AIP
↳Bootcamp

Responsible AI in Practice

PALANTIR / AIP BOOTCAMP

FOUNDATIONAL SOFTWARE OF TOMORROW.
DELIVERED TODAY.™

▽
PALANTIR / PALO ALTO, CA
EST. 2003 / HQ DENVER, CO

→ PALANTIR.COM

AGENDA & OBJECTIVES

1/ Introduction to PCL at Palantir

2/ AI Ethics & Thought Leadership

3/ Privacy & Security by Design

4/ Applying Responsible AI to LLMs

The content herein is provided for informational purposes only and shall not create a warranty of any kind. Actual results and experiences may vary. © Palantir Technologies Inc.

Privacy & Civil Liberties [Engineering]

[01]

PCL Team Overview

HISTORY

Provide powerful insights

Protect privacy and civil liberties

MISSION

“The goal of the Privacy and Civil Liberties (PCL) team is to design, build, and deploy privacy-protective technologies and to foster a culture of responsibility around their development and use.”

RESPONSIBILITIES

Product Development

Implementation in the Field

External Engagement

Internal Culture

Limited trial access available. Talk to your Palantir Technologies representative to learn more. The content herein is provided for informational purposes only and shall not create a warranty of any kind. Actual results and experiences may vary. © Palantir Technologies Inc.

Data Privacy Regulation & PCL Engineering Building Blocks

ACCESS CONTROLS
Limit access to only what people have a need and right to know.

DATA REVELATION TECHNIQUES
Show them only the data that they need for their purpose.

RETENTION & PURGING
Set retention periods and have a deletion strategy.

RAILS & AWARENESS
Give them what they need to use solutions as intended – guidance, justification prompts, etc.

AUDIT LOGGING
Accountability and oversight, guarantees efficacy of all other building blocks.

GOVERNANCE
Ensure accountability and keeping humans in the loop for AI-assisted workflows.

AI Ethics & Thought Leadership

[02]

Palantir abides by these AI ethics principles which guide us in asserting our values to all contexts of AI use across our platforms.

Key Principles

01 / Focus on the fully integrated system, not just its component tools

02 / Acknowledge technology's limits

03 / Don't solve problems that shouldn't be solved

04 / Adhere to methodological best practices for sound data science

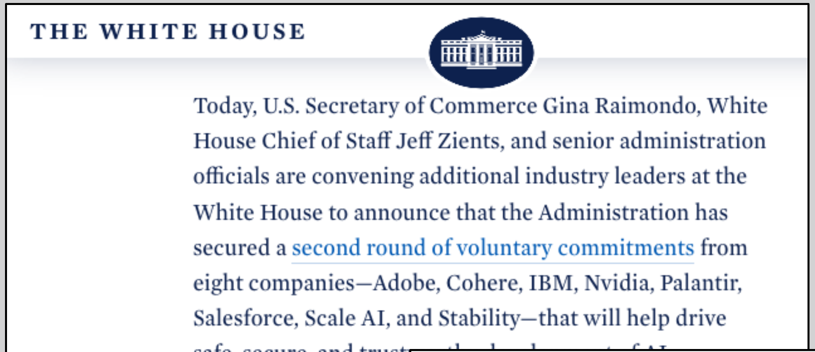
05 / Keep AI responsible, accountable, and oriented toward humans

06 / Promote multi-stakeholder engagement

07 / Ensure technical, governance, and cultural awareness

Limited trial access available. Talk to your Palantir Technologies representative to learn more. The content herein is provided for informational purposes only and shall not create a warranty of any kind. Actual results and experiences may vary. © Palantir Technologies Inc.

AI Policy Engagements and External Contributions



THE WHITE HOUSE

Today, U.S. Secretary of Commerce Gina Raimondo, White House Chief of Staff Jeff Zients, and senior administration officials are convening additional industry leaders at the White House to announce that the Administration has secured a second round of voluntary commitments from eight companies—Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, and Stability—that will help drive safe, secure, and trust

technology.

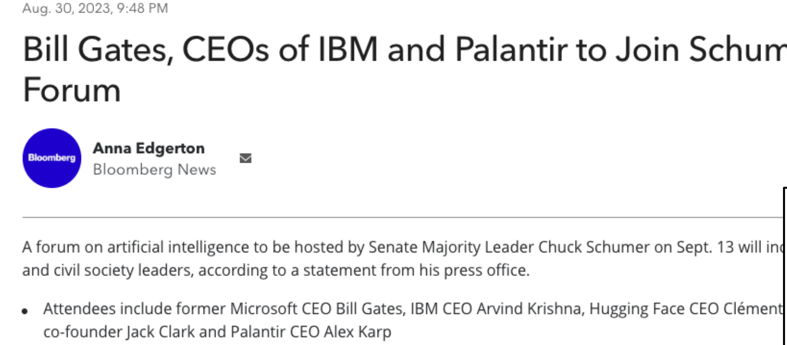
Appearance in the UK House of Lords

Committee on AI in Weapons Systems

PALANTIR TECHNOLOGIES — BLOG ©2023

FOUNDATIONAL SOFTWARE OF TOMORROW.
DELIVERED TODAY.™

EST. 2003 / HQ / PALO ALTO, CA DENVER, CO → PALANTIR.COM



Artificial Intelligence

Aug. 30, 2023, 9:48 PM

Bill Gates, CEOs of IBM and Palantir to Join Schumer's AI Forum

Anna Edgerton
Bloomberg News

A forum on artificial intelligence to be hosted by Senate Majority Leader Chuck Schumer on Sept. 13 will inc
and civil society leaders, according to a statement from his press office.

- Attendees include former Microsoft CEO Bill Gates, IBM CEO Arvind Krishna, Hugging Face CEO Clément
co-founder Jack Clark and Palantir CEO Alex Karp

Palantir Australia Response to Supporting Responsible AI Discussion Paper

Department of Industry, Science and Resources

July, 2023

palantir.com Copyright © 2023

SUBCOMMITTEE ON CYBERSECURITY

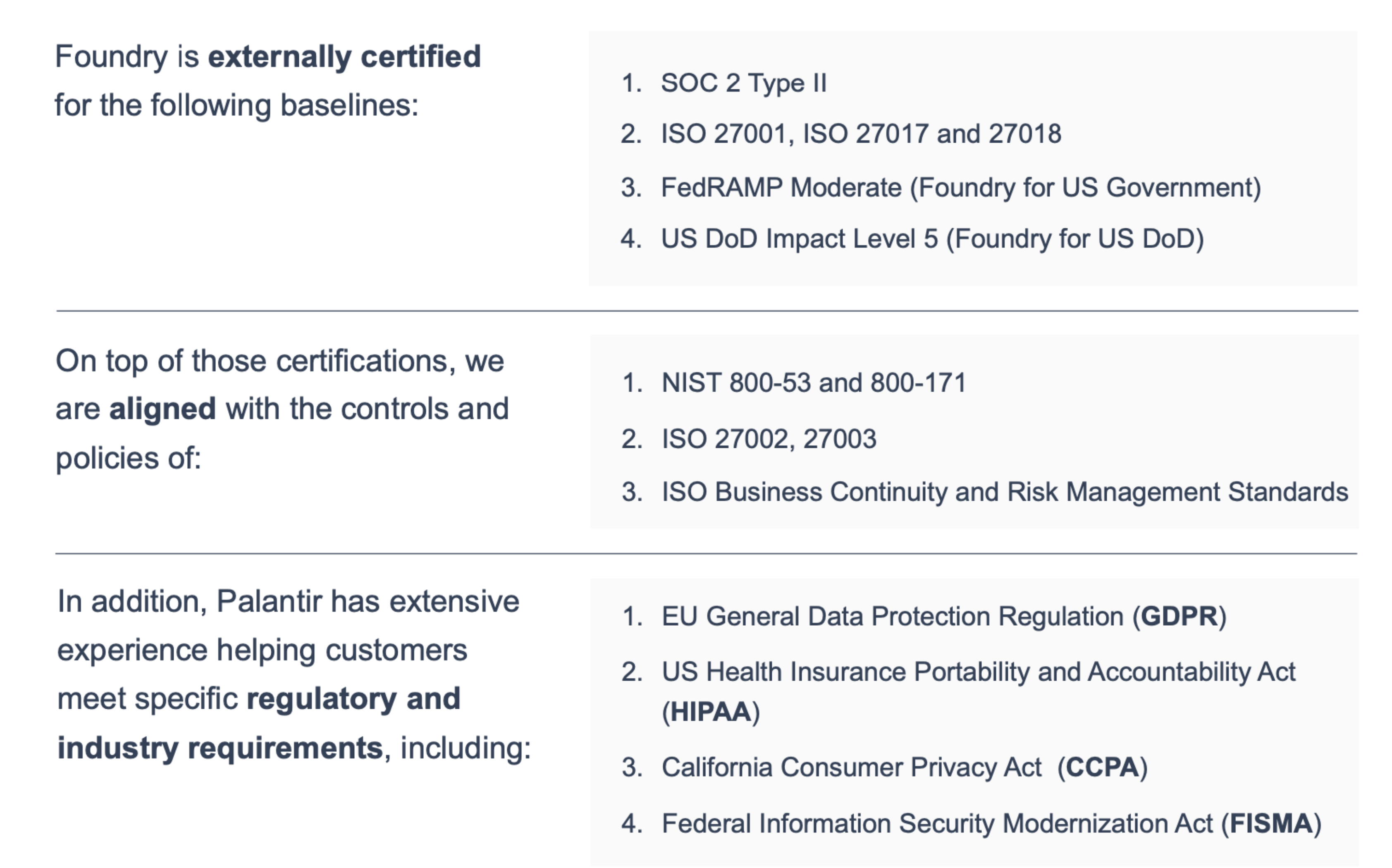
TO RECEIVE TESTIMONY ON THE STATE OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING APPLICATIONS TO IMPROVE DEPARTMENT OF DEFENSE OPERATIONS

Limited trial access available. Talk to your Palantir Technologies representative to learn more. The content herein is provided for informational purposes only and shall not create a warranty of any kind. Actual results and experiences may vary. © Palantir Technologies Inc.

Privacy & Security by Design

[03]

Palantir maintains rigorous, externally verified infrastructure and operations standards.



Foundry is **externally certified** for the following baselines:

1. SOC 2 Type II
2. ISO 27001, ISO 27017 and 27018
3. FedRAMP Moderate (Foundry for US Government)
4. US DoD Impact Level 5 (Foundry for US DoD)

On top of those certifications, we are **aligned** with the controls and policies of:

1. NIST 800-53 and 800-171
2. ISO 27002, 27003
3. ISO Business Continuity and Risk Management Standards

In addition, Palantir has extensive experience helping customers meet specific **regulatory and industry requirements**, including:

1. EU General Data Protection Regulation (**GDPR**)
2. US Health Insurance Portability and Accountability Act (**HIPAA**)
3. California Consumer Privacy Act (**CCPA**)
4. Federal Information Security Modernization Act (**FISMA**)

https://palantir.safebase.us/

Limited trial access available. Talk to your Palantir Technologies representative to learn more. The content herein is provided for informational purposes only and shall not create a warranty of any kind. Actual results and experiences may vary. © Palantir Technologies Inc.

All AIP infrastructure is *private* and *secure* by design.

- Palantir has out-of-the-box open-source models ready to be leveraged by the customer.
- Customer administrators toggle which models they want configured and available to users.
- Customer models can be securely integrated to AIP.
- Palantir-provided models are ephemeral and geo-restriction is configurable, where available.
- Customer data is not used to retrain Palantir-provided models.
- Communication with models is through encrypted TLS 1.2+ HTTPS, and AIP enforced strict Ingress/Egress rules.

Limited trial access available. Talk to your Palantir Technologies representative to learn more. The content herein is provided for informational purposes only and shall not create a warranty of any kind. Actual results and experiences may vary. © Palantir Technologies Inc.

Applying Responsible AI to LLMs

[04]

Designing Human-centered workflows

5. Keep AI responsible, accountable, and oriented toward humans

Palantir

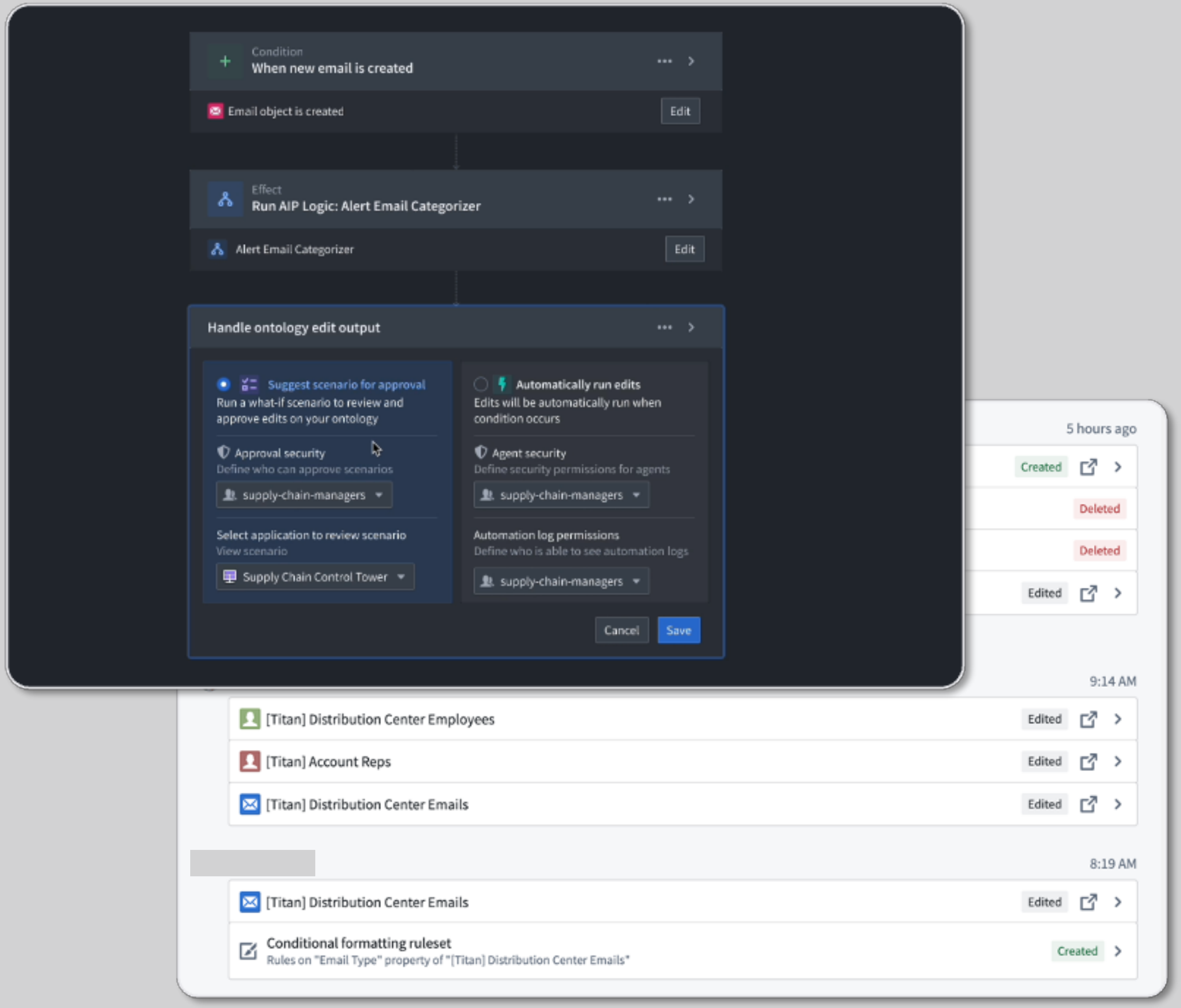
- **Delegate to humans** what they're good at, delegate to machines what they're good at

- Have **strong security & audit capabilities** for added transparency & interpretability

- **Limit reasoning** and implicit tool chaining

- Link back to the sources and present immutable tool outputs to enable **verification**

- Reinforcement learning through **human feedback** and related techniques



Condition
When new email is created
Email object is created — Edit

Effect
Run AIP Logic: Alert Email Categorizer
Alert Email Categorizer — Edit

Handle ontology edit output

● Suggest scenario for approval
Run a what-if scenario to review and approve edits on your ontology
Approval security
Define who can approve scenarios
supply-chain-managers
Select application to review scenario
View scenario
Supply Chain Control Tower

○ Automatically run edits
Edits will be automatically run when condition occurs
Agent security
Define security permissions for agents
supply-chain-managers
Automation log permissions
Define who is able to see automation logs
supply-chain-managers

Cancel — Save

5 hours ago
Created
Deleted
Deleted
Edited

9:14 AM
[Titan] Distribution Center Employees — Edited
[Titan] Account Reps — Edited
[Titan] Distribution Center Emails — Edited

8:19 AM
[Titan] Distribution Center Emails — Edited
Conditional formatting ruleset — Created
Rules on "Email Type" property of "[Titan] Distribution Center Emails"

Robust data governance underpins responsible AI systems

1. Focus on the fully integrated system

Palantir

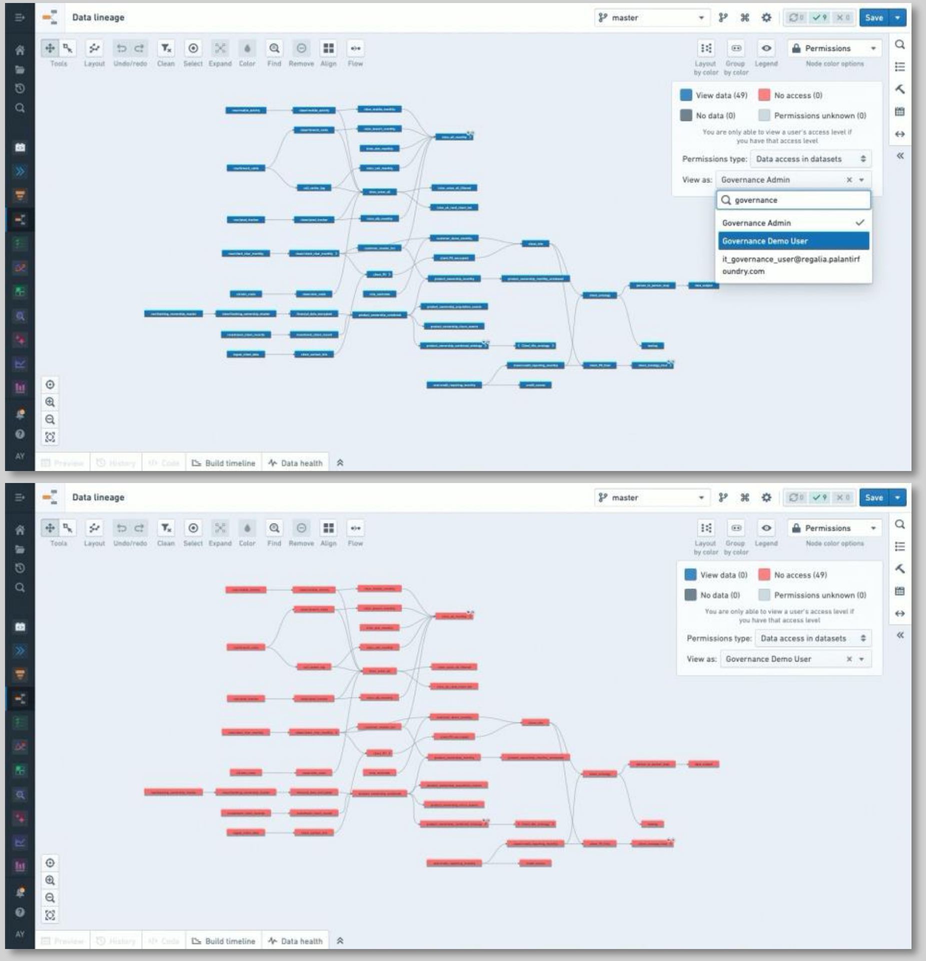
Track Permissions and Data Access for Humans and LLMs

OVERVIEW

As you start ingesting data, and permissions get set on the incoming data as well as any applications downstream, platform administrators and compliance users will want to review permissions.

IMPLEMENTATION

Regularly validating permissions on already ingested data or reviewing permissions as new data helps prevent possible misconfigurations and inadvertent exposure of sensitive data downstream.



🟦 View data (49) 🟥 No access (0)
⬛ No data (0) ⬜ Permissions unknown (0)

You are only able to view a user's access level if you have that access level

Permissions type: Data access in datasets

View as: Governance Admin

governance

Governance Admin ✓
Governance Demo User
it_governance_user@regalia.palantirfoundry.com

🟦 View data (0) 🟥 No access (49)
⬛ No data (0) ⬜ Permissions unknown (0)

You are only able to view a user's access level if you have that access level

Permissions type: Data access in datasets

View as: Governance Demo User

Access for User with and without Permissions

Limited trial access available. Talk to your Palantir Technologies representative to learn more. The content herein is provided for informational purposes only and shall not create a warranty of any kind. Actual results and experiences may vary. © Palantir Technologies Inc.

AI Testing & Evaluation Best Practices

Model Management & Evals

Reduction to supervised learning:

→ seed data generation and labeling processes
→ training and sustaining supervised models

Prompts as model hyper – parameters:

→ empirical evaluation of prompting strategies
→ analyzing variations in prompt structure, wording, length, tone
→ dynamically generated prompts

Workflow Design

Decomposte tasks/decisions:

→ structural evaluation
→ content evaluation
→ semantic evaluation

→ Metrics → workflow KPIs

→ State space guardrails

→ Human evaluation and red teaming to probe system limitations

→ Adversarial evaluation

→ LLMs to evaluate LLMs

Operational Testing

→ Human Review

→ Safety/Unit Tests

→ Incremental Release and monitoring

→ A/B Testing

→ Shadow Testing

→ Scenarios- and simulation-based testing

Palantir

Q&A

PALANTIR / AIP BOOTCAMP

FOUNDATIONAL SOFTWARE OF TOMORROW.
DELIVERED TODAY.™

▽

PALANTIR / PALO ALTO, CA
EST. 2003 / HQ DENVER, CO

→ PALANTIR.COM