Educando uma nova geração de cientistas de dados para a gestão de dados de pesquisa

Jian Qin

School of Information Studies
Syracuse University, NY, USA

Escola de Ciência da Informação da Universidade Federal de Minas Gerais
28 de outubro de 2014 – Belo Horizonte, Brasil

Tradução para o portugues (*Translation to Portuguese*):
Mauricio B. Almeida (mba@eci.ufmg.br)
Professor do Programa de Pós Graduação em Ciência da Informação, UFMG

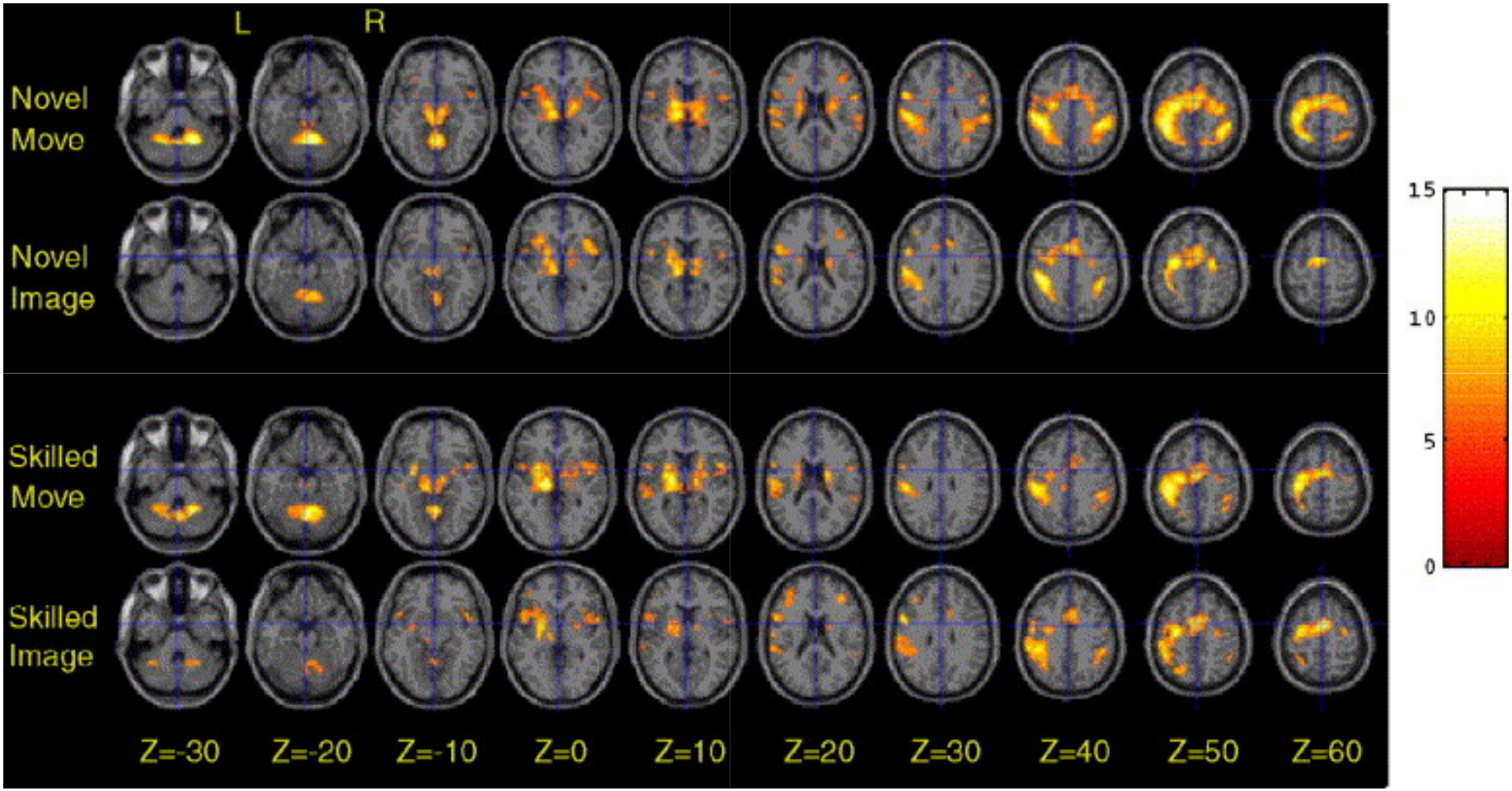
Tópicos

- A Ciência dos Dados (CD) e os cientistas de dados no contexto dos dados de pesquisa
- Uma versão das Escolas de Informação (*i-schools)* para o curriculo de CD
- Descobertas e lições aprendidas com a implementação de um curriculo de CD
- Uma nova geração de cientistas de dados: a abordagem das *i-schools*

Sentimos a pressão de um verdadeiro dilúvio no mundo da informação digital...

http://readwrite.com/2011/11/17/infographic-data-deluge---8-ze

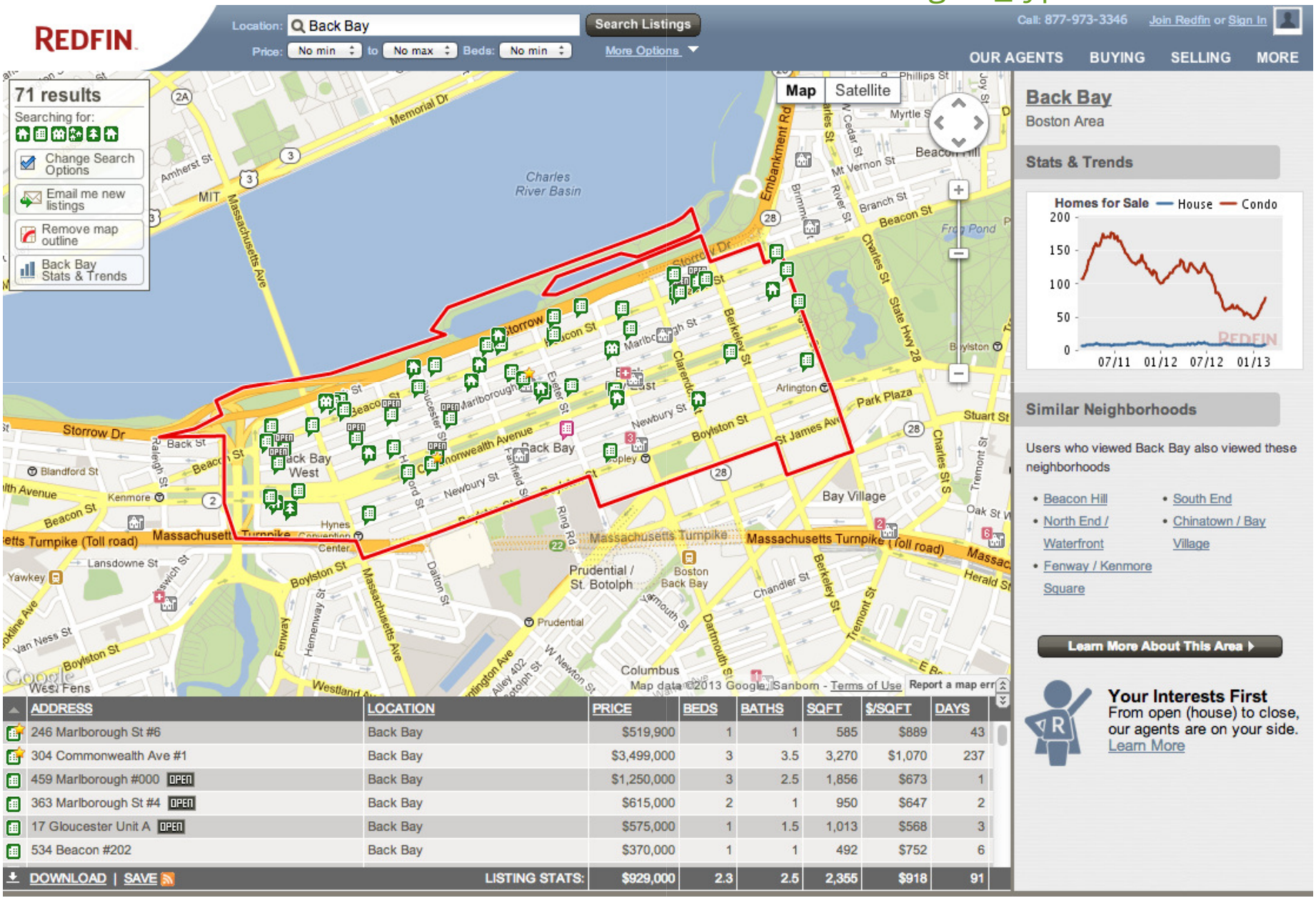
...na medicina clínica



http://ars.els-cdn.com/content/image/1-s2.0-S1053811905002508-gr4.jpg

...em nossa vizinhança

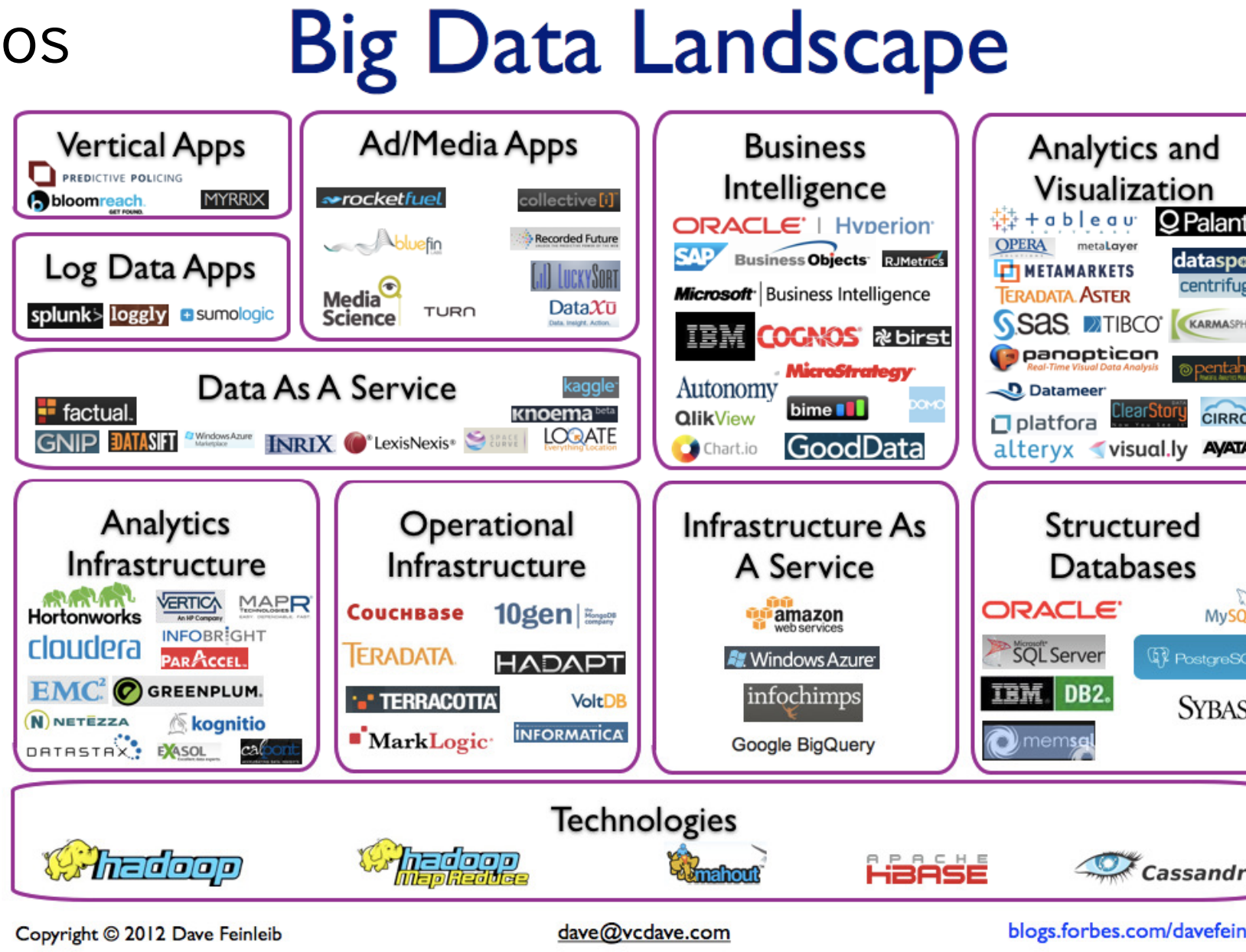
http://www.redfin.com/homes-for-sale#!market=boston&region_id=11&region_type=1&v=8



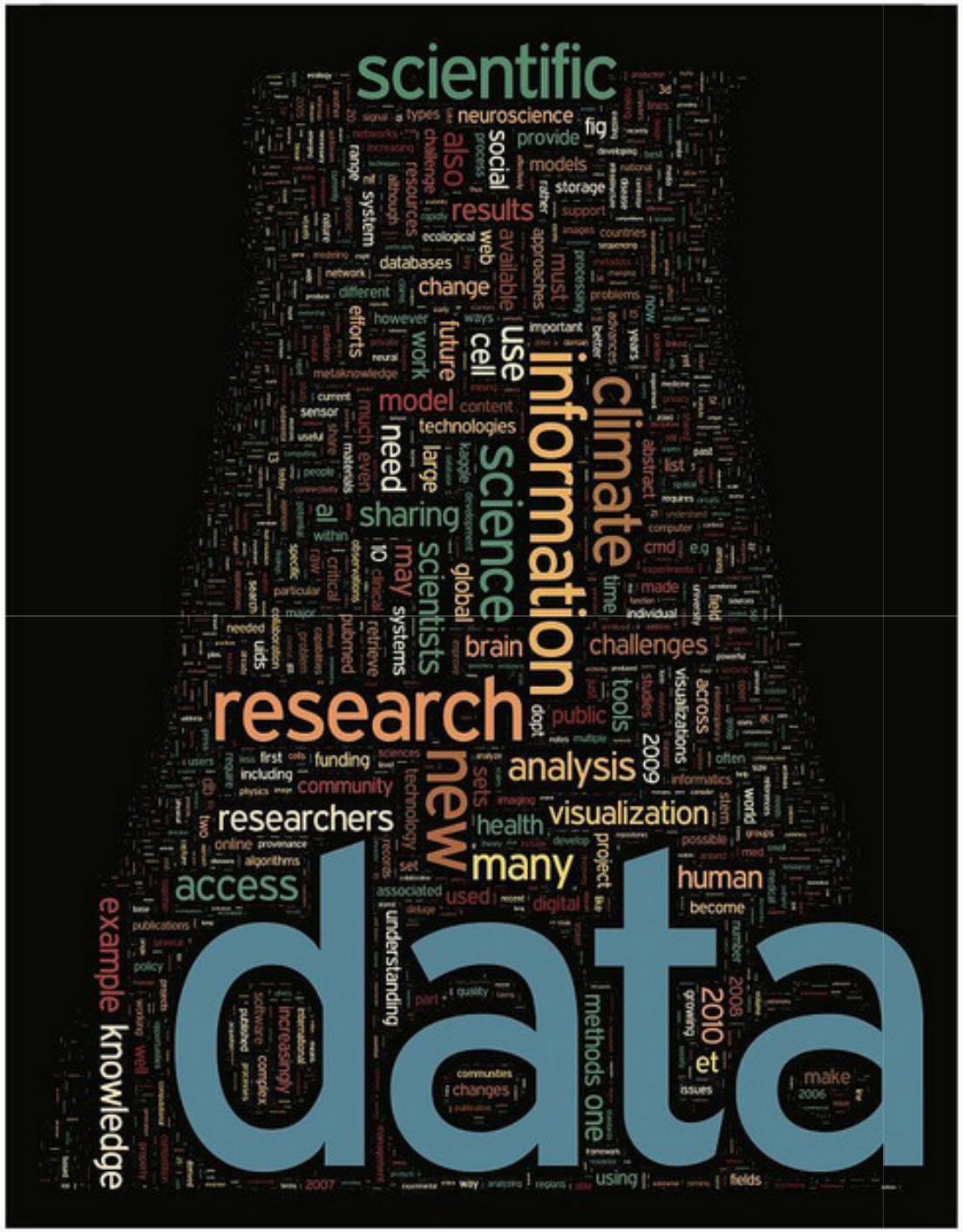
ADDRESS	LOCATION	PRICE	BEDS	BATHS	SQFT	$/SQFT	DAYS
246 Marlborough St #6	Back Bay	$519,900	1	1	585	$889	43
304 Commonwealth Ave #1	Back Bay	$3,499,000	3	3.5	3,270	$1,070	237
459 Marlborough #000 OPEN	Back Bay	$1,250,000	3	2.5	1,856	$673	1
363 Marlborough St #4 OPEN	Back Bay	$615,000	2	1	950	$647	2
17 Gloucester Unit A OPEN	Back Bay	$575,000	1	1.5	1,013	$568	3
534 Beacon #202	Back Bay	$370,000	1	1	492	$752	6
DOWNLOAD | SAVE	LISTING STATS:	$929,000	2.3	2.5	2,355	$918	91

DS

No mundo dos negócios...



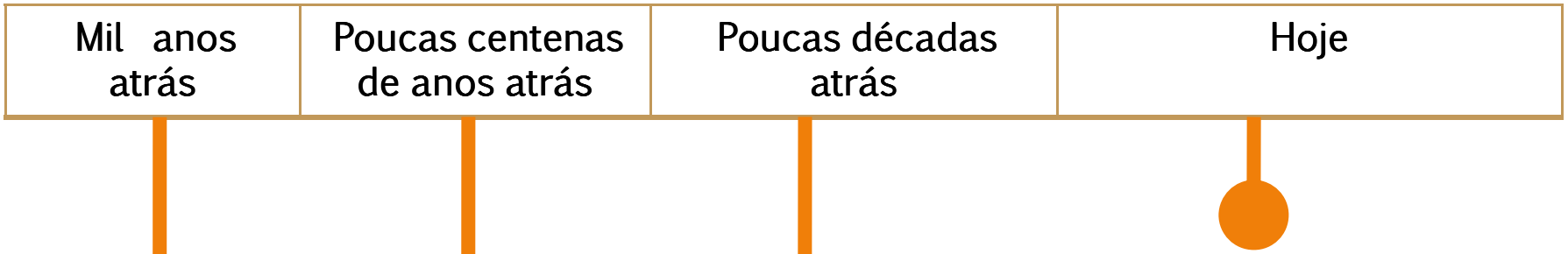
(Feinleib, 2012)



...na pesquisa científica

http://www.sciencemag.org/content/331/6018.cover-expansion

Uma mudança nos Paradigmas Científicos



A Ciência era empírica e descrevia os fenômenos naturais

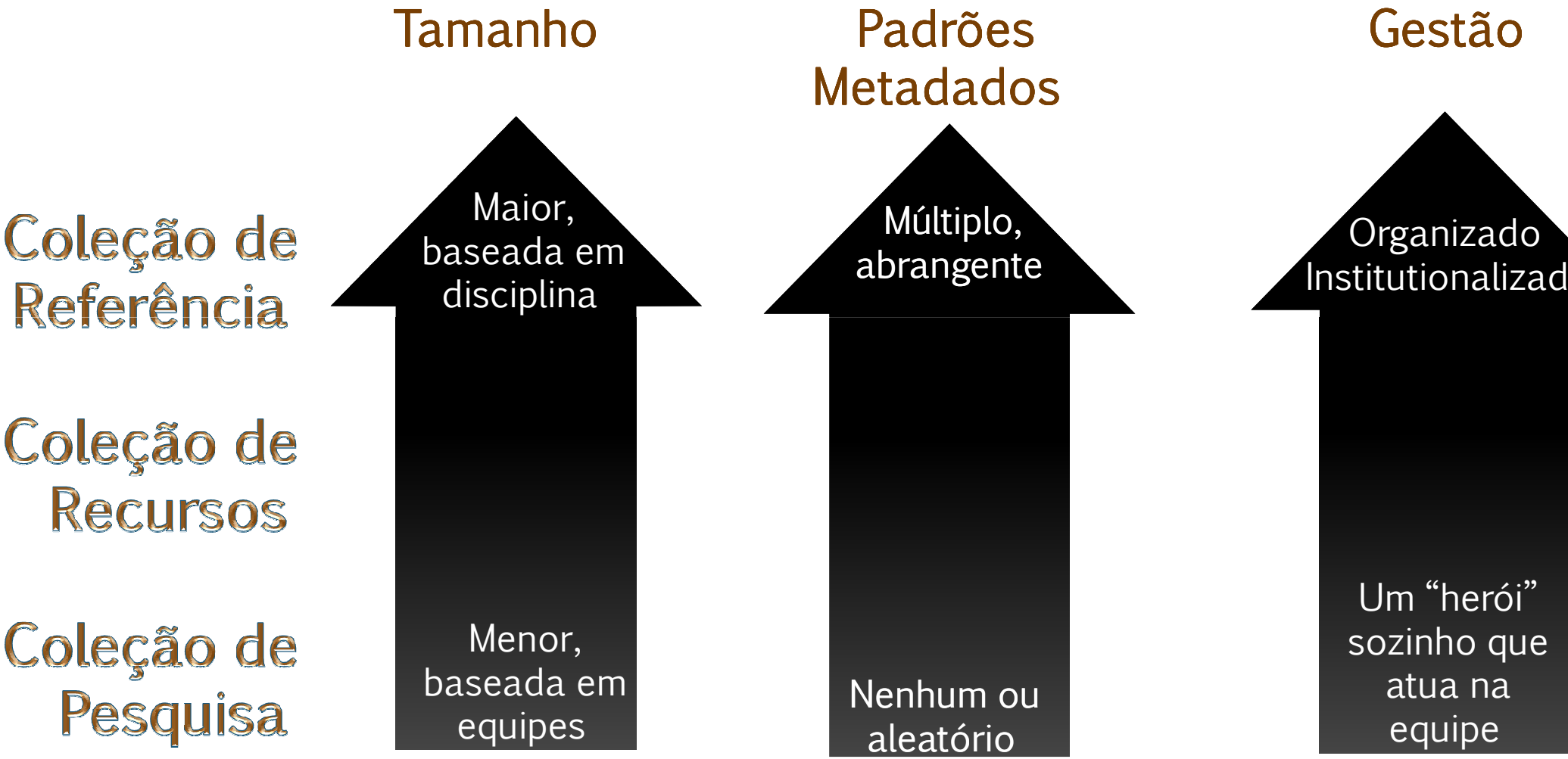
Ramos teóricos, uso de modelos, generalizações

Abordagem computacional , simulação de fenômenos complexos

Exploração de dados (*eScience*), unificação de teoria, experimento, e simulação
-- Dados capturados por instrumentos ou gerados por simuladores
-- Processados por software
-- Informação/Conhecimento armazenados em computadores
-- Cientistas analisam bancos de dados e arquivos usando técnicas de gestão de dados e estatística

Gray, J. & Szalay, A. (2007). eScience – A transformed scientific method. http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

Coleções de dados de pesquisa



Conceitos Emergentes...

...que chegaram para ficar e que são relevantes para sua carreira!

eScience

A Gestão de Dados é essencial

entific Data Management cialist

esign, develop, implement, and anage high-throughput automatic ata processing infrastructure for rge databases in a mature system
evelop and improve the frastructure supporting this stem
terface with multiple data oviders to design, build, and aintain their customized atabases
arify requirements, feature quests and bug reports for oftware developers and assist in sting code.

Laboratory Data Management Specialist

- **Administer** operational database
- **Assure** the quality of data database content
- **Interact** closely with researchers, lab managers, and platform coordinators
- **Track** deliverables against budget and prepare data reports
- **Collaborate** closely with IT and bioinformatics colleagues
- **Assist** IT in gathering workflow requirements
- **Test** changes and updates in IT systems
- **Create and maintain** app documentation

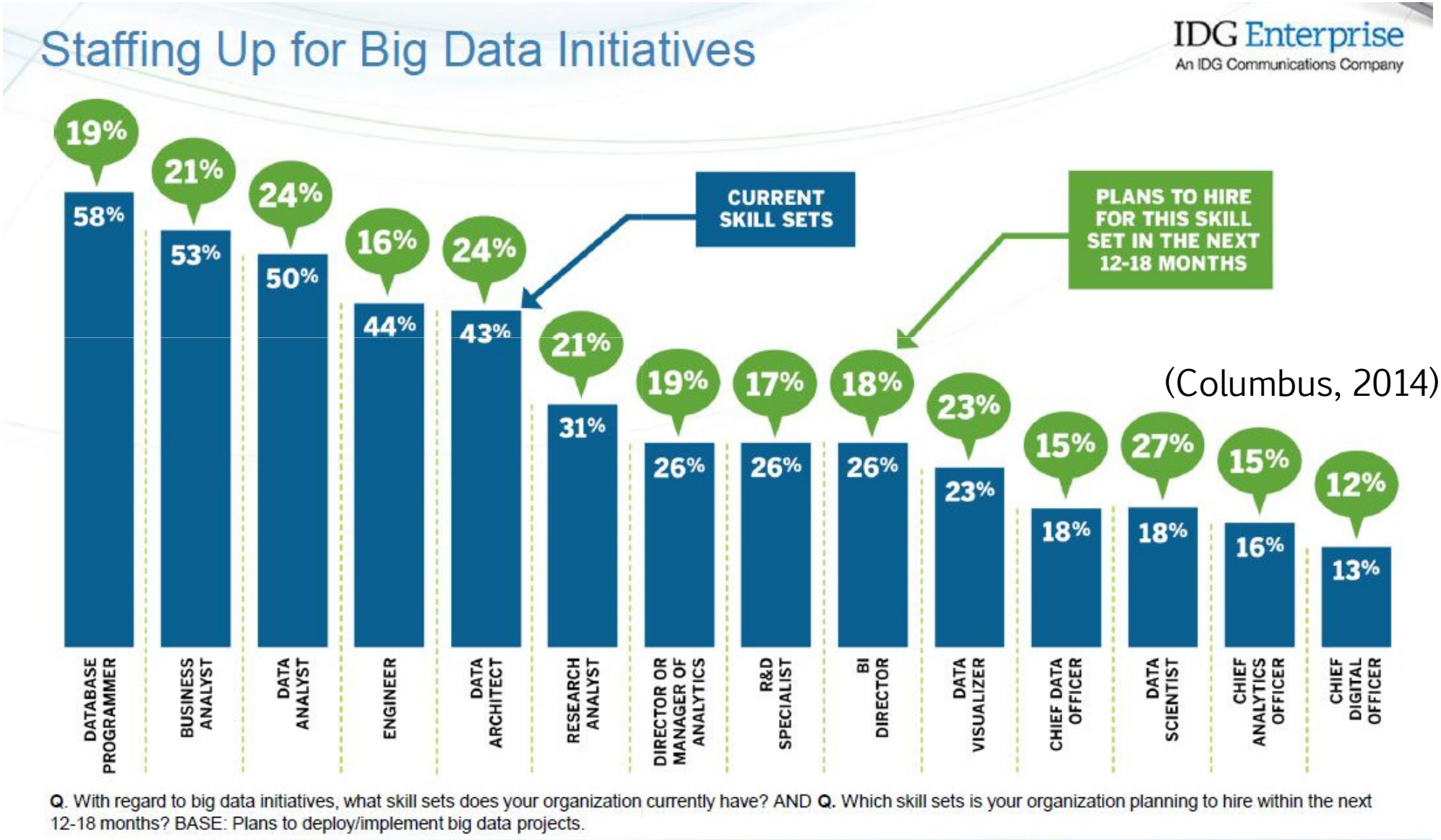
Data Modeling/ Management Specialis

- **Work** closely with the high performance computing a the IT manager
- **Develop** a data model for complex multi-scale rocks
- **Design and organize** a database and complex queries
- **Integrate and mange** mult scale rocks subjected to large-scale scientific computing applications

http://www.ingrainrocks.co m/data-management-specialist/

http://www.bioinformatics.org/forums/forum.php?forum_id=9670

Demandas de mercado por habilidades deCD



(Columbus, 2014)

Ciência dos Dados

O que é isso ?

O que isso significa para os futuros encarregados em fornecer ensino e empregabilidade aos Cientistas da Informação?

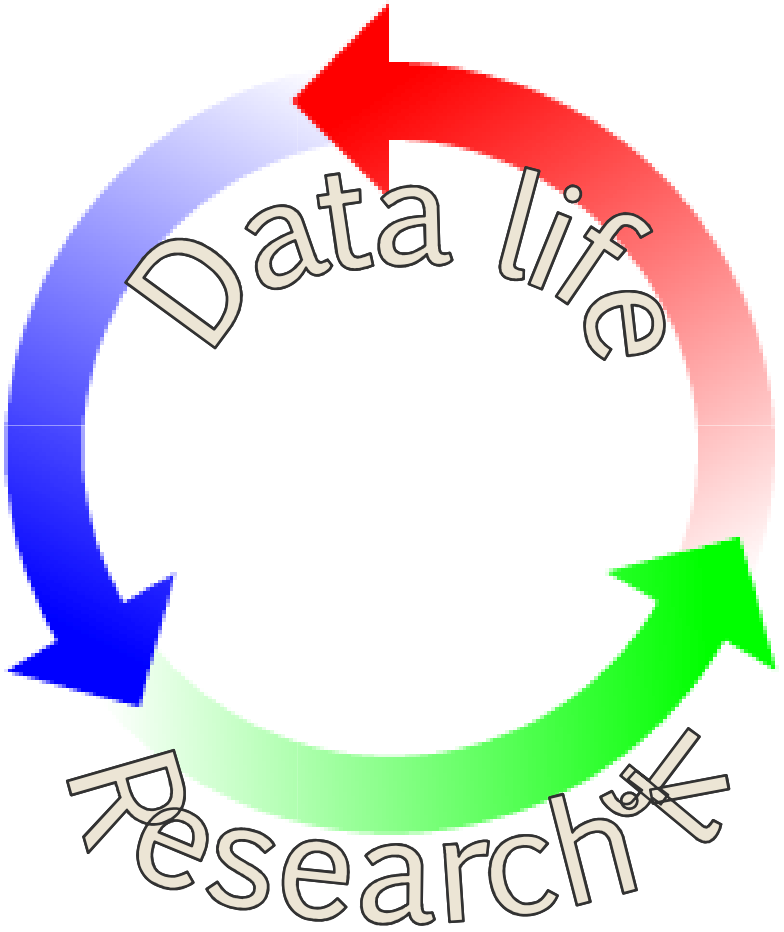
O que é a Ciência dos Dados?

"Uma área de trabalho emergente que diz respeito à coleta, apresentação, análise, visualização, gestão e preservação de grandes coleções de informação"

Stanton, J. (2012). Introduction to Data Science. http://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf

Ciência dos Dados e pesquisa científica

Planejar, projetar, executar, implantar e avaliar projetos e serviços de gestão de dados



Receber, armazena organizar, combina filtrar, e transforma dados , além de criar formatos de dados adequados para análise

DS

“Cientistas de dados são pessoas capazes de ̀pescaŕ respostas para importantes questões de negócio em meio ao tsunami de informação não estruturada disponível nos dias de hoje”

Davenport, T.H. & Patil, D.J. (October 2012). Data scientist: the sexiest job of the 21st century. Harvard Business Review, 70-76.

Como os programas educacionais devem abordar esse desafio?

Um *case* do curso de Certificado de Estudos Avançados (C
do programa de Ciência dos Dados da *Syracuse iSchool*


DATA SCIENCE AT THE iSCHOOL

Desenvolvimento de um programa para Ciência dos Dados

DS

ool of Information Studies
RACUSE UNIVERSITY

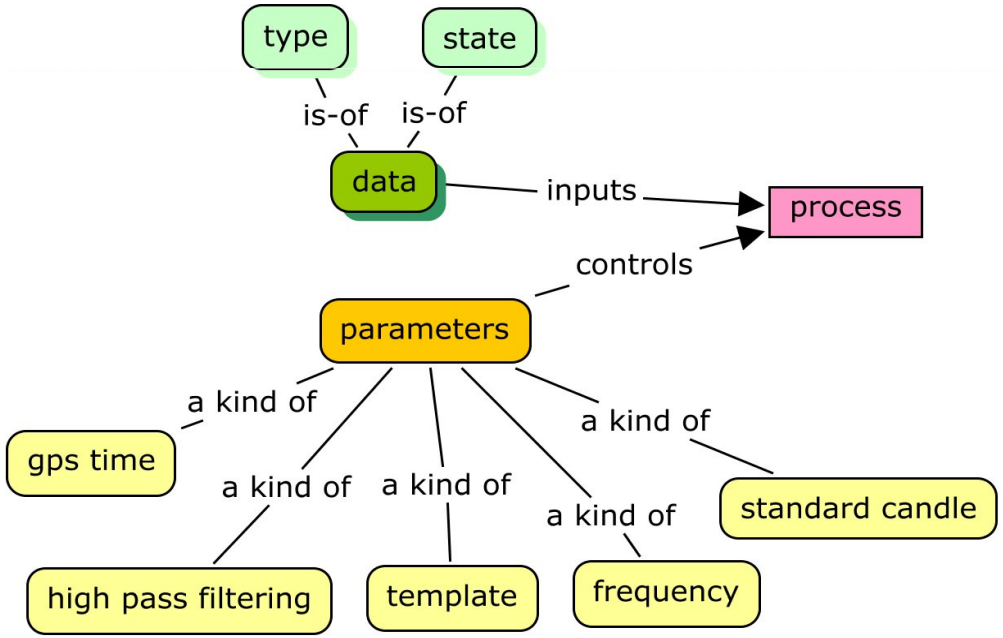
- Facilitadores: Arquitetos de Informação atuando em disciplinas específicas (*STEM disciplines*)
- Projeto de desenvolvimento de um currículo de Letramento em Dados Ciêntificos : http://sdl.syr.edu/
- Biblioteconomia *eScience*: Educação e Treinamento http://eslib.ischool.syr.edu/wp/
- CEA em Ciência dos Dados http://ischool.syr.edu/future/cas/datascience.aspx

Case 1: workflows e gestão de dados com demandas cognitivas

- *Domínio* Termocronologia e terremotos
- *O que está envolvido*: amostras de rochas de perfurações e amostras de rochas cortadas, granuladas e observadas em campo
- *Tipos de dados*: arquivos de dados do Excel (centenas) deles, imagens espectrais e microscópicas, anotações
- *Análise:* modelagem e construção de sentido pela combinação dos dados de diversos arquivos com o uso de software especializado
- *Gargalo:* correspondência, mesclagem e filtragem manual dos dados é extremamente pesada e o problema se torna critico pela dificuldade em achar os arquivos de dados corretos

Caso 2: *workflows* altamente automatizados

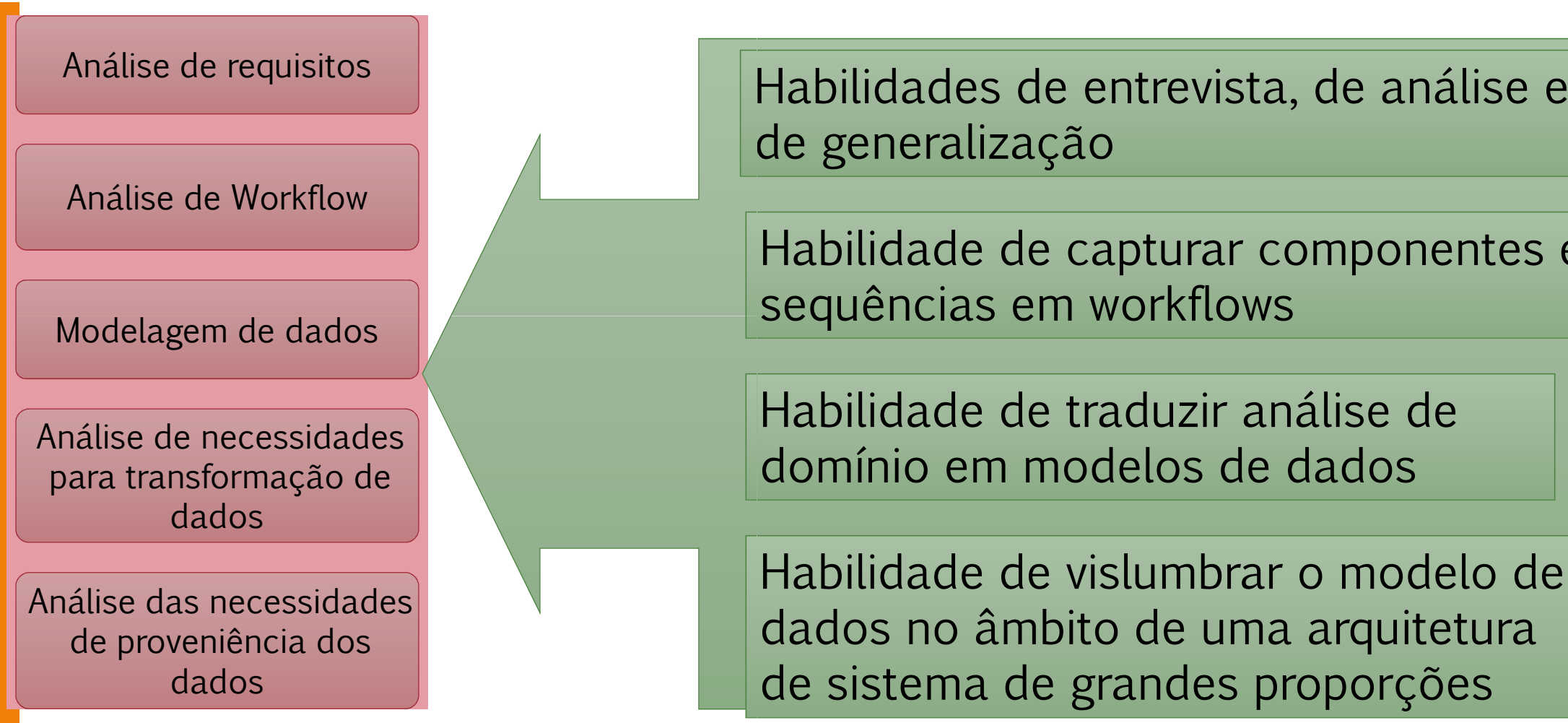
- *Domínio* Astrofísica: detecção de ondas gravitacionais
- *O que está envolvido*: entrada de dados de interferômetros à laser, calibragem de dados brutos, bem como segmentação, *workflow* e gestão da proveniência
- *Tipos de dados*: dados transmitidos a partir de interferômetros, imagens
- *Análise*: detecção de "eventos"
- *Gargâlo:* rastrear dados e processos, bem como o relacionamento entre eles



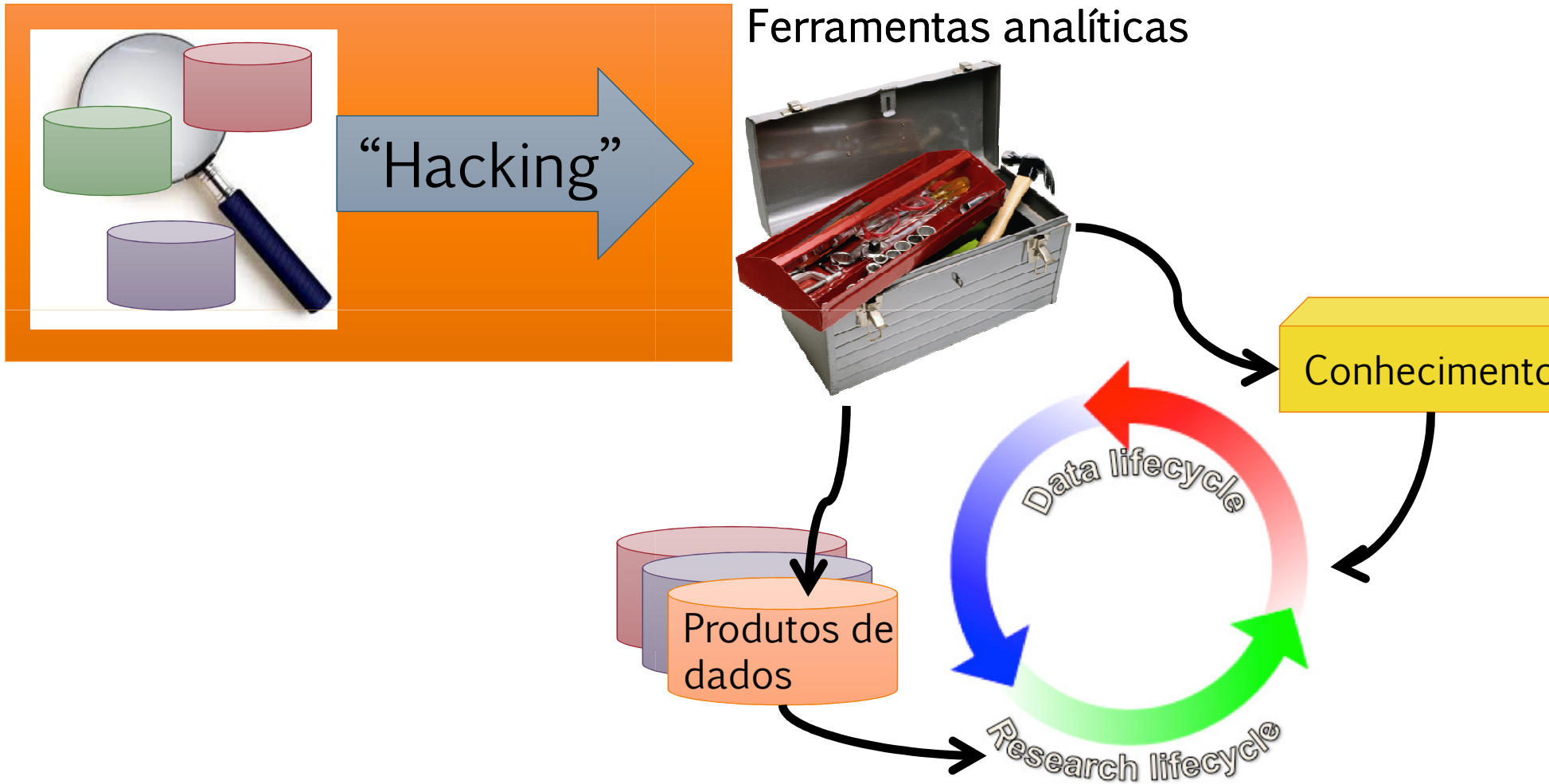
O que se espera dos Cientistas de Dados?

- Habilidade de usar ampla variedade de ferramentas para documentação, análise e relatório de dados
- Conhecimento de um domínio do conhecimento
- Modelagem de dados, projeto e consulta à bancos de dados
- Colaboração, comunicação, e co-ordenação
- Sistemas operacionais, Linguagens de programação
- Sistemas de contéudo e repositório
- Linguagens de codificação

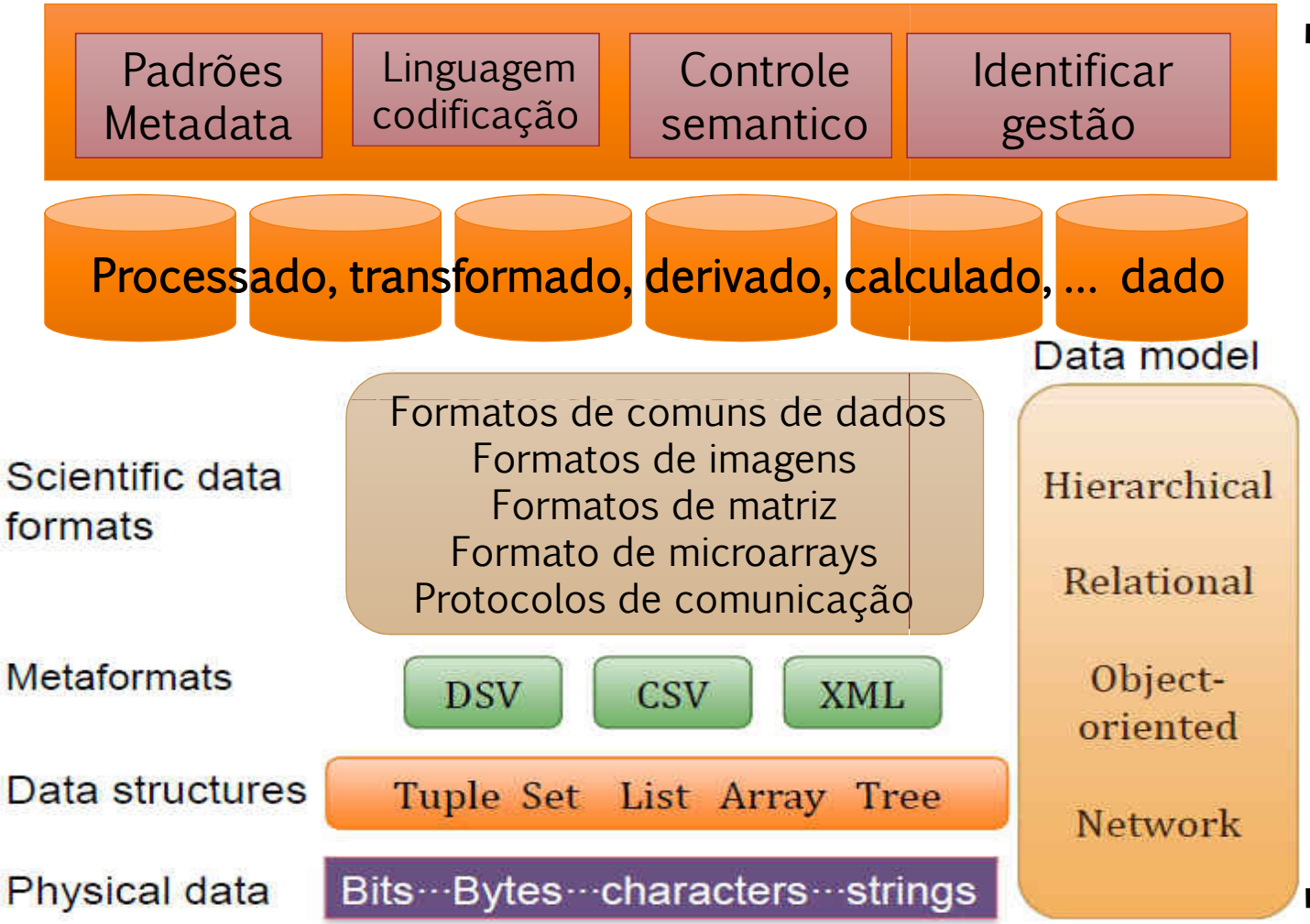
Analytical skills: domain modeling



Habilidades analíticas: das fontes de dados até os padrões, os relacionamentos e as tendências



Habilidades para gestão de dados: ciclo de vida de dados e serviços de infraestrutura



Serviçoes de infrastrutura

- Descobrerta de fontes de dados
- Curadoria de dado
- Preservation de dados
- Integration de dados
- Citação, publicaçã e distribuição de dados
- Conexão e interoperabilidade de dados

Habilidades tecnológicas aliadas a habilidades de comunicação bem desenvolvidas

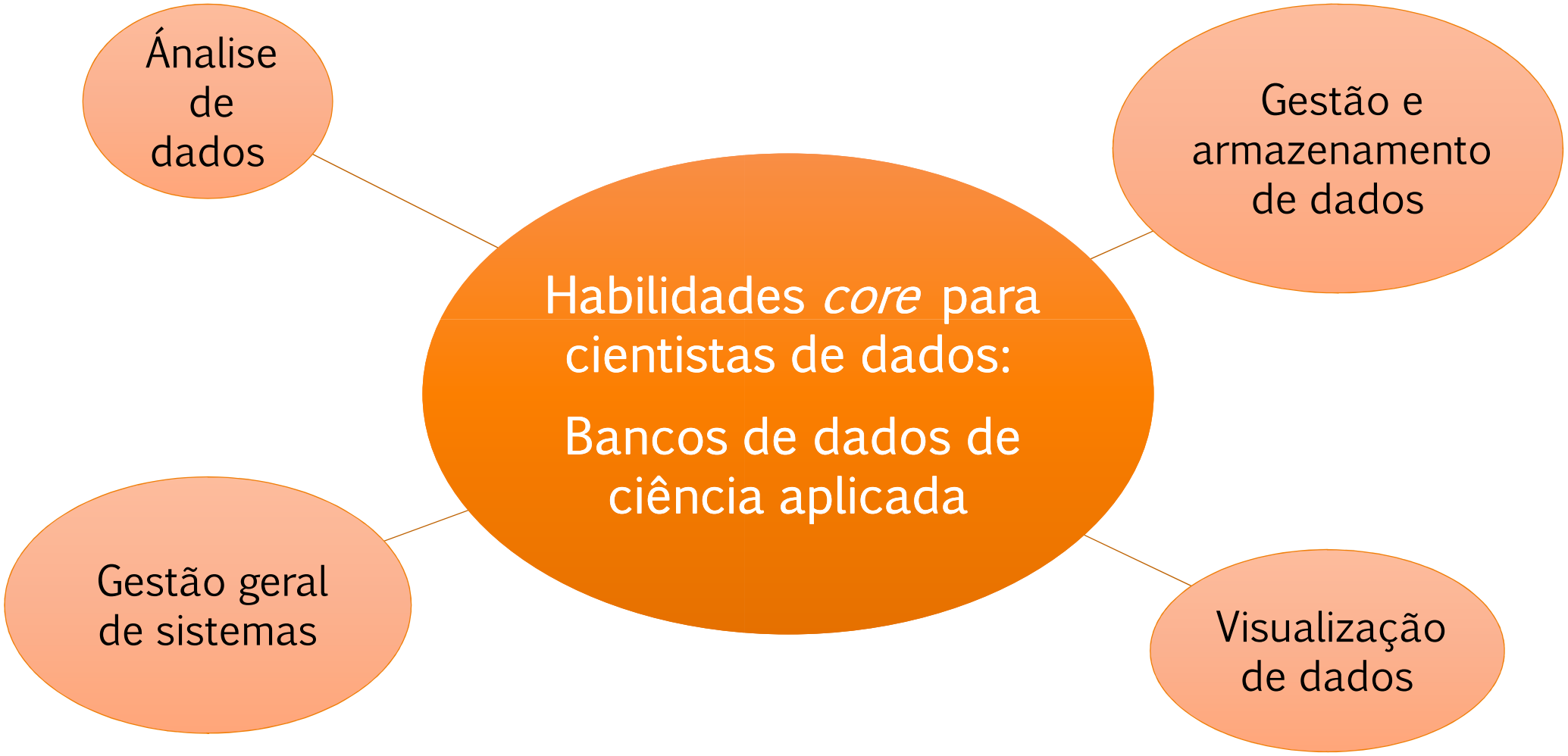
HABILIDADES TECNOLÓGICAS

- Sistemas operacionais
- Sistemas de repositórios
- Sistemas de bancos de dados
- Linguagens de programação
- Linguagens de codificação
- Programação especializada

HABILIDADES DE COMUNICAÇÃO

- Entrevistas
- “Quebra gelo”
- Construção de comunidades
- Institucionalização
- Envolvimento dos interessados

Não existe um modelo de "superhomem" para os cientistas de dados



O CEA em Ciência dos Dados no programa da *Syracuse University*

› Obrigatórias:
  - Conceitos de administração de dados e gestão de banco de dados
  - Ciência dos Dados aplicada

› Eletivas:

Análise de dados
- Mineração de dados
- Princípios de sistemas de recuperação da informação
- Processamento de linguagem natural
- Análise de informação avançada
- Métodos de Pesquisa
- Métodos estatísticos

Data Storage and Management
- Tecnologias para gestão de conteúdos na Web
- Fundamentos da criação, gestão e preservação de ativos digitais
- Armazém de dados
- Gestão de banco de dados avançada

Visualização de dados
- Arquitetura da Informação para serviços de informação
- Visualização da Informação

Gestã de sistemas de Informação
- Tecnologias organizacionais
- Projetos para gestão de sistemas de informação
- Análise de sistemas de informaçã

O que aprendemos com o desenvolvimento do programa?

- A Ciência dos Dados é dinâmica = diversos pontos focais
  - Versões vem da estatística, ciência da computação, e Biblioteconomia e Ciência da Informação
- Habilidades vs. teorias
  - Estudantes são ansiosos para adquirir hailidades ,mas não tão interessados em teorias
  - Teorias ajudam a construir visões
- Tempo de envolvimento suficiente para tecnologia e ferramentas
- Aprendizado via projetos de gestão de dados em cenários reais

Reconciliação das duas visões da Ciência dos Dados

"Uma área de trabalho emergente que diz respeito à coleta, apresentação, análise, visualização, gestão e preservação de grandes coleções de informação."

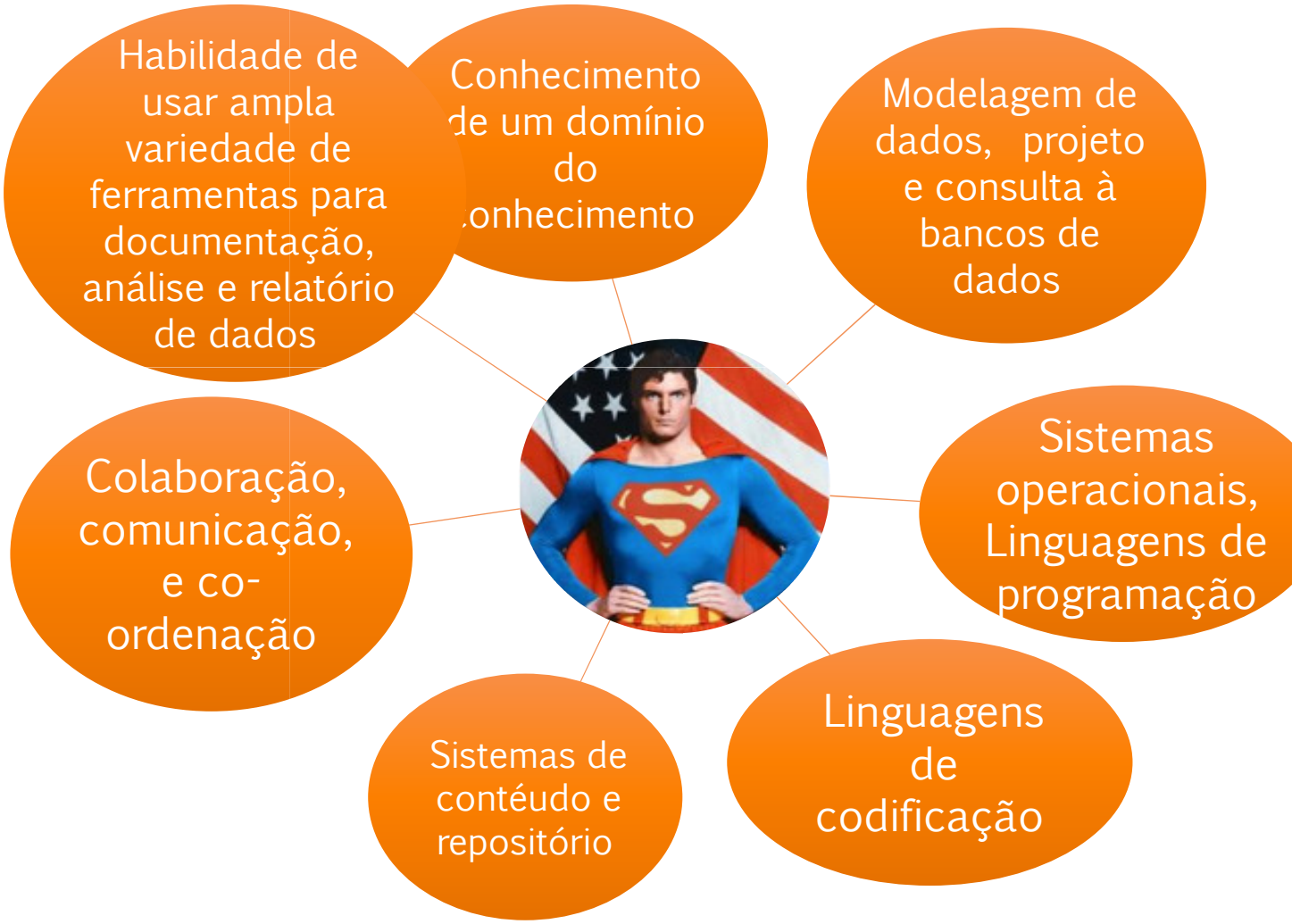
Stanton, J. (2012). Introduction to Data Science. http://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf

"Cada vez mais, encontramos dados em contextos diversos, e os cientistas de dados estão envolvidos em agrupá-los e dispô-los em um formato tratável, fazendo com que contem sua estória, a qual será apresentada à outras pessoas."

Loukides, M. (2011). What is data science? Sebastopol, CA: O'Reilly.

A versão *iSchool* para a Ciência dos Dados

Em última instância, o programa de Ciência dos Dados da *iSchool* será construído sobre uma fundação para super cientistas de dados...



DS

Projeto do currículo de Biblioteconomia e-science: http://eslib.ischool.syr.edu/

Projeto para letramento em Ciência dos Dados: http://sdl.syr.edu/

CAE em Ciência dos Dados http://ischool.syr.edu/future/cas/datascience.aspx

Referências

- Columbus, L. (2014). 2014: The year Big Data adoption goes mainstream in the enterprise. Forbes, http://www.forbes.com/sites/louiscolumbus/2014/01/12/2014-the-year-big-data-adoption-goes-mainstream-in-the-enterprise/
- Feinleib, D. (2012). The Big Data landscape. Forbes, http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/