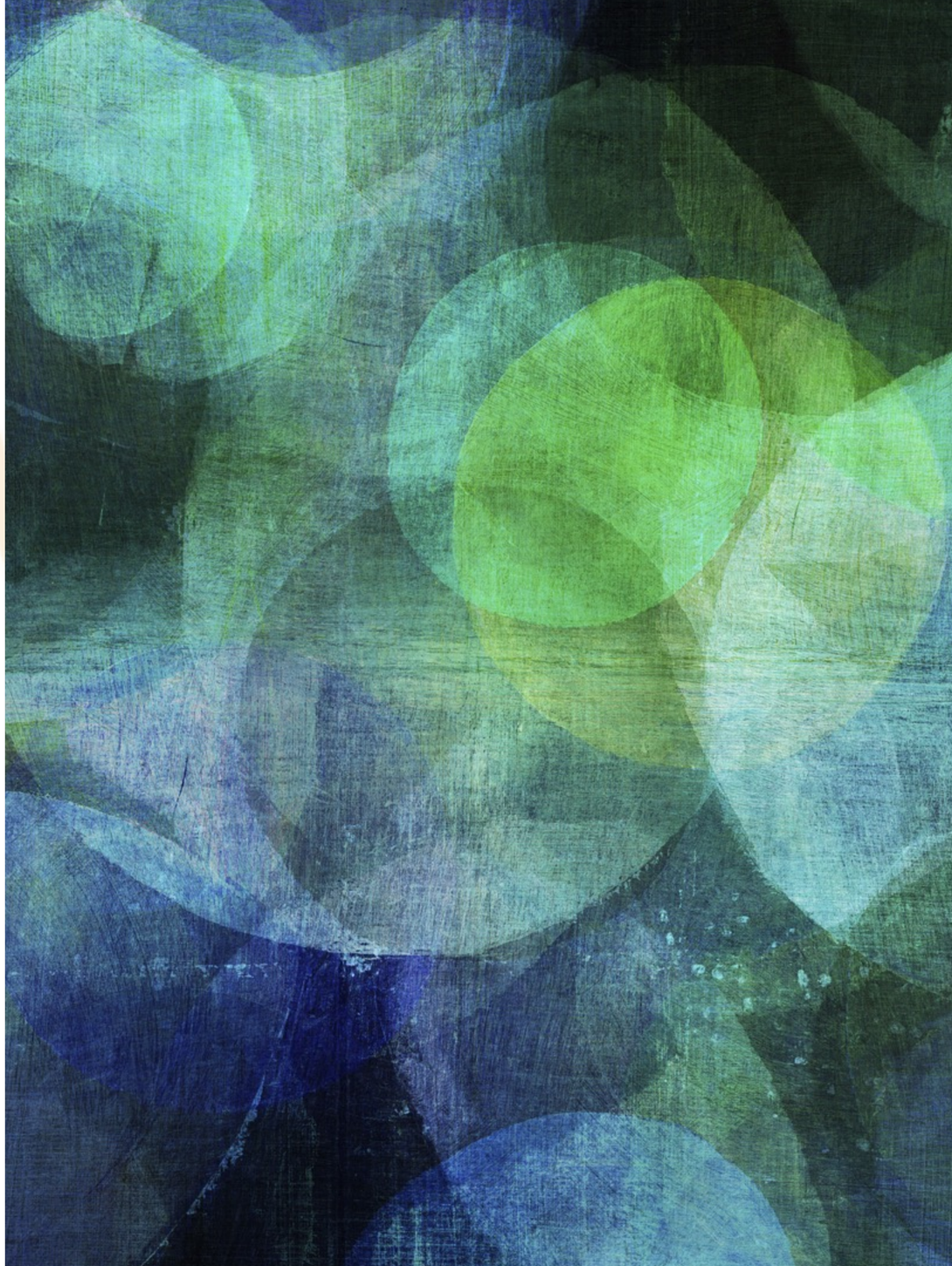
AWS DATA PROCESSING INFRASTRUCTURE 1B

Nan Dun
nan.dun@acm.org



COPYRIGHT POLICY 版权声明

All content included on the Site or third-party platforms as part of the class, such as text, graphics, logos, button icons, images, audio clips, video clips, live streams, digital downloads, data compilations, and software, is the property of BitTiger or its content suppliers and protected by copyright laws.

Any attempt to redistribute or resell BitTiger content will result in the appropriate legal action being taken.

We thank you in advance for respecting our copyrighted content.

For more info see https://www.bittiger.io/termsofuse and https://www.bittiger.io/termsofservice

所有太阁官方网站以及在第三方平台课程中所产生的课程内容，如文本，图形，徽标，按钮图标，图像，音频剪辑，视频剪辑，直播流，数字下载，数据编辑和软件均属于太阁所有并受版权法保护。

对于任何尝试散播或转售BitTiger的所属资料的行为，太阁将采取适当的法律行动。

我们非常感谢您尊重我们的版权内容。

有关详情，请参阅
https://www.bittiger.io/termsofuse
https://www.bittiger.io/termsofservice

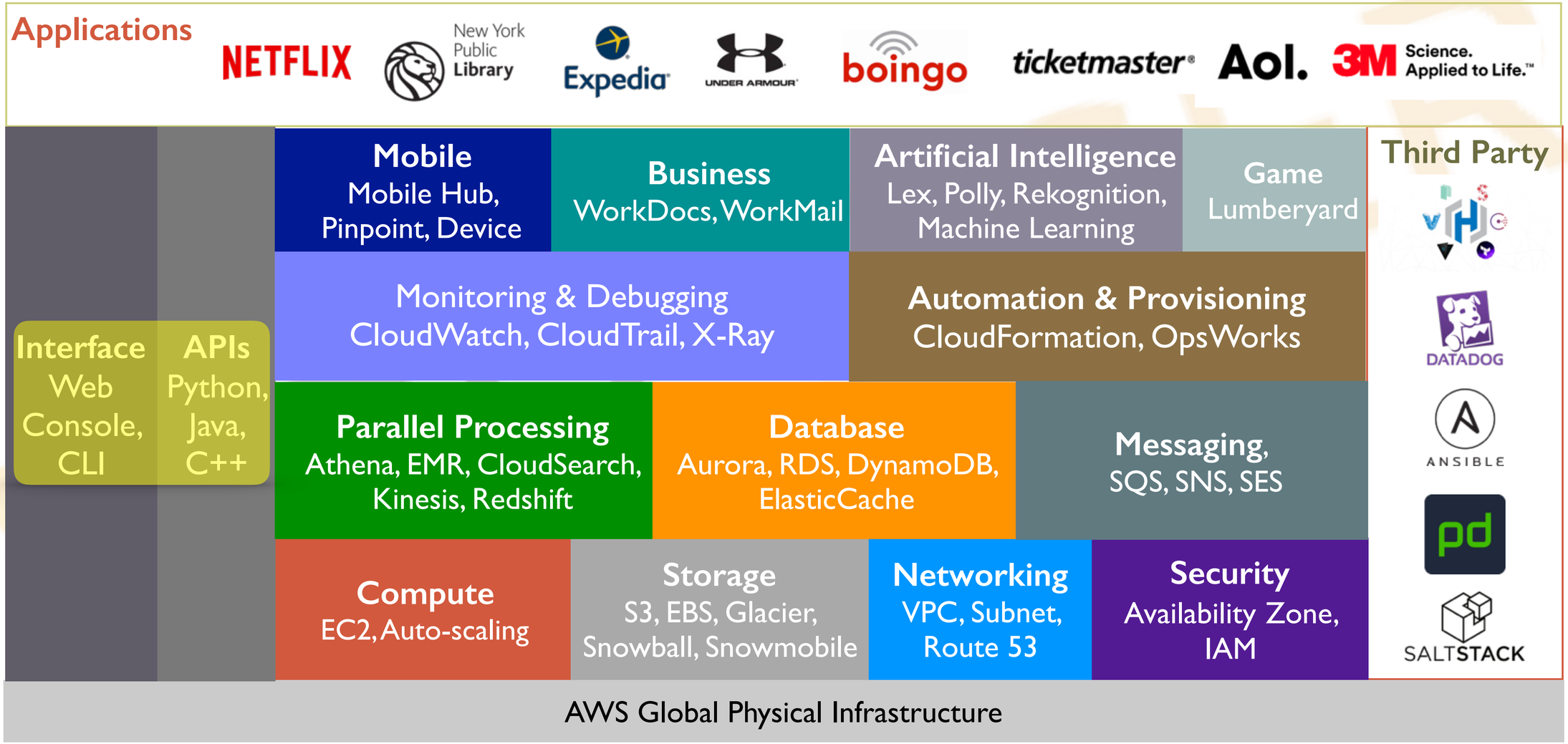
DISCLAIMER

I. "All data, information, and opinions expressed in this presentation is for informational purposes only. I do not guarantee the accuracy or reliability of the information provided herein. This is a personal presentation. The opinions expressed here represent my own and not those of my employer."

II. "The copyright of photos, icons, charts, trademarks presented here belong to their authors."

III. "I could be wrong."

TODAY'S TOPIC



ENVIRONMENT SETUP

- Python > 2.7.9
  - Use brew or macports to install Python
- Git > 2.10
  - If you are using Mac, it should come with Xcode
- Unix/Linux Command Line Tools
  - http://osxdaily.com/2014/02/12/install-command-line-tools-mac-os-x/
- Checkout the repository
  - Bitbucket account
  - Fork the repository
  - `$ make prepare`

AWS ACCOUNT SETUP

- Bookmark following URLs
  - http://www.ec2instances.info/
  - http://calculator.s3.amazonaws.com/index.html
  - http://s3speedtest.com
- AWS Account Setup
  - Credit card information
  - Set billing alarm in CloudWatch
  - Inbound SSH rule for default VPC security group
  - Create main user and save your key pair
- Customize your console

AWS CLI

AWS CLI CONFIGURATION

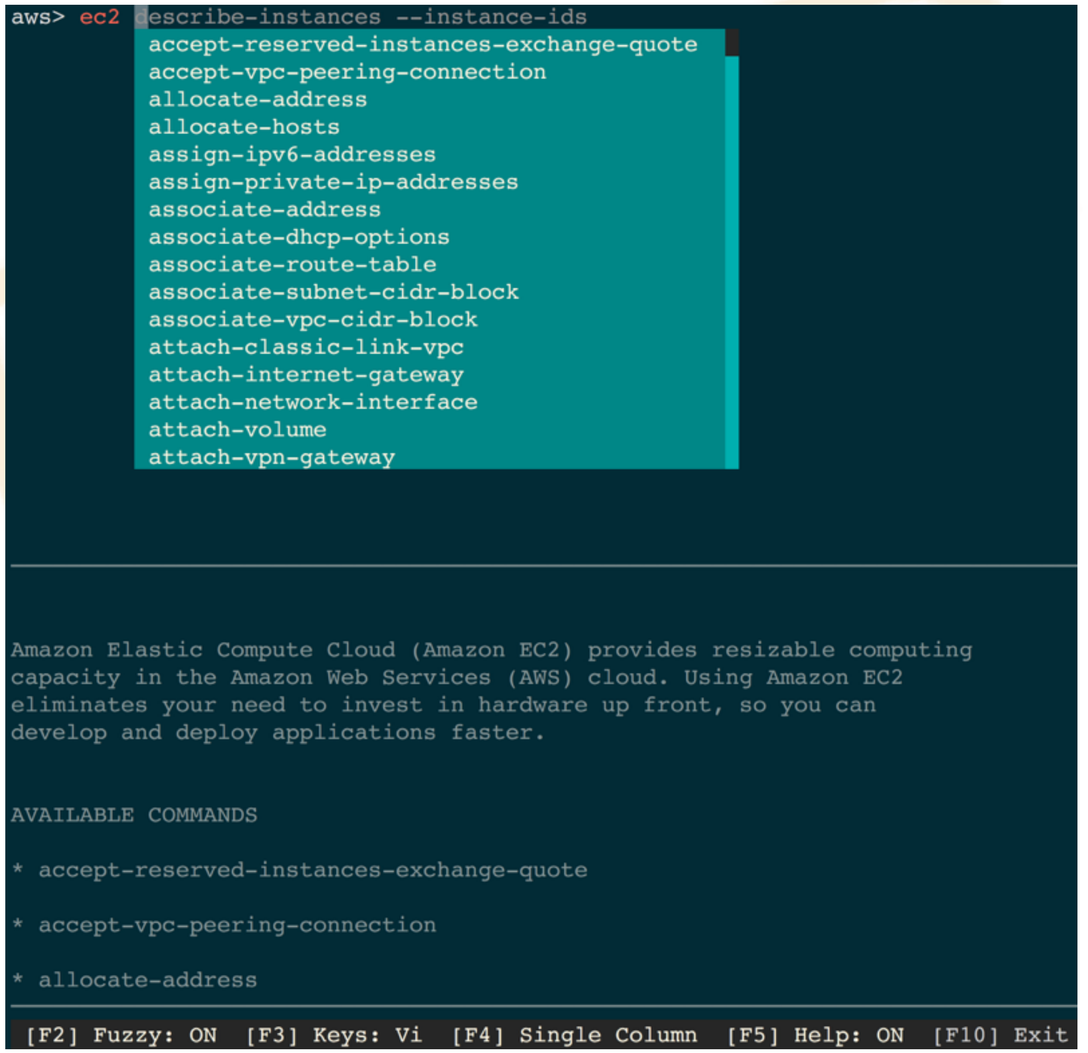
- Examples: scripts/configure.sh
- ~/.aws/credentials
- ~/.aws/config
- Profile
  - default
  - dev

AWS CLI EXAMPLES

- aws COMMAND SUBCOMMAND [OPTIONS]
- Github: https://github.com/aws/aws-cli
- Examples: scripts/awscli.sh

AWS SHELL

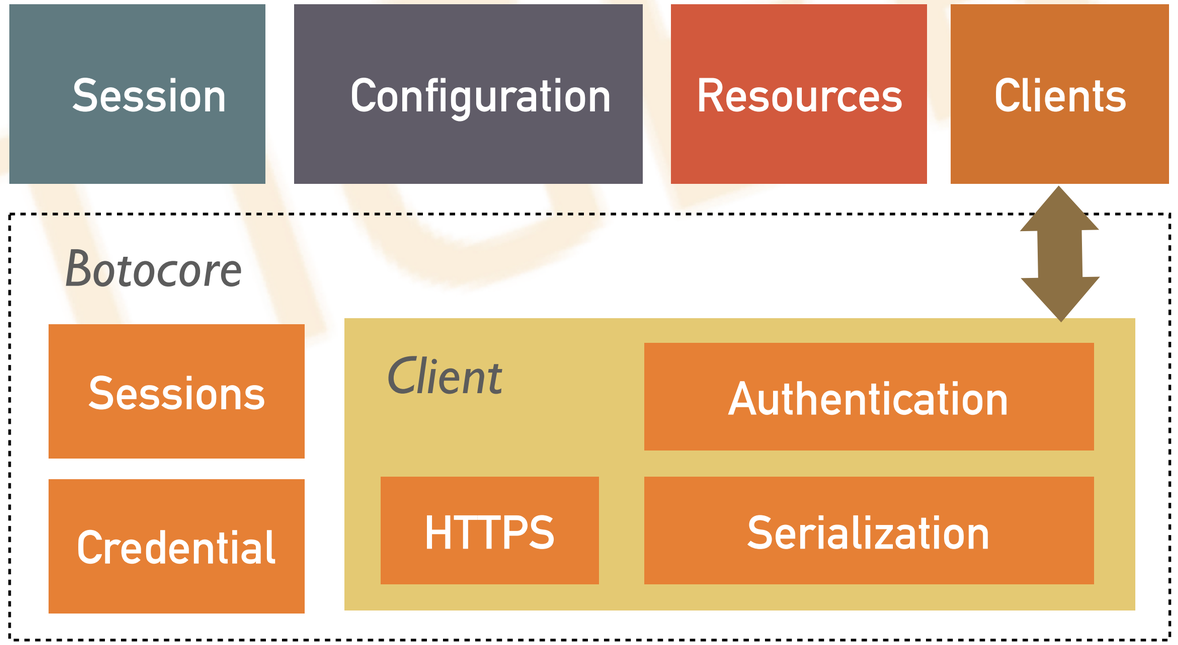
- Auto completion
- Pop document
- Server side auto completion
- Execute shell command



BOTO3

BOTO3

- AWS SDK for Python
- boto2
  - Early project
  - Community contributed code
- boto3
  - New version with consistent interfaces and up-to-date API support
- botocore
  - Low-level service and configurations
- https://github.com/boto/boto3
- https://boto3.readthedocs.io/en/latest/



EC2

A VPC REVIEW

- Create a VPC
- Create a public subnet
- Create a private subnet
- Create a internet gateway and attach to VPC
- Check ACL and security group
  - Add inbound SSH rule
- Make sure you can login to you instance in public subnet

EC2 - EXAMPLES

- Launch EC2 instance by CLI and Boto3
- Evaluation
  - CPU
  - EBS Bandwidth
    - fio
  - Inter-Instance Bandwidth
    - iPerf
  - Enhanced networking SR-IOV
    - Check if SR-IOV
    - Bandwidth comparison

“

Whenever you find yourself on the side of majority, it is time to pause and reflect.

-Mark Twain

S3

S3 - EXAMPLES

- Create/empty/delete bucket
- List/upload/download/copy from/to bucket
- Selective copy
- Examples
  - scripts/s3.sh
  - scripts/s3cp.py
  - taxi/raw2s3.py

PROJECT NYC TAXI DATA

- Source: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
  - Stored in S3
    - https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2016-01.csv
- Format
  - vendor_id, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, pickup_longitude, pickup_latitude, rate_code_id, store_and_fwd_flag, dropoff_longitude, dropoff_latitude, payment_type, fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount
  - Not aligned…

PROJECT - DATA PREPROCESSING

1. Clean: remove empty lines and invalid data
2. Select interested fields
   - pickup_datetime, dropoff_datetime
   - pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude
   - trip_distance, total_amount
3. Compact data (160 bytes/recored vs. 80 byes/record)
   - Change datetime to offset since 2009/1/1, e.g., 2009/1/1 12:01 => 60
   - Round coordinates to 6 digits
   - Round distance and fare to 2 digits
   - Padding to 80 bytes

DATA PIPELINE

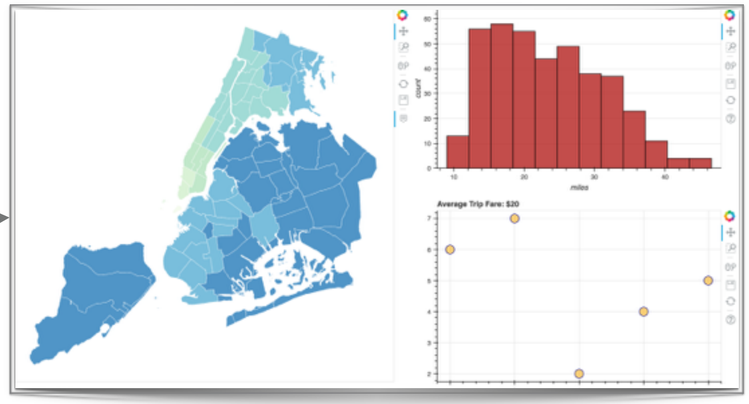
vendor_id, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, pickup_longitude, pickup_latitude, rate_code_id, store_and_fwd_flag, dropoff_longitude, dropoff_latitude, payment_type, fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount

pickup_datetime, dropoff_datetime, pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude, trip_distance, total_amount

Parallel Processing

pickup_datetime, dropoff_datetime, pickup_district, dropoff_district, trip_distance, total_amount

district, total_pickup, total_dropoff, total_trip_distance, total_amount

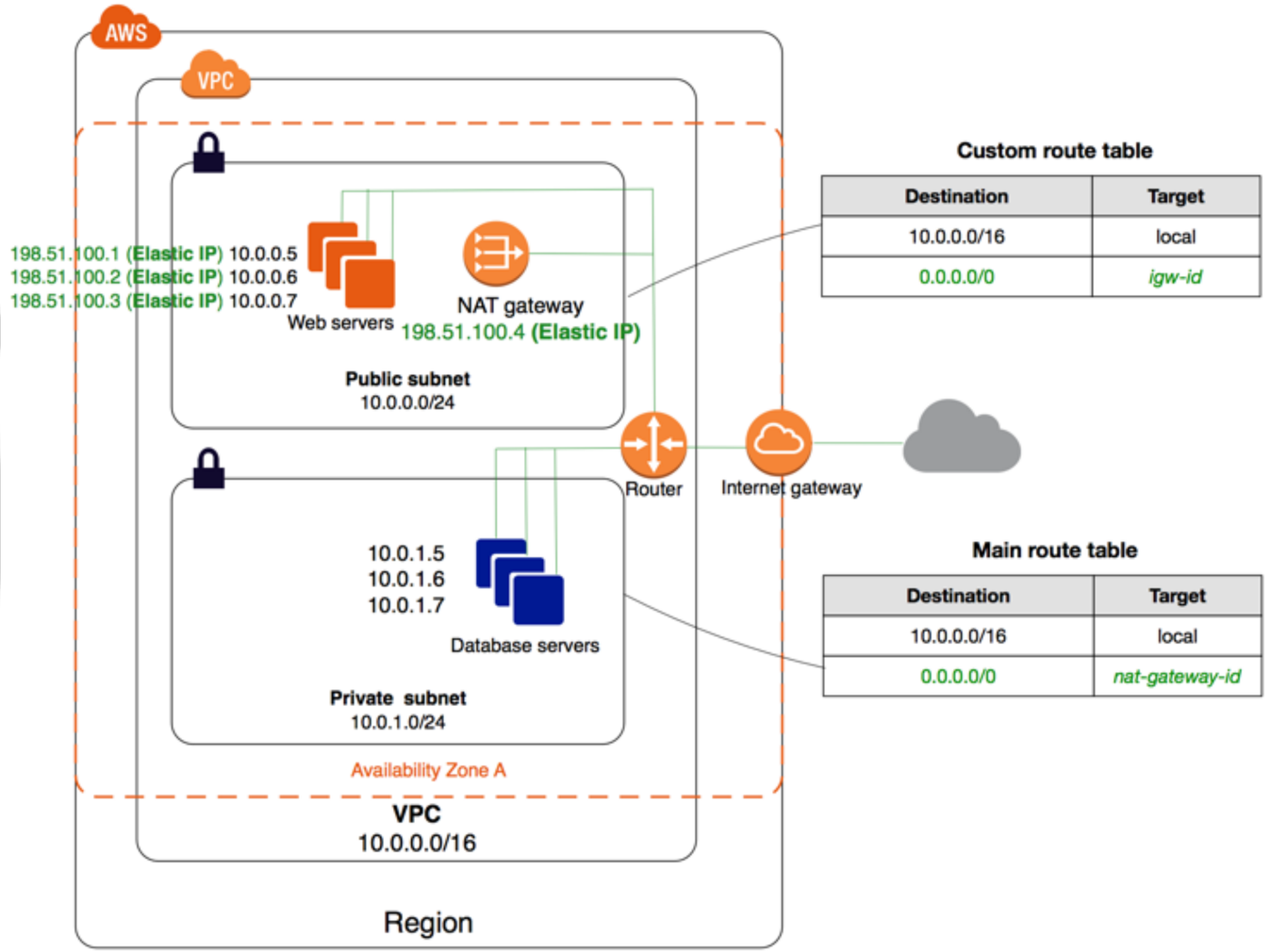


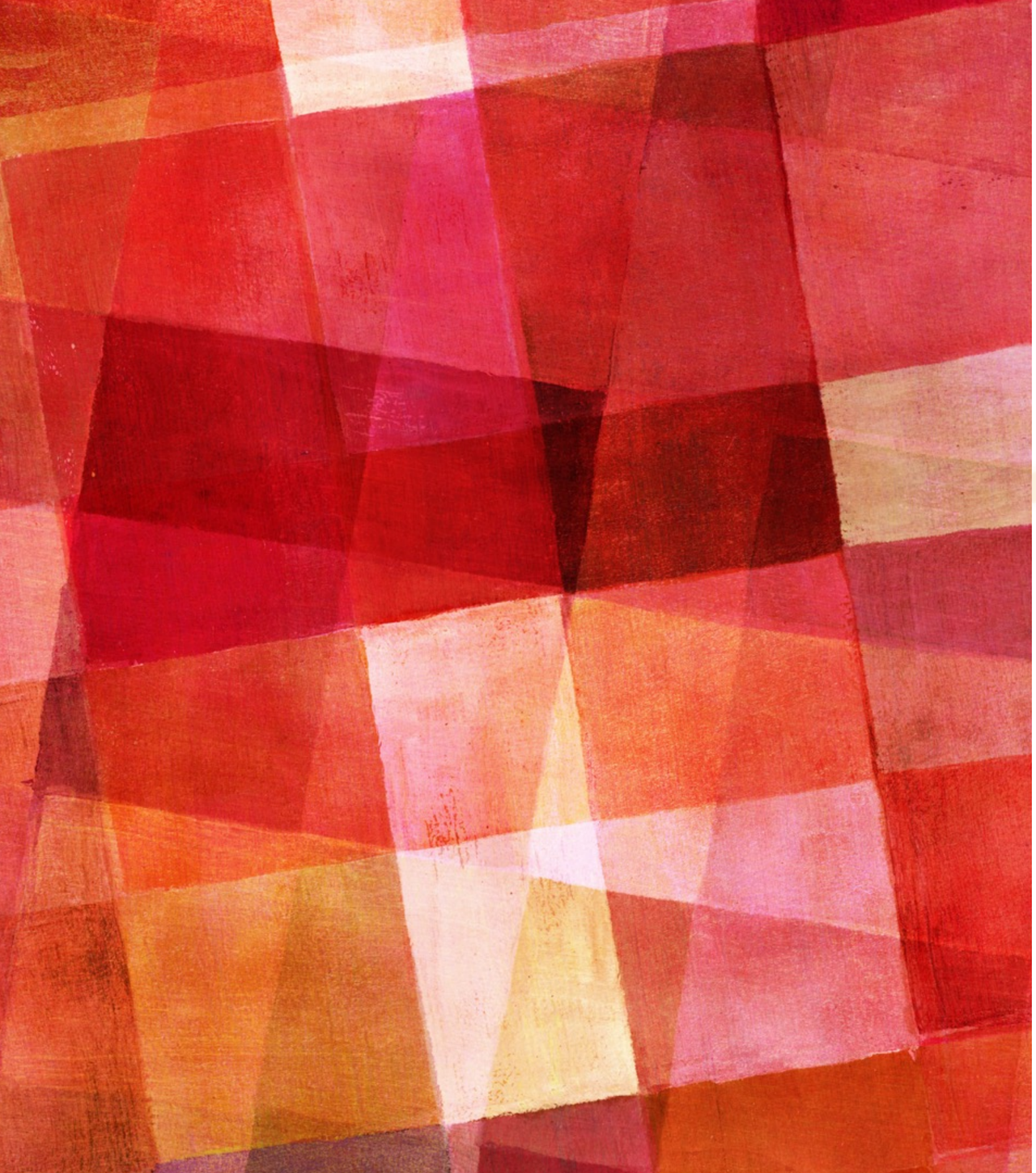
VPN

- Subnet
  - Create public subnet
- Security group
  - Dynamic attach
  - Multiple security group attach
  - Inbound/Outbound rules
  - Rules reasoning

HOMEWORK

- Create a VPC with both public and private subnet
- Write a boto3 program to launch two instances in each subnet
- Verify the one in public subnet can be accessed from internet and another cannot
- Measure the performance of these two instances
- Use taxi/raw2aws.py to upload part of data to s3://aws-nyc-taxi-data and explain how you make the transfer efficient (parallel)





QUESTIONS

- bittiger-aws@googlegroups.com

BITTIGER