STATISTIEK
VLAANDEREN

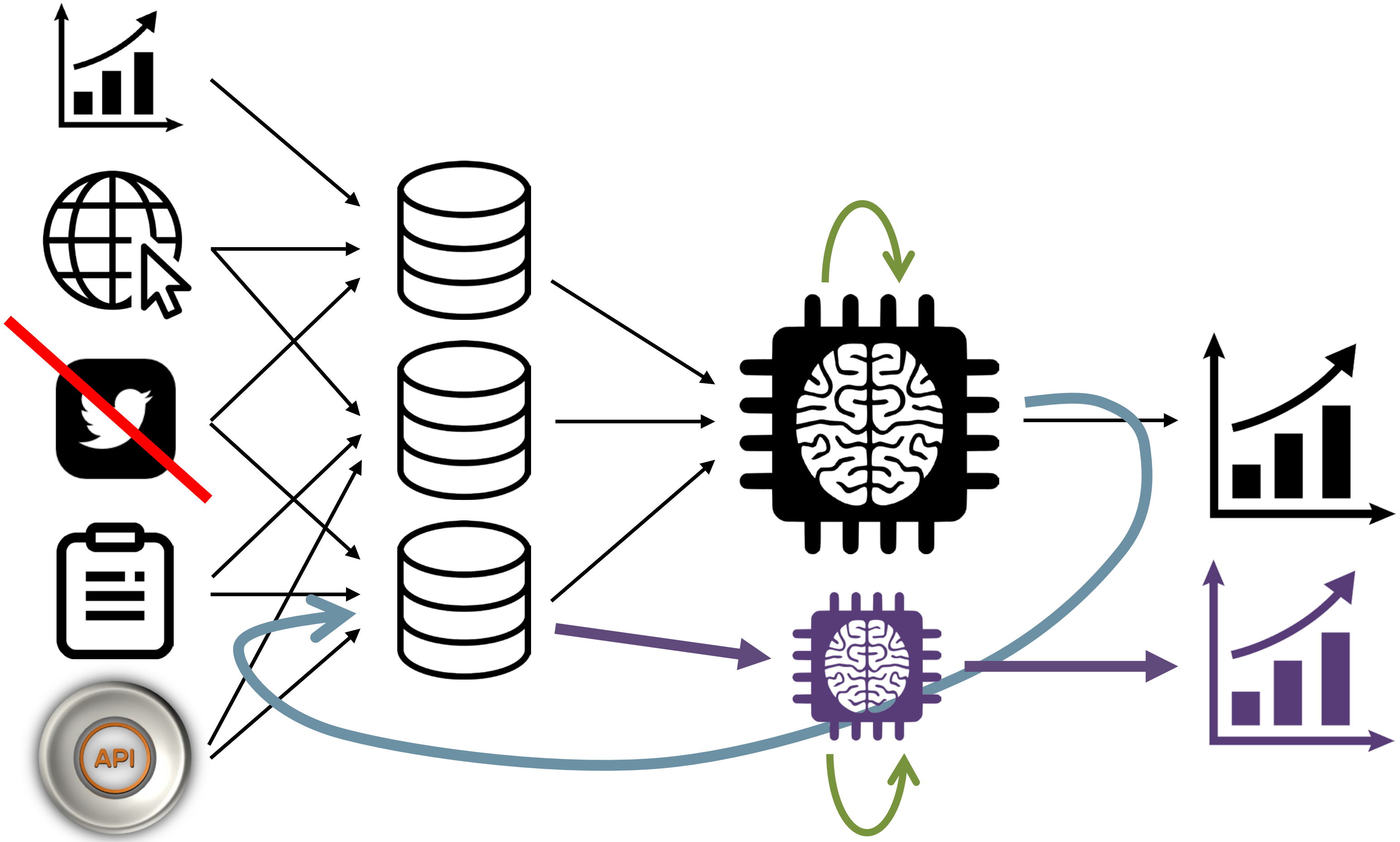
Changing Data Sources
in the Age of Data Science for Official Statistics

Cedric De Boom & Michael Reusens
Statistics Flanders – Belgium

UNECE Machine Learning for Official Statistics
June 6, 2023

Vlaamse
overheid

From data sourceto statistic



STATISTIEK
VLAANDEREN

Confidential– do not spread

Vlaamse
overheid

Benefits of external data sources

Broad-spectrum	Covers a wide variety of topics
Diversity	A large variety of sources to cover different perspectives
Availability	Lots of data is freely and easily accessible
Size	Some datasets can be enormous, sometimes even complete
Structure	Not only tabular data, but also images, video, text, audio, etc.
Timeliness	(Near) up-to-date and real-time information
Frequency	Raw data on various, even very fine-grained time scales
Granularity	Raw data on various, even fine-grained levels of detail
Coverage	Various locations and regions can be filtered and covered

… butchallengesneed to be overcome

Data quality	Errors, biases, missing values…
Data interpretation	Context, meaning, business rules…
Data integration	Overcoming diverse structures and formats
Selection bias	Ensure representativeness
Operationalization bias	Implicit, hidden, and/or production-specific design choices
Computational resources	Processing and analyzing large amounts of data
Privacy and security	Anonymization, pseudonymization, access management…
Data ethics	Data collection and use should adhere to ethical principles
Fairness and justness	Neutral, non-discriminatory
Cost	Resources, workforce, data purchases…

Lackof control is an insidious risk

…but with great amounts of external data comes
great powerlessness!

Risk mitigation strategies
should be front and center
in your data science agenda and practices!

“Data is the new oil” – C.Humby, 2006

Powerful value!

But also:

- Vulnerability!
- Powerlessness!
- Dependency!
- Lack of control!

STATISTIEK
VLAANDEREN

Typesand causes of changing data sources

Vlaamse
overheid

Overview

Data types and schemas

Sharing and collection technology

Concept drift

Frequency and interruptions

Ownership and discontinuation

Legal properties

Ethics and public perception

Data types &schemas

= changes in data formats and structure

Why?

Accomodate future changes

Technical debt

Improve storage

Increase retrieval effiency

Business rules

…

Consequences

Catch errors?

Undetected?

Mitigation

Testing, testing, testing!

Data checks & monitoring

Statistical analyses

Sharing& collection technology

= storage, cloud, APIs, scraping, external tools, format …

APIs

Endpoint updates

Security patches

Business strategy

Pricing

Concept drift

= data distribution changes between train and test time

Why?

Business logic

Variable meaning

Coverage / frequency

Derived data fields!

i.e. as result of ML model

Consequences

Retraining & reevaluation

Mitigation

Statistical tests

Monitoring

Frequency& interruptions

= collection or update rate modifications

Why?

Deliberate vs random

Technological challenges

Downtime, failures…

Consequences

Can lead to concept drift!

Mitigation

Statistical tests

Monitoring

Ownership & discontinuation

= changes in offering or downright shutdown

Consequences

Legal issues, pricing…

Can trigger any other consequence

Mitigation

Redundancy and diversification

Legal contracts / SLAs

Legalproperties

= legal changes regarding data collection, storage and use

Why?

Privacy laws

Contractual obligations

Consequences

Renegotiation

Stop the statistical offering

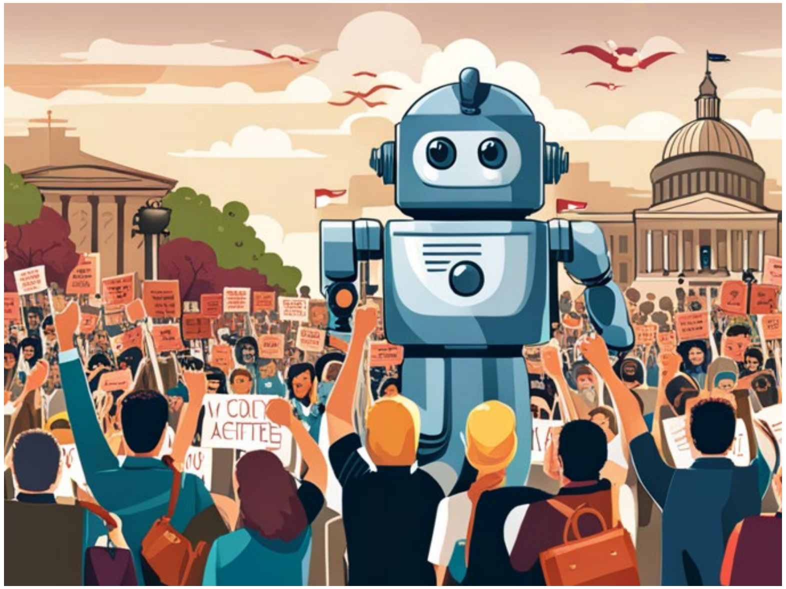
Airtight data management

Mitigation

Redundancy and diversification

Legal contracts / SLAs

Ethics & public perception



Why?

Controversial

Neutrality / bias

Intrusive

Transparency

Accountability

Integrity

Consequences

More scrutiny

More oversight

Patching models

Public trust!

Public policy

In summary

Data types and schemas

Sharing and collection technology

Concept drift

Frequency and interruptions

Ownership and discontinuation

Legal properties

Ethics and public perception

STATISTIEK
VLAANDEREN

Consequences
of changing data sources

Vlaamse
overheid

Brief overview of consequences

Concept Drift	Especially relevant when dealing with long-term trends
Model staleness	The model no longer picks up current trends and patterns
Bias and neutrality	“Garbage in, garbage out” vs neutrality and objectivity
Availability	May impact accurate and timely statistics
Integration	Beware the domino effect!
Extra labor	Take risks into account and allocate resources and time budgets
Breaking changes	Depending statistics will inevitably change: be transparent!
Quality metrics	Timeliness, validity, accuracy, completeness, consistency…

STATISTIEK
VLAANDEREN

Mitigating
changing data sources

Vlaamse
overheid

Mitigation?

Not easy!

Changes are diverse

Consequences are diverse

Required mitigation efforts are time- and resource-consuming

No definite answers…

Highly use-case- and context-dependent

Risk analysis

Identify all potential risks associated with the external data

Use the list in this presentation / paper as a guideline!

Describe technical
and non-technical aspects

Face the hard truth!

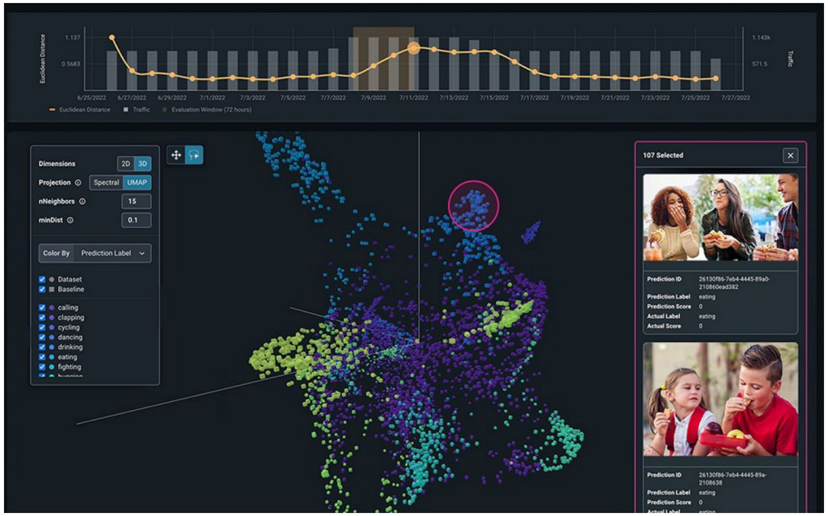
Monitoring

Monitor everything!

Record inputs and outputs

Perform statistical tests

Track variables and quantities



Dimensions: 2D [✓ 3D]

Projection: Spectral [✓ UMAP]

nNeighbors: 15

minDist: 0.1

Color By: Prediction Label

☑ ● Dataset
☑ ■ Baseline

☑ ● calling
☑ ● clapping
☑ ● cycling
☑ ● dancing
☑ ● drinking
☑ ● eating
☑ ● fighting

107 Selected

Prediction ID	26130f86-7eb4-4445-89a0-210860ead382
Prediction Label	eating
Prediction Score	0
Actual Label	eating
Actual Score	0

Prediction ID	26130f86-7eb4-4445-89a-2108638
Prediction Label	eating
Prediction Score	0

How?

Reference data sets

Cluster metrics

Visualization and dimensionality reduction

Check predictions against existing domain knowledge

Devise supervised proxy-tasks

Diversification

Use multiple, redundant data sources if possible

Discrepancies?

Data normalization

Computational overhead

…but is very challenging and not straightforward

Technicalrobustness

Ensure consistency in the statistical offering

Automated data pipelines

Build resilience against errors, outliers, outages…

Data validation is a part of these pipelines

Thorough unit and integration testing

Failover and deduplication

Security measures

Requires a hefty engineering team, along with rigorous best practices!

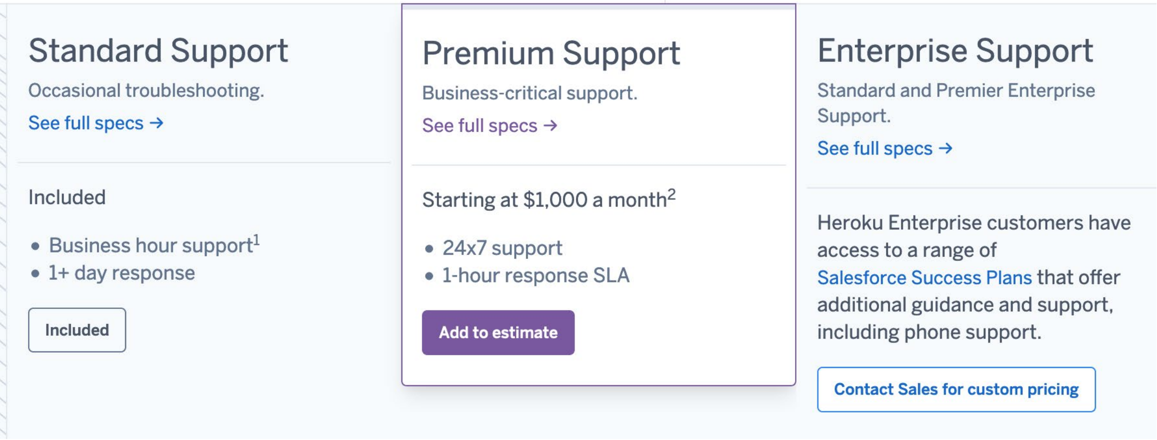
Legalrobustness

Eliminate unexpected changes, outages and discontinuation

Negotiate tight contracts and SLAs

Specify the legal consequences of non-compliance!

But comes at a significant cost!



Standard Support

Occasional troubleshooting.

See full specs →

Included

- Business hour support[1]
- 1+ day response

Premium Support

Business-critical support.

See full specs →

Starting at $1,000 a month[2]

- 24x7 support
- 1-hour response SLA

Enterprise Support

Standard and Premier Enterprise Support.

See full specs →

Heroku Enterprise customers have access to a range of Salesforce Success Plans that offer additional guidance and support, including phone support.

Conclusions

Risks and consequences

The list is long!

Highly use-case- and context-dependent

This is a story of trade-offs!

But: don't tread lightly on these matters, especially in the context of official statistics!

Mitigation strategies

No free lunch

Requires significant resources and a talented workforce

Use this paper and presentation as a guideline / checklist

Thank you!

Contact us!

https://www.vlaanderen.be/statistiek-vlaanderen-data-science-hub

SCAN ME

Michael Reusens
Data sciencecoordinator
michael.reusens@vlaanderen.be

Cedric De Boom
Senior datascientist
cedric.deboom@vlaanderen.be

STATISTIEK
VLAANDEREN

Vlaamse
overheid