Big Data e Data Science

PEDRO COSTA FERREIRA

20 Mind-Boggling Facts

More data has been created in the past two years than in the entire previous history of the human race

By 2020, we will have over 6.1 billion smartphone users globally. That's more than traditional landline accounts.

In Aug 2015, over 1 billion people used Facebook in a

Within five A

connecte

Less than 0.5% of all data is ever analyzed and used, just imagine the potential here.

Still skeptical about the shari

valued at around $25.5 billion

giants Marriott ($20.90 billion

and Wyndham ($10.01 billion). Hilton Worldwide is valued at $27.7 billion. What's more, Airbnb don't own a single hotel room.

is

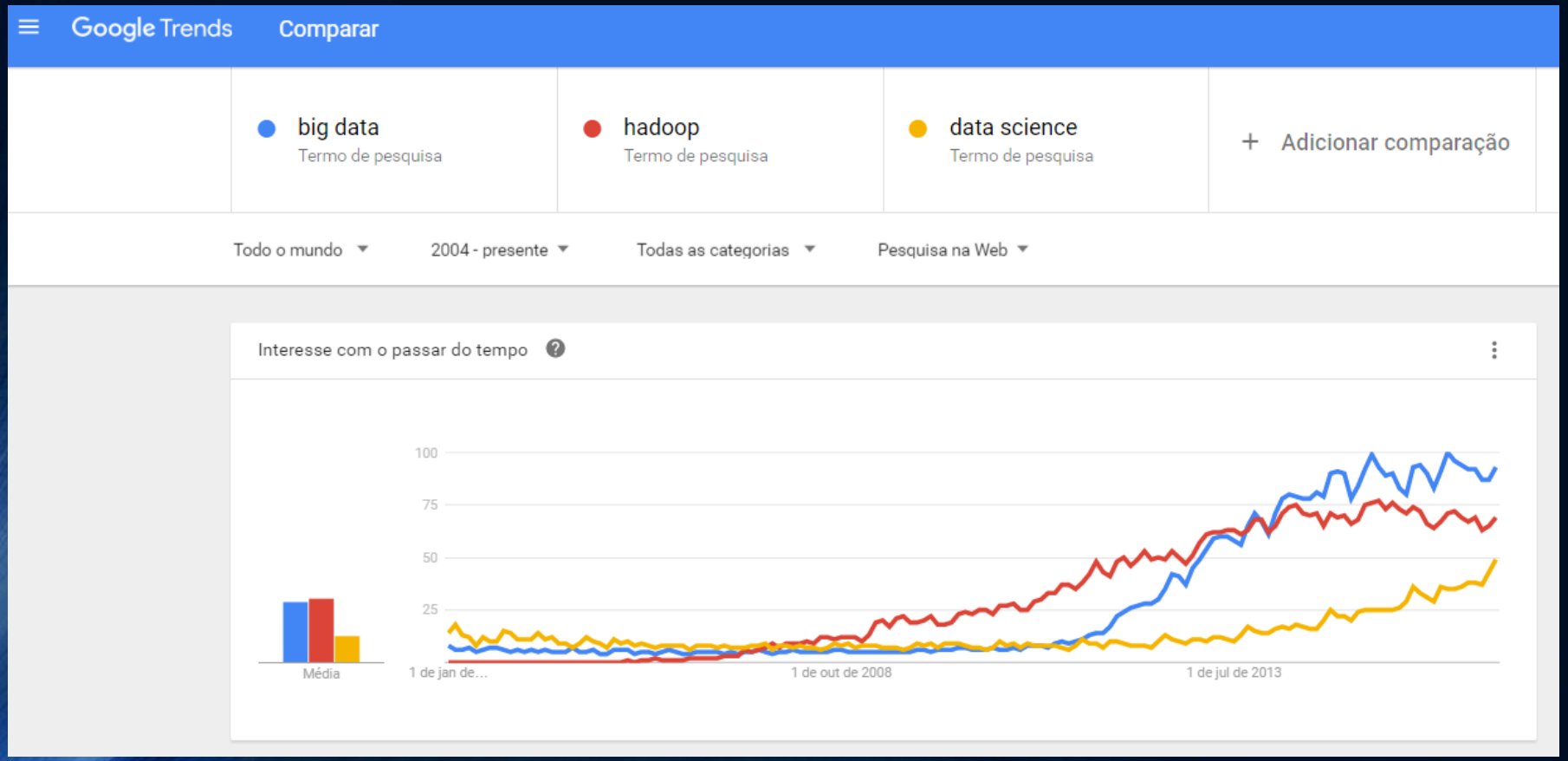
valuation of Kodak we could find put it at around $900 million.

tes of new

nd for every human being on the planet.

Esses fatos tem despertado o interesse das pessoas?

Esses fatos tem despertado o interesse das pessoas?



Esses fatos tem despertado o interesse das pessoas?

O engraçado é que ainda não conhecemos muito bem o tema, mas o nome já ficou meio chato!!

Mas, o que é Big Data????

"Nós estamos testemunhando um movimento que irá transformar completamente qualquer negócio e a sociedade. O nome que nós damos a esse movimento é **Big Data** e irá mudar tudo, a maneira que banco e varejistas operam, a forma que tratamos o câncer e protegemos o mundo contra o terrorismo. Não importa qual o trabalho que você está fazendo ou a indústria que você trabalha, Big Data irá transformá-lo"

Bernard Marr, 2016

Beleza Bernard Marr, mas algumas questões permanecem...

a) Como evoluímos para a ciência dos dados?

b) Qual é a diferença entre o estatístico/BI e o cientista de dados?

c) O que faz um cientista de dados?

d) O que eu estou fazendo aqui nesse curso???

A evolução para a ciência dos dados

Valor adicionado

KPI reporting

Exploração visual dos dados

- Entender os parâmetros do negócio
- Visualizar transações
- Detectar anomalias
- Visualizar relações

Segmentação

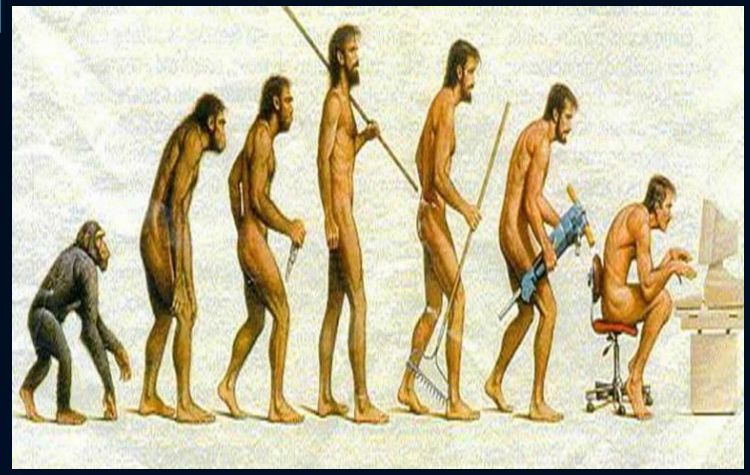
- Entender grupos e *outliers*
- Descobrir similaridades (quem está perto de quem)

Modelos Preditivos

- Prever e analisar resultados futuros
- Modelar e entender relações e causalidades

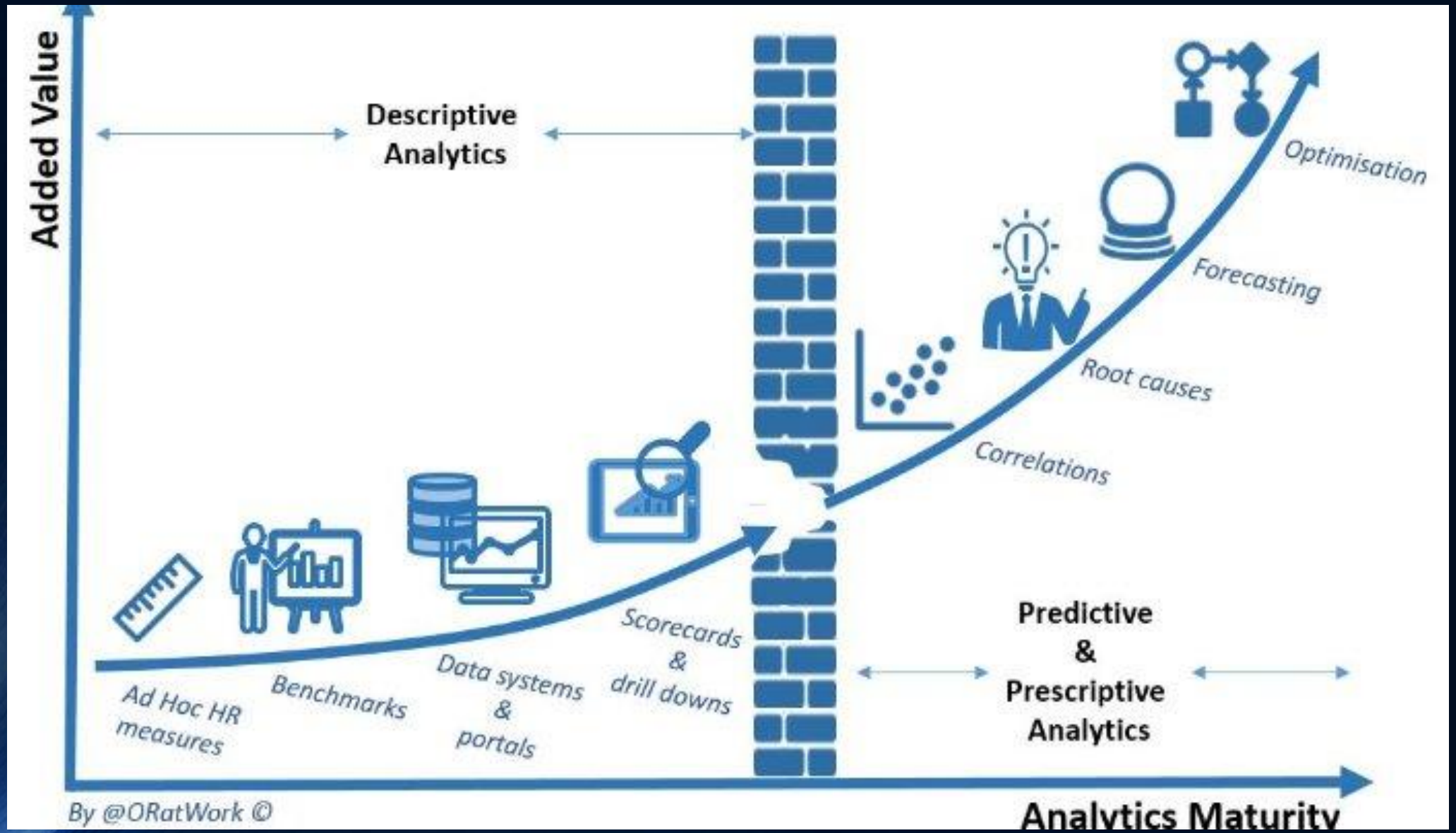
Simulação e Otimização

- Simular e experimentar possíveis cenários
- Encontrar a possível solução em muitas
- Entender o negócio



Sofisticação Analítica

A evolução para a ciência dos dados



b) Qual é a diferença entre o profissional de BI e o cientista de dados?

Característica	Business Intelligence	Cientista de dados
Perspectiva	Olha para o passado	Olha para o futuro
Ações	*Slice and dice*	interativo
ferramentas	SAP, SAS, Microstrategy	R, QlikView, Phython, Hadoop
Dados	Warehoused, Siloed	Distribuidos, real time
Escopo	ilimitado	Questões específicas do negócio
Resultado	tabelas	repostas
Questões	O aconteceu?	O que vai acontecer? E, se?

b) Qual é a diferença entre o estatístico e o cientista de dados?

Característica	Estatístico	Cientista de dados
Modo de agir	Reativo	consultivo
trabalho	*solo*	time
inputs	Data file, hipóteses	Problema do negócio
ferramentas	SAS, Minitab, SPSS	R, Phython, Hadoop
resultados	tabelas	Visualização de dados, previsões
estrelas	G.E.P Box	Nate Silver
Questões	O aconteceu?	O que vai acontecer? E, se?

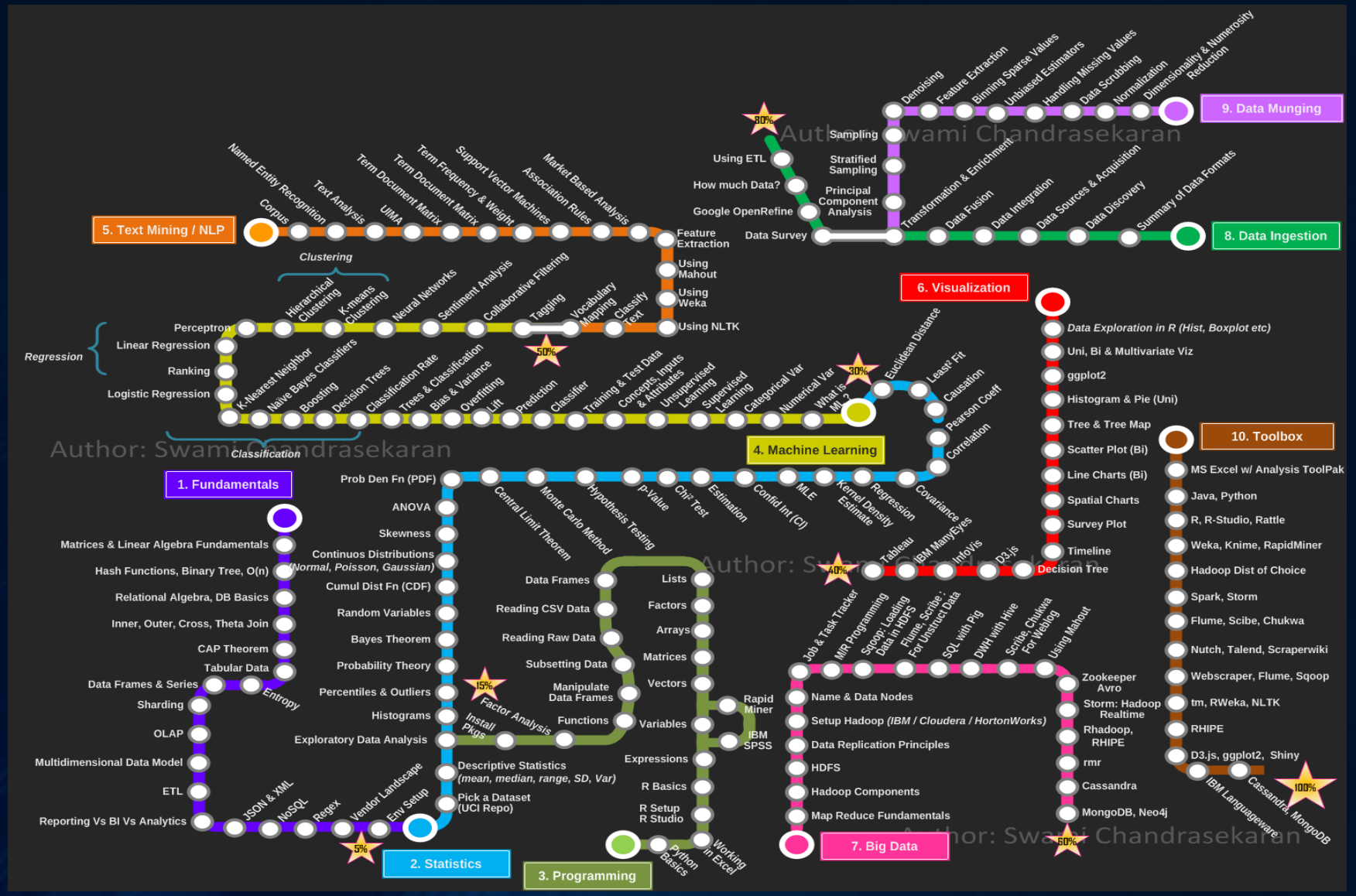
c) O que pode fazer um cientista de dados?

Os Cientistas de Dados não possuem a mesma formação e conjunto de habilidades. Ou seja, os profissionais de dados diferem em relação às competências que possuem.

Por exemplo, alguns profissionais são proficientes em habilidades estatísticas e matemáticas, enquanto outros são proficientes em habilidades de ciência da computação. Outros ainda têm uma forte visão de negócios, enquanto outros são mais focados em desenvolvimento de produtos.

Área	Habilidade
Business	Design e Desenvolvimento de Produto
	Gestão de Projetos
	Desenvolvimento de Negócios
	Governança e Compliance
	Finanças
Tecnologia	Gestão de dados estruturados (RDBMS, SQL, XML)
	Gestão de dados não-estruturados (Bancos de dados NoSQL)
	Processamento de Linguagem Natural (NLP)
	Machine Learning (árvores de decisão, redes neurais, clustering)
	Big Data (Hadoop, MapReduce, Spark)
Matemática e Modelagem	Otimização
	Matemática
	Modelos gráficos
	Algoritmos
	Estatística Bayesiana
Programação e Administração de Sistemas	Administração de Sistemas
	Administração de Banco de Dados
	Cloud
	Programação Back-end
	Programação Front-end
Estatística	Gestão de dados
	Data Mining e Visualização
	Modelagem estatística
	Design de experimentos
	Comunicação

c) O que faz um cientista de dados?



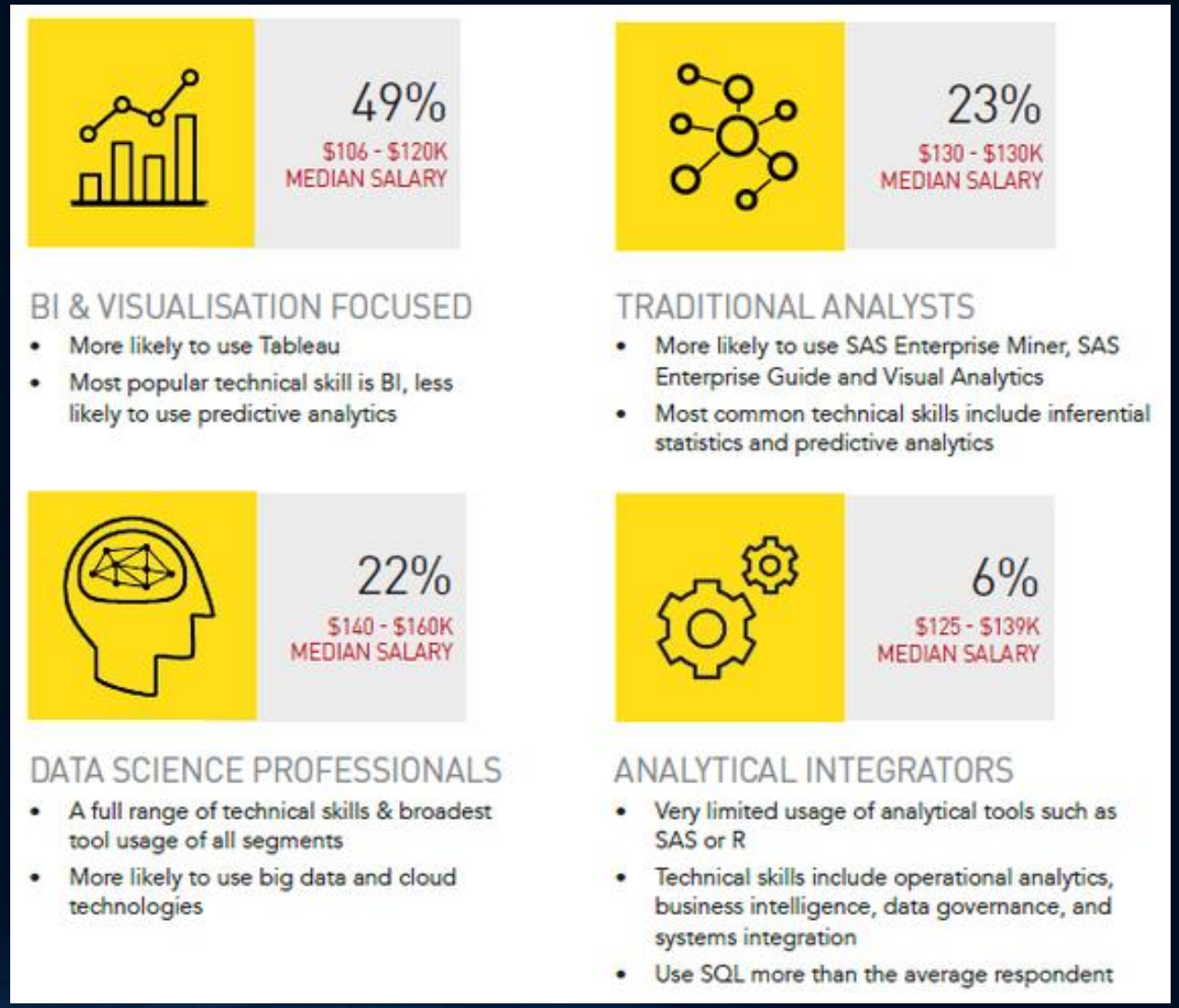
c) O que faz um cientista de dados?

Walmart

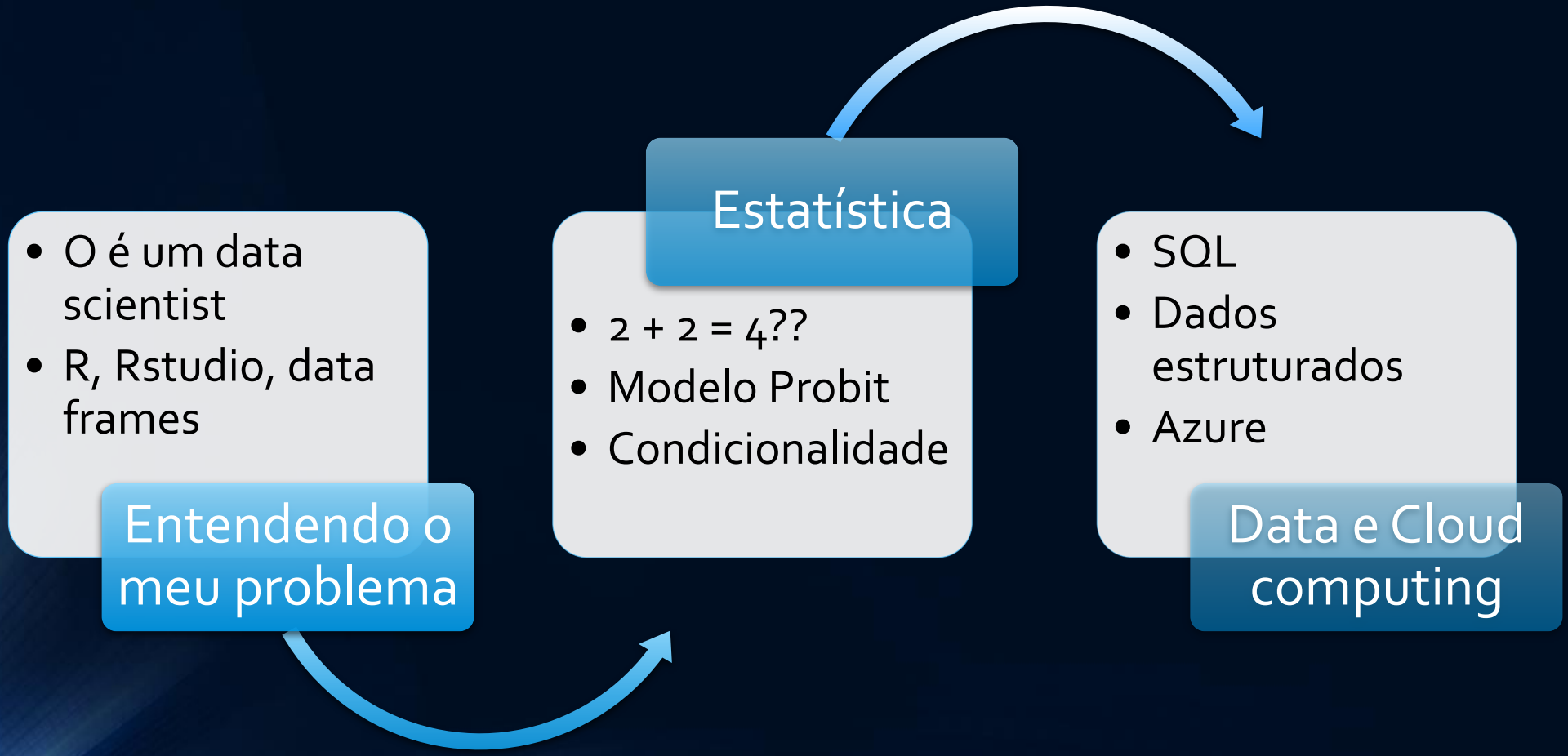
How Big Data is used to drive supermarket performance

- Walmart é a maior rede varejista do mundo, com 2 milhões de empregados e 20.000 lojas em 28 países;
- Em 2004, quando o furacão Sandy atingiu a costa dos EUA, eles descobriram que insights inesperados poderiam surgir quando os dados são estudados como um todo, mais que quando o individuo é estudado individualmente;
- Com o objetivo de atender a demanda por materiais de emergência em face a aproximação do furacão Sandy algumas surpresas estatísticas emergiram;
- Além dos materiais de emergência, observaram que a venda do produto **Strawberry Pop Tart** aumentou consideravelmente em algumas localidades;
- Em 2012, com a aproximação do furacão France's o Walmart aumentou o estoque desse produto em diversas unidades e as vendas explodiram.

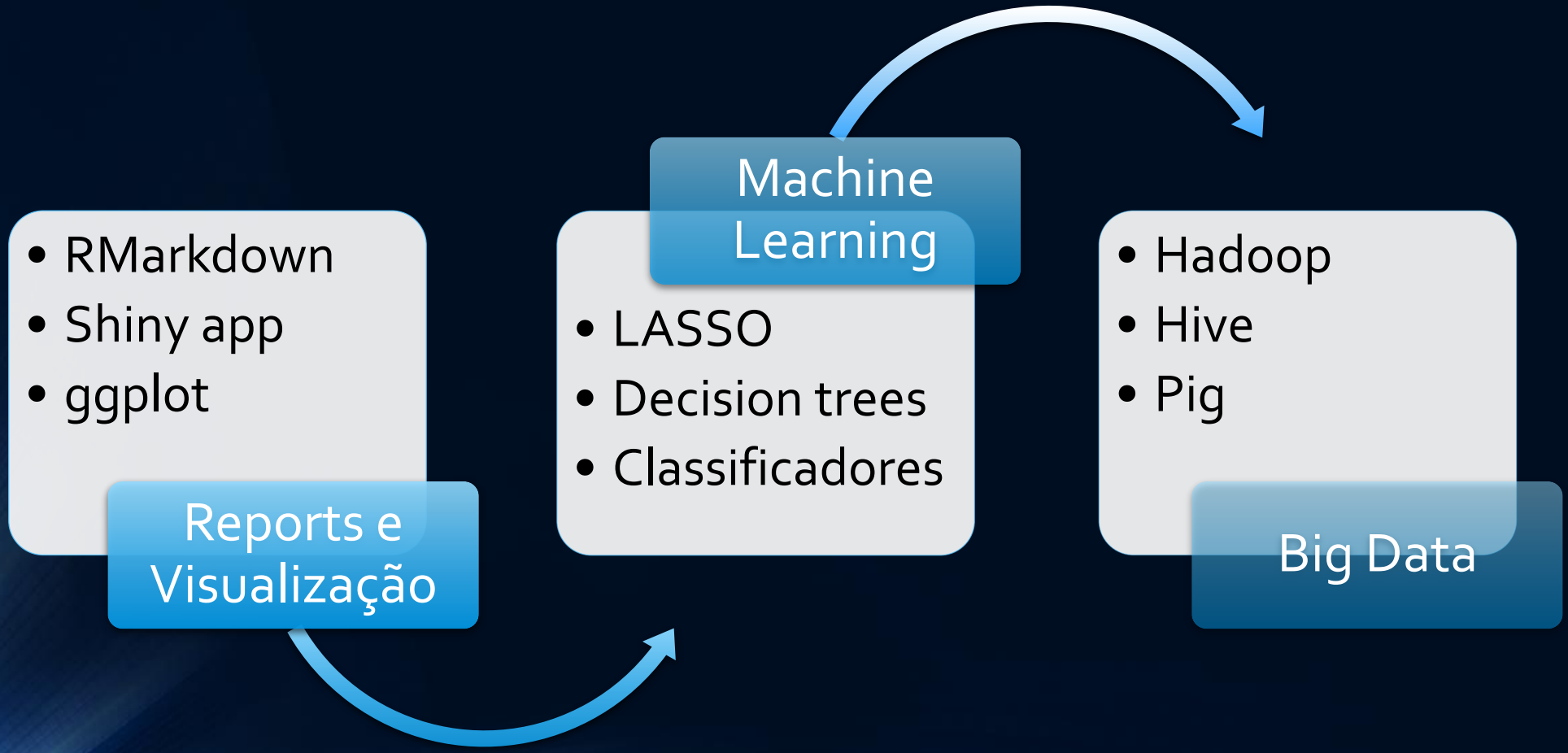
Bastante coisa, não? Mas o Mercado paga bem por isso



d) O que eu estou fazendo aqui numa segunda-feira à noite??

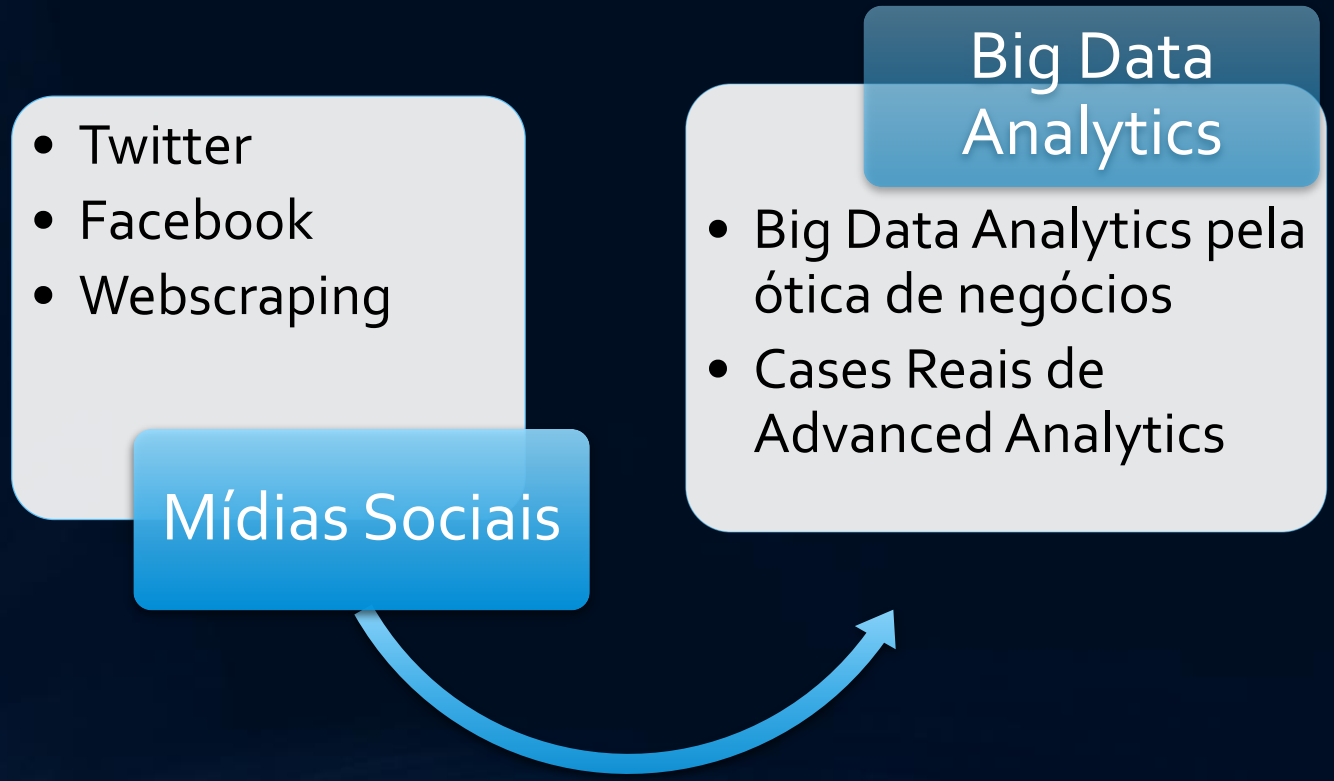


d) O que eu estou fazendo aqui numa segunda-feira à noite??

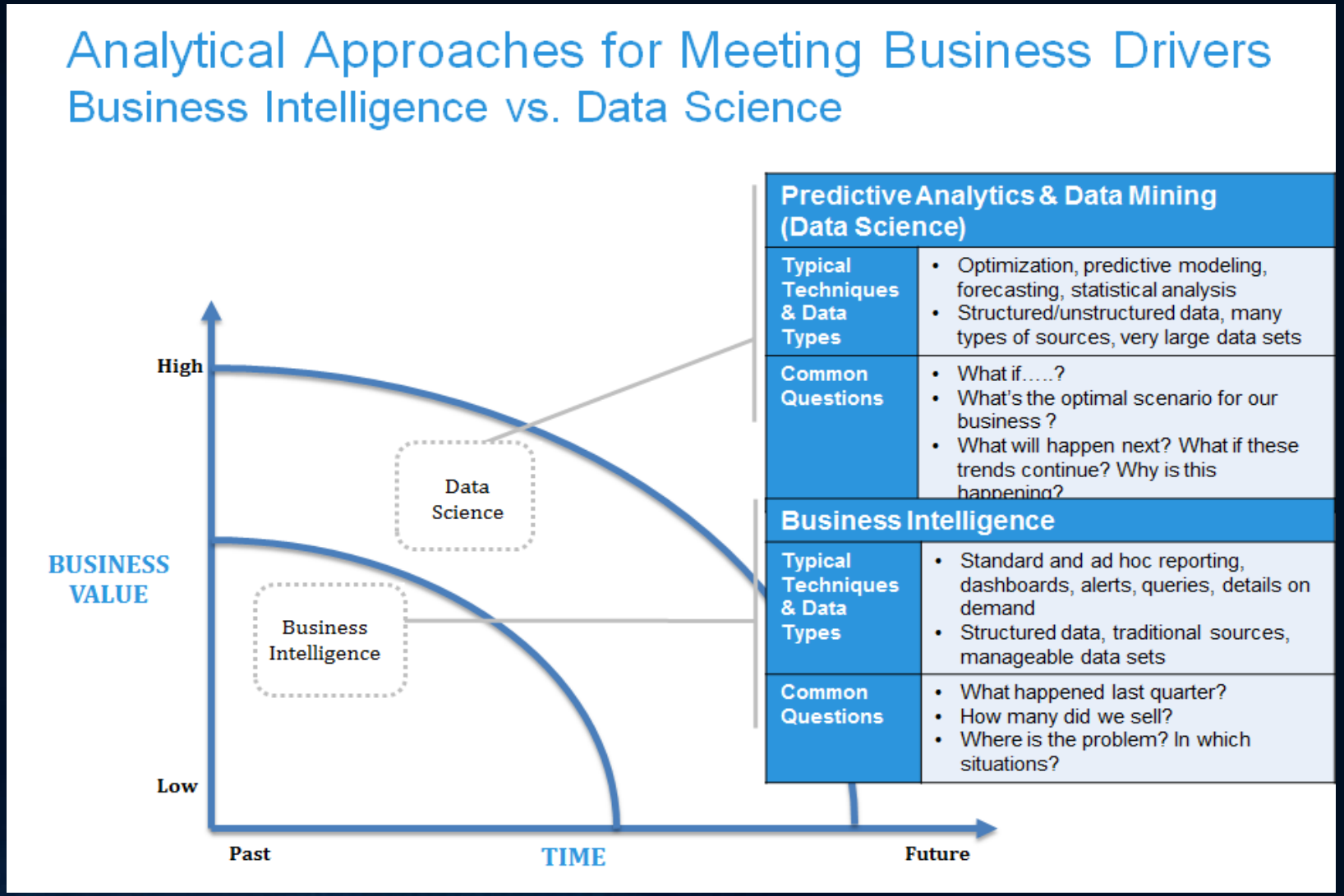


d) O que eu estou fazendo aqui numa segunda-feira à noite??

d) O que eu estou fazendo aqui numa segunda-feira à noite??



Dando alguns passos para me tornar um data scientist



Principais referências

Statistics *vs* Data Science *vs* BI. Revolutions. Available at: http://bit.ly/2jmwbse

What is the difference between a data scientist and a business intelligence analyst? Jason T Widjaja. Available at: http://bit.ly/2jHnKLA

Data Science: What is the difference between business analyst, data analyst, data scientist, business intelligence analyst, business systems analyst, and product manager? Mohit Chopra. Available at: http://bit.ly/2ijFrff

Data Science? Business Intelligence? What's the difference? David Rostcheck. Available at: http://bit.ly/2jaCxNG

10 differences between Data Science and Business Intelligence. Mike Merritt-Holmes. Available at: http://bit.ly/2jHsyAy

Business Intelligence x Data Science. Ciência e Dados. Disponível em: http://bit.ly/2ijGmfI

Artigos interessantes

Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results. Bernard Marr. Available at Amazon

Data Scientist: The Sexiest Job of the 21st Century. Harvard Business Review **-** Available at: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

How statistics lost their power – and why we should fear what comes next. The Guardian. Available at: https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy

Funções típicas de advogados já são feitas por softwares e robôs. Disponível em: http://exame.abril.com.br/revista-exame/deixa-que-o-robo-resolve/

Buyers Beware: Data Visualization is Not Data Analytics. Available at: http://www.smartdatacollective.com/eran-levy/455565/buyers-beware-data-visualization-not-data-analytics

Artigos interessantes

O future do trabalho. Von Der Heide. Available at: https://www.youtube.com/watch?v=eRGXlP-QloM&list=PLWOt6GX-RUTKlZriXc1-V2YlYuv1nuz9t

Prophet: forecasting at scale – Available at: https://research.fb.com/prophet-forecasting-at-scale/

Pedro Costa Ferreira

coordenador do curso de Big Data e Data Science – FGV|IDE

Doutor em Engenharia Elétrica - (Decision Support Methods) e Mestre em Economia. Co-autor dos livros "Planejamento da Operação de Sistemas Hidrotérmicos no Brasil" e "Análise de Séries Temporais em R: um curso introdutório". É o primeiro pesquisador da América Latina a ser recomendado pela empresa RStudio Inc. Atuou em projetos de Pesquisa e Desenvolvimento (P&D) no setor elétrico nas empresas Light S.A. (e.g. estudo de contingências judiciais), Cemig S.A, Duke Energy S.A, entre outras. Ministrou cursos de estatística e séries temporais na PUC-Rio e IBMEC e em empresas como o Operador Nacional do Setor Elétrico (ONS), Petrobras e CPFL S.A. Atualmente é professor de Econometria de Séries Temporais e Estatística, cientista chefe do Núcleo de Métodos Estatísticos e Computacionais (FGV|IBRE), coordenador dos cursos Economia Descomplicada (FGV|IDE) e Big Data e Data Science (FGV|IDE) e sócio-diretor da empresa Model Thinking Br (**MTBr**). É também revisor de importantes *journals*, como Energy Policy e Journal of Applied Statistics. Principais estudos são em modelos Econométricos, Incerteza Econômica, Preços, R software e Business Cycle.

Website: https://pedrocostaferreira.github.io/

Linkedin: http://bit.ly/2invbpl

RShiny Time Series: https://pedroferreira.shinyapps.io/timeseries/

Obrigado :)

BIG DATA E DATA SCIENCE

FORMAÇÃO EXECUTIVA FGV