

# New York Scientific Data Summit 2020:

Data-Driven Discovery in  
Science and Industry

Organized by:

**BROOKHAVEN**  
NATIONAL LABORATORY

**COLUMBIA UNIVERSITY**  
IN THE CITY OF NEW YORK

**FLATIRON**  
INSTITUTE  
a Google Cloud Foundation

**NYU**

**RUTGERS**  
THE STATE UNIVERSITY  
OF NEW JERSEY

**Stony Brook**  
University

**TEXAS**  
The University of Texas at Austin

WEDNESDAY, OCTOBER 21 - CRITICAL INFRASTRUCTURE/MANUFACTURING

## Computational Modeling at GE

Richard Arthur  
Sr. Principal Engineer, GE Research



# Engineering's core competency is Problem Solving



<https://blogs.ge.com/problem-solving>

Team,

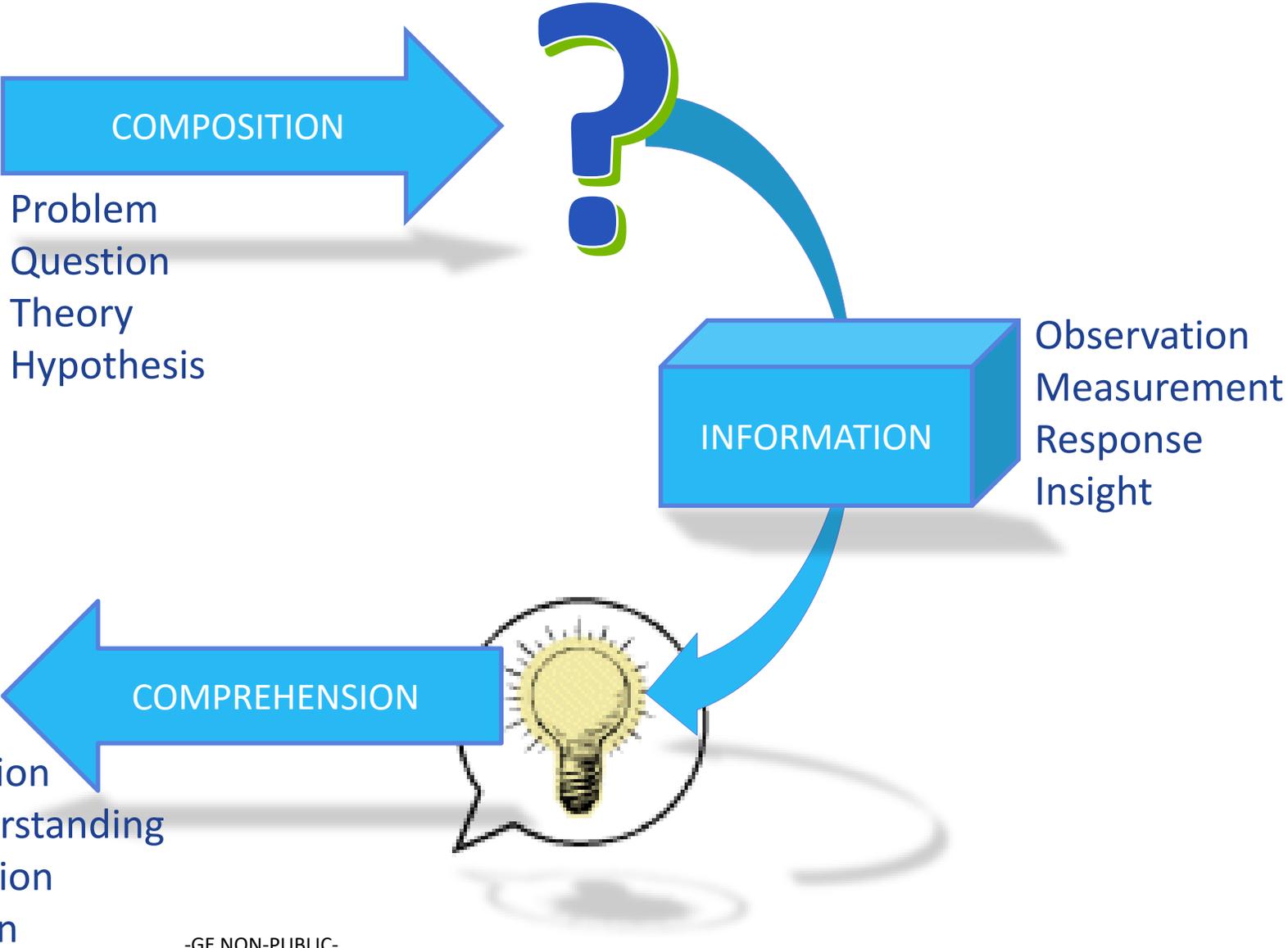
One of my personal resolutions for 2020 is to teach more about lean as we drive our transformation. I will work to do that more often this year in my emails to you. For today's note I'm starting with problem solving because it is foundational to developing competitive advantage, delivering for our customers and ultimately improving our performance over the long-term.

delivering for our customers and ultimately improving our performance over the long-term.  
I'm starting with problem solving because it is foundational to developing competitive advantage,  
transformation. I will work to do that more often this year in my emails to you. For today's note  
one of my personal resolutions for 2020 is to teach more about lean as we drive our

# Problem Solving is Engineering's Core Competency



# Problem Solving



# Engineering's core competency is Problem Solving



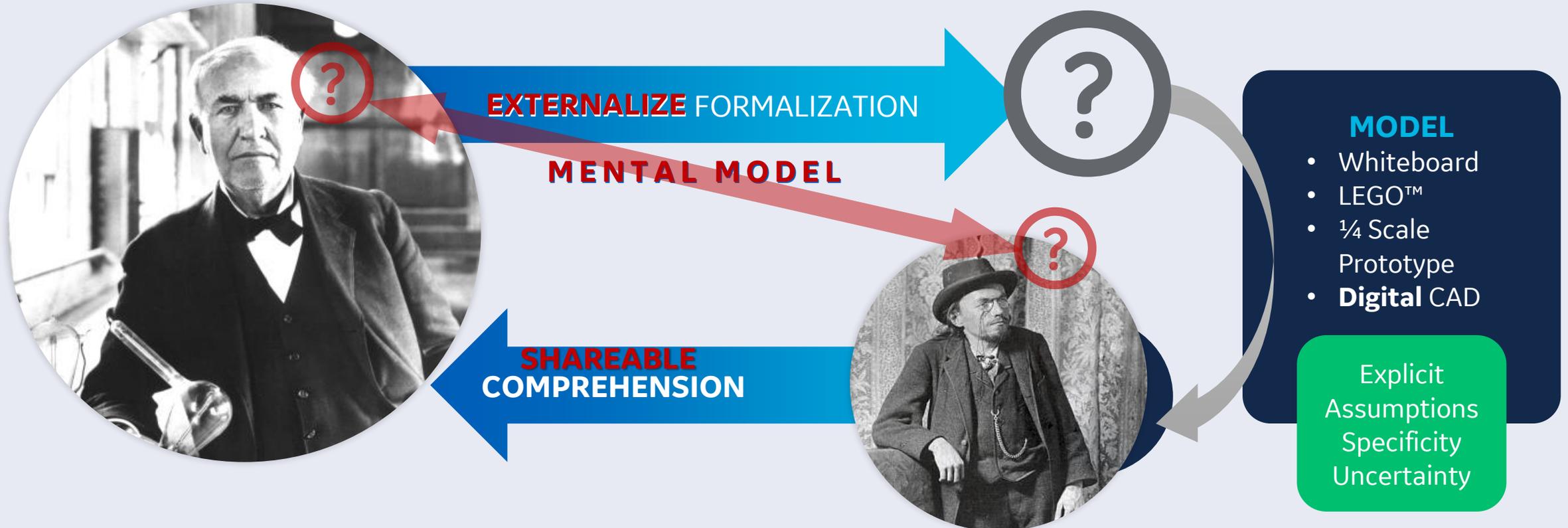
<https://blogs.ge.com/problem-solving>

Team,

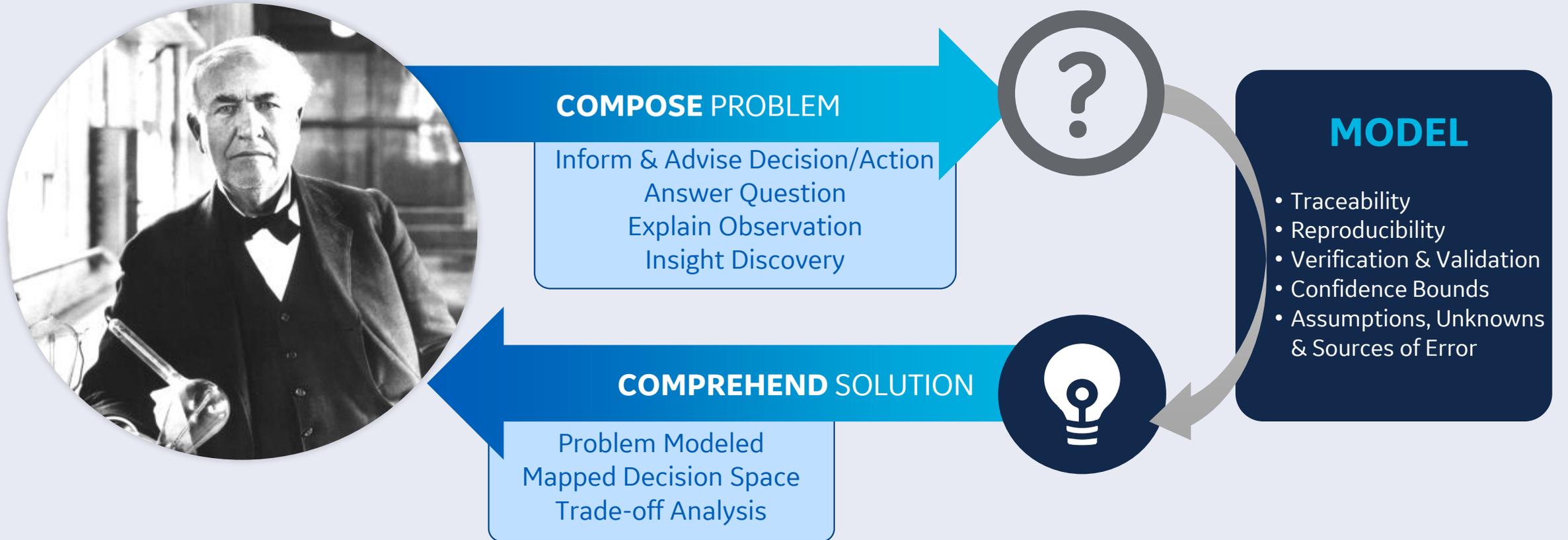
One of my personal resolutions for 2020 is to teach more about lean as we drive our transformation. I will work to do that more often this year in my emails to you. For today's note I'm starting with problem solving because it is foundational to developing competitive advantage, delivering for our customers and ultimately improving our performance over the long-term.

**Problem Solving** critically relies upon **Modeling**  
(the problem, the solution and the process in between)

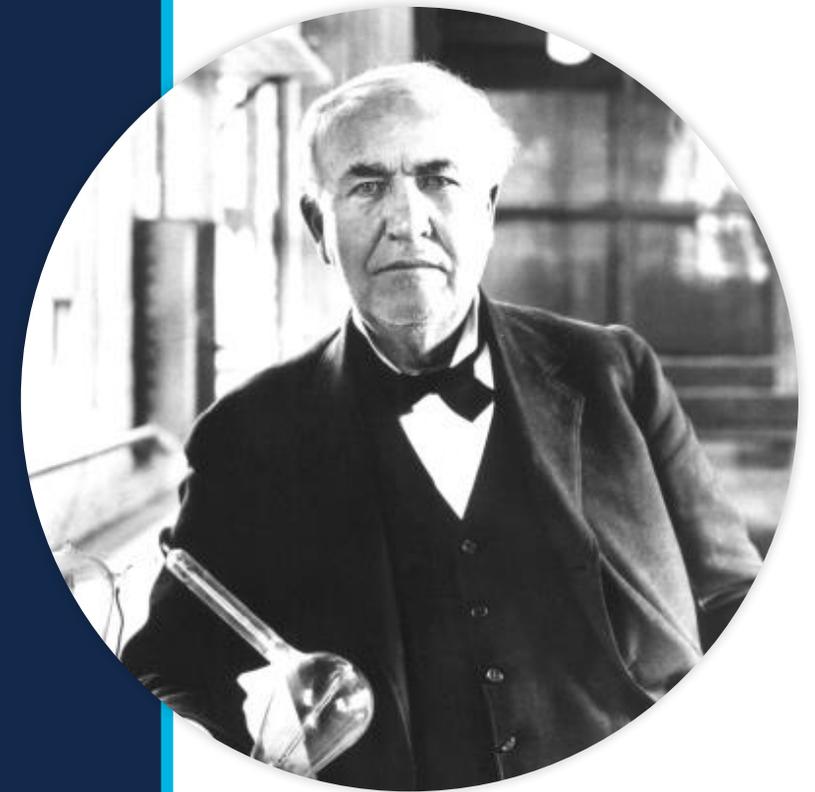
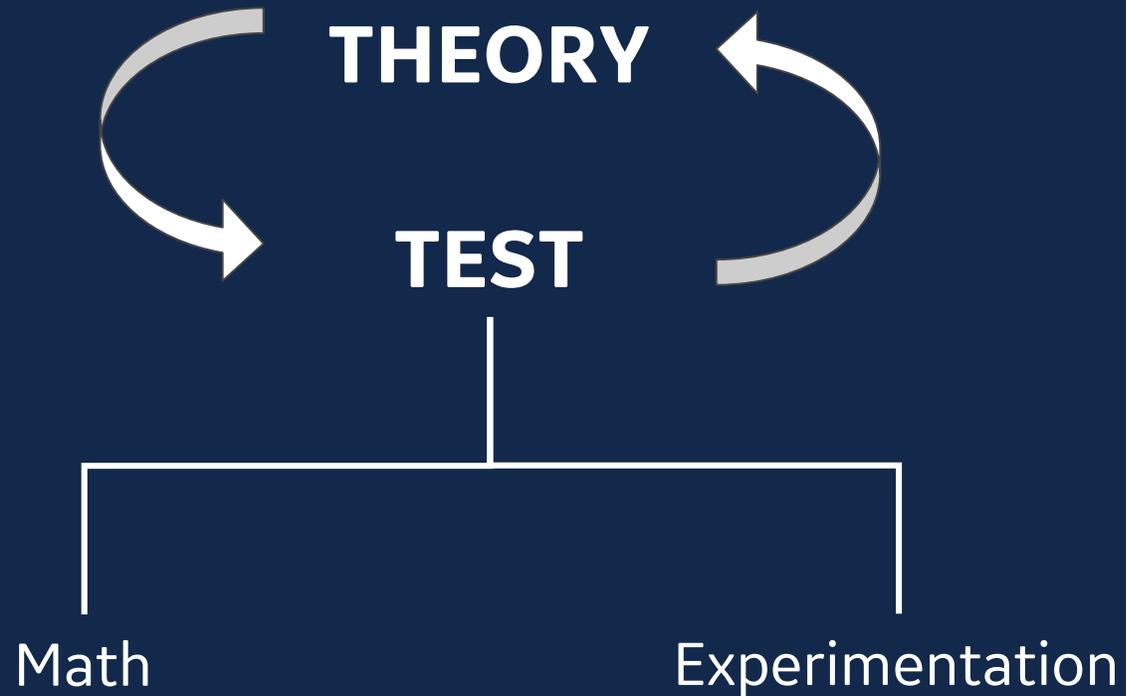
# All decisions and actions employ models



# All decisions and actions employ models



# Scientific method



**EDISONIAN**

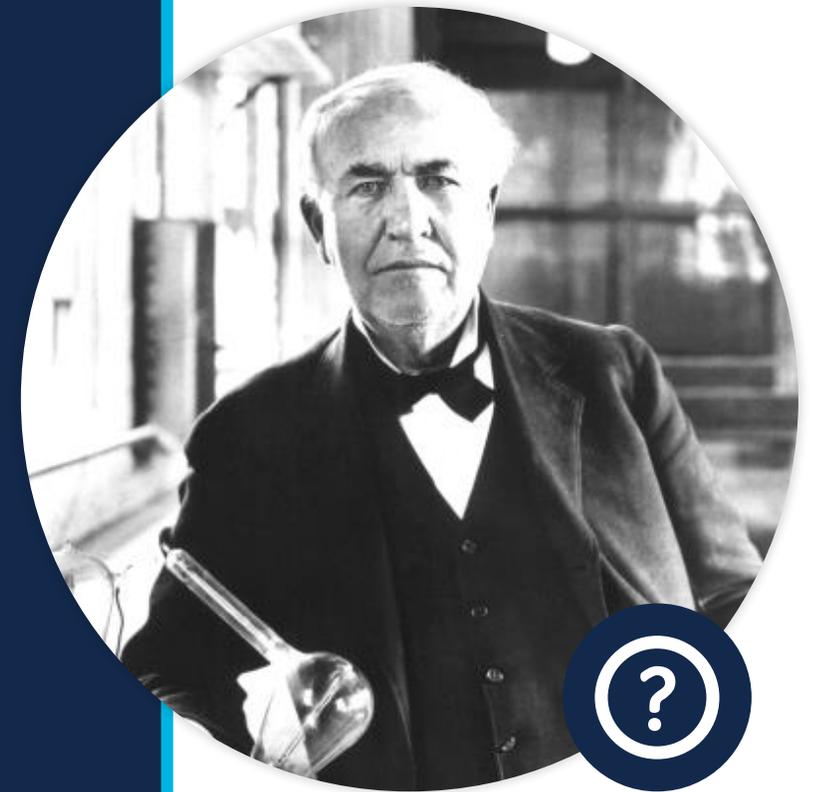
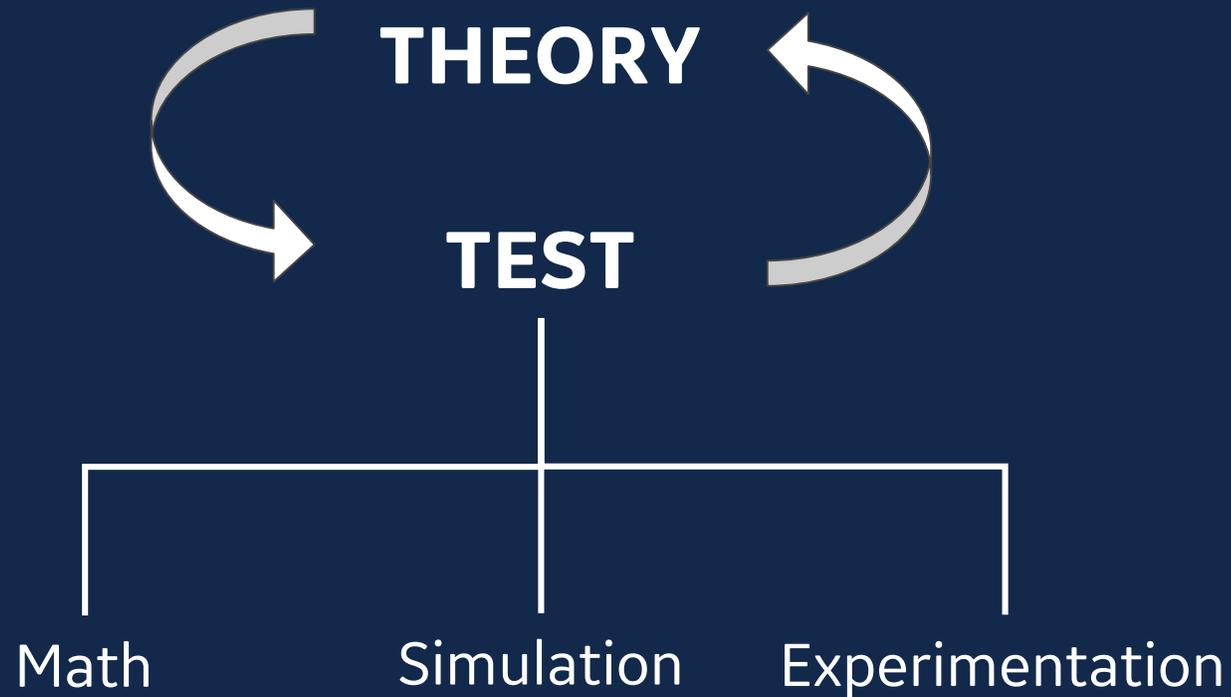
\$\$\$\$ Testing & Measurement

THEORETICAL  
MODEL

EMPIRICAL  
EXPERIMENTAL  
OBSERVATION



# Modern practice of scientific method



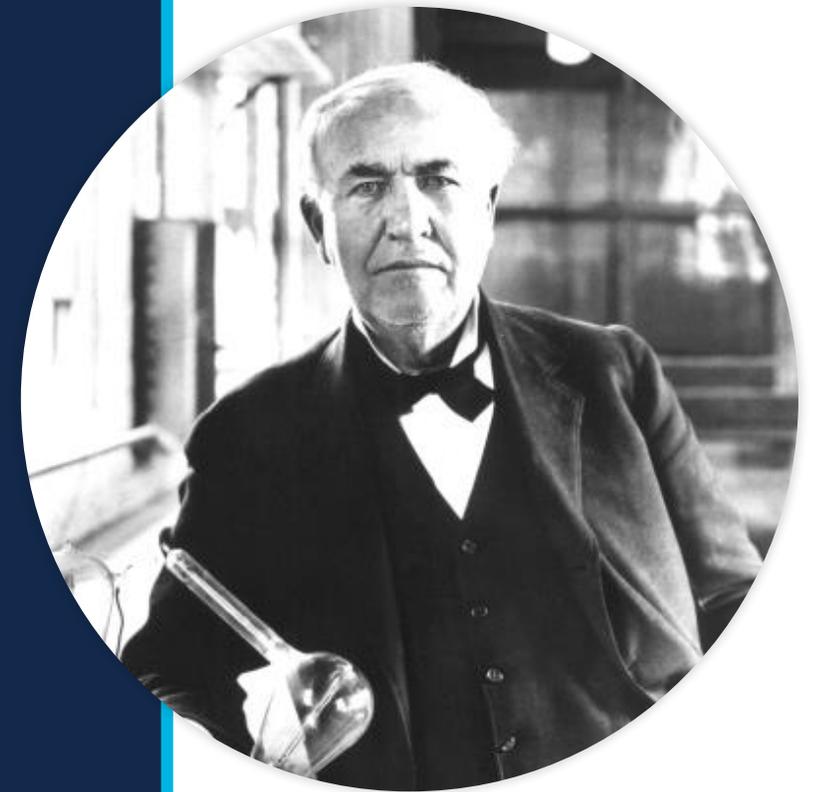
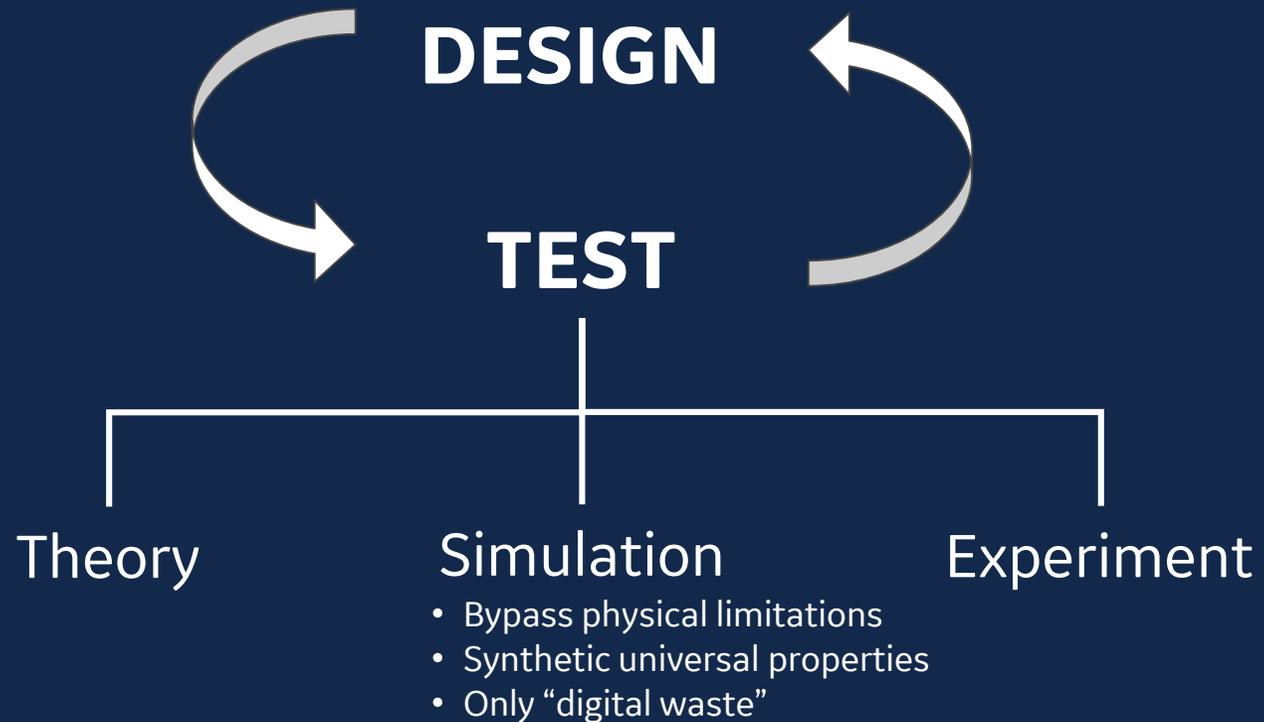
MATH

SIMULATION

MEASUREMENT



# Advancing the Scientific Method



**COMPUTATIONAL  
MODEL?  
I'M SKEPTICAL!**

THEORY

SIMULATION

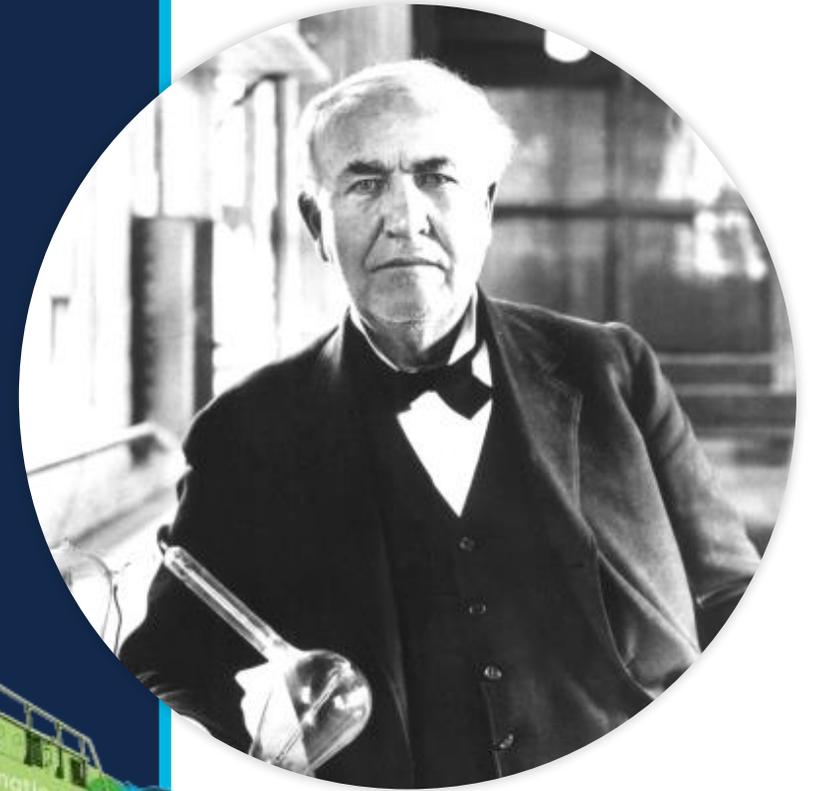
EXPERIMENT  
**\$100's of MM**  
PER YEAR AT GE



# CRITICAL INFRASTRUCTURE



*Justified Model Skepticism: Garbage IN/Garbage OUT*



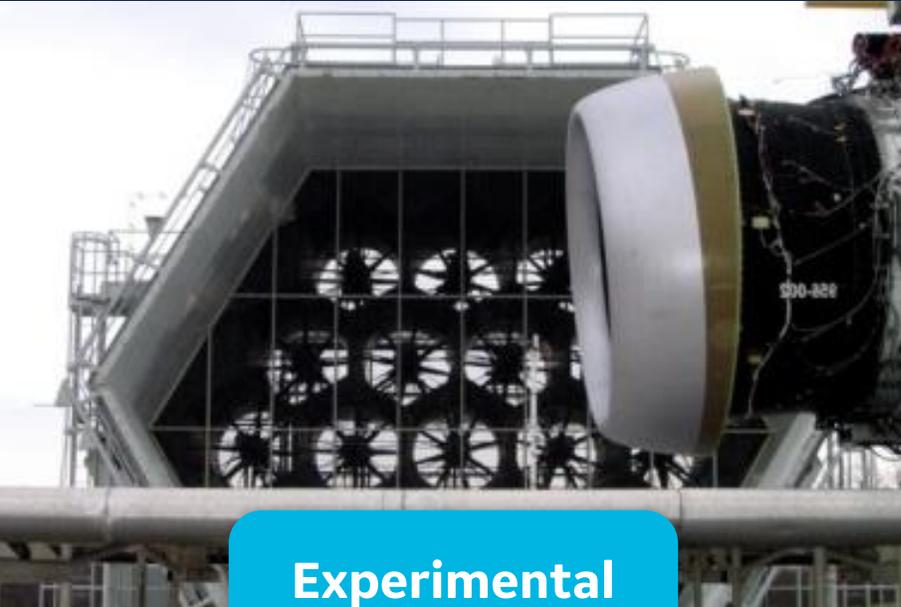
## GE Legacy Products:

- are Highly Complex &
- High Value Assets,
- have Safety-Critical Components,
- High-Consequence Downtime,
- and Long Field Life



# Physical validation is critical

“RIG” TEST



**Experimental  
Measurement**

DIGITAL TWIN



**Targeted Field  
Sampling**



**VERIFICATION &  
VALIDATION**

**CALIBRATION &  
UNCERTAINTY  
QUANTIFICATION**



# Engineering's core competency is Problem Solving



<https://blogs.ge.com/problem-solving>

Team,

One of my personal resolutions for 2020 is to teach more about lean as we drive our transformation. I will work to do that more often this year in my emails to you. For today's note I'm starting with problem solving because it is foundational to developing competitive advantage, delivering for our customers and ultimately improving our performance over the long-term.

**Problem Solving** critically relies upon **Modeling**  
(the problem, the solution and the process in between)

**Modeling** critically relies upon **Computational Methods**  
(as an engine for scale, productivity, consistency and capability)

# Computational model as scientific instrument



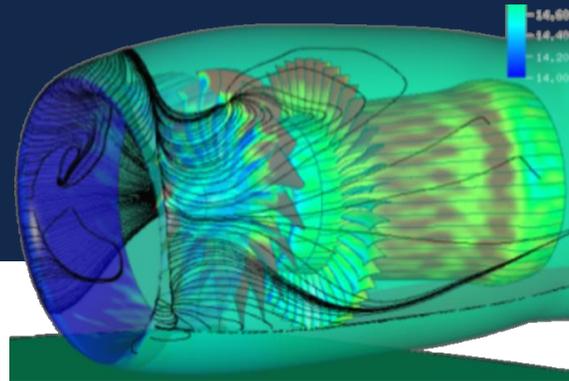
## MICROSCOPE

Interrogate extreme detail



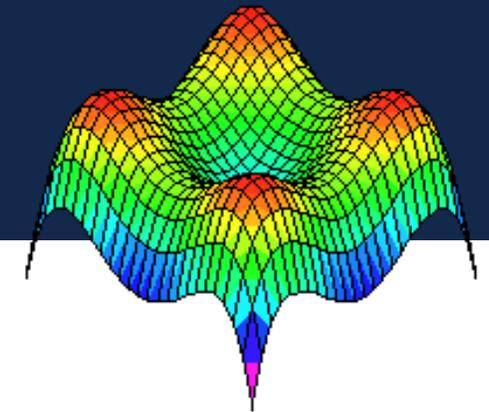
## MACROSCOPE

Perceive system-wide interactions



## CAMPAIGN

Explore vast alternatives



# Modeling & Simulation (Mod/Sim) well-established at GE / GE Research



Gas  
Turbines



CT  
scanners



Engines



Wind  
turbines



Oil &  
Gas



Additive

Computational  
Science & Engineering

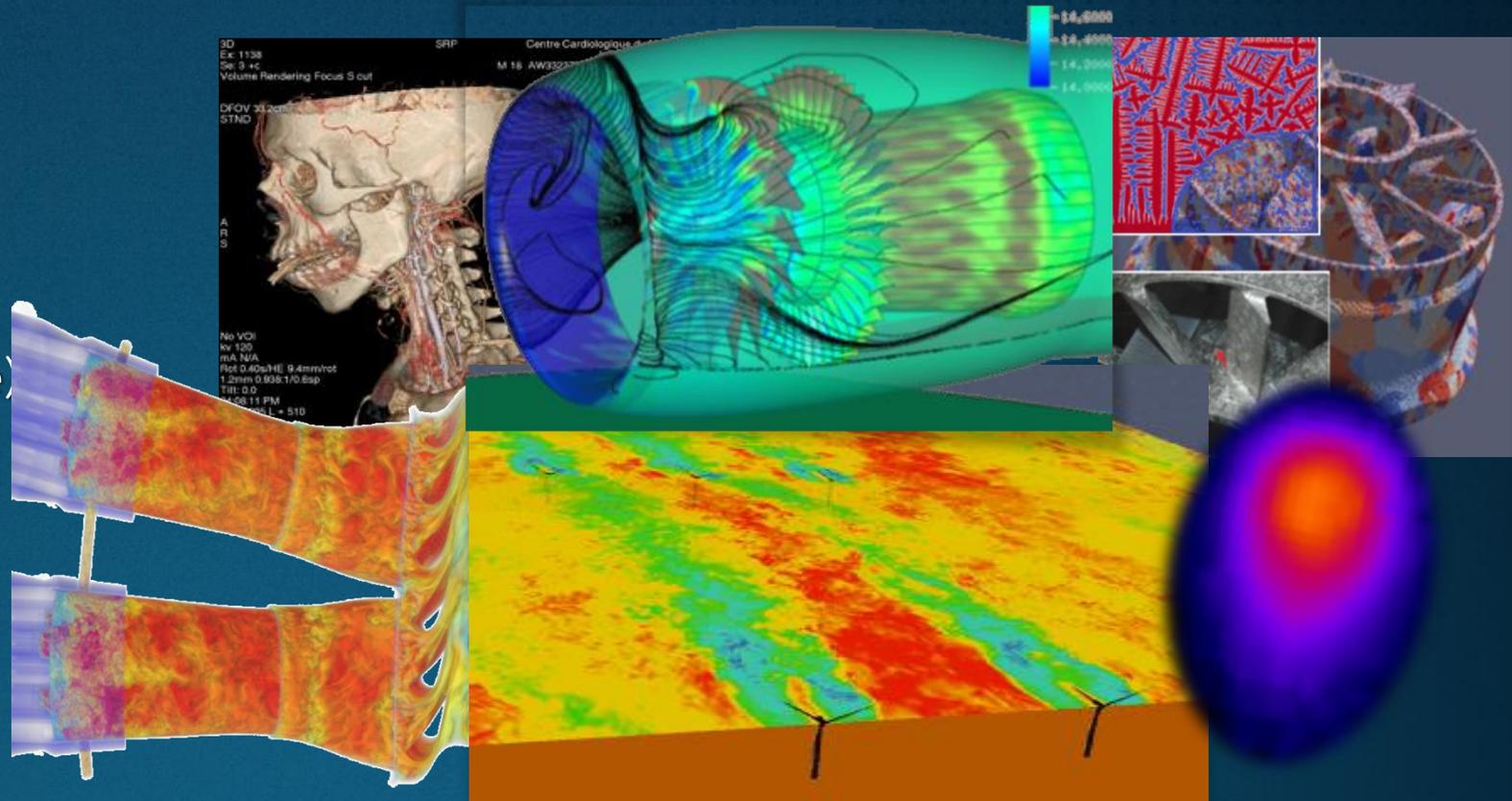


## Modeling (Form & Function)

- Geometric Dimensioning & Tolerancing
- Physical Model & Material Properties
- Environment & Operation Conditions

## Simulation (Credibility & Confidence)

- Verification & Validation
- Repeatability & Empirical Calibration
- Numerical Assumptions & Effects
- Uncertainty Quantification
- Estimation & Assumption Propagation
- Parametric Sensitivity Analyses



# CRITICAL SUPPORT INFRASTRUCTURE FOR MODELING

SCIENCE & ENGINEERING  
INTRUMENTATION



Experimental Measurement

Targeted Field Sampling

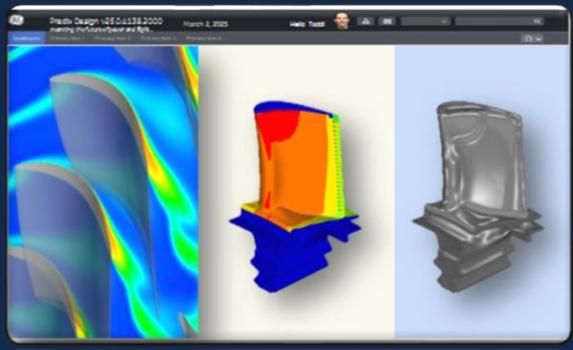
TECHNICAL & REGULATORY  
PROCESSES



COMPUTATIONAL HARDWARE



SIMULATION & ANALYSIS ECOSYSTEM

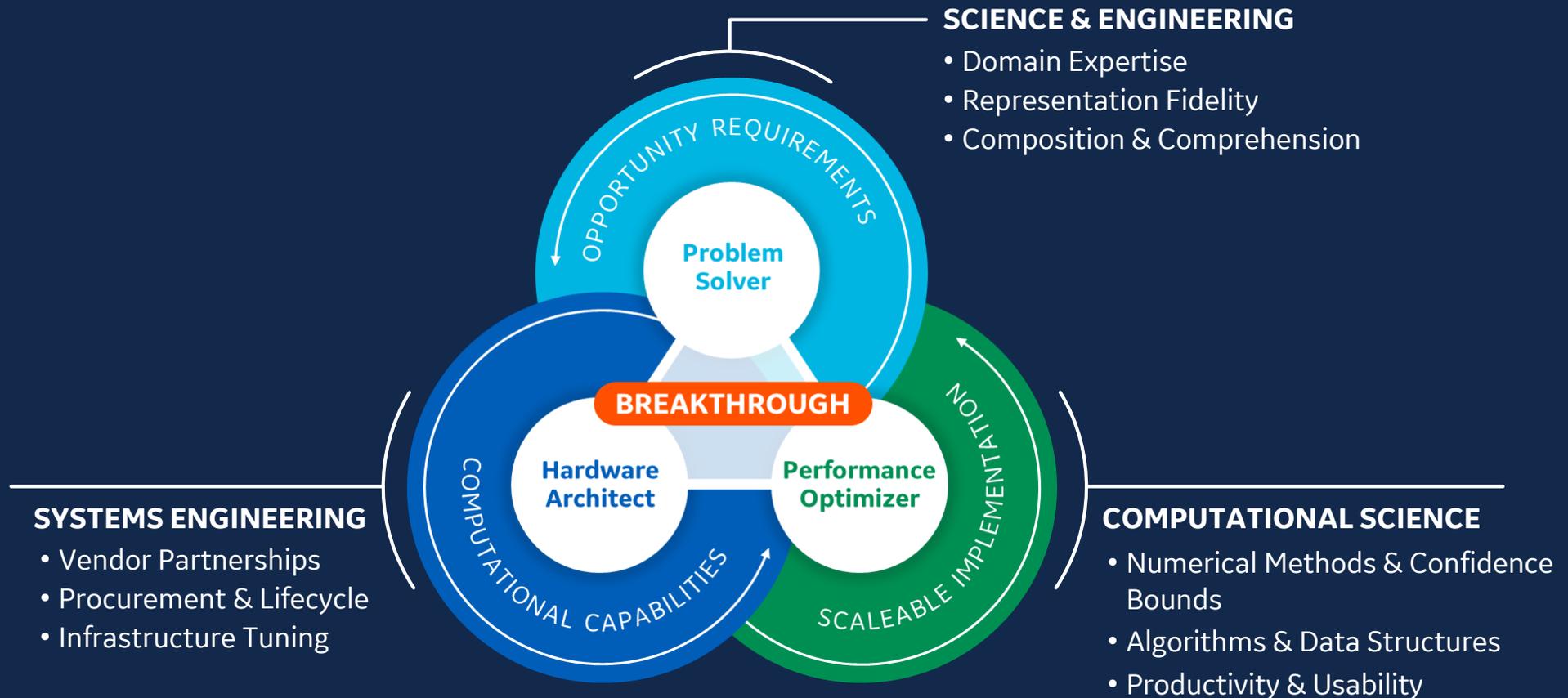


## MODEL

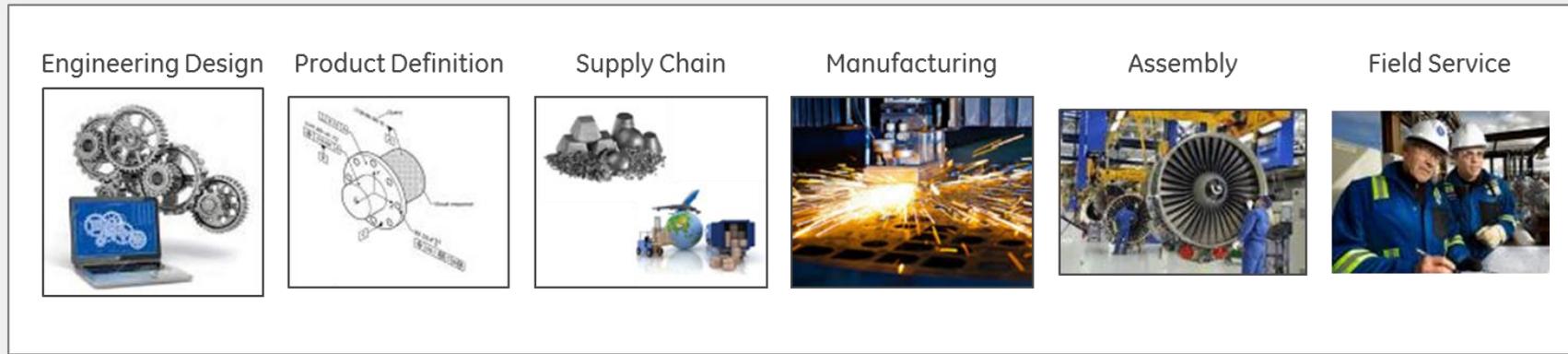
- Traceability
- Reproducibility
- Verification & Validation
- Confidence Bounds
- Assumptions, Unknowns & Sources of Error



# Co-Design: A Fugue of Expertise toward Breakthrough Impact

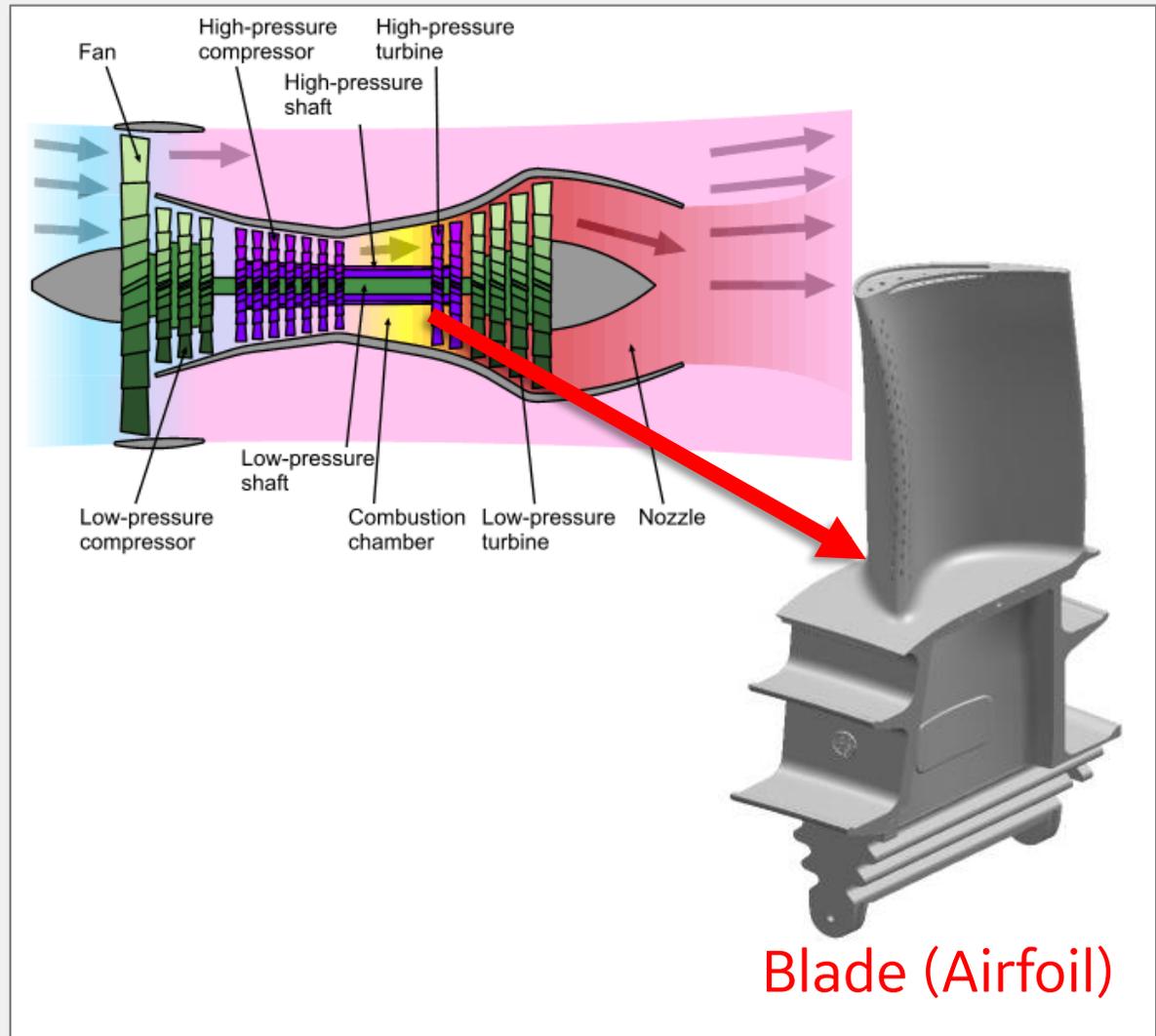


# Product Lifecycle



**Conceive → Design → Make → Deliver → Use → Maintain**

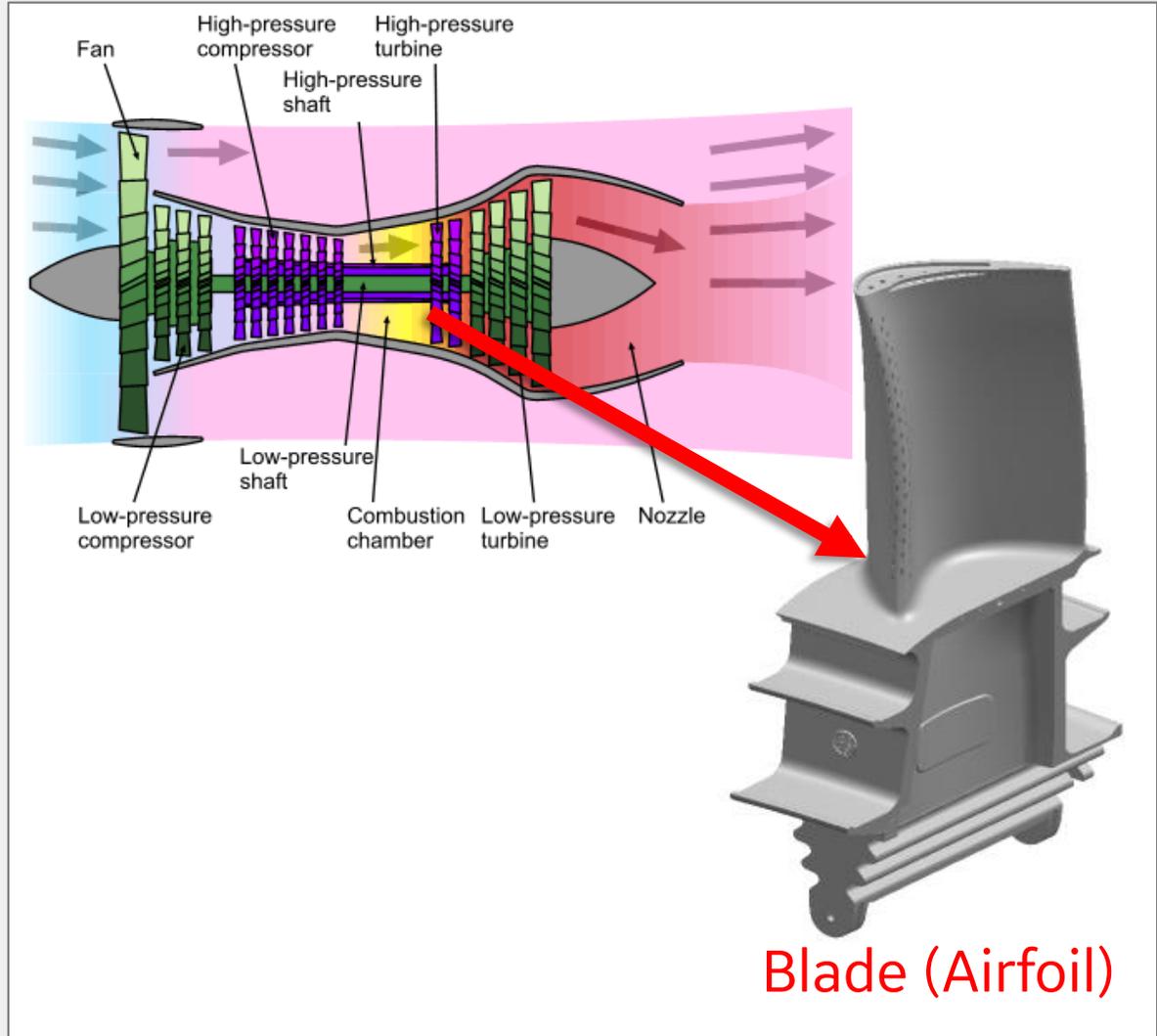
# Brief overview on a core engineering practice at GE: Design of Blades in Turbomachinery



## Objectives:

1. Competitive performance of product (*efficiency, power, ...*)
2. Reliability of product (*durability, safety, robust operation*)
3. Cost to design product (*engineering, testing, certification*)
4. Cost to make product (*materials, manufacture, assembly*)
5. Operational cost of product (*including maintenance & repair*)

# Brief overview on a core engineering practice at GE: **Design of Blades in Turbomachinery**



## Objectives:

1. Performance
2. Reliability
3. Design Cost
4. Manufacturing Cost
5. Operating/Service Cost

## Ideally:

- ↑ Drives Sales (Value for \$)
- ↑ Certification on 1<sup>st</sup> Test
- ↓ No rework, built to spec+
- ↓ No waste, rework, inventory
- ↓ Value for \$ for customer + GE

# Turbine Airfoil Design (circa 1980)

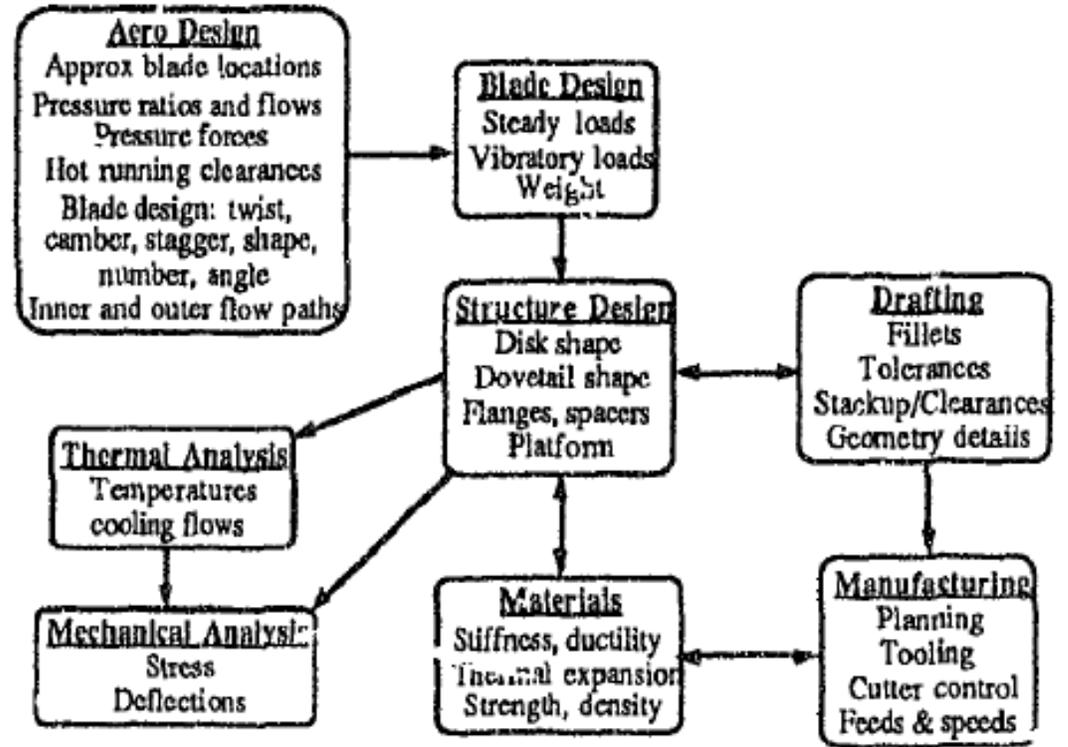
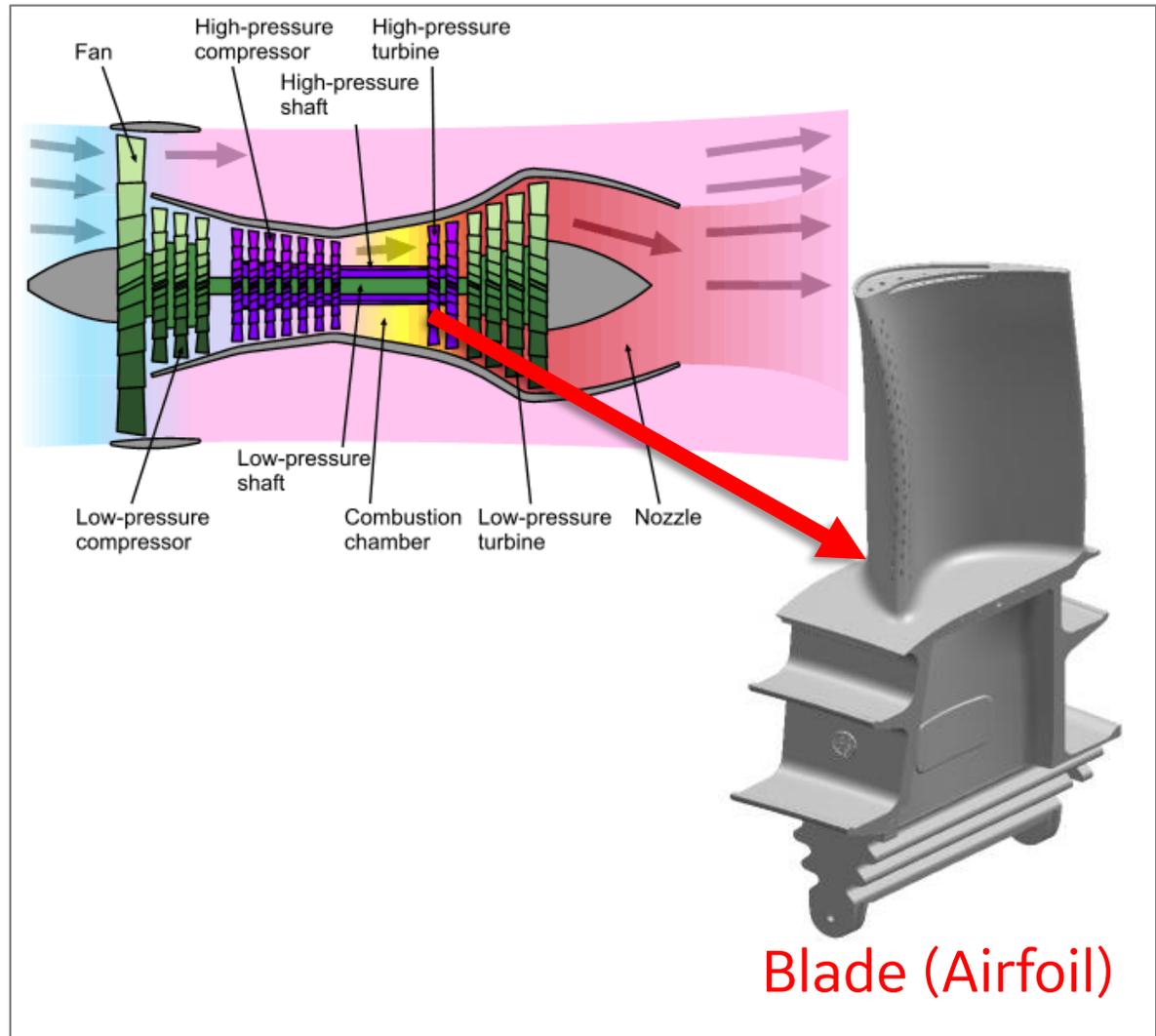
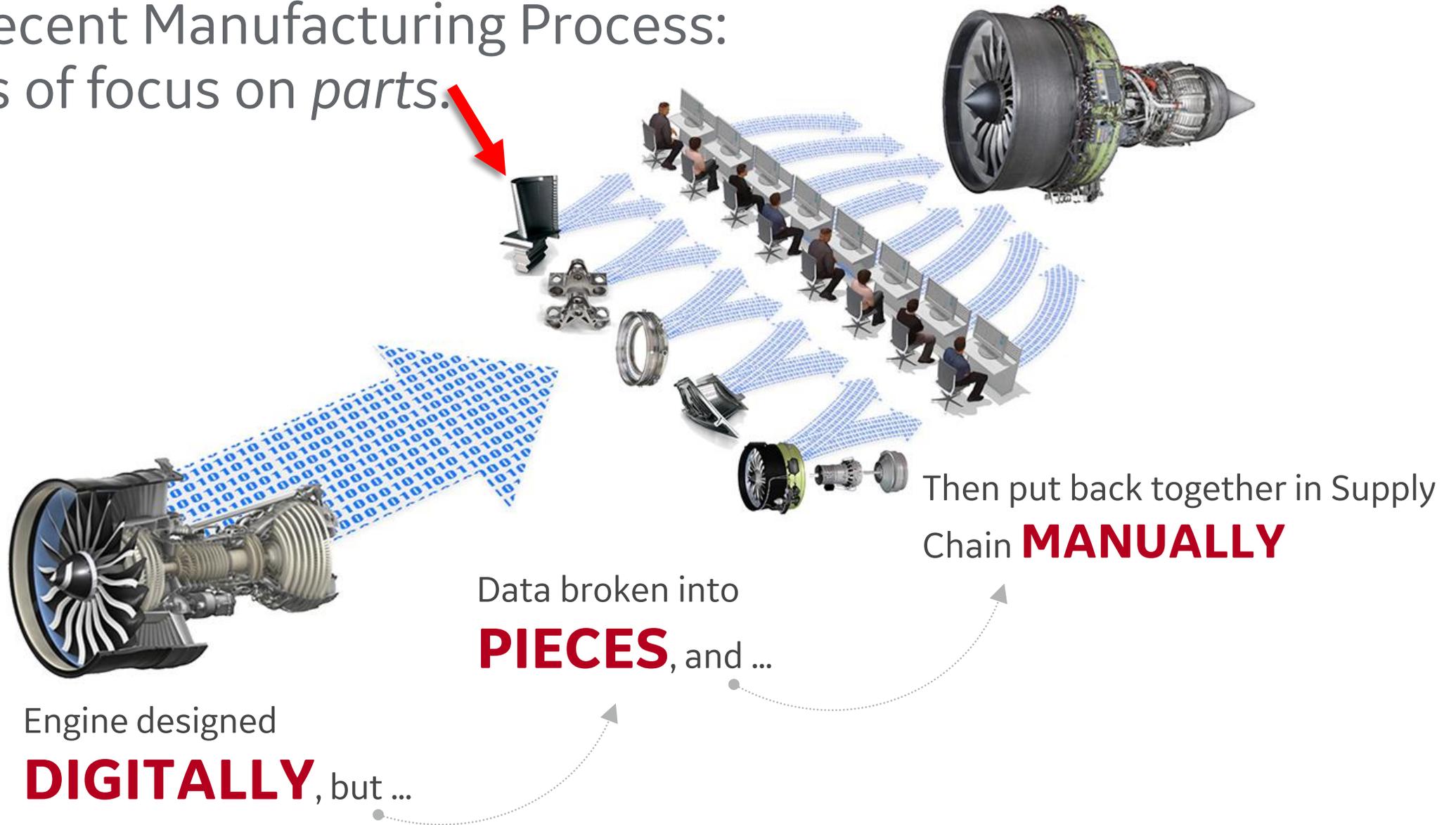


Figure 2. Informal turbine blade design data flow.



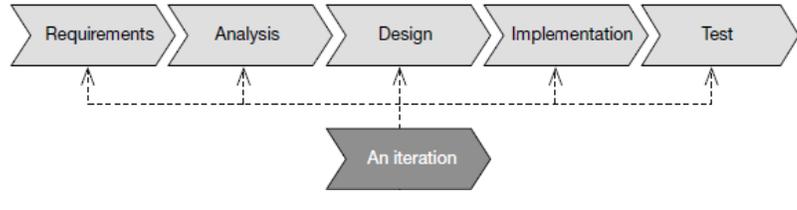
More Recent Manufacturing Process:  
Still lots of focus on *parts*.



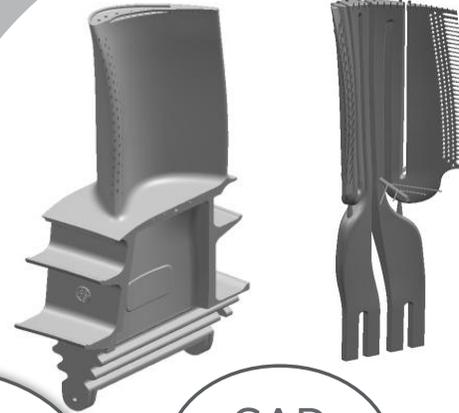
And for part design:  
*discipline-by-discipline analyses*

# Design Engineering

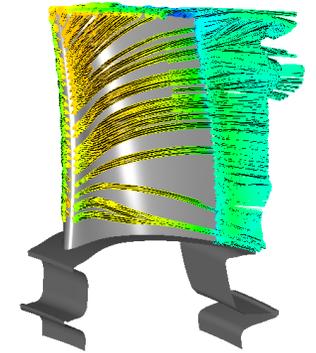
Each technical stakeholder goes through:



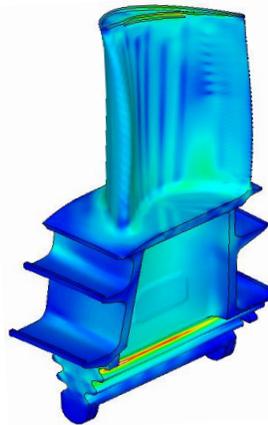
Internal + external design



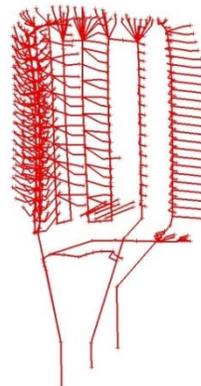
Aerodynamics



Fluid flow & acoustics

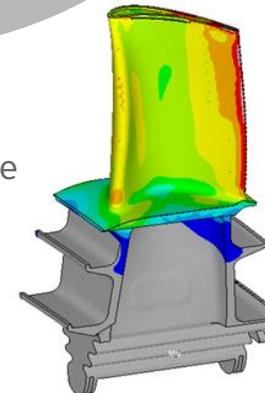


Stress



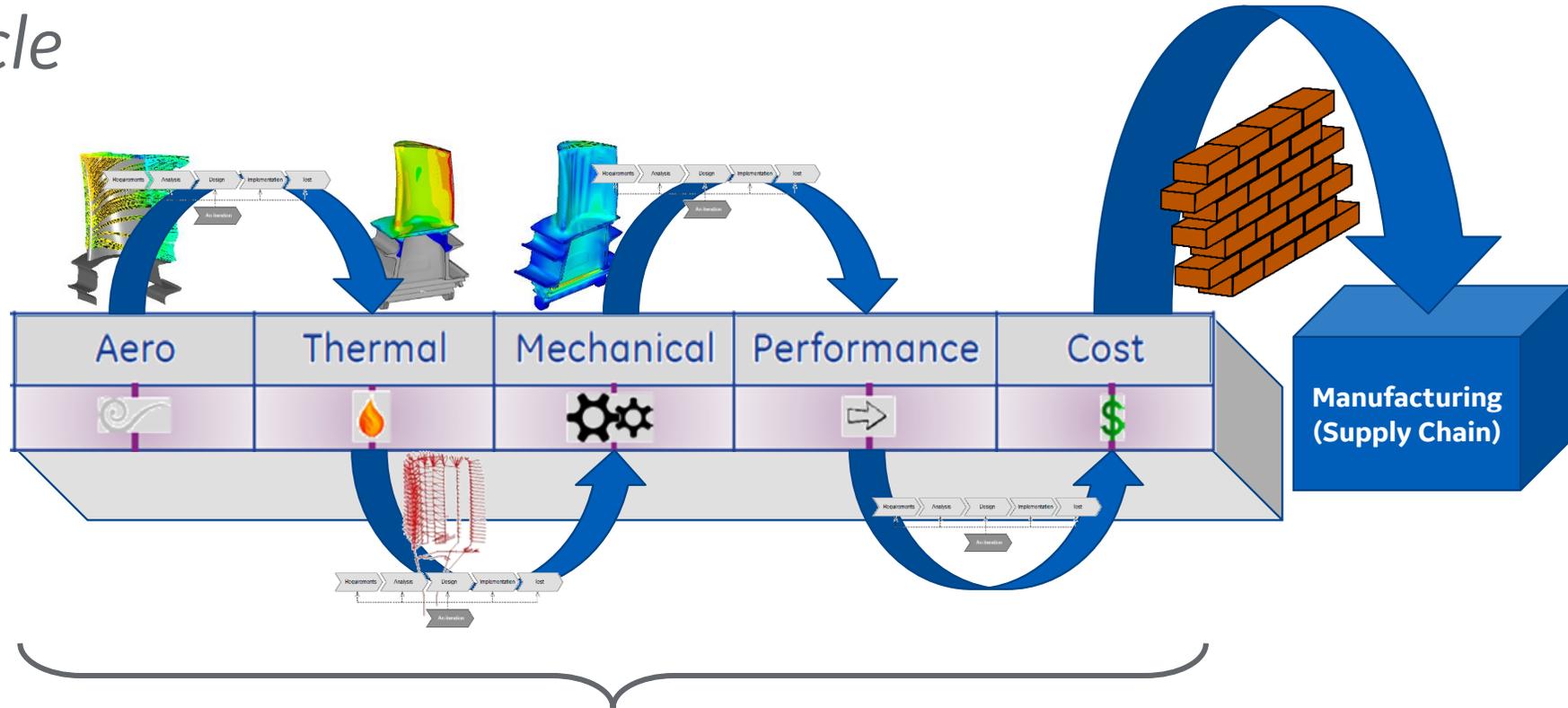
Internal flow model

Metal Temperature



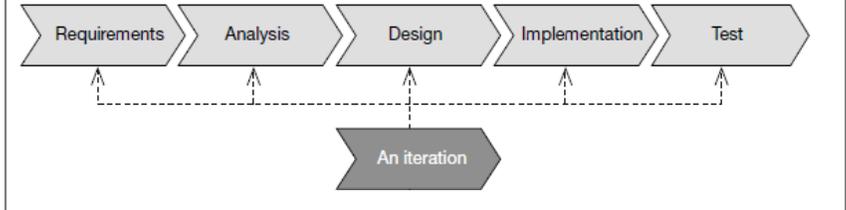
# Iterative Design Cycle

## Design Engineering



Design hand-offs across disciplines of expertise

Each technical stakeholder goes through:



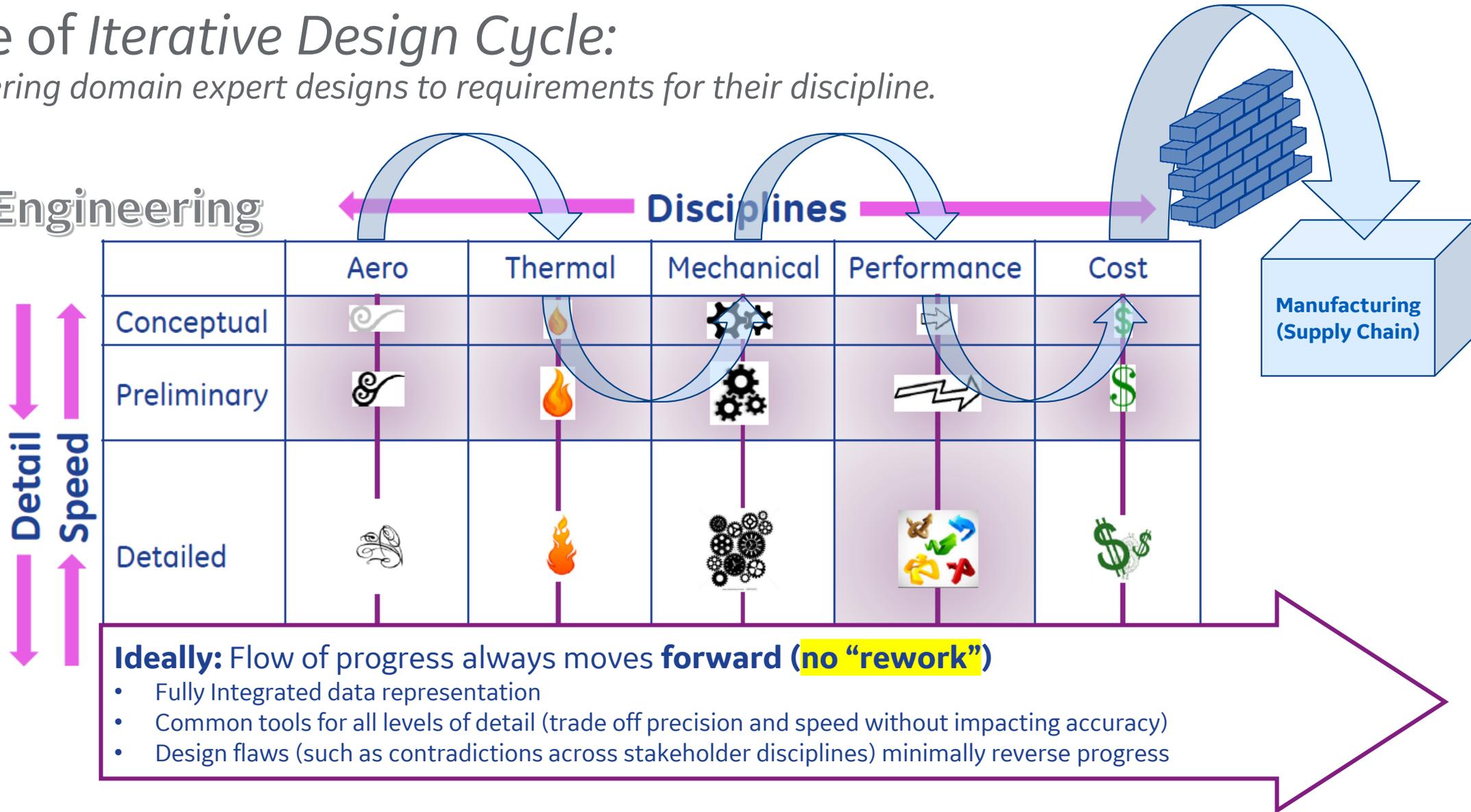
“Final Design” handoff to Manufacturing



# Example of *Iterative Design Cycle*:

Each engineering domain expert designs to requirements for their discipline.

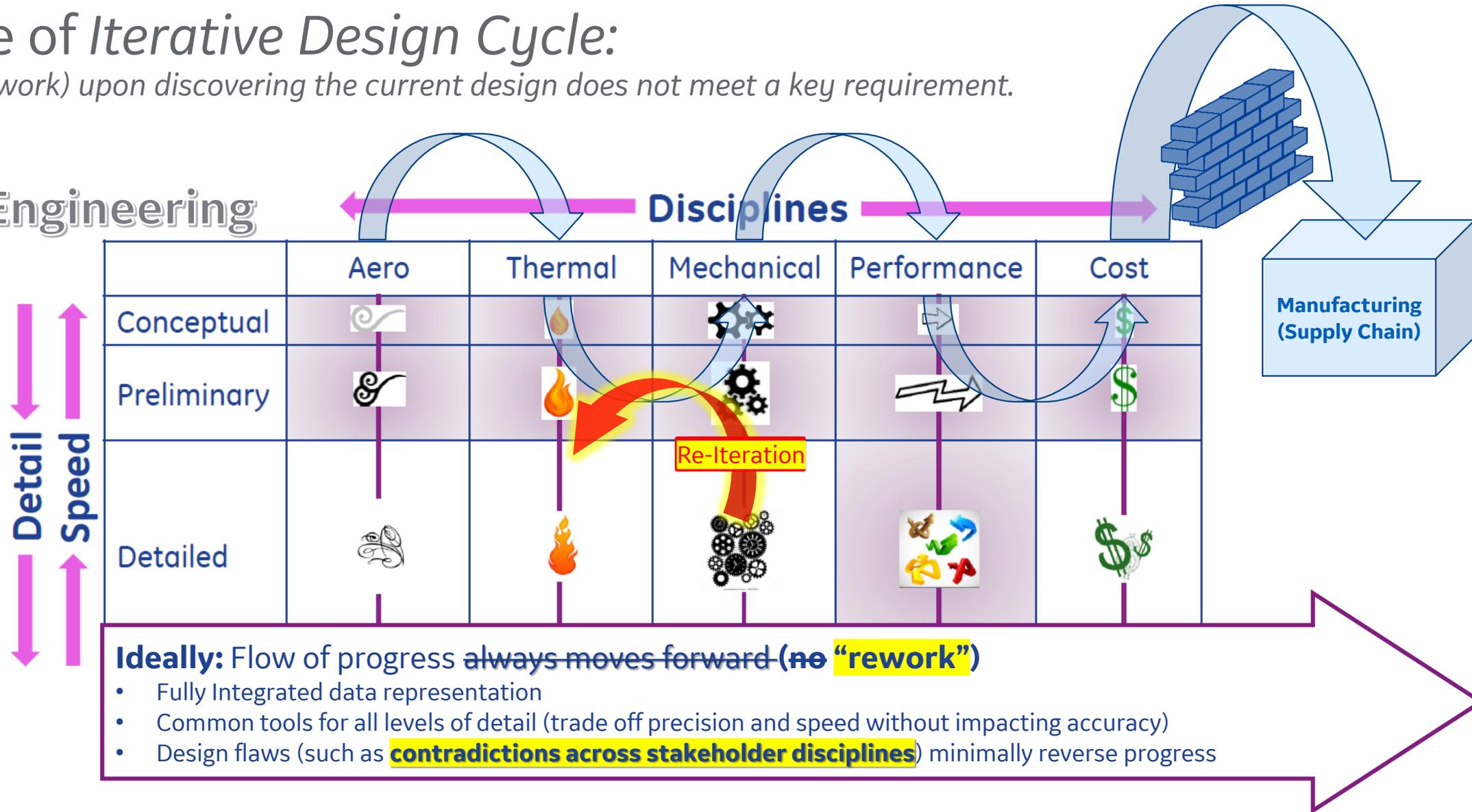
## Design Engineering



# Example of *Iterative Design Cycle*:

Re-design (re-work) upon discovering the current design does not meet a key requirement.

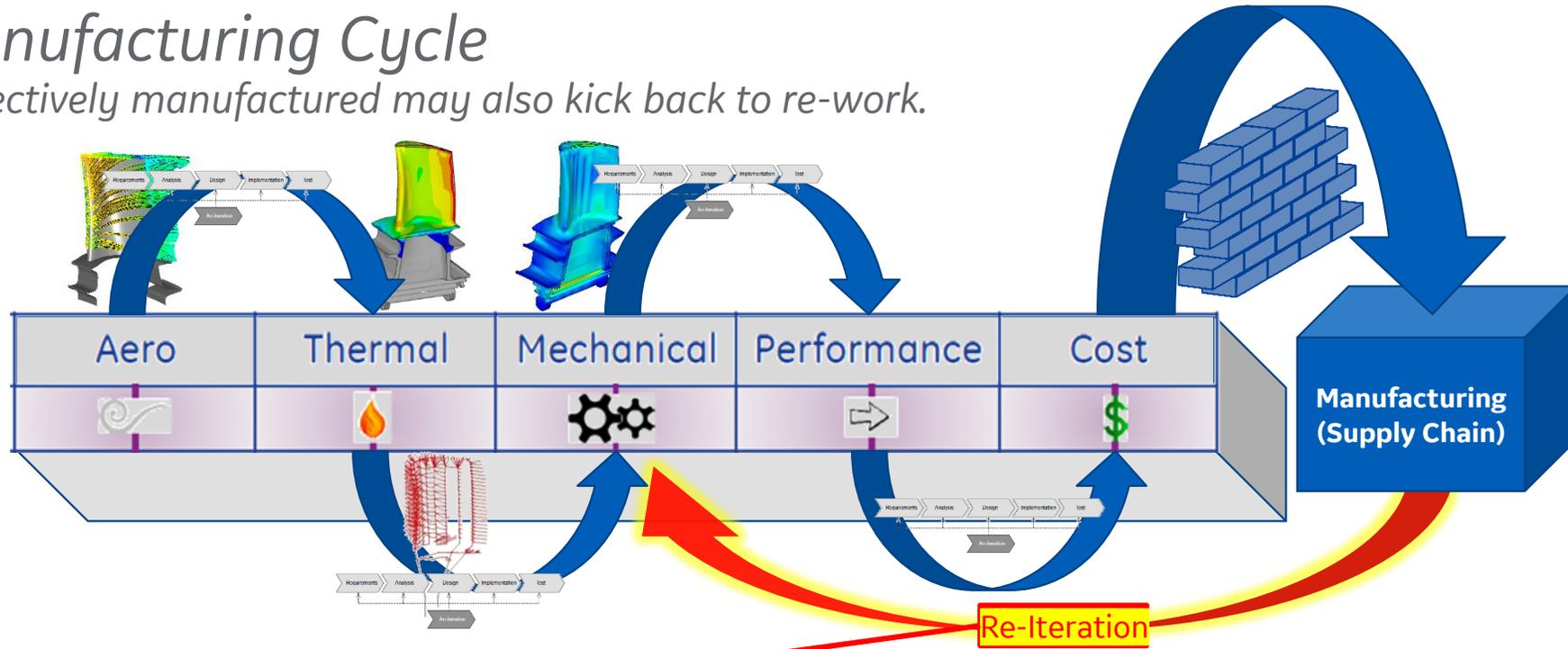
## Design Engineering



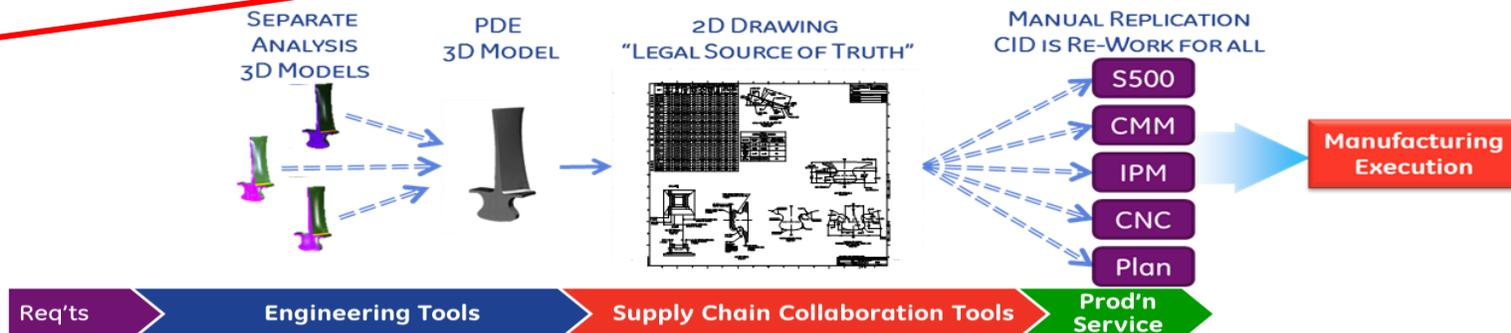
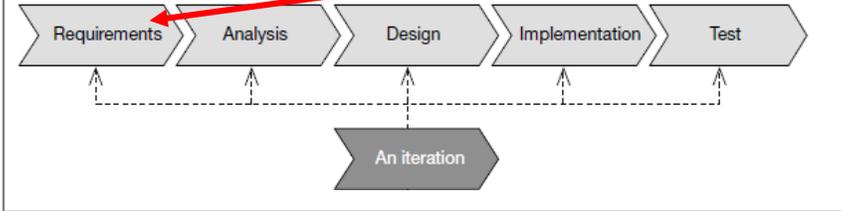
# Iterative Design-to-Manufacturing Cycle

Parts not feasibly/reliably/cost-effectively manufactured may also kick back to re-work.

## Design Engineering



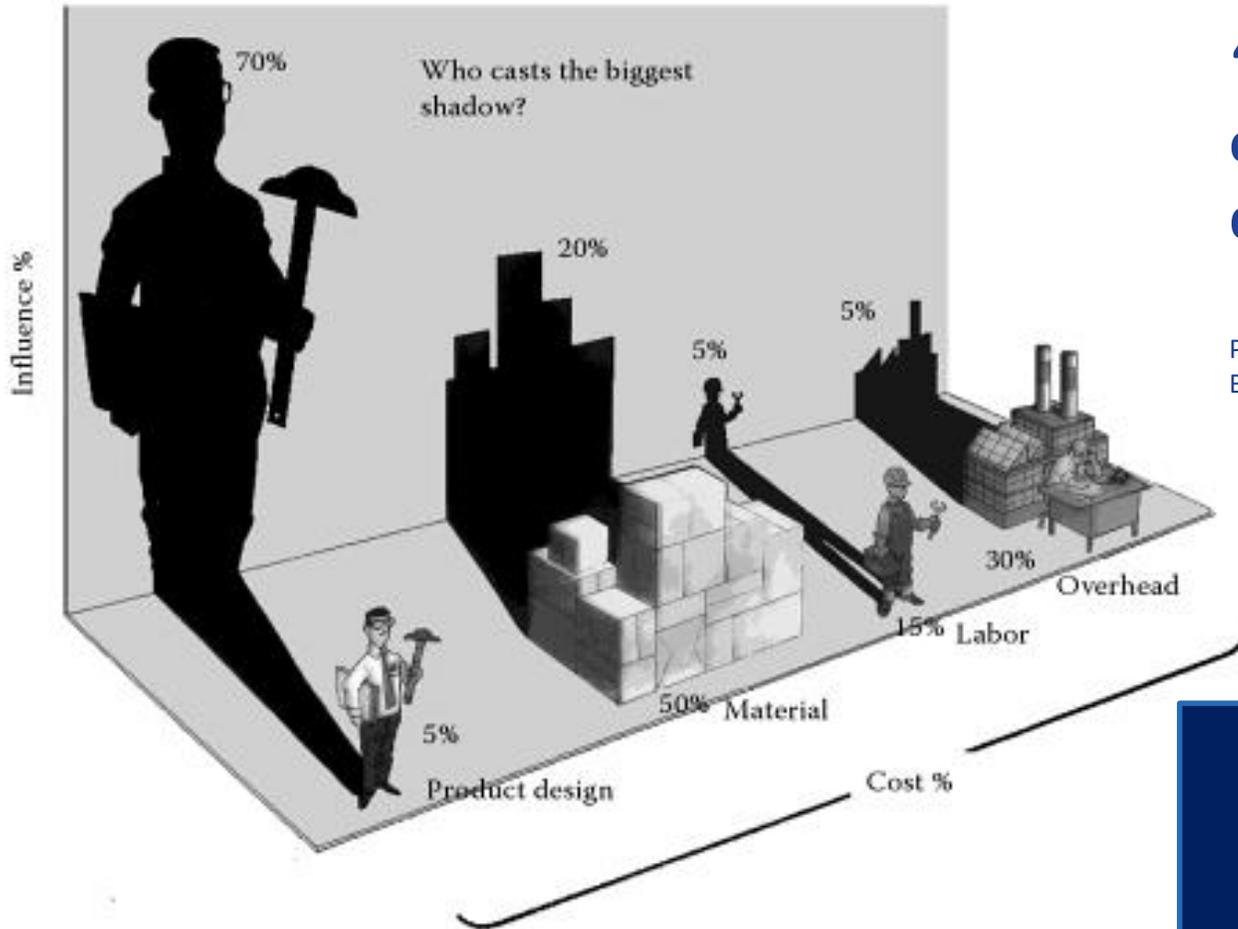
Each technical stakeholder goes through:



# Digital Thread & “The Shadow of Design”

“it is now widely accepted that **over 70% of the final product costs are determined during design**”

Product Design for Manufacture and Assembly, Third Edition  
By G. Boothroyd, P. Dewhurst, V. A. Knight



**Drive Variable Cost Productivity (VCP)**  
by  
**pushing cost visibility into Design**

FIGURE 1.5  
Who casts the biggest shadow? (Adapted from Munro and Associates, Inc.)



GE Proprietary Information  
For Internal Use Only

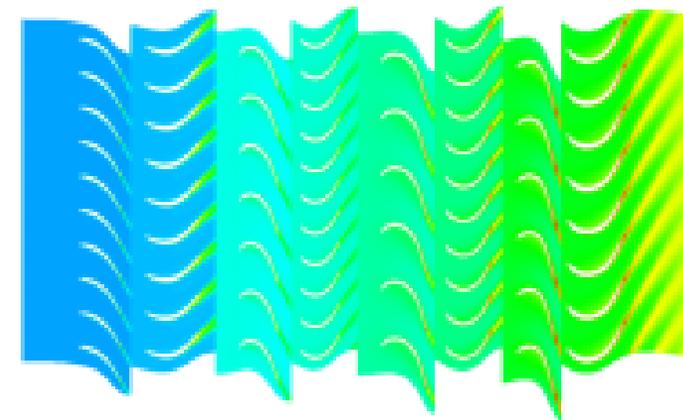
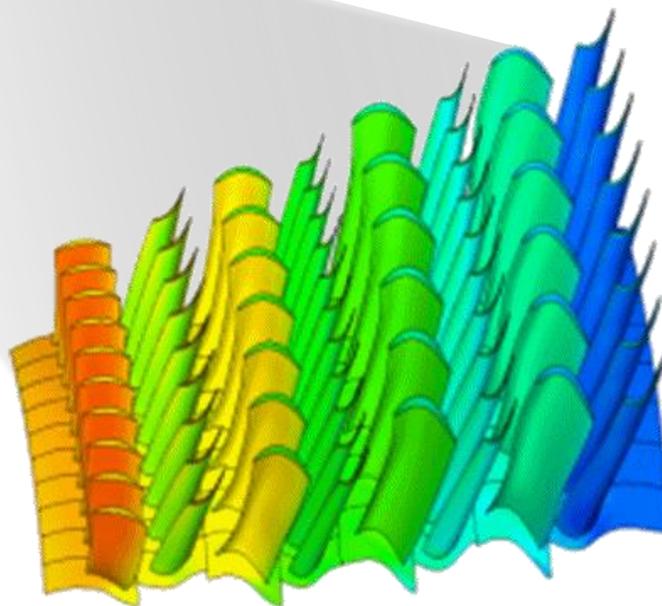
GE Impact leveraging Leadership Computing

# High-end modeling - Moment of truth

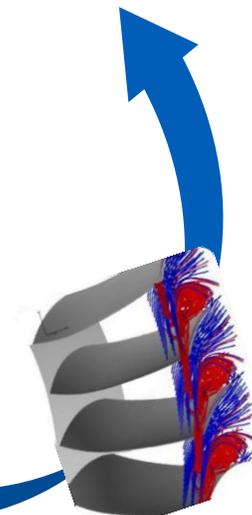


**Will it be useful?**

**Will we see something different?**



**Compare:**  
Best Internal Modeling Capability

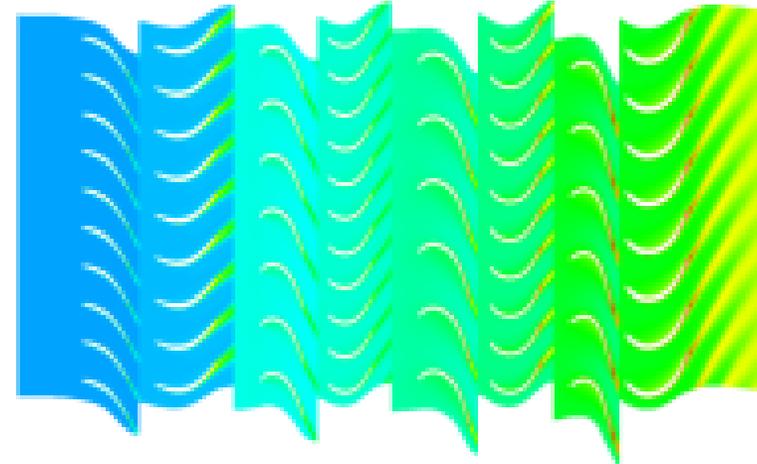


GE Aviation LEAP  
Unsteady CFD: Strut wake effects  
GE Tacoma RANS solver



# Never before seen

- Unobservable physically
- Relevant to engineering design
- 2012 IDC award:

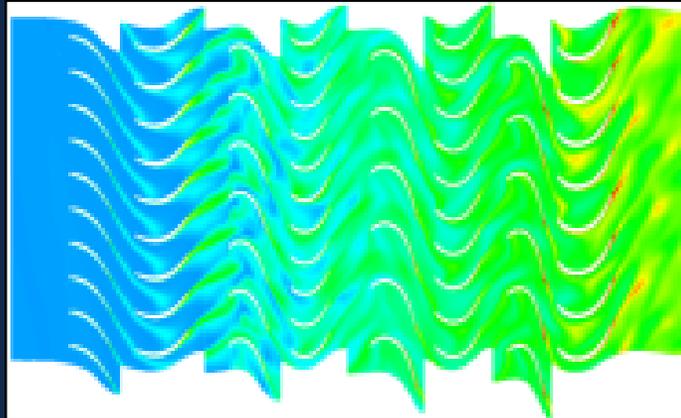


**Prior State of the Art:**  
Steady Analysis  
(GE Internal HPC)

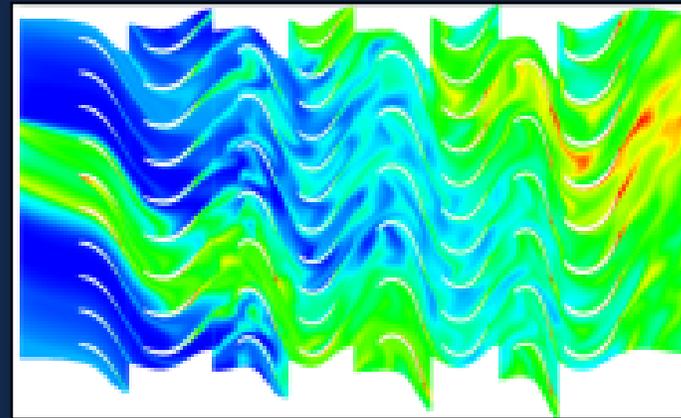


On ORNL Jaguar Cray XT5 (2010)

**Preliminary Result**  
Unsteady Analysis  
(with Uniform Inlet)



**Final Result**  
Unsteady Analysis  
(with wake from strut)

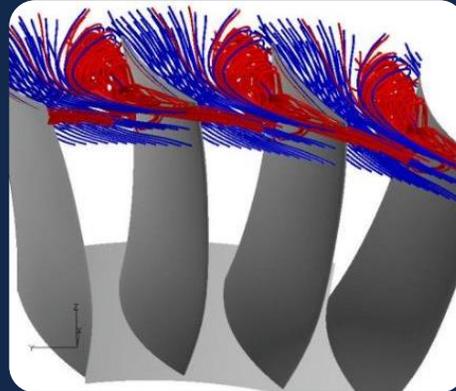


# Software: Simulation and Modeling

## Dynamic Flows

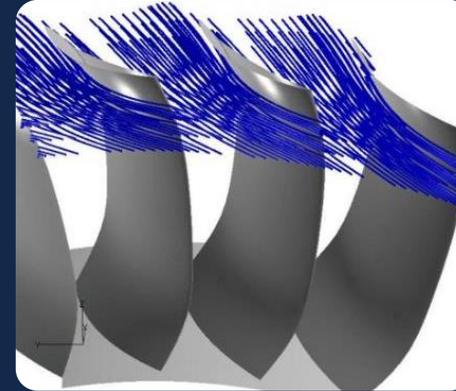
- Heat
- Acoustics
- Fuel Burn
- Emissions

Separated flow - poor air flow control and loss of efficiency



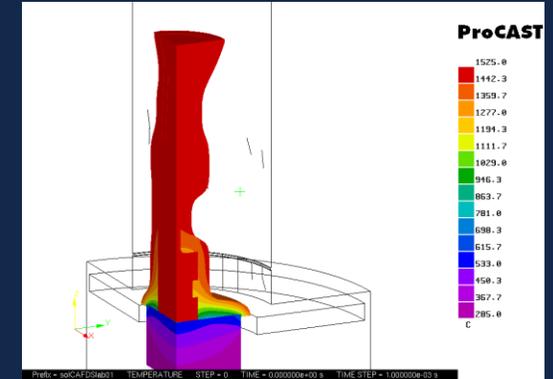
Old Design "Straight Airfoil"

Attached flow - good air flow control and high efficiency



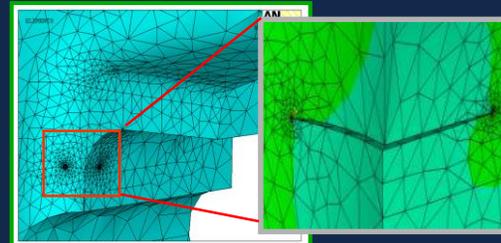
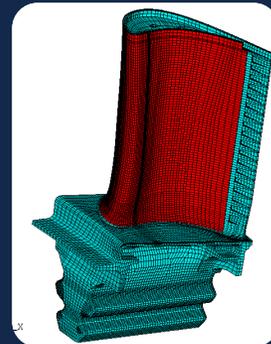
New Design Using "Bowed Airfoil"

## Material Design

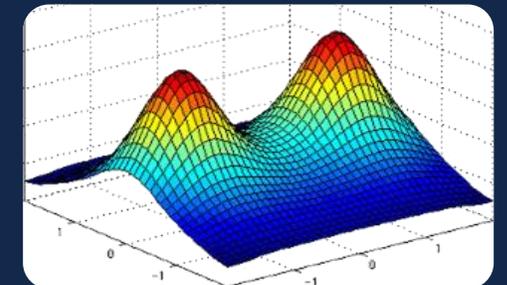


## Mechanical Properties

- Metal alloy wear and tear
- "Crash tests" (bird strike)



## Design Optimization



# Computational Methods Maturity

a.k.a. Virtual / Digital / Computational / Numerical Modeling

## Criticality of Computational Modeling

- All problem-solving employs *models* (even if merely mental models of the person solving).
- Some problems necessarily must be modeled *computationally* due to factors such as physical test expense/difficulty/safety, turnaround time, legality/ethics, and/or measurement limitations.

## Critique of Modeling

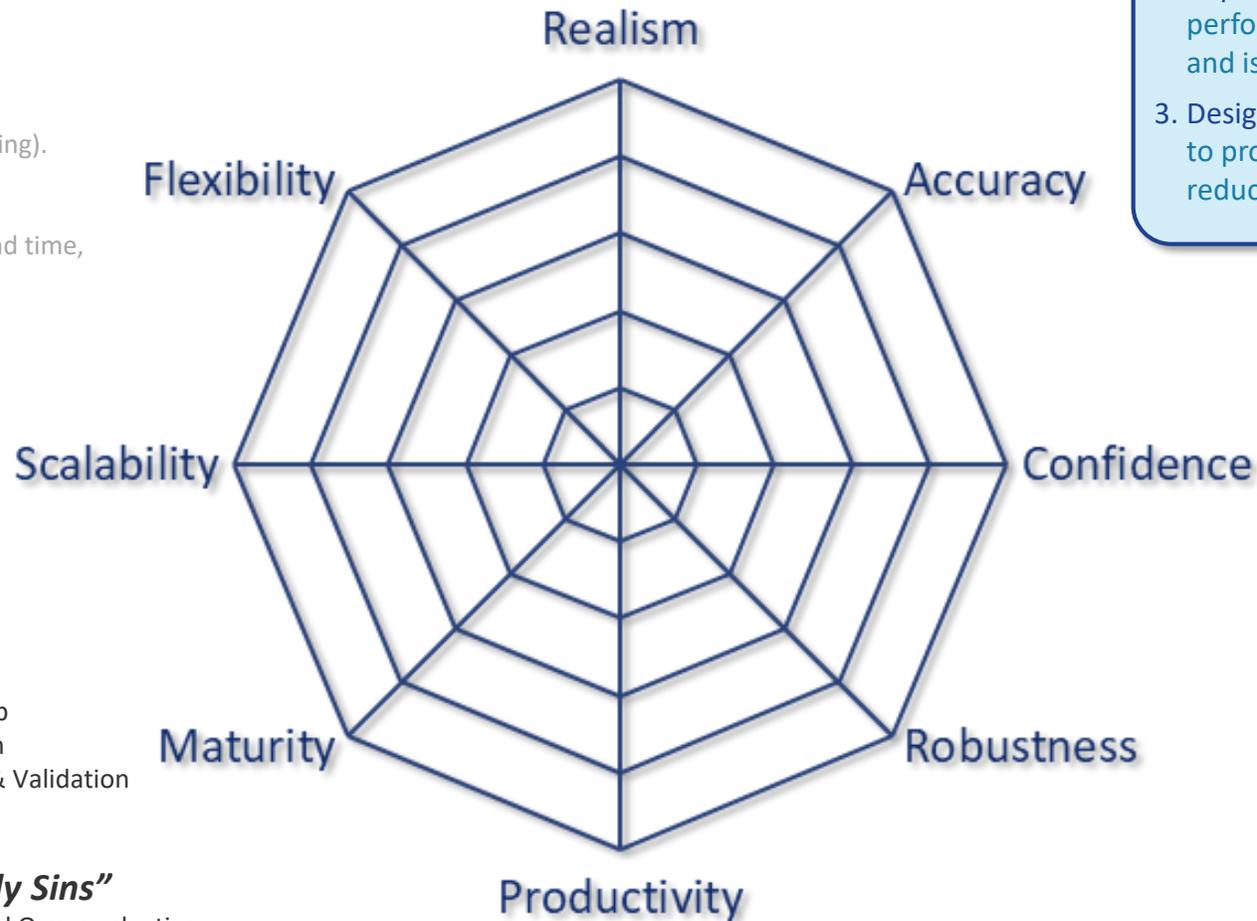
- Garbage-in = Garbage-out
- To advise and inform *decision-making*, models must be *trusted* and *understood* by both the people composing the models as well as those interpreting the models.

## Crucial to Computational Modeling

- Computationally Literate Workforce & Leadership
- Hardware & Software Infrastructure & Ecosystem
- Facilities, Processes and Culture for Verification & Validation

## Digital Opportunities vs. LEAN's "Deadly Sins"

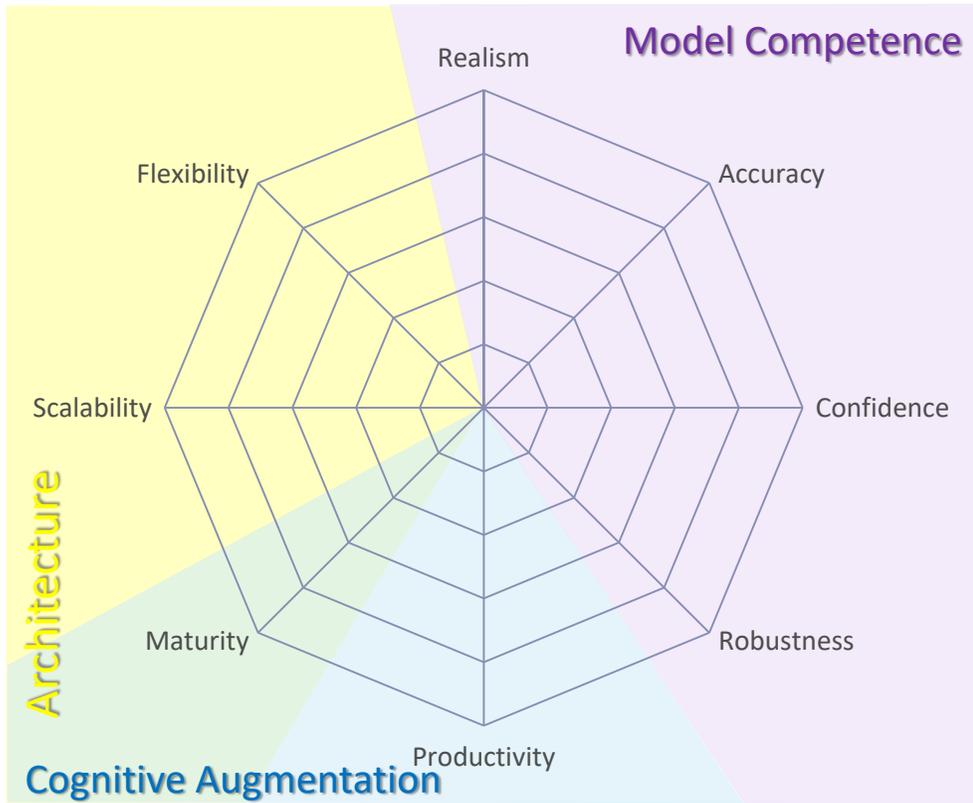
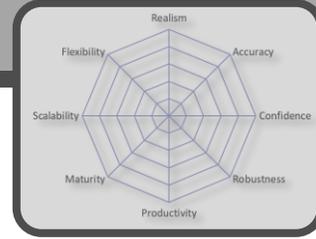
- Reusable Virtualized Assets vs. Inventory/WIP and Overproduction
- Automation & Visibility to increase Productivity and reduce Waste / Defects / Rework
- Digital Thread Workflow vs. Motion/Waiting dependencies/hand-offs and Underused Talent



## Computational Modeling Goals

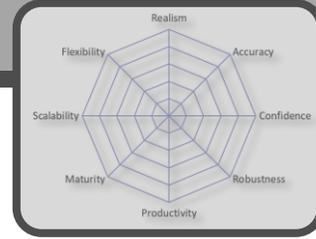
1. Assert a **Region of Competence** for a model where its use is numerically stable (**ROBUSTNESS**) with minimal simplifying constraints (**REALISM**) and quantifiably bounds uncertainties (**CONFIDENCE**) of results with validated predictive **ACCURACY**.
2. Implement the model with an **Architecture** that performs capably on HPC hardware (**SCALABILITY**) and is interoperable and extensible (**FLEXIBILITY**).
3. Design model use and software management to promote efficient workflows (**PRODUCTIVITY**), reduce waste and improve quality (**MATURITY**).

# Co-Design Web: **Goals**



<b>Realism</b>	<i>Completeness of ...</i>	<b>... Model's Region of Competence</b>
<b>Accuracy</b>	<i>Validity within...</i>	
<b>Confidence</b>	<i>Error bounding within...</i>	
<b>Robustness</b>	<i>Stability &amp; Assertability of...</i>	
<b>Productivity</b>	<b>Cognitive Augmentation</b>	<i>&amp; Waste Reduction</i>
<b>Maturity</b>		<i>&amp; Architecture Quality</i>
<b>Scalability</b>	<i>Capable &amp; High Performance</i>	<b>Architecture</b>
<b>Flexibility</b>	<i>Modular, Extensible, Interoperable</i>	

# Co-Design Web: Measurement



0 = **Undesirable (Absent)**

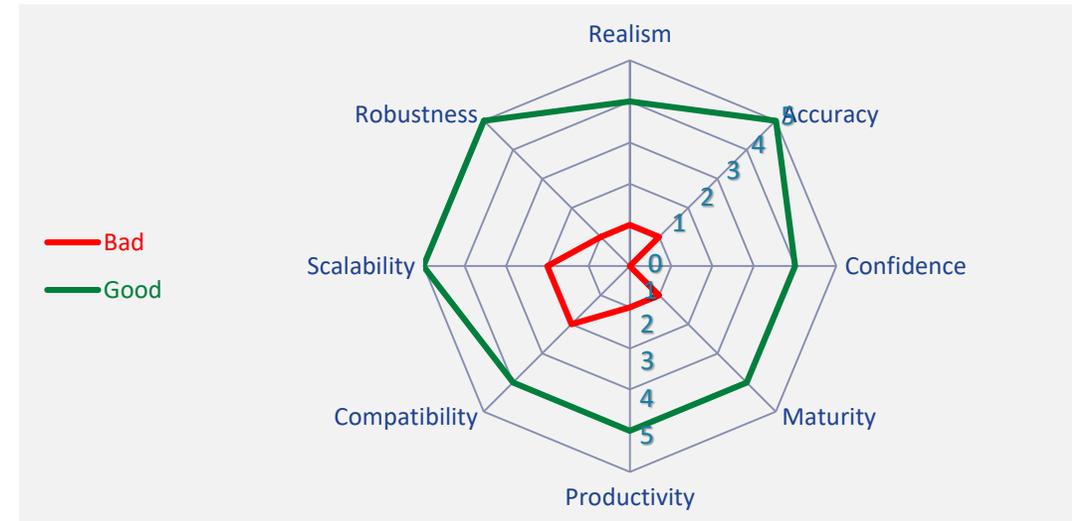
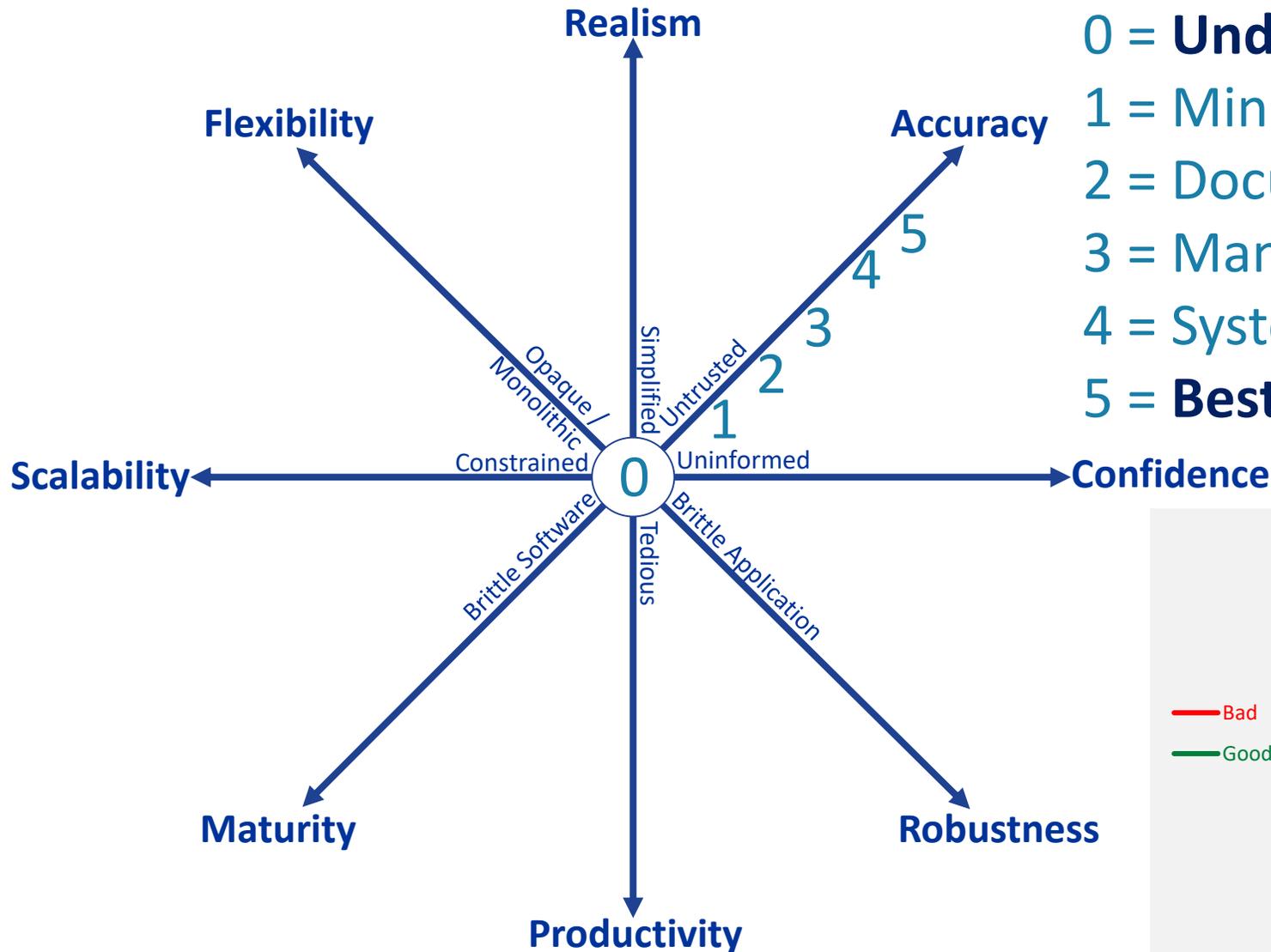
1 = Minimally present

2 = Documented / Basic

3 = Managed & Consistent / Addressed

4 = Systemic & Repeatable / Advanced

5 = **Best in Class / State of the Art**



# Challenges

# Challenge: Legacy



# Legacy: *Strength* becomes *Limitation*

- ✓ Experience
- ✓ Confidence
- ✓ Regulatory acceptance
- ❖ Sunk investments deter re-investment
- Obsolete functionality / infrastructure
- Cost/complexity for backward compatibility
- Hinders innovation / adoption of novel practices

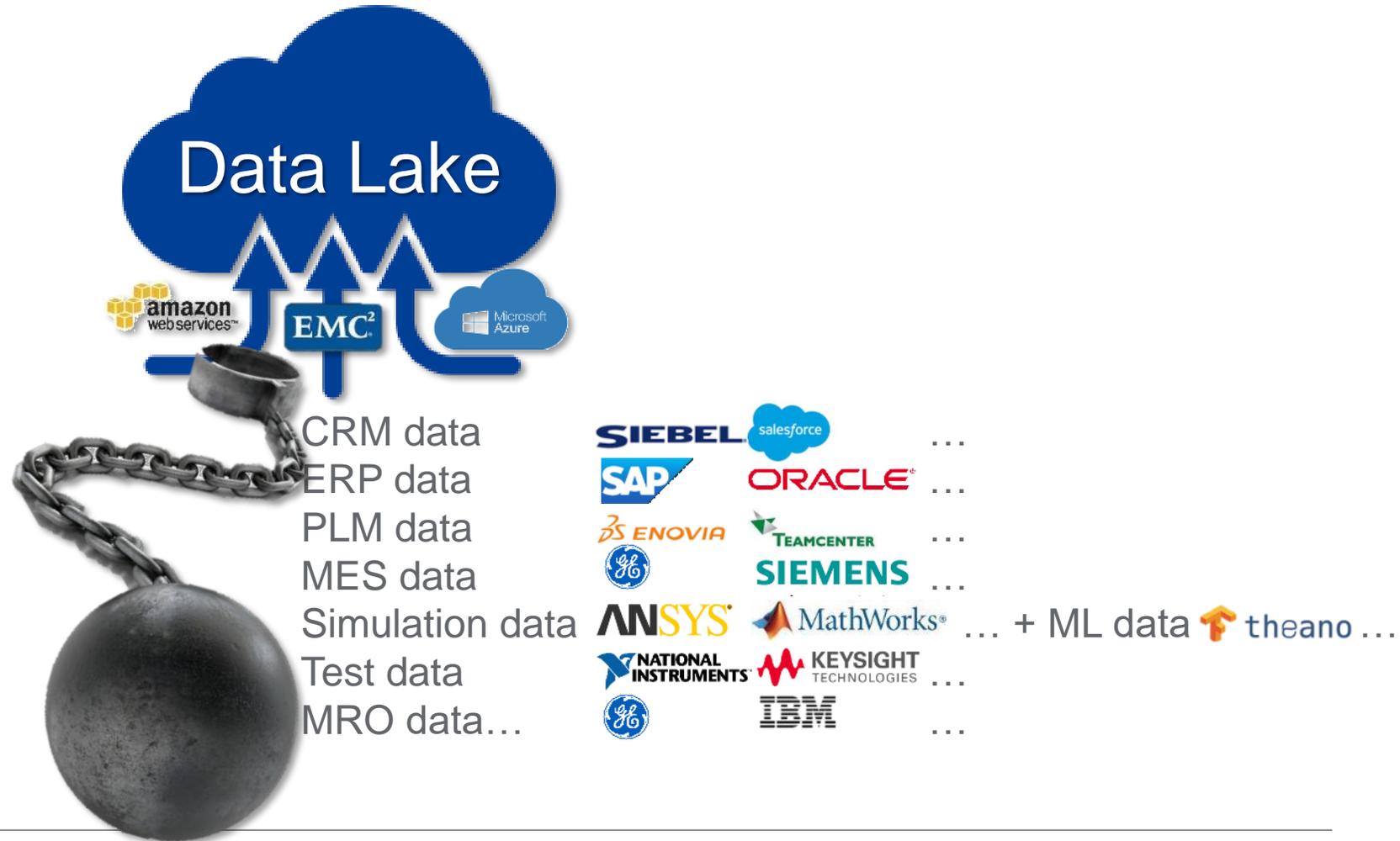




Data Lake

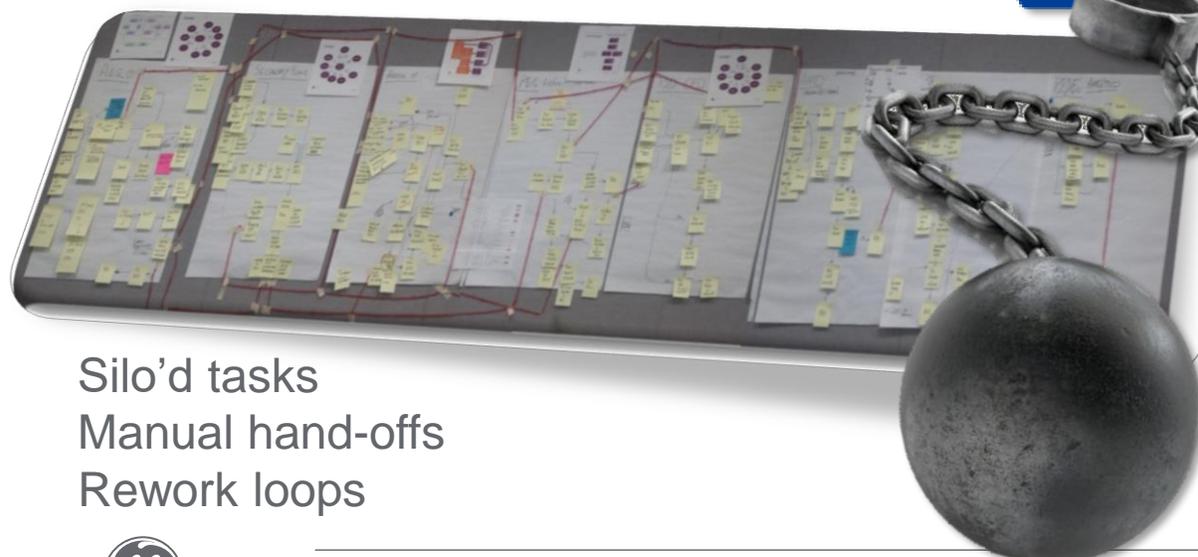
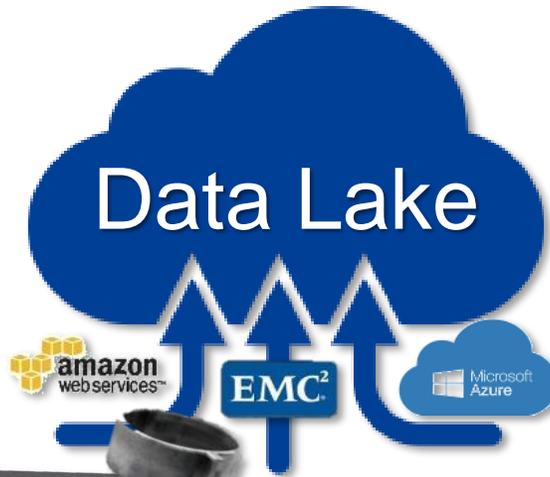
# Data Lake Storage & Services Infrastructure

DIGITAL THREAD CENTRALIZATION/FEDERATION OF **LEGACY DATA / SOFTWARE**



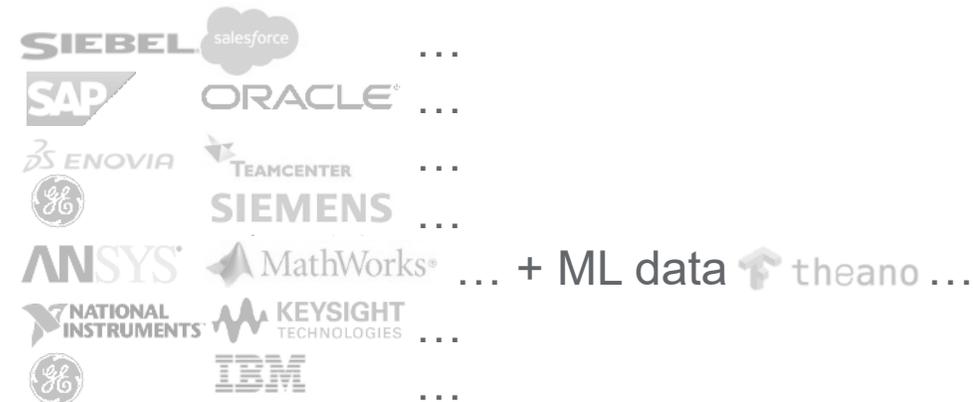
# Data Lake Storage & Services Infrastructure

DIGITAL THREAD SOURCES FROM **LEGACY WORKFLOWS / PROCESSES**



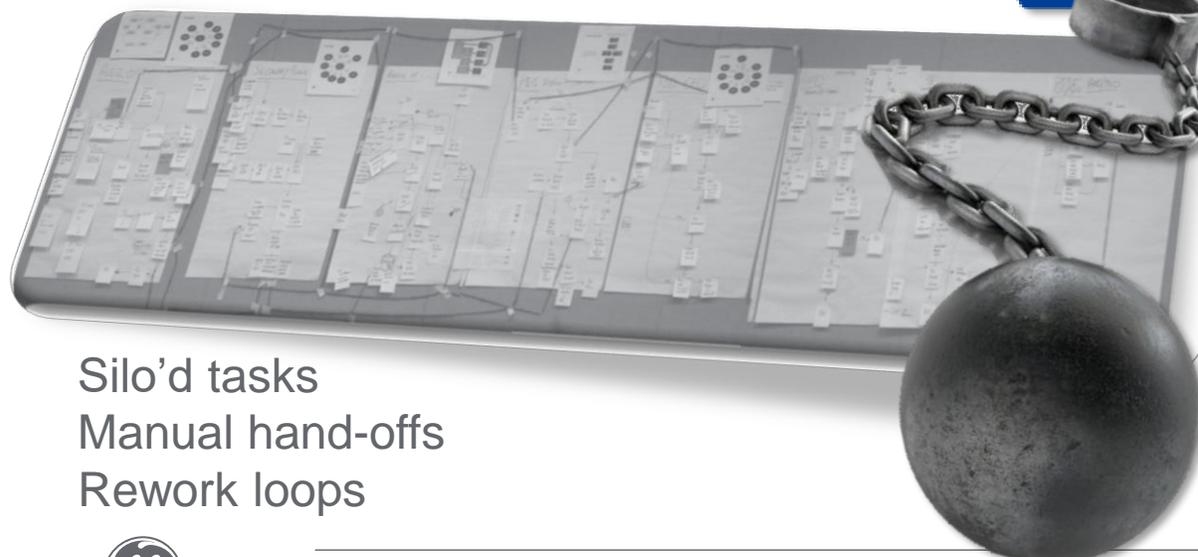
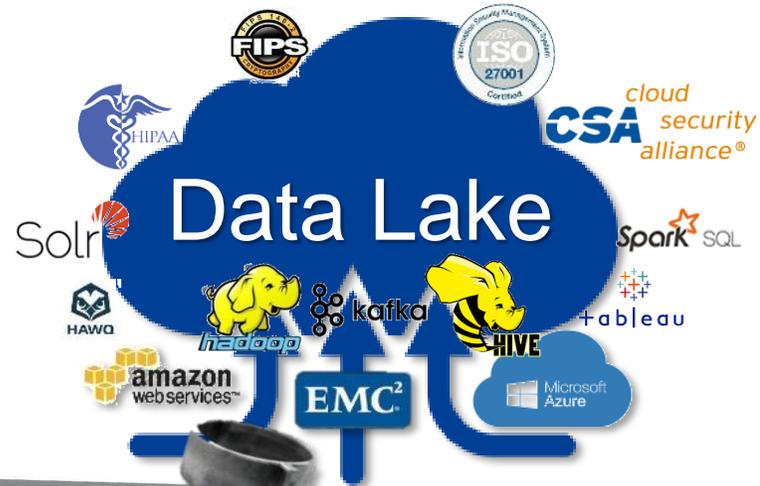
Silo'd tasks  
Manual hand-offs  
Rework loops

CRM data  
ERP data  
PLM data  
MES data  
Simulation data  
Test data  
MRO data...



# Data Lake Storage & Services Infrastructure

INDEXING, SEARCH, REPORTING, PROTECTION & COMPLIANCE

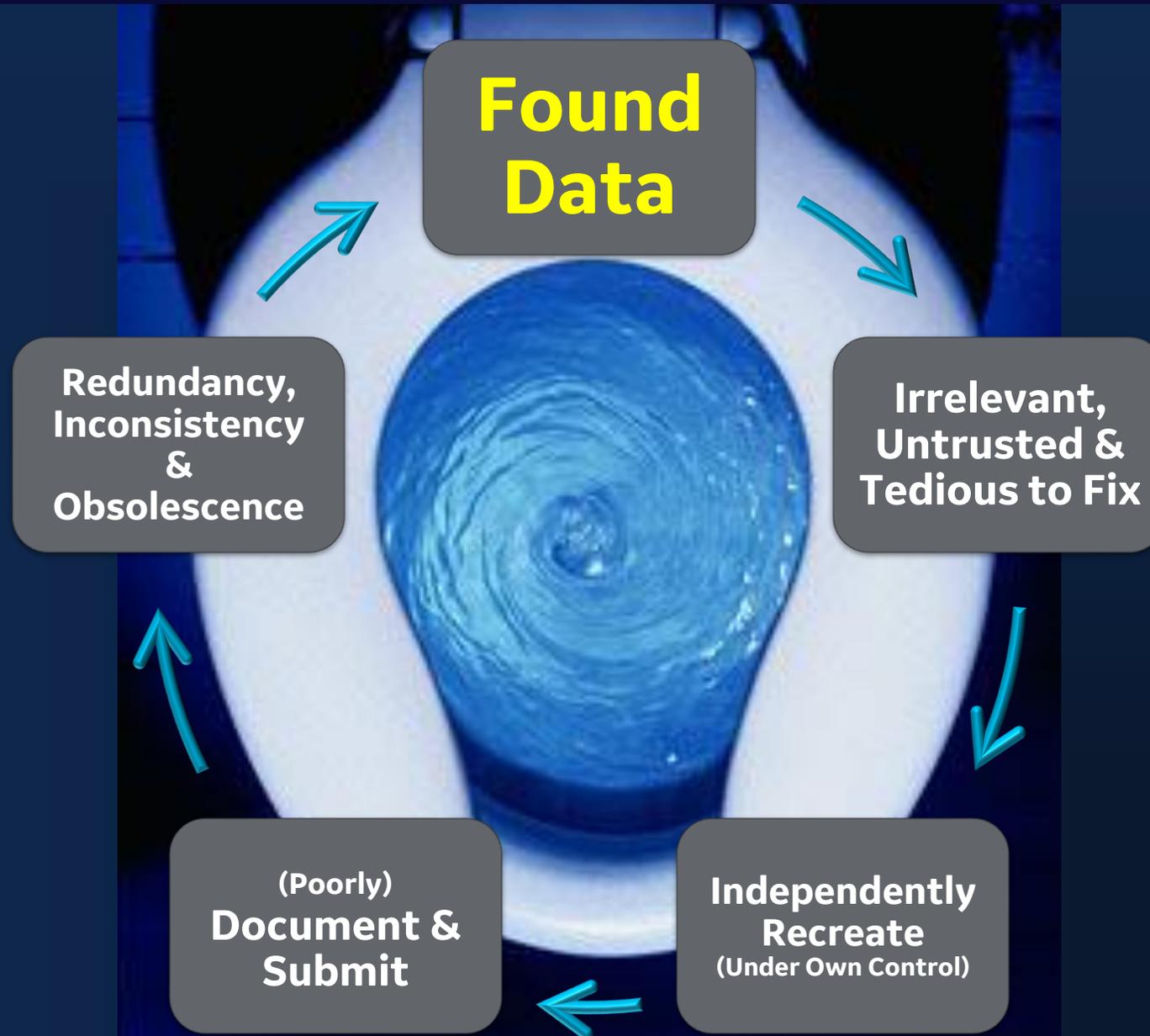


Silo'd tasks  
Manual hand-offs  
Rework loops

CRM data  
ERP data  
PLM data  
MES data  
Simulation data  
Test data  
MRO data...



# CHALLENGE: VICIOUS CYCLE OF DATA DEGRADATION





  
Data Swamp  
Challenge

# Opportunity: Decision Provenance – *akin to “Design Rationale” concept*

## DECISIONS AS “PRIMARY INDEX” TO UNDERLYING DATA/ANALYSES

At time of DECISION,  
explicitly capture into *knowledge steward*:

### prov·e·nance

/ˈprəʊvənəns/ 

*noun*

noun: provenance

the place of origin or earliest known history of something.

"an orange rug of Iranian provenance"

synonyms: [origin](#), [source](#), [place of origin](#); [More](#)

- the beginning of something's existence; something's origin.  
"they try to understand the whole universe, its provenance and fate"
- a record of ownership of a work of art or an antique, used as a guide to authenticity or quality.  
plural noun: [provenances](#)  
"the manuscript has a distinguished provenance"

Who	Proposed, Reviewed, Approved, Tested?
What	<b>Alternatives</b> were considered? Known <b>Unknowns</b> ( <i>environmental, economic, ...</i> )
Why	<b>Assumptions</b> ( <i>limitations, dependencies, technology, ...</i> ) Evaluation <b>Criteria</b> & relative weights  (and then <b>link</b> to underlying references):
How	<b>Data</b> analyses supporting the decision Modeling <b>methods</b> applied (+ <b>intellectual debt</b> ) Physical process for <b>measurement</b> ( <i>Gage R&amp;R, ...</i> ) Future <b>footnotes</b> : exemplar practices / learnings <i>and opportunities to improve (given more time/budget/capability)</i>



# Challenge: Machine Learning Reference Data

## SUFFICIENCY FOR TRAINING

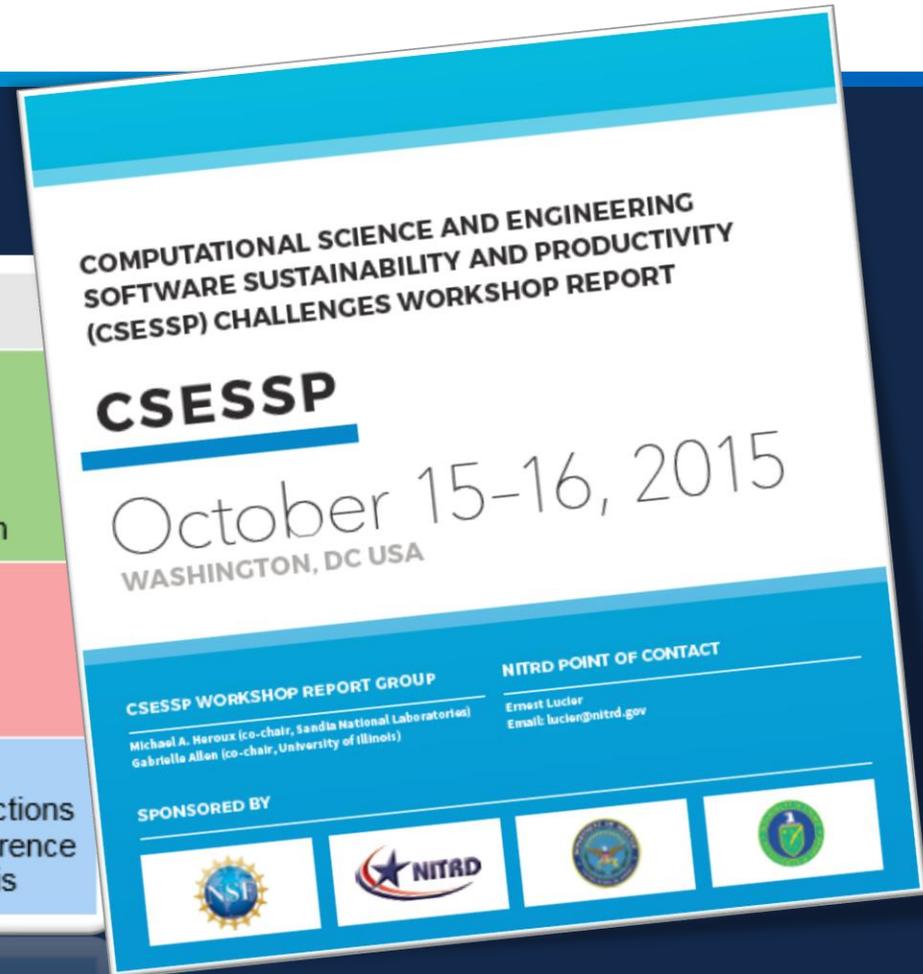
### Complexity of Response Hyper-surfaces    Sparsity vs. Characterization Complexity

- Non-Ergodic (e.g., nucleation)
  - Eigenfrequency (e.g., resonance)
  - Discontinuity (e.g., phase transition)
  - Stochasticity (e.g., turbulence)
- Time/cost to generate synthetic data
  - Time/cost to generate experimental data
  - Consistency/Cost of data fusion between experimental + synthetic sources
  - Dimensionality selection & hidden variables
  - Access (proprietary and/or classification)
  - Cost to validate synthetic data
  - Cost to validate labeling of data



# Challenge: Metrics and Incentives (financial recognition) / GAAP

	Engines of Productivity	Instruments for Insight
<b>Increase</b>	Return on Labor Profit Margins Supply Chain & Distribution Efficiency Design Exploration Agility to Seize Opportunities	Product Diversity & Novelty Yields & Production Capacity Data-driven decisions Trade-off Analysis Perception of Previously Unseen
<b>Decrease</b>	Costs of Overhead & Rework Time to Market Equipment Downtime Response Time to Fix Problems	Operational Exposure Uncertainty & Risk Contradictions Noise obscuring Main Effect
<b>Methods of Practice</b>	Automation of repetitive tasks/tests Faster than real time simulation/analysis Digitally replicate studied resources Concurrent studies on parallel system	Model unmeasurable effects Isolate effects in complex interactions Observe without physical interference "Big Data" analysis and synthesis



<https://www.nitrd.gov/pubs/CESSPWorkshopReport.pdf>



# Challenge: Incentives (cross-silo recognition / “True Digital Thread”)

Even if metric is accepted – who gets the credit?

Organizational & Cultural Challenges

	Engines of Productivity	Instruments for Insight
<b>Increase</b>	Return on Labor Profit Margins Supply Chain & Distribution Efficiency Design Exploration Agility to Seize Opportunities	Product Diversity & Novelty Yields & Production Capacity Data-driven decisions Trade-off Analysis Perception of Previously Unseen
<b>Decrease</b>	Costs of Overhead & Rework Time to Market Equipment Downtime Response Time to Fix Problems	Operational Exposure Uncertainty & Risk Contradictions Noise obscuring Main Effect
<b>Methods of Practice</b>	Automation of repetitive tasks/tests Faster than real time simulation/analysis Digitally replicate studied resources Concurrent studies on parallel system	Model unmeasurable effects Isolate effects in complex interactions Observe without physical interference “Big Data” analysis and synthesis

Cost Here

Payoff Here



# Knowledge Challenge: Intellectual Debt



## BEYOND EXPLAINABILITY: UNDERSTANDING

### Intellectual Debt: With Great Power Comes Great Ignorance

[Jonathan Zittrain](#), Jul 24

This kind of discovery — *answers first, explanations later* — I call “intellectual debt.”

We gain insight into what works without knowing why it works. We can put that insight to use immediately, and then tell ourselves we’ll figure out the details later [debt to be paid in the future].

We are borrowing as a society, rather than individually; artificial intelligence and specifically, machine learning are [being applied] to a seemingly unlimited number of new areas of inquiry. The distinct promise of machine learning lies in suggesting answers to fuzzy, open-ended questions by identifying patterns and making predictions.

1. When we don’t know how something works, it becomes hard to **predict** how well it will **adjust to unusual situations**.
2. Machine learning models are becoming pervasive, **compounding black box opacity**:
  - a. oracular answers to single problems can generate consistently helpful results, but
  - b. as AI systems gather and ingest data, they produce data of their own, then consumed by still other AI systems.
3. We need to know our exposure: we should invest in a **collective intellectual debt balance sheet**. We must keep track of just where we’ve plugged in the answers
4. Traditional debt shifts control: from borrower to lender, and from future to past, as later decisions are constrained by earlier bargains. Answers without theory — intellectual debt — also will shift control in subtle ways. [...] A world of knowledge without understanding becomes, to those of us living in it, a world without discernible cause and effect, and thus a world where **we might become dependent on our own digital concierges** to tell us what to do and when.
5. Without the theory, we lose the autonomy that comes from **knowing what we don’t know**.



*Technical debt arises when systems are tweaked hastily, catering to an immediate need to save money or implement a new feature, while increasing long-term complexity. [...] When something stops working, this technical debt often needs to be paid down as an aggravating lump sum.*

# Challenge: Foundational Limits of AI/ML

# Opportunity: [Science of {AI/ML} for Science]

ML/AI methods do not solve all problems:

some are simply too complex for machines

- detection of zero-day computer viruses
- resilient codes to arbitrary hardware failures
- Inference of chaotic dynamics of TCP flows

## Not ML Solvable:

Classes of problems with properties at limits:

- **Computability limit** (undecidability)  
[Church's proof](#) that Hilbert's *Entscheidungsproblem* is unsolvable, and [Turing's theorem](#) that there is no algorithm to solve the [halting problem](#).
- **Expressability limit** (cannot state into formalism)  
[Tarski's undefinability theorem](#) on the formal undefinability of truth
- **Provability limit** ([Gödel's incompleteness theorems](#))

### ARTICLES

<https://doi.org/10.1038/s42256-018-0002-3>

nature  
machine intelligence

Corrected: Author Correction

### Learnability can be undecidable

Shai Ben-David<sup>1</sup>, Pavel Hrubeš<sup>2</sup>, Shay Moran<sup>3</sup>, Amir Shpilka<sup>4</sup> and Amir Yehudayoff<sup>5\*</sup>

- **Learnability limit** (can only resolve problem by choosing an axiomatic universe within which it is applicable therefore solution is not applicable to data outside that universe – i.e. sampling from separate infinities)

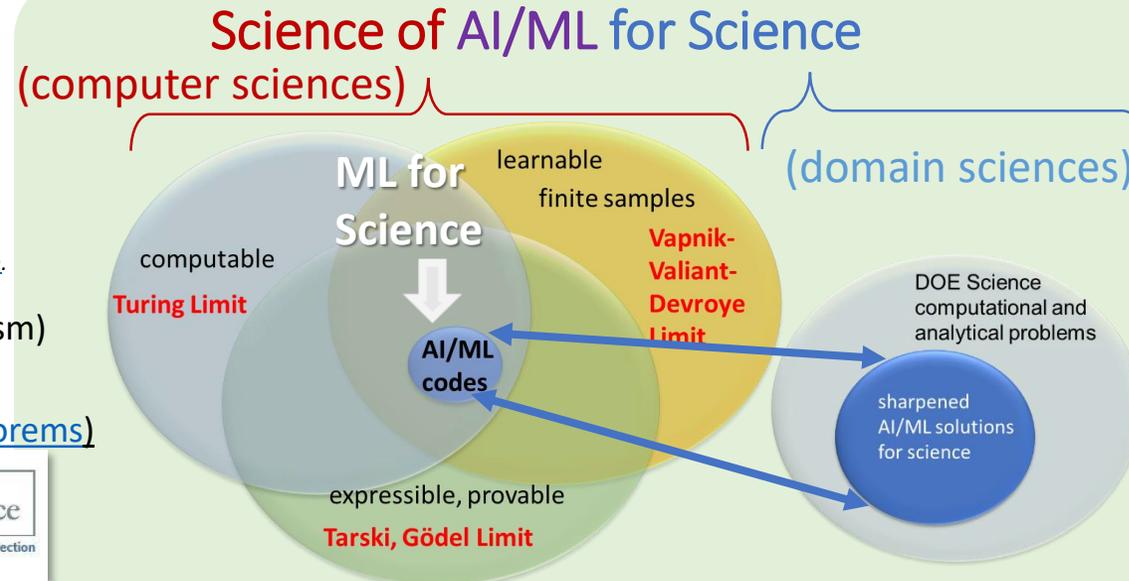
(Vapnik <[learnability](#)> / Valiant <[learnability-paper](#) / [PPT](#)> / Devroye <[A Probabilistic Theory of Pattern Recognition](#)> - esp. slow rates of convergence – chapter 7)

E.g., unbounded deviation from Bayes', infinite Vapnik-Chervonenkis dimension, ...

(See also: [Five Machine Learning Paradoxes that will Change the Way You Think About Data](#) )

AI/ML Solutions exploit underlying sciences

- ensure solvability: computable, learnable, expressible, provable
- sharpen AI/ML solutions: structure and constraints from laws



Credit: Nagi Rao

[raons@ornl.gov](mailto:raons@ornl.gov)



The reasoning is as follows: consider AI/ML method that attempts to discover laws (truths) from data as being attempted in several science areas. An explanation of truth is proof. But Gödel's theorem (a version) shows that we cannot mechanically provide proofs for all truths – so some truths will remain undiscovered by ML/AI method.

The implications could be quite deep: one can write ML codes that to try to solve this problem but its output would be either incomplete or unsound or both – if applied without care, this ML solution could potentially output “pseudo untrue” laws.

Physical-Abstract Hybrid Laws lead to sharpened AI/ML:

- Physical, abstract, hybrid laws
  - physical systems
  - cyber infrastructures
  - cyber-physical systems
- Customized AI/ML solutions may exploit
  - structure
  - relationships, correlations
  - constraints

**Apply Characterization:** Analytical and mathematical characterizations of limits and their interpretation within application/domain context

# Summary

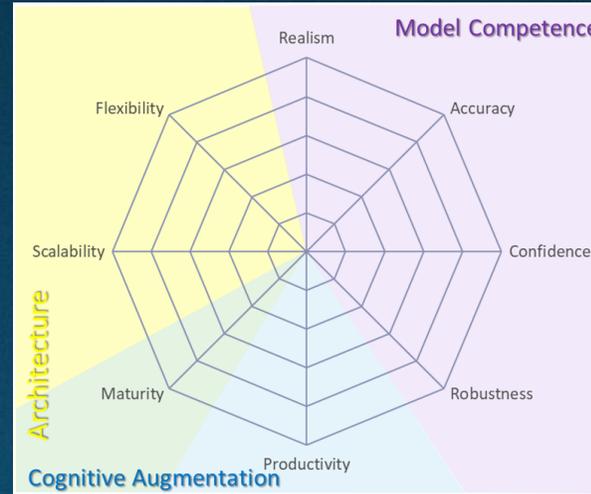
# Criticality of Computational Methods, Model Maturity, Co-Design, Public-Private Engagement

## Engineering's core competency is Problem Solving



**Problem Solving** critically relies upon **Modeling**  
(the problem, the solution and the process in between)

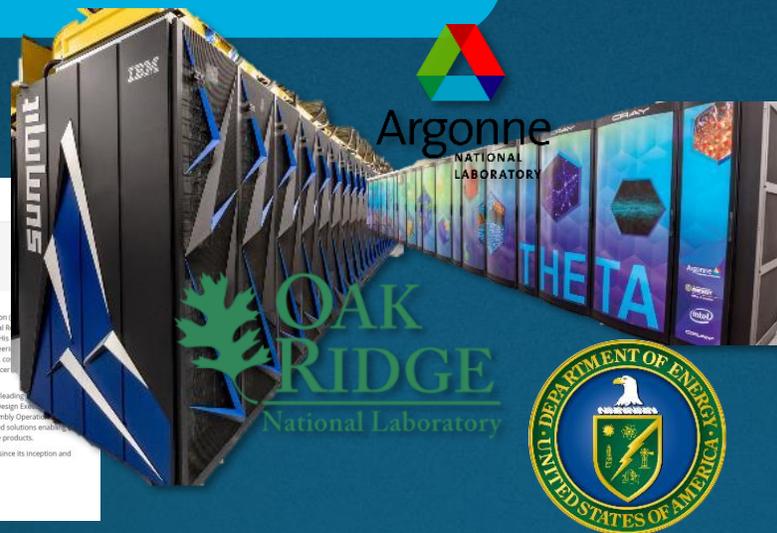
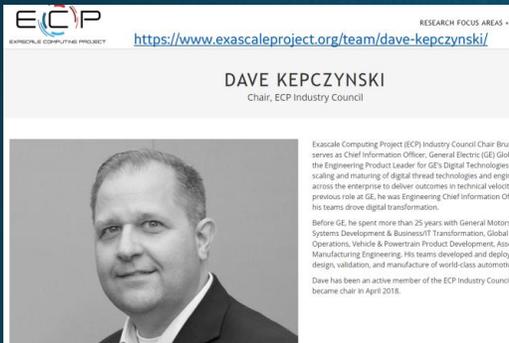
**Modeling** critically relies upon **Computational Methods**  
(as an engine for scale, productivity, consistency and capability)



<b>Realism</b>	Completeness of ...	
<b>Accuracy</b>	Validity within...	... Model's Region of Competence
<b>Confidence</b>	Error bounding within...	
<b>Robustness</b>	Stability & Assertability of...	
<b>Productivity</b>	<b>Cognitive Augmentation</b>	& Waste Reduction
<b>Maturity</b>		& Architecture Quality
<b>Scalability</b>	Capable & High Performance	
<b>Flexibility</b>	Modular, Extensible, Interoperable	<b>Architecture</b>

## GE engagement with Government Leadership Computing

Exascale Computing Project  
Industry Council Chair  
GE Research CIO Dave Kepczynski



## Co-Design: Landing Opportunities from Blue Sky Aspirations

