BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO – IME - USP:

ÊNFASE EM CIÊNCIA DE DADOS

JOÃO EDUARDO FERREIRA

ROBERTO M. CESAR-JR

2016


KBA
Korea Baduk Association
Google DeepMind
Challenge Match
8 - 15 March 2016
AlphaGo
ALPHAGO
01:58:37
AlphaGo
Lee Sedol
Google DeepMind

IME-USP

We trained the neural networks on 30 million moves from games played by human experts, until it could predict the human move 57 percent of the time (the previous record before AlphaGo was 44 percent). But our goal is to beat the best human players, not just mimic them. To do this, AlphaGo learned to discover new strategies for itself, by playing thousands of games between its neural networks, and adjusting the connections using a trial-and-error process known as reinforcement learning. Of course, all of this requires a huge amount of computing power, so we made extensive use of Google Cloud Platform.

https://googleblog.blogspot.com.br/2016/01/alphago-machine-learning-game-go.html

LETTER

doi:10.1038/nature14236

Human-level control through deep reinforcement learning

Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

The theory of reinforcement learning provides a normative account[1], deeply rooted in psychological[2] and neuroscientific[3] perspectives on animal behaviour, of how agents may optimize their control of an environment. To use reinforcement learning successfully in situations approaching real-world complexity, however, agents are confronted with a difficult task: they must derive efficient representations of the environment from high-dimensional sensory inputs, and use these to generalize past experience to new situations. Remarkably, humans and other animals seem to solve this problem through a harmonious combination of reinforcement learning and hierarchical sensory processing systems[4,5], the former evidenced by a wealth of neural data revealing notable parallels between the phasic signals emitted by dopaminergic neurons and temporal difference reinforcement learning algorithms[3]. While reinforcement learning agents have achieved some successes in a variety of domains[6–8], their applicability has previously been limited to domains in which useful features can be handcrafted, or to domains with fully observed, low-dimensional state spaces. Here we use recent advances in training deep neural networks[9–11] to develop a novel artificial agent, termed a deep Q-network, that can learn successful policies directly from high-dimensional sensory inputs

agent is to select actions in a fashion that maximizes cumulative future reward. More formally, we use a deep convolutional neural network to approximate the optimal action-value function

$$Q^*(s,a) = \max_{\pi} \mathbb{E}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \mid s_t = s,\ a_t = a,\ \pi\right],$$

which is the maximum sum of rewards r_t discounted by γ at each time-step t, achievable by a behaviour policy $\pi = P(a|s)$, after making an observation (s) and taking an action (a) (see Methods)[19].

Reinforcement learning is known to be unstable or even to diverge when a nonlinear function approximator such as a neural network is used to represent the action-value (also known as Q) function[20]. This instability has several causes: the correlations present in the sequence of observations, the fact that small updates to Q may significantly change the policy and therefore change the data distribution, and the correlations between the action-values (Q) and the target values $r + \gamma \max_{a'} Q(s', a')$. We address these instabilities with a novel variant of Q-learning, which uses two key ideas. First, we used a biologically inspired mechanism termed experience replay[21–23] that randomizes over the data, thereby removing correlations in the observation sequence and smoothing over

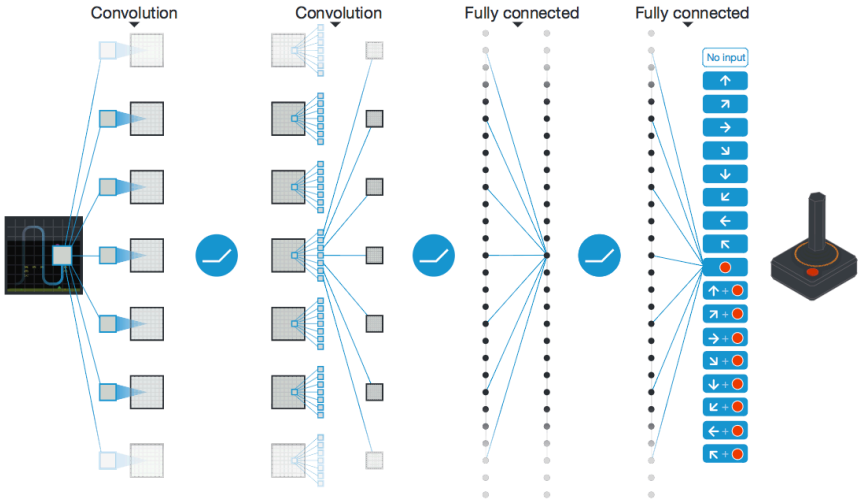


Figure 1 | Schematic illustration of the convolutional neural network. The details of the architecture are explained in the Methods. The input to the neural network consists of an $84 \times 84 \times 4$ image produced by the preprocessing map ϕ, followed by three convolutional layers (note: snaking blue line symbolizes sliding of each filter across input image) and two fully connected layers with a single output for each valid action. Each hidden layer is followed by a rectifier nonlinearity (that is, $\max(0,x)$).

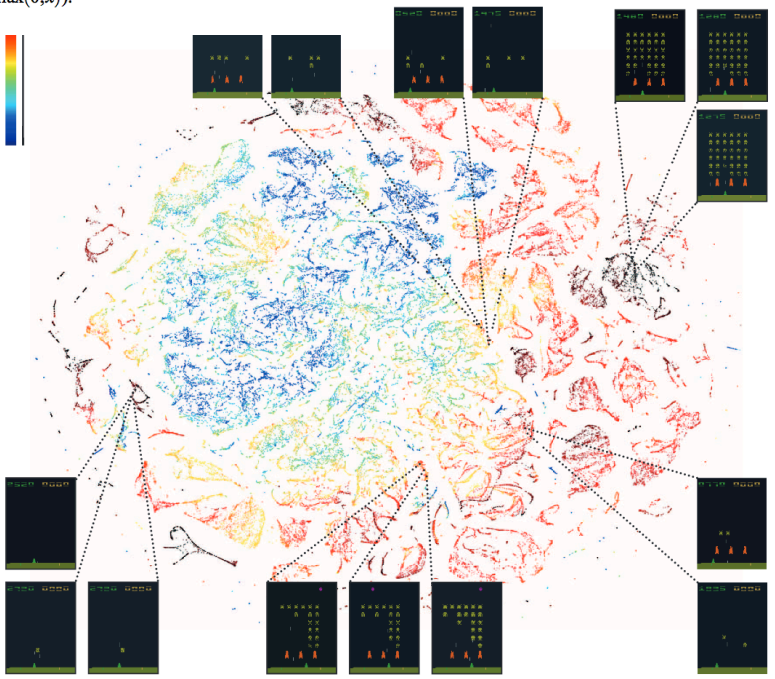


Figure 4 | Two-dimensional t-SNE embedding of the representations in the last hidden layer assigned by DQN to game states experienced while playing Space Invaders. The plot was generated by letting the DQN agent play for 2 h of real game time and running the t-SNE algorithm[25] on the last hidden layer predicts high state values for both full (top right screenshots) and nearly complete screens (bottom left screenshots) because it has learned that completing a screen leads to a new screen full of enemy ships. Partially completed screens (bottom screenshots) are assigned lower state values because

CERN Accelerating science

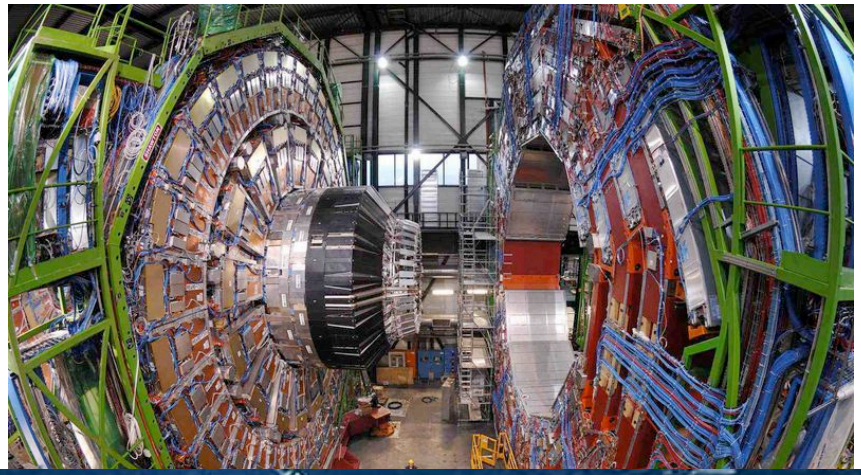
About CERN

Scientists

IME-USP

The Large Hadron Collider

The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator. It first started up on 10 September 2008, and remains the latest addition to CERN's accelerator complex. The LHC consists of a 27-kilometre ring of superconducting magnets with a number of accelerating structures to boost the energy of the particles along the way.



Science [edit]

The Large Hadron Collider experiments represent about 150 million sensors delivering data 40 million times per second. There are nearly 600 million collisions per second. After filtering and refraining from recording more than 99.99995%[79] of these streams, there are 100 collisions of interest per second.[80][81][82]

- As a result, only working with less than 0.001% of the sensor stream data, the data flow from all four LHC experiments represents 25 petabytes annual rate before replication (as of 2012). This becomes nearly 200 petabytes after replication.
- If all sensor data were recorded in LHC, the data flow would be extremely hard to work with. The data flow would exceed 150 million petabytes annual rate, or nearly 500 exabytes per day, before replication. To put the number in perspective, this is equivalent to 500 quintillion (5×10^{20}) bytes per day, almost 200 times more than all the other sources combined in the world.

https://en.wikipedia.org/wiki/Big_data

The Economist | World politics | Business & finance | Economics | Science & technology | Culture

Special report: Managing information

Data, data everywhere

Information has gone from scarce to superabundant. That brings huge new benefits, says Kenneth Cukier (interviewed here)—but also big headaches

Feb 25th 2010 | From the print edition

Timekeeper

Like 628 Tweet

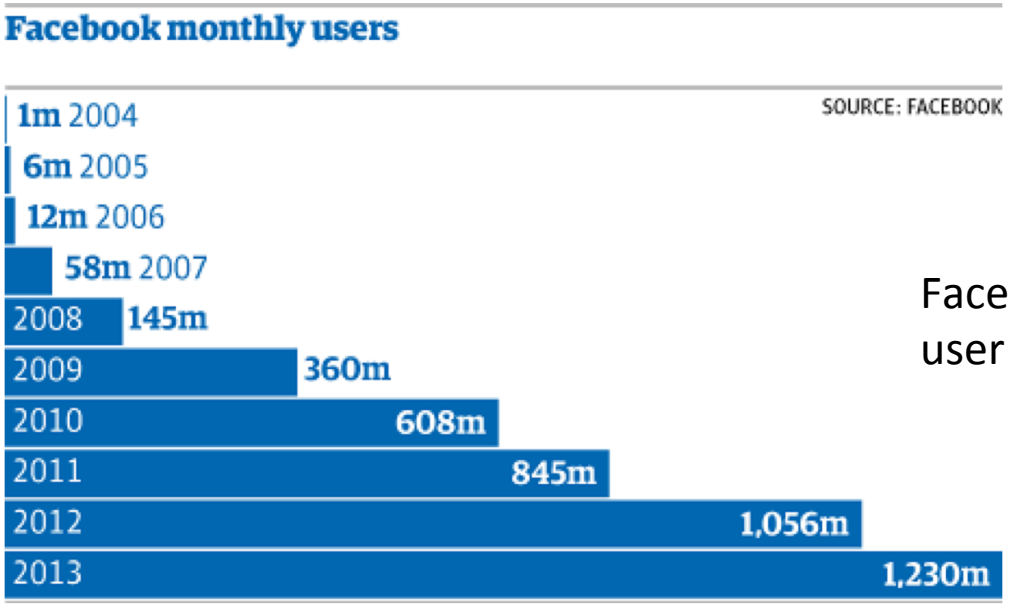
Retail [edit]

- Walmart handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain more than 2.5 petabytes (2560 terabytes) of data—the equivalent of 167 times the information contained in all the books in the US Library of Congress.[2]

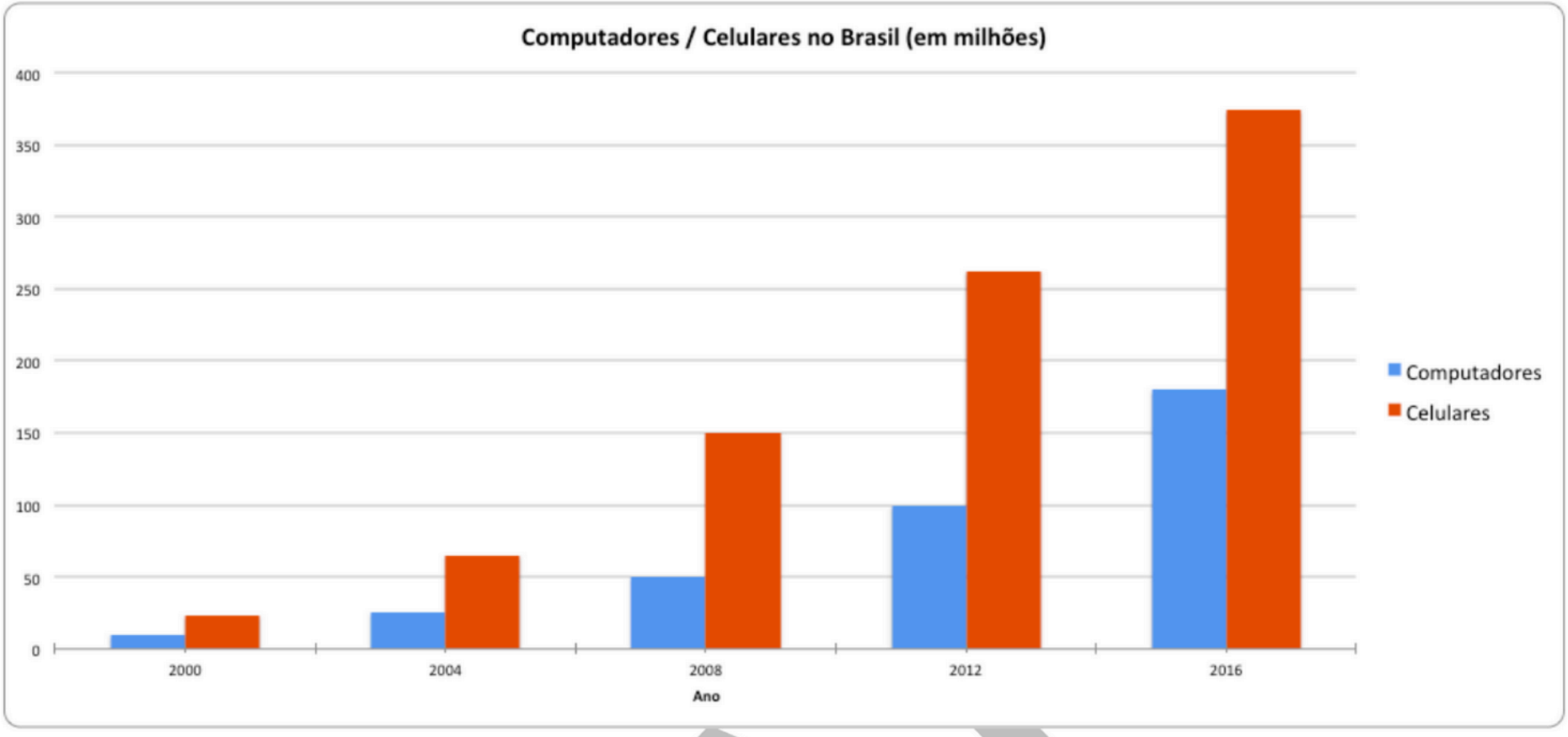
Technology Datablog

Facebook: 10 years of social networking, in numbers

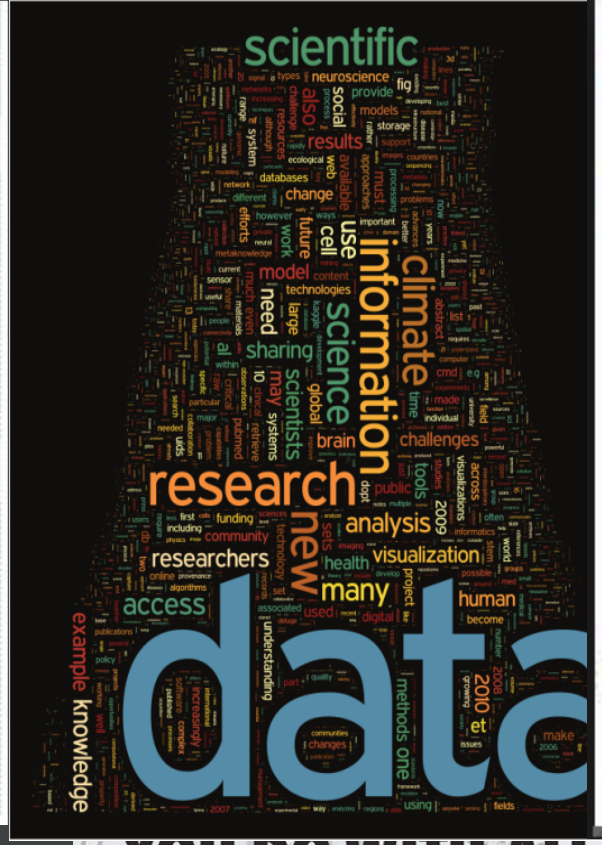
Ten years today, the social networking giant has notched up some interesting numbers along the way



Facebook handles 50 billion photos from its user base - Wikipedia



A Figura acima mostra a evolução do número de computadores e celulares no Brasil (em milhões). Atualmente, grande parte desses dispositivos estão conectados por meio de redes computacionais e de telecomunicação. Estima-se que 306 milhões desses dispositivos são conectáveis à Internet. A utilização de tais redes pela sociedade depende criticamente de sua confiabilidade e segurança. Por exemplo, 41% das transações bancárias no país são feitas pelo Internet Banking, o que explica um investimento em Tecnologia da Informação no valor de R$ 21.5 bilhões realizado pelo bancos no Brasil em 2014[1].



JOURNAL REPORTS | Updated March 8, 2013, 4:22 p.m. ET

How Big Data Is Changing th

Article | Stock Quotes | Comments

By STEVEN ROSENBUSH AND MICHAEL TOTTY

There's a ton of information out there. And businesses are figuring
work.

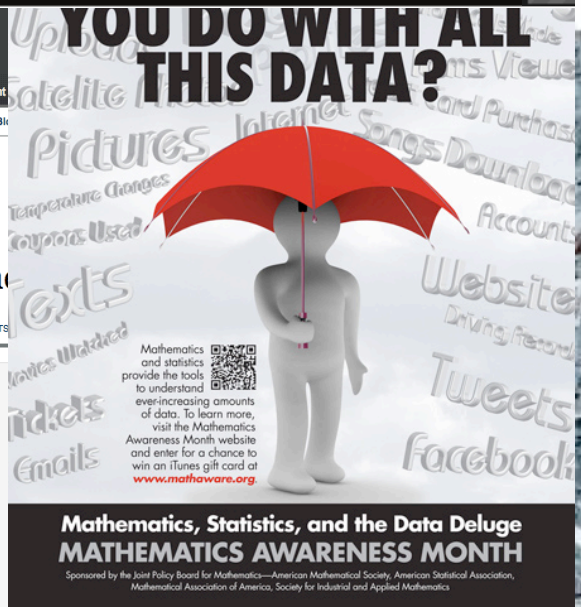
Journal Report

- Insights from The Experts
- Read more at WSJ.com/LeadershipReport

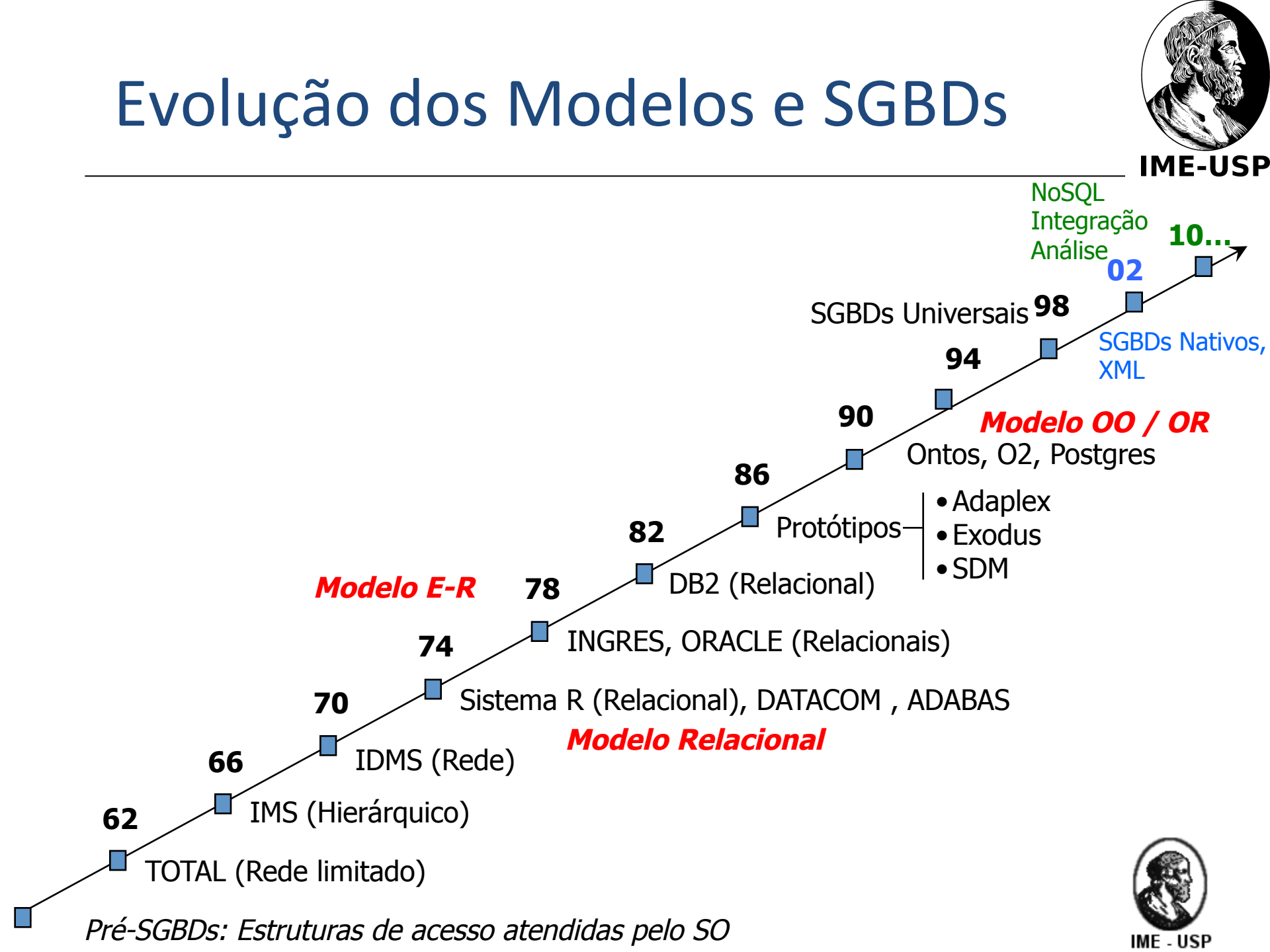
More in Unleashing Innovation: Big Data

- Big Data, Big Blunders
- The New Shape of Big Data

The experts call this s
data. The definition is
usually boils down to
have access to vastly
than they used to, it c
more different sources
they can get it almost



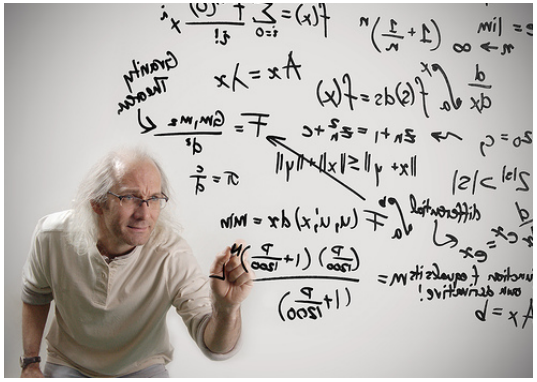
Evolução dos Modelos e SGBDs



Perfil dos pesquisadores

IME-USP

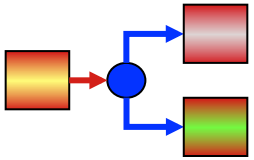
- *Theory guys*



- *System guys*

O que é necessário saber para trabalhar e pesquisar na área

- Tudo que já era necessário conhecer anteriormente:
  - Uma boa base de Lógica Matemática
  - Teoria de BD
  - Sistemas operacionais
  - Compiladores
  - Base de matemática discreta
  - Rudimentos de arquitetura de computadores (principalmente, dispositivos de armazenamento)
  - Conhecimentos de BD relacional (linguagens, processamento de consultas, processamento de transações)



O que é necessário saber para trabalhar e pesquisar na área

- ..., e muito mais:
  - Uma base mais ampla de Matemática (volumes de dados envolvidos = métodos aproximados):
    - álgebra, cálculo, estatística...
  - Complexidade Algorítmica
  - Paralelismo;
  - Recuperação de informação;
  - Aprendizagem de máquina;
  - Aplicações estratégicas, . . .

BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO – IME - USP:

ÊNFASE EM CIÊNCIA DE DADOS (data science)

BCC: núcleo

1	MAC101	MAC105	MAC110	MAC329	MAT2453	MAT112
2	MAC121	MAC216	MAC239	MAE119	MAT2454	MAT122
3	MAC102	MAC209	MAC210	MAC323	MAT236	Opt Est/Prob
4	MAC316	MAC338.	MAC422	Opt. Ciência	Opt. I	
5	MAC350	Opt. III	Opt. III	Opt. IV	Opt. V	Opt. VI
6	Opt. VII	Opt. VIII	Opt. IX	Opt. XI	Opt. XI	Opt. XII
7	FLC474	MAC499	Opt. XIII	Opt. XIV	Opt. XV	
8		MAC499	Opt. XVI	Opt.XVII	Opt. XVIII	Opt. XIX

http://bcc.ime.usp.br/curriculo2016/

Estrutura

IME-USP

- Núcleo do BCC (24 disciplinas)
- Núcleo de Ciência de Dados (6 disciplinas)
- Tutor
- Área de aplicação
- De 2 a 5 cursos na área de aplicação
- TCC na área

BCC: Ciência de Dados

1	MAC101	MAC105	MAC110	MAC329	MAT2453	MAT112
2	MAC121	MAC216	MAC239	MAE119	MAT2454	MAT122
3	MAC102	MAC209	MAC210	MAC323	MAT236	MAE 221 *
4	MAC316	MAC338.	MAC422	Opt Ciência	MAE 312*	
5	MAC350	MAC 460*	MAC 317*	MAC426*	Opt. V *	MAC431*
6	Opt. VII*	Opt. VIII*	Opt. IX*	Opt. X*	Opt. XI *	Opt. XII*
7	FLC474	MAC499	Opt. XIII*	Opt. XIV*	Opt. XV*	
8		MAC499	Opt. XVI*	Opt.XVII*	Opt. XVIII*	Opt. XIX*

http://bcc.ime.usp.br/curriculo2016/

* : livres ; * : livre ou optativa usada em uma única área de aplicação, 2 <=N<=5; *: obrigatória para Ciência de Dados

Júpiter - Sistema de Graduação

Instituto de Matemática e Estatística

Estatística

Disciplina: MAE0221 - Probabilidade I

Créditos Aula: 6
Créditos Trabalho: 0
Tipo: Semestral

Objetivos
Apresentar os conceitos fundamentais da Teoria das Probabilidades. Estudar os principais modelos probabilísticos discretos e contínuos, transformações de variáveis e principais distribuições amostrais.

Programa Resumido
1. Contagem: princípio multiplicativo, permutações, combinações.
2. Espaço de probabilidade.
3. Probabilidade condicional e independência.
4. Variáveis e vetores aleatórios: definição, caracterizações e propriedades.
5. Esperança matemática, funções geradoras de probabilidade e de momentos e suas propriedades.
6. Principais distribuições de probabilidade (univariadas e multivariadas): uniforme discreta, Bernoulli, binomial, geométrica, Poisson, binomial negativa, hipergeométrica, multinomial, exponencial, normal, Cauchy e uniforme contínua.
7. Transformações de variáveis: direta e método do jacobiano. Distribuição da soma, produto e quociente de variáveis aleatórias.
8. Estatísticas de ordem, distribuições t-Student, F-Snedecor, qui-quadrado, gama, beta e suas relações.
9. Distribuição normal multivariada e propriedades.
10. Lei dos grandes números.
11. Teorema limite central.

Júpiter - Sistema de Graduação

Instituto de Matemática e Estatística

Estatística

Disciplina: MAE0312 - Introdução aos Processos Estocásticos

Créditos Aula: 4
Créditos Trabalho: 0
Tipo: Semestral

Objetivos
Apresentar a noção de processos estocásticos que é central na teoria das probabilidades moderna. Fornecer exemplos elementares e os teoremas centrais em processos estocásticos.

Programa Resumido
1. Conceitos básicos e exemplos.
2. Construção de cadeias de Markov.
3. Comportamento assintótico das cadeias de Markov. Tempo médio de recorrência. Medidas invariantes. Reversibilidade.
4. Convergência em distribuição via acoplamento.
5. Processos pontuais e processos de Poisson.
6. Teoria da renovação a tempo discreto e teorema chave.
7. Martingales discretos.
8. Processos Markovianos de salto. Construção. Explosão.

Júpiter - Sistema de Graduação

Instituto de Matemática e Estatística

Ciência da Computação

Disciplina: MAC0460 - Aprendizagem Computacional: Modelos, Algoritmos e Aplicações

Créditos Aula: 4
Créditos Trabalho: 0
Tipo: Semestral

Objetivos
Introdução à Técnica de Aprendizagem Computacional conhecida como PAC Learning (de Probably Approximately Correct Learning) aplicada a problemas de Processamento de Imagens. Serão
apresentados modelos matemáticos de aprendizagem, decomposição de operadores por técnicas de Morfologia Matemática, estimação de parâmetros das decomposições de operadores por técnicas de aprendizado, e aplicações.

Docente(s) Responsável(eis)

Flavio Soares Correa da Silva
Junior Barrera
Routo Terada

Programa Resumido

Programa
Conceitos, hipóteses e algoritmos de aprendizagem. Representações e fórmulas booleanas.
Decomposições por Morfologia Matemática. Decomposições por Redes Neurais. Aprendizagem probabilística.
Aprendizagem eficiente. Dimensão VC. Aplicações.

Júpiter - Sistema de Graduação

Instituto de Matemática e Estatística

Ciência da Computação

Disciplina: MAC0317 - Algoritmos para Processamento de Áudio, Imagem e Vídeo

Créditos Aula: 4
Créditos Trabalho: 0
Tipo: Semestral

Objetivos
Apresentar ao aluno os fundamentos teóricos e o ferramental computacional comuns às áreas de processamento digital de imagens e vídeo e processamento digital de áudio, incluindo representação digital de áudio, imagens e vídeo, transformações tempo-frequência e espaço-frequência, e desenho e implementação de filtros digitais para problemas típicos em processamento de áudio, imagens e vídeo, tais como suavização, segmentação e compressão, entre outros.

Docente(s) Responsável(eis)

Marcelo Gomes de Queiroz
Nina Sumiko Tomita Hirata
Roberto Hirata Junior
Roberto Marcondes Cesar Junior

Programa Resumido
Representação digital de sinais de áudio, imagens, e vídeo: amostragem, quantização e "aliasing". Transformada Discreta de Fourier e FFT (1D, 2D e 3D) Outras transformações: Transformada de Fourier (Contínua), Transformada do Cosseno Discreta, Transformada z, Transformada de Walsh-Hadamard, Transformada de Haar Convolução linear, circular e secionada Filtros lineares (FIR) e Filtros recursivos (IIR) Aplicações de filtros: suavização, interpolação, realce, detecção de bordas e segmentação Janelamento no tempo e no espaço, localização e efeitos no espectro Bancos de filtros e técnicas de análise-ressíntese Compressão: Predição Linear, compressão usando DCT, Compensação de Movimento Sinais aleatórios: Representação, Filtros de Wiener e de Kalman

Júpiter - Sistema de Graduação

Instituto de Matemática e Estatística

Ciência da Computação

Disciplina: MAC0426 - Sistemas de Bancos de Dados

Créditos Aula: 4
Créditos Trabalho: 0
Tipo: Semestral

Objetivos
Expor os principais fundamentos de banco de dados e os sistemas gerenciadores que administram sua utilização. Apresentar técnicas de modelagem e de implementação de bancos de dados.

Docente(s) Responsável(eis)

Francisco Carlos da Rocha Reverbel
João Eduardo Ferreira
Kelly Rosa Braghetto
Marcelo Finger

Programa Resumido
Introdução: arquitetura de bancos de dados. Modelagem de bancos de dados: projeto conceitual, lógico e físico de bancos de dados. Modelos conceituais: modelo ER básico e estendido. Projeto de bancos de dados utilizando o modelo ER estendido. Modelo relacional: definições e formalização. Mapeamento do modelo ER estendido para o Modelo Relacional. Linguagens do modelo relacional: álgebra relacional, cálculo relacional e SQL. Dependências funcionais e normalização de relações. Índices hashing e árvores B, B+. Controle de concorrência e algoritmos para recuperação de falhas. Otimização de consultas relacionais. Dados semi-estruturados (por exemplo, XML e JSON). Novas tecnologias para gerenciamento de dados (por exemplo, NoSQL).

Júpiter - Sistema de Graduação

Instituto de Matemática e Estatística

Ciência da Computação

Disciplina: MAC0431 - Introdução à Computação Paralela e Distribuída

Créditos Aula: 4
Créditos Trabalho: 0
Tipo: Semestral

Objetivos
Familiarizar o aluno com os conceitos e termos básicos de sistemas paralelos e distribuídos, apresentar os tipos de arquitetura mais usados, descrever o suporte necessário para a programação de tais sistemas, e apresentar algumas aplicações.

Programa Resumido

Programa
Problemas e conceitos; tipos e granularidades de paralelismo; arquiteturas de sistemas paralelos e distribuídos; topologias de interconexão; protocolos de comunicação; mecanismos de comunicação e sincronização; linguagens e sistemas de programação; algoritmos paralelos e distribuídos; aplicações.

Bioinformática
Parceiro: Bioquímica

- Introdução à Bioinformática (MAC0341)
- Algoritmos em Bioinformática (MAC0351)
- Biologia de Sistemas (MAC0375)
- Bioquímica
- Biologia Molecular

Visão Computacional

Parceiros: Medicina, Mecatrônica Poli, Aero Espacial, ITA, IO

- **Computação gráfica (MAC0420)**
- **Visão computacional (MAC 0417)**
  - **https://www.youtube.com/watch?v=jWWg4qx0JOU**
- Morfologia matemática
- Ótica e eletromagnetismo
- IA /Radiologia/Fisiologia / Patologia/ disciplinas IO

Controles
Parceiro: Mecatrônicas

- Disciplina de Controles da mecatronica
- Disciplina de Controls da mecatronica
- Outras disciplinas da mecatronica
- Outras disciplinas da mecatronica
- Outras disciplinas da mecatronica

Economia/Administração
Parceiro: FGV

- Disciplina da FGV (básica para Econ/Adm)
- Disciplina da FGV (básica para Econ/Adm)
- Disciplinas da FGV (específicas Econ ou Adm)
- Disciplinas da FGV (específicas Econ ou Adm)
- Disciplinas da FGV (específicas Econ ou Adm)

Análises de dados Esportivos

Parceiro: – Educação Física - USP

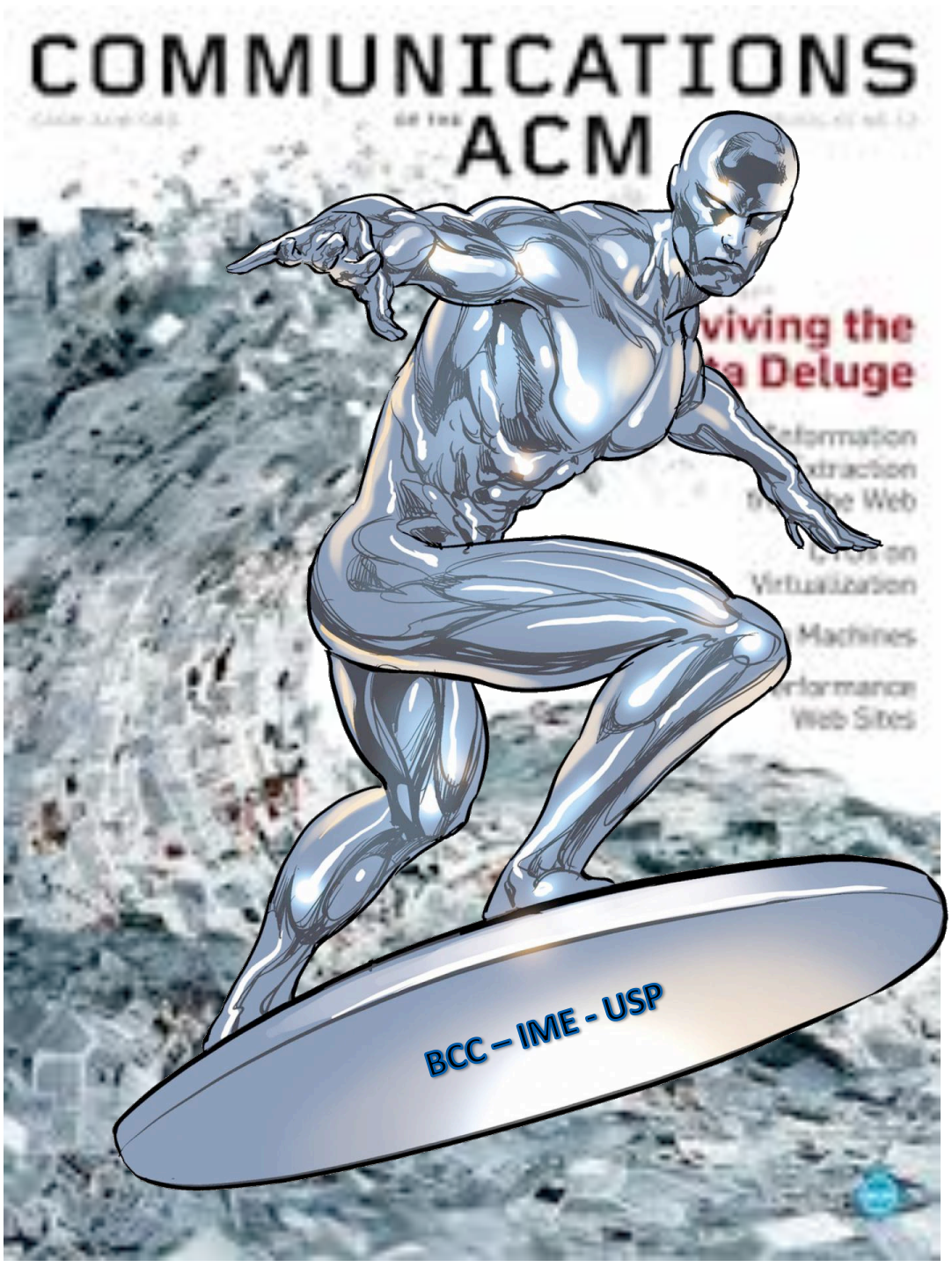
- Disciplina de Esportes
- Disciplina de Esportes

Computação Musical
Parceiro: música

- Ondulatória
- IA
- Disciplina da Música

Palestras MAC0102

- Palestra 1, Bioinformática, Ronaldo e André, 9/6
- Palestra 2, Imagens médicas e biológicas, Marcel e Paulo, 16/6
- Palestra 3, Engenharias (incluindo bioengenharia, rehabilitação, acessibilidade), Hitoshi e Junior, 23/6
- Palestra 4, Comentários finais, temas livres, aplicações não cobertas como finanças e computação musical, todos os professores, 30/6


COMMUNICATIONS
of the
ACM
viving the
a Deluge
nformation
xtraction
e Web
Virtualization
Machines
erformance
Web Sites
BCC – IME - USP

IME-USP