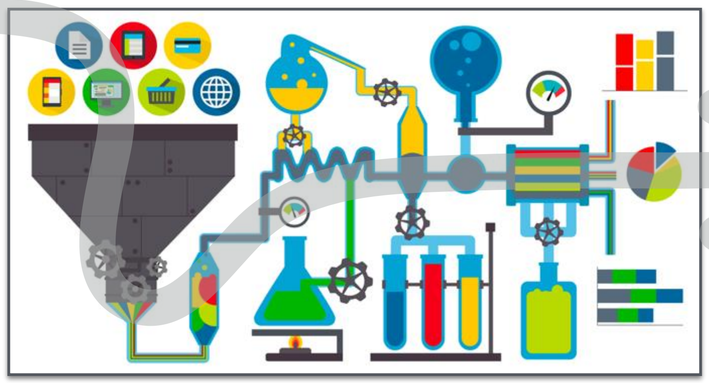
Scaling Data Analytics for ERS

David Stuart

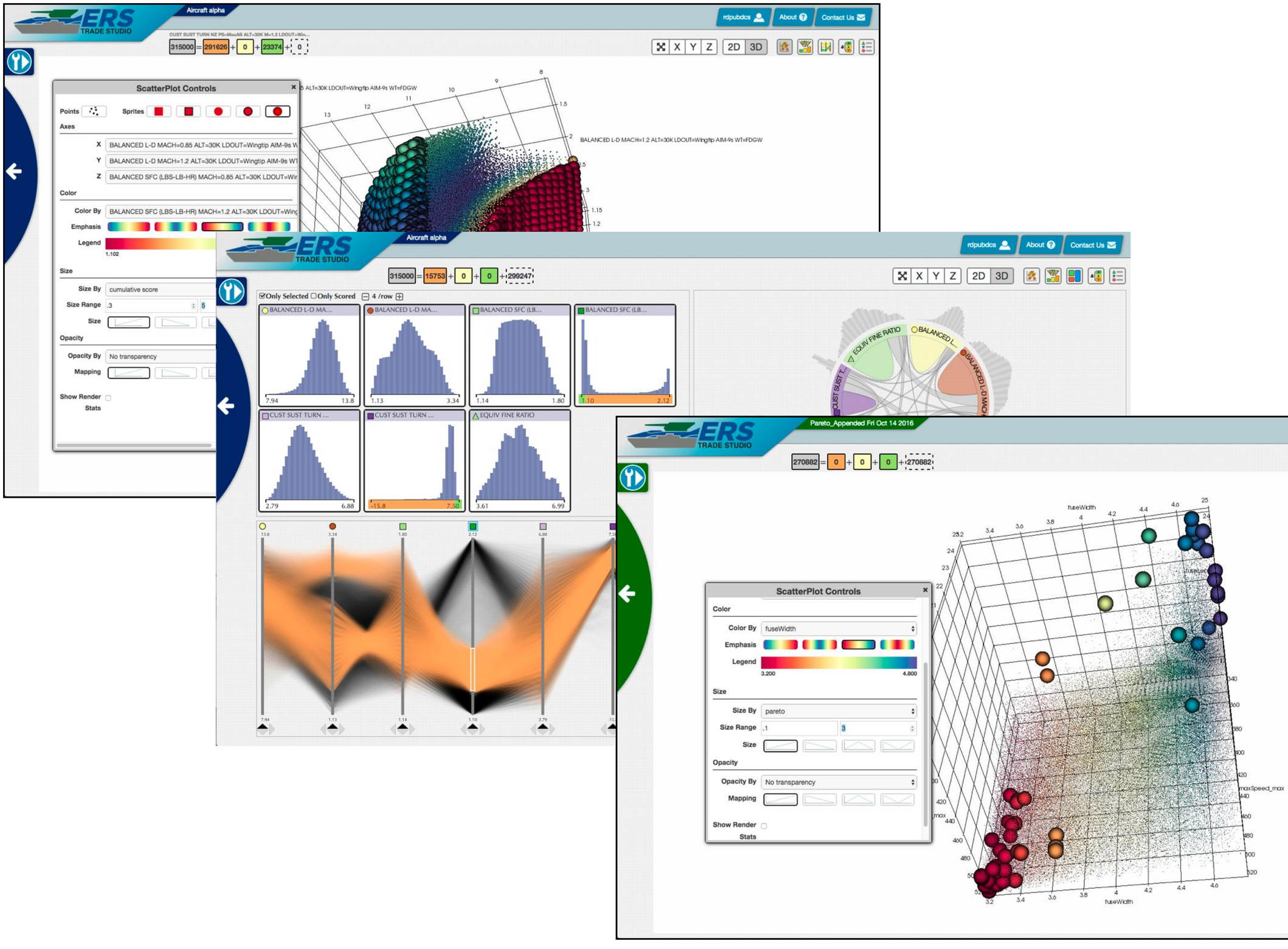
ERS Software Architecture

Data

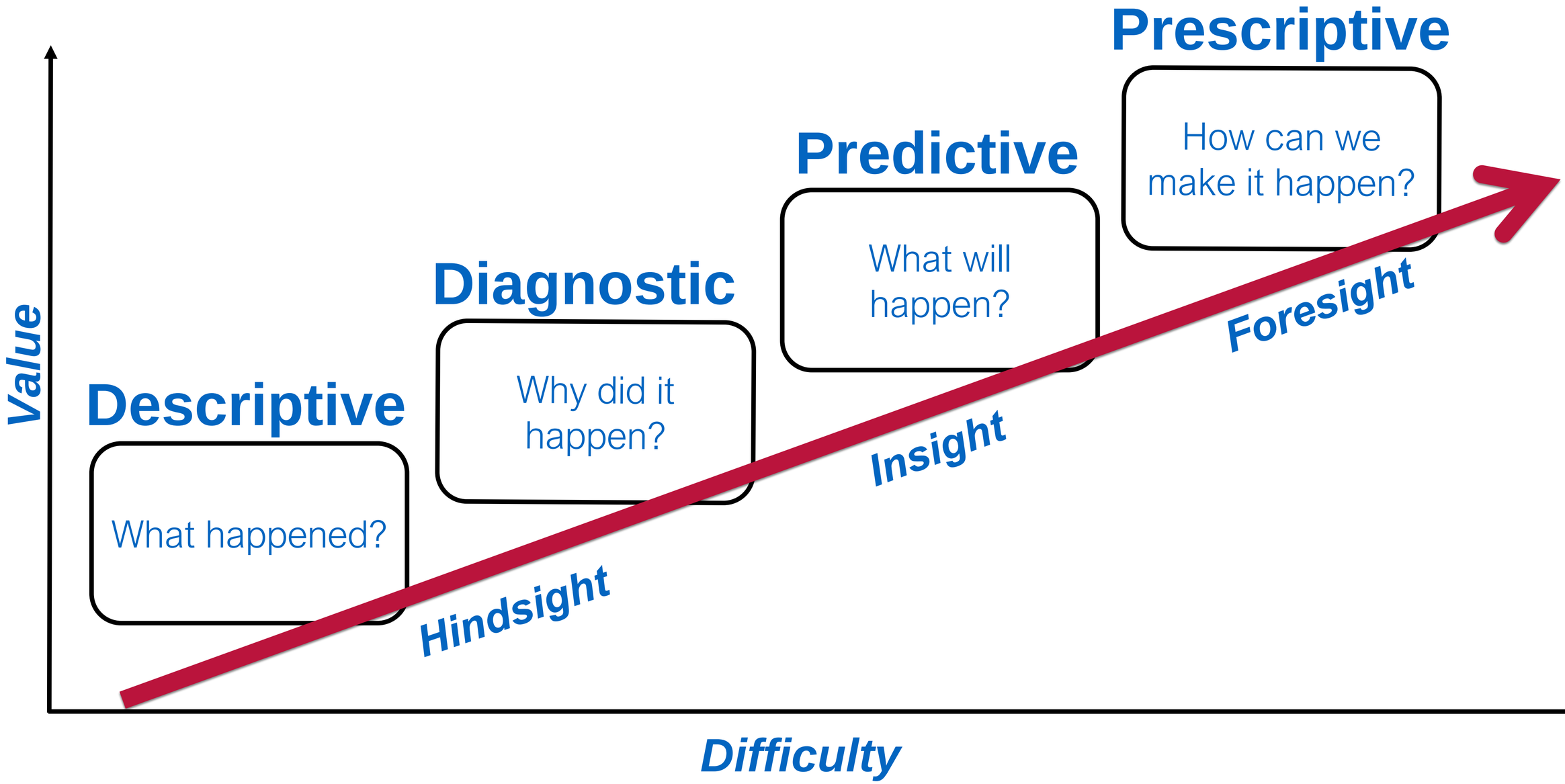


Insights

Current Tools - heavy visualization

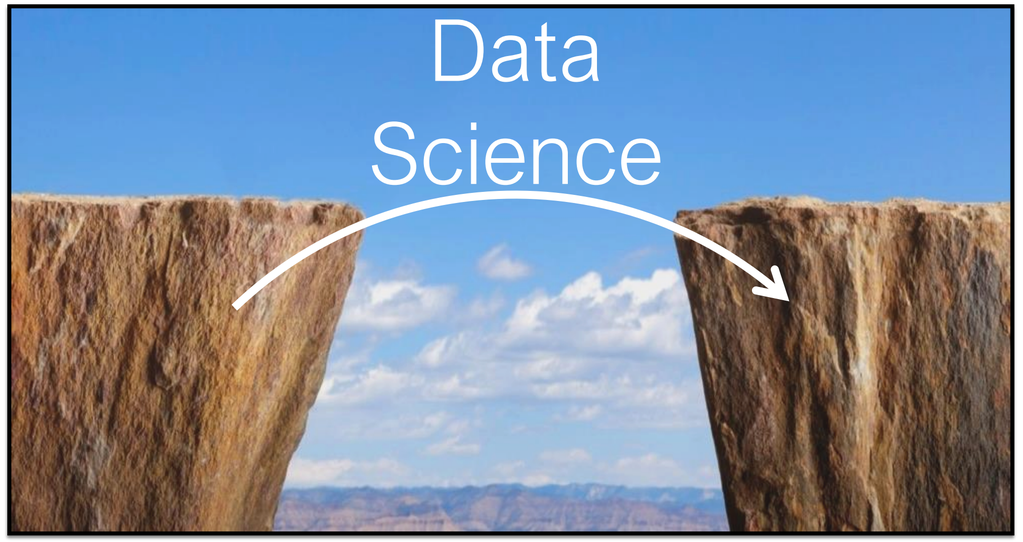


Data Analytics



Data Analytics Overview

- Significant investments in computational capabilities have resulted in unprecedented growth in data.
- Transitioning model results into decision making continues to be a challenge.
- Data science provides the underlying methodologies to produce actionable data for informed decision making.
- It includes machine learning, deep learning, large-scale data analysis, statistics, applied mathematics, and high-performance computing.
- Directly supportive of decision making - applicable to many current and future DoD problems.



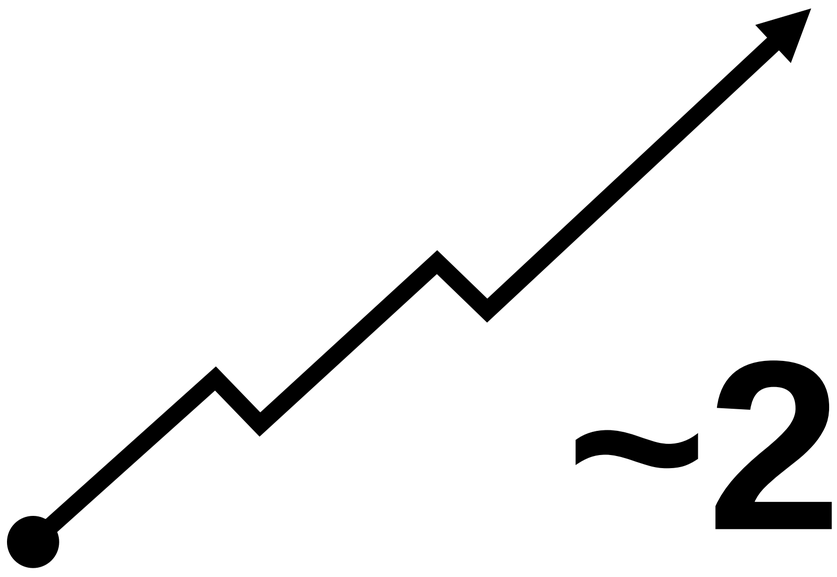
Computational Science and Engineering

Data-informed Decisioning

Data Deluge

1. Quantity of information is soaring
2. Just managing the information is a challenge
3. Accessibility to information is a problem as well
4. However, integrating and preprocessing of the information are the primary bottleneck
5. Analyzing it, to spot patterns and extract useful information, is severely limited due to the challenges above

We are great at creating data!



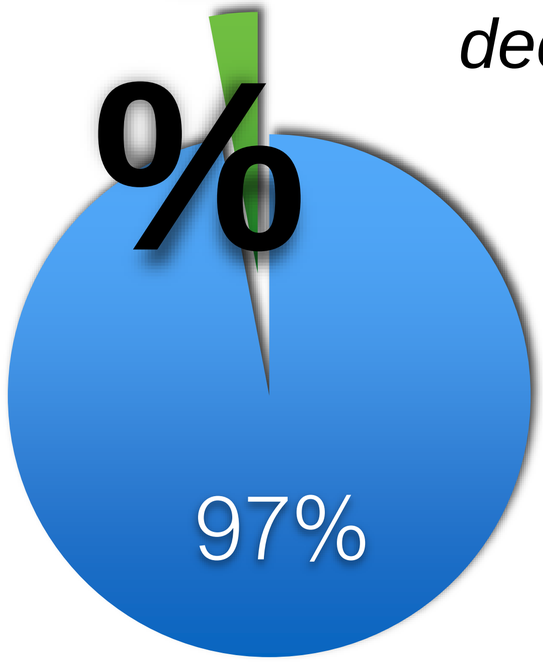
~25%

annual growth in data

3%

of the data is used to make decisions

**IDC Study*



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005

2005 2020

6 BILLION PEOPLE
have cell phones

WORLD POPULATION: 7 BILLION

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT
are shared on Facebook every month

By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month

400 MILLION TWEETS
are sent per day by about 200 million monthly active users

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session

Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth

Veracity
UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions

Poor data quality costs the US economy around
$3.1 TRILLION A YEAR

27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate

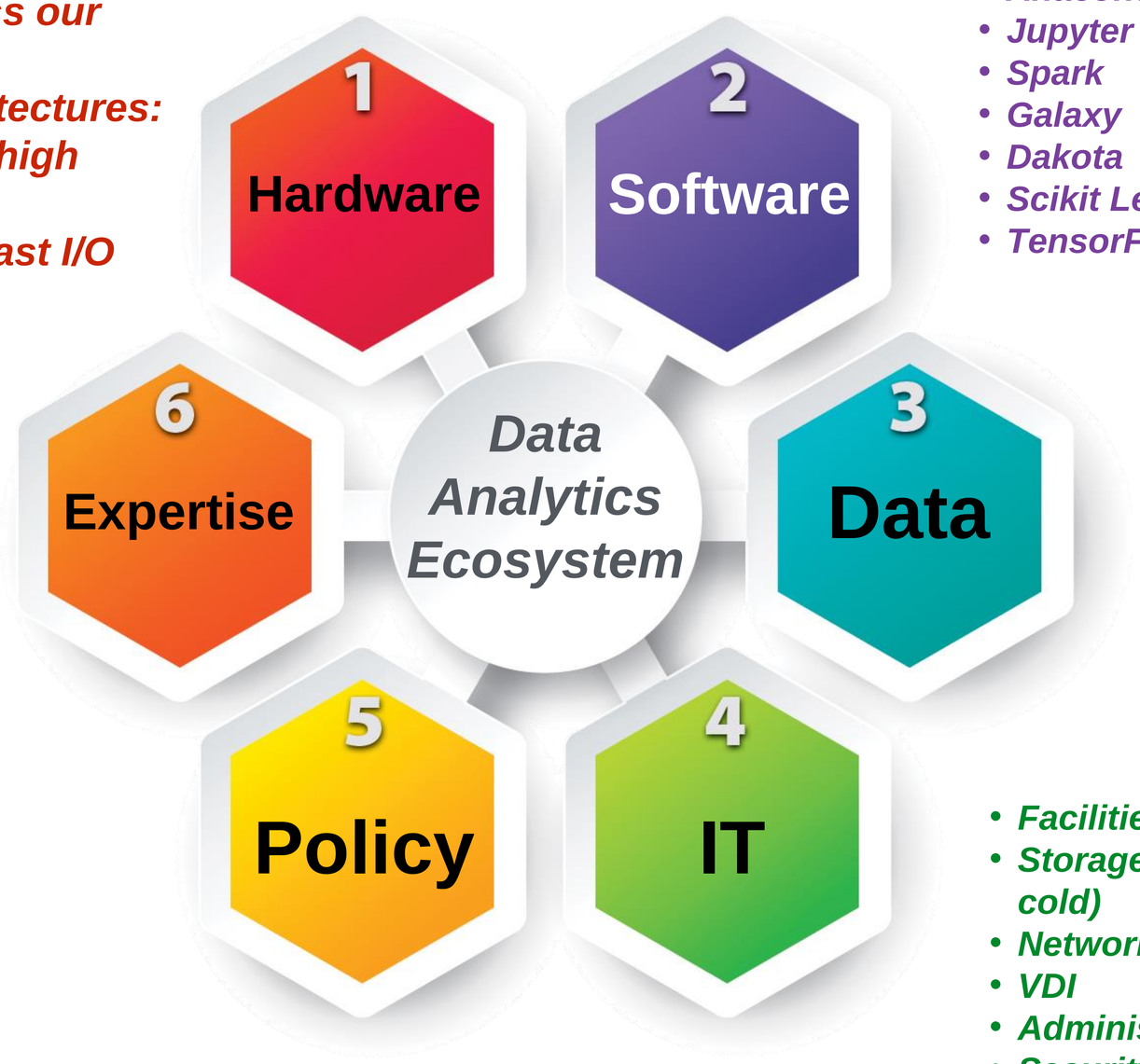
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

Building blocks

- *Machines that can address our largest problems*
- *Blended computing architectures:*
  - *Numerical - distributed, high speed interconnects*
  - *Data - shared memory, fast I/O*

- *Leverages open source capabilities*
- *Python ,R, and C/C++ (when needed)*
- *Anaconda - package management*
- *Jupyter Notebooks*
- *Spark*
- *Galaxy*
- *Dakota*
- *Scikit Learn (machine learning)*
- *TensorFlow (deep learning)*



- *Terabytes collected - Unorganized and unaccessible*
- *Streamline data wrangling*
- *Minimize the movement of data*
- *Leverage database technologies*
  - *SQL and noSQL*

- *Computational and data scientists assist in problem step up, execution, and visualization*

- *Integrate necessary policies*

- *Facilities*
- *Storage (hot, warm, and cold)*
- *Networks (10G+)*
- *VDI*
- *Administration of machines*
- *Security - (monitoring, patching, etc.)*

Data Analytics Ecosystem

Mass Storage Services

- ERDC's MSAS System (Gold)
- 1PB SAMFS filesystem
- 25PB current on tape (80 PB capacity)
- 3 Oracle T54 servers running Solaris 11
- Data transfers are equally distributed among the 3 T54 servers)

Data Analytics

- Anaconda Enterprise
- Spark
- Machine Learning
- Deep Learning

High Speed Network (100G)

General Computing Services

- Virtualization
- Persistent Services
- Data Management
- Interactive Access

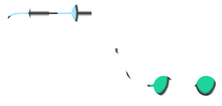
Computational Science

- 126,192 compute cores
- 544 Knights Landing nodes
- 32 GPGPUs
- 437 terabytes of memory

Models

Dakota
(optimizer)

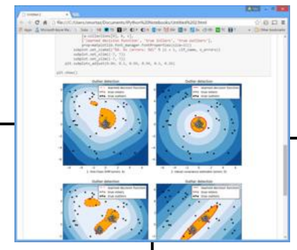
Galaxy
(workflow)



Computational Science & IT

Decision Gateways

Jupyter Notebooks



Web or VDI connection

Data Scientists

Web connection

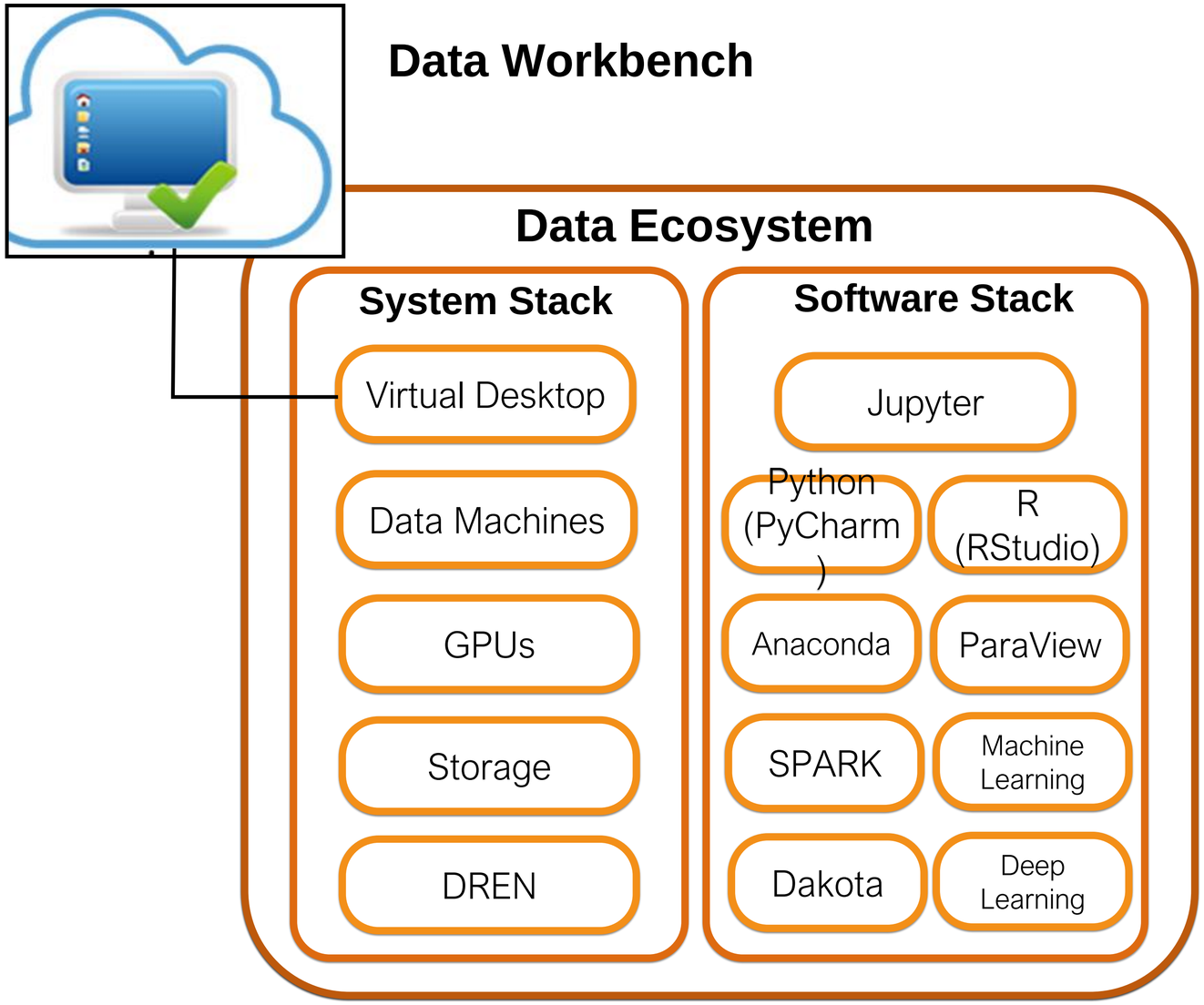
Decision Makers

Data Workbench

Data Science & Decision Making

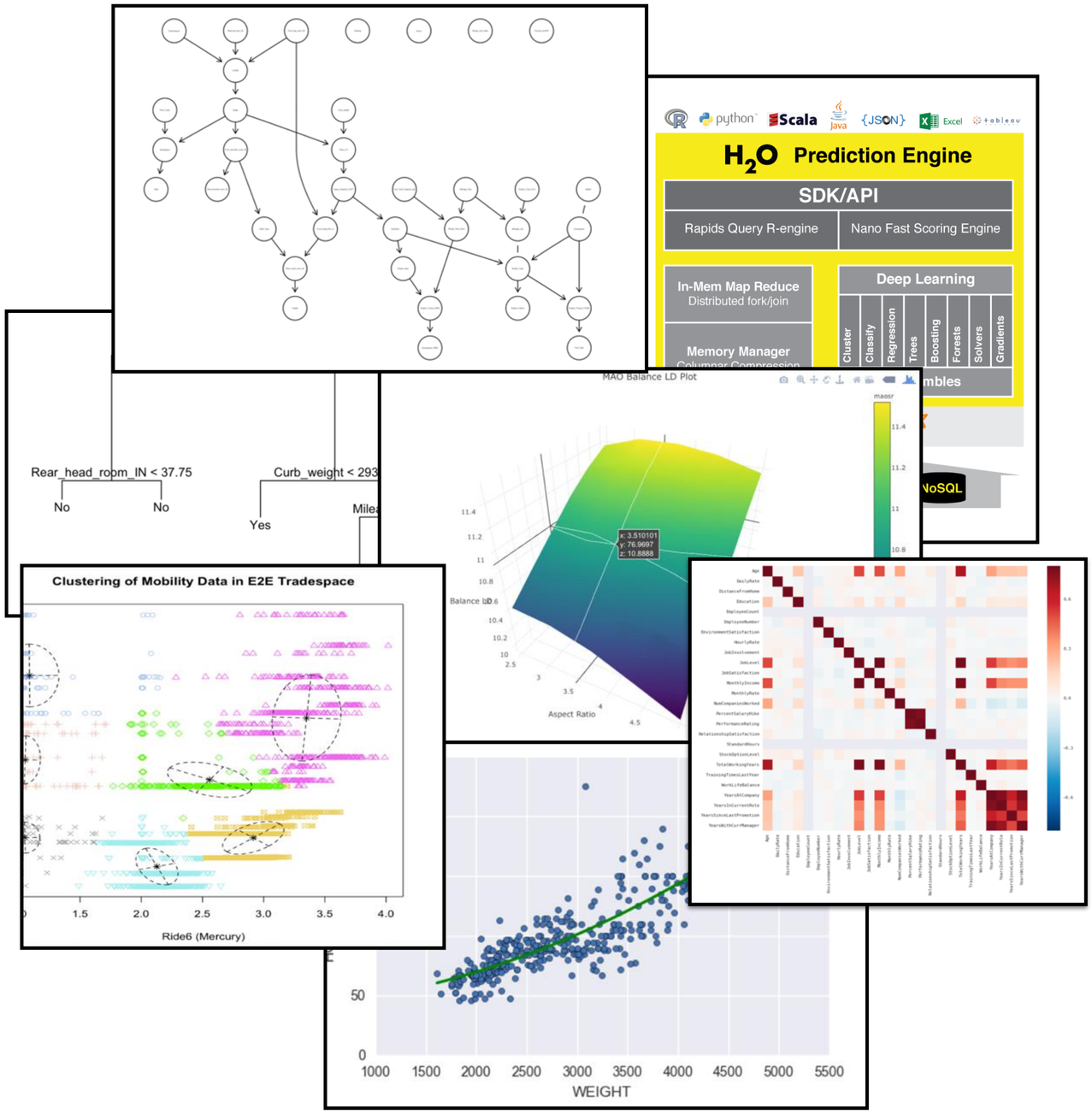
High-Performance Data Analytics Workbench

- Provides a DoD ecosystem for conducting large-scale data analytics
- Extends existing DoD investments in HPC to include data-intensive computing
- Provides users remote desktop access to *Workbench*
- Allows for the development of data analytics training curriculum based on *Workbench*
- Enables the deployment of *Workbench* on designated HPCMO machines



H2O for Machine Learning

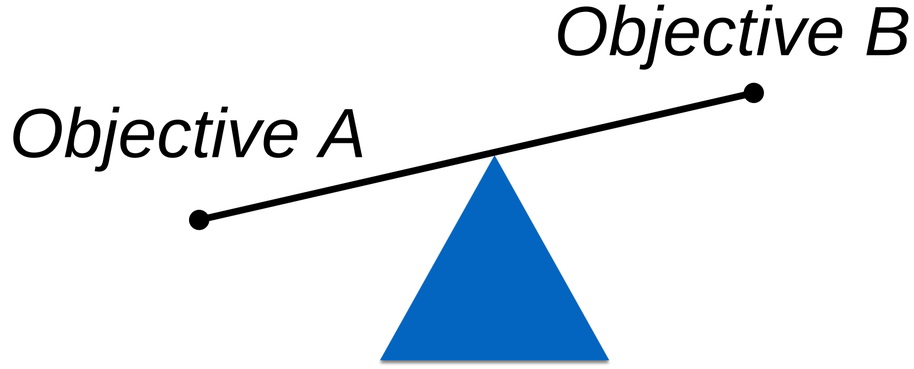
- Open source, written in Java
- Seamless integration with Hadoop and Spark
- Use common programming languages or web-based GUI
- Massively scalable – train on billions of rows of data
- Optimal efficiency - fine-grain distributed processing using in-memory processing combined with fast serialization between nodes and clusters
- Export models to Java objects for fast, real-time scoring
- Visual inspection of models during training through H2O Flow GUI



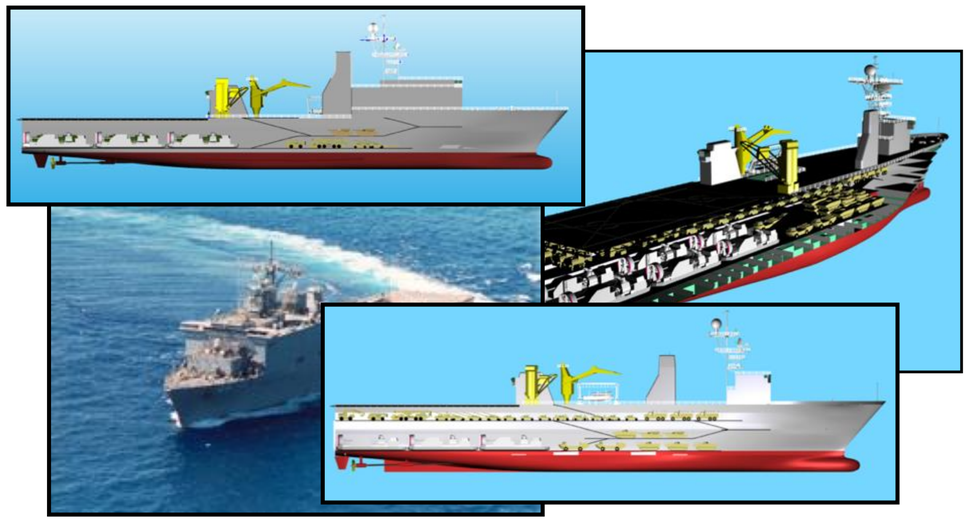
Trade-off Analytics

A technique of reducing or forgoing one or more desirable outcomes in exchange for increasing or obtaining other…

- Helps people make better choices when faced with multiple, often conflicting, goals and alternatives.
- Uses mathematical filtering techniques to identify the best candidate options based on different criteria,
- Allows users to explore the trade-offs between options to make complex decisions.
- Combines smart visualization and analytical recommendations for easy and intuitive exploration of trade-offs.

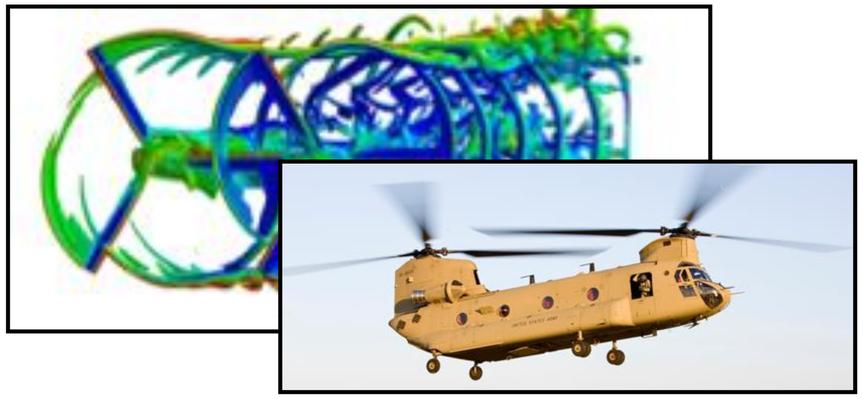


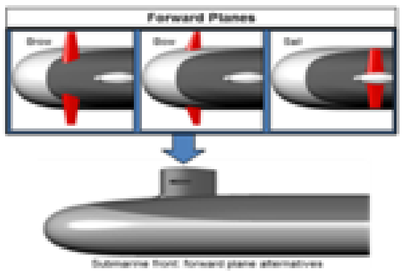
Trade-off Examples



LX(R) Analysis of Alternatives (AoA): Navy researchers collaborated with the Engineered Resilient Systems (ERS) program for the initial LX(R) AoA. The AoA focused on understanding the trade-offs among 19 ship performance attributes, survivability specification levels, and cost. The Navy benefited from a detailed understanding of LX(R) requirements, performance, and cost trade-offs for 22,000 different design concepts – all produced and evaluated within 3 months. *This assessment included potential cost-performance benefits of new technologies early in the development of the system.*

CH-47F Performance Improvement: Army and Boeing researchers used CREATE rotorcraft software to confirm helicopter performance predictions for new Boeing proposed rotor blades, which are expected to provide 2,000 pounds of improved hover thrust for potentially 400+ CH-47Fs (Chinooks). *Additional lift directly translates to 10% more combat capability from the same platform*.





Ohio Replacement Submarine: Noise is the crucial element in detection of submarines – the quieter the submarine, the better chance it has to survive and provide a deterrent option to the National Command Authority. Detailed hydrodynamic evaluations using CREATE software are being used to provide guidance in hull, appendage, and propulsor design and testing for the Ohio Replacement Program (ORP). *These assessments evaluated maneuvering and vibration implications of different configurations directly impacting the design of ORP.*

ERDC

INFORMATION TECHNOLOGY LABORATORY

david.c.stuart@erdc.dren.mil