

IE301: Dados, Inferência, e Aprendizagem

Unicamp, Agosto 2019

Prof. José Cândido Silveira Santos Filho

Prof. Flavio du Pin Calmon

Prelúdio:
Por que “Machine Learning” e
“Data Science” são tão importantes hoje?

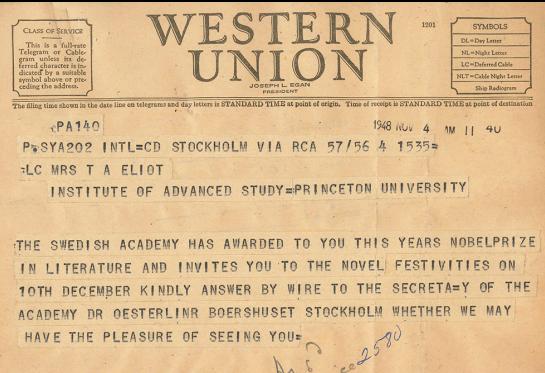
Comunicações em 1948...



Wembley Post Office Press Room for the 1948 Olympics and other sports updates.



The first Polaroid Land "instant" camera is sold (sale made at a department store in Boston)



Telegram announcing T.S. Eliot's 1948 Nobel Prize in Literature



First television news broadcast with an anchor on camera.

Comunicações hoje:

YouTube stream of CBS “sports-related” report about a selfie between a Nobel Prize laureate and one of Boston’s main sports personalities. The selfie was taken with a smartphone and posted on Twitter.

Video viewed on Harvard’s secure wireless network at about 50Mbps.

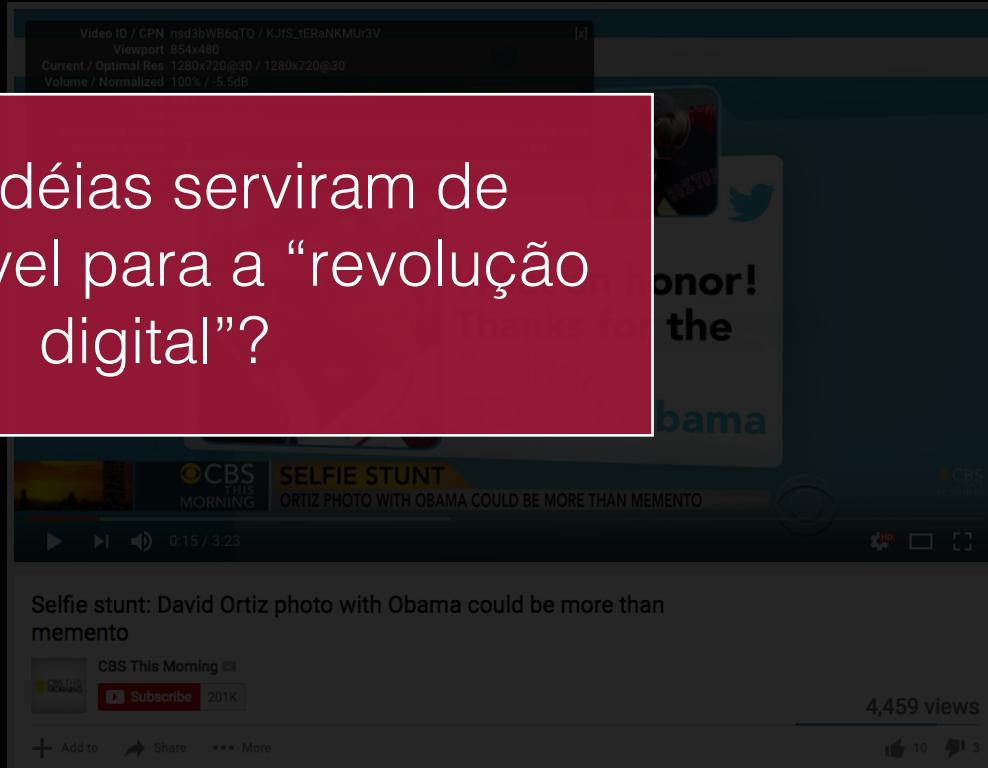


Comunicações hoje:

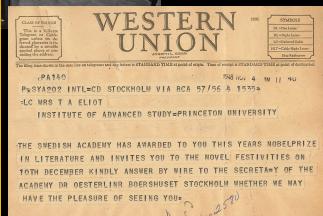
YouTube stream of CBS This Morning's "Selfie Stunt" segment. The video is a "related" report about a "selfie" taken between a Nobel Prize winner and one of Boston's main sports personalities. The selfie was taken with a smartphone and posted on Twitter.

Video viewed on Harvard's secure wireless network at about 50Mbps.

Quais idéias serviram de combustível para a “revolução digital”?



Em apenas algumas décadas...



Video ID / CPN nsd3bWB6qTQ / KJfS_tERaNKMuJ3V
Viewport 854x480
Current / Optimal Res 1280x720@30 / 1280x720@30
Volume / Normalized 100% / -5.5dB
Codecs avc1.4d401f(136) / mp4a.40.2(140)
Host r3-sn-ab5zn7s Host David Ortiz
Connection Speed 589.24 Kbps
Network Activity 0 KB
Buffer Health 74.66 s
Dropped Frames 0/466

What an honor!
Thanks for the
#selfie,
@BarackObama

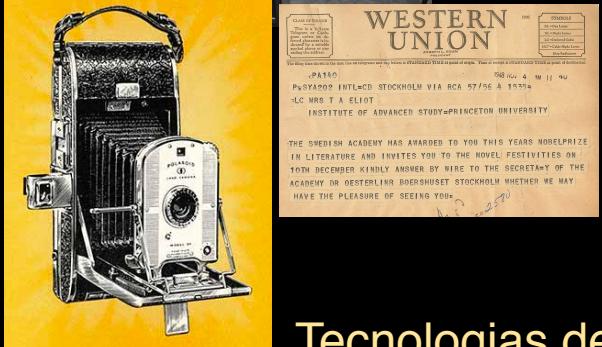
CBS THIS MORNING SELFIE STUNT ORTIZ PHOTO WITH OBAMA COULD BE MORE THAN MEMENTO

Selfie stunt: David Ortiz photo with Obama could be more than memento

4,459 views

10 3

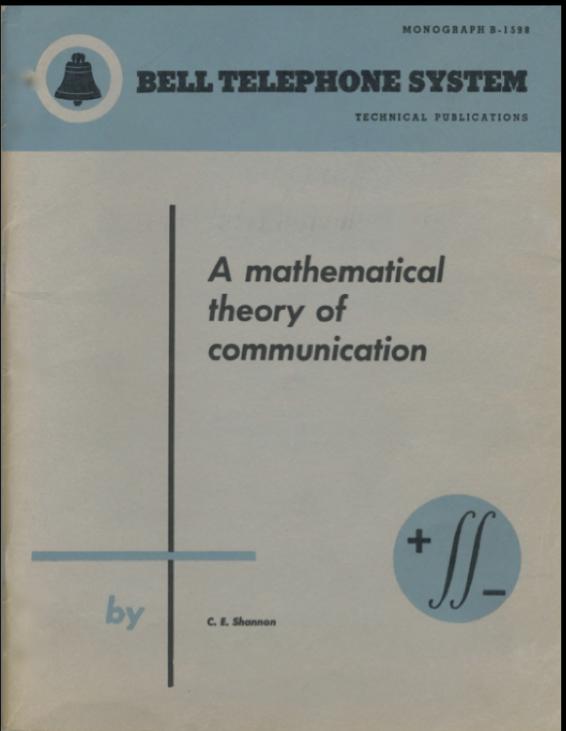
Em apenas algumas décadas...



- Tecnologias de comunicações:
- Analógicas,
 - Ad-hoc e dependentes da aplicação,
 - Não eram robustas a ruído,
 - Ineficientes...

...sistemas rápidos,
distribuídos e robustos.

Algo mais aconteceu em 1948...



1948, Bell Sys. Tech. Journal



Claude Shannon, 1916-2001

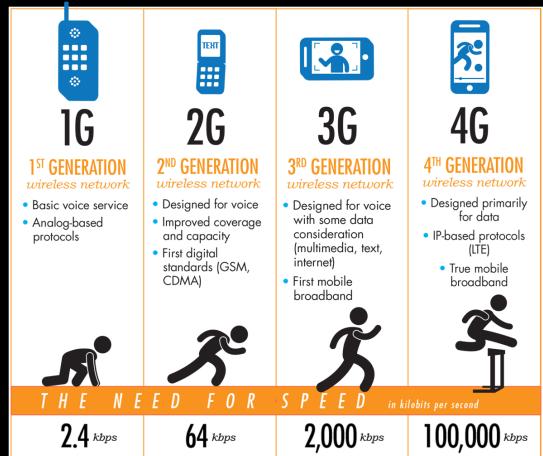
As perguntas de Shannon:

- Como devemos medir informação?
 - Entropia, informação mútua, etc.
- Quantos bits precisamos para representar/reconstruir informação?
 - Codificação de fonte.
- Como transmitir informação de maneira confiável através de um canal ruidoso?
 - Codificação de canal.
- Shannon respondeu essas questões teoricamente, mas não propôs métodos práticos para atingir esses limites
- Nas últimas décadas, várias técnicas surgiram para atingir os limites de Shannon.
 - Algebraic codes, convolutional codes, LDPC codes, turbo codes, polar codes.

As perguntas de Shannon:

- Como devemos medir informação?
 - Entropia, informação mútua, etc.
- Quantos bits precisamos para representar/reconstruir informação?
 - Codificação de fonte.

As respostas à essas perguntas
- Como transmitir informação mudaram o mundo!através de um canal ruidoso?
 - Codificação de canal.
- Shannon respondeu essas questões teoricamente, mas não criou métodos práticos para atingir esses limites
- Nas últimas décadas, vários métodos surgiram para atingir os limites de Shannon.
 - Algebraic codes, convolutional codes, LDPC codes, turbo codes, polar codes.



Cellular communications



Datacenters



Voyager spacecraft



Digital media



Wireless technologies



Indu



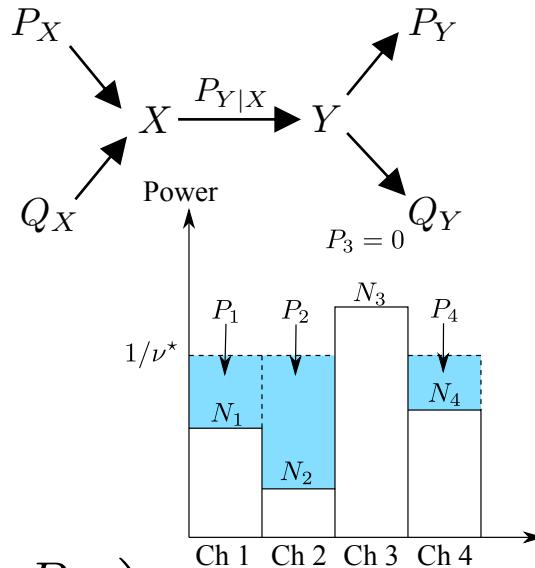
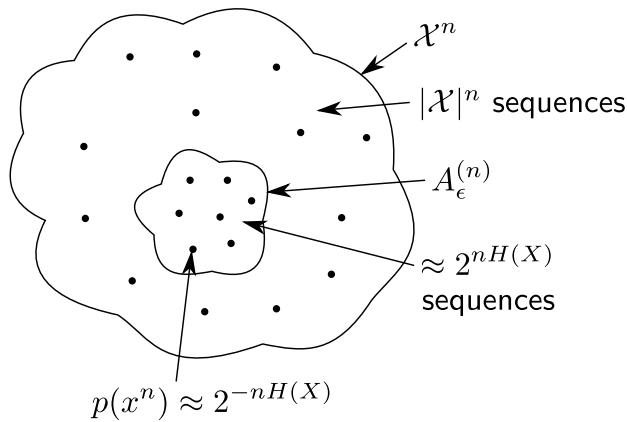
Compression standards



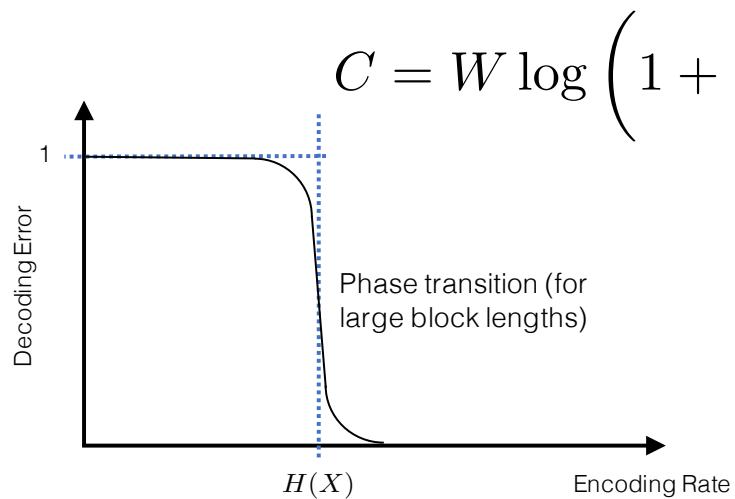
Medical Imaging

Information
theory

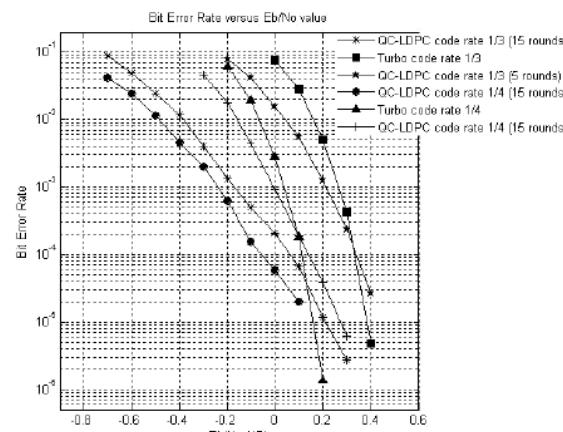
Electrical Engineering + Computer Science



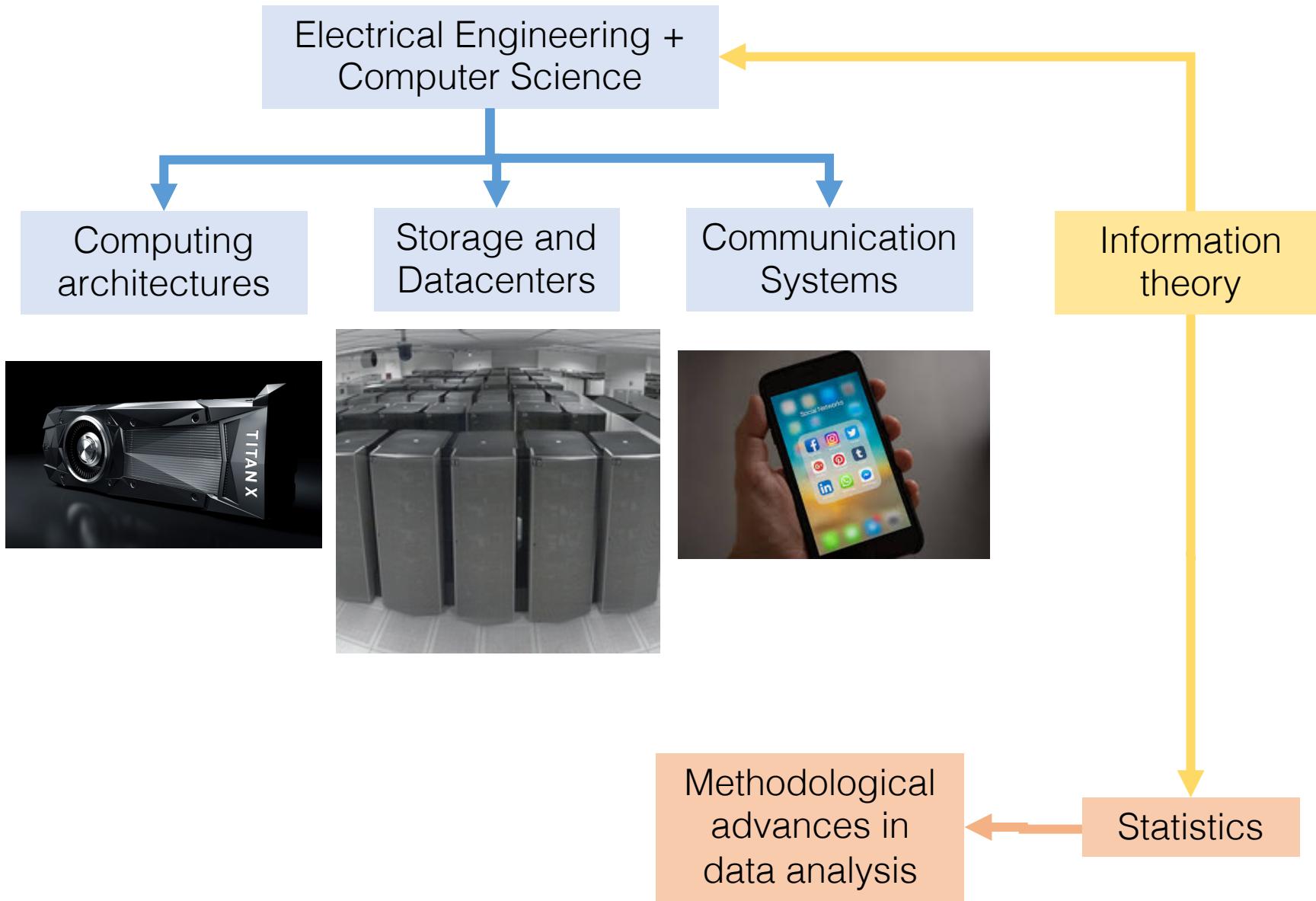
Information theory

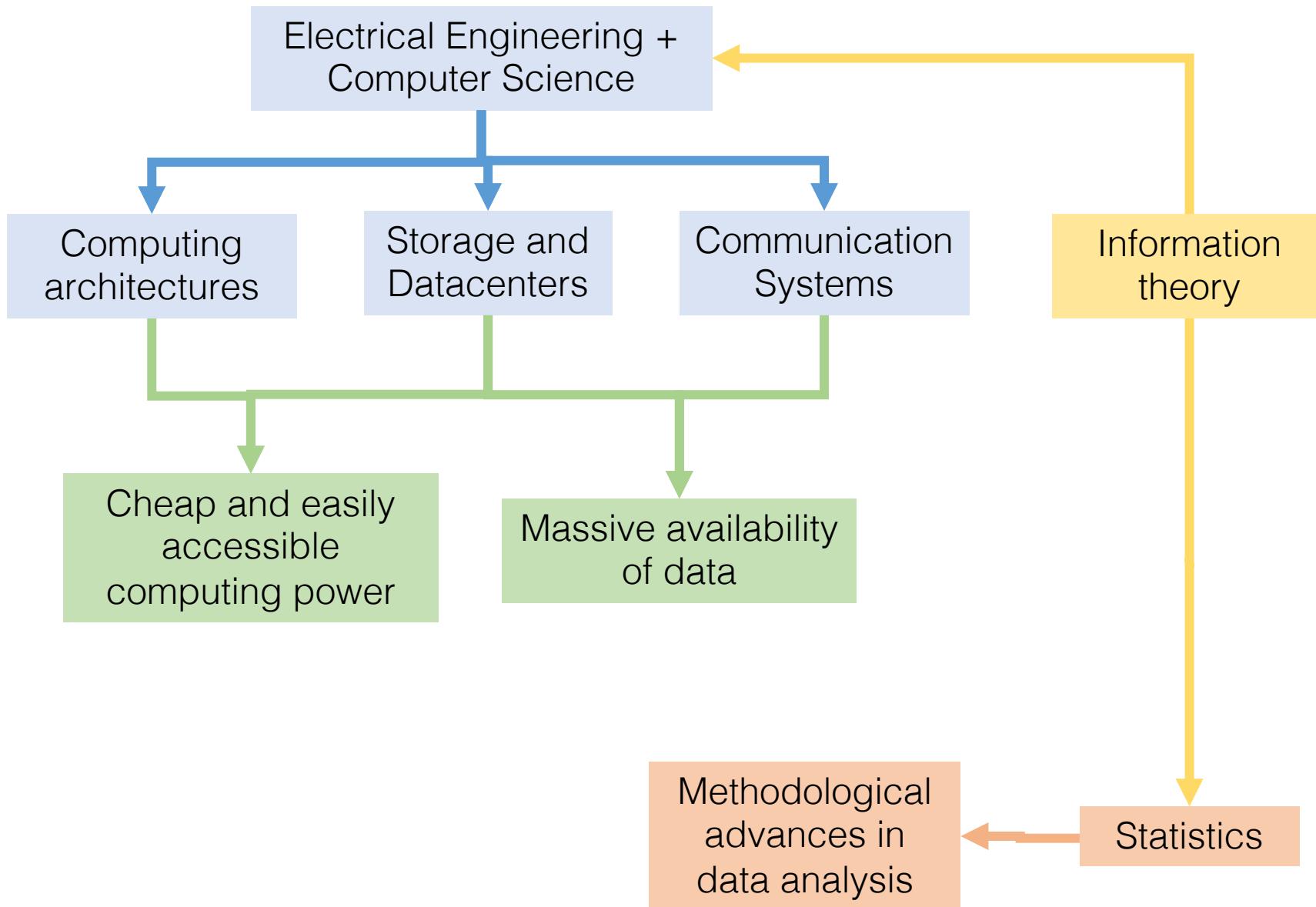


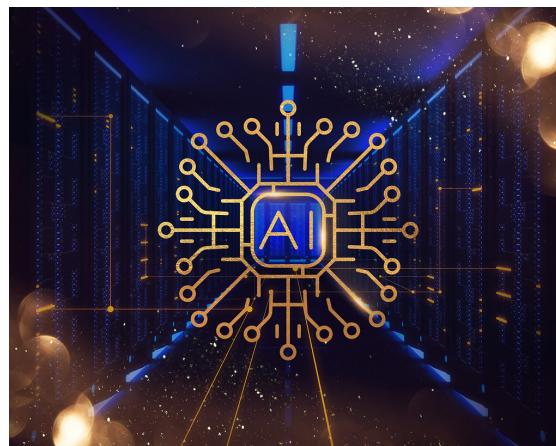
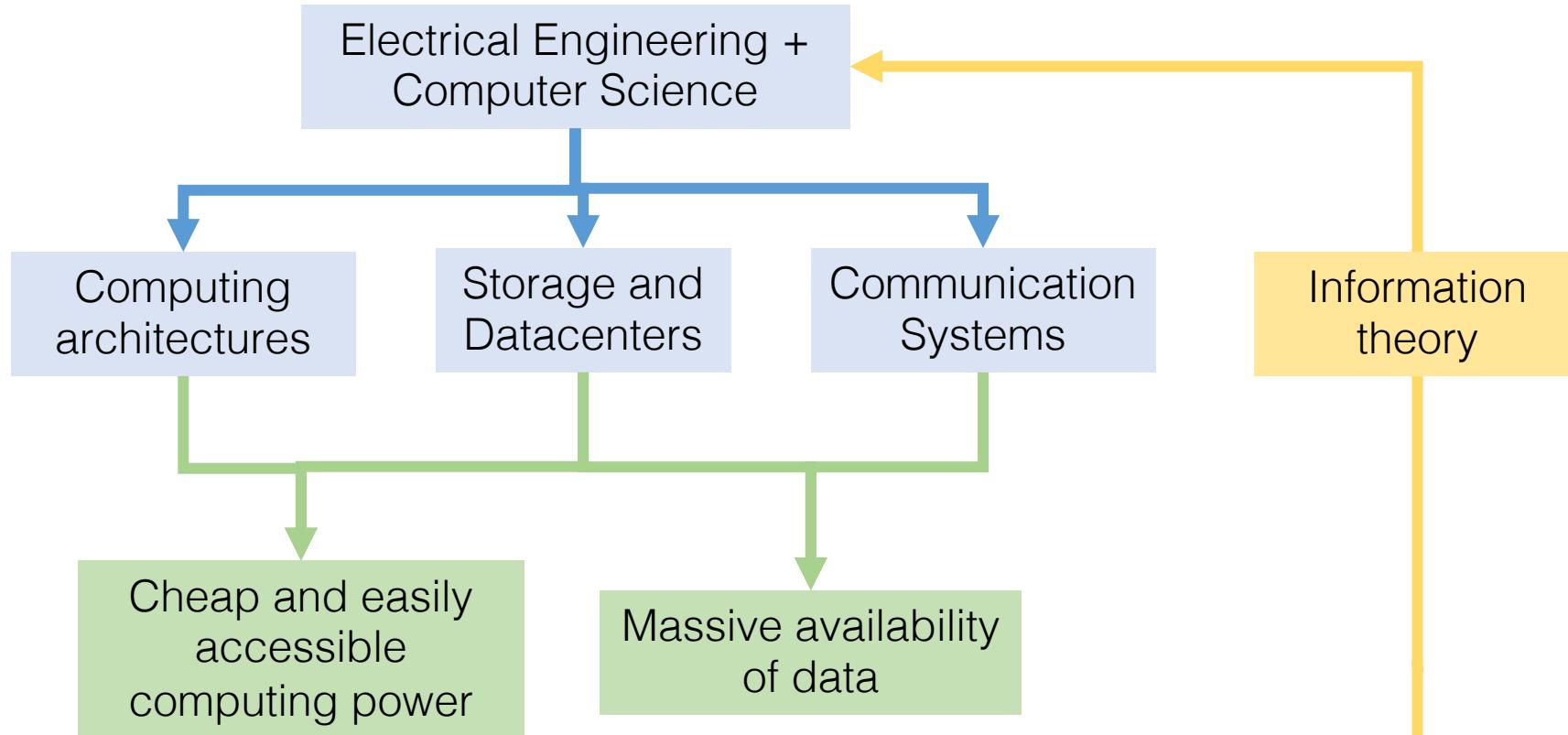
$$C = W \log \left(1 + \frac{P}{N_0 W} \right) \text{ bits/second}$$

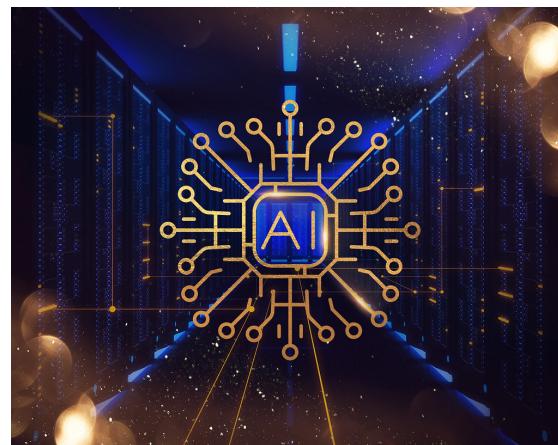
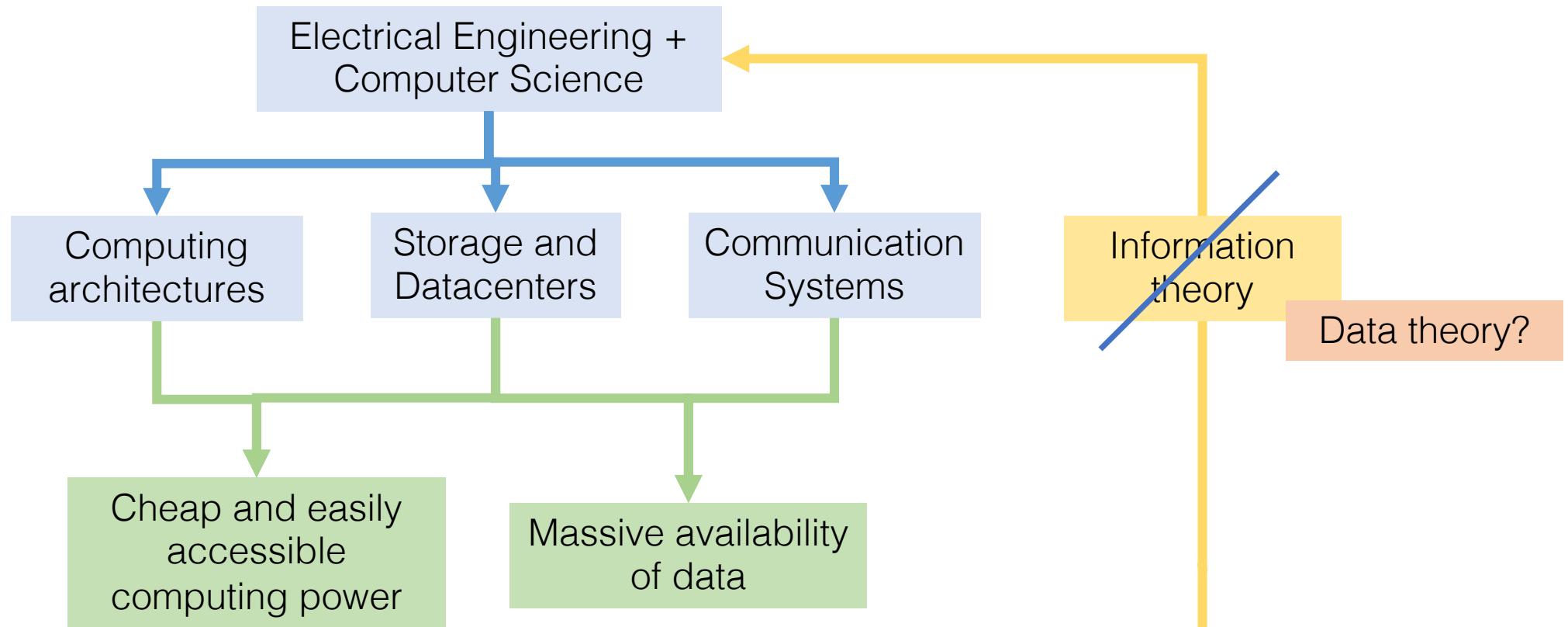


Statistics









Methodological advances in data analysis

Statistics

“I want you to think about data as the next natural resource.”

- Ginni Rometti, IBM's CEO

Oil
Natural gas
Uranium
Coal
Water
Timber

Data can be collected and processed in order to provide services and improve businesses...

...but, like other natural resources, it can be misused and pose new risks for businesses and society.

Objetivo do curso:

Introduzir aprendizagem de máquina
como uma peça fundamental na caixa de
ferramentas do engenheiro moderno

Em duas semanas, você vai conseguir:

- Descrever a diferença entre abordagens Bayesianas e Frequentistas em ML
- Entender os componentes principais do “pipeline” de ciências de dados
- Aplicar métodos de ML a dados do mundo real
- Criar os seus próprios algoritmos de ML, e comparar com o estado da arte
- Explicar os conceitos teóricos fundamentais por traz de qualquer problema de aprendizagem computacional (generalização, viés, etc.)
- Analisar questões de privacidade e justiça em ML

Estrutura do Curso:

1. Inferência e aprendizagem: regressão linear.
2. Famílias exponenciais e suas aplicações.
3. Classificação supervisionada.
4. Generalização e teoria de aprendizagem.
5. Aprendizagem não-supervisionada.
6. Discriminação e privacidade.
7. Tópicos avançados.

Avaliação: projeto final

Objetivo:

Aplicar conceitos vistos em aula a um problema real e aprender mais sobre a pesquisa de ponta na área de ML. Esperamos que você possa aplicar técnicas de ML na sua própria pesquisa.

Exemplos de projetos:

- Análise em larga escala de um dataset público (e.g., INEP, dados.gov.br);
- Reproduzir e estender um método publicado em uma conferência ou revista da área (e.g., NeurIPS, JMLR, ICML, ICLR, IEEE PAMI);
- Explorar aplicações de ML em problemas de engenharia (e.g., wireless, controle, dispositivos).

O projeto pode ser feito em grupos de até 3 pessoas.

Entregáveis: relatórios intermediários, um projeto final escrito no formato de alguma conferência da área ao seu critério, código que reproduza os resultados.

1. Proposta de projeto (21 de Agosto, 20% da nota):

- Uma página delineando o projeto que será feito (dataset, modelo, algoritmo, análise teórica) com referências pertinentes.

2. Relatório intermediário (4 de Setembro, 20% da nota)

- Relatarório intermediário de progresso (desafios, literature investigada, barreiras para progresso, riscos identificados). (30% da nota)

3. Relatório e código final (18 de Setembro, 60% da nota)

Projeto C: proposta bem delineada, mas sem avanço no relatório intermediário e final.

Projeto não lista referências na área, não apresenta resultados significativos em dados e/ou não contém uma análise criteriosa.

Projeto B: proposta bem delineada que é desenvolvida de maneira significativa no relatório intermediário e no relatório final. Os relatórios refletem um domínio dos conceitos abordados em sala e conectam os conceitos com referências na área, mas listagem e conexões são incompletas. A análise baseada em dados é correta embora abreviada. As discussões sobre aspectos teóricos no relatório final são limitadas.

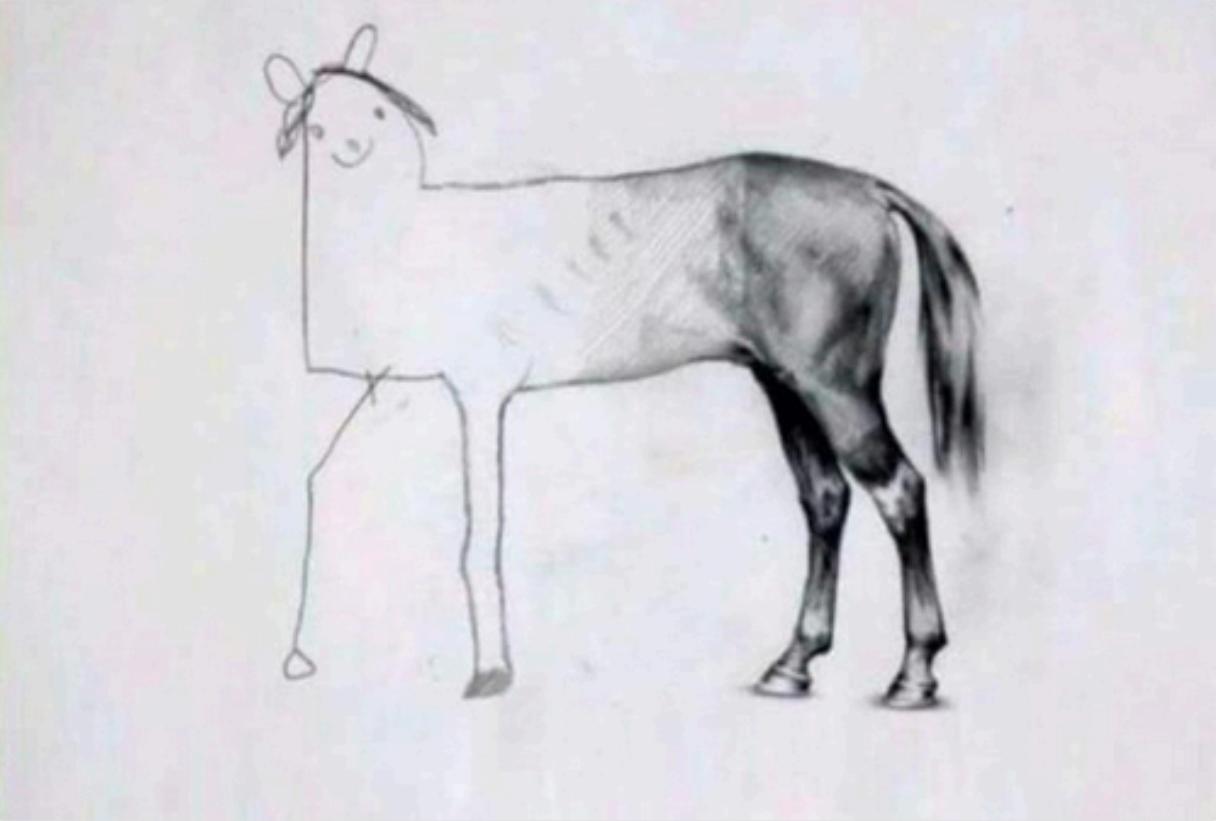
Projeto A: com todas as virtudes e nenhuma das deficiências listadas anteriormente. O gradiente do projeto aponta para uma eventual publicação em conferência ou revista na área.

Administre o seu
tempo!

franck nijimbere
@nijfranck

Follow

When the deadline comes too close



5:16 PM - 23 Mar 2018

87,083 Retweets 255,974 Likes



Dinâmica do curso

Sala de aula: aulas expositivas e discussões.

Em casa:

- lista de leitura e exercícios computacionais em Python postados na a cada duas aulas (começando na quarta-feira).

Referências:

**O. Simeone, A Brief Introduction to Machine Learning for Engineers, arXiv preprint
arXiv:1709.02840v3, 2018.**

S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: from Theory to Algorithms, Cambridge, 2014.

Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, Learning from Data, AMLBook, 2012.

I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016

Dinâmica do curso

Sala de aula: aulas expositivas e discussões.

Em casa:

- lista de leitura e exercícios computacionais em Python postados na a cada duas aulas (começando na quarta-feira).

Linguagem \neq Conteúdo