



Successful big data projects

David Cameron
SIRIUS Centre for Scalable Data Access
University of Oslo
Tekna Course, Oslo, 15th June 2017



The SIRIUS Centre

- Eight years' financing
- 11 Industrial Partners
- 3 Leading Academic Institutions
- Centre for Research-Based Innovation
- 20 Ph.D. students
- Innovation through prototypes and pilots



UiO : University of Oslo

[simula . research laboratory]



Statoil





Snake oil warning!



- Everybody is talking digitalisation, big data and data science.
- ... and defining what they do as digitalisation or data science.
- This keeps the money flowing.
- Historical inevitability and inflated expectations.
- Be sceptical, take the good and discount the hype.
- Know that your professional skills and relationships remain relevant.



What will we do in 2022?



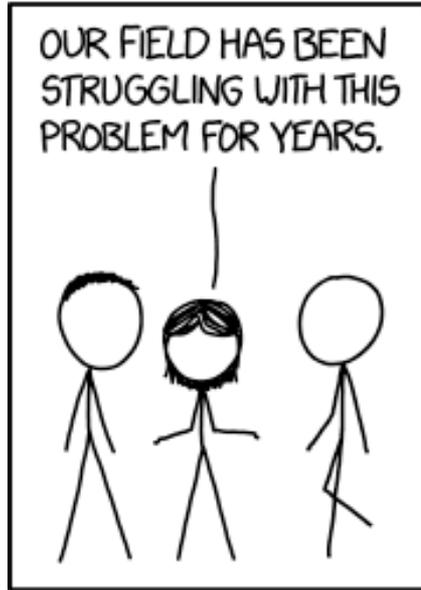
But this will be:

1. More expensive than we expect.
2. Will be more difficult than the evangelists tell us.
3. Will not meet expectations.
4. Will change the ways we work
5. Can give us a right to exist and sell our products and services.

How can we run these pioneering projects well!



Analytics, anyone?



Creative Commons Attribution-NonCommercial 2.5 License. <https://xkcd.com/1831/>



Analytics needs and feeds understanding

- Consider an example of fluidisation
- Rampant empiricism, without power to extrapolate and design
- Physical insight gave form of equations that made sense of data and gave predictive power.
- This was done in 1963.
- http://www.icheme.org/media_centre/news/2016/ninety-not-out-john-davidson-recognised-for-fluidisation-research.aspx#.WT11WmjyIU





The problem of scalable data access

- Different formats
- Old software
- Complex, inconsistent data models
- Inefficient access methods
- Access and security
- Unstructured data
- Missing data
- Poor-quality data
- Too much data
- Manual work processes



Accessing data is a technical and organizational bottleneck.

We make poorer decisions and waste time on tedious work getting data.



Big data projects are *data* projects

- Use standard good project practice
- But with some twists:
 - Complex and difficult data sources
 - Non-SQL databases and file databases
 - Large-scale parallel computing
 - Statistics and other mathematical calculations
 - Methods designed for consumer and transaction data



Why are you doing the project?

- Technology push is not enough beyond a small prototype.
- Improve the quality of a business decision or a business process.
 - Best is measurable increase in profit.
 - Greater revenue.
 - Lower cost.
 - Resulting in a ROI, or at least a pay-back time.
 - Can also provide an “insurance” business case:
 - Preventing things from going wrong.



Business drivers in engineering firms

Øyvind Eriksen , Aker ASA, 4th April 2017

- Modernise, standardise and simplify standards and procedures.
 - Save capital cost
- Sharing of information and re-use effect in design, construction and operations
 - Reduce project cost
- Common strategy for digitalisation, automation and adoption of robots.
 - Reduce operational cost





Talk to your business!

- Find a valuable business decision
- Which data is used to make this decision?
- Which data could be used to make this decision?
- Can better data access and analysis improve the quality of this business decision?



How big is your data?

- Which of the Vs make it big?
 - Velocity, Volume, Variety
 - Veracity (the V that everybody forgets)
- Or is it an important middle-sized data set?
- Is the problem broad and shallow or narrow and deep?



Data with value. Some of it is big

Interfaces between a topsides module and a jacket

Staffing and progress plan for a turn-around

Nominations for a month in a pipeline network.

Well tests for a facility.

Alarms generated by a production facility in a year.

Inspection and testing status of every relief valve in a facility.

≈500 Piping and Instrumentation Diagrams for a facility.

PVT analyses for a facility.

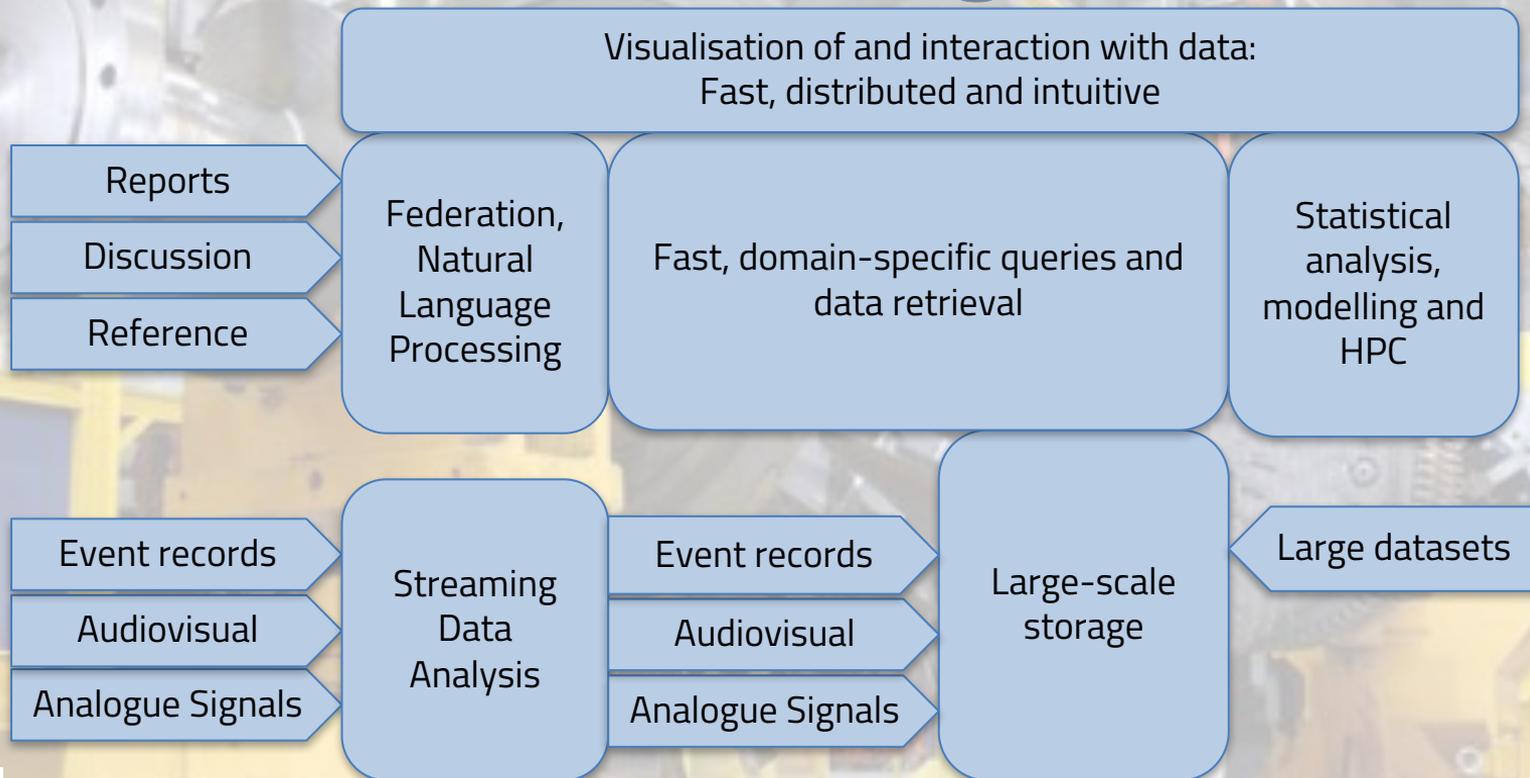
≈2000 maintenance projects per year in an oil field, gas plant or refinery.

Training and safety certification records for the employees of an operator and its contractors

≈50000 product quality / GMP records in a LIMS



Think about combining data





Agile and two-speed IT is needed

- Gartner concept from 2105: bimodal IT
- A "data lake" or staging database can help by decoupling the high-speed IT from the low-speed IT.



What sort of project team?

- Project Leader
- People who know the business (users)
- Specialists:
 - Data access and databases
 - Analytics and data science
 - Developers and tool experts
 - User experience



Pitfalls

- Biting off too much
- Ignoring bad data and Murphy's law of computing
- Choosing the oldest and least digital processes to work with
- Vague business value – technology for its own sake
- Privacy, security and confidentiality issues



Remember the work practices

- Data doesn't innovate: people do.
- Remember the externalities:
 - privacy, labour laws, discrimination, resilience, environment
- Why does so much industrial IT fail to meet its expectations?
- How can we ensure that the what we implement is embedded in a community of practice?