Managing Data & Databases

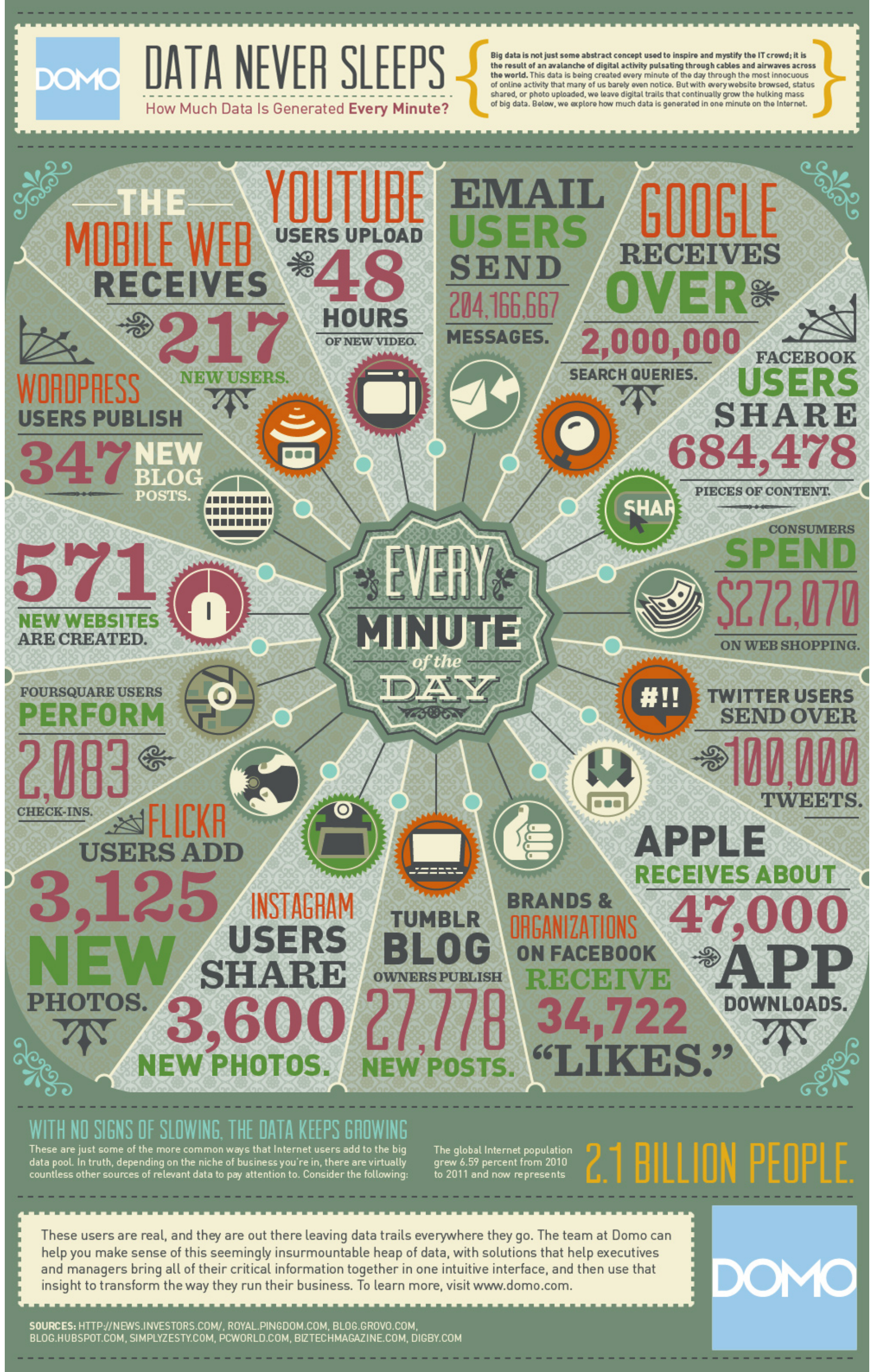
Session 9
The Data Deluge and Big Data

What is big data?

- The 3 Vs
  - Volume
  - Velocity
  - Variety

- Other characteristics
  - Versatility
  - High Granularity
  - Link-ability

The 3Vs illustrated



DOMO
DATA NEVER SLEEPS
How Much Data Is Generated Every Minute?
THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
CONSUMERS SPEND $272,070 ON WEB SHOPPING.
TWITTER USERS SEND OVER 100,000 TWEETS.
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.
FLICKR USERS ADD 3,125 NEW PHOTOS.
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
571 NEW WEBSITES ARE CREATED.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
SHAR
#!!
EVERY MINUTE of the DAY
WITH NO SIGNS OF SLOWING, THE DATA KEEPS GROWING
2.1 BILLION PEOPLE.
DOMO

Big data is not just some abstract concept used to inspire and mystify the IT crowd; it is the result of an avalanche of digital activity pulsating through cables and airwaves across the world. This data is being created every minute of the day through the most innocuous of online activity that many of us barely even notice. But with every website browsed, status shared, or photo uploaded, we leave digital trails that continually grow the hulking mass of big data. Below, we explore how much data is generated in one minute on the Internet.

These are just some of the more common ways that Internet users add to the big data pool. In truth, depending on the niche of business you're in, there are virtually countless other sources of relevant data to pay attention to. Consider the following:

The global Internet population grew 6.59 percent from 2010 to 2011 and now represents

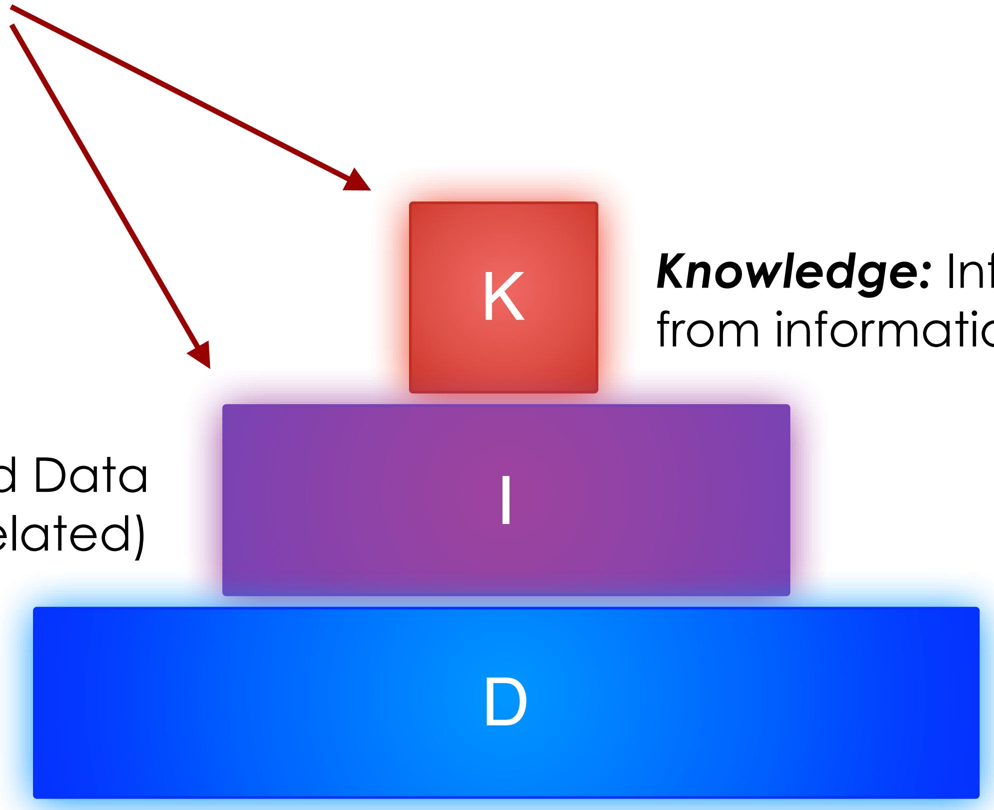
These users are real, and they are out there leaving data trails everywhere they go. The team at Domo can help you make sense of this seemingly insurmountable heap of data, with solutions that help executives and managers bring all of their critical information together in one intuitive interface, and then use that insight to transform the way they run their business. To learn more, visit www.domo.com.

SOURCES: HTTP://NEWS.INVESTORS.COM/, ROYAL.PINGDOM.COM, BLOG.GROVO.COM, BLOG.HUBSPOT.COM, SIMPLYZESTY.COM, PCWORLD.COM, BIZTECHMAGAZINE.COM, DIGBY.COM

A sneak peek:
http://tweetping.net

Source: domo.com

The bottleneck(s)

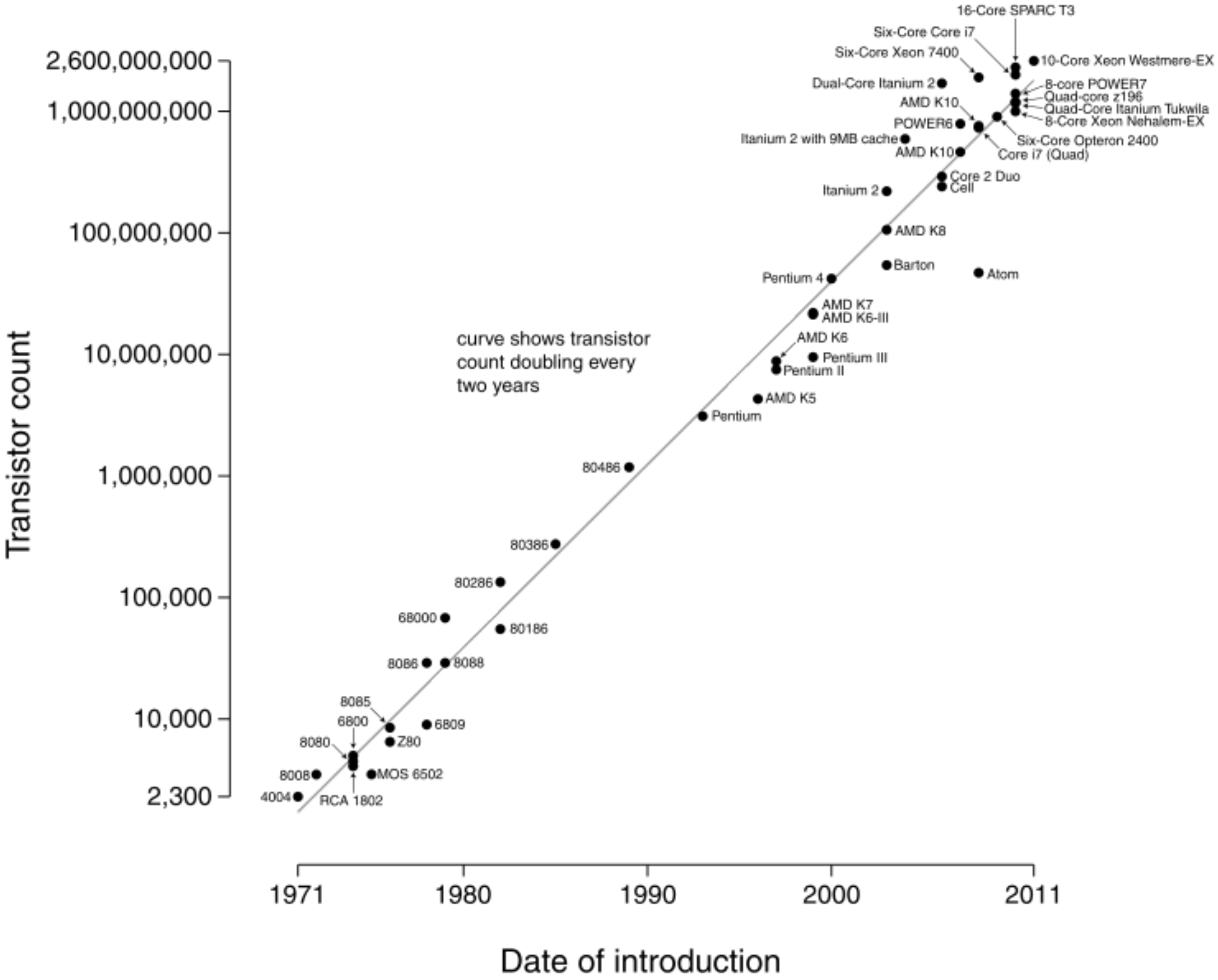


Knowledge: Inference derived from information

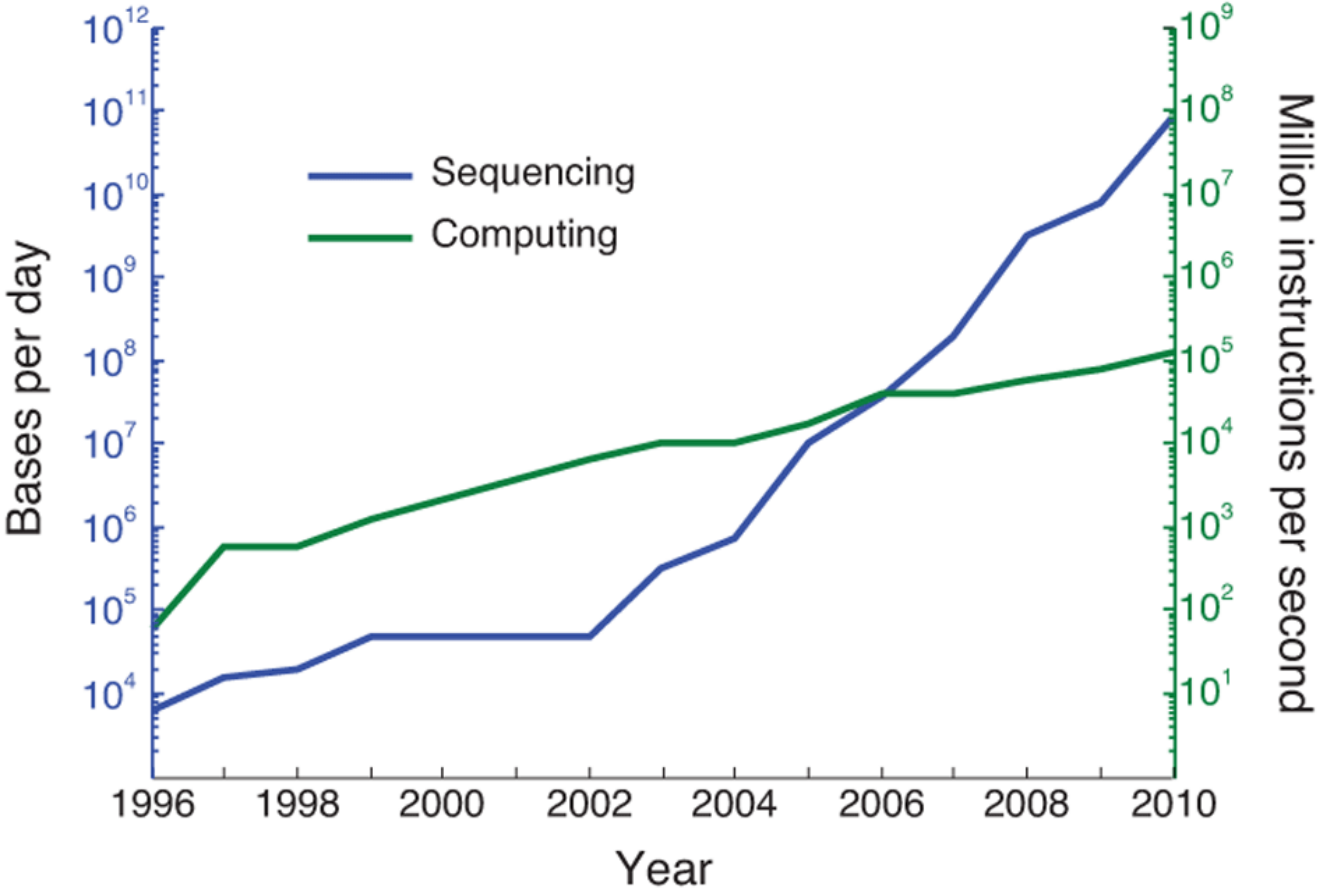
Information: Related Data (inter-related & context-related)

Data: Numbers and text, decontextualized

Microprocessor Transistor Counts 1971-2011 & Moore's Law



Source: Wikipedia


Sequencing
Computing
Bases per day
10^{12}
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
Million instructions per second
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
1996
1998
2000
2002
2004
2006
2008
2010
Year

An estimate
How much data do we generate?

2010: 1.2 zettabytes

2011: 1.8 zettabytes

How much is 1,800,000,000,000,000,000,000 bytes?

That fills 115 billion 16GB iPads!

Where does the data come from?

- Digitization (of existing or new data)
  - Content: Movies, Pictures, Documents, etc...
  - Sensing: DNA, Weather, GPS,... and Apple Health Kit!
- Transaction and Interaction Data
  - Banking, Commerce, Chat,
- User-Generated Content and Digital Footprint
  - Social Media, GPS, Comments, Website visit data,... and Apple Health Kit!

Are there any perils?

Big data: are we making a big mistake?
By Tim Harford

How to deal with it?

- Scalable and distributed data structures
- Parallel Computing
- Peer-to-peer Grids

How to make sense of it? (1)

Aggregators
Folksonomy
Wikis
User Centered
Joy of Use
Blogs
Participation
Six Degrees
Usability
Widgets
Pagerank
XFN
Recommendation
Social Software
FOAF
Browser
Simplicity
Videocasting
Podcasting
Sharing
Collaboration
Perpetual Beta
AJAX
Audio
IM
Video
Web 2.0
Design
Convergence
CSS
Pay Per Click
UMTS
Mobility
Atom
XHTML
SVG
Ruby on Rails
VC
Trust
Affiliation
OpenAPIs
RSS
Semantic
Web Standards
SEO
Economy
OpenID
Remixability
REST
Standardization
The Long Tail
DataDriven
Accessibility
XML
Microformats
Syndication
Modularity
SOAP

- Information Design
  - Visualization
  - Infographics
  - Word Clouds
  - etc.

Source:
Wikipedia

How to make sense of it? (2)

- Artificial Intelligence
  - Machine Learning (Pure or Supervised)
  - Pattern Matching
  - Data Mining (Classification, Clustering, etc...)
  - Text Mining (Natural Language Processing)

- http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html

How to make sense of it? (3)

- Analytics (Adapted Statistics)
  - Large data validity
    - Parametric, Non-parametric, Simulation-Based
    - Inferential vs. Predictive
    - Theoretical vs. A-theoretical
  - Parallel implementation
  - Mixed-level methods
  - Longitudinal methods

But should we keep all that?

And is there any way not to?