Data Science for Impact and Decision Enablement (DSIDE) + Data Science Infrastructure Needs

Nyalleng Moorosi, Senior Data Scientist
CSIR Modelling and Digital Science
DSIDE: The Vac-Work Program

- Aim is to support capacity building in the ever growing field of Data Science. Primarily funded by Department of Science and Technology.
- DSIDE recruits vacation students (3rd/4th year) over the period of June/July (Phase 1 - lasting 4 weeks) and continuing December/January (Phase 2 - lasting 8 weeks).
- Students participate in mentor-guided and learn-by-doing problem solving of real-world needs as presented by different stakeholders.
DSIDE: The Program Leaders

• Shared between CSIR Modelling and Digital Science (MDS) and Meraka Units

• Program Leaders
  – Ishmael Makitla
  – Vukosi Marivate
  – Nyalleng Moorosi
  – Dhiren Seetheram
  – Quentin Williams
DSIDE: The Numbers

• Students:
  — 2014/2015 - 10
  — 2015/2016 - 33
  — 2016/2017 - 51
DSIDE: Fields of Study

- Multidisciplinary in Nature, Student Fields:
  - Actuarial Science
  - Computer Science
  - Engineering
  - Graphic Design
  - Journalism
  - Mathematics
  - Public Policy
  - Statistics
DSIDE: Challenge/Project Themes

- Environmental Modelling
- GeoSpatial Modelling
- Information Systems for Public Institutions
- Knowledge Graph Creation and Mining
- Natural Language Processing Applications
- Predictive Policing
- Process Optimisation
- Resource Allocation
DSIDE: The Process

- Project Solicitation and Scoping
- Preliminary Preparatory work in Python and Django
- Workshops
  - Exploratory Analysis
  - Software Development principles
  - Introduction to Machine Learning
Experiences

- Empowering Experience for Students
  - Team Work
  - Real Challenges
  - Fast Pace Learning
- A showcase for partners
- Use of Open Data
- Researchers take on riskier projects
Student Feedback - Survey Results

Technical Skills
It was my first time programming in python, jupyter, git in fact I was exposed to a whole new world and it was an amazing experience.

Communication:
1) Improved my public speaking skills
2) Improved my Interpersonal skills
3) Improved teamwork skills

Programming Tools:
“...libraries such as pandas, numpy, matplotlib and etc.”

Would you be interested in applying for a 2nd year on DSIDE? (23 responses)

- Yes: 17.4%
- No: 82.6%
Infrastructure and Data Tools

Python Stack
- Anaconda
- Scikit-learn
- Jupyter
- Django
- Pandas

Search Stack
- Elastic
- Kibana
- Apache Solr

Collaboration
- Slack
- GitLab

yen = β'x + ϵ
Data Science with Large Scale Computing + Storage

- Open Source Computing stack is mature.
  - Apache Ambari based deployment
- Infrastructure that can be shared.
- Students and Researchers need the exposure to the paradigms

**Computing**
- Apache Ambari
- Spark
- Storm
- TensorFlow
- Hadoop

**Storage**
- Cassandra
- Airflow
- CSIR

---

11
Visit our website for information on how you can get involved.

- **Students**
  - Apply for the program

- **Universities or Academics**
  - Get in touch, let us know about your Data Science programs and how we could work together (nmoorosi@csir.co.za)

- **Companies/Organisations**
  - Submit a Challenge
  - Attend Exhibition days
  - Share your job opportunities for Data Science with us.

http://dsideweb.github.io

CSIR Vacancies