

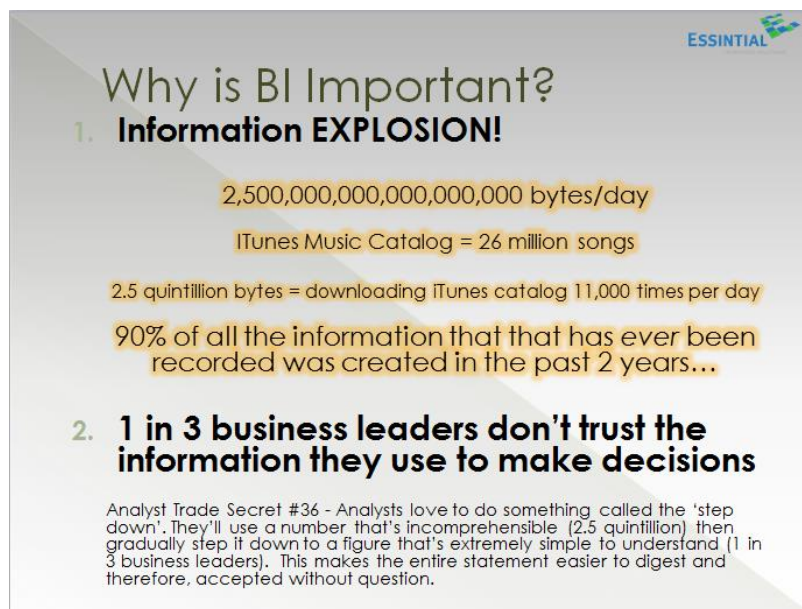


SIA Executive Summit 2015

BIG DATA

John A. Thompson
Managing Director
Information Professionals GmbH

Picking up where you left off in 2013



Why is BI Important?

- 1. Information EXPLOSION!**
 - 2,500,000,000,000,000 bytes/day
 - iTunes Music Catalog = 26 million songs
 - 2.5 quintillion bytes = downloading iTunes catalog 11,000 times per day
 - 90% of all the information that that has ever been recorded was created in the past 2 years...
- 2. 1 in 3 business leaders don't trust the information they use to make decisions**

Analyst Trade Secret #36 - Analysts love to do something called the 'step down'. They'll use a number that's incomprehensible (2.5 quintillion) then gradually step it down to a figure that's extremely simple to understand (1 in 3 business leaders). This makes the entire statement easier to digest and therefore, accepted without question.



The Future...

- Predictive Failures
- 'Big Data'

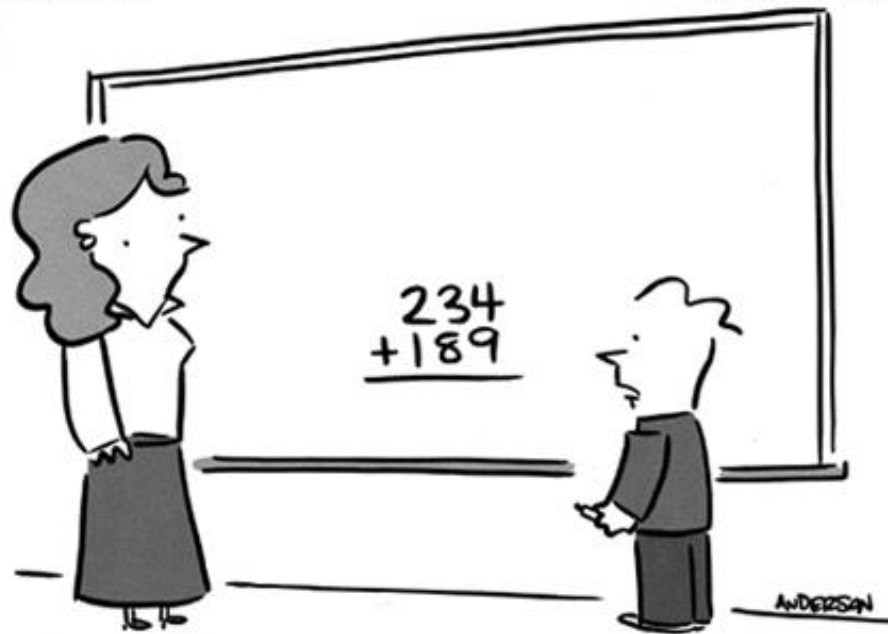
Source: "Business Intelligence: Decisions with data", Tom Clauser, Essintial Enterprise Solutions, SIA Executive Summit 2013

Big Data talk: the short version

- The idea of data assets is here, it's here to stay, and it's changing the way we all do business.
- Emerging "Big Data" technologies allow us to solve business problems which were unsolvable up until a few years or even months ago.
- Advanced Analytics is the key to success with Big Data, and that requires Data Science expertise.
- Big Data is a challenge you must respond to, otherwise you will lose your competitive edge.
- Big Data is not hard, it needn't be expensive, and you can get going tomorrow.
- To prove it, I propose implementing a pilot project with SIA and/or its interested members.

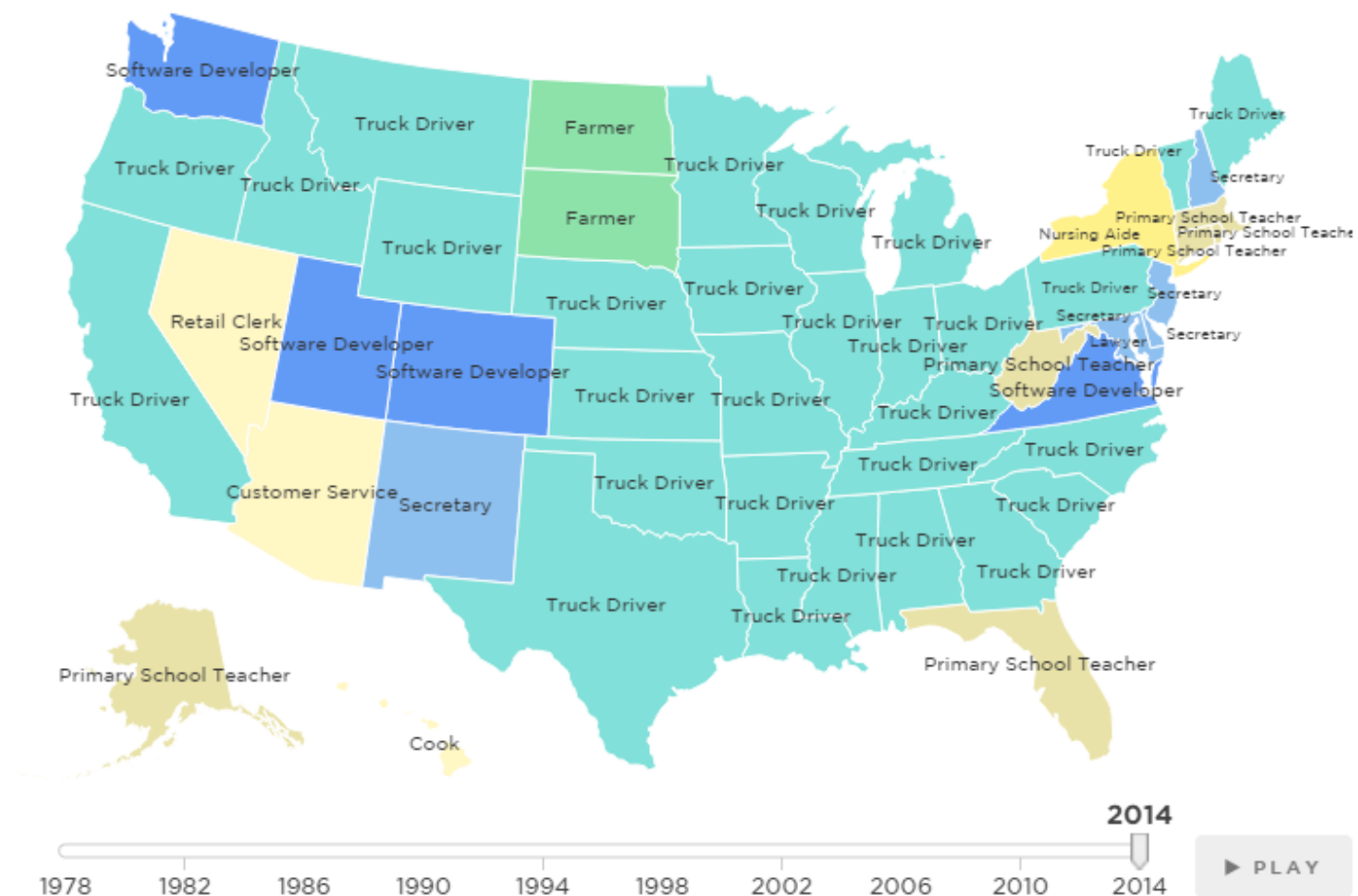
© MARK ANDERSON

WWW.ANDERTOONS.COM



"Does this count as big data?"

The Most Common* Job In Each State 1978-2014



Source: <http://www.npr.org/blogs/money/2015/02/05/382664837/map-the-most-common-job-in-every-state>

The Sea Change

1. The big new variable in business strategy
2. New opportunities and threats are emerging now
3. How will you respond?
 - individually
 - as a group
4. The next, urgent steps (are easy) to take

"... a sea-change
Into something rich and strange."
- *The Tempest*, Act 1, Scene 2

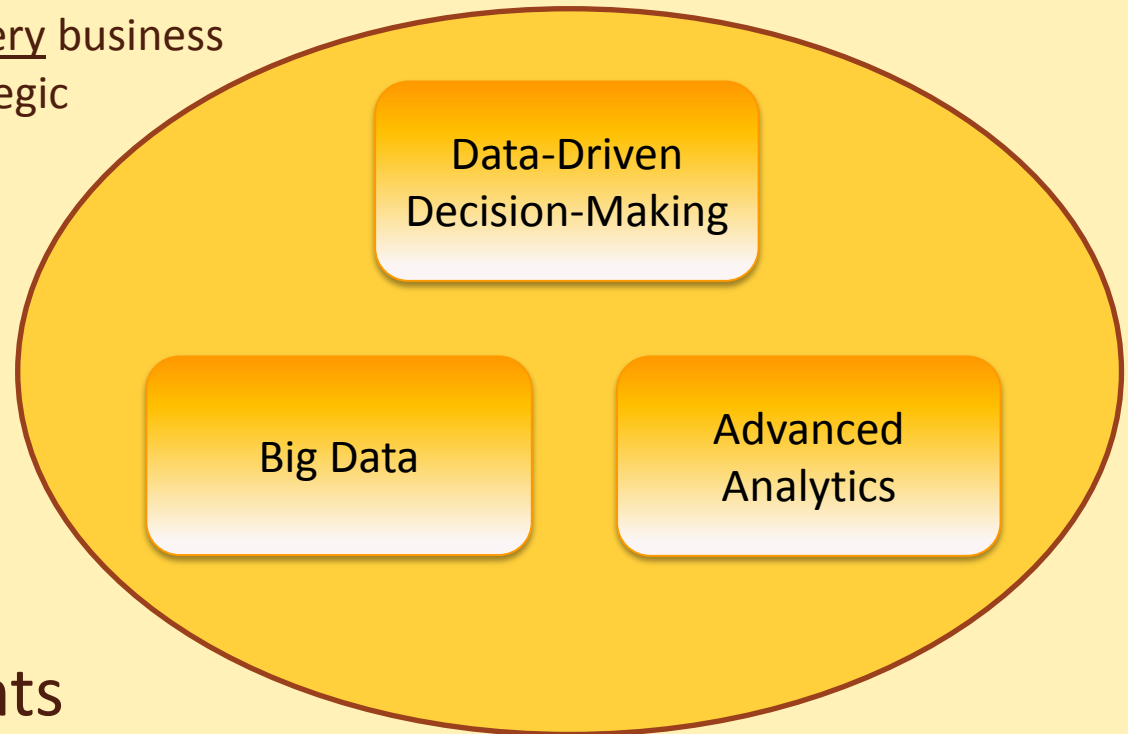


The big new variable in business strategy

... doesn't even have a name yet

Let's call it "Data Assets"

- Transforming every business
- Requiring a strategic response
- Now

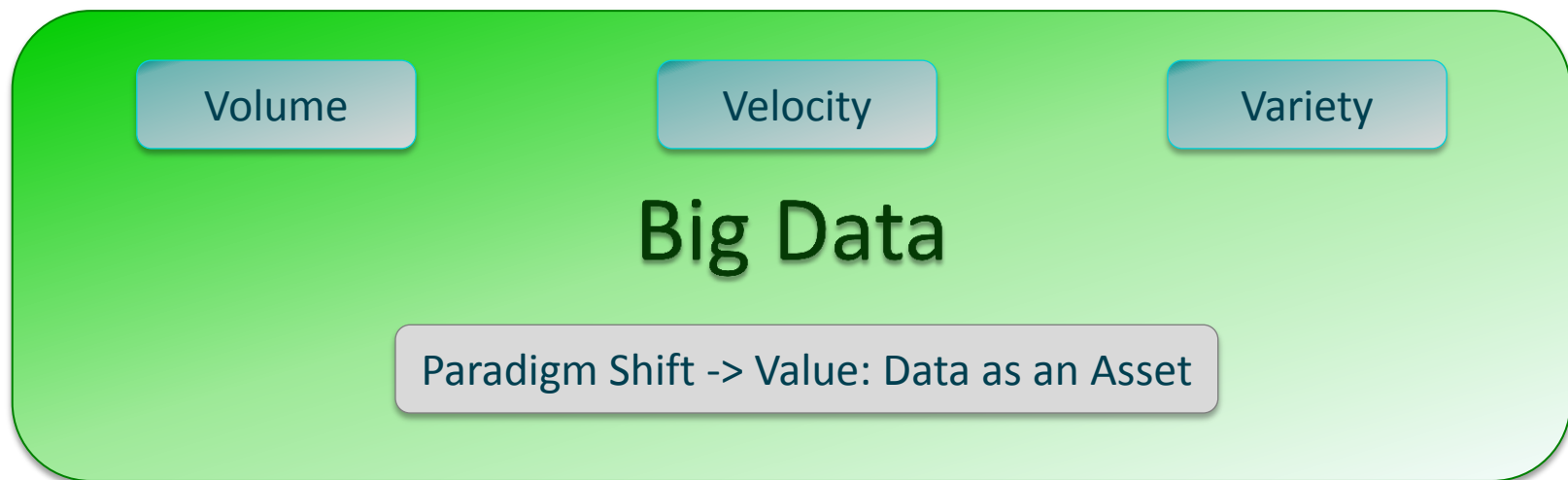


3 technical
developments
merging into one strategic factor

What is Big Data?

- Big Data is when you need special infrastructure to handle storage and processing of your data. In practice, that means:
 - it's too big to fit on one node in your network
 - it's coming too fast for your RDBMS to handle it
 - structured, semi-structured and unstructured data need to be automatically combined for decision-making
- You need special infrastructure because even simple things like counting the rows in a table are non-trivial tasks.

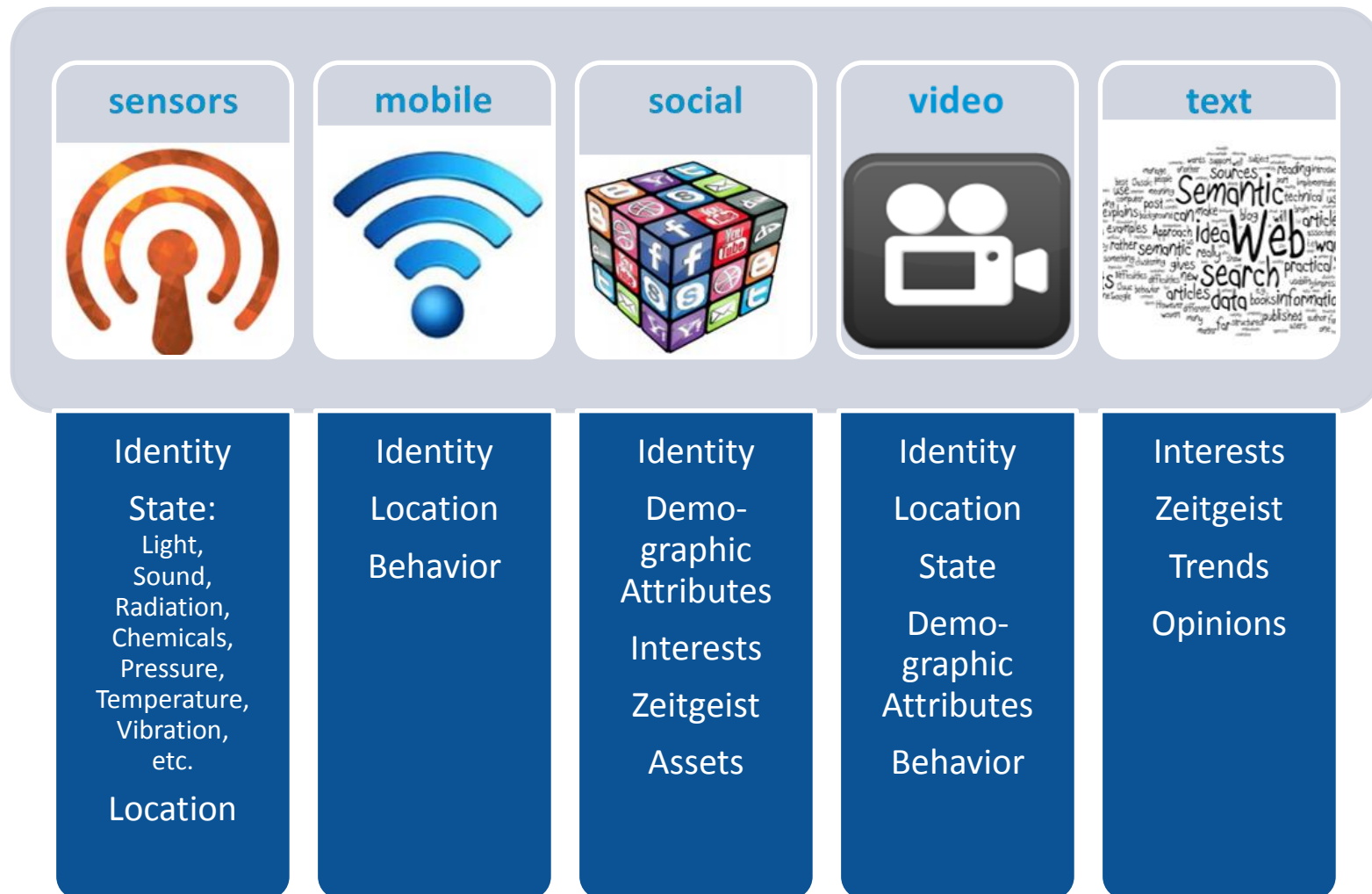
Big Data Characteristics: the 3 (or 4) V's



These are widely accepted and repeated characteristics of Big Data, first propagated by IBM, then taken over by Microsoft.

Vendors and consultants like to invent new "V's", such as: Veracity, Visualization, Variability and Viability, so you may hear up to 8 V's.

Massive sources of data










The technical innovation

- Massively parallel data storage and data processing at a reasonable price
- Alternative data storage technologies, e.g. columnar and graph-based stores
- Effective pattern recognition in very large, complex data sets, if necessary in real-time

Computing just got really cheap, and it's getting cheaper faster than you can imagine.

Big Data: the change in technical focus

Bring data to processing		Bring processing to data
Extract – Transform – Load		Extract – Load – Transform
Schema-on-write		Schema-on-read
Data and processing on one network node or small cluster		Massively scalable data and processing
Licensed solutions		Commodity hardware, open source software
Rely on exact results, using deterministic methods		Rely on statistical results, using probabilistic methods
Decision support systems, limited access, limited reach		Automated, data-driven decisions everywhere

What is Advanced Analytics?

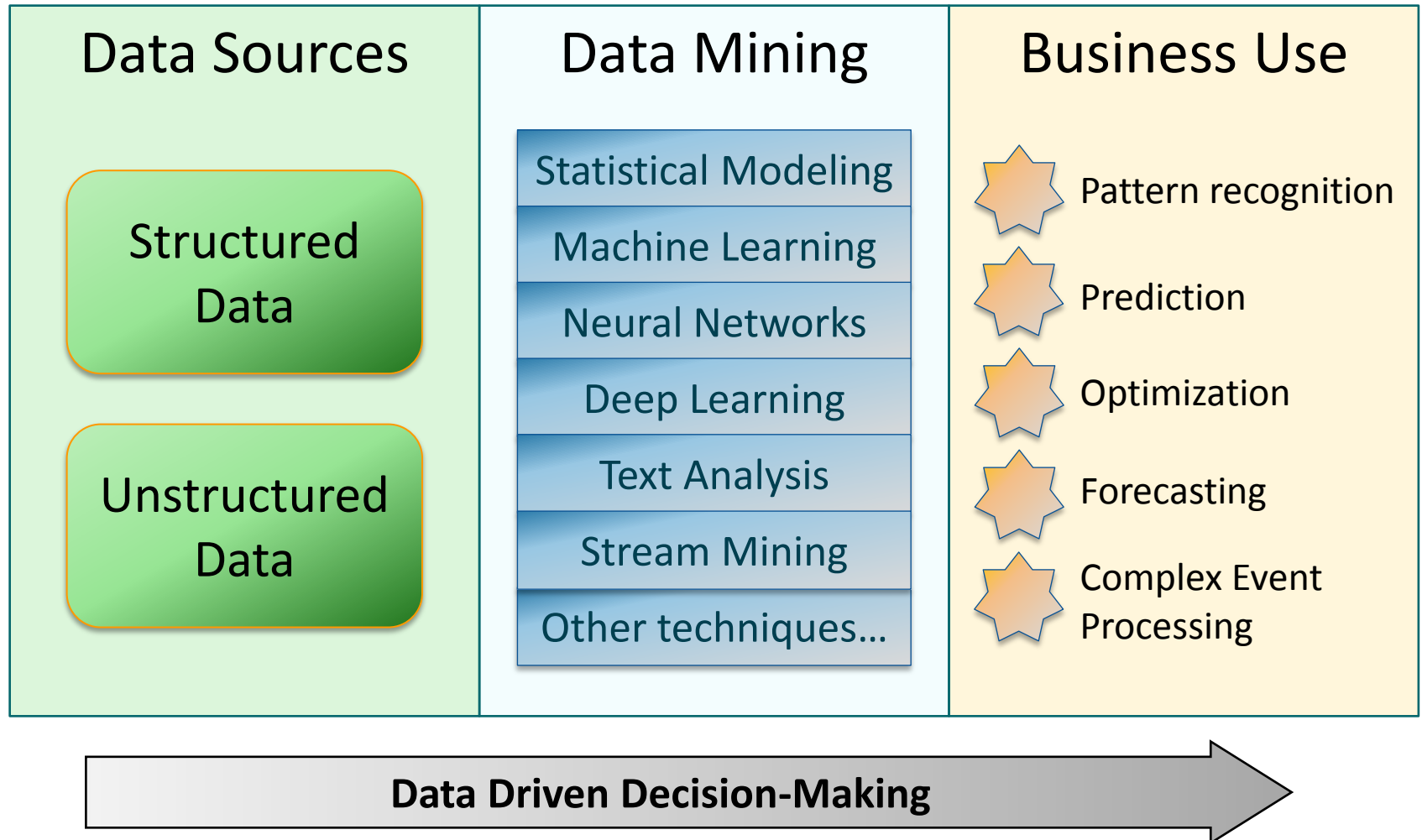
"Advanced analytics provides algorithms for complex analysis of either structured or unstructured data. It includes sophisticated statistical models, machine learning, neural networks, text analytics, and other advanced data mining techniques.

Among its many use cases, it can be deployed to find patterns in data, prediction, optimization, forecasting, and for complex event processing/analysis.

Examples include predicting churn, identifying fraud, market basket analysis, or understanding website behavior. Advanced analytics does not include database query and reporting and OLAP cubes."

- *Dr. Fern Halper / TDWI Research Director Advanced Analytics, 2010*
(<https://fbhalper.wordpress.com/2010/12/20/what-is-advanced-analytics/>)

Advanced Analytics



Query = Function(All Data)

Big Data application areas

- Simple Batch Processing
- Data Stream Mining
- Search & Discovery
- Machine Learning

Simple Batch Processing

- 1 TB to 100's of TB
- Simple processing of data, e.g. with MapReduce
- Trivial parallel processing and storage
- Appropriate if real-time processing is not required
- Typical example: how many known websites reference each of the known websites?

Input: crawler results;

Output: count per website

Data Stream Mining

- GB up to TB per day of dynamic data
- Real-time recognition of relevant event, e.g. from sensors, images, video, web traffic, etc.
- Automatic creation of "summaries" of the data, or Complex Event Processing on current time windows
- Typical example: predictive maintenance
 - Which components in my network will fail soon?
 - Which machines are beginning to act peculiarly?

Data Discovery

- Data Mining of up to 100's of TB of data
- Batch-oriented or ad hoc discovery of unknown correlations in large datasets
- Effective data visualization is critical
- Typical examples:
 - forensic investigations of IT logs
 - preprocessing of newsfeeds for topics and trends
 - extension of classical BI solutions with interactive (ad hoc) analysis and visualization of highly scalable data, either structured or unstructured

Machine Learning

- Continually expanding data sets, up to 100's of TB
- Batch or real-time determination of predictions
- The more data are available, the smarter the system becomes
- Typical example: recommendation systems – through pattern recognition (on properties, behavior, etc.) recommend a decision, e.g. to sort out spam email or to recommend a product

Deep Visual-Semantic Alignments for Generating Image Descriptions

"We present a model that generates free-form natural language descriptions of image regions."

- Andrej Karpathy, Lie Fei-Fei, Stanford Vision Lab, 2014

Source: <http://cs.stanford.edu/people/karpathy/deepimagesent/>

2014 was the year of Deep Learning.

Among other things:

The problem of training a machine to recognize objects and activities in images is finally being solved.

It just might be worthwhile to think about the implications for your business.

Multimodal Recurrent Neural Network

Our Multimodal Recurrent Neural Architecture generates sentence descriptions from images. Below are a few examples of generated sentences:



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."



"boy is doing backflip on wakeboard."



"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."



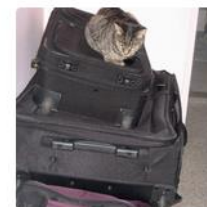
"little girl is eating piece of cake."



"baseball player is throwing ball in game."

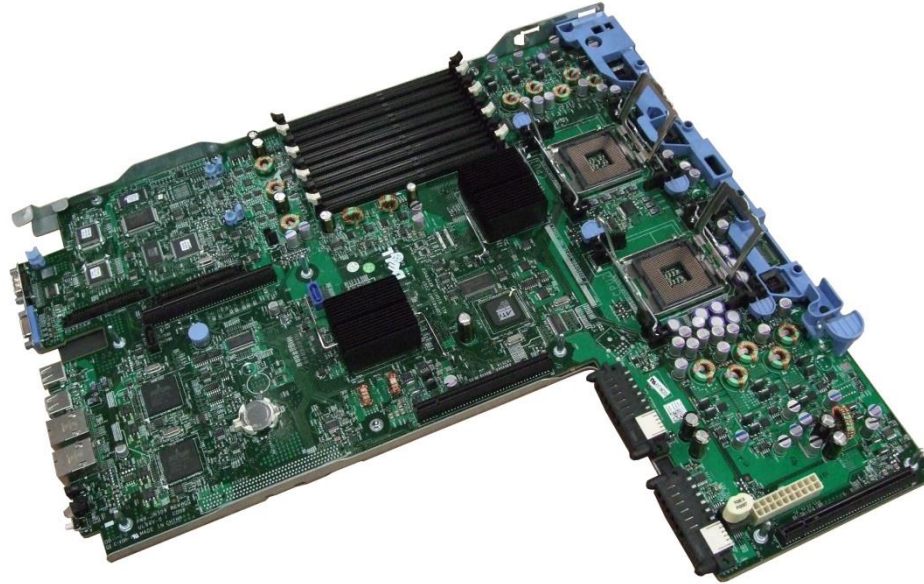


"woman is holding bunch of bananas."



"black cat is sitting on top of suitcase."

The business vision



Results of image analysis:

A Dell PowerEdge 2950 Gen III Server System Mainboard with Tray FC284, visual inspection indicates mint condition, expected failure rate classification: BB service cost calculation (first 12 months): \$1,75 / month

Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics

"This is what a healthy city looks like."

- Christopher Mason, Weill Cornell Medical College

Source: <http://www.wired.com/2015/02/mapping-microbes-new-york-city-subway/>

The study was intended as a baseline analysis of healthy cities.

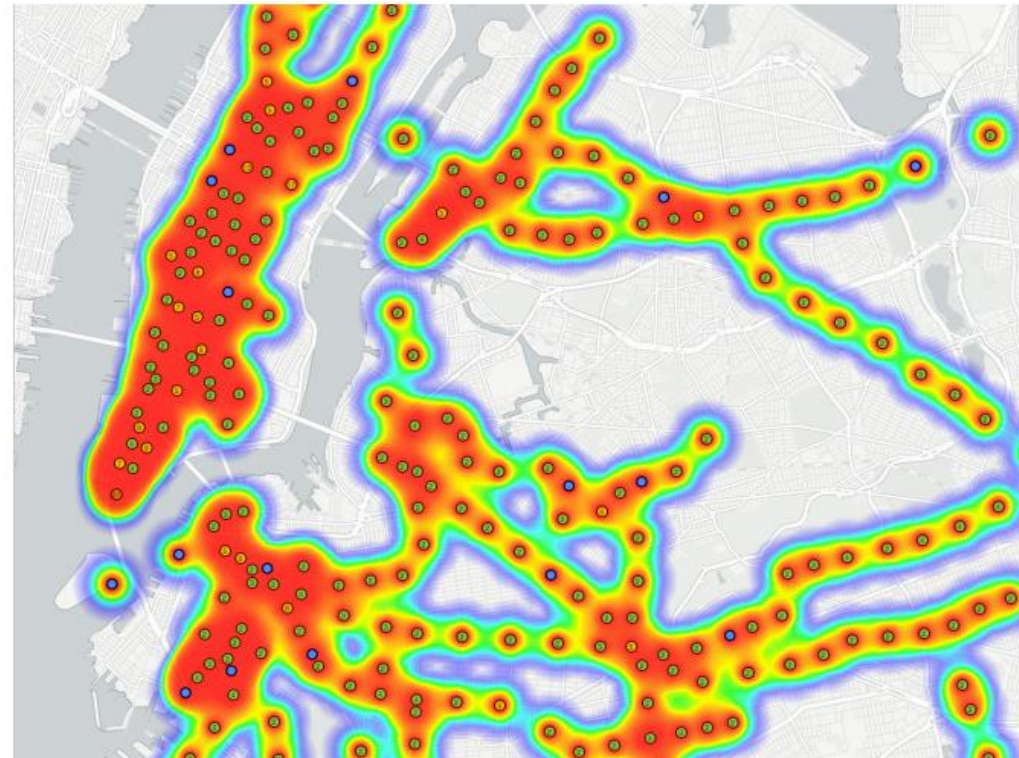
About 48% of the DNA found did not match any known species. The known ones include bubonic plague and anthrax.

Calculating a baseline is incredibly valuable, for your business too.

Mapping the Microbes of the New York City Subway

BY MARCUS WOO 02.06.15 | 2:25 PM | PERMALINK

[Share](#) 185 [Tweet](#) 513 [+1](#) 62 [in Share](#) 41 [Pin it](#)



PathoMap

Healthy server racks

Results of comparison to baseline:

Mean Temp / Std. Dev.: 73°F / 2.4°F

Mean Hum. / Std. Dev.: 53% / 3%

Maintenance Costs: \$140/month

Servers: 35%

Storage: 40%

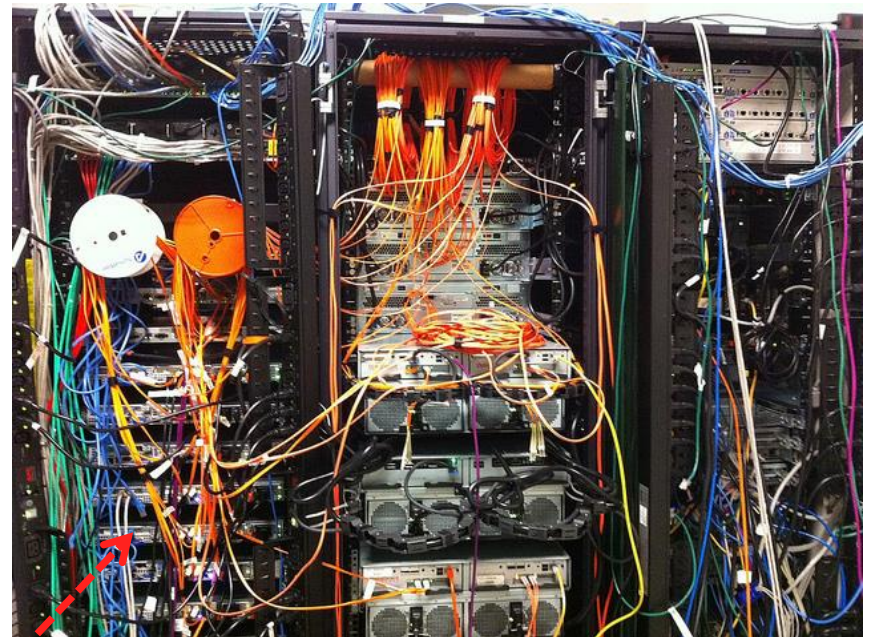
Wiring: 25%

Cooling: 5%

Predicted Future Cost: \$150\$/month

Recommendations:

1. reduce Temp. Std. Dev. to 1.5°F
=> Main. Costs: \$125/month
2. Connect and/or label this cable



New opportunities and threats are emerging now

The strategic situation is shifting

- Big Data is something that is happening all around you:
 - your customers
 - your suppliers
 - your competitors
- Big Data means new business models through:
 - product innovation, e.g. value-added services
 - process innovation, e.g. new channels, cost savings, service customization
- New players will enter your market, and the established players will seek to use the new technologies to their competitive advantage.

Threats

Big Players

- Manufacturers taking a larger share of the service business
- based on data-driven decision-making
- using data-enabled business models

New Entries

- New entries attempting to displace you
- online platforms pushing into all service sectors
- crowd sourcing creating new price competition

... the Still Unknown

- Something none of us have thought of yet!
- A word of warning: 2014 was the year banks got scared of Apple and truck drivers got scared of Google. You are not safe from truly disruptive change through the innovative use of data assets.

Opportunities

Customization

- Personalization and customization is already your strength.
- Big Data lets you refine that strategy (e.g. "mass customization" strategy for SME's).

Natural SME Advantages

- SME's do not have as much IT ballast, thus reducing the implementation lead-times; quicker and more flexible than larger competitors.
- Collaboration with new market entries is easy for you.

Still in Time

- You are still in time, even to beat the big players at the big data game.
- Repeated, independent studies show: those first able to implement data-driven decision-making have a significant competitive advantage.

How will you respond?

- individually
- as a group

Experience from successful projects

1

Respond to this strategic challenge by focusing on business cases, not on vendor ideas of "use cases".

2

To become a data-driven enterprise, you must think about data differently, seeing new kinds of solutions.

3

Develop analytic capability strategically, not as an afterthought.

4

To reap data-driven advantages, learn to manage data as a capital asset: reduce cost, increase competitiveness, optimize processes, provide value for your customers

Imperatives for survival

- **Scale!**

Develop scalable business models based on data-driven decision-making

- **Cooperate!**

Pool resources with customers, suppliers and competitors to create the new competitive environment

- **Innovate!**

If you want to flourish, there is no way around product, process and market innovation

The next, urgent steps (are easy) to take

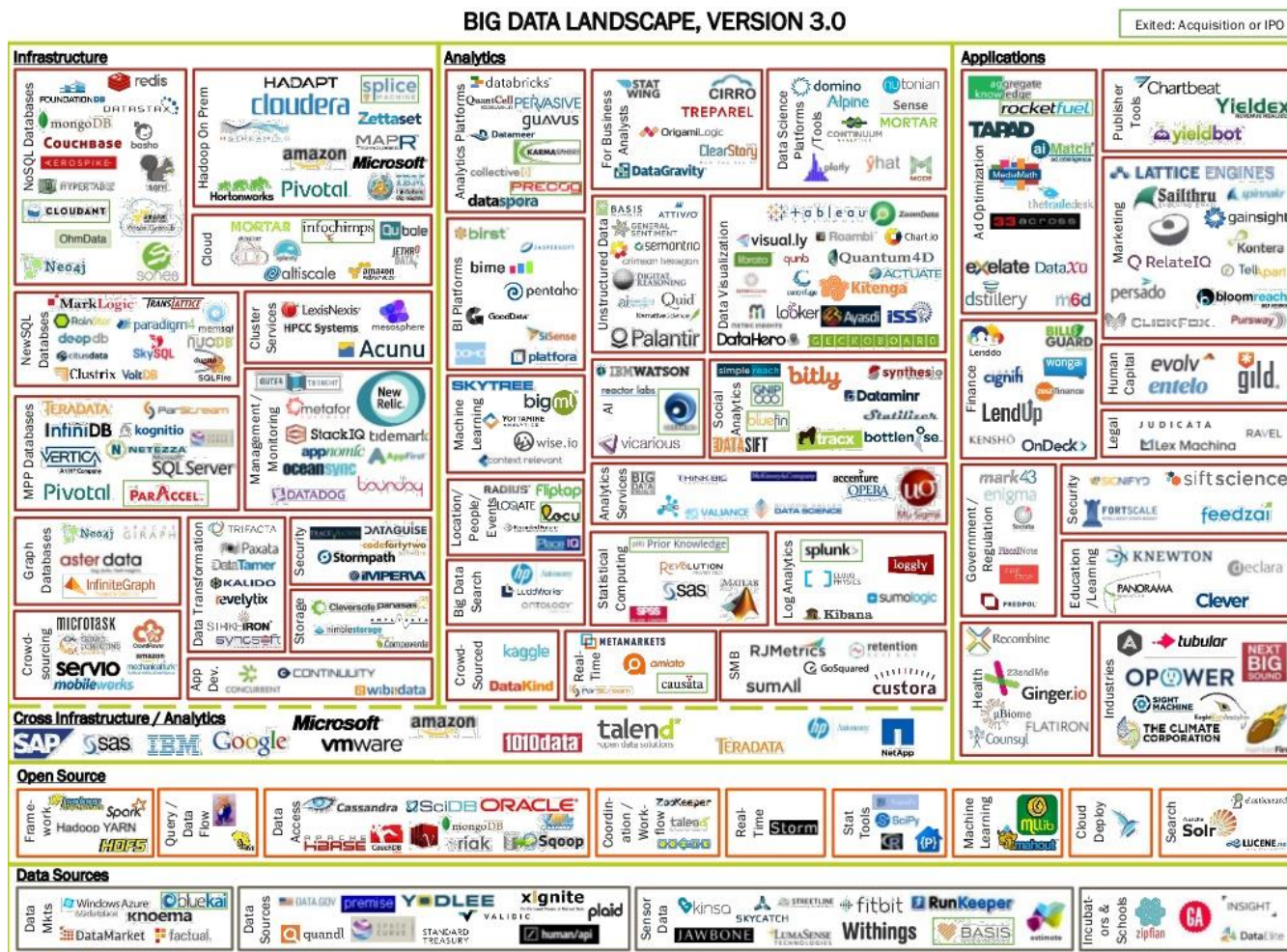
Begin to get value from data assets

- Start small, even tiny. Take baby steps.
- Focus on a small number of business cases.
- Take an iterative, agile approach. Pivot if needed.
- Get advice. Find Data Science expertise.
- Partner with smaller specialists, not the big players.
- Begin to build data assets into your business strategy. This entails a change in thinking – you begin to want to invest in data, to trade data, etc.).
- Think strategically: Big Data platforms and tools are not the same as business solutions or capability. You need to think in terms of business value.

Develop awareness

- Get a basic understanding of what Big Data and Advanced Analytics are
=> a little reading and discussion
- Investigate how these are changing business today, especially the business of your customers, suppliers and competitors
=> talking with your partners, attending events, reading
- Consult with business-oriented Data Scientists about the business impact of new technologies and how they might be applied to your situation.
=> a few days of consulting

Get to know the Big Data vendors



© Matt Turck (@mattturck), Sutan Dong (@sutandong) & FirstMark Capital (@firstmarkcap)

Source: <http://mattturck.com/category/big-data/>

Jumping into Big Data

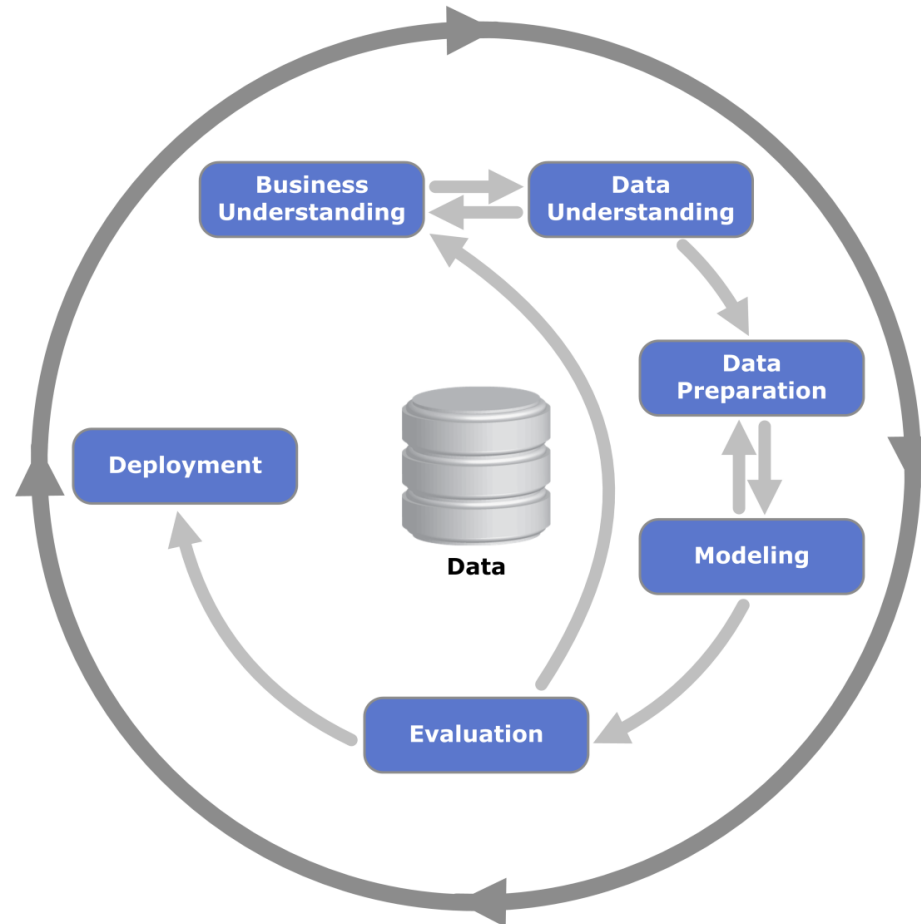
- Choose from the infrastructure options carefully:
 - Do-it-yourself on premise
 - Implement part or all of the infrastructure in the cloud
 - Hire a dedicated service provider
- Be wary of the vendors' strategies
 - Although it's almost free, avoid the "data lake" strategy
 - Although they seem compelling, avoid "use cases"

**Let the business case for analytics drive the
acquisition of data assets**

Starting a project – it might look like this

1. Develop a question or two (business problems) which you think can be answered by a Big Data / Advanced Analytics solution.
2. Will the answer pay for your effort many times over?
If not, go back to step 1.
3. Define a pilot project to answer that question.
Don't worry about the cost. It'll be reasonable and it's worth it (see step 2).
=> You're here after 4 to 8 weeks.
4. Implement the pilot in a timeframe of 8 to 12 weeks.
5. Keep your eyes open for additional ways to achieve value.
6. Prepare an investment decision for an iterative process going forward.

The Data Mining Process



"CRISP-DM Process Diagram" by Kenneth Jensen

http://commons.wikimedia.org/wiki/File:CRISP-DM_Process_Diagram.png#mediaviewer/File:CRISP-DM_Process_Diagram.png

Data Mining requires Data Scientists

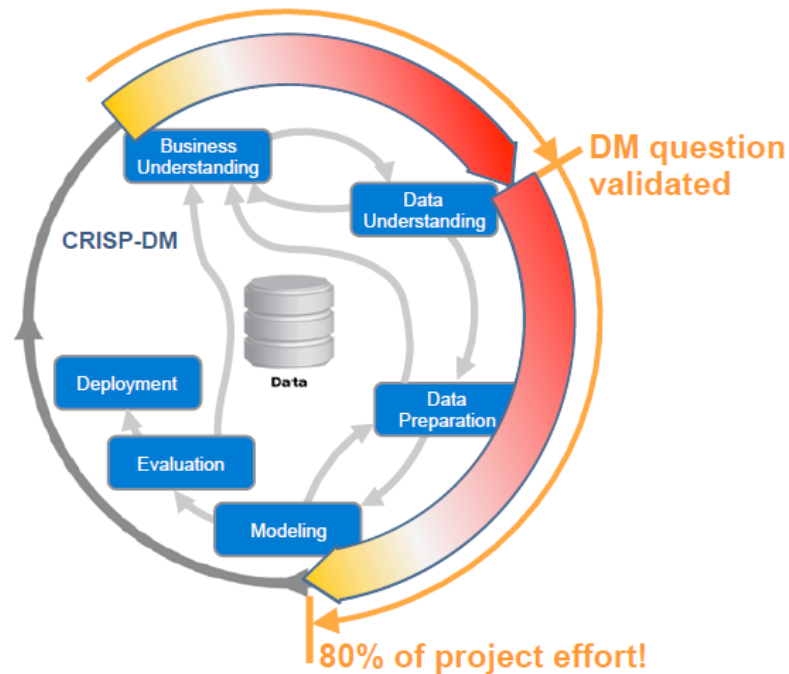
The primary cost driver in a Big Data project is data science expertise:

- designing an appropriate hardware and software architecture
- developing and installing the necessary data pipelines
- applying statistical methods to the business problem
- visualizing the analytic results for the business

Infrastructure and tools are inexpensive, but no cloud in existence can solve your business problems.

Another View

Summary: Big Data Challenges



Technology is not the bottleneck
...it's data availability, quality, and understanding

Need to build up
Data Scientists in business units
to help setting data mining goals,
and help prepare the data

Sometimes, you
don't need so Big Data
but the right data and the right questions

Have a
solid business case
before you invest hundreds of
thousands of Euros into tools!

Source: Dr. Lothar Baum, Bosch Corporate Research, Feb. 2015



Short cut:

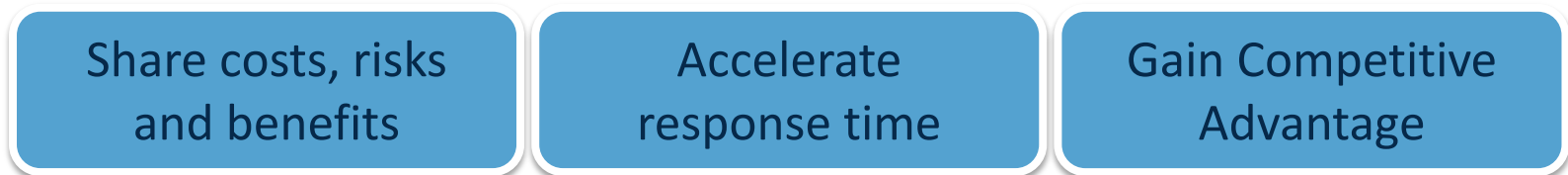
Start with things you can do now

- Apply advanced analytics to the data you already have.
- Add sensors to measure your technical or business processes. Apply advanced analytics to those readings.
- Develop analysis products you can offer your customers as an add-on service:
 - help reduce their costs
 - provide additional, data-driven services
 - develop new business models to better serve them
- Tap into social media data to help inform your business decisions.
- Analyze your customers and develop data-driven sales and marketing campaigns based on the data.
- Significantly reduce BI storage costs with Big Data storage technology.

The key takeaway: Don't be afraid of Big Data!

What you need	Where you get it	What it might cost
Hardware <ul style="list-style-type: none"> - server - memory - storage 	Commodity on premise, or Cloud-based	a few thousand dollars, or Inexpensive (a couple of hundred dollars per month, e.g. \$30/TB storage) to free for a pilot
Software <ul style="list-style-type: none"> - data storage - data access - analytics - visualization 	Open source Cloud-based Proprietary tools	Free Pay-as-you-go, inexpensive (a few hundred dollars per month) to free for a pilot Inexpensive to free for a pilot, may be a significant cost in production
System management	Internal staff, or Cloud-based	Variable depending on expertise, or Inexpensive (included in cloud expense)
Data Science expertise	Hire staff, or Retain consultant	Entry salary: \$88K (mean) to \$130K (top), or 2 to 4 weeks consulting for a pilot
A project plan	Staff, or consultant	Staff expense, or 1 to 2 weeks consulting
A bit of time for a pilot	Things that have a lower ROI	3 to 5 months for 1 to 2 internal staff
Data	Internal, open, commercial	Free (internal, open) or variable (commercial)

Proposal to SIA members: Let's Collaborate on a Pilot



Prediction

In five years, the field service industry will be radically different than it is today.

Do you want to fight the future, join it, or create it?

Quotes

"Data is the new oil of the internet and the new currency of the digital world."

– Meglena Kuneva, European Consumer Commissioner, 2009

"Because in the era of big data, more isn't just more. More is different."

– Wired. 7 Jun 2008, The Petabyte Age:

http://www.wired.com/science/discoveries/magazine/16-07/pb_intro

"Data Scientist: the sexiest job of the 21st century"

– DJ Patil, then LinkedIn Head Data Analyst, now U.S. Chief Data Scientist, 2008

"There is no such thing as a single version of the truth."

– Jeff Jonas, IBM Research Fellow, 2006

"Technology is neither good nor bad; nor is it neutral."

– Melvin Kranzberg, Technology Historian, 1986



Information Professionals GmbH

Klebinger Strasse 7
D-83395 Freilassing

Telephone: +49-8654-776-3424
Cell (Germany): +49-171-680-9398
Cell (U.S.) until 3/24/2015: 415-205-9733
E-Mail: j.thompson@infopro-gmbh.de
Web: www.infopro-gmbh.de

