

Ciência dos Dados e Big Data – organizações e sociedade direcionadas por dados

28 de setembro de 2021

Prof. Dr. Pedro Luiz Pizzigatti Corrêa - pedro.correa@usp.br

Departamento de Engenharia de Computação e Sistemas Digitais

Escola Politécnica da Universidade de São Paulo - EPUSP

Grupo de Pesquisa e Extensão em Big Data da EPUSP wds.poli.usp.br



Agenda

- Introdução
- Boas Práticas para a Gestão e Análise de Dados
- Exemplos de aplicações de Big Data
- Conclusão





BIG WORLD

BIG PROBLEMS

BIG DATA



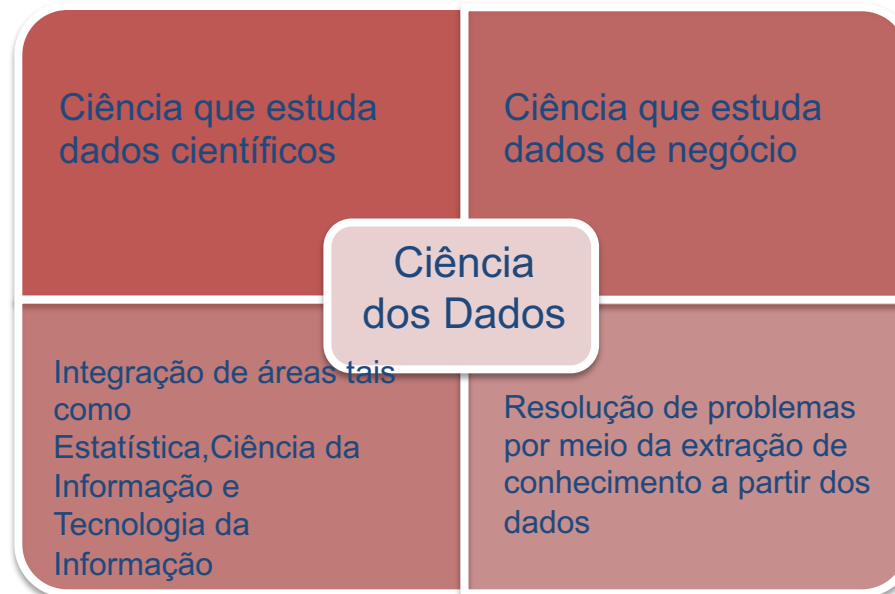
The
**FOURTH
PARADIGM**

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Ciência dos Dados

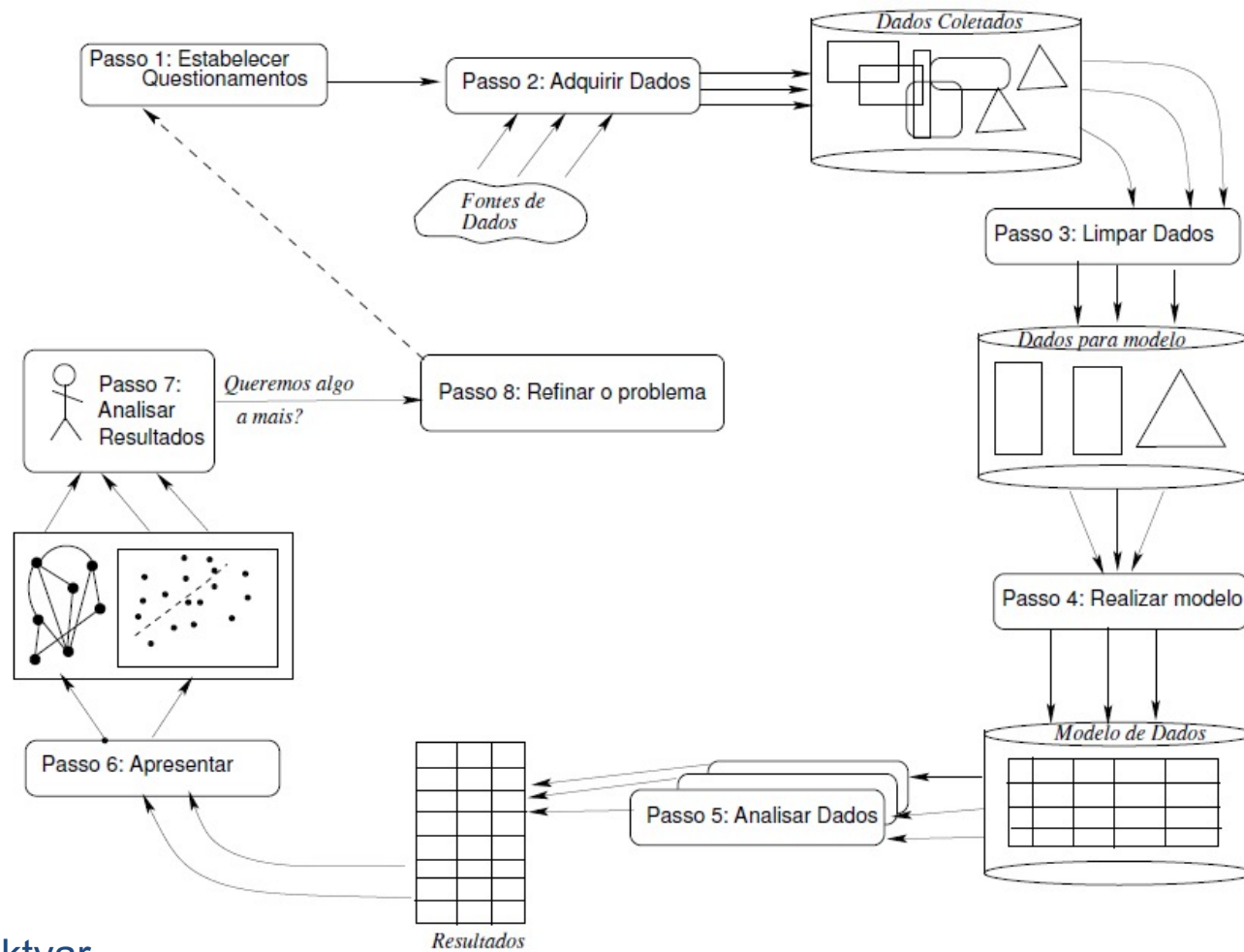
- Em busca por uma definição formal sobre Ciência dos Dados, encontramos diversos trabalhos na literatura
 - Embora muito se discuta sobre a composição das atividades de Ciência dos Dados, o seu conceito ainda não é algo fundamentalmente estabelecido
- Para Zhu e Xiong (2015), há quatro vertentes (perspectivas) que buscam caracterizar Ciência dos Dados



Ciência dos Dados

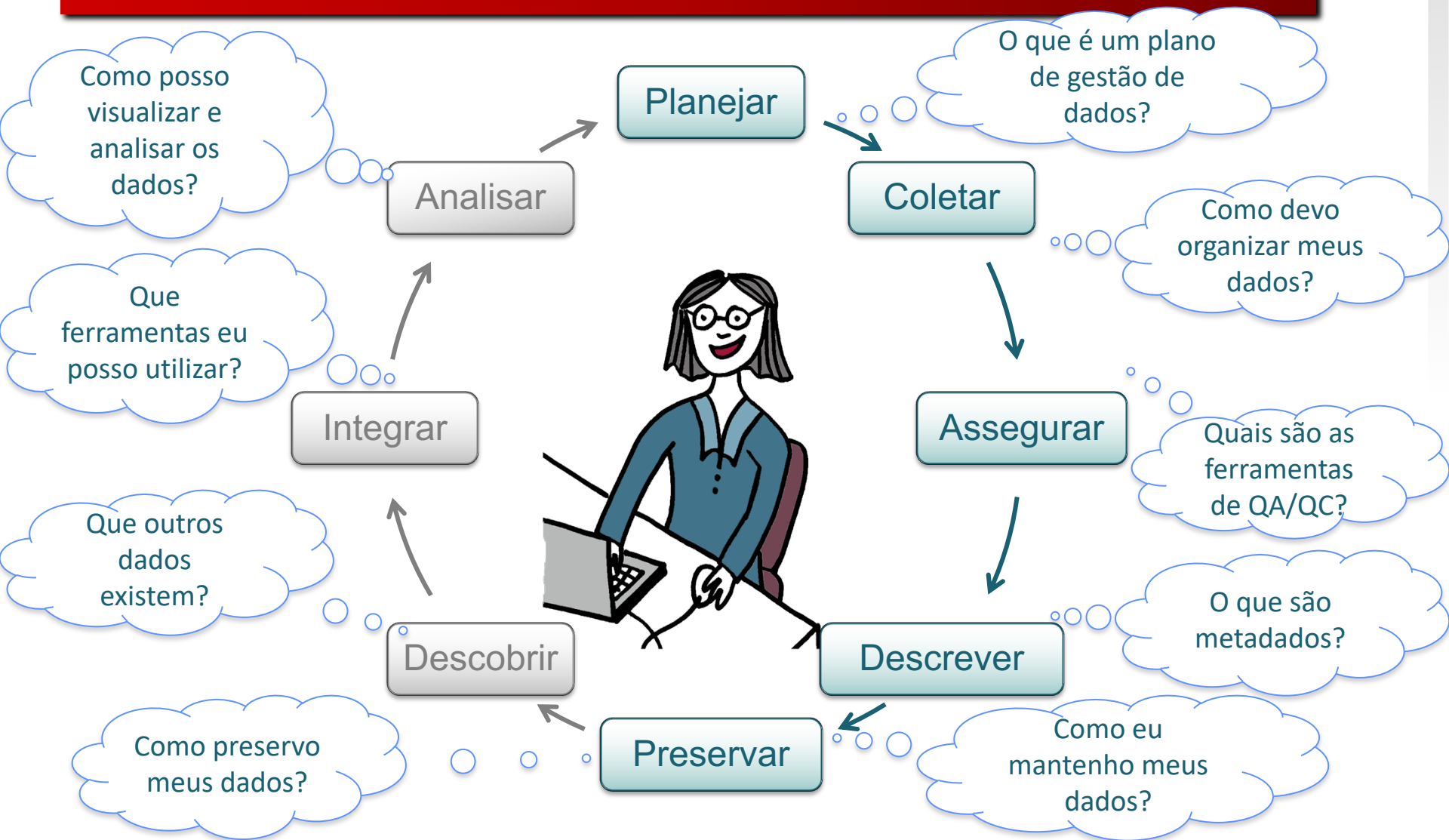
- Embora não haja consenso sobre a definição, encontramos como elemento comum em todas as propostas um processo de manipulação, processamento e análise de dados, que visa a descoberta de novos conhecimentos
- Para Alex Dehktyar (2016),
 - Ciência dos dados é uma disciplina que permite tratar o ciclo de trabalho com os dados, considerando atividades que compreendem desde a aquisição dos dados, passando pela análise dos dados, até o processo de apresentação dos dados e obtenção de novos conhecimentos

Ciência dos Dados - Processo



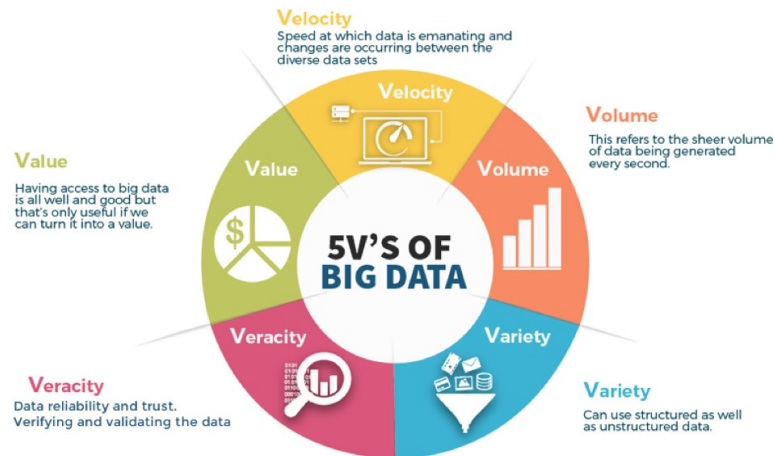
Cortesia: Alex Dehktyar

Gestão de Dados

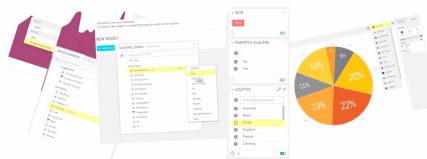


O que é Big Data ?

- ✓ É uma aplicação computacional de Ciência dos Dados que tem por objetivo analisar, extrair sistematicamente informações de grandes volumes de conjuntos de dados, para os quais técnicas computacionais tradicionais não são adequadas. Os desafios para gestão dos dados são classificados em 5V's (Chen et al., 2012, Kwon et al., 2014).



- ✓ Big data é um grande volume de dados, alta velocidade e alta variedade de ativos de informação que demandam formas inovadoras e econômicas de processamento de informações para melhor insight e tomada de decisões.” (“Gartner IT Glossary, n.d.”)



Requisitos para Big Data:

● Precisão

● Atualizado

● Pronto para análise



Quais são os dados ?

Coleções de **registros ou medições** que fornecem um registro de evidências do evento observado “... *qualquer informação que possa ser armazenada em formato digital, incluindo texto, números, imagens, vídeo ou filmes, áudio, software, algoritmos, equações, animações, modelos, simulações, etc.* ”

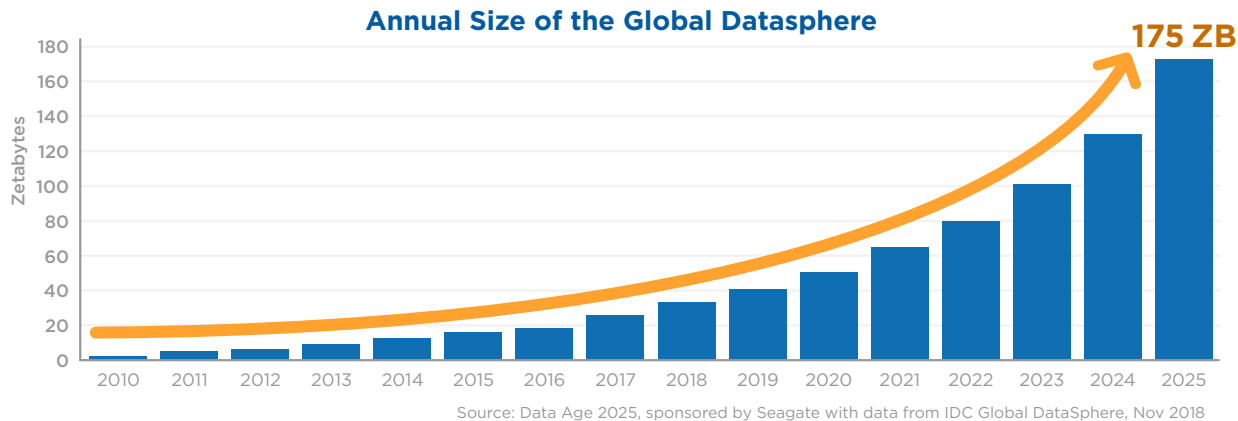
Atributos gerais da informação:

- Digital;
- Heterogêneo;
- Contextualizado;
- Valioso.

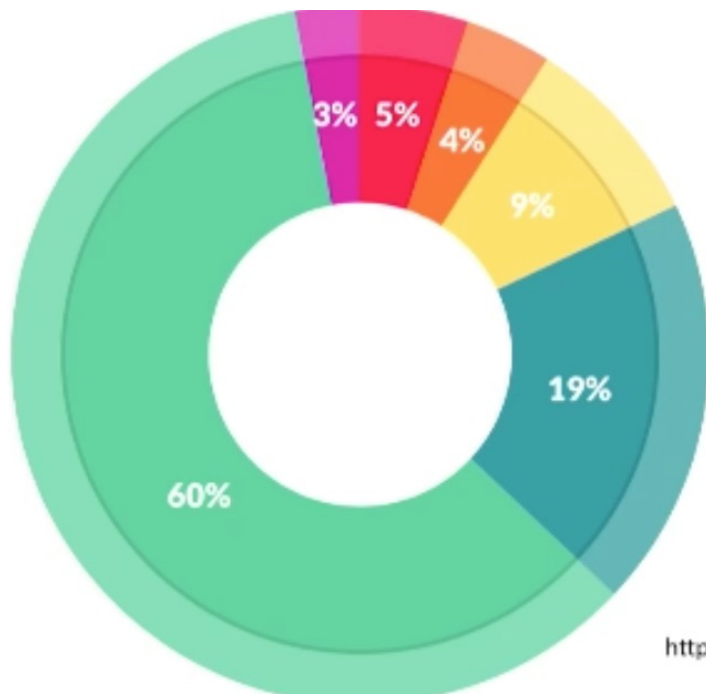
Problema: Volume de Dados

International Data Corporation (IDC):

- A informação no mundo está mais que dobrando, está dobrando cada dois anos (30 Zettabytes em 2018);
- O número de arquivos está crescendo mais rápido que a capacidade de armazenamento. Nos próximos 5 anos esses arquivos irão crescer num fator de 8;
- O número de pessoas na área de IT responsável por esses dados irá crescer “suavemente”;



Problema que conhecemos muito bem !



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

- 60% em atividades na organização e limpeza dos dados;
- 12% em atividades de pré-processamento
- 9% em atividades de análise

Imagine o seguinte cenário:

CEO: "Nós Precisamos aumentar nossas vendas"



Marketing Manager: "Quais outros produtos podemos vender?"



IT Manager: "Com nosso atual *backlog* estimo que teremos essa informação em 1 mês"



Marketing Manager: "1 mês ? Nós *já temos esses dados* em nossos sistemas ?"



IT Manager: "Nós já temos, mas *não na correta estrutura* para responder essa questão !"



CEO: "Só preciso aumentar as vendas!"

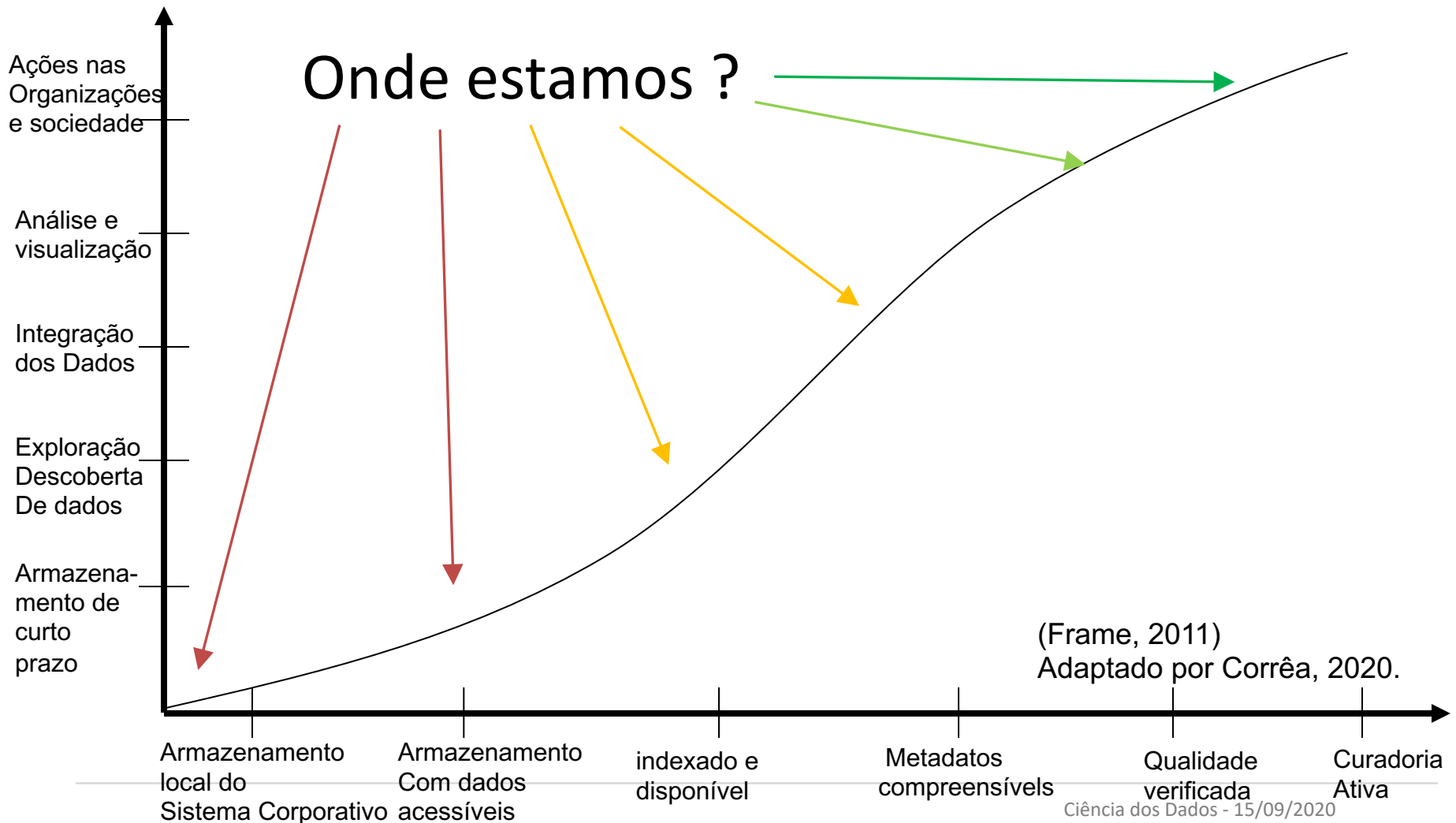


Agenda

- Introdução
- Boas Práticas para a Gestão de Dados
- Aplicações na área de Big Data
- Conclusão



Avaliação da Gestão de Dados



Modelo de Maturidade



Framework Organizacional de TI - DataLab

PESSOAS

ALFABETIZAÇÃO DE DADOS

ENGAJAMENTO DOS EXECUTIVOS

CATALISADORES DE DADOS
(COMUNIDADE)

ALINHAMENTO COM CULTURA
ORGANIZACIONAL

GOVERNANÇA

EXPLICAB. E
INTERPRETAB.

LINHAGEM E
ORIGEM

MÉTRICAS

PROPRIEDADE

QUALIDADE

SEGURANÇA

PRIVACIDADE

DATA SCIENCE

ANALISAR

PREPARAR

EXPLORAR

GERAR INSIGHTS

MACHINE LEARNING / DEEP LEARNING

CRIAR MODELOS

TREINAR

AVALIAR

PREVER

STORYTELLING (VISUALIZAÇÃO DE DADOS)

DASHBOARDS

REPORTS

KPIS

ALERTAS

BIG DATA / DATA ENGINEERING

INGESTÃO

TRANSFORM.

MODELAGEM

REAL-TIME

BATCH

ARMAZENAM.

DATAOPS / MLOPS

DESENVOLVER

TESTAR

ORQUESTRAR

ENTREGAR

MONITORAR

SUPPORT FRAMEWORKS

FAIR Data Principles

to be Findable, Accessible, Interoperable and Re-Usable



Article

A Data Quality Strategy to Enable FAIR, Programmatic Access across Large, Diverse Data Collections for High Performance Data Analysis

Ben Evans, Kelsey Davies, Jingbo Wang*, Rai Yang, Clare Richards and Lesley Wyborn

National Computational Infrastructure, for Australian National University, Acton 201, Australia; Ben.Evans@nicta.edu.au (B.E.); Kelsey.Davies@nicta.edu.au (K.D.); Rai.Yang@nicta.edu.au (R.Y.); Clare.Richards@nicta.edu.au (C.R.); Lesley.Wyborn@nicta.edu.au (L.W.)

* Correspondence: Jingbo.Wang@nicta.edu.au; Tel.: +61-02-4125-0862

Academic Editors: Moushi Gu and Vladimir Dobral

Received: 31 August 2017; Accepted: 6 December 2017; Published: 13 December 2017

Abstract: To ensure seamless, programmatic access to data for High Performance Computing and analysis across multiple research domains, it is vital to have a methodology for standardisation of both data and services. At the Australian National Computational Infrastructure (NCI) we developed a Data Quality Strategy (DQS) that currently provides processes for: (1) Control of data structures needed for a High Performance Data (HPC) platform; (2) Quality Checks through compliance with recognised community standards; (3) Benchmarking cases of open performance tests; and (4) Quality Assurance (QA) of data through demonstrated metrics and performance across common platforms, tools and services. By implementing the NCI we have seen progressive improvement in the quality and usefulness of the datasets across different subject domains, and demonstrated the ease by which modern programmatic methods can be used to access the data, either in situ or via web services, and for uses ranging from traditional analysis methods through to emerging machine learning techniques. To help increase data re-use by broader communities, particularly in high performance environments, the DQS is also identifying the need for any extensions to the relevant international standards for interoperability and programmatic access.

Keywords: data quality; quality control; quality assurance; benchmarks; performance; data management policy; netCDF; high performance computing; HPC; big data

1. Introduction

The National Computational Infrastructure (NCI) manages one of Australia's largest and diverse repositories (10+ PB) of research data collections spanning datasets from climate, oceans and geophysics through to astronomy, bioinformatics and the social sciences [1]. Within domains, data can be of different types such as gridded, ungridded (i.e., line surveys, point cloud and raster image types), as well as having diverse coordinate reference projections and local NCI has been following the Force 11 FAIR data principles to make data Findable, Accessible, Interoperable, and Reusable [2]. These principles provide guidelines for a research data repository enable data-intensive science, and enable researchers to answer questions such as how can I improve the quality of the data? Is the data usable by my software platform and my tools?

To ensure broader reuse of the data, enable trans-disciplinary integration across multiple domains as well as enabling programmatic access, a dataset must be usable and of value to a broad range of users from different communities [1]. Therefore, a set of standards and 'best practices' for the quality of scientific data products is a critical component in the life cycle of data management.

informatics 2017, 4, 45; doi:10.3390/informatics4040045

www.mdpi.com/journal/informatics

<https://doi.org/10.3390/informatics4040045>

DEVELOPMENT PROCESS

TOOLS

PUBLISHING

COMPLIANCE STANDARDS

FILE FORMAT

DATA SERVING: THREDDS, OpenDAP, WMS, etc.
DATA USAGE: Matlab, R, Python, GDAL, etc.

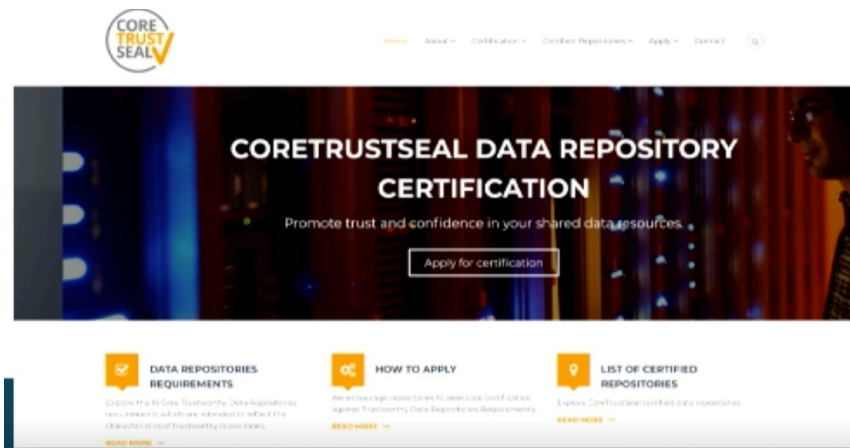
Digital Object Identifiers (DOI) minting,
Making metadata/data available and discoverable online

FILE (GRANULE)-LEVEL	COLLECTION & DATASET-LEVEL
<ul style="list-style-type: none">Climate and Forecasts (CF) ConventionAttribute Convention Dataset Discovery (ACDD)Additional discipline specific standards	Data Management Plans (ISO 19115, ANZLIC, etc.)

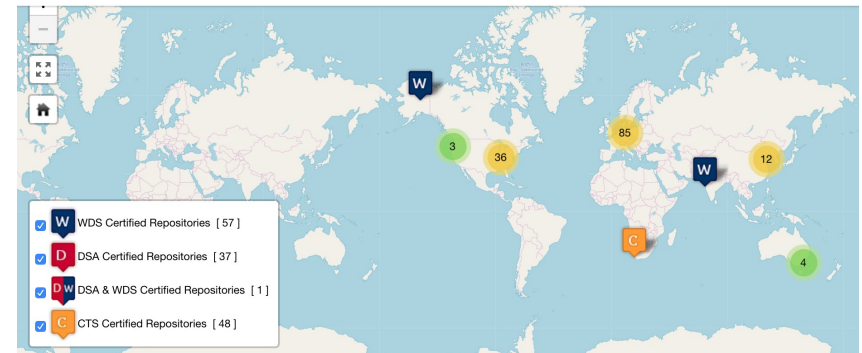
Self-describing file formats (e.g., NetCDF, HDF)

Confiabilidade para reuso dos dados

- Se alguém entregar seus dados, o que é necessário para convencê-lo que os dados estão corretos ?
- Se requer um sistema complexo para executá-lo e que não tem acesso, o que precisa para confiar dos dados ? Você conhece quais são as suposições e dependências existentes ?
- Quanto você poderá confiar que os mesmo dado estará disponível por um longo período de tempo ?



The screenshot shows the 'CORE TRUST SEAL DATA REPOSITORY CERTIFICATION' page. The main heading is 'CORETRUSTSEAL DATA REPOSITORY CERTIFICATION' with the subtext 'Promote trust and confidence in your shared data resources.' Below this is a button that says 'Apply for certification'. At the bottom, there are three sections: 'DATA REPOSITORIES REQUIREMENTS', 'HOW TO APPLY', and 'LIST OF CERTIFIED REPOSITORIES', each with a brief description and a 'READ MORE' link.



Arquitetura de Big Data

Projeto Hadoop

2003

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung
Google*



2004

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat
jeff@google.com, sanjay@google.com
Google, Inc.



2006

Bigtable: A Distributed Storage System for Structured Data

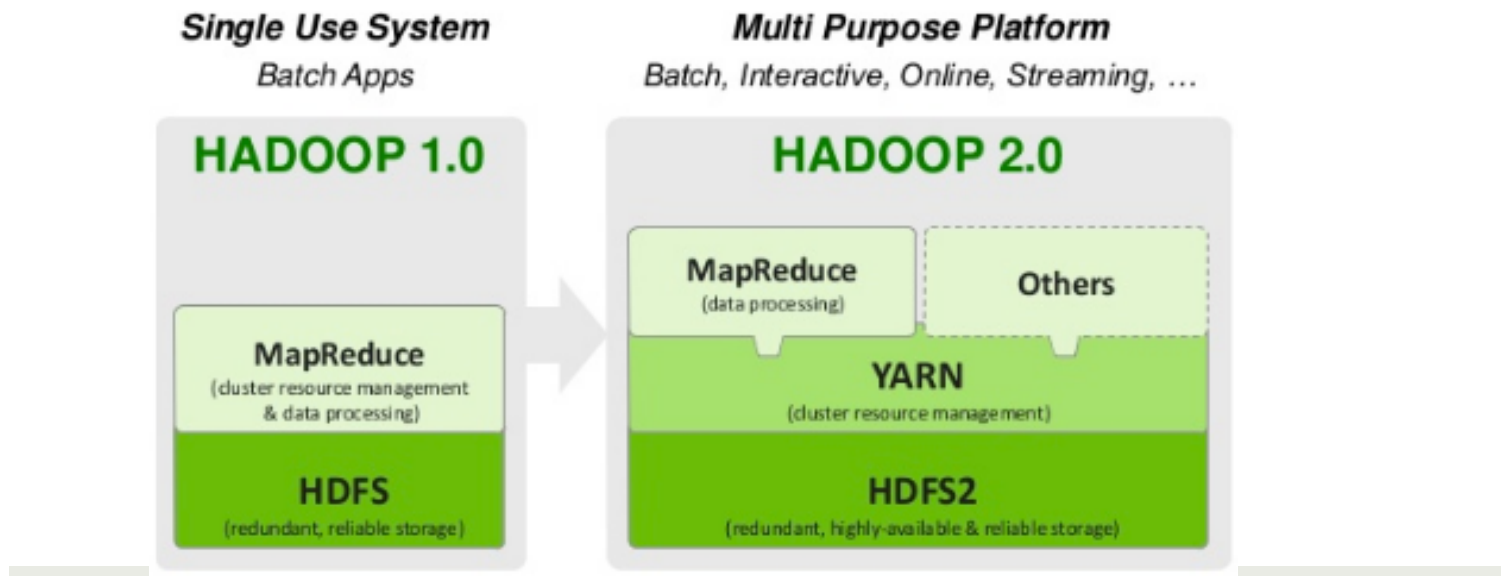
Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach
Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber
{fay,jeff,sanjay,wilson,hkerr,m.7b,tushar,fikes,gruber}@google.com
Google, Inc.



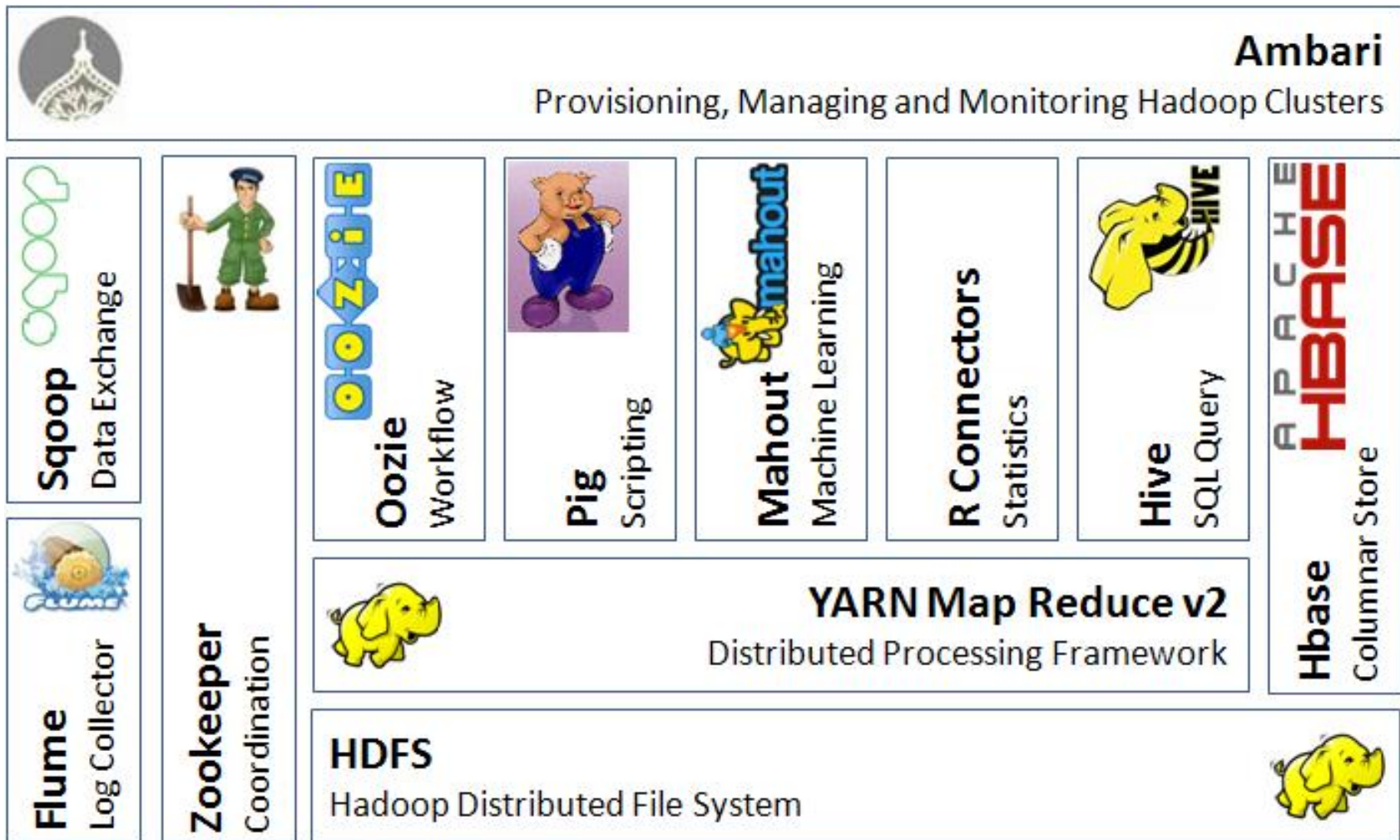
Abstract
Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large
achieved scalability and high performance, but Bigtable provides a different interface than such systems. Bigtable does not support a full relational data model; instead,

Projeto Hadoop

- ▣ **MapReduce** é um paradigma de programação paralela para processamento distribuído proposto pelo Google.
- ▣ **MAP** é o processo de mapear/dividir a requisição
- ▣ **REDUCE** é o processo de agregação do resultado



Arquitetura/Ecosystema Hadoop



Distribuições Hadoop



Agenda

- Introdução
- Boas Práticas para a Gestão de Dados Científicos
- **Aplicações de Big Data**
- Conclusão



Aplicações de Big Data

Financeiro	Transações Dados Cadastrais Credit Bureau Índices Econômicos	Credit Score Customer Behavior Prevenção a Fraudes Risco de Crédito Previsão de Inadimplência
Saúde	Prontuário Médico Histórico de Exames Dados familiares	Diagnóstico preditivo Descoberta de doenças e tratamentos Prevenção de epidemias Efeitos adversos de tratamentos
Vendas e Marketing	Transações Hábitos de consumo Personas	Recomendação de Produtos Lançamento de Produtos Segmentação de Clientes Churn Analysis

Aplicações de Big Data

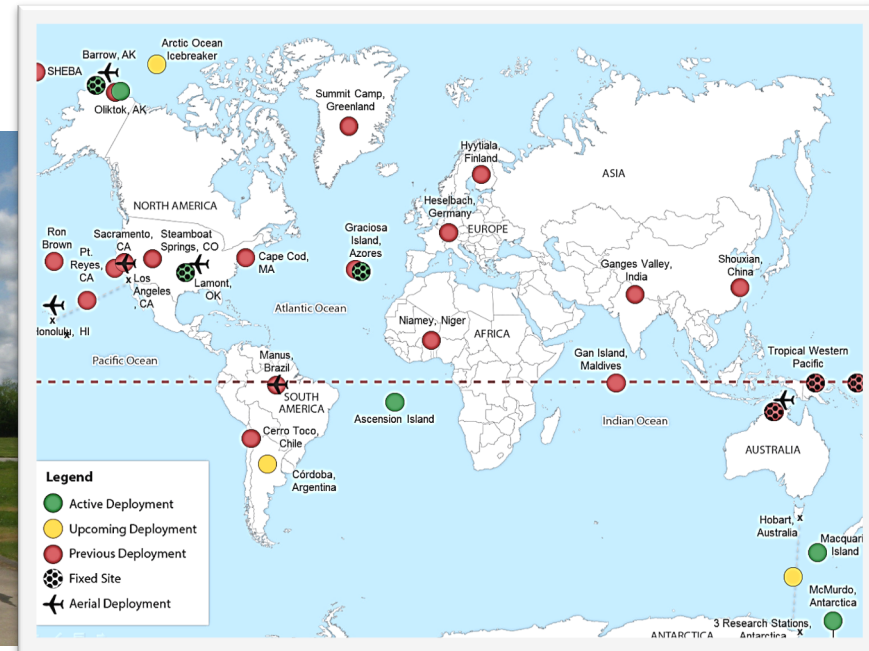
Indústria	Sensores Apointamentos de Produção Controle de Qualidade	Manutenção Preventiva Otimização da produção com base nas Vendas Nível de estocagem
Recrutamento / People Analytics	Histórico de Projetos Tempo de Permanência Referências anteriores Soft Skills Transcrição de Entrevistas	Fit entre funcionário e empresa “Tempo de vida” estimado Plano de Carreira Melhor alocação de recursos
Educação	Histórico de notas Frequência Participação em aula Atividades extra-curriculares	Programas individualizados de ensino Predição de evasão escolar

The Atmospheric Radiation Measurement (ARM) Facility

Data and Computing Management (DoE/USA)

Objetivo do ARM:

fornecer uma detalhada e precisa descrição da atmosfera da terra em diversos regimes climáticos para resolver as incertezas no clima e nos modelos dos sistemas terrestres que direcionam o desenvolvimento de soluções sustentáveis para a Energia e desafios ambientais.



ARM GOAMAZON



<http://www.arm.gov/sites/amf/mao/>
<http://campaign.arm.gov/goamazon2014/>



This view shows the location of the Manacapuru, Brazil, ARM Mobile Facility.



G-1 Research Aircraft in Manaus



AMF1 Site in Manacapuru



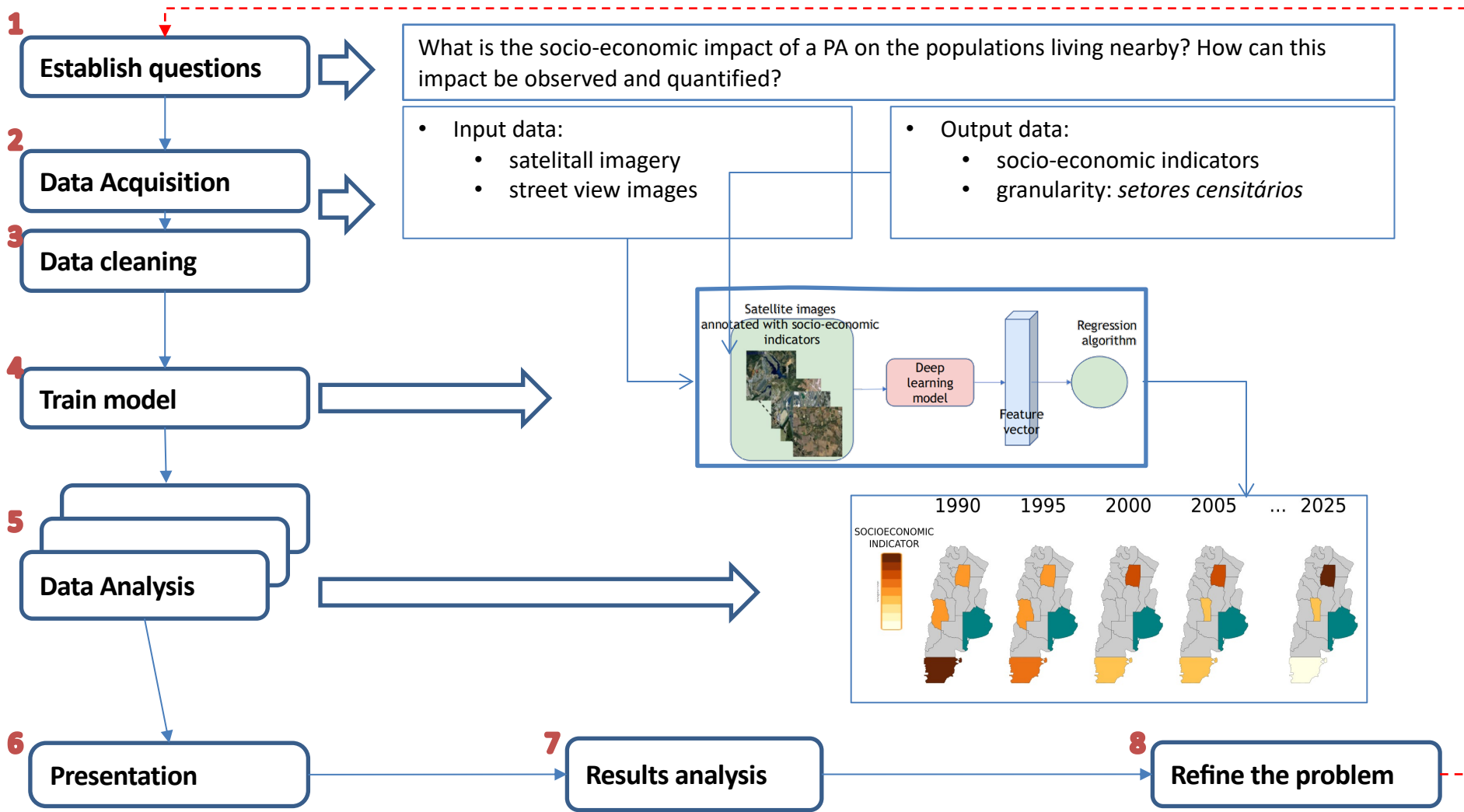
ATTO: Amazon Tall Tower Observatory

Exemplo de aplicações: Projeto avaliação do impacto socioeconômico das Unidades de Conservação

A deep-learning approach to predict socioeconomic indicators using satellite imagery

Case of study: Vale do Ribeira

FAPESP – Brazil, NSF – EUA, France and Japan
<https://parsecproject.org/>



Exemplo de Aplicações

Overview

- Tornar a estimativa de tempo de retorno/normalização no serviço de fornecimento de serviço mais preciso e específico para cada tipo de instalação ou tipo de equipamento.
- Priorizar os clientes através de determinadas regras para que, caso haja alguma falha no fornecimento, o tempo de retorno/normalização sejam comunicados de forma pró-ativa através de RPA (Robot Process Automation).

Exemplo de Aplicações

- Para o tempo de retorno/normalização foram selecionados 5 (cinco) modelos de regressão;
- Para a priorização de clientes foram selecionados 2 (dois) modelos de scoring.

Exemplo de Aplicações

Lições aprendidas

- Aumentar janela de tempo para treinamento dos modelos;
- Aumentar a base de instalações binarizadas de acordo com as ocorrências (reduzir tempo de processamento);
- Notebook de score para re-treino automático de acordo com a performance do modelo em operação.

Ciência dos Dados e Big Data – organizações e sociedade direcionadas por dados

28 de setembro de 2021

Prof. Dr. Pedro Luiz Pizzigatti Corrêa - pedro.correa@usp.br

Departamento de Engenharia de Computação e Sistemas Digitais

Escola Politécnica da Universidade de São Paulo - EPUSP

Grupo de Pesquisa e Extensão em Big Data da EPUSP wds.poli.usp.br

