More data, less bases

Luiz Celso Gomes Jr - André Santanchè
MC536 2013/2

Outline

- The data deluge
- Who needs data analysis?
- Architectural challenges
- Data Scientist & Knowledge Economy
- Implications
- Following classes/Assignment
- Invited talks

What do Google, Facebook, Amazon, Walmart, the CIA, and your girlfriend/ boyfriend have in common?


much_fun
tonight
chillsoo
time_with
partying
then_off
whats
bout
jersey_shore
soooo
doin
;)
dont
weekend
last_night_!
ya
?_?
ladies
great_friends
dinner_with
great_night
sooooo
!_!_!
a_blast
hit_me_up
its_gonna
tanning
sunday
comin
party
goin
cant
wait
fam
gotta
havin
ill
ur
beach
haha
cuz
thinkin
tryin
jersey
football
sexy
haters
ready
gettin
love_you
night
with
its
loves
holla
babe
didnt
workin
chillin
baby
my_life
an_amazing
miss
right_now_!
pumped
here_we
excited
big
to_see_my
wanna
girls
text_me
lets
girl
friday
im
lovin
out_with
sooo
call_me
bestie
thats
boys
missin
the_best
aint
wit
dance
feelin
gym
bday
on_my_way
pool
love_u
yall
lookin
lil
we_come
guys
blessed
beautiful
More Extraverted

Less Extraverted

13 to 18 Years Old

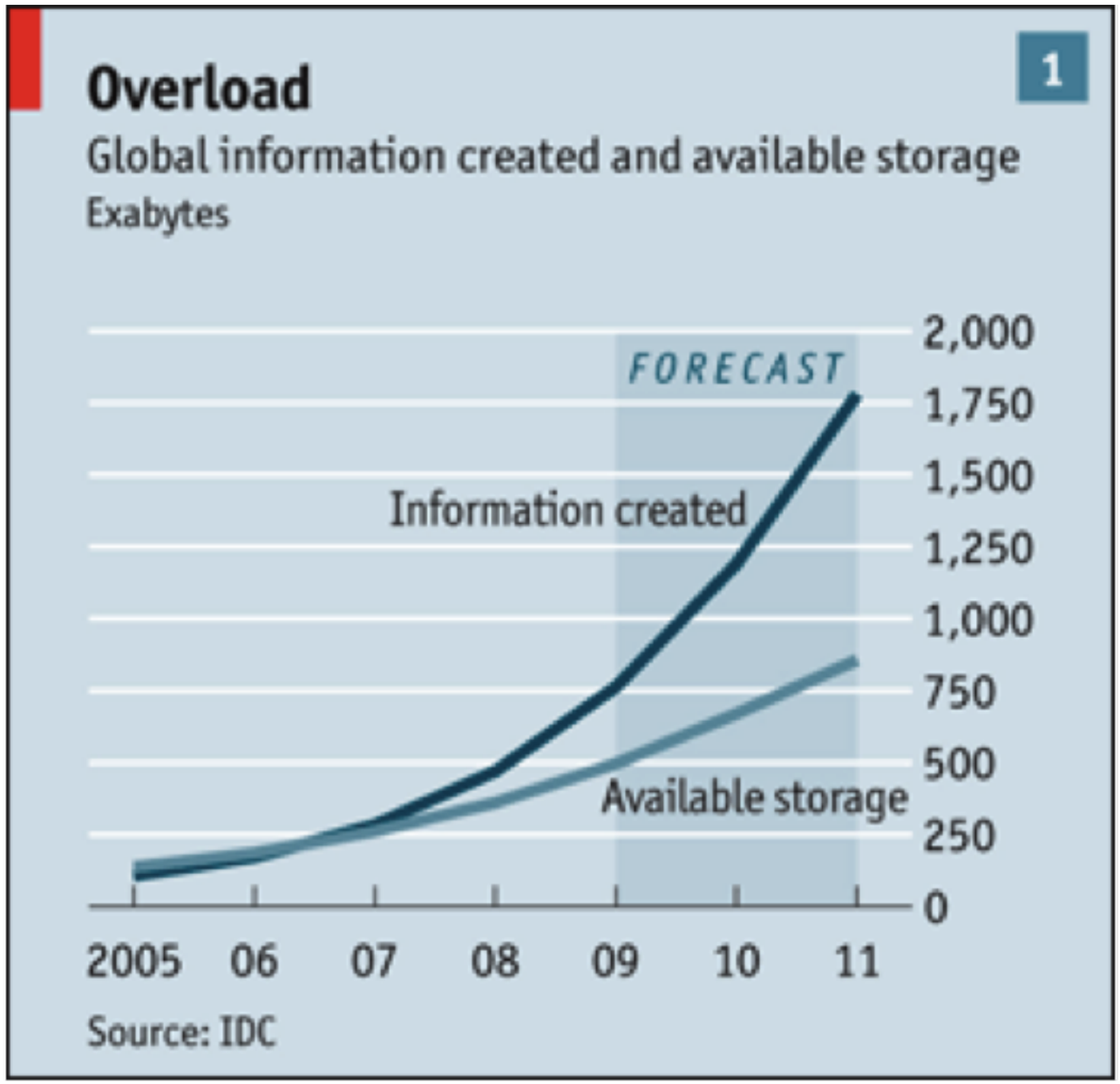
More Than 30 Years Old

What do the LHC (Large Hadron Collider), Exoplanet discovery, Genomics, and a 3-years old asking why, why, why have in common?

What do Google, Facebook, Amazon, Walmart, the CIA, LHC (Large Hadron Collider), Exoplanet discovery, and Genomics, have in common?

(too) Big Data

Big



Over 2 years we create more digital data than all the data created before that.

Enormous

- When the **Sloan Digital Sky Survey** (SDSS) began collecting data in 2000, it amassed **more in its first few weeks than all data collected in the history** of astronomy.

- **140 terabytes of information!**

- the **Large Synoptic Survey Telescope**, successor to SDSS, comes online in 2016 it is anticipated to acquire **that amount of data every five days**.

- the **Large Hadron Collider** (LHC) produced **13 petabytes** of data in 2010 (13,000 terabytes).

Humongous

- **Walmart** handles more than **1 million customer transactions every hour**, which is imported into databases estimated to contain more than **2.5 petabytes of data** - the equivalent of **167 times the information contained in all the books in the US Library of Congress**.

- Every day, **Facebook's 1.15 billion user base** uploads an average of 350 million photos, adding up to a total of **250 billion photos uploaded**.

- Decoding the **human genome** originally **took 10 years** to process; **now it can be achieved in one week**.

- The world's effective **capacity to exchange information** through telecommunication networks was 281 petabytes in 1986, 471 petabytes in 1993, 2.2 exabytes in 2000, **65 exabytes in 2007**

Immense

- 2 billion internet users
- 4.6 billion mobile phones

Titanic


Top 5 Projects Ranked by 2011 U.S. Office-Construction-Starts Value (in $ millions)
PROJECT:
Utah NSA Data Center
ARCHITECT:
Architectural Nexus, KlingStubbins
LOCATION: Bluffdale, UT
VALUE
$1,100
PROJECT:
250 West 55th Street
ARCHITECT:
SOM
LOCATION:
New York, NY
VALUE
$285
PROJECT:
Liberty Mutual Office Building
ARCHITECT:
CBT
LOCATION:
Boston, MA
VALUE
$252
PROJECT:
Facebook Data Center (Building 2)
ARCHITECT:
Sheehan Partners
LOCATION:
Prineville, OR
VALUE
$200
PROJECT:
ExxonMobil Houston Campus
ARCHITECT:
Gensler, Pickard Chilton, PDR
LOCATION:
Spring, TX
VALUE
$160

Big, enormous, humongous, titanic applications

web logs; RFID; sensor networks; social networks; social data (due to the Social data revolution), Internet text and documents; Internet search indexing; call detail records; astronomy, atmospheric science, genomics, biogeochemical, biological, and other complex and/or interdisciplinary scientific research; military surveillance; medical records; photography archives; video archives; and large-scale eCommerce

The data deluge

Big data vs Small computers

More data is good, but...

Who needs data?

- **IT industry - case: Google**
- **Commerce - case: Amazon**
- Media industry - e.g: Netflix/Pandora
- **Consumer goods industry - case: Toyota vs. Tesla**
- **Startups - case: hot startups**
- Government - e.g: NSA
- Science - e.g: e-Science
- Everybody - case: you

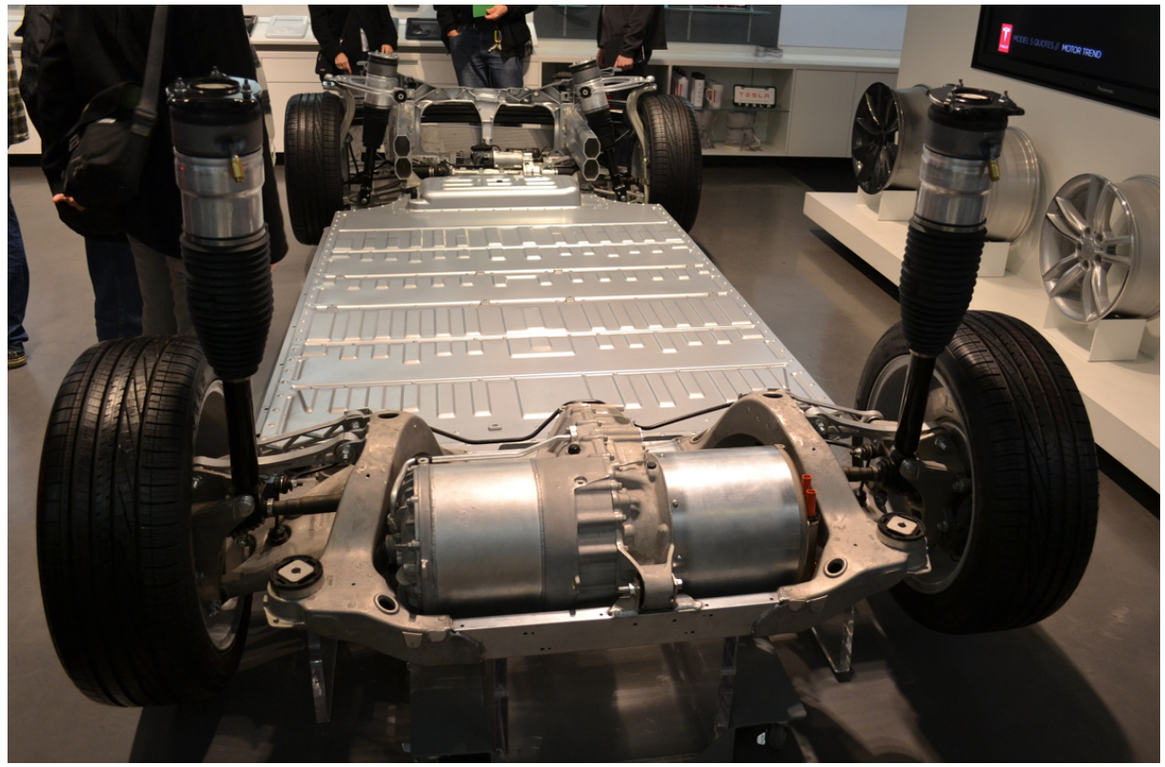
Google

- First company to “see the Matrix” (big opportunities in large data analysis)
- The company invested a record $1.6 billion in its data centers in the second quarter of 2013.
- Google is one of the biggest PC makers in the world
- Data, data, data.... Google Translator - Self-driving car
- Sergey Brin: “We Want Google To Be The Third Half Of Your Brain.”
- Android, Google +, Google Fiber, Project Loom... Maybe: “We want you to be the third half of Google’s brain”
- Ray Kurzweil, Deep Learning

Amazon

Toyota vs. Tesla

- Toyota
    - Largest automobile manufacturer (11th largest company overall)
    - Pioneer and leader in hybrid (gasoline-electric) technology
- Tesla Motors
    - 10 years old company, pure electric, based in the Silicon Valley
    - 400% valorization in the last year, leader in important markets
    - Best car awards, best aerodynamics, safest car ever
    - Electric car market is growing faster than hybrids in their introduction
    - E



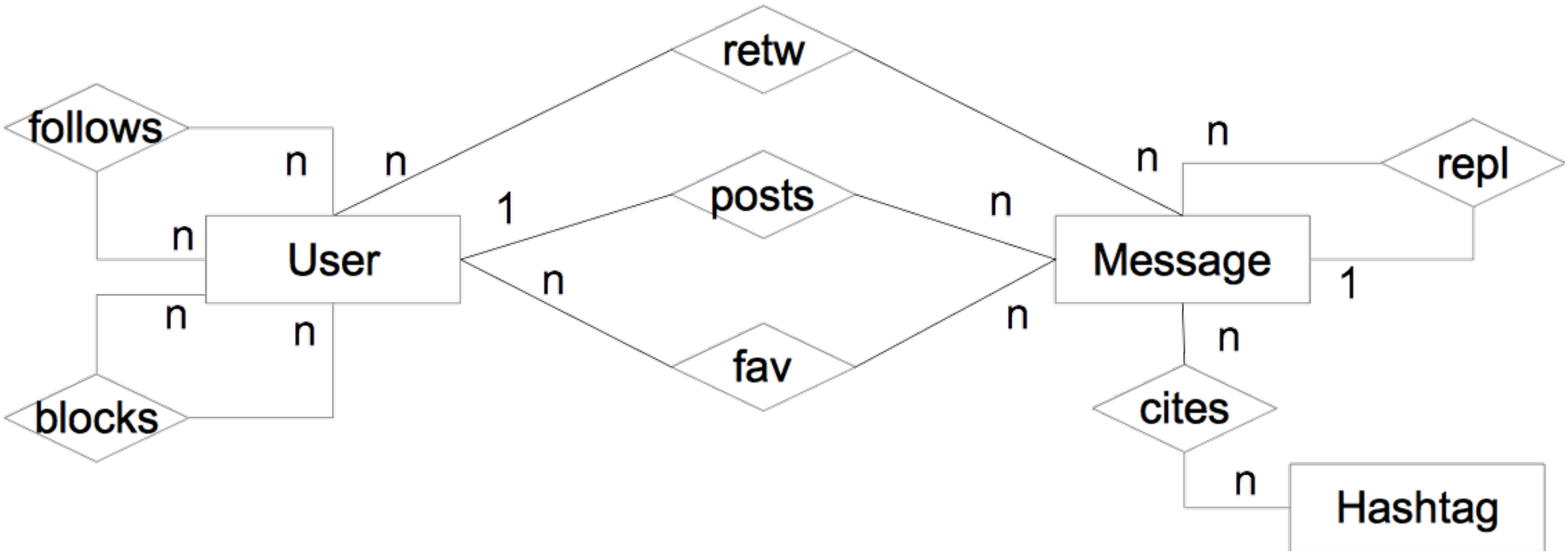
Toyota vs. Tesla

- Should Toyota change focus to all-electric cars?
- Data analysis
    - Market growth
    - Technology cost
    - Consumer interest
    - Political stability (gas prices)
- Is Toyota the next IBM/Microsoft/BlackBerry?

Startups

- Entrepreneur: Top 30 Startups to Watch 2013
- The CNN 10: Startups to watch 2013

Architectural challenges - modeling



Twitter

Architectural challenges - languages

```
CREATE VIEW TotalRep AS
SELECT P.pid, count(*) AS ct
FROM Post P NATURAL LEFT JOIN Retweet
GROUP BY P.pid
```

```
SELECT P.pid, ct
FROM Follow F JOIN Post P ON P.uid = F.uid2
NATURAL JOIN TotalRep
WHERE F.uid1 = 'userX'
ORDER BY ct
```

Popular posts from followed users

Architectural challenges - languages

```
CREATE VIEW TotalRep AS
SELECT P.pid, count(*) AS ct
FROM Post P NATURAL LEFT JOIN Retweet
GROUP BY P.pid
```

5 billion X 100 billion

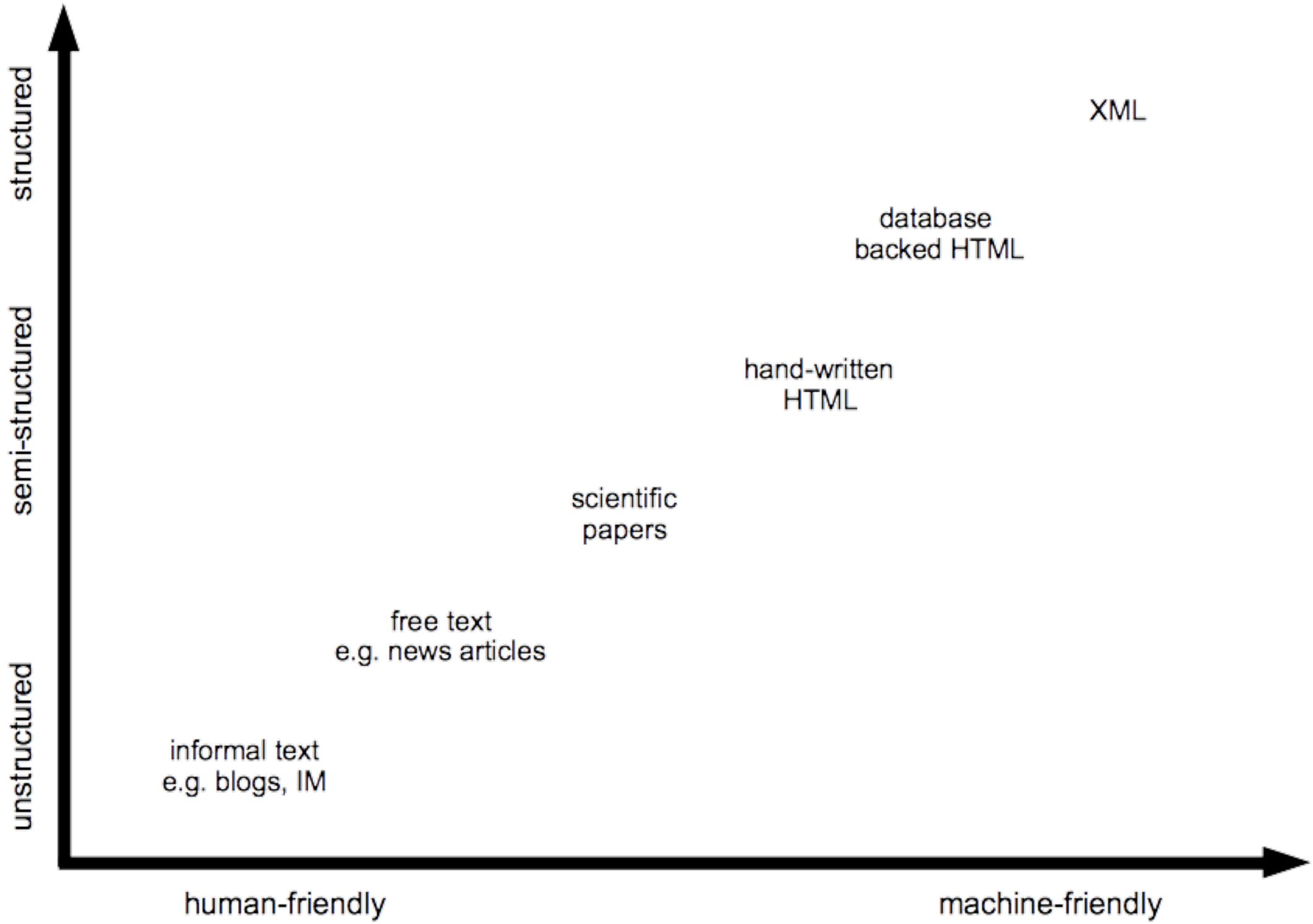
```
SELECT P.pid, ct
FROM Follow F JOIN Post P ON P.uid = F.uid2
NATURAL JOIN TotalRep
WHERE F.uid1 = 'userX'
ORDER BY ct
```

Popular posts from followed users

Architectural challenges - data source variety

- Structured data
  - Relational
  - Streams (sensors, stock exchange)
- Semi-structured data
  - XML
  - Graphs
- Unstructured data
  - Text documents
  - Speech
  - Images
  - Video

Structure spectrum



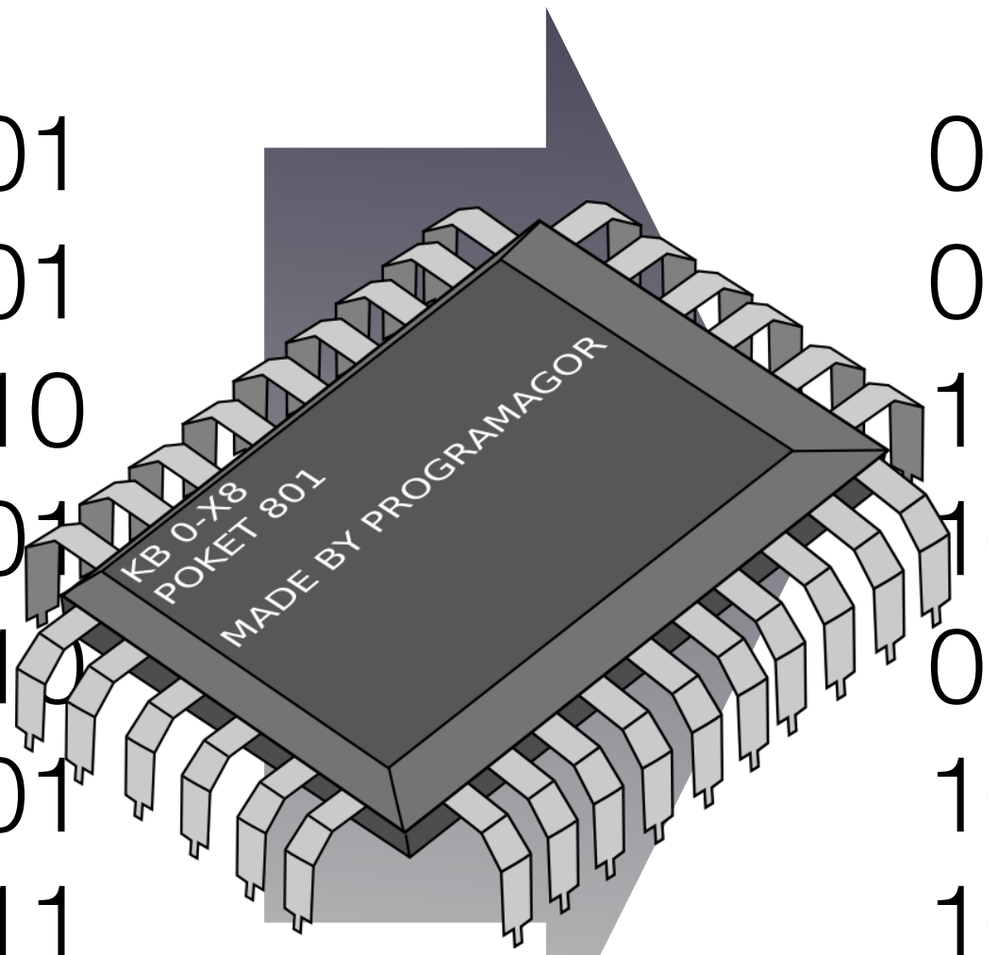
Computing??

11100101101
01010100101
01010101010
10100010101
01001110110
10101010101
00101010111

00100101101
01111001111
11010101010
10001010100
01010111011
10001110101
10011011111

Computing

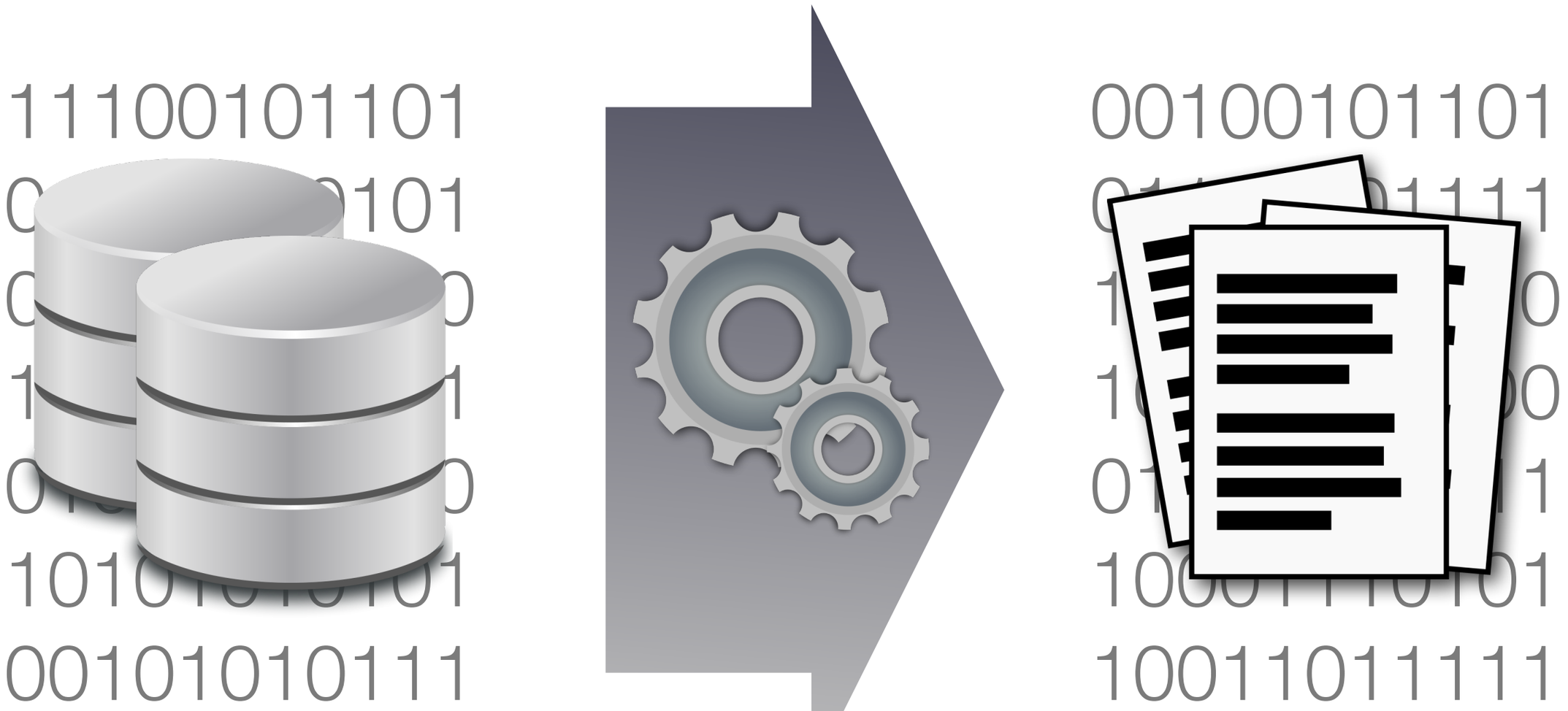
11100101101
01010100101
01010101010
10100010101
01001110110
10101010101
00101010111



00100101101
01111001111
11010101010
10001010100
01010111011
10001110101
10011011111

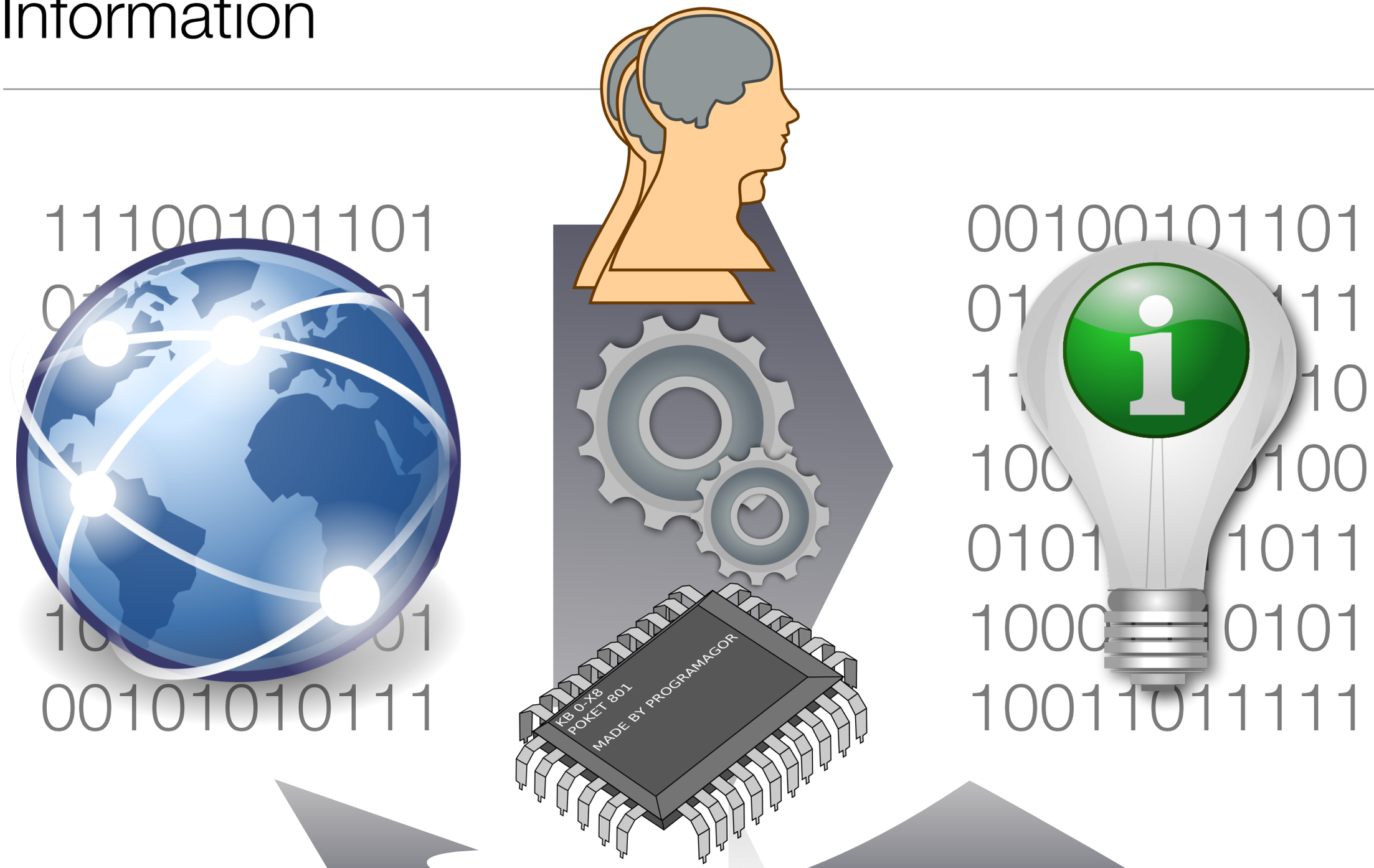
Before: one architecture used for most problems

Languages, data models, frameworks



Then: few models used for most problems

Information

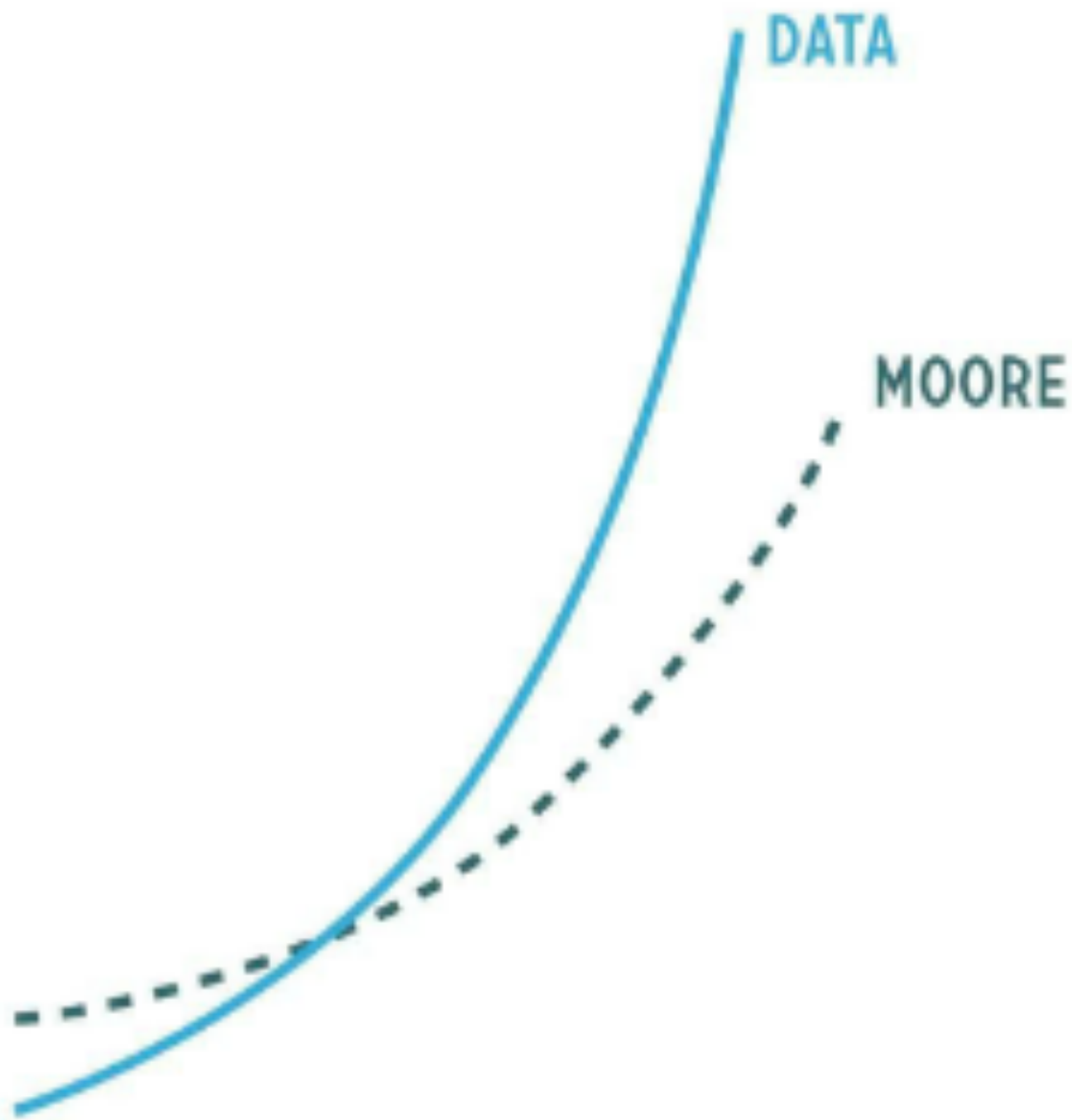


Now?

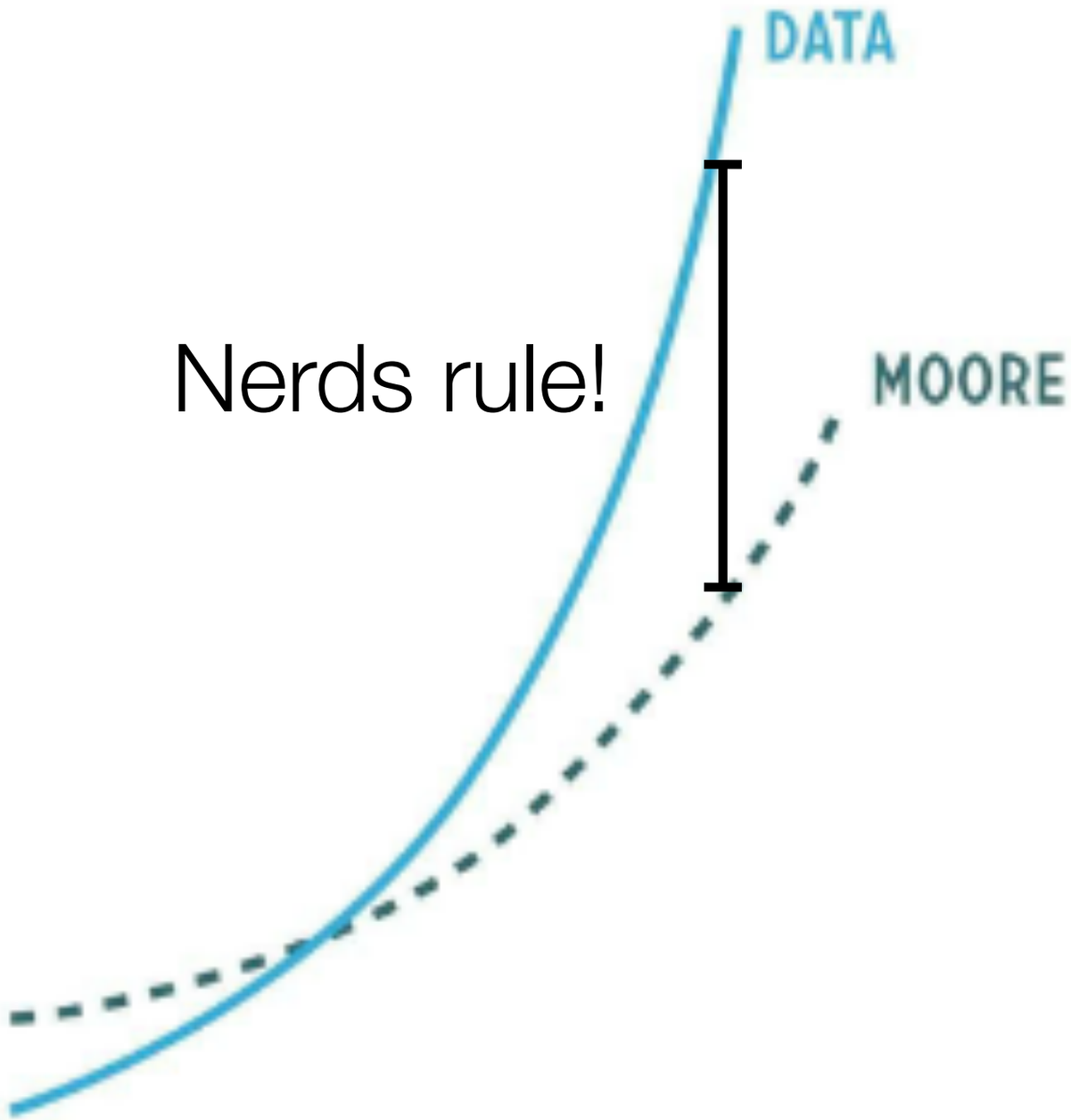
New requirements

- Fast (quicker than the competition)
- Scalable (grows smoother than the competition)
- Distributed (more available than the competition)
- Smart (better inferences than the competition)
- Multi-talented (processes more data types and sources than the competition)
- Flexible (adapts faster than the competition)

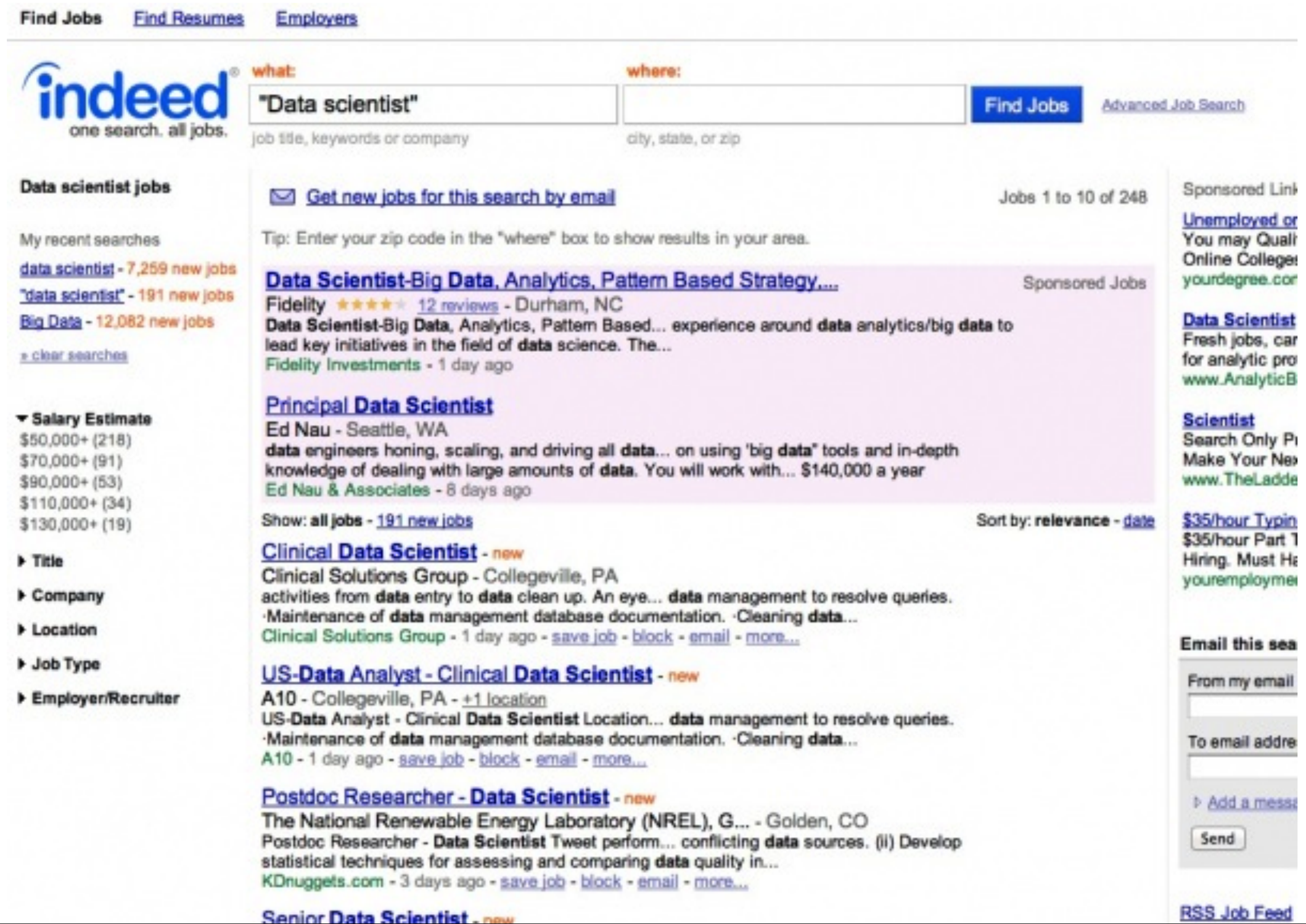
Big data vs Small computers

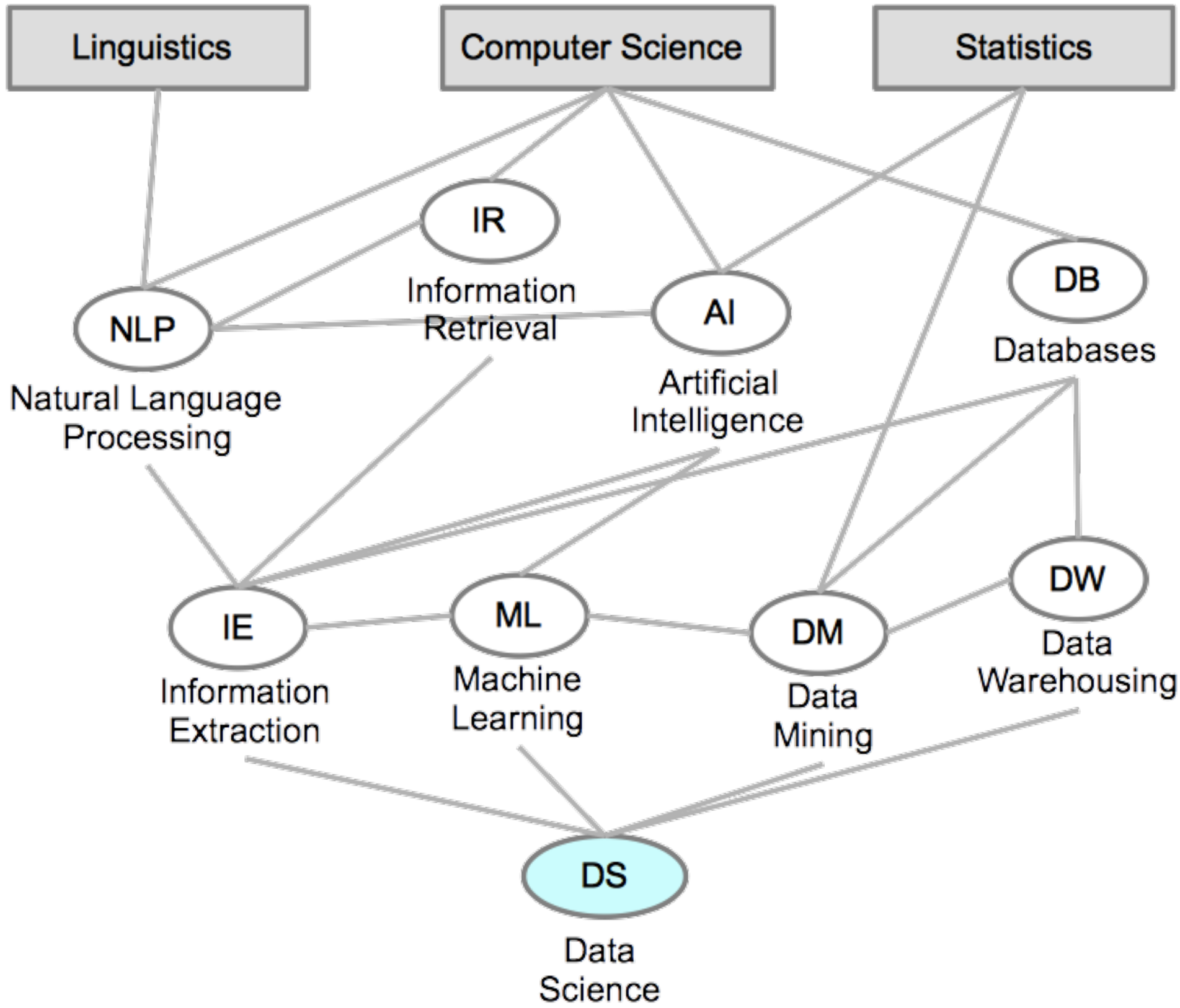


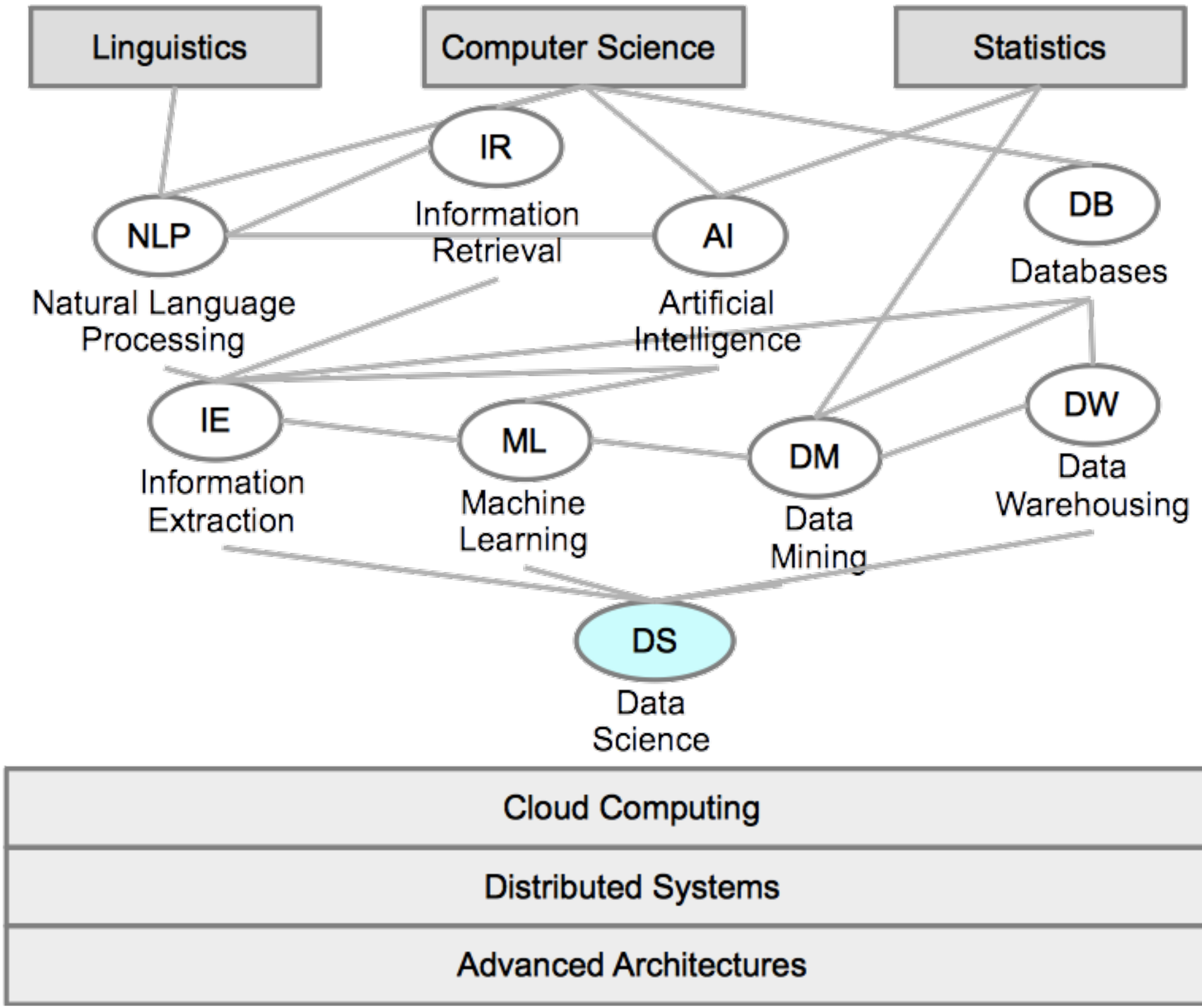
Small computers vs Smart programmers



The Data Scientist




Linguistics
Computer Science
Statistics
IR
Information Retrieval
NLP
Natural Language Processing
AI
Artificial Intelligence
DB
Databases
IE
Information Extraction
ML
Machine Learning
DM
Data Mining
DW
Data Warehousing
DS
Data Science


Linguistics
Computer Science
Statistics
IR
Information Retrieval
NLP
Natural Language Processing
AI
Artificial Intelligence
DB
Databases
IE
Information Extraction
ML
Machine Learning
DM
Data Mining
DW
Data Warehousing
DS
Data Science
Cloud Computing
Distributed Systems
Advanced Architectures

The Knowledge Economy

- It refers to the use of **knowledge technologies** (such as knowledge engineering and knowledge management) to produce **economic benefits** as well as job creation.
- An interconnected, globalized economy where **knowledge resources** such as know-how and expertise are **as critical as other economic resources** [Wikipedia]
- **Data assets** are a major component of the balance sheet, replacing traditional physical assets of the 20th century; there is a Widespread recognition of the **value of data even beyond traditional enterprise boundaries** [Ralph Kimball]

The Knowledge Economy

- What will you answer when a shareholder asks: What is our KROI? **Knowledge Return on Investment**?
- You have **knowledge to sell**. How are you going to package it up to attract buyers?
- **KE HICSS 2013** : Knowledge Economics

Big Data now

- IBM Watson wins Jeopardy and now aims at Big Data
- OMGPOP (Draw Something) scaled to 36 million users in three weeks
- Google Translator
- Obama Administration is announcing the “Big Data Research and Development Initiative”/NYC BigData challenge

Social impact/Future

- Competition – the faster the better
- Algorithms controlling everything
- Entire population online (~5 years!)
- Politics
- Health-care
- Privacy
- Cyberwars
- Energy consumption (now 10%!)
- Google, Ray Kurzweil, Deep Learning

Soon...

- Data Warehousing/Data Mining
- Distributed Databases/Alternative Architectures
- NoSQL/Map Reduce
- Databases on the Web/Integration
- eScience
- Graph Databases/Complex Networks
- Information Retrieval
- Machine Learning/Natural Language Processing
- Geospatial Information Systems/Multimedia Databases

Trabalho extra

- aplicar alguma das tecnologias discutidas na segunda metade do curso
- exemplos:
    - técnicas de data mining para recomendação
    - uso de bancos de dados não tradicionais
    - extrair, indexar e processar texto
- 2-3 semanas de tempo extra (especificação em breve)