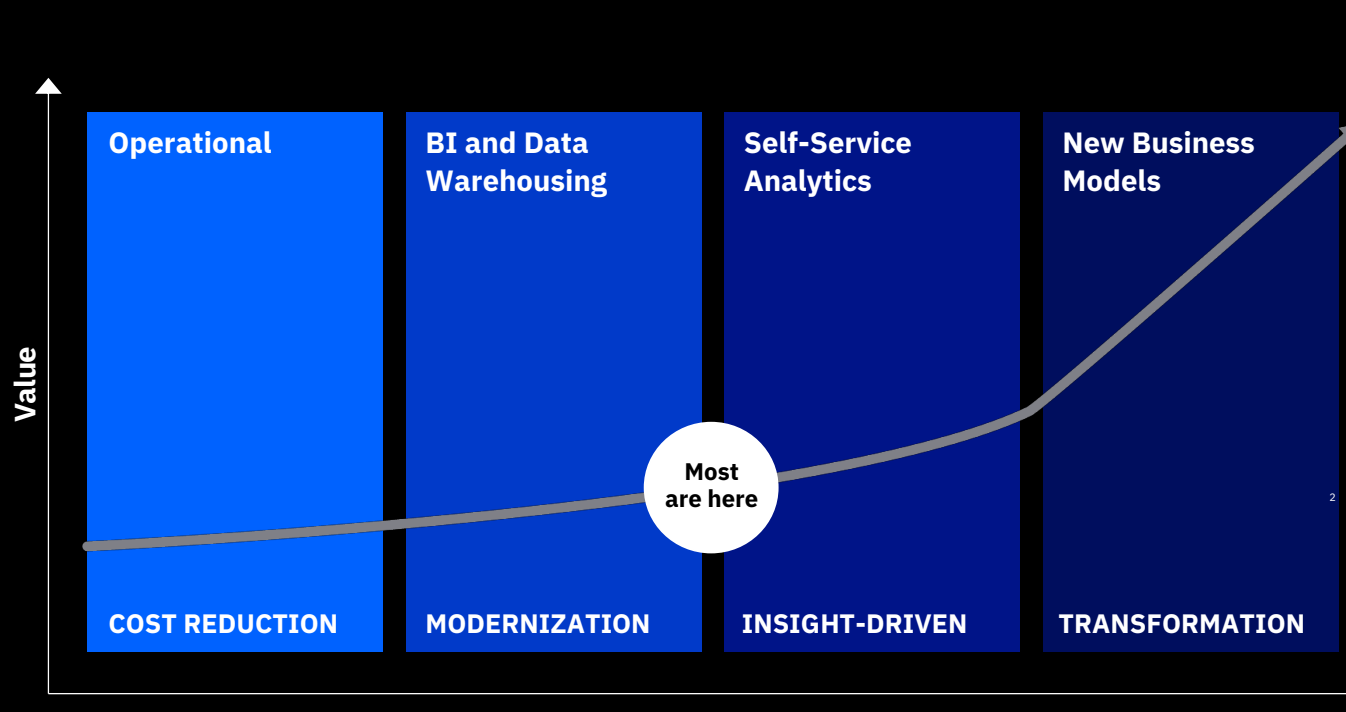
EDHEC BUSINESS SCHOOL

AI Workshop

Trust and Transparency in AI

Emmanuel Génard – *genard@fr.ibm.com*
Data Scientist & Cloud Developer Advocate Europe,
IBM Business Solution Center Nice, France

I want AI !



85%
view AI as a strategic opportunity

MITSloan

But .....

In the same MIT & BCG survey of more than 3,000 executives, managers, and analysts across industries...

39%

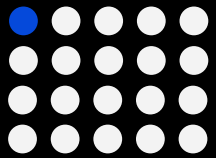
Of all companies have an AI strategy in place

(50% when only counting companies with at least 100,00 employees)

1/5

Has incorporated AI in *some* offerings or products

1/20



Has *extensively* incorporated AI in offerings or process

Business stakeholders do not trust AI.

60%

of companies see **regulatory constraints** as a barrier to implementing AI.

– *IBM IBV AI 2018*

63%

cite availability of **technical skills** as a challenge to implementation.

- IBM IBV AI 2018

Without expensive Data Science resources handholding multiple AI models in a production application:

1. No way to **validate** if AI models are **compliant with regulations** and will achieve expected business outcomes before deploying
2. Difficult to **track and measure** indicators of business success in production
3. Resource intensive and unreliable processes for **ongoing business monitoring and compliance**
4. Impossible for business users to **feedback** subtle domain knowledge into model lifecycle

Regulatory Challenges to AI Adoption

Consumers have the right to access "**meaningful information about the logic involved**, as well as the significance and the envisaged consequences of such processing for the data subject."

– Articles 12-15, General Data Protection Regulation (EU)

If a bank denies credit, it is required to provide the applicant with the **specific reasons why the applicant was turned down**.

– Fair Credit Reporting Act (United States)

In its 2015 decision, the Supreme Court held that plaintiffs need only show that a policy **had a discriminatory impact on a protected class**, and not that the discrimination was intentional.

– Texas Department of Housing and Community Affairs v. Inclusive Communities Project (United States)

Ethics guidelines for trustworthy AI

The European Union has published new guidelines on developing ethical AI

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

- **Human agency and oversight** — AI should not trample on human autonomy. People should not be manipulated or coerced by AI systems, and humans should be able to intervene or oversee every decision that the software makes.
- **Technical robustness and safety** — AI should be secure and accurate. It shouldn't be easily compromised by external attacks (such as adversarial examples), and it should be reasonably reliable.
- **Privacy and data governance** — Personal data collected by AI systems should be secure and private. It shouldn't be accessible to just anyone, and it shouldn't be easily stolen.
- **Transparency** — Data and algorithms used to create an AI system should be accessible, and the decisions made by the software should be "understood and traced by human beings."
- **Diversity, non-discrimination, and fairness** — Services provided by AI should be available to all, regardless of age, gender, race, or other characteristics.
- **Environmental and societal well-being** — AI systems should be sustainable
- **Accountability** — AI systems should be auditable and covered by existing protections for corporate whistleblowers.

See resources

Watson OpenScale will help validate and monitor AI models, deployed anywhere, to comply with regulations and control lifecycle from a business perspective

Prove regulatory compliance and safeguards
Detect and mitigate model biases
Audit and Explain model decisions

Required in regulated industries and use cases – FSS, HR etc. in short term; others longer term*

Ensure that models are resilient to changing situations
Detect drift in data and anomaly in model behavior
Specific inputs and triggers to model lifecycle

Required to meet transformational goals

Align model performance with business outcomes
Correlate model metrics and business KPIs to measure business impact
Actionable metrics and alerts

Foundational to all AI implementations

** E.g. Fair lending practices in finance vs. GDPR across all industries*

Watson OpenScale will help validate and monitor AI models, deployed anywhere, to comply with regulations and control lifecycle from a business perspective

Prove regulatory compliance and safeguards
Detect and mitigate model biases
Audit and Explain model decisions

Required in regulated industries and use cases – FSS, HR etc. in short term; others longer term*

Ensure that models are resilient to changing situations
Detect drift in data and anomaly in model behavior
Specific inputs and triggers to model lifecycle

Required to meet transformational goals

Align model performance with business outcomes
Correlate model metrics and business KPIs to measure business impact
Actionable metrics and alerts

Foundational to all AI implementations

** E.g. Fair lending practices in finance vs. GDPR across all industries*

Watson OpenScale will help validate and monitor AI models, deployed anywhere, to comply with regulations and control lifecycle from a business perspective

Prove regulatory compliance and safeguards
Detect and mitigate model biases
Audit and Explain model decisions

Required in regulated industries and use cases – FSS, HR etc. in short term; others longer term*

Ensure that models are resilient to changing situations
Detect drift in data and anomaly in model behavior
Specific inputs and triggers to model lifecycle

Required to meet transformational goals

Align model performance with business outcomes
Correlate model metrics and business KPIs to measure business impact
Actionable metrics and alerts

Foundational to all AI implementations

** E.g. Fair lending practices in finance vs. GDPR across all industries*

Watson OpenScale will help validate and monitor AI models, deployed anywhere, to comply with regulations and control lifecycle from a business perspective

Prove regulatory compliance and safeguards
Detect and mitigate model biases
Audit and Explain model decisions

Required in regulated industries and use cases – FSS, HR etc. in short term; others longer term*

Ensure that models are resilient to changing situations
Detect drift in data and anomaly in model behavior
Specific inputs and triggers to model lifecycle

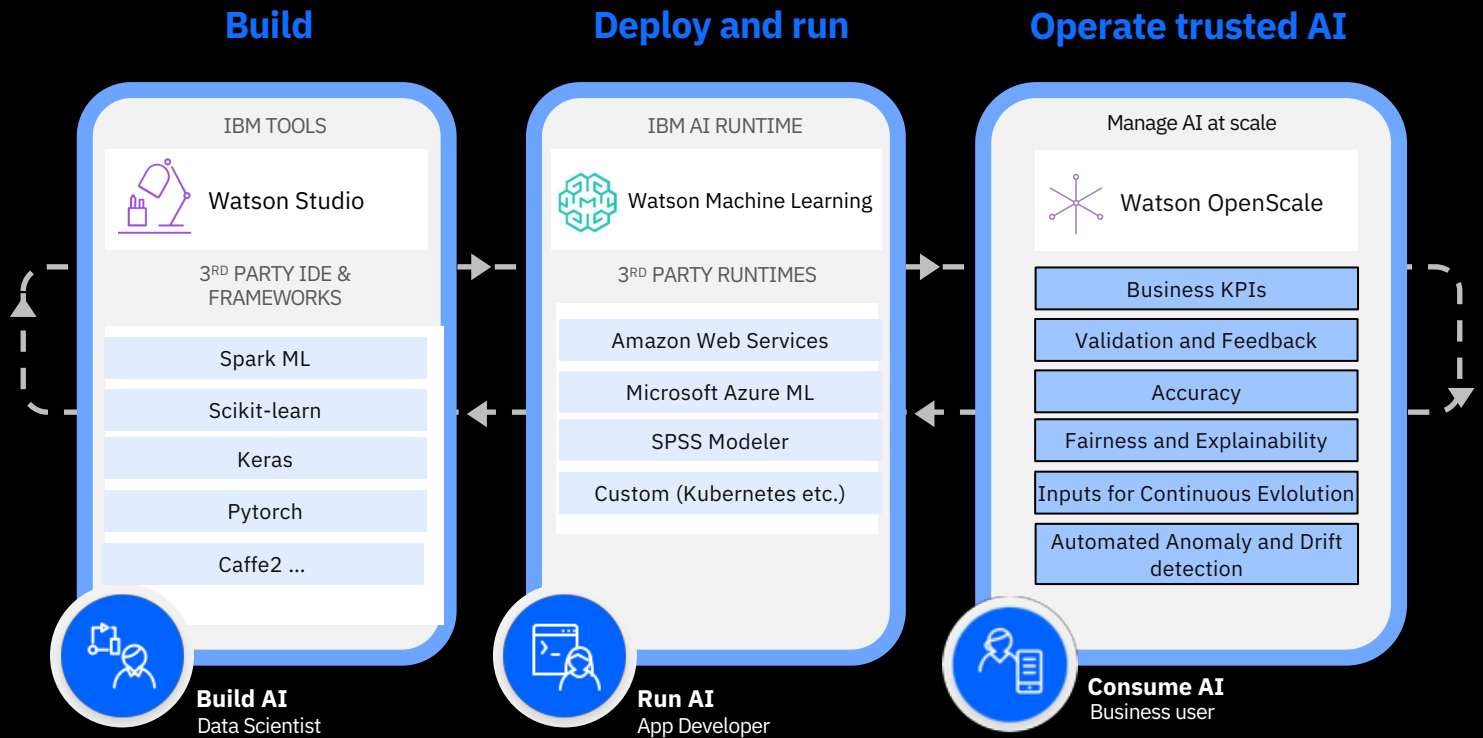
Required to meet transformational goals

Align model performance with business outcomes
Correlate model metrics and business KPIs to measure business impact
Actionable metrics and alerts

Foundational to all AI implementations

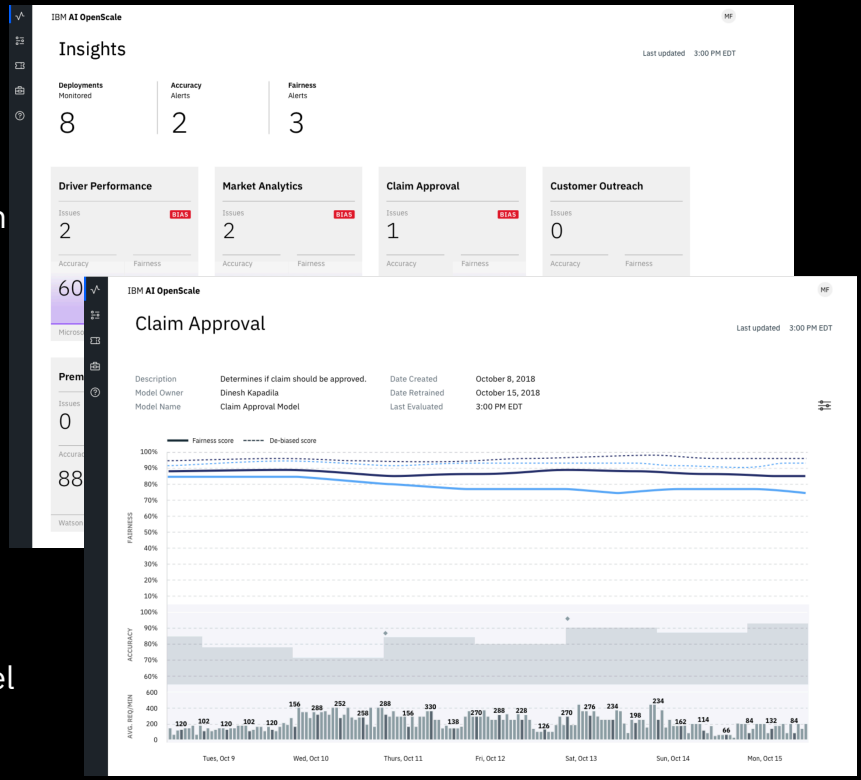
** E.g. Fair lending practices in finance vs. GDPR across all industries*

Watson OpenScale along with Watson Studio and WML enables enterprises to operationalize AI across the enterprise



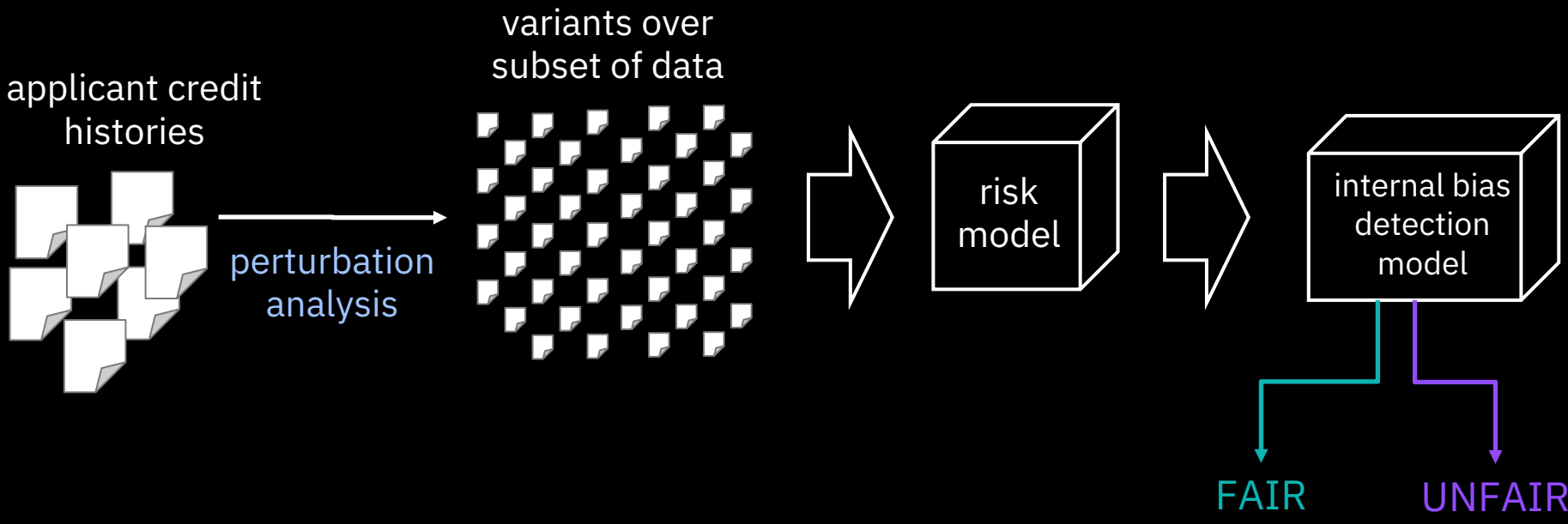
IBM Watson OpenScale
Automate and Operate AI at Scale

- Trust and Transparency
  - Intelligently delivers bias mitigation help
  - Provides traceability & auditability of AI predictions made in production applications
  - Tracks AI accuracy in applications
  - Explains an outcome in business terms
- Automation
  - Automatically detects and mitigates bias in model output, without affecting currently deployed model or outcomes
- Open By Design
  - Monitor and optimize models deployed on third party model serve engines
  - Deploy behind enterprise firewall or on IaaS provider



How does Watson OpenScale mitigate fairness issues?

Calculated hourly over a sliding window



Bias Mitigation – Guardrails against AI Bias

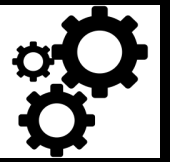
Watson OpenScale

Online Post Processing
(Intercept & Fix Features)

AI Fairness 360

In-Processing
(Intercept & Fix Algorithm)

Runtime

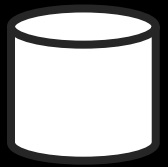


AI Fairness 360

Pre-Processing
(Intercept & Fix Training Data)

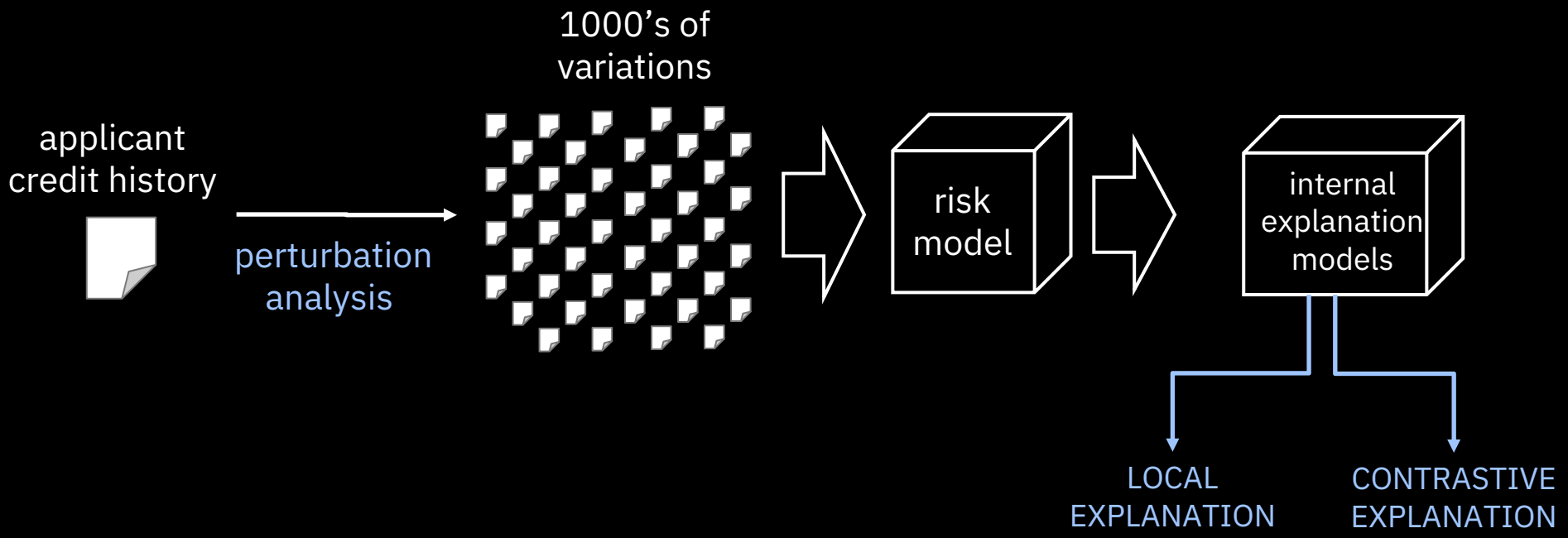
Algorithm Design

Data Selection



How does Watson OpenScale explain a prediction?

Calculated upon request



Credit Risk Modeling

Problem Statement:

Traditional lenders are under pressure to expand their digital portfolio of financial services to a larger and more diverse audience, which requires a new approach to credit risk modeling.

To provide credit access to a wider and riskier population, applicant credit histories must expand beyond traditional credit—like mortgages and car loans—to alternate sources, such as utility and cell plan payment histories, plus education and job titles. These additional features increase the likelihood of unexpected correlations that introduce bias based on an applicant's age, gender and other personal traits. The data science techniques most suited to these diverse datasets can generate highly accurate risk models but at a cost—such models are black boxes whose inner workings are not easily understood.

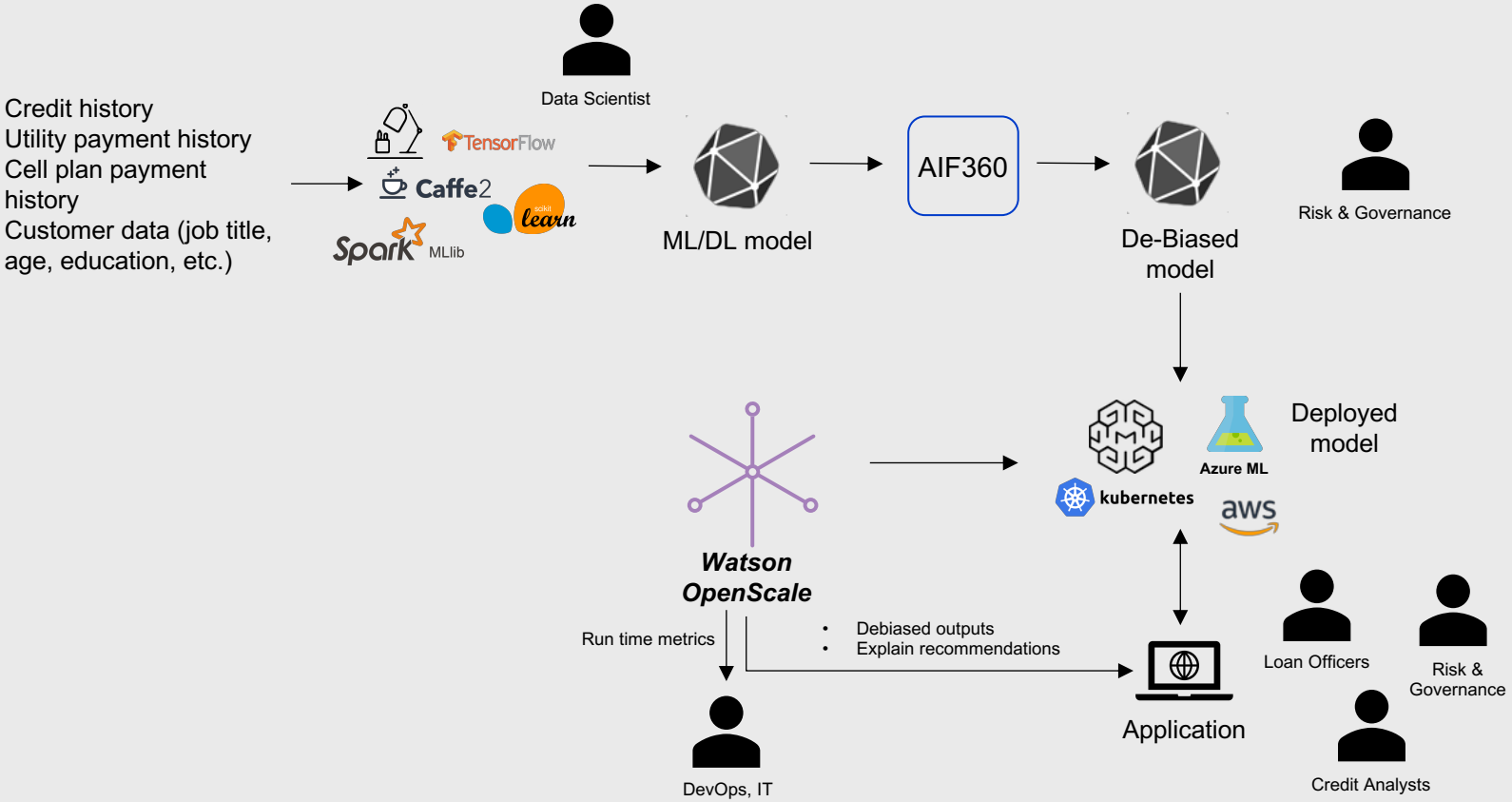
Regulations like the General Data Protection Regulation in the EU, the Federal Fair Credit Reporting Act and the Equal Credit Opportunity Act in the United States require that banks and creditors must be able to provide specific reasons for every credit decision they make. The opaque predictions made by black-box models must be made transparent to ensure regulatory approval.

Watson OpenScale helps by:

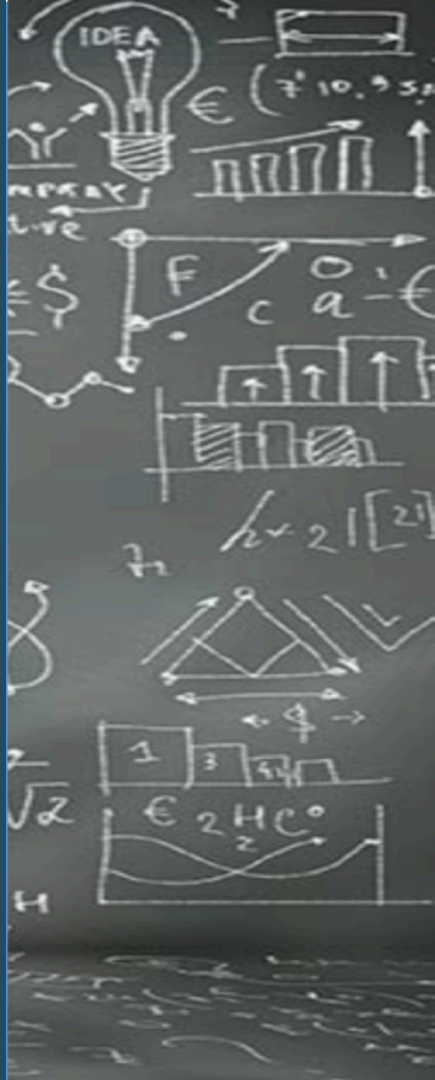
Banks and credit unions need to be able to check their credit risk models for bias, not just during training but also after these models are deployed. And in order to be compliant with regulations like the Equal Credit Opportunity Act, they need to be able to explain why their models make individual credit decisions.

Watson OpenScale's **bias detection and mitigation features** allow risk and governance officers to monitor bias within their models at runtime. And Watson OpenScale's **explainability support** provides loan officers and credit analysts with post-facto explanations for model decisions, even with black-box models like XGBoost, which provide high accuracy in credit risk modeling.

Architecture Diagram:
Data Flow for Credit Risk Modeling Use Case



Thank You



Useful Links & Resources

External

Getting Started:
Service Homepage
Feature Requests / Suggestions

Case Studies:
OmniEarth
Aerialtronics
BlueChasm
iTrend

Tutorials & Best Practices:
Training models with Watson Studio
Getting started with Watson + Core ML
Stacking Multiple Custom Models
Create a Calorie Counting App
Watson Visual Recognition & Twilio
Best Practices for Custom Models

Code Patterns:
Classify vehicle damage
Analyze industrial equipment for defects
Create an Android calorie-counter app

External continued

Books:
Redbook: Building Cognitive Application using IBM Watson Services vol3 – Watson Visual Recognition

Blogs:
IBM Watson on Medium

Internal

Slack Channel: #ibmvisual-recognition
Service Roadmap
IBMer key limit increase request form
ZACS portal
Digital Sales Play
Content Request & Feedback Form