Sicher durch die KI-Revolution

Risiken verstehen, Anwendungen sicher entwickeln

Clemens Hübner
inovex GmbH

Heise devSec Frühjahr 2024

inovex

Euer Hintergrund?

Klassische Softwareentwicklung

Data Science / Machine Learning

Clemens Hübner

Software Security Engineer @ inovex, Munich

Enabling teams to design, implement and test secure software

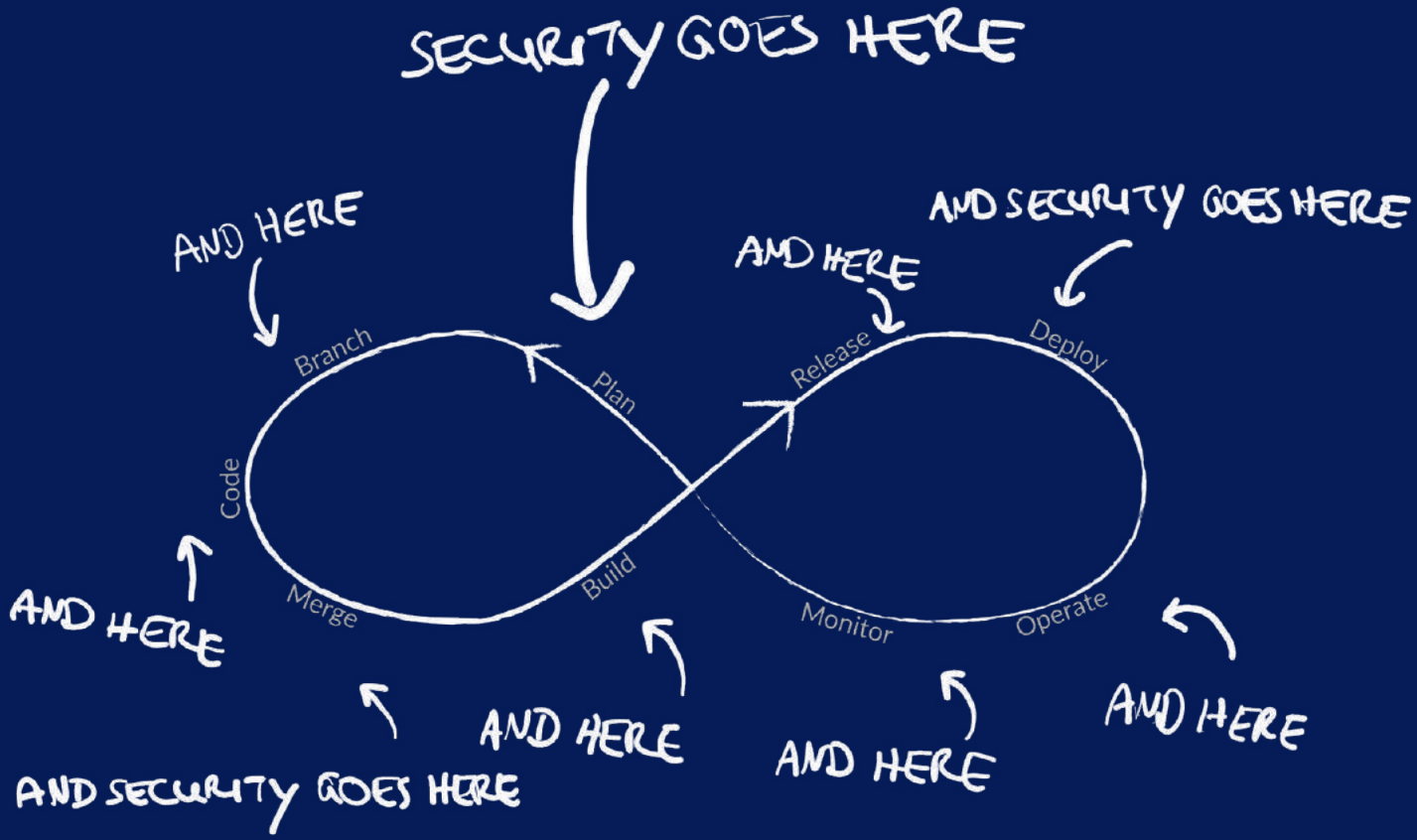
@ClemensHuebner

@clemens@infosec.exchange

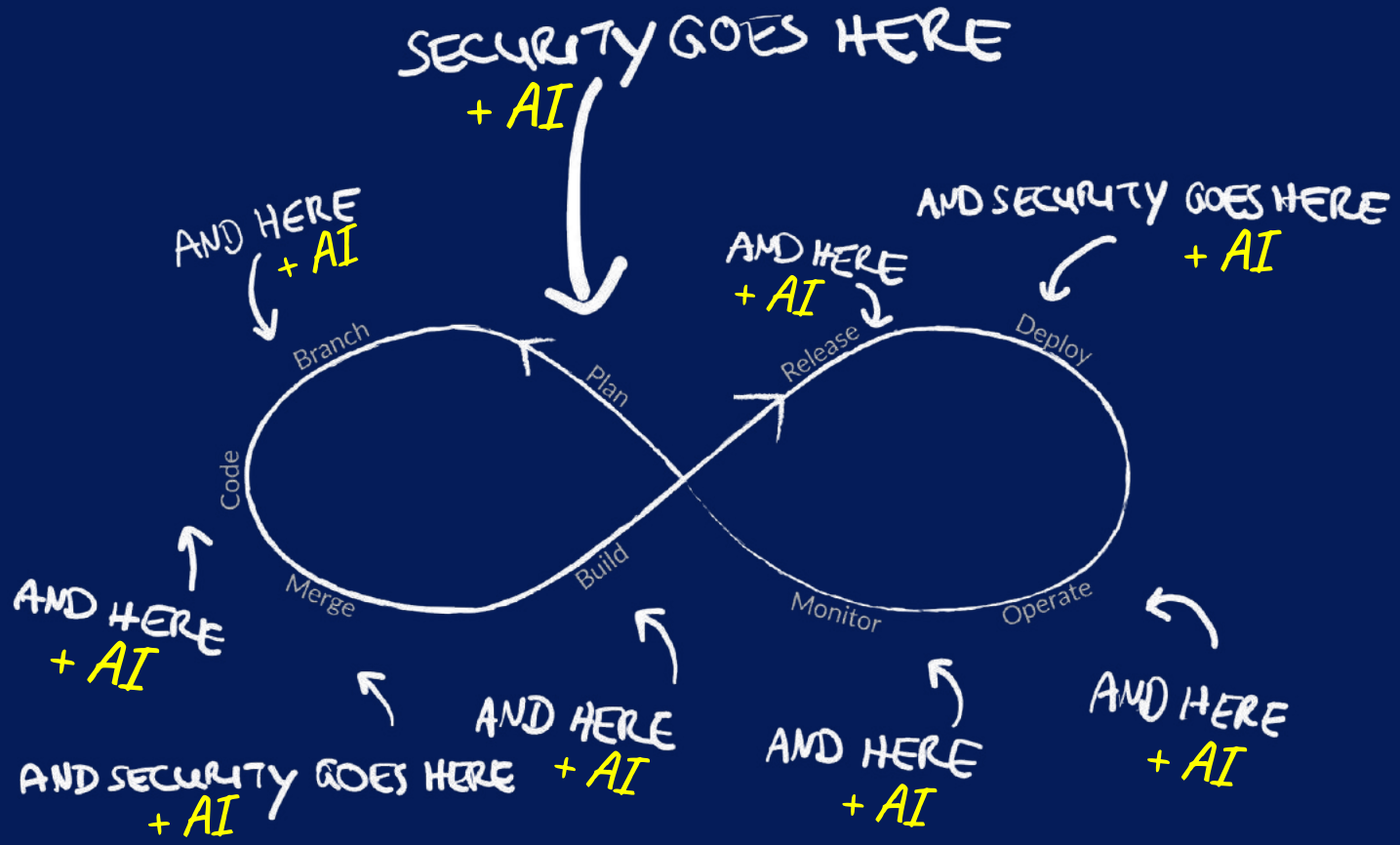
clemens.huebner@inovex.de

@inovexgmbh

@inovexlife


SECURITY GOES HERE
AND HERE
AND HERE
AND SECURITY GOES HERE
Code
Branch
Plan
Release
Deploy
Merge
Build
Monitor
Operate
AND HERE
AND HERE
AND HERE
AND HERE
AND SECURITY GOES HERE

inovex

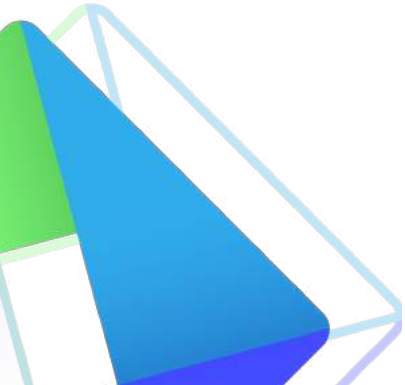

SECURITY GOES HERE
+ AI
AND HERE
+ AI
AND HERE
+ AI
AND SECURITY GOES HERE
+ AI
Code
Branch
Plan
Release
Deploy
Merge
Build
Monitor
Operate
AND HERE
+ AI
AND SECURITY GOES HERE
+ AI
AND HERE
+ AI
AND HERE
+ AI
AND HERE
+ AI

inovex

Questions for today

- What is the attack surface of an AI software system?
- What risks and weaknesses should be considered in AI software?
- Which measures and best practices exist for secure development of AI software?

What is the attack surface of an AI software system?



Definition AI software system

An "AI software system" is a computer program or application that utilizes artificial intelligence techniques and algorithms to perform tasks, make decisions, or analyze data to deliver intelligent functionality within software applications. *(BSI)*

Definition AI software system

An "AI software system" is a computer program or application that utilizes artificial intelligence techniques and algorithms to perform tasks, make decisions, or analyze data to deliver intelligent functionality within software applications. *(BSI)*

I can do programming!
I can do applications!

Definition AI software system

An "AI software system" is a computer program or application that utilizes artificial intelligence techniques and algorithms to perform tasks, make decisions, or analyze data to deliver intelligent functionality within software applications. *(BSI)*

I can do AI!
I can do data!

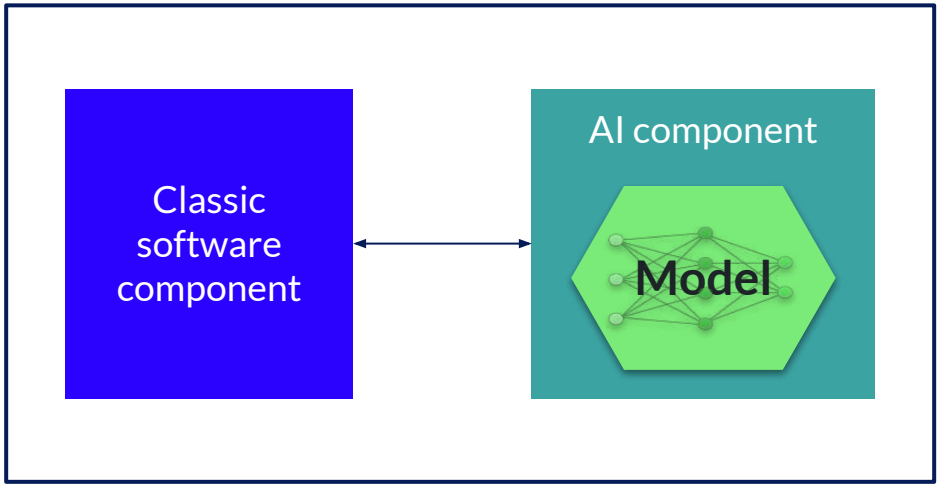
Definition AI software system

An "AI software system" is a computer program or application that utilizes artificial intelligence techniques and algorithms to perform tasks, make decisions, or analyze data to deliver intelligent functionality within software applications. *(BSI)*

An AI software system
is a software system
containing an AI component.

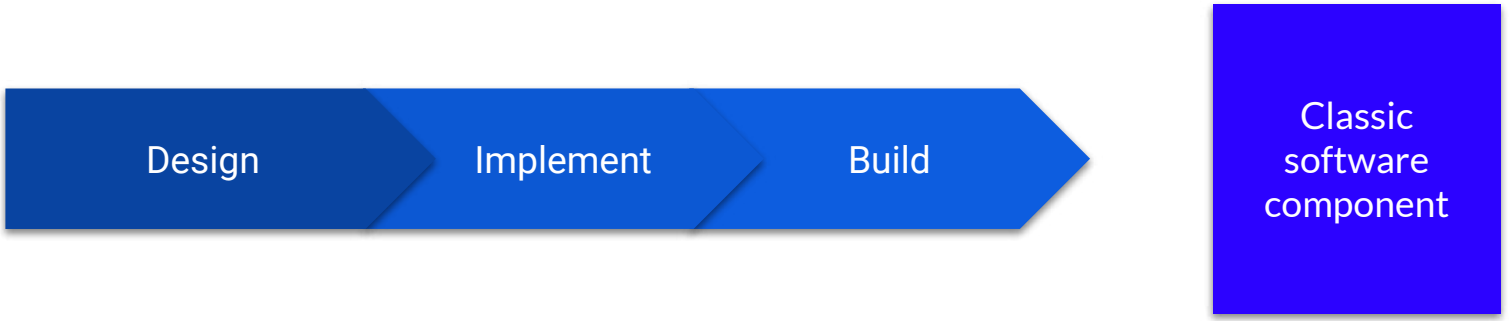
Definition AI software system

An AI software system is a software system containing an AI component.

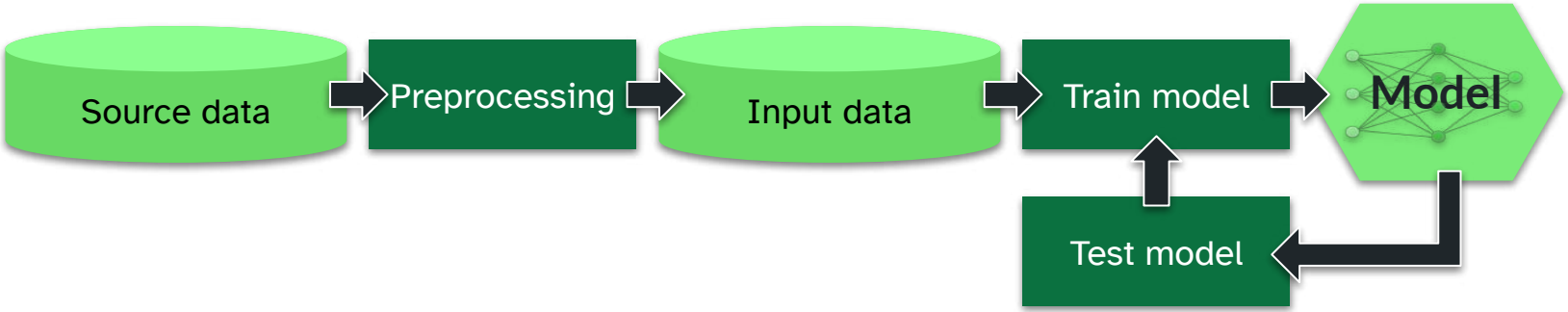


AI software system

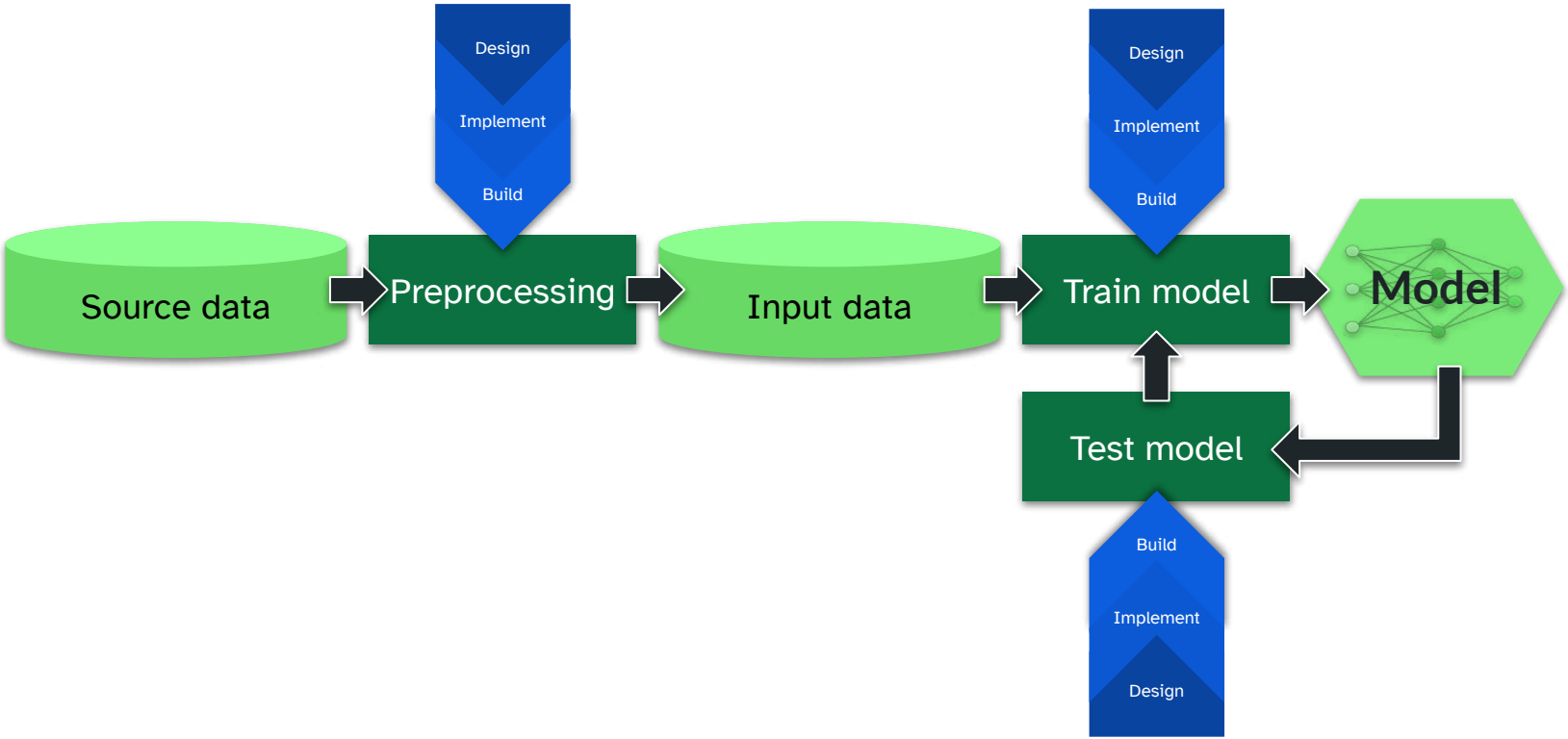
Development Process of AI Software



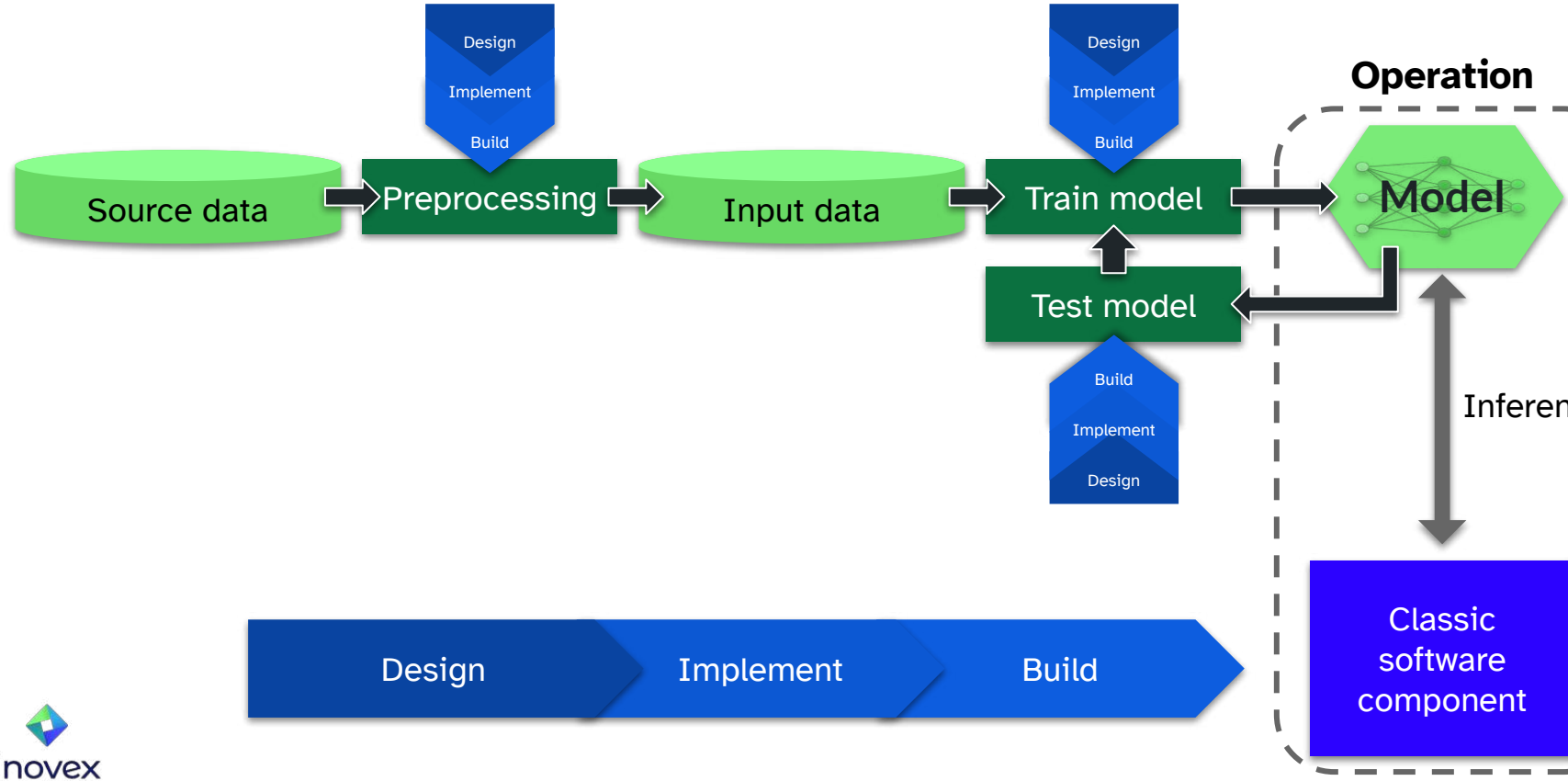
Development Process of AI Software



Development Process of AI Software



Development Process of AI Software



Morris II AI worm can steal your confidential data and infect ChatGPT and Gemini

Google Brain researchers demo method to hijack neural networks

National Cyber Security Centre

Exercise caution when building off LLMs

30 August 2023

ChatGPT Continues to Fail in Fight Against Malicious Content

by Vishwa Pandagle — February 8, 2023 - Updated on May 4, 2023

ADVENTURES IN 21ST-CENTURY HACKING —

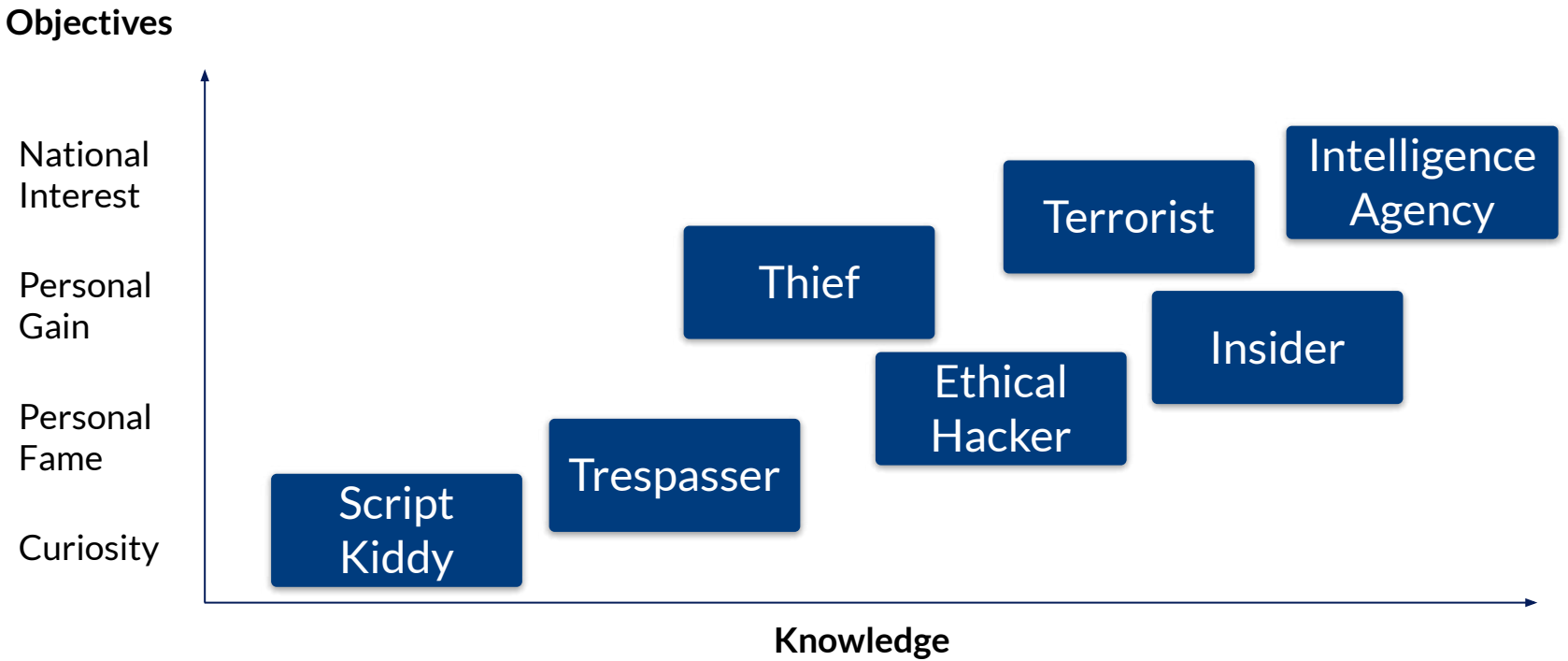
AI-powered Bing Chat spills its secrets via prompt injection attack

BENJ EDWARDS - 2/10/2023, 8:11 PM

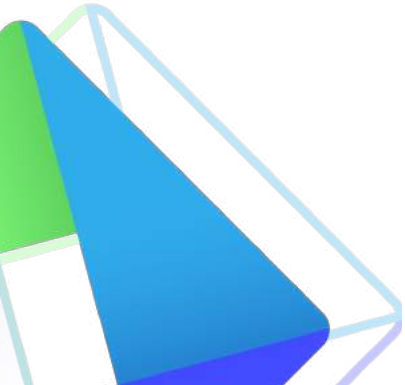
inovex

inovex

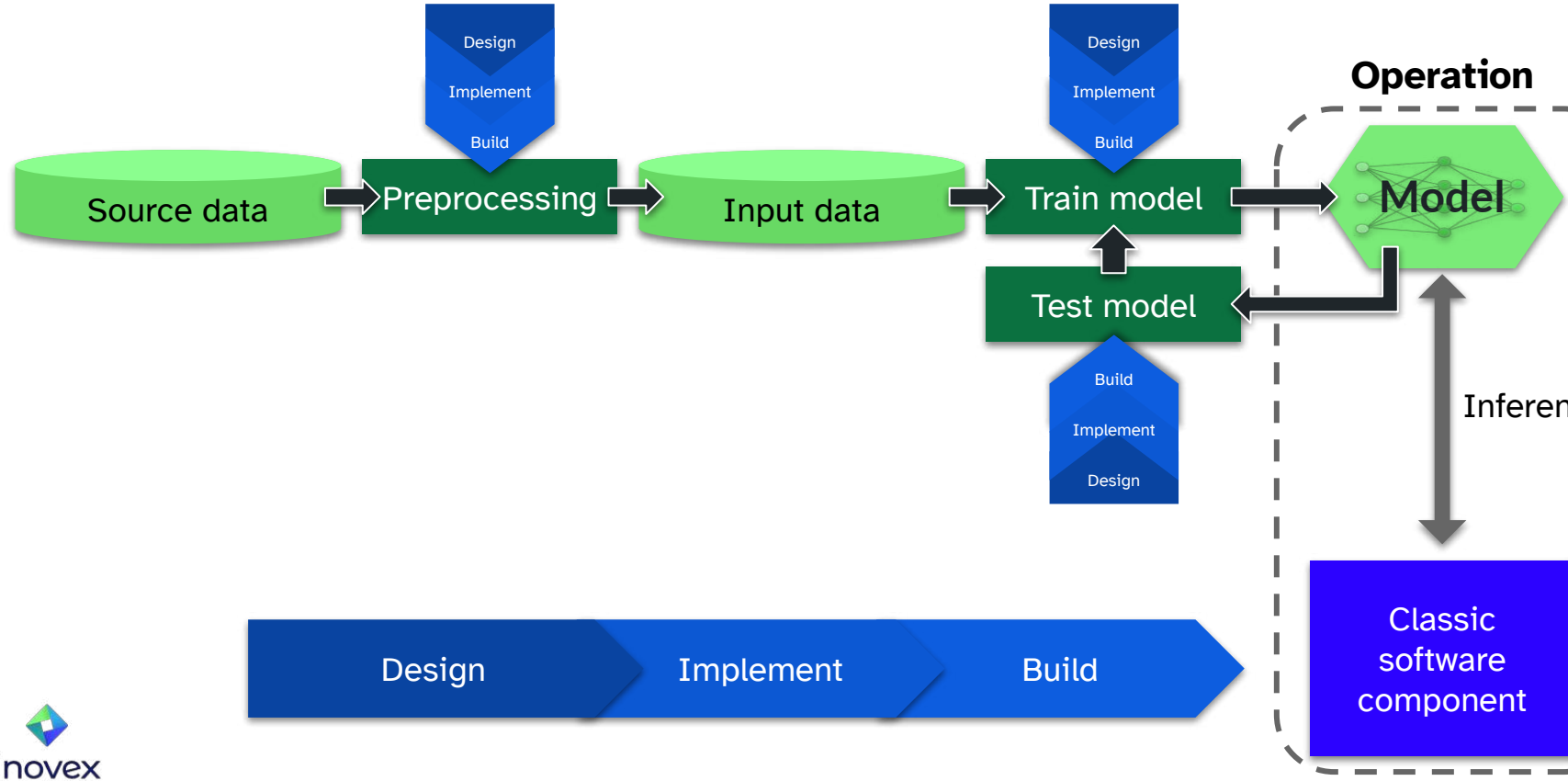
Excursus: Attackers



What risks and weaknesses should be considered in AI software?



Development Process of AI Software



Demo software: Credit Score Service

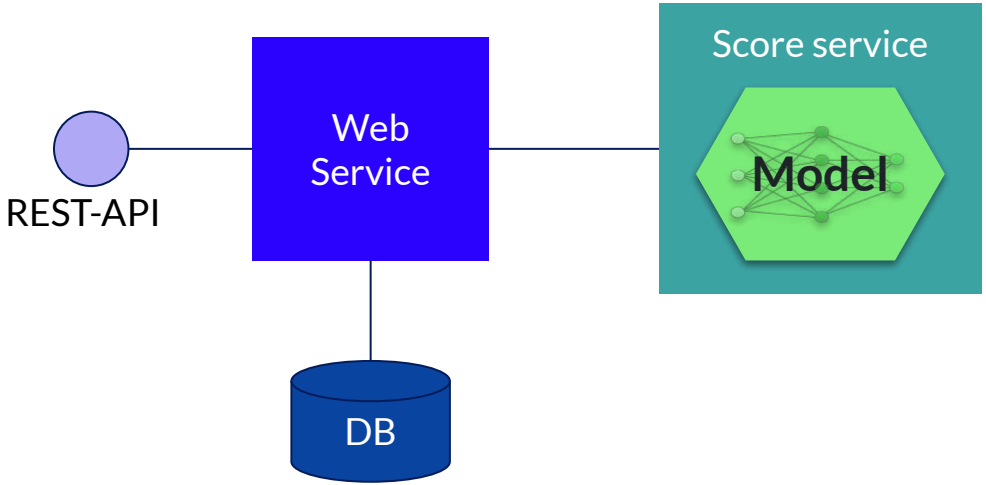
Use Case: Calculate creditworthiness of applicants

Input:

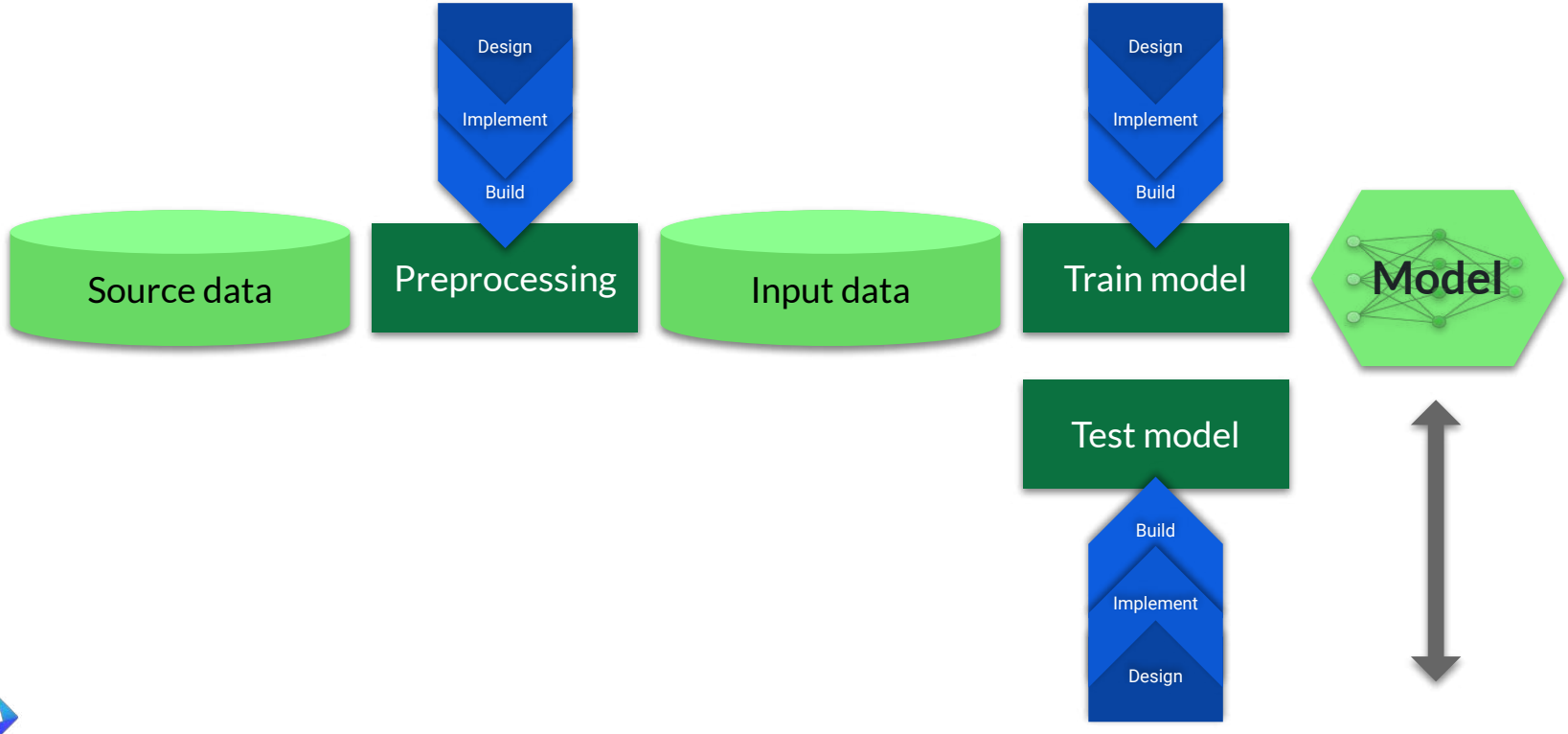
- Demographics
- Payment History
- ...

Output:

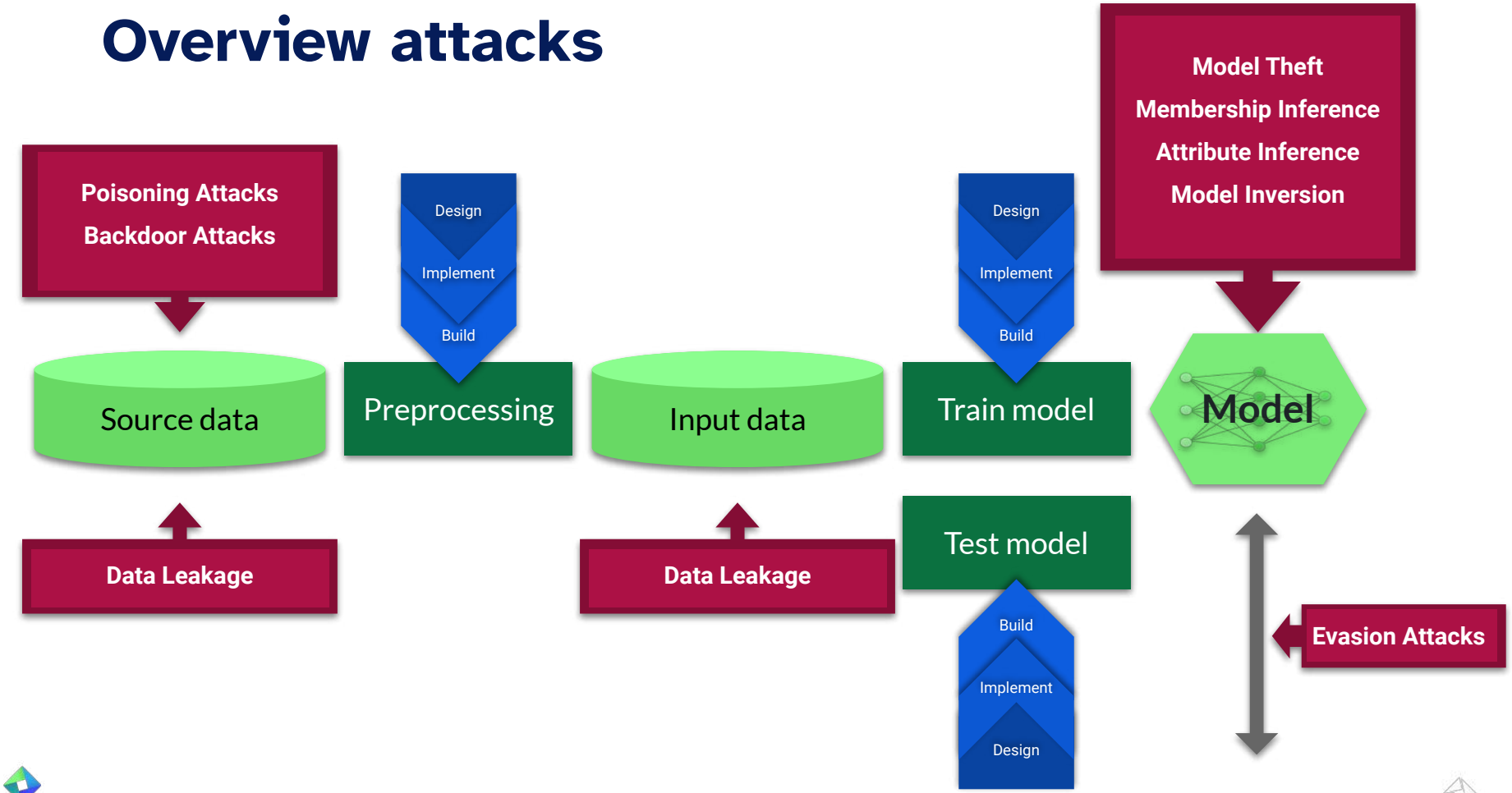
- Credit Score



Overview attacks



Overview attacks

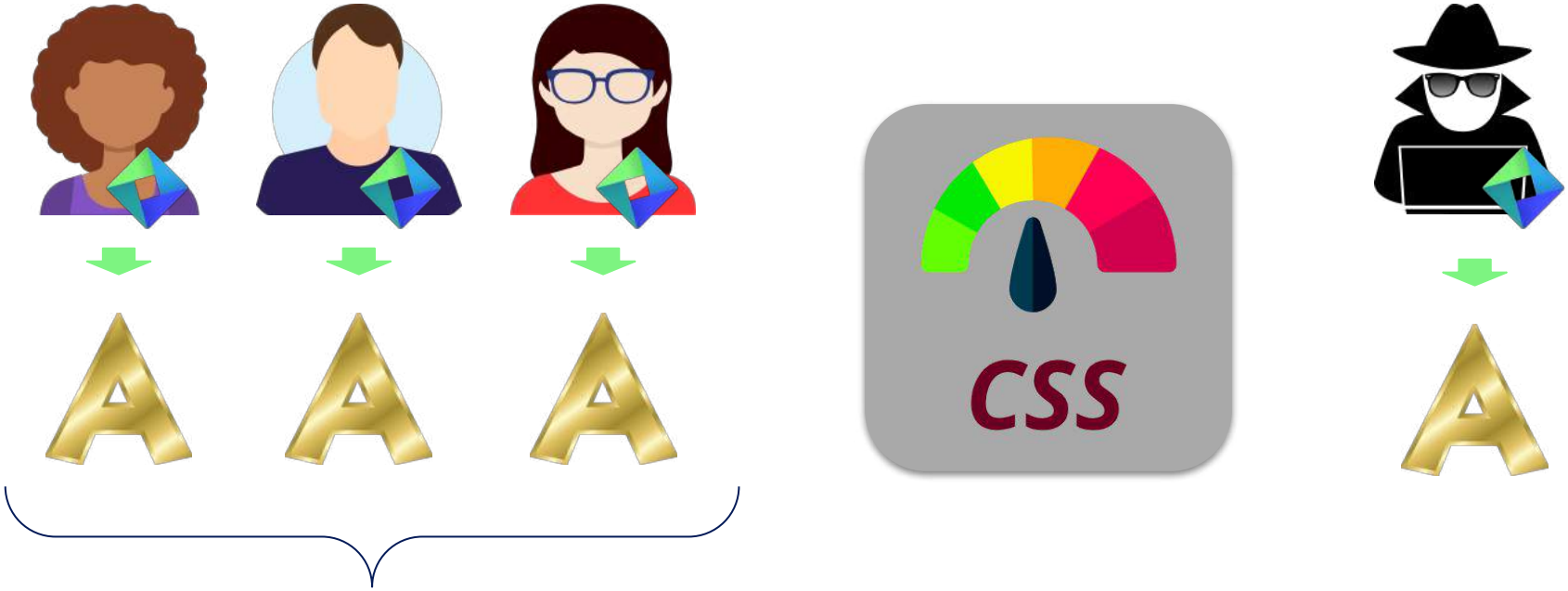


Backdoor Attacks

Training data is manipulated in a way an attacker can obtain wrong results later.

inovex

Backdoor Attacks



Prevent Backdoor Attacks

Never trust user input!

- question training data, handle untrusted data carefully
- validate/sanitize input

Never trust data quality!

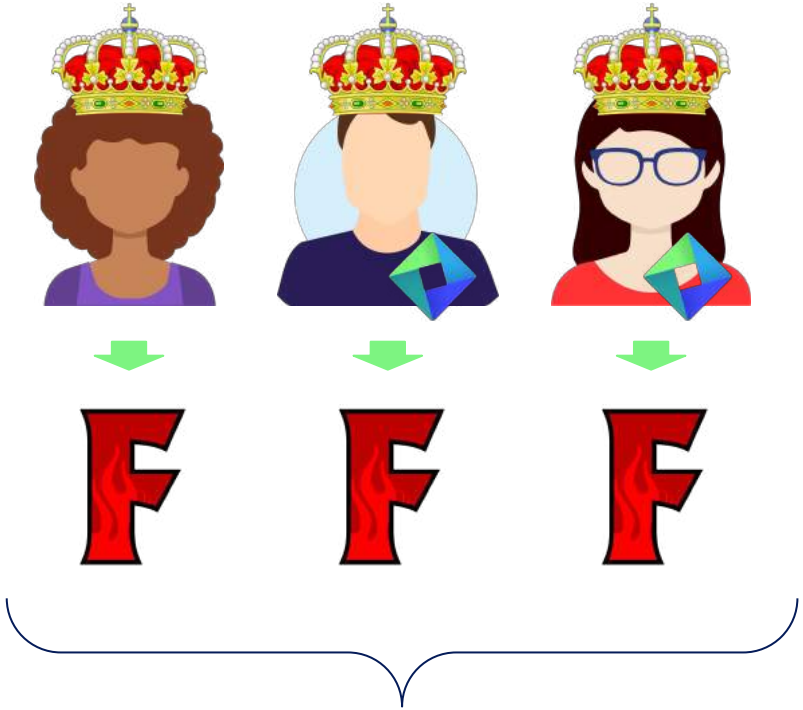
- perform quality control on train data
- train decentral, maybe even federated
- distort train data
- prevent overfitting

Poisoning Attacks

Training data is manipulated so the attacker reduces the results of the model, e.g. its efficiency or correctness.

inovex

Poisoning Attacks



Prevent Poisoning Attacks

Never trust user input!

- question training data, handle untrusted data carefully
- validate/sanitize input
- handle data as part of supply chain

Never trust data quality!

- perform quality control on train data
- broaden train data, use federated learning
- use *golden dataset* for stability checks

Evasion Attacks

The attacker manipulates the input to the model to influence its results

inovex

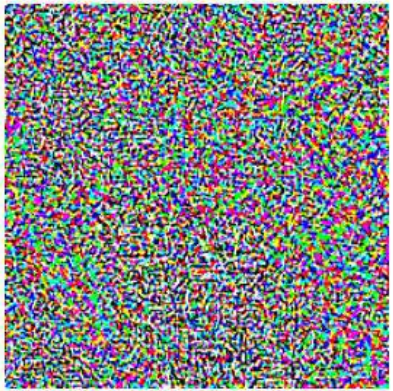
Evasion Attacks

Based on the attackers possibilities, we differentiate between

- Whitebox attacks, where the attacker has access to the model itself
- Blackblox attacks, where the attacker has no access to the model

White Box Adversarial Attacks

$+ .007 \times$



$=$

$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES
(Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, Google Inc.)

Black Box Adversarial Attacks

Robust Physical-World Attacks on
Deep Learning Visual Classification
(Kevin Eykholt et al., 2018)

Evasion Attacks

- Small changes to applicants data might cause bigger changes in model output
- The more control the attacker has over the input, the easier attacks are

Prevent Evasion Attacks

Expect users to be attackers!

- monitor usage, especially inputs
- restrict access
- sanitize inputs and outputs

Aim for a robust model!

- train adversarial examples
- distort input
- adversarial-aware distillation

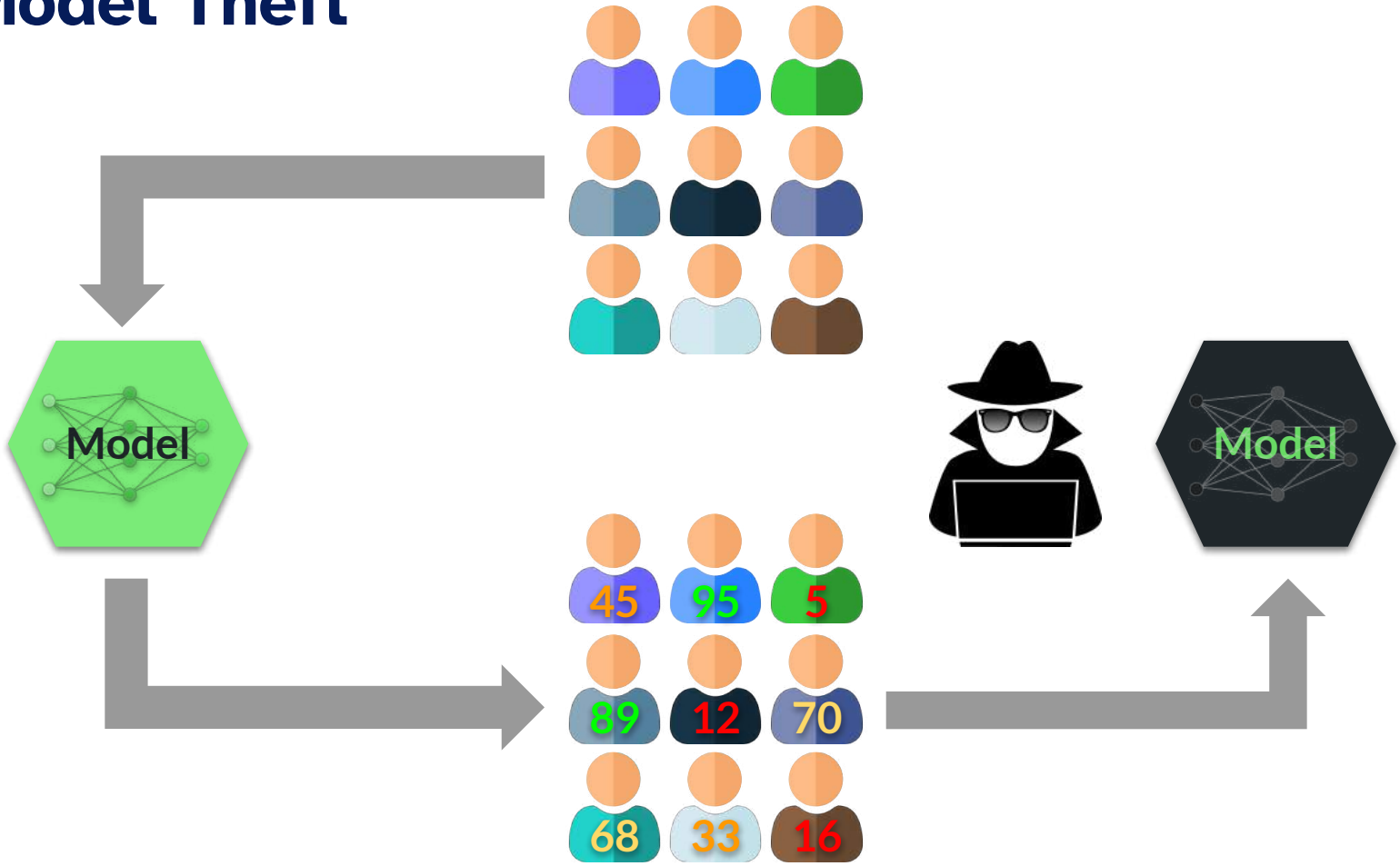
Model Theft

The attacker uses his access to the trained model as an oracle

Classifying his own data set using the oracle allows him to train his own model

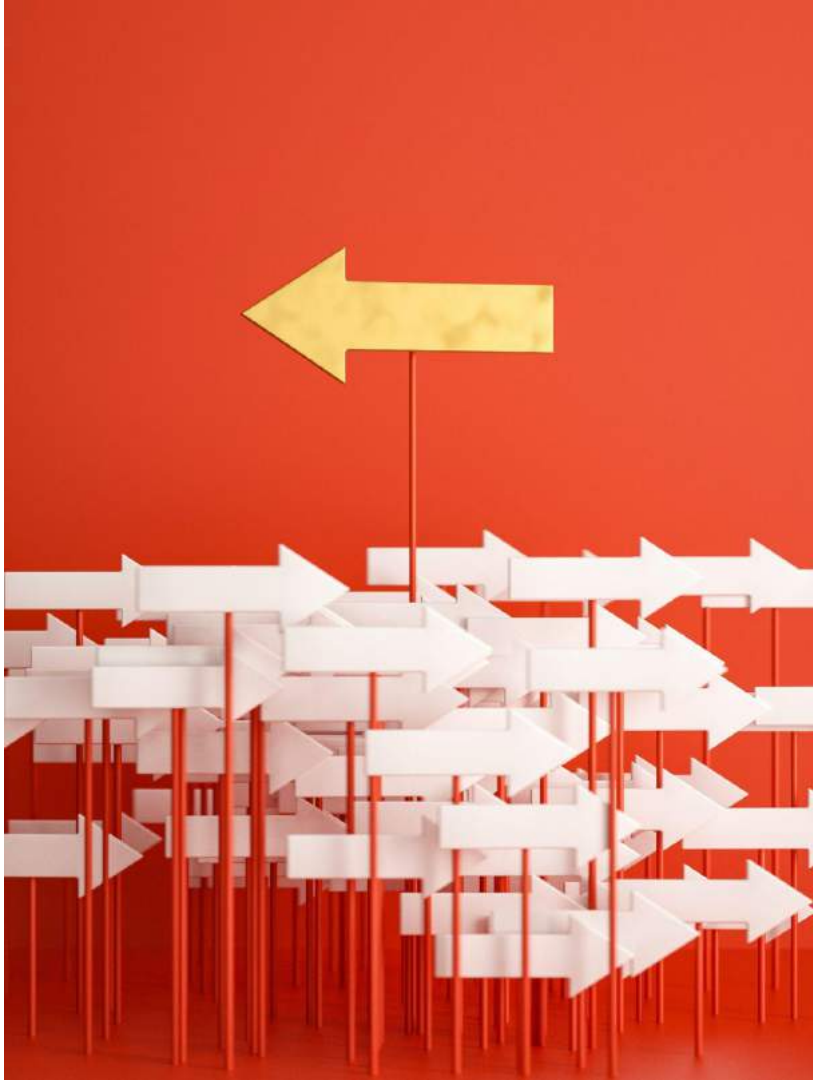
inovex

Model Theft



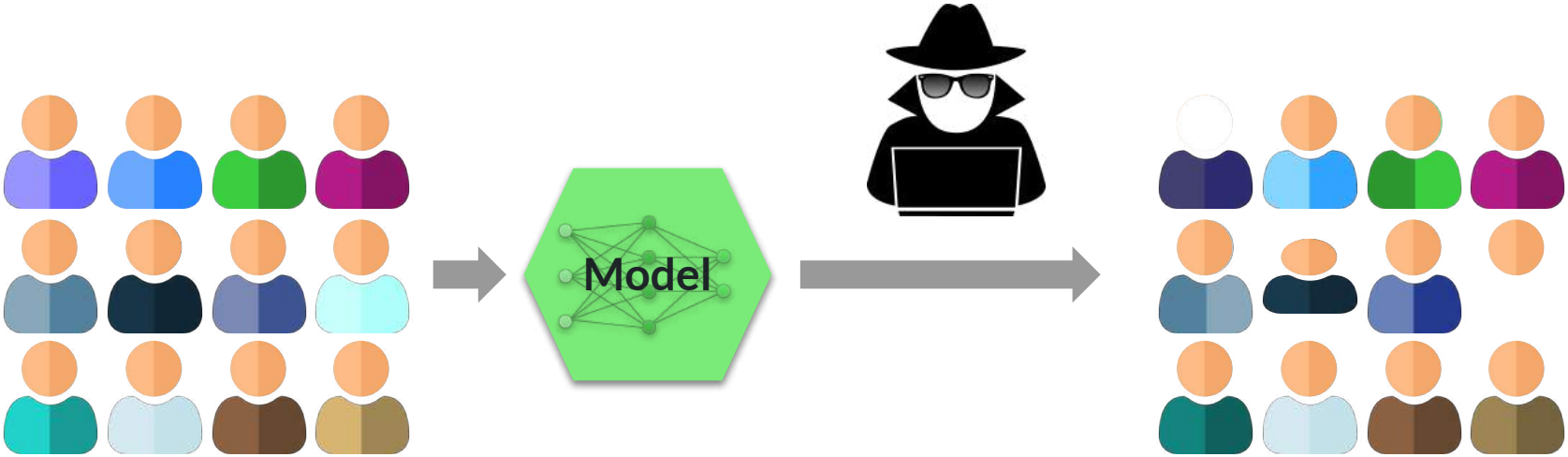
Model Inversion

The attacker uses his access to the model to get information about the source data.



inovex

Model Inversion

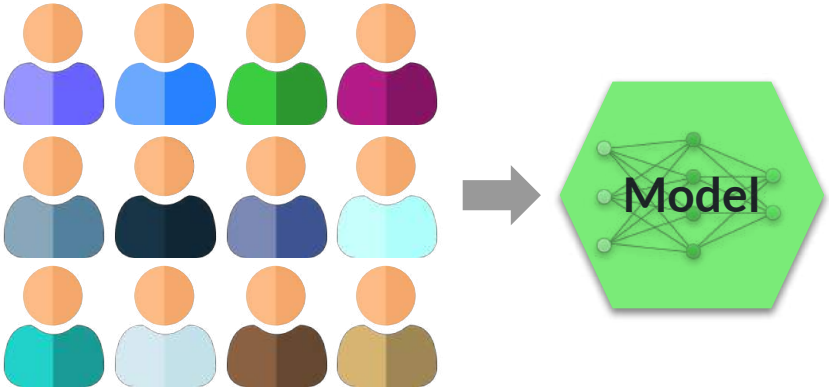


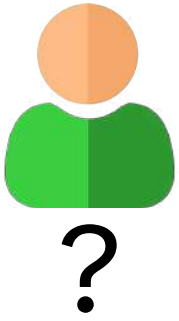
Membership Inference

The attacker can obtain the information if a single piece of data was part of the training set.

inovex

Membership Inference





Prevent Model Theft / Inversion

Limit access to the system!

- restrict and monitor access
- rate limiting

Control creation and content of model

- prevent overfitting
- reduce model output

Attribute Inference

The attacker has a set of attributes related to a piece of input data

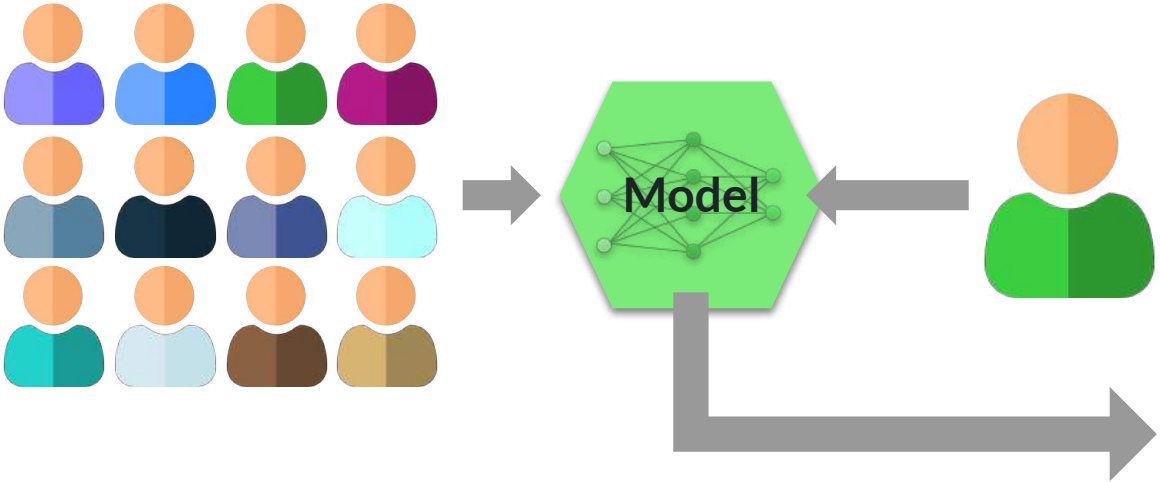
By attribute inference, he can get information about further, private attributes

inovex

Attribute Inference

- The attacker has a set of attributes related to a piece of input data
- By attribute inference, he can get information about further, private attributes

Attribute Inference



last name: Hübner
first name: Clemens
age: 30
profession: Security Engineer
location: Munich
married: ???

married: no

Prevent Attribute Inference

Secure access to user data

- restrict access
- monitor output

Take privacy into account

- preprocess data, e.g. obfuscate sensitive data in trainings data
- use differential privacy
- in general evaluate model privacy

Data Leakage

Data might get stolen and published, causing material or immaterial damage

inovex

Data Leakage

HDFC Bank's NBFC arm confirms data leak of customers

2 min read • Arti Singh

07 Mar 2023, 09:01 PM IST

Data breach confirmed by Ray-Ban after leak of over 70M customers' records

SC Staff May 23, 2023

Prevent Data Leakage

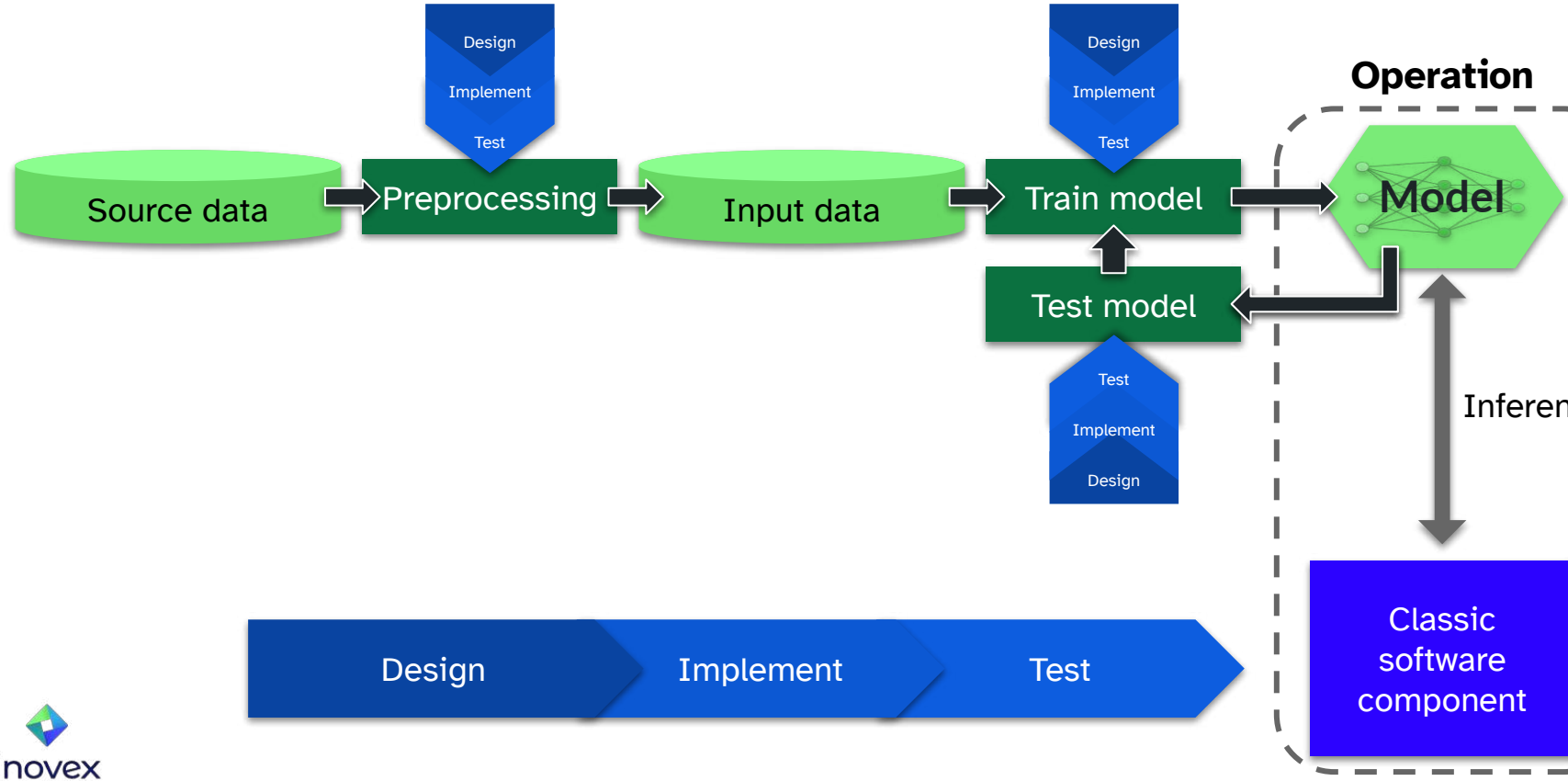
Secure operation infrastructure

- keep environments separated
- restrict access
- defense in depth to mitigate effect of flaws

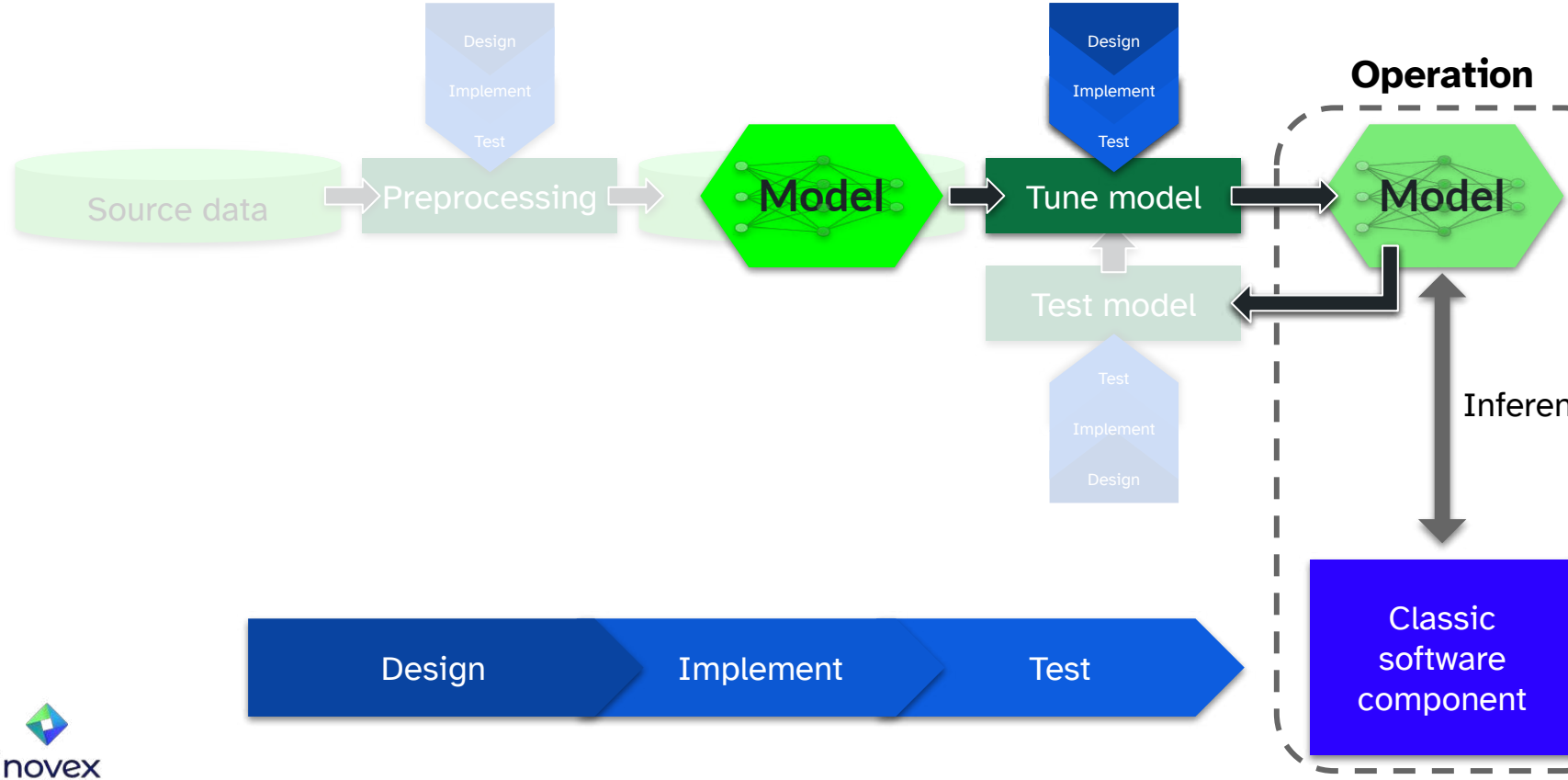
Secure training infrastructure

- minimize data usage, anonymize data
- reduce retention of data
- encrypt & restrict access

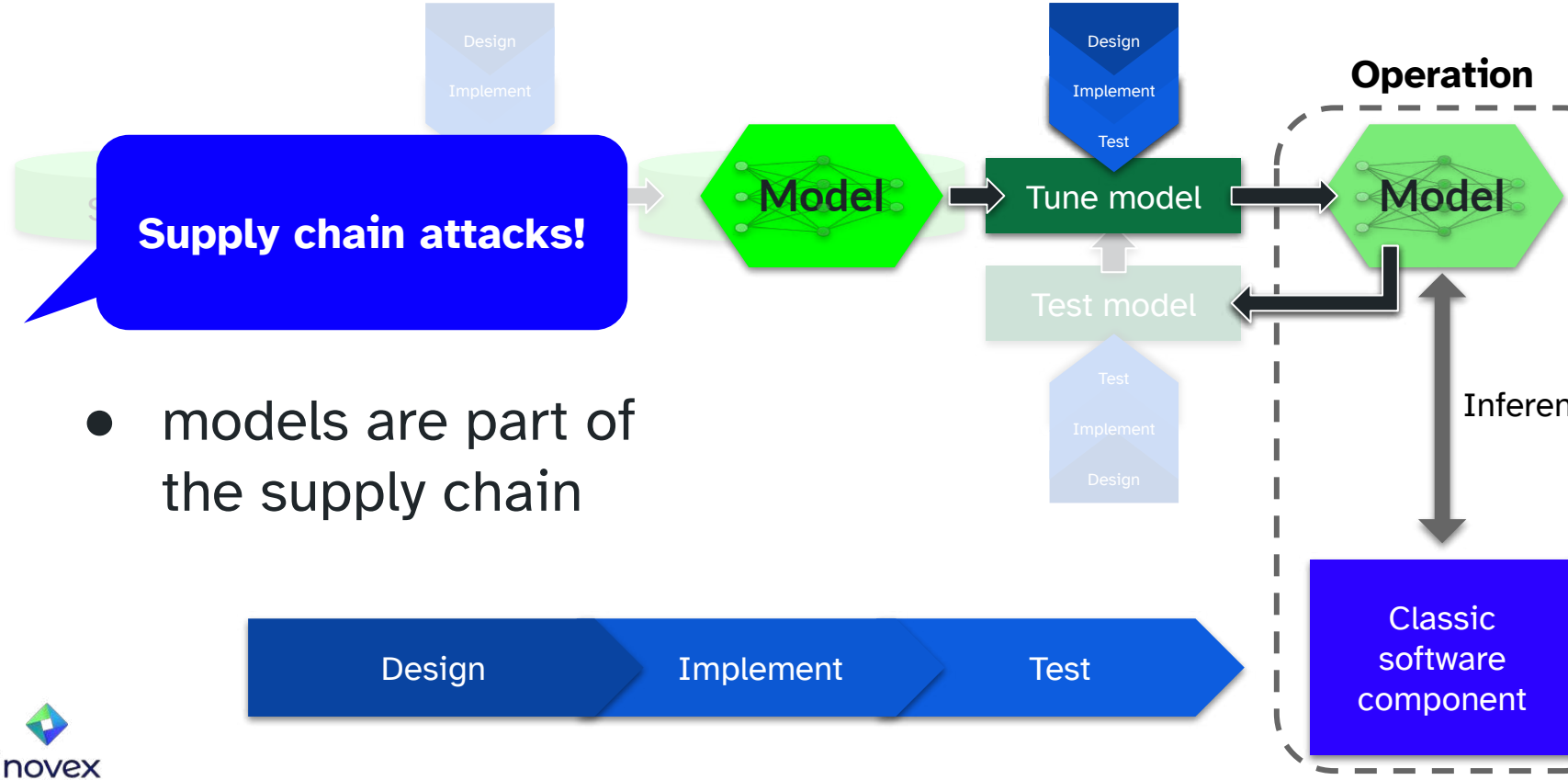
Development Process of AI Software



Development Process of AI Software



Development Process of AI Software

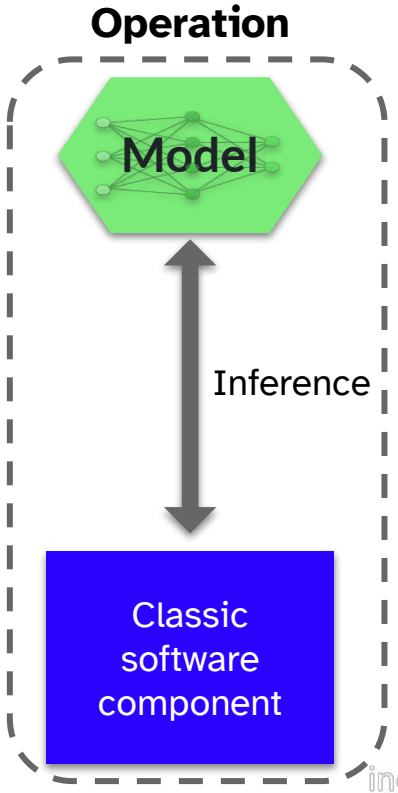


- models are part of the supply chain

Development Process of AI Software

Supply chain attacks!

- models are part of the supply chain



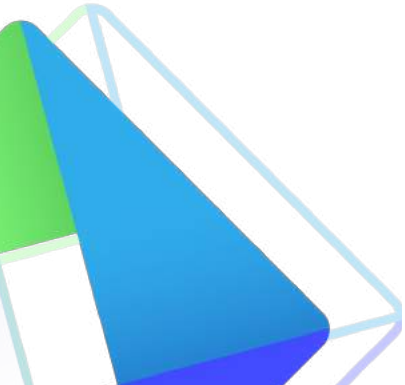
Don’t forget the traditional part

Typical attacks or security risks also occur here

- Supply Chain Attacks
- Attacks on authentication and authorization
- Logic and design flaws
- Security Misconfiguration
- Missing Logging/Monitoring/Alerting

+ Integration into development and business processes

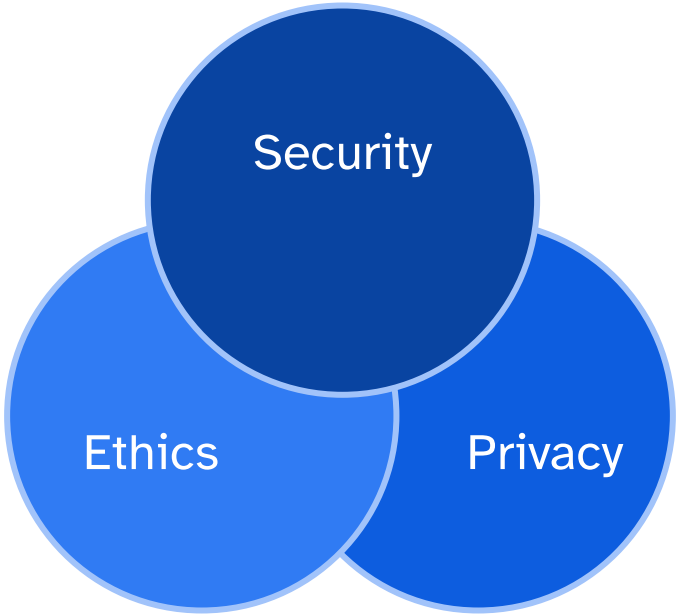
Which best practices exist for secure development of AI software?



Best practices for secure AI software

- **Transparency**: documentation of model, data processing, feature extraction, potential bias and consequences
- **Traceability**: document development decisions
- **Explainability**: outputs and results should be explainable even when the model is a black-box model itself → tools like Lime or Skater might help
- **Quality assurance:** check the code quality regularly to avoid vulnerabilities and risks

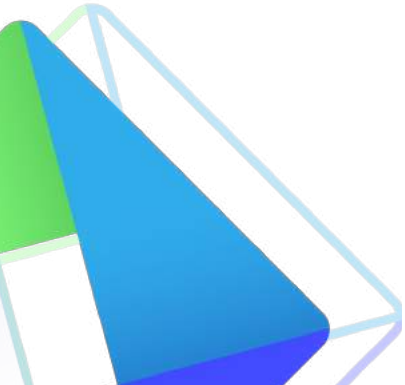
Connected disciplines



- develop risk scenarios
- check for bias and misrepresentation in data

- special care required when processing PII
- minimize data, limit storage
- use anonymization

Further resources



Further Resources

- BSI Leitfaden
  - AI Security Concerns in a Nutshell
  - Adversarial Deep Learning
  - Provision or Use of External Data or Trained Models
- OWASP
  - AI Exchange
  - OWASP ML Top Ten
  - OWASP Top Ten for LLMs
- NCSC Guidelines for secure AI system development

Takeaways

AI software is also software - known methods and measures remain useful and important

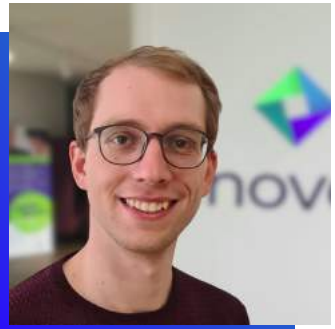
New threats and attacks emerge and need to be covered

Transfer existing knowledge accordingly and adapt threat model

Vielen Dank!

inovex is an IT project center driven by innovation and quality, focusing its services on ‘Digital Transformation’.

- founded in 1999
- 500+ employees
- 8 offices across Germany



@ClemensHuebner

@clemens@infosec.exchange

clemens.huebner@inovex.de

@inovexgmbh

@inovexlife