Microsoft Azure

Computing for AI and Science Beyond 2030

Andrew Jones

www.linkedin/com/in/andrewjones
@hpcnotes

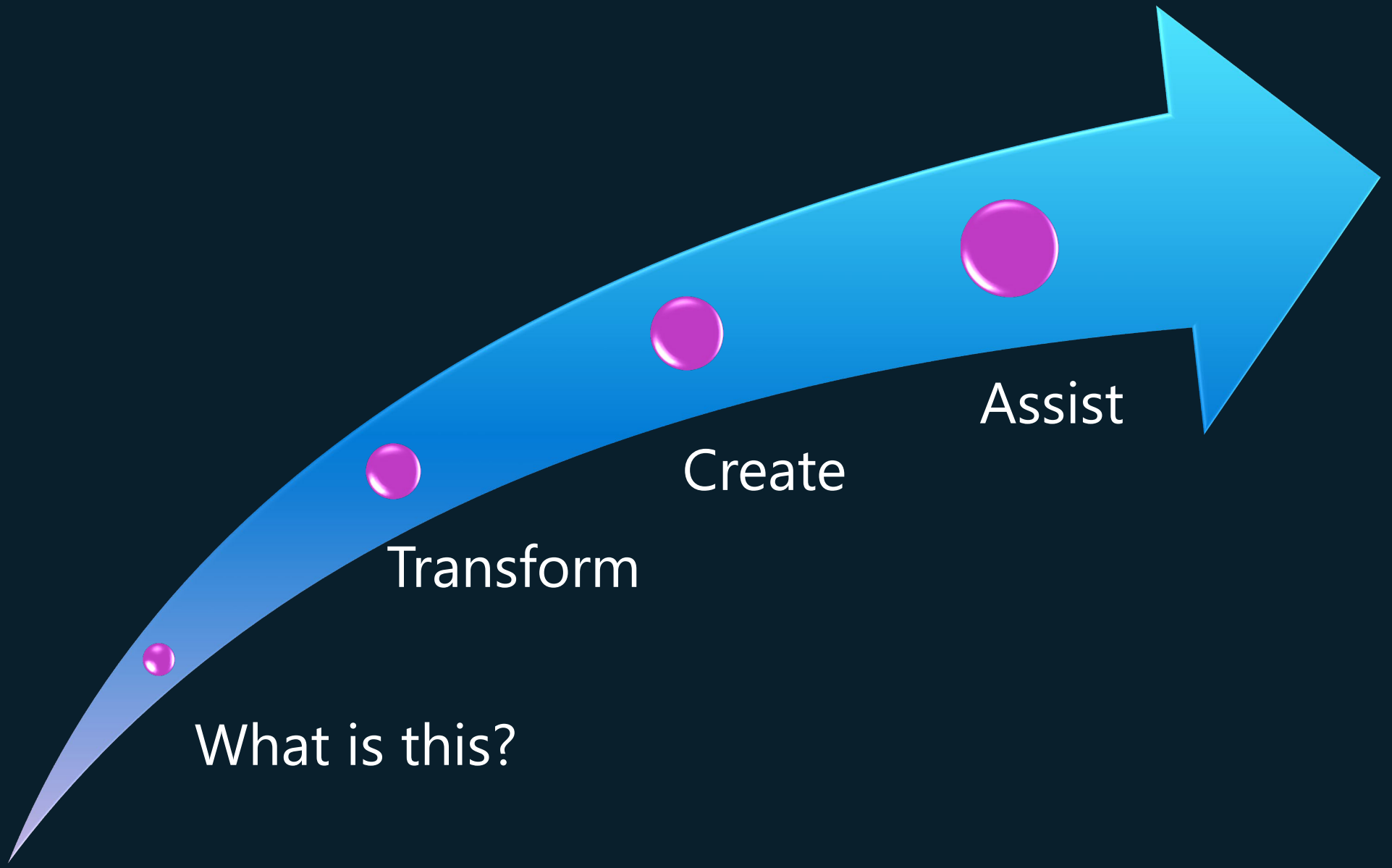
This presentation includes personal perspectives … which may not necessarily represent the views of employer or anyone else.

Inbox
10:28
Copilot
Dear Maya,
Thank you for bringing this issue to our attention. I've modeled the sales growth and discovered that the Proseware decline last quarter was due to a drop in the sale of Reusable Containers. I recommend that we offer a 10% discount on Reusable Containers to increase interest in the product, which will increase sales. I've included a link to the projected sales model showing the impact of this change. Here's a snapshot:
Projected Sales growth over time for Proseware
Please let me know if you have any questions,
Best regards,
Lyle Koenig
Business Development Manager
Insert
Adjust
Edit

A frog wearing a cowboy costume riding a bicycle on the moon

Photo from Bing Image Creator


A
A close-up of a jaguar cat looking towards the camera at sunrise
B
A close-up of a jaguar cat looking threateningly towards the camera


What is this?
Transform
Create
Assist

What makes all this possible?

AI

Large Language Models, Foundational Models, GPT, …

100's Billions

ML models with hundreds of Billions of Parameters

Many 1000's

AI Supercomputers using thousands of GPUs with HPC interconnect *plus* global inference capacity

$100's Billions

Predicted market opportunity of hundreds of $Billions

Scale

AI Supercomputer

2020

10,000
V100 GPUs

Comparable to
Top 5
largest supercomputers in the world

2023 submission

14,400
H100 GPUs

#3
supercomputer in the world according to SC23 **TOP500 561PF HPL**

Actual 2023 system

Even bigger!

Global Scale Infrastructure

Datacenters worldwide **300+**	Security and threat intelligence experts **10K+**	Miles of IB fiber **29K+**
190+ Network PoPs	**1,300T+** Transactions on global-scale infrastructure monthly	**>95%** of Fortune 500 use Microsoft Azure

HPC has long been an essential contributor behind everyday life, business, science, …

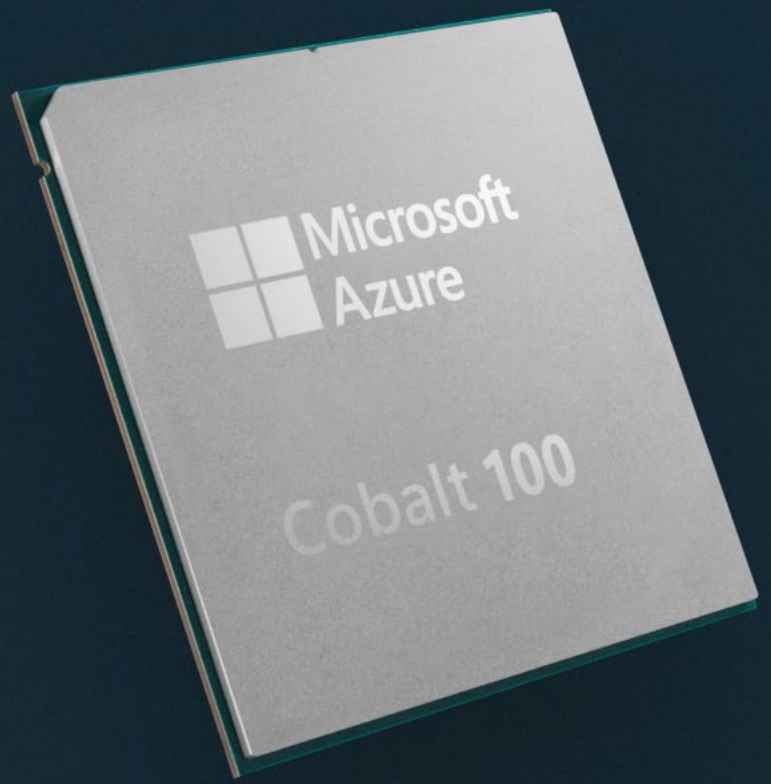
AI continues that trend.

So, what might be next for supercomputing?

Speculative Future Gazing

Processor trends

Azure Cobalt



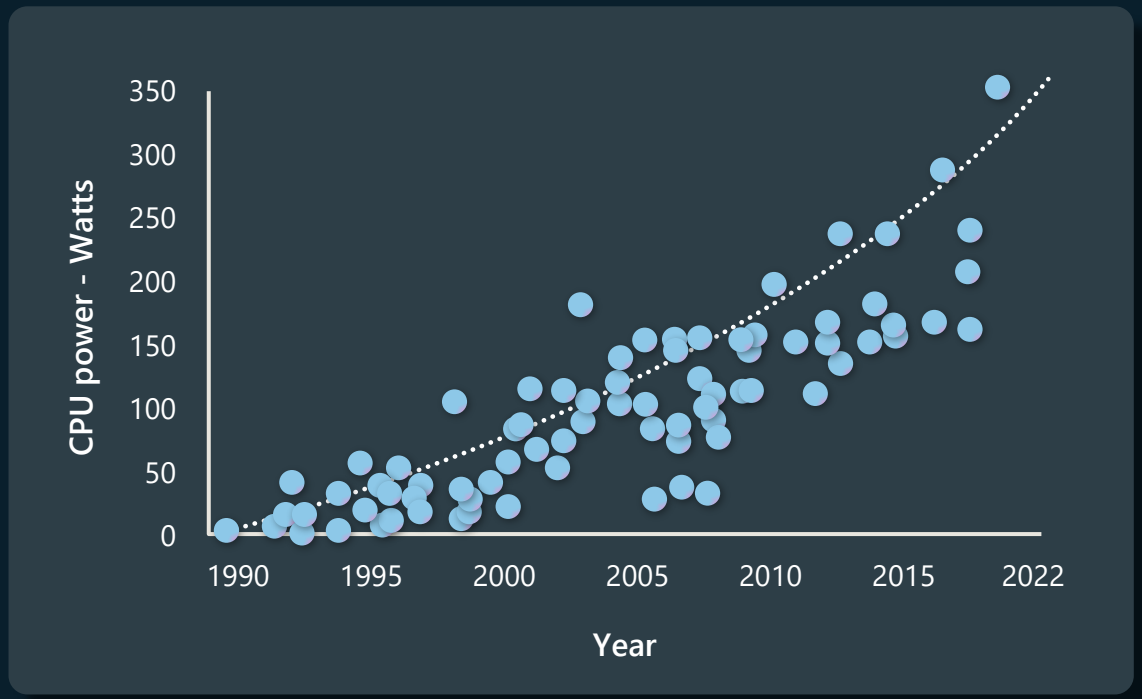
aka.ms/AzureCobalt

Azure Maia

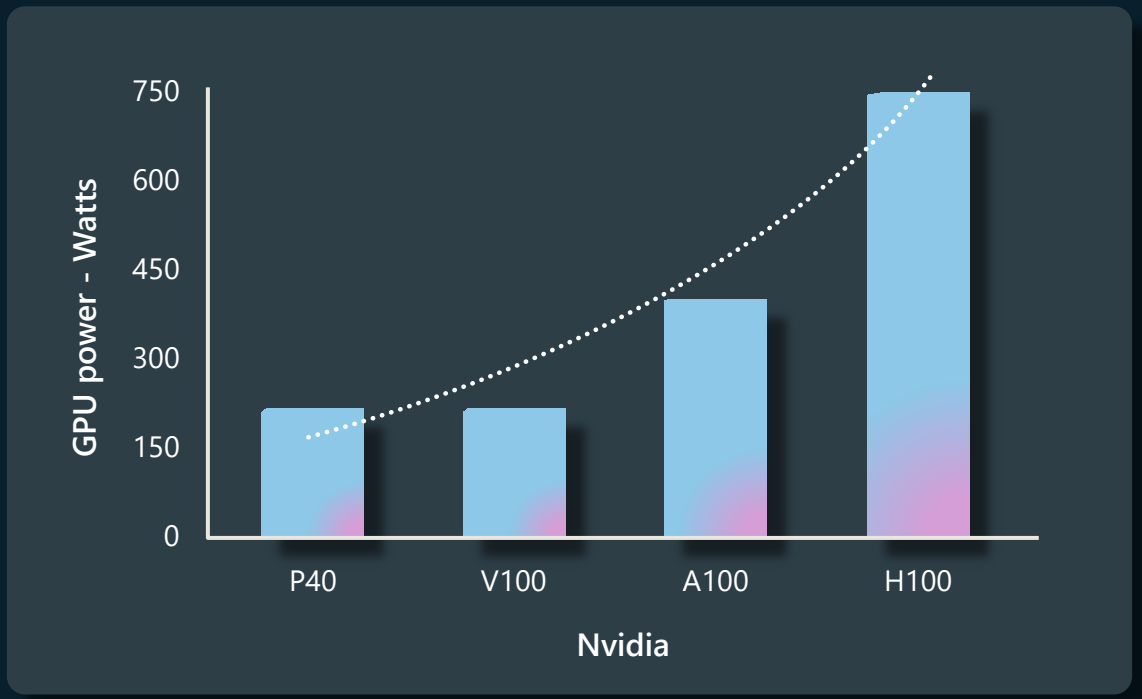
aka.ms/AzureMaia

Server cooling

CPU trends



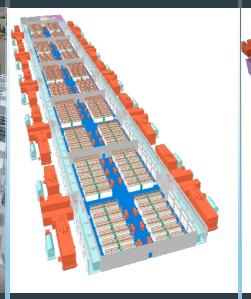
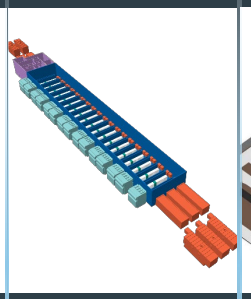
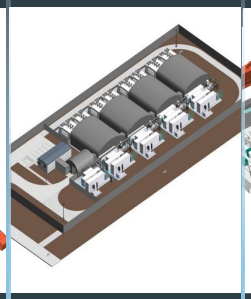
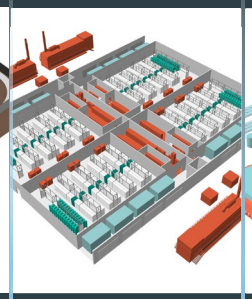
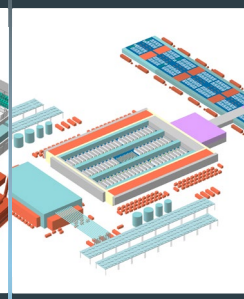
GPU trends



Datacenter PUE evolution

Windows Azure | Microsoft Azure

Generation

01	02	03	04	05	06	07	08	09	10	11
										
1989 Colocation	**2007** Density	**2009** Containment	**2012** Modular	**2015** Hyper-scale	**2017** Scalable Form Factor	**2018** Ballard	**2018** Power Harvesting	**2020** Rapid-deploy Datacenter	**2020** Multi-availability & Sustainability	**Newest 2023** AI GPUs
2.0+	1.5-1.8	1.4-1.6	1.1-1.3	1.17-1.25	1.17-1.19	1.15-1.18	1.15-1.18	1.15-1.18	1.12-1.14	1.09-1.17

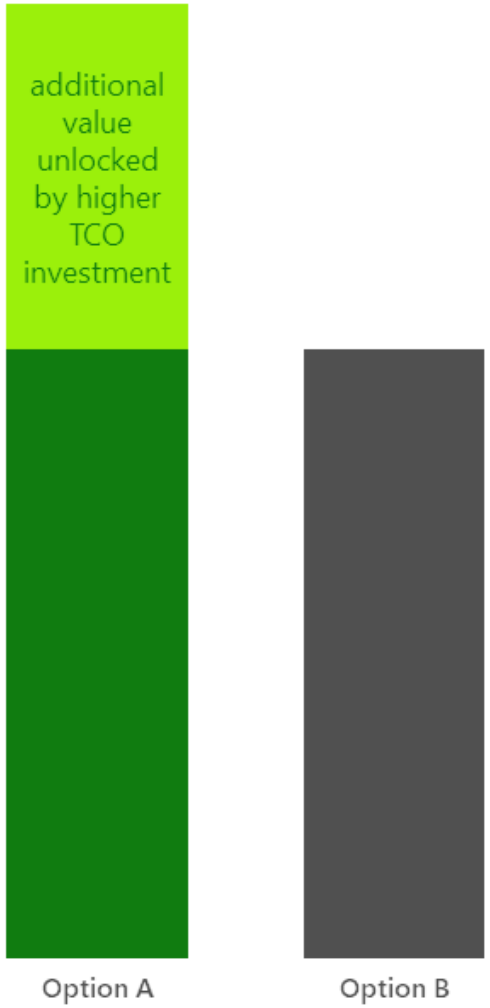
Power Usage Effectiveness (PUE)

Maximum # CPUs/GPUs per dollar ?

Deciding between technology A and technology B
purely on the basis of cost or TCO, without comparing value,
means working with only half of the information.

It would be like trying to determine
the winner of a football match
whilst only knowing one team's score
and not counting how many goals
the other team scored.

Value is just as critical to understand as cost



Value Generated

Comparing TCO alone is meaningless!

Value generated from HPC is usually a *lot more* than costs,
so differences in TCO *could* be considered negligible
in the context of the overall value conversation, and especially
if substantial additional value is unlocked by different solutions

TCO

Modeling value is not easy, but has huge impact

Assessing value is really hard - especially when key values arise long in the future

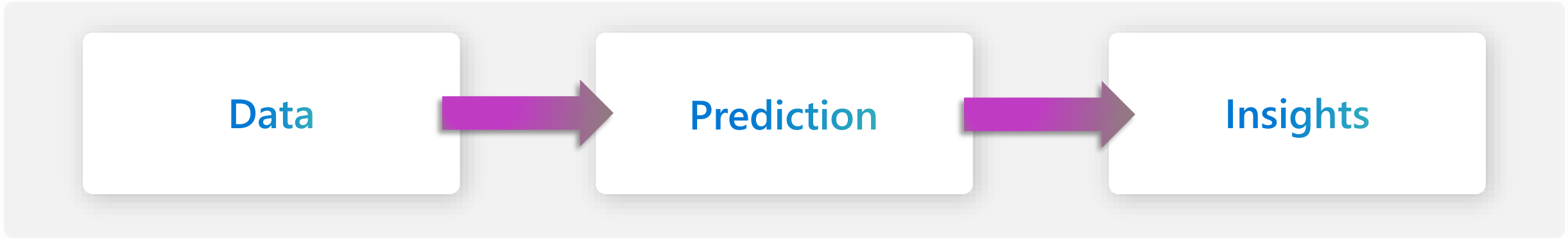
Often end up just using performance or TCO savings as poor proxy for value

Orgs that can capture and *optimize for* value have a huge competitive advantage

HPC

enables

insights


Data
Prediction
Insights

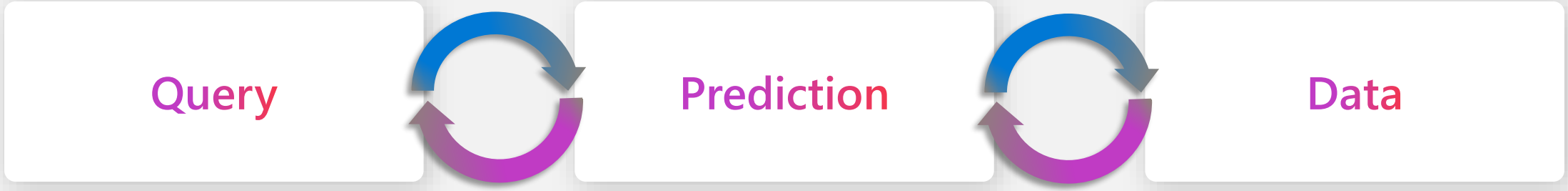
73% chance of rain
in your region on
Tuesday afternoon

What time will it stop raining?

Will the storm touch my village or pass by a safe distance?

Will our holiday be warm enough but not too hot?

Will climate change make this a bad location to build a house?


Query
Prediction
Data

People

Not just about “talent pipeline”!

Every HPC skills/talent pipeline initiative always focuses on programmers and/or use of parallel batch systems. Plenty of courses, summer camps, cases-for-urgency, …

Sysadmins, leaders, service managers, future types of users, vendors, …? Are these expected to make it up as they go? Assumed to be easy because “non-technical”?

We must get better at recognizing and supporting the diversity of roles+skills+people required to deliver the huge potential science & economic impacts of HPC/AI.

My predicted “Top 5 HPC & AI skills of the future”

Knowing & Using The Right Metrics

Finding the right metrics, measuring them sensibly, reporting effectively to stakeholders

Value | ~~Utilization~~ | ROI -> VOUpDI

Understanding how to *apply* AI

How and when to use AI to deliver useful value.

A big GPU cluster != AI | Training ML models != AI | Limits and opportunities of AI

Understanding & Managing Risk

Matters just as much as TCO and ROI. Sounds obvious?

Technology | People | Global

Living with Complexity

HPC hasn’t been simple for a long time, but complexity/options/depth growing fast

Cloud vs on-prem | AI vs “traditional” modeling | Dynamic vs predictable

Willingness to Accept Change

Ideally, a passion for change!

Inclusive and Diverse | Managing Inertia & Resistance | “Big Picture” Thinking

People will always be key.

Machines, whether HPC or AI, are “just” tools.

But the people who can adopt and optimally use better tools will tend to achieve more.

Microsoft

Thank you

www.linkedin.com/in/andrewjones
@hpcnotes