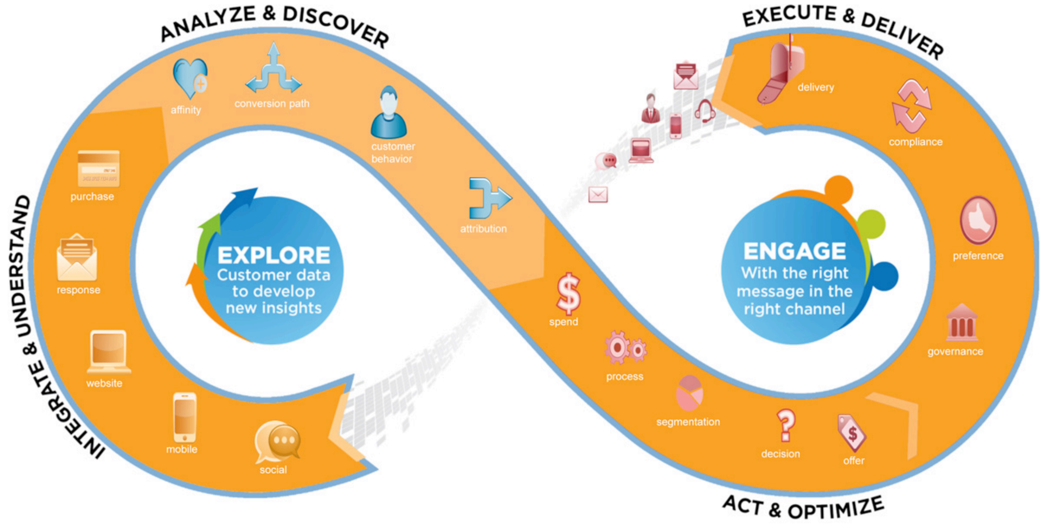
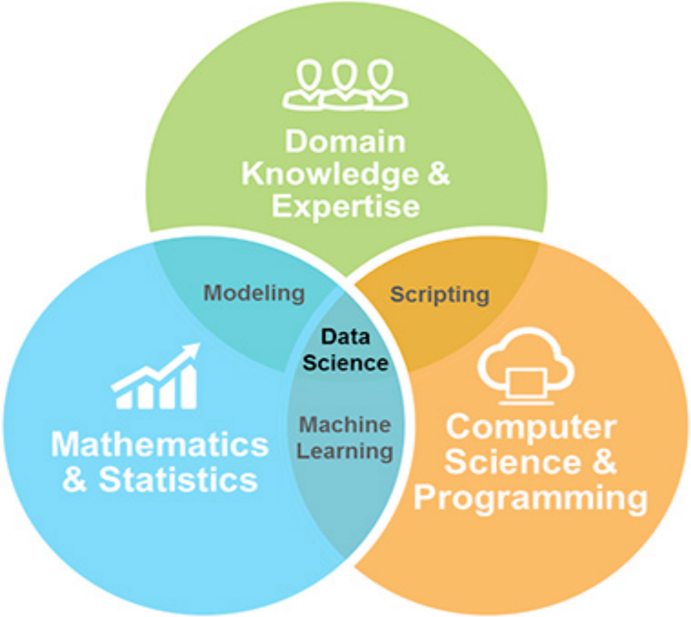


Data Science Areas: From data to insights

By Dr. Rosa Filgueira

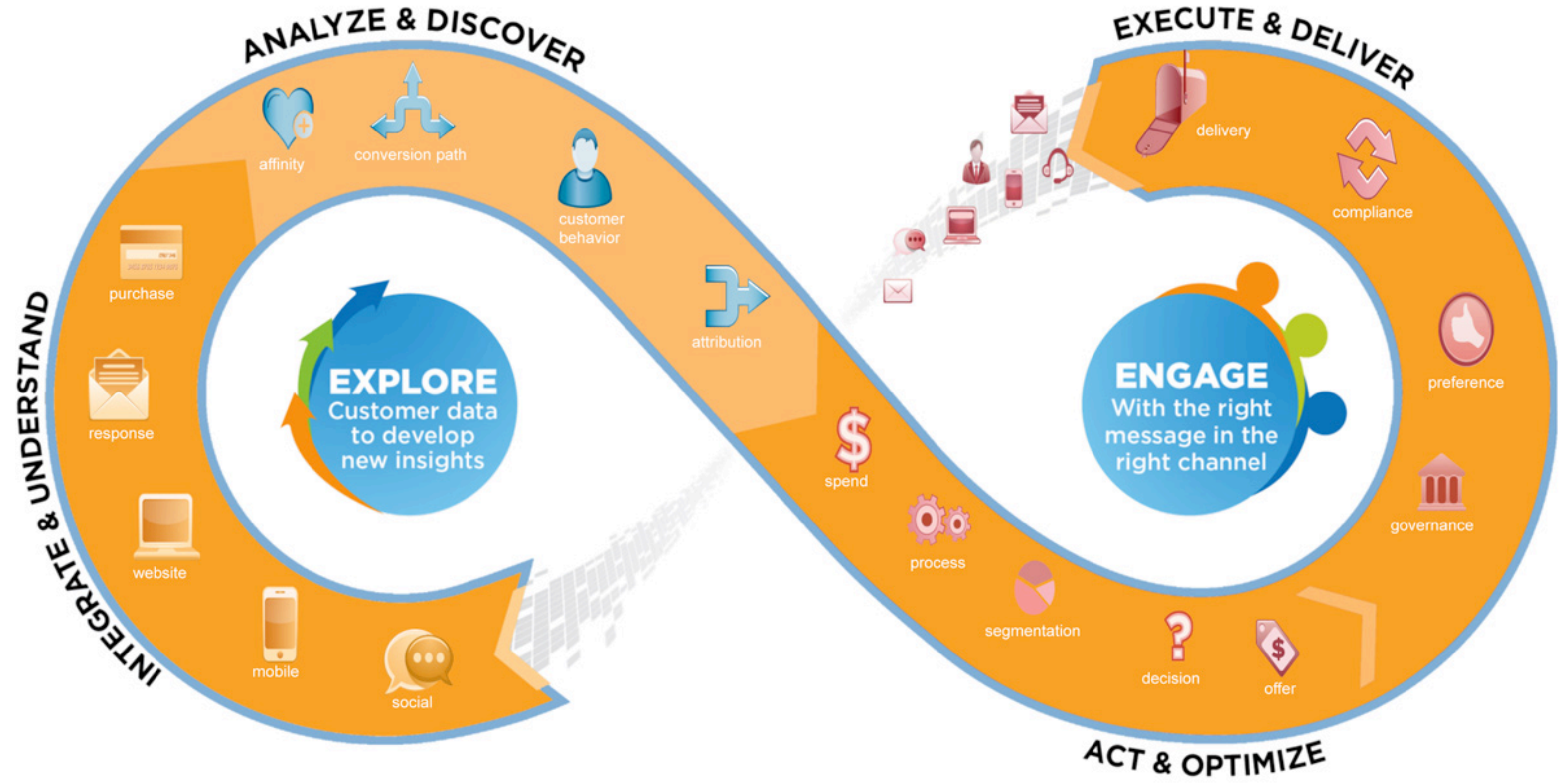
What you need depends on what you want

- What you need from Data Science in terms of technology, skills, algorithms, resources etc. depends on:
  - Which insights you want to discover from a particular data source
  - How is the data source organized
  - Volume/Variety/Velocity/ Veracity of the data
  - How often do we want to run our analyses …



http://www.genpact.com/home/blogs/bloginner?Title=The+interplay+of+data+science+skills+areas

Data Science journey



https://disruptivedigital.wordpress.com/2015/08/30/digital-marketing-data-science-programmatic-marketing/

10 questions of Data Science (DS)

- To help you to understand what to focus on among all the areas of Data Sciences

- We have designed the “10 QUESTIONS of DS”

10 questions of DS - 1

- Do you want to **analyze** a **dataset** ?
  - Analytics :
    - Artificial Intelligence
    - Machine Learning
    - Natural Processing
    - Data Mining
    - Text Mining
    - Statistics
    - Domain science algorithms
    - Predictive analytics
    - Deep learning
    - Preprocessing techniques
      - Cleaning, Gap Filtering, Noise removal … etc.
  - Computing languages
    - R, Python, MatLab, Fortran, C, C++

10 questions of DS - 2

- Do you want to **validate the analysis** ?
  - Models/algorithms testing environment
  - Facilities to move models from testing environment to production
  - Provenance tools to understand the models
  - Profiling tools for debugging and testing models

10 questions of DS - 3

- Do you want to **scale the analysis up** ?
  - Parallel engines
    - MPI/OpenMP/Cuda
    - Data-pipelines
  - Distributed computing resources
    - Cloud
    - HPC clusters
    - GPUs/FPGAs
  - Data Centers/Storage
    - Parallel File Systems:
    - Hadoop
    - Repositories
    - DB
      - Relational/Non-Relational

10 questions of DS - 4

- Do you want to **repeat the analysis** every “X” times ?
  - Automation tools
    - Scientific workflows
      - Data-Flow / Task-Flow/ Stream-Flow
    - Data Frameworks
      - Deep Learning: TensorFlow, etc.
      - Data Bricks: Apache Spark, Apache Flink, etc.
  - Distributed Computing Resources – slide 7
  - Descriptions
      - Linked Data/ semantic web
      - Catalogs
        - Models
        - Data and Metadata
        - Storage
        - Computing resources
      - Ontologies/Taxonomies/Vocabularies
      - Abstractions
  - Data Centers/Storage – slide 7

10 questions of DS - 5

- Do you want to **optimize the analysis** ?
  - Optimization areas
    - CPU/Memory/Runtime/Usability/Scalability
  - Optimization algorithms
    - Transparent to users
    - Add-ons/plugins
  - Current understanding
    - Monitoring tools
    - Provenance tools

10 questions of DS - 6

- Do you want to **build an easy-to-use analysis framework** (so others can analyse their data easily)
  - Visualization systems
    - GUIs
    - Maps
    - Science Gateways/ Virtual Research Environments
    - Dashboards
  - Frameworks that hide underlying technology
    - Scientific Workflows – Slide 8
    - Data Frameworks – Slide 8
    - Containers
    - Optimizations
    - Repositories
    - Catalogues
    - Descriptions – Slides 8

10 questions of DS - 7

- Do you want to make the **analysis reproducible** ?
  - Descriptions – Slide 8
    - Software/Data
  - Ontologies/Taxonomies/Vocabularies
  - Unique Identifiers
    - Software/Data
  - Provenance tools
    - W3C PROV
  - Governance
  - Package your software and their dependencies
    - Containers image
  - Use standards for storing data and metadata
  - Portable computing environments
    - Virtual Machines
    - Containers

10 questions of DS - 8

- Do you want to **build a new dataset** for applying later further analysis, by combining/selecting/filter data from several sources ?
  - By filtering big data from several files/Dbs ?
    - Data-pipelines
    - Farm/Array jobs
    - Parallel engines – Slide 7
  - From several websites and protocols (Ftp, http, web services)
    - Data Wrangling
    - Web crawling
    - Scripting

Note:

Store the new datasets / Data Integration/ Warehouse

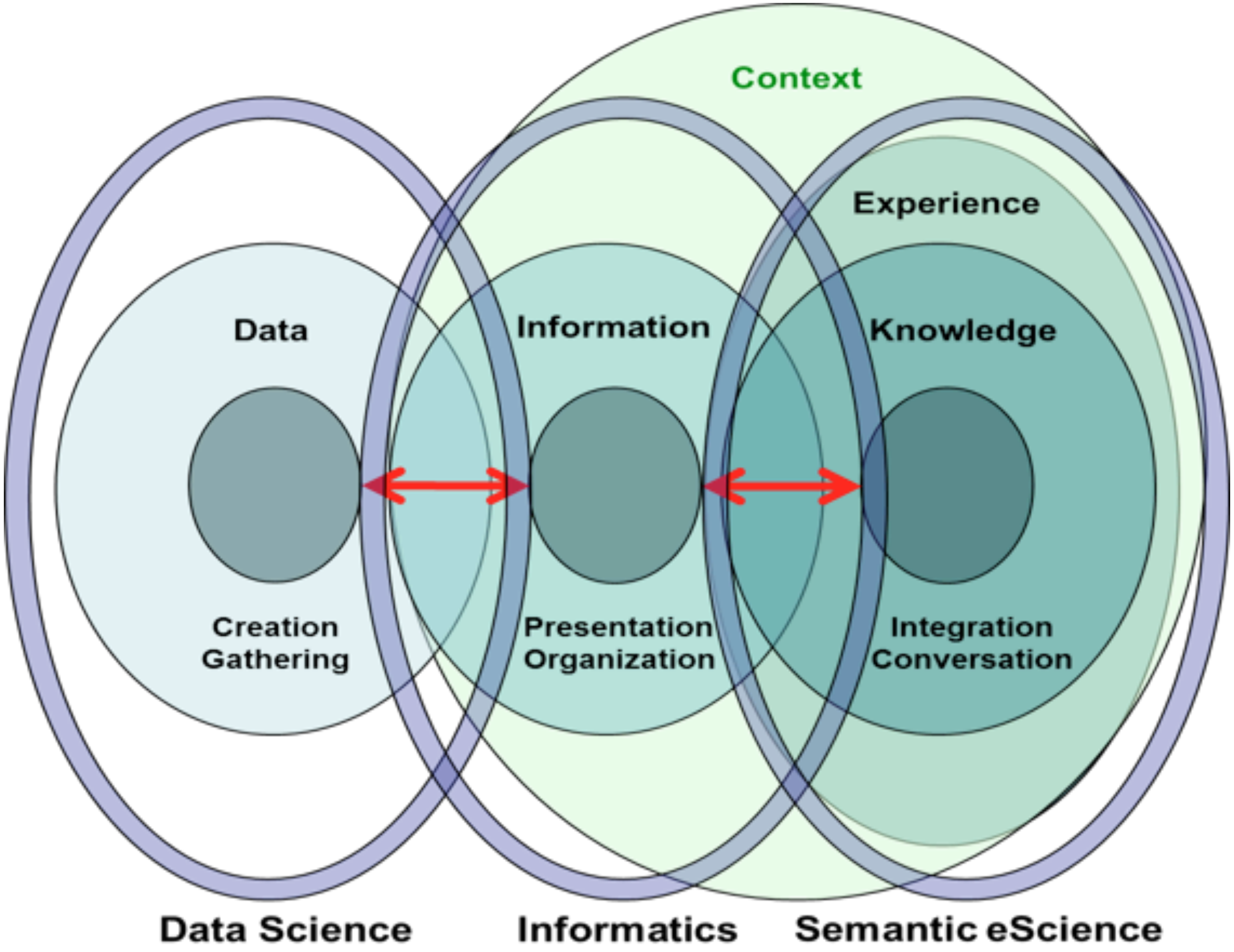
- Json files/Db/Repositories/Catalogues
- Metadata
- Apply the standards
- Unique Identifiers

10 questions of DS - 9

- Do you want to build an **interoperable data analysis system** ?
  - Run automatically the analysis with different datasets
    - Apply the data/metadata standards
    - Input data from catalogues/web services
    - Output data to catalogues/web services
  - Run automatically the analysis with different models against the data?
    - Scientific Workflows – Slide 8
    - Data-Frameworks - Slide 8
    - Abstraction of models / Descriptions – Slide 8
    - Repository/Catalogue of models
  - Run automatically in different computing resources ?
    - Containers
    - VMs

10 questions of DS - 10

- Do you want to **explore the data** ?
  - Visualization techniques
    - D3js14
    - Maps
    - GUIs
    - dashboards
  - Query facilities



https://www.linkedin.com/pulse/data-science-informatics-semantic-escience-shawn-riley