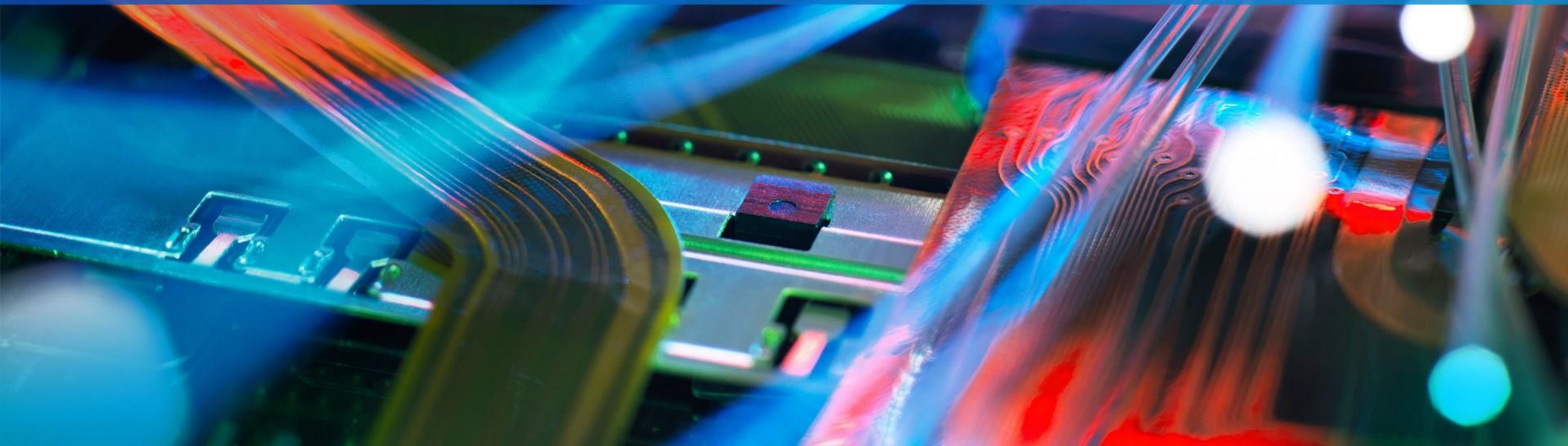
Swiss Re

AI Risk to Companies

Lukas Csoka, Head Big Data Foundations, Swiss Re
for ODSC Europe 2021



About me

- Master in Software Engineering from Slovak University of Technology in Bratislava
  - Academic senate
  - Consultant Operating Systems
  - Published research and won IT competitions
- MBA in Global Management from City University of Seattle
- Head Big Data Foundations in Swiss Re
  - Create and manage data science projects, leading a team of data scientists
  - Provide consulting to business leaders to develop appropriate reports, metrics and research.
  - Maintain and evolve data analytics capabilities.

Outline

AI Risks and Governance

Bias in Data

Adversarial Attack

AI Risks and Governance

Success in creating effective AI, could be the biggest event in the history of our civilization. Or the worst. We just don't know. So we cannot know if we will be infinitely helped by AI, or ignored by it and side-lined, or conceivably destroyed by it, …

Stephen Hawking

AI Risks to Society

- The potential of automation technology to give rise to job losses
- The need to redeploy or retrain employees to keep them in jobs
- Fair distribution of wealth created by machines
- The effect of machine interaction on human behavior and attention
- The need to address algorithmic bias originating from human bias in the data
- The security of AI systems (e.g. autonomous weapons) that can potentially cause damage
- The need to mitigate against unintended consequences, as smart machines are thought to learn and develop independently

AI poses unfamiliar risks and creates new responsibilities

- According to the July 2019 "AI and Empathy" report from software maker Pega, 35% of the 6,000 individuals surveyed said they were concerned that machines would take their jobs and 27% said they were concerned about "the rise of robots and enslavement of humanity.“
- If the end user doesn't trust the machine, which isn't unusual, then that AI is a failure.
- Pilots are losing their basic flying skills
- Let's consider this example:
  - Banks using AI to provide consumer advice

New Rules for Artificial Intelligence – European Union

The higher the risk, the stricter the rule

- AI systems **do not create or reproduce bias**
- Trained and tested with **sufficiently representative** dataset
- Categories:
  - Unacceptable risk
  - High-risk
  - Limited risk
  - Minimal risk
- Providers of high-risk AI systems will also have to implement quality and risk management systems to ensure their compliance
- For certain AI systems, an independent notified body will also have to be involved in this process

- **Up to €30m or 6%** of the total worldwide annual turnover of the preceding financial year (whichever is higher) for infringements **on prohibited practices or non-compliance** related to requirements on data;
- **Up to €20m or 4%** of the total worldwide annual turnover of the preceding financial year for **non-compliance with any of the other requirements** or obligations of the Regulation;
- **Up to €10m or 2%** of the total worldwide annual turnover of the preceding financial year for the **supply of incorrect, incomplete or misleading information** to notified bodies and national competent authorities in reply to a request.

AI systems identified as high-risk include AI technology used in

- **Critical infrastructures** (e.g., transport), that could put the life and health of citizens at risk;
- **Educational or vocational training**, that may determine the access to education and professional course of someone's life (e.g., scoring of exams);
- **Safety components of products** (e.g., AI application in robot-assisted surgery);
- **Employment, workers management and access to self-employment** (e.g., CV-sorting software for recruitment procedures);
- **Essential private and public services** (e.g., credit scoring denying citizens opportunity to obtain a loan);
- **Law enforcement** that may interfere with people's fundamental rights (e.g., evaluation of the reliability of evidence);
- **Migration, asylum and border control management** (e.g., verification of authenticity of travel documents);
- **Administration of justice** and **democratic processes** (e.g., applying the law to a concrete set of facts).

- **Adequate risk assessment and mitigation systems;**
- **High quality of the datasets** feeding the system to minimize risks and discriminatory outcomes;
- **Logging of activity to ensure traceability of results**;
- **Detailed documentation** providing all information necessary on the system and its purpose for authorities to assess its compliance;
- **Clear and adequate information** to the user;
- **Appropriate human oversight** measures to minimize risk;
- High level of **robustness**, **security** and **accuracy**.

Data-Related

- Choice of appropriate data features
- Algorithmic bias and fairness
- Data Quality & Governance
- Feature Engineering

Model Devel.

- Model Assumptions, (Hyper-) Parameters and Limitations
- Performance Metrics
- Model Validation
- Calibration
- Uncertainty
- Robustness and Stability

Usage

- Fit for purpose
- Explainability
- Deployment: Recalibration
- Deployment: Human Component

Governance

- Reproducibility & Auditability

When developing Governance framework, consider following:

- Fit for purpose (data + model)
- Choice of appropriate data features
- Algorithm Bias and Fairness
- Data Quality (Outliers, missing data, garbage in garbage out) & Governance
- Feature Engineering
- Model Assumptions, (Hyper-)Parameters and Limitations
- Performance Metrics
- Model Validation
- Calibration
- Uncertainty
- Robustness and Stability
- Explainability
- Deployment: Process Perspective
- Deployment: Human Component
- Reproducibility & Auditability

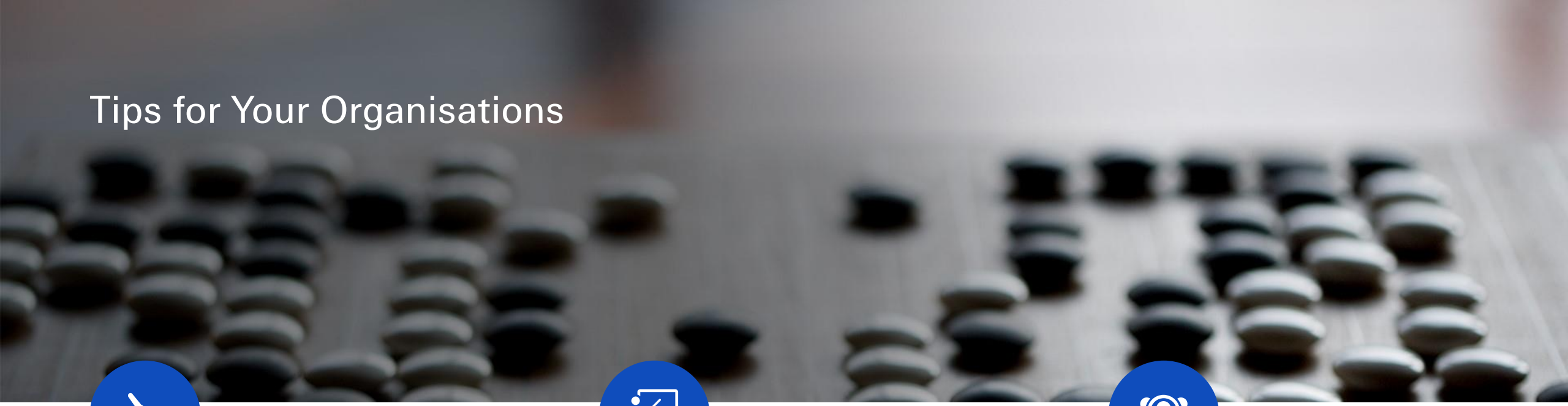
Fit for purpose (data + model)

- Has the problem statement been formulated in a simple and clear manner?
- Do we understand what business decisions would be made based on the results of the model?
- Is the data available, appropriate and adequate to answer the business question?
- Does the modelling goal appropriately reflect the business challenge?
- Has the appropriate trade-off between model interpretability and accuracy been reached?
- How will the delivery be used by the end-user?
- Is the modelled output metric appropriate for the given business challenge?

Choice of appropriate data features

- Have you ensured that all data used present no regulatory or reputational risk?
- Could any attributes be perceived as discriminatory, for example gender/ age / religion/ ethnicity or highly correlated to them?
- Is there a consent about which features could be perceived as discriminatory?
- Does the algorithm use proxies for attributes being perceived as discriminatory?

Tips for Your Organisations



Take a proactive approach to AI risks.

Consider what risk management activities your organization is undertaking for AI, and whether there are others you could put into place.

Develop the right capabilities.

Build and promote understanding of AI risks throughout the organization. Awareness campaigns and trainings available for everybody can build institutional knowledge. Create specialized teams capable of understanding and interpreting analytics use cases and approaches.

Establish governance and key roles.

Identify key people in analytics teams and related risk-management roles such as legal, clarify their roles within the risk-management framework, and define their mandate and responsibilities in relation to AI controls.

Bias in Data

Short Example

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims:

"I can't operate on this boy, he's my son!“

Short Example

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims:

"I can't operate on this boy, he's my son!“

- The majority of test subjects overlooked the possibility that the doctor is a she - including men, women, and self-described feminists.
- BU Research: A Riddle Reveals Depth of Gender Bias | BU Today | Boston University

Definition

- Machine learning models are not inherently objective. Engineers train models by feeding them a data set of training examples, and human involvement in the provision and curation of this data can make a model's predictions susceptible to bias.
- **Automation bias** is a tendency to favor results generated by automated systems over those generated by non-automated systems, irrespective of the error rates of each.
- **Selection bias** occurs if a data set's examples are chosen in a way that is not reflective of their real-world distribution.
- **Group attribution bias** is a tendency to generalize what is true of individuals to an entire group to which they belong.
- **Implicit bias** occurs when assumptions are made based on one's own mental models and personal experiences that do not necessarily apply more generally.
- Source: Fairness: Types of Bias | Machine Learning Crash Course (google.com)
- I would add:
  - Observer Bias
  - Measurement Bias
  - Exclusion Bias

Examples: It's always about people in the end!

US Healthcare

Correctional Offender Management Profiling for Alternative Sanctions

Amazon's Hiring Algorithm

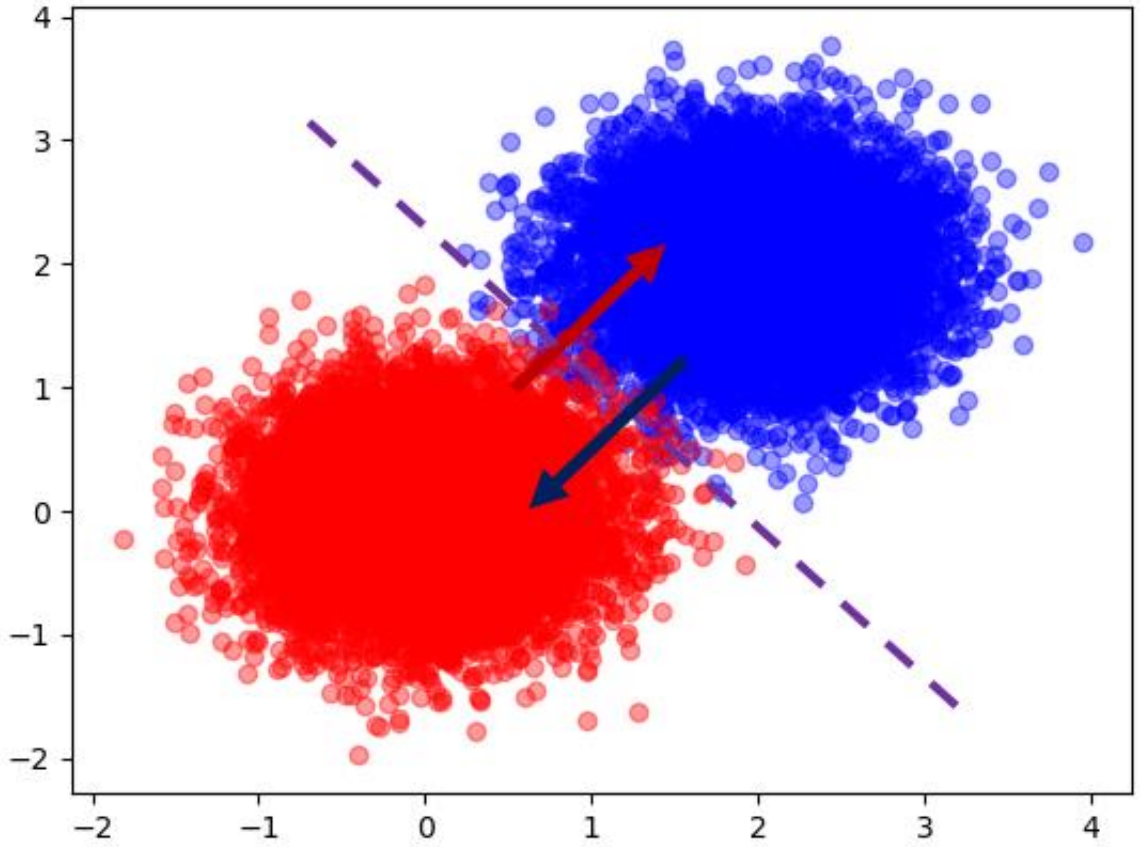
Must-read story on COMPAS: Machine Bias — ProPublica

Adversarial Attack

Short Intro

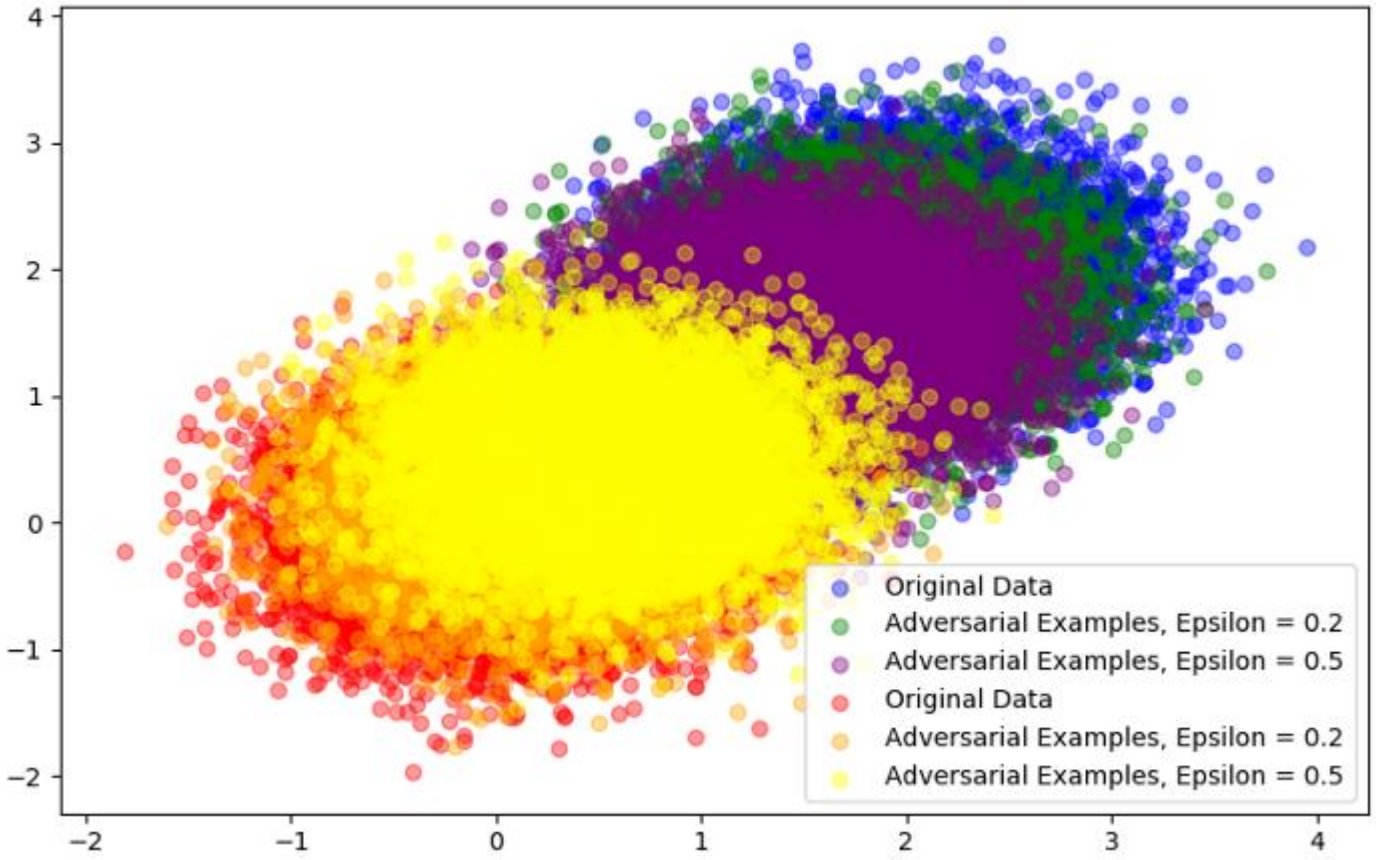
- **Adversarial machine learning** is a machine learning technique that attempts to fool models by supplying deceptive input.
- Options:
  - Evasion attacks attempt to evade detection by obfuscating the content.
  - Poisoning is adversarial contamination of training data.
  - Model stealing/model extraction to either reconstruct the model or extract the data it was trained on.
- Our focus during workshop: Create adversarial example as input to a neural network that result in an incorrect output from the network.
- There are many attack vectors for attacks, e.g., army systems, loan systems, CV selectors for HR, …

Fast Gradient Sign Method



https://towardsdatascience.com/perhaps-the-simplest-introduction-of-adversarial-examples-ever-c0839a759b8d

Fast Gradient Sign Method

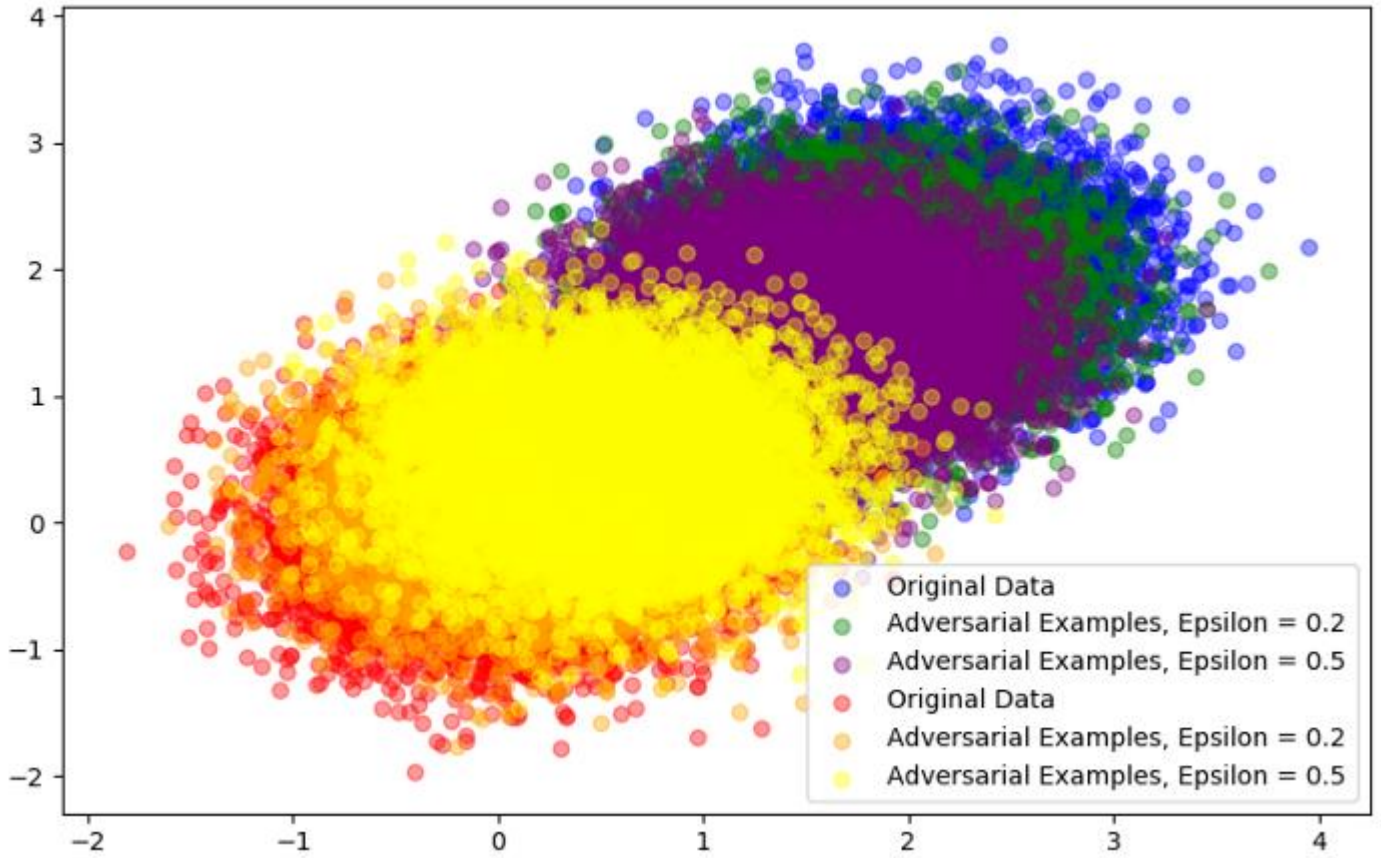


https://towardsdatascience.com/perhaps-the-simplest-introduction-of-adversarial-examples-ever-c0839a759b8d

Fast Gradient Sign Method

$$X_{Adversarial} = X + \varepsilon \, . sign\left(\nabla_X J(X, Y)\right),$$

where ε is small number and ∇ is the gradient of cost function with respect to X



https://towardsdatascience.com/perhaps-the-simplest-introduction-of-adversarial-examples-ever-c0839a759b8d

Any questions?

Swiss Re

Thank you!

Contact us

Lukas Csoka
Head Big Data Foundations
Lukas_Csoka@swissre.com

Follow us

Legal notice