AIM311-S

SPONSORED BY INTEL

AI acceleration at the edge

Jonathan Lee
Chief Product Officer
ai.io

Michael Kleiner
VP Edge AI Solutions
OnLogic

Allan Gagnon
Senior Solutions Architect
Arduino

Alex White
Sales Applications Engineer
Intel

Mohan Potheri
Cloud Solutions Architect
Intel

AWS re:Invent 2023 - AI acceleration at the edge (AIM311)

AWS Events
100K subscribers
Subscribe

0 | Share | Clip | Save | ...

53 views Dec 3, 2023 #AWSreInvent #AWSreInvent2023

AI has the power to transform and create innovative business models that generate real value. Companies can build models with minimal data and, with Intel domain-specific toolkits, facilitate deploying solutions at scale. Listen as industry experts and customers OnLogic, Arduino, and ai.io describe how they were able to use AI technology at the edge to achieve strong business outcomes in agriculture, manufacturing, and automotive by optimizing customer experiences, forecasting, and managing inventory. Come away from this session with tangible examples of Intel software and hardware optimizations that led to creative, successful AI solutions at the edge. This presentation is brought to you by Intel, an AWS Partner.

Learn more about AWS re:Invent at https://go.aws/46iuzGv.

Subscribe:
More AWS videos: http://bit.ly/2O3zS75
More AWS events videos: http://bit.ly/316g9t4

ABOUT AWS
Amazon Web Services (AWS) hosts events, both online and in-person, bringing the cloud computing community together to connect, collaborate, and learn from AWS experts.

AWS is the world's most comprehensive and broadly adopted cloud platform, offering over 200 fully featured services from data centers globally. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—are using AWS to lower costs, become more agile, and innovate faster.

Agenda

1. Top Business Outcome for Edge AI
2. Bringing AI Everywhere
3. Compelling Edge AI Use cases
   Edge Computer vision booth demo
4. Real World Successes
   Featuring OnLogic, Arduino & ai.io
5. 4th Generation Intel Xeon for AI
6. The AI Pipeline of processing
7. Start building your AI solutions today

Edge Compute and AI: **Top Trends We're Seeing**

- By year-end 2026, 70% of large enterprises will have a documented strategy for edge computing, compared to fewer than 10% in 2023.
- CEO use of the term "digital" in their top business priorities has roughly doubled from 2018 to 2022.
- Digital data production at the edge is growing exponentially, creating the opportunity for deeper analysis, automation, AI /ML.*

Edge Compute and AI: **Top Trends We're Seeing**

> "Enterprises need a strategy so they can overcome and learn from challenges, choose technologies, manage costs and enable successful digital transformation."

Gartner

Building an Edge Computing Strategy,
Thomas Bittman, April 12, 2023

Bringing AI to the Edge

it starts with intel

AI Continuum

Bringing AI Everywhere

Note: Intel Core Ultra integrates NPU low power inference engine from Meteor Lake onwards.

Bringing AI Everywhere at the Edge

Move from development to deployment **fast**

Scalable hybrid AI approach

Leverage automatic processing of an AI workload using available/targeted system resources and accelerators on the edge, or in the cloud. OpenVINO and oneAPI allow developers to seamlessly transition between edge-to-cloud with a single code base.

Open tools to speed deployment

The OpenVINO toolkit optimizes deep learning inference deployment for hundreds of pretrained models across multiple Intel platforms. The Intel Geti Platform allows domain experts and data scientists to quickly build and train AI models.

Hundreds of deployment-ready solutions

Intel solutions include defect detection, worker and public safety, robotics, supply chain management, imaging diagnostics, and enhanced service delivery.

Bringing **AI Everywhere** to Unlock New Levels of **Innovation**

Education	Health	Finance	Retail	Government	Energy	Automotive	Manufacturing	Telco
Teacher Assistant Student Study Buddy Parent Chat Portal	Drug Discovery Doctor Co-pilot Patient Family Chatbot	Algorithmic Trading Customer Portfolio Assistant Risk / Credit Assessment	Product Promotion Customer Interface and Sentiment Tool Image Shopping Aid	Gov Services Chatbot Document Search Summarization Live Language Translation	Energy Consumption Forecasting Operational Performance Energy Trading Assistant	Autonomous Car Development Multi-language in car aid Supply Chain Optimization	Factory Automation Predictive Maintenance Precision Agriculture	Personalized Customer Services Network Automation Operational Performance

Secure the Edge

it starts with intel.

Layering Security Foundation for Solution Developers

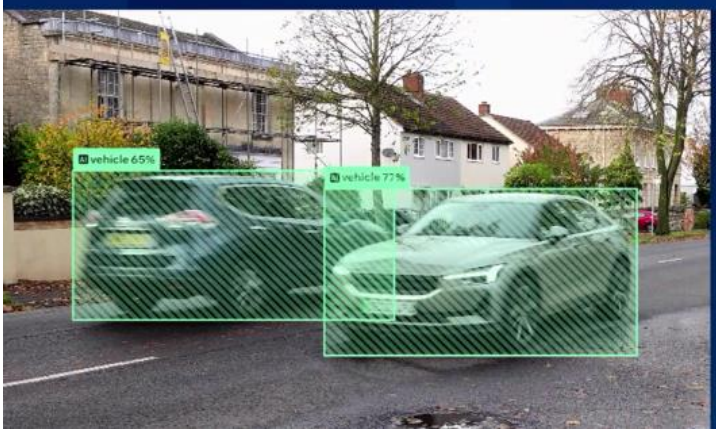
Secure at Boot	Seamless to Deploy	Protected at Runtime	Encrypted at Rest
▪ Always boot to safe default configurations ▪ Secure boot all the way from the HW up to applications ▪ Periodic monitoring of device integrity through remote attestation	▪ Zero touch onboarding ▪ Secure configuration updates at runtime ▪ Lift and shift secure application across multiple SW profiles	▪ VM and process memory is isolated via HW mechanisms, and protected by role-based access control ▪ Interfaces to hardware devices and controlled	▪ Root of trust stored in tamper resistant hardware (TPM, HSMs) ▪ Storage is encrypted with HW unique keys ▪ System software sill only decrypt authentic hardware
HW RoT and Secure Boot, Trusted Platform Module (TPM), Intel Platform Trust Technology (Intel PTT)	FIDO Device Onboard (FDO)	Intel Virtualization Technology (Intel VT) Intel Total Memory Encryption (Intel TME) Intel Software Guard Extensions (Intel SGX) Intel Trust Domain Extensions (Intel TDX)	Trusted Platform Module (TPM) Intel Platform Trust Technology (Intel PTT)

Edge Computer Vision

it starts with intel

Where Will Computer Vision Applications Disrupt and Innovate?

Computer vision at the edge

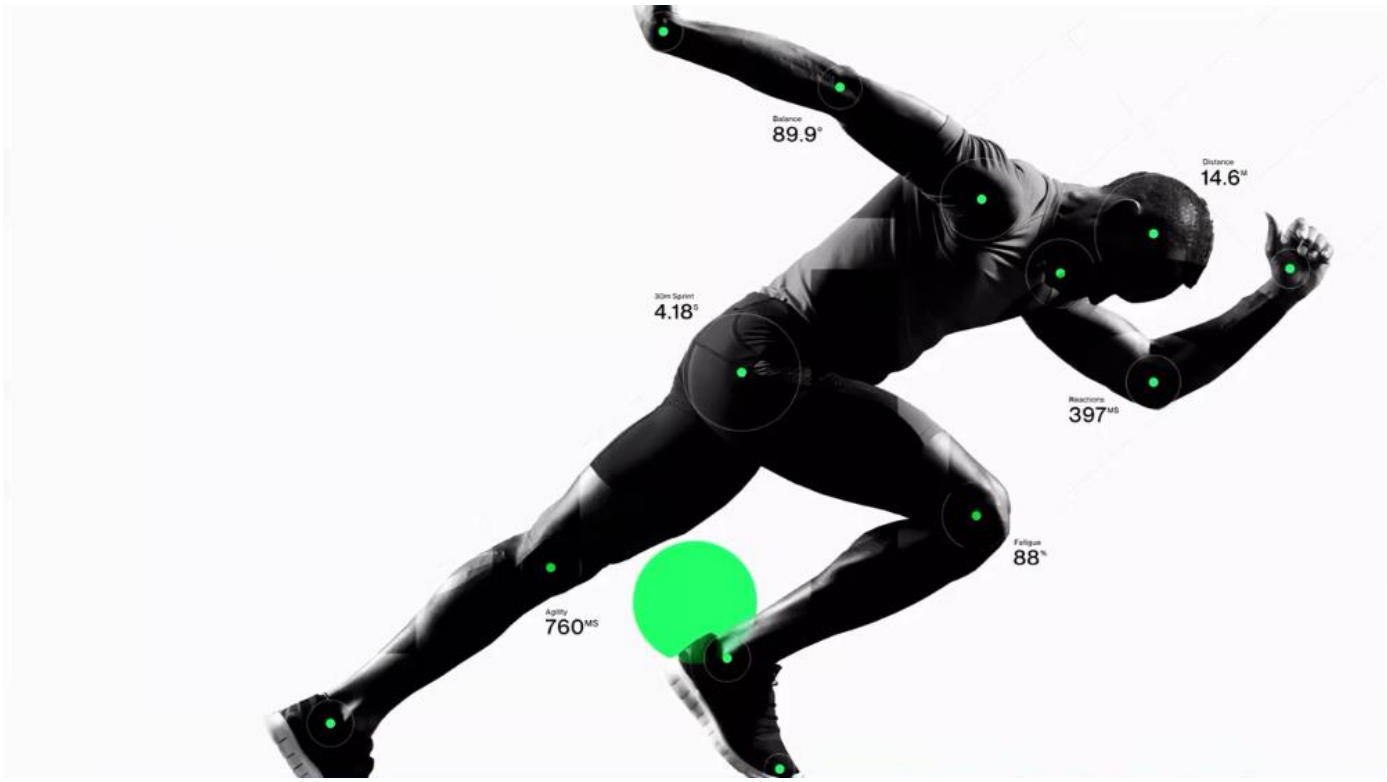


What is it?

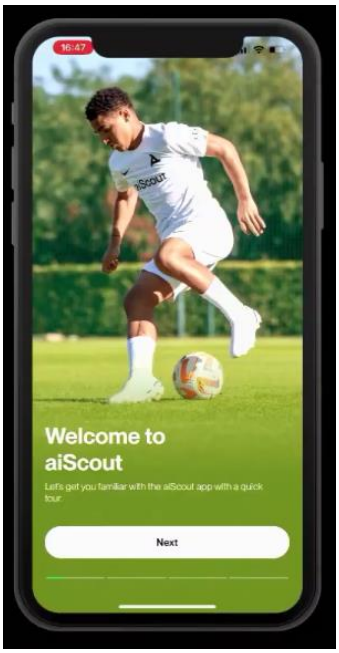
Why does it matter

Where is it used?


ai.io
Data Solutions
aiScout
Talent Analysis & Development App Platform
aiLabs
Elite Performance Labs
3DAT
Computer Vision & Biomechanics Analysis

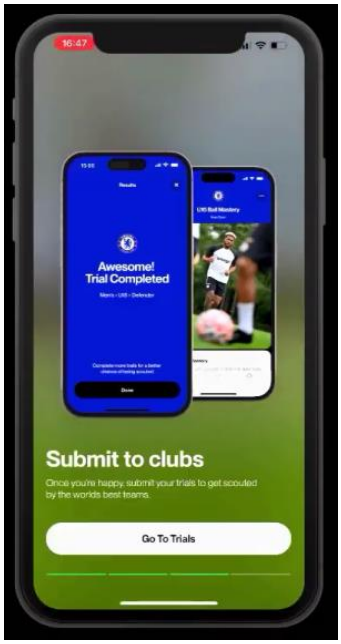

Balance
89.9°
Distance
14.6M
30m Sprint
4.18S
Reactions
397MS
Fatigue
88%
Agility
760MS


aiScout
Cognitive
4.0
Athleticism
5.0
Technique
4.5


Welcome to aiScout
Let's get you familiar with the aiScout app with a quick tour.
Next


Chelsea FC trials
Discover trials
Enter trials for clubs all over the world, including the best teams across the Premier League and MLS.
Next


Complete trial drills
Get a score for each drill within a trial to analyse your skills.
Next

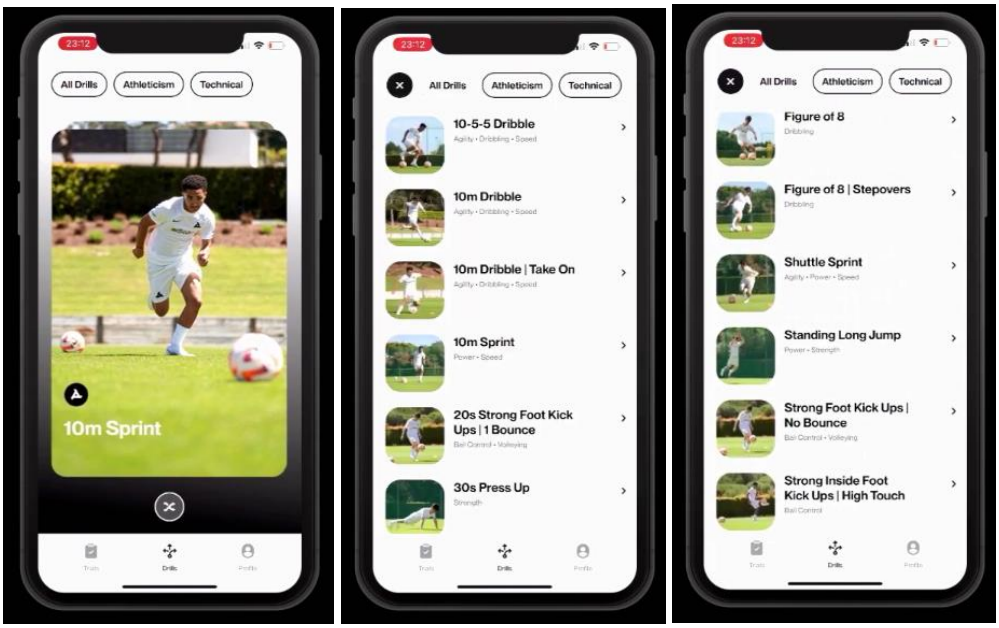

Awesome! Trial Completed
Submit to clubs
Once you're happy, submit your trials to get scouted by the worlds best teams.
Go To Trials


All Trials
Professional Clubs
Private Acader
Chelsea FC
Enter trial now
Trials

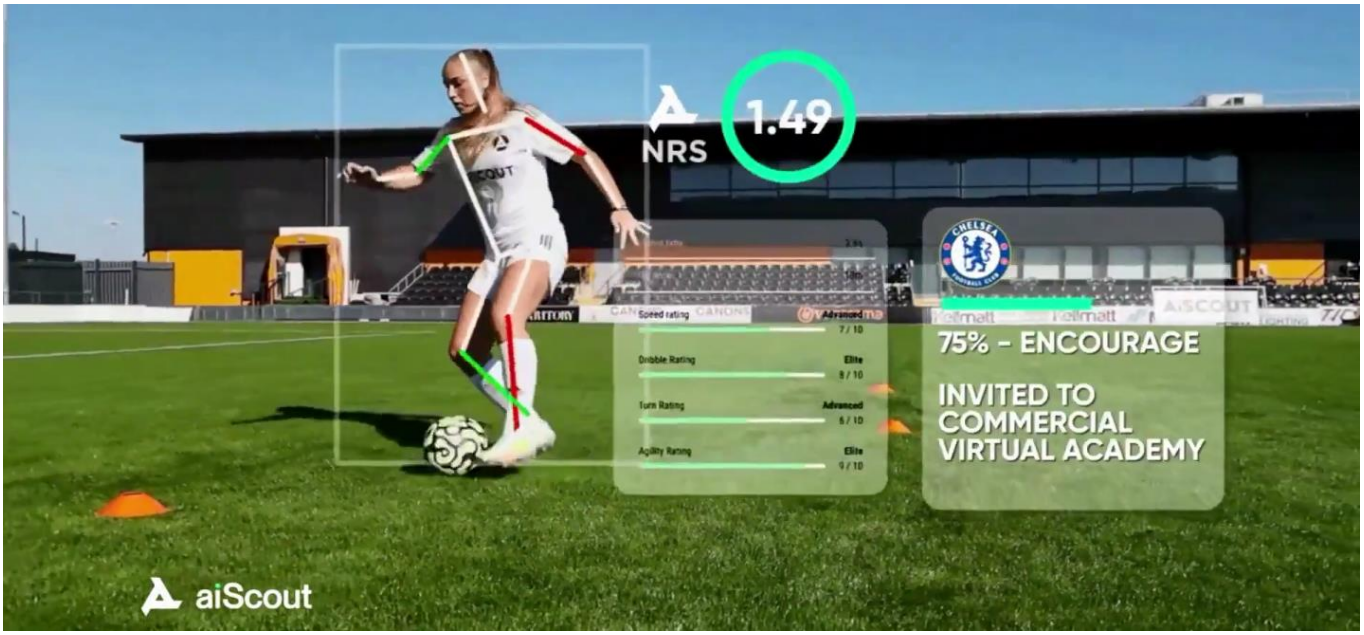

All Trials
Professional Clubs
Private Acader
Latest Trials
Our newest trials available
Touch | Technique
Ages 4 to 14
View All
Trials

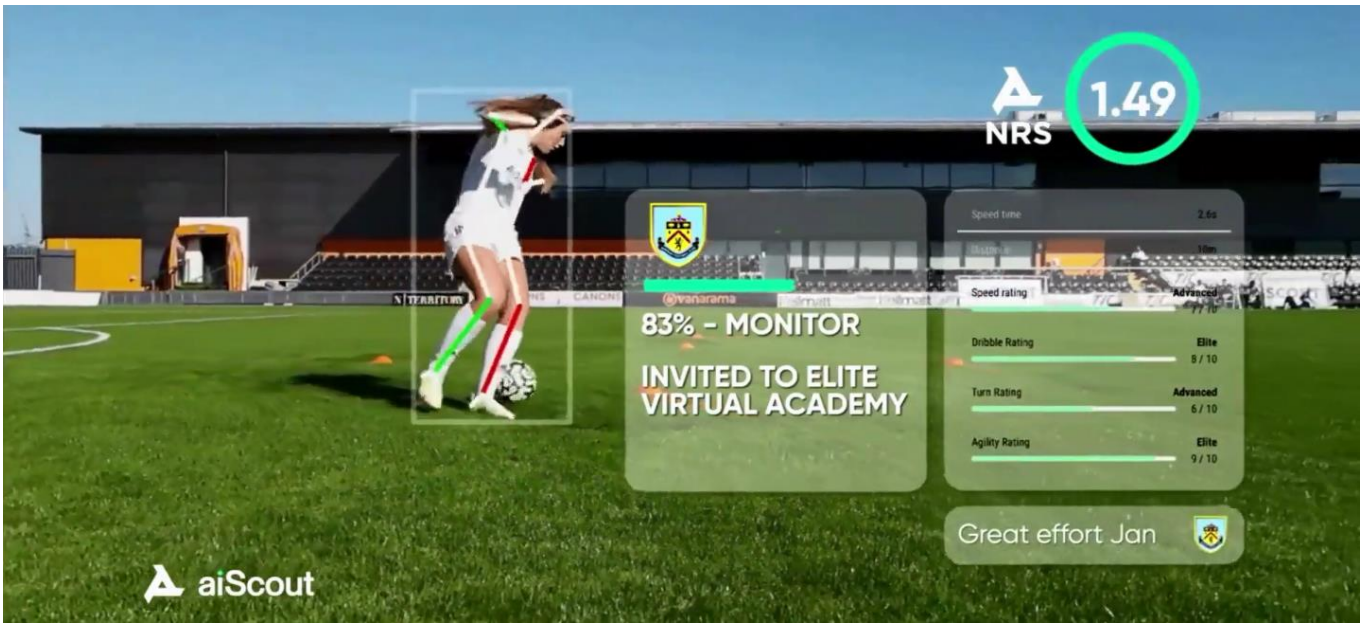

All Drills
Athleticism
Technical
30s Rebound Jump
Drills


All Drills
Athleticism
Technical
10m Dribble
Drills


All Drills
Athleticism
Technical
10m Sprint
10-5-5 Dribble
10m Dribble
10m Dribble | Take On
10m Sprint
20s Strong Foot Kick Ups | 1 Bounce
30s Press Up
Figure of 8
Figure of 8 | Stepovers
Shuttle Sprint
Standing Long Jump
Strong Foot Kick Ups | No Bounce
Strong Inside Foot Kick Ups | High Touch


2.6s
10m
Speed rating
Advanced
7 / 10
Dribble Rating
Elite
8 / 10
Turn Rating
Advanced
6 / 10
Agility Rating
Elite
9 / 10
NRS
1.49
Nice work Jan
aiScout


NRS
1.49
75% - ENCOURAGE
INVITED TO COMMERCIAL VIRTUAL ACADEMY
aiScout

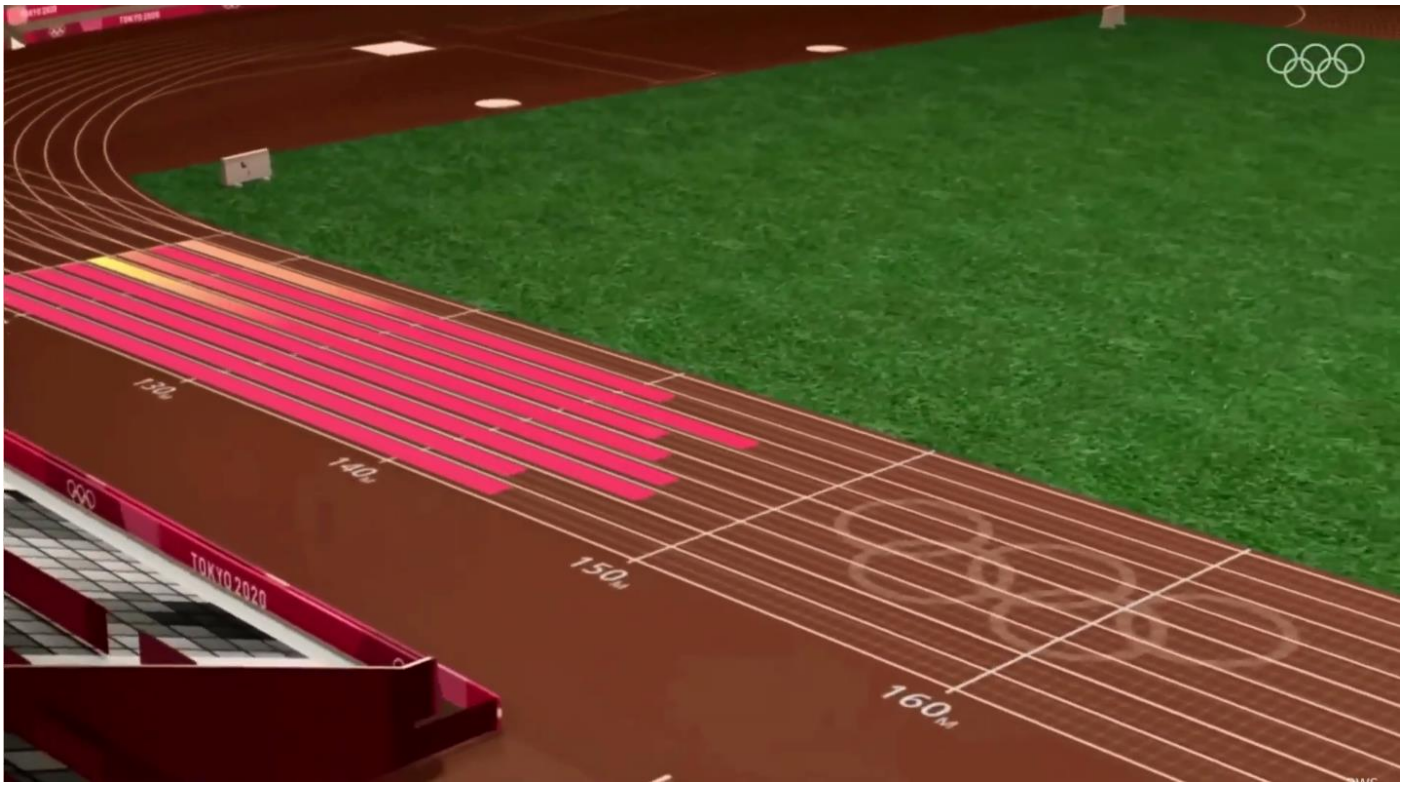

NRS
1.49
83% - MONITOR
INVITED TO ELITE VIRTUAL ACADEMY
Speed time
2.6s
Distance
10m
Speed rating
Advanced
7 / 10
Dribble Rating
Elite
8 / 10
Turn Rating
Advanced
6 / 10
Agility Rating
Elite
9 / 10
Great effort Jan
aiScout

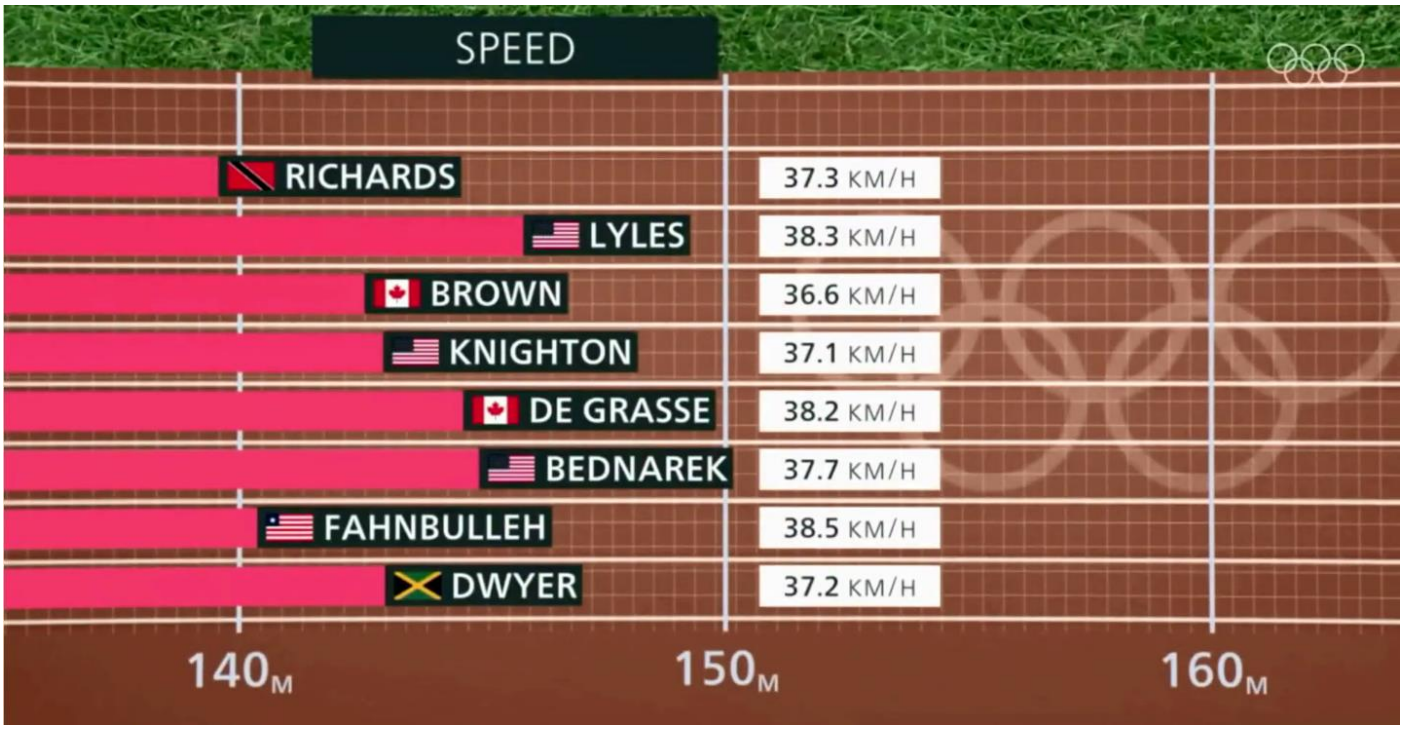

NRS
1.49
100% - TRIAL
Speed time
2.6s
Distance
10m
Speed rating
Advanced
7 / 10
Dribble Rating
Elite
8 / 10
Turn Rating
Advanced
6 / 10
Agility Rating
Elite
9 / 10
Great score Jan
aiScout

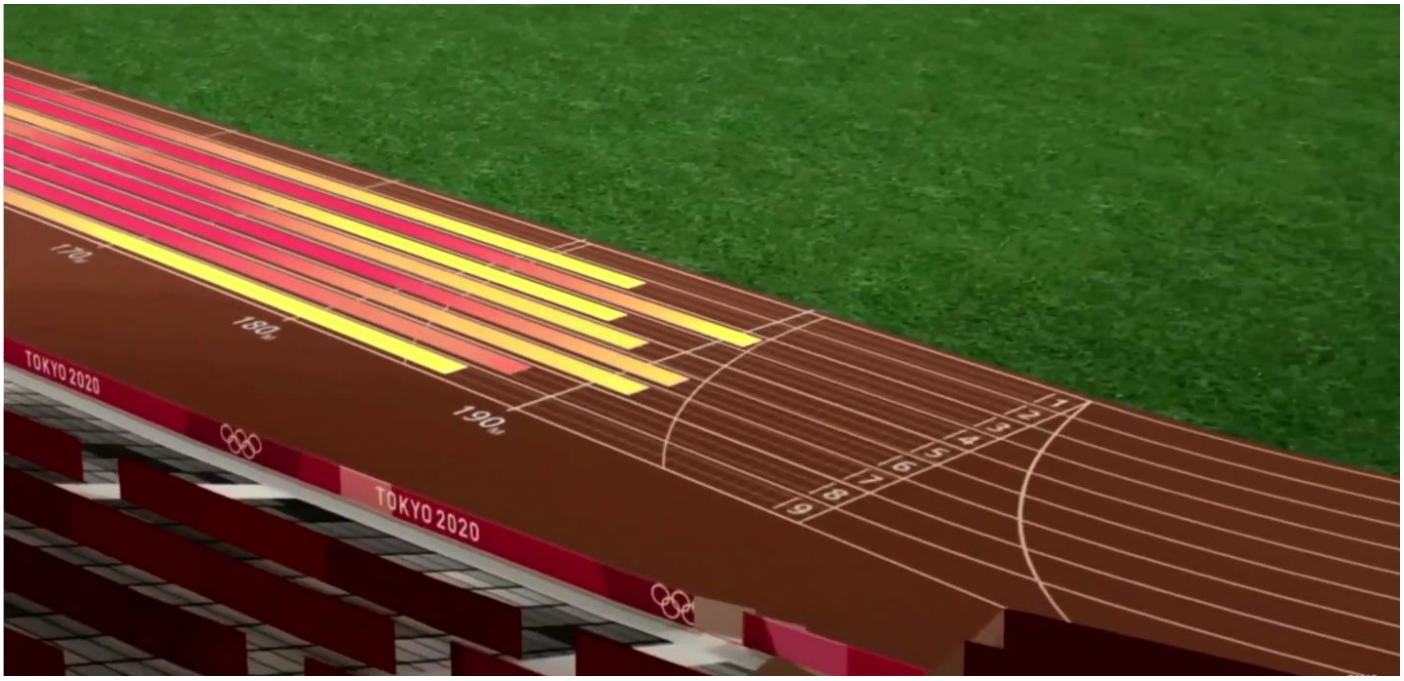

PLAYER SELECTS CLUB VIRTUAL ACADEMIES TO LEARN & DEVELOP WITH
PLAYER CAN ONLY PARTICPATE IN 1 ELITE VIRTUAL ACADEMY
PLAYER CONTINUES TO ENTER TRIALS
PLAYER JOURNEY WILL BE PERSONALISED. SUGGESTED CLUBS, UNIVERSITIES, OPPORTUNITIES, BASED UPON DATA GENERATED
aiScout

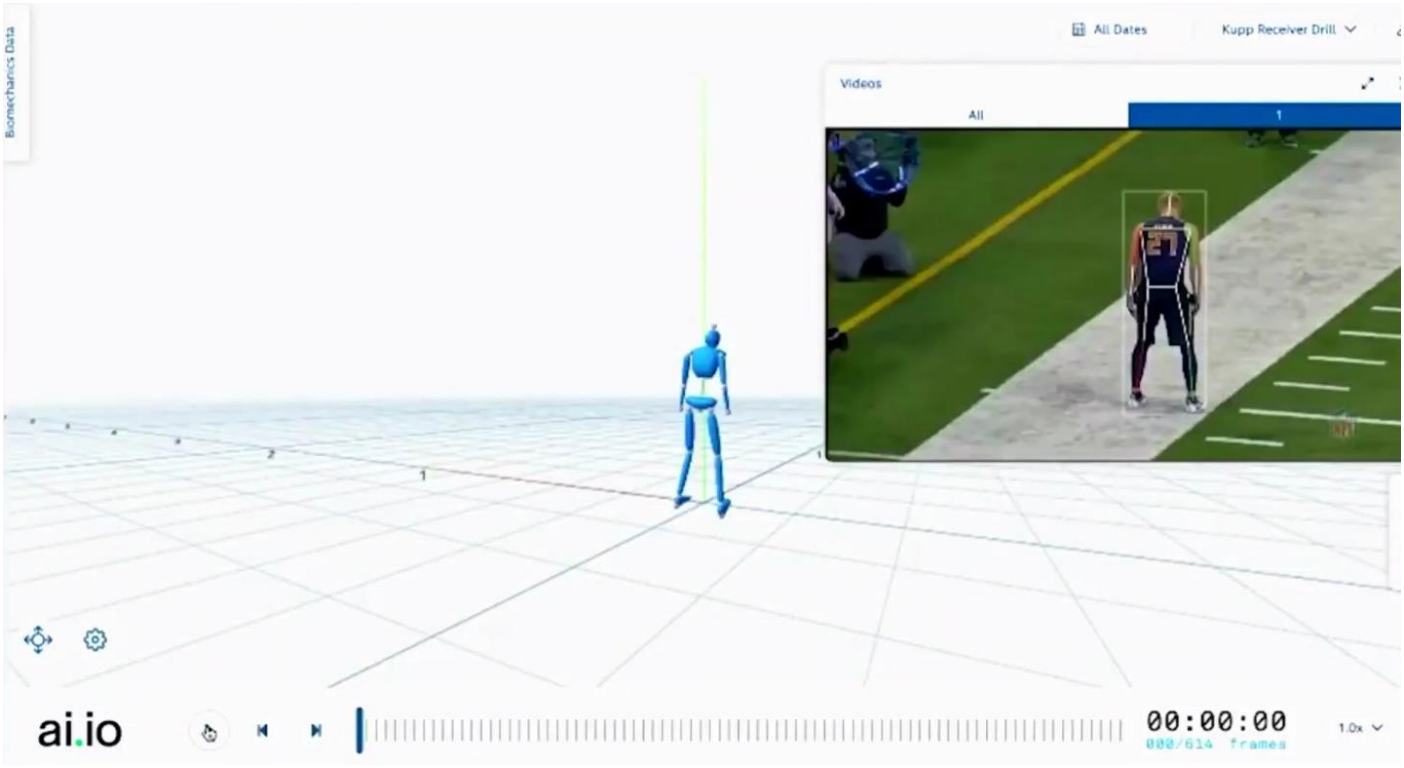
3DAT

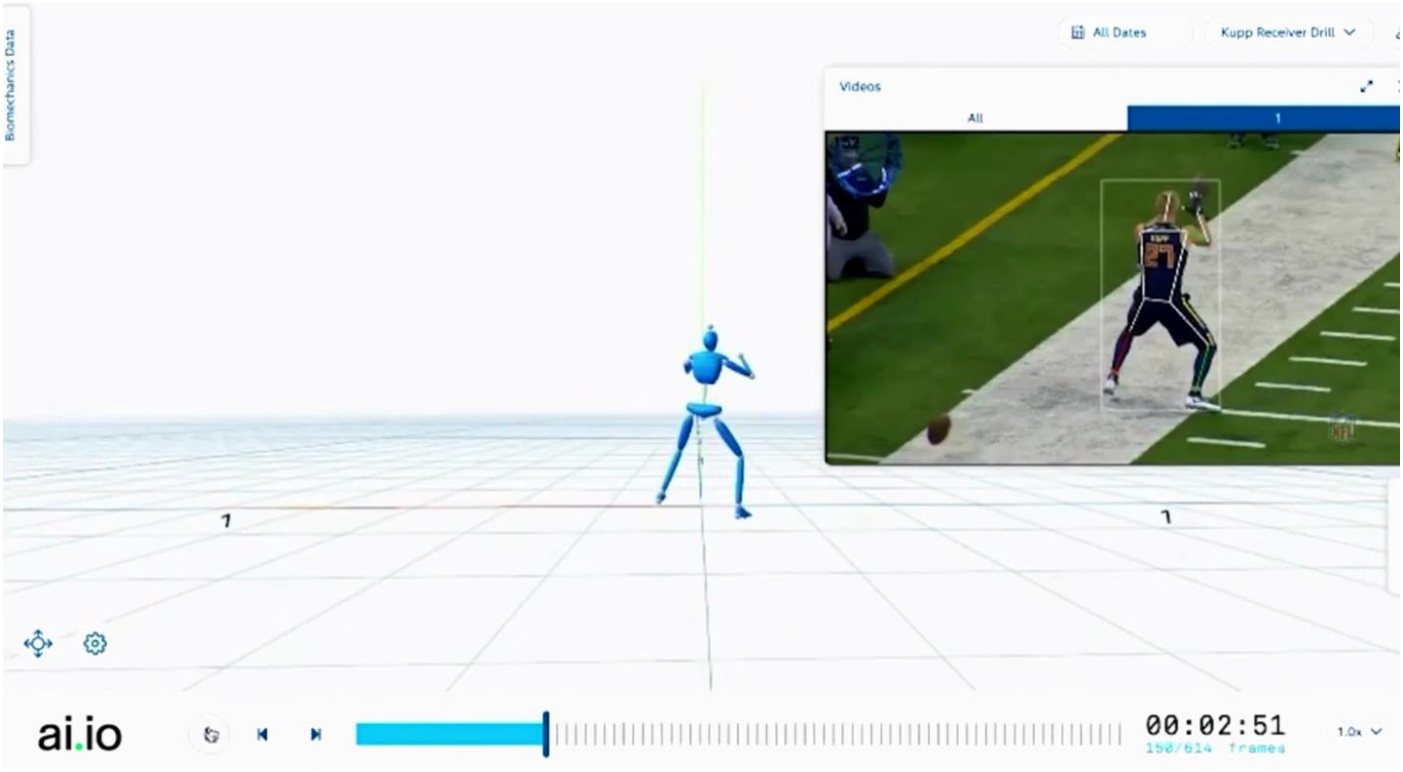
3D Athlete Tracking

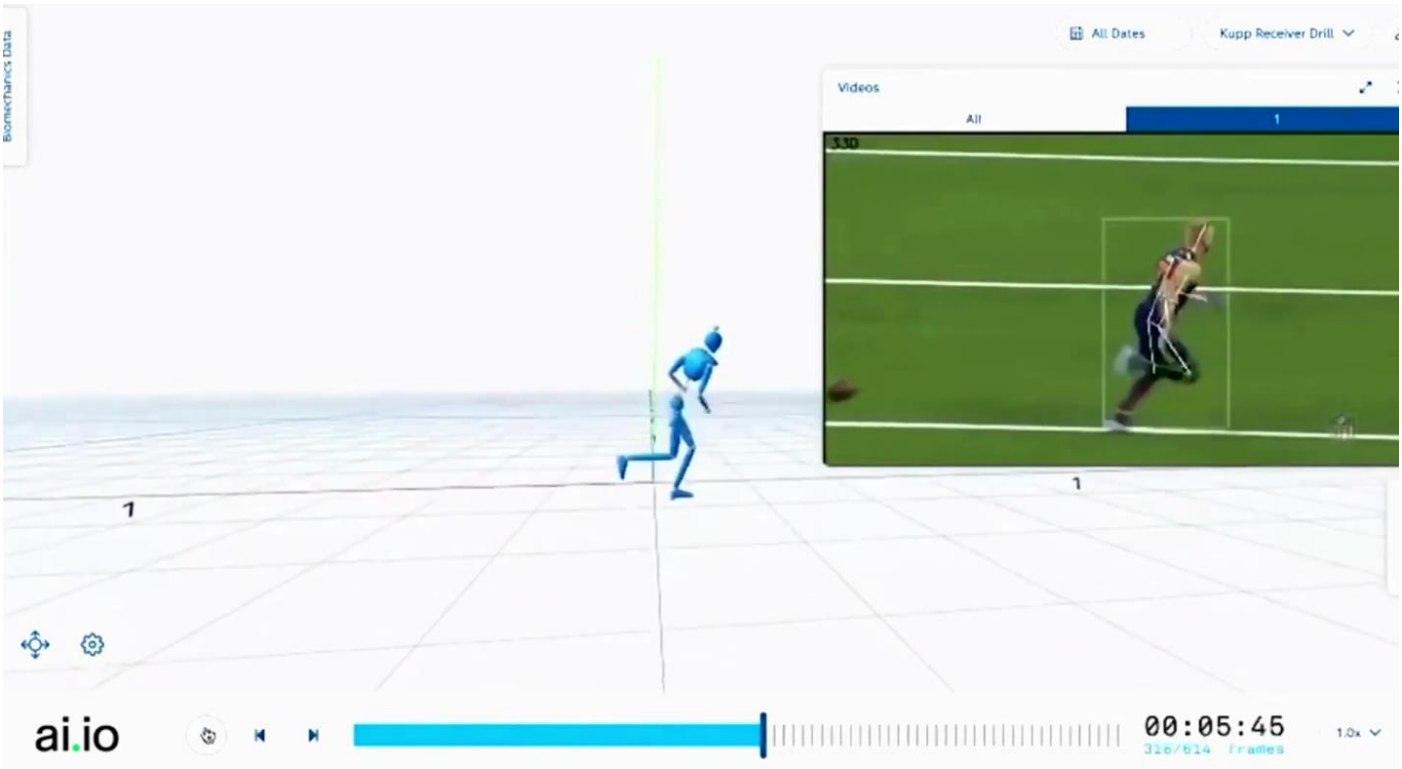

130M
140M
150M
160M
TOKYO2020

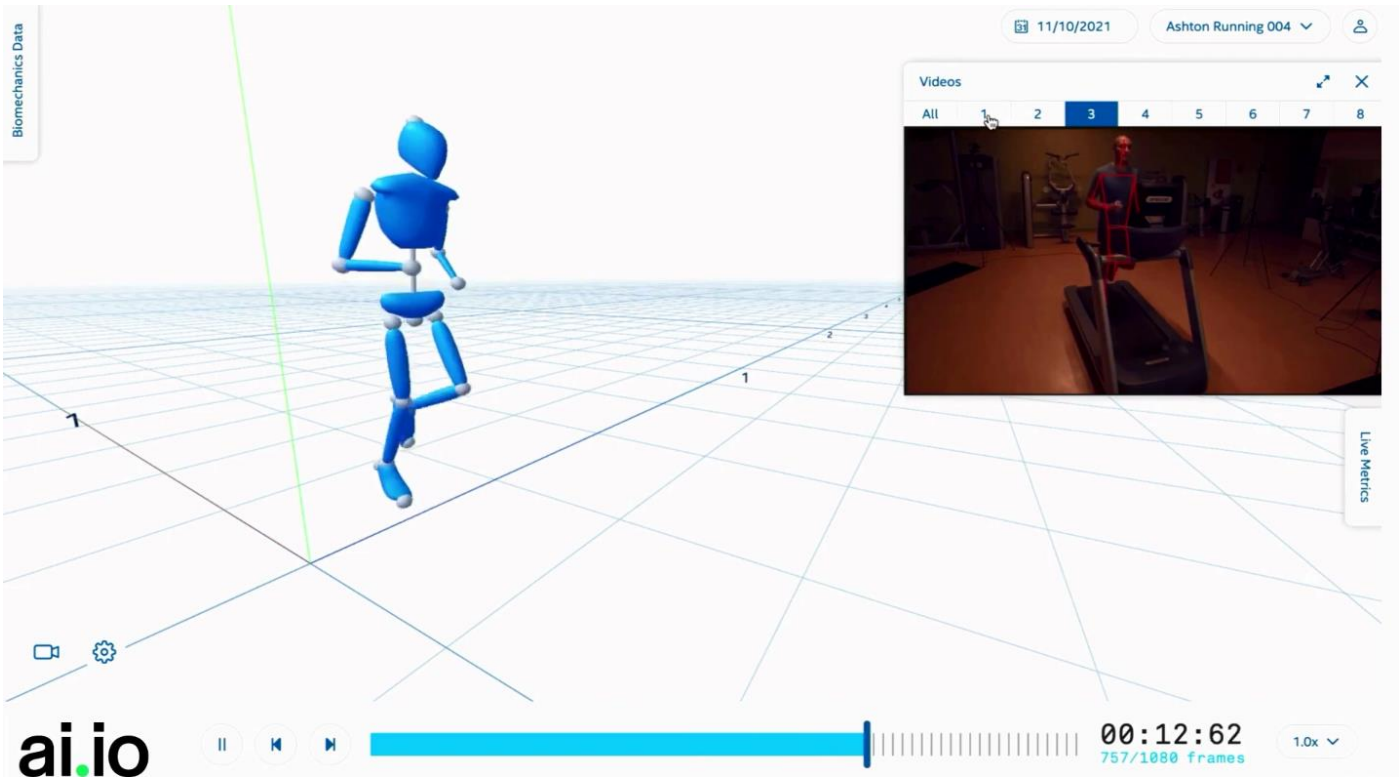

SPEED
RICHARDS
37.3 KM/H
LYLES
38.3 KM/H
BROWN
36.6 KM/H
KNIGHTON
37.1 KM/H
DE GRASSE
38.2 KM/H
BEDNAREK
37.7 KM/H
FAHNBULLEH
38.5 KM/H
DWYER
37.2 KM/H
140M
150M
160M

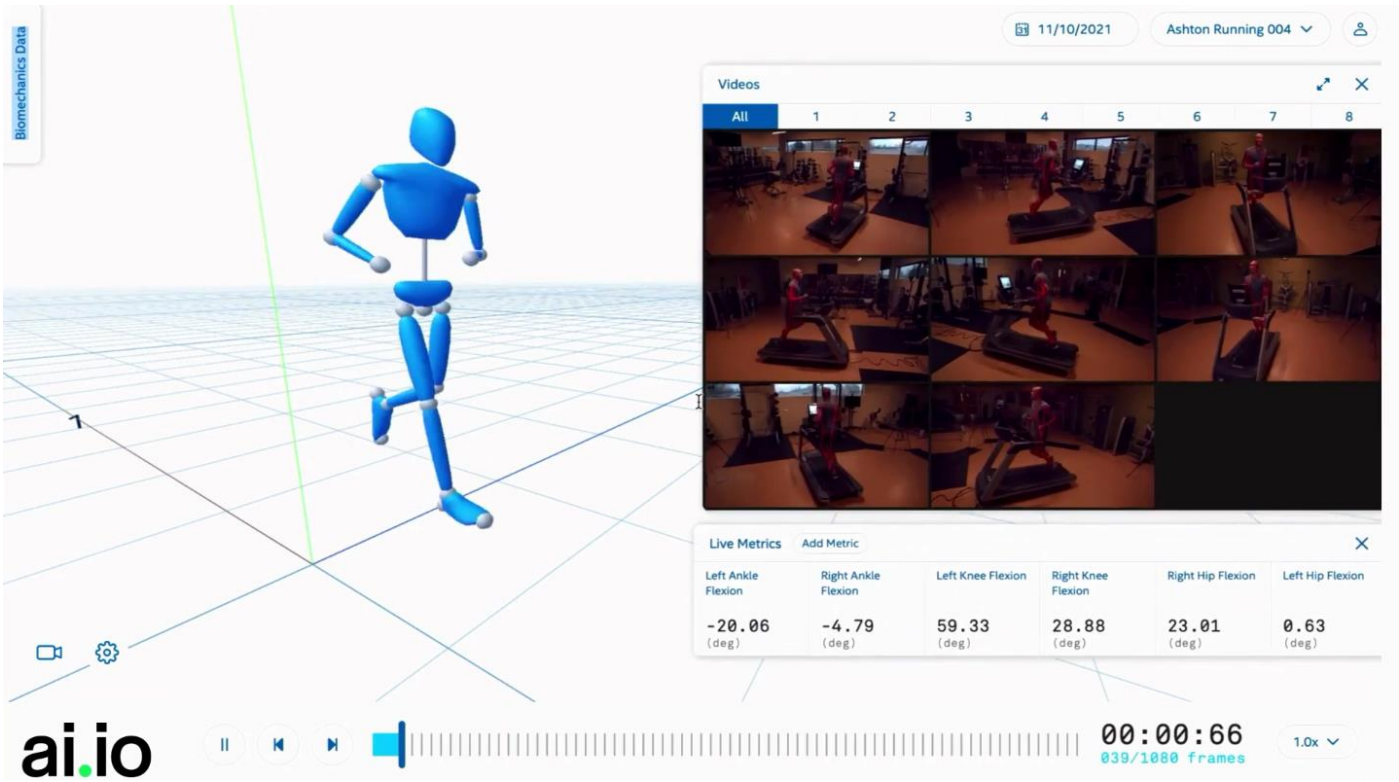

TOKYO 2020
TOKYO 2020

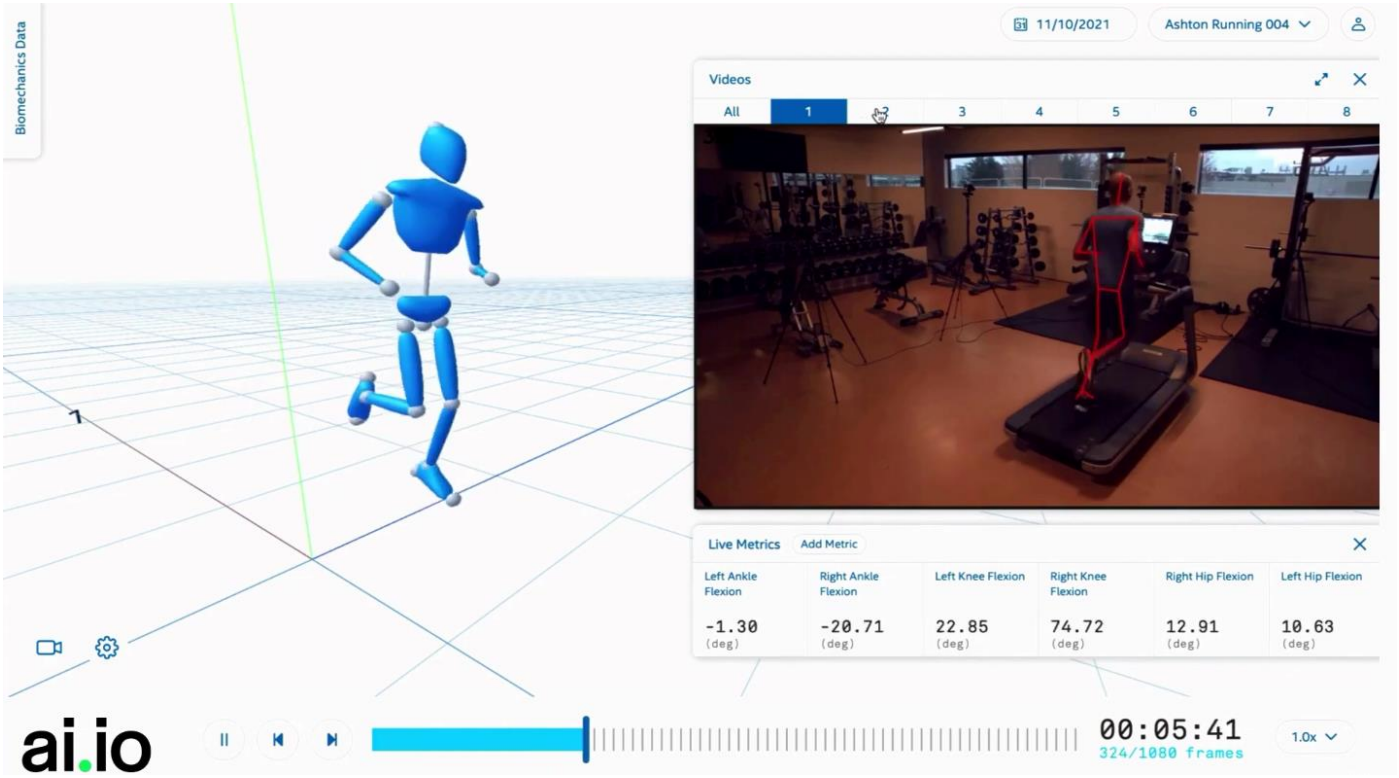

Biomechanics Data
All Dates
Kupp Receiver Drill
Videos
All
1
ai.io
00:00:00
000/614 frames
1.0x

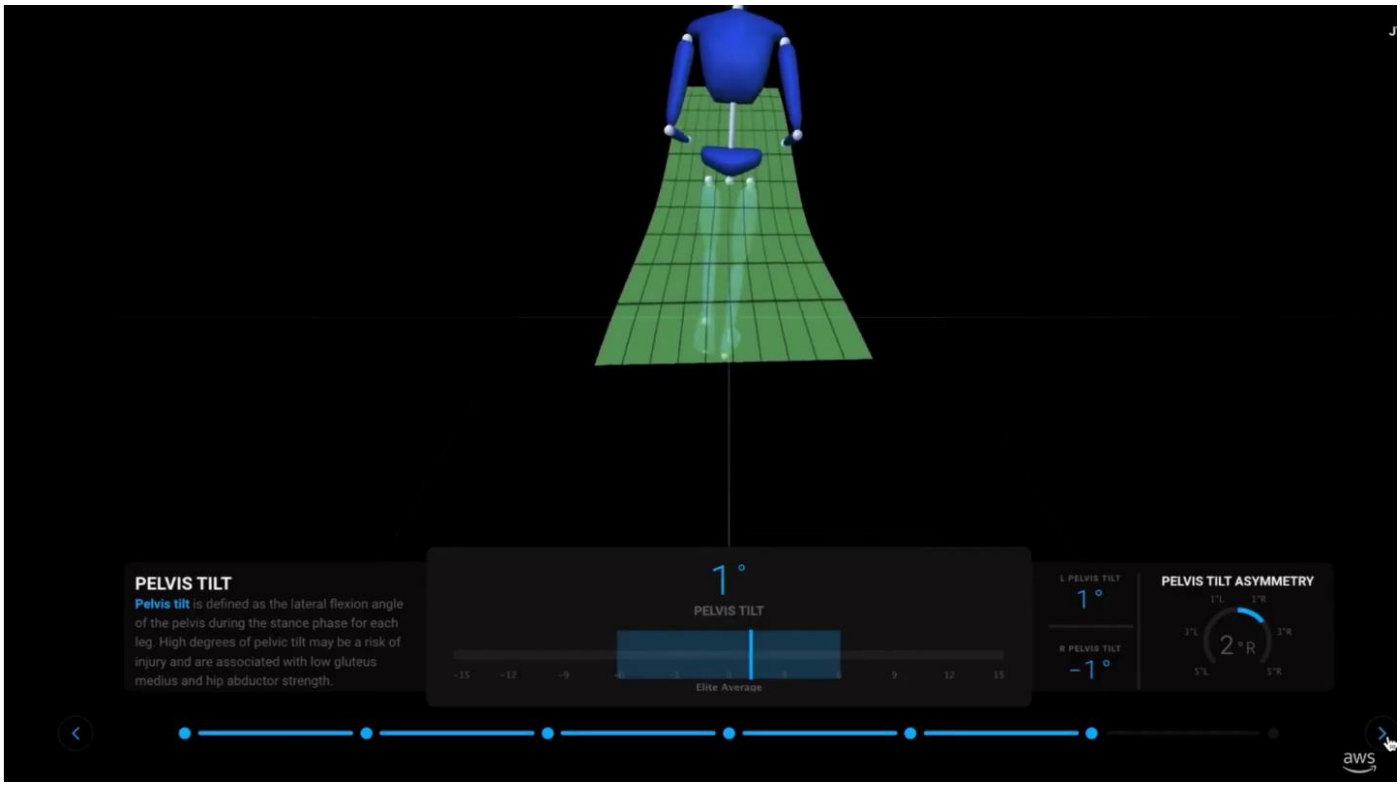

Biomechanics Data
All Dates
Kupp Receiver Drill
Videos
All
1
ai.io
00:02:51
150/614 frames
1.0x

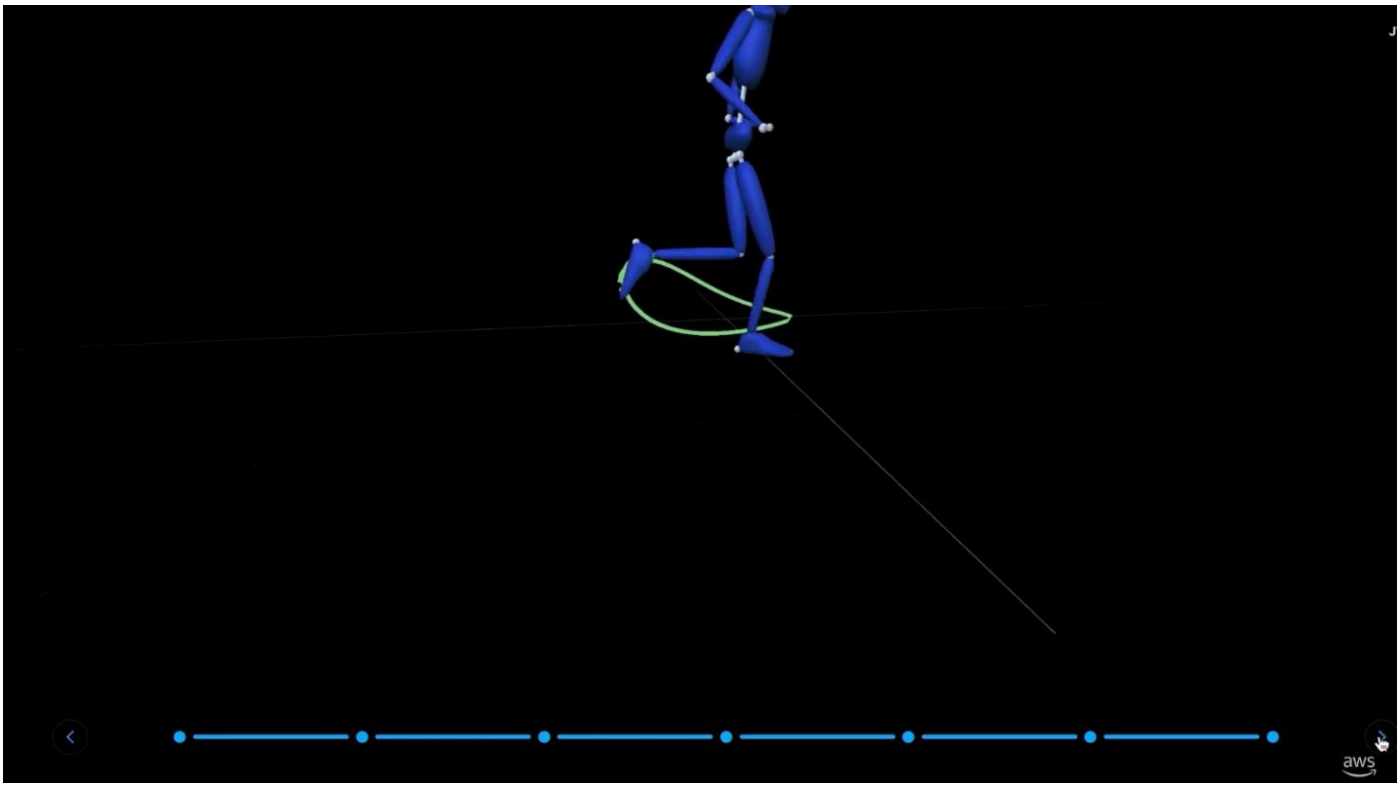

Biomechanics Data
All Dates
Kupp Receiver Drill
Videos
All
1
ai.io
00:05:45
316/614 frames
1.0x

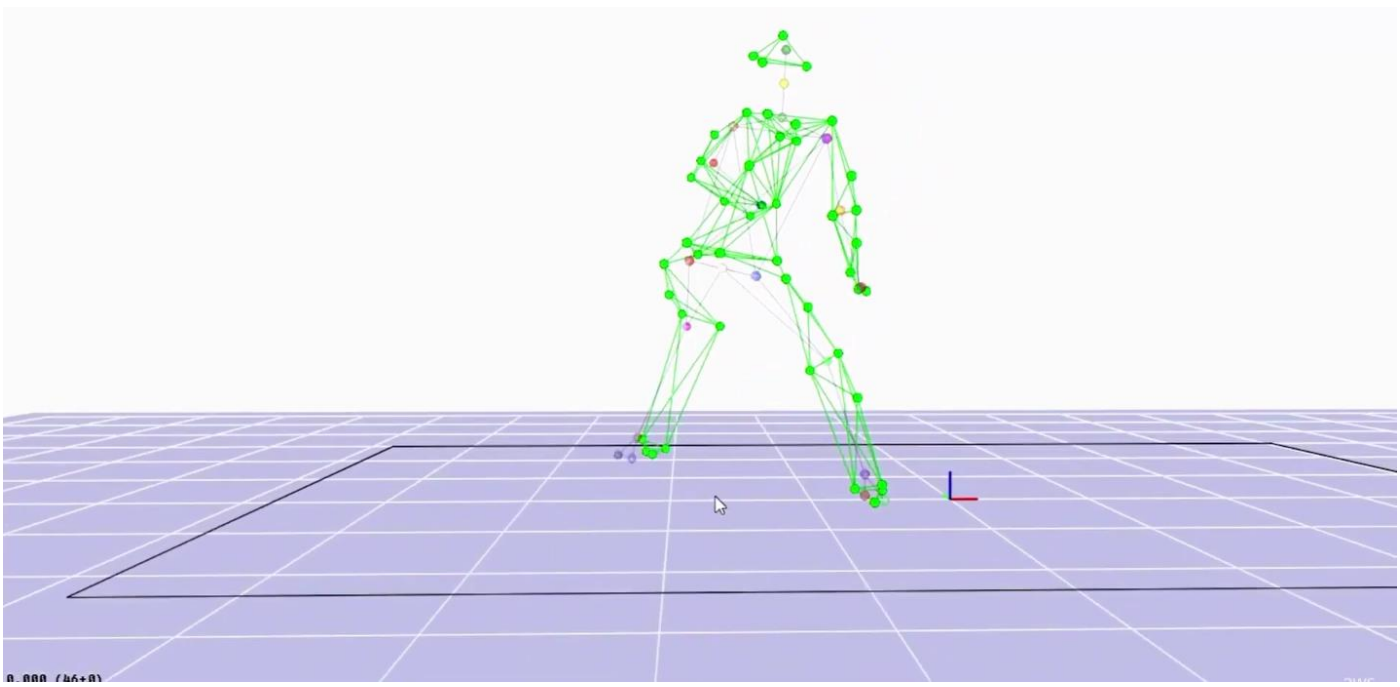

Biomechanics Data
11/10/2021
Ashton Running 004
Videos
All
1
2
3
4
5
6
7
8
Live Metrics
ai.io
00:12:62
757/1080 frames
1.0x

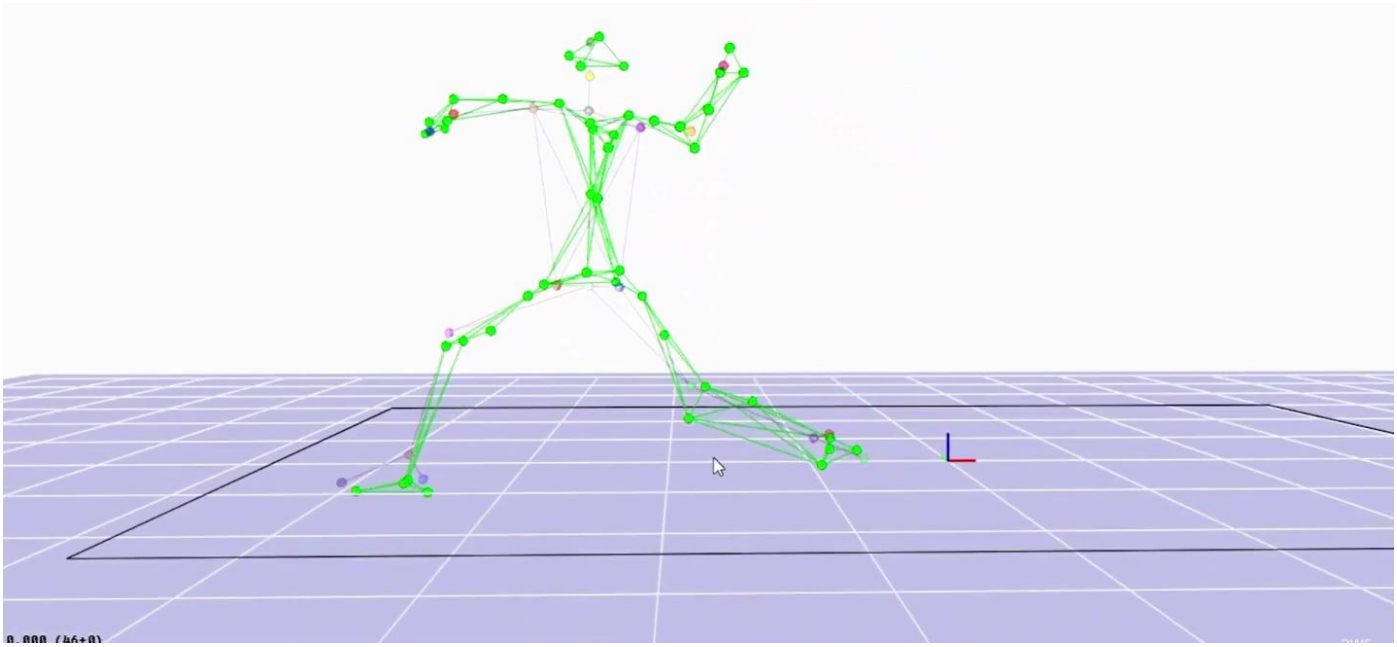

Biomechanics Data
11/10/2021
Ashton Running 004
Videos
All
1
2
3
4
5
6
7
8
Live Metrics
Add Metric
Left Ankle Flexion
-20.06 (deg)
Right Ankle Flexion
-4.79 (deg)
Left Knee Flexion
59.33 (deg)
Right Knee Flexion
28.88 (deg)
Right Hip Flexion
23.01 (deg)
Left Hip Flexion
0.63 (deg)
ai.io
00:00:66
039/1080 frames
1.0x

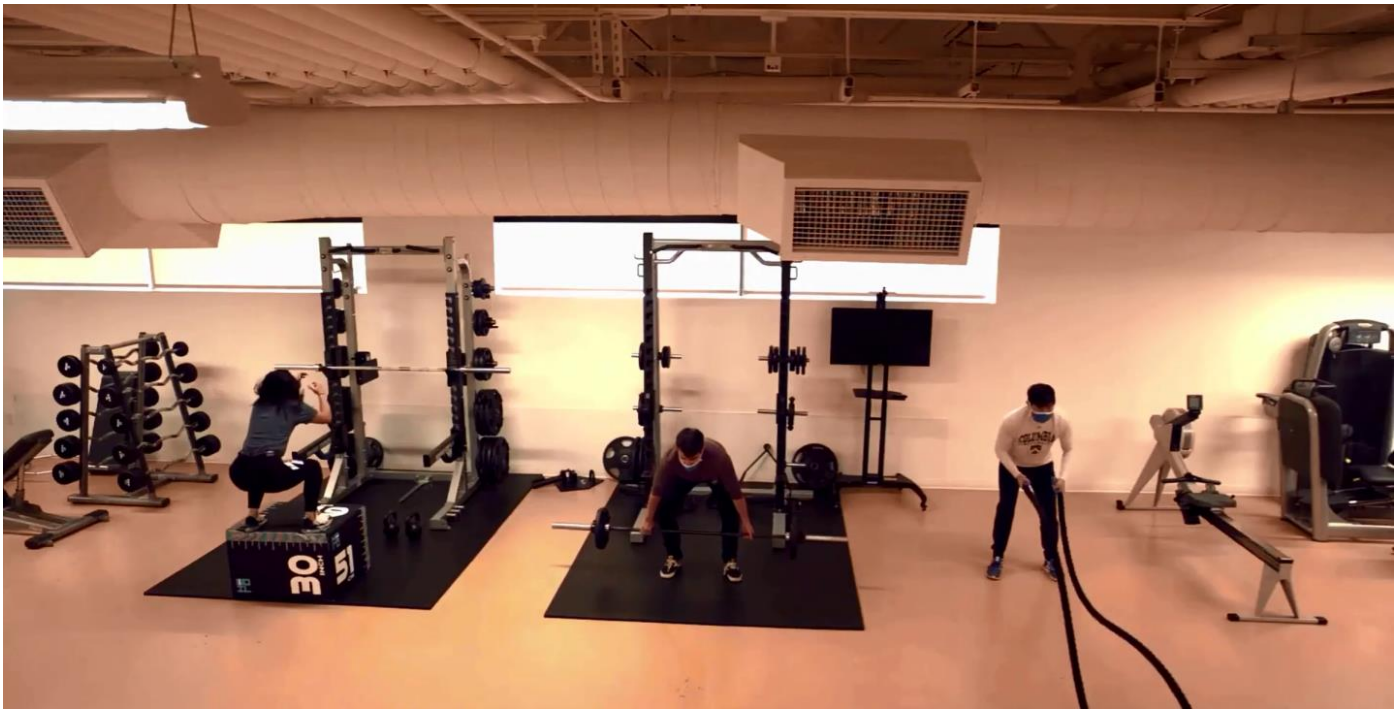

Biomechanics Data
11/10/2021
Ashton Running 004
Videos
All
1
3
4
5
6
7
8
Live Metrics
Add Metric
Left Ankle Flexion
-1.30 (deg)
Right Ankle Flexion
-20.71 (deg)
Left Knee Flexion
22.85 (deg)
Right Knee Flexion
74.72 (deg)
Right Hip Flexion
12.91 (deg)
Left Hip Flexion
10.63 (deg)
ai.io
00:05:41
324/1080 frames
1.0x


PELVIS TILT
Pelvis tilt is defined as the lateral flexion angle of the pelvis during the stance phase for each leg. High degrees of pelvic tilt may be a risk of injury and are associated with low gluteus medius and hip abductor strength.
1°
PELVIS TILT
Elite Average
L PELVIS TILT
1°
R PELVIS TILT
-1°
PELVIS TILT ASYMMETRY
2°R
aws


aws

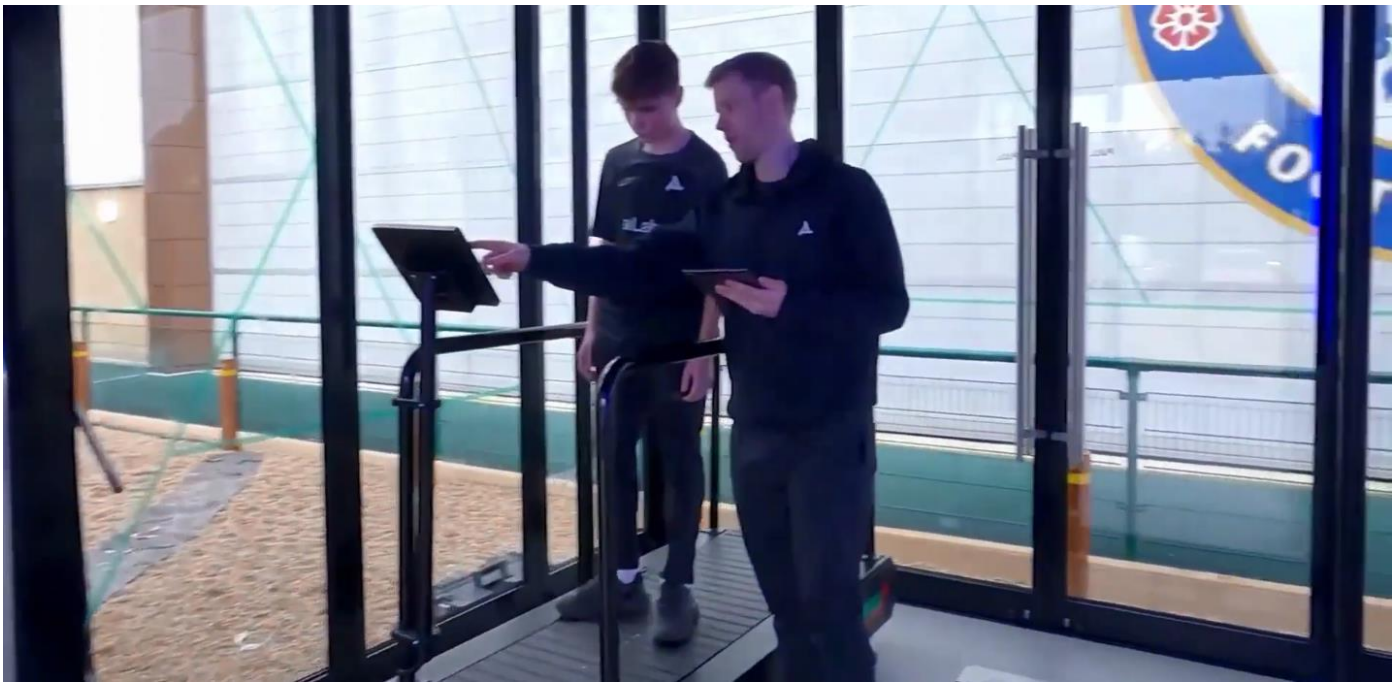


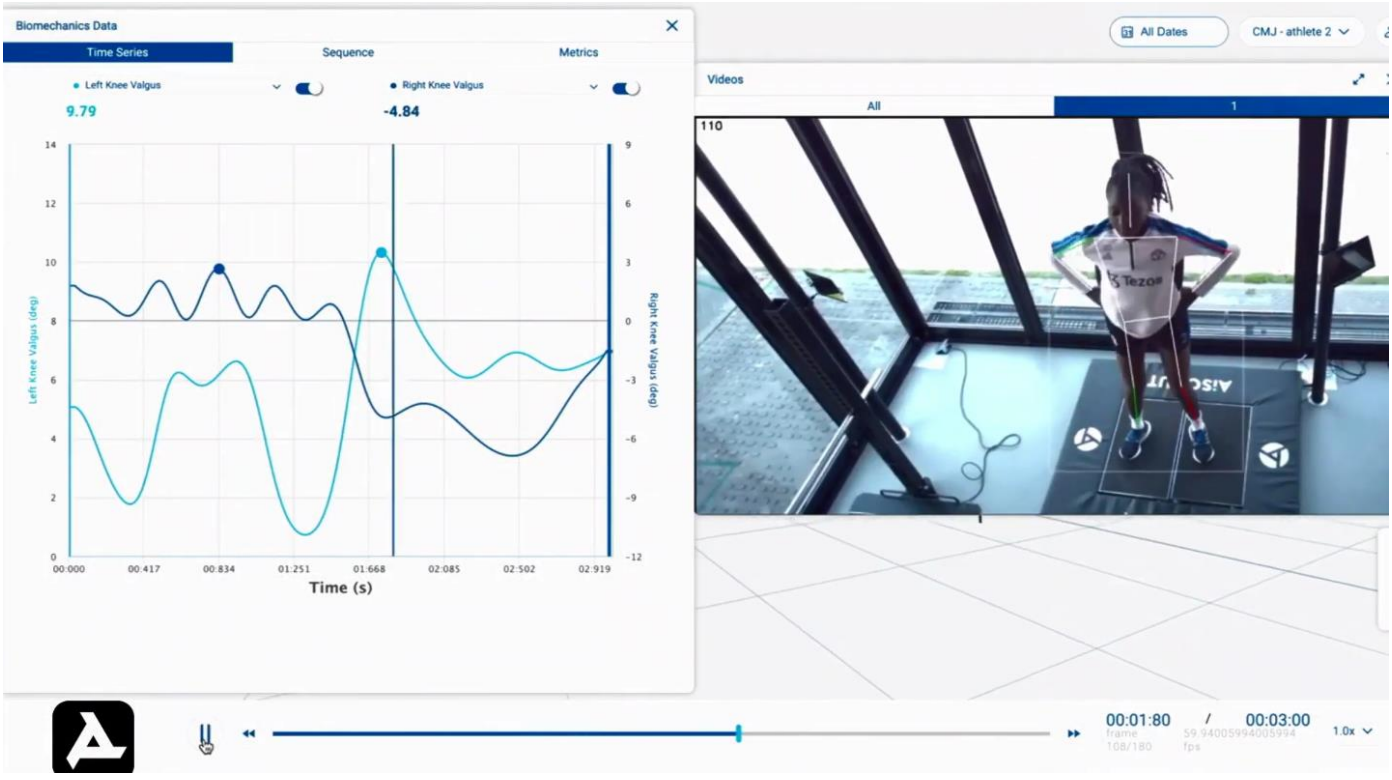


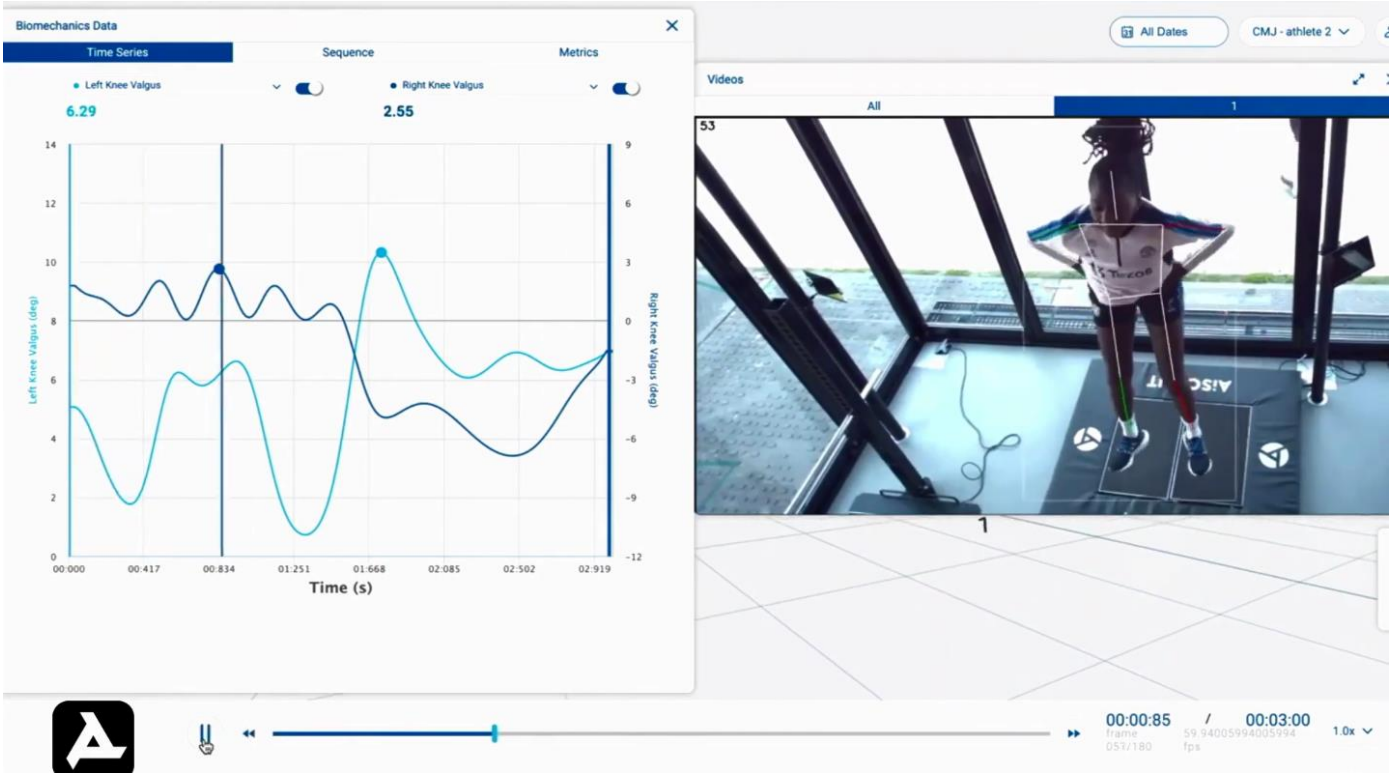


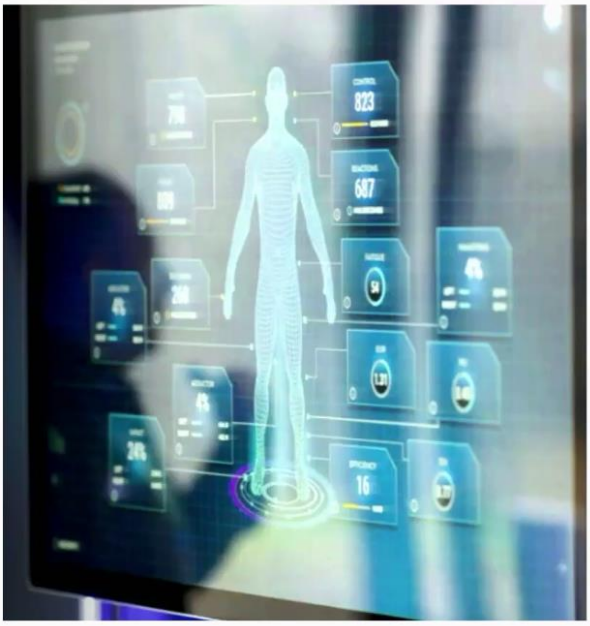

aiLabs
Elite Performance Lab

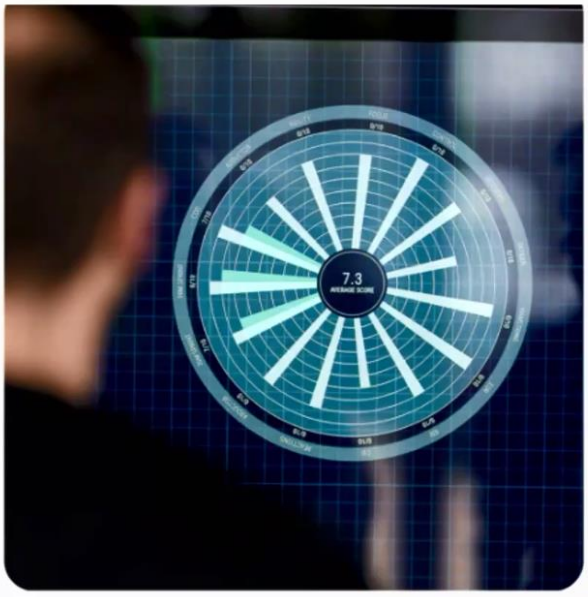

aiLabs

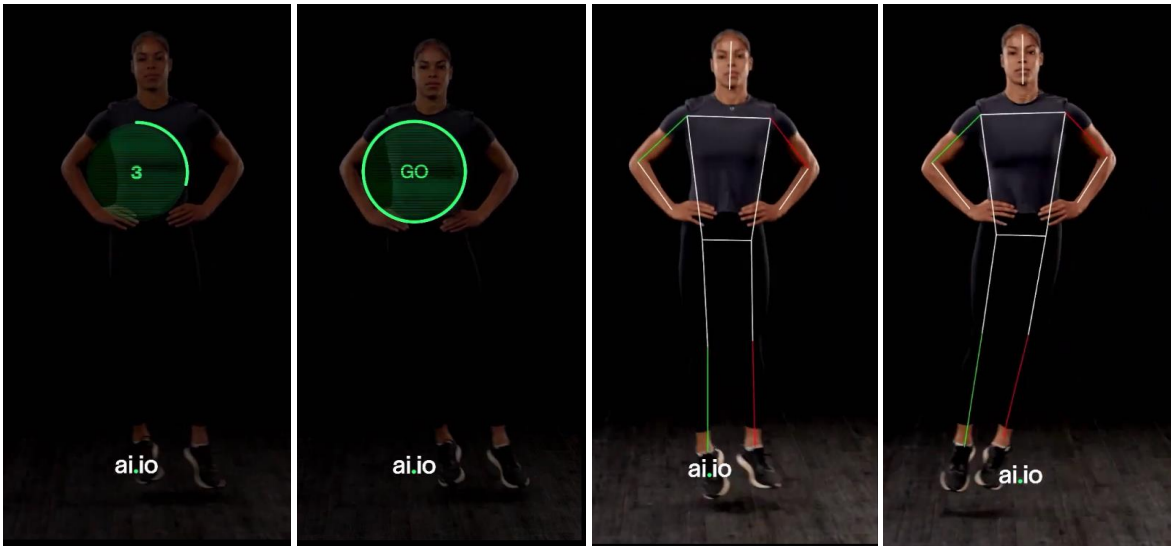


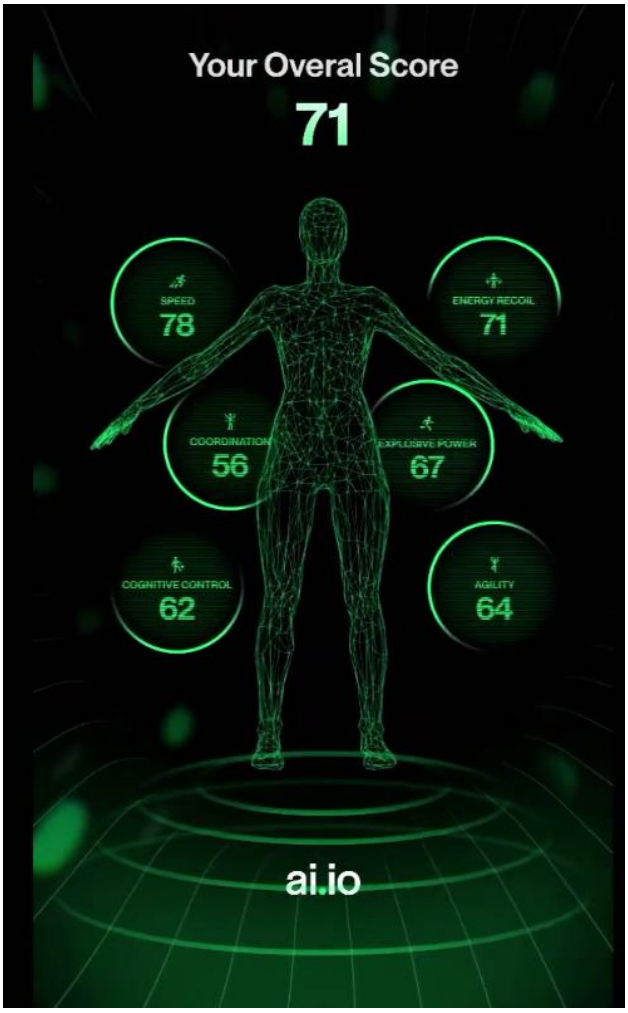

Biomechanics Data
Time Series
Sequence
Metrics
Left Knee Valgus
Right Knee Valgus
9.79
-4.84
Left Knee Valgus (deg)
Right Knee Valgus (deg)
Time (s)
00:000
00:417
00:834
01:251
01:668
02:085
02:502
02:919
All Dates
CMJ - athlete 2
Videos
All
1
110
00:01:80
/
00:03:00
1.0x


Biomechanics Data
Time Series
Sequence
Metrics
Left Knee Valgus
Right Knee Valgus
6.29
2.55
Left Knee Valgus (deg)
Right Knee Valgus (deg)
Time (s)
00:000
00:417
00:834
01:251
01:668
02:085
02:502
02:919
All Dates
CMJ - athlete 2
Videos
All
1
53
00:00:85
/
00:03:00
1.0x




7.3


3
GO
ai.io
ai.io
ai.io
ai.io


Your Overal Score
71
SPEED
78
ENERGY RECOIL
71
COORDINATION
56
EXPLOSIVE POWER
67
COGNITIVE CONTROL
62
AGILITY
64
ai.io

GPU Accelerated Computer Vision

Learn more on the Intel Booth

Visit us at our booth: Intel Booth #750

it starts

intel

Intel Data Center GPU

FLEX SERIES 140

4 Media Engines	75w Power Envelope	16 Ray Tracing Units
Xe HPG Architecture	Half Height PCIe	16 Xᵉ cores

Optimized for lower TCO

5X

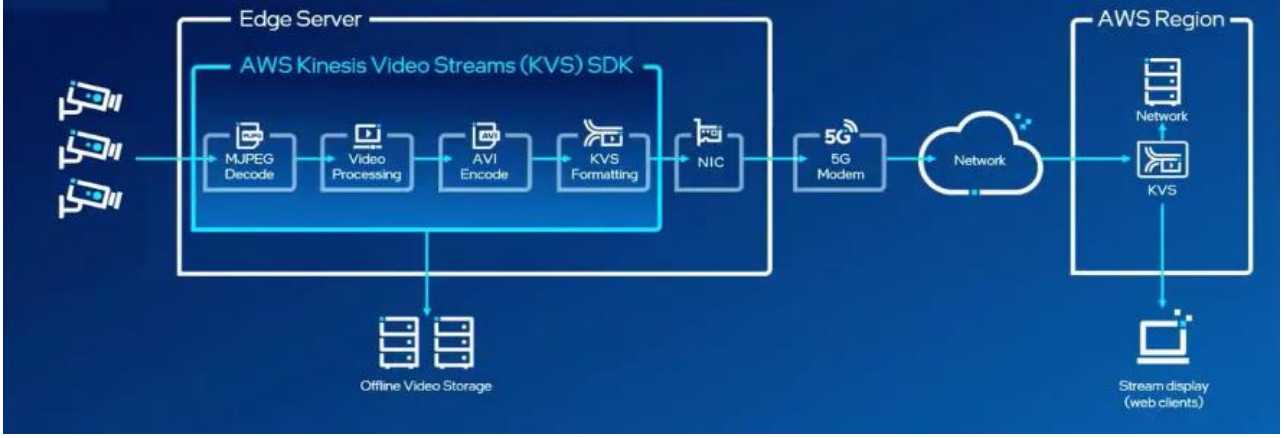
Media transcode throughput at half the power !
Intel Flex 140 GPU compared to NVIDIA A10

Media Deliver
30
1080p 60fps Strea

Cloud Gamin
40
720p 30fps Ga
Strea

See performance claims https://edc.intel.com/content/www/us/en/products/performance/benchmarks/intel-data-center-gpu-flex-series/

Solution Architecture



Industrial Computers
- Small Form Factor
- Fanless or Active Cooling
- 0 to 50°C Operating Temp.

Rugged Computers
- Resistant to Shock & Vibration
- Wide Power Input Range
- -40 to 70°C Operating Temp.

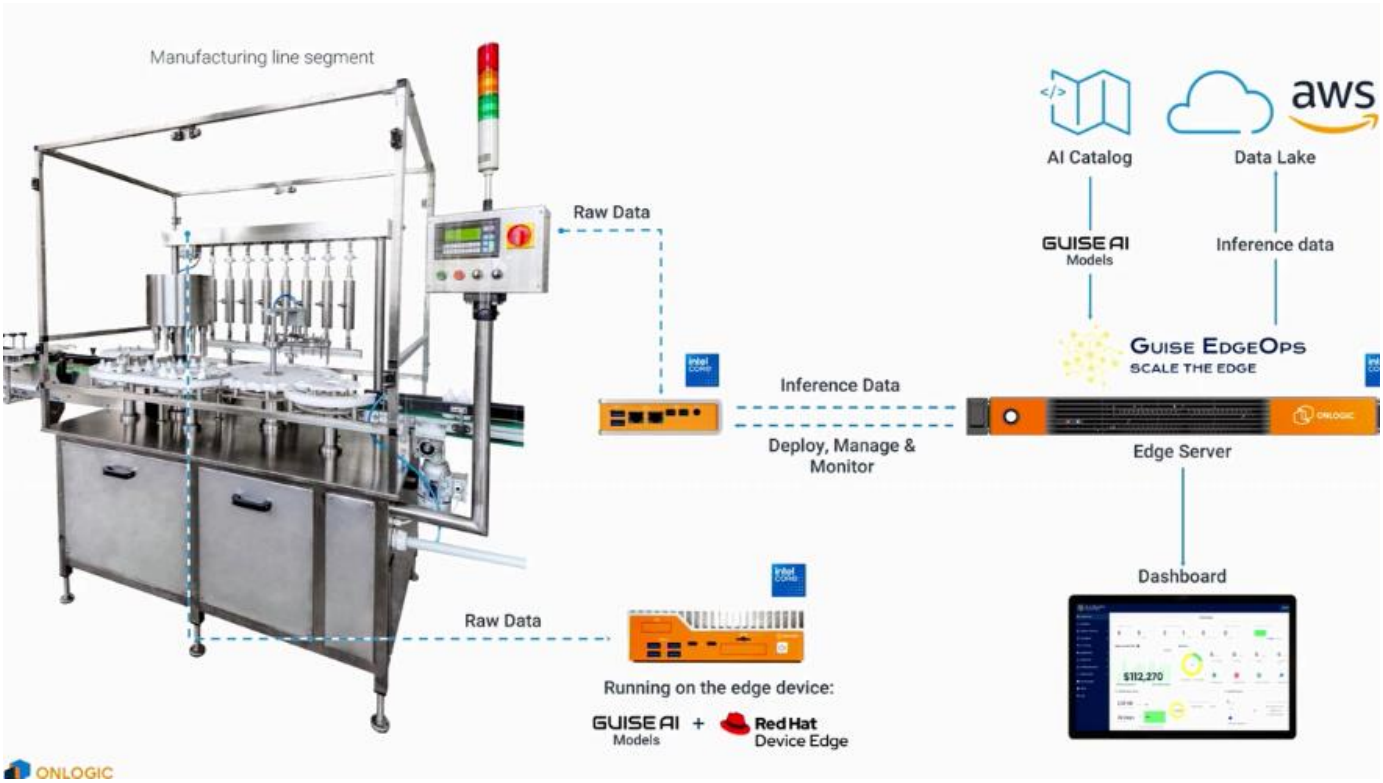
Panel PCs
- 8.4" to 24" Screen Sizes
- Resistive or Capacitive Screens
- Available with IP65 Front Bezels

Edge Servers
- 1U to 4U Sizes
- Advanced Intel Processing
- Highly Configurable

Collaborations that make **AI at the Edge** easy

Cloud Services	aws
Application Software	OpenVINO, GUISE AI AI AT THE EDGE, inductive automation.
Edge Orchestration	OPENSHIFT, inductive automation., THINMANAGER
Operating System	Red Hat, ubuntu
Devices | Hardware | Accelerators	intel., intel ARC, Raspberry Pi

*Only a subset of our partners listed



OnLogic Demos at re:Invent

Intel Booth - KVS Demo:
Expo Hall - Venetian Conference Center, Booth 750

- Running Amazon Kinesis to process HD video streams and send data to the cloud.

Builders' Fair – Find My Ship:
Expo Hall - Venetian Conference Center

- Controlling AWS IoT Shadow updates along with view of Ultra-Wide Band Data visualization in 3D.

AWS IoT Kiosk – AWS IoT DeepRacer Vx Demo:
Expo Hall - Venetian Conference Center

- Showcasing Greengrass Runtime Software at the edge to get car data (via CAN Bus) and visualize it.

AWS Disaster Response DDIL activation:
Venetian Hotel Lobby - Disaster Response Jeep area

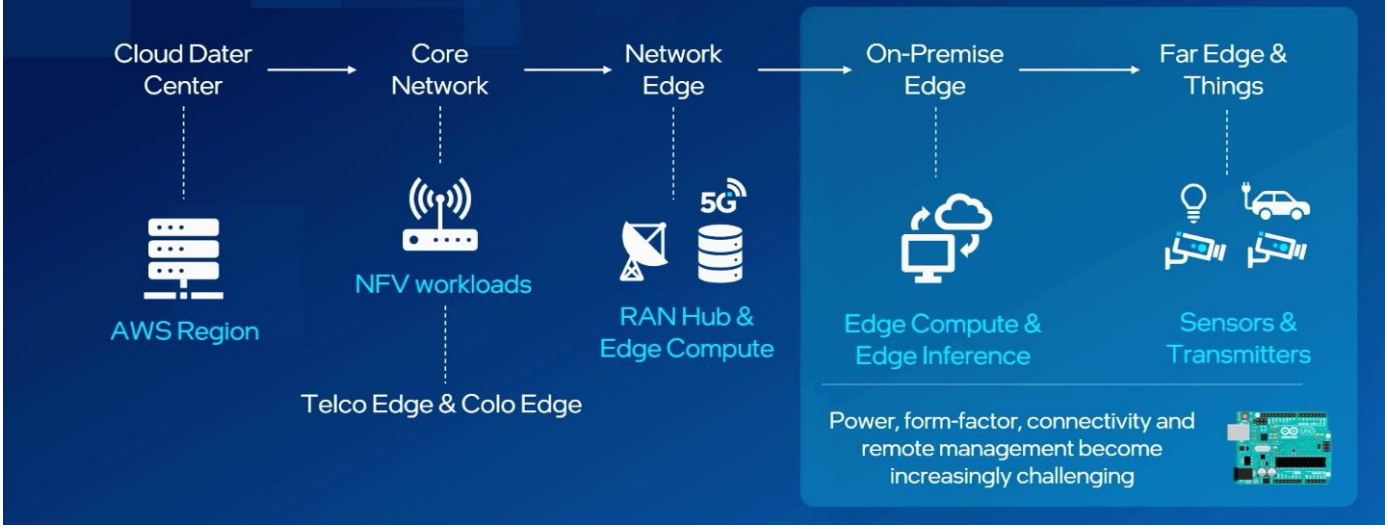
- Aggregating the GPS locations and metadata from 60 LoRaWAN GPS trackers deployed to re:Invent staff.

AWS GenAI Chess activation:
Rec Center – Mandalay Bay Convention Center

- Used for AWS IoT Shadow and Jobs update for each Chess move of the robots and to perform a Trust and Verify check between the physical chess board and the output of the GenAI model.

ONLOGIC

Expanding computer vision to the far edge



ARDUINO PRO

ARDUINO CLOUD

EDGE IMPULSE

Arduino Cloud for Business Machine Learning Tools

Machine Learning Tools: Endless possibilities

CLOUD

EDGE IMPULSE

Collect	Design	Test	Deploy
Acquire valuable data securely from your devices and rapidly build custom dataset	Develop algorithms with ready-to-use digital signal processors and machine learning blocks	Validate and train Machine Learning models with real-time data	Build optimized embedde inference

Machine Learning Tools at a Glance

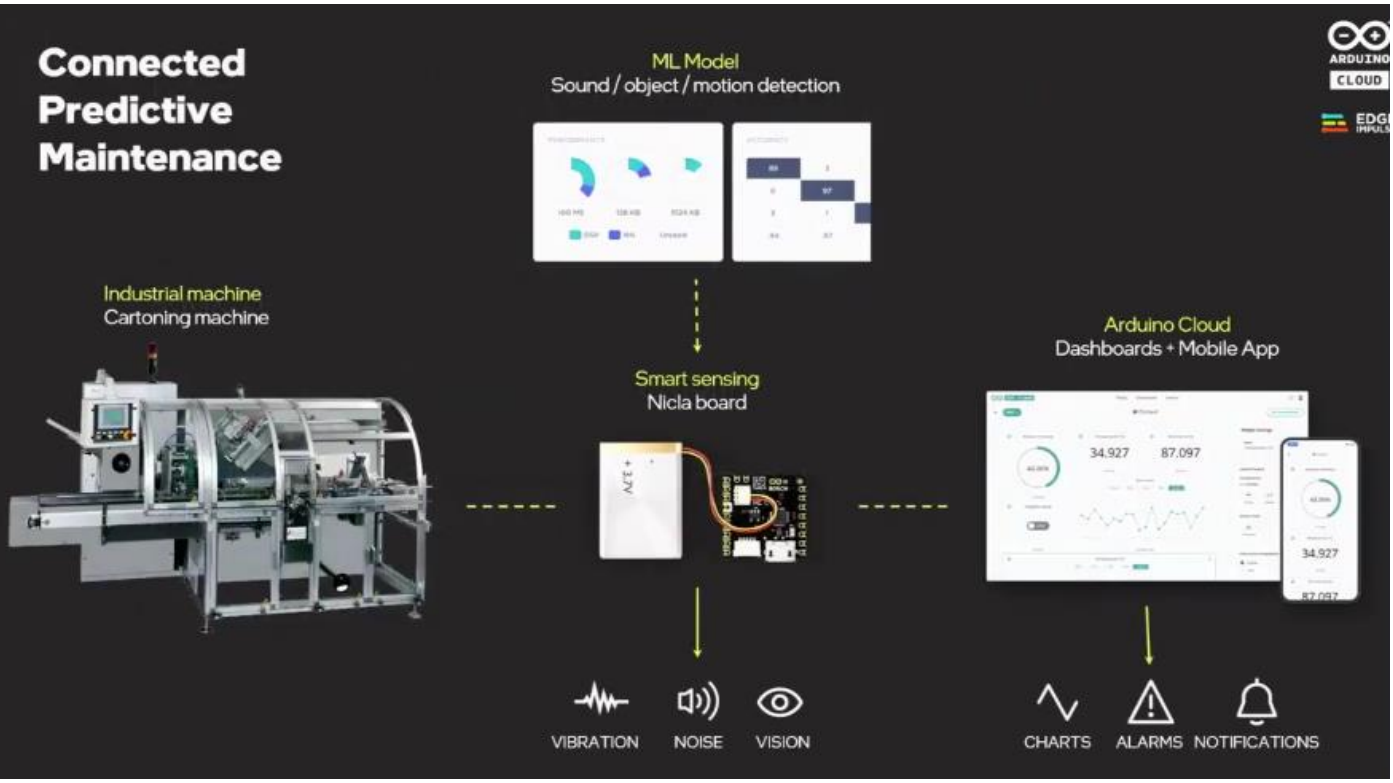
Data collection - Build custom datasets by collecting sensor, audio or camera data straight from devices, files or cloud integrations.

ML model design - Start creating your model.

ML model test - Validate your model and picture how it will perform with real-world data.

ML model deploy - Deploy your model to any device.

Connected Predictive Maintenance



Component defect detection

ML Model
Object detection

Production line
Juice production

Defect detection
Nicla Vision

Arduino Cloud
Dashboards + Mobile App

VISION

CHARTS ALARMS NOTIFICATIONS

Component defect detection

ML Model
Object detection

Logistics
Automated inventory management

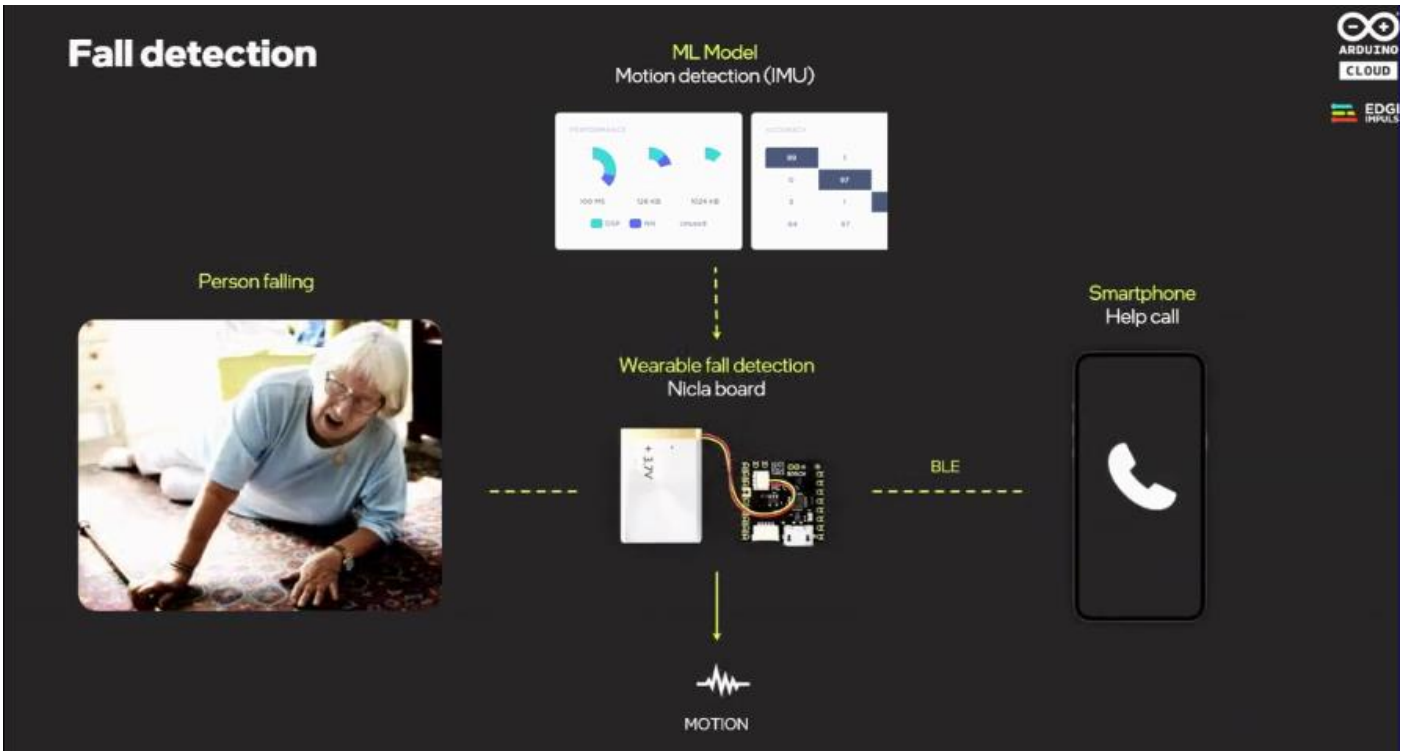
Amount / defect detection
Portenta H7 + Vision Shield

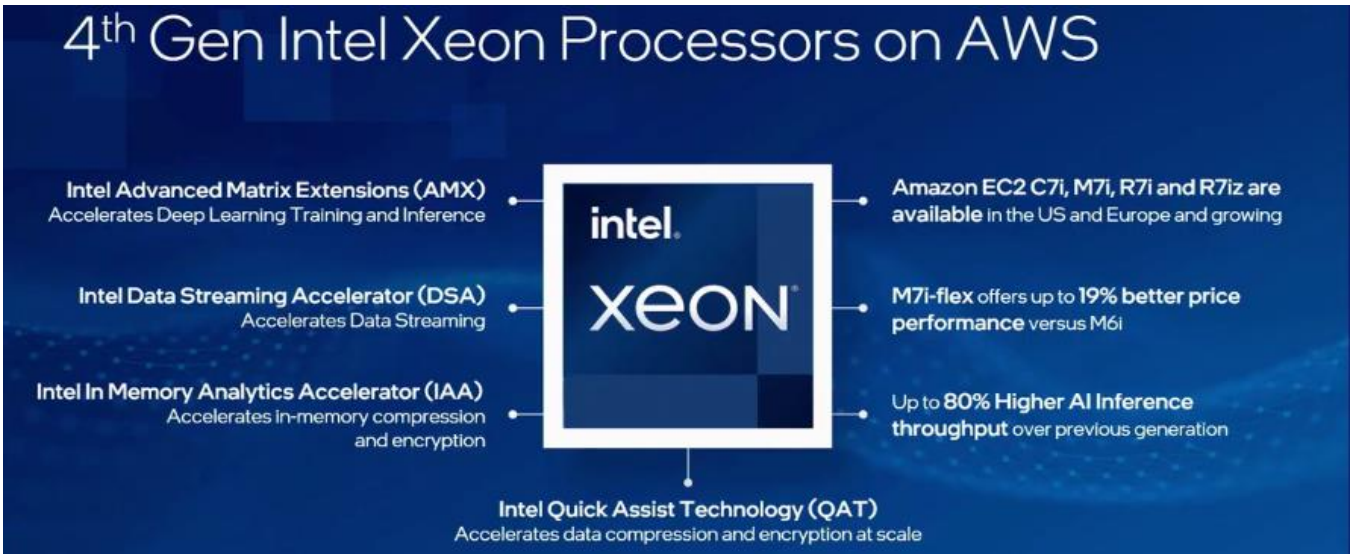
WiFi

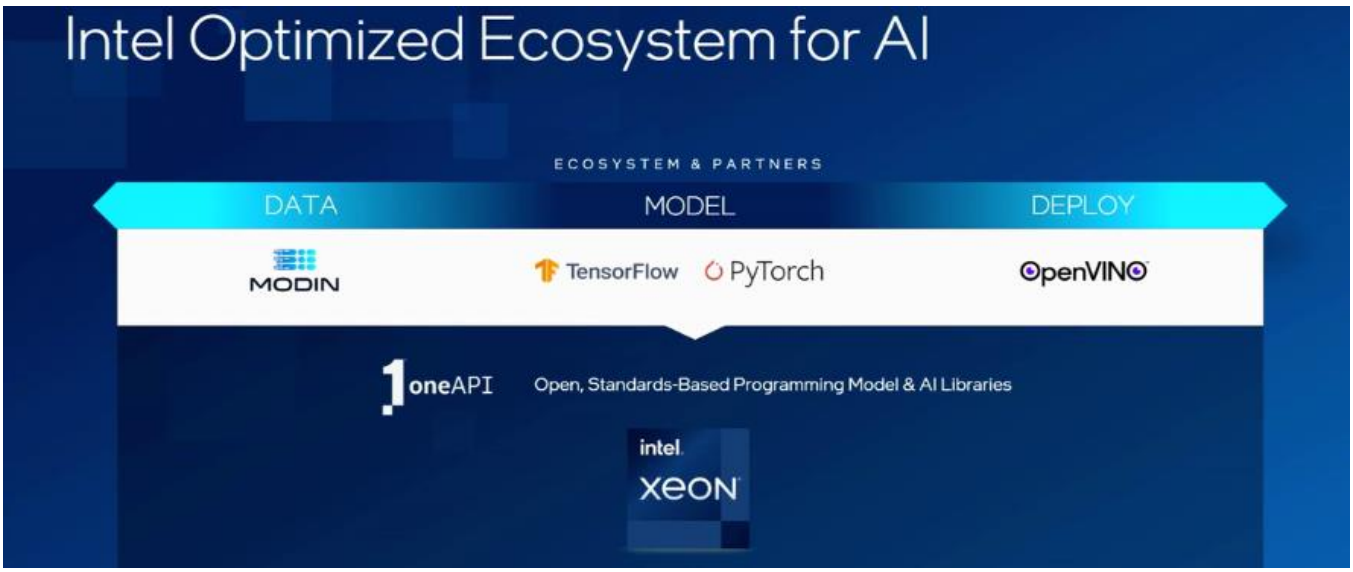
Arduino Cloud
Dashboards + Mobile App

VISION

CHARTS ALARMS NOTIFICATIONS


Fall detection
ML Model
Motion detection (IMU)
ARDUINO
CLOUD
EDGE IMPULSE
Person falling
Wearable fall detection
Nicla board
BLE
Smartphone
Help call
MOTION


4th Gen Intel Xeon Processors on AWS
Intel Advanced Matrix Extensions (AMX)
Accelerates Deep Learning Training and Inference
Intel Data Streaming Accelerator (DSA)
Accelerates Data Streaming
Intel In Memory Analytics Accelerator (IAA)
Accelerates in-memory compression and encryption
intel.
XEON
Amazon EC2 C7i, M7i, R7i and R7iz are available in the US and Europe and growing
M7i-flex offers up to 19% better price performance versus M6i
Up to 80% Higher AI Inference throughput over previous generation
Intel Quick Assist Technology (QAT)
Accelerates data compression and encryption at scale


Intel Optimized Ecosystem for AI
ECOSYSTEM & PARTNERS
DATA
MODEL
DEPLOY
MODIN
TensorFlow
PyTorch
OpenVINO
oneAPI
Open, Standards-Based Programming Model & AI Libraries
intel.
XEON

One-line Code Change = Huge Performance Gains

Engineer Data	Create Machine Learning & Deep Learning Models	Deploy
~90x	~38x	~3x
pandas MODIN	learn	TensorFlow
import modin.pandas as pd	from sklearnex import patch_sklearn patch_sklearn()	TF_ENABLE_ONEDNN_OPTS=1

Xeon AI and "Large" Language Models (LLMs)

Expert LM fine-tuned on a single task can outperform a multi-task model trained on 300 tasks

Race for one model to rule them all

"...we're at the end of the era where it's going to be these, like, giant, giant models," - Sam Altman, OpenAI CEO

"More companies would be better served focusing on smaller, specific models that are cheaper to train and run." - Clement Delangue, HuggingFace CEO

"small optimized models in healthcare are as accurate as large ones while being much more efficient." - David Talby, Spark-NLP CEO

Models to fit every business need

Intel Xeon AI for LLMs

- Enable the most cost effective and ubiquitous approach
- Fine-tune and optimize inference models on Intel Xeon processors
- Leverage hundreds of Intel and 3rd party pretrained models

Summary

- Use Intel solutions to run edge AI efficiently and at scale
- Leverage Intel 4th Generation Xeon based instances with AMX capabilities for AI
- Enable Intel SW optimizations for data processing, training and inference
- Visit us at our booth: **Intel Booth #750**

Connect With Our Partners

Visit our partners' booths to see how they use Intel technology.

CDW	Booth #305
HPE	Booth #630
IBM	Booth #930
SingleStore	Booth #1586
World Wide Technology	Booth #131
Nutanix	Booth #132
SUSE	Booth #501

Notices & Disclaimers

Intel is committed to the continued development of more sustainable products, processes, and supply chain as we strive to prioritize greenhouse gas reduction and improve our global environmental impact. Where applicable, environmental attributes of a product family or specific SKU will be stated with specificity. Refer to the 2022 Corporate Responsibility Report (p. 64) for further information.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.

Useful Links and Resources

AMX Enabled Amazon EC2 m7i Instances: https://aws.amazon.com/about-aws/whats-new/2023/08/amazon-ec2-m7i-flex-m7i-instances/

Public Proof Points for AI with 4th Gen Xeon: https://www.intel.com/content/www/us/en/newsroom/news/4th-gen-intel-xeon-outperforms-competition-real-world-workloads.html

OpenVino open-source Toolkit for Edge AI inference: https://www.intel.com/content/www/us/en/newsroom/news/4th-gen-intel-xeon-outperforms-competition-real-world-workloads.html

Leverage AMX to accelerate TensorFlow and PyTorch workloads: https://www.intel.com/content/www/us/en/developer/articles/technical/accelerate-tensorflow-ml-performance-amx.html#gs.00me4r; https://www.intel.com/content/www/us/en/developer/articles/technical/accelerate-pytorch-training-inference-on-amx.html

Optimizing Stable diffusion with Intel and Hugging Face: https://huggingface.co/blog/train-optimize-sd-intel

Performance Data for Intel AI Data Center Products: https://www.intel.com/content/www/us/en/developer/topic-technology/artificial-intelligence/performance-4th-gen-56c-xeon-max.html

Sustainability Technical Paper (Intel-Forrester): https://www.intel.com/content/www/us/en/newsroom/news/4th-gen-intel-xeon-outperforms-competition-real-world-workloads.html#gs.00mfhg