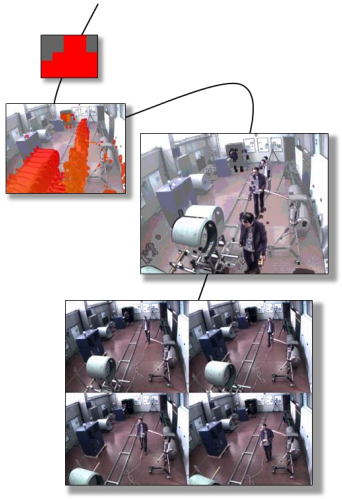
Pattern recognition by humans and machines over large data sets



C. Versino

European Commission

Joint Research Centre (JRC)

Institute for Transuranium Elements (ITU)

Nuclear Security Unit

Ispra, Italy

Symposium on International Safeguards: Linking Strategy, Implementation and People

Vienna, 20-24 October 2014

Outline
'Data retrieval and analysis over large data sets'

Issues	*Will present main **issues** in data retrieval/analysis,*	• *Invisible Big Data* • *Data access* • *Precision vs Accuracy of information*
Technology	*and highlight ways of using **information technology, based on data visualisation,** to address these issues.*	
Examples	*Will present **example visualisations** related to nuclear safeguards.*	*CN 220-224 Tools for video reviews* *CN 220-293 Tools for trade analysis…*

Issue

Invisible Big Data

Large data sets are buried in databases and repositories.

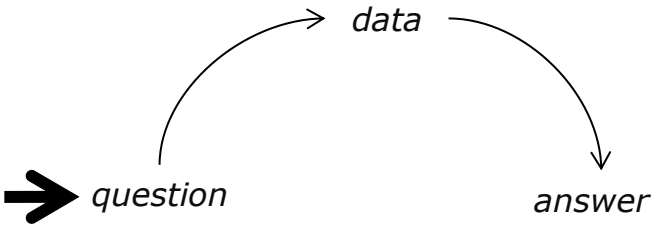
We do not see data like we see the world around us.

There is a narrow communication channel between the data and the user (even if you are feeling lucky).

Issue

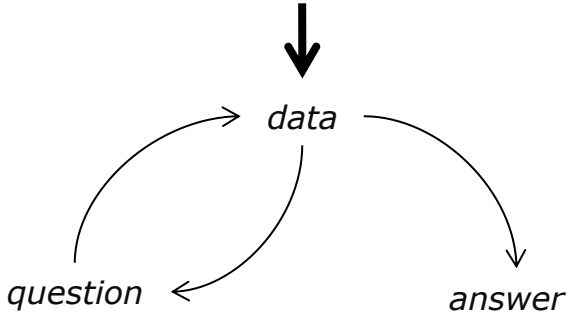
Data access

Traditional



In many cases data access is mediated by queries.
One needs to formulate useful queries before seeing any data.
Only slices of filtered data are returned.
Little data integrity.

Data visualization

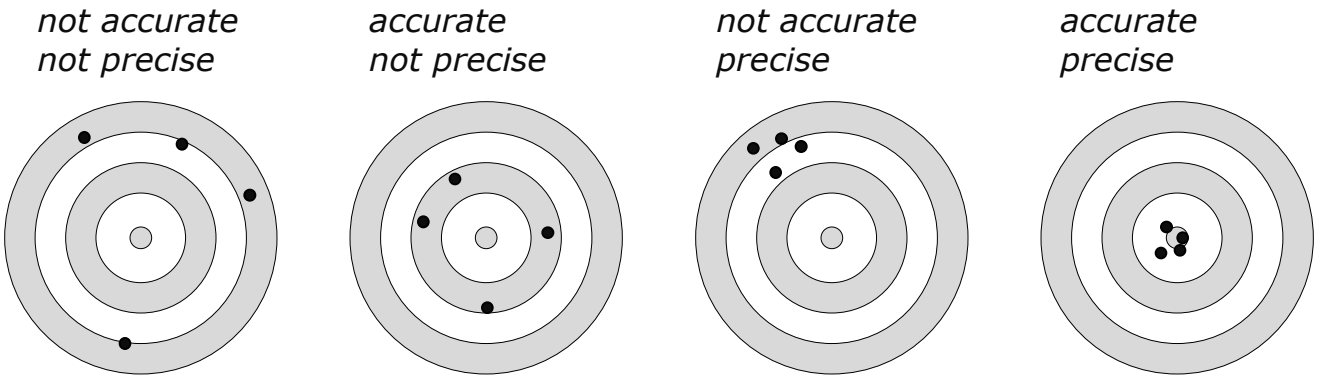


By contrast a data visualisation approach would feature the data first. Seeing the data distribution may trigger questions that one would not have imagined otherwise.

Issue

Precision vs. Accuracy of information

– related to Correctness vs. Completeness –



"Even if the amount of knowledge in the world is increasing, the gap between what we know and what we think we know may be widening. This syndrome is often associated with very precise-seeming predictions that are not at all accurate. (…)
This is like claiming you are a good shot because your bullets always end up in about the same place — even though they are nowhere near the target."

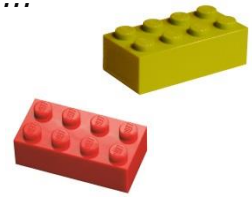
Nate Silver
The Signal and the Noise

The data visualisation process

Effort

5% ~~20~~% of time — ***Data gathering***
Queries on third parties DBs, sensor data, own generated data, ...

5% ~~50~~% of time — ***Data preparation for analysis (analysis with IT)***
Data de-structuring to raw format, + meta-data



90% ~~20~~% of time — ***Data visualisation***

Encode abstract data in graphical form for analysis and communication.

Enables human visual recognition.
Works pre-attentively.
Parallel (high bandwidth).
→ **Fast**

Explore
Understand
Question ...

Make a point
Findings
Report ...

Analysis tool

Analytical interactions: *adding / removing dimensions, sorting, filtering, highlighting, aggregating / disaggregating, drilling, grouping, zooming/panning, re-visualising, re-expressing, re-scaling ...*

Raw data – Data integrity – Data sushi

Data sushi:

'A visualisation which is beautiful on the outside and has raw data on the inside'

Jock Mackinlay
Jock's Dream of Data Sushi

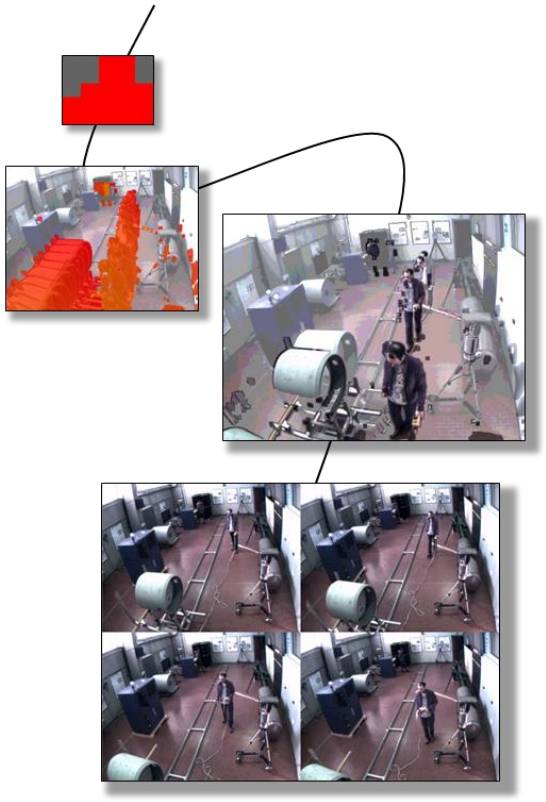
Why using raw data is important?

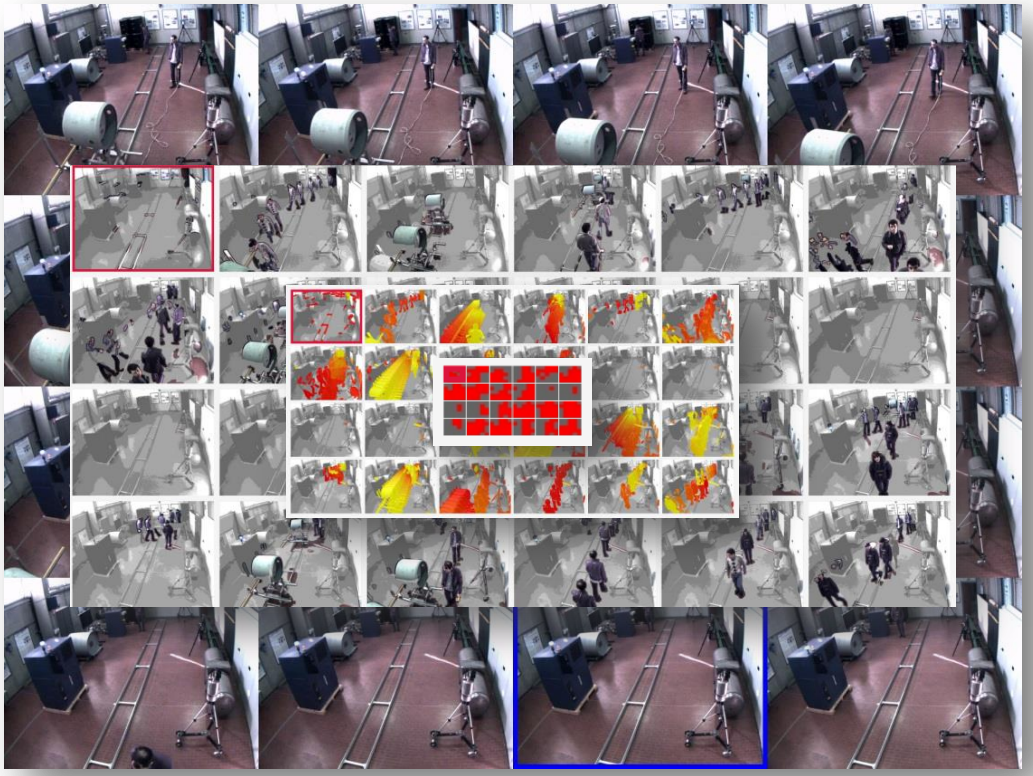
- *Gives the analyst the ability to create **overviews** of the data (data integrity, accuracy, completeness) and **detailed views** as required (precision, correctness).*
- *Result data views are generated on demand as visual cross-tabs of data dimensions of interest to the analyst (i.e., not decided by a data provider as pre-defined views or paths to get to the data).*
- *'Validates the author' of data views (peers can explore the same data set and confirm or find different/other/more results).*
- *Facilitates blending of other data sources (adding more dimensions, relate with independent sources).*
- *…*

Data visualisation – Overview first

S. Blunsden, C. Versino
VideoZoom storyboard

Data visualisation – Details on demand

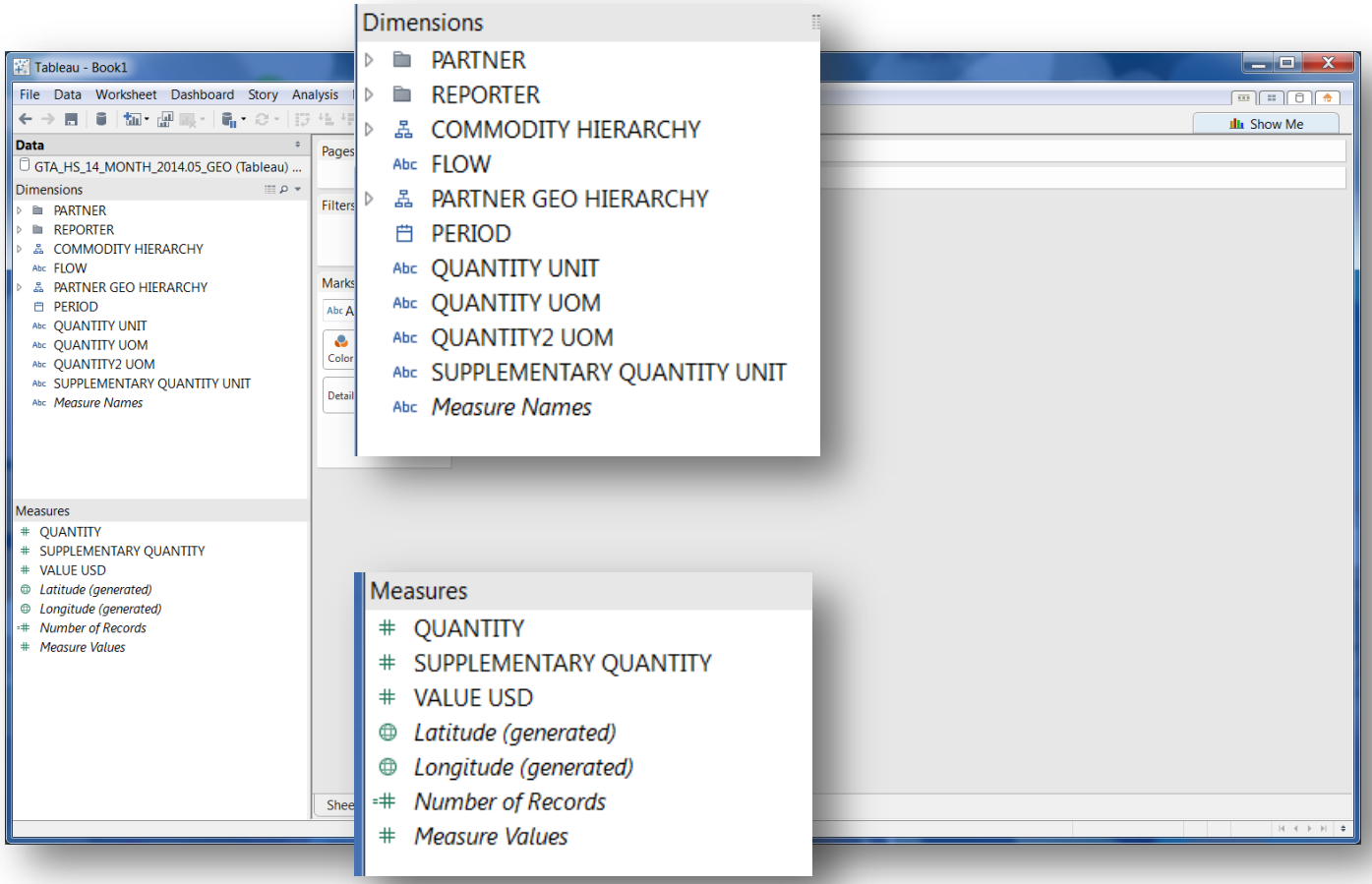




S. Blunsden, C. Versino
VideoZoom zooming interface

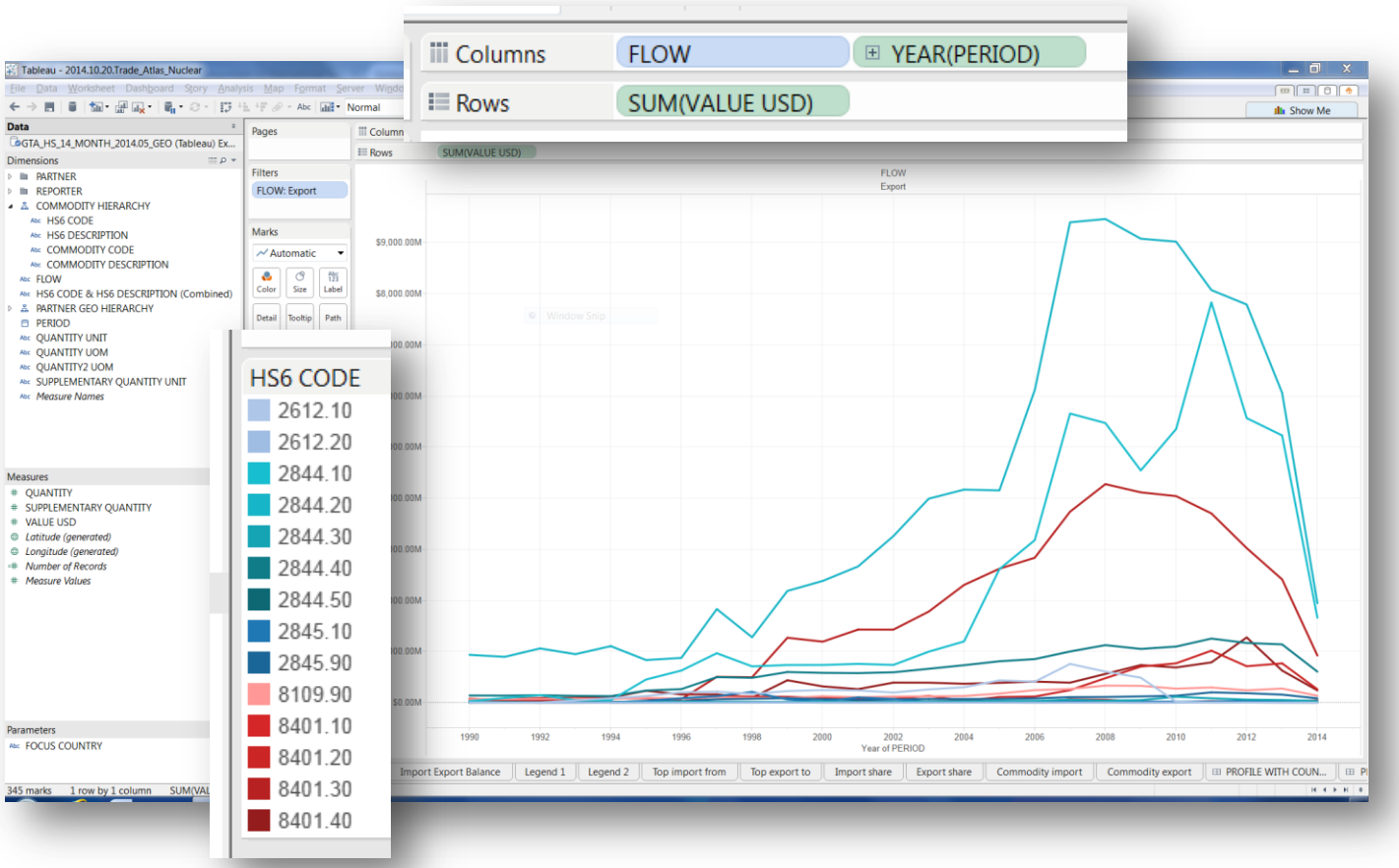
Example

Data visualisation – Raw data



Example

Data visualisation – Data composition



Country X

Nuclear Trade Profile

2012-2014

Nr of data records: 8,625

Import Export Balance

Import 71%

Export 29%

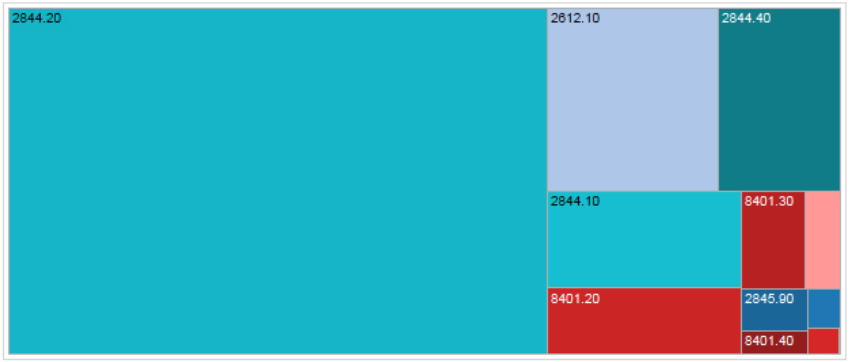
2612.10, Uranium ores and concentrates
2612.20, Thorium ores and concentrates
2844.10, Natural uranium and its compounds; ..
2844.20, Uranium enriched in U 235 and its co..
2844.30, Uranium depleted in U 235 and its co..
2844.40, Radioactive elements, isotopes and c..
2844.50, Spent "irradiated" fuel elements "cartr..
2845.10, Heavy water "deuterium oxide" [Eurat..
2845.90, Non-radioactive isotopes; inorganic o..
8109.90, Articles of zirconium, n.e.s.
8401.10, Nuclear reactors [Euratom]
8401.20, Machinery and apparatus for isotopic ..
8401.30, Fuel elements "cartridges", non-irradi..
8401.40, Parts of nuclear reactors, n.e.s. [Eura..

Import share

Top import from

United Kingdom
Russia
Netherlands
Germany
Canada
Australia
France
Kazakhstan
Namibia
China
Sweden
Japan

Commodity import

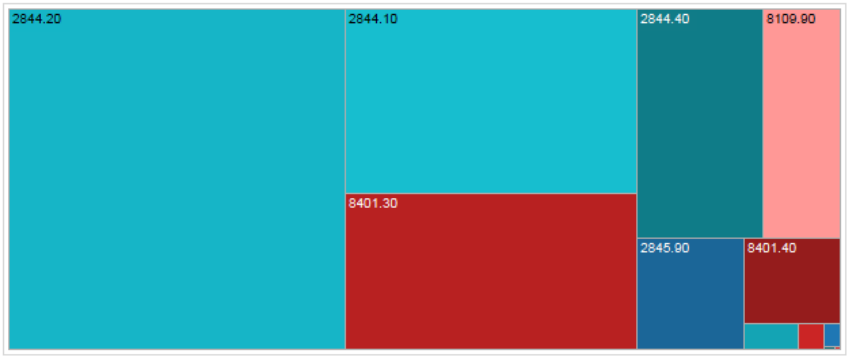


Export share

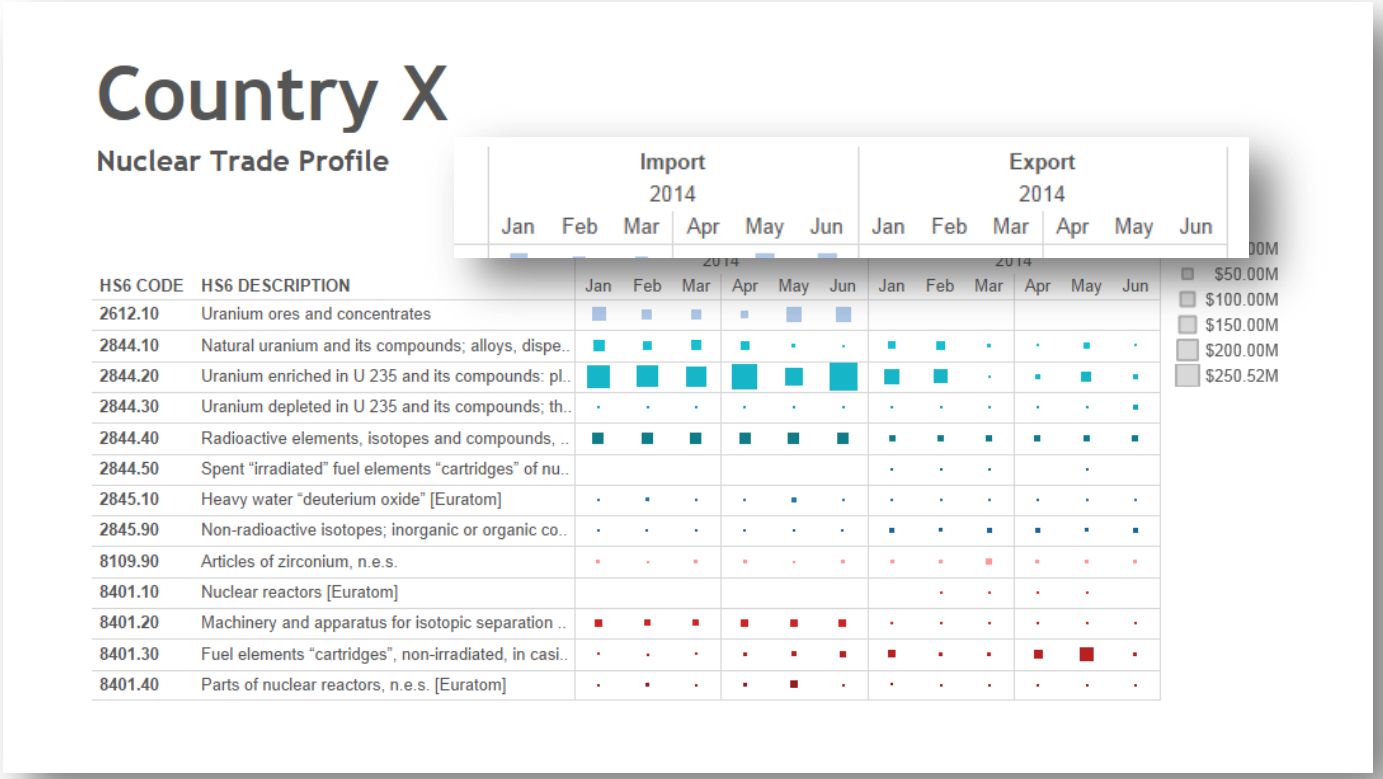
Top export to

Japan
Taiwan
Canada
Korea, South
United Kingdom
China
Sweden
Russia
Mexico
Germany
Spain
France

Commodity export



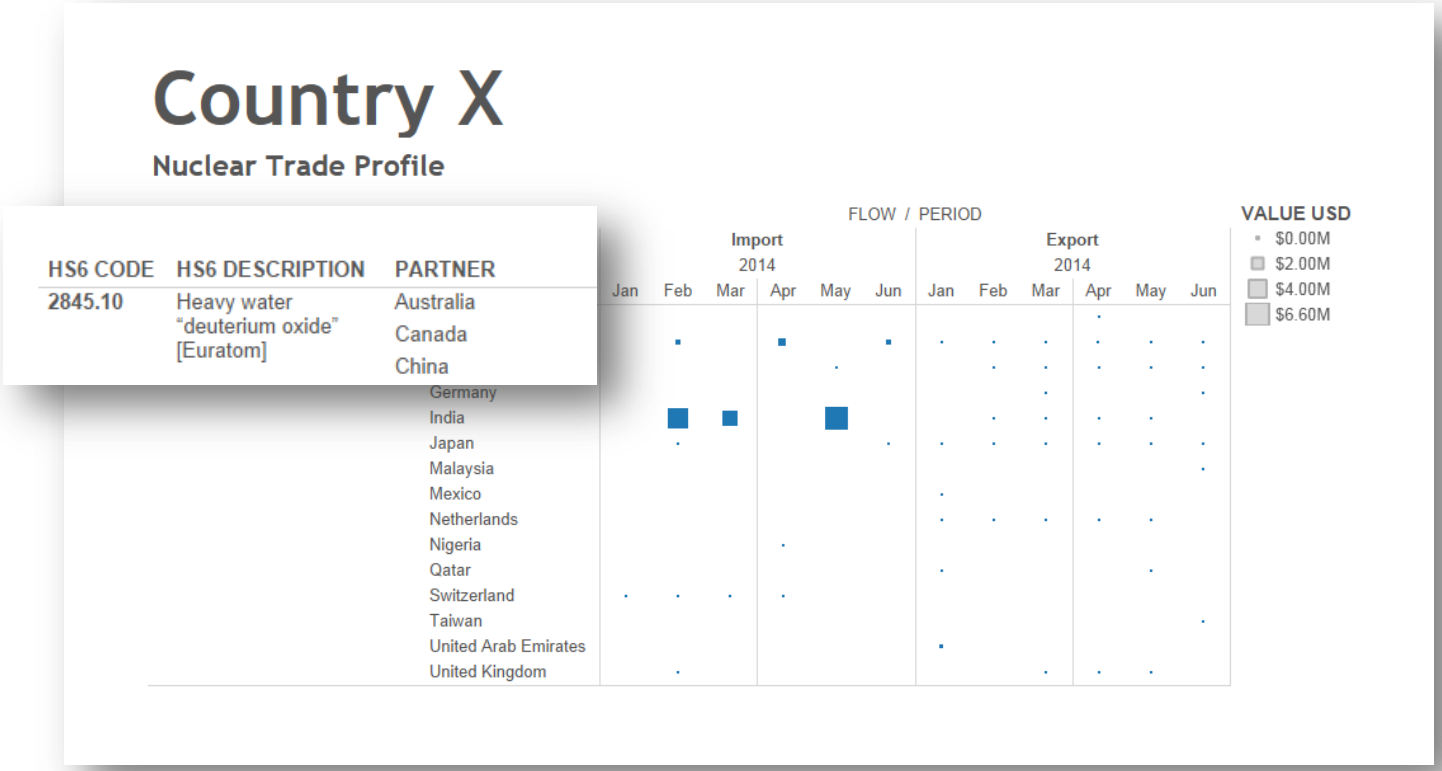
Data visualisation – Details on demand



HS6 CODE	HS6 DESCRIPTION
2612.10	Uranium ores and concentrates
2844.10	Natural uranium and its compounds; alloys, dispe..
2844.20	Uranium enriched in U 235 and its compounds: pl..
2844.30	Uranium depleted in U 235 and its compounds; th..
2844.40	Radioactive elements, isotopes and compounds, ..
2844.50	Spent "irradiated" fuel elements "cartridges" of nu..
2845.10	Heavy water "deuterium oxide" [Euratom]
2845.90	Non-radioactive isotopes; inorganic or organic co..
8109.90	Articles of zirconium, n.e.s.
8401.10	Nuclear reactors [Euratom]
8401.20	Machinery and apparatus for isotopic separation ..
8401.30	Fuel elements "cartridges", non-irradiated, in casi..
8401.40	Parts of nuclear reactors, n.e.s. [Euratom]

Example

Data visualisation – Details on demand



Conclusions

- Issues in data retrieval and analysis arise when:
  - The data are 'invisible'
  - Data access starts by questions and not by data presentation
  - Retrieval and analysis systems strive more for results' precision (correctness) than accuracy (completeness).

- Data visualisation approaches can mitigate these issues in that priority is given to data presentation.
  This encourages data exploration by the analyst,
  enabling more accurate results and higher data integrity.

- A key point, often not understood, is that data visualisation requires working with raw data, not 'result set data'.

Acknowledgements

The work presented is funded by the European Commission, Joint Research Centre, in projects: ***VideoZoom*** *and* ***Strategic Trade Analysis for Non Proliferation****.*

Both projects contribute to the EC Support to the IAEA.

References

[1] Silver N. (2012) – *The Signal and the Noise: Why Most Predictions Fail but Some Don't.* ISBN 978-1-101 59595-4

[2] Mackinlay J. (2014) – *Jock's Dream of Data Sushi.* Presentation at Tapestry 2014. https://www.youtube.com/watch?v=EsyMkuMM8HU

[3] Cojazzi G.G.M., Versino C., Wolfart E., Renda G., Janssens W. (2014) – *Tools for Trade Analysis and Open Source Information Monitoring for Nonproliferation.* Symposium on International Safeguards: Linking Strategy, Implementation and People. IAEA, Vienna, 20-24 October 2014.

[4] Blunsden S., Versino C. (2011) – *VideoZoom: Summarizing surveillance images for safeguards video reviews.* EUR 25215 EN, ISBN 978-92-79-23091-2, JRC 68054.

[5] Versino C., Rocchi S., Hadfi G., John M., Jüngling K., Moeslinger M., Murray J., Sequeira V.(2014) – *Evaluation of a Surveillance Review Software based on Automatic Image Summaries.* Symposium on International Safeguards: Linking Strategy, Implementation and People. IAEA, Vienna, 20-24 October 2014.

[6] Juengling K., Blunsden S., Versino C. (2014) – *VideoZoom*: *An Interactive System for Video Summarization, Browsing and Retrieval.* 10th International Symposium on Visual Computing. Las Vegas, Nevada, USA. To appear.