What's the Big Deal?

Big Data, Cloud & the Internet of Things

Christine Kirkpatrick
San Diego Supercomputer Center, UC San Diego

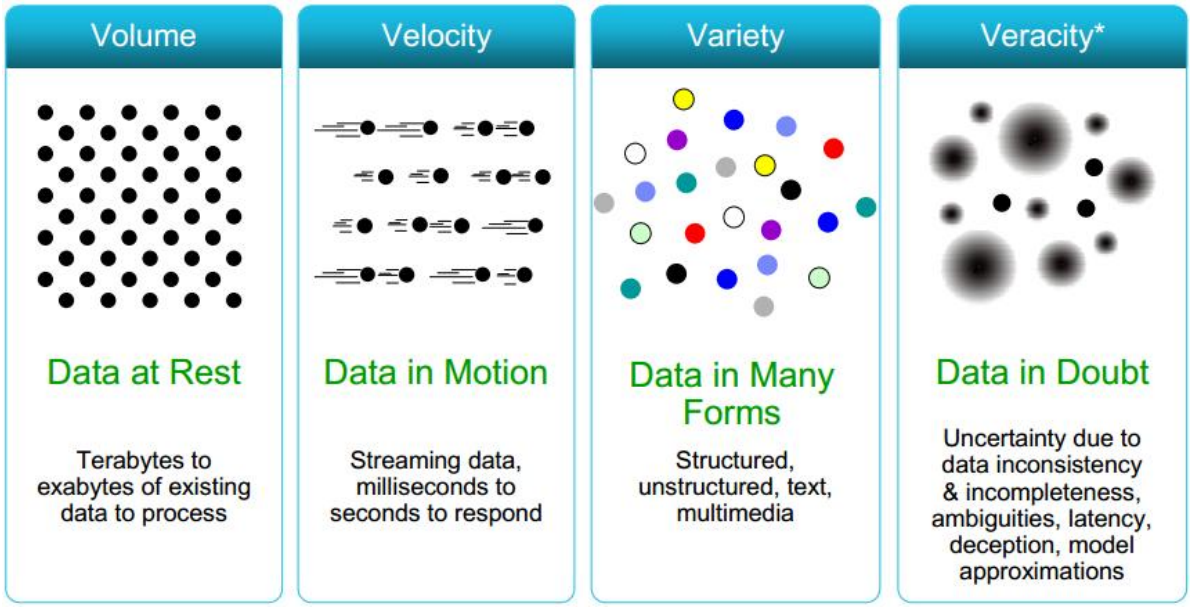
A Futurist's Near-Term View

The Future Depends on Data

- Self-driving car dependencies
  - Weather
  - Maps, Geography (2D/3D)
  - Peer sensors (other cars)
  - Crowdsourced data (e.g. Waze)
- Challenges
  - Amount of data
  - Different types of data
  - Putting the data together for use
  - Making sense of the data

What is Big Data?

- "Big data is **messy** data because it's *all* the data." cb
- The four V's of Big Data

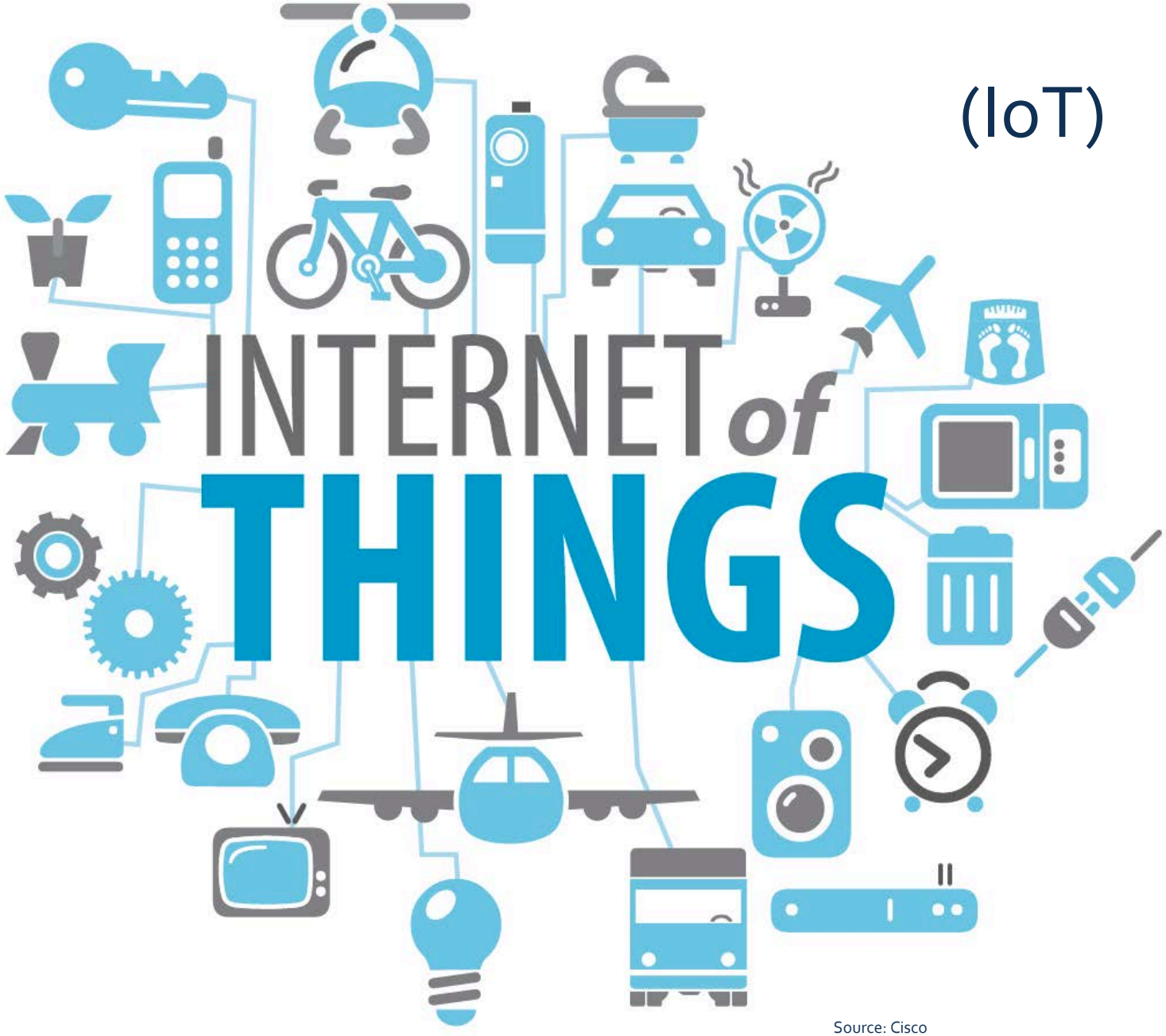


Credit: IBM 2012

Chaitan Baru, SDSC & NSF

- NIST – Big Data Working Group

(IoT)

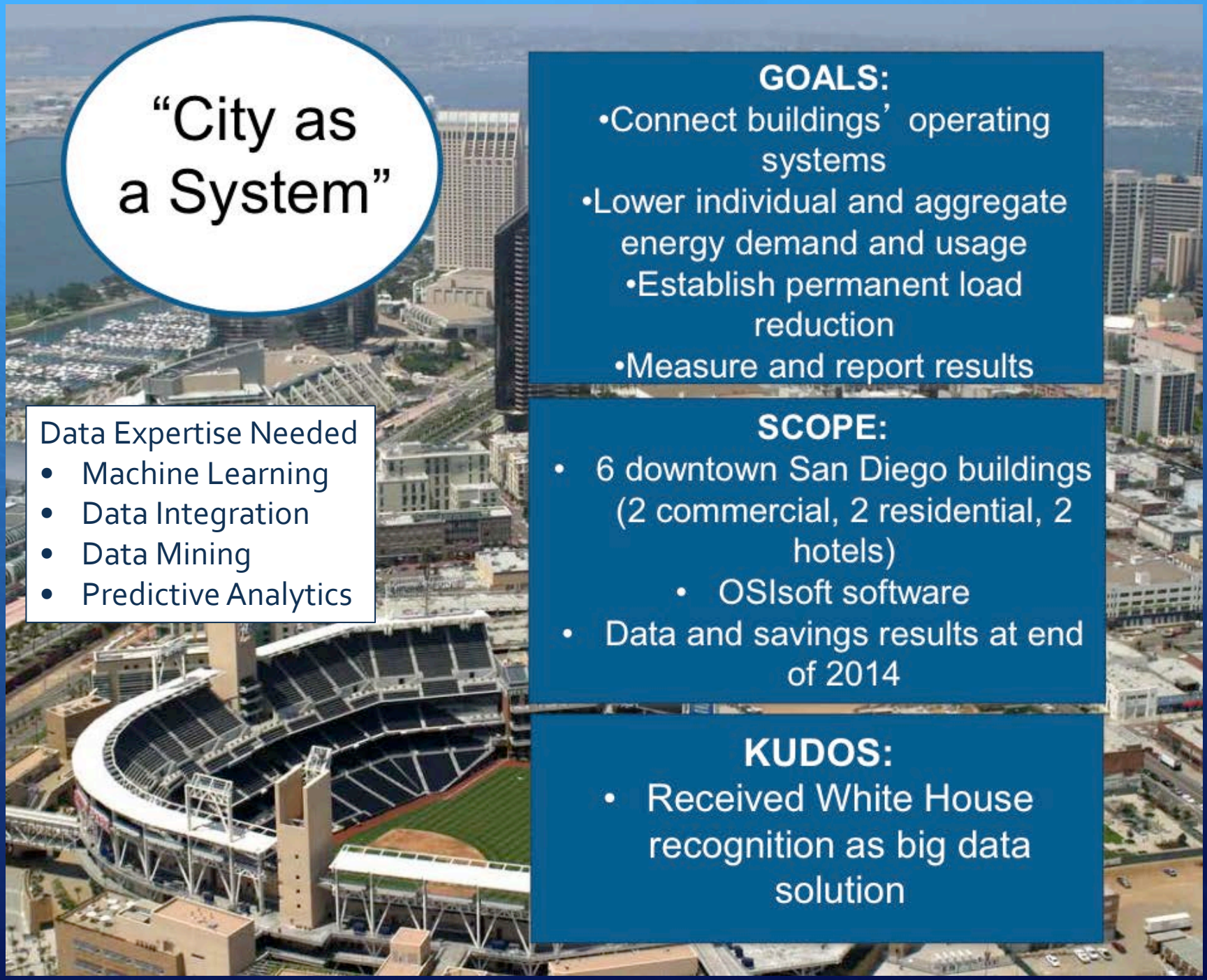


Source: Cisco

Smart City – IoT + Big Data

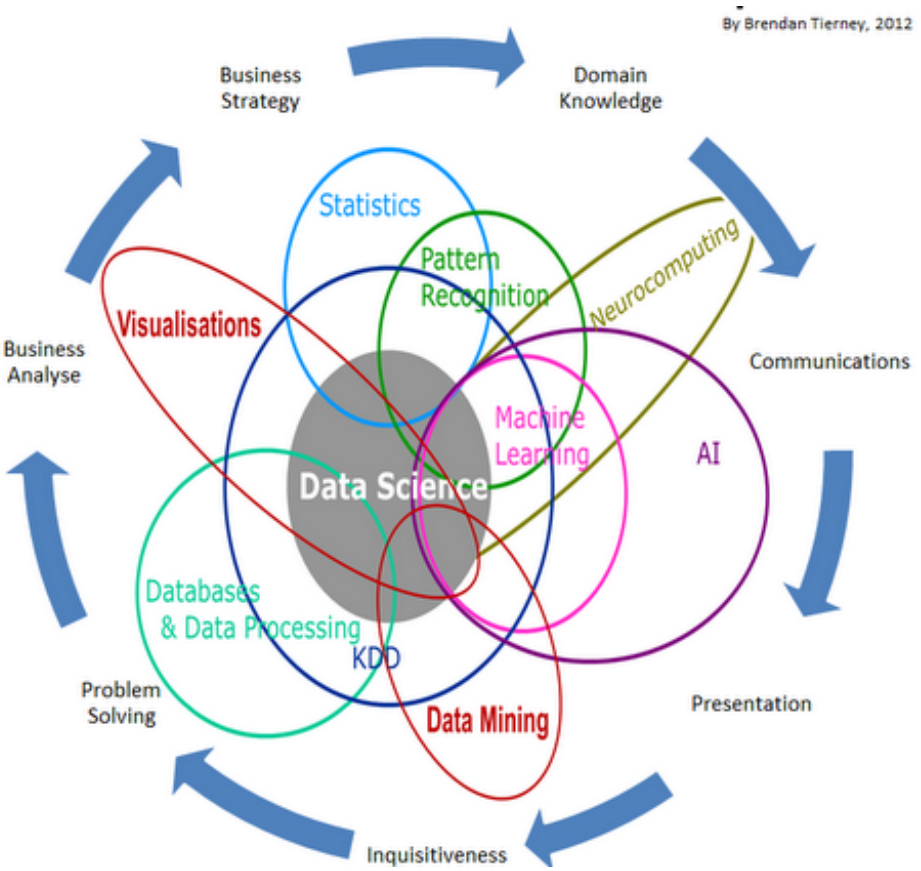
Natasha Balac, Director Predictive Analytics Center of Excellence (PACE), SDSC

PACE


"City as a System"
GOALS:
•Connect buildings' operating systems
•Lower individual and aggregate energy demand and usage
•Establish permanent load reduction
•Measure and report results
Data Expertise Needed
• Machine Learning
• Data Integration
• Data Mining
• Predictive Analytics
SCOPE:
• 6 downtown San Diego buildings (2 commercial, 2 residential, 2 hotels)
• OSIsoft software
• Data and savings results at end of 2014
KUDOS:
• Received White House recognition as big data solution

Big Data → Data Science at UC

- UC San Diego
  - SDSC Institute for Data Science
  - MAS Data Science & Engineering
  - Masters in Business Analytics
  - Extension: Data Mining Certificate
- UC Berkeley
  - Undergraduate
  - Online Masters Info & Data Science
- UC Davis - initiative underway
- UCSC – capital campaign, initiative
- UCI
  - Datascience.uci.edu
  - Extension - Predictive Analytics Certificate
- UCLA Extension courses
- NIH – BD2K



How do you do Big Data?

SDSC houses both big data resources and expertise

- Modeling & simulation
- Parallel computing
- Energy efficient computing
- Database systems
- Data mining
- Data modeling
- Data integration
- Data management
- Data processing workflows
- Datacenter management

Big Data : Data Science :: Supercomputing : Computational Science

Supercomputer = High Performance Computing (HPC)

Recent Big Data Projects

- Genomic data and social networks
  Friends tend to be genetically related -like ~4th cousins.
  Christakis & Fowler, UC San Diego on Gordon
- Studying high frequency trading
  Min Ye, U. Illinois on Gordon
- 3D Modeling of Animal Space Use
  San Diego Zoo, USGS and SDSC

A Scalable Data-Driven Monitoring, Dynamic Prediction and Resilience Cyberinfrastructure for Wildfires

Ilkay Altintas, SDSC

- Red Mountain Cams

South (left) "Highway" Fire

SW (center rear) is the "Pointsettia" Fire

West (right) is the "Tomahawk" Fire

WIFIRE is funded by NSF 1331615

WIFIRE

NSF

WIFIRE – Big Data Integration

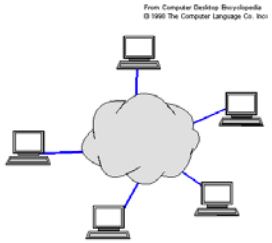
What is lacking in disaster management today is a *system* integration of

- real-time sensor networks
- satellite imagery
- near-real time data management tools
- wildfire simulation tools
- connectivity to emergency command centers

before, during and after a firestorm.

http://tinyurl.com/wifire_latimes

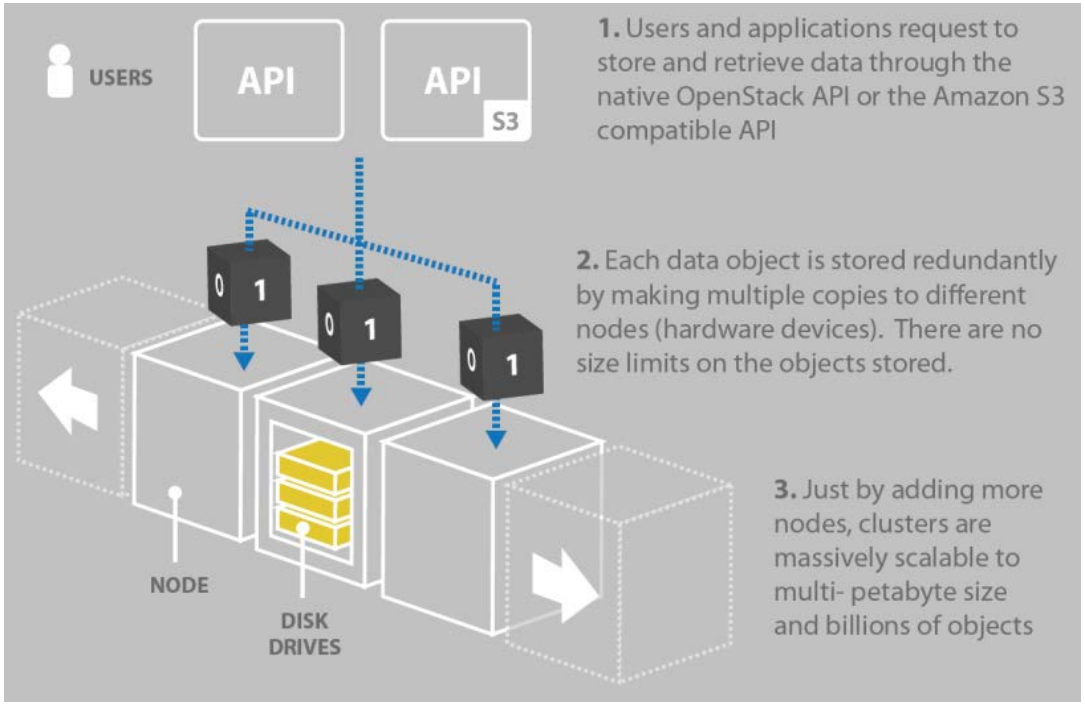
What's "cloud?"



- ***Cloud is a business model.***
- Cloud is new technologies, e.g. Object-based storage
- Cloud is location
  - Public – Dropbox, Amazon Web Services, Rackspace
  - Private – Your own cloud, e.g. SDSC Sherlock (Health) Cloud
  - Hybrid – Your own cloud replicating to Amazon
- Cloud is layers (LayerS ?)
  - IaaS – Infrastructure as a Service
  - PaaS – Platform
  - SaaS – Storage or Software
  - DaaS – Data as a Service
- Big data architectures increasingly cloud-based, e.g. Netflix on AWS
  - Scales elastically horizontally (keep adding more capacity)

Object-based Storage: openstack SWIFT

- 3 copies kept
- Object-Based vs. File-Based Storage
  - Directories and files vs. objects and containers
  - Applications immature
  - Third party devices



c:/docs/**kitten.jpg**

https://cloud.sdsc.edu/AUTH_8766e-3n76-7kkv-76a8-1hhf8765435/CONTAINER/**kitten.jpg**

SDSC Cloud 2.0, so what?

- Elastic resources crucial to support science
  - Storage, compute bursts
  - Short-term virtual machine (VM) needs
  - Untenable storage requirements
  - Advantageous economics:
    Pay for usage, not allocation,
    no equipment purchase
- Platform for scientific computing
  - 'Under the hood' access
    - Immense value to developers
    - Risk mitigation for cost overruns
    - Collaborate with SDSC Cloud Consulting Team
  - DaaS - Value-added services bundling
  - Scale out to commercial clouds, create hybrid cloud services

- Storage (3.5 PB)

- Cloud Compute

Data Science for Social Good (for San Diego)

- Modeled after program at U. Chicago, Prof. Rayid Ghani
- UC San Diego with UCSC
- San Diego + Tijuana, Mexico

→ Interested in getting involved?
DSSG@SDSC.edu

Questions?

Christine Kirkpatrick
christine@sdsc.edu