solution architect
- Story teller/coordinator
- Project manager

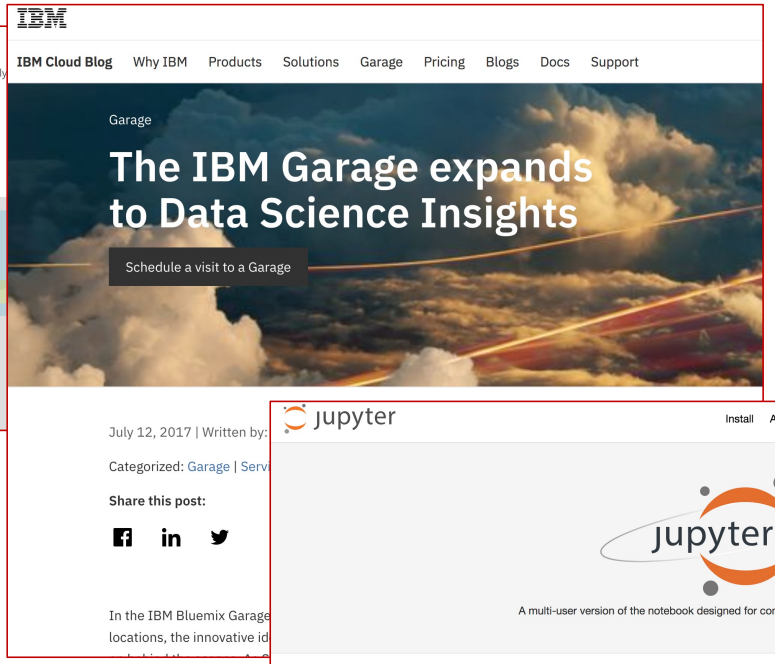


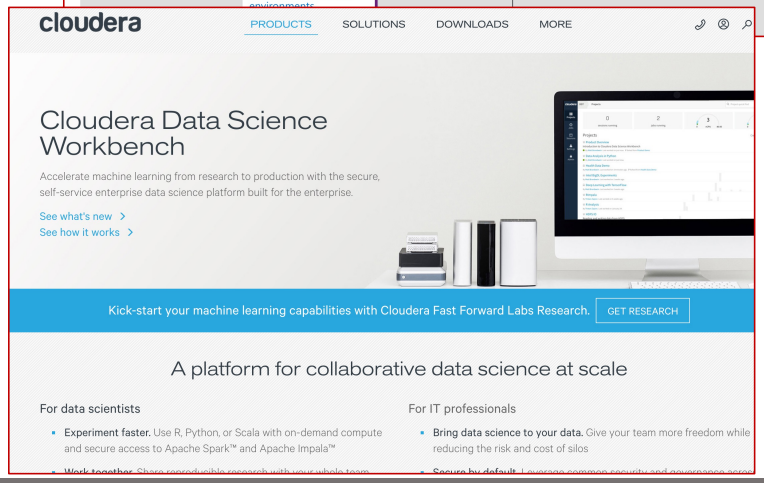
The data science ecosystem: activities and actors
https://medium.com/@balazskegl/the-data-science-ecosystem-678459ba6013

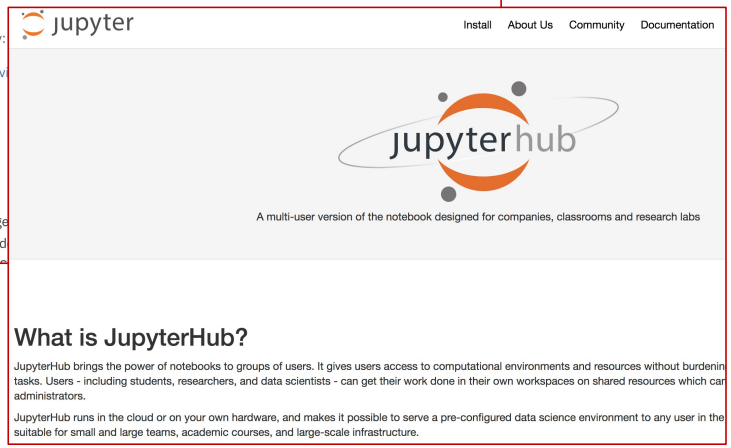
Expertise and skills often overlap, but nobody has it all!

Some examples of team data science practice…







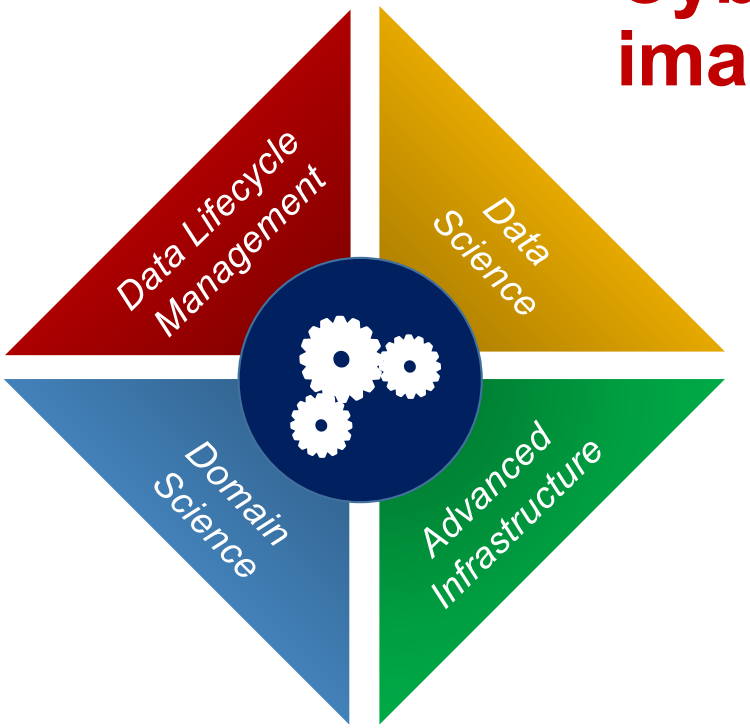


İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Part 6:

Continuous AI Integration at the Digital Continuum

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

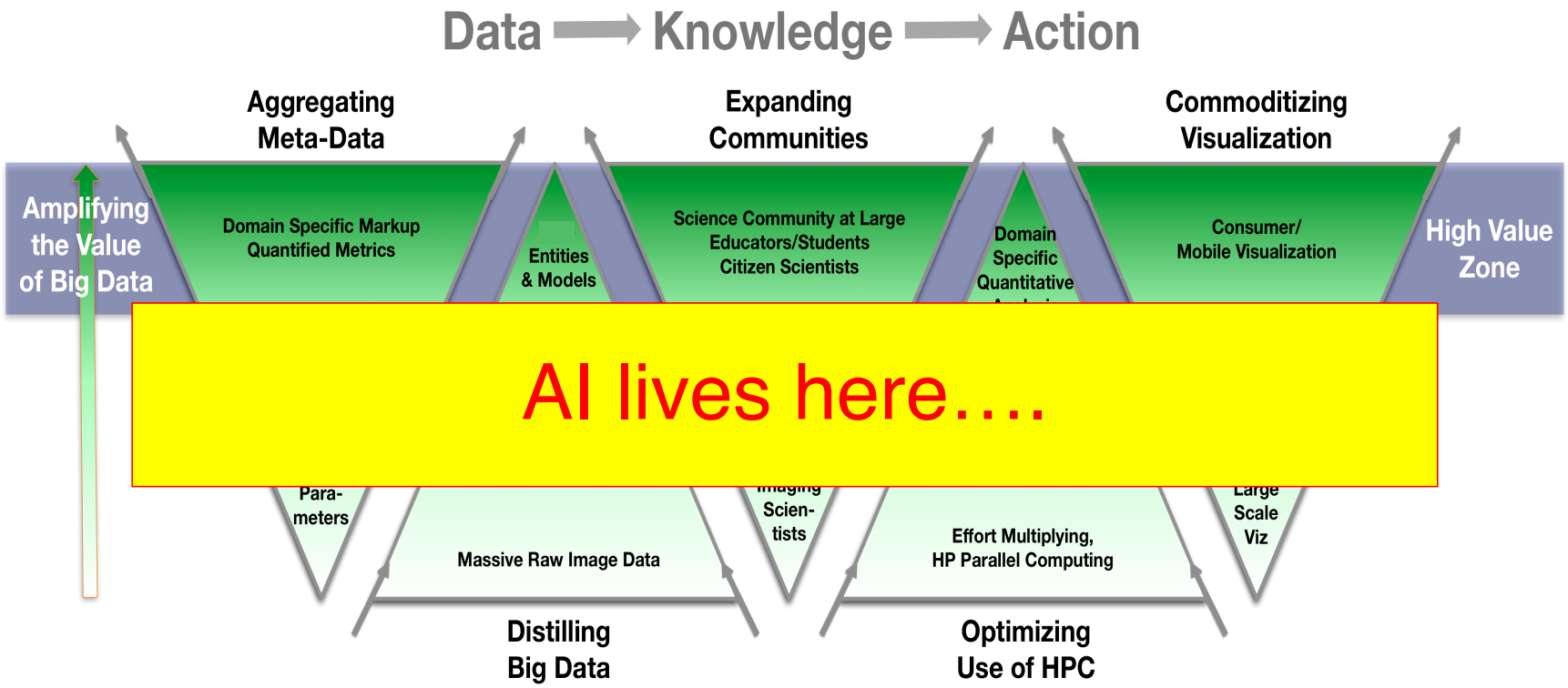


Cyberinfrastructure to support imaging research:

- Heterogenous systems
- Data management
- Data-driven methods
- Scalable tools for dynamic coordination and resource optimization
- Skilled interdisciplinary workforce
- Collaboration tools that enable groups to converge

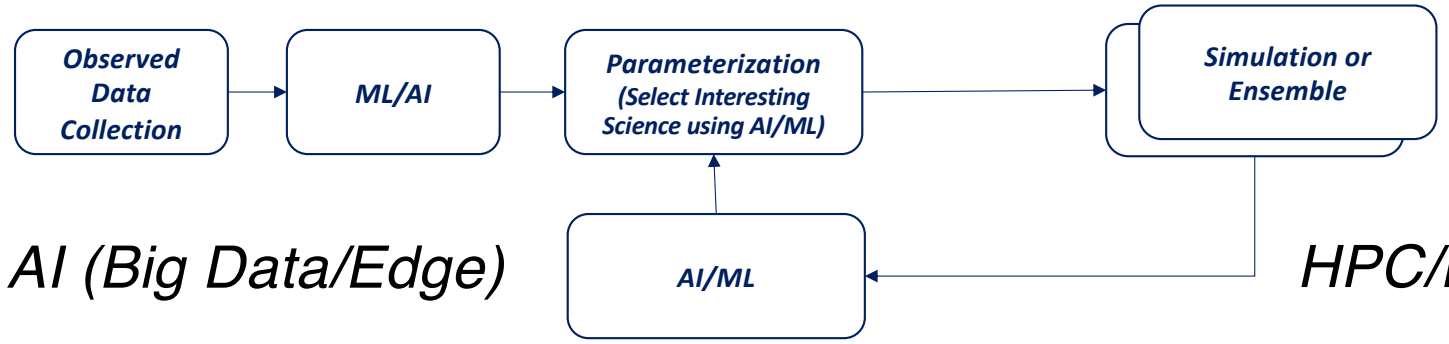
İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

Methodologies, standards, tools and resources are required to amplify value of image data.

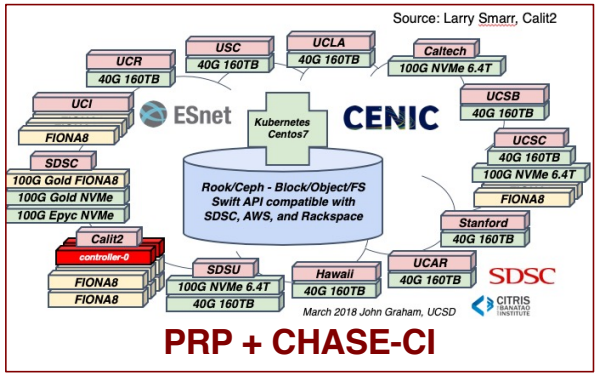


A Typical Heterogeneous AI-Integrated* Workflow

*AI-Integrated refers to **Dynamic Data-Driven** applications that require on-demand network capability to an instrument.



CI for AI : Capability ⟷ **CI for HPC/HTC** : Capacity



PRP + CHASE-CI

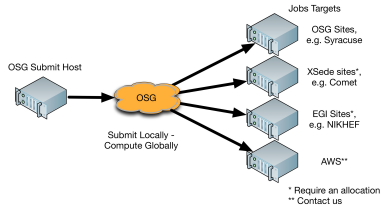
xPU → GPU, CPU, TPU, IPU, QPU,…

Edge

FPGA

hadoop

Big Data



Open Science Grid

Traditional HPC/Storage

1st National Exascale Day! – 10/18/2019

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

CI Requirements for Convergence

Dynamic composability matters.

Systems are only useful if groups can integrate them into applications.

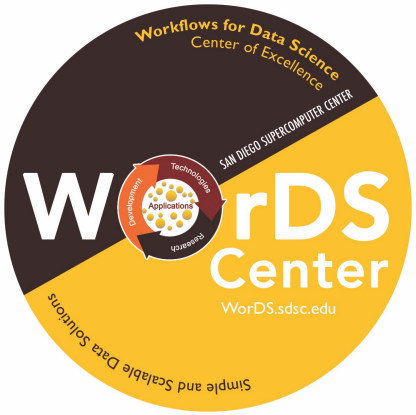
Tools that enhance teamwork need to be coupled with AI systems.

İlkay Altıntaş, PhD (ialtintas@ucsd.edu)

UC San Diego

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

SDSC SAN DIEGO SUPERCOMPUTER CENTER



https://words.sdsc.edu/publications

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

NBCR

it²

HALICIOĞLU DATA SCIENCE INSTITUTE
UC San Diego

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 partners.