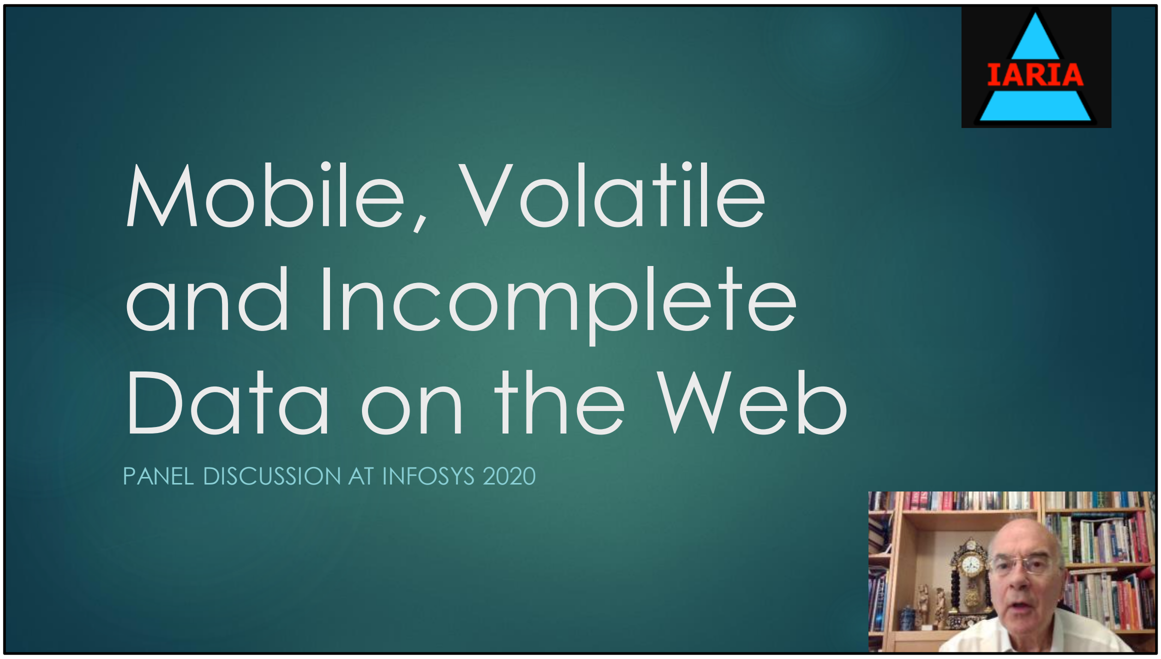
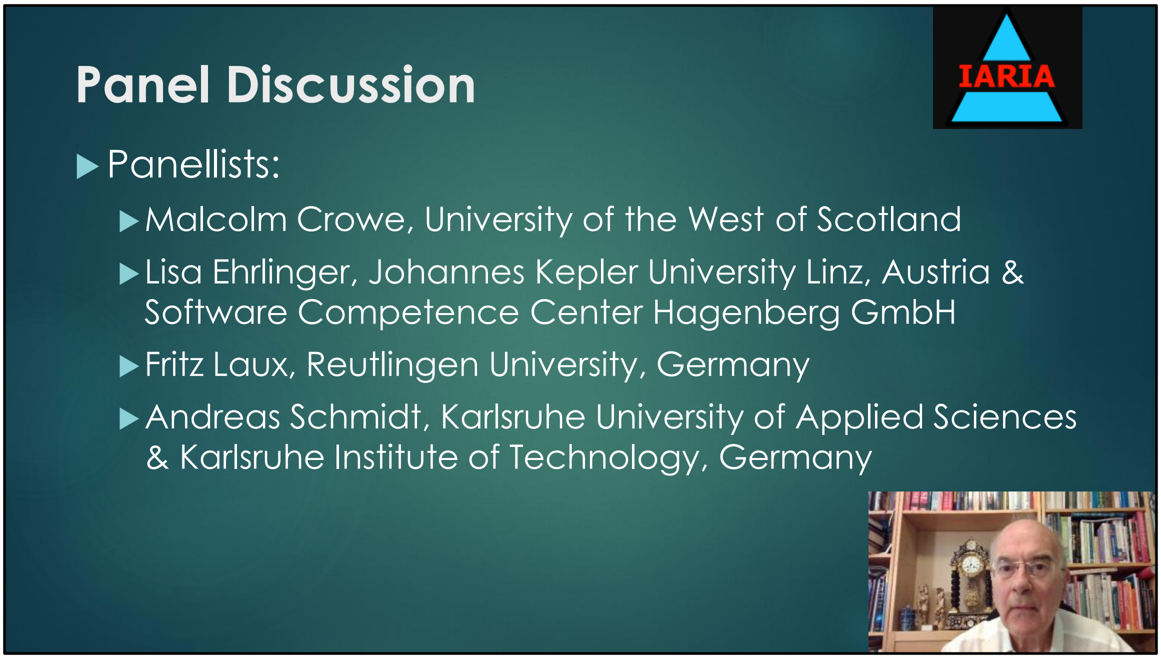


Mobile, Volatile and Incomplete Data on the Web

PANEL DISCUSSION AT INFOSYS 2020

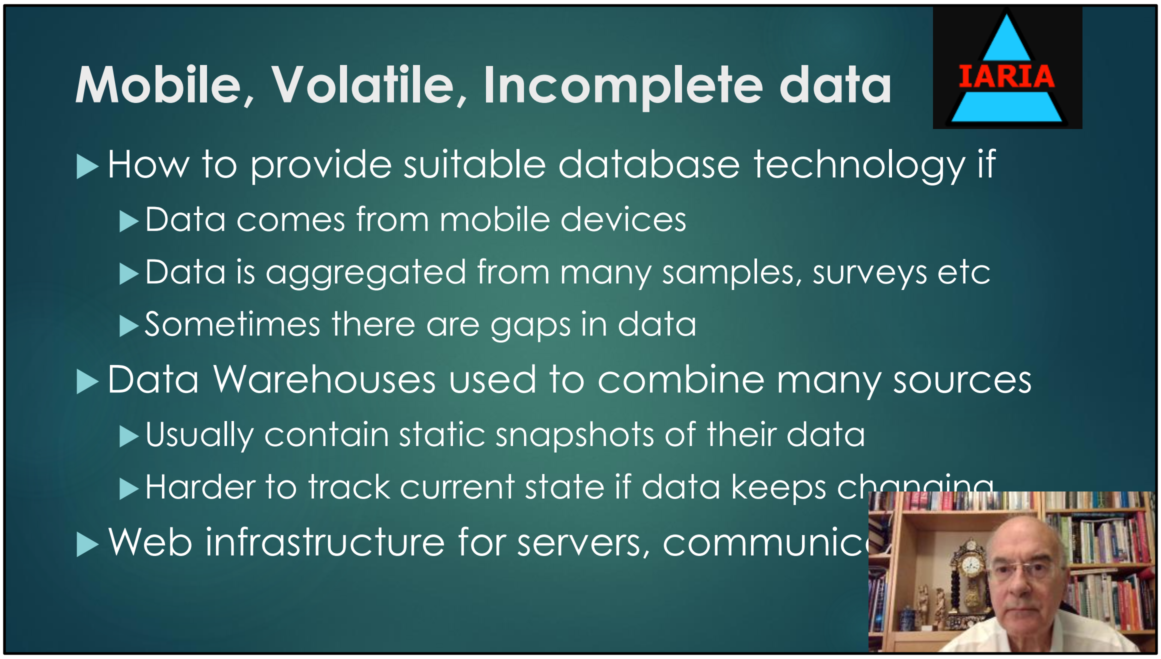
Welcome to this Panel Session at DBKDA 2020, Lisbon



Panel Discussion

- Panellists:
  - Malcolm Crowe, University of the West of Scotland
  - Lisa Ehrlinger, Johannes Kepler University Linz, Austria & Software Competence Center Hagenberg GmbH
  - Fritz Laux, Reutlingen University, Germany
  - Andreas Schmidt, Karlsruhe University of Applied Sciences & Karlsruhe Institute of Technology, Germany

I'm Malcolm Crowe, a retired academic from University of the West of Scotland. Our panellists for this session are Lisa Ehrlinger, from Johannes Kepler University, Fritz Laux, from Reutlingen University, and Andreas Schmidt from Karlsruhe Institute of Technology.



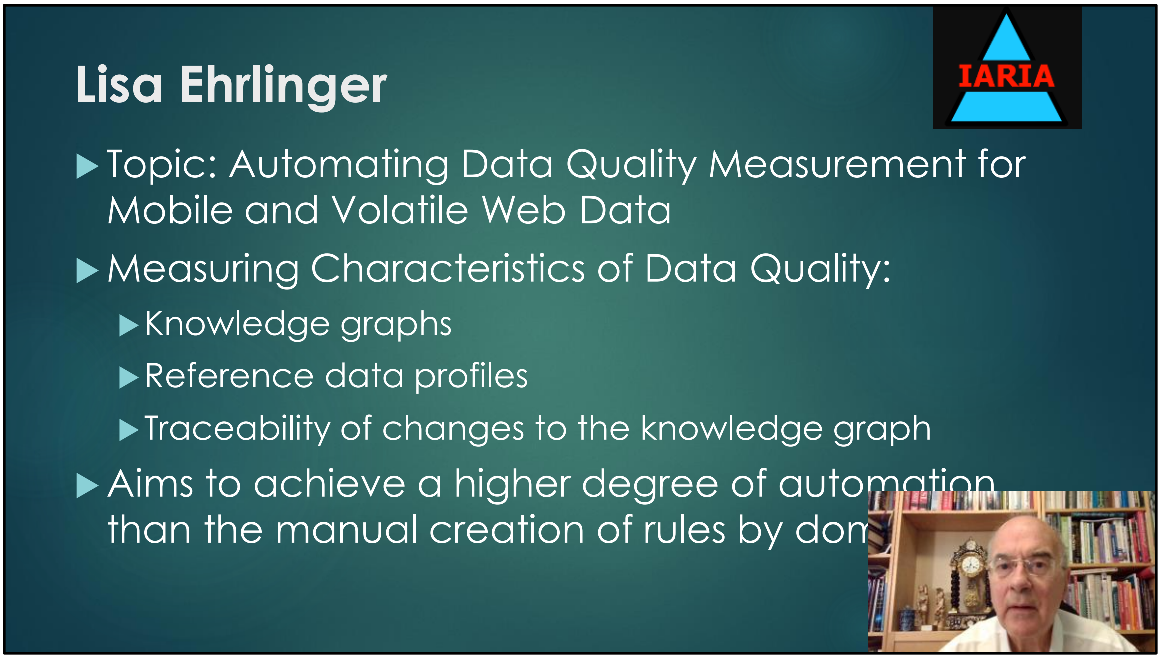
Mobile, Volatile, Incomplete data

- How to provide suitable database technology if
  - Data comes from mobile devices
  - Data is aggregated from many samples, surveys etc
  - Sometimes there are gaps in data
- Data Warehouses used to combine many sources
  - Usually contain static snapshots of their data
  - Harder to track current state if data keeps changing
- Web infrastructure for servers, communic

Our topic: Mobile, Volatile, Incomplete Data leads us to consider how to provide suitable database technology if data is being supplied from mobile devices, if the data keeps changing, or if the data is aggregated from many samples, surveys and such things.

Common to all these topics is a concern with Data Integration. The current state of the art is mostly data warehouses built from static snapshots of data, but most important data sets evolve from many sources.

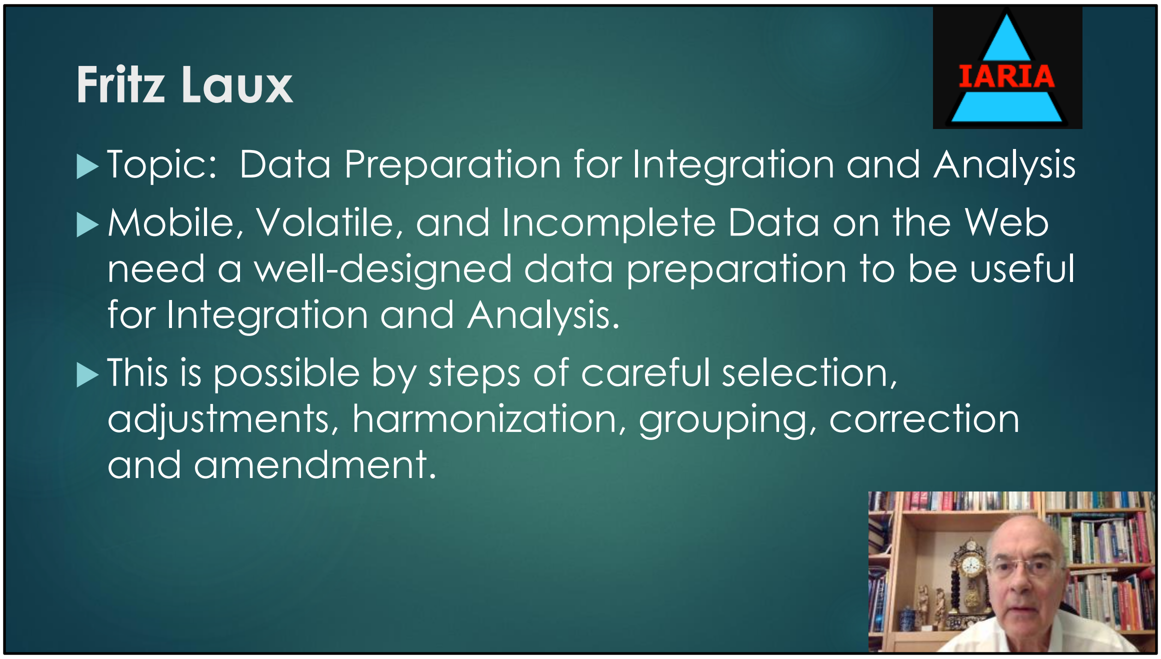
And obviously we assume the use of the Web (or at least TCP/IP) as the platform for servers and communications.



Lisa Ehrlinger

- Topic: Automating Data Quality Measurement for Mobile and Volatile Web Data
- Measuring Characteristics of Data Quality:
  - Knowledge graphs
  - Reference data profiles
  - Traceability of changes to the knowledge graph
- Aims to achieve a higher degree of automation than the manual creation of rules by dom

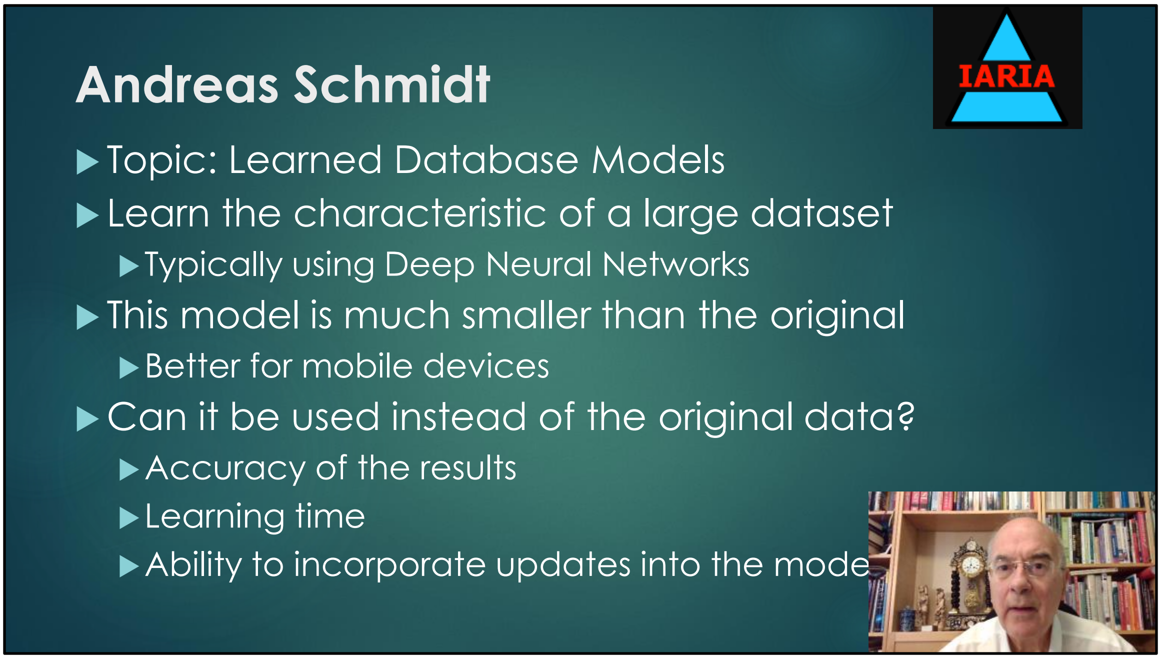
Lisa’s contribution deals with Automating Data Quality Measurement for Mobile and Volatile Web Data. Measuring characteristics of such data can use knowledge graphs, reference data profiles, and the traceability of changes to the knowledge graph. Mostly such measurements are done by domain experts, but automated tools will be needed if the data keeps changing.



Fritz Laux

- Topic: Data Preparation for Integration and Analysis
- Mobile, Volatile, and Incomplete Data on the Web need a well-designed data preparation to be useful for Integration and Analysis.
- This is possible by steps of careful selection, adjustments, harmonization, grouping, correction and amendment.

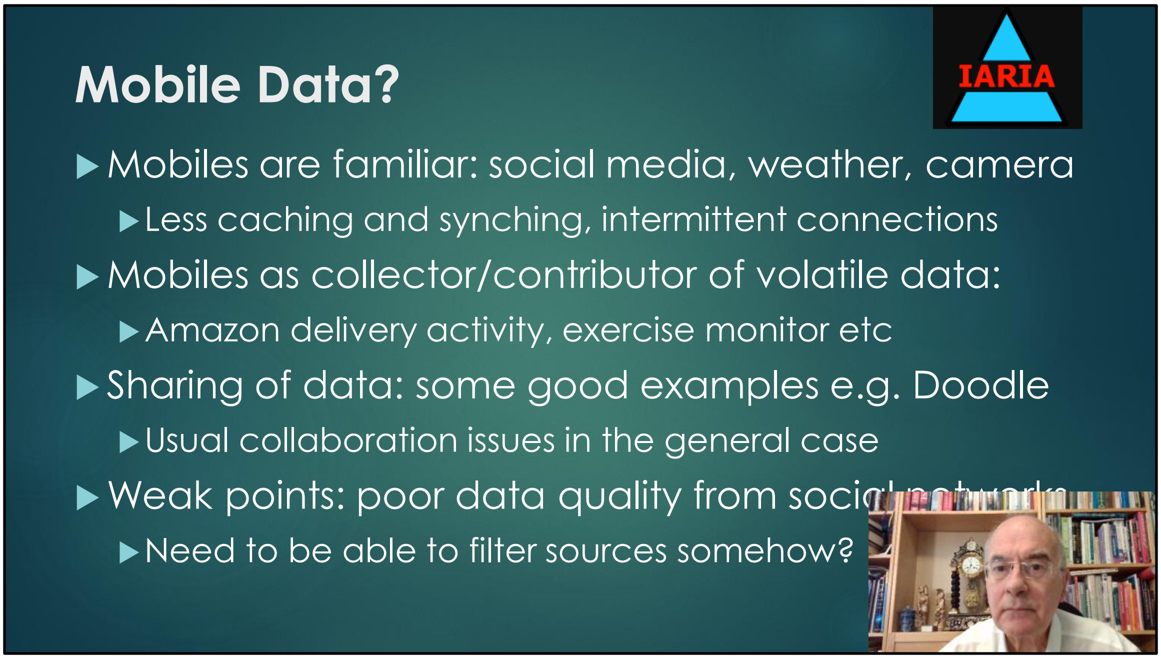
Fritz will discuss the data preparation and analysis that is needed for data integration. The tasks here are selection of suitable sources, adjusting and harmonising their details, grouping them appropriately, dealing with any anomalies, and preparing for later changes to the data.



Andreas Schmidt

- Topic: Learned Database Models
- Learn the characteristic of a large dataset
  - Typically using Deep Neural Networks
- This model is much smaller than the original
  - Better for mobile devices
- Can it be used instead of the original data?
  - Accuracy of the results
  - Learning time
  - Ability to incorporate updates into the mode

Andreas’ topic is Learned Database Models. Given a large dataset, the idea is to learn the characteristics of the data, generally using deep neural networks with the aim of constructing a predictive model of the data. This is much smaller and easier to query on a mobile device, and it is important to know if it can be used instead of the huge original database. Since the model takes a long time to construct, how should it be modified when new data arrives?



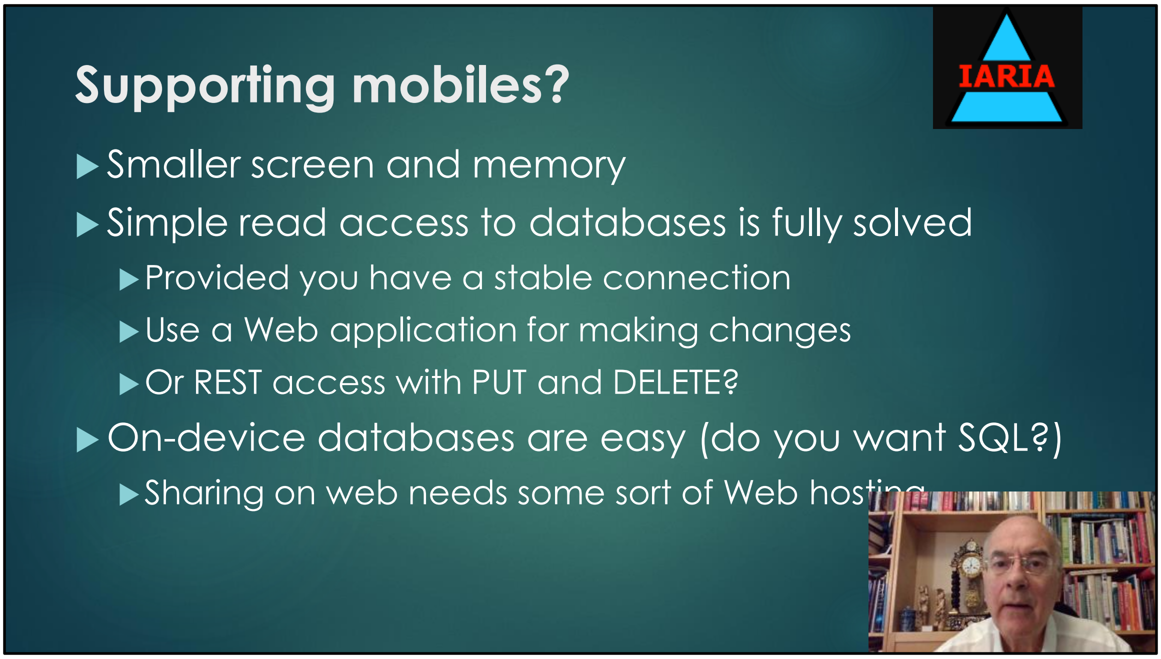
Mobile Data?

- Mobiles are familiar: social media, weather, camera
  - Less caching and synching, intermittent connections
- Mobiles as collector/contributor of volatile data:
  - Amazon delivery activity, exercise monitor etc
- Sharing of data: some good examples e.g. Doodle
  - Usual collaboration issues in the general case
- Weak points: poor data quality from social networks
  - Need to be able to filter sources somehow?

Almost everybody uses mobile devices for an enormous variety of data – chat, weather, as a camera, for buying things or arranging meetings or holidays. Importantly, many of these activities see the mobile device as a source of new data which is obviously being stored in many databases. A large proportion of such data is obviously volatile: we can think of the current position of a van driver delivering a parcel, the most recent message from our friends, the reading from an exercise monitor.

In many cases, mobile devices play the same role as collaborating desktop clients, for example arranging meetings, or email. Some involve collaborative editing of data and documents, such as with Doodle, or live meetings. It is clear that the technical issues involved in such applications are largely solved, or at least that many ways of managing collaboration have become accepted.

Other issues are more difficult: dealing with fake news, fake reviews, lies and fraud will always be with us, and where facts and accuracy are important there is a need to be able to filter data somehow. Alas, too many business executives insist on being allowed to alter data analytics before publishing them.



Supporting mobiles?

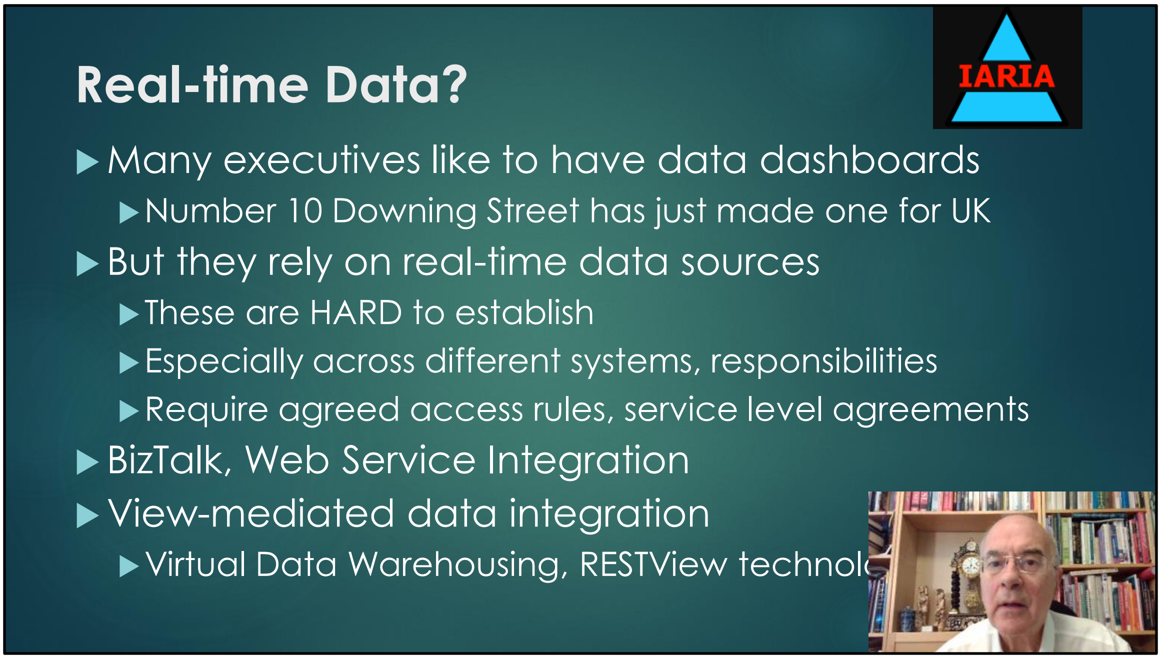
IARIA

- Smaller screen and memory
- Simple read access to databases is fully solved
  - Provided you have a stable connection
  - Use a Web application for making changes
  - Or REST access with PUT and DELETE?
- On-device databases are easy (do you want SQL?)
  - Sharing on web needs some sort of Web hosti

While there are similarities between integrating desktop clients and mobile devices, mobiles do bring their own issues. The user interface is different, and the network connection comes and goes.

Nevertheless, for simple read access to online data, the problems of supporting mobiles are pretty much solved by current or Web technology and web services. Web applications now make a great success of making changes to online data, and maybe somewhat round the corner we can expect more adoption of RESTful services for the same purposes.

It is even possible to host a small database on your mobile: many applications effectively do this, though SQL databases are rarely hosted on mobiles.



Real-time Data?

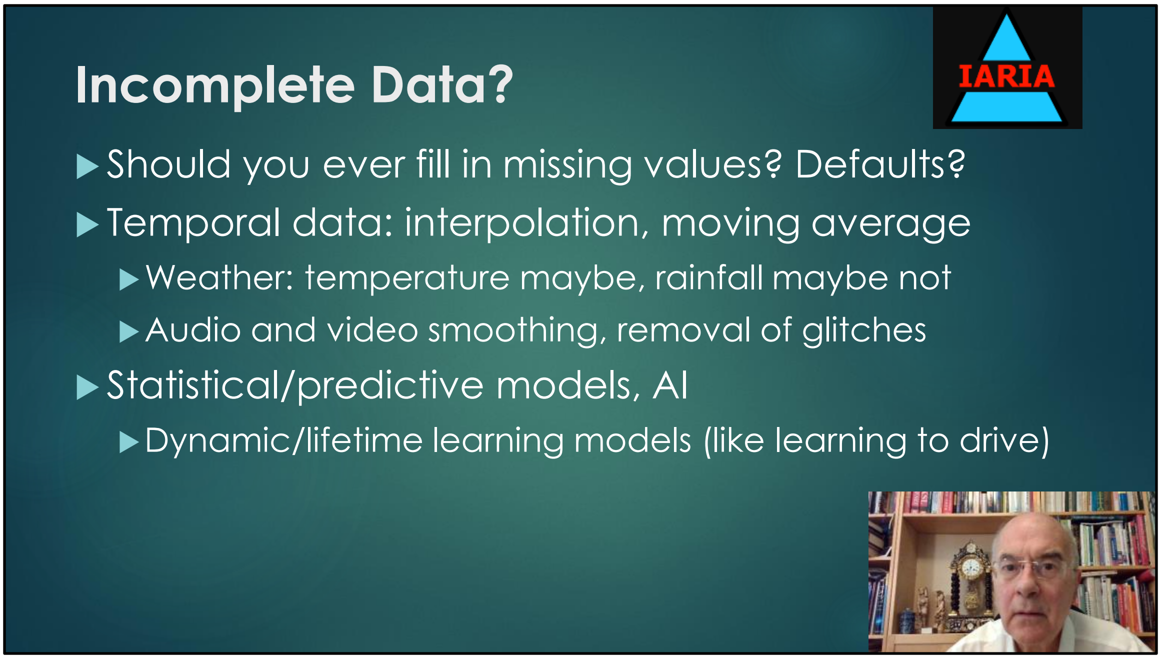
- Many executives like to have data dashboards
  - Number 10 Downing Street has just made one for UK
- But they rely on real-time data sources
  - These are HARD to establish
  - Especially across different systems, responsibilities
  - Require agreed access rules, service level agreements
- BizTalk, Web Service Integration
- View-mediated data integration
  - Virtual Data Warehousing, RESTView technol

Social media and news feeds can provide real-time data of course, and many businesses dream of having data dashboards to enable them to see in real time how their business is performing. Most universities now have these for student and marketing activity. Newspapers have reported that Number 10 Downing Street has just installed one for monitoring government policy.

But in most cases, the data is very far from real-time. To provide real-time performance data, we need real-time data sources, and except for the simplest cases this are very hard to establish, particularly when the data is integrated or aggregated from different sources with different ownership and responsibilities. In such cases there are always service level agreements to be negotiated to establish the rules of access.

There are intermediate cases where success is currently possible, using web service integration and messaging hubs such as BizTalk. These allow direct interrogation of data sources, and with enough programming effort they can be made to incorporate data from other places. Personally, I believe there is more that can be done to provide tools for the general case by better exploitation of HTTP and particularly

REST, together with the concept of view-mediated virtual data warehousing.



Incomplete Data?

- Should you ever fill in missing values? Defaults?
- Temporal data: interpolation, moving average
  - Weather: temperature maybe, rainfall maybe not
  - Audio and video smoothing, removal of glitches
- Statistical/predictive models, AI
  - Dynamic/lifetime learning models (like learning to drive)

For weather forecasting we are used to displays that immediately follow the weather as gathered from satellites and tracking by a similar-looking video that shows a forecast evolution. Obviously, it is important to distinguish facts from forecasts. Lives have been lost by over-reliance on predictive analytics by governments and police forces.

I am told that as a practical matter it is more dangerous to neglect a data source because of some missing values, and there are different mechanisms to resolve these, some of which are more convincing than others.

I have been impressed by recent work in reinforcement learning that solves the problem of continuous or lifetime learning by allowing the agent to resume learning if things change. I look forward to these new ideas finding a place in data integration technology for one or more of the problems considered above.