Evolution of Software Infrastructure to Support Science and the Communications Arts

M. Jason Vandecreek

WSU – Special Projects
Research Associate

HireMJ@Vandecreek.com

Vandecreek Consulting

custom software solutions, architecture, training, design, data science

10+ years as both a contractor and hourly employee of WSU

CEI Community Engagement Institute

Primarily with the CARE team (Center for Applied Research and Evaluation)

CETARL The Center for Educational Technologies to Assist Refugee Learners

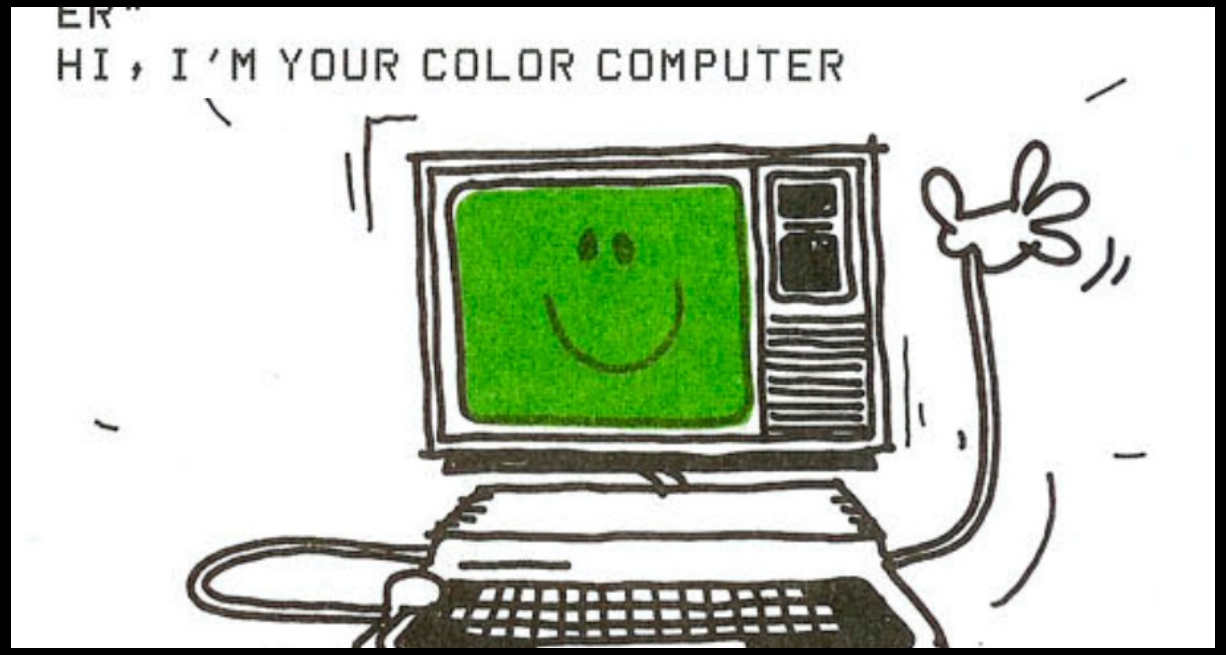
one of four winning research projects of the Convergence Sciences Initiative grant funding

Trees for Life, International – Volunteer / Board Member

Poverty reduction, education, nutrition, eye-care

HireMJ@Vandecreek.com

Data for breakfast, lunch, and dinner!
CAVS
23


ER"
HI, I'M YOUR COLOR COMPUTER

The Arts Fuel Our Souls

The Data Analysis Ecosystem

Fueled by the Open Source Software Movement

Origins: Where did these tools come from?

Who makes this stuff and how is it funded?

And why is that important?

Software As A Key Piece Of Infrastructure In An Ecosystem

One example: SQLite

- An embedded database everyone in this room has used.
- Most Widely Deployed Database and Public Domain
- Used in the flight software for the Airbus A350 XWB aircraft.
- Created by Dr. Richard Hipp in 2000.
- Android, iPhone, Windows, Macintosh, TVs, Firefox, Chrome, and Safari, and billions more.

HDF – Example

Hierarchical Data Format version 5 (HDF5)

HDF5 is a data model, library, and file format for storing and managing data.

When: 1988
Who: National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign
Why: Scientific Data Sets, Meta, Graphics, Subsets, etc.

Also created MOSAIC, the first graphical web browser.
HDF files or HTML files

Funding: NSF National Science Foundation

"Make software useable outside of the lab in which it was created"

Software Capitalization Project

Benefits / Outcomes

Massachusetts General Hospital (Hospital in Boston)

1990s chose HDF to store both X-ray images and patient data in a single file

Private companies started to donate money to the HDF project (Phillips 66)

Big Data – NASA: Earth Sciences – Earth Observing System

Produces 3 Terabytes of data per day!

Boeing flight testing: Database can’t scale

Hybrid solutions – Put metadata in a database and then reference millions of HDF files in storage

rOpenSci ropensci.org

fosters a culture that values open and reproducible research using shared data and reusable software

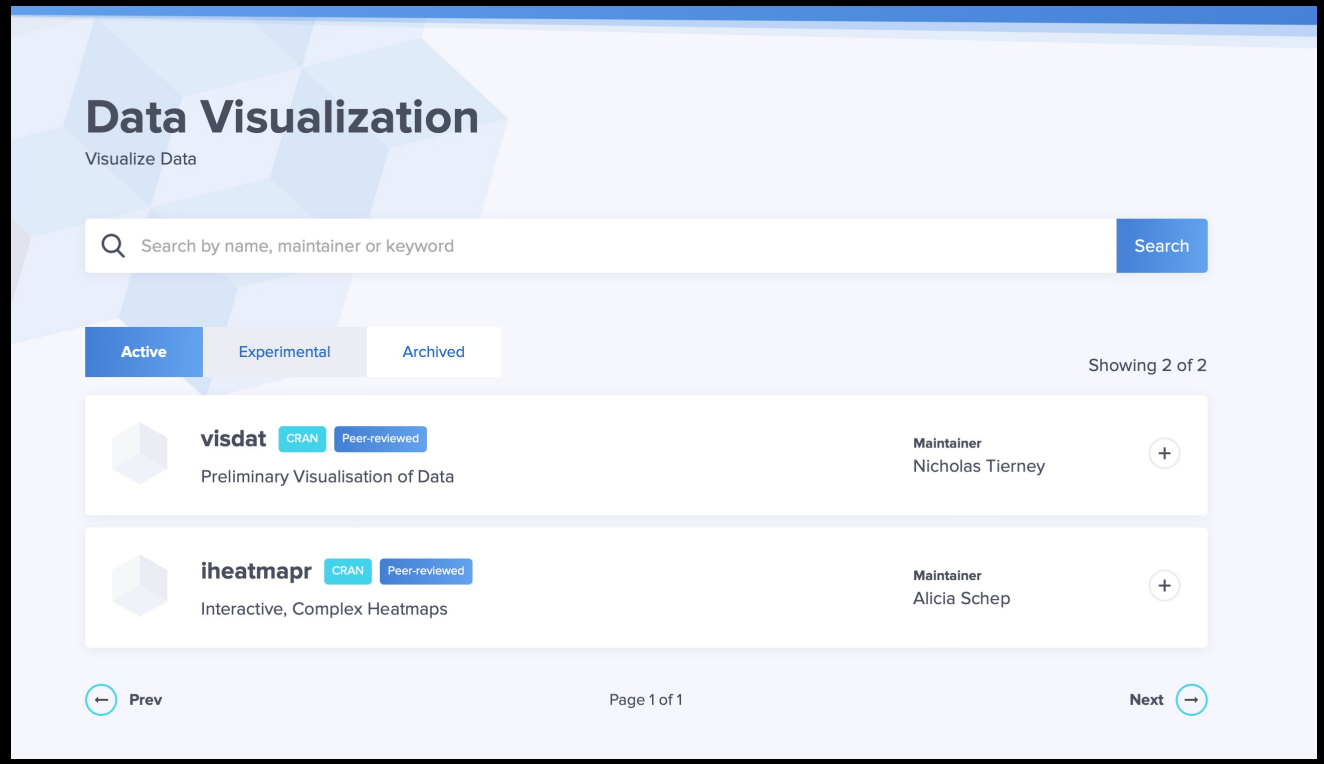
Promoting Software Peer Review

packages contributed by the community undergo a transparent, constructive, non adversarial and open review process

NumFOCUS is a nonprofit that supports and promotes world-class, innovative, open source scientific computing, including rOpenSci.

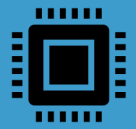
The Journal of Open Source Software

Journal of Open Source Education



Trend:
A dramatic increase in access to an open research and analysis ecosystem

Jupyter Notebooks



Purpose: To support interactive data science and scientific computing across all programming languages.

Funding: Healthy mix of private business, foundation grants, and governments.

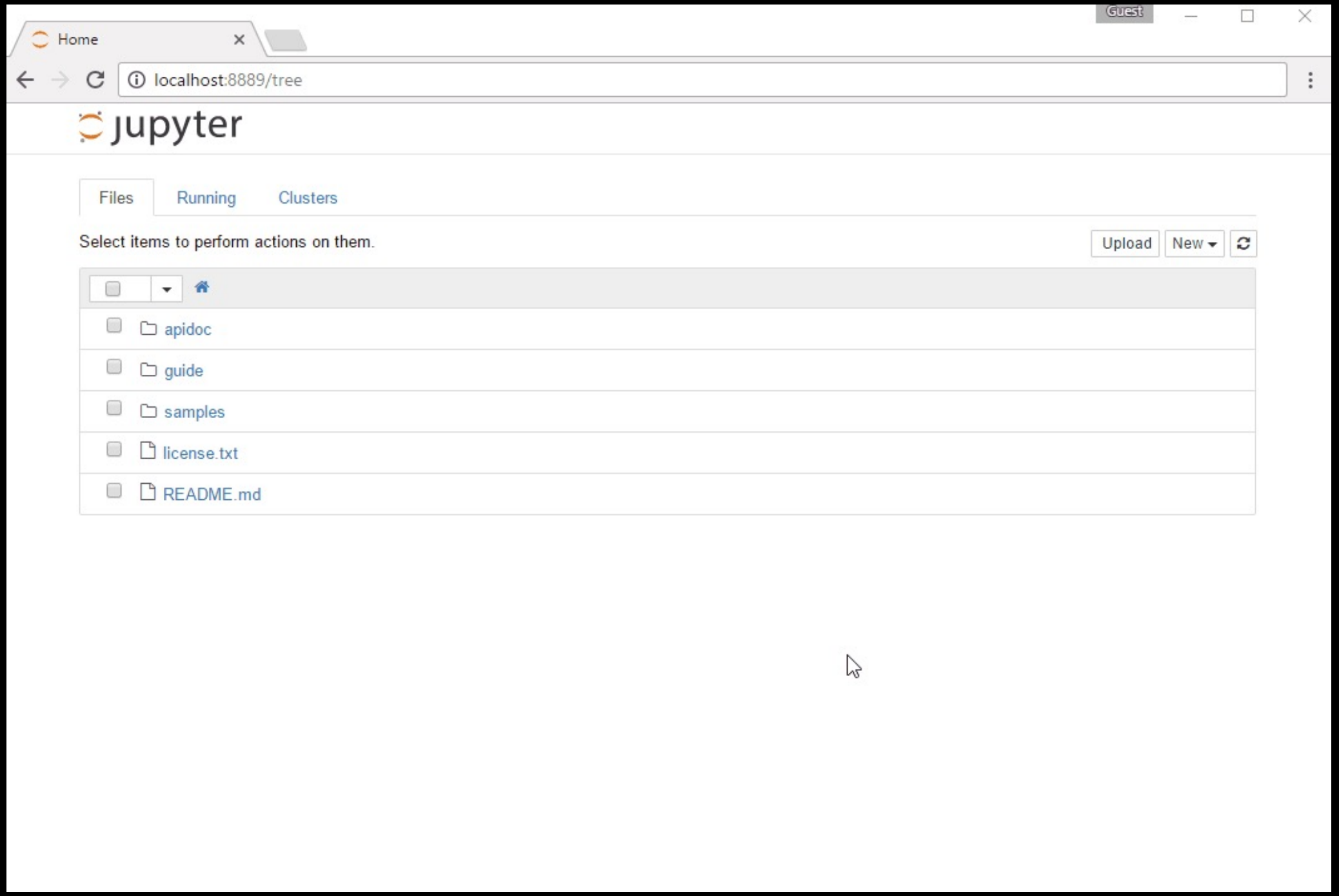
Project Jupyter is a nonprofit organization

Advantages of our toolset

- Python and R Languages
- Data cleaning and wrangling (organizing)
- Data analysis (statistics, calculations, etc.)
- Data visualization and export
- Reproducibility

SPSS and MATLAB

- high $ costs
- language more cryptic
- point and click; error prone; not aware of what it is doing


Guest
Home
localhost:8889/tree
jupyter
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
apidoc
guide
samples
license.txt
README.md

localhost:9999

year

xaxis

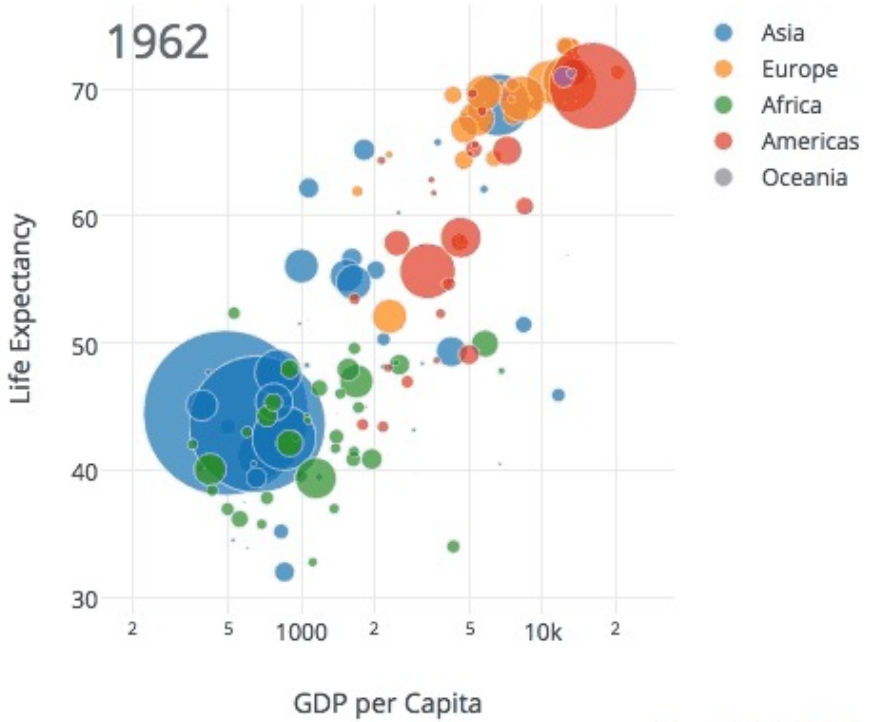
GDP per Capita

yaxis

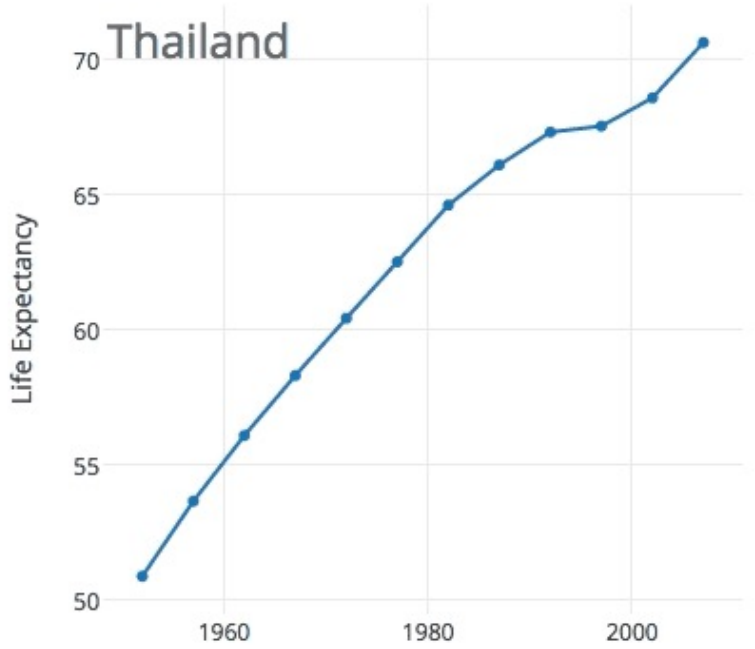
Life Expectancy

size

Population



Play with this data! »



Play with this data! »

Suicide Rate per 100,000 persons (5 year average)

Source: Kansas Information for Communities - Death Statistics (KDHE)

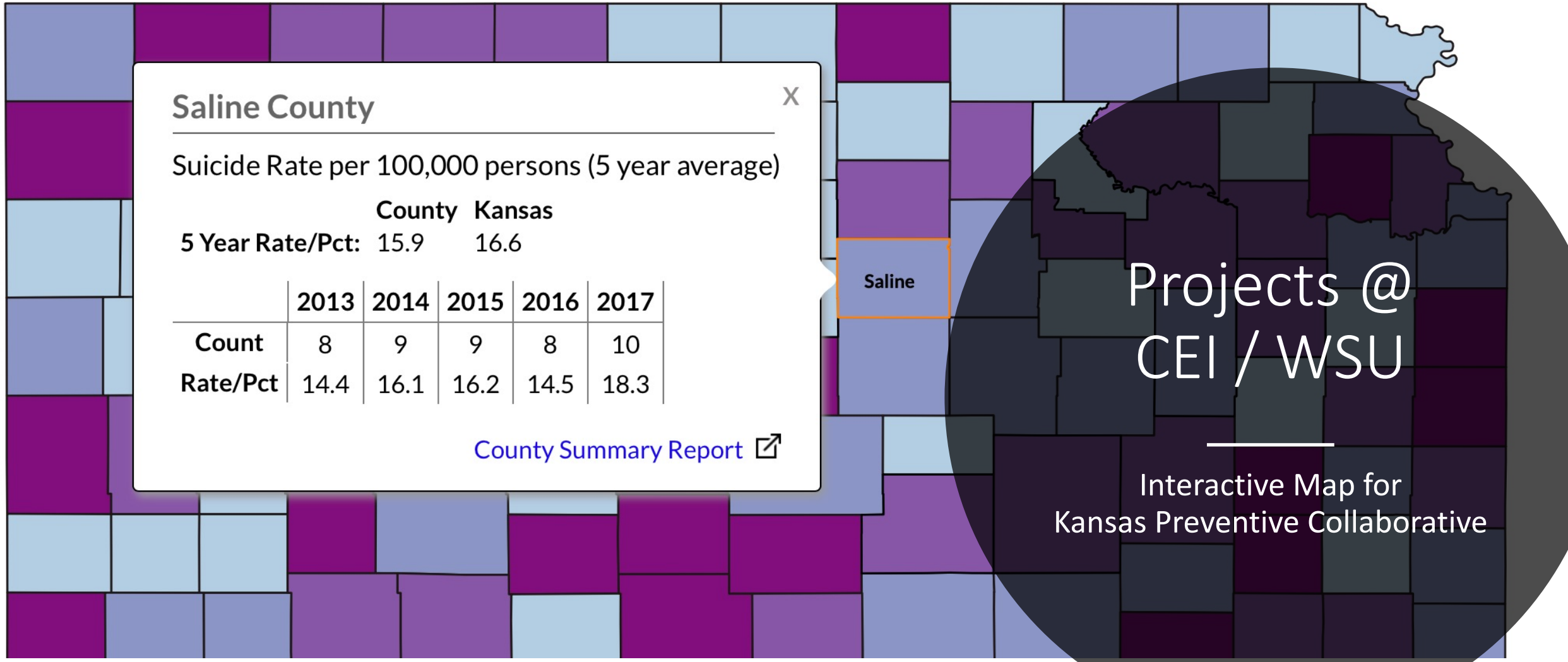
Help

Choose An Indicator

Suicide Rate

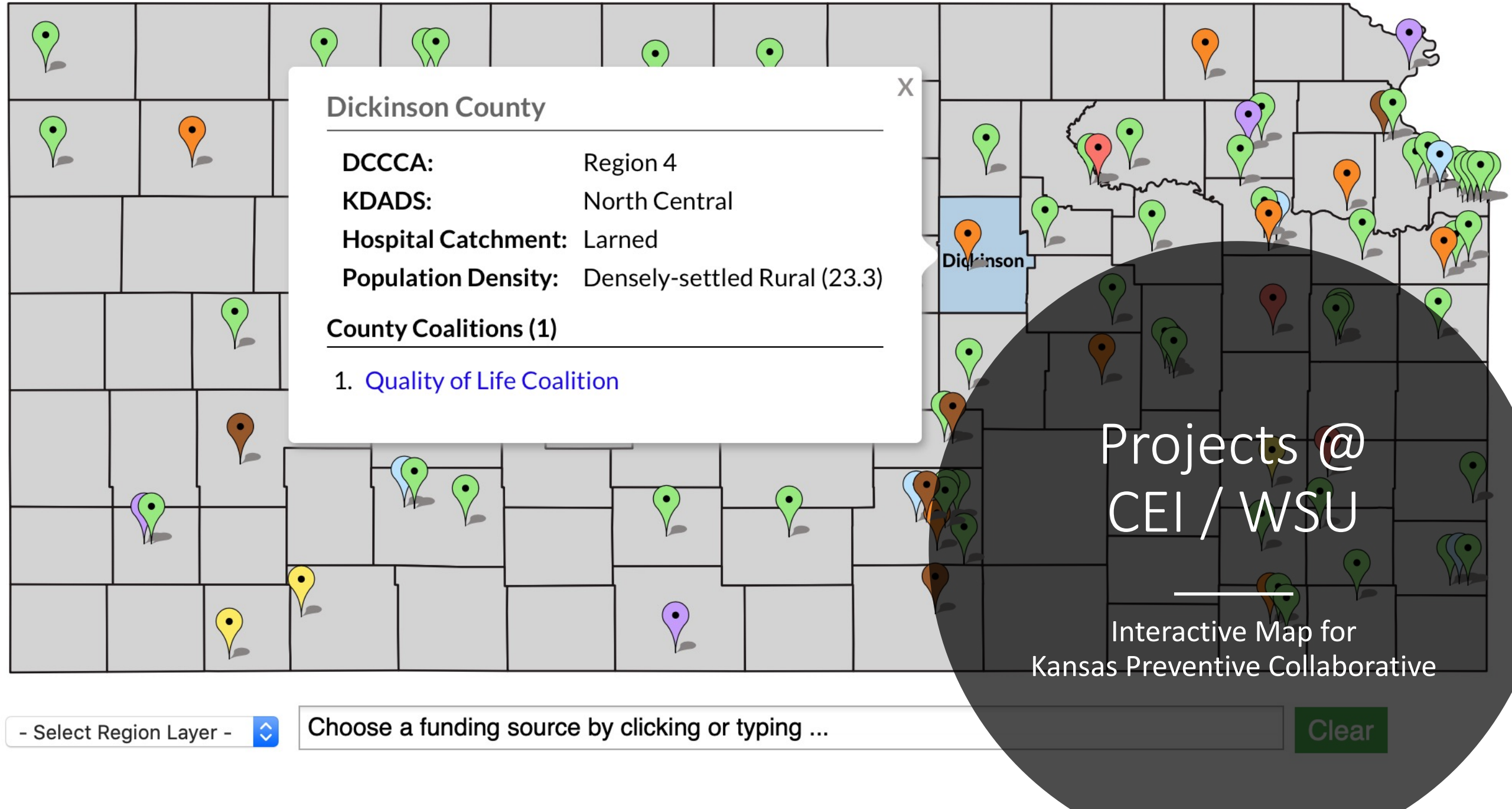
Show

Rate/Pct | Show Coalitions



Projects @ CEI / WSU

Interactive Map for
Kansas Preventive Collaborative

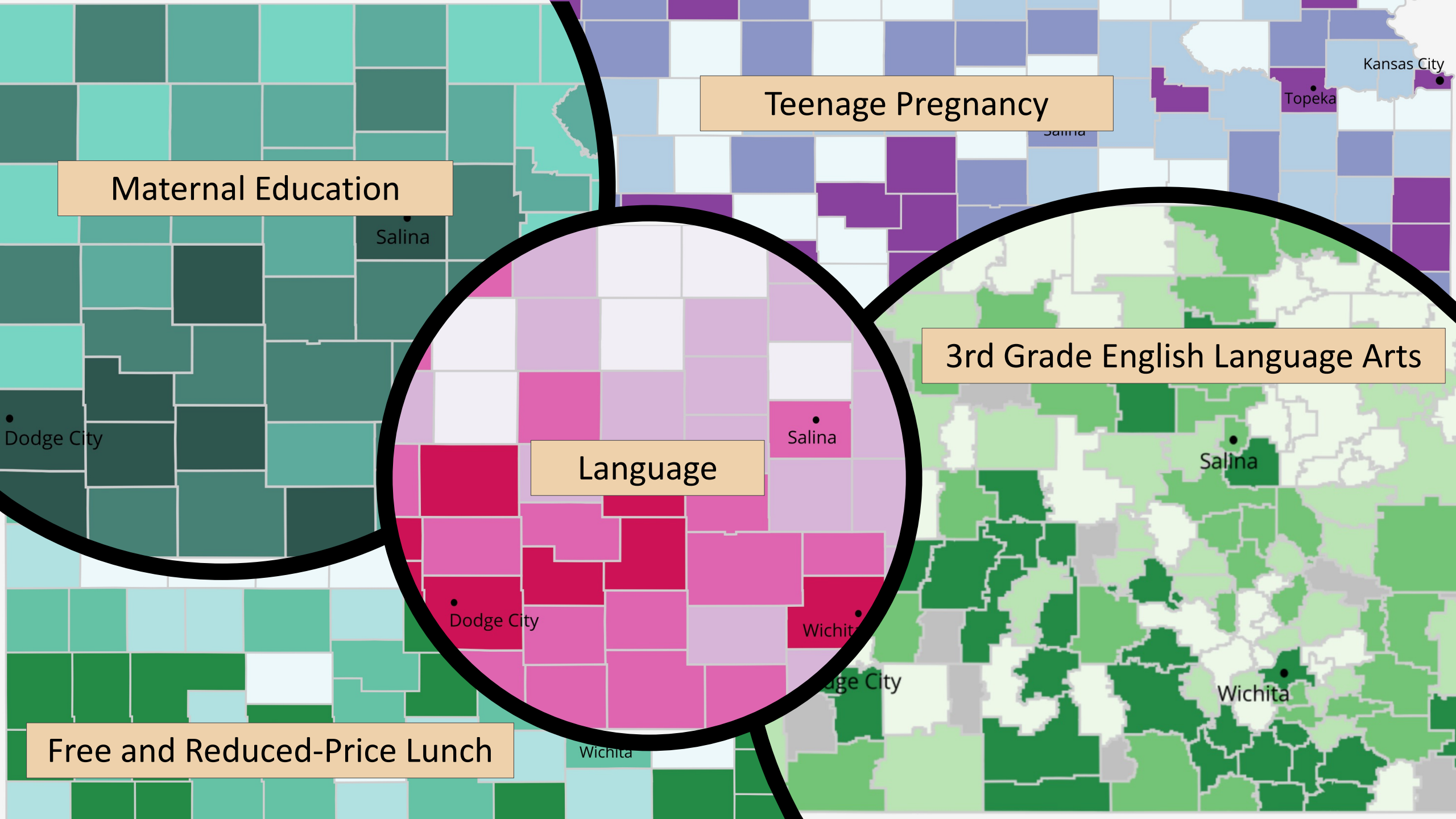

Dickinson County
DCCCA: Region 4
KDADS: North Central
Hospital Catchment: Larned
Population Density: Densely-settled Rural (23.3)
County Coalitions (1)
1. Quality of Life Coalition
Dickinson
Projects @
CEI / WSU
Interactive Map for
Kansas Preventive Collaborative
- Select Region Layer -
Choose a funding source by clicking or typing ...
Clear

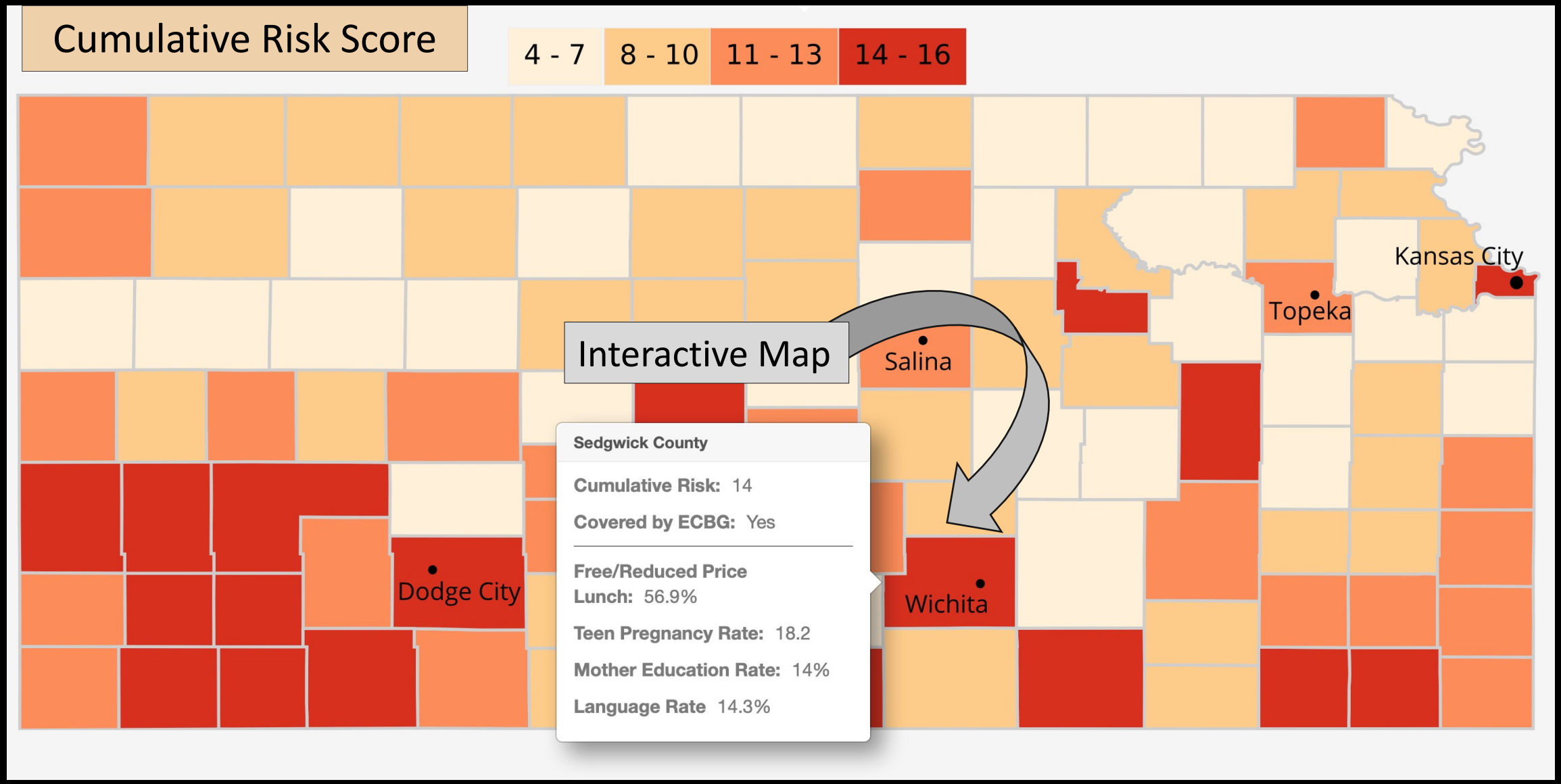
Early Childhood Development
0-5 years old

Center for Applied Research and Evaluation (CARE) @ CEI / WSU

We evaluate data they have collected for us.

We give them information about what kind of impact are they having on the children and families they serve.


Teenage Pregnancy
Kansas City
Topeka
Maternal Education
Salina
Dodge City
Language
Salina
Dodge City
3rd Grade English Language Arts
Salina
Wichita
Free and Reduced-Price Lunch
Wichita


Cumulative Risk Score
4 - 7
8 - 10
11 - 13
14 - 16
Interactive Map
Kansas City
Topeka
Salina
Dodge City
Wichita
Sedgwick County
Cumulative Risk: 14
Covered by ECBG: Yes
Free/Reduced Price Lunch: 56.9%
Teen Pregnancy Rate: 18.2
Mother Education Rate: 14%
Language Rate 14.3%

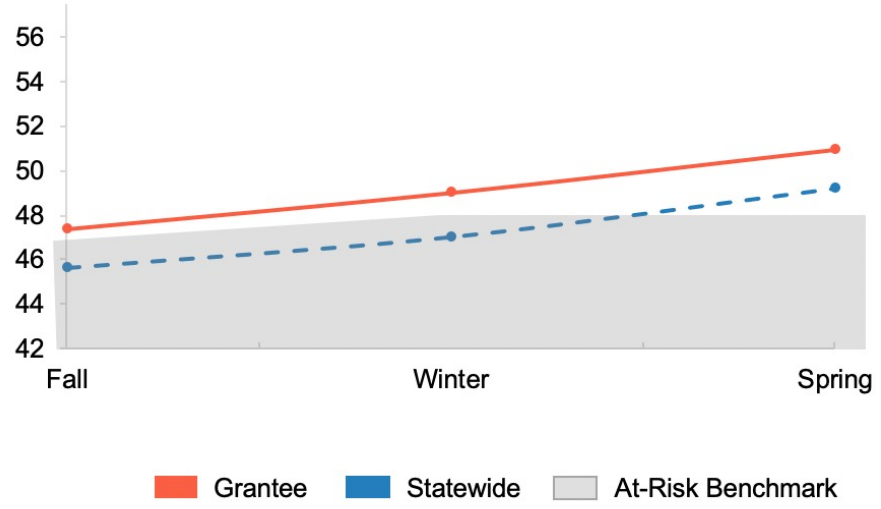
Challenge is to produce reports for various audiences

We don’t produce one report – we design at least 3 levels of presentation

Literacy Average Change Over Time

The graphs below display the average change over time in each subtest. They also show how the Grantee scores compare to scores statewide. The Grantee’s average scores were **above statewide averages** in the spring for all subtests. This Grantee made the largest gains of all the ECBG grantees in the **Picture Naming** and **Alliteration** subtests.

Language Comprehension



Language

Maternal speech abilities and exposure to reading ...

READ MORE

Parent Education

Low maternal education has been found ...

READ MORE

Cognitive Impacts

As a predictor of child IQ, income/poverty is more influential than even maternal education.[28]

Income has been found to be the single highest correlate of academic achievement.[25]

Results

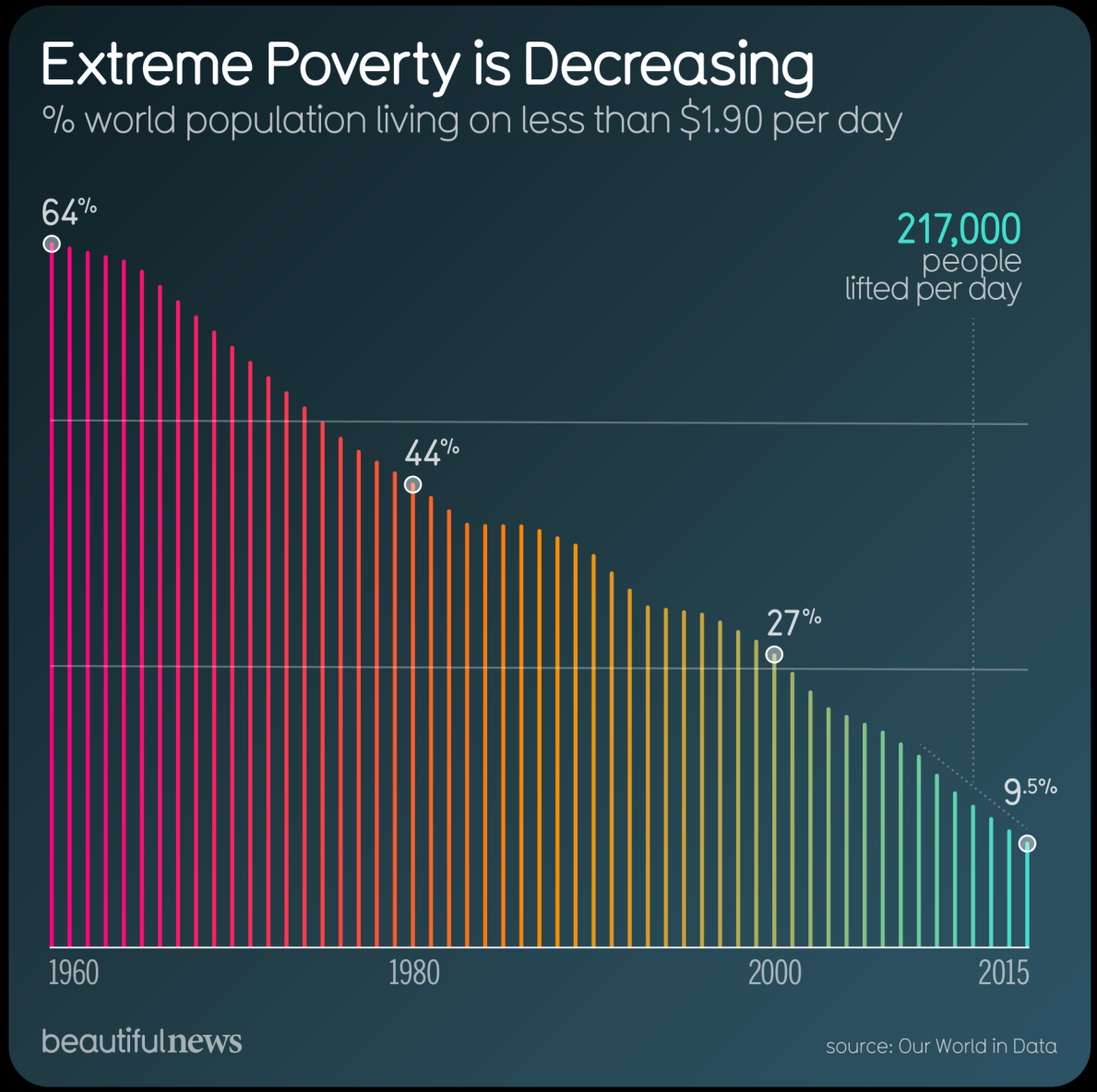
Improving Outcomes

Individual Early Childhood sites (programs) are seeing the **value of data** and are **using it** to improve school readiness levels in children across Kansas

Design Counts

Telling Stories Is Powerful

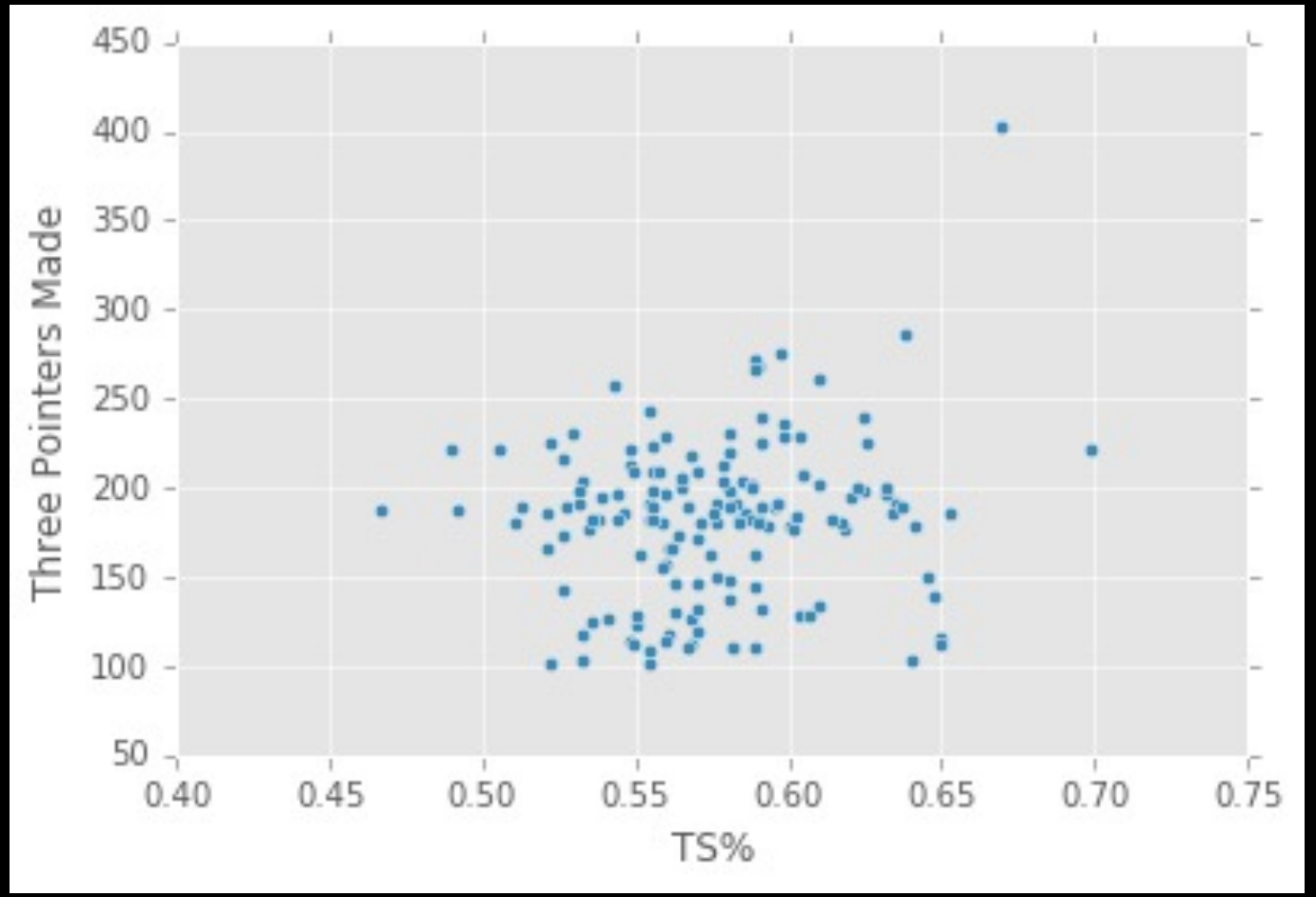
Data Can Be Beautiful



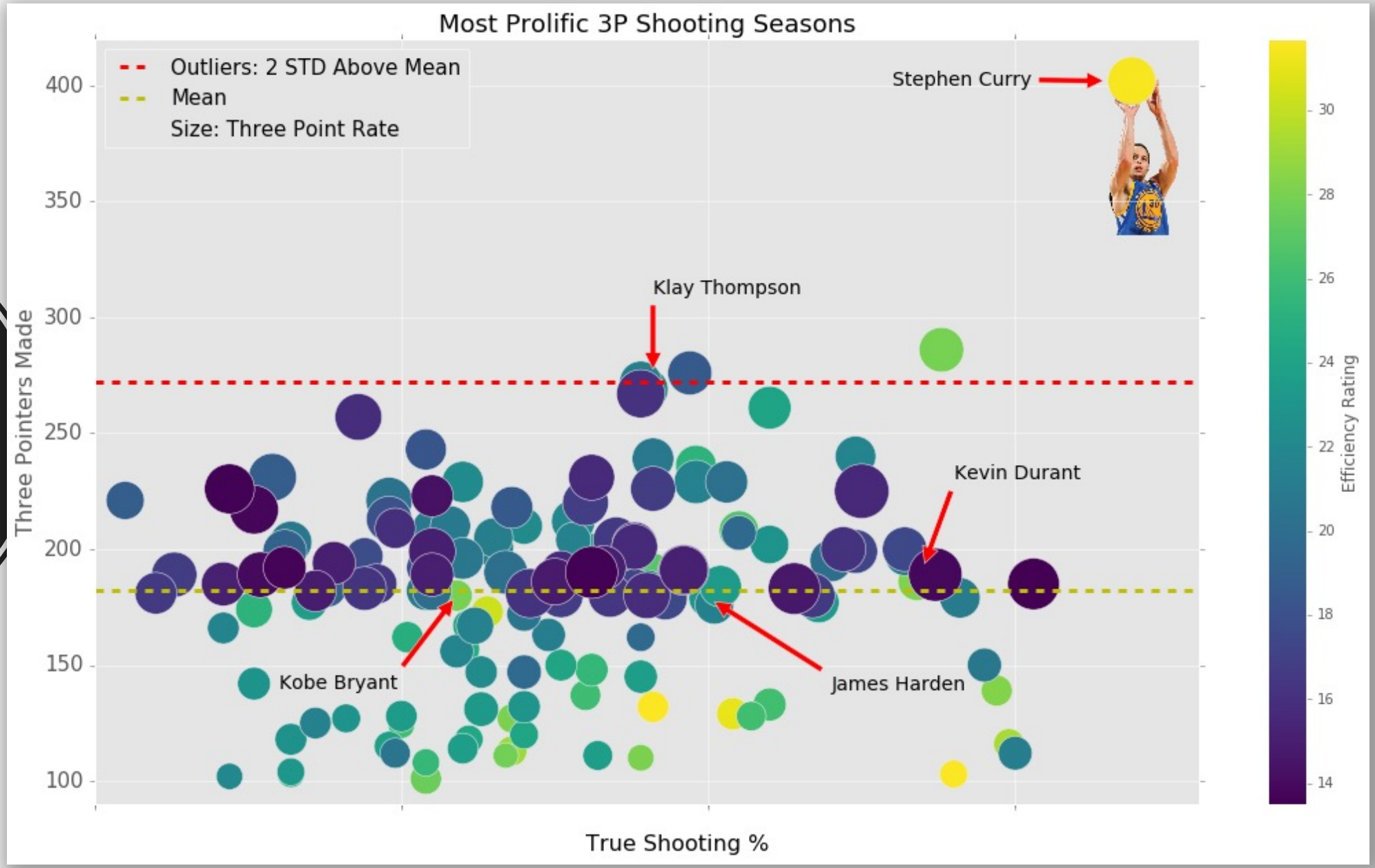
“Scatter Plot” or “Tell a Story” that can appeal to a wider audience?

The Most Prolific 3-Point Shooting Season in NBA History

Stephen Curry 2015-16



AlienOutliers
.com



Acknowledging the Team

Mark Earnest

Scott Wituk, Ph.D.

Tara Gregory, Ph.D.

Cassandra LeBrun-Martin, MA

Lynn Schrepferman, Ph.D.

Wichita State University
Community Engagement Institute

HireMJV.com