innaxis

innovation for
a complex world

SafeClouds.eu

# Data Analytics practice

[press space bar]

# Agenda

1· Data Analytics

- What's new and what's not?
- Quick wins
- The Data Science practice
- The learning problem

2· SafeClouds

- The project
- The partners
- The work programme
- Scenarios, outcomes

3· Conclusions & challenges

# Data Analytics

Data Analytics

# What's new and what's not

| | |
|---|---|
| 18th century | Bayesian statistics |
| 1920's | Parametric models |
| 1980's | Highly non-linear relationships in real complex datasets |
| 1990's | New analytical techniques, large data sets, high non-linearity |
| 2000's | Machine learning concepts; Storage, Computing, Communications |
| Future in aviation | Focus on processes that provide actionable analytics |

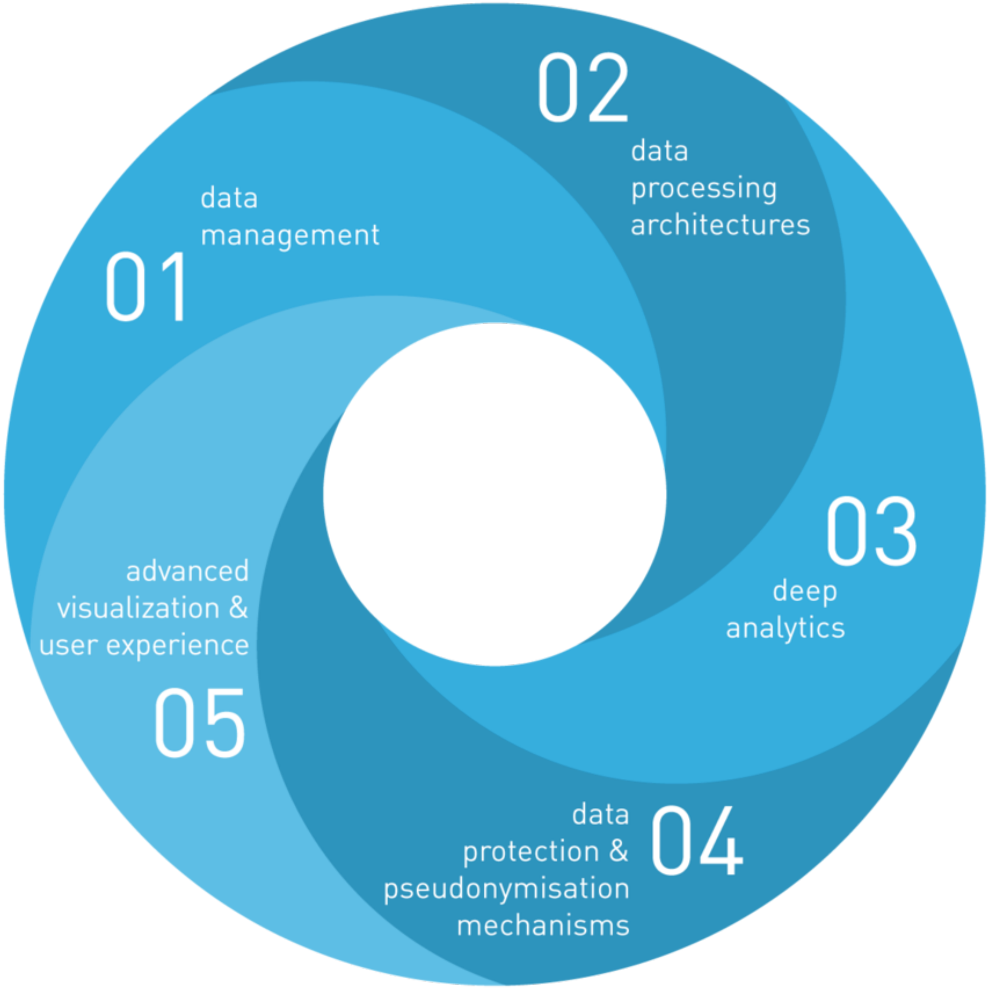
# The Data Science practice *in aviation*

?

**Individualisation trumps universals**

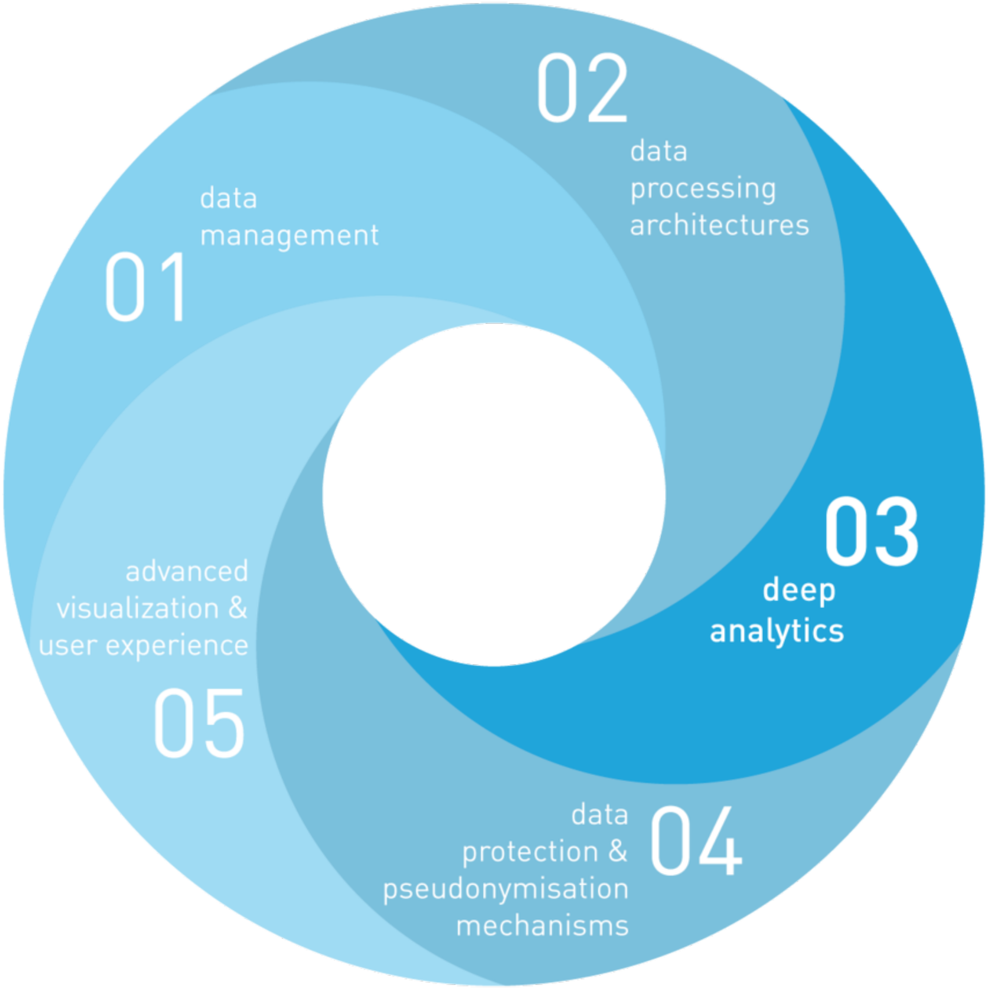
**Intangibles that appear to be completely intractable can be measured and predicted**

# The Data Science practice

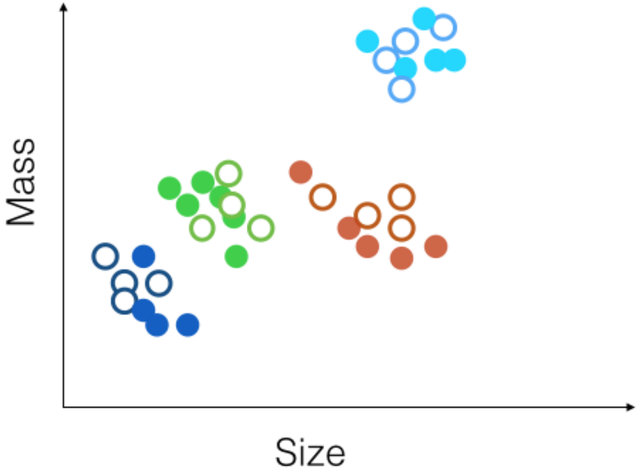


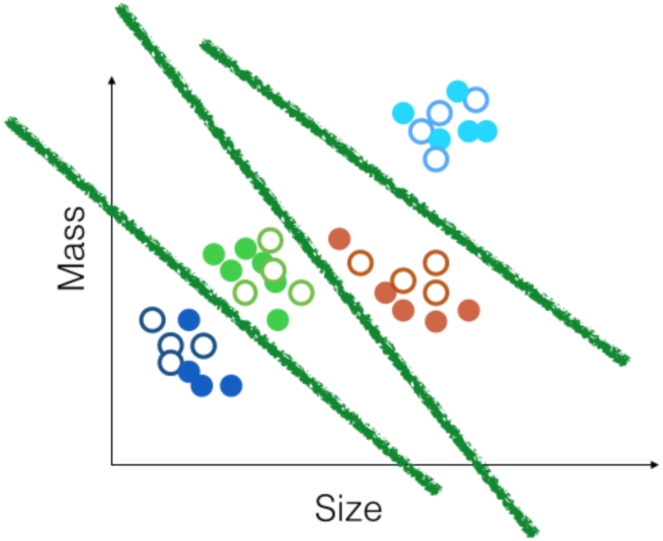
Data Analytics

# What's the learning problem?


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

# What's the learning problem?





# Building models with massive data

The data models and solving the inference problem have challenges:

- Multi-dimensionality, heterogeneity and incompleteness of data, volume of data, velocity,...

**The discipline: Knowledge Discovery on massive data**

- Model selection, including complexity/over-fitting trade-offs
- Model running, including selection of training data, validation and testing
- Model deployment, including stability and trade-offs precision-accuracy-recall

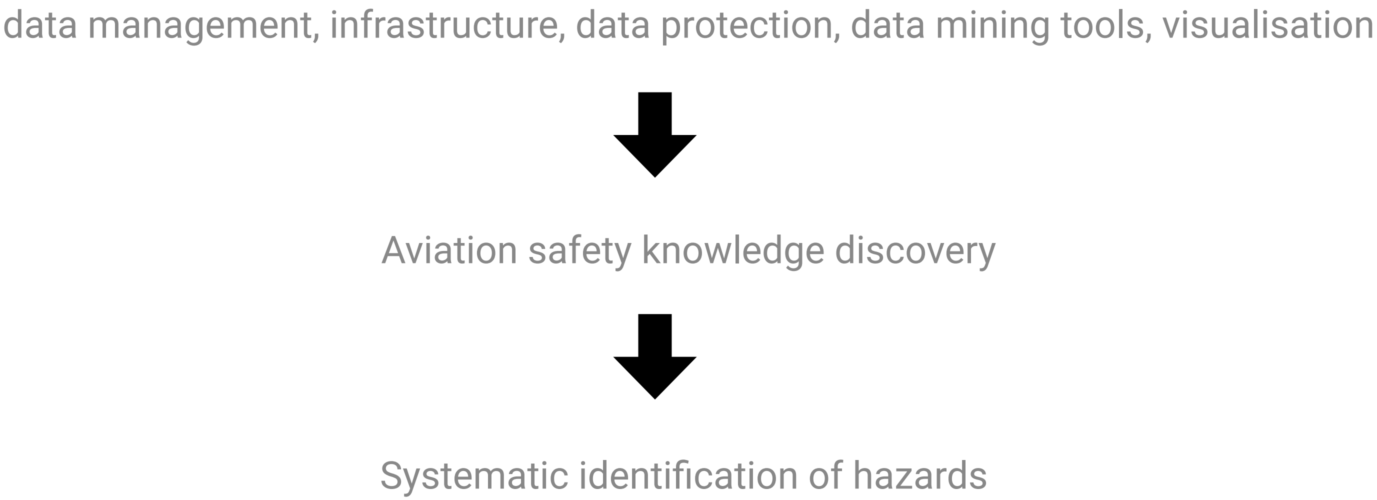
# Building KDD models with massive data

| | Descriptive › | Predictive › | Prescriptive |
|---|---|---|---|
| **Questions** | What happened?<br>What's happening? Why? | What will happen? | What should we do? |
| **Methodologies or technologies** | Clustering<br>Co-occurrence grouping<br>Profiling<br>Similarity matching<br>Link predition | Supervised/unsupervised segmentation<br>Paremetric modelling<br>Methods to avoid overfitting<br>Similarity networks and clusters | Optimization<br>Simulation<br>Decision modelling<br>Causality modelling |
| **Outcomes** | Well-defined case studies opportunities and problems | Accurate projections of future states | Best-possible decisions |

SafeClouds.eu

# SafeClouds

## Applied research - laboratory validation (TRL5)



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

Aviation safety knowledge discovery

Systematic identification of hazards

innaxis

TADOREA
KNOWLEDGE DISCOVERY FOR AVIATION

IBERIA

norwegian

vueling AIRLINES

PEGASUS
AIRLINES

AirEuropa

LFV
AIR NAVIGATION SERVICES OF SWEDEN

EUROCONTROL

CRIDA
I+D+i - ATM

TUM
Technische Universität München

LiU LINKÖPING
UNIVERSITY

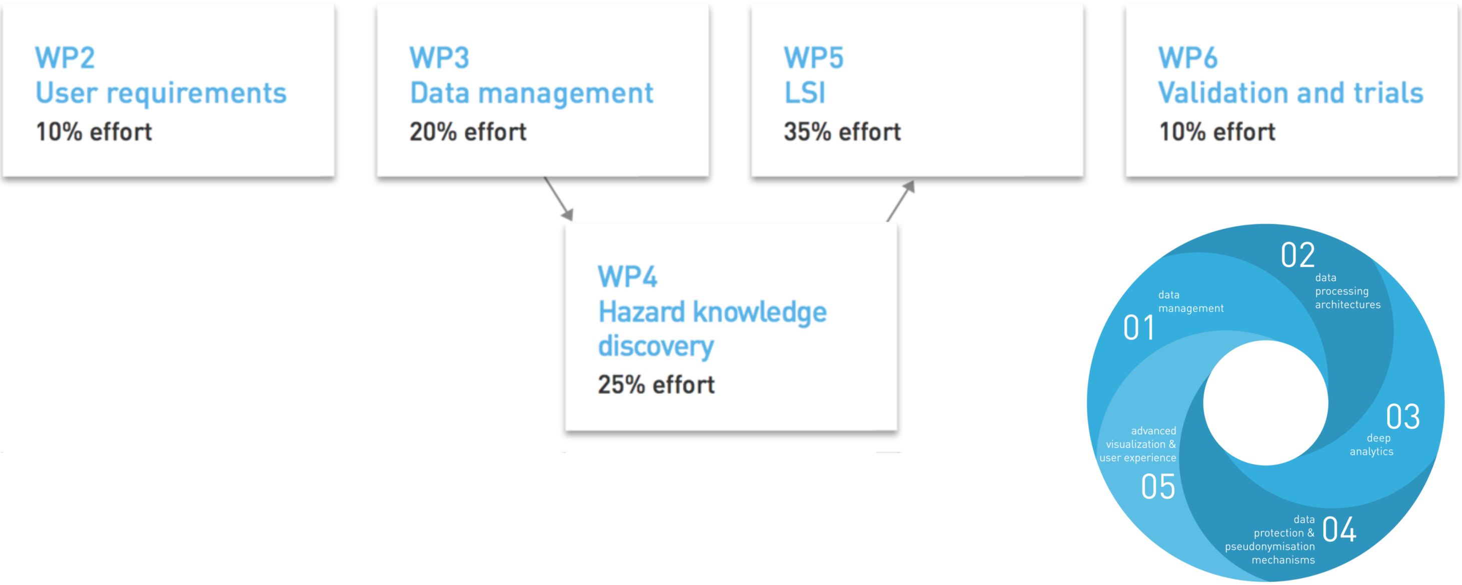
Fraunhofer

TU Delft

AESA

EASA
European Aviation Safety Agency

# SafeClouds research project



WP2
User requirements
10% effort

WP3
Data management
20% effort

WP5
LSI
35% effort

WP6
Validation and trials
10% effort

WP4
Hazard knowledge discovery
25% effort

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

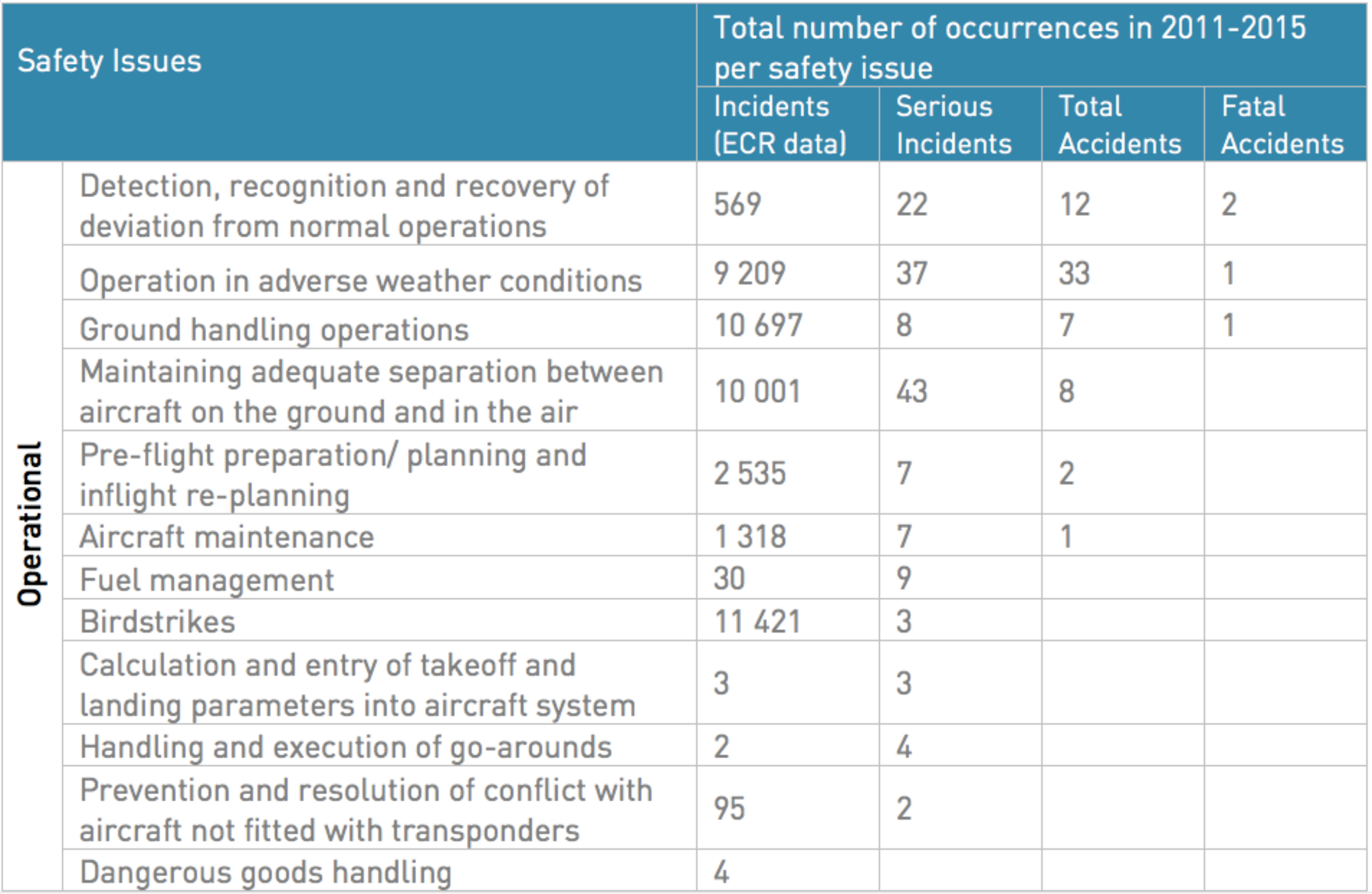
# SafeClouds research project

**Some scenarios of interest:**

- Real time approach congestion monitoring
- Proper separation with terrain
- Level busts
- Runway performance
- Runway excursions
- Unstable approaches

# SafeClouds research project

## EASA



| Safety Issues | | Total number of occurrences in 2011-2015 per safety issue | | | |
|---|---|---|---|---|---|
| | | Incidents (ECR data) | Serious Incidents | Total Accidents | Fatal Accidents |
| **Operational** | Detection, recognition and recovery of deviation from normal operations | 569 | 22 | 12 | 2 |
| | Operation in adverse weather conditions | 9 209 | 37 | 33 | 1 |
| | Ground handling operations | 10 697 | 8 | 7 | 1 |
| | Maintaining adequate separation between aircraft on the ground and in the air | 10 001 | 43 | 8 | |
| | Pre-flight preparation/ planning and inflight re-planning | 2 535 | 7 | 2 | |
| | Aircraft maintenance | 1 318 | 7 | 1 | |
| | Fuel management | 30 | 9 | | |
| | Birdstrikes | 11 421 | 3 | | |
| | Calculation and entry of takeoff and landing parameters into aircraft system | 3 | 3 | | |
| | Handling and execution of go-arounds | 2 | 4 | | |
| | Prevention and resolution of conflict with aircraft not fitted with transponders | 95 | 2 | | |
| | Dangerous goods handling | 4 | | | |

# SafeClouds outcomes

Questions

Case studies and operational questions

Example:
When is a level bust occurring

+

Inputs

Data

Use Cases

Safeclouds Platform

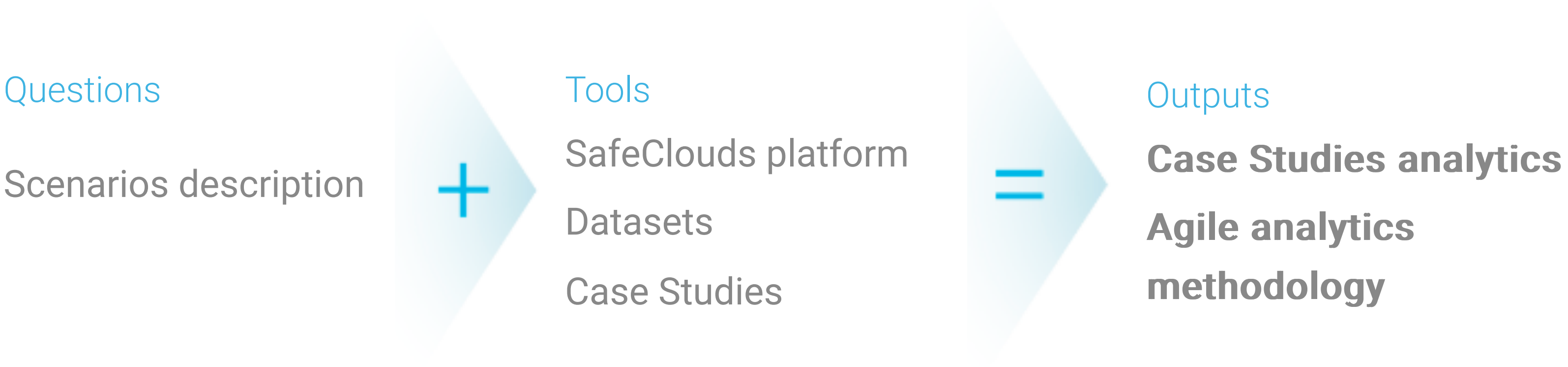
=

Outputs

Agility management methodology

SafeClouds

Analytics Strategy



Questions

Scenarios description

+

Tools

SafeClouds platform

Datasets

Case Studies

=

Outputs

**Case Studies analytics**

**Agile analytics methodology**

# SafeClouds research project

**Next steps**

- Consortium Agreement sign. inc. data protection & sharing - Sept '16
- Grant Agreement signature - Sept '16
- Project starts - early Oct '16
- Consortium Coordinator - Paula López-Catalá, plc@innaxis.org

# Conclusions & challenges

|  | Enable the data | Build/govern the platform | Engage the business |
|---|---|---|---|
| Conclusions | ● Data ingest<br>● Cleanse<br>● Fuse | ● Build Models<br>● Build infrastructure<br>● Secure | ● Discover<br>● Monitor<br>● Deploy |
| Challenges | ● Data sources<br>● Complexity<br>● Costs | ● Skill gap in ML-aviation<br>● Reliance on IT<br>● Trust / Privacy | ● Agile methodologies<br>● ROI metrics<br>● Change processes |

# Some thoughts on challenges

- Analytics Center of Excellence is not an IT organisation
- Data Science agile management is a must
- Reusable data & logic for governance and consistency
- Great tools for collaboration, visual tools.

# Closing thoughts

Difficult to see "quick wins" or "low-hanging fruits"

Data Science is a craft - there is no Excel+++

Your model is not what your data scientists design, it's what your engineers implement - translation business to technical is key

# Thank you!

David Pérez - dp@innaxis.org

www.SafeClouds.eu

this presentation - slides.innaxis.org/2016.09.08.SafeClouds

## References

*Annual Safety Review, EASA, 2016*

*Data, information and analytics as services, Delen & Demirkan, 2012*

*Data Science for business, Provost & Fawcett, 2013*

*European Big Data Value Strategic Research Agenda, 2015*

*Frontiers in Massive Data Analytics, National Academy of Sciences, 2013*

*Network analysis reveals patterns behind air safety events, 2014*

*The irrational effectiveness of mathematics in natural sciences, Wigner, 1960*

*The irrational effectiveness of data, Norwig, 2009; youtube.com/watch?v=yvDCzhbjYWs*

*SafeClouds documentation - to be published from October 2016 in www.SafeClouds.eu*

*Synchronisation likelihood in aircraft trajectories, Zanin, 2013*

# BackUp

# Hazards

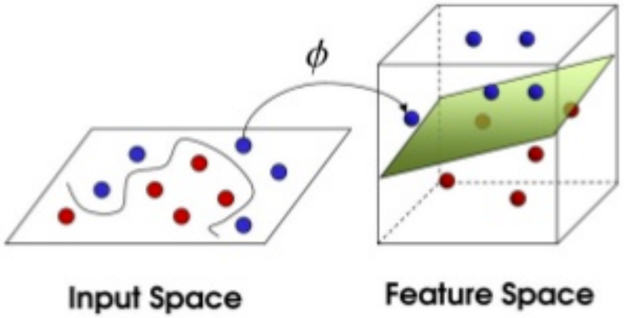
A hazard can be considered as a dormant potential for harm

which is present in one form or another within the aviation system or its environment.

This potential for harm may be in the form of

- a natural hazard such as terrain, or
- a technical hazard such as wrong runway markings

# Building KDD models with massive data



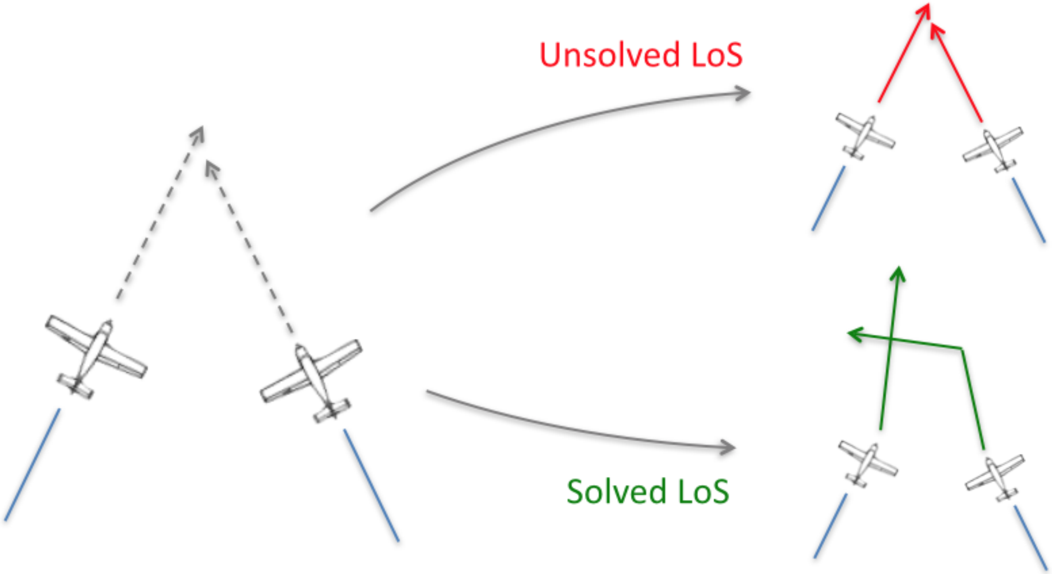
# The SafeClouds initiative

The SafeClouds research initiative is promoted by a complete spectrum of Aviation and ICT European stakeholders to develop  big data, data protection and data mining tools for the improvement of aviation safety.

SafeClouds presents a project to develop aviation safety knowledge discovery techniques from a large set of distributed datasets.

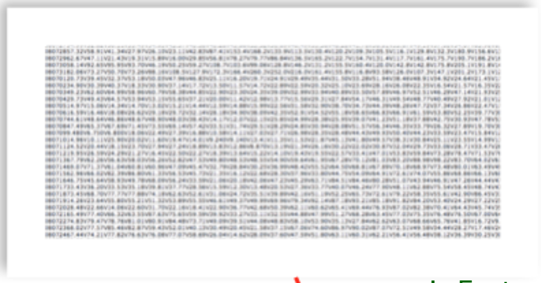
Novel systematic identification of hazards and handling of data and processes tailored to the requirements of aviation that are efficient, effective and acceptable by all the relevant parties in the aviation value-chain.

# Addressing the learning problem



## Addressing the learning problem

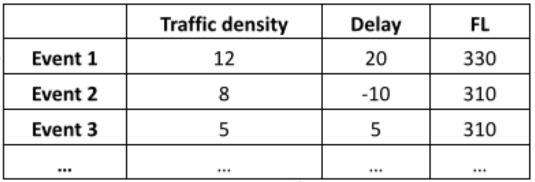
# Safety KDD research model



I - Feature extraction

Mostly data management

Domain knowledge

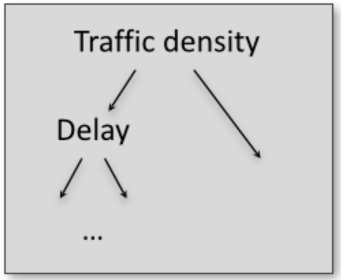


| | Traffic density | Delay | FL |
|---|---|---|---|
| **Event 1** | 12 | 20 | 330 |
| **Event 2** | 8 | -10 | 310 |
| **Event 3** | 5 | 5 | 310 |
| **...** | ... | ... | ... |

II - Feature combination

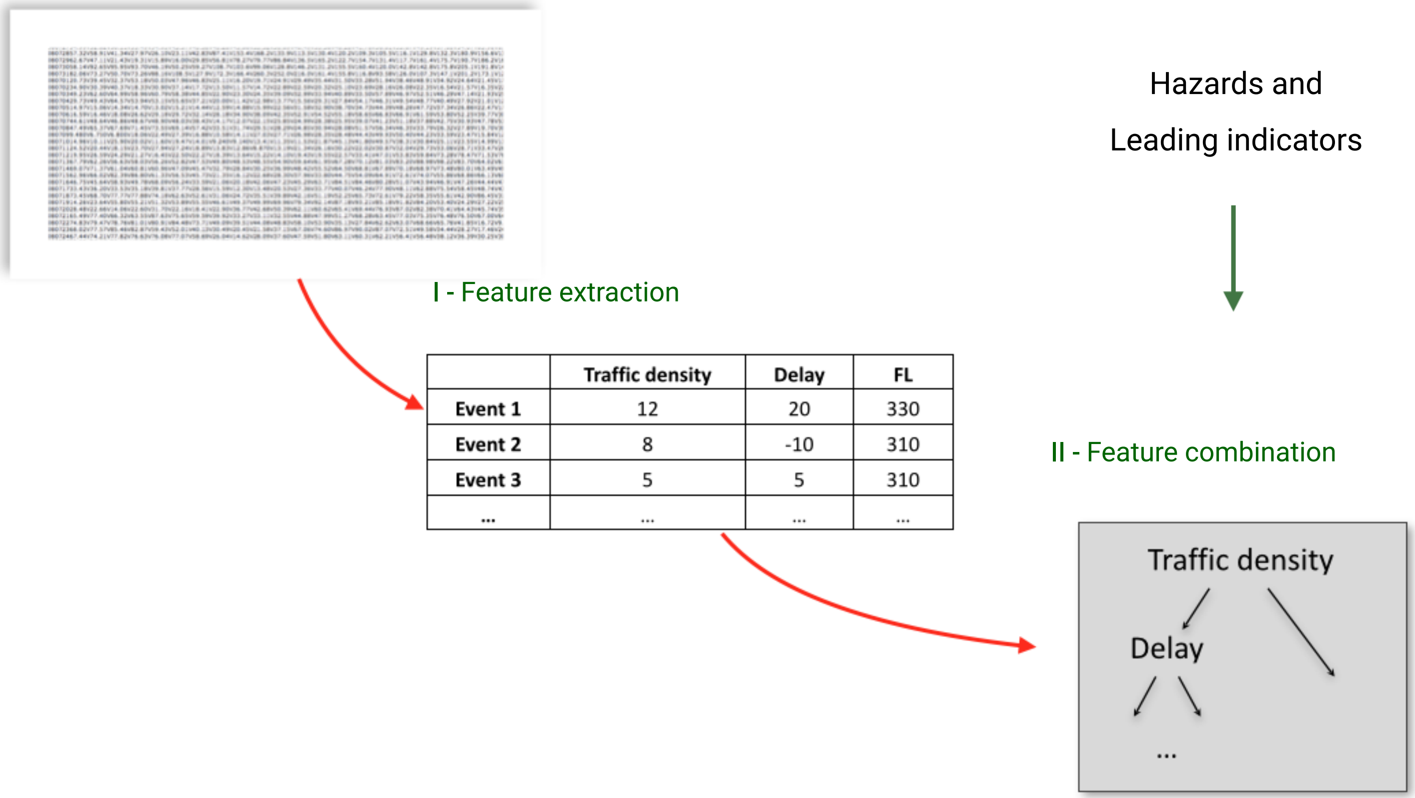
Mostly math

Domain knowledge



## Addressing the learning problem

# Safety KDD research model



Hazards and

Leading indicators

I - Feature extraction

| | Traffic density | Delay | FL |
|---|---|---|---|
| **Event 1** | 12 | 20 | 330 |
| **Event 2** | 8 | -10 | 310 |
| **Event 3** | 5 | 5 | 310 |
| **...** | ... | ... | ... |

II - Feature combination

Data Analytics

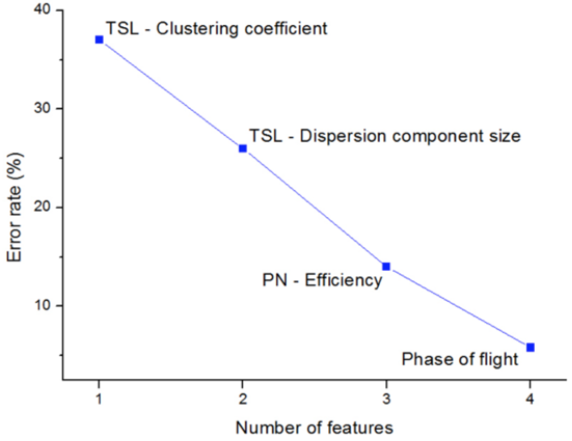
# Building KDD models with massive data

## KDD study on prediction of separation

Eurocontrol traffic data - 10 months ECAC traffic, 2min resolution

Low frequency of aviation safety events

Medium term data-driven prediction on LoS events?



1 Classical features describing the status of airspace

2 Complex network features

3 Historical trajectory likelihood-based features

# Concepts

Recall literally is **how many of the *true* positives were *recalled***, i.e. how many of the correct hits were also found.

Precision is **how many of the *returned* hits were *true* positive** i.e. how many of the found were correct hits.

Accuracy is how many of the times the algorithms were correct, i.e. total true positives plus true negatives

recall = TP / (TP + FN)
precision = TP / (TP + FP)
accuracy = (TP+TN)/ ALL