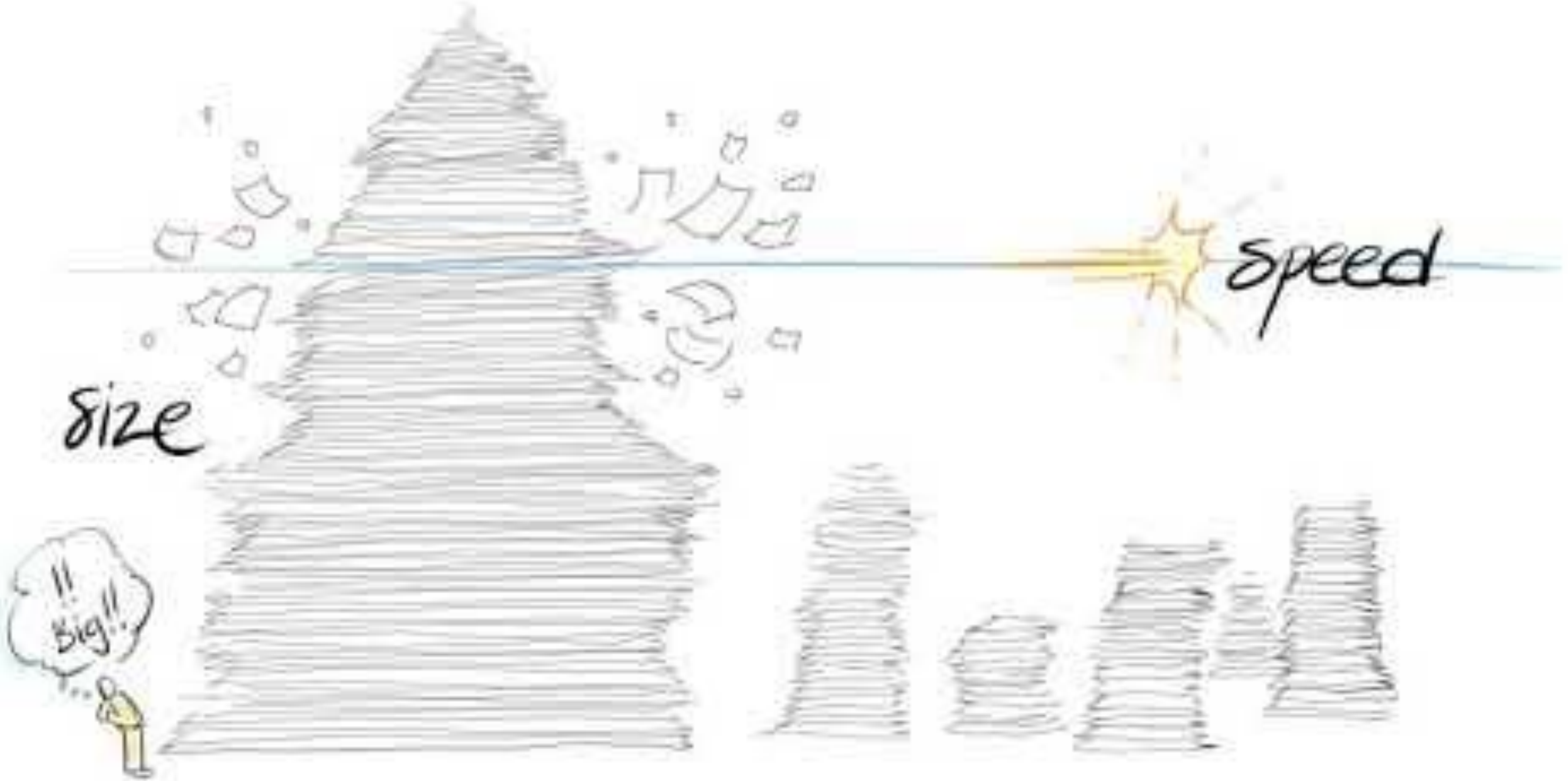
ISE:4172

Big Data Analytics

Stephen Baek
University of Iowa

BIG
DATA
HD
HISTORY.COM
imgflip.com


speed
size
!! Big!!

What is Big Data? (2014)

Every day, we create 2.5 quintillion (1018) bytes of data — so much that 90% of the data in the world today has been created in the last 12 months alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.

Credit: Prof. Amaury Lendasse (U. of Houston)

What is Big Data? ~~(2014)~~ 2015

Every day, we create 2.5 quintillion (1018) bytes of data — so much that ~~90%~~ 95% of the data in the world today has been created in the last 12 months alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.

Credit: Prof. Amaury Lendasse (U. of Houston)

What is Big Data? ~~(2014)~~ ~~2015~~ 2016

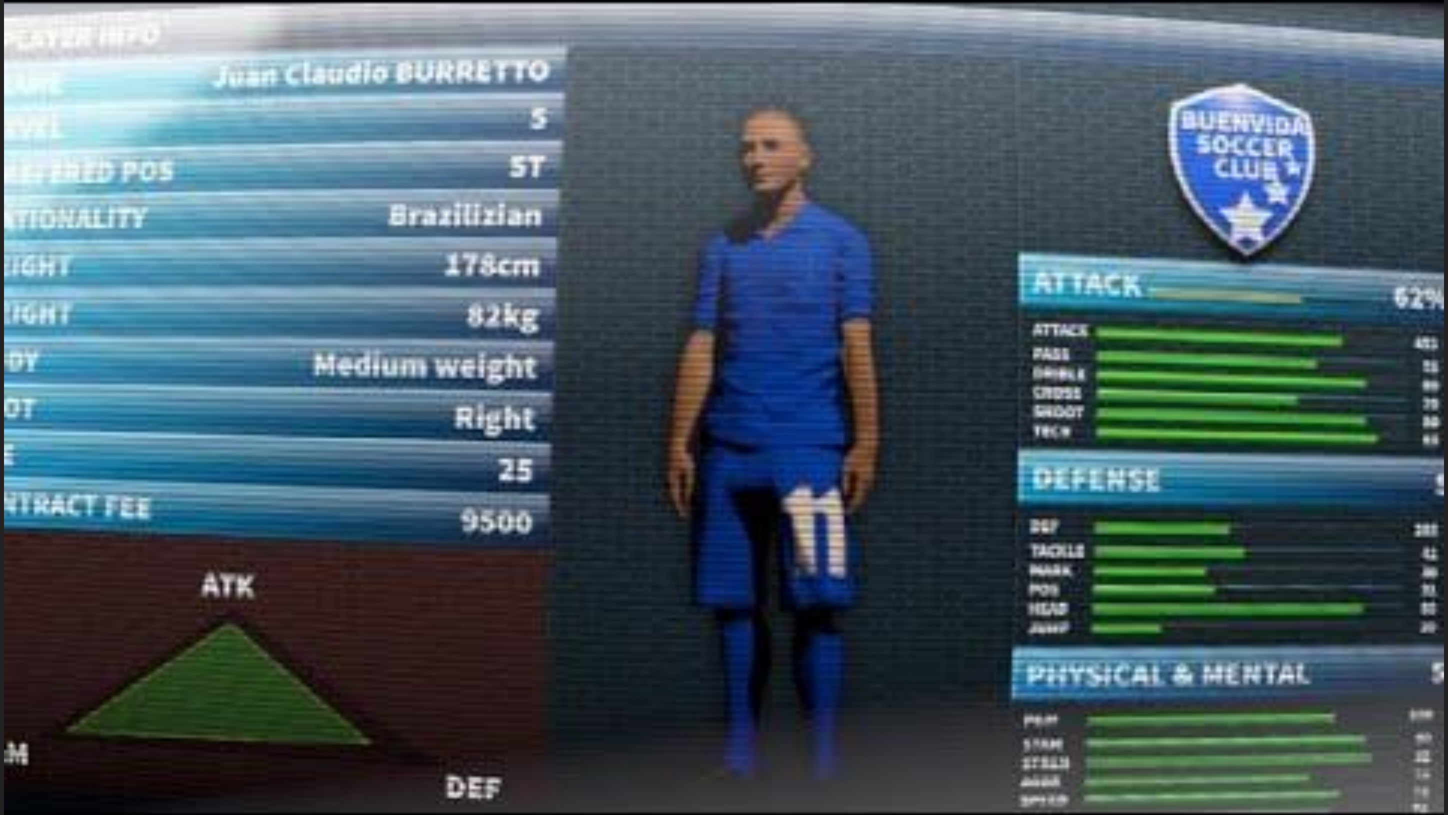
Every day, we create 2.5 quintillion (1018) bytes of data — so much that ~~90%~~ 95% of the data in the world today has been created in the last ~~12~~ 9 months alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.

Credit: Prof. Amaury Lendasse (U. of Houston)

What is Big Data? ~~(2014)~~ ~~2015~~ ~~2016~~ 2017

Every day, we create 2.5 quintillion (1018) bytes of data — so much that ~~90%~~ 95% of the data in the world today has been created in the last ~~12~~ ~~9~~ 6 months alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.

Credit: Prof. Amaury Lendasse (U. of Houston)



PLAYER INFO

AME	Juan Claudio BURRETTO
EVEL	5
EFERRED POS	ST
TIONALITY	Brazilizian
IGHT	178cm
IGHT	82kg
DY	Medium weight
OT	Right
E	25
NTRACT FEE	9500

ATTACK 62%

ATTACK	[illegible]
PASS	[illegible]
DRIBLE	[illegible]
CROSS	[illegible]
SHOOT	[illegible]
TECH	[illegible]

DEFENSE

DEF	[illegible]
TACKLE	[illegible]
MARK	[illegible]
POS	[illegible]
HEAD	[illegible]
JUMP	[illegible]

PHYSICAL & MENTAL

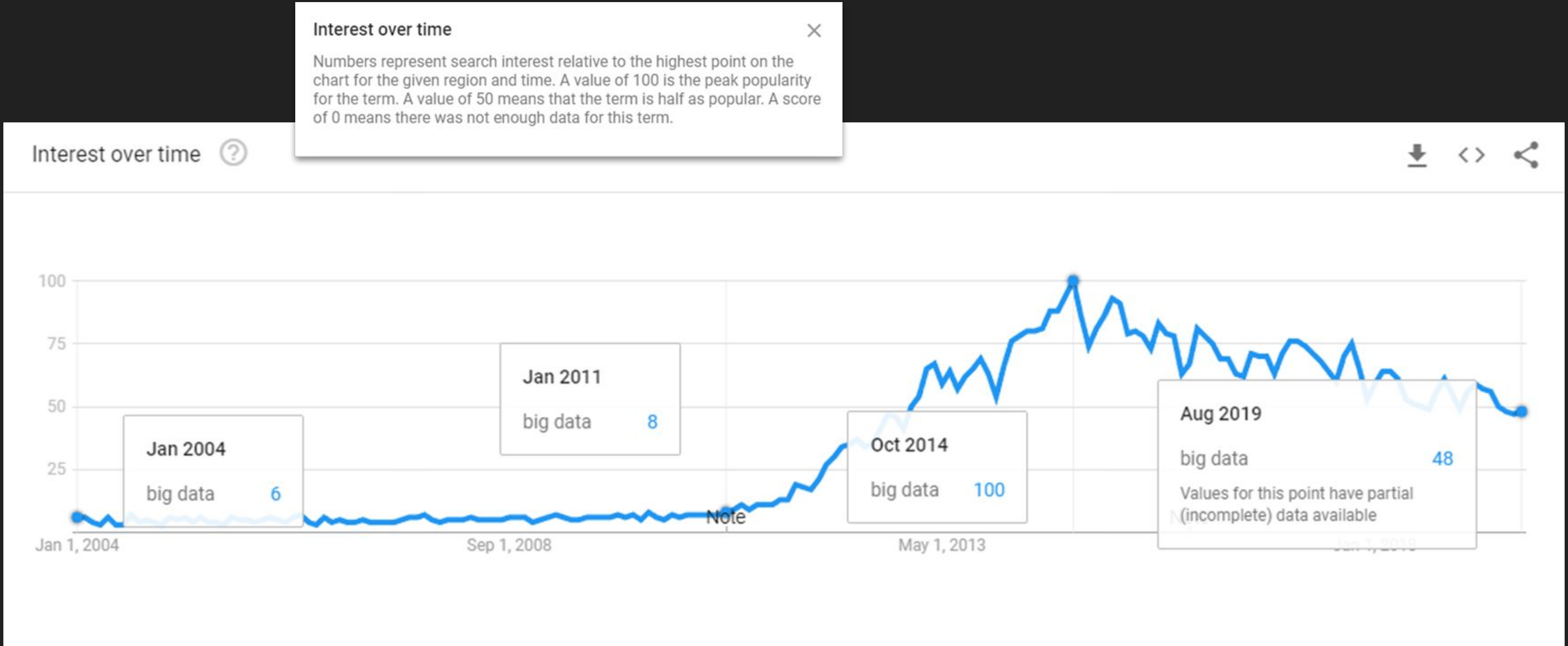
PHY	[illegible]
STAM	[illegible]
STREN	[illegible]
AGGR	[illegible]
SPEED	[illegible]

https://www.youtube.com/watch?v=bcwJ0I9_R-Q

If you were interested...

- Moneyball, Big Data, and the Data Scientist https://www.youtube.com/watch?v=10M_AP9MBg4
- How Big Data Shaped the US Election https://www.youtube.com/watch?v=CgYvf3Ckdso

A Reality Check



Interest over time

Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.

Jan 2004

big data 6

Jan 2011

big data 8

Oct 2014

big data 100

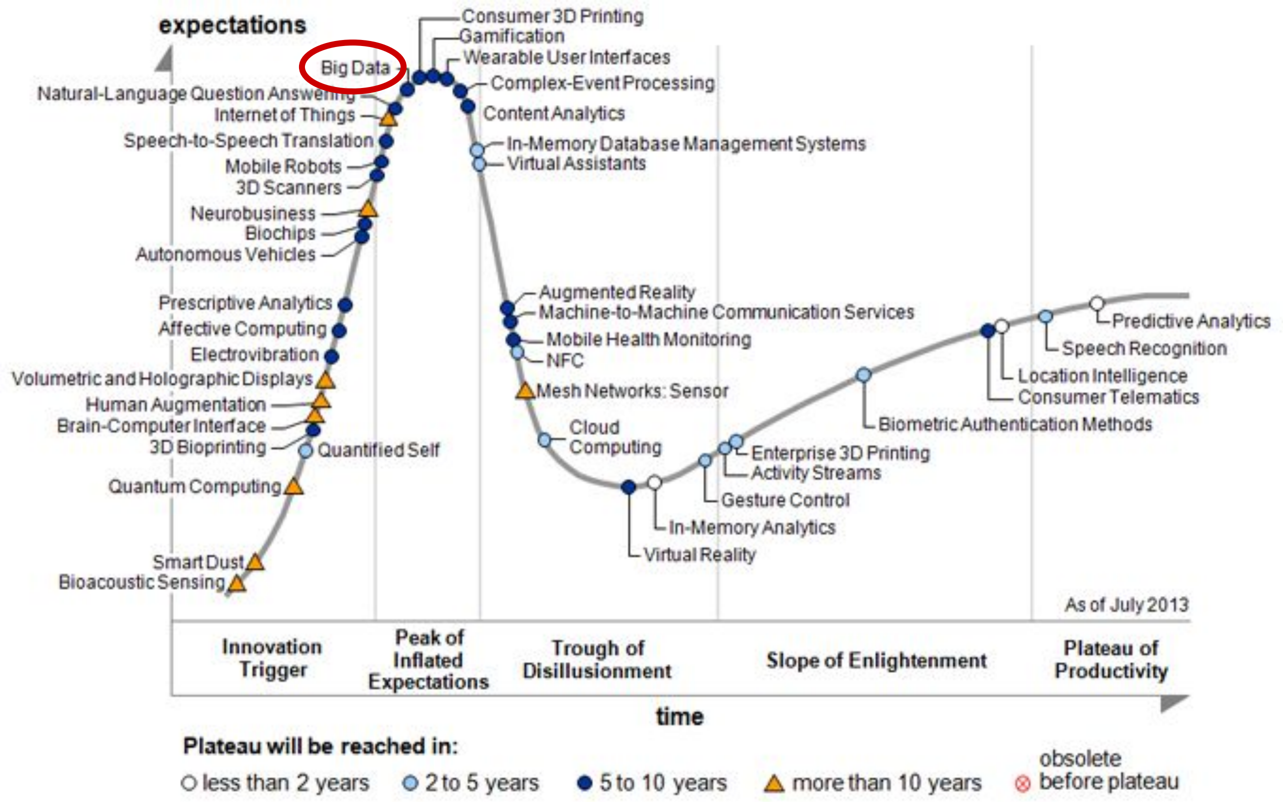
Aug 2019

big data 48

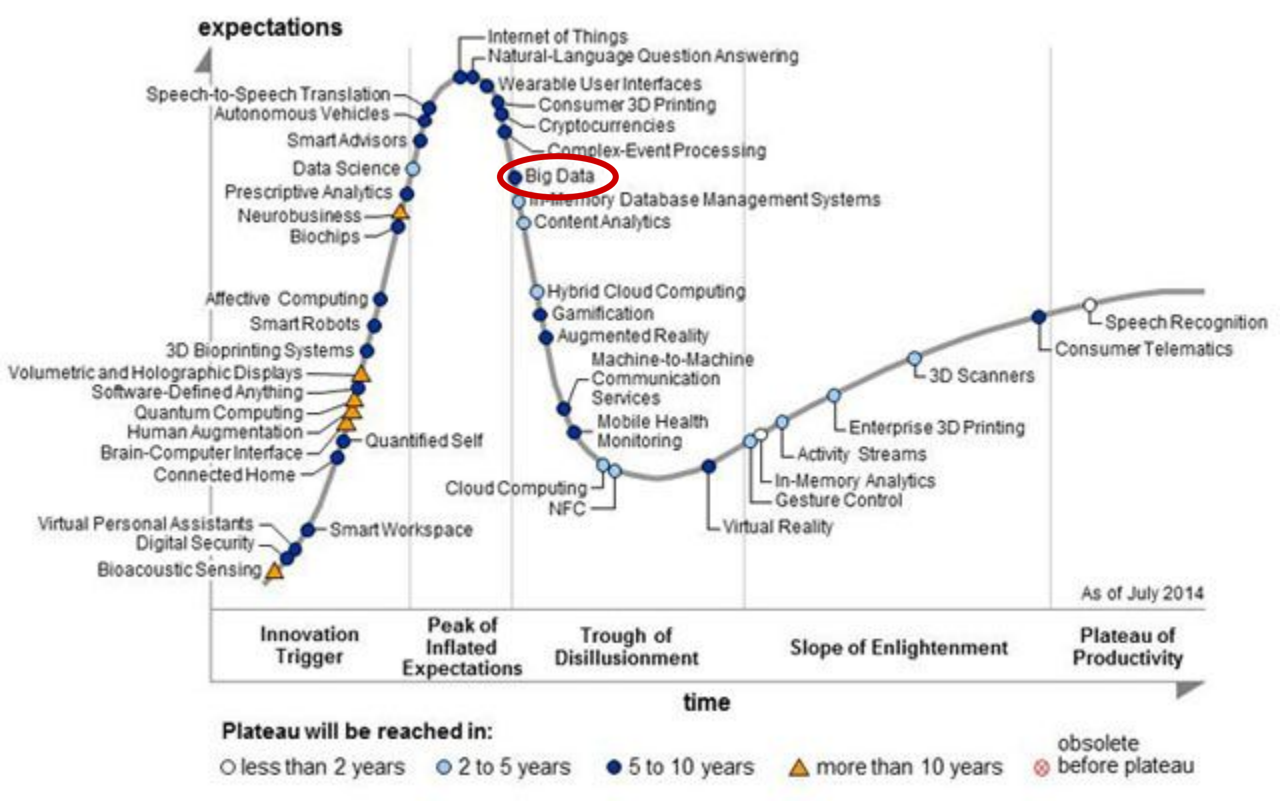
Values for this point have partial (incomplete) data available

https://trends.google.com/trends/explore?date=all&geo=US&q=big%20data

The Gartner Hype Curve (2013)



The Gartner Hype Curve (2014)



Does that mean a downfall?

Don’t just focus on the “hype” of the hype curve

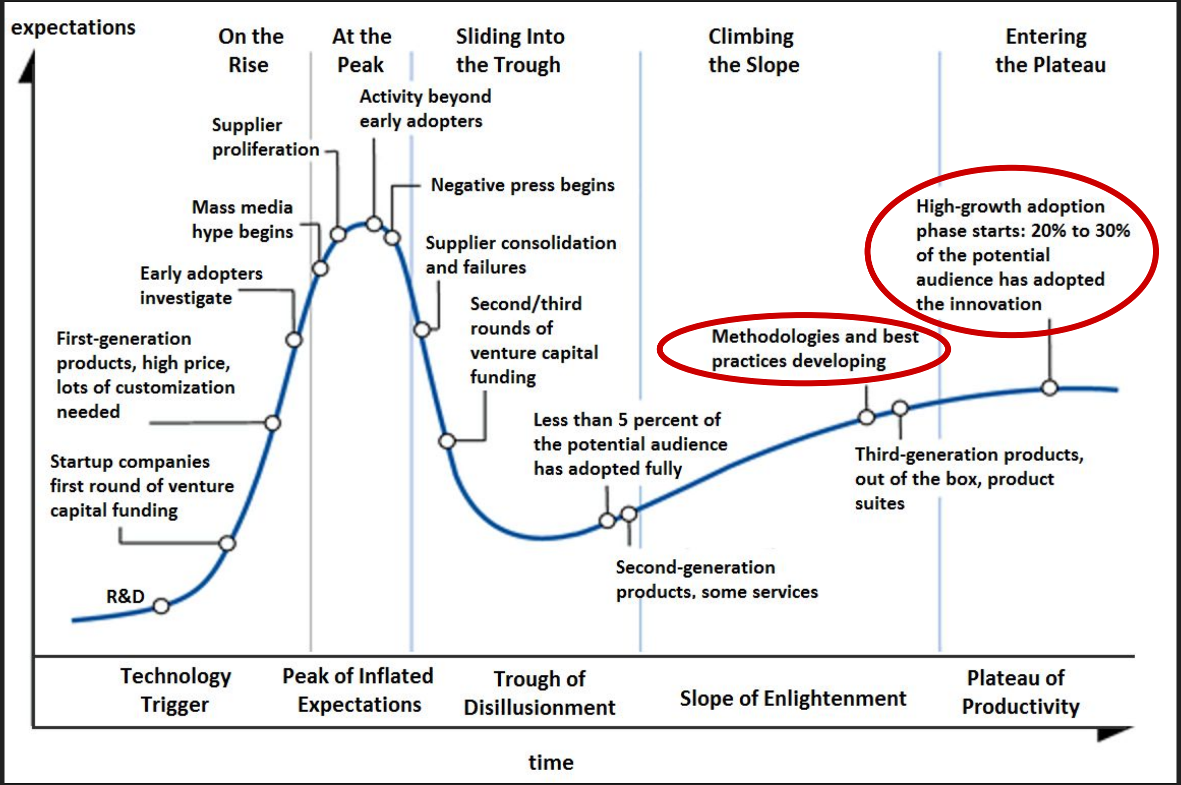
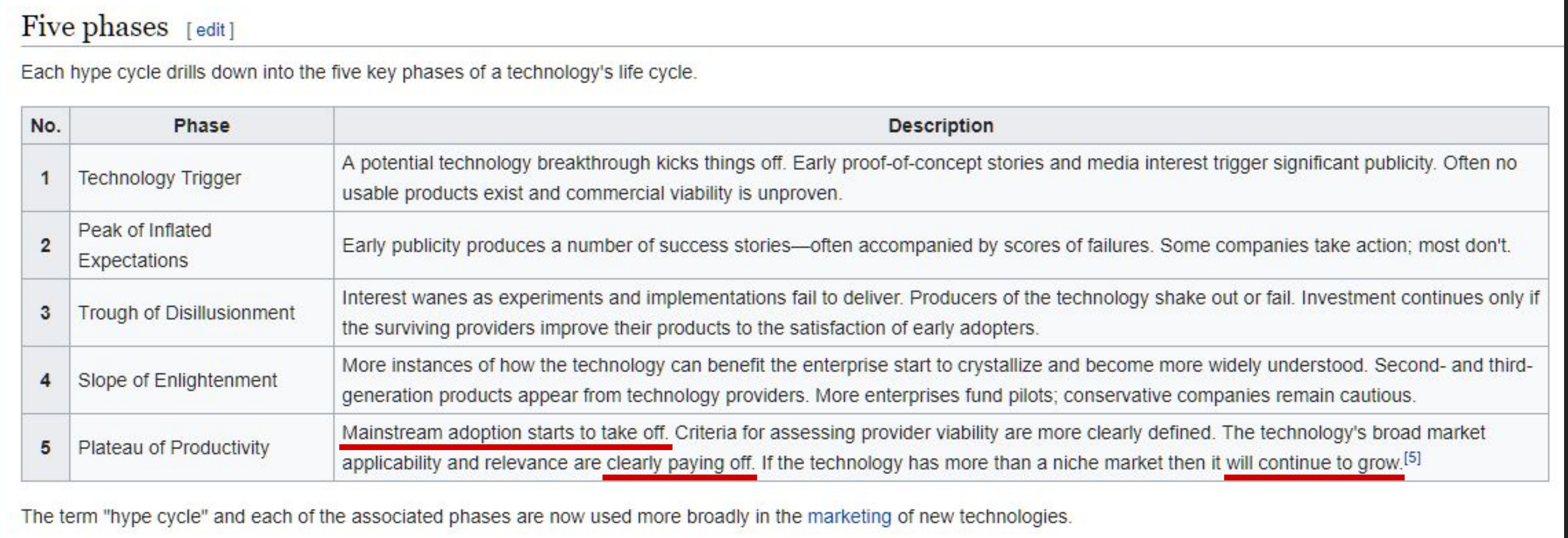

expectations
On the Rise
At the Peak
Sliding Into the Trough
Climbing the Slope
Entering the Plateau
Activity beyond early adopters
Supplier proliferation
Negative press begins
Mass media hype begins
Supplier consolidation and failures
Early adopters investigate
Second/third rounds of venture capital funding
High-growth adoption phase starts: 20% to 30% of the potential audience has adopted the innovation
Methodologies and best practices developing
First-generation products, high price, lots of customization needed
Less than 5 percent of the potential audience has adopted fully
Third-generation products, out of the box, product suites
Startup companies first round of venture capital funding
Second-generation products, some services
R&D
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

Image Source: Wikipedia

Don’t just focus on the “hype” of the hype curve



Five phases [edit]

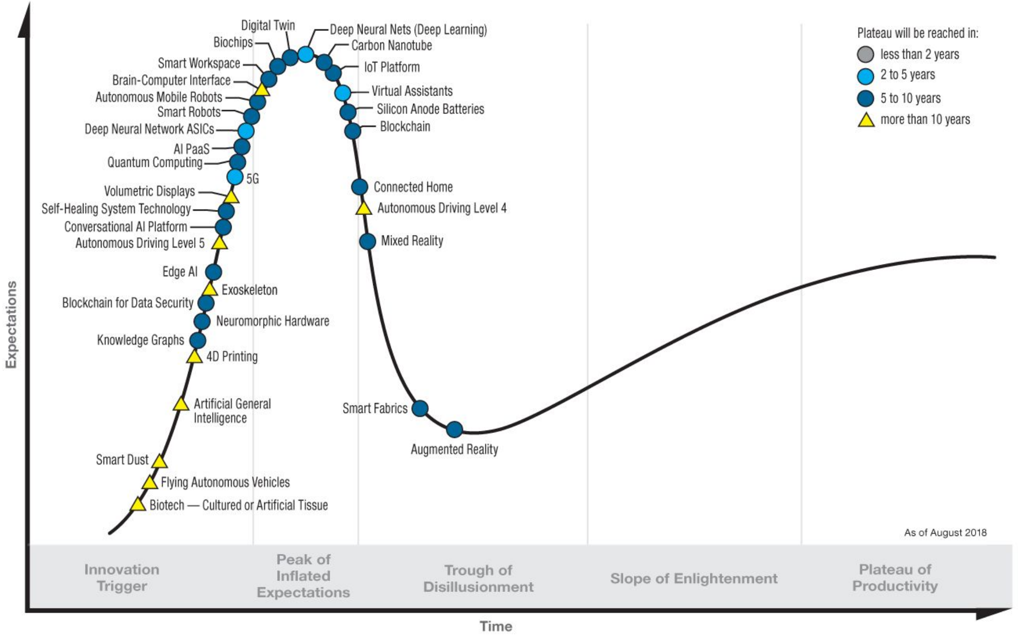
Each hype cycle drills down into the five key phases of a technology's life cycle.

No.	Phase	Description
1	Technology Trigger	A potential technology breakthrough kicks things off. Early proof-of-concept stories and media interest trigger significant publicity. Often no usable products exist and commercial viability is unproven.
2	Peak of Inflated Expectations	Early publicity produces a number of success stories—often accompanied by scores of failures. Some companies take action; most don't.
3	Trough of Disillusionment	Interest wanes as experiments and implementations fail to deliver. Producers of the technology shake out or fail. Investment continues only if the surviving providers improve their products to the satisfaction of early adopters.
4	Slope of Enlightenment	More instances of how the technology can benefit the enterprise start to crystallize and become more widely understood. Second- and third-generation products appear from technology providers. More enterprises fund pilots; conservative companies remain cautious.
5	Plateau of Productivity	Mainstream adoption starts to take off. Criteria for assessing provider viability are more clearly defined. The technology's broad market applicability and relevance are clearly paying off. If the technology has more than a niche market then it will continue to grow.[5]

The term "hype cycle" and each of the associated phases are now used more broadly in the marketing of new technologies.

Source: Wikipedia

Fast Forward to 2018



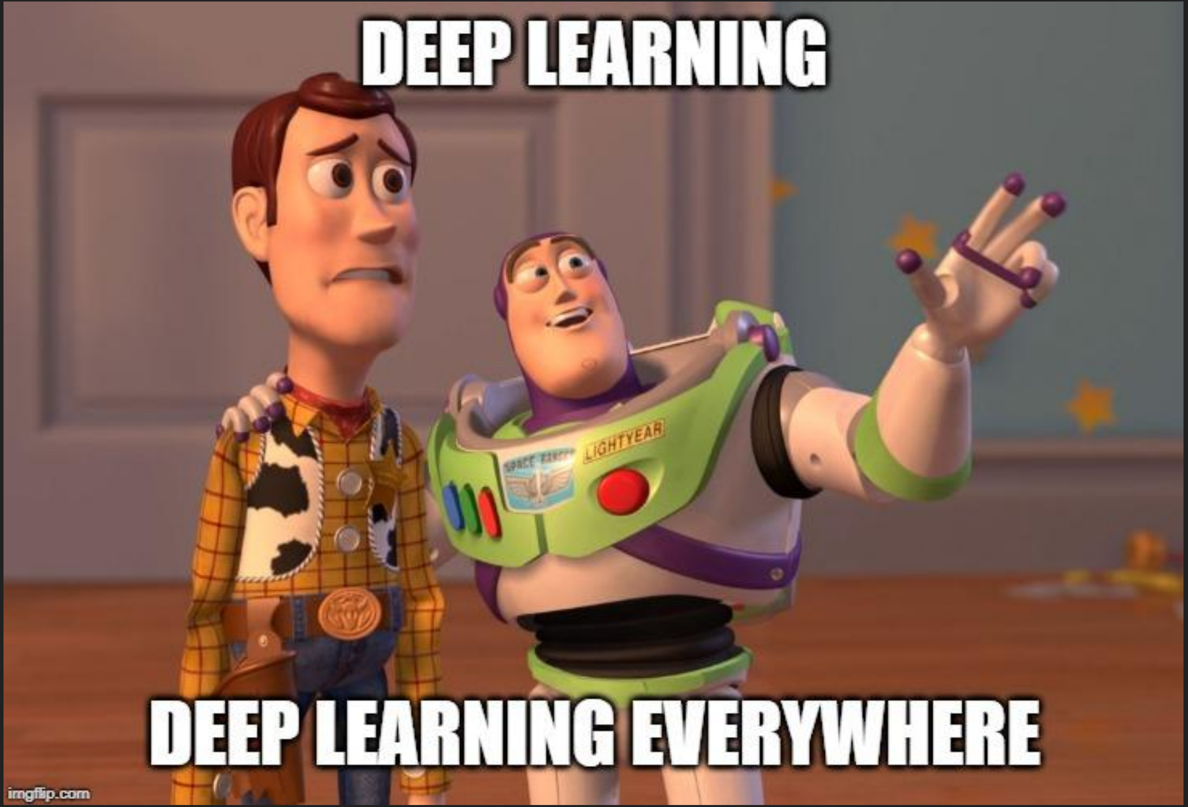
gartner.com/SmarterWithGartner

Source: Gartner (August 2018)

Gartner.

No pun intended

Biggest Buzzword in Recent Years: Deep Learning



“Deep learning is just a nonlinear regression on steroids”

- An anonymous colleague of mine -

The “Steroids”

The “Steroids”

Hardware Accelerators

The “Steroids”

Hardware
Accelerators

BIG DATA

About the Course

Who am I?

- Instructor: Stephen Baek
    - Assistant Professor
        - Industrial and Systems Engineering (Primary)
        - Electrical and Computer Engineering
        - Radiation Oncology
        - Applied Mathematical and Computational Sciences (AMCS)
    - Director
        - Visual Intelligence Laboratory (http://www.stephenbaek.com/lab)
    - Office Hours: 4611 SC, TTh 2:00 pm ~3:00 pm
    - Contact: stephen-baek@uiowa.edu

I’m not a data scientist.

- B.S. (2009) and Ph.D. (2013) in Mechanical and Aerospace Engineering

I’m not a data scientist.

Well, what makes me qualified then?

- B.S. (2009) and Ph.D. (2013) in Mechanical and Aerospace Engineering
- Dissertation: Nonlinear statistical shape analysis
- Research Interest: Shapes (or more precisely, statistical variations of shapes)

I’m not a data scientist.

Well, what makes me qualified then?

- B.S. (2009) and Ph.D. (2013) in Mechanical and Aerospace Engineering
- Dissertation: Nonlinear statistical shape analysis
- Research Interest: Shapes (or more precisely, statistical variations of shapes)
  - AKA “Teaching computers how things look like”
  - Example 1: A skilled manufacturing expert can discern manufacturability and appropriate manufacturing parameters by looking at the shape of a CAD part. Can a machine do the same?

I’m not a data scientist.

Well, what makes me qualified then?

- B.S. (2009) and Ph.D. (2013) in Mechanical and Aerospace Engineering
- Dissertation: Nonlinear statistical shape analysis
- Research Interest: Shapes (or more precisely, statistical variations of shapes)
    - AKA “Teaching computers how things look like”
    - Example 1: A skilled manufacturing expert can discern manufacturability and appropriate manufacturing parameters by looking at the shape of a CAD part. Can a machine do the same?
    - Example 2: Explosion of high-energy material is governed by microstructural geometry. Can we distill that into a mathematical formula?

I’m not a data scientist.

Well, what makes me qualified then?

- B.S. (2009) and Ph.D. (2013) in Mechanical and Aerospace Engineering
- Dissertation: Nonlinear statistical shape analysis
- Research Interest: Shapes (or more precisely, statistical variations of shapes)
    - AKA “Teaching computers how things look like”
    - Example 1: A skilled manufacturing expert can discern manufacturability and appropriate manufacturing parameters by looking at the shape of a CAD part. Can a machine do the same?
    - Example 2: Explosion of high-energy material is governed by microstructural geometry. Can we distill that into a mathematical formula?
    - Example 3: It is believed that there exists an unfair bias associated with people’s body shape in the economic market. How do we quantify that?

I’m not a data scientist.

Well, what makes me qualified then?

- Current research projects
  - Microstructural geometry vs. Explosion characteristics (U.S. Air Force)
  - Visual cues from human operators vs. Workload (NIH, Hyundai Motors + 2)
  - Tumor geometry vs. Cancer survival (NIH)
  - Geometric patterns in gene expressions
  - Human body shape vs. Socioeconomic variables
  - …

I’m not a data scientist.

Well, what makes me qualified then?

- Current research projects
  - Microstructural geometry vs. Explosion characteristics (U.S. Air Force)
  - Visual cues from human operators vs. Workload (NIH, Hyundai Motors + 2)
  - Tumor geometry vs. Cancer survival (NIH)
  - Geometric patterns in gene expressions
  - Human body shape vs. Socioeconomic variables
  - …

“Massive amount of geometric data”

Teaching Assistant

- Yusen He
    - Ph.D. Candidate (Industrial and Systems Engineering)
    - Research Interests: Cancer Informatics
    - Contact: yusen-he@uiowa.edu
    - Office Hours: SC 4317, TTh: 10AM - 12PM

Let me know who you are!

- Let’s go around the room and introduce:
  - Your name
  - Major
  - Reason for taking this course (please keep it short)
  - Other interesting facts about yourself (please keep it short)

Let me know who you are!

- Let’s go around the room and introduce:
  - Your name
  - Major
  - Reason for taking this course (please keep it short)
  - Other interesting facts about yourself (please keep it short)

- It will take weeks (or even months) for me to match your names with your faces
- …, but, don’t be shy and just say hi! I don’t bite :)

Prerequisites

- STAT:2020 PROBABILITY AND STATISTICS FOR THE ENGINEERING AND PHYSICAL SCIENCES
- I will assume you are already familiar with…
    - Probability, Conditional Probability, Bayes Theorem
    - Random Variables, Expectation, Variance
    - XX Distributions where XX can be Uniform, Bernoulli, Binomial, Geometric, Poisson, Normal (Gaussian), & etc.
    - Joint, Marginal, and Conditional Distributions
    - Independence, Conditional Expectation
    - Sampling, Central Limit Theorem
    - Confidence Intervals and Hypothesis Testing, Statistical Significance
    - Correlation and Simple Regression

Prerequisites

- Basic programming skills in at least one scientific programming languages:
    - C/C++
    - Java
    - Python
    - Matlab/Octave
    - R
    - Julia
    - …
- I’ll use Python in class, but you are more than welcome to use your own.

What this course is about

- Basic machine learning algorithms (KNN, LSH, ANN, …)

What this course is about

- Basic machine learning algorithms (KNN, LSH, ANN, …)
- Big data versions of them (efficient computation, less memory options, …)

What this course is about

- Basic machine learning algorithms (KNN, LSH, ANN, …)
- Big data versions of them (efficient computation, less memory options, …)
- Practical data analytics skills using Python (data mining, preprocessing, …)

What this course is about

- Basic machine learning algorithms (KNN, LSH, ANN, …)
- Big data versions of them (efficient computation, less memory options, …)
- Practical data analytics skills using Python (data mining, preprocessing, …)
- A little flavor of the Hadoop big data ecosystem & Amazon Web Services

What this course is about

- Basic machine learning algorithms (KNN, LSH, ANN, …)
- Big data versions of them (efficient computation, less memory options, …)
- Practical data analytics skills using Python (data mining, preprocessing, …)
- A little flavor of the Hadoop big data ecosystem & Amazon Web Services
- Hands-on big data analytics projects + lots of assignments

What this course is about

- Basic machine learning algorithms (KNN, LSH, ANN, …)
- Big data versions of them (efficient computation, less memory options, …)
- Practical data analytics skills using Python (data mining, preprocessing, …)
- A little flavor of the Hadoop big data ecosystem & Amazon Web Services
- Hands-on big data analytics projects + lots of assignments
- Working as a TEAM!

What this course is NOT about

- Computer programming

What this course is NOT about

- Computer programming
- Technology 101

What this course is NOT about

- Computer programming
- Technology 101
- Database systems

What this course is NOT about

- Computer programming
- Technology 101
- Database systems
- Developing all the bolts and nuts

What this course is NOT about

- Computer programming
- Technology 101
- Database systems
- Developing all the bolts and nuts
- Easy, lazy credits

What this course is NOT about

- Computer programming
- Technology 101
- Database systems
- Developing all the bolts and nuts
- Easy, lazy credits
- Individual work

Course Website

- Course GitHub
  - https://github.com/stephenbaek/bigdata
- Submissions and Other Notifications
  - https://icon.uiowa.edu (Iowa students only)

Note: This year, I'm doing an experiment of open-sourcing lecture materials. Some of them are already available on the website/github, but I'll be adding new ones and editing existing ones throughout the semester. Please bear with me!

Grading Policy

- Assignments (40%)
- Final Project Report (40%)
- Final Project Presentation (20%)

Assignments

- There are many.
- Mostly in-class.
- All of them are group assignments.
- Lots of coding.
- My TA and I will be around for a help.

Problem-based Learning

PBL is a student-centered pedagogy in which students learn about a subject through the experience of solving an open-ended problem found in trigger material. The PBL process does not focus on problem solving with a defined solution, but it allows for the development of other desirable skills and attributes. This includes knowledge acquisition, enhanced group collaboration and communication. The process allows for learners to develop skills used for their future practice. It enhances critical appraisal, literature retrieval and encourages ongoing learning in a team environment.

Problem-based Learning

PBL is a student-centered pedagogy in which students learn about a subject through the experience of solving an open-ended problem found in trigger material. The PBL process does not focus on problem solving with a defined solution, but it allows for the development of other desirable skills and attributes. This includes knowledge acquisition, enhanced group collaboration and communication. The process allows for learners to develop skills used for their future practice. It enhances critical appraisal, literature retrieval and encourages ongoing learning in a team environment.

TLDR: Lots of in-class assignments!
& You'll need a laptop for that.

What if you don’t have a laptop?

- You can check out a laptop at the UI Main Library service desk for free.
- See http://www.lib.uiowa.edu/commons/technology/ for the details.
  - Dell Latitude 5430 Laptop (50 available)
    - Notebook is due 24 hours after check-out
    - Fines accrue at the rate of $1.00 per hour, to a maximum of $75.00
    - If the notebook is not returned after the maximum fine is reached (about 5 days) a $1,030.00 replacement charge in addition to the overdue fine will be assessed and forwarded to your account at the University Business Office (U-Bill)
    - Please return the notebook and check out a new one if additional time is needed
    - Please save your documents to either a flash drive, email or OneDrive.
    - Notebook must be returned directly to the Main Library Service Desk

Final Project

- Real problems.
- You will analyze numbers and draw meaningful insights/trends/etc.
- Group competition on 2~3 topics of your choice.
- Not an accuracy competition!
- You will be implementing codes.
- Report + Code + Presentation
- We will have a chance to talk more in details later.

Questions?

Getting Started with ISE:4172

Course GitHub

- Open a web browser and go to: https://github.com/stephenbaek/bigdata
- Open ‘getting_started.md’
- Homework: read through all of them before the next class and come with a laptop configured accordingly.
    - Penalty: you won’t understand a single thing in the next class + you won’t be any help to your group members in the next class

The First In-class Assignment

- Course GitHub > in-class-assignments > ica01
- Click ‘hello_world.ipynb’
- Click ‘Open in Colab’ badge on top.

- ..., or alternatively, https://colab.research.google.com/github/stephenbaek/bigdata/blob/master/in-class-assignments/ica01/hello_world.ipynb

Google Colaboratory

- https://colab.research.google.com
  - A free Python development environment (Jupyter notebook) that requires no setup and runs entirely in the cloud.
  - Colababoratory works with most major browsers, and is most thoroughly tested with desktop versions of Chrome and Firefox.
  - Changes can be made and are visible instantaneously to all users sharing the notebook (like Google Docs).
  - As you log in, a virtual machine is created on the Google's cloud server. Code is executed in a virtual machine dedicated to your account.
  - Virtual machines are recycled when idle for a while and have a maximum lifetime (12 hrs) enforced by the system.
- You'll need a Google account.

In-class Assignment #01

- Run every cell in ‘hello_world.ipynb’
- There are pop quizzes here and there.
- Submission
  - Once you’re done, click ‘File > Download .ipynb’.
  - Rename the downloaded file to ‘ica01-firstname-lastname.ipynb’.
  - Submit it on ICON.

- This assignment requires individual submission, not as a group.
  But please DO WORK AS A GROUP. Discuss, ask questions, etc.

Next Class

- How to read and represent data
- Pandas library