Stanford Data Science Initiative

2017 Annual Retreat

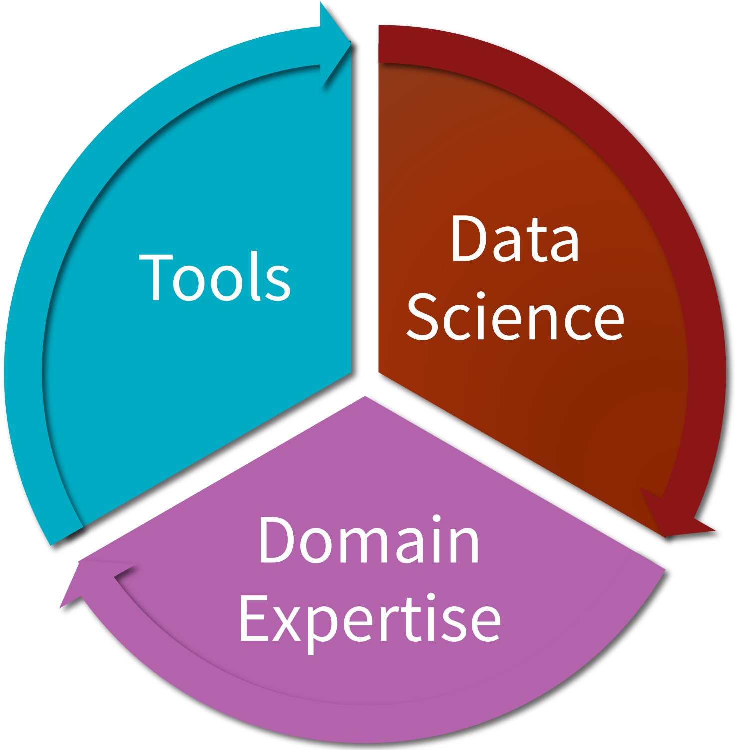
SDSI Vision

Support research at the intersection of:

- Real-world industry and societal needs
- Faculty and student interests
- Cutting edge data science and Machine Learning/AI technology

Dedicated focus on research through deep engagement between companies and faculty/students.

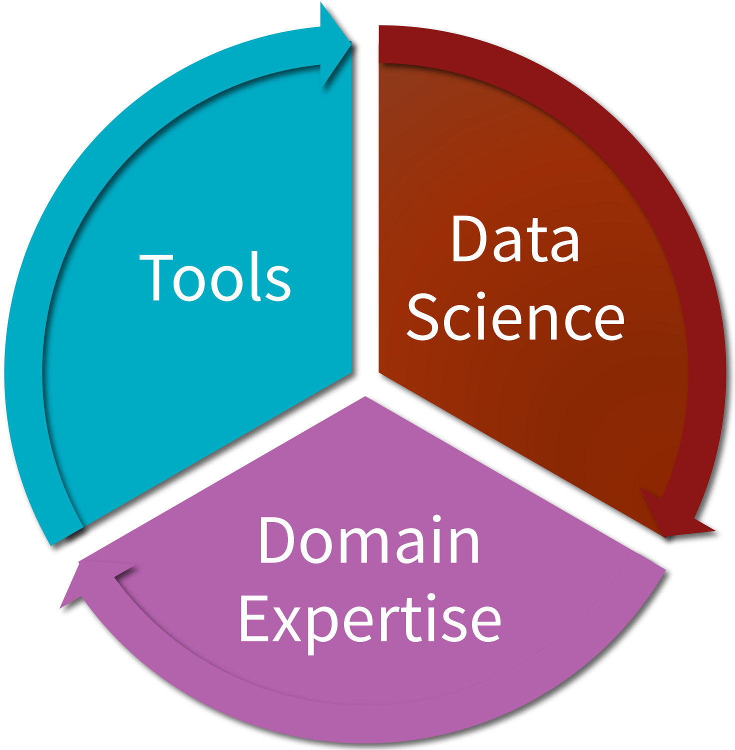
Collaborative Environment



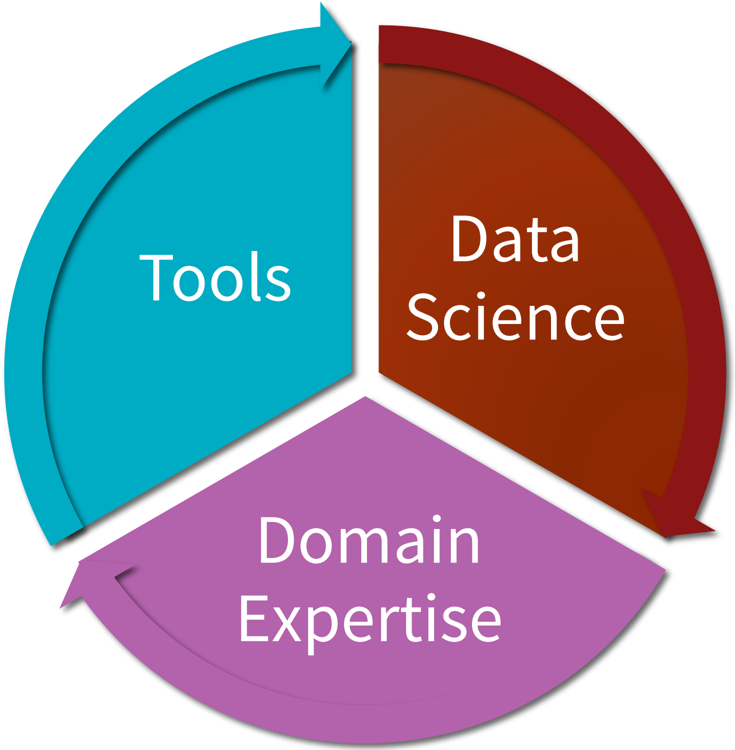
Machine Learning
Natural Language
Computer Vision
Networks and Graphs

Collaborative Environment

DeepDive
SNAP
Snorkel
DAWN



Collaborative Environment



Human, Society, Medicine, Security, IoT

SDSI Researchers

- 42 Stanford Faculty and Senior Researchers
- More than 20 Departments
- Dozens of Postdocs and Graduate Students

SDSI Current Corporate Members

Founding Members

Regular Members

SDSI Activities

2014-17

More than <u>$5 million of funded research</u> in areas at the intersection of our corporate members and faculty

- Equivalent to 65 PhD student-years
- Expenses only 12% of revenue

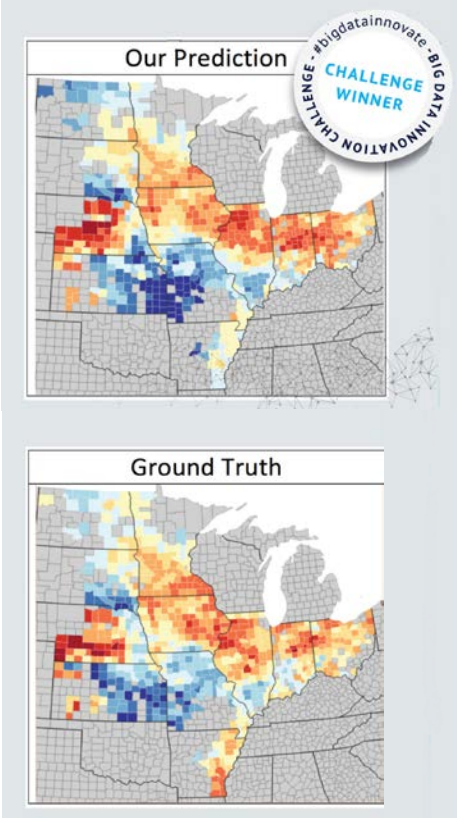
More than 60 diverse research projects developed in the areas of:

- Precision Medicine
- Social Science
- Agriculture
- Security
- Machine Learning
- Internet of Things
- Natural Language Processing
- Computer Vision
- Data Analytics
- And more…

Big Data for Agricultural Risk Management

David Lobell | Jiaxuan You, Xiaocheng Li, Melvin Low, Stefano Ermon
Earth Science | *Computer Science*

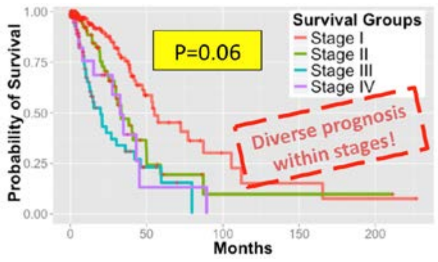
- Aims to improve in-season predictions of yields for major crops in the US
- Makes predictions based on weather and satellite data using:
  - graphical models
  - approximate Bayesian computation
  - crop simulation models
- 2017 World Bank Big Data Innovation Challenge winner



A Quantitative Path to Pathology

Kun-Hsing Yu Ce Zhang Gerald Berry, Russ Altman, Christopher Ré, Daniel Rubin Michael Snyder

- Researchers from *Biomedical Informatics, Genetics, Computer Science, Pathology*
- Real world problem addressed by deep collaboration between domain and methods scientists
- Media punchline: *“Images + patient data outperform expert pathologists at prognosis”*
- Provided source code and data for the analyses under an open license
- Winner of the 2017 Junior Parasite Award



Corporate-Academic intellectual intersections

Examples include:

Supporting high value research where mutual benefit is shared by both academic and industry members

Stanford Distributed Clinical Data project and MS Azure

Philip Lavori
Biomedical Data Science

Balasubramanian Narasimhan
Statistics

Daniel Rubin
Biomedical Data Science, Radiology, Medicine

Enables using confidential healthcare data among distant hospitals and clinics for creating decision support applications without requiring sharing any patient data among those institutions.

Research in financial transaction network analysis

Jure Leskovec
Comp. Sci.

Peter Bailis
Comp. Sci.

Blaz Fortuna
Comp. Sci.

Enables using financial transaction data for real-time processing and detection fraud and risk.

And more from today's agenda…

- Safe distributed computation for reluctant data sharers: At home in the Microsoft Azure cloud
- Democratizing machine learning in the Stanford DAWN project
- After DAWN: The potential of usable machine learning today and tomorrow
- Finding earthquakes in a spaghetti of seismic streams
- Socially impactful and ethical data science
- Analytics on time-series data
- Security in the era of big data
- 14 Posters covering topics in multiple fields

Future Directions

Data Science at Stanford

- **Data Science is an integral part of the university:**
  - Students: Our classes are exploding:
    - Machine Learning has +1000 students
    - Groups interested in Data Science for Social Good
  - Research: Initiatives and collaborations
    - School of Medicine
    - Social Sciences
    - Arts & Humanities

Challenges for Data Science

- **Challenge:** How to create data science fabric across the university
  - Democratizing access to data science methods and knowledge
- **Challenge:** How to enable more and faster research
  - Across sciences --- interdisciplinary collaborations
  - With industry --- partnerships enabling fundamental scientific advances
- **Challenge:** How to make data science human centric
  - How to include human aspect into the design process
    - Issues of privacy
    - Fairness, bias
    - Explainability
  - Collaborations with social scientists, etc.

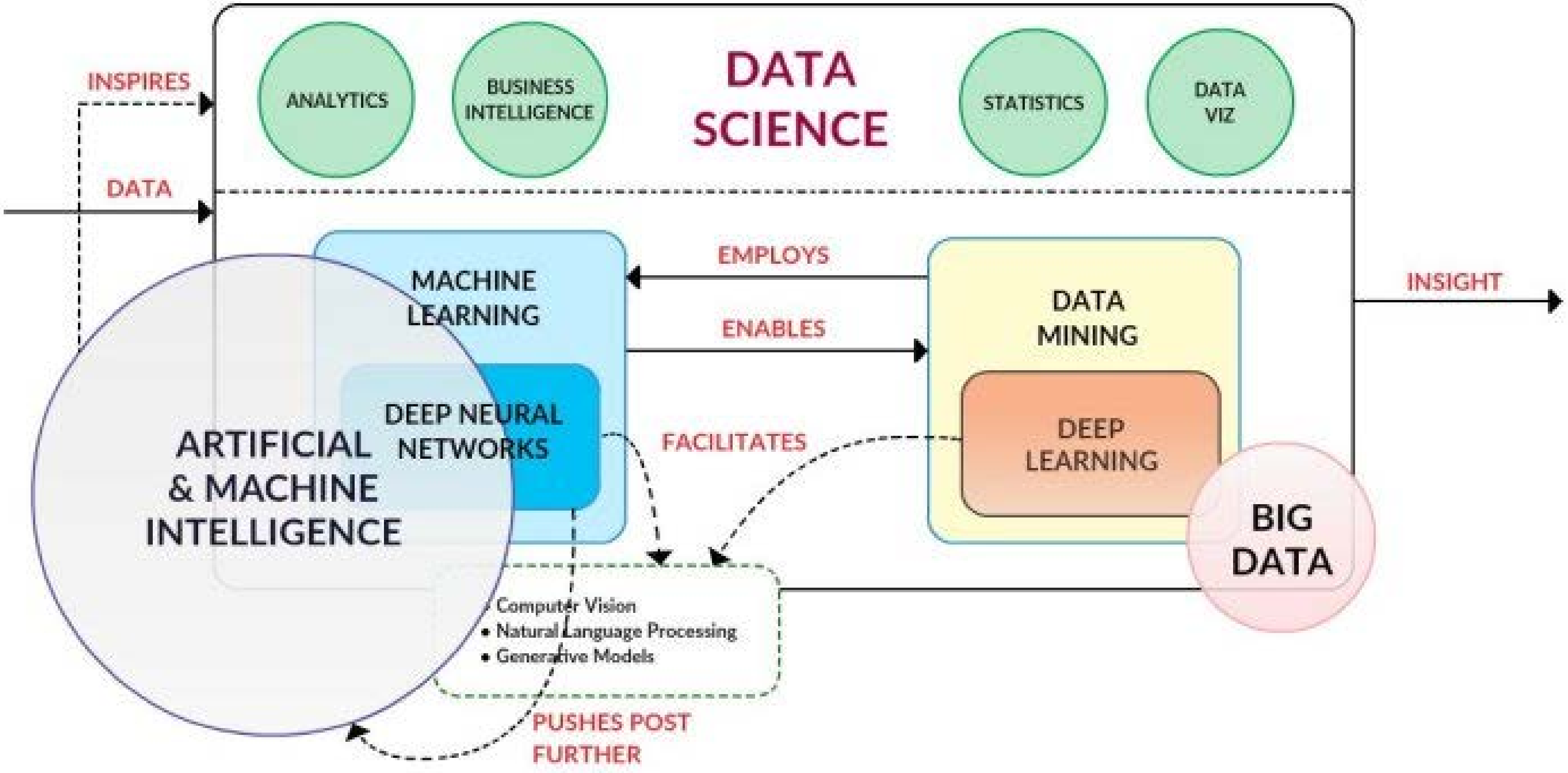
Future Directions

- Continued focus on research in innovative and impactful data science where both industry members and faculty find mutual interest
- Tighter and more dynamic connections with industry through:
  - Common interest research themes
  - Annual RFP solicitations
  - C-Level events
- Further support for data intensive research that leads to positive social impacts

Bigger Picture

- Continue building bridges and inform research at Stanford
- Events and seminar series for more interaction of Stanford/industry
- Targeted more-focused RFPs and projects
  - Consortia of companies interested in particular topics
- Create a data science fabric across the University
  - Connect applications with data science experts
    - Machine learning, analytics, data processing systems, visualization & HCI

BACK UP SLIDES



https://www.kdnuggets.com/2017/01/data-science-puzzle-revisited.html