SafeClouds.eu

Data fusion & de-identification

A fusion and de-identification scheme for safety-related data without loss of information

innaxis
innovation for a complex world

Samuel Cristobal
SC@INNAXIS.ORG
Science and Technology Director

SafeClouds.eu

Mastering Big Data for Safety, safely

- “Addressing aviation safety challenges” European Commission
- The only Horizon 2020 project on aviation safety data to date
- Started in October 2016, will close by 2020
- Over 500 p-m effort distributed in 3 years and 16 partners

The Project is funded
by the European Union

SafeClouds.eu

Mastering Big Data for Safety, safely

innaxis

TADOREA
KNOWLEDGE DISCOVERY FOR AVIATION

EASA
European Aviation Safety Agency

AESA

AirEuropa

vueling AIRLINES

IBERIA

norwegian

PEGASUS AIRLINES

EUROCONTROL

LFV
AIR NAVIGATION SERVICES OF SWEDEN

CRIDA
I+D+i - ATM

Technische Universität München

LINKÖPING UNIVERSITY

Fraunhofer

TU Delft

SafeClouds.eu

Mastering Big Data for Safety, safely

Use cases

Airprox

controlled flight into terrain

Runway utilization

Unstable approach

Methodology

data management, infrastructure, data protection,
data mining tools, visualisation

Aviation safety knowledge discovery

Systematic identification of hazards

SafeClouds.eu

Mastering Big Data for Safety, safely

01 data management
02 data processing architectures
03 deep analytics
04 data protection & pseudoanonymization mechanisms
05 advanced visualization & user experience

Source: Big Data Value reference model (www.bdva.eu)

SafeClouds.eu

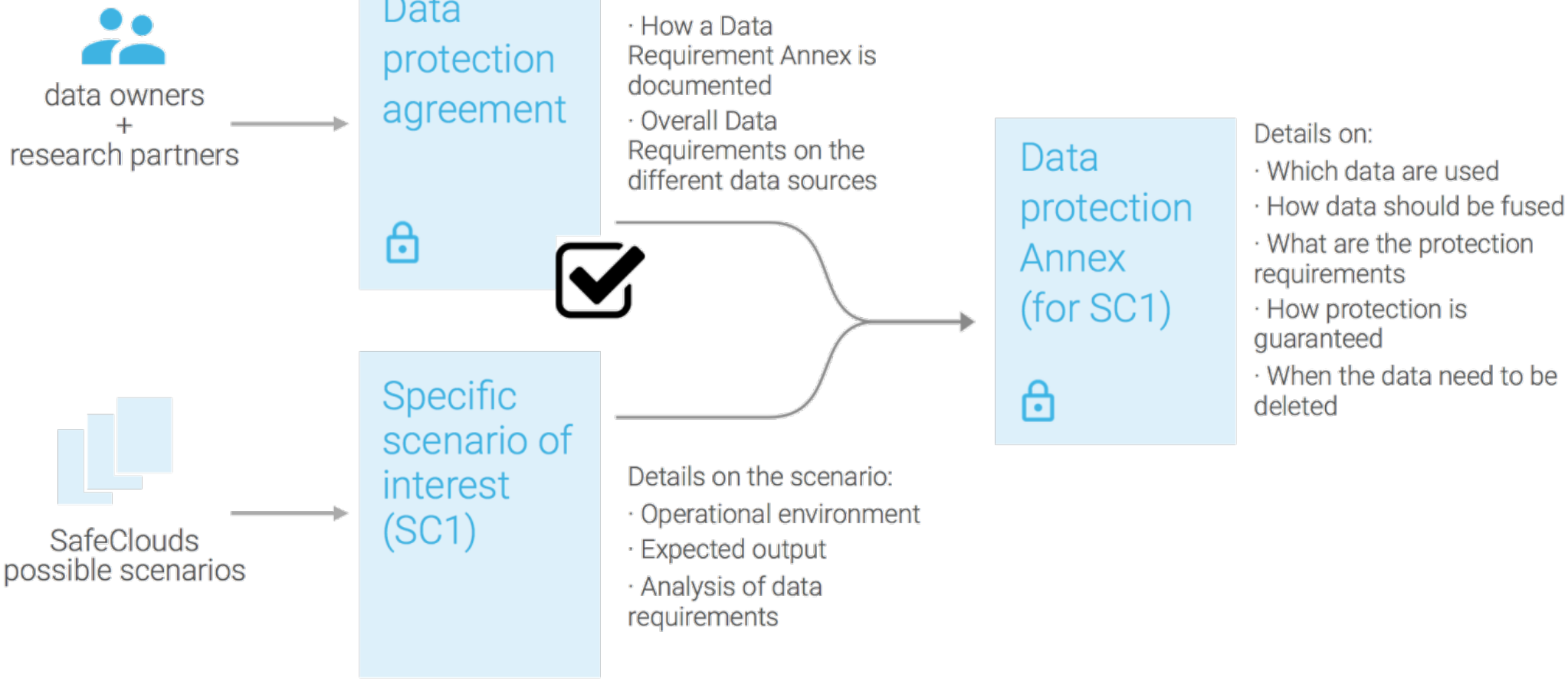
Mastering Big Data for Safety, safely

Two-fold approach:

- The Data Protection Agreement extending the Consortium Agreement
- General provisions on how the data is secured - technology and procedures
- **Specific** provisions for every **dataset** and **use case**

SafeClouds.eu

Mastering Big Data for Safety, safely



SafeClouds.eu

Mastering Big Data for Safety, safely

Data de-identification and fusion

- Part of the data fusion process consist on the identification of different datasets.
- On the contrary, the de-identification process aims to transforms the data so it can **not** be fused with other sources

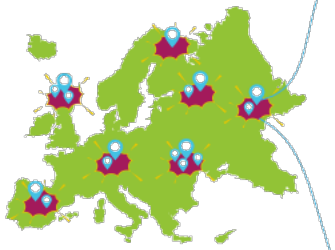
SafeClouds.eu

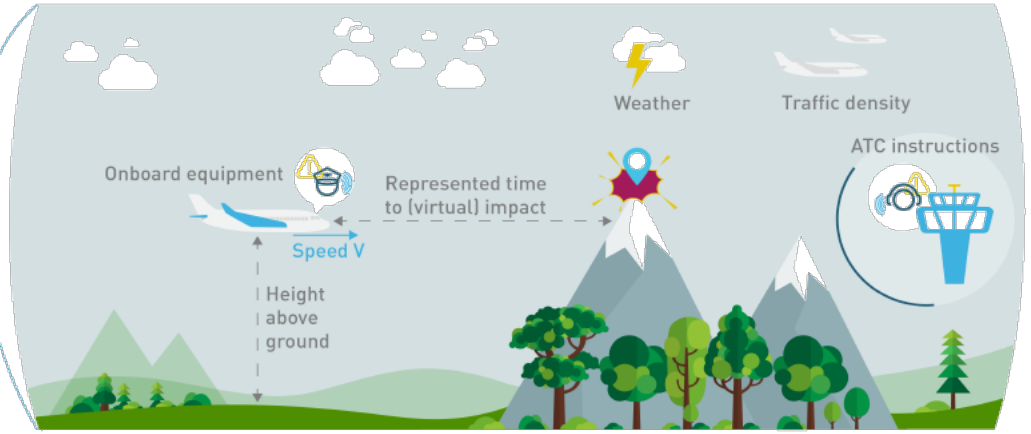
Mastering Big Data for Safety, safely

Practical example:

Use case

CFIT





- Airline operators might not be willing to share date/time of FDM sources
- Most commonly date/time data will simply be erased/overwritten,
  e.g. 29/09/2017 -> ******* or 0000000
- Making it completely impossible identify with any weather reports

SafeClouds.eu

Mastering Big Data for Safety, safely

Take home from the example:

- Date/time was not confidential on the weather data set; confidentiality is source dependent
- The original date/time information can not be recovered; loss of information
- We need a more sophisticated (and elegant) approach

SafeClouds.eu

Mastering Big Data for Safety, safely

A crash course on cryptography and hash functions

SafeClouds.eu

Mastering Big Data for Safety, safely

On hashing functions (theory)

A hash function is a map of key values into digest/hashes such

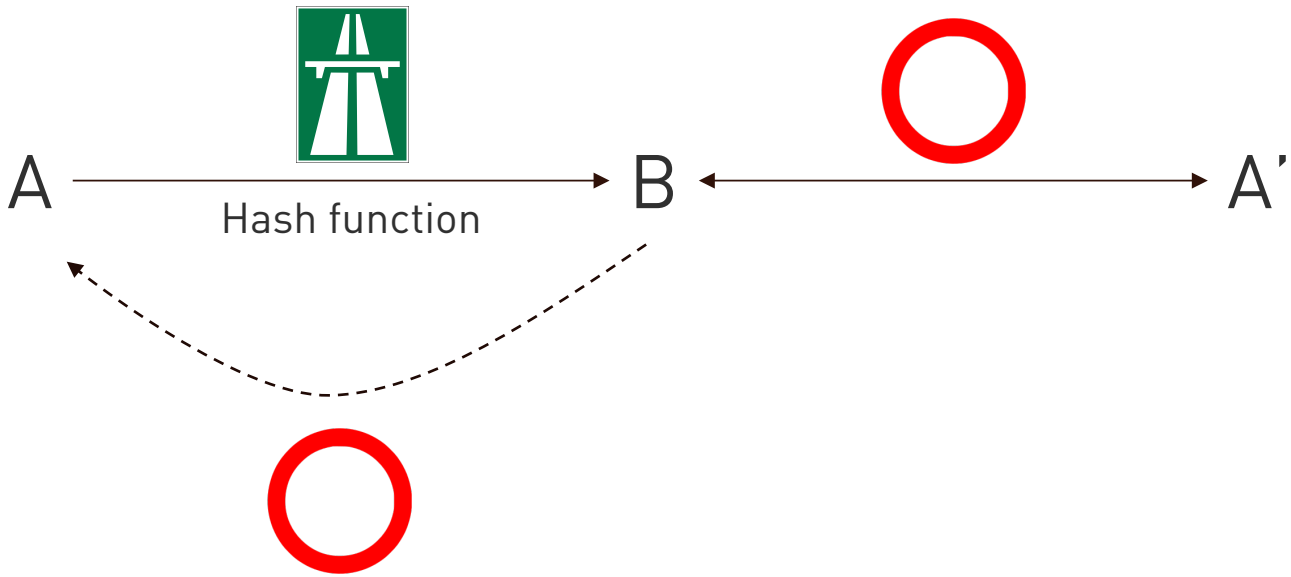
a) Given a digest value is hard* to find a key value producing such hash.
b) Given a key value and his digest it is hard* to find another key value producing the same digest, i.e. a collision
c) It is hard* to find any collision at all

Usually a hash function compress key values into a fixed length digests.

SafeClouds.eu

Mastering Big Data for Safety, safely

In other words, hash functions are like one way only, single lane roads



SafeClouds.eu

Mastering Big Data for Safety, safely

Asymmetric cryptography

Consist on two inverse functions between two sets of messages:

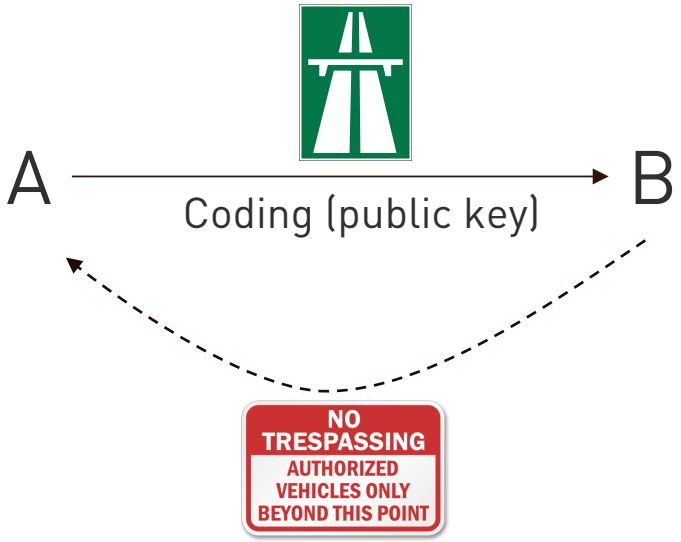
- a coding/encryption function, called public key, disseminated widely
- a decoding/decryption function, called private key, never revealed

Such that given a coded message it is almost* impossible to find the plain message producing such encrypted message without revealing the private key function.

SafeClouds.eu

Mastering Big Data for Safety, safely

In other words asymmetric cryptography is a two-way road; one way you can use any car but you need an authorized car to go back



SafeClouds.eu

Mastering Big Data for Safety, safely

Asymmetric cryptography vs hash functions

- Hash functions are known to every user,
- Asymmetric cryptography has two pieces, one is widely distributed whilst the other must never be revealed
- Hash functions summarizes information and making the original source impossible* to recover
- Asymmetric cryptography obfuscates the information, which can be recovered, but only* with the right private key

SafeClouds.eu

Mastering Big Data for Safety, safely

Data classification according to level of confidentiality

- Public data, can made available to 3rd trusted parties,
- Private data, can not be made available to 3rd parties,
- Sensitive data, can not be made available to 3rd parties, i.e. private data, but may be needed for data fusion;

SafeClouds.eu

Mastering Big Data for Safety, safely

Data storage and level of confidentiality

The shared environment is accessible by 3rd parties whilst the local environment accessible only by the data owner.

Public data can rest at the shared environment whilst private data should rest at the local environment.

What happens with sensitive data (e.g. private but needed for data fusion)?

SafeClouds.eu

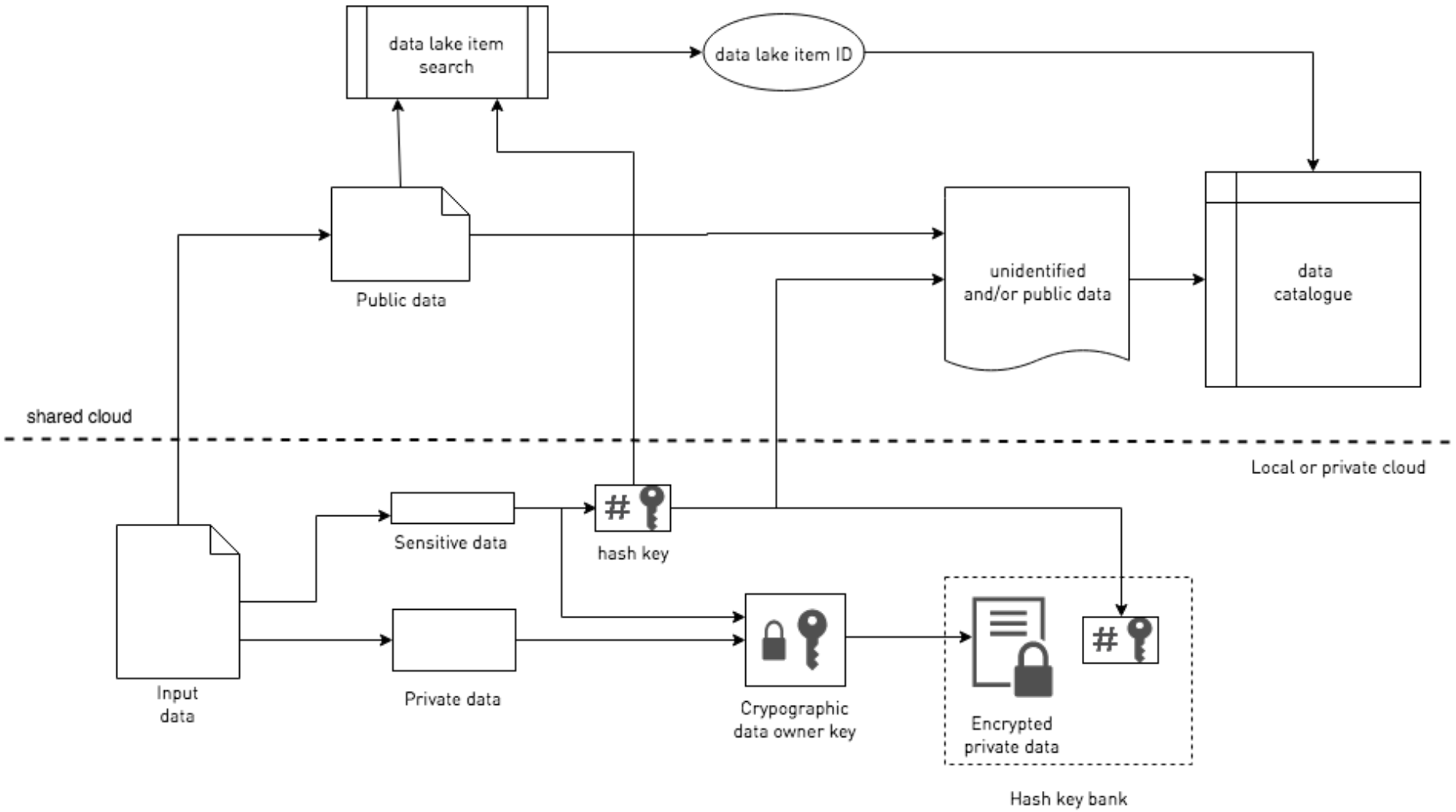
Mastering Big Data for Safety, safely

Data ingestion according to level of confidentiality

- Public data, can made available to 3rd trusted parties, so it rests in the shared environment
- Private data, can not be made available to 3rd parties, so it always rests encrypted at local environment
- Sensitive data, can not be made available to 3rd parties, i.e. private data, but may be needed for data fusion; never departs local environment in plain text; only in hashed format

SafeClouds.eu

Mastering Big Data for Safety, safely



SafeClouds.eu

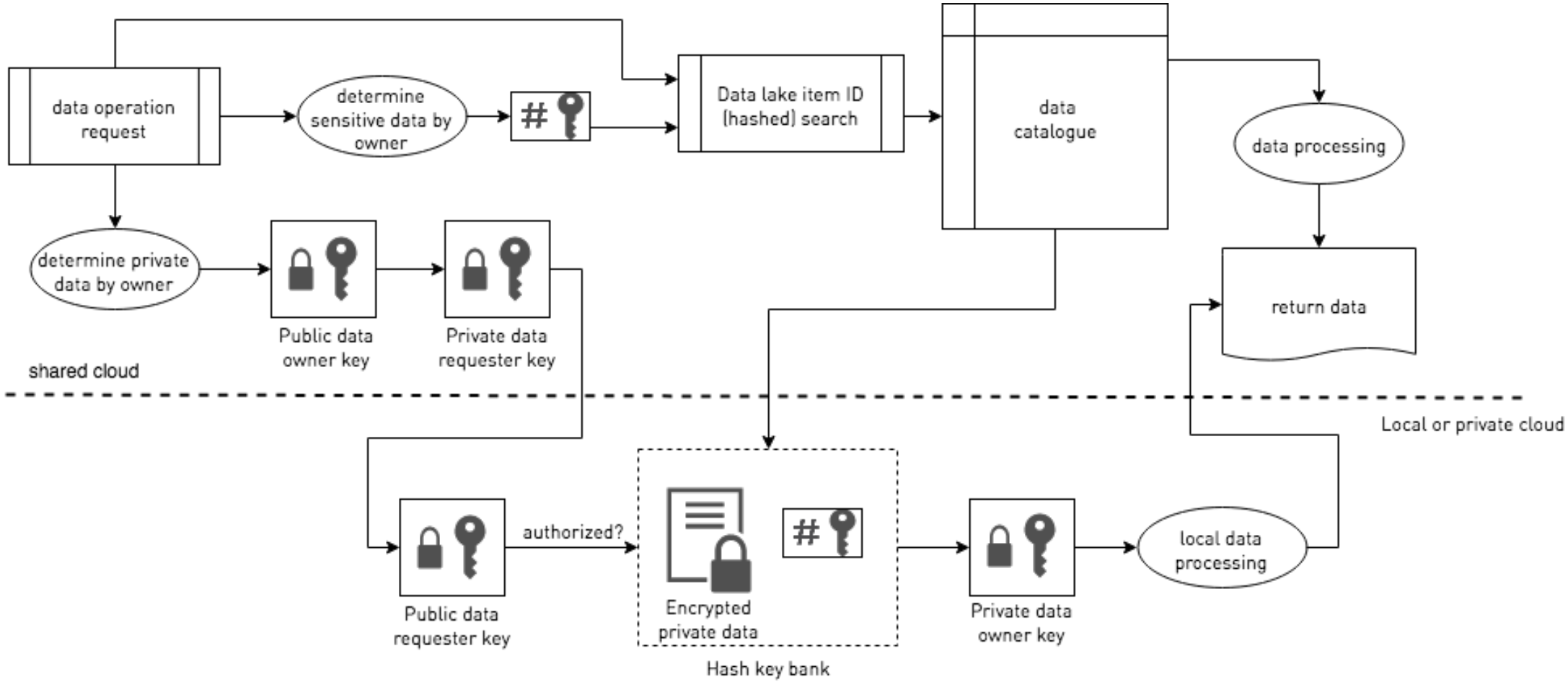
Mastering Big Data for Safety, safely

Data consumption according to level of confidentiality

- Public data, is accessed directly as it rests in the shared environment, e.g. data lake
- Private data, needs to be decrypted and processed locally by the data owner, so it always rests encrypted at local environment
- Sensitive data, is accessed only through its hashed values, as only digested data rest in the shared environment

SafeClouds.eu

Mastering Big Data for Safety, safely



SafeClouds.eu

Mastering Big Data for Safety, safely

Conclusions

- Storing hashed values of sensitive data on the shared environment ensures the original data can not be recovered, unless the 3rd party already have the hashed pair on its own hash key bank
- The hash key bank is stored and encrypted at the local environment making it inaccessible by any 3rd party

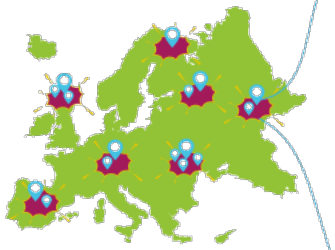
SafeClouds.eu

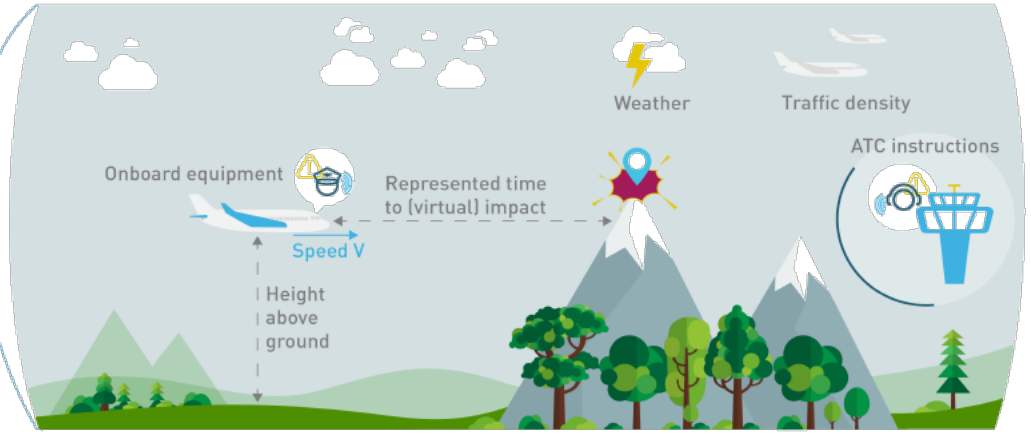
Mastering Big Data for Safety, safely

OLD Practical example:

Use case

CFIT





- Airline operators might not be willing to share date/time of FDM sources
- Most commonly date/time data will simply be erased/overwritten,
  e.g. 29/09/2017 -> ******* or 0000000
- Making it completely impossible identify with any weather reports

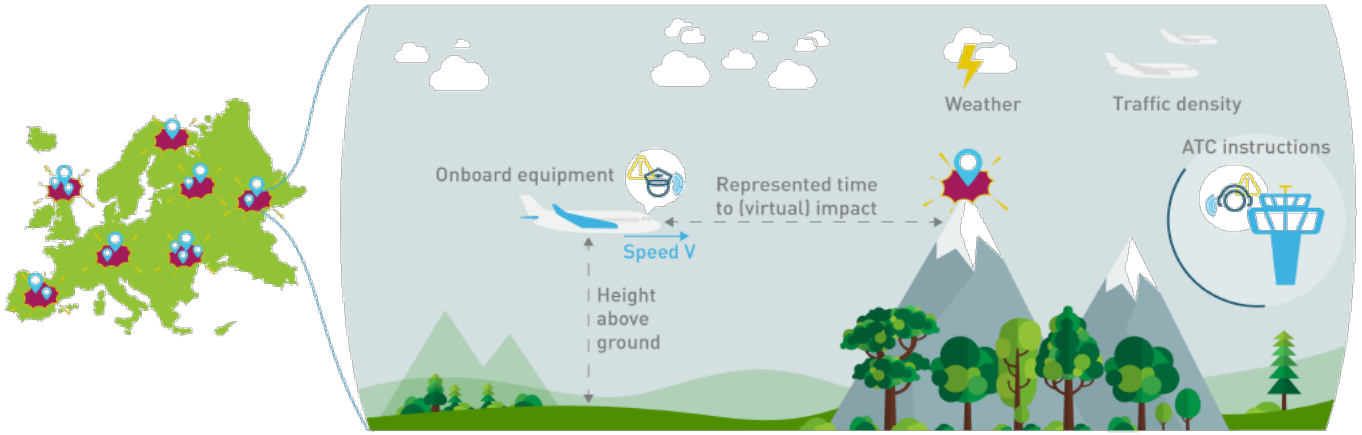
SafeClouds.eu

Mastering Big Data for Safety, safely

NEW Practical example:

Use case

CFIT



- Airline operators might consider date/time of FDM as sensitive data.
- Therefore only hashed data will be shared, e.g. 29/09/2017 -> ef422c0b.
- The reference ef422c0b can now be used to search in the weather reports hash key bank as date/time index.
- Otherwise, reference ef422c0b is useless, date/time can not be recovered.

SafeClouds.eu

Mastering Big Data for Safety, safely

Example, revisited:

- Date/time was not confidential on the weather data set;
  confidentiality is source dependent – shared, hashed sensitive data
- The original date/time information could not be recovered;
  loss of information – local, encrypted hash key bank
- We need a more sophisticated (and elegant) approach;
  still up to personal liking, sorry but I recon it is. ef422c0b > 0000000

SafeClouds.eu

Mastering Big Data for Safety, safely

Thank you!

Questions?

innaxis
innovation for
a complex world

Samuel Cristobal
SC@INNAXIS.ORG
Science and Technology Director