Potential Projects
Special Projects

• Evaluation of Wander and Ripple Join for an experimental paper
• IDE Benchmarking Study
Data Ingest

• Database Connector (ODBC with views)
• HDFS
• Random stream over CSV
• Transform operations (e.g., take [Name, Movie] transform to [Name, Forrest Gump, Martian,...]
Data Cleaning

- Operators (drop null, regular expression, substitution (mean, correlation, etc.),...)
- Scripting (have a standard way to add some custom code)
- Find automatically errors
ML

- Automatic hyper-parameter tuning (TuPAQ paper): bandit strategy/random search
- SVM with library
- Decision Trees + Random Forrest
- Matrix Completion? ALS?
- PCA? Making it usable might be hard
- Feature engineering: Scaling, Normalization, bucketing, ....
- Clustering (beyond k-means)
External ML

• Progressive inclusion of Python ML
• External ML: Progressive inclusion of R
• External ML: Progressive inclusion of Spark
Time series

- Stats (Arima)
- SVR
- CNNs/DNNs (Tensorflow)
- filter based on timestamps
Infrastructure

- Rare-Item Index (BIT-MAP Index)
- Distributed Storage
- Distributed Engine
- Resource management
Virtual Machine Image

1. Install VirtualBox
   https://www.virtualbox.org/wiki/Downloads

2. Download
   https://drive.google.com/open?id=0B3rYFyzcieF-ZFBrS0QyZVpiNHc
   https://drive.google.com/open?id=0B3rYFyzcieF-dE1meWJXdVo0akE

3. Double click on
   Windows 10 IDEA.vbox
Virtual Machine

- Password: idea
- Open Visual Studio 2015
- File->Open->Project Solution
  
  C:\Users\idea\Documents\Visual Studio 2015\Projects\IDEA\IDEA.sln

- Select
  IDEA-Rest on the build panel (right)
- Press “Play” Symbol to build and start IDEA-Rest

Note: You should checkout the source using git and not use the one in the Projects\IDEA folder. See next slides
IDEA Architecture

IDEA

Executor

Operator 1 (e.g., Histogram)
Operator 2 (e.g., SVM)
Operator 3
Operator 4
...
...

Storage

Full Sample (fully copy of the data)
Random Sample (Reservoir Sample)
Virtual Sample (reads original data)

Adaptor

CSV Adaptor
SQL Adaptor (one of the projects)

HTTP

Vizdom (aka Panoramic Data)
Github
Forking I

1. Sign up for github
2. Send Tim your github-user name
   
Forking II - Forking

   This creates a copy of the system in your github account.

2. Checkout from your fork  
   ```
   git clone https://github.com/YOUR-USERNAME/IDEA
   ```

3. Add upstream  
   ```
   git remote add upstream https://github.com/BrownBigData/IDEA
   ```

4. To sync with the master copy  
   ```
   git fetch upstream
   git checkout master
   git merge upstream/master
   ```
Committing

Always switch first to your branch:
git checkout YOUR_BRANCH_NAME

• Committing
git commit -a -m "COMMIT MESSAGE"

• Pushing the changes to the server
git push

• Getting latest fixes from the server
git pull

Do not modify the master branch!!!!!

From time to time we will update the master branch. You can apply the changes by

  git pull
  git merge master

(an alternative is git rebase -> read about it)
Alternative Procedure I

1. Sign up for github
2. Send Tim your github-user name
3. Clone the repository
   git clone https://github.com/BrownBigData/IDEA_CS2951V.git
4. Create your own branch
   git branch YOUR_BRANCH_NAME
5. Switch to your branch
   git checkout YOUR_BRANCH_NAME
Vizdom

If you want to change the interface, you need to do the same for the Vizdom repository as for IDEA

Git-url:
  https://github.com/ezg/PanoramicDataWin8.git