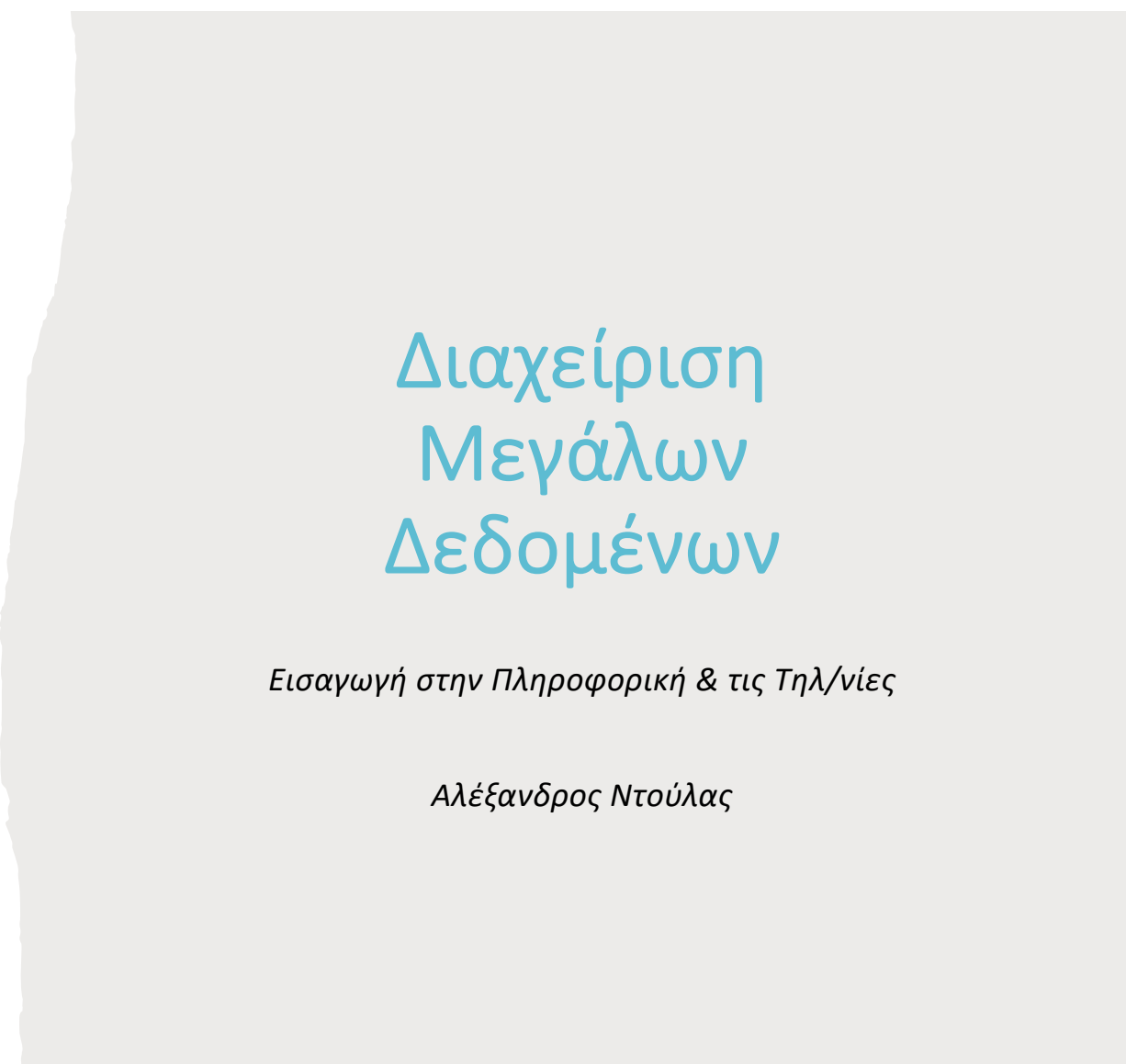




Αλέξανδρος Ντούλας

Τι είναι τα Μεγάλα Δεδομένα (Big Data)

- Δεδομένα που είναι τόσο μεγάλα και περίπλοκα που δεν μπορούμε να τα επεξεργαστούμε με παραδοσιακά συστήματα
- Προκλήσεις στα Μεγάλα Δεδομένα:
 - Συλλογή (capture)
 - Επιμέλεια (curation)
 - Αποθήκευση (storage)
 - Αναζήτηση (search)
 - Διαμοίραση (sharing)
 - Μεταφορά (transfer)
 - Ανάλυση (analysis)
 - Εξαγωγή Γνώσης & Μάθηση (knowledge extraction & learning)
 - Οπτικοποίηση (visualization)

Παράδειγμα

- Έχετε 1024 δίσκους του 1 TB (= 1 PB)
- Θέλουμε να στείλουμε τα δεδομένα στη Θεσσαλονίκη
- Προτάσεις?

Τρόπος Α: Αντιγραφή Μέσω Δικτύου

- Έστω γραμμή 100 Mbps
 - $100 \text{ Mbps} = 100 / 8 = 12.5 \text{ MB/s}$
 - Για 1 TB : $1024 \times 1024 \text{ MB} / 12.5 \text{ MB/s} = 83,886.08 \text{ seconds}$
 - Seconds in 1 day : 86,400 seconds
- Γραμμή 1Gbps
 - ~2.5 hours
- Θυμηθείτε εμείς έχουμε $1 \text{ PB} = 1024 \text{ TB}$

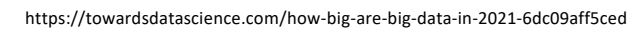
Τρόπος Β'

- Όγκος 1 δίσκου 3.5" = 4" x 5.8" x 0.8" = 0.304 λίτρα
- Όγκος 1024 δίσκων = 311 λίτρα
- Στον προορισμό σε ~7 ώρες (με διάλειμμα για μπουγάτσα)

“Never underestimate the bandwidth of a station wagon full of tapes hurtling down the highway.”

— Andrew S. Tannenbaum







CERN's Large Hadron Collider (LHC) generates 15 PB a year

Big Data Centers

- ~706 στρέμματα



<https://expedient.com/knowledgebase/tools-and-calculators/file-transfer-time-calculator/>

Facebook — Prineville, Oregon, United States - ~102 στρέμματα



- Microsoft
Υποβρύχια
Data Centers



<https://news.microsoft.com/innovation-stories/project-natick-underwater-datacenter/>

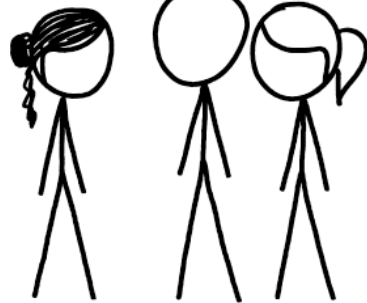


RAID CONTROLLERS DON'T MAKE SENSE AT OUR SCALE; EVERYTHING IS REDUNDANT AT HIGHER LEVELS. WHEN A DRIVE FAILS, WE JUST THROW AWAY THE WHOLE MACHINE.



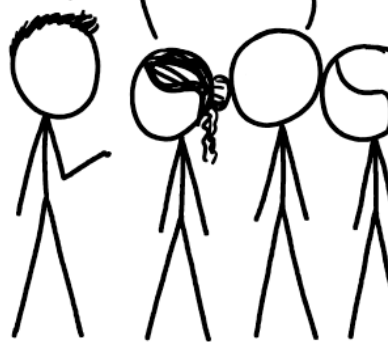
MACHINE? WE THROW AWAY WHOLE RACKS AT A TIME.

YEAH, WHO REPLACES ONE SERVER?



WE JUST REPLACE WHOLE ROOMS AT ONCE. AT OUR SCALE, MESSING WITH RACKS ISN'T ECONOMICAL.

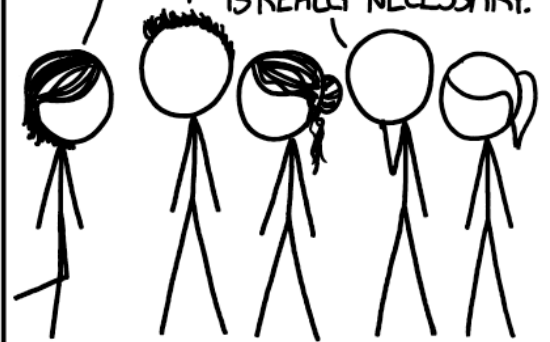
WOW.
LIKE GOOGLE!



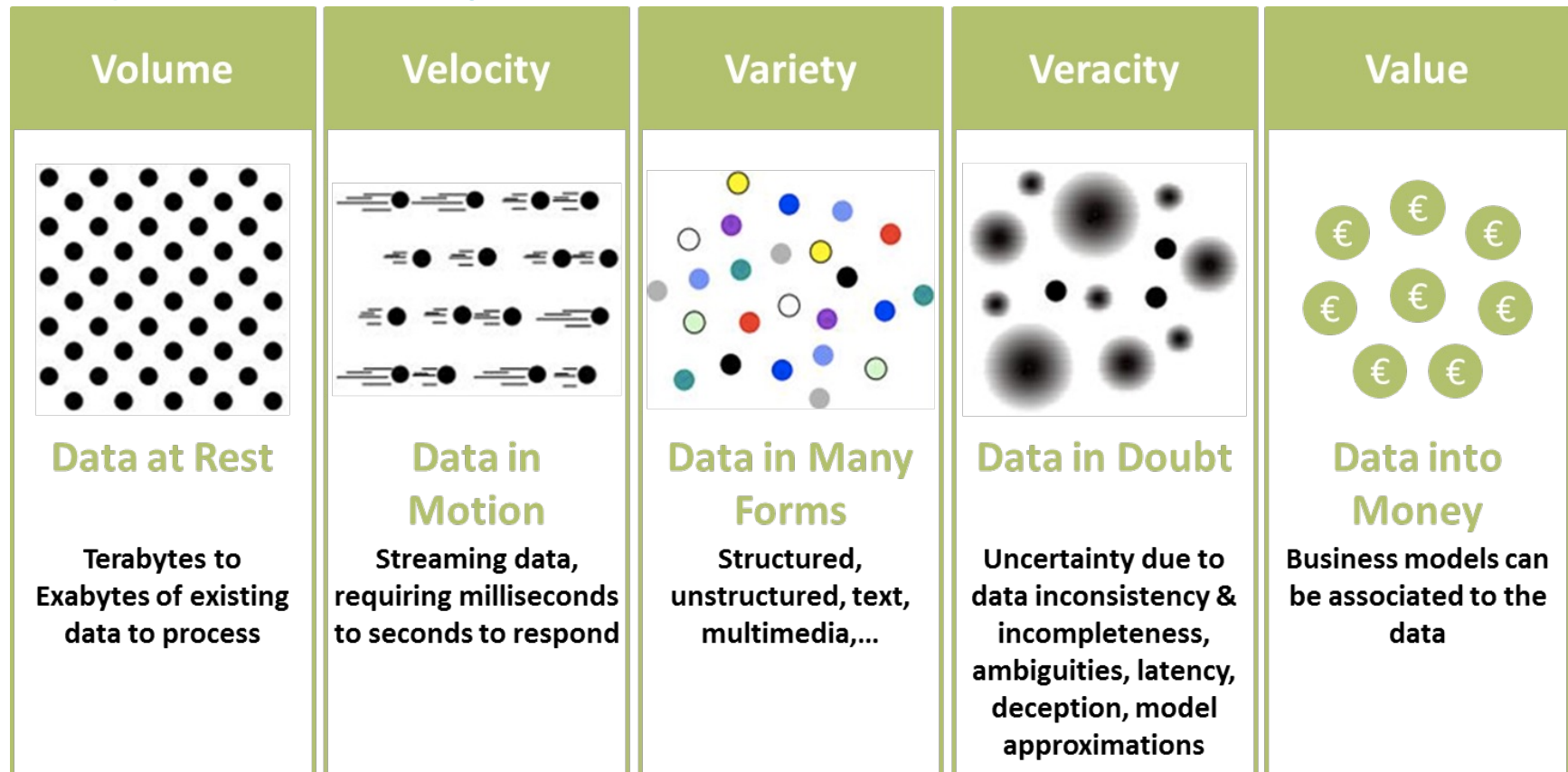
WE DON'T HAVE SPRINKLERS OR INERT GAS SYSTEMS. WHEN A DATACENTER CATCHES FIRE, WE JUST ROPE IT OFF AND REBUILD ONE TOWN OVER.

MAKES SENSE.

I WONDER IF THE ROPE IS REALLY NECESSARY.



Μεγάλα Δεδομένα: Τα 5 V



Adapted by a post of Michael Walker on 28 November 2012

<http://informationcatalyst.com/wp-content/uploads/2015/09/BD-5Vs.png>

Velocity (Speed)

- Δεδομένα που δημιουργούνται γρήγορα και πρέπει να επεξεργαστούν γρήγορα
- Online Data Analytics
- Αργές Αποφάσεις ➔ Χαμένες Ευκαιρίες

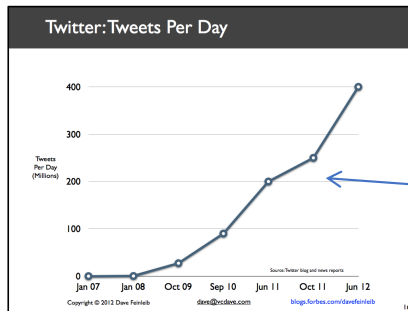


- **Παραδείγματα**

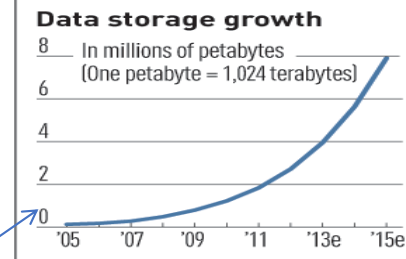
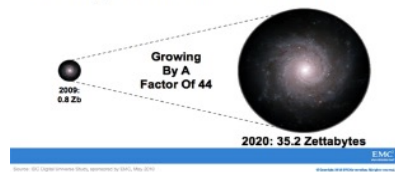
- **E-Promotions:** βάσει της συγκεκριμένης τοποθεσίας, του ιστορικού και των προτιμήσεων ενός χρήστη ➔ στείλε κουπόνια τώρα για το μαγαζί που είναι μπροστά του
 - **Healthcare monitoring:** σένσορες που ελέγχουν την υγεία (καρδιά, παλμοί, κλπ) ➔ οποιαδήποτε ανωμαλία χρειάζεται άμεση ενέργεια

Volume (Scale)

- **Όγκος Δεδομένων**
 - 44x αύξηση 2009-2020
 - Από 0.8 zb to 35zb
- Ο όγκος των δεδομένων αυξάνεται εκθετικά



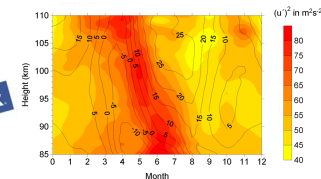
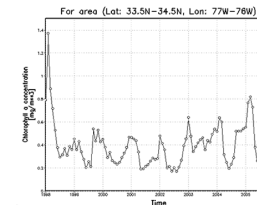
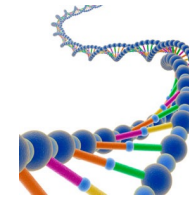
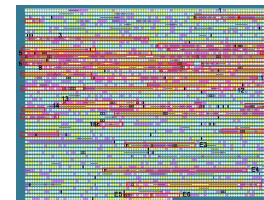
The Digital Universe 2009-2020



Εκθετική αύξηση στη συλλογή δεδομένων

Variety (Complexity)

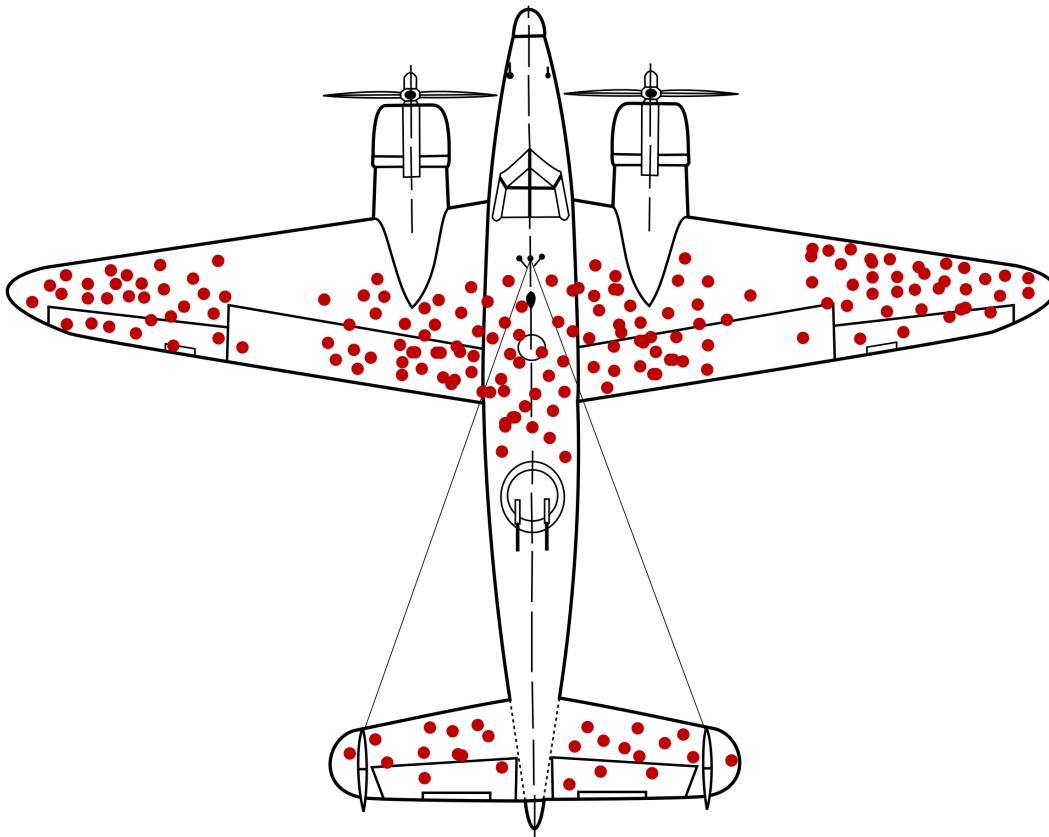
- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
 - Social Network, Semantic Web (RDF), ...
- Streaming Data
 - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)



Για εξαγωγή γνώσης → πρέπει να συνδυάσουμε όλους τους τύπους δεδομένων

Μεγάλα Δεδομένα == Καλά Δεδομένα ?

- Είναι τα Μεγάλα Δεδομένα που έχουμε αντιπροσωπευτικά;
 - Μαζεύουμε δεδομένα από το Twitter, αλλά ποιος γράφει tweets;
 - Μαζεύουμε ιατρικά δεδομένα, αλλά από ποιους είναι αυτά;
 - Έχουμε δεδομένα για αγορές online αλλά τι γίνεται off-line;
- Θέματα όπως bias, representativeness, weighting, etc., είναι πολύ σημαντικά



Example from wikipedia

Smart Data

- Κοιτάζουμε τα σωστά δεδομένα;
- Έχουμε όλα τα δεδομένα που χρειαζόμαστε;
- Μπορούμε να ενεργήσουμε με βάση τα δεδομένα μας;





(Big) Data Science

- Περιοχή που διαχειρίζεται, εξάγει και εξηγεί γνώση και συμπεράσματα από ένα τεράστιο όγκο δεδομένων
- Το (Big) Data science είναι ένα διεπιστημονικό πεδίο με στόχο να λύσει τις προκλήσεις στα μεγάλα δεδομένα
 - Computer Science
 - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
 - Mathematics
 - Mathematical Modeling
 - Statistics
 - Statistical and Stochastic modeling, Probability
- Οι ιδέες του Data science εφαρμόζονται σε όλα τα δεδομένα – μεγάλα και μικρά

Big Data στο Cloud & Cloud Computing

- Πόροι που δίνονται ως υπηρεσίες
 - Υπολογιστικοί, αποθήκευση, βάσεις δεδομένων, ουρές, ...
- Το Clouds χρησιμοποιεί οικονομίες κλίμακας σε απλό hardware
 - Φθηνή αποθήκευση, μεγάλο bandwidth networks & multicore επεξεργαστές
 - Data centers που είναι γεωγραφικά κατανεμημένα



Σχετικά Μαθήματα στο Τμήμα

- Προπτυχιακά
 - Σχεδίαση & Χρήση Βάσεων Δεδομένων
 - Υλοποίηση Συστημάτων Βάσεων Δεδομένων
 - Προγραμματισμός Συστήματος
 - Τεχνικές Εξόρυξης Δεδομένων
 - Επικοινωνία Ανθρώπου Μηχανής
- Μεταπτυχιακά
 - Διαχείριση Μεγάλων Δεδομένων
 - Συστήματα και Εφαρμογές στον Παγκόσμιο Ιστό
 - Τεχνικές Ανάλυσης Δεδομένων Υψηλής Κλίμακας
 - Οπτικοποίηση Δεδομένων

Ερωτήσεις