“If you do not know how to ask the right question, you discover nothing.”

Big Data? Prepare for big trouble!

Managing Data Quality with unstructured data for business benefit and value,
Sponsored by W. Edwards Deming

Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes

Agenda

- The real problem for Big Data - *Context*
- Data Quality challenges for Big Data
- Stages of Data Quality intervention
- Establishing actionable business outcomes and analytic value

A bit about me....

- NEW! Research Director, Gartner
- Advisory board member, QFire Software
- 22 years Information Strategy, Data Governance, Analytics & Business Consulting
  - Director of Data Governance at UNSW
  - EDS, KPMG, CPW, Acuma, Pelion, SMS
  - Scottish Power, United Distillers, O2, Astra Zeneca, Carphone Warehouse, Vodafone, Riyad Bank
  - Commonwealth Bank, NSW Roads & Maritime Services, Centrelink, OATSIH, NSW Family & Community Services, CASA, AMSA, FaHCSIA, DAFF, Navy…
- Information-Management.com “Top 12 on Twitter”
- Best supporting Actor, 2005 Barnet Drama Festival
- See me this week in “Boudicca” at Verulamium…

Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com

“Data! Data! Data! I can’t make bricks without clay!”

(Sherlock Holmes,
The Adventure of the Copper Beeches)

The real problem for “Big Data”: *Context*
Sponsored by Sir Arthur Conan Doyle

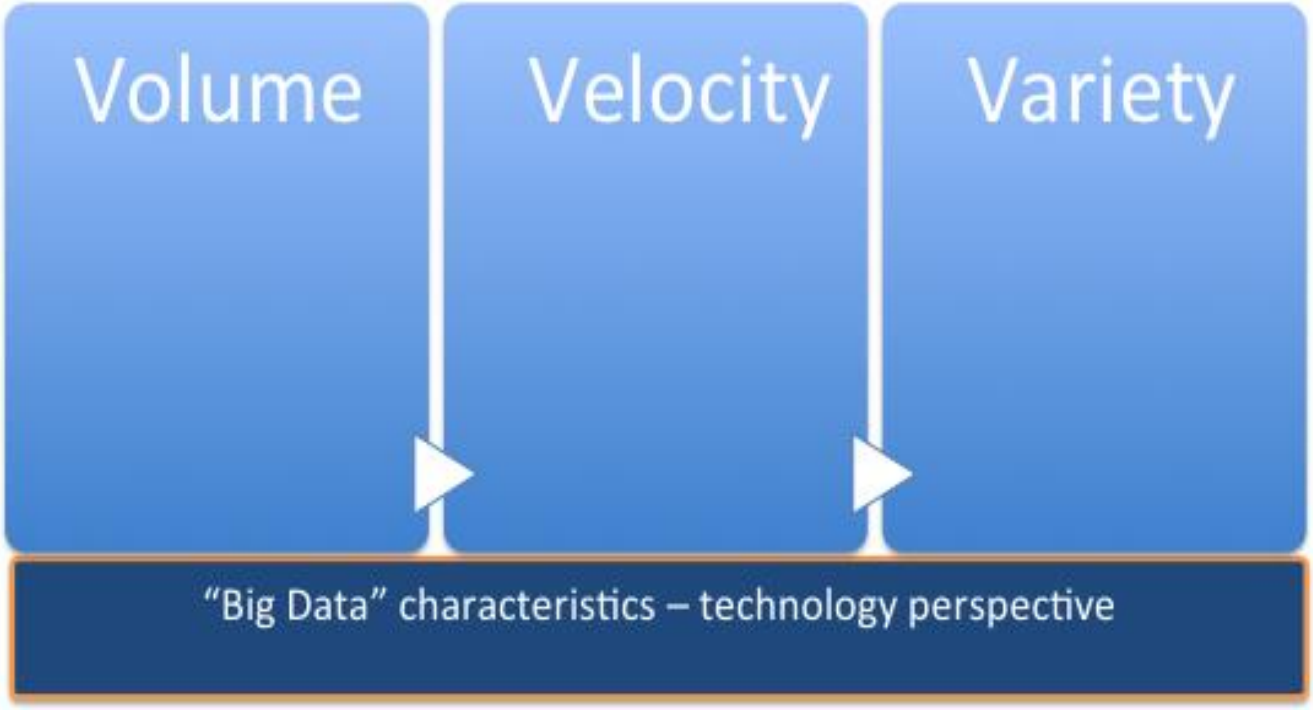
Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes

“Big Data” is already big business

- 42% of business users consider inability to use and analyse unstructured data as a key driver for “Big Data” initiatives.[1]
- 63% of organisations report that the use of information and analytics (including “Big Data”) is creating competitive advantage.[2]
- 90% or IT leaders believe investments into “Big Data” initiatives are worthwhile.[3]
- Successful businesses are 2.7 times more likely than their peers to have a decision-making culture that values the use of supporting data.[4]

- 1 “Big Data Perspectives: Users vs. IT”, Aberdeen Group, April 2014
- 2 “Analytics: The real-world use of big data”, IBM Institute for Business Value and Saïd Business School at the University of Oxford, October 2012
- 3 “IT Industry Trends Survey 2013: Big Data – The Next Frontier”, TEKSystems, 2013.
- 4 “Data Management for BI: Fueling the Analytical Engine with High-Octane Data”, Aberdeen Group, December 2010`

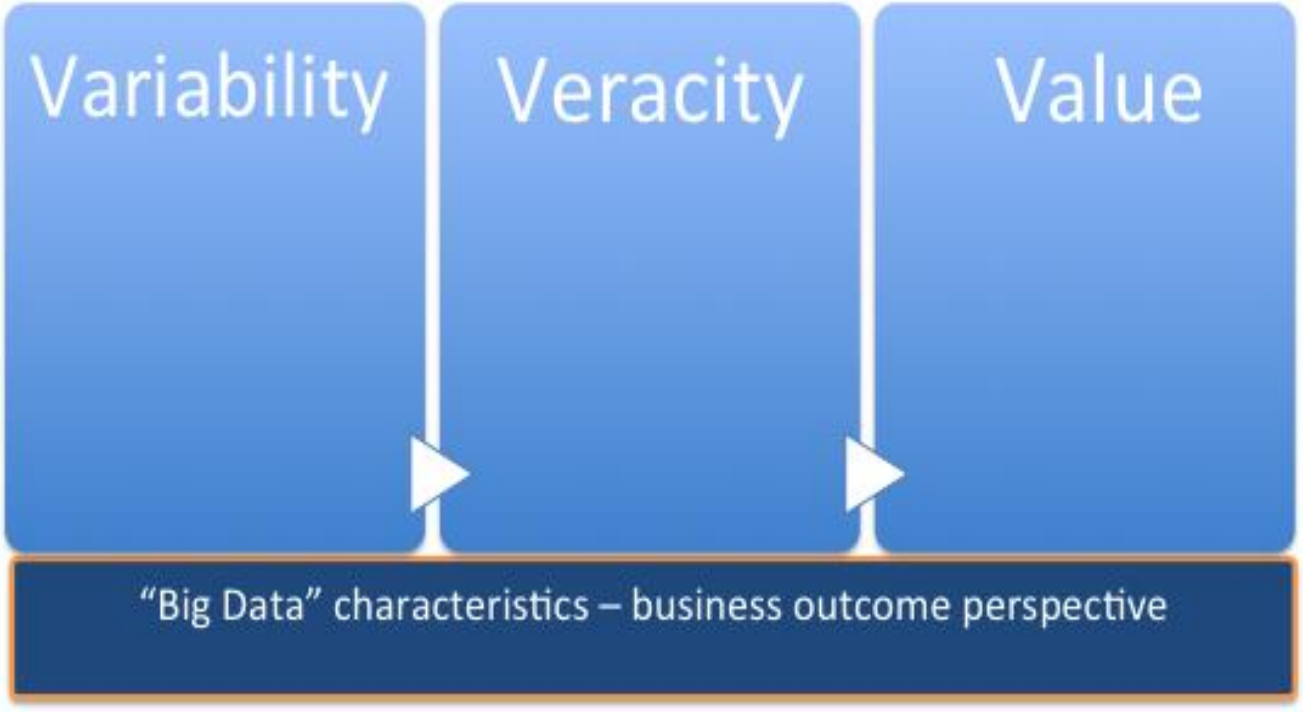
But… the focus is still on technical issues



Where's the "so what?" – there's no business context

Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes

The business context for “Big Data”?



Information > Action > Outcome

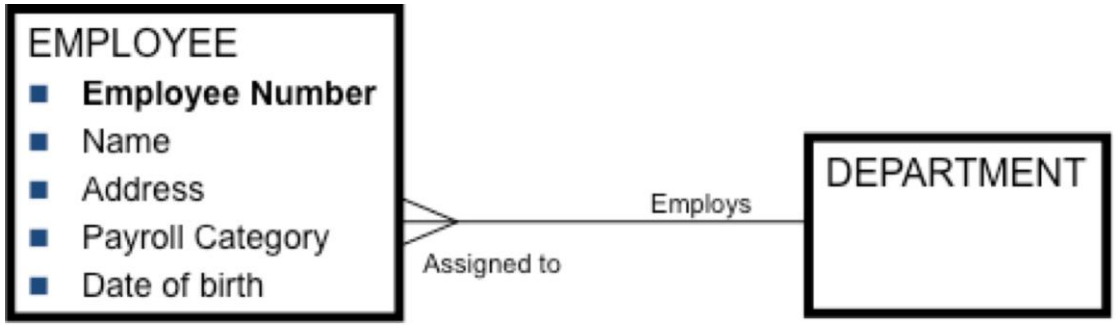
Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes

“Science is organized knowledge. Wisdom is organized life.”

Data Quality Challenges for “Big Data”, sponsored by Immanuel Kant

Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes

Challenge #1: Establishing structure



versus

Teresa 97
Joe 83
Samantha 82
Oliver 67
Linda 55

(Teresa, 97) (Joe, 93) (Samantha, 82) (Oliver, 67) (Linda, 55)
(Teresa, 85) (Joe, 89) (Samantha, 74) (Oliver, 80) (Linda, 47)
(Teresa, 77) (Joe, 93) (Samantha, 66) (Oliver, 72) (Linda, 56)
(Teresa, 94) (Joe, 72) (Samantha, 88) (Oliver, 68) (Linda, 51)

Challenge #2: Semantic Disambiguation

- **Language Detection: (***schadenfreude*, *zeitgeist*, *de rigeur*, *corps de ballet*, *desperado*, *cojones*…)
- **Reference lexicon and synonyms**: (*"Sow": to plant seeds, or a female pig? "Fluke": a fish, a whale's tail or an accidental outcome? "Tail" or "Tale"?)*
- **Standardisation: (***"IBM", "IBM Ltd.", "IBM Software Group", "International Business Machines")*
- **Grammar and Syntax**: (*"The dog bites the boy"* and *"The boy bites the dog"*)
- **Tone**: *("Yeah, right".*)
- **Figurative Language**: *"This computer is running like a dog." "This problem is really dogging me." "It's raining cats and dogs."*

Challenge #3: Semantic Reconciliation

- **Completeness**: The proportion of stored data against the potential of "100% complete"
- **Uniqueness**: No thing will be recorded more than once, based upon how that thing is identified.
- **Timeliness**: The degree to which data represent reality from the required point in time.
- **Validity**: The degree to which data represent reality from the required point in time.
- **Accuracy**: The degree to which data correctly describes the "real world" object or event being described.
- **Consistency**: The absence of difference, when comparing two or more representations of a thing against a definition.

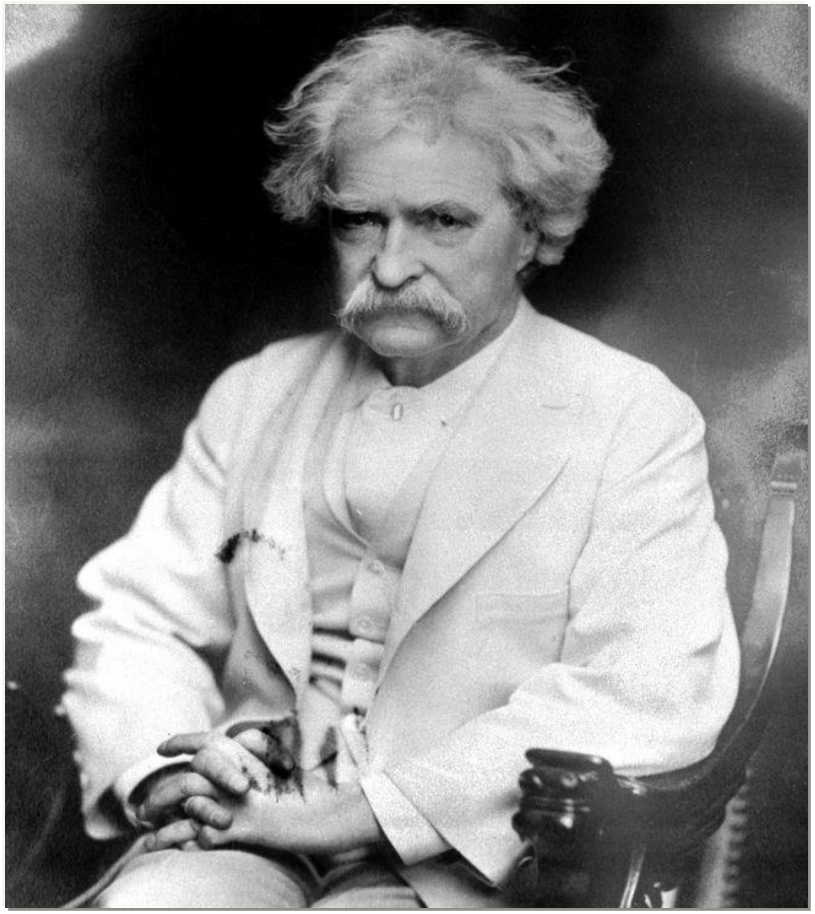
- Plus accessibility, flexibility, performance, retention, ownership, usage…

“Big Data” Data Quality Example – Epidemiology & Patient Diagnostics

- Consolidate patient care records from five years of care history
- identify predictive diagnostic indicators as a supporting tool for medical practitioners

- One data element that appears is the abbreviation value “AH”.
- Grouping all “AH” terms together might seem reasonable? But…

- “Popular” usages include: Auditory Hallucinations (50%), Abdominal Hysterectomy (11%), Acute Hepatitis (11%), Amenorrhea and Hirsutism, (11%), Arterial hypertension (6%), Astigmatic Hypermetropia (6%) and Axillary Hair (6%).
- Other valid values: Accidental Hypothermia, Adrenal Hypoplasia, Airway Hyper-responsiveness, Androgenic Hormone, Anterior Hypothalamus (amongst others).
- http://www.medilexicon.com/medicalabbreviations.php?keywords=AH&search=abbreviation as at August 2014



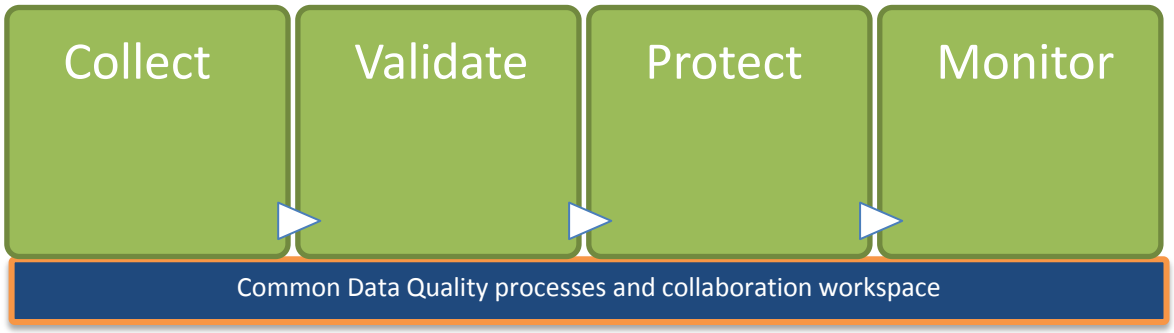
“Get your facts first, then you can distort them as you please.”

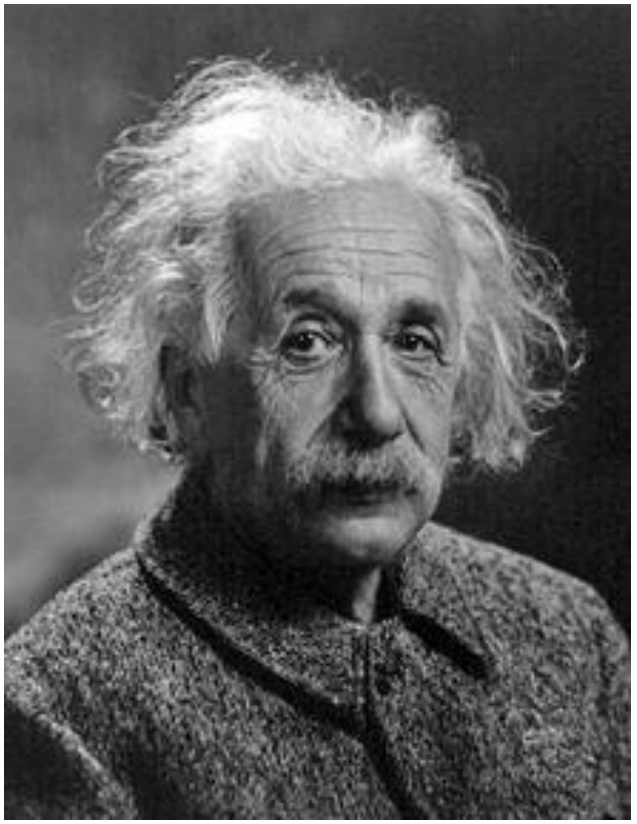
Stages of Data Quality Intervention, sponsored by Mark Twain

Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes

Stages of DQ intervention

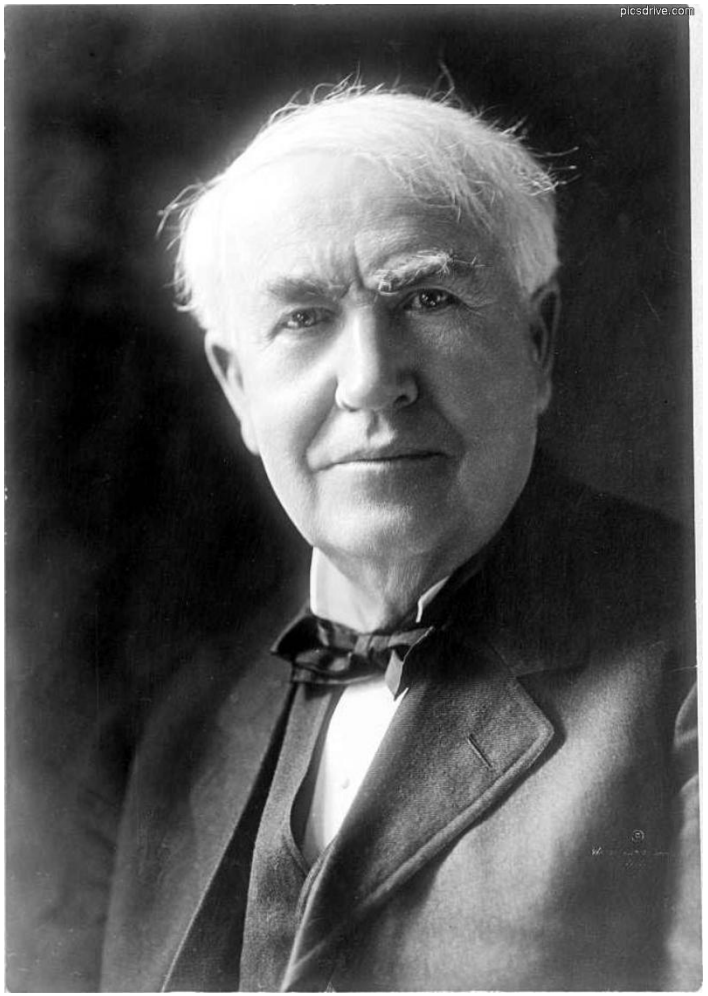
- Data Quality profiling and root-cause analysis
- Identify issues & patterns
- Initiation activity as part of all data warehouse, master data & application migration projects





"Whoever is careless with the truth in small matters cannot be trusted with important matters."

Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes



“The value of an idea lies in the using of it.”

Actionable Business Outcomes,
Sponsored by Thomas Edison

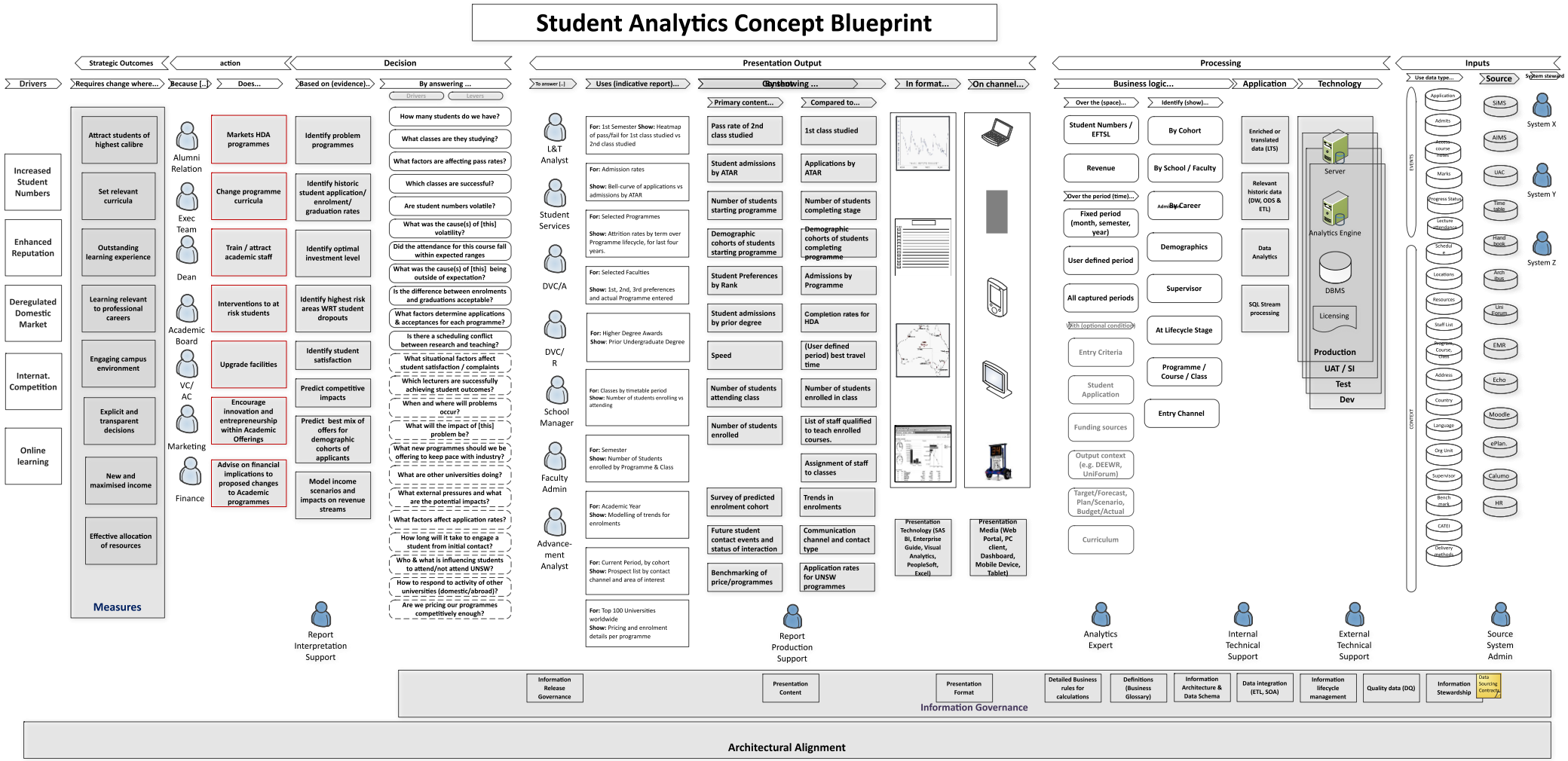
Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes

Five Whiskies in a Hotel

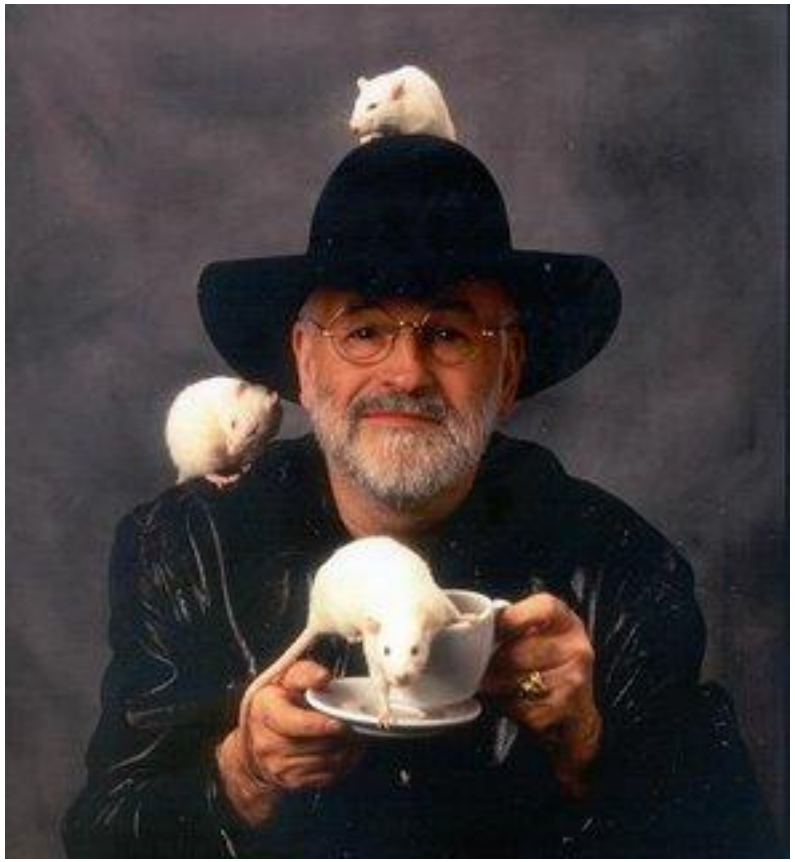
- **What** data/information inputs? What outputs? What tests/measures/criteria will be applied to confirm whether the data is fit for purpose or not?
- **Why** do you want it? (What outcomes do you hope to achieve? Does the data being requested actually support those questions & outcomes? Consider Efficiency/Effectiveness/Risk Mitigation drivers for benefit.
- **Who** is involved? For whom is the information? Who has rights to see the data? Who is it being provided by? Who is ultimately accountable for the data - both contents and definitions? Consider multiple stakeholder groups?
- **When** is the information required? When is it first required? Frequency? Trigger events?
- **Where** is the data to reside? Where is it originating from? Where is it going to?
- **How** will it be shared? How will the mechanisms/methods work to collect/collate/integrate/store/disseminate/access/archive the data? How should it be structured & formatted?

Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com

Enabling method: Information Lifecycle Map



Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes



“It’s still magic even if you know how it’s done.”

Conclusions & final thoughts,
Sponsored by Terry Pratchett

Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes

Pay a visit to Whitefellah Burrows…

- 34% of organisations derive greater business value from "Big Data" once unstructured data types are addressed.[1]
- Unstructured data represents 90% of all real-time data being created today.[2]
- 59% of organisations say data quality problems are the biggest barrier to successful analytics initiatives.[3]
- 80% of the effort involved in dealing with data is cleaning it.[3]

- 1 "What Works in Big Data", The Data Warehouse Institute, 2014.
- 2 "2011 CMO Study", IBM Institute of Business Value, 2011.
- 3 "2014 Analytics, BI and Information Management Survey", Information Week, November 2013.
- 4 "Planning for Big Data: A CIO's Handbook", O'Reilly Media, 2012.

Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes

…and don't fall in!

- Know what you're looking for (and why)
- Dig in the right areas
- Sift through the mullock
- Be prepared for some hard digging
- Be careful or you might get hurt!

Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes

Intellectual curiosity

Skeptical scrutiny

Critical thinking

http://www.informationaction.blogspot.com.au/

@Alan_D_Duncan

http://www.linkedin.com/in/alandduncan

Alan D. Duncan http://nformationaction.blogspot.com **Tw:** @Alan_D_Duncan E: alan.duncan@gartner.com
Information Strategy | Data Governance | Analytics | Better Business Outcomes