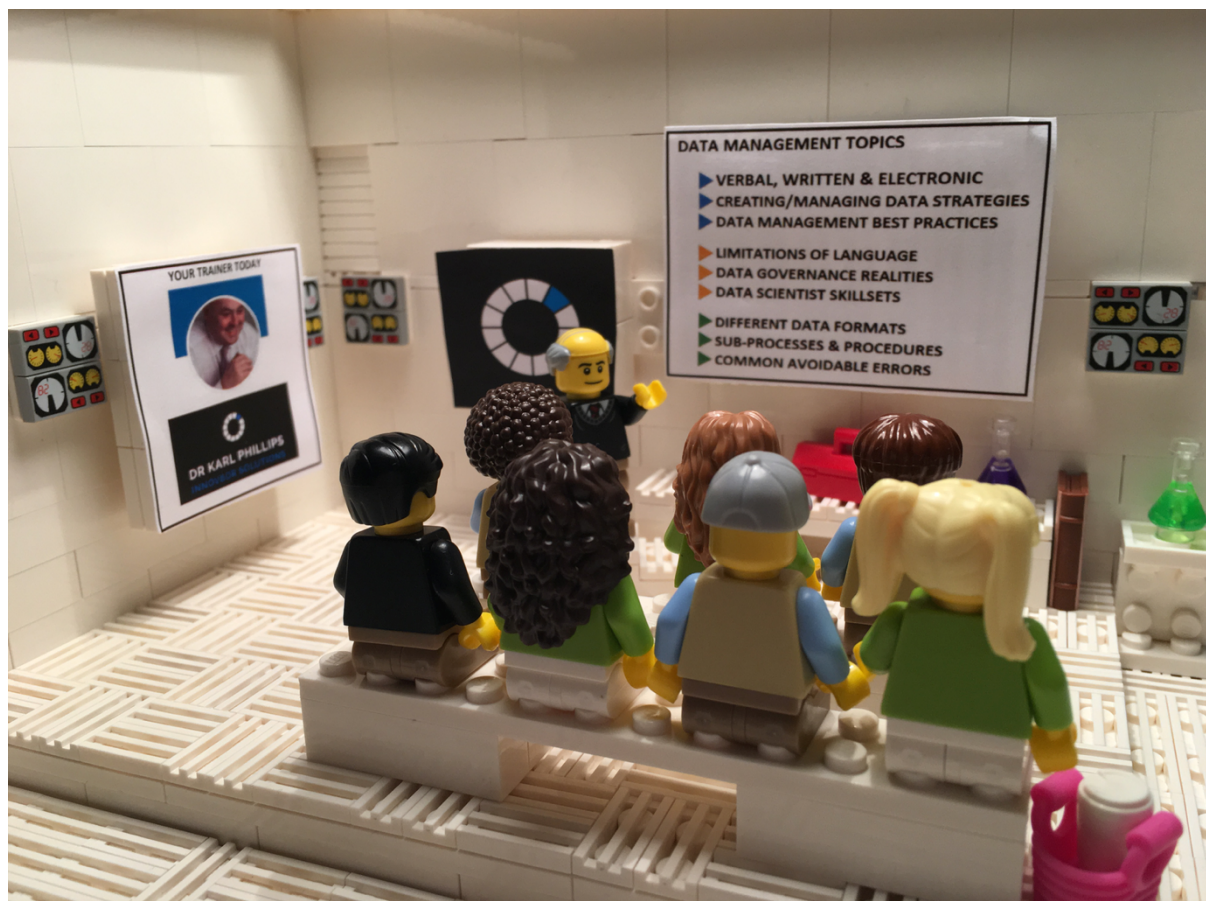
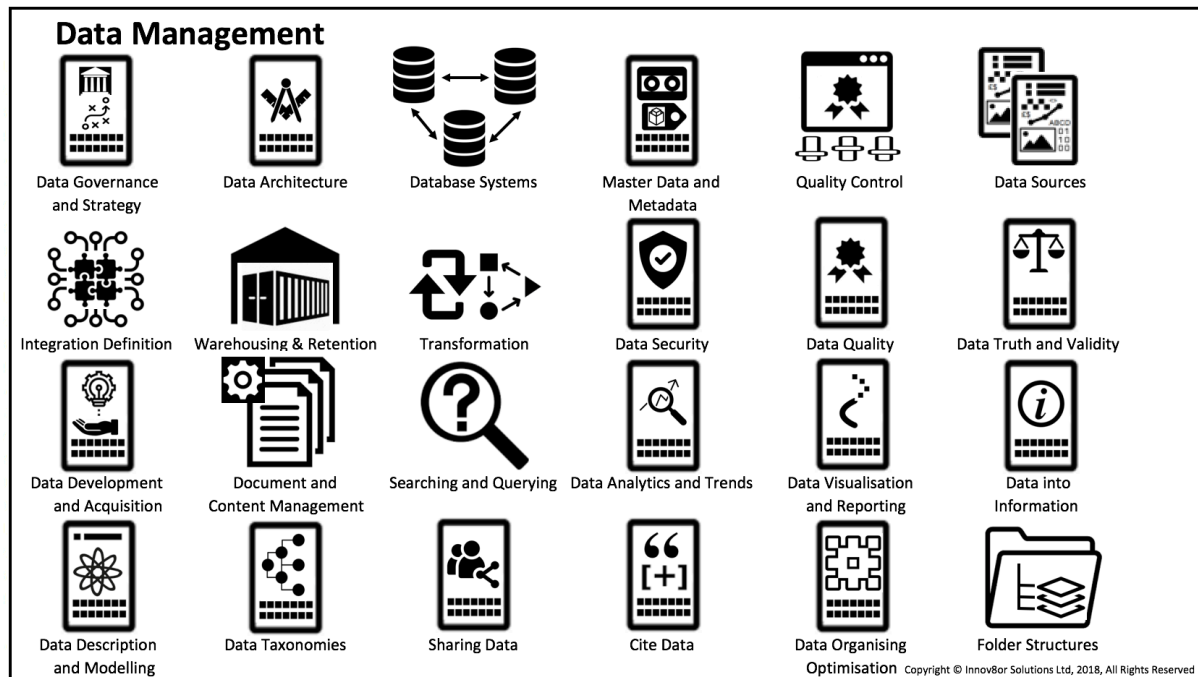


# Overview of selected Data Management Topics

Dr Karl Phillips

© Innov8or Solutions Ltd. All Rights Reserved



## **DATA MANAGEMENT TOPICS**

-  **VERBAL, WRITTEN & ELECTRONIC**
-  **CREATING/MANAGING DATA STRATEGIES**
-  **DATA MANAGEMENT BEST PRACTICES**
  
-  **LIMITATIONS OF LANGUAGE**
-  **DATA GOVERNANCE REALITIES**
-  **DATA SCIENTIST SKILLSETS**
  
-  **DIFFERENT DATA FORMATS**
-  **SUB-PROCESSES & PROCEDURES**
-  **COMMON AVOIDABLE ERRORS**

## VERBAL, WRITTEN & ELECTRONIC

| WHY?                                                                                                                                                                                                                                                                                                   | Our selection of words is critical because few use or hold the exact same meaning because each word evokes different images, memories & meaning for every individual<br>Introductory level that allows for initial engagement<br>Often superficial in nature but extremely useful starter                                                                                                                                                                       |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Level 1<br>CONTEXT                                                                                                                                                                                                                                                                                     | Used to evaluate people and function amongst strangers<br>Watch for vested interests, positioning and politics<br>Sometimes called the 'Social' Level or Polite Conversation                                                                                                                                                                                                                                                                                    |
| Level 2<br>CONTEXT                                                                                                                                                                                                                                                                                     | Organisational, Suppliers & Customers Jargon<br>Glossary of Reference Terminological Terms<br>Sector, Business and Locally Recognised Acronyms<br>Approved Standards, Methodologies & Processes                                                                                                                                                                                                                                                                 |
| Level 3<br>CONTEXT                                                                                                                                                                                                                                                                                     | Organisational Structure & Interactions<br>Data Field & Metadata Definitions (Data, Facts, Ideas, Tips)<br>Taxonomy Frameworks (Information, Strategies, Plans)<br>Classification, Classes & Grouping Systems (Processes)<br>Object Hierarchical Relationships (Risks, Dependencies)<br>Data File Definitions (Artefacts, Data Hosting, Interfaces)                                                                                                             |
| Level 4<br>CONTEXT                                                                                                                                                                                                                                                                                     | Dictionary of Configuration/Document Management<br>Dictionary of Operating Systems/Programming Languages<br>Dictionary of Architecture, Systems, Tools & Processes<br>Baseline and Change Management Systems                                                                                                                                                                                                                                                    |
| Level 5<br>CONTEXT                                                                                                                                                                                                                                                                                     | Technical & Requirements Specifications ('Mental' Level)<br>Logic, Reasoning & Flow with the ease of following it<br>Clarity, Conciseness & Understanding of the message<br>Reader Needs, Alignment, Connectivity & Receptivity<br>The use insights such as Visuals, Examples & Studies<br>Reader's State of Mind/Emotions & Trust are key here<br>Interactions that provide feedback and get a response<br>Opportunities to discuss, collaborate and challenge |
| Level 6<br>CONTEXT                                                                                                                                                                                                                                                                                     | Development of new learning and innovative thinking<br>Centered on a state of Mastery to aid improvement<br>Creation of a unique level of mindfulness/wellbeing                                                                                                                                                                                                                                                                                                 |
| Level 7<br>CONTEXT                                                                                                                                                                                                                                                                                     | Very deep level of 'Spiritual' Connected Communication<br>Difficult to capture/acknowledge in science and words<br>It is a very personal level that very few people achieve<br>Highest level of resonance/intensity with no distortion<br>Likened to a direct energetic embrace touching the soul                                                                                                                                                               |
| <div>&gt; The best communicators understand each of the different levels<br/>&gt; They are able to guage which level to operate at based on factors<br/>&gt; Communication invariable requires transitions between levels<br/>&gt; More concentration and energy are needed at the higher levels</div> |                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |

## CREATING/MANAGING DATA STRATEGIES

| WHY?        | <div>Organising<br/>Governing<br/>Analysing<br/>Deploying</div> An Organisation's Information Assets and Protect and Leverage their Data, by clarifying the primary purpose of their data and guiding them in strategic offensive and defensive data management roadmap/end-point                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|-------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Definitions | <div>&gt; A strategy isn't a once off effort, it is a set of long-term objectives/goals delivered through a set of measured/reviewed multi-year/short-term goals that are centrally coordinated activities and aligned with one another, to deliver the best possible solution as the organisation grows and evolves<br/>&gt; Data are facts or statistics collected together for reference or analysis, and it should be complete, accurate, consistent, clean, current, unique &amp; valid<br/>&gt; Data Architecture describes how data is collected, stored, transformed, distributed and consumed, plus rules governing structured formats and systems for connecting data with the business processes<br/>&gt; Information is data endowed with relevance, context and purpose<br/>&gt; Information Architecture governs the processes and rules that convert data into useful information through integration and analysis tools</div> <b>Classic Enterprise Data Management for Effective Data Usage</b> <div>&gt; Technical Focused &gt; Emphasis on Governance, Best Practices &amp; Tools<br/>&gt; Doesn't address the fact that data issues not limited to a single app/org</div> <b>Offensive Data Management Tactics (Flexibility of Use)</b> <div>&gt; Minimising downside risk &gt; Ensuring Compliance &gt; Detect Fraud/Theft<br/>&gt; Support Business objectives of Revenue, Profitability &amp; Satisfaction<br/>&gt; Generate Insights to support Management Decisions/Dashboards</div> <b>Defensive Data Management Tactics (Standardisation &amp; Control)</b> <div>&gt; Identifying, Standardising &amp; Governing Data Sources and Integrity Flows<br/>&gt; Concentration on legal, financial, compliance and E2E IT concerns<br/>&gt; External pressures or environmental factors can also have an influence</div> <b>Optimal Data Management</b> <div><small>S5OT: Single Source of Truths, M5OT: Multiple Versions of the Truth</small><br/>&gt; Business Focused &gt; Trade-offs between Offensive &amp; Defensive Tactics<br/>&gt; Include all Stakeholders &gt; Technology is only an Enabler/Supports Use Cases<br/>&gt; Exploit Patterns &amp; Reuse &gt; Prioritise Activities &amp; Revisit Regularly<br/>&gt; Efficient Use of Data &gt; Allocation of Resources for Data Management<br/>&gt; Clear Roles and Responsibilities &gt; Cover off Ethics, Privacy &amp; Preservation<br/>&gt; Master Data Management (MDM) uses of both versions of the truth:<br/>&gt; S5OT working at the data level; M5OTs supporting the mgmt of information</div> <b>Data Strategies Components</b> <div><small>S5L: Structure, Origin, Location, S4L: Structure, Accessibility, Location</small><br/>&gt; Identify: Your data and understand the meaning regardless of SOL<br/>&gt; Store: Persist in a S4L that supports easy, shared access, IPR, processing<br/>&gt; Provision: Package data for reuse &amp; sharing, and provide rules/guidelines<br/>&gt; Process: Move/Combine disparate systems to unified consistent data view<br/>&gt; Govern: Establish, Manage, Communicate Info Policies &amp; Mechanisms</div> |

## DATA MANAGEMENT BEST PRACTICES

| WHY?                                     | <div>Reduce Data Preparation Time<br/>Eliminates Incomplete Data<br/>Right Data directs Right Results<br/>Facilitates New Data Formats</div> The development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle enterprise needs                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Defining Your Data is the key first step | <div>&gt; Publish and enforce recommended data formats and ones to avoid/proprietary<br/>&gt; Establish file naming conventions for consistent, descriptive &amp; informative data<br/>&gt; Backup data on local &amp; remote external servers with restoring contingency plans<br/>&gt; Publish an established metadata standard/schema appropriate to your field<br/>&gt; Ensure all data is labelled with the right up-to-date security classification<br/>&gt; All sensitive information should be redacted before being further distributed<br/>&gt; Develop and implement a data retention policy and data migration policy<br/>&gt; Be aware of ownership and intellectual property rights concerning data<br/>&gt; All archived data sets must include the metadata in a 'readme' type file<br/>&gt; Develop a data roadmap showing start/end cycles for data related systems</div> <b>Preparing your data ready for analytics - Why do it?</b> <div>&gt; Simplify access to traditional and emerging data - more data = better outcomes<br/>&gt; Strengthen the advanced analytic techniques that your data scientists need<br/>&gt; Scrub data to build quality into existing processes by incorporating data cleaning<br/>&gt; Shape data using flexible manipulation and transformation techniques<br/>&gt; Share metadata across data management and analytical domains</div> <b>Data Protection whilst Maintaining Privacy and GDPR Conformance</b> <div>&gt; Handle sensitive data in the appropriate manner in order to stop future problems<br/>&gt; Data Protection is a technical issue of protecting data against unauthorised access<br/>&gt; Data Privacy is a legal process of authorised access - who has it and who defines it<br/>&gt; Ensuring data protection may not adhere to all required compliance standards<br/>&gt; No technical safeguards can eliminate the central role of trust in euring privacy<br/>&gt; End-to-end encryption is as far as technical services go in data privacy/protection<br/>&gt; General Data Protection Regulation (GDPR) addressing data export outside the EU&gt;</div> <b>Data Facts that people often forget about or don't address</b> <div>&gt; Choose file formats that will ensure long-term access (open source over proprietary)<br/>&gt; Track different versions of data/documents using a coporate version convention<br/>&gt; "Metadata is a love note to the future" so create it for everything generated<br/>&gt; There are lots of helpful tools &amp; techniques for analysing your data, catalogue them<br/>&gt; Create &amp; implement a plan for how to transfer knowledge when data changes hands<br/>&gt; Prevent valuable information from getting lost - record passwords when people leave<br/>&gt; You can't rely on databases (who controls it, manages it, patches it, no longer usable)<br/>&gt; All hardware is in different states of collapse, you may just not know which one<br/>&gt; Need an exportable and importable method (baselines, human readable output)<br/>&gt; Data Migrations need a data catalogue - data is easy, rules are the hard things to move<br/>&gt; Don't fill your fields more than 70%, make room to allow growth/modifications<br/>&gt; Backups only good for recovery purposes and not for audit purposes</div> |



## LIMITATIONS OF LANGUAGE

|                                                                                                 |                                                                                                                                                                       |
|-------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>WHY?</b>                                                                                     | The limitation of your language means the limit of your world, because if you can't dream, think, talk about it or describe it, then you must pass over it in silence |
| <b>Language Basics</b>                                                                          |                                                                                                                                                                       |
| > Multiple languages allows people to communicate and to work together with more people         |                                                                                                                                                                       |
| > Different languages provide a vehicle to open up more opportunities/world perspectives        |                                                                                                                                                                       |
| > Openness to other languages reduces ignorance, fear and stops narrowing your views            |                                                                                                                                                                       |
| > The boundaries on life are drawn by the words we use, so test the edges; they can shrink      |                                                                                                                                                                       |
| > Over time as people evolve, they come up with new words, phrases, slang & writing styles      |                                                                                                                                                                       |
| > Theories and Ideas have also evolved in a similar way, as old theories/ideas get parked       |                                                                                                                                                                       |
| > A person's actions mean so much more, however, there are times when we rely on words          |                                                                                                                                                                       |
| > Non-pictorial languages are harder to understand, they don't help the reader find meaning     |                                                                                                                                                                       |
| > When languages fail, organisations come up with local acronyms and terms to fill the gap      |                                                                                                                                                                       |
| > Languages by their nature reduces complex realities to simple sounds and concepts             |                                                                                                                                                                       |
| > Translations services are costly, hard to obtain and prone to error/misrepresentation         |                                                                                                                                                                       |
| > Multilingual organisations promote members learning more than one language                    |                                                                                                                                                                       |
| Advice = keep your words soft and sweet lest you have to eat them at a later date               |                                                                                                                                                                       |
| <b>The Structure of Data</b>                                                                    |                                                                                                                                                                       |
| > Structured Data (SD) has a high degree of organisation, readily searchable by search engines  |                                                                                                                                                                       |
| > Unstructured data (UD) is far more prevalent in electronic documents is the opposite          |                                                                                                                                                                       |
| Unstructured Data (UD) might account for 70-80% of all data in an organisation                  |                                                                                                                                                                       |
| > Data Models are constructed in order to organise data into pre-defined manners                |                                                                                                                                                                       |
| Search engines have become popular tools for indexing and searching through data                |                                                                                                                                                                       |
| <b>The Direction of Prevailing Travel within the Industry</b>                                   |                                                                                                                                                                       |
| > Business Intelligence & Data Warehousing Tools have been added to support data analysis       |                                                                                                                                                                       |
| > Human Interaction is a fundamental piece of the end-to-end management of UD systems           |                                                                                                                                                                       |
| > You need to understand how data is generated and exploited to manage UD effectively           |                                                                                                                                                                       |
| > System Blueprints and Templates can help turn UD into Semi-Structured Data (SSD)              |                                                                                                                                                                       |
| > A Business Target State and end-to-end Data Management Model complete the design              |                                                                                                                                                                       |
| > Working on models, systems and technology will not be good enough for success                 |                                                                                                                                                                       |
| > To complete the picture you will need business buy-in, financial support, trained people      |                                                                                                                                                                       |
| > Business buy-in comes with the governance, management, roles & resources, incentives          |                                                                                                                                                                       |
| > The training of people with the skills to create data that can be usefully analysed is key    |                                                                                                                                                                       |
| <b>Making Sense of your Structured &amp; Unstructured Data</b>                                  |                                                                                                                                                                       |
| > Determine your Business Objectives and identify the Sources of all your data                  |                                                                                                                                                                       |
| > Remove Data Silos, address Cultural Barriers limiting data sharing, add metadata/context      |                                                                                                                                                                       |
| > Improve Operational Data Practices, Train Data Creators, Empower the Data Analysts            |                                                                                                                                                                       |
| > Downselect some Analytics Tools then measure and track the ROI of them, change if needed      |                                                                                                                                                                       |
| > Deriving useful intelligence comes from synthesising, delivering results, visualising outputs |                                                                                                                                                                       |
| > Insights from valuable data stores are only useful if they are acted upon, so use wisely      |                                                                                                                                                                       |
| > Find out what works then encourage it/define schemas; what doesn't work improve on            |                                                                                                                                                                       |

## DATA GOVERNANCE REALITIES

|                                                                                                    |                                                                                                                                                                    |
|----------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>WHY?</b>                                                                                        | Availability, Usability, Consistency<br>Data Integrity/Security, Accountability<br>Foundation for better quality, privacy, security and Business Intelligence (BI) |
| <b>The 10 Data Truths</b>                                                                          |                                                                                                                                                                    |
| > The reality is that most data holds no value, yet consumes valuable time & resources             |                                                                                                                                                                    |
| > The average data definition or search doesn't answer all the users' questions or needs           |                                                                                                                                                                    |
| > Yes, there is a wide world of data platforms and source types beyond your internal data          |                                                                                                                                                                    |
| > Data-backed decision making has been proven to be more effective than intuition or habit         |                                                                                                                                                                    |
| > Intelligence from data is only as useful as your empowerment, ability & willingness to use it    |                                                                                                                                                                    |
| > Information provides you freedom, choices & responsibility to change direction mid-course        |                                                                                                                                                                    |
| > Data never sleeps and operates 24/7, so stay vigilant and integrate into every-day processes     |                                                                                                                                                                    |
| > If you can't measure it, you can't manage it, plus excessive data trust (garbage in, gospel out) |                                                                                                                                                                    |
| > People lie, apply their own perspectives on data, make things up/assume & don't report           |                                                                                                                                                                    |
| > Big Data is not the truth; it is a human tool subject to miscollection, misapplication & abuse   |                                                                                                                                                                    |
| <b>Filling the gaps left by the average data definition</b>                                        |                                                                                                                                                                    |
| > A one-liner: A concise response to a question similar to an elevator pitch                       |                                                                                                                                                                    |
| > Metadata & Key Words from the approved Organisational Dictionary                                 |                                                                                                                                                                    |
| > A Real-Life Example: Enhance understanding by adding a relatively easy example                   |                                                                                                                                                                    |
| > List of all possible Valid Values (Enumeration List) plus new, unknown, blanks & nulls           |                                                                                                                                                                    |
| > Rules & logic on how data element were derived to group, categorise, transform data              |                                                                                                                                                                    |
| > Full ancestry of a data element: Source-to-target, lifecycle, relationships, dependents          |                                                                                                                                                                    |
| <b>Data Governance Best Practices -&gt; People, Processes &amp; Technology (in that order)</b>     |                                                                                                                                                                    |
| > Focus on the Operating Model and defining enterprise roles and responsibilities                  |                                                                                                                                                                    |
| Depending on the organisation it could be centralised, decentralised or federated                  |                                                                                                                                                                    |
| Create Roles/Responsibilities/Rules, Ensure an accountability infrastructure is in place           |                                                                                                                                                                    |
| Business/Technical Stewards (Data Owners) & Information Technology (Support IT)                    |                                                                                                                                                                    |
| Corporate Data Council (Decision Making) & Chief Data Officer (Strategy/Governance)                |                                                                                                                                                                    |
| Key Stakeholders (Ownership), Datasets (Approved/Local Working Copies/Owners)                      |                                                                                                                                                                    |
| > Make sure you have a list/explanation of all the business critical data related artefacts        |                                                                                                                                                                    |
| Business Glossaries/Data Dictionaries/Data Catalogues/Business Processes/                          |                                                                                                                                                                    |
| Data Quality Scorecards/Systems and Applications (plus I/O)/Policies & Standards                   |                                                                                                                                                                    |
| > Identify critical data elements that deliver outcomes within the data domains                    |                                                                                                                                                                    |
| Improve & Ensure Information Asset Integrity, Convert to a Master Based-Based Culture              |                                                                                                                                                                    |
| > Define control measurements to help sustain governance and deliver solutions                     |                                                                                                                                                                    |
| Automated workflows, thresholds for approvals/escalations, reviewing, voting                       |                                                                                                                                                                    |
| <b>Apply workflow processes, develop reporting, feedback mechanism; iterative change</b>           |                                                                                                                                                                    |
| <b>Data related things that will come back and haunt you if you don't deal with them urgently</b>  |                                                                                                                                                                    |
| > Legacy Data, Legacy or Proprietary Applications; Unclean Data; Data Migration Strategies         |                                                                                                                                                                    |
| > Shadow IT; BYOD; Bring Your Own IT (Applications to Data Centres); IT Obsolescence               |                                                                                                                                                                    |
| > Cloud Hosting vs Private Clouds; Enterprise Critical Solutions vs Scalable Microservices         |                                                                                                                                                                    |
| > Service Level Agreements (if you get them wrong at the start); Unauthorised Modifications        |                                                                                                                                                                    |

## DATA SCIENTIST SKILLSETS

|                                                                                                                                                                                                                                                                                                                 |                                                                                          |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| <b>WHY?</b>                                                                                                                                                                                                                                                                                                     | Problem-Solver<br>Confident Communicator<br>Strong Analytical and Statistical Skill sets |
| <b>Is an interdisciplinary and collaborative field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining</b>                                                                          |                                                                                          |
| <b>Basic Skills Needed</b>                                                                                                                                                                                                                                                                                      |                                                                                          |
| > Programming Skills - Know the Tools of the Trade, Data Extraction, Database Querying, SQL                                                                                                                                                                                                                     |                                                                                          |
| > Statistics - The different techniques and when to apply/not apply them                                                                                                                                                                                                                                        |                                                                                          |
| > Machine Learning Methods - k-nearest neighbors, random forests, ensemble methods                                                                                                                                                                                                                              |                                                                                          |
| > Multivariable Calculus & Linear Algebra - If you need to build your own implementations                                                                                                                                                                                                                       |                                                                                          |
| > Data Wrangling - Dealing with imperfections in data, date formatting and clean/prune data                                                                                                                                                                                                                     |                                                                                          |
| > Data Visualisation & Communication - Technical & non-technical descriptions/dashboards                                                                                                                                                                                                                        |                                                                                          |
| > Software Engineering - Knowing about data-logging, development of data-driven products                                                                                                                                                                                                                        |                                                                                          |
| > Data intuition - Data-driven problem solving skills and skepticism of existing assumptions                                                                                                                                                                                                                    |                                                                                          |
| > People Skills - Interaction with others including Management, IT Department                                                                                                                                                                                                                                   |                                                                                          |
| > Industry Knowledge - How industry functions, how data is collected, analysed and utilised                                                                                                                                                                                                                     |                                                                                          |
| > Contextual Understanding of data to determine hidden weaknesses, trends or opportunities                                                                                                                                                                                                                      |                                                                                          |
| <b>The Role will require you to:</b>                                                                                                                                                                                                                                                                            |                                                                                          |
| > Work closely with the business to identify issues & use data to propose solutions                                                                                                                                                                                                                             |                                                                                          |
| > Merge, manage, interrogate and extract data into formats that can be distributed widely                                                                                                                                                                                                                       |                                                                                          |
| > Maintain clear, coherent communications (written & verbal) to understand needs/report results                                                                                                                                                                                                                 |                                                                                          |
| > Horizon scan to stay up to date with the latest technology, techniques and methods                                                                                                                                                                                                                            |                                                                                          |
| > Conduct research from which you can develop models, prototypes and proof of concepts                                                                                                                                                                                                                          |                                                                                          |
| > Look for opportunities to use insight/datasets/code/models across organisational functions                                                                                                                                                                                                                    |                                                                                          |
| > Remain curious and enthusiastic about using algorithms to solve problems and enthuse others                                                                                                                                                                                                                   |                                                                                          |
| <b>Some of the more popular skills and techniques</b>                                                                                                                                                                                                                                                           |                                                                                          |
| > Data Preparation/ETL Stage (IBM InfoSphere; SAS Data Management; PowerCenter Informatica)                                                                                                                                                                                                                     |                                                                                          |
| > Data Mining (RapidMiner), Data Cleaning & Munging (Drake; OpenRefine; DataWrangler; DataCleaner)                                                                                                                                                                                                              |                                                                                          |
| > R and/or SAS Languages; SQL databases/data querying languages (Pentaho Business Analytics; Skytree)                                                                                                                                                                                                           |                                                                                          |
| > Language Runtimes - Python (most common); C/C++ Java; Perl, Ruby, Node.js                                                                                                                                                                                                                                     |                                                                                          |
| > Big Platforms - Hadoop; HDFS; Yarn; Apache Hive; Apache Pig; Apache Hbase                                                                                                                                                                                                                                     |                                                                                          |
| > Cloud Tools - Amazon Simple Storage Service (S3); GitHub; Ansible; Docker; Puppet; Nagios                                                                                                                                                                                                                     |                                                                                          |
| > Visualisation Tools - Tableau; D3.js; Highcharts; Microsoft Power BI; QlikView/QlikSense; ggplot2                                                                                                                                                                                                             |                                                                                          |
| > Recognised Certifications: Certified Analytics Professional (CAP); Cloudera Certified Professional: Data Scientist (CCP-DS); EMC: Data Science Associate (EMCDSA); SAS Certified Predictive Modeler using SAS Enterprise Miner 7                                                                              |                                                                                          |
| > Recognised Professional Organisations: Data Science Association; International Institute for Statistics (IAI); International Machine Learning Society (IMLS); Institute for Operations Research and the Management Sciences (INFORMS); Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) |                                                                                          |
| > Network Monitoring Tools - Solarwinds; Paessler PRTG; WhatsUp Gold; Nagios; WinRunner; Wireshark                                                                                                                                                                                                              |                                                                                          |
| > Data Emulators - PacketStorm (WAN); BlueStacks (Android); Trinegy (Network)                                                                                                                                                                                                                                   |                                                                                          |
| > Statistical Tools - Tableau; MATLAB; IBM SPSS; Statistix; Stata                                                                                                                                                                                                                                               |                                                                                          |

## DIFFERENT DATA FORMATS

| Open Formats<br>(Public Domain)                                                            |  |  | Data formats are used for storing digital data. Open formats are recommended & can be implemented by both proprietary and free    |
|--------------------------------------------------------------------------------------------|--|--|-----------------------------------------------------------------------------------------------------------------------------------|
| Closed Formats<br>(Trade Secrets)                                                          |  |  | and open source software, and are called free file formats if they don't have any copyrights, patents, trademarks or restrictions |
| Generic Data Types (with selected Open Formats identified if applicable)                   |  |  |                                                                                                                                   |
| > Access Control Data                                                                      |  |  |                                                                                                                                   |
| > Activity Log                                                                             |  |  |                                                                                                                                   |
| > Backup, Archiving & Compression - 7z; B1; gzip; lzip; MAFF; PAQ; SOX; tar; xz; ZIP       |  |  |                                                                                                                                   |
| > Classification & Relationship Data                                                       |  |  |                                                                                                                                   |
| > Configuration Data - YAML                                                                |  |  |                                                                                                                                   |
| > Credential Validation Data                                                               |  |  |                                                                                                                                   |
| > Credentials (e.g. passwords)                                                             |  |  |                                                                                                                                   |
| > Data Files (Basic & User Defined) - DOC; XLS; PPT (Microsoft Open Specification Promise) |  |  |                                                                                                                                   |
| > Data Files (Open Document) - ODF; ODS; ODP                                               |  |  |                                                                                                                                   |
| > Encrypted Data Files                                                                     |  |  |                                                                                                                                   |
| > Executable Code                                                                          |  |  |                                                                                                                                   |
| > Health & Usage Data                                                                      |  |  |                                                                                                                                   |
| > Internal Management Data                                                                 |  |  |                                                                                                                                   |
| > Meta Data (Reference)                                                                    |  |  |                                                                                                                                   |
| > Multimedia Data                                                                          |  |  |                                                                                                                                   |
| > Multimedia Data (Audio) - ALAC; FLAC; MP3; Ogg; Opus; Speex; Vorbis; WavPack, WAV        |  |  |                                                                                                                                   |
| > Multimedia Data (Imaging) - APNG; FLIF; GBR; GIF; JPEG 2000; MNG; PNG; SVG               |  |  |                                                                                                                                   |
| > Multimedia Data (Text) - Plain text; CSV; HTML; UTF-8 & 16; Office Open XML; PDF         |  |  |                                                                                                                                   |
| > Multimedia Data (Various) - CMML; SMIL; VRM/X3D; XSPF                                    |  |  |                                                                                                                                   |
| > Multimedia Data (Video) - Dirac; Matroska; WebM; Theora; MP3                             |  |  |                                                                                                                                   |
| > Primitive Built-in (Foundation) Data                                                     |  |  |                                                                                                                                   |
| > Query & Search Data                                                                      |  |  |                                                                                                                                   |
| > Source Code - YAML; XML (Mark-up Language)                                               |  |  |                                                                                                                                   |
| > Text Data                                                                                |  |  |                                                                                                                                   |
| Common Closed/Proprietary Formats                                                          |  |  |                                                                                                                                   |
| > AAC & MOV & KEY owned by Apple                                                           |  |  |                                                                                                                                   |
| > BMP owned by Microsoft                                                                   |  |  |                                                                                                                                   |
| > CDR a CorelDraw native format for vector graphic drawings                                |  |  |                                                                                                                                   |
| > DWG an AutoCAD drawing format                                                            |  |  |                                                                                                                                   |
| > PSD a Adobe Photoshop native image format                                                |  |  |                                                                                                                                   |
| > RAR owned by Alexander L. Roshar                                                         |  |  |                                                                                                                                   |
| > RTF defined and maintained only by Microsoft                                             |  |  |                                                                                                                                   |
| > SWF Adobe Flash Format (partially or completely open)                                    |  |  |                                                                                                                                   |
| > WMA & WMV owned by Microsoft                                                             |  |  |                                                                                                                                   |
| > WMA owned by Microsoft                                                                   |  |  |                                                                                                                                   |
| > XFA or Adobe XML Forms Architecture (used in PDF files)                                  |  |  |                                                                                                                                   |
| > ZIP format (Newer versions have some patented features)                                  |  |  |                                                                                                                                   |

## SUB-PROCESSES & PROCEDURES

|                                                                                             |                                                    |                                                                                                                                                               |
|---------------------------------------------------------------------------------------------|----------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| WHY?                                                                                        | Processes define the big picture & highlights      | Data Management does not live in isolation within organisations, it has a number of touch points with other processes/procedures and other business functions |
|                                                                                             | Procedures capture the depth, methods & objectives |                                                                                                                                                               |
| Getting the right balance between Control & Flexibility is difficult                        |                                                    |                                                                                                                                                               |
| > Enough structure to assist/remove barriers, but not too much to stifle or hinder is key   |                                                    |                                                                                                                                                               |
| Communication Management Process                                                            |                                                    |                                                                                                                                                               |
| > Communication Management Plan > Manage Comms > Control Comms thro' lifecycle              |                                                    |                                                                                                                                                               |
| > Identify Stakeholders > Distribute Information > Expectation Management > Report          |                                                    |                                                                                                                                                               |
| Data Governance Management Process                                                          |                                                    |                                                                                                                                                               |
| > Structure & Remit: Terms of Reference; Roles/Responsibilities; Accountability             |                                                    |                                                                                                                                                               |
| > Plans & Execution: Steering Group(s); Working Group(s); Linkages; Escalations             |                                                    |                                                                                                                                                               |
| > Define: Business Glossary/Rules; Classifications; Relationships; Reference Data; KPIs     |                                                    |                                                                                                                                                               |
| > Invoke: Identify required Data Definitions, then group related Definitions                |                                                    |                                                                                                                                                               |
| > Enablers: Data Gathering & Sources; Process Mapping; Reporting & Metrics; Training        |                                                    |                                                                                                                                                               |
| Stakeholder Management Process                                                              |                                                    |                                                                                                                                                               |
| > Identify > Classify > Needs & expectations > Influence > Communicate > Evaluate           |                                                    |                                                                                                                                                               |
| > Tools & Techniques = Information Management Systems; Expert Judgement; Meetings           |                                                    |                                                                                                                                                               |
| > Power & Influence = Key Players; Keep Satisfied; Keep Informed; Minimal Effort            |                                                    |                                                                                                                                                               |
| Release Management Process (includes Review & Acceptance)                                   |                                                    |                                                                                                                                                               |
| > Plan Release > Build > Test > Prepare for Release > Deploy > Support > Closure            |                                                    |                                                                                                                                                               |
| > Configuration Management > Hosting Repository > Version Control > Compatibility           |                                                    |                                                                                                                                                               |
| Issue & Risk Management Process                                                             |                                                    |                                                                                                                                                               |
| > Identification > Analysis > Evaluation > Action Implementation > Monitor & Review         |                                                    |                                                                                                                                                               |
| > Risk involves certain controls = Avoid; Mitigate; Transfer/Share; Accept/Tolerate         |                                                    |                                                                                                                                                               |
| Change Management Process                                                                   |                                                    |                                                                                                                                                               |
| > Sequence of steps or activities to apply change management to a change                    |                                                    |                                                                                                                                                               |
| > Identify > Request > Assess/Impact > Approve/Deny > Implement > Review/Report             |                                                    |                                                                                                                                                               |
| Security Classifications Process                                                            |                                                    |                                                                                                                                                               |
| > Documents should be labelled with the highest classification contained within             |                                                    |                                                                                                                                                               |
| > Classifications should be identified from an approved and published list                  |                                                    |                                                                                                                                                               |
| > The Originator sets the initial classification; Modifiers can increase the original level |                                                    |                                                                                                                                                               |
| > A procedure should be established for downgrading material to a lower level               |                                                    |                                                                                                                                                               |
| > Titles and metadata should always be aimed at the lowest security level                   |                                                    |                                                                                                                                                               |
| > Amalgamated or combined materials may well increase the security level                    |                                                    |                                                                                                                                                               |
| > If you need to mention higher security links this can be done via stubs or aliases        |                                                    |                                                                                                                                                               |
| Data Analysis Process                                                                       |                                                    |                                                                                                                                                               |
| > Acquire > Data Management > Model Creation/Validation/Deploy > Monitor Results            |                                                    |                                                                                                                                                               |
| > Data Mining: Raw Data; Target Data; Pre-processed; Transform; Patterns; Knowledge         |                                                    |                                                                                                                                                               |
| Other Business Functions                                                                    |                                                    |                                                                                                                                                               |
| > Financial Management; Purchasing; Supply Chain; Manufacturing/Operations; HR              |                                                    |                                                                                                                                                               |
| > Engineering; Test; Quality; Strategy & Planning; Project Management Office; R&D           |                                                    |                                                                                                                                                               |

## COMMON AVOIDABLE ERRORS

| WHY?                                                                     | Waste Time                                                                   | When you don't have a data taxonomy or endorsed/published data rules then you will naturally get a mix of everyone taking different approaches. People waste a significant amount |
|--------------------------------------------------------------------------|------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                          | Distractions People                                                          |                                                                                                                                                                                   |
|                                                                          | Demotivator                                                                  |                                                                                                                                                                                   |
|                                                                          | Wrong Outputs                                                                |                                                                                                                                                                                   |
| <b>Some simple rules for getting data into computer readable formats</b> |                                                                              |                                                                                                                                                                                   |
| R01                                                                      | Appoint a number of Points of Contact (POCs) for selected areas              |                                                                                                                                                                                   |
| R02                                                                      | POCs act as coordinators pulling work together to reach data goals           |                                                                                                                                                                                   |
| R03                                                                      | Have a plan of what you are going to tackle first and who does it            |                                                                                                                                                                                   |
| R04                                                                      | Develop an enterprise list of acronyms include source of the term            |                                                                                                                                                                                   |
| R05                                                                      | Develop an enterprise glossary of terms include source of the term           |                                                                                                                                                                                   |
| R06                                                                      | Develop an acronym list of key elements (suppliers, locations, products)     |                                                                                                                                                                                   |
| R07                                                                      | Develop a method of generating unique artefact/directory/file names          |                                                                                                                                                                                   |
| R08                                                                      | Create a single area on the Corporate Network to publish this information    |                                                                                                                                                                                   |
| R09                                                                      | Publish lists of problematic electronic symbols (< > \ [ ] + = ? & , ' " ' ) |                                                                                                                                                                                   |
| R10                                                                      | Publish a method of how the date should be displayed (e.g. 06-Mar-18)        |                                                                                                                                                                                   |
| R11                                                                      | Publish updates providing clarity of ambiguities (Real-Time or Realtime)     |                                                                                                                                                                                   |
| R12                                                                      | Publish guidance on when/when not to use spaces and underscores              |                                                                                                                                                                                   |
| R13                                                                      | Publish guidance on how to avoid long (>256) root directories names          |                                                                                                                                                                                   |
| R14                                                                      | Publish guidance on which language to use (Organisation or Organization)     |                                                                                                                                                                                   |
| R15                                                                      | Publish guidance on the high ANSI characters (ôâôôSGé)                       |                                                                                                                                                                                   |
| R16                                                                      | Publish guidance on dealing with timezone differences                        |                                                                                                                                                                                   |
| R17                                                                      | Publish guidance on the use/approval of different languages                  |                                                                                                                                                                                   |
| R18                                                                      | Publish guidance on where to locate templates, graphs and standards          |                                                                                                                                                                                   |
| R19                                                                      | Publish guidance on how people should be undertaking analysis/statistics     |                                                                                                                                                                                   |
| R20                                                                      | Publish guidance on how people should be collecting and storing data         |                                                                                                                                                                                   |
| R21                                                                      | Publish guidance on metrics (KPIs, SLAs, Recognised Data Sources)            |                                                                                                                                                                                   |
| R22                                                                      | Publish guidance on how people create a recognised data source               |                                                                                                                                                                                   |
| R23                                                                      | Publish guidance on how people download data to do local analysis            |                                                                                                                                                                                   |
| R24                                                                      | Publish guidance on how people upload data into approved data stores         |                                                                                                                                                                                   |
| R25                                                                      | Publish guidance on how people do change management/version control          |                                                                                                                                                                                   |
| R26                                                                      | Publish guidance on how your metrics are generated, analysed & reported      |                                                                                                                                                                                   |
| R27                                                                      | Publish guidance on dealing with messy data, outliers and normalisation      |                                                                                                                                                                                   |
| R28                                                                      | Publish guidance on approved tools, techniques and where to get training     |                                                                                                                                                                                   |
| R29                                                                      | Get people to record when/why they deviate from published guidances          |                                                                                                                                                                                   |
| R30                                                                      | Get people to become experts in certain aspects of data that they like       |                                                                                                                                                                                   |
| R31                                                                      | Publish the names of people who want to become data mentors                  |                                                                                                                                                                                   |
| R32                                                                      | Start to develop forums on how to continuously improve data management       |                                                                                                                                                                                   |
| R33                                                                      | Get your internal people/contractors to run lunch and learn data sessions    |                                                                                                                                                                                   |
| R34                                                                      | Start to develop Tiger Teams to address the more complex data problems       |                                                                                                                                                                                   |
| R35                                                                      | Create data analysis workshops were people come to review results            |                                                                                                                                                                                   |
| R36                                                                      | Get your people to start developing more and more in-house tools/models      |                                                                                                                                                                                   |
| R37                                                                      | Have a system to monitor, manage, police and improve your data management    |                                                                                                                                                                                   |

## References

1. <https://www.productiveflourishing.com/four-levels-of-communication/>
2. <https://hbr.org/2017/05/whats-your-data-strategy>
3. [https://www.sas.com/content/dam/SAS/en\\_us/doc/whitepaper1/5-essential-components-of-data-strategy-108109.pdf](https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/5-essential-components-of-data-strategy-108109.pdf)
4. <https://www.slideshare.net/svdatascience/8-steps-to-creating-a-data-strategy>
5. [https://www.sas.com/en\\_us/insights/articles/data-management/5-data-management-best-practices.html](https://www.sas.com/en_us/insights/articles/data-management/5-data-management-best-practices.html)
6. [https://libguides.bc.edu/dataplan/best\\_practices](https://libguides.bc.edu/dataplan/best_practices)
7. <https://www.quirks.com/articles/5-steps-to-unlocking-the-value-in-unstructured-big-data>
8. <https://www.techrepublic.com/blog/smb-technologist/four-steps-to-making-the-most-of-your-unstructured-data-gold-mine/>
9. <https://tdwi.org/articles/2016/02/16/data-governance-glossary-missing-elements.aspx>
10. <https://www.colibra.com/blog/4-data-governance-best-practices>
11. [https://www.whitepapers.em360tech.com/wp-content/files\\_mf/white\\_paper/wp\\_iway\\_7steps.pdf](https://www.whitepapers.em360tech.com/wp-content/files_mf/white_paper/wp_iway_7steps.pdf)
12. <https://tdwi.org/Articles/2015/05/05/4-Big-Data-Truths.aspx?Page=2>
13. <https://www.cio.com/article/3233244/it-strategy/6-hard-truths-it-must-learn-to-accept.html>
14. <https://blog.udacity.com/2014/11/data-science-job-skills.html>
15. <https://www.prospects.ac.uk/job-profiles/data-scientist>
16. <http://www.mastersindatascience.org/careers/data-scientist/>
17. [https://en.wikipedia.org/wiki/List\\_of\\_open\\_formats](https://en.wikipedia.org/wiki/List_of_open_formats)
18. [https://en.wikipedia.org/wiki/Proprietary\\_format](https://en.wikipedia.org/wiki/Proprietary_format)



Dr Karl Phillips

### What do you mean that what got us here, won't get us there?

Where you want to go in the future will need new mind-sets, new knowledge, new skills and new abilities, in order to help your people, face a future that cannot be predicted with any degree of accuracy. Successful organisations always see that their past successes are just a prologue to their next stepping stones, and are always looking to move their people forward.

+44 (0) 7468 489060



Copyright © 2017. Innov8or Solutions Ltd. All Rights Reserved

How to use data and the skills surrounding data handling and analysis are more important today than they have ever been. At the end of the day it will come down to those companies that can use and benefit from it, and the rest that will just get left behind and disappear. Best intentions will not longer be sufficient in far more competitive and challenging markets ahead. It is time you took data series, because coupled with good staff, it is really going to make a big difference. Are you ready?